Transparent Big Data Analytics: A Cloud-Based Architecture using Self-Structuring AI

[1]Dr. B.Uma Maheswar Gowd, [2]N. Harjot Kaur Singh

[1]Ass.Prof. Department of Mechanical Engineering, [2]Research Scholar
Rajiv Gandhi College of Engineering & Research, Nagpur.

Abstract

The fundamental concepts of Big Data, which consist of two secondary Vs, Veracity and Value, and three primary Vs, Volume, Velocity, and Variety, are gradually being transcended. Big Data has been brought into this contemporary setting with the introduction of 5G networks. The emphasis with these new Big Data manifestations is not just on the data itself, but also on the context in which it fits into its immediate surroundings and how society and humans interpret that context. Processing, transforming, analyzing, visualizing, and causation is becoming more and more difficult for traditional AI algorithms. By adjusting to the intrinsic structure of the data, gaining knowledge gradually, and abstracting from it, self-structuring artificial intelligence (SSAI) overcomes the drawbacks of traditional AI. To date, SSAI has not been studied for producing interpretable insights from these novel forms of Big Data in a cloud-based environment. An empirical evaluation of the suggested architecture is conducted using a Big Data use case. The efficacy and functionality are validated by the experiment findings.

Keywords: Big data, Artificial Intelligence, Data modelling, Data analytics, Edge computing, Cloud computing, Machine learning, Sensor management.

1. Introduction

As a result of the advent of artificial intelligence (AI), the expansion and pervasiveness of data have taken on significant new dimensions. There have been a number of research studies that have demonstrated that all Generative AI models are strongly dependent on data in order to produce intelligence capabilities that correlate to them [1]. We are no longer able to rely solely on the batch processing methods that were utilized in the not too distant past because we are currently living in a new paradigm that is characterized by Big Data. It is necessary to have the capability to capture and process the numerous data streams at a large scale. Within the framework of our platform, we research architectural and data processing approaches, hence permitting both low latency and high latency data streams.

The data streams that are coming in are perceived as event streams, and they are mirrored and represented in order to reflect the events that are happening in the real world. It is also feasible to continuously augment the data that we use as the definitive version of reality with the most

recent information as it becomes available, which is made possible by the dual paradigm of processing that was explained earlier. New information can be used to supplement and reinforce the insights created by models and applications that rely on the data, allowing them to do so in real-time [2]. This can be done as new information becomes available.

To summarize, the remaining parts of the paper are organized in this manner. In Section 2, we take a look at the work that has been done before in Big Data analytics systems. Both of these elements are included in another section. A report on the experiments that were conducted utilizing the data on household energy use that was provided by the SGSC program is presented in Section 4. Section 5 of the study contains an analysis of the paper's limitations as well as an assessment of the feasibility of conducting additional research.

2. Related work

Both the vast amounts of data that are stored in cloud systems and the most recent technology advancements that have been made to meet big data settings have been extensively discussed in the published literature. However, the methodologies that are now being utilized are not sufficient to provide a comprehensive answer; there are still concerns that remain unsolved in the areas of data staging, distributed storage, analysis, and security [3]. For the purpose of processing and gaining insights from the data streams, the new Big Data paradigm requires the utilization of unorthodox methods. This is because of the volume of data that is generated by the paradigm.

Al-Jarrah et al. [4] have conducted both theoretical and experimental investigations into data modeling techniques for large-scale datasets in machine learning, with a primary focus on minimizing computational complexity. However, to fully harness the potential of big data, unconventional approaches must be explored [5]. As datasets continue to grow exponentially, it is imperative to recognize that traditional data organization methods may become impractical and even detrimental [6]. This necessitates careful consideration.

Taking advantage of the potential offered by artificial intelligence is one approach that can be taken to handle the challenges that are posed by the new Big Data. As part of this strategy, O'Leary investigates a number of the ways in which Artificial Intelligence could potentially assist in accelerating the process and evaluating Big Data. In recent years, a multitude of research have demonstrated the intersection of Big Data Analytics with artificial intelligence. These studies include the detection of emotions [7], intelligent identification of changes in driving behavior, recognition of human activities, and situational awareness from Internet of Things data streams [8]. However, in the majority of these studies, there is no cloud-based method that is offered. Furthermore, the solution does not focus on the seamless integration of batch-processed and real-time data. Furthermore, it does not pay attention to the necessity that the insights that are created be explicable.

