Real-Time Stream Data Ingestion With Warehouses

HydPy Meetup 21st June

Sourav Roy

About Me

- Data Engineering Dev Team Lead at S&P Global Enterprise Data Organization, specializing in Data Warehouse Solutions Development, REST API, and Microservices.

- With 12 years of experience in Python backend development, my interest and expertise area involves cloud-native and server-less data pipeline solutions, driven by a passion for all things data

Agenda

- Datawarehouse Vs Databases
- Commonly used Datawarehouses
- Generic Architecture of Streaming & Ingestion
- Feature Highlights
- Best Practices

Database - DataWarehouse - DataLake - DataLakeHouse

- **77%** business relies on real time data which includes Change Data Capture(CDC) and Stream data.

- Optimized for fast transactions (**OLTP**); stores current operational data (e.g., MySQL, PostgreSQL).

- Optimized for analytics (**OLAP**); structured, cleaned, and aggregated data for BI (e.g., Snowflake, Redshift).
-
- Stores raw, unstructured/semi-structured data at scale; schema-on-read (e.g., S3, HDFS).
-
- Combines warehouse performance with lake flexibility; supports BI + ML workloads on a unified platform (e.g., Databricks, Snowflake with Iceberg).

snowflake®

databricks

amazon REDSHIFT

BigQuery

Commonly Used Datawarehouses & Streaming Platforms

Stream Services: Kafka, Kinesis, Dataflow, EventHubs

Streaming Engine: Spark Structured Streaming

Data Warehouses: Databricks, Redshift, Snowflake, BigQuery, Data Fabric

Analysis Tools: Power BI, Tableau

Stream services -

- Acts as the source for incoming data streams.

Streaming processing engine -

- Processes data in real-time.

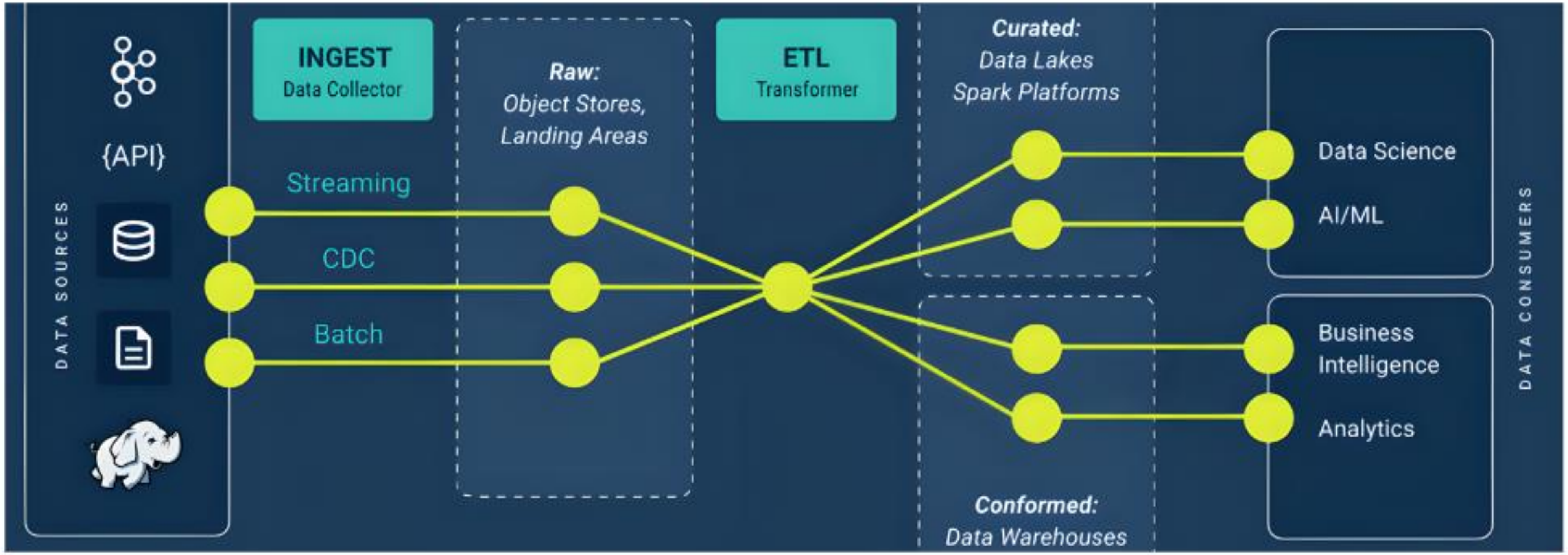
Data warehouses -

- Stores processed data.

