A Survey on Recommendation system in Big Data environment

M.Nandakumari[1], K.Boobalsingh[2]

[1] *PG Scholar, Sri Manakula Vinayagar Engineering College, Pondicherry, India*

[2] *Specialist, Temenos, Bangkok, Thailand*

***Abstract*—As the number of data sets increasing larger, the scalability and efficiency problems increases in fields like commerce and business, society administration, scientific researches. Recommendation system is one of the application being affected in the problem as scalability and efficiency. From the survey, Collaborative Filtering is one of the most frequently used filtering technique for recommendation system which can be implemented on a cloud computing tool named Hadoop, a widely-adopted distributed computing platform using the MapReduce parallel processing.**

***Keywords*—Big Data, Collaborative filtering, Hadoop, MapReduce, Recommendation System.**

I. INTRODUCTION

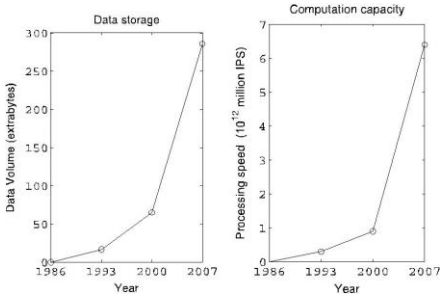
Big Data is the ocean of information we swim in every day vast zeta bytes of data flowing from our computers, mobile devices, and machine sensors. With Big Data solutions, we can dive into all data and gain valuable insights that were previously unimaginable[1]. In recent years, the amount of data in our world has been increasing explosively, and analyzing large data sets the so-called "Big Data" becomes a key basis of competition underpinning new waves of productivity growth, innovation, and consumer surplus. Then, what is "Big Data"?, Big Data refers to datasets whose size is beyond the ability of current technology, method and theory to capture, man-age, and process the data within a tolerable elapsed time. With the growing number of alternative services, effectively recommending services that users preferred have become an important research issue. Service recommender systems have been shown as valuable tools to help users deal with services overload and provide appropriate recommendations to them. Service recommender systems have been shown as valuable tools for providing appropriate recommendations to users. In the last decade, the amount of customers, services and online information has grown rapidly, yielding the big data analysis problem for service recommender systems. Consequently, traditional service recommender systems often suffer from scalability and inefficiency problems when processing or analyzing such large-scale data. Moreover, most of existing service recommender systems present the same ratings and rankings of services to different users without considering diverse users' preferences, and therefore fails to meet users' personalized requirements. Current recommendation methods usually can be classified into three main categories: content-based, collaborative, and hybrid recommendation approaches. Content-based approaches recommend services similar to those the user preferred in the past. Collaborative filtering (CF) approaches recommend services to the user that users with similar tastes preferred in the past. Hybrid approaches combine content-based and CF methods in several different ways. In CF based systems, users receive recommendations based on people who have similar tastes and preferences, which can be further classified into item-based CF and user-based CF. In item-based systems; the predicted rating depends on the ratings of other similar items by the same user. While in user-based systems, the prediction of the rating of an item for a user depends upon the ratings of the same item rated by similar users[2].

II. BIG DATA PROBLEM IN SEVERAL RESEARCH FIELDS

As more and more fields involve Big Data problems, ranging from global economy to society administration, and from scientific researches to national security, we have entered the era of Big Data. Recently, a report from McKinsey institute gives transformative potentials of Big Data in five domains: health care of the United States, public sector administration of European Union, retail of the United States, global manufacturing and personal location data.

Big Data has a deep relationship with e-Science, which is computationally intensive science which usually is implemented in distributed computing systems. Many issues on Big Data applications can be resolved by e-Science which require grid computing . e-Sciences include particle physics, bio-informatics, earth sciences and social simulations. It also provides technologies that enable distributed collaboration, such as the Access Grid. Particle physics has a well-developed e-Science infrastructure in particular because of its need for adequate computing facilities for the analysis of results and storage of data originating from the European Organization for Nuclear Research (CERN) Large Hadron Collider, which started taking data in 2009. e-Science is a big concept with many sub-fields, such as e-Social Science which can be regarded as a higher development in e-Science. It plays a role as a part of social science to collect, process, and analyse the social and behavioural data.

