2015 International Conference on Soft Computing Techniques and Implementations- (ICSCTI)
Department of ECE, FET, MRIU, Faridabad, India, Oct 8-10, 2015

Data Integration of Cloud-based and Relational Databases

Shweta Malhotra, M.N Doja, Bashir Alam, Mansaf Alam
Department of Computer Engineering, Department of Computer Science
Jamia Millia Islamia University
New Delhi, India

Abstract- Today managing data is very critical for all kind of users. Some users use Cloud Database Services for managing their data but at the same time for the security reasons they want to keep some crucial private data at their own end so they face problem like how to integrate both kind of data one located at Local RDBMS and the other at Cloud repositories. In this paper we have analyzed that Hadoop can be used to solve such kind of problems. Hadoop platform is used to integrate data one is located at the Cloud databases and the other located at the Local RDBMS. Hadoop components not only process huge amount of data but with the help of inbuilt security mechanisms one can also securely store their data.

Keywords: Cloud Databas, Sqoop, HIVE, CDBMS, Hadoop, Data Integration.

I. INTRODUCTION

One of the leading service that cloud service provider provides is the Database as a Service (DaaS). As in RDBMS, it consist of three layerd architecture, we have also proposed5 layer-architecture of Cloud Database Management system[3] the layers are External Layer, Conceptual Middle ware layer, Conceptual Layer, Physical Middle ware Layer and Physical Layer as shown in Fig 1.At conceptual Layer, Cloud service providerdeals with internal processing of data.

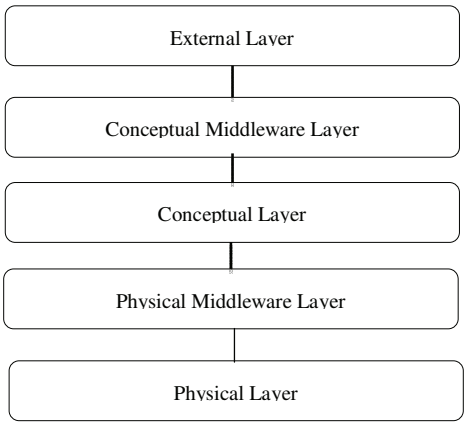

Fig 1. Layered Architecture of Cloud Database Management System [3].

Some of the Cloud users wants to keep their Common data at the cloud repositories as Cloud provides many advantages in terms of cost and time and other confidential data they want to store at their local stores i.e. at Local RDBMS. Now the problem comes how to integrate these data i.e. one at the Cloud repositories and the other at Local RDBMS [1]. Authors in [1] described one system "Big integrator" to enable general queries that combine data in cloud based systems and the Local databases and they have described the system with the help of one general scenario. Its Client server Architecture with the help of plugins absorber and finalizers allow better query capabilities.

In this paper we have analyzed that Hadoop – a framework can also be used for such kind of data integration. Hadoop not only provides easy data integration but it also many advantages as compare to the "Big Integrator" [1].

- It is an Open source software – which is freely available.
- It is used for processing large huge amount of data.
- It is used for processing structured and unstructured data.
- Hadoop provides many components with large query capabilities which are used for large data Processing like HIVE , Pig Map Reduce codes
- Its works as master slave.
- It imparts security because of Map Reduce technique. Data gets encrypted with the help of inbuilt codes of Java Map Reduce codes.

For analysis purpose we have taken the same data as in paper [1].

Table 1: Tables in Cloud Database and Local Database

Type of Database for Storage	Table Name	Primary Key attribute
Local RDBMS	Operator(PID, Name, Skill, Operates)	PID
Cloud Database	Machine(Model, Name, Manufacturer)	Model
	MachineInstaller(MID, Model, SID)	MID
	Site(SID, Name, Country, Region)	SID

978-1-4673-6792-9/15/$31.00©2015 IEEE

Model	Name	Manufacturer
1	M1	Volvo
2	M2	Volvo
3	M3	Volvo
4	M4	Volvo
5	M5	Volvo

Machine table

SID	Name	Country	Region
1	Uppsala	Sweden	Uppland
2	Chengdu	China	Si Chuan
3	Campinas	Brazil	Sao Paulo
4	Chapaevsk	Russia	Samara
5	Monki	Poland	Bialystok

Site table

MID	Model	SID
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5
6	1	1
7	2	2
8	3	3
9	4	4
10	5	5

MachineInstallation table

PID	Name	Skill	Operates
1	John	Operation	1
2	Oliver	Operation	2
3	Bruce	Operation	3
4	Carl	Operation	4
5	Thomas	Operation	5
6	Jens	Operation	6
7	Lucas	Operation	7
8	Alex	Operation	8
9	Ryan	Operation	9
10	Wes	Operation	10

Local table Operator

Fig 2. Data in Cloud and Local RDBMS [1].

