**REPORT**

# Batch Processing

Batch processing is a data processing method in which data is collected, processed, and stored in large batches, rather than processing each data record individually or as it appears.

Batch processing is often used in large data processing systems, where data is processed at fixed intervals or when a certain amount of data is reached.

## Main features

* Processing time: Data is processed in large batches, often at specific times, such as hourly, daily, or weekly.
* Data volume: Process large amounts of data in each batch.
* Latency: There is higher latency than streaming processing, because the data has to wait until enough batches are processed.
* Suitable for: Tasks such as reporting, historical analysis, processing transactions that do not require real time.

## Example

* Payroll calculations: Payroll processing systems collect and process employee salary data at the end of each month.
* Financial reports: The system generates financial reports that summarize transaction data during the month and generate reports at the end of the month.
* Order management: The order management system processes and updates order information every night to summarize the day's transactions.

# Streaming Processing

Streaming processing is a method of continuously processing data as it is created or received, often in near real time. This allows systems to react and analyze data as it arrives, providing rapid response and continuous updates.

## Main features

* Processing time: Data is processed immediately upon receipt or near real time.
* Data volume: Continuously process each small data record.
* Latency: Very low latency, almost real time.
* Great for: Tasks that require real-time response such as fraud detection, system monitoring, event handling.

## Streaming Processing example:

* Fraud detection: Bank transaction monitoring system to detect unusual and fraudulent transactions as soon as they occur.
* System monitoring: The system monitors servers and applications to detect and react to performance issues or errors in real time.
* Real-time analytics: Social media data analytics system to track trends and respond to customers instantly.

# Compare Batch Processing and Streaming Processing in Spark

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Batch Processing**  **(Spark Core)** | **Streaming Processing (Structured Streaming)** |
| Main API | Spark Core (RDDs, DataFrames, Datasets) | Structured Streaming (DataFrames, Datasets) |
| Processing Model | Processes large batches of data | Processes data continuously in micro-batches |
| Processing Time | Processes data in batches, not real-time | Near real-time with low latency |
| Data Sources | HDFS, S3, databases, file systems | Kafka, Kinesis, socket, file systems |
| Triggers | Not applicable | Can be periodic or continuous |
| Stateful Processing | Limited | Strong support, maintains state between batches |
| Use Cases | ETL, historical analysis, reporting | Real-time analytics, system monitoring, event detection |