Getting started with Apache Flink streaming API

Preetdeep Kumar

18th Jan, 2020

https://www.linkedin.com/in/preetdeep-kumar/

https://github.com/preetdeepkumar/flink-tutorials

https://www.meetup.com/Hyderabad-Apache-Flink-Meetup-Group/

Agenda

- Streaming
  - Introduction
  - Architecture
- Flink
  - Design
  - Typical DataStreaming API workflow
- Demo

Streaming – high level summary

- Streaming refers to data that is **continuously generated**, usually at **high velocity** and in **small sizes (KBs)**.
- Common examples of streaming data include:
  - IoT Sensor events
  - Server logs
  - Click-stream data from apps and websites
  - GPS co-ordinates from a ride
  - Social media

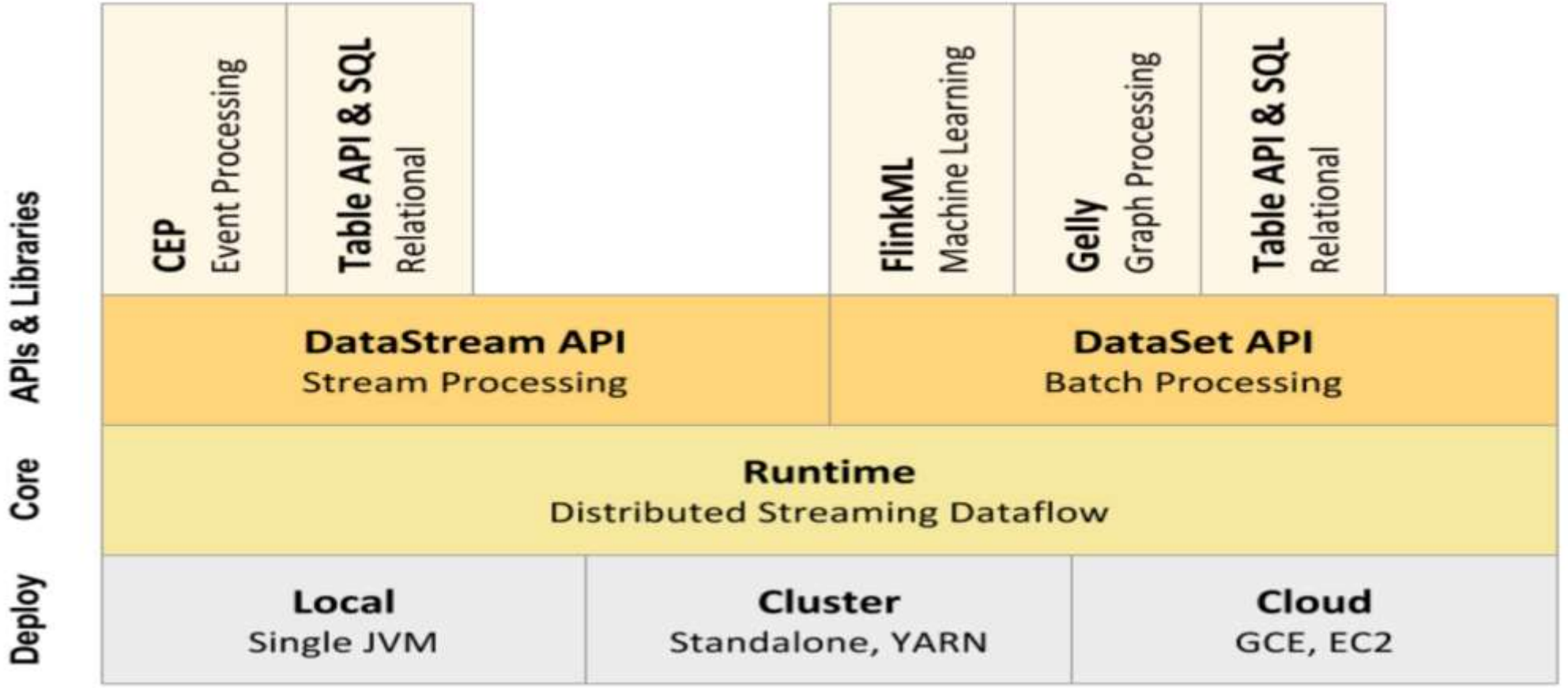
	Batch data processing	Stream data processing
Data scope	Queries or processing over all or most of the data in the dataset	Queries or processing over data within a rolling time window, or on just the most recent data record
Data size	Large batches of data	Individual records or micro batches consisting of a few records
Latency	Minutes to hours	Seconds(near real time) or milliseconds (real time)
Analysis	Complex analytics	Simple response functions, aggregates, and rolling metrics

Source: https://aws.amazon.com/streaming-data/

Typical Streaming data architecture

Data sources	Collection	Ingestion	Process	Storage	Visualize / Analyze
IOT Devices (Sensors) Apps (GPS, Tweets, Clickstreams) Server Logs	Logstash Kinesis Agent YourOwnAgent	Kafka Kinesis Stream	Kstream Kinesis Analytics Flink Spark Streaming	S3 Elasticsearch DB	Kibana Grafana Athena

Flink – High level design



https://ci.apache.org/projects/flink/flink-docs-release-1.9/internals/components.html

Flink’s DataStream typical workflow

1. Create a StreamExecutionEnvironment
2. Add a source which will produce data into Flink
3. Create a DataStream
4. Partition the stream using a key
5. Define a window
6. Provide business logic on the data within a Window
7. Send the result of a window to a source