Amazon Product Co-Purchasing Network -Using Hadoop Framework

Leena Prajapati, Shubhi Shrivastav, Ruchi Agarwal

***Abstract*: These days in advanced age, any online shop you visit uses type of recommendation system. The recommendation system basically is data filtering devices that make use of estimations and data to endorse the most huge things to a particular user. The recommendation system deals with the huge amount of data and hence there is a need of Hadoop platform to manage and process the data. In this paper, MapReduce algorithm is implemented to process the data which is present in the form of ID, title, ratings, categories, ASIN (Amazon Standard Identification Number) etc. and proposed a content based recommendation system using the Hadoop framework to recommend the similar items as per user's liked and purchased items.**

***Index Terms*: Amazon co-purchase network, MapReduce, Hadoop Framework, Recommendation System, Big Data Analysis.**

I. INTRODUCTION

Shopping is a need of each individual and now in this growing age of internet, Online shopping has developed as one of the most popular Internet activities as in [5], providing a variety of products for consumers and a ton of sales challenges for e-commerce players. We will in general purchase items suggested by individuals since we confide in the individual. Also, these days in the advanced age, any online shop you visit uses a type of recommendation system. Recommendation systems are defined as a personalized data filtering technology used to either predict whether a particular user will like a particular item (prediction problem) or to identify a set of N items that will be of interest to a certain user as in [6]. We can think of these systems as a mechanized type of a "shop counter person". You approach him for an item, he demonstrates that item, as well as the related ones which you could purchase.

Amazon uses a recommendation system to suggest other frequently co-purchased items during sale of an item [1]. Like the frequently bought together is an option that is present on the page showing the combos and other special offers that makes you to buying those. The main goal of this recommendation is to increase sales by providing suggestions based on their purchasing history, their Wishlist and the items they have searched for. So, the information about the product can be used to identify other products. For this, categories, content similarities, and so on can be used to identify products that are similar and can be recommended to the users who have already brought one.

Based on the reviews and ratings of the product given by the customers Amazon display the products in the trending lists and in the recommended section. These attributes are then used to build a user's profile or model having the user' interests [10]. Amazon wants to make you shop more instead of just one product. It is just like Supermarket where you are lured to purchase more products by providing exciting offers.

The difference is Amazon is providing this facility in online mode. It will use the recommendations from the engine to email and keep you connected to the current trend of the product/ category.

The Dataset is collected from Amazon, which contains information about the various products. In the dataset, each product lists similar items pre-calculated by Amazon. Each data entry includes an ID, title, category, similar items to this item, and information about users who brought the item and each line of the result includes the customer IDs and product recommendations for that customer.

For such huge amount of data, we use MapReduce, which is a programming paradigm that runs in the background of Hadoop to provide scalability and easy data-processing solutions.

A. Big Data

'Big Data' is defined as a term which depicts a collection of information or data that is tremendous in size, but then developing exponentially with time or we can say when there the capacity of system is less than the amount of data needs to be processed. [4], then it refers to Bigdata. To put it plainly, such information is so extensive (in Petabytes and Zettabytes) and complex that none of the conventional methods or management tools can store it or process it effectively.

Big Data advances can be utilized for making organizing zone or landing zone for new information before distinguishing what information ought to be moved to a data warehouse. What's more, such mix of Big Data advancements and data warehouse causes association to offload rarely gotten with information.

B. Hadoop

Big data solutions (GOOGLE'S SOLUTION)

Dealing with huge set of scalable data is a frantic job and to operate on such data using a single database is a bottleneck. Google tackled this issue using an algorithm called MapReduce, which divides a task into sub-tasks and relegates these sub-tasks, to numerous PCs associated over the system, then gathers the outcomes to frame the last outcome dataset.

Revised Manuscript Received on October 22, 2019.

Leena Prajapati, Assistant System Engineer-Trainee at TATA CONSULTANCY SERVICES Ltd.,

Shubhi Shrivastav, Associate Software Engineer at Birlasoft Ltd.

Ruchi Agarwal, Associate Professor and HOD of BCA Department.

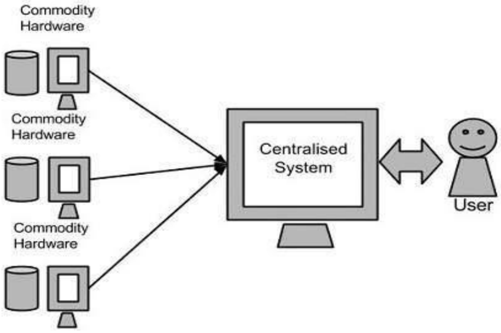


Fig. 1.1 Google's Solution

Above diagram (Fig. 1.16) shows various commodity hardwares which are connected to a single CPU machine or servers with higher capacity.

