Mengqi Yang

BIG DATA: ISSUES, CHALLENGES, TOOLS

Thesis
CENTRIA UNIVERSITY OF APPLIED SCIENCES
Information Technology
1/7/2020

Centria
UNIVERSITY OF APPLIED SCIENCES

ABSTRACT

Centria University of Applied Sciences	Date 1/7/2020	Author Mengqi Yang
Degree programme Information Technology		
Name of thesis BIG DATA: ISSUES, CHALLENGES, TOOLS		
Instructor Kauko Kolehmainen		**Pages** 60
Supervisor Kauko Kolehmainen		

This thesis starts with the basic structure of big data, introduces its characteristics, history, function, and importance. This thesis also describes the development prospects of big data. The focus of the thesis is on the analysis techniques of big data, including different data analysis tools and data mining technologies. It gives an example of the important role that the Internet of Things plays in life. Finally, the application of big data in the field of artificial intelligence is described, and the application research of big data in the intelligent transportation systems and healthcare is demonstrated.

Key words
Big data, data analysis tools, data mining, IoT, artificial intelligence.

1 INTRODUCTION

With the rise of the Internet, mobile Internet, sensors, Internet of Things, social networking sites, cloud computing and so on, most aspects of society have been digitized, resulting in a lot of new real-time data that was previously not collected. There is no doubt that people are already in a sea of big data. Since its introduction, the term or concept of big data has been proposed. After several years of development, there are media rumors and rumors. This is a familiar concept. When browsing popular information technology courses in major universities today, big data and other types of existence have firmly occupied the homepage. This science, which has only appeared in recent years, has attracted many researchers. The name of big data is very familiar to people, but people can only know from its name that it is a data processing technology, but its definition and data analysis about big data belong to a typical knowledge blind area. Big data plays an important role in life, but how big data deals with and solves problems is unknown. In the era of big data with the explosive growth of data, the understanding and popularization of big data is the focus of information technology talents.

This article will introduce big data in detail so that people can have a preliminary understanding and experience of the concept of big data ambiguity. The purpose of this thesis is to introduce the characteristics and functions of big data so that readers can have a basic understanding of the vague concept of big data and understand that big data plays an important role in today's information explosion society. Firstly, it introduces the definition, historical development, and usefulness of big data. Then it gives a comprehensive overview of big data analysis, data mining, Internet of Things and brings a real-life case to the readers to feel more specific and closer to life.

2 BASIC THEORY OF BIG DATA

With the increasing popularity of the era of network information, mobile interconnection, social network, and e-commerce have greatly expanded the boundaries and application fields of the Internet. We are in an era of "big data" with the explosive growth of data. Big data has a far-reaching impact on social economy, politics, culture and people's lives. (Beyer 2011.)

2.1 Definition

Whether it is the germination of big data technology proposed by Doug Lenny, an analyst at Meta Group in 2001, or the first time in 2008 that Smith of IBM initially defined the meaning of big data in terms of "BIG DATA". Up to now, the meaning of big data can be summarized as follows: Big data is a technological concept for human beings to recognize the world. It is a technological process of extracting valuable information from massive, complex and scattered data sets utilizing new data analysis and processing methods supported by information technology. Its core is to mine data intelligently and plays its role. Secondary development of big data is the strong point of successful network companies. For example, Facebook creates a highly personalized user experience and a new advertising model by combining a large amount of user information. It's no coincidence that this business practice of creating new products and services through big data, Google, Yahoo, Amazon, and Facebook are all innovators in the Big Data era. (Beyer 2011.)

In today's era full of digital data, data processing has become easier and faster. People can process thousands of huge amounts of data in an instant. To understand the content of information and discover the relationship between information and information in data, human beings have never had such a profound understanding of data today. We should re-recognize the characteristics of data: Volume. At present, any single device is difficult to directly store, manage and use the amount of data. The "big" in big data also includes the comprehensiveness of data; Fast data flow and dynamic data change (Velocity). Data changes over time and environment; Variety. Data describing the characteristics or laws of a thing exists in many forms, and then a kind of emerging concept veracity. In short, big data refers to increasingly large and increasingly complex data sets, especially data sets from new data sources. The sheer size of traditional data processing software can help us solve past business puzzles. (Beyer 2011.)

2.2 3V characteristics of big data

Big data, in a narrow sense, can be defined as a collection of large amounts of data that are difficult to manage with existing general technologies. The reason why big data is difficult to manage can be described as Volume, Variety, and Velocity by 3V. Broadly speaking, big data can be defined as data that is difficult to manage because of its 3V characteristics, the technology of storing, processing and analyzing these data, and the concept of talents and organizations that can gain practical significance and viewpoints by analyzing these data. (Pinal 2013.)

2.2.1 Volume

The "big" of big data is first reflected in the amount of data. In the field of big data, you need to deal with a large amount of low-density unstructured data, whose value may be unknown, such as Twitter data flow, web page or mobile app click flow, and data captured by device sensors, etc. In practical application, the data volume of big data is usually as high as tens of TB or even hundreds of PB. (Pinal 2013.)

2.2.2 Velocity

High speed describes the speed at which data is created and processed. In the high-speed network era, creating real-time data streams has become a popular trend through high-speed computer processors and servers based on software performance optimization. Not only do companies need to understand how to create data quickly, but they must also know how to quickly process, analyze, and return to users to meet their real-time needs. According to IMS Research's survey of data creation speed, it is predicted that there will be 22 billion Internet-connected devices worldwide by 2020. (Pinal 2013)

2.2.3 Variability

Big data has a multi-layered structure, which means that big data can take on a variety of forms and types. A wide range of data sources determines the diversity of big data forms. Any form of data can play a role. At present, the most widely used recommendation system is Taobao, NetEase cloud music, etc. These platforms will analyze the user's log data and further recommend what users like. Log data is

structured obviously, and some data are not structured obviously, such as pictures, audio, video, etc. These data have weak causality, so it needs to be labeled manually. (Pinal 2013.)

2.2.4 Veracity

In addition to the three V's, some add a fourth to the big data definition. Veracity is an indication of data integrity and the ability for an organization to trust the data and be able to confidently use it to make crucial decisions. The importance of data lies in the support of decision-making. The scale of data does not determine whether it can help decision-making. The authenticity and quality of data are the most important factors for obtaining model knowledge and ideas and the most solid foundation for making successful decisions. The pursuit of high data quality is an important big data requirement and challenge. Even the best data cleansing methods cannot eliminate the inherent unpredictability of certain data, such as human feelings and honesty, weather conditions, economic factors, and the future. (Villanova University 2014.)

Data in the era of big data also presents three other characteristics. A kind of the first feature is that there are many types of data. Including network logs, audio, video, pictures, geographic location information and so on, multi-type data put forward higher requirements for data processing ability. A kind of the second feature is that the data value density is relatively low. For example, with the wide application of the Internet of Things, information perception is ubiquitous, and information is massive, but the value density is low. How to accomplish the value "purification" of data more quickly through powerful machine algorithms is an urgent problem to be solved in the era of big data. A kind of third feature is fast processing speed and high timeliness requirements. This is the most prominent feature that distinguishes big data from traditional data mining. (YonghongTech 2017.)

2.3 Historical development

Although the concept of big data was recently proposed, the origins of big data sets can be traced back to the 1960s-70s. At the time, the data world was still in its infancy, and the world's first data centers and the first relational database appeared in that era. American statistician Hermann Hollers invented a motor to count the number of holes in the card to count the 1890 census data. The equipment allowed

the United States to complete the eight-year census in a year. The annual census activity has triggered a new era of data processing on a global scale. (Bernard 2015.)
Around 2005, people began to realize that users generated massive amounts of data when using Facebook, YouTube, and other online services. In the same year, Hadoop, an open-source framework developed for storing and analyzing data sets, was introduced, and NoSQL began to spread in the same period. The Hadoop project was born in 2005. Hadoop was originally a project that Yahoo! used to solve web search problems. Later, due to its technical efficiency, Hadoop was introduced by the Apache Software Foundation and became an open-source application. (Bernard 2015.)

At the end of 2008, “Big Data” was recognized by some well-known computer science researchers in the United States. The industry organization Computing Community Consortium published an influential whitepaper: “Big Data Computing: In Business, Science and Social Fields create a revolutionary breakthrough. It makes people's thinking not only limited to the machine of data processing but also suggests that what is important for big data is new uses and new insights, not the data itself. The organization can be said to be the first institution to propose the concept of big data. (Bernard 2015.)

In 2009 the government of India set up a biometric database for identity management, and the UN's global pulse program has studied how to use data from mobile phones and social networking sites to analyze and predict problems from spiraling prices to disease outbreaks. (Bernard 2015.)

In mid-2009, the U.S. government further opened the door to data by launching http://Data.gov, a website that provides the public with a wide variety of government data. The site's data set of more than 44,500 is used to ensure that websites and smartphone apps keep track of everything from flights to product recalls to unemployment rates in specific areas, a move that has inspired governments from Kenya to Britain to launch similar initiatives. (Bernard 2015.)

In February 2010, Kenneth library Kerr on the economist published for 14 pages of big data project report data, “ubiquitous data”. "The world has an incredible amount of digital information and it is growing at an incredible rate," Kuker said in the report. The impact of this huge amount of information has been felt in many fields, from the economy to the scientific community, from the government to the arts. Scientists and computer engineers have coined a new word for the phenomenon: "big data." As a result, Kuker became one of the first data scientists to see the trend of the big data era. (Bernard 2015.)

In May 2011, the McKinsey&Company global institute (MGI) released a report named "big data: the next new field of innovation, competition, and productivity", which attracted much attention. This is the

first comprehensive introduction and the prospect of big data for professional institutions. The report points out that big data has penetrated every industry and business functional area and become an important production factor. The report also notes that "big data" stems from a dramatic increase in the ability and speed of data production and collection -- and the ability to generate, transmit, share and access data as more people, devices and sensors are connected through digital networks. (Bernard 2015.)

The emergence of open-source frameworks such as Hadoop and later Spark is of great significance for the development of big data because they reduce the cost of data storage and make big data easier to use. In the following years, the number of Big Data has further exploded. Today, "users" around the world -- not just people but also machines -- continue to generate vast amounts of data. (Bernard 2015.)

With the rise of the Internet of things (IoT), more and more devices are now connected to the Internet and they collect a large amount of data about customers' usage patterns and product performance, and the emergence of machine learning further accelerates the growth of data volume. However, although it has been around for a long time, the use of big data has only just begun. Today, cloud computing further unleashes the potential of big data. (Bernard 2015.)

2.4 The significance and importance

The research and analysis and application of big data are of great significance and value. In his book "The age of big data", Victor Maier Schoenberg, known as the prophet of the big data era, listed many detailed big data application cases, analyzed and predicted the development status and future trend of big data, and put forward many important views and development ideas. He believes that "big data has ushered in a major transformation of The Times", pointing out that big data will bring huge changes, changing the way we live, work and think, changing our business model and affecting all aspects of our economy, politics, technology, and society. (Schoenberg 2012.)

As the application demand for big data industry is growing day by day, more and more research and application fields in the future will need to use big data-parallel computing technology, which will permeate every application field involving large-scale data and complex computing. Not only that, for the center with big data processing computing technology will be a revolutionary influence on the traditional computing technology widespread impact computer architecture, operating system, database, programming technology, programming technology, and method, software engineering technology, multimedia information processing technology of artificial intelligence and other computer application technology,

and combined with the traditional technique of computer-generated a lot of new research hotspot and subject. Big data has brought many new challenges to traditional computer technology. (Schoenberg 2012.)

As a student of information technology, I think the existence of big data has its unique significance, I can use it to achieve a lot of industrial practice. One of the great benefits of technological progress is that it saves labor costs. When people are freed from repetitive manual labor, they have more time to think and develop new technologies. Some actual big data industrial projects that help people to liberate the labor force, such as figure 1: Google search engine autocomplete. Build n-gram Library through the Wiki dataset to realize the automatic completion function of the search engine. Help you quickly find what you want from a huge search library. People bid farewell to the library card retrieval, into the era of machine retrieval. (Sullivan 2018.)

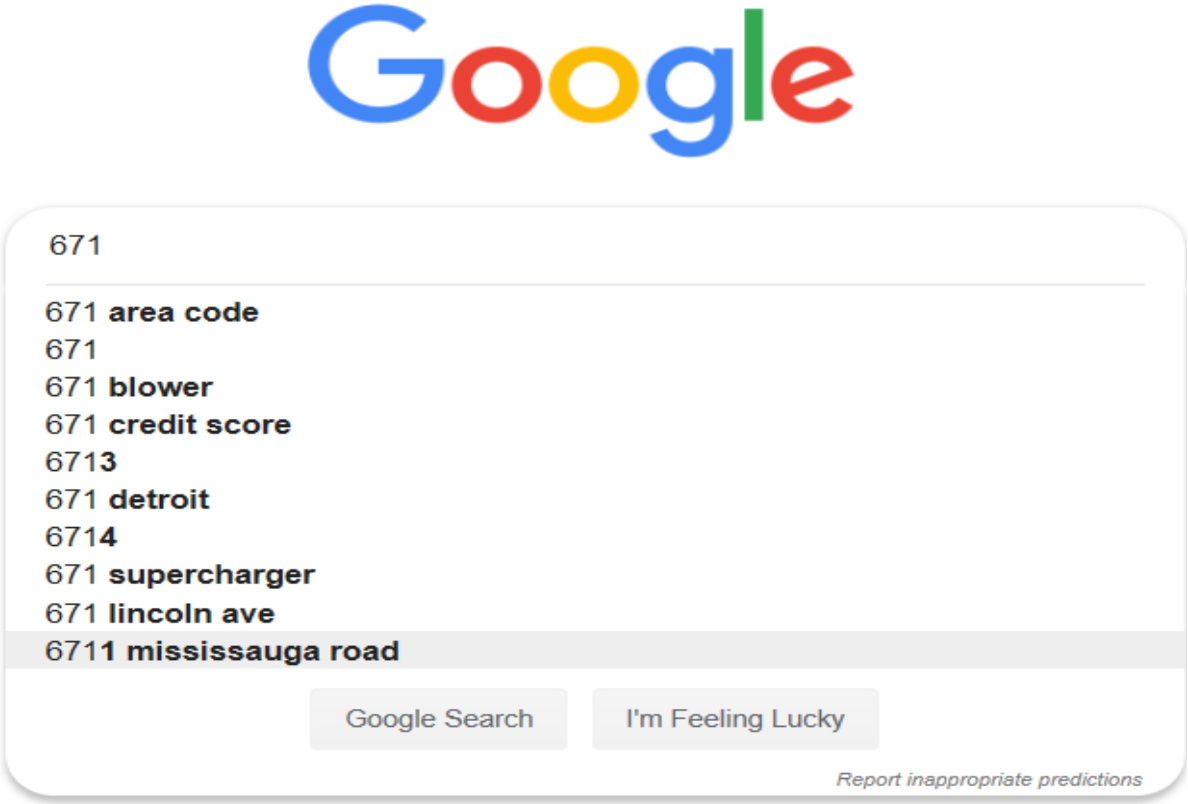


Figure 17. Google search engine autocomplete (Sullivan 2018.)

Through the statistical article emotion keywords analysis article expression emotion. Text sentiment analysis: also known as opinion mining, orientation analysis, etc. In simple terms, it is the process of analyzing, processing, summarizing, and reasoning subjective texts with emotions. The Internet (such as blogs and forums and social service networks such as public commentary) generates a large amount of user-involved valuable commentary information such as people, events, products, and the like. These commentary messages express people's various emotional colors and emotional tendencies, such as hi,

anger, sadness, joy and criticism, praise and so on. Based on this, potential users can view the opinions of public opinion on an event or product by browsing these subjective comments. (Sullivan 2018.)

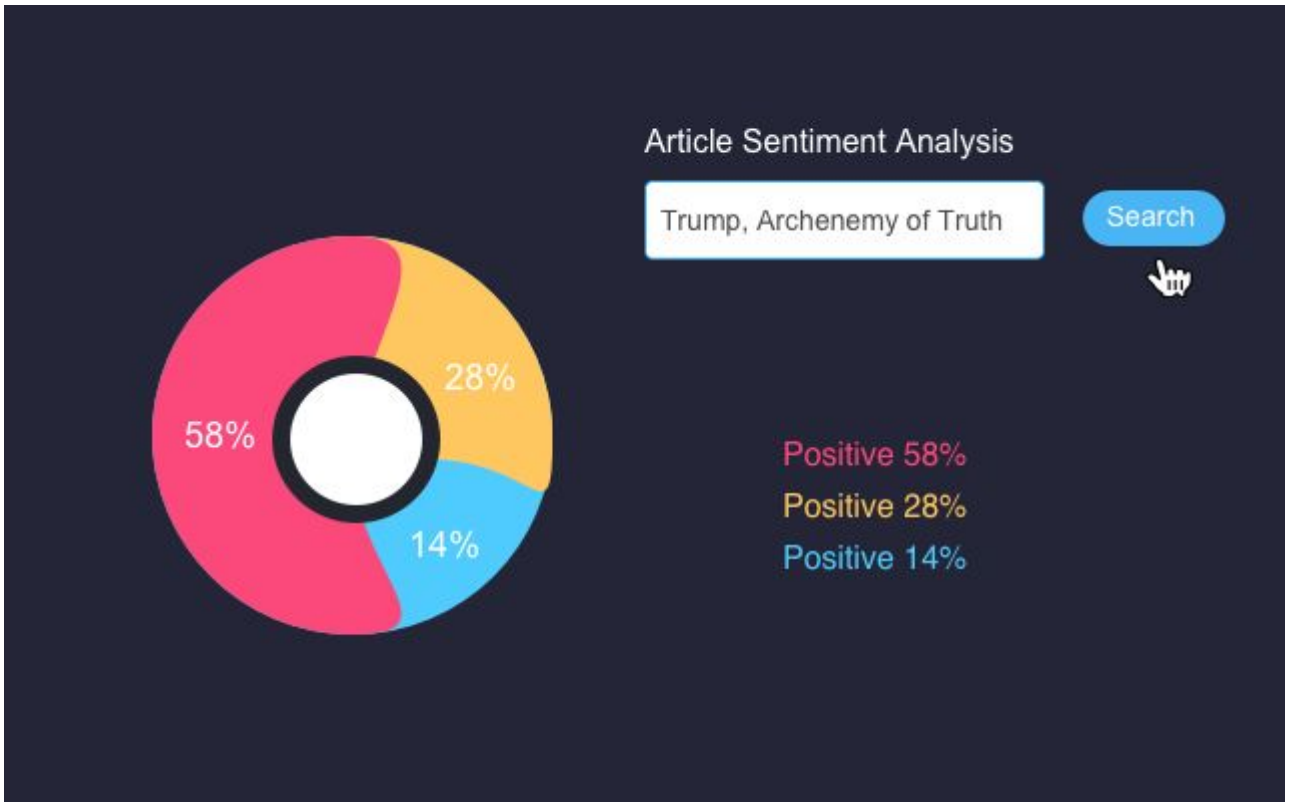


Figure 18. Article Sentiment Analysis (Sullivan 2018.)

