NOVATEUR PUBLICATIONS
INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING RESEARCH AND TECHNOLOGY [IJIERT]
ISSN: 2394-3696
VOLUME 5, ISSUE 10, Oct.-2018

USAGE OF HADOOP AND MICROSOFT CLOUD IN BIG DATA ANALYTICS: AN EXPLORATORY STUDY

Sudhir Allam,
Sr. Data Scientist, Department of Information Technology, USA

ABSTRACT

This paper explores the how Hadoop and Microsoft cloud are used in big data analytics and how it changes in the big data sector. It also discusses these initiatives and how programs address large-scale data analysis and address the concerns of increasingly increasing data files. Big data is a method used to process, transmit and high-speed evaluate large amounts of data [1]. Big data may be informed by an understanding of organized, unstructured, and semi-structured data, which leads to the decline of traditional methods of data processing. Different inputs and the device are used to produce the data at various speeds. After reviewing the basics, further research (Hadoop and Cloud services) is carried out by exploring complementary tools and programs that promote cloud Hadoop [1]. Hadoop is planned to reach thousands of machines from single servers. In the early days of the Internet, the duo decided to invent a means of returning site search results more quickly by the distribution of data and measurements across various machines so that several functions could be performed simultaneously. A cloud-based architecture applies to aggregating elements such as software programs, middleware, and storage servers used for cloud computing. This helps to design, implement and manage cloud-based systems rigorously and is also an effective model for large-scale-up and computerization of the automatically allocated resources. Big Data Analytics (BDA) provides cloud architecture data storage tools to store, analyze and process an enormous amount of data [2]. This paper provides an explorative-based study of the Hadoop and Microsoft cloud framework to understand how it can be used in Big Data analytics. Major companies such as Amazon, IBM Google, and Microsoft have adopted these technologies and their frameworks ideally tailored to their work for researchers, IT analysts, readers, and market users to ensure a successful outcome.

KEYWORD: Big data, Hadoop, Microsoft cloud, MapReduce, HDFS, Cloud services

INTRODUCTION

This huge quantity of data generated at an extremely fast speed and in all types of formats is what we now term big data. However, it is not possible to save these details on the standard structures we have for 40 years [2]. We need a much more complicated architecture to manage this vast data comprising not just of one, but of many modules handling various operations. This vast amount of data, provided by messages, satellite imagery, social networking, email, and much more, can be organized and unstructured, called big data. [2]. as a consequence, different companies have a difficult job to monitor and manage such huge amounts of data that the high-priced data warehouse built by these companies is burdened, culminating in significant processing burdens [3]. To clear this limitation, big data analytics (BDA) is being used, in which organizations are acquiring multiple methods, strategies, and methodologies for getting the correct data from all the vast unstructured data sets [2,3]. Apache Hadoop, an open source, is one of the tools used by companies to provide the latest solutions in warehouse data as well as the processing of data. When combined with different data centers, Hadoop will work out to be inexpensive and efficient. Pig, Jaql, Hive, R-programming, and several others are other systems implemented by companies. The analysis of these broad data sets by BDA promotes IT companies in advance, enhances their sales growth, and enables them to gain a competitive market advantage [4]. However, the increasing prevalence and the tremendous adoption of BDA in the IT industry increases many problems and difficulties, through resolving large-scale and large-scale prices of Big Data [5]. The cloud solves this dilemma. Cloud computing is a wall-to-wall archetype that allows users to quickly provide services over the Internet and ensures access to a shared pool of cloud infrastructure with limited distribution management on request. Even as the expense of acquisition [9] is minimized ultimately, and the advantages of cloud-based big data are already addressed before the game, several leading organizations

provide cloud-based and Hadoop Big Data frameworks. This paper provides a better understanding of the usage of Microsoft cloud and Hadoop frameworks in Big Data Analytics.

