Flight Data Analysis Using Big Data Tools

ANVESH NAYAN LINGAMPALLI[1,*]

[1] *School of Informatics and Computing, Bloomington, IN 47408, U.S.A.*
[*] *Corresponding authors:anveling@indiana.edu*

project-S17-IR-2016, September 10, 2017

Analysis of flight data provides insights on the United States of America's Airline data by using Hadoop in the cloud environment. The On-time performance of flights operated by large air carriers are tracked and made as a report, Air Travel Consumer Report, which is a big data set. Hive component of Hadoop ecosystem, is utilized to process the big data in distributed environment. Efficient accessing and processing of the user queries is acheived by this analysis on flight data.

Keywords: Apache, Hive, Ansible, Pig, I524

Report: https://github.com/cloudmesh/sp17-i524/tree/master/project/S17-IR-2016/report/report.pdf
Code: https://github.com/cloudmesh/sp17-i524/tree/master/project/S17-IR-2016/code

1. INTRODUCTION

Real world data is large and growing exponentially from several years. This data can be in structured or unstructured format which is popularly known as Big Data[1]. Aviation industry manages enormous amount of data, which consists of the information regarding the delayed, cancelled, diverted or on-time flights by large air-carriers[2]. This is essentially a big-data set, where statistics are publicly available as the Air Travel Consumer Report.

Access to multiple clouds is provided by a cloud manager known as Cloudmesh Client[3]. With the help of this cloud manager, Hadoop cluster is built with neccessary add-ons such as Apache Spark[4], Hive[5], Pig[6]. Cloudmesh Client is also used in the deployment of the cluster on various clouds such as Chameleon cloud[7] or Jetstream[8] cloud. Deployment part is automated with the help of ansible[9] scripts where data is extraced, stored and analysed automatically to produce the results.

Big Data analysis of this data will provide a consistent understanding and importance of the given data. With 35 million flight departures per year, data is critically important for any planning decision made by airlines and airports. The results of analysis has benefits which can help airline operations to predict and reduce redundance[10].

2. INFRASTRUCTURE

Cloudmesh client and Chameleon cloud forms the infrastructure of this analysis project.

Cloudmesh client is a toolkit which provides a client interface for accessing different clouds and clusters. It includes a commandline interface to provide abstraction from backend databases. Simplicity is one of the key and powerful features of the cloudmesh client. It makes switching between various clouds easy by providing a convenient programmable interface.

Chameleon cloud is a large scale platform which is an open research community for development of programmable cloud services. It provides wide range of services such as Infrastructure-as-a-service, platform-as-a-service and delivery of high functioning cloud environment.

3. DEPLOYMENT TOOLS

3.1. Ansible

Ansible is an open source software that provides automation for configuration management and application deployment. It facilitates a simple automation platform that makes the application easier to deploy. It also handles ad-hoc task execution and multinode orchestration. Ansible is a software which has an agent-less architecture. This is because there are no deamon processes running in the background. Components of Ansible comprises of modules, playbooks,inventory and ansible towers. Modules can control system resources like services, packages, or files. Inventory is configuration file that reflects the nodes that are available for access. Nodes are represented by hostnames or IP addresses. Playbooks are Ansible's configuration, deployment and orchestration language. These are in the YAML format. Playbooks are generally used to manage configurations and deployment on remote machines. Ansible tower makes Ansible a center for automating tasks by providing a web based console[11].

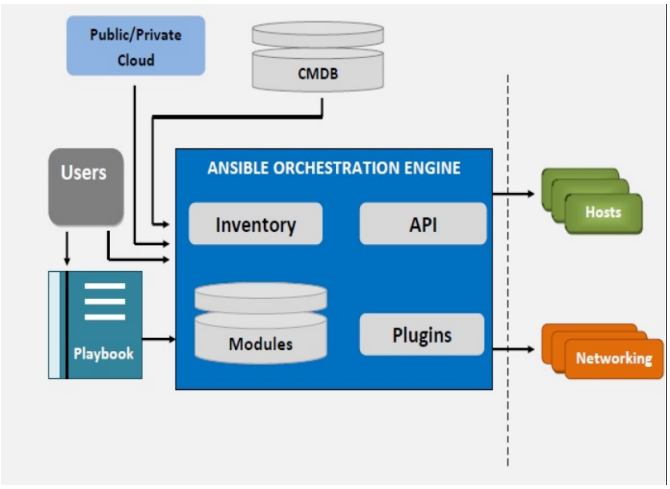

Fig. 1. Architecture of Ansible [12]