Other Big Data applications lies in many scientific disciplines like astronomy, atmospheric science, medicine, genomics, biologic, biogeochemistry and other complex and interdisciplinary scientific researches. Web-based applications encounter Big Data frequently, such as recent hot spots social computing (including social network analysis, online communities, recommender systems, reputation systems, and prediction markets), Internet text and documents, Internet search indexing. Alternatively, There are countless sensor around us, they generate sum less sensor data that need to be utilized, for instance, intelligent transportation systems (ITS) are based on the analysis of large volumes of complex sensor data. Large-scale e-commerce are particularly data-intensive as it involves large number of customers and transactions. In the following subsections, we will briefly introduce several applications of the Big Data problems in commerce and business, society administration, scientific researches and future research field(cloud computing)[3].



- ***Big Data in commerce and business***

According to estimates, the volume of business data worldwide, across almost companies, doubles every 1.2 years. Taking retail industry as an example, we try to give a brief demonstration for the functionalities of Big Data in commercial activities. There are around 267 million transactions per day in Wal-Mart's 6000 stores worldwide. For seeking for higher competitiveness in retail, Wal-Mart recently collaborated with Hewlett Packard to establish a data warehouse which has a capability to store 4 petabytes of data, i.e., 4000 trillion bytes, tracing every purchase record from their point-of-sale terminals. Taking advantage of sophisticated machine learning techniques to exploit the knowledge hidden in this huge volume of data, they successfully improve efficiency of their pricing strategies and advertising campaigns. The management of their inventory and supply chains also significantly benefits from the large-scale warehouse.

In the era of information, almost every big company encounters Big Data problems, especially for multinational corporations. On the one hand, those companies mostly have a large number of customers around the world. On the other hand, there are very large volume and velocity of their transaction data. For instance, FICO's falcon credit card fraud detection sys-tem manages over 2.1 billion valid accounts around the world. There are above 3 billion pieces of content generated on Face-book every day. The same problem happens in every Internet companies. The list could go on and on, as we witness the future businesses battle fields focusing on Big Data.

- ***Big Data in society administration***

Public administration also involves Big Data problems. On one side, the population of one country usually is very large. For another, people in each age level need different public services. For examples, kids and teenagers need more education, the elders require higher level of health care. Every person in one society generates a lot of data in each public section, so the total number of data about public administration in one nation is extremely huge. For instance, there are almost 3 terabytes of data collected by the US Library of Congress by 2011. The Obama administration announced the Big Data re-search and development initiative in

2012, which investigate addressing important problems facing the government by make use of Big Data. The initiative was constitutive of 84 different Big Data programs involving six departments. The similar thing also happened in Europe. Governments around the world are facing adverse conditions to improve their productivity. Namely, they are required to be more effective in public administration. Particularly in the recent global recession, many governments have to provide a higher level of public services with significant budgetary constraints. Therefore, they should take Big Data as a potential budget resource and develop tools to get alternative solutions to decrease big budget deficits and reduce national debt levels. According to McKinsey's report, Big Data functionalities, such as reserving informative patterns and knowledge, provide the public sector a chance to improve productivity and higher levels of efficiency and effectiveness. European's public sector could potentially reduce expenditure of administrative activities by 15–20 percent, increasing 223 billion to 446 billion values, or even more. This estimate is under efficiency gains and a reduction in the difference between actual and potential aggregate of tax revenue. These functionalities could speed up year productivity growth by up to 0.5 percentage points over the next decade.

- ***Big Data in scientific research***

Many scientific fields have already become highly data-driven with the development of computer sciences. For instance, astronomy, meteorology, social computing , bioinformatics and computational biology are greatly based on data-intensive scientific discovery as large volume of data with various types generated or produced in these science fields. How to probe knowledge from the data produced by large-scale scientific simulation? It is a certain Big Data problem which the answer is still unsatisfiable or unknown.