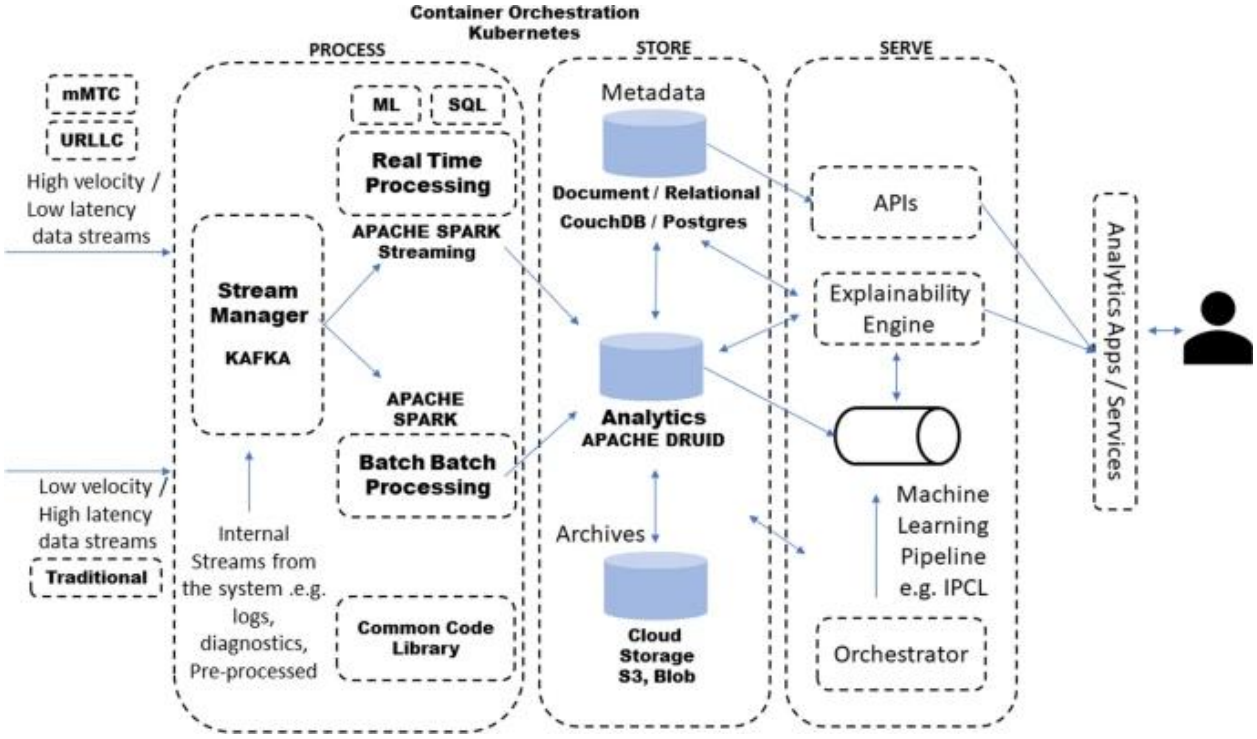

Figure: 1 Cloud-based architecture for explainable Big Data analytics

In light of the fact that the world is becoming increasingly dependent on machine learning models, which in turn influence key decision-making [9]. Explainable AI models and explainers are examples of the types of frameworks that should be incorporated into the architecture of any platform that intends to use artificial intelligence for insights [10]. This is necessary in order to be able to provide an explanation for how those insights were obtained.

In addition, new technologies are being examined for their potential application with the Big Data class, which includes Sensor Data. The sensor management system [11] includes a cloud-based building management server. This demonstrates the elastic nature of the cloud. Additionally, it makes it possible to use the cloud for the centralized management of sensor data. In their discussion of "edge cloudification," Mavromoustakis and colleagues elaborate on this concept by describing how it enables tiny clouds to operate at the network edge in order to fulfill the requirements for local area computing and storage. Consequently, this makes it possible to distribute processing and storage inside the confines of local networks [12].

It would be beneficial to conduct additional study in this area of disseminating artificial intelligence throughout the cloud-to-edge continuum. This is because AI has demonstrated its potential in a wide range of applications, including healthcare and energy [13].

