Toward Scalable Systems for Big Data Analytics: A Technology Tutorial (I - III)

ABSTRACT Recent technological advancement have led to a deluge of data from distinctive domains (e.g., health care and scientific sensors, user-generated data, Internet and financial companies, and supply chain systems) over the past two decades. The term big data was coined to capture the meaning of this emerging trend. In addition to its sheer volume, big data also exhibits other unique characteristics as compared with traditional data. For instance, big data is commonly unstructured and require more real-time analysis. This development calls for new system architectures for data acquisition, transmission, storage, and large-scale data processing mechanisms. In this paper, we present a literature survey and system tutorial for big data analytics platforms, aiming to provide an overall picture for nonexpert readers and instill a do-it-your self spirit for advanced audiences to customize their own big-data solutions. First, we present the definition of big data and discuss big data challenges. Next, we present a systematic framework to decompose big data systems into four sequential modules, namely data generation, data acquisition, data storage, and data analytics. These four modules form a big data value chain. Following that, we present a detailed survey of numerous approaches and mechanisms from research and industry communities. In addition, we present the prevalent Hadoop framework for addressing big data challenges. Finally, we outline several evaluation benchmarks and potential research directions for big data systems.

INDEX TERMS Big data analytics, cloud computing, data acquisition, data storage, data analytics, Hadoop

---

I. INTRODUCTION

The emerging big-data paradigm, owing to its broader impact, has profoundly transformed our society and will continue to attract diverse attentions from both technological experts and the public in general. It is obvious that we are living a data deluge era, evidenced by the sheer volume of data from a variety of sources and its growing rate of generation. For instance, an IDC report predicts that, from 2005 to 2020, the global data volume will grow by a factor of 300, from 130 exabytes to 40,000

exabytes, representing a double growth every two years. The term of “big-data” was coined to capture the profound meaning of this data-explosion trend and indeed the data has been touted as the new oil, which is expected to transform our society. For example, a Mckinsey report states that the potential value of global personal location data is estimated to be $100 billion in revenue to service providers over the next ten years and be as much as $700 billion in value to consumer and business end users. The huge potential associated with big-data has led to an emerging research field that has quickly attracted tremendous interest from diverse sectors, for example, industry, government and research community. The broad interest is first exemplified by coverage on both industrial reports and public media (e.g., the Economist, the New Your Times, and the National Public Radio (NPR)). Government has also played a major role in creating new programs to accelerate the progress of tackling the big data challenges. Finally, Nature and Science Magazines have published special issues to discuss the big-data phenomenon and its challenges, expanding its impact beyond technological domains. As a result, this growing interest in big-data from diverse domains demands a clear and intuitive understanding of its definition, evolutionary history, building technologies and potential challenges.

This tutorial paper focuses on scalable big-data systems, which include a set of tools and mechanisms to load, extract, and improve disparate data while leveraging the massively parallel processing power to perform complex transformations and analysis. Owing to the uniqueness of big-data, designing a scalable big-data systems faces a series of technical challenges, including:

- First, due to the variety of disparate data sources and the sheer volume, it is difficult to collect and integrate data with scalability from distributed locations. For instance, more than 175 million tweets containing text, image, video, and social relationship are generated by millions of accounts distributed globally.
- Second, big data systems need to store and manage the gathered massive and heterogeneous datasets, while provide function and performance guarantee, in terms of fast retrieval, scalability, and privacy protection. For example, Facebook needs to store, access, and analyze over 30 pertabytes of user generate data.
- Third, big data analytics must effectively mine massive datasets at different levels

in realtime or near realtime - including modeling, visualization, prediction, and optimization - such that inherent promises can be revealed to improve decision making and acquire further advantages.

These technological challenges demand an overhauling re-examination of the current data management systems, ranging from their architectural principle to the implementation details. Indeed, many leading industry companies have discarded the transitional solutions to embrace the emerging big data platforms.

However, traditional data management and analysis systems, mainly based on relational database management system (RDBMS), are inadequate in tackling the aforementioned list of big-data challenges. Specifically, the mismatch between the traditional RDBMS and the emerging big-data paradigm falls into the following two aspects, including:

- From the perspective of data structure, RDBMSs can only support structured data, but offer little support for semi-structured or unstructured data.
- From the perspective of scalability, RDBMSs scale up with expensive hardware and cannot scale out with commodity hardware in parallel, which is unsuitable to cope with the ever growing data volume.

