Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 9, No. 3, September 2021, pp. 732~746
ISSN: 2089-3272, DOI: 10.52549/ijeei.v9i3.3170

Open-Source Big Data Platforms and Tools: An Analysis

Yassine Benlachmi[1], Moulay Lahcen Hasnaoui[2]
ENSAM Moulay-Ismail University, LMMI Laboratory, Institute, Meknes, 50000, Morocco

Article Info

Article history:

Received Jun 16, 2021
Revised Jul 31, 2021
Accepted Aug 30, 2021

Keywords:

Big Data
Open-Source Tools
Open-Source Platforms

ABSTRACT

Big data is attracting an excessive amount of interest in the IT and academic sectors. On a regular basis, computer and digital industries generate more data than they have space to store. In the current situation, five billion people have their own mobile phone, and over two billion people are linked globally to exchange various types of data. By 2020, it is estimated that about fifty billion people will be connected to the internet. During2020, data generation, use, and sharing would be forty-four times higher than in previous years. A variety of sectors and organizations are using big data to manage various operations. As a result, a thorough examination of big data's benefits, drawbacks, meaning, and characteristics is needed. The primary goal of this research is to gather information on the various open-source big data tools and platforms that are used by various organizations. In this paper we use a three perspective methodology to identify the strength and weaknesses of the workflow in a open source big data arena. This helps to establish a pipeline of workflow events for both researcher and entrepreneur decision making.

Corresponding Author:

Yassine Benlachmi,
ENSAM Moulay-Ismail University,
LMMI Laboratory,
Institute, Meknes, 50000, Morocco.
Email: yassin040@gmail.com

1 INTRODUCTION

With the proliferation of data on the Internet, in the cloud, in data centres, on smartphones, in the Internet of things, and in sensors, the idea of big data emerged. Big data is characterized by its volume, velocity, and variety [1,2]. Traditional computing models are made obsolete because of these characteristics. The motivation for big data exploration and exploitation assumes enormous value in large datasets. The implementation of big data in some contexts has practically changed the practices of a variety of fields and these applications have the potential to revolutionize several areas [3,4]. Data has infiltrated every industry and all business functions, and it is now regarded as a critical component of production as described in [5]. In future, big data will spawn a slew of new business models, products, and services. This is because it has strategic insight into the information technology (IT) industry and companies. The IT industry will develop new products and consumer segments that had previously been untapped. Established companies will optimized, and new business models will emerge. Academic research is conducted on big data applications, methods, techniques, and architectures. Data science is an interdisciplinary study. Given the interest of various research fields in big data, a strong and intuitive understanding of its meaning, development, constituent technologies, and challenges becomes critical [6].

Big Data Platforms are a collection of hardware infrastructures and software tools designed to quickly collect, store, and analyse data. They allow individuals and businesses to derive value and insight from data produced both inside and outside their company or field of interest [7]. In business settings, these platforms assist managers in getting a clearer picture of their operations, leading to increased efficiency, increased innovation, and a stronger market position. The potential value and information contained inside data is enormous, and it is of interest to all businesses. Unfortunately, few managers are trained to recognize its significance, and far fewer businesses have the budgets and capital to tap into that source of strength. According to OECD (Organization for Economic Cooperation and Development), ninety five percent of businesses are

Small and Medium Enterprises (SMEs) who do not possess the capability of researching the potential of data. 8]. In order to educate these businesses about their potential, realistic strategies must be presented that enable businesses to see immediate results and value. The abundance of Big Data-related applications further complicates the situation. The reason for this is that dealing with Big Data, or data in general, is not a one-step process. From the moment data is obtained until it is analysed, the process is dynamic and never-ending, and certain steps will need to be repeated until a satisfactory result is achieved. To add to the complexity of the issue, each built platform takes a different approach to the conceptual and technological challenges that each step of the process presents. Furthermore, research in this field is ever-changing, so solutions implemented today will be obsolete within the next few years. What is considered Big Data today will not be considered Big Data in the future. Companies who want to maximize their money cannot afford to spend time researching all of the options; they need assistance and education to find the best answer for their particular set of requirements. For the vast majority, it is not a question of wanting or not wanting, but of sheer impossibility due to a lack of resources and expertise. It must be acknowledged that there is no such thing as a single definitive forum that can be named as the best and most suitable for everyone's needs. In this case, there is no such thing as a one-size-fits-all solution. There has also been a lot of research conducted on Big Data and Big Data Platforms [9,10]. Academia and Industry are segregated when it comes to big data market demand. We need to give an easier way for both researchers and entrepreneurs to complete their solutions for any real time problem on big data. In our study we encompass on the big data tools through three major perspectives, they are Open Source big data technologies, Distributed Queuing Management technologies and Big data storage platforms. The main contribution of this paper is as follows

a. To map the pros and cons for the usability of the technologies and platforms considered.
b. To deploy a model for the three perspectives of Big data pipeline along Queue, Engine and Storage that can prove to be a best option for the researchers in this field.
c. To establish a relationship between academic and industry by creating the pipeline which will give both researchers and entrepreneurs decision making to a big data problem.
d. To provide a future direction to both researchers and entrepreneurs to create their own open source alternative to solve real time problems in big data.

The rest of the paper discusses the three perspectives mentioned. Section 3 focusses on Open source big data technologies. The distributed queuing management technologies are elaborated in Section 4 and in Section 5 we try to find the relationship of the above two perspective with the Big data storage platforms. Finally, we conclude our paper in Section 6.

2 OPEN SOURCE BIG DATA TECHNOLOGIES

Big data systems have no predefined concept; however, they can be thought of as a specific framework that can accommodate broader data sets that are not processed through conventional database technologies and techniques [11]. The characteristics are playing an increasingly key role in the growth of big data technologies and platforms. As shown in Figure 1, the characteristics are followed up on at least one stage of the important Big Data chain [7,6]. From the generation to the death phase of any data, this chain separates the distinct phases of the data. The supply chain of big data includes measures such as data generation, data acquisition, data storage, and data analysis (Figure 1).

The first phase in the value chain is data production, which can come from a variety of sources like human participation, social media, events, records, and media, among others. This data may be either qualitative or quantitative. Data acquisition is the next phase in the value chain, and it is described as the process of filtering, cleaning, and collecting data before storing it in a storage location or data warehouse.

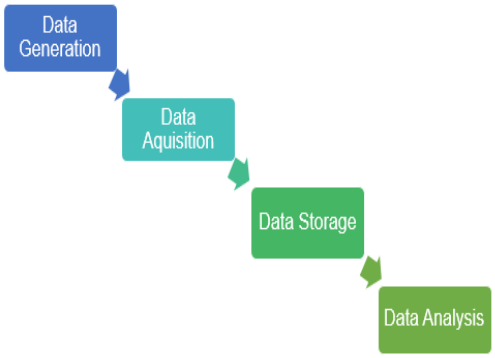

Figure 1. Value Chain of Big Data.

