Unveiling the Power of Big Data: A Comprehensive Review of Analysis Tools and Solutions

Paril Ghori
Country: India
E-mail: parilghori@gmail.com

Article History

Article Submission
26 February 2021

Revised Submission
20 March 2021

Article Accepted
25 May 2021

Article Published
30 June 2021

Abstract

With the development of technology and the increase in communication networks, the size and value of information has come to the forefront today. In today's technology, many devices can produce and record instant data, and as these records increase, the data is considered unimportant and sent to the data warehouse. Many institutions and organizations, who realized that there are rich underground resources under complex and diverse data by conducting research and development studies in this direction, have revealed a structure called "Big Data". It is seen that meaningful, useful and important data emerges thanks to big data. Institutions and organizations, who conducted their research and development studies in this direction, have revealed the structure we call Big Data. Big Data does not only mean large amounts of data, but also data that cannot be processed with traditional methods. As technology develops every day, companies are literally competing with each other. The development of these technologies leads to the production of a lot of data. Data interaction is increasing every moment, and it is becoming difficult to control and analyze this increasing data. Today's big technology companies are creating many big data analysis tools to control and analyze data. In this paper, different types of tools used in the analysis of big data are explained comparatively.

Keywords- *Amazon EMR, Apache Impala, Apache Spark, Hadoop MapReduce, IBM BigInsights, KNIME Big Data Connector, MapR Platform, QlikView, SAP HANA, Talend Studio*

I. INTRODUCTION

Big Data continues to grow in a way that cannot be controlled in proportion to today's technology. It is data that cannot be stored in classical relational databases and is difficult to make structural with today's data processing methods because it is very diverse. Examples of this data are social media sharing, audio and video transmissions, instant records on the network, data received from sensors, web server and client logs, statistical information, news, log files, archive systems and data obtained from other channels. The biggest fundamental reasons for the continuous increase of big data can be shown as the increase in social media use, the increase in personal computers, smartphones, tablets and even smart watches, the increase in sensors that store instant data, the widespread use of the internet and the development of cloud technology. The proliferation of these technologies leads to the formation of billions of data per second in the world. The big data process is the analysis of this data and making it meaningful so that it can be used. Tools are used to analyze such large data. Big data analysis is the process of performing operations on large data sets using analytical techniques [1]. Big data is divided into three as structured, unstructured and semi-structured data. Structured data: Data that is kept

ISSN: 2250-0839

in a fixed manner. Row/column data in relational database systems is an example of this. Also other examples: Exam results, news headlines, population records, etc.
Unstructured data: Data that is not kept in a fixed form but shows variety. For example, e-mail content, images, videos, audio, social media interactions, articles and books.

Semi-structured data: It is not kept in fixed areas but elements such as labels are used to categorize the data [5]. Examples include XML, HTML and JSON.
In order to analyze big data, the components of the big data phenomenon must first be well known. Big data consists of five basic components. These are: volume, variety, value, velocity, and veracity of the data. This is why it is known as the 5V rule in the literature. In addition, variability has recently been included as the sixth rule [3]. As seen in Figure 1, a cycle is seen.

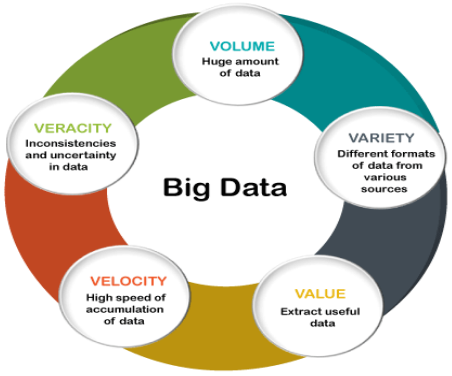

Figure 1: Big Data Components [2]

- Data Volume: Big data is directly proportional to the volume component. Analyzing, processing and recording big data with traditional solutions is insufficient. In 2021, a total of 16.1 ZetaBytes of data was created in the world, and it is expected that 163 ZetaBytes of data will be created in 2025 [4].
- Variety: The data produced does not consist of only one category, the data consists of more than one category. Web interactions, log files, news sites, e-mail flows, published documents, phone records, television and radio archives, social media sharing, etc. categories constitute the variety in big data, and it is quite difficult to analyze this meaningless data with today's techniques.
- Value: It is the ability to produce information from the filtered data by going through the stages in the cycle above with the collection of the obtained data. It is the most important element of the cycle, and the value or worth of the produced data is revealed at this stage.
- Velocity: With the advancement of technology, the increase in devices communicating with the internet, and the widespread use of social media, data is increasing very rapidly. Rapidly increasing data requires the rapid development and operation of both software and hardware products.
- Veracity: It is a very important issue that data comes from accurate and secure channels. At the same time, when data arrives securely, confidentiality and integrity of data must be protected outside of authorization.

Data Dynamics (Variability): It is an additional stage added during the processing or life cycle of data. The continuous increase in data makes it difficult to protect data. For example, with the increase in social media use, it becomes difficult to protect data integrity [2].

II. IMPORTANCE AND BENEFIT OF BIG DATA

In order to get efficiency from the phenomenon of big data, analysis should be done by applying new technologies rather than traditional techniques. When big data is analyzed well, it is an effective data science in finding the most appropriate solutions, making mathematical modeling, and establishing the mechanism of decision support systems. Big data, when used with the right methods and techniques, can enable institutions and governments to make the right decisions strategically, analyze risks effectively and carry out studies such as innovation in technology [6].

