Efficient processing and utilizing Big Data in E-commerce

Xiaohui Pan

Modern Education Technology Center, Shanghai University of Political Science and Law, Shanghai, China 201701

Email: panxiaohui@shupl.edu.cn

Keywords. Big data; E-commerce; Processing;Utilizing

Abstract. Today Big Data is a business imperative and is providing solutions to long-standing business for companies of various domains around the world. Despite of its potential benefits, the efficiency and effectiveness to process and utilize Big Data are still challenging. In this paper, we summarize existing technologies and propose new technologies. Specifically, we categorize Big Data processing technologies and propose a new platform architecture for E-commerce Big Data processing. We also discuss the applications to utilize Big Data and propose a six-layer architecture for utilizing Big Data.

Introduction

The term “Big Data” is one of the hottest words today. So what is Big Data? Wikipedia defines it as “a massive volume of both structured and unstructured data that is so large that it’s difficult to process using traditional database and software techniques.” An IDC report on the study of the digital universe shows that from 2005 to 2020, data will grow by a factor of 300, from 130 exabytes to 40,000 exabytes, or 40 trillion gigabytes. In other words, the digital universe will double in about every two years. These data have been generated in our daily lives from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few.”

Big data can help enterprises improve E-commerce operations in several key areas. For example, structured data such as past orders can inform marketing decisions and product recommendations. Unstructured data such as social media interactions and product reviews may influence the items merchants stock. Not only do enterprises need to do a better job of collecting accurate data, they must find ways to convert it into actionable insight. Whether merchants are big or small, information will be a crucial part of successful E-commerce operations, so they need to learn how to use Big Data.

Challenges in exploiting Big Data

Big data can be categorized into structured and unstructured parts which need different technology to handle. The “structured” portion of Big Data refers to the regular data such as name, address, preferences, sex, age and so on that can be stored within a database. The “unstructured” part encompasses email, video, tweets, and Facebook Likes. None of the unstructured data resides in a fixed database that’s accessible to merchants. But the feedback from, say, social media has become a very useful research tool for businesses.

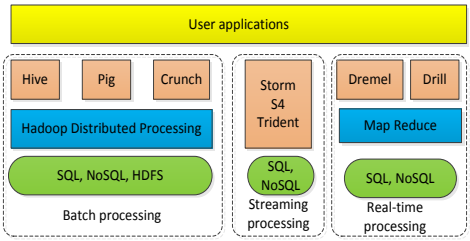

Fig.1 Big data processing technologies

Enterprises have access to a tremendous amount of information about their customers. However, data is useless if it's not leveraged properly. This is where many enterprises fail – they have access to all this information, or Big Data, but they struggle to process it in a meaningful way. Either that, or they simply get overloaded with all the information. In fact, one report from the Edgell Knowledge Network found that while 80 percent of retailers have heard of Big Data, only 47 percent say they can actually use it effectively to fuel their E-commerce operations. More efficient and convenient technologies are needed to process and utilizing Big Data effectively in business.

Processing Big Data

The Big Data processing technologies can be grouped into three categories as shown in Fig.1: batch processing, streaming processing and real-time processing. Batch processing can handle large amount of static data in parallel. However, batch processing often involves significant latency. In many circumstances, for instance, detection of credit-card fraud, algorithmic stock-trading, screening spam emails, and business activity monitoring, data must be processed at real time. These activities are termed complex event processing/event stream processing (CEP/ESP). This section will present a complete technology landscape so enterprises will be able to pick the appropriate technologies as their data solutions.

A. Categorization of processing technologies

1) Batch processing

Given lots of static data, we can process them off-line in a batch processing manner using distributed computing. Apache Hadoop [1] is a distributed computing framework modeled after Google MapReduce [2] to process large amounts of data in parallel. A Hadoop cluster can process data stored on the Hadoop Distributed File System (HDFS) [3], which is modeled after Google GFS [4]. HDFS works more efficiently with a few large data files than numerous small files. A real-world Hadoop job typically takes minutes to hours to complete, therefore Hadoop is not for real-time analytics, but rather for offline, batch data processing. Recently, YARN (Yet Another Resource Negotiator) was proposed which is based on Hadoop but aims to decouple Hadoop from MapReduce paradigm to accommodate other parallel computing models, such as MPI (Message Passing Interface) and Spark.

Hadoop APIs are often considered low level and not easy to program with. Researchers thus designed various abstraction techniques to encapsulate the low-level Hadoop APIs and provide easy programming interfaces. Pig, Hive, Crunch, Scoobi are all such efforts. In Pig, Crunch, and Cascading, data transformation, splitting, merging, and join may be conducted individually in a pipeline manner. In contrast, Hive works like a data warehouse and offers a SQL compatible interactive shell. Programs developed for these platforms are compiled with native Hadoop low level APIs. Given the simplified programming interfaces in conjunction with libraries of reusable functions, development productivity is greatly improved.

