A Framework for Big Data Analytics as a Scalable Systems

Himanshu Shekhar
Department of Computer Science and Engineering
Jaipur National University, Jaipur, Rajasthan, India
hshekhar0@gmail.com
Manoj Sharma
Department of Computer Science and Engineering
Jaipur National University, Jaipur, Rajasthan, India
manoj186@yahoo.co.in

---**ABSTRACT**---

Rapid technological advancements in recent technologies have led to inundation of data from diverse domains like scientific generated data, web, business intelligence and medical science over the past few years. The concept of big data was put to capture the meaning of this emerging vision. In comparison to the traditional data, big data exposes other unique characteristics apart from its massive volume. For instance, big data is fully distributed and unstructured. This emerging trend forces to create a new architecture for data acquisition, transmission, storage and analytics. In this paper, we present a framework for big data analytics in the form of a scalable systems aiming to provide an overall dimensions of big data system and thus to advance the big data challenges. First, we present a definition and background of big data. Next, we present a framework to define the architecture of big data. Finally, we delineate the challenges and future work to be carried out for big data systems.

Keywords- Big data analytics, data acquisition, data storage, data transmission, data analytics.

I. INTRODUCTION

Over the next decade, digital data in India will grow from 40,000 petabytes to 2.3 million petabytes, twice as fast as the world wide rate and will continue to grow exponentially and attract various diverse attentions. The term "big-data" was coined to capture this data explosion trend and certainly the data has been considered as the buzzword which is expected to explore our society. According to various industry estimates, the big-data industry crossed US$8 billion by 2012, and will reach US$16.9 billion by 2015 globally, growing at 7x the ICT industry growth rate. An IDC report [1] predicts that, the "digital universe" will grow to 2.7ZB in 2012, up 48% from 2011 and rocketing toward nearly 8ZB by 2015. According to NASSCOM [2], the big data market in India will grow at 83% annually to US$1 billion by 2015. The immense potential affiliated with big-data has led to an emerging research field that has promptly attracted enormous interest from diverse sectors, for example, government, industry and research community. The extensive interest is first illustrated by comprehending on both industrial reports [3] and public media (e.g., the Economist [4], [5], the New York Times [6]). The role of government played a crucial role in promoting new ideologies [7] and to stimulate the progress of tackling the big-data challenges. Finally, Nature and Science Magazines have published special issues to consider the big-data phenomenon and its challenges, flourishing its impact beyond technological sphere. Consequently, this emerging concern in big-data from various domains urges a clear and intuitive understanding of its definition, architecture, technologies and its various challenges.

This review paper focuses on scalable big-data systems, which include a set of techniques to explore, evaluate, validate, implement and deploy heterogeneous data while leveraging the massively multi processing power to perform complex analysis. Although, the uniqueness of big-data lies underneath, designing a big-data scalable system faces numerous technical challenges, including:

- Firstly, disparate data sources, related high costs and infrastructure bottlenecks are the key concern identified by companies for effectively managing unstructured data streams.
- Second, it need to capture and manage the heterogeneous datasets, while providing guaranteed services like scalability, privacy protection, and fast retrieval.
- Third, Datasets at different levels in real-time must be effectively mined by big data analytics including prediction, modeling, visualization and optimization.

These technological challenges call for modernizing the current data management systems, ranging from their architectural design to the execution details. Indeed, many leading industry [8] have disposed the transitional solutions to comprehend the emerging big data systems.

However, traditional data management, mainly based on relational database management system (RDBMS), is quite ineffective in handling the aforementioned big-data challenges. Thus the aspect of traditional RDBMS falls into the following two aspects:

- RDBMSs can only support structured data, while offering little support for unstructured or semi-structured data.

- RDBMSs scale up with expensive hardware and cannot scale out with current technology components in parallel.

To tackle these challenges, research community has suggested various solutions for big-data systems in an ad-hoc manner. Cloud computing can be considered as the infrastructure for big data platform to meet requirements, such as scalability, composability, reliability and ubiquitous access. For persistent storage and management of massive datasets, distributed file systems [9] and NoSQL [10] are employed. MapReduce [11], a software framework, has been hailed in processing group-aggregation tasks, such as website ranking. Hadoop [12] has also gained grip over distributed processing of data along with data storage and system management in order to build a powerful system-level solution which is becoming the linchpin in handling big-data challenges. We can fabricate various big-data applications based on these innovative technologies. The proliferation of big-data technologies demand a systematic framework that should be in order to capture the evolving big-data research and development efforts and apply those advancement in different area of subjects.

II. BIG DATA: DEFINITION AND PARADIGMS

A. BIG DATA DEFINITION

The definition of big-data is actually quite diverse. Essentially, big-data can be outlined as not only a large volume of data but also other features that distinguish it from the concepts of "massive data" and "very large data". In fact we came across several definitions of big-data and three types of definition play a crucial role in designing the shape of big-data:

- *Attributive Definition:* IDC can be considered as the innovator in big data domain and its impact. It defines big data in a 2011 report [13]: "Big data platform describe a new generation of technologies and architectures, designed to extract value economically from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis." The above "4Vs" definition thus clearly specifies the salient features of big-data, i.e., volume, variety, velocity and value and is widely adopted to qualify big data.
- *Comparative Definition*: According to 2011, Mckinsey's report [3] defines: "Big data can be the datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze." This definition integrates such aspects by which a dataset can be considered as big data.
- *Architectural Definition:* The National Institute of Standards and Technology (NIST) [14] states: "Big data is where the data volume, acquisition velocity, or data representation limits the ability to perform effective analysis using traditional relational approaches or requires the use of significant horizontal scaling for efficient processing."

Table 1. Big data attributes.

Parameters	Big Data
Structured	Semi or unstructured
Volume	Constantly updated
Data Source	Fully distributed
Data Store	NoSQL, HDFS
Data integration	Difficult
Generated Rate	Rapid
Access	Batch or near real time

B. BIG-DATA PARADIGMS: STREAMING VS. BATCH

Big data defines the use of analysis algorithms running on powerful supporting domains to uncover potentials concealed in big data, such as hidden patterns. According to the processing time requirement, big data analytics can be classified into two paradigms:

- *Streaming Processing:* The central point of streaming processing [15] lies in the fact that the potential value of data depends on data freshness. It thus analyzes data as soon possible to derive its results. The data arrives as continuous streams and thus the order in which data arrives cannot be controlled and result in scale of vast unending data which can thus be stored in limited memory.
- *Batch Processing:* The batch-processing defines the way to first store the available data and then perform analysis. MapReduce [11] has become the prominent example of batch-processing model. In MapReduce the data are first divided into small chunks and these chunks are then processed in distributed manner in order to generate intermediate results. The accumulation of all the intermediate results derives the final result.

