Impact of Weather on Taxi Transportation in Chicago

Divyansh Anand
School of Computing
National College of Ireland *Dublin, Ireland*
x22240217@student.ncirl.ie

***Abstract*—This research explains the complex relationship between Chicago's weather and taxi services, using techniques such as Hadoop and MapReduce to filter through large datasets. This report focuses on four key aspects taxi demand, journey duration, fare dynamics and seasonal changes.**

Keywords— Data processing, Hadoop, Python, Map-reduce

I. INTRODUCTION

Weather data is always interesting to explore and analyze, so many things including the retail sector, farming sector, medical sector, transportation sector and so on are affected due to changes in weather. One can observe behavioral change and different patterns in humans when they are involved in the above-stated sectors under different weather conditions. For instance, with variations in weather, different types of clothes and merchandise are sold in the retail sector. Farming season in the USA starts in the spring and ends in the fall season, crops and farmers depend on the rainy season each year for high yielding. The medical sector is sensitive to weather variations as they influence patient visits. People are more prone to viral and flu during the winter season than in summer, respiratory problems rise due to changes in air quality. Additionally, seasonal changes may bring allergies and infections.

Another sector that is significantly impacted due to weather conditions is the transportation sector. Road and rail transport services can be disrupted due to heavy snowfall and rainfall. Dense fog or thunderstorms can lead to flight delays and cancellations thus affecting air transport.

Out of these, it will be intriguing to explore and understand how weather can affect taxi trips for a city, have also studied from a paper.[1] There are so many factors that correlate with Taxi trips and weather conditions. One can compare taxi trip demand during different seasons, as well as the impact of weather on duration and fare. Additionally, one can analyze and understand the pattern between taxi trips and weather, so that availability can be maintained throughout the year. Here in this research paper, we will further discuss the impact of weather on Chicago taxis.

Some objectives that this paper will discover while comparing weather data with taxi trip data are as follows:

A. *What effect do weather conditions have on taxi demand in Chicago?*

By comparing weather data with taxi trip data one can understand how changes in weather can influence the demand for taxi services. For example, this research investigates whether the demand for taxis rises during severe weather such as heavy rain or snow.

B. *What impact do weather conditions have on taxi trip duration in Chicago?*

The duration of taxi rides can be considerably influenced by weather conditions. When we compare weather data with trip length data, we can see how bad weather, traffic jams, or slippery road conditions can contribute to extended travel times. Affecting both passengers and drivers.

C. *What impact do weather conditions have on taxi trip fare in Chicago?*

Weather conditions may impact taxi fares. For instance, extreme weather conditions may boost the demand for taxis in a particular location which may lead to surging in prices.

D. *How do weather conditions affect drop-off/destination location?*

Weather patterns may impact drop-off or destination locations, commuters may choose alternate destinations like shopping malls or transportation hubs where they can find shelter. One more insight that can be observed here would be does sunny weather leads to longer trips or drop-off locations far away from the city or not. This information can also help city planners to plan constructions according to the behavioral pattern.

E. *Understanding the trend between seasons and taxi trips.*

We can understand the trends between different seasons and taxi requirements. For example, in the winter season taxi demand may increase due to difficult road conditions and freezing temperatures whereas, in the summer season demand may be higher for recreational purposes (outdoor activities like beaches, concerts and festivals, boating, farmers market, etc). Recognizing these tendencies one can assist taxi services in optimizing their operations all year. One more thing to work

II. DATA

A. *Chicago Taxi Trips Data*

The dataset contains information on taxi trips in Chicago from 2013 onwards till present, reported to the city's regulatory agency. To ensure privacy while allowing for aggregate analysis, the dataset hides specific taxi medallion numbers, suppresses some Census Tract data, and rounds trip times to the nearest 15 minutes. Not all trips are reported, but the city believes that most of them are included in the dataset. For our research we have taken data from 2020 onwards and extracted from public datasets using big query.

Columns of interest: trip_start_timestamp, trip_end_timestamp, trip_seconds, trip_miles, fare, tips, tolls, extras, trip_total, payment_type.

Importance of columns: It will help us calculate the duration of the trip, demand for a taxi on that day, expenses for the trip on that particular date (using trip start timestamp column).

