BIG DATA PROJECT BASED ON FLIGHT DELAYS

By
Sarvani Keella
Meng Chenghong
Jaya Krishna Vadlamudi
Leela Chakravarthy Pedapudi

ABSTRACT

- Flight delays are quite frequent and are a major source of frustration and cost.
- The Objective of our project is to estimate the delay probability distribution using Hadoop Ecosystem.
- The outcomes are converted into different visualized graphs.

SAMPLE DATA SET

AIRLINE Airline Identifier	FLIGHT_NUMBER Flight Identifier	TAIL_NUMBER Aircraft Identifier	ORIGIN_AIRPORT Starting Airport	DESTINATION_AIRPOR Destination Airport	SCHEDULED_DEPARTU Planned Departure Time	DEPARTURE_TIME WHEEL_OFF - TAXI_OUT	DEPARTURE_DELAY Total Delay on Departure
AS	98	N407AS	ANC	SEA	0005	2354	-11
AA	2336	N3KUAA	LAX	PBI	0010	0002	-8
US	840	N171US	SFO	CLT	0020	0018	-2
AA	258	N3HYAA	LAX	MIA	0020	0015	-5
AS	135	N527AS	SEA	ANC	0025	0024	-1

AIR_TIME The time duration between wheels_off and wheels_on time	DISTANCE Distance between two airports	WHEELS_ON The time point that the aircraft's wheels touch on the ground	TAXI_IN The time duration elapsed between wheels-on and gate arrival at the destination airport	SCHEDULED_ARRIVAL Planned arrival time	ARRIVAL_TIME WHEELS_ON+TAXI_IN	ARRIVAL_DELAY ARRIVAL_TIME-SCHEDULED_ARRIVAL
169	1448	0404	4	0430	0408	-22
263	2330	0737	4	0750	0741	-9
266	2296	0800	11	0806	0811	5
258	2342	0748	8	0805	0756	-9
199	1448	0254	5	0320	0259	-21

APACHE SPARK

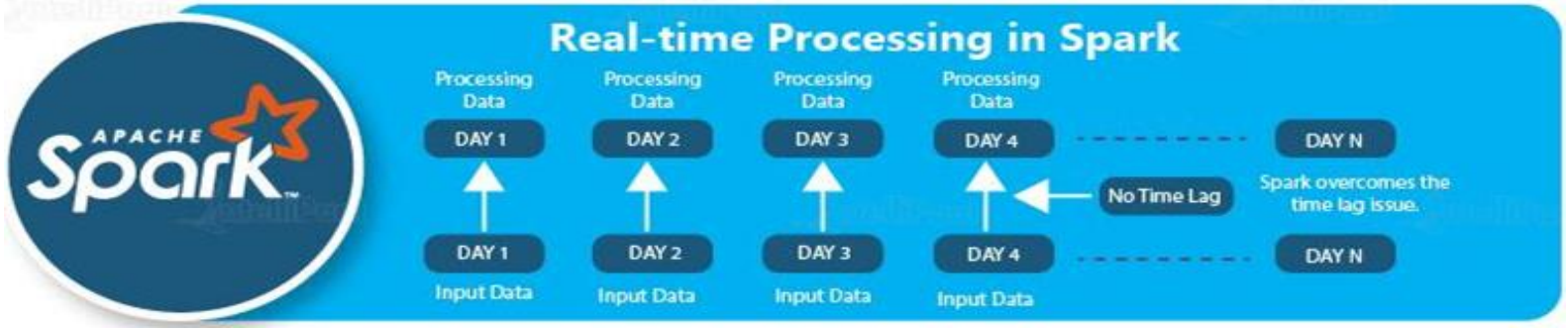
- Spark is originally developed at university of california, but currently maintained by Apache Software Foundation.
- Apache spark is a fast, large scale in-memory data processing engine which can work along with the hadoop.
- It supports different programming API’s such as Java, Python, Scala, SQL and R language.

WHY APACHE SPARK?

