

Data Mesh for Smart Cities

Marcin Krystek

A thesis submitted for the degree of
Doctor of Philosophy
Poznań University of Technology

September 2024

© Marcin Krystek 2024

Except where otherwise indicated, this thesis is my own original work.
Large Language Models were utilized to verify linguistic correctness, maintain text
coherence and enhance stylistic quality of the text.

Marcin Krystek
2 September 2024

To my Parents

Acknowledgments

First and foremost I am extremely grateful to my supervisors dr hab. inż. Mikołaj Morzy Prof. PP and dr inż. Cezary Mazurek for their invaluable advice and support during the completion of this research project. Their vast knowledge and experience enabled me to successfully overcome all the challenges I encountered along the way.

I would like to extend special thanks to Bartosz Lewandowski for introducing me to the topics related to Smart City and shedding light on the specific problems that municipalities face in the area of digital transformation, as well as to Juliusz Pukacki for his constructive criticism of the developed concepts and for providing additional motivation to maintain the practical dimension of this research.

I sincerely thank my colleagues at PSNC, designers and developers, with whom I had the pleasure of collaborating during the first implementation of the data platform for the city of Poznań. Your knowledge and professionalism help turn concepts into real products.

I also express my gratitude to my family for their support and perseverance, especially during the final weeks of writing this dissertation.

Abstract

This thesis, produced as part of an industrial PhD program aimed at integrating academic research with practical business applications, examines the implementation of the Data Mesh architecture within the context of Smart Cities. This innovative framework offers a novel approach to managing and utilizing complex data systems, which is particularly relevant to the evolving digital infrastructure of urban environments.

The work is structured to explore the multifaceted relationship between technology, data management, urban development, and business applications. It addresses the fundamental principles of Data Mesh, including its potential to transform data management practices by addressing socio-technical and organizational challenges often overlooked by traditional data platform approaches.

The thesis emphasizes the need for a thorough understanding of Smart City concepts and the necessity of developing a data management strategy. The analysis focuses on how the strategy should address the specific needs and challenges of urban environments like Poznań, and evaluates the practical implications of applying modern technologies, such as Large Language Models, to enhance the functionality of data platforms.

By bridging theoretical insights with practical experiences, this research aims to provide a comprehensive understanding of how innovative data management solutions can be effectively integrated into Smart City initiatives. The ultimate goal is to advance the field of urban data platforms and facilitate their successful deployment in real-world scenarios.

Streszczenie

Niniejsza praca, powstała w ramach doktoratu wdrożeniowego mającego na celu integrację badań akademickich z praktycznymi zastosowaniami biznesowymi, bada możliwość wdrożenia architektury Data Mesh w kontekście inteligentnych miast. To innowacyjne podejście oferuje nowatorski sposób zarządzania i wykorzystywania złożonych systemów danych, co jest szczególnie istotne w kontekście rozwijającej się infrastruktury cyfrowej w środowiskach miejskich.

Podjętowany problem badawczy dotyczy ustalenia wieloaspektowego związku między technologią, zarządzaniem danymi, rozwojem miast a zastosowaniami biznesowymi. W pracy omówione zostały fundamenty koncepcji Data Mesh, w tym ich potencjał do przekształcania praktyk zarządzania danymi poprzez rozwiązanie wyzwań społeczno-technicznych i organizacyjnych, które często są pomijane przez tradycyjne podejścia do projektowania platform danych.

Praca podkreśla potrzebę dogłębnego zrozumienia koncepcji inteligentnych miast oraz konieczność opracowania strategii zarządzania danymi. Przedmiotem analizy jest sposób w jaki strategia powinna adresować specyficzne potrzeby i wyzwania środowisk miejskich, takich jak Poznań, oraz ocena praktycznych implikacji zastosowania nowoczesnych technologii, takich jak wielkie modele językowe, w celu ulepszenia funkcjonalności platform danych.

Łącząc teoretyczne spostrzeżenia z praktycznymi doświadczeniami, niniejsze badanie ma na celu dostarczenie kompleksowego zrozumienia tego w jaki sposób innowacyjne rozwiązania zarządzania danymi mogą być efektywnie zintegrowane z koncepcją inteligentnych miast. Ostatecznym celem jest rozwój dziedziny miejskich platform danych i ułatwienie ich pomyślnego wdrożenia w praktyce.

Contents

Acknowledgments	vii
Abstract	ix
Streszczenie	xi
1 Introduction	1
1.1 Institutional synergy	1
1.1.1 About PSNC	1
1.1.2 About Poznań	2
1.1.3 Role of partners	5
1.2 Data context	6
1.2.1 Poznań Smart City infrastructures and systems	6
1.2.2 Poznań Open Data	8
1.3 City data in the era of generative AI revolution	9
1.4 Thesis	10
1.5 Research framework	10
2 Smart City	13
2.1 Definition of Smart City	13
2.2 Poznań Smart City strategy	14
2.3 Poznań Smart City data assets	16
2.4 Challenges in implementing Open Data platform for Smart City	18
3 Data Mesh	21
3.1 Data platform architectures	21
3.2 Design pattern weaknesses	22
3.3 Data Mesh principles	24
3.4 Data Mesh approach for data platform design	27
3.4.1 Data domains and data sharing	28
3.4.2 Data product	29
3.4.3 Self-serve data platform	30
3.4.4 Federated computational governance	32
4 Smart City data landscape using Data Mesh	33
4.1 Urban data platforms in literature	33
4.2 Smart City Data Platform pillars	35
4.2.1 Technology	35

4.2.2	Usability	37
4.2.3	Sustainability	38
4.2.4	Governance	39
4.3	Basic tenets of Data Mesh in the context of Smart City	40
4.3.1	Sociological level	40
4.3.2	Technology level	42
4.3.3	Infrastructure level	43
4.4	Lessons learned from implementing Data Mesh in a Smart City context	44
4.4.1	Partnership	45
4.4.2	Small steps	45
4.4.3	Understanding	46
4.4.4	Data value	47
4.4.5	Promotion and dissemination	47
4.4.6	Driving scenario	47
4.4.7	Task management	48
4.5	Technical aspects	48
5	Large Language Models in the service of Open Data platform for Smart City	53
5.1	Motivation	53
5.2	Large Language Models in structured data analysis	55
5.3	Experiment 1: identify and label concepts	56
5.4	Experiment 2: generate human like descriptions	57
5.5	Experiment 3: support in decision making	64
5.6	Experiment conclusions	65
6	ODSC Deployment	69
6.1	Deployment approach in Poznań	69
6.1.1	System layers	69
6.1.2	Data management	71
6.2	Poznań Open Data portal	72
6.2.1	Data products catalog	72
6.2.2	Conversational interface for open data	75
6.3	Data analytics	77
6.4	ODSC Concierge	79
6.4.1	Toolset	79
6.4.2	Dataset reverse engineering	81
6.4.3	Data characterisation	82
7	Conclusion	87

Introduction

This thesis, written as part of the industrial PhD program designed to bridge the gap between academia and business, researchers the Data Mesh architecture for Smart Cities, an innovative approach to managing and utilizing complex data systems in our progressively digital urban landscapes. The program aims to foster cutting-edge research and facilitate the direct transfer of this knowledge into the business sector. The first chapter of the thesis is organized as follows. The section 1.1, provides an overview of the Poznan Supercomputing and Networking Center (PSNC), the city of Poznań, and the roles of partners in the development of Smart City solutions. The next section, 1.2, examines Poznań’s Smart City strategy and its Open Data initiatives. Section 1.3 outlines the possible usage of latest AI achievements in context of Smart City data. Finally, the 1.4 defines the research problem, the results of which will have a direct impact on the implementation of the final solution. Embedding the research within the context of a specific field of study is given in 1.5.

1.1 Institutional synergy

1.1.1 About PSNC

The Poznań Supercomputing and Networking Center (PSNC), affiliated with the Institute of Bioorganic Chemistry of the Polish Academy of Sciences, stands as a prominent node within the European Research Area, particularly in the realm of scientific IT infrastructure [1]. As a leading research and development center in the field of information and communication technologies (ICT), PSNC has played an important role in the conceptualization and construction of critical infrastructure such as the Metropolitan Network POZMAN, the High-Performance Computing Center, and the national broadband network PIONIER, all of which are maintained and continuously expanded by PSNC.

PSNC is integral to the global research and development landscape, engaging primarily in projects under the European Union Framework Programmes while also collaborating with over a thousand partners worldwide. To date, PSNC has been involved in 224 such projects, coordinating 20, and over its 25+ years of operation, has implemented 350 research and structural initiatives, leading 48 of them—amounting to one-seventh of all PSNC’s projects [2].

As an institution with significant developmental potential, PSNC has established broad international collaborations as a direct result of its success in securing and executing projects. PSNC's initiatives involve partnerships with nearly a thousand entities from over 60 countries across six continents. These projects span a wide spectrum of IT-related domains, including next-generation networks, grids, portals, digital libraries, IT energy efficiency, climate and environmental monitoring, renewable energy, digital humanities, personalized medicine, smart agriculture, Industry 4.0, astronomy, bioinformatics, Big Data analytics, artificial intelligence, New Media, education, Smart City initiatives, and cybersecurity. The scope of PSNC's work continues to expand, consistently addressing emerging technological challenges.

Since its inception in 1993, PSNC has been dedicated to its mission of "integrating and developing information infrastructure for science" [1]. Its diverse range of activities reflects a commitment to innovation driven by ICT, resulting in implementations that advance digital science and industry.

PSNC positions itself at the intersection of science and business, where innovation impacts society, and research converges with the responsible development of new technologies. Digital science, with its transformative approach to research through the application of ICT, has become more open, creative, global, and collaborative, fostering closer ties with society. From the beginning, PSNC has embraced new research and development challenges, seeking to create optimal solutions through experimental projects, applying scientific expertise to translate results into effective commercial prototypes [3].

The declining cost of computing, the rise of the Internet as a communication medium, the growth of mobile connectivity, and the increasing influence of social networks and digital platforms have profoundly affected economies and reshaped the operations of businesses, public institutions, and personal life. PSNC addresses the challenges of the digital economy by developing innovative digital prototypes in areas such as next-generation networks, the Internet of Things, Artificial Intelligence, and Big Data, thereby fostering innovation, testing modern products, and building extensive collaborations with industry.

In the realm of social innovation, PSNC is dedicated to enhancing the quality of life for individuals, organizations, and communities by introducing novel solutions, ideas, products, services, or models that address a wide array of social needs. These efforts are implemented on a technological foundation across various social groups, with a focus on deploying new ICT applications, raising awareness of contemporary digital opportunities, and actively combating digital exclusion.

1.1.2 About Poznań

Poznań, one of Poland's most historically prominent cities, is located in the heart of the country's west-central region, strategically situated along the Warta River. It serves as the capital of the Greater Poland Voivodeship and stands as a regional center of culture, economy, and academia. According to report [4], with a population surpassing half a million, Poznań is characterized by a rich demographic diversity.

The city's youthful and vibrant population is largely shaped by its role as a major academic hub, home to numerous universities and institutions of higher learning. This academic presence infuses the city with a dynamic cultural energy and a progressive atmosphere.

Poznań's central mission is to cultivate an environment where all residents can actively participate in the co-creation of the city. This mission underscores the principles of shared responsibility, emphasizing a collective commitment to aligned aspirations while ensuring support for the vulnerable and disabled members of the community.

The city's industrial potential is underpinned by a diversified economic framework, which encompasses several vital sectors that not only provide significant employment opportunities but also offer considerable prospects for future development. Among these sectors, the automotive industry stands out, with major companies such as Volkswagen operating large manufacturing facilities in Poznań. This industry drives local employment and creates a network of suppliers and ancillary services, significantly boosting the region's economic activity.

Poznań's strategic location, bolstered by excellent transportation links including highways, rail networks, and proximity to international borders, has established the city as an important logistics and distribution center. This sector plays a vital role in the city's economic expansion, offering numerous jobs and strengthening broader economic growth. The city's information technology (IT) and business process outsourcing (BPO) sectors are also rapidly advancing, benefiting from the strong academic foundation that supplies a highly educated workforce. These industries are increasingly integral to Poznań's economy, providing essential services in software development, IT, and outsourcing.

The chemical and pharmaceutical industries are another cornerstone of Poznań's economic landscape, with several major companies operating in these fields and contributing significantly to employment and export revenues. Additionally, Poznań has a well-established tradition in the food processing industry, which continues to thrive due to growing demand both within Poland and internationally.

The creative industries and tourism are also integral to Poznań's economic fabric. The city actively encourages the growth of creative sectors, including design, media, and culture, while its rich cultural and historical heritage supports a flourishing tourism industry. These industries generate substantial employment and stimulate local businesses, with innovation and a skilled workforce driving their continued development. The growth in medical tourism and advanced healthcare technologies is driven by city's healthcare and wellness facilities focused on improving infrastructure and services. Finally, the creative and cultural sectors contribute significantly to Poznań's vibrant economy, with the city supporting a range of creative enterprises and cultural initiatives. These sectors, particularly digital media and creative startups, offer high development potential, fostering innovation and creating employment opportunities across the city's diverse economic landscape.

In terms of business services, Poznań offers a wide array of opportunities across several sectors, each with significant potential for growth. The financial services sec-

tor, including banking, insurance, and investment services, is well-established, with major institutions providing critical support to local businesses and international trade. The city's consulting services are also robust, covering management consulting, legal advisory, and business strategy, with firms offering expertise in navigating regulatory challenges, optimizing operations, and driving strategic initiatives. The demand for such specialized services, particularly in areas like sustainability and digital transformation, is expected to increase.

Educational services in Poznań are strong, with numerous universities and vocational schools offering a diverse range of programs. The city places a strong emphasis on lifelong learning and professional development, with opportunities for growth through enhanced collaboration between academia and industry, especially in science, technology, engineering, and mathematics disciplines.

Poznań's future development is guided by a strategy [5] that emphasizes several key priorities aimed at enhancing the city's overall quality of life and global standing.

A primary focus is on cultivating a "Green, Mobile City" which aims to expand green spaces, encourage eco-friendly transportation, reduce reliance on cars, and enhance the appeal of public transport. By developing urban green areas and improving infrastructure for cycling and walking, this initiative seeks to improve air quality, alleviate traffic congestion, and promote healthier lifestyles. To effectively implement this, data on current traffic patterns, pollution levels, public transport usage, and the distribution of green spaces is essential, along with insights into residents' transport preferences.

Another strategic priority is fostering "Smart Entrepreneurship". This initiative is dedicated to supporting start-ups, creative industries, and small and medium-sized enterprises (SMEs) by promoting innovation and collaboration between education, science, and business sectors. The goal is to generate new job opportunities, cultivate a culture of innovation and attract investment, thereby creating a vibrant local economy that benefits residents. Data on the current business landscape, innovation metrics, and the effectiveness of education-business partnerships will be crucial in driving this priority forward.

"Community and Social Dialogue" represents a commitment to enhancing social cohesion, improving public safety, and supporting a strong sense of identity and shared responsibility among residents. Strengthening social integration and ensuring public safety are expected to improve the quality of life, making residents feel more connected to their community and more engaged in the city's development. To support these efforts, accurate data on social integration, public safety statistics and resident participation in community activities is necessary.

The priority of building a "Strong Metropolis" focuses on raising Poznań's international profile, developing its metropolitan area and intensifying collaboration with other cities and regions. Enhancing Poznań's role on the global stage is expected to create better job opportunities, improve infrastructure and attract both tourism and investment. Data on international collaborations, economic performance and infrastructure development, as well as comparisons with other metropolitan areas, will help identify areas for growth and improvement.

Lastly, the priority of creating "Friendly Residential Estates" is aimed at improving the living conditions in residential areas by ensuring better accessibility and enhancing local services and amenities. The objective is to increase residents' satisfaction and well-being through improved housing, local services and recreational facilities. Data on current housing conditions, the availability of local services and resident satisfaction with their living environments will be vital in informing and guiding these enhancements.

The fulfillment of these priorities, the achievement of goals and the attainment of sustainable development are therefore inextricably linked to the city's ability to provide well-organized and efficient access to the data it possesses. Appropriate policies addressing the openness of urban data, along with an organizational culture that supports data sharing, must be reinforced with the right tools to achieve these goals. However, developing a coherent IT system concept will encounter many practical challenges, ranging from the city's complex organizational structure and the lack of a unified IT infrastructure to insufficient competencies in using contemporary tools and technologies. The direct implementation of technological solutions known from literature and practice in such an uncertain and diverse environment carries a high risk of failure understood as the tool's inability to support the city's key processes and to provide a foundation for developing applications for residents. Therefore, there is a need for research that will result in the development of a modern and flexible approach for managing large volumes of heterogeneous data in a Smart City environment.

1.1.3 Role of partners

The collaboration between Poznań City Hall and the PSNC serves as a driving force in advancing the Smart City strategy for Poznań. Both entities contribute uniquely to this partnership, playing complementary roles that drive the city's digital transformation. The collaboration details and the role of each partner were elaborated on further in [6].

Poznań City Hall assumes the mantle of administrative and strategic leadership, directing the Smart City initiatives in alignment with the broader objectives outlined in the City Development Strategy leading up to 2030. The City Hall not only sets the strategic course but also ensures that these initiatives are integrated into the city's overarching development framework. Moreover, the City Hall plays a key role in providing access to essential datasets that fuel the development of smart applications and services. It is also responsible for crafting policies that govern data openness, ensuring that data sharing complies with public information regulations and promotes transparency. In terms of service integration, the City Hall is instrumental in embedding new technologies into the existing urban infrastructure. It actively engages with citizens to ensure that these services are responsive to their needs, as evidenced by the development of electronic services like the Business Activity Register and school recruitment systems.

On the other hand, PSNC contributes its technological expertise and plays a cru-

cial role in the development of the city's IT infrastructure. PSNC's involvement in constructing the fiber-optic network and the Municipal Multimedia Guide (MIM) has been fundamental to the success of many Smart City applications. Furthermore, PSNC is at the forefront of developing and supporting various IT applications that enhance urban living, from transport management systems to environmental monitoring tools, which provide residents with real-time information and contribute to improved city management. The center's research and innovation capabilities ensure that cutting-edge advancements in information and communication technologies are harnessed within the Smart City projects, incorporating developments in areas such as machine learning, big data processing, and data visualization.

The contributions from both Poznań City Hall and PSNC are mutually reinforcing. The City Hall facilitates the necessary access to critical data and establishes policies that support the seamless adoption of smart technologies across the city. Concurrently, PSNC develops and maintains the technical infrastructure, creating innovative applications that leverage the data provided by the City Hall to address real-world urban challenges. This synergy enhances the city's efficiency in managing resources and services, leading to improved administrative operations and increased citizen satisfaction.

The benefits of this partnership are manifold. Poznań City Hall experiences heightened efficiency in its operations, along with national and international recognition for its pioneering efforts in Smart City initiatives. For PSNC, the partnership offers a valuable opportunity to apply its cutting-edge research in a practical, real-world environment, enabling collaborative innovation that addresses the unique challenges of urban living.

The partnership between Poznań City Hall and PSNC stands as a cornerstone of Poznań's Smart City strategy, blending administrative acumen with technological prowess to create a city that is more efficient, responsive and also increasingly livable.

1.2 Data context

1.2.1 Poznań Smart City infrastructures and systems

The strategy of the city of Poznań regarding the implementation of Smart City IT solutions was the subject of previous considerations [6].

The general strategy for implementing the Poznań Smart City concept revolves around harnessing advanced information and communication technologies (ICT) and innovative management methods to improve the quality of life for its residents. This participatory approach leverages the potential of social capital by involving residents in the creation of smart urban solutions, ensuring that the initiatives align with their needs and aspirations.

The strategy begins with the establishment of a robust technological and data infrastructure, which forms the backbone of the Smart City. This includes the development of extensive fiber optic networks, the implementation of wireless zones, and the deployment of a CCTV system. The fiber optic network, consisting of the POZ-

MAN metropolitan area scientific network and the network connecting the city's monitoring cameras, provides high-speed connectivity essential for real-time data transmission and processing. The Wireless@Poznań Zone ensures widespread internet accessibility, covering key areas like the city center and major commuting routes. The CCTV system, although primarily used for security, also contributes valuable data for urban management and research projects.

The Intelligent Transport System (ITS) is a critical component of Poznań's Smart City infrastructure, aimed at optimizing transportation through real-time monitoring and data sharing. ITS facilitates efficient traffic management, improves public transportation services, and provides information on parking availability, thereby enhancing mobility for residents. Additionally, the PEKA system, which manages electronic cards for various city services, integrates transportation data with other civic amenities, promoting seamless service delivery and user convenience.

In tandem with these infrastructural developments, Poznań has introduced a range of smart applications designed to improve urban living. These applications, such as SmartZOO, Atmosphere for Poznań, and Mobile Poznań, offer interactive and real-time information to residents, enhancing their engagement with city services. SmartZOO acts as a mobile guide for zoo visitors, providing educational content, while Atmosphere for Poznań monitors air quality and informs residents about environmental conditions. Mobile Poznań consolidates public transport timetables, parking information, and cultural events, making it easier for residents to navigate the city.

A crucial element in the successful implementation of Poznań's Smart City strategy is the availability and quality of data. High-quality data underpins all Smart City initiatives, enabling informed decision-making, efficient resource allocation and effective service delivery. Data from various sources, such as transportation systems, environmental sensors and public service databases, needs to be accurate, up-to-date, and accessible. This ensures that the applications and systems built on this data provide reliable and useful information to both city administrators and residents.

The collaboration between Poznań City Hall and the Poznań Supercomputing and Networking Center (PSNC) exemplifies the integration of data and technology in the Smart City strategy. City Hall provides administrative leadership, policy direction, and data resources, while PSNC contributes technological expertise and infrastructure. Together, they create a synergistic environment where data is leveraged to develop practical and impactful smart applications.

Data quality is paramount in this context. High-quality data ensures that the insights derived are accurate and actionable, leading to better outcomes. For instance, in the ITS, real-time traffic data must be precise to effectively manage congestion and improve public transport schedules. Similarly, accurate air quality data in the Atmosphere for Poznań application is essential for making informed decisions about public health interventions.

The strategy also highlights the importance of continuous data monitoring and evaluation. By regularly assessing the performance of Smart City initiatives through data analysis, Poznań can identify areas for improvement and make necessary adjust-

ments. This iterative process ensures that the city's smart solutions remain effective and responsive to changing conditions and resident needs.

In conclusion, the implementation of the Poznań Smart City concept is grounded in the development of a technological infrastructure and the strategic use of high-quality data. This approach enhances urban management, improves public services, and fosters a participatory environment where residents are actively involved in shaping their city's future. The availability and quality of data are critical to the success of these initiatives, providing the foundation for informed decision-making, efficient resource utilization, and continuous innovation. Through this data-driven strategy, Poznań aims to create a more livable, sustainable, and resilient urban environment for its residents.

1.2.2 Poznań Open Data

The implementation of the open data platform in the city of Poznań is driven by two main motivations. The first stems directly from the obligations imposed on the city administration by legal acts in force in Poland. The foundation for all actions in this area is the Public Information Access Act of September 6, 2001. Additionally, the legal framework for making data available and reusing it is defined by the Act of February 25, 2016, on the Reuse of Public Sector Information, which incorporates the directives of the European Parliament and Council 2003/98/EC of November 17, 2003, on the reuse of public sector information, as amended by Directive 2013/37/EU of June 26, 2013. Both of these legal acts regulate the obligations of the city administration regarding ensuring public access to the data under its jurisdiction and the scope of the data that should be made available.

The second motivation arises from the desire to systematize the data management process within the City Hall. The city's previous efforts in implementing the Smart City concept focused on deploying IT systems dedicated to specific applications. However, these systems operate independently, and access to them is structurally challenging. Moreover, only a few initiatives have been undertaken within the City Hall to publish specialized data, such as those related to spatial planning. As the city aspires to further digitization and seeks to create conditions for the development of digital services for residents, the need has arisen to develop a cohesive solution that addresses the needs of both officials and external users. Leveraging the potential of data, particularly the knowledge contained within it about the functioning of the city, will form the basis for making better-informed decisions related to ongoing management and planning. An indirect goal of the implemented open data platform is to conduct a detailed inventory of datasets that could play a crucial role in the aforementioned decision-making process as well as act as an accelerator for the development of digital services. In many cases, it is necessary to establish actual data ownership by identifying a clear owner responsible for maintaining and ensuring the quality of the data. This is also an excellent opportunity to develop a consistent and systematic process for data sharing. Beyond the technical aspects, the essence of this process will primarily involve policies on data access, quality maintenance, and

consistency.

The need to create a unified data platform for the city is a result of both legal requirements and the growing awareness of the potential for data utilization. This presents a unique opportunity to propose a solution that will contribute to a digital breakthrough in city management and establish a foundation for the development of digital services for residents.

1.3 City data in the era of generative AI revolution

As we begin to consider how a data platform for the city of Poznań should be constructed, it is essential to examine the broader technological and organizational context. Data processing technologies have been evolving since the 1960s, with each successive technological breakthrough bringing a significant increase in processing efficiency by removing barriers related to computational scalability. Increased computing power and lower memory costs have allowed for greater flexibility in data storage formats and the degree of data normalization. Previous efforts to ensure a globally consistent data model are being replaced by an approach that creates dynamic models generated to meet the needs of a specific analytical scenario.

In recent years, we have witnessed the rapid development of technologies in the fields of machine learning and artificial intelligence. Since 2018, Large Language Models (LLMs) have been made available for widespread use, offering unique capabilities in data processing. For the first time in history, a tool has emerged that allows for the processing of text, structured data, and semi-structured data in a way that gives the impression that the tool "understands" the informational content contained within the data. Importantly, both the data itself and its structure can be interpreted. The ability of LLMs to highly abstractly model the reality presented to them for interpretation, as well as their ability to generalize and find connections with concepts defined outside the presented data set, offers enormous potential for use in the context of creating a data platform.

The goal of this doctoral project is therefore to explore areas where the use of artificial intelligence (AI), particularly LLMs, can bring significant value in solving complex and highly practical problems related to the creation and maintenance of a data platform for the city.

