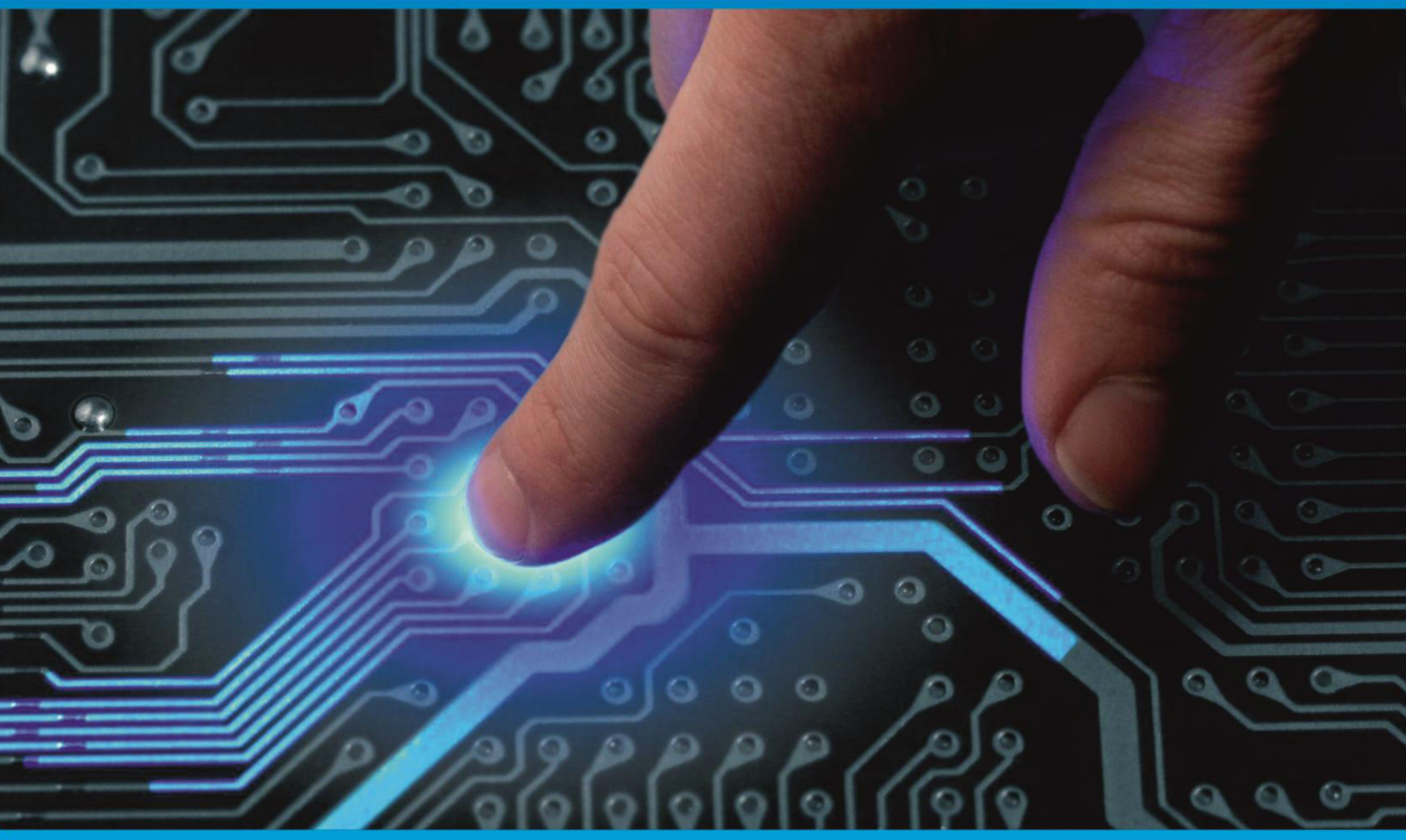
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 4, April 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904168|

Data Analytics as a Service for Business Models using Cloud Computing Environments

