Log Mining Based on Hadoop's Map and Reduce Technique

Anuja Pandit
Department of Computer Science, PVPIT
PVPIT
PUNE, INDIA
anujapandit25@gmail.com

Amruta Deshpande
Department of Computer Science, PVPIT
PVPIT
PUNE, INDIA
amrutadeshpande1991@gmail.com

Prajakta Karmarkar
Department of Computer Science, PVPIT
PVPIT
PUNE, INDIA
prajdodiedo@gmail.com

ABSTRACT:

In the world of cloud and grid computing Virtual Database Technology (VDB) is one of the effective solutions for integration of data from heterogeneous sources. Hadoop is a large-scale distributed batch processing infrastructure and also designed to efficiently distribute large amounts of work across a set of machines. Hadoop is an implementation of Map Reduce. This paper proposes application for inauguration of new branch of pizza in particular area according to hits from customers. In this paper we will take the log files for the particular website which will be stored on web mining server. These data will be passed on to the cloud server for region wise distribution on the virtual servers. Mapping and reduction will be done on these region wise data. The final output is then sent back to the server and client. This paper utilizes the parallel and distributed processing capability of Hadoop Map Reduce for handling heterogeneous query execution on large datasets. So Virtual Database Engine built on top of this will result in effective high performance distributed data integration

KEYWORDS: Database Integration; Virtual Database Technology; Hadoop; Map Reduce; Heterogeneous Databases; Query Optimization.

1. INTRODUCTION:

Today for distributed systems like Cloud and Grid data integration from heterogeneous data sources is unavoidable. The data is made available in several structured formats (HTML, XML, etc), as well as in tables, spreadsheets and statistical tools. Integration of these data is carried out at a tremendous cost, often at 35% of IT budget.VDB technology is one of the effective solutions for integration of data from heterogeneous databases. This technology is used in many data integration applications like facebook, Google etc.

This paper proposes application for inauguration of new branch of pizza in particular area according to hits from customers. For this we are going to mine the log files. After that we are going to use map reduce technique to reduce that data area wise. We will consider that area which will have maximum number of hits, as the area to inaugurate a new dominos branch.

Generally while opening new pizza restaurant owner set various discounts and offers on pizzas. It can be good marketing but not best. Using our system instead of giving offers we calculate hits from people and decide where to open new restaurant and this survey is done accurately. Map Reduce is a new framework specifically designed for processing huge datasets on distributed sources. Apache's Hadoop is an implementation of Map Reduce. Currently Hadoop has been applied successfully for file based datasets. The execution engine that is developed on top of Hadoop applies Map and Reduce techniques to break down the parsing and execution stages for parallel and distributed processing. Map Reduce will provide the fault tolerance, scalability and reliability because its library is designed to help process very large amount of data using hundred and thousands of machine, which must tolerate the machine failure. This paper proposes to utilize the parallel and distributed processing capability of Hadoop MapReduce for handling heterogeneous query execution on large datasets. So Virtual Database Engine built on top of this will result in effective high performance distributed data integration.

2. HDFS:

The Apache Hadoop project defines the hadoop distributed file system (HDFS) as: "the primary storage system used by Hadoop applications. HDFS creates multiple replicas of data blocks and distributes them on compute nodes throughout a cluster to enable reliable, extremely rapid computations."

