A Review of Big Data Techniques

Akash Kumar, Ruchi Gairola

Abstract-In today's era everyone has a lot of data. Data size is not only in TB(Terabytes), but it has reached up-to ZB(ZetaByte) and PB (Peta Byte). So, this different and enormous amount of data requires some best technique to store and manage and retrieve the data smoothly and correctly. Researches were on their way from a long time but now the waiting is over and the new revolutionary technology has entered in the world of data storage and management and i.e. "Big-Data".

Index Terms— Big Data, Hadoop, Map Reduce, Hadoop Distributed File System, Big Data Challenges

1 INTRODUCTION

In early time, people store data in files and papers. Then came the small storage devices having storing capacity in MB, and with time also reached up to GB (Giga Byte) and TB(Tera Byte).But now this is also not enough for us. So, the biggest revolution in the field of data storage is BIGDATA.As the name suggest it can store enormous amount of data. Today Data Storage has reached its new dimensions, which are:

1 ZetaByte(ZB)=10^21 Bytes
1 YottaByte(YB)=10^24 Bytes
1 XenottaByte(XB)=10^27 Bytes
1 Shilentnobyte(SB)=10^30 Bytes
1 Demegegrottebyte=10^33 Bytes
These are those storage challenges which led to the evolution of this new technology i.e. Big-Data. Data stored in Big-Data is mostly in ZB. Most common data sources are:

2 SOLUTION OF BIG-DATA

Today, the best solution for BIG-DATA is HADOOP

2.1 HADOOP

At first Google launched CFS and MapReduce technique which was published in 2004 to introduce this new technology. After that, Doug Cutting developed an open source version of MapReduce and i.e. called the Apache-Hadoop. Today, Hadoop plays the role of backb0ne in the companies and websites like Facebook, Twitter, LinkedIn and Yahoo. ASF (Apache Software Foundation) is governing the Apache Hadoop which is its core part. It can manage you to access correct data from a huge bulk of data either structured or unstructured without much investment. It is functionally designed to work on innumerous data and maintain its organization for the easy access of the users.

2.2 Apache HIVE

It is DW (Data Warehouse) information built on the top most level in Apache Hadoop. Functions of Hive are ad_hoc query, data summarization and large data sets analysis. Hives helps in integration between tools for data visualization and business approach. Data can be accessed by using an easy query language known as Hive-QL. Hive allows appending, overwriting but not deleting. It supports primitive data type such as string, Boolean, float, binary, decimal-integer, double, smallest and big-int.

2.3 Apache Pig

It allows the users to write compound MapReduce conversion using a simple scripting language. It translates the Pig Latin script into MapReduce to execute it in Hadoop. Pig Latin language defines many data set transformations such as sort, join and aggregate. Java-script helps the users to call functions from the Pig-Latin directly.

2.4 Apache Mahout

A large amount of data in the form of clusters is stored at the servers. In the hub of artificial intelligence. Use MapReduce

- *Akash Kumar is currently pursuing bachelor degree program in Computer Science & Engineering from Shivalik College of Enigeering, Dehradun (Affiliated to Uttarakhand Technical University), Uttarakhand, India.*
- *Ruchi Gairola is currently pursuing M.Tech in the Department of Computer Science & Engineering from Uttarakhand Technical University Dehradun, Uttarakhand.*

paradigm, some algorithm also designed for the better performance, they are Distributed Item-based Collaborative filtering Canopy Clustering, Hierarchical clustering, k-means clustering, Spectral clustering machine. Machine learning is a field of Artificial Intelligence(AI) hub. These machine Works accordingly without being explicitly programmed. 4.Apache Hbase: It is the Hadoop database, a distributed and scalable big data store tools. The main aim of Apache HBAS project is to host a very large table containing billions of row and millions of columns, a top cluster of commodity hardware. It is used for read and write access on the Big-Data. It provides the facility of both read and write to the users. It is best suited for those who works on sparse datasheets. 2.5. Map Reduce Framework: It is the heart of Apache-Hadoop. Due to this a huge number of user can access a large number of servers in a single cluster. The MapReduce Framework become easy for those who know the scale-out data processing solution in clustered. MapReduce refers to two combined works: the first is the map job and second is reduce job. In first step one set of
data is converted into another set of data where individual elements are broken down into key/value pairs Reduce job takes the output from a map as input and combine those data tuples into a smaller set of tuples.

