YEDİTEPE
ÜNİVERSİTESİ

T.C. YEDITEPE UNIVERSITY
FACULTY OF COMMERCIAL SCIENCES

DEPARTMENT OF MANAGEMENT INFORMATION SYSTEMS

GRADUATION THESIS

"BIG DATA TOOLS AND TECHNOLOGIES"

BY

BATUHAN ÖZYURT

20161314018

Submitted to Graduate Faculty of Commerce

in Partial Fulfillment of the Requirements

for the Degree of Bachelor in

Management Information Systems

HAZİRAN 2022

APPROVED BY

Asst. Prof. Dr. Engin Kandıran

ACKNOWLEDGMENT

Firstly, Yeditepe University Management Information Systems Department lecturer Res. Asst. I would like to thank Engin Kandiran. For his guidance, advice, patience, understanding and trust in me in completing my graduation thesis.

At the same time, Yeditepe University Management Information Systems department head dear Assoc. Dr. prof. Askin Demirağ, Lecturer Erhan Tümsa and Dr. Instructor I would like to thank its member, Elif Kartal, for they interest in data and data analysis in her lectures. In addition, thanks to his guidance and lectures throughout my education life, Dr. Instructor I would like to thank my professor Çağla Özen.

Finally, I would like to thank my brother, mother, father who have always supported and helped me.

ABSTRACT

Big data is a new driver of global economic and social change. Global data collection is reaching a tipping point for major technological changes that can open up new ways of making decisions, managing our health, our cities, finance and education. As data complexity increases, including the volume, variety, velocity, and accuracy of data, the real impact depends on our ability to discover the "value" of data through big data analytics techniques. Big data analytics poses significant challenges in designing highly scalable algorithms and systems to integrate data and discover enormous hidden value from diverse, complex, and large-scale datasets. Potential breakthroughs include new algorithms, methods, systems and applications in big data analysis that can effectively uncover useful and hidden knowledge from big data.

Big data analysis is closely related to Turkey, path towards a digital economy and society. Turkey has become a world leader in big data analysis, holding senior positions such as chairman and editor-in-chief in major conferences and journals in the field of big data. But to maintain such leadership, Turkey's universities, government and industry must move quickly to address a number of major challenges. These challenges include "Foundation," which involves new algorithms, theories, and methods for discovering knowledge from big data, and "Systems and Applications," which involves innovative applications and systems that help support big data practices. Big data analytics must also be a team effort across academic institutions, government, society and industry, and researchers from multiple disciplines, including computer science and engineering, health, data science, and society and politics.

Keywords: big data, data, analysis, big data technology, visualization, tools

ÖZET

Büyük veri, dünyadaki ekonomik ve toplumsal değişimlerin yeni itici gücüdür. dünyanın verileri koleksiyon, yenilikler getirebilecek büyük teknolojik değişiklikler için bir devrilme noktasına ulaşıyor. Karar verme, sağlığımızı yönetme, şehirler, finans ve eğitim yolları. İken Verinin hacmi, çeşitliliği, hızı ve doğruluğu dahil olmak üzere veri karmaşıklıkları artıyor, gerçek etki, Büyük Veri aracılığıyla verilerdeki "değeri" ortaya çıkarma yeteneğimize bağlıdır. Analitik teknolojileri. Büyük Veri Analitiği, tasarım konusunda büyük bir zorluk teşkil ediyor. verileri entegre etmek ve büyük gizli bilgileri ortaya çıkarmak için yüksek düzeyde ölçeklenebilir algoritmalar ve sistemler çeşitli, karmaşık ve büyük ölçekli veri kümelerinden elde edilen değerler. Potansiyel atılımlar, Big'deki yeni algoritmaları, metodolojileri, sistemleri ve uygulamaları içerir. Büyük Veriden faydalı ve gizli bilgileri verimli bir şekilde keşfeden Veri Analitiği ve etkili bir şekilde.

Big Data Analytics, dijital ekonomiye doğru ilerlerken Hong Kong ile ilgilidir ve toplum. Hong Kong, Büyük Veri Analitiğinde zaten dünyanın en iyileri arasında yer alıyor. önemli konferansların şeflerinde başkan ve editör olarak liderlik pozisyonlarını yükseltmek ve Büyük Veri ile ilgili alanlarda dergiler. Ancak bu tür liderlik pozisyonlarını sürdürmek için Hong Kong üniversiteleri, hükümeti ve endüstrisi, bir dizi soruna yönelik olarak hızlı hareket etmelidir. büyük zorluklar. Bu zorluklar, yeni konuları ilgilendiren “temelleri” içerir. büyük miktarlardaki bilgi keşfinde algoritmalar, teoriler ve metodolojiler yenilikçi uygulamalar ve sistemlerle ilgili veriler ve “sistemler ve uygulamalar” Büyük Veri uygulamalarını desteklemek için kullanışlıdır. Büyük veri analitiği de ekip çalışması olmalıdır akademik kurumlar, hükümet ve toplum ve endüstri arasında kesişen ve bilgisayar bilimi ve mühendisliği, sağlık, veri bilimi ve sosyal ve politika alanları.

Anahtar kelimeler: büyük veri, analiz, büyük veri teknolojisi, görselleştirme, araçlar.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

BI	Business Intelligence
XML	Extensible Markup Language
IOT	Internet of Things
GPS	Global Positioning System
IT	Information Technology
AMAZON EMR	Elastic Map Reduce
YARN	Yet Anohter Resource Negotiator
SQL	Structured Query Language
KPI	Key Performance Indicator
XML	Extensible Markup Language
JSON	JavaScript Object Natation
LAN	Local Area Network
HDFS	Hadoop Distributed File System
ACID	Atomicity, Consistency, Isolation, Durability

INTRODUCTION

The ability to leverage the ever-increasing amount of business data will provide insight into what is unfolding in the world. In this context, Big Data is a of the current main fashion words. Big Data (BD) is the technical term used. Reference to a large number of heterogeneous datasets created and generated processing, analysis of rapidly spreading and traditional techniques, Retrieving, storing and visualizing such large datasets is now convenient and insufficient. This can be seen in many areas such as sensor-generated data, social media. digital media upload and download. Advanced, unconventional and adaptive analytics are required to address these issues. challenges of managing and analyzing a wide variety of BD islands. It is expanding exponentially as a result of the large amount of data generated by tracking sensors, social media, transaction logs and metadata to name just a few It is one of many data sources. For example, 8,910 new Tweets, 89,845 GB internet traffic usage, 81,734 Google searches etc.Also, as shown in Figure 2.1, digital universe data is expected to increase 10-fold in the 2020s. This huge amount of data generated has implications for businesses and how they work In particular, business firms are extremely fond of providing information that encourages decision makers to adapt to rapid growth. in data volume. However, out of 85 percent of those aiming to become a data-driven company, only 37 percent proved successful. Most Information overload2 has made the decision-making process much more difficult. This led to the need for a shutdown. Reviewing internal business processes and tools used to collect Transferring, storing and analyzing the data stream generated by a company's internal and external sources.

CHAPTER 1

1. BİG DATA

1.2 Definition of Big Data

Big Data- in Turkish, Big Data- is a conception that first surfaced in the field of astronomy and genetics. The conception of big data has started to be used for the internet over time, so it has come a part of our diurnal life that we unconsciously enter and contribute to it constantly. although there are numerous dictionary delineations of big data, as the most explicatory description; “ The social media accounts that people use every day, hunt machines, the traces they leave before during their internet browsing and the huge data mound that brings together all the relations of individual druggies with the internet ” can be used.

As a end result of the reality that the laptop is so powerful in all regions of our lives, a whole lot of facts is saved, processed and managed. With the massive use of the Internet through companies, establishments and people, the circulation, processing and speedy proliferation of

this facts withinside the digital surroundings has found out some other end result. The facts we're speakme approximately consists of facts this is entered and saved as a demand of a service, in addition to a huge variety of facts that appear to be extraordinarily useless and useless, and those are developing like an avalanche at a completely excessive speed. With 2012 figures, 2.five quintillion bytes of facts are produced in an afternoon withinside the world. It is predicted that the entire facts length will attain forty four instances the modern time in 10 years. Thus, we see that a rubbish sell off such as unstructured facts has emerged. It did not take lengthy to understand that this phenomenon, which become formerly known as an facts sell off as it become now no longer based and couldn't be used, definitely contained a notable treasure. As a be counted of fact, this sell off such as facts along with net server logs, social media sharing and publications, blogs, micro blogs, net information ought to definitely be made very functional. If interpreted with the proper evaluation methods, those facts have to have made a completely critical contribution to creating very critical choices correctly, coping with dangers correctly, and signing new discoveries and discoveries.So what are the data components in big data?

There are five components in the big data platform. Since these components are expressed with 5 words starting with the letter V, they are called 5v.

5V that makes up the Big Data Basic dimensions:

- Volume
- Velocity
- Variety
- Veracity
- Value

In 2001, Gart analyst Doug Laney defines big data as higher-paced growth and more comprehensive datasets. And this definition has 3 basic dimensions. These are called 3V of big data. This is 3V: Volume (Volume), Velocity (Speed) and Variety. However, 2V has been added as the final (big data) dimensions and Value Veracity.

1.2.1 Volume

It refers to the huge amount of data stored. Traditional data is measured in sizes like megabytes, gigabytes, terabytes, while big data is stored in petabytes and zettabytes. To see

the difference, let's look at this comparison made by the University of California, Berkeley School of Information: 1 gigabyte is equivalent to 7 minutes of HD-TV video, while one zetabyte is equivalent to 250 billion DVDs.

According to the report by EMC, the size of the digital universe doubles every 2 years. Considering that the data is increasing day by day, considering the future developments, how to deal with these large data piles should be considered and plans should be made accordingly. Without suitable solutions for storage and processing, it will be very difficult to derive appropriate insights from this data.

1.2.2 Velocity

The continuous increase in the amount of data causes the number and variety of operations to be performed on the data to increase at the same rate. For data to be useful, companies and organizations must have the ability to leverage it in real time and gain insights simultaneously. While some of the data may be aggregated and remain relevant over time, most big data requires immediate action for the best results. Sensor data from health devices can be a good example in this regard. In order to get a positive result and prevent possible problems, it is necessary to have the ability to process health data instantly.

1.2.3 Variety

Big data; It covers structured, semi-structured and unstructured data. It's roughly 95% unstructured data and doesn't fit easily into the traditional model. The data produced is generally unstructured. Data from many different sources or languages, ranging from e-mails, videos, financial transactions and scientific data, lead to the emergence of different formats. To be meaningful and workable, they must be interchangeable.

1.2.4 Veracity

Today, thousands of people are sharing from different media. Or companies data data in many different formats. When the data is obtained and analyzed, the data must be real data for correct analysis. Big data has now started to become capital. Companies are in pursuit of producing new and need-oriented services. However, in order for this capital to be usable, it must consist of correct and real data.

1.2.5 Value

Recently, many companies have created their own data platforms. And analysis is made with the data obtained. And it is desired to obtain value with these analyzes.

1.3 History of Big Data

Although the idea of huge information is fairly new, the origins of massive datasets date returned to the 1960s-1970s, whilst the information global turned into simply rising with the improvement of the primary information facilities and databases.

Around 2005, human beings started out to recognize and have a look at how plenty information customers have been producing via Facebook, YouTube, and different on line services. In the identical year, Hadoop (an open supply gadget created mainly for storing and studying massive datasets) turned into developed. Meanwhile, NoSQL, a database application, started out to advantage popularity.

The improvement of open supply structures like Hadoop – and greater lately Spark – turned into vital to the evolution of huge information due to the fact they make it inexpensive to coordinate and shop huge information. Since then, the quantity of huge information has grown tremendously. Users are nonetheless growing massive quantities of information, however it might be a mistake to assume that simplest people do information creation.

The emergence of the Internet of Things (IoT) has ended in greater items and gadgets being related to the Internet via the gathering of information on purchaser utilization styles and product performance. In addition, the effect of gadget mastering on information technology continues to be noticeable.

Cloud computing has in addition accelerated the opportunities of huge information. The cloud computing surroundings gives a exceedingly bendy scalability platform wherein builders can without difficulty method brief units to check a dataset.

