

DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**CLOUD-BASED DATA MINING TOOL FOR
BIG DATA**



by
Gamze ÖZÇELİK

June, 2019
İZMİR

CLOUD-BASED DATA MINING TOOL FOR BIG DATA

**A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master of Sciences
in Computer Engineering**

by


Gamze ÖZÇELİK

June, 2019


İZMİR

M.Sc THESIS EXAMINATION RESULT FORM


We have read the thesis entitled “**CLOUD-BASED DATA MINING TOOL FOR BIG DATA**” completed by **GAMZE ÖZÇELİK** under supervision of **PROF. DR. ALP KUT** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.


Prof. Dr. Alp KUT

Supervisor


Assoc. Prof. Dr. Derya BİRANT

(Jury Member)


Asst. Prof. Dr. Özgü GAN

(Jury Member)


Prof. Dr. Kadriye ERTEKİN

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I thank to my advisor Prof. Dr. Alp Kut for his support, patience, guidance and knowledge.

I present my thanks to İsmail Yürek who introduced me to Apache Spark.

I also thank my family for their support and tolerance.

Gamze ÖZÇELİK



CLOUD-BASED DATA MINING TOOL FOR BIG DATA

ABSTRACT

Nowadays, in parallel with the increase in the amount of data, the need for data mining has increased and the process of converting data into information has become inevitable in all areas of life. Depending on the fact that the data is kept digitally, there is a large amount of data waiting to be converted into information. With this study, it is aimed that the data mining process, which has become such a general need nowadays, is easily realized by non-experts. Within the scope of the study, "Sparkle Mining" tool has been developed which enables analysis with good performance in big data.

Within the scope of the project, an alternative tool for data analysis has been developed. Before this tool was developed, the technology and libraries to be used were investigated extensively. In data analysis, the technologies and data mining library that will enable us to achieve the best performance have been used by our system. The studies conducted in order to determine the data mining library to be used are also presented within the scope of the thesis.

The data set loaded into the system is analyzed according to the rules set by the user. Data and analysis results are stored in the cloud area defined for the user. With this study, users will be able to do their analysis easily also they will be able to view the data and the results of their analysis at any time and place.

Keywords: Apache Spark, azure, data mining, cloud computing

BÜYÜK VERİ İÇİN BULUT TABANLI VERİ MADENCİLİĞİ ARACI

ÖZ

Günümüzde veri miktarındaki artışa paralel olarak veri madenciliğine duyulan ihtiyaç da artmış, veriyi bilgiye dönüştürme operasyonu hayatın her alanında kaçınılmaz hale gelmiştir. Verilerin dijital ortamda tutuluyor olmasına bağlı olarak, bilgiye dönüştürülmeyi bekleyen çok miktarda veri bulunmaktadır. Bu çalışma ile günümüzde bu kadar genel ihtiyaç haline gelen veri madenciliği sürecinin, uzman olmayan kişiler tarafından da kolaylıkla yapılıyor olması hedeflenmiştir. Çalışma kapsamında büyük veriler üzerinde performanslı analiz yapılmasına imkan sağlayan ‘Sparkle Mining’ aracı geliştirilmiştir.

Proje kapsamında veri analizinde alternatif olarak kullanılabilecek bir araç geliştirilmiştir. Bu araç geliştirilmeden önce, kullanılacak teknoloji ve kütüphaneler kapsamlı olarak araştırılmıştır. Veri analizinde en iyi performansı yakalamamızı sağlayacak teknolojiler ve veri madenciliği kütüphanesi sistemimiz tarafından kullanılmıştır. Kullanılacak veri madenciliği kütüphanesini belirlemek amacıyla yapılan çalışmalar da tez kapsamında sunulmuştur.

Projede belli formata getirilmiş veri seti, web uygulama aracılığıyla ile buluta yüklenmektedir. Sisteme yüklenen veri seti kullanıcının belirttiği kural setlerine göre analiz edilmektedir. Veri setleri ve analiz sonuçları kullanıcı için tanımlanan bulut alanında saklanmaktadır. Bu çalışma ile veri madenciliğinde uzman kişilerin yapabildiği analizleri, kullanıcılar kolaylıkla yapabilecek, bununla birlikte verilerine ve elde ettikleri analiz sonuçlarına istedikleri yer ve zamanda ulaşabileceklerdir.

Anahtar kelimeler: Apache Spark, azure, veri madenciliği, bulut bilişim

CONTENTS

	Page
M.Sc THESIS EXAMINATION RESULT FORM.....	ii
ACKNOWLEDGEMENTS.....	iii
ABSTRACT.....	iv
ÖZ.....	v
LIST OF FIGURES.....	x
LIST OF TABLES.....	xi
 CHAPTER ONE - INTRODUCTION.....	 1
1.1 General	1
1.2 Purpose	2
1.3 Organization of the Thesis	2
 CHAPTER TWO - LITERATURE REVIEW	 4
2.1 Cloud Computing Studies	4
2.2 Apache Spark Studies	4
2.3 Data Mining Tools and Data Mining Technique Studies	5
2.4 Data Mining Studies in the Cloud Computing	6
 CHAPTER THREE - CLOUD COMPUTING.....	 8
3.1 Deployment Models	9
3.1.1 Public Cloud	9
3.1.2 Private Cloud	9
3.1.3 Hybrid Cloud	9
3.1.4 Community Cloud	9
3.2 Service Models	10
3.2.1 Infrastructure as a service (IaaS)	10

3.2.2 Platform as a service (PaaS).....	10
3.2.3 Software as a Service (SaaS).....	10
3.3 Benefits of Cloud Computing	11
3.4 Disadvantages of Cloud Computing	11
CHAPTER FOUR – DATA MINING PROCESS.....	12
4.1 Data Preparation	12
4.1.1 Data selection	12
4.1.2 Data pre-processing	13
4.1.3 Transformation	13
4.2 Data Mining	13
4.2.1 Association Rule Mining	13
4.2.2 Clustering	14
4.3 Interpretation Evaluation	15
CHAPTER FIVE – DATA MINING TOOLS.....	16
5.1 Weka	16
5.2 Apache Spark	16
5.2.1 Apache Spark Use Cases	17
5.2.1.1 Spark Streaming	17
5.2.1.2 Machine Learning	18
5.2.1.3 Data integration	18
5.2.1.4 Interactive analytics	18
5.2.2 Apache Spark Usage Techniques	18
5.2.3 Apache Spark Architecture... ..	19
5.2.3.1 Spark Core	20
5.2.3.2 Spark SQL	20
5.2.3.3 Spark Streaming	20
5.2.3.4 MLlib	20
5.2.3.5 GraphX	20
5.2.3.6 Spark R	20

5.2.4 Resilient Distributed Datasets (RDDs).....	21
CHAPTER SIX - WEKA / SPARK BIG DATA ANALYSIS.....	22
6.1 Data Preparation	22
6.2 Data Mining	23
6.2.1 K-Means Algorithm	23
6.2.2 Fp-Growth Algorithm	24
6.3 Weka / Spark Comparison and Evaluation	24
6.3.1 K-Means Algorithm Results	24
6.3.2 Fp-Growth Algorithm Results	27
CHAPTER SEVEN - THE PROPOSED PROJECT.....	29
7.1 Sign up and Login	29
7.2 Data Upload on Web Application	33
7.3 Defining Data Mining Rule	33
7.4 Data Mining	36
7.5 Representation of Analysis Results	37
CHAPTER EIGHT - USED TECHNOLOGIES.....	39
8.1 The Proposed Project Technology Architecture	39
8.2 ASP.NET	40
8.2.1 ASP.NET MVC	40
8.3 Spring Framework	41
8.3.1 Spring Boot	42
8.4 Azure Storage	42
8.4.1 Azure Storage Naming Rules	43
8.4.2 Azure Storage Tables	44
8.4.3 Azure Blobs	45

CHAPTER NINE - CONCLUSION AND FUTURE WORK.....46

9.1 Conclusion 46

9.2 Future Work 46

REFERENCES.....48



LIST OF FIGURES

	Page
Figure 3.1 Cloud computing	8
Figure 4.1 Data mining process steps	12
Figure 5.1 Apache Spark architecture	19
Figure 6.1 K-Means algorithm operation time	25
Figure 6.2 K-Means algorithm sum of squared error	26
Figure 6.3 Customer clustering result	26
Figure 6.4 Spark F-Growth algorithm operation time	27
Figure 6.5 Spark F-Growth algorithm number of rules	28
Figure 7.1 The proposed project flow diagram	29
Figure 7.2 Sign up page	30
Figure 7.3 Login page	31
Figure 7.4 My data page	32
Figure 7.5 Cloud url save popup	33
Figure 7.6 Defining data mining rule popup	34
Figure 7.7 Defining association rule popup	35
Figure 7.8 State of data mining process page	36
Figure 7.9 Analysis result page for first data set	37
Figure 7.10 Analysis result page for second data set	38
Figure 8.1 Proposed project technology architecture	39
Figure 8.2 Azure storage on azure storage explorer	43
Figure 8.3 Azure storage tables diagram	44

LIST OF TABLES

	Page
Table 6.1 Characteristics of data set	23
Table 8.1 Azure storage naming rules	44



CHAPTER ONE

INTRODUCTION

1.1 General

Nowadays, with the fact that information systems play an active role in almost every area of life, there have been serious increases in the data sizes. With this large increase in data size, data mining has also evolved naturally and still continues to evolve. Nowadays, the interest in this field has increased with the fact that popular commercial companies and public security institutions make serious investments.

Data mining is a set of statistical and mathematical methods used to obtain hidden patterns / models in large data systems. These methods play a very important role in the decision-making process. Nowadays, the data mining that has become indispensable has become more difficult with the increase in data sizes. Existing systems are not enough to encourage people who want to discover information to analyze large data and find new methods.

In the world of software, it is thought that the knowledge discovery process can be performed with better performance if the technologies that provide performance in the calculation methods are used in data mining process. In this direction, technological developments such as distributed systems and cloud technology have been used in conjunction with data mining techniques. In data mining processes, a lot of studies have been done using cloud technology and distributed systems and new ones are added to these studies every day.

1.2 Purpose

A cloud-based, dynamic data mining tool has been developed that can be easily used by people who are not experts in data mining. Our goal is to be able to address people who do not have much knowledge about data mining, so a web application with very simple interface is designed. By developing a dynamic system, it is ensured that the data set of each sector can be easily analyzed.

The performance problem of the data mining process is considered at every step of application development. For this purpose, research has been done and the most appropriate technologies have been tried to be used to improve performance. The results of the study comparing the accuracy rates and performances of the Weka, Spark libraries are presented and the reason why Spark is used is given.

Cloud was used as storage to ensure the easy accessibility of data sets and discovered information. In this way, the support of cloud computing was taken at both easy accessibility and speed points.

1.3 Organization of the Thesis

This thesis contains nine chapters and this thesis is organized as follows.

In Chapter 2, the related literature and previous studies about data mining, cloud computing, Apache Spark and Weka are summarized.

In Chapter 3, cloud computing technology are described.

In Chapter 4, data mining process and techniques are described.

In Chapter 5, information about Weka and Spark tools are given.

In Chapter 6, data mining operations with Weka, Spark tools are summarized and compared their results.

In Chapter 7, information about the developed application has been given and explained how to use the application.

In Chapter 8, technological architecture and technologies are summarized.

In Chapter 9, the conclusion and future works are described.



CHAPTER TWO

LITERATURE REVIEW

We have briefly explained the works that guide the decisions we made while developing our thesis. These are the studies that are successful in cloud computing, data mining, Weka and Spark.

2.1 Cloud Computing Studies

Györödi, Pavel, Györödi & Zmaranda (2017) aim to answer the question of why the cloud is being used. For this purpose, they compared the response times of on premises databases and cloud databases. Microsoft Azure is selected because it allows the use of the online portal for the management of the services offered from the cloud options. In order to compare the test results, a special test architecture was established and this architecture was also presented. At the end of the study, it is determined that cloud computing provides the best solution in terms of usability, security, scalability, performance, but its cost is higher than on premises solutions. However, it was stated that the cost could be ignored considering that there should be people in the on premises solutions who would be interested in the servers.

Srivastava & Khan (2018) have examined many articles about cloud computing and provided information about cloud computing. The types, advantages, services and evolution of cloud computing are emphasized. It is aimed to show the situation of IT sector before and after cloud computing as the output of the study.