RESEARCH PROBLEM

The main problem that this explorative study aims at solving is how Hadoop and Microsoft cloud are used in big data analytics. This topic is important because, for the last 40 years, organizations have used conventional methods to collect and process their data. However, today's data can't be managed by such databases, since the majority of today's data is semi-structured or unstructured. Nevertheless, traditional frameworks only manage organized data in well rows and columns [6]. Relationships The databases are internally scalable because further computation, memory, and storage have to be added to the same structure. This may prove to be quite costly. Today's data were housed in various silos. It can be a challenging job to get them together and analyze them for trends. While the computing ability of application servers has increased over the years, owing to their restricted capacities and speed, databases have lagged [7]. Today, however, when many apps generate big data. Hadoop and Microsoft cloud is essential in providing the database environment with a much-needed upgrade.

LITERATURE REVIEW

BRINGING BIG DATA AND CLOUD TOGETHER

Whilst cloud offers paid-for-application, portable and elastic platforms on request, BDA emphasizes revolutionizing their information properties, which are 3 V's, into another 5 that symbolize value (for businesses) [5]. Any of these big data measurements are shown below. [8]

- Volume: data quantity spread
- Velocity: the speed at which the data is spread.
- Variety: the heterogeneity of the propagated data form

Cloud computing provides the opportunity to adapt large data volumes across the Internet through virtualization by hardware, thereby increasing the availability, scalability, and usability of Big Data[3, 9]. Besides, cloud storage also offers exclusive computational resources for resourceful handling and large data analysis through a program called the Big Data as a Service (BDaaS) [6, 7]. Subsequently, both Large Data and Cloud add benefit to businesses through maximizing versatility, elasticity, usability, and ease of cloud-powered Big Data analytics and by reducing control and the complexities of big data solutions deployment. [7,8].

MICROSOFT CLOUD SUPPORTED BIG DATA

Microsoft Azure provides a broad range of cloud resources, allowing both developers and IT practitioners to design, deliver and manage apps that range from mobile devices to ISC technologies across its global datacenter's grid, with DevOps support and advanced software. Here, too, BDA options.

AZURE HDINSIGHT

Apache HDInsight Hadoop has become a huge thing for Big Data over the last decades, and although its usage is shrinking, it remains extremely strong. It enables users to execute dynamic, distributed analytical activities on almost every data volume. HDInsight lets users build large data clusters with Hadoop [8] and scaling them up or down according to their needs. It combines with Azure resources such as Data Factory or Data Lake Storage, enabling them to use Hadoop data analysis on your existing data. Like Apache Spark, Apache Kafka, HBase, Hive, and Storm, HDInsight has a wide range of common Hadoop tools. It offers customized services from Azure redundancy solutions as the result of technological advancement, protection, enforcement, and data integrity. Unlike many other Manufacturing Facilities, SLA provides 99.99 percent on one instance of a virtual computer just on the essential virtual machines. For example, it offers optimized cluster development for Spark, Hadoop, Kafka, HBase Storm, as well as 99.9% SLA-supported Microsoft R Servers [9]. Any of its key characteristics include worldwide accessibility, high protection and enforcement, highly efficient research, and development network, cost-effectivity, and extensibility.

DATABRICKS OF AZURE

Databricks is an Apache Spark-based analytics service. Apache Spark is a seasoned platform for processing large quantities of unstructured data at fast velocities. Databricks complements languages such as, Scala, Python, Java, SQle, and R and also AI/ML libraries such as TensorFlow and PyTorch that help users to use some of such frameworks with Spark data[9,10]. Databricks also implements Azure Machine Learning, allowing users to have access to a vast range of pre-trained machine implementations. Databricks enables a setup of controlled auto-scale and auto-termination Apache Spark clusters to eliminate the complexities of Spark in the centralized model.

AZURE STREAM ANALYTICS

Azure Stream Analytics enables users to create a stream network. It is built on technologies without servers. Stream Analytics allows users to describe the stream processing analytics pathway, including data processing described utilizing SQL syntax and output in minutes. It grows exponentially dynamically based on the stream bandwidth utilization and performance. Although streaming data also needs high efficiency and fast and accurate response, Azure Stream Analytics provides sub-second latency with assured "precisely once" event handling. It also provides accessibility of 99.9 percent[11].

HADOOP SUPPORTS ADVANCED ANALYTICS

In comparison to standard tools, Hadoop has more precise facts and figures. Hadoop embraces innovative features such as data analysis and predictive analytics to include and depict practical information in the most visually appealing way possible. It may aid in the optimization of results utilizing a single server and the handling of large amounts of data.

