RECOMMENDATION SYSTEM USING APACHE APARK

MS. ABINAYA, ASSISTANT PROFESSOR, M.KAILAS, KATHIRVEL, S.LOGAMBAL

[1]Assistant Professor, Department of CSE, Sri Ramakrishna Institute of Technology, Tamil Nadu, India.
[234]UG Student, Department of CSE, Sri Ramakrishna Institute of Technology, Tamil Nadu, India.

Abstract

Product database are repetitively becoming larger, making it progressively difficult for impartial systems to process products data. A vast amount of data is available on the internet in the form of ratings, ranks, assessments, thoughts, complaints, explanations, responses, and comments about any product, event, distinct, and services that may be used to make accurate decisions. Furthermore, there are numerous blog forums on the internet where web users can offer their opinions, assessments, and remarks about the things. For decision-making, a recommendation based on the rating and summary of relevant language about the items might be employed. The rise of e-commerce sites and online transactions is increasing the need for a robust recommendation system. Many people now purchase things from internet shopping platforms.. Approaches of praising innovative things take their boundaries, particularly for information discovery. Recommender systems have converted extremely common in current ages and are exploited in a various area some applications which are movies, music, news, books, research objects, search enquiries, social codes, and products. These systems are advantageous substitute to search algorithms as they benefit users to determine stuffs, they might not have set up by themselves.

I. INTRODUCTION

A. Background History

Recommender systems are computer programmes that make recommendations to users based on a variety of parameters. These algorithms forecast the product that people are most likely to buy and are most interested in. The recommender system works with a vast amount of data by filtering the most important information depending on the information provided by the user and other criteria take care of the user's preference and interest. It determines the compatibility of the person and the object, as well as the similarities between users and products, in order to make suggestions. These types of systems have helped both the users and the services delivered. The quality and decision-making process has also improved through these kinds of systems.