To address these challenges, the research community and industry have proposed various solutions for big data systems in an ac-hoc manner. Cloud computing can be deployed as the infrastructure layer for big data systems to meet certain infrastructure requirements, such as cost-effectiveness, elasticity, and the ability to scale up or down. Distributed file systems and NoSQL databases are suitable for persistent storage and the management of massive scheme free datasets. MapReduce, a programming framework, has achieved great success in processing group-aggregation tasks, such as website ranking. Hadoop integrates data storage, data processing, system management, and other modules to form a powerful system-level solution, which is becoming the mainstay in handling big data challenges. We can construct various big data applications based on these innovative technologies and platforms. In light of the proliferation of big-data technologies, a systematic framework should be in order to capture the fast evolution of big-data research and development efforts and put the development in different frontiers in perspective.

In this paper, learning from our first-hand experience of building a big-data solution on our private modular data center testbed, we strive to offer a systematic tutorial for scalable big-data systems, focusing on the enabling technologies and the architectural principle. It is our humble expectation that the paper can server as a first stop for domain experts, big-data users and the general audience to look for information and guideline in their specific needs for big-data solutions. For example, the domain experts could follow our guideline to develop their own big-data platform and conduct research in big-data domain; the big-data users can use our framework to evaluation alternative solutions proposed by their vendors; and the general audience can understand the basic of big-data and its impact on their work and life. For such a purpose, we first present a list of alternative definitions of big data, supplemented with the history of big-data and big-data paradigms. Following that, we introduce a generic framework to decompose big data platforms into four components, i.e., data generation, data acquisition, data storage, and data analysis. For each stage, we survey current research and development efforts and provide engineering insights for architecture design. Moving toward a specific solution, we then delve on Hadoop - the de facto choice for big data analysis platform, and provide benchmark results for big-data platforms.

The rest of this paper is organized as follows. In Section II, we present the definition of big data and its brief history, in addition to processing of big data and its brief history, in addition to processing paradigms. Then, in Section III, we introduce the big data value chain (which is composed of four phases), the big data technology map, the layered system architecture and challenges. The next four sections describe the different big data phases associated with the big data value chain. Specifically, Section IV focuses on big data generation and introduces representative big data sources. Section V discusses big data acquisition and presents data collection, data transmission, and data preprocessing technologies. Section VI investigates big data storage approaches and programming models. Section VII discuss big data analytics, and several applications are discussed in Section VIII. Section IX introduces Hadoop, which is the current mainstay of the big data movement. Section X outlines several benchmarks for evaluating the performance of big data systems. A brief conclusion with recommendations for future studies is presented in Section XI.

II. BIG DATA: DEFINITION, HISTORY AND PARADIGMS

In this section, we first present a list of popular definitions of big data, followed by a brief history of its evolution. This section also discusses two alternative paradigms, streaming processing and batch processing.

A. BIG DATA DEFINITION

Given its current popularity, the definition of big data is rather diverse, and reaching a consensus is difficult. Fundamentally, big data means not only a large volume of data but also other features that differentiate it from the concepts of "massive data" and "very large data". In fact, several definitions for big data are found in the literature, and three types of definitions play an important role in shaping how big data is viewed:

- Attribute Definition: IDC is a pioneer in studying big data and its impact. It defines big data in 2011 report that was sponsored by EMC (the cloud computing leader): "Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis." This definition delineates the four salient feature of big data, i.e., volume, variety, velocity and value. As a result, the "4Vs" definition has been used widely to characterize big data. A similar description appeared in 2011 research report in which META group (now Gartner) analyst Doug Laney noted that data growth challenges and opportunities and three-dimensional, i.e., increasing volume, velocity, and variety. Although this description was not meant originally to define big data, Gartner and much of the industry, including IBM and certain Microsoft researchers, continue to use this "3Vs" model to describe big data 10 years later.
- Comparative Definition: In 2011, Mckinsey's report defined big data as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze." This definition is subjective and does not define big data in terms of any particular metric. However, it incorporates an evolutionary aspect in the definition (over time or across sectors) of what a dataset must be to be considered as big data.
- Architectural Definition: The National Institute of Standards and Technology (NIST)