4. ANALYSIS TOOLS

4.1. Apache Hive

Hive is one of the ecosystems in Hadoop[13] framework which is built to analyze the data on hadoop cluster. Syntax of Hive is based on SQL[14], which is also known as HiveQL. MySQL or PostGreSQL[15] can be used for implementation of the queries. Hive provides tools which enable easy data extraction, transformation and data loading. Files can be stored in Hadoop Distributed File System(HDFS)[16] and accessed by Hive efficiently.

Schema of the Hive tables is stored in Hive Metastore. Metastore holds the information about tabes and partitions which are present in the data warehouse. In Hive the default Metastore is DerBy Database, which is a relational database management system provided by Apache Software. There are two components of Hive, HCatalog and WebHCat. HCatalog is a storage management layer for Hadoop which provides data processing tools such as Pig and MapReduce. WebHCat provides a service that is used to run Hadoop MapReduce, Pig, Hive jobs using REST interface[17].

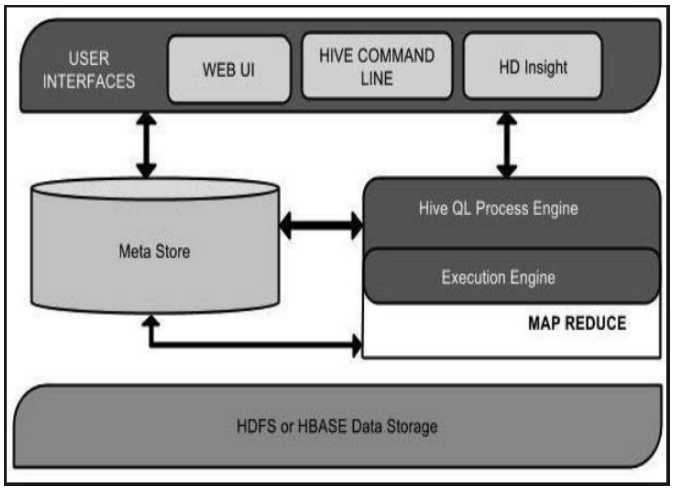

Fig. 2. Apache Hive Architecture [18]

Hive's SQL provides the basic SQL operations, such as,

- Filtering the rows from the table using WHERE clause.
- Selecting columns from table using SELECT clause.
- Joining two tables
- Aggregations of the data using 'group by' clause.
- Storing the results in a hdfs directory.

4.2. Hadoop Distributed File System

Hadoop Distributed File System provides a distributed file data storage system which spans large clusters of servers. It distributes -the storage and computation across many servers which maintains economy of the storage[16].

The file system is designed to be fault-tolerant. When HDFS takes data, the information is broken into pieces and distributed them to different nodes in a cluster. This provides parallel processing on clusters. MapReduce programming model is implemented when the applications are executed.

HDFS uses master/slave architecture, where each cluster consists of a Namenode, which manages file system operations and supports Datanodes, which manage data storage on individual compute nodes. Namenodes monitor the Datanodes in creating, deleting and replicating data by mapping them into data nodes.