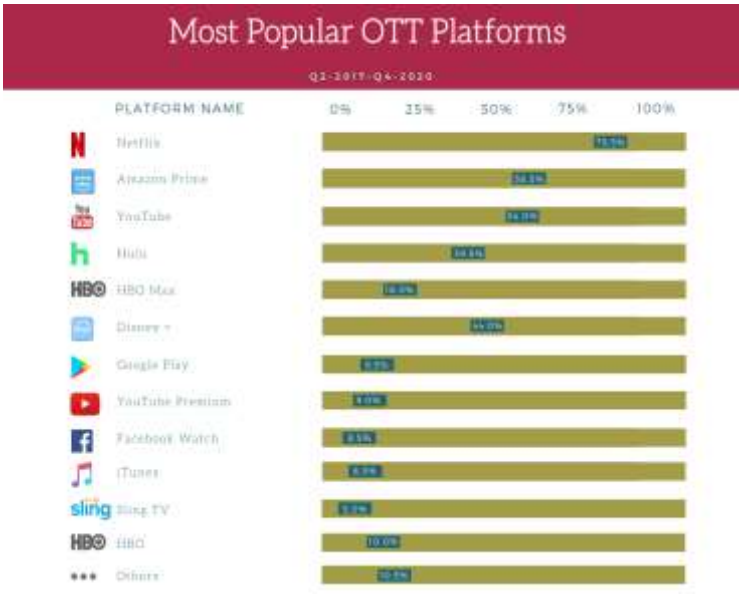

Figure 1.1

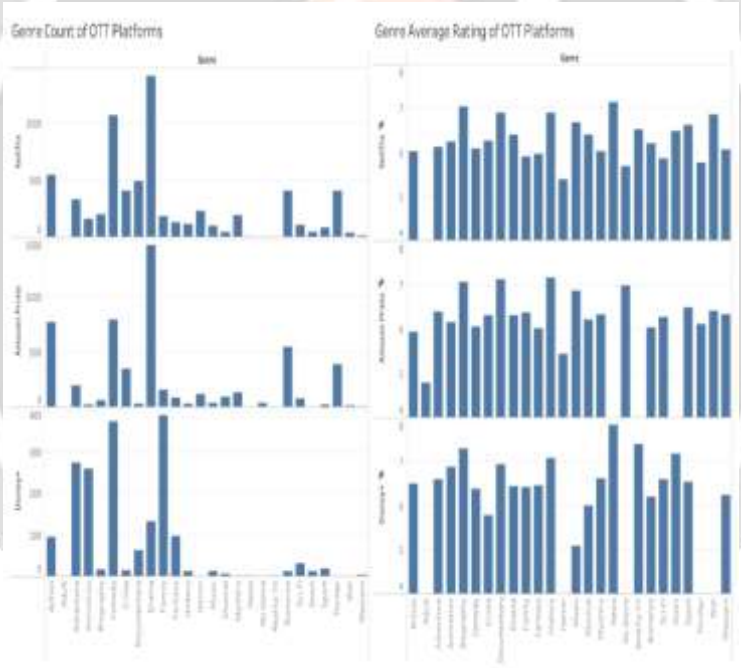

Figure 1.2

Problem Statement

Recommendation system has been used by enormous users. It used commonly in recent years and are used in a variable area in many popular applications which comprise of movies, songs, bulletin, files, research courses, online shopping, social networking sites, and products recommendation. Therefore, the goal of our project is to apply principles of machine learning using Apache spark's ML libraries to develop a recommendation model. When processing or analysing texts on a large scale, traditional recommender systems usually suffer from scalability, efficiency, and real-time recommendation issues. To get rid from these problems, a recommendation system is implemented in Apache Spark.

Scope

Recommender systems are critical in some industries as they Playlist generators for video and music services, product recommenders for online businesses, content recommenders for social media platforms, and open web content recommenders are all instances of recommender systems. The main aim of the project is to develop a recommendation model which can be scalable for small businesses and websites. The project primary focuses on movie recommendation but can be trans-compiled to other areas of focus like songs, articles, etc.

B. Existing System

In existing system conventional recommender systems frequently suffer from deficiency of scalability, efficiency, and real time recommendation problems while processing or analyzing documents taking place at huge scale. Existing system suffer from the problem of content-based filtering methods which are limited for small scale recommendation. The existing system cannot be considered as a template model, so that it cannot be applied to different area of use other than its trained field area. Also, existing system has cold start and data sparsity problems.

C. Proposed System

The content-based filtering, collaborative filtering, and hybrid recommender systems are the three types of recommender systems. The objective of project is to develop a hybrid recommendation system using apache spark. We develop a recommendation model which can be scaled for businesses. We aim to remove the hurdles Faced by conventional systems, that is cold start and data sparsity problem. We use apache spark because our recommendation model is applicable for large scale modeling. We have built a hybrid recommendation model for suggesting movies. We have clubbed together popularity-based model (using KNN) and collaborative filtered model (using spark' ALS). So that our model can be able to recommend both similar and popular movies around the time and genre-oriented movies. Therefore, we have used movie lens dataset consisting of around 27,00,000+ movies and user's ratings. We pre-process the data, build insights and then build and test our recommendation system. The final output we gain is a fully optimized and tested recommendation model working on movie recommendation.