Movie recommendation system: Using algorithms, Netflix can tailor the entire user experience to the needs of each user, including the home page, title, visuals for each movie, and more. Netflix recommends content that users might be interested in by collecting each user's rating for each video (scores 1-5). However, as Netflix has more user behavior data (including user-viewed content, device usage, viewing time, viewing frequency, viewing location), machine learning is used to build recommendation algorithms to capture more rule-based algorithms that may leak. Information that is useful for predicting preferences, such as the order in which the videos are viewed, and the interaction between different factors. (Ziyan 2018.)

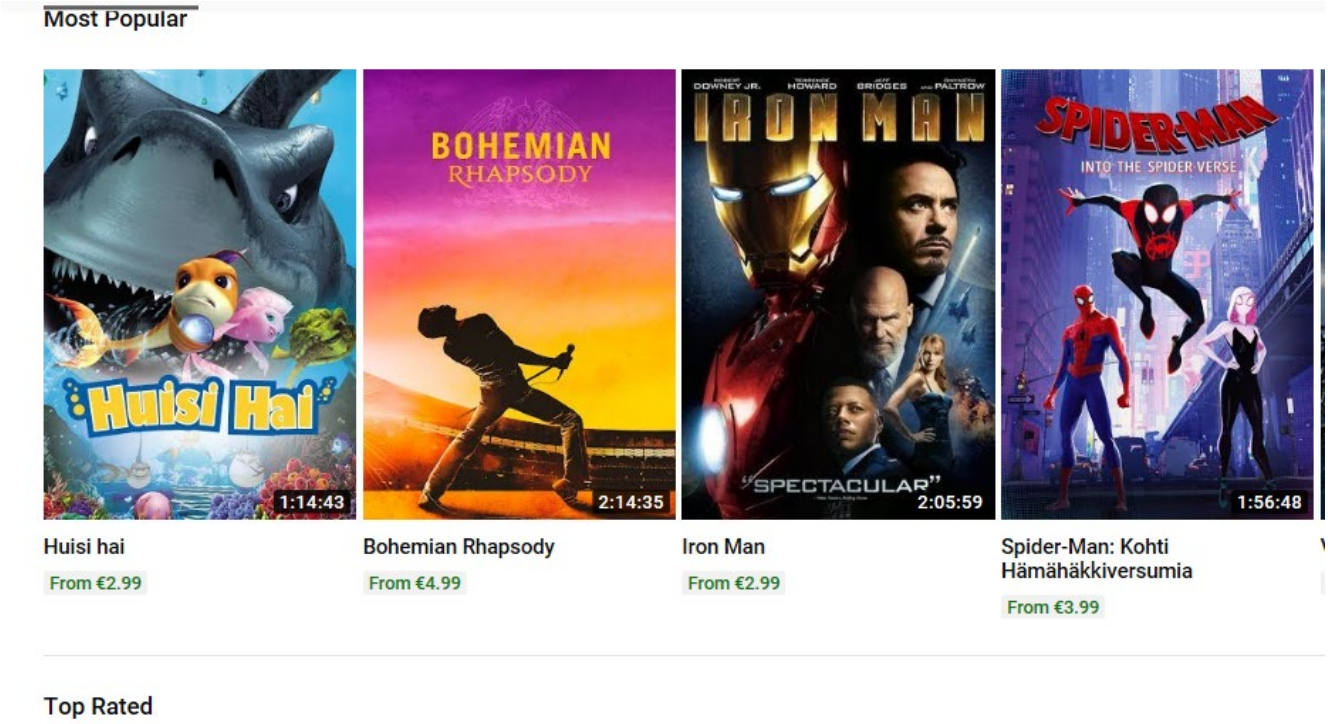


Figure 19. Movie recommendation (Windows store 2019.)

The following are the recommended entries for the Netflix homepage. The algorithms behind them can be summarized into two categories: Content-based filtering and Collaborative filtering. The content filtering algorithm finds similar videos and recommends them to users according to the characteristics of the film itself. This is what most video websites can do now to recommend similar works to users. (Ziyan 2018.)

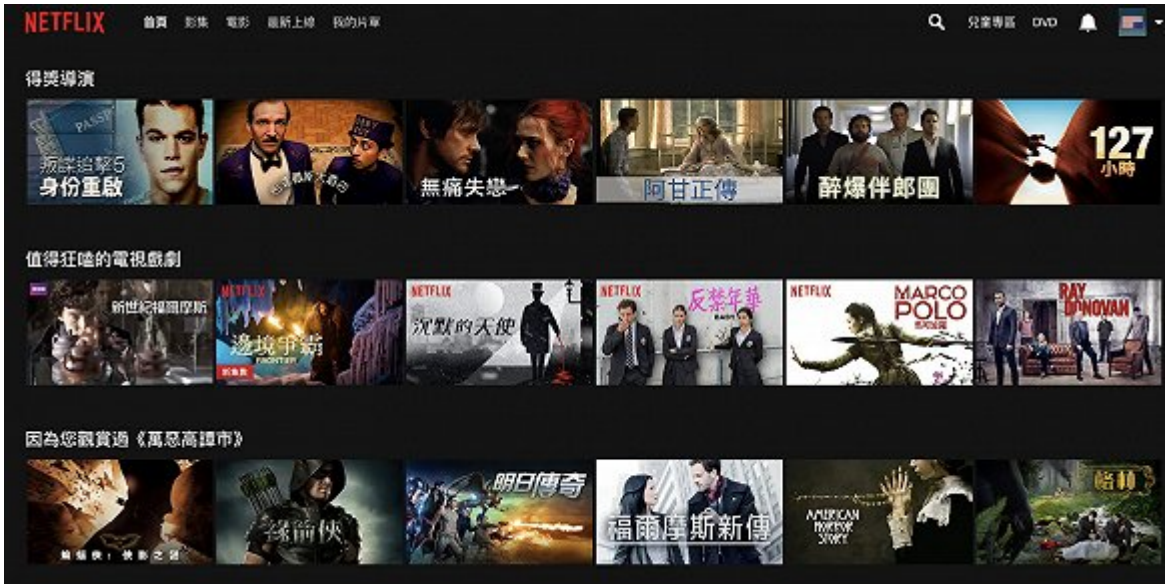


Figure 20. Netflix's algorithm-based unique home page for each user (Ziyan 2018.)

The latter is to find out similar users, so that user A may like the videos that user B has seen. Specific to Netflix, content recommendation based on this algorithm is provided by the assumption that similar

browsing patterns represent similar user tastes. The concept of “Taste Communities” plays an important role in these recommendation algorithms. "The same taste user group" is a group of users who like to see the same content. Netflix has identified 2,000 such as user groups. (Ziyan 2018.)

Jason	Yes	Yes	Yes	Yes
Andi	No	Yes	No	Yes
Sarah	Yes	No	Yes	No
Sam	No	No	Yes	Yes
Scaz	Yes	??	??	??

Figure 21. Simple concept recognition of content-based filtering (Ziyan 2018.)

We live in a world that is constantly generating massive amounts of data. According to IBM, more than 90 percent of the existing data was generated in the last two years. IBM estimates that there are now millions of terabytes of data per day -- the equivalent of about a billion gigabytes of data every 24 hours, or a whole bunch of hard drives every day. Data can be generated in a variety of ways. Every keystroke, every purchase, and every sensor activation generate data. And the data is not simply numerical. The data can also be extracted from comments, tweets, and photo-sharing posted on social media, turning consumers' emotions into useful information that can be quantified and measured. Each isolated piece of data by itself is of little value. But when linked to other data, seemingly worthless data becomes valuable insights into human behavior and the business environment of companies themselves and their competitors. In business, understanding what big data means is more important than defining it. Big data is not a disruptive technology. Having big data doesn't immediately improve business performance. (Keith 2013.)

Big data is nothing new. Technological advances have spawned vast amounts of data before we feel we can manage or consume it. History abounds with examples. Today, humans are only able to realize the benefits of accumulating large amounts of data by using more and more visualization tools. By moving

over a simple table, QlikView, SAP visualization intelligence, and other tools can provide us with dynamic visualization of big data, presenting the myriad links and interconnections behind the information. This is a critical step in the transformation of unstructured big data into what we call operational business intelligence. (Keith 2013.)

2.5 The future of big data

From the perspective of resources, big data is a new resource, which reflects a new view of resources. Since 1990, the ability to compute, store and transfer data has grown exponentially, driven by Moore's law. Since 2000, represented by the Hadoop distributed storage and computing technology rapid development, greatly improved the Internet enterprise data management capabilities, Internet companies the mining use of "data exhaust" great success, caused the whole society began to re-examine the value of "data", began to treat data as a unique strategic resource. The so-called 3V features of big data are mainly described from this perspective. (CAICT 2016)

From a technical perspective, big data represents a new generation of data management and analysis technology. Traditional data management and analysis technology takes structured data as the management object, carries on the analysis on the small data set, takes the centralized architecture as the main, the cost is high. For multi-source heterogeneous data, on the very large-scale data set is analyzed, mainly distributed architecture of a new generation of data management technology, and open-source software stack tide, while greatly improve processing efficiency, reduces the cost data applications into one hundred times. (CAICT 2016)

Big data opens a new way of thinking. The application of big data endows "seeking truth from facts" with a new connotation. One is "data-driven", that is, operation and management decisions can be driven by data from the bottom up, or even directly made by machines according to data, as in quantitative stock trading, real-time bidding advertising and other scenarios. The second is "data closed loop". When observing big data cases in the Internet industry, they can often construct a complete "data closed-loop" including data collection, modeling analysis, effect evaluation, feedback, and correction, to continuously upgrade themselves and spiral up. At present, many "big data applications" either have insufficient data or do not have to use the new generation technology but embody the thinking of data-driven and data closed-loop and improve the production management efficiency, which is the embodiment of the application of big data thinking concept. (CAICT 2016)

In the future, data may become the largest commodity to be traded. But a large amount of data is not big data. The characteristics of big data are large amounts of data, many kinds of data and maximizing the value of non-standardized data. Therefore, the value of big data is to obtain the greatest value of data through data sharing and cross-multiplexing. Big data will be like infrastructure in the future, with data providers, managers, regulators, and data cross-multiplexing will turn big data into a big industry. (Schroeck & Shockley & Dr. Smart & Professor Dolores & Tufano 2012.)

2.5.1 Development Direction of Big Data

After 20 years, the Internet has undergone tremendous changes. Mobile Internet, social network and e-commerce have greatly expanded the boundaries and application fields of the Internet. To buy goods, people must first browse, compare and enquire; to engage in activities, customers must first solicit, discuss and plan; on the Internet, it is precise because of the large amount of data generated, through the collection and analysis of these data, Internet enterprises can predict the future behavior of human beings in the physical world. The big data technology is the ability to collect and analyze vast amounts of various types of data and quickly acquire information that will affect the future. The sources of big data are very wide. Satellites in the sky, cars on the ground and sensors buried in the soil always generate a large amount of data. When these data are used together, the social value and economic value will be incalculable. (Hao 2014.)

Around data and end-users, there are three main directions for the development of the computer industry: “the first, application software will be pan-Internet. Second, the industry will be vertically integrated.” The closer companies are to end-users, the greater the voice they will have in the industry chain. Third, data will become assets. Pan-Internet is an important channel to collect data. Without pan-Internet application software, companies can hardly obtain user's behavior data. At the data application level, the trend of vertical integration in the industry is to gather many user data, to be closer to users, to understand users better and to provide more appropriate services for them. Data becomes the strategic significance of data more emphasized by assets. (Hao 2014.)

2.5.2 Development Trend

Trend 1 is to become an important strategic resource. In the future, big data will become an important strategic resource at the corporate, social and national levels. Big data will continue to be an important

asset of various institutions and a powerful weapon to enhance the competitiveness of institutions and companies. Enterprises will be fonder of user data, make full use of the data generated by customers interacting with their online products or services, and get value from it. Besides, big data will also play an important role in market impact - affecting advertising, product promotion, and consumer behavior. (Hao 2014.)

Trend 2 is the data privacy standards will be introduced. Big data will face a major challenge of privacy protection. The existing privacy protection laws and technologies are difficult to adapt to the big data environment. It is more and more difficult to protect personal privacy, and there may be paid privacy services. (Hao 2014.)

Trend 3 is the analysis method has changed. Big data analysis will bring about a series of major changes. Like computers and the Internet, big data may be a new technological revolution. Data mining, machine learning, and artificial intelligence-based on big data may change many algorithms and basic theories in small data, which may lead to theoretical breakthroughs. (Hao 2014.)

Trend 4 is a deep integration with cloud computing. Big data processing and cloud technology complement each other. Cloud computing provides a flexible and scalable infrastructure support environment and efficient mode of data services for big data. Big data provides new business value for cloud computing. Therefore, from 2013, big data technology and cloud computing technology inevitably enter a more perfect combination period. Overall, cloud computing, the Internet of Things, mobile Internet and other emerging computing forms are not only the place where big data are generated but also the field where big data analysis methods are needed. (Hao 2014.)

Trend 5is network security issues highlight. The security of big data is worrying, and the protection of big data is becoming more and more important. With the increasing of different data, the physical security of data storage and the security of cloud storage will be higher, which will put forward higher requirements for multiple copies of data and disaster tolerance mechanisms. Internet and digital life make it easier for criminals to get information about people, and there are more criminal means that are not easily tracked and prevented, which may lead to more clever deception. (Hao 2014.)

Trend 6is the birth of big data discipline. Data science will emerge as an emerging discipline related to big data. At the same time, many Monographs on data science will be published. Trend 7: Promote data analysts and other professions. Big data will spawn new jobs, such as data analysts and data scientists.

Data analysis talents with rich experience become scarce resources, and data-driven job opportunities will show explosive growth. (Hao 2014.)

2.6 Where and how big data is used

Throughout history, the emergence of every revolutionary technology will greatly increase social production efficiency. For example, the invention of the wheel has improved the efficiency of human transportation, and the invention of the telegraph telephone has improved the efficiency of human communication. In recent years, the rapid development of big data technology and the integration of various professional fields have become more and more intensive and will inevitably penetrate all walks of life, like metallurgy and printing technology, and comprehensively improve social productivity. (Ding 2019.)

2.6.1 Integration of big data and AI at the application level

Before big data, intelligence was hard to come by machine. The wisdom of each industry depends on experts in each industry. The ability of an expert is closely related to his accumulation of experience and knowledge. The more knowledge he accumulates, the more likely he is to make the right choice. However, the work of human experts is not only inefficient but also inaccurate. Especially in the case of lack of experience or data, experts often rely on intuition to make judgments, which exacerbates the inaccuracy of the results. Recently, the very popular artificial intelligence, which is characterized by many scientific and engineering calculations by computers, is faster and more accurate than the human brain. The core of artificial intelligence is that computers constantly acquire knowledge and learn strategies from experience. When encountering similar problems, they use empirical knowledge to solve problems and accumulate new experiences, just like ordinary people. The more experienced, the better the ability of artificial intelligence to solve problems. Experience is essentially data. When the amount of data is large, it needs to be processed by big data technology. Therefore, artificial intelligence cannot be separated from big data technology. (Ding 2019.)

There are two separate but related technology paths to gaining intelligence from data. One is the big data analysis and mining technology, and the other is artificial intelligence (AI) technology based on machine learning. Both analytical mining technology and machine learning technology rely on massive data for modeling and ultimately output wisdom. From the perspective of the application, these two routes gradually converge and can be evolved and replaced on the technical level. (Wang & Shen 2019.)

For example, in the field of intelligent operation and maintenance in the telecommunications industry, it is currently based on the analysis and mining technology to analyze network data, locate network faults, and form a closed loop of automatic operation and maintenance. In the future, deep learning technology can be used to model network data and locate network faults more accurately. From the perspective of automated operation and maintenance, it does not care whether the underlying technology is analysis and mining or deep learning, but only the improvement of analysis accuracy can be perceived. In the future various application systems, the co-evolution of analysis and mining and AI will be a universal phenomenon. Most application systems based on big data technology are likely to evolve into an AI system in the future. For example, financial anti-money laundering and anti-fraud, intelligent medical treatment and Internet public opinion monitoring can all benefit from the technological progress of big data and AI, providing increasingly intelligent analysis and improving production efficiency. (Wang & Shen 2019.)

2.6.2 Big data and the combination of various fields

In terms of product development, currently, companies like Netflix and Procter & Gamble use big data to predict customer needs. They classify the key attributes of the past and current products or services, model the relationship between these attributes and the successful products of the business, and then build the prediction model of new products and new services. P&G also plans, produces and releases new products based on data and analysis from focus groups, social media, test marketing and pre-launch. (Sravanthi & Reddy 2015.)

Predictive maintenance means tear of various structured data (such as equipment, brand, type, and other information) and unstructured data (including millions of log entries, sensor data, error messages, and engine temperature) are often hidden for prediction of mechanical failure information, through the analysis of these data, the enterprise can identify potential problems before the accident, more economical and efficient to arrange maintenance activities, maximum extend the uptime of components and equipment. (Sravanthi & Reddy 2015.)

Customer experience means the core of today's market competition is to win customers. Companies are in a better position to understand the customer experience clearly than they were in the past. Among them, big data enables you to collect data through social media, website visits, call records and other sources, to improve customer interaction, provide personalized products for customers, reduce customer

churn rate, take the initiative to solve problems, and finally create more value with an excellent experience. (Sravanthi & Reddy 2015.)

Machine learning is a hot topic these days, and data (big data in particular) is one of the big drivers behind this phenomenon. By using big data to train machine learning models, we can "train" machines to specific capabilities without having to program them. (Sravanthi & Reddy 2015.)

Operational efficiency is not usually a hot topic, but big data has the most profound impact in this field. With big data, you can drill down and evaluate production, customer feedback, return rates, and more to reduce shortages, predict future demand, and use big data to make better decisions based on current market needs. (Sravanthi & Reddy 2015.)

Big data is also conducive to promoting innovation. Big data helps you study the interrelationships among people, organizations, entities, and processes to drive innovation in new ways based on deep insights. With the help of big data, you can effectively improve financial and enterprise planning decisions, validate trends and customer needs, better provide customers with new products and services, and implement dynamic pricing to maximize revenue. In short, big data will open the door to the world of innovation and bring you endless possibilities. (Sravanthi & Reddy 2015.)

