Information Technology Journal, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.

Research on the Method of Big Data Analysis

H.F. Qin and Z.H. Li
Department of Computer Science, ChuXiong Normal University, China

Abstract: With the development of society, the relational database facing to the great opportunities and challenges, how to store big data, analysis big data is become a hot issue. This article from the traditional data analysis start, find out the traditional data analysis situation and the trend of data analysis. Big data is facing a lot of issues, such as architecture, analysis technical, storage, privacy and security. Due to the method of analysis, the article mainly introduced to the structure data analysis, uncertainty analysis and others. The aim is for the study of big data prepare. To improve the ability of big data analysis and let the big data service for us.

Key words: Big data, analysis architecture, storage, uncertainty data analysis

INTRODUCTION

March 22, 2012, Obama government announced the launch of "Big Data Research and Development Initiative", this is in 1993 the United States announced the "information superhighway" program of science and technology after again, is also a sign to the rise of big data research never had height. With the appearance of the Internet, there is no limit of time, place and object communication between people. With the emergence of the Internet of things, the people and the things communication becomes more and closer, the diversification of communication has become more and more complex. Take a look at a friend, a day of traffic, SMS and micro blogging, you will know how much data in a day in the world? The data growth at 50% speed, two years doubled, like a piece of paper, the number of times fold you can arrive to the moon, the growth of the data is out our imagination (Lin, 2012).

Big data is more lager and complex, the common method of data analysis, software and tool can't meet the people needs. The big data face to a lot of issue, such as management, storage, search, analysis and so on. The characteristics of big data can be summarized as 4 V, Volume, Variety, Velocity and Velocity (Li and Cheng, 2012). The value of the large data is huge, often need to mass of data mining, the user value will become found. How to remove the noise data from big data and mining, extract valuable information is big data analysis solving the problem.

As to the problem of big data, the article from the situation of the big data start, find out the problem of big data facing and introduce some method of data analysis. The aim of study is find out the problem of big data, introduce some technology in the big data, for the study of big data prepare.

The situation of the big data analysis: Data analysis is use appropriate statistical methods to analysis the large number of first-hand information and secondary data. In order to using data maximize and play the role of the data. Research data in detail and summary, extract useful information and form the conclusion. The data is also known as the observation, is the result of the experiment, measurement, observation, survey, appearance quantity form. Big data for the system with a very high demand, big data platform is consisting of technologies of computing, transmission, storage, interaction. Now, the technology of the data center is difficult to meet the demands of big data. The architecture of entire IT will be developed. Storage capacity growth lags far behind the growth of the data, design the rational architecture has become a key in information system.

The traditional data analysis problems: The appearance of big data let the traditional relational database encountered unprecedented impact. In the traditional database applications, the pursuit of the relational model of data management is the consistency and accuracy of the data. For big data analysis needs, longitudinal scale up system, that is, through increased or replace CPU, memory, hard disk to expand the ability of a single node, will encounter the bottleneck; horizontal scale out system, that is, through the increase computing nodes connected into clusters and rewrite the software, the cluster in the parallel execution, is the economical solution. For the big analysis needs, some people think out a lot of methods. Some people from longitudinal scale up, according to

Corresponding Author: H.F. Qin, Department of Computer Science, ChuXiong Normal University, China

increase and replace hardware, such as CPU, memory, hard disk. But the method is expanding the ability of a single node, so it will encounter the bottleneck. Another people from horizontal scale out, according to improve the ability of software, such as rewrite the software, increase computing nodes, cluster and parallel execution. The method is complex but is an economical solution. Adopting larger-scale cluster manage and analyst big data require face to variety challenger, the availability of the system is in key role (Schroeder and Gibson, 2007).

The traditional analysis method is put data into the database, according to the data collation, cleaning, analysis and meet to OLAP analysis require. For the big data import from a file system into the database needs high hardware support, for the need of analysis, sometimes, Analysts must to import the data into the analysis software SAS or SPSS to analyze, software requirements are higher, the analyst had wanted to more than 70,000 records imported into SPSS for analysis in 2005 but did not succeed, because using the SPSS software version is too low, the data is too lager. Finally, import the data into SAS for analysis, although a success but the efficiency is a rather low and store into EXCLE file, a single file cannot be stored, the analysis software cannot analysis direct in database. Facing big data at present, the analyst cannot success trying to store large data into the database and move the data from the database into the analysis software, it is unrealistic. For unstructured or semi-structured data, relational database is also unable to play its efficiency.

