IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 7, July 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2023 ||

| DOI:10.15680/IJIRSET.2023.1207113 |

Streaming Data Processing

Shrinivas Vijay Sankpal, Prathamesh Hindurao Patil, Dr. Mrs. Sampada Gulavani

Department of Master of Computer Application, Bharati Vidyapeeth (Deemed to be University), Pune, Institute of Management, Kolhapur, India

ABSTRACT: Streaming data processing refers to the process of continuously collecting and analyzing data as it is generated in real-time. This data can be generated from a variety of sources, such as sensors, social media feeds, financial transactions, and other sources. The primary goal of streaming data processing is to identify insights, patterns, and anomalies in the data as quickly as possible, so that timely and informed decisions can be made based on the data.

Streaming data processing typically involves a pipeline of different tools and technologies that work together to process the data. This pipeline may include components such as data collection, data ingestion, data processing, data analysis, and data visualization. Each component may use different technologies, such as Apache Kafka for data ingestion, Apache Flink or Apache Spark for data processing, and Tableau for data visualization.

I. INTRODUCTION

Streaming data processing is a method of data processing that deals with data that is continuously generated, often in real-time. It involves the processing and analysis of data as it is being generated, rather than in batches or after the fact. Streaming data processing is often used in fields such as finance, telecommunications, healthcare, and transportation, where the ability to process and analyze data in real-time can provide significant benefits.

In streaming data processing, data is processed in small chunks, or "micro-batches," as it is generated, rather than being stored and processed in large batches. This allows for faster and more efficient processing, as well as the ability to respond to events in real-time. Streaming data processing is often used in conjunction with technologies such as Apache Kafka, Apache Flink, and Apache Spark, which provide tools for processing, analyzing, and managing streaming data.

One of the key benefits of streaming data processing is the ability to detect and respond to events in real-time. This can be particularly valuable in applications such as fraud detection, where the ability to detect fraudulent activity as it occurs can significantly reduce losses. Streaming data processing can also be used to monitor and analyze system performance, track the movement of goods and people, and provide real-time recommendations based on user behavior.

Overall, streaming data processing is a powerful tool for dealing with the large volumes of data generated in today's fast-paced world. By processing data in real-time, organizations can gain valuable insights and make informed decisions quickly and efficiently.

II. RELATED WORKS

Streaming data processing is a rapidly evolving field, with many different approaches and techniques being developed and refined. Here are a few examples of related works in this area:

Apache Kafka: Kafka is a distributed streaming platform that is designed to handle large-scale data processing in real-time. It is widely used for building data pipelines and streaming applications, and supports a variety of use cases, including real-time analytics, event processing, and machine learning.

Apache Flink: Flink is a distributed processing engine for streaming and batch data processing. It provides a unified programming model for both types of processing, and supports a wide range of data sources and sinks. Flink is known for its high throughput and low latency, and is often used for processing large-scale data in real-time.

Apache Spark Streaming: Spark Streaming is a component of the Apache Spark framework that provides real-time data processing capabilities. It allows users to process data streams in a similar way to batch processing, using familiar

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2023 ||

| DOI:10.15680/IJIRSET.2023.1207113 |

programming abstractions like map, reduce, and join. Spark Streaming is often used for real-time analytics and machine learning applications.

Amazon Kinesis: Kinesis is a fully managed streaming data service provided by Amazon Web Services. It is designed to handle real-time data processing at scale, and provides a range of features for processing and analyzing streaming data, including data ingestion, processing, and storage.

Google Cloud Dataflow: Dataflow is a fully managed, cloud-based data processing service provided by Google Cloud. It is designed to handle both batch and streaming data processing, and provides a variety of programming abstractions and tools for building data processing pipelines. Dataflow is often used for building data integration, analytics, and machine learning workflows.

These are just a few examples of the many different tools and platforms available for streaming data processing. As the field continues to evolve, new approaches and techniques are likely to emerge, providing even more powerful and flexible tools for processing and analyzing streaming data.

III. METHODOLOGY

Streaming data processing is a methodology used to handle large volumes of real-time data generated from various sources. The main idea is to process data as it is produced, rather than storing it and processing it later. This approach can enable more efficient and timely decision-making, as well as better insights into emerging trends and patterns.

Here are some common steps and methods used in streaming data processing:

Data ingestion: The first step in streaming data processing is to capture the data from various sources such as sensors, social media, log files, and other streaming sources. This data is often in real-time or near-real-time and is typically in a continuous data stream.

