RESEARCH ARTICLE OPEN ACCESS

A Comparative Study on Operational Database, Data Warehouse and Hadoop File System

T.Jalaja[1], M.Shailaja[2]

[1,2](Department of Computer Science, Osmania University/Vasavi College of Engineering, Hyderabad, India)

Abstract:

A Computer database is a collection of logically related data that is stored in a computer system, so that a computer program or person using a query language can use it to answer queries. An operational database (OLTP) contains up-to-date, modifiable application specific data. A data warehouse (OLAP) is a subject-oriented, integrated, time-variant and non-volatile collection of data used to make business decisions. Hadoop Distributed File System (HDFS) allows storing large amount of data on a cloud of machines. In this paper, we surveyed the literature related to operational databases, data warehouse and hadoop technology.

***Keywords* — Operational Database, Data warehouse, Hadoop, OLTP, OLAP, Map Reduce.**

I. INTRODUCTION

One of the largest technological challenges in software systems research today is to provide mechanisms for storage, manipulation, and information retrieval on large amounts of data. A database is a collection of related data and a database system is a database and database software together. Operational Databases are transactional databases, which supports on-line transaction processing (OLTP) that includes insertions, updates, deletions and also supports information query requirements.

Operational Database is designed to make transactional systems run efficiently. It is used to store detailed and current data. The main emphasis of this system is on very fast query processing, maintaining data integrity in multi-access environments. Thus databases must strike a balance between efficiency in transaction processing and supporting query requirements .They can't further be optimized for theapplications such as OLAP, DSS and data mining.

A Data Warehouse [1] is a database that is designed for facilitating querying and analysis. In contrast to databases, data warehouses generally contain very large amounts of data from multiple sources that may include databasesfrom different data models and sometimes files acquired from independent systems and platforms. Often it is designed as OLAP (On-Line Analytical Processing) systems.

Big data comes from relatively new types of data sources like social media, public filings, content available in the public domain through agencies or subscriptions, documents and e-mails including both structured and unstructured texts,digital devices and sensors including location based smart phone, weather and telemetric data .There are several approaches to collecting, storing,processing, and analyzing big data.Hadoop [5] is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. Hadoop is not atype of database, but rather a software ecosystem that allows massive parallel computing. The main focus of this paper is to present a clear understanding on operational database, data warehouse and Hadoop technology.

This paper is organized as follows: In section II, we present literature survey on Operational database, Datawarehouse and Hadoop Distributed File System. In Section III, we present the comparative studies among the works cited above. In section IV, we summarize the work.

II. LITERATURE SURVEY

Day-by-day technology in software systems is improving and is becoming very challenging in terms of storage, manipulation, and information retrieval. There are different types of storage systems having the capability of storing data from few Megabytes to Peta bytes.In this paper we discuss about the traditional/operational database, which is capable of storing the transactional data of an organization in part A. Then in part B,we discuss about the datawarehouse, which is capable of analyzing the business data which is used for decision making. Later in part C we come across the way to store large volume of structured as well as unstructured data, which is further utilized for analytics.

A. Operational Database (OLTP)

An operational database contains data about the things that go on inside an organization or enterprise. For example, an operational database might contain fact and dimension data describing transactions, data on customer complaints, employee information, taking order and fulfilling them in a store , track of payments and inventory ,information and amounts from credit cards etc.

Operational systems maintain records of daily business transactions.Traditional databases support on-line transaction processing (OLTP), which includes insertions, updates, and deletions, while also supporting information query requirements. An operational database contains enterprise data which are up to date and modifiable. As the name implies, it is the database that is currently and progressive in use capturing real time data and supplying data for real time computations and other analyzing processes.

A database structure commonly used in GIS in which data is stored based on 2 dimensional tables where multiple relationships between data elements can be defined and established in an adhoc manner. Relational Database Management System - a database system made up of files with data elements in two-dimensional array (rows and columns).This database management system has the capability to recombine data elements to form different relations resulting in a great flexibility of data usage. SQL is used to manipulate relational databases. The relational model contains the following components:

•Collection of objects or relations.
•Set of operations to act on the relations.
•Data integrity for accuracy and consistency.

