International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), Volume 5, Special Issue 2, May 2015)
International Conference on Advances in Computer and Communication Engineering (ACCE-2015)

A Comprehensive Survey on Big Data Issues and Alternative Approaches to Hadoop MapReduce

Gauri S. Rapate[1], Nandita Yambem[2]
[1,2]*Assistant Professor, Vemana Institute of Technology, Bangalore*
[1]gauri.rapate@gmail.com, [2]nanditayambem@gmail.com

***Abstract*—as we all are aware that [Big] Data is created by human beings and gadgets every second, is ever growing in phenomenal rate. These big Data needs to be stored and analyzed for reference purposes & to gain Intelligence or knowledge for futuristic decision making. We conducted exhaustive technological survey to cover the aspects like Cloud computing - its advantages and disadvantages, Big Data & its issues, Tools & Techniques like – Hadoop, MapReduce, stages, Open source alternatives to Hadoop etc..**

***Keywords*— BigData, Cloud Computing, Hadoop, MapReduce, Open source.**

I. Introduction To Cloud Computing

Cloud computing refers to a paradigm, in which a n number of systems are connected in a network which can be private or public, to provide dynamically scalable infrastructure for application, data and file storage by making use of a service over the internet, at another location[1].

This will reduce the computation cost, hosting an application, content storage and delivery significantly.

Forrester defines cloud computing as:

"A pool of abstracted, highly scalable, and managed compute infrastructure capable of hosting end-customer applications and billed by consumption."

The cloud computing concept is based on the principal of 'reusability of IT capabilities'. The uniqueness of cloud computing is that if compared to traditional concepts it broadens horizons across organizational boundaries.

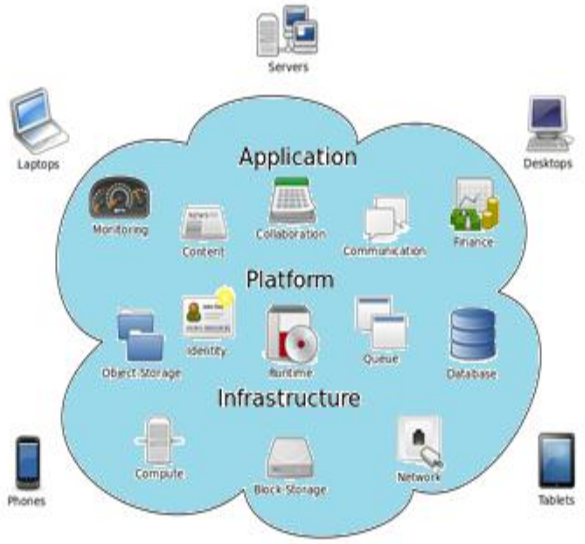

Fig 1:- Cloud Computing [7]

A. Advantages:-

1. Cloud users need not have to invest in information technology infrastructure, purchase hardware, or buy software licenses, the benefits are low up-front costs, rapid return on investment, rapid deployment, customization, flexible use, and solutions that can make use of new innovations.
2. Personal information may be better protected in the cloud. Using better security mechanisms cloud computing may improve efforts to build privacy protection into technology.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), Volume 5, Special Issue 2, May 2015)
International Conference on Advances in Computer and Communication Engineering (ACCE-2015)

3. The cloud may also encourage open standards for cloud computing which establishes baseline data security features common across different services and providers.
4. Information in the paper documents or hard drives can be easily lost compared to cloud.

B. Disadvantages: -

1. *Security:* Cloud Computing may give rise to certain privacy implications as data is stored in remote locations. Cloud providers often serve multiple customers simultaneously which raises the scale of exposure to possible breaches.
2. *Costing Model:* Migration to the Cloud significantly reduces the infrastructure cost and raises the cost of data communication, like transferring an organization's data to and from the public and community Cloud especially in the hybrid cloud deployment model and per unit of computing resource's cost are on the higher side.
3. *Service Level Agreement (SLA):* Cloud consumers need to ensure the quality, availability, reliability, and performance of the computing resources when consumers have migrated their core business functions onto their entrusted cloud, i.e. it is vital for consumers to obtain guarantees from providers on service delivery.

