Data Engineering

LTAT.02.007

Ass Prof. Riccardo Tommasini

Assistants: Fabiano Spiga, Mohamed Ragab, Hassan Eldeeb

https://courses.cs.ut.ee/2020/dataeng

Forum

Moodle

Data Pipeline

A pipeline is a set of data processing elements connected in series, where the output of one element is the input of the next one.

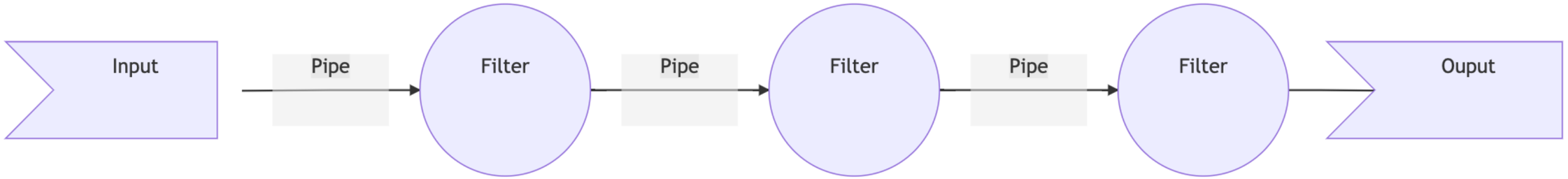
The elements of a pipeline are often executed in parallel or in time-sliced fashion; in that case, some amount of buffer **storage** is often inserted between elements.

Data pipelines Then

The concept of a pipeline began from the good old Unix "Pipe" symbol (|).

The output of one "process" (on left side of the pipe) to be given as "input" to another process (which was on the right side of the pipe).

Elements of a Data Pipeline



Pipes are connectors which send data from one component (filter) to another.

Filters do actual data "processing" (transformation/cleansing/scrubbing/munging... whatever)

Input or **Source** is the actual data source (database output/text file/SQL resultset/raw text)

Output or **Sink** is the final output at the end of this chain.

Big Data Pipeline[1]

The big data world brings velocity, which forced a paradigm shift in data architectures.

Streaming data gave rise to some interesting data pipeline implementations.

[1] The reason as to why Airflow does not support streaming is that there are no obvious behavior rules that could be set so that the airflow scheduler could deterministically check if it has been completed or not.

Basic Operation of Data Pipelines

Basic Operation of Data Pipelines

- store schema and access information of different data sources

Basic Operation of Data Pipelines

- store schema and access information of different data sources
- extract discrete data elements of the source data

Basic Operation of Data Pipelines

- store schema and access information of different data sources
- extract discrete data elements of the source data
- copy extracted data from a data source to another

Basic Operation of Data Pipelines

- store schema and access information of different data sources
- extract discrete data elements of the source data
- copy extracted data from a data source to another
- transform data

Basic Operation of Data Pipelines

- store schema and access information of different data sources
- extract discrete data elements of the source data
- copy extracted data from a data source to another
- transform data
  - correct errors in data elements extracted from source data

Basic Operation of Data Pipelines

- store schema and access information of different data sources
- extract discrete data elements of the source data
- copy extracted data from a data source to another
- transform data
  - correct errors in data elements extracted from source data
  - standardize data in data elements based on field type

Basic Operation of Data Pipelines

- store schema and access information of different data sources
- extract discrete data elements of the source data
- copy extracted data from a data source to another
- transform data
  - correct errors in data elements extracted from source data
  - standardize data in data elements based on field type
- join or merge (in a rule driven way) with other data sources

Big Data Pipeline - Batch

Big Data Pipeline - Batch

- [[MapReduce]]

Big Data Pipeline - Batch

- [[MapReduce]]
- [[FlumeJava]]

Big Data Pipeline - Batch

- [[MapReduce]]
- [[FlumeJava]]
- [[Apache Airflow]]

MapReduce

MapReduce

- A simple abstraction for distributed data processing

MapReduce

- A simple abstraction for distributed data processing
- Invented by Google

MapReduce

- A simple abstraction for distributed data processing
- Invented by Google
- Two operations

MapReduce

- A simple abstraction for distributed data processing
- Invented by Google
- Two operations
  - Map: think of it as a transformation either 1:1 or 1:M that will be applied to each element of your data

MapReduce

- A simple abstraction for distributed data processing
- Invented by Google
- Two operations
  - Map: think of it as a transformation either 1:1 or 1:M that will be applied to each element of your data
  - Reduce: think of it as a form of aggregating or compacting the data M:1

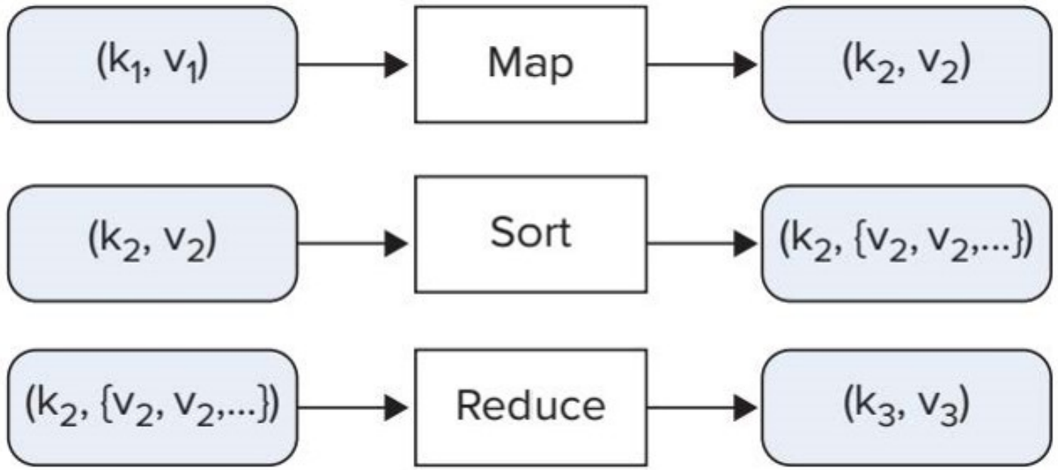
MapReduce

- A simple abstraction for distributed data processing
- Invented by Google
- Two operations
  - Map: think of it as a transformation either 1:1 or 1:M that will be applied to each element of your data
  - Reduce: think of it as a form of aggregating or compacting the data M:1
- Hadoop provides the open source implementation of that model

MapReduce

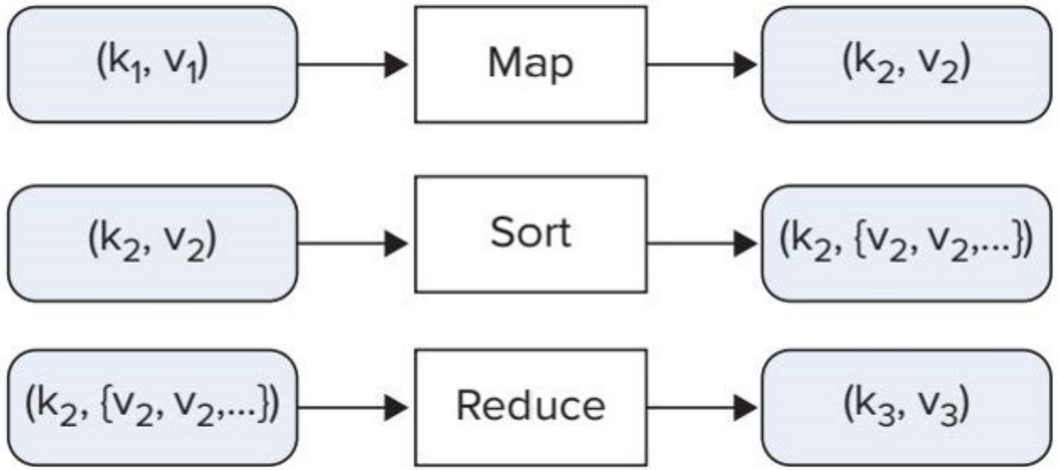
- A simple abstraction for distributed data processing
- Invented by Google
- Two operations
  - Map: think of it as a transformation either 1:1 or 1:M that will be applied to each element of your data
  - Reduce: think of it as a form of aggregating or compacting the data M:1
- Hadoop provides the open source implementation of that model
- It was believed that all sorts of large-scale data processing can be tackled by MapReduce, [this is not true as you will see in future lectures]{style="color: red"}

Overview


(k_1, v_1)
Map
(k_2, v_2)
(k_2, v_2)
Sort
$(k_2, \{v_2, v_2, \ldots\})$
$(k_2, \{v_2, v_2, \ldots\})$
Reduce
(k_3, v_3)

