

A”

Aalto University
School of Science

Streaming Data Processing with Apache Flink^{1 2}

Hong-Tri Nguyen

Department of Computer Science

hong-tri.nguyen@aalto.fi

CS-E4640 Big Data Platforms, Spring 2025
March 20, 2025

¹<https://flink.apache.org/>

²<https://github.com/rdsea/bigdataplatforms/blob/master/tutorials/streamingwithflink/>

Outline

Flink

Dirty your hand

Peparation

At home

Slide streaming process

Download and run

- ▶ Flink -v 1.19.2 (binary file)
- ▶ RabbitMQ (docker image) – not work for Flink 1.20
- ▶ Kafka -v 3.9.0 (binary file)
- ▶ BTS (csv dataset)

At class

Dirty hand with Flink

- ▶ Flink setting
- ▶ An example usecase with a dataset

Introduce the concept

Apache Flink is a framework and distributed processing engine for stateful computations over unbounded and bounded data streams.

- ▶ Stream Processing with Unbounded Streams, Bounded Streams
- ▶ Stateful Computations: it can maintain and update state information as events are processed.
- ▶ Exactly-Once Semantics: in the event of a failure, it guarantees state is neither duplicated nor lost, providing strong reliability guarantees.

Why need this tools³

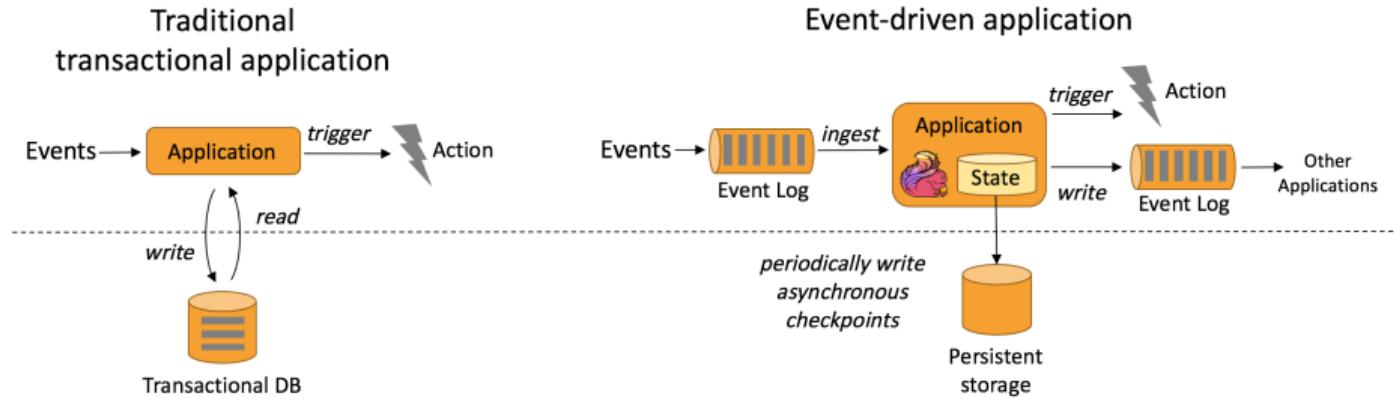


Figure: Event-driven applications

Why need this tools³

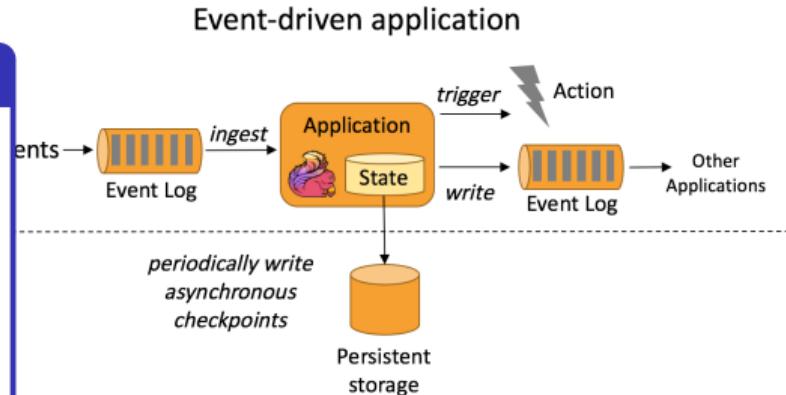
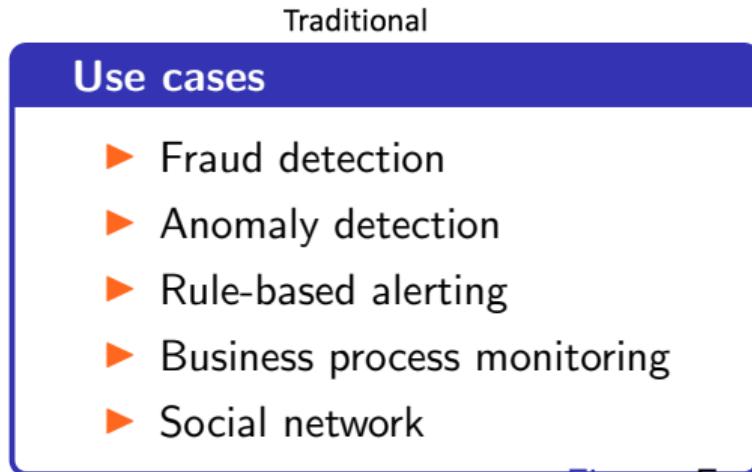
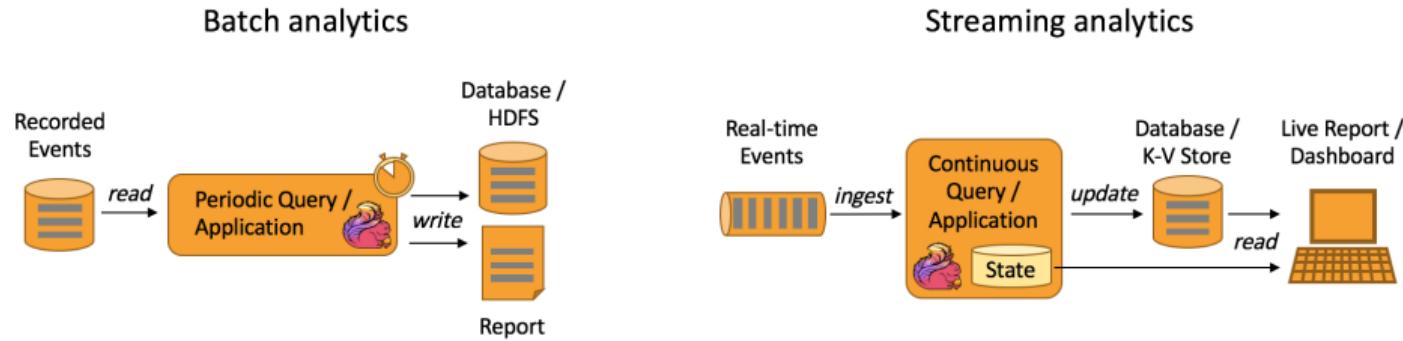


Figure: Event-driven applications

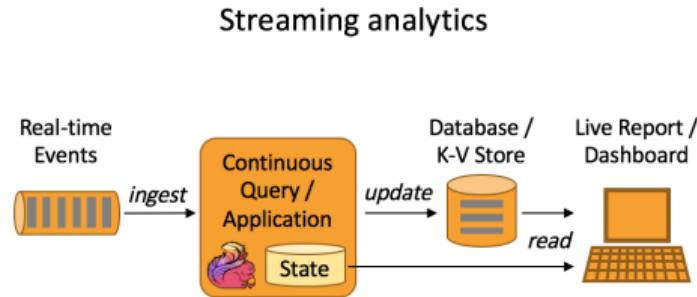
Why need this tools⁴



Why need this tools⁴

Use cases for data analytic

- ▶ Quality monitoring of Telco networks
- ▶ Analysis of product updates & experiment evaluation in mobile applications
- ▶ Ad-hoc analysis of live data in consumer technology
- ▶ Large-scale graph analysis