   These guarantees are provided through Service Level Agreements (SLAs) negotiated between the providers and consumers.

   Different cloud offerings (IaaS, PaaS and SaaS) have to define different SLA meta specifications which raises a number of implementation problems for the cloud providers.

   User feedback should be constantly incorporated in Advanced SLA mechanisms
4. *Cloud Interoperability Issue:* Interoperability's main goal is to realize the seamless fluid data across clouds and between cloud and local applications. There are a number of levels that interoperability is essential for cloud computing.

   First, to optimize the IT asset and computing resources and secondly for the purpose of optimization, an organization may need to outsource a number of marginal functions to cloud services offered by different vendors. To address the interoperability issue standardization is a good solution

C. Security Measures

The security measures vary from one cloud provider to another cloud provider and among the various types of clouds. Like, encryption methods the providers have in place, methods of protection they have for the actual hardware where the data will be stored on, backups of the data and firewalls set up.

For community cloud, there should be barriers to keep information separate from company to company.

II. BIGDATA

Big Data is a term for datasets which describe enormous volumes of structured and unstructured data which is very difficult to process using traditional databases and software technologies.

The Big Data have the following properties:

a) Volume: Factors like storing transactions data, live streaming data and data collected from sensors contribute towards increasing Volume.

b) Variety: There are different formats of data available today like traditional databases, text documents, emails, video, audio, transactions etc.

c) Velocity: Velocity means the need of speed to process the data and to produce the data.

d) Variability: The variations in data flows can be highly inconsistent.

e) Complexity: Complexity of the data increases as the data comes from multiple sources and then the data has to be linked, matched, cleansed and transformed into required formats for processing.

The real time examples of bigdata are credit card transactions, Walmart customer transactions, and Facebook users generating social interaction data.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), Volume 5, Special Issue 2, May 2015)
International Conference on Advances in Computer and Communication Engineering (ACCE-2015)

Fig 2:- Bigdata

A. Bigdata Issues

1. The biggest challenge of big data is to go through the large volumes of data and access it in detail at a high speed. The challenge is directly proportional to the degree of granularity.
2. The next issue deals with the understanding of data to visualize it as part of data analysis.
3. Displaying meaningful result can again become a challenge for extremely large data and different categories of information. Eg: - Plotting points on a graph for analysis becomes difficult [6].

III. TOOLS AND TECHNIQUES

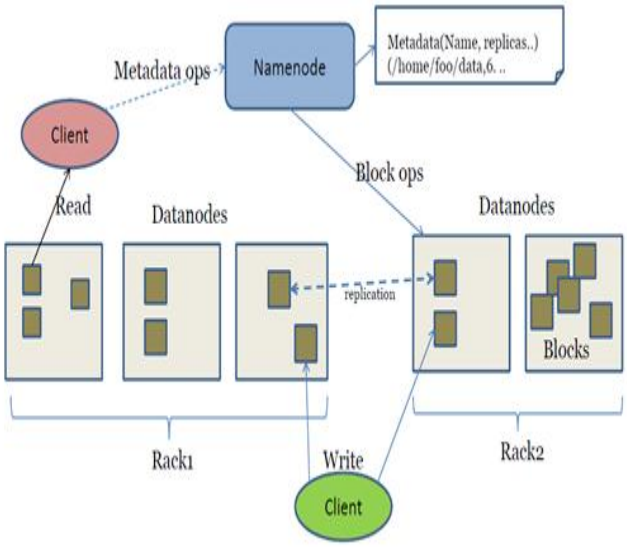

Fig 3:- HDFS Architecture

A. Hadoop

Hadoop is a free, Java-based programming framework hosted by the Apache Software Foundation which supports the processing of large sets of data in a distributed computing environment.

Hadoop framework is used to develop applications capable of running on clusters of computers and perform complete statistical analysis for huge amounts of data.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), Volume 5, Special Issue 2, May 2015)
International Conference on Advances in Computer and Communication Engineering (ACCE-2015)

It provides two things:

1. A distributed filesystem called HDFS (Hadoop Distributed File System)
2. A framework and API for building and running *MapReduce* jobs.

