Architecture of Data Fusion System Based on Big Data Technology

Chunhui Yang[a], Bo Cheng

State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China

[a]chuenfaiy@163.com

Keywords: data fusion, big data, system architecture, expansibility, real-time.

Abstract. Data fusion technology has been widely used in many fields, but the existing data fusion system is faced with the test of the system scalability, real-time processing and fusion abundant dimension. This paper tries to design a data fusion system architecture which can be able to support massive data processing. This paper begins with an analysis of the data fusion system, the development status quo, key techniques and problems currently facing. Then, we make an analysis combined with the characteristics of big data about what the design principles should the new data fusion system follow. Finally, we focus on data access, data computing and data storage in three aspects to explain in detail the design of the system architecture.

1. Introduction

Multi-sensor data fusion technology is a variety of disciplines in the forefront of technology, has been widely used in battlefield surveillance, automatic target recognition, industrial process control, robotics, remote sensing, medical diagnosis and other fields.

In civilian areas, sensor technology for the indicators to monitor the production equipment in the automatic production process and timely feedback the production process or state of alarm, ensure production equipment for the right working state; In the military field, the application of the sensor is more important, through the development of a large number of sensors for monitoring and artificial detection in distance persistence and other aspects of the short board, it has important significance to the development of national defense industry. Now with the fusion strategy is more and more complex, more and more large scale data fusion system, the current is a severe test time, expansibility, big data analysis techniques are applied to the data fusion system has become a trend.

2. Overview of data fusion technology

2.1 Definition of data fusion

Data fusion is an integrated automation information processing technology, compared with the exact definition can be summarized as [1]: make full use of multi-source heterogeneous data complementarity and computer high-speed operation and intelligent, comprehensive treatment data of different time and space, resulting in the reality environment a more precise description, to achieve more accurate recognition and judgement, improve the information quality. Therefore, a variety of sensor is the basis of information fusion. The multi-source heterogeneous data is the object of information fusion. Optimization and coordination of comprehensive treatment is the core of information fusion.

2.2 Process of data fusion

The process of data fusion mainly includes multi sensors (signal acquisition), data preprocessing, data fusion center (feature extraction, data fusion calculation) and the result output [2] and so on. The process of data fusion is shown in figure 1.

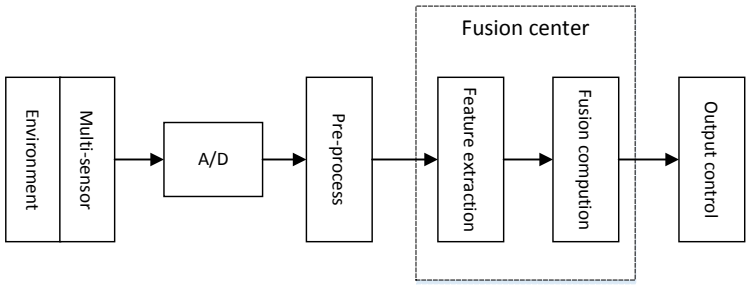

Fig. 1 The process of fusion system

2.3 Key technology

Data fusion is the key technology of data conversion, data association, database and fusion calculation, which is the core technology of multi-sensor data fusion system, the specific introduction of each key technology is as follows:

- **Data conversion.** Data conversion not only to convert the information of different levels, but also need to be converted to the environment or the objective description of the difference and similarities. Even if the same level of information, there are different descriptions.
- **Data association technology.** In the process of data association, the core problem is to overcome the uncertainty of sensor measurement and interference caused by the relevance of the two meaning, in order to maintain the consistency of data.
- **Situation database.** Situation database is divided into real - time database and non - real - time database, which requires large capacity, quick search, good open connection and good user interface.
- **Fusion calculation.**

3. The problem of data fusion system

Data fusion technology has been successfully applied to a variety of scenarios, and it has played an important role, but with increase in the sensor device and the improvement of sensor performance, different scenarios put forward more demand for the data fusion system design. As a new subject, there are still a lot of problems [3] to be solved and improved in the theoretical research level and engineering practice level. We will analyze the current problems faced by the system from a few angles.

