ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 12, June 2015

Analysis and Design of ETL process using Hadoop

Shruti Tekadpande, Leena Deshpande
Student, Assistant professor

Abstract— ***Extract-transform-load periodically transfers data from source system to the datawarehouse with data having different format. An increasing challenge before ETL system is that source data is available in many different forms. It may be text file, sensor data etc. Also, the volume of such kind of data is increasing day by day. It is today's requirement to process data with different format. There are many programming paradigm which can handle this issue .Among various technique, one relevant and emerging programming paradigm is MapReduce. It has been used in data-intensive area at various companies like Facebook, yahoo etc. Hadoop, distributing computing platform, provides platform to analyze data .Hive is an emerging data warehousing platform build over HDFS. It can be used as ETL platform to populate datawarehouse at organization .In this paper, we will analyze the data processing capacity of HIVE by performing extraction, building various transformations and load operation on data. The aim of this project is to develop dimensional modeling technique for Hadoop and design ETL process to load the datawarehouse system.***

Index Terms— **Datawarehouse, ETL, Hadoop, Hive**

I. INTRODUCTION

The ETL system plays an important role in building datawarehouse. The ETL system typical extract different form of data from different source system ,various transformation and cleansing rule are apply on data with business rules and requirements.[16] There are various traditional open source ETL tools available in market to meet the requirements. But increasing amount as well as unstructured nature of data places a new challenge before traditional ETL tool to draw business conclusion .It has become common for an enterprise to collect data in unstructured form with increasing amount to process each day. Unstructured nature and increasing amount makes ETL very time consuming but the time window assigned to ETL process remains the same .In addition to this due to rapidly changing nature of business compels users to demand the data as soon as possible. To parallelize the task is key to achieve better performance and solves the issue of scalability.

Different technologies have been emerging in recent years .The novel cloud Computing and MapReduce has been used for parallel computing in data intensive area. The key concept behind the MapReduce technology is that it implements the map and reduce function, which eventually processes the key/value pairs and it can be executed on different instances. MapReduce provides scalability and capacity on commodity Machines. It has fault tolerance, provide load balancing, task scheduling to a parallel program. It is very interesting to develop ETL programming using Map Reduce technology.

The processing of data in ETL shows the composable property. This can be explained as follows the data processing of dimension tables and fact table can be divided into smaller computation unit and partial results from these small computations can be combined to develop the final results in datawarehouse. This resembles the functionality of MapReduce in term of Map and reduce function .Hence MapReduce can be a good platform to migrate the ETL process.

This paper is structured as follows: The section describe the datawarehouse and necessity to migrate its platform to another platform The section 3 describes the MapReduce programming technique to process data .Section 3 describe the dimension processing and fact processing respectively in MapReduce. Section 4 describes the hive system with different transformations with implementation in Hadoop platform. The comparison with Pentho data integration tool is described in the next section .The last section concludes the paper and provides the future work for this project.

II. DATAWAREHOUSE SYSTEM

In the early 1990, a new trend emerged .The enterprises wanted other database known as datawarehouse in which data from various operational databases are gathered and stored to perform business intelligence. The datawarehouse and operational database performs different activity in organization. OLTP optimizes for updates .Operational database store the daily activity in an organization. On the other hand, datawarehouse process ad-hoc queries, which are sometimes complex. Also periodic loading of datawarehouse with slowly changing data is taken place in the enterprises. Data is stored in different schema. [1]

A. Data ware house platform and ETL process

A data warehouse platform consists of an operating system, a database management system (DBMS), and data storage and hardware servers. A datawarehouse platform plays an important role in datawarehouse /analytical system. Different categories of datawarehouse platform from existing to future solutions are [2]:

1. Traditional RDBMS database
2. Column oriented database
3. In memory databases
4. Cloud based datawarehouse system
5. Hadoop based datawarehouse system

Nowadays study has been shown that there are some disadvantages of traditional datawarehouse platform .A survey conducted in [2] describe the problem and the percentage of user facing the problem.

