CENTER FOR DEVELOPMENT OF
ADVANCED COMPUTING

Flight Delay Analysis using Apache Spark and Building Flight Delay Prediction Machine Learning Model

Presented By

1508 Sandeep Bayas

1520 Nipun Jaiswal

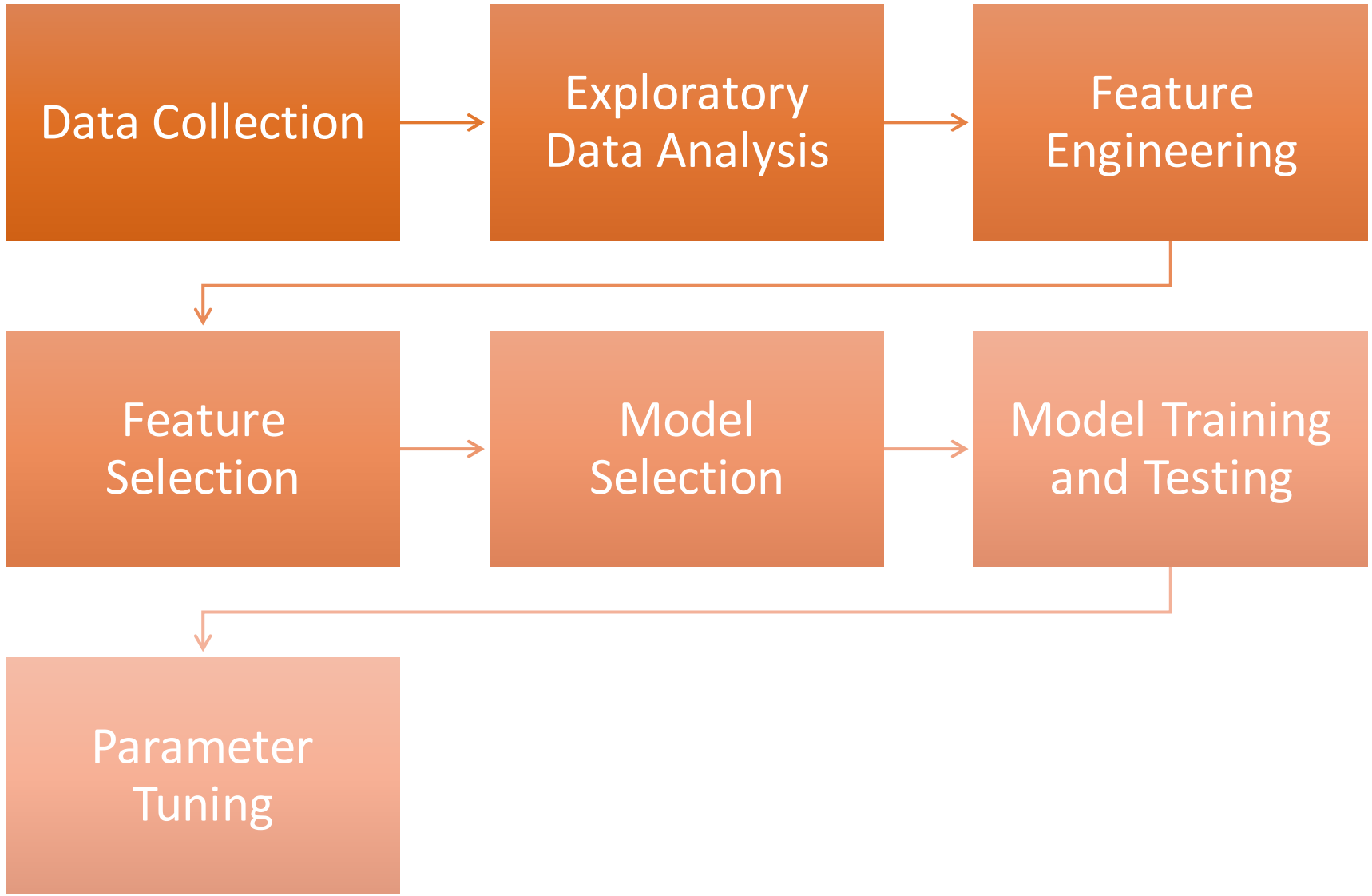
Prashant Karhale
Center coordinator

External guide
Akshay Tilekar

Introduction

- Flight delays has become a very important consideration for air transportation all over the world .
- These delays not only cause inconvenience to the airlines but for the passengers also .
- So the primary goal of this project is to collect and analyze the data.
- Provide insights from data to Business Stakeholders and help them for making a well-assessed decision to increase operation efficiency and growth in business.
- Alongside building a predictive machine learning model to reduce inconvenience happens to passengers while travelling.