2) Streaming processing

In Streaming processing, data flows continuously through a topology such as a network of transformation blocks. Twitter Storm [5] is an open-source, Big Data processing platform for distributed, real-time streaming processing. Storm implements a data flow model, in which the slices of data being analyzed at any moment in an aggregate function is specified by a sliding window. A sliding window may be like "last hour", or "last 24 hours", which is constantly shifting over time. The size of a sliding window cannot grow infinitely. Data can be fed to Storm through distributed messaging queues.

Stream data processing is not intended to analyze a full big data set, nor is it capable of storing that amount of data. There are other similar streaming processing platforms for Big Data, including Apache S4 [6] and Trident. Apache S4 is a product for distributed, scalable, continuous, stream data processing. Trident is an abstraction API of Twitter Storm that makes it easier to use.

3) Real-time processing

In contrast to the batch processing and streaming processing which analyzes Big Data off-line, some business applications such as credit-card fraud require the data to be processed immediately. This is called Big Data OnLine Analytical Processing (OLAP). OLAP is extremely data and compute intensive in that terabytes (or more) of data are scanned to compute arbitrary data aggregates within seconds. Note indexing is indeed not helpful in a full "table" scan; in addition, building an index on a big data set is costly and slow.

Google Dremel [7] and Apache Drill [8] are representative real-time platforms for Big Data processing. In these platforms, coordination, query planning, scheduling, and execution are all distributed throughout nodes in a cluster to maximize parallelization. These platforms favor a query-efficient columnar storage format. the existing row-based data are transformed using a MapReduce job before

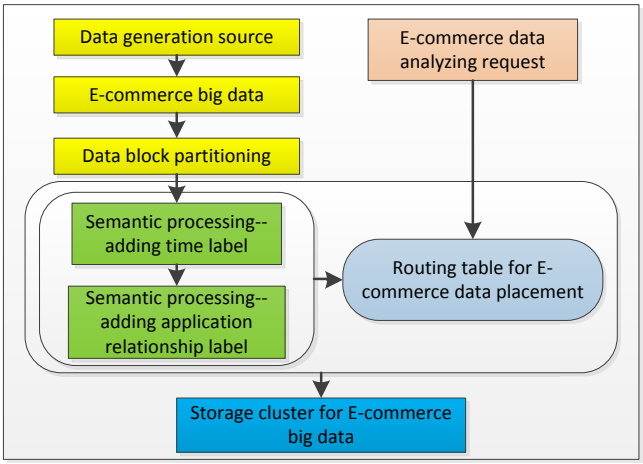

Fig.2 A platform architecture for E-commerce Big Data processing.

saving directly in a columnar format in a data store (NOSQL, NewSQL, Relational, and more). These platforms are not a replacement of Hadoop since they have limitations in full "table" scan-based ad-hoc queries. With the data source being an OLTP database (BigTable, HBase), a write made by an end user is reflected instantaneously in an analysis report. Such architecture brings you a Big Data OLAP system with typical latency in seconds range. To save resources, it is recommended to build a backup view on an analysis job, and return that view if no changes are made on the result. Impala and Drill have nice integration with commercial business intelligence (BI) tools like Tableau.

B. A new platform architecture for E-commerce Big Data processing

As can be observed, current Big Data processing platforms such as Hadoop, Hadoop++ [9] and CoHadoop [10] are inefficient under certain circumstances. For example, these platforms need a shuffle stage between the map state and reduce stage to put the related data together. Therefore, in this paper, we propose a novel platform architecture for E-commerce Big Data processing.

1) Overview

The architecture of the new platform is shown in Fig. 2. In our architecture, these data will be split into blocks according to the Hadoop mechanism. Then these data blocks are marked by time labels. In addition, data blocks can be marked by application relationship labels which are related to the preliminary analyzing requirements in E-commerce. For example, if dataset A and dataset B are connected in a joint query, we can conclude that some data blocks in dataset A and B are related to each other. Therefore, we can add application relationship labels onto these related data blocks. Finally, data blocks that are closely related will be stored onto a same node. After the semantic processing of all data blocks, we can get the routing table for E-commerce Big Data that contains the data placement information. With this new architecture, the shuffle stage can be removed and the efficiency of computation especially the relation query computation is improved.