Figure 1.3

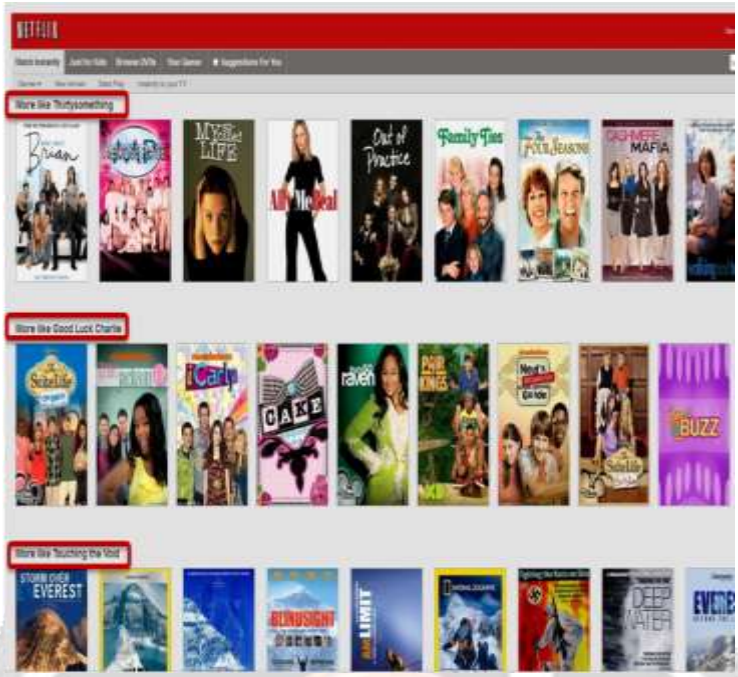

Figure 1.4

I. LITERATURE REVIEW

1) **A Review Paper on Product Ranking Based on Big Data**

Smita M. Deshpande R. S. Shirsath, In this paper based on distributed learning implemented on Apache Hadoop and spark. Here on Apache spark they are implementing our algorithms. They are using Apache spark over Hadoop system because Apache spark runs 100x faster than Hadoop MapReduce. Then we use several recommendation algorithms, such as collaborative filtering, content-based, hybrid, SVD, trustSVD, and so on, to make suggestions. They are using linear regression method for providing recommendations to users. Using this algorithm, they are predicting user ratings. To predict user ratings, they are factoring our matrix which was in HDFS into vectors and values. Then They associate this factorized data with test user input data, this value which

2) **Ranking Online Consumer Reviews**

Sunil Saumya Prakash Singh Abdullah Mohammed, Nripendra P. Rana. The random-forest classifier is used to categorise reviews as low or high quality. Only the gradient boosting regressor is used to predict the helpfulness scores of high-quality reviews. Low-quality reviews' helpfulness scores are not calculated because they will never be among the top k reviews. They are simply added to the review-listing website at the conclusion of the review list. The suggested system ensures that all high-quality reviews appear at the top of review listing pages and that all low-quality reviews appear at the bottom.

3) **Sentiment analysis using product review data**

Xing Fang Justin Zhan. They attempt to solve the problem of sentiment polarity categorization, which is one of the most difficult challenges in sentiment analysis, in this study. A generic process for categorising sentiment polarity is proposed, along with comprehensive process descriptions. The data for this study came from Amazon.com's online product reviews. Experiments on sentence-level categorization as well as review-level categorization have yielded promising results.

4) **Using Big Data to predict Amazon product ratings**

Jongwook Woo Monika Mishra. This paper aims to apply several machine learning (ML) models to the massive dataset present in e-commerce from Amazon to analyze and predict ratings and to recommend products. We provide a Big Data architecture suitable for big datasets for storing and computation, which is not achievable with standard architecture due to the size of the Amazon product review dataset. Furthermore, the dataset has roughly 7 million records and 15 attributes. With the dataset, they develop several models in Oracle Big Data and Azure Cloud Computing services to predict the review rating and recommendation for the items at Amazon.

5) **An intelligent approach to design of Ecommerce metasearch and ranking system using next-generation big data analytics**

Dheeraj Malhotra Omprakash Rishi. The goal of this study is to look into the limits of traditional search and page ranking methods in an ECommerce setting. The main goal is to help clients make an online purchase decision by delivering a tailored page ranking order of E-Commerce web links in response to an E-Commerce enquiry analyzing the customer preferences and browsing behavior.

6) **Review on the Product Ranking Methods**

Ahmad choirun Najib Nur Aini Rkahmawati. This paper aims to develop a Systematic Literature Review (SLR) to summaries existing research and finding new gaps in product ranking research. They develop SLR by defining inclusion criteria, initiating 7 preliminary findings, selecting primary studies, and summarizing the outcome of results. We proposed three dimensions as research questions. It includes product ranking item categories, product ranking methodologies, and dataset features for each study.