The four Vs (Volume, Velocity, Variety, and Value) play a significant role in data acquisition. The next task is data management, which involves storing data in any storage infrastructure that can accommodate big data or enormous amounts of data. The data is stored in a format that allows big data applications and services to access, process, and use it. In the final stage, data is analysed to discover useful patterns, data correlations, and other websites.

Any agency, company, or person that plans to use a big data platform in the future should research and execute a big data strategy before selecting and implementing a big data platform. Platforms in big data can help consumers understand their needs [12]. The big data strategy is divided into three stages:

1. Big Data Basics: In this stage the acknowledgement of various data types is represented such as pre-processed data, structured data, social data, or unstructured data.
2. Big Data Assessment: Different aspects of the data are assessed and evaluated at this stage such as source, usage, security, privacy regulation and future growth of the data.
3. Big Data Strategy: In this stage, the big data strategy is itself a study about the big data impact on the organization, which advantages can be taken by an organization from big data, and most important about the economic impacts.

An enterprise can select the best big data platform based on their requirements and functionalities after analysing and documenting their big data strategy. A common goal of such stages is to provide the ability to combine secretly obtained and publicly accessible Big Data with information produced within a company, as well as to break down the joint set for esteem extraction [13]. More specifically, Big Data Platforms should have the following features:

- It should be in a comprehensive and ready state for enterprise use.
- A platform must be flexible and scalable with respect to the requirements.
- Data must be updated with low computing power.
- A platform should have fault tolerance and robust quality.
- The platform can be open source and corporate so that the development and investment teams can take more interest.
- Maintenance process should

The applications that make up a big data platform are those that are not solely based on functionality. It does not have to be made with the most cutting-edge technology. Often all it takes is a new design or a different approach to an old problem. All is dependent on the needs of the organization or person who will be using the platform. This arrangement, however, is not widely used because updating several years old structures to meet modern needs comes at a considerable cost. As a result, new phases are evolving to meet new requirements. Big data platform specifications are divided into three stages: data acquisition, data organization, and data analysis. Data collection systems typically use less processing power for data capture, shorter data query times, data distribution environments with support for various data structures, and shorter data capture times. All of these features are provided by the no structured query language NoSQL database tool, but this platform focuses more on data capture than data categorization [14].

In the Big Data age, the preferred method of storing all data in a unique location is no longer viable. Large organizations must be able to move copious amounts of data while maintaining their integrity. The basic requirements for the frameworks are that they have a limit in terms of vertical and horizontal scaling, that they allow circulated programming, that they have high throughput of data in the centre of the base, and that they help with both structured and unstructured data. Hadoop and its Hadoop Distributed File System (HDFS) are currently the most widely used solutions for this era of data processing. However, as the worldview shifts from cluster preparation to continuous stream handling, more fitting phases are replacing Hadoop, such as Spark, Storm, or S4. Data analysis, like knowledge coordination. It necessitates the use of disparate circumstances to comprehend the undertakings of in-depth research and observations into a broad range of data forms. Platforms must accept facts such as data being stored in different frameworks and scaling up in terms of volume. They should be able to communicate responses and react to changes in knowledge behaviour more quickly. Information mining and examination assignments performed by Big Data Platforms are urged to be done in various areas with the middle of the road results being sent to a focal area that bunches them back for another procedure of concatenation in some situations where information moving from one spot to another raises security concerns but also restrictive expenses for associations to help. However, because middle-of-the-road results are not as precise as crude information, this results in a significantly more multifaceted nature because commotion and discretion can be presented to security support, and perusing in addition to including them may result in both significant information and a faulty translation [15].

3 DISTRIBUTED QUEING MANAGEMENT SYSTEM

A Big Data Platform is a collection of resources and technologies that are used in ecosystems to perform various types of data analysis, including complexity, volume, and dynamic data. As a result, scaling

up the hardware becomes a necessity, and selecting the appropriate hardware or technological advancements becomes a crucial decision if the client's requirements are to be met in a reasonable amount of time [16]. The time for a decision on the client's requirements usually falls on three important properties they are engine, queue and storage. The following are some open -source big data systems and their comparisons:

3.1 Cloudera

Cloudera is a multi-environment platform controlled by open-source technology that helps clients extract valuable business insights from their data, regardless of where it resides. In an endeavour knowledge cloud, information is readily accessible across the board with the adaptability and flexibility to handle any remaining mission. Because of its transparent architecture, it provides clients with transparency into the entire knowledge lifecycle as well as customization flexibility. The Cloudera Company was the first to build and distribute Apache Hadoop-based applications. For interested users, this platform made big data analytics more user-friendly and usable. Through this network, several important open-source projects have been integrated with Hadoop. Cloudera has built a framework with advanced features to implement the end-to-end Big Data workflow. Cloudera ecosystem is made up of various projects that perform various Big Data tasks such as analysis, storage, user web interfaces, searching, and message passing. CDP Data Center, Enterprise Data Hub, and heavy-duty professionals HDP Enterprise Plus are the three forms of annual subscriptions available for this platform. Every subscription's price and components are determined by the amount of storage space, number of nodes, and computing power available. Cloudera and Hortonworks Company merged in 2019 to offer multi-cloud, comprehensive, and end-to-end hybrid services. This platform came in second place in an analytic report with a ranking of almost 50 points. Cloudera received 85 points in the report. The platforms' main features are as follows [17,31] (Figure 2).

Figure 2. Cloudera Platform.

Open-Source Platform: The user can have access to the source code of Apache and this code can be customized, updated and adjusted according to the requirements.

Hybrid Deployment: This platform provides the deployment facilities to users with flexibility and accessibility. The user can access the data and perform the read write operations from the cloud or open premises storages. The meaning of hybrid architecture of this platform is that users can choose between the IaaS and PaaS platforms.

Data Warehouse: Data can be gathered from structured, unstructured and edge sources. They have optimized infrastructure that moves the workload on other platforms for large amounts of data analysis so that queries are answered instantly by auto scaling data warehouses.

Machine Learning: Cloudera empowers venture information science in the cloud with self-administration access to represented information. It conveys AI workspaces with flexible auto-suspending asset utilization guardrails that can give start to finish AI instruments in one durable condition.

Operation database DB: The operational database DB guarantees both high simultaneousness and low inactivity, handling enormous heaps of information all the while immediately. It can separate ongoing bits of knowledge and empower adaptable information-driven applications.

