T.C.

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL

BIG DATA ANALYTICS AND MANAGEMENT HEAD OF THE DEPARTMENT

SUSTAINABLE CITY MANAGEMENT WITH BIG DATA AND DATA ANALYTICS IN METROPOLITAN CITIES, ISTANBUL EXAMPLE

MASTER'S THESIS

METIN HANEDAN

ISTANBUL 2024

T.C.

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL

BIG DATA ANALYTICS AND MANAGEMENT HEAD OF THE DEPARTMENT

SUSTAINABLE CITY MANAGEMENT WITH BIG DATA AND DATA ANALYTICS IN METROPOLITAN CITIES, ISTANBUL EXAMPLE

MASTER'S THESIS

THESIS ADVISOR

Dr. BURCU OZDEMIR

ISTANBUL 2024

T.C.

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL

MASTER THESIS APPROVAL FORM

Program Name:	BIG DATA ANALYTICS AND MANAGEMENT
Student's Name and Surname:	Metin HANEDAN
Name Of The Thesis:	SUSTAINABLE CITY MANAGEMENT WITH BIG DATA AND DATA ANALYTICS IN METROPOLITAN CITIES, ISTANBUL EXAMPLE
Thesis Defense Date:	18.01.2024

This thesis has been approved by the Graduate School which has fulfilled the necessary conditions as Master thesis.

Assoc. Prof. Dr. Yucel Batu SALMAN

Institute Director

This thesis was read by us, quality and content as a Master's thesis has been seen and accepted as sufficient.

	Title/Name	**Institution**	**Signature**
Thesis Advisor's	Dr. Burcu OZDEMIR	ITU	
Member's	Prof. Dr. Sureyya OZOGUR AKYUZ	BAU	
Member's	Prof. Dr. Arif Cagdas AYDINOGLU	GTU	

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: Metin HANEDAN

Signature:

ABSTRACT

SUSTAINABLE CITY MANAGEMENT WITH BIG DATA AND DATA ANALYTICS IN METROPOLITAN CITIES, ISTANBUL EXAMPLE

Metin HANEDAN

Master's Program in Big Data Analytics and Management

Supervisor: Dr. Burcu OZDEMIR

January 2024, 71 pages

Today, with the rapid development of information and communication technologies, data has begun to be defined as the mine of our age. As data-driven decision-making mechanisms have come to the forefront in many areas, data-driven governance in cities that produce the most data has taken its place as an important tool in sustainable city management. This study aims to investigate the approach of sustainable city management with big data and data analytics in metropolitan cities by taking Istanbul, one of the largest and most complex megacities in the world, as an example for the establishment of a big data infrastructure, as well as sample studies on data analytics in order to understand, predict and produce effective solutions to the problems arising in metropolitan cities with big data and data analytics. As a result, this study aims to offer new perspectives for sustainable city management in metropolitan areas and to provide practical recommendations to both the scientific world and local governments. The complexity and size of Istanbul makes this thesis important and reveals how big data and data analytics can transform sustainable city management. This study will be an important step in strengthening strategic thinking on sustainable urbanization and shaping the future of metropolitan cities, and will serve as a foundation for future studies on the subject.

Key Words: Big Data, Data Analytics, Sustainable City, Data-Driven Decision-Making

ÖZ

BÜYÜKŞEHİRLERDE BÜYÜK VERİ VE VERİ ANALİTİĞİ İLE SÜRDÜRÜLEBİLİR ŞEHİR YÖNETİMİ, İSTANBUL ÖRNEĞİ

Metin HANEDAN
Büyük Veri Analitiği ve Yönetimi Yüksek Lisans Programı
Tez Danışmanı: Dr. Burcu ÖZDEMİR

Ocak 2024, 71 sayfa

Günümüzde bilgi ve iletişim teknolojilerinin hızlı gelişimi ile, veri çağımızın madeni olarak tanımlanmaya başlamıştır. Veri odaklı karar alma mekanizmaları bir çok alanda ön plana çıktığı gibi, en çok veriyi üreten şehirlerde veri odaklı yönetişim konusu sürdürülebilir şehir yönetiminde önemli bir araç olarak yerini almıştır. Bu çalışma, büyükşehirlerde büyük veri ve veri analitiği ile sürdürülebilir şehir yönetimi yaklaşımını dünya üzerindeki en büyük ve karmaşık megakentlerden biri olan İstanbul'u örnek alarak büyük veri altyapısının kurulması için bir yöntem araştırmasını ortaya koymayı amaçlamanın yanında büyük veri ve veri analitiği ile metropollerde ortaya çıkan sorunları anlamak, öngörmek ve etkili çözümler üretmek adına veri analitğiyle ilgili örneklem çalışmalarını içermektedir. Sonuç olarak, bu çalışma, büyükşehirlerdeki sürdürülebilir şehir yönetimi için yeni bakış açıları sunarak, hem bilim dünyasına hem de yerel yönetimlere pratik öneriler sunmayı hedeflemektedir. İstanbul'un karmaşıklığı ve büyüklüğü, bu tez çalışmasını önemli kılmakta ve büyük veri ile veri analitiğinin sürdürülebilir şehir yönetimi açısından nasıl bir dönüşüm sağlayabileceğini ortaya koymaktadır. Bu çalışma, sürdürülebilir kentleşme konusundaki stratejik düşünceyi güçlendirmek ve büyükşehirlerin geleceğini şekillendirmek adına önemli bir adım olacak ve konu ile ilgili gelecek çalışmalara atlık oluşturacaktır.

Anahtar Kelimeler: Büyük Veri, Veri Analitiği, Sürdürülebilir Şehir, Veri Odaklı Karar Verme

To my wife Ayşin and daughter Karmen…

ACKNOWLEDGEMENTS

I would like to express my endless gratitude to my esteemed teacher Dr. Burcu Ozdemir, who has shown interest and support in the planning, research, execution and formation of this thesis study, from whose vast knowledge and experience I benefited, and who shaped my study in the light of scientific foundations with her guidance and information.

I would like to thank my beloved wife and daughter with all my heart for their understanding, support and patience.

I would like to thank my colleagues and everyone for their help.

TABLE OF CONTENTS

ETHICAL CONDUCT iii

ABSTRACT iv

ÖZ v

DEDICATION vi

ACKNOWLEDGEMENTS 7

LIST OF TABLES 10

LIST OF FIGURES 11

Chapter 1 13

Introduction 13

1.1 Statement of the Problem 14

1.2 Purpose of the Study 14

1.3 Method and Organization of the Study 14

1.4 Limits of the Study 16

1.5 Scope of the Study 17

Chapter 2 19

Big Data Analytics and Sustainable City Management 19

2.1 Big Data and Data Analytics 19

2.2 Sustainable City Management 23

2.3 Data-driven City Management 25

Chapter 3 28

Big Data and Data Analytics at IMM 28

3.1 The Design of the Methodology 28

3.2 Big Data at IMM 31

3.2.1 How was it decided to establish a big data platform at IMM? 32

3.2.2 Big data and data analytics tools and services. 38

3.3 Data Analytics at IMM 45
3.3.1 Development processes for analytical work 45
Chapter 4 47
Analytic Scenarios at IMM 47
4.1 İstanbul Card Segmentation Case Study 47
4.1.1 Problem definition and objective of the case study 47
4.1.2 Scope of the case study. 47
4.1.3 Processes of the case study 47
4.1.4 Datasets of the case study 51
4.1.5 Selected algorithms and results of the case study 51
4.2 ALO153 Call Center Sentiment Analysis Case Study 55
4.2.1 Problem definition and objective of the case study 55
4.2.2 Scope of the case study 55
4.2.3 Processes of the case study 55
4.2.4 Datasets of the case study 57
4.2.5 Selected algorithms and results of the case study 58
4.3 Traffic Intensity Forecasting Case Study 61
4.3.1 Problem definition and objective of the case study 61
4.3.2 Scope of the case study 61
4.3.3 Processes of the case study 62
4.3.4 Datasets of the case study 64
4.3.5 Selected algorithms and results of the case study 65
Chapter 5 68
Conclusions 68
REFERENCES 72

LIST OF TABLES

TABLES

Table 1 Istanbul Card Segmentation, Free Card, Frequent users 48
Table 2 Istanbul Card Segmentation, Free Card, Frequent users, Travel Percentages by time zone 48
Table 3 Istanbul Card Segmentation, Discounted Card, Subway preference users ... 49
Table 4 Istanbul Card Segmentation, Discounted Card, Subway preference users, Travel Percentages by time zone 49
Table 5 Istanbul Card Segmentation, Full Card, Frequent users 50
Table 6 Istanbul Card Segmentation, Full Card, Frequent users, Travel Percentages by time zone 50
Table 7 RFM Segments and Sample Actions ……....56
Table 8 Metrics of Equally distributed data 59
Table 9 Metrics of Unsampled under-represented class data..... 60
Table 10 Metrics of Equally distribution kept data..... 60

LIST OF FIGURES

FIGURES

Figure 1. Flowchart of the Study 16
Figure 2. Four Data Analytics Types 23
Figure 3. Data Analytic Steps 30
Figure 4. Big data reference architecture 32
Figure 5. Data collection to decision support cycle 33
Figure 6. What can be observed with big data? 36
Figure 7. What can be done with big data? 37
Figure 8. Big Data Infrastructure 40
Figure 9. Citizen 360 View with big data 42
Figure 10. Outputs of big data platform 43
Figure 11. Outputs of big data platform – Traffic Density 43
Figure 12. IMM Wifi – Real Time Dashboard 44
Figure 13. IMM Open Data Portal 44
Figure 14. Free Cards RFM Segments 53
Figure 15. Discounted Cards RFM Segments 53
Figure 16. Full Cards RFM Segments 54
Figure 17. Most Common Unigrams 56
Figure 18. Most Common Bigrams 57
Figure 19. Most Common Trigrams 57
Figure 20. SadedeGel Library Infrastructure 59
Figure 21. Processes of the case study 62
Figure 22. Data preparation and exploration 63
Figure 23. Traffic Density Forecasting – Speed and Index Prediction 66
Figure 24. District and route based traffic density forecast 67

LIST OF ABBREVIATIONS

IMM	Istanbul Metropolitan Municipality
IETT	Istanbul Electric Tramway and Tunnel
ISKI	Istanbul Water and Sewerage Administration
NLP	Natural Language Processing
EDA	Exploratary Data Analytics
CRM	Customer Relation Management
KVKK	Personal Data Protection Law
IOT	Internet of Things
POC	Proof of Concept
SPSS	Statistical Package for the Social Sciences
DSX	Data Science Experience (IBM)
OBIEE	Oracle Business Intelligence Enterprise Edition
CDH	Cloudera Hadoop Distribution
SQL	Structured Query Language
HA	High Availability
CKAN	Comprehensive Knowledge Archive Network
ML	Machine Learning
ETL	Extract Transform Load
PCA	Principal Component Analysis
SVM	Support Vector Machine
KNN	K-Nearest Neighbor
ARIMAX	Auto Regressive Integrated Moving Average Exogenous
ESM	Exponential Smoothing Model
IDM	Intermittent Demand Model
UCM	Unobserved Components Model
MAPE	Mean Absolute Percentage Error
RFM	Recency Frequency Monetary
MERNIS	Central Population Administration System
URL	Uniform Resource Locator
HTML	Hypertext Markup Language
TF-IDF	Term Frequency-Inverse Document Frequency

Chapter 1

Introduction

Today, the management dynamics of cities are changing rapidly and data, as the key to this change, has gained an important position at the center of our information age. The rapid evolution in information and communication technologies has brought data-based decision-making mechanisms to the forefront in many areas, and in cities that produce the most data, data-based management has become the main tool of sustainable city management. In this context, this study aims to investigate the approach to sustainable city management with big data and data analytics in metropolitan cities. Rapid developments in information and communication technologies have led us to define data as the mine of our age. This evolution has led to the prominence of big data and data analytics in shaping city management processes. Cities have the potential to create fast, effective and sustainable solutions using these technologies. Istanbul, one of the largest and most complex megacities in the world, is the focus of this study. Istanbul stands out not only for its physical size, but also for its cultural diversity, historical richness and management challenges. This megacity serves as a laboratory for the establishment of big data infrastructure and the study of data analytics applications. The main purpose of this study is to understand how big data and data analytics can be used for sustainable city management in big cities. Analyzing the case of Istanbul will provide new perspectives and shed light on feasible solutions for sustainable city management. This study aims to provide concrete recommendations based on real-world applications, rather than just a scientific research. This study focuses on a literature review on big data, data analytics techniques, urban management, and the necessity and trends of big data-oriented studies. Moving on to the analysis of Istanbul, the current situation analysis, needs analysis, and the establishment of big data infrastructure are discussed in detail. Despite the wide range of big data analytics techniques and their potential to provide effective solutions in areas such as population growth, limited resource use, climate change and urbanization in the city, this study faces certain limitations. These limitations stem from the fact that the study focuses on specific areas of practice and relies on a limited sample to ensure general validity. This study consists of four main

sections. In the introduction, the importance of big data and data analytics in sustainable city management is emphasized and the methodology and application content is summarized, focusing on big data and data analytics platforms to be established in Istanbul.

1.1 Statement of the Problem

Today, the rapidly growing population of metropolitan cities brings the need for sustainable city management to the forefront. Advances in information and technology offer the potential to produce faster, more effective and sustainable solutions for city management. In this context, the main motivation of this study is to emphasize the critical importance of data and technological advances brought by the information and technology age in terms of providing insights and foresight to city management by using big data and data analytics in a large metropolis like Istanbul.

1.2 Purpose of the Study

While this study aims to identify methods that can be used to establish big data infrastructure in metropolitan municipalities, it also aims to contribute to sustainable and integrated city management through the application of data analytics techniques and the results obtained. In this context, it aims to shed light on how big data and data analytics processes can be applied in the perspective of sustainable city management by bringing a new perspective to the literature and to contribute to the applicability of these techniques in other metropolitan municipalities.

1.3 Method and Organization of the Study

Within the scope of this study, literature research has focused on the concepts of big data, data analytics techniques, types and sustainable city management, and the need for and trends in big data-oriented studies in city management have been examined in detail.

In the application study, the city of Istanbul, one of the largest and most complex megacities in the world, was taken as the center of the research. In the first stage, a current situation analysis was conducted in Istanbul Metropolitan Municipality and existing data sources, related departments and hardware were identified. After the

needs analysis, face-to-face interviews were conducted with more than 20 companies specialized in big data platform and the appropriate technical architecture was decided.

Finally, 3 different applications were included to compare the before and after of the big data system. In these applications, segment analysis, sentiment analysis and forecasting applications were selected and how big data and data analytics can contribute to sustainable city management was evaluated in detail.

As shown in the flowchart of the study (Figure 1):

The literature review on improving sustainable urban governance through big data analytics first focused on elucidating the basic concepts of big data, its various types and analytical techniques, including its applications in urban governance. The necessity of big data-centric studies in urban governance is explored and current trends are examined, highlighting the integral role of data-driven decision-making in the development of sustainable urban environments.

The application study started with the selection of Istanbul Metropolitan Municipality as the study area, as one of the most comprehensive and multifaceted megacities in the world. The current situation analysis identified existing data sources, relevant municipal units and infrastructure. A needs analysis then identified the key elements needed to integrate a comprehensive big data system into the city's administrative framework.

Stakeholder consultation was rigorously conducted through structured interviews with more than twenty organizations specialized in big data platforms, yielding valuable insights. Based on the culmination of the needs assessment and stakeholder feedback, a technical architecture was meticulously designed to fit the specific requirements of the municipality.

This was followed by implementation by deploying the selected technical architecture into the operational environment of Istanbul Metropolitan Municipality. Three different applications were developed to evaluate the effectiveness of this implementation: Segment Analysis, Sentiment Analysis and Prediction Applications.

In the evaluation phase, operational effectiveness was compared before and after the big data system implementation, utilizing the three applications to identify performance improvements. The study then assesses the contribution of big data and analytics to sustainable city management, revealing tangible benefits and strategic insights.

The study concludes with a comprehensive conclusion that summarizes key findings, articulates the significant impact of big data on advancing sustainable practices within the complex urban fabric of Istanbul, and sets a precedent for future urban management paradigms.

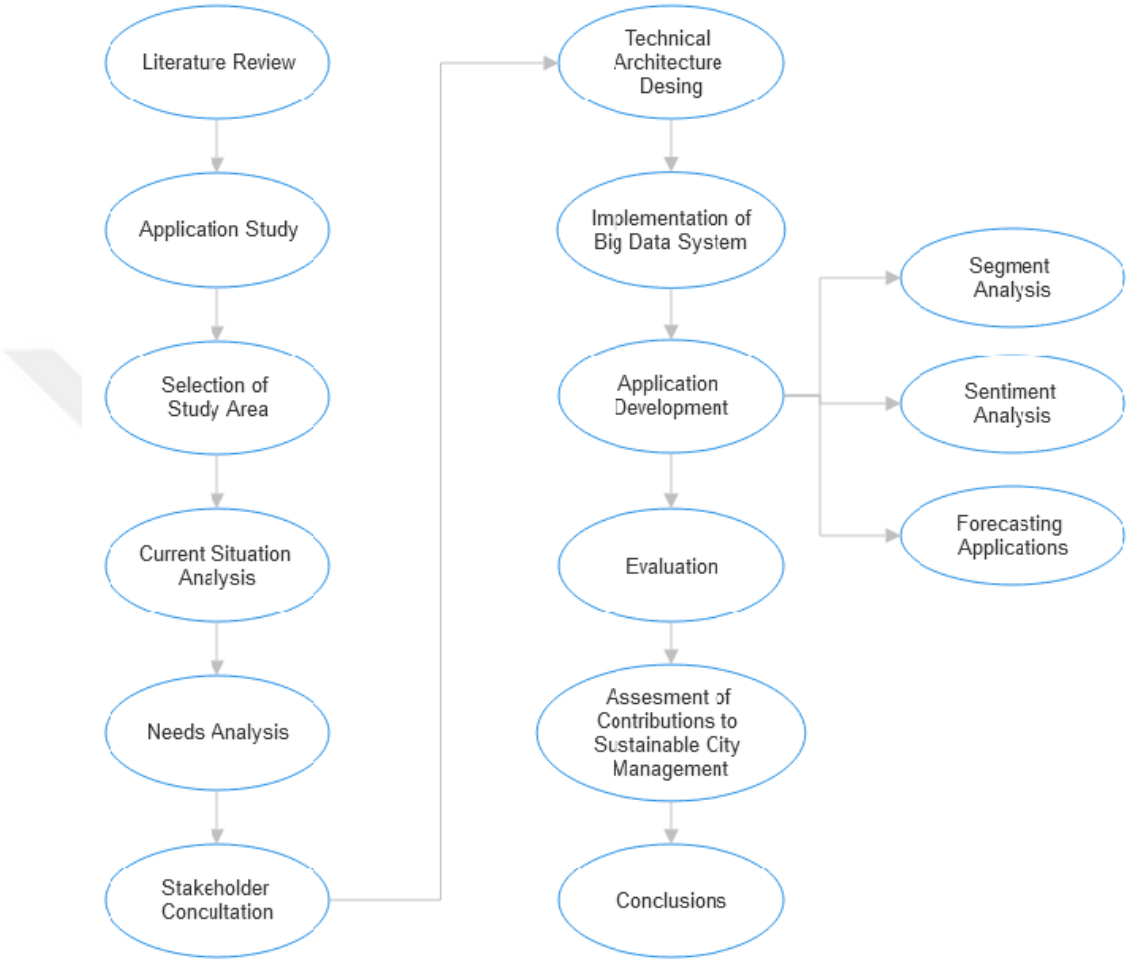


Figure 1. Flowchart of the Study

1.4 Limits of the Study

This study faces a number of limitations, given the wide range of big data analytics techniques and their potential to provide effective and efficient solutions in areas such as urban population growth, limited resource use, climate change and urbanization processes.