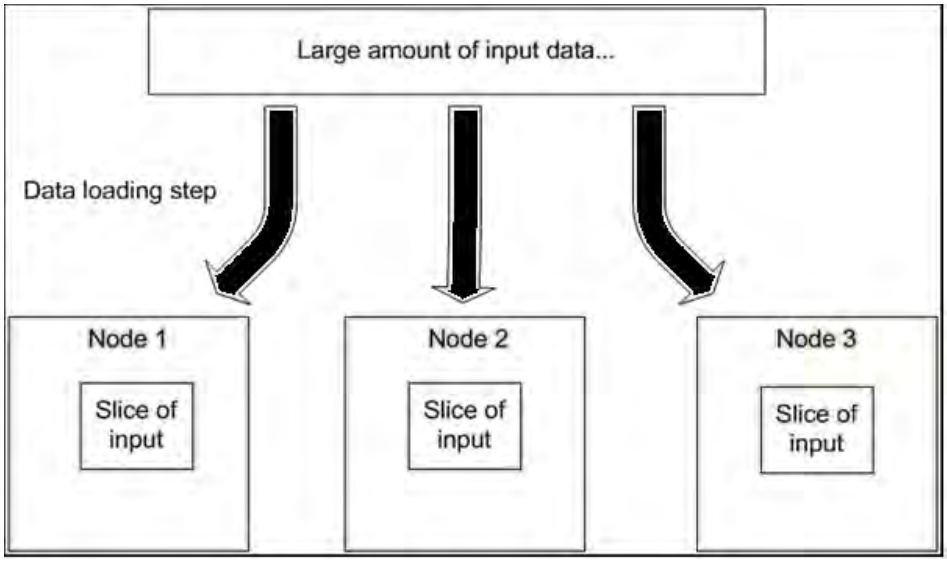


Figure 1: Hadoop Disributed File Sysetm

2.1 Writing Files To Hdfs:

1. HDFS stores file system metadata and application data separately
2. HDFS stores metadata on a dedicated server, called the Name Node.
3. Application data are stored on other servers called Data Nodes.
4. All servers are fully connected and communicate with each other using TCP-based protocols

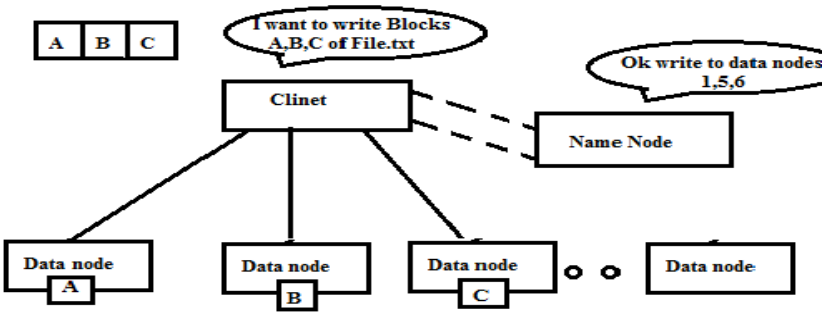


Figure 2: Writing Files To Hdfs

3. MAPREDUCE

MapReduce is a programming model for expressing distributed computation on massive amount of data and an execution framework for large-scale data processing on clusters of commodity servers. It was originally developed by Google and built on well known principles in parallel and distributed processing. Hadoop is the open source implementation of MapReduce written in java which provides reliable, scalable and fault tolerance distributed computing. Hadoop environment set up involves a great number of parameters which are crucial to achieve best performance. It allows programmers to develop distributed applications without any distributed knowledge.

Key-value pair forms the basic data structure in MapReduce. Keys and values may be primitives such as integers, floating point values, strings, and raw bytes or they may be arbitrary complex structures. Programmers typically need to define their custom data types. The map function takes the input record and generates intermediate key and value pairs. The reduce function takes an intermediate key and a set of values to form a smaller set of values. Typically just zero or one output value is produced by the reducer. In MapReduce, the programmer defines a mapper and reducer with the following

Signature: Map (k1, v1) → [(K2, v2)]
Reduce (K2, [v2]) → [(k3, v3)]
[….] denotes the list.