A Timeline of the Evolution of Big Data

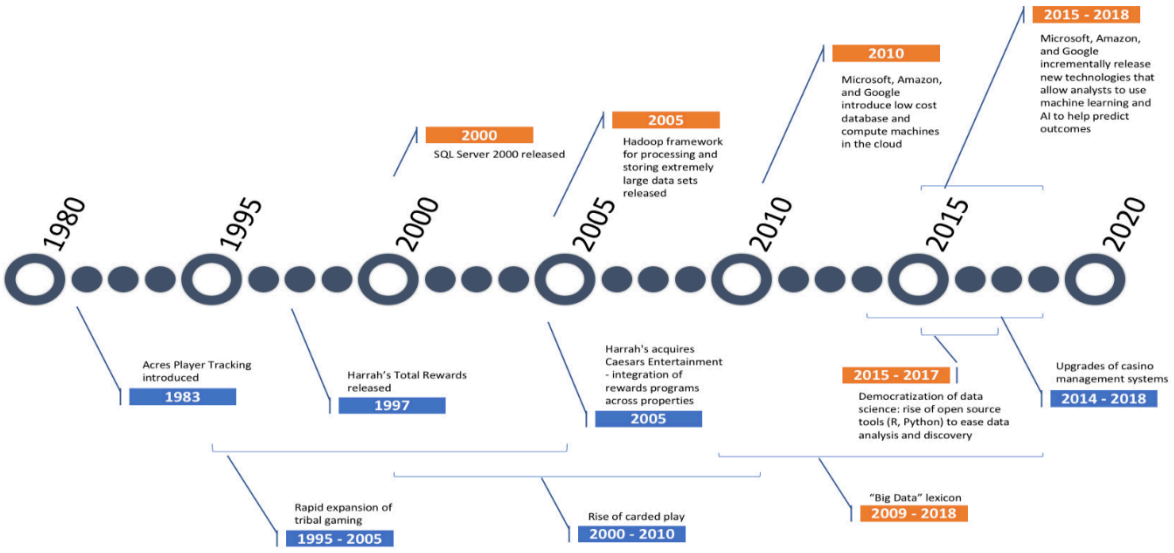


Figure 1 A Timeline of the Evolution of Big Data (Timeline of the Evolution)

1.4 Big Data Technology

1.4.1 Why is Big Data Important?

Companies use big data of their structures to enhance operations, offer higher boss service, create personalized advertising and marketing campaigns and take different movements that, ultimately, can growth sales and profits. Businesses that use it efficaciously keep a capability aggressive benefit over the ones that do not due to the fact they are capable of make quicker and extra knowledgeable enterprise decisions.

For example, big data offers treasured insights into clients that corporations can use to refine their advertising and marketing, marketing and marketing and promotions so one can growth boss engagement and conversion rates. Both ancient and real-time data may be analyzed to evaluate the evolving possibilities of purchasers or company buyers, permitting corporations to come to be extra attentive to patron wishes and needs.

Big data is likewise utilized by clinical researchers to pick out sickness symptoms and symptoms and threat elements and through docs to assist diagnose illnesses and clinical situations in. In addition, a mixture of data from digital fitness data, social media sites, the internet and different reassets offers healthcare groups and authorities groups updated data on infectious diseases threats or outbreaks.

Here are some more examples of how big data is used by organizations:

- In the energy industry, big data helps oil and gas companies identify potential drilling locations and monitor pipeline operations; likewise, utilities use it to track electrical grids.
- Financial services firms use big data systems for risk management and real-time analysis of market data.
- Manufacturers and transportation companies rely on big data to manage their supply chains and optimized delivery routes.
- Other government uses include emergency response, crime prevention and smart city initiatives.

1.4.2 The human side of big data management and analytics ?

Ultimately, the enterprise cost and blessings of big data tasks rely upon the people tasked with dealing with and reading the information. Some big data equipment permit much less technical customers to run predictive analytics programs or assist agencies set up a appropriate infrastructure for big data projects, whilst minimizing the want for hardware and allotted software program know-how.

Big information may be contrasted with small information, a time period it truly is from time to time used to explain information units that may be without problems used for self-carrier

BI and analytics. A typically quoted axiom is, "Big information is for machines; small information is for people."

1.4.2 How Big Data works?

1.4.2.1 Combining

Big data brings together the data obtained from different sources and applications. Any time for data such as correction, payment and scheduling is useless. Analyzing big data in a terabyte or even petabyte requires strategies and technologies. After the integration and processing of the data in the integration, it must be demonstrable in a way that the commissioned analysts can work with.

1.4.2.2 Management

Big data is stored. The storage process can occur from a major event within the system in the cloud. It is possible in both sides at the same time. You can apply your data in any way and request its necessary and necessary engines for actionable operations. The cloud storage system available is filling up for increasing popularity with many users choosing from storage options based on where their data resides.

1.4.2.3 Analyzing

Investing in great roads gives analysis and verda. Various data can be achieved by visualizing, more data to do new things, you can make your plan and share the machine with more data.

1.5 What is the Relationship between Big Data and Cloud?

Cloud computing is an net-primarily based totally data garage provider that lets in data sharing among all gadgets inclusive of computers, phones, pills and servers, no matter time and place.

Big data, on the alternative hand, is data produced over the years from special reassets or the equal source. The quantity of data that the laptop can not technique is known as large data or large data in English.

Big data management, which gives many benefits in particular in advertising and income activities, has a totally complicated shape in phrases of utility and technique.

We are in a virtual age in which the net of factors and large data have an effect on each component of life. Mobile gadgets, software program data, social media, cameras, microphones, in short, all our moves at the net are saved for processing within the data glide. In fact, it's miles foreseen that every one of our moves might be directed to servers as data glide withinside the very close to future.

The non-stop increase of data in phrases of size, variety and complexity, and that it'll retain to grow, has come to be an answer middle instead of a hassle with cloud computing. Especially with cloud computing possibilities disposing of the bounds of garage and computing power, the manner for large data has been cleared even more.

By allowing businesses to create the suitable surroundings to control and examine large quantities of data in actual time, cloud computing additionally gives the productiveness and enterprise flexibility vital for aggressive advantage.

Companies inclusive of Amazon and Walmart, which might be the various great customers of large data and cloud technology, have multiplied their income photographs to first-rate dimensions, in particular with the data they acquired from consumer alternatives on social networking sites.

It is plain that with the brand new technology to be developed, large data and cloud computing will retain to provide first-rate possibilities to establishments and customers.

1.6 Uses of Real-Time Big Data in Various Domains?

1.6.1 Big Data Applications

The number one intention of Big Data programs is to assist agencies make extra informative commercial enterprise choices with the aid of using studying massive volumes of facts. It should encompass internet server logs, Internet click on circulation facts, social media content material and pastime reports, textual content from client emails, cellular telecall smartphone name info and device facts captured with the aid of using a couple of sensors.

Organizations from specific area are making an investment in Big Data programs, for analyzing massive facts units to discover all hidden patterns, unknown correlations, marketplace trends, client possibilities and different beneficial commercial enterprise information. For example:

- Big Data Applications in Healthcare
- Big Data Applications in Manufacturing
- Big Data Applications in Media & Entertainment
- Big Data Applications in IoT
- Big Data Applications in Government

1.6.1 Big Data Applications in Healthcare

The degree of facts generated inside healthcare structures isn't always trivial. Traditionally, the fitness care enterprise lagged in the use of Big Data, due to confined cappotential to standardize and consolidate facts. But now Big facts analytics have progressed healthcare via way of means of imparting personalized medicinal drug and prescriptive analytics. Researchers are mining the facts to look what remedies are greater powerful for precise conditions, perceive styles associated with drug facet effects, and profits different critical data that may assist sufferers and decrease costs.

With the delivered adoption of mHealth, eHealth and wearable technology the extent of facts is growing at an exponential rate. This consists of digital fitness file facts, imaging facts, affected person generated facts, sensor facts, and different styles of facts.

By mapping healthcare facts with geographical facts sets, it is feasible to are expecting employment so one can strengthen in particular areas. Based of predictions, it is less difficult to strategize diagnostics and plan for stocking serums and vaccines.

1.6.2 Big Data Applications in Manufacturing

Predictive production offers near-0 downtime and transparency. It calls for an big quantity of statistics and superior prediction equipment for a scientific technique of statistics into beneficial information.

Goal for the goal of using Big Data for manufacturing:

- Product quality and tracking
- Supply
- Manufacturing defect
- Predictable
- Increased energy
- Test and simulation for new production
- Support for mass customization of production

1.6.3 Big Data Applications in Media & Entertainment

Various organizations withinside the media and amusement enterprise are dealing with new commercial enterprise models, for the manner they create, marketplace and distribute their content material. This is going on due to cutting-edge consumer's seek and the requirement of having access to content material anywhere, any time, on any device.

Big Data presents actionable factors of data approximately tens of thousands and thousands of individuals. Now, publishing environments are tailoring commercials and content material to enchantment consumers. These insights are accumulated through diverse data-mining activities.

Big Data applications benefits media and entertainment industry by:

- Predicting what the audience wants
- Scheduling optimization
- Increasing acquisition and retention
- name targeting
- Content monetization and new product development

1.6.4 Big Data Applications in Internet of Things

Data extracted from IoT devices provides a mapping of device inter-connectivity. Such mappings have been used by various companies and to increase efficiency. IoT is also increasingly adopted as a means of gathering sensory data, and this sensory data is used in medical and manufacturing contexts.

1.6.5 Big Data Applications in Internet of Things

Data extracted from IoT gadgets affords a mapping of tool inter-connectivity. Such mappings had been utilized by diverse corporations and to boom efficiency. IoT is likewise an increasing number of followed as a way of accumulating sensory facts, and this sensory facts is utilized in scientific and production contexts.

1.6.5.1 Cyber Security & Intelligence

The federal authorities released a cyber safety studies and improvement plan that is predicated at the capacity to investigate big facts units if you want to enhance the safety of U.S. pc networks. The National Geospatial-Intelligence Agency is developing a "Map of the World" which could collect and examine facts from a huge type of reassets consisting of satellite tv for pc and social media facts. It carries a whole lot of facts from classified, unclassified, and top-mystery networks.

1.6.5.2 Crime Prediction and Prevention

Police departments can leverage advanced, real-time analytics to offer actionable intelligence that may be used to recognize crook behavior, discover crime/incident patterns, and discover location-primarily based totally threats.

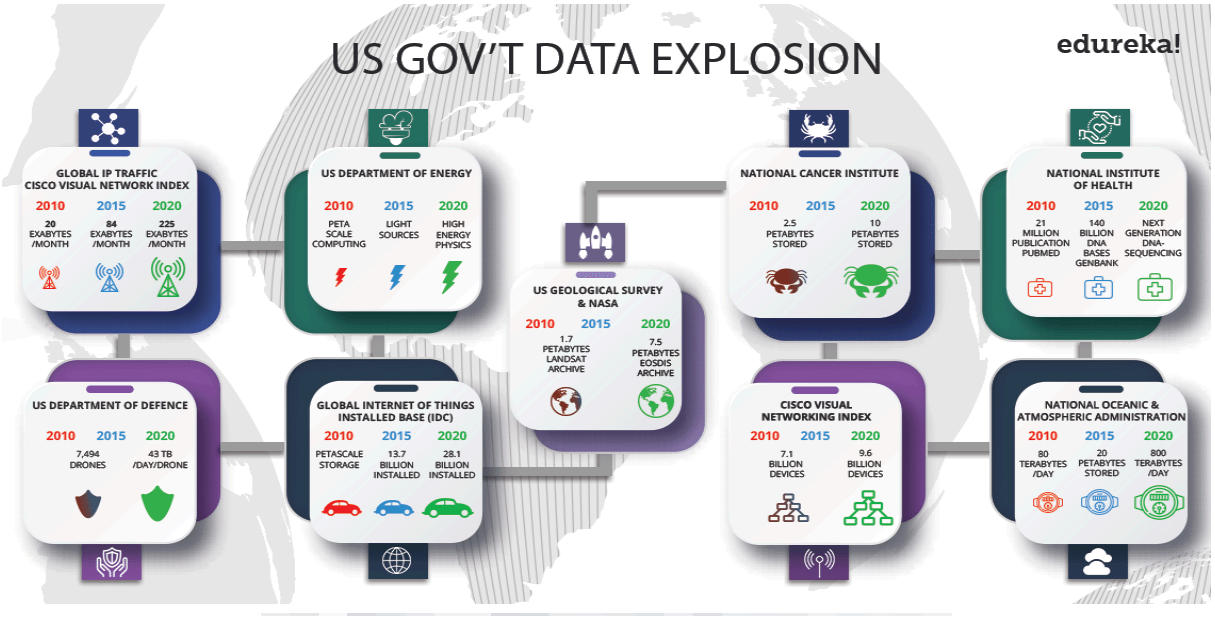
1.6.5.3 Pharmaceutical Drug Evaluation

According to a McKinsey report, Big Data technologies could reduce research and development costs for pharmaceutical makers by $40 billion to $70 billion. The FDA and NIH use Big Data technologies to access large amounts of data to evaluate drugs and treatment.

