Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 18 Issue 3 Version 1.0 Year 2018
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

A Review of Real World Big Data Processing Structure: Problems and Solutions

By Khalid Imtiaz & M. Junaid Arshad

University of Engineering and Technology

Abstract- Information sort and sum in human culture is developing in astonishing pace which is brought about by rising new administrations as distributed computing, web of things and area-based administrations, the time of enormous information has arrived. As information, has been principal asset, how to oversee and use enormous information better has pulled in much consideration. Particularly, with the advancement of web of things, how to handling huge sum continuous information has turned into an extraordinary test in research and applications. As of late, distributed computing innovation has pulled in much consideration with elite, yet how to utilize distributed computing innovation for substantial scale ongoing information preparing has not been contemplated. This paper concentrated the difficulties of huge information firstly and finishes up every one of these difficulties into six issues. Keeping in mind the end goal to enhance the execution of constant handling of substantial information, this paper manufactures a sort of real-time big data processing (RTDP) design considering the distributed computing innovation and after that proposed the four layers of the engineering, and various leveled figuring model.

Keywords: big data; cloud computing; data stream; hardware software co-design; CEP; big data analytics as a service; big data cloud architecture.

GJCST-C Classification: H.3.m

Strictly as per the compliance and regulations of:

A Review of Real World Big Data Processing Structure: Problems and Solutions

Khalid Imtiaz [α] & M. Junaid Arshad [σ]

Abstract- Information sort and sum in human culture is developing in astonishing pace which is brought about by rising new administrations as distributed computing, web of things and area-based administrations, the time of enormous information has arrived. As information, has been principal asset, how to oversee and use enormous information better has pulled in much consideration. Particularly, with the advancement of web of things, how to handling huge sum continuous information has turned into an extraordinary test in research and applications. As of late, distributed computing innovation has pulled in much consideration with elite, yet how to utilize distributed computing innovation for substantial scale ongoing information preparing has not been contemplated. This paper concentrated the difficulties of huge information firstly and finishes up every one of these difficulties into six issues. Keeping in mind the end goal to enhance the execution of constant handling of substantial information, this paper manufactures a sort of real-time big data processing (RTDP) design considering the distributed computing innovation and after that proposed the four layers of the engineering, and various leveled figuring model. This paper proposed a multi-level stockpiling model and the LMA-based application organization technique to meet the continuous and heterogeneity necessities of RTDP framework. We utilize DSMS, CEP, group-based Map Reduce and other handling mode and FPGA, GPU, CPU, ASIC advancements contrastingly to preparing the information at the terminal of information gathering. We organized the information and afterward transfer to the cloud server and Map Reduce the information consolidated with the effective processing abilities cloud design. This paper brings up the general structure for future RTDP framework and computation techniques, is right now the general strategy RTDP framework outline.

Keywords: big data; cloud computing; data stream; hardware software co-design; CEP; big data analytics as a service; big data cloud architecture.

I. Introduction

With the advancement of web of things, different substantial scale continuous information preparing in view of ongoing sensor information are turning into the key of the development of EPC (epcglobal arrange) application at present. The scholarly community, the industry and even the administration establishment have as of now gave careful consideration to enormous information issues and created a distinct fascination.

In May2011, the world-renowned consulting firm McKinley released a detailed report on big data Big data: The next frontier for innovation, competition, and productivity[1], sparking a broad discussion of Big Data. The report gave a detailed analysis of the impact of big data, key technology and application areas. In January 2012, Davos World Economic Forum released a report entitled "Big Data, Big Impact: New Possibilities for Corresponding author e-mail: *pwang@ss.pku.edu.cn* International Development" [2], raising a research boom of Big Data. The report explored how to make a better use of the data to generate good social profits in the new data generation mode, and focused on the integration and utilization of mobile data produced by individual and other data. In March, the U.S. government released "Big Data Research and Development Initiative" [3] to put there search of Big Data on the agenda, and officially launched the "Big Data development plan".

In this manner, the examination of ongoing huge information has incredible application prospect and research esteem. Because of the continuous and the expansive size of information handling and different elements that ongoing enormous information requires make the review for constant huge information preparing testing, for the most part progressively, strength and extensive scale and so on.

Real time: Real-time processing of big data mainly focuses on electricity, energy, smart city, intelligent transportation, and intelligent medical fields. During the information processing it needs to be able to make quick decisions, and feedback relevant instructions to the sensing terminal input within a very short time delay. For instance, in Fire monitoring and rescue system, its processing center needs to be able to analyze and process the data collected by sensors in the site of the incident in a very short period, to make integrated decision by comprehensively considering site information such as the movement of persons and the site form and meantime to issue the corresponding instructions to the site sensing terminals, such as what extinguishing agent used for rescue, how to protect people's safety in the site of the incident and how to help the firemen to rescue. At the same time, the information gathered by sensing terminals and instruction information must arrive information gathering or processing terminal in real time and make relevant decisions. The loss caused by that decision-making information can't be conveyed Stability: In real time is also incalculable, so real-time processing of large data is particularly important. the areas covered by Real-time

Author α σ: University of Engineering and Technology, Lahore Campus Pakistan, Computer Science & Engineering Department.
e-mails: engr.khalidimtiaz@gmail.com, mjunaiduet@gmail.com

processing of big data are mostly closely related to the people such as Smart City 's intelligent transportation systems and high-speed train control system, and mostly highly associated with infrastructure which also determines the real-time data processing system in a large system architecture, hardware and software equipment and other aspects must possess high stability.

Large-scale: As talked about above, continuous huge information preparing frameworks are frequently firmly identified with urban foundation and real national application, so its application is regularly a tremendous scale. For example, shrewd city astute transportation framework, once the biggest ongoing information examination and choice are made, it frequently goes for transportation basic leadership at a commonplace and city level or even a national level, and significantly affects the national life.

Thus, this paper highlights the big data processing architecture under the cloud computing platform. It presents a data storage solution on heterogeneous platforms in real-time big data processing system, constructs a calculation mode for big data processing and points out the general framework for real-time big data processing which provides the basis for the RTDP (Real-Time Data Processing). Section 2 in this paper gives an overview of big data; section 3 discusses the differences and challenges between big data and real-time big data; section 4 points out the current deficiencies of cloud computing technology; section 5 combine cloud computing technology with the feature of real time big data to architect the processing platform for real-time big data; section 6 gives a demo about how the RTDP system is used in smart grid system and finally summarize this article.