7) **Product ranking using hierarchical aspect structures**

Si Li Zhaoyan Ming Yan Leng Jun Guo. We introduce a novel hierarchical aspect-based product rating approach in this research. They first mine aspect-based pairwise comparative opinions from both user reviews on multiple review websites and community-based question answering pairs containing product comparison information. Next, they use hierarchical structure-based model to propagate and reassign the aspectbased comparative opinions by using the parent-child and sibling relations between aspects in the product aspect hierarchy.

8) **GroupLens: A Collaborative Filtering Architecture for Netnews**

Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom and John Riedl. People can use collaborative filters to make decisions based on the opinions of others. Group Lens is a system for collaborative filtering of Netnews, to help people find articles they will like in the huge stream of available articles. Clients for news readers show expected scores and make it simple for users to review items after reading them.

9) **Recommendation Systems: Principles, methods, and evaluation**

F.O.Isinkaye, Y.O.Folajimi and B.A.Ojokoh. On the Internet, where the number of choices is overwhelming, there is need to filter, prioritize and efficiently deliver relevant information to alleviate the problem of information overload, which Many Internet users may have a problem as a result of this. Recommendation systems solve this problem by searching through large volume of dynamically generated information to provide users with personalized content and services. This paper explores the different characteristics and potentials of different prediction techniques in recommendation systems to serve as a compass for research and practice in the field of recommendation systems.

Requirements specification

D. Hardware requirements

1. CPU: Intel i5 or above / AMD Ryzen -3 or above
2. Ram: 8GB or above 9
3. Hard-disk capacity: 20 GB
4. GPU: Nvidia GT- 740M or higher / AMD – RX 560 or higher
5. Other peripherals: Keyboard, Trackpad/Mouse, VGA-Color Monitor. quirements

E. Software requirements

1. Operating system: Linux or Windows or MacOS Platform architecture: x86 – 64
2. Programming languages: Python 3.x
3. Databook: Google Colab
4. Libraries needed: pandas, Streamlit, ScikitLearn, Matplotlib, Seaborn, Numpy, Streamlit, PySpark.

II. METHOLOGY

a. Information collection (IC) phase

This phase collects user's details to produce a model for prediction tasks which includes user's attributes, behaviors, or content of the resources. The system must gather much information about the user to provide best recommendations. A recommendation system will not produce correct and accurate output until the user model is well constructed. We have collected data form MovieLens.org consisting of around 27,00,000+ movies and user's ratings.

b. Learning Phase

This phase applies learning algorithms on the user's data which are obtained from the feedbacks in IC phase. The learning algorithms are the methods which are helpful in drawing out the patterns appropriate for application in certain situations. We apply and test ML algorithms. Particularly we apply KNN and spark's ALS Algorithms. After building the ML model we stress test it and tune it if necessary.

c. Prediction/recommendation phase

By analyzing the patterns which are obtained from learning phase, this phase provides recommendation or predictions for the given data. The trained data in learning phase provides certain patterns and then which are subjected to envision the user's course of action or future interests. The model built in learning phase (after testing and tuning) is deployed for user's action through a webapp.

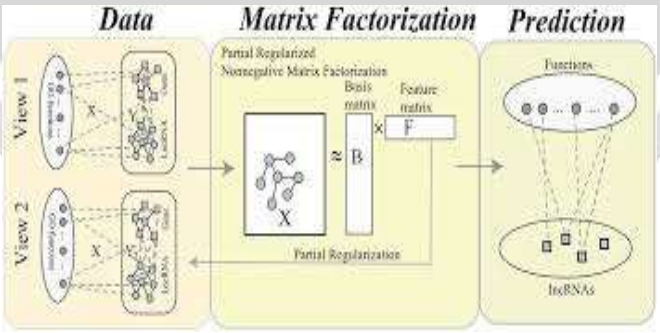

Figure 2.1

Data Collection

Data collection is the systematic gathering and measurement of information on variables of interest in order to answer research questions, test hypotheses, and evaluate outcomes.

```
In [229]: 1 t_ratings = pd.read_csv("ratings.csv")
          2 t_ratings.drop('timestamp', axis=1).head(10)
```

	userId	movieId	rating
0	1	1	4.0
1	1	3	4.0
2	1	6	4.0
3	1	47	5.0
4	1	50	5.0
5	1	70	3.0
6	1	101	5.0
7	1	110	4.0
8	1	151	5.0
9	1	157	5.0

