Journal of Analysis and Computation (JAC)
(An International Peer Reviewed Journal), www.ijaconline.com, ISSN 0973-2861
National Conference on: Research Challenges in Information Systems of Computer Science (NCRCISCS)"

A STUDY ON BIG DATA HADOOP AND ITS DATABASE TOOLS

R. NANDHAKUMAR[1] AND ANTONY SELVADOSS THANAMANI[2]

[1]Acadamitian -*Department of Computer Science ,NGM College,Pollachi-642001,India*

[2]Associate Prof & Head, *Department of Computer Science,NGM College,Pollachi-642001,India*

E-Mail-nkumarram@gmail.com

ABSTRACT-

In the information era, wide variety of data have become available on hand to organizational decision makers. Big data refers to datasets that are not only big, but also high in variety, volume and velocity, which makes them difficult to handle using traditional data base tools and techniques. Due to the rapid growth of such data, solutions need to be analyzed and presented in order to handle and extract information and knowledge from these datasets. Furthermore, decision makers of organization need to be able to gain valuable imminent from such varied and rapidly changing data, ranging from daily transactional process to customer interactions and social media network data. Such data can be provided using big data analytics, which is the application of advanced analytics techniques on big data. This paper aims to analyze and compare some of the different database tools which can be applied to big data, as well as the occasions provided by the application of big data analytics in various organizational domains.

***Keywords*—**Big data analytics, Big data Database, Big data database tools.

I. INTRODUCTION

The term "Big Data" was first introduced to the computing world by Roger Magoulas from O'Reilly media in 2005 in order to define ahuge amount of data that traditional database management techniques cannot manage, access and process due to the scalability, flexibility and size of the data. A study on the Evolution of Big Data as a Research and Scientific Topic shows that the term "Big Data" was present in research starting with 1970s but has been comprised in publications in 2008. Nowadays the Big Data concept is treated from different points of view covering its implications in many fields. According to MiKE 2.0, the open source standard for Information Management system, Big Data is defined based on the size which consists of large, complex and independent collection of data sets, each with the potential for interaction of data. In addition to this, an important aspect of Big Data is the fact that it can't be handled with standard data base management techniques due to the versatility,inconsistency and unpredictability of the possible combinations.

Big Data has four aspects as shown in fig 1:

Volume: refers to the quantity of data captured by the company. This data must be used further to obtain important knowledge about the organization.
Velocity: refers to the time in which Big Data can be processed and produce the results. **Variety:** refers to the type of data that Big Data can integrate. This data can be structured as well as unstructured;
Veracity: refers to the degree in which a leader trusts the used information in order to take decision. So getting the right correspondence in Big Data is very important for the business in future performance.

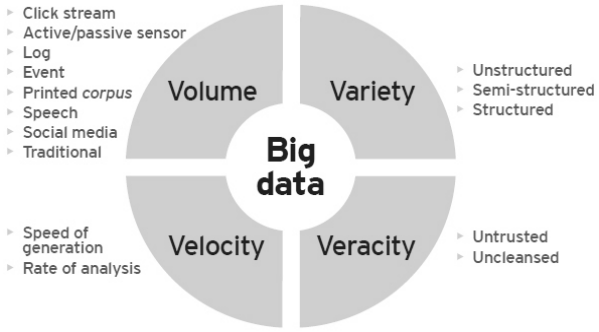

Fig 1. V's of Big Data

The amount of data stored in various sectors can vary in the data stored and how they created, i.e., images, audio, text information etc., from one organization to another. From the practical point of view, the graphical interface

used in the big data analytics tools leads to be more efficient, faster and better decisions and performance which are massively preferred byanalysts, business users and researchers [1].

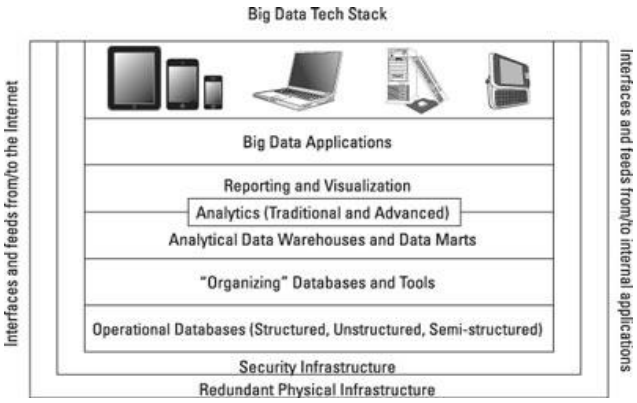

Fig 2. Big Data Architecture

Here's a closer look at what's in the image and the relationship between the components:

