Research on Big Data Computing Model based on Spark and Big Data Application

Sida Chen1*

1 School of computer science and technology, TIANGONG University, Tianjin, China, 300387

*Corresponding author’s e-mail: gaoming@cas-harbour.org

Abstract. Under the background of the information age, with the rapid development of cloud computing and Internet of things technology, all kinds of data information grow rapidly. How to transform massive data information into effective resources is the key point of big data technology research. This paper introduces the inherent laws of big data mining technology, determine and use valuable information data, which can open a new thinking and cognitive perspective, and it is of great significance to the development of the social economy. It can be concluded that the big data ecosystem based on spark platform and application of big data still have a lot of room for development, but there are also some problems. Nowadays, as a low latency cluster distributed computing system for big data collection, spark platform can provide more support for improving the efficiency of big data mining, but some of its methods are not perfect.

1. Introduction

Hadoop platform and spark platform, which are specially responsible for big data processing, play an important role in accelerating the efficiency of big data computing and processing and ensuring the full utilization of big data resources. The core technology of Hadoop technology, spark technology and the understanding of its ecosystem are also the fundamental premises to realize the in-depth application of the two technologies. This paper will introduce the application of big data in our life from two aspects: the big data computing model based on the spark platform and the big data ecosystem based on the spark platform. And data mining technology can use algorithms to mine valuable information and interesting patterns in massive data. Therefore, this study can not only provide theoretical reference to deepen the research on Hadoop technology and spark technology in the era of big data, but also provide some guidance and help for the practical application of the two technologies.

2. Big data computing model based on spark platform

Apache Spark is a text-based framework that was presented at the University of Berkeley in 2009. The existence of multiple advantages has made the processing engine a powerful and useful process for macro processing, and distinguishes it from other tools, such as Hadoop and Storm [1, 2, 3-6]. All features provided by Apache Spark are built on top of the core[7]. For Spark, there are three APIs -Java, Python and Scala. The Spark core is an API location that defines the resilient distributed RDD dataset, which is the concept of Spark's original programming [8].

Apache spark is also an open-source general purpose processing engine that allows data scientists to build and run fast and sophisticated applications on Hadoop. Spark provides a set of simple and
easy-to-understand programming APIs that are employed to build applications at a rapid pace in Scala. The Spark Engine supports a set of high-level tools that support SQL like queries, streaming data applications, complex analytics such as machine learning, and graph algorithms. The state and mutation in Scala make the evaluation of expressions, especially in parallel systems, much easier than imperative languages like Java or Python, which leads for harnessing modern hardware that has more cores, more nodes, and more RAM, rather than faster clock speeds. It has the powerful property based testing DSL to write an exhaustive test suite. In PBT this paper defines properties, not cases. Spark’s rich API facilitates writing MapReduce stages as elegant linear, and short code doesn’t mean short run times, which is the advantage of functional programming, especially from algebraic type.

2.1. Big data ecosystem based on spark platform

2.1.1. Spark Runtime. When analyzing spark runtime, the first thing that should be determined is its functionality, such as task scheduling and memory management. When using RDD structure to transmit data inside spark, the premise is to determine the core logic data information of spark, which will have certain similarity with object concept under some conditions. The first step is to divide all the data into several subsets, and at the same time, each subset should be able to be transmitted to any node in the cluster for further processing. The second step is to provide reliable protection for the intermediate results of the calculation, through this protection it can obtain the same file content as the calculation results, and at the same time backup the file content in the subset node. The third step is in the case of any data subset calculation error, the rest of the subset should be sorted again to meet the fault tolerance mechanism.

2.1.2 Spark SQL. Spark SQL is developed from shark. In order to realize hive compatibility, shark reuses the logic of HQL parsing, logical execution plan translation and execution plan optimization in hive. It can be approximately considered that only the physical execution plan is replaced by spark job (supplemented by various optimizations such as memory column storage, which have little relationship with hive).

