Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 4, Issue 1, pp: (35-46), Month: January - April 2017, Available at: www.noveltyjournals.com

21st Century Present Condition and Challenges Related To Large-Scale Data Processing In Smart Grid and Optional Framework for Large Data Storage and Analysis

[1]Arslan Habib, [2]Hafiz Muhammad Hafeez

[1,2]Northwestern Polytechnical University, Xian 710072, China

Abstract: **Panoramic state (Analog or Digital) data is one of the most critical requirement for Smart grid operation, and during the management, operation and upkeeping of smart grid heavy heterogeneous and multi-state data, called the large (Big) data, are generated. In 21th century, how to store this data more reliably and more cheaply and data-approach and analyze them briskly are critical research subjects. In the first phase, the source of this large data generated in diverse process of smart grid, for example power generation, transmission, transformation and utilization and the elements of this data are analyzed. In second phase, the present large data processing methods adopted in the fields of industrial monitoring, internet and business are summarized, and the positive consequences and drawbacks of these methods in coping with the construction of smart grid and large data processing are analyzed in detail. Lastly, in the aspects of large data storage, real-time data operation and fusion of heterogeneous multi-data sources and visual resolving of large data, the chances and challenges brought by smart grid large data are clarified in detail.**

Keywords: **Parallel Database; Large/Big Data; smart grid; cloud computing.**

1. INTRODUCTION

Since last decade, with the global energy problems are increasingly grim, the world has launched a smart grid research [1]. The ultimate goal of the smart grid is to build a panoramic real-time system that covers the whole production process of the power system, including power generation, transmission, sub-station, distribution and dispatching [2]. The foundation for smart grid security, self-healing, green, robust and reliable operation is grid-wide real-time data acquisition, transmission and storage, as well as a rapid analysis of accumulated mass multi-source data. With the deepening of the smart grid construction and the advancement of power grid operation and apparatus inspection, the large amount of data is generated, and gradually constitute the information of today's concern to large data, which requires the appropriate storage and fast processing technology as support.

People need to study large data analysis techniques and theory to accumulate the massive multi-source heterogeneous data caused by the extensive application of cloud computing platform. Currently, the large data has become the research topic, which the academia and the industry pay close attention to [3]. It has obtained the application in many domains and has the immense application prospect. In 2009 alone, Google's contribution to the US economy through large data services was 54 billion USD, which is only a fraction of the cost of the large data [4]. Taobao company through a large sum of changes in transaction data analysis, can be predicted 6 months ahead of global economic trends. IBM uses up to 4PBs of climate and environmental history data to design a fan (wind turbine) location model and to determine the optimum location for fan installation, resulting in up surged fan productivity and prolonged service life [5].

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 4, Issue 1, pp: (35-46), Month: January - April 2017, Available at: www.noveltyjournals.com

In May 2011, McKinsey published "The Big Data: The Next Frontier for Competition, Innovation, and Productivity" [6], which sheds light on the status of large data research and its importance for the society. In 21th century, Data science has become the urgent need for social development and technological progress.

In the smart grid system, large data generated in the whole system at all stages. For example, in the electricity side, with the installation and deployment of a large sum of smart meters and smart terminals, the interaction between power companies and users. Time to time, these companies access to the consumer information, thus collecting the detailed mass of electricity consumption data more granular than ever, which adds to the user-side data [7]. By analyzing this data, we can better understand the electricity consumption behavior of power consumers [8], rationally design the power demand response system [9] and short-term load prediction system [10] and so on.

To understand and view the large data properly in grid system, it is necessary to provide useful references to show application status and future challenges for the construction of smart grid application. This paper attempts to summarize the research and application status and 21th century challenges of large data in smart grid, and presents an optional framework for large data storage and analysis.

2. SMART GRID DATA AND ITS CHARACTERISTICS

2.1 Large-Scale data in the smart grid: Power business data can be break downed into three categories: one is the grid operation and apparatus testing or monitoring data; the second is the power enterprise marketing data, such as trading price, electricity sales, electricity customers and other data; third power enterprise management data.

Depending on the intrinsic structure of the data, the data can be further subdivided into structured data and unstructured data. Structured data mainly includes the data stored in the relational database, most of the current power system data is in this form. With the development of information technology, this part of the data is growing very fast. Compared with structured data, the data that is inconvenient to use the 2-D logical table of the database is called unstructured data, including video surveillance, graphics, image processing and other data generated. According to a study administered by the Internet data center (IDC), this data is growing very fast, with 80% of the data in the enterprise being unstructured and growing exponentially by 60% annually [11]. In the power system, unstructured data accounts for an enormous portion of smart grid data.

According to processing time requirements, structured data can be divided into real-time data and quasi-real-time data, such as grid scheduling, and this real-time data needs to be controlled quickly and accurately; and a large sum of state monitoring data on real-time requirements are relatively low, can be used as quasi-real-time data processing.

The dissimilarity between smart grid and traditional power grid is a high level of intelligence, and intelligent premise is a large sum of real-time status data acquisition. The modern large-scale smart grid data is mainly due to the following aspects:

1) In order to get the running state information of the apparatus in real time, more and more points are collected. The conventional dispatch automation system containing hundreds of thousands of collection points and with electricity the data center will reach one million or even ten million level [12]. The number of appliances and instruments that need to be monitored is humongous, and each appliances and instruments are equipped with sensors, the monitoring device connects these sensors together through an appropriate communication channel and are uploaded to the data center by a data collection server at the substation in accordance with a unified communications standard. This actually constitutes a matter of networking. The back-end of things using cloud-computing platform has been considered the development trend of the future. Smart grid devices are interconnected with the infrastructure layer of the cloud computing plateform for data exchange.