Why need this tools⁵

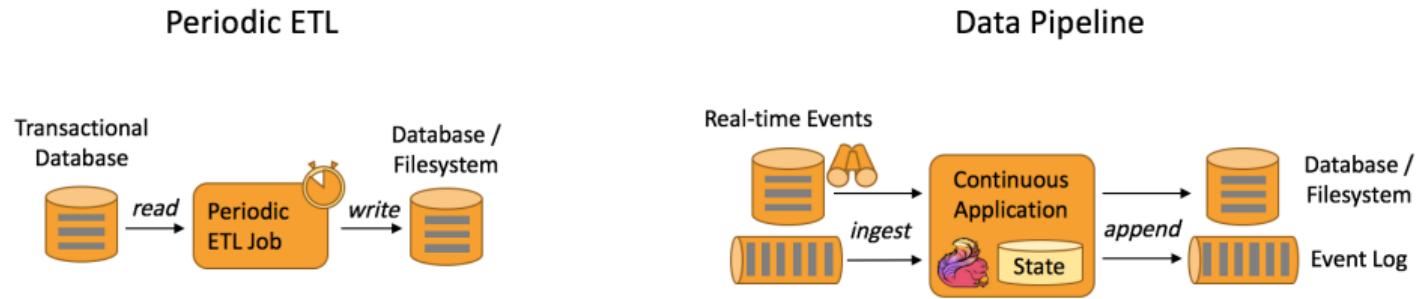


Figure: Data pipeline applications

Why need this tools⁵

Use cases for data pipeline

- ▶ Real-time search index building in e-commerce
- ▶ Continuous ETL in e-commerce

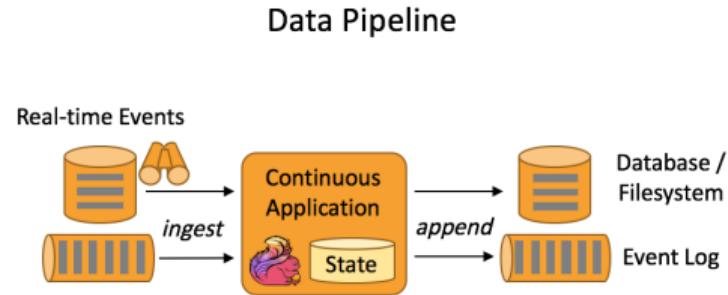


Figure: Data pipeline applications

Errors

- ▶ Dependencies of versions
- ▶ Kafka⁶

⁶ <https://nightlies.apache.org/flink/flink-docs-release-1.20/docs/connectors/stream/kafka/>

Ready for Dirty Hand?



