Significance of developing Multiple Big Data Analytics Platforms with Rapid Response

Sudhir Allam, *Data Engineer, Department of Information Technology, USA*

Abstract—*This paper looks at how multiple big data platforms can be very important in achieving the best outcome when used in rapid response operations. Big data analytics and technological innovation place a strain on traditional data governance processes, creating a slew of complex issues inside corporations. However, a governed data lake demonstrates how this can be accomplished [1]. The most difficult aspect of integrating different systems is figuring out how to adjust to their unique programming capabilities to complete tasks as quickly as possible to achieve the best results. This paper introduces the latest big data analytics platforms for a rapid response which integrates large-scale data frameworks like RHadoop and SparkR to build high-performance, multi-platform, Big Data analysis for rapid data collection and R-programming analysis [1]. The goal is to improve work scheduling optimization using the big data analytics platforms and to apply the customized device selection based on ascertaining capabilities to significantly boost system performance. Furthermore, rather than running Java or Scala programs, users will issue R commands to conduct data extraction and analytics in the presented platforms [1]. As a consequence, while the configured data analytics platform selection will greatly minimize data extraction and analysis period, and as per the output index measured for different approaches, planned optimization almost certainly improves device reliability considerably.*

Keywords: Big data, Analytics, Hadoop, Spark, RHadoop

I. INTRODUCTION

Data is important in today's market and technological environment. Although the idea of big data has been around for years, most companies today recognize that by capturing all data that enters their industries, they can implement analytics and derive substantial benefit. Big data applications and projects are increasing to explore these data to obtain information that can help make strategic choices [2]. Big Data is a term that emerged at the dawn of the twenty-first century, and it is still used by every major technology company [2,3]. Big Data is a term that applies to large, voluminous data collections, which can be structured or unstructured. This vast volume of data is generated by companies and consumers every day. Big Data Analytics is the method of analysing broad data sets to reveal insights and trends [3]. The world of data analytics is enormous in and of itself. Big data, with large volume, high speed, high variety, and increasing unparalleled volumes of data, often confront new challenges with various problems in big data, including storage, backup, maintenance, processing, searching, visualization, practical implementation, and others [3]. Unfortunately, conventional techniques cannot be overcome and it is thus worthwhile that we try to explore how useful knowledge can be extracted from the massive quantities of data.

There is a no better approach than by applying them to data you recognize and can take insights into practice can be found in studying the fundamental skills and knowledge around data, data analytics. Big Data analytics is, without a doubt, a breakthrough in the world of information technology [4]. Every year, businesses are increasing their usage of data analytics. The company's main emphasis is on clients. The sector is also prosperous in the applications of Business to Consumers (B2C) [5].

The main aim of this paper is to explore the significance of big data analytics platforms with rapid response. This will involve looking at how various big data systems ensure consistency for all current business intelligence (BI) along with associated analytical tools so that the company would not need to update vast quantities of software for certain platforms [5]. Thus, the upper-level tools that focus on the relational database that originally held the data will operate with little to no alteration on the added systems, gaining the benefits of high performance, highly available, and easy implementation. To speed up the processing of a vast volume of data, I/O delay time can be transmitted via a reliable and consistent distributed file system [6]. This research seeks to share various big data platforms

to achieve convergence with any established business intelligence (BI) along with similar advanced analytics such that the businesses may not need to update massive volumes of software for those platforms by developing a collection of several big data analytics platforms with high performance, high reliability, and high availability.

II. RESEARCH PROBLEM

The problem that this paper will solve is to explore how Multiple Big Data Analytics Platforms with Rapid Response can be useful in dealing with the current environment of massive data.. When an organization has monthly TBs data then they need to review the data and get desirable outcomes from it, however, operating on such a large volume of data with standard tools is not easy [6]. Nevertheless, if somehow to get to operate with those basic tools will take days to get reliable results. That is why big data systems are used to manage these data over a limited amount of time to provide reliable information. Multiple Big Data Analytics Platforms may be useful in processing large data. This includes Hadoop, Spark, and their expanded RHadoop and SparkR frameworks to address all types of relatively Big Data Analytics issues [7]. Hadoop distributes large data sets through many nodes, allowing big data mining and analytics much more efficiently than was available in the past. Spark, on the other side, is not capable of storing data in a distributed fashion [8]. Spark is ideal for analyzing streaming data, or running needed multiple operations. Spark was first created for Hadoop, but data scientists believe that they work well together in the modern world for several large data applications.

