Journal of Digital Media & Culture Technology Vol.1, No.1, June (2021), pp.1-12
ISSN 2983-0664(p) 2983-0850(e) *http://dx.doi.org/10.29056/jdmct.2021.06.01*

Big Data Service Design : From Collection to Analysis in the Cloud Computing Ecosystem

D. Jaswanth[1], B. Dinesh Reddy[2]*

Abstract

As one of the leading evolutionary trajectories in the information domain, big data innovation has profound applicability for information mining, data analysis, and data sharing within voluminous datasets. It leverages the latent value of data to generate substantial economic benefits while informing decision-making strategies for socio-economic development. Big data solution design, a novel service model that views data as a resource, orchestrates the collection and processing of data from diverse sources. This paper aims to provide a concise overview of the essential elements of big data solution design and its technical processing framework with a focus on data collection and storage. For this, this paper embarks on a comprehensive review of big data processing and analysis strategies, cognizant of the distinct service requirements that aim to provide invaluable information to solution consumers. It then presents an all-inclusive cloud computing solution system predicated on big data with a summary of a variety of big data application scenarios across disparate fields.

Keyword : Big data, decision-making, visualization, service, cloud computing

1. Introduction

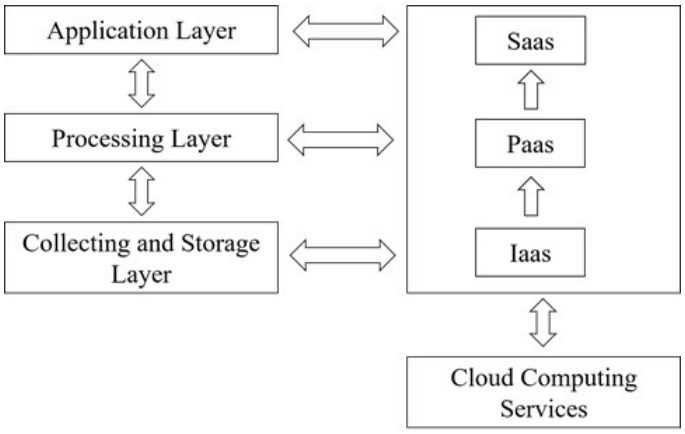
Ever since the concept of big data was first expounded in the esteemed journal, *Nature*, it has been characterized as colossal data sets that overwhelm existing technologies, methodologies, and paradigms in terms of storage, processing, and analysis [1]. Industry reports reveal a notable surge in the economic scale of the global big data market, rising to US$ 58.9 billion in 2017, which signified a robust growth of 29.1%. Further projections indicate that by 2020, this market would be poised to generate revenue surpassing 121.4 billion US dollars. Consequently, it is of pressing importance to foster the development of technologies and systems that boast superior efficiency for the computation, processing, and analysis of voluminous data. Moreover, the implementation of big data technology stands to bolster social governance and operational efficiency, while also facilitating advancements in scientific research [2][3]. However, the intricacy of tasks involved, which are beyond the capability of traditional reasoning and

1 Department of Computer Science and Engineering, Vignan's Institute of Information Technology, India [Professor]
e-mail: jashu3032@gmail.com
2 Department of Computer Science and Engineering, Vignan's Institute of Information Technology, India [Professor]
e-mail: dinesh4net@gmail.com (Corresponding author)

Received(March 15, 2021), Review Result(1st: April 12, 2021), Accepted(June 4, 2021), Published(June 30, 2021)

learning methodologies, necessitates the advent of cutting-edge strategies, algorithms, and infrastructures. Accordingly, in the face of these emergent challenges posed by big data technologies, this study aims to provide a comprehensive exploration of the existing big data service framework.

This study focuses on scrutinizing the existing big data service framework, encompassing three primary layers. Firstly, the data collection and storage layer necessitates data sources for big data services to be gathered using suitable devices, with the collected data subsequently being stored and processed in a distributed data or database system in a pre-processed state. Secondly, in the data processing layer, diverse processing frameworks are employed based on the data type. A profound analysis of big data is currently primarily grounded on large-scale machine learning technologies, adept at unlocking the latent value within the data. This layer also employs visualization tools to convey results to data service consumers [4]. Finally, the application layer demonstrates the use of big data technology across various domains. In addition, big data-driven cloud computing services use software and infrastructure developed around the cloud model, including SaaS, PaaS, and IaaS, for big data processing. A graphical representation of the big data service framework is showin in [Fig. 1].



