Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

Data Stream Pre-processing for Real Time Analytical Processing

Dr.V.Valli Mayil
Associate Professor and Head, Department of Computer Science and Applications, Periyar Maniammai University, Thanjavur, TamilNadu, India

Abstract: Real-time analytics is the process of continuous dynamic analysis of data stream flowing continuously minute by minute from multiple real time sources. Streaming data is a continuous flow of data in high speed from sources like sensors, machines, vehicles, mobile phones, social media networks, and other real-time sources. The organizations and enterprises are focussing on real time data stream analytics rather than batch processing to understand the future patterns to predict the hidden knowledge for decision making. The continuous flows of data are needed to be integrated and pre-processed for real time systems. In this paper, various pre-processing phases for real time analytics are discussed.

I. INTRODUCTION

Real time data stream is a high velocity, high volume of structured as well as unstructured data flowing through real-time sources such as financial & market data, Internet of Things, network monitoring, mobile, sensors, clickstream, and transactions. Analysing these continuous multiple types, time varying data would produce the knowledge used for decision making in various enterprises and organization. Various technology or method of integrating, analysing and predicting the future pattern would be the new challenges and issues for research analyst.

In the real world, a number of key real-time applications e.g., various management, smart spaces, green buildings, health monitoring, electric grid management, and military surveillance, are in situation to take strong decisions based on real time data. They are required to deal with the fast-increasing size and speed of real-time data streams. Thus, processing big real-time data streams in a timely manner to extract valuable information from them is a key issue.

Streaming processing is the suitable platform to process data streams or sensor data. The traditional data model is earliest one, in which data is stored in a data store and query is used to manipulate the data. Different types of unstructured and semi structured from different sources is also processed under stream processing. In traditional model the stored large volume of data are processed and analysed to fine the knowledge for decision making. In stream data model the movement of data is taken for analysis as the continuous ingestion and continuous complex analysis is taken place to provide continuous intelligence for powerful analytics.

Different challenges in Stream Data are

- Dissemination of large volume of streaming data from different sources
- Rapid integration of various sources of data into unified structured schema
- Updating data for scalability as data volumes increase in size and complexity
- Processing massive amounts of streaming events (filter, aggregate, rule, automate, predict, act, monitor, alert)
- Analytics: Live data discovery and monitoring, continuous query processing, automated alerts and reactions
- Real-time Analytics for pattern prediction to make necessary decision
- Visualizing or integrating output patterns into different types such as statistical reports or messages

The different requirements for a stream data platform are that the platform must be

- Reliable to handle critical updates such as searching and indexing without loss of data.
- Support enough to handle large volume log or event data streams.
- Buffer is used to keep the data for long time
- Provide data for real-time applications.
- Scalable system which supports the organization to include more modules in order to operate more applications.
- These requirements are necessary to handle flow of data analytics. The goal of the stream data platform is to provide full support for large volume data.

The real time analytical platform combines both streaming data integration and streaming operational intelligence in a single platform. The efficient framework is required to acquire , integrate, transform, organize, process and analyses the real-time data in a single platform. The platform must addresses the model of integrating multiple heterogeneous data into decision support system. The framework can be categorized into 3 main components that interact with each other, namely Data Ingestion, Data Transformation and Aggregation, Data Store and Data Analytics.

II. PREPROCESSING A REAL TIME DATA STREAM DATA

The different sources of data available for a stream model are Web, Data bases, Sensor, Mobile, e-mail etc. The different types of data are text, video, images are in structured and unstructured format. The continuous data must be pre-processed for integration, cleaning and transformation

There are five main phases are considered for the architecture of sream model

- Data Integration and Pre-processing : Data from heterogeneous sources are collected and stored in a unified format for future use. In includes data Integration, filtering and Transformation
- Stream Data Storage: Data that are integrated must be stored in a scalable storage
- Information Retrieval: Useful information can be searched and retrieved from the storage by the appropriate query

- Data Analytics Methods – Distributed Data Processor- Collected data can be mined and interesting pattern or trends can be extracted.
- Decision Making – Analysis of data is used for various decision making

In this paper we discussed various methods for pre-processing multiple sources of data and method for preparing for further analysis. Pre-processing phase of real time data analytical method consists of the processes such as Data ingestion and Integration, Data filtering, Data transformation

A. Data Integration and ingestion

Data integration is the process of combining data from different sources and providing the user in a common format.

Challenges in Data Integration of streaming data

Streaming data are extremely heterogeneous, both structure and unstructured in nature, which creates a need for data integration and ingestion into scalable data storage. The types of streaming data in the form of information in text format, Message data from mobile device, Web data such as text, pdf or word format, Databases, data in blogs. All these data are in many different format, different characteristics and are received from different sources, also follows many uncertainty. The availability of such data poses a challenge for design and development of different strategies for acquisition, integration and ingestion. The author Naumann F., Raschid, L., 2006 provides the framework for data integration.

Lenzerini M., 2002., uses OWL 2QL ontology language to model global schema. He presents on overview on modelling a data integration application, processing queries in data integration and inconsistent data sources. The author explains how OWL or SPARK SQL can be used for data integration.

A. Arasu, S. Babu, and J. Widom STREAM's formal continuous query model takes its basis from the well-understood relational model. More specifically, STREAM's Continuous Query Language (CQL) is an extension to SQL:1999. First, it introduces "stream" as a second data type in addition to "relation". Second, in addition to the "relation-to-relation" operations of the relational algebra, CQL introduces "stream-to-relation" operations for constructing windows on streams as well as "relation-to-stream" operations to convert results of relational operations back into the stream data type

Simon Beckstein., et al., explains the method of Integrating Semantic Knowledge in Data Stream Processing. The author discussed Complex Event Processing (CEP) technology for processing high-frequent data streams. The intelligent stream based systems are integrated with semantically background knowledge. The method of adding an ontology access mechanism to a common Continuous Query Language (CQL) is compared with C-SPARQL, a streaming extension of the RDF query language SPARQL.

