Chapter 5
Big Data Analysis

Abstract In this chapter, we introduce the methods, architectures and tools for big data analysis. The analysis of big data mainly involves analytical methods for traditional data and big data, analytical architecture for big data, and software used for mining and analysis of big data. Data analysis is the final and the most important phase in the value chain of big data, with the purpose of extracting useful values, providing suggestions or decisions. Different levels of potential values can be generated through the analysis of datasets in different fields.

5.1 Traditional Data Analysis

Traditional data analysis means to use proper statistical methods to analyze massive first-hand data and second-hand data, to concentrate, extract, and refine useful data hidden in a batch of chaotic data, and to identify the inherent law of the subject matter, so as to develop functions of data to the greatest extent and maximize the value of data. Data analysis plays a huge guidance role in making development plans for a country, as well as understanding customer demands and predicting market trend by enterprises.

Big data analysis can be deemed as the analysis of a special kind of data. Therefore, many traditional data analysis methods may still be utilized for big data analysis. Several representative traditional data analysis methods are examined in the following, many of which are from statistics and computer science.

- *Cluster Analysis*: cluster analysis is a statistical method for grouping objects, and specifically, classifying objects according to some features. Cluster analysis is used to differentiate objects with certain features and divide them into some categories (clusters) according to these features,such that objects in the category will have high homogeneity different categories will have high heterogeneity. Cluster analysis is an unsupervised study method without the use of training data.

M. Chen et al., *Big Data: Related Technologies, Challenges and Future Prospects*, SpringerBriefs in Computer Science, DOI 10.1007/978-3-319-06245-7_5,

- *Factor Analysis*: factor analysis is basically targeted at describing the relation among many indicators or elements with only a few factors, i.e., grouping several closely related variables and then every group of variables becomes a factor (called a factor because it is unobservable, i.e., not a specific variable), and the few factors are then used to reveal the most valuable information of the original data.
- *Correlation Analysis*: correlation analysis is an analytical method for determining the law of correlations among observed phenomena and accordingly conducting forecast and control. There are a plentiful of quantitative relations among observed phenomena such as correlation, correlative dependence, and mutual restriction. Such relations may be classified into two types: (a) function, reflecting the strict dependence relationship among phenomena, which is also called a definitive dependence relationship, among which, every numerical value of a variable corresponds to one or several determined values; (b) correlation, under which some undetermined and inexact dependence relations exist, and a numerical value of a variable may correspond to several numerical values of the other variable, and such numerical values present regular fluctuation surrounding their mean values. A classic example is that customers of many supermarkets purchase beers while they are buying diapers.
- *Regression Analysis*: regression analysis is a mathematical tool for revealing correlations between one variable and several other variables. Based on a group of experiments or observed data, regression analysis identifies dependence relationships among variables hidden by randomness. Regression analysis may change complex and undetermined correlations among variables into simple and regular correlations.
- *A/B Testing*: also called bucket testing. It is a technology for determining plans to improve target variables by comparing the tested group. Big data will require a large number of tests to be executed and analyzed, to ensure sufficient scale of the groups for detecting the significant differences between the control group and the treatment group.
- *Statistical Analysis*: Statistical analysis is based on the statistical theory, a branch of applied mathematics. In statistical theory, randomness and uncertainty are modeled with Probability Theory. Statistical analysis can provide description and inference for large-scale datasets. Descriptive statistical analysis can summarize and describe datasets and inferential statistical analysis draws conclusions from data subject to random variations. Analytical technologies based on complex multi-variate statistical analysis include regression analysis, factor analysis, clustering, and recognition analysis, etc. Statistical analysis is widely applied in the economic and medical care fields [1].
- *Data Mining*: Data mining is a process for extracting hidden, unknown, but potentially useful information and knowledge from massive, incomplete, noisy, fuzzy, and random data. There are also some terms similar to data mining, e.g., discovering knowledge from databases, data analysis, data fusion, and decision support.

