International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 6 454 - 456

Apache Hive

1st Haresh Masand, 2nd Deep Mandani, 3rd Aman Dikshit, 4th Ameya Parkar
MCA Department,
Affiliation of Mumbai University

Abstract: The size of data coming from various has increased rapidly. Within few seconds; terabytes of data is collected by servers today. Sources includes data from Internet, satellites, social networking sites, mobile phones, etc. So processing such colossal amount of data with relational database is proving costly and impacting performance. Hadoop is a popular open- source framework used for processing such large data sets. Hadoop uses Map-Reduce programming for processing the datasets. Map reduce is a low level and requires to write their own custom mapreduce tasks. This requires knowledge of programming language either c++, Python, Java or Ruby. So to avoid this problem Hive was introduced. Hive is an open source data warehousing tool that is built on top of hadoop. It is SQL-Like Language which is very useful for non technical or a person is not into development but still can process data using hadoop framework.

*****

1. Introduction

Apache hive is an open source data warehousing tool built on top of the hadoop framework. It is an SQL-Like language used for processing of large data sets. Analysis on large datasets has been one of the main focus of many companies like Google, Yahoo, Facebook, etc. For example, Facebook in every second, generates phetabytes of data, so processing such huge amount of data becomes almost impossible with relational databases. Hive is much similar to SQL in terms of syntax but works differently internally. Hive converts simple looking queries into number of mapreduce jobs. When the Hive query complies, the query compiler creates a directed acyclic graph of map reduce tasks. Based on the graph the query gets executed. Following are the main components of Hive System:

- MetaStore: The component which stores the metadata of tables, columns, partitions, etc.
- Query Compiler: It transforms the query into a tree representation, which then transforms to a number of map reduce jobs.
- Driver: It manages the lifecycle of the HQL statements.
- Execution Engine: Executes the tasks produced by the compiler in a proper order.

Hive Server: It provides an interface for connecting to other applications.

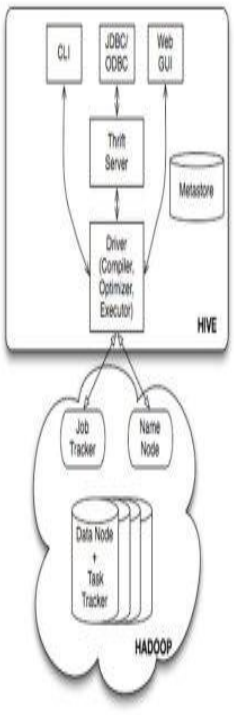
2. Problems with map reduce approach

Map reduce is a low level programming model used in handling big data in hadoop. Map reduce is very useful when you have to write complex business queries, custom partitioners or when developers need definite program control. Following are problems of using Map reduce approach:

- Knowledge of programming language is must. It can be Java or C or Python.
- Writing nested queries in Map reduce is difficult.
- Joining two datasets is another difficult task in map reduce.
- The time taken to write program in map reduce is more.

3. Hive architecture

Hive is one of the important component of hadoop. The below diagram shows abstract architecture of how hive works with hadoop.



The diagram shows how Command line interface, JDBC/ODBC and Web GUI connects to Hive. When user comes with CLI, it directly gets connected to Hive Drivers. User gets connected to Hive by using API of thrift server when it comes with JDBC/ODBC and users gets connected directly when comes via Web GUI. The Hive driver receives

query from the user and sends it to hadoop architecture. The Hadoop architecture uses namenode, datanode, job tracker and task tracker to execute the give query.

4. Benefits of Hive

- Knowledge of programming language is not required.
- It is simple to execute queries in Hive as it is similar to SQL.
- It is very to write queries containing joins.
- It has good execution speed and high throughput.
- It supports partitioning of data at the level of tables to improve performance.
- It has a rule based optimizer for optimizing logical plans.

It supports external tables which makes it possible to process data without actual storing in into HDFS.