Poor query response	45%
Can't support advanced analytics	40%
Can't scale to large data volumes	37%
Current platform is a legacy we must phase out	23%

As ETL is an indispensible step in building datawarehouse system [15].Migration of ETL to Hadoop and loading data itself in Hadoop can be prove advantageous .The ETL steps of joining number of tables, lookups, cleaning of data can be performed with effectively less amount of time on Hadoop platform [10].[16] paper also recommends the migration of ETL 20% workload to Hadoop gains maximum while building datawarehouse

- Relatively high elapse processing times
- Very complex scripts like change data capture, joins, cursors etc
- File processing and semi structured data

III. HADOOP AND DISTRIBUTED SYSTEM

Hadoop, an open source framework provides the distributed file system.[6] It has been used to process and analyze the unstructured and large amount of data in a very short time. It also transforms the data using MapReduce paradigm. The key characteristic of Hadoop is that it partitioned the data over the different node of cluster .The cluster provides the computational capacity, large amount of storage for continually generating data .It also adds IO bandwidth by adding commodity servers.[4] The case study shows small scale industries can also be benefitted by usage of Hadoop[9] The organization should consider the Hadoop as a viable option when there is unexpectedly increase in data. [3] Explains this fact by giving a suitable example. According to[13] Hadoop exhibits the MAD characteristics. The M stands for Magnetic, which attracts all nature of data .The A for agility and D for deep analytics.

A. MapReduce programming module

MapReduce is a parallel programming paradigm taken from functional programming concepts. This programming paradigm is first proposed by Google for processing large amount of data in distributed environment. [5]It performs the computations by means of two functions namely map and reduce. In [17], MapReduce based datawarehouse framework named Cheetah has been developed .it can process the data at 1GB/sec. The problems of searching, indexing, data mining have been handled effectively in this framework.

B. Hive

The Hive is an open source data warehousing platform on top of the hdfs .The driver, Metastore, query compiler are the main building blocks of the HIVE system. The driver is used to manage the lifecycle of HIVEQL statement. The Metastore is a system catalogue .It stores the metadata about the tables, columns and partition etc.The HIVEQL queries are being transformed into DAG of map/reduce tasks by the query compiler. There are various other components such as command Line Interface, the web UI and JDBC/ODBC driver [11]

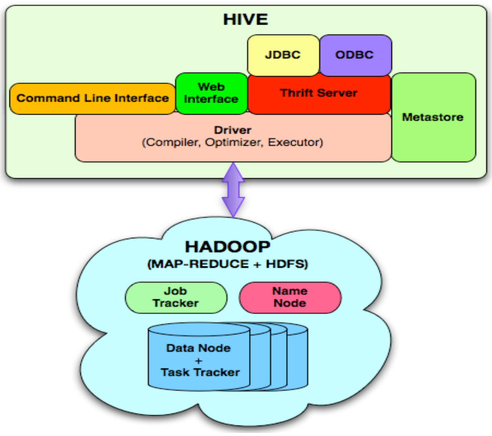

Fig1.Architecture of Hive

IV. PROPOSED SYSTEM

As per [7], modelling methodology in Hadoop for modelling datawarehouse needs to be studied. In this paper, we design the ETL process in HIVE and apply traditional dimensional modelling to Hadoop. The paper makes several contributions: we build star schema to process data and build various transformations, data cleaning, and filtering method to populate the data. This paper addresses the processing of dimensional schema in Hive system. The results have been compared with the Pentho data integration tool explicit ETL tool used with relational database .The evaluation shows that Hive achieves good scalability over the relational tool with some disadvantages.

A. The running example

To compare the results with traditional ETL tool and Hive system, we use the running example throughout this paper. In this example, there are various tables interconnected to each other. The examples include the information about customers, products, their purchasing details etc. To process, analyze and store the information in the datawarehouse system it required the ETL process.

While storing the data into datawarehouse, data modeling technique has to be applied on existing schema to transform it into another schema .In the above example, We have chosen the star schema to model the data .The star schema consists of number of dimensions table connected to centrally situated fact table .The fact table stores the foreign keys which are nothing but the primary keys of dimension tables.

