International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 6, Issue 4, July- August 2017 ISSN 2278-6856

Modern Technologies of BigData Analytics: Case study on Hadoop Platform

Dharminder Yadav[1], Umesh Chandra[2]

[1]Research Scholar, Computer Science Department, Glocal University, Saharanpur, UP, India

[2]PhD, Assistant Professor, Computer Science Department, Glocal University, Saharanpur, UP, India

Abstract

Data is growing in the worldwide by daily activities, by using the hand-held devices, the Internet, and social media sites.This paper main discusses about data processing by using various tool of Hadoop.This present study cover most of the tools used in Hadoop that help in parallel processing and MapReduce. The day since BigData term introduced to database world , Hadoop act like a savior for most of the large, small organization. Researchers will definitely found a way through Hadoop to work huge data concept and most of the researchers are being done in the field of BigData analytics and data mining with the help of Hadoop.

Keywords— Big Data, Hadoop, HDFS (Hadoop Distributed File System), NOSQL

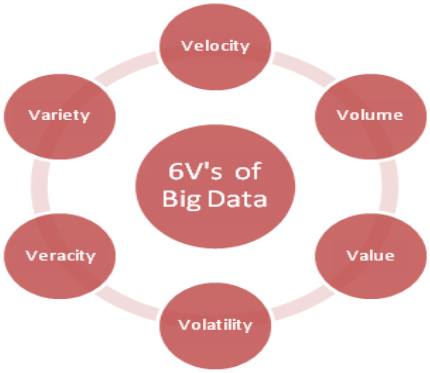
1.INTRODUCTION

Big Data provide storage and data processing facilities to Cloud computing [26]. Big data comes around 2005 but now it is used everywhere in daily life, which alludes to an expansive scope of informational collections practically difficult to manage, handle and prepare utilizing accessible regular apparatuses and information administration devices, because of their size and intricacy [10].Apache Hadoop distributed file system (HDFS) is developing as a preferable software component for cloud computing combined with integrated part such as MapReduce, Hive, NOSQL. Hadoop, MapReduce provides to the application programmer the abstraction of the map and reduce.

Fig.1. Type of Data and Data Growth

Map algorithm divides the large data sets into smaller chunks, process parallel and Reducer algorithm combine these small chunks in sequence to produce output. With the help of Hadoop very large volumes of data being generated by daily activities accessible for the organization [4, 10]. Big Data can be watched in the finance, business and where a colossal amount of data generated, stock exchange, banking, on-line and on-site procuring [2]. Enormous Information as an examination subject from a few focuses course of events,

geographic yield, disciplinary output, types of distributed papers, topical and theoretical advancement. The Big Data challenges define in 6V's that are variety, velocity, volume, value, veracity, and volatility [23].



Volume: Data is growing exponentially by daily activities which we handle. Data not only in the form of text, but also in the form of music, video, HD videos, Simple and HD image files. Data size in different organization stored in terms Terabytes, Petabytes and Zettabytes. To handle this huge database, we need to re-evaluate the architecture and application [4, 26].

Velocity: Velocity refers to the speed of Data. Data comes in unperformed speed and processed in a timely manner. The data is huge and flow of data is continues, comes from different sources like social media network, hand-held devices, stock exchange and sensors. This real-time data help organization and researchers in decision making and competitive strategies, if organization and researchers are able to process velocity of data. The velocity of data is challenging for organizations [4, 10, 26].

Value: Every data have some hidden value. The extracting valuable information from unformatted non-traditional data challenges for organizations. This information is used by organizations for business strategies.

Variety: Data can be any format, structure, unstructured, semi-structured. Unstructured data is text documents, email, video, audio, financial transaction and stock trading. To organize, merging and managing these varieties of huge data is big challenges for organizations [10].

Variability: With the increasing velocities and varieties of data, data flows can be inconsistent with periodic peaks. To manage these events triggered peak and unstructured data loads can be challenging for an organization [26].

Volatility: Volatility refers to how quickly data change, how long data is going to be valid and how much time data store.

Complexity: Data comes from multiple sources to link, match, cleanse, transform data, connect and correlation relationships, hierarchies and multiple data linkages across systems out of control [23].