suggests that, “Big data is where the data volume, acquisition velocity, or data representation limits the ability to perform effective analysis using traditional relational approaches of requires the use of significant horizontal scaling for efficient processing.” In particular, big data can be further categorized into big data science and big data frameworks. Big data science is “the study of techniques covering the acquisition, conditioning, and evaluation of big data,” whereas big data frameworks are “software libraries along with their associated algorithms that enable distributed processing and analysis of big data problems across clusters of computer units”. An instantiation of one or more big data frameworks is known as big data infrastructure.

Concurrently, there has been much discussion in various industries and academia about what big data actually means.

However, reaching a consensus about the definition of big data is difficult, if not impossible. A logical choice might be to embrace all the alternative definitions, each of which focuses on a specific aspect of big data. In this paper, we take this approach and embark on developing an understanding of common problems and approaches in big data science and engineering.

C. BIG-DATA PARADIGMS: STRAMING VS. BATCH

Big data analytics is the process of using analysis algorithms running on powerful supporting platforms to uncover potential concealed in big data, such as hidden patterns or unknown correlations. According to the processing time requirement, big data analytics can be categorized into two alternative paradigms:

- Streaming Processing: The start point for the streaming processing paradigm is the assumption that the potential value of data depends on data freshness. Thus, the streaming processing paradigm analyzes data as soon as possible to derive its results. In this paradigm, data arrives in a stream. In its continuous arrival, because the stream is fast and carries enormous volume, only a small portion of the stream is stored in limited memory. One or few passes over the stream are made to find approximation results. Streaming processing theory and technology have been studied for decades. Representative open source systems include

Storm, S4, and Kafka. The streaming processing paradigm is used for online applications, commonly at the second, or even millisecond, level.

- Batch Processing: In the batch-processing paradigm, data are first stored and then analyzed. MapReduce has become the dominant batch-processing model. The core idea of MapReduce is that data are first divided into small chunks. Next, these chunks are processed in parallel and in a distributed manner to generate intermediate results. The final result is derived by aggregating all the intermediate results. This model schedules computation resources close to data location, which avoids the communication overhead of data transmission. The MapReduce model is simple and widely applied in bioinformatics, web mining, and machine learning.

There are many differences between these two processing paradigms. In general, the streaming processing paradigm is suitable for applications in which data are generated in the form of a stream and rapid processing is required to obtain approximation results. Therefore, the streaming processing paradigm's application domains are relatively narrow. Recently, most applications have adopted the batch-processing paradigm; even certain real-time processing applications use the batch-processing paradigm to achieve a faster response. Moreover, some research effort has been made to integrate the advantages of these two paradigms.

Big data platforms can use alternative processing paradigms; however, the differences in these two paradigms will cause architectural distinctions in the associated platforms. For example, batch-processing-based platforms typically encompass complex data storage and management systems, whereas streaming-processing-based platforms do not. In practice, we can customize the platform according to the data characteristics and application requirements. Because the batch-processing paradigm is widely adopted, we only consider batch-processing-based big data platforms in this paper.

III. BIG-DATA SYSTEM ARCHITECTURE

In this section, we focus on the value chain for big data analytics. Specifically, we describe a big data value chain that consists of four stages (generation, acquisition, storage, and processing). Next, we present a big data technology map that associates the leading technologies in this domain with specific phases in the big data value chain and a

time stamp.

A. BIG-DATA SYSTEM: A VALUE-CHAIN VIEW

A big-data system is complex, providing functions to deal with different phases in the digital data life cycle, ranging from its birth to its destruction. At the same time, the system usually involves multiple distinct phases for different applications. In this case, we adopt a systems-engineering approach, well accepted in industry, to decompose a typical big-data system into four consecutive phases, including data generation, data acquisition, data storage, and data analytics. Notice that data visualization is an assistance method for data analysis. In general, one shall visualize data to find some rough patterns first, and then employ specific data mining methods. I mention this in data analytics section. The details for each phase are explained as follows.