The applications uses MapReduce algorithm, which uses parallel programmimg model. In simpler words, the data is processed simultaneously on different machines where total factual and statistical calculation for an immense amount of information is performed.

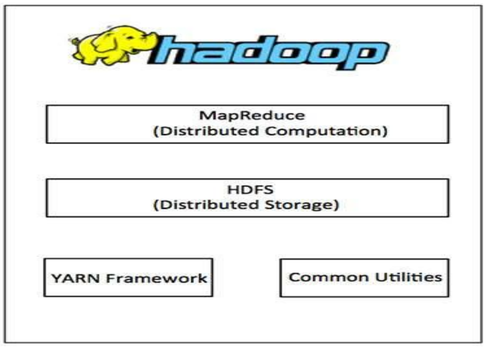


Fig. 1.2 Hadoop Framework

The Hadoop framework has following four modules as shown in Fig. 1.2:

- *Hadoop Common:* Hadoop module requires these Java libraries and utilities. These libraries give file system and OS level reflections and contains the important Java documents and contents required to begin Hadoop.
- *Hadoop YARN:* used for cluster resource management and job scheduling.
- *Hadoop Distributed File System (HDFS):* A distributed file system that provides high-throughput access to application data.
- *Hadoop MapReduce:* YARN-based framework for parallel processing of vast amount of data.

C. MapReduce

MapReduce is a programming model where large sets of data is distributed among the nodes. This helps in processing the Big Data parallelly on multiple nodes. This algorithm gives diagnostic abilities to examining gigantic volumes of complex information. MapReduce assembles an application into a sequence of pairs of Map and Reduce functions as in [11]. The input for the functions is present in HDFS file(s) and output is saved into HDFS files.

The MapReduce algorithm consists of two important tasks, Map and Reduce.

- Mapper Class is used to perform map tasks. This is the first step, which converts the input into tuples which further broke down into key-value pairs.
- Reducer Class performs the reduce task. The output of the previous task (i.e. Map task) is considered as input for reducer task, and hence the reducer task takes place after the Map task.

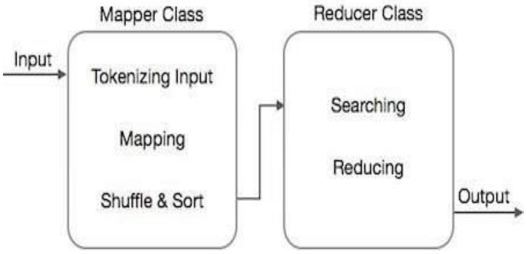


Fig. 1.3 Mapper and Reducer

As shown in Fig. 1.3 the input data of mapper class goes under tokenization, mapping and then sorting. The result of this class is considered as an input for Reducer class, which performs reducing operation on the imtermediate results to obtain final result.

Both the input information and output are put away in a document. The algorithm deals with planning, scheduling, observing and re-executing the unsuccessful tasks again. The shuffling process is the only mode of intercommunication between the nodes as in[9].

The MapReduce system consists of one master Job Tracker and slave Task Tracker per cluster-node. The master is in charge of managing resources,tracking their utilization and for scheduling the jobs on the slaves, checking these resources and executing the unsuccessful jobs again. The slaves runs the jobs as coordinated by the master and respond to the master occasionally.

D. Content-based Filtering

This is an approach to make recommendation system. These filtering methods are take in account the user's profile having his interests and the description of the products. [3].In these systems, keywords are used to describe the products and the user's profile is built based on the products a user likes.

The holistic approach of this method is that a customer is most likely to have more interest for the products he usually buys as in[8]. In other words, this algorithm basically tries to list out the items similar to the items a user liked as it is considered that the item similar to the item which has already been purchased or liked would be of user's interest. The content features of the user and the product are not considered by the recommendation algorithm. It regard these as a node as in [7]. For example, when we are prescribing a similar sort of thing like a motion picture or melody proposal.

II. LITERATURE REVIEW

The recommendation system by using AHP and hybrid collaborative filtering technique. After collecting data from websites, the analytics hierarchy process (AHP) to find a set of closest friends used and the hybrid collaborative filtering technique to generate the recommendation list applied as in [6]. A system using the Apriori algorithm or Association rule to understand the purchasing behavior of the buyers for product analysis by providing statistical data analysis using MapReduce technique to boost their sales as in [5]. The content based recommendation algorithm using Mapreduce parallel programming model. In this paper, two variations implemented, first variation generates the recommendation list and second variation generates the best recommendation for each customer as in [8]. A Personalized Recommendation Engine uses the hybrid filtering approach in item based recommendation algorithm. From this, dummy user records are obtained using Selenium in video or image format, and then data are extracted from images and stored in CSV files using MongoDB and HIVE. The recommendation algorithm is run on Mahout and stores the results in MongoDB database a in [9]. A recommendation system, for huge amount data present on the web using the Hadoop framework and the Mahout interface to analyze the data provided about movies in the form of ratings and reviews implemented as in [3].