3.2 Azure Hadoop and distributed insight HDInsight

It is a cloud-based full-spectrum, managed, and open-source analytics solution for businesses. This framework supports a number of open-source frameworks, such as Apache Storm, line long and process LLAP, Apache Spark, and others. Azure Hadoop and distributed insight HDInsight is another name for Hadoop

components cloud delivery [18]. Processing large amounts of data has become easy, fast, and cost-effective. A wide variety of scenarios, including machine learning, IoT, data warehousing, and extract, transform and load ETL, can be handled using various open-source frameworks (Extract, Transform and Load). In the analytics report of the "Selecthub" website, Azure HDInsight received 82 points, making it one of the top 5 big data platforms. There are the following key features [17] (Figure 3).

- Monitoring: This platform provides the facility to monitor all the clusters on a single interface using the Azure Monitor logs.
- Cost effective and Scalable: The workload on this platform can be up and down according to the need. Users just pay for these nodes which they are using. Flexibility and performance can be increased by decoupling the computer and storage.
- Globally Accessible: Compared with other platforms, it is available in most regions. There is also a version of the Azure platform available in China, Germany and Azure government, which provides an option to meet the user's needs on key sovereign territories.
- Extendable: By using the script action, adding edge nodes or integration with other big data tools, the cluster of this platform can be extended with installed components. This platform provides the facility to integrate with other famous big data tools on a click deployment.

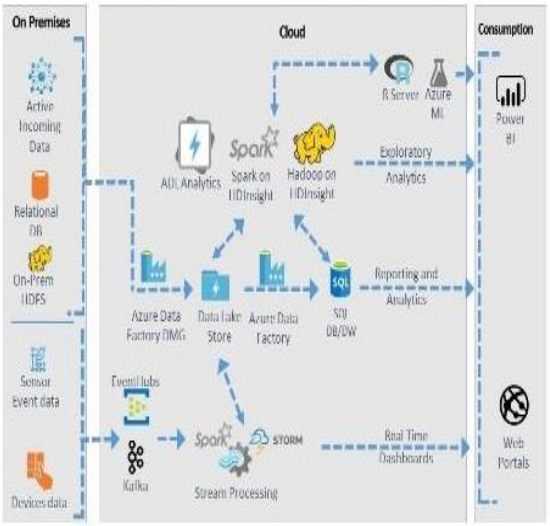

Figure 3. Azure Architecture.

- This platform can store data of several types and sizes. Poly Base, a component of SQL Server 2012 Parallel Data Warehouse, will help you discover new bits of information by combining social and non-social data directly inside Microsoft Excel by effectively breaking down Hadoop data. By using power business intelligence (BI), internal and external data can be combined to answer new types of questions. A consumer can get Hadoop without hesitation on this site. HDInsight can deploy an Apache Hadoop cluster in under a minute.

3.3 Massive Online Analysis (MOA)

This is an open-source framework that presents conceptual drift in the data mining stream. For the assessment, MOA employs a variety of online and offline methods. It is a software-oriented environment for implementing algorithms and running experiments with the goal of online learning from growing data streams. The Naive Bayes classifier can be used at the leaves to enforce boosting, Hoeffding Trees, and bagging in the MOA. Using WEKA (Waikato Environment for Knowledge Analysis), bi-directional interaction is supported in MOA. WEKA is an open-source workbench that is used to implement the batch machine learning methods of a wide range. The MOA is released with the license of GNU general public license (GPL). As compared to traditional batch learning methods, requirements are different in the data stream environments [19]. These requirements are postulated below and presented in Figure 4.

Examples are processed one at a time and only inspected one time at the most.

a) Memory is extremely limited.
b) Time is also limited to process an example.
c) Regardless of the time, predictions can be made anytime.

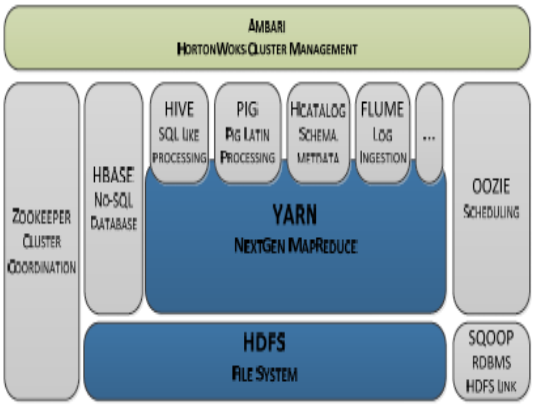

Figure 4. MOA GUI.

Classification of data streams is a relatively new field, with little research being done in this area compared to conventional batch methods. The majority of experiment evaluations have less than one million training examples. If an algorithm can handle a large number of stream instances, data stream classification is useful; otherwise, the results are disappointing [20]. MOA is a platform that allows a large number of data streams to be analysed using classification algorithms, with tens of millions of examples analysed in a finite amount of memory [5]. If anyone tests the algorithm with less examples than these, it is not a realistically challenging process. This platform is written in Java, which has the advantage of portability since it can run on any platform that has a Java Virtual Machine (JVM) and libraries that are compatible. Different classifiers, stream generators, and assessment techniques are available in the MOA. MOA is used in a graphical user interface (GUI) as well as a command line interface (CLI) (CLI). This platform is extremely easy to use and it would be beneficial to expand this classifier - AbstractClassifier" to make writing new classifiers easier. The classification is the new priority of MOA. This platform will eventually include periodic pattern learning, data stream clustering, and regression. The latest version of the MOA is “MOA 19.04”.

3.4 PEGASUS (A Peta-Scale Graph Mining System)

Precision engineered geometrically advanced suspension PEGASUS is an open-source stage developed by Carnegie Mellon University's information mining community and designed specifically for data mining in diagram structures ranging in size from a few gigabytes to petabytes. Since datasets of this scale can no longer be prepared on single-hub computers, PEGASUS, which runs on the Hadoop brain, employs parallel programming. Since it deals with information that exists in the framework or charts and structures of piles of hubs and connections, it is a more explicit information mining technique or network than others.

This is a significant improvement over earlier implementations, which could only work in millions of sizes, while this platform will work in billions. Graph or Diagram structures are increasing in number and importance in a variety of areas, including portable systems, informal associations, and clinical fields, such as protein guidelines [35]. This framework provides a quick and pervasive user interface to help complex applications achieve their full potential. PEGASUS combines various diagram mining tasks, such as processing the chart distance across, registering the period of each hub, and finding connections between diagram hubs, by using a matrix-vector increase speculation known as the goodwill industries of monocacy valley GIM-V [35]. Page Rank, Random Walk with Restart, and distance through estimate were some of the diagram mining activities we worked on. It provides direct scaling in terms of the number of edges to break down, making it suitable for use with any number of machines. The architecture is not fully developed yet, and current efforts are underway to expand the library to include more up-to-date data mining and AI calculations, as well as to include more efficient chart ordering strategies. It is written in Java and runs on any system capable of running Hadoop, with a preference for uniplexed information and computing system UNIX machines. It has a lot of programming requirements because it needs Hadoop, as well as Apache Ant, Java, Python, and Gnuplot. The Apache License adaptation 2.0 authorizes it, and the most recent stable version is 2.0 [21].