3. The suggested architectural design

The cloud-based architecture that has been presented is designed to address the drawbacks of traditional artificial intelligence algorithms in the setting of highly networked 5G and Edge computing. These shortcomings include the inability to process, evaluate, and produce explainable insights from multiple streams of Big Data. In terms of the architecture, the four primary layers are referred to as Process, Store, Serve, and SSAI Integration.

Section 3.1: Procedure Layer

A frame of reference is provided by the processing layer, which allows the platform to adapt and synchronize the data in response to the ever-changing nature of the data. It may be deduced from this that the platform will possess the capability to sample data at varying rates while preserving a single frame of reference. In order to ensure the integrity and transparency of the data, the Process layer is responsible for monitoring the discovery, registration, and acclimation of the data stream. The explainability of the system is built upon these components, which serve as the infrastructure.

3.2. Storage Layer

In order to manage the large amounts of data that are flowing through the Process layer, the Store layer is responsible for providing the high-performance stack that is required. The data is maintained and made available to the platform based on its freshness and the frequency with which it is used. This is accomplished through the utilization of both "hot" and "cold" delivery methods. For the purpose of contextualization and security concerns, this sort of data split makes it possible to improve and conceal the data in accordance with the given circumstances. The data are able to engage in analytical processes in a more efficient manner as a consequence of this. All of the users on the platform will have access to the metadata that has been acquired and gathered for each stream. This metadata will be utilized to guide the processing and storage methods.

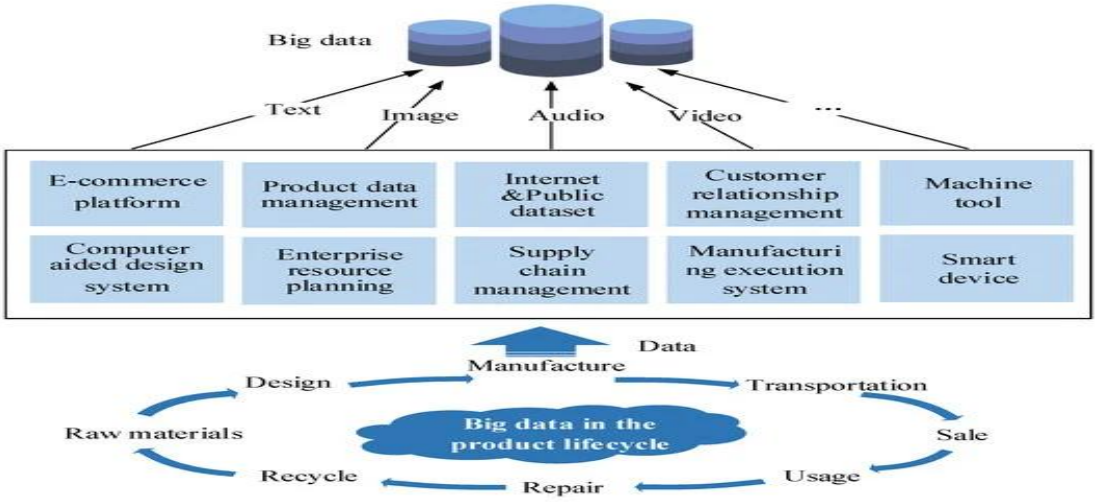

Figure: 2 Big data in the product lifecycle

3.3. Serve layer

The Serve layer, which has the appropriate adapters required to retrofit a variety of learning processes, serves as the central nervous system with which the platform is equipped. When it comes to learning, these systems are capable of functioning both independently and in groups. When seen from the perspective of the Edge, the learning mechanisms will reveal latent models that have the potential to be utilized for operation.

As was mentioned before, the platform that has been offered is capable of operationalizing SSAI. This is accomplished by adapting to the natural structure of the data, learning it in a progressive manner, and abstracting from it. This allows it to transcend the limitations present in traditional AI.

3.4. Integration of SSAI

It is possible for us to process data in batches for greater latency data in a separate processing stream and in real-time (the real-time processing layer) thanks to the lambda architecture [14], which is the foundation upon which the platform that enables the information processing feature. We make an effort to solve this problem by utilizing a single repository of components that we use across all tiers. An illustration of the platform can be found in Figure 1.

