Airline Dataset Analysis using Hadoop, Hive, Pig and Impala

Patel Ayushi - A20407392

Bhargavi Deshpande - A20395720

Abstract

This paper is about the analysis of the airline data set which is performed using Cloudera ,it delivers the modern platform for analytics optimized for the cloud. The project is performing big data analysis on airline dataset using Hadoop and tools like Pig, Hive and Impala. At the end analysis will be visualized using Excel spreadsheet.

Keywords: Hadoop, HDFS, Pig, Hive and Impala, Data Analysis

Introduction

There is a lot of excitement that exists with the term Big Data. In simple words, Big Data can be large-scale data which does not have a well-defined structure. The size of the data is so huge that it is not practically easy for a single computer to store and process all the data. In traditional computing approach there are many problems in a different way and the focus was always to increase the processing speed and power of the computer. As the data grows exponentially, the processing power of the single computer becomes a bottleneck and thus a new approach was needed to address the issue at hand.[3] A new way was developed wherein many non-expensive commodity computers were working together in harmony with each other, in order to store and process this big data in parallel. This allows us to extract meaningful information from a large data set.[3] In addition, by using the cloud technology, it is easy to create cluster, compute and release the computing resources when it is not needed. So, from the cloud technology we get the computing power of the cluster of computers with minimal investment. The draft mainly contains details about Hadoop, Hive, Pig and Impala and gives vague idea of the project flow.

Practicum

1. Datasets

The airline dataset contains flight details, carriers, airport details and plane-data. All data are in .csv file format.[2]

The data consists of flight arrival and departure details for all commercial flights , from 2003 to 2008 and each file are sized with – 115 MB to 125 MB, The airport details file is sized with 239 KB and plane-data is with 419 KB.[2]

2. Cloudera quick-start and Impala

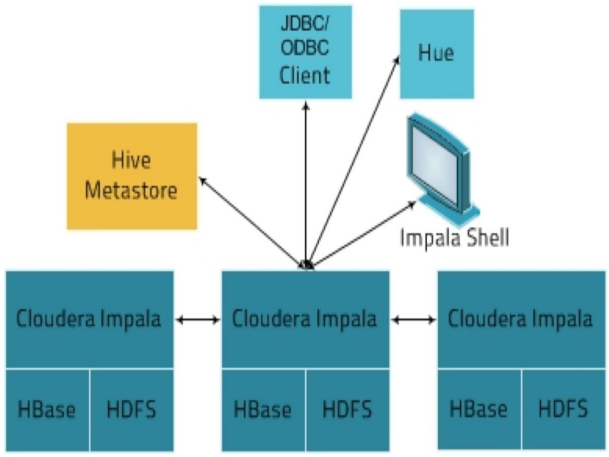
Cloudera provides a scalable, flexible, integrated platform that makes it easy to manage rapidly increasing volumes and varieties of data in your enterprise. Cloudera products and solutions enable you to deploy and manage Apache Hadoop and related projects, manipulate and analyze your data, and keep that data secure and protected[4]. Cloudera provides many products and tools. We are using CDH and Apache Impala.

CDH: The most complete, tested and popular distribution of Apache Hadoop and other related open-source projects, including Apache Impala and Cloudera Search. CDH also provides flexibility,security and integration with numerous hardware and software solutions.

Apache Impala: A massively parallel processing SQL engine for interactive analytics and business intelligence. Its highly optimized architecture makes it ideally suited for traditional BI-style queries with joins, aggregations, and subqueries. It can query Hadoop data files from a variety of sources, including those produced by MapReduce jobs overloaded into Hive tables. The YARN resource management component lets Impala coexist on clusters running batch workloads concurrently with Impala SQL queries. You can manage Impala alongside other Hadoop components through the Cloudera Manager user interface and secure its data through the Sentry authorization framework.[5]

How Impala Works with CDH.

The following graphic illustrates how Impala is positioned in the broader Cloudera environment[5]:



3. Data Preprocessing with pig/Spark

- *Purpose:* Give unstructured data some structure, Handle bad data, Integrate data with values from external system, Preprocess binary content,
- Writing directly to the data warehouse.