1.6.5.4 Scientific Research

The National Science Foundation has initiated a long-term plan to:

- Implement new methods for deriving knowledge from data
- Develop new approaches to education
- Create a new infrastructure to "manage, curate, and serve data to communities".



1.6.5.5 Weather Forecasting

The NOAA gathers information each minute of each day from land, sea, and space-primarily based totally sensors. Daily NOAA makes use of Big Data to research and extract price from over 20 terabytes of information.

1.6.5.6 Tax Compliance

Big Data Applications may be utilized by tax companies to investigate each unstructured and dependent facts from reassets which will become aware of suspicious conduct and more than one identities. This could assist in tax fraud identification.

1.6.5.7 Traffic Optimization

Big Data facilitates in aggregating actual-time visitors data collected from street sensors, GPS gadgets and video cameras. The ability visitors troubles in dense regions may be avoided via way of means of adjusting public transportation routes in actual time.

1.7 How Big Data Analytics works?

To get legitimate and applicable consequences from big data analytics applications, data scientists and different data analysts have to have an in depth knowledge of the to be had data and a feel of what they are seeking out in it. That makes data preparation, which incorporates profiling, cleansing, validation and transformation of data sets, a critical first step withinside the analytics process.

Once the data has been collected and organized for analysis, numerous data technology and superior analytics disciplines may be implemented to run extraordinary applications, the usage of gear that offer large data analytics capabilities and capabilities. Those disciplines encompass system mastering and its deep mastering offshoot, predictive modeling, data mining, statistical analysis, streaming analytics, textual content mining and more.

Using customer data as an example, the different branches of analytics that can be done with sets of big data include the following:

- Comparative analysis. This examines customer behavior metrics and real-time customer engagement in order to compare a company's products, services and branding with those of its competitors.
- Social media listening. This analyzes what people are saying on social media about a business or product, which can help identify potential problems and target audiences for marketing campaigns.
- Marketing analytics. This provides information that can be used to improve marketing campaigns and promotional offers for products, services and business initiatives.
- Sentiment analysis. All of the data that's gathered on customers can be analyzed to reveal how they feel about a company or brand, customer satisfaction levels, potential issues and how customer service could be improved.

1.8 Big Data Management Technologies

Hadoop, an open supply allotted processing framework launched in 2006, to begin with became on the middle of maximum massive data architectures. The improvement of Spark and different processing engines driven MapReduce, the engine constructed into Hadoop, extra to the side. The end result is an environment of massive data technology that may be used for unique packages however frequently are deployed together.

Big data platforms and managed services offered by IT vendors combine many of those technologies in a single package, primarily for use in the cloud. Currently, that includes these offerings, listed alphabetically:

- Amazon EMR
- Cloudera Data Platform
- Google Cloud Dataproc
- HPE Ezmeral Data Fabric (formerly MapR Data Platform)
- Microsoft Azure HDInsight

For organizations that want to deploy big data systems themselves, either on premises or in the cloud, the technologies that are available to them in addition to Hadoop and Spark include the following categories of tools:

- storage repositories, such as the Hadoop Distributed File System (HDFS) and cloud object storage services that include Amazon Simple Storage Service (S3), Google Cloud Storage and Azure Blob Storage;
- cluster management frameworks, like Kubernetes, Mesos and YARN, Hadoop's built-in resource manager and job scheduler, which stands for Yet Another Resource Negotiator but is commonly known by the acronym alone;
- stream processing engines, such as Flink, Hudi, Kafka, Samza, Storm and the Spark Streaming and Structured Streaming modules built into Spark;
- NoSQL databases that include Cassandra, Couchbase, CouchDB, HBase, MarkLogic Data Hub, MongoDB, Neo4j, Redis and various other technologies;
- data lake and data warehouse platforms, among them Amazon Redshift, Delta Lake, Google BigQuery, Kylin and Snowflake; and
- SQL query engines, like Drill, Hive, Impala, Presto and Trino.

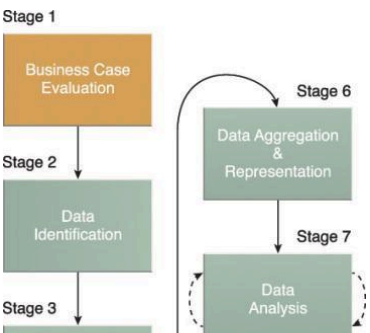
1.9 Big Data Analytics Lifecycle

Big Data evaluation differs from conventional information evaluation frequently because of the volume, speed and range traits of the information being processes. To deal with the awesome necessities for appearing evaluation on Big Data, a step by step technique is wanted to prepare the sports and obligations worried with acquiring, processing, studying and repurposing information. The upcoming sections discover a particular information analytics lifecycle that organizes and manages the obligations and sports related to the evaluation of Big Data. From a Big Data adoption and making plans perspective, it's miles vital that similarly to the lifecycle, attention be made for problems of training, education, tooling and staffing of a information analytics team.

The Big Data Analytics Lifecycle is divided into nine phases, named as follows:

- Business Case Evaluation
- Data Identification
- Data Acquisition & Filtering
- Data Extraction
- Data Validation & Cleansing
- Data Aggregation & Representation
- Data Analysis
- Data Visualization
- Utilization of Analysis Results

1.9.1 Business Case Evalutaion



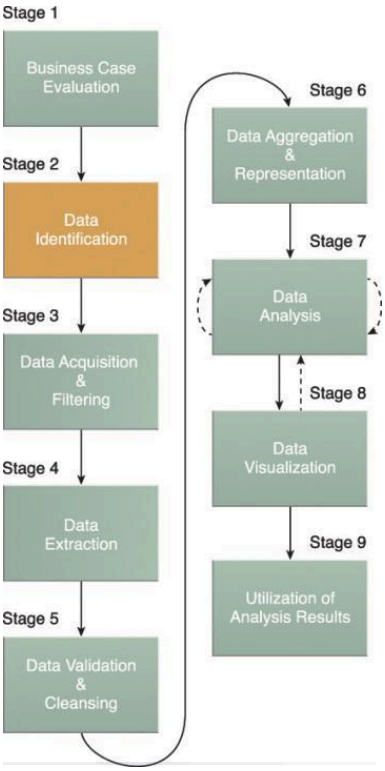
Each Big Data analytics lifecycle should start with a well-described commercial enterprise case that offers a clean information of the justification, motivation and desires of

sporting out the evaluation. The Business Case Evaluation level proven in Figure four calls for that a commercial enterprise case be created, assessed and accredited previous to intending with the real hands-on evaluation tasks. An assessment of a Big Data analytics commercial enterprise case allows decision-makers recognize the commercial enterprise sources as a way to want to be applied and which commercial enterprise demanding situations the evaluation will tackle. The similarly identity of KPIs throughout this level can assist decide evaluation standards and steerage for the assessment of the analytic results. If KPIs aren't with no trouble available, efforts must be made to make the desires of the evaluation undertaking SMART, which stands for specific, measurable, attainable, applicable and timely.

Based on commercial enterprise necessities which can be documented withinside the commercial enterprise case, it is able to be decided whether or not the commercial enterprise issues being addressed are definitely Big Data issues. In order to qualify as a Big Data hassle, a commercial enterprise hassle wishes to be at once associated with one or greater of the Big Data traits of volume, velocity, or variety.

Note additionally that some other final results of this level is the dedication of the underlying finances required to perform the evaluation undertaking. Any required purchase, along with tools, hardware and education, should be understood earlier in order that the predicted funding may be weighed in opposition to the anticipated advantages of accomplishing the desires. Initial iterations of the Big Data analytics lifecycle would require greater up-the front funding of Big Data technologies, merchandise and education in comparison to later iterations in which those in advance investments may be again and again leveraged.

1.9.2 Data Identification



The Data Identification stage shown in Figure 5 is dedicated to identifying the datasets required for the analysis project and their sources.

Identifying a wider variety of data sources may increase the probability of finding hidden patterns and correlations. For example, to provide insight, it can be beneficial to identify as many types of related data sources as possible, especially when it is unclear exactly what to look for.

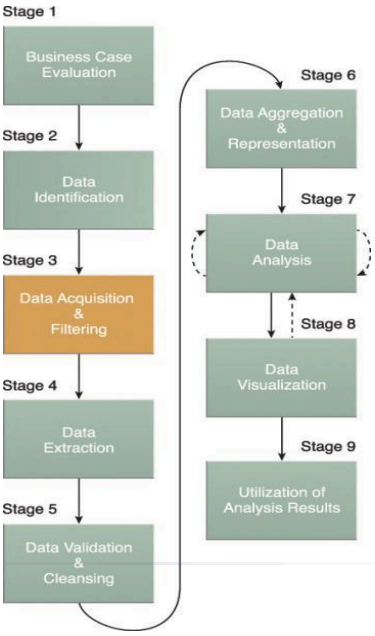
Depending on the business scope of the analysis project and nature of the business problems being addressed, the required datasets and their sources can be internal and/or external to the enterprise.

In the case of internal datasets, a list of available datasets from internal sources, such as data marts and operational systems,

are typically compiled and matched against a pre-defined dataset specification.

In the case of external datasets, a list of possible third-party data providers, such as data markets and publicly available datasets, are compiled. Some forms of external data may be embedded within blogs or other types of content-based websites, in which case they may need to be harvested via automated tools.

1.9.3 Data Acquisition and Filtering

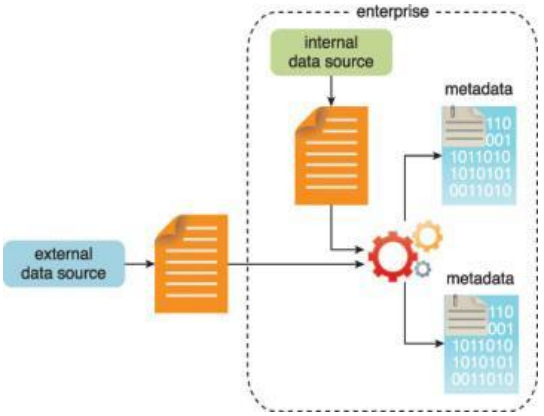


The Data Acquisition and Filtering stage, proven in Figure 6, the information is accrued from all the information reassets that have been diagnosed at some point of the preceding stage. The obtained information is then subjected to automatic filtering for the elimination of corrupt information or information that has been deemed to don't have any fee to the evaluation objectives.Depending at the form of information source, information can also additionally come as a group of files, inclusive of information bought from a third-celebration information many cases, mainly wherein outside, unstructured information is concerned, a few or maximum of the obtained information can be irrelevant (noise) and may be discarded as a part of the filtering process.

Data categorized as “corrupt” can encompass information with lacking or nonsensical values or invalid information types. Data this is filtered out for one evaluation can also additionally probable be precious for a distinct form of evaluation. Therefore, it's far really useful to save a verbatim replica of the unique dataset earlier than intending with the filtering. To reduce the specified garage space, the verbatim replica may be compressed.

Both inner and outside information wishes to be continued as soon as it receives generated or enters the organizational boundary. For batch analytics, this information is continued to disk previous to evaluation. In the case of realtime analytics, the information is analyzed first after which continued to disk.

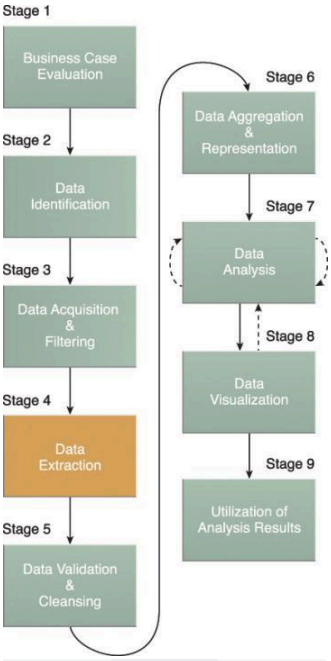
As evidenced in Figure 7, metadata may be delivered thru automation to facts from each inner and outside facts reassets to enhance the class and querying. Examples of appended metadata encompass dataset length and structure, supply information, date and time of introduction or series and language-particular information. It is essential that metadata be machine-readable and



handed ahead alongside next evaluation stages. This facilitates hold facts provenance for the duration of the Big Data analytics lifecycle, which facilitates to set up and keep facts accuracy and quality.