2.PROBLEM BIG DATA

1) Information Security

Today devices connected to each other and other devices through Internet, volume of data collection, storage, velocity of data and variety of data bring new challenge in form of information security [1]. The current security mechanism such as firewall, DMZs are not sufficient for Big Data Architecture. So different organization, governments, CSA, journal and research paper proposed new security, privacy policies and technologies to fulfil the Big data privacy and security demands. These challenges divided in four Big data mien such as infrastructure security, Data privacy, Data management, integrity and reactive security [2, 27].

Fig.3. Challenges of Information Security [27]

These four categories are also subdivided into ten distinct security challenges. Big data security aspect also preserved data confidentiality, integrity, and availability [4].

2)Heterogeneity and Incompleteness

Heterogeneity is the big challenge in Bigdata analysis. The computer works more efficient if data identical in size and structure and complete [8].In Big data, data may be structure, semi-structure and unstructured is the big challenge in data analysis. Consider an example of the patient. A patient has the different record such as one record for Medical report/Lab test, one record for surgical operations, one record for each admission at the hospital and one record for a lifetime hospital interaction with the patient. Each and every patient has different surgical operations and lab tests records. These data design does not well structure, so managing these unstructured and incomplete data is required further works.

3)Privacy

Privacy of data related to personal information customer, how personal information of customer stored, processed and used. Privacy of data is the major problem with big data. Identification of personal information during transmission over the Internet by untrusted partner and malicious insider is the serious privacy issue. Data privacy ensure that consumer personal information is collected, stored shared and utilized in an apt way, so information is safe from other parties[26]. In some country have the law about data privacy such as in USA lusty law for health records, but another country has no such rules.

4)Scale

Big Data is collection large data sets. Managing these large data sets and brisk increasing volumes is a big problem from last ten years. In past years, this problem was solved by increasing the speed of the processor and using parallel processing but now the quantity of data is huge and processor become static [8].Now it's time for cloud computing technology, due to this data is generated with high speed. This huge volume of data sets and velocity of data becoming a challenging problem to the data analysts.

5)Timeliness

The other side of size is processing very much speed. The processing huge data sets, the longer time it will take to analyse. We will require the new design of a system that adequately deals with huge size is likely also to result in a system that can process huge size of data sets in the faster way [9].This is not meant by speed, someone says velocity in the context of Bigdata, rather there is acquisition rate as a challenge.

6)Human Collaboration

Even with the technical advancement in computing analysis, human inputs remains vital. The incredible advances made in computing analysis, their debris many patterns that human can easily understand but the computer cannot or computer algorithm take time. CAPTCHA and Wikipedia are a good example [10].Big data analysis system support input from multiple human experts, process and shared for analysis of results. These multiple experts may be separated in space and time; it is very expensive to bring together under one roof.

3.TECHNOLOGY FOR BIG DATA PROCESSING

For the processing, the huge amount of data various technologies has been introduced for manipulating, analysing and visualizing the Big data. We examine the two classes of technology, (1) Operational Big Data include the system like MongoDB, NoSQL that provide operational capabilities for real-time, interactive workload where data is primarily captured and stored. (2) Analytical

Big Data includes systems like Massively Parallel Processing (MPP) database systems and MapReduce that provide analytical capabilities for reflective and complex analysis that may touch all most of the data [14].There are more than 40 Big data platform and software are available for providing efficient analytics for large data sets (IBM Bigdata analytics, HP Bigdata, SAP Bigdata analytics, Microsoft Bigdata etc), but the Hadoop is mostly used.

HADOOP: SOLUTION FOR BIG DATA PROCESSING

Hadoop is an Apache open source framework created by Doung Cutting and Michael J. Cafarella. Hadoop is written in Java that supports the processing of large data sets in a distributed computing environment [11]. MapReduce was developed by Google, but now it is part of Apache Hadoop, MapReduce is a programming framework model where an application breaks into the various part. It is developed to scale up from a single server to thousand of the machine, each offers local computation and storage [5]. The current Apache Hadoop scheme consists of the Hadoop Kernel, MapReduce, HDFS and numbers of assorted elements like Apache Hives, Base and Zookeeper.

