Flink Training for Real-Time Data Engineering

Mastering Connectors for Your Data Architecture

AGENDA

- The role of connectors in a Flink ecosystem.
- Change Data Capture (CDC): Implementing real-time pipelines from OLTP databases.
    - Connecting to PostgreSQL for real-time ingestion.
- Connecting to your destinations: Writing transformed data to Snowflake and PostgreSQL.
- Exploring the landscape of managed vs. open-source connectors

The Role of Connectors in Apache Flink

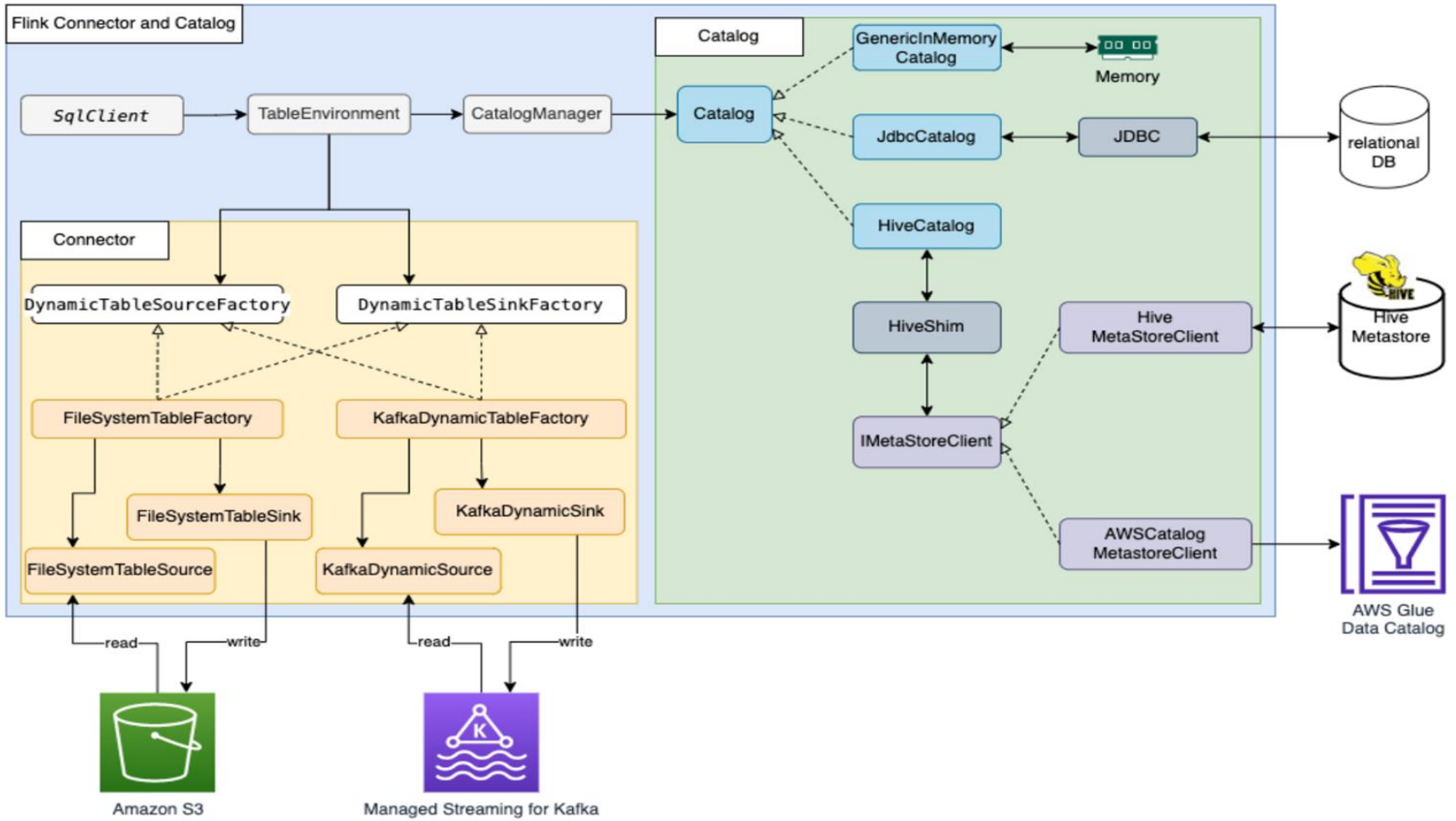
What are Connectors?

- Interfaces that let Flink integrate with external systems.
- Act as **sources** (input) and **sinks** (output) for data.

Why Important?

- Enable real-time & batch processing across diverse data systems.
- Ensure seamless **end-to-end data pipelines.**

Architecture



https://aws.amazon.com/blogs/big-data/build-a-unified-data-lake-with-apache-flink-on-amazon-emr/

Ecosystem of Flink Connectors

- **Databases:** MySQL, PostgreSQL, MongoDB, etc.
- **Message Queues:** Apache Kafka, RabbitMQ, Pulsar.
- **Cloud Storage:** Amazon S3, GCS, Azure Blob.
- **Analytical Stores:** Elasticsearch, ClickHouse, Snowflake, etc.

This broad support makes Flink a **unified data processing engine.**

Applications of Flink Connectors

Streaming Analytics

- Real-time dashboards, monitoring, alerts.