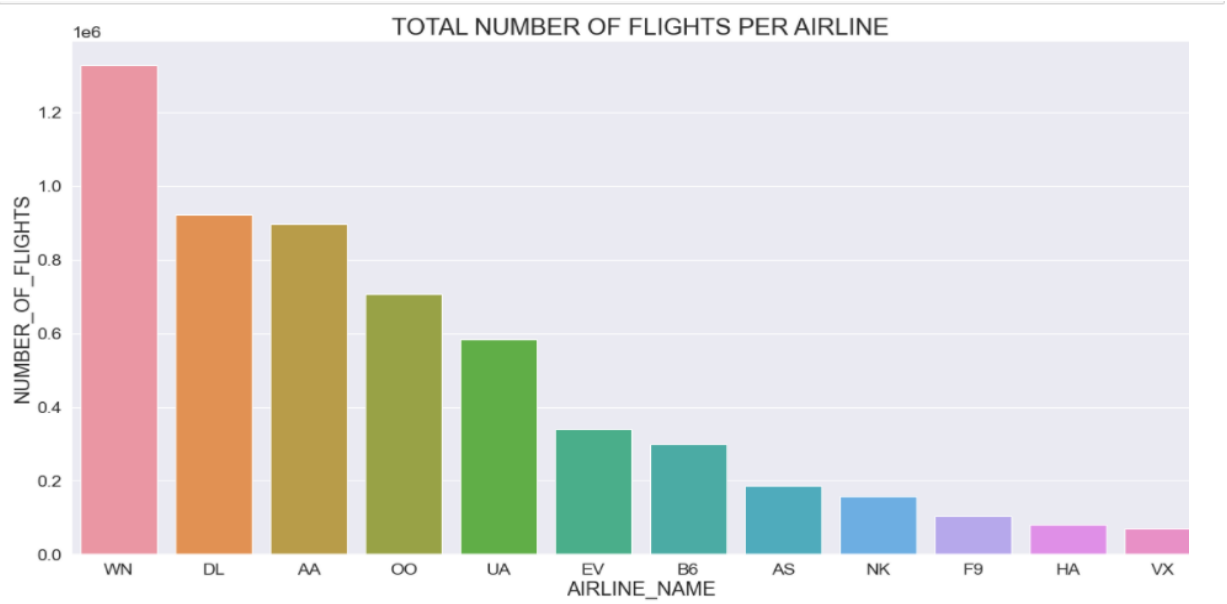
Proposed Methodology

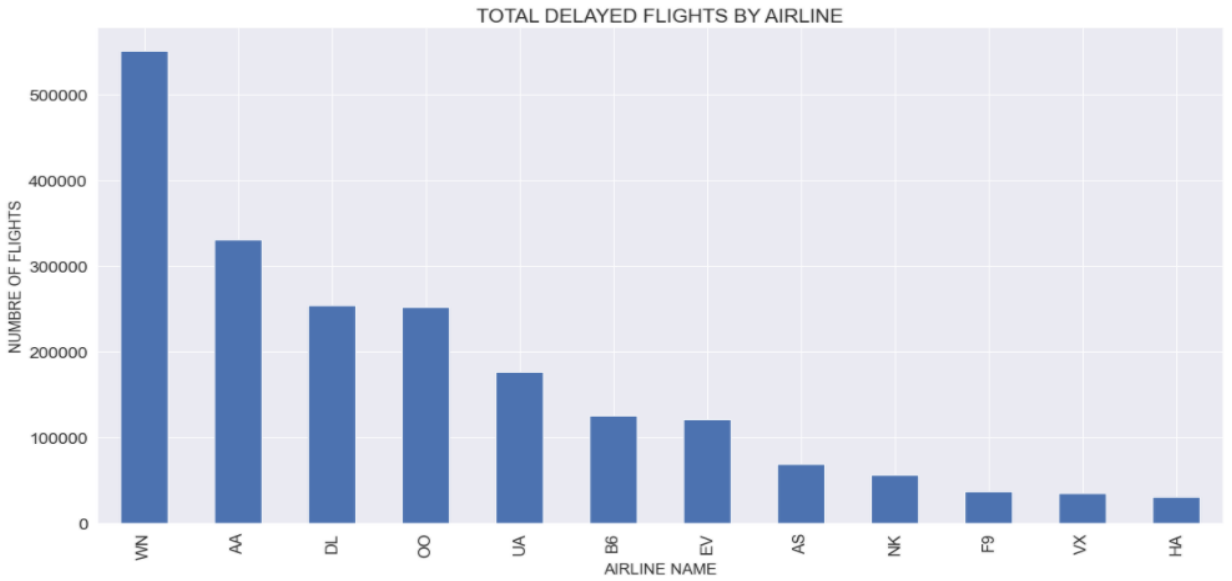


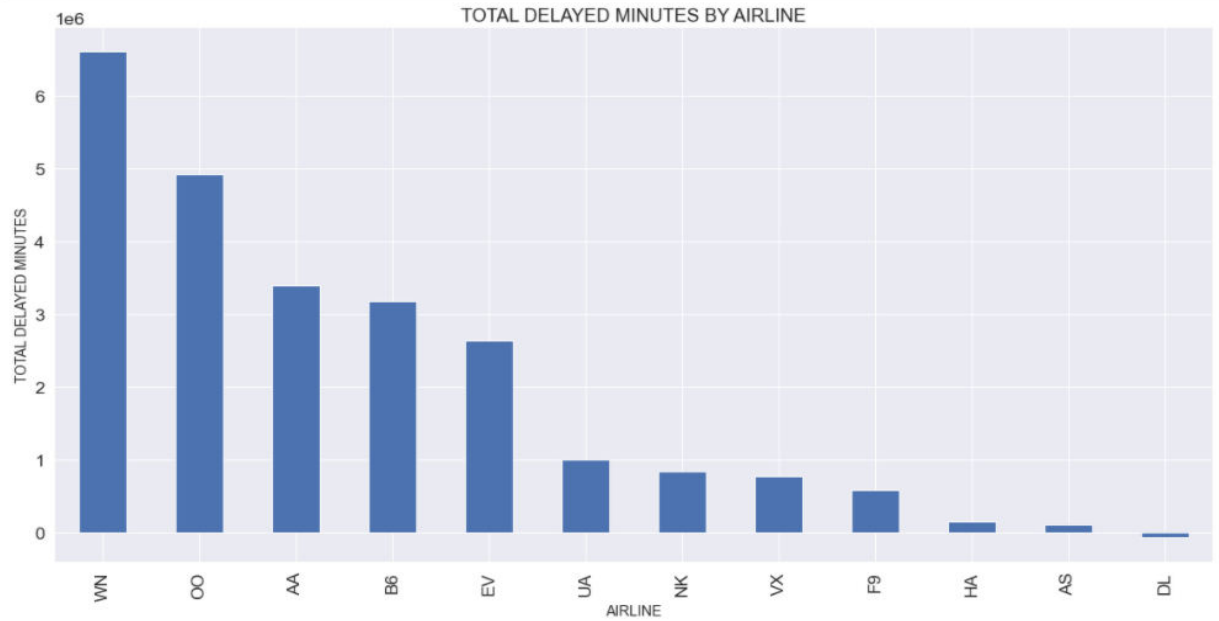
```
FL_DATE                object
OP_CARRIER             object
OP_CARRIER_FL_NUM       int64
ORIGIN                 object
DEST                   object
CRS_DEP_TIME            int64
DEP_TIME              float64
DEP_DELAY             float64
TAXI_OUT              float64
WHEELS_OFF            float64
WHEELS_ON             float64
TAXI_IN               float64
CRS_ARR_TIME            int64
ARR_TIME              float64
ARR_DELAY             float64
CANCELLED             float64
CANCELLATION_CODE      object
DIVERTED              float64
CRS_ELAPSED_TIME      float64
ACTUAL_ELAPSED_TIME   float64
AIR_TIME              float64
DISTANCE              float64
CARRIER_DELAY         float64
WEATHER_DELAY         float64
NAS_DELAY             float64
SECURITY_DELAY        float64
LATE_AIRCRAFT_DELAY   float64
Unnamed: 27           float64
dtype: object
```

