www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 4 April 2015, Page No. 11224-11228

A Comparative Analysis of Traditional RDBMS with MapReduce and Hive for E-Governance system

Mr.Swapnil A. Kale [1], Prof.S.S.Dandge [2]

[1] M.E. (CSE), Second Year, Dept. of Computer Science, Prof. Ram Meghe Institute Of Technology and Research, Badnera,Amravati Sant Gadgebaba Amravati University, Amravati, Maharashtra, India – 444701.
swapnil2304@gmail.com

[2] Assistant Professor, Dept. of Computer Science, Prof. Ram Meghe Institute Of Technology and Research, Badnera Amravati. Sant Gadgebaba Amravati University, Amravati, Maharashtra, India – 444701.
sangramdandge@gmail.com

Abstract: ***India is moving towards digitization. Most of the state and central governments are digitizing the departments due to which the use of E-governance applications has increased. As a result data is getting added in large size daily. Hadoop and MapReduce are used to handle these large volumes of variable size data. Processing and sharing of such a large data by traditional methods is difficult by the use of traditional methods. What is important is the speed at which one can process that data and extract the actionable business intelligence information. In this paper we present the comparison of traditional RDBMS with MapReduce technique and use of HiveQL to implement MapReduce on large databases. The paper also focuses on use of Sqoop connector to connect SQL and Hadoop that illustrates how big data can result in the transformation of the government by increased efficiency and effectiveness in the E-governance service.***

Keywords: Hadoop, Map-Reduce, Sqoop, Hive

1. Introduction

E-Governance applications are widely used these days due to awareness of citizens and also because of availability of data. Because of wide variety of data it becomes very difficult to process data by traditional methods and analyze it. The traditional methods are not optimized for high performance analysis. In India E-Governance services and applications are increasing radically [1]. Large team of professionals and data analytics is working behind the scenes to process and filter key points emerging from debates on different government portals and determine popular frame of mind about particular issues from social media websites. Effective management and analysis of huge data set introduces a critical challenge. Hadoop and its Map-Reduce framework along with the use of Sqoop and Hive can be efficient to process large volume of data.

The remainder of this paper is organized as follows. Section III provides necessity of Hadoop and Section IV describes the Architecture of Data Storage. Section V describes Comparison of Map-Reduce and Traditional RDBMS, Section VI gives idea about use of Sqoop connector and Section VII use of HiveQL to run Map-reduce jobs, Finally section VIII concludes this paper.

2. Literature Review

Various Ministry departments are being asked to reply with an action report on these ideas and policy suggestions. The biggest issue for Government is how to be pertinent? If all citizens are treated with self-respect and invited to team up, it can be an easier administration [2]. Recent development in the area of Big Data has attracted Government sector. The Prime Minister's Office is using different Big Data techniques to process ideas given by citizens on its crowd sourcing platform to generate actionable reports for various departments for consideration and implementation.

A large number of e-governance applications are already in operation in most of the state and central departments. So it is necessary to create a Business Intelligence infrastructure at the head quarter. The principle of Hadoop is to process data in a distributed file system manner. So, a single file is split into blocks and the blocks are spread in the cluster nodes. Hadoop applications require vastly available distributed file systems with unrestricted capacity. Hadoop is also used for batch processing purpose by Yahoo and Facebook. Hadoop and Hive form key parts of the storage and analytics infrastructure of Facebook[3], [6].

3. Necessity of Hadoop

Hadoop has the advantages of self-control, scalability, fault tolerance, effortless programming and most important is the storage method of Hadoop which is very suitable for the e-government data storage requirements [4].

The necessity can be analyzed by considering a simple example of development of an application for weather information management. Such application contains distributed and heterogeneous data sources including GIS maps, satellite images and temperature measurement stations, precipitation, atmospheric pressure, and wind speed.

A capability matrix based on different criteria to judge the strengths and weaknesses of open source cloud technologies having impact on the development of a user base are compared

as follows:

1) Management Tools

Eucalyptus – The version 1.6 of Eucalyptus comes along with Ubuntu server. However the system must be configured by hand, using command-line instructions and configuration files.

Django-Python – The creation of web application and its back-end database configuration must be done with Python modules and command-line tools. The application can be managed with a web interface once this is done.

Hadoop – Using Hadoop the system must be configured by hand, with a combination of text-based configuration files and XML files.

2) Development Tools

Eucalyptus - Eucalyptus can be extended to add new features but there are no tools present to help in this case other than importing the project into the NetBeans or Eclipse IDE.