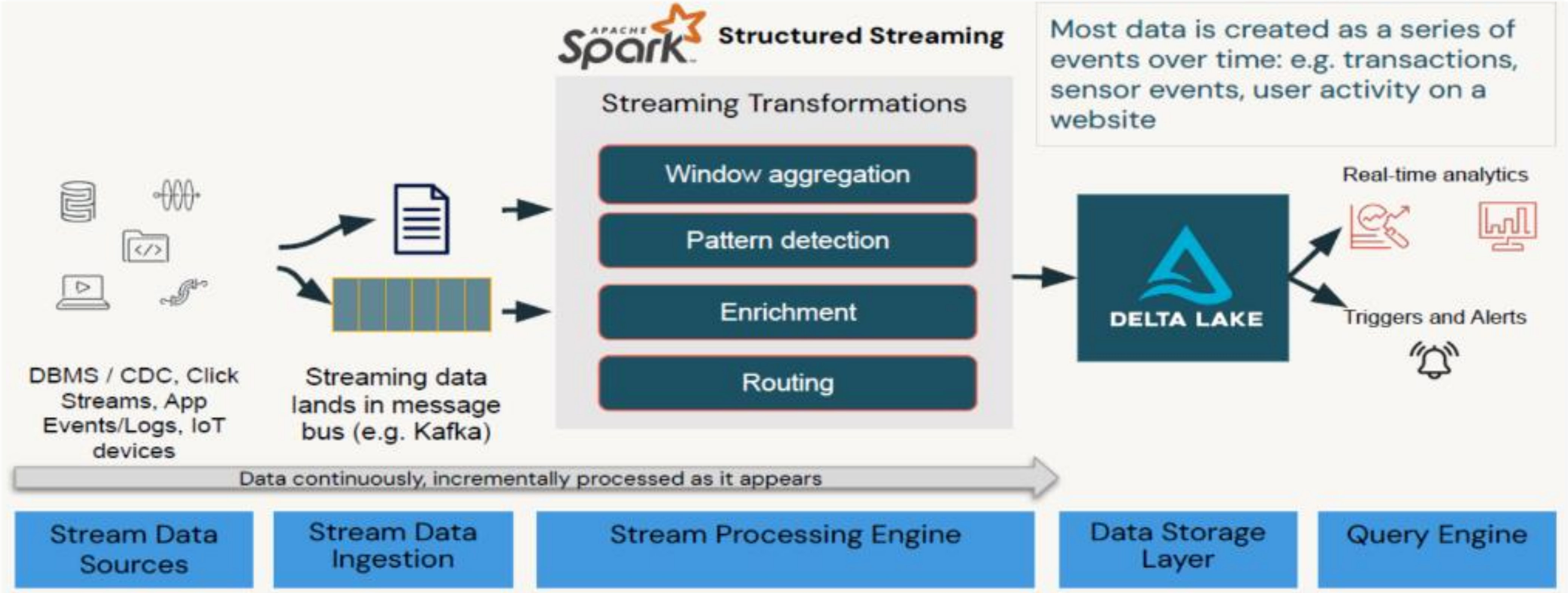
Business Intelligence tools -

- Analysis, forecasting nd decision-making.