- Data storage is less of a problem than efficient data retrieval

4. Working with Hive vs. Impala or both

- Apache Hive and Impala, are used for running queries on HDFS.
- Hive generates query expressions at compile time whereas Impala does runtime code generation. Apache Hive is not that ideal for interactive computing whereas Impala is meant for interactive computing.
- Hive is batch based Hadoop MapReduce whereas Impala is more like Massively Parallel Processing (MPP) database.
- Impala offers fast interactive SQL queries directly on our Apache Hadoop data stored in HDFS or HBase.
- Impala needs more time than Hive LLAP; Impala supports Text,RCFile,Sequence File,Avro and Parquet file formats.When it comes to parquet files Impala writes data files with a default block size of 1GB.

To execute queries both Hive and Impala have a strong MapReduce foundation. However, when we use both together, we get the best out of both the worlds. Such as compatibility and performance.

5. Hive/Impala partitioning and clustering

- Hive is used for partitioning.Two ways of partitioning dynamic and static partitioning both are being implemented on dataset to know the difference.
- We are using using Static partitioning as we get data year by year.
- Dynamic partitioning is also shown for the dataset as with dynamic partitioning we can add any number of partitions with single SQL execution[6].
- Hive is used for clustering. Clustering is used to do sampling and for bucket sort/join.
- Data sampling is done on carrier,origin and time using bucketed tables[7].
- Impala does not support Clustering.

Purpose: To get efficient spread of the data so that data does not get skewed by randomness.

6. Data Compression, tuning and query optimization

Hive

- Data Compression :Hive uses MapReduce and so data is compressed using Hive to save space on disk and network[8].
- Execution Engine used is hive.execution.engine.

Impala

- COMPUTE STATS to improve query performance while performing joins[9].
- File Formats :Statistic is computed on various impala file formats to compare performance.

7. Using database views to represent data

- Purpose of representing data using View.
    - Security
    - Projection to hide complexities

8. Visualizing data using Microsoft Excel Via ODBC

- Storing and visualizing our data using Microsoft Excel.

9. QlikView (Extra):

- Planning to visualize our dataset using other visualization tool like QlikView- To load, explore data and perform analysis.

Conclusion

This practicum is all about the analysis of airline dataset. The main purpose of the practicum is - after preprocessing the data sets, information and description part of three datasets namely airport, carriers and plane details can be summarized, i.e. city, state or country of the airport, code or description of carriers and details about the planes like type of planes, aircraft type, manufacturer and date when issue was reported in planes. Moreover, by using "join" feature, we can write queries to find out the best time to fly and number of people flying between different locations.

References

[1]https://www.cloudera.com/
[2]http://stat-computing.org/dataexpo/2009/
[3]https://www.ijcsmc.com/docs/papers/June2017/V6I6201764.pdf
[4]https://www.cloudera.com/documentation/enterprise/latest/PDF/cloudera-quickstart.pdf
[5]https://www.cloudera.com/documentation/enterprise/5-9-x/PDF/cloudera-introduction.pdf
[6]https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#Language_Manual DDL-Dynamic Partitions
[7]https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL+BucketedTables
[8]https://www.cloudera.com/documentation/enterprise/5-9-x/topics/introduction_compression.html
[9]https://www.cloudera.com/documentation/enterprise/5-9-x/topics/impala_compute_stats.html

Working List

No.	Task	Assigned To	Status (Date) - 2018
1	**Gathering Datasets**	Bhargavi Deshpande	Completed
2	**Cloudera quick-start and Impala**	Team	Completed
3	**Data Preprocessing with pig/Spark**	Ayushi Patel	In Progress
4	**Working with Hive vs. Impala or both**	Team	11/10 - 11/15
5	**Hive/Impala partitioning and clustering**	Ayushi Patel	11/10 - 11/16
6	**Data Compression, tuning and query optimization**	Bhargavi Deshpande	11/17 - 11/20
7	**Using database views to represent data**	Ayushi Patel	11/21 - 11/24
8	**Visualizing data using Microsoft Excel Via ODBC**	Bhargavi Deshpande	11/24 -11/25
9	**QlikView (Extra)**	Team	11/26 - 11/27
10	**Report**	Team	11/28 - 11/29