3.1 Data size

In the current background of rapid development of information technology and networking technology, more and more number of sensors in different scenarios, sensor types are more and more, and the performance of the sensor has been greatly improved compared to the previous day. So now the data acquisition system of the received data into the explosive growth trend, the scale of the data from the past every "GB" level for the current "TB" or "PB" level, so the current data fusion system is facing the problems of massive data storage and rapid retrieval, how can the massive data storage effectively supporting the source constantly, how to meet the real-time demand of various queries based on mass on the data, is an very important things for current system designers to think about.

3.2 Fusion dimension

Now, heterogeneous sensor number of kinds of rapid growth, perhaps a generic application scene contains the 10 to 20 kinds of sensors and different types of data generated by these sensors are not the same, there are a lot of structured data, such as temperature, humidity data, a lot of non- structured data, such as video surveillance, high-definition camera, and so on. In the face of such a rich variety of data, how to use, how to for users provide more effective scene portrait, require data analysts to make more reasonable choices based on specific occasions and the users' actual demand.

3.3 Fusion algorithm

Faced with massive data, the integration of the system needs to do three things: how to extract useful information from large scale data, how to make data association more effective with the extracted information; and how to ensure the real-time nature of data association in large scale data.

So the fusion system put forward higher request to the existing data mining algorithm and the fusion algorithm in the performance or the final effect:

- The hidden feature extraction, which requires the design algorithm has more effective learning ability and feature extraction, and relates to the algorithm actually contains the image processing technology, speech recognition, natural language processing technology;
- The search and combination of historical information;
- The combination of heterogeneous data, such as pictures, video and so on.

4. Architecture design of data fusion system based on big data technology

4.1 Design proposal and principles

There are two very important starting points about why to use big data technology to reconstruct the existing data fusion system.

The scene meets the features of big data, as shown in the figure 2, big data has four important features: Volume, Variety, Velocity, and Value.

Problems encountered in the scene is very suitable for big data technology to solve, big data technology is designed for under the condition of affordable by very fast (velocity) acquisition, discovery and analysis, from the large volume, multi category (variety) data to extract valuable information (value). And it will be a new generation of technology and architecture in the field of IT.

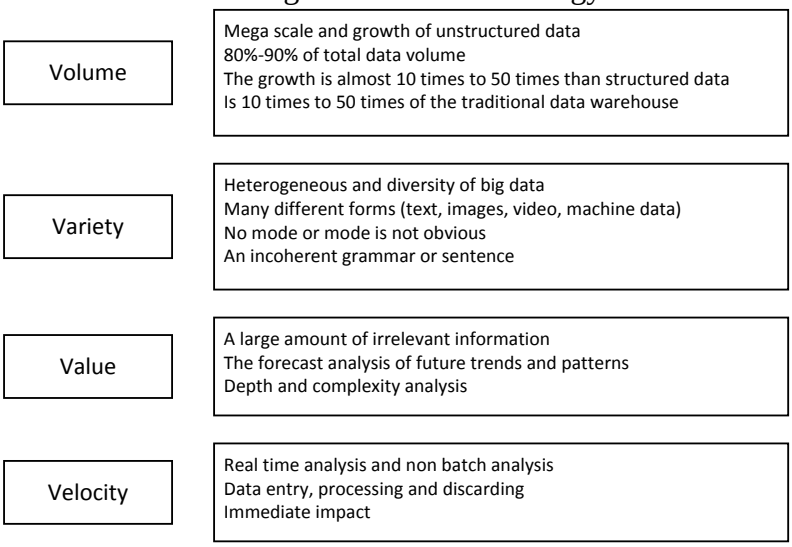

Fig. 2 The characteristics of big data

The goal of data fusion is to help senior executives make more accurate decisions, and to provide the decision makers with a full range, multi means of target association information, so the design needs to obey the following principles [4].

Data is the next "Inside Intel", so the acquisition and storage should be considered together. The core value of data fusion system is distributed storage and analysis of massive data, fusion system has continued acceptance of heterogeneous sensor networks collecting various types of data, these data is the key to support fusion, so detection of large data storage presents higher technical requirements, next generation system must have good design of big data infrastructure to ensure secure and flexible storage.

Scalable data fusion model and algorithm modules are expected to meet the changing needs. Data driven business characteristics will make data analysis be more important, and the key is the fusion algorithm. Designing scalable algorithm modules, especially for large data parallel algorithms is needed.