Generic Architecture of Streaming DataWarehouse Ingestion



Generic Stream Processing Architecture

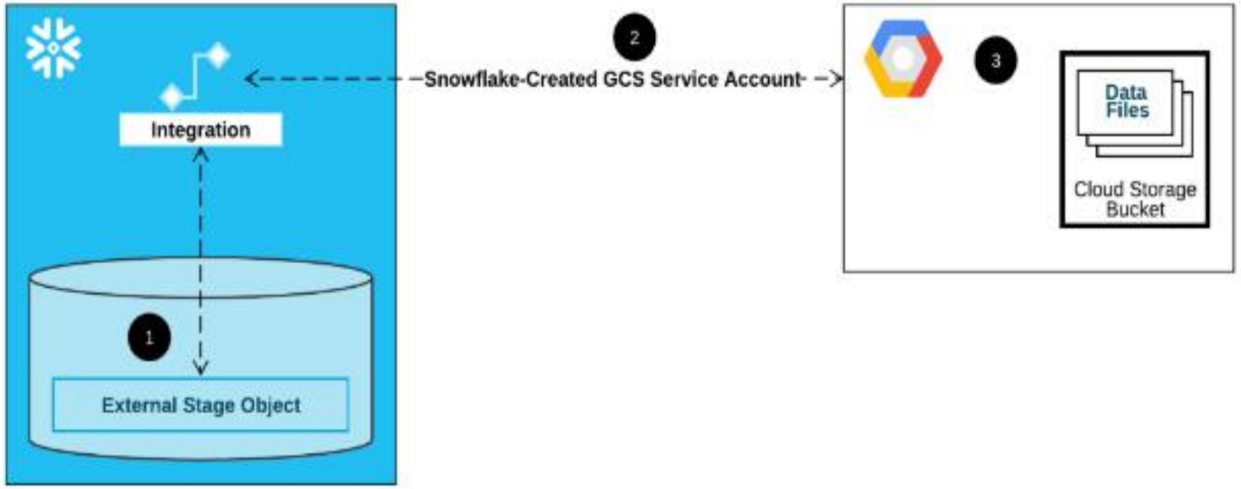


Most data is created as a series of events over time: e.g. transactions, sensor events, user activity on a website

Spark Structured Streaming
Streaming Transformations
Window aggregation
Pattern detection
Enrichment
Routing
DELTA LAKE
Real-time analytics
Triggers and Alerts
DBMS / CDC, Click Streams, App Events/Logs, IoT devices
Streaming data lands in message bus (e.g. Kafka)
Data continuously, incrementally processed as it appears
Stream Data Sources
Stream Data Ingestion
Stream Processing Engine
Data Storage Layer
Query Engine