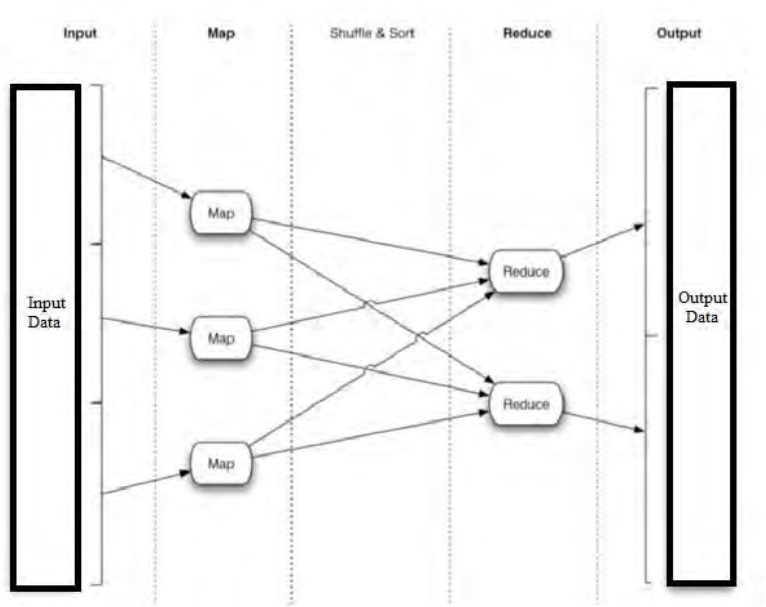


Figure 3: Map Reduce Technique

3.1 Mapper- .

The map is a function that splits up the incoming work.

3.2 Reducer-

The function that receives the mapped work and produces final result is the reducer.

3.3 Master-

Monitoring the execution of mappers and reducers as to re-execute them when failures are detected.

MapReduce framework is responsible for automatically splitting the input, distributing each chunk to workers (mappers) on multiple machines, grouping and sorting all intermediate values associated with the intermediate key, passing these values to workers (reducers) on multiple resources,. Monitoring the execution of mappers and reducers as to re-execute them when failures are detected is done by the master.

4. SYSTEM DESIGN

4.1 System Architecture

Systems architecture is the conceptual model that defines the structure, behavior, and more views of a system.

Overall Flow Of Application:

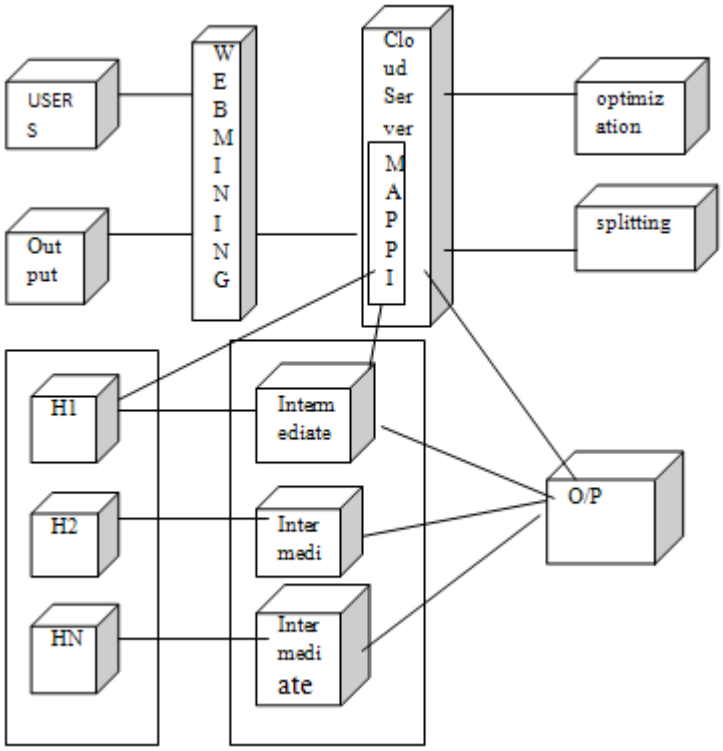


Figure 4:System Architecture

5. QUERY OPTIMIZATION

The purpose of grouping together similar queries to same data source, a queue is introduced in the Execute Engine. The incoming sub queries from the Query Parser are queued in this queue. A Query Combiner acts on this queue in a periodic interval to look for similar queries to same data source, and groups them together. This grouped query is now sent to a Grouped Query Queue. The Map Reduce Master removes the grouped queries from this queue and processes using different workers. This effectively optimizes the time spent in processing and executing duplicate queries.