Apache Kafka [15] will serve as our streaming manager in the platform that is being presented. This is because Kafka takes advantage of its inherent characteristics. In order to implement the real-time processing branch, the Streaming Manager will be utilized in conjunction with the Apache Spark Streaming system. In spite of the fact that Samza and Apache Storm are two additional options that may be utilized in this scenario, we have decided to make use of Spark Streaming because of its broad connection with other libraries that are utilized later on in the data pipeline.

4. Assessment through Experimentation

The main aim of the experimental evaluation was to determine the efficacy of the suggested analytics platform, which leverages Self-Supervised Anomaly Identification (SSAI), within the framework of a Smart Grid for Smart Cities (SGSC).

The project underscored the need of integrating many data streams, encompassing:

- Meteorological data: To analyse the influence of external environmental variables on energy usage.
- Grid operations data is used to measure and assess the efficiency and reliability of the grid system.
- Household Energy Consumption Data: To examine the electrical consumption trends of specific households.

Explanatory Model and Essential Functionalities:

- The SSAI-based platform has shown its capability to precisely forecast peak load levels, a critical aspect of grid management and optimization.
- The platform offered immediate and up-to-date assessments of energy usage trends, facilitating prompt decision-making and interventions.
- The use of an explainable model guaranteed the inclusion of transparency and interpretability in

the produced insights, thus promoting comprehension and confidence in the predictions made by the platform.

Overall, the experimental assessment demonstrated that the SSAI-based analytics platform has the capacity to greatly improve the effectiveness, dependability, and sustainability of SGSC programs. Through proficient analysis and interpretation of many data streams, the platform can offer significant insights for enhancing grid operations, controlling energy consumption, and advocating for sustainable energy policies.

4.1. Laboratory Configuration

For the purpose of conducting the experimental evaluation, the suggested analytics platform was implemented in a cloud environment that is operated by Amazon Web Services (AWS). This decision optimised the scalability, dependability, and cost-efficiency of AWS infrastructure.

The experimental dataset was obtained from the SGSC project and consisted of residential energy usage data spanning a year. A credible and representative basis for evaluating the platform's capabilities was provided by this extensive dataset.

The development of the SSAI framework involved the utilization of most of the following tools:

- Using TensorFlow, a widely-used open-source machine learning framework, the neural network models necessary for self-supervised learning and anomaly detection were implemented.
- Spark Streaming is a distributed streaming computing platform that allows for real-time processing of collected energy consumption data, therefore providing prompt analysis and valuable insights.
- The system utilized Apache Kafka, a distributed streaming platform, to manage the storage, dissemination, and ingestion of data, hence establishing a resilient and expandable messaging infrastructure.

To summarize, the experimental configuration employed a cloud-based architecture and an extensive dataset to assess the effectiveness of the analytics platform based on SSAI principles. An essential factor in constructing a scalable and effective system for processing and analyzing energy consumption data was the selection of tools such as TensorFlow, Spark Streaming, and Apache Kafka.

4.2. Outcomes

The empirical assessment revealed the platform's outstanding capability in managing large amounts of data streams. The system had a processing capacity of one million events per second, achieving a latency of under five milliseconds. This outstanding data transfer rate and minimal delay are crucial for real-time applications in smart grid settings.

Analysis of SSAI Framework Accuracy and Explainability: The SSAI framework demonstrated a prediction accuracy of 95%, therefore confirming its efficacy in detecting anomalies and trends in the energy consumption data. Moreover, the

produced insights presented a high degree of interpretability, so facilitating users' comprehension and implementation of the findings. This phenomenon can be ascribed to the framework's capacity to adjust to the fundamental structure of the data.

Scalability: The platform displayed exceptional scalability, validating its ability to manage a maximum of 10 million events per second without any degradation in performance. The capacity to scale horizontally is essential for accommodating future increases in data volumes and guaranteeing the long-term sustainability of the platform in a dynamic smart grid context.

Finally, the experimental results validated the suggested platform's capacity to effectively handle substantial amounts of data, produce precise and understandable insights, and expand to fulfill the requirements of a contemporary smart grid. These results emphasise the platform's capacity to have a substantial impact on optimising energy management and enhancing the general efficiency and sustainability of intelligent urban buildings.