II. Big Data Overview

Big data itself is a relatively abstract concept, so far there is not a clear and uniform definition. Many scholars, organizational structure and research institutes gave out their own definition of big data [18] [19] [20]. Currently the definition for large data is difficult to reach a full consensus, the paper references Academician Li Guojies definition for big data: in general sense, Big Data refers to a data collection that cant be obtained within a tolerable time by using traditional IT technology, hardware and software tools for their perception, acquisition, management, processing and service[21]. Real-time data is a big data that is generated in real time and requires real-time processing. Per the definition of Big Data, Big Data is characterized by volume, velocity and variety where traditional data processing methods and tools cannot be qualified. Volume means a very large amount of data, particularly in data storage and computation. By 2010 the global amount of information would rapidly upto 988 billion GB[22].Experts predict that by 2020 annual data will increase 43 times. Velocity means the speed of data grow this increasing, mean while people's requirements for data storage and processing speed are also rising. Purely in scientific research, annual volume of new data accumulated by the Large Hadron Collider is about 15PB [23]. In the field of electronic commerce, Wal-Mart's sells everyday more than 267 million(267Million)products[24].

Data processing requires faster speed, and in many areas data have been requested to carry out in real-time processing such as disaster prediction and rapid disaster rehabilitation under certain conditions need quickly quantify on the extent of the disaster, the regional scope impacted etc. Variety refers to the data that contains structured data table, semi-structured and unstructured text, video, images and other information, and the interaction between data is very frequent and widespread. It specifically includes diverse data sources, various data types, and a strong correlation between the data.

With the advancement of PC and system innovation, and additionally astute frameworks is regular utilized as a part of present day life, enormous information has turned out to be progressively near individuals' everyday lives. In 2008, Big Data issue released by "Nature" pointed out the importance of big data in biology, and it was necessary to build biological big data system to solve complex biological data structure problem [25]. Paper [25] pointed out that the new big data system must be able to tolerate various structures of data and unstructured data, has flexible operability and must ensure data reusability. Furthermore, Big Data plays an important role in the defense of national network digital security, maintaining social stability and promoting sustainable economic and social development [26]. With the development of big data technology, Big Data also plays an important role in creating a smart city, and has important applications in urban planning, intelligent traffic management, monitoring public opinion, safety protection and many other fields [27].

III. Difference and Challenges between Big Data and Real-Time Big Data

Huge information is trademark by multi-source heterogeneous information, broadly circulated, dynamic development, and "information mode after the data"[28] [29]. Notwithstanding having every one of the qualities with huge information, constant enormous information has its own attributes. Contrasted and the huge information, with regards to information reconciliation ongoing enormous information has higher prerequisites in information procurement gadgets, information examination devices, information security, and different angles. The accompanying presents from information

incorporation, information investigation, information security, information administration and benchmarking.

a) Data Collect

With the improvement of web of things [30] and Cyber Physical System (CPS) [31], the ongoing of information handling requires ever more elevated. Under the enormous information condition, various sensors and portable terminals scatter in various information administration frame work which makes information accumulation itself an issue. In RTDP framework, its ongoing information accumulation confronted makes information mix confronting many difficulties.

i. *Extensive heterogeneity*

In huge information framework, the information created by versatile terminals, tablet PCs, UPS and different terminals is frequently put away in reserve, yet in RTDP framework it requires information synchronization which conveys huge difficulties to the remote system transmission. When managing handling heterogeneity, huge information framework can utilize NoSQL innovation and other new stockpiling techniques, for example, Hadoop HDFS. However, the constant requires low in this sort of capacity innovation, where the information is frequently put away once yet read commonly. However, this sort of capacity innovation is a long way from fulfilling the necessity of ongoing enormous information framework that requires information synchronization. Because of broad heterogeneity of enormous information, information transformation must be done amid information incorporations preparing, however conventional information distribution center has clearly deficient to address the issues of time and scale that huge information requires [32] [33] [34].

ii. *Data quality protection*

In the time of huge information, it is a marvel frequently creates the impression that valuable data is being submerged in a substantial number of futile data [6]. The information nature of Big Data has two issues: how to oversee substantial scale information and how to wash it. Amid the cleaning procedure, if the cleaning granularity is too little, it is anything but difficult to sift through the valuable data; if the cleaning granularity is excessively coarse, it can't accomplish the genuine cleaning impact. So, between the amount and quality it requires cautious thought and measured which is more apparent progressively enormous information framework. From one perspective, it obliges framework to synchronize information in a brief span; then again, it additionally requires the framework to make a fast reaction to information progressively. The execution necessities of the speed of information transmission and information investigation are expanding. In addition, the information might be separated at once hub may get to be distinctly basic post handling information. Consequently, how to get a handle on the connection amongst information and precisely decide the convenience and viability of information turns into a genuine test.

b) Data Analytics

Information examination is certainly not another issue. Customary information examination is principally propelled for organized information source, and right now has a total and successful framework. On the premise of the information distribution center, it assembles an information solid shape for online logical preparing (OLAP). Information mining innovation makes it conceivable to discover further learning from a lot of information. Be that as it may, with the entry of the time of enormous information, the volume of various semi-organized and unstructured information quickly develops, which conveys gigantic effect and difficulties to the customary examination strategies and existing procedures are no longer relevant. It for the most part reflects in convenience and file outline under element condition.

i. *Timeliness of information preparing*

In the period of huge information, time is esteem. As time passes by, the estimation of information contained in the information is too weakening. Progressively information frameworks, time is required higher. For instance, in an information preparing of debacle examination, continuous fast prepares, airplane and other high opportuneness execution gadget, time has gone past monetary esteem. Harms brought about by absurd postponement would be difficult to assess. The period of constant huge information proposes another and higher necessity to the courses of events of information handling, for the most part in the choice and change of information preparing mode. Continuous information handling modes fundamentally incorporates three modes: gushing mode, clump mode and a mix of two-a blended handling mode. Albeit as of now numerous researchers have made an incredible commitment to continuous information preparing mode, yet there is no regular structure for constant handling of expansive information.

ii. *Record plan under element condition*

The information design in the period of enormous information might change always as information volume shifts and existing social database file is no longer relevant. Step by step instructions to plan a basic, proficient and ready to rapidly make an adjustment has turned into a one of the real difficulties of enormous information preparing when information mode changes. Current arrangement is essentially fabricated a list by NoSQL databases to take care of this issue, yet they have been notable take care of the demand for constant handling of enormous information.

