E-COMMERCE ANALYSIS USING INCREMENTAL MAP REDUCE WITH MAP REDUCE RECOMMENDATION SYSTEM

Jhansi Rani Sivasubramanian,
Final Year, M.E. Computer Science and Engineering.
K. Ramakrishnan College of Technology, Trichy, India
subu.janu2010@gmail.com

Girija Kirupakaran
Assistant Professor (Department of CSE)
K. Ramakrishnan College of Technology, Trichy, India
girijajoy2000@gmail.com

***Abstract* - As computer systems create and collect growing amounts of data, analyze it becomes a basic part of improving the services provided by Internet companies. A vital property of the workloads method by Map Reduce applications is that they are often incremental by nature; i.e., Map Reduce jobs often run frequently with small changes in their input. In this paper, explain the architecture, implementation, and evaluation of a vital Map Reduce framework, named I^2 map reduce framework, for incremental computations. I^2 map reduce notice changes to the inputs and allow the automatic update of the outputs by employing an efficient, fine-grained result re-use mechanism. To attain efficiency without give up transparency, accept recent advances in the area of programming languages to identify methodically the shortcomings of task-level memorization approaches, and address them using numerous novel techniques such as a storage system to store the input of consecutive runs, a reduction phase that make the incremental computation of the reduce tasks more efficient, and a scheduling algorithm for Hadoop that is aware of the location of previously computed results.**

***Keywords*-- E-commerce, Map Reduce framework, Incremental processing, Job level task, Fine grained result.**

I. INTRODUCTION

A data is a collection of details from web servers usually of unstructured form in the digital universe. A large quantity of the data accessible in the internet is generated either by individuals, groups or by the organization over a meticulous period of time. The volume of data becomes bigger day by day as the procedure of World Wide Web makes an interdisciplinary part of human activities. Rise of these data leads to a novel technology such as big data that acts as a tool to method, control and direct very large dataset along with the storage space required. Big Data is large volume, large velocity and variety information assets that insist cost-effective, inventive forum of information processing for improved insight and decision making. Big data, a buzz word that can be handle peta bytes or terabytes of data in a reasonable amount of time. Big data is separate from large existing database which uses Hadoop framework for data intensive distributed applications. Big Data analytics apply higher analytical techniques of big datasets to find out hidden patterns and other useful information. It is performed using software tools mainly for predictive analysis and data mining. The growing number of technologies is used to aggregate, manipulate, manage and analyze big data. The basic flow is described in figure 1.

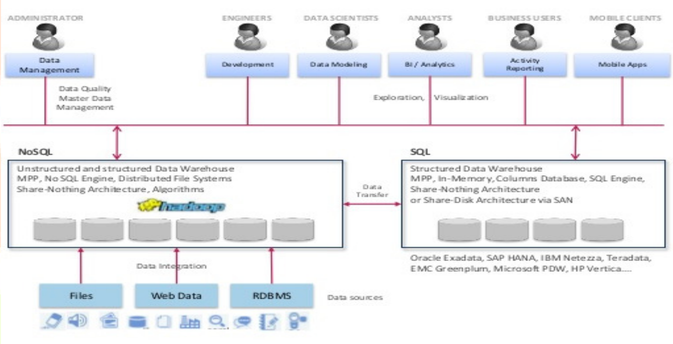
ADMINISTRATOR
ENGINEERS
DATA SCIENTISTS
ANALYSTS
BUSINESS USERS
MOBILE CLIENTS
Data Management
Development
Data Modeling
BI / Analytics
Activity Reporting
Mobile Apps
Data Quality
Master Data
Management
Exploration, Visualization
NoSQL
Unstructured and structured Data Warehouse
MPP, No SQL Engine, Distributed File Systems
Share-Nothing Architecture, Algorithms
hadoop
Data Transfer
SQL
Structured Data Warehouse
MPP, In-Memory, Columns Database, SQL Engine,
Share-Nothing Architecture
or Share-Disk Architecture via SAN
Oracle Exadata, SAP HANA, IBM Netezza, Teradata,
EMC Greenplum, Microsoft PDW, HP Vertica....
Data Integration
Files
Web Data
RDBMS
Data sources

Figure 1.1 Basic structure of E-commerce analytics

II RELATED WORK

P. Bhatotia et.al…, [4] present a system called Incoop, which permit existing MapReduce programs, not calculated for incremental processing, to execute visibly in an incremental manner. In Incoop, calculation can respond repeatedly and professionally to modifications to their input data by reusing middle results from previous runs, and incrementally inform the output according to the modify in the input.