III. LITERATURE REVIEW

A. Multiple Big Data Analytics Platforms

A data analytics platform is an array of resources and technology that helps one to access, integrate, interact with, discover, and simulate data from different sources that an organization might have. A robust big data analytics platform combines a variety of technologies for different features, such as statistical big data analytics and visualization, as well as location intelligence, natural language processing, and content analytics. Its key goal is to transform every type of data into meaningful observations that lead to real operational efficiency [8]. In software engineering, modern methods to programming, concurrent processing and subsequent distribution of computer programming models, and new programming systems were built in helping software developing companies. This is where Hadoop architecture, an open-source version of the MapReduce programming model that often uses a distributed file structure, takes the lead. Nevertheless, following its release, there have been evolutions to the MapReduce programming model as well as alternative programming models implemented by Spark and Storm architectures, all of which have seen positive results [9]. Hadoop and Spark are the two major large data analytics platforms that will be addressed in this paper.

1. Hadoop

The Hadoop Open Source distributed computing architecture offers stable, flexible, distributed computing, large-scale cluster computing analyses, data storage, including distributed file system HDFS, NoSQL [10]. database, distributed computation MapReduce and distributed HBase. Though the Hadoop framework can distribute data and hold vast volumes of data, a significant amount of data may require technical analysis [10].

However, R is unable to read data that is larger than the amount of memory available on the computer, so there is a limit on the amount of data when a large amount is to be processed. Hence, integrating Hadoop and RHadoop is very important in making sure that the distribution and processing of the vast amount of data go smoothly [11]. In this manner, R can not only manage technical analytics but will also make it simple to use Hadoop functions, like the ability to access HDFS via the rhdfs module and to leverage MapReduce for distributed computing via the rmr2 package [12].

Some tools are built on top of Hadoop applications. First, using a basic scripting language named Pig Latin, Apache Pig can solve routine MapReduce conversions on massive amounts of data [12]. Secondly is the Apache Hive which is a data warehouse framework that uses a SQL-style language named HiveQL to query and handle massive databases in distributed storage. Another one is Apache Sqoop is a platform for moving vast volumes of data as quickly and easily as possible through Hadoop and organized data storage. Apache Flume is also a distributed log storage framework that is extremely versatile and can be used to record data, process log data, and transmit log data [12]. Apache Zookeeper is another distributed technology for operation synchronization that is mostly used to manage clustered applications that are often found in data management challenges. Final Apache Avro is a data serialization

framework aimed at supporting large volumes of data sharing and intensive data.

2. Spark

Spark platform can be described as an open-source parallel computing system that facilitates in-memory processing and enhances the speed of analyzing data. Spark is also optimized for high reliability, fault tolerance, and quick computation Spark is an excellent choice for graph computing, machine learning, as well as big data analytics. The principal functions and orientation are identical to those of Hadoop MapReduce. It aims to reduce the I/O delay generated by a large number of loop files switched between memory and disk throughout MapReduce using in-memory cluster computing [12,13]. The processor speed may theoretically be numerous times faster than Hadoop. Spark utilizes the Scala language, but can also support Java, and Python, while its internal storage structure is HDFS-compatible [13].

a. Spark's Features and Functions

The key benefits of Spark over Hadoop MapReduce are the ability to write jobs with several phases using its functional programming API and the capacity to retain intermediate memory space. Spark's capabilities were shown using its built-in libraries, which were implemented using the Spark Core Engine, as well as the ability to access additional libraries from the Spark Packages repository [13].