iii. *Absence of earlier learning*

From one viewpoint, since semi-organized and unstructured information proliferate, it is hard to construct its inside formal relations while dissecting the information; On the other hand, it is troublesome for this information should have been handled continuously to have adequate time to set up from the earlier learning because of the happening to the information stream in the type of a perpetual stream.

c) Data Security

Data privacy issues associated with the advent of computers has been in existence. In the era of big data, the Internet makes it easier to produce and disseminate data, which makes data privacy problems get worse, especially in real-time processing of large data. On the one hand, it requires data transmission real-time synchronization; on the other hand, it demands strict protection for data privacy, which both raise new demands to system architecture and computing power.

i. *Expose hidden data*

With the appearance of the Internet, especially the appearance of social networks, people are increasingly used to leave data footprints. Through data extraction and integration technology, accumulate and associate these data footprints may cause privacy exposure. In real-time big data processing, how to ensure the speed of processing a data as well as data security is a key issue which has troubled many scholars. Data disclosure conflicts with privacy protection by hiding data to protect privacy it will lose the value of data; thus it is essential to public data. Especially by digging accumulated real-time large-scale data, we can draw a lot of useful information, which has a great value. How to ensure the balance between data privacy and data publicly is currently in research and application a difficulty and hot issue. Therefore, the data privacy in the era of big data is mainly reflected in digging data under the premise of not exposing sensitive information of the user. Paper [35] proposed privacy preserving data mining concept, and many scholars have started to focus on research in this area. However, there are is a conflict between the amount of information and the privacy of data, and that's why so far it has not yet a good solution. A new differential privacy method proposed by Dwork may be a way to solve the protection of data privacy in big data, but this technology is still far from practical applications [36].

d) Usability issue of data management

Its difficulties for the most part reflect in two viewpoints: gigantic information volume, complex examination, different outcome shapes; various businesses required by huge information. However, examination specialist's absence of learning of both perspectives generally. Accordingly, the ease of use of constant huge information administration principally reflects in simple to find, simple to learn and simple to utilize [37]. Along these lines with a specific end goal to accomplish ease of use of enormous information administration, there are three essential standards to be minded as takes after:

i. *Visibility*

Deceivability requires the utilization of the information and the outcomes be indicated unmistakably in an exceptionally instinctive manner. The most effective method to accomplish more techniques for vast information handling and instruments rearrangements and mechanization will be a noteworthy test later on. Ultra-substantial scale information representation itself is an issue, while constant perception of huge scale information will spend a considerable measure of registering assets and GPU assets. Subsequently how to upgrade the execution and use of the GPU is an intense test.

ii. *Mapping*

Instructions to coordinate another huge information preparing strategy to handling strategies and techniques individuals have turned out to be usual to and accomplish quick writing computer programs is an extraordinary test to information ease of use later on. For Map Reduce needs SQL-like standard dialect, the scientists built up a more elevated amount dialects and frameworks. Run of the mill agents are the Hadoop Hive SQL [32] and Pig Latin [38], Google's Sawzall [39], Microsoft's SCOPE [40] and DryadLINQ [41] and also MRQL [42], and so on. Be that as it may, how to apply these dialects and frameworks to ongoing enormous information preparing still stay huge difficulties.

iii. *Feedback*

Criticism configuration permits individuals to monitor their working procedures. Works about this perspective is few in Big Information field [43] [44] [45]. In the period of huge information, the inner structure of many apparatuses is exceptionally mind boggling. Also, in programming investigating it is like Black Box troubleshooting for the typical clients and the strategy is unpredictable and additionally need of input. On the off chance that later on human-PC collaboration innovation can be presented in the weight of huge information, individuals can be all the more completely required in the entire investigation prepare, which will successfully enhance the client's criticism sense and extraordinarily enhance the usability.

An outline meets the over three standards will have the capacity to have a decent convenience. Perception, human-PC collaboration and information because systems can successfully improve ease of use. Behind these innovations, gigantic metadata administration needs our unique consideration [46]. So how to accomplish a proficient administration of the enormous metadata in a huge scale capacity framework

will have an import effect on the ease of use of ongoing huge information.

e) Test benchmark of performance

Advantages A critical perspective for enormous information administration is the quality affirmation, particularly for ongoing administration of extensive information as catastrophe created by information mistake will be intense and even limitless. The initial phase in quality affirmation is to do execution testing. There is not yet a test benchmark for the administration of enormous information. Principle challenges confronted by building huge information benchmarks are as followings [47]:

i. *High many-sided quality of framework*

Constant enormous information is exceedingly heterogeneous in information design and in addition equipment and programming and it is hard to model every single huge dat items with a uniform model. Continuous enormous information framework requires high opportuneness which makes it difficult to remove a delegate client conduct progressively. What's more, information size is substantial and information is extremely hard to imitate which both make the test more troublesome.

ii. *Rapid upset of framework*

The customary social database framework design is moderately steady, however the information continuously enormous information preparing is in a consistent condition of development, and there is a sure relationship be tween's the information, which makes the benchmark test comes about got soon not mirror the present framework real execution. Continuously enormous information framework test results are required to be finished inside a brief timeframe delay with high exactness, which in the equipment and programming angles is a genuine test to the test benchmark.

Reconstruct or reuse existing test benchmark Extend and reuse on the current benchmarks will significantly decrease the workload of building another vast information test benchmark. Potential applicant's principles are SWIM (Measurable Workload Injector for Map Reduce) [48], MRBS [49], Hadoop possess Grid Mix [50], TPC-DS [51], YCSB++ [52], and so on. In any case, these benchmarks are no longer relevant continuously enormous information preparing. Presently there are as of now some explores concentrating on the development of huge pieces of information test benchmark, yet there is additionally a view which thinks its untimely to talk about that at present. By following and investigating the heaps of seven items which are connected with Map Reduce innovation, Chen et al [47] [53] think it is difficult to decide ordinary client situations in the period of huge information. When all is said in done, building huge information and constant huge information test benchmark is vital. In any case, the difficulties it will face are a considerable measure, and it is extremely hard to construct a perceived testing models like TPC.