Hadoop is a cost-effective approach for both large and small businesses, making it a compelling solution for limitless possibilities. Companies and businesses are becoming more familiar with Hadoop as time goes by. They're working on using big data to help with marketing and other tools. The following are some of its most important tools.

HADOOP DISTRIBUTED FILE SYSTEM

The HDFS file system is written in Java and is designed to be distributed, scalable, and portable. It achieves so by offering shell commands and java user interface approaches that are close to those used by other file systems, rather than keeping data in a data store due to the absence of POSIX enforcement. A Hadoop cluster that needs replication solutions and is accessible for the name node due to its essential circumstances can be hosted on a single name node as well as a cluster of data nodes[12]. Thus every data node can be supported by the HDFS-specific block protocol. The file system communicates through sockets referred to as TCP/IP sockets. Remote calls (RPCs) are used to connect with customers. Multiple computers can keep huge files in the region of gigabytes to terabytes. It ensures durability by reproducing the data through different hosts. It needs no overlapping array of separate risks for data storage in hosts. Any raid setups are used to improve input and output efficiency. When the replication value is set to 3, data is saved on three nodes. The data from the first and second nodes will be processed on the same rack, while the data from the third nodes will be placed on a separate rack[14]. Data nodes can communicate with one another to reconfigure data and transfer backups around to maintain data replication high.

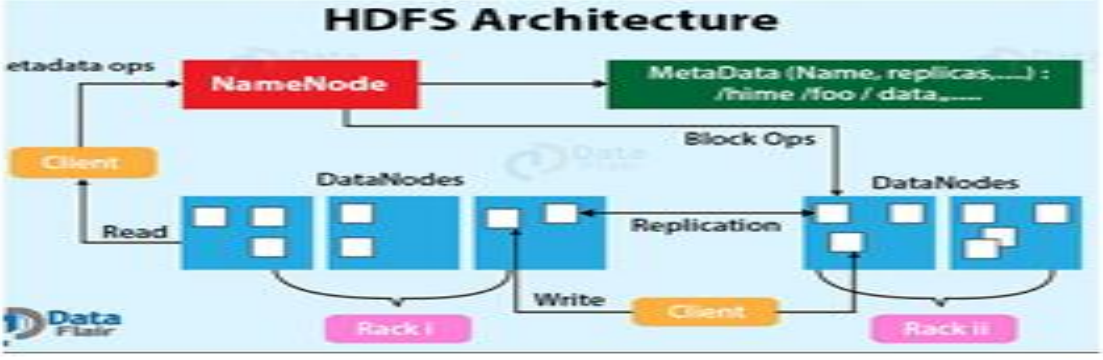

Fig I: HDFS architecture

HADOOP MAPREDUCE

Map-reduce is a program that dependably writes massive amounts of data in tandem or on huge clusters of commodity hardware. It is focused on a distributed computing architecture and a programming environment. The MapReduce algorithm is made up of two main tasks: map and reduces [14,15]. By using a map, a collection of data can be transformed into another set of data, with each variable being broken down into tuples. The tuples may be used to represent both core and value components. Reduce uses the data of a chart as input and condenses the tuples of the data into a smaller collection of tuples[16]. When the chart has completed its mission, the minimize task is completed. MapReduce allows data processing to be scaled through a large number of computing nodes with ease. The terms mappers and reducers are used to describe data processing primitives. When the decomposition of data processing applications takes place in Mappers and Reducers, it would be a non-trivial operation according to the map-reduce model[17]. When we compose a request in the MapReduce method and size the query over dozens of devices in a cluster, the configuration shifts. That's the process that draws a large number of users to this model.

