Team 4 Final presntation

15146311 Jun Nyeong Nam

16102284 Sung Ho Lee

18102077 Jung Ha Park

17102070 Jae Hun Hwang

-page 3 (introduction)

Our team was looking for a topic suitable for big data management, so we decided to deal with flight data at domestic airports from 2017 to June 2019. We give you brief explanation of this data, the data includes the departure time information, arrival time information, airline, flight name, delay status, scheduled time, actual arrival time, cancellation, reason for delay, etc. are included The total data is about 987,000.

-page 4 (introduction)

We are going to analyze airplane delays and cancellations caused by various factors such as airports, airlines and seasons based on the flight data, and show the probability of delays or cancellations based on the analysis.

-page 5(introduction)

Motivation: Continuously increasing demand for air traffic services has also increased inconvenience to passengers due to delays and cancellations of aircraft.

Expected Effect:

1. Increase user benefit by showing previous flight data more intuitively and efficiently.

2. Help airlines and airports become more efficient and cost-effective.

-page 6

We used mysql to process the query. Tableau was used to visualize based on this result and all these results were posted on the web using php and ajax.

-page 7

Then let me explain our service process in more detail.

1. Process the initial data table using **mysql**

2. Using **mysql**, test all data table and connect with **tableau**

3. The data are visualized through **tableau.**

4. The results are shown in the form of bulletin boards in Web services created

using **PHP and AJAX**.

5. The web service also provide fields to type seasons, hours, airlines, airports,

etc., and show the probability of delays and cancellations based on entered

information.

-page 8

We put the file AFSNT.csv in table In order to manage data, we put the csv files in each table about the reason for delay, airport information, airline information.

-page 9

We came up with the desired result through the join method of four tables. In this process, the actual time of the planned time was saved as (yyyy-mm-dd-hh-mm-ss)date type in the new column on the existing AFSNT table, and the columns that were difficult to identify, which is the reference airport and the relative airport sorted by departure and arrival, were stored in the origin arrival column. And we got the delay time, put it in the delay time column, also we put it in the delay group.

-page10

From now on, I will explain while looking at the actual mysql code. We tried to analyze the original flight data by identifying number of delays caused by various factor (annual, seasonal, monthly, day of the week, time, delays by Delay Code). First, we changed the data’s columns (Airport, Airline, Delay reasons)

-page 11

* ‘Date’: Merge ‘Year’, ‘Month’, ‘Day’ columns and change data type ‘Datetime’ for calculating
* ‘Origin’ & ‘Destination’: The original data is a mix of departure and arrival airports, so we separate these data for recognize easily.

-page 12

* ‘Scheduled Time’: Using ‘Origin’ & ‘Destination’, we remake scheduled time at ‘scheduled’ to calculate finding out how late it was.
* ‘Actual Time’: Using ‘Origin’ & ‘Destination’, we remake scheduled time at ‘scheduled’ to calculate finding out how late it was.

-page 13

* ‘Delay Time’: Using ‘Scheduled time’ and ‘Actual time’, we calculate how long the flight has been delayed.
* ‘Delay Time Group’: Make a group ‘0 ~ 30’, ‘30 ~ 45’, ‘45 ~ 60’ etc. to show how serious the delay is.

-page 14

* We made several SQL queries statements to analyze and visualize the data as we did earlier. For example, we created seasonal delay rates, time-to-time delay rates, and so on, etc. But we've come to use 'tableau' for 'better' data visualization than using php and JavaScript!

-page15

* First, we link MySQL and Tableau for visualization, so we can use pre-processed database before pages.
* And we can make graph like under picture using MySQL database.

-page16

* To show the delay result, we make homepages input user’s data and show result using database.
* To show our analyzed data and visualization using tableau, we make homepages with analysis by categories.

-page 17

We will tell you about the algorithm that runs the homepage.

Index.php

* Platform that connects various homepage
* Input data and hand over these data that will be use to query at ‘flightno.php’

flightno.php

* Showing result ‘delay rate’, ‘estimated delay time’, table of the number of delays by delay time group etc using post ‘index.php’ and running query.
* If you want to know why this result is showed, you can see the whole data table made by query

Airline.php, delayreason.php airport.php

* We used it as a php to post the results of classifying the data by airline by airport and other reasons for delays by season and year-by-year.
* I also expressed more results in tableau.

Showmore.php:

* Showing a whole data table which is made of ‘flightno.php’ query

-page 18

We want to show analyzing the original flight data at web, so we make five homepages.