- **Interfaces and feeds:** On either side of the diagram are specification of interfaces and feeds into and out of both internally managed data and data feeds from external sources. To understand how big data works in the real world, start by understanding this necessity of the data.
- **Redundant physical infrastructure:** The supporting physical infrastructure is fundamental necessity for the operation and scalability of big data architecture. Without the availability of robust physical infrastructures, big data would not have emerged as an important trend.
- **Security infrastructure:** The more important big data analysis becomes to companies, the more important it should be need to secure the data. For example, in a healthcare company,we will probably want to use big data applications to determine changes in demographics or shifts in patient needs and treatments.
- **Operational data sources:** When we think about big data, understand that we have to incorporate all the data sources that will give us a complete picture of business and see how the data impacts the way we operate the business.

II.BIG DATA DATABASE TOOLS

A. Hadoop

The name Hadoop has become synonymous with big data. It's an open-source software framework for distributed storage of very large datasets on computer clustering networks. All that means that we can scale our data up and down without having to worry about hardware and network failures. Hadoop provides massive amounts of storage for any kind of data, massive processing power and the ability to handle practically limitless simultaneous tasks or jobs.Hadoop is not for the beginner. To truly strip up its power, we really need to knowbasics of Java. It might be anassurance, but Hadoop is certainly worth the effort – since many other companies and technologies run off of it or integrate with it. Hadoop involves a cluster of storage/computing nodes (or machines) out of which one node is assigned as master and other as slave nodes. The HDFS [18] maintains each file in the chunk of same size blocks or nodes (except the last block). Also, various replications of these blocks are maintained on various nodes in the cluster for the sake of reliability and fault tolerance. The Map-Reduce function computing technique divides the whole task of processing into smaller blocks and assign it to various slave machines which are the required data is available and executes computing right at that node. In this way it saves significant time and cost involved in transferring data from data server to the computing machine. Following are the advantages, disadvantages and latest version of Hadoop.

i. Advantages of Hadoop

- Open source: Being an open source, Hadoop is freely available online [3].
- Cost Effective: saves cost as it utilizes cheaper, lower end cluster of commodity of machines instead of costlier high end server. Also, distributed storage of data and transfer of computing code rather than data saves high transfer costs for large data sets [3].
- Scalable: To handle larger data, and to maintain performance and is capable to scale linearly by putting additional nodes in clusters [3].
- Fault Tolerant and Robust: It replicates data block on multiple nodes that facilitates the recuperation from a single node or machine failure. Also, Hadoop's architecture deals with frequent malfunctions in hardware. If a node fails the task of that node is reassigned to some other node in the cluster [4].
- High Throughput: Due to batch processing high throughput is achieved[4].
- Portability: Hadoop architecture can be effectively ported [5] while working with several commodities of operating systems and hardwares that may be assorted [6].

ii. Disadvantages of Hadoop

- Single Point Failure: Hadoop's (version up to 2.x) HDFS as well as MapReducefunction suffer from single points of failure [7].
- Low Efficiency/ Poor Performance than DBMS [7]: Hadoop shows lower efficiency due its inability to switch to the next stage before completing the previous stage tasks which causes Hadoop unsuitable for pipeline parallel processing, runtime scheduling
  that causes degraded efficiency per node. Unlike RDBMS, it has no specific optimization of execution plans that could minimize the transfer of data among various nodes.
- Inefficient Dealing with Small Files: As HDFS is meant for high throughput optimization [8], it does not suit to random reads on small files [9].
- Not Suitable for Real Time Access: MapReduce and HDFS employ batch processing architecture and it does not fit for real-time accesses [8].