We are aware of the current state of knowledge and recommendations presented in publications such as [7], [8] regarding how to create urban data platforms. The confrontation of these literary considerations with the practical experiences of PSNC and the awareness of the specific nature of the platform being developed for the city obliges us to take a critical look at the solutions used to date. We recognize aspects in them that are not addressed, and the omission of which may contribute to the failure of achieving the long-term goal of supporting the city's sustainable development.

The synergy resulting from the close cooperation between the city of Poznań and PSNC compels us to seek solutions that go beyond the patterns described in the literature. Fulfilling the city's ambition to achieve full digital transformation and the

objectives that fall within the concept of Smart City 4.0 (defined in [9] and [10]) requires an innovative approach. At the same time, the rapid pace of technological development necessitates the search for solutions that are stable and will provide technological advantages for many years. Otherwise, by the time the platform is fully implemented, it may exhibit signs of technological debt. To meet these challenges, we propose in this work the concept of an Open Data platform for Smart Cities (ODSC). It combines an innovative approach for creating data platforms with the latest technological advancements in artificial intelligence and Large Language Models.

1.4 Thesis

One of the overarching goals of ODSC is to support the sustainable development of the city by creating the technical and organizational conditions necessary for the sharing and reuse of datasets managed by the city administration. In addition to internal use for the needs of the city administration and urban management processes, one of the main consumers of this data will be services and applications designed to meet the vital needs of residents. It is impossible to predict all the potential ways the data might be used. Nevertheless, we know that for datasets to be utilized, they must first be discovered and then correctly interpreted. Therefore, it is crucial to provide additional information that includes the characteristics of the dataset, its broader context and potential applications, as well as technical aspects related to the data structure. In practice, datasets are often lacking in detailed documentation that could be used to prepare such metadata. We believe that the absence of key metadata can be supplemented automatically or semi-automatically by leveraging LLMs. Presenting representative values from a dataset to an LLM, along with providing a minimum context derived from the data structure, should be sufficient to generate a valuable and human-understandable description of the dataset and its broader context.

1.5 Research framework

The ODSC concept, due to its nature, is a multifaceted task. The proposed solution will operate in an environment with many diverse stakeholders, requiring the reconciliation of often conflicting requirements regarding the platform's assumptions and functioning. Stakeholders will play a crucial role not only in defining the requirements but also in continuously verifying the implemented elements of ODSC. This also applies to the design phase, where a new model for data management and perception will be developed through theoretical exercises. A key aspect will be balancing the representation of stakeholders' interests so that ODSC ultimately supports the overall goals expressed in the sustainable development strategy of the city of Poznań.

The preparation of the ODSC platform's assumptions will result from a research

process. However, the practical implementation of ODSC will be a long-term and iterative process. It will involve not only launching the appropriate infrastructure and providing the necessary software. The key to the long-term success of the developed solution will be sociological and organizational changes. These will guarantee the proper use of ODSC and the ultimate achievement of its set goals.

One very important, though difficult to capture, aspect of the designed platform is risk management. This involves addressing fundamental issues regarding the potential of the proposed solutions to maintain operational efficiency in the long term during the project phase. Additionally, the competence of individual stakeholders and the implementation of their business processes must also be taken into account. How ODSC's implementation will impact the execution of these processes will be crucial for its acceptance and long-term maintenance. Furthermore, the fact that public funds are being invested in the development of the platform is not insignificant. All design decisions should minimize the risk of ODSC not being able to support the sustainable development process. Otherwise, the allocation of funds for the development and implementation of ODSC could be considered wasteful.

Although the ODSC concept tries to take a broader view of the issue of creating data platforms beyond just the technological perspective, these aspects should also be appropriately addressed at the implementation stage. This includes infrastructure issues, covering the target runtime environment, virtualization technologies, management of the software development and deployment process, and specific analytical tools supporting the automation of data processing and the delivery of valuable knowledge embedded in the data.

Considering the range of issues that must be addressed to successfully implement a data platform for the city, we are convinced that development and implementation of the ODSC concept requires a systematic approach defined by [11] as systems engineering - a "technique of using knowledge from various branches of engineering and science to introduce technological innovations into the planning and development stages of a system."

Smart City

When embarking on a discussion about creating a data platform for a city aspiring to become a Smart City, it is important to first pause and reflect on the concept of a Smart City itself. A proper understanding of the foundational principles of this concept and the strategy adopted by the city of Poznań for its implementation are crucial for comprehending the broader context that will influence how the platform is defined. Equally significant are the challenges posed to designers by the unique characteristics of the city of Poznań.

This chapter is organized as follows. In the section 2.1 the nuances of various definitions of a Smart City are considered, and a common denominator is sought. The section 2.2 is dedicated to presenting the general principles and the adopted strategy for implementing a Smart City. In the section 2.3 an overview of the data that will be processed on the platform is given. The identified challenges are summarized in the section 2.4

The issues covered in this chapter have been discussed previously by the author in the following publications [12] and [13].

2.1 Definition of Smart City

The concept of a Smart City can vary significantly depending on the organization defining it, the context in which it is introduced, or the terminology employed to describe it. The literature reflects this diversity, with at least 24 distinct definitions of the "Smart City" concept available [14], illustrating the ambiguity that surrounds the term. Over time, the fundamental principles of the Smart City concept have also evolved, progressing from version 1.0 to 4.0 [15], with each iteration emphasizing different facets of technology integration.

For the purposes of this discussion, we decided to adapt definition provided by the European Commission: "A Smart City is a place where traditional networks and services are made more efficient with the use of digital solutions for the benefit of its inhabitants and business" [16]. Although this definition remains broad, it underscores the importance of utilizing digital technologies to enhance existing processes, ultimately leading to a more livable and welcoming environment. Achieving such

improvements necessitates the initial collection of detailed source data that characterizes the phenomena under investigation from various perspectives. This data then undergoes analysis to yield new insights, which inform subsequent actions. This cyclical process may be repeated until the desired optimization objectives are achieved. It is evident that without the foundational data, no optimization efforts can be undertaken. From this perspective, the ability to collect, deliver, process, and analyze data is fundamental to the creation of a Smart City.

In a data-centric view, a Smart City is defined by its capacity to harness diverse data types that describe and characterize ongoing processes, transforming this data into actionable knowledge for future governance or new business opportunities. Supporting this process presents considerable challenges, requiring continuous innovation, the resolution of emerging obstacles, and the adoption of cutting-edge technological advancements.

Despite the coexistence of various Smart City definitions, they all share a common foundation: the use of digital data to improve the city and enhance the quality of life for its residents. As data serves as the driving force behind optimization processes and business strategies, effective data management, governance, and efficient acquisition and distribution are crucial to the success of a Smart City. Over recent years, numerous approaches to constructing Smart City solutions have been implemented, providing valuable experience and insights into the primary challenges faced.

The development of Smart Cities is inextricably linked to the data collected within urban spaces and its use in process optimization, governance, or revenue generation. All smart applications, analytical models, and decision-support systems are built on a foundation of data. Consequently, data is the central focus for all stakeholders involved. The capacity to efficiently collect and distribute data is a hallmark of a modern, digitized Smart City. The rapid pace of technological innovation has led to the creation of vast amounts of digital data, which can be leveraged for the benefit of citizens. The evolution of Smart City data platforms must keep pace with these technological changes and be equipped to support new user scenarios.

2.2 Poznań Smart City strategy

This section is based on Poznań City Hall development strategy [5] and other official publications [17].

The strategic vision for Poznań as a modern city is shaped by a commitment to several foundational principles. These principles, which guide the city's transformation into a Smart City, underscore the importance of sustainability, technological innovation, economic vitality, community inclusiveness, and digital advancement.

Central to Poznań's ambition is the principle of sustainable development, which aims to transform the city into an eco-friendly urban environment. This involves enhancing environmental conditions, creating blue-green infrastructure and optimizing the use of resources through a circular economy. The city places significant emphasis

on environmental education to ensure that these initiatives lead to long-term sustainability, seeding a culture of ecological responsibility among its residents.

Technological advancement is also a cornerstone of Poznań's modern city framework. The city's Smart City concept revolves around the integration of new technologies to optimize urban management and improve the quality of life for its citizens. This includes the implementation of advanced traffic management systems, the expansion of digital services and the seamless integration of data into decision-making processes.

Economic growth and innovation are equally critical to Poznań's strategy. The city is dedicated to fostering a business-friendly environment that supports the development of advanced technological industries, research and development initiatives and the MICE sector (meetings, incentives, conferences and exhibitions). By encouraging entrepreneurship, attracting foreign investments, and nurturing creative and innovative industries, Poznań aims to establish itself as a hub of economic activity and innovation.

At the heart of Poznań's vision lies a commitment to community and social inclusion. The city endeavors to create an inclusive, safe and socially integrated urban environment. This involves the development of attractive and livable urban spaces, the promotion of social responsibility among businesses and the enhancement of public services to meet the diverse needs of its residents. The goal is to build a city where all citizens feel valued and are able to actively participate in the life of the community.

Digital transformation is another key element of Poznań's strategy. The city is making significant strides in digitalization by offering a range of online public services, supporting digital education and developing applications that simplify daily life for its residents. Ensuring the security and accessibility of digital data is also a critical focus, as these elements are essential for maintaining trust and ensuring the success of the city's digital initiatives.

Poznań's Smart City strategy is built on the belief that a modern city must be both innovative and responsive to the evolving needs of its inhabitants, while also being environmentally conscious and economically vibrant. According to this strategy, a Smart City is defined as a new-generation intelligent city that utilizes advanced information and communication technologies (ICT) and modern management methods to enhance the quality of life. This approach is inherently participatory, drawing on the potential of social capital and actively involving residents in the development of smart urban solutions.

The strategy outlines several key aspects, each designed to improve the quality of life for residents by addressing specific challenges and opportunities within the city. Central to this effort is the effective use of data, which plays a critical role in informing decision-making processes. By providing insights into the needs and behaviors of residents, data enables city planners and officials to develop policies and initiatives that are more closely aligned with the realities of urban life, thereby enhancing the overall impact of the Smart City strategy.

In addition to its role in governance, data-driven approaches can significantly

enhance public services such as healthcare, education, and public safety. For instance, predictive analytics in healthcare can lead to better resource allocation and improved patient outcomes, while data-driven educational initiatives can help tailor learning experiences to the needs of individual students.

Citizen engagement is another critical component of Poznań's Smart City strategy. Open data initiatives empower residents by providing them with the information needed to participate in civic life and contribute to decision-making processes. This creates a more inclusive and participatory approach to city governance, ensuring that the voices of all citizens are heard and considered.

Finally, access to data can drive innovation and economic growth by attracting businesses, particularly in the technology and startup sectors. This, in turn, creates job opportunities and supports economic growth, thereby improving the standard of living for all residents.

In conclusion, the successful implementation of Poznań's Smart City strategy hinges on the effective use of data. By enabling more informed decision-making, optimizing city operations, enhancing public services, and promoting citizen engagement, data plays an important role in driving the city's transformation into a modern, dynamic and livable urban environment. Ultimately, these efforts aim to improve the quality of life for all residents, making Poznań a better place to live, work and thrive.

2.3 Poznań Smart City data assets

In 1997, the City Council of Poznań, in collaboration with the Poznań Supercomputing and Networking Center, initiated a partnership aimed at digitizing city assets. This collaboration led to the creation of the Municipal Multimedia Guide (MIM), an early communication tool designed to facilitate interaction between the city, its residents, and visitors. By 2019, MIM had become a vital resource, serving nearly six million users annually and aggregating data contributed by approximately 2,500 editors. It provided access to over twenty data collections via REST APIs, which enabled the development of various services for residents, tourists, and city officials. These services included digital city plans, platforms for public consultations, civil budget management, air quality monitoring, intervention reporting, and applications such as the smart zoo feature. The architecture of MIM was instrumental in supporting numerous Smart City applications that utilized its data, reflecting the broader concept of a Smart City as a network of innovative services impacting key areas of urban life and decision-making.

The Poznań Smart City initiative encompasses a diverse array of applications and infrastructures, which can be categorized into three principal areas: interactive applications, environmental and transportation management systems, and foundational data and infrastructure networks. Each category plays a critical role in improving urban life by utilizing and generating specific types of data, thereby affecting daily activities and decision-making processes.

Interactive Applications focus on enhancing user engagement through real-time, context-sensitive information. These applications cater to a range of needs including navigation, educational content, and location-based services. They produce and consume data related to user interactions, geographic locations, and real-time updates, which contribute to the convenience and quality of life for residents. By offering features such as interactive guides and real-time transit updates, these applications simplify access to city amenities, streamline daily tasks, and personalize user experiences.

Environmental and Transportation Management Systems are crucial for optimizing urban operations and managing environmental conditions. These systems handle data related to air quality, traffic patterns, and transportation logistics. For instance, air quality monitoring systems generate data that informs public health policies, while transportation management systems provide real-time traffic and transit information, aiding in efficient navigation and reducing congestion. These systems are essential for promoting urban sustainability, improving transportation efficiency, and ensuring public safety.

Foundational Data and Infrastructure Networks form the technological core of the Smart City. This category includes high-speed connectivity systems, wireless networks, and geospatial information systems. These infrastructures manage substantial volumes of data related to network connectivity, public safety surveillance, and spatial information. By enabling high-speed internet access and reliable data transmission, they support the operation of other Smart City applications and services. Their role is to ensure smooth data flow, integrate systems and maintain the overall operational efficiency of the city's smart initiatives.

The successful integration and management of these applications and infrastructures underscore the importance of establishing a unified data platform. Such a platform would facilitate the consolidation and processing of diverse data streams from multiple sources, ensuring coherence and enhancing decision-making capabilities. To support integrated and cross-functional data analysis, Smart City platforms must evolve to handle complex data interactions and predictive modeling. This includes developing systems that capture the behavior of individual urban components and their interdependencies, enabling automated controls and interactions. Addressing the needs of decision-makers at various levels—from individual residents to large-scale service providers—requires robust and flexible data platforms. The experience gained from deploying existing smart applications highlights the complexities involved in integrating new scenarios within the current framework. As technological advancements and the sophistication of future scenarios continue to grow, there is a clear need for next-generation data platforms to meet emerging challenges and leverage new data sets and artificial intelligence solutions effectively.

2.4 Challenges in implementing Open Data platform for Smart City

The construction of a ODSC presents a range of challenges, primarily due to the inherent complexity and heterogeneity of the data it aims to manage. A critical obstacle is that data within a ODSC originate from multiple sources, each employing different protocols, formats, and technologies [18]. This diversity is compounded when the platform has to accommodate multiple tenants, each with its management strategies and requirements. Consequently, large-scale management and harmonization of data within a ODSC remain a significant challenge, with implications for efficiency and scalability [19].

One of the challenges in building a ODSC lies in the complexities of the data itself [20]. The main issue is the lack of standardized descriptions, which causes ambiguity and hinders effective data utilization. Furthermore, inconsistencies in the naming of features exacerbate the difficulty of data handling and analysis, leading to inefficiencies in the data integration and application development processes. Another challenge is the task of identifying specific data subsets necessary for the delivery of particular services within the Smart City framework. This identification is crucial for operational efficiency and facilitation of targeted service provisions; however, it remains an obstacle due to the varied nature of the data involved.

Maintaining fresh data in real time across city operations, from traffic management systems to public utilities, is another challenging task. The sheer size of the data collected and the processing power required mean that conventional data processing methods are often inadequate, leading to potential lags in data freshness that could compromise decision-making and citizen services. Reacting to attribute updates presents another layer of complexity. Attributes of urban data (e.g., traffic flow, energy consumption rates) are highly dynamic, changing not just daily but often in real time. Adjusting to these changes promptly requires highly adaptive algorithms and substantial computational resources, complicating the deployment and scaling of the platform. Lastly, producing universal APIs that address the diverse needs of city stakeholders is a significant technical hurdle. These APIs must be robust, secure, and flexible enough to handle a wide array of data types and usage scenarios, ensuring that they can effectively serve the unique demands of each user group without compromising the overall performance of the platform.

Another challenge of implementing a ODSC is the developers' struggle to grasp the global structure of the complex data ecosystem. Currently, developers lack the tools to understand the overarching data architecture, forcing them to rely on the cumbersome and inefficient method of querying administrators. This approach is not only difficult and expensive, but also depends on the availability and expertise of specific individuals. The enforcement and maintenance of coherent dataset descriptions within the ODSC is also problematic. Given the sheer number and diversity of data sources, along with their varying levels of quality, establishing and maintaining uniformity and clarity across dataset descriptions is a difficult task.

Beyond the challenges related to the nature of data, its diversity, and the tech-

nical issues associated with its acquisition and effective processing, there are also several organizational matters that play a significant role in ensuring data accessibility. One of the first challenges we encountered during the implementation of the ODSC platform for the city of Poznań was the organizational complexity of the city administration. The Poznań city administration comprises 333 [21] organizational units, each possessing considerable autonomy in the tasks they perform. Some of these units undertake the development of extensive IT systems to support their internal business processes. These units are also responsible for managing IT systems that support urban infrastructure. Examples of such infrastructure include traffic control systems, public transportation passenger control systems, vehicle location systems for public transport, and others.

Most of these systems were implemented as a result of initiatives aimed at delivering the required functionalities by the contractor. However, the procurement conditions did not previously include requirements related to ensuring access to the data managed by these systems. Consequently, as of today, the city of Poznań has many data silos that cannot be easily or systematically accessed. In extreme cases, access to data is nearly impossible due to the closed nature of the delivered IT systems. Organizational autonomy and the lack of vested interest among individual units in ensuring access to the data they hold mean that this situation persists and is even maintained. Moreover, due to restrictions in information flow between units, employees are often unaware of the existence of data that could be useful in their daily work. As a result, there is no internal pressure to improve the data exchange process between organizational units.

Additionally, data related to the administration's activities are often treated by officials as a measure of their effectiveness and efficiency. According to this interpretation, making data available could be discrediting for them and expose them to professional consequences. The fear of data being used to assess work quality creates an additional psychological barrier, hindering any activity in this area. Individual officials are not aware of the objective value of the data they possess and manage. They are even less able to assess the increase in value that would result from combining their data with other data, the existence of which they often do not even know. Therefore, there is a lack of understanding of the concept of data sharing and the potential for such activities.

All of this leads to the conclusion that, in our view, the city administration does not possess the organizational and cultural conditions typical of a corporation. Dispersed competencies, a complex system of dependencies, formalized collaboration between units, a lack of cohesive IT infrastructure, and difficulties in implementing new recommendations that expand the scope of duties are just some of the significant challenges the city faces in considering the creation of a unified data platform. The success of the ODSC platform implementation will largely depend on how effectively these challenges are addressed.

Smart City data systems developed over the past two decades are now burdened by technological debt, making them less adaptable to rapid changes. As technological advancements rapidly alter how data is generated, transferred, and processed, both

city governments and their technological partners must explore new methodologies for developing next-generation data platforms that can meet future challenges. The Data Mesh concept, which shifts the focus from technology to the data itself, may offer a promising alternative for designing modern Smart City data platform.

Data Mesh

Various approaches for building data platforms have been proposed in the literature. Most solutions are optimized for performance, bringing existing technologies closer to the physical limits imposed by hardware. However, there is serious doubt as to whether these platforms and architectures actually enable digital transformation and the transition to a data-driven decision-making management model. In response to these doubts, a new concept for building data platforms - Data Mesh has been proposed. Its goal is to address the socio-technical and organizational issues that are typically overlooked in previous approaches.

The concept of Data Mesh is described in detail in [22]. In this chapter, we quote and summarize the key elements of Data Mesh. Their adaptation to the conditions and specific requirements of the urban environment will form the foundation upon which we will develop the idea of ODSC.

This chapter is organized as follows. Section 3.1 summarise former approaches to design data platforms. Their problems and associated weaknesses are presented in section 3.2. Section 3.3 introduce Data Mesh definition and the most important aspects of this architecture are given in section 3.4.

3.1 Data platform architectures

The development of data platform architectures has been significantly influenced by the evolving demands of data utilization, technological advancements and the increasing intricacies of data management. Over time, three distinct generations of architectures have emerged, each addressing specific needs and challenges in the realm of data processing and analysis.

The first generation, known as data warehouse architecture, originated in the 1960s, focusing primarily on the management of structured data for business intelligence (BI) purposes. The core objective of data warehouses was to consolidate data from various operational systems into a centralized repository, facilitating access for reporting and analytical visualization [23]. This architecture adhered to a systematic approach where data was extracted from diverse sources, transformed into a standardized format and then loaded into warehouse tables for querying through SQL-like languages. Designed to cater to data analysts and management teams,

data warehouses provided historical insights and supported operational planning. However, the extensive ETL (Extract, Transform, Load) processes required by data warehouses and described in details in [24], became increasingly complex and costly, often leading to significant technical debt and high maintenance burdens. Furthermore, their integration with modern engineering practices, such as continuous integration and continuous deployment (CI/CD) explained in [25], proved challenging, necessitating frequent migrations to more advanced solutions.

In response to the limitations of data warehouses, the second generation, known as data lake architecture, emerged around 2010. This generation, as described in [26], was designed to meet the specific needs of data scientists engaged in machine learning and big data analytics. Unlike data warehouses, data lakes retained raw data in its original form, allowing data scientists to work with untransformed, granular data that more accurately reflected real-world scenarios. This architecture relied on scalable object storage, enabling the ingestion of data from various sources with minimal upfront transformation. The architecture's design supported massively parallel data reads, which were crucial for training machine learning models. Data lakes also introduced zones for organizing data at varying stages of cleansing and transformation, leading to the creation of lakeshore marts and feature stores tailored for specific analytical tasks. Despite their flexibility and scalability, data lakes often became difficult to manage over time, with complex pipelines and unmanaged datasets posing challenges to maintaining data quality, lineage, and trust. The complexity inherent in data lakes sometimes led to their gradual deterioration in usability and effectiveness.

The most recent generation of data platform architecture, commonly referred to as multimodal cloud architecture, represents an evolution that builds upon the strengths of its predecessors while addressing their shortcomings. This architecture integrates both batch and stream processing capabilities, enabling real-time data availability and transformation. By utilizing frameworks such as Apache Beam, multimodal architectures unify these processes and leverage cloud-based managed services to optimize elasticity and cost [27]. Modern implementations often combine the functionalities of data warehouses and data lakes, resulting in platforms that support embedded machine learning training and real-time analytics. These platforms make use of cloud-native technologies, which separate compute and storage resources to enhance scalability and efficiency. However, despite the advancements of third-generation architectures, they still encounter challenges inherited from earlier models, including the complexity of managing large-scale data infrastructures and ensuring data integrity and accessibility across diverse use cases.

3.2 Design pattern weaknesses

A common misconception in data platform architectures is the belief that data must be centralized under the control of a single organization to be effective, with a unified enterprise-wide taxonomy. This assumption has a detrimental effect on the overall availability and performance of the platform. Centralizing data management in-

roduces a significant bottleneck, hampering the efficient integration and utilization of diverse data sources. This centralization causes delays in adapting to new data sources and responding to consumer demands, thereby diminishing the platform's agility and stifling innovation.

Another misguided assumption is that data architecture, technology and organizational structures should be monolithic, with tightly integrated components. However, as the number of data sources and use cases grows, a monolithic approach becomes increasingly complex and challenging to sustain. This complexity leads to organizational inefficiencies, longer response times, and difficulties in scaling and adapting to new requirements, ultimately compromising the platform's flexibility and overall performance.

Additionally, an overemphasis on technology-driven design is a flawed assumption, where the architecture is shaped by the available technologies rather than being aligned with business domains. This approach results in a fragmented and cumbersome system that incurs high synchronization costs and is slow to adapt to changes. Moreover, it fosters organizational structures that are misaligned with the goal of efficiently delivering business outcomes, further impairing the platform's effectiveness and agility.

The centralized and monolithic nature of many commonly adopted data platform architectures introduces significant organizational complexity and inefficiency. These structures create a challenge in synchronizing and prioritizing tasks, making it difficult for organizations to adapt their shifting priorities to the needs of the data platform. A major consequence is that data engineers, often distanced from the business domains they are supposed to support, encounter difficulties in delivering relevant and high-quality data. Their lack of domain-specific knowledge, compounded by the existence of organizational silos, impairs their ability to meet the nuanced requirements of the business effectively.

This centralized and monolithic approach also severely limits the platform's ability to respond swiftly to changes. The time required to make new data sources available or to accommodate new use cases is unacceptably long, stifling the ability to experiment and innovate rapidly—an essential capability in today's data-driven enterprises. The need for extensive coordination and synchronization across different teams and systems causes delays, creating a sluggish and inflexible response to evolving data needs.

The quality and trustworthiness of data often diminish in such architectures due to the centralization of data management away from its original context. When the central data team is distanced from the data's source domains, it becomes increasingly difficult to preserve the inherent meaning and context of the data. This distance hinders effective data cleansing and enrichment, leading to a decline in data quality that can erode trust among users.

The scaling of data platforms also presents challenges within a monolithic framework. As the number of data sources and use cases expands, the architecture becomes progressively harder to scale. The growing complexity of data pipelines, coupled with tightly interdependent components, leads to mounting maintenance costs

and inefficiencies.

Another significant issue arises from the misalignment between technical and organizational structures. In an effort to manage complexity, technical partitioning often results in teams that focus on activities rather than on achieving business outcomes. This misalignment generates friction within the organization, ultimately slowing down the delivery of new, high-quality data and compromising the platform's effectiveness.

The prevalent architectural approaches characterized by centralization, monolithic designs, and technology-centric frameworks are unsuitable to the demands of the modern data landscape. These architectures struggle with scalability, responsiveness, and the maintenance of data quality, thereby limiting the ability of data platforms to support rapid innovation and diverse analytical needs. A shift towards more decentralized, domain-oriented architectures could alleviate these challenges, enhancing the agility and efficiency of data platforms in meeting contemporary business demands.

