Real-Time data analysis with Apache Spark

Shree Shingre,Tanmay Y,Ninad Jamkar,Nishant Baruah[1,]

[1]

Introduction

In the realm of real-time data analysis with Apache Spark, the pursuit of optimized performance stands as a critical imperative. As organizations increasingly rely on Apache Spark to glean timely insights, the challenge lies in enhancing resource efficiency and reducing latency without compromising the quality of service. In the era of the Internet of Things (IoT), where the demand for real-time processing surges, Apache Spark serves as a pivotal framework. Its capabilities promise to address latency concerns, yet the dynamic nature of real-time data analysis environments necessitates sophisticated strategies. This literature review delves into the field, exploring how Apache Spark, characterized by its decentralized architecture, addresses challenges in real-time data analysis. Through a systematic examination of current research, the review aims to identify trends, challenges, and gaps in the existing body of knowledge, providing a comprehensive overview of the state-of-the-art in real-time data analysis with Apache Spark.

Literature Review

[1]The paper presents a framework for real-time generation of network traffic statistics using Apache Spark Streaming and demonstrates its suitability for network traffic monitoring, while also proving the feasibility of implementing basic methods for NetFlow data analysis in the stream processing framework. Moreover, the authors highlight that their stream processing implementation uncovers new information not available with traditional network monitoring approaches. Stream processing systems like Apache Spark Streaming provide enough throughput to process a large volume of NetFlow data, making them suitable for network traffic monitoring. - The integration of Apache Spark Streaming into a current network monitoring architecture is possible and allows for the implementation of the same basic methods for NetFlow data analysis as traditional approaches. - The stream processing implementation discovers new information that is not available when using traditional network monitoring approaches. The methodology involves the development and integration of a framework for real-time generation of network traffic statistics using Apache Spark Streaming, and the implementation of basic methods for NetFlow data analysis in the stream processing framework. The limitations of the study include the current state of the Apache Spark Streaming system, which is still under development and lacks critical features such as the recovery of lost data in the stream.

[2]The paper focuses on real-time analysis of streaming Twitter data, including hashtag analysis and measurement of happiness, using Spark streaming and an authentication framework, in an experimental environment using Linux, Ubuntu 16.04 LTS, Spark 2.2.0, and Scala 2.11 version. The paper presents an approach to analyzing streaming data in real-time, specifically focusing on real-time Twitter data using Spark streaming for processing and analyzing hashtags related to specific keywords. - The paper highlights the importance of text mining and natural language processing techniques in deriving useful information from Twitter streaming data. - The paper provides insights into the process of gathering and analyzing Twitter streaming data using Spark and the specific tools and technologies used for the analysis.

[3]The summary of the paper is the need for a model that can perform real-time analysis of data from various sources, including social media, and make it useful for analysis. The current analytical models are not ideally suited for real-time analysis of data. The proposed model will collect data from structured and unstructured sources, filter relevant data in real-time or from stored data, and make it useful for analysis and processing. This model is expected to be more successful in performing real-time analysis compared to current models. The methodology involves collecting data from structured and unstructured sources, filtering relevant data in real time or from stored data, and making it useful for analysis and processing.

[4]The paper discusses an open-source framework for stream processing and big data, an in-memory handling model with machine learning algorithms, a comparison between non-distributed and distributed data usage, and the capability of the Apache Spark platform in handling big data sets with parallel speedup. The Apache Spark platform achieves immaculate parallel speedup for handling big data sets. - In-memory processing model with machine learning algorithms is implemented for stream processing and big data. - The data used in a subset of non-distributed mode is found to be better than using all data in distributed mode.

[5]The paper discusses the advantages of Apache Spark over Hadoop MapReduce, emphasizing lower latency queries, iterative computations, and real-time processing, as well as

the focus on time-series analysis in both environments. Apache Spark ensures lower latency queries, iterative computations, and real-time processing on similar data. - The paper focuses on time-series analysis in Hadoop and Spark environment, which processes and analyzes real-time data to generate patterns for a clearer glimpse of the statistics and characteristics of data, making Spark more efficient over MapReduce. The methodology involves discussing the advantages of Apache Spark over Hadoop MapReduce, analyzing real-time data using time-series analysis, and comparing the time-series analysis in the Hadoop and Spark environment.