Due to the current Internet domain is data driven and the fastest area of technology updating, so during the architecture design and technology selection, we mainly refer to the some excellent architecture and combine the current research needs with some of the current relatively mature big data technology framework, and finally we got a set of solutions.

4.2 Overall technical architecture

Large data processing including data source, data access, data cleaning, data cache, computing and storage, data service, consumption data and so on, each link is highly available and scalable. Input of the previous link is as the output of the next phase of data processing, and the data flow process is monitored by data quality control system. Abnormal data will trigger an automatic warning, and alarm service. Data flow architecture based on big data technology is shown in the figure 3.

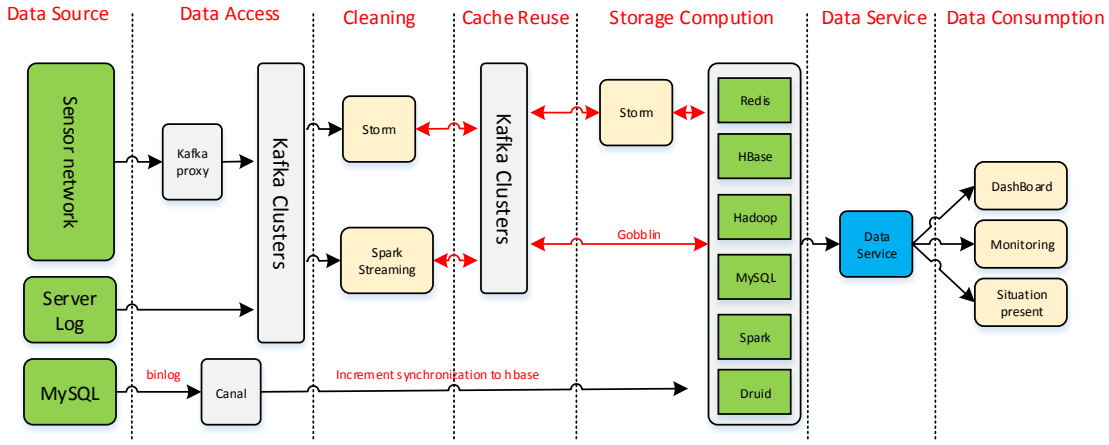

Fig. 3 Data flow architecture

According to the analysis of data flow architecture, we can get a data fusion system architecture based on big data technology. As shown in the figure 4, each layer can be a service component and each service component is scalable.

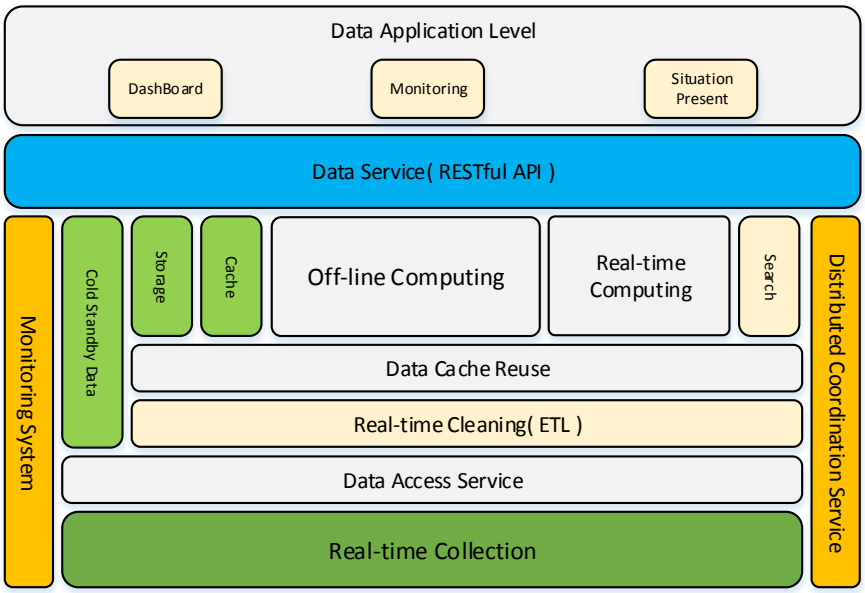

Fig. 4 Data fusion system architecture

4.3 Main components

The core of the system architecture is divided into three parts: data access services, data computing services, data storage services. We will give a description about them in the following part in detail.