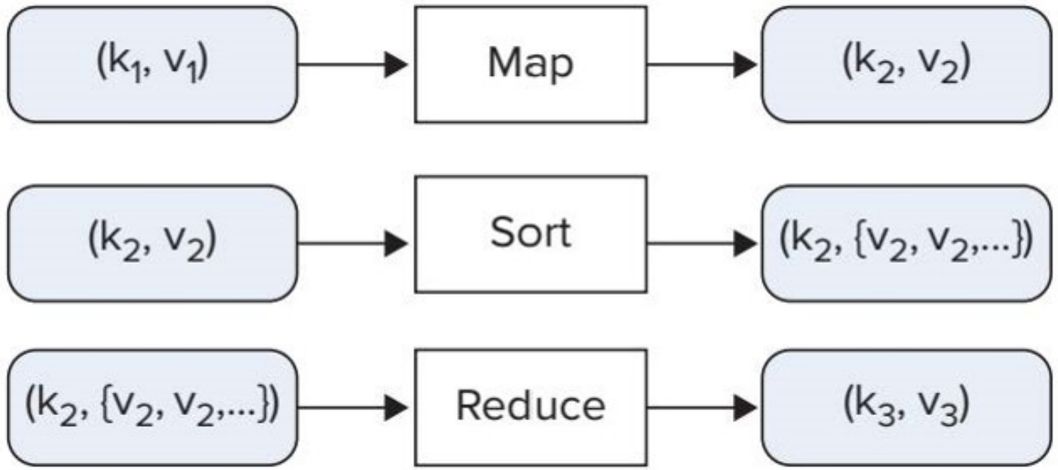
Overview

- You need to implement map, reduce or both



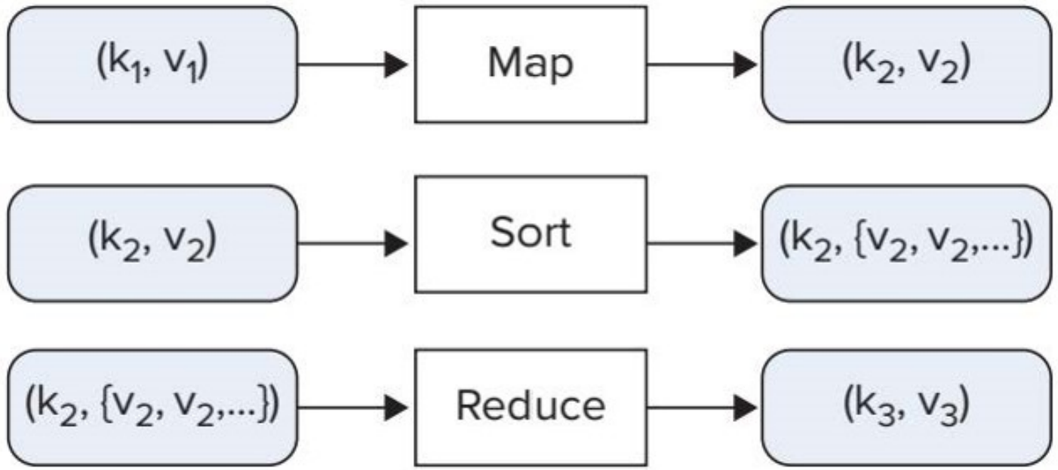
Overview

- You need to implement map, reduce or both
- Map



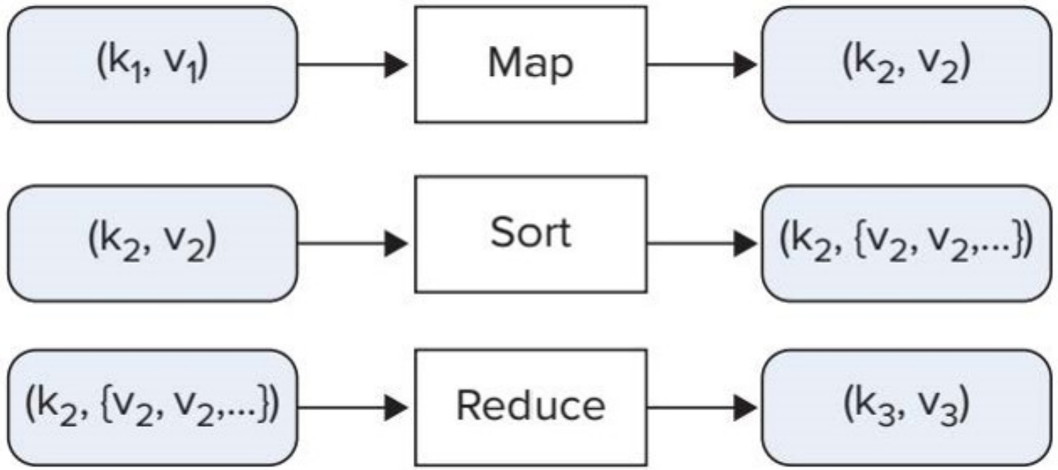
Overview

- You need to implement map, reduce or both
- Map
  - Can be used to split elements,



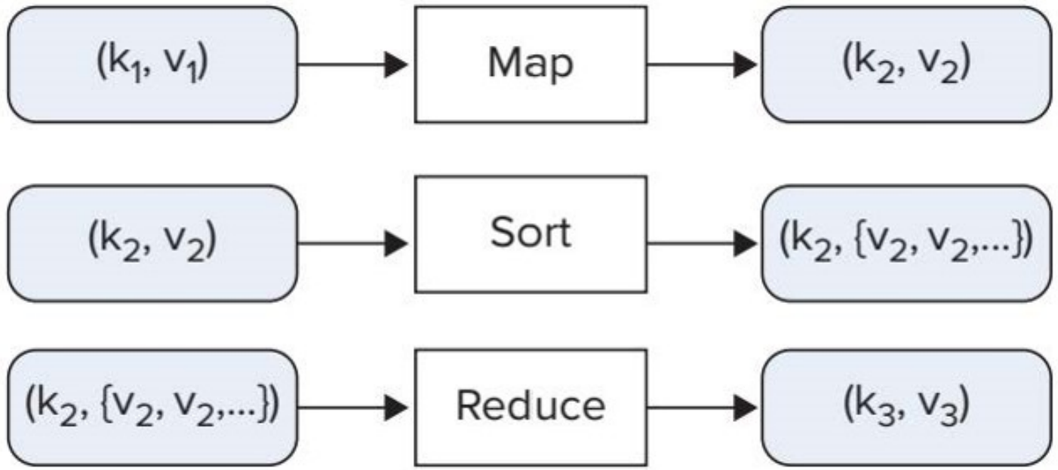
Overview

- You need to implement map, reduce or both
- Map
  - Can be used to split elements,
  - Can be used to filter elements



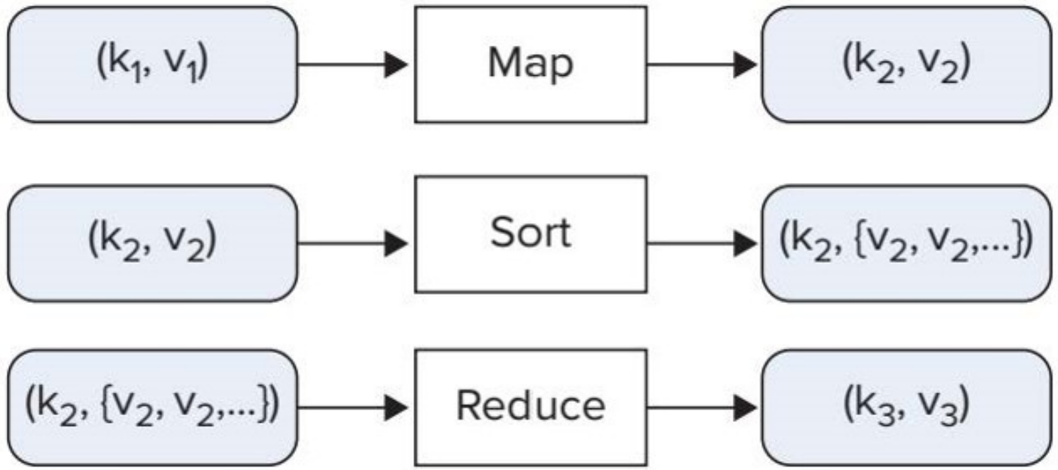
Overview

- You need to implement map, reduce or both
- Map
  - Can be used to split elements,
  - Can be used to filter elements
- Reduce



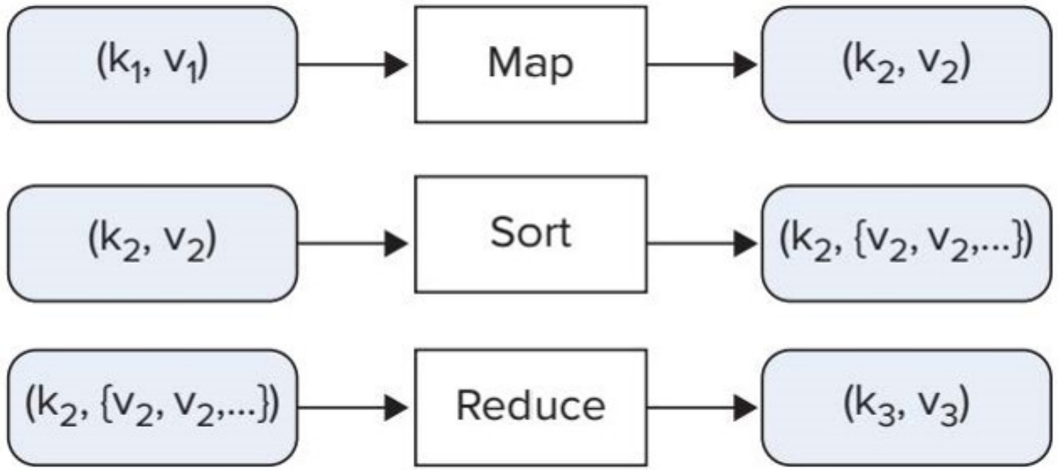
Overview

- You need to implement map, reduce or both
- Map
  - Can be used to split elements,
  - Can be used to filter elements
- Reduce
  - To compute aggregates