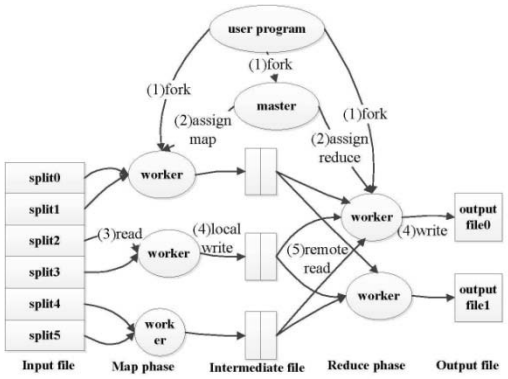

Fig. 1: The framework of the Map Reduce Model

IV. Shortcomings of Cloud Computing Architecture

a) Cloud Computing Overview

Cloud computing is the product of the traditional computer technology and network technology development integration such as grid computing, distributed computing, parallel computing, utility computing, network storage, virtualization, load balancing, etc., which aims at integrating multiple relative low-cost computing entities into one perfect system with powerful computing ability via the network and with the help of the SaaS, PaaS, IaaS, MSP and other advanced business models distributing this powerful computing ability to the hands of the end user.

b) Shortcomings of cloud computing architecture

MapReduce model is simple, and in reality, many problems can be represented with MapReduce model. Thus MapReduce model has a high value as well as many application scenarios. But MapReduces achievement is mainly relying on the Hadoop framework while the data processing method of Hadoop is" Store first post-processing" which is not applicable to real-time large data processing. Though currently there are some improved algorithms able to make Hadoop-based architecture almost real-time, for example, some latest technology like Cloudera Impala is trying to solve problems of processing real-time big data on Hadoop the batch processing of Hadoop and its structural features make Hadoop defective in processing big data in real time. Hadoops defect in real-time big data processing mainly reflects in data processing modes and application deployment. This paper will discuss these two aspects separately in the following.

Big data processing mode can be divided into stream processing and batch processing. The former is store-then-process, and the latter is straight-through-processing. In stream processing, the value of data reduces as time goes by which demanding real-time; in batch processing, data firstly is stored and then can be

processed online and offline [46]. MapReduce is the most representative of the batch processing method.

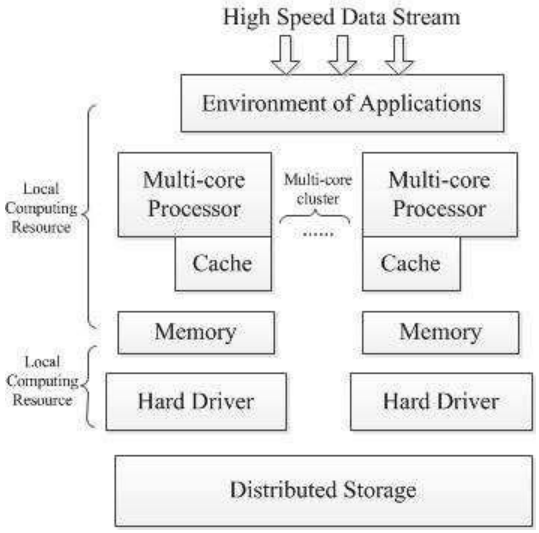

Fig. 2: Supporting Environment

c) Other architectures

Except the Hadoop-based real-time big data processing architecture, researchers already design an architecture to deal with a streaming data based on the way to process the stream-oriented data, noticing that batch processing in Hadoop can't meet the feature of real-time big streaming data. For example, Twitters Storm processing mode, Apaches Spark and LinkedIn In-stream.

Spark, as an advanced version of Hadoop, is a cluster distributed computing system that aims to make super-big data collection analytics fast. As the third generation product of Hadoop, Spark stores the middle results with internal storage instead of HDFS, improving Hadoops performance to some extent with a higher cost. Resilient Distributed Dataset, RDD, is an abstract use of distributed memory as well as the most fundamental abstract of Spark, achieving operating the local collection to operate the abstract of a distributed data set. Spark provides multiple types operations of data set which is called Transformations.

Storm cluster has some similarity with Hadoop. The difference is that its Job in MapReduce running in Hadoop cluster and Topology in Strom. Topology is the highest-level abstract in Storm. Every work process executes a sub-set of a Topology, which consists of multiple Workers running in several machines. But naturally the two frameworks are different. Job in MapReduce is a short-time task and dies with the tasks ending but Topology is a process waiting for a task and it will run all the time as system running unless is killed explicitly. In Storm cluster, it also has Master node and Worker node.

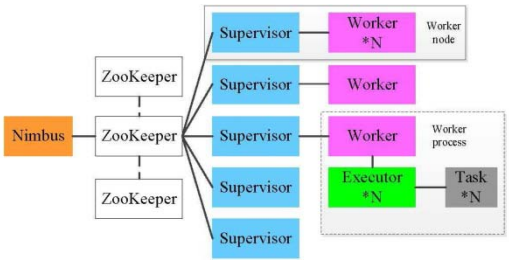

Fig. 3: Physical Architecture of Storm

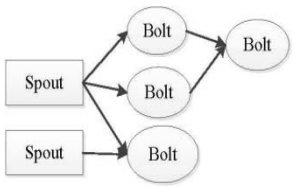

Fig. 4: Physical Architecture of Storm

V. Real-Time Big Data Processing Framework

In addition to powerful computing ability, real-time big data processing system must have strong timeliness which means it must quickly respond to the request from system terminals in a very short time delay. So at first, real-time big data processing system must have powerful computing ability for big data. A traditional method to process big data is to rely on the powerful computing capabilities of the cloud computing platform to achieve, while for the timeliness it must rely on the ability of the rapid data exchange between system's internal and nodes.

RTDP (Real-Time Data Processing) framework into four layers–Data, Analytics, Integration and Decision from a functional level. Shown in Figure 5.

a) Data

This layer mainly charges for data collection and storage, but also including data cleaning and some simple data analysis, preparing data for Analytics. At the terminal of data collection, it needs to manage all terminals. For example, the FPGA commonly used in Data Stream Management System, DSMS; the ASIC used in Complex Event Processing, CEP; and CPU and GPU (Graphic Processing Unit) in batch processing system represented by MapReduce. Data storage module is responsible for the management of large-scale storage systems. Thanks to the heterogeneity of real-time data sources and the large data processing platform, RTDP systems can handle data from various data sources, including Hadoop for unstructured storage, the data warehouse system for structured storage and analysis, SQL databases, and some other data source system.

b) Analytics

This layer is the core of RTDP system and the critical layer to determine the performance of RTDP system. This layer is mainly responsible for data structure modelling, data cleansing and other data analysis processing, preparing data for the algorithm integration layer.

c) Integration

d) Decision

This layer makes decisions with the results of data analysis which is the highest layer of data processing system as well as the ultimate goal of data analysis process. RTDP is a procedure involving numerous tools and systems interact with each other iteratively. At every level, the definition of" Big data" and" Real time" is not immutable. They have their own unique meaning at every level due to the functional association at each level. The four layers will be general process of RTDP in the future as well as the basic framework of the RTDP in this paper. Here we are going to discuss each layer in detail from the functionality, processing methods, related tools and deployment aspects of the system.

