CSP554 – Project draft & Literature Review

Team Name: Panda

Members: Arinjay Jain (A20447307), Ayush Dadhich (A20449379), Keerthana N (A20443972), Parth Gupta(A20449774)

- **Project Topic: Develop a Data handling pipeline with the aid of Big Data SQL tools**
- **Application subject area:** Movies & Entertainment
- **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

The following papers were researched while preparing for this project. They cover movie data analysis, big data technology tools and a combination of the two.

1. **Literature Review**

   a. **Mining Chinese social media UGC: a big-data framework for analyzing Douban movie reviews [1]**

   This paper addresses common Big Data problems of time constraints and memory costs involved with using standard single-machine and software. The authors propose a novel big data processing framework to investigate a niche subset of user-generated popular culture content on Douban, a well-known Chinese-language online social network. Huge data samples were harvested via an asynchronous scraping crawler and the rest of the research was implemented on big data technologies. Their major contributions were:

   i. An efficient framework implemented for large volumes of social media data processing based on the Hadoop platform. User-generated contents were collected, distributed, stored and processed on the Hadoop distributed file system (HDFS)

   ii. An asynchronous scraping crawler was implemented via a multiple-task queue to collect data in an efficient and simultaneous manner

   iii. A novel extraction, transformation and load (ETL) process was introduced

   iv. An improved Apriori algorithm based on MapReduce was proposed to increase the flexibility and efficiency of Big Data Mining.

   The authors conclude that the proposed framework offers a flexible capability and efficient applicability for the processing of large amounts of social media data that in turn can be fed back to producers and distributors of both commercial and user-generated digital media contents.

   b. **Scalable sentiment classification for Big Data analysis using Naive Bayes Classifier [2]**

   This paper evaluates the scalability of Naive Bayes classifier in large datasets to achieve fine-grained control of the analysis procedure. A Big Data analyzing system was also designed to help this study. A standard approach is to use Mahout, a machine learning library for clustering, classification and

filtering and is implemented on top of Hadoop. The authors compare the performance of Naive Bayes with the implementation using Mahout and built a big data analyzing system. This system adds four modules on Hadoop: work flow controller, data parser, the user terminal and the result collector. Each of these modules were implemented using MapReduce frameworks and data transfer to and from HDFS. The model was implemented on Virtual Hadoop cluster to enable testing of the Hadoop program in the cloud. The authors conclude that as data increases, the Naive Bayes Classifier model has 82% accuracy. The Hadoop implemented model shows faster performance and lesser processing time as the amount of data used crosses 2000K cases.

c. Movie Popularity Classification based on Inherent Movie Attributes using C4.5, PART and Correlation Coefficient [3]

This paper talks about the surplus research being done on classifying movies for future selection based on the attributes of already released movies. The authors mention that much can be predicted by considering parameters such as actors, directors, languages, countries, etc. and is an aspect of movie data that can be taken advantage of in prediction how a movie would perform. In this paper, the authors propose to analyse details of movies prior to their release and predict the success, revenue, ratings, etc. of a movie and compare it to the same post-release. This would be greatly useful to producers, financiers, academics and even viewers to understand the contributing factors that lead to a movie's success. The objective of this paper was to provide a suitable approach along with the necessary factors that were to be considered for developing pre-release and post-release movie datasets using Internet Movie Database (IMDB), classify data and interpret future predictions. This was accomplished using tools JMDB, SQL, WEKA, PART and Decision Trees. The results showed that decision trees perform well in prediction, directors and budget together play an important role in the success of a movie and the correlation between budget and foreign revenues is significantly high.

d. A Review Paper on Big Data: Technologies, Tools and Trends [4]

This paper talks about the rapid increase in generation of data and increase in internet population, thus explaining the need for Big Data Technologies. This paper covers the history of Big Data, Definition, Characteristics and Generation of Big Data. This paper also mentions the various categories of Big Data, its Management and details the major tools of Big Data: Hadoop, HDFS, MapReduce Frameworks, YARN, Hbase, Pig, ZooKeeper, HCatalog, Hive, Mahout, Oozie, Kafka, Spark and many more. The authors conclude by mentioning the applications of Big Data and the various analyses that can be done on Big Data.

e. Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data [5]

This paper aims to bridge the gap between 'real time monitoring' and 'early predicting' by building a minimalistic predictive model for the financial success of movies based on collective activity data of online users. The authors show that the popularity of a movie can be predicted much before its release by measuring and analyzing the activity level of editors and viewers of the corresponding entry to the movie in Wikipedia and Twitter. THe tools used in this research were Wikipedia, Toolserver, Mojo, Bots and used a 10-fold cross-validation to validate the multivariate linear regression model. They concluded that their model works more accurately for movies that are more popular and the volume of the related data is larger. They also concluded that most of the movies predicted by the Twitter method were among the successful ones.

2. Project Milestones

The following table enumerates the proposed milestones for this project. We are aiming to complete at least ~8 of these comparisons, leaving the remaining to be completed if time permits.

Sl. no.	Task	Team Member In charge	Due date
	Big Data Tools utilized:		
	Hadoop / HDFS		
1.	Data cleaning (Using R/Python)	**Keerthana**	**11/12**
2.	Data Imputation	**Parth**	**11/14**
3.	Moving data to HDFS	**Arinjay**	**11/16**
	DataBase: NOSQL HBase		
4.	Using HBase to run SparkSQL queries	**All**	
	Apache SparkSQL (DataFrame) / Python / R		
5.	Querying using SparkSQL (DataFrame). Compare:		
	a. Revenue vs Release Date	**Ayush**	**11/20**
	b. Revenue vs Ratings	**Ayush**	**11/20**
	c. Different Genres	**Ayush**	**11/20**
	d. (Vote_avg/Vote_count) vs Popularity	**Parth**	**11/20**
	e. Revenue vs Production Company	**Parth**	**11/20**
	f. Revenue vs Directors	**Parth**	**11/20**
	g. Revenue vs Run Time	**Arinjay**	**11/20**
	h. Run time vs Votes	**Arinjay**	**11/20**
	i. Revenue vs Budget Analysis	**Arinjay**	**11/20**
	j. Production Company vs Profit	**Keerthana**	**11/20**
	k. Months that have the maximum revenue	**Keerthana**	**11/20**
	Using Databricks		
6.	Build Visualizations - (Plots, graphs, heat maps, histograms)	**All**	**11/23**
7.	Final Review	**All**	**11/24**

8.	Complete Report	**All**	**11/24**

3. References

[1] Jie Yang & Brian Yecies “Mining Chinese social media UGC: a big-data framework for analyzing Douban movie reviews”
https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0037-9

[2] Bingwei Liu, Erik Blasch, Yu Chen, Dan Shen and Genshe Chen “Scalable sentiment classification for Big Data analysis using Naïve Bayes Classifier”
https://ieeexplore.ieee.org/abstract/document/6691740

[3] Khalid Ibnal Asad, Tanvir Ahmed and Md. Saiedur Rahman “Movie Popularity Classification based on Inherent Movie Attributes using C4.5, PART and Correlation Coefficient”
https://arxiv.org/ftp/arxiv/papers/1209/1209.6070.pdf

[4] Anurag Agrahari and Prof D.T.V. Dharmaji Rao “A Review paper on Big Data: Technologies, Tools and Trends”
https://www.irjet.net/archives/V4/i10/IRJET-V4I10112.pdf

[5] Márton Mestyán, Taha Yasseri and János Kertész “Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data”
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0071226

[6] Jaehui Park and Su-young Chi “An Implementation of a High Throughput Data Ingestion System for Machine Logs in Manufacturing Industry - IEEE Conference Publication”
https://ieeexplore.ieee.org/abstract/document/7536997

[7] Dataset : https://www.kaggle.com/rounakbanik/the-movies-dataset