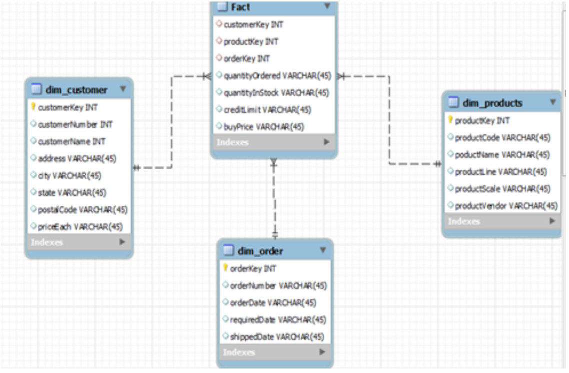

Fig.2 Star schema design

B. Hardware configuration

The experiments have been performed on single node cluster having 8GB of memory. It has 64 bit operating system with 1TB of hard disc. The cloudera 4.7.0 has been used to set up a single node cluster .To test the results with RDBMS, the Pentho data integration (PDI) 5.2.0 has been configured with MYSQL.

The first step is to extract the required data from the number of table and decide dimension table according to the requirement. DIMCUSTOMERS, DIMORDERS, DIMPRODUCTS and the FACT_CENTRAL have been decided. The dimension table stores the basic information about the entity whereas the fact stores the calculated measures and aggregated data (if required) in denormalized form. In our example, primary keys namely customerKey, productKey, orderKey will serve as surrogate keys in the fact table. Also the quantity Order, quantittyInstock etc. are the measures to be store in the fact table.

The power of Hive is in parallel processing of data. Also, it uses novel MapReduce programming approach. The processing of ETL and dimensions are as follows [8]

1. Loading of raw files into the hdfs
2. Partition the input data sets
3. Read the input data and provide the data to the map function in the map readers
4. Process dimension data, perform transformations and load it into dimension stores
5. Prepare fact processing
6. Read the input data for fact processing and
7. Bulk-load fact data into the DW

C. Fact Processing

This Fact processing includes the looking up of dimension keys, aggregation of measures and then loading of data into the fact table. According to process Fact algorithm[8],measures needs to stored in face table first then dimension look ups needs to be apply on dimension table to get the surrogate key

D. Experimental Evaluation

To judge the performance of hive and MYSQL, we have integrated the Pentho data integration solution with MYSQL. The procedure to develop star schema with transformations has been carried out with PDI. The two data sets of small size and large size have been used .For this purpose , data set has been generated using SQL script and loaded into both the systems. The aim of this experiment is to analyse the performance of two systems in terms of processing time of dimension and fact table.The following table shows the time taken by Hive to perform various transformations on small size and large size data set and its comparison with the Pentho system.

TABLE I. EXPERIMENT WITH SMALL DATASET

No .of rows	No. of jobs	Time taken by hive	Time taken by Pentho
Customer Dimension with 220 rows	4	28s	1m 10s
Order dimension with 330 rows	4	22s	60s
Product Dimensionwith 450	4	24s	52s
Fact table	4	5m 1s	5m 10s

TABLE IIIII. EXPERIMENT WITH LARGE DATASET

No. of rows	No. of jobs	Time taken by Hive	Time taken by Pentho
Customer dimension with 1,50,000	4	52s	3 m 19 s
Order dimension with 2,50,000	4	60s	3m 30s
Product dimension with 1,60,000	4	55s	2m 48 s
A large Fact table	4	6m 10s	7 m 10 s

Third data set contains 300 MB of data in each table. This experiment is carried out with Hive and observed that it takes 5 min to process each table on single node and 10 min to process fact table.

A. The observed features of Hive and Pentho are as follows:

The Hive uses the Hive shell to perform all ETL operations while Pentho has Spoop which provides drag-and-drop like facility to build the ETL. ETL in hive requires extensive knowledge of HiveQl and programming as opposed to the Pentho data integration tool. The Pentho provides easy user interface to develop ETL. There is a facility to define user defined functions (UDF) in both systems. The Hive provides the integration with Python and java. We can perform various transformations in hive as well as in Pentho such as filter, aggregations and joins. In case of multipoint, Pentho performs better than the hive. In Pentho, we can perform both SCD-1 and SCD-2 whereas hive does not provide SCDs .We can implement data modelling technique (star schema in our case) by means of handcode while Pentho is integrated with MYSQL provides explicit data modelling technique. Processing of an unstructured data in hive is the advantages over Pentho

V. DIRECTION FOR FUTURE REASEARCH

Future research directions can be recommended as follows:(1)Performing ETL operations on than 1 GB of data .(2) The SCD-I and SCD-II should be performed on Hive to meet the today's requirement .(3) Performance evaluation with other ETL tool is recommended.