First, this study focuses on 3 specific application areas and thus does not claim to provide an overall scope. Therefore, it could not cover other potential application areas in order to reach a broad urban management perspective.

Secondly, the sampling to ensure the general validity of the effectiveness of the algorithms and techniques used is limited given the different city conditions, cultural differences and city structures. This imposes certain limitations in drawing general conclusions without taking into account the unique conditions in various cities.

Finally, this study was conducted in an environment of rapid technological and algorithmic evolution. This risks making the results of the study obtained in a given time period outdated over time.

1.5 Scope of the Study

The study consists of four sections. In the introduction, the importance of data and technology in sustainable city management is mentioned and the content of the method to be followed and the applications that can be developed while aiming to establish big data and data analytics platforms that can be established in metropolitan cities for this purpose are mentioned.

In the second section, the literature on the use of big data and data analytics in sustainable city management is reviewed. The concepts of big data and data analytics are explained in detail, the types of data analytics are mentioned and the process from data to information and from information to prediction is detailed. In addition, the importance and possible areas of big data use in the principles of environment, economy, social, governance and participation, which are essential for sustainable city management, are evaluated.

In the third chapter, the concept of big city and the organizational structure of the Metropolitan Municipality that manages the mega city of Istanbul are discussed. The stages of deciding to establish a big data platform in IBB and the needs analysis are mentioned, information about the reference architecture of the big data platform that can be established is given, and after these planning and analysis studies, the problems that the big data platform can solve in IBB and what its outputs will be are explained in detail. Prior to the establishment of the platform, current big data and data analytics systems were examined in detail and appropriate architectural proposals were investigated so that all promising technological services and products could be included in the project. At the end of this process, the architecture and details of the

big data and data analytics system are mentioned. The outputs of the platform and the gains obtained are detailed. In addition, the problem definition of 3 different analytical applications, the development processes and the work carried out until the final stage are mentioned in detail and the outputs are presented.

In the fourth and final section, IBB's work with the limited data available before having a big data and data analytics platform, the gains achieved with the establishment of the platform, and examples of possible analytical projects that can be developed on the platform are mentioned and the road map of Istanbul's big data and data analytics journey is tried to be determined.

The introduction sets the stage by outlining the objectives of the study and highlighting the significance of big data and analytics in modern urban governance. The thesis at hand examines the transformative influence of big data and data analytics within the Istanbul Metropolitan Municipality (IMM), providing valuable insights into the municipality's progression towards intelligent and sustainable urban governance.

Chapter 2

Big Data Analytics and Sustainable City Management

Today, sustainable urban management in rapidly growing metropolitan cities is one of the most important challenges of our age. Increasing population, limited use of resources, climate change and urbanization processes lead city governments to produce more effective, efficient and sustainable solutions. In this context, big data and data analytics provide an important tool for city governments. Big data enables city governments to make more effective and faster decisions based on current and historical data of urban systems. In this way, city governments can use resources more efficiently and achieve sustainable development goals more effectively (Li et al., 2017). Big data in city management enables better planning in areas such as traffic management, energy consumption, water resources management. This contributes to making cities more sustainable and livable (Hassanzadeh and Wang, 2016). Big data analytics is an important tool that can be used in urban planning to predict future needs and manage existing infrastructure more effectively. In this way, cities become more sustainable and adaptable (Batty et al., 2012). Big data analytics can accelerate emergency response by assessing the flow of real-time data in urban planning, enabling safer management of cities (Zhang et al., 2014). Big data analytics plays a critical role in improving urban services. City governments can use analytics to regulate traffic flow, optimize garbage collection, and monitor energy consumption to develop more sustainable energy policies (Batty et al., 2012). Big data analytics can improve citizen satisfaction and increase the competitive advantage of cities by enabling more effective delivery of city services.

In this section of the study, big data and data analytics techniques will be discussed and the relationship between the concept of sustainable city management and data analytics will be explained.

2.1 Big Data and Data Analytics

Big data refers to large-scale data sets that cannot be processed by traditional data processing techniques due to their volume, variety and velocity. These data sets are often generated at high speed, come from different sources, and can exist in various

formats (Chen et al., 2014). Big data refers to large data sets that are high-volume, fast-changing and contain various types of data. These data sets pose a challenge to traditional data processing tools in terms of processing, analyzing and extracting meaningful information (Manyika et al., 2011). Big data is usually characterized by three V's: volume (large amount of data), variety (different types of data) and velocity (speed of data creation and processing). These characteristics indicate the complexity and difficulty of processing big data (Davenport and Dyche, 2013). From the perspective of an ecosystem where technology is rapidly developing and data is growing rapidly, the concept of big data is the general nomenclature given to the model that represents the categorization of raw data, structured, semi-structured and unstructured data from different sources, the transformation and storage of structured, semi-structured and unstructured data into meaningful and processable information, and the interpretation and transformation of this information into systems that will take action or make decisions (Katerina et al., 2020). In order for data to be defined as big data, it is expected to have some characteristics. These characteristics are as follows;

- Volume; the capacity of data is increasing day by day and organizations need to design in detail how they will deal with this growing data volume, processing, archiving, storing, deleting, integrating.
- Variety refers to the diversity of data.
- Velocity refers to the high and increasing speed at which data is generated.
- Value is the step where the data creates and generates value. It is extremely important to process the data in a way that creates added value for the organization.
- Verification; It is the step of making sure that the data is secure. It is the step of monitoring and accessibility at the security level required from the right layer during data flow (URL 1).

Data analytics is the process of using statistical and mathematical methods to analyze large amounts of data, recognize patterns, extract meaningful information and predict future events. This process is used to strengthen decision-making processes and provide strategic guidance (Delen and Demirkan, 2013). Data analytics is used to discover regularities, trends and relationships in data sets. These discoveries are used to optimize business processes, gain competitive advantage and make better decisions

(Witten and Frank, 2005). By supporting decision-making processes, data analytics helps organizations use their data more effectively and achieve strategic goals. Analytical results enable informed decisions (Davenport et al., 2013).

Data analytics is a discipline that combines a wide range of techniques that can be used for different data types and analysis objectives. Data analytics is a discipline that consists of collecting, storing, categorizing, analyzing and finally producing descriptive, explanatory reports and predictive, prescriptive conclusions from data. The main purpose of data analytics is to find trends and cause and effect relationships, solve problems, and make predictions and forecasts. In this process, statistical analysis methods and techniques, machine learning methods and, if necessary, optimization and simulation techniques are applied to the data (Stefani et al., 2018). The most commonly used techniques in the literature;

- Statistical Analysis: Data analytics describes, measures and models data sets using statistical analysis tools. This technique is used to identify patterns and relationships in data sets (Hair et al., 2019).
- Machine Learning: Machine learning is the process of learning and making predictions from data sets through the use of algorithms, models and artificial intelligence techniques. This technique is used to predict future events, recognize patterns and perform classification (Bishop, 2006).
- Data Mining: Data mining is the use of statistical and mathematical techniques to discover hidden information and patterns in large data sets. This technique is used in data analytics processes to explore data sets in depth (Han et al., 2011).
- Natural Language Processing (NLP): Natural Language Processing is the process of understanding, interpreting and inferring text-based data. This technique is used to analyze and understand text data such as social media, customer feedback (Manning et al., 2008).
- A/B Testing: A/B testing is a technique used in comparative analysis between two or more groups. This test is used to identify differences between variables and evaluate which strategy is more effective (Kohavi et al., 2009).
- Regression Analysis: Regression analysis is a statistical technique used to determine the relationship between dependent and independent variables.

This technique is used to understand the correlation and interactions between variables (Montgomery et al., 2012).

- Temporal Analysis: Temporal analysis is used to identify trends, seasonal effects and cycles by examining data sets that change over time. This technique is important for understanding changes over time (Chatfield et al., 2004).

When analyzing the types of data analytics, 6 different types of data analytics come to the fore. Each type covers different analytical purposes and application areas, and is used depending on the needs and goals of an organization. These are;

1. Descriptive Data Analytics:

Descriptive data analytics aims to identify key characteristics, patterns and trends by summarizing existing data sets. This type of analytics is used in data discovery and understanding processes (Tukey, 1977).

2. Predictive Data Analytics:

Predictive data analytics uses statistical and mathematical models to predict future events from historical data sets. This type of analytics focuses on predicting future trends and possible scenarios (Han et al., 2011).

3. Diagnostic Data Analytics:

Diagnostic data analytics aims to understand cause-and-effect relationships and the origin of events in existing data sets. This type of analytics is particularly used to understand the causes of a particular event and to evaluate past performance (Provost and Fawcett, 2013).

4. Prescriptive Data Analytics:

Prescriptive data analytics recommends actions to achieve specific goals. This type of analytics assesses the current situation, evaluates alternative scenarios and identifies the most effective courses of action (Davenport and Harris, 2010).

5. Exploratory Data Analytics:

Exploratory data analytics is a method used to discover unknown patterns and relationships. This type of analytics usually aims at in-depth exploration and analysis of large and complex data sets (Tukey, 1977).

6. Interactive Data Analytics:

Interactive data analytics is an approach that allows the user to interact directly with data sets. This type of analytics allows users to visually look at and analyze data in depth (Buja, A., Cook, D., & Swayne, D. F 1996).

The most commonly used types of data analytics are descriptive, diagnostic, predictive and prescriptive analytics. Figure 2, summarizes which type of data analytics answers which types of questions.

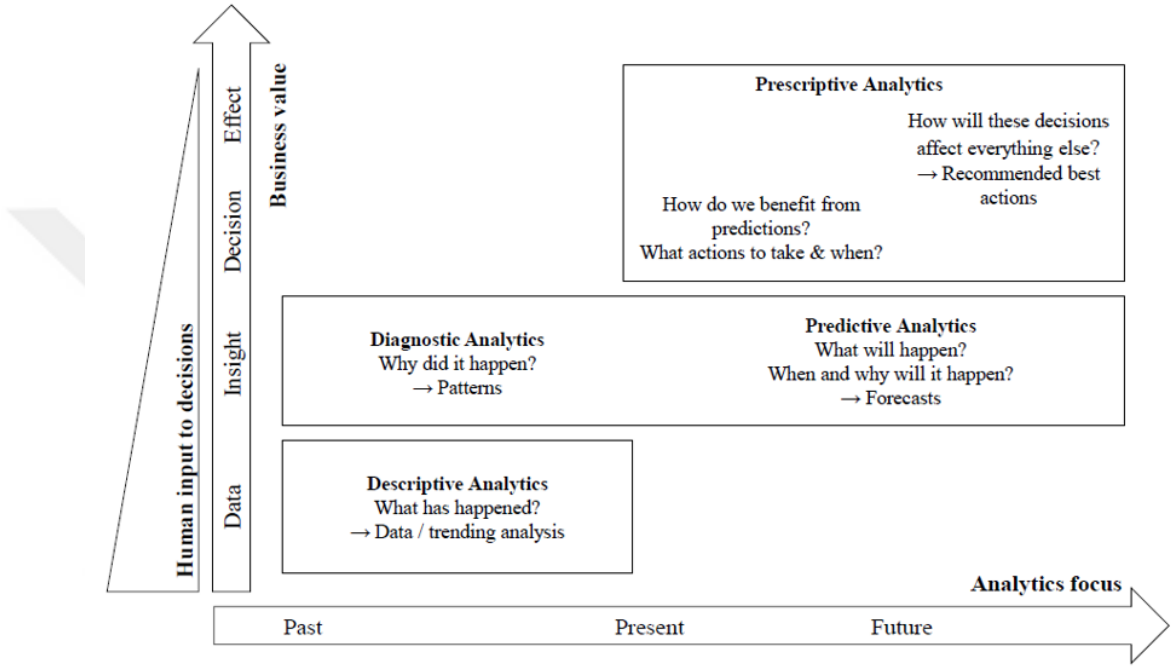


Figure 2. Four Data Analytics Types (Lily Koops, 2020)

2.2 Sustainable City Management

Sustainable city management refers to a comprehensive approach that aims to achieve cities' long-term sustainability goals in environmental, economic and social dimensions. This approach includes not only economic growth but also factors such as protecting environmental resources, ensuring social justice and maintaining the balance of local ecosystems. In short, the concept of sustainable city management refers to a management approach that focuses on achieving the long-term sustainability goals of cities by balancing environmental, economic and social factors.

Bulkeley et al. (2000), through the environmental dimension, states that the concept of sustainable city management adopts the principle of environmental sustainability and aims to use natural resources in a sustainable manner and to

minimize the impacts of cities on ecosystems (Bulkeley, 2000). Sustainable city management through the Economic Dimension aims to strengthen the local economy, reduce income inequality and increase the economic resilience of cities by adopting the principles of economic sustainability (Betsill and Bulkeley, 2006). In terms of the Social Dimension, sustainable city management aims to meet the social needs of all individuals living in cities by prioritizing social justice and participation. It supports equal and fair access to basic services such as education, health and housing (Lejano and Stokols, 2013). In the governance and participation dimension, sustainable city management emphasizes effective governance and participation and is defined as involving city residents in decision-making processes and adopting a transparent and accountable management approach (Robinson, 2006). In the Infrastructure Planning dimension, sustainable city management aims to reduce the environmental impact of cities by focusing on renewable energy sources, energy efficiency and low-carbon transportation systems in infrastructure planning (Girardet, 1999). Finally, in terms of Green Spaces and Biodiversity, sustainable city management aims to manage the natural environment of cities in a sustainable way by focusing on factors such as protecting and increasing green spaces and supporting biodiversity (Beatley, 2000).

In general terms, sustainable city management refers to a holistic approach that aims to plan, manage and develop cities based on the principles of environmental, economic and social sustainability. This concept aims to address the challenges of cities such as rapid growth, resource use, environmental impacts and social inequality. Sustainable city management focuses on factors such as strengthening the local economy, improving social services and engaging local residents, while promoting the efficient use of environmental resources. It also aims to ensure that cities leave a livable environment for future generations by adopting long-term sustainability goals. Sustainable city management can therefore take advantage of the opportunities provided by the big data platform to manage all these dimensions in an integrated manner. For example;

- Environmental Sustainability: Big data analytics processes large-scale environmental data from various sensors and measurements, enabling a better understanding of environmental factors such as air quality, water management, waste management. This data supports decision-making processes towards environmental sustainability goals.

- Economic Sustainability: Big data can analyze city economic data to assess business trends, economic performance and employment status. This data can be used to increase local job opportunities, support sustainable economic growth and reduce income inequality.
- Social Sustainability: Big data analytics helps us understand social dynamics in the city by examining data on social mobility, population density, education and health. This information can be used to plan education and health services according to needs, strengthen social services and reduce inequalities.
- Governance and Participation: Big data can be used to support decision-making and increase transparency in city governance. In addition, data from social media and other digital platforms can be used to more effectively engage the public in city governance.
- Infrastructure Planning and Energy Efficiency: Big data can contribute to more effective planning of city infrastructure and the development of sustainable energy policies by analyzing data such as traffic flow, energy consumption and infrastructure usage.

With its in-depth analysis and predictive capabilities in these areas, big data can provide a more information-driven, efficient and sustainable approach to city management. However, it is also important to establish appropriate security measures and data privacy policies in order to use this data effectively.

2.3 Data-driven City Management

Data-driven city management is a management model that aims to make more effective and sustainable decisions using large amounts of data collected in various service areas in cities (Kitchin, 2014). Data-driven city management continuously monitors and analyzes various factors in the city through sensors, smart devices and other data sources. This process includes data collection and analysis in areas such as traffic management, energy use, environmental impacts and safety (Caragliu et al., 2011). Data-driven city management aims to use collected data to make quick decisions and manage city services more effectively. This can lead to improvements in areas such as emergency response, traffic management, environmental sustainability

and public safety (Zanella et al., 2014). Data-driven city management encourages feedback and participation from city residents. This aims to adopt a more transparent management approach and respond more responsively to the needs of the public (Townsend, 2013). By adopting sustainability principles, data-driven city management aims to increase energy efficiency, reduce environmental impacts and improve the quality of city life. It also aims to make life in the city more comfortable and sustainable by encouraging technological innovations (Giffinger et al., 2007).

Many cities around the world are putting forward various projects to make their cities more sustainable, effective and livable through data-driven city management practices.

Seoul is one of the pioneering cities adopting a data-driven approach to city management. It has integrated various technologies to monitor traffic flow, optimize energy consumption, and manage environmental sustainability projects using big data analytics. It also uses mobile applications and online platforms to increase public participation and involvement in city management.

By adopting smart city practices, Singapore has moved to a sustainable and data-driven governance model. Using sensor networks, big data analytics and IoT (internet of things) technologies, it has implemented various projects to improve traffic management, energy efficiency and disaster management.

New York City is improving service quality by implementing data-driven solutions in various areas of the city. For example, it uses big data analytics to identify crime precursors and develop programs to prevent them, improving city safety. It also uses data analytics in areas such as traffic management, garbage collection and emergency response.

Barcelona is another notable example of a smart city. Using big data and sensor technologies, the city has focused on improving traffic management, energy efficiency, environmental sustainability and public health. It also increases participation in city governance through public online platforms.

Data-driven city management practices demonstrate the significant advantages that big data and data analytics bring to governance processes in cities.

Traffic Management: Big data and data analytics are used to understand and manage traffic flow. In this way, cities can reduce traffic congestion, optimize transportation systems and facilitate the daily life of the public. For example, cities

such as Seoul and Singapore have created intelligent transportation systems by analyzing traffic data and have created more effective solutions to traffic problems.

Energy Efficiency and Sustainability: Big data can be used to monitor, manage and optimize energy consumption. By analyzing energy use, cities can support sustainable energy projects. For example, cities such as Singapore and Barcelona have adopted big data and smart energy management practices to improve energy efficiency.

