CPSC 531 - Advance Database Management System

Final Project Proposal

Airline Delay Management System

Team members

Lency Lakhani

CWID=885196055

Email ID= lencylakhani@csu.fullerton.edu

Hetal Patel

CWID= 885868455

Email ID = hetal-patel.1994@csu.fullerton.edu

Mentor

Prof. Tseng-Ching Shen

Email ID = jtcshen@fullerton.edu

Problem statement

Data is a vital system component and should be processed efficiently within the specified time frame. In recent years, data has been remarkably grown (structured, semi-structured, and unstructured) from various sources, including social media, banking data, health data, agricultural data, and so on. This data can be in Zettabytes, and it is hard to maintain such massive data. Studying these datasets can assist in using the information as the foundation for new goods or services. One of these analyses focuses on the airline delay system. Around 20% of airline flights are canceled or delayed yearly, costing passengers more than $20 billion in money and time. For a pleasant user experience, the system should adhere to "ilities" such as maintainability, usability, availability, security, and performance. This project aims to create a model for analyzing flight system delays. It will assess the likelihood of any flight being more than 'x' minutes late using available software tools for analyzing and storing massive data.

Dataset

We have taken dataset from Kaggle website. We have taken data of Airline details of 2018. This data comprises the flight date, airline name, cancelled flights, diverted flights, departure city, arrival city, flight number, departure time, departure delay, taxi out, taxi in, wheel off, wheel on, arrival time, arrival delay, airtime, distance, tail number, etc. we are planning to analyze departure delay time in given number of minutes.

Architecture of Hadoop

Architecture of Hadoop has been divided into three component.1) HDFS (Hadoop Distributed File System) 2) Map Reduce 3) Apache Spark/Yarn. HDFS is used to store data in multiple computers by dividing big, massive data into a number of parts. HDFS has fault tolerance capacity by providing replication of each part of datasets. HDFS has two node one is **Namenode** and other is **Datanode.** Namenode has a master dataset. Datanode has distributed data of datasets. Moreover, Map Reduce helps to process and analyze big data. First of all, dataset is divided into number of part then data has been sorted among the parts. Furthermore, data has been reduced in the process of Map Reduce. Spark is an open-source framework to workload the distributed data.

Implementation guide

In this project, we will try to build OLAP (Online Analytical Processing) solution. We are planning to use Amazon Web Services for our project environment.

- We have taken dataset from Kaggle website.
- We are planning to Set up Hadoop on AWS using EMR, S3.

- We are using S3 for storage purposes.
- We want to build AWS EMR to analyze and process data from S3.
- We will run the Spark job on the EMR and process the tasks to the EMR core nodes for analysis and we will end up with results which, we are looking for.

Technical Perspective

- **Hadoop**

  Hadoop is an open-source java framework which can help to analyze big data by dividing data into multiple storage clusters. Moreover, Hadoop follows HDFS (Hadoop distributed file services) which helps to sperate one Hadoop cluster into multiple clusters. Hadoop comes with Map Reduced programming model which helps to provide distributed processing among the number of clusters in Hadoop.

- **Amazon Web Services**

  Amazon Web Services is a web service for cloud computing to analyze gigabytes of bigdata. Amazon comes with S3, EC2 and EMR services.

- **S3**

  S3 is a storage infrastructure to provide scalable and secure to analyze big, massive data. S3 provides security from unauthorized users. S3 gives access to use data anywhere. S3 has fault tolerance capacity because it saves the replication of each data part to each other.

- **Amazon Elastic MapReduce**

  Amazon EMR is an Amazon Web Services tools used for analyze and process big distributed datasets. EMR works as Map Reduce where, dataset is distributed in multiple parts and analyze solution at the end.

Technologies and tools to use:

- Hadoop
- Spark
- AWS services like EC2, EMR etc.

References

- https://hadoop.apache.org/
- https://aws.amazon.com/console/
- https://www.kaggle.com/
- https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022?select=Combined_Flights_2018.csv