Mr. Rahul U. Patil, Mr. Atul U. Patil, Mr.A.P.Pande

M.Tech, Dept. of CSE, BVIT Navi Mumbai, India

M.E, Dept. of CSE, PVPIT Budhgaon Sangli, India

M. E, Dept. of Electronics, PVPIT Budhgaon Sangli, India

ABSTRACT: This paper proposes a model of Data analytics service forvarious IT Organizations and business models using cloud computing service environment. As millions of gigabytes of data is produced daily by each of organizations managing analysing such huge data is very tedious tasks for all organizations and some conventional data analysis services available but with lots of drawbacks or lack of full proof services. Cloud computing service having services that mixes On-Demand provisioning of resource nodes with improved utilization, withopportunistic provisioning of cycles from unutilized cloud nodes to other different processes. As data analytics is big vast area and it require lots of attention also this particular area of information technology world requires lots of computing power which now a days purchasing and establishing a such huge infrastructure in not possible for lots of IT organizations or businesses we have proposed such Analytics service for data analysis in cloud computing environment which encompasses and proposes a model which is not only cost effective but also secured in terms of provisioning data analysis service along with high end performance with Big data processing capabilities.

KEYWORDS: Data Analytics as a service,CPU utilization, Storage utilization, Map-Reduce, Splitting algorithm, Big Data, Cloud Technology.

I. INTRODUCTION

Big Data get generated to an enormous volume of organized, unstructured, semi-organized, or blended sorts ofdata which is hard to manage and process and analyse with conventional database technique. Big data analytics or BigData Hadoop technology havingimportant role in analysing cleaning and processing such large amounts of data but such technology also and some kind of advantages and disadvantages associated with it just like Hadoop requires huge amounts of commodity hardware and variety of servers to process ,tackle and manage that data. So for the Mid level to small scale organizations its very Costly work to set up an Infrastructure which can handle all these requirements. So one of the solutions is provided with Cloud computing technology along with Big data analytics technology which have a mechanism where special analytic procedures work on big data. Steps to be followed in Big Data Hadoop are Acquiring, Recording, Extracting (data-cleaning), i.e Extracting, Cleaning and transforming of data for the purpose of data analytics.

The process of data analysis also known as Analytic Workflow, broadly involves a chain of data cleansing and integration tasks. In light of the scientific categorization and an investigation of the existing analytical software's and frameworks, we introduced the architecture of Data Analytics-as-a-Service . Data Analytics-as-a-Service speaks to the way to deal with an extendible stage that can give cloud-based Infrastructure and services over an assortment of businesses and use cases .Through a practical point of view, the platform includes end to end abilities of an analytic result, from data acquiring to end-client perception, revealing and association. Over and above this conventional usefulness, it grows the run of the mill procedure with imaginative thoughts, as Analytic. Talking about the Cloud Computing working environments and services which helps us to do these analytics activities are as follows.

Cloud computing is considered as a highlyfuture coming new geek for servicing computing needs as a utility. In cloud computing variety of cloud users need various kinds of services as per demands for the completing different tasks. So it is the job of cloud computing to avail all the demanded services to the cloud consumers. But due to the availability of finite resources it is very difficult for cloud providers to provide all the demanded services. From the cloud providers' perspective cloud resources allocation done with some algorithm with utilization of resources is first priority. So, that is critical issue to meet cloud clients' Agreed terms and conditions. To ensure need based

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904168|

availability of computing resources a cloud service provider requires to overprovision: store a huge proportion of servers idle so they get utilized to satisfy an demand based request, which may arrive at any instant of time. If all servers get not utilized its goes in underutilization. The way to handle such condition means subsequently disallowing huge sum of requests to a point at which a provider no longer provides demand based computing [2].

Different set of opportunities are available in field of cloud computing, which arevarious services are get provided over internet and use of computer technology. Like Software as a service in which computing tasks, are transforming data centres into pools of computing service on a huge scale on cloud servers. So, the huge network bandwidth and reliable network service makes it possible that users can avail to reliable cloud services from remote where all resources kept in cloud and data centres. In recent IaaS acquired so much market which replacesing the traditional ways to maintain physical infrastructure for computing and also reducing costs associated. A advantage of this service cloud is servicing clients demand based provisioning to cloud resources. But to do this activity cloud service provider either keep ideal lots of computing power or deny service requests of user requests (in this case service provider fail to achieve SLA). Also lots of user's doesn't require demand based service but full flagged service which consideration is also must take in to consideration. Many applications and workflows are designed for recoverable systems where interruptions in service are expected.