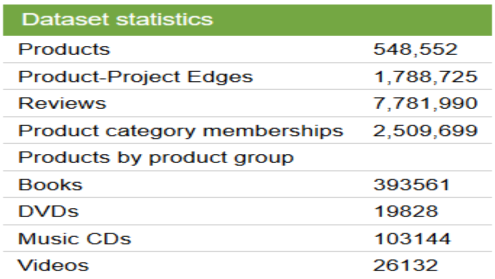
From these papers it can be stated that there are different approaches that have been taken to propose a recommendation system and various algorithms have been implemented using the Hadoop framework.

III. ANALYSIS & RESULTS

In this paper, we have done Content Based Recommendation, i.e., one could use information about the product to identify similar products and recommend them to the users who have already bought one.

We have used dataset collected from Amazon about products for making content based recommendations. Each product in dataset has pre-calculated similar item's list by Amazon. This dataset has the following fields:

- *Id:* Product Id (number 0, ..., 548551)
- *ASIN:* Amazon Standard Identification Number
- *Title:* Name of the product
- *Group:* Product group (Book, DVD, Video or Music)
- *Sales Rank:* Amazon Sales Rank
- *Similar:* ASINs of co-purchased products (people who buy X also buy Y)
- *Categories:* product category what kind of product is whether a book or a DVD etc(separated by |, category id in []
- *Reviews:* review information of product: user id, rating, votes on reviews,number og helpful reviews.



Dataset statistics	
Products	548,552
Product-Project Edges	1,788,725
Reviews	7,781,990
Product category memberships	2,509,699
Products by product group	
Books	393561
DVDs	19828
Music CDs	103144
Videos	26132

Fig. 3.2 Dataset Statistics

The statistics of the dataset are as shown in Fig. 3.2 which contains the product metadata and other information for about 548,552 different products (Books, music CDs, DVDs and VHS video tapes).

For doing the Content Based Recommendation, we have used MapReduce Algorithm. This means that there will be two jobs running for the complete analysis of the data i.e. Mapper and Reducer.

- *Mapper:* We have a class named 'MostFrequentUserFinder' which will execute the mapper's task
- *Reducer:* The second MapReduce job, i.e. Reducer is run and its class name is 'ContentBasedRecommendation'.

We have debugged the MapReduce program using Intellij IDEA, which is a Java integrated development environment (IDE). The JDK and JRE had already installed on Hadoop so, we set the project SDK as JDK to run Java programs on Intellij. All the required Jar files are added to provide the Hadoop module dependencies. Now the first mapper job is executed and the output of the mapper acts as an input to the second job (i.e., reducer), the output of this job will be the final desired output.

The map task of the first MapReduce job receives data about each product in a log file as a distinct key-value pair. When the product's data is send to map task, it emits the customer IDs as key and product's information as value.

Run the first MapReduce job '*MostFrequentUserFinder*' and after successful execution the Mapper job , the purchase data of each customer is extracted and the results will look like the following (as shown in Fig.3.3)

```
customerID=ATVPDKIKX0DER,review=A
SIN=0140445684#title=Capital : A Critique
of Political Economy (Penguin
Classics)#salesrank=17693#group=Book#rat
ing=5#similar=0451527100|0140445757|0879
757051|0553585975|1573921394|,review=ASI
N=0195110382#title=War at Sea: A Naval
History of World War
II#salesrank=631564#group=Book#rating=5
#similar=1585741485|0140246967|155750428
8|0374205183|0553802577|
```

Fig. 3.3 Intermediate output

Hadoop divides the input data into key-value pairs, then it runs reducer task by collecting all the values (i.e. products bought by customer) for each key (i.e. customer). Then the output of reducer task will contain customer and the list of products bought by them. For limiting the size of the dataset, the reducer will not emit any customer who has brought less than five products.

Run the second MapReduce job ContentBasedRecommendation with the output of the first MapReduce job and after successful execution, we will see the results as follows (as shown in Fig.3.4). Each row of the result will have a customer ID and the list of product recommendations.

```
ATVPDKIKX0DER [0451527100,
0140445757, 0879757051, 0553585975,
1573921394, 1585741485, 0140246967,
1557504288, 0374205183, 0553802577]
```

Fig. 3.4 Final output

The steps taken by the MapReduce task to make recommendations are:

- Hadoop create a list of similar items of the products bought by the user.
- Then, during the mapper phase, the product bought by the user gets eliminated from the list of similar items.
- Then ten items are selected as recommendations.

