ISSN: 2348 9510
International Journal Of Core Engineering & Management (IJCEM)
Volume 3, Issue 3, June 2016

A REVIEW ON BIG DATA ENVIRONMENT ON DIFFERENT FRAMEWORKS, TECHNIQUES AND TOOLS

Ankush Verma
Research Scholar, Pacific University, Udaipur
ankush.verma08@rediffmail.com

Ashik Husain Mansuri
Research Scholar, Pacific University, Udaipur
aashiqmansuri@gmail.com

Dr. Neelesh Jain
Professor , Sagar Institute of Technology, Bhopal
neeleshcmc@gmail.com

Abstract

The term 'Big Data' is not just the data in terabytes, but also used to get information and knowledge in a response time. The Big Data has spread in the frameworks of cloud computing and business intelligence for the organization. Thru this paper we have tried to explore Big Data and the research work going on this field. Also we have tried to discuss different programming frameworks, techniques and tools used by organization for analysis its large-scale data where as analytics almost all the framework are based on the MapReduce scheme, Hadoop and open-source implementation.

Keywords: Big Data, frameworks, Hadoop, MapReduce, cloud computing.

INTRODUCTION

Today the big data have a lot of attention in term of techniques to process, analyze, visualize and storing potentially in a large data sets at reasonable time. For the last years, the field of "Big Data" has emerged as the new frontier in the wide spectrum of IT-enabled innovation and opportunities allowed by the information revolution. Through the beginning of the 21st century and the information explosion accelerates each year by 40 percent. This is called the "Big Data revolution" and it is not only big in volume, it is also big in variety and velocity meaning different types of data at a wide range of input speeds and refresh frequencies. Every day, we create trillions of data all over the world.

ISSN: 2348 9510
International Journal Of Core Engineering & Management (IJCEM)
Volume 3, Issue 3, June 2016

Big Data analytics is the process of analyzing and mining Big Data – can produce operational and business knowledge at an unprecedented scale and specificity. The advanced technology is to be used for storage, processing, and analysis of Big Data. Large firms such as Google, Yahoo, Oracle, Amazon, IBM, Microsoft, Twitter, Facebook and Amazon are providing cloud models that incorporate sound data storage solutions. Hadoop MapReduce clusters are one of the most popular cluster deployments in the cloud for big data [1]. Large firm has its own framework, tools and techniques for Big Data according to the use of its firm for future predicts to help them business growing organization. The data is from social networking sites, ecommerce, scientific experiments, mobile conversations, sensor networks, government, banking/insurance, manufacturing and various other sources. We have new tools and techniques to organize, manage, store, process and analyze Big Data.

LITERATURE REVIEW ON TECHNIQUES AND TOOLS

Hadoop - The initial version of Hadoop was created in 2004 by Doug Cutting it is a framework that can be installed on a commodity Linux cluster to permit large scale distributed data analysis. It is based on the MapReduce programming model consists of several modules which provide different parts of the necessary functionality to distribute tasks and data across a cluster [2]. The Hadoop framework is designed to provide shared storage and analysis infrastructure. The storage portion of the Hadoop framework is provided by a distributed file system solution such as HDFS, while the analysis functionality is presented by MapReduce.

MapReduce - MapReduce programming model has been used at Google for many different purposes for the generation of data of Google's production web search service, sorting, data mining, machine learning and many other systems with parallel and distributed systems. Implementation of MapReduce for large clusters of machines comprising thousands of machines [3]. MapReduce adopts a flexible computation model with a simple interface consisting of map and reduce functions. MapReduce is well-recognized for its scalability, flexibility, fault tolerance and a number of other attractive features [4]. . There are two basic operations for MapReduce on the data i.e. Map and Reduce function.

Mappers are applied to all input key-value pairs, which generate an arbitrary number of intermediate key-value pairs. Reducers are applied to all values associated with the same key. Between the map and reduce phases lies a barrier that involves a large distributed sort and group by.

Map : (k1; v1) ! [(k2; v2)]

Reduce : (k2; [v2]) ! [(k3; v3)]

HDFS - The Hadoop Distributed File System (HDFS) is designed to store very large data sets and to stream data sets to user applications. In a large cluster, thousands of servers both host directly attached storage and execute user application tasks [5]. Hadoop provides the robust, fault-tolerant Hadoop Distributed File System (HDFS), inspired by Google's file system as well as a Java-based API that allows parallel processing across

ISSN: 2348 9510
International Journal Of Core Engineering & Management (IJCEM)
Volume 3, Issue 3, June 2016

the nodes of the cluster using the MapReduce paradigm .