1.9.4 Data Extraction

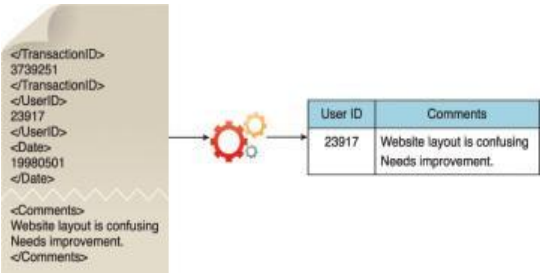
Some of the facts recognized as enter for the evaluation might also additionally arrive in a layout incompatible with the Big Data answer. The want to cope with disparate styles of facts is much more likely with facts from outside sources. The Data Extraction lifecycle stage, proven in Figure 8, is devoted to extracting disparate facts and remodeling it right into a layout that the underlying Big Data answer can use for the motivation of the facts evaluation.



The extent of extraction and transformation required depends on the types of analytics and capabilities of the Big Data solution. For example, extracting the required fields from delimited textual data, such as with webserver log files, may not be necessary if the underlying Big Data solution can already directly process those files.

Similarly, extracting text for text analytics, which requires scans of whole documents, is simplified if the underlying Big Data solution can directly read the document in its native format.

Figure 9 illustrates the extraction of feedback and a consumer ID embedded inside an XML data with out the want for similarly transformation.



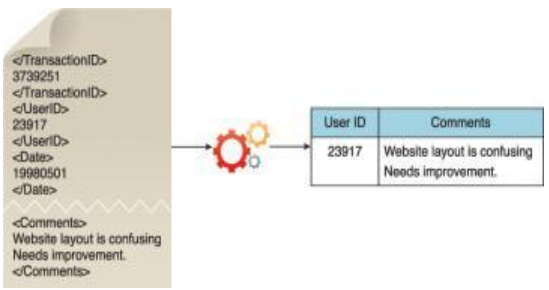
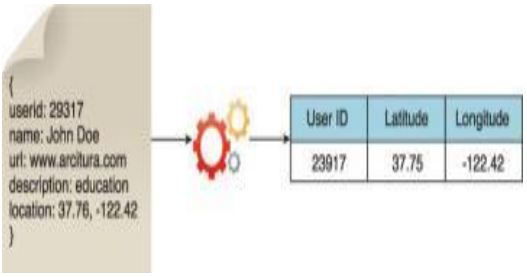
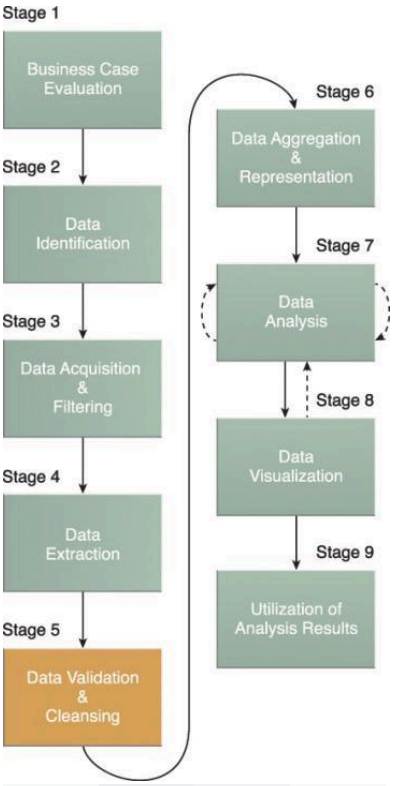


Figure 10 demonstrates the extraction of the range and longitude coordinates of a person from an unmarried JSON field.



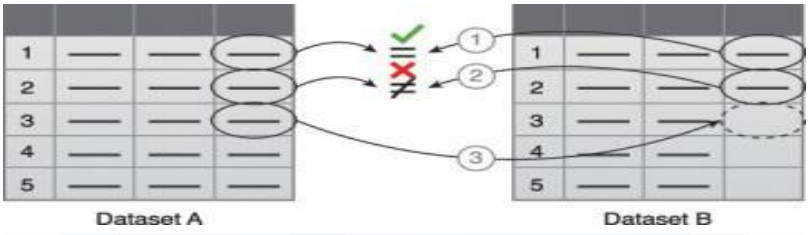
Further transformation is wanted with a purpose to separate the information into separate fields as required through the Big Data solution.

1.9.5 Data Validation and Cleansing



Invalid information can skew and falsify evaluation results. Unlike conventional agency information, in which the information shape is pre-described and information is pre-validated, information enter into Big Data analyzes may be unstructured with none indication of validity. Its complexity can in addition make it tough to reach at a fixed of appropriate validation constraints. The Data Validation and Cleansing degree proven in Figure 12 is devoted to setting up regularly complicated validation policies and casting off any recognized invalid statistics.

Big Data answers regularly acquire redundant statistics at



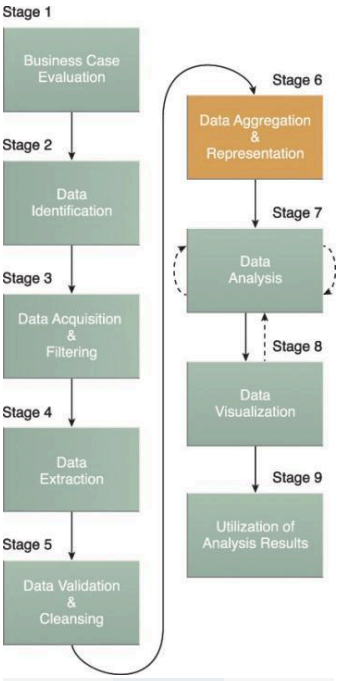
some stage in precise datasets. This redundancy can be exploited to find out interconnected datasets on the way to accumulate validation parameters and fill in missing valid statistics. For example, as illustrated in Figure 13 :

- The first value in Dataset B is validated against its corresponding value in Dataset A.
- The second value in Dataset B is not validated against its corresponding value in Dataset A.
- If a value is missing, it is inserted from Dataset A

For batch analytics, facts validation and cleaning may be completed thru an offline ETL operation. For realtime analytics, a greater complicated in-reminiscence device is needed to validate and cleanse the facts because it arrives from the source. Provenance can play an critical position in figuring out the accuracy and excellent of questionable facts. Data that looks to be invalid can also additionally nonetheless be treasured in that it could own hidden styles and trends, as proven in Figure 14.

1.9.6 Data Aggregation and Representation

Data can be unfold throughout a couple of datasets, requiring that datasets be joined collectively thru not unusualplace fields, as an instance date or ID. In different cases, the equal statistics fields can also additionally seem in a couple of datasets, which includes date of birth. Either way, a technique of statistics reconciliation is needed or the dataset representing the suitable price wishes to be determined.



The Data Aggregation and Representation stage, shown in Figure 15, is dedicated to integrating multiple datasets together to arrive at a unified view.

Performing this degree can end up complex due to variations in:

Data Structure – Although the statistics layout can be the identical, the statistics version can be unique.

Semantics – A cost this is categorized in another way in unique datasets may also imply the identical thing, for example "surname" and "final name."

The big volumes processed via way of means of Big Data answers could make statistics aggregation a time and effort-in depth operation. Reconciling those variations can require complicated good judgment this is finished mechanically with out the want for human intervention.

Future statistics evaluation necessities want to be taken into consideration at some stage in this degree to assist foster statistics reusability. Whether statistics aggregation is needed or not, it's far critical to recognize that the identical statistics may be saved in lots of unique forms. One shape can be higher acceptable for a selected kind of evaluation than another. For example, statistics saved as a BLOB could be of little use if the evaluation calls for get entry to person statistics fields.

A data shape standardized with the aid of using the Big Data answer can act as a not unusual place denominator that may be used for a variety of evaluation strategies and projects. This can require setting up a central, popular evaluation repository, inclusive of a NoSQL database, as proven in Figure 16.

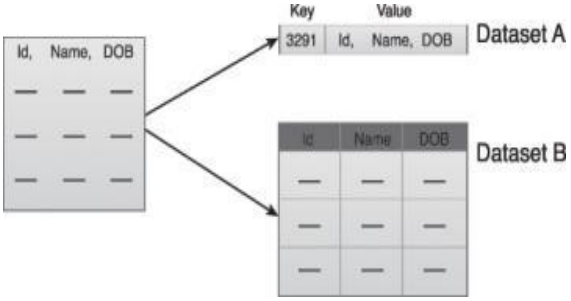
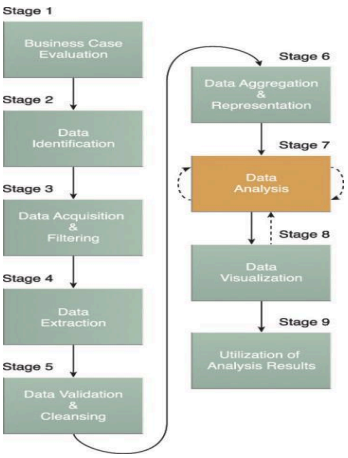


Figure 17 suggests the equal piece of statistics saved in special formats. Dataset A consists of the favored piece of statistics, however it's far a part of a BLOB that isn't always effortlessly

available for querying. Dataset B consists of the equal piece of statistics prepared in column-primarily based totally storage, permitting every discipline to be queried individually.

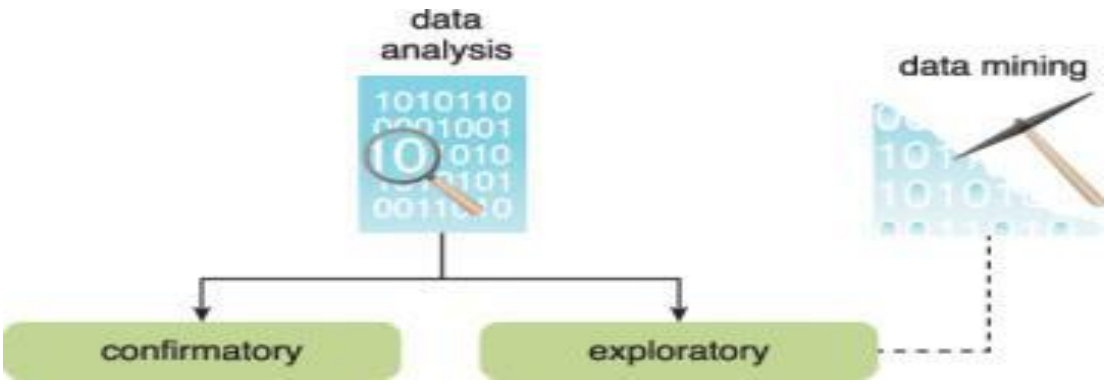
1.9.7 Data Analysis



The Data Analysis stage shown in Figure 18 is dedicated to carrying out the actual analysis task, which typically involves one or more types of analytics. This stage can be iterative in nature, especially if the data analysis is exploratory, in which case analysis is repeated until the appropriate pattern or correlation is uncovered. The exploratory evaluation method can be defined shortly, in conjunction with confirmatory evaluation.

Depending at the form of analytical end result required, this degree may be as easy as querying a dataset to compute an aggregation for comparison. On the opposite hand, it may be as tough as combining facts mining and complicated statistical evaluation strategies to find out styles and anomalies or to generate a statistical or mathematical version to depict relationships among variables.

Data analysis can be classified as confirmatory analysis or exploratory analysis, the latter of which is linked to data mining, as shown in Figure 19.



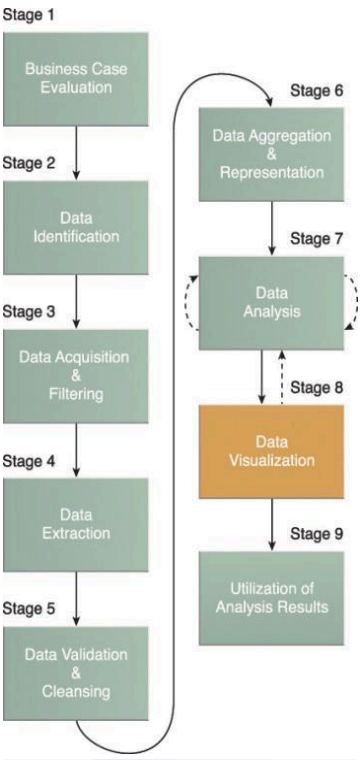
Confirmatory information evaluation is a deductive method wherein the reason of the phenomenon being investigated is proposed beforehand. The proposed reason or assumption is referred to as a speculation. The information is then analyzed to show or disprove the speculation and offer definitive solutions to particular questions.

