Big Data Management

Putu Suta Adya Dharma Rahadi [1], Kheri Arionadi Shobirin [2], Sri Ariyani[3]
[1][2] Department of Electrical and Computer Engineering, Post Graduate Program, Udayana University,
Email: suta.rahady@gmail.com
[3]Department of Electrical and Computer Engineering, Udayana University

Abstract - In recent years, big data is a term representing large and complicated datasets that traditional data processing, including acquisition, pre-processing, storage, analytics, and visualization, etc., are not capable of tackling. Big data is a term that describes a large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. The usage of big data nowadays has deeply penetrated into various industries, such as e-commerce, retail, manufacturing, media, transportation, health-care, education, etc. In e-commerce, big data can be used in product ranking solution which offers solution to provide accurate information to the user based on their behavior when they are accessing any e-commerce website.

Keywords - Big data, big data management

I. INTRODUCTION

A hot topic in recent years, big data is a term representing large and complicated data sets that traditional data processing, including acquisition, pre-processing, storage, analytics and visualization, is not capable to tackle. [1]. The birth of big data, as a concept if not as a term, is usually associated with a META Group report by Doug Laney entitled "3-D Data Management: Controlling Data Volume, Velocity, and Variety" published in 2001 [2]. Further developments now suggest big data problems are identified by the so-called "5V": volume (quantity of data), variety (data from different categories), velocity (fast generation of new data), veracity (quality of the data), and value (in the data) [3].

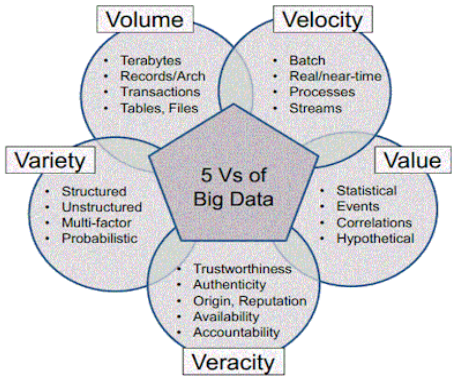

Image 1: 5V Big Data

The usage of big data nowadays has deeply penetrated into various industries, such as e-commerce, retail, manufacturing, media, transportation, health-care, education, etc. This paper will focus on management big data (e-commerce)

E-commerce nowadays has tremendously changed people's consumption mode. Many shopping experiences has shifted from off-line shopping mall to online virtual stores. Through website and app endpoint usage to do online shopping, ecommerce company can accumulate and capture huge amount of user's online behavior, including browsing, adding to cart, adding to favorite, purchasing, etc. All of these online behaviors can reflect user's purchasing preference, inclination, and purpose.

As user's online behavior embodies big data's five "V" features, traditional data analysis based on relational database is unable to efficiently and effectively process so huge amount of data. With the help of big data technology, the hidden valuable user preference information can be efficiently mined from online behavior, which is of high value for ecommerce company to make targeted marketing-campaign, provide users with personalized service and increase online shopping experience.

This paper is structured in the following formats: Section II describes the concept of the Big Data, Big Data Tools. Section III describes the application design implementation a big data based product ranking solution. Section IV concludes the concept and implementation a big data based product ranking solution for further research.

II. CONCEPT, TOOLS AND TECHNOLOGY USED

A. Big Data

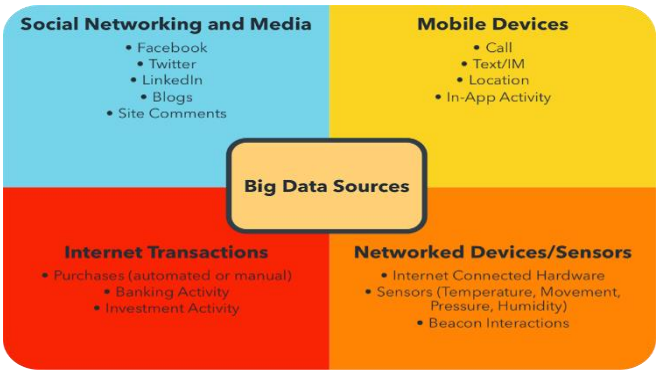

Image 2: Big Data Sources

Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it's not the amount of data that's

important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

B. Big Data Tools

Working with Big Data, always to be thinking about how to store it. Part of how Big Data got the distinction as "Big" is that it became too much for traditional systems to handle. A good data storage provider should offer you an infrastructure on which to run all your other analytics tools as well as a place to store and query your data.

