Apache Beam in the Data Analytics Life Cycle

Sachin Agarwal

Group Product Manager - Google Cloud

http://linkedin.com/in/sachinag

Data industry trends

Happening Now

- Migration to the Cloud
- Massive amounts of (raw) data
- Emergence of new regulations
- Need to reduce time to insights

Emerging Trends

- Data reliability
- Real-time analytics
- Governed data democratization
- AI/ML operationalization

Data analytics & data processing

The data analytics discipline

Data analytics is an overarching practice that encompasses the complete life cycle of insight generation, from collection to quality and access control.

Data processing is a component of the Data Analytics practice. It transforms raw data to power valuable insights and access to the right information.

Data processing is done in three phases:

Ingestion

Processing

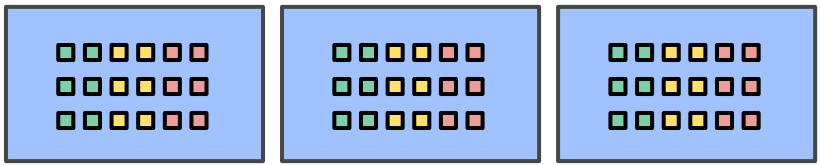
Activation

There are two types of data processing

Batch

Data is collected and processed in chucks. Processing is triggered after a time window has passed or a given amount of data has been collected.

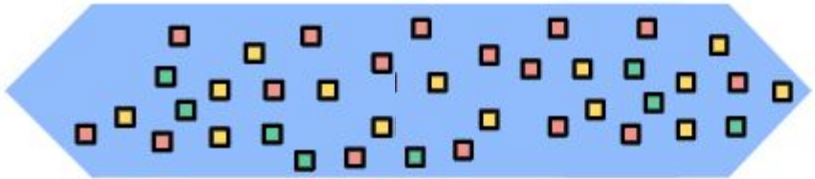
E.g.: payroll systems, preventive manufacturing maintenance, insurance billing, etc.



Streaming (Real-Time)

It is the continuous processing of data that aims to derive insights or new information shortly after each new data point enters a system for the first time.

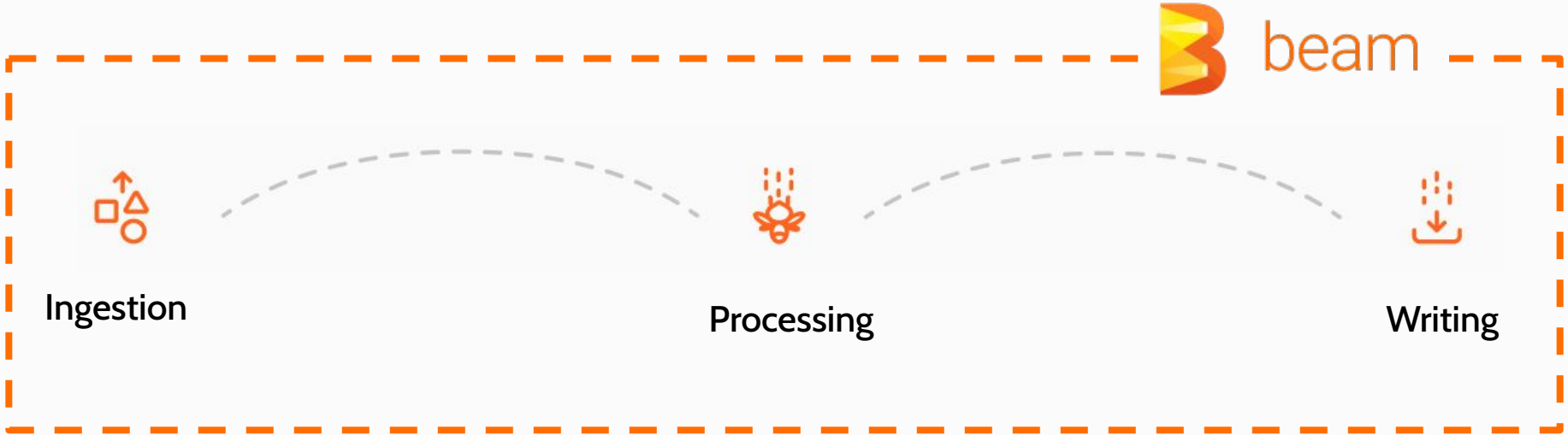
E.g.: experience personalization, anomaly detection, malfunction alerting system, etc.



Where does Apache Beam fit in?

Apache Beam is a unified programming model

Apache Beam is a unified programming model
to build batch and streaming data processing pipelines



Building Apache Beam pipelines requires 3 steps

Step 1. Choose your runner, Apache Beam is portable!

Changing only one line of code, you can run Beam pipelines in any supported Beam runner including Dataflow, Spark, or Flink

Step 2. Choose your favorite language

You can develop Beam pipelines in your **favorite language**:

Java, Scala, Python, Go, and SQL (Typescript too!)

Step 3. Use I/O connectors and transforms to implement your biz logic

Input Connectors

- File systems
- Messaging services
- Databases
- Caches
- Others

Built-in Transforms

- GroupByKey
- CoGroupByKey
- Combine
- Flatten
- Partition
- ...

Activation / Output Connectors

- File systems
- Messaging services
- Databases
- Caches
- Others

Recap

- ✅ Beam is a unified model to build **batch and streaming** data pipelines
- ✅ **Change a single line of code** to port your Beam pipelines to any runner
- ✅ Code in your **favorite language**
- ✅ A large collection of built-in input/output connectors and transforms
- ✅ It's easy to build your own connectors and transforms!

In today's module

- Apache Beam Overview
- Real-time semantic enrichment and online clustering of text content
- Workshop: My First Batch Pipeline with Apache Beam

Thank You!

BEAM COLLEGE