Basic cases

1. Setup Flink for practices
2. Play around with a csv dataset (BST dataset)

1. Setup Flink for practices

Download and run

Down flink v 1.19.2

start ./bin/start-cluster.sh

stop ./bin/stop-cluster.sh

job ./bin/flink run

examples/streaming/WordCount.jar

Kafka producer and consumer

Producer topic and address

Consumer topic and address

2. Play around with a CSV dataset (BST dataset)

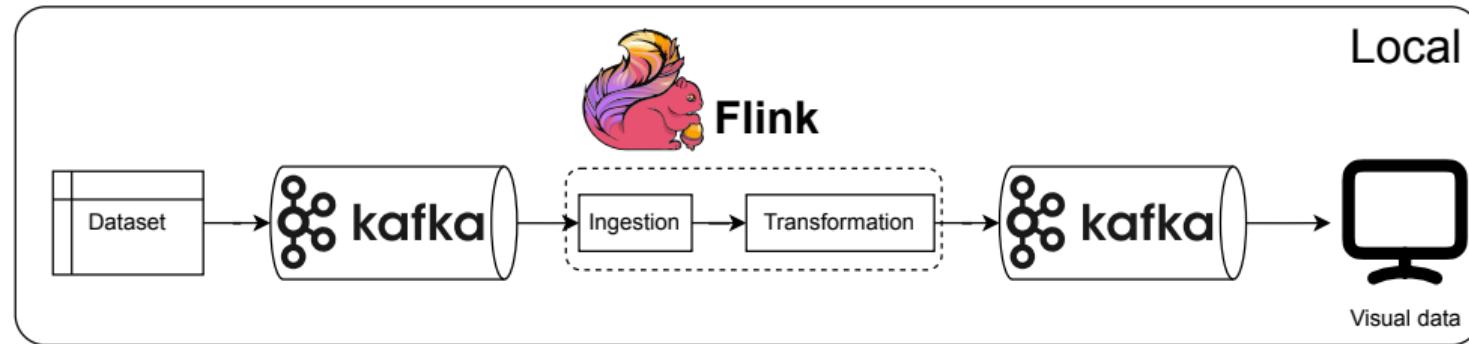


Figure: All local setting

3. Play around with a CSV dataset (BST dataset)

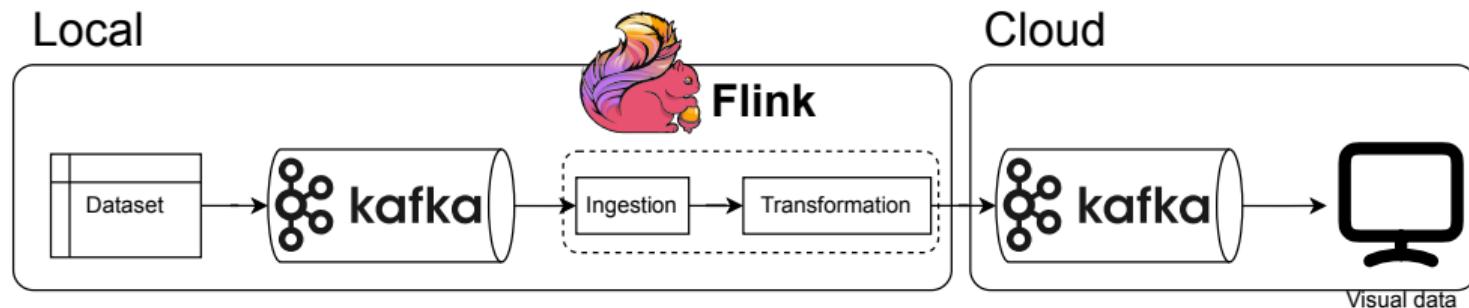


Figure: Work with a Kafka-based messageQ in cloud

4. Play around with a CSV dataset (BST dataset)

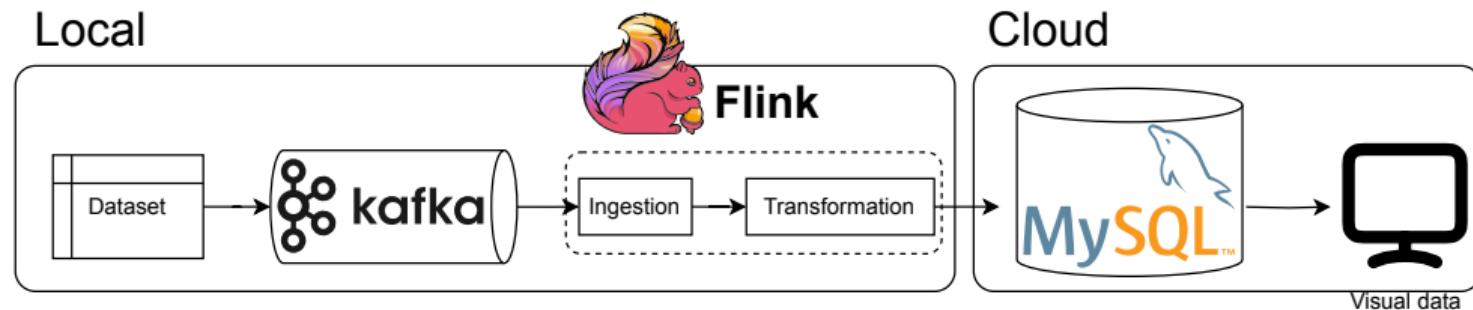


Figure: Work with a mySQL in cloud