Hadoop MapReduce Model

The Hadoop MapReduce act as a master - slave architecture where one master node manages a number of slave nodes and its components are.

- **Name Node** – Is the Master node which is responsible for storing the meta-data for all the files and directories. It has information such as the blocks that make a file, and where are those blocks located in the cluster.
- **Data Node** – It is the Slave node that contains the actual data. It reports information of the blocks it contains to the Name Node in a periodic [6].
- **Job Tracker** – Job Tracker is a centralized program that keeps track of the slave nodes. Schedules, allocates and monitors job execute on slaves. It runs MapReduce operations.

- **Task Tracker** - executes on each of the slave nodes run MapReduce operation.

Other then Hadoop, MapReduce model which is widely used by almost in all the framework and techniques for big data analysis, data mining as:

Pig - Pig Latin that have designed to spot between the declarative style of SQL which is familiar and the low-level procedural style of map-reduce [7]. Pig is a high level data flow system with custom map and reduces functions or executables. Its program is compiled into sequence of map-reduce jobs and execute in the Hadoop environment an open source implementation using by Yahoo [8].

DryadLINQ - Microsoft has DryadLINQ for academic use, allowing users to new programming model and runtime that is capable of performing large scale data intensive analyses. In this they represent in applying DryadLINQ for a series of scientific data analysis applications, identify their mapping to the DryadLINQ programming model. Microsoft DryadLINQ use parallel processing. In these frameworks, the data is in nodes of clusters to perform data processing [9].

HaLoop - HaLoop support MapReduce for improves their efficiency by making the task scheduler loop aware and by adding various caching mechanisms. HaLoop, a modified version of the open source MapReduce implementation Hadoop. It utilizes map-reduce pairs with a single pipeline in the loop body, uses caches data to disk. HaLoop relies on a distributed file system [10]. It evaluate on real queries and data sets.

Pregel - Many practical computing problems that concern large graphs, such as the Web graph and various social networks. Pregel is a distributed programming framework, focused on providing users with a natural API for programming graph algorithms while managing the details of distribution invisibly, including messaging and fault tolerance [11].

ISSN: 2348 9510
International Journal Of Core Engineering & Management (IJCEM)
Volume 3, Issue 3, June 2016

Starfish - Starfish, a self-tuning system for big data analytics. Build on Hadoop to get good performance automatically throughout the data lifecycle in analytics without any need on their part to understand and manipulate the many tuning in Hadoop [12].

Hive - Hive, an open-source data warehousing solution built on top of Hadoop developed by Facebook. It supports queries expressed in a SQL-like declarative language - HiveQL, which are compiled into MapReduce jobs that are executed using Hadoop [13]. Hive is designed to handle large amounts of data and store them in tables like a relational database management system or in a data warehouse while using the parallel and batch processing functionalities of the Hadoop

Cloud Computing - The Big Data processing in the context of cloud computing including cloud storage and computing architecture with parallel processing framework, major application and optimization of MapReduce. The key issues of big data processing, including cloud computing platform, cloud architecture, cloud database and data storage scheme [14].

Spark - A framework called Spark that supports fault tolerance of MapReduce. To achieve these goals, Spark introduces resilient distributed datasets (RDDs). An RDD is a read-only collection of objects partitioned across a set of machines that can be rebuilt if a partition is lost this means that RDD is reconstructed if node fails and reuse it in multiple MapReduce [15]. Several parallel also performed on RDD for reduce to combine dataset, collect it and passes to user provider functions.

SQL-MapReduce - Analysts to data scientists in this discovery process we feature rich, fast, and flexible discovery platform that is SQL-MapReduce. Data scientists consists of four steps: data acquisition, data preparation, data analysis, and data interpretation support parallelization of Complex Operations, Simplification of Queries, Big Data Access, Multi data store access, High-speed analysis [16].

Grid Computing - Grid Computing offered the advantage about the storage capabilities, the processing power, concept of distributed computing and the Hadoop technology is used for the implementation purpose of Big Data. The benefit of grid computing center is the high storage capability and the high processing power which makes the big contributions among the scientific research, help the scientists to analyze and store the large and complex data [17].

