# BIG ANALYTICS OVER ENORMOUS UNSTRUCTURED & SEMI-STRUCTURED MULTIDIMENSIONAL KNOWN-UNKOWN DATA - THE BIG DATA HADOOP

Mr.Rajkumar G Verma
Network Administrator
Symbiosis International University
Pune, Maharashtra

***Abstract*-Known-Unknown data is continuously increasing at a high speed. As a consequence of which, producing it is getting difficult to store or analyze such vast amount of data.Exabyte's, for instance.Collecting this large amount of known-unknown data requires high speed of computing resources, which are evidently considered to be main hurdles in present scenario.**

**Many organizations have limitations to buy machines that comprise of processing throughputs and memory for storing the data, which is expensive solution, but the fact remains that the business processes no longer fit on a single cost effective computer. Therefore, it becomes imperative extracting useful knowledge or information from the huge large-scale data that is enormous in size which is the data repositories.Like an open internet known-unknown big data repositories. These data repositories can be carried out with the mean of analytics which may be the complex procedure.**

**From openinternet orintranet enormous amount of unstructured-structured known-unknown data is continuously produced by various means like high real time performance applications, streaming media, voice applications, research organizations applications, social networks,government sites, health organization, manufacturing industries' application's data bases that leads to a huge amount of raw data which refers to Big data.**

**There is a great scope of development in big data concepts, thatare constantly changing to conduct research studies on amountof unstructured-structured, known-unknown, multidimensional data and to some extent the raw hidden data.**

***Keywords: Qubits, Bigdata, MapReduce, Hadoop Distributions, Quantum Computations.***

## INTRODUCTION

With the advancement of technology, large amount of known-unknown data that is generated from various means of digitaldevices, decillion internetconnected devices,sensors node, and distribution system irrespective of the departments that is associated with it. For instance, water purification department, electricity department, commodity markets,share markets,constructions industries,mining industries, medical science,software based application,e-commercesites ,space center data , there are abundant studies generating huge amount of data leading to expensive computational resources comprising of adequate memory, storing space, distribution technique,new technologies architecture computational resources which altogether need big analytics tools like big data-Hadoop.

## SENSING KNOWN-UNKNOWN BIG DATA ON, INTRANET,OPEN ENVIRONMENT

Continually generating, vast amount of multidimensional, known-unknown big data on Internet or intranet is not limited to finite few big players such as Google, Yahoo, Microsoft in latest dynamic technology trends. Evensmall organizations are generating huge amount of known-unknown data.

However with recent technology transformation, cloud computing, quantum computations, data centres, hardwarestorage, cputhroughputs increases and the cost comes down rapidly. As a result all this, domainsstart producing a massive amount of data and organizationsneed to store them for a long time due to inexpensive online-offline storage and processing throughputs capabilities. These stored known-unknown multidimensional big data gives rise to number of challenges such as processing, accuracy, integrity, and throughputs, analyzing, monitoring.

Sensing data, the digital sensors capture information like proper supply of water incites, powerdistribution, gasoline distribution ,agro-industries' situations in order to utilize is in the most favorable way and monitor the resources and manage supply chain, and logistics. Coming to latest technologies likequantum computations where information can store as 0's or 1's or the both at the same time qubits which leads to a drastically huge data collection.

## COMMON SPECIFIC CHARACTERISTICS

- Distribution of data repositories and the actual size is refers to large-scale data.
- All real time running applications on large scope, enormous data repositories having capabilities to scale over growing in size inputs rapidly terms as a scalability attributes.
- Support advance machine learning algorithms process from low-level known-unknown data to some extent structured information.
- Inventing,developing robust and interpretable analysis over the big data repositories in order to derive intelligence and extract authentic useful information from them.
- Data sizes such as Terabytes (TB),petabytes(PB),zettabytes(ZB) relates to volume of data.
- How frequently known-unknown data is generating relates to velocity of data. For example every fraction of time seconds.
- Decillion digital devices are connected with the internet or intranet generating different types of data relates to variety of data from the different sources produce bigdata such as digital sensors, digitaldevices, social networks, and streamed data.
- Complexity of data different network topology multivariate protocol layer.