Recently, real-time processing applications have also adopted the batch processing techniques to achieve a faster response.

Table 2. Comparison between streaming and batch processing

Parameters	Streaming processing	Batch processing
Input	Stream of new data	Data chunks
Storage	Not store	Store
Hardware	Typical single limited amount of memory	Multiple CPUs
Data size	Infinite or unknown	Finite and known
Time	A few seconds	Much longer
Processing	A single passes over data	Processed in multiple rounds
Applications	Web mining, sensor networks	Widely adopted in almost every domain

III. BIG DATA SYSTEM ARCHITECTURE

A. BIG DATA SYSTEM: A VALUE-CHAIN VIEW

A big-data system involves complex architecture which provides purpose and role to deal with various phases in the digital data life cycle, ranging from its birth to its destruction. To deal with different applications [16], [17] the system usually involves multiple distinct phases. In this

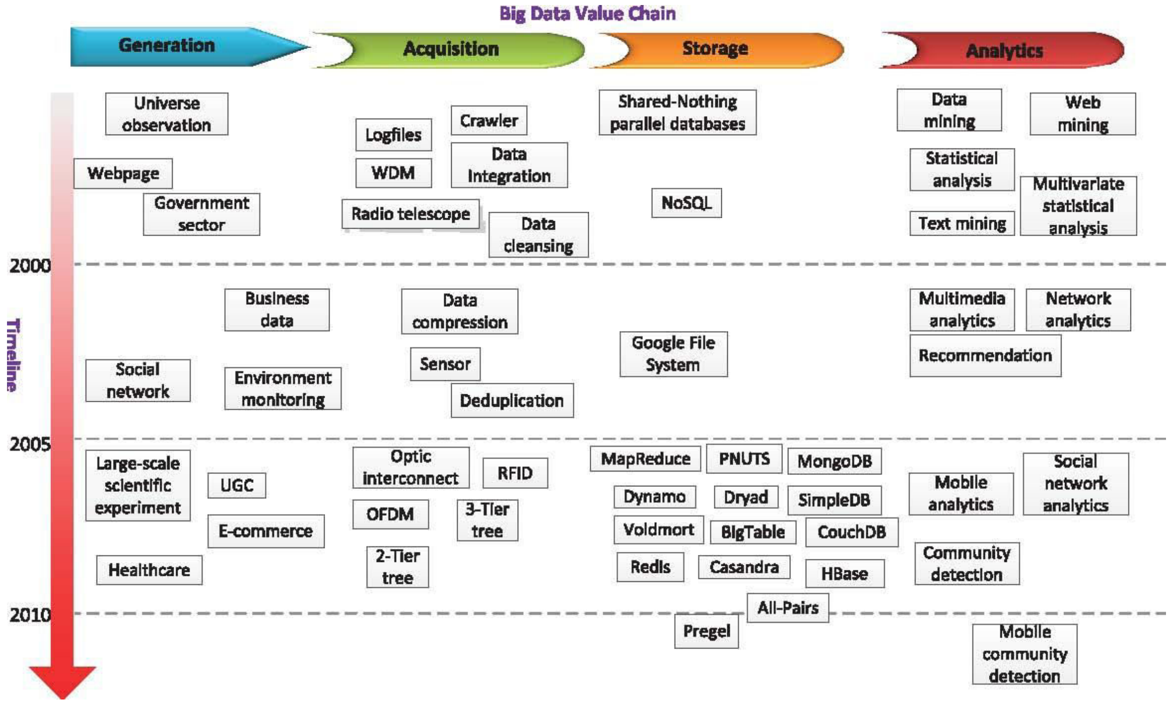


Figure 1. Big data technology map [60].

paper, we follow a systems-engineering pattern, well accepted in industry, [18], [19] to disintegrate a distinctive big-data system into four successive phases i.e., data generation, data acquisition, data storage, and data analytics.

Data generation refers how data is generated. In this context the term "big-data" is intended to consider the complex datasets that are generated from various distributed heterogeneous data source, including sensors, click streams, and videos. In this paper we consider data from three domains namely from business, Internet and scientific research.

Data acquisition concerns with the process of obtaining useful information from the generated data. It is further classified into data collection, data transmission, and data pre-processing. Firstly, data is collected from diverse source. Data collection thus defines the technology to obtain raw data from a specific data production environment. Secondly, the collected data uses high transmission mechanism to transmit the data for the purpose of storage for various types of analytical applications. Finally, data pre-processing operations is applied to eliminate the data redundancy for efficient storage and mining.

Data storage concerns storing and managing large datasets in a persistent storage system. A data storage system can be further subdivided into two parts: hardware infrastructure and data management. Hardware infrastructure consists of shared resources organized in a very flexible manner. Data management system is positioned on top of the hardware infrastructure to maintain large-scale datasets.

Data analysis pertained to the use of assorted analytical methods and tools to scrutinize, transform, and simulate data to extract value. Various field lay different application requirements and features. Emerging analytics research areas can be grouped into six technical domains: structured data analytics, text analytics, multimedia analytics, web analytics, network analytics, and mobile analytics.

B. BIG-DATA SYSTEM: A LAYERED VIEW

The big data system can be decomposed into a layered structure. To emphasize the complexity of a big data system, this layered view provides a conceptual hierarchy and is further classified into three layers, i.e., the infrastructure layer, the computing layer, and the application layer.

- The *infrastructure layer* consists of a cluster of ICT resources, which can be coordinated by cloud computing and altered by virtualization technology. A specific service-level agreement (SLA) exhibits these resources to upper-layer systems in a fine-grained manner.
- The *computing layer* encapsulates several big data tools including data management, data integration, and the programming model into a middleware layer that runs over raw ICT resources.
- The *application layer* feat the interface rendered by the programming models in order to enforce various data analysis functions, including querying, clustering, classification, and statistical analyses. McKinsey report presented five potential big data application domains: retails, health care, public sector administration, personal location data, and global manufacturing.