2.5 Hadoop Distributed File System

HDFS is the distributed file system. It is used to store large datasets simply and reliably for those dataset at high bandwidth to user application.

2.6 Apache Flame

Flame is a faithful, distribute and available service for efficient collection and aggregation Of huge amount of data. The architecture of Flame is Flexible and simple. It has some very good tolerating mechanism for recovery and failure. It is used in online Analytic application.

2.7 Apache Sqoop

Data between database and Hadoop is transferred by a command line interface, Apache Sqoop. It is free from SQL query. Hive and Hbase is use to import data from Hadoop into relational DB. Sqoop became the best Apache project in 2012.

3 Module of HADOOP

Hadoop Apache Hadoop Contains different modules to provide different functions. Four main modules of Hadoop are: Hadoop common,HDFS,MapReduce and YARN.Firstly, HADOOP Common is used for proving support other HADOOP modules. Second module is HADOOP distributed File System for storing various types of data and last is Hadoop MapReduce for parallel processing engine and Hadoop YARN provides a framework for job scheduling and cluster resource management. These all are used for organizing BIG-DATA.

3.1. Hadoop Distributed File System

HDFS is used in Fault tolerance and scalability. It stores large file system by breaking them into small parts of 64 or 128MB and create a replica at three or more servers. HDFS also provides an API to read and write in parallel. Performance and capability can be calculated by adding data nodes, and a single name node technique manages data monitors and data placement serves availability. The HDFS detect and compensate for server failure and disk failure.

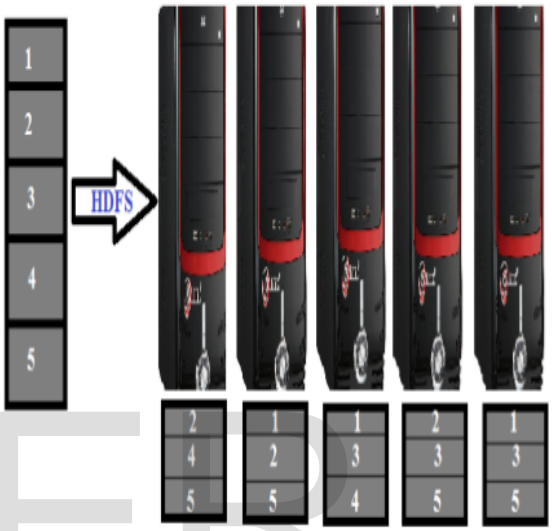


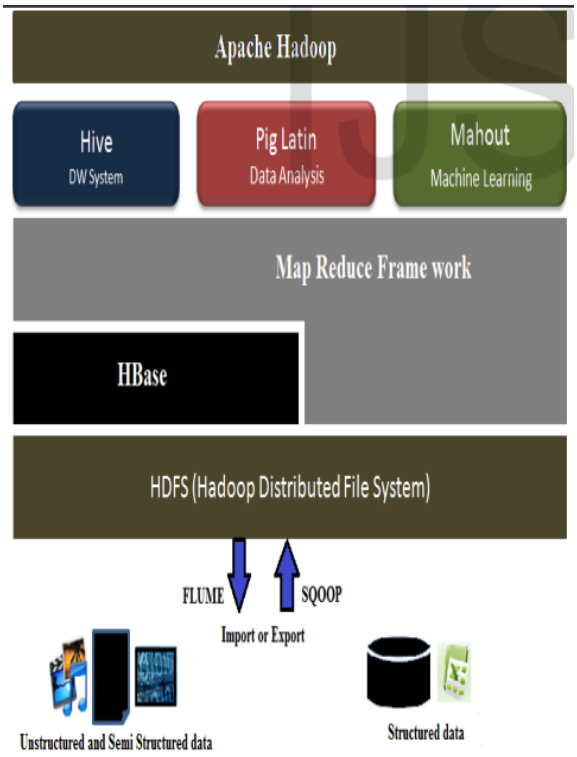
Fig. 4 Apache Hadoop distributed file system

3.2. Apache MapReduce

Centre of Hadoop is the distributed processing frame work known as MapReduce. It corrects data parallel problem for which data can be Subdivided and processed individually. It splits the whole data into sub-parts and then process it and
after processing integrates them as final output. The system is distributed into input data set and multiple chunks; all assigned a map work that can process the data in parallel. The task of each map reads the input as a set of
pairs (value & key) and transforms it into a paired set as output. The intermediate results are also combined to the output for the final result.