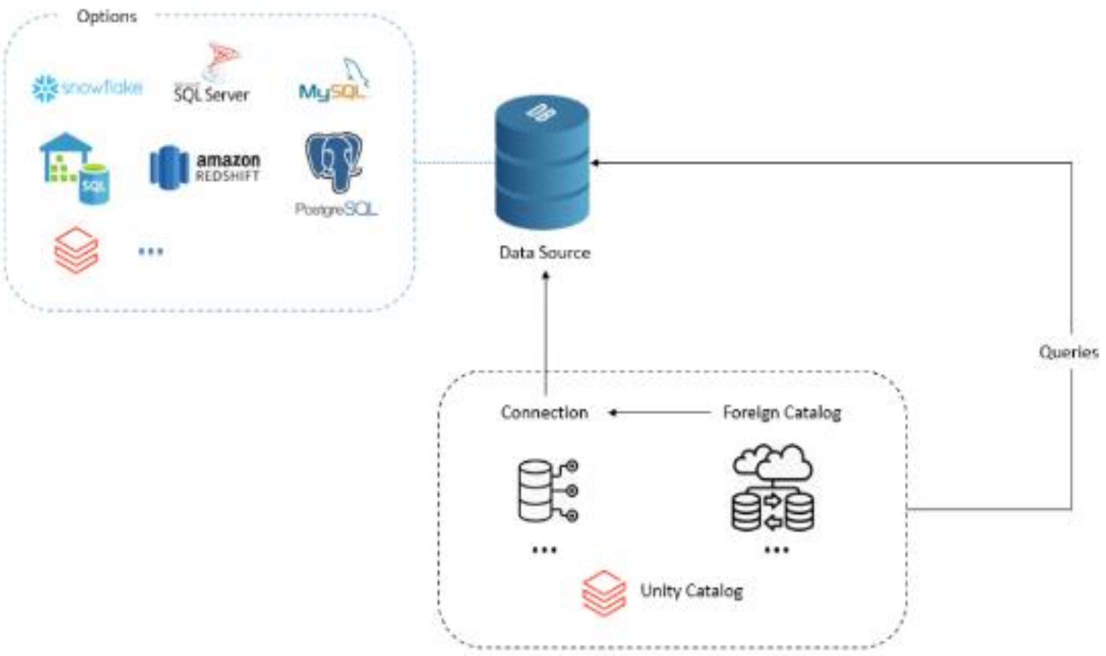
Snowflake Snowpipe

- □ **Auto-ingests** new data files in near real-time.
- □ Detects and loads data with minimal manual effort.
- □ **Scales automatically** for varying data loads.
- □ **Pay only for data loaded**, not continuous compute.
- □ Easy integration and robust load monitoring.



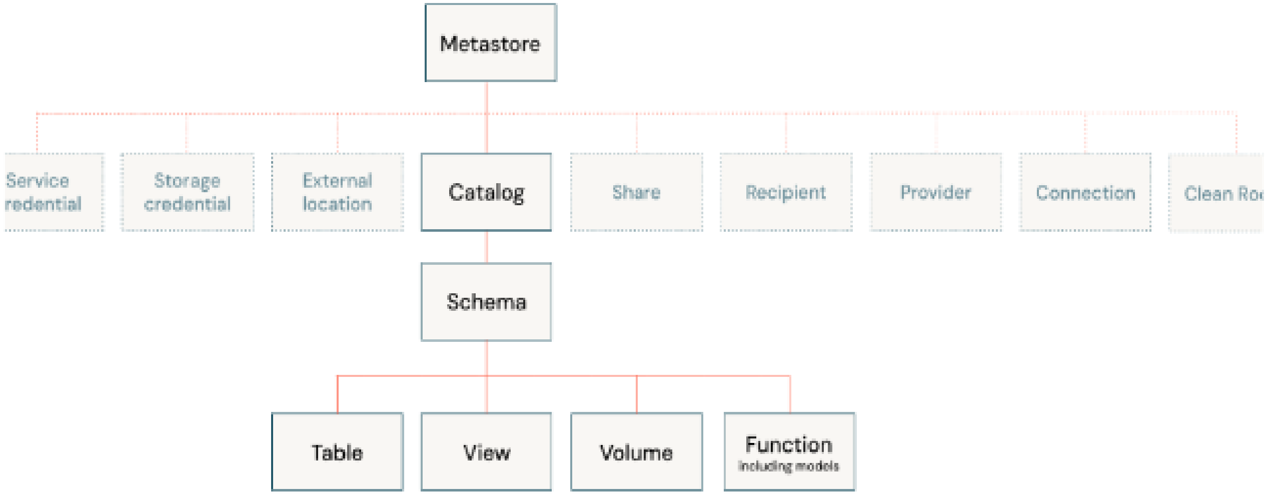
Query Federation

- □ Query **multiple data sources** from one SQL endpoint.
- □ Analyze external data **without moving** or copying it.
- □ **Unified security and governance** for all data sources
- □ Integrate structured and unstructured data in queries.
- □ Faster insights by reducing data silos in analytics.



Catalogues

- □ **Centralized** access control for all data assets.
- □ **Fine-grained** permissions at table, column, and row.
- □ Automated **data lineage** and audit logging.
- □ Simplifies secure **data sharing** across workspaces.
- □ Consistent governance across clouds and teams.



Miscellaneous Features

- □ **Delta** tables in databricks.
- □ **CDF**(Change-Data-Feed).
- □ **Staging** tables in snowflake.
- □ **Dataflow** in gcp.
- □ Data sharing.
- □ Delta, direct, reader, analyticshub share
- □ Simplex & Multiplex stream
- □ Bronze, Silver, Gold tables

Data Security & Best Practices

□ **PII** data and regulatory compliance.
□ Data audit at regular intervals.
□ **Vaccum**-ing warehouse.
□ Data isolation.
□ Granular access control through ACL
□ Metadata management and data lineage.
□ Data retention policy
□ **Clustering** mechanisms(Z-order, Liquid Clustering).