2) In order to capture all kinds of state information and meet the needs of the application system, the sampling frequency of the apparatus is getting higher and higher. For example, in the state monitoring system of power transmission and transformation apparatus, in order to diagnose the condition such as the insulation discharge, the sampling frequency of the signal must be above 200 kHz (UHF detection needs GHz sampling rate). Thus, for a smart grid equipment-monitoring platform, the amount of data stored in the monitoring or testing is very large.

3) Record every detail of the production run for the real, complete reflection of the production process, required to achieve "real-time change sampling" [13].

In the smart grid, large data generated in all aspects of the power system, including:

1) Generation side. With the development of digital construction of large power plants [14], the vast amount of process data is saved. These data are rich in information, which is of great significance for analyzing the status of production operation, providing control and optimization strategies, defect diagnosis, and knowledge uncovering and data mining [15]. The data-driven defect diagnosis method has been proposed [16], which can fix the problem of defect diagnosis, optimization configuration and evaluation of the production process and apparatus, which cannot be fixed by the former method based on analysis and the monitoring method based on qualitative experience knowledge using massive process data. In addition, the real-time monitoring and control of a large sum of distributed energy sources need to be carried out in order to grasp the equipment and operation state of distributed power supply in a tolerable and accurate manner [17]. To support wind turbine siting optimization, the collected weather data for modeling is growing at 80% per day [5].

2) The power transmission side. In 2006, the US Department of Energy and the Federal Energy Commission recommended the installation of a synchronous phasor monitoring system. At present, United States uses 100 phase measurement devices (phasor measurement unit, PMU) to collect 6.2 billion data points a day, the data is about 60 GB, if the monitoring device is increased to 1 000 sets then data points collected per day will be 41.5 billion, the amount of data will 402 GB [18]. Phasor monitoring is only a small part of smart grid monitoring.

3) Electric Power side. To accurately obtain the consumer's electricity data, the power company has deployed a large sum of smart meters with the capability of two-way communication. These meters can send electricity information to power grid in every 5 minute. Pacific Gas & Electric collects more than 3TB of data each month from nine million smart meters [19]. Disordered charge and discharge behavior of meters will bring trouble to the grid operation, if we can reasonably arrange the smart meters charging and discharging time, it will bring advantages to the grid. Along advantages, large state of monitoring for these meters will produce large data.

2.2 Characteristics of Large data in the smart grid: The large data in the smart grid has "4V" characteristics, namely 1. Large Volume, 2.multi type variety, 3. Low value density, 4. Fast change.

1) The data volume is humongous. From the TB level, jumped to PB level. Conventional SCADA System 10,000 telemetry points, generated at a sampling interval of 3-4 s, producing 1.03 TB of data per year (1.03 TB = 12 bytes / frame □ 0.3 frames / s □ 10,000 Telemetry point □ 86400 s / day 365 days), and 10,000 telemetry of the Wide Area Phasor Measurement System (WAMS) Point, the sampling rate can reach 100 times / s, according to the above formula, the annual production of 495 TB of data [13].

2) Numerous data types. There are sorts of structured and semi-structured data, such as real-time data, historical data, text data, multimedia data and time series data, as well as unstructured data. Numerous categories of data query and processing frequency and performance requirements are not the same. For example, oil chromatogram data in power plant condition monitoring is sampled once every 0.5 h, while insulation discharge data is sampled at rates of several hundred kHz or even GHz.

3) Low value density. In video, for example, the continuous monitoring process, the data may be useful only 1-2 s. The same problem exists in the condition monitoring of power transmission and transformation apparatus. Most of the data collected are normal data, only a very tiny amount of abnormal data and this abnormal data is the most important basis for condition-based maintenance.

4) Processing speed. Analyze large amounts of data in a small part of second to support decision-making. On-line state data processing performance requirements are much higher than offline data. This on-line stream data analysis and mining are fundamentally different from traditional data mining capacity and skills [20].

In addition, the data processing in the smart grid has certain requirements on the quality of the data, which can be considered as a new attribute for all kinds of smart grid data. The authenticity of the data assigns to the reliability level associated with the particular type of data [5]. High quality data has an important impact on the correctness of data analysis results. Nevertheless, even the finest data cleaning methods cannot remove the deep-rooted unpredictability of certain data. It is feasible to acknowledge the uncertainty requirements and to use the authenticity of the data as a dimension of smart grid large data.

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 4, Issue 1, pp: (35-46), Month: January - April 2017, Available at: www.noveltyjournals.com

Turbulent large data in smart grid has brought new challenges and opportunities in smart grid construction. China Xintong Company (State Network ICT Corporation) established the large data team to deal with the construction of smart grid in the large data problem [21]. IBM collects and model large data, serving a variety of energy industries and utilities such as smart meter analysis, decision-based operation and maintenance, wind turbine location based on weather data, load forecasting and dispatch, and so on [5].

3. LARGE DATA PROCESSING TECHNOLOGY

3.1 The value and complexity of large data processing: In present years, large data has become a common concern of the scientific and technological circles and industry. March 2012, the US government announced 200 million USD investment to start the "big data research and development plan". For the US government, the large data is "The future of new oil," and "Big data research" will have a profound impact on the future of science and technology and economic development. The ability of a country to have the large size of the data and the ability to use it will be an important part of the overall national strength, and the possession and control of data will also become a new focus of competition between countries and enterprises [3].

The present global data storage and processing capacity has lagged far behind the data growth rate. For example, Taobao daily increase in transaction data is up to 10 TB; eBay analysis platform daily processing data is up to 100 PB, More than the US NASDAQ stock market all-day data processing capacity; WAL-MART is one of the first companies to use large data analysis and thus benefit, has created a classic business case called "Beer and Diapers". Now Wal-Mart deals with one million transactions per hour, with about 2.5 petabytes of data in the database, which is 167 times that of the Library of Congress; Microsoft spent 20 years and spends millions of dollars on Office Spell Checking function, Google use large data directly to achieve statistical analysis.