All other database structures can be reduced to a set of relational tables. It is easy to use when compared to other database systems.Traditional databases are optimized to process queries that may touch a small part of the database and transactions that deal with insertions or updates of a few tuples per relation to process.

Because of the very dynamic nature of an operational database, there are certain issues that need to be addressed appropriately. An operational database can grow very fast in size and bulk so database administrations and IT analysts must purchase high powered computer hardware and top notch database management systems. The operational database is the source of data for the data warehouse.

B. Datawarehouse (OLAP)

According to Surajit Chaudhuri and Umeshwar Dayal as in [6], a data warehouse is a “subject-oriented, integrated, timevarying,non-volatile collection of data that is used primarily in organizational decision making as in [7]. Typically, the data warehouse is maintained separately from the organization’s operational databases. The data warehouse supports on-line analytical processing (OLAP), the functional and performance requirements of which are quite different from those of the on-line transaction processing (OLTP) applications traditionally supported by the operational databases.

A data warehouse is a centralized repository that stores data from multiple information sources and transforms them into a common, multidimensional data model for efficient querying and analysis as shown in Figure 1.

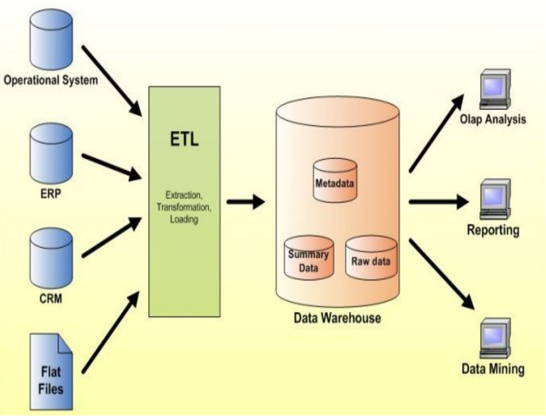

Figure 1: Data Warehouse Architecture

Historical, summarized and consolidated data is more important than detailed, individual records. Since data warehouses contain consolidated data, perhaps from several operational databases, over potentially long periods of time, they tend to be orders of magnitude larger than operational databases; enterprise data warehouses are projected to be hundreds of gigabytes to terabytes in size. The workloads are query intensive with mostly ad hoc, complex queries that can access millions of records and perform a lot of scans, joins, and aggregates. Query throughput and response times are more important than transaction throughput. To facilitate complex analyses and visualization, the data in a warehouse is typically modeled multidimensional.

The 3-D data cube of the sales data with respect to time, item, and location is shown in Figure 2.

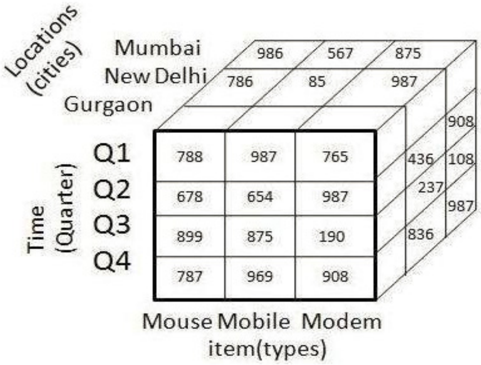

Figure 2: Multidimensional data (Data Cube)

Typical OLAP operations include rollup (increasing the level of aggregation) and drill-down (decreasing the level of aggregation or increasing detail) along one or more dimension hierarchies, slice and dice (selection and projection), and pivot (re-orienting the multidimensional view of data).

Decision support usually requires consolidating data from many heterogeneous sources: these might include external sources such as stock market feeds, in addition to several operational databases. The different sources might contain data of varying quality, or use inconsistent representations, codes and formats, which have to be reconciled. Finally, supporting the multidimensional data models and operations typical of OLAP requires special data organization, access methods, and implementation methods, not generally provided by commercial DBMSs targeted for OLTP. It is for all these reasons that data warehouses are implemented separately from operational databases.

A popular conceptual model that influences the front-end tools, database design, and the query engines for OLAP is the multidimensional view of data in the warehouse.