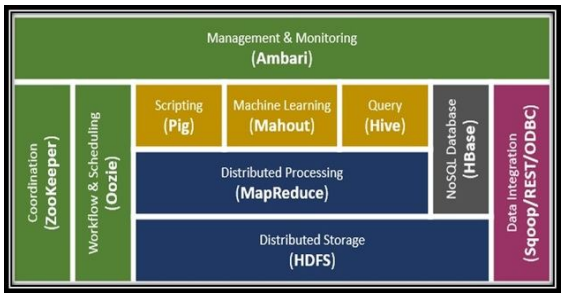


Fig.4. Ecosystem of Hadoop [3, 10]}

Hadoop consists of two major part HDFS (Hadoop Distributed file system) and MapReduce. It is unfeasible for storing a large quantity of data in existing file system, so Hadoop uses a brand new file system known as HDFS that split data into several smaller components and distribute each part redundantly across multiple nodes.

HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

HDFS is a distributed file system to run on commodity hardware. HDFS is versatile, highly fault-tolerant, clustered way to handling and designed to be deployed on low-cost hardware [11].

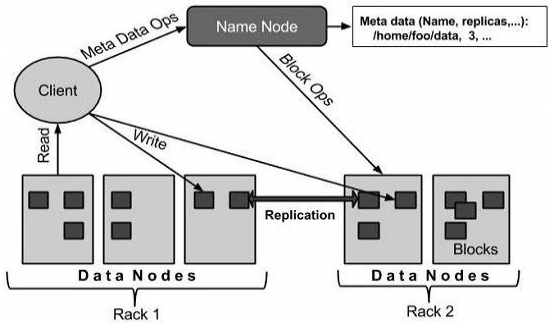


Fig.5. HDFS Architecture [8]

HDFS designed to store huge amount of data and survive the failure of storage infrastructure without losing data. It creates clusters of inexpensive machine and coordinates work among them. It works in distributed system, one of the nodes fail Hadoop continues to run the cluster without losing data or interrupting work, by shifting work to other machine working in the cluster [22].Hadoop distributed file system manage the storage huge data sets on the cluster by splitting incoming data sets into small pieces, called blocks, each block size is 64 MB. It stores the blocks redundantly across the pool of servers. These data block stored by data node, these data node store and manage by NameNode. HDFS is divided into three parts such as NameNode, DataNode, and HDFS Clients/EdgeNode.

NameNode

HDFS contain metadata in a dedicated server called NameNode. It is a single and centrally placed node in a cluster, which holds information about Hadoop file system. It is also called Namespace, which contains the hierarchy of files and directories. These are shown on NameNode by inodes, which contain records attributes like permissions, modification, and access time and disk space quotas. The file split into blocks, each block of a file is independently placed on multiple DataNodes. It contains information about free blocks which are allocated next and maintains the mapping of blocks to DataNodes[15].HDFS client wants to read a file, first, it contacts to NameNode for the locations of data blocks comprising the files and then reads content from the DataNodes closest to the client. When client writing data then client requests the NameNode to allocate a set of DataNode to host block, then client writes data to DataNodes in pipeline way.

DataNode

DataNode is slave node of NameNode and responsible for storing actual data on HDFS. Each data block has two files in host original file system, one file contains data and other contains metadata. The communication between DataNode and NameNode at start-ups [22]. Each data block have Namespace ID when it is formatted, in communication these Namespace Id and DataNode is verified with NameNode and if it does not match with NameNode then that DataNode shuts down. When a new DataNode without Namespace ID join cluster then they assign cluster Namespace ID. Each DataNode has storage ID which helps to identify it after restarting it with a different IP address or port. To identify the duplicate blocks in the cluster each DataNode sends block report to NameNode, first block report send during DataNode Registration and consecutive block reports are send at every hour. This helps the NameNode to track the duplicate nodes in the cluster. Each DataNode send heartbeat to NameNode to confirm that it is operating and its duplicate blocks are available, default heartbeat interval is 3 seconds and if no heartbeat signal is received at NameNode within ten minutes then DataNode not available mark by NameNode. Actual data store in DataNode and it is configured on the hard disk [15].

DataNode has mainly two tasks such as, block storing in an HDFS and providing the platform for the running job.

HDFS clients/Edge nodes

HDFS clients are acts as an interface between the Hadoop cluster and the outside network. It is also referred to as edge node or gateway node. Edge nodes are used to run client application and cluster administration tool [7].Each Hadoop cluster has one client and acts as a communicator between DataNode and NameNode.