3.5 Hortonworks Data Platform (HDP)

Hortonworks Data Platform (HDP) is an open-source framework for transferring data and creating massive, multi-source informational indexes. The HDP team helps clients modernize their IT base and secure their data, whether in the cloud or on-premises, as well as generate new revenue sources, improve customer experience, and reduce costs [26]. HDP enables agile technology organization, AI and deep learning for high-value tasks, continuous data warehousing, and security and administration. Currently, it is an essential part of

information design for information that is quite still[22]. It is a modern information design that conveys prompt incentive by slicing stockpiling costs as it incorporates Yarn into its server farm, and by advancing Enterprise Data Warehouse costs by offloading low esteem registering undertakings, for example, ETL to Yarn. Yarn enables HDP to integrate all data preparation engines across the network and business biological system to provide consistent mutual administrations and assets across the platform. Ambari is a user-friendly Web user interface UI and a powerful representational state transfer application programming interface REST API that makes HDP easier, more predictable, and secure. Furthermore, HDP is a complete solution that includes not only data preparation and executives, but also the ability to organize an undertaking's demands through defence, administration, and other activities.

As shown in Figure 5, Hortonworks is a set of open-source platforms such as Hive, Hbase, Pig, Yarn, and Hadoop. The components of this platform are represented by different colours in the figure: blue represents Hadoop core stack components, grey colour represents Hadoop ecosystem components, and green colour represents HDP components. The HDP encourages the use of Apache total exclusion zone Tez to improve efficiency and address performance-related issues [23]. “Because this platform does not see Hadoop as a replacement for conventional data management systems, it focuses on providing integration elements for those platforms” [24] (HDP, 2016). HDP sees Hadoop as a tool to complement the existing data platform, a similar vision to that of the Proprietary Software vendors.

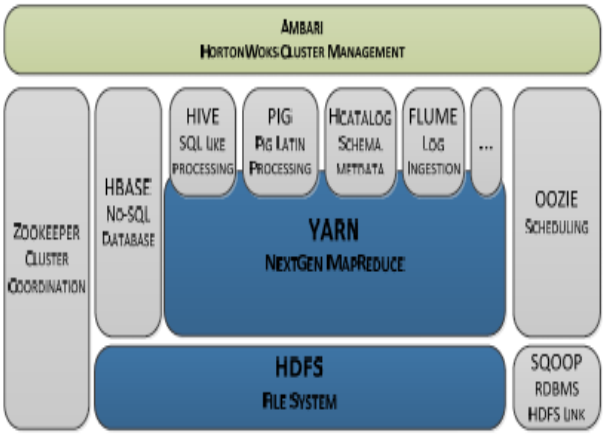

Figure 5. Distribution of Hortonworks.

3.6 Pivotal Big Data

Pivotal Big Data is a cloud-based tool that is used to develop the company's future. Small services of this platform enable the company's employees to design compostable services for independent implementation, scaling, and recovery. The fully automated and runtime deployment services are available in production ready applications (Figure 6). This framework is open-source, and it is built on Cloud Foundry, which is itself an open-source platform as a service (PaaS), with contributions from over 40 members of the Cloud Foundry Foundation as described in [1]. Pivotal Data Suite provides the fundamental components of a cutting-edge, cloud-based data architecture for assembling and executing the correct calculations for sophisticated applications. It comprises the fundamental components for cluster and stream examination architectures and can be delivered both on-premises and in the open clouds.

The Pivotal Data Suite provides you with access to and support for our executive's company contributions. According to the site, Pivotal Data Suite allows you to mix and match different portfolio products as needed [2,22]. According to Pivotal, the Pivotal Big Data suite is a mix of traditional and emerging technology. "The Pivotal Big Data Suite fills a much-needed gap in the market," Pivotal said, "by providing a multi-faceted data portfolio with a 'use it as you need it' pricing model, enterprise companies can capitalize on exponential data growth. Per-core pricing guarantees that data that is actually being processed is not taxed," according to the company. "This will be critical as businesses continue to merge more and more data into a Business Data Lake." Most references along businesses uses big data approach are addressed only based on Programming paradigm and languages [40 ,41, 44]. But there is less preferences on a pipeline for a decision-making process. We imply queue, engine, storage pipeline to address this decision-making process. We shall first look at the various available platforms in this section which helps us to find a better direction towards the pipeline approach.

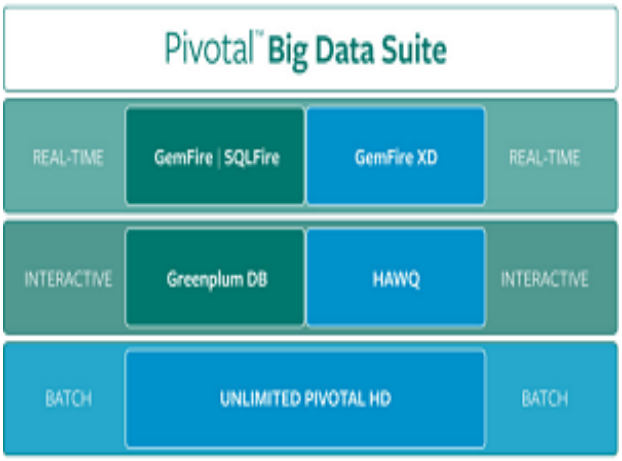

Figure 6. Pivotal Big Data Architecture

3.7 Apache Hadoop

The Apache Hadoop framework is a free and open-source project developed by the Apache Foundation. This platform implements the MapReduce paradigms as well as the "Hadoop Distributed File System" (HDFS). The platform employs thousands of servers, which also function as the slaves, but each server on this network has its own storage and processing. A large amount of data is processed in a distributed environment using basic programming models where one group of servers is the master and the other is the slave of the network [3]. In this platform, MapReduce and HDFS are the two main components where the Hadoop distributed file system HDFS is used for the data storage purpose and MapReduce for the data analysis [4]. HDFS is the foundation of Apache Hadoop. The key advantage of MapReduce is that it supports parallelization and can break work across several units in case of a node failure. MapReduce allows a non-experienced user to quickly perform operations on substantial amounts of data as mentioned in [5,21].

HDFS is a file system that is used in Apache Hadoop and Google File System (GFS). The scalable distributed file system HDFS stores a large amount of data. The HDFS stores data copies on various nodes for fault tolerance and parallel processing. Name node and data node, also known as master node and slave node respectively, are two types of nodes.