Compared with the extensive research and application of large data in the field of Commerce and the Internet, large data in the smart grid construction research needs to be further strengthened. Because the cloud-computing platform has the advantages of large storage capacity, low cost, high reliability and high scalability, it is not appropriate for the main system of power dispatching and automation, but can be used in the background of dispatching automation system, and can also be used in Smart Grid Data Center (marketing, management and apparatus condition monitoring). Under the environment of cloud platform, large data processing and display tools are emerging; in order to diminish the software development work has brought benefits. Nevertheless, data mining is usually associated with a specific application object, and large data mining is not a small challenge. Such as the initial screening of defected data [22] and some other applications based on clustering methods, in the face of massive data, traditional clustering algorithms cannot be completed on ordinary computing system. In addition, data processing is facing large-scale challenges, while diversification of data processing needs gradually revealed. In contrast to the data processing services that support a single service type, the large data to be processed by the common data processing platform involves a variety of online / offline, linear / non-linear, flow data and graph data and other complex hybrid computing. The following is the summary of present mainstream data processing technologies and pointing out the limitations of these technologies to explore possible solutions when dealing with large-scale smart grid data.

3.2 Parallel database: Relational databases (such as Oracle, etc.) mainly store structured data, provide appropriate and beneficial data query and analysis capabilities, the capability to quickly deal with transactions, concurrent access to multiple consumers and data security based on strict rules. It is extensively used because, through the SQL query (Structured Query Language) language and powerful data analysis capabilities and high procedural and data independence and other advantages. Nevertheless, with the acceleration of the construction of smart grid, data has been far beyond the management of relational databases, geographic information systems, as well as pictures, audio and video and other unstructured data has become an important part of the large data that needs to be stored and processed. The relational database for structured data storage cannot satisfy the requirement of large-scale data access and large-scale data analysis. Mainly in the:

1) Limited data storage capacity. Relational database can effectively deal with TB data, but when the amount of data reaches the PB level then mainstream database is difficult to deal with. To avoid this problem, the present power enterprises use "Raw data" to extract the "Cooked data" storage way, this allows you to diminish the amount of data that is transmitted by the network and the database, but inevitably loss of "Raw data" of crucial information, such as the discharge spectral of the insulation.

2) Fast access capability of relational model to large data. The relational model is a content-based model [23]. That is, in the traditional relational database, this locates the corresponding line according to the value of the column. This kind of access model introduces the time-consuming input and output in the process of data access, which affects the ability of fast access. Although, the traditional database system can be partitioned (horizontal partition and vertical partition)to diminish the number of data input and output in the query process by reducing the response time, Improve the data processing capacity. Nevertheless, in the massive data size, the partition brought about by the performance improvement is not significant.

3) Lack of ability to deal with unstructured data. The traditional relational database processing of data is limited to certain data types, such as numbers, characters, strings, etc., on unstructured data (pictures, audio, etc.) the support is poor. Nevertheless, with the increase of consumer application requirements, the development of hardware technology and the promotion of multimedia communication on the Internet, the consumer's demand for multimedia processing from simple storage to identification, retrieval and in-depth processing, in the face of growing processing of huge sound, Image, video, E-mail and other complex data types, this traditional database has become inadequate.

4) Poor scalability. In the massive or large scale, the traditional database of a fatal weakness is its scalability difference. Usually there are two ways to clear up the problem of database scalability: scale up and scale out. In the face of massive data processing, by improving the performance of the server scale up way in terms of cost and processing capacity cannot meet the requirements, the only practicable way is to carry out scale out. The method of relational database management system scale out is to deploy the entire database to a cluster by vertical and horizontal cutting. The advantage of this approach is that the RDBMS can be used as a mature technology, but the drawback is that it is special fixed application, the application of different cutting methods are not the same [24].

3.3 Cloud-Computing Technology: The demand for large data technology is accompanied by the emergence of cloud computing platform, it is necessary to introduce the cloud computing technology. In fact, the present cloud computing technology is an crucial part of large data storage and processing technology. Due to the large amount of data and the characteristics of distributed data, for the traditional data management technology is difficult to manage this massive data.

Cloud computing is the core of mass data storage and data parallel processing technology. Its core ideas include, distributed file system (DFS) and MapReduce technology, the main idea proposed by Google. DFS is characterized by high fault/defect tolerance and is designed for deployment on inexpensive hardware, and it provides high-throughput data access for applications that are appropriate for programs with large data sets. Hadoop provides an open source execution of DFS (HDFS), the distributed file system to relax the requirements of POSIX and access the data in the file system can be realized in the form of a flow (streaming access), and has high reliability, high scalability and load balance capability.

MapReduce [25] is a parallel programming model proposed by Google in 2004 for parallel processing and generation of large data sets. Hadoop includes an open source execution of MapReduce [26], which is one of the large data-processing skills that draw attention. In order to make the MapReduce parallel programming model easier to use, there are a variety of large data processing advanced query language, such as FACEBOOK's Hive [27], YAHOO's Pig [28], GOOGLE's Sawzall [29], etc. These high-level query languages parse the query statement into a series of MapReduce jobs through a parser and execute it in parallel on a distributed file system. Compared with the basic MapReduce system, high-level query language is more acceptable for large-scale data parallel processing [30]. MapReduce and high-level query language are also exposed to real-time and efficiency deficiencies in the application, so there are many studies for them to optimize. Cloudera released a real-time query for the open-source project Impala 1.0 Beta, which showed a 3 to 90-fold improvement over the original MapReduce-based Hive SQL query [31]. Mahout is a parallel-based MapReduce developed by Apache Data mining projects, compared to traditional data mining algorithms, performance increased significantly [32].