Many methods, tools and techniques have been developed for big data analysis. Using these methods, businesses can increase their profit image, make predictions according to customers' behaviors and determine

which product they buy the most, which colors they like and at what time they shop. In order to get efficiency from the big data phenomenon, in-depth analysis should be done with new technologies rather than traditional techniques. When the data at hand is defined and analyzed well enough, it provides significant benefits to businesses and organizations. Obtaining strong prediction ability, cost and time savings are the most important benefits.

III. AREAS OF USE OF BIG DATA

Big data has many application areas in the world such as retail sales, e-commerce, banking, social media, business, advertising, weather conditions, stock exchange, health, telecommunication, intelligence activities, education, e-government applications, transportation and cargo transactions and the energy sector. The application areas can be associated as follows:

- Telecommunications: Fast solutions can be produced by collecting subscriber data and analyzing consumer data. Fast answers can be produced by performing real-time analysis. They also record the mobile data of their subscribers. Analysis processes are performed on this data and they can offer special discounted packages and suitable tariffs for postpaid or prepaid according to the data used by their customers.
- Banking: Banks model user behavior with the data they collect from their customers. With this model, they have the ability to make their customers feel special. Early warning detection and credit risk detection can be made for securities and card fraud.
- Education: With the collection of education system data, positive developments can be created on students thanks to big data. The big data obtained by collecting education data can be analyzed and students who are behind the system can be found. In this way, a correct method can be selected for students and a great contribution can be made to their progress in their courses.
- Retail: There is unused data obtained from customer cards, POS devices, RFID tags and personnel entrance cards. By analyzing these data, timely stock analysis, personnel optimization, customer behavior analysis, correct product placement, product variety and price optimization can be done.
- Health: Individual data such as blood stock tracking, early diagnosis and status tracking of diseases, DNA analysis can be used using records obtained in the health system.
- Business: Finding the cause of customer losses, determining customer behaviors.
- Public: Integration and ability to work together between government departments and different organizations. Governments can provide appropriate services for citizens by analyzing issues such as energy service areas, fraud detection, research and development projects, and environmental protection.
- Energy: Energy companies obtain big data by storing their subscribers' consumption data. In today's technology, energy companies can analyze this data and offer different price tariffs to their customers individually.
- Social Media: Data is produced instantly on social media platforms used by billions of people. To give examples of this data; structured and unstructured data such as location notification, photo and video sharing, person statuses, personal information, comments and likes are published on social media on a daily basis. Social media analysis is performed using this data and user behaviors can be determined.
- Transportation: In the transportation sector, mostly the sender and receiver information held by cargo companies is analyzed and advantageous plans can be made based on this analysis data. Bus, airline, train and ship companies can improve their relations with customers by collecting passenger information. Airline companies use big data to improve their dialogue with their customers and make them feel special. Companies have found ways to analyze the social status of passengers and create an opportunity. There is a lot of data from personal information, travel plans, preferences, hotel accommodation and car rental habits. Using this data, they make passengers feel special and help/advantage passengers [28].

Big data methods and technologies Technologies have been developed by large companies due to the proliferation, unstructured and diverse nature of data and the ability to process big data. Many structures and methods have been developed to obtain, collect, process and analyze big data [2].

3.1 Big Data Analysis Methods

There are many techniques and methods to analyze data sets. These methods are mostly used in data mining applications. Some techniques used in big data analysis are described below. Researchers continue to develop new techniques and improve existing techniques when it comes to the need to analyze data sets [5]. Not all of these techniques can be used for big data, and some are effectively applied to smaller data sets (A/B testing, regression analysis) [5].

- A/B testing: This method is also known as split testing or bucket testing. For example, it determines the rate at which the color of buttons, text content, fields and images on an e-commerce site should increase conversions [5].
- Association rule method: It is a technique that allows finding relationships between variables in large databases [5, 7]. For example, it is a technique that determines which products a consumer frequently buys when buying a product and applies this determined data as marketing analysis [5]. Relationships can be discovered by comparing data from a POS device in a shopping store [7]. Association rules are one of the data mining methods.
- Classification: It is a method of dividing data in a database into certain classes. It is the process of determining the class of the data added to the database by passing it through a certain algorithmic filter. It is frequently used in data mining.
- Cluster analysis: It is a technique used to classify various data groups into smaller groups according to their similar characteristics [5].
- Genetic algorithms: It is a search and optimization algorithm inspired by the working system of evolution. In this technique, parameters are used in a coded form.
- Machine learning: It is a sub-specialist branch of computer science. It is the design and development of algorithms that allow a system to develop behaviors based on experimental data using mathematical methods [5]. Examples of machine learning are face recognition systems, handwriting recognition, natural language processing, search engines, pattern recognition, and voice recognition [8].
- Natural language processing: It is a sub-branch of artificial intelligence. It is a branch of science that aims to process, analyze and use natural languages (daily spoken language). It uses computer algorithms when analyzing human language. Natural language processing is a form of machine learning. The process of conducting sentiment analysis of how consumers react to a chain store's current campaign on social media can be shown as an example of natural language processing [5].
- Regression analysis: It is a technique that detects the connection between two or more types of data and allows us to make predictions based on this connection. Such as the time spent in the store and the connections between elements such as background music, age and height.

3.2 Big Data Technologies

Many techniques, methods and technologies are used to reveal data. There are currently countless data flows in the world and the data that is constantly being generated can be analyzed. Analyzing and processing this data requires a system and technology infrastructure that requires serious performance. In this section, the technologies used in big data analysis and processing are explained.

- Distributed System: It is the communication of multiple computers on a network without being dependent on each other and without being aware of it. It is the system in which more than one computer works for a job.
- MapReduce: It is a programming model used to process and analyze large data sets in parallel [9]. It is a data analysis algorithm consisting of two functions, Map and Reduce. It was first proposed by Google. When the map function, which filters the data to be processed, and the Reduce function, which returns the analysis of the processed data as a result, come together, MapReduce emerged.
- Google file system (GFS): Google designed the Google File System to meet the rapidly increasing demands, thinking that existing file systems do not meet the data processing needs [10]. Google keeps the data and analyzes of billions of web pages on the Google File System [11].