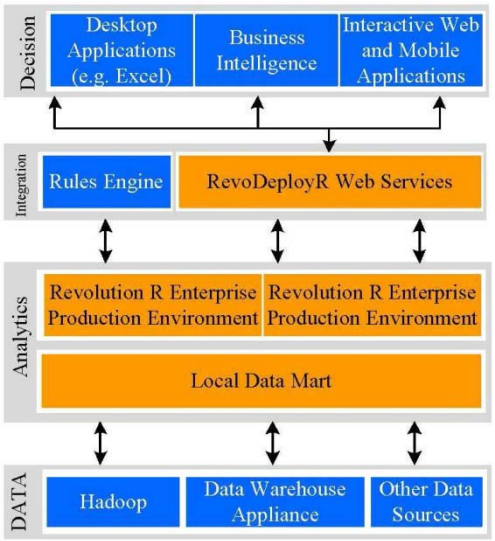

Fig. 5: Architecture of Real-Time Data Processing

e) Data Layer

Since the data collected by sensors is rough and messy, and original data often contain too much useless data, modeling and data analysis for the tremendous difficulties, so the data collection process must be preliminary data analys is and filtering. First need to extract the data features, integrated data sources, extraction points of interest, select the characteristic function to determine the data formats and extract useful information from data marts, and several steps in which the data feature extraction for unstructured text data, etc. of data is very important, therefore, makes the feature extraction for data collection and storage is an important part of the process.

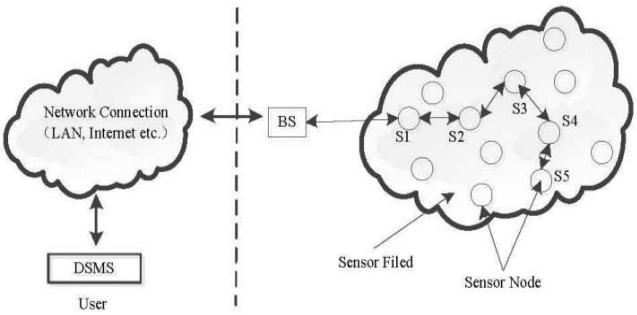

Fig. 6: Adaptive Task Allocation of Data Stream Management System

f) Data Collection and Data Base

RTDP systems heterogeneous platforms and performance makes RTDP system's data source contains a variety of ways, according to the data processing mode can be roughly divided into the CEP, DSMS, DBMS, based on a variety of ways such as MapReduce batch for each treatment have their different data acquisition techniques, such as remote medical field for surgical treatment of complex event processing scenarios for data acquisition ASIC, decoding audio and video coding in an FPGA, etc. Thus, during the data collection and management there are certain rules that must be collected on the side of the device identification, and can be based on different device programming overhead deployment and management nodes.

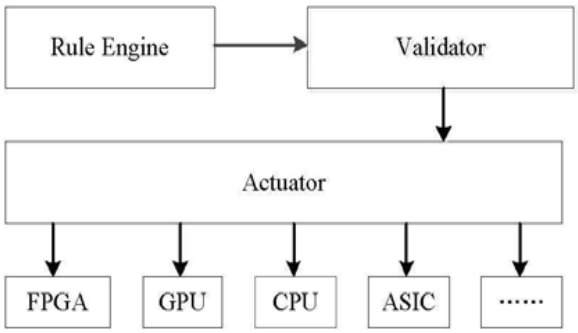

Fig. 7: Structure of Logical Management Adapter

g) Data Storage and Cleaning

In RTDP, data comes from a wide range of sources, unstructured and structured data mixed, so Hadoop and other unstructured storage system in RTDP system has a natural advantage, but Hadoop itself does not achieve full real-time requirements, which determines our in real-time using Hadoop big data storage process Hadoop first need to solve real-time problems in the framework of the proposed RTDP use of multi-level storage architecture to solve the problem, its architecture is shown in Figure 8.

In RTDP multi-level storage system data through a lot of the local server first preliminary processing, and then uploaded to the cloud server for in-depth analysis and processing. Such architectural approach to solve the data filtering is how to determine the relevance of the issue of data is an important means, Since the real time processing of large data nodes need to collect data in the shortest possible time for rapid processing, but also need to filter out unwanted data, but the data collection process, we can confirm the current data be collected for post data key input, for data-dependent judgment is an extremely complex task.

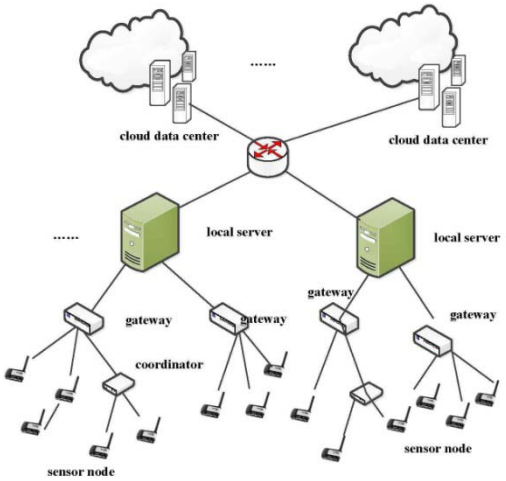

Fig. 8: Multi-Level Data Storage Model

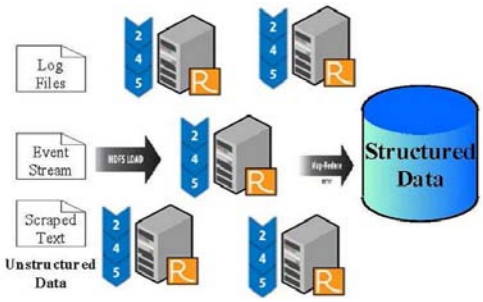

Fig. 9: Structured Data

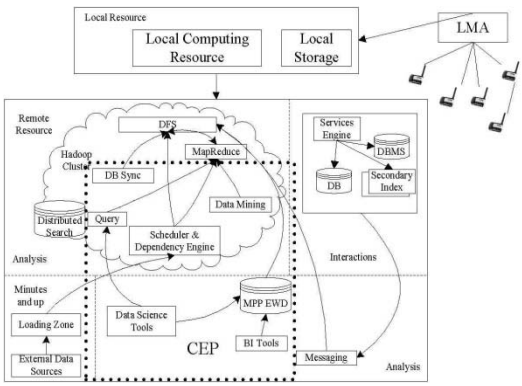

Fig. 10: Schematic of Storage and Analysis

h) Date Processing Algorithm

Along with architectural patterns, algorithm framework also plays an important role in RTDP systems computation results. In recent years, many research has done in big data processing algorithm, but the research about real-time big data has not been taken into account.