B. *Weather Data*

All of the daily weather observations from weather stations throughout the world are compiled into the Global Surface Summary of the Day (GSOD) dataset. It contains data on temperature, precipitation, wind direction and speed, humidity, atmospheric pressure, and other things. Researchers studying climate change and meteorology will find great use for this dataset, as will companies and organizations that use weather information for a range of purposes, such as energy management, transportation, and agriculture. It offers an extensive global database of everyday meteorological conditions, facilitating a multitude of studies and research projects. For our research we have taken data from 2020 onwards and extracted from public datasets using big query.

Columns of interest: stn, wban, date, year, visib, prcp, sndp, fog, rain_drizzle, snow_ice_pellets, hail, thunder, tornado_funnel_cloud.

Importance of columns: It will help us understand on which days (using date column) fog, rain, snow ice pellets, hail, thunder and tornado were there.

With the objectives proposed choosing above 2 datasets looks promising to accomplish our objectives.

III. METHODOLOGY

In this research, after copying data from hdfs directory to the data frames, several checks have been performed on the data to ensure data quality such as checking for null values and checking for redundant data.

A. *Checking for null values*

In the NOAA GSOD data-frame this research found 3 columns namely flag max, flag min and flag precipitation having null values as shown in fig 3.1. This was done using isnull function.

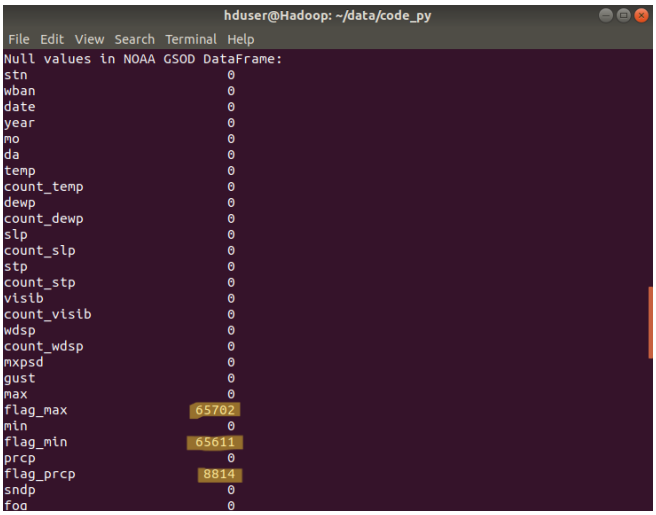


Fig 3.1: Check for null values in Noaa Gsod

Since these three columns flag max, flag min and flag precipitation are not required for any research question, have dumped them from the dataframe. Fig 3.2 shows the columns of Noaa gsod after dumping the columns with null values.

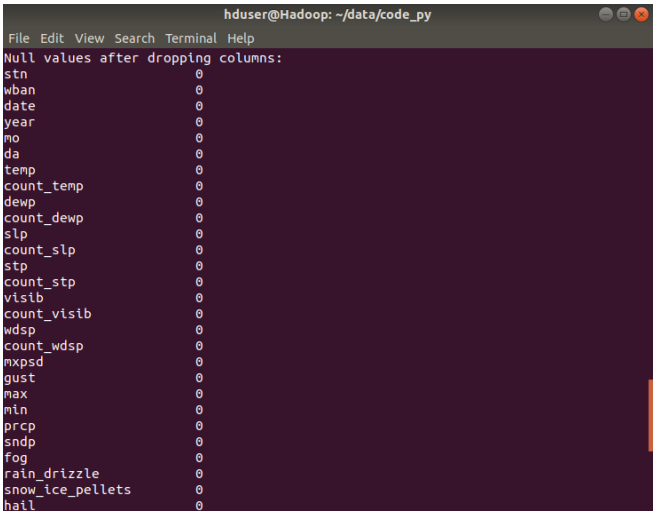


Fig 3.2: After removing columns with null values

In case of taxi trip dataframe, we don't have null values in the columns, please refer fig 3.3.