3.4 The application of Cloud computing in Smart Grid: The largest amount of data in the smart grid should belong to the power apparatus condition monitoring data. The status monitoring data not only includes the online condition monitoring data (timing data and video), but also includes the basic information of the apparatus, experimental data, defect records, etc. The data quantity is very large, the reliability is high, and the real-time requirements are higher than business management data.

The application of cloud computing technology in the domestic power industry is still in the exploratory stage, the research content is mainly focused on the system design, the recognition of ideas, prospects and so on. In [33], Hadoop is used to store and manage data by means of virtualization technology, distributed redundant storage and data storage based on column storage

to ensure the reliable and efficient management of power grid mass data, is still only a framework. In order to fix the low utilization rate of resources, power system disaster recovery center, disaster recovery complex business process and a series of problems, the [34] cloud computing resource management platform framework and some modules are designed, the goal is to achieve power enterprise ERP data backup, but not yet achieved. In the literature [35], the system architecture and the hierarchy of the cloud computing center of the power system simulation are designed, such as infrastructure cloud, data management cloud, simulation computing cloud, etc. In literature [36], discusses the challenges of the future smart grid control center, proposed the combo of networking and cloud computing technology is the technical support of future control center. The Hadoop cloud-computing platform was designed and implemented in the laboratory and the Hadoop-based power apparatus condition monitoring and storage system was designed and executed [37], the dynamic timing data, static data and video data are stored, keyword inquired and processed in parallel. The system has been tested to verify the cloud-computing platform and advantages were high reliability, good scalability and data parallel access performance.

In some countries, cloud computing applications have been used for large data storage and simple processing. Cloud has been implemented for the actual operation of the system. In [38] analyzed the real-time query requirements of different consumers in the power system, and the smart grid data cloud model is designed, which is especially acceptable and appropriate for dealing with mass flow data generated in the smart grid. It is based on this model of real-time data intelligent measurement and management System. Cloudera designed and implemented the Hadoop platform-based smart grid project on the Tennessee Valley Authority (TVA) [39] to help the US grid manage hundreds of Terabytes of PMU data, highlighting Hadoop's high reliability and with the advantage of low price, In addition TVA developed the superPDC on the basis of the project, and open source through openPDC project, this work will facilitate the large-scale analysis of measurement data, It can also provide a general platform for other time series data processing. Japan Kyushu Power Company using Hadoop cloud computing platform for rapid parallel analysis of massive user data related to power usage [40]. Based on the platform, variety of distributed batch application software is developed to improve the speed of data processing and effectiveness.

In [41], the cloud-computing platform is applied to the smart grid in detail, and the conclusion is that the present cloud computing platform can meet the reliability and scalability of the smart grid monitoring software, but the real-time, data privacy and security requirements cannot be met, pending further study.

4. OPPORTUNITIES AND CHALLENGES FOR SMART GRID LARGE DATA

4.1 Large data transmission and storage technology: With the gradual advancement of the construction of the smart grid, the operational data, apparatus status and online monitoring data are recorded in all aspects of the power system. The large data transmission and storage problems not only place a great burden on the monitoring devices, but also restricts the leapfrog development of intelligent power system. [42].

Through data compression, we can effectively diminish the amount of network data transmission, improve storage efficiency. Therefore, the data compression technology has received broad attention. In [43], the real time data compression and reconstruction algorithm based on lifting scheme for transient defect signal is discussed, the real-time data of the power system is compressed and decompressed by the combo of linear integer transform wavelet biorthogonal filter and Huffman coding. In [44], the paper studies the periodical data compression algorithm of thermal power plant based on 2-D lifting wavelet. In [45], the parametric compression algorithm of steady-state data in power system is studied. In the transmission line condition monitoring system, in order to find the insulator discharge, the leakage current sampling frequency will be relatively high, the data quantity is big. Currently, such systems generally use wireless communication, network bandwidth is limited, so the need for data compression. In [46], an adaptive multi-level tree partitioning algorithm (SPIHT) is proposed, which can adaptively divide the set according to the significance of wavelet coefficients, especially for high-noise signals such as compressed leakage current. Data compression on the one hand to reduce the storage space, on the other hand compression and decompression resulting in a lot of CPU resources consumption. After the data reaches at monitoring center, it is necessary to decompress the data and need the appropriate calculation and storage platform.

In data storage, large amounts of data in smart grids can be stored using distributed file systems, such as Hadoop's HDFS. Nevertheless, these systems can store large amounts of data, but it is difficult to meet the real-time requirements of power systems [47]. Thus, it is necessary to classify the data in the system according to the performance and analysis requirements: Real time database system is used for real-time data with high performance requirements; using traditional parallel data

Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 4, Issue 1, pp: (35-46), Month: January - April 2017, Available at: www.noveltyjournals.com

warehouse system for core business data; large sum of historical and unstructured Data should be in distributed File system. This paper presents a multi-tier storage system for large data in the smart grid, as shown in Figure 1. It should be noted that, in view of the present cloud platform to receive real-time monitoring of smart grid data cannot be guaranteed, In front of the data access and information integration in Figure 1, we can set up a sum of front-end computers, responsible for receiving the alarm information or monitoring data sent to the communication network in real time and being responsible for temporary storage when the cloud platform cannot respond.