VI. CONCLUSION

In this paper , different datawarehouse platforms with its limitations has been studied .Also an indispensable step of Extract-transform-load (ETL) process required to develop the datawarehouse system has also been studied. It is claimed in various papers massive data processing capacity, ability to process unstructured data , capacity to perform complex transformations makes Hadoop as viable alternative to migrate the ETL process of data warehouse system.

The main objective of this paper is to perform dimensional modeling in hive and load the results in DW in less time. We have developed the dimensional modeling technique and perform various transformations in hive .We have successfully load the results into star schema by performing transformations in hive. The results have been compared with traditional ETL tool. The results shows that Hive perform better than PDI in terms of processing time.

REFERENCES

[1] Michael Stonebraker and Çetintemel "One Size Fits All: An Idea Whose Time Has Come and Gone", Proceedings of the 21st International Conference on Data Engineering (*ICDE* 2005).

[2] Philip Russom, Next generation Datawarehouse platforms, The Data warehousing Institute, 2009.

[3] Awadallah Amar, Graham.Dan, Hadoop and the Data Warehouse- When to use which, Cloudera Inc and Teradata Corporation, 2011.

[4] J. Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1):107–113, 2008.

[5] Hadoop Map Reduce. http://hadoop.apache.org/mapreduce.

[6] Hadoop distributed file system (hdfs). http://hadoop.apache.org/hdfs.

[7] kuldeep deshpande, and dr. Bhimappa desai **"**limitations of datawarehouse platforms and Assessment of hadoop as an alternative", *IJITMIS* ,Volume 5, Issue 2, pp. 51-58.

[8] Xiufeng Liu, Christian Thomsen, and Torben Bach Pedersen "ETLMR: A Highly Scalable Dimensional ETL Framework Based on MapReduce", pp. 1–31, Springer.

[9] Hollingsworth. A, Graham.D, Hadoop and Hive as scalable alternatives to RDBMS –A Case Study, Boise State University Scholarworks, 2012.

[10] T.K.Das and Arati Mohapatro, A Study on Big Data Integration with Data Warehouse, International Journal of Computer Trends and Technology (*IJCTT*) – volume 9 number 4, Mar 2014.

[11] Thusoo, A. Sarma, J.S., Jain, N., Shao, Z., Chakka, P. Zhang, N.,Antony, S., Liu, H., and Murthy, R. Hive – A Petabyte Scale Data Warehouse Using Hadoop. Proc. of ICDE, 2010.

[12] Yongqiang He et all " RCFile: A Fast and Space-efficient Data Placement Structure in MapReduce-based Warehouse Systems", ICDE 2011.

[13] Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., and Welton, C. MAD Skills: New Analysis Practices for Big Data. PVLDB 2(2), 2009.

[14] Clark Bradley, Ralph Hollingshead, Scott Kraus, Jason Lefler, Roshan Taheri ,Data Modeling Considerations in Hadoop and Hive, Technical paper, SAS,2013.

[15] Liu Z.H., Krishnamurthy.V. , Towards Business Intelligence over Unified Structured and Unstructured Data using XML, edited volume "Business Intelligence-Solution for Business Development", Intech Publisher, 2011.

[16] Offload your Data warehouse with Hadoop, Syncsort publication, 2014.

[17] Chen Songting. "Cheetah – Ahigh performance custom Datawarehouse on top of MapReduce" Proceedings of VLDB, Vol 3, No. 2, 2010.

[18] M. Stonebraker, D. Abadi, D.J. DeWitt, S. Madden, E. Paulson, A. Pavlo, and A. Rasin. MapReduce and parallel DBMSs: friends or foes? Communications of the ACM, 53(1):64–71, 2010.

AUTHOR BIOGRAPHY

Shruti Tekadpande is studying master degree at VIIT, Pune. Her area of interest is ETL, datawarehouse.

Mrs. Leena Deshpande is an assistant professor at VIIT, Pune. Her area of interest is data mining and business intelligence.