**IS669 – Group Project:**

**Part 1: Design phase.** For a maximum of **50** points

Use the site below for the data that will be needed for this part of this project:

<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7>

Variable descriptions

|  |  |  |
| --- | --- | --- |
|  | **Name** | **Description** |
| 1 | Year | 1987-2008 |
| 2 | Month | 1-12 |
| 3 | DayofMonth | 1-31 |
| 4 | DayOfWeek | 1 (Monday) - 7 (Sunday) |
| 5 | DepTime | actual departure time (local, hhmm) |
| 6 | CRSDepTime | scheduled departure time (local, hhmm) |
| 7 | ArrTime | actual arrival time (local, hhmm) |
| 8 | CRSArrTime | scheduled arrival time (local, hhmm) |
| 9 | UniqueCarrier | [unique carrier code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 10 | FlightNum | flight number |
| 11 | TailNum | plane tail number |
| 12 | ActualElapsedTime | in minutes |
| 13 | CRSElapsedTime | in minutes |
| 14 | AirTime | in minutes |
| 15 | ArrDelay | arrival delay, in minutes |
| 16 | DepDelay | departure delay, in minutes |
| 17 | Origin | origin [IATA airport code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 18 | Dest | destination [IATA airport code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 19 | Distance | in miles |
| 20 | TaxiIn | taxi in time, in minutes |
| 21 | TaxiOut | taxi out time in minutes |
| 22 | Cancelled | was the flight cancelled? |
| 23 | CancellationCode | reason for cancellation (A = carrier, B = weather, C = NAS, D = security) |
| 24 | Diverted | 1 = yes, 0 = no |
| 25 | CarrierDelay | in minutes |
| 26 | WeatherDelay | in minutes |
| 27 | NASDelay | in minutes |
| 28 | SecurityDelay | in minutes |
| 29 | LateAircraftDelay | in minutes |

Note below site where you have auxilliary files on the site to convert the **carrier** code and **airport** names.

**http://stat-computing.org/dataexpo/2009/supplemental-data.html**

Each member of each team will be assigned a year and will do these tasks with the file

1. extract the data from the site and **download the cvs file from the website**
2. create a Hadoop table and load all the file data into it. Prefix the table name with your name.
3. display a sample of the loaded rows
4. determine the three **airports** with the highest delay time (in hours)
5. determine the three **carriers** with the highest delay time (in hours)
6. determine overall which type of delay (arrivals or departures) is the largest for **airports**
7. collaborate on the presentation deck and a conclusions slide for the group

Each team member will capture the member individual results and put them in Powerpoint slides and present them in class with the rest of the team.

Use legible screenshots of the successful execution of the work necessary for each task and formatted results.

Each team member will upload to Classes the group Powerpoint deck including all the individual member slides plus a conclusions slide. **Prefix your file with your name or you will lose 5 points.**

Example of some records in spreadsheet format:

