ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

Big Data Implementation Using Hadoop and Grid Computing

Yuvraj S. Sase , Pratik A.Yadav

ME, Vishwabharti Academy's College of Engineering, India.

ME, Vishwabharti Academy's College of Engineering, India

Abstract: Big data is business transformation. So every organization is trying to analyze their big data. Big data poses implementation problems in extreme conditions. This paper focuses on some methods to use grid computing along with hadoop. Grid computing provide large storage capability and computation power. This paper listed some open source toolkit use to implement solution such as Hadoop, Globus Toolkit.

Keywords: grid computing, hadoop, storage capacity, processing capacity

I.INTRODUCTION

Big Data is energy source of present world. It refashions future of Global Economics. Big data revolution changes way of thinking in business[5]. It was waste material in past and now it helps organization to progress in great extent. Big data affect decision making from the bottom-up and the top-down. Big Data speed up discoveries and small predictions in daily activities.Data with great volume that is unstructuredand having great velocity are main identities of big data. Oracle add the fourth, value[4]. For getting valuable data from big data, system analyzes large volume of irrelevant data.

Hadoop, an open source tool, can be used to analyze big data.Hadoop consist HDFS and map-reduce functionalities. Hadoop also provides reliability by replicating data over multiple nodes. Map function analyze data and produce result for each entity and reduce function reduces number of results[7]. Big data analysis call for large storage capacity and great processing power which can be satisfied by grid computing[2].Computing tools like globus toolkit is available for grid computing. Globus toolkithas gram service and job manager respectively to control job execution and scheduling best node for execution. Condor, fork, pbs are examples of Job managers in globus.

Integration of Hadoop and Globus Toolkit can provide solution on storage and analysis of big data.

II. HADOOP ARCHITECTURE

Map reduce and hadoop distributed file system (HDFS) services are provided by Hadoop. In HDFS file system, one NameNode and multiple DataNodes are present.Map reduce engine have job tracker present onNameNode and task trackers on DataNodes. Job tracker is responsible for job submission on DataNodes and DataNode actually executejobs[7].

A.Namenode

Namenode is present on master system of hadoop. Files are divided into number of data blocks and this data blocks are saved on DataNode. For security and reliability this datablocks are replicated on different DataNodes.All records are maintained by NameNode. NameNode contains metadata about all datablocks stored on datanodes. Any client application which require data, contacts to NameNode. Namenode provide information about location of data on DataNodes. Secondary NameNode is maintained to avoid problem which can occur due to failure of primary NameNode.

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

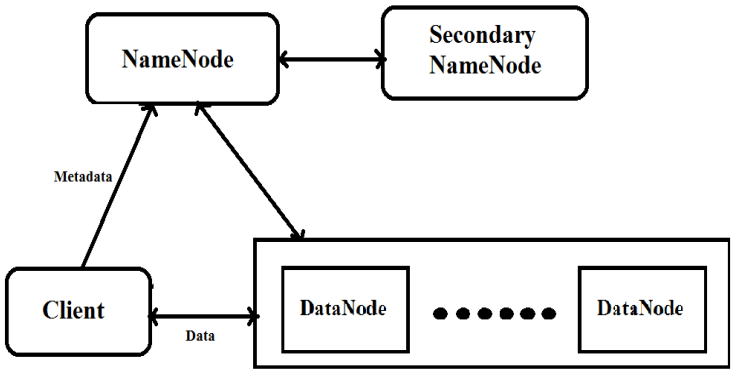

Fig. 1 Architecture of NameNode

B. DataNode

DataNodes store application data in the data bock format. When DataNode starts, first it does handshake with NameNode. Handshake process registers DataNode to NameNode.DataNode registered with unique storage id. DataNode identifies replicas present on it and send block report to NameNode. NameNode stores metadata of datablocks present on that DataNodefrom block reports.DataNode have task trackers which executes jobs on DataNode.

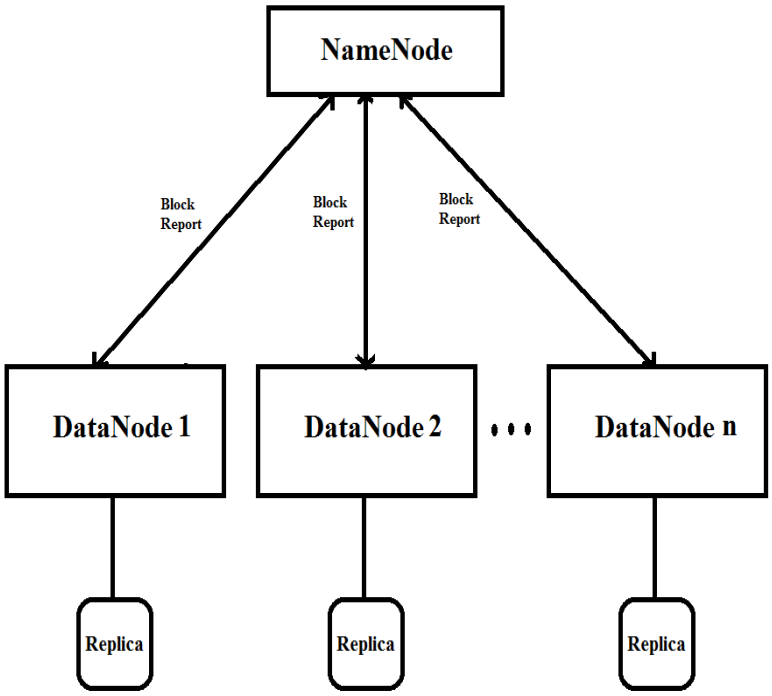

Fig.2 Architecture of Datanode.

C. Job tracker and Task Tracker

Hadoop has mapreduce engine which consist of job tracker and task trackers. Main function of job tracker and task tracker is to execute map and reduce functionalities on different nodes. Job is submitted to node on which job-tracker runs. Job tracker has information like on which node data blocks are present and it send jobs for execution on

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

respective task trackers. Job tracker always tries to execute job on nodes which are near to data. Task trackers execute jobs sent by job tracker. Job tracker checks status of task tracker after every few minutes.

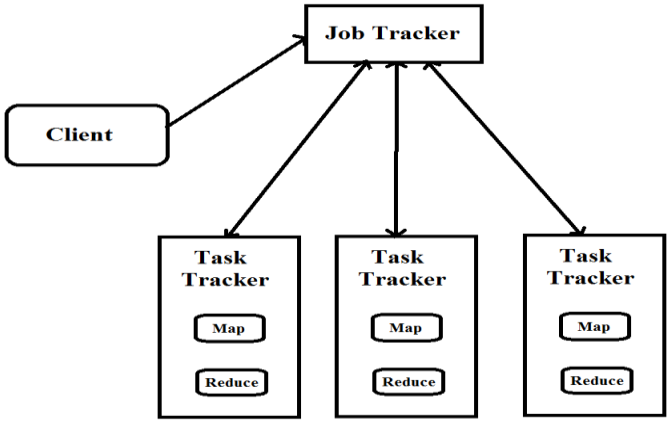

Fig. 3 Architecture of Job tracker and task tracker

III. PROBLEMS IN HADOOP

Hadoop provide different nodes and services for storing and analyzing big data. But in extreme conditionshadoop can experience problems. Such problems affect overall performance of Hadoop and can be listed as below

Problem 1:NameNode failure due to limited capacity to store the metadata

Problem 2 :Searching the huge metadata at NameNode may go inefficient due to its size

Problem 3: No inbuilt facility for parallelization of jobs submitted to tasktrackers

IV. GRID COMPUTING ARCHITECTURE

Grid computing is collection of computers connected together using network and sharing resources with each other. Grid Computing enables to use underutilized resources on network to increase computational power and decrease time requirement for job execution. Grid computing integrates different computer recourses such as storage, processor. This enables execution of different application which is impossible on one system. Difference between Grid and distributed computing is jobs executing on grid are non interactive. Jobs are executing in parallel way [8]. Grid computing creates a virtual computer whose hardware and software resources are scattered. Grid computing can contain geographically dispersed computer nodes. Grid computing consists of one control node which controls all other nodes. Control node is responsible for resource sharing and job execute on different number of nodes.