## ANALYTICS OVER BIG DATA

Unstructured semi-structured multidimensional known-unknown DataScale-out instead of scale-up.There arises a frequent need to buy bigger machine with respect to speed, memory to run bigger database, which is very expensive, in fact the need of the hour is for purchase of the "database-class servers" for scaling commercials relations database. The high-end machines are not cost effective for many applications.

Data acquisition is the first step. Most of the sources produce staggering amounts of raw data much of this data is of no use, further can be filtered and compressed by orders of magnitude. Defining these filters in such a way that they do not discard useful information is one of the big challenges.

For many organizations data storage and processing is not considered to be a new problem. Fraud detections, cyber crime, data integrity, authenticity, accuracy and many other applications have to deal with these issues for decades. Dramatically what happened when the variety, velocity, volume of this data has changed .This makes sense as many algorithms efficiency helps in knowing more about the problem leads to better decisions which leads to revenue generation, desired result, reduced fraud ,safer conditions .

Simply identifying, locating, understanding, citing data is not only the question more challenging way is Data analysis.

In Big data computing environment's data mining require efficiently accessible data, declarative query, integrated,trustworthy scalable mining algorithms. There
is multi-step pipeline required to extract value from data. Incompleteness, scale, timeliness, integrity and complexity of process give rise to all challenges at all phases of the pipeline.

## TOOLS AND TECHNIQUES AVAILABLE FOR ANALYZING THE KNOWN-UNKNOWN BIG DATA

The following tools and techniques are available

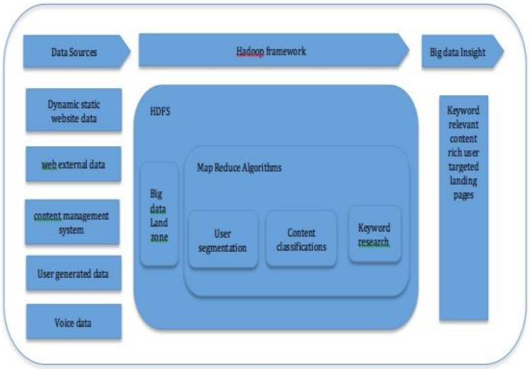
### HADOOP

Open source platform apachehadoop provides scalablecoding , cost-effective, scalableinfrastructure, pragmatic for developing, designing,batch data processing system for many types of applications.Apache hadoop consist of a distributed file system called the hadoop distribution filesystem(HDFS) and set of computational layer that implements a process paradigmcalled Mapreduce.

With existing old commodity server machine with non-specialized network or hardware infrastructure which may be simple, multiple individuals or groups cluster can be shared to form a single,storage, parallel processing, logical and compute platform or cluster form with the hadoop system.

Hadoop runs on individual machines memory,cpu,storage where the data lives rather than putting data across the network,which leads to a greater performance in hadoop environment.

### BIG DATA HADOOP ARCHITECTURE



Hadoop Cluster consists of master/slave architecture for both distribution storage and distribution computation. Distribution of storage system on different machine is called the Hadoop distribution file system or HDFS.HDFS consists of master namenode that directs the slave DataNodedeamons to perform the low level input output tasks.Namenodekeeps track how your files are broken down into the file blocks,which nodes store those file block in a HDFS and keeps track of overall health of the distribution file system.

As server hosting the Namenode,doesn't store any user data or user interaction,computation for a MapReduce program to lower the workload on the machine.Hence the NameNode server space does not double as a DataNode or a TaskTarcker.

### HADOOP COMPONENTS IN DETAIL

On a network set of daemonsor resident programs ,batch programs ,running on a different server, some daemon have specific roles, some reside only on one sever or across multiple servers. There are some specific daemons whichrun using hadoop.

