Getting started with Apache Beam

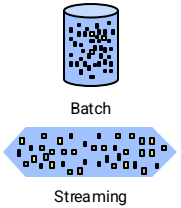
Sascha Kerbler

What is Apache Beam?

Apache Beam is a **unified model** for defining both **batch and streaming data-parallel processing pipelines**, as well as a set of language-specific SDKs for constructing pipelines and Runners for executing them on distributed processing backends

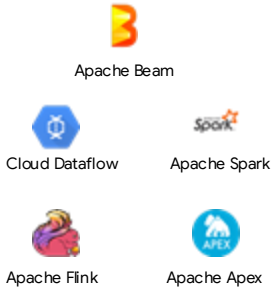
What is Apache Beam?

One Model



Multiple Modes

Multiple SDKs

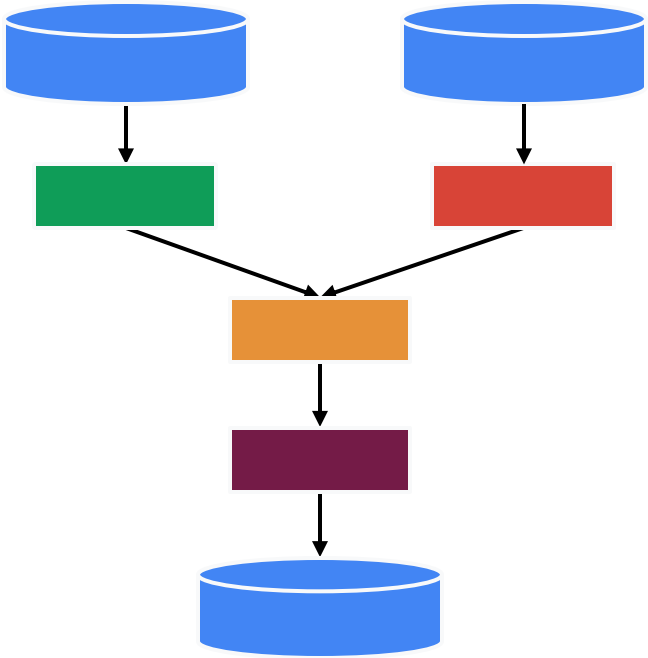


Multiple Runners

What is a pipeline?

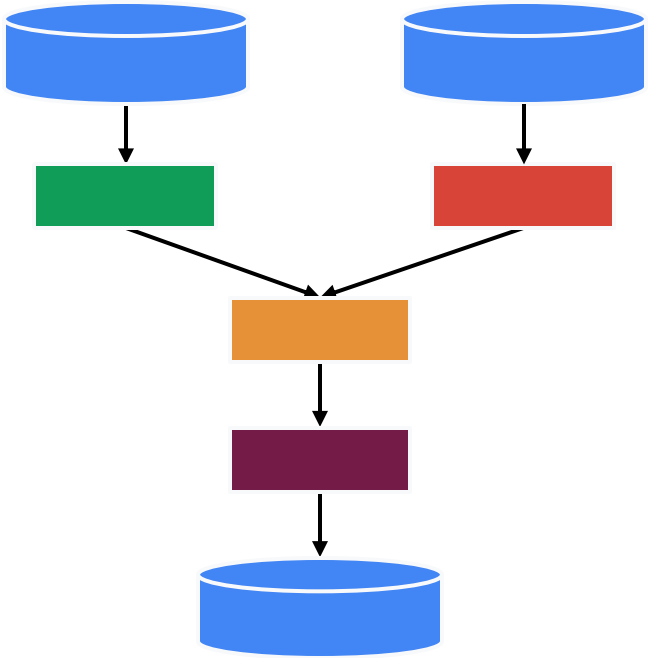
A Directed Acyclic Graph of data **transformations** applied to one or more **collections** of data

- May include multiple sources and multiple *sinks*
- Optimized and executed as a unit

