## Assignment – Part A

Data about domestic flights in the US is publicly available from http://stat-computing.org/dataexpo/2009/the-data.html and http://stat-computing.org/dataexpo/2009/supplemental-data.html. Descriptions of the data are also available from these links.

You can use MapReduce, Pig, or Spark to analyze the data.

1. Explain your choice of processing framework briefly.

Answer the following questions. Explain your strategies, show your programs, and up to the first 20 records in the results. List your assumptions.

2. How many flights were there from JFK to LAX?
3. What was the sum and average of all arrival delays for all delayed flights?
4. What was the average departure delay for each state?
5. Which airline performed the worst seen from a customer perspective? (This can be answered in many different ways. Describe what you consider, e.g., arrival delays, cancelations, …)
6. Which airport performed the worst seen from a customer perspective? (This can also be answered in many different ways. Describe what you consider.)

You choose which year(s) you want to consider. On appserver2, I have downloaded the files 1987.csv (122MB), 2007.csv (671MB) 2008.csv (658MB), carriers.csv, and airports.csv. The files can be found in `/home/chr/data/flights.`

7. On appserver2 (and possibly your laptop), these files are just stored as ordinary files in the OS-managed file system. How would they be stored in HDFS running on a cluster? Which advantages/disadvantages would that give?