**Data Migration and Report Generation**

**Project Description:** Case study contains 2 sets of datasets. Each set contains College, Student and Apply datasets for 2010 and 2011 year.

Field (Column) information of datasets is as –

* College -> cName, State, enrollment, year
* Student -> sID, sName, GPA, year
* Apply   -> sID, cName, major, decision, year

**Download the attached file to get the dataset files.**

**Data Sets file naming format:**

Student\_yyyy, College\_yyyy, Apply\_yyyy

Requirement is to come up with script which does the following

a.     Ingest datasets from local server/edge node in to HDFS based on year. Let’s call this HDFS region as RAW region.

b.    Move the data to different region (Lets call this as validated region). Create external hive table for each datasets with proper partitions.

c.     Data/file pointed by Hive table should be encrypted. (optional)

d.    Create following views for reporting purposes

                 i.        **GPA** -> GPA's of students applying CS in CA (GPA-Grade Point Average)

               ii.        **Reject** -> Names, Year, ID's and  GPA's of students who were applied to CS and were rejected

              iii.        **Accept** -> Names, Year, ID's and college of students who were applied to any course and were accepted

**Proficiency level:** Level L2(Intermediate) case study for the experienced users who want to gain more practical knowledge of Hive and Hadoop eco system.

**Pre-requisite skills:** Proficiency in Hive, Hive Query Language (HQL), understanding of Hadoop Framework.

Architecture:

HDFS HIVE Tables

LFS

Student

Apply

College

Valid

Raww