So model is proposed, a cloud service with Big Data based service configuration that combines demand based provisioning of computing power or services with scheduling based provisioningof cloud machines to other clients. The objective is to handle larger data in less amount of time and keep utilization of all idle cloud nodes through splitting of larger files into smaller one using Map-Reducing algorithm, also increase the CPU utilization and storage utilization for uploading files and downloading files. To keep data and services trustworthy, security is also maintain using RSA algorithm which is widely used for secure data transmission.

II. RELATED WORK

There is much research work in the field of Big Data Hadoop over the past decades. Some of the work done has been discussed, this paper researched blend of architecture and its safety, proposed a new Data Ananytics as service architecture which encomposses a various services and infrastructure requires to manage process and analyse data getting produced in Various businesses , SaaS model was used to deployed the related software on the cloud platform, so that the resource utilization and computing of scientific tasks quality will be improved [2].Cloud servicewhich serves demand based resources to underlying users and make provisioning of resouces with the help of efficient scheduling algorithm from cloud machines to users by deploying backfill virtual machines from store.

III. THE PROPOSED SYSTEM

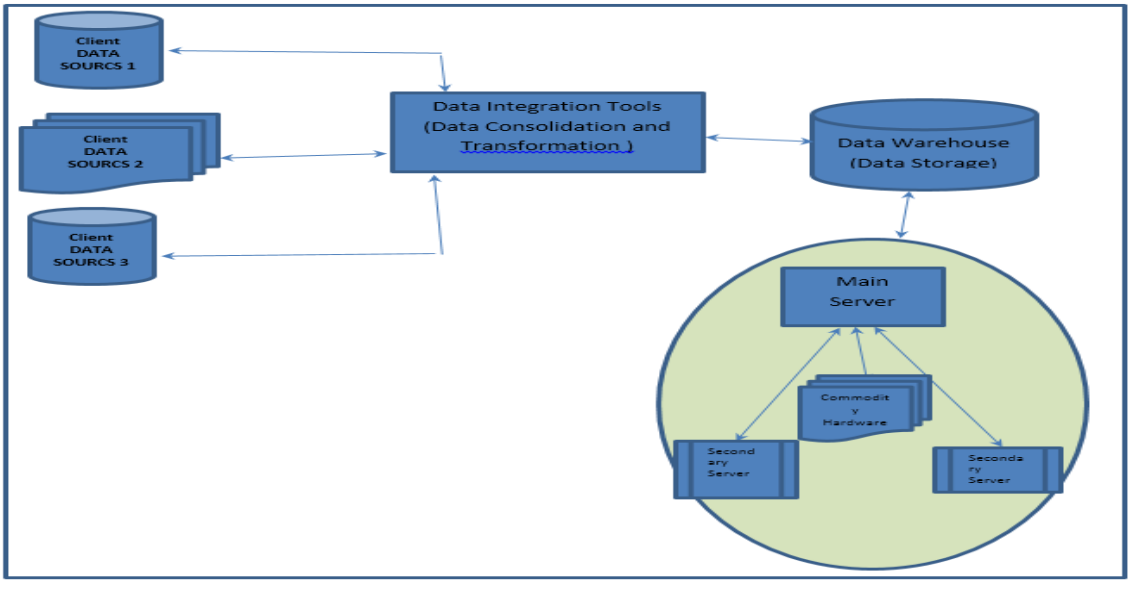


Fig 01. Data Analytics Service Cloud Based Architecture

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904168|

Cloud services becomesmost, prominent solution for tasks and handling and data processing, also storage and distribution, but moving huge amount of data in and out for computing makes an tremendous challenge . Cloud services are successful model of services oriented computing and has many future opportunities to clients with its various services like explained below.

Three most popular cloud paradigms include:

1. Platform as a Service
2. Infrastructure as a Service
3. Software as a Service

The model also be expanded to database systems as a Service or Storage as a Service. ExpandedDBMS both for uploading intensive application data loads, as well as decision support systems are critical parts of the cloud infrastructure. For thissystem include distributed DBMS for updating intensive workload data and parallel DBMS for analysis workload handling. Change in toaccess data patterns of application and the need to scale out to thousands of commodity machines led to the birth of a new class of systems referred to as Key-Value stores.

