**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:

[**http://stat-computing.org/dataexpo/2009/the-data.html**](http://stat-computing.org/dataexpo/2009/the-data.html)

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 that you have ancillary files on the site to convert the **carrier** and **airport** names.

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

Each member of each team will

* select a different year to extract the data from the site
* load it into Hadoop
* determine the five carriers with the highest delay time (in hours)
* determine the five carriers with the lowest delay time (in hours)

**Capture the output results as follows:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Year** | **Carrier** | **Total Delay** | **Arrival delay** | **Departure Delay** | **Carrier Delay** | **Weather**  **Delay** | **NAS Delay** | **Security Delay** | **Late Aircraft Delay** |

**Upload to Blackboard:**

1. **the output result as well as captured legible screenshots of the successful completion of the preparation steps:**
   1. **download of the cvs file from the website,**
   2. **loading of the data into Hadoop,**
   3. **any queries necessary for the analysis and summarization of the above results.**

**Example of an output in spreadsheet format:**