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

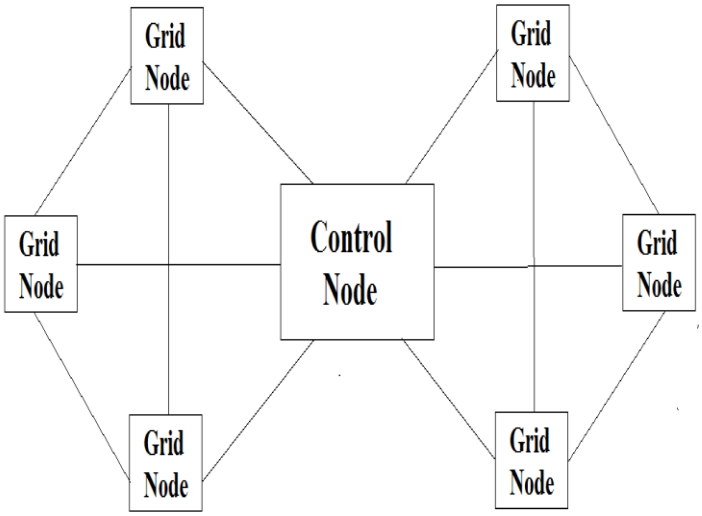

Fig. 4 Grid Computing Architecture

As shown in figure, Grid computing has one control node and all others are grid nodes connected to each other. User can submit job from any node and control node decides node on which job is going to execute. User feels single system when user access grid network. Open source tools such as Globus Toolkit can be use to create computer grid network. Globus toolkit has GRAM, GSI, GridFTP components to provide all services required for grid computing. GRAM service controls all job execution tasks. GRAM service uses job manager such as condor to locate job executing node GSI handles security and GridFTP is to transfer data between different nodes.

V. SOLUTION FORHADOOP'S PROBLEMS USING GRID

Section III defines some problems in hadoop during big data analysis. It is conclude that Hadoop face problems in storage and computation power. These are points where grid is strong. So that combination of Hadoop and grid can provide solution for these problems[1]. Architecture for Hadoop and grid is,