MapReduce Architecture

MapReduce is a framework originally developed by Google. Now MapReduce algorithm is part of Hadoop and managed by Apache. MapReduce is a programming model for processing and generating a huge amount of structured and unstructured datasets stored in HDFS. It works on the model of split-apply-combine strategy for data analysis. MapReduce program is composing of Map and Reduce function. The map function divides the data into multiple small blocks, each block assigned to Map function that can processed datasets in parallel. Each process read input as a (key, value) pair to generate a moderate (key, value) pair as output. The Reduce function merge all the moderate values correlated with same moderate (key, value) which group than into final results produced just in one or zero output value per Reduce task [14]. The running system guard all internal details like partitioning the input data, scheduling the program's execution, handling machine failure.

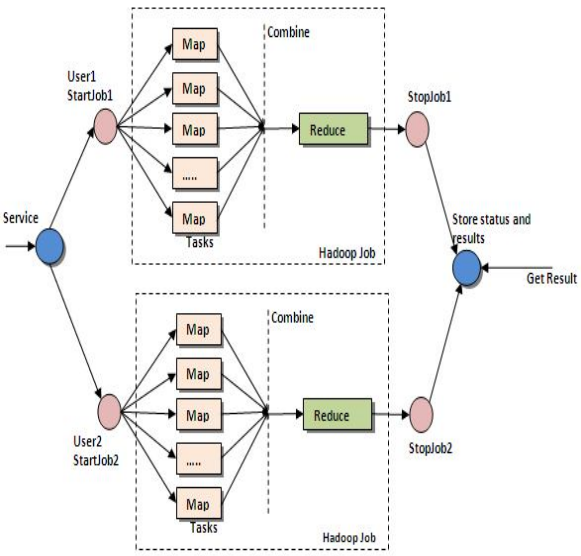


Fig.6. MapReduce Architecture [22]

MapReduce is working on all data inside our cluster in batch processing. Pig and Hives are also part of Hadoop MapReduce framework for job submitting to the cluster. HIVE is a SQL interpreter which is developed by facebook and Pig is another subproject build by yahoo which is used the pull out data from the cluster.

SOME ESSENTIAL PART OF HADOOP PROJECT

YARN (Yet Another Resource Negotiator)

It is a part of Apache Hadoop project. YARN (Yet Another Resource Negotiator) is a cluster management technology and its latest version is 2.5.0 which support, scheduler pre-emption, Security for Timeline Server and act as a resource manager application. It is one of the key features in the second-generation Hadoop version of the Apache Software Foundation's open source distributed processing framework [2, 3]. The next generation of Hadoop computes platform known as YARN, which departs from its familiar, monolithic architecture. Programming frameworks running on YARN coordinate intra-application communication, execution flow, and dynamic optimizations as they see fit, unlocking dramatic performance improvements .Apache Yarn Framework divides into two parts master and slave, the master is known as Resource Manager, a slave called node manager and application manager. Each slave node has one node manager and each application has one Application master. Resource Manager manages the overall assignments of benefits (CPU and memory) among each one of the applications. It has two noteworthy parts, Scheduler and Application manager.

Amberi

Amberi is a top level sub part of Apache Hadoop project. Amberi support selected 64-bit operating systems such as RHEL (Red hat Enterprise Linux), CentOS 6 and 7, OEL (Oracle Enterprise Linux) 6 and 7, SLES (SuSE Linux Enterprise Server) 11, Ubuntu 12 and 14, and Debian 7. Its latest version is Amberi 2.5.0.It is a web-based gizmo for managing and tracking Apache Hadoop clusters. It includes MapReduce, Pig, Hives, HDFS, Hbase, Zookeeper and Sqoop. It is very simple and interactive user interface to install various tools, perform sundry management and monitoring tasks. It is responsible for providing dashboard for viewing cluster health. It has two components such as Amberi server and Amberi Agent. Amberi server is a master process to communicate with Amberi agents [10, 22].It is responsible for maintaining all cluster metadata. Amberi agent installed on each node and sends periodically own health status according to master.