C. BIG-DATA SYSTEM CHALLENGES

The proposed definition of big data suggests that the big data is beyond the capability of current hardware and software platforms. Thus, designing and deployment of big data analytics is an unmanageable task. The present infrastructure in turn addresses wide range of challenges. In this paper, we exert much pressure to classify these challenges into three categories: data collection and management, data analytics, and system issues.

Data collection and management covers massive amount of disparate and complex data. It describes the following big data challenges:

- *Data Representation:* An effective data presentation should be designed to reflect the structure, hierarchy, and diversity of the heterogeneous data.
- *Redundancy Reduction and Data Compression:* Redundancy reduction and data compression are efficient way to reduce overall system overhead.
- *Data Life-Cycle Management:* Analysis value must associate data importance principle to decide which parts of data should be accepted and which parts should be discarded.
- *Data Privacy and Security:* To eliminate privacy leakage and to facilitate various analysis privacy supports must be provided at platform level.

The advances in big data analytics pose a significant impact including modeling, prediction, simulation, and interpretation. Unfortunately, data from the disparate sources present tremendous challenges:

- *Approximate Analytics:* Approximate analysis specify approximate query to handle real time requirement as data sets grow.
- *Connecting Social Media:* Connecting social media with inter-field data, application can achieve high levels of precision and distinct points of view.
- *Deep Analytics:* Advanced analytical technologies, like machine learning are requisite to unlock novel insights.

Finally, large-scale systems generally face several common issues, in particular.

- *Scalability:* All the components in big data system must be scalable to address the ever-growing size of complex heterogeneous data.
- *Energy Management:* From the point of economic perspective energy consumption has attracted greater concern as data volume increases. Hence, system-level power control and management mechanisms must be implemented.
- *Collaboration:* An extensive big data infrastructure is required to allow engineers and scientist to access the disparate data, and apply their respective skills to accomplish the goal of data analytics.

IV. PHASE I: DATA GENERATION

A. DATA SOURCES: TRENDS AND EXEMPLARY CATEGORIES

Recent trends of big data can be specified by the data generation rate. Specifically, the technological advancement leads to the increasing rate of data generation. IBM estimates that 90% of the data in the world today has been created in the past two years [20]. In this paper, therefore we classify data generation patterns into three consecutive phases:

- *Phase I*: The first phase began around 1990s and many organizations adopted database systems to store large volumes of data.
- *Phase II:* The second phase set out with the uprising of web systems. The Web 1.0 systems, leads to the growing popularity of search engines and ecommerce which in result generated huge amount of unstructured data. Further Web 2.0 created copious user generated data such as social networking.
- *Phase III:* The third phase sparked by the emergence of smart phones, tablets and sensor devices.

In this paper we present datasets from three distinctive domains.

1) BUSINESS DATA

Business data include the business transaction on the Internet which is estimated to double every 1.2 years [21] across all companies worldwide and is expected to reach 450 billion per day [22]. For example, every day, Amazon handles millions of back-end queries from more than half a million third party sellers [23].

2) NETWORKING DATA

Networking data includes data ranging from mobile network, social network to the websites and click streams. The advancement in the internet is generating the networking data at record speeds. For example, Facebook analyze access and store more than 30 PBs of user generated data. Over more than 30 billion searches were performed per month on Twitter [24].

3) SCIENTIFIC DATA

The advancement in scientific domain ensues generation of massive data from the scientific applications. Here, we highlight three crucial domains that are heavily relying on big data analytics. They are Astronomy [25], Computational Biology [26] and lastly High-Energy Physics [27].

B. DATA ATTRIBUTES

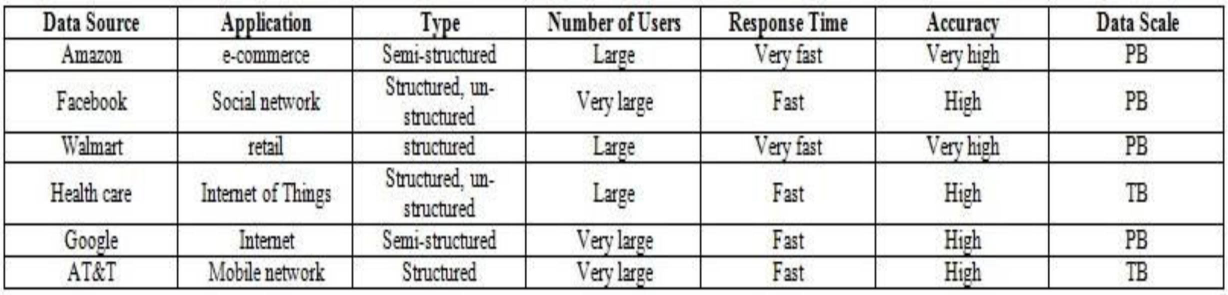
The progressive computing across internet including business sector, social environment, and government sectors is generating heterogeneous data with unprecedented rate and complexity. These datasets possesses distinctive characteristics. According to NIST [14] five attributes can be used to classify big data, and they are.

- *Volume* is the absolute volume of datasets.
- *Variety* refers to the structured, unstructured, and semi-structured data form.
- *Velocity* defines the rate of data generation.
- *Relational Limitation* includes particular queries and special forms of data.
- *Horizontal Scalability* represents the ability to join multiple datasets.

V. PHASE II: DATA ACQUISITION

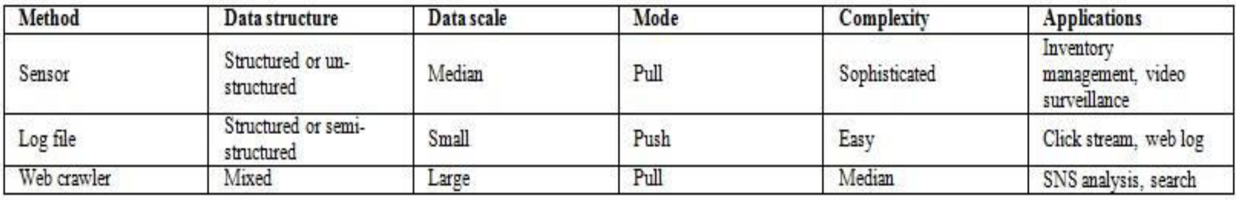
The undertaking of the data acquisition phase is to aggregate information in a digital form for further storage and analysis. This phase consist of three sub-phase, data collection, data transmission, and data pre-processing.

