ONE PAGE SUMMARY

IT & Software › Data Engineering › Batch Processing

Introduction to Batch Processing

Batch Processing: Efficient Data Handling for Large Datasets

DEFINITION

> Batch processing is a method of executing a series of jobs or tasks on a dataset without manual intervention. It is ideal for processing large volumes of data at once, often scheduled during off-peak hours.

KEY CONCEPTS

Data Batches

Data is grouped into batches for processing, which improves efficiency and resource utilization.

Scheduled Jobs

Batch processes are often scheduled to run at specific times, reducing the impact on system performance.

Throughput

Throughput refers to the amount of data processed in a given time frame, a key metric in batch processing.

Latency

Latency is the delay before a batch job starts processing, which can affect the timeliness of data availability.

Error Handling

Batch processing systems must have robust error handling to manage failures without manual intervention.

EXAMPLES

→ Monthly payroll processing for employees.

→ Daily sales report generation from transaction data.

→ Yearly data archiving for compliance purposes.

MEMORY TIPS

★ Think of 'batch' like baking cookies: you prepare a batch at once instead of one cookie at a time.

★ Remember 'scheduled' like a train timetable; it runs on a fixed schedule.

★ Use 'throughput' as 'through the roof' to visualize high data processing rates.

COMMON MISTAKES

✗ Overlooking error handling, leading to job failures.

✗ Not scheduling jobs during off-peak hours, causing system slowdowns.

✗ Failing to monitor batch job performance metrics.

QUICK RECAP

Batch processing is an efficient way to handle large datasets by executing jobs in groups. It involves scheduled tasks, error management, and focuses on maximizing throughput while minimizing latency.