**Big Data – Group Project – Phases**

**Project Phase I [10 points]. Delivery: 1-2 page document with:**

1. Description of the project
2. Identifying the business problem
3. Data source:
   1. Data collection strategy (e.g., API, available data)
      1. API information (e.g., Quota, endpoints)
      2. Regular expressions involved to parse data of interest?
   2. Size of the data (e.g., number of rows)
   3. Type of data (e.g., structured, unstructured, semi-structured)
   4. Quality of the data (e.g., missing values, duplicates)?
4. Big Data pattern: Alpha, Beta, etc.
5. What are the tasks involved?
   1. Data aggregation
   2. Data Visualization
   3. Data Mining
      1. Classification
      2. Regression
6. Technology/Libraries you plan to use:
   1. Apache Spark
   2. Apache NiFi
   3. Apache Hadoop
   4. Which AWS services did you used?

**Data:** Use open data.

Sources: Twitter data API, Postman APIs, “Finance data APIs”, “Baseball data API”, “Soccer data API”, Kaggle datasets (filter by size > 100 Mb), etc.