2.2 Apache Spark Studies

Researchers in their work aimed to determine the risks of miscarriage of pregnant women and inform them via mobile devices. The data set used is collected via mobile devices. While age, body mass index and abortion rate were taken from the users,

acceleration and GPS information are collected from smart phone sensors and clustering was performed with Apache Spark. Using the K-Means algorithm, women are divided into three groups according to their risk status. Women at risk of miscarriage have informed via mobile devices (Asri, Mousannif & Moatassime, 2017).

Researchers within the scope of the study, have run the Eclat algorithm on four separate data sets using Apache Spark. In the study, the Eclat algorithm was run with different minimum support values and speed evaluation was performed. This study is an important study since it is the first study that has been encountered related to the Eclat algorithm, although studies have been conducted using Spark's Fp-Growth and Apriori algorithms. (Mohamed, Abdel-fattah & El-Gaber, 2017)

Rathee, Kaul & Kashyap (2015) propose a parallel Apriori algorithm running on Spark. The R-Apriori algorithm was developed by utilizing Spark's in-memory, distributed platform. The proposed algorithm was found to be highly effectiveness, efficiency and scalability according to the standard Apriori algorithm according to many experiments performed with different data sets.

2.3 Data Mining Tools and Data Mining Technique Studies

Naive Bayes algorithm has been run with iris data set on RapidMiner, Orange and Weka tools. The accuracy rates of the results of three different data mining tools were compared. As a result of the study, it was determined that the error rate of the results produced by Weka tool was the lowest. In our study, Spark has been found to work with less error than Weka and in this direction, the Spark which has a good performance, was analyzed with less error than Weka, Orange and RapidMiner. (Ahmed, 2017)

The performance of classification algorithms in preventing the infecting of malware to the computer system were examined. This study aims to identify weaknesses in existing algorithms and to help them to create more efficient algorithms. Within the scope of the study, J45, LMT, Naïve Bayes, Random Forest, RBF Network, MLP Classifier, Random Tree, REP Tree, SimpleLogistic, IBK, LWL, Bagging, AdaBoost, KStar, SVM algorithms were examined. Weka is used in the execution of algorithms. At the end of the experiments, it was found that Random Forest algorithm produces the highest accuracy rate when detecting malware (Dada, Bassi, Hurcha & Alkali, 2019).

Bharati & Ramageri (2010) focused on the practices of the institutions that applied data mining processes in their study. The problem definition of many important firms in the world has been made and then they have been informed about the results of data mining.

2.4 Data Mining Studies in the Cloud Computing

Zeng (2018) states that a lot of problems will be solved when the information discovery process on big data sets is done with the cloud computing platform in his study. Within the scope of the study, firstly information was given about cloud computing and data mining methods and then the application areas of data mining were mentioned. It is stated that it will be advantageous to use cloud computing with data mining for reasons such as low cost of cloud computing, strong data processing capability, consistent data system. It was emphasized that this study could guide future data mining studies.

Data mining in cloud computing is the acquisition of data on web resources as structured data. Integration of data mining with cloud computing enables infrastructure and storage costs to be reduced. The biggest problem of cloud computing is security, but the security of the cloud is better than the security provided by small firms

themselves. If cloud is to be used in data mining, companies should consider security. "DMCloud" provided ad data mining in cloud by Microsoft (Dhote & Deshpande, 2016).

Data mining infrastructure offered by cloud computing provides the user with information processing services ready. Thus, the user does not deal with the complex system that he / she stores and manages his data, user can only focus on the data mining process. In data mining with cloud computing, users only use the data mining algorithms they need and pay for it. It does not deal with complex processes. At the same time, the customer doesn't need hardware infrastructure because he can do data mining via browser (Chaudhari, 2015).

In big data analysis,

- Spark is used
- Comparing the performance of different data mining tools
- Comparing the results of different algorithms on the same data mining tool

although there are many studies, there is no study examine the performance or error rate of Spark and other data mining tools. Therefore, we have done a study comparing Weka and Spark tools in this thesis scope.

CHAPTER THREE

CLOUD COMPUTING

Cloud computing is an Internet-based system based on pay-as-you-use logic. Cloud computing provides infrastructure, software, platforms and devices to users through virtual servers. Cloud computing only rents the services to its customers and tenants pay for the time they use. Cloud computing users can better concentrate on their business without dealing with technical details. The speed of application development of customers not interested in infrastructure increases, while the cost may increase or decrease. However, when we think about the cost of the personnel who are interested in the infrastructure, it can be thought that cloud computing will be advantageous for the tenants in every situation. Figure 3.1 shows services and models offered to customers within the scope of cloud computing.

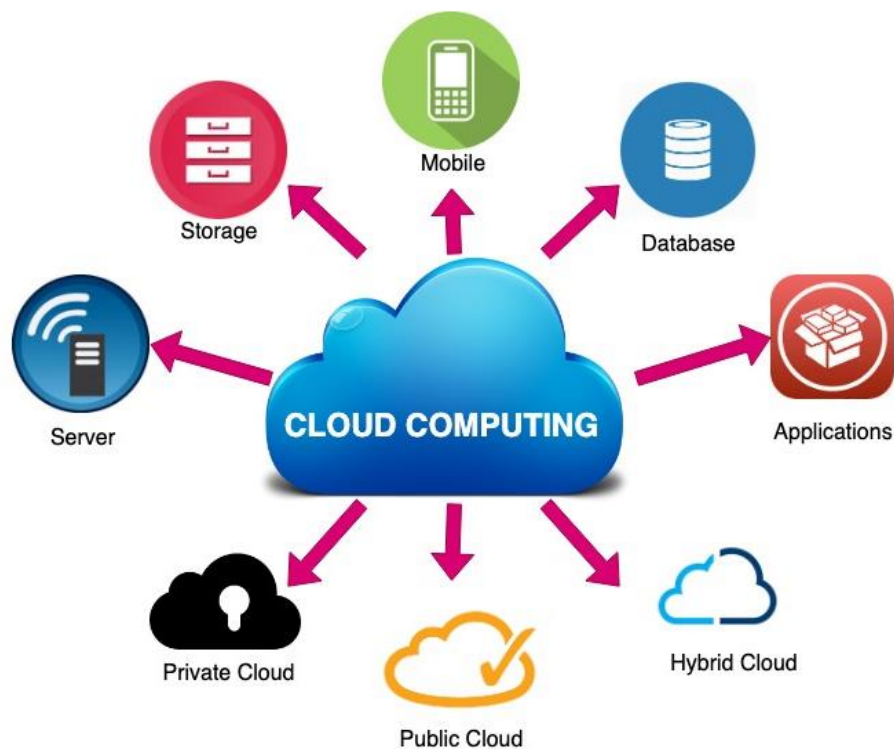


Figure 3.1 Cloud computing

3.1 Deployment Models

Distribution models are four types according to accessibility.

3.1.1 Public Cloud

The public cloud is the traditional cloud computing distributed model. This cloud is accessible to all customers via the Internet. Private cloud is more secure than public cloud because public cloud is accessible to all internet users. For example; azure.

3.1.2 Private Cloud

Private cloud is more secure than other cloud deployment models. It is prepared for an organization and the organization can manage all the resources it has in the cloud. Provides resources to people who determined by organization.

3.1.3 Hybrid Cloud

Hybrid cloud is the distribution model where general, private is used integrated. Where security is important, private cloud is used, while other areas use public cloud. Hybrid cloud is the solution according to our needs and cost.

3.1.4 Community Cloud

Organizations with similar needs can come together to create a community cloud. These organizations build the cloud infrastructure according to their needs. Cost is more expensive than the general cloud. But the community cloud is safer than the general cloud.

3.2 Service Models

Cloud computing's service models are three types.

3.2.1 Infrastructure as a service (IaaS)

IaaS users can easily access many computer resources via the Internet. Storage, operating system, network, hardware, storage device are examples of computer resources. These resources are rented to users and the user pays for the time he uses these resources. Google Compute Engine, Windows Azure, Amazon EC2 are examples for IaaS (Malik, Wani & Rashid, 2018).

3.2.2 Platform as a service (PaaS)

PaaS offers its users the application development and distribution platform as a service. Thus, users do not have to pay serious money to servers, power and equipment. They can use this service only with internet access. Force.com, Windows Azure, Google are examples for PaaS (Malik, Wani & Rashid, 2018).

3.2.3 Software as a Service (SaaS)

SaaS is a cloud service service that allows the user to use the software via the internet. Users can use the software without installing them on their computers. Thus, users do not have to deal with software purchase, maintenance, and managing updates. Microsoft Office 365 is example for SaaS (Malik, Wani & Rashid, 2018).

3.3 Benefits of Cloud Computing

- **Eco-friendly:** It provides 30% less energy use and carbon production than local server use.
- **Scalability and adjustable capacity:** Allows users to shape their consumption according to their needs.
- **Ease of access:** Cloud computing provides easy access to all types of resources via the Internet.
- **Low cost:** Cloud computing enables companies to meet the services they need via the Internet, rather than building their own service infrastructures. This means less cost for firms.
- **No fixed investment cost:** No need to purchase hardware. You can rent for the time required.
- **Flexibility and efficiency:** Capacity increase and reduction can be made upon request.

3.4 Disadvantages of Cloud Computing

- **Service continuity and availability:** In the event of a problem with the Cloud Services service providers, all companies receiving service from this service provider will be affected and become unable to serve their customers.
- **Data security and privacy:** The use of cloud computing services by many users at the same time has risks for data privacy and security.
- **Service provider dependency and data lockdown:** Data loss may occur due to malfunctions and attacks that may occur in a Cloud Computing service provider.
- **Remote access:** Given the weakness of remote access, cloud computing has a high security risk.
- **Data Transfer:** When applications are starting to use more intense data, it can make it difficult to migrate data from the user to the cloud computing.

CHAPTER FOUR

DATA MINING PROCESS

The discovery of the information on the data set we used was completed following the steps in the figure. Figure 4.1 shows steps of data mining process.

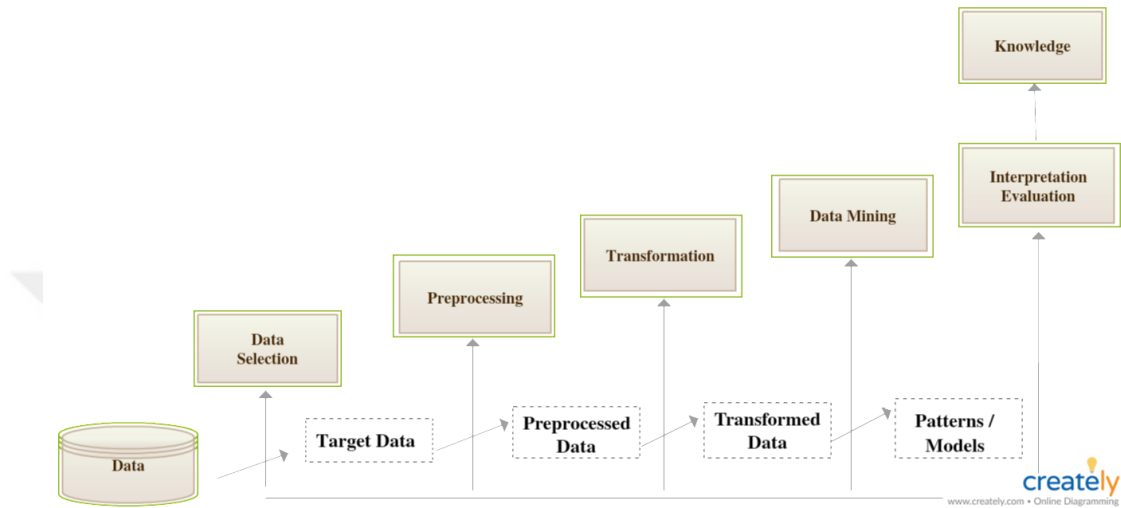


Figure 4.1 Data mining process steps

4.1 Data Preparation

Data preparation operation aims to improve the data that are incomplete, noisy and inconsistent. In order for data mining algorithms to be executed, the data is made ready for analysis by passing through certain stages. The preparation of the data includes data selection, data preprocessing and data conversion steps, as shown in the picture above.

4.1.1 Data selection

Determining the data to be needed in the analysis.

4.1.2 Data pre-processing

Completion of missing data, extraction of outliers and noisy data.

4.1.3 Transformation

Changing, discretizing and normalizing data formats.

