**TASKS TO BE COMPLETED:**

* Utilize nosql technology or Python bigquery with distributed big data models to handle the storage, cleansing, processing and retrieval of unstructured datasets.
* Understand the given dataset variables,
* Develop a data model based on Cassandra nosql or any other distributed big data models
* Implement the database model using Data Definition Language (DDL)
* Perform data loading/batch processing and other necessary Data Manipulation Language (DML) including the use of non tabular models (Document/Column/ Map) and collections.
* Perform read and write operations using CQL or bigquery via an API.

**TASKS DONE INDIVIDUALLY:**

* Data Understanding
* Data pre-processing if any
* NoSQL Database modelling
* Develop the NoSQL database model in Cassandra or other big data distributed technologies
* Load instances and perform read and write operations using clusters
* Design a number of queries to process certain ad hoc queries
* Link your database model with an API and execute the queries
* Communicate the results
* Use Visualization method to present some of the useful patterns and results discovered and
* communicate the importance of results visualization.

**Possible Data Analytics and Big Data Tools:**

* Cassandra or Apache Hbase (master slave):
* R
* Hadoop/Spark

**Project Document:**

* Executive summary
* Introduction
* Data model
* Database structure of key spaces, tables, collections, etc
* Database implementation
* Data manipulation and scripting
* Sample of CQL queries at least 4 queries 2 of which using collections
* Descriptive Analysis using API
* Reflection

NOTE: A very good work will receive between 70%-79% marks. A very good portfolio incorporates :

* Identifying issues in the dataset,
* The use of at least three data analytics models
* Experimental results analysis,
* Data and results visualization,
* Data cleansing and pre-processing,
* A very good structure of the report,
* A very good recommendation to the different stakeholders. All the items are properly linked.

NOTES:

[Through packages Revolution contributed to open source including rhdfs and rhbase, R users can directly ingest data from both the hdfs file system and the hbase database subsystems in Hadoop.](http://blog.revolutionanalytics.com/2015/06/using-hadoop-with-r-it-depends.html)

[Apache Hadoop: is a Big Data framework which uses HDFS(Hadoop Distributed File System) to Store the data and MapReduce framework to process that data.Java is used as native language to write MapReduce programs.](https://www.quora.com/What-is-the-relationship-between-Apache-Hadoop-HBase-Hive-and-Cassandra)

[Apache Hive is a batch processing framework which is used to process the data using a language called Hive Query Language(HQL).HQL is a sql wrapper on top of HDFS which prevents writing Mapreduce programs in Java.Instead one can use SQL like language to do their daily tasks.](https://www.quora.com/What-is-the-relationship-between-Apache-Hadoop-HBase-Hive-and-Cassandra)

[How APIs Work](https://readwrite.com/2013/09/19/api-defined/) eg: Facebook users undoubtedly appreciate the ability to sign into many apps and Web sites using their Facebook ID—a feature that relies upon Facebook APIs to work.

[3 easy ways to get your data into R](https://www.import.io/post/3-easy-ways-to-get-your-data-into-r/)