In the retail industry, technology and means of data analysis are widely used. Traditional enterprises such as Wal-Mart reshape and optimize the supply chain through data mining, and emerging e-commerce companies such as Amazon and Taobao grasp and analyze massive data to provide users with more professional and personalized services. In this era, if enterprises and manufacturers are indifferent to their opinions, they will lose many attention groups, and the influence of the traditional marketing model of communication will be greatly reduced. In terms of personal privacy, large amounts of data often contain some detailed and potential information about us, which gradually raises our concerns about personal privacy. Some big data companies need to take this seriously. (Schroeck & Shockley & Dr. Smart & Professor Dolores & Tufano 2012.)

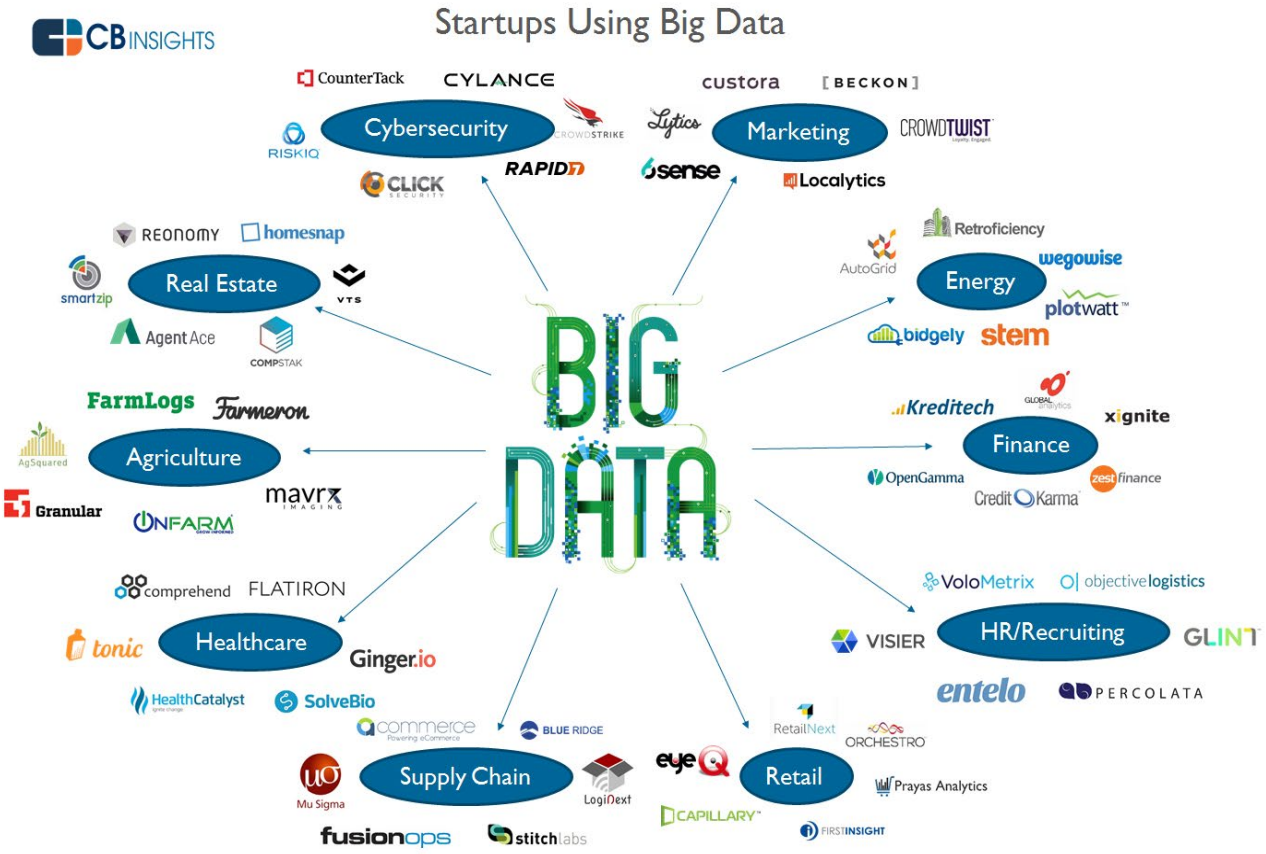


Figure 22. Startups Using Big Data (Schroeck & Shockley & Dr. Smart & Professor Dolores & Tufano 2012.)

3 ABOUT DATA ANALYSIS

Big data is the hottest word in the IT industry. The following business value of big data, such as data warehouse, data security, data analysis, and data mining, has gradually become the focus of profit sought after by industry professionals. With the advent of the era of big data, big data analysis comes into being. The processing and analysis of big data are becoming the node of the new generation of information technology fusion applications. Mobile Internet, Internet of things, social network, digital home, e-commerce and other forms of application of the new generation of information technology, these applications continue to produce big data. Cloud computing provides the storage and computing platform for these massive and diversified big data. Through the management, processing, analysis, and optimization of data from different sources, the results are fed back to the above applications, which will create great economic and social value. (Sriram 2017.)

3.1 Basic Definitions of Data Analysis

Data analysis refers to the process of analyzing a large amount of collected data with appropriate statistical analysis methods, extracting useful information and forming conclusions, and making a detailed study and summary of the data. This process is also the supporting process of the quality management system. In practice, data analysis can help people make judgments to take appropriate action. The mathematical basis of data analysis was established in the early 20th century, but it was not until the advent of computers that practical operations became possible and data analysis was promoted. Data analysis is a combination of mathematics and computer science. (Galetto 2016.)

In the field of statistics, some scholars divide data analysis into descriptive data analysis, exploratory data analysis, and verifiable data analysis. Among them, exploratory data analysis focuses on finding new features in the data, while verifiable data analysis focuses on verifying the authenticity of existing hypotheses. From another point of view, descriptive data analysis belongs to primary data analysis. Common analysis methods include comparative analysis, average analysis, cross-analysis, etc. Exploratory data analysis and confirmatory data analysis belong to advanced data analysis. Common analysis methods include correlation analysis, factor analysis, regression analysis and so on. (Galetto 2016.)

Because data analysis is mostly done by software. This requires data analysts not only to master various data analysis methods but also to be familiar with the operation of mainstream data analysis software.

General data analysis can be done by Excel, while advanced data analysis should be carried out by professional analysis software, such as SPSS Statistics, a data analysis tool. (Galetto 2016.)

3.2 Theory and technologies

As all know, big data is not simply just the definition of the data itself, but the most important reality is to analyze big data. Only through analysis can get a lot of intelligent, in-depth and valuable information. So, more and more applications involve big data, and the attributes of these big data, including quantity, speed, diversity and so on, all present the increasing complexity of big data, so the analysis method of big data is particularly important in the field of big data, which can be said to be the decisive factor to determine whether the final information is valuable or not. (Sriram 2017.)

3.2.1 Five basic aspects of Big Data analysis

The first aspect is the analysis visualization. Data visualization is the most basic requirement of data analysis tools, whether for data analysis experts or ordinary users. Visualization can intuitively display the data, let the data speak for themselves, and let the audience hear the results. (Oracle 2013.)

The second aspect is the data mining algorithm. Visualization is for people to see, and data mining is for machines to see. Clustering, segmentation, outlier analysis and other algorithms allow us to dig deep into the data and find value. These algorithms not only deal with the amount of big data but also the speed of processing big data. (Oracle 2013.)

The third point is the prediction and analysis function. Data mining can enable analysts to better understand data, and predictive analysis can enable analysts to make some predictive judgments based on visual analysis and data mining results. (Oracle 2013.)

The fourth basic aspect is prediction and analysis function. We know that the diversity of unstructured data brings new challenges to data analysis. We need a series of tools to parse, extract and analyze data. Semantic engines need to be designed to extract information intelligently from "documents". (Oracle 2013.)

The last aspect is data quality and master data management. Data quality and data management are some management best practices. Data processing through standardized processes and tools ensures a predefined high-quality analysis result. (Oracle 2013.)

3.2.2 Basic Big Data processing flow

The first process is data collection. Big data acquisition refers to the use of multiple databases to receive data from the client (in the form of Web, App or sensor, etc.), and users can conduct simple queries and processing through these databases. In big data collection process, its main characteristic and challenge is high concurrency, because at the same time may have tens of thousands of users to access and manipulate, such as train ticketing website, they reached millions of concurrent traffic during peak, so need to deploy many databases on the acquisition end to support. And how to carry out load balancing and sharing among these databases needs in-depth thinking and design. (Oracle 2013.)

Then import preprocessing. Collection side, although there will be a lot of the database itself, to the analysis of these huge amounts of data efficiently or should the data import from front end to a centralized large distributed database, or distributed storage cluster, and can be based on the import do some simple cleaning and pretreatment. Some users will also use Storm from Twitter to perform the streaming calculations on the data when importing, to meet the real-time computing needs of some businesses. The characteristics and challenges of the import and preprocessing process are a large amount of data imported, which often reaches the level of 100 megabytes or even 1000 megabytes per second. (Oracle 2013.)

Then perform statistical analysis on the data. Statistics and analysis the main use of distributed database, or distributed computing cluster to store huge amounts of data in its ordinary analysis and classification summary, etc., in order to satisfy the demands of most common analysis, in this regard, some real-time demand would use the EMC Greenplum, Oracle Exadata does, as well as the column type based on MySQL storage Info bright, etc., and some of the batch or demand can use Hadoop based on semi-structured data. The main feature and challenge of statistics and analysis is a large amount of data involved in the analysis. (Oracle 2013.)

Finally, data mining is carried out. Different from the previous statistics and analysis process, data mining generally does not have any preset topics, but mainly carries out a calculation based on various

algorithms on existing data, to predict and realize some requirements for high-level data analysis. Typical algorithms are k-means for clustering, SVM for statistical learning and Naive Bayes for classification. Hadoop Mahout is the main tool used. (Oracle 2013.)

3.3 Big Data analysis tools

Big data analytics is a process that examines a wide variety of data sets to discover unknown correlations, hidden patterns, market trends, customer preferences, and most useful information that can help organizations make business decisions. More information on big data analytics. In today's era, many companies implement big data analytics to make life easier. Data can be processed very quickly and efficiently. This includes analyzing the data and using the results. This reduces workload and increases efficiency when it is not possible with more traditional business intelligence solutions. (Bappalige 2014.)

3.3.1 Apache Hadoop

Hadoop is a software framework for distributed processing of large amounts of data. It allows distributed processing of big data sets across mainframes using simple programming models. Applications that work with the Hadoop framework can work in environments that provide distributed storage and computing across computer clusters. Hadoop aims to expand from a single server to thousands of machines, each providing local computing and storage. Hadoop is reliable it maintains multiple copies of working data, ensuring that processing can be redistributed against failed nodes. Hadoop is efficient because it works in parallel, speeding up processing through parallel processing. Hadoop is also scalable and able to handle petabytes of data. Besides, Hadoop relies on a community server, so it costs less and can be used by anyone. (Bappalige 2014.)

The Hadoop framework transparently provides reliability and data movement for applications. It implements a programming paradigm called MapReduce: applications are divided into many small parts, and each part can be executed or re-executed on any node in the cluster. Besides, Hadoop also provides a distributed file system to store data from all computing nodes, which brings a very high bandwidth to the entire cluster. MapReduce and distributed file system design enables the whole framework to automatically handle node failures. It enables applications to work independently with thousands of computers and PB-level data. It is now generally accepted that the entire Apache Hadoop "platform" includes the Hadoop kernel, MapReduce, Hadoop Distributed File System (HDFS) and some related projects,

such as Apache Hive and Apache HBase. The core design of Hadoop's framework is that HDFS and MapReduce. HDFS provides storage for massive data, while MapReduce provides computation for massive data. (Bappalige 2014.)

Hadoop is a distributed computing platform that makes it easy for users to build and use. Users can easily develop and run applications handling massive amounts of data on Hadoop. It mainly has the following advantages: High reliability. Hadoop's ability to store and process data bit by bit is well respected; High scalability. Hadoop distributes data and performs computing tasks between clusters of available computers that can easily scale to thousands of nodes; High efficiency. Hadoop can move data between nodes dynamically and keep each node in dynamic balance, so processing speed is very fast; High fault tolerance. Hadoop automatically saves multiple copies of data and automatically redistributes failed tasks. Hadoop comes with a framework written in the Java language, so running on a Linux production platform is ideal. Applications on Hadoop can also be written in other languages, such as C++. (Bappalige 2014.)

3.3.2 HPCC

HPCC (High-Performance Computing Cluster), also known as DAS (Data Analysis Supercomputer), is an open-source data-intensive computing system platform developed by LexisNexis Risk Solutions. The HPCC platform uses a software architecture implemented on a commercial computing cluster to provide high-performance data-parallel processing for applications that leverage big data. The HPCC platform includes a system configuration that supports parallel batch data processing (Thor) and high-performance online query applications using index data files (Roxie). The HPCC platform also includes a data-centric declarative programming language for parallel data processing called ECL. (Middleton 2011.)

The HPCC software architecture includes Thor and Roxie clusters as well as common middleware components, an external communication layer, a client interface that provides end-user services and system management tools, and data from external sources that support monitoring and facilitating the loading and storage of file system auxiliary components. Typically, the HPCC environment includes only Thor clusters, or Thor and Roxie clusters, although Roxie is occasionally used to build its index. (Middleton 2011.)

The entire HPCC software architecture is shown in Figure 6:

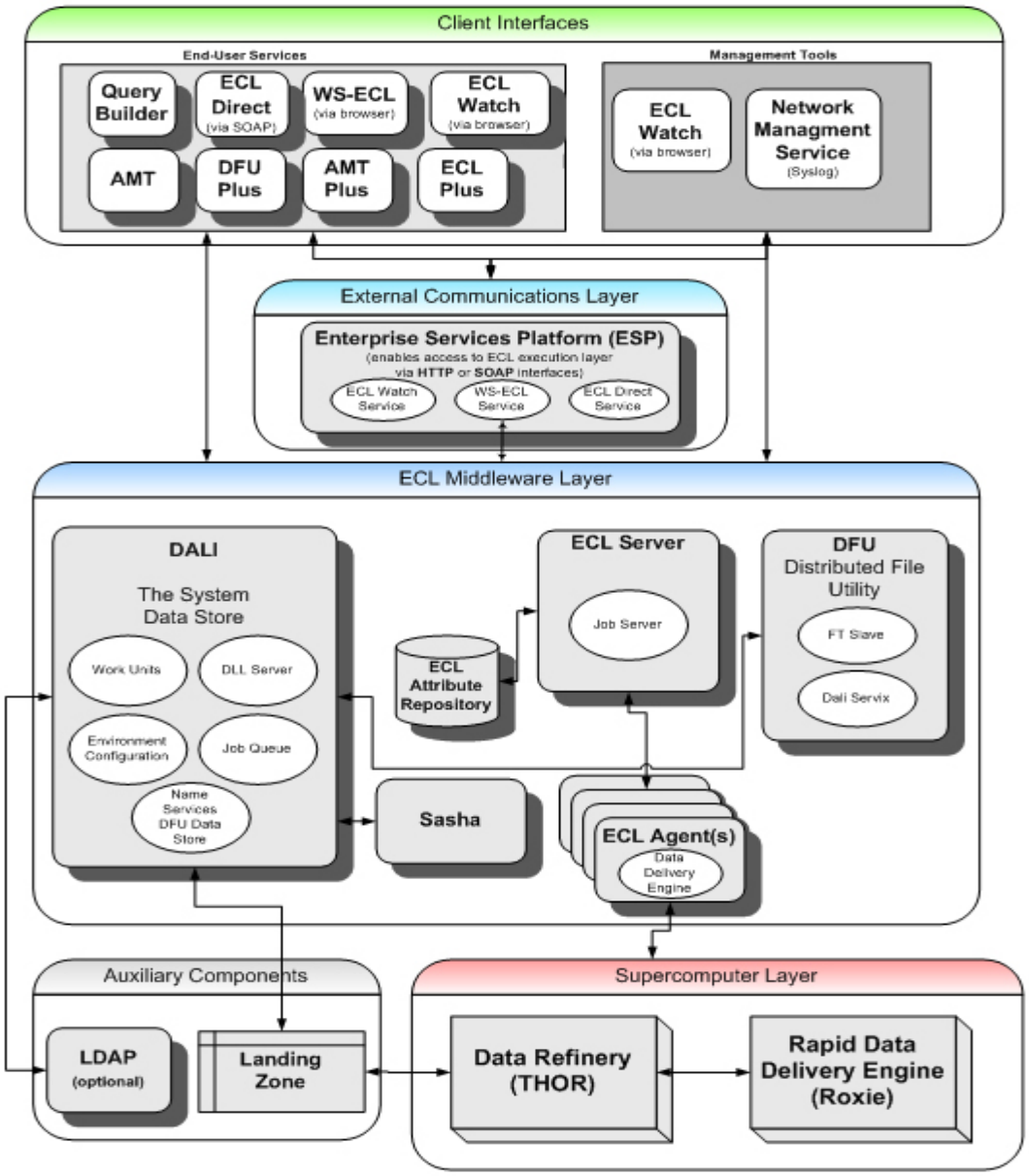


Figure 23. HPCC software architecture (Middleton 2011.)

3.3.3 SPSS

SPSS integrates data entry, data editing, data management, statistical analysis, report making, and graphic drawing. In theory, if the computer's hard disk and memory are large enough, SPSS can handle data files of any size, no matter how many variables are contained in the file or how many cases are contained in the data. SPSS is a widely used statistical analysis program for social sciences. It is also used by market researchers, health researchers, research companies, governments, education research-

ers, marketing organizations, data mining, and others. In addition to statistical analysis, data management (case selection, file sharing) and data documents are the characteristics of basic software. (Stauber 2017.)

Statistical information contained in basic software has descriptive statistics: cross-tabulation, frequency, description, exploration, descriptive ratio statistics; Bivariate statistics: mean, t-test, variance analysis, correlation, nonparametric test, Bayesian; Prediction of numerical results: linear regression; Prediction of recognition group: factor analysis, cluster analysis (two-step, K-means, hierarchical), discrimination Geospatial analysis, simulation; R Extension (GUI), Python. (Stauber 2017.)

