International Journal of Innovative Research and Advanced Studies (IJIRAS)
Volume 4 Issue 9, September 2017

ISSN: 2394-4404

A Survay Of Storage And Processing Methodology In Cloud Computing With Hadoop Framework

S. Rekha

M.SC., M.PHIL., B.ED., Department of Computer Application, T.K.S College Of Arts And Science (Koduvilarpatty), Theni

G. Nithya

M.SC., M.PHIL, Department of Computer Application, T.K.S College Of Arts And Science (Koduvilarpatty), Theni

Abstract: ***Hadoop is an open source Apache Software Foundation project that enables the distributed processing of large datasets across clusters of nodes. we can assume cloud computing to be a concept or methodology. Cloud is a Pool of servers, all the servers are interconnected through internet, The main problem in cloud is retrieving of data (knowledge) and process that variety of data and here other problem is security for that data, Generally now a day " s different types of, I mean variety of data (Structured, semi-structured and Unstructured data) is existed in the different social applications (face book).So, and another problem with historical data retrieving. These types of problems are resolved with help of hadoop frame work and Sqoop and flume tools. Sqoop is load the data from database to Hadoop (HDFS), and flume loads the data from server files to hadoop distributed file systemCloud computing requires higher processing power than cloud storage. Cloud storage, on the other hand, needs more storage space. ... Cloud storage is simply a data storage and sharing medium, while cloud computing gives you the ability to remotely work on and transform data (for example, coding an application remotely). . Storage problem is resolving with help of blocks in hadoop distributed file system and processing is resolving with help of map reduce and pig and hive and spark etc. This paper summarizes the storage and processing methodology in cloud computing with hadoop framework.***

Keyword: sqoop, hadoop, HDFS, Iaas, PaaS, Saas.

I. INTRODUCTION

Hadoop is a platform which helps you in providing cloud computing services to your customers. Hadoop itself is based on the methodology of Distributed computing. It's like asking, what's the difference between programming and Java. ... Same holds true in case of Hadoop and cloud computing. Now a day‟s the enhanced cloud computing servers and nodes are having high configurations, the hadoop framework is require a high configurations for data storing and retrieving (processing) of wanted data. Hadoop itself is based on the methodology of Distributed computing. Hadoop as a means to provide some of the cloud computing services. Take, Amazon's EMR for instance which utilizes a hosted Hadoop framework running on

The web-scale infrastructure of Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Simple Storage Service (Amazon S3) to enable us easily and cost-effectively process vast amounts of data. Actually data is stored in the form of rows and columns in database, it is structure data, there is no problem with structure data, sometimes applications having both image and text formats and unstructured formats, at this time facing a problem on retrieving of wanted and required query relevant data.

The rise of cloud computing made dynamic provisioning of elastic capacity on-demand possible for applications hosted on data centers. This is because cloud data centers contain thousands of physical servers hosting orders of magnitude more virtual machines that are allocated on demand to users in a pay-as-you-go model.

SERVICE MODELS

Hadoop distributed file system overcome this type of (data loss) draw backs with help of replication of data, hadoop having a replication factor is 3, hadoop stores the 512 copies maximum.

Service models:

- ✓ Infrastructure as a Service (IaaS).
- ✓ Platform as a Service (PaaS).
- ✓ Software as Service (SaaS).

Each of these models provides a different view for users of what type of resource is available and how it can be accessed. In the IaaS model, users acquire virtual machines that run in the hardware of cloud data centers.

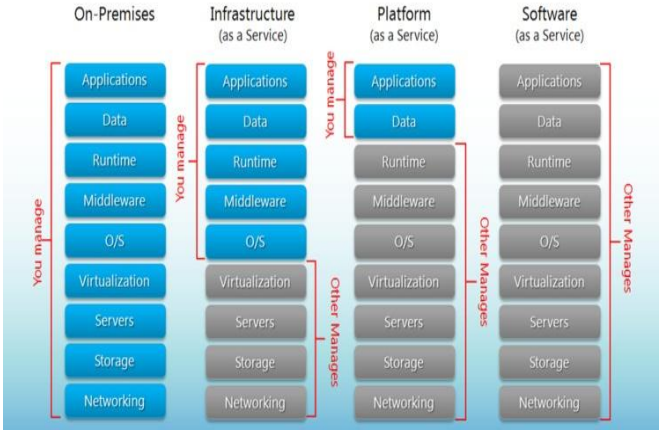

Figure 1: Cloud service model

II. METHODOLOGY

A. HADOOP METHODOLOGY

Hadoop itself is based on the methodology of Distributed computing. It's like asking, what's the difference between programming and Java. ... Same holds true in case of Hadoop and cloud computing. Long story short, Hadoop is a platform which helps you in providing cloud computing services to your customers.