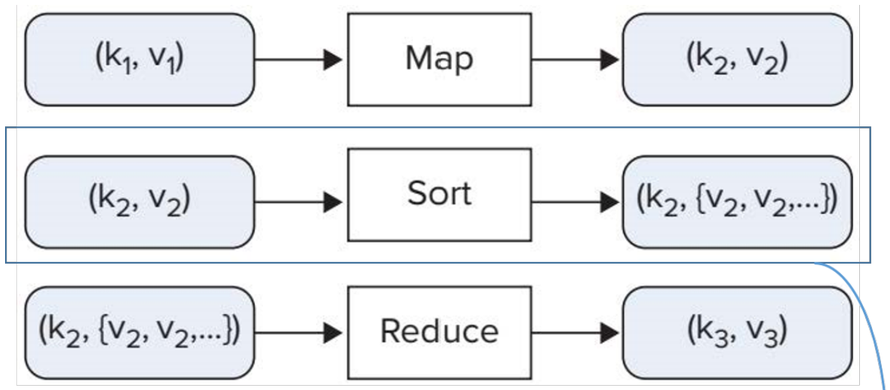


Overview

- You need to implement map, reduce or both
- Map
  - Can be used to split elements,
  - Can be used to filter elements
- Reduce
  - To compute aggregates
  - To combine results

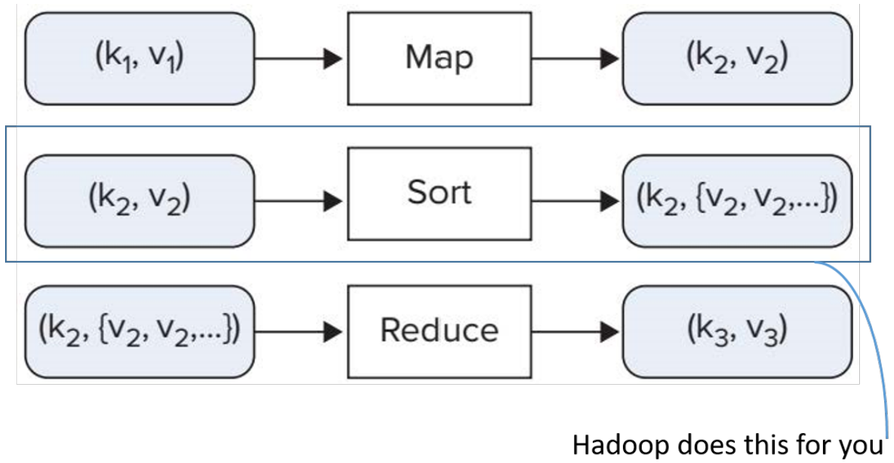


Overview


(k_1, v_1)
Map
(k_2, v_2)
(k_2, v_2)
Sort
$(k_2, \{v_2, v_2, ...\})$
$(k_2, \{v_2, v_2, ...\})$
Reduce
(k_3, v_3)
Hadoop does this for you

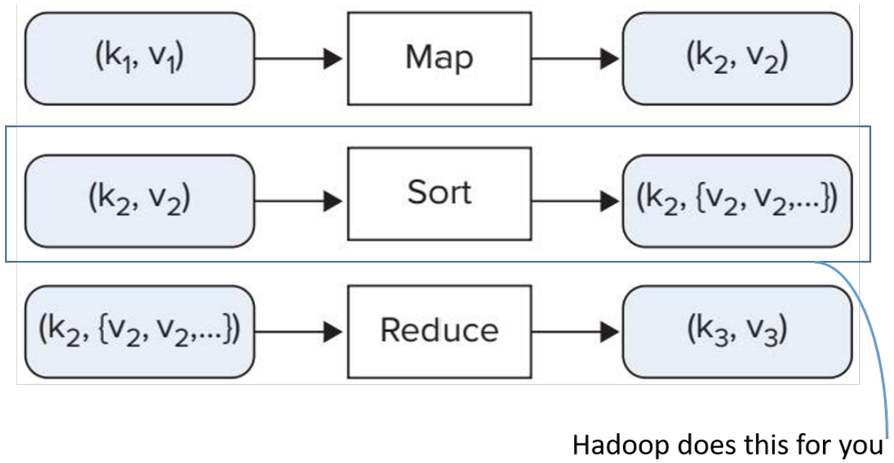
Overview

- You need to implement map, reduce or both



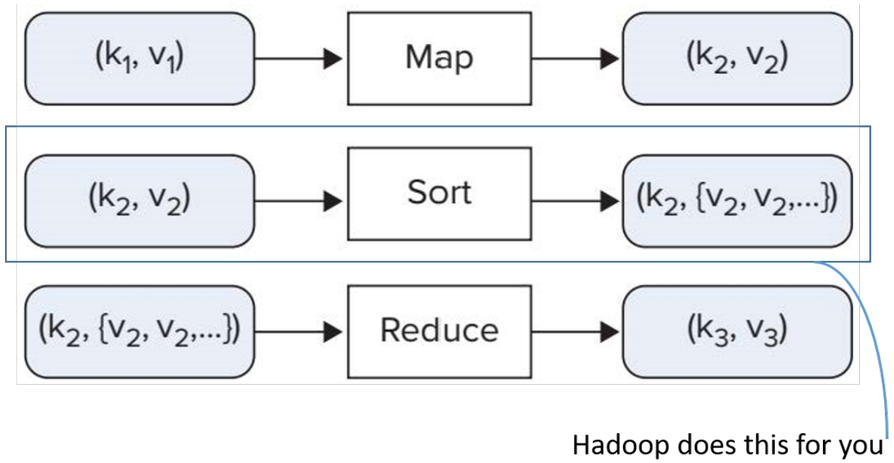
Overview

- You need to implement map, reduce or both
- Map



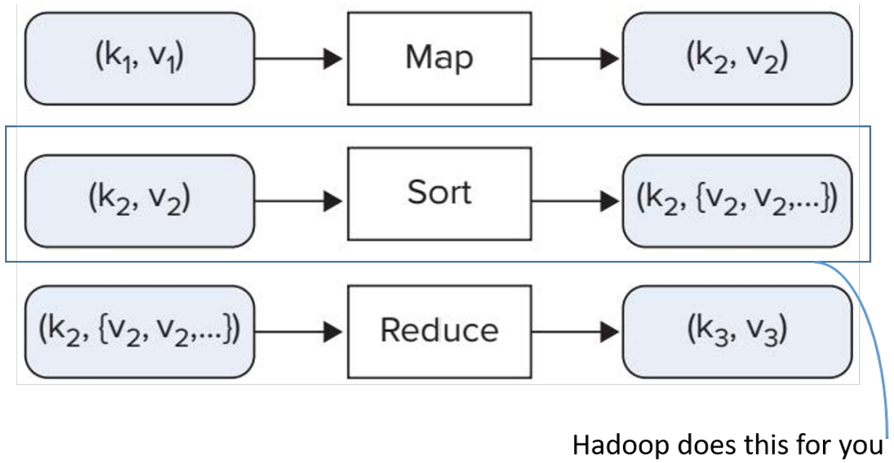
Overview

- You need to implement map, reduce or both
- Map
  - Can be used to split elements,



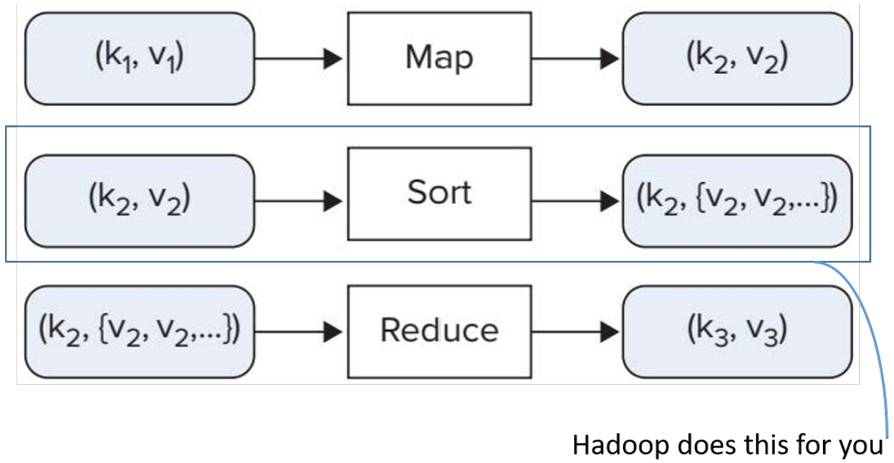
Overview

- You need to implement map, reduce or both
- Map
  - Can be used to split elements,
  - Can be used to filter elements



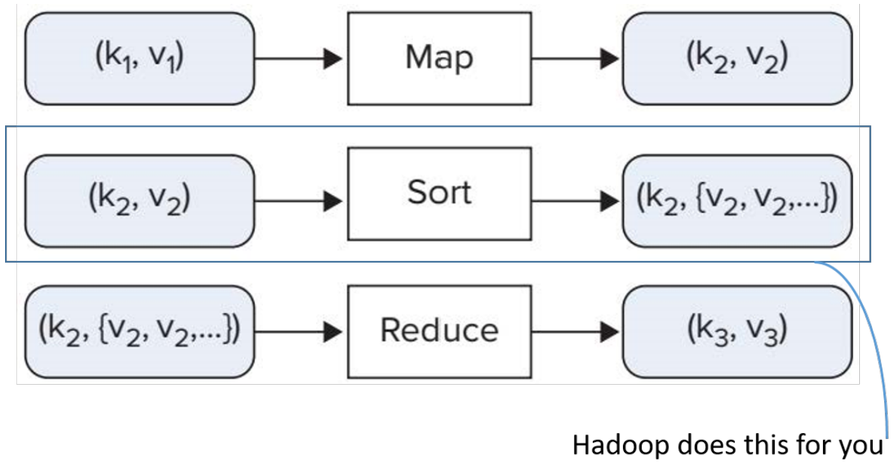
Overview

- You need to implement map, reduce or both
- Map
  - Can be used to split elements,
  - Can be used to filter elements
- Reduce