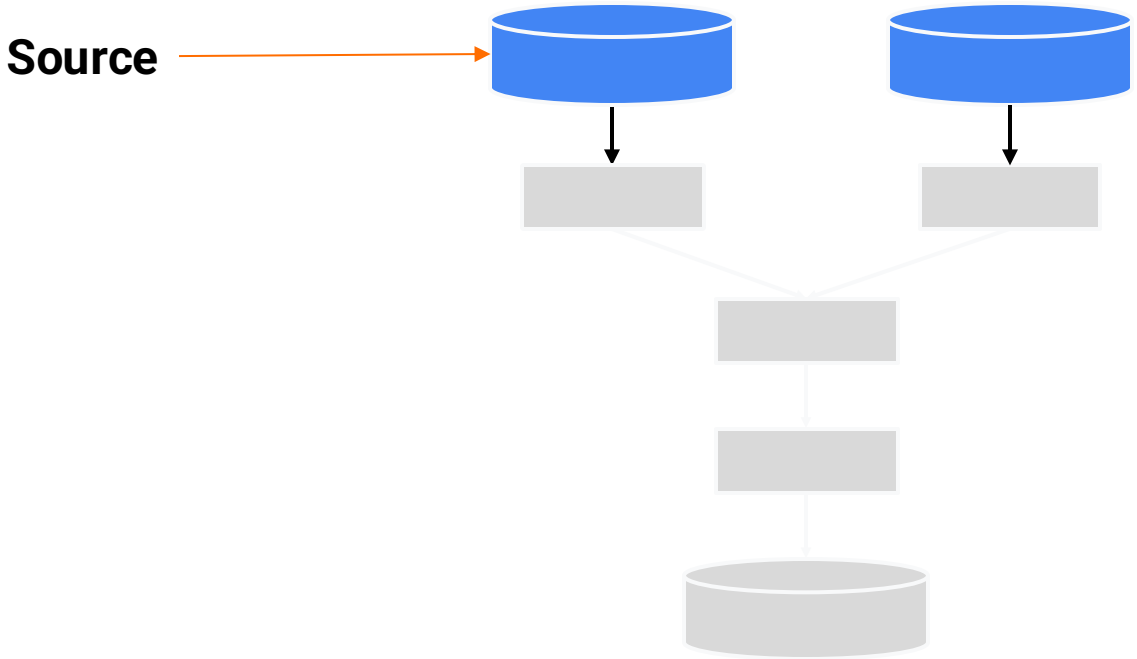


What is a pipeline?

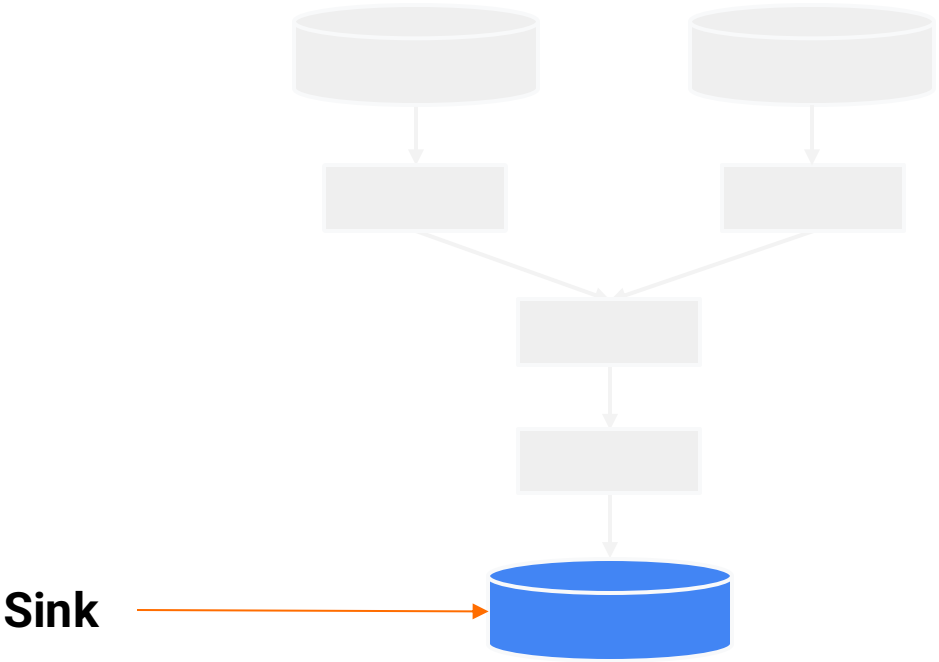
Beam represents datasets using an abstraction called **PCollection**



