Describe the difference between batch and streaming data

3 minutes

Data processing is simply the conversion of raw data to meaningful information through a process. Depending on how the data is ingested into your system, you could process each data item as it arrives, or buffer the raw data and process it in groups. Processing data as it arrives is called *streaming*. Buffering and processing the data in groups is called *batch processing*.

Understand batch processing

In batch processing, newly arriving data elements are collected into a group. The whole group is then processed at a future time as a batch. Exactly when each group is processed can be determined in a number of ways. For example, you can process data based on a scheduled time interval (for example, every hour), or it could be triggered when a certain amount of data has arrived, or as the result of some other event.

An example of batch processing is the way that votes are typically counted in elections. The votes are not entered when they are cast, but are all entered together at one time in a batch.

Advantages of batch processing include:

- Large volumes of data can be processed at a convenient time.
- It can be scheduled to run at a time when computers or systems might otherwise be idle, such as overnight, or during off-peak hours.

Disadvantages of batch processing include:

- The time delay between ingesting the data and getting the results.
- All of a batch job's input data must be ready before a batch can be processed. This means data must be carefully checked. Problems with data, errors, and program crashes that occur during batch jobs bring the whole process to a halt. The input data must be carefully checked before the job can be run again. Even minor data errors, such as typographical errors in dates, can prevent a batch job from running.

An example of an effective use of batch processing would be a connection to a mainframe system. Vast amounts of data need to be transferred into a data analysis system and the data is not real-time. An example of ineffective batch-processing would be to transfer small amounts of real-time data, such as a financial stock-ticker.

Understand streaming and real-time data

In stream processing, each new piece of data is processed when it arrives. For example, data ingestion is inherently a streaming process.

Streaming handles data in real time. Unlike batch processing, there's no waiting until the next batch processing interval, and data is processed as individual pieces rather than being processed a batch at a time. Streaming data processing is beneficial in most scenarios where new, dynamic data is generated on a continual basis.

Examples of streaming data include:

- A financial institution tracks changes in the stock market in real time, computes value-at-risk, and automatically rebalances portfolios based on stock price movements.
- An online gaming company collects real-time data about player-game interactions, and feeds the data into its gaming platform. It then analyzes the data in real time, offers incentives and dynamic experiences to engage its players.
- A real-estate website that tracks a subset of data from consumers’ mobile devices, and makes real-time property recommendations of properties to visit based on their geo-location.

Stream processing is ideal for time-critical operations that require an instant real-time response. For example, a system that monitors a building for smoke and heat needs to trigger alarms and unlock doors to allow residents to escape immediately in the event of a fire.

Understand differences between batch and streaming data

Apart from the way in which batch processing and streaming processing handle data, there are other differences:

- *Data Scope*: Batch processing can process all the data in the dataset. Stream processing typically only has access to the most recent data received, or within a rolling time window (the last 30 seconds, for example).

- *Data Size*: Batch processing is suitable for handling large datasets efficiently. Stream processing is intended for individual records or *micro batches* consisting of few records.

- *Performance*: The latency for batch processing is typically a few hours. Stream processing typically occurs immediately, with latency in the order of seconds or milliseconds.

Latency is the time taken for the data to be received and processed.

- *Analysis*: You typically use batch processing for performing complex analytics. Stream processing is used for simple response functions, aggregates, or calculations such as rolling averages.

Next unit: Knowledge check

Continue >