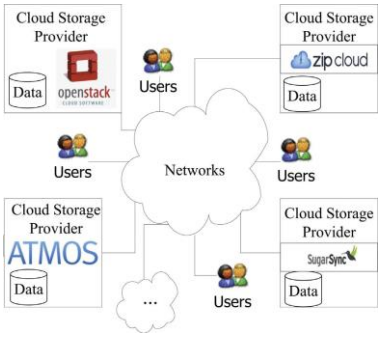
For instances, a sophisticated telescope is regarded as a very large digital camera which generate huge number of universal images. For example, the Large Synoptic Survey Telescope (LSST) will record 30 trillion bytes of image data in a single day. The size of the data equals to two entire Sloan Digital Sky Surveys daily. Astronomers will utilize computing facilities and advanced analysis methods to this data to investigate the origins of the universe. The Large Hadron Collider (LHC) is a particle accelerator that can generate 60 terabytes of data per day. The patterns in those data can give us an unprecedented understanding the nature of the universe. 32 petabytes of climate observations and simulations were con-served on the discovery supercomputing cluster in the NASA Center for Climate Simulation (NCCS). The volume of human genome information is also so large that decoding them originally took a decade to process. Otherwise, a lot of other e-Science projects are proposed or underway in a wide variety of other research fields, range from environmental science, oceanography and geology to biology and sociology. One common point exists in these disciplines is that they generate enormous data sets that automated analysis is highly required. Additionally, centralized repository is necessary as it is impractical to replicate copies for remote individual research groups. Therefore, centralized storage and analysis approaches drive the whole system designs.

- ***Big data in cloud computing***

The development of virtualization technologies have made supercomputing more accessible and affordable. Powerful computing infrastructures hidden in virtualization software make systems to be like a true physical computer, but with the flexible specification of details such as number of processors, memory and disk size, and operating system. The use of these virtual computers is known as cloud computing, which has been one of the most robust Big Data techniques. The name of cloud computing comes from the use of a cloud-shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams. It entrusts remote services with a user's data, software and computation. The combination of virtual machines and large numbers of affordable processors has made it possible for internet-based companies to invest in large-scale computational clusters and advanced data-storage systems.

As illustrated in above figure, cloud computing not only delivers applications and services over the Internet, it also has been extended to infrastructure as a service, for example, Amazon EC2, and platform as a service, such as Google AppEngine and Microsoft Azure. Infrastructure vendors provide hardware and a software stack including operating system, database, middleware and perhaps single instance of a conventional application. Therefore, it shows out illusion of infinite resources without up-front cost and fine-grained billing. It leads to the utility computing, i.e., pay-as-you-go computing.

Surprisingly, the cloud computing options available today are already well matched to the major themes of need, though some of us might not see it. Big Data forms a framework for discussing cloud computing options. Depending on special need, users can go into the marketplace and buy infrastructure services from providers like Google and Amazon, Software as a Ser-vice (SaaS) from a whole crew of companies starting at Salesforce and proceeding through NetSuite, Cloud9, Jobscience and Zuora-a list that is almost never ending. Another bonus brought by cloud environments is cloud storage which provides a possible tool for storing Big Data. Cloud storage have good extensibility and scalability in storing information.



Cloud computing is a highly feasible technology and attract a large number of researchers to develop it and try to apply to Big Data problems. Usually, we need to combine the distributed MapReduce and cloud computing to get an effective answer for providing petabyte-scale computing. Cloud View is a framework for storage, processing and analysis of massive machine maintenance data in a cloud computing environment, which is formulated using the Map/Reduce model and reaches real-time response. In, the authors extended Map/Reduce's filtering aggregation programming model in cloud environment and boosts the performance of complex analysis queries.

Apart from its flexibility, cloud computing addresses one of the challenges relating to transferring and sharing data, be-cause data sets and analysis results held in the cloud can be shared with others. There are a few disadvantages in cloud computing. The obvious one is the time and cost that are required to upload and download large quantities of data in the cloud environment. Otherwise, it becomes more difficult to control over the distribution of the computation and the under-lying hardware. Furthermore, there are privacy concerns relating to the hosting of data sets on publicly accessible servers, as well as issues related to storage of data from human studies. It is right to say that Big Data problems will push the cloud computing to a high level of development.