Django-Python - The PyDev plug-in must be added to Eclipse to enable development and deployment of Django-based Web applications and Python applications.

Hadoop - An Eclipse plug-in helps in the development and debugging of Hadoop-based applications. It also helps in easy deployment of the application on the cloud.

4. Architecture of Massive Data Storage

Large data storage take up the hierarchical and distributed structure in design, and it can be divided into the different tier as shown in the figure

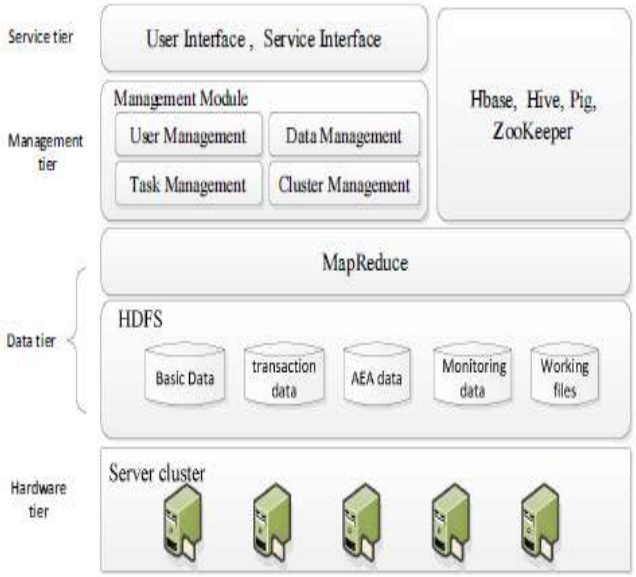

Figure 1: Four Tier Data Storage Architecture

1) Hardware tier

Hadoop cluster consists of a sequence of servers, which is configured as NameNode and Data Node, providing computing and storage services to the higher tier.

2) Data tier

Different kinds of data is stored in the data tier, including transaction data, basic data, the administrative examination, working files, monitoring data, etc. After processing is done by the MapReduce function, the data is stored on the server in the hardware tier and it can be managed by HDFS.

3) Management tier

Management tier mainly provides data management, user management, cluster management and task management.

4) Service tier

The service tier provides services to upper tier of management platform. Users could execute huge data processing and storage management using the GUI interface; the high-level software could directly use Hive, Pig, ZooKeeper or other data processing and management tools provided by Hadoop.

5. MapReduce vs Traditional RDBMS

MapReduce component of Hadoop plays an important role by using the Data Locality property where it arranges the data with the node itself so that the data access is fast [6], [7].There are different areas where MapReduce framework is found to have advantage over traditional database processing methods. In Grid computing work is distributed across clusters and they have a common shared File system hosted by SAN. The jobs in Grid computing are mainly computation intensive but in case of Big data where access to larger volume of data as network bandwidth is the main hurdle and because of that nodes start becoming idle.

MapReduce is considered to run jobs on trusted, dedicated hardware operating in a single data centre with very high collective bandwidth interconnections. [6], [7]. The traditional database works on data size in range of Gigabytes as compared to MapReduce dealing in petabytes. The Scaling in MapReduce is linear compared to traditional database. RDBMS however differs structurally, in updating, and access techniques from MapReduce. Comparison of RDBMS over some general properties of big data is as shown below.

Table 1: Traditional RDBMS compared with MapReduce

	Traditional RDBMS	***MapReduce***
Data Size	Gigabytes size	Petabytes size
Access	Interactive as well as batch	Batch
Updates	Many times Read and write	Read many times but Write once
Structure	Schema is Static	Schema is Dynamic
Scaling	Nonlinear	Linear
Integrity	High	Low

MapReduce can simply be solution to some problems that includes: distributed grep, counting URL access frequency, term-vector per host, inverted index, a variety of representations of the graph structure of web documents, etc.

Grid computing can be considered a substitute for MapReduce processing but Grid computing actually makes a use of message passing Interface (MPI). It provides great control to the user providing mechanism for handling the data flow. On the other hand Map Reduce works at the higher level where the data flow is implied and the programmer just looks for key-value pairs. Map Reduce handles the problem of coordination of jobs in distributed system easily being based on shared-nothing architecture. Its implementation itself detects the failed tasks and reschedules them on healthy machines. Thus the order of tasks running never really matters for a programmer. But in case of MPI, a clear management of check pointing and system recovery should to be done by the program [6], [7].