Y. Bu, et.al…, [6] present Pregelix, a large-scale graph analytics system that began in 2011. Pregelix obtain a novel set-oriented, iterative dataflow approach to apply the user level Pregel programming model. It achieves so by treating the messages and vertex states in a Pregel calculation like tuples with a well-defined schema; it then employ database-style query evaluation techniques to execute the user's program.

B. Howe, et.al…, [7] provides HaLoop, a customized version of the Hadoop MapReduce framework that is planned to serve these applications. HaLoop not only extends MapReduce with programming support for iterative applications, it also considerably improves their efficiency by making the task scheduler loop-aware and by adding various caching mechanisms.

J. Ekanayake, et.al…, [10] implements Twister framework which is an improved MapReduce runtime with an extensive programming model that supports iterative MapReduce computations efficiently. It uses a publish/subscribe messaging infrastructure for communication and data transfers, and supports long running map/reduce tasks, which can be used in "configure once and use many times" approach.

S. Ewen, et.al…, [1] propose a method to mix incremental iterations, a form of work-set iterations, through parallel data flows. After presentation how to mix bulk iterations into a dataflow system and its optimizer, current an extension to the programming model for incremental iterations.

III. E-COMMERCE ANALYZER USING DATA MINING TECHNIQUES

E-commerce, generally referred to as blogs, are considered as online diaries published and maintained by individual users (bloggers), bloggers' daily activities reports. In existing system, analyze the web logs using k means clustering. Clustering is an unsupervised classification and widely used for mining web usages with main objective to grouping a known collection of unlabeled objects into evocative clusters. In web domain, clusters may be web documents and web links. K- means clustering algorithm as follows:

Step 1: Choose K cluster centers randomly from n points
Step 2: Assign each point to clusters
Step 3: Compute new cluster centers
Step4: If termination criteria satisfied, stop otherwise continues from step 2.

Then cluster the web logs using page rank algorithm which play a major role in making the user search Navigation easier in the results of a search engine, which helps in best utilization web resources by providing required information to the Navigator. This algorithm follows the following steps:

Step 1: As visitor connects to the internet, web analyzer counts the visitors.
Step 2: E-commerce web analyzer does the task of authentication function with the help of previously recorded information.
Step 3: E-commerce web analyzer identifies and records the behavior of the customer if visitor visits the site first time and if the customer is an old, it will find the previous records of the customer and will present the site in accordance with the likening of the customer.
Step 4: E-commerce web analyzer stores time stamp of the customer visit.
Step 5: E-commerce web analyzer observes the total online expenditure divided by the number of clicking.
Step 6: E-commerce web analyzer keeps record of the request status.
Step 7: The same analyzer analyzes marketing investment.
Step 8: Web analyzer reduces the size of log files
Step 9: Updates the visitor record.
Step 10: End

IV. E-COMMERCE ANALYZING USING I^2 MAP REDUCE APPROACH

Website personalization is the process of customizing the content and structure of a website for specifically needs. Steps of personalization as

a) The collection of web data
b) Modeling and categorization of these data.
c) Analysis the collected data
d)Determination of the actions that should be performed.

We can analyze web logs using i^2 map reduce approach. To run an incremental iterative step Ai, i^{2}MapReduce care for each iteration as an incremental one step job and viewed in fig 2. In the first iterative, delta input is produce delta structure data. The preserved MRBGraph reproduce the last iteration in job Ai−1. Only the Map and Reduce example that are precious by the delta input are re-computed. The output of the major Reduce is the delta state data. Apart from the computation, i^{2}MapReduce revive the MRBGraph with the newly calculate intermediate states. Then denote the state as updated MRBGraph. In the j-th iteration, the structure data ruins the same as in the (j − 1)-th iteration, but the loop-variant state data has been updated. Using the preserved MRBGraph j−1, i^{2}MapReduce recomputes only the Map and Reduce instances that are affected by the input change.