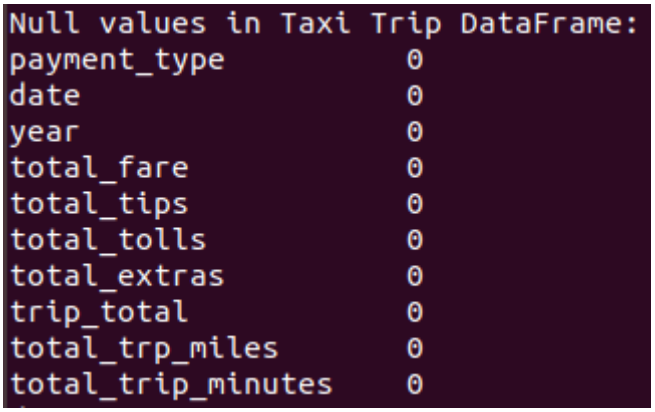


Fig 3.3: Check for null values in Taxi trip

B. Check for duplicates

This research has checked for duplicates using duplicated function in both Noaa gsod and taxi trip dataframes and found out that none of them have duplicate rows, please refer below fig 3.4.

```
Duplicate records in NOAA GSOD DataFrame:
('Number of duplicate rows:', 0)

Duplicate records in Taxi Trip DataFrame:
('Number of duplicate rows:', 0)
```

Fig 3.4: Check for duplicate rows

IV. ARCHITECTURE AND IMPLEMENTATION

In this section the paper shares insights of how everything was implemented using the architecture given in the fig 4.1.

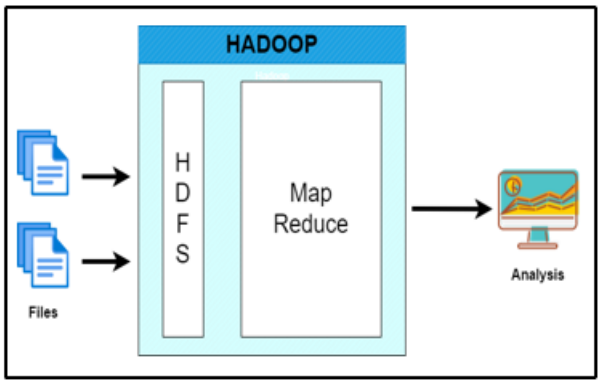


Fig 4.1: Architecture diagram

A. Architecture overview

The architecture this research followed for the analysis is given in fig 4.1, where files were loaded to hdfs first. Then code executes, it takes csv files from hdfs directory as input and then do the necessary data quality checks- a) checking for null values and b) duplicate rows in the data [D]. Then the cleaned datasets are merged on the basis of date column. Resulted merged dataframe is given as an input to map reduce functions to extract the meaningful insights from the datasets. Results of map reduce functions is converted into csv files which is finally visualized using graphical representation.

B. Implementation

Implementation phase involves 5 steps right from reading data from hdfs to visualizing the data.

Before the code starts working on the csv files, csv files need to be uploaded to Hadoop hdfs directory for that we use below command-

hdfs dfs -copyFromLocal local/path/*.csv destination/hdfs/path/

B.1) Pre-requirements

Both the csv files should be present in the hdfs directory, and all the nodes- data nodes, name node, secondary name node needs to be running in the process as well as yarn needs to be activated for the resource management.

B.2) Reading files

Process reads the file directly from hdfs and loads it into a dataframes using pandas library.

B.3) Data quality check

Before loading the data further, research makes sure that the data is cleaned and insightful. This phase of processing takes care of the null values found in the input data and redundant(repeated) data. If not removed carefully, it may impact the results later. More details regarding data cleaning particularly on the datasets used in this research have been discussed under III) Methodology section in this paper.

B.4) Data Transformation

After cleaning the data, will be converting and reformatting data from its original form into a different structure or format to make it suitable for analysis, reporting and other purposes. This research has merged both the dataframes on the basis of date.

Merging datasets on the date column allows for a correlation analysis between variables in different datasets. For example, this research can explore how weather conditions (from one dataset) correlate with taxi demand (from another dataset) over time.

This also helps in efficient analysis as it simplifies filtering, sorting and grouping based on time.

B.5) Data Analysis (MapReduce)

Map reduce is a programming model used for the parallel processing of huge datasets in a distributed environment. It consists of two phases, the mapping and reducing phase and processes data in key-value pairs.