4.2 Data Mining

The data mining stage is the operation of algorithms on the cleaned data set with the help of data mining tools. Hidden patterns and models are discovered as a result of data mining.

There are different data mining techniques such as Classification, Clustering, Association Rule Mining and Regression Analysis. We were interested in Clustering and Association Rule Mining within the scope of the thesis and summarized briefly about these techniques below.

4.2.1 Association Rules Mining

Association rules mining is the discovery of interesting relationship sets within the data set. If the obtained rules provide both the minimum support value and the confidence value, it is considered as the association rule. These minimum support and confidence values are expressed as formal definition in the following.

Let $\mathfrak{I} = \{i_1, i_2, \dots, i_m\}$ be a set of items. Let D , the task-relevant data, be a set of database transactions where each transaction T is a set of items such that $T \subseteq \mathfrak{I}$. Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is implication of the form $A \Rightarrow B$, where $A \subset \mathfrak{I}$, $B \subset \mathfrak{I}$, and $A \cap B = \emptyset$. The rule $A \Rightarrow B$

holds in the transaction set D with support s , where s is the percentage of transactions in D that contain $A \cup B$ (i.e., both A and B). This is taken to be the probability, $P(A \cup B)$. The rule $A \Rightarrow B$ has confidence c in the transaction set D if c is the percentage of transactions in D containing A that also contain B . This is taken to be the conditional probability, $P(B|A)$ (Györödi, Györödi & Holban, 2004, s.2).

4.1 equation indicates the support value and 4.2 equation shows confidence value.

$$\text{Support } (A \Rightarrow B) = P(A \cup B) \quad (4.1)$$

$$\text{Confidence } (A \Rightarrow B) = P(B | A) \quad (4.2)$$

4.2.2 Clustering

The clustering process is the collection of similar objects in the data set in the same group. A small number of data clusters can result in the loss of fine details (Rai & Singh, 2010).

Known traditional clustering techniques:

- Hierarchical Methods
- Partitioning Methods
- Density-Based Algorithms
- Grid Based Clustering

4.3 Interpretation Evaluation

After terminating the algorithm, the model is produced and this model needs to be interpreted. Algorithm results are interpreted together with the data set and the process of information discovery is completed. In this process, the excel can be used to visualize and interpret your analysis.



CHAPTER FIVE

DATA MINING TOOLS

In this section, our researches about the two different data mining libraries used in the analysis of our data set were shared. It is aimed to give general information about Weka and Apache Spark before explaining our analysis process.

5.1 Weka

Weka is a data mining tool and that contains data mining algorithms for data analysis. You can analyze your data through the desktop application or by referring to the Weka library in your code. Weka use cases are data pre-processing, regression, classification, clustering, association rules, visualization (Weka 3: Machine learning software in java, n.d.).

In previous studies, It has been found that Weka works with less error rate than RapidMiner and Orange. The most sensitive issue in data analysis is the low error rate. However, it is a known fact that Weka does not work well on large data sets. For this reason, we have searched for a new data mining library and as a result of our research we encountered Apache Spark.

5.2 Apache Spark

Spark is a general-purpose data processing engine, suitable for use in a wide range of circumstances. That was designed to run in memory, and this allowed Spark to process data much faster. Spark began life in 2009 as a project within the AMPLab at the University of California, Berkeley (Scott, 2015).

Spark has expansive libraries and APIs, and that supports for different programming languages such as Java, Python, R, SQL and Scala. Spark is often used Hadoop Distributed File System for storage data, but can also integrate with HBase,

Amazon's S3, MongoDB, Cassandra, MapR-DB. Well-known companies such as IBM, Huawei, Chinese search engine Baidu, Alibaba Taobao, social networking company Tencent and pharmaceutical company Novartis are using Apache Spark (Scott, 2015).

In our study, if we explain the reasons why Spark is used in data processing:

- **Simplicity:** We can easily use Spark's capabilities with collection of APIs. These APIs are designed and documented to be easy to use by developers.
- **Speed:** Spark was developed to shorten data processing time. Spark won the Daytona Gray Sort benchmarking challenge in 2014, by processing 100 terabytes of data in 23 minutes and did not leave any doubt about its fast work (Scott, 2015).
- **Support:** Spark supports many programming languages and storage systems. The Apache Spark community is also large, active and international.

5.2.1 Apache Spark Use Cases

Spark is an API-aided tool that integrates into applications to quickly analyze users' data. Use cases of Spark:

5.2.1.1 Spark Streaming

It is very difficult for developers to deal with data streams, such as log files and sensor data. These data usually have a continuous flow from multiple sources. Although it is difficult to store and process this data on disk, it is very important to obtain meaningful information.

5.2.1.2 Machine Learning

Machine learning has become indispensable with increasing data sizes. Spark's ability to store data in memory and respond quickly to queries creates an environment suitable for the machine learning process.

5.2.1.3 Data integration

Data generated by different systems need to be combined to report and analyze. However, since the data are dirty and incompatible, it is not easy to combine these data. In order to analyze the data, data integration is done first. Data integration includes data collection, cleaning, standardization and analysis.

5.2.1.4 Interactive analytics

Spark can quickly respond to and adapt to the interactive query process.

5.2.2 Apache Spark Usage Techniques

Apache Spark is open source software and requires at least Java version 6 and Maven version 3.0.4. You can use the spark with the help of spark shell or APIs.

We used the Apache Spark in the Java Spring boot project via Apache Maven by adding dependency. Apache Maven is a software project management and comprehension tool. Based on the concept of a project object model (POM.xml), Maven can manage a project's build, reporting and documentation from a central piece of information (Maven – Welcome to Apache Maven, n.d.).

Pom.xml dependency for adding Apache Spark.

```
<dependency>
  <groupId>org.apache.spark</groupId>
  <artifactId>spark-mllib_2.12</artifactId>
  <version>2.4.1</version>
</dependency>
```

Recently, many of big data technology has been involved in our lives. In our work, we chose Apache Spark because the speed of the data analysis process is quite good and we gave information about Apache Spark in this part. Spark, like other big data technologies, is not the best technique for every data processing process.

5.2.3 Apache Spark Architecture

Spark is a open source project that developed for use in various architectures with various programming languages. The Spark project stack contains Spark Core and four libraries that optimized to meet the requirements of four different usage situations. Applications must contain Spark Core and at least one of these libraries (Scott, 2015). Figure 5.2 shows project stack of Apache Spark.

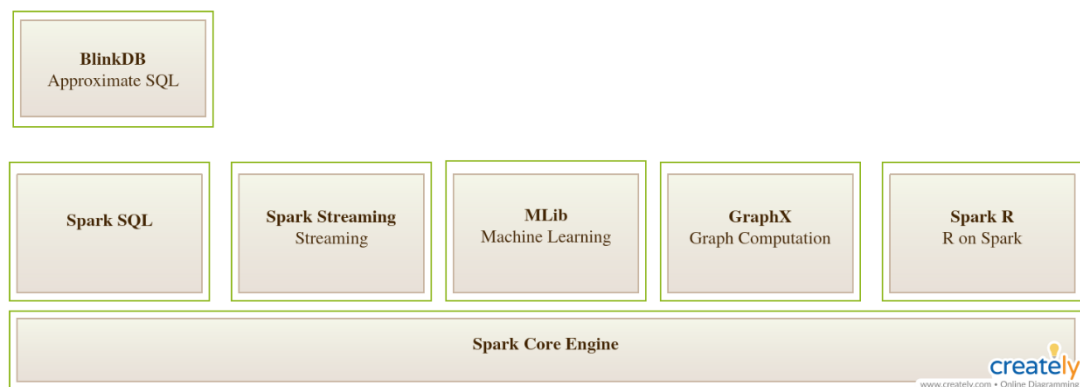


Figure 5.1 Apache Spark architecture

5.2.3.1 Spark Core

It is the foundation stone of Spark and is responsible for management functions such as task scheduling. Spark Core implements and depends upon a programming abstraction known as Resilient Distributed Datasets (Scott, 2015).

5.2.3.2 Spark SQL

Designed to work with structured data. Spark SQL supports the Hive project and the HiveQL query language. Spark can be integrated with SQL databases, data warehouses and business intelligence tools and supports JDBC and ODBC connections.

5.2.3.3 Spark Streaming

The purpose of this module is to process the streaming data.

5.2.3.4 MLlib

It is machine learning library of Spark and contains classification, correlations and hypothesis testing, regression, clustering and principal component analysis.

5.2.3.5 GraphX

Designed to support the analysis on graphs of data and includes many graphical algorithms.

5.2.3.6 Spark R

It was developed for data scientists and statisticians who use R programming language to benefit Spark.

5.2.4 Resilient Distributed Datasets (RDDs)

RDD (Resilient Distributed Dataset) is a very important concept for Spark. It is designed to support in-memory data storage. RDDs are collections of distributed objects and are divided into several parts, each of which is calculated on different nodes. Efficiency is increased by running operations in parallel across multiple nodes in the cluster and minimizing data repetition between these nodes. Two basic operations are performed on the data in the RDD (Scott, 2015).

- **Transformations:** Creating a new RDD with techniques such as mapping, filtering.
- **Actions:** Performing various measurements without changing data.

The original RDD does not change during the process. RDD conversions are logged and If any data loss occurs in cluster nodes, they are repaired immediately. Transformations are not executed if they are not needed by the next process, and this increases efficiency. Because unnecessary data processing is not done. Rdds remains in memory, thus dramatically enhancing performance in repetitive queries and transactions.

CHAPTER SIX

WEKA / SPARK BIG DATA ANALYSIS

In this section, clustering and association rule mining were run on customer transaction data set with both the Weka and Apache Spark and the results were shared. The results of this study contributed to the determination of the data mining library to be used by the system to be developed.

Data analysis is the process of extracting information from large-scale data. There are many libraries to be used during data analysis, and these libraries show differences in working. In the study, the same two algorithms in two different libraries were run on the same data set. The aim is to determine whether the decision to use Apache Spark in the tool to be developed is correct.

The steps we take to analyze our data set and the results for each step are listed below.

6.1 Data Preparation

The data set used in the analysis is the real data set obtained from the ERP system. It includes store products, customers and customer purchasing transactions. Since the data set we have contains customer buying transactions, we aimed to cluster the customers according to the products they bought and to identify the products sold together. In line with this goal, we have also performed data cleaning. In the process of data preparation, Visual Studio 2017, Mssql Server tools and C# programming language were used.

During the data preparation process; It was decided that it would be more meaningful to analyze the products according to their categories because of the large number of product types. Thus, 90 product categories (Jackets, Pants, etc.) were used

in the clustering process. It was found meaningful to determine customer profiles according to the categories of the products they received.

The same data set was also used in a different version in association rule mining. Table 6.1 shows characteristics of the data set before and after data preparation. Since the number of the product code is quite many, the products to be used in the association analysis are referred to as "Gender-Color-Category". A total of 2133 product's category were analyzed in the association rule mining.

Table 6.1 Characteristics of data set

	Data set	Cleaned data set
The number of customers	3,119,918	718,903
The number of sales transaction	4,275,910	4,275,910

6.2 Data Mining

The data mining step is to run data mining algorithms on the cleaned data set. Two different data mining libraries were used in our study. Although the algorithms we use are the same, the number of repeats they make up to the end of the analysis varies. In this context, the accuracy rates of the results they produce differ.

6.2.1 K-Means Algorithm

K-Means algorithm is one of the most widely used clustering algorithms. The K value indicates how many clusters we can allocate to our customers. In our study, the K-Means algorithm was used to express the similar customers with the same profile. Customers are divided into profiles according to the products they have buy or not buy from 90 product categories in the store.

6.2.2 Fp-Growth Algorithm

The main goal in association analysis is to find patterns and relationships that are frequently repeated in big data sets. When we discover patterns that are frequently repeated in our data set, we will identify the products sold together. The Fp-Growth algorithm was used to determine the products sold together in the data set. In the 2133 product category, the product categories sold were determined by the result of the Fp-Growth algorithm.

6.3 Weka / Spark Comparison and Evaluation

K-Means and Fp-Growth algorithms were run with Weka and Apache Spark. The results were shared.

6.3.1 K-Means Algorithm Results

Sum of Squared Errors is the sum of the squared differences between each observation and its group's mean. n is the number of observations x_i is the value of the i th observation and \bar{x} is the mean of all the observations (Error Sum of Squares (SSE), n.d, s.1).

Equation 6.1 shows “Sum of Squared Errors”.

$$\text{Sum of Squared Errors} = \sqrt{\sum_{i=1}^K (x_i - \bar{x})^2} \quad (6.1)$$

K-Means algorithm was run with different cluster numbers (in the range of 2 to 20) on the cleaned data set in Spark and Weka. Considering the working time and error rates of the algorithm, it was observed that Spark was working longer time and the error rate was less than Weka. Although the same algorithm is used, the end points of the analysis tools and the error rates are differ. As both data mining tools have a result generation time in seconds, it will be the most meaningful result of the number of tools and clusters with the lowest error rate. The execution times for the result of the k-means algorithm are shown in Figure 6.1 and the sum of squared errors are shown in Figure 6.2.

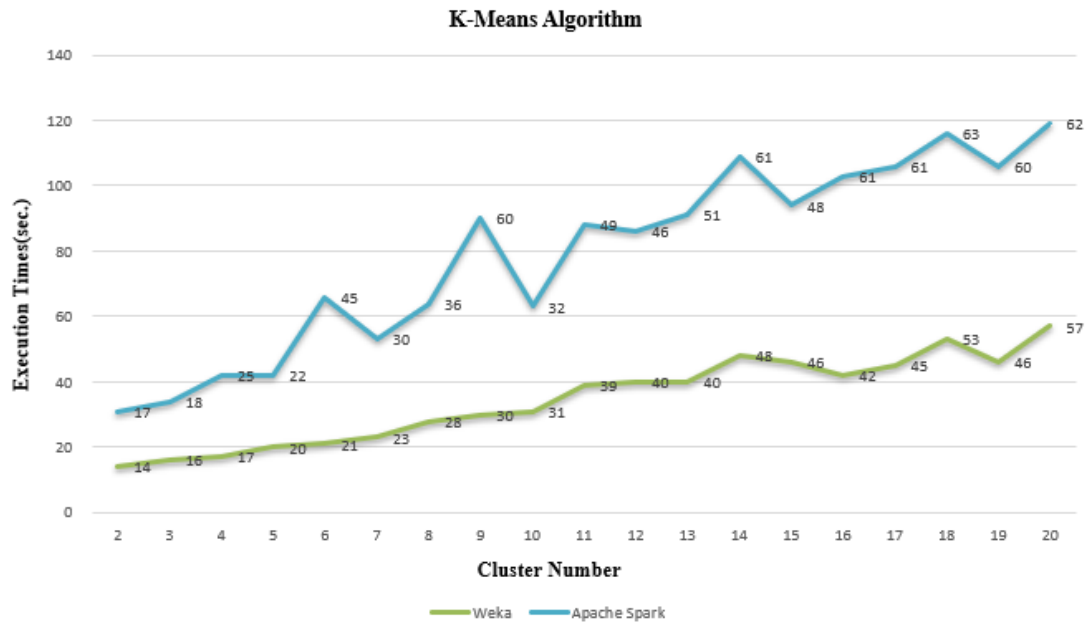


Figure 6.1 K-Means algorithm operation time

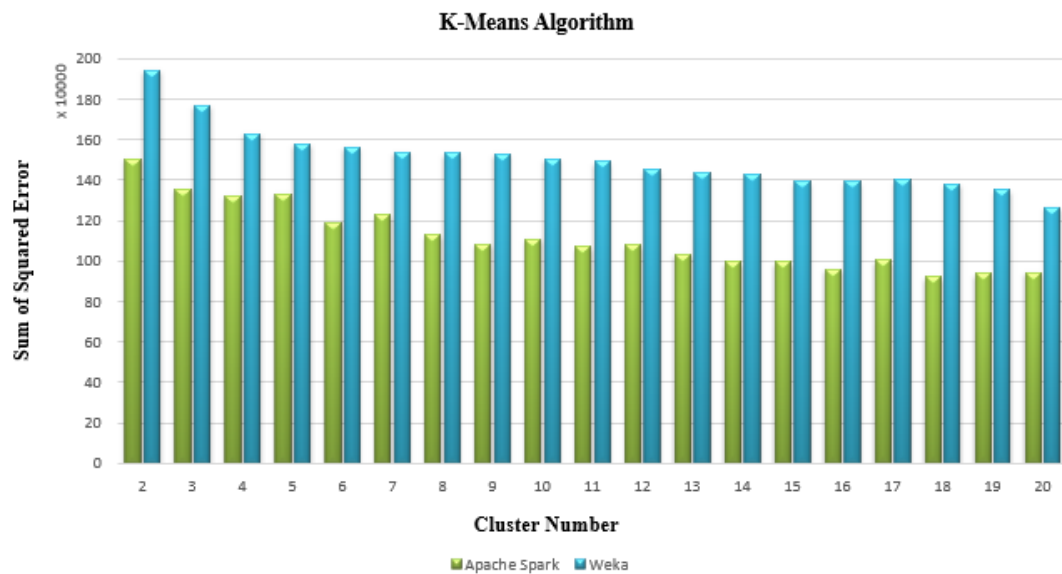


Figure 6.2 K-Means algorithm sum of squared error

We selected Spark's result for 18 clusters as example. The result of clustering of Spark is shown in Figure 6.3.

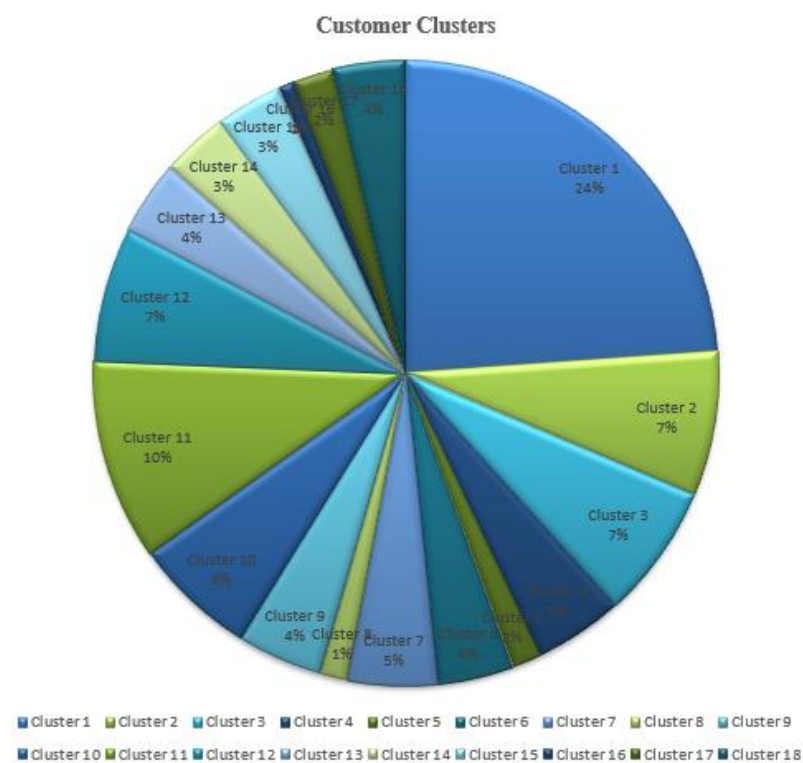


Figure 6.3 Customer clustering result

6.3.2 Fp-Growth Algorithm Results

Due to the size of the data size, Weka Fp-Growth Algorithm could not produce results. The execution times of the different minimum support values for the Spark Fp-Growth algorithm are given Figure 6.4.

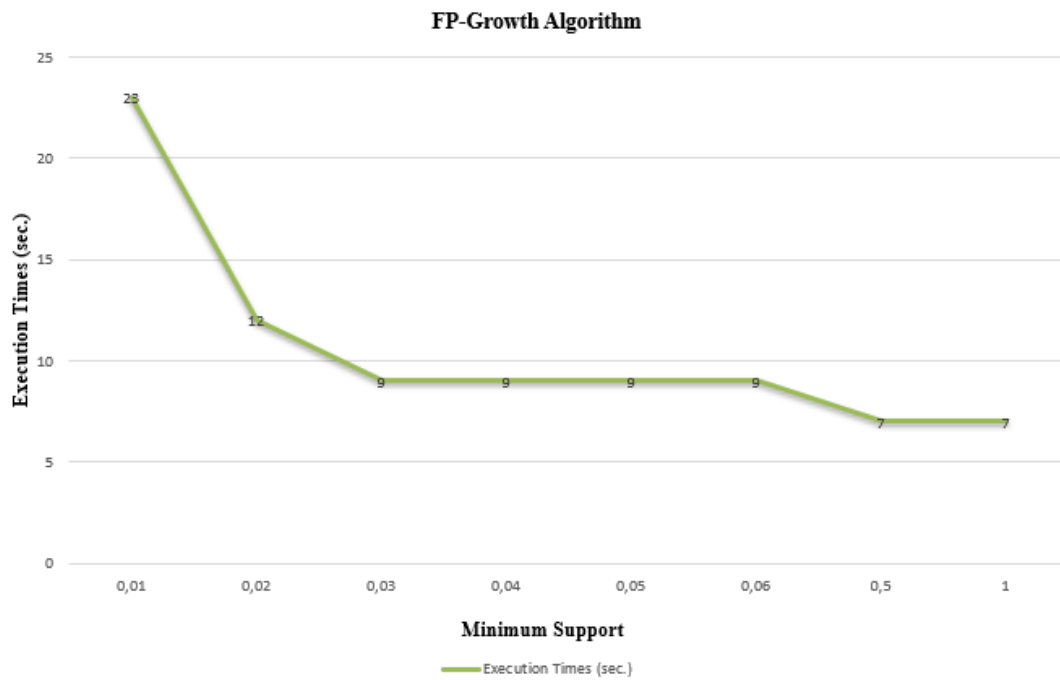


Figure 6.4 Spark Fp-Growth algorithm operation time

Number of rules different minimum support values for the Spark Fp-Growth algorithm are given Figure 6.5.

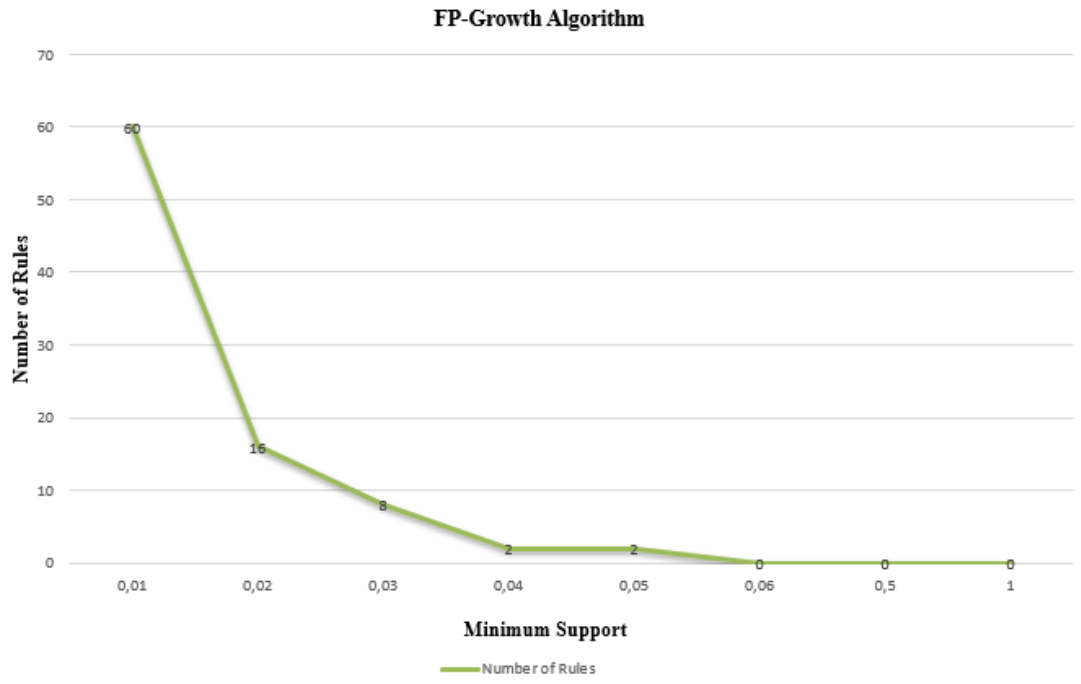


Figure 6.5 Spark Fp-Growth algorithm number of rules

Using the Fp-Growth algorithm, 60 rule with 0.01, 16 rule with 0.02, 8 rule with 0.03, 2 rule with 0.04 and 0.05 minimum support value were obtained. No rule set was found for minimum support value of 0.05 and greater. According to these results; A black belt area should be offered to a customer black socks and black shoes. If we give a few examples of the rules set:

- men's black straight suit, men's white cvc shirt
- men's blue shirts, men's white shirts
- men's navy blue suit, men's white shirt
- men's black shirt, men's white shirt
- men's black belt, male black socks
- men's black shoes, men's black belt

CHAPTER SEVEN

THE PROPOSED PROJECT

We named the tool as “Sparkle Mining”. The user completes the analysis process by performing the steps in the diagram below with using the Sparkle Mining. Figure 7.1 shows the analysis steps followed by the user of the developed tool.