Table 2. Advantage and Disadvantage of Open-Source Big Data Platforms.

Platforms	Advantages	Disadvantages
Cloudera	User Friendly Interface Useful tools like Cloudera Impala Secure by Design Deployable Everywhere SQL tools for real-time analytics	Slower than Hadoop distribution Support less tools as compared to HDP Complications in UI
Azure HDInsight	Provide Cloud native Platform as a Service Lower Cost Scalability	On demand scalability is limited No single Host designs Vender Management is separate
MOA	Portable Perform Big Data streams mining Perform large scale machine learning Multi label Classification	Works in serial computing
PEGASUS	Support Cross platform Store large scale of data Large Graph Mining Open Source Parallel Algo on Hadoop and Spark	Interface is not user friendly
HDP	Support Windows OS Strong Security Low Complexity Stability	Not rich features in Ambari Management interface.
Pivotal Big Data	Lightweight API Gateway Open Source Independent Deployment Lower cost	N/A
Apache Hadoop	Lower Cost of Storage. Data Processing workload optimization. Flexible "Schema-on-Read" access to all enterprise data. Largest Community	I/O operation are not enough Optimize Disk space issues

3.8 Open-source big data platforms comparison

Different open-source big data systems were explored in depth in the previous section. To compare all open-source big data systems, ten attributes were chosen for comparison [7]. These characteristics have been explored in the literature [30] and [31] to assist business managers in selecting the best big data platform for their needs. Developers, Year, Programming Language, User Interface, Operating System, and User Rating, as well as the most recent version, programming paradigm, storage space, software requirement, and algorithms, are among these attributes. These characteristics not only characterize platforms functionally, but also determine whether they can be used with other systems or whether a platform is appropriate for small, medium, or large businesses.

The comparison begins with the developer and developmental year of open-source big data systems to determine the platform's strong position, as Cloudera was created by three engineers affiliated with Facebook, Yahoo, and Google. The first big data platform appeared at the turn of the century, and Apache Hadoop, which was created in 2006, is the oldest and most widely used open-source big data platform. Next, the programming language is examined so that a manager can quickly determine which type of programmers are required to operate the system. Apache Hadoop is the platform that supports the greatest number of programming languages. The user interface characteristics of any platform are critical because they provide users with ease of use. Most of the platforms have nice and user-friendly interfaces, making it easy for users to communicate with the system. The next step is to evaluate the operating system to inform users which operating system they need to use. Azure HDInsight, MOA, Pegasus, and Pivotal big data are examples of platforms that support Microsoft Windows, Linux, and UNIX. The programming paradigm is a function that is linked to the company's size and computer infrastructure. The MOA platform excels at serial computing, while all other systems, such as Pegasus, Apache Hadoop, and Cloudera, excel at parallel computing.

Software specifications and storage space characteristics are also critical factors to consider when choosing a device. Cloudera is a network that offers additional storage space at a low cost. MOA is a platform that has less storage capacity than other platforms. Finally, the algorithm attribute is examined to determine which platform supports the greatest number of algorithms. This property also shows which platform is more comprehensive than the others.

Machine learning algorithms such as the Apache Hadoop, MOA, and Pegasus platforms support classification, clustering, and regression. Each platform's advantages and disadvantages and weak points are discussed in detail in Table 1. The aim of this paper is to compare the top open-source big data platforms. The strengths and weaknesses of each platform will help SMEs choose the right big data platform for their day-to-day operations. Apache Hadoop, Cloudera, and Pivotal Big Data are the ones who support bigger datasets while MOA is the weaker one in this field supporting only datasets with a size of a few megabytes.

In conclusion, MOA is unquestionably the best forum for small businesses with limited computing infrastructures and limited data volumes. It is difficult to choose between Apache Hadoop and Cloudera for larger organizations with larger infrastructures that require parallel computing. While the first has the advantage of supporting more data types, the second is more user-friendly due to the presence of a graphical user interface (GUI) that helps monitor the work being performed.

4 BIG DATA STORAGE PLATFORMS

In the field of information technology, data is everything these days. Furthermore, this data is increasingly expanding daily. Data is now measured in terabytes and petabytes, instead of megabytes and kilobytes. The data in its raw form is meaningless; but, after some processing, it becomes valuable information and knowledge that aids in analysis and decision-making. For this reason, there are numerous open-source big data resources on the market. These programs allow you to analyse, sort, report, and do a lot more with data. In this section, different open-source big data tools will be discussed with its advantages and disadvantages [25].

4.1 Knime

It isa free and open-source tool for analysing enormous amounts of data. Knime, which stands for "Konstanz Information Miner," is a platform for interactive data pipeline execution and visual assembly. It is used in analysis, data analytics, market intelligence, enterprise reporting, CRM, and text mining, among other fields (Figure 7). Some of the world's most prestigious firms, such as Johnson & Johnson, Comcast, and Canadian Tire, use this method to mine their data [32]. Its key features include a simple user interface, easy integration with new nodes, and the ability to explore qualified models and study results. Knime 4.7 is the latest release update (as from 2018). It is a powerful platform for data analytics tasks that integrates with the powerful libraries WEKA and R-Statistic.

4.2 MongoDB

This tool is a non-rational database (NO-SQL) written in C, C++, and JavaScript with a complex schema of JSON documents. It was created in 2009 and is still being developed and expanded. MongoDB is a database that is used by both small and large businesses with thousands of users. Many well-known companies, like the New York Times, eBay, Facebook, Google, and Foursquare, use this tool. The data is stored as a binary JavaScript object notation BSON document in this database, rather than a table with a predefined schema[6]. It is a free to use open-source platform that supports a variety of operating systems including Windows, Linux, Free Berkeley software distribution BSD, and Solaris (Figure 8). BSON format, Replication, file storage, Capped Collection, Indexing, Aggregation, Load balancing, and MongoDB management service are the key features of this database (MMS).

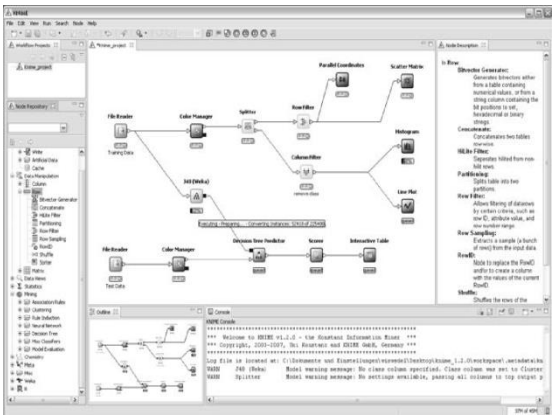

Figure 7. Data flow Analysis in KNIME.