Facing telecommunications millions of data analysis, the analyst adopting data warehouse, construct the analysis theme-oriented data cube. According to the data aggregation, analysis, from the upper layer, find the theme analyze and analysis of the variables. The main method is adopting OLAP data on exploration, drilling down; identify the theme, for this theme, layers and layers analysis. Finally, as to the data analyze deeply. This purpose just determine subject and reduced the data but it cost lager time and this method relies on the database, data warehouse and data analysis software, as to the hardware is high. Facing to the current data, this analysis is small. Had also trying to use the high-dimensional clustering solve this but existing a lot of problems.

The data analysis trend: Compared to traditional data analysis, big data analysis has a large amount of data, query and analysis complex. To discover knowledge from data and guide people's decisions, it is necessary to analysis the data deeply, rather than generate reports only. Complex analysis must depend on the model of analysis complexity, it is difficult to use SQL to express and it is deep analysis (Qin *et al*., 2012). A typical OLAP data analysis operations (data gathering, aggregating, slicing and rotation) was not enough, relational database analysis, also need to path analysis, time series analysis, graph analysis, What-if analysis, as well as hardware/the software restriction never tried complex statistical analysis model (Li and Cheng, 2012).

Big data analysis are faced with the problem: Big data is with a vocabulary of culture gene and marketing concepts but also reflects the trend in field of science and technology, this trend opens the door to understanding the world and making decisions. The data has not only the traditional structured and unstructured data in big data era. Data processing is not entirely the issue of the size of the data but did not appear for the unstructured approach which is caused by the concept of large data reasons. Structured data depend on processes, enterprise deal with data according processing, data is not driven process but a process-driven data and data are dependent processes. But the information processing structure and the application of non-structural data is not such a way. Big data handling requirements are completely different. Analysts are not going to a complete reversal of data management and processing system but add new processing mode in the original basis, form more perfect more complete system.

According to data mining, collation, analysis, use and realize the data value as to big data, is a hot issue in the industry. The appearance of big data makes many fields toward discovery knowledge and decision direction change in the data driven. The appearance of big data to our traditional data analysis has brought great challenges. It also contains structure, storage, analysis technology, security and privacy.

Big data analysis architecture: In traditional data analysis method, collect data and processing data is through the database and even preliminary analysis can through the database joint data warehouse to solve. The purpose of constructing data warehouse is oriented to the analysis of integrated data environment for enterprises, to provide Decision Support. In fact, the data warehouse itself is not “production” any data and at the same time, it not need “consumption” any data. Data from the outside and open to external application, this is reason why called “warehouse”, not that “factory”. So the basic architecture of data warehouse mainly contains the data inflow and outflow, can be divided into three layer-source data, data warehouse, data applications. In the past ten years, data warehouse and database connection can be regarded as seamless, relational database management

system is also provide interface to the customer, for instance SQL SEVER, whether SQL SEVER 2000, SQL SEVER 2005, or SQL SEVER2008, provide kinds of service for people's decision provides as convenient as possible, such as analysis service and reporting services. Relational database architecture easy is to understand and analysis. So it is very popular.

However, the emergence of big data, the people is facing challenge on relational databases variously. How to analyze and store data become wanting to solve the problem in numerous data analysis. Whether it should be negative the modal and schema are built so many years, that it is should be completely negative database, data warehouse, I think all is worth careful thinking. The data warehouse is born to the data analysis; meet the people needs diversely. But in the era of big data, adopting previous database, data warehouse architecture is in troubles. Wang *et al.* (2011) lists some key features for big data analytic. summarizes current main implementation platforms, parallel databases, Map Reduce and hybrid architectures.

But facing big data, each analysis platform is not perfect, has a long way to go. Large data analysis forces us to reflect on the architecture of traditional data warehouse, with an open mind and to stand in a higher level to thinking, study such as Map Reduce new platform, so as find out the architecture that analysis and application.

The technical of the big data analysis: Due to the special nature of data, data analysis technology is in the development stage still, the old technology is maturing and the new technology will be more. Mainly in the visual analysis, data mining algorithms, predictive analysis, semantic engines, data quality and data management.