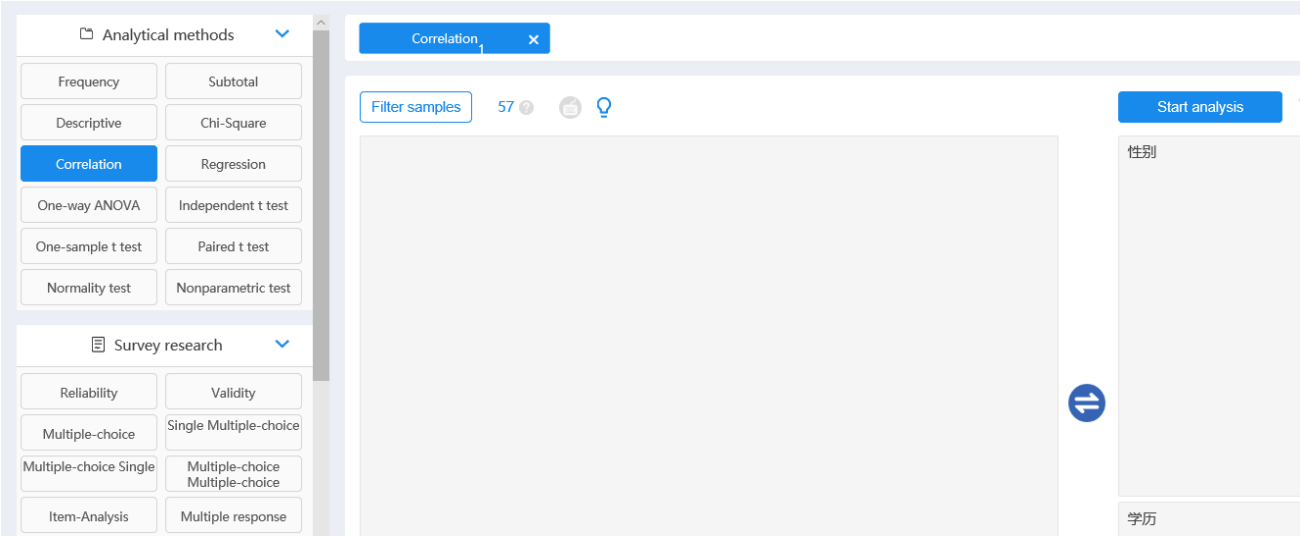


Figure 24. SPSS General Statistical Method (Stauber 2017.)

Statistical functions include conventional centralized and variance numbers, correlation analysis, regression analysis, variance analysis, chi-square test, t-test and non-parametric test, as well as recently developed multivariate statistical techniques, such as multivariate regression analysis, cluster analysis, discriminant analysis, principal component analysis and factor analysis, and can be displayed on the screen, such as normal distribution map, histogram, scatter point. Charts and other statistical charts. (Stauber 2017.)

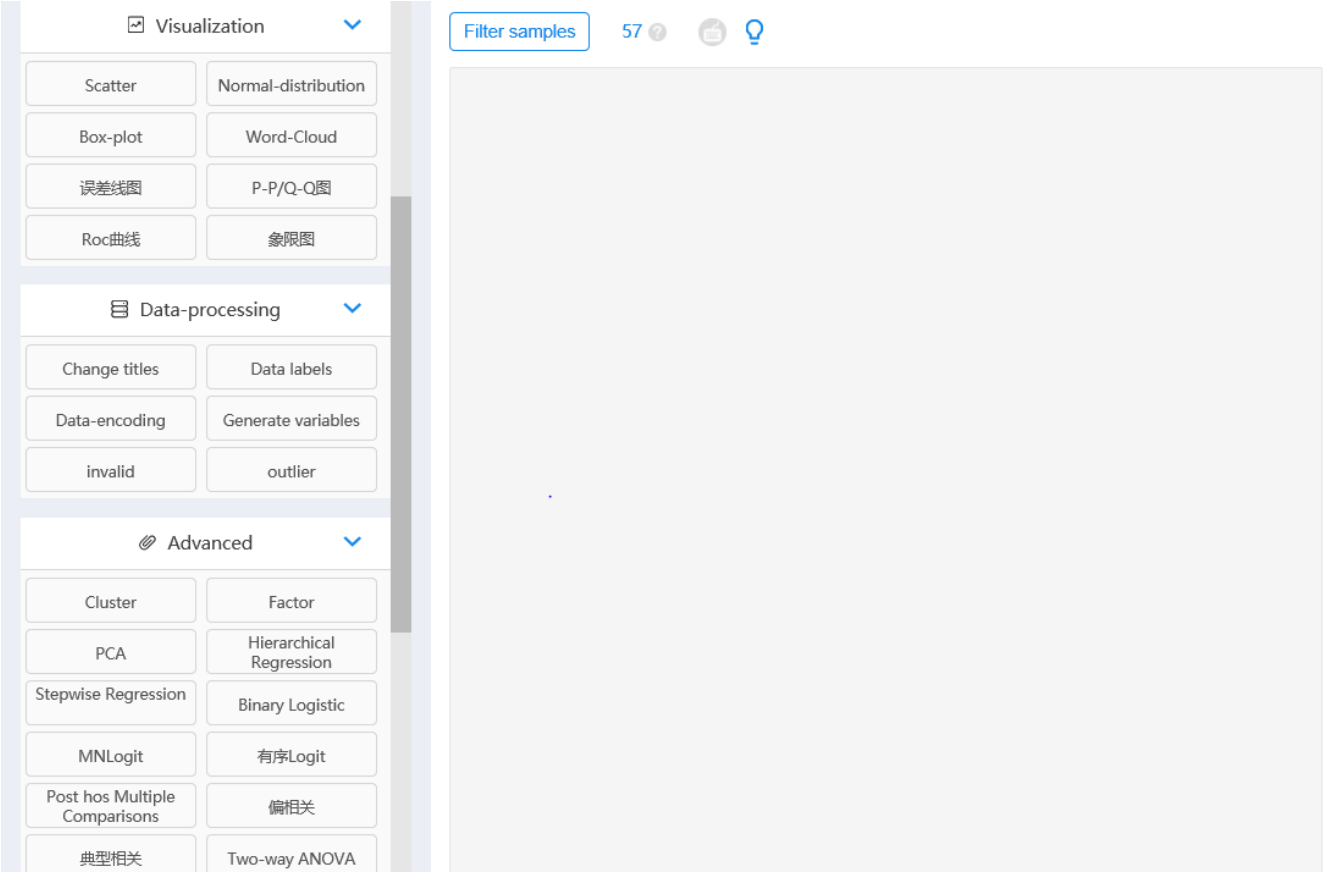


Figure 25. SPSS Data Analysis and Processing (Stauber 2017.)

GNU PSPP is an open-source alternative for SPSS (Proprietary Statistical Software), It provides similar functionalities to SPSS with addition to full support of open source generated and open-data file extensions as Gnumeric, LibreOffice and OpenOffice. It works smoothly with big datasets. (Medevel 2019.)

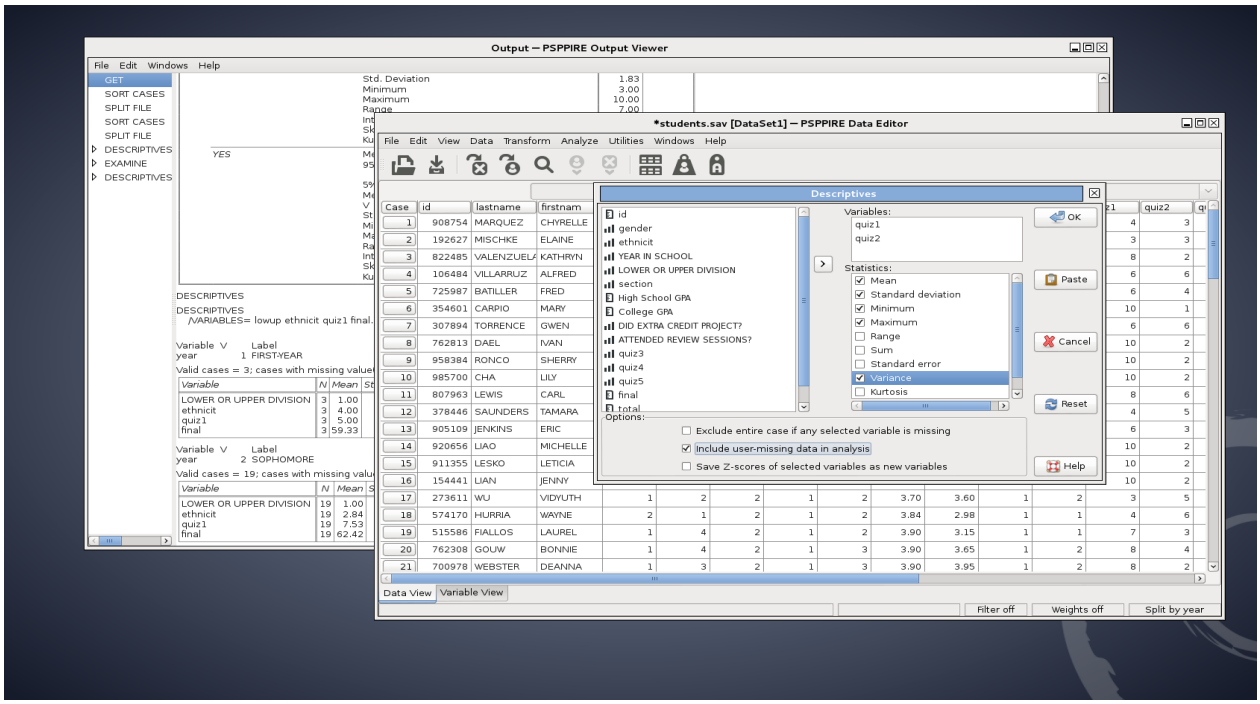


Figure 26. GNU PSPP (Medevel 2019.)

3.4 Data mining

Data Mining is the process of extracting hidden, unknown, but potentially useful information and knowledge from many practical, noisy, fuzzy and random application Data. This definition has several implications: the data source must be real, large, and noisy; Discover the knowledge that the user is interested in; The discovered knowledge should be acceptable, understandable and applicable; Discovery is not required to be a one-size-fits-all knowledge, only to support specific discovery problems. Synonyms of data mining include data fusion, artificial intelligence, business intelligence, pattern recognition, machine learning, knowledge discovery, data analysis, and decision support. (Clifton & Christopher 2010.)

3.4.1 Technical definition and meaning

Data and information are also forming of knowledge, but people tend to regard concepts, rules, patterns, rules and constraints as knowledge. People see data as a source of knowledge as if mining ore or panning for gold. Raw data can be structured, such as data in a relational database; It can also be semi-structured, such as text, graphics, and image data; Even heterogeneous data distributed over a network. The method of discovering knowledge can be mathematical or non-mathematical; It can be deductive or inductive.

The discovered knowledge can be used for information management, query optimization, decision support, and process control, as well as for the maintenance of data itself. Therefore, data mining is an interdisciplinary subject, which improves people's application of data from a low-level simple query to mining knowledge from data and providing decision support. Driven by this demand, researchers from different fields, especially scholars and engineers in database technology, artificial intelligence technology, mathematical statistics, visualization technology, parallel computing, and other aspects, have devoted themselves to the emerging research field of data mining and formed new technical hotspots. Here, knowledge discovery is not about discovering the universal truth, discovering new natural science theorems and pure mathematical formulas, or even proving machine theorems. All discovered knowledge is relative, domain-specific, with specific premises and constraints, and yet easily understood by the user. It is best to express the findings in natural language. (Clifton & Christopher 2010.)

Data mining is a new business information processing technology. Its main characteristic is to extract, transform, analyze and model a large amount of business data in the business database, and extract the key data to assist the business decision. In short, data mining is a kind of deep data analysis method. Data analysis itself has a history of many years, but in the past, the purpose of data collection and analysis was for scientific research. Also, due to the limitation of computing power at that time, the complex data analysis methods for the analysis of large amounts of data were greatly limited. At present, due to the realization of business automation in various industries, a lot of business data was generated in the commercial field, which was no longer collected for analysis but was generated by Opportunistic commercial operation. The analysis of these data is no longer purely for the research needs, but mainly to provide really valuable information for business decisions, to make profits. However, a common problem faced by all enterprises is that the amount of enterprise data is very large, but there is very little valuable information on it. Therefore, through deep analysis of a large amount of data, information beneficial to business operation and improving competitiveness can be obtained, just like panning for gold from ore, hence the name of data mining. (Clifton & Christopher 2010.)

Therefore, data mining can be described as an advanced and effective method to explore and analyze a large amount of enterprise data according to the established business objectives of the enterprise, reveal hidden, unknown or verified known laws, and further model them. (Clifton & Christopher 2010.)

3.4.2 Common methods of data mining

Common methods for data analysis by data mining include tracking patterns, classification, clustering, association rules, characteristics, and Web page mining, etc., which mine data from different perspectives. (Alton 2017.)

The first thing to describe is data tracking. One of the most basic techniques in data mining is learning to recognize patterns in data sets. This is usually the identification of certain distortions in periodic data or the fluctuation of a variable over time. For example, you may see that sales of one of your products seem to soar before the holidays, or you may notice that warm weather will encourage more people to visit your website. (Alton 2017.)

Classification is to find out the common characteristics of a group of data objects in the database and divide them into different classes according to the classification model. The purpose is to map the data items in the database to a given category through the classification model. It can be applied to customer classification, customer attributes, and feature analysis, customer satisfaction analysis, customer purchase trend prediction, etc. For example, a food retailer divides customers into different categories according to their preferences for different foods, this allows marketers to mail advertising brochures directly to customers with this preference, greatly increasing business opportunities. (Alton 2017.)

Clustering analysis divides a group of data into several categories according to similarity and difference. Its purpose is to make the similarity between data belonging to the same category as large as possible and the similarity between data belonging to different categories as small as possible. It can be applied to the classification of customer groups, customer background analysis, customer purchase trend prediction, market segmentation and so on. (Alton 2017.)

Association rules are rules that describe the relationship between data items in the database, that is, according to the appearance of some items in a transaction, some items can be exported in the same transaction, that is, hidden associations or mutual relations between data. In customer relationship management, through a lot of customer database, data mining, can from many records found in the interesting relationship, find out the key factors influencing the effect of marketing, product positioning, pricing, and customized customer base, customers seek, segmentation and maintain, marketing and sales, marketing, risk assessment and fraud prediction, etc. Provided a reference for decision support. Characteristics: Feature analysis is to extract the characteristic expressions about the data from a group of data in the database, which express the overall characteristics of the data set. (Alton 2017.)

With the rapid development of Internet and the global popularity of the Web, made of extremely rich information available on the Web, through to the Web mining, analyzes the huge amounts of data can make use of the Web, collection of politics, economy, policy, technology, financial, market, competitors, such as supply and demand information, customer information, focus on analysis and deal with those to the enterprise or the potential for significant influence of the external environment and internal business information, and according to the analysis results to find the problems appeared in the process of enterprise management and may cause early signs of crisis, to these information analyses and processing, In order to identify, analyze, evaluate and manage crises. (Alton 2017.)

3.4.3 Data mining function

Data mining makes proactive, knowledge-based decisions by predicting future trends and behaviors. The goal of data mining is to discover implicit and meaningful knowledge from the database, which has the following five main functions. (Han & Kamber 2000.)

Data mining can automatically predict trends and behaviors. Data mining automatically searches for predictive information in big databases. A lot of problems that need manual analysis in the past can be quickly and directly concluded from the data itself. A typical example is market forecasting. Data mining uses past promotional data to find the most rewarding users of future investments. Other predictable issues include forecasting bankruptcy and identifying the groups most likely to respond to designated events. (Han & Kamber 2000.)

Data mining can do a correlation analysis. Data association is a kind of important discoverable knowledge in a database. If there is some regularity between the values of two or more variables, it is called correlation. Relevance can be divided into simple correlation, temporal correlation, and causal correlation. The purpose of association analysis is to find out the hidden connections in the database. Sometimes the association function of data in the database is not known, even if it is known, it is uncertain, so the rules generated by association analysis have credibility. (Han & Kamber 2000.)

Records in the database can be divided into a series of meaningful subsets, namely clustering. Clustering enhances people's understanding of objective reality and is a prerequisite for conceptual description and deviation analysis. Clustering technology mainly includes traditional pattern recognition methods and mathematical taxonomy. The key point of clustering technology is not only considering the distance

between objects but also requiring the classes to have some connotation description, thus avoiding some one-sidedness of traditional technology. (Han & Kamber 2000.)

Then concept Description. A conceptual description is to describe the connotation of a certain kind of object and summarize its characteristics. The conceptual description can be divided into feature descriptions and distinctive descriptions. The former describes the common features of a class of objects, while the latter describes the differences between different classes of objects. Generating a class's feature description involves only the generality of all objects in that class. (Han & Kamber 2000.)

Data mining can detect a deviation in the database. Data in the database often have some abnormal records, so it is very meaningful to detect these deviations from the database. Deviations include many potential bits of knowledge, such as abnormal instances in classification, special cases that do not satisfy rules, deviations between observations and model predictions, and changes in quantities over time. The basic method of deviation detection is to find meaningful differences between observation results and reference values. (Han & Kamber 2000.)

Differences between Data Mining and Traditional Analytical Methods: The essential difference between data mining and traditional data analysis (such as query, report, online application analysis) is that data mining is to mine information and discover knowledge without definite assumptions. The information obtained from data mining should have three characteristics: unknown, effective and practical. (Han & Kamber 2000.)

Previously unknown information means that the information is unpredictable. Data mining is to discover information or knowledge that cannot be found by intuition, or even against intuition. The more unexpected the information is, the more valuable it may be. The most typical example is a business application that is a chain store passing through. Data mining has found an amazing connection between diapers and beer for children. (Han & Kamber 2000.)

3.4.4 Typical problems solved by data mining

It should be emphasized that data mining technology is application-oriented from the beginning. Currently, in many areas, data mining is a very fashionable word, especially in business areas such as banking, telecommunications, insurance, transportation, and retail (such as supermarkets). Typical business

problems that data mining can solve include: Database Marketing, Customer Segmentation & Classification, Profile Analysis, and Churn Analysis, Customer Credit Score (Credit) Scoring) and so on. (Xia 2004.)

Data mining technology has been widely used in enterprise marketing. By collecting, processing and processing a large amount of information related to consumer behavior, we can determine the interest, consumption habits, consumption tendency and consumption demand of specific consumer groups or individuals, and then infer the next consumption behavior of corresponding consumer groups or individuals. Then, based on this, we can identify the next consumption behavior of the corresponding consumer groups or individuals. Compared with the traditional large-scale marketing methods which do not distinguish the characteristics of consumers, the targeted marketing of specific content by the coming consumer groups greatly saves the marketing cost and improves the marketing effect, thus bringing more profits to the enterprises. (Xia 2004.)

Business consumption information comes from various channels in the market. For example, when we use credit cards to consume, commercial enterprises can collect business consumption information in the process of credit card settlement, record the time, place, goods or services we are interested in, price level and payment capacity we are willing to receive, etc. When we apply for credit cards, apply for car driving licenses, fill in commodity warranty forms and other needs to be filled in. Our personal information is stored in the corresponding business database on the grid; besides collecting relevant business information on their own, enterprises can even purchase such information from other companies or institutions for their use. (Xia 2004.)