Our Proposed system encompasses various clients which are required to process and analyse their historical data over the period of time for the business growth so such business models incorporate their databases with our cloud based service which having data warehouse where data get first normalized and stored as data get generated at huge direct data processing is not possible also number of clients are more so data get stored in Data Warehouse. After this process from the accounts of clients they can do data analysis tasks. So client can choose data set to get processed and Primary server where Map - Reduce Programming model of Hadoop Implemented which takes data and Divides it in equal tasks and assign then to number of secondary servers which then processes i.e reduces the tasks and whichever gives out partial results of Main server and then with help of communications lines and Data Warehouse results provided back to client accounts. In the same way data security is also provided by Secondary Servers which Encrypts data and again stores back data in Data Warehouse More detailed explanations of various models are provided further.

In the domain of data analysis, we propose a blend of the Hadoop Map Reduce paradigm and its open-source implementation Hadoop, in terms of usability and performance along with the Cloud computing various services to encounter the tasks of data Analytics. The algorithm has six modules:

1. Cloud Servers Setup (Hadoop based Infra)
2. Client registration and Login facility
3. Cloud Service Provider
4. Data Analytics Service
5. Encryption/Decryption of Data for security
6. Administration of client files(Third Party Auditor)

3.1 Cloud Servers Setup (Hadoop ManagementInfrastructure)

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures [6]. Hadoop implements Map Reduce, using the Hadoop Distributed File System (HDFS).The HDFS allows users to have a single addressable namespace, spread across many hundreds or thousands of servers, creating a single large file system. Hadoop has been demonstrated on clusters with 2000 nodes. The current design target is 10,000 node clusters.

Hadoop was inspired by MapReduce, a framework in which an application is broken down into numerous small parts. Any of these parts (also called fragments or blocks) can be run on any node in the cluster. The current Apache Hadoop ecosystem consists of the Hadoop kernel, MapReduce, the Hadoop distributed file system (HDFS).

JobTracker is the daemon service for submitting and tracking MapReduce jobs in Hadoop. There is only One Job Tracker process run on any hadoop cluster. Job Tracker runs on its own JVM process. In a typical production cluster itruns on a separate machine. Each slave node is configured with job tracker node location. The JobTracker is a single point of failure for the Hadoop MapReduce service. If it goes down, all running jobs are halted. JobTracker in Hadoop

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904168|

performs, Client applications submit jobs to the Job tracker. The JobTracker talks to the NameNode to determine the location of the data The JobTracker locates TaskTracker nodes with available slots at or near the data The JobTracker submits the work to the chosen TaskTracker nodes. The TaskTracker nodes are monitored. If they do not submit heartbeat signals often enough, they are deemed to have failed and the work is scheduled on a different TaskTracker. A TaskTracker will notify the JobTracker when a task fails. The JobTracker decides what to do then: it may resubmit the job elsewhere, it may mark that specific record as something to avoid, and it may even blacklist the TaskTracker as unreliable. When the work is completed, the JobTracker updates its status [9].

A TaskTracker is a slave node daemon in the cluster that accepts tasks (Map, Reduce and Shuffle operations) from a JobTracker. There is only One Task Tracker process run on any hadoop slave node. Task Tracker runs on its own JVM process. Every TaskTracker is configured with a set of slots, these indicate the number of tasks that it can accept. The TaskTracker starts a separate JVM process to do the actual work (called as Task Instance) this is to ensure that process failure does not take down the task tracker. The TaskTracker monitors these task instances, capturing the output and exit codes. When the Task instances finish, successfully or not, the task tracker notifies the JobTracker. The TaskTrackers also send out heartbeat messages to the JobTracker, usually every few minutes, to reassure the JobTracker that it is still alive. These messages also inform the JobTracker of the number of available slots, so the JobTracker can stay up to date with where in the cluster work can be delegated [9].

Namenode stores the entire system namespace. Information like last modified time, created time, file size, owner, permissions etc. are stored in Namenode. The fsimage on the name node is in a binary format. Use the "Offline Image Viewer" to dump the fsimage in a human-readable format. When the number of files is huge, a single Namenode will not be able to keep all the metadata. In fact that is one of the limitations of HDFS [9].

The current Apache Hadoop ecosystem consists of the Hadoop kernel, MapReduce, the Hadoop distributed file system (HDFS).