IV. FUTURE SCOPE

There are two future prospects that can be achieved later to make this project more feasible and efficient as: (1) This project is a basic recommendation system which recommends the products on the basis of gave data set information which gives the result as recommended product IDs instead of their names, which makes the output unreadable by a user as it could only be understood if we know every product ID which is not possible. Therefore, for future scope of the project we can try to get the names of the recommended products instead of their IDs. (2) We have used the content based filtering technique in which, to describe the items or products cetain keywords are used and besides this, a user's profile is built to state the type of item this user likes. While making this project more feasible and efficient, there are two more filtering methods that could be used are Collaborative filtering technique and Hybrid filtering technique.

V. CONCLUSION

The data are in the form of reviews and ratings which cannot be used directly. This kind of huge data needs to be first filtered and then used for recommendation system. In this paper, we have used Hadoop framework for implementing content-based recommendation. We have used Mapeduce algorithm which recommends similar products to the customers based on their previous buying patterns. This will increase the sales and revenue. The customer satisfaction will lead to customer retention.

REFERENCES

1. Basuchowdhuri, P., Shekhawat, M. K., & Saha, S. K., "Analysis of product purchase patterns in a co-purchase network," *Fourth International Conference on Emerging Applications of Information Technology,*2014.
2. Kadam, S. D., Motwani, D., &Vaidya, S. A., "Big Data analytics-recommendation system with Hadoop framework," *International Conference on Inventive Computation Technologies (ICICT),* 2016.
3. Verma, J. P., Patel, B. & Patel, A., "Big Data Analysis: Recommendation system with Hadoop framework.,"*International Conference on Computational Intelligence & Communication Technology (IEEE),* 2015.
4. Inodhini, S. Rajalakshmi, V. & Govindarajalu, B., "Building a personalized recommendation system with big data and Hadoop MapReduce.," *International Journal Of Engineering Research And Technology (IJERT), 3* (4), 2014.
5. Jaju, K., Nehe, V., &Konduri, A., "Commercial product analysis using Hadoop, MapReduce," *International Research Journal Of Engineering And Technology (IRJET), 3*(4), 2016.
6. Raval, K. P. & Tanna, S., "Data analytics using the Hadoop framework for effective recommendation in e-commerce based on social network knowledge," *IJARIIE-ISSN(O)*, 2 (3), 2016.
7. Wang, Q., "Design and implementation of recommender system based on Hadoop," *International Conference on Computational Intelligence & Communication Technology (IEEE)*, 2016.
8. Saravanan, S. , "Design of large-scale content-based recommender system using Hadoop MapReduce framework.," *International Conference on Computational Intelligence & Communication Technology (IEEE)*, 2015.
9. Sahu, U., Tripathy, A. K., Chitnis, A., Corda, K. A. & Rodrigues, S., "Personalized recommendation engine using Hadoop." *International Conference on Technologies for Sustainable Development (ICTSD)*, 2015.
10. Pessemier, T. D., Vanhecke, K., Dooms, S. & Martens, L., "Content-based recommendation algorithms on the Hadoop Map Reduce framework," *7th International conference on web information systems and technologies*, 2011.
11. Kang, S. J., Lee, S. Y. & Lee, K. M., "Performance comparison of openmp, mpi, and MapReduce in practical problems," *Advances in Multimedia Journal, Hindawi Publishing Corporation*, 2014.

AUTHORS PROFILE

Leena Prajapati is working as an Assistant System Engineer-Trainee at TATA CONSULTANCY SERVICES Ltd. She has completed her graduation in B. Tech (Computer Science) from JIMS Engineering Management Technical Campus in 2019. She has taken training in Big Data Hadoop, Python for Data Science and Machine Learning based on which she made various projects.

Shubhi Shrivastav is working as an Associate Software Engineer at Birlasoft Ltd. She has completed her graduation in B. Tech (Computer Science) from JIMS Engineering Management Technical Campus in 2019. She has taken training in Big Data Hadoop, Python for Data Science and Machine Learning based on which she made various projects.

Dr. RuchiAgarwal is an Associate Professor and HOD of BCA Department. She has done Ph.D. (Computer Science) from Birla Institute of Technology (BIT), Mesra, Ranchi in the field of data mining. She has more than 15 years of teaching experience. She has published various papers in international journals and conference proceedings. She has guided various B Tech and M.Tech level projects. She is guiding Ph.D. students in the area of Big Data Analytics. Her research interest areas are Big Data Analytics, Data Mining and Customer Analytics.