- Big Table: It is a structure designed as a distributed system to manage very large data. Satellite images, web pages, financial data and news on Google are kept on the big table. This much data is kept on it in real time. Big table controls all this data by providing a flexible and high-performance solution [12]. Google; runs all of its mapreduce, Google File System and Big Table solutions on clusters formed by thousands of low-cost server computers [11].
- Business intelligence: These are tools that collect and process large amounts of unstructured data from internal and external systems, such as books, journals, documents, health records, images, files, e-mails and other sources [13].
- In-Memory Database Technology: Unlike traditional databases, it does not keep data on disk. It is a technology that allows data warehouses to perform fast operations by storing them on main memory (RAM). Considering that the memory cost has decreased, it can be foreseen as an efficient approach.
- NoSQL: It is a structure that emerged as an alternative to relational databases in order to store the ever-growing data and to meet the needs of high-traffic systems [14]. NoSQL means not just SQL.
- Cassandra: It is an open source database that has adopted the NoSQL database structure by Apache. It is used by institutions such as CERN, GoDaddy, GitHub, Instagram, Netflix, Reddit [15].
- MongoDB: It is a database management tool that has implemented the NoSQL database architecture, is non-relational, open source, free, document-based and used for high-volume content.
- DynamoDB: It is a distributed database developed by Amazon and a database service that has adopted the NoSQL architecture completely, provides fast, high performance and has no limit on the amount of data [16]. It is a structure that has been used within Amazon for a long time before serving the end user.
- Hadoop: It is a software framework developed to process certain types of large data sets on a distributed system. It was developed inspired by Google MapReduce and the Google file system [5]. It was developed to process large-scale data between clustered computers. It is an open-source ecosystem that combines the features of Hadoop MapReduce with a distributed file system called Hadoop Distributed File System (HDFS). It is also a software development structure consisting of HDFS and MapReduce components [17].
- YARN (Yet Another Resource Negotiator): It is designed to provide resource management (such as CPU, memory) for big data, large-scale applications such as MapReduce, and distributed systems. It is distributed as open source by the Apache Software Foundation. It provides resource management and scheduling in the background [18].
- Apache Spark: A project claimed to be faster than Hadoop, which allows operations on large data.
- MLlib: A library that includes Apache Spark's machine learning algorithms.
- Apache Hive: A data warehouse package used to perform queries and data analysis on Hadoop. It is developed by Apache. It uses a structure similar to SQL, such as HiveQL, to manage and query structured data. The most important feature of Hive is that it converts the codes written with Hive into Java MapReduce codes in the background, so there is no need to learn Java.
- Apache Pig: A project developed for data processing like Hive. MapReduce operations are performed more easily and quickly with Pig. It is an open source project developed by Yahoo.
- Storm: A system that can perform real-time calculations and is distributed free of charge. Storm performs Hadoop's batch calculations in real time. It processes unlimited data streams reliably [19].
- Cloud Computing: Cloud computing is a service that can store data on the Internet via web services and also allows for collaborative information sharing [20].
- ElasticSearch: It is a powerful and fast tool for document indexing, full-text searching and data analysis within large data warehouses.
- Redis: Redis is an open source database that is kept in memory as a data structure and used as a cache and intermediary. It supports data structures such as strings, lists, sets, sorted sets, images, and geographic locations. Redis is a NoSQL database designed with key/value design and is increasingly popular.
- Apache Oozie: It is a workflow scheduler tool that manages all jobs in the Hadoop system in periodic time periods, developed open source and free of charge.

- Apache Kafka: It is a messaging system required to collect large data sets quickly, error-free and securely from various channels and send them to other distributed systems. The streamed data is thrown into a queue structure and transferred to tools such as Hadoop, pig, hive, spark. It is a tool that provides reliable real-time data flow between systems or applications of collected data [21].
- Apache Mahout: It is an open source library developed by Apache to create compatible and high-performance machine learning applications [22].
- Sqoop: It is an open source structure that allows data transfer between relational databases and Hadoop (HDFS).
- Apache Flume: Flume is designed as a distributed system to efficiently collect, keep together and reliably transport large amounts of log data [23]. In short, it is defined as a log (log data) collection tool.

Hbase: It is a database programmed in Java, running on the Hadoop file system, inspired by Google Big Table, non-relational and adopted NoSQL architecture.

IV. BIG DATA PROCESSING TOOLS

4.1 Apache Hadoop / MapReduce

Hadoop is a software framework developed to process certain types of large data sets on a distributed system. It was developed inspired by Google MapReduce and Google file system [5]. It was developed to process large-scale data between clustered computers. It is an open source ecosystem that combines Hadoop MapReduce features with a distributed file system called Hadoop Distributed File System (HDFS). It is also a software development structure consisting of HDFS and MapReduce components [17]. With HDFS, large data can be stored and managed on more than one computer thanks to the distributed system structure. It is a programming model used to process and analyze large data sets in parallel [9]. It is a data analysis algorithm consisting of two functions, Map and Reduce. It was first proposed by Google. When the map function, which filters the data to be processed, and the Reduce function, which returns the analysis of the processed data as a result, come together, MapReduce emerged.

