On the Application of Big Data and its Intelligent Processing Technology in Internet of Things Industry

Bin Huang\textsuperscript{1,2,*}

\textsuperscript{1}Harbin Engineering University, Harbin, 150001, China
\textsuperscript{2}CSIC Information Technology Co., Ltd, Beijing, 100089, China

*Corresponding author e-mail: huangbin0501@hrbeu.edu.cn

Abstract. Big data and the Internet of Things are developing at a rapid pace, bringing great opportunities for the application of various technology fields, including smart cities, smart transportation, smart grid and telemedicine consultation. The wide application of the Internet of Things makes big data analysis challenging, as data is collected and processed by different sensors \cite{1}. Based on this, this paper explores the application of big data and its intelligent processing technology in the Internet of Things industry.

Keywords: Big Data Technology, Internet of Things Industry, Applied Analysis

1. Introduction
Through the Internet of Things, machines, equipment and personnel can be centrally managed and controlled, but also remote control of home equipment, remote control of cars, and even search for locations and prevent stolen items \cite{2}. By collecting data from these small things and aggregating it into big data, and even redesigning roads to reduce car accidents, disaster prediction and crime prevention, and epidemic control, we can connect things to things \cite{3}.

2. Characteristics of big data and related techniques
Table 1. Specific description of big data features

| Character | Explain                                |
|-----------|----------------------------------------|
| Volume    | (1) Big data, (2) Broad data sources for big data, (3) Data is often distributed |
| Variety   | Big data comes from a wide variety of sources and analytical methods |
| Velocity  | Online or real-time analysis requires a short period of time for rapid access, processing, analysis and reporting of massive data |
| Value     | The value density is low, once the information needed is obtained from the massive data analysis, the value will often be very high |
| Veracity  | (1) The whole life cycle assurance of big data is real, its data source is trusted, the processing process is based on trusted computing, the storage is protected and the data is authentic; (2) Big data should be used with authorization and access control, and its modifications can be traced back to |
| Vulnerable| (1) The big data analysis structure has high value and is easy to stimulate the attacker's profit-seeking nature and launch an attack; (2) Big data is not mature enough, and there are still many loopholes in data acquisition, storage, processing and management, which are vulnerable to attack; (3) Attackers can use big data to launch large-scale sustained attacks, which are more destructive |

Table 2. Big data related technical methods

| Type/level of analysis | Usage mode                              | Existing architecture/tools       | Advantage / use of attribution type |
|------------------------|-----------------------------------------|-----------------------------------|------------------------------------|
| Actual time            | Analysis of data collected from sensors | Greenplum HANA                    | Collateral execution               |
| Off-line               | When a quick response is not required   | Scribe Kafka Timetunnel Chukwa    | Efficient data collection and cost reduction |
| Memory level           | Application when data size is smaller than cluster memory | MongoDB                           | Real-time analysis                 |
| Business Intelligence Level | When the size of the data is larger than the memory level | Data analysis methodology         | Real-time and offline              |
| Large scale            | As the amount of data exceeds the full capacity of BI analytical products and traditional databases | MapReduce | Off-line |

3. Big data intelligent processing technology

In the processing of big data, the key role is the big data framework, through the big data system framework, to achieve the integration of large-scale data processing. From manual statistical analysis to computer, to today's distributed computing platform, the rapid improvement of data processing speed is the continuous evolution of the overall architecture. Nowadays, there are many big data frameworks available on the market, the most popular is Hadoop, Spark and Storm, Hadoop are the mainstream, but the two rising stars of Spark and Storm are also developing rapidly [4,5].

3.1. Hadoop

When it comes to big data, the first thing to think about is Hadoop, because Hadoop is the most widely used big data tool in the world. It has good cross-platform and can be deployed in cheap computer clusters. It is widely used in the industry. It is synonymous with distributed computing architecture. With extremely high fault tolerance and very low hardware prices, the big data market is booming. Nearly all mainstream vendors develop and deliver services around Hadoop, such as Google, Baidu, Cisco, Huawei, Alibaba and Microsoft, which support Hadoop. Hadoop has become a huge ecosystem so far and has implemented a large number of algorithms and components. HDFS, MapReduce and Yarn. are the main individual components of the Hadoop framework HDFS is responsible for data
storage, MapReduce data calculation and resource scheduling in big data processing [6]. Based on these three core components, Hadoop can achieve efficient processing of large-scale data. Hadoop excellent fault handling mechanism, support high scalability, fault tolerance, high availability, more suitable for big data platform research and development. However Hadoop there is a big limitation that the processing of data is mainly offline processing, which has a set of large-scale offline data processing, but for the data processing task with high timeliness requirements, it can not be completed very well.