Users keep some data at the cloud repositories and some at the Local databases. Table 1 summarized the types of tables on Local RDBMS and Cloud Databases and Fig 2. [1] Shows the actual data in all the tables.

Whole paper is summarized as follows: Section 2 describes the current state of work. In Section 3, Basics of Cloud Database, Hadoop and components like Hive and Sqoop are been discussed. Then in Section 4, data integration of Cloud Databases and Local RDBMS are described through Hadoop. Lastly Section 5 outlines the conclusion.

II. CURRENT STATE OF WORK

Author in [1] described about the system "BigIntegrator"- A client server architecture based system used to combine the queries for data located at Cloud and Local Database repositories. With the help of plugins like wrappers and finalizers it provides scalability in terms of queries.

We have analyzed that Hadoop [4][5][6][7][10] provides many advantages as it can be used for processing huge amount of data. It is used for processing both structured and unstructured data moreover, it also imparts inbuilt security because of Map Reduce technique. Google Big table [9], HDFS (Hadoop Distributed File Sytem [7] used for storing Large amount of data. Hadoop[10] provides both storage and processing capabilities. With SQLMR[8], HIVE[5](SQL-style) query language provides a SQL kind of interface where users can write logics in SQL construct which will be converted into Map Reduce job with the help of Hadoop. Sqoop[6] one of the component of Hadoop is used for imporing and exporing data from and to local RDBMS to Cloud Repositories.

III. CLOUD DATA BASE AND HADOOP

A. *Cloud Database*

Cloud provides database as a service that typically works on Cloud based platform. Leading Cloud Database Providers includes [2][3]Amazon Web Services, Google SQL Cloud, Microsoft Azure, Mongo Lab, Rackspace etc which provides DBaaS to their users in two forms either Database as a Virtual image or directly provide database service and maintain and manage all the requirements of the users. Users for security or some other reasons keep their confidential data with them i.e. on local RDBMS and other data they keep on the Cloud.

Now one problem comes when they want to integrate both the data. Hadoop is one platform used for processing and storing a large amount of data and is used for such type of data integration.

B. *Hadoop*

Hadoop is a platform which provides both storing and processing capability for large amount of data. Hadoop, as shown in Fig[3], is based on Master-Slave kind of Configuration where Storage is provided by HDFS(Hadoop Distributed File System) and Data Processing is provided by YARN (Yet Another Resource Negotiator) are described below.

- HDFS is used to provide storage. In HDFS Master i.e. NameNode stores all the metadata information i.e. which block is located in which Datanode (Slave). Which block is full, which block is free etc. Actual storage is being done on the Datanode.
- YARN is used to Process large amount of data. At master level Resource Manager handles all the incoming data processing requests. It sends request to NodeManager and Actual processing is done at Slave i.e. by Node Manager. Resource Manager allocates necessary resources for the processing of data to Node Manager's container component and then Processing is done.

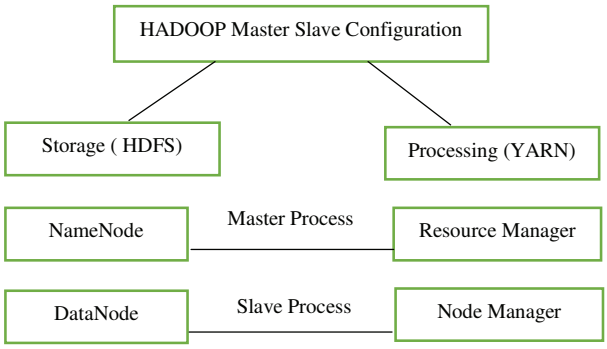

Fig 3. Hadoop Master Slave Configuration

C. *HIVE & Sqoop [5][6]*

HIVE (SQL- like) Language fills up the gap between the tools available in the market like Hadoop with Map Reduce and the kind of expertise users are having like user understand SQL kind of Languages very easily. Hive provides a SQL kind of interface where users can write logics in SQL construct which then converted into Map Reduce job with the help of Hadoop. SQOOP one of the component of Hadoop is used to import and export data from Local RDBMS to HDFS

IV. DATA INTEGRATION EXPERIMENT

We have analyzed Hadoop provides so many advantages in terms of Structured or unstructured Data storage, Huge Data Processing and in this paper Hadoop is used for Data Integration. Fig 4 explained the actual Data Integration Process where some data is located at the Cloud repositories and the other data is located at Local Databases such as RDBMS. Many Cloud Providers manages their huge amount of data on Distributed File system such as Hadoop's HDFS. Cloud users manages their Personal Data on Local RDBMS for security reasons.