HDFS is a file system where data storage is distributed across several machines. It links together file systems on local nodes to make it into one large file system. To overcome node failures HDFS replicates data across multiple sources, thereby improving reliability.

There are two and a half types of machine in a HDFS cluster:

- Datanode – these are the nodes for HDFS to actually store the data, there are usually quite a few of these.
- Namenode – It is the 'master' machine to control all the meta data for the cluster.
- Secondary Namenode – it is a separate service that keeps a copy of the edit logs, and file system image and appends periodically to keep the size reasonable.
- If the backup node and the checkpoint node is used then Secondary Name Node can be deprecated but the functionality remains similar.

HDFS has unique features that make it ideal for distributed systems:

- ***Failure tolerant*** – To avoid failures data should be duplicated across multiple data nodes. Generally, the industry maintains the standard of replication factor of 3.
- ***Scalability*** - For the read/write capacity to scale fairly well with the number of datanodes data transfers happen directly with the datanodes
- ***Space*** – If there is need of more disk space one can add more datanodes and re-balance

B. Mapreduce

MapReduce is a programming model where the parallel and distributed algorithms on a cluster and are used to process and generate large data sets.

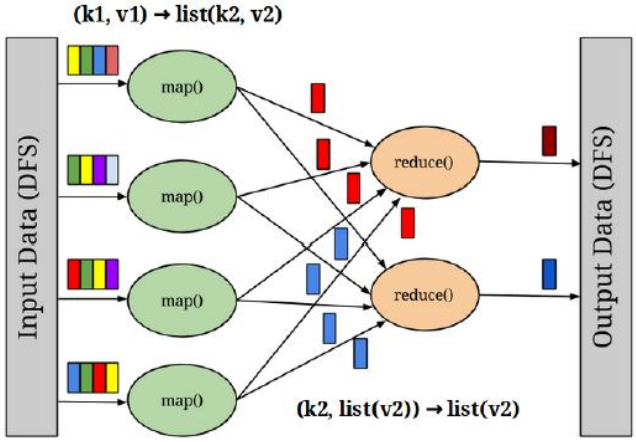

Fig 4 :- MapReduce

1) Different stages of MapReduce:

Here we have two stages; map and reduce which acts like two operators and form a static pipeline.

Mapper Operation: first the input is read and parsed from the distributed file system. This parsing is in the form of key-value pairs.

These key-value pairs are mapped using user-defined function to generate intermediate key-value pairs, which are then sorted and grouped by key.

Reducer Operation: the groups obtained by the mapper are then processed by parallel reduce tasks. In the framework partitioning stage pairs sharing the same key will be processed by same reduce task, then each group will be assigned a user defined function to produce the output, thereby creating a file in the distributed file system with results.

2) Advantages of MapReduce:

- The software where MapReduce is implemented i.e.Hadoop is free.
- It is low cost storage distributed file system.
- It can speed up certain data processing operations.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), Volume 5, Special Issue 2, May 2015)
International Conference on Advances in Computer and Communication Engineering (ACCE-2015)

- Here the data is distributed and computations are loc to data which avoids network overload.
- The programming model is very simple. Here the end user programmer has to only write map-reduce task. Independent task makes it easy to handle partial failure.
- The entire node can fail and restart.
- Reliability is achieved by parceling out many operations on the set of data to each node in the network.
- Each node reports back periodically with completed work and status updates. Dead node is recorded by a master node if a node falls silent for longer than specified interval.