Input Data



Output Data



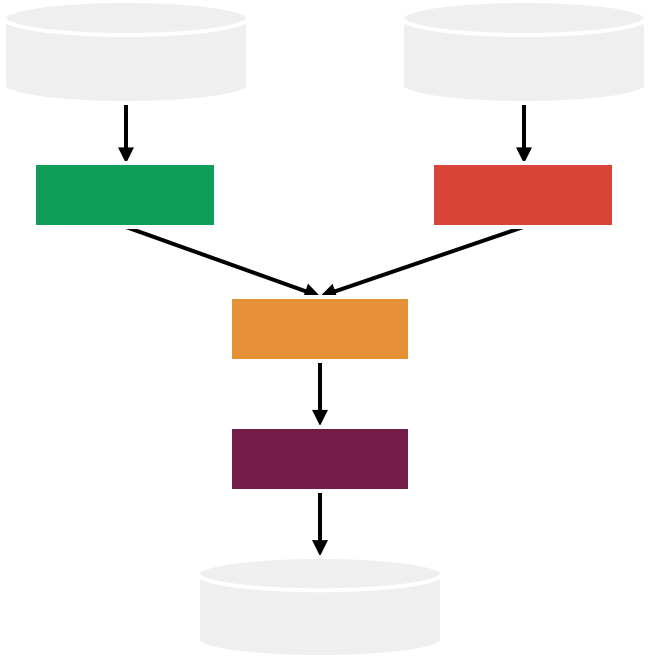
I/O Connectors

Name	Description	Javadoc
FileIO	General-purpose transforms for working with files: listing files (matching), reading and writing.	org.apache.beam.sdk.io.FileIO
AvroIO	PTransforms for reading from and writing to Avro files.	org.apache.beam.sdk.io.AvroIO
TextIO	PTransforms for reading and writing text files.	org.apache.beam.sdk.io.TextIO
TFRecordIO	PTransforms for reading and writing TensorFlow TFRecord files.	org.apache.beam.sdk.io.TFRecordIO
XmlIO	Transforms for reading and writing XML files using JAXB mappers.	org.apache.beam.sdk.io.xml.XmlIO
TikaIO	Transforms for parsing arbitrary files using Apache Tika.	org.apache.beam.sdk.io.tika.TikaIO
ParquetIO (guide)	IO for reading from and writing to Parquet files.	org.apache.beam.sdk.io.parquet.ParquetIO
ThriftIO	PTransforms for reading and writing files containing Thrift-encoded data.	org.apache.beam.sdk.io.thrift.ThriftIO

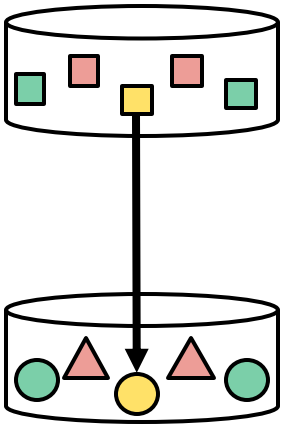
Full List of Connectors: https://beam.apache.org/documentation/io/connectors/

What is a pipeline?

