**Big Data Mining in Commerce**

Harrison Huynh

Research Student  
Atlanta, Ga 30302  
Georgia State University  
+1 470.599.1126

hhuynh12@student.gsu.edu

Jarrod Vega

Research Student  
Atlanta, Ga 30302  
Georgia State University  
+1 404.201.3107

jvega7@student.gsu.edul

**ABSTRACT**

Big data is a fairly new idea and is evolving at a rapid pace to fit the needs of an ever increasing enterprise. Harnessing the power and effectiveness of Big Data is challenging and can be very beneficial to large retail companies looking to maximize customer buying practices. A retail website can take in the large data of previously purchased items from a user’s account and suggest what to buy that is similar to the previously bought item, combined with what others bought when they purchased something similar.. But to suggest a full array of items matching to the product is one of the most challenging problems. In this paper, Big data mining requires a lot of data, therefore data is mined from a smaller source of purchases history from a database to further understand the complexity of data mining algorithms. We will be finding the correlations between purchased products and suggest other products that pertain to the correlated items. We will look at which item attribute is the best to compare to other product, in order to find the best correlation. In return we will improve the data mining capabilities to suggest for users their next item to purchase.

**CCS Concepts**

**• Applied computing ➝ Electronic commerce**

**• Human -centered computing ➝Human computer Interaction**

**• Information systems ➝ Information retrieval**

**Keywords**

BigData,SequentialPatternMining, multi-petabyte,Apriori,FP-Growth Algorithm

# INTRODUCTION

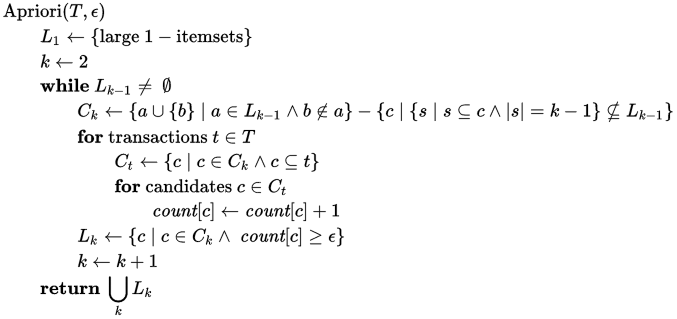
Due to an alarming increase in the Big Data movement, there is a huge demand for mining data that companies have collected over the past decade. Companies have more data that they can imagine, without any idea in how to mine and find the relationship in order to utilize it to their advantage.

There are two major problems with Big Data. First the storage and management of a very large volume of information that companies have collected. Commercial systems have been report to scale very well, being capable of handling “multi-petabyte” databases, but in addition to their "cost" in terms of price and hardware resources, they do not have the capability of importing data into a “native representation.”

Secondly, there’s is not a good process for implementing the exploration of the large volume of data, which in turn is difficult to suggest products for user in a timely manner. In this paper we are going to explore the different algorithms and technologies to implement the best algorithms for finding the best relational correlation between products.

One of the biggest challenge in consumer targeting is to provide the users the right message at the right time. In order to provide the right information at the right time is to collect the right data, but not only that, the quality of data is really important. It is important to measure the importance of data. Their context is vital to understanding the strengths and weaknesses of the data. It is easy to see patterns within data collections, but we need to know if the patterns are real. In this paper we will explore what kind of data we need to collect and determine their importance relating to the consumers need.

# Key Algorithms/Technologies

**Figure 1**

“Apriori[[1]](https://en.wikipedia.org/wiki/Apriori_algorithm#cite_note-apriori-1) is an algorithm for frequent item set mining and [association rule learning](https://en.wikipedia.org/wiki/Association_rule_learning) over transactional [databases](https://en.wikipedia.org/wiki/Databases). It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine [association rules](https://en.wikipedia.org/wiki/Association_rules) which highlight general trends in the [database](https://en.wikipedia.org/wiki/Database): this has applications in domains such as [market basket analysis](https://en.wikipedia.org/wiki/Market_basket_analysis).”[1]