At the same time, it also relies on history Metastore and Hive serde (used to be compatible with various existing history storage formats).

Spark SQL only relies on HQL parser, history meta store and Hive serde in the aspect of history compatibility. In other words, since HQL is parsed into an abstract syntax tree (AST), it is all taken over by spark SQL. Catalyst is responsible for execution plan generation and optimization. With the help of scala's functional language features such as pattern matching, it is much simpler to develop execution plan optimization strategy with catalyst than hive.

Sparksql has many advantages:

The first advantage is integration. Seamlessly, the Sparksql mixes SQL queries with spark programs. Spark SQL allows participants to use SQL or familiar dataframe API to query structured data in Spark Program. It can be used in Java, Scala, Python and r. This advantage can make our research more efficient, using our skilled API to operate on the data.

The second advantage is hive integration, running SQL or hiveQL queries on existing warehouses.

The third advantage is unified data access, which connects to any data source in the same way. Dataframes and SQL provide common methods to access various data sources, including hive, Avro, parquet, ORC, JSON and JDBC. You can even add data across these sources.

The fourth advantage is unified data access, data connection of standard connection, through JDBC or ODBC connection. Server mode provides industry standard JDBC and ODBC connection for business intelligence tools.

2.1.3 Graph X. Graph x is a key subproject of spark. It is built on the basis of spark. Based on large-scale graph computing, and supported by graph x derived features, it can realize the rich computing of large graph data processing in spark ecosystem. When other related components are
integrated, with the support of powerful data processing capability, all applications can be obtained through multiple scenarios. By analyzing the functionality of graph x, it can be determined that it can provide rich graph data operators, such as core and optimization operators [3]. Moreover, graph x can also meet the graph operation of multiple distributed clusters, and has sufficient API interfaces. Especially when it reaches a certain graph scale, it can carry out a lean algorithm to promote the large-scale processing of distributed atlas. As an important part of spark, the key of graph x is to further improve the data absorption and scale.

2.1.4 Spark Streaming. In addition to the good ability to expand spark data, spark system can also divide spark streaming data stream strictly according to time mode, and finally form RDD. Processing streaming data with relatively small time intervals will be affected by processing delay to a certain extent, so it can be regarded as just in time processing system [4]. Moreover, spark streaming has strong fault tolerance, including error handling and recovery. And spark streaming can also seamlessly connect with relevant spark ecological modules. In addition to completing flow data together, it can also effectively deal with some complex phenomena.

2.1.5 MLlib-Algorithm library. MLlib algorithm library analyzes the process of calculation algorithm, and it can be determined that it has high complexity. In iterative calculation, all calculations need to be put into disk to wait for the start and processing of tasks, and the whole process needs a lot of CPU. Based on spark platform, part of the work can be done directly in memory. The corresponding iteration part of the calculation task is directly transferred to the memory because the further efficiency improvement of iterative calculation. Under the certain conditions, it can also realize the operation of disk and network. In summary, spark has more remarkable advantages in iterative computing, which can exist as a distributed machine learning platform. The application of collaborative filtering algorithm is to select and determine a reliable idea first, and then provide the idea to users. The steps of the algorithm can be divided into the following processes. First, system filtering. The users with common interests are selected, and then the items are selected and classified according to their preferences, and a new set or sequence is formed. In this process, users can be defined as neighbors, and corresponding users should be organized and utilized to determine the most effective execution method. Second, collaborative filtering. In order to complete the steps of user preferences collection, similarity analysis and recommendation based on the calculation results, the integration of user preferences is the core element of the final recommendation. First of all, we need to select a user system, and then group it according to user behavior to collect data behavior reliably, and then do further data preprocessing to recommend items that users may like through their preferences. Many applications mentioned in the big data application section also use this algorithm.