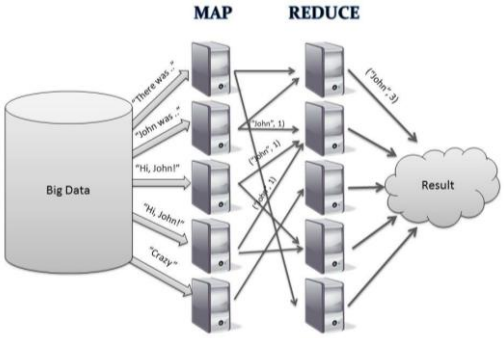

Figure 2: Map/Reduce Process [24]

4.2 Talend Open Studio for Big Data

This big data processing tool developed by Talend software developers simplifies ETL (Data collection, data transformation and loading to target) [25] processes for data sets with the speed and capacity of Hadoop [26]. Talend Open Studio for Big Data is an eclipse-based tool with easy, wizard applications. It also enables us to benefit from the full power of Hadoop by using NoSQL, HDFS, Hbase, Hive, Pig, YARN technologies in an integrated way for your data [26].

4.3 SAP HANA

SAP Company has introduced SAP HANA to process big data in real time [31]. It provides fast processing by using the in-memory database technology that it hosts with the constantly increasing large volumes of data. Thanks to its in-memory technology, it has the ability to make instant, comparative, fast decisions and analyze big data in real time [32]. SAP HANA is not a stand-alone software that performs data control, analysis, and

integration operations, but it is a data platform that comes with hardware. It has the ability to analyze structured, unstructured, and semi-structured data.

4.4 KNIME Big Data Connector

It is an open source data analysis platform. It contains two tools related to Big Data. The Knime Big Data Connector tool provides easy access to Hadoop data from KNIME Server and KNIME Analytics. It is accessed via libraries via Hive or Impala to access Hadoop/HDFS. In addition, after accessing this data, data is added to Knime nodes and operations are performed. The KNIME platform performs many operations. These can be specified as document reading and parsing, entity recognition, filtering and modification operations, word counting, keyword extraction, transformation, and visualization operations [27].

4.5 Apache Spark

It is a platform that can process data 10 times faster than Hadoop Mapreduce on disk or 100 times faster in memory. Apache Sparki is a fast engine that can process large-scale data. It has the ability to easily process streaming data instantly, and it can also process data faster and easier by writing less code than the codes written in Hadoop MaprReduce process. It can be considered as an alternative to MapReduce structure. In-memory technology is used in Spark processes. It can run on Hadoop or cloud alone [30]. It can access HDFS, Cassandra, HBase and many other data sources. Spark includes powerful libraries such as MLib for machine learning, Spark Streaming, SQL, GraphX for streaming data [30].

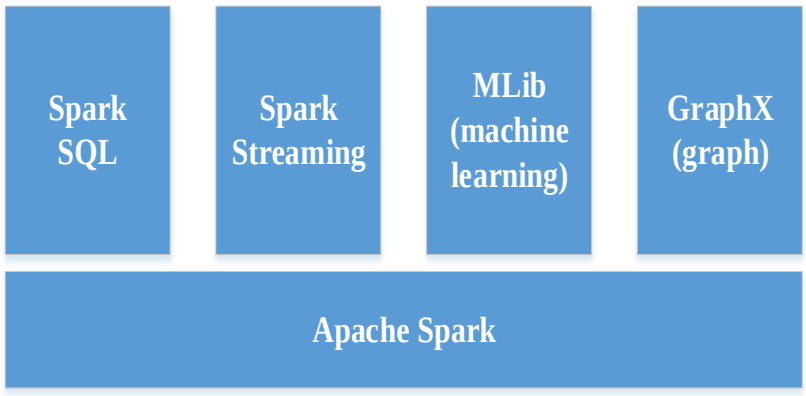

Figure 3: Apache Spark Ecosystem [30]

4.6 Amazon EMR (Elastic MapReduce)

A service of Amazon, which has the ability to perform data analysis, can manage data using platforms such as Apache Hadoop and Spark to process and analyze large amounts of data [35]. It offers the ability to transfer and manage large-scale data to data storage storages such as Amazon S3 (Simple Storage Service), DynamoDB, HDFS to store data.

4.7 QlickView

QlikView has introduced a tool that produces solutions for big data. Data analysis can be done without using a data model and programming language [29]. Data can be followed instantly with smart visuals and can be accessed via your mobile devices. Data can be compressed up to 10 times with in-memory database technology. For users who have a big data infrastructure, a big data warehouse but do not want to store QlikView data sets, QlikView provides direct access to in-memory and out-of-memory data sources using the Direct Discovery approach using specific analyses using in-memory technology [29].

ISSN: 2250-0839

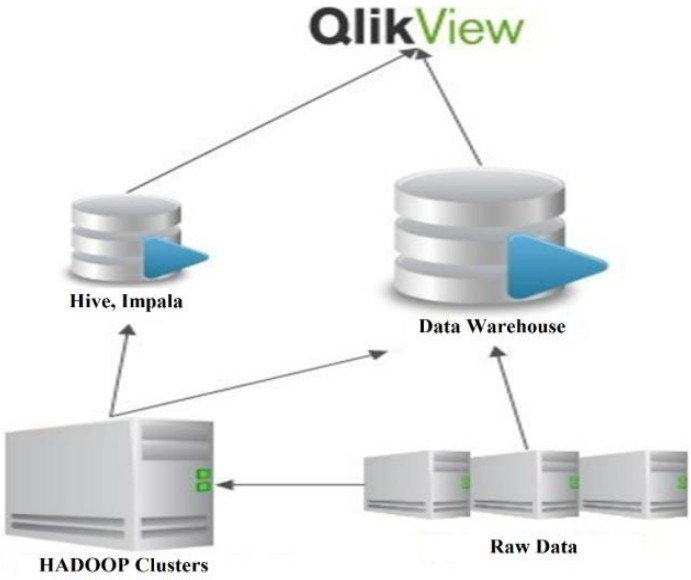

Figure 4: QlikView Working Logic [29]

4.8 IBM BigInsights

It offers big data analysis operations and processes that appeal to corporate scales as a platform. This flexible product adopts an enterprise structure that combines the features of Hadoop and Spark architecture to provide big data analysis. With the BigInsights product, it has the ability to easily integrate and analyze structured and unstructured data. These capabilities include many topics such as Spark, query sentences and text analysis. The ready provision of the Hadoop infrastructure can reveal valuable information in less time [33]. It can analyze data with the JAQL query language.