In Section 1 we have shown that three tables i.e. Machine, MachineInstaller and Site are stored at Cloud Repositories and Operators table is stored at Local RDBMS. Data Integration of Cloud and Local Database is being done with the help of Hadoop.

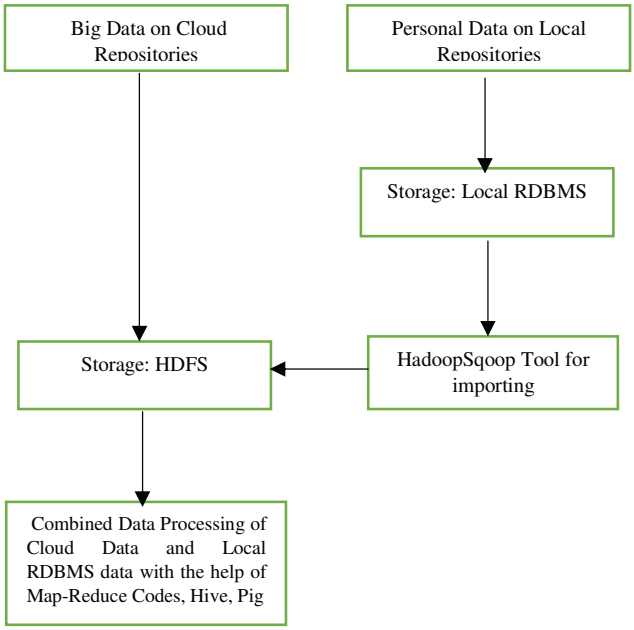

Fig 4. Data Integration Process

Experiment Details:

MySQL 5.6 is used to store the Operator table as shown in Fig 5. Other tables namely Machine, MachineInstaller and Site are

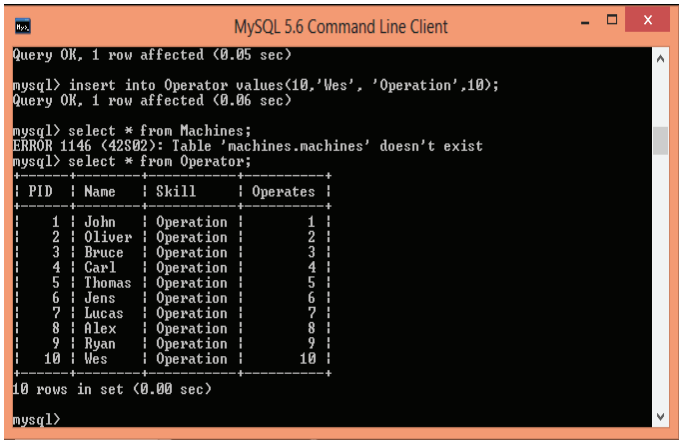

Fig 5. Data at Local RDBMS i.e. MySQL 5.6

Stored at HDFS and these tables are being created under HIVE environment as shown in Fig. 6

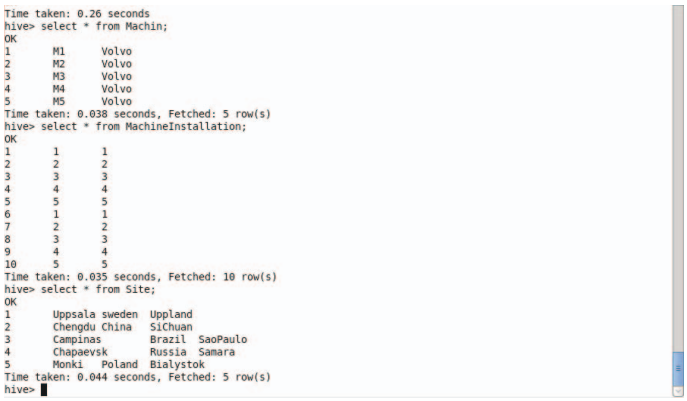

Fig 6. HIVE Enviornment for storing three tables

Fig 7 shows the result of Sqoop component which is used for importing data from MySQL 5.6 to Hadoop's HDFS. With the help of following command local table operator is imported from MySQL5.6 to HDFS

sqoop import –connect jdbc:oracle:thin:system/ system@ 192.168.93.1:1521:xe --username system -P --table system. Operator -- columns "PID" --target-dir /sqoopoutput1 -m 1