3. Application of big data

3.1 Combination of medical platform

With the development of new technologies, health care data is increasing. Estimated in 2012, the data is about 200 petabytes(PB), estimated to reach 250000 PB by 2020[9]. Analysis of these data is very important for acquiring knowledge, extracting useful information, and discovering hidden data patterns. And in the area of health care will improve the quality of services, reduce costs and reduce errors [10, 11]. This kind of data has many features: including high volume, variety, scalability, complexity, high-speed production and uncertainty, which makes it possible to use common data mining techniques and typical software and hardware to analyze this type of data [12-15].

A Big Data analysis is a process that organizes the data collected from various sources and then analyzes the data sets in order to discover the facts and meaningful patterns [16].

Large-scale data analyzes have many uses in the field of health. For example, early diagnosis of diseases such as breast cancer, in the processing of medical images and medical signals for providing high-quality diagnosis, monitoring patient symptoms, tracking chronic diseases such as diabetes,
preventing the incidence of contagious diseases, education through social networks, genetic data analysis and personalized (precision) medicine. Examples of this type of data are omics data, including genomics, transcriptomics, proteomics and pharmacogenomics, biomedical data, web data and data in various electronic health records (EHRs) and hospital information systems (HISs). Data contained in the EHR and HIS contain rich data including demographic characteristics, test results, diagnosis and information of each individual [17-25]. Therefore, analysis of health data has been considered with regard to its importance, so that it has led scholars and scientists to create structures, methodologies, approaches and new approaches for managing, controlling and processing Big Data [26]. In recent years, many tools have been introduced for Big Data analysis. It is intended to introduce the tools provided by the Apache Software Foundation and then compare them with each other after defining the Big Data and its feature. Some of the tools available for Big Data analysis are Apache Hadoop [27], Spark [28], and Flink [29], and the focus of these tools is on batch processing or stream processing. Most batch processing tools are based on the Apache Hadoop Infrastructure such as Apache Mahout.

Datastream analysis programs are often used for real-time analysis. With these tools, disease and medical plans can be prevented and made and so on. With those effective tools, in future the disease could be prevented and the useful medical plans would be made as well.

3.1.1 Disease prediction. Different from Google's prediction of influenza trend through search terms, disease prediction is based on the hospital's big data analysis of past patients' diseases and causes, which diseases a person will get and when he will get in his present life.

3.1.2 Disease prevention. Johnson & Johnson otdms diabetes management software, through the intuitive data map, shows the blood glucose data of patients with type 2 diabetes, helps patients with diabetes to better control blood glucose, and strengthen self-management and self prevention. It has played a very good effect . Through the integration of big data, we can analyze the condition of patients and provide them with correct medical advice, which is helpful for patients to realize self-service conveniently and quickly, and can play a good role in preventing diseases and their concurrent diseases [30].

For example, we used otdms management software to compare the levels of trace elements, blood glucose and blood lipid between the observation group and the control group.

Compared with the control group, the serum Zn and Mg contents in T2DM group were significantly decreased, while the FBG and TG contents were significantly increased (P < 0.05), and there was no significant difference in other indexes between the two groups (P > 0.05; Table1)