Hadoop architecture greatly improves the computing storage performance and reduces the hardware input cost of the computing platform. However, because the calculation process is placed on the hard disk and limited by the hardware conditions, the speed of data throughput and processing is obviously not as fast as using memory, especially when using Hadoop for iterative calculation, it is very resource-consuming. And in the development process need to write a lot of relatively underlying code, not efficient [7,8].

3.2. Spark

On the basis of the limitation of Hadoop in real-time data processing, the Spark and Storm framework emerges as the times require, and the batch processing framework with improved data flow processing is realized through memory calculation. Based on the Hadoop architecture, it makes up for the deficiency of Hadoop in real-time data processing. To make the program run faster, Spark provides memory computing and reduces the I/O overhead of iterative computing. Spark not only has the advantages of Hadoop MapReduce, but also solves its defects, and gradually becomes the most popular computing platform in today's field. As a rising star of big data framework, Spark has more efficient and fast computing power [9].

We know that there are four main computing modes, except for the special type of graph computing, the other three are sufficient to cope with most application scenarios, because these are the three main types of processing in practical applications: complex batch data processing, interactive query based on historical data and data processing based on real-time data flow. Hadoop MapReduce are mainly used for computing, Hive and Impala for interactive queries, Storm mainly for streaming data processing. All of these can only be used for one application, but if there are three application requirements at the same time, Spark is more appropriate. Because the Spark's design concept is "a software stack meets different application scenarios", it has a complete ecosystem that can provide both a memory computing framework and a variety of types of computing (which can support both streaming computing and interactive queries), providing one-stop solutions [10].

Moreover, Spark can be well compatible with Hadoop ecosystems, Hadoop applications can be easily migrated to Spark platforms. the main functional components of data storage, in addition to Hadoop HDFS or Amazon S3, include Spark Core( basic general functions that can perform complex batch computing), Spark SQL( support interactive query computing based on historical data), Spark Streaming( support real-time streaming computing), providing common machine learning, supporting data mining based on historical data), and GraphX( support graph computing) [11].

Though Spark has many advantages, it can not completely replace Hadoop, but mainly replace MapReduce computational model. Spark do not have tens of thousands of levels of clusters like Hadoop, so in practical applications, Spark are often used in combination with Hadoop. It can use YARN to realize resource scheduling management and distributed storage with the help of HDFS. Moreover, compared with Hadoop can use a large number of cheap computer clusters for distributed storage computing (low cost), Spark has higher hardware requirements and higher costs.

3.3. Storm

Different from Hadoop batch mode, Storm use a flow computing framework, Twitter open source, Hosted on the GitHub. And like Hadoop, Storm also proposed two computational roles, Spout and Bolt. If Hadoop were a bucket, One bucket in one well at a time, And then Storm is a faucet, It can be opened to continuously produce water. Storm also support many languages, Java、Ruby、Python, etc.
Because Storm is a flow computing framework, it uses memory. This has a great advantage in terms of delay. But Storm don't persist data. But the downside of Storm is, Both offline, high latency. Or interactive queries, and it is not as good as Spark frame. Different mechanisms determine the different scenarios they apply. Like stocks, Stock price changes are not in seconds. As a result, it is suitable to adopt the Spark frame of calculating delay of second level; And in high-frequency trading, High frequency profits are often between one ms, A Storm framework for real-time delay calculation is more suitable [12].

The meaning of Storm for real-time computing is similar to that of Hadoop. It can simply, efficiently and reliably process streaming data and support many languages. It can be integrated with many systems to develop a more powerful real-time computing system. Like the most popular hybrid cloud in the cloud computing market, more and more organizations and individuals use hybrid big data platform architecture because each architecture has its own advantages and disadvantages. Hadoop, for example, its data processing speed and difficulty are far less than Spark and Storm, but because its data can be saved for a long time after the hard disk is cut off, it is also necessary to use it process the data that needs to be stored for a long time. Nevertheless, because Hadoop have very good compatibility, it is also very easy to use in combination with Spark and Storm to meet the differentiated needs of different organizations and individuals. Taking into account the scenarios used in the network security situation, that is, most of them are complex batch data processing (log events) and interactive queries based on historical data, as well as data mining. There will also be some requirements for real-time streaming data processing (such as session flow detection and analysis). It is suggested that its big data platform adopt the construction mode of Hadoop and Spark [13].

The framework of big data processing is constantly updated and optimized. No structure can realize the perfect processing of big data. In the development of real big data platform, it needs to be considered according to the actual needs.