Data visualization is the most basic functions for normal users and data analysis expert. Data visualization can let the data speak for themselves, allow users intuitive feel results.

Visualization is a machine language translation to see and data mining is machine language. Let segmentation, clustering, isolated point analysis as well as a variety a variety of algorithms refined data mining value. These algorithms must be able to cope with the large amount of data and also has a very high processing speed.

Data mining allows analysts to understand faster and better and thus enhance the accuracy of the judgments of data carrying information and predictive analysis allows the analyst to make some forward-looking judgments based on the results of image analysis and data mining.

Diversification of unstructured data has brought new challenges to the data analysis, we need a set of tools to analyze and refine the data. The semantic engine needs to be designed to have enough artificial intelligence to be sufficient to take the initiative to extract information from the data.

Data quality and management is the management of the best practice, through standardized process and machine processing of data can ensure that you get a default quality analysis results.

OLAP technology: In recent years, data analysis, according to the entire OLAP, identify factors and data mining partial deeply.

The problem of OLAP, business agile changeable, inevitably leads to business model then change constantly. Business dimensions and measure change once, technical personnel need to put the whole Cube (multidimensional Cube) to define and to generate. Business personnel, only in the Cube on the multidimensional analysis, limit the problem rapid changed, so that the so-called BI system becomes rigid daily report system.

OLAP analysis requires a large number of data segment and the relationship between tables which is obviously not No-SQL and traditional database strengths, often must use a specific database optimized for BI. Example, the vast majority of database optimized for BI using the column is stored or mixed storage, compression, lazy loading, statistics, pre-stored data block slice index technology.

MapReduce technology: In 2004, Google Company (Dean and Ghemawat, 2004) Ghemawat first to take advantage of the MapReduce (Almeida and Calistru, 2012) technologies, clustering solve the problem of large data processing as a parallel computing model for data analysis and processing. MapReduce is not compatible with the basic existing BI tools. Because in its design is not to become a database system, so it does not provide a SQL interface. It has been committed to the SQL statement and MapReduce tasks conversion work (hive) and thus the possible of MapReduce with existing BI tools compatible. Data scientists are exploring a new road outside the traditional database and business intelligence tools, Hadoop MapReduce as a data analysis of the most powerful tools of the era of big data.

Hadoop is a special deal with large data technology, particularly in unstructured data, such as communication, Web applications, text, applications, network and security log data, etc. There are many open source technologies; the scope of the BI architecture has played a positive role in the expansion. Including BI itself, data integration, data profiling, data modeling, data cleansing and master data

management tool. Open source software development speed compared to traditional commercial software can be said to be fast a lot, because there are a large number of users as a basis, you need to continue to learn new knowledge, in order to be able to use them to design, develop and deploy your BI system.

OLAP analysis on Hadoop platform, this problem also exists, the Facebook for the Hive development of RCFile data format, that is, above some optimization techniques to achieve better performance.

However, for Hadoop platform, only through the use of the Hive imitate the SQL, for data analysis is far from enough, the first Hive although will Hives translation when MapReduce optimized but still low efficiency. Multidimensional analysis is to do the fact table and dimension table association; more than one dimension performance will decline sharply. Secondly, RCFile ranks hybrid storage mode, in fact limit data format, that is to say, the data format for a particular analysis advance design well, once the analysis of business model varies, mass data format conversion cost is very huge. Finally, HiveQL for OLAP business analyst is still very unfriendly, dimension and measurement is the direct business personnel analysis language.

Data mining technology: Big data is as important as natural resources, human resources, strategic resources, is a national digital sovereignty reflect. “Big data” is an opportunity and a challenge in a society. It will put all aspects requirements to limit the information management. The large amount of data is more convenient to discover the laws, such as probability in statistics about coin, you only throw a few do not see it the law but if throwing more, then both sides of probability is similar. One discipline alone to solve the problem for the vast amounts of unstructured data is unrealistic to rely on multi-discipline integrated. Data mining is a combination of machine learning, statistics and database technology interdisciplinary. From the database perspective, data mining is the process of discovering knowledge from large amounts of data stored in the database, data warehouse or other repository. From the machine learning perspective, data mining is extracted from the data implied in which people do not know in advance but potentially useful information and knowledge of the process. From the statistical perspective, data mining is to analyze the data sets (often a lot of) to find unexpected relationships and presented to the user of the data understandable and useful process. Comprehensive data mining perspective of the three disciplines, use database solve the OLTP (online online processing); use machine learning to solve the problem of data reduction; use the statistical point of view to predict. But all are from data analysis is comprehensive statistical analysis, machine learning, artificial intelligence, database and many other aspects of the research.