| Index     | T2DM group (n=116) | Control group (n=75) | P     |
|-----------|--------------------|----------------------|-------|
| Fe (μmol/L) | 12.54±4.07         | 13.08±5.80           | 0.105 |
| Zn (μmol/L) | 12.11±1.70         | 13.32±1.89           | <0.001|
| Mg (μmol/L) | 0.84 (0.79, 0.88)  | 0.88 (0.83, 0.92)    | 0.001 |
| Cu (μmol/L) | 15.30 (13.30, 17.30) | 14.70 (12.30, 17.30) | <0.336|
| FBG (mmol/L) | 6.69 (5.79, 8.07)  | 5.05 (4.87, 5.44)    | <0.001|
| TC (mmol/L) | 4.60 (3.72, 5.46)  | 4.43 (3.87, 5.31)    | 0.861 |
| TG (mmol/L) | 1.49 (1.06, 2.19)  | 1.25 (0.94, 1.72)    | 0.032 |
| HDL (mmol/L) | 1.08 (0.92, 1.27)  | 1.12 (0.98, 1.27)    | 0.335 |
| LDL (mmol/L) | 2.54 (1.95, 3.15)  | 2.49 (2.09, 3.06)    | 0.867 |
3.1.3 Speech recognition designated medical scheme. By guiding voice communication, the system can roughly diagnose the disease, and then provide medical guidance programs. This can help realize the automated treatment of minor diseases, and also solve the problem of the low ratio of doctors to patients in some places. Through voice guidance, such as the first step to know the symptom of the patients, according to the communication with patients to narrow a certain range. Patient's answer to further narrow a certain range. Next, through further inquiry, we can get the possible case analysis according to the background big data system, and give the corresponding score. Such as 80% of gastroenteritis, 50% of gastric ulcer. The other symptoms corresponding to each result are given, and the recommended treatment scheme or further examination scheme is given. Thus, it can reduce the doctor's diagnosis and treatment of minor diseases, reduce the doctor's burden, and reasonably use and allocate hospital resources.

3.2 Natural disaster

Natural disaster big data is the result of an effective response to the "application problems of disaster prevention, disaster relief and disaster emergency management brought by massive disaster complex information in the Internet of things era"

A new way of thinking, technical system and innovation ability to deal with complex disaster information can effectively deal with comprehensive disaster prevention and reduction, and improve the ability and level of disaster prevention and reduction and relief. Its strategic significance and core values are mainly reflected in the following three aspects: first, on the level of strategic thinking, the disaster problem has surpassed the ability of contemporary scientific understanding, "high frequency and high intensity of disasters are the basic problems of the world", and comprehensive disaster reduction has risen to the level of national strategy. Second, at the level of innovation and development of information science and technology, disaster big data has brought all-around challenges to the traditional information science and technology system. Big data science is accelerating the formation of a new theory and technology system with complex information (or irregular information) as the core. Third, at the level of social innovation and development, disaster big data is the core capability to ensure the realization of the world's "Internet +" and "smart security, smart disaster reduction" strategies.

Disaster cloud computing refers to efficient disaster information monitoring, data standardization, abnormal prediction and early warning of all kinds of disaster factors, all kinds of disaster risk assessment, all kinds of disasters

Risk early warning, emergency management analysis, emergency information services and other analysis and calculation models. The goal is to build a super large distributed ring in the way of shared infrastructure

Its main function is to provide fast and safe disaster cloud data processing, prediction and early warning model, disaster risk assessment model, emergency management service model and other services.

![Figure 1 Schematic diagram of emergency prevention path for major natural disasters](image-url)
3.3 A popular application

3.3.1 Mobile app usage. Both app store and Google play have more than 1.5 million applications, and the number of relevant developers has exceeded 300000 respectively. According to the survey of flurry analytics, the usage of global mobile applications will increase by 58% in 2020. In a large number of applications, only a small part of the applications are often used by people. How to stand out from the many applications becomes a difficulty for every developer and entrepreneur.

3.3.2 Next super app + "guess what you like". Develop an application that meets the needs of users, is independent of existing software, and provides a convenient life for future users. This application meets the needs of people's lives and may become the next super application. How to develop an application, adapt to market dynamics, user demand changes, maintain the characteristics of the application, in the same type of application, maintain their own style, is more important. The term "guess what you like" is known for the first time on the home page of Amazon app. According to your browsing history, you can judge what you may want to buy or be interested in. Different from aimless promotion, it greatly improves the proportion of product publicity and sales success. Guess what you like is not only applicable to shopping apps, such as reading, video websites, personalized push can not only facilitate the choice of users, but also help enterprises understand users and enhance their adhesion. On a video website, you can know which type of video you are more interested in. Today's video websites not only integrate video content, but also produce more video content created by enterprises, which can produce more high-quality and user-friendly videos.