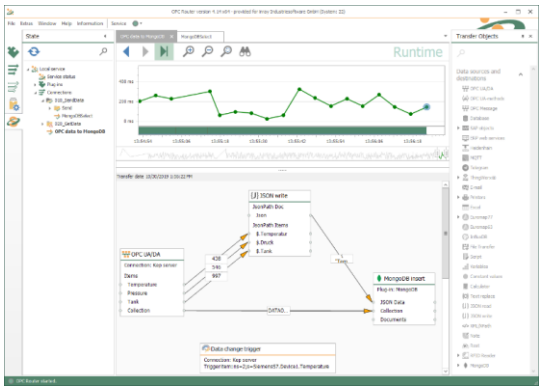

Figure 8. Data flow Analysis in MongoDB.

4.3 HIVE

It is a Java-based open-source cross platform that allows for data summarization, data interpretation, and querying. This application is based on Hadoop. It was created by Facebook in January 2007 and was later used by Hadoop users in August 2008 after becoming an open-source tool. Query language for Hive is similar to SQL declarative language. The hive query language (HiveQL)is used to create map to reduce tasks that are executed by Hadoop [33]. This tool allows users to run queries on Hadoop data clusters in the same way that they would on conventional databases. This tool has made Hadoop more accessible to users of business intelligence. It is a user-friendly framework with a SQL interface and a logical model. This application is built on top of Hadoop and allows for the termination and review of related queries.

Table 2. Open-Source Big Data Tools Functionality.

Tools	Functions
KNIME	Understanding of data, workflow designs of data sciences, Access to reusable components
MongoDB	Allow dynamic queries and Schema, implement aggregation functions.
Hive	Structured Query Language
MapReduce	It provides fault tolerance and distributed processing
Lumify	2D and 3D graph visualization, analysis of link between graph entities.
Rapid Miner	Access, process and analyse any type of data. it can clean the data for analytics at an expert level.

From Table 2 it is evident the queuing strategies are available various forms with respect to language, process and computing. These available functions need to be segregated for the applications in both industry and research. Hence, we focus on how we can find different relationships with application for each platform and technologies we have discussed in this section.

4.4 MapReduce

Because of its rich features, such as communicative manners and simplicity, it is an incredibly powerful and efficient method for big data analysis. The main goal of this tool is to sort out and generate large datasets in a cluster using parallel distributed computing. It is a Java-based programming model. MapReduce consists of three operations: Map, Shuffle, and Reduce. The Map function is used to sort and filter data, producing key values as output, such as sorting employee names alphabetically [34]. The shuffle operation operates with the Map- output keys, which means that all relevant data for one key is stored on the same worker node. The reduce operation is used to generate one key against each group of data.

4.5 Lumify

This tool was developed by Altamira, a company that is struggling to handle large amounts of data. It is a free and open-source platform for analysing, integrating, and visualizing big data. Graph visualization in 2D and 3D, relationship analysis between graph entities, multimedia analysis, full text search integration with mapping, and real-time collaboration with the combination of project and workspaces are the key features of this method [7]. Eric Schmidt, Google's executive chairperson, has stated:
"The biggest disrupter (of 2014) that we're sure about is the arrival of big data and machine intelligence everywhere".

Table 3. Advantage and Disadvantage of Open-Source Big Data Tools.

Platforms	Advantages	Disadvantages
Knime	It is a set of rich algorithms. Easy to set up. User friendly interface. Easily integrate with other languages and technologies. A lot of manual work is automated	Too much RAM No integration with graph databases. Data capacity should be increased.
Mongo database DB	Low Cost and easy to use. Reliable. No issues in installation and maintenance. Support many platforms and technologies.	Slow in some use cases. Analytics are limited.
HIVE	It is similar to SQL. It provides fast response time even on large data sets. It can be extended without compromising the performance. This application is scalable as volume and variety of data can be increased.	Navigation is not in Hive. Dependent tasks can be created. Deletion of file is permanent. Search function is not available on each function.
MapReduce	Execute the process again automatically if failed. Locality Optimization High Security and Authentication. Highly scalable toll.	Latency is the biggest drawback that makes it unable to be used in real time applications. In the PSO algorithm it required a large amount of data.
Lumify	Scalable Cloud environments support such work with AWS. Secure	NA
Rapid Miner	Open source No need to know about programming No complex mathematical calculations Good customer service and technical support API and cloud can be integrated with it	Need to improve the online data services.

4.6 Rapid Miner

This tool provides data mining, machine learning, and data science capabilities. This platform is cross-platform compatible. It is a free, open-source platform with all the requisite features. This Java-based tool can visualize, validate, optimize, and perform in-depth data analysis. For beginners, this is the easiest tool, but the relation of each node is important and must be understood [36]. In terms of word processing, rapid miner has more features.

Table 4. Decision making through pipelines involving Queue, Engine and Storage.

End to End Implementation	Purpose	Similar Platforms	Queue/Engine/Storage
Knowledge	Machine Learning Deep Learning	Hadoop Storm Samza Spark Flink	Engine
Processing Model	1-dimensional data Real Time data	Storm Samza Spark Flink Pivotal Big Data	Queue
Data Model	RDBMS JSON	KNIME, Hadoop, Hive, MongoDB Cassandra HBase [44]	Storage
Computing	Standalone Cloud Edge	Azure, HDP	Engine
Memory Process	Stored Online	Cloudera, Hadoop, Storm, Samza, Spark Flink[44]	Storage
Query Instance	Classifying Cleaning	Cloudera Spark SQL, Table API, StreamCQL, SamzaSQL, Squall and Athenax[37], [44]	Queue

Table 4 discusses the decision making process thereby establishing a relationship between End to End implementation of each requirements mapped to similar platforms. This helps an entrepreneur to decide how his/her industrial big data problems can be addressed through a dedicated pipeline along Queuing, Engine and Storage. Consequentially we can infer from the decision making process that whenever the Engine is of an important criteria we need to be careful with the computing. Computing here refers to the infrastructure available to handle streaming jobs [39], [40]. To discover the failure of the jobs is crucial in this environment. Hence both memory and queuing strategies in Big data pipelines play a major role in decision making. Table 5 which is a mapping of our reviewed platforms to its usage in both Research and Industry. The table shows how the three important categories classify between a decision making for a researcher or an entrepreneur. Industries with big data problems can decide based on this classification for a better decision in the future. The pros and cons of various platforms as listed in Table 2 & 3 helps in identifying various decision making process for better implementation of solutions regarding.