Data mining is mainly used to complete the following six different tasks, with corresponding analytical methods: Classification, Estimation, Prediction, Affinity grouping or association rules, Clustering, and Description and Visualization. Original data is deemed as the source to form knowledge and data mining is a process of discovering knowledge from the original data. Original data may be structured data, e.g., data in relational databases, or semi-structured data, e.g., text, graphical, and image data, or even heterogeneous data distributed in the network. Methods to discover knowledge may be mathematical or non-mathematical, and deductive or inductive. Discovered knowledge may be used for information management, query optimization, decision support, and process control, as well as data maintenance.

Mining methods are generally divided into machine learning methods, neural network methods, and database methods. Machine learning may be next divided into inductive learning, example-based learning, and genetic algorithms, etc. Neural network methods may be divided into feedforward neural networks and self-organizing neural networks, etc. Database methods mainly include multi-dimensional data analysis or OLAP (On-Line Analytical Processing), as well as attribute-oriented inductive method.

Various data mining algorithms have been developed, including artificial intelligence, machine learning, mode identification, statistics and database community, etc. In 2006, The IEEE International Conference on Data Mining Series (ICDM) identified ten most influential data mining algorithms through a strict selection procedure [2], including C4.5, k-means, SVM, Apriori, EM, Naive Bayes, and Cart, etc. These ten algorithms cover classification, clustering, regression, statistical learning, association analysis, and linking mining, all of which are the most important problems in data mining research. In addition, other advanced algorithms such as neural networks and genetic algorithms can also be applied to data mining in different applications. Some prominent applications are gaming, business, science, engineering, and supervision, etc.

5.2 Big Data Analytic Methods

In the dawn of the big data era, people are concerned with how to rapidly extract key information from massive data so as to bring values for enterprises and individuals. At present, the main processing methods of big data are shown as follows.

- *Bloom Filter*: Bloom Filter is actually a bit array and a series of Hash functions. The principle of Bloom Filter is to store Hash values of data other than data itself by utilizing a bit array, which is in essence a bitmap index that uses Hash functions to conduct lossy compression storage of data. It has such advantages as high space efficiency and high query speed, but also with some disadvantages like having a certain misrecognition rate and deletion difficulty. Bloom Filter applies to big data applications that allow a certain misrecognition rate.

Table 5.1 Comparison of MPI, MapReduce and Dryad

	MPI	MapReduce	Dryad
Deployment	Computing node and data storage arranged separately (Data should be moved computing node)	Computing and data storage arranged at the same node (Computing should be close to data)	Computing and data storage arranged at the same node (Computing should be close to data)
Resource management/ scheduling	–	Workqueue(google) HOD(Yahoo)	Not clear
Low level programming	MPI API	MapReduce API	Dryad API
High level programming	Null	Pig, Hive, Jaql, ⋯	Scope, DryadLINQ
Data storage	The local file system, NFS, ⋯	GFS(google) HDFS(Hadoop), KFS Amazon S3, ⋯	NTFS, Cosmos DFS
Task partitioning	User manually partition the tasks	Automation	Automation
Communication	Messaging, Remote memory access	Files(Local FS, DFS)	Files, TCP Pipes, Shared-memory FIFOs
Fault-tolerant	Checkpoint	Task re-execute	Task re-execute