Then finally same query which is used to combine data from Cloud and Local RDBMS that is performed under HIVE enviornment to retrieve the existence of machine of Model ‘M1’ along with the operator’s name where the Manufacturer of

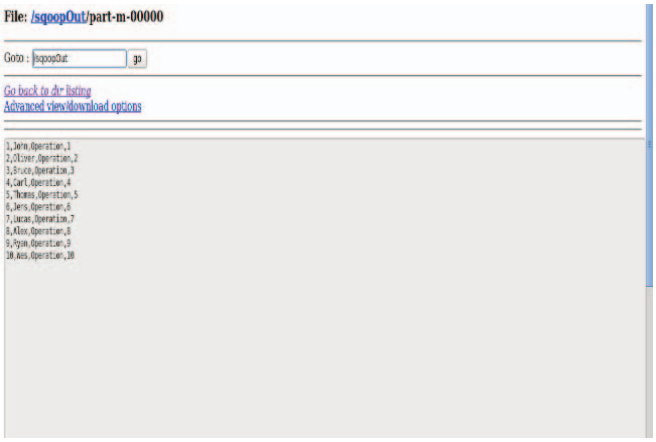

Fig 7. Sqoop Tool used for importing Data

machine names starts with ‘v’ , its installed region is ‘Uppland’ and te site identity ids one.

Query in HIVE : Select i.Mid, o.Name from Machine m, Machine Installation I, Site s, Operator o where m. Name=’M1’ and m. Manufacturer Like ‘V%’ and s.Region=’Uppland’ and s.Sid=1 and m. Model=i Model and i. Sid =s. Sid and i.Mid=o. Operates;

Result is provided under HIVE enviornment as shown in Fig. 8.

V. CONCLUSION

In this paper we have analyzed that Hadoop provides many advantages in terms of Huge Data Storage , Data Processing, Data integration etc. Storage is provided by Hadoop distributed File System and Data Processing is provided by Yet another Resource allocator (YARN). In this paper for data integartion Hadoop’s HIVE and SQOOP tools are being used. HIVE provides SQL kind of interface to interact with the data stored at HDFS and Sqoop is used for importing data that is stored at Local RDBMS i.e. MySQL 5.6 to HDFS. Once the data is imported to the HDFS then HIVE is used to intgrate data.

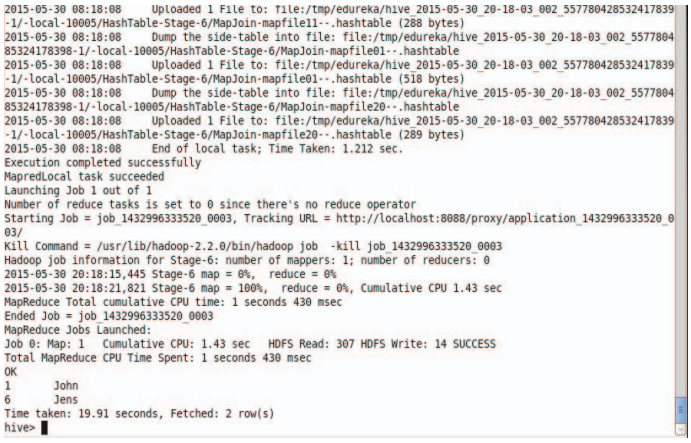

Fig 8. HIVE Query Result

REFERENCES

[1] M. Zhu and T. Risch. “Querying Combined Cloud-Based and Relational Databases”. International Conference on cloud and service computing (CSC) 2011, page 330-335.

[2] Mansaf Alam and Kashish Ara Shakil, “Cloud Database Management System Architecture”, UACEE International Journal of Computer Science and its Applications, Volume 3(1), 2013, page 27-31.

[3] S.Mongia,M.N.Doja,B.Alam,M.Alam,“5 layered Architecture of CloudDatabase Management System”, AASRI DCS2013 Conference.

[4] https://hadoop.apache.org/

[5] https://hive.apache.org/

[6] https://sqoop.apache.org/docs/1.4.2/SqoopUserGuide.html

[7] hadoop.apache.org/docs/r1.2.1/hdfs_design.html.

[8] M. Hsieh, C. Chang, L. Ho, J. Wu and P. Lui. “SQLMR: A Scalable Database Management System for Cloud Computing”. International Conference on Parallel Processing (ICPP) 2011, page315-324.

[9] F. Chang et al, “Big table: A distributed storage system for structured data,” in OSDI, 2006, pp. 205-218

[10] Devakunchari R, “Handling big data with Hadoop toolkit” IEEE International communication and Embeded Systems(ICICES), 2014, pages 1-5.