International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

# Big Data Analytics

**Leeza Sharma[1], Ankita Gupta[2]**

[1, 2]Computer Science and Engineering (Information Security), PEC University of Technology, Chandigarh, India

**Abstract:** ***With advancement of technologies and services, data with high velocity, variety and volume is produced which cannot be handled by traditional architectures, algorithms or databases. So, there is a need of new architecture that finds the hidden threads and trends from different structured or unstructured sources and that technique is called BIG DATA. This Review paper presents different Methodologies to implement Big Data that is HPCC (older one) and HADOOP. The whole process of deployment can be divided into 5 phases- Data Distillation, Model Deployment, Validation and Deployment, Real time Scoring, Model Refresh. Along with that it concentrates on comparing HPCC , HADOOP and their components also.***

**Keywords:** Big data, Hadoop, HPCC, Map reduce, HDFS

## 1. Introduction

Huge information is created at all times by each and everything which is around us.It can be transmitted by Sensors, Systems and Mobile devices.Big information come in from numerous sources at various Velocity, Volume and Variety. Optimal Processing power and techniques are required to extract meaningful value from Big data. Huge Data can be created on enormous scale by large number of online communications among individuals and exchanges amongst individuals and systems.

Investigation of substantial information sets to reveal concealed examples, client inclinations, showcase patterns, obscure relationships, and other helpful business related data is the procedure of Big information examination. Aptitudes required for Big Data is Management and handling of dispersed information and new apparatuses for information investigation and representation of unstructured information. Today's intricate world once in a while require various specialists from various fields to comprehend what precisely going on. Contribution from numerous human specialists and shared examination of results must be bolstered by a Big Data investigation framework. These distinctive specialists might be isolated in space and time when it is too expensive to unite a whole group in one room. The information framework needs to take this circulated master info, and bolster their affiliation.

“Real-time big data isn’t just a process for storing petabytes or exabytes of data in data warehouse,”says Michael Minelli,co-author of Big Data,Big Analytics.“It’s about the ability to make better decisions and take meaningful actions at the right time. It’s about detecting fraud while someone is swiping a credit card,or triggering an offer while shopper is standing on a checkout line, or placing an ad on a website while someone is reading a specific article.It’s about combining and analyzing data so you can take the right action, at the right time and at the right place.”

Lexis Nexis invented HPCC in 1999 to solve different graph problems. The business model is based upon consuming large volumes of structured and unstructured data and converting it into a massive social graph of people and businesses. The architecture was based on the data flow paradigm that supports three types of parallelism:Data Parallelism, Pipeline Parallelism, System Parallelism.

Hadoop is a Programming platform started in 2006, used to process huge amount of data sets in a distributed computing environment. Hadoop was developed by Google’s MapReduce.The Hadoop platform incorporates different parts like HDFS and MapReduce which each play out an exceptional errand .

## 2. Deployment Phases

**Process of deployment of big data can be divided into 5 phases[3] :**

### 2.1 Data Distillation

The data distillation phase includes obtaining features for disorganized text, refining for areas of interest, selecting similar characters or components and results for modeling, and exporting sets of distilled data to a local data mart.

### 2.2 Model Deployment

Processes in this phase include feature chosing, sampling and accumulation, different transformation; model appraisal; model ractification and model benchmarking. The goal at this phase is creating a prognostic model that is strengthy, robust, intelligible and executable. The key requirements for data scientists at this phase are rapidness, resilience, productivity, and reproducibility.

### 2.3 Validation and Deployment

The goal at this phase is validating the model to confirm that in the real world,it works well. The validation process involves obtaining fresh data, executing it against the model, and comparing results with outcomes run on data that’sbeen withheld as a validation set. If the model works, it can be deployed into a production.

### 2.4 Real time Scoring

In the scoring phase,some real-time systems will use the same hardware that’s used in the data layer, but they will not use the similar data. At this phase of the process, the

deployed scoring rules are “divorced” from the data in the data layer.

### 2.5 Model Refresh

Data is always changing, so there is need to refresh the data and refresh the model that is built on the actual data. To refresh the models, The existing programs which are used to run the data and build the models can be re-used.It also includes periodic model refreshes

## 3. Methodology

### 3.1 HPCC

It is High Performance Computing Cluster.It is utilized to give elite and data parallel processing for applications that incorporates Big Data. It was concocted in 1999 by LexisNexis. Extensive volumes of organized and unstructured information is utilized by business model to change over it into a tremendous social chart of individuals and business.

The HPCC framework design incorporates two distinctive group handling situations, both are enhanced freely and afterward utilized for its parallel data processing purpose. The one of these stages is known as an data refinery(Thor) which is utilized for general handling of enormous volumes of crude information of any sort furthermore utilized for various purposes which incorporates information purifying and cleanliness, remove, change, stack preparing of the crude information, record connecting and entity resolution, large-scale ad-hoc complex analytics, and creation of keyed data and indexes to support high-performance structured queries and data warehouse applications.

