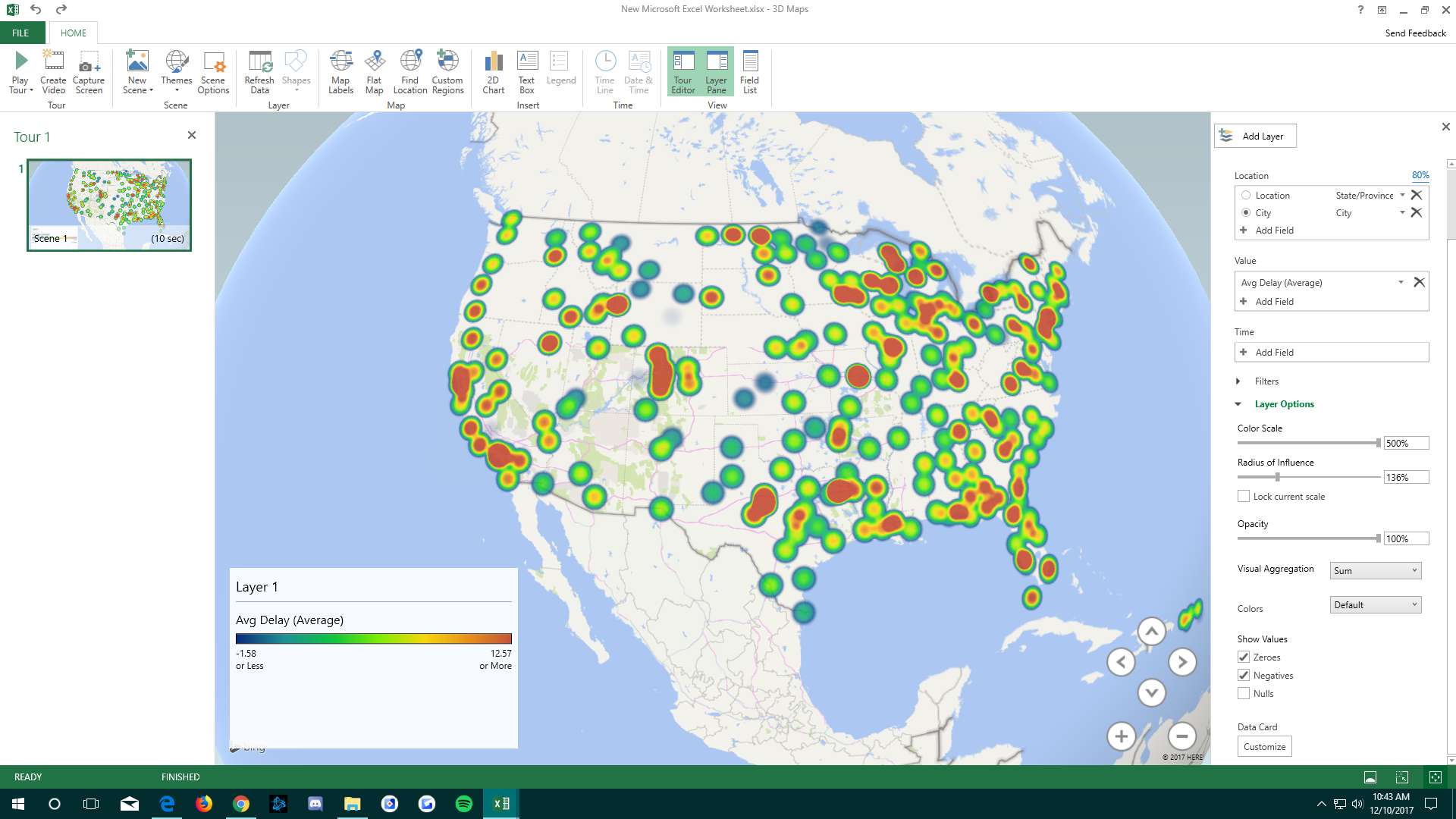
**Air Traffic Statistics and Predictions**





**Abstract:** The purpose of our research was to compile Air Traffic Statistics in order to do a comparison between all the different carriers and airports in order to predict which of these best and worst based on the typical trends extrapolated from the data. The United States was chosen in order to narrow down the information while at the same time keeping the information really relevant for the authors and people living in the United States. In order to conduct our research we used a 2 node cluster running on version IOP 4.2, in Big Insights. Our implementation of Big Insights allowed us to handle and deal with a very large set of information and formulate our analysis through what we saw in the information that was gathered. The data discovered revealed several trends, some that would be extremely useful for people traveling by airport.