To substantiate the argument regarding the inefficacy of current approaches to building data platforms, it is worth paying attention to reference findings from the NewVantage "Partners Releases 2021 Big Data and AI Executive Survey" conducted among Fortune 1000 companies and industry leaders 28. The results of this survey highlight a concerning gap between the investments made in data initiatives and the actual outcomes achieved by organizations. In the 2021 report, only 24.4% of companies indicated that they had successfully established a data-centric culture. Similarly, a mere 24.0% of firms reported that they had become data-driven, while just 41.2% stated that they were leveraging data and analytics to compete in the market. These figures are alarmingly low, especially given the significant financial commitments made by these companies. Notably, 99% of the surveyed organizations have invested in big data and AI, with 62% of them reporting investments exceeding \$50 million. 28 concludes: "Cultural change, however, seems a more significant barrier to long-term success. Over the years this survey has been conducted, we have commented on the challenge of changing corporate cultures in a more data-driven direction. This year's findings exhibit that challenge to an even greater degree. All questions relating to the long-term progress of corporate data initiatives exhibited declines from 2019 and 2020 levels, a disappointing development."

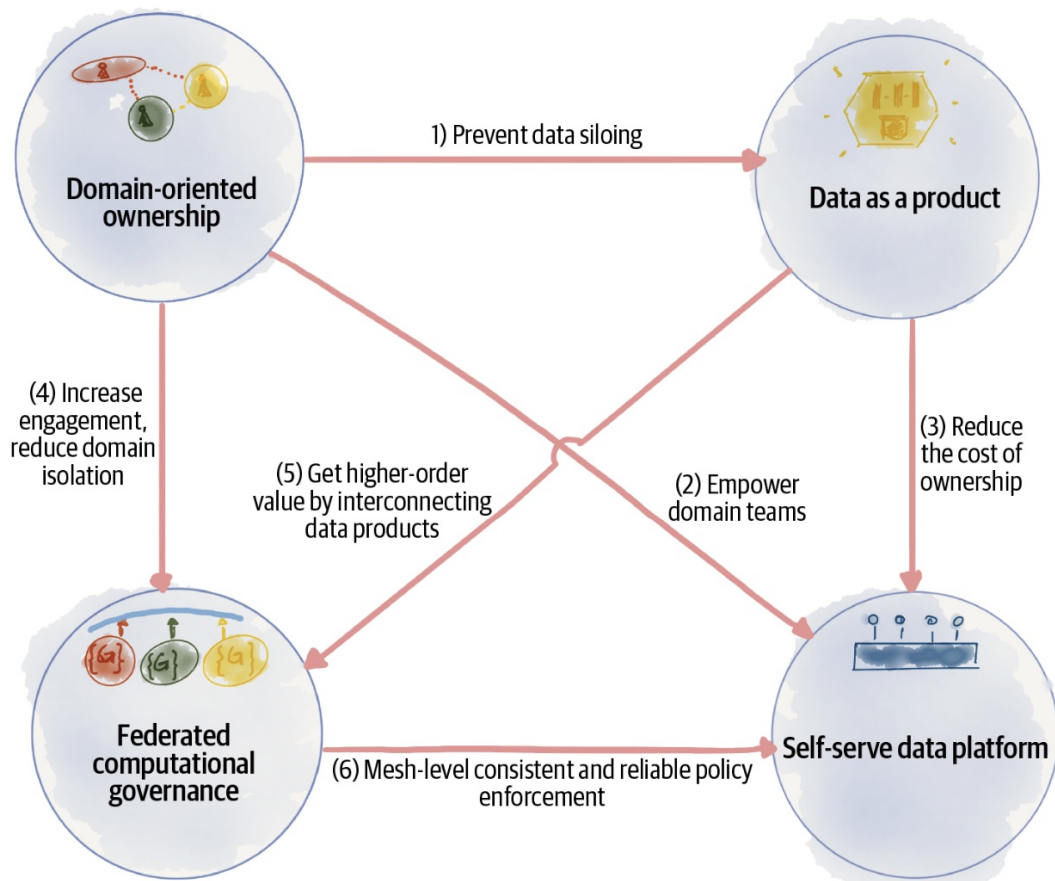
3.3 Data Mesh principles

The concept of Data Mesh was proposed as an alternative to previously used architectures. It is designed to address the issues identified as key weaknesses in the approaches that have been used so far. As defined by 22 "Data Mesh is a decentralized sociotechnical approach to share, access, and manage analytical data in complex and large scale environments within or across organizations."

The concept's goal is to improve the relationships between domain teams and eliminate any friction that slows down existing processes and limits the ability to

adapt. It also identifies the need for a fundamental shift in organizational culture, moving away from dividing teams by competencies and towards organizing competencies around data domains. Again, in the words of [22] "I have decided to classify Data Mesh as a sociotechnical paradigm: an approach that recognizes the interactions between people and the technical architecture and solutions in complex organizations. This is an approach to data management that not only optimizes for the technical excellence of analytical data sharing solutions but also improves the experience of all people involved: data providers, users, and owners. Data Mesh can be utilized as an element of an enterprise data strategy, articulating the target state of both the enterprise architecture and an organizational operating model with an iterative execution model."

Data Mesh concept is defined by four principles. Their relations are presented on Figure 3.1.



* Direction of the arrow shows the dependency from one principle to another; implementing the from principle creates the challenge that the to principle addresses.

Figure 3.1: Data Mesh principles and their relationships. Original image published in [22]

The principle of Domain Ownership within a Data Mesh architecture involves shifting the responsibility for managing analytical data to the business domains that are most connected to the data, whether as its source or primary consumers. This approach segments and administers data according to the business domain it pertains to, creating an alignment between business operations, technology, and analytical data. The intent is to enable data sharing that scales with organizational growth, localize changes within specific business domains, minimize the need for cross-team coordination, and enhance data accuracy by eliminating the complexities associated with intermediary pipelines.

Data as a Product principle emphasizes that domain-specific data should be treated as a product, directly accessible to data users such as analysts and scientists. This principle ensures that data products are discoverable, addressable, comprehensible, reliable, accessible, interoperable, valuable, and secure. Each data product, referred to as a data quantum, operates independently and includes all the necessary structural elements for data sharing. The goal is to avoid the formation of data silos within domains, encourage a culture of data-driven innovation, and ensure robustness against changes by establishing clear data-sharing agreements.

The Self-Serve Data Platform principle supports the development of a platform that allows domain teams to efficiently share data through self-service mechanisms. This platform reduces the complexities involved in data management, streamlines the lifecycle of data products, and maintains a reliable network of interconnected data assets. It simplifies the processes for data users to find, access, and utilize data products, while also facilitating data providers in the creation, deployment, and upkeep of these products. The platform's objective is to reduce the cost associated with decentralized data ownership, abstract underlying complexities, empower generalist developers, and automate governance policies to ensure security and compliance.

Federated Computational Governance principle introduces a governance model that relies on decentralized decision-making and accountability, bringing together teams composed of domain representatives, data platform experts, and specialists in legal, compliance, and security matters. This model maintains a balance between the autonomy of individual domains and the overall interoperability within the Data Mesh. It focuses on codifying and automating governance policies for every data product to streamline compliance and minimize manual synchronization efforts. The goal is to maximize the value derived from data aggregation, address issues of domain incompatibility, and implement cross-domain governance requirements effectively.

These four principles—Domain Ownership, Data as a Product, Self-Serve Data Platform, and Federated Computational Governance—are designed to work in harmony with one another. While domain ownership might risk creating data silos, the principle of treating data as a product ensures the availability of high-quality data across domains. The self-serve data platform alleviates the cognitive burden and costs associated with data product ownership for domain teams, thereby boosting productivity. Federated computational governance offers a structured approach to managing data interoperability and compliance, addressing challenges that arise

from decentralized domain ownership. Collectively, these principles establish a resilient and scalable Data Mesh architecture that facilitates efficient data sharing and utilization across an organization.

3.4 Data Mesh approach for data platform design

The principles discussed in section 3.3 significantly influence the overall architecture, offering practical guidance for their implementation. Adopting a Data Mesh architecture requires substantial changes across multiple aspects of an organization's data management strategy.

From an organizational perspective, the conventional approach of centralized data ownership by specialized teams is replaced by a decentralized model as shown on Figure 3.2. In this framework, the responsibility and accountability for data are transferred to the business domains where the data originates or is primarily used.

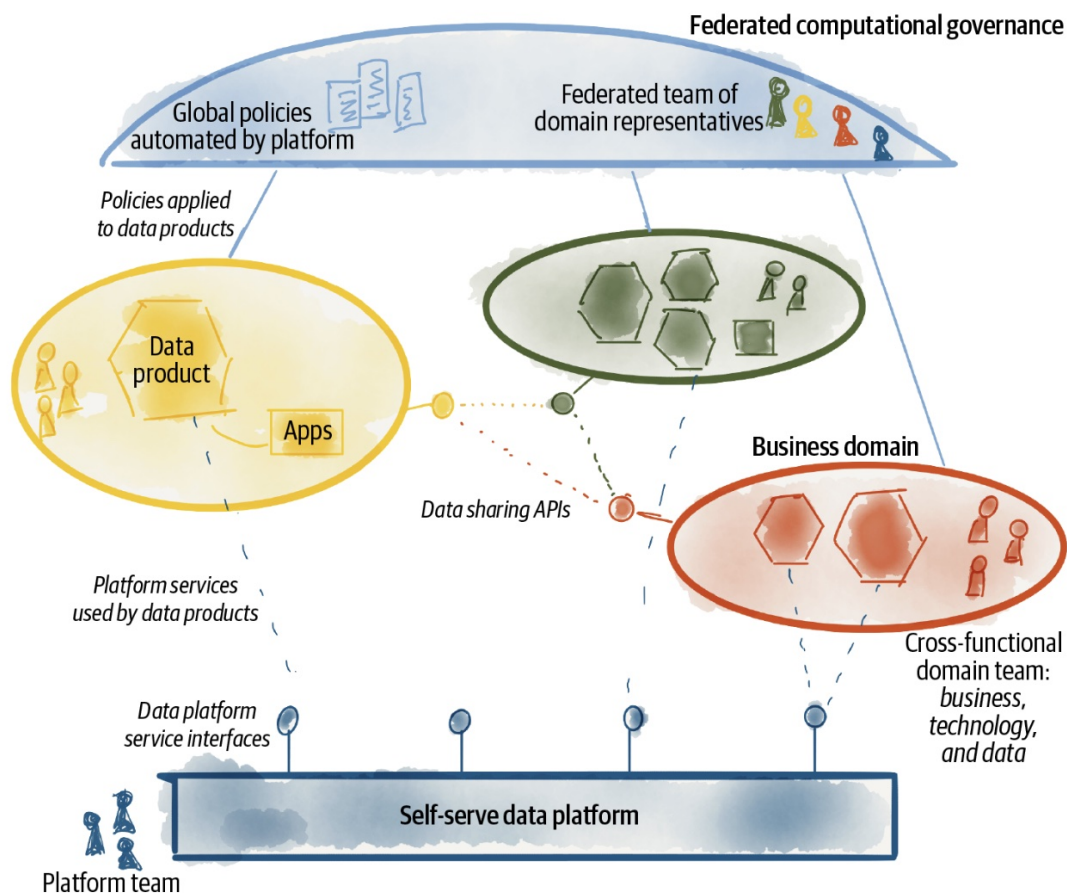


Figure 3.2: Data Mesh operational model. Original image published in [22]

Architecturally, rather than consolidating data into centralized warehouses or lakes, Data Mesh enables connectivity through a distributed network of data products. These products are interconnected and accessed using standardized protocols, thereby enhancing interoperability and scalability.

Technologically, Data Mesh redefines the relationship between data and the code that manages it, treating them as a unified, autonomous entity. This marks a departure from traditional systems where data is often a mere byproduct of pipeline execution, allowing for a more integrated and holistic approach to data management.

Operationally, governance within a Data Mesh moves away from a centralized, top-down model that depends heavily on manual oversight, toward a federated system. This system embeds computational policies within the nodes of the Mesh, increasing automation and ensuring consistency across the platform.

Principally, there is a shift in the underlying values, moving from the notion of data as a static asset to be accumulated, toward viewing data as a product designed to meet the needs of its users, both within and outside the organization.

Infrastructurally, Data Mesh brings together previously fragmented infrastructure services, uniting the domains of data and analytics with operational systems. This integrated infrastructure supports both operational and data systems in a cohesive manner, enabling more seamless and efficient data operations.

Transitioning to a Data Mesh approach requires a fundamental reconsideration of how data is managed, utilized, and governed, leading to a more agile, user-focused, and efficient data ecosystem.

The principles outlined in section 3.3 provide a foundational understanding of Data Mesh and the assumptions upon which it is based. A more detailed exploration of these principles will facilitate the development of a logical architecture, serving as the foundation for the physical implementation of a data platform aligned with the Data Mesh model.

3.4.1 Data domains and data sharing

In the context of Data Mesh architecture, domain-oriented analytical data sharing plays an important role, focusing on the decentralization of data ownership and management. This model redefines the structure of data ownership by aligning analytical data with specific business domains, thereby allowing these domains to manage and share the data they produce. Such an approach significantly enhances the accessibility and scalability of data across the organization.

Traditionally, domains serve their users through operational APIs—such as REST, GraphQL or gRPC—that handle functions like resource creation or updates, typically managing smaller data volumes to provide real-time system snapshots. The Data Mesh framework does not disrupt these operational interfaces but rather integrates with them, adding another layer of responsibility for domains: the sharing of analytical data. Through newly established analytical interfaces, APIs are made available that facilitate the discovery, understanding and access to data products. These interfaces govern the sharing process and may direct users to underlying storage systems,

including blob storage, event streams, or database tables.

In this architecture, each domain maintains complete control over its operational and analytical data, ensuring a decentralized approach to data ownership. This autonomy empowers domain teams to independently manage the release and deployment of their applications and data products, which enhances both scalability and agility. While operational APIs remain in use, the Data Mesh architecture also allows these APIs to serve as sources for the creation of data products, enabling a seamless transition between operational and analytical data flows.

The adoption of standardized protocols for analytical data interfaces ensures uniformity and ease of access across different domains, promoting interoperability and simplifying the complexities associated with data consumption throughout the organization. These analytical interfaces are instrumental in making data products easily discoverable and comprehensible, reinforcing a user-centric perspective where data is treated as a product tailored to meet specific user needs.

Domains often depend on data from other domains. The Data Mesh framework ensures that such interdependencies are explicitly defined and managed by the data products themselves. Each data product governs its data consumption from upstream sources, thereby ensuring that dependencies are clearly established and data consumption is effectively controlled.

By structuring data around business domains and providing comprehensive interfaces for both operational and analytical data, the Data Mesh architecture advocates for a decentralized and scalable approach to data management. This model not only improves the accessibility and usability of data but also grants domains greater autonomy and control, resulting in a more efficient and dynamic data ecosystem.

3.4.2 Data product

Within the framework of Data Mesh architecture, a data product serves as a fundamental architectural unit — an autonomous entity that can be independently deployed and managed. This concept ensures that each data product possesses high functional cohesion, containing all the necessary components to operate self-sufficiently. By integrating transformation code, data, metadata, policies, and infrastructure dependencies, data products become fully encapsulated, enhancing their usability, discoverability, and secure sharing capabilities.

The core components of a data product include transformation code that handles domain-specific data processing, producing analytical outputs ready for sharing. Unlike traditional systems, where transformation logic may reside outside the data unit, Data Mesh embeds this code directly within the data product, ensuring its autonomous functionality. In order to share its content, data product provides structured APIs that allow for seamless access, discovery, and interaction with the data. These APIs adhere to predefined contracts, facilitating both synchronous and asynchronous communication while ensuring consistent data handling.

Policy as Code within a data product enforces behavioral and structural rules, such as encryption, access control, and compliance measures, safeguarding data in-

egrity and security. The management of data and its associated metadata is another important element, encompassing everything from documentation and semantic definitions to Service Level Objective (SLO) metrics. This organization enhances the transparency and usability of data across the organization. Additionally, platform dependencies define the infrastructure needs and provisioning requirements necessary for each data product to operate in isolation, further reinforcing its autonomous nature.

Data products are designed to interact with other systems and products through well-defined data sharing mechanisms. Input Data Ports allow a data product to ingest data from various sources, including upstream products or operational systems, whether through near-real-time event streams or periodic data aggregation. Conversely, Output Data Ports provide externally accessible APIs for sharing the product's output, enforcing stringent access policies and guiding users to the appropriate underlying storage solutions for secure data retrieval.

To ensure that data products are easily discoverable and their operations are transparent, Discovery APIs offer comprehensive metadata, including unique identifiers, ownership details, semantic descriptions, and searchable tags. These APIs facilitate efficient data discovery and understanding. In parallel, Observability APIs provide monitoring and auditing tools, such as logs, data lineage, and performance metrics, supporting ongoing operations and enabling effective troubleshooting.

Each component of a data product—ranging from its structural elements to its data sharing interfaces and discovery mechanisms—contributes to its status as a fully independent and cohesive unit. The integration of transformation code and policy management directly within the data product ensures that it can govern its lifecycle and processes without external dependencies. The APIs and interfaces promote fluid interoperability and data sharing, while the discovery and observability features enhance transparency and user engagement.

Data products engage in dynamic relationships through their input and output data ports, enabling data flow and consumption across different domains. These interactions can be tailored to specific needs, whether through synchronous methods like API queries or asynchronous techniques such as event subscriptions. By clearly defining dependencies and interaction points, data products maintain manageable and well-structured relationships, supporting a scalable and interconnected data ecosystem.

The data product model within Data Mesh architecture guarantees that each unit operates as a self-contained, highly cohesive entity, capable of independent functioning and seamless integration with other data products. This model significantly enhances data accessibility, usability, and governance, laying the foundation for a robust and scalable data platform.

3.4.3 Self-serve data platform

In the context of Data Mesh architecture, the self-serving platform plays a critical role by breaking down complex functionalities into modular, API-driven services

with open interfaces. This design departs from traditional, monolithic platforms by emphasizing scalability, adaptability, and seamless integration across diverse environments, including multiplatform and multihosting settings. By enabling cross-functional domain teams to autonomously manage their data products, the platform also ensures secure and compliant data discovery, learning, and consumption. The architecture is divided into distinct planes, each dedicated to specific roles within the Data Mesh framework.

At the core of the architecture is the Data Infrastructure (Utility) Plane, which manages the essential infrastructure components necessary for constructing and operating the Data Mesh. This layer is responsible for provisioning and maintaining resources such as storage, compute power, and identity management systems. It is closely integrated with the broader digital application infrastructure that supports operational systems, including shared continuous integration and deployment (CI/CD) pipelines that serve both data products and applications. By overseeing common storage solutions like object stores used in both analytical and operational contexts, this plane ensures a robust and efficient foundation for the entire Data Mesh.

Building on this foundation, the Data Product Experience Plane offers higher-level tools and abstractions that facilitate the development, maintenance, and consumption of data products utilizing the resources from the infrastructure plane. It provides support to data product developers throughout the entire lifecycle of a data product, from initial setup through building, testing, deployment, and monitoring. This plane also caters to data consumers by simplifying access to and interaction with data products, allowing for tasks such as subscribing to data outputs and retrieving data. By streamlining the creation and management of data products, this plane enhances developer productivity and ensures that data products are readily accessible and usable.

The Mesh Experience Plane serves as the overarching layer that integrates and manages functionalities across multiple data products, enhancing the overall coherence and effectiveness of the data ecosystem. It enables operations that span the entire Data Mesh, such as searching for specific data products and tracking their lineage to understand the interconnections between different data inputs and outputs across the Mesh. By consolidating features from individual data products, this plane provides a unified view and management capability for the Data Mesh, ensuring that all interconnected data products work together smoothly and efficiently. This facilitates user navigation, comprehension, and utilization of the Data Mesh as a whole.

In conclusion, the self-serving platform in a Data Mesh architecture is structured across three key planes. The Data Infrastructure Plane manages the fundamental infrastructure resources, the Data Product Experience Plane simplifies the development and use of data products, and the Mesh Experience Plane ensures cohesive functionality and management across the entire Data Mesh. Together, these planes form a flexible and scalable architecture designed to meet the complex demands of modern data ecosystems.

3.4.4 Federated computational governance

In the Data Mesh architecture, governance is reinforced by embedding computational policies directly within data products, leveraging a set of logical components to ensure consistent application across the entire data ecosystem. This design approach relies on the concept of a data quantum, which enables the uniform definition and execution of policies as code throughout the Mesh. As a result, each data product is equipped to autonomously evaluate and enforce these policies as necessary. The core components facilitating this governance framework include the data product sidecar, the computational container, and the control port.

The data product sidecar is integral to the execution of policies and the standardization of various data product aspects. Typically implemented as a separate process adjacent to the data product, the sidecar addresses cross-cutting concerns that require uniformity across the Mesh. Its primary function is to enforce policies within the local context of each data product, thus avoiding centralized points of failure or operational bottlenecks. This localized enforcement includes critical functions such as access control and encryption. Additionally, the sidecar enhances interoperability by offering standardized APIs that govern interactions between data products, including the management of service level objective (SLO) metrics.

The computational container of a data product consolidates all structural elements necessary for independent operation, including the sidecar, policy configurations, transformation code, data, and interfaces for data sharing. This encapsulation allows each data product to function autonomously while maintaining adherence to governance standards. Consistency across data products is further ensured by the data product experience plane, which upholds the standardized governance framework within the Mesh.

The control port serves as a dedicated interface that provides APIs for configuring policies and performing privileged operations on data products. It supports both localized and centralized policy configurations, enabling data product owners to establish specific policies, such as data anonymization, while allowing central authorities to enforce broader access controls. Designed for extensibility, the control port can accommodate new policy types as they emerge. It also facilitates the execution of critical operations, such as ensuring compliance with GDPR's right to be forgotten, across all data products in the Mesh. This mechanism ensures that governance protocols are uniformly applied and effectively managed across the entire architecture.

Smart City data landscape using Data Mesh

The range of challenges and risks associated with designing a data platform for a modern city is extraordinarily broad. At the same time, it is believed that the real difficulty does not lie in the technical aspects but rather in the necessity of cultural and organizational changes at the level of all stakeholders. The concept of Data Mesh, in its fundamental principles, appears to address similar issues. It defines a minimal set of changes that must occur to fully realize the data-driven management paradigm. However, adapting a concept that originates from a corporate environment to the organizational realities of a city administration requires a series of assumptions and a well-thought-out strategy. In this chapter, the journey undertaken in implementing the ODSC concept in the city of Poznań is presented, along with the assumptions made and the experiences gained.

The chapter is organized in the following way. Section 4.1 presents state of the art in solutions used for building urban data platforms. 4.2 section lays the foundation for modern Smart City data platform. The assumptions regarding the implementation of Data Mesh in the context of Smart City are presented in the section 4.3. The experience gained during the implementation in the city of Poznań are summarized from sociological perspective in 4.4 and from technical perspective in 4.5.

Findings included in this chapter were previously published by the author in [12].

4.1 Urban data platforms in literature

The history of the Smart City concept dates back to 1994 with one of the first attempts to use information technologies by the city of Amsterdam 29. The concept itself has been developed and redefined, as described by [30], from Smart City 1.0, driven by new technologies, through Smart City 2.0 and 3.0, which considered the greater influence of local authorities and communities on the directions of development and methods of technology application, to Smart City 4.0, focusing on the use of technology and data to achieve sustainable development, as pointed out in [31]. From the perspective of Smart City development, emphasizing the necessity of

data exchange and utilization to improve residents' quality of life is a relatively new concept. The emergence of new requirements for managing urban data has thus necessitated the development of urban data platforms, which are intended to become tools for achieving the goals set by Smart City 4.0. The existing literature available in 2024 on creating urban data platforms predominantly focuses on the conceptual aspects of what an urban data platform should be and what requirements it should meet 32, 33.

Among the publications, there are studies and reports presenting how data platforms are implemented by various cities around the world: 34, 35. These reports focus on showcasing the diversity of approaches, from applications to "data marketplaces," the roles of different stakeholders, and the potential for cross-sector integration. They typically conclude with findings from the conducted analyses and recommendations for the development of urban platforms.

Very few publications present the actual implementation details of an urban data platform. These publications focus exclusively on technical aspects. They demonstrate how technologies known from the Big Data (explained in details in [36]) and Cloud Computing (described in [37]) domains are used as the foundation upon which the functionalities required by the city are delivered. The article 38 details the use of Hadoop [39] and Apache Spark [40] technologies for processing data from IoT [41]. The authors also discuss the data model and its storage using NoSQL databases. The potential for using Big Data technology is also mentioned in the strategy for the city of Warsaw 42.

A separate group of publications consists of materials published within the framework of EU policy on promoting sustainable development. In this context, it is important to mention the "European data space for smart communities" initiative 43. Publications related to this initiative extensively justify and promote the implementation of information technologies to build a European data space. Soft educational and promotional activities are complemented by efforts towards standardizing and universalizing the implemented technologies. For this reason, the "European data space" also provides detailed specifications describing possible architectures and suggested technologies 44. Among the available and promoted tools are solutions developed as part of previous competitions and programs implemented within the EU framework. Among the promoted technologies, tools managed by the FIWARE foundation occupy a special place.

An important element of the literature review also includes municipal publications, which usually describe strategies for implementing the Smart City 4.0 concept, including urban data platforms 45, 46. These strategies naturally focus on the goals that are to be achieved through such implementations. They also address issues important from the city's perspective, identifying areas where improvements are key to achieving success. Therefore, the strategies define the path that a given city intends to follow from the date of the document's publication to achieve the overarching goal of sustainable development.

The existing literature on the construction and operation of urban data platforms focuses on the general strategies related to the implementation of both the Smart

City 4.0 concept and the data platforms themselves. Numerous social and sectoral issues are discussed, and conclusions and recommendations are presented based on the experiences of the most advanced cities in implementing these concepts. There are few publications addressing technological aspects. Technical documentation and specifications promoted within EU activities focus on tools and standardization. In our opinion, the discussion on effective ways to implement urban data platforms needs a new perspective from data engineering or more broadly, from information systems engineering point of view. Experiences discussed in section 3.2 indicate that simple adaptation of technological solutions developed by corporations for their own purposes is risky in itself and without adequately reorganized processes may not yield the expected results. In commercial applications, this leads to a lack of return on investment. In the case of municipal investments, where public funds are involved, the risks associated with ineffective implementations must be minimized to avoid accusations of mismanagement. The application of the Data Mesh concept in the realization of urban data platforms, in our view, should help minimize risks by focusing more on data processing processes than on the technologies themselves.