Fig i: Spark Packages repository

Apache Spark is a highly customizable cluster computing platform for handling real-time data. Apache Spark's primary functionality is in-memory cluster computation, which significantly improves a program's processing capability. Spark offers a programming framework for whole clusters with tacit parallel computing and routing protocols [13]. It encompasses a wide variety of workloads such as batch processes, iterative algorithms, dynamic searches, and streaming.

Fig ii: Features of Spark

SparkR is a Spark platform that is integrated with R suite designed to support Resilient Distributed Dataset (RDD) API which enables it to perform distributed computation utilizing Spark. To deploy SparkR one needs to download Spark 1.4 or a new version later as well as R-related modules such as rJava and rhdfs. RJava makes it less complicated for R to call Java-owned tools, like Spark and Hadoop, and rhdfs, including RHadoop, to link with HDFS, when R calls artifacts, objects, and functions which are java - Based Figure 5 depicts the SparkR system [14]. While RHadoop will enable distributed computing with R language, it is not as efficient as SparkR. SparkR, which uses in-memory cluster computation, requires additional capacity than RHadoop [14]. To prevent task shut down due to hardware resource limitations, both RHadoop and SparkR may be built together and used synonymously at the same location. Furthermore, it makes it possible to carry out a distributed computation, we need a corresponding algorithm to evaluate the most appropriate computational methods [14].

B. Significance of Multiple Big Data Analytics Platforms

I. Using Hadoop and Spark together

Combining Hadoop and Spark is a great way to bring more data out of your data. There are some scenarios in which you might want to combine the two tools. Even though some have speculated that Spark would completely replace Hadoop due to the latter's processing capacity, the two are intended to balance rather than clash [14]. A simpler representation of the Spark-and-Hadoop design is seen below:

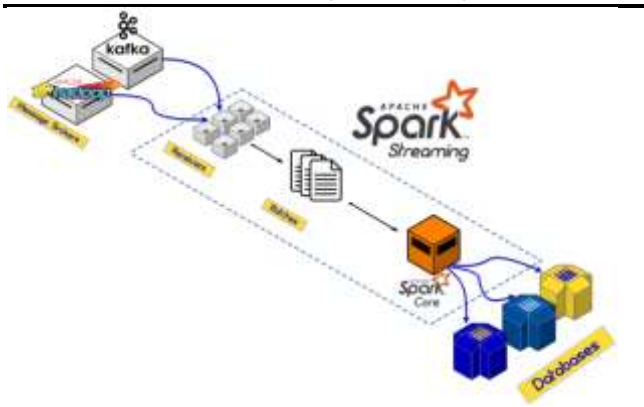

Fig iii: How Spark works together with Hadoop and Kafka

The advantage of having both methods can be seen by companies requiring batch analysis and stream analyses for various resources. Hadoop can handle heavier operations at a cheaper price, while Spark processes several smaller workstations that need immediate retrofitting. YARN also allows it possible, though it is not with Apache Spark, to store and analyze archived information [15]. As a result, Hadoop and, in particular, YARN, became vital threads for linking together real-time computation, deep learning, and recursive graph processing.

II. SparkR and RHadoop enhances Data Science Workflow

The use of SparkR will help to solve the scalability problem that single machine R has and speed up the data analysis workflow. R has more than 10,000 packages and is one of the most important software packages in data analytics. R is an open-source program and is commonly taught in universities and colleges by statistics and IT curricula. R uses a data frame as an API that conveniently handles the data [15]. R has a potent framework for visualization that allows data scientists to easily visualize data. Data processing using R is therefore restricted to storage on a single server and it is also not feasible to use R in massive data sets because R is a single thread [15]. The Spark Group develops the SparkR package to solve the R scalability issue. This package is built on a distributed data structure that allows for the structured processing of information with a user-friendly syntax. Spark includes a distributed computing engine, a data source, and data constructs that are not stored in memory. R offers a diverse atmosphere, interactivity, bundles, and visualization capabilities. SparkR incorporates the benefits of both Spark and R into a single package [16].