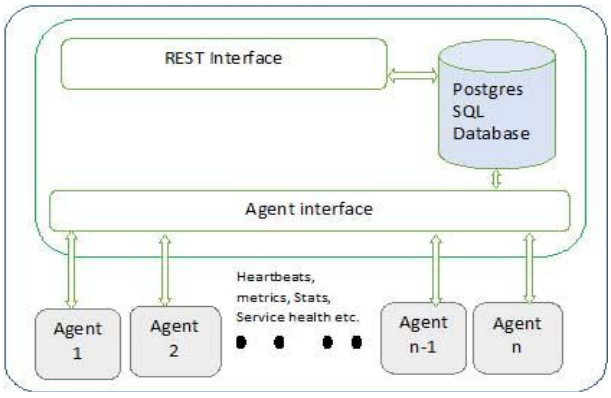


Fig.7. Amberi Architecture [15]

Pig

Apache Pig is a tool for analysing large data sets for data flows and performing all data manipulation operation in Hadoop. The latest version of the Pig is 0.16 and released on dates 08-06-2016.The pig has high-level programming language just like SQL was known as Pig Latin and it is an alternative to Java, includes with Azure HDInsight. It supports for parallel processing that enables to handle large data sets. Pig provide the facilities to define the own function, easy programming environment and permit the system their execution automatically.

Hive

Hive initially developed by Facebook but now it is part of Apache Hadoop project. It is Fast, scalable and provides the SQL type querying interface called HiveQL which supports select, project, join, aggregate, union, and sub-queries in the from clause. HivesQL is supports data definition language (DDL), data manipulation language (DML).It is data warehouse software to store schema in the database, processing data in HDFS. It is developed to process OLAP operation [26].Components of Hives are metastore, compiler, deriver, optimizer, and executor. Metastore is software run on an RDBMS and uses a Data Nucleus to convert object schema into a relational schema and vice versa. It stores all information about tables such as table location, schema, type of columns and partition. The compiler uses metadata to generate execution plans. It has three stages such as parse, Type checking and Semantic Analysis and Optimization. The driver that manages the lifecycle of a HiveQL and maintains a session handle. Execution that executes the tasks produced by the compiler and interacts with Hadoop instance.

Mahout

Mahout is a scalable machine level algorithm target on collaborative filtering, clustering, and classification on top of Apache Hadoop. It provides Java library for mathematical operation. Its first released version is 0.10.0, released on 11 April 2015 and the latest version is 0.13.0, released on 17 April 2017[3]. Mahout utilized the concept of artificial intelligence to improve the performance. Mahout 0.13.0 includes Vienna CL3, Java CPP, GPU, OpenMP, and Mahout JVM. It supports GPU, OpenMP Matrix-Matrix and Matrix-Vector Multiplication on spark. Mahout uses k-Means, fuzzy k-Means, Canopy, Dirichlet, and Mean- Shift for clustering in MapReduce.

Avro

Avro is a remote procedure call, serialization framework and provides data exchange services for Hadoop. It is developed by Doug cutting. Avro uses JSON format to exchange between programs written in any language [11, 17].It stores data definition and data together in one message and store in binary format for making compact and efficient. It provides both serialization format and a wire format for communication between Hadoop nodes. Avro have API that support languages like Java, Python, Ruby, C, C++ and more. It supports schema evaluation procedure to handle schema changes. JASON specifies the remote procedure call for Avro.

Spark

Spark is an open source data processing framework, developed in 2009 at UC Berkeley's AMP Lab by Matei Zaharia. It was open source in 2010 and in 2013 it becomes part of Apache foundation. Now it is Apache spark on the top in Hadoop architecture. It is fast, in-memory data processing computing in the cluster, which increases the data processing of the application[3,11].Spark running on Hadoop YARN, which provides the facilities to developer create application anywhere and provide the platform to spark and other application to share cluster and data sets. Apache spark has two components such as spark core and the set of libraries. It speeds up the iterative data processing. Start Core is the establishment of the general venture. It gives circulated undertaking dispatching, scheduling, and essential I/O functionalities, uncovered through partner application programming interface focused on the RDD abstraction. RDDs will contain any style of Python, Java, or Scala objects. Spark SQL is an important part on top of Spark Core. It introduced the concept of data abstraction known as data Frames. Spark SQL manipulates DataFrames in Scala, Java, or Python through DSL (Domain-specific language) and dataFrame support for structured and unstructured data. It also supports SQL language and ODBC / JDBC Server. Spark Streaming provide fast scheduling capability and GraphX is a graph processing framework.