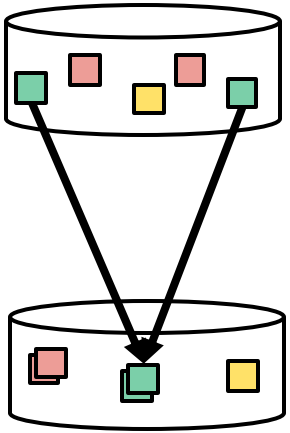
Data transformations are represented by an abstraction called **PTransform**



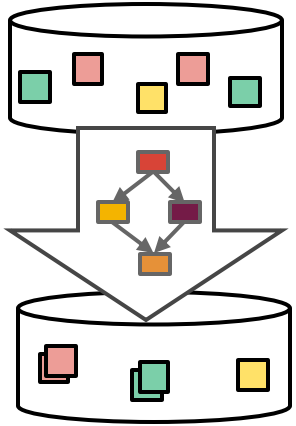
Transform Types



Element-Wise (map)

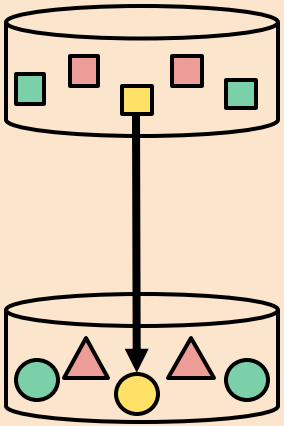


Aggregating (reduce)

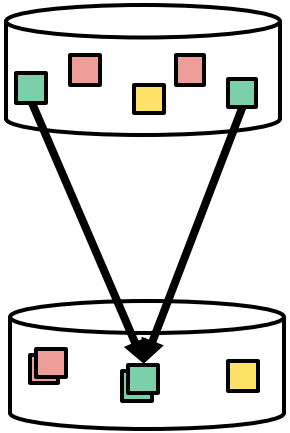


Composite (reusable combinations)

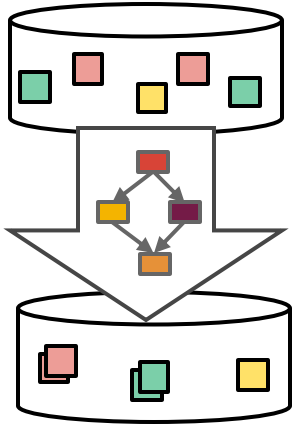
Transform Types



Element-Wise (map)



Aggregating (reduce)

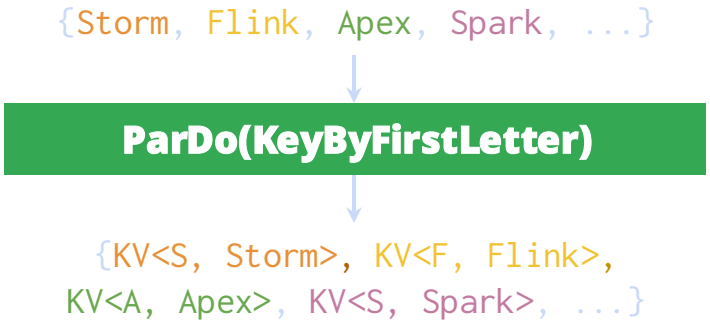


Composite (reusable combinations)