**1. Introduction**

Our research led us to several air traffic statistics spanning from the 1990’s to the present, with information and statistics recorded down to the month on RITA. While browsing the information we decided this would be the ideal data set to work with because of its size and found that the information available would prove to be invaluable to most if not all people traveling. Convenience is such a huge issue when it comes to traveling especially when it comes to traveling by the often overcrowded airports, we discovered information that may help make traveling by airport feel less of a drawback on people’s vacations. Prior to our research we would just book flights based on price without much more thought into it, however moving forward we are confident that we can see past the price of the plane ticket moving forward. Many people do not realize that by booking on specific days or times and with certain airlines can really assist in preventing any kind of derailment from your strict itinerary because history often proves through trends that certain things happen time and again. Knowing the best days to travel or which airports and carriers tend to have delays can help let a traveler know if you should avoid the airport or carrier all together depending on the amount of time you can adjust in your itinerary. It can help prevent a chain reaction which may or may not ruin your trip all together. With this knowledge it also allows travellers to make much more informed decisions specifically for anything regarding the airport. Plans in regards to flying with specific carriers or airports and making all while making the optimal choices based on the statistics will not always guarantee a satisfied flight, it will greatly increase your odds of having one.

Section 2 touches on our analysis model and describes more in detail of how we chose to approach the analysis of the data through our use of Big Insights.

**2. Analysis Model**

Our approach to analyzing the data can be broken down into six major steps: Data Gathering, Extraction of compressed data, compiling of data, platform through Big Insights, table creation through HiveQL, and our analysis through the information in front of us. Each step will be touched on below, and is visualized in figure 1 below.

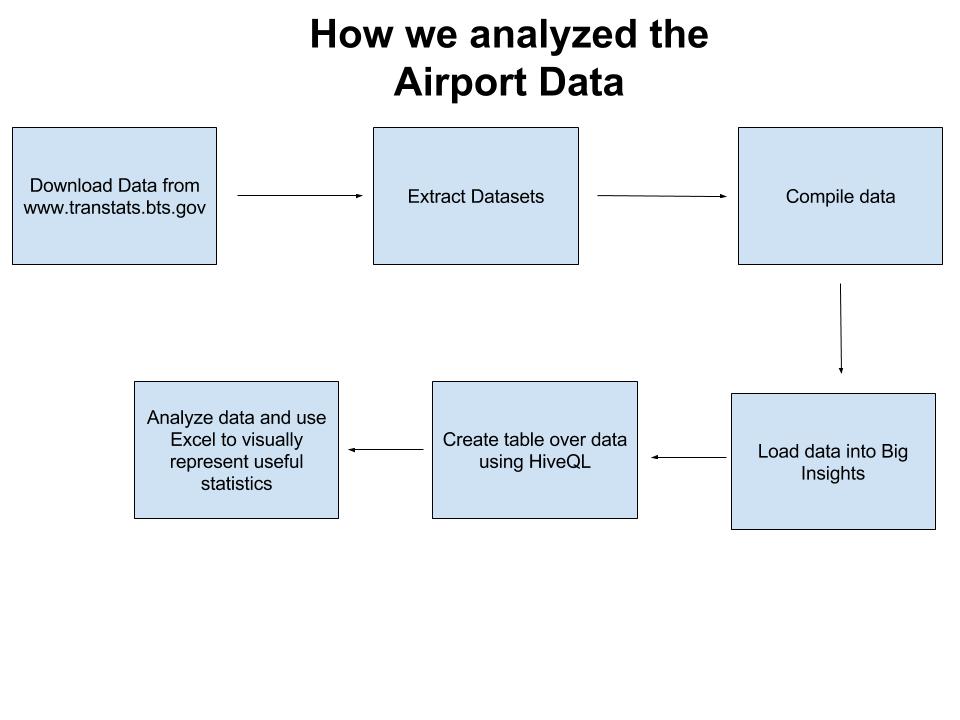
The first step I mentioned was data gathering, and this step was an extremely difficult step. This is so because of the sheer amount of data available across the vast and many platforms. We knew from the start what data we wanted to create analysis for and that was through our interest in RISC. Searching through different data repositories and warehouses was no difficult task, going through sites such as kaggle, google aws, socarata, and then finally settling with Data.gov. Data.gov linked us to the transtats website shown in diagram 1. After downloading the data we had to decompress the file.

The second step we took was the extraction of compressed data. The files we downloaded were compressed because of the sheer size of the information, and in order to do so we had to extract the data in order to narrow down the years to giving us the opportunity to compile them in specific years in order to do comparative analysis. We opted to work with the latest couple of years to keep the data relevant in the present day because over time different trends come and go for different reasons. As soon as we chose the years, we then proceeded to compile the data.