Data scientists use a common workflow to solve challenges as the volume of data grows, along with additional technological advancements like SparkR. SparkR, on the other hand, cannot cover all of R's functions, nor is it appropriate to do so since not all functionalities need robustness, but not every dataset is massive. For instance, if the dataset contains 1 billion files, one would not need to work with the entire dataset if the framework is as easy as regression analysis with a few hundred variables, but a random forest algorithm with dozens of design elements may profit from additional details [16]. They must use the appropriate method at the appropriate time. We'll go through some common situations and best practices for combining SparkR and R in the next portion.

Users usually begin with a huge dataset stored as JSON, CSV, ORC, or parquet files on HDFS, AWS S3, or a relational database management system (RDBMS) [16]. The growing use of cloud-based big data computing is a good example of this. Joining the necessary datasets is the first phase, followed by • Data cleaning operations to delete invalid rows or columns; and • Selecting particular rows or columns. This move decreases the size of the dataset since these users usually aggregate or sample their results. These moves are often referred to as data wrangling since they include handling a huge dataset, and SparkR is the best tool for the job. The pre-processed data is then compiled into a local directory, where it is used by a single computer R to create models or conduct other statistical tasks. A traditional data scientist should be comfortable with this, and thousands of CRAN kits are available to them. As data scientists, we have previously used the famous dplyr program to conduct exploratory data analysis and manipulation. In the single-machine age, dplyr offers several fantastic, simple-to-use functions for data processing. SparkR offers the same functions and essentially the same API in the big data age to manage bigger-size datasets [16]. It should be a relatively painless transition for standard R apps.

III. Performance

Spark is 100 times faster in terms of in-memory and ten times better on disk. Additionally, it's been used to process 100 TB of data three times as quickly as Hadoop MapReduce on a tenth of the computers. Spark is quicker in particular for machine learning algorithms like Naive Bayes and k-means [17].

Spark efficiency is superior to Hadoop in terms of processing speed for many reasons:

1. When Spark runs a certain part of a MapReduce task, it is not constrained by input-output considerations. It is much quicker for apps.
2. The DAGs in Spark allow for optimizations in-between stages. Hadoop doesn't have a cyclical

relation between MapReduce stages, so there's no way to tune output at that stage.

If Spark is used in conjunction with other shared resources on YARN, though, the output can suffer and memory leakage can occur. As a result, whenever a customer has a batch processing use case, Hadoop has been discovered to be the more powerful method [16,17].

IV. Costs

Spark and Hadoop are both open-source Apache applications, which means one could theoretically run them with no installation costs. It is however necessary to remember the overall ownership expense, including upkeep, acquisitions of hardware and software, and the recruitment of a team who knows cluster management. On-premises, the main consideration is that Hadoop needs more storage memory and Spark needs more RAM, which means that establishing Spark clusters can be costlier [17]. Furthermore, because Spark is a newer framework, specialists in it are harder to come by and are much costlier. Another alternative is to use a distributor like Cloudera for Hadoop or Spark for DataBricks to download and install, or to execute EMR/MapReduce operations in the cloud using AWS.

The difference of extract pricing information may also be complex because Hadoop and Spark work together, even on EMR cases that are required to run on Spark. If users pick a computer configured EMR cluster for Hadoop for a rather high comparative factor, the cost for the least example, c4. large becomes $0.026 per hour. The most affordable memory-optimized Spark cluster will cost $0.067 every hour. As a result, Spark is much costlier per hour, but if users optimize for computing time, equivalent tasks can require less time on a Spark cluster [17].

Since Hadoop was developed to replicate data through several nodes, it is extremely fault-tolerant. Each file is partitioned into blocks and repeated several times through several machines, meaning that even though a single computer fails, the file may be restored using blocks from other machines.

The fault tolerance of Spark is generally accomplished by RDD operational activities. Data-at-rest is initially located in HDFS, which would be fault-tolerant thanks to Hadoop's framework. When an RDD is developed, a lineage is often developed, which records how the dataset was built and, because it is permanent, maybe rebuilt from scratch if necessary. The DAG can also be used to reassemble data via Spark directories through data nodes. Data is distributed across executor nodes and is susceptible to corruption if a node or contact amongst executors and drivers crashes. Kerberos authentication is supported by both Spark and Hadoop, although Hadoop's HDFS security mechanisms are more fine-grained. Apache Sentry is also another platform accessible explicitly for HDFS-level authentication. It is a framework for implementing fine-grained metadata accessibility. Spark's security architecture is generally limited, but it does support mutual key authentication.