These data information from various channels is combined and processed by supercomputers, parallel processing, neuron networks, modeling algorithms, and other information processing techniques, from which the decision-making information used by businessmen for directional marketing to specific consumer groups or individuals can be obtained. In countries and regions with a more developed market economy, many companies begin to process business information deeply through data mining because of the original information system, to build their competitive advantage and expand their turnover. American Express has a database for recording credit card business with a data volume of 5.4 billion characters and is still updating as the business progresses. Through mining these data, Express has formulated the promotion strategy of "Relationship Billing" preference, that is, if a customer buys a fashion set with Express Card in one store, then another pair of shoes in the same store can get a larger discount,

which can not only increase the sales of the store but also increase the sales of Express Card in that store utilization rate. (Xia 2004.)

Data mining-based marketing can often send out marketing materials to consumers related to their previous consumption behavior. Kraft Food Company has established a database of 30 million customers. The database is established by collecting customers and sales records that respond positively to other promotional means such as coupons issued by the company. Kraft Food Company understands the interests and tastes of specific customers through data mining and sends coupons of specific products to them on this basis. Recommend Kraft product recipes that meet customer's tastes and health status. Reader's Digest Publishing Company in the United States runs a business database that has accumulated for 40 years. It contains more than 100 million subscribers 'data all over the world. The database runs 24 hours a day continuously to ensure that the data is constantly updated in real-time. It is precisely based on the advantages of data mining on the customer data database that enables Reader's Digest Publishing Company to be popular. The magazine has expanded its business to the publishing and distribution of professional magazines, books and audio-visual products, A conceptual expanding its own business. (Xia 2004.)

3.4.5 How data mining relates to data analysis

Data analysis is simply to say, data analysis is to analyze data content. Professionally speaking, data analysis refers to the use of appropriate statistical analysis methods and tools to process and analyze collected data, extract valuable information and play the role of data according to the purpose of analysis. It mainly achieves three functions: current situation analysis, cause analysis and forecasting analysis. The goal of data analysis is clear. First, make assumptions, and then verify the hypothesis through data analysis, to get the corresponding conclusions. Contrast analysis, grouping analysis, cross-analysis, regression analysis, and other commonly used analysis methods are mainly used. Data analysis usually obtains the results of an index statistic, such as sum, average, etc. These index data need to be interpreted in combination with business to play the value and role of the data. (Netease Cloud 2018.)

Data mining refers to the process of mining unknown and valuable information and knowledge from a large number of data through statistical, artificial intelligence, machine learning, and other methods. Data mining mainly focuses on solving four kinds of problems: classification, clustering, Association and prediction, the focus of data mining is to find unknown patterns and laws; as is often said, data mining cases: beer and diapers, which are previously unknown, but also very valuable information. The

main methods of mining are decision tree, neural network, Association rules, cluster analysis, artificial intelligence, machine learning and so on. Results Output model or rules, and corresponding model scores or labels, such as loss probability, total score, similarity, predictive value, labels such as high and low-value users, loss and non-loss, good and bad credit, etc. (Netease Cloud 2018)

The essence of data analysis and data mining is the same, which is to discover knowledge about business (valuable information) from data, to help business operations, improve products and help enterprises make better decisions. Therefore, data analysis and data mining constitute a broad data analysis. Data analysis is an operational means of data. Or algorithms. The goal is to sort out, filter and process the data according to prior constraints to obtain information. Data mining is the value analysis of the information after the means of data analysis. (NetEase Cloud 2018.)

The biggest difference between data analysis and data mining is that data analysis is based on input data and processed by prior constraints, but not by adjusting conclusions. For example, researchers need image recognition, which belongs to data analysis. Face analysis is needed. Data is obtained through a priori approach, regardless of the outcome. There's no problem with your data analysis. You need to bear the consequences in silence and respect the facts. Therefore, data analysis focuses on data validity, authenticity and the correctness of prior constraints. Data mining is different. Data mining is the acquisition of the value of information. Valuation naturally does not consider the data itself, but the value of the data. Thus, a batch of data, when trying to do different value mining on it. Assessment is data mining. (NetEase Cloud 2018.)

Big data is the massive data mining of the Internet, and data mining is more for the internal enterprise industry of small-scale data mining, data analysis is to make targeted analysis and diagnosis, big data needs to be analyzed is the trend and development, data mining mainly discovers problems and diagnosis. (NetEase Cloud 2018.)

3.5 Internet of Things

The development of the Internet of Things will completely change people's way of life and greatly improve people's quality of life and efficiency. Logistics is related to the clothing, food, housing, and transportation of modern people, and its development is related to all aspects of the social economy. Extensive promotion and application of the Internet of Things technology can not only improve and optimize the logistics supply chain management system and rationalize logistics management but also play an active

role in improving logistics efficiency, reducing logistics costs and optimizing resource allocation. (Wigmore 2014.)

3.5.1 The concept of the Internet of Things

The Internet of things (IoT) is the extension of Internet connectivity into physical devices and everyday objects. Embedded with electronics, Internet connectivity, and other forms of hardware (such as sensors), these devices can communicate and interact with others over the Internet, and they can be remotely monitored and controlled. The definition of the Internet of things has evolved due to the convergence of multiple technologies, real-time analytics, machine learning, commodity sensors, and embedded systems. Traditional fields of embedded systems, wireless sensor networks, control systems, automation (including home and building automation), and others all contribute to enabling the Internet of things. In the consumer market, IoT technology is most synonymous with products pertaining to the concept of the "smart home", covering devices and appliances (such as lighting fixtures, thermostats, home security systems and cameras, and other home appliances) that support one or more common ecosystems, and can be controlled via devices associated with that ecosystem, such as smartphones and smart speakers. (Wigmore 2014.)

3.5.2 Internet of Things technology

Internet of Things technology is defined as: through radio frequency identification (RFID), infrared sensors, global positioning system, laser scanner, and other information sensing equipment, according to the agreement, any items will be connected to the Internet, information exchange and communication, to achieve intelligent identification, location, tracking, monitoring, and management. Network technology. The key technologies of the Internet of Things include radio frequency identification technology, sensor technology, network and communication technology, and data mining and fusion technology. (Wigmore 2014.)

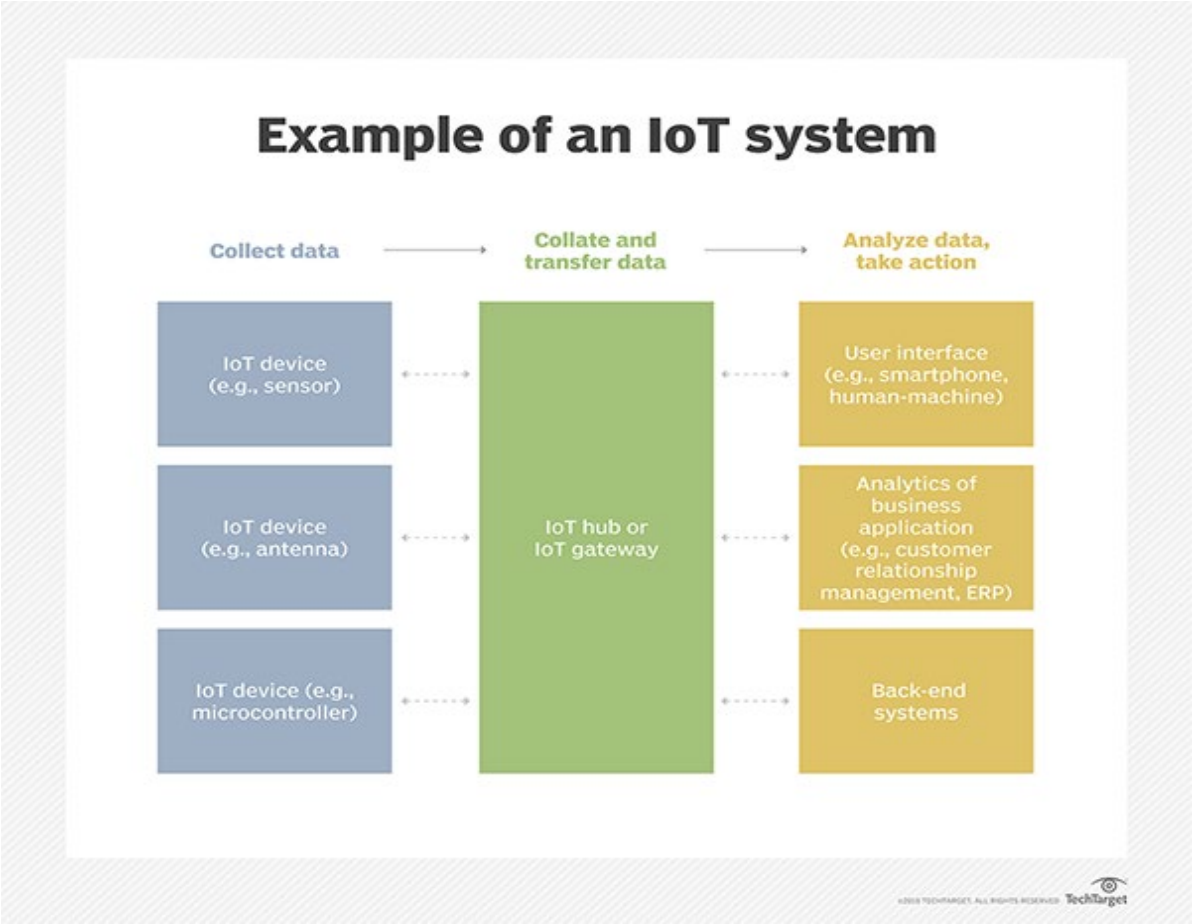


Figure 27. Example of an IoT system (Wigmore 2014.)

Radio Frequency Identification Technology, Radio Frequency Identification (RFID), also known as an electronic tag, is a wireless communication technology that uses radio frequency signals and its spatially coupled transmission characteristics to realize automatic identification of static or mobile objects to be identified and read and write related wireless data. The information reading and writing device sends radio frequency signals to the products with electronic tags, activates the electronic tags, and releases the information stored in the chip by the energy obtained from the induction current, thus completing the management and control of the items. (Wigmore 2014.)

Another key technology used in the Internet of Things is sensor technology. As an important means of information acquisition, it is mainly accomplished by sensors, sensor nodes, and electronic tags. Integrated miniaturized sensors can be embedded in any object, and cooperate to monitor it in real-time, and then upload the collected information wirelessly, to realize ubiquitous sensing. Sensor nodes have the ability of sensing, computing, and communication, and can collect, process and transmit data. (Wigmore 2014.)

Another aspect of the Internet of things technology is network and communication technology. The reliable transmission of information security in the Internet of Things involves two aspects: long-distance communication and short-distance communication. In the aspect of remote communication, it mainly includes IP Internet, 2G/3G mobile communication, satellite communication, Internet networking, and gateway technology. The short-range technology mainly includes WIFI, Bluetooth and so on. (Wigmore 2014.)

Data Mining and Fusion Technology are distributed computing technologies such as P2P and cloud computing provide a new and efficient computing model for the Internet of Things. They can mine hidden and effective information from massive data in time. They can also solve the problem of data fusion between heterogeneous networks or multiple systems. They have relatively reliable data centers and can easily realize data and application sharing among different devices. (Wigmore 2014.)

3.5.3 Application of Internet of Things in real life

There are numerous real-world applications of the Internet of things, ranging from consumer IoT and enterprise IoT to manufacturing and industrial IoT. IoT applications span numerous verticals, including automotive, telco, energy and more. In the consumer segment, for example, smart homes that are equipped with smart thermostats, smart appliances, and connected heating, lighting and electronic devices can be controlled remotely via computers, smartphones or other mobile devices. (Wigmore 2014.)

Wearable devices with sensors and software can collect and analyze user data, send messages to other technologies about the users to make users ‘lives easier and more comfortable. Wearable devices are also used for public safety -- for example, improve first responders' responses during emergencies by providing optimized routs to a location or by tracking construction workers ‘or firefighters' vital signs at life-threatening sites. (Wigmore 2014.)

In health care, IoT offers many benefits, including the ability to monitor patients more closely to use the data that's generated and analyzed. Hospitals often use IoT systems to complete tasks such as inventory management, for both pharmaceuticals and medical instruments. (Wigmore 2014.)

Urban Operation Management would be monitoring and management of urban network and components, such as well covers, public facilities, key facilities such as urban water, electricity, gas, heat and other

low-rise pipelines, and the status of various vehicles and personnel, limited monitoring of daily sanitation operations, snow-sweeping and ice-shoveling, garbage dregs removal. (Wigmore 2014.)

The application of the Internet of Things in the field of health care can effectively detect the indicators of patients. It can be used in medical supervision, drug supervision, medical electronic file management, plasma collection, and monitoring, etc. Providing support for patient monitoring, telemedicine, and assistance for the disabled and providing timely and warm care for the vulnerable is one of the application areas of the Internet of Things, which has received unprecedented attention. (Joyia & Liaqat & Farooq & Rehman 2017)

Application in Agriculture can be widely used in monitoring and controlling crop growth environment, animal health monitoring and animal slaughter monitoring. Functional monitoring through sensing technology can timely sense the changes in soil composition, moisture, and fertilizer, dynamically track the growth process of plants and provide a scientific basis for real-time adjustment of farming methods. In every link of food processing, through the Internet of Things, we can track the growth, processing and marketing process of animal and plant products in real-time and monitor the quality and safety of products. (Meola 2016.)

In the ecological environment, through intelligent perception and transmission of information, the Internet of Things can play a huge role in atmospheric and soil management, forest and water resources protection, climate change and natural disasters. It can help improve the living environment and form closed-loop management of monitoring, early warning and control of pollution sources by using the Internet of Things technology. Sensors are used to strengthen air quality and urban noise monitoring, on-site information publicity in public places and the mobile communication system is used to strengthen the linkage with supervision departments. Strengthen the construction of a water quality detection network system for reservoirs, rivers, and residential water, and form real-time monitoring. Strengthen the construction of the sensing system for natural resources such as forest greenbelt and wetland and control the situation of greenbelt resources in time with the geospatial database. Sensor technology and communication technology are used to rationally allocate and use water, electricity, natural gas, coal and oil resources. (Jane & Kirk 2015.)

3.5.4 The connection between the Internet of Things and Big Data

Internet of Things, Big Data and Cloud Computing, as the representative technologies of the third information tide, will have a wide impact in the future. Internet of Things focuses on the interconnection of things, big data focuses on the value of data, while cloud computing provides services such as computing resources for big data and the Internet of Things. The connection between the Internet of Things and big data is still very close, mainly reflected in the following aspects: (JunMing 2019.)

First, the Internet of Things is an important foundation for big data. There are three main data sources of big data, namely the Internet of Things, Web system and traditional information system. The Internet of Things is the main data source of big data, accounting for more than 90% of the total data source, so there is no big data without the Internet of Things. (JunMing 2019.)

Second, big data is an important part of the Internet of Things system. The architecture of the Internet of Things is divided into six parts: equipment, network, platform, analysis, application, and security. The main content of the analysis part is big data analysis. Big data analysis is one of the most important means to realize the value of big data. There are two methods of analysis, one is based on statistics, the other is based on machine learning. When big data is combined with artificial intelligence technology, the agent can send the decision to the terminal through the Internet of Things platform, of course, the decision can also be made manually. (JunMing 2019.)

Third, the development of the Internet of Things platform further integrates big data and artificial intelligence. At present, the research and development of the Internet of Things platform are in the developing stage. With the formulation of relevant standards, the future Internet of Things platform will further integrate big data and artificial intelligence. The future of the Internet of Things is bound to be data and intelligent. (JunMing 2019.)

4 REAL CASE ANALYSIS OF BIG DATA

Artificial intelligence has made such a big breakthrough in recent years, one of the important driving forces is big data. Before the concept of big data came into being, computers couldn't solve some problems that need to be judged. So today's AI is more like data intelligence. AI uses a lot of data as a guide so that the problems that need to be judged by machines can finally be transformed into data problems. If artificial intelligence feels far from life, then traffic information should be the closest part of people's lives. Traffic management is also inseparable from the existence of big data analysis. Through the data collection and processing of urban traffic, big data technology can realize the optimization of urban traffic. The next section will explain the application of big data in ITS in detail. (Lingling & Tianpeng 2015.)

4.1 Application of Big Data in Artificial Intelligence

Recently, artificial intelligence (AI) is very popular. Its remarkable feature is that a computer can complete a lot of scientific and engineering calculations faster and more accurately than the human brain. The core of AI is that computers constantly acquire knowledge and learning strategies from experience. When encountering similar problems, they use empirical knowledge to solve problems and accumulate new experiences, just like ordinary people. The more experienced, the better the ability of AI to solve problems. Experience is essentially data. When the amount of data is large, it needs to be processed by big data technology. Therefore, AI can not do without big data technology. (Morris & Schlenoff & Srinivasan 2017.)

Artificial intelligence has been developed based on various knowledge such as psychology, mathematics, and informatics, which can summarize and analyze the laws of human activities in the society. Big data is based on a vast amount of information. Through scientific summary and classification, it can predict the possible events. The application of big data in artificial intelligence is mainly to realize data. To transform knowledge and promote the further improvement of technology. (Morris & Schlenoff & Srinivasan 2017.)

4.1.1 The core technology of big data in the AI domain

The first is to collect the data. Due to the continuous development of computer technology, a very large number of new data are also generated at all times. The current growth rate has reached 50% per year. The application of big data can make a detailed analysis of the specific movement status or location of some equipment, which accelerates the development of traditional information technology and makes data processing work consume manpower and material resources. The source is smaller, while AI reads the relevant information and intellectualizes the analysis through probability analysis or statistics, which improves the overall accuracy. (Oracle 2013.)

Big data storage applications are parallel databases. Through parallel processing for multiple nodes, the implementation of database tasks is realized. Because of its high performance, the current practical application is universal. Over the years, the performance of the system has been continuously improved, and the specific results cache and database index have been continuously improved. Because of its series of problems, many people choose to store data on Intelligent terminals. The application of AI robots to extract the core content information can save a lot of storage space and reduce storage risk. (Oracle 2013.)

Data representation, retrieval, and random access. Big data has its characteristics, and the representation of data is more complex. In the past system, keywords are published to all servers in data retrieval to achieve parallel retrieval. The retrieval results sometimes can not meet the actual needs. So some people have applied the HDFS system under the Apache Hadoop framework to open source information for big data, and finally. Random access is realized. (Oracle 2013.)

The final data is used and mined again. The application of big data has been very extensive. Today's e-shopping, publishing pictures or videos through social media and so on all use this technology. The core content of this technology is the mining technology of big data. The effective information searched in the huge, incomplete and random database can reduce various risks and finally get the scientific judgment. Big data mining technology is summarized as data classification, summary analysis, clustering, web data mining and so on. (Oracle 2013.)