Data sampling strategies are generally used. Unexpected findings or anomalies are typically overlooked in view that a predetermined reason turned into assumed. Exploratory information evaluation is an inductive method this is carefully related to information mining. No speculation or predetermined assumptions are generated. Instead, the information is explored

via evaluation to increase an knowledge of the reason of the phenomenon. Although it could now no longer offer definitive solutions, this technique offers a trendy path that may facilitate the invention of styles or anomalies.

1.9.8 Data Visualization

The capacity to research large quantities of data and discover beneficial insights incorporates little fee if the best ones which can interpret the consequences are the analysts.



The Data Visualization stage, proven in Figure 20, is devoted to the usage of information visualization strategies and equipment to graphically speak the evaluation consequences for powerful interpretation through enterprise users.

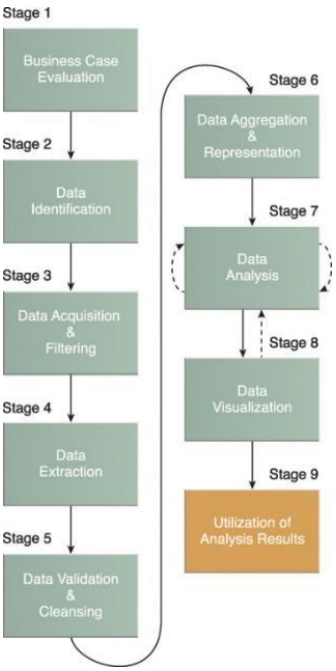
Business customers want so as to apprehend the outcomes for you to gain cost from the evaluation and finally have the capacity to offer feedback, as indicated via way of means of the dashed line main from level eight again to level 7.

The outcomes of finishing the Data Visualization level offer customers with the to carry out visible evaluation, taking into account the invention of solutions to questions that customers have now no longer but even capacity. Visual evaluation strategies are protected later on this book.

The identical outcomes can be provided in some of distinct ways, that may impact the translation of the outcomes. Consequently, it's far essential to apply the maximum appropriate visualization method via way of means of maintaining the commercial enterprise area in context.

Another aspect to keep in mind is that providing a method of drilling down to comparatively simple statistics is crucial, in order for users to understand how the rolled up or aggregated results were generated.

1.9.9 Utilization of Analysis Results



Subsequent to evaluation consequences being made to be had to enterprise customers to aid enterprise decision-making, which include through dashboards, there can be in addition possibilities to make use of the evaluation consequences. The Utilization of Analysis Results stage, proven in Figure 3.23, is devoted to figuring out how and in which processed evaluation information may be in addition leveraged.

Depending on the character of the evaluation issues being addressed, it's far viable for the evaluation consequences to produce “models” that encapsulate new insights and understandings approximately the character of the styles and relationships that exist inside the information that turned into analyzed. A version might also additionally appear like a mathematical equation or a hard and fast of rules. Models may be used to enhance enterprise manner common sense and alertness machine common sense, and they are able to shape the idea of a brand new machine or software program program.Common areas that are explored during this stage include the following:

- Input for Enterprise Systems: The statistics evaluation outcomes can be routinely or manually fed at once into employer structures to beautify and optimize their behaviors and performance. For example, an internet keep may be fed processed customer-associated evaluation outcomes that could effect the way it generates product recommendations. New fashions can be used to enhance the programming good judgment inside present employer structures or might also additionally shape the premise of latest structures.
- Business Process Optimization: The identified patterns, correlations and anomalies discovered during the data analysis are used to refine business processes. An example is consolidating transportation routes as part of a supply chain process. Models may also lead to opportunities to improve business process logic.
- Alerts: Data analysis results can be used as input for existing alerts or may form the basis of new alerts. For example, alerts may be created to inform users via email or SMS text about an event that requires them to take corrective action

CHAPTER 2

2. BIG DATA TOOLS AND VISULATION

2.1 Big Data Tools

Big data is certainly too big and complicated statistics that can not be handled the usage of conventional statistics processing methods.

Big Data calls for a fixed of gear and strategies for evaluation to advantage insights from it.

There are some of huge statistics gear to be had withinside the marketplace together with Hadoop which enables in storing and processing big statistics, Spark enables in-reminiscence calculation, Storm enables in quicker processing of unbounded statistics, Apache Cassandra gives excessive availability and scalability of a database , MongoDB gives cross-platform capabilities, so there are specific capabilities of each Big Data tool

Imagine you being on the pinnacle of the sport withinside the area of Big Data and your enterprise on cloud nine, much like Sachin Tendulkar in the sport of cricket.

The solution is an excellent set of Big Data gear.

Analyzing and processing Big Data isn't a clean task. Big Data is one huge hassle and to address it you want a fixed of remarkable huge statistics gear with a purpose to now no longer most effective resolve this hassle however additionally assist you in generating considerable results.

2.2 Top 10 Big Data Tools

- Apache Hadoop
- Apache Spark
- Flink
- Apache Storm
- Apache Cassandra
- MongoDB
- Kafka
- Tableu
- RapidMiner
- R Programming

Big Data is an crucial a part of nearly each company in recent times and to get great consequences thru Big Data Analytics a fixed of gear is wanted at every segment of statistics processing and evaluation. There are some elements to be taken into consideration whilst choosing the set of gear i.e., the dimensions of the datasets, pricing of the tool, type of evaluation to be done, and plenty of more. With the exponential increase of Big Data, the marketplace is likewise flooded with its diverse gear. These gear utilized in large statistics assist in bringing out higher value performance and therefore will increase the velocity of evaluation.

2.2.1 Apache Hadoop

Apache Hadoop is one of the most used equipment in the Big Data industry. Hadoop is an open source framework from Apache and runs on commercial hardware. It is used to hold procedure and examine Big Data.

Hadoop is written in Java. Apache Hadoop allows parallel processing of statistics as it runs on several machines at the same time. It uses clustered architecture. A cluster is a set of structures that can be associated via LAN.

It consists of 3 parts:

- Hadoop Distributed File System (HDFS) – It is the storage layer of Hadoop.
- Map-Reduce – It is the data processing layer of Hadoop.
- YARN – It is the resource management layer of Hadoop.

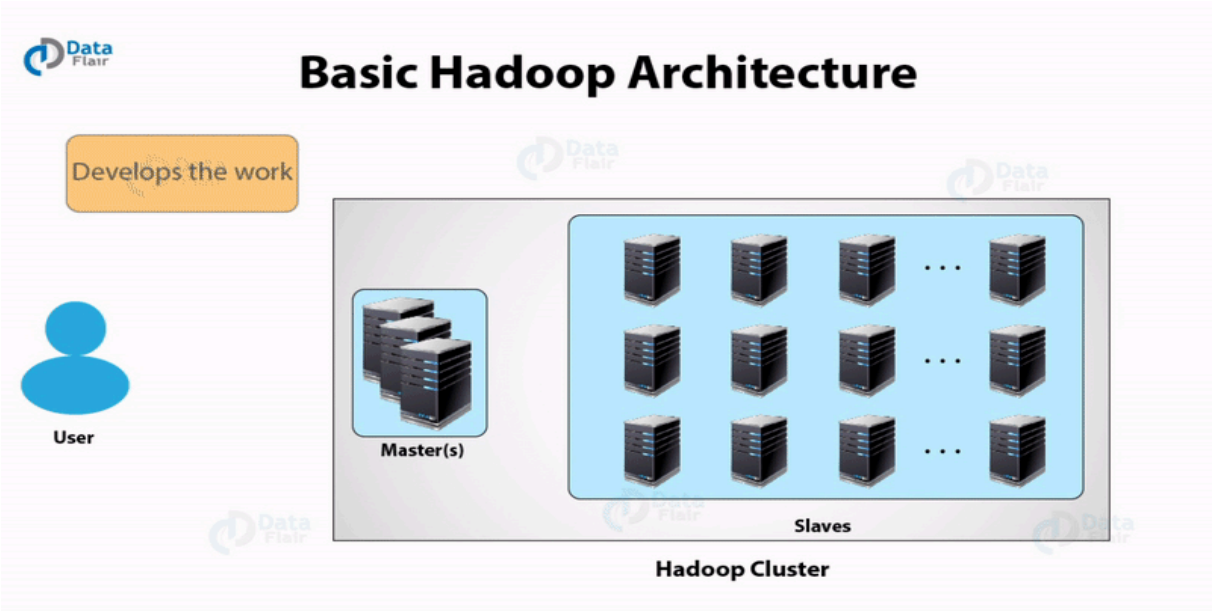


Figure 22 Basic Hadoop Architecture

Everything that has evolved also comes with a few dangers. Here are some about Hadoop:

- Hadoop does not support real-time processing. It only supports batch processing.
- Hadoop cannot do in-memory calculations.

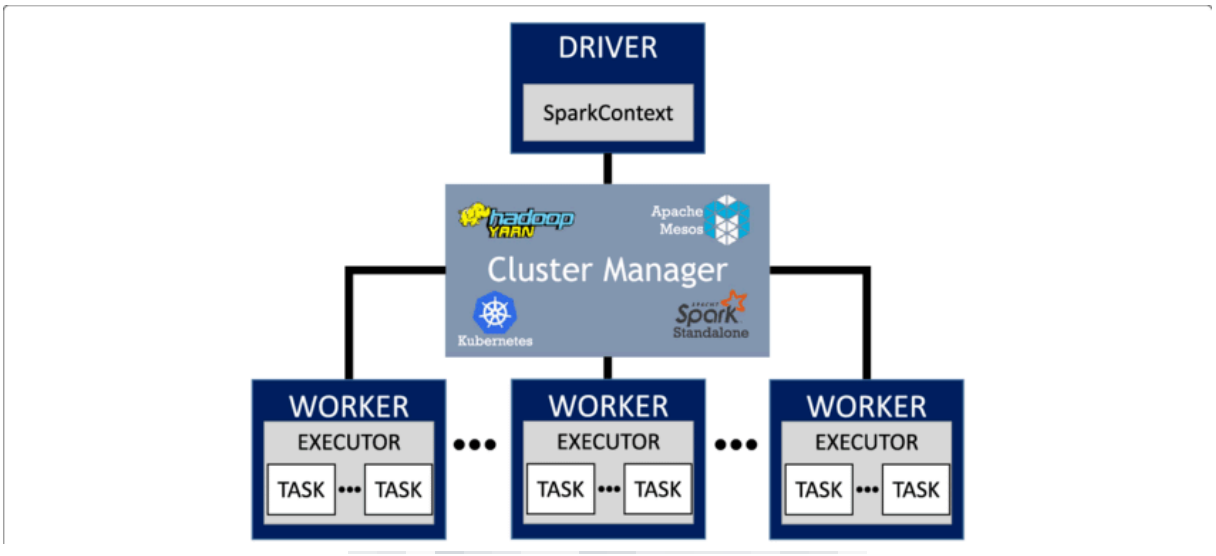
2.2.2 Apache Spark

Apache Spark may be taken into consideration because the successor of Hadoop because it overcomes the drawbacks of it. Spark, in contrast to Hadoop, helps each real-time in addition to batch processing. It is a general-cause clustering system.

It additionally helps in-reminiscence calculations, which makes it one hundred instances quicker than Hadoop. This is made feasible via way of means of lowering the wide variety of read/write operations into the disk.

It offers greater flexibility and flexibility in comparison to Hadoop because it works with distinctive information shops inclusive of HDFS, OpenStack and Apache Cassandra.

I It gives high-stage APIs in Java, Python, Scala and R. Spark additionally gives a full-size set of high-stage gear consisting of Spark SQL for dependent statistics processing, MLlib for system learning, GraphX for graph statistics set processing, and Spark Streaming. It additionally includes eighty high-stage operators for green question execution.



2.2.2 Apache Storm

Apache Storm is an open-supply large facts tool, disbursed real-time and fault-tolerant processing system. It techniques successfully unbounded streams of facts.

By unbounded streams, we check with the facts this is ever-developing and has a starting however no described end. The largest benefit of Apache Storm is that it is able to be used with any of the programming languages and it similarly helps JSON primarily based totally

protocols. The processing pace of Storm could be too excessive. It is without difficulty scalable and additionally fault-tolerant. It is a good deal less complicated to use.

On the alternative hand, it ensures the processing of every facts set. It's processing pace is fast and a widespread found became as excessive as 1,000,000 tuples processed in keeping with 2nd on every node.

2.2.4 Apache Cassandra

With out compromising overall performance efficiency. It is one of the satisfactory large statistics gear that could accommodate all styles of statistics units particularly structured, semi-structured, and unstructured.

It is the ideal platform for mission-crucial statistics and not using a unmarried factor of failure and presents fault tolerance on each commodity hardware and cloud infrastructure.