	FL_DATE	OP_CARRIER	OP_CARRIER_FL_NUM	ORIGIN	DEST	CRS_DEP_TIME	DEP_TIME	DEP_DELAY	TAXI_OUT	WHEELS_OFF	...
0	2017-01-01	AA	1	JFK	LAX	800	831.0	31.0	25.0	856.0	...
1	2017-01-01	AA	2	LAX	JFK	900	934.0	34.0	34.0	1008.0	...
2	2017-01-01	AA	4	LAX	JFK	1130	1221.0	51.0	20.0	1241.0	...
3	2017-01-01	AA	5	DFW	HNL	1135	1252.0	77.0	19.0	1311.0	...
4	2017-01-01	AA	6	OGG	DFW	1855	1855.0	0.0	16.0	1911.0	...

The dataset come from the website of bureau of transportation and statistics United States (www.bts.gov) OR you can get it on Kaggle also. The dataset provide information about flight records per airline across United States. which include 28 features and around 56 million rows.



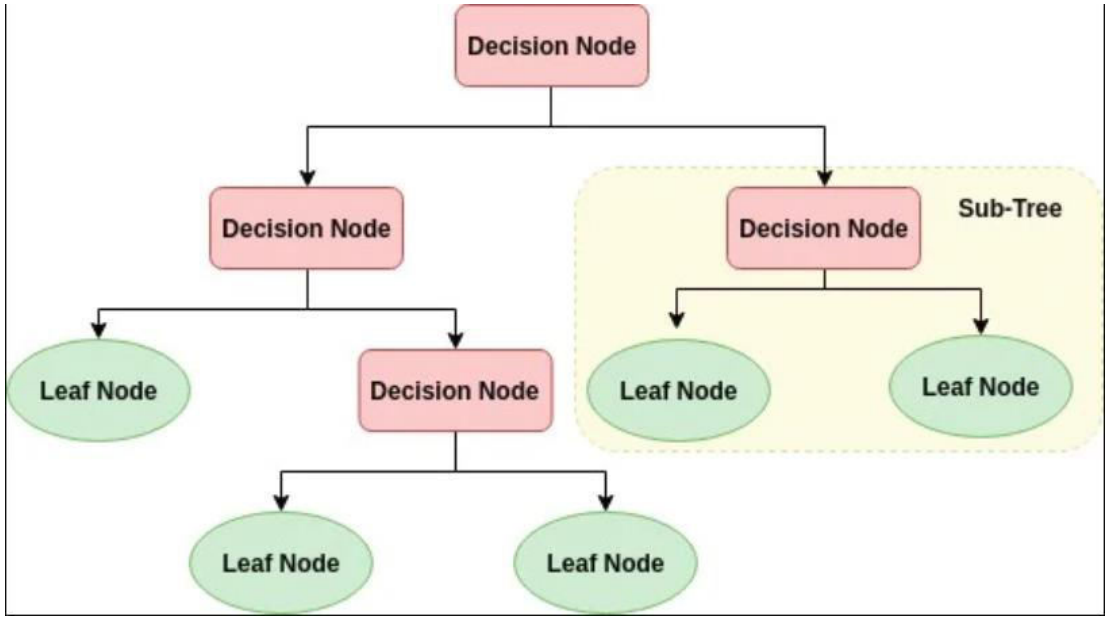




With reference of above plots we come to know that B6 (JetBlue Airways) operates less number of flights as compare to others(Top 5) . Along with there flights are delayed also in large number as compare to other 4 flight operators. So passengers must be cautious while selecting the airlines

Feature Engineering and Data Preprocessing

- Putting NA where there is missing values
- Remove unnecessary columns (as per the requirement)
- Format or Binarize the target or unknown variable
- Check datatypes, handle columns with Object / string type (either converting to number / removing them)
- Removing columns which have all missing / NA values
- Removing rows with missing / NA values greater than 50%