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

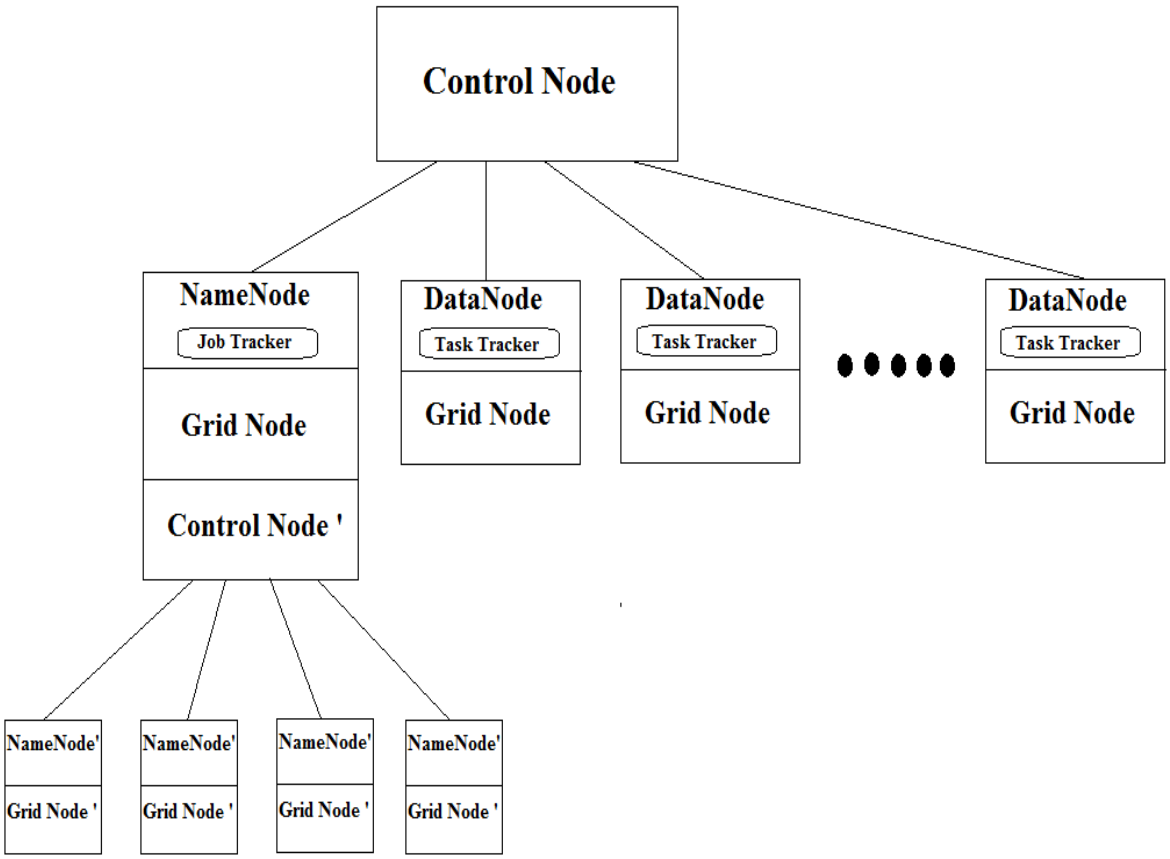

Fig. 5 Combinearchitecture of Hadoop and Grid

Above architecture shows grid and hadoop node. First grid node act as NameNode of Hadoop and other grid nodes are DataNodes. One more sub grid is form with NameNode as control node and different number of grid nodes. Solution for first problem is solved by this sub grid. Each grid node of sub grid shares storage resources. Therefore NameNode can use this storage memory to store metadata. All storage memory treated as if it is of only one system. User is unaware about grid storage. Whole sub grid is treated as one Name Node. So problem of storage is removed by combination of Hadoop and grid.Second problem of computation power is also solved by sub grid. Job of searching metadata of particular data block can be parallelized on number of grid nodes which minimizes time require to searching in metadata. Sub grid increases computation power of NameNode. Client application also can efficiently access NameNode.

Third problem is parallelizing map reduce jobs on DataNodes. Hadoopdoes notparallelize map reduce problem. But grid can parallelize it. If one job has to process more than one data block on one DataNode, then it can be parallelized by transferring next data blocks to any idle systems and execute same job on that system. So that number of data blocks processed parallel depend on number of idle systems. This transfer and execution of job cannot be done by only Hadoop. For this gridhelps hadoop to parallelize jobs. Grid handles transfer of data between DataNodes and execution of job. So third problem also solved by integration of grid and hadoop system.

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

VI. CONCLUSION

Hadoop analyze big data but faces some problem of storage and computation power whenever NameNode is overloaded. Grid helps hadoop to overcome these problems. So combination of Hadoop and grid for implement big data can be used.

REFERENCES

[1] Garlasu,D.Sandulescu,V.; Halcu,I.;Neculoiu,G.; Grigoriu,O.; Marinescu,M.; Marinescu, V., "A big data implementation based on Grid computing" , Roedunet International Conference (RoEduNet), 2013 11th
[2] "A big data implementation using grid computing" ,www.pantechproed.com/download_project.php
[3]C.Chandhini, MeganaL.P , "Grid Computing-A Next Level Challenge with Big Data" , International Journal of Scientific & Engineering Research Volume 4, Issue3, March-2013 1ISSN 2229-5518
[4]An Oracle White Paper March 2013 , "Big Data analytics : Advanced analytics in oracle database"
[5] "IDEAS ECONOMY: FINDING VALUE IN BIG DATA", www.oracle.com/us/technologies/big-data/index.html
[6] Ann Chervenak; Ian Foster; Carl Kesselman; Charles Salisbury; Steven Tuecke, "The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientifc Datasets"
[7] "Apache Hadoop" http://en.wikipedia.org/wiki/Apache_Hadoop
[8] "Grid Comuting" http://en.wikipedia.org/wiki/Grid_computing