Table 5. Options available for two different dimensions (Research and Industry)

Reviewed Platforms	Research			Industry		
	Queuing	Engine	Storage	Queuing	Engine	Storage
KNIME	Yes	No	Yes	No	No	Yes
MongoDB	Yes	Yes	Yes	Yes	No	Yes
Hive	Yes	Yes	Yes	Yes	No	Yes
MapReduce	Yes	Yes	Yes	Yes	No	Yes
Lumify	Yes	No	No	Yes	No	Yes
Cloudera	Yes	Yes	No	Yes	No	Yes
Azure HDInsight	No	Yes	No	Yes	No	Yes
MOA	Yes	Yes	No	Yes	Yes	No
PEGASUS	Yes	No	Yes	No	Yes	Yes
HDP	Yes	Yes	No	No	Yes	Yes
Pivotal Big Data	Yes	No	Yes	Yes	Yes	No
Apache Hadoop	Yes	Yes	Yes	No	Yes	No

In conclusion, open source tools were reviewed for their approach towards the pipeline comprising of queuing, engine and storage for both research and industrial purposes. As we can infer there is no exact appropriation but one can take any path as mentioned in Table 5 to reach their destination. This helps in better decision making process. Each tool has its own set of features, benefits, and drawbacks, which are listed in Table 3. The best way to handle big data is to select the tool wisely. Lumify is a good choice if the data volume is not excessive, and it allows you to analyse the data in a graph form. Hadoop-based tools are the perfect way to develop a great analytics team, perform complex analysis for market development, and make informed decisions, but this team needs technical expertise [38]. Rapid miner is the best choice for beginners because it provides quick results and eliminates the need for infrastructure.

5 CONCLUSION AND FUTURE WORK

Wearable computers and the Internet of Things (IoT) are currently producing a large volume of heterogeneous data on a regular basis. Many tools and frameworks needed complex architecture to process and analyse this structured and unstructured data, resulting in productive outcomes and learnings. Even if there were only a few choices from which to choose, putting together such an architecture would be incredibly difficult. The open science data group is large and diverse, offering many opportunities and choices.

In this paper, we discuss some common open-source big data frameworks and tools as well as their features, functionality, benefits, and drawbacks. Big data analytics provide a diverse collection of processes for handling massive and complex data sets. In the current situation, research into the advancement of big data tools and frameworks is necessary [37]. The concept of big data, its features, and implementations are discussed in depth at the beginning of the paper. The paper's main goal is to address and explore the most widely used open-source big data tools and platforms in all their aspects. The benefits, drawbacks, features, and functionality of open-source big data tools and frameworks are examined. With the rise of Big Data and more people and organizations recognizing the value and opportunities it provides, the number of Big Data systems and platforms, both open source and proprietary, should grow rapidly in the coming years. Putting in place a framework and a platform that answers all the questions raised by Big Data is extremely time-consuming, expensive, and unlikely to be feasible by any stretch of the imagination. The more common approach is for organizations to spend resources on developing their own set of advancements that are tailored to their specific needs, either by leveraging internal resources or by calling upon the network. This results in a large number of available options, which not only complicates the decision-making process for those looking for a stage for their project, but also adds to the amount of excess and coverage in the arrangements offered. It is critical to educate market leaders as well as individuals in general about Big Data's capabilities to ensure the success of platform research and implementation.

Cloudera, Azure HDInsight, Massive Online Analysis (MOA), PEGASUS, Horton Data Platform, Pivotal Big Data, and Apache Hadoop are among the seven common open-source big data systems discussed in this paper. According to Table 1, the Apache Hadoop, MOA, and Pegasus platforms support a wide range of machine learning algorithms, including grouping, clustering, and regression. Each platform's strong and weak points are discussed in detail in Table 1. MOA is the forum for small businesses with minimal computing resources and a small amount of data. It is difficult to tell which platform, Apache Hadoop, or Cloudera, is better for larger infrastructure-based businesses that need parallel computing and operate with large data sets. Hadoop's key advantage is that it can handle a wide range of data types, while Cloudera offers a user-friendly interface that makes it easy to keep track of tasks.

In addition, we investigated and evaluated six Big Data Open-Source Platforms: Konstanz information miner KNIME, MongoDB, Hive, MapReduce, Lumify, and Rapid Miner. For the study of big data, there are several resources available on the market, some of which are open source and others which are not. Tables 2 and 3 demonstrate the basic features, advantages, and drawbacks of each method. Choose wisely the tools to perform various operations on big data, such as Lumify, which is a decent option if the data volume is not drastic and allows you to analyse the data in the form of a graph. Hadoop-based tools are the perfect way to develop a great analytics team, perform complex analysis for market development, and make informed decisions, but this team needs technical expertise. Rapid miner is the best choice for beginners, academics, and those who do not want to deal with infrastructure. If a user is already familiar with Rapid Miner, they can use KNIME to advance their knowledge, and it can also be used with Hadoop. Big data technologies and platforms will be examined in real-time experiments in the future. The use of open-source resources and platforms will be investigated further.

ACKNOWLEDGMENTS

This publication has emanated from my laboratory ENSAM Moulay-Ismail University, LMMI Laboratory, Institute, Meknes. I would like to thank my supervisor and reviewers for their valuable comments without which this would not have been possible.

REFERENCES

[1] S. M. Borodo, S. M. Shamsuddin, and S. Hasan, “Big data platforms and techniques,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 1, no. 1, pp. 191–200, 2016.
[2] D. Laney and Others, “3D data management: Controlling data volume, velocity and variety,” META group research note, vol. 6, no. 70, p. 1, 2001.
[3] C. L. Philip Chen and C.-Y. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on Big Data,” Inf. Sci. , vol. 275, pp. 314–347, Aug. 2014.
[4] T. Huang, L. Lan, X. Fang, P. An, J. Min, and F. Wang, “Promises and Challenges of Big Data Computing in Health Sciences,” Big Data Research, vol. 2, no. 1, pp. 2–11, Mar. 2015.