- Filling missing / NA values using central tendencies



- Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems.
- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.

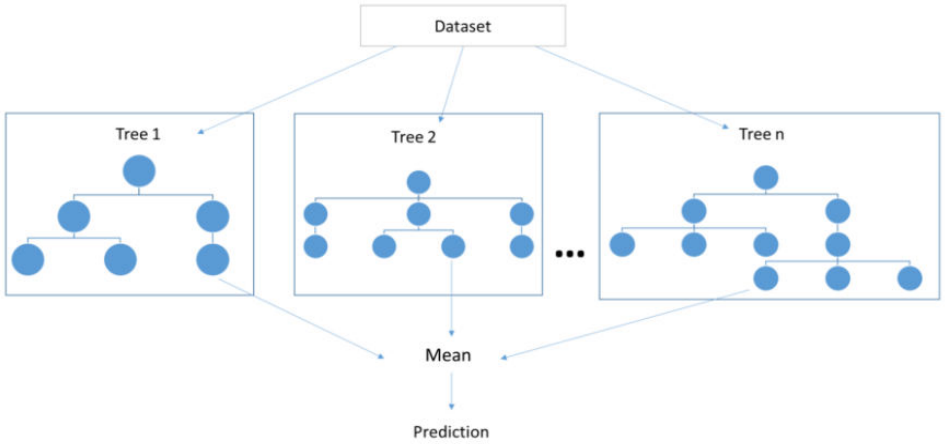
- After applying Decision Tree algorithm on our dataset i.e both classification as well as regression model we got predicated and actual values as follows(Actual and Predicted).

	Actual	Delay/No_Delay
4429905	0	0
340258	0	0
4363275	0	0
2563623	0	0
4098989	0	0
...	...	...
1056306	0	0
4812823	0	0
4649165	0	0
354559	0	0
802708	0	0

1673824 rows × 2 columns

	Actual	Predicted
1390887	-9.0	-9.0
2775323	-20.0	-20.0
4127958	19.0	19.0
2173009	-14.0	-14.0
4362732	-30.0	-29.0
652116	57.0	60.0
1669628	-21.0	-21.0
4306366	-19.0	-19.0
3625714	22.0	22.0
3730946	-19.0	-19.0
3177473	-9.0	-9.0
2905544	96.0	94.0
12408	-9.0	-9.0
388642	-10.0	-10.0
2530425	70.0	71.0
1056306	-16.0	-14.0
4812823	-18.0	-18.0
4649165	-3.0	-3.0
354559	-15.0	-15.0
802708	-16.0	-17.0

Random Forest :



- Random forest is a supervised learning algorithm .
- The forest it builds, is an ensemble of decision trees, usually trained with the bagging method.
- The general idea of the bagging method is that a combination of learning models increases the overall result.
- Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

- After applying Decision Tree algorithm on our dataset i.e both classification as well as regression model we got predicated and actual values as follows(Actual and Predicted).