Table 3. Typical big data sources



Data Source	Application	Type	Number of Users	Response Time	Accuracy	Data Scale
Amazon	e-commerce	Semi-structured	Large	Very fast	Very high	PB
Facebook	Social network	Structured, un-structured	Very large	Fast	High	PB
Walmart	retail	structured	Large	Very fast	Very high	PB
Health care	Internet of Things	Structured, un-structured	Large	Fast	High	TB
Google	Internet	Semi-structured	Very large	Fast	High	PB
AT&T	Mobile network	Structured	Very large	Fast	High	TB

Table 4. Comparison for three data collection methods.



Method	Data structure	Data scale	Mode	Complexity	Applications
Sensor	Structured or un-structured	Median	Pull	Sophisticated	Inventory management, video surveillance
Log file	Structured or semi-structured	Small	Push	Easy	Click stream, web log
Web crawler	Mixed	Large	Pull	Median	SNS analysis, search

A. DATA COLLECTION

Data collection refers to the process of retrieving raw data from real-world objects in well designed manner. Along with the characteristics of data sources, the objectives of data analysis must also be considered while applying data collection methods. Here, we focus on three most common methods for big data collection.

1) SENSOR

Sensors are device used to capture a physical quantity from hardware units and convert it into a digital signal for further processing. Using wired or wireless networks, the converted information can then be transferred to a data collection point.

2) LOG FILE

Log files are generated by source systems in order to record activities in a specialized file format for later analysis. It is one of the most widely adopted data collection methods. Almost all digital devices uses log file for its running application.

3) WEB CRAWLER

A web crawler [28] is a program that downloads and stores web-pages for a search engine. Initially, a crawler inspects set of URLs in a queue and prioritized accordingly. The crawler then fetches a URL that has a certain priority, downloads the page, identifies all the URLs and then adds the new URLs to the queue. This process is repeated.

B DATA TRANSMISSION

The raw data collected must be transferred into a data storage infrastructure such as data center for subsequent processing. The transmission process can be further categorized into two stages, IP backbone transmission and data center transmission.

1) IP BACKBONE

The IP backbone, provide a high-capacity trunk route to channelize big data from its source to a data center at the Internet scale. The capacity and the transmission rate are determined by the physical media and the link management methods.

- *Physical Media* are primarily composed of many optical fiber cables bunched together to increase capacity.
- *Link Management* relates to how the signal is transmitted over the physical media.

2) DATA CENTER TRANSMISSION

Data center transmission refers to the process of analyzing, placement adjustment, and processing of big data after the big data is transmitted into the data center. It always associates with data center network architecture and transportation protocol:

- *Data Center Network Architecture*: A data center is a collection of server hosted in multiple racks connected via the data center's internal connection network. Most current data center internal connection networks follow a 2-tier or 3-tier architecture approach.
- *Transportation Protocol:* For the purpose of data transmission the most important network protocols are TCP and UDP; however, when huge amount of data to be transferred their performance degrades. Enhanced TCP improves link throughput while providing a small inevitable TCP flow latency. UDP is suitable for transferring a huge volume of data but lacks congestion control.

C. DATA PRE-PROCESSING

Since the data collected from diverse sources, they may have different levels of quality in terms of noise, consistency, redundancy, etc. In this section, we introduce three data pre-processing techniques.

1) INTEGRATION

Data integration refers to the process of combining data residing in diverse source and providing users with a unified view of the data [29]. Earlier, two approaches existed, the data warehouse method and the data federation method. The data warehouse method [29], highlights the process of extraction, transformation and loading of disparate data. The data federation method creates a virtual database to query and aggregate data from disparate sources.

2) CLEANSING

The data cleansing technique refers to the process of specifying and removing incomplete and inaccurate data with the objective to improve the quality of data. A universal framework consists of five steps for data cleansing:

- Define and determine error types.
- Search and identify error instances.
- Correct the errors.
- Document error instances and error types.
- Modify data entry procedures.

3) REDUNDANCY ELIMINATION

Data redundancy is the repetition of such data which is serving no useful purpose. It unnecessarily increases transmission overhead and results in concerning issues like data inconsistency, low reliability, wasted memory space and data corruption. Various redundancy reduction methods were put to use in different datasets, such as redundancy detection [30] and data compression [31].

VI. PHASE III: DATA STORAGE

The data storage scheme coordinates and organizes the collected information for the analysis and value extraction in a very practical and convenient manner. To carry out this purpose, the data storage scheme should offer two sets of features:

- The storage infrastructure must reconcile information reliably and persistently.
- To analyze and query immense quantity of data, the storage system must provide a scalable access interface.

The data storage system can be divided into two components: Hardware infrastructure and data management.

A. STORAGE INFRASTRUCTURE

Hardware infrastructure is used for physically storing the collected information. The storage infrastructure uses storage devices which can be classified based on the specific technology. Typical storage technologies include:

- *Random Access Memory (RAM):* RAM is data storage device associated with volatile types of memory, as it closes its information when powered-off. Modern RAM includes SRAM, DRAM, and PRAM.
- *Magnetic Disks:* Magnetic disks include hard disk drive (HDD) which is a non-volatile memory, as it retains its data even when powered off with lower per capacity cost, but the read and write operations are much slower.
- *Storage Class Memory:* Storage class memory relates to non-mechanical storage device, such as flash memory. Flash memory is used to fabricate solid-state drives (SSDs). SSDs have no mechanical components, have less latency and lower access times.

The storage infrastructure can also be interpreted from a networking architecture perspective [32] and can be organized in different modes, which are as follows:

- *Direct Attached Storage (DAS):* DAS is a storage system consisting of a collection of data storage devices. Using the host bus adapter (HBS), these devices are connected directly to a computer with no storage network between them.
- *Network Attached Storage (NAS):* NAS is file-level storage consisting of many hard drives ordered into logical, redundant storage containers. NAS provides storage as well as file system, and can be considered as a file server.
- *Storage Area Network (SAN):* SANs possesses the most complicated architecture and work as dedicated networks that provide block-level storage to a group of computers. SANs can merge several storage devices and make them accessible to computers.