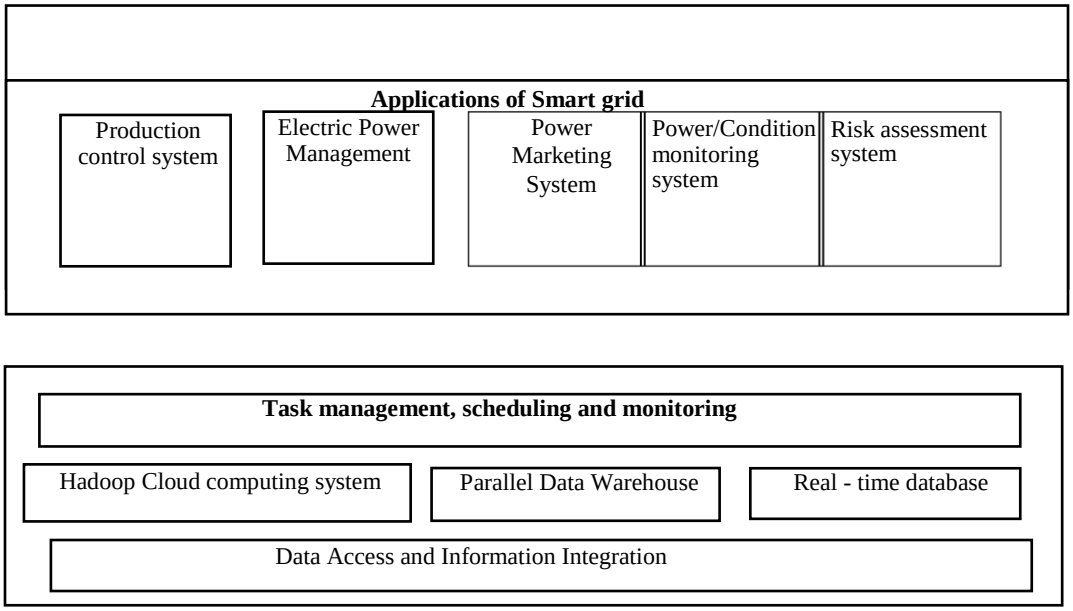

Figure 1: Multi-Level storage system for large data from smart grid

In addition, the smart grid data format and traditional business data is very different, has its own characteristics. For example, in the fault recording and transmission apparatus condition monitoring, the waveform data is more, and with the characteristics of fast data generation, the waveform data and the traditional commercial data are essentially different. Thus, it is necessary to study the format of large data storage for smart grid, which is helpful for subsequent data analysis and calculation.

All kinds of heterogeneous data in smart grid environment cannot be described by the existing simple data structure, Computer algorithms are relatively inefficient in dealing with complex structural data, nevertheless in dealing with homogeneous data is very efficient [48]. Thus, how to organize the data into a reasonable homogeneous structure, large data storage is a critical issue. In addition, there are a large sum of unstructured and semi-structured data in the smart grid, how to convert these data into a structured format, is a major challenge.

4.2 Real-time data processing technology:

4.2.1 Data processing timelines for large data: The data processing speed is very crucial. In general, the larger the data size, the longer the processing time will be. Conventional data storage schemes are designed for a certain amount of data, and can be processed very quickly within their design limits, but cannot accommodate large data requirements. The future of smart grid environment, from the power generation, power transmission link and to the electricity link, need real-time data processing. The present cloud computing system can provide fast service, but it may be affected by transient network congestion, or even a single server failure, but cannot guarantee the response time [41].

Memory-based databases are gaining increasing attention. A memory database is a database that stores data directly in memory, compared to the disk, the memory data read and write speed to be higher than a few orders of magnitude, storing data in memory rather than accessing it from disk can greatly improve application performance. Currently, the power system has begun to use the memory database to improve real-time. For example, in response to last year's electricity shortage in some parts of China, SAP introduced and launched a smart meter analysis solution based on HANA memory database [49], In order to achieve the analysis of power consumption situation and to do the corresponding preventive measures, hoping to integrate and analyze the data involved in the smart grid and the data of the large power consumers.

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 4, Issue 1, pp: (35-46), Month: January - April 2017, Available at: www.noveltyjournals.com

Querying keywords in a large data set is also an critical challenge. It is not feasible to scan the whole dataset to find a record that meets the requirements, even through parallel processing methods such as MapReduce, it is not reasonable to speed up scanning. It is a kind of method to save the system resources by establishing index structure for data in advance. Currently, the general index structure is designed to support only a few simple data types, large data is required for the complex structure of the data to establish the appropriate index structure [50], which is a humongous challenge. For example, the multidimensional data collected by the networking (IoT), its data volume is growing, the query time-limit is required, and the design of the index needs to be updated. The design of the index is very challenging. The following analysis from the power generation, power transmission and power consumption, represent the challenges during processing the large-scale data in the smart grid.

4.2.2 Power generation link: Power generation companies are characterized by continuous production process, High-speed real-time data processing, long-term historical data storage and production information integration and sharing. Studies have shown that the normal operation of the SCADA system receives monitoring data delay if more than 50 ms, it will lead to the wrong control strategy [51-52]; studies also have shown that the SCADA system in the use of Internet environment, the most common TCP/IP protocol failure occurs [52-53], the particular reason is the TCP protocol in the flow control and data error correction, resulting in data delay. Future smart grid solutions will require real-time response, even if a node failure occurs. Present relational database systems and cloud computing systems are designed to handle permanent, stable data. The relational database emphasizes the integrity and consistency of the maintenance data. The cloud-computing system emphasizes reliability and scalability; however, it is difficult to take into account that the relevant data and its processing timing constraints, cannot meet the needs of industrial production management real-time applications

4.2.3 Power transmission link: Condition monitoring has a high requirement for the performance and real-time performance of data storage and processing platform, while the cloud computing technology can effectively handle the large data, but it needs to further improve the access capability of the cloud platform to mass data, requirements. In the past, large-scale blackouts [54] were initially caused by a number of environmental factors, such as line trip caused by wind and so on. The monitoring range of the present SCADA system is limited to the main parameters of the system, the lack of information about the health status of the important apparatus of the system makes it difficult for the operator to deal with the problem correctly. The future smart grid requires a fault self-healing function, the SCADA system must have the whole network monitoring data and it is necessary to take the state data of power apparatus into account, this is a real requirement for real-time processing of the platform.