These are as follows:

- NameNode
- DataNode
- Secondary NameNode
- JobTracker
- TaskTracker

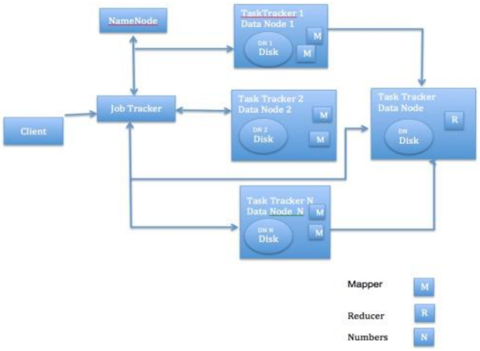
Especially NameNode installed on dedicated special machine,which stores all the metadata for the filesystem across the cluster and regulates access to files by clients. It is a single arbitrator and repository for all HDFS metadata.But on the other hand, it also creates a single point of failure - losing the NameNodeeffectively means losing the HDFS. A way out for this is through Hadoop, that implements a *Secondary name Node.*

Secondary NameNode(SNN) is not a "backup NameNode."It serves the checkpoint mechanism for the primary NameNode.Basically HDFSSNN store the HDFS Metatdata state at a point in time and edit log is a transactional log of every metadata change since the image file was created.

Secondary NameNode(SNN) holds a copy (out-of-date) of the primary’s persistent state in the form of last image.There is ongoing work to create a true backup NameNode which would be able to take over in the event of primary node failure.

DataNodeStores the actual blocks of a file in the HDFS on its own local disk on network each slave node in HDFS cluster will host a DataNode daemon which process the distribution filesystem means writing and reading the HDFS Intermediate daemons *JobTracker*plays a very important rolebetween the application and Hadoop cluster. It is one of the master components,responsible for managing overall execution of a job. Various functions performed by the jobtracker like scheduling child tasks (individual Mapper and reducer) to individual

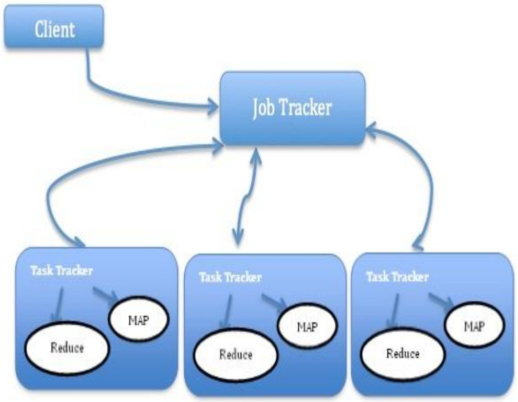
blocks to actual files on the local machine file system. When ever there is a transaction of reading or writing a HDFS file ,the file is split into the blocks and same time NameNode will intimate your client which Datanode each blocks resides in. Client directly communicates with the DataNode to process the local file system corresponds to the blocks. Replication of file data blocks to another DataNode can be raised. DataNode provides continuous heartbeat to NameNode regarding local changes as well as receive instructions to update, move, create or delete the blocks from the local disk.
nodes,monitoring health of each task and node,even rescheduling failedtasks. On every hadoop cluster there is only one jobtrackerdeamon.



*JobTracker : Process Flow for Job submission*

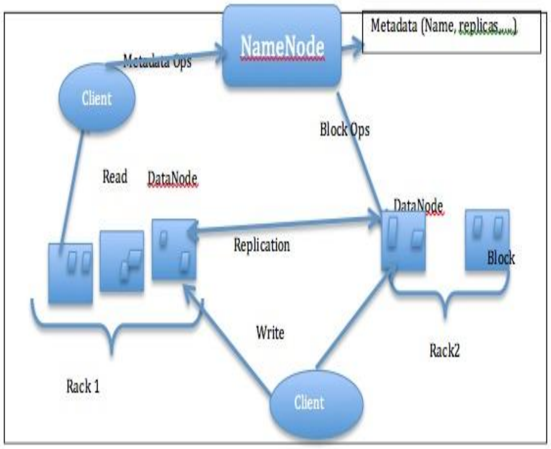
Managing individual task execution on each slave machine is done by the *TaskTrakcer*Tasktracker accepts requests for individual tasks such as Map,Reduce and Suffle operations.TaskTracker further configured with a set of slots that usually depends upon the total number of cores available on the machine.When a request is initiated by the jobTracker to launch a new task the TaskTrackerinitiates new (Java virtual Machine)JVM for the task.