4.1.2 Artificial intelligence robots

Using the perception level, operation level or cognitive level of AI robots to set, the robot can play a practical role, such as playing selected music content through software, quickly finding the phone number needed, providing nutritional meals matching actual requirements, and organically integrating AI

technology and big data technology so that the robot can perform like human beings. Decision-making or thinking, transferring a large amount of information through information sensors, using a pattern recognition engine to analyze big data structurally or systematically, and using data feedback or learning algorithm to deepen the skill set of robots. Through practical application, it is found that the more training corpus data, the more needs of neuron nodes, and the knowledge of specific semantics. Do not be able to be more accurate, through scientific calculations, the gap between the overall recognition rate of 10 million and 1 million neurons has reached 10%, and the gap between the overall recognition rate of 10 billion and 1 billion neurons has been higher than 20%, so the optimization of big data applications is an inevitable trend. (Manohar 2018.)

4.1.3 Intelligent manufacturing

Intelligent manufacturing is produced based on artificial intelligence. Knowledge is the basis for promoting the development of intelligence. Intelligence is a level of how to apply knowledge. Intelligent manufacturing includes intelligent manufacturing systems and intelligent manufacturing technology. A series of similar analysis, reasoning, and decision-making activities are carried out in the specific application process. Based on intelligent manufacturing, related automation concepts are innovated, and the development is increasingly highly integrated, intelligent and flexible. Many years of research and development have had an impact on the manufacturing industry. Data acquisition and management in the manufacturing industry, order management, intelligent manufacturing, and customization platform are all related to big data. After in-depth mining, more accurate matching can be achieved and the risk of manufacturers can be reduced. (Manohar 2018.)

4.1.4 Intelligent agriculture

The intelligent agriculture is a modern advanced agricultural production that can achieve high-efficiency, intensive and sustainable development through industrialized production under the condition of artificial management. It can carry out large-scale operations against the seasons, all-weather and anniversary. Based on modern agriculture, it can be applied to agricultural engineering, biotechnology, new materials, and other disciplines, the enable science and technology. It is a breakthrough reform and innovation that the technological level rises to a new height, which improves the productivity of land and the working efficiency of the working people. Combining with the actual situation of different regions, controlling specific instructions based on accurate data analysis and building a mobile big data system

for agriculture, agricultural workers can quickly understand the specific industry dynamics, grasp the growth status on time, and achieve scientific agricultural management. (Meola 2016.)

4.2 Application of Big Data in the Intelligent Transportation System

In the process of life and work, the fast pace of the city shows our dependence on traffic, the rapid growth of employment and the continuous increase of population density, which leads to the continuous emergence of traffic congestion. Nowadays, with the rapid development of cities, people are demanding more and more traffic trips. Faced with these normalized problems, such as slow urban traffic operation, serious traffic safety problems and frequent traffic accidents, the traditional way of thinking has been unable to solve these problems. Under the guidance of geographic information, communication, sensors, and computer technology, the new thinking mode of Intelligent Transportation has gradually changed from conceptual assumption to a new round of leapfrog development, which makes the management of Intelligent Transportation more efficient, informative and extensive. (Zipei 2012.)

In 1991, Bill Enman came up with a new term: Data Warehouse. Data Warehouse is a subject-oriented, integrated, non-Volatile and Time-Variant data set. The biggest difference between data warehouses and databases is that the data warehouse is designed for analyzing data for decision-making purposes, while a database is designed for efficient transaction office. Store and query data. With the application and development of data warehouse and database technology, the rapid accumulation and extensive use of data, we have entered the data age. We need powerful tools urgently to discover valuable information in data. Data mining technology can not only extract and transform the past data but also identify the possible connections between these data. The challenge of data mining in the future is that data is no longer a small amount of precise, sample-based and randomized data, but a large amount of mixed and massive data. Big data is developed from cloud computing, artificial intelligence, and data mining. (Zipei 2012.)

4.2.1 Data between Intelligent Transportation Systems

Intelligent Transport System (ITS) uses advanced detection, communication and computer technology to transform the traditional transportation system, thereby enhancing the efficiency of the system, improving the safety and efficiency of the ground transportation network, reducing energy consumption and environmental pollution and other integrated transportation and management systems. The application and integration technology of ITS enables system operators and users to better manage and optimize

the transportation system. ITS allows the use of information technology to collect state data for roads, traffic signals, buses, trucks, and trains; and integrate data to influence and improve the operating system. (Weidong & Xiangnong 2005.)

ITS is a complex integrated transportation system. Its operation needs to be realized through these subsystems. In road traffic system, people, vehicles, roads, and goods are important components. The main purpose of the system is to achieve the effective movement of these components. If the road traffic system is equipped with intelligent traffic information centers, traffic management centers, traffic control centers and other road traffic infrastructures such as intelligent vehicle-mounted facilities, various detection facilities, information dissemination facilities, it will constitute a complete intelligent transportation system. To realize the functions of ITS, it is necessary to integrate information among subsystems. The information-sharing platform is the main means of information fusion among subsystems. The platform will provide engine and scheduling of data resources and information sharing services for relevant subsystems. It will standardize the nature, organization structure, function, and transmission mode of the shared information of the whole urban traffic information system and form a data warehouse system by using an effective information circulation mechanism, and integrate, store and access the shared data and so on. (Weidong & Xiangnong 2005.)

By using big data technology, shared data can be extracted from the data information of various subsystems of dynamic intelligent transportation and integrated with cross-regional and cross-domain "data warehouse"; historical data can be migrated to the data platform while ensuring data integrity. It can also provide users with data information services according to the needs of each subsystem and the internal relations among them and organize the direct output of the internally stored data, while the related data stored by other subsystems are queried and supported by the information-sharing platform. (Weidong & Xiangnong 2005.)

4.2.2 Big Data Application of Information Acquisition Technology

Data is the basis and lifeblood of intelligent transportation. All intelligent applications in the system are realized through the real-time collection, analysis, prediction and scientific management of massive data. Average speed, average Lane occupancy, traffic flow, and speed are all very important traffic parameters. In traditional ITS, the main way to collect these data is to use static traffic detection methods of fixed-point detection equipment such as optical detectors, cameras to obtain real-time driving speed and travel conditions. With the development of mass storage technology, wireless communication technology, and real-time dynamic positioning technology, video surveillance, GPS data, and mobile data will gradually replace static detection data and become the basic data source of intelligent transportation. (Zipei 2012.)

In the era of big data, intelligent transportation integrates different real-time data systems, such as automobile navigation systems, traffic signal control system and global positioning systems. At the same time, it combines with parking guidance and information system, weather information system and so on. It realizes the connection between vehicle and vehicle, person and vehicle at any time and anywhere, and then collects data through such systems as a traffic information collection system based on IPv6. Data information, analysis of traffic behavior and state, integration of multi-system information to make decisions, and rapid response to emergencies, which greatly speeds up the frequency of real-time data acquisition, improves the accuracy of traffic behavior prediction and provides the factual basis for traffic management and service. For example, the information acquisition system is used to collect traffic status information, vehicle location information, accident information and other information for the traffic command center and driver to provide information exchange and data analysis, to ensure the rational allocation of traffic routes and control of traffic flow. (Zipei 2012.)

4.2.3 The architecture of Big data Analysis Platform

The creation of big data is accelerating, and a large amount of data, the variety, the rapid changes and the authenticity of these big data, etc., reflect the complexity of big data, so I want to be big, miscellaneous and fast. It is especially important to find out certain rules and trends in big data, that is, the analysis of big data. All we must do is to tell the big data reaction situation, that is, to analyze big data and mine more intelligent, deep and useful information. At present, some intelligent traffic data analysis systems have many shortcomings, such as inadequate use of traffic data, easy to ignore the potential value of traffic data, and difficult to obtain road information in time. With the continuous development

of society, we urgently need high-speed data mining and analysis methods for real-time and reliable analysis of traffic data. (Lingling & Tianpeng 2015.)

The data analysis platform consists of three parts: traffic data acquisition module, traffic data analysis module, and traffic data processing module. The relationship between the modules is shown in figure 11:

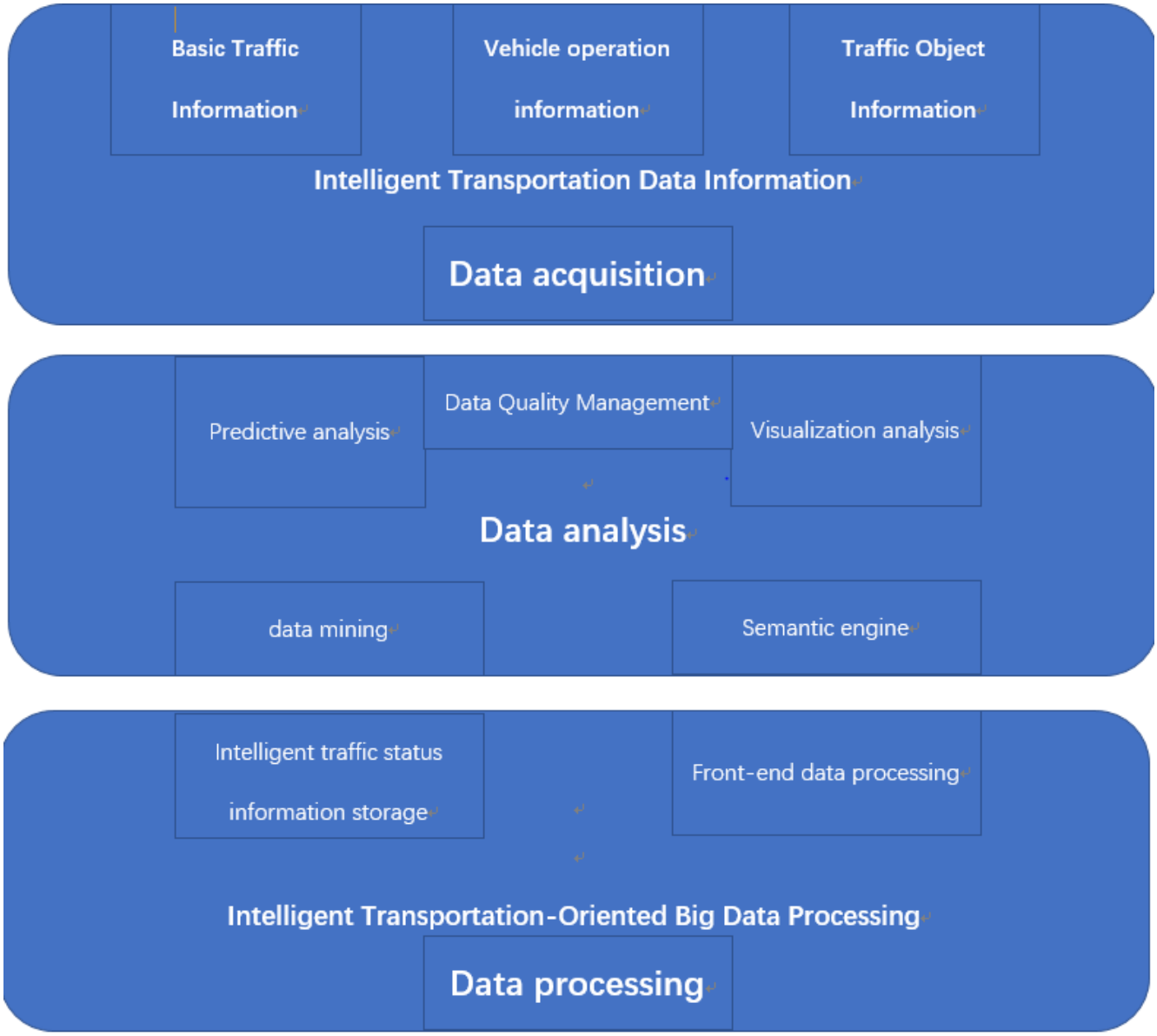


Figure 28. Intelligent Transportation Big Data Analysis Platform Architecture (Lingling & Tianpeng 2015.)

Traffic Information Acquisition Module means in the intelligent transportation system, traffic data information is the core content. The traffic data information acquisition module is the basis of the other two modules. It uses cloud computing, high-definition monitoring, mobile communication technology, vehicle networking and so on to realize the all-round collection of traffic data information. Front-end

equipment can acquire vehicle information, vehicle driving status information, surrounding environment information and so on, at the same time, it can determine the road condition and detect the environment. Front-end and back-end software supports a variety of video transmission protocols, has the search and storage function of traffic information management, and can support many front-end card devices, providing a piece of basic information for various needs of services. This module has many intelligent transportation systems, and the data update frequency is faster. At the same time, the data information perceived by each subsystem can be fused by wireless communication technology to obtain more accurate traffic information. (Lingling & Tianpeng 2015.)

Traffic data analysis module integrates structured, unstructured and multi-structured data into the platform by using visualization analysis, general database, data mining and other data analysis methods, and produces information conducive to decision-making and judgment of traffic agents through real-time analysis of data. The accuracy of the data analyzed should be strictly verified by the big data analysis platform, and the accuracy of the analysis method should be evaluated periodically. At the same time, the platform should be able to provide users with an analysis of traffic congestion status and traveler information service needs. While reducing data redundancy, support multi-dimensional data access and hierarchical storage. (Lingling & Tianpeng 2015.)

Traffic Data Processing Module means that in Intelligent Transportation System (ITS), we need to deal with a large amount of data, which is widely distributed. The data processing module is to achieve powerful data processing and analysis with high performance through a certain scale of a computing center and a complete computing framework. Users only need to submit relevant computing tasks and related data. Because traffic data is divided into real-time traffic data and historical analysis data, traffic data processing module is divided into two frameworks, real-time computing framework deals with real-time traffic data flow, non-real-time computing framework deals with batch data such as historical analysis. (Lingling & Tianpeng 2015.)

4.2.4 Technology realization

The analysis idea of intelligent traffic data analysis platform is to use visualization analysis and data mining technology to analyze and mine the correlation and similarity of data in different feature dimensions, and to convey the analysis results to managers or users. Improve the perception ability of the traffic management department, planning department and the public to the road condition and traffic

condition, improve the quality of traffic information service, and realize traffic intelligence. (Lingling & Tianpeng 2015.)

For the real-time collection of traffic data, it can be achieved in many ways. The static traffic information acquisition method mainly uses fixed position induction coil or video surveillance, relying on one or more induction coils installed under the road to generate electromagnetic induction, to detect the passing vehicle information. Dynamic traffic information acquisition method mainly collects real-time traffic flow, vehicle speed, time, traffic accident and other traffic parameters automatically through magnetic frequency, micro-frequency, photoelectric detectors, road conditions, and weight sensors. Besides, the use of aerial photography technology to track the movement of vehicles can predict the trend of congestion and calculate travel time. (Lingling & Tianpeng 2015.)

Integrating multi-channel and format inconsistent data information of each system, establishing a unified platform for video, graphics and image access. Based on these data information, extracting, integrating and in-depth analysis are carried out to obtain available information and knowledge. Aiming at the problems of large range fluctuation of mobile traffic data parameters and lack of discontinuity of traffic data in the traffic systems, the theory, and technology of data integration technology, decision support, and expert mathematical model are used to analyze traffic volume, congestion trend and traffic incidents. Besides, data mining technology is used for traffic flow prediction, traffic congestion analysis, road traffic safety analysis, and intelligent computing technology such as artificial intelligence and neural networks is used to provide effective technical support for traffic data acquisition, data information management, and intelligent analysis. (Lingling & Tianpeng 2015.)

Intelligent technologies such as dynamic traffic data processing and distributed processing are combined to deal with disorderly and seemingly irregular traffic data, which reflects the "intelligence" aspect of intelligent traffic. Data processing technology of distributed dynamic intelligent transportation systems is the preliminary progress of traffic data in the field of decision support, such as optimizing signal timing, traffic accident detection and so on. Dynamic traffic data processing technology can provide support for traffic information management such as short-term traffic prediction, abnormal behavior in road traffic, and retrieve relevant data from the data platform for processing and compare with the collected real-time data to judge the current traffic situation. (Lingling & Tianpeng 2015.)

4.3 Big Data on Medical Health

Informatized medical data, medical research data, patient characterization data, and medical health-related data generated by mobile devices, social networks, and sensors provide new ideas for healthcare practitioners to exploit potential relationships with big data technologies. Models that help physicians improve diagnostic accuracy, predict treatment outcomes, reduce healthcare costs, help pharmaceutical companies identify potential adverse drug reactions, and help public health departments identify potential epidemics promptly. The following is a description of the usefulness of big data in terms of public health, drug side-effect assessment, treatment prediction, and lowering medical costs, assisted diagnosis, and personalized treatment. (Dong & lin & jinhai & Liao 2015.)

4.3.1 Characteristics of medical and health Big Data

Big data can effectively help doctors make a more accurate clinical diagnosis. More accurately predict the cost and efficacy of treatment regimens; Integrating patient genetic information for personalized treatment; Analysis of population health data to predict disease outbreaks. Using big data can also reduce healthcare costs. The McKinsey global institute estimates that using big data analytics will save the United States $300 billion a year. The two areas with the greatest potential for savings include clinical operations and research and development. Examples of using big data to help healthcare companies do business are proliferating. For example, ActiveHealthManagement collects data about users' health to help users to realize health management. The integration of clinical data and genetic data in CancerIQ helps to realize the risk assessment, prevention, and treatment of cancer. clinical USES big data to predict treatment outcomes and lower costs. (Dong & lin & jinhai & Liao 2015.)

Besides the characteristics of big data, medical big data also has the characteristics of polymorphism, timeliness, incompleteness, redundancy, and privacy. Polymorphism means that doctors' descriptions of patients are subjective and difficult to achieve standardization. The timeliness index data is only useful for a period; Incompleteness refers to the deviation and absence of the state description of patients in medical analysis. Redundancy refers to the existence of a large amount of duplicate or irrelevant information in medical data. Privacy refers to the high privacy of users' medical and health data, and the leakage of information will cause serious consequences. (Yan & qin & fan 2014.)

4.3.2 Help find drug side effects

The detection of adverse reactions after drug marketing generally relies on passive detection and active detection. Passive testing relies on adverse reaction reports from doctors, patients and pharmaceutical companies. The biggest problem of passive detection is underreporting. According to reference (Karimi; Wang & Jimenez 2015), 94% of adverse reactions are not reported. Active detection of USES text mining and data mining technologies to discover potential adverse reactions caused by drugs from EHR, EMR, social network and search engines (Karimi & Wang & Jimenez 2015). References (Jin & Chen & He 2008) explore possible adverse drug reactions in electronic cases by taking advantage of the time sequence of adverse drug reactions. References (Chazard & Ficheur & Bernonville 2011) divided the conditions causing adverse reactions into the use of one drug, two drugs, one drug and one patient's characteristics, one drug, and one drug allergic event and found the relationship between conditions and adverse reaction results according to data mining methods such as decision tree and clustering. When there is a low-frequency causal relationship between drug use and adverse reactions, it is difficult for general data mining algorithms to distinguish causal relationships from accidental events.