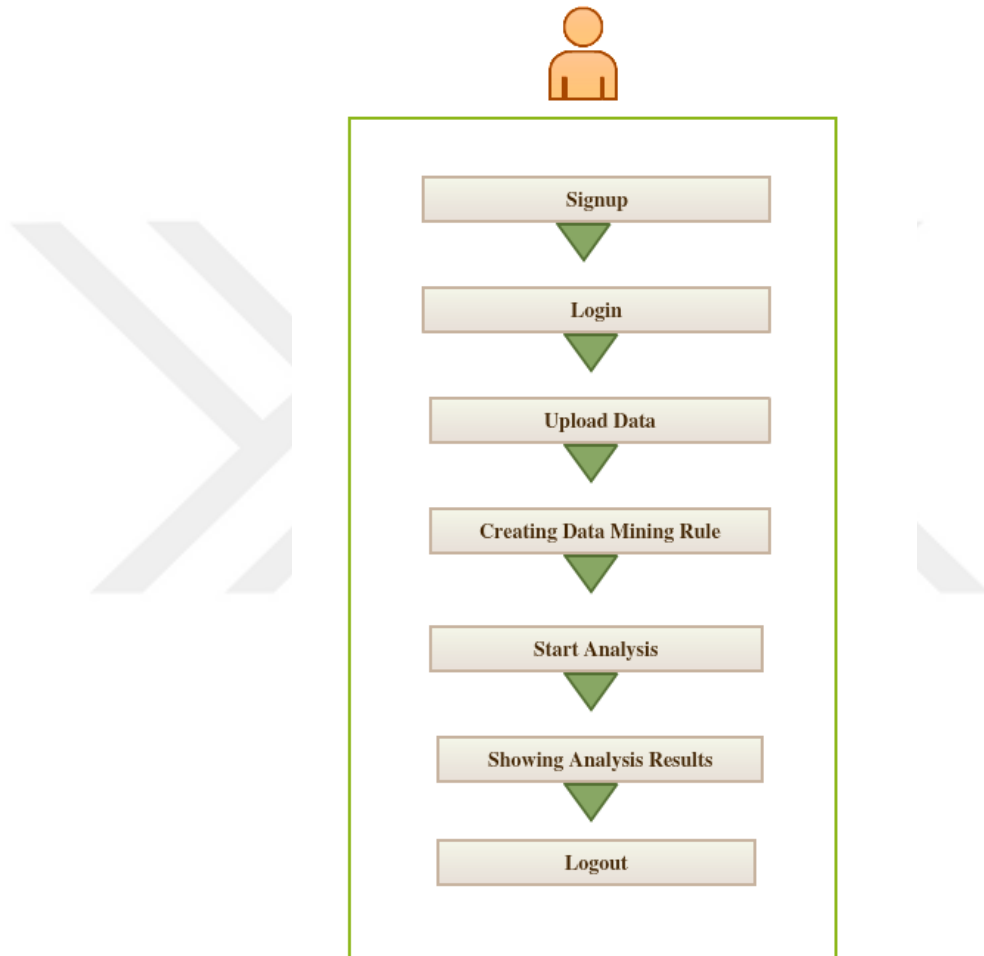


Figure 7.1 The proposed project flow diagram

7.1 Sign up and Login

The user can register the system by entering the user name, first name, last name and password through the screen shown in Figure 7.2. User’s username and e-mail must be unique. For each registered user, a partition and blob container is created in

the azure storage table so that data and analysis results can be found quickly. Partition Key and blob container's name is username for each user. All information of the user is queried with Partition and Row Key. If the user has already registered to the system, click on the "I am already a member" button and can be redirected to the Login page.

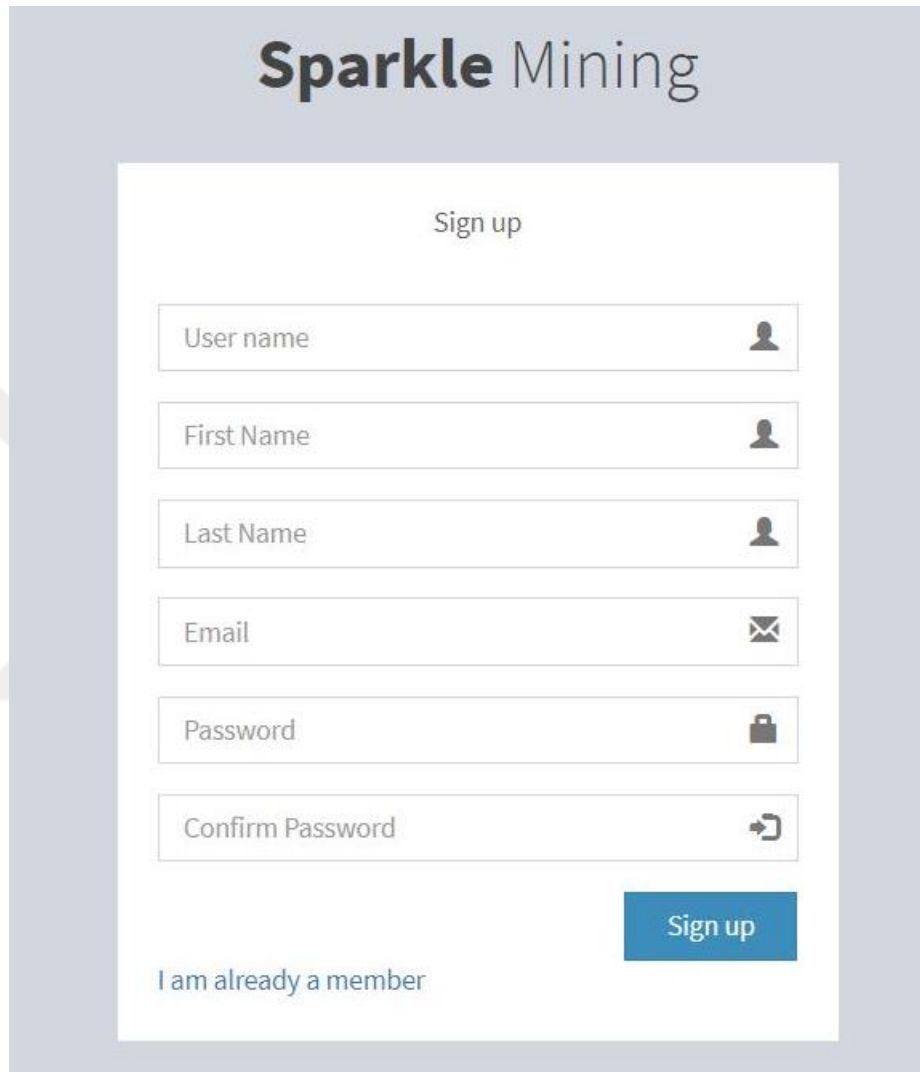
The image shows a web form titled "Sparkle Mining" with a subtitle "Sign up". The form contains six input fields: "User name", "First Name", "Last Name", "Email", "Password", and "Confirm Password". Each field has a corresponding icon on the right: a person icon for the first three, an envelope icon for Email, a lock icon for Password, and a circular arrow icon for Confirm Password. A blue "Sign up" button is located at the bottom right of the form. Below the button, there is a link that says "I am already a member". The entire form is set against a light blue background with a subtle geometric pattern.

Figure 7.2 Sign up page

By completing the login process by entering the user name and password, the user will be using the application actively. Figure 7.3 shows login screen of the developed tool.

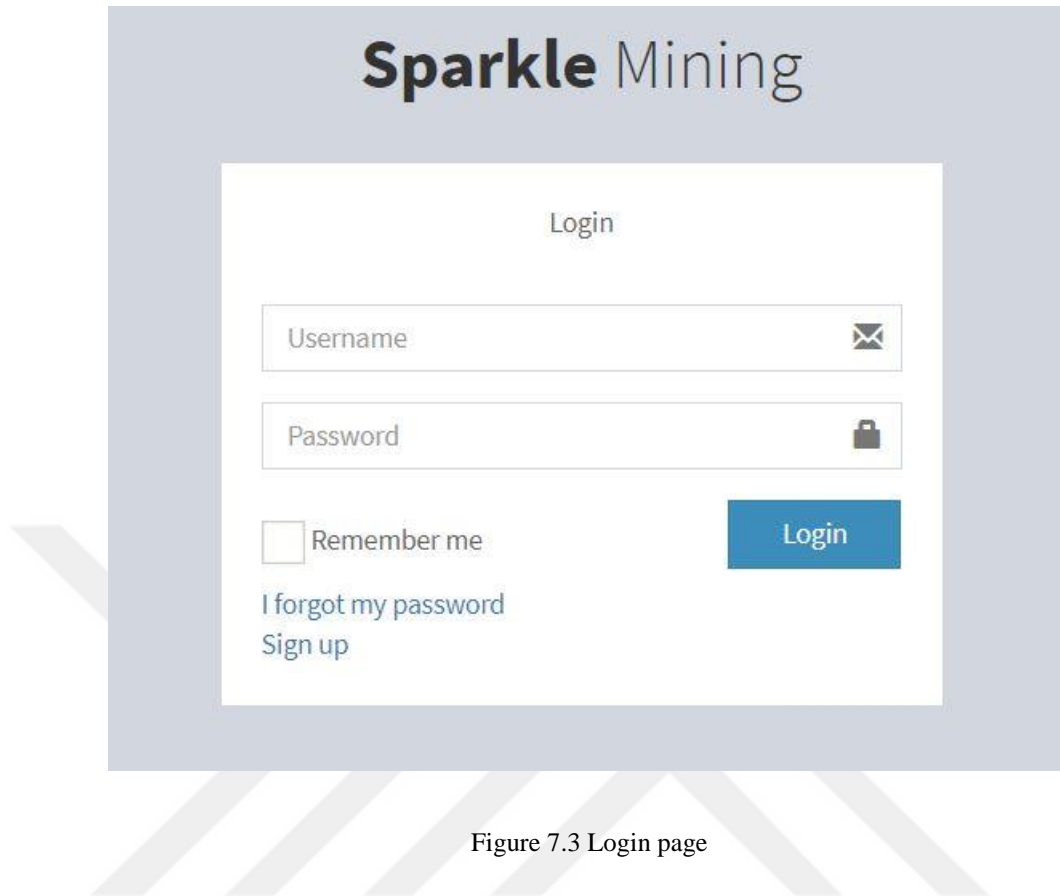


Figure 7.3 Login page

The user who completes the entry will meet with the “My Data” page. Figure 7.4 shows “My Data” screen. The screen is designed in a very simple structure. The goal is to enable the user to analyze the data set with little effort. In the example shown, two analysis files are uploaded into the system. Since the system belongs to the gamzeozcelik, the files are located in the gamzeozcelik blob container. We can understand it from the cloud url information. Using this screen, you can also see the type of files uploaded, when it is uploaded and delete what you want to delete. The page consists of two tabs, "My Data" and "Analysis Results".

By using the "My data" tab;

- can upload data,
- list the data sets uploaded,
- can save the cloud url of the data sets in the cloud
- can create a rule of analysis for the data sets it has uploaded to the system.

By using the "Analysis Results" tab;

- can see the rules and conditions (Created, Ongoing, Finished, Failed) of analysis,
- can initiate analysis
- can see the results of analysis.

