Iraqi Journal for Computer Science and Mathematics

Journal Homepage: http://journal.esj.edu.iq/index.php/IJCM

ISSN: 2788-7421

Big Data Streaming Platforms: A Review

Harish Kumar[1], Mohd Arfian Ismail[2,*]

[1]Department of Computer Science, College of Computer Science, King Khalid University, Abha 61413, Saudi Arabia
[2]Faculty of Computing, College of Computing and Applied Sciences, Universiti Malaysia Pahang, Malaysia

*Corresponding Author: Mohd Arfian Ismail

DOI: https://doi.org/10.52866/ijcsm.2022.02.01.010
Received February 2022; Accepted March 2022; Available online April 2022

ABSTRACT: Yesterday's "Big Data" is today's "data." As technology advances, new difficulties and new solutions emerge. In recent years, as a result of the development of Internet of Things (IoT) applications, the area of Data Mining has been confronted with the difficulty of analyzing and interpreting data streams in real time and at a high data throughput. This situation is known as the velocity element of big data. The rapid advancement of technology has come with an increased use of social media, computer networks, cloud computing, and the IoT. Experiments in the laboratory also generate a large quantity of data, which must be gathered, handled, and evaluated. This massive amount of data is referred to as "Big Data." Analysts have seen an upsurge in data including valuable and worthless elements. In extracting usable information, data warehouses struggle to keep up with the rising volume of data collected. This article provides an overview of big data architecture and platforms, tools for data stream processing, and examples of implementations. Streaming computing is the focus of our project, which is building a data stream management system to deliver large-scale, cost-effective big data services. Owing to this study, the feasibility of large-scale data processing for distributed, real-time computing is improved even when the systems are overwhelmed.

Keywords: big data; streaming processing; Kafka; Spark

1. INTRODUCTION

As a result of recent technical advancements, what has come to be known as the "Data Era" [1] is currently occurring at the same time as a remarkable increase in the mobility of digital devices is observed. The ubiquitous interconnectedness of modern technologies enables real-time access to an unprecedented quantity of information and more importantly, the sources of that information [2, 3]. Almost every activity performed on an electronic device creates data, which are then archived for future use due to their anticipated informative value. With the remarkable increase in the amount of big data that has been noticed recently [4, 5], the challenge of mining and extracting information from this type of data is gaining traction. Among the many different properties of big data, velocity is undoubtedly one of the most important, and this trait provides big data streams their first-class status among other types of data. Hence, mining large amounts of data from a variety of sources is essential and required, as demonstrated by recent attempts in this field. Another substantial difficulty that occurs in conjunction with the mining problem is the challenge of maintaining the privacy of gigantic data stream sources while mining massive volumes of information. A wide variety of applications that create big data exist today, including social media platforms such as Facebook, Twitter, and Google [6], and broad public websites, scientific investigations, commercial applications, cloud computing devices, Internet of Things (IoT) devices, e-governance applications, bio-medical applications, and a variety of other things. Big data [7] is characterized by the properties of volume, validity, and diversity of information included inside it. It likewise reflects the data acquired in a methodical manner, but it has a storage and power capacity that exceeds that of conventional devices in an enterprise. As data volume continues to expand at an alarming rate, handling petabytes of data every day is becoming necessary. It

represents the exponential increase in the amount of organized, semi structured, and unstructured data that are becoming available. Furthermore, the data include music, video, photos, and other types of media.

Processing them correctly will enable reaching the most accurate and quick judgments. To grasp the remarkable insights concealed inside such data, such data must be analyzed and extracted, which is known as big data analytics. Clearly, the problem has a wide range of real-world application scenarios, spanning from trajectory data stream management to electronic health data stream processing and from fraud detection and analysis of corporate data streams to surveillance and emergency management. As a consequence of media convergence, an increase in the use of the Internet has been observed, which has resulted in the development of social media, which is an online networking platform.

