Report of Assignment-4

Bishwajeet Dey, Rashmi Dwaraka, Futai Lin

October 13, 2017

1. Introduction

This report discusses the architecture decisions and the corresponding runtimes used to solve On-Time Performance Data (http://janvitek.org/pdpmr/f17/task-a4-delay.html) across a multi node cluster in AWS.

1.2 Dataset description

The dataset is on-time airline performance data hosted by Bureau of Transportation Statistics. It contains detailed facets of all the domestic flights during 1987 to 2015 in USA.

1.3 Execution environment

We executed the jobs on a 4 node in a master slave configuration of EMR cluster. Each node is a general purpose M4 instances.

Features:

Processor: 2.3 GHz Intel Xeon® E5-2686 v4 (Broadwell) processors or 2.4 GHz Intel Xeon® E5-2676 v3 (Haswell)
VCPU: 4
Memory: 32GiB
SSD Storage: EBS

2. Design

We have 2 jobs.

1. The first job sanity checks each flight record and aggregates the flight count and flight mean delay for all the flights in a month. Each flight is consider for an airline flying to a particular destination. Since we have restricted 1 file per map, we are able to calculate the mean and count aggregated month-wise. This job doesnot require a reducer as we aggregate all the values in map.
2. The second job calculates the top 5 active airlines and airports.

3. Performance and Result

The execution of the job on th 16GB data, the execution time is 19 minutes.

Top 5 active Airports and Airlines over the years

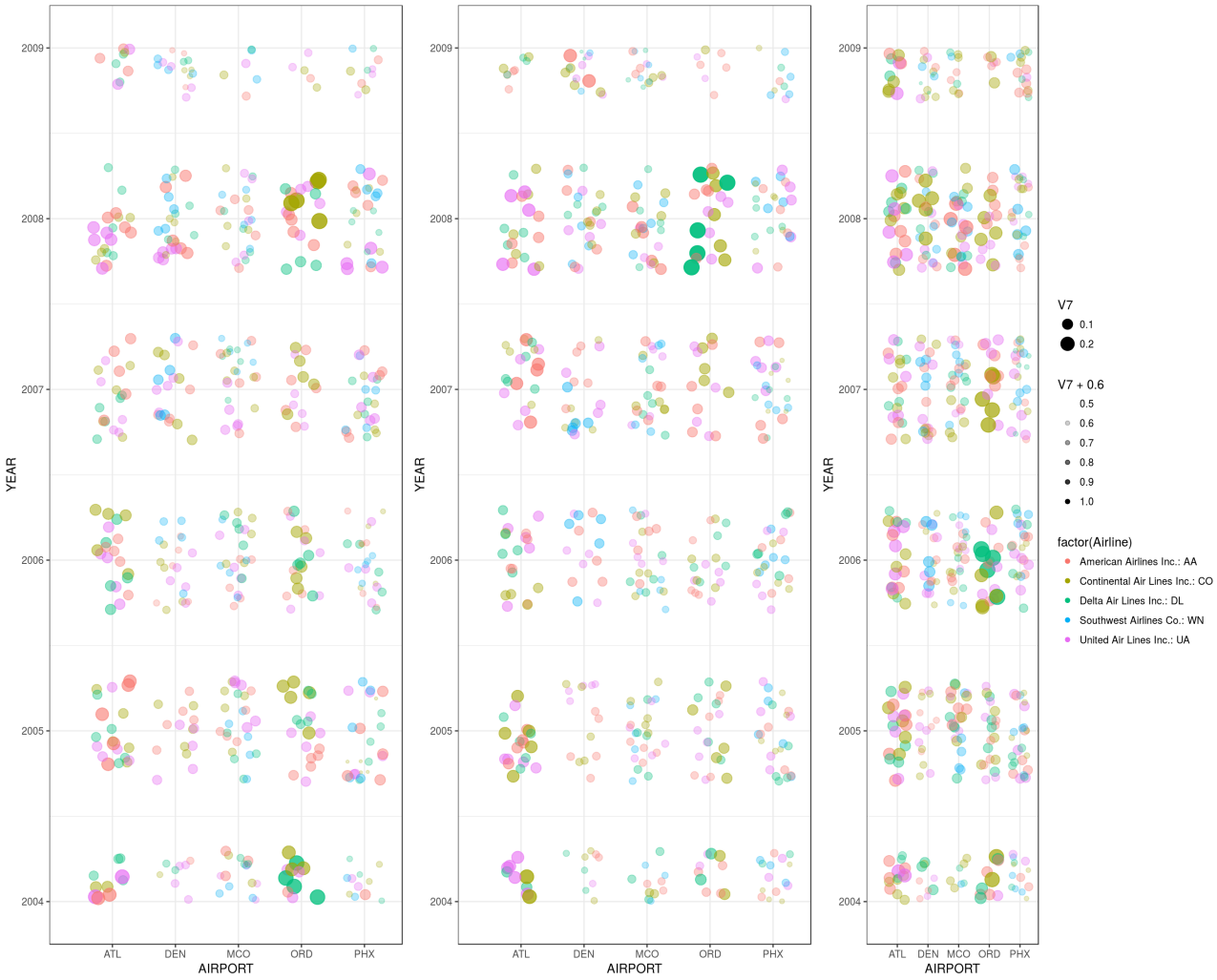
Airport_Code	Airport	Flight_Count
13930	Chicago, IL: Chicago O'Hare International	1518875
10397	Atlanta, GA: Hartsfield-Jackson Atlanta International	1230636
11292	Denver, CO: Denver International	1047164
13204	Orlando, FL: Orlando International	714296
14107	Phoenix, AZ: Phoenix Sky Harbor International	711443

Airline_Code	Airline	Flight_Count
19977	United Air Lines Inc.: UA	3122427
19393	Southwest Airlines Co.: WN	3024422
19790	Delta Air Lines Inc.: DL	2912467
19805	American Airlines Inc.: AA	2090660
19704	Continental Air Lines Inc.: CO	1541720

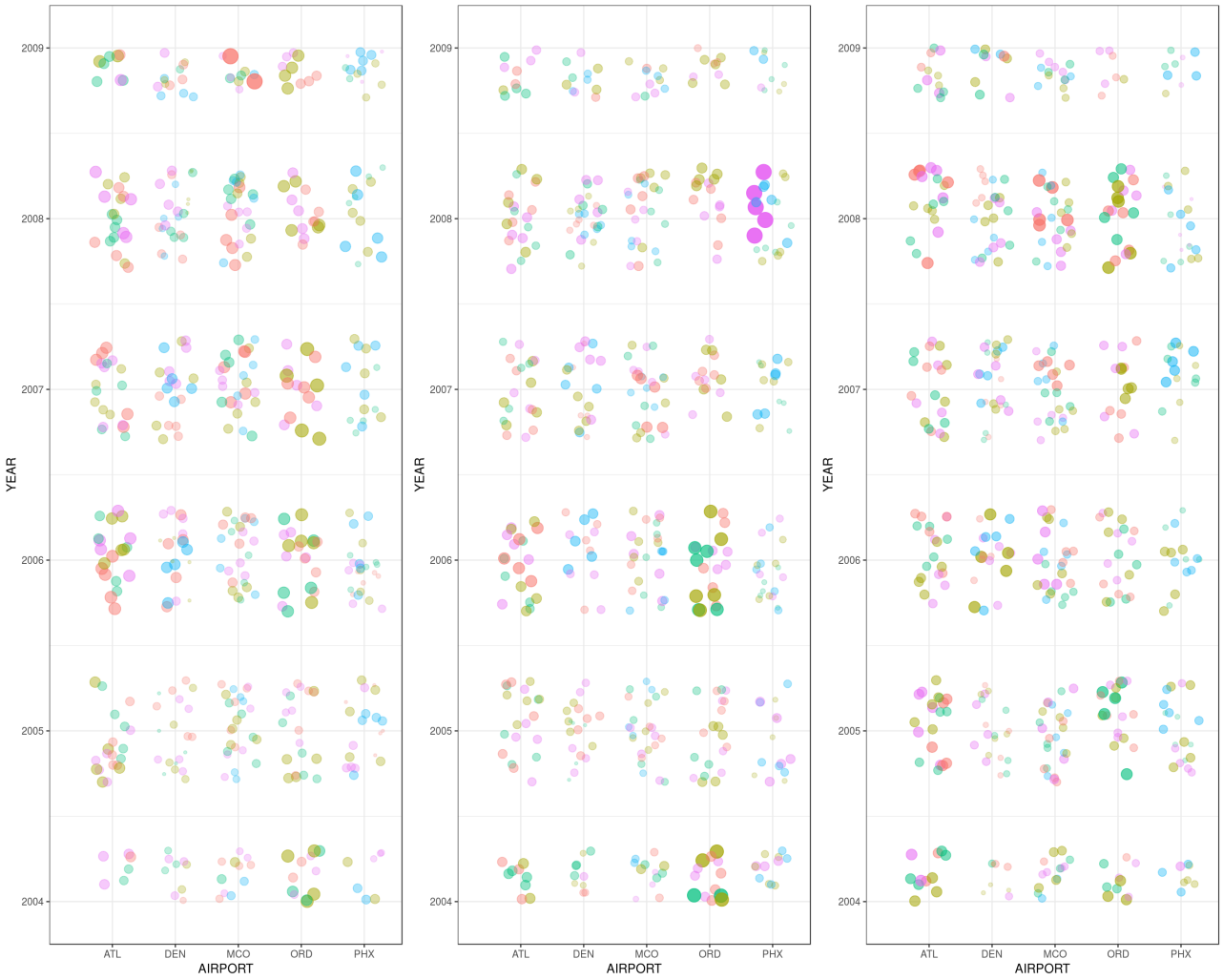
Analysis on the flight delays over the years (2004-2009)

We have reduced the dataset to the years 2004 to 2009 as the flight count was maximum during these years.
The below graphs were plotted for every quarter to analyse seasonal flight delays.
The colors in the graph represent the top 5 airlines. The X-axis represent top 5 Airports, and Y axis represents the years. The size of the bubble represent the mean flight delay. The hgher opacity represents higher normalized mean flight delay.

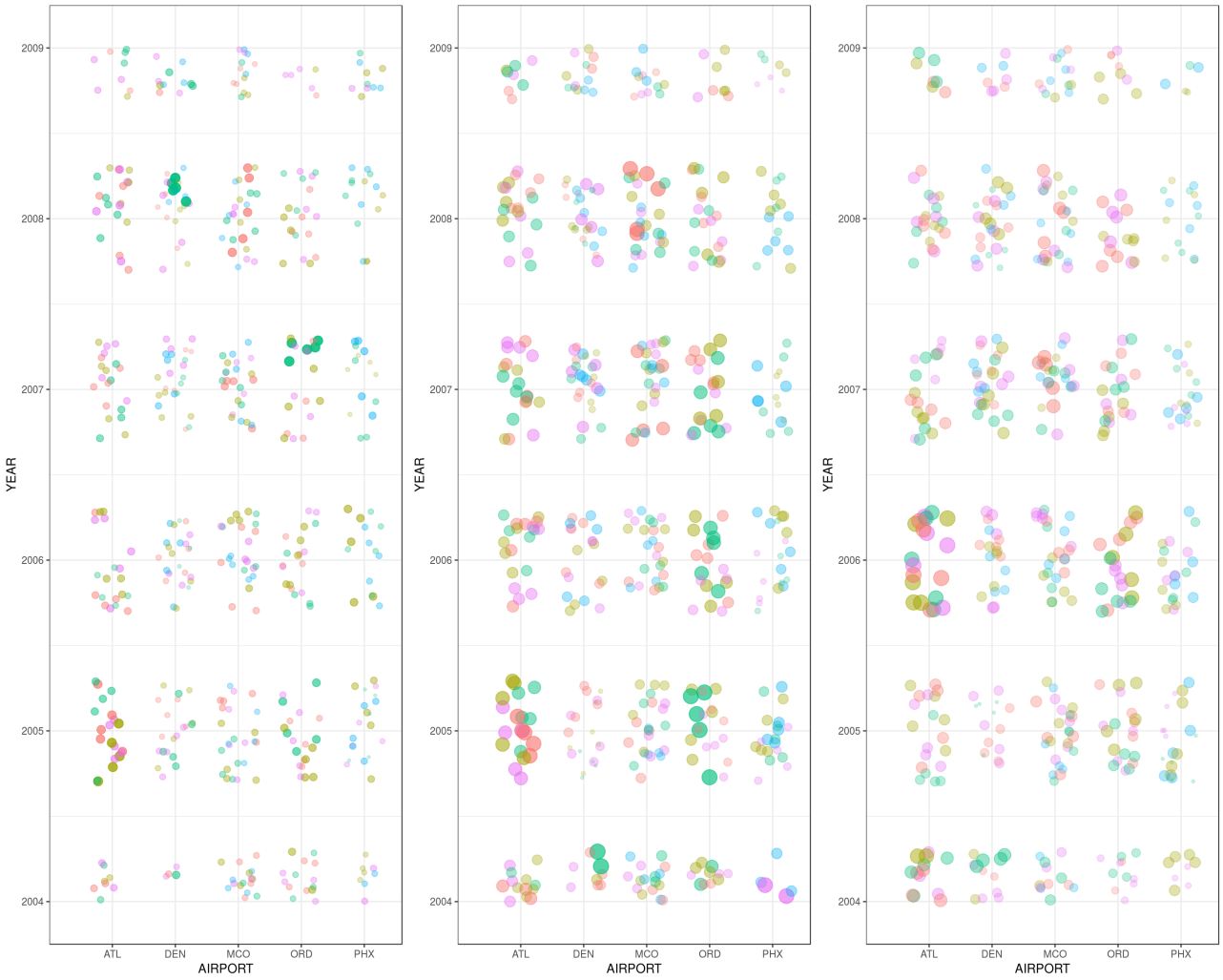
Mean Delay of top 5 airports and airlines over the years during Jan-Mar



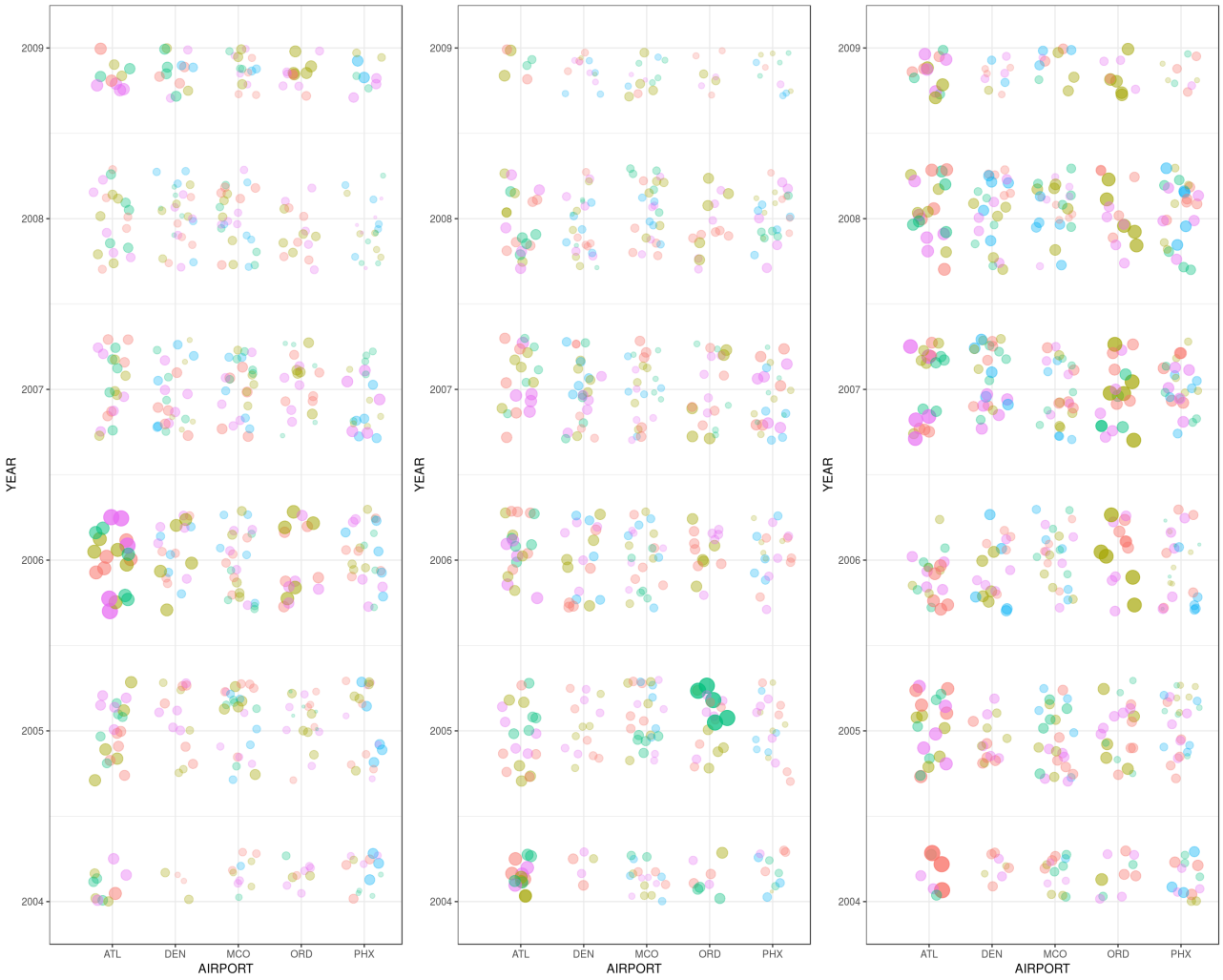
Mean Delay of top 5 airports and airlines over the years during Apr-June



Mean Delay of top 5 airports and airlines over the years during July-Sept



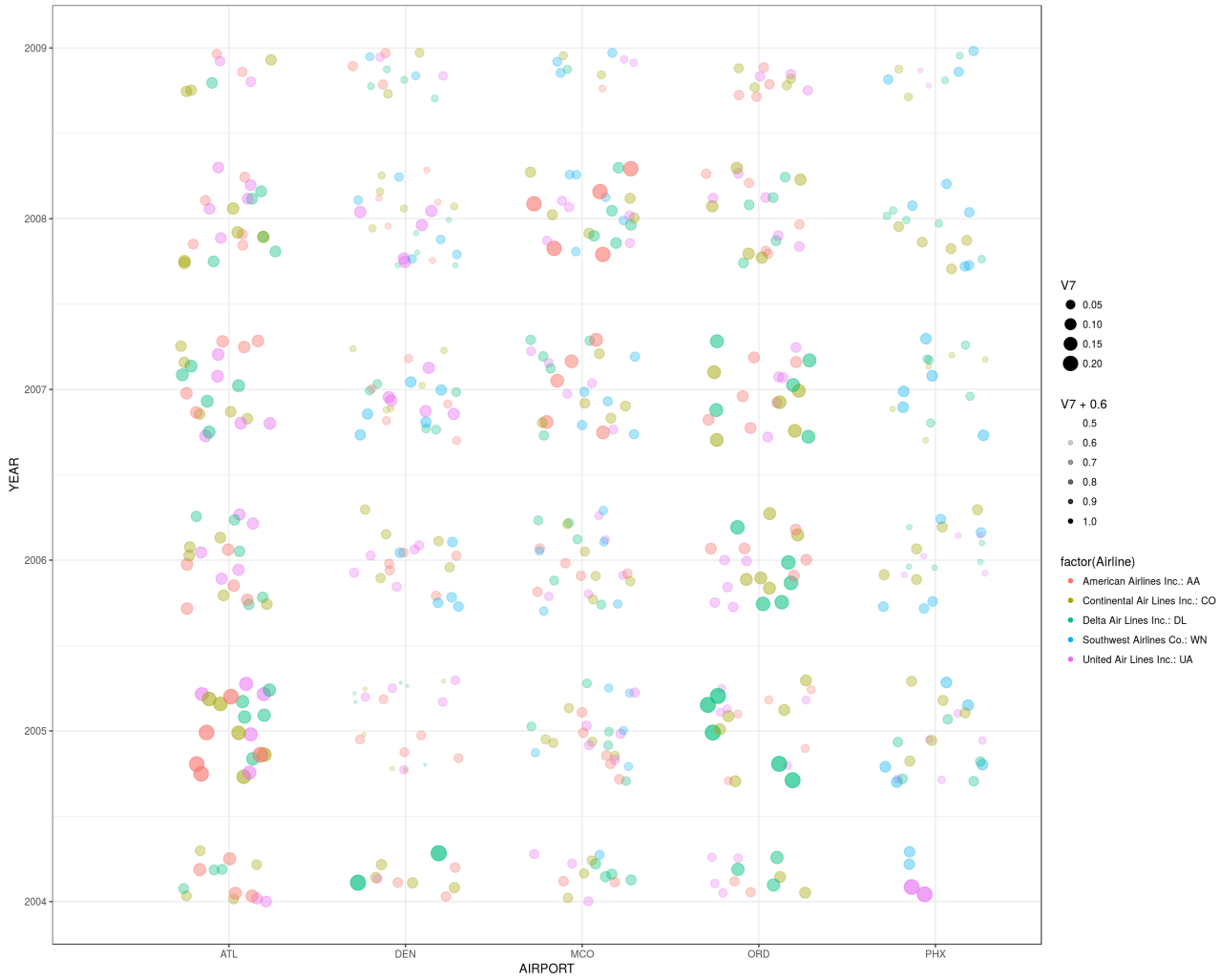
Mean Delay of top 5 airports and airlines over the years during Oct–Dec



Mean Delay of top 5 airports and airlines over the years during Aug

In this graph we found unusually high fligh delays for atlanta airport in August,2005. Later we found resources online suggesting the reasons for high flight delay and cancellation was due to hurricane Katrina. Below is a news link http://money.cnn.com/2005/08/30/news/katrina_air/index.htm (http://money.cnn.com/2005/08/30/news/katrina_air/index.htm). This event had a cascading effect on other airports as well. Similar delays were seen in Chicago.

Also, we can infer there were high delays during the july-august-september range every year, as it is the season for hurricanes. https://en.wikipedia.org/wiki/Timeline_of_the_2006_Atlantic_hurricane_season (https://en.wikipedia.org/wiki/Timeline_of_the_2006_Atlantic_hurricane_season)


2009
2008
2007
2006
2005
2004
YEAR
ATL
DEN
MCO
ORD
PHX
AIRPORT
V7
0.05
0.10
0.15
0.20
V7 + 0.6
0.5
0.6
0.7
0.8
0.9
1.0
factor(Airline)
American Airlines Inc.: AA
Continental Air Lines Inc.: CO
Delta Air Lines Inc.: DL
Southwest Airlines Co.: WN
United Air Lines Inc.: UA