- *Hashing*: it is a method that essentially transforms data into shorter fixed-length numerical values or index values. Hashing has such advantages as rapid reading, writing, and high query speed, but a sound Hash function is hard to be found.
- *Index*: index is always an effective method to reduce the expense of disc reading and writing, and improve insertion, deletion, modification, and query speeds in both traditional relational databases that manage structured data, and technologies that manage semi-structured and unstructured data. However, index has a disadvantage that it has the additional cost for storing index files and the index files should be maintained dynamically according to data updates.
- *Triel*: also called trie tree, a variant of Hash Tree. It is mainly applied to rapid retrieval and word frequency statistics. The main idea of Triel is to utilize common prefixes of character strings to reduce comparison on character strings to the greatest extent, so as to improve query efficiency.
- *Parallel Computing*: compared to traditional serial computing, parallel computing refers to utilizing several computing resources to complete a computation task. Its basic idea is to decompose a problem and assign them to several independent processes to be independently completed, so as to achieve co-processing. Presently, some classic parallel computing models include MPI (Message Passing Interface), MapReduce, and Dryad. A qualitative comparison of the three models is presented in Table 5.1.

Although the parallel computing systems or tools, such as MapReduce or Dryad, are useful for big data analysis, they are low levels tools that have a steep learning curve. Therefore, some high-level parallel programming tools or languages are being developed based on these systems. Such high-level languages include Sawzall, Pig, and Hive used for MapReduce, and Scope and DryadLINQ used for Dryad.

5.3 Architecture for Big Data Analysis

Due to the wide range of sources and variety, different structures, and the broad application fields of big data, different analytical architectures shall be considered for big data with different application requirements.

5.3.1 Real-Time vs. Offline Analysis

Big data analysis can be classified into real-time analysis and off-line analysis according to the real-time requirement. Real-time analysis is mainly used in E-commerce and finance. Since data constantly changes, rapid data analysis is needed and analytical results shall be returned with a very short delay. The main existing

architectures of real-time analysis include (a) parallel processing clusters using traditional relational databases, and (b) memory-based computing platforms. For example, Greenplum from EMC and HANA from SAP are all real-time analysis architectures.

Offline analysis is usually used for applications without high requirements on response time, e.g., machine learning, statistical analysis, and recommendation algorithms. Offline analysis generally conducts analysis by importing big data of logs into a special platform through data acquisition tools. Under the big data setting, many Internet enterprises utilize the offline analysis architecture based on Hadoop in order to reduce the cost of data format conversion and improve the efficiency of data acquisition. Examples include Facebook's open source tool Scribe, LinkedIn's open source tool Kafka, Taobao's open source tool Timetunnel, and Chukwa of Hadoop, etc. These tools can meet the demands of data acquisition and transmission with hundreds of MB per second.

5.3.2 Analysis at Different Levels

Big data analysis can also be classified into memory level analysis, Business Intelligence (BI) level analysis, and massive level analysis, which are examined in the following.

- *Memory-Level*: Memory-level analysis is for the case when the total data volume is within the maximum level of the memory of a clusters. The memory of current server cluster surpasses hundreds of GB while even the TB level is common. Therefore, an internal database technology may be used and hot data shall reside in the memory so as to improve the analytical efficiency. Memory-level analysis is extremely suitable for real-time analysis. MongoDB is a representative memory-level analytical architecture. With the development of SSD (Solid-State Drive), the capacity and performance of memory-level data analysis has been further improved and widely applied.
- *BI*: BI analysis is for the case when the data scale surpasses the memory level but may be imported into the BI analysis environment. Currently, mainstream BI products are provided with data analysis plans supporting the level over TB.
- *Massive*: Massive analysis for the case when the data scale has completely surpassed the capacities of BI products and traditional relational databases. At present, most massive analysis utilize HDFS of Hadoop to store data and use MapReduce for data analysis. Most massive analysis belongs to the offline analysis category.

5.3.3 Analysis with Different Complexity

The time and space complexity of data analysis algorithms differ greatly from each other according to different kinds of data and application demands. For example, for applications that are amenable to parallel processing, a distributed algorithm may be designed and a parallel processing model may be used for data analysis.

5.4 Tools for Big Data Mining and Analysis

Many tools for big data mining and analysis are available, including professional and amateur software, expensive commercial software, and free open source software. In this section, we briefly review the top five widely used software, according to a survey of "What Analytics, Data mining, Big Data software you used in the past 12 months for a real project" of 798 professionals made by KDNuggets in 2012 [3].