4.2 Smart City Data Platform pillars

When developing a new data platform, it's crucial to consider several foundational aspects to ensure its effectiveness and long-term viability. These aspects include technology, usability, sustainability, and governance, each addressing different dimensions of the platform's design and operation. The technology pillar focuses on managing data volume, velocity, variety, and silos, and integrating emerging processing paradigms. These elements are essential for building a scalable and adaptable architecture capable of handling diverse data types and high-speed processing requirements. The usability aspect emphasizes the platform's ability to adapt to various data scenarios, support future user needs, and manage a dynamic range of stakeholders. This ensures that the platform remains relevant and effective as user requirements and technological advancements evolve. The sustainability dimension looks at the platform's capacity to handle new data types, deliver innovative products, accommodate diverse scenarios, and integrate with new stakeholders. This focus helps maintain the platform's utility and adaptability over time. Lastly, the governance aspect involves deciding between centralized and decentralized governance, ensuring active stakeholder involvement, and setting flexible rules for data management. Effective governance is crucial for maintaining data quality and aligning with the needs of all users. Together, these aspects form the basis of a robust data platform architecture, ensuring that it can meet current demands and adapt to future challenges.

4.2.1 Technology

The technology pillar of a modern data platform must address several key aspects to ensure its effectiveness and scalability: data volume, data velocity, data variety, data

silos, and emerging processing paradigms. Each of these aspects plays a critical role in shaping the architecture and operational capabilities of the platform.

The rapid expansion of digital technologies and the proliferation of IT systems lead to massive increases in data volume. Effective management of this data is essential for supporting both real-time operations and long-term strategic analysis. To accommodate large datasets, a modern data platform must incorporate scalable storage solutions and robust processing capabilities. Failure to address the challenges associated with high data volume can result in system overloads, diminished performance, and significant data management issues.

Another challenge is the high velocity at which data is generated and needs to be processed, especially from real-time IT systems. The ability to handle this data flow efficiently is crucial for providing timely insights and maintaining operational effectiveness. An effective data platform architecture must be capable of processing and analyzing data streams with minimal latency. By addressing data velocity, the platform ensures that it can deliver up-to-date information and support dynamic decision-making processes.

The diverse nature of data—encompassing numerical values, text, images, video streams, and both structured and unstructured formats—presents a challenge for data management. The platform must be versatile enough to handle a wide range of data types and formats, incorporating tools for data integration, transformation, and storage. This flexibility is vital for supporting various use cases and applications, enhancing the platform's ability to adapt to different scenarios and requirements.

The issue of data silos, where data is confined within isolated systems, impedes accessibility and integration. This problem is particularly acute with legacy systems and proprietary solutions that lack interoperability. To mitigate this risk, the new architecture must emphasize data sharing and integration capabilities from the outset. Ensuring that data can be accessed and shared across different systems fosters a more interconnected and efficient data ecosystem.

The advent of IoT devices and new processing models such as edge and fog computing introduces additional complexities. These paradigms enable data processing closer to the source, reducing latency and bandwidth demands while enhancing real-time processing. A modern data platform must be designed to support these evolving paradigms, integrating them into the architecture to harness their benefits. Addressing these new processing models is crucial for building a platform that remains adaptable and future-proof in the face of ongoing technological advancements.

The technology pillar of a new data platform architecture for Smart Cities must address the challenges of data volume, velocity, variety, data silos, and emerging processing paradigms. By focusing on these aspects, the architecture can ensure scalability, flexibility, and the ability to integrate and process diverse data types in real time. These considerations are crucial for building a platform that not only meets current needs but is also adaptable to future technological advancements and the growing demands of Smart City management.

4.2.2 Usability

In designing a modern data platform, addressing key usability aspects is essential for ensuring that the platform remains effective and relevant over time. As the digital landscape evolves, a data platform must be equipped to handle diverse and unforeseen data scenarios, adapt to future user needs, and manage a dynamic array of stakeholders. The following aspects—adaptability to diverse data scenarios, flexibility in supporting future requirements, and dynamic stakeholder management—are critical in shaping a data platform architecture that is both resilient and forward-compatible. Each aspect plays an important role in enhancing the platform’s ability to manage a wide range of data types, accommodate evolving demands, and integrate with a changing user base, forming the foundation for a robust and enduring data management solution.

Adaptability involves the platform’s capability to handle various data scenarios without being restricted by predefined schemas or structures. Unlike traditional data warehouses, which require strict schema definitions before data insertion, or data lakes, which impose transformation and quality control at the consumption phase, a modern data platform must be able to accommodate and manage data in a way that supports a wide range of scenarios. This adaptability is crucial for the platform to remain relevant and useful as new data types and user scenarios emerge. It allows the platform to store and process data without being constrained by rigid schemas, facilitating better handling of data variety and unexpected use cases. This characteristic ensures that the platform can effectively support evolving requirements and scenarios that were not anticipated during its initial design. To achieve adaptability, the architecture should support flexible data storage and processing mechanisms that do not rely on fixed schemas. This could involve using semi-structured or schema-on-read approaches that allow data to be dynamically interpreted and utilized based on context, rather than requiring upfront transformation.

Flexibility refers to the platform’s ability to evolve and accommodate future user needs and requirements. Given the unpredictable nature of user scenarios and the constant evolution of data utilization practices, the platform must be designed to be easily adaptable to new requirements and emerging technologies. Supporting future requirements is essential for the longevity and usefulness of the data platform. It ensures that the platform can integrate new data sources, support new types of analysis, and adapt to changes in technology and user expectations. This flexibility helps avoid the pitfalls of obsolescence and ensures the platform remains a valuable asset over time. The architecture should include modular components and flexible interfaces that allow for the integration of new data sources and functionalities. This can be achieved through the use of open standards, APIs, and a modular design that facilitates easy upgrades and extensions.

Dynamic stakeholder management involves designing the platform to accommodate a changing landscape of users and stakeholders. This includes not only current stakeholders such as data owners, analytical teams, and infrastructure maintenance teams but also future stakeholders who may emerge as technology evolves. Effective

stakeholder management ensures that the platform remains relevant and useful to its users, regardless of changes in technology or user roles. By focusing on standards and rules for data access and utilization rather than supporting specific stakeholders, the platform can adapt to new users and roles as they arise. The architecture should emphasize flexible access controls, role-based permissions, and standards for data governance that can accommodate new stakeholders. This approach allows the platform to integrate new users and roles without requiring significant redesigns or manual adjustments.

The main usability aspects that the new data platform architecture must address include adaptability to diverse data scenarios, flexibility in supporting future user requirements, and dynamic stakeholder management. By focusing on these aspects, the architecture will ensure that the platform can effectively manage a wide range of data types and scenarios, evolve to meet future needs, and accommodate a changing set of stakeholders, thereby providing a robust and future-proof solution for data management and utilization.

4.2.3 Sustainability

The data platform must be designed to handle a diverse range of data types without imposing restrictions or limitations. This aspect ensures that the platform can integrate various forms of data, including structured, semi-structured, and unstructured formats, as well as emerging data types. The capacity to accommodate new data types is important because it allows the platform to stay relevant as data sources evolve and new technologies emerge. By avoiding rigid constraints on data types, the platform can remain adaptable and continue to provide valuable insights from a wide array of data sources, thereby extending its utility and lifespan.

The platform should support the development and deployment of new products derived from data processing and analytical processes. This capability involves enabling the creation of new data products, services, or features based on the data handled by the platform. This flexibility is important as it allows for continuous innovation and the adaptation of the platform to meet changing needs and opportunities. The ability to materialize new results and insights ensures that the platform can evolve in response to user demands and technological advancements, maintaining its relevance and enhancing its value over time.

A sustainable data platform must be versatile enough to accommodate a variety of user scenarios without imposing limitations. This includes supporting different types of data utilization and analytical processes that may arise. The platform should facilitate the implementation of new use cases and adapt to evolving requirements, allowing users to explore and exploit data in novel ways. By providing this level of flexibility, the platform can support a broad spectrum of applications and scenarios, ensuring that it remains useful and adaptable in the face of changing user needs and emerging trends.

The platform must be capable of accommodating new stakeholders, including future users or entities that may interact with the platform. This involves providing

technical solutions that can integrate with and serve new stakeholders, as well as reflecting their interests in the governance rules. Ensuring that the platform can support a dynamic range of stakeholders is essential for its sustainability, as it allows the platform to adapt to changes in the user base and incorporate diverse perspectives and needs. This aspect ensures that the platform remains inclusive and relevant, fostering a robust and collaborative data ecosystem.

Addressing the ability to consume new types of data, deliver new products, serve new scenarios, and support new stakeholders forms the foundation for a sustainable data platform architecture. These aspects are crucial for maintaining the platform's flexibility, adaptability, and relevance over time, ensuring that it can effectively manage evolving data requirements and continue to provide value in a rapidly changing technological landscape.

4.2.4 Governance

When designing a data platform, addressing data governance aspects is critical to ensure effective management and utilization of data. The main governance aspects to consider include centralized vs. decentralized governance, stakeholder involvement, and rule enforcement. Each of these elements plays a key role in shaping how data is managed, controlled, and utilized across the platform.

Data governance can be approached in two main ways: centralized and decentralized. A centralized approach consolidates decision-making and rule-setting within a single governing body. This model is often effective for smaller, more straightforward platforms with a single data owner and a unified dataset, such as traditional data warehouses. However, this method may fall short in larger, more complex platforms with multiple data owners and diverse user scenarios, such as data lakes. In such cases, a decentralized approach might be more suitable, distributing governance responsibilities among various stakeholders. This model allows for more nuanced rule-setting that can accommodate the varied requirements and expectations of different users. The choice between centralized and decentralized governance impacts how effectively the platform can manage data and meet diverse needs.

Effective data governance requires active involvement from all relevant stakeholders. This includes data owners, users, and other parties who interact with the platform. Governance rules should reflect the needs and interests of these stakeholders to ensure their acceptance and adherence. This collaborative approach helps create a governance framework that is practical and responsive to real-world scenarios, enhancing data quality and utility. Involving stakeholders in the governance process ensures that the rules are not only applicable but also respected, leading to better data management and more effective use of the platform.

Data governance involves defining rules and making decisions about how data should be managed and used. However, rather than enforcing rigid rules, the focus should be on creating rules that align with the needs and expectations of all stakeholders. This approach ensures that governance rules are not only appropriate but also supported by the platform's users. Effective rule-setting involves balancing con-

trol with flexibility, allowing the platform to adapt to changing requirements while maintaining oversight. This balance is crucial for fostering a data governance framework that is both robust and adaptable, promoting high data quality and effective data management.

The main data governance aspects that should be addressed in a new data platform architecture include the choice between centralized and decentralized governance, active stakeholder involvement, and thoughtful rule enforcement. Addressing these aspects ensures that the platform can effectively manage data across diverse scenarios, reflect the needs of its users, and maintain high standards of data quality and governance. This approach lays the foundation for a data platform that is well-managed, flexible, and capable of meeting the evolving demands of its users.

4.3 Basic tenets of Data Mesh in the context of Smart City

Proper design of a data platform to support Smart City development needs to address number of different aspects in order to deliver a functional platform, prove its usability and make it successful in a long term. Fostering this long-term perspective seems to be crucial as urban data platform can be considered as a public investment, and approaches used so far for implementing similar cases in the corporate environments frequently fail for reasons mentioned previously. Data Mesh concept for building next generation data platforms proposes an alternative approach. According to the Data Mesh paradigm, technology is no longer in the centre of concern. By focusing on the data, all other aspects become a means to deliver a valuable product, as the data itself should be considered. This paradigm shift, as well as new concepts introduced by Data Mesh (like data domain, distributed data governance), are trying to address the most important weaknesses of the first and second generation data platforms (warehouses and data lakes respectively), the weaknesses that lead them to final failure, understood as an inability to deliver expected values.

Paradigm shift is not only about changing the focus from technology to data. It introduces a spectrum of changes on different levels starting from a sociological perspective involving human interactions, through technical aspects of how specific tools are used, how infrastructure is delivering computing power and storage, on the data governance ending.

4.3.1 Sociological level

Implementing Data Mesh in a Smart City environment requires significant sociological shifts to ensure the successful adoption of this modern data management approach. These shifts help align the behaviors, attitudes, and practices of various stakeholders to the decentralized, product-oriented nature of Data Mesh, enabling the creation of a robust and sustainable data platform.

One of the primary shifts necessary is the redefinition of data from a byproduct of operations to a valuable product in itself. This change in perception is essential for transforming how data is treated within the city's ecosystem. Viewing data as

a product implies that it must be managed with the same rigor and strategic intent as any other marketable product. This shift necessitates the development of well-defined processes for data production, emphasizing quality, usability, and value. It moves away from the traditional, ad-hoc approaches to data handling and requires data owners and teams to focus on delivering high-quality, well-documented, and user-friendly data sets. This transformation is crucial in fostering a culture of accountability and excellence, ensuring that data is not only accessible but also valuable and reliable.

Another significant shift involves fostering a culture of ownership and accountability among data producers. In a Data Mesh framework, data ownership is decentralized, meaning that various departments or entities within the city must take responsibility for their respective data products. This decentralization is essential for allowing flexibility and innovation in data management. However, it also requires a strong commitment to governance and quality assurance. Data owners must understand their role in maintaining the integrity, security, and interoperability of their data. This shift is vital in ensuring that the decentralized model functions effectively, with each unit contributing high-quality data that can be seamlessly integrated and utilized across the city's systems.

The adoption of Data Mesh also necessitates a shift towards decentralized governance while maintaining centralized standards. In a complex environment like a Smart City, where data is generated and managed by numerous departments and organizations, it is unrealistic to impose a one-size-fits-all governance model. Instead, a balance must be struck between allowing autonomy and ensuring consistency. Centralized standards for data security, interoperability, and discoverability must be established, but the application of these standards should be flexible to accommodate the diverse needs and practices of different entities. This shift is crucial in preventing data silos and ensuring that all data products adhere to a common framework that facilitates integration and collaboration.

Equally important is the shift in mindset regarding data sharing. Traditionally, data sharing might be viewed with suspicion, particularly concerning concerns over security or the potential exposure of sensitive information. However, in the context of Data Mesh and Smart Cities, data sharing should be re-envisioned as a means of creating value rather than a risk. This shift encourages stakeholders to see the broader benefits of data sharing, such as improved public services, enhanced decision-making, and increased innovation. By adopting this perspective, data owners can be more open to collaboration and more willing to contribute to a shared data ecosystem that benefits the entire city.

Finally, there is a need for a collaborative and cross-functional approach to data management. The organizational complexity of a Smart City, with its numerous departments and varying responsibilities, requires that all stakeholders work together to achieve common goals. This collaborative spirit is essential for overcoming the challenges associated with decentralized data ownership and governance. By fostering a culture of collaboration, the city can ensure that data products are not only produced and managed effectively but also utilized to their fullest potential across

different applications and services.

The main sociological shifts necessary for implementing Data Mesh in a Smart City include redefining data as a product, fostering ownership and accountability, balancing decentralized governance with centralized standards, promoting data sharing as a value-creating activity, and encouraging cross-functional collaboration. These shifts are critical for building a data platform that is resilient, adaptable, and capable of supporting the complex and dynamic needs of a modern Smart City.

4.3.2 Technology level

To successfully implement Data Mesh within a Smart City environment, it is crucial to leverage advanced technologies that can support the decentralized and product-oriented approach inherent in Data Mesh. These technologies play a pivotal role in enabling the creation, delivery, and management of data products, ensuring that they meet the diverse and evolving needs of a Smart City.

Cloud platforms and resource virtualization technologies form the backbone of the technological infrastructure needed for Data Mesh. These platforms provide a scalable and flexible environment that can easily accommodate the dynamic demands of a Smart City. By utilizing cloud services, Smart City data platforms can benefit from on-demand access to computational power and storage resources, which can be scaled up or down based on real-time requirements. This flexibility is essential for handling the vast amounts of data generated in a Smart City, enabling efficient data processing and storage without the need for extensive physical infrastructure. Moreover, cloud platforms facilitate the deployment of a wide range of software services in a Software as a Service (SaaS) model, which significantly reduces the time and cost associated with setting up and maintaining the necessary software stacks.

The adoption of cloud technologies also allows for the efficient management of software resources through DevOps practices as described by [47]. By delegating software management tasks to specialized DevOps teams, data management teams can focus exclusively on their analytical tasks. This division of responsibilities is crucial in preventing operational bottlenecks and ensuring smooth collaboration between different teams involved in the Data Mesh. The ability to provide custom services tailored to specific user needs further enhances the agility and efficiency of the data platform, enabling it to quickly adapt to new challenges and opportunities.

In addition to cloud platforms, Big Data technologies are integral to the implementation of Data Mesh in a Smart City. Technologies such as Hadoop Distributed File System (HDFS) and Apache Spark provide the necessary infrastructure for distributed storage and processing, which are essential for managing large-scale data across multiple nodes in a decentralized environment. HDFS offers a robust solution for storing vast amounts of data across multiple locations, ensuring data availability and fault tolerance, while Apache Spark enables fast and efficient data processing, supporting the complex analytics required in a Smart City.

Furthermore, task management and user control systems are critical for maintaining the operational efficiency of the Data Mesh. These systems ensure that data

processing tasks are executed in a coordinated manner, and that user access to data is managed securely and effectively. By implementing comprehensive task management and user control mechanisms, the Smart City data platform can maintain high levels of security and reliability, which are essential for safeguarding sensitive data and ensuring compliance with regulatory requirements.

Object storage technologies also play a role in the Data Mesh architecture. These technologies provide a scalable and cost-effective solution for storing unstructured data, which is increasingly prevalent in Smart City environments. With their ability to handle large volumes of diverse data types, object storage systems offer virtually limitless scalability, making them ideal for the ever-expanding data needs of a Smart City.

The successful implementation of Data Mesh in a Smart City environment hinges on the integration of cloud platforms, Big Data technologies, and advanced resource management systems. These technologies provide the necessary foundation for a scalable, flexible, and secure data platform that can support the decentralized, product-oriented approach of Data Mesh. By leveraging these technologies, Smart Cities can effectively manage the complexities of their data ecosystems, enabling them to deliver high-quality data products that drive innovation, improve public services, and enhance overall city management.

4.3.3 Infrastructure level

When implementing Data Mesh within a Smart City environment, it is essential to address specific infrastructure aspects that will support the decentralized and domain-oriented nature of this approach. The infrastructure must be designed not merely as a passive environment hosting services but as a dynamic, responsive system that enables data management teams to operate independently and efficiently.

A key requirement for the infrastructure is that it must support a self-service model from the outset. This means that data management teams must be empowered with the necessary tools and resources to manage their workflows without relying heavily on centralized infrastructure teams. By providing access to Continuous Integration/Continuous Deployment (CI/CD) pipelines, reusable processing templates, and general-purpose services such as a common storage layer, data teams can execute tasks autonomously. This autonomy is crucial in avoiding potential bottlenecks that could arise if infrastructure teams are overburdened with requests from multiple development projects. Moreover, as the complexity of data scenarios increases, it becomes increasingly important for decisions to be made by those with deep domain knowledge, which resides primarily within the data teams themselves.

This self-service infrastructure model also reduces friction in the development and deployment processes, allowing for more agile and responsive project execution. By equipping data teams with the ability to manage technical tasks independently, the Smart City data platform can achieve faster iteration cycles, enabling it to quickly adapt to new requirements and challenges. This flexibility is particularly important in a Smart City environment, where the data landscape is continually evolving, and

the ability to respond swiftly to changes is critical for maintaining the efficiency and effectiveness of city management processes.

Furthermore, the infrastructure must be robust and scalable to accommodate the diverse and growing data needs of a Smart City. This involves integrating technologies that support distributed storage, processing, and task management, as well as implementing security and access control mechanisms that ensure the integrity and confidentiality of the data. The infrastructure must be designed to handle the high volume and variety of data generated across different city departments and services, ensuring that data is stored, processed, and accessed in a manner that is both efficient and secure.

Effective infrastructure management is also critical for maintaining the overall health and performance of the Smart City data platform. By adopting modern DevOps practices, infrastructure teams can automate many of the routine maintenance tasks, freeing up resources to focus on more strategic initiatives. Automation not only improves efficiency but also reduces the risk of human error, ensuring that the infrastructure remains stable and reliable even as the demands on it grow.

The infrastructure for implementing Data Mesh in a Smart City environment must be designed to support a self-service model, enabling data management teams to operate independently and efficiently. This approach minimizes bottlenecks, reduces friction in the development process, and allows for more agile and responsive project execution. By integrating scalable and secure technologies, and adopting modern infrastructure management practices, the Smart City data platform can effectively manage the complex and dynamic data environment, driving innovation and improving the overall efficiency of city operations.

4.4 Lessons learned from implementing Data Mesh in a Smart City context

Data Mesh is not a ready-made architecture, nor is it a technology. It is based on the observation that existing technologies are sufficiently advanced and deliver excellent results on their own. These results can be measured in various ways: the amount of data processed per unit of time, the ability to adapt datasets with different formats, or the capacity to scale infrastructure according to the size of the problem being addressed. Currently available technological solutions seem to allow for the design and implementation of IT systems that can be optimal according to any chosen evaluation criterion. However, corporate practice shows that technological capabilities and appropriate infrastructure alone are often not enough to achieve satisfactory business goals. This experience demonstrates that the key to success lies in the proper organization of processes related to data processing at the level of specialist teams, removing barriers, improving communication, and fundamentally changing the approach to data. We believe that the direct application of corporate experiences in the context of a city is not possible due to a completely different organizational structure, a complex system of dependencies, a different work culture, and a lack of technolog-

ical consistency across the entire organization. Furthermore, the objectives of each city department are focused on providing services to citizens rather than generating commercial profit. The urban environment might be managed differently at different levels, involving different competencies and levels of understanding among officials. It also includes subsidiary companies operating on behalf of the city council but having their own status and rules. Therefore, the image of the city as an institution appears more like a mosaic of diverse teams rather than a cohesive corporate entity. Experience from the project shows that implementing a new data processing concept and vision of their use in a broader context within a city government involves overcoming a number of challenges. As we anticipated, the direct adaptation of corporate patterns proved impossible. Therefore, the implementation of ODSC (Open Data platform for Smart City) in the City of Poznań required adopting a strategy composed of the following elements.

4.4.1 Partnership

The design and implementation of the ODSC platform became possible thanks to the longstanding partnership between the city of Poznań and PSNC. Years of collaboration have built significant trust on the city's part in PSNC's competencies in information technology. On the other hand, PSNC has developed a deep understanding of the specific nature of the city's operations, its digitization needs, and the potential challenges involved. A critical factor enabling the creation of a data platform based on the latest technologies and built according to the newest standards is the strong self-awareness among city officials. The recognition of the enormous potential within the data generated in the urban space, combined with the legal obligation to publish this data, inspired the initiation of another joint project. PSNC's experience in data management and processing, along with its knowledge of contemporary technologies, helped outline the preliminary vision of what ODSC should be and what requirements it should meet. It is important to emphasize that these efforts required significant openness and collaboration from both sides. Implementing ODSC involves changing the way many internal processes are carried out within the city administration and necessitates a fundamental shift in how the data at the city's disposal is perceived. Accepting these challenges was essential to building a modern and future-proof solution. In our view, the close cooperation between partners and mutual inspiration were key to achieving the initial successes in implementing ODSC.

4.4.2 Small steps

Implementing an IT system like ODSC is a significant organizational challenge. From the IT system's perspective, the platform consists of many tools that must be developed and/or integrated into a cohesive whole. Therefore, the contractor is expected to have a high level of proficiency in executing IT projects and a broad knowledge of technology. From the administration's perspective, implementing ODSC represents

a major shift in organization and work culture. Experience gained from implementing ODSC for the city of Poznań shows that creating a comprehensive project that encompasses all possible system elements is extremely difficult. This difficulty arises because some of the client's requirements and needs only become apparent during the project's execution. We believe that an effective way to address such challenges is through an incremental approach. The starting point for implementation should be a carefully crafted document that outlines the vision, assumptions, and goals set for the platform. Based on this document, it is possible to develop the technical specifications for the platform, which should provide a general framework for the IT work. Understanding the objectives and the means to achieve them allows the work to be divided into stages, during which ODSC will be enhanced with new functionalities. This approach gives the city considerable flexibility in prioritizing tasks depending on the importance of business requirements and strategic needs. Moreover, tools that are developed and implemented in stages reach future users relatively quickly. This enables feedback from users at a very early stage of the project, helping to validate the initial project assumptions. It also allows for the commencement of promotional and outreach activities, which are crucial for building understanding and proper perception of ODSC among users. Implementing the work in stages also offers the opportunity to spread the project's costs over time. However, for the contractor, staging the work introduces additional challenges in executing the IT project. Therefore, the contractor's experience in managing such projects and the mutual understanding of both the contractor's and the client's needs and limitations are crucial.

4.4.3 Understanding

The partnership between the client, the city of Poznań, and the contractor, PSNC, proved to be extremely helpful in the implementation of the ODSC project. The design and execution of such a platform should be seen as a process that partners must navigate together. The implementation of ODSC in Poznań began with an effort to understand the city's immediate needs and to identify the goal that should be achieved. Implementing a new IT system also requires understanding its impact on the current functioning of the city administration. Awareness of these relationships and a willingness to take actions that alter the fundamental perceptions and understanding of data paved the way for implementing the platform according to a concept that offers a practical opportunity to realize Smart City 4.0 in Poznań. The journey to a fully functioning ODSC, in addition to having a clear understanding of the direction of development and the ultimate goal, requires close monitoring of progress and proper interpretation of emerging circumstances. A partnership based on understanding each other's needs and mutual acceptance in the spirit of achieving a common goal is essential for the successful completion of the entire endeavor.

4.4.4 Data value

The ODSC platform should be seen as a tool for the development of the city and its transformation towards Smart City 4.0. At the core of this transformation is the recognition of the real value that data generated in the urban space represents. Data owners—municipal units, domain teams, city-owned companies—often view the data they possess as merely a byproduct of the processes they carry out. They fail to see the broader perspective of how their resources can be utilized and are often skeptical about the need to share data, seeing it as an additional obligation and an unnecessary expense. Therefore, it is essential to build awareness among data owners and providers of the valuable asset they hold. The ODSC platform serves as a tool to facilitate this awareness and to systematize the process of achieving it. A fundamental shift in the approach to data sharing—from viewing it as an unpleasant obligation to an internal commitment to maintaining one's own "data product"—can be seen as a measure of success in implementing ODSC in the urban environment.

4.4.5 Promotion and dissemination

One of the goals of carefully phasing the work on ODSC is to enable the early initiation of tasks related to promoting and spreading the concept of ODSC among future users. These users can be divided into internal users, consisting of employees of the municipal office and its subordinate units, and external users, such as residents, institutions, or companies that rely on city data to carry out their business processes. Internal efforts should primarily focus on changing the way data is perceived by those who manage it. Effectively building user awareness should be done by promoting examples that illustrate the advantages of the new approach. Theoretical arguments should be supported by demonstrating the tools developed within ODSC as aids for carrying out new tasks. It's also important to provide users with a comprehensive perspective on the use of ODSC so that they do not see their participation in this initiative as a mandated obligation, but as an opportunity to streamline internal processes within the office. Promotional activities aimed at external users should include showcasing the potential for utilizing the data provided by the city. This particularly concerns consistent policies related to the security and quality of published information, as well as data access mechanisms. External consumers play a stabilizing role and drive the further development of ODSC, both in terms of functional expansion and the volume of data published in this manner.

4.4.6 Driving scenario

For both implementation and promotional efforts, it is highly beneficial to define a so-called "driving scenario". This is an example scenario of a user who utilizes ODSC resources to carry out a complete data processing workflow and publish the results of their work. Having such a scenario is helpful in the early identification of problems that users might encounter with the system. It allows for quick intervention and appropriate adjustments to the assumptions. However, it is crucial to exercise

caution so that the final implementation of ODSC does not become overly tailored to the specific requirements of this scenario. Its support should be seen as an inspiration and a testing tool, rather than an end goal in itself. A well-defined driving scenario is also a significant asset for promotional and outreach activities, as it can effectively demonstrate most of the details related to how ODSC operates, including the benefits of using the platform.

4.4.7 Task management

When embarking on the implementation of ODSC for the city of Poznań, we were aware that not all the final elements of the platform were defined in sufficient detail. The final shape of individual functional modules is determined based on ongoing analysis of needs. The key here is to maintain alignment with the overall technical design and the goals outlined in the implementation strategy. This approach requires a delicate balance between the necessity of delivering the required functionality and the creation of solutions that serve as the foundation for further development. Proper planning, as well as close collaboration and mutual understanding between partners, play a vital role in this process. Given that the implementation timeline is extended and many requirements are defined during the project's progression, skillful management of the task list is extremely important. Special attention should be paid to the relationships and constraints between various stages. Any potential delays in execution, even if justified, could significantly impact the platform's development potential in the near future. Minor remarks and requirements affecting the scope of ODSC functionality should be recorded and assigned to the appropriate stage of implementation. The ultimate goal is to deliver ODSC in a form that meets the needs of specific users. The previously mentioned driving scenario mechanism, which allows for verifying the completeness of the implementation, can be helpful in properly maintaining the task list.

The key elements of the strategy developed during the implementation of ODSC for the city of Poznań have been presented above. These elements stem from direct experiences gained through collaboration with the city and have proven effective in the initial stages of the platform's deployment. We are confident that these experiences are universal in nature and can serve as practical guidance for future implementations.

4.5 Technical aspects

As mentioned in the previous section, the concept of Data Mesh avoids promoting specific technological solutions. Instead, it focuses on defining the requirements for how technology should be used by all stakeholders. In practice, an IT system aligned with the Data Mesh concept can be implemented using any technology suitable for solving a given problem. In the case of developing a data platform for the city of Poznań, the final product had to take the form of an IT system that met the business

requirements defined by the city and provided the necessary functionalities. The implementation process was divided into stages, according to the priorities set by the city. Below, we present selected technical aspects that we identified as the most significant in the development of the ODSC platform.

One of the key issues to address in the early stages of the project was the decision regarding the target environment on which the platform would be deployed. To maintain full control over the infrastructure and minimize maintenance costs, the implementation was carried out using computing resources provided by PSNC. Among the possible infrastructure models—bare metal, Infrastructure as a Service (IaaS), or Platform as a Service (PaaS)—we chose to base the deployment on the PaaS model. This decision was primarily driven by the clear delegation of responsibilities related to infrastructure maintenance and continuity to dedicated administrators. In this model, computing, storage, and network resources are treated as services that can be freely utilized within the allocated pool. At the same time, technical users, especially those responsible for the development and deployment of services, remain independent in utilizing these resources, reducing the need to involve the administration team to a minimum. Along with well-prepared deployment and service maintenance automation processes, this solution forms the foundation for implementing the principle of the Self-Serve Data Platform, as formulated in the Data Mesh concept. The use of infrastructure in the PaaS model also allows for the optimal use of resources by enabling the allocation of the appropriate amount of resources necessary for the efficient operation of ODSC. Given the iterative and extended process of developing the ODSC, this solution is crucial for minimizing and properly controlling the costs associated with resource utilization.

The second essential element in implementing the Self-Serve Data Platform principle is the automation of all processes related to deploying and maintaining the services available within the ODSC platform. For services developed from scratch to meet the platform's needs or to perform specific tasks and functions according to the "data domains" principle, the entire software development process must be automated. This particularly applies to the stages of building applications and services, automated testing, and deploying and launching services in the target environment. Our experience from the project indicates that the time invested in preparing and executing the first deployments of this type pays off multiple times with each subsequent deployment. This is made possible by developing CI/CD process configuration templates, which can be successfully reused across multiple applications. Similarly, services created according to these same patterns allow for a drastic reduction in the time required to prepare the initial configuration. Reusing existing configurations, application templates, and software components is, in our view, key to achieving full flexibility in implementing the requirements arising from the needs of "data domains," while simultaneously minimizing the number of work hours necessary to accomplish them. This approach is crucial for providing the proper IT support in the form of APIs for "data products," a key element in the Data Mesh concept.

According to the Data Mesh concept, a "data domain" is a virtual space defined

by the domain-specific context of the data processed within that domain. This space encompasses not only the physical data but also interdisciplinary specialists responsible for the proper interpretation and effective processing of the data, as well as the processes defined by these specialists. The final output of these processes is the generation of information and services, which constitute the essence of the "data product" concept. A complete ODSC (Open Data platform for Smart City) should therefore provide the necessary functionalities to support the processes related to data preparation and analysis. In our view, it is crucial to deliver these functionalities in an integrated manner so that users can complete their tasks without needing to transfer any data outside the platform and can utilize the computational resources available within the platform. To achieve this, we decided to base this element of the platform on a standardized methodology known as MLOps (Machine Learning Operations) introduced by [48]. MLOps is a set of practices that combines the development of machine learning (ML) models with their operational deployment and maintenance in production. This concept draws inspiration from DevOps (Development Operations) and aims to streamline and automate ML-related processes. The main principles of MLOps include:

- automated deployment and maintenance of ML models, enabling quick and reliable implementation of new model versions
- ensuring the reproducibility of ML experiment results, which is crucial for validation and audits
- the ability to scale models and infrastructure to handle increasing amounts of data and computational demands
- continuous monitoring of model performance in production to detect issues and avoid negative impacts on system operations
- facilitating collaboration between data scientists, data engineers, and operations engineers.

The functionalities resulting from MLOps practices have been integrated into the ODSC platform using available open-source software. These functionalities allow for the execution of any data processing tasks, even those unrelated to machine learning. Thanks to the platform's modular design, it is possible to easily expand this element with new tools and libraries necessary for data analysis.

A crucial element influencing the perception of ODSC as an integrated data platform is the thoughtful implementation of user identity management. The modular structure and the awareness of the need to build a range of applications and services that will utilize the platform's resources in the near future led to the adoption of the Single Sign-On (SSO) concept. This service is responsible for authenticating and authorizing user access to individual services. From the perspective of effective integration, these processes are based on widely accepted and commonly used standards, such as OpenID Connect for authentication and OAuth2 for access authorization. Experience gained during the project shows that it is very important for

all elements of the platform, services, and applications to use the same, consistent method of user authentication and authorization. This approach lowers implementation costs by enabling the development of standardized procedures and helps simplify the management of applications and their code. This also applies to system services, such as data (file) storage services. Ensuring vertical integration from user applications, through services (APIs), down to system services (storage) significantly simplifies the overall system architecture. It also unlocks the potential for using and sharing data within the platform without the need to copy it or maintain consistency between services.

The implementation of ODSC for the city of Poznań is iterative in nature. In each stage, the functionalities most needed by the city at that moment are delivered. The first to be implemented was the resource catalog. Within the platform, the resource catalog acts as an integrating element, consolidating information about the datasets available within the urban space. Each of these datasets is interpreted as a separate "data product" according to the DM concept. Starting early with the catalog and observing the processes required to prepare a dataset for publication provided excellent insights into how the concepts of "data domain" and "data product" could function in the context of the city. For the officials working with the platform, the initial delivery of the catalog helped build awareness of the value of the data they possess and initiated a mental shift in how they perceive it. This is one of the key changes advocated by the "DM" concept. Additionally, the operational catalog serves as a promotional tool that aids in spreading the knowledge and ideas of ODSC within the office. It also facilitates public communication and disseminates knowledge about the data available within the city's resources, thus fulfilling the office's obligation to provide data access. From an implementation perspective, providing the resource catalog to users first allowed for the systematization and prioritization of requirements that will be addressed in subsequent stages. Ultimately, we are convinced that the decision to implement the resource catalog first was beneficial, as it met expectations, initiated positive changes within the office, and positively influenced the platform's development process.

The second set of functionalities we decided to implement within ODSC is a tool that supports collaboration between users. As previously mentioned, the "data domain" concept requires close cooperation among users within interdisciplinary teams, with the ultimate goal of creating new "data products." These teams include data analysts, data scientists, programmers, domain experts, and DevOps specialists. For data processing by analysts, dedicated tools were provided to create the analytical part of the platform. To facilitate team organization, improve communication, and simplify the process of managing permissions, a project management service was implemented. From the perspective of ODSC, a project integrates data, models, and processes necessary to effectively carry out data analysis or build a machine learning model. Within this element, the possibilities of applying a consistent access control mechanism to all services and resources that will be used within the platform in the future were implemented and tested. The development of this module is based on the experience gathered by PSNC in projects focused specifically on data analysis.

We are confident that such functionality will find its place within ODSC and will be planned for the next stages of deployment.

At the beginning of the implementation of an open data platform, it is important to consider which approach will be optimal from the city's perspective. The choice typically boils down to deciding whether to purchase a ready-made solution or to design and build a system tailored to the specific requirements and needs of the city. It is difficult to definitively assess which option is more advantageous, as it depends on the local conditions of the particular city. In the case of the city of Poznań, a decision was made to collaborate with PSNC as a technology partner, proven in the execution of previous projects. This approach allowed the project to be divided into stages. The first step was to define the vision of the platform as a whole, identifying basic business needs and functional requirements. At this stage, the city recognized the need for organizational changes and a new approach to data management. Building awareness among city officials was seen as an ongoing process that began as early as possible in the project. The second phase of the platform's implementation focused on delivering functionalities that were key from the city's perspective. The chosen functionalities were those that supported promotional activities within the city office, aiming to build a community of informed users who see the value and necessity of collaborating within the platform. The subsequent stages of the platform's development will focus on implementing and integrating new functionalities identified by the city. The success of the entire process will hinge on achieving a scale effect, where, after surpassing a threshold of users and resources on the platform, its further use will become essential for the city's internal processes and its communication with the outside world.

Large Language Models in the service of Open Data platform for Smart City

It is understandable and partially acceptable that implementing a data platform for a modern city requires a significant effort. However, in order to use the final system effectively on the daily basis, it must not cause the slightest difficulty for stakeholders. Data providers, being key stakeholders, must feel fully supported in the processes they undertake, and they are entitled to expect this level of assistance. This necessitates the automation of processes wherever possible. Solutions that simplify the most challenging and time-consuming tasks will be of particular value. In the context of ODSC, one of the most critical elements that can be addressed in this way is the automation of generating metadata related to the data products themselves.

This chapter is organised as follows. The section 5.1 explains reasons why it is necessary to automate repetitive tasks in large scale data platform. Section 5.2 gives a short overview of the general ability of LLM's to process data. Sections 5.3, 5.4, 5.5 presents in details conducted experiments designed to verify if LLM's have ability to support particular ODSC user scenarios. The final findings are summarized in 5.6.

The results of the experiment were previously published by the author in [13].

5.1 Motivation

European Union (EU) official documents propose following definition of open data: "Open (Government) Data refers to the information collected, produced or paid for by the public bodies (also referred to as Public Sector Information) and made freely available for re-use for any purpose" [49]. EU document refers to further definition of openness given by [50]. The practical implementation of the openness concept can be carried out according to the FAIR principles, which seems to be complementary to open data concept by improving the accessibility and usability of data. Open Data focuses on removing access barriers, while FAIR ensures that data, once accessible,

is usable and beneficial.

The FAIR principles described in 51 are a set of guidelines designed to ensure that data is shared in a way that maximizes its utility and accessibility. FAIR stands for Findable, Accessible, Interoperable, and Reusable. The principles have the following interpretation:

- Findable - data should be easy to find for both humans and computers. This is achieved through the use of rich metadata and unique identifiers, such as DOIs, ensuring that datasets can be located and referenced accurately.
- Accessible - once found, data should be easy to access, ideally through standardized communication protocols. This includes clear guidelines on who can access the data and under what conditions, ensuring that even when data is behind access controls, the metadata remains accessible to enable discoverability.
- Interoperable - data should be compatible with other datasets and tools, facilitating integration and analysis across different platforms. This involves the use of shared vocabularies, ontologies, and standards to ensure that data can be combined and understood in different contexts.
- Reusable - data should be well-documented and richly described so it can be used in future research and applications.

This requires clear information on the data's provenance, usage licenses, and the conditions under which it can be reused, ensuring that data can be applied in new and varied ways beyond its original purpose.

Data Mesh data product concept is a further extension on FAIR principles, being their specification and clarification. These include attributes of data product such as: discoverable, addressable, understandable, trustworthy, natively accessible, interoperable, valuable and security 22.

A review of the above principles, both general ones derived from FAIR and specific ones describing the Data Mesh data product, clearly shows the great importance of proper preparation of data for publication. Apart from the technical aspects, the implementation of which is necessary for the proper implementation of data sharing, a proper understanding of the characteristics, context and interpretation of the data being shared is crucial to the fulfilment of the above principles. The quality of metadata describing data sets determines the majority of processes that are necessary to achieve even the most basic goals, such as data discoverability or reusability. In the context of the city data platform, this means that it is necessary to generate extensive and precise descriptions for a huge number of data sets, for which detailed documentation is often lacking. Knowledge regarding the proper interpretation of the data set at best comes directly from the data owner, and at worst is lost as a consequence of HR activities. In view of the above, there is a need for a tool that would provide support in the implementation of basic, repeatable tasks related to the preparation of valuable metadata. This applies to both the generation of human-understandable descriptions and information useful in further machine processing

of data sets. The primary goal of such a tool is, among other things, to improve the quality of the information generated and to shorten the time needed to prepare it. As a result, the general accessibility barrier of the data platform for its future users in terms of creating and publishing new data sets using it will be lowered. Ease of use and minimization of workload are therefore key to the good reception of the platform by city officials and employees. Their positive reaction will translate into an increase in the potential success of the data platform as whole.

Recently developed Large Language Models (LLM) have so far shown great effectiveness in generating valuable content based on presented source data. The ease of use of LLM models and their flexibility in terms of generated texts may prove useful in automating the process of publishing on the open data platform. At the same time, their potential can be used to implement further application scenarios with minimal implementation costs. The aim of the experiments described below is to investigate the possibilities of practical use of LLM to support users of the open data platform in implementing their individual tasks.

5.2 Large Language Models in structured data analysis

The ongoing generative AI revolution marks a transformative period in the evolution of large language models (LLMs), reshaping the landscape of data analysis, content creation, and artificial intelligence research. This revolution commenced with the release of GPT (Generative Pretrained Transformer) by OpenAI in June 2018, followed by its more powerful successor, GPT-2, in February 2019. The trend of exponential growth in capabilities and sophistication continued with GPT-3, launched in June 2020, setting new benchmarks for the artificial generation of human-like text. Google, for instance, introduced Meena in 2020, which demonstrated advanced conversational capabilities, and followed with the release of LaMDA in May 2021. Meta (formerly Facebook) also entered the arena by releasing OPT-175B in May 2022, with the aim of democratizing access to powerful LLMs by making it open source. This move towards open-source models was further underscored by the release of BLOOM by Hugging Face in July 2022, which embodied the collaborative spirit of the AI research community. Following these, last two years saw the release of Jurassic-1 Jumbo from AI21 Labs, a massive LLM trained on a vast dataset of text and code. More recently, 2023 brought Bard and Gemini from Google AI, Claude from Anthropic AI, and LLaMA 2 from Meta, all showcasing further advances in reasoning, dialogue, and content creation. More recently, models like GPT-4 and others from various organizations have pushed the boundaries even further, although specific release dates for some remain under wraps due to the competitive nature of the field.

Large language models (LLMs) have demonstrated remarkable utility in analyzing structured data, a domain traditionally navigated through more conventional statistical and machine learning techniques. LLMs, through their deep understanding of natural language, can infer relationships, trends, and insights from structured data presented in natural language form or converted into such information through

preprocessing steps. They accomplish this by essentially translating the structured data into a "language" they can understand, thereby leveraging their natural language processing capabilities to predict outcomes, generate summaries, or identify anomalous patterns within the data.

LLMs break barriers with nonstandard tasks that necessitate a more "human" grasp of dataset semantics, beyond mere statistical correlations. This ability comes from their training on diverse datasets that include factual information, reasoning, and contextual situations. LLMs can engage with datasets that require an understanding of the underlying principles or concepts, identifying patterns, or generating predictions that are not based solely on numerical analysis but also on logical or thematic coherence. Through this advanced understanding, LLMs are not just analyzing data; they are interpreting it, providing insights that are both quantitatively robust and qualitatively useful.

The capability of LLMs to comprehend, process, and generate human-like text has been thoroughly examined and leveraged in many applications. However, a less explored terrain is their potential proficiency in handling tabular data, characteristic of the ODSC scenario. We evaluate the extent to which LLMs possess an "understanding" of tabular data structures and content to the extent that they can function effectively as custodians of large datasets. Specifically, we explore the capability of LLMs to perform tasks such as providing succinct summaries of datasets, recommending more intuitive feature names based on contextual understanding and suggesting datasets that are pertinent to specific queries or business needs.

5.3 Experiment 1: identify and label concepts

The goal of the initial experiment is to determine whether the LLM can accurately identify and label the concept represented in a database based on a small sample. The LLM only receives raw data, without any clues about the importance of individual attributes, their semantics, and their relationships. To ensure these conditions, pseudonymization was used to disguise the names of individual attributes and their random placement in the data structure. The original attribute names were replaced with four-character codes using base64 encoding, which was created with the `openssl` tool. To reduce the likelihood of hallucinations in the model, the general context in which a particular dataset operates was specified.

In the experimental setup, `gpt-4-turbo` [52] was presented with a dataset comprising 50 randomly selected tuples from a table, along with a minimalist description that imparts basic context about the table. The primary objective assigned to the LLM was to autonomously discern the underlying semantics of the features contained within these tuples and subsequently propose relevant names for these features. LLM was tasked with identifying and delineating the relationships between features, which required a degree of 'understanding' of the structure and content of the data. The LLM was also asked to synthesize a concise and informative description of the dataset as a whole, with the objective of providing users with a clear and

immediate understanding of the nature and scope of the data. The gpt-4-turbo model was accessed through OpenAI's API, using the langchain programming library. The experiment was conducted in late March 2024. The LLM was tested on four datasets from the ODSC of the city of Poznań:

- *Graves*: information about the graves located in the city cemetery,
- *Address Points*: address points, their locations, categories and affiliation to various functional areas of the city
- *Bike Stations*: location and current occupancy status of city bike stations,
- *MPK Stops*: information about public transport stops, their location, categories, types of vehicles and possible transfers.

For each dataset, we have crafted a query with the following requirements. LLM was asked to provide the characteristics of the dataset, but the query explicitly forced the model to focus on feature values and their relationships, while ignoring the order of features when determining the relationship between them. The query provided requirements for the output format and contents. Ultimately, the sample query took the following form:

Analyze following JSON document where each feature represents description of single address point in the city of Poznań, Poland. Analyze values and relationships between all attributes. The position of the attributes in the document does not matter. For all attributes in section properties suggest new descriptive names. New names must explain meaning of the attribute value and make it easier to understand what each attribute represents. For each attribute suggest the most appropriate data type. Leave the old name if there is lack of data or there is a very low certainty of the concept identification. As a result print triples using following pattern: old attribute name, new attribute name in snake case, data type, attribute interpretation.

followed by a random sample of data, formatted as a JSON document (see Fig. 5.1).

Table 5.1 presents the summary of meta-data generation for all tables. The detailed results for individual tables are presented in Table 5.2 (*Graves*), Table 5.3 (*MPK Stops*), Table 5.4 (*Bike Stations*), and Table 5.5 (*Address Points*).

5.4 Experiment 2: generate human like descriptions

The goal of the second experiment was to test the ability of the LLM model to generate an informative description of the dataset in a way that would mimic the description prepared by a domain expert. Such a description should include a summary and explanation of the meaning of individual attributes, identify the domain area to which the dataset applies, and identify potential areas of application in future

```
{
  "features": [
    {
      "geometry": {
        "coordinates": [
          16.7779975682727,
          52.4801856747057
        ],
        "type": "Point"
      },
      "id": 88835,
      "properties": {
        "Hi1E": "Sektor III",
        "IdFd": "60-480",
        "JGa4": 1,
        "Ki0W": null,
        "Lihs": "turystyczna 1",
        "MiAs": "I-Kiekrz",
        "RjaC": "Turystyczna",
        "UUBF": "VI",
        "V9C8": "Kiekrz",
        "W0bd": 1202,
        "ZE3B": "Aglomeracja Poznań",
        "dkJA": "Komisariat Policji Poznań-Jeżyce",
        "epFx": "Zurbanizowany",
        "lWQ9": "turystyczna",
        "lsMf": "Turystyczna",
        "nDES": 28,
        "nt1q": "ul.",
        "q4tJ": "pw. Świętego Michała Archanioła i
          Wniebowzięcia Najświętszej Maryi Panny
          w Poznaniu - Kiekrzu",
        "r4Lg": "turystyczna",
        "rswz": 0,
        "tCdY": 188,
        "xKXa": "1"
      },
      "type": "Feature"
    },
    [...]
  ]
}
```

Figure 5.1: Example of a single dataset record

analyses. The LLM was presented with a description of the dataset structure containing information about the names of individual attributes, their type, and a short description indicating the potential meaning of the attribute if its name was not self-descriptive. The description of the dataset was prepared in the form of a JSON document (see Fig. 5.2):

Specifically, for the experiment, the expert created descriptions for 18 datasets, which were generated by or for urban services in Poznań, Poland. Datasets used

Table 5.1: Generation of feature descriptions using the LLM, A: number of features, B: number of empty features, C: number of recognized features, D: number of correctly recognized features, E: percent of recognized features, F: percent of non-empty recognized features, G: percent of correctly recognized features

dataset	A	B	C	D	E	F	G
Graves	18	5	11	8	61,11%	84,62%	72,73%
Address Points	22	1	14	13	63,64%	66,67%	92,86%
MPK Stops	5	0	5	4	100,00%	100,00%	80,00%
Bike Stations	5	0	5	3	100,00%	100,00%	60,00%

in the experiment are listed in Table 5.6. A set of 15 queries, categorized into four groups, was created to produce a description of the dataset that aligns with the experiment’s assumptions. These queries were parameterized to enable dynamic referencing of the dataset they would be applied to. The query responses were simultaneously stored in a text file for validation purposes and in a vector database for use in the subsequent experiment. The queries were constructed so as to discover the structure of the data and possible access techniques, to generate feature descriptions and interpretations, to recognize the domain of the data, and to discover future analytic opportunities with respect to the data. Below are the LLM’s responses to selected queries for the *Bike Stations* dataset.

Question: Identify all features available in the Bike Stations dataset. Use feature titles to understand their meaning. Based on the feature meaning discover possible categories to which each feature could be assigned. The example categories could be: location address or geospatial characteristics. List the categories together with city bike station dataset features assigned to each category.

Answer:

- 1 Features assigned to categories:
- 2 1. Geospatial characteristics:geometry (coordinates)
- 3 2. Bike station details: bike_racks, bikes, free_racks, label
- 4 3. Time-related information: updated

Question: Considering type and nature of each identified features in Bike Stations dataset, for each feature suggest two different tools or programming libraries which are most suitable for processing this kind of data. Suggest also an analytical technique which is the most relevant for particular feature.

Answer:

- 1 1. For processing the coordinates feature in the city bike station dataset, suitable tools or programming libraries could include:
- 2 - GeoPandas: A Python library for working with geospatial data that provides easy-to-use tools for handling geometric data structures.
- 3 - Leaflet: A JavaScript library for interactive maps that can be used to visualize and interact with the geographic coordinates.
- 4 Relevant analytical technique: Spatial analysis, which involves analyzing patterns and relationships within geographic data.
- 5 2. For processing the bike racks feature in the city bike station dataset, suitable tools or programming libraries could include:

Table 5.2: Names and descriptions of the *Graves* dataset features generated by the LLM

attribute name	description	type	code	LLM name	LLM description	LLM type	uniq values
cm_id	graveyard identifier	number	0kqa	graveyard_id	The id of the graveyard where the grave is located	int	1
cm_nr	graveyard number	number	gawU	grave_id	Unknown attribute due to lack of data.	int	50
g_date_birth	person's date of birth	date	wRc2	birth_date	The date when the person was born	date	50
g_date_burial	person's burial date	date	sAK1	burial_date	The date when the person was buried	date	50
g_date_death	person's date of death	date	clgE	death_date	The date when the person died	date	50
g_family	type of the grave	string	dXUz	-	Unknown attribute due to lack of data.	string	0
g_field	graveyard field identifier	string	O8m2	-	Unknown attribute due to lack of data.	string	0
g_name	person's name	string	Q+R2	first_name	The first name of the person buried in the grave	string	48
g_place	place number	string	pleb	grave_location	The location of the grave in the graveyard	string	50
g_quarter	graveyard quarter	string	ambd	grave_type	The type of the grave (P for private, L for public)	string	21
g_row	row identifier inside quarter	string	PBeA	grave_number	The number of the grave in the graveyard	string	48
g_size	size of the grave	string	6lfn	-	Unknown attribute due to lack of data.	string	0
g_surname_name	surname and name	string	dsfp	-	Unknown attribute due to lack of data.	string	0
g_surname	person's surname	string	dyjH	last_name_lowercase	The last name of the buried person in lowercase	string	50
paid	is the fee paid	number	PaI5	-	Unknown attribute due to lack of data.	int	1
print_name	capital name	string	nqu	full_name	The full name of the person buried in the grave	string	48
print_surname_name	capital surname and name	string	vWTD	-	Unknown attribute due to lack of data.	string	0
print_surname	capital surname	string	dK4E	last_name	The last name of the person buried in the grave	string	50

Table 5.3: Names and descriptions of the *MPK Stops* dataset features generated by the LLM

title	description	code	LLM name	LLM description	LLM type	uniq values
stop type	distinguish between tram and bus stop	Dlkr	direction_count	The number of directions at the stop	int	2
lines	list of public transport lines using this stop	EkmW	line_numbers	List of lines that stop at the stop	List[int]	43
name	common name	clND	stop_name	Name of the stop	string	49
string stop type	distinguish between tram and bus stop	PVtE	vehicle_type	Type of the vehicle that stops at the stop (T for tram, A for bus)	string	2
description	stop type description	IOLd	stop_type	Type of the stop (tram or bus)	string	3

Table 5.4: Names and descriptions of the *Bike Stations* dataset features generated by the LLM

attribute name	code	LLM name	LLM description	LLM type	uniq values
bike_racks	yeNA	total_docks	The total number of docks at the bike station. It is a string that can be converted to an integer.	int	11
bikes	IXWC	broken_bikes	The number of broken bikes at the bike station. It is a string that can be converted to an integer.	int	1
free_racks	Xc78	available_bikes	The number of bikes currently available at the bike station. It is a string that can be converted to an integer.	int	14
label	Hlkk	station_name	The name of the bike station	string	50
updated	Ebor	last_update	The date and time of the last update of the bike station's data. It is a string in the format "YYYY-MM-DD HH:MM"	date	1

Table 5.5: Names and descriptions of the *Address Points* dataset features generated by the LLM

attribute name pl	code	LLM name	LLM description	LLM type	uniq values
adres	Ljhs	secaddress_lowercase	represents the address in lowercase	string	50
gimnazjum	rswz	-	unknown attribute, lack of data to identify the concept.	string	5
goap	Hi1E	sector	represents the sector of the city where the address point is located	string	3
goap_odpady	ZE3B	agglomeration	represents the agglomeration where the city is located.	string	2
kod	IdFd	postal_code	represents the postal code of the address point	string	8
komisariat	dkJA	police_station	represents the nearest police station to the address point	string	2
krajobrazowa	epFx	urbanization_status	represents the urbanization status of the address point.	string	2
nazwa_parafii	q4j	parish	represents the parish the address point belongs to	string	4
nr	xKXa	house_number	represents the house number of the address point	string	36
obwoady_ro	JGa4	-	unknown attribute, lack of data to identify the concept	int	3
obwoady	iCdY	-	unknown attribute, lack of data to identify the concept.	int	5
okregi_ro	MiAs	district	represents the district of the city where the address point is located	string	5
okregi	UUBF	-	unknown attribute, lack of data to identify the concept	string	2
parafie	Wobd	-	unknown attribute, lack of data to identify the concept	int	4
podstawowa	nDES	-	unknown attribute, lack of data to identify the concept	int	4
rejon	KlOW	-	unknown attribute, lack of data to identify the concept	-	0
samorzady_lokalne	V9C8	neighborhood	represents the neighborhood of the city where the address point is located	string	4
typ	nt1q	street_prefix	represents the prefix of the street name (e.g., "ul." for "ulica" which means "street" in Polish).	string	1
wg_imienia	r4Lg	-	unknown attribute, lack of data to identify the concept.	string	38
wg_imienia_wydruk	RjaC	street_name	string, represents the name of the street of the address point	string	38
wg_nazwiska	IWQ9	street_name_lowercase	represents the name of the street of the address point in lowercase	string	38
wg_nazwiska_wydruk	IsMf	street_name_capitalized	represents the name of the street of the address point with the first letter capitalized	string	38

```

{
  "city_bike_station": {
    "type": "object",
    "title": "city bike station",
    "service_url":
      "https://www.poznan.pl/mim/plan/map_service.html?
      mtype=pub_transport&co=stacje_rowerowe",
    "properties": {
      "geometry": {
        "type": "Point",
        "crs": "EPSG:4326",
        "coordinates": {
          "title": "coordinates",
          "type": "double",
          "description": "coordinates of exact point"
        }
      },
      "bike_racks": {
        "title": "bike racks",
        "type": "number",
        "description": "total number of bike racks available
          at the station"
      },
      "bikes": {
        "title": "bikes",
        "type": "number",
        "description": "number of bikes parked at the station"
      },
      "free_racks": {
        "title": "free racks",
        "type": "number",
        "description": "number of free racks available
          at the station"
      },
      "label": {
        "title": "label",
        "type": "string",
        "description": "human readable name of the station"
      },
      "updated": {
        "title": "updated",
        "type": "timestamp",
        "description": "represents date when station
          status was updated"
      }
    }
  }
}

```

Figure 5.2: Example of the *Bike Stations* dataset description.

- 6 - Pandas: A Python library for data manipulation and analysis that can handle numerical data efficiently.
- 7 - Matplotlib: A Python library for creating static, animated, and interactive

Table 5.6: Datasets used in Experiment 2

Address Points	address points, their location, categories and affiliation to various functional areas of the city
Bike Stations	city bike stations, their location and current occupancy status
Election Districts	names of all election districts in the city
Env Decisions	case number, status and reference to parcel covered by the environmental decision
Graves	description of the graves located in the city cemetery
Graveyards	names of graveyards and API parameters which can be used to interact with services
Historic Buildings	address, ways of contact, description and classification of historic buildings in the city
MPK Lines	description of each public transport line available in the city
MPK Stops	public transport stops, their location, categories, types of vehicles and possible transfers
MPK Tickets	all locations where public transport tickets are sold including category and opening hours
Parking Machines	location of the machine and accepted payment methods
Polling Places	location, district name and number and accessibility for disabled people
Properties Sale	properties for sale including their location, parcel details, comments, contacts, type of offer and links to other services containing further information
School Districts	names and numbers of school districts, together with recruitment campaign codes
SD Plans	spatial development plans, references to official documents and dates indicating history of the plan development
SD Plans Progress	spatial development plans which are under development, references to official documents and date when work was initially started
Streets	different categories of places: streets, squares and others, including patron names and street length.
Wireless	locations and names of all wireless access points available in the city

visualizations in Python.

- 8 Relevant analytical technique: Descriptive statistics, which can be used to
summarize and analyze the distribution of bike racks at different stations.
- 9 3. For processing the updated feature in the city bike station dataset, suitable
tools or programming libraries could include:
- 10 - `datetime`: A Python module for manipulating dates and times that can be used to
handle timestamps.
- 11 - `Moment.js`: A JavaScript library for parsing, validating, manipulating, and
formatting dates and times.
- 12 Relevant analytical technique: Time series analysis, which involves analyzing data
collected over time to identify patterns and trends.

Question: *Elaborate Bike Stations dataset domain context description. Explain relationship between Bike Stations dataset and urban lifestyle or city inhabitants habits.*

Answer:

- 1 The city bike station dataset provides information about bike stations in a city, including details such as the coordinates of the station, the number of bike racks available, the number of bikes parked at the station, the number of free racks available, the name of the station, and the last time the station status was updated. This dataset is relevant to urban lifestyle and city inhabitants' habits as it reflects the availability and usage of bike-sharing services in the city.
- 2 Bike stations play a crucial role in promoting sustainable transportation options, reducing traffic congestion, and improving air quality in urban areas. City inhabitants who use bike-sharing services can benefit from a convenient and environmentally friendly mode of transportation for short trips around the city. The dataset can also provide insights into the popularity of cycling as a means of commuting or recreation in the city, as well as the overall interest in promoting a healthier and more active lifestyle among residents.

- 3 By analyzing the data from city bike stations, urban planners and policymakers can make informed decisions about expanding bike-sharing networks, improving infrastructure for cyclists, and promoting initiatives to encourage more people to use bicycles as a mode of transportation. Ultimately, the city bike station dataset contributes to creating a more sustainable and livable urban environment that supports healthy and active lifestyles for city inhabitants.

Question: Considering Bike Stations dataset domain context and data origin describe two possible analytical scenarios which can be carried out by professional data scientist. Each scenario should include problem statement and goal of the analysis.

Answer:

- 1 Scenario 1: Problem Statement: A city wants to optimize the distribution of bikes among its bike stations to ensure that there are enough bikes available for users at all times.
- 2 Goal of the Analysis: The data scientist can analyze the historical data of bike racks, bikes, and free racks at each station along with the timestamps of when the station status was updated. By identifying patterns in bike usage and station availability, the goal is to create a predictive model that can recommend the optimal distribution of bikes among stations to minimize the chances of stations running out of bikes or racks.
- 3 Scenario 2: Problem Statement: A city wants to improve the efficiency of its bike sharing system by identifying stations that are consistently underutilized or overutilized.
- 4 Goal of the Analysis: The data scientist can analyze the data on bike racks, bikes, and free racks at each station to identify stations that have a consistently low or high number of bikes available. By clustering stations based on their usage patterns, the goal is to recommend adjustments to the bike distribution system to ensure a more balanced utilization of all stations and improve the overall efficiency of the bike sharing system.

5.5 Experiment 3: support in decision making

Experiment 2 resulted in the generation of detailed descriptions of 18 datasets, including their individual attributes and overall characteristics. These descriptions provided a domain context and suggested potential applications for future analyzes. The purpose of the third experiment was to verify whether, based on the generated information, the LLM can support the analyst in selecting datasets that are best suited for use in a given research scenario. The experiment was designed with the assumption that the primary user interacting with the system would be a professional analyst or statistician seeking to perform an analytical task. The user's objective is to identify datasets that could aid in the execution of this task. Nine system queries were created, each representing a potential research issue. Each query was divided into three sections: the first defined the user's role; the second detailed the problem and the objective of the analytical task; the third provided guidelines for the model about the range of source information to be used for the response and the format of the response itself. The questions were formulated in a manner that ensures that the analytical task's content does not directly refer to the names of the datasets.

An example query: *As a statistician, I would like to analyze the frequency of occurrence of city residents' names and surnames. I am interested in identifying and tracking general trends over a long period of time and how names and usernames function in public space. Suggest the most relevant datasets from the context which can support this analytical task. List dataset names and justify the choice explaining why particular datasets are suitable for my use case.*

Model response: *Based on your analytical task of analyzing the frequency of occurrence of city residents' names and surnames over a long period of time to track trends and understand how names and surnames function in the public space, the most relevant datasets from the context provided would be the street dataset and the grave dataset. 1. Street Dataset: - Justification: The street dataset contains information about patron surnames and names associated with various streets in the city. By analyzing this dataset, you can track the frequency of occurrence of different surnames and names over time, providing insights into the historical and cultural significance of these names in the public space. This dataset can help you understand naming trends, patterns, and changes in the city's streets, reflecting the evolution of the urban environment and its inhabitants. 2. Grave Dataset: - Justification: The grave dataset can also be valuable for your analysis as it contains information about the names and surnames of deceased individuals buried in the city. By studying this dataset, you can track the frequency of occurrence of names and surnames associated with burial sites over time, providing insights into population demographics, historical naming conventions, and cultural practices related to death and remembrance. This dataset can complement the street dataset by offering a different perspective on the usage and significance of names and surnames in the public space. By combining insights from both the street dataset and the grave dataset, you can gain a comprehensive understanding of how names and surnames have been used, evolved, and functioned in the city's public space over an extended period of time. This analysis can help you identify trends, patterns, and correlations related to city residents' names and surnames, contributing to a deeper understanding of urban lifestyle and habits.*

5.6 Experiment conclusions

The range of tasks performed by the LLM model within the proposed experiments can be divided into two main groups: concept identification - involving naming attributes based on their values; and content generation - involving generating descriptions based on the provided context. The context included both metadata describing the dataset and hints provided in the model's instructions. During the experiment, attention was paid to the differences in the effectiveness of performing specific tasks depending on the model used. The task of concept identification, which also involved searching for relationships between attributes, was more effectively performed by the model with a larger context and trained on more recent data. For generating descriptive content based on the provided information, a model with a smaller context achieved sufficiently good results. Therefore, it is possible and justifiable to select an appropriate LLM model based on the task it is to perform. This is particularly significant in the context of using commercial models, where payment is made based

on the amount of data (tokens) processed by the model.

Experiment 1 demonstrated the model's ability to effectively interpret the data presented to it and assign names that define the context and meaning. The language model showed high effectiveness in interpreting attributes whose values are categorical and represent concepts that the model might have previously identified in a different context. Examples of such attributes include names, surnames, addresses including street names, and proper names of objects or categories. The model also demonstrated very good effectiveness in identifying names of alphanumeric attributes with a structured form, such as dates, times, and geographical coordinates. In these cases, the interpretation of the attribute's meaning results from recognizing the convention appropriate for that type of data. The same characters recorded according to different conventions will be interpreted differently. The language model clearly performs worse in interpreting purely numerical attributes. This applies to situations where there is no context arising from the format or the relationships between other attributes. In such cases, the model usually cannot unequivocally indicate the interpretation of the analyzed attributes, which directly results from the lack of context that could be interpreted.

In the case of analyzing a dataset consisting of many attributes, the model can consider the relationships between individual attributes in its interpretation. One premise for such interpretation can be the order of attributes in the dataset. To exclude the possibility of misinterpretation and potential model hallucinations, it is suggested that in tasks involving the discovery of attribute names, the model should be presented with code names consisting of random alphanumeric strings. At the query engineering level, the model should also be instructed to ignore the order of attributes in the dataset for the purpose of interpretation. This approach helps to limit incorrect interpretations based on misleading premises. Further query engineering allows for prompting the model to interpret the semantic relationships between attributes based solely on their values. An example of such an interpretation is three dates from the "graves" dataset, which, thanks to a properly constructed query, were correctly interpreted by the model as the date of birth, death, and burial, respectively, based on ordering the attribute values from oldest to newest.

An additional piece of information that the LLM model can effectively use when interpreting source data is the context provided by the user at the time of presenting the dataset. Indicating to the model the subject matter of the presented data clearly aids in the interpretation of individual attributes. This is particularly true for attributes whose interpretation is ambiguous due to a narrow range of values or their numerical nature. An example of such a model response is the case of address data, where the model correctly identified the attribute `house_number` as the only numerical attribute in the dataset. In this case, the context of the address point helped to clarify the interpretation of this attribute.

The results of the experiments clearly indicate the potential for significantly improving the effectiveness of the language model through proper and thoughtful query engineering. The use of static, predefined query templates that are independent of the specifics of the datasets limits the model's ability to generate valuable

information for the user and increases the risk of errors by the model. The effectiveness of the model can be enhanced by expanding the context provided by the user. This context may include conclusions from prior data analysis using statistical techniques or data mining. By pre-suggesting all relationships between attributes, identifying empty or constant value attributes, and presenting the model with representative values by filtering out outliers and erroneous values, the quality of the results generated by the model can be significantly improved.

The proposed experiments demonstrated the LLM model's ability to interpret source data and assign proper meanings to individual attributes. Creating the correct context is crucial for the LLM model to subsequently generate comprehensive descriptions that include the semantic interpretation of individual attributes, the relationships between them, and the semantic characterization of the entire dataset. Based on the provided information, the model can place the dataset in a broader context, explaining its origin and significance and its potential for use. It is also possible to obtain guidance on how to access the data and the appropriate techniques for analyzing specific datasets and attributes. Using the generated information, the LLM model can effectively perform decision support tasks. The experiment proved the model's ability to identify datasets most suitable for generally formulated search criteria. The proposed "decision support" scenario is just one possible way to utilize the LLM and the previously prepared information set. Implementing new functionalities based on the use of the LLM requires formulating a new set of commands for the model and appropriately interpreting the results. Consequently, this will significantly reduce the time needed to implement a new product for the end user.

ODSC Deployment

The conducted experiments have demonstrated that Large Language Models (LLMs) can effectively assist in the execution of repetitive tasks that are part of the processes carried out by data platform users. ODSC, as a concept for creating data platforms, provides guidelines that we believe should be followed when developing such solutions. In this chapter, we present specific conceptual, technological, and application solutions that we applied or developed during the implementation in the city of Poznań.

Section 6.1 presents the deployment concept and the key technologies used for the implementation of ODSC in the city of Poznań. Section 6.2 outlines the functionalities implemented in the first phase of the deployment work, focusing on data set cataloging. The software providing data analysis functionalities is presented in section 6.3. The final section 6.4, is dedicated to prototypes of tools created to automate the processes of generating valuable descriptions for data sets.

6.1 Deployment approach in Poznań

As part of the ODSC definition, we have established the principles that should guide the creation of a modern data platform for the city. However, the method of their practical implementation remains an open issue. The selection of appropriate technologies should depend on the available tools, competencies, and actual needs. All these elements must be carefully analyzed before implementation to ensure that appropriate design decisions can be made.

6.1.1 System layers

When preparing the implementation of ODSC for the city of Poznań, we paid special attention to issues related to the optimal preparation of the infrastructure supporting processing tasks. Ultimately, the proposed solution consists of four layers of abstraction: Infrastructure, Virtualization, Application, and Processes, the relationships of which are depicted in Figure 6.1.

The foundation of the implemented platform is a dedicated hardware infrastructure, consisting of servers that provide computing power in the form of processors

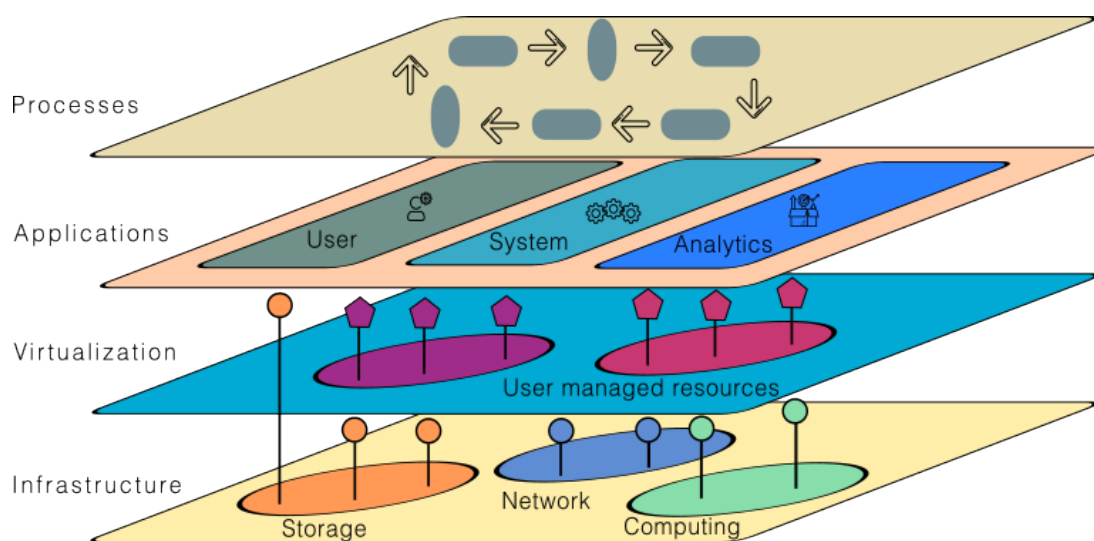


Figure 6.1: ODSC resource delivery

and volatile memory, a network infrastructure that ensures communication between computing nodes, and storage infrastructure that allows for the permanent storage of data. One of the most important features provided by the storage infrastructure is access to its resources through a well-defined and standardized high-level API. This is crucial for the implementation of future applications and data-sharing processes. A dedicated team of administrators is responsible for maintaining the infrastructure and the virtualization layer built on top of it.

The virtualization of resources is handled by Kubernetes. The mechanisms provided by Kubernetes allow users to access a dedicated pool of infrastructure resources. Users can independently launch applications according to their needs, without the involvement or knowledge of infrastructure administrators. This approach eliminates the bottleneck that infrastructure or operating system administrators typically represent in a traditional approach. Granting users full freedom to manage resources and removing the need to handle dependencies at the operating system level fosters experimentation and rapid prototyping. Full support for automation mechanisms in application deployment and software development management further reduces the time needed to launch new applications. This is crucial for implementing the Data Product concept, which often involves the creation of a service API that guarantees access to the newly developed product.

The application layer consists of all applications launched by or for the users. In our approach, we distinguish three main groups of applications:

- System applications serve globally unique functionalities and services provided for all other applications. This group includes services for user authentication and authorization, billing systems, monitoring, Open Data portal and others.
- Analytical applications are set of applications used for data processing tasks,

from basic cleaning and transformation to building analytical models. The use of these tools allows users to efficiently develop new products and analyses for city management processes. These functionalities are provided by the Kubeflow tool.

- User applications are all applications developed by and for individual users. These include services providing access to developed data products, machine learning models built on available data, or other applications utilizing previously developed data and intended for a broader audience.

The final and highest layer of abstraction is the process layer. The goal of this layer is to enable the definition of high-level processes based on services provided by the application layer. These processes allow for the execution of cyclical tasks, such as reporting, monitoring phenomena, or anomaly detection. This layer is implemented using Apache Airflow.

6.1.2 Data management

A flexible and efficiently functioning technology stack is an essential environment for achieving the key goal of proper data management and processing. The ODSC concept addresses the most significant challenges associated with this goal by systematizing and optimizing processes. The data flow diagram between the key components of the system is presented in Figure 6.2.

Available data sources can be divided into two groups:

- Legacy systems - these are all the existing IT systems in the urban space, built without any requirements for data exchange. Acquiring data from such systems and integrating them into the entire ODSC ecosystem requires the creation of agents that need to be embedded in the infrastructure of these systems. The agents' task is to deliver data to the rest of the platform in a standardized manner. The specific logic of each agent depends on the system it supports and the nature of the data. An agent can query data from a common repository in the form of snapshots, or if the data is highly volatile, it can pass it directly to analytical applications or Data Services.
- Modern systems - these are IT systems built with the necessity of providing data access to external users in mind. Such systems offer APIs through which the relevant data can be retrieved. For these systems, there is no need to implement Data Agents. Instead, we define processes that ensure data delivery to the repository or directly to analytical services.

Analytical services are the central point of data processing. They allow users to access data stored in the Common Data Repository or directly from the sources if necessary. Depending on the nature of the analysis results, they may be stored again in the Common Data Repository or shared directly through Data Services.

From the perspective of ODSC, and particularly in the context of implementing the Data Mesh concept, the components highlighted in blue represent the practical

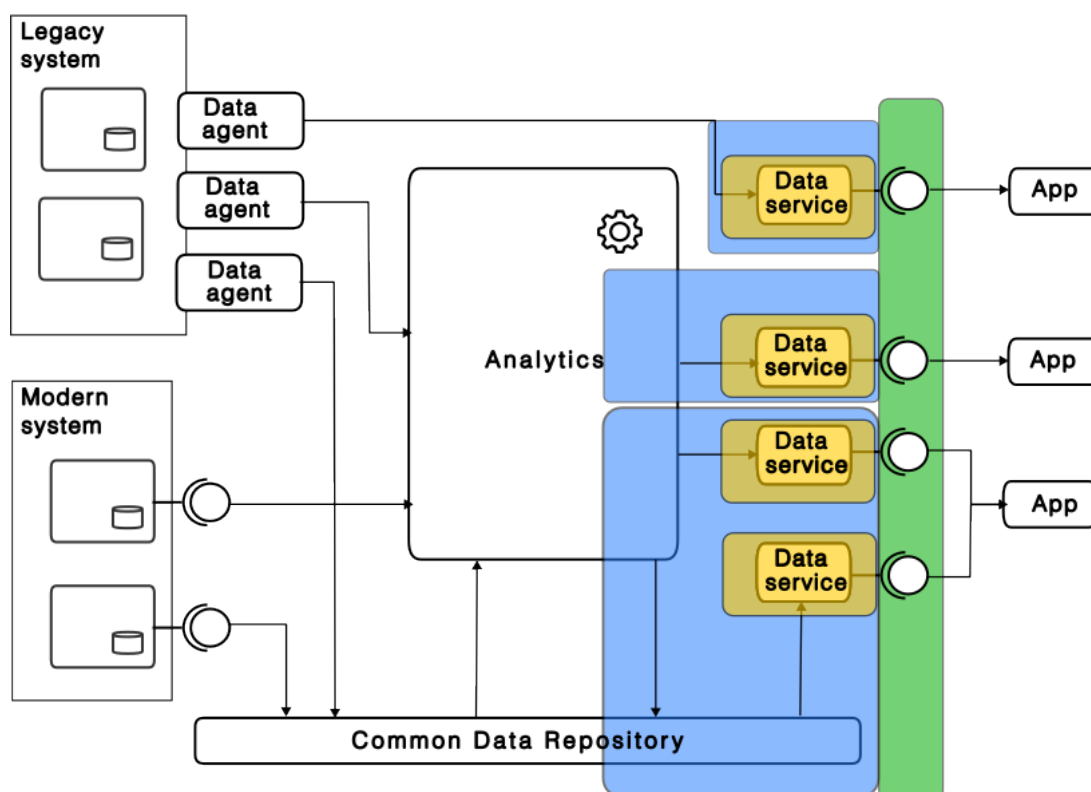


Figure 6.2: ODSC data flow

realization of the Data Domain concept. There is a clear connection between the ability to utilize computational resources at no cost to process data and ultimately make it available as a product. Data Products, highlighted in yellow, are services that draw static resources from the repository and dynamic ones directly from the analytical services. Each service has its own API, but its format is standardized to the extent necessary to ensure proper Data Governance — highlighted in green.