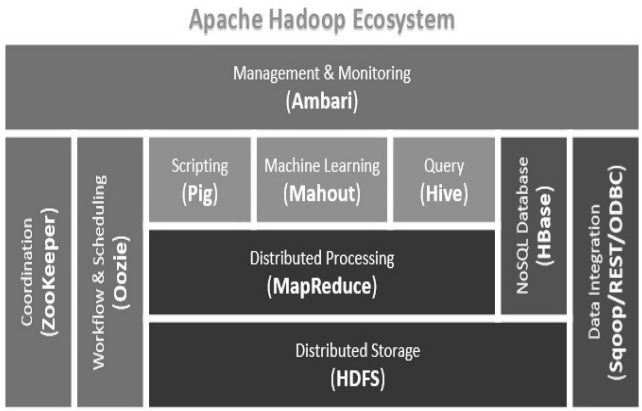

Fig 3. Hadoop Architecture

B. Cassandra

Cassandra is an Apache free and open-sourceand distributeddatabase management systemdesigned to handle large amount of data across many commodity hardware and serversby providing high availability of performance with no single point of failure. It offers robust support for clusters spanning multiple datacenters,[1] with asynchronous master less replication allowing low latency operations for all clients.

Cassandra also places a high value on performance. In 2012, University of Toronto researchers studying NoSQL systems concluded that "In terms of scalability, there is a clear winner throughout our experiments. Cassandra achieves the highest throughput for the maximum number of nodes in all experiments" although "this comes at the price of high write and read latencies.

Main features

- Decentralized
  Every node in the cluster has the same role. There is no single point of failure. Data is distributed across the cluster (so that each node contains different data), but there is no master as every node can service any request as master node. Supports replication and multi data center replication strategies are configurable[17]. Cassandra is designed as a distributed system, for deployment of large numbers of nodes across various
- Scalability
  Read and write throughput both increase linearly as new machines are added, with no downtime or interruption to applications.
- Fault-tolerant
  Data is automatically replicated to various nodes for fault-tolerance. Replication across multiple data centers is supported. Failed nodes can be replaced with no latency time.
- Tunable consistency
  Writes and reads offer a tunable level of consistency, all the way from "writes never fail" to "block for all replicas to be readable", with the quorum level in the middle.[10]
- MapReduce support
  Cassandra has Hadoop integration, with MapReduce support. There is support also for Apache Pig and Apache Hive.

C. HBase

HBase is a data model that is similar to Google's big table designed to provide quick random access to huge amounts of structured data. This tutorial provides an introduction to HBase, the procedures to set up HBase on Hadoop File Systems, and ways to interact with HBase shell. It also describes how to connect to HBase using java, and how to perform basic operations on HBase using java.

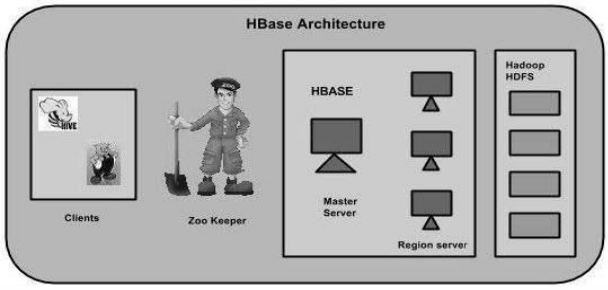

Fig 4. HBase Architecture

HBase is a distributed column-oriented database built on top of the Hadoop file system. It is an open-source project and is horizontally scalable. Base is a data model that is similar to Google's big table designed to provide quick random access to huge amounts of structured data. It leverages the fault tolerance provided by the Hadoop File System (HDFS).It is a part of the Hadoop ecosystem that provides random real-time read/write access to data in the Hadoop File System.One can store the data in HDFS either directly or through HBase. Data consumer reads/accesses the data in HDFS randomly using HBase. HBase sits on top of the Hadoop File System and provides read and write access.

HBase is a **column-oriented database** and the tables in it are sorted by row. The table schema defines only column families, which are the key value pairs. A table have multiple column families and each column family can have any number of columns. Subsequent column values are stored contiguously on the disk. Each cell value of the table has a timestamp. In short, in an HBase:

- Table is a collection of rows.
- Row is a collection of column families.
- Column family is a collection of columns.
- Column is a collection of key value pairs.

Features of HBase

- HBase is linearly scalable.
- It has automatic failure support.
- It provides consistent read and writes.
- It integrates with Hadoop, both as a source and a destination.
- It has easy java API for client.
- It provides data replication across clusters.

Applications of HBase

- It is used whenever there is a need to write heavy applications.
- HBase is used whenever we need to provide fast random access to available data.
- Companies such as Facebook, Twitter, Yahoo, and Adobe use HBase internally.
- It hosts very large tables on top of clusters of commodity hardware.