Public Health and Safety: Big data analytics can make significant contributions to public health and safety. Cities can use big data to track disease outbreaks, plan emergency responses and identify crime precursors. For example, New York City uses crime analytics to monitor the security situation in the city and take preventive measures.

Public Engagement and Transparency: Big data enables city residents to participate more effectively in governance processes. In addition, public data platforms enable city governments to be more transparent and residents to access information more easily. This allows for more robust communication between the city government and the public. Cities such as Barcelona are leading the way with various public engagement platforms and open data initiatives.

These applications demonstrate how big data and data analytics provide critical support to city management in achieving sustainable development goals. By making cities more sustainable and resilient, these technologies contribute significantly to the goal of leaving a liveable environment for future generations.

The literature review provides a comprehensive overview of the theoretical foundations and practical applications of big data and analytics in urban management. It synthesizes existing research, emphasizing the potential of data-driven decision-making in enhancing service provision, citizen engagement, and urban sustainability.

Chapter 3

Big Data and Data Analytics at IMM

There are 30 metropolitan cities in Turkey. Cities with a population of over 750,000 are considered metropolitan municipalities under the law. These metropolitan municipalities include the district municipalities within their borders and provide a wide range of services. These services include infrastructure, transportation, environmental protection, education, health, culture and social services.

With 39 districts, Istanbul is the metropolitan city with the largest number of districts in Turkey. Istanbul Metropolitan Municipality is a large organization with more than 90 thousand employees, including a Mayor, a Secretary General, 7 Deputy Secretary Generals, 28 Departments, 112 Directorates and 29 subsidiary companies. It also includes the Istanbul Electric Tramway and Tunnel (IETT) and Istanbul Water and Sewerage Administration (ISKI). The volume of data generated in such a large organization is very high. One of the areas of responsibility of the municipality's IT and Smart City directorate is to store and process big data and produce data-driven decision-making reports and analytical solutions to provide services in line with sustainable urban management.

3.1 The Design of the Methodology

In the intricate urban landscape of Turkey encompassing 30 metropolitan cities, the Istanbul Metropolitan Municipality (IMM) confronted the intricate challenges associated with the management of extensive datasets and the constraints posed by prevailing analytical tools. In response, IMM embarked on a deliberate and strategic endeavor to establish a robust Big Data platform.

The primary objective of this initiative was to harness a myriad of data sources, including but not limited to traffic sensors, social media platforms, and corporate applications, with the intent of addressing multifaceted objectives spanning from disaster management to environmental planning and the optimization of transportation systems. This initiative unfolded through a meticulously planned sequence of strategic

steps, commencing with the delineation of objectives, formulation of preservation policies, and culminating in the successful execution of a Proof of Concept (PoC).

As the Big Data platform materialized, IMM achieved significant milestones, encompassing the explicit definition of objectives, the institutionalization of a managerial framework, and the formulation of a comprehensive reference architecture. Preservation and retention policies were methodically instituted, and valuable insights were gleaned from the successful execution of the PoC. Stakeholder engagement, characterized by audits and surveys, served to fortify collaborative efforts involving institutional entities, consultants, and suppliers.

Persistent endeavors were directed towards the management of data quality, provision of administrative support, and the strategic advancement of data science and analytics. The concluding reflections underscored the pivotal role of the platform in shaping business intelligence.

The subsequent phase of project development witnessed the fruition of conducted PoCs and the initiation of comprehensive project development endeavors, with a specific focus on realizing objectives related to intelligent and sustainable city management.

Confronting the challenges encountered by IMM, a methodical and systematic approach unfolded for each case study, with a particular emphasis on the datasets pertinent to Istanbul Card and ALO153 Call Center applications. This entailed meticulous operations related to data preprocessing.

The Istanbul Card Segmentation Case Study involved the development of a nuanced behavioral segmentation model, integrating approximately 100 variables and employing Recency, Frequency, Monetary (RFM) analysis for the evaluation of customer value. The ALO153 Call Center Sentiment Analysis Case Study entailed the construction of a sentiment analysis model employing TF-IDF embeddings and the SadedeGel Library for the numerical representation of textual data.

The overarching methodology championed a data-driven decision-making approach, with insights derived from segmentation models steering actionable recommendations aimed at optimizing services and augmenting citizen satisfaction. Detailed reports comprehensively documented the entire process, encompassing data preprocessing steps, model development, and ensuing results. Stakeholder engagement manifested in the presentation of findings and collaborative discussions on their implications for decision-making.

A culture of continuous improvement was instilled through the systematic gathering of feedback and a commitment to remain abreast of emerging techniques and technologies within the realm of data science. This comprehensive and meticulous approach seeks to empower IMM in the pursuit of informed decision-making, sustainable urban planning, and the optimization of service delivery.

Each step of the study is summarized in the Figure 3

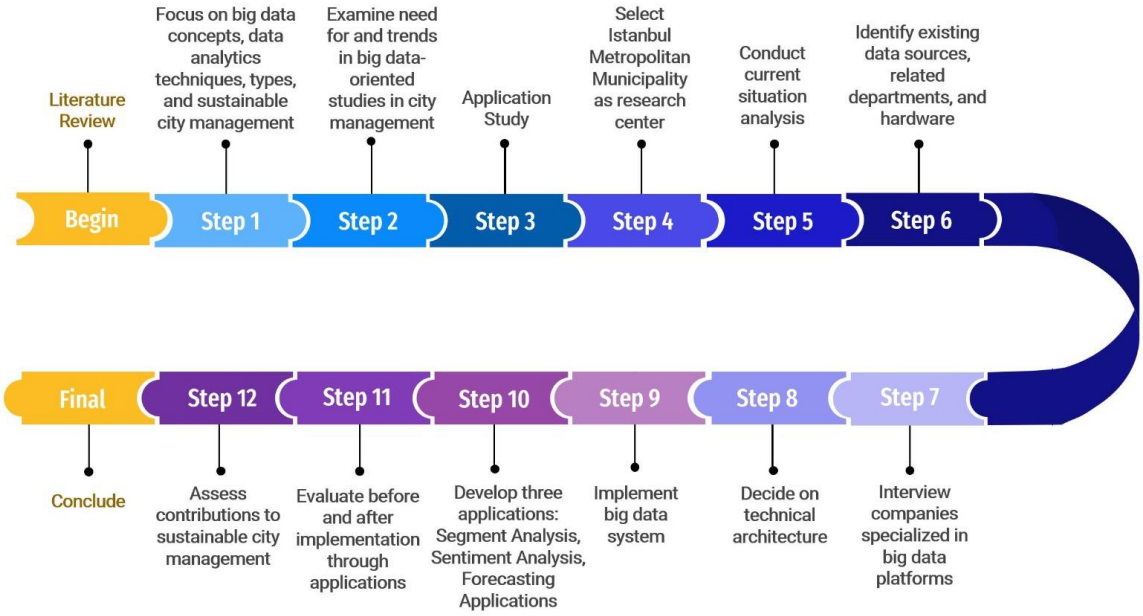


Figure 3. Data Analytic Steps

3.2 Big Data at IMM

IMM's departments, subsidiaries and affiliate companies generate a wide variety and large volumes of data. There are hundreds of different data sources, from corporate applications and CRM systems to social media and sensors. All this data is kept on different platforms for their own purposes and can be analyzed in a way that is usually limited to the relevant field. These analyses are in the form of current situation analysis with data warehouse platforms and are not suitable for making predictions and inferences for the future.

These data sources; traffic sensor data, icing sensors, flood warning sensors, air pollution measurement sensors, meteorology data, vehicle tracking systems, IstanbulSenin super mobile application and mini applications within it, traffic cameras, security cameras, tourist cameras, traffic and road information screens, smart meter systems data, scada systems data, Istanbul card pass data, Ibb-Wifi data etc. In addition to the data, corporate application data; geographical data systems, call center data, social services data, financial services data, license, cemetery, e-municipality application data, cultural services applications, suspended invoice etc. It consists of many more application data such as. Examples of non-structural data include scanned documents, video archive data, camera recording systems, orthophoto maps, system logs, drawing projects, zoning plans, etc.

In general, when IMM is considered, a big data pool is formed when the data provided by affiliated organizations and subsidiary companies, data not included in the KVKK process, and data provided through protocols with external institutions are brought together.

Big data is the platform that IMM try to use in processes such as disaster management, environmental management, zoning management, cultural services management, transportation services management, health and social services management, and in the local management of the brand city that facilitates life with sustainable and innovative solutions and produces global value for urbanism and civilization.

3.2.1 How was it decided to establish a big data platform at IMM? The big data platform has emerged in line with IMM's objectives stated below**.**

- Developing next generation capabilities using emerging big data foundations, techniques and technologies,
- Supporting argument to explore and understand the reliability of data and the resulting knowledge to make better decisions and support important discoveries,
- Enhancing the value of data through policies that support data sharing and management,
- Understand the big data environment, its sharing and use in accordance with privacy, security and ethical rules
- Training and implementation of big data training to meet the growing demand for both deep analytical capability and analytical capacity to create a broader workforce.

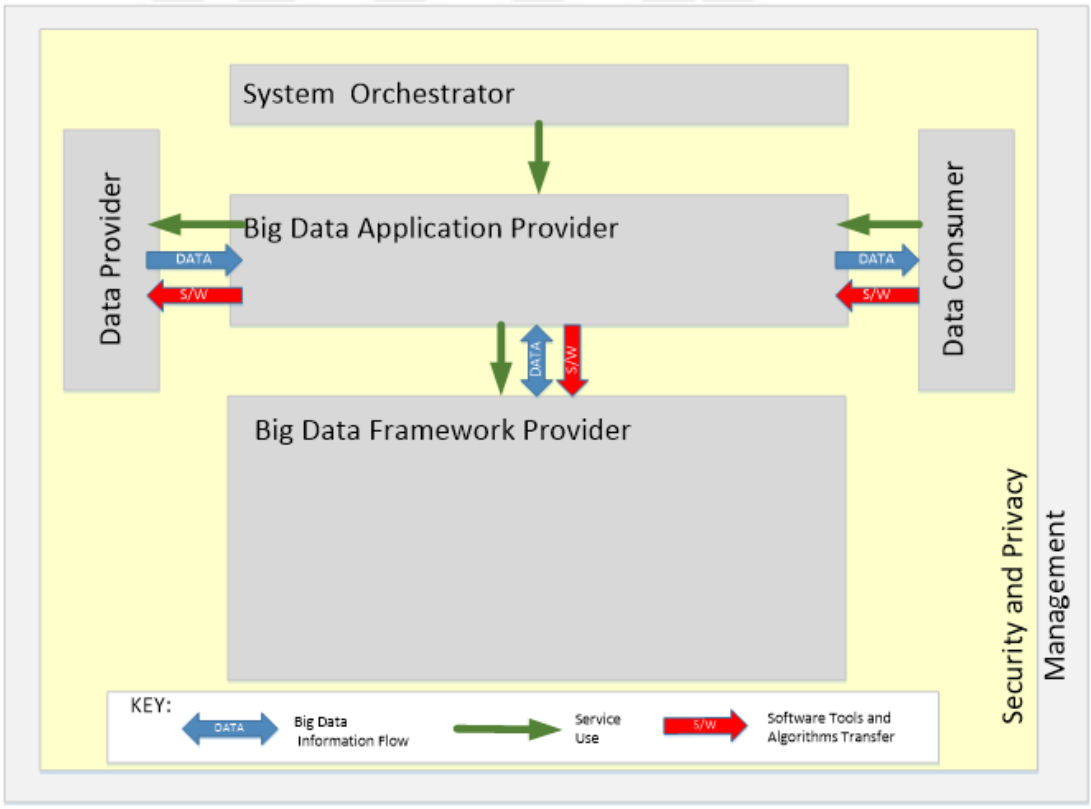


Figure 4. Big data reference architecture

The tasks defined in this reference architecture (Figure 4) are:

- System Planner: Identifies and integrates data application activities required in an operational vertical system,
- Data Provider: Brings new streams of data or information into the Big Data system,
- Big Data Application Provider: Manages a data lifecycle to meet the requirements defined by the System Planner and the security and privacy requirements,
- Big Data Framework Provider: Creates a computational framework to execute specific transformation applications while maintaining the confidentiality and integrity of the data,
- Data Consumer: Includes end users and other systems that use the results of the Big Data Application Provider.

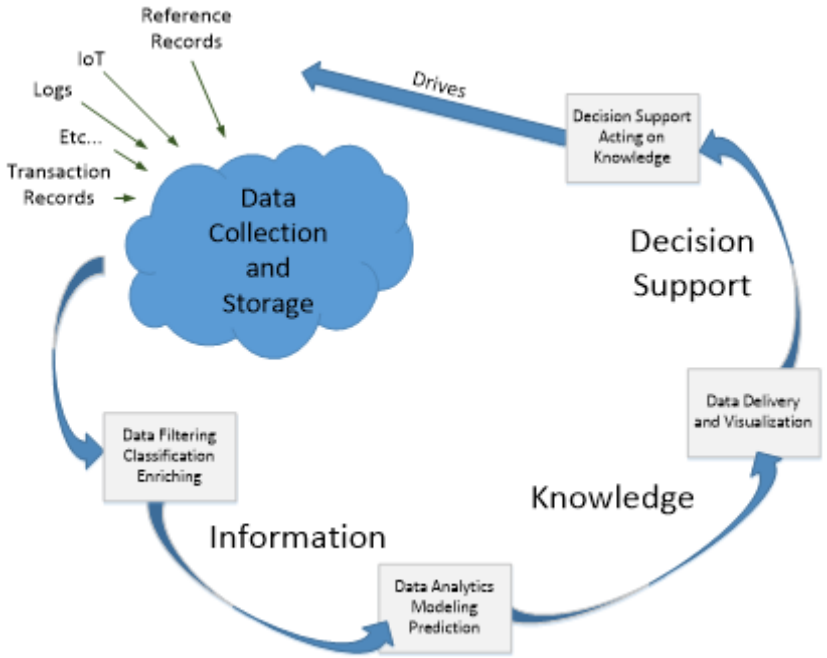


Figure 5. Data collection to decision support cycle

As shown in the Figure 5, the transition from raw data to information, and subsequently to knowledge through data analytics, decision support, and action, can be easily facilitated through data collection, storage, filtering, classification, and enrichment.

The integration of comprehensive big data systems facilitates the establishment of a robust data infrastructure essential for the development of intelligent and sustainable urban systems. This infrastructure serves multiple service domains, including but not limited to urban planning, citizen engagement, and crisis management. It enables multifaceted analyses ranging from cross-sectional and historical descriptive studies to causative and predictive modelling thereby enhancing decision-making processes. Furthermore, it supports the development of responsive disaster and event management systems by allowing real-time situational analysis and predictive assessments, leading to proactive measures and informed urban development strategies.

- Identification of existing resources (e.g. databases in IMM main data center or in independent institutional facilities, etc.)
- Estimation of resource requirements and planned scale to meet 1-3-5 years big data targets. Source databases for scaling, data transfer rates, etc.
- Establish licenses and maintenance-service agreements for the products to be used
- Creating a test environment
- Personnel trainings (Technical and theoretical trainings)
- Long-term storage and archive storage planning

After all this planning and analysis;

- Big data goals, objectives and strategy have been defined.
- A big data management structure was established.
- IMM data lifecycle definition was defined both within the standard data architecture and how the data will be associated with Big Data and Decision Support Systems.
- As part of IMM's management process, a methodology was defined to identify and classify potential sources of data inputs from IoT and other applications. The identified data sources are cataloged and made available to other agencies and, in due course, to citizens as open data.
- A Big Data Reference Architecture was created to ensure that IMM's independent organizations fully understand the mobility of data, the

participants in the Big Data Architecture, the processing of data, and the escrow management of data.

- Specific preservation and data retention policies were established to ensure that data is archived and preserved or deleted in due course during its lifecycle.
- Private, public and industry Big Data Reference Architectures were analyzed.
- Identified and interviewed potential consulting companies and consulting partners that could assist IMM in the early stages (1~5 years) of the Big Data Project.
- An architectural framework and process was selected to guide the Big Data Project and other subsequent technology initiatives. This was revised to incorporate all elements of subsequent project development, project management and lifecycle management of the technology used within IMM.
- IMM has developed a standard data reference model or IMM information exchange model that will be a catalog or reference for data structures and formats. It is essential to support data sharing, including Big Data / IoT data shared across government agencies, and to support the possibility of making this information public.
- A Proof of Concept was prepared as soon as appropriate, with the goal of providing a "quick win" or positive view of how big data can support more effective decision support systems in one or several IMM administrations.
- All IMM institutions were audited and surveyed on basic operating applications. Selected common applications used throughout IMM and worked on a transition plan to transfer common applications that are completely self-running to a centralized application platform. (Centralized data management project)
- IMM is working with IMM stakeholders, consultants and suppliers on a proposal for Information Technologies to develop appropriate data processing models to meet the needs of all IMM organizational stakeholders. This will also depend on policies related to data lifecycle management and the types or structures of data generated from IoT devices and applications.
- Efforts are being made to establish data quality management, assessment and standards using international standards or requirements developed by IMM stakeholders within the organization, or national/local security requirements.

- Full administrative support was provided for the Big Data project.
- The need to develop Big Data, data analytics, and data science as a strategic requirement for Istanbul has been identified. Efforts are being made to increase data science and data analytics projects as an output. Solutions to unit needs are being worked on.

As a result of all these studies, the steps of first conducting a POC on big data and then tendering and starting the development of the project were realized. Figure 6 and Figure 7 show possible scenarios.