The Hadoop Distributed File System (HDFS)

HDFS is a fault tolerant and self-healing distributed file system designed to turn a cluster of industry standard servers into a massively scalable pool of storage. Developed specifically for large-scale data processing workloads where scalability, flexibility and throughput are critical, HDFS accepts data in any format regardless of schema, optimizes for high bandwidth streaming, and scales to proven deployments of 100PB and beyond [8].

Key HDFS Features:

- Scale-Out Architecture - Add servers to increase capacity
- High Availability - Serve mission-critical workflows and applications
- Fault Tolerance - Automatically and seamlessly recover from failures
- Flexible Access – Multiple and open frameworks for serialization and file system mounts
- Load Balancing - Place data intelligently for maximum efficiency and utilization
- Tunable Replication - Multiple copies of each file provide data protection and computational performance
- Security - POSIX-based file permissions for users and groups with optional LDAP integration [8].

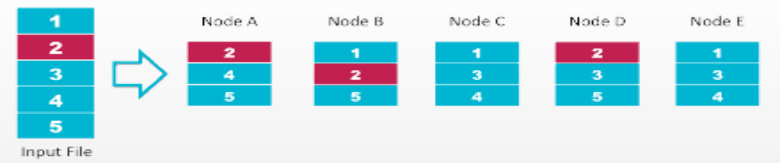


Fig.2 hdfs data distribution [8]

Data in HDFS is replicated across multiple nodes for compute performance and data protection.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904168|

3.2 Client registration and Login facility

It provides an Interface to Login. Client can upload the file and download file from cloud and get the detailed summary of his account. Generally In Various IT Organizations Data gets generated at tremendous rate so as per the business model requirements data is stored in Data Warehouse before processing it. For that purpose various cloud based login accounts to each organization are provided for the purpose of choosing different analytics services.

3.3 Cloud Service Provider(Administrator)

Administration of User and Data.Authority to Add/Remove user. Generally this cloud service provider can provide different facilities to these business models or It organizations like providing SaaS, PaaS, IaaS services and different analytics services.

3.4 Data Analytics Service

Map-Reduce is a programming model and an associated implementation for processing and generating large datasets that is amenable to a broad variety of real-world tasks. Users specify the computation in terms of a map and a reduce function also Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key. Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system [7]. MapReduce is a massively scalable, parallel processing framework that works in tandem with HDFS. With MapReduce and Hadoop, compute is executed at the location of the data, rather than moving data to the compute location; data storage and computation coexist on the same physical nodes in the cluster. MapReduce processes exceedingly large amounts of data without being affected by traditional bottlenecks like network bandwidth by taking advantage of this data proximity [8].

Our implementation of File Splitting Map-Reduce Algorithm runs on a large cluster of commodity machines and is highly scalable. Map-Reduce is Popularized by open-source Hadoop projects. Our File Splitting Map-Reduce algorithm works on processing of large files by dividing them on a number of chunks and assigning the tasks to the cluster nodes in hadoop multi node configuration. In these ways our proposed File Splitting Map-Reduce algorithm improves the Utilization of the Cluster nodes in terms of Time, CPU, and storage.

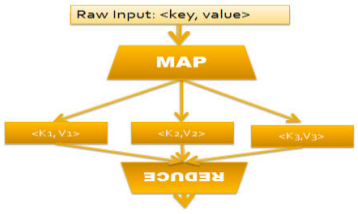


Fig.3 programming framework

Applying a map operation to each logical ‘record’ in our input in order to compute a set of intermediate key/value pairs, and then applying a reduce operation to all the values that shared the same key, in order to combine the derived data appropriately. Our use of a programming model with user specified map and reduce operations allows us to parallelize large computations easily [7]. It enables parallelization and distribution of large scale computations, combined with an implementation of this interface that achieves high performance on large clusters of commodity PCs.

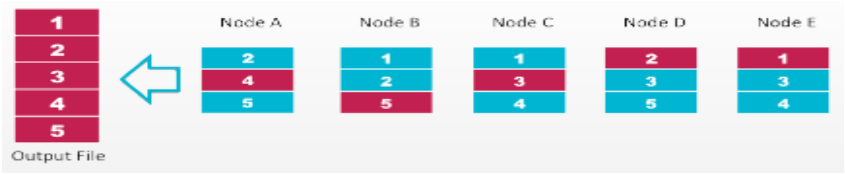


Fig.4 Map-Reduce compute distribution [8]

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904168|

3.4.1 Programming Model

File Splitting Map-Reduce Algorithm-

In this scenario clients are going to upload or download files from the main server where the file splitting map-reduce algorithm is going to execute. On the main server the mapper function will provide the list of available cluster I/P addresses to which tasks are assigned so that the task of file splitting gets assigned to each live cluster. File splitting map-reduce algorithm splits file according to size and the available cluster nodes.