Web-based social networking is a concept that has provided individuals a worldwide platform for disseminating their news, viewpoints, and ideas on events taking place in their communities. Using web-based social networking has become more accessible and necessary in today's culture. Clients can communicate with other users in the system by exchanging ideas, images, status updates, postings, and other matters on social networking sites. It has proven to be one of the most extensive platforms for users to voice their ideas. To grasp the worth of public mood, social media is regarded as a gold mine. In the social media world, often known as the Internet, developers create applications modeled after the web 2.0 ideological and inventive establishment, and allow clients to transact in real time. Owing to the widespread usage of Twitter information, it has become especially well-known as a data source for research, development, and applications across a wide range of fields. Using data from Twitter as part of a request to provide an approach to establish business is one area where data from Twitter are often used as a part of market analysis. Researchers have used Twitter to predict the success of online services and goods as well as identify potential customers who follow to obtain information about products and services. The study of data mining has resulted in the development of several tools, methods, and algorithms that may be used to manipulate enormous amounts of data to solve real-world issues. The data mining case objectives are to deal with large amounts of data and patterns successfully, and to maximize the amount of insightful learning available to the user. Big data is used to depict the exponential growth of information as well as the accessibility of both ordered and unstructured information, and it is becoming increasingly popular. It facilitates connections in selecting business patterns and the form of research. Dealing with big data is complicated because of the large amount of resources required. The organization of enormous data in this study is based on the consideration of big data applications and dependencies.

1.1 BACKGROUND

Stream computing [8] describes the real-time processing of large amounts of data generated at fast rates from various sources with minimal latency. It is a new paradigm needed by the proliferation of new data-generating situations, The widespread usage of location services, mobile devices, and the pervasiveness of sensors are a few examples. It may be defined as the high-velocity data flow via real-time sources such as the IoT, sensors, market information, smart applications, and clickstreams (Figures 1 and 2) [9]. The essential premise of this paradigm is that the potential value of data is derived from the fact that it is current. The outcome is those data are analyzed immediately upon arrival in contrast to batch computing, where results are generated in a stream [10], and data are initially stored before examination. Computing systems with scalable parallel architectures and high performance are critical for the future of technology. The features of stream computing enable organizations to examine and respond in real time to constantly changing and fluctuating data. The accessible stream processing frameworks include Storm, S4, Kafka, and Spark [11]. The true disparities between batch processing and stream processing concepts are illustrated here. A programming paradigm known as stream computing is necessary to include flowing data into the decision making. With stream computing, queries that are relatively static may be assessed on data in motion (i.e., real-time data) in a continuous manner.

1.2 BIG DATA STREAMING

Metrics that are not reliant on the system's intended use or purpose may characterize data storage and processing in a given system. The "3D/3Vs" of big data and more recently, the "4Vs" of big data measure the data's volume, velocity, diversity, and value/veracity [12]; each metric corresponds to a certain component of information; volume, velocity, variety, and value/veracity of data volume measure the order of magnitude in which a piece of data is measured in terms of its size, regardless of how it is encoded or where it is stored. The pace at which a specific volume of data moves is measured by data velocity [13]. This metric adds a temporal dimension to data volumes, which reveals itself in the frequency with which events occur, the time they arrive, and the speed at which data are sent (throughput and bandwidth). Cloud services are now widely available and may be utilized by anybody from any place at any time of day. As a result, speed is becoming increasingly vital and required for success. This limitation and the interaction of such systems with human users result in major ramifications in terms of data rate volatility because the data rate changes are directly tied to our human habits and limits as well as the sort of habit they are built to cope with. All of the data are of a certain

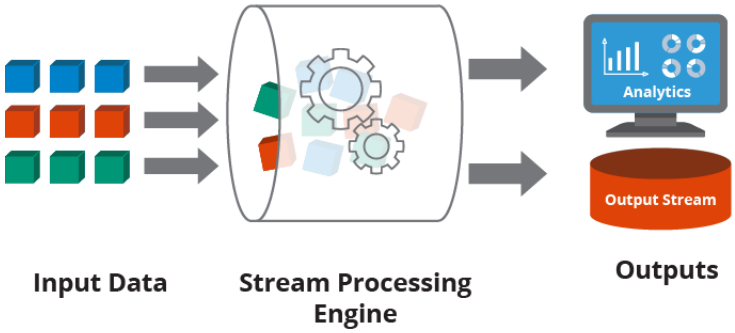

FIGURE 1. Streaming processing structure

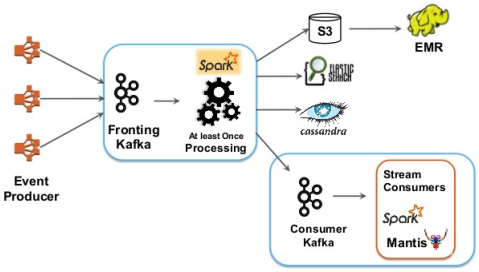

FIGURE 2. New big data streaming tools

type. As a result of technology advancements in recent decades, the number of data sources available and the diversity of data available have increased. Consequently, unstructured data are identified, created, and stored. Unstructured data are information pieces created less purposefully; hence, they are not always ideal for relational database models (as opposed to structured data). Unstructured data, which are frequently utilized in conjunction with structured data, originate from any sensor that interacts with the natural world and can originate from any sensor that interfaces with the natural world (biosensors, scanners, and cameras). As a result, "Big Data" storage and processing facilities are likely to encounter this paradigm and today need to be capable of dealing with it in the future. This result relates to the data's informative value in a given context and the direct implications for the analytical tools in place, and can include a wide range of data, from user usage statistics that provides valuable market insight into payment details and/or financial data that serve as the backbone of an industry. Another V to consider is veracity, which serves as a dependable indicator of the data's integrity and dependability. Decision making is one of the primary components influenced by big data processing, and making judgments based on trustworthy data is critical; hence, veracity is a value in and of itself in this context.