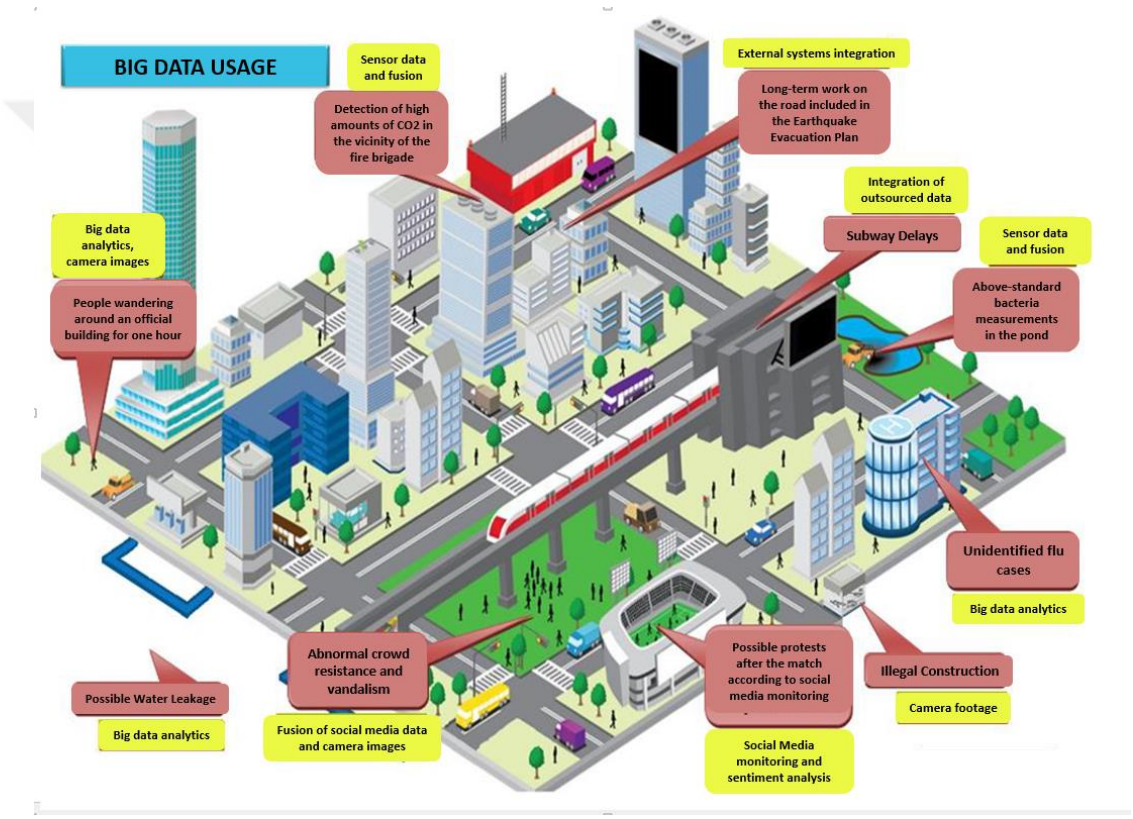


Figure 6. What can be observed with big data?

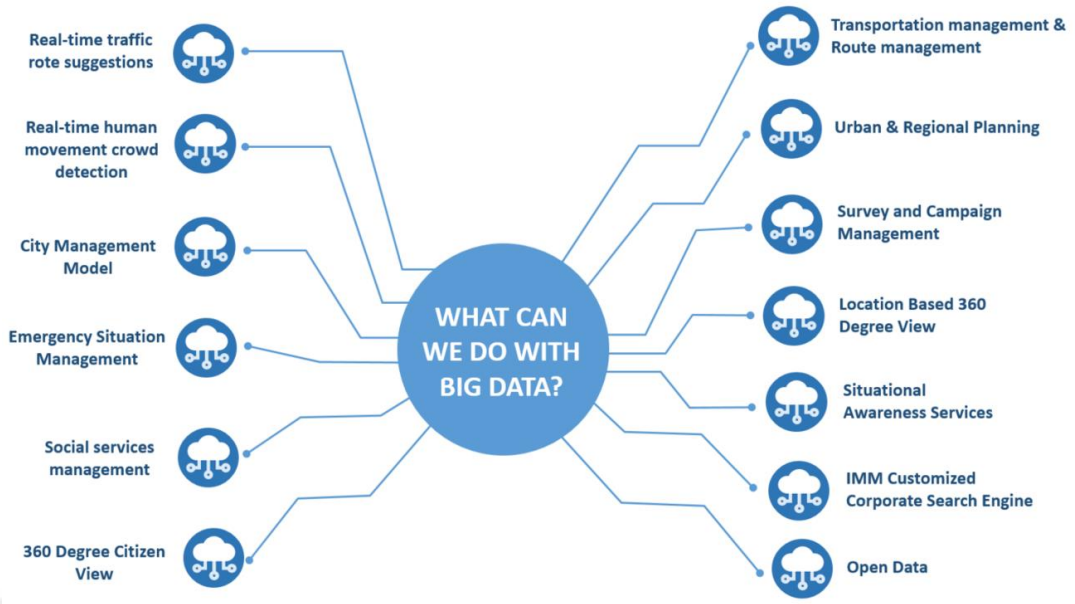


Figure 7. What can be done with big data?

The big data platform is an essential system to support the decision process, 360-degree citizen view, in-house data portal, open data portal, data infrastructure in disaster crisis management, geospatial analysis and analytics, sensor data and central IoT platform, data infrastructure of the digital twin platform, safe society, clean environment, carbon footprint reduction, traceable and measurable city, effective disaster management, traffic congestion reduction and effective management, safe networks, effective use of resources in IMM's targets for smart and sustainable city management.

3.2.2 Big data and data analytics tools and services. What were the big data and data analytics tools and services researched and discussed during the big data platform selection process? For this question meetings were held with technology companies with big data solutions such as Cloudera (URL 3), Hortonworks (URL 4), MAPR (URL 5), Oracle Big Data Appliance (URL 6), IBM big data appliance (URL 7), Teradata (URL 8), Pentaho (URL 9), Huawei FusionInsight (URL 10) etc. and subcontractor companies that will act as their consultants. The big data reference architectures of other companies and public institutions were reviewed. Then, technical specifications were prepared and tendered for all possible technologies through the tender process. Other services and technologies researched here are grouped below.

- Databases;

Databases will be part of the big data system to be established. Infrastructures running Oracle (URL 11) and MS SQL (URL 12) databases used within the organization can be used as structured databases. Neo4j (URL 13), Allegrograph (URL 14), Objectivity InfiniteGraph (URL 15), Oracle Graph (URL 16), etc. as graph databases; HP Vertica (URL 17), Teradata Aster (URL 18), Greenplum (URL 19), Actian (URL 20), etc. as MPP (massively parallel processing) tools; databases such as NuoDB (URL 21), ScaleDB (URL 22) as NewSQL databases that can run on big data platforms will be evaluated. In addition, GPU-based databases such as Kinetica (URL 23), MapD (URL 24), etc. will be evaluated as databases and infrastructures that provide high performance and low operational costs by solving software layer solutions at the hardware layer and providing real-time parallel data processing capability without the need for software layer solutions such as indexing, pre-aggregation or data sharding.

- Development and deployment;

The following tools can be used to accomplish these tasks:
Data Storage; Hadoop (URL 25), HBase (URL 26), Oracle, MySQL (URL 27) etc.
Data Processing; Hadoop, MapReduce (URL 28) etc.
Data Access; Hive (URL 29), Pig (URL 30), Mahout (URL 31), Avro (URL 32), Sqoop (URL 33), Spark (URL 34) etc.

Management; Oozie (URL 35), ZooKeeper (URL 36), Flume (URL 37), Chuwa (URL 38) etc.

Data Integration; Presto (Facebook) (URL 39), Oracle Big Data Sql, Oracle Big Data Connectors, Oracle Data Integrator, IBM Big SQL, Sqoop, Spark, Nifi (URL 40), Flume, Striim (URL 41), Talend (URL 42), Kafka (URL 43) etc.

- Analytics

The following tools can be used for these technologies:

Analytics Platforms: 1010 Data (URL 44), MapR Technologies, Action, Infobright (URL 45), Pivotal (URL 46), HP Haven (URL 47), etc.

Data Science tools: Mahout, Spotfire (URL 48), SAS (URL 49), RapidMiner (URL 50), Weka (URL 51), Predixion (URL 52), R (URL 53), Oracle Advanced Analytics, Oracle R Advanced Analytics for Hadoop, SparkML (URL 54), SPSS (URL 55), DSX (URL 56) etc.

Business Analytics tools: Tableau (URL 57), OBIEE (URL 58), Oracle Data Visualization, Microstrategy (URL 59), Treasuredata (URL 60), Qlik (URL 61), Pentaho, BigML (URL 62), Lavastorm Analytics (URL 63), Cognos Analytics (URL 64), etc.

Social media and sentiment analysis tools: Lexalytics (URL 65), Attensity (URL 66), Sprout Social (URL 67), etc., which can be integrated into data analytics and storage.

After all these preliminary researches, in December 2018, with the "Establishment of Big Data Infrastructure and 360-degree Citizen View Project", structural, semi-structural and non-structural data owned by IMM units were collected on an umbrella platform, relevant decision support reports were produced and analytical scenarios were developed.

Some of the outputs of this process are listed below, Figure 8.

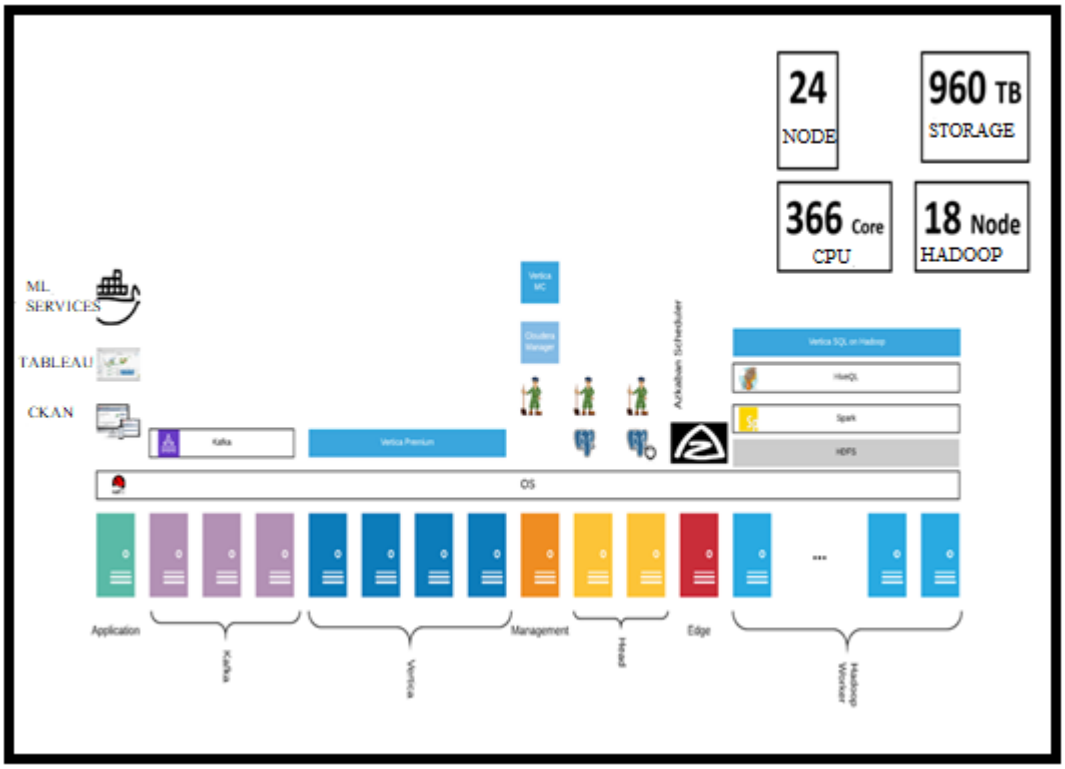


Figure 8. Big Data Infrastructure

The components of the established IMM Big Data Platform are as follows,

- 12xCDH (Cloudera Hadoop Distribution) Worker Servers: Hadoop's machines that perform the actual data processing. It includes the Hadoop software as well as the Vertica SQL on Hadoop software, a SQL-2003 compatible query engine.
- 2xCDH Head Servers: These are the servers on which HDFS name nodes and other Hadoop services that require redundancy (HA) run.
- 1xCDH Management Server: It is the server that runs management software such as Cloudera Manager and Vertica Management Console, as well as software such as ZooKeeper that requires 3-way mirroring.
- 1xCDH Edge Server: It is the server used by the software that provides access to Hadoop and hosts a temporary data storage area (Staging Area) on it.
- 4xVertica Premium Servers: Servers where Vertica software is installed.
- 3xKafka Servers: The servers where the Kafka software is installed, which is the environment where the vast majority of the data that enters the system other than relational databases enters.

- 1xApplication Server: It is the server where software such as Tableau Server, Restful AI services and CKAN (open-source open data portal for the storage and distribution of open data) server are installed.

Explanations about the services working here;

- ML Services: It includes predictive APIs that open to the outside world through Docker containers.
- Tableau Server: It is the software that constitutes the infrastructure of all the panels prepared within the scope of the project. It is the platform where reports prepared with Tablea Prep software are presented to end users.
- CKAN: It is the framework on which the Open Data Portal software is built.
- Vertica Premium: It is the analytical database that directly distributes most of the analytical services provided under the project.
- Cloudera Hadoop Distribution: It is the platform through which most of the ETL flows performed within the scope of the project flow. Apart from this, it is the software umbrella where long-term data storage is realized.

As a result; Structured, semi-structured and unstructured data generated by different applications, services and systems within IMM in processes related to citizens were gathered and processed in a big data pool, the relationships between the data were determined and visualized with visualization tools to provide a 360-degree view of the citizen, and analytics on this data were used to better understand the relations between IMM and citizens, improve IMM services and produce new services.

It was made possible to get to know the citizen very closely, to understand his/her variable needs, to interact instantly and to carry out large works that cannot be done with human resources with artificial intelligence methods.

A big data technology infrastructure that can process, store and run advanced analytical processes on the extraordinary amount of data that will be generated by IOT (Internet of Things) sensor systems that IMM aims to install throughout Istanbul in the near future has also been established.

As in many developed cities, a data laboratory was established to analyze the data to be collected from different sources by data scientists with advanced analytics and to obtain new valuable outputs that will contribute to our institution.

Furthermore, an open data portal has been established to facilitate the development of various applications for Istanbul by students, mobile application developers, as well as local and foreign developers. This initiative enables the utilization of anonymized data for external dissemination, contributing to the recognition of Istanbul as a prominent smart city, akin to other developed urban centers such as New York, London, and Chicago.

This platform enables Smart Istanbul to manage its management more effectively, to know its citizens better, to carry its services to advanced levels, and creates the data infrastructure of the Central IoT Platform and projects such as the Digital Twin.

Some outputs of the platform are shown in Figure 9, 10, 11, 12 and 13.