4.3.3 Assisted treatment prediction and cost reduction

Part of the current high cost of health care comes from medical errors and waste. Preventable medical errors cause 98, 000 deaths a year in the United States alone, according to the American medical association. The United States spends more than 170 billion dollars on health care, while China spends more than 300 billion yuan on health care every year. Against this background, many countries are reforming their health care systems to reduce medical errors and waste, and ultimately cut health care costs. The HITECH act on medical and health information technology passed by the United States in 2011 announced that it decided to invest 50 billion dollars to use information technology to solve the existing problems in the medical industry within five years. In 2009, China announced the first part of a 10-year plan to reform its health care system for 120 billion yuan. (Chen & Compton & Hsiao 2013)

In reference (Srinivasan & Arunasalam 2013), the Australian medical insurance industry was analyzed, and it was believed that the current verification technology could not effectively detect fraud, abuse, waste, error and other phenomena in medical services because the old verification technology only focused on a single case and could not make use of the connection between multiple cases. Taking medical bills as the data source, the author established a prediction model for data such as treatment cost and hospital stay and used data mining techniques to discover abnormal data in bills. Use the rule base established by the domain expert to analyze the abnormal bill, find the possible problems and give warning.

Typical application environments include misuse of medical devices, surgical procedures that do not match the diagnosis of the disease, and excessive charges. Early detection of problems in the medical process could save state insurers, patients and private insurers a lot of money. (Srinivasan & Arunasalam 2013.)

4.3.4 Help with public health testing

In 2009, Google predicted the outbreak of a/H1N1 influenza 1-2 weeks earlier than the centers for disease control and prevention in the United States, which shocked the scientists in the medical and computer fields. Google's research report was published in the journal Nature. Google USES big data to predict flu outbreaks from relevant searches. Later, Baidu also launched "Baidu disease prediction", which USES user search to predict disease outbreaks. Predicting the outbreak of influenza with the help of big data can be divided into an active collection and passive collection. Passive collection analyzes the current situation and trend of influenza with the data submitted by users in the cycle, while active collection analyzes and forecasts with the records of users in tweets and search engines. (Davidson & Haim & Radin 2015.)

Flu Near You predicted the outbreak of influenza with the help of self-flu testing submitted by user cycles. First, people sign up for Flu Near You, and then each week they receive an email directing them to the site. On the site, users fill out a survey about whether they have flu symptoms. Eventually, Flu Near You collects information and USES big data technology to generate visualizations of current and future Flu disease forecasts. (Chunara & Aman & Smolinski 2013.)

In the early stage of an influenza outbreak, users usually search-relevant content in search engines or post relevant content on social networks, which can serve as the early warning of an epidemic outbreak. References (Lampos & Bie & Cristianini 2010.) by taking the tweets of users on Twitter and the urban influenza-like illness rate published by British health care bureau as the data source, the author selected characteristics and keywords of tweets through LASSO algorithm to establish the prediction model of influenza-like case rate in the next few days and obtained relatively accurate results. Long exposure to pathogens increases the chance of infection during disease transmission, so tracking population exposure and population location will help understand epidemic behavior. References (Kostkova 2013.) epidemiological data sources can be divided into media (including official media), mobile devices, social networking, Pro - Med mailing lists, and hospital laboratory data, and according to the different data, source

to design a set of data collection and analysis data, test data, data visualization system, intuitive performance of an epidemic. (Dong & lin & Jinhai & liao 2015.)

4.3.5 Source of medical health data

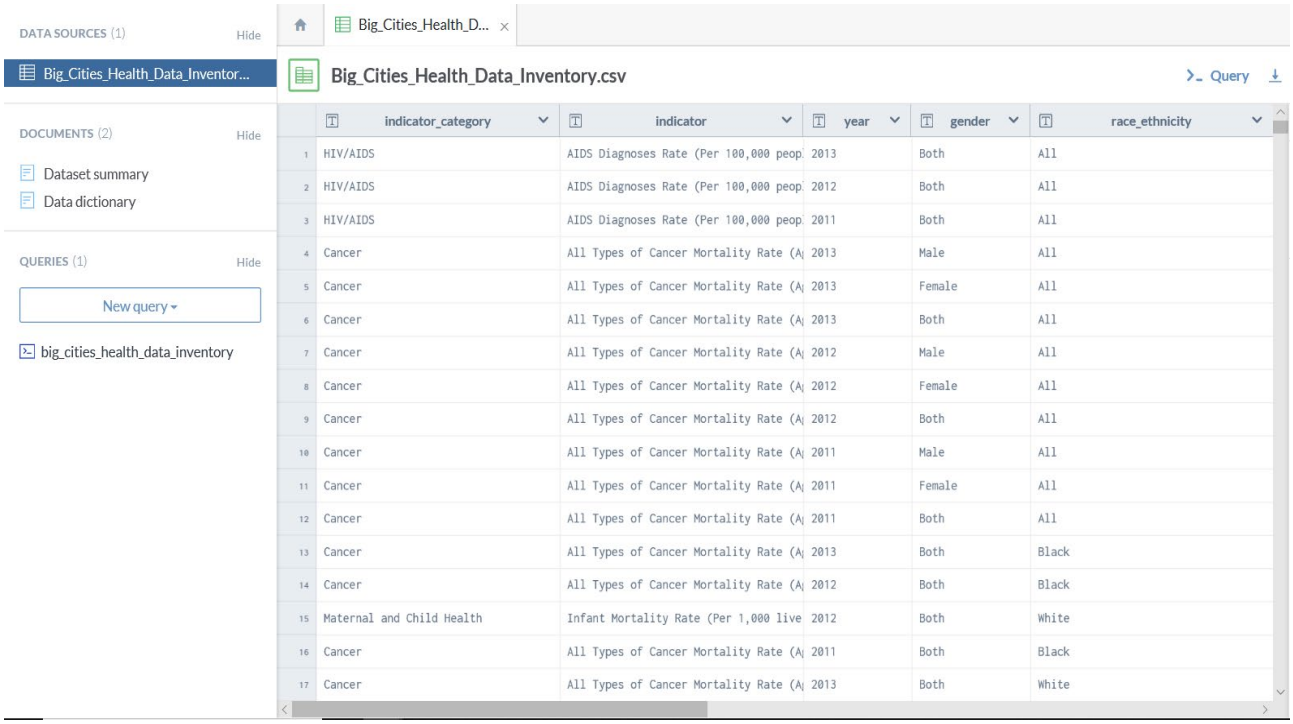
The sources of medical and health big data mainly include three aspects: personal health data, medical data, and population health data. In terms of personal health data, data sources are mainly sensor information and online information. Using visualization techniques to process personal health data and personal disease data can help users more easily realize health management and disease management. Processing personal diet and exercise data can help users intuitively understand the physical condition, and help users keep healthy. In terms of medical data, data sources are mainly medical research data and electronic case data. With doctors unable to keep up with the pace of discovering new medical knowledge from this data and applying it to patient care, medical visualization will provide doctors with the opportunity to intuitively understand new knowledge. Population health data and disease monitoring data can help users understand the population health status and disease outbreak status through visualization technology. (Groves & Kayyali & Knott 2013.)

Hospital information system (HIS) is an important source of medical data. Hospital information systems include Electronic medical record system (EMRS), laboratory information system (LIS), picture archiving &communication system (PACS), radiological information system (radiology) RIS, clinical decision support system (CDSS), etc. Also, various health devices can help collect vital signs such as blood oxygen concentration, respiration, blood pressure, temperature, pulse rate, and physical activity. Social networks and search engines also contain potential population health information. (Dong & lin & Jinhai & Liao 2015.)

4.3.6 Big cities health

The health care industry is shifting its focus from health care to prevention. The following chart, provided by the centers for disease control and prevention, shows the health status of 26 of the largest and most modern cities in the United States, including 34 health indicators. These indicators reflect some of the major causes of morbidity and mortality in the United States and the major priorities of national, state and local health institutions. Public health data were collected in nine major categories: HIV/AIDS, cancer, nutrition/physical activity/obesity, food safety, infectious diseases, maternal and child health,

tobacco, injury/violence, and behavioral health/substance abuse, and different individuals were recorded by indicator category, indicator, year, gender, race-ethnicity, and place. This data is helpful for public health detection and disease prevention at any time. (Dataquest 2017.)



	indicator_category	indicator	year	gender	race_ethnicity
1	HIV/AIDS	AIDS Diagnoses Rate (Per 100,000 peop	2013	Both	All
2	HIV/AIDS	AIDS Diagnoses Rate (Per 100,000 peop	2012	Both	All
3	HIV/AIDS	AIDS Diagnoses Rate (Per 100,000 peop	2011	Both	All
4	Cancer	All Types of Cancer Mortality Rate (A	2013	Male	All
5	Cancer	All Types of Cancer Mortality Rate (A	2013	Female	All
6	Cancer	All Types of Cancer Mortality Rate (A	2013	Both	All
7	Cancer	All Types of Cancer Mortality Rate (A	2012	Male	All
8	Cancer	All Types of Cancer Mortality Rate (A	2012	Female	All
9	Cancer	All Types of Cancer Mortality Rate (A	2012	Both	All
10	Cancer	All Types of Cancer Mortality Rate (A	2011	Male	All
11	Cancer	All Types of Cancer Mortality Rate (A	2011	Female	All
12	Cancer	All Types of Cancer Mortality Rate (A	2011	Both	All
13	Cancer	All Types of Cancer Mortality Rate (A	2013	Both	Black
14	Cancer	All Types of Cancer Mortality Rate (A	2012	Both	Black
15	Maternal and Child Health	Infant Mortality Rate (Per 1,000 live	2012	Both	White
16	Cancer	All Types of Cancer Mortality Rate (A	2011	Both	Black
17	Cancer	All Types of Cancer Mortality Rate (A	2013	Both	White

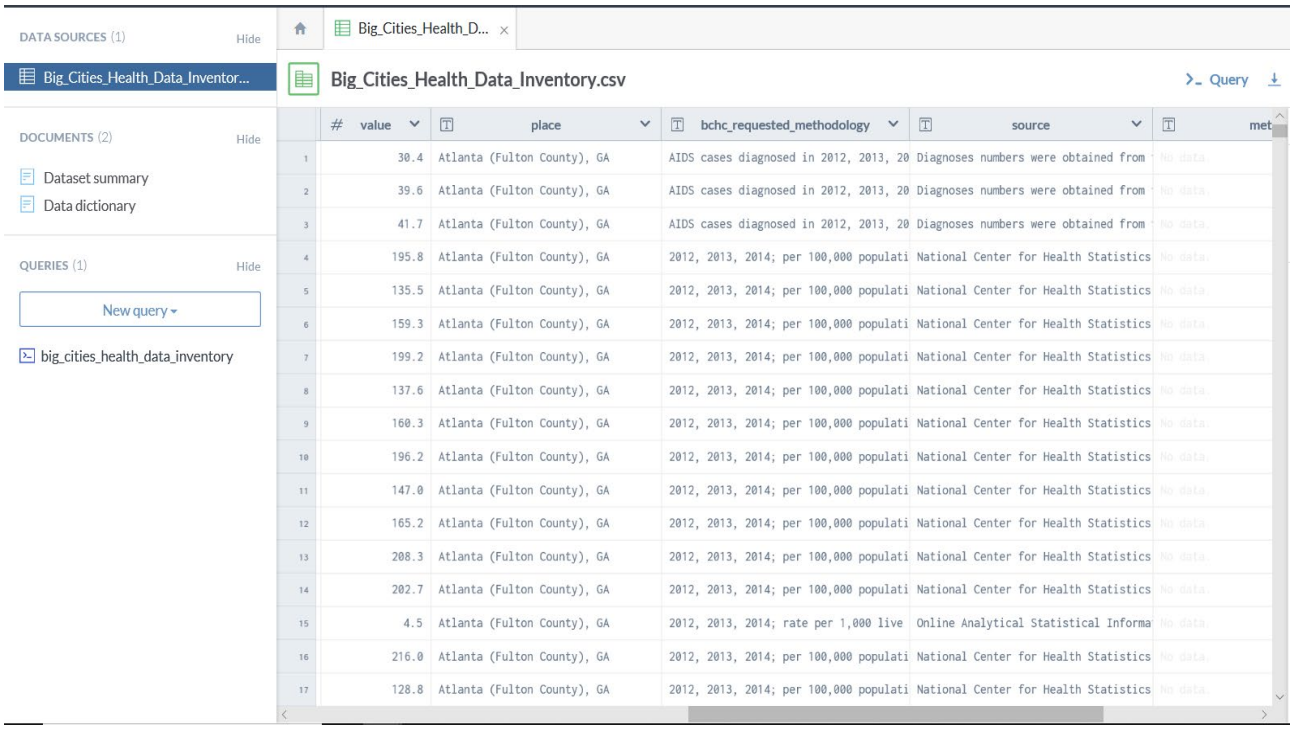
Figure 29. Big cities health data in an indicator, year gender and race-ethnicity (DataWorld 2017.)

The metropolitan health checklist (BCHI) is the open-access data platform of the metropolitan health alliance (BCHC). It is a standardized data collection center that provides a "snapshot" of health in the 30 largest and most prosperous cities in the United States and makes health indicators comparable in most BCHC member jurisdictions. The platform contains 18,000 data points across more than 50 health, socioeconomic, and demographic indicators across 11 categories in the United States. (BCHC 2016.)

Most of this data comes from cities, while others are protected by the U.S. census or other similar public data sets available for cities. When sample size allowed, the indicators were divided into racial and ethnic subgroups. City-specific data in the BCHI can help guide urban health policies and priorities and allow for comparative comparability between major U.S. urban centers. The data can be organized by city, by indicator, and by year, race, and gender. (BCHC 2016.)

Data are critical to practicing public health and help ensure that plans are responsive to the health needs of communities. Public health agencies are responsible for collecting and analyzing data on health problems and opportunities regularly and making it publicly available. Data is collected at the state, county,

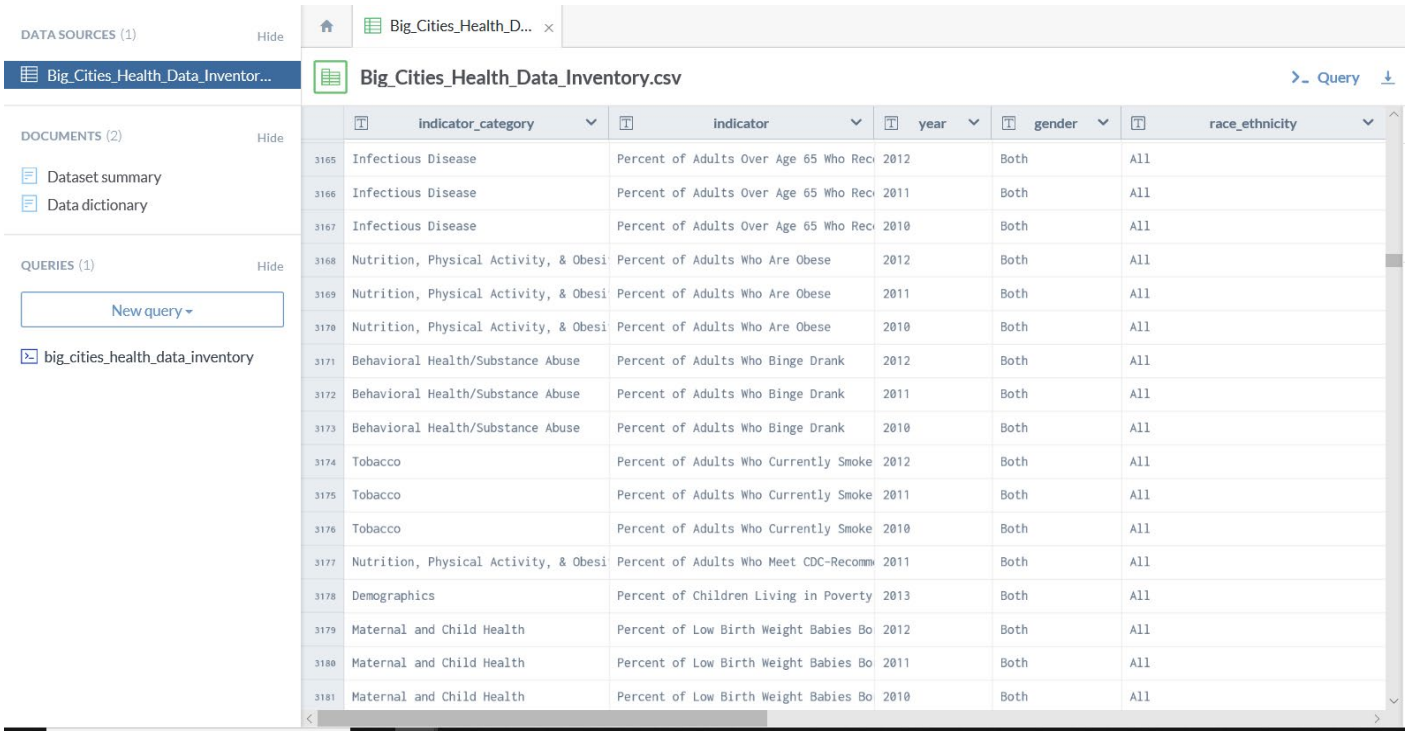
and city levels, and can be collected from many federal government surveys. Big cities' health data in place, value, bchc requested methodology, and source are shown in figure 14. (BCHC 2016.)