The second architecture is based on an extension of SPARQL to process RDF data streams. C-SPARQL allows integrated rules that process stream data and query RDF triple stores containing static background knowledge.

LUO Jinga, DANG An-ronga, attempted to adopt the method of multi-layer ontology sharing mechanism to implement data sharing, and put forward approaches of constructing ontology data integration platform and then use this platform to realize data integration.

B. Proposed solution for data integration

In the basic data stream model, the input data is a continuous flow of data that could be read sequentially. The continuous data is originated from multiple sources/formats including databases, documents, web, e-mail, etc. This scenario is relevant for a large number of applications where massive amounts of data need to be processed. The effective and efficient algorithms are needed to address the data integration and ingestion.

Employing Data Stream query languages

Patrick Lehti and peter Fankhaser explains SWQL A query language for data integration based on OWL this semantic web query language integrates the heterogeneous data. This approach combines query language and meta data integration system to build a unified data model.

The data sources for data stream processing are extremely heterogeneous. In multiple heterogeneous systems, different name references are used for same entity. The data integration system receives data from a given source at an indeterminate and varying rate. The data integration system is needed to accommodate the various meaning of same entity and should produce the common entity name. The similarity of two different entity names is estimated by the technique of information retrieval [Coh98] and on probabilistic metrics [PR01]. The common format such as XML is used to represent the semantic entities. The continuous data from different sources are converted into XML representation as it has the vital characteristics to represent structured, semi structured and unstructured data such as relational and object-oriented data, word text documents, spreadsheets and graphics images.

The data integration module standardise the data from multiple sources and convert the data into common schema called XML.

StreamSQL is a structured query language developed to manipulate real time stream of events. The streams contains a sequences of tuples generated continuously. Hence the operations over streams must be monotonic. StreamSQL executed on streams returns the result as increment updation.

The operations in StreamSQL includes SELECT and WHERE clause, that are used to filter unwanted tuples. The Windowing operations in a stream are used to limit the creation of tuples. The aggregation is used to find analytical operations such as count, average, max, etc. **Windowing and Joining** operations are used to limit the stream data over a period of time and joined the stream of separate windows.

Modern Tool for Knowledge Extraction

SQLstream is a streaming analytics platform providing tool and services for data stream applications.

Knowledge extraction and transformations is the type of knowledge discovery whose goal is to extract structured information from the integrated unified schema. **Knowledge extraction** is the generation of knowledge from structured XML sources. The multi source data are integrated and stored in the XML format. Hence, obtaining result from XML using query is essential.

Oracle Data Integration solutions supports ETL based data integration tool. It enables the user to manipulate stream oriented realtime analytics. It support a variety of big data standards includs MapReduce, Pig Latin, HiveQL and Spark solutions.

The Kafka framework from Apache foundation based on distributed, partitioned and replicated system which supports the context of streaming process. The streaming data are collected as logs and collected and published to Kafka continuously as a pipelining process. The stream data are then collected aggregated for future process. An another alternative open source stream processing tools include Apache Storm and Apache Samza.

The open source software for batch and streaming are MapReduce and Storm. Flink and Spark are general-purpose data processing platforms for top level projects of the Apache Software Foundation (ASF).

Flink is a data stream engine that supports continuous data stream and process the stream with the properties of distribution, fault tolerance computation. It has many API such as DataStreamAPI, DataSetAPI, TableAPI. Flink performs stream processing with **Event Time** semantics. It applies widowing process which accumulates the data when it arrives for particular time events.

Spark platform combines SQL, streaming, and complex analytics. Spark supports various technologies such as SQL data frames, Mlib for machine learning, Spark streaming. Spark can be installed and implemented in on Hadoop, Mesos, standalone, or in the cloud. It can access data sources such as HDFS, Cassandra, HBase, and S3

CONCLUSION

Data preprocessing in real time data analytics is a critical phases in data stream applications. Data integration from multiple sources, cleaning and transformation takes important contribution for data analytics. In this paper we discussed the contribution of XML unified model and semantic information to integrate and transform the data into inified XML structure. Recently many innovative tools have been developed to address the problem of data integration. In this paper we highlighted the tools and its importance shortly.

References

[1] Lenzerini M., 2002., Data Integration : Theretical perspective. In.PODS 2002, 233-246.

[2] Naumann F., Raschid, L., 2006., Information Integration and DisDM workshop on Information Integration, Philadephia, USA, Oct 25-27

[3] LUO Jinga, *, DANG An-ronga, MAO Qi-zhia, The Study Of Integration Of Multi-Sources Heterogeneous Data Based On The Ontology

[4] WU Hao, XING Gui-fen. Research on technology of information integration based on ontology, Computer Applications, 2005-02-01

[5] H.Wache, T.Vogele, U.Visser, H.Stuckenschmidt, G.Schuster, H.Neumann, and S.Hübner. "Ontology-based Integration of Information-A Survey Existing Approaches," In: Proceedings of IJCAI_01 Workshop: Ontologies and Information Sharing, Seattle, WA, 2001, Vol.pp.108-117.

[6] Simon Beckstein, Ralf Bruns, J¨urgen Dunkel, Leonard Renners, Integrating Semantic Knowledge in Data Stream Processing