Element-Wise Transforms

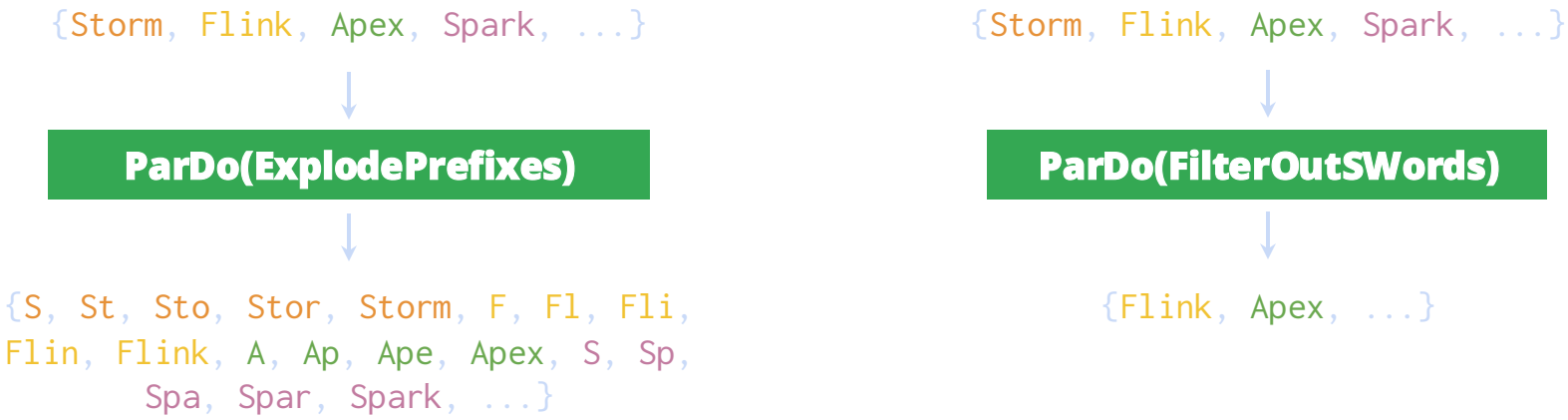
(ParDo = “Parallel Do”)

Performs a user-provided transformation on each element of a PCollection independently



Element-Wise Transforms

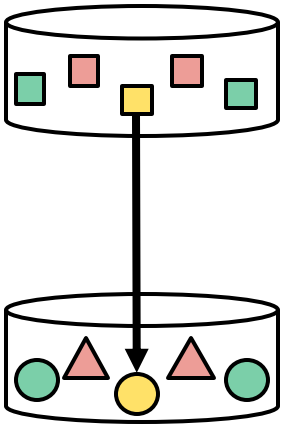
Can output 1, 0 or many values for each input element



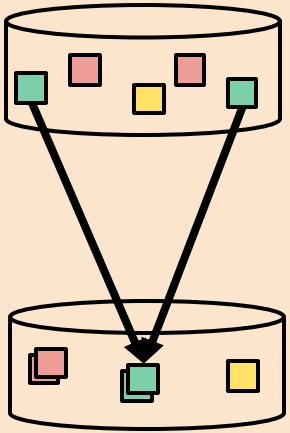
Element-Wise Transforms

ParDo	1-input to (0,1,many)-outputs
Filter	1-input to (0 or 1)-outputs
MapElements	1-input to 1-output
FlatMapElements	1-input to (0,1,many)-output
WithKey	value -> KV(f(value), value)
Keys	KV Pair -> Keys
Values	KV Pair -> Values

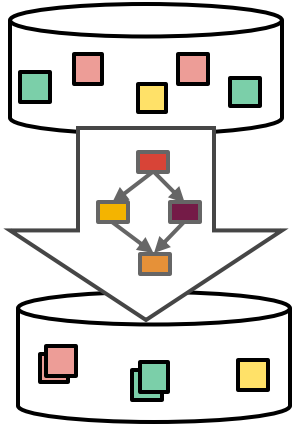
Transform Types



Element-Wise (map)



Aggregating (reduce)



Composite (reusable combinations)