Data access service is the entrance of the whole fusion system, as shown in the figure 5, the module shields off the complexity of the underlying heterogeneous sensor network data, and providing a unified data access service. All raw data are unified distributed to the protocol stack, and the data which is parsed will be pushed to the specific data subject pipeline in pub-sub system.

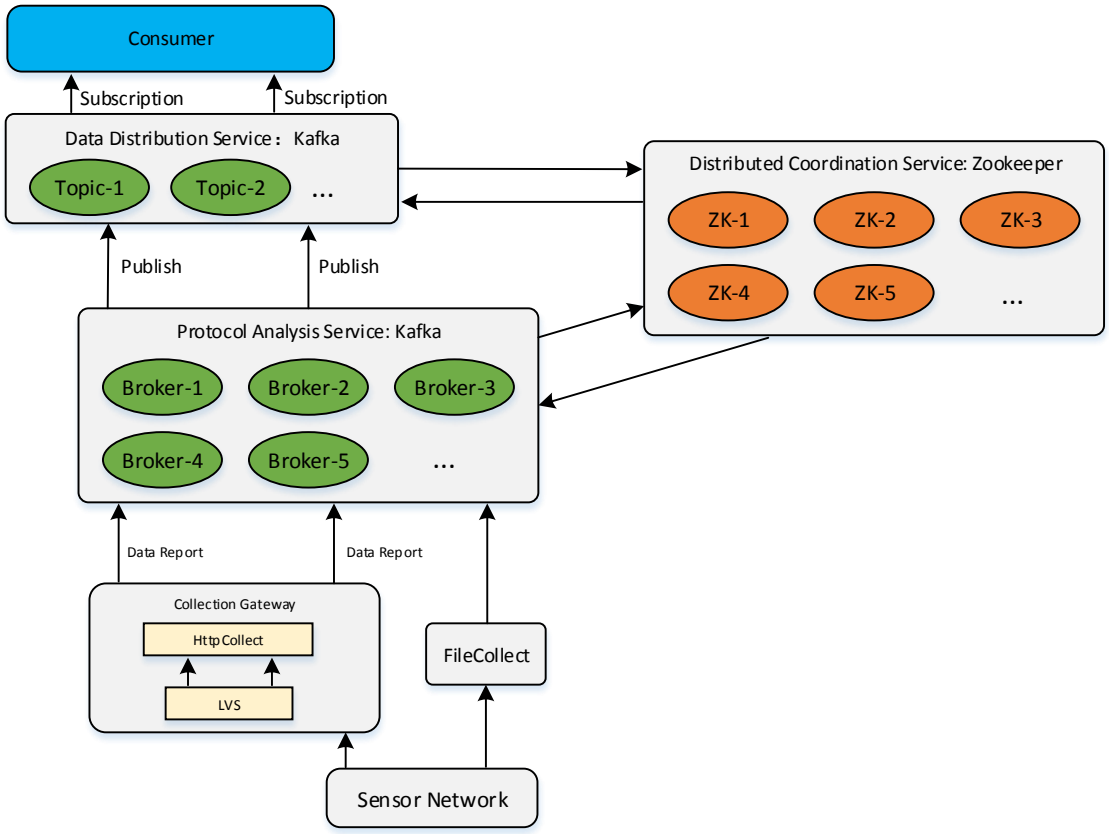

Fig. 5 Data access service

Data computing services is strictly divided into two blocks, as shown in the figure 6, one block is about real-time computing, and another is about off-line computing. We use spark as the main computing framework. This way in the overall performance is higher than the hive by 5-10 times. We also use Hive or Spark framework to do something for the part of the complex operations; off-line calculation is mainly used for target feature mining, extract available information extracting from the historical database, while real-time computing platform based on storm framework, is mainly used in data cleaning, real-time track correlation, real-time target found and real-time situational statistics etc.