[Fig. 1] Architecture of Big Data

In the subsequent portions of this study, Section-2 delineates the infrastructure of big data service architecture, encompassing the collection and storage of vast amounts of data. Section-3 introduces technologies for processing and analyzing large-scale data. Section-4 details the cloud computing service models predicated on voluminous data, and the fusion of cloud computing and big data technologies. Finally, in the closing section, we encapsulate several practical scenarios demonstrating the application of big data services.

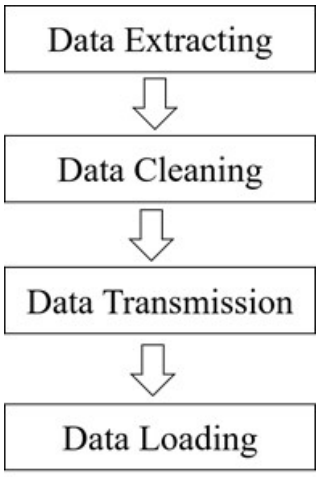
2. Storage and Data collection

In the era of big data, data integration generally involves extracting and loading large amounts of data from massive data sources. These dispersed data need to be collected by suitable hardware or software, and data storage management schemes should be prepared for these massive data in subsequent processing stages [5].

2.1 Data Collection

Big data types primarily consist of static batch data and dynamic stream data. Batch data is retained in a static format, whereas stream data represents a continuous, real-time data sequence. Notably, not all streaming data is stored; many components are discarded immediately post-processing.

Given the instability of stream data transmission, collection methods differ from those used for traditional batch data. For batch data from various sources, ETL (Extract-Transform-Load) tools are routinely employed to enable the transmission and collection of differing data types. As shown in [Fig. 2], data extraction begins from the sources, followed by transformation and loading into the designated storage targets. ETL procedures also expunge corrupted or 'dirty' data through data processing operations such as linkage, transformation, and maintenance. Pot, Data Stage, Informatica, among others, are some frequently utilized ETL tools.



[Fig. 2] Data Collection procedure

Real-time collection of stream data necessitates a tool offering timeliness, fault tolerance, stability, and reliability. Flume, a reliable and fault-tolerant distributed stream processing system, collects, aggregates, and transfers voluminous log data from disparate sources to a centralized storage location. Typically, Flume operates within the Hadoop ecosystem, serving as a middleware between data sources and receivers. Additionally, Kafka, a widely used open-source messaging system, primarily constructs real-time data pipelines and streaming applications. To bolster control and expedite processing of stream data, Kafka employs queues, mitigating any processing speed asynchronism between data generation and processing. Other systems of note include Facebook's Scribe and Taobao's Time-Tunnel.

2.2 Data Storage

Over the past decades, relational databases and structured data management techniques have been widely used. However, according to the characteristics of big data, the storage systems primarily adopted include distributed file systems, NoSQL, NewSQL, and other data management systems.

In 2003, Google developed the Google File System (GFS), a scalable file system designed for large-scale distributed storage. It offers high aggregation efficiency for substantial data and satisfies users' requirements for extensive storage capacity. The Hadoop Distributed File System (HDFS), a component of the Apache Hadoop core project, adopted its design philosophy from GFS. Presently, HDFS is viewed as the most commonly utilized big data tool, furnishing redundancy, reliability, and scalability for parallel distributed architecture systems [6].

NoSQL, a collective term for databases that manage data non-relationally, comprises various data models including key-value pairs, column families, graphs, or documents [7]. For instance, the key-value pair model's simplistic structure reduces the likelihood of data collisions, and its computing programming model is easier to implement. Another model, document-based, utilizes a key to identify each document. Unlike key-value storage, data within a document can be searched. Column family stores, influenced by Google's Bigtable, store their data based on the column family model. NewSQL denotes a novel category of relational database that not only delivers scalability akin to NoSQL but also preserves ACID and SQL services for operations. Spanner and MemSQL are currently the most commonly deployed NewSQL databases.