3.4 Face recognition

3.4.1 Business big data analysis system based on face recognition. Firstly, face detection, face recognition, gender recognition, age recognition, facial expression recognition and gaze duration statistics are processed by face recognition technology to collect the information of users watching advertising machines and digital signs. After collecting a large amount of data, big data processing technology is used to analyze users' preferences for content, so as to provide more humane services for users of different gender, age and emotional state. In addition, we can put different contents in different periods to make reference. It is called "guess what you like" for short. It emphasizes the relationship between people and things. It can dynamically recommend advertisements and messages to different people.

3.4.2 Non manual application. It is not used in traditional applications such as click, gesture and other operation modes. It can realize non-manual and automatic application by recognizing face, eyes and other information and predicting the next operation. For example, we can judge whether we are interested in the fixation time and facial expression of a text or a picture, directly enter the article or enlarge a picture, and so on. Through face recognition and big data features, the application is fully automatic.

4. Conclusion

First, this paper describes the big data computing model based on spark platform, and summarizes its related applications: Spark SQL, graph x, mllib algorithm library, spark streaming. But there are still some problems, for example, platform stability. Therefore, it is necessary to be cautious in the commercialization process and pay more attention to testing. The data mining algorithms that have been implemented are relatively few. The data mining algorithms that have been implemented in mllib are still relatively limited, and a large number of data mining algorithms have not been parallelized. Therefore, in specific applications, it is necessary to implement data mining algorithms that are not parallelized on specific problems.
Second, this paper introduces the application of big data, and illustrates the importance of big data technology from medical platform combination, natural disaster prediction, popular applications, face recognition and other aspects.

Third, big data computing model still needs to be improved, the real-time data processing methods should be enhanced and innovated by us. In the future social development trend, according to the characteristics of the development trend of big data technology, in the future society, big data application technology needs to develop in the direction of refinement, comprehensiveness and scientific. For example, future data processing methods will be more refined. The current big data application processing technology is mainly batch processing, which has some limitations and can not improve the frequency of data processing. In some important industries, if the big data processing technology can not meet the requirements of accuracy and precision, the economic and social benefits will be greatly reduced. Therefore, in the future social development, real-time data processing can be used to continuously optimize data processing to meet the development needs of big data technology.

Acknowledgement
First and foremost, I would like to show my deepest gratitude to my professor and teaching assistant in my university, who have provided me with valuable guidance in every stage of the writing of this thesis. Further, I would like to thank all my thesis teachers for their encouragement and support. Without all their enlightening instruction and impressive kindness, I could not have completed my thesis.

References
[1] Verma A, Mansuri AH, Jain N. Big data management processing with Hadoop MapReduce and spark technology: A comparison. Symposium on Colossal Data Analysis and Networking. 2016.
[2] Zaharia M, Xin RS, Wendell P, Das T, Armbrust M, Dave A, et al. Apache spark: A unified engine for big data processing. Communications of the ACM. 2016;59(11):56-65
[3] Estrada R, Ruiz I. Big data SMACK: A guide to apache spark, mesos, akka, cassandra, and kafka. Apress; 2016.
[4] Meng X. Mlib: Scalable machine learning on spark. Spark Workshop; 2014.
[5] Zaharia M, Chowdhury M, Das T, Dave A, Ma J, McCauley M, et al. Fast and interactive analytics over hadoop data with Spark. Usenix Login. 2012; 37(4): 45-51.
[6] Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I, et al. Spark: Cluster computing with working sets. Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing; 2010.
[7] Team, D. Apache spark ecosystem: Complete spark components guide[Internet]. 2017[cited: 1 Dec2018]. Available from: https://data-flair.training/blogs/apache-spark-ecosystem-components
[8] Safabakhsh M. Apache spark [Internet]. 2018[cited: 1 Jul 2019]. Available from: http://myhadoop.ir/?page_id=131
[9] Hermon R, Williams PA. Big data in healthcare: What is it used for? 3rd Australian eHealth Informatics and Security Conference; 2014.
[10] Chen M, Mao S, Liu Y. Big data: A survey. Mobile Netw Appl. 2014; 19(2): 171-209.
[11] Ristevski B, Chen M. Big data analytics in medicine and healthcare. J Integr Bioinform. 2018; 15(3): 1-5. PMID: 29746254 DOI: 10.1515/jib-2017-0030 [PubMed]
[12] Mooney SJ, Pejaver V. Big data in public health: Terminology, machine learning, and privacy. Annu Rev Public Health. 2018; 39:95-112.
[13] Jin X, Wah BW, Cheng X, Wang Y. Significance and challenges of big data research. Big Data Research. 2015; 2(2): 59-64.
[14] Bello-Orgaz G, Jung JJ, Camacho D. Social big data: Recent achievements and new challenges. Information Fusion. 2016; 28:45-59.
[15] Arockia PanimalarS, Varneka ShreeS, Veneshia KathrineA. The 17 V’s of big data. International Research Journal of Engineering and Technology. 2017;4(9): 329-33.