IV. FUTURE IN THE U.S

With big data growing faster in the United States, multiple big data analytics platforms will be implemented to improve the operations of many companies. The amount of data being stored will keep growing and move to the cloud. Most big data scientists agree the volume of data generated would increase massively in the future. As bandwidth requirements continue to increase, a significant difference exists between the need for and supply of data practitioners.

V. ECONOMIC BENEFITS TO THE U.S

Approximately 500,000 Big Data jobs are currently available in the United States. However, according to McKinsey, there are between 140,000 and 190,000 employees with graduate degrees in statistics, software engineering, and other related science who are in short supply [18]. Even more concerning is the lack of 1.5 million administrators and analysts who are already employed in conventional roles but can incorporate Big Data analytics systems into their decision-making. Workers' efficiency and wages are likely to rise as a result of the need to comprehend and act on better data. Data collection, transmission, storage, and analysis would become much simpler as a result of ongoing technical advancements [18]. In conjunction with the associated innovations in material science, bioengineering, IT, and nanotechnology, a wide range of emerging technologies will become possible.

According to a study by the McKinsey Global Institute, Big Data may contribute $3 trillion to the value of only seven sectors next year. The United States will benefit from $1.3 trillion of this [18]. Additionally, the study projected that over half of this benefit would be passed on to consumers in the form of reduced road congestion, simplified market comparisons, and improved matchmaking between academic institutions and students. It's worth noting that many of these advantages have little effect on GDP or investment income in the sense that we quantify them. However, they do suggest a higher standard of living.

However, the effect is felt by more than just customers. Businesses who utilize data-driven decision-making enjoy a 5–6% increase in efficiency and production development compared to their rivals, often when accounting for other acquisitions in information technology use. Asset use, return on equity, and market value all showed significant variations. A report on the effect of Open Data initiatives on governance was recently published by the Omidyar Network. According to the survey, implementing these policies may increase annual income in the G20 by $700 billion to $950 billion [18,19]. Positive effects include decreased corruption, enhanced working standards, greater energy usage, and increased international trade.

VI. CONCLUSION

This paper explored how multiple big data platforms with rapid response are significant. The findings show that big data analytical platforms have the same setup and efficiency. The quality in the different operational conditions in the application of optimizing the scheduling of various big data analysis platforms has always resulted in different efficiencies. Hadoop and Spark are two of the most popular data handling distributed systems in the world today. With both the MapReduce paradigm, Hadoop is mostly used for disk-intensive operations, while Spark is a more versatile but more expensive in-memory processing framework. Both are Apache top-level solutions that are often used together and have certain parallels, but it's necessary to know the differences between them before choosing to use them. By optimizing task scheduling, automatically detecting clustering conditions, and then selecting a suitable platform for word processing, performance reliability can be greatly increased. The fundamental central database of a Big Data analytics application cannot be minimized in any way. Data networks, on the other hand, can be viewed as service environments that must be constantly rebalanced. Because of the significance of big data analytics platforms, there is a lot of competition and a lot of requests for big data experts. Data science and analytics is a rapidly developing field with a lot of promise. Analyzing the supply chain of a company and gaining insights are both possible with data analytics. The use of analytics will help analysts gain a better understanding of their business. Specialists in data analytics offer businesses the opportunity to learn about emerging markets. Big data analytics has a wide range of applications and is extremely important in many fields and businesses. As a result, a practitioner needs to stay up to date on these strategies. Companies, on the other hand, will benefit greatly from the proper use of these analytics methods.

References

[1] R. Margolis, L. Derr, M. Dunn, M. Huerta, J. Larkin, J. Sheehan, M. Guyer and E. Green, "The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data", Journal of the American Medical Informatics Association, vol. 21, no. 6, pp. 957-958, 2014.