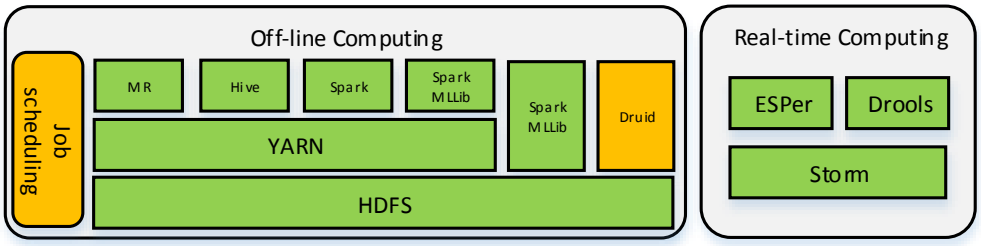

Fig. 6 Data computing service

Data storage service includes two major aspects: data storage and data cache. Data storage service construction is based on the data storage model [5] of big data platform as shown in figure 7.

Data storage model is divided into 7 parts: data cache layer (DCL), data detail layer (DDL), common data layer (common), data summary layer (DSL), data application layer (DAL), data analysis layer (Analysis), temporary data layer (Temp). Each part will be described in detail below:

- Data cache layer: storing the original data collected by the sensor network.
- Data detail layer: storing the detail data that is pre-processed in the data cache layer.
- Common data layer: storing the two-dimensional tables and business system data.
- Data Summary Layer: storing the target association data, target feature data and so on.
- Data Application Layer: storing the target analysis, situation analysis and warning analysis data.

- Data analysis layer: storing the result data of model calculations, these results are used to support data fusion and data mining.
- Temporary data layer: storing the temporary data produced by the modeling calculation, quality verification and so on.

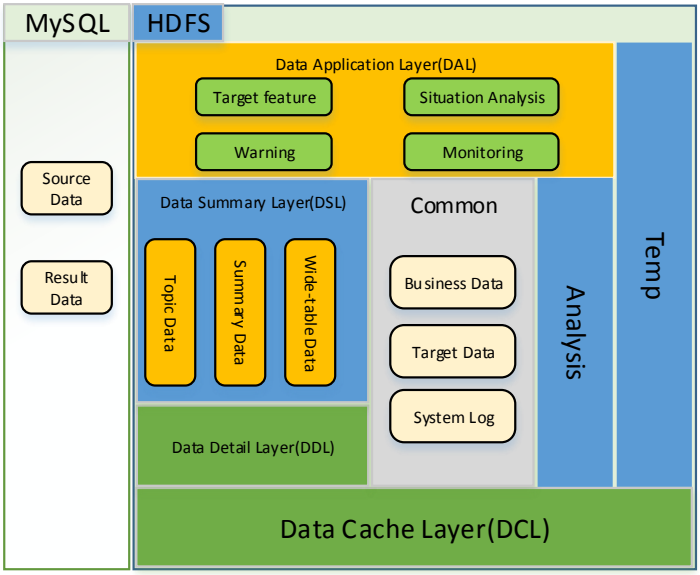

Fig. 7 Data storage model

5. Conclusion

This paper begins with an analysis of the data fusion system, the development status quo, key techniques and problems currently facing. Then, we make an analysis combined with the characteristics of big data about what the design principles should the new data fusion system follow. Finally, we focus on data access, data computing and data storage in three aspects to explain in detail the design of the system architecture.

Acknowledgement

This paper is supported by the National Natural Science Foundation of China under Grant No. (61372115, 61132001), "973" program of National Basic Research Program of China Grant No. 2012CB315802, and National High-tech R&D Program of China (863 Program) under Grant No. 2013AA102301.

References

[1] Hall D L, Llinas J. An introduction to multisensory data fusion [J]. Proceedings of the IEEE, 1997, 85(1): 6-23

[2] PANG M, ZHU W. Application of data fusion based on RBF neural networks in waste gas data processing [J]. Transducer and Microsystem Technologies, 2007, 4: 030.

[3] Dong H, Evans D. Data-Fusion Techniques and Its Application[C]. Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007). 2007.

[4] Feng X, Ming-Hua Y U, Xiao-Ling M A, et al. Learning analytics system architecture based on big data technologies [J]. Journal of East China Normal University, 2014, 31(2): 20-29.

[5] Jian HOU, Shuai R. Wen HOU. Massive Data Storage Model based on Cloud Computing [J]. Communications Technology, 2011, 5: 056.