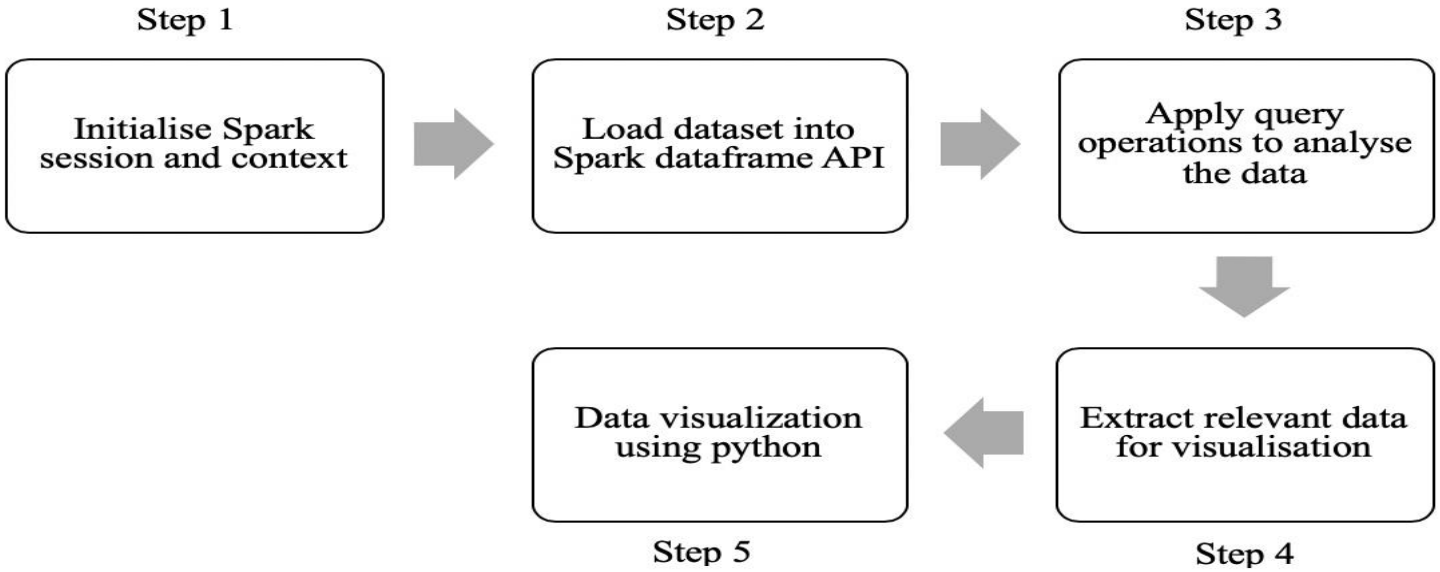
- Spark does both batch-processing and in-memory processing of data, for batch-processing it is 10 times faster than mapreduce and upto 100 times faster with in-memory processing .
- Less latency when compared to hadoop mapreduce since it does in-memory processing where most of the input data is stored in cache.

WHY APACHE SPARK?

- It supports stream processing (real-time processing) which involves continuous input and output of data (data parallelism).