B. DATA MANAGEMENT FRAMEWORK

The data management framework relates how to organize the information in a convenient manner for effective processing. In this paper, we adopt a layered view to classify the data management framework into three layers that consist of file systems, database technology, and programming models.

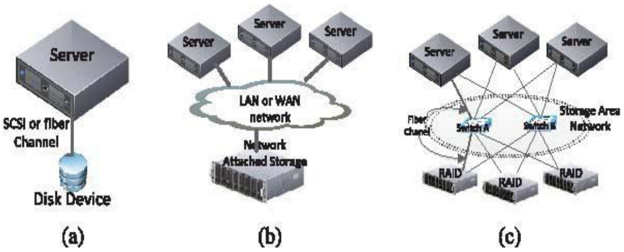


Figure 2. Network architecture of storage systems.

1) FILE SYSTEMS

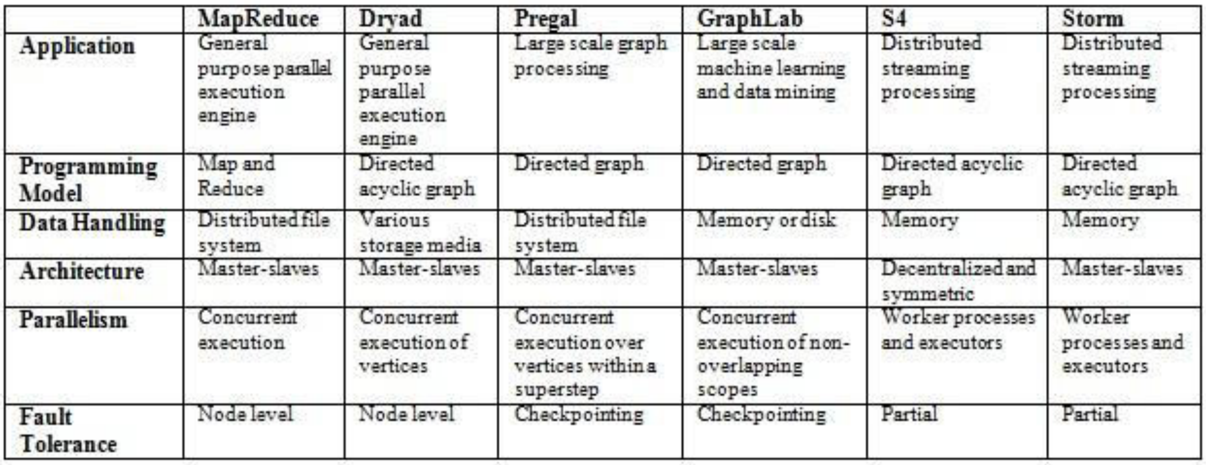
The file system provides the basis of data storage for both academy and industry. We consider such system which are open source or designed for enterprise use.

Google implemented GFS as distributed file system [33] for large distributed data intensive applications. It uses relatively low commodity servers to provide high performance and fault tolerance to a large number of clients. HDFS [34] and Kosmosfs [35] are open source derivatives of GFS. Facebook designed Haystack [36] to store a huge amount of small-file photos. Taobao proposed Tao File System (TFS) [37] and FastDFS [38] much similar to Haystack.

2) DATABASE TECHNOLOGIES

Different scale of datasets from diverse applications enforced the proposal of various database systems. Traditional database system can no longer handle the variety and scale challenges demanded by big data. The NoSQL database is becoming touchstone to cope with big data problems as it includes certain essential characteristics, including being schema free, eventual consistency, supporting easy replication, possessing a simple API, and supporting huge amount of data. We present three primary types of NoSQL databases organized by data model. They are key-value stores, column-oriented databases, and document databases.

Table 3. Feature summary of programming models



	MapReduce	Dryad	Pregal	GraphLab	S4	Storm
Application	General purpose parallel execution engine	General purpose parallel execution engine	Large scale graph processing	Large scale machine learning and data mining	Distributed streaming processing	Distributed streaming processing
Programming Model	Map and Reduce	Directed acyclic graph	Directed graph	Directed graph	Directed acyclic graph	Directed acyclic graph
Data Handling	Distributed file system	Various storage media	Distributed file system	Memory or disk	Memory	Memory
Architecture	Master-slaves	Master-slaves	Master-slaves	Master-slaves	Decentralized and symmetric	Master-slaves
Parallelism	Concurrent execution	Concurrent execution of vertices	Concurrent execution over vertices within a superstep	Concurrent execution of non-overlapping scopes	Worker processes and executors	Worker processes and executors
Fault Tolerance	Node level	Node level	Checkpointing	Checkpointing	Partial	Partial

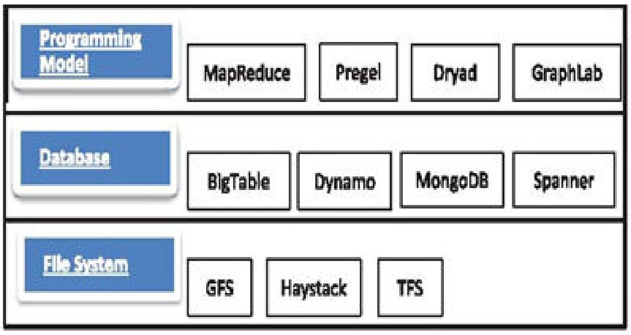


Figure 3. Data Management Technology [60]

A) KEY-VALUE STORES

In key-value data model, data are stored in the form of key-value pair. Each of the keys is unique, and clients request values for each key. The emergence of key-value databases in recent years have been heavily influenced by Amazon's Dynamo [39].

B) COLUMN ORIENTED DATABASES

Column oriented databases stores and process data in column form. Rows and column will be split over multiple nodes to attain scalability. The idea of column oriented databases was adopted from the Google's Bigtable [40].

C) DOCUMENT DATABASES

Document databases support more complex data structures than key-value stores. There is no strict schema to which documents must conform. The major representative of document databases includes SimpleDB, MongoDB [41], CouchDB.