FlexAnalytics - To reduce data movement and release severe I/O performance bottleneck a flexible data analytics framework and placement strategies are built .It is useful for data pre-processing, runtime data analysis, visualization and for large-scale data transfer with features profiling and real-time system resource status or for monitoring [18] .

RFM - Intelligent precision marketing framework has been designed combined with smart technology for its functional structure and operational processes for Big Data into E-commerce and traditional retail industry. Analyze the data of with the improved RFM model framework. divided into three: the data layer, analysis layer and decision-making layer [19].

ISSN: 2348 9510
International Journal Of Core Engineering & Management (IJCEM)
Volume 3, Issue 3, June 2016

DDP - The Hadoop data placement policy (DDP) is applied in a homogeneous cluster, the Hadoop uses of the resources of each node to allocate data blocks, thereby improving data locality and reducing the additional overhead to enhance Hadoop performance [20].

Big Data Deep Learning - As the data keeps getting bigger, deep learning is coming to play a key role in providing Big Data deep learning is providing big data predictive analytics solutions, particularly with the increased processing power and the advances in graphics processors analytics solutions with learning techniques [21].

Sensor Data Management - Model-view sensor data management, which stores the sensor data in the form of key-value stores segments, namely KVI-index- Scan–MapReduce. KVI-index consists of two interval indices on the time and sensor value dimensions respectively, each of which has an in-memory search tree and a secondary list materialized in the key-value store [22].

HBase - Combine cloud computing and big data analysis technology to provide image processing cloud infrastructure and HBase based Big Data processing engine to satisfy the needs**.** HBase using Hadoop for increase in performance. It use cloud architecture to handle random real-time reading and writing of Big Data [23].

Storm - Storm is a real time fault-tolerant and distributed stream data processing system handle real-time stream data management tasks that are crucial to provide Twitter services. Storm is designed for easily add or remove nodes from the Storm cluster with user interactions on Twitter [24].

YARN - The YARN was started to give Hadoop the ability to run non MapReduce jobs within the Hadoop framework. YARN provides a generic resource management for implementing distributed applications, resource management and job scheduling/monitoring [25].

AWS - The Amazon Web Service (AWS) ecosystem is specifically designed to handle this growing amount of data and provide ways to your business can collect and analyze it. The Amazon Web Services ecosystem of analytical solutions is designed to handle growing amount of data and analyze it for cost-effective, fast query, easy-to-use cloud computing platform using parallelizing and distributing queries across multiple nodes [26]. Using Amazon Elastic MapReduce adopting Apache Hadoop across many compute nodes in a cluster.

MRAM - Processing of big data using a powerful machine is not efficient solution. a new framework is proposed to improve big data analysis overcome Hadoop. The proposed framework is called MapReduce Agent Mobility (MRAM). MRAM is developed by using mobile agent and MapReduce paradigm under Java Agent Development Framework (JADE). MRAM send met-data contains map of network and all data about tasks and dependences between them [27].

Kafka - Apache Kafka is an Open Source messaging system used for large data transfer across several data sources via a centralized cluster with efficient transfer of Video / Image data from multiple image sources.

ISSN: 2348 9510
International Journal Of Core Engineering & Management (IJCEM)
Volume 3, Issue 3, June 2016

Apache Kafka provide a high-throughput, low-latency platform for handling real-time data feeds transfer live data using Connected Component Labeling (CCL) [28]. The CCL Image analytics is implemented in Storm topology.

CONCLUSION

Traditional data processing, analyze, visualize and storing approaches are facing many challenges in continuously increasing computing demands of Big Data. Issues and challenges in frameworks of MapReduce, Hadoop, real time faces when dealing with Big Data are identified and categorized. Organization has to work on design and develop a new, updated tools and technologies which effectively handle the processing of Big Data. They also Plan to build a cost based optimizer and adaptive optimization techniques with more efficient plans.

REFERENCES

[1] Xu Zhao, Ling Liu, Qi Zhang and Xiaoshe Dong, "*Improving MapReduce Performance in a Heterogeneous Cloud: A Measurement Study*", China.

[2] Nils Braden, "The Hadoop Framework".

[3] Jeffrey Dean and Sanjay Ghemawat, "*MapReduce: Simplified Data Processing on Large Clusters*", Communications of the ACM, Vol. 51, Jan 2008.

[4] Feng Ii, Beng Chin, M Tamer and Sai Wu, "*Distributed Data Management Using MapReduce*".

[5] Konstantin Shvachko, Hairong Kuang, Sanjay R and Robert C, "*The Hadoop Distributed File System*".