* + ‘index.php’: This home page is our base homepage. You can go to another homepage to use navigation bars or click button under the homepage. Moreover, you can input data your flights schedule, and we will post input data and show a result to analyze probability, the most delayed reason, etc, at homepage

-page 19

* Flightno.php: We make box which is showed result of SQL queries to count the number of columns in conformity with queries and calculate delay rate
* We make box which is showed received data at ‘index.php
* We make box which is showed result of SQL queries the top 3 reasons and time group

-page20

* Airline.php: This page shows various statistical results related to airlines.
* Airport.php: This page shows the various statistics related to the airport.
* Delayreason.php: This page shows various statistical results related to the reason for the delay.

-page21

* The data.php have a 3-table containing the actual name corresponding to the code (airline, airport, delay reason). This table is used in index php to change the result (code, symbol) to a real name for users to easily recognize.

-page22

* To take care of user’s input, we had to consider several issues. In this case, user might select the route which not actually exists. For example, flight originated from Sacheon (사천) airport has only one route to Kimpo (김포) airport. To handle the problem, we use Ajax technique. We send chosen origin airport to ‘getdest.php’ via getDest() function, and search available destination for the origin. As a result, users are only able to choose destination within existing route information.

-page23

* We also had to consider date of user’s input. Since we only have data between 2017 and 2019, we could not allow users searching out of range dates. In addition, even though the months are always between January to December, the days available for each month are different. (February has 28 days and June has 30 days etc.). Also if leap year (윤년) is involved, the length of February becomes 29 days. To tackle the problem easily, we decide to make user choose days by selected year and month. Specifically, we send year and month to ‘getday.php’ getDay() function, and create options of select tag on ‘index.php’ based on the result come from query getting available days.

-page24

* This php code is a user input(departure airport, arrival airport, date and time) processing query that lists the most frequently appeared flight delay data of same departure, arrival airport +-3days and +-3hours based on the input and show the most frequently featured delay group, Top3, most frequently featured delay reason top 3

-page 25

* This php code informs users when there is no same data in our table corresponding user input

-page 26

* This code inform user Indicates whether the flight corresponding to the user input has been delayed, if delayed, the corresponding delay group, the proportion of flight data being delayed during the results of the 23page php code.

-page27

* Page23, page25 php code result

-page28

* we show visualizations of data using Tableau in front of pages, through 'airline.php’, ‘airport.php, ‘delayreason.php’. For these, we upload visualizations ‘Tableau public server’ and call up the saved contents and show them on our homepage.

-page29~31

29 to 31 pages is a more clearly visual representation of delayed data classification by airport.

The first picture is a bar graph showing the delay/delay rate and time per airport.

The second picture shows the departure delay time by route through the size of the circle.

The third photo shows the arrival delay time by route through the size of the circle.

The fourth photo showed the delay ratio by airport route.

The 5th photo shows the delay detail factor ratio by airport in circle graph.

The last photo shows the proportion of delay/cancellation factor groups by airport in a pie chart.

-page32-35

The 32-35 pages are a more clearly visual representation of the results of each airline's classification of delayed data.

The first picture shows the airport usage rate and delay rate of each airline in a one graph.

The second photo shows the frequency of the delay group by airline.

The third picture shows the delay detail factor ratio and delay rate of each airline in a circle graph.

The fourth photo showed the delay rate by airline by day.

The 5th photo shows the average number of flights, delays, and delays per month by airline.

The 6th photo shows the delay rate by group, which is the delay factor by airline.

The last photo shows the delay rate by group, which is the delay factor by airline.

-page36

The 36 pages are a more clearly visual representation of the results of each airport's classification of delayed data.

The first photo showed the delay rate by delay detail factor.

The second photo shows the delay time by detail factor per month.

-page37~40

**The steps of analyzing the problems and debugging them**

Initially, our team was targeted to make service to predict the probability of future flight delay or cancellations based on past data through tensorflow.js. However, there were many difficulties with only the data we had. Firstly, we could not confirm that the input data by the user was the actual flight data. Because we cannot to get flight information from the actual flight reservation service (such as sky scanner). Second, to predict delays or cancelations of flight for a specific day, numerous factors must be considered such as weather, wind, rainfall. But it’s difficult to interconnect a lot of data files. So, we decided to bypass by getting user input within our data and providing information of delay based on our analysis not prediction. Although current service is simply to show past flight delay information for the user input, but it can be developed into more useful service if we change the server language, build predictive models, and link more data together.

<Try to Use Hadoop>

We putted original csv file as a table from local to hdfs using sqoop, and then tried to preprocess the table using impala *alter table, update set* commands. However, the *update set* command was only supported in Impala 2.8 or higher versions, while the impala in our Cloudera environment was 2.2.0 version. So, we tried various solutions such as updating Impala and installing the cloudera trial version. But it was hard to achieve. Also, Impala does not support datetime data types, but our main parts of preprocessing are the calculation and merging of date and time. In the process of importing the table to hdfs through sqoop, importing did not work because the date data was not recognized in table. Therefore, we use mysql for both data preprocessing and executing query.