Aggregations

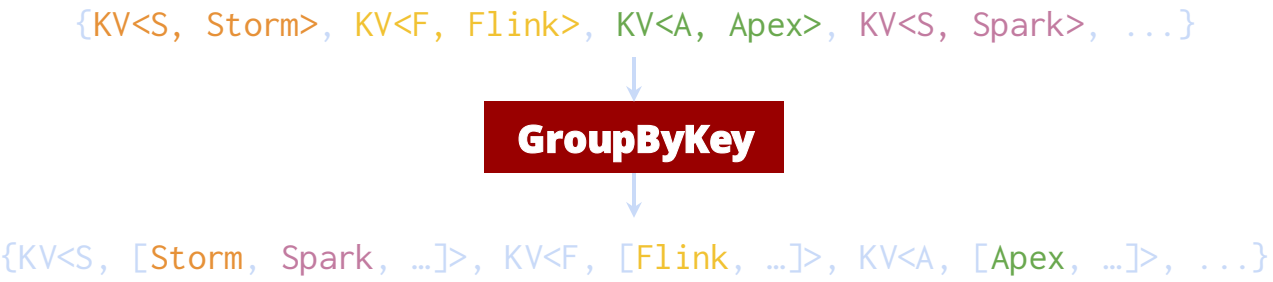
Count: computes the count of all elements in the aggregation

Max: computes the maximum element in the aggregation

Sum: computes the sum of all elements in the aggregation

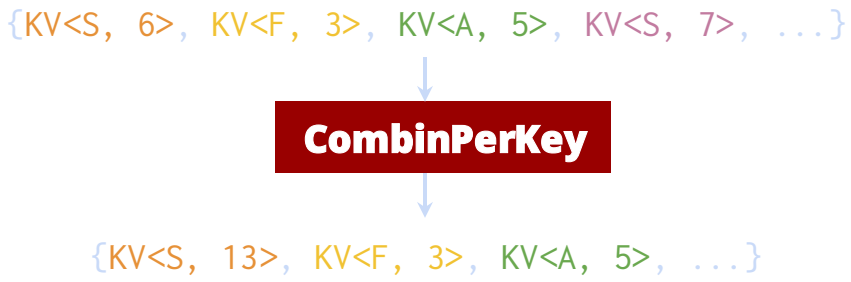
Aggregations

GroupByKey: Takes a PCollection of **key-value pairs** and **groups** all values with the **same key**

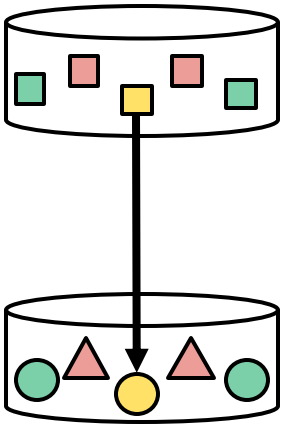


Aggregations

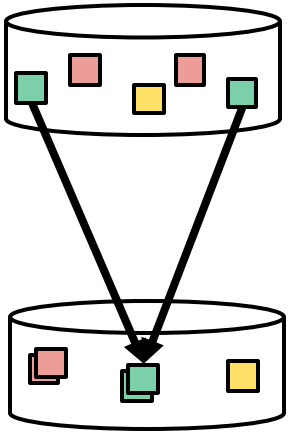
CombinePerKey: CombinePerKey is a type of aggregation that applies a CombineFn (such as summation) to elements with the same key, resulting in a significantly smaller output than the input.



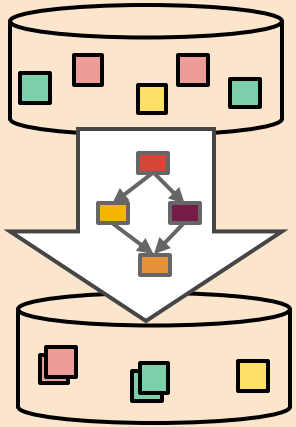
Transform Types



Element-Wise (map)



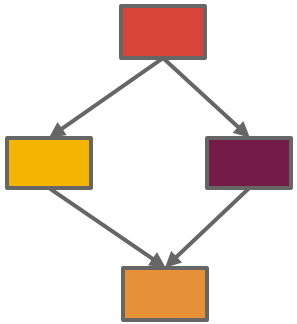
Aggregating (reduce)