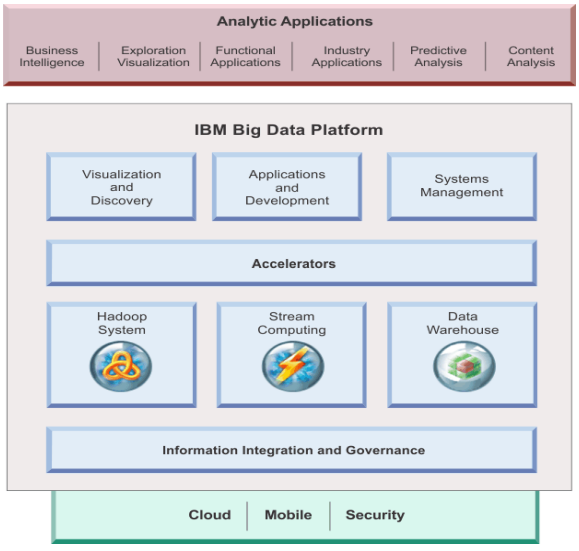

Figure 5: IBM Big Data Structure [34]

4.9 Apache Impala

Impala is an engine that can perform real-time, high-performance and distributed queries that provide results in a short time by developers on the Apache Hadoop ecosystem, which was previously developed by Cloudera and later included in the Apache incubation project [23]. Impala is an open source application that is accessible to everyone. Impala is a free application and is offered as a package that can perform real-time queries or queries compiled from source. It has also been supported by Cloudera, MapR, Oracle and Amazon recently.

4.10 MapR Converged Data Platform

MapR is a platform that can perform real-time data analysis using technologies such as Hadoop, Spark and Apache Drill, which it hosts to contribute to new generation data applications, and provides an enterprise data management and storage platform using a scalable and own file system. The MapR platform significantly reduces both hardware and operational costs by providing enterprise security, reliability and real-time processing for your important data and applications [36]. It uses Apache Kafka application interfaces for streaming data analysis, and also includes a unity HDFS structure with its own file structure system. It can run on cloud services such as Amazon Web Service, Microsoft Azure.

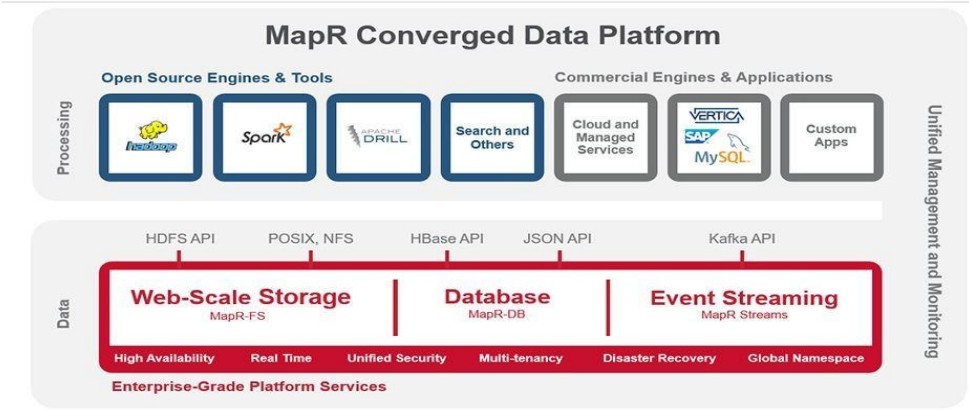

Figure 6: MapR Architecture [36]

V. BIG DATA LAYERED ARCHITECTURE

Big data architecture includes components such as data storage, processing, analysis and management. As seen in Figure 7, it is shown in a layered structure.

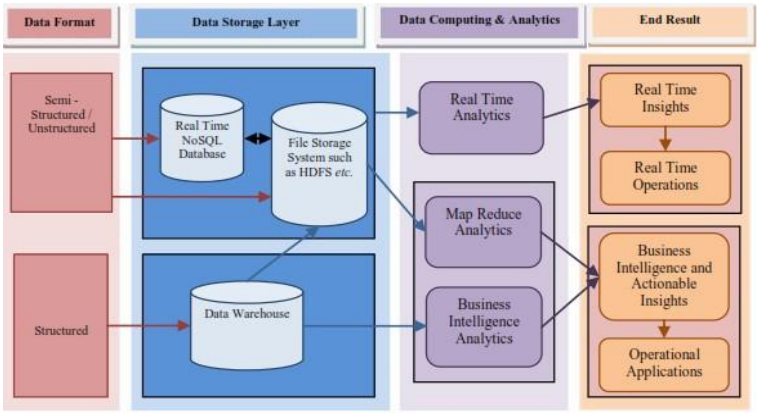

Figure 7: Big Data Layered Architecture [37]

Data analysis and direction can be made by looking at the layers in the big data architecture seen in Figure 7. According to this layered architecture, the processes that the data at hand has been passed through can be determined by looking at its complexity. All processes can also be processed separately.

VI. COMPARISON OF BIG DATA PROCESSING TOOLS

Big data tools are compared in terms of speed, scalability, calculation, technology and method in Table 1, Table 2, Table 3 and Table 4 below.