6. Using Sqoop

RDBMS is perfect fit for scale and speed in processing massive relational data i.e structured and static data sets. Big Data analytics Stack can act together with RDBMS through connector – Sqoop. ***Sqoop*** is a command-line interface application for transferring data between relational databases and Hadoop.

Using Sqoop connector and the Hadoop's Map-Reduce framework, structured data is processed using Hive programming. So that user specific analysis can be carried out easily. The Business Intelligence tools can be used for complex data analysis on the huge database. The storage of data among the Hadoop cluster would help in easy access and availability of data. Sqoop uses the primary key column to determine how to divide the source data across its mappers. Sqoop also has significant integration with Hive, allowing it to import data from a relational source into either new or existing Hive tables. [7]

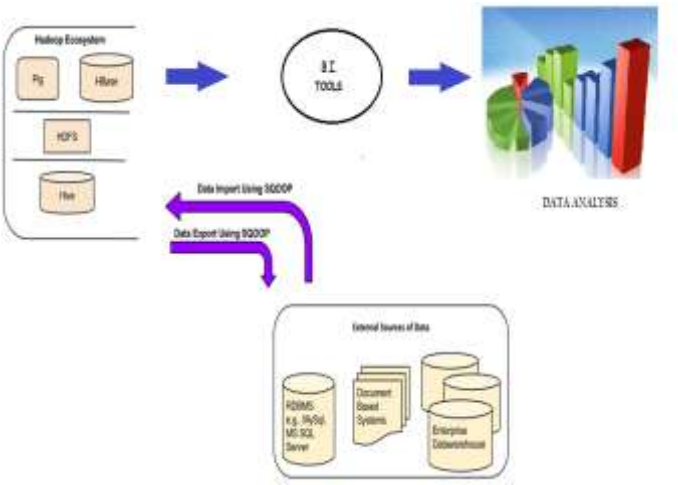

Figure 2: Sqoop to connect RDBMS and NoSQL database

7. HiveQL to run MapReduce Jobs

A programmer comfortable with SQL language will definitely prefer to express data operations with SQL language. Hive is Hadoop's data warehouse system that provides mechanism to project structure onto data stored in HDFS or a compatible file system [8].

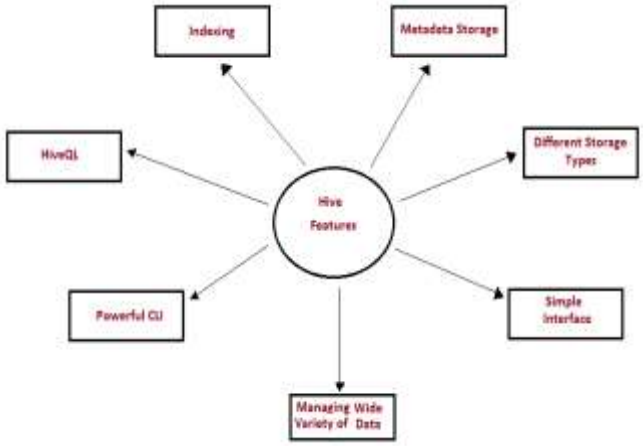

Figure 3: Hive Features

Hive enables users to plug in conventional map-reduce jobs into queries. The language includes a type system with support for tables containing primitive types, collections like maps, and nested compositions of maps. Hive query language (HiveQL) runs over Hadoop map-reduce framework, but conceals complexity from the developer, and it has some useful extensions that are helpful for batch processing systems.[10]

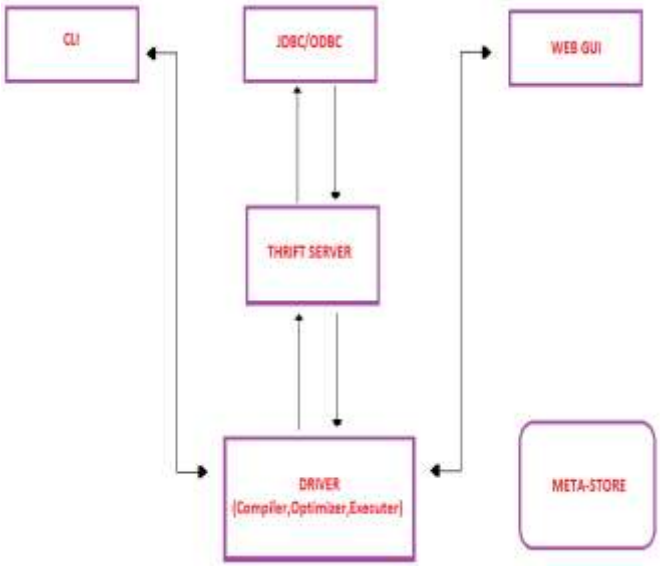

Figure 4: Working of Hive Queries

Advantages of Apache Hive

- Perfectly useful for low level interface requirement of Hadoop
- It supports external tables and ODBC/JDBC
- Intelligence Optimizer
- It supports Table-level Partitioning and speed up the query execution time
- Metadata store is a big advantage in the architecture that makes the lookup easy