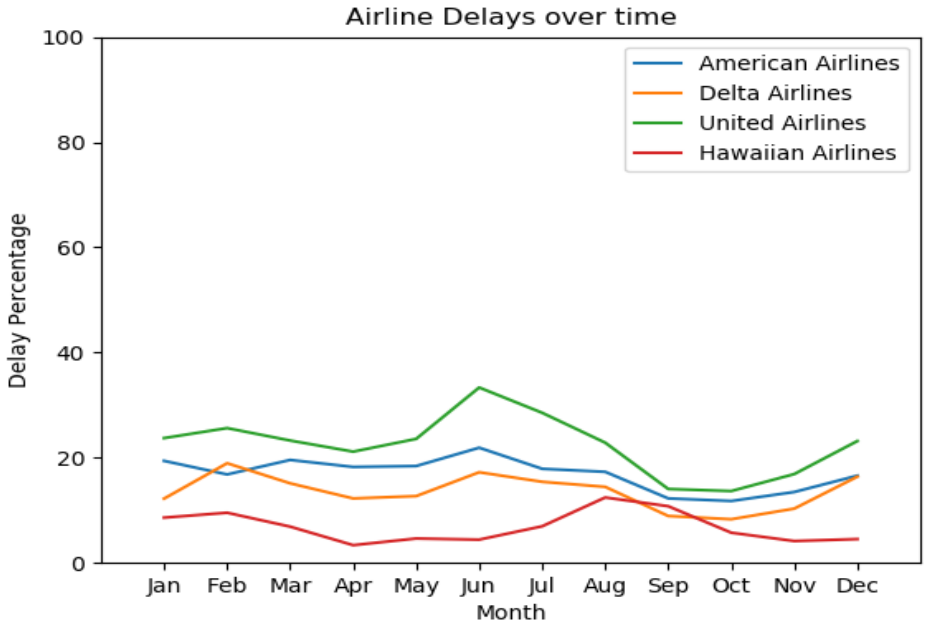


DATA ANALYSIS USING SPARK



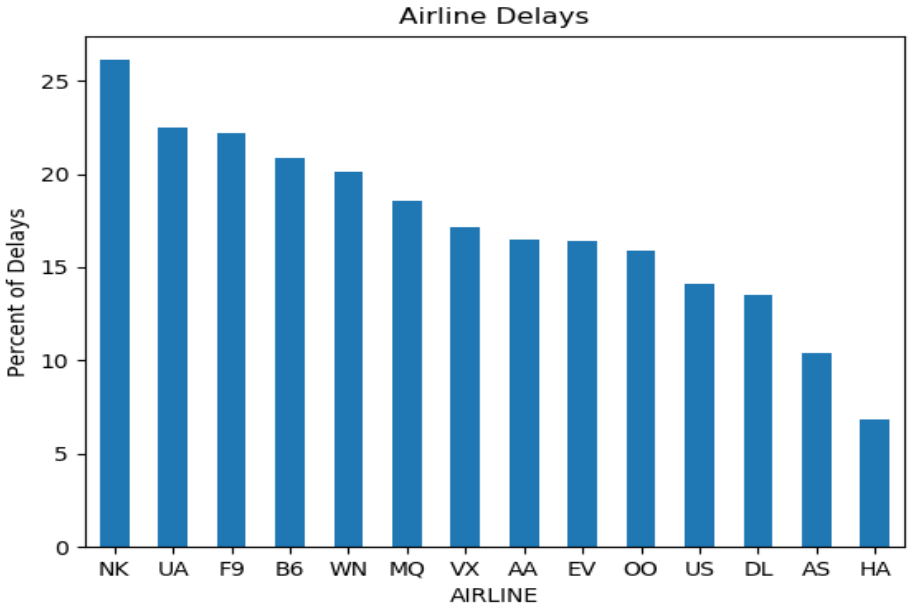
VISUALIZED RESULTS

- The following line graph summarizes the flight delays over time based on the specific airlines.
- By looking at the graph, we can observe that most delays are caused by **“United Airlines”** and the least by **“Hawaiian Airlines”**.