Table 1: Features of Big Data Tools

Tool / Platform	Speed	Scalability	Fault Tolerance	In-Memory Database Technology	Open Source
Hadoop MapReduce	Slow, lower performance	Scalable	Yes	Data is processed on disk	Yes
Talend Studio	Fast	Scalable	Yes	Uses In-Memory technology with Apache Spark infrastructure	As a platform, it is not open source, its components are made up of open source projects
QlikView	Fast	Scalable	Yes	Uses In-Memory technology with Apache Spark infrastructure	As a platform, it is not open source, its components are made up of open source projects
Amazon EMR	Fast	Dynamic scalable	Yes	Uses In-Memory technology with Apache Spark infrastructure	As a platform, it is not open source, its components are made up of open source projects
Apache Spark	Fast	Scalable	Yes	Yes	Yes
SAP HANA	Fast	Scalable	Yes	Yes	As a platform, it is not open source, its components are made up of open source projects
IBM BigInsights	Fast	Scalable	Yes	Uses In-Memory technology with Apache Spark infrastructure	As a platform, it is not open source, its components are made up of open source projects
Apache Impala	Fast	Scalable	Yes	Supports in-memory data processing	Yes
KNIME Big Data Connector	Fast	Scalable	Yes	Uses In-Memory technology with Apache Spark infrastructure	As a platform, it is not open source, its components are made up of open source projects
MapR Platform	Speed	Scalable	Yes	Uses In-Memory with MapR-DB	As a platform, it is not open source, its components are made up of open source projects

Table 2: Features of Big Data Tools

Tool / Platform	Distributed File System	Supported Data Format	Real-time analysis	License Type	Parallel Processing
Hadoop MapReduce	Yes	Structured, Semi-Structured Unstructured	No	Free	Yes
Talend Studio	Yes	Structured, Semi-Structured Unstructured	Yes	Paid	Large-scale parallel processing
QlikView	Yes	Structured, Semi-Structured Unstructured	Yes	Paid	Large-scale parallel processing
Amazon EMR	Yes	Structured, Semi-Structured Unstructured	Yes	Paid	Yes
Apache Spark	Yes	Structured, Semi-Structured Unstructured	Yes	Free	Yes
SAP HANA	Yes	Structured, Semi-Structured Unstructured	Yes	Paid	Massively Parallel Processing
IBM BigInsights	Yes	Structured, Semi-Structured Unstructured	Yes	Paid	Distributed parallel processing
Apache Impala	Yes	Structured, Semi-Structured	Yes	Free	Massively parallel processing
KNIME Big Data Connector	Yes	Structured, Semi-Structured Unstructured	Yes	Paid	Yes
MapR Platform	Yes	Structured, Semi-Structured Unstructured	Yes	Lite version is free	Parallel and iterative processing

Table 3: Features of Big Data Tools

Tool / Platform	Supported Operating System	Online Workability	Algorithm / Library Support	Purpose of Use	Extract, transform, and load (ETL) Support
Hadoop MapReduce	Windows, Linux, Mac OS	Yes (On the Cloud)	Yes	Data analysis, data processing	Yes
Talend Studio	Windows, Linux, Mac OS	Yes (On the Cloud)	Yes	Data analysis, data management,	Yes

QlikView	Windows, Linux, Mac OS	Yes	Yes	Data analysis, data integration	Yes
Amazon EMR	Debian/Squeeze	Yes	Yes	Data analysis, data processing	Yes
Apache Spark	Windows, Linux, Mac OS	Yes (On the Cloud)	Yes	Data processing, data management	Yes
SAP HANA	SUSE Linux Server, Red Hat Linux Server, Windows	Yes (On the Cloud)	Yes	Data analysis, fast reporting, data transfer	Yes
IBM BigInsights	SUSE Linux Server, Red Hat Linux Server, Windows	Yes (On the Cloud)	Yes	Data analysis, data processing, data management	Scalable ETL platform
Apache Impala	Linux	Yes (On the Cloud)	Yes	Data processing	Not required
KNIME Big Data Connector	Windows, Linux, Mac OS	Yes (On the Cloud)	Yes	Data analysis, data processing	Yes
MapR Platform	Red Hat, CentOS, Ubuntu, SuSe, Oracle Enterprise Linux	Yes (On the Cloud)	Yes	Data processing, improvement and analysis	Yes

Table 4: Features of Big Data Tools

Tool / Platform	**Data Storage Technology**	**File System**	**Streaming Data support**	**Cloud workability**	**Programming Language Support**
Hadoop MapReduce	Cassandra, Couchbase, DynamoDB, MongoDB, Neo4j, HBase	HDFS, Amazon S3	None	Yes	C++, Java, Python, PHP, Ruby, R
Talend Studio	Cassandra, Couchbase, DynamoDB, MongoDB, Neo4j	HDFS	Yes (Spark Stream)	Yes	C++, Java, Python, PHP, Ruby, R
QlikView	Neo4j, MongoDB, Cassandra, DynamoDB	HDFS	Yes (Spark Stream)	Yes	No programming language required
Amazon EMR	Cassandra, Couchbase, DynamoDB, MongoDB, Neo4j, HBase	EMRFS, HDFS, Amazon S3, EFS	Yes (Amazon Kinesis Stream)	Yes	Java, Perl, Python, Ruby, C++, PHP and R
Apache Spark	HBase, MongoDB	HDFS, MapR File System, Amazon S3	Yes (Spark Stream)	Yes	Java, Python, R, Scala
SAP HANA	HDBLCM (HANA Database LifeCycle Manager)	Unix File System, Clustered File System, Shared File System, GPFS	Yes (Smart Data Streaming)	Yes	ABAP, SQLScript, R
IBM BigInsights	IBM Big SQL	GPFS, Unix File System	Yes (InfoSphere Streams)	Yes	Annotation Query Language (AQL), C++, Java, R, Python
Apache Impala	HiveQL, HBase	Does not host a file system Runs on Hadoop	None	Yes	Supports all languages that support JDBC/ODBC application interfaces (APIs).
KNIME Big Data Connector	Cassandra, Couchbase, DynamoDB, MongoDB, Neo4j	HDFS	Yes (KNIME Streaming Execution)	Yes	R, Python, Java
MapR Platform	ANSI-SQL, MapR-DB, MongoDB	MapR File System	Yes (MapR Event Streaming)	Yes	Java, Python, Ruby, R, Scala

Table 1, Table 2, Table 3 and Table 4 have examined the specific strengths and weaknesses of the big data tools whose features are specified and compared.