[2] W. Fan and A. Bifet, "Mining Big Data: current status and forecast to the future", *ACM sIGKDD Explorations Newsletter*, vol. 14, pp. 1-5, 2013.

[3] S. Hoffman, Apache Flume: Distributed Log Collection for Hadoop, Packt Publishing Ltd, Maharashtra, India, 2013.

[4] M. Adnan, M. Afzal, M. Aslam, R. Jan, and A. M. Martinez-Enriquez, "Minimizing big data problems using cloud computing based on Hadoop architecture," in Proceedings of the 2014 11th Annual High Capacity Optical Networks and Emerging/Enabling Technologies (Photonics for Energy), HONET-PfE 2014, pp. 99–103, Charlotte, NC, USA, 2014.

[5] P. Mika and G. Tummarello, "Web semantics in the clouds," IEEE Intelligent Systems, vol. 23, no. 5, pp. 82–87, 2008.

[6] A. Gahlawat, "Big data analytics using R and Hadoop," International Journal of Computational Engineering and Management, vol. 1, no. 17, pp. 9–14, 2013.

[7] Kala Karun and K. Chitharanjan, "A review on Hadoop—HDFS infrastructure extensions," in Proceedings of the 2013 IEEE Conference on Information and Communication Technologies, ICT 2013, pp. 132–137, Tamil Nadu, India, 2013.

[8] J. Lin and D. Ryaboy, "Scaling Big Data mining infrastructure: the twitter experience", ACM SIGKDD Explorations Newsletter, vol. 14, no. 2, pp. 6-19, 2013.

[9] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: from big data to big impact,"

MIS Quarterly: Management Information Systems, vol. 36, no. 4, pp. 1165–1188, 2012.

[10] D. Wickens, "Processing resources in attention dual-task performance and workload assessment," Office of Naval Research Engineering Psychology Program, No. N-000-14-79-C-0658,1981.

[11] PeterAugustine, "Leveraging Big Data Analytics and Hadoop in Developing India's Healthcare Services", International Journal of Computer Applications, vol. 89, no. 16, pp. 44-50, 2014..

[12] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: cluster computing with working sets," in Proceedings of the 2nd USENIX Workshop on Hot Topics in Cloud Computing, pp. 95–101, Portland, Ore, USA, 2010.

[13] S. Chaudhuri, U. Dayal, and V. Narasayya, "An overview of business intelligence technology," Communications of the ACM, vol. 54, no. 8, pp. 88–98, 2011.

[14] M. Maurya and S. Mahajan, "Performance analysis of MapReduce programs on Hadoop cluster," in Proceedings of the 2012 World Congress on Information and Communication Technologies, WICT 2012, pp. 505–510, Trivandrum, India, 2012.

[15] Thusoo, J. S. Sarma, N. Jain, et al., "Hive: a warehousing solution over a map-reduce framework," Proceedings of the VLDB Endowment, vol. 2, no. 2, pp. 1626–1629, 2009.

[16] G. Li, J. Kim, and A. Feng, "Yahoo audience expansion: migration from Hadoop streaming to spark," in Proceedings of the Spark Summit 2013, San Francisco, Calif, USA, 2013, Yahoo, 2017, https://spark-summit.org/2013/li-yahoo-audience-expansion-migration-from-hadoop-streaming-to-spark/.

[17] M. Zaharia, M. Chowdhury, T. Das, et al., "Fast and interactive analytics over Hadoop data with Spark," USENIX Login, vol. 37, no. 4, pp. 45–51, 2012.

[18] R. Chang, H.-F. Tsai, and Y.-C. Tsai, "High-performed virtualization services for in-cloud enterprise resource planning system," Journal of Information Hiding and Multimedia Signal Processing, vol. 5, no. 4, pp. 614–624, 2014.

[19] H. Topcuoglu, S. Hariri, and M. Wu, "Performance-effective and low-complexity task scheduling for heterogeneous computing," IEEE Transactions on Parallel and Distributed Systems, vol. 13, no. 3, pp. 260–274, 2002.