D) OTHER NoSQL AND HYBRID DATABASES

Apart from the aforementioned data stores, many other types have been implemented to support different types of data stores such as DEX [42], PNUTS [43] and Neo4j [44].

3) PROGRAMMING MODELS

The programming model is designed to implement the application logics and facilitate the data analysis applications. Many programming models have been proposed to resolve the domain-specific applications. We briefly introduce three types of process models: the generic processing model, the graph processing model, and the stream processing model.

- *Generic Processing Model:* This processing model deals with the general application problems and is used in MapReduce [11] and its variants, and in Dryad.
- *Graph Processing Model:* This type of processing model expresses the growing class of applications in terms of related entities and implemented using graphic models in one or other iterative forms. Among this model, most popular are Pregel [45] and GraphLab [46].
- *Stream Processing Model:* Storm [47] and S4 [48] are two of the most distributed stream processing model that are implemented on the JVM. The major difference between them is in their architecture i.e., Storm is a master-slave system such as MapReduce, whereas S4 follows a decentralized and symmetric architecture.

VII. PHASE IV: DATA ANALYSIS

The most important phase of the big data value chain is data analysis, which emphasize the aim to extract useful data, propose conclusions and support decision making.

A PURPOSE AND CATEGORIES

Data analytics deals with the selective information incurred through observation, measurement, or experiments about a subject of interest. Data analytics aim to extract useful information from the subject that is targeted under consideration. The subject nature and purpose may vary greatly. Few of the potential purposes are listed below:

- To generalize and interpret the data,
- To suggest and assist decision-making,
- To diagnose and derive reasons for fault,
- To check whether the data are legitimate, and
- To predict future occurrence.

According to the depth of analysis, Blackett et al. [49] classified data into three levels: descriptive analysis, predictive analysis, and prescriptive analysis.

- *Descriptive Analysis:* In this historical data is exploited to trace what occurred.
- *Predictive Analysis:* It aims to focus on predicting future trends and probabilities.
- *Prescriptive Analysis:* It supports decision making and efficiency.

B. APPLICATION EVOLUTION

More recently, the explosion of massive data urges big data analytics to describe the advanced analysis methods or mechanisms. In the following, we reveal different application evolution of data analysis.

1) BUSINESS APPLICATION EVOLUTION

The earliest business data were unstructured data, and were stored in relational database management system. The analysis techniques which came to use were quite simple and intuitive. The common business methods include dashboards, scorecards, data mining, search-based intelligence, reporting, and online transaction processing. Currently, the web offered a unique chance for organizations to present their businesses online. Various web mining techniques can be applied to inspect product placement optimization, product recommendations, customer transaction analysis and market structure analysis.

2) NETWORK APPLICATION EVOLUTION

The earlier network mainly offer web service and email facility. These services were carried out using data mining, text analysis and webpage analysis techniques. Currently, network data prevails and capture the global data volumes as almost all application run on network regardless of their domains. The web intertwined diverse type of data, including text, audio, video and photos.

3) SCIENTIFIC APPLICATION EVOLUTION

Current areas of scientific field are gleaning a huge volume of data from high throughput instruments ranging from astrophysics to genomics. Recently, National Science Foundation (NSF) declared the *BIGDATA* program solicitation [50] to ease information sharing and data analytics. Previously, several scientific research domains has developed huge data platform and reaped the resulting benefits.

C. COMMON METHODS

Some of the common methods applied to almost all of the analysis are:

- *Data visualization:* It relates closely to information visualization and information graphics. The data visualization aim to transmit information distinctly and effectively through graphical means [51]. Charts and graphs help people understand quickly and easily.
- *Statistical analysis:* It is based on statistical theory which employs probability theory to model randomness and uncertainty. For large datasets, statistical analysis can serve for description and interference.
- *Data mining:* It is relatively a process of discovering hidden patterns in large data sets. Various data mining algorithms have been developed in the artificial intelligence, pattern recognition, machine learning and statistics. Some of the most influential algorithms include C4.5, k-means, SVM, a priori, EM, PageRank, AdaBoost, kNN, Naive Bayes, and CART.

VIII. CASES IN POINT OF BIG DATA ANALYTICS

We present six types of big data application, organized by data type: structured data analytics, text analytics, web analytics, multimedia analytics, network analytics, and mobile analytics.

A. STRUCTURED DATA ANALYTICS

A huge amount of structured data is generated from the scientific research field and business sector. Management of these data depends on the mature RDBMS, OLAP, and data warehousing. Recently, deep learning, a set of machine- learning methods is becoming an active research area. Statistical machine learning, based on exact mathematical models has already been applied in anomaly detection [52]. Currently, process mining [53] has emerged as a new research area that focuses on event data and process discovery and conformance-checking techniques.

B. TEXT ANALYTICS

Text Analytics, also known as text mining, is the process of extracting useful information and knowledge from unstructured text. Text includes webpages, e-mail communication, social media content and corporate documents. Hence, it is considered to be of higher commercial potential. Text mining is an interdisciplinary field at the intersection of computational linguistics, information retrieval, machine learning, statistics, and data mining. Text analytics is particularly based on text representation and natural language processing (NLP) which can enhance the available information related to text terms.

C. WEB ANALYTICS

Web Analytics aims to retrieve, extract, and evaluate information for knowledge discovery from web documents and services automatically. Web analytics is based on several other platforms such as databases, NLP, text mining, and information retrieval. Web analytics is categorized into three areas of interest: web content mining, web structure mining, and web usage mining [54].

Web content mining is the discovery of useful information from website content which includes data such as text, image, audio, video, and hyperlinks.

Web structure mining is the discovery of the model underlying link structures on the web. Here, structure represents the graph of links in a site.

Web usage mining refers to mining secondary data generated by web sessions. It includes data from proxy server logs, browser logs, web server access logs, user profile, user sessions, and cookies.

D. MULTIMEDIA ANALYTICS

Multimedia analytics refers to extracting interesting knowledge and understanding the semantics captured in multimedia data. Recent area of subjects under multimedia analytics includes multimedia annotation, multimedia summarization, multimedia indexing and retrieval.

Multimedia annotation refers to assigning images and videos a set of labels that depict their contents at syntactic levels.