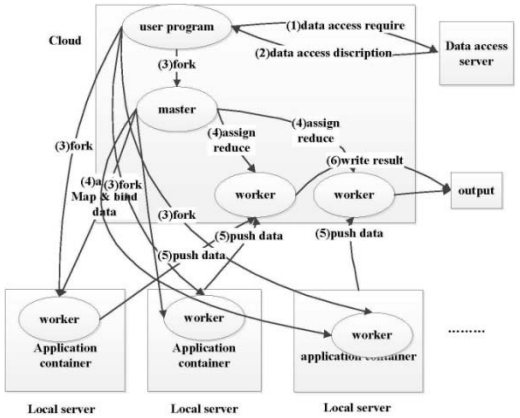

Fig.11: New Big Data Processing Model

i) Calculation Implementation Method

In previous chapters, a two layers' calculation mode has been proposed, first, the local server choose local node management and calculation procedures on the local node management, and simple data cleansing and structured modeling according to LMA. Unstructured data collected by data collector will be transformed into structured data and then uploaded to cloud memory systems and mapped to different management servers. Superstardom makes use of the computing power of cloud terminal to carry through real-time computation and analyze.

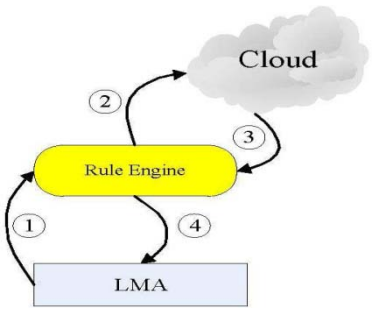

Fig.12: Application Deployment Process

j) Decision Making Layer

Truth be told, basic leadership layer incorporates two sections of ideas, to begin with, the test and refresh the model, the second is to give administrators to basic leadership. RTDP framework amid the procedure of information handling, with the stream of information, the information at various circumstances with a specific changeability, and between information likewise has a specific pertinence. subsequently change with time and profundity

information preparing, information investigation layer information show made may not meet the present needs, so we have to keep the information handling while the refresh information and refresh the information model to adjust to changes in the information then again, choice bolster layer is the most elevated amount of RTDP framework, the reason for existing is to complete information preparing related choices, so the layer must picture the created yield brings about request to give leaders to oversee related basic leadership exercises. Next a capacity diagram will be settled on about model approval and choice support.

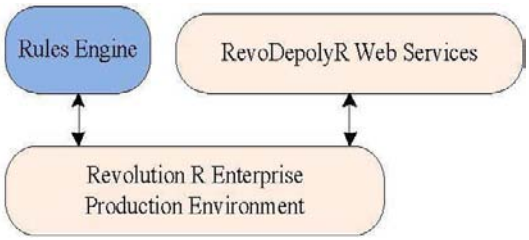

Fig.13: Application Deployment Process

VI. Application Examples

What's more, application area related RTDP explore has likewise got a great deal of consideration, and have made some preparatory research. IBM in 2010 formally proposed "shrewd city" vision and from that point forward a vast continuous information and research to get savvy city boundless consideration as urban framework concentrates canny transportation, brilliant matrix, and urban administrations identified with therapeutic insight has likewise been an incredible improvement. Different applications have their own particular one of a kind quality, many key issues still uncertain, a hefty portion of the current zones most important reviews likewise stay in the research facility stage. A case of RTDP shrewd matrix keen lattice framework in the application review will be made underneath.

Savvy lattice through countless time sensor detecting framework operation state changes, can give quicker element assessment of constant blame conclusion and vitality tracking in covering areas, districts and even across the nation dynamic direction of vitality to accomplish conveyance vitality creation with sensible dispatch, for enhancing vitality effectiveness, enhancing urban foundation assumes an essential part.

The various specialized challenges are all inside the extent of the entire system and information identified with checking and constant computation, so this paper the answer for the present field of keen lattice of the key issues, we should depend RTDP handling structures for expansive scale organize wide continuous information joining the proposed RTDP structure to fabricate another sort of shrewd network design, appeared in Figure 15.

Quick reenactment and demonstrating is the center programming of ADO, including hazard evaluation, self-mending and other propelled control and advancement programming framework for the savvy lattice to offer help and prescient numerical capacity, keeping in mind the end goal to accomplish enhanced network strength, security, dependability and operational productivity. Dispersion quick recreation and demonstrating need to bolster arrange reconfiguration, voltage and responsive power control; blame area, separation and reclamation of power; when the framework topology changes taking after the security re-tuning four self-mending abilities. Above capacity interconnectedness, coming about DFSM turn out to be extremely convoluted, for instance, either a network reproduction requires another hand-off with voltage control or the new celebration program likewise incorporates capacities to reestablish control. DFSM by means of dispersed insightful system specialists to accomplish hierarchical limits crosswise over geological limits and canny control framework keeping in mind the end goal to accomplish self-mending capacities of these keen system operators, ready to gather and trade data and frameworks, (for example, the accompanying such electrical insurance operation) nearby control choices, while as indicated by the framework prerequisites to facilitate these projects.

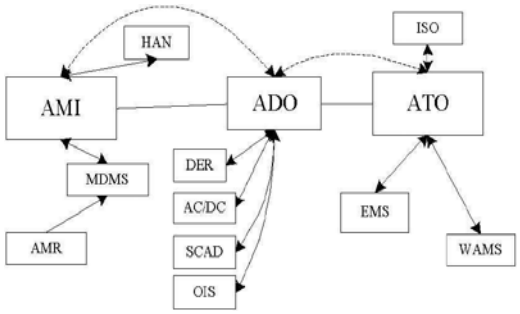

Fig.14: Technical Composition and Functionality of Smart Grid

Shrewd Grid unwavering quality issues contemplated by the AC transmission hardware for blunder infusion approach the gear disappointment mode investigation. Disappointment of the hardware and on its impact and power network unwavering quality were surveyed. writing demonstrates that the present development of brilliant framework mix of utilizations and innovation arrangements, including savvy meters, correspondence systems, metering database administration (MDMS), client premises organize (HAN), client benefit, remote turn on or off, as appeared in Figure 14, keen lattice advancements has achieved a specific level of accessibility and adaptability, however in the force of constant sending, administration, blame location and recuperation, there are still deficiencies.