III. TYPES OF FILTERING APPROACHES:

1. Content based filtering

Filtering that is widely used in recommender systems is content-based filtering. Content based filtering methods are based on the information about the items that are going to be recommended. In other words, these algorithms try to recommend the items similar to those that a user liked in the past. In particular, various candidate items are compared with items previously rated by the user and the best-matching items are recommended. This approach has its roots in information retrieval and information filtering research. Basically those methods use an item profile i.e. a set of attributes (features) characterizing the item within the system. The system creates a content based profile of users based on a weighted vector of item features. The weights denote the importance of each feature to the user and can be computed from individually rated content vectors using a variety of techniques. Simple approaches use the average values of the rated item vector while other sophisticated methods use Bayesian Classifiers (and other machine learning techniques, including clustering, decision trees, and artificial neural networks) in order to estimate the probability that the user is going to like the item. Content based system has features such as simplicity and effectiveness, but also some drawbacks: It is difficult to distinguish the quality

of the filtering results from the same subject. Since the quantity of information increases rapidly, the information of the same subject increases too, making the efficiency and quality of the content-based system much reduced in a long term.

2. *Collaborative filtering*

Collaborative filtering technology is one of the method widely used in recommender systems. Compared with the content-based filtering system, collaborative filtering system could automatically filter the information that the system could not analyze and represent, and recommend up-to-date information. Collaborative filtering methods are based on collecting and analyzing a large amount of information on users' behaviour, activity or preferences and predicting what users will like based on their similarity to other users. One of the most common types of Collaborative Filtering is item-to-item collaborative filtering (people who buy x also buy y), an algorithm popularized by Amazon.com recommender system. User-based collaborative filtering attempts to model the social process of asking a friend for a recommendation. A particular type of collaborative filtering algorithms uses matrix factorization, a low-rank matrix approximation technique. A key advantage of the collaborative filtering approach is that it does not rely on machine analyzable content and therefore it is capable of accurately recommending complex items such as movies without requiring an "understanding" of the item itself[4].

Many collaborative filtering techniques have been developed. They can be categorized into two types:

- ***Memory Based Collaborative Filtering:***

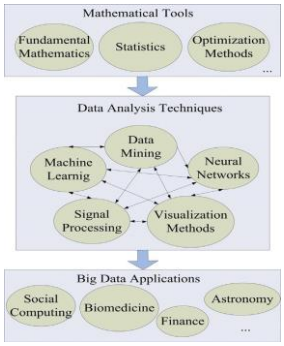
Memory-based CF uses user-to-user or item-to-item correlations based on users' rating behaviour to recommend or predict ratings for users on future items. Correlations can be measured by various distance metrics, such as Pearson correlation coefficient, cosine distance, and Euclidean distance. Memory-based collaborative filtering uses the whole training set each time it computes a prediction, which makes it easy to incorporate new data but suffers slow performance on large data sets. Speedup can be achieved by pre-calculating correlations and other needed information and incrementally updating them. For some applications, however, the size requirement makes the approach infeasible. It can perform with high recommendation accuracy, and new data can also be easily applied into recommendation. However, it is costly in computing and with bad scalability[4].

- ***Model Based Collaborative Filtering:***

Unlike memory-based CF, model-based approach does not use the whole data set to compute a prediction. Instead, it builds a model of the data based on a training set and uses that model to predict future ratings. For example, clustering based CF method builds a model of the data set as clusters of users, and then uses the ratings of users within the cluster to predict. A very successful model -based method is the Singular Value Decomposition (SVD), which represents the data by a set of vectors, one for each item and user, such that the dot product of the user vector and the movie vector is the best approximation for the training set. Typical the model building process is computationally expensive and memory intensive. After models are constructed, predictions can be done very fast with small memory requirement. Model-based CF methods usually achieve less accurate prediction than memory-based methods on dense data sets where a large fraction of user-item values are available in the training set, but perform better on sparse data sets[4].

IV. TECHNIQUES AND TOOLS IN BIG DATA

Big Data needs extraordinary techniques to efficiently process large volume of data within limited run times. Reasonably, Big Data techniques are driven by specified applications. Big Data techniques involve a number of disciplines, including statistics, data mining, machine learning, neural networks, social network analysis, signal processing, pattern recognition, optimization methods and visualization approaches. There are many specific techniques in these disciplines, and they overlap with each other hourly.