Multimedia summarization includes audio and video summarization that extracts and synthesizes a portion of data from original contents.

Multimedia indexing and retrieval employs the description, storage, and organization of multimedia information for convenient and fast retrieval.

E. NETWORK ANALYTICS

The origin of the network analysis include sociology network analysis [55], bibliometric analysis [56] and is deviating towards emerging social network analysis due to rapid growth of online social networks. Social networks contain enormous amount of linkage and content data, where linkage data represent the communication between graph structures among entities and content data contains text, images and other multimedia content in the networks. Social networks can be visualized as graphs, in which a vertex represents a person, and an edge corresponds to associations between the corresponding persons.

The revolutionary development of Web 2.0, leads to the explosion of user-generated content on social networks. The term *social media* is coined to describe such user-generated content, including images, videos, blogs, micro blogs, social networking sites, social news, social book marketing and wikis. However, social media analytics face certain challenges. First, there are tremendous and ever-growing social media data, which must be analyzed within a reasonable time constraint. Second, social networks are dynamic, ever-changing and are updated rapidly. Third, social media data contains many noisy data.

F. MOBILE ANALYTICS

With the rapid advancement of mobile computing [57]-[59], more mobile terminals and applications including mobile phones, RFID and sensors are positioned globally. Currently, mobile analysis is far from mature but is facing certain critical challenges caused by the inherent characteristics of mobile data, such as redundancy richness, noisiness, mobile awareness, and activity sensitivity.

Recent advances in mobile technology and wireless sensors have triggered the deployment of body sensor networks for real time-monitoring of an individual health.

IX. CONCLUSION

The prevalence of big data in present era brings an urgent need for the advanced data acquisition, management, and analysis mechanisms. In this paper, we have presented the concept of big data and highlighted the big data value chain. The big data value chain consists of four stages: data generation, data acquisition, data storage, and data analysis. We have further presented numerous mechanisms in different stages. In generation phase, we have listed several data sources and data attributes. In acquisition phase, data collection methods were checked, followed by data transmission and data pre-processing methods. In storage phase, NoSQL stores were introduced. Finally, in the big data analytics phase, we have investigated various data analytics methods organized according to data characteristics.

REFERENCES

[1] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east,'' in *Proc. IDC iView, IDC Anal. Future*, 2012.

[2] (2014). Big Data - The Next Big Thing [Online]. Available: http://www.crisil.com/global-offshoring/gra- nasscom.html

[3] J. Manyika *et al., Big Data: The Next Frontier for Innovation, Competition, and Productivity*. San Francisco, CA, USA: McKinsey Global Institute, 2011, pp. 1-137.

[4] K. Cukier, "Data, data everywhere," *Economist*, vol. 394, no. 8671, pp.3-16, 2010.

[5] T. economist. (2011, Nov.) *Drowning in Numbers—Digital Data Will Flood the Planet-and Help us Understand it Better* [Online]. Available: http://www.economist.com/blogs/dailychart/2011/11/bigdata-0.

[6] S. Lohr. (2012). The age of big data. *New York Times* [Online]. *11*. Available: http://www.nytimes.com/2012/02/12/sunday-review/big-datasimpact-in-the-world.html?pagewanted=all&r=0.

[7] W. House. (2012, Mar.). *Fact Sheet: Big Data Across the Federal Government* [Online]. Available: http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data%_fact_sheet_3_29_2012.pdf.

[8] Wiki. (2013). *Applications and Organizations Using Hadoop* [Online]. Available: http://wiki.apache.org/hadoop/ PoweredBy

[9] J. H. Howard *et al.*, ''Scale and performance in a distributed file system,'' *ACM Trans. Comput. Syst.*, vol. 6, no. 1, pp. 51–81, 1988.

[10] R. Cattell, ''Scalable SQL and NoSQL data stores,'' *SIGMOD Rec.*, vol. 39, no. 4, pp. 12–27, 2011.

[11] J. Dean and S. Ghemawat, ''Mapreduce: Simplified data processing on large clusters,'' *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[12] T. White, *Hadoop: The Definitive Guide*. Sebastopol, CA, USA: O'Reilly Media, 2012.

[13] J. Gantz and D. Reinsel, ''Extracting value from chaos,'' in *Proc. IDC iView*, 2011, pp. 1–12.

[14] M. Cooper and P. Mell. (2012). *Tackling Big Data* [Online]. Available: http://csrc.nist.gov/groups/SMA/forum/documents/june2012presentations/f%csm_june2012_cooper_mell.pdf

[15] N. Tatbul, ''Streaming data integration: Challenges and opportunities,'' in *Proc. IEEE 26th Int. Conf. Data Eng. Workshops (ICDEW)*, Mar. 2010, pp. 155–158.

[16] E. B. S. D. D. Agrawal *et al.*, ''Challenges and opportunities with big data—A community white paper developed by leading researchers across the united states,'' The Computing Research Association, CRA White Paper, Feb. 2012.

[17] D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker, ''Interactions with big data analytics,'' *Interactions*, vol. 19, no. 3, pp. 50–59, May 2012.

[18] F. Gallagher. (2013). *The Big Data Value Chain* [Online]. Available: http://fraysen.blogspot.sg/2012/06/big-data-value-chain.html.

[19] M. Sevilla. (2012). *Big Data Vendors and Technologies, the list!* [Online]. Available: http://www.capgemini.com/blog/capping-it-off/2012/09/big-data-vendors-a%nd-technologies-the-list

[20] (2013). *What is Big Data*, IBM, New York, NY, USA [Online]. Available: http://www-01.ibm.com/software/data/bigdata/

[21] Knowwpc. (2013). *eBay Study: How to Build Trust and Improve the Shopping Experience* [Online]. Available:http://knowwpcarey.com/article.cfm?aid=1171

[22] J. Gantz and D. Reinsel, ''The digital universe decade-are you ready,'' in *Proc. White Paper, IDC*, 2010.

[23] J. Layton. (2013). *How Amazon Works* [Online]. Available:http://knowwpcarey.com/article.cfm?aid=1171

[24] Wikibon. (2013). *A Comprehensive List of Big* Data Statistics [Online]. Available: http://wikibon.org/blog/big-data-statistics/

[25] (2013). *SDSS* [Online]. Available: http://www.sdss.org/

[26] R. E. Bryant, ''Data-intensive scalable computing for scientific applications,'' *Comput. Sci. Eng.*, vol. 13, no. 6, pp. 25–33, 2011.