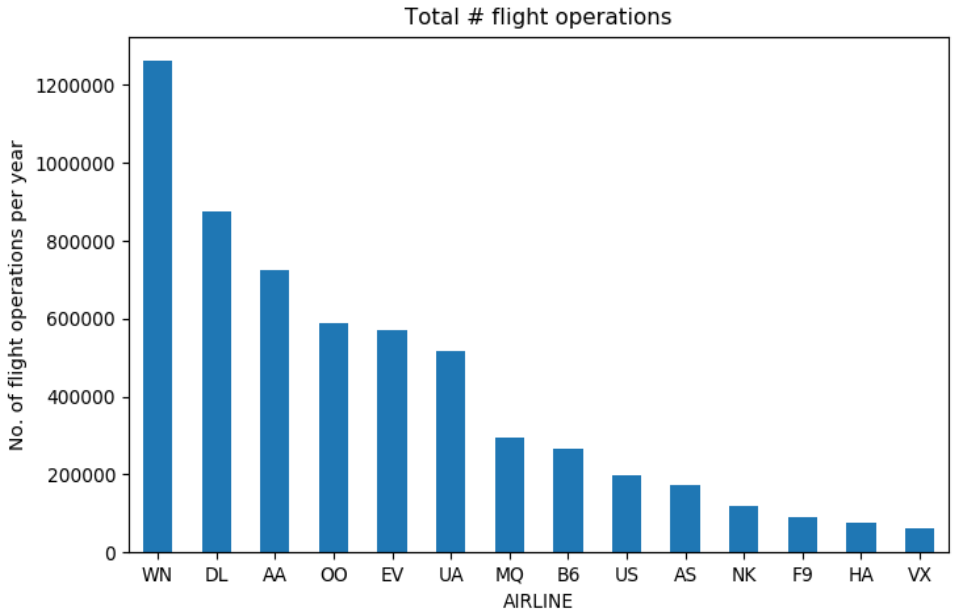
VISUALIZED RESULTS

- The following is bar graph, which depicts the percentage of total delays caused by specific airlines.
- Most delays caused by **“spirit airlines”** (NK) and the least delays by **“Hawaiian airlines”** (HA).



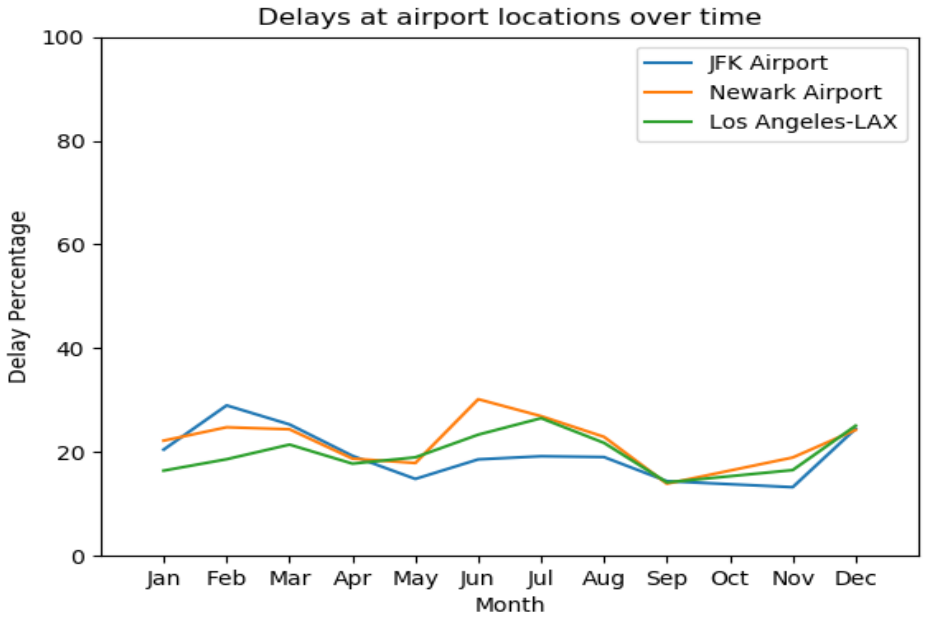
VISUALIZED RESULTS

- This bar graph describes the total number of flights operated by every individual airlines over the year.
- **“Southwest airlines”** (WN) operates more number of flights over the year.



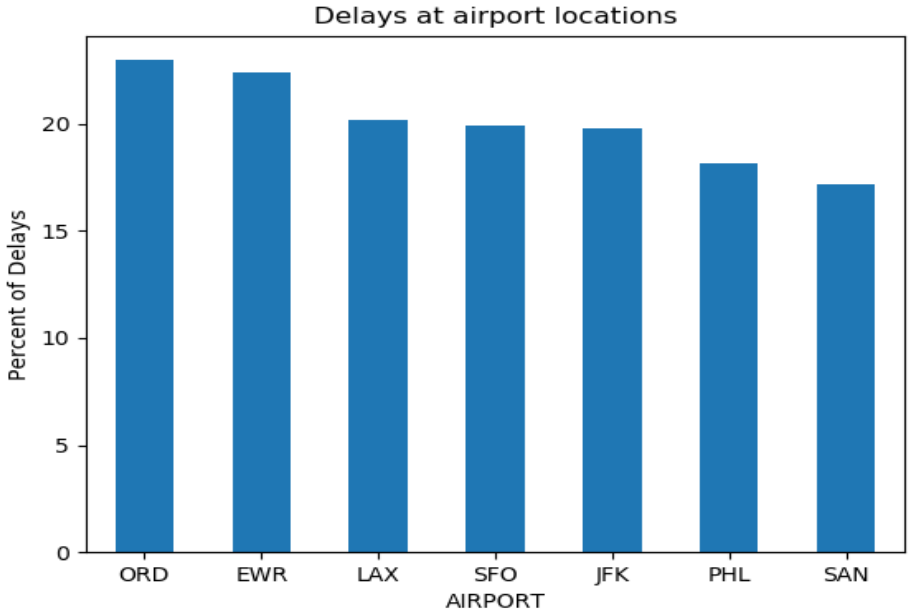
VISUALIZED RESULTS

- The following line graph shows the flight delays over time at particular airport locations.
- We can observe that there is no huge difference in delays over time, they we actually overlapping.



VISUALIZED RESULTS

- This bar graph describes the percentage of delays caused at particular airport locations.
- **“Chicago o’hare international airport”** (ORD) stands at the top with the most flight delays.



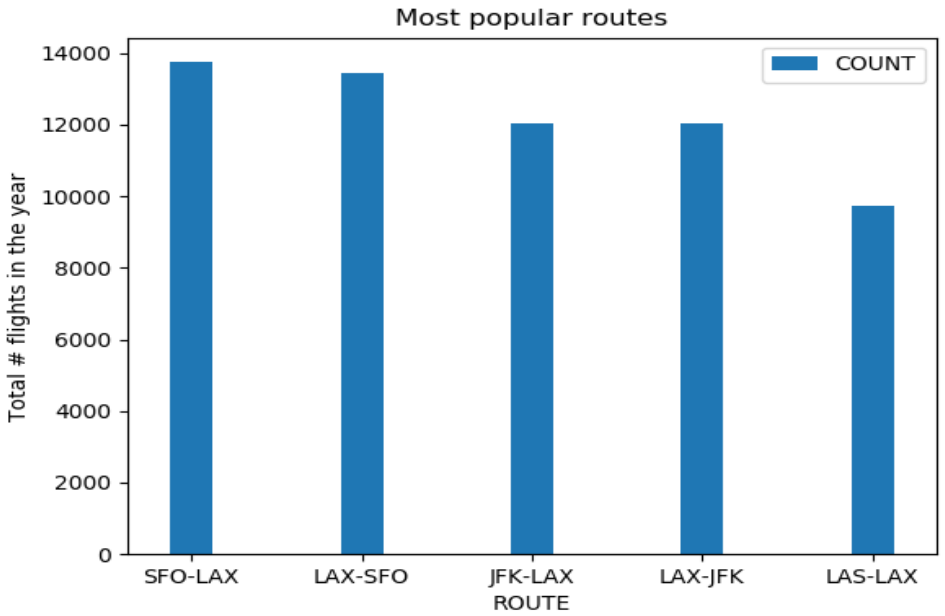
VISUALIZED RESULTS

- This bar graph describes the most popular flight routes based on total number of flights operated between airports.

SFO - San Francisco International Airport,

LAX - Los Angeles International Airport,

JFK - John F. Kennedy International Airport.



DEMONSTRATION

PROBLEMS FACED

- Initially we have started working with Hadoop using MapReduce, Hive but as it is running on virtual machine and size of data set is huge we have encountered multiple crashes of VMware. So, we have used Apache Spark which has more advantages over MapReduce.
- Tried linear regression as a part of prediction, but it does not worked as the accuracy rate showed up to be only 11 percent.

CONCLUSION

- We have performed in-depth analysis on the data set and generated visualized graphs as a part of analysis.
- These graphs provides an insight about the performance of the various airlines based on delays along with details of delays at airport locations.
- As a future enhancement, we can try a different predicting algorithm to get more accurate results for prediction.

REFERENCES

- Source for data set: https://www.kaggle.com/usdot/flight-delays
- https://en.wikipedia.org/wiki/Apache_Spark
- https://intellipaat.com/blog/what-is-apache-spark/
- https://www.infoworld.com/article/3014440/five-things-you-need-to-know-about-hadoop-v-apache-spark.html