1.3 SCALABILITY

Scalability is one of the most challenging problems to overcome in big data streaming analysis because of the vast amount of data involved. The amount of big data generated is increasing at an exponential pace, which is substantially faster than the rate at which computer resources are being added. Moore's law is followed by processors, but the amount of data is growing at an exponential rate. To deal with the ever-increasing size and complexity of data, the primary focus of research efforts should be the development of scalable frameworks and algorithms that will accommodate the data stream computing mode, the development of an effective resource allocation strategy, and the resolution of parallelization issues.

1.4 INCOMPLETENESS AND HETEROGENEITY

Massive amounts of data are generated regularly and constantly changing in terms of structure, organization, semantics, accessibility, and granularity. How to cope with an ever-increasing volume of data, extract useful information from them, and aggregate and correlate streaming data from various sources in real time, all while maintaining high performance, is the challenge in this case. To develop a competent data presentation, it should be designed in such a manner that accurately represents the structure, diversity, and order of the arriving streaming data.

2. WEAKNESSES AND STRENGTHS OF BIG DATA STREAMING TOOLS

According to the research, individual big data streaming technologies may not deliver the complete range of necessary capabilities. Finding a specialized big data solution that combines important features such as scalability, integration, fault tolerance, timeliness, consistency, heterogeneity and incompleteness management, and load balancing is difficult. While Spark streaming and Sonora are great, efficient checkpoint systems, the amount of operator space accessible to user programs in both situations is restricted. The S4 protocol cannot provide a fault-tolerant persistent state in its entirety. Storm's "at-least-once" technique for record delivery does not ensure that messages are sent in the correct sequence. Trident's ability to function depends on the strict transaction sequencing required. In contrast to streaming SQL, it provides simple, succinct solutions when dealing with numerous streaming challenges. Complex application logic (such as matrix multiplication) and intelligible state abstractions are expressed using the operational flow of an imperative language rather than a declarative language such as SQL. Furthermore, BlockMon makes advantage of batching and cache locality optimization techniques to improve the efficiency of memory allocation and the speed at which data are accessed. However, deadlock may arise if data streams are enqueued at a faster pace than the rate at which blocks are consumed.

A flow control layer on top of Apache Samza is required to handle batch latency processing difficulties, which is why it is included in the Apache Samza distribution. The Flink framework, despite its limits in terms of scalability, is well suited for large-scale stream processing as well as batch-oriented tasks. Redis' in-memory data store enables it to be incredibly quick but because of this, the size of the Redis data store is determined by the amount of RAM available on the system. In designing big data benchmarks applicable to a wide range of workload scenarios, the variety of large data presents a considerable challenge. In the case of big data, keeping a single benchmark is difficult because utilizing the same benchmark on various data sets does not produce the same results. Thus, benchmark testing should be performed on applications specific to their company or organization. Moreover, identifying the workload associated with a certain application domain is necessary before investigating a big data system. The vast majority of the big data benchmarks that are now available are designed to assess a certain type of system or architecture. For example, the HiBench benchmarking tool is well suited for assessing Hadoop, Spark, and streaming workloads, whereas GridMix and PigMix are well suited for analyzing MapReduce Hadoop systems. BigBench may be used to test Teradata Aster DBMS, MapReduce systems, Redshift database, Hive, Spark, and Impala. It can also be used to test other types of databases. According to the information now available, BigDataBench appears to be the only big data benchmark that can currently assess a hybrid of multiple big data systems. As a result, several scholars have examined their work and consider synthetic and real- world data in their analyses, but the scientific community does not appear to be unanimously in favor of the use of a standard benchmark dataset for large data streaming analysis. Only a handful of the studies that made use of standardized benchmarking are described in further detail further down the page. Word Count and Grep are used as benchmarks to determine the overall quality of the work. According to the findings, the suggested approach can deal with uncertain input and ensure that the overall latency of the event remains within a reasonable range of expectations. On two datasets, the car dataset and the Wikinews5 dataset, the algorithm created by was evaluated, and it outperformed sequential processing in both cases. Their technique (pipeline implementation) was more effective and faster than the competition.