In this research, to answer all the objectives, merged data frame is used as in input on which map-reduce method worked. This research involved below three phases in the map-reduce process for the given datasets-

B.5.1) Mapping, Key-value pair generation

The mapper function emits key-value pairs, where keys represent the date and values involve weather conditions.

B.5.2) Grouping data

In the map reduce process, data is grouped on the basis on common keys which is date in our case.

B.5.3) Reducer phase

In the reducer function, data was aggregated with the same key, performing computations such as averaging fare, duration, taxi demand for each date.

This step helps in deriving meaningful insights from the large amount of data generated during the mapping phase [G]

B.6) Data Visualization

This part is very crucial for conveying insights from the derived from the analysis of weather and taxi data.

In this research, the output received after all the KPI's from the map reduce functions is saved in different csv files based on the objectives.

This data can be further visualized which will make people easier to understand, helping them making better decisions. Visualizations can reveal patterns, trends and insights that might be challenging to understand from raw data.

V. RESULTS

For this research, will start discussing about results by answering the objective questions first and then will take it ahead about what was expected and what was surprising. Lastly will discuss about the challenges faced and solved in this research.

5.A) What effect do weather conditions have on taxi demand in Chicago?

	average_taxi_demand	bad_condition	timestamp
0	52105.661250	0	2020-02-08 00:00:00
1	28925.761429	1	2021-04-12 00:00:00
2	32473.374286	0	2021-05-07 00:00:00
3	70356.245714	0	2022-12-02 00:00:00
4	50925.620000	0	2021-12-13 00:00:00
5	27768.936667	0	2021-04-17 00:00:00
6	10108.745714	1	2020-08-09 00:00:00
7	13571.340000	0	2020-08-14 00:00:00
8	50935.494286	1	2021-09-08 00:00:00
9	103722.041429	1	2023-06-02 00:00:00
10	58641.436667	0	2023-01-09 00:00:00
11	58477.528750	1	2020-01-26 00:00:00
12	70871.242857	0	2021-11-29 00:00:00
13	15630.521429	0	2021-01-25 00:00:00
14	15985.535000	1	2020-10-26 00:00:00
15	74855.342857	1	2023-08-14 00:00:00
16	48510.141429	0	2022-11-23 00:00:00
17	13988.934286	0	2020-11-04 00:00:00
18	20651.655714	0	2021-03-08 00:00:00
19	93838.208750	0	2020-03-02 00:00:00

Fig 5.1: Average taxi demand in bad weather (1) and good weather condition (2)

Observation: From fig 5.1, one can observe that average taxi demand has increased during bad weather conditions.

5.B) *What impact do weather conditions have on taxi trip duration in Chicago?*

	average_taxi_duration	bad_condition	timestamp
0	39312.341667	0	2020-02-08 00:00:00
1	22577.483333	1	2021-04-12 00:00:00
2	27851.507143	0	2021-05-07 00:00:00
3	53854.240476	0	2022-12-02 00:00:00
4	36167.321429	0	2021-12-13 00:00:00
5	21322.013889	0	2021-04-17 00:00:00
6	8977.538095	1	2020-08-09 00:00:00
7	13236.909524	0	2020-08-14 00:00:00
8	34937.759524	1	2021-09-08 00:00:00
9	86203.721429	1	2023-06-02 00:00:00
10	42128.330556	0	2023-01-09 00:00:00
11	34923.254167	1	2020-01-26 00:00:00
12	48703.578571	0	2021-11-29 00:00:00
13	14231.285714	0	2021-01-25 00:00:00
14	12618.920833	1	2020-10-26 00:00:00
15	52696.030952	1	2023-08-14 00:00:00
16	39431.500000	0	2022-11-23 00:00:00
17	12288.783333	0	2020-11-04 00:00:00
18	17490.909524	0	2021-03-08 00:00:00
19	65782.477083	0	2020-03-02 00:00:00

Fig 5.2: Average taxi duration in bad (1) and good (0) weather conditions

Observation: From fig 5.2, one can understand trip duration tends to increase during bad weather conditions. But there are some instances where trip duration tends to increase during good weather conditions.