TaskTracker is responsible for sending the heartbeat message & number of free available slots to the JobTracker.



Above figure depicts the JobTarcker and TaskTracker interaction.Client calls the job tracker to begin a data processing job,the job tracker partitions work and assigns different map and reduce tasks to each tasktracker in the hadoop cluster.

## HADOOP DISTRIBUTED FILE SYSTEM



MapReduce : Functional programming concepts operate on one record at a time  builds data processing applications.MapReduce further divided two distinct things ,the programming model and the specific implementation of the framework.MapReduce is the most prominent way of writing applications through which is the process simplified by the development of large scale, fault-tolerant ,distributed data processing applications.

In MapReduce, Programmers writes codes that consist primarily of a map function and a reduce function, and framework handles scheduling parts of the job on slave machine, parallelizingthe work, monitoring for and recovering from failures and so forth. Framework is invoked by the user-provided code.

MapReduce provides simplicity for development of no socket programming, no threading, synchronization logic, no special techniques to deal with enormous amount of data.

MapReduce design to share nothing system, task do not share state each other.
They can execute in parallel and on separate machines. Hot swappable hardware can be done in Mapreduce systems this is one of the best scalability of the MapReduce.

MapReduce consists of fault tolerance and automatic parallelization and distribution of work, which is done by the map, and reduces functions that process individual records in dataset. Framework is responsible for splitting a MapReduce job into tasks.

A MapReduce program process data by combination key/value pairs in the general form.

Map:(k1, v1) ->list(k2,v2)
Reduce: (k2, list (v2))->list (k3, v3)

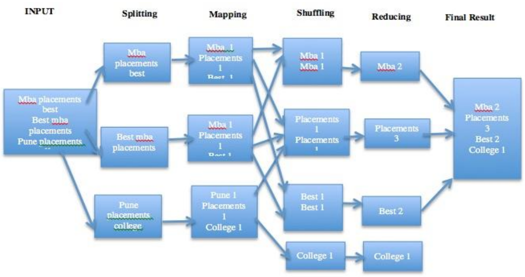
In a Hadoop Mapreduce there are two major daemons namely jobtracker and the tasktracker .

The general MapReduce data flow. Input data distributing to different nodes, at “shuffle” step nodes can communicate with each other. Restriction on communications greatly helps in scalability.

There are various component of MapRedcue, includingthese:

- Mapper
- Reducer
- Combiner
- Partitioned

For instance, the overall MapReduce word Count process for a website page info which will increase the visibility with respect to best MBA, Placements, Institute keywords on the Internet.


INPUT
Splitting
Mapping
Shuffling
Reducing
Final Result
Mba placements best Best mba placements Pune placements
Mba placements best
Best mba placements
Pune placements college
Mba 1 Placements 1 Best 1
Mba 1 Placements 1 Best 1
Pune 1 Placements 1 College 1
Mba 1 Mba 1
Placements 1 Placements 1
Best 1 Best 1
College 1
Mba 2
Placements 3
Best 2
College 1
Mba 2 Placements 3 Best 2 College 1

## DISCUSSION

Authentic actionable results or outcomes can be carried out by analyzing huge known-unknown multidimensional structured-unstructured data through the proper usage of big data tool-hadoop.

Actionable result or outcomes leads to revenue generation for the organization who have planned and implemented the big data processing.Scalable data can be streamlined for proper computation process.

Investing huge cost on a big server can be minimized with the help of distribution storage and MapReduce functionality.Investing in the hadoop tool, Giant organization or Turnkey type organization will make their foremost decisions.Attributes like speed and accuracy, and data integrity can be carried out by the sensitive data analyzing.

Top IT leaders realize that they need to invest on big analytics programs to hold the wealth of client data they are capturing, relevant patterns, insights to aggressively attack the market and establish dominance.