The final consumers of the Data Product APIs are client applications, which implement individual business models.

6.2 Poznań Open Data portal

6.2.1 Data products catalog

The Poznań Open Data Portal is one of the first ODSC components launched and made available to a broader audience. According to information available on [53] it contains 55 physical registered datasets, as well as one dataset available exclusively through an API. The role of this application is to provide information about the datasets, which are an early version of "data products." On the one hand, the Open Data Portal serves an informational function, fulfilling the obligation of the City Hall

to make municipal data available for reuse. On the other hand, it is a tool for the early promotion of the ODSC concept within the office.

Due to the primary function of the portal, which is cataloging resources, the main page focuses on search functions and a summary of current user activity on the portal. Therefore, as shown on 6.3, the primary space is occupied by the full-text search form. Below, the main categories to which datasets are assigned are listed.

Otwarte Dane

To publiczna platforma udostępniająca dane instytucjonalne miasta Poznania dla mieszkańców i wszystkich zainteresowanych ich wykorzystaniem.

Wpisz frazę i wyszukaj dane Szukaj

43 DANYCH 1 API

<p>ŚRODOWISKO</p> <p>Zieleń miejska i dane urbanistyczne</p> <p>4</p>	<p>MOBILNOŚĆ</p> <p>Komunikacja miejska, mobilność, drogi</p> <p>23</p>	<p>SPOLECZNOŚĆ</p> <p>Aktywność i dostęp do usług publicznych</p> <p>2</p>	<p>JAKOŚĆ ŻYCIA</p> <p>Aktywność i dane w zakresie turystyki</p> <p>7</p>
<p>CYFROWE MIASTO</p> <p>Dane udostępnione przez UM Poznań</p> <p>6</p>	<p>GOSPODARKA</p> <p>Liczba urodzeń, zgonów oraz inne dane demograficzne mieszkańców</p> <p>2</p>		

Figure 6.3: Poznań Open Data portal - main page search

At the bottom of the page, users can view a list of the most popular and newest datasets added by users, as presented on 6.4.

The search results view visualized on 6.5 presents users with a list of datasets that match the search criteria. As part of the dataset summary, a brief description, a short statistic of views and updates, and available visualization methods and formats are displayed. List of filters allow further narrowing of search results. In addition to the main categories, filtering can be done by data presentation method, format, or the organization providing the data. Users can also describe their resources with dy-

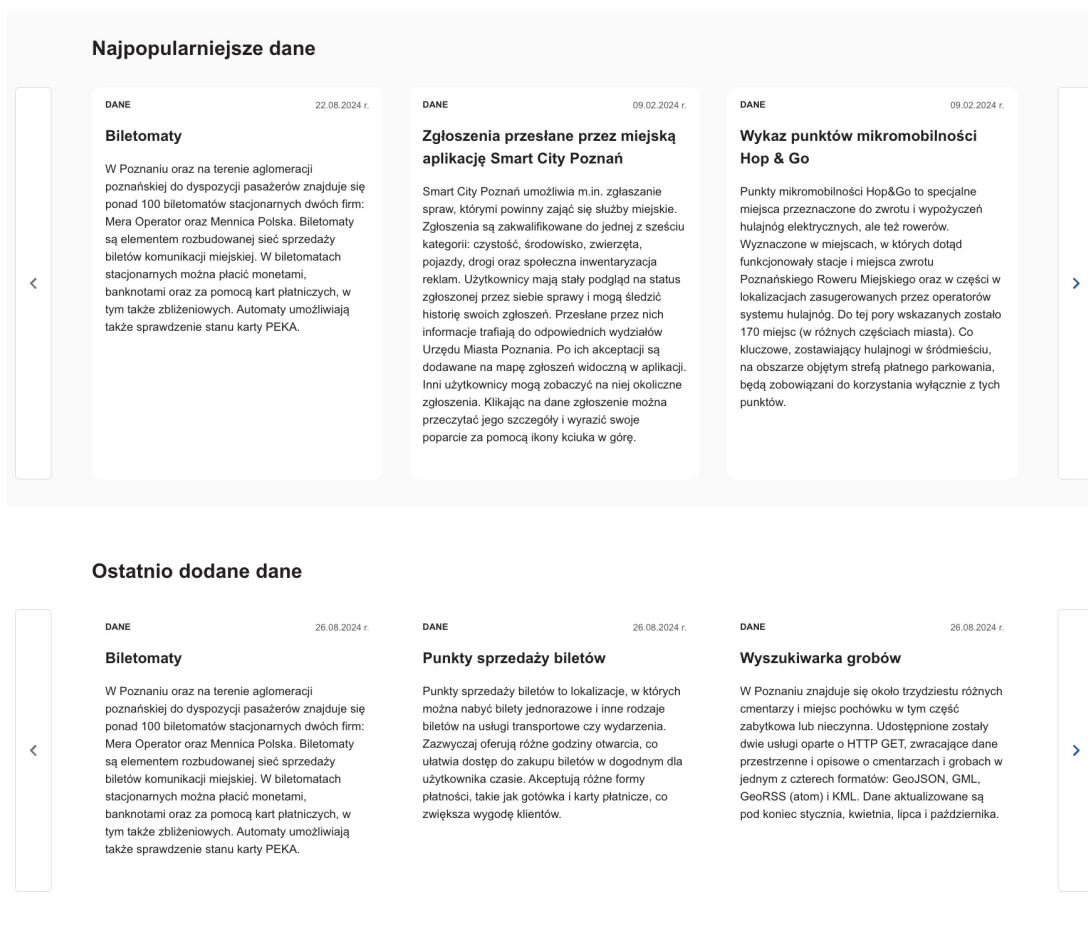


Figure 6.4: Poznań Open Data portal usage

namically defined tags, which allows for an additional dimension of result grouping.

In the dataset details view, users can access an extensive description of the dataset. Each dataset can consist of multiple resources, and therefore, a list of resources that make up the dataset is presented below the description. Various forms of visualizing the content of each resource are available. Resources with spatial attributes can be displayed on a map - 6.6.

Numeric attributes describing changes in a phenomenon over time or its proportions can be presented on a chart as shown on 6.7.

A comprehensive summary of attribute lists, their types, and meanings, along with values, is available to users in a tabular view presented on 6.8.

For visualization purposes, only a subset of available records is used. The aim of this functionality is to provide a general presentation of the dataset's basic features rather than a detailed analysis.

Below the visualization section, a summary of characteristic features such as

Kategorie

- Mobilność (23)
- Jakość życia (7)
- Cyfrowe miasto (6)
- Środowisko (4)
- Gospodarka (2)
- Społeczność (2)

Sposób prezentacji danych

- Mapa (10)
- Wykres (9)
- Tabela (9)
- Publikacja (1)

Format dostarczonych danych

- URL (17)
- JSON (14)
- CSV (8)
- XLSX (2)

Organizacje udostępniające dane

- Biuro Cyfryzacji i Cyberbezpieczeństwa (43)

Tagi

- infrastruktura (4)
- strefy (3)
- ZTM (3)
- środowisko (3)
- samochody (3)

Biletomaty

W Poznaniu oraz na terenie aglomeracji poznańskiej do dyspozycji pasażerów znajduje się ponad 100 biletomatów stacjonarnych dwóch firm: Mera Operator oraz Mennica Polska. Biletomaty są elementem rozbudowanej sieci sprzedaży biletów komunikacji miejskiej. W biletomatach stacjonarnych można płacić monetami, banknotami oraz za pomocą kart płatniczych, w tym także zbliżeniowych. Automaty umożliwiają także sprawdzenie stanu karty PEKA.

TAGI: biletomaty, bilety, infrastruktura, PEKA
KATEGORIA: Mobilność

AKTUALIZACJA: 22.08.2024 r. SPOSÓB PREZENTACJI:
WYSWIETLENIA: 1043 DOSTĘPNE FORMATY PLIKÓW: csv

Wykaz punktów mikromobilności Hop & Go

Punkty mikromobilności Hop&Go to specjalne miejsca przeznaczone do zwrotu i wypożyczeń hulajnog elektrycznych, ale też rowerów. Wyznaczone w miejscach, w których dotąd funkcjonowały stacje i miejsca zwrotu Poznańskiego Roweru Miejskiego oraz w części w lokalizacjach zasugerowanych przez operatorów systemu hulajnog. Do tej pory wskazanych zostało 170 miejsc (w różnych częściach miasta). Co kluczowe, zostawiający hulajnogi w śródmieściu, na obszarze objętym strefą płatnego parkowania, będą zobowiązani do korzystania wyłącznie z tych punktów.

TAGI: UTO, mikromobilności, hulajnogii, strefy, pojazdy elektryczne
KATEGORIA: Mobilność

AKTUALIZACJA: 26.07.2024 r. SPOSÓB PREZENTACJI:
WYSWIETLENIA: 392 DOSTĘPNE FORMATY PLIKÓW: csv

Parkingi buforowe Park&Go

Parkingi Park&Go są integralnym elementem Strefy Płatnego Parkowania (SPP). Ich zadaniem, podobnie jak SPP, jest ograniczenie liczby samochodów wjeżdżających i parkujących na ulicach w ścisłym centrum miasta. Parkingi zlokalizowane zostały w bezpośrednim sąsiedztwie Śródmiejskiej Strefy Płatnego Parkowania Centrum. Są one dedykowane osobom dojeżdżającym do śródmieścia spoza Poznania lub z peryferyjnych dzielnic miasta. Ich lokalizacja umożliwia szybki dostęp do wielu urzędów, sklepów i punktów usługowych, mających swoje siedziby na Starym Mieście. Kierowcy, którzy zdecydują się zostawić swoje auta na parkingu buforowym mogą kontynuować podróż przede wszystkim pieszo.

TAGI: infrastruktura, parking, samochody, SPP, strefy
KATEGORIA: Mobilność

AKTUALIZACJA: 26.07.2024 r. SPOSÓB PREZENTACJI:
WYSWIETLENIA: 247 DOSTĘPNE FORMATY PLIKÓW: csv

Parkingi Park&Ride

Figure 6.5: Poznań Open Data portal - search results

owner, license, category, tags, and others is provided - see 6.9. The broader context of the dataset is also presented through references to relevant external publications.

6.2.2 Conversational interface for open data

Inspired by the promising results of experiments using LLMs for decision support, we designed an additional interface to allow for more dynamic interaction with the open data portal. We expect that some users might want to select datasets not only based on specific keywords but also on the broader context in which a dataset exists. Existing conversational solutions have already familiarized users with dynamic forms of interaction with applications. Therefore, we proposed a new, dynamic interface, shown at 6.10, 6.11 and 6.12, that can adapt to the content of responses generated by the model through active components. The aim of this solution is to attempt to visualize the broader context of the data to the user while maintaining

Biletomaty

W Poznaniu oraz na terenie aglomeracji poznańskiej do dyspozycji pasażerów znajduje się ponad 100 biletomatów stacyjnych dwóch firm: Mera Operator oraz Mennica Polska. Biletomaty są elementem rozbudowanej sieci sprzedaży biletów komunikacji miejskiej. W biletomatach stacyjnych można płacić monetami, banknotami oraz za pomocą kart płatniczych, w tym także zbliżeniowych. Automaty umożliwiają także sprawdzenie stanu karty PEKA.

Dane i zasoby



 biletomaty_dataset.csv
 5 B
CSV
Pobierz plik

Prezentacja i dostęp do danych

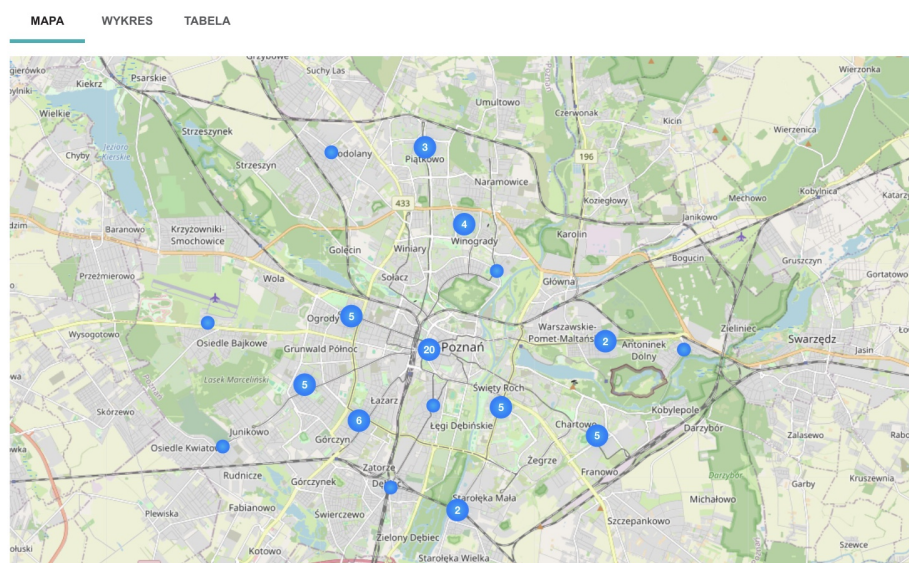


Figure 6.6: Poznań Open Data - dataset details, map

clarity in the information presented.

From a technical perspective, implementing such functionality requires ensuring the stable operation of the application as the number of datasets grows. Furthermore, the production application must be properly secured against data loss in the event of a failure. To meet these requirements, the prototype of the conversational interface was based on the use of Langchain tools—providing API access to LLM models and a local repository — Elasticsearch, serving as a vector database. The use of Elasticsearch allows for the reuse of a service that is also used for full-text indexing, minimizing the number of software components needed to achieve the required functionality. Key to the implementation is the use of the Retrieval Augmented Generation (RAG) mechanism available in the Langchain package, which allows the use of LLM models to infer from users' private data. For the prototype, we used the GPT-4 model from OpenAI.

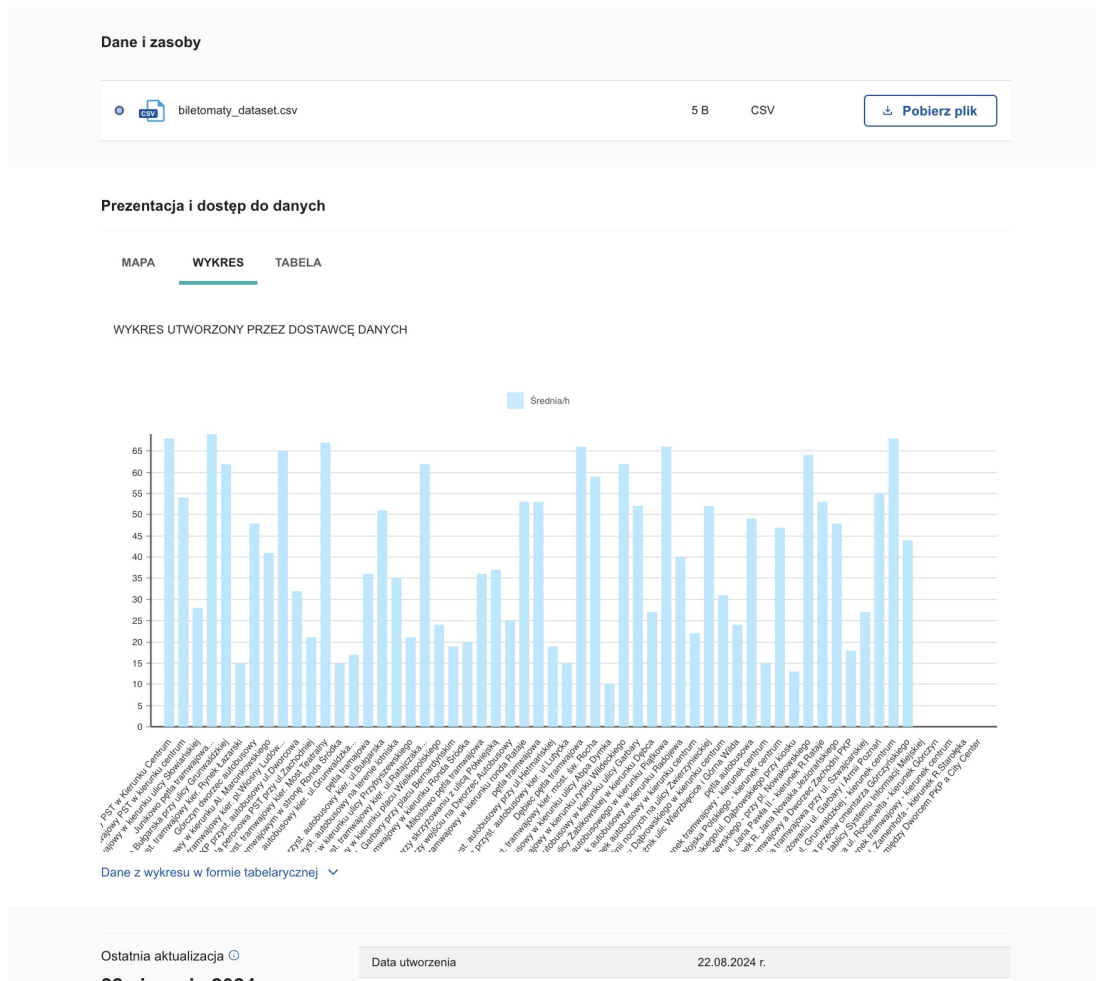


Figure 6.7: Poznań Open Data - dataset details, chart

6.3 Data analytics

In line with the concept proposed within the ODSC framework, the delivery of data analysis functionalities should be grounded in the MLOps approach. Kubeflow is one of the key tools that embodies this approach in practice. It offers a scalable, Kubernetes-based environment for the development, deployment, and monitoring of machine learning models. Kubeflow enables seamless orchestration of end-to-end workflows, encompassing data preparation, model training, and deployment, while also providing essential tools for versioning, experiment tracking, and model serving. By utilizing Kubeflow, we can achieve a more efficient, reliable, and scalable machine learning pipeline, ensuring that models remain accurate and relevant over time. This, in turn, fosters more effective and sustainable data analytics practices, ultimately driving improved business outcomes.

It is important to highlight that Kubeflow offers tools that are not only benefi-

Dane i zasoby

biletomaty_dataset.csv 5 B CSV
Pobierz plik

Prezentacja i dostęp do danych

MAPA WYKRES TABELA

Srednia/h	Nazwa	Współrzędne	Opis
68	Aleje Solidarności	52.413877311805436,16.91270705846174	przystanek tramwajowy PST w kierunku Centrum
54	Szymanowskiego	52.46044182062915,16.917067766191867	przystanek tramwajowy PST w kierunku centrum
19	Kurpińskiego	52.45064804083454,16.9195032119774	przystanek tramwajowy PST w kierunku centrum
54	Lechicka/Poznan Plaza	52.44214700052922,16.920018196108323	przystanek tramwajowy PST w kierunku centrum
19	Słowiańska	52.42720735313947,16.919288635256223	przystanek tramwajowy PST w kierunku centrum
28	Rondo Solidarność	52.43431675093359,16.937679265545988	przystanek tramwajowy w kierunku ulicy Słowiańskiej
69	Junikow	52.38335669322764,16.83485269546826	Junikowo pętla tramwajowa...
62	Stadion INEA	52.39525719919365,16.862581372263776	ulica Bułgarska przy ulicy Grunwaldzkiej

< 1 2 3 4 5 ... 8 >

KOLUMNY W TYM ZBIORZE DANYCH

Nazwa kolumny / Nazwa pola API	Opis	Rodzaj
Opis	Opiss	Liczba
Średnia/h	Opis	Liczba
Nazwa	Opis	Zwykły tekst
Współrzędne	Opis	Współrzędne

Figure 6.8: Poznań Open Data - dataset details, attributes summary

cial for building and maintaining machine learning models but also for conducting fundamental analyses. The solution's high modularity allows for the flexible expansion of its functionalities in line with user expectations, while also enhancing the automation of processes.

In the context of the deployment, an initial integration of the Kubeflow tool with the infrastructure supporting the development of the ODSC was successfully launched and executed. These tools are currently in the testing phase, and their full integration with the ODSC will require the completion of additional implementation tasks. A critical aspect that demands special attention is the provision of appropriate tools to users for facilitating interdisciplinary collaboration in the development of data domains. As part of the deployment, a service has been designed and the first version implemented, enabling the definition of analytical projects, management of project participants, and control of access rights to resources. This service is currently undergoing integration with other components of the Kubeflow tool, laying

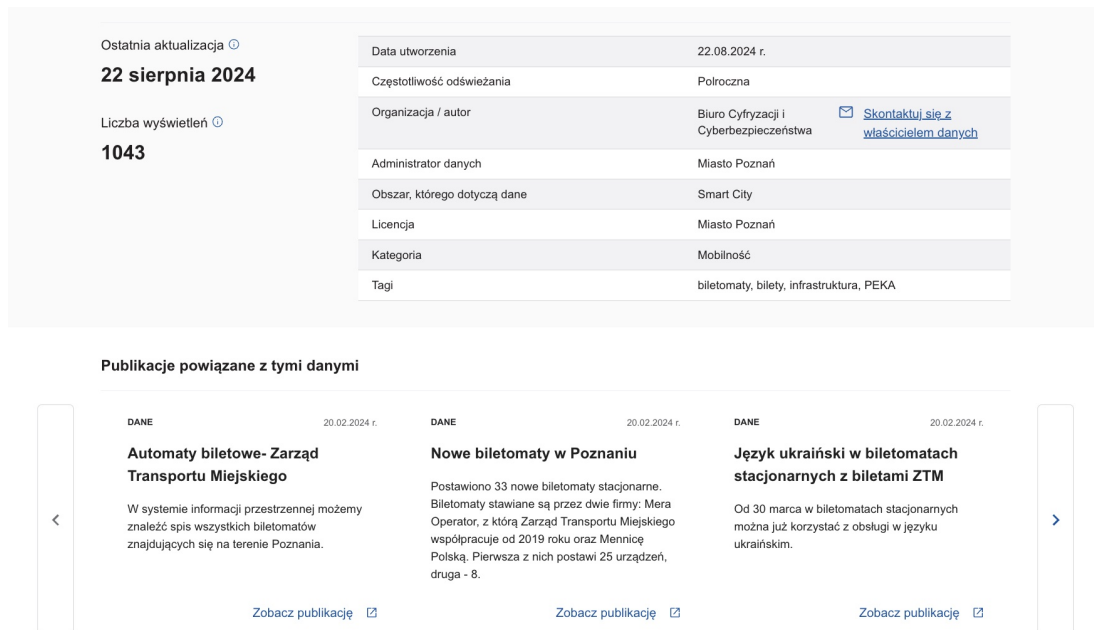


Figure 6.9: Poznań Open Data - dataset features summary

the groundwork for the continued development of future functionalities.

6.4 ODSC Concierge

The ability to use LLMs to generate valuable information about datasets, as demonstrated in the experiments presented in the previous chapter, has enabled us to create application prototypes that could be used for the automation of processes related to publishing information about datasets on the Open Data portal. The developed prototypes are intended to explore how the designed functionalities can be effectively implemented. Below, we provide a detailed overview of the tools used to create the prototypes and the functionalities that have been successfully implemented.

6.4.1 Toolset

The implementation of the ODSC Concierge functionality was carried out using the Langchain framework, the ChromaDB vector database, and language models provided by OpenAI. Below we provide a brief overview of each of these tools:

Langchain is a framework designed to streamline the development of applications that leverage large language models (LLMs). It provides a set of tools and abstractions to build complex, multi-step workflows that involve natural language understanding, data processing, and model interactions. Langchain allows developers to integrate various data sources, apply reasoning and decision-making processes, and

manage state and context across interactions, making it easier to create sophisticated, AI-driven applications.

A brief descriptions of the Python libraries, which were used for implementing ODSC concierge application prototypes.

- langchain - a basic framework tools and abstractions
- langchain_core - contains the fundamental building blocks and utilities necessary for creating and managing workflows involving language models.
- langchain_openai - integrates OpenAI's language models, making it easier to use OpenAI's models within Langchain-powered applications.
- langchain_community - includes community-contributed tools, integrations, and extensions, enabling collaborative enhancements and use of the Langchain framework.

A brief descriptions of the Java libraries, which were used for implementing Open Data Portal conversational interface.

- langchain4j - a Java implementation of the Langchain framework, designed to facilitate the development of applications that utilize large language models. It provides tools for creating complex workflows, similar to its Python counterpart but tailored for the Java ecosystem.
- langchain4j-open-ai - a Java library that integrates OpenAI's language models into the Langchain4j framework, allowing developers to leverage OpenAI's capabilities within Java-based applications.
- langchain4j-easy-rag - a Java library that simplifies the implementation of Retrieval-Augmented Generation (RAG) workflows within the Langchain4j framework. It enables developers to build applications where language models generate responses augmented with retrieved information from a knowledge base.
- langchain4j-embeddings-all-minilm-l6-v2 - a Java library that integrates the "all-MiniLM-L6-v2" embeddings model into Langchain4j, allowing for efficient creation of text embeddings. These embeddings can be used for tasks such as semantic search, clustering, and other NLP-related operations within the Langchain4j framework.
- langchain4j-elasticsearch - a Java library that provides integration between Langchain4j and Elasticsearch, enabling the use of Elasticsearch as a vector database or search engine within Langchain4j workflows. This is particularly useful for tasks involving large-scale data retrieval and management.

ChromaDB is an open-source, AI-native vector database designed for managing and querying high-dimensional vector embeddings. It is optimized for use in machine learning and AI applications, particularly those involving large language

models (LLMs) and other neural network-based systems. ChromaDB enables efficient storage, retrieval, and similarity search of vector data, making it useful for tasks like semantic search, recommendation systems, and other applications where understanding and organizing complex, unstructured data is crucial.