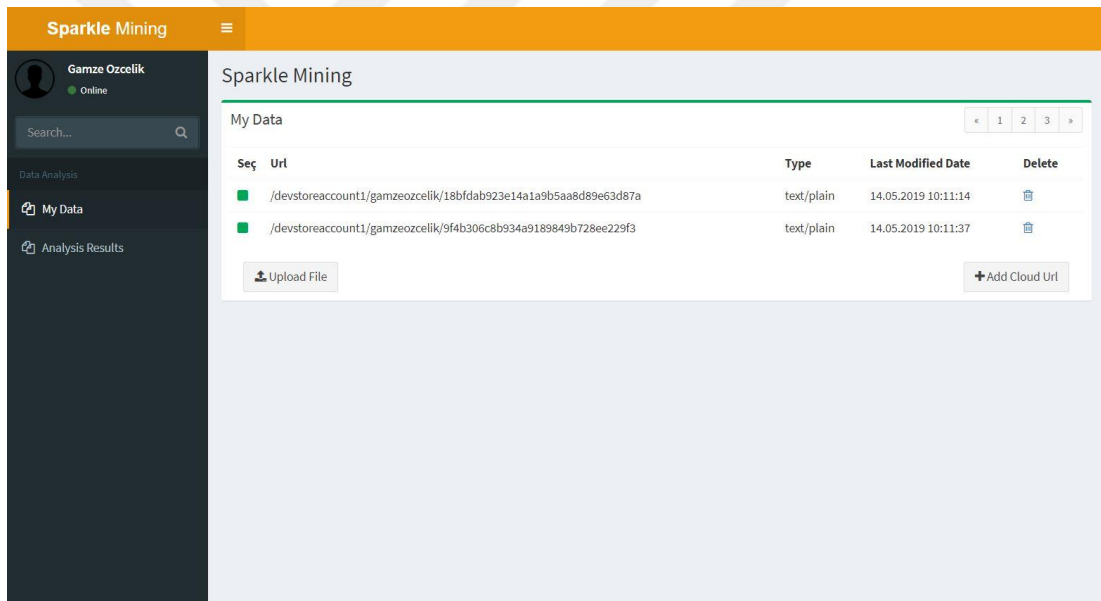


Figure 7.4 My data page

7.2 Data Upload on Web Application

Data sets can be analyzed in .csv format as application standard. Therefore, the user must convert the data to a .csv format, either manually or through a code. The user can upload his data to the system by using the "Upload File" button. The data uploaded to the system is stored under the user's container that is under Azure Blobs.

If the user does not want to upload his / her data to the system, he / she can save his / her cloud information and analyze his / her data in this way. In this case, only the user's cloud information is stored in the Azure Storage Tables and no user data is stored in Azure Blob. All of the cloud addresses of the user's data stored in the system and uploaded to the system are listed on the main screen. Figure 7.5 shows popup where the user saves the cloud information.

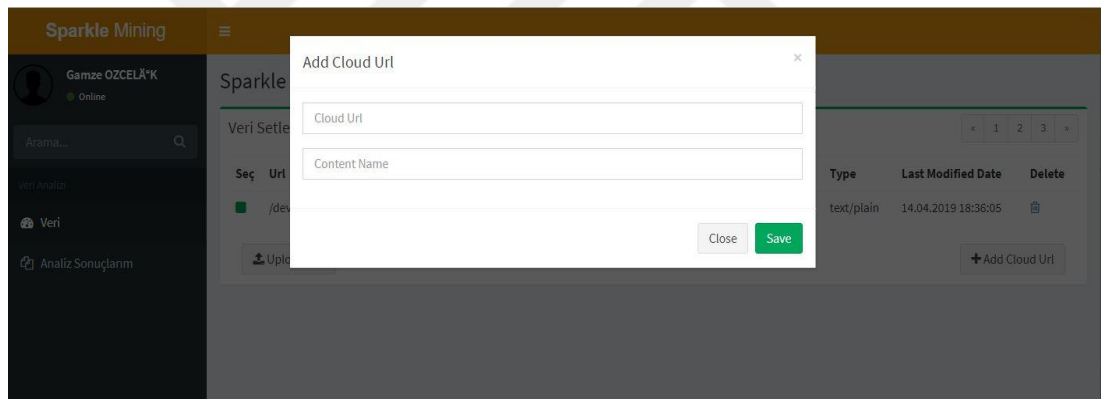


Figure 7.5 Cloud url save popup

7.3 Defining Data Mining Rule

Data mining rule popup opens when the user clicks on any data set from the list. Figure 7.6 shows data mining rule popup. The user should specify which data mining technique to use on this popup and a name for the analysis to be made. Sparkle Mining includes only the Association Rule Mining technique. In future studies, other data mining techniques and different algorithms for each data mining technique will be added to the system scope. The design of the system was designed with these in mind.

The parameters that the system expects to run 4 different data mining techniques:

- **Classification operation parameters:** target attribute, maximum iteration.
- **Regression operation parameters:** target attribute, maximum iteration.
- **Clustering operation parameters:** cluster number, maximum iteration.
- **Association rule mining parameters:** minimum support, confidence, maximum iteration.

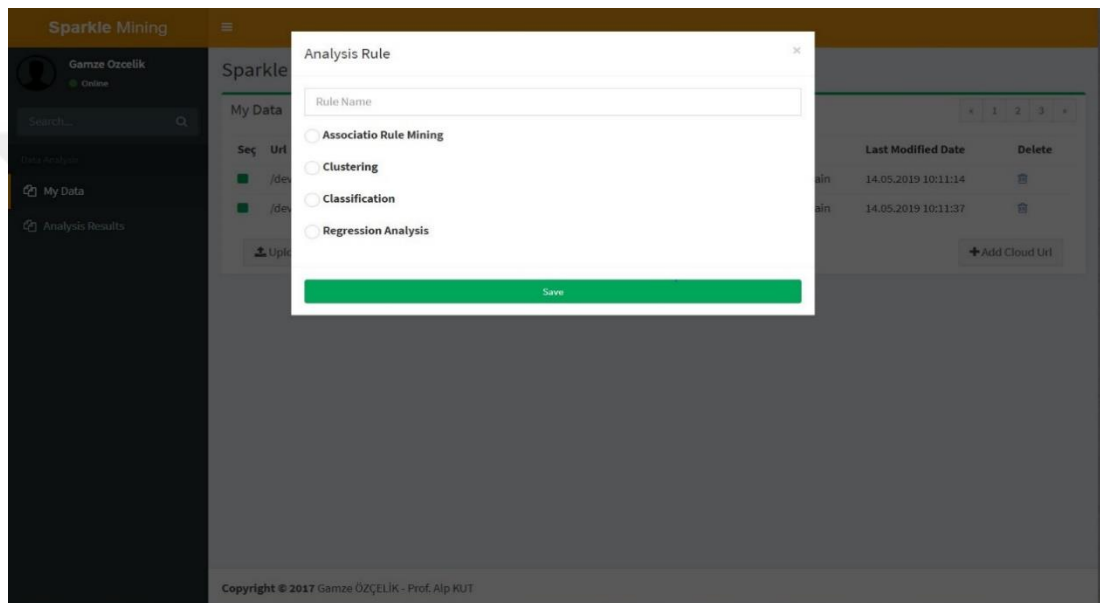


Figure 7.6 Defining data mining rule popup

The user-defined rule set is kept in XML format in order to provide dynamic in application. An example of the set of rules defined for association rule mining is shared below.

```
<?xml version="1.0" encoding="utf-16"?>
<AnalysisRule>
  <RuleName>Rule 2</RuleName>
  <CloudUrlOfFile>/devstoreaccount1/gamzeozcelik/18bfdab923e14a1a9b5aa
8d89e63d87a</CloudUrlOfFile>
  <AnalysisType>AssociationRule</AnalysisType>
```

```

<MinSupport>0.02</MinSupport>
<Confidence>0.1</Confidence>
<AssociationMaxIteration>50</AssociationMaxIteration>
<ClusteringMaxIteration>0</ClusteringMaxIteration>
<ClassificationMaxIteration>0</ClassificationMaxIteration>
<RegressionMaxIteration>0</RegressionMaxIteration>
<ClusterNumber>0</ClusterNumber>
<TargetAttribute4Regression>0</TargetAttribute4Regression>
<TargetAttribute4Classification>0</TargetAttribute4Classification>
</AnalysisRule>

```

The parameters for the analysis are taken from the user via the popup in Figure 7.7. The rule set defined by the user with these parameters is converted to xml and stored in Azure storage tables. The first status of the saved rule set is "Created". Minimum support, Confidence and maximum iteration value are taken for association rule mining.

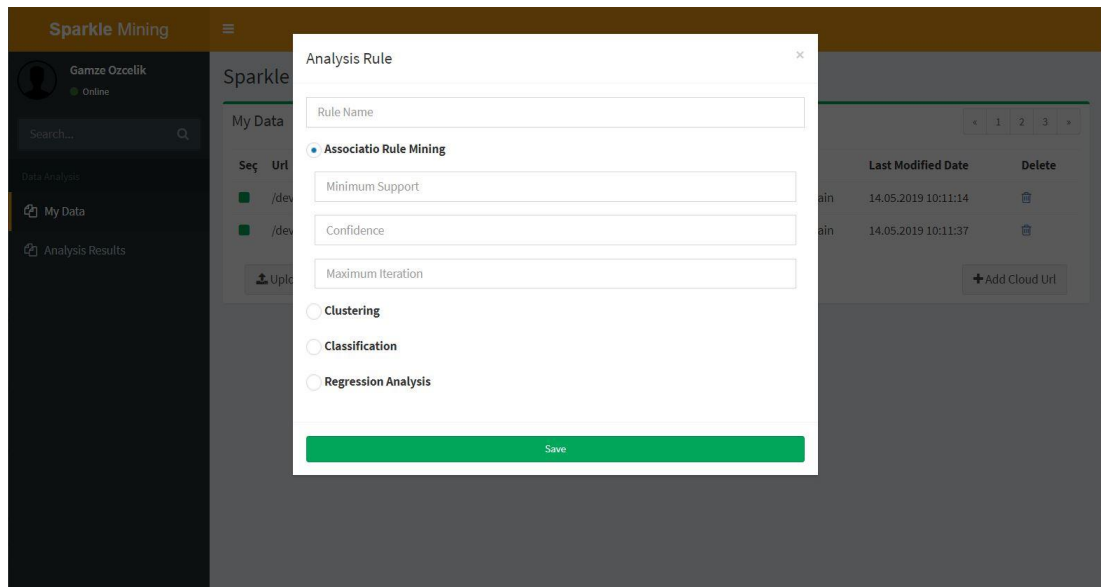


Figure 7.7 Defining association rule popup

7.4 Data Mining

We were prepared data and created our rule for analyzing and we ended our work on the "My data" tab. We can examine the status of the analysis rules that we define from the "Analysis Results" tab in Figure 7.8. The following screen shows four different colors and icons. As you can see from the colors and icons:

- **Yellow icon:** created analysis operation.
- **Blue icon:** ongoing analysis operation.
- **Green icon:** finished analysis operation.
- **Red icon:** failed analysis operation.

