Research on Key Technologies of E-commerce Big Data Analysis Platform

Allam Maalla*, Qiwei Jia
Research Institute of Open Data Commercial Application, School of Information Technology and Engineering, Guangzhou College of Commerce, Guangzhou, China, 511363
Wah Yuet Hardware (Shenzhen) CO., LTD, Shenzhen, China,518118
allamamaalla@yahoo.com

Abstract. This paper designs and implements the E-commerce big data analysis platform, which is mainly based on the data of commodities and sales of E-commerce sales platforms, providing a platform for intelligent analysis. The platform provides an interface for users to operate. It can efficiently obtain the necessary business information for e-commerce users and provide decision support. The system can be divided into four parts: crawler system, storage system, offline data analysis system and user interaction system. The storage system adopts the Hadoop ecosystem. Hadoop's HDFS are highly fault-tolerant, suitable for processing GB, TB, or even PB-level data, and can be expanded horizontally, and deployed on multiple inexpensive machines, the platform is Hadoop-based. HBase and Hive will handle and use data more efficiently. The offline analysis system adopts the Spark framework. That Spark is based on RDD (Elastic Distributed) set data processing. It can connect various data sources of HBase and Hive. Because the data are loaded into the memory of the cluster host, and quickly iterated, it is suitable for multiple rounds of computing tasks such as machine learning. Spark-SQL can provide SQL-like operations on structured or semi-structured data, which can greatly improve the efficiency of off-line data analysis.

1.Introduction
In the area of big data, data are produced in large quantities every day, and data are more and more complex. The scale of data in various fields is growing with a rapid attitude, and the whole society has entered the era of big data. At the same time, the economic status of e-commerce in the Internet has also increased because of big data, and vigorously developing e-commerce has become the primary task of countries and regions.

At present, the domestic e-commerce enterprises overall on a smaller scale. Its development areas are mainly concentrated in the aspect of market development, with the lack of the cognition to the big data. Although the development of e-commerce has an indispensable relationship with big data, many e-commerce enterprises lack of big data analysis and forecasting ability that causes the enterprise marketing strategy is difficult to keep up with the development of big data, therefore its validity is restricted. There are traditional electricity data analysis methods: heat network marketing, customer value analysis, product analysis, etc., but because the required data on the dimensions and quantity demand is higher, and many small shops lack of technology, capital, talent and many other aspects of the amount of data, it is more difficult on the electric data analysis work. Therefore, building the e-commerce big data analysis platform, it can help many small e-commerce enterprises, enable them...
to analyze customers' demands through data, and make predictions for future business in advance, so as to reduce costs and increase profits. It can obtain better analysis results with the input of low cost and play a positive role in its store development and marketing decisions.

2. Functional Requirements

2.1. Company Analysis
1) Natural exposure index of store PC: calculate the index of the display volume in the search results page based on the keyword index and keyword ranking (use keywords to search on the e-commerce platform of PC terminal, and the ranking of the store listed by the integrated keyword heat itself as the store exposure index).

2) 30-day sales index: the sales volume of commodities in the past 30 days is estimated according to the daily evaluation number through the model (taking the store as the dimension, establishing the model through the relationship between the comment data of other e-commerce platforms such as Jindong, and predicting the sales status based on the comment data of Taboo).

3) Natural exposure index of store APP: calculate the index of the display volume in the APP search results page based on the keyword index and keyword ranking (search on the e-commerce platform of APP terminal with keywords, and the ranking of the store with comprehensive keyword heat as the exposure index).

2.2. Commodity Analysis
1) natural exposure index of commodity PC, the index of the display volume in the search results page is calculated according to the keyword index and keyword ranking (search by keywords on the e-commerce platform of PC terminal, and the ranking of the product is the heat of the keyword itself as the exposure index of the store).

2) The 30-day sales index of goods: the sales volume of goods in the past 30 days is estimated according to the daily evaluation number through the model (taking the store as the dimension, establishing the model through the relationship between the comment data of other e-commerce platforms such as Jindong, and predicting its sales status based on the comment data of Taboo).

3) Natural exposure index of commodity APP, Calculate the index of the display volume in the APP search results page based on the keyword index and keyword ranking (search on the APP e-commerce platform with keywords, and the ranking of the goods with the comprehensive keyword heat as the exposure index of the store).