3. Data Processing

At the beginning of data processing, raw data needs to be cleaned, cropped, and integrated to provide

high-performance big data services for consumers.

3.1 Data Processing

Due to the massive volume of data, the cost of data processing and analysis by accessing all the data is considerable, and completion within a specified time may not be feasible. Thus, initial data processing involves cleaning, cropping, and integrating raw data to provide consumers with high-performance big data services.

3.1.1 Batch Data Processing

Batch data, being static and extremely voluminous, is typically processed using a distributed offline computing method capable of parallel computation [8]. In 2004, Google devised MapReduce, a popular distributed programming model for large data set processing. MapReduce enables users to construct intricate calculations on extensive data sets without concerns of synchronization, fault tolerance, reliability, or availability. This method partitions the input data set into independent blocks processed by parallel Map tasks, and the output results are subsequently processed by Reduce tasks. During the Reduce phase, shuffle processing dynamically distributes data partitions to Reduce nodes.

3.1.2 Stream Data Processing

The stream data processing pattern is apt for data requiring real-time responses, necessitating a stream processing structure with low latency. Storm, an open-source distributed real-time stream data processing system, operates similarly to a real-time Hadoop computing system. It eliminates the need for complex task scheduling, thus reducing processing latency. Storm is characterized by its simplistic programming model, scalability, high reliability, and robust fault tolerance.

Apache Samza is another stream processing framework adept at efficiently processing large volumes of data. LinkedIn's largest Samza implementation can manage millions of messages per second during peak times. The integration of Samza with Kafka capitalizes on the strengths of both architectures. Kafka offers Samza fault tolerance, data buffering, state storage, and other technologies, akin to the dependence of the MapReduce engine on HDFS.

3.1.3 Hybrid Data Processing

Apache Spark is a novel batch data processing framework possessing stream data processing capabilities. Employing a methodology known as Resilient Distributed Dataset (RDD), Spark accelerates

batch data processing by executing the entire procedure in memory. Without memory capacity constraints, Spark outperforms Hadoop in data processing. Furthermore, its Spark Streaming component implements a Micro-batch method, treating continuous streams as a series of micro-batch data and continuously processing these micro-batch jobs. Although Spark Streaming has excellent fault tolerance and load balancing, its performance lags behind pure stream processing structures.

Flink, an open-source data analysis framework overseen by Apache, handles both batch and stream data processing. Compared with other big data processing frameworks, Flink employs unique data processing techniques, using a consistent message queue to process data at different points in persistent streams. While both Spark and Flink can manage hybrid data, the micro-batch processing framework utilized by Spark incurs longer data processing times than Flink.

3.1.4 Graph Data Processing

In large data sets, certain information, referred to as graph data, is interconnected in graph or network form. With the number of vertices and edges of these graphs reaching several millions, traditional graph data processing structures can no longer accommodate the substantial computing requirements. At present, two primary types of graph processing frameworks handle this large-scale graph data: one is a graph database capable of real-time data processing; the other is a computing engine capable of parallel batch processing.

Google's Pregel, a popular batch simultaneous parallel computing system, introduced a vertex-centric large graph computing model. More efficient than MapReduce in handling iterative graph data computation, Pregel delivers an exceptional performance computing engine for traversal, shortest path, and PageRank calculations for large graph data.

3.2 Data Analysis and Visualization

Technologies for big data analysis, such as machine learning, are employed to extract valuable data from vast information repositories, thereby enabling the prediction and analysis of future trends and patterns [9][10]. The resultant insights must then be presented to data service consumers through data visualization [11].

Machine learning, increasingly utilized for comprehensive big data analysis, is an analytical domain concentrating on the principles, learning systems, and algorithmic characteristics. It spans fields such as artificial intelligence, information theory, optimal control, cognitive science, mathematics, engineering, data mining, control systems, identification systems, informatics, and more [12]. In the sphere of big

data analysis, machine learning has garnered substantial attention. For instance, for image data analysis, Niu et al. proposed a unique multi-scale depth model for extracting rich, discriminative features representative of varying visual concepts [13]. A novel unsupervised feature learning strategy was proposed for mapping pixel reflectance to lighting stable codes, while an efficient graphic retrieval method was advanced that improved the overall retrieval rate. In natural language processing, Liu et al. [3] recommended a feature selection method based on correlation analysis and Fisher, capable of eliminating redundant features, and an unsupervised learning method was employed to estimate the acoustic model of speech recognition without needing training data labeling.