Data generation concerns how data are generated. In this case, the term “big data” is designed to mean large, diverse, and complex datasets that are generated from various longitudinal and/or distributed data sources, including sensors, video, click streams, and other available digital sources. Normally, these datasets are associated with different levels of domain-specific values. In this paper, we focus on datasets from three prominent domains, business, Internet, and scientific research, for which values are relatively easy to understand. However, there are overwhelming technical challenges in collecting, processing, and analyzing these datasets that demand new solutions to embrace the latest advances in the information and communications technology (ICT) domain.

Data acquisition refers to the process of obtaining information and is subdivided into data collection, data transmission, and data pre-processing. First, because data may come from a diverse set of sources, websites that host formatted text, images and/or videos - data collection refers to dedicated data collection technology that acquires raw data from a specific data production environment. Second, after collecting raw data, we need a high-speed transmission mechanism to transmit the data into the proper storage sustaining system for various types of analytical applications. Finally, collected datasets might contain many meaningless data, which unnecessarily increases the amount of storage space and affects the consequent data analysis. For instance, redundancy is

common in most datasets collected from sensors deployed to monitor the environment, and we can use data compression technology to address this issue. Thus, we must perform data pre-processing operations for efficient storage and mining.

Data storage concerns persistently storing and managing large-scale datasets. A data storage system can be divided into two parts: hardware infrastructure and data management. Hardware infrastructure consists of a pool of shared ICT resources organized in an elastic way for various tasks in response to their instantaneous demand. The hardware infrastructure should be able to scale up and out and be able to be dynamically reconfigured to address different types of application environments. Data management software is deployed on top of the hardware infrastructure to maintain large-scale datasets. Additionally, to analyze or interact with the stored data, storage systems must provide several interface functions, fast querying and other programming models.

Data analysis leverages analytical methods or tools to inspect, transform, and model data to extract value. Many application fields leverage opportunities presented by abundant data and domain-specific analytical methods to derive the intended impact. Although various fields pose different application requirements and data characteristics, a few of these fields may leverage similar underlying technologies. Emerging analytics research can be classified into six critical technical areas: structured data analytics, text analytics, multimedia analytics, web analytics, network analytics, and mobile analytics. This classification is intended to highlight the key data characteristics of each area.

B. BIG-DATA TECHNOLOGY MAP

Big data research is a vast field that connects with many enabling technologies. In this section, we present a big data technology map. In this technology map, we associate a list of enabling technologies, both open source and proprietary, with different stages in the big data value chain.

This map reflects the development trends of big data. In the data generation stage, the structure of big data becomes increasingly complex, from structured or unstructured to a mixture of different types, whereas data sources become increasingly diverse. In the data acquisition stage, data collection, data pre-processing, and data transmission

research emerge at different times. Most research in the data storage stage began in approximately 2005. The fundamental methods of data analytics were built before 2000, and subsequent research attempts to leverage these methods to solve domain-specific problems. Moreover, qualified technology or methods associated with different stages can be chosen from this map to customize a big data system.

C. BIG-DATA SYSTEM: A LAYERED VIEW

Alternatively, the big data system can be decomposed into a layered structure. The layered structure is divisible into three layers, i.e., the infrastructure layer, the computing layer, and the application layer, from bottom to top. This layered view only provides a conceptual hierarchy to underscore the complexity of a big data system. The function of each layer is as follows.

- The infrastructure layer consists of a pool of ICT resources, which can be organized by cloud computing infrastructure and enabled by virtualization technology. These resources will be exposed to upper-layer system in a fine-grained manner with specific service-level agreement (SLA). Within this model, resources must be allocated to meet the big data demand while achieving resource efficiency by maximizing system utilization, energy awareness, operational simplification, etc.
- The computing layer encapsulates various data tools into a middleware layer that runs over raw ICT resources. In the context of big data, typical tools include data integration, data management, and the programming model. Data integration means acquiring data from disparate sources and integrating the dataset into a unified from with the necessary data pre-processing operations. Data management refers to mechanisms and tools that provide persistent data storage and highly efficient management, such as distributed file systems and SQL or NoSQL data stores. The programming model implements abstraction application logic and facilitates the data analysis applications. MapReduce, Dryad, Pregel, and Dremel exemplify programming models.
- The application layer exploits the interface provided by the programming models to implement various data analysis functions, including querying, statistical

analyses, clustering, and classification; then, it combines basic analytical methods to develop various field related applications. McKinsey presented five potential big data application domains: health care, public sector administration, retail, global manufacturing, and personal location data.