[27] (2013). *Atlas* [Online]. Available: http://http://atlasexperiment.org/

[28] J. Cho and H. Garcia-Molina, ''Parallel crawlers,'' in *Proc. 11th Int. Conf. World Wide Web*, 2002, pp. 124–135.

[29] M. Lenzerini, ''Data integration: A theoretical perspective,'' in *Proc. 21st ACM SIGMOD-SIGACT-SIGART Symp. Principles Database Syst.*, 2002, pp. 233–246.

[30] Y. Zhang, J. Callan, and T. Minka, ''Novelty and redundancy detection in adaptive filtering,'' in *Proc. 25th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retr.*, 2002, pp. 81–88.

[31] D. Salomon, *Data Compression*. New York, NY, USA: Springer-Verlag, 2004.

[32] U. Troppens, R. Erkens, W. Mueller-Friedt, R. Wolafka, and N. Haustein, *Storage Networks Explained: Basics and Application of Fibre Channel SAN, NAS, ISCSI, Infiniband and FCoE*. New York, NY, USA: Wiley, 2011.

[33] S. Ghemawat, H. Gobioff, and S.-T. Leung, ''The Google file system,'' in *Proc. 19th ACM Symp. Operating Syst. Principles*, 2003, pp. 29–43.

[34] (2013). *Hadoop Distributed File System* [Online]. Available: http://hadoop.apache.org/ docs/r1.0.4/hdfsdesign.html

[35] (2013). *Kosmosfs* [Online]. Available: https://code.google.com/p/kosmosfs/

[36] D. Beaver, S. Kumar, H. C. Li, J. Sobel, and P. Vajgel, ''Finding a needle in Haystack: Facebook's photo storage,'' in *Proc. 9th USENIX Conf. Oper. Syst. Des. Implement. (OSDI)*, 2010, pp. 1–8.

[37] (2013). *Taobao File System* [Online]. Available: http://code.taobao.org/p/tfs/src/

[38] (2013). *Fast Distributed File System* [Online]. Available: http://code.google.com/p/fastdfs/

[39] G. DeCandia *et al.*, ''Dynamo: Amazon's highly available key-value store,'' *SIGOPS Oper. Syst. Rev.*, vol. 41, no. 6, pp. 205–220, 2007.

[40] F. Chang *et al.*, ''Bigtable: A distributed storage system for structured data,'' *ACM Trans. Comput. Syst.*, vol. 26, no. 2, pp. 4:1–4:26, Jun. 2008.

[41] (2013). *MongoDB* [Online]. Available: http://www.mongodb.org/

[42] (2013). *Dex* [Online]. Available: http://www.sparsitytechnologies. com/dex.php

[43] B. F. Cooper *et al.*, ''PNUTS: Yahoo!'s hosted data serving platform,'' in *Proc. VLDB Endowment*, vol. 1, no. 2, pp. 1277–1288, 2008.

[44] (2013). *Neo4j* [Online]. Available: http://www.neo4j.org/

[45] G. Malewicz *et al.*, ''Pregel: A system for large-scale graph processing,'' in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, Jun. 2010, pp. 135–146.

[46] Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola, and J. M. Hellerstein, ''Distributed graphlab: A framework for machine learn learning and data mining in the cloud,'' *Proc. VLDB Endowment*, vol. 5, no. 8, pp. 716–727, 2012.

[47] (2013). *Storm* [Online]. Available: http://storm-project.net/

[48] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari, ‘‘S4: Distributed stream computing platform,’’ in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Dec. 2010, pp. 170–177.

[49] G. Blackett. (2013). *Analytics Network-O.R. Analytics* [Online]. Available: http://www.theor society.com/Pages/SpecialInterest /AnalyticsNetwork_anal%ytics.aspx

[50] N. S. Foundation. (2013). *Core Techniques and Technologies for Advancing Big Data Science and Engineering* [Online]. Available: http://www.nsf.gov /pubs/2012/nsf12499/nsf12499.htm

[51] V. Friedman. (2008). *Data visualization and infographics* [Online]. Available: http://www. smashingmagazine.com/2008/01/14/monday-inspiration-data-visualization-and-infographics/

[52] G. K. Baah, A. Gray, and M. J. Harrold, ‘‘On-line anomaly detection of deployed software: A statistical machine learning approach,’’ in *Proc. 3rd Int. Workshop Softw. Qual. Assurance*, 2006, pp. 70–77.

[53] W. van der Aalst, ‘‘Process mining: Overview and opportunities,’’ *ACM Trans. Manag. Inform. Syst.*, vol. 3, no. 2, pp. 7:1–7:17, Jul. 2012.

[54] S. K. Pal, V. Talwar, and P. Mitra, ‘‘Web mining in soft computing framework: Relevance, state of the art and future directions,’’ *IEEE Trans. Neural Netw.*, vol. 13, no. 5, pp. 1163–1177, 2002.

[55] D. J. Watts, *Six Degrees: The Science of a Connected Age*. New York, NY, USA: Norton, 2004.

[56] J. E. Hirsch, ‘‘An index to quantify an individual’s scientific research output,’’ *Proc. Nat. Acad. Sci. United States Amer.*, vol. 102, no. 46, p. 16569, 2005.

[57] H. Zhang, Z. Zhang, and H. Dai, ‘‘Gossip-based information spreading in mobile networks,’’ *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, pp. 5918–5928, Nov. 2013.

[58] H. Zhang, Z. Zhang, and H. Dai, ‘‘Mobile conductance and gossip-based information spreading in mobile networks,’’ in *IEEE Int. Symp. Inf. Theory Proc. (ISIT)*, Jul. 2013, pp. 824–828.

[59] H. Zhang, Y. Huang, Z. Zhang, and H. Dai. (2014). Mobile conductance in sparse networks and mobility-connectivity tradeoff. in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)* [Online]. Available: http://www4. ncsu.edu/?hdai/ISIT2014-HZ.pdf .

[60] Han Wo. , Yonggang Wen. (2014). “Toward scalable systems for big data Analytics”, vol . 2, pp. 657.