TABLE I: COMPARISON OF CASSANDRA AND HBASE

	Cassandra	HBASE
Data Model	Columnar Database	Columnar Database
Interface	HTTP/REST	HTTP/REST
Object Storage	Database contains data in columns(key-value pair)	Database contains data in columns(key-value pair
Query Method	Map/Reduce+CQL	Map/Reduce + Drill
Replication	peer – to –peer with Multiple data centers	Cluster Replication
Concurrency	Atomicity, Isolation	MVCC
Written in	Java	Java

D. CouchDB

Apache CouchDB is open source database software that spotlight on easiness of use and having an architecture that completely hold close the Web.[11] It has a document-orientedNoSQL database architecture and is implemented in the concurrency-oriented Erlang language; it uses JSON to store data, JavaScript is a query language used withMapReduce, and HTTP for an API.[11]It wasfirst released in 2005 and later became an Apache Software Foundation project in 2008. Unlike a relational database, this database does not store data and relationships in a table format. Instead of this, each database is a collection of independent document data. Each document preserves its own data and self-contained structure definition. An application may access and manipulate multiple databases, such as one stored on a user's mobile phone and another on a server. Document metadata contains updated information, making it possible to merge any differences that may have occurred while the databases were disconnected. It implements a form of multi-version concurrency control (MVCC) so it does not lock the database file during writes. Disagreements are left to the application to resolve. Resolving a conflict usually involves first merging data into one of the documents, then deleting the old one [11].

Main features

- ACID Semantics

CouchDB provides ACID semantics.[11] It does this by implementing a form of Multi-Version Concurrency Control, meaning that it can handle a large volume of concurrent readers and writers without conflict.

- Built for Offline

  CouchDB can replicate to devices (like smartphones) that can go offline and handle data sync for us when the device is back to online.
- Distributed Architecture with Replication

  CouchDB was designed with bi-directional replication (or ynchronization) and off-line operation in mind. This means multiple replicas can have their own copies of the same data, modify it, and then sync those changes later.
- Document Storage

  CouchDB stores data as documents, key/value pairs expressed as JSON. Field/value paircan be simple things like characters, numbers, or dates; but should be ordered lists and associative arrays can also be used. Every document in a CouchDB database has a unique id and there is no required document schema definition.
- Eventual Consistency

  CouchDB guarantees eventual consistency to be able to provide both availability and partitional tolerance.
- Map/Reduce Views and Indexes

  The stored data is structured using views. In CouchDB, each view is constructed by a JavaScript function that acts as the Map half of a map/reduce functional operation.
- HTTP API

  All items should have a unique URI that gets exposed via HTTP. It uses the
  HTTP methods POST, GET, PUT and DELETE for the four basic Create, Read, Update, Delete operations on all resources.

E. MongoDB

MongoDB is a free and open-sourcecross-platformdocument-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with schemas. This data base is developed by MongoDB Inc. and is free and open-source, published in combination with the GNU Affero General Public License and the Apache License.Any relational database has a typical schema design that shows number of tables and the relationship between these tables. While in MongoDB, there is no concept of relationship.

Main features

- Ad hoc queries

  MongoDB supports field, join and range queries, regular expression searches [12]. Queries can return specific fields of documents and also include user-defined JavaScript functions. Queries can also be configured to return a random sample of results of a given size.
- Indexing

  Fields in a MongoDB document can be indexed with primary and secondary indices.
- Replication

  MongoDB provides high availability with replica sets [12].A replica set consists of two or more copies of the data and each replica set member may act in the role of primary or secondary replica at any time.

  All writes and reads are done on the primary replica by default. Secondary replicas maintain a copy of the data of the primary using built-in replication. When a primary replica fails, the replica set conducts an election process to determine which secondary should become the primary. Secondary's can optionally serve read operations, but that data is only eventually consistent by default.
- Load balancing

  MongoDB scales horizontally using slicing [12]. The user chooses a slice key, which determines how the data in a collection will be distributed. The data is split into ranges and distributed across multiple networks.Alternatively, the shard key can be hashed to map to a shard – enabling an even data distribution and it can also run over multiple commodity servers, balancing the load or duplicating data to keep the system up and running in case of hardware or network failure.