Big Data Techniques

- ***Big Data tools based on batch processing***

Apache Hadoop and map/reduce:

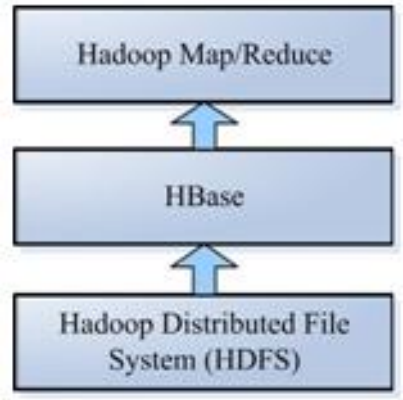
Apache Hadoop is one of the most well-established software platforms that support data-intensive distributed applica-tions. It implements the computational paradigm named Map/Reduce. Apache Hadoop platform consists of the Hadoop kernel, Map/Reduce and Hadoop distributed file system (HDFS), as well as a number of related projects, including Apache Hive, Apache HBase, and so on.

Map/Reduce , which is a programming model and an execution for processing and generating large volume of data sets, was pioneered by Google, and developed by Yahoo! and other web companies. Map/Reduce is based on the divide and conquer method, and works by recursively breaking down a complex problem into many sub-problems, until these sub-problems is scalable for solving directly. After that, these sub-problems are assigned to a cluster of working notes, and solved in separate and parallel ways. Finally, the solutions to the sub-problems are then combined to give a solution to the original problem. The divide and conquer method is implemented by two steps: Map step and Reduce step. In terms of Hadoop cluster, there are two kinds of nodes in Hadoop infrastructure. They are master nodes and worker nodes. The master node takes the input, divides it into smaller sub-problems, and distributes them to worker nodes in Map step. Afterwards, the master node collects the answers to all the sub-problems and combines them in some way to form the output in Reduce step.

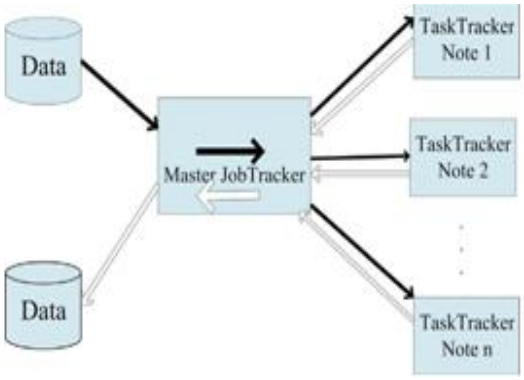
With the addition of Map/Reduce, Hadoop works as a powerful software framework for easily writing applications which process vast quantities of data in-parallel on large clusters (perhaps thousands of nodes) of commodity hard-ware in a reliable, fault-tolerant manner. We give a famous and prototypical example that counts the occurrence number of each word in a set of documents for Map/Reduce framework, where the two main functions Map () and Reduce () are given in the following. The Map steps are implemented on Hadoop cluster in a parallel way, and a large number of lists of intermediate data pairs with the form (key, c) are produced, where key represents a specified word, and the parameter c indicates the count of the word appearance. In Reduce steps, those lists of data pairs are integrated to the final results recursively in the main function.

There are a master (JobTracker) and a number of slaves (TaskTracker) in the Map/Reduce frame-work. The master node is in charge of job scheduling and task distribution for the slaves. The slaves implement the tasks exactly as assigned by the master. As long as the systems start to run, the master node keeps monitoring all the data nodes. If there is a data nodes failed to execute the related tasks, the master node will ask the data node or another data node to re-execute the failed tasks. In practice, applications specify the input files and output locations, and submit their Map and Reduce functions via interactions of client interfaces. These parameters are important to construct a job configuration. After that, the Hadoop job client submits the job and configuration to the JobTracker. Once JobTracker receive all the necessary information, it will distribute the software/configuration to the TaskTrckers, schedule tasks and monitor them, provide status and diagnostic information to the job-client. From the foregoing, we know that coordination plays a very important role in Hadoop, it ensures the performance of a Hadoop job.

Andrew Pavlo gave an overall discussion on properties of Map/Reduce framework, as well as other approaches to large-scale data analysis. Many data mining algorithms have been designed to accommodate Map/Reduce. For example, data cube materialization and mining, efficient skyline computation and scalable boosting methods.