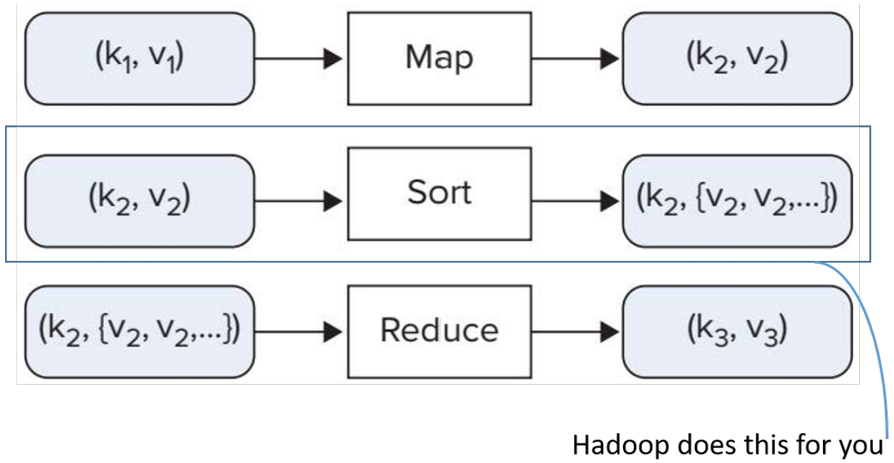
Overview

- You need to implement map, reduce or both
- Map
  - Can be used to split elements,
  - Can be used to filter elements
- Reduce
  - To compute aggregates

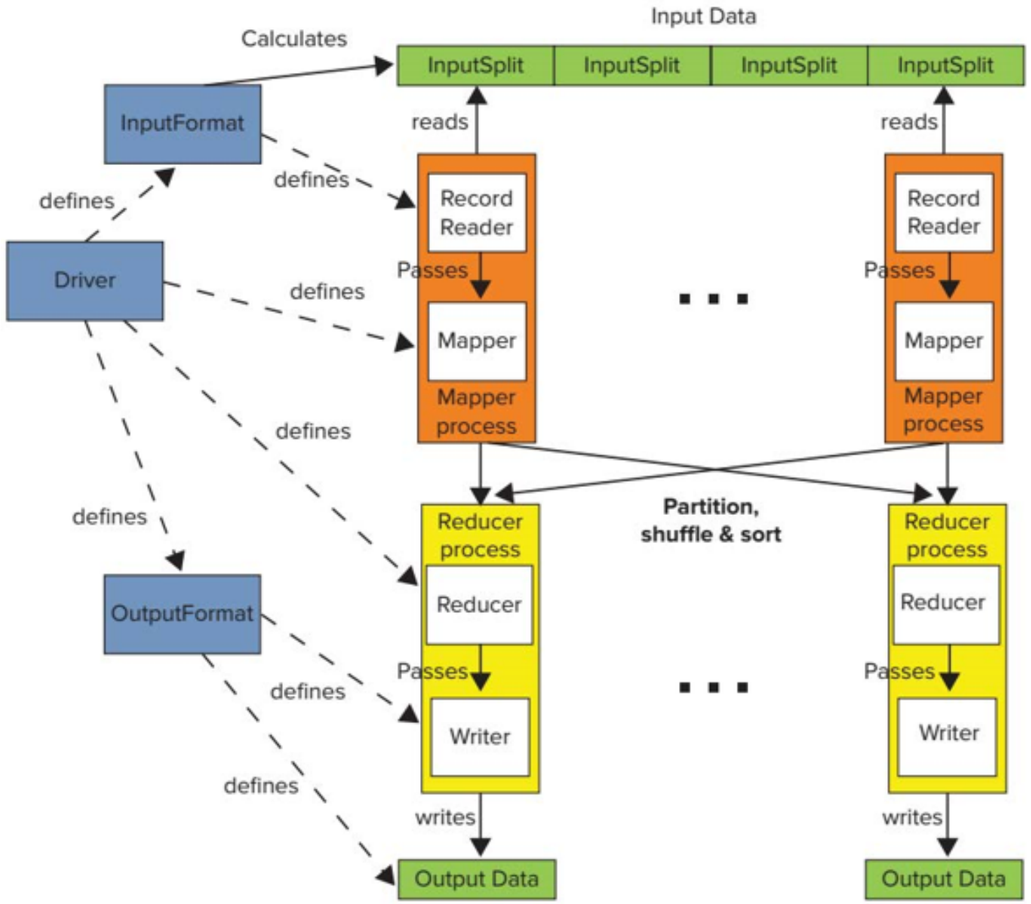


Overview

- You need to implement map, reduce or both
- Map
  - Can be used to split elements,
  - Can be used to filter elements
- Reduce
  - To compute aggregates
  - To combine results

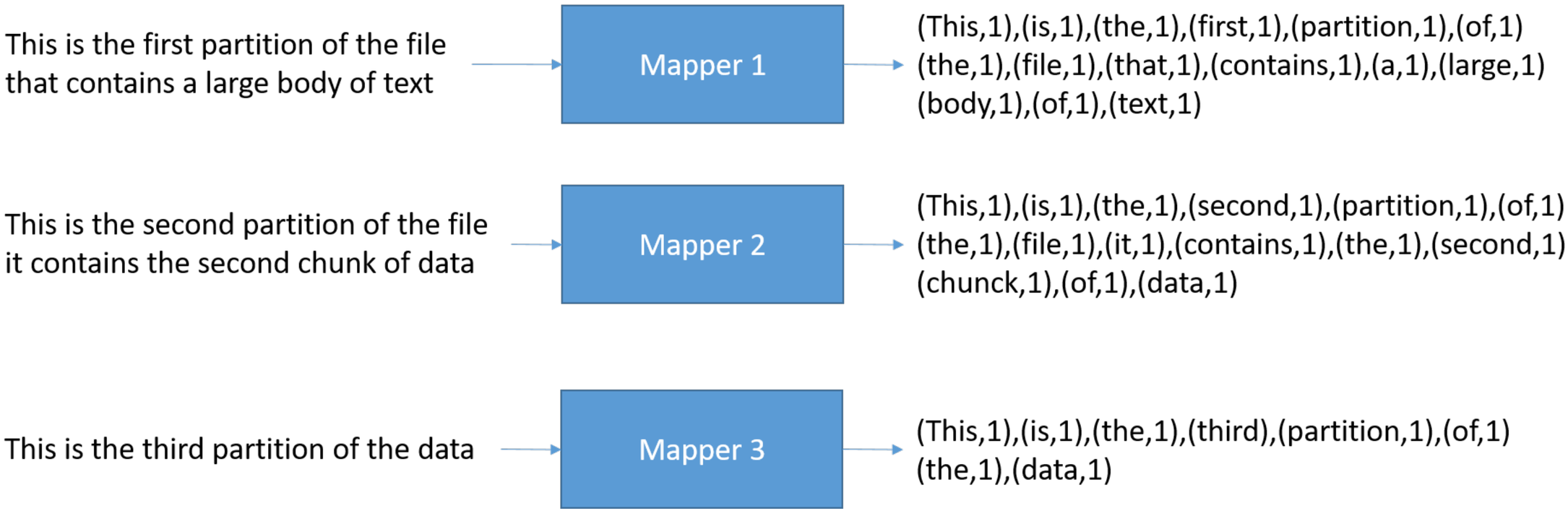


Workflow

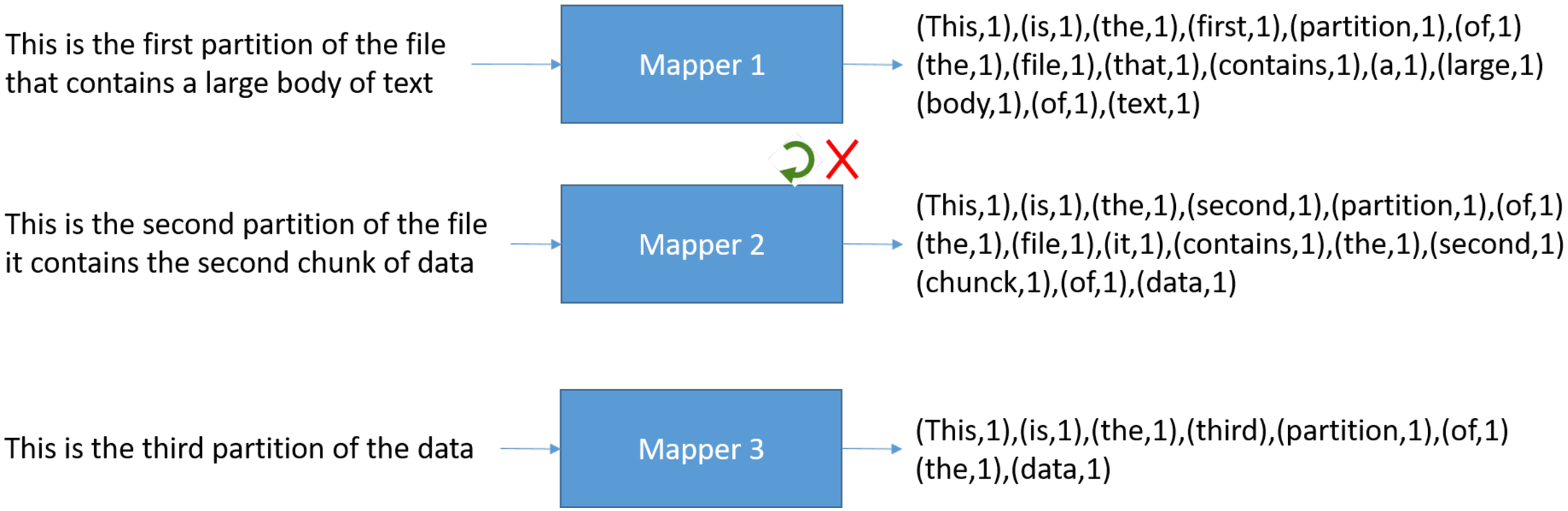

Input Data
Calculates
InputSplit
InputSplit
InputSplit
InputSplit
InputFormat
reads
reads
defines
defines
Record Reader
Record Reader
Driver
Passes
Passes
defines
Mapper
Mapper
Mapper process
Mapper process
defines
defines
Partition, shuffle & sort
Reducer process
Reducer process
OutputFormat
Reducer
Reducer
Passes
Passes
defines
Writer
Writer
defines
writes
writes
Output Data
Output Data