[5] J. Manyika et al., Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey, 2011.
[6] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," IEEE Access, vol. 2, pp. 652–687, 2014.
[7] P. D. Coimbra de Almeida and J. Bernardino, "Big Data Open Source Platforms," in 2015 IEEE International Congress on Big Data, 2015, pp. 268–275.
[8] O. for Economic Cooperation and D. (oecd) Staff, OECD Factbook 2014: Economic, Environmental, and Social Statistics. OECD, 2014.
[9] V. Mayer-Schönberger and K. Cukier, Big Data: A Revolution that Will Transform how We Live, Work, and Think. Houghton Mifflin Harcourt, 2013.
[10] M. Barata, J. Bernardino, and P. Furtado, "YCSB and TPC-H: Big Data and Decision Support Benchmarks," in 2014 IEEE International Congress on Big Data, 2014, pp. 800–801.
[11] R. Gupta, S. Gupta, and A. Singhal, "Big Data: Overview," arXiv [cs.OH], 16-Apr-2014 [Online]. Available: http://arxiv.org/abs/1404.4136
[12] M. G. Huddar and M. M. Ramannavar, "A survey on big data analytical tools," International Journal of Latest Trends in Engineering and Technology (IJLTET), pp. 85–91, 2013.
[13] J.-P. Dijcks, "Oracle: Big data for the enterprise," Oracle white paper, vol. 16, 2012.
[14] V. Abramova and J. Bernardino, "NoSQL databases: MongoDB vs cassandra," Proceedings of the international C* conference on computer science and software engineering, pp. 14-22, 2013 [Online]. Available: https://dl.acm.org/doi/abs/10.1145/2494444.2494447
[15] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 97–107, Jan. 2014.
[16] D. Singh and C. K. Reddy, "A survey on platforms for big data analytics," J Big Data, vol. 2, no. 1, p. 8, 2015.
[17] Y. Liu, T. Teichert, M. Rossi, H. Li, and F. Hu, "Big data for big insights: Investigating language-specific drivers of hotel satisfaction with 412,784 user-generated reviews," Tourism Manage., vol. 59, pp. 554–563, Apr. 2017.
[18] V. Yadav, Processing Big Data with Azure HDInsight: Building Real-World Big Data Systems on Azure HDInsight Using the Hadoop Ecosystem. Apress, Berkeley, CA, 2017.
[19] A. Bifet et al., "MOA: Massive Online Analysis, a Framework for Stream Classification and Clustering," in Proceedings of the First Workshop on Applications of Pattern Analysis, 2010, vol. 11, pp. 44–50.
[20] R. B. Kirkby, "Improving hoeffding trees," The University of Waikato, 2007 [Online]. Available: https://researchcommons.waikato.ac.nz/handle/10289/2568
[21] U. Kang, C. E. Tsourakakis, and C. Faloutsos, "PEGASUS: APeta-Scale Graph Mining System Implementation and Observations," in 2009 Ninth IEEE International Conference on Data Mining, 2009, pp. 229–238.
[22] R. Menon, Cloudera Administration Handbook. Packt Publishing Ltd, 2014.
[23] T. W. Dinsmore, Disruptive Analytics: Charting Your Strategy for Next-Generation Business Analytics. Apress, Berkeley, CA, 2016.
[24] H. Zhang, Q. Wei, and Y. Luo, "A novel infinite-time optimal tracking control scheme for a class of discrete-time nonlinear systems via the greedy HDP iteration algorithm," IEEE Trans. Syst. Man Cybern. B Cybern., vol. 38, no. 4, pp. 937–942, Aug. 2008.
[25] T. L. C. da Silva et al., "Big Data Analytics Technologies and Platforms: A Brief Review," in LADaS@ VLDB, 2018, pp. 25–32.
[26] "Pivotal Data Suite Info — VMware Tanzu Network." [Online]. Available: https://network.pivotal.io/products/big-data/info. [Accessed: 31-May-2021]
[27] W. U. Hadoop, "Welcome to ApacheTM Hadoop®!," 2016.
[28] B. Saraladevi, N. Pazhaniraja, P. V. Paul, M. S. S. Basha, and P. Dhavachelvan, "Big Data and Hadoop-a Study in Security Perspective," Procedia Comput. Sci., vol. 50, pp. 596–601, Jan. 2015.
[29] J. A. Miller, C. Bowman, V. G. Harish, and S. Quinn, "Open Source Big Data Analytics Frameworks Written in Scala," in 2016 IEEE International Congress on Big Data (BigData Congress), 2016, pp. 389–393.
[30] A. Fernández et al., "Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 4, no. 5, pp. 380–409, Sep. 2014.
[31] X. Liu, N. Iftikhar, and X. Xie, "Survey of real-time processing systems for big data," in Proceedings of the 18th International Database Engineering & Applications Symposium, Porto, Portugal, 2014, pp. 356–361.
[32] M. R. Berthold et al., "KNIME: The Konstanz Information Miner," in Data Analysis, Machine Learning and Applications, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 319–326.
[33] C. Győrödi, R. Győrödi, G. Pecherle, and A. Olah, "A comparative study: MongoDB vs. MySQL," in 2015 13th International Conference on Engineering of Modern Electric Systems (EMES), 2015, pp. 1–6.
[34] S. Pandey and V. Tokekar, "Prominence of MapReduce in Big Data Processing," in 2014 Fourth International Conference on Communication Systems and Network Technologies, 2014, pp. 555–560.
[35] A. Thusoo et al., "Hive - a petabyte scale data warehouse using Hadoop," in 2010 IEEE 26th International Conference on Data Engineering (ICDE 2010), 2010, pp. 996–1005.
[36] G. Ertek, D. Tapucu, and I. Arin, "Text mining with rapidminer," RapidMiner: Data mining use cases and business analytics applications, p. 241, 2013.
[37] S. Radhya, John G. Breslin, and Muhammad Intizar Ali. "Big data and stream processing platforms for Industry 4.0 requirements mapping for a predictive maintenance use case." Journal of Manufacturing Systems, 2020, vol.54, pp.138-151.
[38] A. Amado, P. Cortez, P.Rita, & S. Moro, "Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis.", European Research on Management and Business Economics, 24(1), 2018, pp. 1-7.[39] A.

Oussous, F.Z. Benjelloun, A.A. Lahcen, and S. Belfkih, "Big Data technologies: A survey." Journal of King Saud University-Computer and Information Sciences, 30(4), 2018, pp.431-448.
[40] ur Rehman, Muhammad Habib, Ibrar Yaqoob, Khaled Salah, Muhammad Imran, Prem Prakash Jayaraman, and Charith Perera. "The role of big data analytics in industrial Internet of Things." Future Generation Computer Systems 99 2019, pp: 247-259.
[41] Gepp, Adrian, Martina K. Linnenluecke, Terrence J. O'Neill, and Tom Smith. "Big data techniques in auditing research and practice: Current trends and future opportunities." Journal of Accounting Literature, 40, 2018, pp.102-115.
[42] S. Landset, T.M. Khoshgoftaar, A.N. Richter and T. Hasanin, " A survey of open source tools for machine learning with big data in the Hadoop ecosystem", Journal of Big Data, *2*(1), 2015, pp. 1-36.
[43] P.D.C. de Almeida, and J., Bernardino, " Big data open source platforms". 2015 IEEE international congress on big data, 2015, June, (pp. 268-275). IEEE.
[44] U. Chandrasekhar, A. Reddy, and R., Rath, "A comparative study of enterprise and open source big data analytical tools." In 2013 IEEE Conference on Information & Communication Technologies 2013 April. , (pp. 372-377). IEEE.