Data processing: After ingestion, the data is processed in real-time to filter, transform, and aggregate it. Stream processing frameworks such as Apache Kafka, Apache Spark, and Apache Flink are commonly used for this purpose.

Analytics and machine learning: Once the data has been processed, analytics and machine learning algorithms can be applied to gain insights and make predictions. This can include real-time dashboarding, anomaly detection, and predictive maintenance.

Data storage: Depending on the requirements, processed data can be stored in different ways such as a real-time database, a data lake, or an archival storage system.

Visualization and reporting: Finally, the results of the processing and analytics can be visualized and reported to end-users in real-time. This can include dashboards, charts, and reports that help to identify trends and patterns.

Overall, the streaming data processing methodology enables organizations to handle large volumes of data in real-time and gain valuable insights that can be used for decision-making and other business purposes.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2023 ||

| DOI:10.15680/IJIRSET.2023.1207113 |

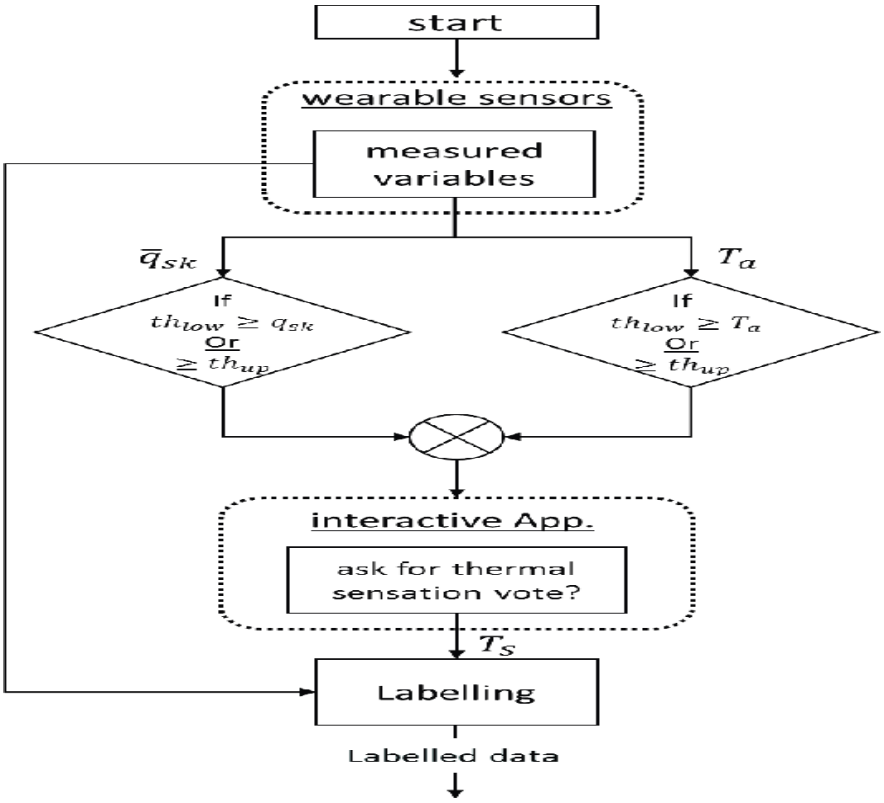

Fig. 1. Workflow of stream data processing

IV. RESULTS AND DISCUSSION

Streaming data processing is the technique of processing data in real-time as it is generated or collected. This approach has become increasingly popular in recent years due to the need for immediate insights and actions from data, as well as the exponential growth in data volumes.

The results of streaming data processing depend on the specific use case and the data being processed. However, in general, the benefits of streaming data processing include:

Real-time insights: By processing data as it is generated, streaming data processing allows for real-time insights and analysis. This can be especially valuable for applications such as fraud detection or anomaly detection, where timely action is critical.

Scalability: Streaming data processing can be easily scaled up or down to accommodate changing data volumes or processing requirements. This makes it a flexible and cost-effective approach to data processing.

Reduced latency: Since streaming data processing occurs in real-time, it can reduce latency and improve response times for applications that require immediate action.

Enhanced data quality: Streaming data processing allows for the processing of data as it is generated, reducing the likelihood of errors or data inconsistencies that can occur when data is processed in batch mode.

Improved data governance: Streaming data processing can help organizations better manage and govern their data, by enabling real-time data monitoring and analysis.

