Apache Beam in the Data Analytics Life Cycle

Griselda Cuevas

Product Manager - Google Cloud

http://linkedin.com/in/griscz

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

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 into valuable insights and information.

The data analytics practice

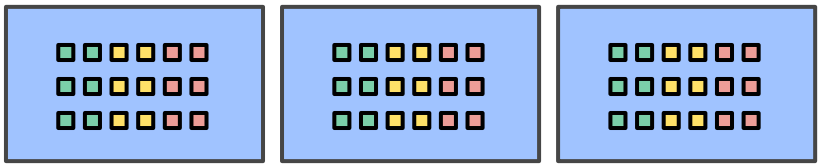
Data processing is done in three phases:

There are two types of data processing

Batch

Data is collected and processed in chunks.
It is Used for large amounts of data.

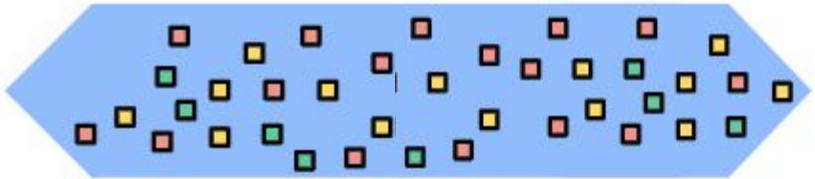
E.g.: payroll systems, preventive manufacturing maintenance, insurance billing, etc.



Streaming (Real-Time)

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

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



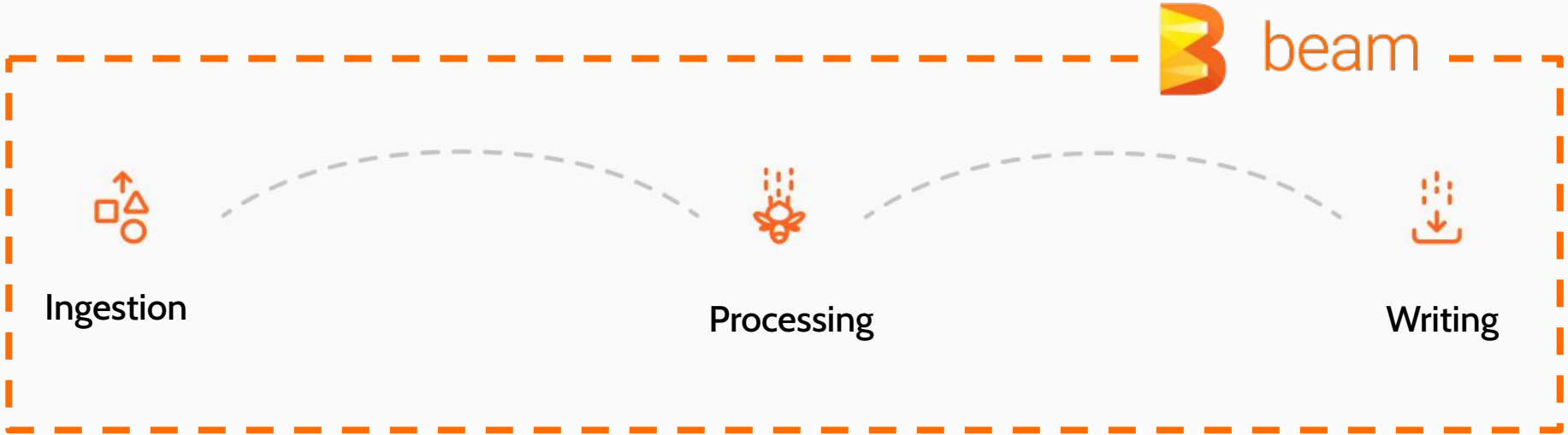
Where does Apache Beam fits in?

A common misconception...

Apache Beam is a substitu pache Spark or Apache Flink

Truth is...

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



Building Apache Beam pipelines in 3 steps

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

You can run Apache Beam pipelines in any supported runner

including Apache Spark, Apache Flink and Dataflow

Step 2. Choose your favorite language

You can develop Apache Beam pipelines in your language of choice: Java, Python, SQL and Go

Step 3. Use I/O connectors and transforms to solve your use case

Input Connectors

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

Operations

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

Output Connectors

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

Recap

✅ Beam is a **unified model** to build batch and streaming data pipelines

✅ Beam pipelines are portable and can run in different runners changing only a single line of code

✅ You can code in your favorite programming language

✅ A large collection of IO connectors and operators is available

✅ It's easy to build your own connectors and operators!

In today's module

- Apache Beam in action
- Apache Beam Overview
- Defining a directed acyclic graph
- Runner specific overview: Architecture, management and autotuning
- Putting it all together with a Python demo

Thank You!

BEAM COLLEGE