**CS595 – Initial Project Proposal**

* **Project Topic:** Build a Data processing pipeline using Big Data SQL tools
* **Application subject area:** Entertainment & Movies
* **Data set Source:** IMDB Movies dataset from Kaggle

<https://www.kaggle.com/tomiandrep/imdb-filmid/data>

Data set of 1,000 most popular movies on IMDB in the last 10 years. The data points included are: Title, Genre, Description, Director, Actors, Year, Runtime, Rating, Votes, Revenue, Metascore.

<https://www.kaggle.com/rounakbanik/the-movies-dataset/data>

The main Movies Metadata file. Contains information on 45,000 movies featured in the Full MovieLens dataset. Features include posters, backdrops, budget, revenue, release dates, languages, production countries and companies.

* **Relevant Question/Problem:**

Build a data processing pipeline using Big Data technologies to ingest, curate, transform and use techniques to explore some interesting data and derive insights on Movies data from IMDB.

* **Proposed Approach: Tools for** 
  + **Data Ingestion:** HDFS commands to move data on to Hadoop environment
  + **Data Curation & Transformation:** SparkSQL (using DataFrames)
  + **Database:** NoSQL HBASE DB
  + **Data Mining & Analysis:** SQL and R on Big Data

* **Reference resources:**
* Some Kernels and existing sample code available on the respective datasets Kaggle link:
  + <https://www.kaggle.com/tomiandrep/imdb-filmid/data>
  + <https://www.kaggle.com/rounakbanik/the-movies-dataset/data>