4 CHALLENGES AND ITS SOLUTION THROUGH HADOOP

The enormous amount of data is the biggest problem of each and every company. Now, finding correct data is becoming a big problem. Companies have started tighten the snuggle over the data load. Major challenges are storage capture, search, sharing, transferring, analysis and visualization.



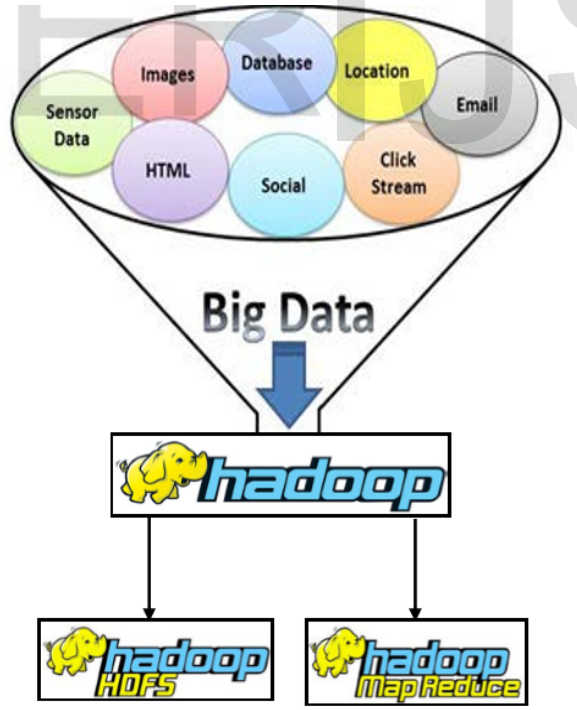
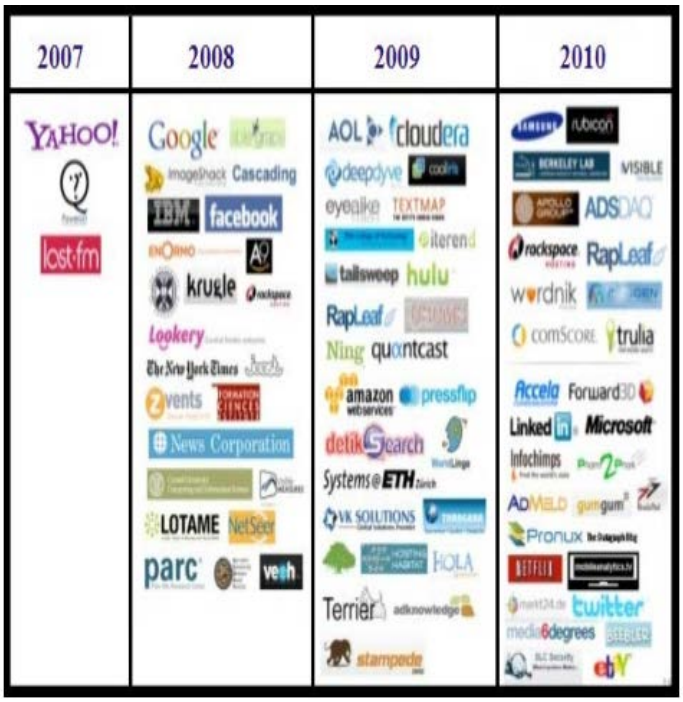


Fig. 3 Big data is the problem and Hadoop is the solution. (With two main techniques) [28][29][30][31]

5 HADOOP USED BY DIFFERENT COMPANIES



In this paper, I explicitly explained the problem of storing and managing data. And the best solution for this is the latest technology HADOOP. Today, the huge amount of data of different companies, government and organizations. Everyone is facing the same problem of storing and the challenge of managing their data. In current scenario Apache Hadoop is the best solution for it. I also explained whole HADOOP ecosystem and its different parts with live example using Oracle application.