To compile the data we chose to group the data by year. We had to search through each dataset to confirm each of them had the information we were looking for. As soon as it was confirmed, we proceeded to clean the data by removing any null columns and any properly labeling the important ones. We just made sure all the important aspects of data cleaning were checked: Validity, accuracy, completeness, consistency, and uniformity. This was an extremely important step so the data was much easier to navigate when uploaded into Big Insights. After cleaning and compiling the data, we then proceeded to upload the data to Big Insights.

Uploading the data to Big Insights was the much more straightforward step. We ran the typical queries -wget -O year (to download the data on the big data platform), -tar (in order to extract the data), -mkdir (to create a directory for the data in order to make the data much more easily accessible, and finally we ran to -ls commands to confirm that the data was in the correct directory. After confirming all the data was uploaded we proceeded to run Hive in order to create a table to structure the information.

In order to create the table we had to choose all the relevant columns in each table so we could add structure to all of the information. By structuring the data we made it much more digestible to each of the group members who wanted to run queries. We structured the data through making each column definitions specific in order to narrow our research target, specifying the file formats, and data types. After creating the tables we then proceeded with our analysis which will be touched on.



(Figure 1: Our analysis depicted.)

**3. Data Analysis**

Air traffic data usually consists of a collection of flight trajectories of different aircraft. Each flight trajectory usually contains information about the type of aircraft, origin and destination airports, followed by a series of entries that records the time, location, and altitude of the aircraft. The flight tracks are usually recorded in 10 second intervals. Other information such as date, heading, velocity, are generally recorded as well. This data set is comprised of uniquely identified flight paths, each containing latitude, longitude, and altitude information at 10 second intervals for the duration of the flight within the area of interest..

In this paper, we are interested in analyzing air traffic congestion and air traffic delays. Based on our data we can surmise that the lowest average departure time was on Friday. The highest departure delay was on Wednesday and Thursday, followed by Sunday, Saturday, Monday then Tuesday. The lowest average arrival delay is on Friday as well. The highest arrival delay being Thursday, followed by Wednesday, Sunday, Saturday, Tuesday and then Monday. top of the first page, followed by the authors’ names and their affiliations. Long title should be typed on two lines without a blank line intervening. Leave approximately 1 cm (0.39 in) between the title and the body of the first page.

**Abstract:** Type the abstract at the beginning of the first column. The abstract should be no longer than 200 words.

**Text:** Begin typing the main body of the text immediately after the abstract, observing the two-column format as shown in this example.

**2.2 Sections**

**Headings:** Type and label section and subsection headings in the style shown on these pages. Use numbered sections, in order to facilitate cross references.

**References:** Citations within the text appear in brackets as [ref. number]. Gather the full set of references together under the heading **References**; place the section before any **Appendices**, unless they contain references. Arrange the references in the order that they are cited in the text. Provide as complete a citation as possible, using a consistent format.

**Appendixes:** Appendixes, if any, directly follow the text and the references (but see above). Letter them in sequence and provide an informative title: **Appendix A Title of Appendix.**

**2.3 Footnotes**

Put footnotes at the bottom of the page. They may be numbered or referred to by asterisks or other symbols.[[1]](#footnote-0) Footnotes should be separated from the text by a line.[[2]](#footnote-1)

**2.4 Graphics**

**Illustrations:** Place figures, tables, and photographs in the paper near where they are first discussed, rather than at the end, if possible. Wide illustrations may run across both columns.

**Captions:** Provide a caption for every illustration; number each one sequentially in the form: "Figure 1. Caption of the Figure." "Table 1. Caption of the Table." Type the captions for figures below the figures. Type the captions for tables above the tables.

**3. Length of Camera-ready Manuscript**

For the length of camera-ready manuscripts, **a paper is limited up to 4 pages**. All illustrations, references, and appendices must be accommodated within these page limits. Any extra page beyond the first four pages will be deleted. **Please DO NOT put a page number in each pag**

**4. Submission Process**

1. Format your paper using this template.

2. Turn the hardcopy by Dec 4th before the lecture starts

### **References**

[1] T.A. Jones, “Writing a good paper,” *IEEE Trans. on General Writing*, Vol. 1, no. 2, pp.1-10, May 2002.

[2] K. Hwang, *Computer Arithmetic*, John Wiley, 1997.

1. This is how a footnote should appear [↑](#footnote-ref-0)
2. Note the line separating the footnotes from the text [↑](#footnote-ref-1)