1. **Hadoop**, the name Hadoop has become synonymous with big data. It's an open-source software framework for distributed storage of very large datasets on computer clusters. All that means you can scale your data up and down without having to worry about hardware failures. Hadoop provides massive amounts of storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.
2. **Cloudera**, Speaking of which, Cloudera is essentially a brand name for Hadoop with some extra services stuck on. They can help your business build an enterprise data hub, to allow people in your organization better access to the data you are storing. While it does have an open source element, Cloudera is mostly and enterprise solution to help businesses manage their Hadoop ecosystem. Essentially, they do a lot of the hard work of administering Hadoop for you. They will also deliver a certain amount of data security, which is highly important if you're storing any sensitive or personal data.
3. **MangoDB**, MongoDB is the modern, start-up approach to databases. Think of them as an alternative to relational databases. It's good for managing data that changes frequently or data that is unstructured or semi-structured. Common use cases include storing data for mobile apps, product catalogs, real-time personalization, content management and applications delivering a single view across multiple systems. Again, MongoDB is not for the data newbie. As with any database, you do need to know how to query it using a programming language.
4. **Talend**, talend is another great open source company that offers a number of data products. Here we're focusing on their Master Data Management (MDM) offering, which combines real-time data, applications, and process integration with embedded data quality and stewardship. Because it's open source, Talend is completely free making it a good option no matter what stage of business you are in. And it saves you having to build and maintain your own data management system – which is a tremendously complex and difficult task.

C. Data Cleansing

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 and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Here are some tools for data cleaning:

1. **OpenRefine** (formerly Google Refine) is an open source tool that is dedicated to cleaning messy data, explore huge data sets easily and quickly even if the data is a little unstructured.
2. **DataCleaner,** DataCleaner recognizes that data manipulation is a long and drawn out task. Data visualization tools can only read nicely structured, "clean" data sets and many others data cleaning tools.

D. E-commerce

E-commerce (electronic commerce or EC) is the buying and selling of goods and services, or the transmitting of funds or data, over an electronic network, primarily the internet. These business transactions occur either as business-to-business, business-to-consumer, consumer-to-consumer or consumer-to-business [4].

III. A BIG DATA BASED PRODUCT RANKING SOLUTION

Big Data Based Product Ranking Solution is one of the uses of big data management. Product ranking, based on user's online behavior analysis. It aims to order products according to relevance degree for a certain query. In online product search scenario, given product search words, and the search result is the sorted product list relevant to search words. In product online assortment scenario, given a product classification, i.e., brand, hierarchical category and

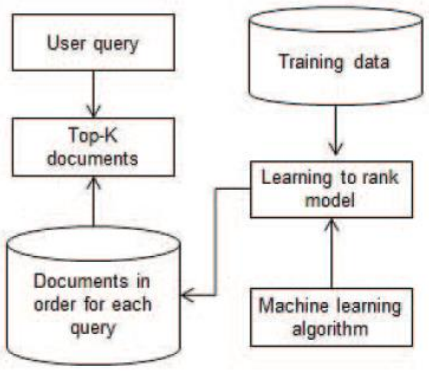

Image 3: Mechanism for learning to rank

production area, etc. The assortment result is a sorted product list related to a product classification. Learning to rank is a supervised or reinforcement machine learning technology in information retrieval field (The main mechanism

is referred to image. 2). Each training data consists of a pair of query and document which is characterized by a relevance degree matching the query. The relevance degree for each document related to a query can be manually rated by human experience. As it is impossible to rate all documents, only the top few documents, a subset of documents, are pooled by existing ranking models.

Case example: when a user who does online shopping is showing behaviors of looking for cosmetic product's information, the type of advertising shown on the web-site will automatically relate to cosmetics. This is possible because the machine learning processes the behavior done by the user.

IV. CONCLUSION

Mining users 'online behavior, including click-though and purchasing, helps e-commerce companies to know users' preference, inclination and purpose. Learning to rank is a widely-used approach to mine the document ranking matching each query. Rather than applying technology to extract text features of products, we propose an approach to extract the inner product line characteristics as the features for training data. To implement this approach, learning to rank scenario under big data, we propose Also a big data-based architecture to process and analyze huge number of users' behavior data.

V. ACKNOWLEDGMENT

The authors wish to thank to I Made Sudarma Lecturer of IT Audit subject in Master's Degree of Electrical Engineering - Information System Management and Computer Udayana University.

VI. REFERENCES

[1] Li, J., Shao, B., Xu, J., Li, H., & Wang, Q. (n.d.). A Big Data Based Product Ranking Solution. Industry & Solutions Research, IBM Research-China.

[2] Laney, D. (2001). 3D data management: Controlling data volume, velocity, and variety. META Group.

[3] O. Terzo, P. Ruiu, E. Bucci, and F. Xhafa, "Data as a service (DaaS) for sharing and processing of large data collections in the cloud," in Proc. Int. Conf. Complex Intell. Softw. Intensive Syst., 2013, pp. 475–480.

[4] Rouse, M. (2016, December 21). e-commerce (electronic commerce or EC) . Retrieved from Techtarget: http://searchcio.techtarget.com/definition/e-commerce