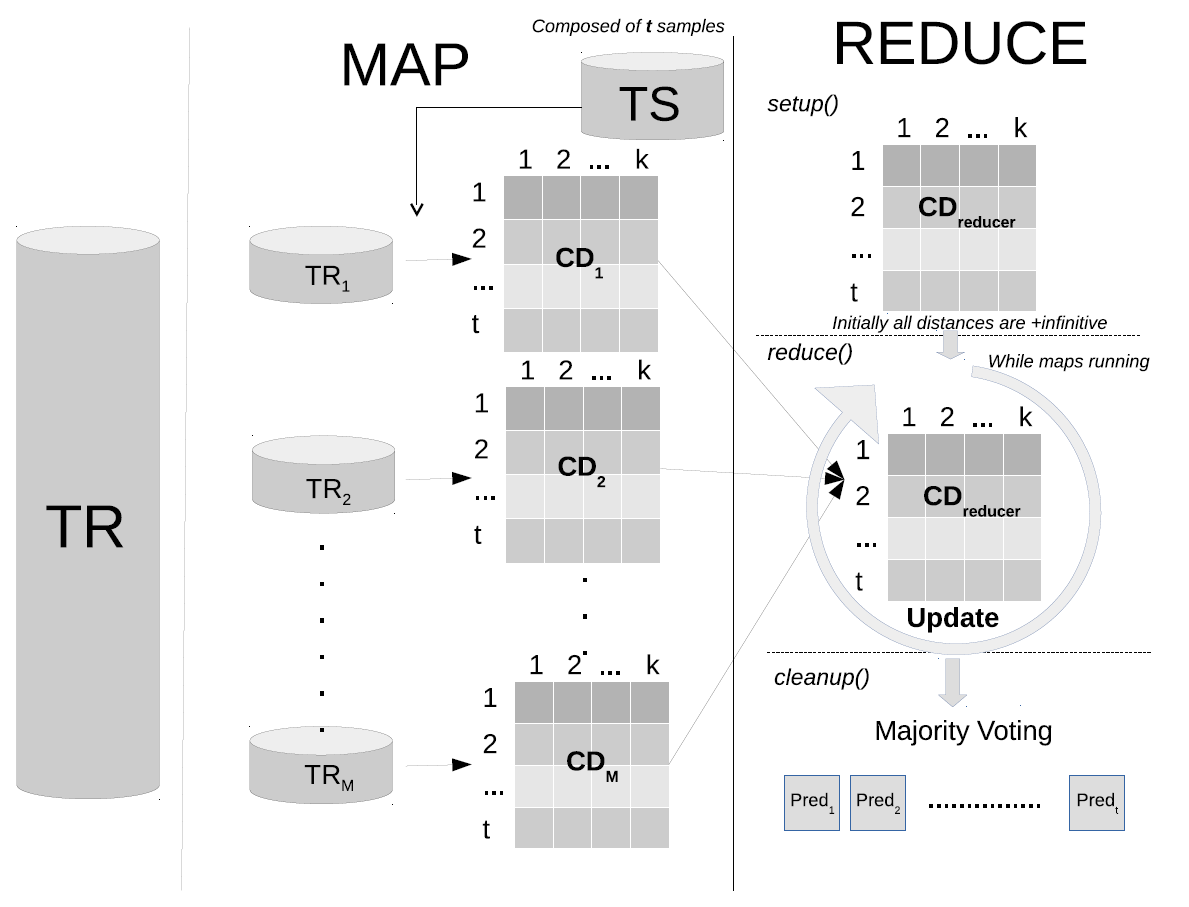
### FP-Growth Algorithm

For FP-Growth Algorithm first need to construct an FP-tree with the data. To do this we must first have a database and a minimum support point. To construct the tree you must scan the database. Collect set frequent item. Sort the info into descending order of frequent items. Root of the FP- tree needs to be null.Select the frequent items sort according to the order of F List. Let the sorted frequent-item list in Trans be [ p | P], where p is the first element and P is the remaining list. Call insert tree([ p | P], T ). Call FP- growth(Tree,a){Tree contains a single prefix-path part of Tree; Let P be the single prefix-path part of Tree; Let Q be the multipath part with the top branching node replaced by a null root; for each combination denoted as B of the nodes in the path P do; generate pattern B with a support = minimum support of nodes in B; let frequent pattern set P be the set of patterns so generated;} else let Qq be the Tree; for each item in Q do {generate pattern B = ai U a with support = ai; construct B’s conditional pattern-base and then B’s conditional FP-tree B; if Tree B does not equal null then call FP-growth(Tree B, B); let freq pattern set(Q) be the set of pattern so generated; }return(freq pattern set(P) ∪ freq pattern set(Q) ∪ (freq pattern set(P) × freq pattern set(Q))).

### 

The frequent-pattern tree is a compact structure that stores quantitative information about frequent patterns in a database.

**Figure 2**



The Big Data movement is indistinctly related to the open source software revolution. Big companies such as Facebook, Yahoo!, Twitter, LinkedIn can benefit from using open source software to develop. One of the big open source software that companies mentioned from above is Apache Hadoop.[4]

Hadoop allows for the coding of applications that is required to rapidly process a large amount of data to be in parallel of large clusters of compute nodes.[4]