A platform-as-a-service deployment may be ideal for a platform with a high spike load profile, such as a gaming platform. Customers are only billed when data are processed if platform distribution can be achieved on an infrastructure-as-a-service cloud platform, according to the company. When dealing with predictable or continuous demands, on-premise deployment may be a practical choice. If the workloads are mixed, a hybrid cloud and an on-premise solution may be explored for implementation (i.e., consisting of consistent flows and spikes). This setup enables the seamless integration of web-based services or applications as well as remote access to important activities while on the road without compromising performance.

3. CONCLUSION AND FURTHER WORK

Large-scale data streaming analytics has emerged as the next frontier of competitiveness and innovation as a result of the difficulties and possibilities from the information technology revolution. Small- and medium-sized businesses that use big data streaming analytics obtain insights that enable them to make more informed real-time decisions, providing them a competitive advantage over their competitors. This study analyzed the contemporary triumphs and creative obstacles that have emerged in the studied domain while considering the key advancements in the context of privacy- preserving large data stream mining, which were discussed in the previous study. As application scenarios such as social network analysis, scientific tools and systems, census data mining, and other related areas grow in relevance, the importance of this scientific issue is expected to increase in importance in the next years. This conclusion is logical to obtain as soon as possible. With the help of a comprehensive literature review, the authors attempted to present a holistic view of big data streaming analytics by understanding and identifying the tools and technologies, methods and techniques, benchmarks

Table 1. Open source tools and technology for large data stream analysis

No.	Instruments and technological advances	Source
1.	Spark streaming	**[14]**
2.	Apache Kylin	**[15]**
3.	Yahoo! S4	**[16]**
4.	Kafka	**[17]**
5.	CQELS	**[18]**
6.	NoSQL	**[19]**
7.	XSEQ	**[20]**
8.	EsperTech	**[21]**
9.	ETALIS	**[22]**
10.	BlockMon	**[23]**
11.	Apache storm	**[24]**
12.	Apache Aurora	**[25]**
13.	SAMOA	**[26]**
14.	Apache Samza	**[27]**
15.	Redis	**[28]**
16.	C-SPARQL	**[29]**
17.	Splunk stream	**[30]**
18.	Photon	**[31]**
19.	MavEStream	**[32]**

or methods of evaluation used, and key issues in big data stream analysis to showcase the signposts of future research directions.

ACKNOWLEDGMENT

The moral support from the universiti malaysia pahang.

CONFLICT OF INTERESTS

The authors declare no conflict of interests.

REFERENCES

[1] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, "Deep learning for IoT big data and streaming analytics: A survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 2923–2960, 2018.
[2] R. Ranjan, "Streaming big data processing in datacenter clouds," *IEEE cloud computing*, vol. 1, no. 1, pp. 78–83, 2014.
[3] A. A. Safaei, "Real-time processing of streaming big data," *Real-Time Systems*, vol. 53, no. 1, pp. 1–44, 2017.
[4] N. D. Zaki, N. Y. Hashim, Y. M. Mohialden, M. A. Mohammed, T. Sutikno, and A. H. Ali, "A real-time big data sentiment analysis for iraqi tweets using spark streaming," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 4, pp. 1411–1419, 2020.
[5] A. H. Ali, "A survey on vertical and horizontal scaling platforms for big data analytics," *International Journal of Integrated Engineering*, vol. 11, no. 6, pp. 138–150, 2019.
[6] S. Shahrivari, "Beyond batch processing: towards real-time and streaming big data," *Computers*, vol. 3, no. 4, pp. 117–129, 2014.
[7] R. Agerri, X. Artola, Z. Beloki, G. Rigau, and A. Soroa, "Big data for Natural Language Processing: A streaming approach," *Knowledge-Based Systems*, vol. 79, pp. 36–42, 2015.
[8] A. H. Ali and M. Z. Abdullah, "Recent trends in distributed online stream processing platform for big data: Survey," *2018 1st Annual International Conference on Information and Sciences (AiCIS)*, pp. 140–145, 2018.
[9] M. D. Smith and R. Telang, "Streaming, sharing, stealing: Big data and the future of entertainment, Mit Press," 2016.
[10] A. Kumari, S. Tanwar, S. Tyagi, and N. Kumar, "Verification and validation techniques for streaming big data analytics in internet of things environment," *IET Networks*, vol. 8, no. 3, pp. 155–163, 2019.
[11] R. Scolati, I. Fronza, N. E. Ioini, A. Samir, and C. Pahl, "A Containerized Big Data Streaming Architecture for Edge Cloud Computing on Clustered Single-board Devices," *Closer*, pp. 68–80, 2019.
[12] W. Inoubli, S. Aridhi, H. Mezni, M. Maddouri, and E. Nguifo, "A comparative study on streaming frameworks for big data," *VLDB 2018-44th International Conference on Very Large Data Bases: Workshop LADaS-Latin American Data Science*, pp. 1–8, 2018.
[13] L. R. Nair, S. D. Shetty, and S. D. Shetty, "Applying spark based machine learning model on streaming big data for health status prediction," *Computers & Electrical Engineering*, vol. 65, pp. 393–399, 2018.
[14] S. Chintapalli, "Benchmarking streaming computation engines: Storm, flink and spark streaming," *2016 IEEE international parallel and distributed processing symposium workshops (IPDPSW)*, pp. 1789–1792, 2016.