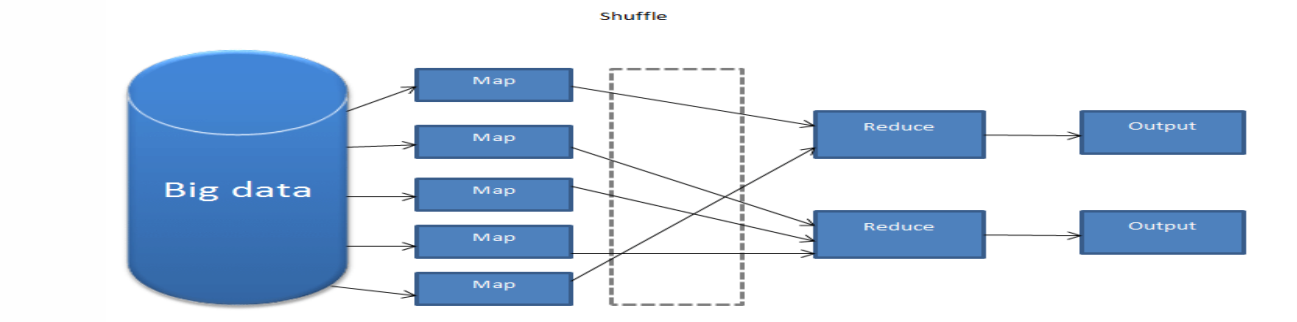

Fig ii: The MapReduce architecture

HOW DOES HADOOP HELP IN BIG DATA ANALYTICS?

Standard databases have two major problems that Hadoop solves:

CAPACITY: HADOOP IS CAPABLE OF STORING VAST AMOUNTS OF DATA.

The data is divided into blocks and saved through clusters of commodity servers utilizing a distributed file system named HDFS (Hadoop Distributed File System). These commodity systems are cost-effective and quickly scalable as data expand since they are designed with basic hardware configurations[13,14].

QUICKER DATA STORAGE AND RETRIEVAL: HADOOP IS FASTER AT STORING AND RETRIEVING DATA.

Hadoop performs concurrent computation across data sets using the MapReduce functional programming framework [17]. As a result, rather than processing data sequentially, functions are separated and executed simultaneously through distributed networks when a question is submitted to the database. Finally, the results of both assignments are compiled and submitted back to the application, resulting in a significant increase in processing time.

FUTURE IN THE UNITED STATES

While several of Microsoft's apps are built on Azure, the business is progressively introducing Azure services to allow consumers to extend and configure their products. Since public clouds are already in their development, critical areas such as reliability and security will eventually shift. Data indicates that public cloud emails are more reliable than other on-site options. Because the majority of abused vulnerabilities target out-of-date applications, automating patching and updating cloud services greatly improves the security of all data and software in PaaS [17]. Many technology experts agree that there are no fundamental reasons why public clouds would be less reliable; in fact, they are expected to become more secure than on-premises due to the intense scrutiny operators will place on security and the depth of information they are accumulating. Thousands of small and large companies in the United States and around the world have now gained from Microsoft's technologies, software, services, and network of partners as they focus on their digital

transformation journeys. In the future, they want to bolster our innovation capabilities by developing business clouds for automobiles, retail, financial services, non-profits, and healthcare.

ECONOMIC BENEFITS TO THE UNITED STATES

Many companies in the United States are using Hadoop and the Microsoft cloud to extend their operations and reach new markets. Microsoft's overall costs for analytics and business intelligence offerings were lowered by an average of 21.9 percent as compared to rival alternatives. The Hadoop big data analytics market was projected to grow at a CAGR of 13.0 percent during the forecast period, from USD 12,797 million in 2020 to USD 23,526 million in 2025 [18]. The US will contribute around 30.6 percent of the world market share in 2020. By 2020, the Hadoop market in the United States is estimated to be worth $9.3 billion. The country presently has a 30.55 percent share of the global economy. China, the world's second-largest economy, is forecast to reach a market cap of US$116.6 billion in 2027, a pace of annual growth of 54.9 percent.

CONCLUSION

This thesis delves into the use of Hadoop and Microsoft Cloud in big data analytics. Technological advancement is always introduced while handling large quantities of data. Analytical expertise is needed to evaluate the large data. The method is economical and applicable to a broad variety of technological realms. Although some major obstacles to cloud adoption exist today, they are projected to diminish over time. Although new, unexpected obstacles to public cloud adoption can exist, the public cloud's economic advantage may improve over time when cloud providers recognize the economic benefits. This Hadoop and Microsoft cloud-based big data framework paradigm envisions its usage as a means of offering reasonable and efficient computational solutions to problems in contrast to computationally unavailable conventional machine learning algorithms, allowing improved predictive modeling and decision making, as well as quick and accurate real-time research with expanded utilization out of the same crate.