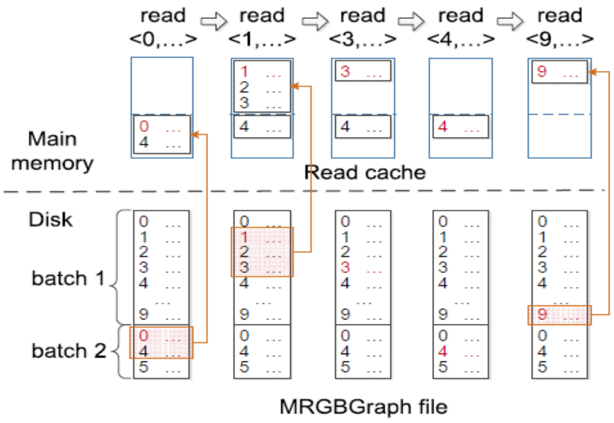

Figure 1.2 Job sequences

Servers store following information for every request. IP address, Date/time stamp, Status of request, Referring URL, Status of request, Type of user agent used software manufacturer and version no, Type of operation system, Network location and IP address: can include country, city or any other geographic data as well as the host name, Time of visit, Page visited, Time spent on each page of the website, Referring site statistics: can include the website you can through to reach this website and search engine query that brought there. The framework of the work is plotted in fig 3.

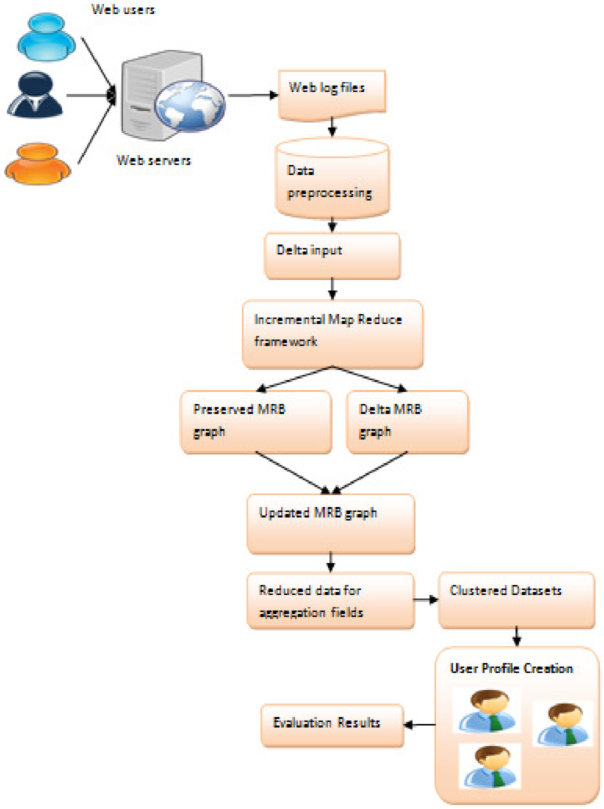

Figure 1.3 System Architecture

SYSTEM DESCRIPTION

The various components used here are: (A) Preprocessing, (B) Incremental Map reduce, (C) Mined Data, (D) Performance Evolution.

A. PREPROCESSING

Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant, etc. parts of the data and then replacing, modifying, or deleting this dirty data or coarse data. After cleansing, a data set will be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleansing differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at entry time, rather than on batches of data. Pattern Matching is used to separate the fields, and only log line of successful status count is used for further processing. The processed line is checked to find out 'GET' method and records with processing data in the requested URL are identified and removed. The preprocessed log value includes the all the field and sent for a request.

(B) INCREMENTAL MAP REDUCE:

The original input interface of Hadoop Map Reduce requires users to specify the input files. In incremental map reduce, the concept of job is extended. There are four kinds of job submission, one-time job, initial job, incremental job, and continuous incremental job.