Hbase

HBase is an open source software and part of Hadoop. HBase is a non relational database that maintains read/write access for huge data sets and run on top of HDFS. It is fast, fault tolerant and useable[12,18].It creates a huge table for storing multi-structure data and table acts as an input and output for MapReduce job in Hadoop access through Java API. Phoenix provides an SQL layer for Hbase to access classical SQL database. Hbase, provide a messaging platform to Facebook and other organization also uses Hbase are yahoo!, Adobe, Netflix, flurry, Rocket Fuel etc. The Apache Trafodion is a component of Hbase that uses it as a Storing engine and provides a SQL query engine with ODBC and JDBC drivers. Hbase is coming after Google BigTable and written in Java however not support SQL scripting.

Zookeeper

Zookeeper is an open source contour, distributed and provides synchronization services across the cluster. It is the centralized system where distributed system import and export data from it. It stores data in znodes instead of files or directories. Znode is two type, Persistent and Ephemeral znodes. Default znode in Zookeeper is Persistent znode, which contain configuration details and a new node is always added in persistent znode. Ephemeral is a session node which created when an application is started and deleted when application finished. Zookeeper acts as an

admin tool for managing, coordinating huge cluster of machine and jobs [15].It follows the client-server model, where client nodes use services and server nodes provide services. Zookeeper process all requests in FIFO order that received from the client. Its application is fetching services, Katta and Yahoo! Message Broker. The fetching Service is a vital part of search engine and Yahoo! Crawlers. Katta provides the coordination services for non-Yahoo! Application. Yahoo! Message Broker is distributed system and client can send and received the message from [25].Zookeeper conjointly provides the distributed lock services.

Cassandra

Cassandra is developed by Data Stax Enterprise and integrated with Hadoop and MapReduce [9]. Cassandra 0.6 provides support to Hadoop and MapReduce to retrieve data from Cassandra and Cassandra 0.7 provide the mechanism to export data from Cassandra. Apache Cassandra could be a NoSQL database ideal for high-speed, on-line transactional data, whereas Hadoop could be a huge data analytics system that focuses on knowledge warehousing and knowledge lake use cases.

Hcatalog

Hcatalog is an open source metadata and licensed under the Apache license [13]. It's provided the table management framework that works with Hadoop HDFS data. Its initial stable version is 0.4 and released in May 2012[20]. HCatalog makes data within the Apache Hadoop framework available to users within and outdoors the system. It provides interoperability through table abstraction and present common table layer to identical data model.

Sqoop

Sqoop is an element of the Apache project since may 2012, latest stable version is 1.4.6. Sqoop could be a tool that to transfer huge data between Hadoop HDFS and relational database (RDBMS) like Oracle, MySql, Teradata, Postgres [21]. Sqoop perform import and export operation with the assistance of MapReduce Framework, import operation performs parallel row by row, export operation is serialized and fault tolerance. Sqoop conjointly import data to other Hadoop components like HBase, Hive through HBase client and Hives client. Most of the processes of Sqoop are automated. Sqoop uses JDBC Connector to Connect with Relational Database and it provides a command line interface to end user.

Oozie

Apache Oozie the latest version is 4.2.0 run underneath the Apache License 2.0. It is an open source Java web application job scheduler system to run and manage Hadoop jobs Apache Oozie is strongly connected with Hadoop component like Hive, Pig, Sqoop and also connected with system jobs Java, Shell[14]. In Oozies Callback and Polling are two operations to detect completion of job. There are three type of jobs in Oozie are Workflow jobs, Coordinator jobs, and Oozie Bundle. Workflow jobs use Directed Acyclic Graph to define the sequence of the action plan and running on demand. Coordinator jobs running periodically depending upon time and data available. Oozie Bundle jobs are the combination of coordinator jobs and workflow jobs manage as a single job[19]. Hue and Oozie Eclipse plugin is the editor for Oozie. Hue uses drag and drops action to specify jobs, but Oozie Eclipse plugin uses the graphical method to specify the jobs.