4. The relationship between big data and the internet of things

![Figure 1. The relationship between big data and the Internet of things](image)

5. Application of big data and its intelligent processing technology in internet of things industry

5.1. News media
In this information scale expands sharply, the type is rich and diverse, and accelerates the development, the large-scale each kind of data, causes the human to enter the so-called "big data age ".

The role of big data in news media: fast and accurate automatic tracking, collecting thousands of network media information, expanding news clues, improving the speed of collection; supporting the effective capture of tens of thousands of news every day. The depth and breadth of the monitoring
scope can be set by itself; support the intelligent extraction and audit of the required content; realize
the integration of Internet information content collection, browsing, editing, management and publishing [14].

5.2. Telecommunications
Big data is essential in the field of telecommunications. The application content of big data in the field
of telecommunications is: collecting data of hardware equipment such as base station, analyzing
equipment load condition, generating equipment expansion, optimization, quality checking, expansion
and other suggestions to achieve the purpose of balancing network traffic; analyzing user's single data,
defining user, defining user's attributes, analyzing the characteristics of mobile phone terminal, thus
forming decision-making such as package recommendation and terminal recommendation; making
more comprehensive user behavior analysis and user preference analysis according to the app software
used by users and the visited web pages; Collect social network data such as Weibo, understand user's
evaluation and opinion to operator, public opinion analysis [15].

5.3. Electronic commerce
The role of big data in the field of e-commerce: shopping behavior and sales forecast analysis;
commodity correlation analysis; whole network product information collection, product material
acquisition; through the analysis of product prices and sales, guide the strategy of new products; Build
and maintain cloud comment system; e-commerce channel distribution.

6. Conclusion
Internet of things and big data are interdependent and have a great influence on each other. With the
development of the Internet of Things, it has created a demand for big data capabilities. The increasing
volume of daily data requires more advanced and innovative storage solutions, which in turn requires
updating the organization's big data storage infrastructure. Big data and the Internet of Things are
closely linked. Clearly, these two areas will generate new solutions and opportunities that will have a
long-term impact. Big data and the Internet of Things are playing an important role in technological
advances and transforming our lives to be faster and smarter. The Internet of Things can connect
anything that generates data to the Internet, such as wearables, video games, cars, devices, airplanes,
etc., to collect data. Companies can use big data to better understand their customers' preferences and
behaviors, thereby improving business performance and thus saving time and money.

References
[1] Zhou Xingyuan, on the development of Internet of things under the background of big data,
information construction. 2018,9.20.
[2] Lu Denglong, Zhu Shibaing, big data and its Architecture and key Technologies, Journal of the
Institute of equipment 2(28):87-95.
[3] Chen Hui, Gong Tingyu. Application of big data Technology in Internet of things Industry
[J].A Brief Discussion Jiangxi Communication Technology. 2017(01)
[4] Yu Jing, Wang Hui. Application of big data and its Intelligent processing Technology in
Internet of things Industry [J]. Silicon Valley. 2013(17)
[5] Wang Jiesong. Application of big data and Internet of things Technology in Smart City [J].2
Wireless Internet technology. 2018(03)
[6] Shen Weijie, Ni Hao, Zhang Xingjian, et al. [J]. of Sampling and Sealing Sample Based on
Internet of things Computer Science and Applications ,2019,9(11):2174-2181.
[7] Wan Guohai. Extension Strategy of wood-based panel enterprise based on Internet of things
big data J]. Forest Industry. 2020,47(4):71-73.
[8] Yang Zhongjie, Zhao Yifeng, Wan Zhengxi, et al. Development and Application of Tooling
System Based on Internet of things Technology [J].and Hydropower and pumped
storage. 2019,5(6):16-19,151.
[9] Jiang Wei. Application of big data and Internet of things technology in the construction of intelligent oil field [J]. Science and Wealth 1(28):390.

[10] Lu Xiaohuang. The Application of 5G Communication Technology in the situation of Internet of things [J]. and Digital Communications World, 2020(4):160.

[11] Ren Dong, Dong Xuejian, Cao Reform, etc. Application and Exploration of Internet of things Technology in Statistical data acquisition [J]. Research World, 2020(4):62-65.

[12] Zhang Baoyan. Application of big data in the Internet of things [J]. A Shanxi Electronic Technology, 2019(6):94-96.

[13] Niu Xiaoli. J]. on Internet of things big data Storage and Management Technology Computer programming skills and maintenance 2020(2):67-6871.

[14] Ding Jixiang. Application of Internet of things Technology in Smart City based on big data era [J]. Time farm machinery, 2019, 46(9):54-55.

[15] Wang Yizhong. Application [J] of Internet of things Technology based on big data Encyclopedia Forum Electronic Journal, 2019(8):522.