# Big Data Analytics Project Task List

| | |
|---|---|
| **Project Name:** | *Predicting Crime Rates using Taxi rides data* |
| **Team Members:** | 1) Carlos Petricioli (cpa253) |
| | 2) Valerie Angulo (vaa238) |
| | 3) Varsha Muralidharan (vm1370) |

| Task | Who | Start Date | End Date | Comments |
|---|---|---|---|---|
| **Data Planning Stage** | | | | |
| **Identify data sources** | ***All*** | Oct 23 | Oct 24 | |
| **Plan where data will reside** | ***All*** | Oct 24 | Oct 27 | *• We will initialy have about 250GB of data*<br>*• We will save it on Dumbo* |
| **Taxi Rides Processing** | | | | |
| **Write code to ingest data source 1** | Carlos | Oct 27 | Oct 31 | *•In this step, you'll read the data from the source and write it or copy it into HDFS* |
| **Write code to profile data source 1** | Varsha | Nov 2 | Nov 9 | *•This is to characterize the data and the range of values in each column*<br>*•You might notice unexpected values in a column - you may decide to normalize the values (e.g. Street vs. St. vs. street) in the ETL stage*<br>*•Find min, max, and averages*<br>*•Find min and max length of text fields* |
| **Write code to clean/format (ETL) data source 1** | Valerie | Nov 2 | Nov 9 | |
| **NYPD Crimes Processing** | | | | |
| **Write code to ingest data source 2** | Carlos | Oct 27 | Oct 31 | *•In this step, you'll read the data from the source and write it or copy it into HDFS* |
| **Write code to profile data source 2** | Varsha | Nov 2 | Nov 9 | *•This is to characterize the data and the range of values in each column*<br>*•You might notice unexpected values in a column - you may decide to normalize the values (e.g. Street vs. St. vs. street) in the ETL stage*<br>*•Find min, max, and averages*<br>*•Find min and max length of text fields* |
| **Write code to clean/format (ETL) data source 2** | Valerie | Nov 2 | Nov 11 | |
| **Weather data Processing** | | | | |
| **Write code to ingest data source 3** | Carlos | Oct 27 | Oct 31 | *•In this step, you'll read the data from the source and write it or copy it into HDFS* |
| **Write code to profile data source 3** | Varsha | Nov 2 | Nov 9 | *•This is to characterize the data and the range of values in each column*<br>*•You might notice unexpected values in a column - you may decide to normalize the values (e.g. Street vs. St. vs. street) in the ETL stage*<br>*•Find min, max, and averages*<br>*•Find min and max length of text fields* |
| **Write code to clean/format (ETL) data source 3** | Valerie | Nov 2 | Nov 11 | |

| Task | Who | Start Date | End Date | Comments |
|---|---|---|---|---|
| **Develop, Test, and Refine the Analytic** | | | | |
| **Design the analytic(s)** | ***All*** | Oct 23 | Nov 2 | *Based on our model* |
| **Code the analytic(s)** | ***All*** | Nov 11 | Nov 21 | *To try to relate crime, weather and taxi data* |
| **Test the analytic(s)** | ***All*** | Nov 21 | Nov 25 | *To see what patterns we get* |
| **Analyze results of analytic(s)** | ***All*** | Nov 25 | Dec 3 | *•Are the results what you expected?* |
| | ***All*** | Nov 25 | Dec 3 | *•Do you need to adjust the analytic(s)?* |
| **Iterate on the analytic** | ***All*** | Dec 3 | Dec 10 | *•To improve results, and/or to better understand results* |
| **Final analytic code due** | ***All*** | - | ***15-Dec-17*** | |