Example: Word Count

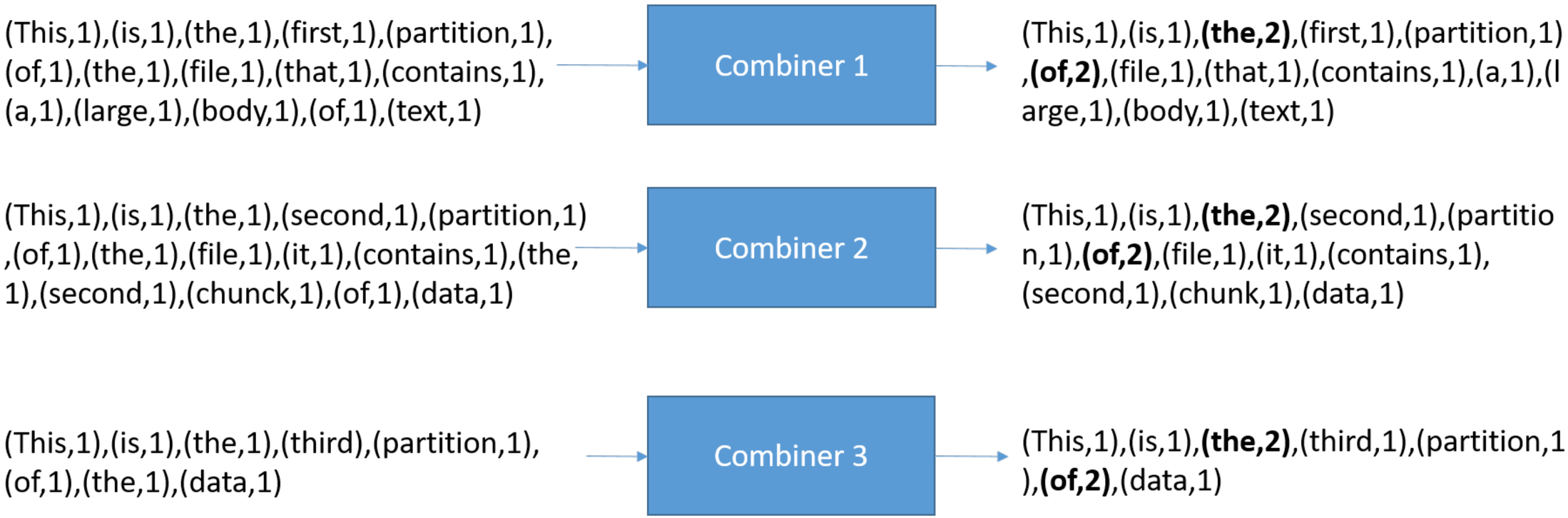
Word Count Mapper



Word Count Mapper



Word Count Combiner (optional) (local reducer)



Combiners run locally on the node where the mapper runs. They can be used to reduce the number of records emitted to the next phase

Word Count Partitioning/Shuffling/Sorting

Word Count Partitioning/Shuffling/Sorting

- Partitioning

Word Count Partitioning/Shuffling/Sorting

- Partitioning
  - Distributes the key-space to the number of reducers available.

Word Count Partitioning/Shuffling/Sorting

- Partitioning
  - Distributes the key-space to the number of reducers available.
    - Number of reducers need not be the same as number of mappers

Word Count Partitioning/Shuffling/Sorting

- Partitioning
  - Distributes the key-space to the number of reducers available.
    - Number of reducers need not be the same as number of mappers
    - Assume two reducers

Word Count Partitioning/Shuffling/Sorting

- Partitioning
  - Distributes the key-space to the number of reducers available.
    - Number of reducers need not be the same as number of mappers
    - Assume two reducers
  - Makes sure that same keys go to the same reducer

Word Count Partitioning/Shuffling/Sorting

- Partitioning
  - Distributes the key-space to the number of reducers available.
    - Number of reducers need not be the same as number of mappers
    - Assume two reducers
  - Makes sure that same keys go to the same reducer
  - Usually implemented via a hash function

Word Count Partitioning/Shuffling/Sorting

- Partitioning
  - Distributes the key-space to the number of reducers available.
    - Number of reducers need not be the same as number of mappers
    - Assume two reducers
  - Makes sure that same keys go to the same reducer
  - Usually implemented via a hash function
- Shuffling

Word Count Partitioning/Shuffling/Sorting

- Partitioning
  - Distributes the key-space to the number of reducers available.
    - Number of reducers need not be the same as number of mappers
    - Assume two reducers
  - Makes sure that same keys go to the same reducer
  - Usually implemented via a hash function
- Shuffling
  - Preparing partitioned keys for the reduce step

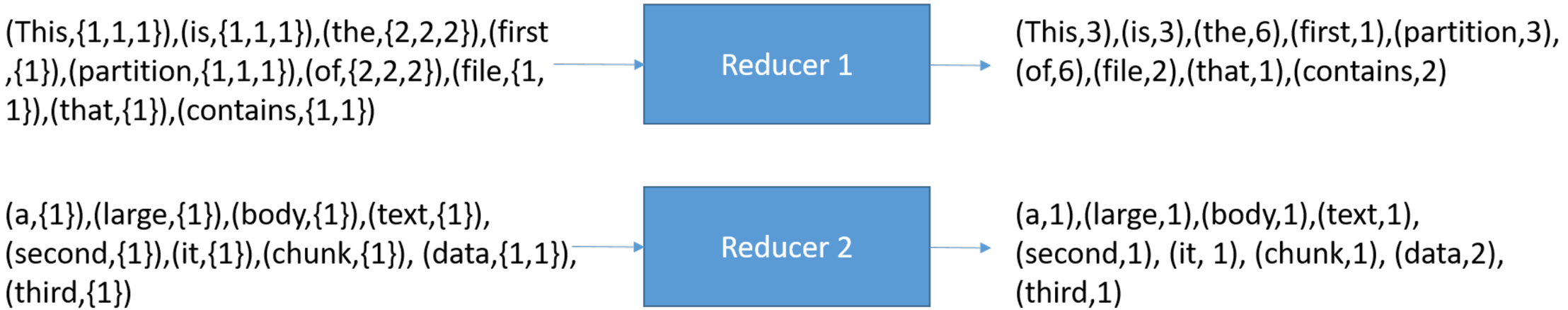
Word Count Partitioning/Shuffling/Sorting

- Partitioning
  - Distributes the key-space to the number of reducers available.
    - Number of reducers need not be the same as number of mappers
    - Assume two reducers
  - Makes sure that same keys go to the same reducer
  - Usually implemented via a hash function
- Shuffling
  - Preparing partitioned keys for the reduce step
- Sorting

Word Count Partitioning/Shuffling/Sorting

- Partitioning
  - Distributes the key-space to the number of reducers available.
    - Number of reducers need not be the same as number of mappers
    - Assume two reducers
  - Makes sure that same keys go to the same reducer
  - Usually implemented via a hash function
- Shuffling
  - Preparing partitioned keys for the reduce step
- Sorting
  - Transforms (k,v1), (k,v2), (k,v3) (kx,{v1,v2,v3})

Word Count Reducer



There will be two files written to the specified output path. The number is equivalent to the number of reducers.

Other Uses for MapReduce

Other Uses for MapReduce

- Reduce phase is optional in a MapReduce job

Other Uses for MapReduce

- Reduce phase is optional in a MapReduce job
- Running face-recognition on millions of images

Other Uses for MapReduce

- Reduce phase is optional in a MapReduce job
- Running face-recognition on millions of images
  - Map-only job

Other Uses for MapReduce

- Reduce phase is optional in a MapReduce job
- Running face-recognition on millions of images
  - Map-only job
  - Input (Image ID, Image)

Other Uses for MapReduce

- Reduce phase is optional in a MapReduce job
- Running face-recognition on millions of images
  - Map-only job
  - Input (Image ID, Image)
  - Output (Image ID, list of features)

Other Uses for MapReduce

- Reduce phase is optional in a MapReduce job
- Running face-recognition on millions of images
  - Map-only job
  - Input (Image ID, Image)
  - Output (Image ID, list of features)
  - Features are loaded to distributed cache

Other Uses for MapReduce

- Reduce phase is optional in a MapReduce job
- Running face-recognition on millions of images
  - Map-only job
  - Input (Image ID, Image)
  - Output (Image ID, list of features)
  - Features are loaded to distributed cache
  - The results of each mapper will be written in a separate file in the output folder

Other Uses for MapReduce

- Reduce phase is optional in a MapReduce job
- Running face-recognition on millions of images
  - Map-only job
  - Input (Image ID, Image)
  - Output (Image ID, list of features)
  - Features are loaded to distributed cache
  - The results of each mapper will be written in a separate file in the output folder
  - What to do if we want all results in a single file?