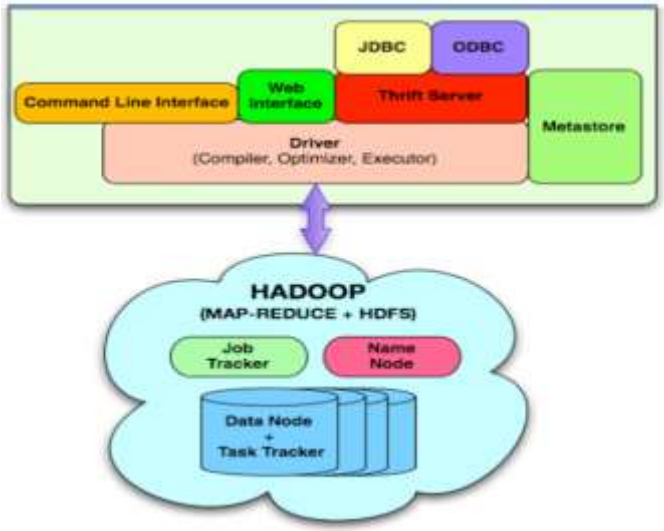

Figure 5: Hive System Architecture

SQL is a set based keyword based language, declarative programming language and not an imperative programming language like C, for accessing and manipulating database systems. Internally, a compiler translates HiveQL statements into a directed acyclic graph of MapReduce (MR) jobs, which are submitted to Hadoop for execution. Hive is one of the easiest to use high-level MapReduce (MR) frameworks. [9] Hive also maintains metadata in a metastore, which is stored in a relational database, as well as this metadata contains information about tables. Particular strength of HiveQL is in offering ad-hoc querying of data, in contrast to the compilation

requirement of Pig. Hive is a starting point for more full-featured BI (business intelligence) systems which offer a user friendly interface for common users.

In order to incorporate the BI system, we need strong data processing query language and then BI system can be built over it. HiveQL allows creation of new tables according to partitions as well as buckets (The data in partitions is further divided as buckets) and allows insertion of data in single or multiple tables. Following table shows comparison of SQL and HiveQL considering operations they perform.

Table 2: SQL vs Hive Query Language

Sr. No.	Operations & Functions	SQL	Hive Query Language
1.	Select	SQL-92 supports it.	Single table or view in the FROM clause. For partial ordering SORT BY is used. To limit number of rows returned LIMIT operations is used. HAVING clause is not supported.
2.	Updates	UPDATE, INSERT, DELETE	INSERT OVERWRITE TABLE(It populates complete table or partition)
3.	Data types present	Integral, floating point, fixed point, text and binary strings, temporal	Integral, floating point, boolean, string, array, map, struct
4.	Default Join Types	Inner Join	Equi Join
5.	Built-In Functions	Built-in functions are in Hundreds.	Dozens of built-in Functions present.
6.	Multiple table inserts	Not supported in SQL	Supported in HiveQL
7.	Create table as select	Not valid in SQL but may be found in some databases	Supported by HiveQL
8.	Extension points	User-defined functions and Stored procedures.	User-defined functions and Map-Reduce scripts.

Considering the time analysis of Map-Reduce jobs performed on the same system by Hive, Pig and simple java program we can say that Hive outperforms Pig in most cases. The number of Map and Reduced tasks carried out by Hive can be more suitable for smaller data sets. Writing mapper and reducers, compiling, debugging, submitting the jobs, and retrieving the results using the Map-Reduce approach takes developer's time. On the other hand, it is much easier to describe data operations with Hive for a user less familiar to Java programming language. Users familiar with SQL language may prefer to use Hive.