[6]The paper discusses the implementation of a real-time data pipeline architecture using Apache Flume, Apache Spark, and Apache Hive for collecting, processing, and storing students' activities data from an E-Learning platform, emphasizing the importance of real-time analytics, big data, and the use of open source technologies like Apache Hadoop, Apache Hive, Apache Spark, and Apache Flume in this context. It also highlights the revolutionary nature of E-Learning and the significance of learning analytics in understanding learners' interactions with educational resource.

[7]The paper aims to predict the total traffic count of streaming data in various routes to reduce traffic congestion and inform the public about the current traffic condition by displaying it on a dashboard. Real-time traffic monitoring is achieved using sensor-connected devices, Apache Kafka, and Spark streaming engine, with improvements in real-time traffic prediction and dashboard updating. The main findings are the limitations of the existing traffic prediction system and the improvements proposed for real-time traffic prediction using live streaming data and in-memory processing. The methodology involves the use of sensor connected devices, Apache Kafka, Spark streaming engine, connected vehicles, Apache Hadoop, and spring boot for real-time traffic monitoring and prediction.

[8]The paper discusses the widespread use of Big Data, the implementation of Hadoop MapReduce and Apache Spark for real-time data processing, experimental simulations comparing the two implementations, and the drawbacks of using Hadoop for real-time processing. The paper highlights the implementation of Hadoop MapReduce and Apache Spark for real-time data processing, conducts experimental simulations to analyze real-time data streams using Spark and Hadoop, and introduces a comparison of the two implementations in terms of architecture and performance, along with a discussion of the results of simulations and the drawbacks of using Hadoop for real-time processing.

[9] The paper discusses the growing demand for efficient methods for processing large-scale heterogeneous data in real-time, particularly in the context of vehicle traffic data streams. It also highlights the challenge of performing low-latency analysis with real-time data and introduces a Big Data cloud platform with ingestion, analysis, storage, and data query APIs. The methodology involves evaluating existing methods and tools in distributed and parallel computing, storage, query, and ingestion, as well as introducing a new Big Data cloud platform. The paper discusses the challenges of performing low-latency analysis with real-time data, particularly in the context of vehicle traffic data streams, and introduces a Big Data cloud platform for analytics system development and evaluation.

[10] The paper discusses the significance of data in the internet world, the challenges of gathering data from social networking websites, and proposes a framework for real-time data streaming of Twitter data using Apache Spark and Scala IDE, with the potential for statistical analysis and report generation. The paper proposes a framework for real-time data streaming of Twitter data using Apache Spark, introduces a GUI for direct tweeting into Twitter, and highlights the potential use of obtained data for statistical analysis and report generation. The methodology involves proposing a framework for real-time data streaming of Twitter data using Apache Spark and Scala IDE, direct data ingestion from Twitter, GUI for direct tweeting, and categorization of tweets using '#' tags.

[11] The paper provides an overview of the book "Pro Spark Streaming" by Zubair Nabi, which focuses on leveraging Spark Streaming for real-time, streaming applications in various industries such as social media, finance, online advertising, and IoT, emphasizing the use of DStreams and micro-batch processing to support streaming analysis. It highlights the book's practical approach with ready-to-deploy examples and actual code. The book covers end-to-end real-time application development using real-world applications and datasets, emphasizes the importance of DStreams and micro-batch processing, and targets data scientists, big data experts, BI analysts, and data architects.

[12] The paper presents the development of a new system in Spark for extracting packet features with less memory consumption and at a faster rate, addressing the need for network monitoring systems to capture network packets and provide packet features in near real time to protect from attacks. The system described in the paper focuses on real-time stream processing using Spark streaming technology, which enables traffic analysis and extraction of packet features in a distributed computation environment. The methodology involves the development of a new system in Spark for extracting packet features with less memory consumption and at a faster rate, utilizing the streaming capability inherent in Spark for traffic analysis and extraction of packet features, with a focus on real-time stream processing using Spark streaming technology.

tical constraints based on extreme value theory to address the ultra-reliable low-latency needs of mission-critical applications. It aims to minimize power consumption by optimizing the balance between local computation and task offloading, considering wireless channel dynamics and server computation capabilities. A user-server association policy, guided by channel quality and server workload, is proposed alongside a two-timescale mechanism combining Lyapunov optimization and matching theory for dynamic task offloading and resource allocation. Simulations show this approach significantly improves task computation reliability and reduces delays compared to traditional methods.