5. Limitations of Hive

- Cannot implement complex queries.
- Hive is useful only if data structured.
- Debugging in Hive is difficult.
- Correlated queries are not supported.
- It does not support update and delete.
- It does not support single insert. The data is required to be loaded from file

6. Applications of Hive

- Hive can be used for reporting. We can generate many types of different reports that suits user requirements.

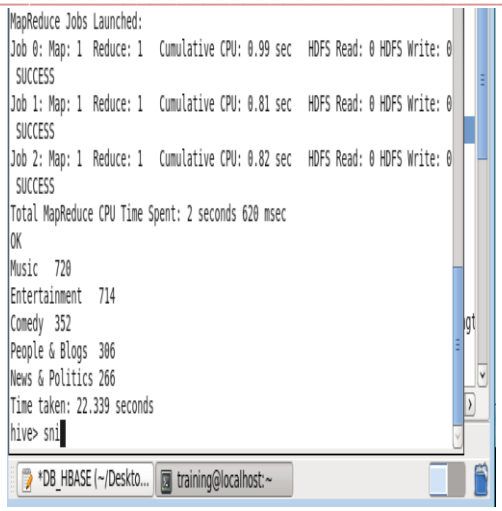
- Ad-hoc analysis.
- Machine Learning
- Data mining
- Research and Development
- Real Time Web analytics
- Log Data Analysis

7. Example on Hive

Below is the example of how hive can be used for data processing. We have taken youtube data and performed analytics on it using Hive.

- Find out the top 5 categories with maximum number of videos uploaded.

Query: select Category, count (*) as video_count from youtube group by Category sort by video_count desc limit 5;

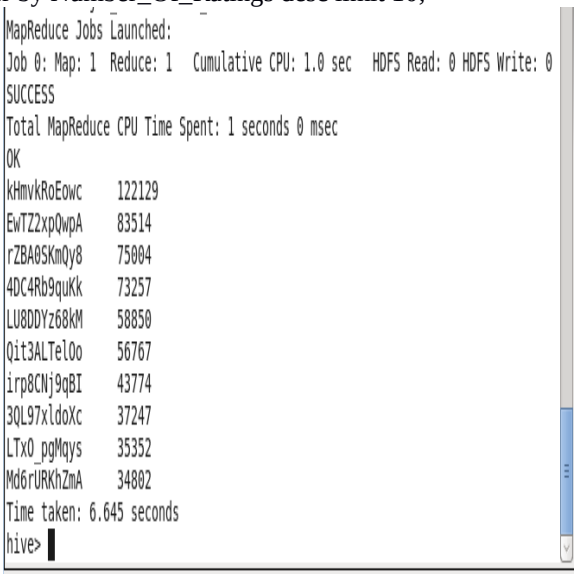


- Find out the most viewed videos

Query: select Video_Id, Views from youtube order by Views desc limit 1;

- Find out the top 10 rated videos

Query:select Video_Id,Number_Of_Ratings from youtube order by Number_Of_Ratings desc limit 10;



8. Conclusion

Today data is not just a data, it is a big data. So conventional relational database is not efficient enough to handle such large

amount of data. Hadoop is a framework for big data processing. Hive is one of the important component of hadoop ecosystem. Hive provides simpler way of performing data analytics on large data sets without having to write complex programs in map reduce. One of the advantage of using Hive is performance and execution speed. There
are still improvements need to be made in Hive like processing unstructured data, executing correlated sub queries, etc.

9. Acknowledgement

We would like to take this opportunity to express our profound gratitude and deep regard to Prof. Ameya Parker, for his exemplary guidance, valuable feedback and encouragement throughout the duration of the research paper. His valuable suggestions
were of immense help throughout our research work. His perceptive criticism kept us working to make this project in a much better way. Working under him was a great experience for us. We would also like to give my sincere gratitude to all the friends and colleagues who filled in the survey, without which this research would have been incomplete.

10. References

[1] Dean Wampler, “Programming Hive”
[2] http://www.google.com
[3] http://www.wikipedia.com
[4] http://www.hortonworks.com
[5] http://www.cloudera.com