# Data Source

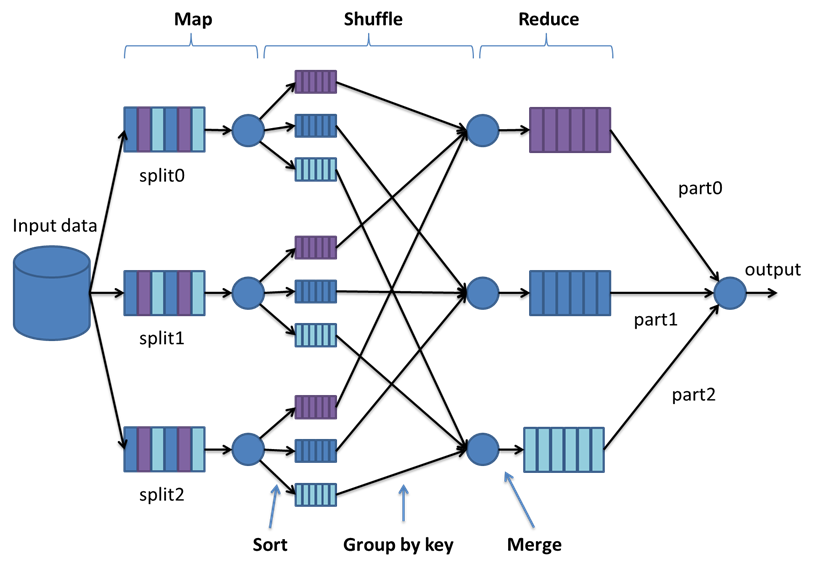
Our data will be mined from a smaller database that will represent a large scale database similar to of Amazon. The reason we are doing this is because we do not have the capability to host such a large amount of data, therefore we are scaling down the data to better implement our algorithm.

# Related Work

Macy’s is a well known mid -ranged to upscale department store in America. Right now they are using and implicating big data into their daily routine in order to get a better personalized shopping experience for their customer. [5]

One important thing to keep in mind is the design of the data model. We need to be able to design a data model that will allow for scalability. To understand what is needed to make a scalable data model we need to first understand the the strength and weaknesses of the data we collect. We need to also take into consideration of the cost and benefits of the collection of data. The only way to really determine which model is best for the situation is by trial and error. [7]

**Figure 2**

****

**The best known example of Big Data execution environment is probably Google MapReduce[3]**

# What’s Expected to be Turned In

At the end of our project, we are expected to turn in an algorithm that will mine data from a created database. With the algorithm we are expected to be provided with information that we could utilize to suggest more products of the same nature/correlation to each other. We aim to provide an algorithm that is precise and reliable so that we could provide a response in a timely manner. Along with the algorithm we will be giving a detailed report of how making a proper data model can greatly improve the scalability and relevance of the data.

# REFERENCES

1. Apriori algorithm. (n.d.). Retrieved September 22, 2016, from https://en.wikipedia.org/wiki/Apriori\_algorithm
2. By Doing So, a Complete Library of Methods Would Be Available for Researchers and Experts to Take the Full Advantage of This Novel Work Area. Additionally, This Can Be Used as Basis for the Development of Novel and More Powerful Fuzzy Algorithms. However, the Transition from Standard Learning Approaches towards the MapReduce Parallel Model Is Not Straightforward, and Requires a Thorough Analysis of the Designer. "Soft Computing and Intelligent Information Systems." *Big Data: Algorithms for Data Preprocessing, Computational Intelligence, and Imbalanced Classes*. N.p., n.d. Web. 22 Sept. 2016.
3. By Using This Algorithm, the FP-tree Is Constructed in Two Scans of the Database. The First Scan Collects and Sort the Set of Frequent Items, and the Second Constructs the FP-Tree. "Data Mining Algorithms In R/Frequent Pattern Mining/The FP-Growth Algorithm." *- Wikibooks, Open Books for an Open World*. N.p., n.d. Web. 22 Sept. 2016.
4. Fan, Wei, and Albert Bifet. "Mining Big Data." *SIGKDD Explor. Newsl. ACM SIGKDD Explorations Newsletter* 14.2 (2013): 1. Web.
5. Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.
6. Kleissner, Charly. "Data mining for the enterprise." *System Sciences, 1998., Proceedings of the Thirty-First Hawaii International Conference on*. Vol. 7. IEEE, 1998.
7. Kohavi, Ron. "Mining e-commerce data: the good, the bad, and the ugly."*Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001.
8. Macy's Is Changing The Shopping Experience With Big Data Analytics. (n.d.). Retrieved September 22, 2016, from <https://datafloq.com/read/macys-changing-shopping-experience-big-data-analyt/286>
9. Masseglia, Florent, Maguelonne Teisseire, and Pascal Poncelet. "Sequential Pattern Mining." *Encyclopedia of Data Warehousing and Mining* (2005): 1028-1032.
10. Sarwar, Badrul, et al. "Analysis of recommendation algorithms for e-commerce." *Proceedings of the 2nd ACM conference on Electronic commerce*. ACM, 2000.
11. Rajpurohit., B. A. (n.d.). KDnuggets. Retrieved September 22,2016,from<http://www.kdnuggets.com/2014/12/interview-daqing-zhao-macys-advanced-analytics-marketing.html>