The instability of the new green energy power generation caused the fluctuation of the power grid, which formed great pressure at overall grid. Currently, the power grid scheduling and control model is not able to deal with the large sum of small power generation system fluctuations and unpredictable behavior. To support this, a new state-of-the-art monitoring system for power grids is needed to track power grid real-time status more finely, [41]. So future SCADA systems require real-time processing which should be highly effective than the present monitoring data.

4.2.4 Electricity links: In the future of smart grid environment, the family may be equipped with a variety of power, power monitoring apparatus, to achieve low-cost electricity, and with the power grid to match the load. For example, electric water heaters may choose to run this low-power consumption during the night; air conditioning will be based on user comfort, real time automatic adjustment of electricity price and load. To some extent, we can think that the SCADA system has entered the ordinary family, and the real-time data processing of the electric power is becoming more and more important.

4.3 Heterogeneous multi data source processing technology:

4.3.1 Integration of heterogeneous information: The future smart grid requires power generation, transmission, substation, power distribution, electricity, scheduling and other links, to achieve a comprehensive collection of information, smooth transmission and efficient processing, highly integrated power flow, information flow, and business flow. Therefore, the primary function is to achieve large-scale multi-source heterogeneous information integration for the data center to provide resource-intensive configuration for smart grid. It is urgent to fix the problem, for massive heterogeneous data, how to construct a model to regulate the expression, how to realize the data fusion based on the model, and carries on the effective storage and query.

The information system of the power grid is based on the needs of the business or the department, and there are different platforms, application systems and data formats, leading to information and resource dispersion, heterogeneity, horizontal cannot be shared, the upper and lower levels of vertical difficulties, Such as: power system monitoring, energy management, distribution management, market operations and other types of information systems, mostly independent of each other, so the data cannot be shared. The use of cloud platform to achieve the integration of the independent system, which can realize the information exchange between these isolated systems.

In addition, the smart grid infrastructure are in large-scale, for large public and distributed in different locations. For example: In State Grid Corporation China, they are establishing two levels of data sources for its headquarters and for the network of companies to achieve the company's headquarters, the provincial network companies, cities and counties of the company's 3 tier applications.

4.3.2 Efficient management of various types of smart grid data: In the heterogeneous multi-source information fusion and management of smart grid, It is necessary to establish a model of information interoperability like IEC 61850 or IEC 61970. Since the data types in the smart grid are more than those covered by IEC 61850, it is an optional method to build domain oriented analysis model and semantic based service model. This paper studies the integration scheme of heterogeneous data fusion and mining, and the real-time mining algorithm, the theory of statistical learning; support vector machine, relevance vector machine and association rule mining are applied. Because the deterioration of the apparatus condition is a process from variable to qualitative change, it is more meaningful to mine time series data such as oil chromatogram, which has accumulated for many years. Although there are some research results on this kind of data mining, the practicality is not high.

4.4 Large data visualization analysis technology: Faced with vast amounts of smart grid data, how to present to users in an intuitive, easy-to-understand manner under limited screen space is a challenging chore [56]. The visualization method has been proved to be an effective method to solve the large-scale data analysis and has been broadly used in practice [57]. Large-scale data sets generated by numerous types of smart grid applications, including high-precision, high-resolution data, time-varying data and multivariable data. A typical data set can reach the TB quantity set. How to excerpt useful information from these large and complex data quickly and effectively becomes a key technical difficulty in the application of smart grid. Visualization draws data into high precision, high-resolution images through a series of sophisticated algorithms and provides interactive tools for efficient use of the human visual system and allows real-time data processing and algorithmic parameters to be observed by qualitatively and quantitatively Analysis [58].

The challenges include the scalability of the visualization algorithm, the parallel image synthesis algorithm, the extraction and display of crucial information, and so on [59].

5. CONCLUSION

The future of smart grid will be relying on panoramic large state (Analog or Digital) data processing and analysis technology. Cloud computing provides a platform for storage and analysis of heterogeneous and diverse data. Platform operation will inevitably produce large data after a period operation. Large data analysis will be the status of power apparatus maintenance, power grid self-healing, and isolated information systems to support interoperability. This will become an important candidate for a low-cost, good system Scalability (unlimited storage capacity), high reliability, parallel analysis and other advantages. Several systems have been put into practice in the world, but there are still many challenges in data consistency, privacy and security, and need to find the appropriate solution. Large data processing technology is still lacking, and requires to be explored by us.

REFERENCES

[1] Xi Fang, Satyajayant Misra, Guoliang Xue, et al. Smart Grid, the new and improved power grid : a survey[J] . IEEE Communications Surveys and Tutorials (COMST), 2012, 14(4) : 944-980.

[2] Zhang Wenliang, Tang Guangfu, Zha Kunpeng, et al. Application of advanced power electronics in smart grid[J]. Proceedings of the CSEE, 2010, 30(4) : 1-7(in Chinese).

[3] Li Guojie . The scientific value of big data[J] . Research Communications of The CCF, 2012, 8(9) : 8-15(in Chinese).

[4] Divyakant Agrawal, Philip Bernstein, Elisa Bertino, et al. Challenges and opportunities with big data[J] . Proceedings of the VLDB Endowment, 2012, 5(12) : 2032-2033.