6.1 Speed, Scalability and In-Memory Technology

In terms of speed, all tools except the Hadoop MapReduce method work fast. However, it should not be forgotten that some tools use the Hadoop infrastructure. They use the Apache Spark platform together with the

Hadoop architecture. For this reason, tools that use the Apache Spark platform and infrastructure, which are faster than the Hadoop MapReduce method in terms of speed, should be selected as a priority. Considering that big data is constantly increasing, the scalability feature of the tools examined in this paper should be common to all tools when the necessary computer infrastructure and technologies are provided. In-memory technology processes data directly on the main memory (RAM), while in traditional data processing technology, it can be said that in-memory technology is faster due to the processing of data stored on the disk. For this reason, among the tools we examined, tools other than Hadoop MapReduce have faster data processing features.

6.2 Open Source, License, Operating System Support and Real-Time Operation

Apache Hadoop, spark and impala are less costly than other tools because they are open source and free of charge in terms of license. MapR, IBM BigInsights, SAP HANA, Talend Studio and KNIME Big Data Connector tools include many open source projects, but the integrated platform is different from others in terms of license type. Hadoop does not support real-time analysis feature unlike other tools. Thanks to the real-time analysis feature, the data processed in the tools can be processed instantly and can be monitored in seconds. In terms of supported operating systems, Apache Hadoop, spark, talend studio, QlikView and KNIME tools are supported by Microsoft Windows, Linux distributions and Mac OS operating systems, Amazon EMR platform is supported by Debian/Squeeze (Linux distribution) and Windows operating system, SAP HANA and IBM BigInsights tools are supported by SUSE Linux Server, Red Hat Linux Server operating systems, and MapR big data platform is supported by Red Hat, CentOS, Ubuntu, SuSe, Oracle Enterprise Linux operating systems.

6.3 Online Access, Mobile, Cloud-Based Operation, Programming Language Support and Usage Area

QlikView and Amazon EMR big data tools work on online platforms and QlikView also offers the ability to report and track your analyses via the mobile platform. Apache Hadoop, spark, impala, MapR, SAP HANA, IBM BigInsights, KNIME and Talend studio tools can work with cloud-based systems. Apache Hadoop MapReduce, spark, MapR, Talend, Impala, SAP HANA, IBM BigInsights tools support many programming and query languages. In addition, when the infrastructure and technology for big data processing are provided, QlikView can perform data analysis, fast reporting and data visualization without a programming language. Since Apache Impala is a query engine, it supports all languages and databases that support JDBC / ODBC application interfaces (APIs). The most important common feature of big data processing tools, which is expressed comparatively, is that they have the ability to analyze and process big data.

6.4 Streaming Data Support, File System Structure and Database Technology

There is no library for Hadoop MapReduce and Impala for the analysis of continuously streaming data. Apache Spark, Talend Studio and QlikView platforms use Spark Streaming, Amazon EMR uses Amazon Kinesis Stream, SAP HANA uses Smart Data Streaming, IBM BigInsights uses InfoSphere Streams, KNIME Big Data Connector uses KNIME Streaming Execution, and MapR Event Streaming structures are used on the MapR platform. Hadoop MapReduce HDFS, Amazon S3; Talend Studio HDFS; QlikView HDFS; KNIME Big Data Connector HDFS; Amazon EMR EMRFS, HDFS, Amazon S3,EFS; Apache Spark HDFS, MapR File System, Amazon S3; SAP HANA Unix File System, Clustered File System, Shared File System, GPFS; IBM BigInsights GPFS, Unix File System; MapR Platform developed MapR File System; works with file systems while Apache Impala does not have a file system but since it works integrated with Hadoop, it can be applied on HDFS.

Hadoop MapReduce, Talend Studio, QlikView, Amazon EMR, Apache Spark, KNIME Big Data Connector tools can work with many databases. MapR platform can work with MapR-DB and other databases. Apache Impala uses a language similar to SQL since it is a query engine. Hive works with HBase. IBM BigInsights' BigSQL can work with HDBLCM (HANA Database LifeCycle Manager) in SAP HANA platform.

VII. CONCLUSION

There has been a great increase in the amount of data with the development of technology. Most of this data is kept aside as it is considered unusable. This data that is kept both takes up a lot of space and is not considered as unimportant, thus increasing the data and causing the formation of big data. Technology companies have developed tools for big data analysis to use the information in big data, to provide benefit, and to reveal the

valuable information in it. Thanks to big data tools, operations such as data analysis, data management, and processing of data and revealing valuable information can be performed.

In today's technology, many tools and infrastructures have been developed for big data analysis, operation and management. In order to use big data tools efficiently, your system infrastructure must be strong. The most commonly used big data analysis tools are explained in this paper. These tools are compared in terms of many features such as operating system support, speed, and real-time analysis, streaming data support, license type, scalability, in-memory database technology support and programming language support.