	Actual	Predicted
2249516	0.0	0.0
1594843	11.0	13.2
673049	-8.0	-8.4
1373709	-10.0	-9.6
4111358	-4.0	-4.0
2212830	9.0	9.0
948423	299.0	306.2
3053181	3.0	2.8
4256245	-23.0	-23.0
1027461	-31.0	-29.2
4115331	-7.0	-7.0
5129801	-26.0	-26.0
2992352	-14.0	-11.2
3209844	-3.0	-3.4
2286056	-3.0	-2.6
3407081	0.0	-0.4
1702443	-1.0	-0.8
1280839	8.0	8.4
4846176	59.0	60.8
2091501	11.0	10.4

	Actual	Predicted
4429905	0	0
340258	0	0
4363275	0	0
2563623	0	0
4098989	0	0
...	...	...
1056306	0	0
4812823	0	0
4649165	0	0
354559	0	0
802708	0	0

1673824 rows × 2 columns

- Hyperparameters are important because they directly control the behavior of the training algorithm and have a significant impact on the performance of the model is being trained
- A hyperparameter is a hyperparameters whose value is used to control the learning process and increase the model performance by choosing best parameters

```
param ={
    'criterion' :['gini', 'entropy'],
   'max_depth':[1,2,3,4,5,None]        #max split
}
```

```
from sklearn.model_selection import GridSearchCV
GCV_grid=GridSearchCV(estimator=model,param_grid=param,cv=10,verbose=2)
```

```
GCV_grid.fit(X_train, Y_train)
```

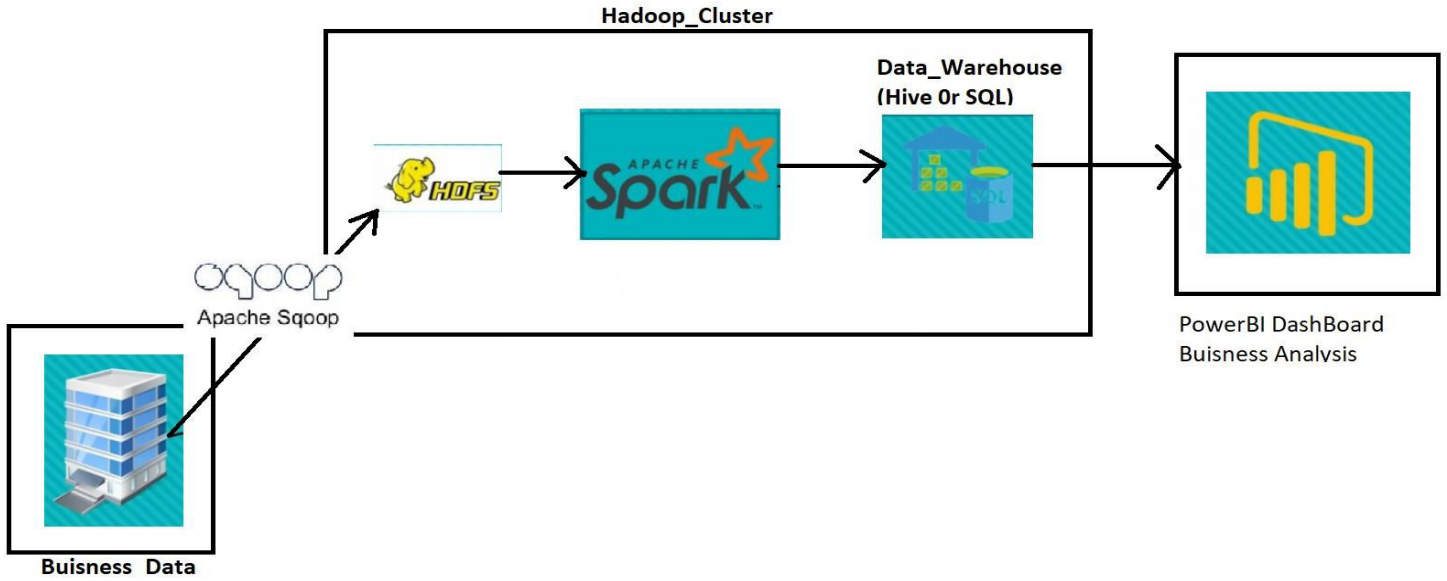
```
{'criterion': 'gini', 'max_depth': 1}
[0.99999974 0.99999974 0.99999974 0.99999949 0.99999949 0.99999949
 0.99999974 0.99999974 0.99999974 0.99999949 0.99999949 0.99999949]
```

```
param1 ={
    'n_estimators':[10,50,100],
    'criterion' :['gini', 'entropy'],    #how many decision tree to keep
   'max_depth':[1,2,3,4,5,None]         #max split
}
```

```
from sklearn.model_selection import GridSearchCV
GCV_grid1=GridSearchCV(estimator=model_Class,param_grid=param1,cv=5,verbose=1)
```

```
{'criterion': 'gini', 'max_depth': 3, 'n_estimators': 100}
[0.96102481 0.90858711 0.88651744 0.99995186 0.96677708 0.9645536
 0.99948868 0.99467916 0.99999974 0.97478381 0.99954783 0.99999974
 0.99999974 0.99999974 0.99999974 0.99999974 0.99999974 0.99999974
 0.95715164 0.90410635 0.88225662 0.99986865 0.9699067  0.97111753
 0.99970248 0.99618444 0.99999974 0.9856398  0.99994726 0.99999974
 0.99999974 0.99999974 0.99999974 0.99999974 0.99999974 0.99999974]
```