Due to parallel distributions storage and processing, managing &monitoring big server infrastructure can be minimized.

## CONCLUSION

Analytics and information, when defined, the traditional measurement mechanisms do not work efficiently. Many organizations may be concerned about the service accuracy and quality in addition to the cost and delivery time.In today’s World, most of the organizations, fundamentally depend on their data and information handling services facilitated by their information technology to capture, store, flow, manage and analyze data in a better way.

## LIMITATIONS AND FUTURE WORK

There are additional and important theories and models that are not addressed. The details on how the distribution, managing & monitoring of resources would impact the operations of specific domain are overlooked.
How to educate new Data Base Administrators, data engineers, data analysts and the users of big data technologies has not been addressed.

The hindrances caused due to the presence of Voice data caused by human beings, or other elements like voice data generated through vehicles, emissions from Industrial hubs, natural calamities etc, that are prominently present in the External atmosphere cannot be destroyed until collapsed with high speed of collisions, that occur, and in a scattered mode.

When data can be collected of the type of huge known-unknown, unstructured-structured & scattered data, it can be figured out with the help of big data tools.

Latest technologies like quantum computation, where information can be stored as 0’s or 1’s or both at the same time, which leads to a very huge data collections. This collection of enormous huge data can be analyzed with big data techniques.

## REFERENCES

[1] Apache Hadoop, http://hadoop.apache.org.
[2] Gridmix.HADOOP-HOME/mapred/src/benchmarks/gridmix in Hadoop 0.21.0 onwards.
[3] Chris Eaton, Dirk Deroos, Tom Deutsch, George Lapis, and Paul Zikopoulos, Understanding Big Data.: McGraw-Hill Companies, April 2012, http://public.dhe.ibm.com/common/ssi/ecm/en/iml14296usen/IML1429 6USEN.pdf [Accessed on: 2012-06-08].
[4] Yuri Demchenko, Zhiming Zhao, Paola Grosso, AdiantoWibisono, Cees de Laat, “Addressing Big Data Challenges for Scientific Data Infrastructure”, IEEE , 4th International

Conference on Cloud Computing Technology and Science, 2012

[5] D.L. Jones et al., "Big data challenges for large radio arrays," , march 2012, pp. 1-6. S. Lohr. The age of big data.New York Times (http://www.nytimes.com/2012/02/12/sunday-review/big-impact-in-the-world.html), Feb 2012.

[6] Apache HBase, http://hbase.apache.org.

[7] P. Allen, S. Higgins, P. McRaie, H. Schlaman, Service orientation: winning strategy and best practices, Cambridge University Press, New York, NY, 2006.

[8] P. Checkland, S. Holwell, Information, systems, and information systems: making sense of the field, Wiley, Chichester, UK, 1998/2005.

[9] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery: an overview, Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996, pp. 1–34.

[10] Borgman CL, Wallis J, and Enyedy N. 2007. Little science con- fronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. Int J Dig Lib 7: 17–30.

[11] G. Ananthanarayanan, S. Kandula, A. Greenberg, I. Stoica, Y. Lu, B. Saha, and E. Harris. Reining in the outliers in map-reduce clusters using mantri. In USENIX OSDI'10, December 2010.

[12] K. Webb, A. Snoeren, and K. Yocum. Topology switching for data center networks.In USENIX Hot-ICE'11, March 2011.

[13] Sachchidanand Singh, Nirmala Singh, "Big Data Analytics", IEEE, International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, 2012.

## Author Profile

**Rajkumar G Verma ,**Currently working with Network Administrator in Symbiosis International university,SIMS.He received his Bachelor of Computer Science, Bachelor of Education, Master in Computer Management Degree from Pune University, MBA from Chennai Board, Currently doing M.Tech level program from National Institute of Electronics & Information Technology (NIELIT).His research Interest includes Networking & Security, Computer Forensic ,Parallel & Distributed Computing ,Gesture Recognition, Computing,Sensor,Wireless Communications.