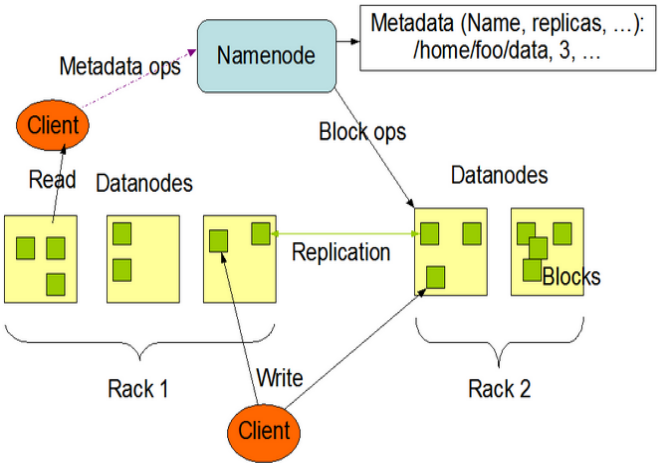

Fig. 3. Architecture of Hadoop Distributed File System [19]

5. IMPLEMENTATION

Implementation of Hive to perform data analysis consists of the following steps.

- Virtual machines are created on Chameleon Cloud with the help of Cloudmesh Client.
- Hadoop cluster is deployed on Chameleon cloud.
- Flight Data is loaded into HDFS.
- Data is then transfered to Hive tables.
- Finally, analysis is done on Hive tables using HiveQL, this can either be done using Hive interface or by installing PostgresQL interface.

5.1. Analysis in Hive Query Language

The following are the queries in Hive QL, which are similar to SQL statements.

What are the total number of flights which are cancelled?

```
SELECT year, month, count(cancelled) as Cancelled-flight
FROM Airline
WHERE cancelled = 1
GROUP BY YEAR, MONTH
ORDER BY YEAR, MONTH
LIMIT 20
```

What are the total number of flights which are diverted?

```
SELECT year, count(diverted) as Diverted-flights
FROM Airline
WHERE diverted = 1
GROUP BY month
ORDER BY month
LIMIT 10
```

Similarly, analysis of the data is done where the queries are as follows.

What is effect of flight distance on cancellations?

What is the effect of flight distance on average departure delay?

What is the monthly average departure delay?

What is the yearly average departure delay?

6. TECHNOLOGIES

- Distributed Computation and Storage:- HDFS and Hive
- Development:- PostgreSQL and Java
- Deployment:- Ansible

7. BENCHMARKING

After the deployment stage, benchmarking is implemented. This process is important as it evaluates the performance of the application and the system. This project is implemented on the chameleon cloud and the results are observed. Different cloud environments are used for creting these benchmarks

Characteristics of Chameleon cloud on which the analysis is implemented are,

Chameleon	VCPU	RAM(GB)	Storage(GB)
m1.small	1	2	20
m1.medium	2	4	40
m1.large	4	8	80

Fig. 4. Different types of chameleon clouds

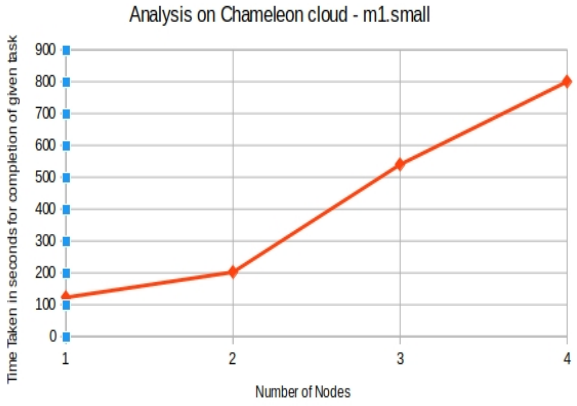

Fig. 5. Analysis time taken for a HiveQL-m1.small

The following are the observations from the analysis on chameleon cloud of different flavors.

These are the observed results when m1.small cloud is used for analysis As the number of nodes in a cluster increase, the amount of time the task takes to complete is increasing. This trend is also observed when m1.medium and m1.large flavors of Chameleon cloud are used.

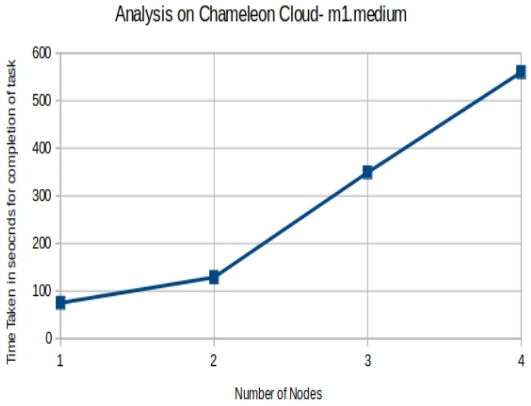

Fig. 6. Analysis time taken for a HiveQL-m1.medium

8. CONCLUSION

Deployment and the analysis of the flight data is implemented by using Apache Hive, Cloudmesh and Ansible automation. The results obtained from the qeries in Hive environment gives insights on the available flight data. Hive uses map and reduce functions internally which is taken care of Hadoop system. Observations such as number of flights cancelled, number of flights departed from an airport are made from Hive Query Language. From these results trends and patterns are observed which provide detailed analysis of the data.

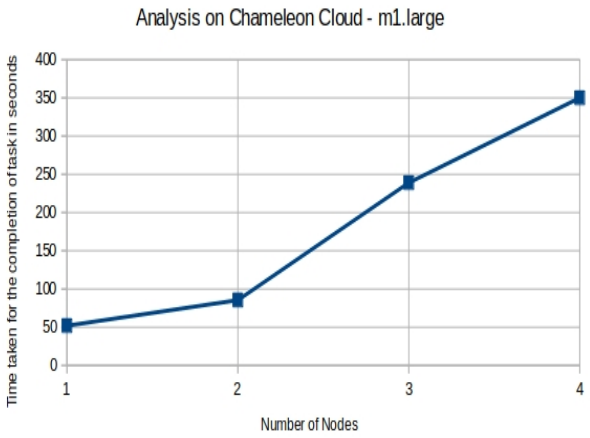

Fig. 7. Analysis time taken for a HiveQL-m1.large

9. ACKNOWLEDGEMENTS

This project is undertaken as part of the I524: Big Data and Open Source Software Projects coursework at Indiana University. We would like to thank our Prof. Gregor von Laszewski, Prof. Gregory Fox and the Associate Instructors for their help and support.

REFERENCES

[1] "Big data," Web Page. [Online]. Available: https://www.sas.com/en_us/insights/big-data/what-is-big-data.html
[2] "Aviation analysis," Web Page. [Online]. Available: https://aviationanalytics.com/airport-analytics/data-analysis/
[3] "Cloudmesh," Web Page. [Online]. Available: https://cloudmesh.github.io/
[4] "Apache spark," Web Page. [Online]. Available: http://spark.apache.org/
[5] "Apache hive," Web Page. [Online]. Available: https://hive.apache.org/
[6] "Apache pig," Web Page. [Online]. Available: https://pig.apache.org/
[7] "Chameleon cloud," Web Page. [Online]. Available: https://www.chameleoncloud.org/
[8] "Jetstream cloud," Web Page. [Online]. Available: https://jetstream-cloud.org/
[9] "Ansible webpage," Web Page. [Online]. Available: https://www.ansible.com/
[10] "Big data in aviation," Web page. [Online]. Available: http://apex.aero/2016/11/30/big-data-aviation-industry-case-becoming-data-driven
[11] "Ansible : Tutorial," Web Page. [Online]. Available: https://serversforhackers.com/an-ansible-tutorial
[12] "Architecture of ansible," Web Page. [Online]. Available: https://devops.com/ansible-automation-provisioning-configuration-management/
[13] "Apache hadoop," Web Page. [Online]. Available: http://hadoop.apache.org/
[14] "Structured query language," Web Page. [Online]. Available: https://en.wikipedia.org/wiki/SQL
[15] "Postgresql," Web Page. [Online]. Available: https://www.postgresql.org/
[16] "Hadoop distributed file system," Web Page. [Online]. Available: https://hortonworks.com/apache/hdfs/
[17] "Information on hive," Web Page. [Online]. Available: https://hortonworks.com/hadoop-tutorial/how-to-process-data-with-apache-hive/
[18] "Architecture of hive," Web Page. [Online]. Available: http://blog.cloudera.com/blog/2013/07/how-hiveserver2-brings-security-and-concurrency-to-apache-hive/
[19] "Architecture of hdfs," Web Page. [Online]. Available: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html#NameNode+and+DataNodes