C. Hadoop Distributed Filesystem (HDFS)

According to P.Sarada Devi, V.Visweswara Rao and K.Raghavender as in [8], Data warehouse is a collection of heterogeneous data which can effectively and easily manage the data from multiple databases in a uniform fashion. In addition to data warehousing ETL, many technologies such as ERP (Enterprise Resource Planning), SCM,SAP, BI tools are used in the world market for handling structured data efficiently and effectively.In order to handle the large amount of structured, semi structured and unstructured data generated by different social network or industries, Hadoop is being used.

Hadoop is a framework developed as an Open Source Software based on papers published in 2004 by Google Inc. that deal with the “Map Reduce” distributed processing and the “Google File System”, a system they had used to scale their data processing needs. Hadoop is an open source framework for writing and running distributed applications that process large amounts of data". It consists of two main components –

Storage: The Hadoop Distributed File System

Processing: "Map Reduce"

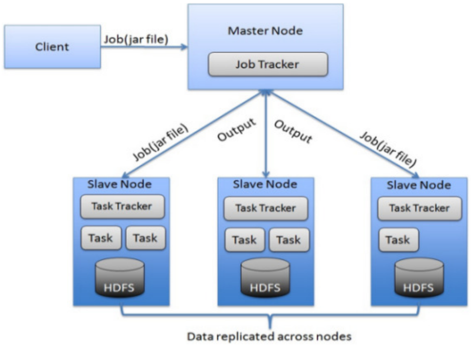

Figure 3: Hadoop Architecture

HDFS is the storage component of Hadoop. It's a distributed file system that's modeled after the Google File System (GFS) paper [11]. Files in HDFS are stored across one or more blocks, and each block is typically 64 MB or larger. Blocks are replicated across multiple hosts in the hadoop cluster to help with availability and fault tolerance.

HDFS is able to store huge amounts of information, scale up incrementally and survive the failure of significant parts of the storage infrastructure without losing data. Hadoop creates clusters of machines and coordinates work among them. Clusters can be built with inexpensive computers. If one fails, Hadoop continues to operate the cluster without losing data or interrupting work, by shifting work to the remaining machines in the cluster. HDFS manages storage on the cluster by breaking incoming files into pieces, called "blocks," and storing each of the blocks redundantly across the pool of servers. In the common case, HDFS stores three complete copies of each file by copying each piece to three different servers [9].

Map Reduce is a batch-based, distributed computing framework modeled after Google's paper on Map Reduce [12]. Map reduce data processing model has two main phases - "MAPPING" and "REDUCING".

Mapping Phase: Map Reduce takes the input data and feeds each data element to the mapper which filters and transforms the input data into the desired format for business use.

Reducing Phase: The reducer processes all the outputs from the mapper and arrives as a final result by performing business logic to get the desired output.

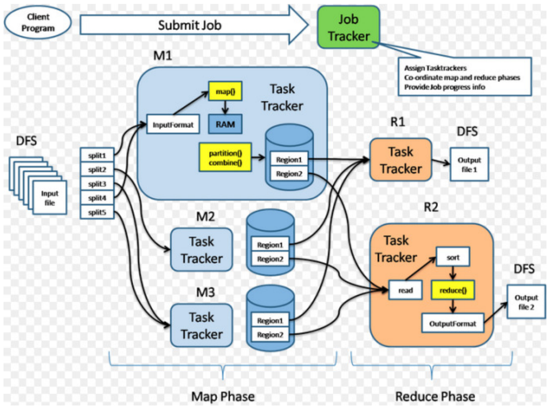

Figure 4: MapReduce Architecture

MapReduce Architecture

The processing pillar in the Hadoop ecosystem is the MapReduce framework. The framework allows the specification of an operation to be applied to a huge data set, divide the problem and data, and run it in parallel. From an analyst's point of view, this can occur on multiple dimensions. For example, a very large dataset can be reduced into a smaller subset where analytics can be applied [10].

In Hadoop, these kinds of operations are written as MapReduce jobs in Java. There are a number of higher level languages like Hive and Pig that make writing these programs easier. The outputs of these jobs can be written back to either HDFS or placed in a traditional data warehouse.