2.3. Brand Value Analysis
1) natural exposure index of brand PC: the index of the display volume in the search results page is calculated according to the keyword index and keyword ranking (search by keywords on the e-commerce platform of PC terminal, and the ranking of the product's brand, the heat of the keyword itself, is used as the exposure index of the store).

2) 30-day sales index of brand corresponding products: according to the daily evaluation number, the sales volume of products corresponding to the brand in the past 30 days can be estimated through the model (taking the store as the dimension, the model can be established through the relationship between the comment data of other e-commerce platforms such as dingdong, and the sales status can be predicted based on the comment data of Taobao).

3). Natural exposure index of brand goods APP: calculate the index of the display volume in the APP search results page based on keyword index and keyword ranking (search by keywords on the e-commerce platform of APP terminal, the ranking of the products corresponding to the brand, the heat of the keyword itself, as the exposure index of the store).

2.4. Keyword Analysis
1). Natural search goods to make the dimension, analyzing what keywords will have the goods,
searching goods in particular keyword how to compete with other similar products, natural search exposure index, position in the search results, natural search exposure index refers to a product under a specific keywords exposure index is.

2). Natural search store display, dimensions by key words, analyzing what keywords will be in the shop, competed with other similar products from goods in a specific keyword, natural search exposure index, position in the search results, natural search exposure index refers to a product under a specific keywords exposure index is.

3). Keyword market analysis, the keyword has the index of search popularity, but the value of a keyword should also be observed. For example, the popularity of a keyword is very high, but the number of product sellers is also high. So analyzing the market value of a keyword can compare the seller situation and the buyer situation (keyword heat).

3. Data Quality Requirements

3.1. Collection Efficiency
1). There are about 100,000 keywords in e-commerce, and the average of each keyword is 10 pages. That is to say, the search page is requested about 50 million times. The detailed information page 60 x 50 million = 300 million times per page. Consider multi-threading and distributed to increase the crawling speed.

3.2. Data Increment
1). The sales volume, comment number, comment content, price, etc. They are changing in real time; this increment should be expressed. Judging whether a goods id crawl, if it is, it's been crawl, in a short time or long ago been crawl, if haven't climb over a period time will be updated, the specific way is to increment the key and add a time attribute, Every time it updates, it inserts a data, export the real-time data is exported last a time.

3.3. Anti-crawler Problem
1) Speed limit: When the speed of collection is too fast, some usual requests for unlimited web pages will take IP measures to prevent crawlers, and for our business creeper, the product details page is a page that we collect very frequently. Therefore, our high-speed visit is likely to be sealed by the IP address of the e-commerce website. So our strategy is to crawl for 30 minutes to 40 minutes in a random time, Crawler threads sleep between 10 and 30 seconds in the middle of a second.

2) User agent restriction: access to the same user agent can also be identified as a crawler by the site, so we need to collect a large number of user agents, whenever a crawler requests, we need to go through a download middleware, adding a forged user agent to requests.

3) Cookie Disablement: Websites can identify the same link based on speed, user agent and cookie. If a crawler carries a cookie on every request, the e-commerce website will determine that the same machine is requesting a large number of websites, so it will identify your request as an exception request, and we have at least one The two solution. The first is to collect a large number of cookie, and the request is to bring random cookie. The second is to disable cookies, which are empty in the requested requests, and to take a second approach in our research.

4. Data Removal
In incremental crawling requirements, we have to judge whether Id has crawled. So you can store a hash structure in Redis, key can save ID, and value saves the last update date. Also need a storage structure stored with id's data structure can also use Redis hash structure, key save id, value deposit in same id (sacrificing memory speed), need to remove the same, when id to come in, whether the query id in the key, proving the goods haven’t crawled if it is not in, if it is in it, check the value corresponding to key, that is, the last update time. If it is less than a certain time, no action is taken; If it is greater than a certain time, it will update its commodity price, comment number, praise number,
middle score, bad comment number and additional comment number. Find and match the item id with the item id, and the crawler requests a price with all the same item id with the comment information returned from the first request.