[6] Online article:- https://learnhadoopwithme.wordpress.com/tag/datanode/

[7] Christopher Olston, Benjamin Reed, Utkarsh Srivastava and Ravi Kumar, "*Pig Latin: A Not-So-Foreign Language for Data Processing*", ACM, Vancouver, BC, Cannada, 2008.

[8] Alan gates, Olga Natkowich, Shubham Chopra and Pradeep Kamanth, "*Building a high level data flow system on the top of Map-Reduce The Pig experience*", ACM, VLDB, August 2009.

[9] Jaliya Ekanayake, Thilina Gunarathn and Geoffrey Fox, "*DryadLINQ for Scientific Analyses*".

ISSN: 2348 9510
International Journal Of Core Engineering & Management (IJCEM)
Volume 3, Issue 3, June 2016

[10] Yingyi Bu Bill Howe & Magdalena Balazinska Michael D. Ernst, "*HaLoop: Efficient Iterative Data Processing on Large Clusters*", 36th International Conference on Very Large Data Bases, Singapore, Vol. 3, September 2010.

[11] Grzegorz Malewicz, Matthew Austern, Aart Bik and James Dehnert, "*Pregel: A System for Large-Scale Graph Processing*", ACM, June 2010.

[12] Harold Lim, Gang Luo, Nedyalko Borisov and Liang Dong, "*Starfish: A Selftuning System for Big Data Analytics*", 5th Biennial Conference, California, January 2011.

[13] Ashish Thusoo, Joydeep Sarma, Namit Jain and Zheng Shao, "*Hive – A Petabyte Scale Data Warehouse Using Hadoop*".

[14] Changqing Ji, Yu Li and Wenming Qiu, "*Big Data Processing in Cloud Computing environments*", International Symposium on Pervasive Systems, Algorithms and Networks, Spain , IEEE, 2012.

[15] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin and Scott Shenker, "*Spark: Cluster Computing with Working Sets*", Berkeley.

[16] Rick F and Van Der Lans "*Discovering Business Insights in Big Data Using SQL-MapReduce*", Teradata, July 2013.

[17] Garlasu and D.Sandulescu, "*A Big Data implementation based on Grid Computing*", Grid Computing, Jan 2013.

[18] Hongbo Zoua, Yongen Yub, Wei Tang and Hsuan-Wei, "*Flex Analytics: A Flexible Data Analytics Framework for Big Data Applications with I/O Performance Improvement*", Elsevier, July 2014.

[19] Jianhui Zhang1 & Junxuan Zh, "*Research Intelligent Precision Marketing of E-commerce Based on the Big Data*", Journal of Management and Strategy, Feb 2014.

[20] Chia-WeiLee, Kuang-YuHsieh, Sun-YuanHsieh, Hung-ChangHsiao, "*A Dynamic Data Placement Strategy for Hadoop in Heterogeneous Environments*", Elsevier, July 2014.

[21] Xue Wen Chen, "*Big Data Deep Learning: Challenges and Perspectives*", IEEE, May 2014.

[22] Tian Guo, Thanasis G.Papaioannou & KarlAberer, "*Efficient Indexing and Query processing of Model-View Sensor Data in the Cloud*", Elsevier, Switzerland, July 2014.

[23] R. Saraswathy, P. Priyadharshini, P. Sandeepa, "*HBase Cloud Research Architecture for Large Scale Image Processing*", IJARCSSE, Vol. 4, Issue-12, December 2014.

[24] Jignesh M. Patel, Ankit, Siddarth & Karthik, "*Storm Twitter* ", SIGMOD', ACM, USA, June 2014.

ISSN: 2348 9510
International Journal Of Core Engineering & Management (IJCEM)
Volume 3, Issue 3, June 2016

[25] Arun Murthy, Jeff Markham, Vinod Kumar Vavilapalli, and Doug Eadline, "*Apache Hadoop Yarn*", 2014.

[26] Erik Swenson, "Big *Data Analytics Options on AWS*", Amazon Web Services, pages 10-12, Dec 2014,.

[27] Youssef M. ESSA, "*Mobile Agent based New Framework for Improving Big Data Analysis*", IEEE, 2014.

[28] Lokesh Babu Rao and C. Elayaraja, "*Image Analytics on Big Data In Motion Implementation of Image Analytics CCL in Apache Kafka and Storm*", Vol.3, Issue-3, Mar 2015.