Data Pipelines: Batch vs Stream

Modern data-driven systems rely heavily on efficient data pipelines to extract, process, and move data from source systems to analytical platforms. Two primary paradigms exist in data processing: **Batch Processing** and **Stream Processing**. Choosing between them depends on latency needs, data volume, complexity, and infrastructure.

2. Batch Processing

How It Works:

Data is collected over a period, stored, and then processed as a group (batch). For example, processing all sales transactions at the end of the day.

Architecture:

- **Ingestion**: Data is read from logs/files (e.g., S3, HDFS).
- **Transformation**: Performed using frameworks like Apache Spark or MapReduce.
- **Storage/Output**: Results are stored in databases, warehouses (like Snowflake, Redshift).

Pros:

- Efficient for large volumes.
- Simpler to implement and maintain.
- Ideal for data analysis that isn’t time-sensitive.

Cons:

- Not suitable for real-time needs.
- Delayed data visibility and insights.

4. Key Design Considerations

Criteria	**Description**
Latency	Real-time systems need low latency; batch can tolerate delay.
Scalability	Both systems should scale, but streaming needs fine-grained resource tuning.
Fault Tolerance	Stream processors must handle retries, checkpoints, and state recovery.
Ordering Guarantees	Stream processing might face out-of-order data; batch is more deterministic.
Throughput	Batch systems can handle more data in bulk; streaming needs to be always on.

6. Hybrid Pipelines: Lambda & Kappa Architectures

- **Lambda Architecture**: Combines batch and stream processing for fault tolerance and real-time analytics.
- **Kappa Architecture**: Simplifies by using stream processing for all workloads with replayable logs (e.g., Kafka).

Conclusion

Both batch and stream processing have unique strengths. Batch is robust and simple for offline analytics, while stream processing is essential for real-time applications. Often, modern data architectures blend both to maximize performance and insight delivery.