# Big Data Computing
# Lab - VIII

**September 28, 2021**

## I. Flight Data Analysis using Apache Spark GraphX

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. The Summary information includes the number of on-time, delayed, canceled, and diverted flights which are published in DOT's monthly Air Travel Consumer Report. In this problem, we consider the dataset of January 2014 flight delays and cancellations. The schema for the dataset is provided below:

| Column | Datatype | Description |
|---|---|---|
| dOfM | String | Day of Month |
| dOfW | String | Day of Week |
| carrier | String | Carrier code |
| tailNum | String | Unique identifier for the plane - tail number |
| flnum | Integer | Flight number |
| org_id | String | Origin airport ID |
| origin | String | Origin airport code |
| dest_id | String | Destination airport ID |
| dest | String | Destination airport code |
| crsdeptime | Double | Scheduled departure time |
| deptime | Double | Actual departure time |
| depdelaymins | Double | Departure delay in minutes |
| crsarrtime | Double | Scheduled arrival time |
| arrtime | Double | Actual arrival time |
| arrdelaymins | Double | Arrival delay minutes |
| crselapsedtime | Double | Elapsed time |

| | | |
|---|---|---|
| dist | Integer | Distance |

Where a sample cross-section of the data is represented below:

```
31,5,AA,N3LBAA,937,14100,PHL,13303,MIA,600,600,0,855,834,0,175,1013
1,3,AA,N3ENAA,942,11298,DFW,13487,MSP,1225,1223,0,1445,1428,0,140,852
3,5,AA,N471AA,193,13198,MCI,13930,ORD,1635,1756,81,1805,1930,85,90,403
3,5,WN,N710SW,1552,14771,SFO,12889,LAS,2050,2348,178,2210,55,165,80,414
18,6,US,N904AW,663,14107,PHX,12758,KOA,1135,1159,24,1514,1532,18,399,2860
……..
…….
```

Taking into consideration the above dataset, you are provided with a source code file that utilizes the Apache Spark GraphX library for solving the following queries:

1. Compute the total number of airports.
2. Calculate the total number of routes existing among the different airports.
3. Outline the flight routes that exceed over a distance of 1000 miles.
4. Display the longest running routes in descending order
5. Display the most busiest airports with highest inward and outgoing traffic
6. Output the top 10 busiest & idle flight routes between airport to airport.
7. Display the most important airports using PageRank algorithm.

Note: Some important instructions:

a. **To copy the file contents directly from local system to VM:**

scp -r /home/iitp/CS-555-Lab-2021/Lab-VIII/flights iitp@172.16.27.166:/home/iitp/

b. **To compile and create jar file: (Browse to flights directory on VM and type)**

mvn clean && mvn compile && mvn package

c. **To run the program and redirecting outcome to a specific output file:**

$SPARK_HOME/bin/spark-submit --master local --class FlightApp /home/iitp/flights/target/sparkgraphx-1.0.jar > output.txt

d. **Displaying the output:**

nano output.txt