5.C) *What impact do weather conditions have on taxi trip fare in Chicago?*

	average_taxi_fare	bad_condition	timestamp
0	44357.762500	0	2020-02-08 00:00:00
1	24562.798571	1	2021-04-12 00:00:00
2	27809.620000	0	2021-05-07 00:00:00
3	57746.062857	0	2022-12-02 00:00:00
4	42746.752857	0	2021-12-13 00:00:00
5	22305.820000	0	2021-04-17 00:00:00
6	8740.621429	1	2020-08-09 00:00:00
7	12144.434286	0	2020-08-14 00:00:00
8	41269.315714	1	2021-09-08 00:00:00
9	81853.640000	1	2023-06-02 00:00:00
10	48316.380000	0	2023-01-09 00:00:00
11	48522.720000	1	2020-01-26 00:00:00
12	57655.315714	0	2021-11-29 00:00:00
13	14164.375714	0	2021-01-25 00:00:00
14	13348.392500	1	2020-10-26 00:00:00
15	59122.704286	1	2023-08-14 00:00:00
16	40111.521429	0	2022-11-23 00:00:00
17	12236.048571	0	2020-11-04 00:00:00
18	18317.924286	0	2021-03-08 00:00:00
19	78796.653750	0	2020-03-02 00:00:00

Fig 5.3: Average taxi fare in bad (1) and good (0) weather conditions.

Observation: From the fig 5.3, highest fare which is $81,853 was collected during bad weather conditions. There are more cases to support this at the same time there are instances showing taxi's earn good fare amount in the good weather conditions as well.

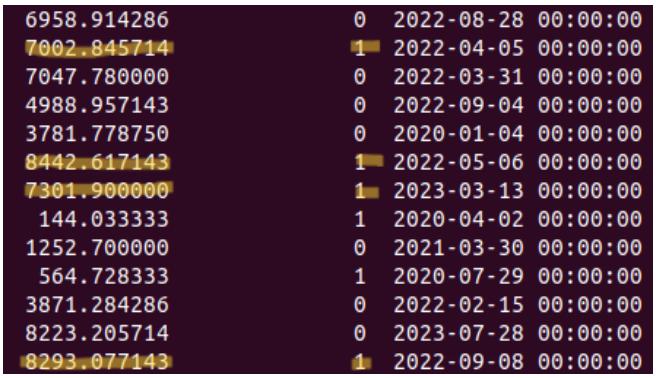
5.D) *How do weather conditions affect drop-off/destination location?*

Observation: We couldn't add location related columns in the taxi trip dataset as it was increasing memory the of the entire dataset in Gigabytes. Which was challenging for data frames to read and process. This will be added in the future scope section of this paper.

5.E) *Understanding the trend between seasons and taxi trips.*

Observation: For this we can refer fig 5.1, and observed that demand for taxi which involves parameters like – total fare earned, total minutes of the trip, total trip miles and so on collected on a particular date has increased on bad weather days which ultimately added to show increase in taxi demand for that particular day.
One can see in fig 5.2, duration of the trips increases on good days which means that sunny days attract people to go for longer trips.

One more interesting observation would be people tend to tip more during heavy rainfall, snowfall, or any bad weather days, please observe fig 5.4. Which attributes to a heightened sense of gratitude towards transportation service providers who ensure safe and reliable transit despite challenging weather.



6958.914286	0	2022-08-28 00:00:00
7002.845714	1	2022-04-05 00:00:00
7047.780000	0	2022-03-31 00:00:00
4988.957143	0	2022-09-04 00:00:00
3781.778750	0	2020-01-04 00:00:00
8442.617143	1	2022-05-06 00:00:00
7301.900000	1	2023-03-13 00:00:00
144.033333	1	2020-04-02 00:00:00
1252.700000	0	2021-03-30 00:00:00
564.728333	1	2020-07-29 00:00:00
3871.284286	0	2022-02-15 00:00:00
8223.205714	0	2023-07-28 00:00:00
8293.077143	1	2022-09-08 00:00:00

Fig 5.4: Average tips collected on good (0) and bad weather days (1)

EXPECTED: The predicted conclusions of this research were based on determining the complex relationship between weather dynamics and Chicago taxi services. Expectations included an increase in taxi demand during adverse conditions, an increase in journey durations, and a potential impact on fare dynamics. Predicting seasonal trends and determining how weather complexities would influence passenger behavior were also important expectations. The integration of Hadoop and MapReduce sought to handle large datasets rapidly and extract relevant insights, encouraging an in-depth understanding of the taxi-weather connection.