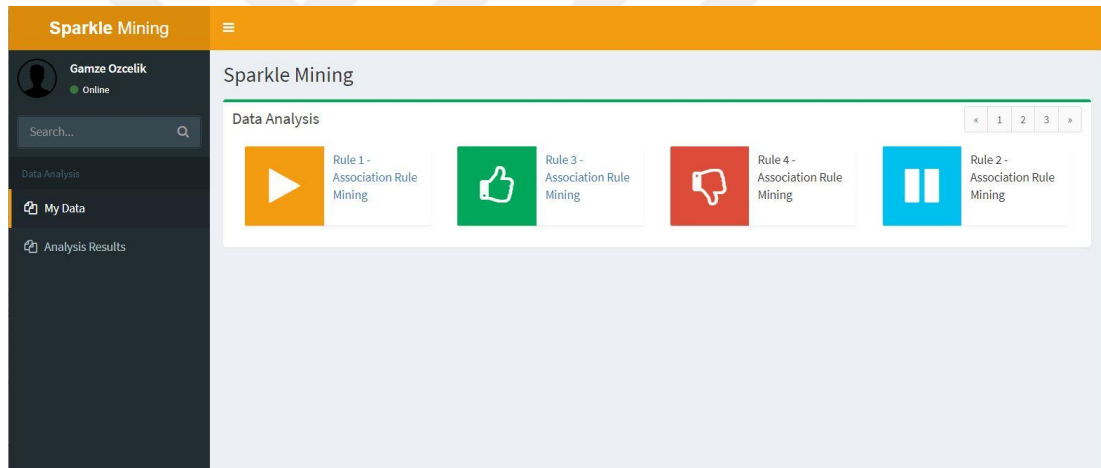


Figure 7.8 State of data mining process page

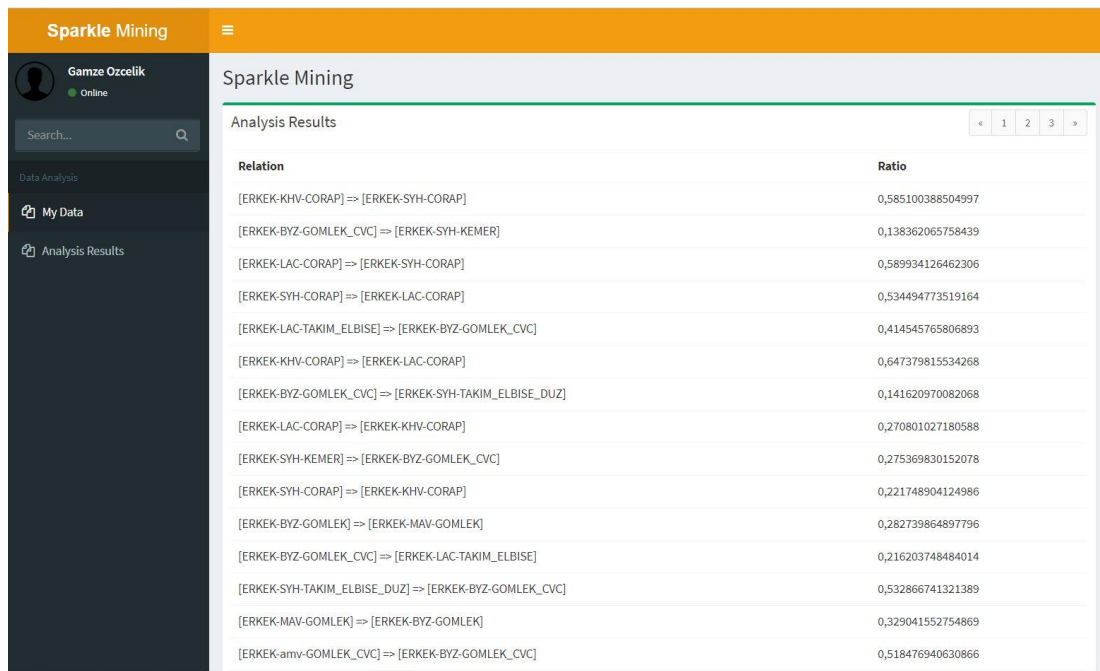
Yellow and green buttons are clickable but blue and red buttons are not clickable. A yellow button indicates that the analysis rule has been created but that the analysis is not started. The user can save the analysis rule to the system but start the analysis later. When the yellow button is clicked, the analysis is started and the button color becomes blue. The blue button is not clickable because it is not allowed to stop the analysis. When the analysis process receives an error, the color of the button becomes red and the color of the button becomes green when the analysis is successful. When

the green button is clicked, the results of the analysis of the data set determined according to the specified rule are shown.

7.5 Representation of Analysis Results

Clicking the green buttons on the “Analysis Results” tab will display the results of the defined analysis. Scope of work; since there is only association rule mining, the relation and ratio information is displayed on the screen. different information will be shown for clustering, classification, regression analysis. The result of two different analyzes with two different data sets is shown below. One of the data sets can be analyzed, while the other could not be analyzed with weka.

The first one of the analysis results is about 4 million 200 thousand sales transaction. This data set could not be analyzed by weka due to its large size. The results taken by Apache Spark are shown in Figure 7.9. The reason for being very low in association is that there must be a large number of associations in order to achieve a rate because of the large number of transaction.



Relation	Ratio
[ERKEK-KHV-CORAP] => [ERKEK-SYH-CORAP]	0,585100388504997
[ERKEK-BYZ-GOMLEK_CVC] => [ERKEK-SYH-KEMER]	0,138362065758439
[ERKEK-LAC-CORAP] => [ERKEK-SYH-CORAP]	0,589934126462306
[ERKEK-SYH-CORAP] => [ERKEK-LAC-CORAP]	0,534494773519164
[ERKEK-LAC-TAKIM_ELBISE] => [ERKEK-BYZ-GOMLEK_CVC]	0,414545765806893
[ERKEK-KHV-CORAP] => [ERKEK-LAC-CORAP]	0,647379815534268
[ERKEK-BYZ-GOMLEK_CVC] => [ERKEK-SYH-TAKIM_ELBISE_DUZ]	0,141620970082068
[ERKEK-LAC-CORAP] => [ERKEK-KHV-CORAP]	0,270801027180588
[ERKEK-SYH-KEMER] => [ERKEK-BYZ-GOMLEK_CVC]	0,275369830152078
[ERKEK-SYH-CORAP] => [ERKEK-KHV-CORAP]	0,221748904124986
[ERKEK-BYZ-GOMLEK] => [ERKEK-MAV-GOMLEK]	0,282739864897796
[ERKEK-BYZ-GOMLEK_CVC] => [ERKEK-LAC-TAKIM_ELBISE]	0,216203748484014
[ERKEK-SYH-TAKIM_ELBISE_DUZ] => [ERKEK-BYZ-GOMLEK_CVC]	0,532866741321389
[ERKEK-MAV-GOMLEK] => [ERKEK-BYZ-GOMLEK]	0,329041552754869
[ERKEK-amv-GOMLEK_CVC] => [ERKEK-BYZ-GOMLEK_CVC]	0,518476940630866

Figure 7.9 Analysis result page for first data set

The second analysis results is about the small size sample data set. Due to the small size of the data set, the number of association rules is quite many and high rate. The second analysis results for small data set are shown in Figure 7.10.

Relation	Ratio
$[q, t, z] \Rightarrow [x]$	1
$[q, x] \Rightarrow [y]$	1
$[y, z] \Rightarrow [x]$	1
$[s, x, z] \Rightarrow [t]$	1
$[t, s, y] \Rightarrow [z]$	1
$[q, x] \Rightarrow [z]$	1
$[y, x, z] \Rightarrow [t]$	1
$[q, y, x, z] \Rightarrow [t]$	1
$[q, t, z] \Rightarrow [y]$	1
$[t, y, z] \Rightarrow [x]$	1
$[q, y] \Rightarrow [z]$	1
$[y] \Rightarrow [t]$	1
$[q, t, y] \Rightarrow [x]$	1
$[s, x, z] \Rightarrow [y]$	1

Figure 7.10 Analysis result page for second data set

CHAPTER EIGHT

USED TECHNOLOGIES

This chapter explains used technologies during the development process throughout the project. Figure 8.1 shows technology architecture of proposed project.

8.1 The Proposed Project Technology Architecture

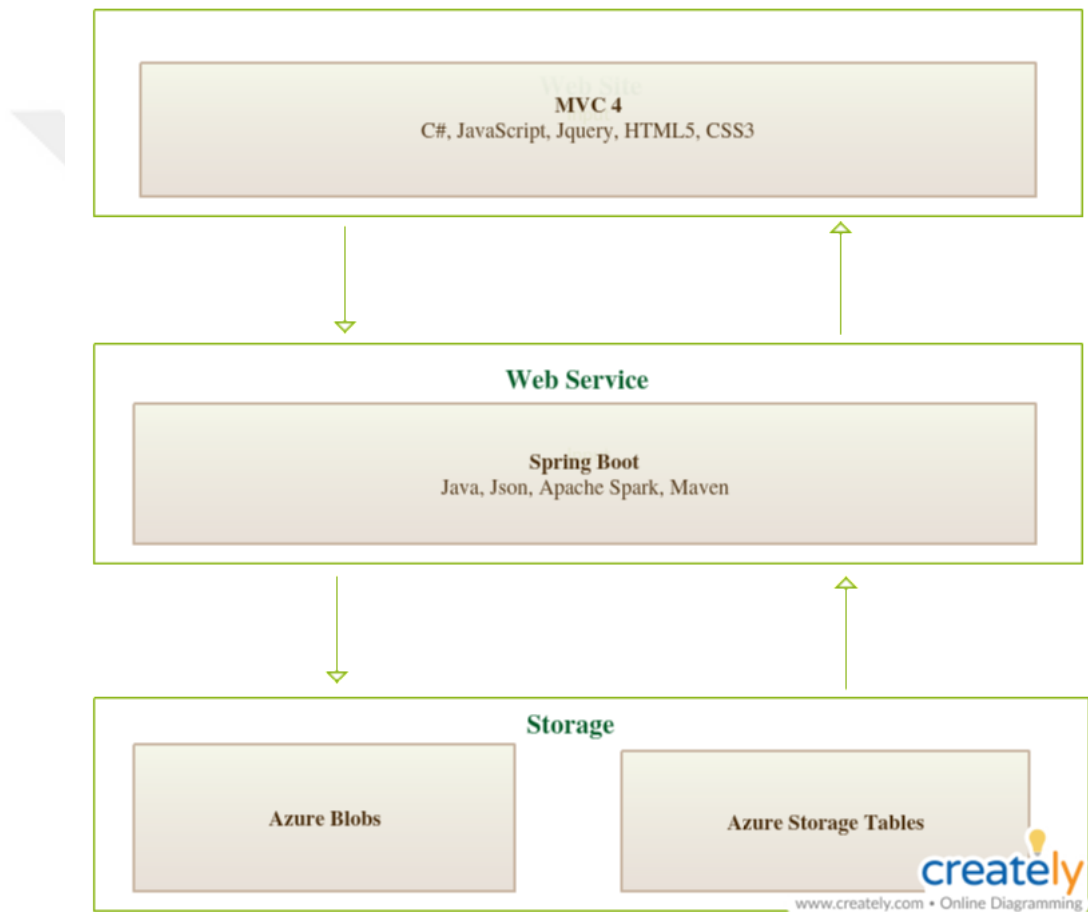


Figure 8.1 Proposed project technology architecture

8.2 ASP.NET

The ASP.NET is a free, cross-platform, open source platform for developing many different types applications. We can use multiple languages, editors and libraries to build for web, mobile, desktop, gaming and IOT applications with .NET. We can write our .NET applications with C#, F# or Visual Basics. Asp.Net supports many databases as SQL Server, MySQL, PostgreSQL, SQLite, MongoDB, Redis, Azure Storage (What is .NET?, n.d.).

ASP.NET framework is used in the web site of the this study. C# was used as a programming language, Nuget used for package management and ASP.NET MVC is used as architectural pattern.

8.2.1 ASP.NET MVC

MVC is a web application architecture and separate the application into three layers as data, user interface, application layer. In this way, layers can be used and updated independently (ASP.NET MVC Pattern, n.d.).

- **Model layer** responsible for the application data. We can say that object definitions are object oriented.
- **View layer** presenting the model data as an interface to the user.
- **Controller layer** responsible for evaluating the user requests and updating the relevant model. It acts as an interface between Model and View.

8.3 Spring Framework

Spring application is open source development framework for enterprise Java. The Spring Framework is very common to use because it provides a high-performance, reusable code, modular structure and easily testing.

Features of Spring Framework:

- **MVC Architecture**
- **Dependency injection (DI):** According to principles of Java, a class should be as independent as possible from other classes. In this way, complex scenarios can be avoided when a class is to be modified. Dependency Injection is applied to minimize this dependence. Dependencies with “Dependency Injection” are injected into runtime not specified in the code (Spring Framework – Overview, n.d.).
- **Inversion of control (IoC):** IoC explained as taken control from the application, is transferred to the Spring Framework. Thanks to IoC dependencies, life cycles, management of objects are left to Spring Framework. “Dependency Injection” is one of the “Inversion of Control” examples (Spring Framework – Overview, n.d.).
- **Aspect oriented programming (AOP):** “Cross cutting concerns” are issues that may be needed in any application, except for the business logic of the application. The best example for “cross cutting concerns” is logging and caching. No matter what our application is, we will need logging and caching. The AOP offers the possibility to distinguish between the requirements that can be evaluated within the concept of cross cutting concerns and the classes affected. In this way, we do not need to deal with anything other than what we need in business logic in our classes, and since we handle the issue of logging in a separate point, we can use it in every point we need in the application. With using AOP, we can improve our application on reusability, extensibility, maintainability, modularity issues (Spring Framework – Overview, n.d.).