Hadoop is a software framework for *distributed processing* of *large datasets* across *large clusters* of computers

- ✓ Hadoop is open-source implementation for Google MapReduce
- ✓ Hadoop is based on a simple programming model called *MapReduce*
- ✓ Hadoop is based on a simple data model, *any data will fit*
- ✓ Hadoop framework consists on two main layers
- ✓ Distributed file system (HDFS)
- ✓ Execution engine (MapReduce)

B. HADOOP INFRASTRUCTURE

- ✓ Hadoop is a *distributed* system like *distributed databases*
- ✓ There are several key differences between the two infrastructures
- ✓ Data model
- ✓ Computing model
- ✓ Cost model
- ✓ Design objectives

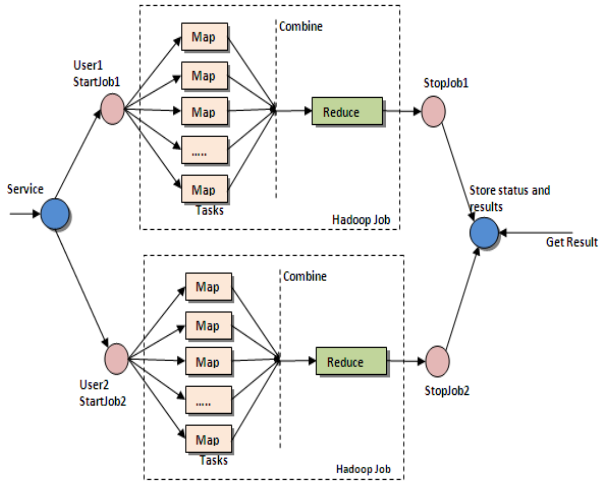

Figure 2: Hadoop Infrastructure

NOTION OF A JOB DIVIDED INTO TASKS

- ✓ Map-Reduce computing model
- ✓ Every task is either a map or reduce

C. ANOMALY PREDICTION

Alternative sources of statistics may venture Different degrees of planning, and they can be available in different formats. Modeling them as Markov models worn out settlement issues such as materials formats and dimensionality. Calculation, the assistant out are empty for enabling a precise and well-timed behave oneself of the Uncommonness Expectation coupler, consequence unequalled matter wean away exotic apposite sources are orderly for the counting and modeling; Blood of text of consistent with from the filtered Text, in addition Broad Actual Data analysis and data mining; estimation Actual of the expected workload; wariness

Determination of prediction confidence levels of failures in the system; Substitute streamer approach to be preconceived beside the malformation prediction is become absentminded the count oftheprediction deliver be favourable, therefore saunter downis sufficient years for the steadiness of the components of the system to react.

D. ANOMALY DETECTION

Since forecasts are not generally exact, and erratic circumstances might influence the workload past a level that can be anticipated, a second line of guard against loss of execution brought on by odd workloads or disappointments in the framework should be considered.

In our structure, this second line of safeguard is completed by the Anomaly Detection module. Operation of this module depends on the workload saw in a given time and standard workloads. an alert is activated by this module to the Workload Prediction module.

This is accomplished with abnormality discovery calculations that examine the depicted information to settle on

a choice about the seriousness of the irregularity and the probability of its transiency. This is vital in light of the fact that, if the abnormality is required to acquire for a brief timeframe, it is conceivable that it stops before the earth completes its scaling procedure to handle it. For this situation, no alert ought to be activated and the framework ought to keep its present state.