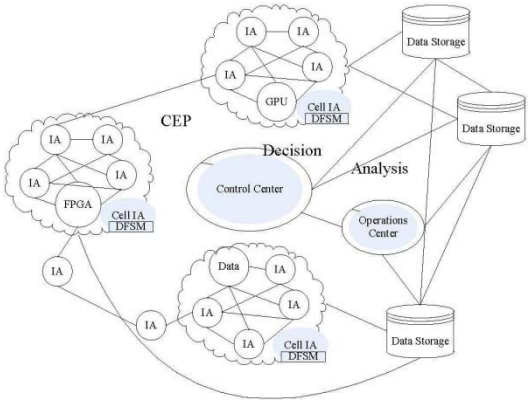

Fig.15: Architecture of Distributed Smart Grid

VII. Conclusion & Future Work

Real-time data processing for large current technology is undoubtedly a huge challenge, there is lack of support for massive real-time processing of large data frame and platform real-time processing of large data processing compared to conventional static data with high data throughput and real-time requirements. cloud computing technology in order to solve massive data processing and developed a series of techniques, however, cloud computing is very suitable for mass static, long-term without the written data has a good effect, but it is difficult to achieve real-time processing.

In this paper, on the basis of cloud computing technology to build a kind of real-time processing of large data frame, the model proposed RTDP four architecture, and hierarchical computing model. RTDP system in order to meet real-time requirements, and to consider different system platforms RTDP structural characteristics, the paper also presents a large data storage for real-time multi-level storage model and the LMA-based application deployment methods of data collection terminal based on the different ways of data processing were used DSMS, CEP, batch-based MapReduce other processing mode, depending on the environment in which the sensor data acquisition and the desired type of difference data collected were used FPGA, GPU, CPU, ASIC technology to achieve data collection and data cleansing and, through a structured process the data structure modeling, uploaded to the cloud server for storage, while the washed structured data on the local server for Reduce, combined with powerful computing capabilities cloud architecture for large-scale real-time computing with MapReduce.

This thesis indicates generally that the basic framework for future RTDP system and basic processing mode, but there are still many issues that need further study. The main point are as follows:

1. How to determine the appropriate mode of calculation in a RTDP system, how to determine the data processing mode and approach is a key factor in determining system performance, so the calculation mode and how to determine the appropriate method of calculating the design of the future core of the work;
2. Calculation models and how to achieve unity between computing technology is currently used mainly batch calculation mode and streaming processing, data computing model in determining how to design the corresponding calculation after the manner and with what kind of hardware implementation is the next big real-time data processing priority;
3. How to ensure the network transmission speed and QoS (Quality of Services); now widely used in a variety of network QoS technology, RTDP not sufficient to ensure a real-time, high reliability requirement. RTDP network QoS issues RTDP with difficulty from the inherent characteristics, so to guarantee QoS of the real-time RTDP sex have a significant impact;
4. How to ensure the system's physical time synchronization. RTDP system involves many interactions between systems and tools, software used for real-time marker approach does not meet the future RTDP high real-time requirements, the interactive how to ensure data during physical time synchronization is the future research directions;
5. How to ensure the correctness of the data processing. Error detection mechanism and automatically repair the computer has long been a difficult area of research, how to handle the data detection and error diagnostic and system repair is a huge project.

RTDP is a complex project involving many disciplines and techniques to be thorough in all aspects of research, pointed out that the article provides an overview of future research directions, and this is our future research subject.

References Références Referencias

1. Zhigao Zheng, Ping Wang, Jing Liu, Appl. Math. Inf. Sci. 9, No. 6, 3169-3190 (2015).
2. Tene O, Polonetsky J. Big data for all: Privacy and user control in the age of analytics[J]. Nw. J. Tech. & Intell. Prop., 2012, 11: xxvii.
3. LohrS.Theage ofbigdata[J].NewYorkTimes, 2012,11.
4. LynchC.Bigdata:How do your data grow ?[J].Nature, 2008, 455*(7209)*:2829
5. Bryant, RE, Katz, RH, Lazowska, ED, Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society [R], Ver. 8, (2008), Computing Research Association, Computing Community Consortium.
6. Jonathan T. Overpeck, Gerald A. Meehl, Sandrine Bony, and David R. Easterling. Climate Data