3) Issues of MapReduce:

- Open-source software quality is highly variable because the economics of open-source development provide no incentive for software suitability or quality. Here, software engineers try their hands at software authoring. Firms selling open-source solutions, provides implementation services, so few incentives will be given to make the software easy to setup and use.
- Built to operate on arbitrary size clusters, so efficient storage was not designed .It has no query optimizer – Developers need to be sure to optimize their own data flow since there is no optimizer. Transaction consistency or recovery checkpoints as it were built to be a file system. So depending on nature of job Hadoop cluster may or may not be 100% accurate.
- HDFS was built without the notion of efficiency, resulting in multiple copies of the data. At a minimum, generally three copies of the data required and for maintaining performance data locality is needed, so often see six copies of the data required. So by definition data is "big".
- Some open source components are available which attempts to set up Hadoop as a query-able data warehouse, but offers very limited SQL support.
- The framework of MapReduce challenging and difficult to leverage for more than simple transformational logic. Some open source components attempts to simplify this, but use proprietary languages.
- Still software is under active development – HDFS has recently added support for append operation
- MapReduce is much less efficient than parallel database query systems the MapReduce model imposes too strong assumptions on the dependence relation among data, and the correctness often depends on the commutativity, associativity, and other properties of the operations because joins operation of multiple datasets are difficult and slow and no indices.
- MapReduce still is a single master, so it requires care and may limit scaling. The master node can easily become a single point of failure.
- MapReduce not suitable
    - for Real Time processing
    - When processing requires a lot of data to be shuffled over the network.

4) Possible solutions to issues of MapReduce:

For dealing with small files on Hadoop

- Change the "feeder" software means if small files are the problem, change the upstream code to stop generating them.
- An offline aggregation process to be run which aggregates the small files and re-uploads the aggregated files ready for processing thereby improving performance.
- At the start of the job flow an additional Hadoop step to be added which aggregates the small files, so reduces the number of additional moving parts.

Possible solutions to aggregate and compact small files on Hadoop

a. Filecrush - a tool developed by Edward Capriolo that crush small files on HDFS and is available as a jarfile such that it is ready to run on your cluster. By default it won't bother crushing a file which is within 75% of the HDFS block size

B. Consolidator - a Hadoop file consolidation tool developed by Nathan Marz.from the dfs-datastores library.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), Volume 5, Special Issue 2, May 2015)
International Conference on Advances in Computer and Communication Engineering (ACCE-2015)

a. S3DistCp - created by Amazon as an S3-friendly adaptation of Hadoop DistCp and if you are running on Elastic MapReduce, this can deal with the small files problem using its groupBy option for aggregating files.

IV. Open Source Alternatives To Hadoop

Some of the popular alternative to Hadoop is Disco, Spark, BashReduce, GraphLab, Storm (tech.backtype.com/), HPCC systems, etc.

- Disco allows developers to write MapReduce jobs in Python and backend is built using Erlang, the functional language, which has built-in support for consistency, fault tolerance and job scheduling. Disco does not use HDFS, rather it uses a fault-tolerant distributed file system DDFS.
- Spark developed by UC Berkeley allows in-memory MapReduce processing distributed across multiple machines and is implemented using Scala.
    - BashReduce developed by Richard Crowley implements MapReduce for standard Unix commands such as sort, awk, grep, join etc. It supports mapping/partitioning, reducing, and merging and has:
        - The ability to pass a filename to each process rather than the actual file data assuming that each machine has a local copy of the data or has access to a shared file system. i.e BashReduce "sort of" handles task coordination and a distributed file system, which greatly reduces the network bandwidth required.
        - Processing of a directory full of files rather than a single file. Files may be compressed using gzip and bashreduce will detect handles the decompression.
        - Instead of the default sort –M option, the -M option allows you to specify your own merge program.

A. GraphLab

GraphLab was developed at Carnegie Mellon and is designed for use in machine learning. The goal of Graph Lab is to make the design and implementation efficient and correct parallel machine learning algorithms easier.

GraphLab update phase can both read and modify overlapping sets of data which is accomplished by allowing the user to specify data as a graph where each vertex and edge in the graph is an associated memory. The update phases can be chained such that one update function can recursively trigger other update functions that operate on vertices in the graph. This makes machine learning on graphs more tractable, but also improves dynamic iterative algorithms.

GraphLab has reduced stage, called the sync operation. The results which are global can be used by all vertices in the graph. The sync operations are performed at time intervals, and there is not as strong of a tie between the update and sync phases are not necessarily dependent on some prior update completing.

