CSP:554 – Project Proposal

Group name: **Panda**
Arinjay Jain: **A20447307**
Ayush Dadhich: **A20449379**
Parth Gupta: **A20449774**
Keerthana Nagula: **A20442972**

- **Project Topic:** Develop a Data handling pipeline with the aid of Big Data SQL tools

- **Application subject area:** Movies & Entertainment

- **Data set Source:** IMDB Movies dataset from Kaggle

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

  The main Movies Metadata file (Banik, "The Movies Dataset"). Files contain information on 45,000 movies utilized in the full MovieLens dataset. Data set include plot keywords, languages, production countries and represents a collection of 26 million ratings from all the 45,000 movies.

- **Project Motive / Question:**

  Develop a data processing pipeline using Big Data technologies to insert, modify, transform and implement Big Data techniques to research relation between proposed data and derive meaningful / relevant understanding of data streaming on movies data from IMDB.

- **Proposed Approach: Tools / Techniques for:**
  - ➢ **Data Insertion:** HDFS commands to move data on to Hadoop environment
  - ➢ **Data Modification & Transformation:** SparkSQL
  - ➢ **Database:** NoSQL HBASE DB
  - ➢ **Data Mining & Analysis:** SQL and Python / R on Big Data

- **Reference resources:**

Tomiandrep. "IMDB Filmid." *Kaggle*, Kaggle, 13 Dec. 2017, https://www.kaggle.com/tomiandrep/imdb-filmid/data.

Banik, Rounak. "The Movies Dataset." *Kaggle*, 10 Nov. 2017, https://www.kaggle.com/rounakbanik/the-movies-dataset/data.