**Hive Data Analysis for Issues faced by people traveling through different Airlines:**

**Introduction:**

Nowadays, people tend to travel by airplanes because they are way faster and less prone to accidents than the other means of travel. Most travelers easily reach their destination within a few hours and seldom it takes more than 24 hours. Sadly, in the past few years, air travel has become more problematic and hence less appealing to the passengers. One of the main reasons for this is rising fuel prices, which directly means a hike in the ticket prices. Another main reason can be tighter security regulations that means passengers have to wait for longer duration.

We will be using Hive to analyze a few such problems. Hive is used for analyzing organized data. The main reason for using hive for analyzing is because it is fault-tolerant and is a very handy tool for big data analysis and interpretation. We will be using SQL(Structured Query Language) for communicating with the database and collecting the required data needed for analysis. HiveQL is similar to the very famous SQL that interacts with the Hive database for this particular problem for analyzing necessary data in a very structured format. Many organizations such as Airbnb and Guardians use Apache Hive to solve such problems within their organization.

By analyzing the dataset, we can provide solutions to both people as well as the airline industry. People will be able to have a comfortable journey and by providing them with such services, airlines will be able to generate more revenue and result in higher customer satisfaction too.

**Problem Statement:**

The dataset that we will be providing information regarding different aspects such as: prices, duration, facilities, age, etc.

We will be using this data set to solve the following problems faced by the different customers as well as airlines too:

1. How many times each passenger travelled?

We are analyzing this because it will help the people in the airline industry to recommend them more often through mails, messages so that one can easily book that flight without searching through different sites.

1. Why is there a difference in the number of passengers travelling in each flight even though all the facilities that are provided are similar? What is one such facility that matters to the passenger the most that is making them change their mind?

This will help the airlines to come up with a solution so that they can increase the number of passengers for their airline.

1. Many times young passengers need window seats and elderly want to sit where there is no trouble for them. So we have to arrange according to the age to provide them the best match. This will help the airlines to provide comfortable journeys to the people from different age groups and hence make it less troublesome for them.

**Methodology:**

Size of data has been increasing rapidly from past few years making it difficult to analyse the data. Traditional warehouse are very expensive to use. Hadoop makes it easy to store and process large sets of data which is written in java language. Hive is a core component of hadoop.Hive is a data warehouse infrastructure that is built on top of Hadoop. It uses map reduce and HDFS for processing and storage. In 2006 Facebook switched from oracle database that uses sql queries to handle the database to Hadoop that understands only map reduce to handle big data. So now they needed interface that can convert sql to map reduce program and run it in cluster and show result. Hive is an interface created by Facebook. And later it was donated to apache as a top-level project. Hive is very much similar to Sql.

**Advantages of hive:**

1.It can be used as an ETL.

2.Provides capability of querying and analysis.

3.Can handle large data sets.

4.Can do sql operations (filters, joins, group by, order by).

In this project we have chosen the domain flights and performed some analysis on the dataset. First we have created the dataset and tables named indigofeb2019, spicejetfeb2019, vistara1, vistara3. We also showed how joins work in this project.

After creating the tables we have loaded the data into tables.

Some of the challenges which we observed in passenger data are:

1. How many times each passenger travelled?

We are analyzing this because it will help the people in the airline industry to recommend them more often through mails, messages so that one can easily book that flight without searching through different sites.

2. Why is there a difference in the number of passengers travelling in each flight even though all the facilities that are provided are similar? What is one such facility that matters to the passenger the most that is making them change their mind?

3. Many times young passengers need window seats and elderly want to sit where there is no trouble for them. So, we have to arrange according to the age to provide them the best match. This will help the airlines to provide comfortable journeys to the people from different age groups and hence make it less troublesome for them.

After all the tables are loaded with data, we have written queries to solve those challenges.

**Query to the first challenge:**

Select count(\*), pname from indigofeb2019 group by name;

**Query to the second challenge:**

Select count(\*) from indigofeb2019;

Select count(\*) from vistara1;

**Query to the third challenge:**

Select bookingid,pname,page from indigofeb2019 order by page desc;

Select bookingid,pname,page from indigofeb2019 order by page asc;

**References:**

<https://www.analyticssteps.com/blogs/what-hive-big-data-and-its-benefits>

<https://www.ijcsmc.com/docs/papers/June2017/V6I6201764.pdf>

<http://www.lifeisafile.com/_posts/2015-08-23-flight-analysis/>

<https://www.washingtonpost.com/travel/tips/unofficial-rules-every-seat-plane-aisle/>

<https://www.ijser.org/researchpaper/HIVE-Processing-Structured-Data-in-HADOOP.pdf>