Apache Drill

Apache drills latest version is Drill 1.10 released on March 2017.It is an open source software framework which has a SQL query engine to process huge data sets and records in seconds [10, 15]. Drill is part of the Google Dremel system which also called Google BigQuery. Drill version 1.10 support, create temporary table as (CTTAS), JDBC connection, web console, INT96 timestamps and Kerberos authentication. Apache Drill support relational, various non relational database and file systems such as HDFS, Map R-DB, Hbase, MongoDB, Swift, and Google Cloud Storage. Drillbit consists of SQL parser, RPC end point, Storage Plugin interface, storage plugin (HBase, Map R-DB, MongoDB, and Cassandra) and third party plugin. Zookeeper help the Drillbit find the other Drillbit in the cluster and submit client query.

3.CONCLUSIONS

This paper concludes that all the tools used in the Hadoop have some outstanding role that help in doing number of parallel processing through clusters and will definitely reduce our processing work while dealing with big data. Today big data is not only a term but it's a complete research area field that can be researched and analysed with technology like Hadoop. Working with Hadoop will definitely contribute a lot to Big Data area for pattern finding and business intelligence.

REFERENCES

[1] Advantech, Enhancing Big Data Security.
[2] Agrawal, D., Das, S., and El Abbadi, A.,Big data and cloud computing. In Proceedings of the 14th International Conference on Extending Database Technology, New York, New York, USA, 2011.
[3] Apache Hadoop, http://Hadoop.apache.org.
[4] E. R. Osawaru and A.H. Riyaz, A Highlight of Security Challenges in Big Data, International Journal of Information Systems and Engineering, Volume 2, Issue 1, 2014.
[5] E. S. A. Ahmed and R. A.Saeed, A Survey of Big Data Cloud Computing Security, International Journal of Computer Science and Software Engineering, vol. 03, Issue 01, pp. 78-85, 2014.
[6] G.kaur and M. Kaur, Review Paper on Big Data using Hadoop, Int. J. of Comp. Engg. and Tech., vol. 6, Issue 12, 2015, pp. 65-71.
[7] Hadoop Distributed File System, http://Hadoop.apache .org/hdfs.

[8] H.S. Bhosale and D.P. Gadekar, A Review Paper on Big Data and Hadoop, Int. J. of Sci. and Research Pub., vol. 4, Issue 10, 2014.
[9] https://cassandra.apache.org/.
[10] https://en.wikipedia.org/wiki/Big_data.
[11] https://en.wikipedia.org/wiki/Apache_Hadoop.
[12] https://en.wikipedia.org/wiki/Apache_Hbase.
[13] https://en.wikipedia.org/wiki/Apache_Hcatalog.
[14] https://en.wikipedia.org/wiki/Apache_Oozie.
[15] https://hortonworks.com.
[16] https://hortonworks.com/apache/amberi.
[17] https://hortonworks.com/apache/avro.
[18] https://hortonworks.com/apache/hbase.
[19] https://hortonworks.com/apache/Oozie.
[20] https://hortonworks.com/apache/hcatalog.
[21] http://sqoop.apache.org/
[22] http:// www.tutorialpoint.com/Hadoop.
[23] K. Zvarevashe, M. Mutandavari, T. Gotora, A Survey of the Security Use Cases in Big Data, Inter. J. of Innov. Res. in Comp. Comm. and Engg., vol. 02, Issue 05, 2014.
[24] Oracle: Big data For enterprise, An Oracle White Paper, 2012.
[25] P. Hunt, M. Konar, F. P. Junqueira and B. Reed, Zookeeper: Wait-freecoordinationforInternet-scalesystems, Yahoo! Grid and Yahoo! Research.
[26] V. B. Bobade, Survey Paper on Big Data and Hadoop, International Research Journal of Engineering and Technology, vol. 03,Issue 01, 2016, pp 861-863.
[27] V.N. Inukollu, S. Arsi, and S.R. Ravuri, Security issue Associated with Big data in Cloud Computing, Int. J. of Network Security and its Application, vol. 6, Issue 3, 2014.
[28] W.K. Chen, Linear Networks and Systems, 1993, pp. 123-135.

AUTHOR

Dharminder yadav received the B.Sc. degree in computer science from GNK Collage, M.C.A degree from JMIT Radour and PhD. (P) from Glocal University in 2007 and 2010, respectively.