[5] IBM Corporation Software Group. IBM big data overview for energy and utilities[EB/OL] . 2011-06[2012] . http://www-01.ibm.com/ software/tivoli/solutions/industry/energy-utilities/.

[6] McKinsey Global Institute. Big data：the next frontier for innovation, competition, and productivity[R].2011.

[7] Peijian Wang . D-pro ： dynamic data center operations with demand-responsive electricity prices in smart grid[J] . IEEE, Transactions on Smart Grid, 2012, 3(4)：1743-1754.

[8] Zhou Hui, Niu Wenjie, Wang Yi. Analysis of clients' credit based ontheirs paying behaviors[J]. Power Demand-Side Management, 2006, 8(6)：12-16(in Chinese).

[9] Conejo A J, Morales J M, Baringo L. Real-time demand response model[J]. IEEE Transactions on Smart Grid, 2010, 1(3)：236-242.

[10] Niu Dongxiao, Gu Zhihong, Xing Mian, et al. Study on forecasting approach to short-term load of SVM based on data mining [J]. Proceedings of the CSEE, 2006, 26(18)：6-12(in Chinese).

[11] Xie Huacheng, Chen Xiangdong. Cloud storage-oriented unstructured data storage[J]. Journal of Computer Applications, 2012, 32(7)：1924-1928(in Chinese).

[12] Li Feng, Xie Jun, Lan Jinbo, et al. Prospect and discussion of relay system configuration for intelligent substation[J]. Electric Power Automation Equipment, 2012, 32(2)：122-126(in Chinese).

[13] 江苏瑞中数据股份有限公司. 海迅实时数据库助力智能电网建设 [EB/OL]. 2011-05[2013 02]. http://hvdc. chinapower.com.cn/mem bercenter/sitebuild4/content.asp.

[14] Hou Ziliang , Pan Gang . Constructing demonstration projects of digitized power plant to speed up the informatization process in fossilfired power plants[J]. Electric Power, 2005, 38(2)：78-80(in Chinese).

[15] Li Han , Xiao Deyun . Survey on data driven fault diagnosis methods[J]. Control and Decision, 2011, 26(1) ：1-16(in Chinese).

[16] Zhou Donghua, Hu Yanyan. Fault diagnosis techniques for dynamics system[J]. Acta Automatica Sinica, 2009, 35(6)：748-758(in Chinese).

[17] Pregelj A, Begovic M, Rohatgi A. Quantitative techniques for analysis of large data set in renewable distributed generation[J]. IEEE Trans on Power Systems, 2004, 19(3)：1277-1285.

[18] Versant. NoSQL and the smart grid big data challenge[EB/OL]. 2012-08[2013 02]. http://www.greentechmedia. com/articles/read/ver sant-nosql-and-the-smart-grid-big-data-challenge/.

[19] David Kligman. PG&E's Austin kicks off conference on dealing with smart grid data[EB/OL] . 2012-08[2013-02] . http://www.pgecurrents.com/2012/08/14/pg-topic-is-dealing-with-data-that-comeswith-smart-grid/.

[20] Jin Cheqing, Qian Weining, Zhou Aoying. Analysis and management of streaming data : a survey[J]. Journal of Software, 2004, 5(8)：1172-1181 (in Chinese).

[21] 国网信通有限公司. 信通公司举办大数据开启智能电网新时代研 讨会[EB/OL] . 2012-07[2013 02] . http:// www.sgit.sgcc.com.cn/newzxzx/gsxw/07/277345.shtml.

[22] Zhang Guangbin, Shu Hongchun, Yu Jilai. Travelling wave field datacontingency screening based on semi-supervised clustering using generalized current modal component[J]. Proceedings of the CSEE, 2012, 32(10)：150-158(in Chinese).

[23] Codd E F . A relational model of data for large shared data banks[J]. Communications of the ACM, 1970, 13(6) ：377-387.

[24] Roland Bouman. Database sharding at Netlog with MySQL and PHP[EB/OL]. 2009-02[2013 02] . http://www. jurriaanpersyn.com/archives/2009/02/12/database-sharding-at-netlog-with-mysql-andphp/.

[25] Jeffrey Dean, Sanjay Ghemawat . MapReduce : simplified data processing on large clusters[C]//OSDI'04 : Sixth Symposium on Operating System Design and Implementation. San Fancisco , California : USENIX Association Berkeley, 2004 : 137-150.

[26] Apache. Apache Hadoop core[EB/OL]. 2012-08[2013-02]. http://hadoop.apache.org/core/.

[27] Thusoo A, Sarma J, Jain N, et al. Hive : a warehousing solution over map-reduce framework[C]//Proc of the 35th Int Conf on Very Large Data Bases (VLDB). Lyon, France : VLDB, 2009 : 1626-1629.

[28] Christopher Olston , Benjamin Reed , Utkarsh Srivastava . Pig latin : a not-so-foreign language for data processing[C]//Proceedings of the 2008 ACM SIGMOD international conference on Management of data . Vancouver, Canada : ACM, 2008 : 1099-1110.

[29] Rob Pike, Sean Dorward, Robert Griesemer, et al. Interpreting the data : parallel analysis with Sawzall[J]. Scientific Programming, 2005, 13(4) : 277-298.

[30] Wang Peng, Meng Dan, Zhan Jianfeng, et al. Review of programming models for data-intensive computing[J] . Journal of Computer Research and Development, 2010, 47(11) : 1993-2002(in Chinese).

[31] Marcel Kornacker, Justin Erickson. Cloudera Impala : real-time queries in Apache Hadoop for real[EB/OL] . 2012-10 [2013-02]. http://blog. cloudera.com/blog/2012/10/cloudera-impalareal- time-queries-in-apache-hadoop-for-real/.

[32] Apache. What is Apache Mahout[EB/OL]. 2011-05[2013-02]. http://mahout.apache.org/.