The computation takes a set of input key/value pairs, and produces a set of output key/value pairs. The user of Map-Reduce library expresses the computation as two functions: Map and Reduce [7].

Map, Written by user, takes an input pair and produces a set of intermediate key/value pairs. The Map-Reduce library groups together all intermediate values associated with the same intermediate key and passes them to the Reduce function[7].

The Reduce function, also written by the user, accepts the intermediate key and a set of values for that key. It merges together to these values to form a possibly smaller set of values. Typically just zero or one output value is produced per Reduce invocation. The intermediate values are supplied to the user's reduce function via an iterator. This allows us to handle lists of values that are too large to fit in memory [7].

The map and reduce functions supplied by the user have associated types:

Map (k1,v1) ✇list (k2,v2)

Reduce (k2, list(v2)) ✇list (v2)

It means the input keys and values are drawn from a different domain than the output keys and values. Furthermore, the intermediate keys and values are from the same domain as the output keys and values [7].

This process is automatic Parallelization. Depending on the size of RAW INPUT DATA ✇instantiate multiple MAP tasks. Similarly, depending upon the number of intermediate <key, value> partitions ✇instantiate multiple REDUCE tasks. Map-Reduce data-parallel programming model hides complexity of distribution and fault tolerance.

3.5 Encryption/decryption for data security

Data security is most concerned for all the big IT Organizations and all business groups as most sensitive data gets exchanged through different servers through the internet. There are high chances of data compromise and misuse. For that purpose we have proposed various encryption decryption algorithms for the data security at server ends. And there are various latest encryption decryption algorithms available to use which provide high level data security but for example we are using here RSA algorithm as information provided further. In this, files get encrypted/decrypted by using the RSA encryption/decryption algorithm. RSA encryption/decryption algorithm uses public key & private key for the encryption and decryption of data. Clients upload the file along with some secret/public key so a private key is generated & the file gets encrypted. At the reverse process by using the public key/private key pair file get decrypted and downloaded.

3.6 Administration of client files(Third Party Auditor)

This module provides facility for auditing all client files, As Various activities are done by Client. Files Log records and got created and Stored on Main Server. For each registered client Log record is created which records the various activities like which operations (upload/download) performed by client.Also Log records keep track of time and date at which various activities are carried out by clients. For the safety and security of the Client data and also for the auditing purposes the Log records helps. Also for the Administrator Log record facility is provided which records the Log information of all the registered clients. So that Administrator can control over all the data stored on Cloud servers. Administrators can see Client wise Log records which helps us to detect the fraud data access if any fake user tries to access the data stored on Cloud servers.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904168|

IV. RESULTS

Our results of the project will be explained well with the help of project work done on number of clients and one main server and then three to five secondary servers so then we have get these results bases on three parameters taken into consideration like

1) Time
2) CPU Utilization
3) Storage Utilization.

Our evaluation examines the improved utilization of Cluster nodes i.e. Secondary servers by uploading and downloading files for Conventional Data Analytics service versus Cloud Based Data Analytics Service from three perspectives. First is improved time utilization and second is improved CPU utilization. The storage utilization also improved tremendously.

4.1 Results for time utilization

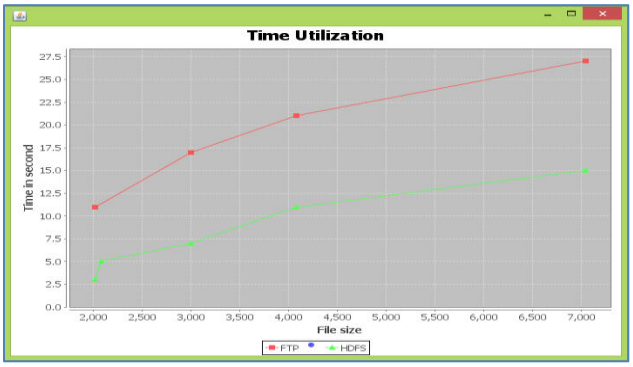


Fig.5 time utilization graph for uploading files

Fig. 5 shows time utilization for Conventional Data Analytics service versus Cloud Based Data Analytics Service for Uploading data for Data Analytics. These are:

Uploading File Size(in Gb)	**Time (in sec) for** Conventional Data Analytics service	**Time (in sec) for** Cloud Based Data Analytics Service
2	10	2.5
3	17.5	7.5
4.2	20	10
7	27	12.5

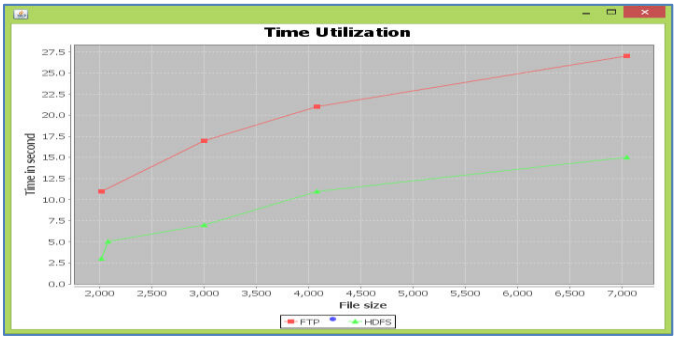


Fig.6 time utilization graph for downloading Analysed data files

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904168|

Fig. 6 shows time utilization of Conventional Data Analytics service versus Cloud Based Data Analytics Service for Uploading data and Data Analysis. These are:

Downloading File Size(in Gb)	**Time (in sec) for** Conventional Data Analytics service	**Time (in sec) for** Cloud Based Data Analytics Service
2	10	2.5
3	17.5	7.5
4.2	20	10
7	27	12.5

4.2 Results for CPU utilization

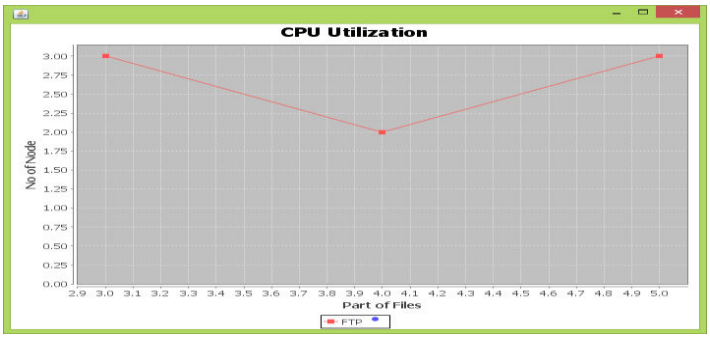


Fig.7 cpu utilization graph for Conventional Data Analytics service

Fig.7 describes the CPU utilization for Conventional Data Analytics service versus Cloud Based Data Analytics Service for Uploading data for Data Analytics.

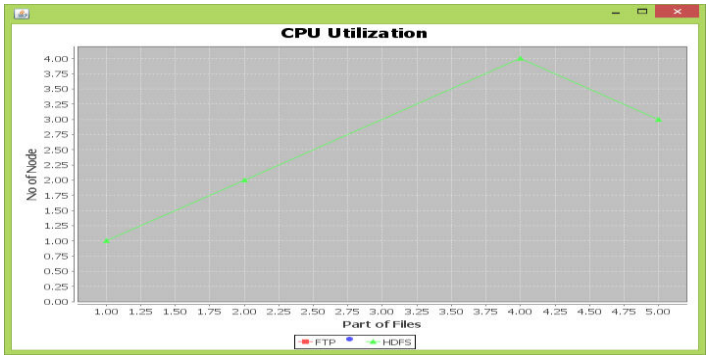


Fig.8 Describes CPU utilization graph on Cloud Based Data Analytics Service on number of Cluster nodes.

V. CONCLUSION

We have proposed improved cloud infrastructure that combines On-Demand allocation of resources with improved utilization, opportunistic provisioning of cycles from idle cloud nodes to other processes. A cloud infrastructure using Hadoop configuration with improved CPU utilization and storage utilization is proposed using File splitting Map-Reduce Algorithm. Hence all cloud nodes which remain idle are all utilized and also improve in security challenges and achieve load balancing and fast processing of large data in less amount of time. We compare the Conventional Data Analytics serviceand Cloud Based Data Analytics service for data analysis and enhance the CPU utilization and storage utilization.