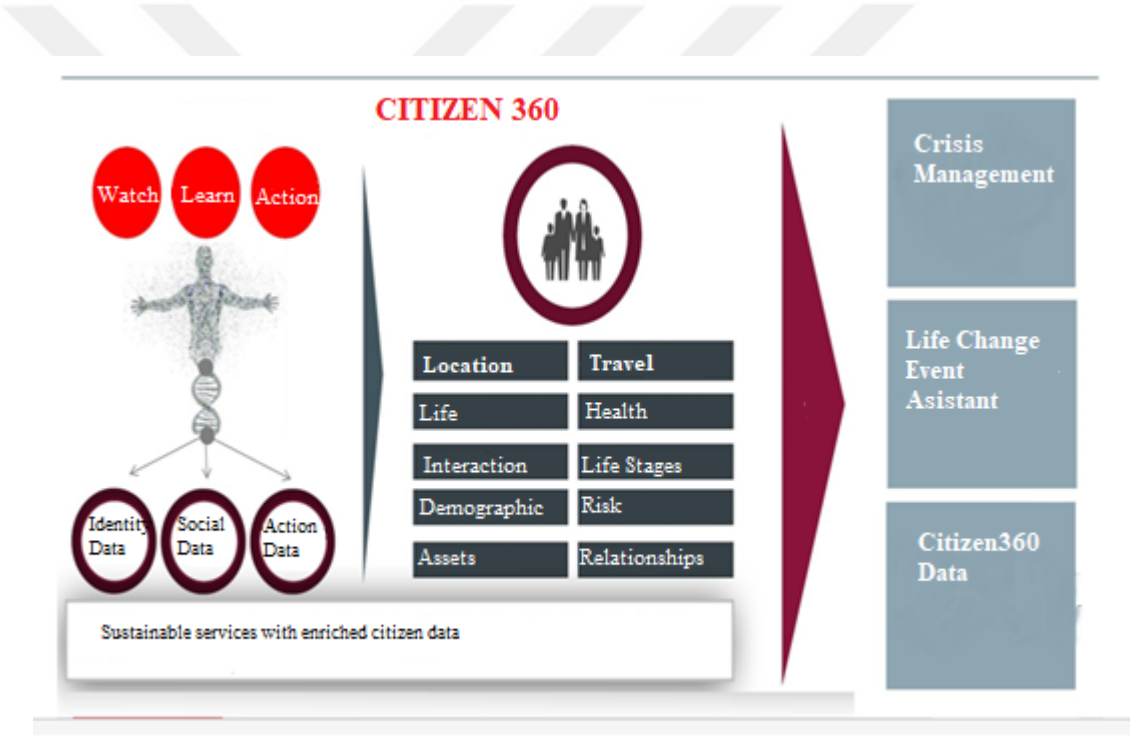


Figure 9. Citizen 360 View with big data

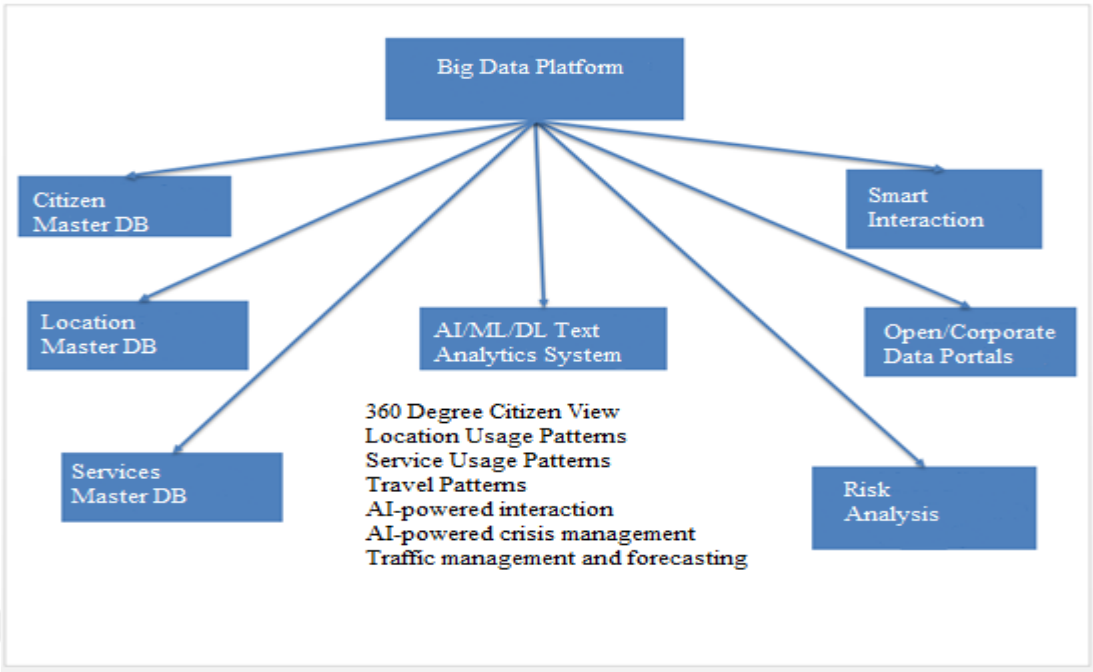


Figure 10. Outputs of big data platform

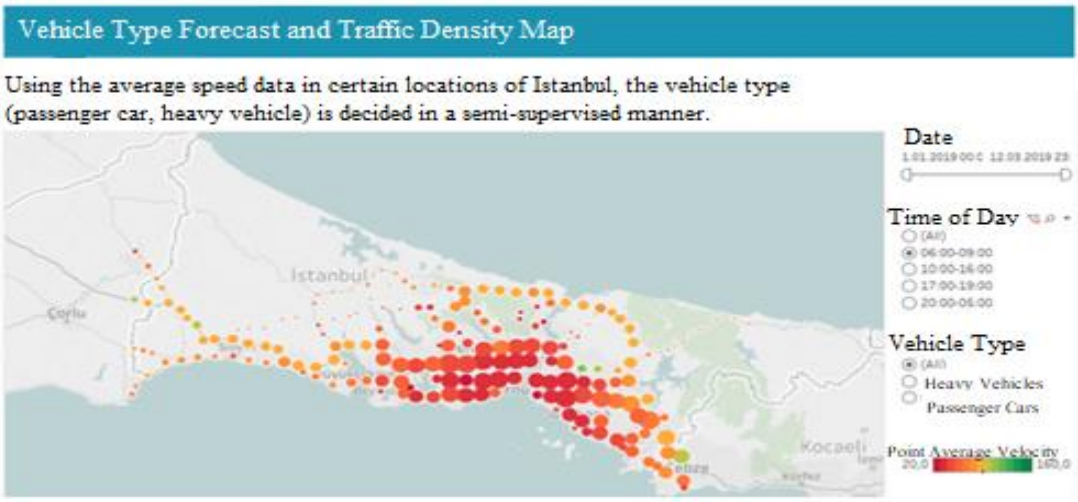
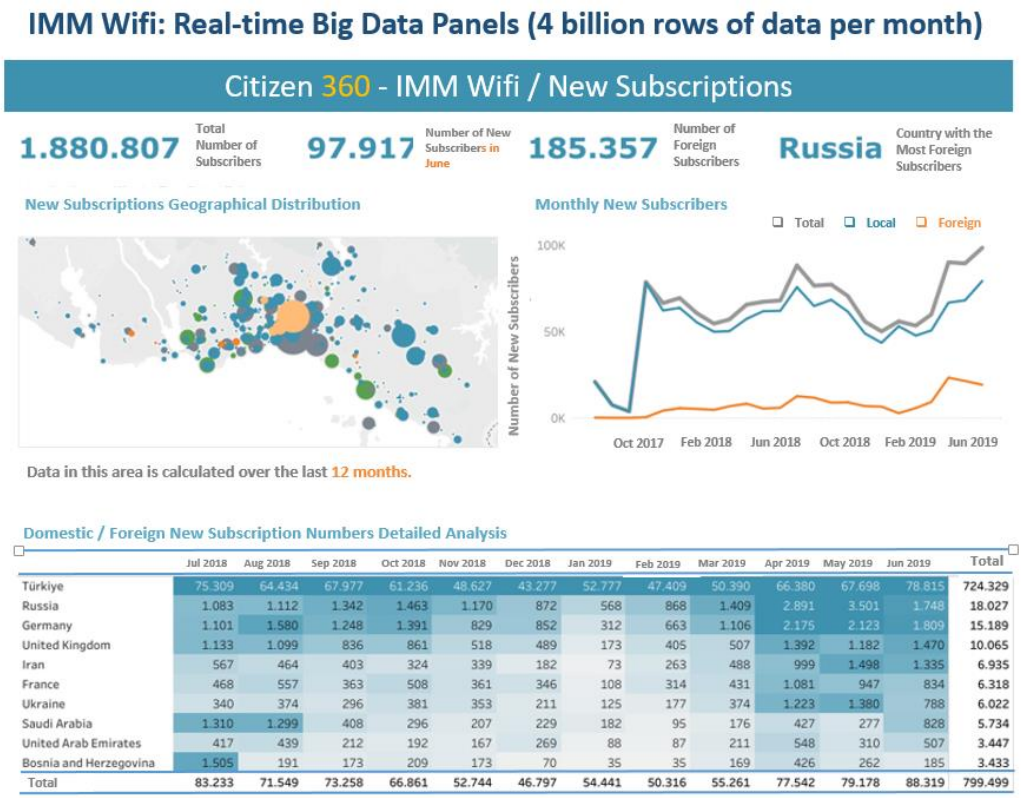


Figure 11. Outputs of big data platform – Traffic Density



	Jul 2018	Aug 2018	Sep 2018	Oct 2018	Nov 2018	Dec 2018	Jan 2019	Feb 2019	Mar 2019	Apr 2019	May 2019	Jun 2019	Total
Türkiye	75.309	64.434	67.977	61.236	48.627	43.277	52.777	47.409	50.390	66.380	67.698	78.815	724.329
Russia	1.083	1.112	1.342	1.463	1.170	872	568	868	1.409	2.891	3.501	1.748	18.027
Germany	1.101	1.580	1.248	1.391	829	852	312	663	1.106	2.175	2.123	1.809	15.189
United Kingdom	1.133	1.099	836	861	518	489	173	405	507	1.392	1.182	1.470	10.065
Iran	567	464	403	324	339	182	73	263	488	999	1.498	1.335	6.935
France	468	557	363	508	361	346	108	314	431	1.081	947	834	6.318
Ukraine	340	374	296	381	353	211	125	177	374	1.223	1.380	788	6.022
Saudi Arabia	1.310	1.299	408	296	207	229	182	95	176	427	277	828	5.734
United Arab Emirates	417	439	212	192	167	269	88	87	211	548	310	507	3.447
Bosnia and Herzegovina	1.505	191	173	209	173	70	35	35	169	426	262	185	3.433
Total	83.233	71.549	73.258	66.861	52.744	46.797	54.441	50.316	55.261	77.542	79.178	88.319	799.499

Figure 12. IMM Wifi – Real Time Dashboard

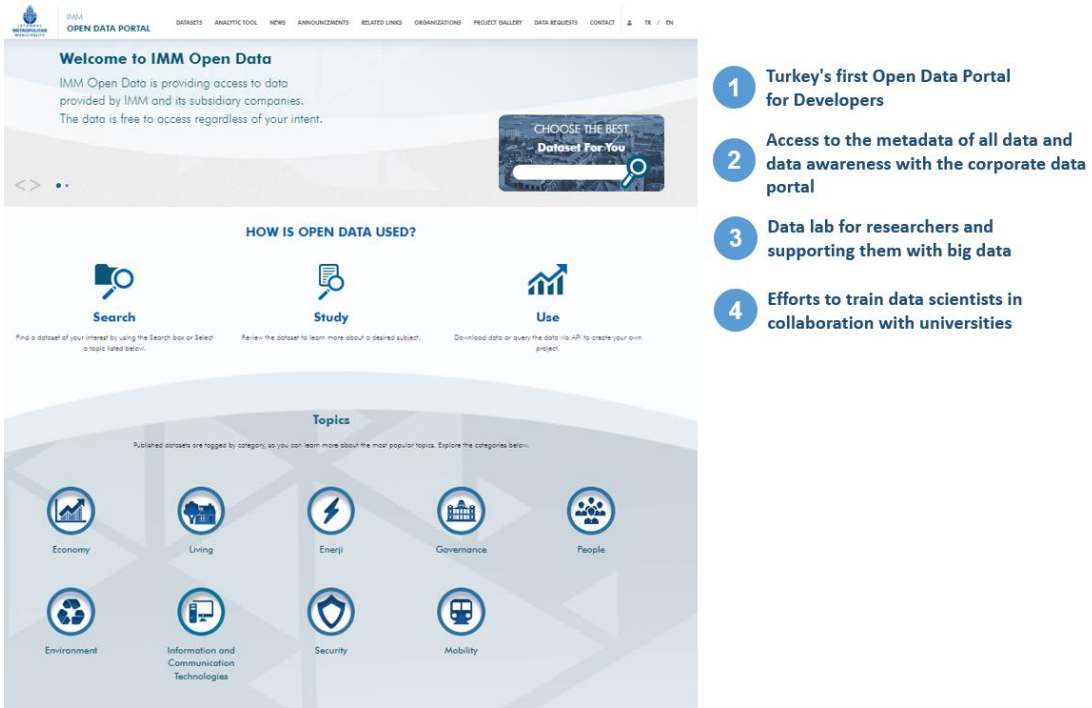


Figure 13. IMM Open Data Portal

3.3 Data Analytics at IMM

IMM has been using existing data warehouse systems within the scope of descriptive and diagnostic analytics for a long time. After the implementation of the big data platform, scenarios were realized on the predictive and prescriptive analytics side thanks to the analytical tools brought to the organization.

In the first version of the big data platform, scenarios such as call categorization, call duration and anomaly detection, call sentiment analysis, vehicle type prediction, Wi-Fi anomaly detection were performed on the analytical database Vertica.

With Vertica, summary statistics, event analysis, graph analysis, spatial analysis, classification, regression, clustering, anomaly detection and dimensionality reduction,

With Tableau, summary statistics, time series regression, trend analysis, data visualization and exploration analysis steps were performed. In addition to the existing data warehouse tools, Tableau was also used within the scope of descriptive analytics.

In the machine learning steps in the analytical architecture, after all the feature engineering steps were completed on Vertica, the final data was modeled on open source Jupyter notebooks and the outputs were re-served on Vertica.

3.3.1 Development processes for analytical work. What are the development processes until all analytical work goes live in a continuous improvement cycle? These processes are; requirement analysis, design cycle, development cycle, test and quality, intake and acceptance processes

a) Requirement Analysis;

- Determining on which applications the data that may be needed to solve the problems related to the problem/scenario are stored and processed,
- Identification of data integrations,
- Examination of data governance methods,
- Determination of technical and functional requirements for the application.

b) Design Cycle

- Interface design and improvements,
- Designing access to data sources and data profiling steps,
- Determination of data policies,

- Determination of modelling and associated data for the scenario to be developed (input data, basic design and results).

c) Development Cycle

- Analyzing the data and the data mining step,
- Data enrichment,
- Data trending and analyzing patterns,
- Modelling,
- Release management,
- Running the model,
- Testing and monitoring,
- Edit and retrain the model.

d) Test and Quality

- Validation of the models,
- Reporting test results,
- Control of output conditions and acceptances,
- Documentation of the work carried out in quality standards,
- Documentation of lessons learned.

e) Intake and Acceptance Processes

- - Bringing models to life,
- - Performance tests,
- - Automation of the common data model,
- - Operation of acceptance processes,
- - Determination of operation processes.

Chapter 4

Analytic Scenarios at IMM

4.1 İstanbul Card Segmentation Case Study

4.1.1 Problem definition and objective of the case study. Istanbul Card is the transportation card used by citizens in Istanbul transportation. Thanks to this card, which is used in transportation vehicles such as buses, subways, metrobuses, ferries, citizens living in the city can meet their transportation needs by topping up the balance on their cards.

In this case study, it is extremely important for a sustainable urban transportation system to bring a perspective on transportation problems with the travel behaviour segmentation of Istanbul card users used for transportation and to ensure more optimized use of resources during transportation hours.

4.1.2 Scope of the case study. The aim of this study is to optimize transportation resources by classifying Istanbul card users into the right classes, to increase citizen satisfaction in transportation and to carry out the necessary work for an efficient transportation system.

4.1.3 Processes of the case study. First of all, Istanbul cards can be divided into 3 main card types as free, discount and full card types. For these card types, daily flowing transportation card pass data (8.5M card issuance) was analyzed and scenarios related to the transportation needs of the institution were realized. The distribution among these card types is free 9.5%, discount 21.9%, full 68.5%.

Approximately 100 different variables were created to segment card usage behaviour. The variables used and generated here can be grouped under 3 main headings such as demographic (age, gender, floor type, etc.), usage information (number of days of card use, total number of card issues, time of card use - morning, daytime, evening, night, weekday, weekend, etc.) and balance loading information (total amount of balance loading, number of balance loading, frequency of balance

loading, use of different locations for balance loading, non-transportation spending trend, etc.).

As a result of the developed behavioral segmentation model, 2 different behavioral segments were created for free cards and 4 different behavioral segments were created for discount and full cards.

a) Free Cards ;

The behavioral segments produced in the study for the free card type were labeled as frequent users and rare users. The distribution here is 20% frequent users and 80% rare users.

Citizens in the free card - frequent users segment prefer the time of day more frequently, and the average number of card issues is 2.6, with the most frequently used public transportation vehicles being bus and metro, respectively.

Citizens in the free card - infrequent users segment; more than half of this segment chooses daytime as their travel time preference and the average number of card issues per day is 1.8. The most frequently used means of public transportation in this group are bus and subway, respectively, as in frequent users.

Table 1

Istanbul Card Segmentation, Free Card, Frequent users

Average Age	Man	Woman	Avg. Number of Cards used per day
48.5	67%	33%	2.6

Table 2

Istanbul Card Segmentation, Free Card, Frequent users, Travel Percentages by time zone

Morning(06-09)	Daytime(10-16)	Evening(17-19)	Night(20-05)
25%	38%	23%	9%

b) Discounted Cards ;

The behavioral segments produced in the study for the discounted card type are distributed as subway preference (11%), bus preference (14%), travelers (10%) and the group that prefers public transportation less (65%).

Citizens in the discounted card - subway preference segment consist of the citizens who print the most cards among the discounted group. The time between two card issuance is approximately 10 hours. They definitely use the subway every day. They rank 3rd in terms of average and total balance top-ups, 2nd in terms of number and 3rd in terms of amount of non-transportation expenditures.

Table 3

Istanbul Card Segmentation, Discounted Card, Subway preference users

Average Age	Man	Woman	Avg. Number of Cards used per day
24.5	47%	53%	3

Table 4

Istanbul Card Segmentation, Discounted Card, Subway preference users, Travel Percentages by time zone

Morning(06-09)	Daytime(10-16)	Evening(17-19)	Night(20-05)
28%	26%	26%	15%

Discounted card - those who prefer buses; this segment type is the 2nd segment that prints the most cards in the discounted card group. The average time between two card issuances is 19 hours. It is the 2nd segment with the highest average and total balance top-up amounts. Although this is the segment with the highest amount of non-transportation expenditures, it is the 3rd segment in terms of number.

Discounted card - travelers; it is the 3rd segment with the highest number of card issuances within the discounted card group. The average time between two card issuances is 27 hours. Compared to other segments, ferry and motor boat usage is the highest. Citizens in this segment make the highest average and total balance top-ups. It is the segment with the 2nd highest amount of non-transportation expenditures.

Discounted card - public transportation distance group; is the group that prints the least number of cards within the discounted card group. The average time between two card issuances is 80 hours. It is the group with the lowest average and total balance loading.

c) Full Cards ;

The behavioral segments produced in the study for the full card type are distributed as frequent users (14%), people coming from Adalar (13%), people who prefer daytime (66%) and people who prefer public transportation less (7%).

Full card - frequent users segment; It is the segment consisting of citizens who print the most cards among full card users. The time between two card issuances is 22 hours, and the segment with the highest average and total balance top-ups.

Table 5

Istanbul Card Segmentation, Full Card, Frequent users

Average Age	Man	Woman	Avg. Number of Cards used per day
39	42%	58%	2,6

Table 6

Istanbul Card Segmentation, Full Card, Frequent users, Travel Percentages by time zone

Morning(06-09)	Daytime(10-16)	Evening(17-19)	Night(20-05)
31%	28%	26%	15%

Full card - People coming from Adalar segment; This segment has the second highest number of card issuances among the full card group. Although the average time between two card issuances is 40 hours, it is the second segment in terms of average and total balance top-ups. They definitely use the subway every day and print cards during daytime hours.

Full card -Intraday preferential segment is the 2nd segment that prints the fewest cards in the full card group. The average time between two card issuances is 109 hours. It is the 2nd segment with the lowest average and total balance top-ups.

Full card -Distanced to public transportation; It is the segment with the least card issuance within the full card group.

4.1.4 Datasets of the case study. In this case study Istanbul Card users demographic information and card types are used in the model. Another dataset is daily card usage data which contains transaction date, route, bus, metro or metrobus usage etc. In addition Istanbul Card fee filling information and refund information is important for the model. And the last dataset for the model contains information about the refills made within the scope of social assistance other than transportation expenditures.

4.1.5 Selected algorithms and results of the case study. In this case study Istanbul Card users demographic information and card types are used in the model. Another dataset is daily card usage data which contains transaction date, route, bus, metro or metrobus usage etc. In addition Istanbul Card fee filling information and refund information is important for the model. And the last dataset for the model contains information about the refills made within the scope of social assistance other than transportation expenditures.

a) What is RFM?

RFM is a method for analyzing customer value and is used to develop service and marketing techniques specific to the groups formed.

Recency: Indicates when the customer last purchased a product or service.

Frequency: Indicates the total number of times the customer purchases a product or service.

Monetary: Indicates the monetary value of the customer's total purchases.