[15] S. V. Ranawade, S. Navale, A. Dhamal, K. Deshpande, and C. Ghuge, "Online analytical processing on hadoop using apache kylin," *International Journal of Applied Information Systems*, vol. 12, pp. 1–5, 2017.
[16] J. Chauhan, S. A. Chowdhury, and D. Makaroff, "Performance evaluation of Yahoo! S4: A first look," *2012 Seventh International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, pp. 58–65, 2012.
[17] F. Kafka, "The Basic Kafka, Simon and Schuster," 1979.
[18] D. L. Phuoc, M. Dao-Tran, A. L. Tuan, M. N. Duc, and M. Hauswirth, "RDF stream processing with CQELS framework for real-time analysis," *Proceedings of the 9th ACM International Conference on Distributed Event-Based Systems*, pp. 285–292, 2015.
[19] J. Han, E. Haihong, G. Le, and J. Du, "Survey on NoSQL database," *2011 6th international conference on pervasive computing and applications*, pp. 363–366, 2011.
[20] X. Meng, Y. Jiang, Y. Chen, and H. Wang, "XSeq: an indexing infrastructure for tree pattern queries," *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pp. 941–942, 2004.
[21] Espertech, "Esper-Complex Event Processing," 2014.
[22] A. L. Blockmon, "Spectroscopic Analysis of Vibronic Relaxation Pathways in Molecular Spin Qubit [Ho (W5O18) 2] 9-: Sparse Spectra Are Key," *Inorganic Chemistry*, vol. 60, no. 18, pp. 14096–14104, 2021.
[23] D. Anicic, S. Rudolph, P. Fodor, and N. Stojanovic, "Stream reasoning and complex event processing in ETALIS," *Semantic web*, vol. 3, no. 4, pp. 397–407, 2012.
[24] M. H. Iqbal and T. R. Soomro, "Big data analysis: Apache storm perspective," *International journal of computer trends and technology*, vol. 19, no. 1, pp. 9–14, 2015.
[25] R. Delvalle, G. Rattihalli, A. Beltre, M. Govindaraju, and M. J. Lewis, "Exploring the design space for optimizations with apache aurora and mesos," *2016 IEEE 9th International Conference on Cloud Computing*, pp. 537–544, 2016.
[26] M. Mead, A. Sieben, and J. Straub, *Coming of age in Samoa*. 1973.
[27] Z. Zhuang, T. Feng, Y. Pan, H. Ramachandra, and B. Sridharan, "Effective multi-stream joining in apache samza framework," *2016 IEEE International Congress on Big Data*, pp. 267–274, 2016.
[28] A. N. Gade, T. S. Larsen, S. B. Nissen, and R. L. Jensen, "REDIS: A value-based decision support tool for renovation of building portfolios," *Building and environment*, vol. 142, pp. 107–118, 2018.
[29] D. F. Barbieri, D. Braga, S. Ceri, E. D. Valle, and M. Grossniklaus, "C-SPARQL: SPARQL for continuous querying," *Proceedings of the 18th international conference on World wide web*, pp. 1061–1062, 2009.
[30] K. Wang and S. Brant, "Building a local passive DNS capability for malware incident response," *Virus bulletin conference*, 2016.
[31] A. B. U'ren, K. Banaszek, and I. A. Walmsley, "Photon engineering for quantum information processing," 2003.
[32] Q. Jiang, R. Adaikkalavan, and S. Chakravarthy, "MavEStream: synergistic integration of stream and event processing," *2007 Second International Conference on Digital Telecommunications (ICDT'07)*, pp. 29–29, 2007.