Other Uses for MapReduce

- Reduce phase is optional in a MapReduce job
- Running face-recognition on millions of images
  - Map-only job
  - Input (Image ID, Image)
  - Output (Image ID, list of features)
  - Features are loaded to distributed cache
  - The results of each mapper will be written in a separate file in the output folder
  - What to do if we want all results in a single file?
    - Add a single reducer

Other Uses for MapReduce

- Reduce phase is optional in a MapReduce job
- Running face-recognition on millions of images
  - Map-only job
  - Input (Image ID, Image)
  - Output (Image ID, list of features)
  - Features are loaded to distributed cache
  - The results of each mapper will be written in a separate file in the output folder
  - What to do if we want all results in a single file?
    - Add a single reducer
    - However, this will come with a large overhead as shuffle and sort will be invoked

FlumeJava[1]

Easy, Efficient Data-Parallel Pipelines

[1] The reason as to why Airflow does not support streaming is that there are no obvious behavior rules that could be set so that the airflow scheduler could deterministically check if it has been completed or not.

Problem

Problem

- MapReduce works well for computations that can be broken down into a map step, a shuffle step, and a reduce step

Problem

- MapReduce works well for computations that can be broken down into a map step, a shuffle step, and a reduce step
- Many real-world computations require a chain of MapReduce works, e.g., graph analytics

Problem

- MapReduce works well for computations that can be broken down into a map step, a shuffle step, and a reduce step
- Many real-world computations require a chain of MapReduce works, e.g., graph analytics
- Additional coordination code chains together the separate MapReduce stages, requiring additional work to manage the intermediate results

Solution

Solution

- FlumeJava is a new system that aims to support the development of data-parallel pipelines.

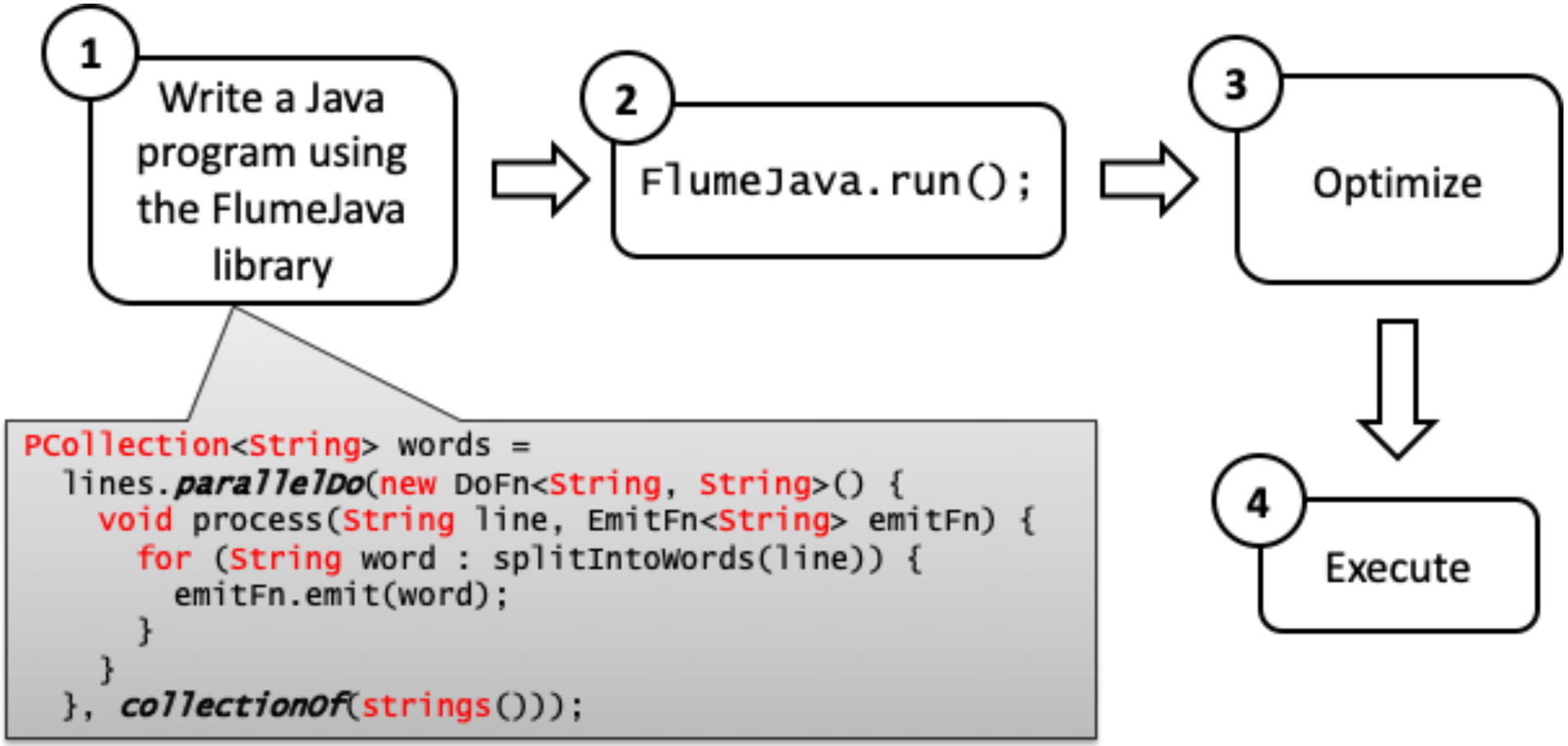
Solution

- FlumeJava is a new system that aims to support the development of data-parallel pipelines.
- abstract away the details of how data is represented, including whether the data is represented as an in-memory data structure

Solution

- FlumeJava is a new system that aims to support the development of data-parallel pipelines.
- abstract away the details of how data is represented, including whether the data is represented as an in-memory data structure
- FlumeJava evaluation is lazy: The invocation of a parallel operation does not actually run the operation, but instead simply records the operation and its arguments in an internal execution plan graph structure.

FlumeJava Workflow



Core Abstractions

Parallel Collections

Parallel Collections

- **PCollection**<T> a (possibly huge) immutable bag of elements of type T. If it has a well-defined order is called a Sequence

Parallel Collections

- **PCollection**<T> a (possibly huge) immutable bag of elements of type T. If it has a well-defined order is called a Sequence
- **PTable**<K, V> represents a (possibly huge) immutable multi-map with keys of type K and values of type V.

Parallel Collections

- **PCollection**<T> a (possibly huge) immutable bag of elements of type T. If it has a well-defined order is called a Sequence
- **PTable**<K, V> represents a (possibly huge) immutable multi-map with keys of type K and values of type V.
  - In Java, PTable<K,V> is a subclass of PCollection<Pair<K,V>>

Parallel Collections

- **PCollection**<T> a (possibly huge) immutable bag of elements of type T. If it has a well-defined order is called a Sequence
- **PTable**<K, V> represents a (possibly huge) immutable multi-map with keys of type K and values of type V.
  - In Java, PTable<K,V> is a subclass of PCollection<Pair<K,V>>
- **PObjects**<T> is a container for a single Java object of type T that acts much like a future. Like PCollections, PObjects can be either deferred or materialized, allowing them to be computed as results of deferred operations in pipelines.

Parallel Operations (Primitives)

Parallel Operations (Primitives)

- **parallelDo**() higher-order function that supports element-wise computation over an input PCollection<T>

Parallel Operations (Primitives)

- **parallelDo**() higher-order function that supports element-wise computation over an input PCollection<T>
- **groupByKey**() converts a multi-map of type PTable<K,V> (which can have many key/value pairs with the same key) into a uni-map of type PTable<K, Collection\<V\>> [2]

---

[2] captures the essence of the shuffle step of MapReduce.

Parallel Operations (Primitives)

- **parallelDo**() higher-order function that supports element-wise computation over an input PCollection<T>
- **groupByKey**() converts a multi-map of type PTable<K,V> (which can have many key/value pairs with the same key) into a uni-map of type PTable<K, Collection\<V\>> [2]
- **combineValues**() takes an input PTable<K, Collection<V>> and an associative combining function on Vs, and returns a PTable<K, V> where each input collection of values has been combined into a single output value[3]

---

[2] captures the essence of the shuffle step of MapReduce.

[3] combineValues() is a special case of parallelDo(). The associativity of the combining function allows it to be implemented via a combination of a MapReduce combiner and a MapReduce reducer.

Parallel Operations (Primitives)

- **parallelDo**() higher-order function that supports element-wise computation over an input PCollection<T>
- **groupByKey**() converts a multi-map of type PTable<K,V> (which can have many key/value pairs with the same key) into a uni-map of type PTable<K, Collection\<V\>> [2]
- **combineValues**() takes an input PTable<K, Collection<V>> and an associative combining function on Vs, and returns a PTable<K, V> where each input collection of values has been combined into a single output value[3]
- **flatten**() takes a list of PCollection<T>s and returns a single PCollection<T>

---

[2] captures the essence of the shuffle step of MapReduce.

[3] combineValues() is a special case of parallelDo(). The associativity of the combining function allows it to be implemented via a combination of a MapReduce combiner and a MapReduce reducer.