7ACKNOWLEDGEMENT

First of all, I would like to thank my mentor Prof. Govind Arya of SCE, Dehradun who gave me this opportunity of writing this paper. I would also like to thank my faculties Prof.Devendra Prasad, Prof. Sandeep Singh Rana who motivated me a lot.

8 REFERENCES

1. International Journal of Scientific & Engineering Research, Volume 5, Issue 6, June-2014 138 ISSN 2229-5518 IJSER © 2014 http://www.ijser.org

2. Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler, "The Hadoop Distributed File System", Shv, Hairong, SRadia, Chansler@Yahoo-Inc.com, IEEE 2010 [2] John Gantz , David Reinsel, "Extracting Value from Chaos", IDC IVIEW, June 2011

3. S. Vikram Phaneendra, E. Madhusudhana Reddy,

"Big Data - Solutions for RDBMS Problems – A Survey", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 9, ISSN: 2278-1021, SEP 2013

4. Contributing Authors, "Big Data Spectrum", Infosys Limited Bangalore India, 2012

5. Papineni Rajesh, Y. Madhavi Latha, "HADOOP the Ultimate Solution for BIG DATA Problems", IJCTT, Vol-4 Issue-4,April 2013

6. Azza Abouzeid, Kamil BajdaPawlikowski, Daniel Abadi, Avi Silberschatz, Alexander Rasin (August 2009), "HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads" VLDB '09, Lyon, France.

7. Jeffrey Dean, Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Google, Inc. , OSDI 2004

8. White Paper Big Data Analytics, " Extract,Transform and Load Big Data with Apache Hadoop", Intel, 2013

9. Umesh V. Nikam, Anup W. Burange, Abhishek A. Gulhane, "Big Data and HADOOP: A Big Game Changer", International Journal of Advance Research in Computer Science and Management Studies, Volume 1, Issue 7, ISSN: 2321-7782, DEC 2013

10. Jaya Singh, Ajay Rana, "Exploring The Big Data Spectrum", IJETAE, Vol-3 Issue-4,April 2013

11. Shilpa, Manjit Kaur, "BIG Data and Methreview", International Journal of Advanced Research in Computer Science and Software Engineering(Ijarcsse), Volume 3, Issue 10, ISSN: 2277 128X ,October 2013

12. Payal Malik, Lipika Bose, "Study and Comparison of Big Data with Relational Approach", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, ISSN: 2277 128X, August 2013

13. http://en.wikipedia.org/wiki/Apache_Hadoop

14. Tom White," Hadoop: The Definitive Guide", O'Reilly Media, Inc,2009

15. http://www.guruzon.com/6/introduction/map-reduce

16. http://hortonworks.com/hadoop

17. http://hbase.apache.org

18. http://www.01.ibm.com/software/data/infosphere/hadoop

19. http://blog.enablecloud.com/2012/06/what-lies-at-core-ofhadoop.html

20. https://oraclecn.activeevents.com/connect/fileDownload/session/4617511313E404EE4C4B45AF6F1DFEC2/CON1476_Gubar.ppt.pptx

21. http://en.wikipedia.org/wiki/Big_data

22. http://www.snia.org/sites/default/education/tutorials/2013/spring/big/RobPeglar_Introduction_to_Analytics_and_Big_Data_Hadoop.pdf

23. http://institute.lanl.gov/isti/irhpit/projects/

24. http://evbdn.eventbrite.com/s3s3/eventlogos/114011/mediaonmobile.jpg

25. http://www.techywood.com/wp-content/uploads/2013/12/wpidsocial-media-logos_15773.png [26] http://uwaterloo.ca/institute-nanotechnology/sites/ca.institutenanotechnology/files/resize/uploads/images/WatLAB-500x332.jpg

26. http://dsn.east.isi.edu/images/soldier_sm.jpg

27. http://www.scaledb.com/images/big_data.jpg

28. http://readwrite.com/files/_hadoopelephant_rgb1.png

29. http://blogs.mulesoft.org/wp-content/uploads/2013/05/hdfs-logosquare.jpg