The goal of data mining is sometimes stated generically as estimating ‘‘useful’’ models from data and this includes, of course, predictive learning and statistical model estimation. However, in a more narrow sense, many data mining algorithms attempt to extract a subset of data samples (from a given large data set) with useful (or interesting) properties. This goal is conceptually similar to exploratory data analysis in statistics (Hand 1998; Hand *et al*., 2001), even though the practical issues are quite different due to huge data size that prevents manual exploration of data (commonly used by statisticians). There seems to be no generally accepted theoretical framework for data mining, so data mining algorithms are initially introduced (by practitioners) and then ‘‘justified’’ using formal arguments from statistics, predictive learning and information retrieval. There is a significant overlap between these methods.

Data mining is important, whether Internet of things or wisdom city. Data mining as an emerging interdisciplinary application fields, are all walks of life decision support activities play a more and more important role. Data privacy and protection is an important issue in the data mining.

Deadlock problem: Data analysis but also solve the problem of instantaneous concurrent deadlock. Solve instantaneous deadlock is a large data analysis requirements of real-time nature of the necessary and sufficient conditions. A set of processes is deadlocked when each process in the set is blocked awaiting an event (typically the freeing up of some requested resource) that can only be triggered by another blocked process in the set. There are three common approaches to dealing with deadlock: prevention, detection and avoidance. An approach to solving the deadlock problem that differs subtly from deadlock prevention is deadlock avoidance. Deadlock prevention, peoples constrain resource requests to prevent at least one of the four conditions of deadlock. This is either done indirectly, by preventing one of the three necessary policy conditions (mutual exclusion, hold and wait, no preemption), or directly by preventing circular wait. This leads to inefficient use of resources and inefficient execution of processes. On the other hand, Deadlock avoidance, allows the three necessary conditions but makes judicious choices to assure that the deadlock point is never reached. As such, avoidance allows more concurrency than prevention. With deadlock avoidance, a decision is made dynamically whether the

current resource allocation request will, if granted, potentially lead to a deadlock. Deadlock avoidance thus requires knowledge of future process resource requests.

In this section, mainly describe two approaches to deadlock avoidance:

- Do not start a process if its demands might lead to deadlock
- Do not grant an incremental resource request to a process if this allocation might lead to deadlock

Big data storage problems: With the explosive growth of data applications, it has spawned its own unique architecture but also directly promotes the development of storage, networking and computing technologies. The development of hardware, storage capacity growth lags far behind the growth of the data; reasonable storage architecture design has become the key to information systems. Increase capacity by adding modules or disk cabinet, or even no downtime. Currently, customers are increasingly popular Scale-out architecture storage. Scale-out cluster structure is each node has their characters. It having a certain storage capacity, owns a data processing capability and the interconnection equipment. The architecture is completely different from traditional storage system. It realizes seamless smooth extensions to avoid storage silos for the storage of large data. "Big data" applications exists a real-time problem. Especially with regard to online trading or financial applications, Scale-out storage system architecture can play the advantage, because each node has a processing and interconnection components, can also increase capacity while processing capabilities simultaneously increased. Object-based storage systems, it is possible to support concurrent data streams, thereby further improving data throughput.

Distributed storage and computing architecture allows processing of large amounts of data in a reliable, efficient and scalable way. Work in a parallel way, the data processing speed is relatively fast and low cost, Hadoop and NoSQL belong to the category of distributed storage technology.

Somebody has proposed to management using data lake or data pool. But the data pool more data-oriented exploration and discovery, rather than the traditional business intelligence reporting and analysis, Evelson added, brought a vicious cycle: data cannot be managed until it is modeled after the data analysis but it is essential to be modeled.

Data management program provides a framework for setting data use policy and implement control to ensure that information to keep accurate consistent and can be accessed. Clearly, in the major challenges in the process, management big data to classification, modeling and data mapping and data capture and storage, especially for a large number of unstructured characteristic information.

Other problem: Big data analysis often need more class data reference each other, in the past does, not have this kind of data hybrid access, so big data applications also gave birth to the new, need to consider safety problem.