Figure 1. Reptilian weight removal process

5. Data Storage
The feature of the data is that the search page data is growing straightly. The data of detailed data page and commodity comment growing slowly, but the data volume is large. In addition, it need to store the cover of each item with the cover of shop. The speed of data access is also high, so the traditional relational database cannot meet this requirement. Compared to the popular open source non-relational databases of recent years, Redis, Mongodb, Hive, Hbase, it found out that Redis is a memory-based key-value database. Its access speed is very fast so that suits for cache system. Mongodb is an excellent open source non-relational database. It uses scenarios event record, content management or blog platform, etc. However, the drawback is that reading performance will decrease on massive data. Hive is suitable for data warehousing, for some real-time data analysis which it is not high. Hbase is suitable for large data volumes (100s TB of data) and requires fast random access without requiring many features in relational databases (such as cross columns, cross tables, transactions, joins, etc.). Combined with the above, HBase is suitable for the storage layer of crawler, but HBase is not conducive to conditional query. Therefore, we added a Mongo-db before HBase to store the crawled data, and wrote program in python to convert the data into HBase in a neat format.

6. Hive Data Warehouse Storage and System Design
The storage system uses Hadoop which based HBase database to store real-time data for online analysis, another database is the hive database based on Hadoop, it stores the data collected by the acquisition system into HBase, it clean and filter the data through custom ETL rules regularly for subsequent offline data analysis.

The reason for choosing HBase is that it is distributed, column-based and column extensible database, it is also a highly reliable and high-performance database. There is no need to worry about the data having one more field in the stored procedure to rebuild the table suddenly, and the real-time query efficiency of HBase database is very high, and the principle of column family query is used internally. The database contains three data acquisition tables, search page “search”, data from e-commerce search engine through the keyword search of the product data; detail page “detail”, data from each product corresponding to the detailed data; comment page “comment”, data source is the latest 5 pages of each product corresponding to the comment data. System design as fellow:
1) Crawler data flow description: the first is the collection of keywords, by searching the Internet (part of the crawl), collected to 200,000 keywords, part of the figure 1.
2) Put the keywords in the Redis database: and use the keywords to search the page crawler, get the product ID into the Redis database. Other data such as trade name, shop name, price and so on are stored in HBase. Another program retrieved the ID from Redis and crawled the details page data. The specific process is shown in figure 2.

![System design Crawler Data and Partial keyword diagram](image)

3) Web test: Through data search page, Input: price range, sales volume, number of search areas, date selection. Output: commodity ID, commodity name, price, sales volume, stock. Storage mode: HBase, Redis, Mongodb. Capture frequency: once a week.

![Web test and analysis results](image)

7. Conclusion
E-commerce big data analysis platform, mainly using crawler monitor data situation, framework for data collection, has the advantage of its built-in threading and scheduling mechanism and engineering. The storage system uses the Hadoop ecosystem, which is suitable for processing GB, TB or even PB data. It can be scaled horizontally and deployed on multiple inexpensive machines. This platform is based on Hadoop and hive that will process and use data more efficiently. It is loaded into the memory of the cluster host, so it can quickly iterate data, which will be suitable for multi-round computing tasks such as machine learning, and can greatly improve the efficiency of data offline analysis, and provides external operation interfaces such as java, python, and R.

Acknowledgments
This research was financially supported by the 2017 Higher Education Teaching Research and Reform Project of Guangdong Province, "The Big Data Talent Cultivation Practice Teaching System of Business Schools" (Project No. 2017SJXJGG01), and Characteristic Key Subject of E-commerce Construction Project of Guangzhou College of Commerce Foundation, (Project No.TSZDXK201601).

References
[1] Chien, C.F. and L.F. Chen, "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry". Expert Systems and Applications, 2015.
34(1): p. 380-290.

[2] Mur-Artal R, Montiel J M M, Tardos J D. ORB-SLAM: a versatile and accurate monocular SLAM system [J]. IEEE Transactions on Robotics, 2015, 31(5): 1147-1163.

[3] Maalla A, Zhou C, Guang-Yu W U. The Construction of Heterogeneous Platform of Unified Service System Based on Cloud Computing [J]. 2017(cmee). DOI: 10.12783/dtce/cmee2017/20056.

[4] Moy, J., 2004. Open Shortest Path First Version 2. RFQ 1583, Internet Engineering Task Force. http://www.ietf.org.

[5] Tambe M, Zhang W (2013) Towards flexible teamwork in persistent teams: extended report. Journal of Autonomous Agents and Multi-agent Systems, special issue on Best of ICMAS 98(3): 159-183.

[6] Maalla A (2016) Construction of Information Ecosystem on Enterprise Information Portal (EIP). J Electr Electron Syst 5: 205. doi:10.4172/2332-0796.1000205.