REFERENCES

1) S. Lee, K. Grover and A. Lim, "Enabling actionable analytics for mobile devices: performance issues of distributed analytics on Hadoop mobile clusters", Journal of Cloud Computing: Advances, Systems and Applications, vol. 2, no. 1, p. 15, 2013.
2) S. Alenezi and S. Mesbah, "Big Data Spatial Analytics in Social Networks using Hadoop", International Journal of Computer Applications, vol. 128, no. 14, pp. 21-26, 2015. M. Xiangning, "Research on Image Storage Based on 3D Point Cloud", Big Data and Cloud Innovation, vol. 1, no. 1, 2017.
3) Bernstein, "The Emerging Hadoop, Analytics, Stream Stack for Big Data", IEEE Cloud Computing, vol. 1, no. 4, pp. 84-86, 2014.
4) L. Greeshma and G. Pradeepini, "Big Data Analytics with Apache Hadoop MapReduce Framework", Indian Journal of Science and Technology, vol. 9, no. 26, 2016.
5) J. Issa, "Performance characterization and analysis for Hadoop K-means iteration", Journal of Cloud Computing, vol. 5, no. 1, 2016.
6) K. Jabeen, "Scalability Study of Hadoop MapReduce and Hive in Big Data Analytics", International Journal of Engineering and Computer Science, 2016.
7) J. Reyes-Ortiz, L. Oneto and D. Anguita, "Big Data Analytics in the Cloud: Spark on Hadoop vs MPI/OpenMP on Beowulf", Procedia Computer Science, vol. 53, pp. 121-130, 2015.
8) J. PJacob and A. Basu, "Performance Analysis of Hadoop Map Reduce on Eucalyptus Private Cloud", International Journal of Computer Applications, vol. 79, no. 17, pp. 10-13, 2013.
9) D. 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.
10) P. Paul and D. Veeraiah, "Multi-Layered Security Model for Hadoop Environment", International Journal of Handheld Computing Research, vol. 8, no. 4, pp. 58-71, 2017.
11) P. Gupta, P. Kumar, and G. Gopal, "Sentiment Analysis on Hadoop with Hadoop Streaming", International Journal of Computer Applications, vol. 121, no. 11, pp. 4-8, 2015.
12) S. Alenezi and S. Mesbah, "Big Data Spatial Analytics in Social Networks using Hadoop", International Journal of Computer Applications, vol. 128, no. 14, pp. 21-26, 2015.

13) R. Singh and P. Kaur, "Analyzing performance of Apache Tez and MapReduce with hadoop multinode cluster on Amazon cloud", Journal of Big Data, vol. 3, no. 1, 2016.
14) Pavlo, E. Paulson, A. Rasin, D. J. Abadi, D. J. DeWitt, S. Madden, and M. Stonebraker. A comparison of approaches to large-scale data analysis. SIGMOD '09, New York, NY, USA, 2009.
15) Jlassi, P. Martineau, and V. Tkindt. Offline scheduling of map and reduce tasks on Hadoop systems. In CLOSER 2015, May 2015.
16) Y. Samadi, M. Zbakh and C. Tadonki, "Performance comparison between Hadoop and Spark frameworks using HiBench benchmarks", Concurrency and Computation: Practice and Experience, vol. 30, no. 12, p. e4367, 2017.
17) P. Vasconcelos and G. de Araújo Freitas, "Evaluating Virtualization for Hadoop MapReduce on an OpenNebula Cloud", International Journal of Multimedia and Image Processing, vol. 4, no. 34, pp. 234-244, 2014.
18) S. Xie, "Investigation on Fast Response Performance of Dam Deformation Monitoring System with Wireless Sensor and Virtualizing Technique", International Journal of Multimedia and Ubiquitous Engineering, vol. 11, no. 7, pp. 193-204, 2016.