WordCount in Primitive FlumeJava

Sentences to Words

```
PCollection<String> words =
    lines.parallelDo(new DoFn<String,String>() {
        void process(String line, EmitFn<String> emitFn) {
            for (String word : splitIntoWords(line)) {
                emitFn.emit(word);
            }
        }
    }, collectionOf(strings()));
```

Words Occurrences

```
PTable<String,Integer> wordsWithOnes =
    words.parallelDo(
        new DoFn<String, Pair<String,Integer>>() {
            void process(String word,
                EmitFn<Pair<String,Integer>> emitFn) {
                    emitFn.emit(Pair.of(word, 1));
            }
    }, tableOf(strings(), ints()));
```

Occurrences to Counts

```
PTable<String,Collection<Integer>> gwOnes = wordsWithOnes.groupByKey();

PTable<String,Integer> wordCounts = gwOnes.combineValues(SUM_INTS);
```

Derived operations

Derived operations

- **count**() takes a PCollection<T> and returns a PTable<T, Integer> mapping each distinct element of the input PCollection to the number of times it occurs. f parallelDo(), groupByKey(), and combineValues()

```
PTable<String,Collection<Integer>> gwOnes = wordsWithOnes.groupByKey();

PTable<String,Integer> wordCounts = gwOnes.combineValues(SUM_INTS);
```

OR

```
PTable<String,Integer> wordCounts = words.count();
```

Derived operations

Derived operations

- **count**() takes a PCollection<T> and returns a PTable<T, Integer> mapping each distinct element of the input PCollection to the number of times it occurs. f parallelDo(), groupByKey(), and combineValues()

Derived operations

- **count**() takes a PCollection<T> and returns a PTable<T, Integer> mapping each distinct element of the input PCollection to the number of times it occurs. f parallelDo(), groupByKey(), and combineValues()
- **join**() implements a kind of join over two or more PTables sharing a common key type. When applied to a multi-map PTable<K, V1> and a multimap PTable<K, V2>, join() returns a uni-map PTable<K,Tuple2<Collection<V1>, Collection<V2\>>>

Join

Join

1. Apply **parallelDo**() to each input PTable<K, Vi> to convert it into a common format of type PTable<K, TaggedUnion2<V1,V2>>.

Join

1. Apply **parallelDo**() to each input PTable<K, Vi> to convert it into a common format of type PTable<K, TaggedUnion2<V1,V2>>.
2. Combine the tables using **flatten**().

Join

1. Apply **parallelDo**() to each input PTable<K, Vi> to convert it into a common format of type PTable<K, TaggedUnion2<V1,V2>>.
2. Combine the tables using **flatten**().
3. Apply **groupByKey**() to the flattened table to produce a PTable<K, Collection<TaggedUnion2<V1,V2>>>.

Join

1. Apply **parallelDo**() to each input PTable<K, Vi> to convert it into a common format of type PTable<K, TaggedUnion2<V1,V2>>.
2. Combine the tables using **flatten**().
3. Apply **groupByKey**() to the flattened table to produce a PTable<K, Collection<TaggedUnion2<V1,V2>>>.
4. Apply **parallelDo**() to the key-grouped table, converting each Collection<TaggedUnion2<V1,V2>> into a Tuple2 of a Collection<V1> and a Collection<V2>.

Derived operations

Derived operations

- **count**() takes a PCollection<T> and returns a PTable<T, Integer> mapping each distinct element of the input PCollection to the number of times it occurs. f parallelDo(), groupByKey(), and combineValues()

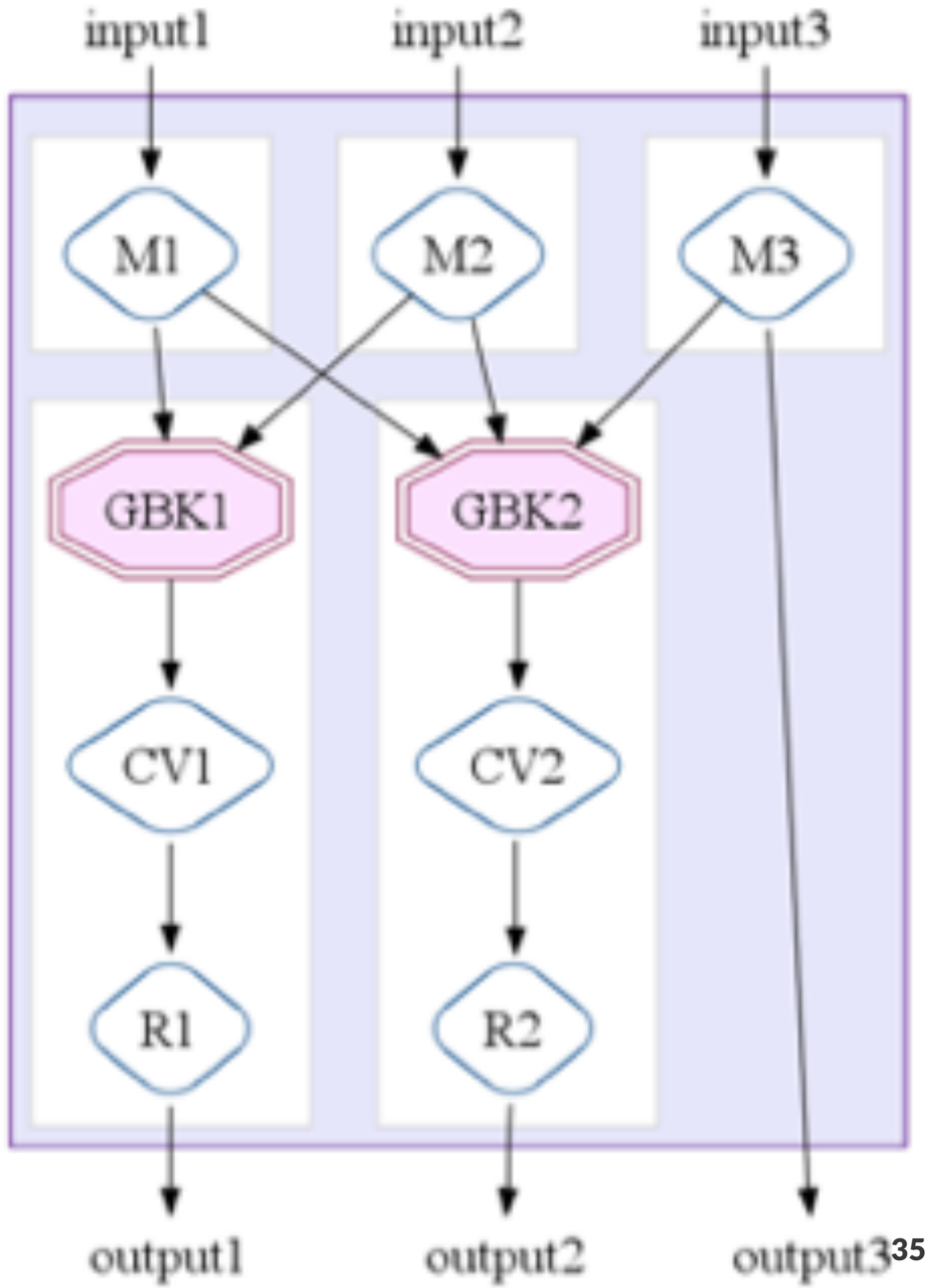
Derived operations

- **count**() takes a PCollection<T> and returns a PTable<T, Integer> mapping each distinct element of the input PCollection to the number of times it occurs. f parallelDo(), groupByKey(), and combineValues()
- **join**() implements a kind of join over two or more PTables sharing a common key type. When applied to a multi-map PTable<K, V1> and a multimap PTable<K, V2>, join() returns a uni-map PTable<K,Tuple2<Collection<V1>, Collection<V2\>>>

Derived operations

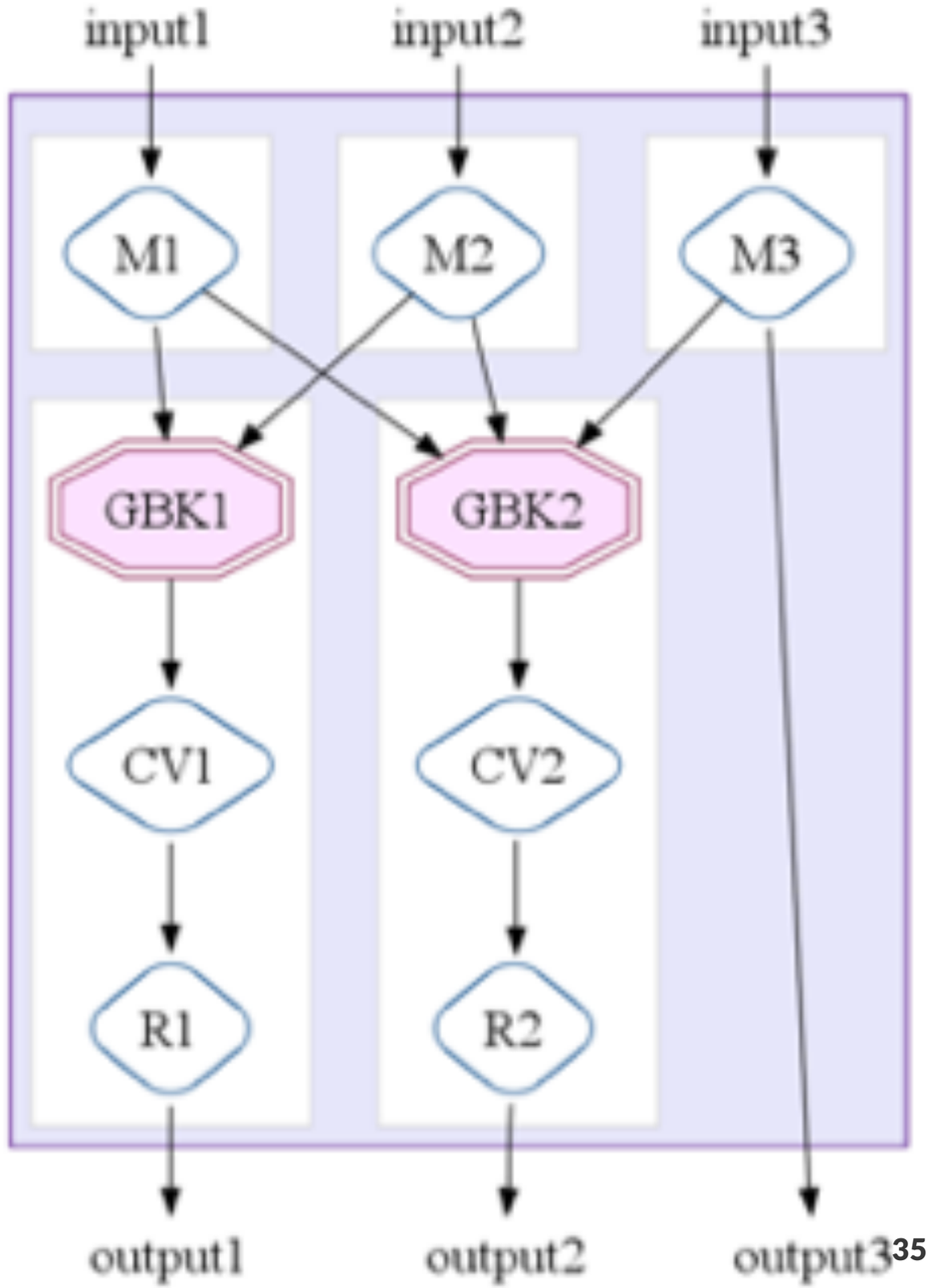
- **count**() takes a PCollection<T> and returns a PTable<T, Integer> mapping each distinct element of the input PCollection to the number of times it occurs. f parallelDo(), groupByKey(), and combineValues()
- **join**() implements a kind of join over two or more PTables sharing a common key type. When applied to a multi-map PTable<K, V1> and a multimap PTable<K, V2>, join() returns a uni-map PTable<K,Tuple2<Collection<V1>, Collection<V2\>>>
- **top**() which takes a comparison function and a count N and returns the greatest N elements of its receiver PCollection according to the comparison function.