SURPRISING: Certain findings that proved wrong the early predictions, adding a degree of complexities to the study. While an increase in taxi demand was expected during bad weather, the magnitude of this spike and its appearance in fare dynamics exceeded expectations. Surprisingly, extended trip durations were observed even in favorable weather conditions, indicating a complex link between weather and travel behavior. The study also discovered an unexpected link between bad weather and increased tipping behavior, indicating a positive attitude toward transportation services in bad weather. These surprises highlight the complexity of the weather-taxi connection and highlight the importance of thorough research in understanding such complexities.

Challenges:

There were few challenges we faced during the entire implementation of this research.

a) Dataset missing: Dataset picked during project proposal phase is no more available on kaggle website, to fix it Big query was used to download the public datasets of noaa gsod and taxi trip. [E]

b) Memory issue: During reading and merging phase of the datasets, have faced memory issue multiple times. To solve this issue have increased RAM size from 4 GB to 12 GB for the oracle virtual box. [F]

VI . CONCLUSION

This research successfully explained the complex relationship between Chicago's weather patterns and taxi services, employing advanced techniques such as Hadoop and MapReduce for robust data analysis. The study focused on important aspects including taxi demand, journey durations, fare dynamics, and seasonal influences. The findings not only confirmed anticipated trends, such as increased demand during adverse weather, prolonged trip durations, and variable fare impacts but also revealed surprising details. Unexpected correlations, like increased tipping behaviour during adverse weather, demonstrated the diverse nature of the weather-taxi dynamic. The architectural framework, data processing methodologies and the output, employed proved crucial in extracting meaningful insights from extensive datasets.

Future work:

Despite the valuable insights gained, future scope for this research still exists. The missing location-related columns in the taxi trip dataset present an opportunity for further research into how weather conditions influence drop-off or destination choices. Additionally, expanding the scope beyond 2020 may provide a more comprehensive understanding of long-term trends. Future work could also incorporate real-time weather data for more dynamic analyses. Continuous advancements in data processing and analytical tools could enhance the depth and efficiency of future investigations in this domain. Overall, the research lays the framework for further detailed investigations into the dynamic relationship between weather and taxi transportation in urban environments.

REFERENCES

A) Nurmi, P., Perrels, A., & Nurmi, V. (2013). Expected impacts and value of improvements in weather forecasting on the road transport sector. Meteorological Applications, 20(2), 217–223. https://doi.org/10.1002/met.1399

B) Chicago taxi trips. (2018, April 18). Kaggle. https://www.kaggle.com/datasets/chicago/chicago-taxi-trips-bq/data

C) NOAA GSOD. (2019, August 30). Kaggle. https://www.kaggle.com/datasets/noaa/gsod

D) Big data engineering: How distributed systems transform data processing? | Data Science Dojo. (n.d.). Data Science Dojo. https://datasciencedojo.com/blog/big-data-engineering/#

E) Google Cloud Platform. (n.d.-b). https://console.cloud.google.com/bigquery?project=rosy-spring-405311&ws=!1m20!1m4!4m3!1sbigquery-public-data!2snoaa_gsod!3sgsod2021!1m4!1m3!1srosy-spring-405311!2sbquxjob_4f3fbd8a_18c253094d7!3sUS!1m4!4m3!1sbigquery-public-data!2snoaa_gsod!3sstations!1m4!1m3!1srosy-spring-405311!2sbquxjob_dd0ef48_18c2beaecea!3sUS!1m0

F) MemoryError when I merge two Pandas data frames. (n.d.). Stack Overflow. https://stackoverflow.com/questions/47386405/memoryerror-when-i-merge-two-pandas-data-frames#:~:text=The%20reason%20you%20might%20be%20getting%20MemoryError%3A%20Unable,%28subset%20%3D%27column_name%27%2C%20keep%20%3D%20False%2C%20inplace%20%3D%20True%29

G) Ray, R. (2023, August 11). Hadoop Streaming: Writing a Hadoop MapReduce program in Python. Edureka. https://www.edureka.co/blog/hadoop-streaming-mapreduce-program/