Elasticsearch can be used as an alternative solution to ChromaDB. Elasticsearch, particularly with its support for dense vector fields and vector similarity search, can store and retrieve vector embeddings, making it suitable for tasks such as semantic search, recommendation systems, and other applications that require similarity matching. When integrated with Langchain, Elasticsearch can be used to store and retrieve embeddings generated by language models, enabling workflows that involve retrieving contextually relevant information based on the similarity of embeddings. This allows for the implementation of Retrieval-Augmented Generation (RAG) workflows, where relevant documents or data points are retrieved from Elasticsearch and used to augment the responses generated by a language model.

OpenAI LLM models - are advanced neural network models designed to understand and generate human-like text based on large-scale data. These models, including well-known versions like GPT-3 and GPT-4, are trained on vast amounts of internet text, enabling them to perform a wide range of language tasks such as text completion, translation, summarization, question-answering, and more. OpenAI's LLMs are notable for their ability to generate coherent and contextually relevant text, making them valuable for applications in AI-driven content creation, conversational agents, coding assistance, and other natural language processing tasks.

6.4.2 Dataset reverse engineering

Dataset reverse engineering is a task that involves discovering the correct interpretation of attributes that either lack clear names or have names that are encoded. This issue arises because, in operational data models, individual attributes may be labeled with codes or abbreviations that provide little meaningful information. Although it is considered poor practice, this approach is still common, especially in older information systems with significant technological legacy. Additionally, if documentation exists, it does not always provide a detailed description of each attribute's meaning. When such data needs to be made available for further use, the lack of knowledge about how to interpret the data becomes a key problem during data retrieval. Furthermore, the need to rediscover the meaning of the data will be a repetitive task every time an attempt is made to utilize such a dataset.

To address this issue, a prototype tool has been created that generates probable attribute names based on the presented records from the dataset. If the record values do not allow unambiguous identification, the user receives a notification indicating that interpretation is not possible. This situation mainly occurs when attributes have empty values or when the distribution of their values does not provide a basis for interpretation.

6.4.3 Data characterisation

Based on the Poznań Open Data portal, it is clear that a rich description of a dataset is essential for properly presenting it to users. The goal is to provide sufficiently detailed information so that users can make informed decisions about the dataset's suitability for their needs. Too short or too technical description can result in poor visibility of the dataset in search results. This applies to both full-text index-based searches and searches supported by language models. Additionally, the process of generating descriptions is extremely laborious and repetitive, which can be a significant barrier for users looking to contribute their resources to the portal and may discourage them from using it.

To address these challenges, a prototype application has been developed that, based on a given data schema—including the dataset name, attribute names, and their types—generates a comprehensive description of the dataset. This description can serve as a general overview used to characterize the dataset. For individual attributes, brief summaries are also generated to provide appropriate interpretations. In the context of searches supported by language models, the application can produce additional content to explain the phenomena represented by the dataset, its interpretation, its role for city residents, and potential ways the dataset could be used in future analyses. These details are collected in the form of vector representations, which can then be used as additional context for a conversational interface application. The integration of the prototyped functionality will be the focus of work carried out in the subsequent stages of the implementation.

Szukaj

Wyświetlam wyniki wyszukiwania dla: Młyny w Poznaniu

Znalezione media

Dane tekstowe

Statystyki na wykresach

Tabele

Wizualizacje

Wyniki wyszukiwania dla: Młyny w Poznaniu

Młyny w liczbach

Liczba młynarzy w XIX wieku: 100

Liczba młynów w XIX wieku: 50

Maksymalna produkcja mąki rocznie: 2000 ton

Średnia liczba pracowników w dużych młynach: 20 osób

Roczna ilość zboża przetworzonego przez młyny: 10 000 ton

Średni wiek młynarzy: 45 lat

Średnia powierzchnia młynów wodnych: 150 m²

Powierzchnia działek pod młynami: 1000 m²

Historia młynów w Poznaniu

Największy młyn w Poznaniu: Młyn XIX wieczny

Spadek liczby młynów po II wojnie światowej: Po II wojnie światowej liczba młynów w Poznaniu drastycznie spadła, z około 50 w 1900 roku do zaledwie kilku działających w latach 60. XX wieku, co było wynikiem zarówno zniszczeń wojennych, jak i zmian technologicznych

Rodzaje młynów w Poznaniu: Wiatrak, wodny, parowy

Znani młynarze w Poznaniu

August Cieszkowski – Cieszkowski był właścicielem młyna, który odgrywał kluczową rolę w zaopatrzeniu Poznania w mąkę podczas II wojny światowej. Po wojnie jego młyn został znacjonalizowany, ale wcześniej był uważany za jeden z najlepiej zarządzanych młynów w okolicy.

Kazimierz Klaczo – Młynarz, który w okresie międzywojennym prowadził jeden z najstarszych młynów w Poznaniu. Jego młyn nazywany był "Młynem nad Cybiną" i był jednym z najważniejszych w regionie, dostarczając mąkę do licznych piekarni w mieście i okolicach.

Rodzina Stefańskich – Ta rodzina była znana z prowadzenia jednego z największych młynów w Poznaniu na przełomie XIX i XX wieku. Młyn Stefańskich był znaczącym dostawcą mąki w regionie, a rodzina była zaangażowana w rozwój lokalnego przemysłu młynarskiego oraz w życie społeczne miasta.

Ciekawostki o młynach w Poznaniu

Dziedzictwo kulturowe: Młyny w Poznaniu są ważnym elementem dziedzictwa kulturowego miasta, stanowiąc świadectwo jego historycznego rozwoju.

Miejsce dla turystów: Młyny w Poznaniu są popularnym celem turystycznym, przyciągającym odwiedzających zainteresowanych historią i przemysłem.

Rola w rozwoju miasta: Młyny odgrywały kluczową rolę w rozwoju Poznania, wspierając rozwój gospodarczy miasta na przestrzeni lat.

Figure 6.10: Poznań Open Data portal - conversational interface 1

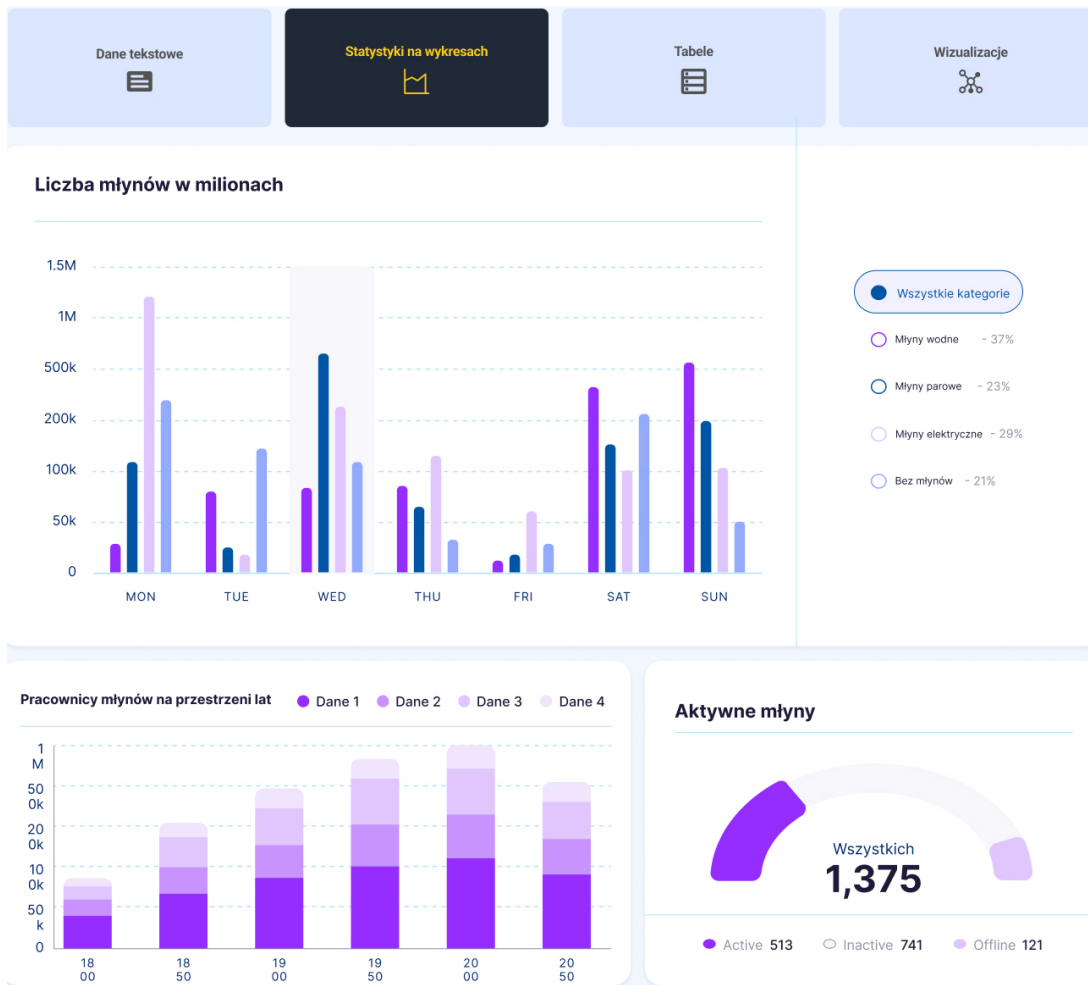


Figure 6.11: Poznań Open Data portal - conversational interface 2

Conclusion

The objective of this industrial PhD project was to develop a data platform concept for a modern city aspiring to become a Smart City. The research was driven by the practical necessity of providing the city of Poznań with tools to fulfill its obligation to open access to municipal data. The collaboration between PSNC and the city of Poznań provided an opportunity to address this practical issue more broadly, enabling significant changes in how data is perceived by the city administration and how city officials process and draw conclusions from the data at their disposal.

We began our considerations by presenting the broad context of the unique relationship between the city of Poznań and its technological partner, PSNC. By highlighting the potentials of both institutions and the objective needs in the field of municipal data processing, we hypothesized the possibility of utilizing the latest artificial intelligence technologies to improve the management process of the city's data platform. To better understand what a municipal data platform should be, we considered the essence of the Smart City concept from the perspective of data and its potential use. Aware of the existence of Poznań's official Smart City strategy, we approached the issue from a practical standpoint, focusing on the role of data in this process and how it could be utilized. This allowed us to define the main challenges associated with building a data platform that could address the city's needs and bring the practical implementation of the Smart City concept closer to reality.

We realized that the nature of the issues to be addressed in this task was technological as well as socio-cultural. Therefore, we closely examined the assumptions of the Data Mesh concept which was recently published in literature. The principles and ideas defined within Data Mesh seemed appropriate for application in the designed platform. Given that the structure, organization, and work culture of the city administration differ significantly from corporate environments, we conducted an analysis of the possibility of adapting the Data Mesh concept to the conditions of a municipal data platform. Ultimately, this led us to describe the fundamental assumptions that became the basis for defining ODSC (Open Data platform for Smart City).

By having the opportunity to observe the practical use of platform elements, we identified areas that pose the greatest barrier for users to fully harness the platform's potential. We conducted several experiments demonstrating the possibility of using LLMs (Large Language Models) to generate information about data sets, which are

crucial for the proper functioning of the platform. Based on the conclusions from the experiment, we developed a prototype service that supports users by automating the most repetitive and highly tedious tasks. At the same time, through a decision-support application, we demonstrated the potential of these technologies to quickly deliver new functionalities for the system.

The starting point for this dissertation was the conviction that it is necessary to develop a data platform concept capable of providing long-term support for the sustainable development of the city. We believe that addressing this issue solely by focusing on the implementation of specific technologies and tools is far from sufficient. Therefore, we focused on seeking solutions that could systematically optimize processes related to the sharing, management, and processing of data. In terms of organization and work culture, the ODSC concept we described appears to meet the expectations placed upon it. The practical implementation on the part of the Poznań city administration is still in its early stages, but we are already seeing a positive reception of this concept among those directly involved. The actual implementation of data analysis processes will be carried out in subsequent stages of platform development, which extend beyond the timeframe of this doctorate. Nonetheless, PSNC's experiences gained from implementing similar concepts in other projects demonstrate the high effectiveness and efficiency of the approach described in this dissertation. In terms of process automation, the conducted experiments revealed the enormous potential of properly utilizing LLMs. The models' ability to generate accurate generalizations addresses challenges related to the correct interpretation of data lacking proper documentation and domain knowledge. Additionally, the model's ability to represent a broader context allows generation of accurate and content-rich data set characteristics. Both of these aspects are crucial to ensure the proper visibility of these resources on the data platform, enabling their reuse in the future. The decision-making process in selecting the appropriate data sets for specific analytical tasks can also be successfully supported by the proper use of LLMs.

Combining theoretical reflections with practical experiences has allowed us to better understand how innovative data management solutions can be effectively integrated with the Smart City concept. In the implementation aspect, the results of the analyses conducted were applied during the realization of the data platform for the city of Poznań. In agreement with the city authorities, these works were divided into stages. The first, concerning the Open Data Portal, has been completed and made available in production. The subsequent stages are being implemented according to the adopted schedule. The tools put into use are actively being used to promote and disseminate the concept of a modern data platform among city officials and employees, in accordance with the assumptions developed in this thesis.

In summary, the primary goal of the industrial PhD project, which was to implement the developed technologies in the form of a practical solution, has been achieved. The solutions proposed in the dissertation were inspired by the real needs of the city and supported by actual usage scenarios. As a result, the conclusions from the analyses conducted, the results of experiments, and the implementation guidelines are universal and can serve as a basis for reuse in future applications.

List of Figures

3.1	Data Mesh principles and their relationships. Original image published in [22]	25
3.2	Data Mesh operational model. Original image published in [22]	27
5.1	Example of a single dataset record	58
5.2	Example of the <i>Bike Stations</i> dataset description.	62
6.1	ODSC resource delivery	70
6.2	ODSC data flow	72
6.3	Poznań Open Data portal - main page search	73
6.4	Poznań Open Data portal usage	74
6.5	Poznań Open Data portal - search results	75
6.6	Poznań Open Data - dataset details, map	76
6.7	Poznań Open Data - dataset details, chart	77
6.8	Poznań Open Data - dataset details, attributes summary	78
6.9	Poznań Open Data - dataset features summary	79
6.10	Poznań Open Data portal - conversational interface 1	83
6.11	Poznań Open Data portal - conversational interface 2	84
6.12	Poznań Open Data portal - conversational interface 3	85

List of Tables

5.1	Generation of feature descriptions using the LLM, A: number of features, B: number of empty features, C: number of recognized features, D: number of correctly recognized features, E: percent of recognized features, F: percent of non-empty recognized features, G: percent of correctly recognized features	59
5.2	Names and descriptions of the <i>Graves</i> dataset features generated by the LLM	60
5.3	Names and descriptions of the <i>MPK Stops</i> dataset features generated by the LLM	60
5.4	Names and descriptions of the <i>Bike Stations</i> dataset features generated by the LLM	61
5.5	Names and descriptions of the <i>Address Points</i> dataset features generated by the LLM	61
5.6	Datasets used in Experiment 2	63

Bibliography

1. PSNC. What's PSNC, 2024. URL <https://www.psnk.pl/about/whats-psnc/>. (cited on pages 1 and 2)
2. PSNC. PSNC Information Folder, 2021. URL https://www.psnk.pl/files/2021/11/Information_Folder_2021.pdf. (cited on page 1)
3. PSNC. PSNC Report 2023, 2023. URL https://www.pcsc.pl/files/2023/12/PCSS_Raport_2023.pdf. (cited on page 2)
4. Urząd Miasta Poznania. Raport o stanie miasta, 2023. URL <https://bip.poznan.pl/public/bip/attachments.att?co=show&instance=1001&parent=129962&lang=pl&id=454378>. (cited on page 2)
5. Urząd Miasta Poznania. Development Strategy for the City of Poznan 2020+. 2017. (cited on pages 4 and 14)
6. C. Mazurek and M. Stroinski. Common data and technological partnership – the foundation for the development of smart cities – Poznań Case Study. *Proceedings of the 52nd Hawaii International Conference on System Sciences*, pages 7390–7398, 2019. doi:<http://hdl.handle.net/10125/60176>. (cited on pages 5 and 6)
7. Ministerstwo Cyfryzacji. Otwieranie danych. Podręcznik dobrych praktyk, 2018. URL <https://dane.gov.pl/media/ckeditor/2018/11/22/otwieranie-danych-podrecznik-dobrych-praktyk.pdf>. (cited on page 9)
8. W. Łachowski. Zarządzanie danymi w miastach. Podręcznik dla samorządów, 2021. URL https://obserwatorium.miasta.pl/wp-content/uploads/2021/10/IRMiR_Zarządzaniemdanymi_20210922-4.pdf. (cited on page 9)
9. Zbigniew J. Makiela, Magdalena M. Stuss, Karolina Mucha-Kuś, Grzegorz Kinel-ski, Marcin Budziński, and Janusz Michałek. Smart City 4.0: Sustainable Urban Development in the Metropolis GZM. *Sustainability*, 14(6), 2022. ISSN 2071-1050. doi:10.3390/su14063516. URL <https://www.mdpi.com/2071-1050/14/6/3516>. (cited on page 10)
10. Yeji Yun and Minhwa Lee. Smart City 4.0 from the Perspective of Open Innovation. *Journal of Open Innovation: Technology, Market, and Complexity*, 5(4), 2019. ISSN 2199-8531. doi:10.3390/joitmc5040092. URL <https://www.mdpi.com/2199-8531/5/4/92>. (cited on page 10)

11. William K. Holstein and Hendrik W. Bode. systems engineering, 2024. URL <https://www.britannica.com/topic/systems-engineering>. (cited on page 11)
12. M. Krystek, C. Mazurek, M. Morzy, and J. Pukacki. Introducing data mesh paradigm for smart city platforms design. *Proceedings of the 56th Hawaii International Conference on System Sciences*, page 6885, 01 2023. URL <https://hdl.handle.net/10125/103468>. (cited on pages 13 and 33)
13. M. Krystek, C. Mazurek, M. Morzy, and M. Basiński. Managing Data Platforms for Smart Cities Using Large Language Models. *Proceedings of the 32th International Conference on Information Systems Development*, 08 2024. (cited on pages 13 and 53)
14. Vito Albino, Umberto Berardi, and Rosa Dangelico. Smart Cities: Definitions, Dimensions, Performance, and Initiatives. *Journal of Urban Technology*, 22:2015, 02 2015. doi:10.1080/10630732.2014.942092. (cited on page 13)
15. Thomas Mueller. Redefining the Smart City Concept: A New Smart City Definition. URL <https://www.beesmart.city/en/smart-city-blog/towards-a-new-smart-city-definition>. (cited on page 13)
16. European Commission. Smart Cities. URL https://commission.europa.eu/eu-regional-and-urban-development/topics/cities-and-urban-development/city-initiatives/smart-cities_en. (cited on page 13)
17. Urząd Miasta Poznania. Smart City Poznań. URL <https://www.poznan.pl/mim/public/smartcity/attachments.att?co=show&instance=1017&parent=81579&lang=pl&id=315643>. (cited on page 14)
18. Bhagya Nathali Silva, Murad Khan, and Kijun Han. Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities. *Sustainable cities and society*, pages 697–713, 2018. (cited on page 18)
19. Chun Sing Lai, et al. A review of technical standards for smart cities. *Clean Technologies 2.3*, pages 290–310, 2020. (cited on page 18)
20. Sarah Barns. Smart cities and urban data platforms: Designing interfaces for smart governance. *City, culture and society*, pages 5–12, 2018. (cited on page 18)
21. Urząd Miasta Poznania. Wykaz miejskich jednostek organizacyjnych, spółek prawa handlowego i spółdzielni z udziałem miasta oraz stowarzyszeń, związków międzygminnych, związków powiatowo-gminnych i fundacji, których miasto jest członkiem bądź fundatorem za 2023 r., 2023. URL <https://bip.poznan.pl/bip/struktura-organizacyjna-miasta-poznania,doc,293/wykaz-miejskich-jednostek-organizacyjnych-spolek-prawa-handlowego-i-spoldzielni-z-udzialem-miasta-oraz-stowarzyszen-zwiazkow-mie,128465.html>. (cited on page 19)

-
22. Zhamak Dehghani. *Data Mesh Delivering Data-Driven Value at Scale*. OREILLY, 2022. (cited on pages 21, 24, 25, 27, 54, and 89)
 23. W.H. Inmon. *Building the Data Warehouse*. Wiley, 2005. (cited on page 21)
 24. Ralph Kimball and Joe Caserta. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. Wiley, 2004. (cited on page 22)
 25. Sander Rossel. *Continuous Integration, Delivery, and Deployment*. Packt Publishing, 2017. (cited on page 22)
 26. Alex Gorelik. *The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science*. O'Reilly, 2019. (cited on page 22)
 27. Apache Software Foundation. Apache beam, 2024. URL <https://beam.apache.org/about/>. (cited on page 22)
 28. Randy Bean. NewVantage Partners Releases 2021 Big Data and AI Executive Survey. URL <https://www.businesswire.com/news/home/20210104005022/en/>, https://c6abb8db-514c-4f5b-b5a1-fc710f1e464e.filesusr.com/ugd/e5361a_76709448ddc6490981f0cbea42d51508.pdf. (cited on page 24)
 29. Peter Van den Besselaar and Dennis Beckers. The Life and Death of the Great Amsterdam Digital City. pages 66–96, 09 2003. ISBN 978-3-540-25331-0. doi:10.1007/11407546_4. (cited on page 33)
 30. Boyd Cohen. The 3 generations of smart cities, 2015. URL <https://www.fastcompany.com/3047795/the-3-generations-of-smart-cities>. (cited on page 33)
 31. Kinelski G. Smart city 4.0 as a set of social synergies. *Polish Journal of Management Studies*, 2022. (cited on page 33)
 32. The United for Smart Sustainable Cities. Redefining smart city platforms: Setting the stage for Minimal Interoperability Mechanisms, 2022. URL https://www.itu.int/dms_pub/itu-t/opb/tut/T-TUT-SMARTCITY-2022-05-PDF-E.pdf. (cited on page 34)
 33. Lou Celi. Smart city solutions for a riskier world. URL <https://thoughtlabgroup.com/wp-content/uploads/2021/03/Smart-City-Solutions-eBook-.pdf>. (cited on page 34)
 34. Paul Box, Ashlin Lee, Gavin Smith, Adrian Mackenzie, Todd Sanderson, Andrew Reeson, Andreas Duenser, and Robert Fleet. Data platforms for smart cities a landscape scan and recommendations for smart city practice. 03 2020. doi:10.25919/5e7504a8de92f. (cited on page 34)

35. Devin Diran, Alexander Woestenburg, Bas Kotterink, Adriaan Slob, and Roland van der Heijden. Guidance on smart city design and decision platform, 2020. URL https://ruggedised.eu/fileadmin/repository/Publications/D1.6_-_Guidance_on_Smart_City_Design_and_Decision_Platform.pdf. (cited on page 34)
36. Viktor Mayer-Schönberger and Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt, 2013. (cited on page 34)
37. Thomas Erl, Ricardo Puttini, and Zaigham Mahmood. *Cloud Computing: Concepts, Technology & Architecture*. Prentice Hall, 2013. (cited on page 34)
38. Bin Cheng, Salvatore Longo, Flavio Cirillo, Martin Bauer, and Ernő Kovacs. Building a Big Data Platform for Smart Cities: Experience and Lessons from Santander. 07 2015. doi:10.1109/BigDataCongress.2015.91. (cited on page 34)
39. Apache Software Foundation. Apache hadoop, 2024. URL <https://hadoop.apache.org/>. (cited on page 34)
40. Apache Software Foundation. Apache Spark. Unified engine for large-scale data analytics, 2024. URL <https://spark.apache.org/>. (cited on page 34)
41. Olivier Hersent, David Boswarthick, and Omar Elloumi. *The Internet of Things: Key Applications and Protocols*. Wiley, 2012. (cited on page 34)
42. Urząd Miasta Warszawy. Warszawa bardziej cyfrowa, 2021. URL <https://um.warszawa.pl/-/warszawa-bardziej-cyfrowa>. (cited on page 34)
43. DS4SSCC-DEP. European data space for smart communities, . URL <https://www.ds4sscc.eu/>. (cited on page 34)
44. DS4SSCC-DEP. European data space for smart communities - catalogue of specifications, . URL <https://inventory.ds4sscc.eu/catalogue-of-specifications>. (cited on page 34)
45. Mayor of London. Smarter London Together. URL https://www.london.gov.uk/sites/default/files/smarter_london_together_v1.66_-_published.pdf. (cited on page 34)
46. Urząd Miasta Poznania. Smart City Poznan: An Innovation Hub In The Greater Poland Region. URL <https://www.poznan.pl/mim/wortals/en/en/news,9560/smart-city-poznan-an-innovation-hub-in-the-greater-poland-region,166160.html>. (cited on page 34)
47. Jennifer Davis and Ryn Daniels. *Effective DevOps*. OREILLY, 2016. URL <https://www.oreilly.com/library/view/effective-devops/9781491926291/>. (cited on page 42)

-
48. Mark Treveil, Nicolas Omont, Clément Stenac, Kenji Lefevre, Du Phan, Joachim Zentici, Adrien Lavoillotte, Makoto Miyazaki, and Lynn Heidmann. *Introducing MLOps*. OREILLY, 2020. URL <https://www.oreilly.com/library/view/introducing-mlops/9781492083283/>. (cited on page 50)
 49. Publications Office of the European Union. What is open data? URL <https://data.europa.eu/en/dataeuropa-academy/what-open-data>. (cited on page 53)
 50. Open Knowledge Foundation. The open definition. URL <https://opendefinition.org/>. (cited on page 53)
 51. Go Fair Initiative. Fair principles. URL <https://www.go-fair.org/fair-principles/>. (cited on page 54)
 52. Josh Achiam, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. (cited on page 56)
 53. Urząd Miasta Poznań. Portal Otwarte Dane miasta Poznań, 2024. URL <https://www.poznan.pl/opendata>. (cited on page 72)