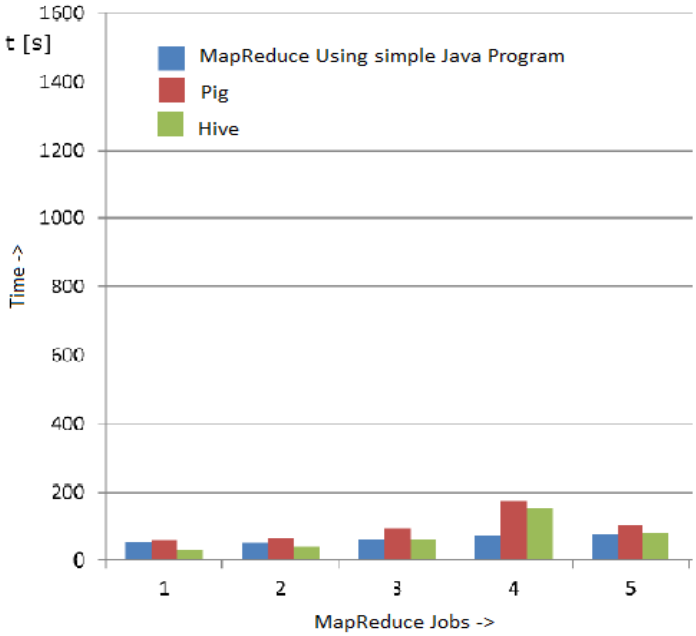

Figure 6: Comparison of Hive with Pig and Simple Java

8. Conclusion

Usually working files data of the e-government are small files, but they are very large in number, that requires special treatment before storing them. Hadoop is one of the best distributed massive data processing framework and it has high performance on distributed computing and distributed storage. Using Hadoop the massive data storage and processing by MapReduce in an e-government system proposes the use of Hive architecture and its components that have advantage over traditional methods of data processing. Also HiveQL can be useful for treatment of small files in the system.

References

[1] Rajagopalan M.R and Solaimurugan vellaipandiyan "*Big data framework for national E-governance plan*" ICT and Knowledge Engineering (ICT&KE), 2013, 11th InternationalConferenceon DOI:10.1109/ICTKE.2013.6756283 Publication Year: 2013, Page(s): 1–5 IEEE CONFERENCE PUBLICATIONS

[2] http://analyticsindiamag.com/pmo-using-big-data-techniques-mygov-translate-popular-mood-government-action/

[3] A. Thusoo, Z. Shao, S. Anthony *et al.*, "Data warehousing and analytics infrastructure at facebook," in SIGMOD 2010, International conference on Management of data, pp. 1013-1020.

[4] Fu Chang-Jian, Leng Zhihua. A Framework for Recommender Systemsin E-Commerce Based on Distributed Storage and Data-Mining. International Conference on E-Business and E-Government (lCEE20 I 0), Volume I, pp.3502 - 3505.

[5] D. Borthakur, J. Gray, J. S. Sarma *et al.*, "Apache hadoop goes realtime at Facebook," in ACM SIGMOD International Conference on Management of Data, 2011,

pp. 1071-1080.
[6] Tom White (2013). Inkling for Web [Online]. Available: https://www.inkling.com/read/hadoop-definitive-guide-tom-white-3rd/chapter-1/comparison-with-other-systems
[7] Katal, A.; Wazid, M.; Goudar, R.H., "Big data: Issues, challenges, tools and Good practices," Contemporary Computing (IC3), 2013 Sixth International Conference on , vol., no., pp.404,409, 8-10 Aug. 2013
[8] "Welcome to Hive!" December, 2012; http://hive.apache.org/
[9] Ivan Tomašić, Aleksandra Rashkovska, Matjaž Depolli and Roman Trobec "A Comparison of Hadoop Tools for Analyzing Tabular Data**",** Jožef Stefan Institute, Slovenia, Informatica 37 (2013) 131–138
[10] Thusoo, A. ; Sarma, J.S. ; Jain, N. ; Zheng Shao ; Chakka, P. ;Ning Zhang ; Antony, S. ; Hao Liu ; Murthy, R. "Hive - Petabyte scale data warehouse using Hadoop" in Data Engineering (ICDE), 2010 IEEE 26th International conference on DOI: 10.1109/ICDE.2010.5447738 Publication Year:2010, Page(s):996-1005

Author Profile

Swapnil A. Kale received the B.E. degree in Information Technology from Pune Institute of Computer Technology, Pune, Pune University in 2008. Now pursuing M.E. from Prof.Ram Meghe Institute of Technology & Research, Badnera-Amravati.

Prof. Sangram S. Dandge, is an Assitantant Professor, in Department of CSE, at Prof. Ram Meghe Institute Of Technology and Research, Badnera,Amravati. Sant Gadgebaba Amravati University, Amarvati, Maharashtra, India - 444701