Hadoop system architecture



Map/reduce overview: solid arrows are for Map flows, and feint arrows are for reduce flows

- ***Big Data tools based on Stream Processing***

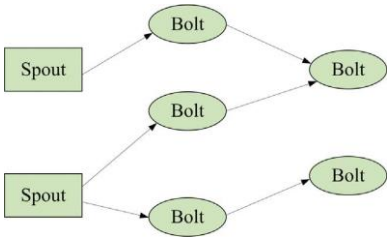
Storm:

Storm is a distributed and fault-tolerant real-time computation system for processing limitless streaming data. It is released as open source and free for remoulding. Storm is specifically designed for real-time processing, contrasts with Hadoop which is for batch processing. It is also very easy to set up and operate, and guarantees all the data will be processed. It is also scalable and fault-tolerant to provide competitive performances. Storm is efficient that a benchmark clocked it at over a million tuples processed per second per node. Therefore, it has many applications, such as real-time analytics, inter-active operation system, on-line machine learning, continuous computation, distributed RPC, and ETL.

A Storm cluster is ostensibly similar to a Hadoop cluster. Whereas on Storm users run different topologies for different Storm tasks. However, Hadoop platform implements Map/Reduce jobs for corresponding applications. There are a number
of differences between Map/Reduce jobs and topologies. The key one is that a Map/Reduce job eventually finishes, whereas a topology processes messages all the time, or until users terminate it.

To implement real-time computation on Storm, users need to create different topologies. A topology is a graph of computation and can be created and submitted in any programming language. There are two kinds of node in topologies, namely, spouts and bolts. A spout is one of the starting points in the graph, which denotes source of streams. A bolt processes input streams and outputs new streams. Each node in a topology contains processing logic, and links between nodes indicate how data should be processed between nodes. Therefore, a topology is a graph represent-ing the transformations of the stream, and each node in the topology executes in parallel.

A Storm cluster consists of two kinds of working nodes. They are only one master node and several worker nodes. The master node and worker nodes implement two kinds of daemons: Nimbus and Supervisor respectively. The two daemons have similar functions with according JobTracker and TaskTracker in Map/Reduce framework. Nimbus is in charge of distributing code across the Storm cluster, scheduling works assigning tasks to worker nodes, monitoring the whole system. If there is a failure in the cluster, the Nimbus will detect it and re-execute the corresponding task. The super-visor complies with tasks assigned by Nimbus, and starts or stops worker processes as necessary based on the instructions of Nimbus. The whole computational topology is partitioned and distributed to a number of worker processes, each worker process implements a part of the topology.



Strom topology

V. CONCLUSION

In this survey paper, I give an overview on Big Data problems, current techniques and technologies. A solution for solving the scalability and efficiency problem of recommendation system in big data environment, a cloud computing tool named hadoop is being used, which is a widely-adopted distributed computing platform using the MapReduce parallel processing paradigm.

VI. REFERENCES

[1] http://www.sap.com/solution/big-data/software/overview.html

[2]IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, TPDS-2013-12-1141 1 (KASR: A Keyword-Aware Service Recommendation Method on MapReduce for Big Data Applications Shunmei Meng, Wanchun Dou, Xuyun Zhang, Jinjun Chen, Senior Member, IEEE)

[3] Data-intensive applications, challenges, techniques and technologies: A survey on Big Data

C.L. Philip Chen, Chun-Yang Zhang

[4]A Survey on Recommender Systems based on Collaborative Filtering Technique

Atisha Sachan, Department of Computer Science And Engineering, LNCT, BHOPAL, MP, INDIA

Vineet Richariya, Head Of The Department Of Computer Science And Engineering, LNCT, BHOPAL, MP, INDIA

[5]http://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=8&cad=rja&uact=8&ved=0CGAQFjAH&url=http%3A%2F%2Fwww.gsma.com%2Fmembership%2Fwp-content%2Fuploads%2F2013%2F07%2FThe-Top-Challenges-of-Big-Data-Analytics.pdf&ei=ucfyU_bDFpC8uATR-4GoBg&usg=AFQjCNE2Pepw3wSXpwc0HWjJRfH40zDRmA&sig2=RXb_FtdVHNy3HYQyDJuOpA&bvm=bv.73231344,d.c2E

[6]http://www.developer.com/db/understand-the-high-availability-and-scalability-challenges-with-big-data.html