Challenges in the 21st Century [J]. Science, 2011, 331(6018): 700-702
7. Labrinidis, Alexandros and Jagadish, H. V. Challenges and opportunities with big data [J]. Proc. VLDB Endow. 2012, 5(12): 2032-2033.
8. WANG Shan, WANG Hui-Ju, QIN Xiong-Pai, ZHOU Xuan. Architecting Big Data: Challenges, Studies and Forecasts [J]. Chinese Journal of Computer. 2011, 34 (10): 1741-1752.
9. Lu Weixing, Shou Yinbiao, Shi Lianjun. WSCC DISTURBANCE ON AUGUST 10, 1996 IN THE UNITED
10. P. Jeffrey Palermo. The August 14, 2003 blackout and Its importance to China [J]. EAST CHINA ELECERIC POWER. 2004, 32(1): 2-6.
11. Li Cuiping, Wang Minfeng .Excerpts from the Translation of Challenges and Opportunities with Big Data[J].e-Science Technology & Application, 2013, 4(1): 12-18.
12. Dobbie W, Fryer Jr RG. Getting beneath the veil of effective schools: Evidence from New YorkCity[R]. National Bureau of Economic Research,2011.
13. H.V.Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, Cyrus Shahabi. Bigdata and its technical challenges[J]. Communications of the ACM,2014, 57(7): 86-94.
14. Flood M, Jagadish H V, Kyle A, et al. Using Data for Systemic Financial Risk Management[C]. Proceedings of The 5th biennial Conference on Innovative Data Systems Research (CIDR 2011). 2011: 144-147.
15. Genovese Y, Prentice S. Pattern-based strategy: getting value from big data[J]. Gartner Special Report G, 2011, 214032: 2011.
16. Albert-Lszl Barabsi. The network takeover. Nature Physics, 2012, 8(1): 14-16.
17. Labrinidis A, Jagadish H V. Challenges and opportunities with big data[J]. Proceedings of the VLDB Endowment, 2012, 5(12): 2032-2033.
18. Lohr S. How big data became so big[J]. New York Times, 2012, 11.
19. Gattiker A, Gebara F H, Hofstee H P, et al. Big Data text-oriented benchmark creation for Hadoop[J]. IBM Journal of Research and Development, 2013, 57(3/4): 10: 1-10: 6.
20. Chen M, Mao S, Liu Y. Big data: A survey[J]. Mobile Networks and Applications, 2014, 19(2): 171- 209.
21. Li Guojie, Cheng Xueqi. Research Status and Scientific Thinking of Big Data [J]. Bulletin of the Chinese Academy of Sciences. 2012, 27(6): 647-657.
22. Yadagiri S, Thalluri P V S. Information technology on surge: information literacy on demand[J]. DESIDOC Journal of Library & Information Technology, 2011, 32(1):64-69.
23. Cohen J, Dolan B, Dunlap M, Hellerstein JM, Welton C. MAD skills: New analysis practices for big data [J]. PVLDB, 2009, 2(2): 14811492.
24. Randal E. Bryant & Joan Digney. Data-Intensive Supercomputing: The case for DISC [R].2007.10: 1-14.
25. John Boyle. Biology must develop its own big-data systems. Nature. 2008, 499(7): 7.
26. Wang Yuan-Zhuo, Jin Xiao-Long, Chen Xue-Qi. Network Big Data: Present and Future [J].Chinese Journal of Computer. 2013, 36(6): 1125-1138.
27. Zou Guowei, Cheng Jianbo. The Application of Big Data Technology to Smart City [J]. POWER SYSTEM TECHNOLOGY, 2013, 4: 25-28.
28. QIN Xiong-Pai, WANG Hui-Ju, DU Xiao-Yong, WANG Shan. Big Data Analysis-Competition and Symbiosis of RDBMS and MapReduce [J]. Journal of Software. 2012, 23(1): 32-45.
29. Tan Xiongpai, Wang Huiju, Li Furong, et al. New Landscape of Data Management Technologies [J]. Journal of Software. 2013, 24(2): 175-197.
30. CHEN Hai-Ming, CUI Li, XIE Kai-Bin. A Comparative Study on Architectures and Implementation Methodologies of Internet of Things [J]. Chinese Journal of Computers. 2013, 36(1): 168-188.
31. Lee E A, Seshia S A. Introduction to embedded systems: A cyber-physical systems approach[M]. Lee & Seshia, 2011.
32. Thusoo A, Sarma JS,Jainn, etal. Hive-Apeta byte scale data ware house using Hadoop [C]. Proc. of ICDE2010. Piscataway, NJ:IEEE, 2010: 996-1005 Piscataway, NJ: IEEE, 2010: 996-1005
33. Abouzied A, Bajda-Pawlikowski K, Huang Jiewen, et al. HadoopDB in action: Building real world applications [C]. Proc. of SIGMOD 2010, New York: ACM, 2010: 1111-1114.
34. Chen Songting. Cheetah: A high performance, custom data warehouse on top of MapReduce [J]. PVLDB, 2010, 3(2): 1459-1468.
35. Agrawal R, Srikant R. Privacy preserving data mining [C]. Proc. of SIGMOD 2000. New York: ACM, 2000: 439-450.
36. Dwork C. Differential privacy [C]. Proc. of ICALP 2006. Berlin: Springer, 2006: 1-12.
37. Norman D A. The Design of Everyday Things [M].New York: Basic Books. 2002.
38. Olston C, Reed B, Srivastava U, et al. Pig Latin: A not-so-foreign language for data processing [C]. Proc of SIGMOD 2008, New York:ACM, 2008: 1099-1110.
39. Pike R, Dorward S, Griesemter R, et al. Interpreting the data: Parallel analysis with Sawzall [J]. Scientific Programming, 2005, 13(4): 277-298.
40. Chaiken R, Jenkins B, Larson P-A, et al. SCOPE: Easy and efficient parallel processing of massive data sets [J]. PVLDB, 2008, 1(2): 1265-1276.

41. Isard M, Yu Y. Distributed data-parallel computing using a high-level programming language [C]. Proc. of SIGMOD 2009. New York: ACM, 2009: 987-994.
42. Fegaras L, Li C, Gupta U, et al. XML query optimization in MapReduce [C]. Proc. of WebDB 2011. New York: ACM, 2011.
43. Morton K, Balazinska M, Grosstman D. Para Timer: A progress indicator for MapReduce DAGs [C]. Proc. of SIGMOD 2010. New York: ACM, 2010: 507-518.
44. Morton K, Friesen A, Balazinka A, et al. KAMD: Estimating the progress of MapReduce pipelines [C]. Proc. of ICDE 2010. Piscataway, NJ: IEEE, 2010: 681-684.
45. Huang Dachuan, Shi Xuanhua, Ibrabim Shadi, et al. MR- scope: A real-time tracing tool for MapReduce [C]. Proc. of HPDC 2010. New York: ACM, 2010: 849-855.
46. Meng Xiaofeng, Ci Xiang. Big Data Management: Concepts, Techniques and Challenges [J]. Journal of Computer Research and Development. 2013, 50(1): 146-169.
47. Chen Y. We dont know enough to make a big data benchmark suite-an academia-industry view[C]. Proc. of WBDB, 2012.
48. Chen Yanpei, Ganapathi A, Griffith R, et al. The case for evaluating MapReduce performance using workload suites [C]. Proc. of MASCOTS 2011. Piscataway, NJ: IEEE, 2011: 390-399.
49. Sangroya A, Serrano D, Bouchenak S. Mrbs: A comprehensive mapreduce benchmark suite[R]. LIG, Grenoble, France, Research Report RR-LIG-024, 2012.
50. Tan J, Kavulya S, Gandhi R, et al. Light-weight black-box failure detection for distributed systems[C]. Proceedings of the 2012 workshop on Management of big data systems. ACM, 2012: 13-18.
51. Zhao J M, Wang W S, Liu X, et al. Big data benchmark- big DS[M]. Advancing Big Data Benchmarks. Springer International Publishing, 2014: 49-57.
52. Patil S, Polte M, Ren K, et al. YCSB++: Benchmarking and performance debugging advance features in scalable table stores [C]. Proc. of SoCC 2011. New York: ACM, 2011.
53. Chen Y, Alspaugh S, Katz R. Interactive query processing in big data systems: A cross-industry study of MapReduce workloads [J]. PVLDB, 2012, 5(12).1802-1813.