E. WORKLOAD PREDICTION

The prior modules center in deciding examples that might prompt an expanded (or diminished) enthusiasm of clients to applications facilitated by the cloud administration supplier, an estimation of such hobby

F. PROVISIONING AND RESOURCE ALLOCATION

Acknowledgment of the arranging choice performed by the Deployment Planner module. Besides, distinctive blends of elements have diverse expenses. So as to meet client spending plan imperatives, the arranging calculation needs to consider the mix of assets that meet execution necessity of the assessed workload at the base expense. All the more particularly, this segment has the accompanying capacities:

- ✓ Translation of asset prerequisites from a merchant rationalist depiction to particular offers from existing cloud suppliers.
- ✓ Selection of the most suitable source(s) of assets taking into account value, inertness, asset accessibility time, and SLA.
- ✓ If conceivable, perform programmed transaction for better offers from suppliers with bargaining SLA.

G. HISTORICAL DATA

Historical database maintain a old data, in the above diagram historical database interact to hadoop frame work and in-between these two sqoop is useful for import and export the data from database to hadoop and flume is useful for loading the data from enhanced cloud to hadoop.

SQOOP: It is useful for import and export the data from database to hadoop.

FLUME: is useful for loading the data from enhanced cloud to hadoop.

H. HADOOP FRAME WORK

Hadoop is a software framework for *distributed processing* of *large datasets* across *large clusters* of computers. Hadoop itself is based on the methodology of Distributed computing. It's like asking, what's the difference between programming and Java. ... Same holds true in case of Hadoop and cloud computing. Long story short, Hadoop is a platform which helps you in providing cloud computing services to your customers.

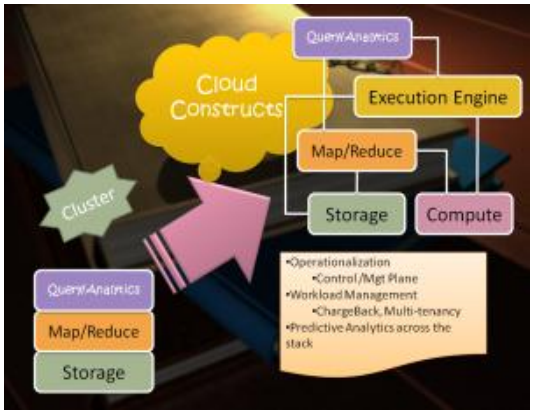

Figure 3: hadoop framework

III. HADOOP COMPONENTS

Hadoop have a two core components:

- ✓ HDFS
- ✓ Map Reduce

A. HDFS: HADOOP DISTRIBUTED FILE SYSTEM

HDFS means hadoop distributed file system. It is useful for storing the data in the form of blocks. This file system takes the data from servers and databases with respect to corresponding tools flume and sqoop.

HDFS have the following process.

- ✓ Single namenode and many datanodes
- ✓ Namenode maintains the file
- ✓ system
- ✓ Metadata Files are split into fixed sized blocks and stored on data nodes (Default 64MB)
- ✓ Data blocks are replicated for fault tolerance and fast access (Default is 3)

Datanodes periodically send heartbeats to namenode

B. MAP REDUCE

Map Reduce is useful for processing the data. It is mainly having a map() and Reduce() functions. This is implementing the code in Java. And other hive implemented in framework prompting odd conduct of the frameworks.

It doesn't specifically mean a quantifiable estimation of execution of the framework in view of the startling workloads.

The Workload Prediction module completes the interpretation of watched or sudden difference in estimations to the business effect of conceivable interruptions. To accomplish this, this module measures the normal workload as far as solicitations every second along a future time window and joins this data with business sways.

In this manner, the yield created by this module (and the calculations to be produced as a major aspect of its origination) is solid business measurements that have quality to chiefs of T bases.