[16] Goga K, Xhafa F, Terzo O. VM deployment methods for DaaS model in clouds. In: Barolli L, Xhafa F, Javaid N, Spaho E, Kolici V. (eds) Advances in internet, data & web technologies. Lecture notes on data engineering and communications technologies, vol 17. Springer, Cham; 2018.

[17] Khan AS, Fleischauer A, Casani J, Groseclose SL. The next public health revolution: Public health information fusion and social networks. Am J Public Health. 2010;100(7):1237-42.

[18] Velikova M, Lucas PJF, Samulski M, Karssemeijer N. A probabilistic framework for image information fusion with an application to mammographic analysis. Medical Image Analysis. 2012;16(4):865-75.

[19] Antink CH, Leonhardt S, Walter M. A synthesizer framework for multimodal cardiorespiratory signals. Biomedical Physics & Engineering Express. 2017;3(3):035028.

[20] Sung W-T, Chang K-Y. Evidence-based multi-sensor information fusion for remote health care systems. Sensors and Actuators A: Physical. 2013;204:1-19.

[21] Benke K, Benke G. Artificial intelligence and big data in public health. Int J Environ Res Public Health. 2018; 15(12): 2796-805.

[22] Kim W-J. Knowledge-based diagnosis and prediction using big data and deep learning in precision medicine. Investig Clin Urol. 2018; 59(2): 69–71.

[23] Lee CH, Yoon H-J. Medical big data: Promise and challenges. Kidney Res Clin Pract. 2017; 36(1): 3–11.

[24] Archenaa J, Anita EM. A survey of big data analytics in healthcare and government. Procedia Computer Science. 2015;50:408-13.

[25] Andreu-Perez J, Poon CCY, Merrifield RD, Wong STC, Yang G-Z. Big data for health. IEEE Journal of Biomedical and Health Informatics. 2015;19(4):1193-208.

[26] Verma A, MansuriAH, JainN. Big data management processing with hadoop MapReduce and spark technology: A comparison. Symposium on Colossal Data Analysis and Networking. 2016;IEEE.

[27] Taylor RC. An overview of the hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinformatics. 2010; 11(12):S1.

[28] Zaharia M, Xin RS, Wendell P, Das T, Armbrust M, Dave A, et al. Apache spark: A unified engine for big data processing. Communications of the ACM. 2016;59(11):56-65.

[29] Carbone P, Ewen S, Haridi S. Apache flink: Stream and batch processing in a single engine. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. 2015.

[30] Yang Zhou, Zhang Yawei, Yuan Tuoping, Chen Siping, Wang Liang. Application of OTDMS diabetes management software in the prevention and treatment of diabetes [J]. seeking medical advice (second half month), 2011,9 (12): 404-405.