The first challenge appears in terms of privacy. The privacy is the most sensitive issue, with conceptual, legal and technological implications. This concern increases its importance in the context of big data.

Another challenge, indirectly related with the previous, is the access and sharing of information.

Considering the size issue, analysts also know that the larger the data set to be processed, the longer it will take to analyze.

Finally, working with new data sources brings a significant number of analytical challenges.

The relevance and harshness of those challenges will vary depending on the type of analysis being conducted and on the type of decisions that the data might eventually inform. The big core challenge is to analyze what the data is really telling us in a fully transparent manner. The challenges are intertwined and difficult to consider in isolation but according to King and Powell (2008), they can be split into three categories: (a) getting the picture right (i.e., summarizing the data) (b) interpreting or making sense of the data through inferences and (c) defining and detecting anomalies.

Data analysis methods: Big data analysis is a complicated problem and the traditional analysis of data, big data must also be combined with field. According to different field and different application, use different analysis methods analysis and interpret the results in different field and different professional knowledge. Sometimes, according to the data characteristics and business characteristics analysis and interpret. For big data analysis, such as classification and application of technical requirements, the people can from the demand analysis of building business model start. First, find out the data model. Second, analyze data model, from data acquisition, data sorting, data cleaning, data integration, data analysis to data feedback is a model analysis process.

Statistical model estimation, based on extending a classical statistical and function approximation framework (rooted in a density estimation approach) to developing flexible (adaptive) learning algorithms (Ripley, 1994; Hastie *et al*., 2001). The classical approach, as proposed by Fisher (1952), specification and estimation.

Specification consists in determining the parametric form of the unknown underlying distributions, whereas estimation is the process of determining parameters of these distributions. Classical theory focuses on the problem of estimation and sidesteps the issue of specification (Cherkassky and Mulier, 2007). It include Density Estimation, Classification, Regression, Solving Problems with Finite Data, Nonparametric Methods and Stochastic Approximation.

Data warehouse is a database used for reporting and data analysis. It is a central repository of data which is created by integrating data from one or more disparate sources. Data warehouses store current as well as historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons. The data warehouse is the heart of the architected environment and is the foundation of all DSS processing.

Data mining methodology is a diverse field that includes many methods developed under statistical model estimation and predictive learning. There exist two classes of data mining techniques, that is, methods aimed at building ''global'' models (describing all available data) and ''local'' models describing some (unspecified) portion of available data (Cherkassky and Mulier, 2007).

Certainty data analysis: Traditional deterministic data management technology has been greatly development. According to the traditional database technology, establish the theme of analysis, extract, transform, load data, built theme-oriented data cube. Adopting OLAP analysis data cube, find out the theme analysis further data mining. In the data mining, there are some analysis will be adopt such as data preprocessing technology, properties Statute technology and so on. But for numeric data and non numeric data, must be using different analysis method to analyze it. For operational treatment is usually the OLTP technique, usually for one or a set of records of querying and modifying, fast respond to user request, the data security and integrity, the consistency of the things, transaction throughput, data backup and restore demanding. To analyze the type of data is usually the data warehouse and OLAP technology, requires a data source, data extraction, conversion and loading tools, data warehouse, OLAP server and statistical data analysis tool.

Uncertainty data analysis: The ubiquity of the idea of uncertainty is illustrated by the rich variety of words used to describe it and related concepts. Probability, chance, randomness, luck, hazard and fate are just a few examples. Uncertainty data causes more complicated. May be the original data itself is not accurate or using coarse granularity data set, also may be in order to meet the specific application purpose or in dealing with missing value, data integration process and produce:

- **The original data:** This is not accurate produce uncertainty data is the most direct factor. First of all, physical instrument collected data by the accuracy of the accuracy of the instrument restricted. Secondly, in the network transmission (especially the wireless network transmission) process, the accuracy of the data by bandwidth, transmission delays, energy etc factors. Also, in sensor network applications and RFID application and so on, the surrounding environment also affects the accuracy of the original data
- **Use coarse grain size data set:** Obviously, from the coarse grain size data set into fine grain size data collection process will introduce uncertainty. For instance, if the population distribution database to township for base unit records the national population and some application but request village-based unit of the population quantity, inquires the results there is uncertainty
- **To meet specific application purpose:** For privacy protection and other special purpose, some application unable to get the original accurate data but only can get the accurate data after transformation
- **Missing value:** Missing value causes a lot of, equipment failure, unable to get information and other fields, such as inconsistent historical reasons may produce missing value. A typical processing method is interpolation, interpolation after data can be regarded as obey certain probability distribution. In addition, also can delete all contain missing value records but also operate in a certain extent, change the original data distribution. For this part of the data, it must determine whether the data existence. If it must, It can re-investigation, man-made or mathematical method (regression, Bayesian, decision tree, the mean or global constants, etc.) to fill. Otherwise can ignore, when taking data can be limited conditions and did not take ignore data. In the analysis, fill the missing values is a complex and difficult work, so the database will adopt invalid value to solve it
- **Data integration:** Different data sources of data information may not be consistent; in data integration process will introduce uncertainty. For

instance, Web contains a lot of information but since the page updated, many factors such as the content of a page is not consistent

Other analysis: There are a lot of methods as to big data analysis, in the seasonal changes, time series analysis is a widely used quantitative analysis method, it is mainly used to describe and explore phenomenon over time the number of change regularity. Time series analysis is its development stage and the use of statistical analysis methods of it, there are the traditional time series analysis and modern time series analysis. The traditional methods are time series smoothing method. Simple average method, moving average, exponential smoothing method, these three methods referred to as the time series smoothing method. The basic idea is that: Apart from some irregular movements, the time series data of the past there is some kind of basic form, assuming this form will not change in the short term, can be used as the basis of the forecast of the next period. Smoothing the main purpose is to eliminate the extreme values of the timing data to the certain smoothing the intermediate value as the prediction is based.

In the era of big data, there are many data analysis methods, the author will research.

CONCLUSION

Big data has entered into our lives, for this big data, how to discovery knowledge from the data and create value for the enterprise, are many analyst eager to solve the problem. The text research the method of big data, however, due to the complexity of big data itself, the BI technology must be improved, the data has only just begun and the method of data analysis, many methods are just trying to, yet to be further valid.

REFERENCES

Almeida, F.L.F. and C. Calistru, 2012. The main challenges and issues of big data management. Int. J. Res. Stud. Comput., Vol. 2, 10.5861/ijrsc.2012.209.

Cherkassky, V. and F.M. Mulier, 2007. Larning from Data: Concepts, Theory and Methods. 2nd Edn., John Wiley and Sons, New York, USA., ISBN: 9780470140512.

Dean, J. and S. Ghemawat, 2004. MapReduce: Simplified data processing on large clusters. Proceedings of the 6th Conference on Symposium on Operating Systems Design and Implementation, December 06-08, 2004, San Francisco, CA., USA., pp: 10.

Fisher, R.A., 1952. Contributions to Mathematical Statistics. John Wiley and Sons, New York, USA.

Hand, D., H. Mannila and P. Smyth, 2001. Principles of Data Mining. MIT Press, Cambridge, MA.

Hand, D.J., 1998. Data mining: Statistics and more? Am. Statistician, 52: 112-118.

Hastie, T., R. Tibshirani and J. Friedman, 2001. The Elements of Statistical Learning: Data Mining, Inference and Prediction. 1st Edn., Springer-Verlag, Canada, ISBN: 0-387-95284-5, pp: 746.

King, G. and E.N. Powell, 2008. How not to lie without statistics. http://gking.harvard.edu/files/abs/nolie-abs.shtml.

Li, G. and X. Cheng, 2012. Research data and scientific thinking. Bull. Chinese Acad. Sci., 27: 647-656.

Lin, H., 2012. Thinking of large data China. Comput. Communi., (In Press).

Qin, X.P., H.J. Wang, X.Y. Du and S. Wang, 2012. Big data analysis competition and symbiosis of RDBMS and MapReduce. J. Software, 23: 32-45.

Ripley, B.D., 1994. Neural networks and related methods for classification. J. R. Stat. Soc. Ser. B, 56: 409-456.

Schroeder, B. and G.A. Gibson, 2007. Understanding failures in petascale computers. J. Physics, 78: 1-11.

Wang, S., H.J. Wang, X.P. Qin and X. Zhou, 2011. Architecting big data:Challenges, studies and forecasts. Chinese J. Comput., 34: 1741-1751.