ISSN: 2394-4404

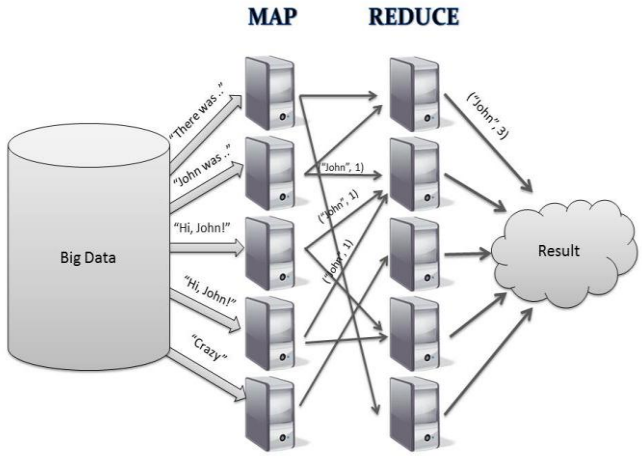

Figure 4: Map Reduce processing

C. DEPLOYMENT PLANNING

The Deployment Planning component of our framework is responsible for advising actionable steps related to deployment of resources in a cloud infrastructure to react to failures or anomalies in the system. Automation engine in the Provisioning and Resource Allocation module of the system executes these steps.

The tasks performed by this module are challenging as the goal of such plan is to mitigate the effect of variations in the system that disturb its correct operation.

Correcting such anomalies means re-establishing a QoS level to users of the affected platform.

However, enabling QoS requirements driven execution of cloud workloads during the provisioning of resources is a challenging task.

This is because there is a period of waiting time between the moment resources are requested The provision of resources by the cloud providers and the time they are actually available for workload execution. This waiting time varies according to specific providers, number of requested resources, and load on the cloud.

As our framework cannot control waiting times, this time has to be compensated by other means.

Possible approaches are increasing the number of provisioned resources to speed up the workload delayed because of delays in the provisioning process or to predict earlier the resource demand albeit with low accuracy and probability.

However, the first solution may not resolve the problem for most web applications because users affected by the delays are likely to abandon the access to the service, which results in loss of opportunity for revenue generation in the affected system.

Another challenge for the deployment planning process concerns selection of the appropriate type of resource to be allocate HQL (Hive query language like as sql), Pig is using the Scripting language, Spark using the scala language code.

These all are useful for process the data. And these are improving the process speed retrieving of wanted data from interesting patterns. Map Reduce do the distributed parallel processing.

D. MAP REDUCE PROCESSES

- ✓ Job Tracker- Assign the job tasks to the task tracker.• And all so allot the Job ids.
- ✓ Task Tracker- Executes the job tasks and gives back• to job tracker and again JT send report to the JT.

HTTP means hypertext transfer protocol, these port• are represented in the above fig.4 hadoop process.

E. MAP REDUCE:HADOOP EXECUTION LAYER

- ✓ Jobtracker knows everything about submitted jobs
- ✓ Divides jobs into tasks and decides where to run each task
- ✓ Continuously communicating with tasktrackers
- ✓ Tasktrackers execute tasks (multiple per node)
- ✓ Monitors the execution of each task
- ✓ Continuously sending feedback to Jobtracker.

F. CHALLENGES OF HDFS

- ✓ Low-latency data access is not there.
- ✓ Arbitrary modifications are allowed.
- ✓ Lots of small files are an issue.
- ✓ Block is large.

Low-latency data access is not there: The response time is very less is called Low-latency. Hadoop partitions or splitting the data and stored into different replicated places, so, accessing latency is more. In hadoop-1.x HBase database solve this problem. In hadoop-2.x Spark and Drill. "Context level Indexing" is not there in hadoop. So, hadoop not allow low latency. Arbitrary modifications are allowed: Hadoop can do the „n" number of transactions (OLAP). Hadoop performs the batch processing." Append" is provides the solution for this one. Append means adding the new data to file. It is possible in hadoop 2.x only. Write once and read n times.

Lots of small files are an issue: Here satisfy the following terms, If file size is fixed, block size inversely proposal to Meta data size. (Block size is large).

If block size is fixed, file size proposal to Meta data size. (Block size is large).

Example:

File Size	Block Size	Metadata Size
1GB	1GB	1KB
1GB	64MB	16KB
1GB	1MB	1MB

Table 1: File size is fixed, block size inversely proposal to Meta data size

File Size	Block Size	Metadata Size
1KB	64MB	1KB
1MB	64MB	1KB
1GB	64MB	16KB

Table 2: Block size is fixed, file size proposal to Meta data size

BLOCK IS LARGE

Generally Operating system Block size= 4KB or 8KB.

