CertDB: A Practical Data Analysis System on Big Data

Jianqiang Li¹,a, Jia Cui¹,b, Bo Wang¹,c, Qi Wang¹,d, Ge Fu¹,e, Bing Jia²,f

¹The National Computer Network Emergency Response Technical Team/Coordination Center of China (CNCERT/CC)
²Henan Worker’s Cultural Palace

¹No. 3A, Yumin Road, Beijing, China
²No. 125, Jianshe West Road, Zhengzhou, China

ªBitJarod@163.com, b,c,d,e{cj, wb, wq, fg}@cert.org.cn, f67449131@163.com

Keywords: Big Data, Data Analysis, Knowledge Output.

Abstract. With the rapid development of big data technologies, more and more companies focus on the integration and analysis of the large scale data which are produced when the companies run, in order to discovering and solving their inner problems. Therefore, a data analysis system is in urgent need to process the big data generated by heterogeneous data sources (e.g. online and offline structured data, semi-structured data and non-structured data). Unfortunately, there is no integrated tool which can universally collect heterogeneous data sources, accomplish data ETL, storage, analysis, and knowledge output. In this paper, an integrated tool: CertDB is proposed, which can provide one-stop services of data collection, analysis, and knowledge output. The CertDB provides services to both expert and inexperienced data analysts, by proving both graphical BI interface and programming interface.

Introduction and Related Work

In the era of internet, precious value in big data becomes the driving force for the creative ability of a company. Mayer-Schönberger pointed out that the three major transitions the era of big data processing: focusing on the entirety but not sampling, the efficiency but not accuracy, the relevance but not causality [1]. The large-scale data processing technique is a great challenge. Nowadays, the technologies for processing big data mainly include batch processing of the offline data, real-time processing of online data [2], interactive processing of massive data and the comprehensive processing of the map data. This section introduces the characteristics of these four data processing techniques and the typical systems.

Batch Data Processing Systems. Batch data processing systems are used widely for offline large data storage and calculation, the requirement of real-time computing is not applicable. For batch data processing, accuracy and comprehension are important criterions. The representative systems include (a) social network systems: Facebook, Weibo, and Wechat; (b) E-business systems: Amazon, eBay, Alibaba; (c) Search Engines: Google, Yahoo!, Baidu.

Stream Data Processing Systems. Stream data processing systems have been widely used in the industry. The typical systems include Apache Storm [3], Facebook Scribe [4], Linkedin Samza[5], Cloudera Flume [6], and Apache Spark Streaming [7].

Interactive Data Processing Systems. Interactive data processing systems are known as Business Intelligence Tools (BI Tools). Compared with non-interactive data processing, interactive data processing is more flexible and easy to control. Typical systems include Apache Spark [8], Google Dremel [9], etc.

Map Data Processing Systems. Main map databases include GraphLab [10], Giraph [11] (based on Pregel, Neo4j, HyperGraphDB), InfiniteGraph [12], Cassovary [13], and Microsoft Trinity [14].

In the era of big data, big data systems make a great development based on the integration of open sourced technologies, in order to achieve the features of large scale, high-throughput, low-cost and strong-expansion abilities. In this paper, we propose an integrated system CertDB. Users can extract
data from heterogeneous data sources, achieve data ETL, analyze data and output knowledge by an one-stop way. The CertDB is applicable to both expert and inexperienced data analysts, by proving both BI interface and programming interfaces.

Main Framework of CertDB

As shown in Fig. 1, the CertDB includes three modules: The Universal Data Storage System (UDSS), Data Analysis Platform (DAP), and Knowledge Output System (KOS). The UDSS provides universal data storage by interacting with the outer systems. It also collects heterogeneous data generated different outer systems and provides service of data storage & universal query. At the same time, the UDSS supports data subscription. The DAP accomplishes complicated data analysis by computing and analysis framework. The KOS provides storage, modification, and analysis result display of the knowledge data.

The UDSS interacts with the outer systems, providing structured data, semi-structured data and unstructured data storage (OLTP and OLAP are supported). The UDSS can deal with both online data and offline data. By using the interactive BI tools provided by the DAP, data analysts request the source data, and achieve the analytical scene modeling & calculation by using the computing and storage resources provided by the DAP. In the end, the knowledge data can be stored and maintained in the KOS. The KOS also provides services of knowledge output and visualization (e.g. figures, tables, etc.).

Universal Data Storage System (UDSS)

As shown in Fig. 2, the Universal Data Storage System (UDSS) contains four modules: Non-Realtime and Realtime Data Load, Query Engine, Computing Engine, and Storage Platform. First, the UDSS loads analytical source data generated by different data sources into the storage system, and then provides services of data query and data maintenance. The Storage Platform supports full-text query, SQL-like query, and multimedia data query. Distributed file system and local file system are both supported.

Data Analysis Platform (DAP)

As shown in Fig. 3, The DAP includes two parts: Storage and Computing Framework and Data Analysis Support System (BI Platform). The Storage and Computing Framework deals with offline data & online streaming data load, data ETL, and provides data analysis & storage services, including map-reduce, Spark & Sparking Streaming, and map data computing services. Furthermore, The BI Platform is constructed upon the Storage and Computing Framework, which allows users to define their own operators, such as the machine learning functions defined in MLlib in Spark. The BI
Platform provides computing tasks management, schedule and security-related services. Meanwhile, it provides an interface for data analysts.

Figure 2. Main Framework of the Universal Data Storage System (UDSS).

Figure 3. Main Framework of Data Analysis Platform (DAP).
Knowledge Output System (KOS)

The Knowledge Output System (KOS) provides knowledge data storage, query and management services. The KOS also support operations of batch data update for large-scale data.

Conclusions

In this paper, we propose a large-scale data analysis systemCertDB. The CertDB provides data load, data ETL, data analysis, knowledge outputs services in a one-stop way. The CertDB is deployed in a platform of 85 servers cluster. The application demonstrates that the infrastructure and the framework of the system is dependable and applicable.

Acknowledgements

This work was financially supported by the Young Fund of National Computer Network Emergency Response Technical Team/Coordination Center of China (Grant Number: 2015QN-010) and the National High Technology Research and Development Program ("863"Program) of China (Grant Number: 2012AA012803).

References