Training data

	movie_id	title	user_id	rating
0	1	Toy Story (1995)	308	4
1	1	Toy Story (1995)	287	5
2	1	Toy Story (1995)	148	4
3	1	Toy Story (1995)	280	4
4	1	Toy Story (1995)	66	3

```
1 new_recommendation = new_test_predictions.join(spark_movies, 'movieId', 'left')\
2                                          .sort('prediction', ascending=False)
3 new_recommendation.show()
```

```
+-------+------+------+----------+--------------------+--------------------+
|movieId|userId|rating|prediction|               title|              genres|
+-------+------+------+----------+--------------------+--------------------+
|    318|   162|   5.0|   5.32176|Shawshank Redempt...|         Crime|Drama|
|    318|    52|   5.0| 5.1278366|Shawshank Redempt...|         Crime|Drama|
|    296|   348|   5.0|  5.078769| Pulp Fiction (1994)|Comedy|Crime|Dram...|
|    318|   122|   5.0| 5.0326896|Shawshank Redempt...|         Crime|Drama|
|    858|   417|   5.0| 5.0283055|Godfather, The (1...|         Crime|Drama|
|   2959|   296|   5.0| 5.0194983|   Fight Club (1999)|Action|Crime|Dram...|
| 112552|   515|   5.0|  5.001726|     Whiplash (2014)|              Drama|
|   1732|   122|   5.0| 4.9857078|Big Lebowski, The...|        Comedy|Crime|
|   1208|   465|   5.0|  4.984627|Apocalypse Now (1...|    Action|Drama|War|
|    364|    43|   5.0| 4.9801106|Lion King, The (1...|Adventure|Animati...|
|   1201|   171|   5.0|  4.975332|Good, the Bad and...|Action|Adventure|...|
|   1136|   348|   4.5| 4.9602985|Monty Python and ...|Adventure|Comedy|...|
|   1213|   220|   5.0|  4.957315|   Goodfellas (1990)|         Crime|Drama|
|   2959|   523|   4.5| 4.9512634|   Fight Club (1999)|Action|Crime|Dram...|
|  58559|    25|   5.0|  4.949626|Dark Knight, The ...|Action|Crime|Dram...|
|    608|   348|   5.0| 4.9489193|       Fargo (1996)|Comedy|Crime|Dram...|
|   1193|   597|   5.0| 4.9469843|One Flew Over the...|               Drama|
|    293|    99|   5.0| 4.9322643|Léon: The Profess...|Action|Crime|Dram...|
|  69844|   491|   5.0| 4.9227095|Harry Potter and ...|Adventure|Fantasy...|
|   2858|   417|   4.0|  4.881535|American Beauty (...|       Drama|Romance|
+-------+------+------+----------+--------------------+--------------------+
only showing top 20 rows
```

Testing data

MODULES

A. Loading data

Transferring information from one electronic file or database to another. Data loading entails transforming data from one format to another, such as from one type of production database to another vendor's decision support database.

B. Exploratory data and data visualization

EDA is simply a tool for better understanding and representing your data, allowing you to develop a more powerful and generalised model. EDA makes data visualisation simple, making it simple to explain our findings to others.

***C.* Training ml models**

Providing training data to an ML algorithm (that is, the learning algorithm) is the first step in training an ML model. The model artefact developed by the training process is referred to as an ML model.

D. Testing models

Model-based testing is a technique in which test cases are produced automatically from application models. It is a modern software testing approach that use a model, which is a secondary, lightweight implementation of a software build.