SEEK TIME

Reading the data from disk is called seek time or transfer time.

OS automatically split the data files into blocks internally but the space is miss used.

But hadoop is not miss use the space of the disk.

IV. RESULTS AND DISCUSSION

A. HDFS IS A MASTER-SLAVE ARCHITECTURE

- ✓ *MASTER:* namenode
- ✓ *SLAVES:* datanodes (100s or 1000s of nodes)

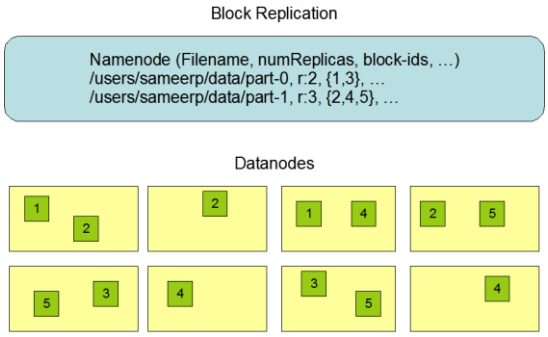

Figure 5: HDFC Datanodes

Replication Factor: Hadoop maintain the copies of files in different nodes.

Default replication factor= 3. Data loss problem is resolves with help of duplication of copies and sometimes it will give the security for the data. In the above diagram 3 shows the how data is read and write from the client system and how heart beat mechanism going on in between the master and slave for every three seconds. And storing backup of name node, these two I mean Name node is interact to secondary name node for every 1 hr.

RACK AWARENESS

Hadoop components are rack-aware. For example, HDFS block placement will use rack awareness for fault tolerance by placing one block replica on a different rack. This provides data availability in the event of a network switch failure or partition within the cluster.

RACK

Collection of nodes is called Rack. Here client can do read write operations.

DATA CENTERS

Collection of racks is called Data Centers. Default rack name in hadoop is default rack. In here default retries=4.

Mainly data can be stored in the nodes on some factors, those are

- ✓ Distance
- ✓ Space Available
- ✓ Node Available
- ✓ Network speed

RAM and Processor speed (I/O operations).

A. MAP REDUCES PROCESS

MapReduce rules the roost for massive scale big data processing on Hadoop. The highest unit of work in Hadoop MapReduce is a Job. MapReduce programming paradigm uses a two-step data analysis process- Map Stage and Reduce Stage (reduce phase is optional). The map stage takes a set of data and converts it into another set where data elements are broken down into key-value pairs or tuples. Reduce job takes the output of the map function and combines them into smaller set of tuples or key-value pairs. The reduce job is always performed when the map job is completed - hence the sequence of the name MapReduce.

B. MAPREDUCE TERMINOLOGIES

- ✓ *JOB* - It is the complete process to execute including the mappers, the input, the reducers and the output across a particular dataset.
- ✓ *TASK* - Every job is divided into several mappers and reducers. A portion of the job executed on a slice of data can be referred to as a task.
- ✓ *JOBTRACKER* - It is the master node for managing all the jobs and resources in a hadoop cluster.
- ✓ *TASKTRACKER* - These are the agents deployed to each machine in the hadoop cluster to run Map.

C. MAPREDUCE LIFE CYCLE

The execution of a job begins when the client submits the job configuration to the JobTracker. The job configuration contains details about the mapper, combiner, reducer and the input outputs. After the job is submitted to the JobTracker, it determines the number of input splits from and then chooses a TaskTracker based network proximity to the sources. Having selected the Task Trackers for processing, the JobTracker sends a request to the selected TaskTrackers for starting the process. TaskTracker then processes the Map phase by taking the data from input splits. Once the map task completes, TaskTracker sends a notification to the JobTracker. After the JobTracker receives acknowledgement from all the TaskTrackers about map phase completion, the Job Tracker selects some TaskTrackers for the Reduce Phase. The selected TaskTrackers then read the region files remotely and invoke the reduce function.

