Difference between Batch processing vs Stream Processing? When to use?

Description

What is Batch processing

Batch processing is a processing which is done either on an ad-hoc basis (by invoking it manually) or at scheduled time against a collection of data. Batch processing is a technique which processes data in large group or volume with minimal interaction, instead of a single data element.

Use cases of Batch processing

If we consider the use case of banks or any other financial domain, there are end of day (EOD) files that arrive for processing. These end of day files contain records in thousands or millions and the file size is also huge. The batch is run either daily or weekly or monthly based on the use case and business domain. Another example of batch processing is the end of month payroll information processing.

Batch processing can be used for use cases like reading and writing to files, transforming data from one form to other form, reading from or writing to databases, creating reports using source data, import data from one format and export to other format from one database to other etc.

Batch processing frameworks will parse your files, process them , and transform them into a specific format and store them into a database. Further we can use analytics to show the processed data in analytical tools like Qlik sense or Tableau.

What is Stream processing

Stream processing is also called as real time processing which processes data instantaneously as they are pushed. Batch processing relies on collection of data, whereas streams rely on continuous stream of data. Stream can be used for fraud detection, decision making, pattern learning etc.

Use cases of Stream processing

Lets consider an example use case of a web application which is being accessed across the globe. We want to know the accessibility information of people visiting the website by country. Whenever someone visits the website, there will be an event that is sent to a Kafka topic containing information like country name, page visited etc. All these information are streamed real time inside our Kafka topic. What does streaming mean here? Here the information is continuously pushed into the Kafka topic. From this Kafka topic we can use different frameworks like Apache Flink or Apache Storm in order to consume the messages in real time and provide the required information.

The end goal of our use case can be to provide more scalable services efficiently to the country from where we are getting the maximum traffic.

Micro batching – Hybrid approach

Micro batching is the process of incorporating both batch and stream processing. the Micro batches are a collection of both batch and streams and you can process data instantaneously but you will batch the data for short span of time, so you have incorporated the scheduled processing feature and also collection of data along with real time processing and continuous processing of data. So micro batches are collection of both batch and streams and you can process data instantaneously but you will batch the data for short span of time. Compared to batches you had to add to do ad-hoc processing or scheduled processing for longer period of time, in micro batches you do in a short span of time. These are the basic difference between batching streaming and micro batching.

Category

1. Design

Date Created
December 26, 2022
Author
kk-ravi144gmail-com