**Programme: Bachelor of Information and Communication Technologies (BICT) and Grad Dips**

**Course: (563.684-18-MC-31) - Big Data Analysis**

**Quarter: 03**

**Assignment type: Individual or Team (this depends on the class size)**

**Purpose**

The purpose of this course is to design, develop and implement a big data operational storage and link it with a big data platform using API.

**Skills**

Technical documentation, research and investigation, experimentations, big data analysis and processing, big data models building and visualization, big data platforms utilization, team work, organisation, future big data technologies, forward looking and innovative thinking, planning, analysis, assessment, integration, abstraction and high-level modelling.

**Weightage, Learning Outcomes, and Submission Dates and Methods**

|  |  |  |  |
| --- | --- | --- | --- |
| **ITEM** | **% Weight** | **Submission Date** | **Submission Method** |
| Project (45%) and Presentation (15%)  (LOs 1-5) | 60% | 5th of September of 2018, 16.00 pm  Presentations are held on September 6th 2018 in the class | One of the Team members is to submit the project document in Canvas. |

**Forming A Team**

A team of two to three will work together as data scientists and the team leader as the lead data scientist to prepare a technical report based on real application data by applying big data models and No SQL. The project background and resources are given later in this document. The team members must conduct an extensive investigation using the dataset provided, especially pre-processing, descriptive analysis, experimentations using a data analytics platform, results generations and analysis, results visualization (graphs & tables) and recommendations.

**Marking:**

See project grade rubric document on canvas

**Feedback:**

Feedback will be provided on the progress document during class time. In addition, a written feedback will be provided if the team submitted a draft to the lecturer before the deadline. This written feedback issues must be addressed in the final submission.

**Project Background**

This project involves processing and manipulating unstructured dataset that contains stories and comments from Hacker News from its launch in 2006. Each story contains a story id, the author that made the post, when it was written, and the number of points the story received. Hacker News is a social news website focusing on computer science and entrepreneurship. It is run by Paul Graham's investment fund and start-up incubator, Y Combinator. In general, content that can be submitted is defined as "anything that gratifies one's intellectual curiosity".

Each story contains a story ID, the author that made the post, when it was written, and the number of points the story received. You can use the BigQuery Python client library to query tables in this dataset in Kernels or use CQL in Cassandra or other open NoSQL platforms. Note that methods available in Kernels are limited to querying data. Tables are at bigquery-public-data.hacker\_news.[TABLENAME]. Fork this kernel to get started.

**Project Aim**

The aim of this project is to utilize NoSQL technology or Python BigQuery with distributed big data models to handle the storage, cleansing, processing and retrieval of unstructured datasets. The students are expected to 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) and then perform data loading/batch processing and other necessary Data Manipulation Language (DML) including the use of non tabular models (Document/Column/ Map) and collections. Finally, perform read and write operations using CQL or BigQuery via an API.

**Project Task**

Develop a technical report based on a full investigation of the Hacker News dataset provided on Canvas using big data technology (experimental). In your investigation, you should in depth perform

1. Data Understanding
2. Data pre-processing if any
3. NoSQL Database modelling
4. Develop the NoSQL database model in Cassandra or other big data distributed technologies
5. Load instances and perform read and write operations using clusters
6. Design a number of queries to process certain ad hoc queries
7. Link your database model with an API and execute the queries
8. Communicate the results
9. 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 / MongoDB/ Other NoSQL database technology
* R /WEKA/Python
* Hadoop/Spark
* Any other open source big data tools

**Resources**

* [Raw Dataset](https://www.kaggle.com/hacker-news/hacker-news/home)
* Technical Report Template
* Grade Rubric

**Presentation**

There will be 10-15 minutes presentation in the final week of the course. The weight of the presentation is 10% of the total assessment weight.

**Assessment Weight**

The project weight is 70% from the total grades of the course. The project is divided into two deliverables

* Report 45%
* Presentation 15%

**Due Date**

5th of September by 16:00 PM via canvas.