8.3.1 Spring Boot

Spring Boot is a framework that makes it easy for us to develop Spring based applications. It provides the necessary infrastructure to focus on our application without wasting much time on infrastructure features. In many applications we create, there are beans we use in general. These come automatically with the Spring Boot project. For example, these beans; Tomcat Server, Spring MVC (Building an Application with Spring Boot, n.d.)

We created the web service part of our application as a Spring Boot project. This is because our web service uses the Apache Spark library to perform our data analysis. Since Apache Spark does not support the C# language, we have developed the web service with the Spring framework, although our website was developed with the .Net framework. In the Spring boot project, because Tomcat server and Spring MVC are already present, we have not added any different dependency to the application except to add the Apache Spark library.

8.4 Azure Storage

Azure Storage is the cloud storage method developed by Microsoft. Azure Storage contains Azure Blobs, Azure Files, Azure Queues, Azure Tables services. Azure Blobs and Azure Tables were used in proposed project.

The following dependencies were added to the application to use Azure storage.

```
<dependency>
  <groupId>com.microsoft.azure</groupId>
  <artifactId>azure-storage</artifactId>
  <version>6.1.0</version>
</dependency>
```

```

<dependency>
  <groupId>com.microsoft.azure</groupId>
  <artifactId>azure-keyvault-extensions</artifactId>
  <version>0.8.0</version>
</dependency>

```

Figure 8.2 shows the storage structure of the project displayed via Azure Storage Explorer.

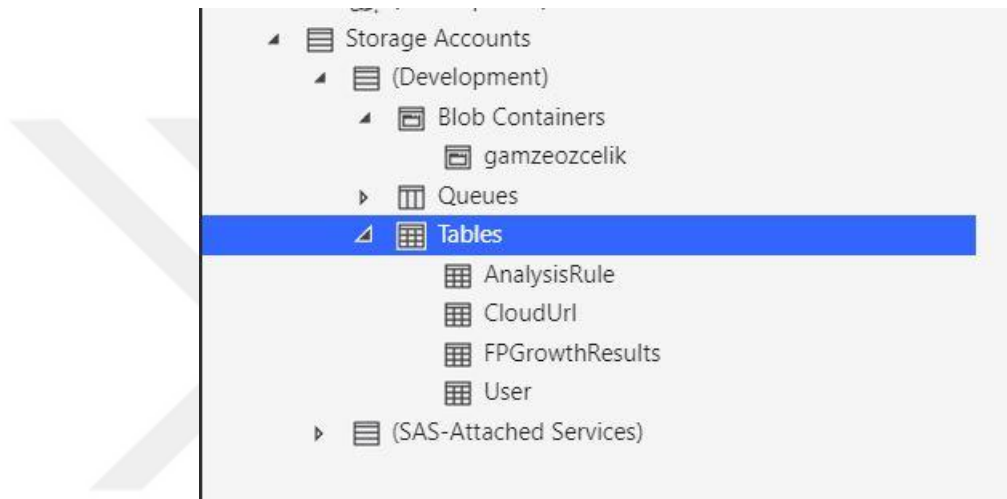


Figure 8.2 Azure storage on azure storage explorer

8.4.1 Azure Storage Naming Rules

Blob container was created for each user who registered to the system. However, since Azure has naming rules, the user name entered during the registration was checked according to the naming rules listed below. These naming rules shown in Table 8.1.

Table 8.1 Azure storage naming rules

Kind	Length	Casing?	Valid chars?
Storage Account	3 - 24	lowercase	alphanumeric
Blob Name	1-1024	case-sensitive	any url char
Container Name	3 - 63	lowercase	alphanumeric and dash
Queue Name	3 - 63	lowercase	alphanumeric and dash
Table Name	3 - 63	case-sensitive	alphanumeric

8.4.2 Azure Storage Tables

Azure Table Storage provides storing structural data in the cloud. It is very easy to update when data needs change in the application. (Introduction to Table storage - Object storage in Azure, n.d.) In our application user information, defined rule sets, analysis results are stored in Azure Storage Tables. The diagram of our Azure Storage Table is shown in Figure 8.3.

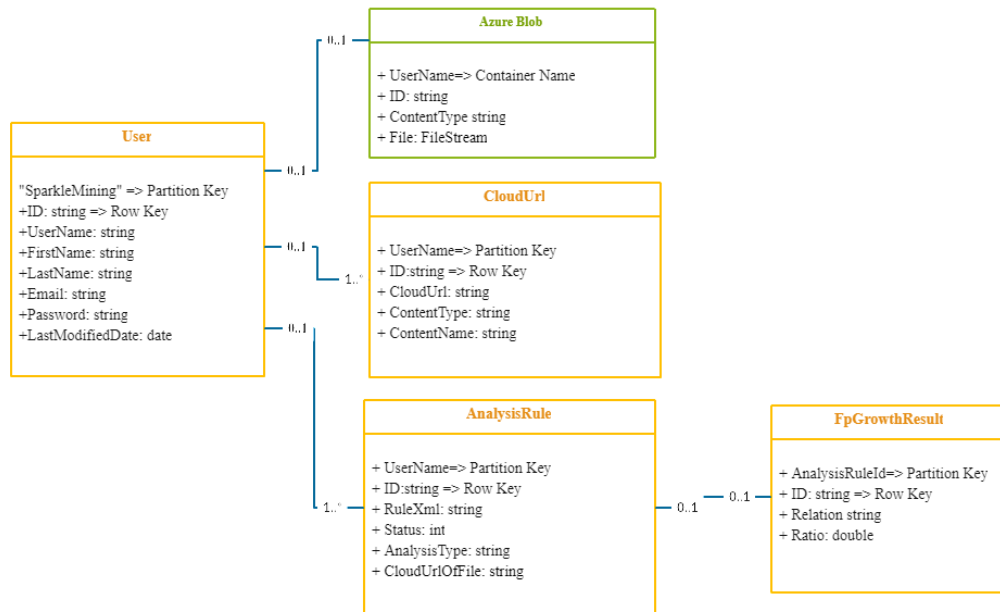


Figure 8.3 Azure storage tables diagram

8.4.3 Azure Blobs

Azure Blob storage is object storage technique for the cloud that was developed by Microsoft. Blob storage is optimized for storing big unstructured data. Unstructured data is data that is not compatible to a particular data model or definition, such as text or binary data. Client applications can access in Blob storage's objects via HTTP/HTTPS, from everywhere. Objects in Blob storage are accessible via the Azure Storage REST API, Azure PowerShell, Azure CLI, or an Azure Storage client library. Client libraries are available for many languages, like PHP, .NET, Python, Java, Node.js (Quickstart: Use .NET to create a blob in object storage - Azure Storage, n.d.).

We have access to Azure Blob Storage via Azure Storage client library with .net and java languages and we stored our text files. We used Azure Blobs for storing text files. In our project, we used Azure Blobs to store files that users upload to the system.

CHAPTER NINE

CONCLUSION AND FUTURE WORK

9.1 Conclusion

When the results of two different algorithms on Weka and Spark are examined; It was observed that the error rate for each cluster number of K-Means algorithm running on Spark was lower than Weka. In the association rule mining stage, Although it is not possible to run the Fp-Growth algorithm used in Weka due to the data size, the result was generated in a short time on Spark. These results show us that the analysis of large-scale data is possible with Apache Spark and that the results produced with Apache Spark are more accurate. While analyzing big data sets, it was determined that analyzing with Apache Spark also provided an advantage in terms of accuracy as well as performance.

“Sparkle Mining” has been developed to enable Apache Spark library to be used by non-experts in code writing and data mining. The developed data mining tool is cloud-based, working with good performance and is quite easy to use. On the other hand, it has a dynamic design that allows users to direct their analysis in the way they want. It is a easily accessible system depending on the storage the data and analysis results in the cloud. The application is web based and the users to reach the results of analysis whenever they want.

9.2 Future Work

There are not all data mining techniques and algorithms in the developed application. However, the system was designed considering that all algorithms will be added during development. In the next step, all algorithms will be added to the application and users will be provided with a comprehensive data mining tool.

In the application, it is intended that the user conduct analysis by both file upload and cloud information recording. However, there is no study based on taking, recording and managing cloud access information from people. In the next stage, people will be able to analyze directly on their own cloud systems without uploading their files.



REFERENCES

- Ahmed, K. P. (2017). Analysis of data mining tools for disease prediction. *School of Computer Science and Engineering*, 9(10), 1886-1888.
- ASP.NET MVC Pattern (n.d.). Retrieved April 22, 2019, from <https://dotnet.microsoft.com/apps/aspnet/mvc>.
- Asri, H., Mousannif, H., & Moatassime, H. A. (2017). Real-time miscarriage prediction with spark. *The 7th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare*, 113, 423-428.
- Bharati M., & Ramageri M. (2010). Data mining techniques and applications. *Indian Journal of Computer Science and Engineering*, 1(4), 301-305.
- Building an Application with Spring Boot (n.d.). Retrieved May 10, 2019, from <https://spring.io/guides/gs/spring-boot>.
- Chaudhari, S. (2015). Data mining using cloud computing. *3rd International Conference, associate with ISTE(New Delhi) & International Journal of Pure & Applied Reserch in Engineering & Technology*.
- Dada, E. G., Bassi, J. S., Hurcha, Y. J., & Alkali, A. H. (2019). Performance evaluation of machine learning algorithms for detection and prevention of malware attacks. *IOSR Journal of Computer Engineering (IOSR-JCE)*, 21(3), 18-27.
- Dhote R. A., & Deshpande S. P. (2016). data mining with cloud computing: - an overview. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 5(1), 211-214.

Error Sum of Squares (SSE) (n.d.). Retrieved July 10, 2019, from https://hlab.stanford.edu/brian/error_sum_of_squares.

Györödi, C., Györödi, R., & Holban, S. (2004). A comparative study of association rules mining algorithms. In *SACI 2004, 1st Romanian-Hungarian Joint Symposium on Applied Computational Intelligence*, 213-222.

Györödi R., Pavel M. I., Györödi C., & Zmaranda D. (2017). Performance of onPrem vs. azure sql server. a case study. *IEEE Access*, 7, 15894-15902.

Introduction to Table storage - Object storage in Azure (n.d.). Retrieved May 12, 2019, from <https://docs.microsoft.com/en-us/azure/storage/tables/table-storage-overview>.

Malik, M. I., Wani, S. H., & Rashid, A. (2018). Cloud computing-technologies. *International Journal of Advanced Research in Computer Science*, 9(2), 379-384.

Maven – Welcome to Apache Maven (n.d.). Retrieved April 20, 2019, from <https://maven.apache.org>.

Mohamed, W., Abdel-fattah, M. A., & El-Gaber, S. A. (2017). An implementation of eclat on spark. *International Journal of Computer Science and Information Security (IJCSIS)*, 15(6), 241-247.

Özçelik, G., & Kut A. (2017). Büyük veri analizinde veri madenciliği araçlarının performansı. *34. Ulusal Bilişim Kurultayı*, 53-57.

Quickstart: Use .NET to create a blob in object storage - Azure Storage (n.d.). Retrieved May 19, 2019, from <https://docs.microsoft.com/en-us/azure/storage/blobs/storage-quickstart-blobs-dotnet?tabs=windows>

Rai, P., & Singh, S. (2010). A survey of clustering techniques. *International Journal of Computer Applications*, 7(12), 1-5.

Rathee, S., Kaul, M., & Kashyap, A. (2015). R-Apriori: an efficient apriori based algorithm on spark. *Proceedings of the 8th Workshop on Ph. D. Workshop in Information and Knowledge Management*, 27-34.

Scott, J. A. (2015). *Getting started with Apache Spark* (1st ed.). United States of America: MapR Technologies.

Spring Framework – Overview (n.d.). Retrieved April 25, 2019, from https://www.tutorialspoint.com/spring/spring_overview.

Srivastava P., & Khan R. (2018). A review paper on cloud computing. *International Journals of Advanced Research in Computer Science and Software Engineering*, 8(6), 17-20.

Weka 3: Machine learning software in java (n.d.). Retrieved April 10, 2019, from <http://www.cs.waikato.ac.nz/ml/weka>.

What is .NET? (n.d.). Retrieved April 21, 2019, from <https://dotnet.microsoft.com/learn/dotnet/what-is-dotnet>.

Zeng, J. (2018). The development and application of data mining based on cloud computing. *First International Journal of Computer Science and Information Technologies*, 1087.