III. COMPARISION

TABLE 1
COMPARISION OF OLTP, OLAP, HDFS

	Database (OLTP)	**Datawarehouse (OLAP)**	**Hadoop (HDFS)**
Type of data	Transactional data, detailed data.	Historical data, Derived data, Metadata	Real time data, derived, enterprise, Simple and complex data
Data Storage	Stores operational data – structured data	Stores historical information – structured data	Stores enterprise data - structured, semi structured and Unstructured data.
Data Modeli	Network , hierarchical,	Conceptual, Logical, And	File system(Sc

ng techniques	object oriented, object relational and Relational modeling techniques (Schema-on-write)	Physical Data modeling techniques.	hema-on-read)
Optimization	Efficient for performing read-write operations of single point transactions.	Efficient for performing Reading/retrieving large data sets and for aggregating data.	HDFS is efficient for performing reading and writing large files (gigabytes and larger). High throughput
Data Access techniques	Access to few records at once by predefined transactions	Access a lot of records in each access by ad hoc queries and periodic reports	Access Streaming data, large volumes of data

Data warehousing is faster and cost effective when compared with Apache Hadoop.Unlike traditional ETL tools, Hadoop persists the raw data and can be used to re-process it repeatedly in a very efficient manner .Hadoop mostly process semi structured and unstructured data also, whereas Data Warehouse is best suited for processing only structured data efficiently.

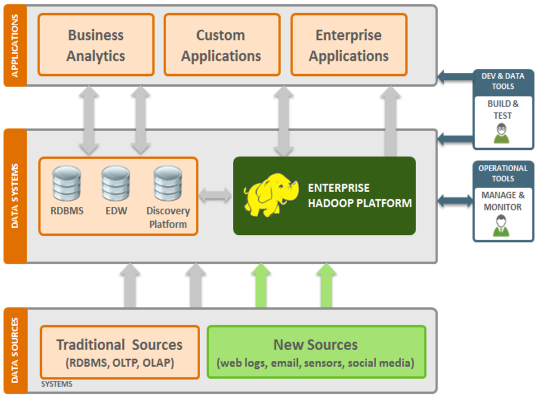

Figure 5: Showing OLTP, OLAP, Hadoop

IV. CONCLUSION

Thus, the operational/traditional databases are used to store an application specific data pertaining to an Organization or Enterprise.This data is used for query purpose.But, in order to store large amount of data compared to Traditional database, which is further used for analysis purpose, based on which the business decisions are made, Datawarehouses are maintained. Incontrast, to store

Zettabytes of structured/unstructured dynamic data, Hadoop distributed File systems are used. These are more efficient and fault tolerant.

REFERENCES

1. *Wided Oueslati and Jalel Akaichi, "A Survey on Data Warehouse Evolution", Vol.2, No.4, November 2010, IJDMS.*

2. *T.K.Das1 and AratiMohapatro "A Study on Big Data Integration with Data Warehouse"International Journal of Computer Trends and Technology (IJCTT) – volume 9 number 4– Mar 2014.*

3. *Dr. Mrs.Pushpa Suri, Mrs. Meenakshi Sharma" A Comparative Studybetween The Performance of Relational& object Oriented Database In Data Warehouse".*

4. *Inmon, W. et.al. (2000) Exploration warehousing: turning business information into business opportunity. John Wiley & Sons.*

5. *Hadoop.* http://hadoop.apache.org

6. *Surajit Chaudhuri and Umeshwar Dayal "An Overview of Data Warehousing and OLAP Technology", ACM Sigmod Record, March 1997).*

7. *Inmon, W.H., Building the Data Warehouse. John Wiley, 1992.*

8. *P.SaradaA Devi, V.Viswewara Rao and K.Raghavender"Emerging Technology Big Data-Hadoop Over Data Warehousing Using ETL" Proceedings of 2nd IRF International Conference, 21st September-2014, Vizag, India.*

9. *MrigankMridul, AkashdeepKhajuria, Snehasish Dutta, Kumar N " Analysis of Bidgata using Apache Hadoop and Map Reduce" Volume 4, Issue 5, May 2014".*

10. *Harshawardhan S. Bhosale, Prof. DevendraP. Gadekar "Harshawardhan S. Bhosale1, Prof. Devendra P. Gadekar" International JournalofScientific and Research Publications, Volume 4, Issue 10, October 2014.*

11. *Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung – "The Google File System.*

12. *Jeffrey Dean and Sanjay Ghemawat ,Google, Inc –"MapReduce: Simplified Data Processing on Large Clusters".*