Big Data Methodology:



- Generally in most of the organizations the crucial data is stored in RDBMS systems (OLTP) and we cant perform analytics there.
- So here we need to import this data from RDBMS to HDFS using Sqoop and performing analysis inside Hadoop cluster.after analyzing the data the output stored into data warehouse(MySQL) and finally we collect the data from warehouse into visualization tool Powerbi for doing the visualizations

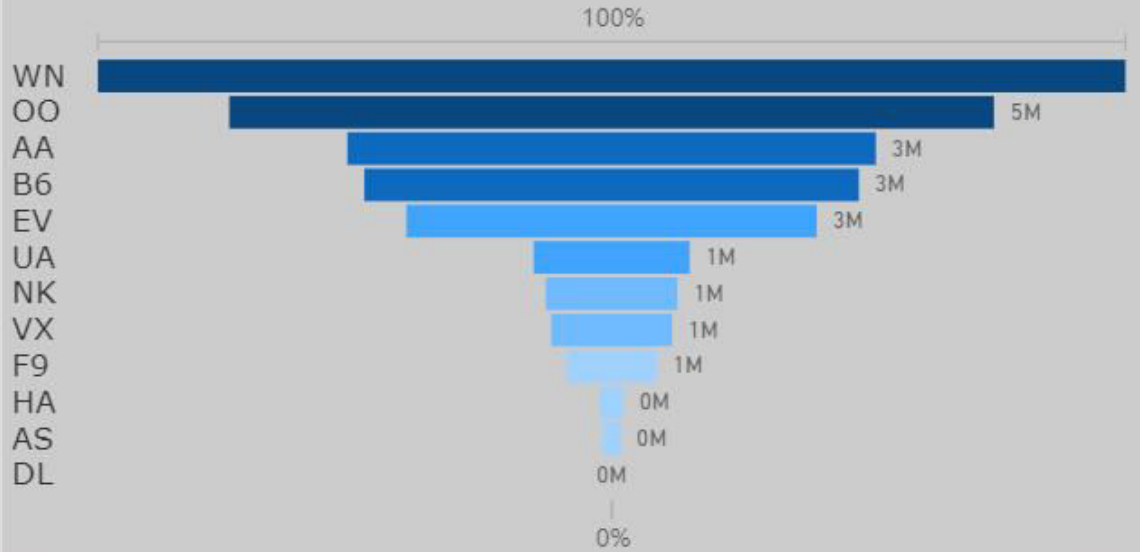
13K
DIVERTED

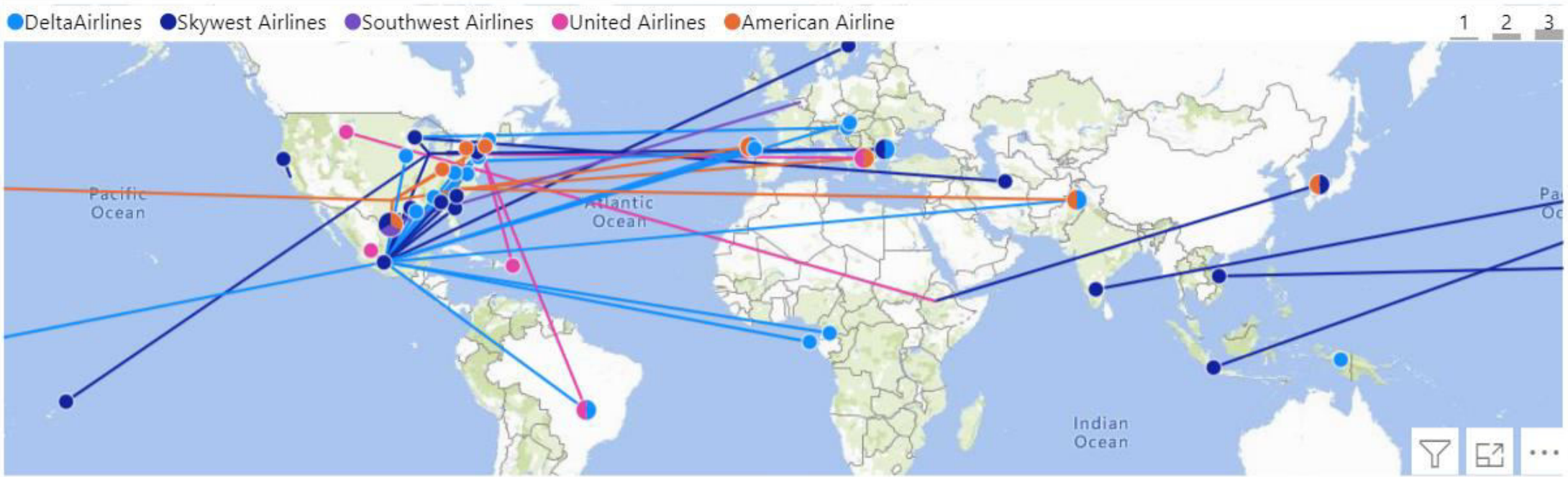
83K
CANCELLED

5bn
DISTANCE

Scheduled Arrival Time VS Actual Arrival Time

Arrival Delay Per Airline





Operating Airlines

Select all	AA	AS	B6	DL	EV	F9	HA	NK	OO	UA	VX	WN

Conclusion

- Here we utilized machine learning capabilities to predicting the flight delays. This model is based on a simple classification and regression techniques using Decision Tree and Random Forest algorithms.
- The model has achieved an overall 98% testing accuracy on publicly accessible dataset,
- It is concluded from accuracy that Random Forest is highly suitable for solving this kind of problem statements.
- For storing and processing this kind of large datasets Spark with Hadoop cluster doing a great job by harnessing the capabilities of Parallel processing. And for better insights and visuals from data Powerbi has done the great job.

Future Scope

- Integrating the ml model with application to predict the real time flight delays
- With new advancement in the field of Deep learning we can use Neural Networks algorithm on the flight and weather data.
- As neural Network works on pattern matching methodology.
- Also the scope of the project is very much confined to flight and weather delay of US so here we can also take into count of other countries like India, Japan and many more

Thank You