D. BIG-DATA SYSTEM CHALLENGES

Designing and deploying a big data analytics system is not a trivial or straigthforward task. As one of its definitions suggests, big data is beyond the capability of current hardware and software platforms. The new hardware and software platforms in turn demand new infrastructure and models to address the wide range of challenges of big data. Recent works have discussed potential obstacles to the growth of big data applications. In this paper, we strive to classify these challenges into three categories: data collection and management, data analytics, and system issues.

Data collection and management addresses massive amounts of heterogeneous and complex data. The following challenges of big data must be met:

- *Data Representation:* Many datasets are heterogeneous in type, structure, semantics, organization, granularity, and accessibility. A competent data presentation should be designed to reflect the structure, hierarchy, and diversity of the data, and an integration technique should be designed to enable efficient operations across different datasets.
- *Redundancy Reduction and Data Compression*: Typically, there is a large number of redundant data in raw datasets. Redundancy reduction and data compression without scarifying potential value are efficient ways to lessen overall system overhead.
- *Data Life-Cycle Management*: Pervasive sensing and computing is generating data at an unprecedented rate and scale that exceed must smaller advances in storage system technologies. One of the urgent challenges is that the current storage system cannot host the massive data. In general, the value concealed in the big data depends on data freshness; therefore, we should set up the data importance principle associated with the analysis value to decide what parts of data should be archived and what parts should be discarded.

- *Data Privacy and Security*: With the proliferation of online services and mobile phones, privacy and security concerns regarding accessing and analyzing personal information is growing. It is critical to understand what support for privacy must be provided at the platform level to eliminate privacy leakage and to facilitate various analyses.
- There will be a significant impact that results from advances in big data analytics, including interpretation, modeling, prediction, and simulation. Unfortunately, massive amounts of data, heterogeneous data structures, and diverse applications present tremendous challenges, such as the following.

Approximate Analytics: As data sets grow and the real time requirement becomes stricter, analysis of the entire dataset is becoming more difficult. One wary to potentially solve this problem is to provide approximate results, such as the following.

- *Approximate Analytics*: As data sets and the real time requirement becomes stricter, analysis of the entire dataset is becoming more difficult. One way to potentially, such as by means of an approximation query. The notion of approximation has two dimensions: the accuracy of the result and the groups omitted from the output.
- *Connecting Social Media*: Social media possesses unique properties, such as vastness, statistical redundancy and the availability of user feedback. Various extraction techniques have been successfully used to identify references from social medial to specific product names, locations, or people on websites. By connecting inter-field data with social media, applications can achieve high levels of precision and distinct points of view.
- *Deep Analytics*: One of the drivers of excitement around big data is the expectation of gaming novel insights. Sophisticated analytical technologies, such as machine learning, are necessary to unlock such insights. However, effectively leveraging these analysis toolkits requires an understanding of probability and statistics. The potential pillars of privacy and security mechanisms are mandatory access control and security communication, multi-granularity access control, privacy-aware data mining and analysis, and security storage and management.

Finally, large-scale parallel systems generally confront several common issues; however,

the emergence of big data has amplified the following challenges, in particular.

Energy Management: The energy consumption of large scale computing systems has attracted greater concern from economic and environmental perspective. Data transmission, storage, and processing will inevitably consume progressively more energy, as data volume and analytics demand increases. Therefore, system-level power control and management mechanisms must be considered in a big data system, while continuing to provide extensibility and accessibility.

Scalability: A big data analytics system must be able to support very large datasets created now and in the future. All the components in big data systems must be capable of scaling to address the ever-growing size of complex datasets.

Collaboration: Big data analytics is an interdisciplinary research field that requires specialists from multiple professional fields collaborating to mine hidden values. A comprehensive big data cyber infrastructure is necessary to allow broad communities of scientists and engineers to access the diverse data, apply their respective expertise, and cooperate to accomplish the goals of analysis.

In the remainder of this paper, we follow the value-chain framework to investigate the four phases of the big-data analytic platform.