E. Fine Tunning if needed and further study

Tuning is the process of improving the performance of a model without overfitting or increasing variance. This is performed in machine learning by picking proper "hyper parameters." Hyper parameters can be viewed of as a machine learning model's "dials" or "knobs."

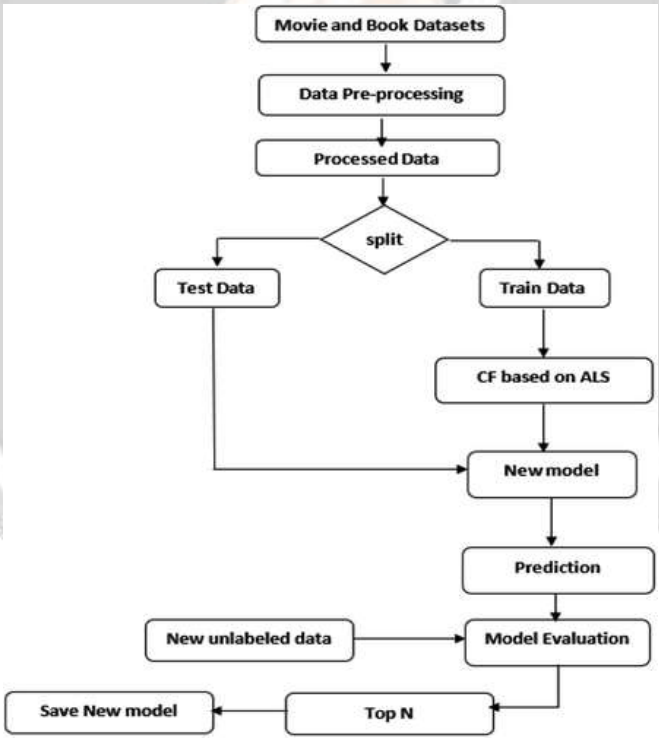

Figure 2.2

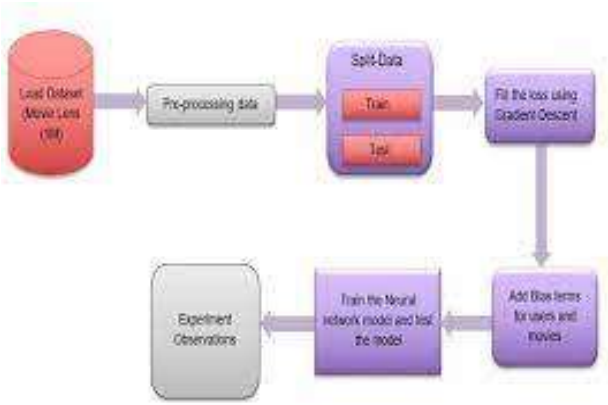

Figure 2.3

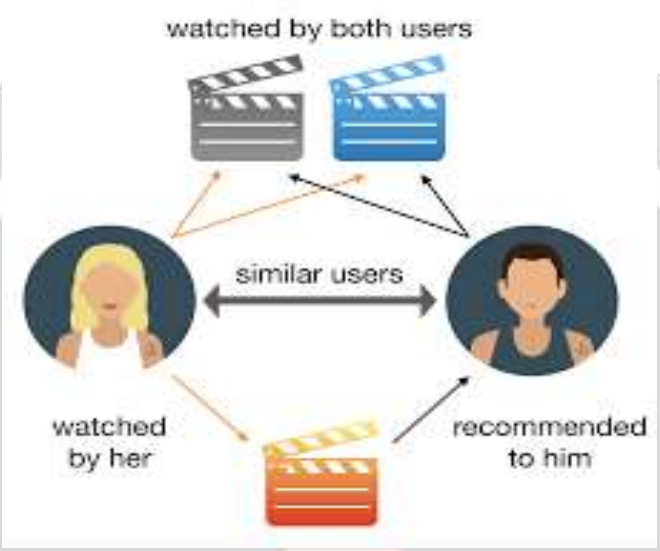

Figure 2.4

III. Results and Discussion

A . Data Analysis

we have made literature survey of different phases and various techniques in the recommender systems. It has been observed from the study that collaborative filtering has a better advantage over the other techniques and the user-user collaborative filtering gives efficient and correct results than the item-item filtering. To increase the quality of the results hybrid filtering is used. We have built a hybrid recommendation system by coupling together two ML models namely KNN, ALS. They provide better results compared to conventional recommendation systems and are scalable than it.