ETL Pipelines

- Ingest → Transform → Load into data lakes/warehouses.

Data Lake Integration

- Manage large-scale batch & streaming data efficiently.

Hybrid Pipelines

- Combine **streaming + batch** for flexible analytics.

Why Real-Time Data Pipelines Matter

Business Agility

Traditional batch ETL creates delays that limit business responsiveness and decision-making speed.

Change Streaming

CDC streams only data changes (inserts, updates, deletes) in real time, reducing system overhead.

Instant Analytics

Enables instant analytics, event-driven applications, and seamless data replication across systems.

Real-world example: Fraud detection systems react instantly to suspicious transactions, preventing losses within milliseconds of occurrence.

What is Change Data Capture (CDC)?

1

Event Streaming

CDC captures and streams database changes as events directly from transaction logs, ensuring complete data lineage.

2

Efficient Processing

Avoids resource-intensive full data snapshots whilst significantly reducing load on source databases.

3

Real-time Operations

Supports insert, update, and delete operations with sub-second latency for immediate downstream processing.

4

Architectural Foundation

CDC serves as the backbone of modern event-driven architectures, enabling reactive systems at scale.

Apache Flink CDC: The Streaming Powerhouse

Unified Processing

Flink CDC integrates seamlessly with Apache Flink's stream processing engine, supporting both streaming and batch workloads.

Fault Tolerance

Built-in fault tolerance with exactly-once processing guarantees ensures complete data integrity across pipeline failures.

Schema Evolution

Automatically captures schema and data changes from multiple database sources without manual intervention.

Architecture Overview

Source Layer

PostgreSQL or MongoDB CDC connectors capture change events from transaction logs and change streams.

Processing Layer

Apache Flink processes, transforms, and enriches data in motion with complex event processing capabilities.

Sink Layer

Data flows to downstream systems including data warehouses, caches, and real-time dashboards.

Note: Kafka often serves as a durable message broker between source systems and Flink for enhanced reliability and scalability.

Handling Schema Evolution & Fault Tolerance

Automatic Schema Detection

Flink CDC automatically detects and propagates schema changes without manual intervention or pipeline restarts.

Zero Downtime

Supports adding or removing tables and columns without pipeline downtime, ensuring continuous data flow.

Exactly-Once Semantics

Guarantees no data loss or duplication during failures with built-in checkpointing and recovery mechanisms.

Real-World Impact & Use Cases

Global E-commerce

Synchronising inventory and orders across multiple regions, enabling real-time stock updates and consistent customer experiences worldwide.

Financial Services

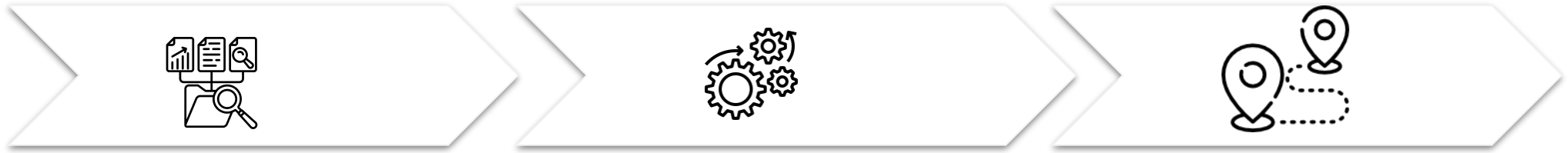
Real-time fraud detection and compliance monitoring, processing millions of transactions with sub-second response times.

SaaS Platforms

Powering live dashboards and user activity tracking, providing instant insights into customer behaviour and system performance.

> **Success Story:** PostgreSQL + Flink CDC pipeline reduced data latency from hours to seconds, enabling real-time decision making for a Fortune 500 retailer.

Data Flow Architecture



Data Sources

Streaming and batch data ingestion from various systems

Apache Flink

Real-time processing engine for transformation and enrichment

Destinations

Snowflake for analytics, PostgreSQL for operational workloads

Flink serves as the central processing engine, enabling real-time data transformation before writing to your chosen destinations.

Why These Destinations Matter

Snowflake

- Cloud-native data warehouse
- Elastic scaling capabilities
- Optimised for analytical workloads
- SQL-based query interface

PostgreSQL

- Robust transactional database
- ACID compliance
- Operational dashboards
- Real-time reporting layer

Connector Comparison

Feature	Managed Connectors	Open-Source Connectors
Ease of Use	Pre-built, minimal configuration	Requires setup & tuning
Support	Enterprise SLA	Community-driven
Flexibility	Limited customisation	Highly customisable
Cost	Usually paid	Free (infrastructure cost only)
Examples	Confluent Cloud, Snowflake Native	Flink CDC, Debezium

Key Takeaways

Destination Strategy

Choose Snowflake for analytical workloads and PostgreSQL for operational systems based on your specific use case requirements.

Connector Selection

Evaluate managed versus open-source connectors based on your team's expertise, budget constraints, and customisation needs.

Hybrid Approach

Consider combining managed and open-source solutions to optimise both cost and functionality across your data pipeline architecture.

ANY SUGGESTIONS?

DATACOUCH
CREATING INFINITE POSSIBILITIES

DSAI
Data Science and AI Academy

Piramal Finance


THANK YOU