Advantages of MongoDB over RDBMS

- Structure of a single object is clear
- **Schema less** − MongoDB is a document database in which one collection holds different documents. Number of fields, content and size of the document can differ from one document to another.
- MongoDB is a document database in which one collection holds different documents. Number of fields, content and size of the document can differ from one document to another
- No complex joins
- Deep query-ability. MongoDB supports dynamic queries on documents using a document-based query language that's nearly as powerful as SQL
- Tuning
- Ease of scale-out − MongoDB is easy to scale

Name of Big data tools	Mode of Software	Types of Data	Language Used	Operating System

- Conversion/mapping of application objects to database objects not needed [12]

TABLE II: COMPARISON OF COUCHDB AND MONGODB

	Couchdb	Mongodb
Data Model	Document-Oriented (JSON)	Document-Oriented (BSON)
Interface	HTTP/REST	Custom protocol over TCP/IP
Object Storage	Database contains Documents	Database contains Collections Collections contains Documents
Query Method	Map/Reduce (javascript + others) creating Views + Range queries	Map/Reduce (javascript) creating Collections + Object-Based query language
Replication	Master-Master with custom conflict resolution functions	Master-Slave
Concurrency	MVCC (Multi Version Concurrency Control)	Update in-place
Written In	Erlang	C++

III. SUMMARIZATION OF BIG DATA TOOLS

The following Table III demonstrates the comparative aspects of the diverse tools and its uses in big data based on data sources and its operating system. It tells about mode of software, types of data, language used and its operating system. The main objective of this comparison of database tools is not to look which is the best tool in big data, but to demonstrate its usage, flexibility, scalabilityand performance to create alertness of big data in various fields.

TABLE III: COMPARISON OF BIG DATA TOOLS

HBase				
Cassandra	Commercial and Open Source	Structured and Unstructured data	SQL	Windows XP, Vista, 7 and 8
CouchDB	Commercial and Open Source	Structured, Semi-Structured and Unstructured data	JavaScript, PHP, Erlang	Windows, Ubuntu
MongoDB	Open Source	Structured, Semi-Structured and Unstructured data	C++	Amazon Linux, Windows Server 2012 & 2012 R2, Debian 7.1

IV. CONCLUSION

In this paper, more than a few big data tools were elucidated along with their features of several tasks. Big data provide vastly effective supporting processes for collection of data sets which is too complex and large in size. This mandatory requirement gives the way for developing many tools in big data research. Whereas these data base tools are generated both in real time and in non-real time and also in very large scale which comes from sensors, web, networks, audio/video, etc. Thus the aim of this survey is to enhance the knowledge in big data tools and their applications applied in various companies. It also provides obliging services for readers, researches, business users and analysts to make enhanced and quicker decisions using data which will promote for development and innovation in the future.

REFERENCES

[1]http://www.slideshare.net/HarshMishra3/harsh-big-data-seminar-report

[2]http://www.infoworld.com/d/business- intelligence/7-top-tools-tamingbig- data-191131.

[3] J. Venner, ―Pro Hadoop‖, a press, (2016).

[4] T. White,‖ Hadoop: The Definitive Guide‖, third ed., O'Reilly Media, Yahoo Press, (2017).

[5] W. Tantisiriroj, S. Patil and G Gibson, ―Data-intensive File Systems for Internet Services‖, A Rose by Any OtherName (CMU-PDL-08-114). Research Centers and Institutes at Research

[6] M. K. McKusick and S. Quinlan, ―GFS: Evolution on Fast-forward‖, ACM Queue, New York, vol. 7, no. 7, (2013).

[7] K. Shvachko, H. Kuang, S. Radia and R Chansler, ―The Hadoop Distributed File System‖, Proceedings of IEEE Conference,978-1-4244-7153-9/10, (2015).

[8] J. Dean and S. Ghemawat, ―Mapreduce: Simplified data processing on large clusters, communCM, vol. 51, no. 1, (2007). 107–113.

[9] J. Dean and S. Ghemawat, ―Mapreduce: Aflexible data processing tool‖, commun. ACM, vol. 53, no. 1, **(2016)**, pp. 72–77.

[10] Hewitt, Eben (December 15, 2015). Cassandra: The Definitive Guide(1st ed.).

[11] Brown, MC (October 31, 2016), Getting Started with CouchDB(1st ed.),O'Reilly Media.

[12]Pirtle, Mitch (March 3, 2017), MongoDB for Web Development (1st ed.), Addison-Wesley Professional.

[13] Holt, Bradley (April 11, 2017), Scaling CouchDB (1st ed.), O'Reilly Media.

[14]J. Cohen, B. Dolan, M. Dunlap, J.M. Hellerstein, C. Welton, MAD skills: new analysis practices for big data, Proceedings of the VLDB Endow 2 (2) (2018).