After calculating the RFM values according to the relevant definitions, these values are converted into scores and called RFM scores. Segments are created based on this score.

b) RFM Segments and Actions

Table 7

RFM Segments and Sample Actions

Segments	Sample Actions
THE BEST Recent use, frequent and high spending users who do.	- Rewarding - Early adopters for new products - Brand ambassadors
Segments	- Sample Actions
LOYALISTS Most recent use, often good spending users.	- High value product proposals - Getting feedback through surveys
POTENTIAL LOYALISTS Recently started to use it but it's good quantities and multiple uses, average and six money spenders.	- Loyalty programs - Personalized recommendations.
NEW COMERS Recent use but frequent use those who are not	- New user orientations (onboarding support) - Relationship development
PROMISING ONES Recently used, averaging about sixpieces, those with average and above average use	- Personalized recommendations - Free product recommendations
ONES TO WATCH OUT FOR Above average proximity, frequency and monetary value, those about to be lost if not activated	- Time-constrained recommendations - Recommendations based on past uses
ONES ABOUT TO SLEEP It has below average proximity, frequency and monetary value, those about to be lost if not activated	- Offer popular services at a discount
RISKY High although in montanes and in frequent use, long time no use	- Personalized communication - Discount coupons
ONES NOT TO LOSE Often very high large-scale utilization and long those that have not been in use for some time	- New service proposals
THOSE IN SLEEP MODE Low In the montanes, there is little use and has been for a long time.	- Creating brand awareness - Free product recommendations

RFM is used in this case study and results are in Figures (14, 15, 16);

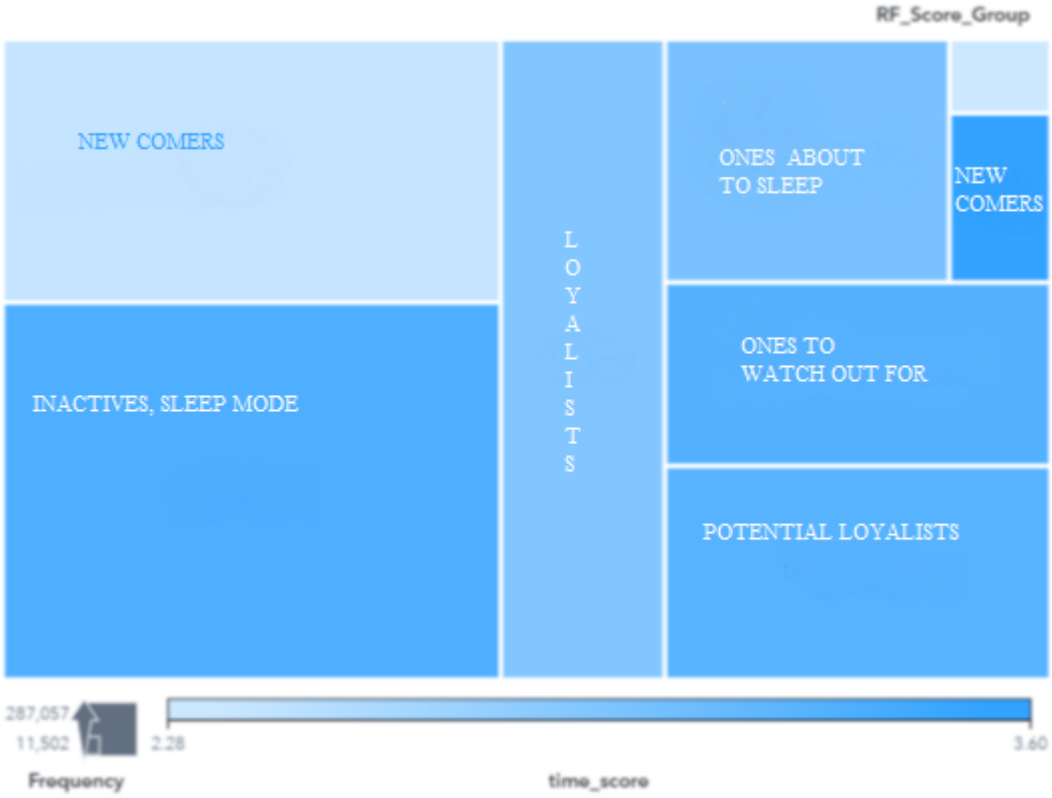


Figure 14. Free Cards RFM Segments

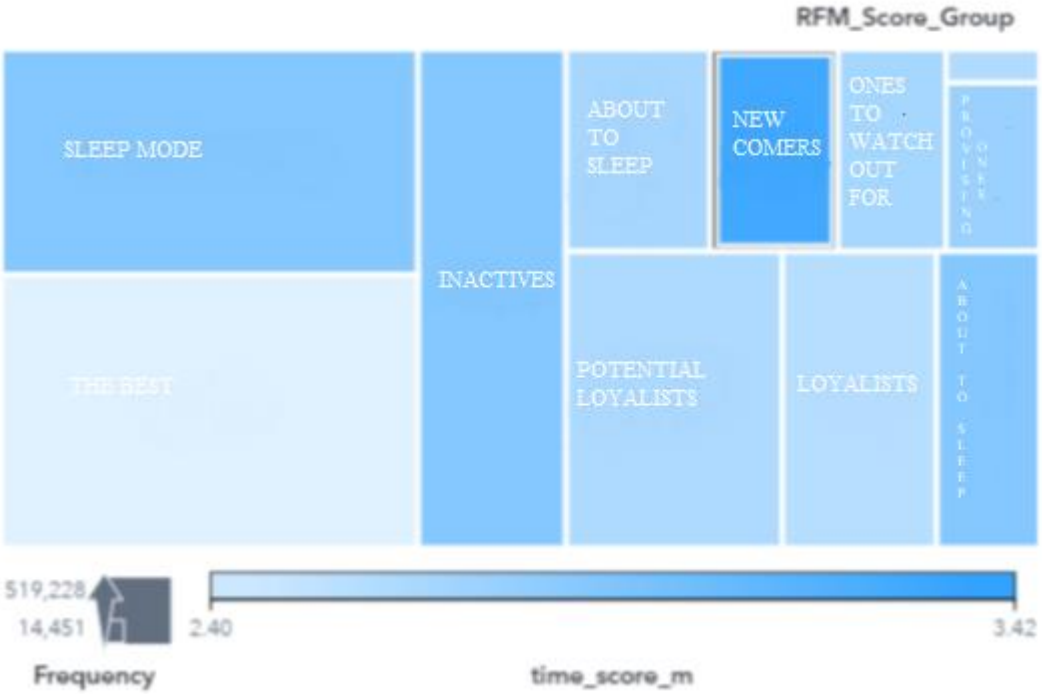


Figure 15. Discounted Cards RFM Segments

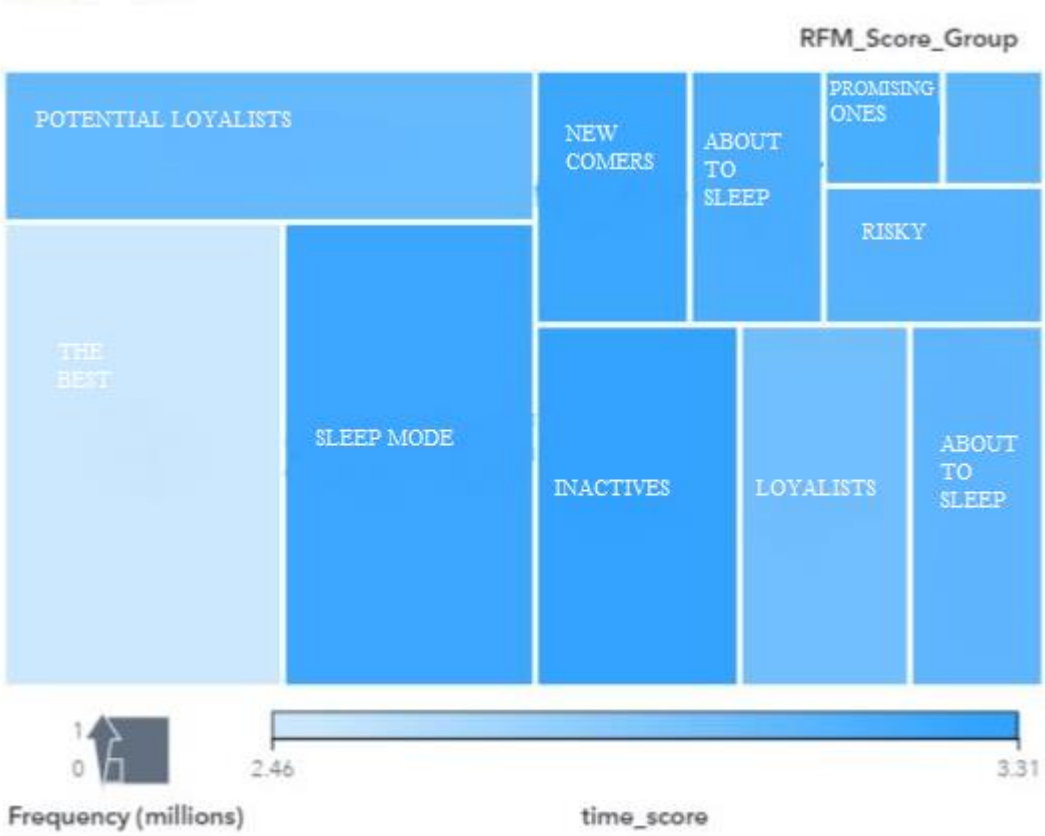


Figure 16. Full Cards RFM Segments

As a result, citizens of Istanbul travel with Istanbul card in public transportation in the city. Segmenting these users is an extremely important study for public transportation planning and vehicle stop optimization. The Istanbul Card Segmentation case study examines the behavioral patterns of card users to optimize public transportation services. It demonstrates the value of RFM analysis in segmenting users and tailoring services to meet their specific needs, thereby enhancing customer satisfaction and efficiency in public transit.

4.2 ALO153 Call Center Sentiment Analysis Case Study

4.2.1 Problem definition and objective of the case study. IMM Call Center is a system through which citizens can communicate their problems related to Istanbul to the municipality through a call center, social media and web page. The verbal, written and social media calls are also processed within the organization and directed to the relevant departments to be resolved as soon as possible. In order for the call center employee to be able to approach the process more positively and harmoniously according to the communication language used by the citizen regarding the previous requests in a written request or a verbal call, it is important to be able to analyze the positive, negative and neutral emotions of the applications regarding the municipality. In this way, the call center employee will be able to choose the right communication language in the process and as a result, it is extremely important in terms of providing sustainable municipal services that make data-driven decisions by increasing citizen satisfaction.

4.2.2 Scope of the case study. The aim of this study is to classify the citizen requests coming to the call center in 3 different statuses such as positive, negative or neutral about the communication language of the citizen about the municipality and to ensure that a correct communication language is preferred with the citizen through the call center. It is aimed to make the correct class assignment by analyzing the application texts coming to the call center with natural language processing methods.

4.2.3 Processes of the case study. In data science projects, we start the process by first separating the working data we receive as train, development and test data. It is extremely important to make this distinction at the beginning of the process in order to ensure that the success of the developed model is at the desired levels and that there is no biased learning and overlearning. So we have 12 partitions of our data and we split it like 7 of them for train, 1 of them for development and 4 of them for test. Every partition has over 1M sentence. After all, we should be able to predict the relevant call text as positive, negative and neutral classes.

When we look at the data set we have and do a preliminary labeling, we see that the vast majority of the data has negative and neutral labeling. In this case we have to deal with classification problems where there is an imbalance between the classes.

- An imbalanced classification problem is an example of a classification problem where the distribution of examples across the known classes is biased or skewed.
- Imbalanced classifications pose a challenge for predictive modeling as most of the machine learning algorithms used for classification were designed around the assumption of an equal number of examples for each class. This results in models that have poor predictive performance, specifically for the minority class.

In the preprocessing steps, we perform clean and normalization operations on the data. The operations performed in this process;

- Free text is not standardized and full of typos, non-alphanumeric characters.
- Standardization is vital for obtaining proper tokens and vocabulary with reduced noise. We make standardization via removing URL's, emojis, HTML tags, whitespaces, empty documents and changing capital letters to lowercase.

In the visualization processes performed after this process; After getting rid of substandard text pieces, we wanted to check cleaned word structure in each sentiment class. And there is some graphics about common words, common n-grams and word count below in Figures (17, 18 and 19).

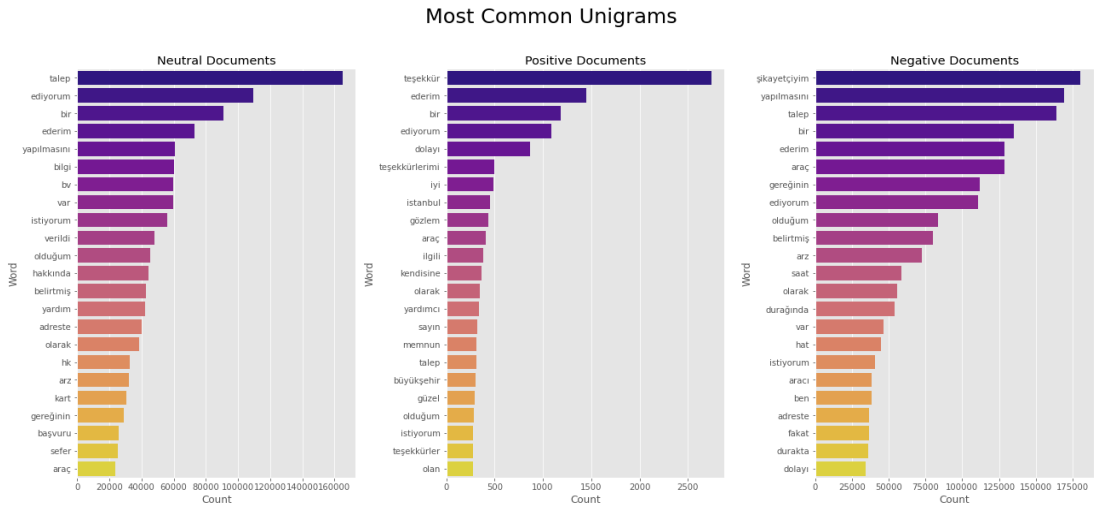


Figure 17. Most Common Unigrams

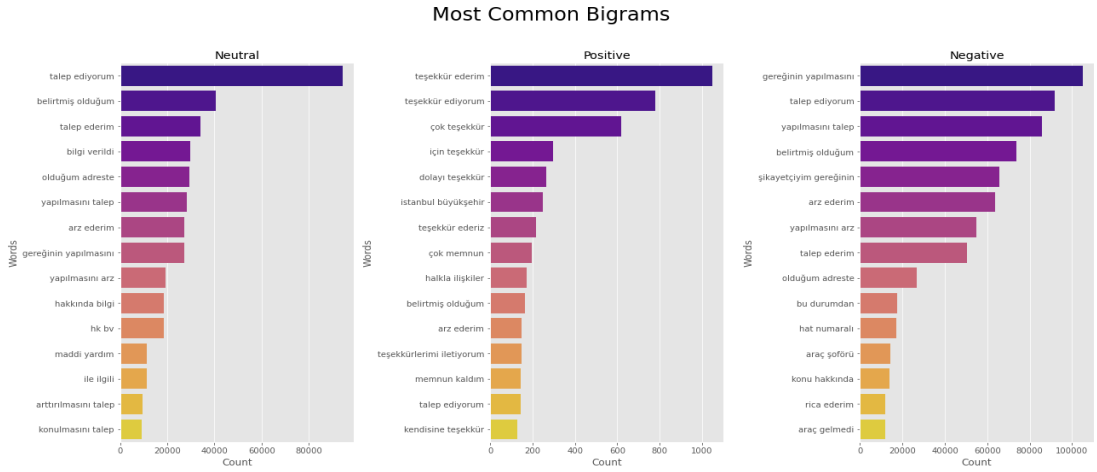


Figure 18. Most Common Bigrams

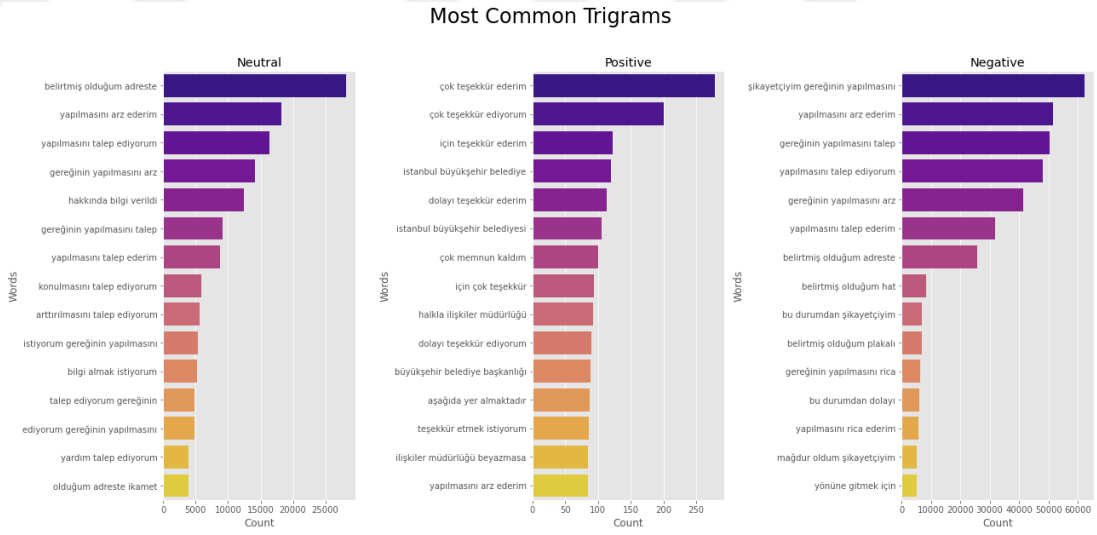


Figure 19. Most Common Trigrams

4.2.4 Datasets of the case study. In the context of the ALO153 solution center, proactive measures are undertaken upon receipt of text-based inputs from citizens, which encompass requests, grievances, and proposals communicated through email, telephone, or social media platforms. Within the dataset, several dimensions are present; however, those pertinent to our project include the descriptive text of the citizen's request and the communication method utilized. These dimensions are instrumental for the generation of district, neighborhood, and street-specific reports, as well as for the construction of predictive models relevant to these forecasts.

4.2.5 Selected algorithms and results of the case study. Examines the selected algorithms and the outcomes of the case study, which are as follows:

a) Term Frequency-Inverse Document Frequency (TF-IDF) Embeddings:

TF-IDF is a statistical measure used to evaluate the significance of a word within a document relative to a collection of documents, commonly applied in text mining and natural language processing domains. It quantifies the uniqueness of a term by comparig its frequency in a specific document to its distribution across other documents. For the numerical representation of texts, TF-IDF embeddings are utilized where:

Term Frequency (TF) refers to the frequency of a term in a document, normalized to account for document length.

Inverse Document Frequency (IDF) assesses the informativeness of a term based on its prevalence across all documents in the corpus.

For example, if the term "Ali" appears twice in a 10-term document, its TF is calculated as 0.2. Assuming "Ali" is found in 3 out of 100 documents, the IDF is log(100/3), approximately equal to 3.5, rendering the TF-IDF score for "Ali" to be 0.7. TF highlights the term's importance within a specific document, while IDF diminishes the weight of terms that occur frequently across the corpus.

b) SadedeGel Library:

SadedeGel is an open-source library designed for Turkish Natural Language Processing (NLP) tasks, particularly specializing in the summarization of Turkish news texts via extraction-based techniques.

The SadedeGel Library (Figure 20) contains approximately 27,000 tokens. During the embedding process, the TF-IDF vector is computed for each sentence and aggregated to represent the document.