Cassandra works pretty correctly beneathneath heavy loads. It does now no longer comply with master-slave structure so all nodes have the equal role. Apache Cassandra helps the ACID properties.

2.2.5 MongoDB

MongoDB is an open-supply data analytics tool, NoSQL database that gives cross-platform capabilities. It is exemplary for a enterprise that desires fast-shifting and real-time data for taking decisions.

MongoDB is best for folks who need data-pushed solutions. It is user-pleasant because it gives simpler set up and maintenance. MongoDB is dependable in addition to cost-effective

It is written in C, C++, and JavaScript. It is one of the maximum famous databases for Big Data because it helps the control of unstructured data or the data that modifications frequently.

MongoDB makes use of dynamic schemas. Hence, you may put together data quickly. This permits in decreasing the general cost. It executes on MEAN software program stack, NET packages and, Java platform. It is likewise bendy in cloud infrastructure.

But a positive downfall within the processing pace has been observed for a few use-cases.

2.2.6 Apache Flink

Apache Flink is an Open-supply data analytics device disbursed processing framework for bounded and unbounded data streams. It is written in Java and Scala. It gives excessive accuracy outcomes even for late-arriving data.

Flink is a stateful and fault-tolerant i.e. it has the capacity to get over faults easily. It gives excessive-overall performance performance at a huge scale, acting on hundreds of nodes. It offers a low-latency, excessive throughput streaming engine and helps occasion time and kingdom management.

2.2.7 Kafka

Apache Kafka is an open-supply platform that turned into created via way of means of LinkedIn withinside the 12 months 2011.

Apache Kafka is a disbursed occasion processing or streaming platform which gives excessive throughput to the systems. It is green sufficient to address trillions of occasions a day. It is a streaming platform this is pretty scalable and additionally gives first-rate fault tolerance.

The streaming manner consists of publishing and subscribing to streams of information alike to the messaging systems, storing those information durably, after which processing those information. These information are saved in companies referred to as topics.

Apache Kafka gives excessive-pace streaming and ensures 0 downtime.

2.2.8 Tableau

Tableau is one of the pleasant facts visualization and software program answer gear within the Business Intelligence industry. It's a device that unleashes the electricity of your facts.

It turns your uncooked facts into treasured insights and improving the decision-making system of the businesses.

Tableau gives a fast facts evaluation system and ended in visualizations are withinside the shape of interactive dashboards and worksheets.

It works in synchronization with other Big Data tools such as Hadoop

- Tableau offered the capabilities of data blending are best in the market. It provides an efficient real-time analysis.
- Tableau is not only bound to the technology industry but is a crucial part of some other industries as well. This software doesn't require any technical or programming skills to operate.

2.2.9 RapidMiner

RapidMiner is a cross-platform device that gives a sturdy surroundings for Data Science, Machine Learning and Data Analytics procedures. It is an incorporated platform for the entire

Data Science lifecycle beginning from facts prep to gadget mastering to predictive version deployment.

It gives diverse licenses for small, medium, and massive proprietary editions. Apparently, it additionally gives a loose version that allows simplest 1 logical processor and as much as 10,000 facts rows.

RapidMiner is an open-supply device this is written in java. RapidMiner gives excessive performance even if incorporated with APIs and cloud services. It gives a few sturdy Data Science equipment and algorithms.

2.2.10 R Programming

R is an open-supply programming language and is one of the maximum complete statistical evaluation languages. It is a multi-paradigm programming language that gives a dynamic improvement environment. As it's far an open-supply mission and lots of human beings have contributed to the improvement of the R.

R is written in C and Fortran. It is one of the maximum extensively used statistical evaluation gear because it presents a full-size package deal ecosystem.

It allows the green overall performance of various statistical operations and allows in producing the outcomes of data evaluation in graphical in addition to textual content format. The images and charting advantages it presents are unmatchable.

2.3 Big Data Visualization

Data visualization is the presentation of the evaluation effects of summary statistics via way of means of developing graphs, diagrams, tables, photographs or animations. The motivation of visualization is to make the complicated statistics offered in classical layout of statistical and variable statistics clean to recognize with without difficulty perceptible graphical interfaces. The human belief gadget is limited to a few dimensions and better dimensional statistics systems cross past the boundaries of human belief. Thanks to visualization, multidimensional statistics is made greater comprehensible via way of means of decreasing the scale to two or three dimensions. Research indicates that human beings reply higher to visuals than to another stimulus. The human mind techniques visible statistics 60,000 instances quicker than text. In fact, visible statistics makes up ninety percentage of the statistics transmitted to the mind.

Today, main media retailers including The New York Times use superior data evaluation and modern pc photographs to offer information insights to dazzle readers with infographics and save you them from getting slowed down withinside the text. Datasets furnished via way of means of America federal authorities and enterprise reassets permit people to create and

proportion graphical interpretations of data on their personal on blogs and social networks. Psychologists and neurosurgeons have notably studied how humans reply to photo stimuli and use short- and long-time period reminiscence to elicit preceding reports in processing statistics. These research have treated expert humans from all walks of life, including doctors, pilots, monetary offerings specialists, regulation enforcement officials and army personnel, and because of the research, it's been visible that visualization significantly helps notion and understanding. However, acquiring new statistics to make selections and broaden techniques has emerge as even extra vital with data visualization. Colin Ware; In his book "Visualizing Knowledge", he states, "We have received extra statistics with visualization than with every other sense". “The 20 billion or so neurons the mind makes use of to investigate visible statistics shape a pattern-locating mechanism that could be a essential element of a lot of our cognitive activity,” he says. Today, programmers and scientists are searching out new visualization tools. For example, corporations including Amazon, Twitter, Apple, Facebook and Google use data visualization to make suitable commercial enterprise selections. These corporations pay billions of bucks to the Business Intelligence enterprise to supply analytical reviews and data visualizations. With Business Intelligence, corporations use any sort of data visualization that offers statistics to make a commercial enterprise selection or take an action. There are predominant motivations why BI is vital in data visualization: First, it visualizes the relationships and hidden systems among operational and commercial enterprise activities. In this way, it's miles feasible to look the connection among the operating situations and paintings overall performance of the customers extra clearly. In ultra-modern noticeably aggressive commercial enterprise environments, it's miles extra vital than ever to discover those relationships amongst data. Second, data visualization is used prominently in measuring consumer delight via way of means of presenting a flexible angle on commercial enterprise and running dynamics. Today, many strategies are utilized in data visualization.

The most commonly used of these techniques are:

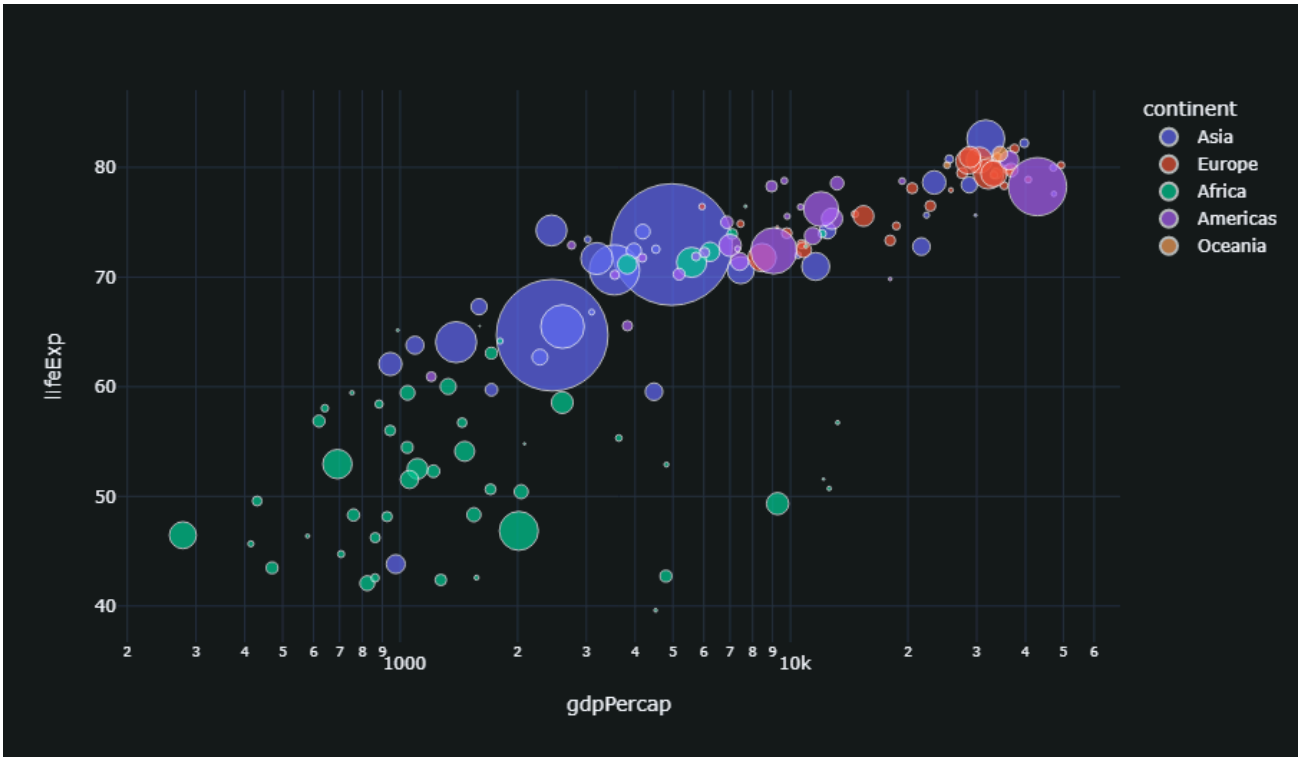
- Motion charts allow effective analysis of large and multivariate data. These charts interact using dynamic two-dimensional bubble diagrams (See Figure 23). Due to the designed variable mapping, blobs (bubbles - central objects of this technique) can be controlled. Motion graphics are provided by graphical data tools, Google, amCharts, and IBM Many Eyes (Olshannikova et al., 2015).

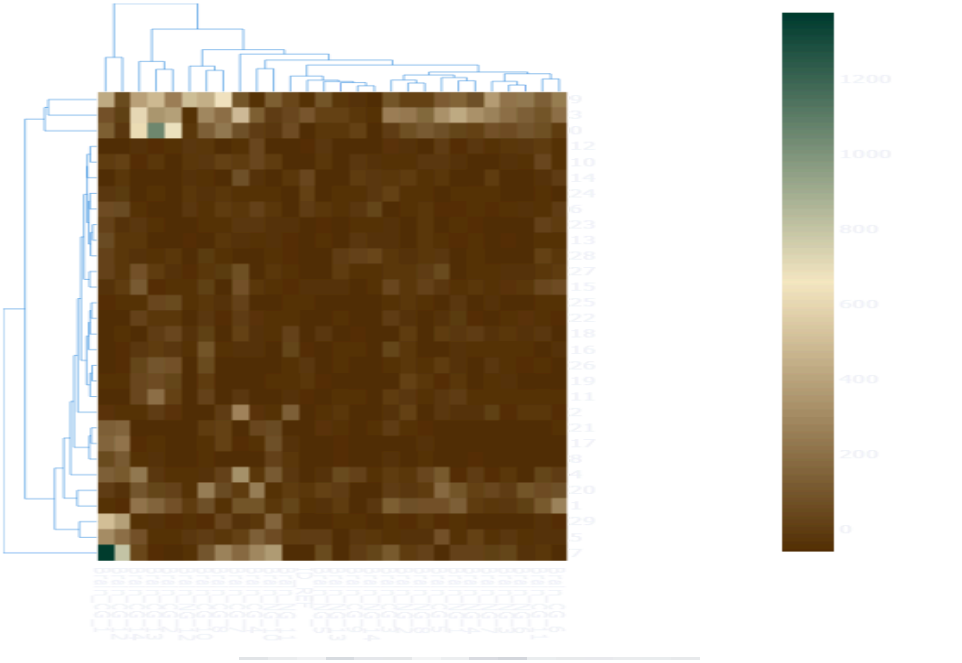
- Word cloud is a technique that brings together the words in a schematic table depending on the frequency of use of the words in the text and shows the words with the highest frequency larger (See Figure 24.). Thanks to this technique, the meanings emphasized by many sites, theses, articles and novels can be easily analyzed (Sönmez, 2009). In Figure 3 below, text visualization was made using the Tableau program.