After processing and analysis, data service consumers require access to valuable data. Data visualization typically employs tables and illustrations to convey data information to users. Through dynamic visual effects, data can be presented in a more understandable and intuitive manner, enhancing its appeal and persuasiveness. Data visualization enables data analysts to delve deeper into data from multiple dimensions, employing data trends, patterns, correlations, and other details, thereby augmenting data analysis.

4. Big Data-Based Cloud Computing Service Systems

The benefits of cloud computing technology lie in its powerful distributed processing engines, distributed databases, cloud storage, and virtualization technologies. The establishment of a cloud computing service system based on big data creates a high-performance data cloud service system.

4.1 Big Data-based Cloud Computing Service Models

Cloud computing presents a unique computational paradigm and enables a shared pool of configurable computing resources inlcuding networks, servers, storage, applications and services, thus facilitating swift and straightforward configuration and deployment of tasks. With big data workloads often experiencing frequent shifts in scope and dimensionality, potent processing systems are necessary to accommodate these changes. Consequently, the elements of big data computing architecture require thoughtful design, the consideration of the system's features, cost, speed, and scalability.

Recently, the cloud computing service model based on big data technologies has attracted substantial research attention. Given the expanding scope of web applications worldwide, cloud computing's role in big data processing grows increasingly crucial. Novel SaaS models have been developed, and semantic models have been designed to guide the data collection process. To accommodate the emerging

requirements of tenant data duplication protection in SaaS, Li et al. proposed a new tenant duplication protection feature, MT-DIPS, based on tuple sampling. A context-aware secure strategy model was also discussed, customizable according to the specific needs of PaaS-based applications [14].

4.2 Big Data and Cloud Computing

Big data and cloud computing are deeply interconnected, with notable cloud service providers like Microsoft Azure and Amazon offering comprehensive big data development environments on the cloud. These environments integrate data computing, algorithm development, and data services to accommodate various data development requirements. In the context of data processing, real-time data streams and historical data are collaboratively managed through the combination of batch processing and stream processing. For algorithm development, the machine learning platform facilitates the implementation of regression, classification, clustering, and other algorithms, also supporting popular deep learning frameworks like TensorFlow. The data cloud service ecosystem features data maps for the exploration of data details, easing data development and maintenance. Furthermore, professional-level examinations ensure financial data security on the cloud, offering comprehensive data management and security solutions.

The confluence of big data and cloud computing technologies has piqued increasing research interest. Sookhak et al. proposed a new data structure based on the algebraic properties of outsourced files for cloud computing in the field of data collection and storage [15]. In contrast, Yang et al. presented a unique hypergraph overlay model-based energy-saving storage strategy for cloud data centers [16]. For data processing, a novel computing structure named Firework was proposed, with Wang et al. suggesting a machine learning structure for cloud computing and complementary resource allocation [17]. Ezenwoke et al. recommended a visual visualization structure for cloud services [18].

5. Big Data Application Scenarios

Big data technologies have permeated every aspect of people's lives and have been applied in various sectors. In the field of the Internet of Things, data collection technology of wireless sensor networks and data processing algorithms of big data can enable functional applications such as the Internet of Vehicles, novel computer architectures, indoor localization, and road anomaly detection.

5.1 Recommendation Systems

As data expansion has led to information overload, recommendation systems employ big data processing technologies to extract potentially useful information, enabling users to make informed decisions. These systems build models based on users' historical behavior and preferences, suggesting items or products of potential interest, and providing personalized lists for users.

As researchers strive to handle the growing data volume intelligently, studies have explored the complexity of revealing the implicit structure of recommendation systems. The contextual operating tensor model proposed has developed a new suggestion method. In application, Chen et al. recommended a time-aware intelligent object suggestion model jointly in the social internet of things [19]. As vehicle trajectory is easily affected by the surroundings and user behavior, trajectory prediction is relatively low. Thus, a prediction model incorporating environmental awareness and behavioral preference has been suggested. Customer data analysis and recommendation are crucial to the growth of businesses in market competition. A new efficient query framework to discover and recommend potential customers for target products has been proposed.