userId	movieId	rating	prediction
107339	148	4.0	3.3288345
93112	148	3.0	2.9263139
106148	148	2.5	2.7871637
234926	148	4.0	2.7707722
253535	148	4.0	2.7711174
207939	148	3.0	2.8659055
220572	148	2.0	2.7842884
244192	148	3.0	3.0389357
102642	148	4.0	3.34515
275860	148	3.0	2.8727908

only showing top 10 rows

Figure 2.4

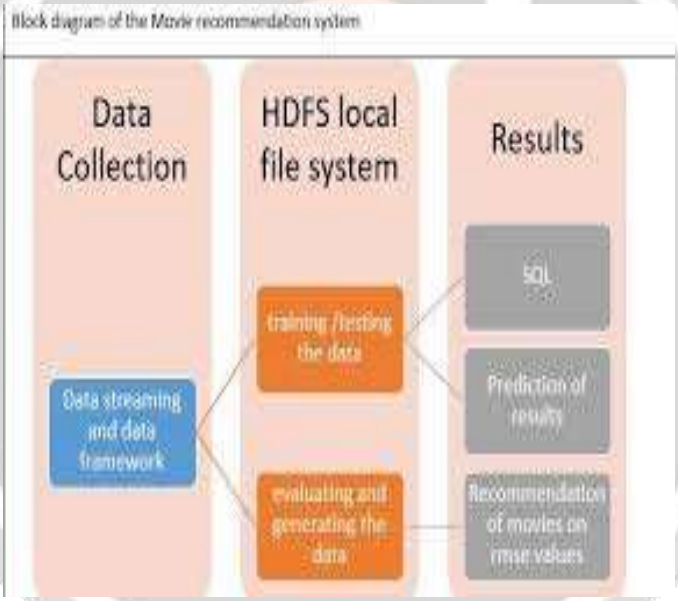

Figure 2.6

Future work

We can expand the data flow by parallel pipelining. This makes the model to alter it with the upcoming new data. We can cluster data into parts using Apache Hadoop (for large chunks of data) and process it parallelly. This can be further deployed into cloud for real-time learning. We can also fuse Deep learning models together for a better performance and wide area of deployment. By these we can be able to achieve a hybrid model which learns when there is a change is field of data.

REFERENCES

1. A Review Paper on Ranking of Product on Big Data - Smita M. Deshpande R. S. Shirsath
2. Ranking Online Consumer Reviews - Sunil Saumya Prakash Singh Abdullah Mohammed, Nripendra P. Rana
3. Sentiment analysis using product review data - Xing Fang Justin Zhan
4. Predicting the ratings of Amazon products using Big Data - Jongwook Woo Monika Mishra
5. An intelligent approach to design of Ecommerce metasearch and ranking system using next-generation big data analytics - Dheeraj Malhotra Omprakash Rishi
6. Review on the Product Ranking Methods - Ahmad choirun Najib Nur Aini Rkahmawati
7. Product ranking using hierarchical aspect structures - Si Li Zhaoyan Ming Yan Leng Jun Guo
8. A.G.Babu, S. S. Kumari, and K. Kamakshaiah, "An experimental analysis of clustering sentiments for opinion mining," in ICMLSC '17 Proceedings of the 2017 International Conference on Machine Learning and Soft Computing
9. Michael D. Ekstrand, John T. Riedl and Joseph A. Konstan., "Collaborative Filtering Recommender Systems", Foundations and Trends R in Human Computer Interaction.
10. Recommendation Systems: Principles, methods, and evaluation – F.O.Isinkaye, Y.O.Folajimi and B.A.Ojokoh
11. GroupLens: An Open Architecture for Collaborative Filtering of Netnews - Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom and John Ried