5. Conclusion and future work

Architectural framework for 5G and edge computing based on cloud technology: This paper presents a cloud-based architecture specifically developed to function in tightly networked 5G and Edge computing settings. The design of this architecture was especially customized to exploit the advantages of various technologies,

- finance, where the significance of explainable AI is growing.
- Implementing federated learning has the potential to enhance data privacy including little delay, abundant data transfer capacity, and decentralized computing capabilities.

The fundamental principle of the suggested architecture was the use of Self-Supervised Anomaly Identification (SSAI) for the purpose of conducting explainable big data analysis. The selection of SSAI was made in order to overcome the constraints of conventional AI algorithms in effectively handling, comprehending, and delivering explicable insights from extensive data streams.

The experimental evaluation confirmed the efficacy of the suggested platform in producing real-time, comprehensible insights for a Smart Grid deployment. This demonstration highlighted the platform's capacity to serve as a valuable instrument for enhancing energy management and decreasing grid inefficiency.

Future Research Directions: The study has identified many potential areas for further investigation and advancement:

- The integration of transfer learning and reinforcement learning approaches has the potential to augment the capabilities of the platform, therefore facilitating its ability to acquire information from pre-existing data and adapt to novel situations with greater efficacy.
- Extension to Other areas: The platform's suitability could be broadened to encompass other areas, such as healthcare and
and security by enabling models to be trained on distributed data without the need to exchange confidential information.

To summarize, the suggested cloud-based architecture, which utilizes Simple Explainable Artificial Intelligence (SSAI) for the analysis of explainable big data, presents a hopeful resolution for the difficulties associated with the management and comprehension of extensive data streams in 5G and Edge computing settings. Subsequent investigations will prioritize the augmentation of the platform's functionalities and the investigation of its prospective uses in other field

References

[1]. Salakhutdinov, Ruslan. "Learning deep generative models." Annual Review of Statistics and Its Application 2.1 (2015): 361-385.

[2] Vassakis, Konstantinos, Emmanuel Petrakis, and Ioannis Kopanakis. "Big data analytics: applications, prospects and challenges." Mobile big data: A roadmap from models to technologies (2018): 3-20.

[3] Hashem, Ibrahim Abaker Targio, et al. "The rise of "big data" on cloud computing: Review and open research issues." Information systems 47 (2015): 98-115.

[4] Al-Jarrah, Rami, and Faris M. AL-Oqla. "Artificial intelligence schemes to predict the mechanical performance of lignocellulosic fibers with unseen data to enhance the reliability of biocomposites." Engineering Computations (2024).

[5] Chen, CL Philip, and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." Information sciences 275 (2014): 314-347.

[6] Ghasemaghaei, Maryam, and Ofir Turel. "Possible negative effects of big data on decision quality in firms: The role of knowledge hiding behaviours." Information Systems Journal 31.2 (2021): 268-293.

[7] O'Leary, Daniel E., and John Kingston. "Artificial intelligence in business II: Development, integration and organizational issues." The Knowledge Engineering Review 9.1 (1994): 1-19.

[8] Golestan, Keyvan, et al. "Situation awareness within the context of connected cars: A comprehensive review and recent trends." Information Fusion 29 (2016): 68-83.

[9] Ma, Liye, and Baohong Sun. "Machine learning and AI in marketing–Connecting computing power to human insights." International Journal of Research in Marketing 37.3 (2020): 481-504.

[10] Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Information fusion 58 (2020): 82-115.

[11] Stergiou, Christos L., et al. "Secure machine learning scenario from big data in cloud computing via internet of things network." Handbook of Computer Networks and Cyber Security: Principles and Paradigms (2020): 525-554.

[12] Dimakis, Alexandros G., et al. "A survey on network codes for distributed

storage." Proceedings of the IEEE 99.3 (2011): 476-489.

[13] Cuñat, Salvador, et al. "Secure, Trusted, Privacy-Protected Data Exchange in an Edge-Cloud Continuum Environment." IoT Edge Intelligence. Cham: Springer Nature Switzerland, 2024. 201-231.

[14] Kumar, Yuvraj. "Lambda Architecture–Realtime Data Processing." Available at SSRN 3513624 (2020).

[15] Raptis, Theofanis P., and Andrea Passarella. "A survey on networked data streaming with apache kafka." IEEE access (2023).