5.2 Smart Grid

In the smart grid network, infrastructure data, required for automated decision support, is generated through wireless sensor networks and similar technologies. This data is processed and stored in real time to facilitate decision making based on historical or real-time data. Machine learning technologies applied in the smart grid can predict power consumption, pricing, power generation estimation, error detection, adaptive control, among others. New electricity price forecast models have been proposed to handle large price data in the power grid. Moreover, researchers have also proposed optimal polynomial running time algorithms to study the complexity of data upload in communication systems between decentralized devices [20][21].

5.3 Emotional Analysis

The advent of social networks in the big data era has heralded large-scale data scenarios involving emotional analysis. This analysis involves scrutinizing individuals' sentiments, opinions, and evaluations to extract valuable information from vast data [22]. High-precision emotional classification is a major challenge in emotional analysis, with current projects exploring a variety of emotion classification

strategies, from rule-based and dictionary-based methods to complex machine learning algorithms. Machine learning-based techniques utilize algorithms like Random Forests, ANN, k-nearest neighbor algorithm, Genetic Algorithm, among others, for data processing in emotional analysis. For instance, Poria et al. proposed a new multimodal emotional analysis method that utilizes audio, video, and text as information sources.

6. Conclusion

In the era of rapidly advancing modern information technologies, data has established itself as an essential cornerstone for the progression of both production materials and technologies. This paper undertakes a comprehensive examination of the architecture of large data services, explores cloud computing services rooted in large data, and delves into contemporary instances of big data applications. Firstly, it scrutinizes the frameworks that cater to large data collection and storage, employing an array of respective technologies and tools. Based on diverse data processing environments, it introduces the technical underpinnings of four prototypical data processing methodologies. Following this, the paper ventures into an exploration of big data applications, particularly focusing on the deployment of machine learning technologies for an in-depth analysis of voluminous data. Moreover, a critical appraisal of big data visualization technologies is presented. Then, it delineates the models of cloud computing services, emphasizing their confluence with big data technologies with a summary of practical scenarios where large data technologies have been successfully deployed.

References

[1] N. A. Ghani, S. Hamid, I. A. T. Hashem and E. Ahmed, “Social media big data analytics: A survey”, Computers in Human Behavior, vol. 101, December 2019, pp. 417-428, doi: 10.1016/J.CHB.2018.08.039.

[2] L. Wei, et al., “Security and privacy for storage and computation in cloud computing”, Inf. Sci., vol. 258, Feburary 2014, pp. 371-386, doi: 10.1016/j.ins.2013.04.028.

[3] C. Liu, et al., “Authorized public auditing of dynamic big data storage on cloud with efficient verifiable fine-grained updates”, in IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 9, September 2013, pp. 2234 - 2244, doi: 10.1109/TPDS.2013.191.

[4] Z. M. Fadlullah et al., “State-of-the-Art Deep Learning: Evolving Machine Intelligence Toward Tomorrow’s Intelligent Network Traffic Control Systems”, IEEE Communications Surveys & Tutorials, vol. 19, no. 4, 2017, pp. 2432-2455, doi: 10.1109/COMST.2017.2707140.

[5] S. R. Salkuti, “A survey of big data and machine learning”, International Journal of Electrical & Computer

Engineering, vol. 10, no. 1, February 2020, pp.575-580, doi: 10.11591/ijece.v10i1.

[6] J. B. Wang, J. Wang, Y. Wu, J. Y. Wang, H. Zhu, M. Lin, J. Wang, "A Machine Learning Framework for Resource Allocation Assisted by Cloud Computing", IEEE Network, vol. 32, no. 2, March-April 2018, pp. 144-151, doi: 10.48550/arXiv.1712.05929.

[7] E. Castaneda et al., "An Overview on Resource Allocation Techniques for Multi-User MIMO Systems", IEEE Commun Surveys & Tutorials, vol. 19, no. 1, 2017, pp. 239-84.