- **One-time run:** To be compatible with the conventional use case, incremental map reduce provides almost the same client submission APIs. No state needs to be saved when the job is one-time run.
- **Initial run:** The job is executed just like one-time run. However, the execute state of the job needs to be saved. If users do not request to terminate the whole job, system will keep the state.
- **Incremental run:** Based on initial run, incremental run requires system to fetch the state of the prior runs and combine it with the new added data. At the same time, the state will be updated according to the current run.
- **Continuous run:** For some data analysis applications, new inputs are continuously added to the system, so the system needs to discover these new data automatically and submits jobs automatically.

(C) MINED DATA

In this module, implement pattern matching approach to construct the profile based on user name and password. User id and user name are used to construct the profile. The user profile is incrementally developed over time and it is stored for use in later sessions. The information exploited for constructing the profile usually comes from various sources, so it relies on different aspects of the user. On the other hand, in ephemeral preferences, the information used to construct each user profile is only gathered during the current session, and it is immediately exploited for executing some adaptive process aimed at personalizing the current interaction. In other words, since each user profile is computed based on term weights in a Web page the user browsed and the browsed pages are different according to each user, the profile and keyword are constructed in the form of a user-terms. This approach allows us to construct a more appropriate user profile and keyword that perform a fine-grained search that is better adapted to each user's preferences.

(D) PERFORMANCE EVALUATION

In this module, evaluate the performance of the system using time and accuracy metrics. The proposed approach presents an incremental data processing model which is compatible with the Map Reduce model and its runtime. It supports Map Reduce-based applications without any modification. A part of this various things may be of help to the system administrator like, analysis of errors helps to know the problems while accessing the website, analysis of references to website during special event will help administrator to know and balance load, analysis of navigational patterns and duration will help the administrator with the knowledge about how to decrease the duration of the user by providing layout change, decrease in duration helps to usage of less bandwidth.

VI. CONCLUSION

In this paper, proposed a technique to observe the users and collect the information of users on the website and then provide the semantic data to the request of visitors. It will work better than the cookies and beacons etc. There is no need to enter user name and password every time. E-commerce analyzer will remember password and user name and also personalization, information, location memory and site understanding. The E-commerce analyzer optimize logs just track what you need to know about visitors without complex filtering. It will also reduce the size of log file. E-commerce analyzer will be used to collect the information across different domains and websites. The percentage of the total number of visitors who make a purchase on the site Log analyzer will enable them to better understand and respond to the interests of visitors to their sites. E-commerce analyzer will allow e-commerce sites to recognize visitor's generated form online and email advertising campaign.

REFERENCES

[1] S. Ewen, K. Tzoumas, M. Kaufmann, and V. Markl. Spinning fast iterative data flows. PVLDB, 5(11):1268–1279, 2012.

[2] T. J¨org, R. Parvizi, H. Yong, and S. Dessloch. Incremental recomputations in mapreduce. In Proc. Of CloudDB '11, 2011.

[3] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In Proc. of VLDB '94, pages 487–499, 1994.

[4] P. Bhatotia, A. Wieder, R. Rodrigues, U. A. Acar, and R. Pasquin. Incoop: Mapreduce for incremental computations. In Proc. of SOCC '11, 2011.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst., 30(1-7):107–117, Apr. 1998.

[6] Y. Bu, V. Borkar, J. Jia, M. J. Carey, and T. Condie. Pregelix: Big(ger) graph analytics on a dataflow engine. PVLDB, 8(2):161–172, 2015.

[7] Y. Bu, B. Howe, M. Balazinska, and M. D. Ernst. Haloop: efficient iterative data processing on large clusters. PVLDB, 3(1-2):285–296, 2010.

[8] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. Technical Report 1999-22, Stanford InfoLab, 1999.

[9] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. In Proc. of OSDI '04, 2004.

[10] J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S.-H. Bae, J. Qiu, and G. Fox. Twister: a runtime for iterative mapreduce. In Proc. of MAPREDUCE '10, 2010.