B. Storm

Storm developed by Nathan Marz is a distributed, reliable, and fault-tolerant stream processing system and processes data in parallel.

Storm is written in Clojure, but can be written in any programming language .Storm is fault-tolerant, horizontally scalable, and reliable, fast.

The key properties of Storm are

- Storm's programming model is simple and dramatically lowers the complexity for doing real time processing.
- Storm runs on the JVM (and is written in Clojure), but any programming language can be used on top of Storm..
- For scaling a real time computation, add more machines and Storm takes care of the rest.
- Storm ensures that each message will be fully processed at least once i.e. exactly once as long as there are no errors.
- Storm is fast so ZeroMQ is used for the underlying message passing, and care is taken so that messages are processed extremely quickly.

C. HPCC Systems

High-Performance Computing Cluster developed by LexisNexis Risk Solutions. HPCC written in C++ makes writing parallel-processing workflows easier by using Enterprise Control Language (ECL). In-memory querying is much faster as there are less bloated object sizes originating from the JVM.

HPCC has two "systems" for processing and serving data

1. Thor Data Refinery Cluster which process data like Hadoop.
2. Roxy Rapid Data Delivery Cluster which supports transactions and similar to a data warehouse (like HBase) and supports transactions

V. CONCLUSION

After conducting our in-depth technological survey, our key findings were:

1. In Big Data applications 4Vs – Volume, Velocity, Variety & Variability are the key issues.
2. Cloud Computing is compulsorily required to store all these BIG DATA. Security & privacy are key concerns.
3. In comparison with Open Source alternatives with Hadoop, Open sources are not trustworthy and reliable for commercial applications. Also, Open source Sofware are freeware, are not supported & updated regularly.
4. MapReduce is a programming model where the parallel and distributed algorithms on a cluster, are used to process and generate large data sets.

REFERENCES

[1] Venkata Narasimha Inukollu, Sailaja Arsi and Srinivasa, Rao Ravuri K," Security issues associated with Big Data in Cloud Computing ". International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014.

[2] Monjur Ahmed1 and Mohammad Ashraf Hossain2, Cloud Computing And Security Issues In The Cloud, International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.1, January 2014

[3] Stephen Kaisler, FrankArmour J. Alberto Espinosa, William Money, "Big Data: Issues and Challenges Moving Forward", 2013 46th Hawaii International Conference on System Sciences.

[4] Kuyoro S. O. Ibikunle F. Awodele O., Cloud Computing Security Issues and Challenges, International Journal of Computer Networks (IJCN), Volume (3): Issue (5): 2011

[5] Guillermo Lafuente, 2014 "Big Data Security - Challenges & Solutions", available at www.mwrinfosecurity.com

[6] Five big data challenges, 2013 And how to overcome them with visual analytics and available at "sas.com/visualanalytics"

[7] Venkata Narasimha Inukollu, Sailaja Arsi and Srinivasa Rao Ravuri K HIGH LEVEL VIEW OF CLOUD SECURITY: ISSUES AND SOLUTIONS

[8] Dealing with Hadoop's small files problem SNOWPLOW, Alex Dean , 2013.

[9] VASILIKI KALAVRI, "Performance Optimization Techniques and Tools for Data-Intensive Computation Platforms" School of Information and Communication Technology, KTH Royal Institute of Technology, Stockholm, Sweden 2014

[10] "A Beginners Guide to Hadoop" available at http://blog.matthewrathbone.com/2013/04/17/what-is-hadoop.html

[11] "MapReduce" available at http://en.wikipedia.org/wiki/MapReduce

[12] Sandya Mannarswamy "Taming the Big Data Beast with Hadoop and Alternatives" ,2012

[13] Hadoops Limitations for Big Data Analytics, Paraccel , 2014

[14] VASILIKI KALAVRI,"Performance Optimization Techniques and Tools for Data-Intensive Computation Platforms" School of Information and Communication Technology, KTH Royal Institute of Technology, Stockholm, Sweden 2014