4_Batch vs Streaming

Tuesday, February 10, 2026 3:23 PM

Batch vs Streaming — Simple, Practical Explanation

1. What is Batch Processing?

Batch means **processing data in bulk at scheduled intervals**.
You wait, collect a chunk of data, and then process it all at once.
Authoritative sources describe batch as:

- "Processing a high volume of data in a batch within a specific time span"
- "Bulk processing of data at predefined intervals"
- "Suitable when immediate results are not required"

When batch makes sense

- Daily/Hourly ingestion
- Historical reporting
- Payroll, billing, month-end processes
- When data size is known and finite
- When latency is not critical

How it feels in real life

You run a job at 2 AM → it processes yesterday's data → loads into warehouse.

2. What is Streaming Processing?

Streaming means **processing data continuously as it arrives**, with very low latency.
Sources describe streaming as:

- "Continuous processing from message buses… low latency"
- "Analyzes data continuously as it arrives"
- "Real-time ingestion and analysis"

When streaming makes sense

- Fraud detection
- Real-time dashboards
- IoT sensor data
- Clickstream analytics
- Alerts and anomaly detection

How it feels in real life

Data arrives → system processes it instantly → dashboard updates in seconds.

3. Key Differences (Clear & Practical)

Latency

- **Batch:** Minutes → hours → scheduled
- **Streaming:** Seconds → milliseconds

Data arrival

- **Batch:** Data is collected first, then processed
- **Streaming:** Data is processed event-by-event

Use cases

- **Batch:** Historical analysis, reporting, large ETL jobs
- **Streaming:** Real-time decisions, monitoring, alerts

Complexity

- **Batch:** Easier to build and maintain
- **Streaming:** More complex (state management, checkpoints, failures)

4. Databricks Perspective (Important for You)

Databricks treats batch and streaming under **one unified engine** (Spark + Structured Streaming):

- "Unified architecture for batch and streaming"
- "Can treat cloud storage and Delta Lake as streaming sources for incremental processing"

This means:

- You can start with batch

- Later convert to streaming with minimal code changes
- Delta Lake makes incremental processing easy