The recommended big data processing tool in this paper is Apache Spark structure, which is open source and free, faster than Hadoop MapReduce structure, fast data processing feature using in-memory database technology, scalable, streaming data and real-time analysis, has many library support such as machine learning, can run on Windows, Linux and Mac OS X operating systems and has cloud support.

REFERENCES

[1] RUSSOM, Philip, et al. "Big data analytics." TDWI best practices report, fourth quarter, 2021, 19: 40.

[2] KATAL, Avita; WAZID, Mohammad; GOUDAR, R. H. "Big data: issues, challenges, tools, and good practices." In Contemporary Computing (IC3), 2013 Sixth International Conference on. IEEE, 2013, pp. 404-409.

[3] DEMCHENKO, Yuri. "Defining the big data architecture framework (BDAF)." Outcome of the Brainstorming Session at the University of Amsterdam. SNE Group, University of Amsterdam, Amsterdam, 2013.

[4] “Data Age 2025” (2021), (online), Available: http://www.seagate.com/our-story/data-age-2025

[5] MANYIKA, James, et al. "Big data: The next frontier for innovation, competition, and productivity." 2021.

[6] Big Data (2021), (online), Available: https://en.wikipedia.org/wiki/Big_data

[7] "Big Data Techniques That Create Business Value" (2021), (online), Available: https://www.firmex.com/thedealroom/7-big-data-techniques-that-create-business-value/

[8] "Machine learning" (2021), (online), Available: https://en.wikipedia.org/wiki/Machine_learning

[9] LÄMMEL, Ralf. "Google’s MapReduce programming model—Revisited." Science of Computer Programming, 2008, 70.1: 1-30.

[10] GHEMAWAT, Sanjay; GOBIOFF, Howard; LEUNG, Shun-Tak. "The Google file system." In ACM SIGOPS Operating Systems Review. ACM, 2003, pp. 29-43.

[11] Big Data (2021), (online), Available: http://devveri.com/big-data

[12] CHANG, Fay, et al. "Bigtable: A distributed storage system for structured data." ACM Transactions on Computer Systems (TOCS), 2008, 26.2: 4.

[13] "What are business intelligence tools?" (2021), (online), Available: https://azure.microsoft.com/tr-tr/overview/what-are-business-intelligence-tools/

[14] NoSQL (2021), (online), Available: http://devveri.com/nosql-nedir

[15] Cassandra, Apache. "Apache Cassandra." Google Scholar (2021).

[16] DynamoDB, Amazon. "Amazon Web Services," (online), 2021.

[17] Hadoop (2021), (online), Available: http://devveri.com/hadoop-nedir

[18] YARN (2021), (online), Available: https://hadoop.apache.org/docs/r2.7.1/hadoop-yarn/hadoop-yarn-site/YARN.html

[19] Storm, Apache. "Storm, distributed and fault-tolerant real-time computation." (2021).

[20] Jararweh, Y., Doulat, A., AlQudah, O., Ahmed, E., Al-Ayyoub, M. and Benkhelifa, E., 2016, May. The future of mobile cloud computing: integrating cloudlets and mobile edge computing. In 2016 23rd International conference on telecommunications (ICT) (pp. 1-5). IEEE.

[21] KAFKA, Apache. "A high-throughput, distributed messaging system." URL: kafka.apache.org as of 2014, 5.1.

[22] Gupta, P., Sharma, A. and Jindal, R., 2016. Scalable machine-learning algorithms for big data analytics: a comprehensive review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 6(6), pp.194-214.

[23] Chauhan, Avkash. Learning Cloudera Impala. Packt Publishing Ltd, 2013.

[24] Kalavri, Vasiliki, and Vladimir Vlassov. "MapReduce: Limitations, optimizations, and open issues." Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on. IEEE, 2013.

[25] Vassiliadis, Panos. "A survey of Extract–Transform–Load technology." International Journal of Data Warehousing and Mining (IJDWM), 5.3 (2009): 1-27.

[26] "Talend Open Studio for Big Data" (2021), (online), Available: https://sourceforge.net/projects/talend-bigdata/

[27] Ramagundam, S. (2021). Next Gen Linear Tv: Content Generation And Enhancement With Artificial Intelligence. *International Neurourology Journal*, *25*(4), 22-28..

[28] "Airlines' Hidden Treasure; Big Data" (2021), (online), Available: http://www.kokpit.aero/havayollarininelindeki-gizli-hazine-buyuk-veri

[29] "Qlikview and Big Data" (2021), (online), Available: http://www.bitechnology.com/qlikview-ve-big-data/

[30] Spark, Apache. "Apache Spark: Lightning-fast cluster computing." (2021).

[31] Trifu, M.R. and Ivan, M.L., 2014. Big Data: present and future. Database Systems Journal, 5(1).

[32] Ercan, Mehmet. "The Secret of Catching Time in SAP HANA." Innova IT Solutions, (2021), (online), Available: www.innova.com.tr/sap-hana-in-memory-technology.asp

[33] IBM BigInsights, (2021), (online), Available: https://www.ibm.com/us-en/marketplace/biginsights

[34] "InfoSphere BigInsights features and architecture" (2021), (online), Available: http://www.ibm.com/support/knowledgecenter/SSPT3X_3.0.0/com.ibm.swg.im.infosphere.biginsights.product.doc/doc/bifeaturesarchitecture.html

[35] Amazon, E. M. R. "Amazon Elastic MapReduce." (2021).

[36] MapR Converged Data Platform, (2021), (online), Available: https://www.mapr.com/products/mapr-converged-data-platform

[37] Prasad, Bakshi Rohit, and Sonali Agarwal. "Comparative study of big data computing and storage tools: A review." International Journal of Database Theory and Application 9.1 (2021): 45-66.