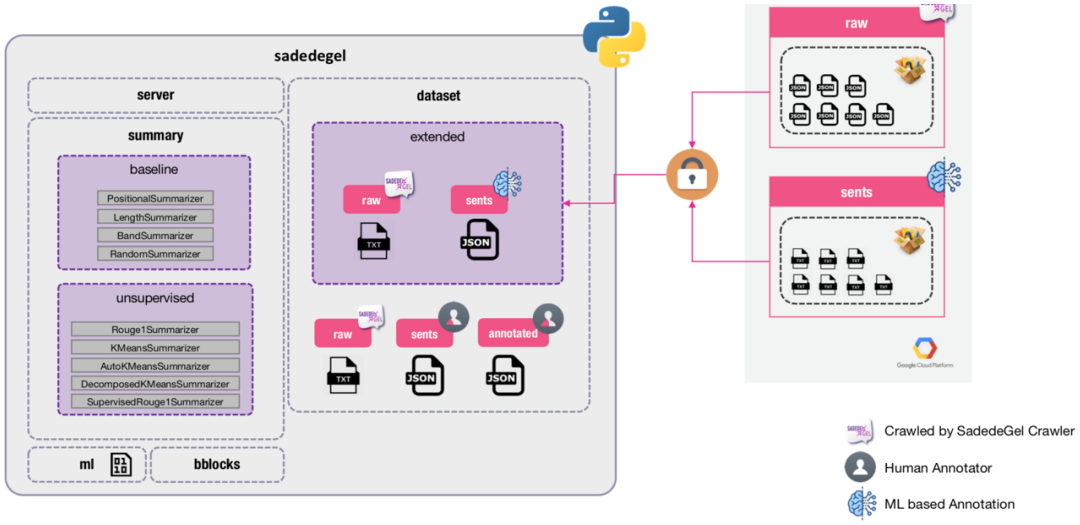


Figure 20. SadedeGel Library Infrastructure (URL 2)

This case study entails a multiclass classification problem, typically assessed using the following metrics:

Average Accuracy

F1 Score

Log-loss

Due to class imbalance, the F1 Score is chosen as it offers a more nuanced evaluation of incorrectly classified cases, particularly in imbalanced datasets.

Tables 8, 9, and 10 present the performance metrics under different data distributions:

Table 8

Metrics of Equally distributed data

Sentiment	Precision	Recall	F1 Score
NEGATIVE	0.74	0.88	0.81
NEUTRAL	0.90	0.68	0.78
Table 8 (Continue)			
Sentiment	Precision	Recall	F1 Score
POSITIVE	0.08	0.92	0.15
Macro Avg.	0.57	0.83	0.58
Weighted Avg.	**0.82**	**0.77**	**0.59**

Table 9

Metrics of Unsampled under-represented class data

Sentiment	Precision	Recall	F1 Score
NEGATIVE	0.87	0.81	0.84
NEUTRAL	0.85	0.89	0.87
POSITIVE	0.41	0.87	0.56
Macro Avg.	0.71	0.86	0.75
Weighted Avg.	**0.85**	**0.85**	**0.85**

Table 10

Metrics of Equally distribution kept data

Sentiment	Precision	Recall	F1 Score
NEGATIVE	0.87	0.81	0.84
NEUTRAL	0.85	0.90	0.87
POSITIVE	0.79	0.62	0.70
Macro Avg.	0.84	0.77	0.80
Weighted Avg.	**0.86**	**0.86**	**0.85**

In an analytical evaluation of classification performance, Tables 8, 9, and 10 demonstrate the outcomes under varying conditions of data distribution. Table 8 reveals that when data is equally distributed, the precision for negative sentiment classification is moderately high, yet the model struggles with precision in classifying positive sentiments, resulting in a lower macro-averaged F1 score. Conversely, Table 9, which presents metrics for data with an under-represented class without sampling adjustments, indicates an improvement in precision and recall across all classes, reflected in a higher macro-average F1 score. Finally, Table 10, which also assesses an equally distributed dataset, shows a significant increase in the F1 score for the positive class and overall precision and recall, leading to the highest recorded macro-averaged and weighted-averaged F1 scores. These findings illustrate the impact of data distribution on model performance, with balanced data yielding a more uniform classification efficacy across different sentiments.

The ALO153 Call Center Sentiment Analysis case study explores the sentiment of citizen requests received by the call center, leveraging TF-IDF embeddings and the SadedeGel Library. It underscores the importance of data preprocessing and machine learning techniques in extracting actionable insights from unstructured textual data, leading to improved crisis management and citizen relations.

4.3 Traffic Intensity Forecasting Case Study

4.3.1 Problem definition and objective of the case study. One of the biggest problems of metropolitan cities is the increasing population and the transportation problems that arise with it. Istanbul is the 1st city with the densest population in Turkey, the registered population in the city is 16.5 million, and the traffic problem is one of the biggest problems of the city in a city where the daily population reaches 20 million with the effect of immigration and tourism.

In order to meet the public transportation needs of the people of Istanbul and to cope with the traffic problem, continuous improvement works are being carried out in the city and solutions that can prevent congestion are being sought and developed. Within IMM, methods including artificial intelligence, machine learning and data analytics technologies have been tried and necessary studies have been carried out and continue to be carried out in order to provide solutions to this problem. Data science techniques have been applied in detail to predict traffic density, identify anomalies in traffic, and optimize the vehicles used in public transportation using transportation data within IMM.

The long time spent in traffic has a large share in the decrease in the quality of life of the person. With the analytical study, it is aimed to reduce travel times thanks to the predicted traffic density, to reduce fuel consumption, thus reducing carbon emissions and, in a holistic approach, to improve the quality of life of citizens.

4.3.2 Scope of the case study. This case study, which was developed to be a solution to Istanbul traffic, focuses on predicting traffic flow and road and district-based density.

By creating predictive models of traffic and traffic flow at important intersections, it includes speed forecasting and forecasting the routes and road sections

where Istanbul traffic is expected to be the busiest on the predicted day, i.e. traffic flow. Traffic density and speed are estimated by districts. By estimating the distribution of the estimated speed value according to the hours of the day, transportation demand and traffic flow are estimated.

In addition, the project aims to reduce travel times, reduce fuel consumption, reduce carbon emissions and thus improve the quality of life of citizens.

4.3.3 Processes of the case study. These processes are shown in Figure 21:

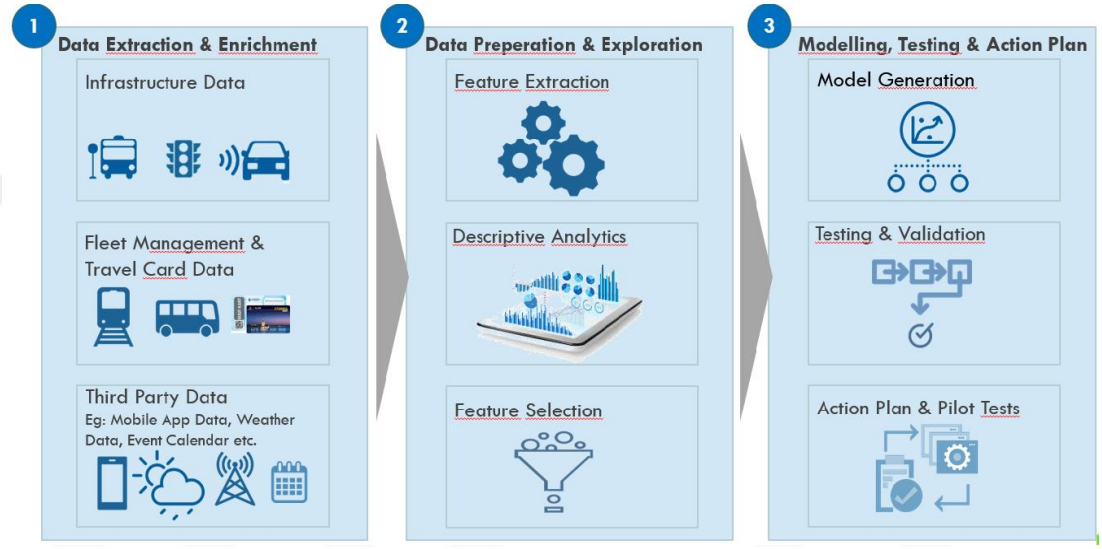


Figure 21. Processes of the case study

Data Extraction and Enrichment (Figure 22);

- Start with data; Review existing sources of information (sensor, smart fleet management data, traffic lights data, IMM traffic data etc.)
- Determine readily available, potentially useful external data sources; mobile application data (velocity), sociodemographic datasets, weather data, calendar events etc.
- Merge data sources; bring all data to one system, determine merge keys, crosscheck accuracy.
- Cleanse the data; remove unreliable and low quality data fields.

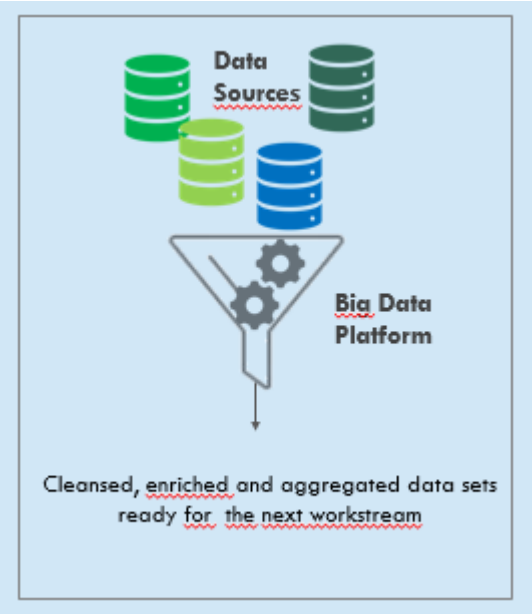


Figure 22. Data preparation and exploration

Data Preparation and Exploration;

- Feature extraction; Transform datasets to generate new structured datasets to be used in analysis and feature selection. For example; bus stop based dataset, route-based dataset, user-based origin-destination dataset, traffic density dataset etc.
- Descriptive Analytics; Conduct descriptive analysis to better understand significant patterns in data. For example, time-based analysis, event/activity based analysis, user behaviour, detecting and excluding outliers.
- Feature selection/Time-period selection; Using descriptive analysis outputs to decide targets and features to be used in the models. Conducting further analysis (classification, PCA, regression etc.) to reduce the number of features to be used in the modelling phase.

Modelling, Testing and Action Plan;

- Define required number of separate models to cover differing conditions; workdays, weekends, school days, specific weather condition, specific events (football matches, concerts, conferences, infrastructure maintenance etc.)
- Multiple modeling techniques will be employed to construct models, aiming to select the one with the highest accuracy. Utilizing a combined approach

enables the exploration of various models concurrently. Potential algorithms include Decision Trees, Support Vector Machines (SVM), Time-series analysis, and K-Nearest Neighbors (KNN), among others. The study will evaluate both agent-based and activity-based approaches, with a plan to integrate findings from both methodologies.

- Design a test and evaluation methodology, Define the criteria for goodness of a model (accuracy, rate, input data reliability, ease of data preparation, ease of interpretation etc. Define the data on which these criteria will be tested.
- Run the alternative models on predefined test datasets
- Rank the models according to the predefined test and evaluation criteria; accuracy rate, ease of interpretation, ease of data preparation.
- Choose the final model or a combination of final models for each target.
- Go back to modelling phase if needed.
- Wrap-up important findings and significant insights.
- Develop an action plan
- Design pilot tests for each proposed item in the action plan.

4.3.4 Datasets of the case study. In this case study the average hourly speed value of each road and segment (a designed area on the road) is used. Also segment direction and vehicle count per segment are important.

In addition, parameters such as speed color of the segment with previous information (fast-green, average speed-yellow, slow-red), district information, tunnel, road type are also important for the model.

Another dataset used in this model is station-based meteorological data such as temperature, humidity, precipitation, wind speed and direction.

And also road credentials, public holidays and the numbers of important points such as education, health, bus stops, trade, parking lots, market places etc., is important for the model.

4.3.5 Selected algorithms and results of the case study. Among the forecasting algorithms, hierarchical model, panel series neural network, seasonal model, non-seasonal model, auto forecasting model algorithms are used.

Among all models, ARIMAX and ESM models are used in the model development phase. Auto forecasting is a structure produced by SAS and develops the best model to produce forecasting models from ESM, IDM, UCM, ARIMAX (URL 68) model sets.

MAPE (Mean Absolute Percentage Error) is chosen as the selection criterion for selecting the best model.

Hierarchical Model; automatically develops predictions for each specified hierarchy.

Model success criteria: Seasonal Model has the lowest MAPE value of 6.6601 on the reference day. Generally, Seasonal Model and Hierarchical Models are selected as the best model.

Hourly speed and index are estimated for the forecasted day and the following 4 days for a total of 5 days. The 15 routes and 15 road segments expected to be the busiest on the predicted day, the district-based traffic density distribution and the change in the speed values on the predicted day during the day are listed.

The prediction for the school opening day was 98.2% correct for the D100 highway Anatolian-European crossing.

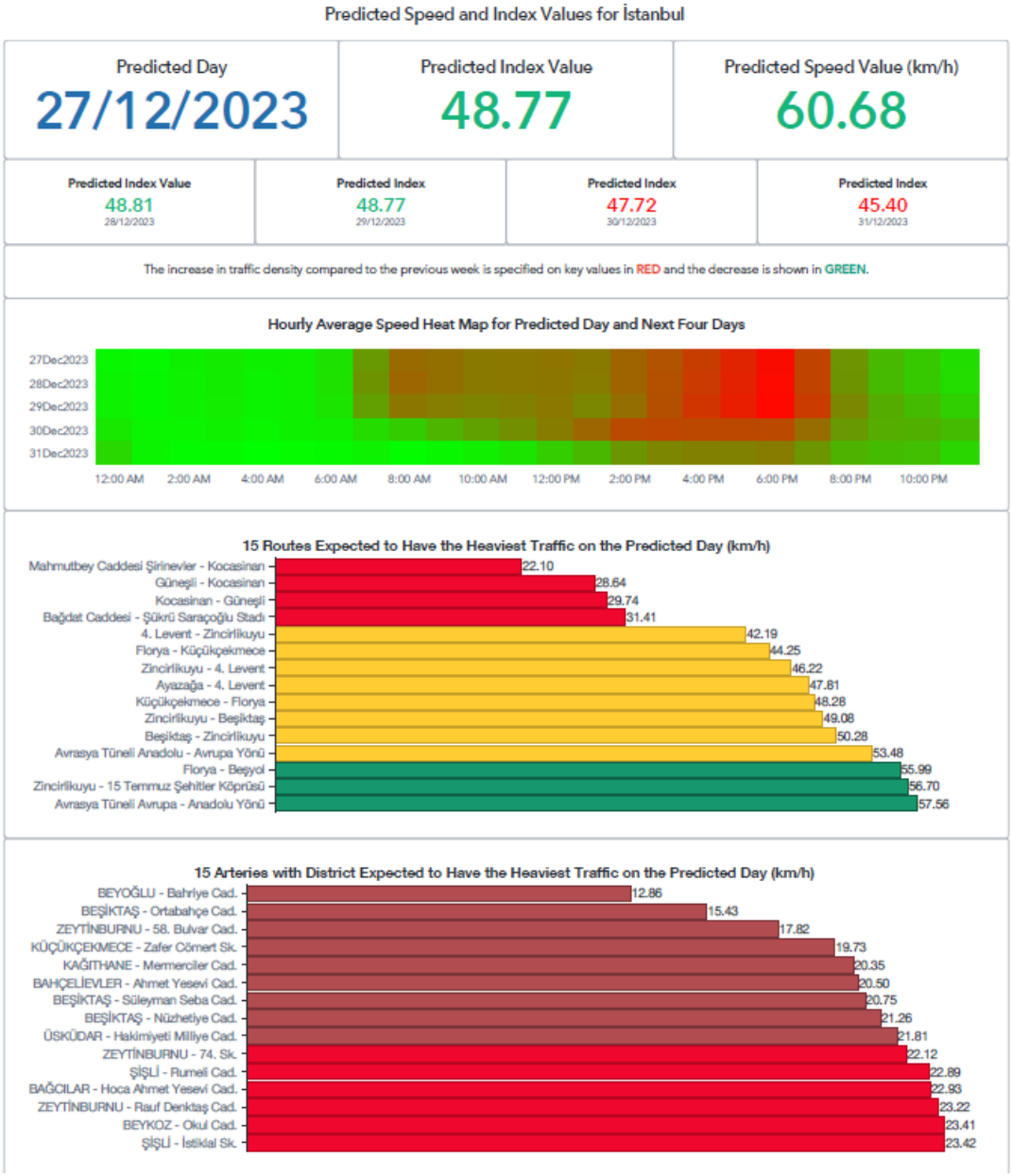


Figure 23. Traffic Density Forecasting – Speed and Index Prediction

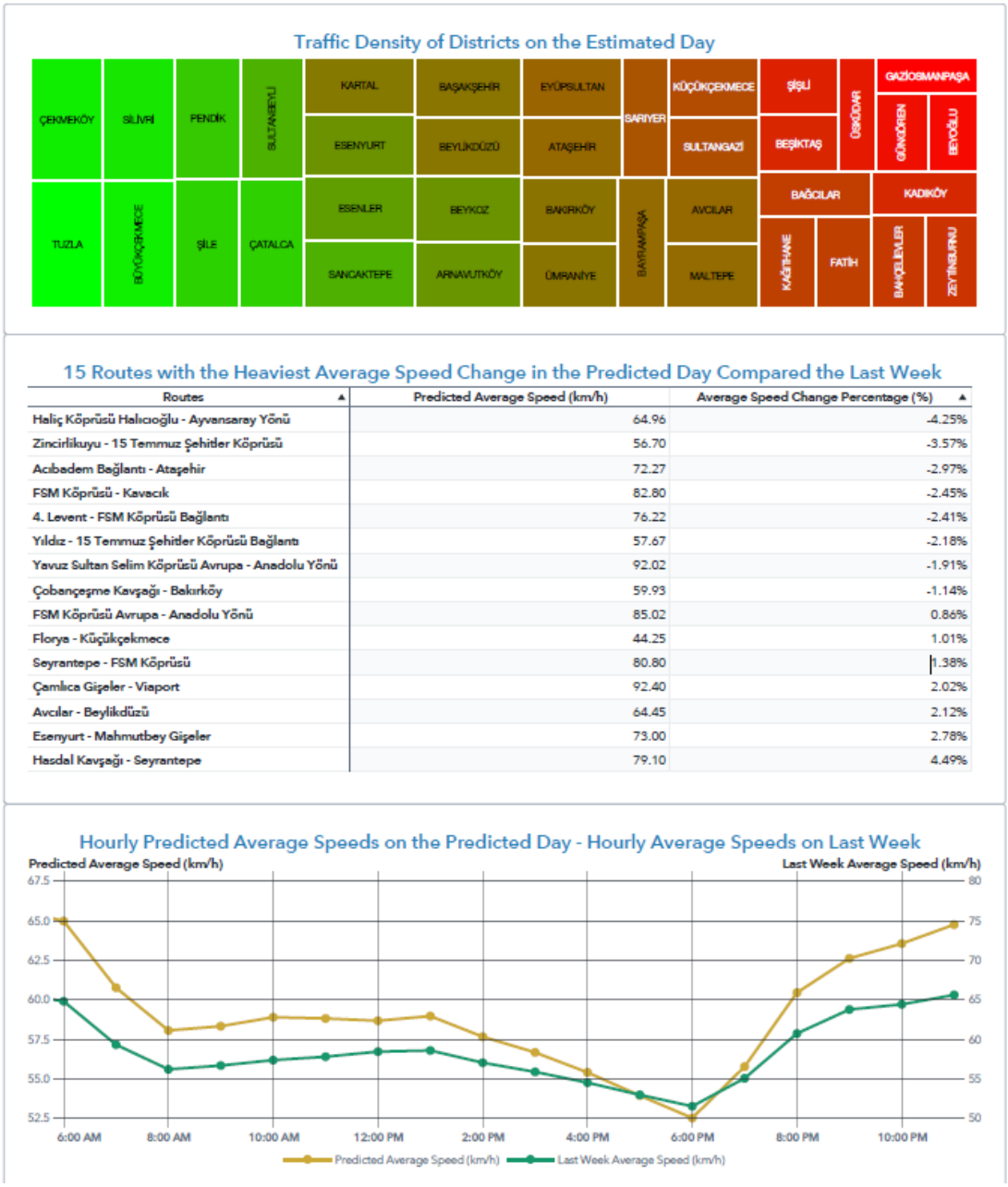


Figure 24. District and route based traffic density forecast

The Traffic Intensity Forecasting case study (Figure 23, 24) focused on predicting traffic flow and density to mitigate congestion and optimize urban mobility. By integrating sensor data and advanced modeling techniques, it demonstrates the municipality's commitment to leveraging predictive analytics for sustainable transportation planning and management.