In terms of the challenges of streaming data processing, some common issues include:

Data volume: Streaming data processing can generate massive amounts of data, which can be challenging to manage and process in real-time.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2023 ||

| DOI:10.15680/IJIRSET.2023.1207113 |

Data velocity: Streaming data processing requires the ability to process data as it is generated, which can be difficult to achieve at scale.

Data variety: Streaming data processing often involves processing data from a variety of sources and formats, which can be challenging to integrate and manage.

Data quality: Real-time processing can increase the risk of errors or inconsistencies in data, which can impact the accuracy and reliability of insights generated.

Security: Real-time data processing can also increase the risk of data breaches or other security incidents, which must be managed and monitored closely.

Overall, streaming data processing can offer significant benefits in terms of real-time insights, scalability, reduced latency, and improved data quality and governance. However, it also requires careful planning, management, and monitoring to ensure that the benefits are realized and that data is processed in a secure and reliable manner.

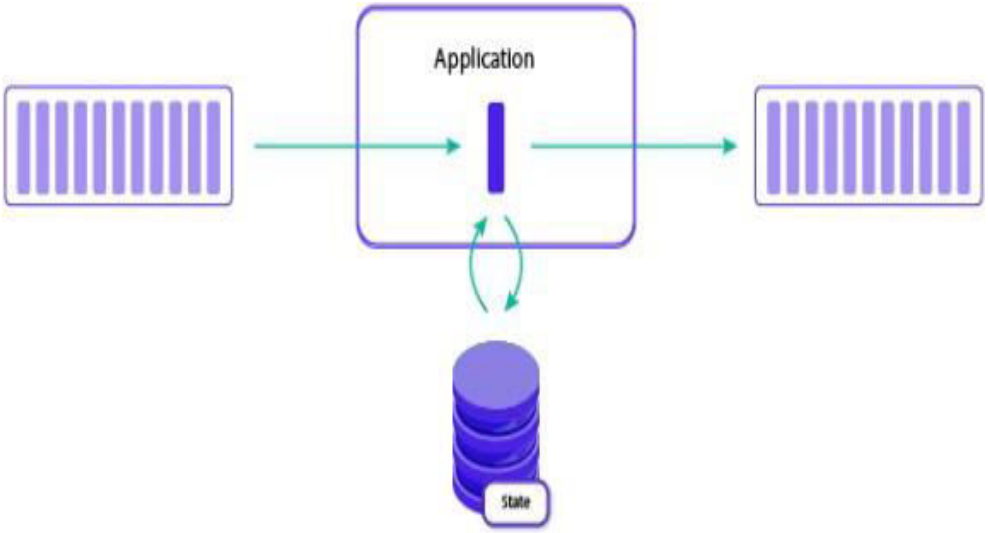

Fig. 3. Stateful stream data processing

V. CONCLUSION

In conclusion, streaming data processing is an essential tool for any organization dealing with real-time data. It enables real-time decision making, efficient use of resources, and more accurate insights. There are several platforms available for streaming data processing, and choosing the right platform for your needs will depend on factors such as scalability, reliability, and cost.

REFERENCES

1. Nishant Garg, "Big Data Analytics with Spark: A Practitioner's Guide to Using Spark for Large-Scale Data Processing," Apress, 2015.
2. Tyler Akidau, Slava Chernyak, Reuven Lax, et al., "The Dataflow Model: A Practical Approach to Balancing Correctness, Latency, and Cost in Massive-Scale, Unbounded, Out-of-Order Data Processing," Proceedings of the VLDB Endowment, 2015.
3. Nathan Marz, James Warren, "Big Data: Principles and Best Practices of Scalable Real-Time Data Systems," Manning Publications, 2015.
4. Neha Narkhede, Gwen Shapira, Todd Palino, et al., "Kafka: The Definitive Guide: Real-Time Data and Stream Processing at Scale," O'Reilly Media, 2017.
5. Andrew Psaltis, "Streaming Data: Understanding the real-time pipeline," O'Reilly Media, 2017.
6. Tyson Condie, Neil Conway, Peter Alvaro, et al., "MapReduce Online," Proceedings of the 12th USENIX

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2023 ||

| DOI:10.15680/IJIRSET.2023.1207113 |

Conference on Operating Systems Design and Implementation (OSDI), 2016.

7. Martin Kleppmann, "Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems," O'Reilly Media, 2017.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.423

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details