	# value	place	bchc_requested_methodology	source	met
1	30.4	Atlanta (Fulton County), GA	AIDS cases diagnosed in 2012, 2013, 20	Diagnoses numbers were obtained from	No data.
2	39.6	Atlanta (Fulton County), GA	AIDS cases diagnosed in 2012, 2013, 20	Diagnoses numbers were obtained from	No data.
3	41.7	Atlanta (Fulton County), GA	AIDS cases diagnosed in 2012, 2013, 20	Diagnoses numbers were obtained from	No data.
4	195.8	Atlanta (Fulton County), GA	2012, 2013, 2014; per 100,000 populati	National Center for Health Statistics	No data.
5	135.5	Atlanta (Fulton County), GA	2012, 2013, 2014; per 100,000 populati	National Center for Health Statistics	No data.
6	159.3	Atlanta (Fulton County), GA	2012, 2013, 2014; per 100,000 populati	National Center for Health Statistics	No data.
7	199.2	Atlanta (Fulton County), GA	2012, 2013, 2014; per 100,000 populati	National Center for Health Statistics	No data.
8	137.6	Atlanta (Fulton County), GA	2012, 2013, 2014; per 100,000 populati	National Center for Health Statistics	No data.
9	160.3	Atlanta (Fulton County), GA	2012, 2013, 2014; per 100,000 populati	National Center for Health Statistics	No data.
10	196.2	Atlanta (Fulton County), GA	2012, 2013, 2014; per 100,000 populati	National Center for Health Statistics	No data.
11	147.0	Atlanta (Fulton County), GA	2012, 2013, 2014; per 100,000 populati	National Center for Health Statistics	No data.
12	165.2	Atlanta (Fulton County), GA	2012, 2013, 2014; per 100,000 populati	National Center for Health Statistics	No data.
13	208.3	Atlanta (Fulton County), GA	2012, 2013, 2014; per 100,000 populati	National Center for Health Statistics	No data.
14	202.7	Atlanta (Fulton County), GA	2012, 2013, 2014; per 100,000 populati	National Center for Health Statistics	No data.
15	4.5	Atlanta (Fulton County), GA	2012, 2013, 2014; rate per 1,000 live	Online Analytical Statistical Informa	No data.
16	216.0	Atlanta (Fulton County), GA	2012, 2013, 2014; per 100,000 populati	National Center for Health Statistics	No data.
17	128.8	Atlanta (Fulton County), GA	2012, 2013, 2014; per 100,000 populati	National Center for Health Statistics	No data.

Figure 30. Big cities health data in place, value, bchc reqyested methodology and source (DataWorld 2017.)

Indicators include nine broad categories of public health importance: behavioral health and substance abuse; Cancer; Environmental health for chronic diseases; Food safety; HIV/AIDS; Infectious diseases; Injury and violence; And maternal and child health. The other two categories include demography and life expectancy/total mortality. These indicators were selected based on their relationship to the leading causes of morbidity and mortality in the United States and their role in creating healthier, safer communities. Some of the indicators are shown in figures 15 and 16. (BCHC 2016.)



	indicator_category	indicator	year	gender	race_ethnicity
3165	Infectious Disease	Percent of Adults Over Age 65 Who Rec	2012	Both	All
3166	Infectious Disease	Percent of Adults Over Age 65 Who Rec	2011	Both	All
3167	Infectious Disease	Percent of Adults Over Age 65 Who Rec	2010	Both	All
3168	Nutrition, Physical Activity, & Obesi	Percent of Adults Who Are Obese	2012	Both	All
3169	Nutrition, Physical Activity, & Obesi	Percent of Adults Who Are Obese	2011	Both	All
3170	Nutrition, Physical Activity, & Obesi	Percent of Adults Who Are Obese	2010	Both	All
3171	Behavioral Health/Substance Abuse	Percent of Adults Who Binge Drank	2012	Both	All
3172	Behavioral Health/Substance Abuse	Percent of Adults Who Binge Drank	2011	Both	All
3173	Behavioral Health/Substance Abuse	Percent of Adults Who Binge Drank	2010	Both	All
3174	Tobacco	Percent of Adults Who Currently Smoke	2012	Both	All
3175	Tobacco	Percent of Adults Who Currently Smoke	2011	Both	All
3176	Tobacco	Percent of Adults Who Currently Smoke	2010	Both	All
3177	Nutrition, Physical Activity, & Obesi	Percent of Adults Who Meet CDC-Recomm	2011	Both	All
3178	Demographics	Percent of Children Living in Poverty	2013	Both	All
3179	Maternal and Child Health	Percent of Low Birth Weight Babies Bo	2012	Both	All
3180	Maternal and Child Health	Percent of Low Birth Weight Babies Bo	2011	Both	All
3181	Maternal and Child Health	Percent of Low Birth Weight Babies Bo	2010	Both	All

Figure 31. Big cities health data for different indicator category (DataWorld 2017.)

Over the years, as the disease burden and priorities changed, BCHI also included indicators that added nutrition, obesity and physical exercise-related indicators. Looking ahead, other indicators will be included. Over the years, the report's contributions have been cited by local health sector professionals, newsletters, academic publications, and the press. Indicators were chosen because they affect mortality and morbidity among residents of large cities and are one of the indicators commonly used in public health. The metrics are shown in figures 15 and 16. (BCHC 2016.)

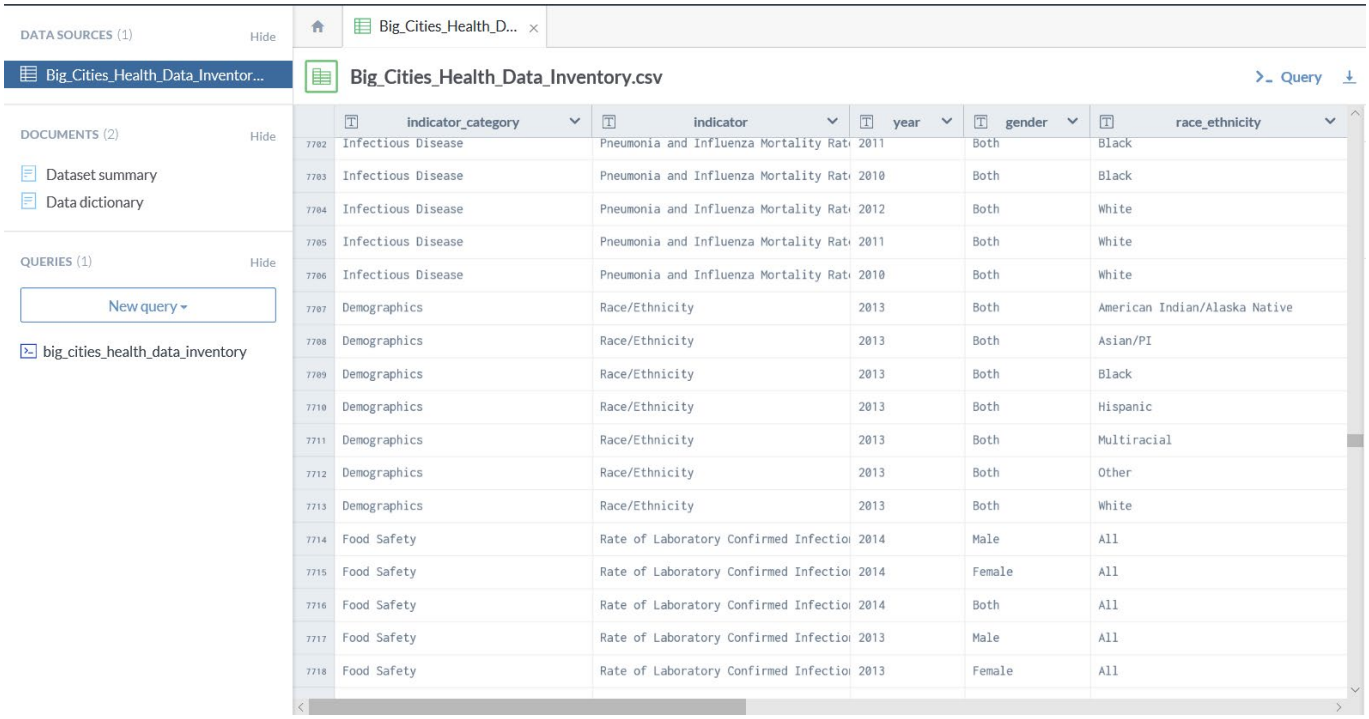


Figure 32. Big cities health data for different indicator category (DataWorld 2017.)

5 CONCLUSION

Through the analysis of references, this thesis elaborates the concept and characteristics of big data, data analysis and the effects of big data in various fields, including the application of big data in artificial intelligence and how big data manages intelligent transportation system, which provides effective help for readers who are not very clear about the concept of big data. Through the study, it is found that although the references on big data have been growing explosively in recent years. But the research and application of big data are still in its infancy. The thesis expounds the complementary existence of big data and artificial intelligence, the development of AI depends on cloud technology and big data technology, based on scientific analysis and research on the development situation, we should seek advantages and avoid disadvantages, and apply artificial intelligence technology to daily life or social development and construction work. However, there are still many problems in practical application. Artificial intelligence cannot make a scientific and reasonable judgment on some special situations occurring in human thought processing, and the protection of data and information security is another major research direction. The arrival of the era of big data opens a new way of thinking. Big data is a new resource. It reflects a new resource view. At the same time, big data represents a new generation of data management and analysis technology.

As mentioned in this thesis, data may become the largest commodity in the future. The application of big data in intelligent transportation and medical health is introduced. The big data analysis platform for intelligent transportation is built. The operation of the analysis platform and the intelligent transportation system can manage and control the urban traffic and road network and improve the construction of urban infrastructure. The application of big data technology in intelligent transportation makes full use of the advantages of data mining and big data processing technology, improves problems such as insufficient flexibility of infrastructure and limited resources, and improves the operating efficiency and core competitiveness of intelligent transportation. In terms of health care, big data on health care is still in its early stages of development, but it has shown the potential to transform medical services. Medical and health service providers can use big data analysis technology to explore potential relationships from clinical data, research data, personal health data, and public health data, providing help for clinical decision-making, public health, and personal health. In the future, medical and health big data will develop rapidly. With the development of technology, the combination of medical technology and big data technology will better provide services for human health.

REFERENCES

An Oracle White Paper, 2013. Big Data Analytics. Available: Big Data Analytics [March 2013]. Accessed: April 2019

Anthony M. Middleton, 2011. Available: "HPCC Systems: Introduction to HPCC (High-Performance Computing Cluster)". 24 May 2011. Accessed: May 2019

Bai Lingling, Han Tianpeng, 2015. Research on the Application of Big Data in Intelligent Transportation System. Available: Research on the Application of Big Data in Intelligent Transportation System ： 1009-3044(2015)10-0204-03 Accessed: May 2019

Bernard Marr, 2015. A brief history of big data everyone should read. Available: https://www.weforum.org/agenda/2015/02/a-brief-history-of-big-data-everyone-should-read/ 25 Feb, 2015. Accessed: March 2019

Beyer, Mark. 2011. Available: Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data. Gartner. 2011-07-13. Accessed: March 2019

CAICT, 2016. Big data white paper, Available: http://www.cac.gov.cn/files/pdf/baipishu/dashuju2016.pdf December 2016. Accessed: October 2019
Cao Weidong; Fang Xiangnong, 2005. Application of Data Mining in Intelligent Transportation System [J]. Computer Engineering, 2005, 31:91-92, 95. Accessed: May 2019

Chazard E, FicheurG, BernonvilleS, et al. Data mining to generate adverse drug events detection rules. IEEE Transactions on Information Technology in Biomedicine, 2011, 15(6): 823~830. Accessed: June 2019

Chen H, Compton S, Hsiao O . DiabeticLink: a Health Big Data System for Patient Empowerment and Personalized Healthcare. Smart Health. Berlin Heidelberg: Springer, 2013. Accessed: June 2019

Chunara R, Aman S, Smolinski M,et al. Flu near you: an online self-reported influenza surveillance system in the USA.Online Journal of Public Health Informatics, 2013, 5(1). Accessed: June 2019
Clifton, Christopher,2010. Data Mining. Available: Encyclopædia Britannica: Definition of Data Mining. Accessed: May 2019

Danny Sullivan, 2018. How Google autocomplete works in Search. Available: https://www.blog.google/products/search/how-google-autocomplete-works-search/ April 20, 2018. Accessed : March 2019

Dataquest, 2017. 19 Places to Find Free Data Sets for Data Science Projects / data world. Available: https://www.dataquest.io/blog/free-datasets-for-projects/ 11 January 2017. Accessed: June 2019

BCHC DATA PLATFORM, 2016. Available: https://bchi.bigcitieshealth.org/indicators/1827/searches / 34446/ Accessed: November 2019

Davidson M W, Haim D A,Radin J M. Using Networks to Combine “Big Data”and Traditional Surveillance toImproveInfluenza Predictions. Scientific Reports, 2015(5): 1~5. Accessed: June 2019
Ding Hui, 2019. The application of big data in the field of artificial intelligence. Available: https://www.zte.com.cn/china/about/magazine/zte-technologies/2017/3/cn_1357/462839 Accessed: March 2019

Dong cheng, lin li, jinhai, liao xiaofei, 2015. Medical health big data: application examples and system analysis. Available: http://www.xjis.org.cn/attachment/file/yiliao.pdf 21 January 2015. Accessed: June 2019

Doug Stauber, 2017. SPSS Predictive Analytics. Available: "What's New in SPSS Statistics 25 & Subscription - SPSS Predictive Analytics". 18 July 2017. Accessed: May 2019

Groves P, KayyaliB, KnottD,et al. The big data revolution in healthcare. McKinsey and Company,2013: 1~19. Accessed: June 2019

Hao ShuXian, 2014. Available: Innovative Analysis of Business Model in the Age of Big Data June, 2014. Accessed: March 2019

Hart Jane K; Martinez Kirk,2015. Available: "Toward an environmental Internet of Things" 1 May 2015. Accessed: May 2019

Hashemian M S, Stanley K G, Knowles D L, et al. Human network data collection in the wild: the epidemiological utility of microcontact and location data. Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, New York, USA,2012:255~264. Accessed: June 2019

J. Han and M. Kamber , 2000. Available: Data Mining: Concepts and Techniques Accessed: May 2019

Jin H D, Chen J, He H X, et al. Mining unexpected temporal associations: applications in detecting adverse drug reactions.IEEE Transactions on Information Technology in Biomedicine, 2008, 12(4): 488~500. Accessed: June 2019

Joyia Gulraiz J.; Liaqat Rao M.; Farooq Aftab; Rehman Saad, 2017. "Internet of Medical Things (IOMT): Applications, Benefits and Future Challenges in Healthcare Domain". Journal of Communications. Available: doi:10.12720/jcm.12.4.240-247. Accessed: May 2019

Karimi S, Wang C, Jimenez A M, et al. Text and data mining techniques in adverse drug reaction detection.ACM Computing Surveys , 2015, 47 (4): 56. Accessed: June 2019

KC Morris, C; Schlenoff and V. Srinivasan, 2017. "Guest Editorial A Remarkable Resurgence of Artificial Intelligence and Its Impact on Automation and Autonomy," in IEEE Transactions on Automation Science and Engineering, vol. 14, no. 2, pp. 407- 409, April 2017. Available: https://ieeexplore.ieee.org/document/7858589 Accessed: May 2019

Keith Carter, 2013. Big data: What is the importance? Available: https://thinkbusiness.nus.edu.sg/article/big-data-whats-the-big-deal-2/ December 9th, 2013. Accessed: April 2019

Kostkova P. A roadmap to integrated digital public health surveillance: the vision and the challenges. Proceedings of the 22nd International Conference on World Wide Web Companion, London, UK, 2013. Accessed: June 2019

Kuchipudi Sravanthi; Tatireddy Subba Reddy, 2015. Applications of Big data in Various Fields. ISSN: 0975-9646. Available: Applications of Big data in Various Fields Accessed: May 2019

Kusnetzky, Dan, 2010. Available: What is "Big Data?" ZDNet. 2010-02-21. Accessed: March 2019
Lampos V, Bie T D, Cristianini N. Flu detector-tracking epidemics on twitter. Machine Learning and Knowledge Discovery in Databases, 2010(6323): 599~602. Accessed: June 2019

Larry Alton, 2017. The 7 Most Important Data Mining Techniques. Available: https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques 22 December 2017. Accessed: May 2019

Lvy Wigmore, 2014. Available: "Internet of Things (IoT)". June, 2014. Accessed: May 2019
Manohar Parakh, 2018. Big Data and Robotics Available: https://dzone.com/articles/big-data-and-robotics 10 July 2018. Accessed: May 2019

Medevel, 2019. +20 Open-source Free Statistical, Data analysis and Notebook Projects for Data Scientistshttps. Available: https://medevel.com/open-source-data-science-analysis/ 9 February 2019. Accessed: June 2019

Meola A, 2016. Available: "Why IoT, big data & smart farming are the future of agriculture". Business Insider. Insider, Inc. Retrieved 26 July 2018. Accessed: May 2019

Michael Schroeck, Rebecca Shockley, Dr. Janet Smart, Professor Dolores Romero-Morales, Professor Molly Galetto, 2016. What Is Data Analysis? Available: https://www.ngdata.com/what-is-data-analysis/ January 20, 2016. Accessed: May 2019

Netease Cloud, 2018. Available: https://www.zhihu.com/question/20127962/answer/432920406 Zhihu. 2018-07-03. Accessed: June 2019

Pengzheng Ziyan, 2018. Netflix's AI Arena. Available: https://baike.baidu.com/tashuo/browse/content?id=19986ed346850a3d67355248 December 25, 2018. Accessed: October 2019.

Peter Tufano, 2012. Analysis: The application of big data in the real world. Available: https://www.ibm.com/downloads/cas/ED0JV08Q Accessed: October 2019

Pinal Dave, 2013. Big data - What is big data – 3Vs of Big Data – Volume, Velocity and Variety. Available: https://blog.sqlauthority.com/2013/10/02/big-data-what-is-big-data-3-vs-of-big-data-volume-velocity-and-variety-day-2-of-21/ October 2, 2013. Accessed: June 2019

Sachin P Bappalige, 2014. An introduction to Apache Hadoop for big data. Available: https://opensource.com/life/14/8/intro-apache-hadoop-big-data 26 Aug 2014. Accessed: May 2019
Srinivasan U,Arunasalam B. Leveraging big data analytics to reduce healthcare costs. IT Professional, 2013, 15 (6): 21~28. Accessed: June 2019

Sriram Vaidhyanathan, 2017. Available: An introduction to Big Data Analytics 14 Aug, 2017. Accessed: May 2019

Wang Dezheng; Shen Shanhong, 2019. Future trends in big data applications. Available: https://www.zte.com.cn/china/about/magazine/zte-technologies/2017/3/cn_1362/462822 Accessed: April 2019

Victor Maier Schoenberg ,2012. Big Data (with Kenneth Cukier), Houghton Mifflin Harcourt, ISBN 978-0544002692. Accessed: April 2019

Villanova University, 2014. What is big data? Available: https://www.villanovau.com/resources/bi/what-is-big-data/ June 4, 2013. Accessed: March 2019

Xia Huosong, 2004. Data Warehouse and Data Mining Technology. Available: http://abook.cn/pdf/2398.0201.pdf March, 2004. Accessed: May 2019

Xu Zipei,2010. The Data Revolution: Big Data [M]. Guangxi Normal University Press, 2012:86-93. Accessed: May 2019

Yan yan, qin xingbin, fan jianping. Review of medical and health big data research. Technology and application of scientific research informatization,2014, 5(6): 3-16. Accessed: June 2019

YonghongTech, 2017. Big data generation, characteristics and data analysis methods. Available: https://m.yonghongtech.com/zx/dsjfx/1344.html. 2017-07-12. Accessed: March 2019