Chapter 5

Conclusions

The conclusion of this study encapsulates the transformative role of big data and data analytics in the Istanbul Metropolitan Municipality (IMM). Prior to the establishment of a big data platform, the municipality's capability was limited to descriptive analysis using traditional data warehouse platforms. The advent of big data and analytics infrastructure has markedly enhanced the municipality's ability to integrate datasets from various directorates, subsidiaries, and affiliates. This integration has facilitated the application of machine learning algorithms, enabling not only future forecasting but also the optimization of data for strategic decision-making and the development of sustainable city management services.

Since its inception in 2019, the big data platform has been instrumental in bolstering IMM's service management through diverse analytical tools, contributing to detailed decision support reports, and laying the groundwork for a central IoT platform and the digital twin of the city. The platform's evolution has been marked by the creation of a multitude of artificial intelligence and machine learning services tailored to the municipality's operational needs. With continual enhancements in digitalization, new applications and potential usage scenarios are being developed, demonstrating the dynamic nature of the data ecosystem within IMM. More than thirty scenarios across various domains such as pricing, parking, transportation, and social media analysis have been proposed, reflecting the comprehensive impact of the big data initiative.

In a systematic and methodological approach, the study highlighted case studies focusing on Istanbul Card segmentation and ALO153 Call Center sentiment analysis, which demonstrated the value of data preprocessing. The former employed a behavioral segmentation model using Recency, Frequency, Monetary (RFM) analysis, while the latter utilized TF-IDF embeddings and the SadedeGel Library for textual data analysis. These applications underscore the importance of a data-driven decision-making paradigm, where insights from data analytics directly inform strategic recommendations to enhance service provision and citizen satisfaction.

The study's final reflections stress the significance of continuous improvement and stakeholder engagement in advancing data science and analytics capabilities

within IMM. The feedback loop and adaptation to emerging data science techniques and technologies are essential for maintaining the momentum in informed decision-making and the pursuit of sustainable urban planning and service optimization.

This expanded discussion and conclusion provide a holistic view of the study's findings, underscoring the critical role .underscoring the critical role of innovative data management strategies in modern urban governance. The empirical evidence obtained from the analytical case studies within the IMM's big data project substantiates the practical benefits of such strategies. The Istanbul Card segmentation analysis, with its incorporation of RFM analysis and over 100 variables, and the ALO153 Call Center sentiment analysis, using TF-IDF embeddings and the SadedeGel Library, are testaments to the power of data in refining service delivery and fostering a customer-centric approach.

The empirical evidence obtained from the analytical case studies within the IMM's big data project substantiates the practical benefits of such strategies. Notably, the Traffic Intensity Forecasting Case Study provides a compelling demonstration of the power of predictive analytics in addressing urban transportation challenges, a significant concern for metropolitan cities like Istanbul. This case study aimed to mitigate traffic congestion by developing predictive models for traffic flow and density, thereby enhancing the quality of life for citizens by reducing travel times and carbon emissions.

The Traffic Intensity Forecasting Case Study utilized an array of data sources, including sensor data, smart fleet management data, traffic light data, and external data sources like mobile application data. By cleaning and enriching this data, the study created structured datasets for analysis and feature selection. The use of descriptive analytics to understand patterns in the data was crucial for feature selection and model development. The study employed various modeling techniques, such as Decision Trees and Support Vector Machines, to construct models with the highest accuracy. This meticulous process involved designing a test and evaluation methodology, running alternative models on test datasets, and ranking the models according to predefined criteria, including accuracy and ease of interpretation.

The datasets in the Traffic Intensity Forecasting Case Study included the average hourly speed value of each road and segment, segment direction, vehicle count,

meteorological data, and other factors affecting traffic flow. The study utilized hierarchical models, panel series, neural networks, and seasonal models for forecasting, with ARIMAX and ESM models being pivotal during the model development phase. The MAPE criterion was used to select the best model, with the Seasonal Model achieving the lowest MAPE value, indicating high predictive accuracy.

The successful prediction of traffic density and speed on the D100 highway during the school opening day, with 98.2% accuracy, highlights the efficacy of the models developed. This capability to predict traffic conditions days in advance allows for proactive measures to manage congestion and optimize public transportation, illustrating IMM's innovative use of data analytics in creating a more sustainable and efficient urban environment

The integration of the Traffic Intensity Forecasting Case Study into the overarching narrative of IMM's big data initiatives showcases the municipality's commitment to leveraging advanced analytics for urban planning and management. The study's outcomes underscore the potential for predictive analytics to contribute significantly to sustainable city management, enhancing the strategic decision-making process within IMM. The incorporation of this case study into the conclusion of the thesis amplifies the comprehensive nature of the big data platform's impact, illustrating its role in enhancing IMM's services and providing a blueprint for other metropolitan municipalities aiming to harness the power of big data for urban management.

The findings suggest that the introduction of the big data platform has significantly increased the municipality’s capacity for predictive and prescriptive analytics. This capability has proved indispensable for intelligent and sustainable city management, evident in the enhanced precision of sentiment classification, as demonstrated by the performance metrics. The application of balanced data distribution yielded improvements in model accuracy and the F1 score, particularly in the positive class of sentiment analysis, showcasing the effectiveness of the methodologies implemented.

However, the study also acknowledges certain limitations. The focus on three specific application areas might not encompass the entire spectrum of urban management, suggesting the need for future research to cover a wider range of

applications. Additionally, the generalizability of the algorithms and techniques to other city conditions and cultural contexts remains a challenge. Moreover, the rapid pace of technological evolution necessitates continuous updates to the system and methodologies to remain relevant and effective.

In conclusion, the integration of big data and analytics into IMM's operations has facilitated a more informed and proactive approach to city management. The successful implementation of the big data platform has enabled the transformation of large volumes of data into actionable insights, thereby significantly contributing to the strategic goals of sustainable and integrated city management. It is recommended that IMM continues to evolve its data science capabilities, keeping pace with technological advancements and expanding the scope of its analytics applications. Further studies could explore the transferability of IMM's approach to other metropolitan contexts, thereby reinforcing the global discourse on sustainable urban management in the age of big data.

REFERENCES

Batty, M. (2013). Big Data, Smart Cities and City Planning. Dialogues in Human Geopraphy, 3(3), 274-279

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., ... & Portugali, Y. (2012). Smart cities of the future. *The European Physical Journal Special Topics*, *214*, 481-518.

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., & Portugali, Y. (2012). Smart Cities of the Future. The European Physical Journal Special Topics, 214(1), 481-518

Beatley, T. (2000). Preserving biodiversity: Challenges for planners. *Journal of the American Planning Association*, *66*(1), 5-20.

Beatley, T. (2012). Green Cities of Europe: Global Lessons on Green Urbanism. Island Press

Betsill, M. M., & Bulkeley, H. (2006). Cities and the multilevel governance of global climate change. *Global governance*, *12*, 141.

Bishop, C. (2006). Pattern recognition and machine learning. *Springer google schola*, *2*, 35-42.

Brenno C. Menezes, Jeffrey D. Kelly, Adriano G. Leal, Galo C. Le Roux, (2019). Predictive, Prescriptive and Detective Analytics for Smart Manufacturing in the Information Age.

Buja, A., Cook, D., & Swayne, D. F. (1996). Interactive high-dimensional data visualization. *Journal of computational and graphical statistics*, *5*(1), 78-99.

Bulkeley, H. (2010). Cities and the governing of climate change. *Annual review of environment and resources*, *35*, 229-253.

Caragliu, A., Del Bo, C., & Nijkamp, P. (2011). Smart Cities in Europe. Journal of Urban Technology, 18(2), 65-82

Caragliu, A., Del Bo, C., & Nijkamp, P. (2013). Smart cities in Europe. In *Creating Smart-er Cities* (pp. 65-82). Routledge.

Chatfield, D. C., Kim, J. G., Harrison, T. P., & Hayya, J. C. (2004). The bullwhip effect—impact of stochastic lead time, information quality, and information sharing: a simulation study. *Production and operations management*, *13*(4), 340-353.

Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information sciences*, *275*, 314-347.

Chen, C., Zhang, C., & Wei, L. (2015). Urban Big Data and Sustainable Development Goals: A Case Study in China Habitat International, 45, 60-67.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: Form Big Data to Big Impact. MIS Quarterly, 36(4), 1165-1188

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile networks and applications*, *19*, 171-209.

Davenport, T. H., Harris, J. G., & Morison, R. (2010). *Analytics at work: Smarter decisions, better results*. Harvard Business Press

Davenport, T., & Harris, J. (2017). *Competing on analytics: Updated, with a new introduction: The new science of winning*. Harvard Business Press.

Deakin, M., & Allwinkle, S. (2007). Urban Regeneration and Sustainable Communities: Role of Networks, Innovation, and Creativity in Building Successful Partnerships. Journal of Urban Technology.

Demirkan, H., & Delen, D. (2013). Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. *Decision Support Systems*, *55*(1), 412-421.

Evans, J. P. (2012). Resilience, Ecology and Adaptation in the Experimental City. Transactions of the Institute of British Geopraphers, 37(1), 77-90.

Giffinger, R., Fertner, C., Kramar, H., & Meijers, E. (2007). City-ranking of European medium-sized cities. *Cent. Reg. Sci. Vienna UT*, *9*(1), 1-12.

Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler-Milanovic, N., & Meijers, E. J. (2007). Smart cities. Ranking of European medium-sized cities. Final Report.

Girardet, H. (1999). Creating sustainable cities.

Hair, J. F., LDS Gabriel, M., Silva, D. D., & Braga, S. (2019). Development and validation of attitudes measurement scales: fundamental and practical aspects. *RAUSP Management Journal*, *54*, 490-507.

Han, J., Pei, J., & Tong, H. (2022). *Data mining: concepts and techniques*. Morgan Kaufmann.

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *ACM sigmod record*, *29*(2), 1-12.

Hassanzadeh, H. R., & Wang, M. D. (2016, December). DeeperBind: Enhancing prediction of sequence specificities of DNA binding proteins. In *2016 IEEE International conference on bioinformatics and biomedicine (BIBM)* (pp. 178-183). IEEE.

Katerina Lepenioti, Alexandros Bousdekis, Dimitris Apostolou, Gregoris Mentzas.(2020). Prescriptive analytics: Literature review and research challenges

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big data & society*, *1*(1), 2053951714528481.

Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures and their consequences. *The Data Revolution*, 1-240. Beatley, T. (2000). Preserving biodiversity: Challenges for planners. *Journal of the American Planning Association*, *66*(1), 5-20.

Kitchin, R. (2014). The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. SAGE Publication

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, *18*, 140-181.

Koops Liliy (2020). Optimized Maintenance Decision-Making – A Simulation-supported Prescriptive Analytics Approach based on Probabilistic Cost-Benefit Analysis

Lafferty, W. M., & Langhelle, O. (1999). Institutional Design for Sustainable Development. Natural Reseources Forum, 23(3), 193-204

Lejano, R. P., & Stokols, D. (2013). Social ecology, sustainability, and economics. *Ecological economics*, *89*, 1-6.

Li, Y., Wang, D., Xiao, H., & Zhang, J. (2017). Sustainable urban management through big data and data analytics. Procedia Computer Science, 122, 469-476.

Lu, Y., Zhou, Y., van den Brink, P. J., & Yuan, J. (2019). Big Data for Global Urbanization: Challenges and Opportunities of Urban Informatics. Science of the Total Environment, 652, 1055-1069.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Xml retrieval. *Introduction to Information Retrieval*.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). Big data: The next frontier for innovation, competition, and productivity.

Marx, S., Weber, E. U., Orlove, B. S., Leiserowitz, A., Krantz, D. H., Roncoli, C., & Mastrangelo, B. (2007). Communication and Mental Processes: Experiential and Analytic Processing of Uncertain Climate Information. Global Environmental Change, 17(1)47-58

Mayer-Schönberger, V., & Cukier, K. (2013). Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

n1pqtrfbcgolcyn13lwsv8hr66tl.htm

Neirotti, P., De Marco, A., Cagliano, A. C., Mangano, G., & Scorrano, F. (2014). Current Trends in Smart City Initiatives: Some Stylised Facts. Cities, 38 25-36

Pacione, M. (2005). Urban Environmental Quality and Human Well-being- A Social Goegraphical Perspective. Landscape and Urban Planning, 74(1), 1-19

Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big data*, *1*(1), 51-59.

Raghupathi, W., & Raghupathi, V. (2014). Big Data Analytics in Healtcare: Promise and Potential. Health Information Science and Systems, 2(1), 3

Robinson, M. (2006). Budget analysis and policy advocacy: The role of non-governmental public action.

Siemens, G., & Baker, R. S. (2012). Learning Analytics and Educational Data Mining: An Overview. Journal of Educational Technology & Society, 15(3), 167-169

Stefani, Karolin, Zschech, Patrick. (2018). Constituent Elements For Prescriptive Analytics Systems.

Townsend, A. M. (2013). *Smart cities: Big data, civic hackers, and the quest for a new utopia*. WW Norton & Company.

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2, pp. 131-160).

URL 1 - ZDNet-TechRepublic, (2019) How to win with prescriptive analytics.

URL 10 - https://www.huawei.com/en/huaweitech/publication/77/big-results-from-big-data

URL 11 - https://www.oracle.com/database/

URL 12 - https://learn.microsoft.com/en-us/sql/relational-databases/databases/databases?view=sql-server-ver16

URL 13 - https://neo4j.com/

URL 14 - https://allegrograph.com/

URL 15 - https://en.wikipedia.org/wiki/InfiniteGraph

URL 16 - https://www.oracle.com/tr/database/graph/

URL 17 - https://www.vertica.com/

URL 18 - https://en.wikipedia.org/wiki/Aster_Data_Systems

URL 19 - https://greenplum.org/

URL 2 - https://sadedegel.ai/

URL 20 - https://www.actian.com

URL 21 - https://www.3ds.com/nuodb-distributed-sql-database

URL 22 - https://www.usenix.org/conference/osdi23/presentation/mehdi

URL 23 - https://www.kinetica.com/

URL 24 - https://www.mapd.com/

URL 25 - https://hadoop.apache.org/

URL 26 - https://hbase.apache.org/

URL 27 - https://www.mysql.com/

URL 28 - https://en.wikipedia.org/wiki/MapReduce

URL 29 - https://hive.apache.org/

URL 3 - https://www.cloudera.com/

URL 30 - https://pig.apache.org/

URL 31 - https://mahout.apache.org/

URL 32 - https://mahout.apache.org/

URL 33 - https://sqoop.apache.org/

URL 34 - https://spark.apache.org/

URL 35 - https://oozie.apache.org/

URL 36 - https://zookeeper.apache.org/

URL 37- https://flume.apache.org/

URL 38 - https://chukwa.apache.org/

URL 39 - http://prestodb.io/

URL 4 - https://en.wikipedia.org/wiki/Hortonworks

URL 40 - https://nifi.apache.org/

URL 41 - https://www.striim.com/

URL 42 - https://www.talend.com/

URL 43 - https://kafka.apache.org/

URL 44 - https://www.1010data.com/

URL 45 - https://ignitetech.com/softwarelibrary/infobrightdb

URL 46 - https://greenplum.org/

URL 47 - https://ravendb.net/articles/ravendb-and-multi-region-setup

URL 48 - https://www.spotfire.com

URL 49 - https://www.sas.com

URL 5 - https://en.wikipedia.org/wiki/MapR

URL 50 - https://altair.com/altair-rapidminer

URL 51 - https://www.cs.waikato.ac.nz/~ml/weka/

URL 52 - https://www.predixionsoftware.com/

URL 53 - https://www.r-project.org/

URL 54 - https://spark.apache.org/docs/latest/ml-guide.html

URL 55 - https://www.ibm.com/products/spss-statistics

URL 56 - https://www.ibm.com/docs/en/ias?topic=dsed-tutorial-getting-started-data-science-experience-integrated-analytics-system

URL 57 - https://www.tableau.com/

URL 58 - https://www.oracle.com/tr/business-analytics/business-intelligence/technologies/bi-enterprise-edition.html

URL 59 - https://www.microstrategy.com/

URL 6 - https://www.oracle.com/database/technologies/bigdata-appliance.html

URL 60 - https://www.treasuredata.com/w

URL 61 - https://www.qlik.com/us

URL 62 - https://bigml.com/

URL 63 - https://en.wikipedia.org/wiki/Lavastorm_Analytics

URL 64 - https://www.ibm.com/products/cognos-analytics

URL 65 - https://www.lexalytics.com/

URL 66 - https://en.wikipedia.org/wiki/Attensity

URL 67 - https://sproutsocial.com/

URL 68 - https://documentation.sas.com/doc/en/vfcdc/8.4/vfug/

URL 7 - https://www.ibm.com/analytics/big-data-analytics

URL 8 - https://www.teradata.com/solutions/big-data

URL 9 - https://support.pentaho.com/hc/en-us/articles/360002913871-Big-Data-and-Pentaho

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & Data, M. (2005, June). Practical machine learning tools and techniques. In *Data mining* (Vol. 2, No. 4, pp. 403-413). Amsterdam, The Netherlands: Elsevier.

Zanella, A., Bui, N., Castellani, A., Vangelista, L., & Zorzi, M. (2014). Internet of things for smart cities. *IEEE Internet of Things journal*, *1*(1), 22-32.

Zheng, Y., Zhang, L., Xie, X., & Ma, W. Y. (2015). Mining Interesting Locations and Travel Sequences from GPS Trajectories. Journal of Visualized Experiments, (97), e52544.

Zhou, K., Fu, C., & Yang, S. (2016). Big data driven smart energy management: From big data to big insights. *Renewable and sustainable energy reviews*, *56*, 215-225.

Zikopoulos, P., Eaton, C., Deroos, D., Deutsch, T., & Lapis, G. (2012). Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill Osborne Media.