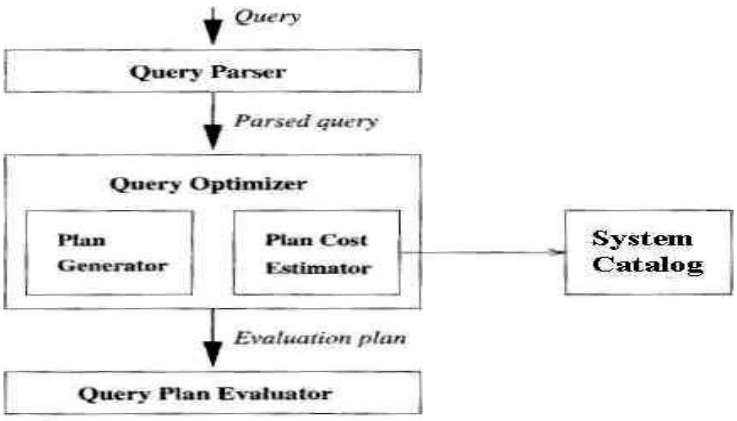


Figure 5 : Query Optimization

6. TECHNICAL SPECIFICATION

6.1. Advantages

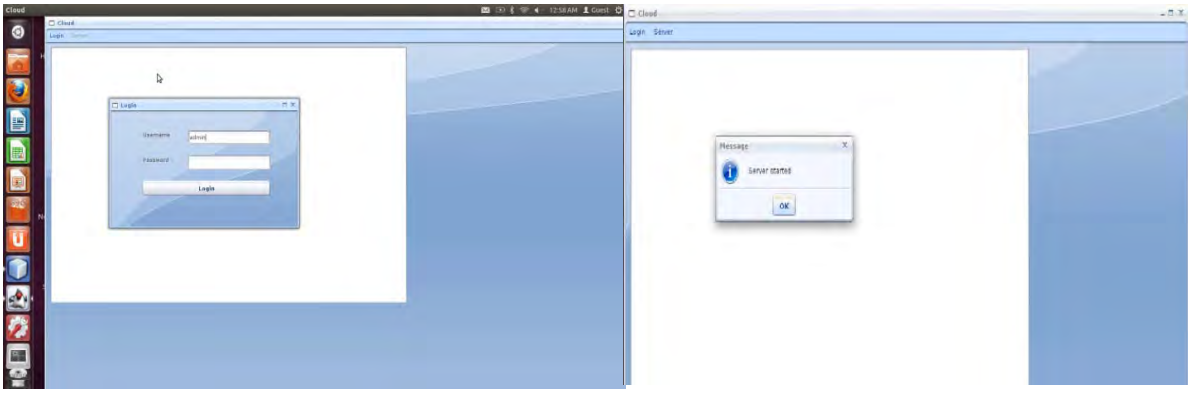
1. It reads data parallel and provide higher output and used for huge processing datasets on distributed sources
2. Designed to run on cheap commodity hardware such as hadoop.
3. Handles data replication and node failure. It is efficient and reliable data processing

6.2. Disadvantages

1. We cannot update data after it is inserted.
2. We can only load data using bulk load, do bulk delete.

7. SERVER (Snapshots):

Hadoop server has started.



7.1 Working of Server

1. Hadoop distributed file 2. Hadoop web analytics. 3. Hadoop graph.

7.2 GUI for application: This is login page for admin and customer. Registered customers can access the account. Customer can buy pizzas and this information is in the form of logs which will be sent to the hadoop server. Admin can see the hits of users. By using this application he will get hits in reduced form according to particular area.

8. REFERENCES

[1] Asish Gupta, Venkey Harinarayan, Anand Rajaraman. Virtual Database Technology, ACM Sigmod Record 26 (4) (1994) 57-61.

[2] Wenhao Xu, Jing Li, Yongwei Wu, Xiaomeng Huang, Guangwen Yang, VDM: Virtual Database Management for Distributed and File System, Grid And Cooperative Computing (2008), IEEE.

[3] Yuji Wada, Yuta Watanabe, Keisuke Syoubu, JunSawamoto, Takashi Katoh. Virtual Database Technology for Distributed Database, 2010 IEEE 24th, International Conference on Advanced Information Networking and Applications Workshop.

[4] Ferreira.R, Moura-ires, J., Martins, R., Pntoquilho.M. XML based Metadata Repository for Information Systems, IEEE Artificial intelligence conference, 2005.

[5] Proceedings of ICETECT 2011 Application of hadoop map reduce technique to VDS.

[6] Hammoud, S., Maozhen Li, Yang Liu, Alham N.K.,Zelong Liu. MRSim: A discrete event basedMapReduce simulator. Seventh International IEEEConference on Fuzzy Systems and Knowledge Discovery (FSKD), 2010.

[7] Apache Hadoop, http://Hadoop.apache.org.