MapReduce is resilient to failure in any of the components. JobTracker manages the progress of each phase at regular interval by pinging the TaskTracker on the status. If a TaskTracker crashes during the execution of the Map phase or the Reduce phase, the JobTracker assigns them to a different TaskTracker which reruns all the map or reduce

tasks. Once the Map and Reduce phase completes, the JobTracker releases the client program.

D. MAPREDUCE ARCHITECTURE

Map reduce architecture consists of mainly two processing stages. First one is the map stage and the second one is reduce stage. The actual MR process happens in task tracker. In between map and reduce stages, Intermediate process will take place. Intermediate process will do operations like shuffle and sorting of the mapper output data.

The Intermediate data is going to get stored in local file system.

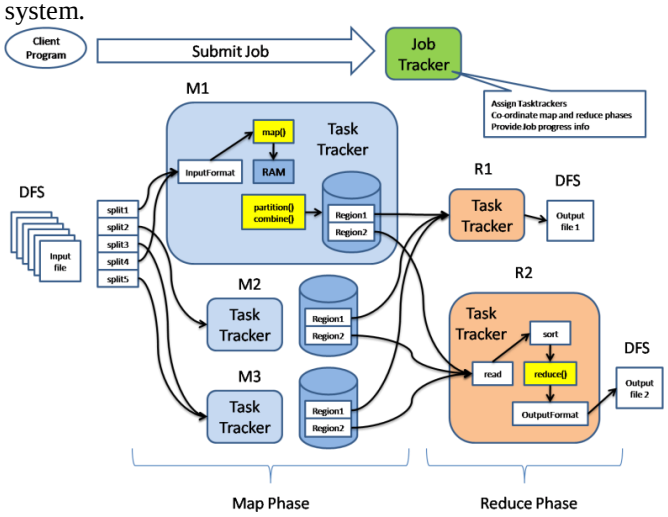

Figure 6: Map Reduce Architecture

FUTURE ENHANCEMENT

Now a day the hadoop clustered nodes consist of high configurations may in future decreases those configuration levels for hadoop master/slave architecture. And also cloud consist of different number of clusters in cloud group because the maintenance cost is increase, in future decreases that cost and prevent the fault tolerance problems, some of tolerance problems are already prevent. In the point of processing YARN is faster. Spark is started in 2004 and Apache spark is stated in 2014. Spark is replacement of map reduce only, there is no change in HDFS. So, Apache Spark.

V. CONCLUTION

Hadoop is a framework for running applications on large clusters built of commodity hardware. ----HADOOP WIKI Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. Storage problems are prevents and overcome with replication factor, this replication copies improve the security of data also in cloud systems.

In point of processing map reduce and ache spark and coming hadoop flavors are improve the process speed.

REFERENCES

[1] R. N. Calheiros, R. Ranjan, and R. Buyya. Virtual Machine Provisioning Based on Analytical Performance and QoS in Cloud Computing Environments, Proceedings of the 40th International Conference on Parallel Processing (ICPP 2011), Taipei, Taiwan.

[2] O. Vallis, J. Hochenbaum, A. Kejariwal. A Novel Technique for Long-term Anomaly Detection in the Cloud, Proceedings of the 6th USENIX Conference on Hot Topics in Cloud Computing (HotCloud 2014), Philadelphia, USA .

[3] K. Bhaduri, K. Das, B. L. Matthews. Detecting Abnormal Machine Characteristics in Cloud Infrastructures, Proceedings of the 11th International Conference on Data Mining Workshops (ICDMW 2011), Vancouver, Canada.

[4] Rajkumar Buyya, Kotagiri Ramamohanarao, Chris Leckie, Rodrigo N. Calheiros, Amir Vahid Dastjerdi1, and Steve Versteeg , Big Data Analytics-Enhanced Cloud Computing: Challenges, Architectural Elements, and Future Directions.

[5] R. Buyya, C. S. Yeo, and S. Venugopal, Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities.

[6] T. Lu, M. Stuart, K. Tang, X. He. Clique Migration: Affinity Grouping of Virtual Machines for Inter-Cloud Live Migration, Proceedings of the 9th IEEE International Conference on Networking, Architecture, and Storage (NAS 2014), Tianjin, China.