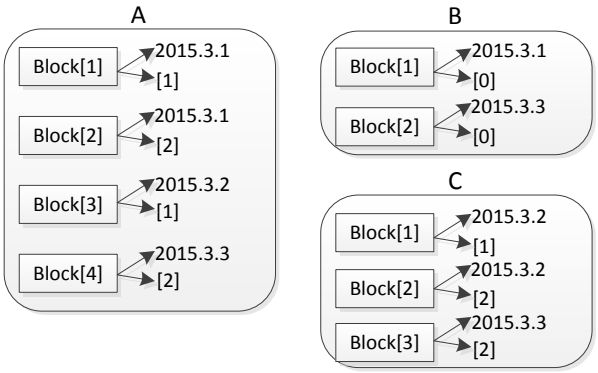

Fig. 3 Semantic processing: adding time and relationship labels

2) Details of semantic processing

In the semantic processing, time label and relationship label can be applied individually or simultaneously onto data blocks so that related data blocks can be arranged to store onto a same node. The labels are defined as following:

Time labels: use the creation time of the data as their time labels;

Relationship labels: label=0 means no relationship; otherwise, data blocks with a same label value have a relationship on computing.

As shown in Fig. 3, there are three datasets: A, B and C. Dataset A has four data blocks: Block [1], Block [2], Block [3] and Block [4]. Dataset B has two blocks: Block [1] and Block [2]. Finally, Dataset C has three blocks: Block [1], Block [2] and Block [3]. A data block is 64 MB by default in Hadoop. We can apply the time label, the relationship label or both labels onto these blocks.

Adding the time labels can improve the efficiency of computation. For example, if a query only requests the information on March 2, 2015 of dataset A, then only Block [3] need to be processed. This significantly reduces the computation time. Similarly, adding the relationship label can also greatly improve the efficiency. For instance, Block [1], Block [3] of dataset A and Block [1] of dataset C have the same relationship label of 1, so they will be stored onto a same node, say Storage Node A. Similarly, Block [2], Block [4] of dataset A and Block [2], Block [3] of dataset C have the same relationship label of 2, so they will be stored onto a same node, say Storage Node B. When a query requests for related blocks, they can be fetched from a same storage node to a compute node, thus the shuffle stage and possibly the reduce stage can be removed, which significantly improves computing efficiency.

Utilizing Big Data

After processing Big Data, how to use the results to benefit E-commerce? This section discusses the applications and the system architecture to utilize Big Data.

C. Applications of Big Data

Since Big Data comes from business transactions, they can be used to improve various aspects of E-commerce: pricing, improving custom experience, business prediction, managing supply chain and so on.

1) Pricing:

Using Big Data, companies can adjust pricing on their numerous products based on customer demands and competitions.

2) Improving custom experience

For E-commerce, two major challenges are user acquisition and user keeping. The competition is tough and most consumers have wide range of choices for the same product. It is of vital importance to let customers feel satisfied with the business. Companies are closely analyzing the buying path for each customer and improving customer experience making it a seamless process. Companies are analyzing data from customer service call and improving processes.

3) Recommendation and advertising

Companies can analyze customer habits and make customized recommendations based their preferences. A typical example is to analyze what consumers bought and what age group they represented. it can provide you with the opportunity to tailor your marketing by approaching them with products that are similar to those they have bought before. Meanwhile, analysis can also look at consumers' overall shopping habits, such as when they go online, which sites they visit, what products they like and what they say on social media. These preferences may be affected by seasons and holidays.

4) Business Analytics

Enterprises can conduct useful analysis based on Big Data. For example, supermarkets can analyze the best sale products and buy more in advance from manufactures. Similarly, manufactures can predict the goods to produce in next month that can gain more profits.

5) Managing supply chain

Big data can also allow businesses to combine historical information with up-to-date analysis to plan for future events. For example, finding out about any ongoing issues that might interrupt the supply chain—such as traffic concerns—can help you not just to respond in the short term but make plans in the future, which could offer you a competitive advantage. E-commerce companies are dealing with lot of moving parts – vendors, logistics, warehousing, delivery, returns etc. They are building efficient systems using analytics to manage the process. Companies are using Internet of things, to collect and communicate data on a wide range of conditions and redefining supply chain intelligence.

D. System architecture of utilizing Big Data

Big data introduces highly specialized features that set it apart from legacy systems. Here we propose a system architecture for utilizing Big Data. The architecture consists of multiple layers of different functionalities, as shown in Fig. 4. The functionalities of layers are described as following:

1) Storage layer:

This layer is in charge of storing large and diverse amount of data on storage devices such as disk arrays. Disk storage becomes more cost-effective since disk technologies

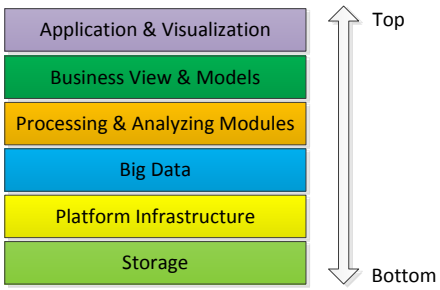

Fig. 4 A system architecture for Big Data utilizing

now evolve very quickly. The design goals of this layer include the scalability to add more storage capacity cheaply and flexibly and the efficiency for the archival and quick retrieval of large amounts of data.