[33] Wang Dewen, Song Yaqi, Zhu Yongli. Information platform of smart grid based on cloud computing[J]. Automation of Electric Power Systems, 2010, 34(22) : 7-12(in Chinese).

[34] Zhu Zheng, Gu Zhongjian, Wu Jinlong, et al. Application of cloud computing in electric power system data recovery[J]. Power SystemTechnology, 2012, 36(9) : 43-50(in Chinese).

[35] Mu Lianshun, Cui Lizhong, An Ning. Research and practice of cloud computing center for power system[J]. Power System Technology, 2011, 35(6) : 170-175(in Chinese).

[36] Wang Guanghui, Li Baowei, Hu Zechun, et al. Challenges and future evolution of control center under smart grid environment[J]. Power System Technology, 2011, 35(8) : 1-5(in Chinese).

[37] Liu Shuren, Song Yaqi, Zhu Yongli, et al. Research on data storage for smart grid condition monitoring using Hadoop[J]. Computer Science, 2013, 40(1) : 81-84(in Chinese).

[38] Rusitschka S, Eger K, Gerdes C. Smart grid data cloud : a model for utilizing cloud computing in the smart grid domain[C]//Smart Grid Communications (SmartGridComm) , 2010 First IEEE International Conference . Gaithersburg, MD : IEEE, 2010 : 483-488.

[39] Christophe Bisciglia. The smart grid : Hadoop at the Tennessee Valley Authority(TVA)[EB/OL]. 2009-06[2013-02]. http://www.cloudera.com/blog/2009/06/smart-grid-hadoop-tennessee-valley-authority-tva/.

[40] Kawasoe S, Igarashi Y, Shibayama K, et al. Examples of distributed information platforms constructed by power utilities in Japan[C]//CIGRE 2012. Paris, France : CIGRE, 2012 : 108-113.

[41] Kenneth P Birman, Lakshmi Ganesh, Robbert van Renesse. Running smart grid control software on cloud computing architectures[C]// Workshop on Computational Needs for the Next Generation Electric Grid, Cornell University. Ithaca, NY : DOE, 2011 : 1-28.

[42] Zhang Baohui. Strengthen the protection relay and urgency control systems to improve the capability of security in the interconnected power network[J]. Proceedings of the CSEE, 2004, 24(7) : 1-6(in Chinese).

[43] Yan Changyou， Yang Qixun， Liu Wanshun． A real-time data compression & reconstruction method based on lifting scheme [J]． Proceedings of the CSEE， 2005， 25(9)：6-10(in Chinese).

[44] Bao Wen， Zhou Rui， Liu Jinfu． A periodical data compression method based on 2-D lifting wavelet transform in thermal power plant [J]． Proceedings of the CSEE， 2007， 27(29)：96-101(in Chinese).

[45] Zhang Bin， Zhang Donglai． Parametric compression algorithm for power system steady data[J]． Proceedings of the CSEE， 2011， 31(1)：72-79 (in Chinese).

[46] Zhu Yongli， Zhai Xueming， Jiang Xiaolei． Adaptive SPIHT algorithm for data compression of insulator leakage currents[J]． Transactions of China Electrotechnical Society， 2011， 26(12)：190-196(in Chinese).

[47] Stonebraker M， Abadi D J， Madden S， et al． MapReduce and parallel DBMSs：friends or foes?[J]． Communications of the ACM， 2010， 53(1)：64-71.

[48] 周晓方，陆嘉恒，李翠平，等．从数据管理视角看大数据挑战[J]．中国计算机学会通讯，2012，8(9)：16-20.

[49] 丁慧茹．SAP 推 HANA 电力行业应用智能电表分析提升服务 [EB/OL]． 2011-12[2013-02]． http://cio.zdnet.com.cn/cio/2011/ 1220/2070971.shtml.

[50] Cooper B F， Neal Sample， Franklin M J， et al． A fast index for semistructured data[C]//Procedings of the 27th VLDB Conference． Roma， Italy：VLDB， 2001：341-350.

[51] Chi Ho， Robbert van Renesse， Mark Bickford， et al． Nysiad：practical protocol transformation to tolerate byzantine failures[C]//USENIX Symposium on Networked System Design and Implementation (NSDI 08)． San Francisco， CA：USENIX， 2008：175-188.

[52] Hopkinson Ken M， Giovanini Renan， Wang Xaioru， et al． EPOCHS：integrated cots software for agent-based electric power and communication simulation[C]//WSC 2003． New Orleans， Louisiana， USA：IEEE， 2003， 2：1158-1166.

[53] Junqueira F P， Reed B C． The life and times of a Zookeeper[C]//SPAA '09． New York， USA：ACM， 2009：46-46.

[54] Wikipedia． Northeast blackout of 2003[EB/OL]． 2003-12[2013 02]． http://en.wikipedia.org/wiki/Northeast_Blackout_of_2003.

[55] Zhang Wenliang， Liu Zhuangzhi， Wang Mingjun． Research status and development trend of smart grid[J]． Power System Technology， 2009， 33(13)：1-11(in Chinese).

[56] Wong P C， Shen H W， Chen C， et al． Top ten interaction challenges in extreme-scale visual analytics[J]． Computer Graphics and Applications， 2012， 32(4)：63-67.

[57] 袁晓如，张昕，肖何，等．可视化研究前沿及展望[J]．科研信息 化技术与应用，2011，2(4)：3-13.

[58] Wong P C， Thomas J． Visual analytics[J]． IEEE Computer Graphics and Applications， 2004， 24(5)：20-21.

[59] Thomas J J， Cook K A． Illuminating the path：the research and development agenda for visual analytics[M]． IEEE Computer Society， 2005：28-32.