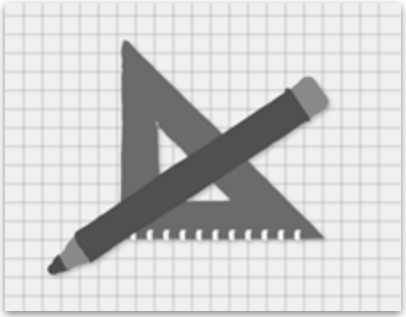


Composite (reusable combinations)

Composite Transform

Composite Transform: A composite transform is a PTransform that combines one or more other PTransforms together to perform a more complex data processing task.





Demo

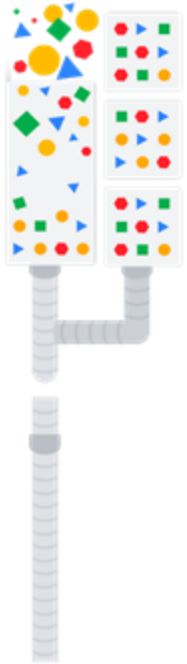
Popular Products

Challenges:

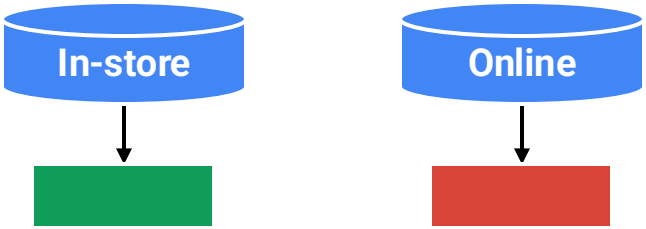
Data is stored in two storage systems

Goal:

Create a pipeline that finds the most popular products

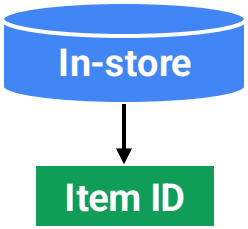


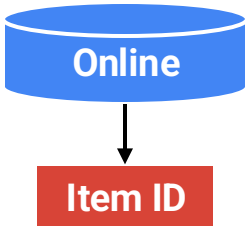
Popular Products



Read in-store & online sales

Popular Products

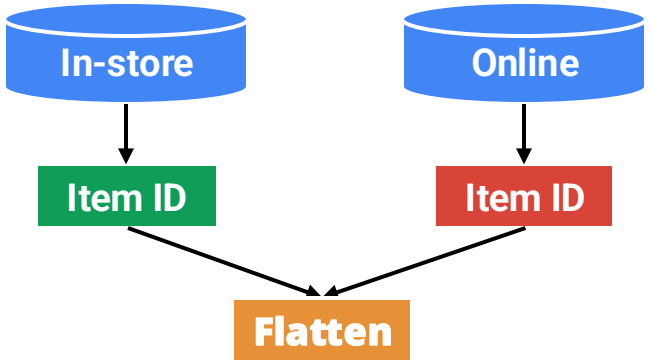




Read in-store & online sales

Extract item ids

Popular Products

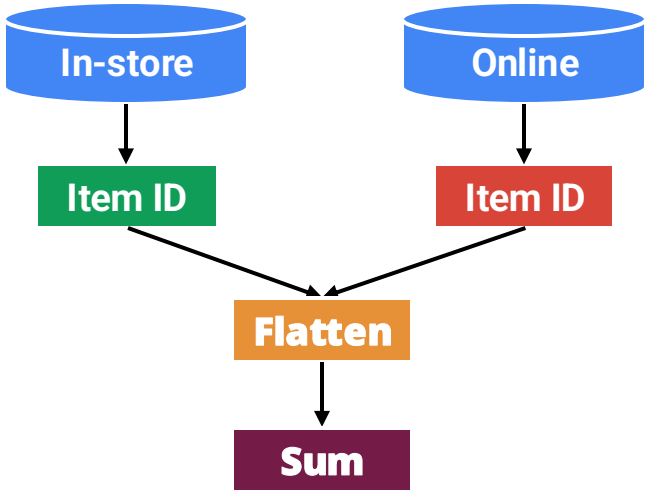


Read in-store & online sales

Extract item ids

Flatten to create a unified view

Popular Products



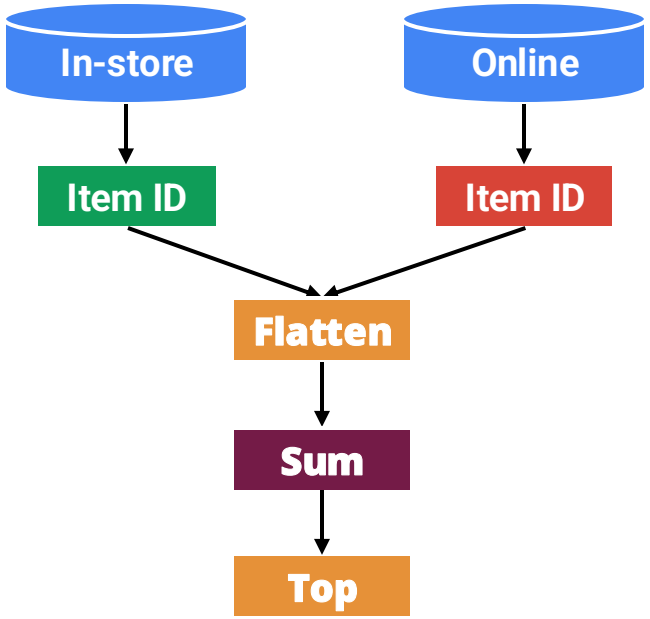
Read in-store & online sales

Extract item ids

Flatten to create a unified view

For each item id, count the # of sales

Popular Products



Read in-store & online sales

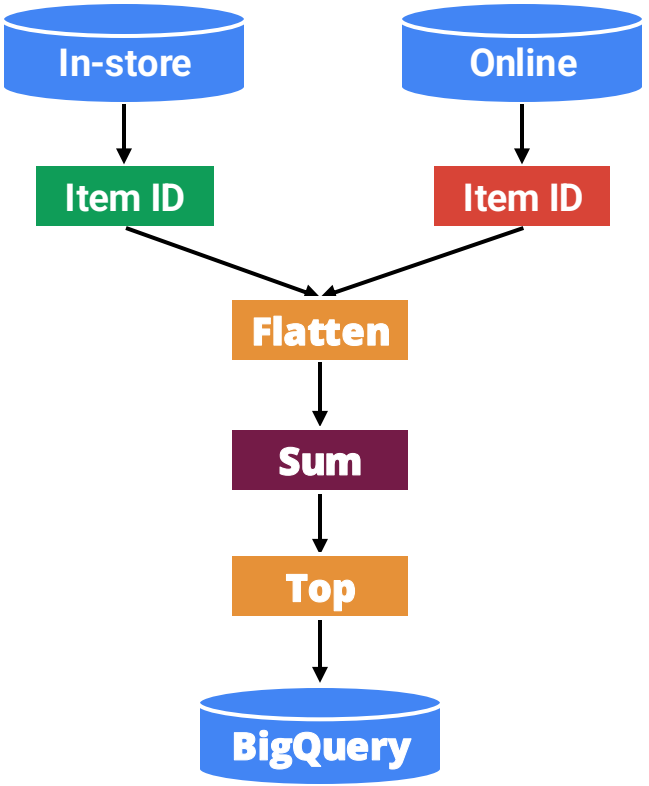
Extract item ids

Flatten to create a unified view

For each item id, count the # of sales

Show most sold products

Popular Products



Read in-store & online sales

Extract item ids

Flatten to create a unified view

For each item id, count the # of sales

Show most sold products

Write the result to BigQuery

Thank you!