2) Platform infrastructure layer:

This layer is essentially the collection of hardware that enables high-performance processing of Big Data. This layer includes capabilities to integrate, manage, and apply powerful computational processing to the data. An example of this layer is a high-performance cluster running a Hadoop platform. As described in previous section, Hadoop was designed and built to optimize complex manipulation of large amounts of data while vastly exceeding the price/performance of traditional databases. Hadoop is a unified storage and processing environment that is highly scalable to large and complex data volumes. In the new world of Big Data, open source projects like Hadoop have become the de facto processing platform for Big Data.

3) Big data layer

The growth of Big Data is as broad and complex as the applications for them. Big data come from numerous sources such as sensor data, bank transaction records, or social media interactions, to name

a few examples. The Big Data layer in the architecture implies that the collection of data is a separate asset, warranting discrete management and governance.

4) Processing and analyzing modules layer

This layer consists of the actual programs to manipulate and process the data, such as a map-reduce program running in a Hadoop platform. Just as Big Data vary with the business application, the programs also need to vary. The map-reduce program not only distribute data across the disks, but to apply complex computational instructions to that data. Programs run in parallel across multiple nodes in the infrastructure layer to process and analyze data.

5) Business view and models layer

Depending on the Big Data application, additional processing via MapReduce or custom Java code might be used to construct an intermediate data structure, such as a statistical model, a flat file, a relational table, or a cube. The resulting structure may be intended for additional analysis, or to be queried by a traditional SQL-based query tool. This business view ensures that Big Data is more consumable by the tools and the knowledge workers that already exist in an organization.

6) Application and visualization layer

This layer considers how to organize and present the result data. One of the more profound developments in the world of Big Data is the adoption of data visualization. Data visualization tools allow the average business person to view information in an intuitive, graphical way.

Conclusion

Even though it hasn't been long since the advent of Big Data, enterprises are rapidly joining the data economy. The primary value from Big Data comes not from the data in its raw form, but from the processing and analysis of it and the insights, products, and services that emerge from analysis. In this paper, we discussed the existing technologies and proposed new technologies for processing and utilizing Big Data in E-commerce. Our paper can help enterprises to exploit Big Data more effectively.

Acknowledgment

The author would like to thank anonymous reviewers for their fruitful feedback and comments that have helped her improve the quality of this work. This work is supported by the 2014 university research program funding of Shanghai University of Political Science and Law (Grant No. 2014XJ18).

References

[1] The Apache Software Foundation, "Hadoop MapReduce", http://hadoop.apache.org/docs/r1.1.1/mapred_tutorial.html.

[2] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Proc. of the 6th symposium on operating System Design and lmplementation (OSDI), USENIX Association, 2004, pp.1-13.

[3] The Apache Software Foundation, "Hadoop Distributed File System", http://hadoop.apache.org/docs/r1.1.1/hdfs_design.html.

[4] S. Ghemawat, H. Gobioff and S-T. Leung, "The Google file system," Proc. of the 19th symposium on Operating Systems Principles, 2003, pp. 29-43.

[5] A. Toshniwal, S. Taneja, A. Shukla, K. Ramasamy, J. M. Patel, S. Kulkarni, J. Jackson, K. Gade, M. Fu, J. Donham, N. Bhagat, S. Mittal and D. Ryaboy, "Store @ Twitter", Proc. of SIGMOD, ACM Press, 2014, pp.147-156.

[6] L. Neumeyer, B. Robbins, A. Nair, A. Kesari, "S4: Distributed Stream Computing Platform," Prof. of IEEE 13th International Conference on Data Mining Workshops, IEEE Computer Society, 2010, pp. 170-177, doi:10.1109/ICDMW.2010.172.

[7] S. Melnik, A. Gubarev, J. J. Long, G. Romer, S. Shivakumar, M. Tolton, T. Vassilakis, “Dremel: Interactive Analysis of Web-Scale Datasets,” Communications of the ACM, Vol. 54 No. 6, pp. 114-123, doi: 10.1145/1953122.1953148.

[8] The Apache Software Foundation, “Apache Drill: Interactive Ad-Hoc Analysis at Scale”, https://drill.apache.org/overview/.

[9] J. Dittrich, J.-A. Quiané-Ruiz, A. Jindal, Y. Kargin, V. Setty, J. Schad, “Hadoop++: Making a Yellow Elephant Run Like a Cheetah”, Proc of the VLDB Endowment, 2010, Vol. 3, No. 1, pp. 518-529.

[10] M. Y. Eltabakh, Y. Tian, F. Ozcan, R. Gemulla, A. Krettek, J. McPherson, “CoHadoop: Flexible Data Placement and Its Exploitation in Hadoop”, Proc. of the VLDB Endowment, 2011, Vol. 4, No. 9, pp. 575-585.