Data Engineering Report - 16164571

Data Exploration
Data Exploring involves the visual exploration of a data set to gain a better understanding for the data set and its characteristics within. In this project we had 2 main types of datasets to be imported. These were movies.dat and ratings.dat . Movies contained the movie id and movie title in my implementation. I explored the data by importing it as a dataframe and viewing the data frame within my code. As there is not much correlation between the data and it is just an id and movie name I felt a graph visualisation wasn't necessary. I used methods such as .show() , .printSchema() and .take() to explore my dataset

I also added a small function to calculate the total number of movies. See def name_retriever(movieId)

The same went for the ratings data. I used dataframes to import the data and view it. I also counted the total number of ratings in the entire data set. I used methods such as .count(), .select() and .distinct() to count the number of items in datasets

Data Processing
Data processing involves cleaning and normalising our data and the product of this processing is the final data set. I initially processed the data into data frames and gave each column a title and a specific type to each column. From here I also added in my own ratings to the ratings data field using the .union() function to connect two datasets I then split the data into its specific training and testing data. I used a split ratio of 60:20:20 . I felt there was no need to normalize the data as the data was already in quite a normalized form.

Data Standardization
Standardization involves the bringing of data into a common format and value range. It is a virtual process for machine learning and it is important to scale your training and testing data. The ratings of the data all range between 0.0 and 5.0, hence there is no need to adjust the datas values. This data should provide a bias free result without the need for normalization or standardization.