The other one of the parallel data processing platforms is called Roxie and functions as a rapid data delivery engine. This platform is designed as an online high-performance structured query and analysis platform or data warehouse delivering the parallel data access processing requirements of online applications through Web services interfaces supporting thousands of simultaneous queries. Roxie utilizes a distributed indexed filesystem to provide parallel processing of queries using an optimized execution environment and filesystem for high-performance online processing. After that hadoop appears.

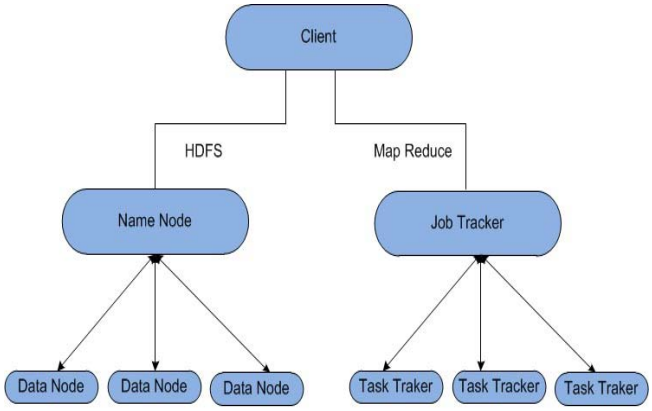
### 3.2 HADOOP

It is an open-source software framework for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. Every one of the modules in Hadoop are planned with a key presumption that equipment disappointments are basic and ought to be naturally taken care of by the system.

Hadoop was developed at Yahoo to list web information. The venture was begun in 2006. In 2008 Yahoo discharged the source as open source.

The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part called MapReduce.

**Hadoop Diagrammatical Overview**




#### 3.2.1 HDFS

Hadoop incorporates a fault-tolerant storage system called the Hadoop Distributed File System, or HDFS. HDFS can store immense measures of data and scale up incrementally. Hadoop creates clusters of machines and coordinates work among them. Clusters can be worked with cheap PCs. In the event that one fails, Hadoop keeps on working the group without losing information or intruding on work, by moving work to the rest of the machines in the Cluster. HDFS oversees storage on the cluster by breaking approaching records into pieces, called "blocks," and putting away each of the blocks redundantly over the pool of servers. In the normal case, HDFS stores three finish duplicates of every document by replicating every piece to three distinct servers.

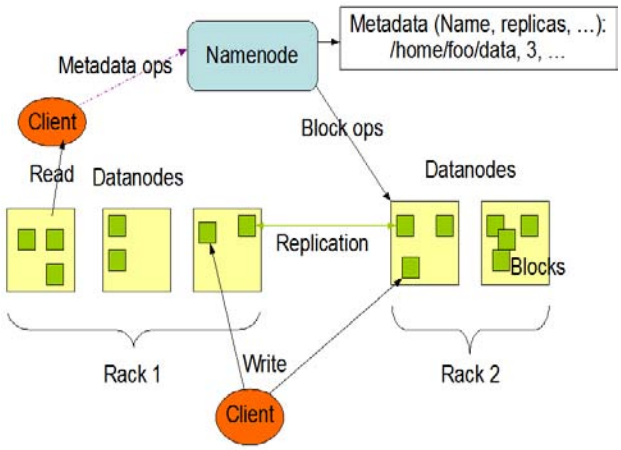
It is intended to store substantial information sets reliably, and to stream those information sets at high data transmission to client applications. By distributing storage and computation over numerous servers, the asset can develop with demand while staying efficient at each size.

In general, divide-and-conquer strategy of processing data is nothing really new, but the combination of HDFS being open source software (which overcomes the need for high-priced specialized storage solutions), and its ability to carry out some degree of automatic redundancy compound annual growth rate (CAGR) of 58 percent until 2018.

HDFS [1] has a master/slave architecture. An HDFS cluster consists of a single NameNode, a master server that manages the file system namespace and regulates access to files by clients. In addition, there are a number of DataNodes, usually one per node in the cluster, which manage storage attached to the nodes that they run on. HDFS exposes a file system namespace and allows user data to be stored in files. Internally, a file is split into one or more blocks and these blocks are stored in a set of DataNodes. The NameNode executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes. The DataNodes are responsible for

serving read and write requests from the file system's clients. The DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode.

HDFS Architecture




**Benefits of HDFS**

**Big data capable**

The Hallmark of HDFS is its ability to tackle big data use cases and most of the characteristics that comprise them (data velocity, variety, and volume). The rate at which HDFS can supply data to the programming layers of Hadoop equates to faster batch processing times and quicker answers to complex analytic questions.

**Portability**

One benefit of HDFS is its portability between various Hadoop distributions, which helps minimize vendor lock-in.

**Cost-Effective**

As previously stated, HDFS is open source software, which translates into real cost savings for its users. As many companies can attest, high-priced storage solutions can take a significant budgets and are many times completely out of reach for small or startup.