- *R* (30.7 %): R, an open source programming language and software environment, is designed for data mining/analysis and visualization. While compute-intensive tasks are executed, code programmed with C, C++, and Fortran may be in under the R environment. In addition, skilled users may directly call R objects in C. R is a realization of the S language. S is an interpreted language developed by AT&T Bell Labs and used for data exploration, statistical analysis, and drawing plots. Initially, S was mainly implemented in S-PLUS, but S-PLUS is a commercial software. Compared to S, R is more popular since it is open source. R ranks top 1 in the KDNuggets 2012 survey. Furthermore, in a survey of "Design languages you have used for data mining/analysis in the past year" in 2012, R was also in the first place, defeating SQL and Java. Due to the popularity of R, database manufacturers such as Teradata and Oracle both released products supporting R.
- *Excel* (29.8 %): Excel, a core component of Microsoft Office, provides powerful data processing and statistical analysis capability, and aids decision making. When Excel is installed, some advanced plug-ins, such as Analysis ToolPak and Solver Add-in, with powerful functions for data analysis are also integrated but such plug-ins can be used only if users enable them. Excel is also the only commercial software among the top five.
- *Rapid-I Rapidminer* (26.7 %): Rapidminer is an open source software used for data mining, machine learning, and predictive analysis. In an investigation of KDnuggets in 2011, it was more frequently used than R (ranked Top 1). Data mining and machine learning programs provided by RapidMiner include Extract, Transform and Load (ETL), data pre-processing and visualization, modeling, evaluation, and deployment. The data mining flow is described in XML and displayed through a graphic user interface (GUI). RapidMiner is written in Java. It integrates the learner and evaluation method of Weka, and works with R. Functions of Rapidminer are implemented with connection of processes of

operators. The entire flow can be deemed as a production line of a factory, with original data input and model results output. The operators can be regarded as specific functions and feature different input and output characteristics.

- *KNIME* (21.8 %): KNIME (Konstanz Information Miner) is a user-friendly, intelligent, and open-source-rich data integration, data processing, data analysis, and data mining platform [4]. It allows users to create data flows or data channels in a visualized manner, to selectively run some or all analytical procedures, and provides analytical results, models, and interactive views. KNIME was written in Java and, based on Eclipse, provides more functions as plug-ins. Through plug-in files, users can insert processing modules to files, pictures, and time series, and integrate them into various open source projects, e.g., R and Weka. KNIME controls data integration, cleansing, conversion, filtering, statistics, mining, and finally data visualization. The entire development process is conducted under a visualized environment. KNIME is designed as a module-based and expandable framework. There is no dependence between its processing units and data containers, making them adaptive to the distributed environment and independent development. In addition, it is easy to expand KNIME. Developers can effortlessly expand various nodes and views of KNIME.
- *Weka/Pentaho* (14.8 %): Weka, abbreviated from Waikato Environment for Knowledge Analysis, is a free and open-source machine learning and data mining software written in Java. Weka provides such functions as data processing, feature selection, classification, regression, clustering, association rule, and visualization, etc. Pentaho is one of the most popular open-source commercial intelligent software. It is a BI kit based on the Java platform. It includes a web server platform and several tools to support report, analysis, chart, data integration, and data mining, etc., all aspects of BI. Weka's data processing algorithms are also integrated in Pentaho and can be directly called.

References

1. Theodore Wilbur Anderson, Theodore Wilbur Anderson, Theodore Wilbur Anderson, and Theodore Wilbur Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958.
2. Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2008.
3. What analytics, data mining, big data software you used in the past 12 months for a real project? http://www.kdnuggets.com/polls/2012/analytics-data-mining-big-data-software.html, 2012.
4. Michael R Berthold, Nicolas Cebron, Fabian Dill, Thomas R Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. *KNIME: The Konstanz information miner*. Springer, 2008.