Till now in many proposed works there is no cloud based data analytics service using BigData Hadoop.But we have proposed and implemented Cloud based data analytics service with Bigdata Hadoop with good results.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904168|

We evaluate the backfill solution using an on-demand user workload on cloud structure using hadoop. We contribute to an increase of the CPU utilization and time utilization between Conventional service and cloud based service. In our work also all cloud nodes are fully utilized , no any cloud remains idle, also data analysis gets at a faster rate so that tasks get processed at less amount of time which is also a big advantage hence improving utilization of underlying infrastructure along with providing full proof security .

REFERENCES

[1]. Cloud Computing Resources Utilization and Cost Optimization for Processing Cloud Assets. 2020 IEEE International Conference on Smart Cloud (SmartCloud) 978-1-7281-6547-9/20/$31.00 ©2020 IEEE DOI 10.1109/SmartCloud49737.2020.00017

[2]. A Systematic Review of the Security in Cloud Computing: Data Integrity, Confidentiality and Availability. 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON) Galgotias University, Greater Noida, UP, India. Oct 2-4, 2020

[3]. Reviewing Cloud Monitoring: Towards Cloud Resource Profiling 2018 IEEE 11th International Conference on Cloud Computing 2018

[4]. IaaS Reactive Autoscaling Performance Challenges 2018 IEEE 11th International Conference on Cloud Computing

[5]. Block Design-based Key Agreement for Group Data Sharing in Cloud Computing

[6]. Naor, M., Rothblum, G.N.: The complexity of online memory checking. In: Proc.
of FOCS 2020, pp. 573–584 (2020) Jian Shen, Member, IEEE, Tianqi Zhou, Debiao He, Yuexin Zhang, Xingming Sun, Senior Member, IEEE, and Yang Xiang, Senior Member, IEEE

[7]. Low-Information-Loss Anonymization of Trajectory Data Considering Map Information 2020 IEEE Explore

[8]. Data Anonymization: K-anonymity Sensitivity Analysis Wilson Santosa,b , Gonçalo Sousaa , Paula Prata a,b, Maria Eugénia Ferrãoa,c a Universidade da Beira Interior, Covilhã, Portugal b Instituto de Telecomunicações c Centro de Matemática Aplicada à Previsão e Decisão Económica, Lisboa, Portugal. 2020 15th Iberian Conference on Information Systems and Technologies (CISTI) 24 – 27 June 2020, Seville, Spain ISBN: 978-989-54659-0-3

. 1, pp.107-113, 2008 January.

[9]. Modelling and Prediction of Resource Utilization of Hadoop Clusters: A Machine Learning Approach Hassan Tariq School of Engineering and Computer Science, Victoria University of Wellington, Wellington 6140, New Zealand Department of Computer Science, University of Agriculture, Faisalabad, Pakistan https://dl.**acm.org**/doi/10.1145/3344341.3368821 UCC '19, December 2–5, 2019, Auckland, New Zealand.

[10]. A Multi-Optimization Technique for Improvement of Hadoop Performance with a Dynamic Job Execution Method Based on Artificial Neural Network Rayan Alanazi1 · Fawaz Alhazmi2 · Haejin Chung3 · Yunmook Nah4 Received: 12 July 2019 / Accepted: 24 April 2020 © Springer Nature Singapore Pte Ltd 2020

[11] J. Dean et al., "MapReduce: Simplified Data Processing on Large Clusters", In CACM, Jan 2008.

[12] J. Dean et al., **"** MapReduce: a flexible data processing tool", In CACM, Jan 2010.

[13] M. Stonebraker et al., "MapReduce and parallel DBMSs: friends or foes?", In CACM. Jan 2010.

[14] A.Pavlo et al., "A comparison of approaches to large-scale data analysis", In SIGMOD 2009.

[15] A. Abouzeid et al., "HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads", In VLDB 2009.

[16] F. N. Afrati et al., **"**Optimizing joins in a map-reduce environment",In EDBT 2010.

[17] P. Agrawal et al., "Asynchronous view maintenance for VLSD databases", In SIGMOD 2009.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.488

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details