[8] Y. Arfat, S. Usman, R. Mehmood and I. Katib, "Big data tools, technologies, and applications: A survey, In Smart Infrastructure and Applications", Springer, Cham. 2020, pp. 453-490.

[9] M. Ambigavathi and D. Sridharan, "A survey on big data in healthcare applications" In Intelligent Communication, 28 August 2019, Springer, Singapore, pp. 755-763, doi: 10.1007/978-981-13-8618-3_77.

[10] H. Herodotou, Y. Chen, and J. Lu, "A survey on automatic parameter tuning for big data processing systems", ACM Computing Surveys (CSUR), vol. 53, no. 2, April 2020, pp. 1-37, doi:10.1145/3381027.

[11] F. Bajaber, S. Sakr, O. Batarfi, A. Altalhi and A. Barnawi, "Benchmarking big data systems: A survey", Computer Communications, 2020, pp. 241-251.

[12] R. H. Hariri, E. M. Fredericks and K. M. Bowers, "Uncertainty in big data analytics: Survey, opportunities, and challenges". Journal of Big Data, vol. 6, no. 1, Jun 2019, pp. 1-16, doi: 10.1186/s40537-019-0206-3.

[13] Y. Niu, Z. Lu, JR. Wen, T. Xiang, SF. Chang, "Multi-Modal Multi-Scale Deep Learning for Large-Scale Image Annotation", IEEE Trans Image Process, vol. 28, no. 4, Apr 2019, pp. 1720-1731, doi: 10.1109/TIP.2018.2881928.

[14] W. Li et al, "Complexity and Algorithms for Superposed Data Uploading Problem in Networks With Smart Devices", in IEEE Internet of Things Journal, vol. 7, no. 7, July 2020, pp. 5882-5891, doi: 10.1109/JIOT.2019.2949352.

[15] M. Sookhak, F. R. Yu, A. Y. Zomaya, "Auditing Big Data Storage in Cloud Computing Using Divide and Conquer Tables", IEEE Transactions on Parallel and Distributed Systems, vol. 29, no. 5, May 2018, pp. 999-1012, doi: 10.1109/TPDS.2017.2784423.

[16] T. Yang, H. Pen, W. Li, D. Yuan, A. Y. Zomaya, "An Energy- Efficient Storage Strategy for Cloud Datacenters Based on Variable K-Coverage of a Hypergraph", IEEE Transactions on Parallel and Distributed Systems, vol. 28, no. 12, December, 2017, pp. 3344-3355, doi: 10.1109/TPDS.2017.2723004.

[17] J. Wang, Y. Yang, T. Wang, R. S. Sherratt and J. Zhang, "Big data service architecture: a survey" Journal of Internet Technology, vol. 21, no.2, pp. 393-405.

[18] A. Ezenwoke, O. Daramola, M. Adigun, "Towards a Visualization Framework for Service Selection in Cloud E-Marketplaces", IEEE 13th World Congress on Services, June 2017, Honolulu, HI, USA, pp. 122-129, doi: 10.1109/ICWS.2017.100.

[19] Y. Chen, M. Zhou, Z. Zheng and D. Chen, "Time-Aware Smart Object Recommendation in Social Internet of Things", in IEEE Internet of Things Journal, vol. 7, no. 3, March 2020, pp. 2014-2027, doi: 10.1109/JIOT.2019.2960822.

[20] Y. Li, J. Ma and Y. Zhang, "Image retrieval from remote sensing big data: A survey.", Information

Fusion, vol. 67, March 2021, pp. 94-115, doi: 10.1016/j.inffus.2020.10.008.

[21] W. Yu, G. Zhao, Q. Liu and Y. Song, “Role of big data analytics capability in developing integrated hospital supply chains and operational flexibility: An organizational information processing theory perspective”, Technological Forecasting and Social Change, vol. 163, February 2021, doi: 10.1016/j.techfore.2020.120417.

[22] S. Poria, E. Cambria, N. Howard, G. B. Huang, A. Hussain, “Fusing Audio, Visual and Textual Clues for Sentiment Analysis from Multimodal Content”, Neurocomputing, vol. 174, January, 2016, pp. 50-59, doi: 10.1016/j.neucom.2015.01.095.