[13] The paper discusses the development of an analytical framework for real-time data analytics of Twitter data, focusing on the quality of captured data and the ability to perform spatial and temporal analyses. The study acknowledges the limitations of social media data as it reflects people's opinions and feelings about ongoing issues.The paper aims to develop an analytical framework for real-time data analytics of Twitter data, including data ingestion, stream processing, and data visualization components with the Apache Kafka messaging system. The proposed framework is designed to not only perform basic processing tasks but also make an infrastructure for performing more sophisticated and complicated analytics on streaming data. The study emphasizes the importance of dealing with social media data, which includes various data types arriving in large volumes every second, and the need for a proper framework that can process data in memory as it arrives.

Conclusions

In conclusion, this literature review provides a comprehensive overview of the state-of-the-art in real-time data analysis with Apache Spark. The exploration of foundational principles, methodologies, and innovations demonstrates the framework's pivotal role in addressing the challenges associated with processing vast datasets in distributed computing environments. As organizations increasingly adopt Apache Spark for timely decision-making, understanding its architecture, programming model, and key features becomes paramount. The critical evaluation of research studies, case implementations, and practical considerations illuminates both the strengths and limitations of Apache Spark in real-time data analysis. The synthesis of existing knowledge not only caters to researchers and practitioners seeking an informed understanding but also highlights potential avenues for future research. With Apache Spark at the forefront of real-time analytics, this literature review contributes to the ongoing discourse, emphasizing the importance of continued exploration and innovation in this dynamic field.

References

1. Real-time analysis of NetFlow data for generating network traffic statistics using Apache Spark Milan Cermák, Tomás Jirsík, Martin Laštovička Published in IEEE/IFIP Network Operations… 25 April 2016
2. Real-time Streaming Data Analysis using Spark Kyeongjoo Kim Published 23 January 2018 Computer Science
3. Analysis of Big Data using Apache Spark Ankur Saxena, Monika Chand, +3 authors Inish Krishna Shreshtha Published 4 April 2020
4. SCSI: Real-Time Data Analysis with Cassandra and Spark A. Chaudhari, Preeti Mulay Published in Studies in Big Data 17 June 2018
5. Survey on high performance analytics of bigdata with apache spark Published in IEEE/IFIP Network Operations… 25 April 2016
6. Real-Time Analysis of Students' Activities on an E-Learning Platform based on Apache Spark. Abdelmajid Chaffai, L. Hassouni, H. Anoun Published 2017
7. Real-Time Traffic Monitoring System Using Spark A. Saraswathi, Mummoorthy A, +1 author KP Porkodi Published in International Conference on… 1 September 2019
8. Real-time data analysis using Spark and Hadoop Khadija Aziz, Dounia Zaidouni, M. Bellafkih Published in OPTIMA 26 April 2018
9. Low Latency analiytics for streaming traffic data with Apacha Spark Khadija Aziz, Dounia Zaidouni, M. Bellafkih Published in OPTIMA 26 April 2018
10. Real-time Data Streaming using Apache Spark on Fully Configured Hadoop Cluster K. Prasad Published in JOURNAL OF MECHANICS OF… 21 December 2018
11. Pro Spark Streaming: The Zen of Real-Time Analytics Using Apache Zubair Nabi Published 13 June 2016
12. Network Data Analysis Using Spark K. Swetha, S. Sathyadevan, P. Bilna Published in Computer Science On-line… 2015
13. Developing a Real-Time Data Analytics Framework for Twitter Streaming Data Babak Yadranjiaghdam, S. Yasrobi, Nasseh Tabrizi Published in BigData Congress [Services… 1 June 2017