**3.2.2 Map Reduce**:

The processing pillar in the Hadoop ecosystem is the MapReduce framework. The framework allows the specification of an operation to be applied to a huge data set,divide the problem and data, and run it in parallel. From ananalyst's point of view, this can occur on multiple dimensions.For example, a very large dataset can be reduced into a smaller subset where analytics can be applied. In a traditional datawarehousing scenario, this might entail applying an ETL operation on the data to produce something usable by the analyst. In Hadoop, these kinds of operations are written as MapReduce jobs in Java. There are a number of higher level languages like Hive and Pig that make writing these programs easier. The outputs of these jobs can be written back to either HDFS or placed in a traditional data warehouse.

There are two functions in MapReduce as follows:

Map – the function takes key/value pairs as input and generates an intermediate set of key/value pairs

Reduce – the function which merges all the intermediate values associated with the same intermediate key be accepted by mapper phase.

Below are steps that happen as part of the MapReduce life cycle.

Step 1. Input request should in the form of .JAR file which contains Driver code, Mapper code and Reducer code.

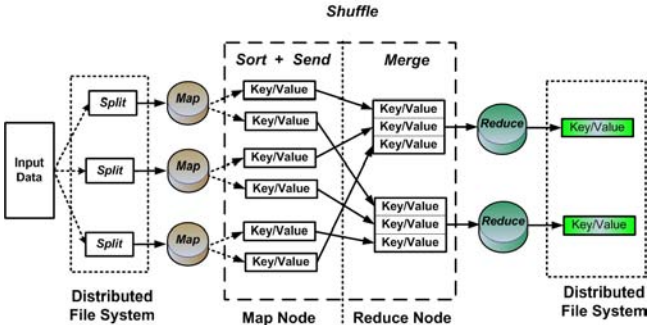
Step 2.Job Tracker assigns the mapper tasks by tracking the business logic from the .JAR file on the all the available task trackers.

Step 3.Once all the task trackers are done with mapper processes, they send the same status back to Job Tracker.

Step 4.All the task trackers do with mapper phase, then job tracker initiates sort and shuffle phase on all the mapper outputs.

Step 5.When the sort and shuffle is done, job tracker initiates reducer phase on all available task trackers.

Step 6.Once all task trackers do with reducer phase, they update the same status back to the job tracker.Mapper and Reducer are user driven phases. Mapper class output filename is "part-m-00000" and Reducer class output filename is "part-r-00000". Job Tracker and Task Tracker are the two daemons which are entirely responsible for map reduce processing.




**Comparison of HPCC and HADOOP [2]**

| *HPCC* | *HADOOP* |
|---|---|
| It uses native filesystem to store files. | It uses Distrubuted File System to store files. |
| Thor is used to provide data warehouse functions. | Hive is used to provide Data Warehouse structure to HDFS files. |
| ECL is used which is a declarative SQL-like language. | Pig provides easy declarative language constructs. |
| Difficult to horizontally sacle by simply adding nodes. | It has horizontal Scalability. |
| It uses C/C++ and FORTRAN | It uses latest languages like java/python. |
| HPC is primarily used for scientific research applications in areas like life sciences, healthcare and mining sectors. | Hadoop is focused on analytics use cases and focus is really commercial. |
| Infrastructure cost is more as compared to Hadoop as it needs high end computing components. | Hadoop runs on commodity hardware which means infrastructure costs can be less. |

## 4. Benefits of Big Data

1) Cost reduction
2) Faster, better decision making
3) Better Data Management
4) Unlimited storage
5) Accessible from any place as normally stored in clouds
6) Speed of Data transmission and processing is very high

## 5. Challenges of Big Data

- Requires special computer power.
- Large noise
- Privacy problems

## 6. Conclusion

We have entered an era of Big Data. The paper describes the concept of Big Data along with 3 Vs, Volume, Velocity and variety of Big Data. The paper describes real time deployment phases of Big data and different techniques such as HPCC and Hadoop which are open source softwares used for processing of Big Data as HPCC is older technique and Hadoop is latest one. Comparison of both the techniques also described in this paper. As information technology systems become less monolithic and more distributed, real-time big data analytics will become less exotic and more commonplace.

## References

[1] International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014 ISSN 2250-3153 **A Review Paper on Big Data and Hadoop**
[2] https://hpccsystems.com/why-hpcc-systems/hpcc-hadoop-comparison
[3] https://www.ijarcsse.com/docs/papers/Volume_5/5_May 2015/V5I5-0720.pdf
[4] https://en.wikipedia.org/wiki/Big_data
[5] http://www.researchgate.net/profile/Bhu
[6] http://www.sciencedirect.com/science/art
[7] http://www.ijetajournal.org/volume-3/iss
[8] **Real-Time Big Data Analytics: Emerging Architecture** by Mike Barlow Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA