MSCI: 6110 Big Data Management and Analytics
Group Project (35% of points)

1. General Description:

The goal of this project is to perform data analysis on a real-life topic. Students should form groups of 3 (2 or 4 only when approved by instructor). Each group should identify a topic of interest based on the given datasets (or identify and collect other data upon approval from instructor), transform the data into suitable schema, load the data into a Hadoop cluster, and perform analysis using Hive, R and the SparkR package. The project has two phases (Data Management and Data Analysis).

2. Specific Requirements and Deliverables

i. Phase 1: Big Data Management

The instructor will give a few candidate datasets. Each group should decide which dataset to analyze and propose their analysis objectives. The dataset should be used for the entire project (Phase 1 and 2). Other datasets can be used but upon instructor's approval.

The following steps should be followed:

(a) **Sign up your team members on ICON and choose top two datasets your team would like to work on (due 9/5 6pm). The instructor will help assign project topics if there are significant overlaps in selections.**

(b) Download/obtain the datasets, understand the formats, schema, and semantics of the data. Identify a set of data summarization tasks (e.g., daily average, histograms, heat map, etc) to better understand the data.

(c) Load the data into the Hadoop file system (HDFS). Create tables and load the data into the tables. Your tables should use partitions, buckets, and other Hive concepts discussed in class.

(d) Write a number of Hive Queries to accomplish the proposed data summarization tasks. Visualize the results (you may download the summary and use your own laptop for visualization)

(e) Propose analysis questions for Phase 2. Note the analysis should answer a real-life question of interest and importance. The analysis should be non-trivial, such as predictive analysis, regression, anomaly detection, clustering, etc. Draw conclusions based on the analysis results.

Deliverables:

(1) **A mid-term project report (Due 10/3, 6PM)**. The report should include detailed illustration of the above steps, including a detailed introduction to the dataset. Each group should submit a **SINGLE PDF** file on ICON. Feedback will be provided and you have a chance to adjust your plan for phase two. The report should be detailed, informative, and self-contained.

(2) **Give a mid-term presentation** (10/3 in class). Schedule will be announced later. Each member MUST attend and participate in the presentation. Each group has 15 minutes plus Q & A.

ii. Phase 2: Big Data Analytics

Use R/SparkR to perform the data analysis tasks proposed in Phase 1 and generate results. Charts, plots, maps, and other types of analysis results are desired. **Bonus will be given** if external data (from Web Service API) in XML or JSON formats are involved in the analysis of the given data. Note the external dataset must be used in a meaningful way.

Deliverable:

(1) **Give a final presentation (10/31 in class)**. Schedule will be announced later. Each member MUST attend and participate the presentation. Each group has 15 minutes plus Q & A.

(2) **A final project report (due 11/7 6PM)**. The report should include your report of Phase 1 and the detailed steps, results, and conclusions of the analysis in Phase 2. Charts, plots, maps and other visualized results should be included. Each group should submit a **SINGLE PDF** file on ICON.

3. Project Evaluation (100pts):

The instructor will evaluate the project based on the following aspects. Each aspect may receive a grade of (A: Excellent 100%, B: Good 90%, C: Fair 80%, D: Poor 60% or lower).

Phase 1 (50 pts):

(1) **Hive Tables (15pts).** Did the group successfully process the data and load the data into Hive? Did they implement the Hive data structures to improve query efficiency? Are they correct?
(2) **Data Summarization (20 pts).** What data summarizations tasks were performed by the group? Are they implemented correctly? Visualized? How informative are they? How challenging?
(3) **Analysis Plan (5 pts)** clearly present your analysis goal and name a few feasible analysis you would like to perform on the data for Phase 2. The analysis should be interesting and feasible.
(4) **Presentation and Reports (10 pts).** Clearly introduce the details of the dataset and the project steps in the reports. Present the project clearly within the required time in the presentation. Address questions properly. Participate in the Q&A discussion and submit peer review timely.

Phase 2 (50 pts + 5 pts):

(1) **Analysis and Conclusions (30 pts).** What analysis have you performed? What kind of results are generated? What are the conclusions? Are they meaningful/logically sound? Are the analysis challenging to implement? Are they correct? Advanced analysis tasks will receive higher rating.
(2) **Presentation (10 pts).** Clearly present your project within time limit. Address questions properly. Participate in the discussion in Q&A for peer review.
(3) **Project report writing (10 pts).** Clearly document all the steps in details. Report analysis results and your conclusion. See the deliverable requirements above.
(4) **Bonus (5pts):** XML or JSON data obtained from Web API are included in a meaningful way.

4. Peer evaluation of contributions

At the end of the semester, each student also needs to submit a document to evaluate how much each member in your team contributed in the project. Please list specifically the contribution of each member in your opinion, and give a final percentage. **For example: Alice:** Discussed the problem, implemented the Hive queries, **Bob:** Wrote the reports, **Carl:** Implemented the R code for Spark data analysis, wrote the reports. Contributions: Alice 33%, Bob 25%, Carl 42% (they should sum up to 100%).