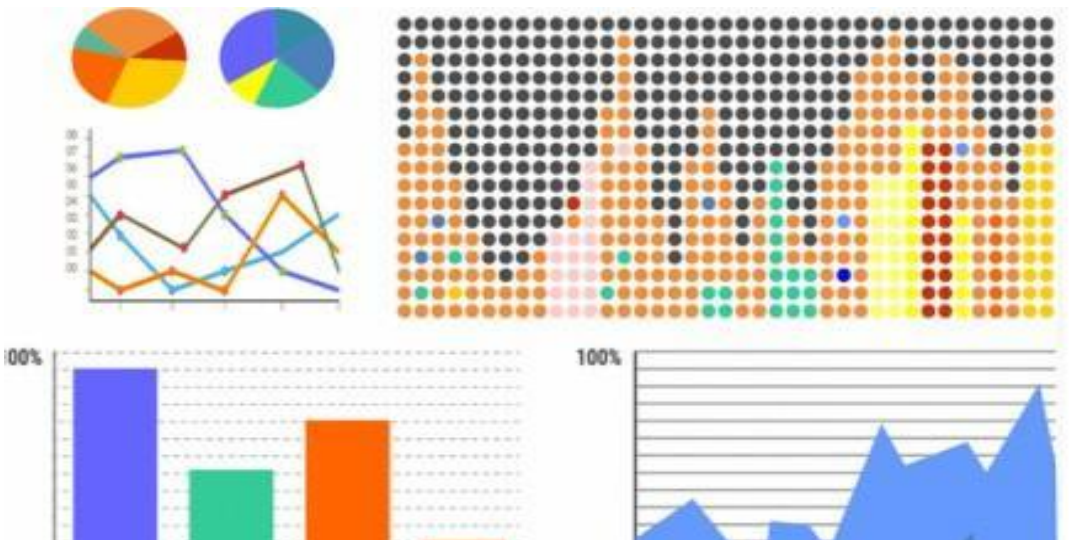
- The Dashboard is a visual representation of the most important information needed to achieve one or more goals. Dashboards are consolidated and organized into a single

screen (See Figure 5). In this way, information can be viewed at a single glance (Sandy Chiang, 2011).

- Clustergram, Hierarchical cluster analysis uses dendrogram charts to visualize how clusters are created. As the number of clusters increases, an alternative graph called "clustergram" is proposed to examine how cluster elements are assigned to clusters (See Figure 25). This graph is useful for exploratory analysis for non-hierarchical clustering algorithms such as k-means, and for hierarchical cluster algorithms when the number of observations is large enough to make dendrograms practical (McKinsey, 2011).




1200
1000
800
600
400
200
0


100%
100%

All of the cutting-edge strategies and gear used for visualisation these days are primarily based totally on simple cognitive

It is primarily based totally at the standards of psychology and the connections among size, color and visible items. Basic standards which includes manipulation are used. In phrases of human cognitive psychology. Gestalt Principles are of splendid importance. Gestalt Principles are psychology's method to visible notion. It says it is a workspace. The people, the world, the founder It is recommended that they understand in a holistically organized configuration as opposed to in parts. (For example, one's notion of the woodland first after which all of the man or woman timber as an entire as possible). Also, the human thoughts fills withinside the blanks, seeking to keep away from ambiguity. Works and effortlessly acknowledges similarities and differences. Law of proximity (forming a set series of items), the regulation of similarity (perceptually, if items are much like every other) grouped), symmetry (the tendency to understand items as symmetrical shapes), closure (Our mind has a trend to shut incomplete items) and the figure-floor regulation (Visual clean and stagnant roles of items) need to be taken under consideration within the visualization of massive information. For this purpose, the simplest visualization technique is the only that makes the first-class use of a couple of standards. Otherwise, too many colors, shapes, and connections can reason problems in knowledge the information or might also additionally reason complexity in spotting a few visible elements.

2.4 Analysis and Visualization with Apache Spark

2.4.1 Analysis with Apache Spark

In this section, we will perform exploratory data analysis using Azure Open Datasets and Apache Spark. We will then visualize our results in a Synapse Studio notebook in Azure Synapse Analytics.

We will analyze the New York City (NYC) Taxi dataset. Data is available through Azure Open Datasets. This subset of the dataset contains information about yellow taxi trips: each trip made, start and end time, and information about locations, cost, and other interesting attributes.

2.4.2 Downloading and Preparing Data

After creating a notebook using the PySpark core, we will use several different libraries to help us visualize the dataset we will be using.

```python
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
```

Libraries we will use:

- Matplotlib
- Seaborn
- Pandas

Since the raw data is in Parquet format, we will pull the file directly into memory as a DataFrame using the Spark context. Create a Spark DataFrame by retrieving the data using the Open Datasets API. Here we will use Spark DataFrame schema in read properties to infer data types and schema.

```python
from azureml.opendatasets import NycTlcYellow
from datetime import datetime
from dateutil import parser

end_date = parser.parse('2018-06-06')
start_date = parser.parse('2018-05-01')
nyc_tlc = NycTlcYellow(start_date=start_date, end_date=end_date)
df = nyc_tlc.to_spark_dataframe()
```

We want to do some initial filtering to clean up the dataset after the data has been read. We can remove unnecessary columns and add columns that extract important information. We will also filter out anomalies in the dataset.

```python
# Filter the dataset
from pyspark.sql.functions import *

filtered_df = df.select('vendorID', 'passengerCount', 'tripDistance','paymentType', 'fareAmount', 'tipAmount'\
                                , date_format('tpepPickupDateTime', 'hh').alias('hour_of_day')\
                                , dayofweek('tpepPickupDateTime').alias('day_of_week')\
                                , dayofmonth(col('tpepPickupDateTime')).alias('day_of_month'))\
                            .filter((df.passengerCount > 0)\
                                & (df.tipAmount >= 0)\
                                & (df.fareAmount >= 1) & (df.fareAmount <= 250)\
                                & (df.tripDistance > 0) & (df.tripDistance <= 200))

filtered_df.createOrReplaceTempView("taxi_dataset")
```

2.4.3 Analyzing Data

We have a huge style of equipment for information analytics to assist us benefit insights from information. We'll overview some useful equipment determined in Azure Synapse Analytics notebooks. In this analysis, we need to recognize the elements that deliver better taxi hints for our selected period.

2.4.3.1 Apache Spark SQL Magic

- We will first carry out exploratory data evaluation through Apache Spark SQL and magic instructions with the Azure Synapse notebook. After developing our query, we can visualize the effects the use of the integrated chart alternatives feature.
- We created a brand new mobileular in our pocket book and the code below. Using this question we need to recognize how the common type quantities have modified over the term we've got chosen. This question may also assist us pick out different beneficial insights, along with the minimum/most quantity of guidelines in keeping with day and the common quantity of fees.

```
%%sql
SELECT
    day_of_month
    , MIN(tipAmount) AS minTipAmount
    , MAX(tipAmount) AS maxTipAmount
    , AVG(tipAmount) AS avgTipAmount
    , AVG(fareAmount) as fareAmount
FROM taxi_dataset
GROUP BY day_of_month
ORDER BY day_of_month ASC
```

- After our question is completed running, we are able to transfer to graph view and visualize the results. This instance creates a line chart day_of_month with the aid of using specifying the sphere as the important thing and avgTipAmount because the value. After making the selections, pick out Apply to refresh your chart.

2.4.4 Visualizing Data

In addition to the built-in notebook graphics options, we can use popular open source libraries to create your own visualizations. We will use Seaborn and Matplotlib. These are common Python libraries for data visualization.

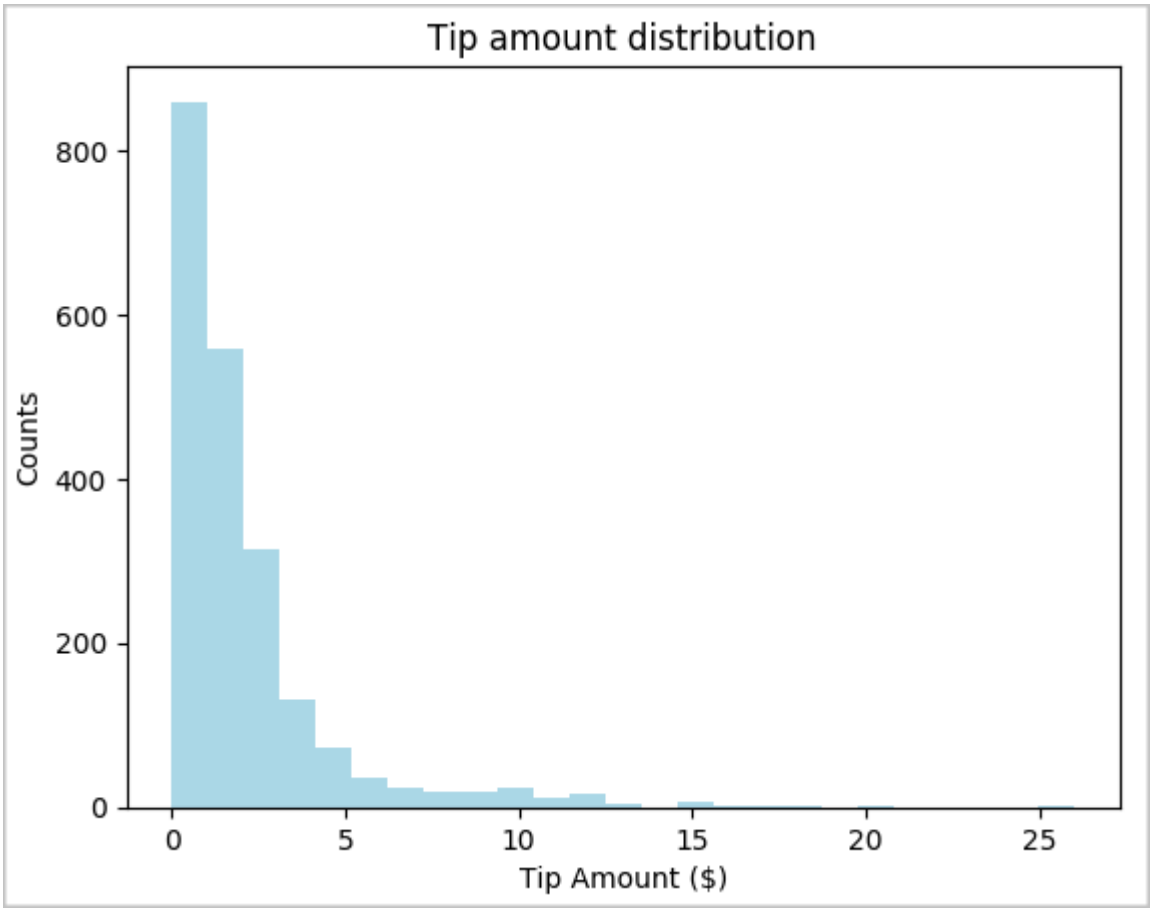
- We will do a small sampling of the dataset to make development easier. We will be using the built-in Apache Spark sampling feature. Also both Seaborn and Matplotlib require Pandas DataFrame or NumPy array. To get Pandas DataFrame transform DataFrame using command toPandas() .

```
# To make development easier, faster, and less expensive, downsample for now
sampled_taxi_df = filtered_df.sample(True, 0.001, seed=1234)

# The charting package needs a Pandas DataFrame or NumPy array to do the conversion
sampled_taxi_pd_df = sampled_taxi_df.toPandas()
```

- We want to understand the distribution of clues in our dataset. Using matplotlib we will create a histogram that shows the distribution of the hint amount and the number.

Depending on the distribution, we may see tips skewed towards amounts less than or equal to $10.

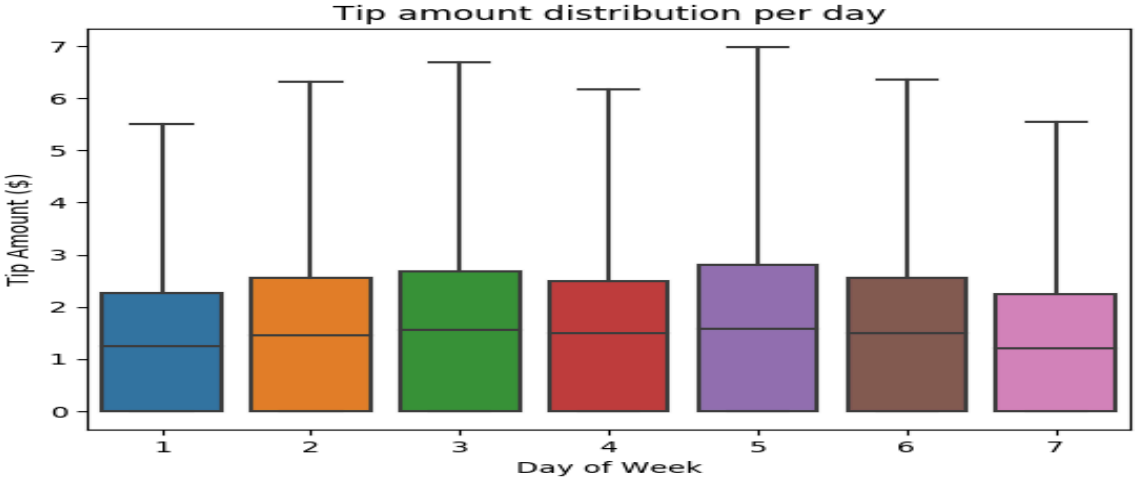
```
# Look at a histogram of tips by count by using Matplotlib

ax1 = sampled_taxi_pd_df['tipAmount'].plot(kind='hist', bins=25, facecolor='lightblue')
ax1.set_title('Tip amount distribution')
ax1.set_xlabel('Tip Amount ($)')
ax1.set_ylabel('Counts')
plt.suptitle('')
plt.show()
```



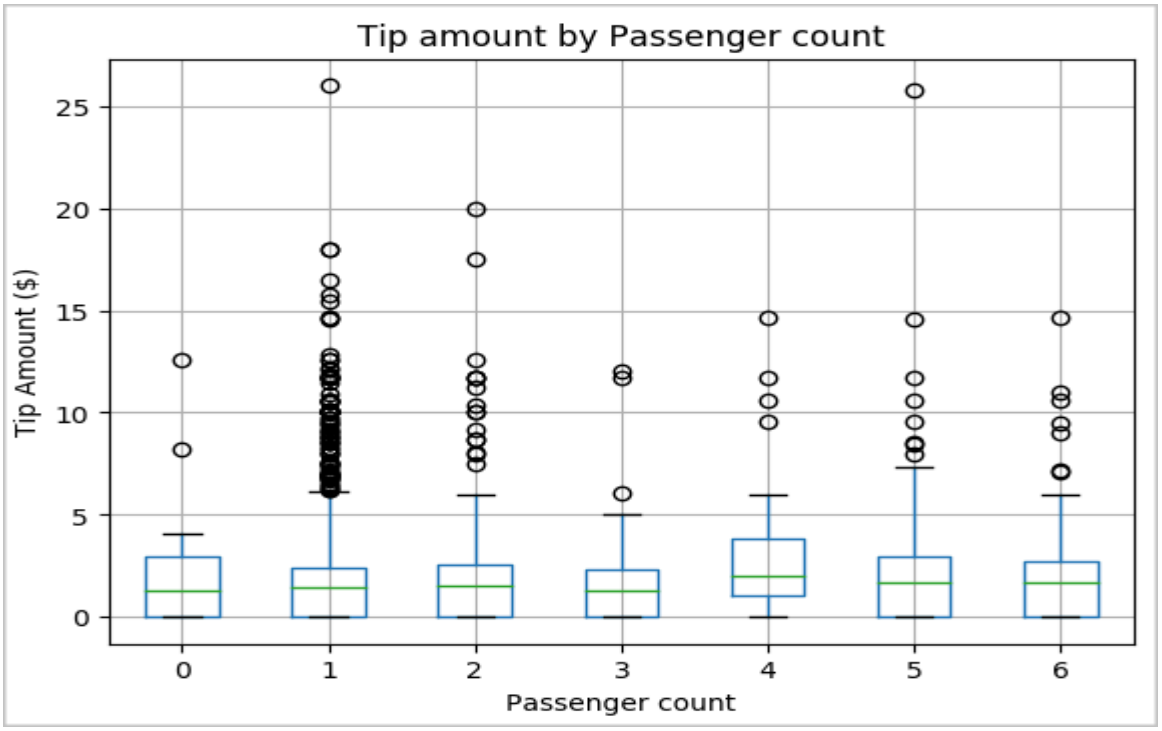
- Next, to understand the relationship between the clues of a particular trip and the days of the week. We will be using Seaborn to create a box plot outlining trends for each day of the week.

```
# View the distribution of tips by day of week using Seaborn
ax = sns.boxplot(x="day_of_week", y="tipAmount",data=sampled_taxi_pd_df, showfliers = False)
ax.set_title('Tip amount distribution per day')
ax.set_xlabel('Day of Week')
ax.set_ylabel('Tip Amount ($)')
plt.show()
```

- Another hypothesis is that we thought that there might be a positive relationship between the number of passengers and the total amount of taxi tips. To verify this



relationship, we run the code below to create a box plot showing the distribution of cues for each passenger number.



- Finally, we want to understand and visualize the relationship between the fee amount and the tip amount. Based on previous data, we can see that there are several

observations that people do not hint at. However, we can also see that there is a positive relationship between the overall fee and the amount of tips.

```
# Look at the relationship between fare and tip amounts

ax = sampled_taxi_pd_df.plot(kind='scatter', x= 'fareAmount', y = 'tipAmount', c='blue', alpha = 0.10, s=2.5*(sampled_taxi_pd_df['passengerCount']))
ax.set_title('Tip amount by Fare amount')
ax.set_xlabel('Fare Amount ($)')
ax.set_ylabel('Tip Amount ($)')
plt.axis([-2, 80, -2, 20])
plt.suptitle('')
plt.show()
```

Using the New York City (NYC) Taxi data we have with Apache Spark, we visualized the tables such as the distribution of the tip amount, the distribution of the daily tip amount, the amount of tip according to the number of passengers and the amount of tip according to the fare amount with the help of python.

3. CONCLUSION

Big data and data evaluation including transportation, electricity, health, aviation, agriculture, finance and retail it has caused superb adjustments in lots of areas. Today, companies, banks and the general public groups come each day from quite a few reassets: finance, mobile,

healthcare, transaction, purchaser studies and it could effortlessly system many data including social media data in databases. Informatics and trends in cloud technology and immediate get entry to on line data reassets it caused the emergence of a brand new era of effective vehicles. With analytical equipment of this electricity formerly inaccessible to lecturers and operators via way of means of merging or it supplied get right of entry to and use of huge data units that had been unavailable. Today, firms and textual content evaluation, system learning, predictive evaluation, data mining, statistics, herbal the usage of strategies including language processing and visualization, you could create huge, formerly unused via way of means of studying the quantity of to be had company data and gaining new insights, you could get higher and quicker they're capable of make decisions. In this study, first of all, trendy data approximately the issue of huge data given after which generalized approximately the generally used visualization strategies via way of means of supplying data, the significance of visualization for groups and states became explained. Data visualization is a reasonably new and promising discipline in pc science. Patterns, use pc photos consequences to show traits and relationships in datasets he makes use of it. Thanks to the growing generation today, huge and complicated data units may be effortlessly analyzed he has grow to be capable of. Previously, whilst huge and complicated datasets had been encountered today, whilst resorting to the sampling method, it's far now feasible to accumulate all of the data (the principle one) in a easy manner we're capable of system it. Especially the 19th. With massive and complicated facts from the sampling technique turned into used whilst encountered. However, the sampling technique does now no longer at a time whilst there are few and high-potential virtual technology aren't widespread, additionally it is herbal greeted him. A new technology of applications advanced these days permits us to apply the entire variety of facts with the aid of using allowing it, it has allowed us to look many info that we might now no longer have visible with a constrained quantity of facts before. In this way, the sub-classes that the pattern can not get right of entry to and it additionally helped us to look a miles clearer view of the infrastructures.

Especially nations way to the technological differencesed with the aid of using huge facts it maintains to create possibilities for Today it's far utilized in packages in lots of industries as may be visible, huge facts, whilst interpreted correctly, gives establishments and pioneering companies growing with new possibilities it gives for its creation. Big facts, weather change, sickness surveillance and herbal catastrophe with the aid of using offering the possibility to discover the primary reasons in their occurrence, transport, fitness and it has had a social life-facilitating impact in lots of regions inclusive of marketing. Other on the alternative hand, the blessings of huge facts to society are typically facts privateness, privateness and safety it's far constrained in such topics as.

Despite the advances in era these days, the huge facts age remains the start of its evolution in a phase. Therefore, huge facts processing strategies to clear up huge facts issues new answers are continuously being advanced and improved. In this statement, the huge facts world, multidisciplinary, which ends up in a higher know-how of complicated systems and Dec among them it could be said that strategies and strategies want to be improved.

REFERENCES

SINHA, S. (2021, November 10) *Real Time Big Data Applications in Various Domains.* 2022 tarihinde edureka! : https://www.edureka.co/blog/big-data-applications-revolutionizing-various-domains/ adresinden alındı

Big Data (Büyük Veri) Nedir? (2021, September 26*)*. 2022 tarihinde Proente Otomasyon*:* https://proente.com/big-data-buyuk-veri-nedir/#:~:text=%7C%20Big%20Data%20Nedir%3F, büyük%2C%20daha%20karmaşık%20veri%20kümeleridir *adresinden alındı.*

Big Data Analytics Life Cycle (2021, September 06). 2022 tarihinde GeeksforGeeks: https://www.geeksforgeeks.org/big-data-analytics-life-cycle/#:~:text=Data%20Munging(Validation%20and%20Cleaning,Preparation%20for%20Modeling%20and%20Assessment) adresinden alındı.

Büyük veri (Big Data) nedir? (tarih yok). 2022 tarihinde Mysoft: https://www.mysoft.com.tr/buyuk-veri-big-data adresinden alındı.

Big Data Nedir? Büyük Veri, Yapay Zekanın Zincirlerini Kırmasını Sağlayacak Anahtar Olabilir mi? (2022, January 15). 2022 tarihinde Evrim Ağacı: https://evrimagaci.org/big-data-nedir-buyuk-veri-yapay-zekanin-zincirlerini-kirmasini-saglayacak-anahtar-olabilir-mi-11347 adresinden alındı.

How Big Data Shaped Today's Casino and Why You Should Care (tarih yok). 2022 tarihinde Raving: https://betravingknows.com/weekly-reports/data-analytics/2018/05/how-big-data-shaped-todays-casino-and-why-you-should-care/ adresinden alındı.

Büyük Veri (Big Data) Nedir? (2019, March 8). 2022 tarihinde Ceyrekmuhendis: https://www.ceyrekmuhendis.com/buyuk-veri-big-data-nedir/ adresinden alındı.

ERL T., KHATTAK W., BUHLER P. (2016). Big Data Fundamentals: Concepts, Drivers and Techniques. *Big Data Analytics Lifecycle* (s. 11-15).

Big Data (2022, January 15). 2022 tarihinde Techtarget: https://www.techtarget.com/searchdatamanagement/definition/big-data adresinden alındı.

Guberman, S. (2015), "Guberman S. On Gestalt theory principles. GESTALT THEORY", 2015;37(1):25–44.

ÇELİK S., AKDAMAR E. (2018). *BIG DATA AND DATA VISUALIZATION.* ISSN:1694-528X University of Economics and Entrepreneurship, Turkish World Kyrgyz – Turkish Institute of Social Sciences, Jalalabad – KYRGYZSTAN

Analyze data with Apache Spark. (2022, May 02). 2022 tarihinde Microsoft: https://docs.microsoft.com/tr-tr/azure/synapse-analytics/spark/apache-spark-data-visualization-tutorial adresinden alındı.

Hitzler, P. and K. Janowicz, *Linked Data, Big Data, and the 4th Paradigm*. Semantic Web, 2013. 4(3): p. 233-235

ABU-SALİH B., WONGTHONGTHAM P., ZHU D., YAN CHAN K. (2021). Introduction to Big Data Technology., (s.1).

Big Data (tarih yok). 2022 tarihinde Techtarget: https://www.techtarget.com/whatis/definition/3Vs#:~:text=3Vs%20(volume%2C%20variety%20and%20velocity)%20are%20three%20defining%20properties,or%20dimensions%20of%20big%20data.&text=According%20to%20the%203Vs%20model,of%20data%20to%20be%20managed adresinden alındı.

Big Data (Büyük Veri) Nedir? (2019, April 29*).* 2022 tarihinde Medium*:* https://medium.com/dusunenbeyinler/big-data-büyük-veri-analizi-d53d8f8ab52b adresinden alındı.

Bayrakçı, S. (2015), *Büyük Veri Nedir* Serkanbayrakci: https://serkanbayrakci.wordpress.com/tag/bigdata adresinden alındı.

Kaya İ., Akbulut H. D., Özener K. (2018). *BIG DATA ANALYTICS IN INTERNAL AUDIT.,* (s. 260)