MapShuffleCombineReduce (MSCR)



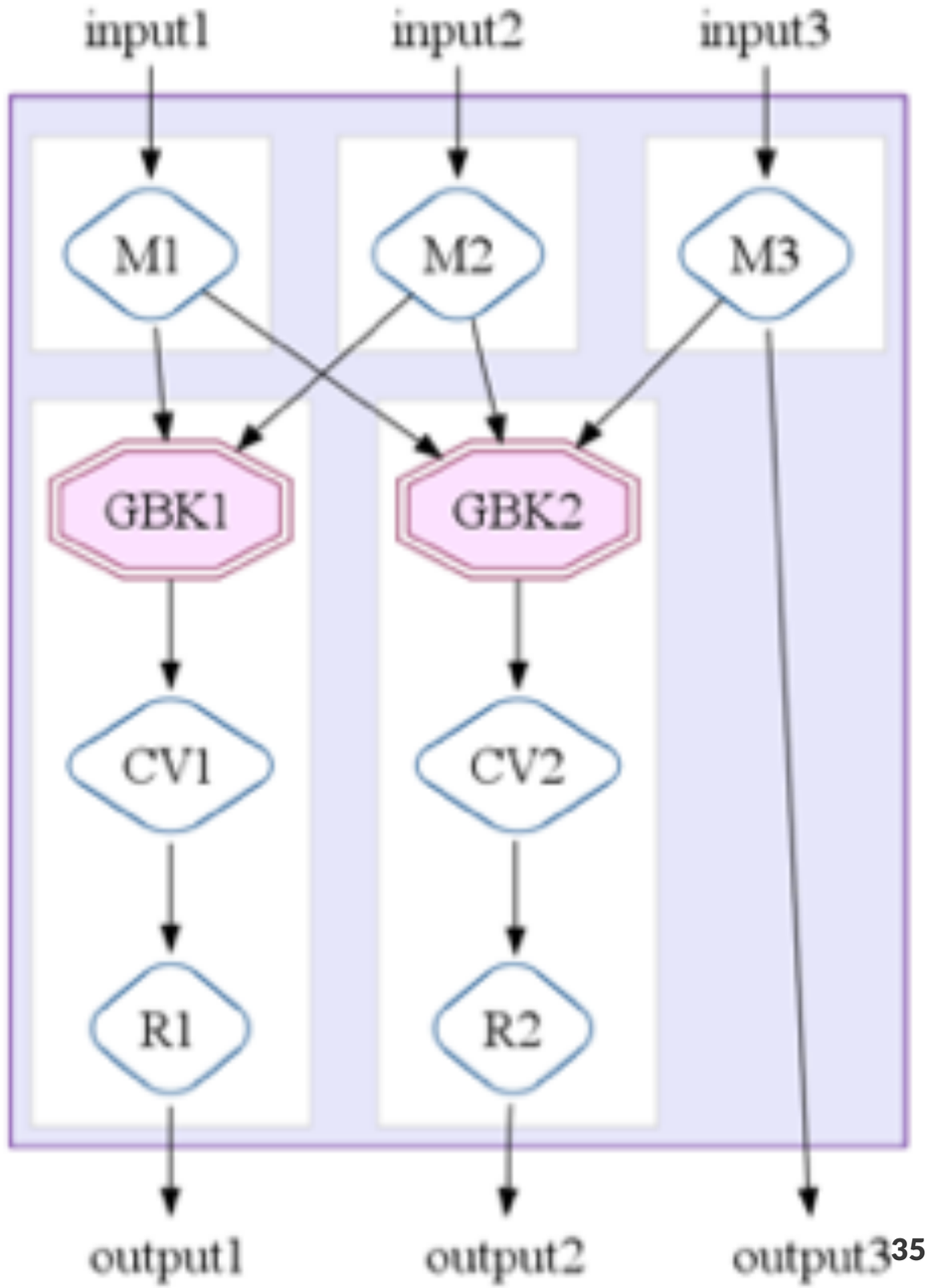
MapShuffleCombineReduce (MSCR)

- Transform combinations of the four primitives into single MapReduce



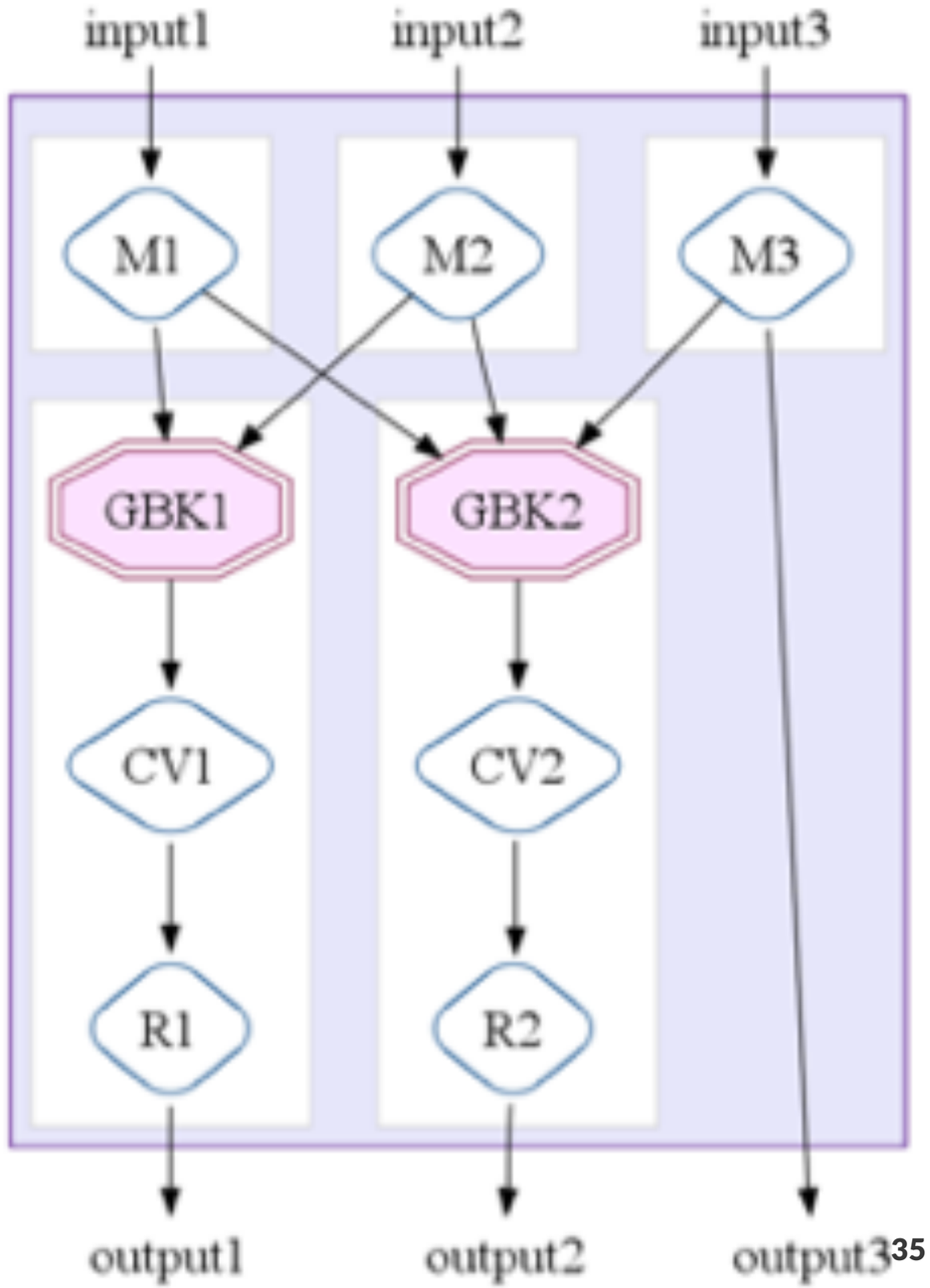
MapShuffleCombineReduce (MSCR)

- Transform combinations of the four primitives into single MapReduce
- Generalizes MapReduce



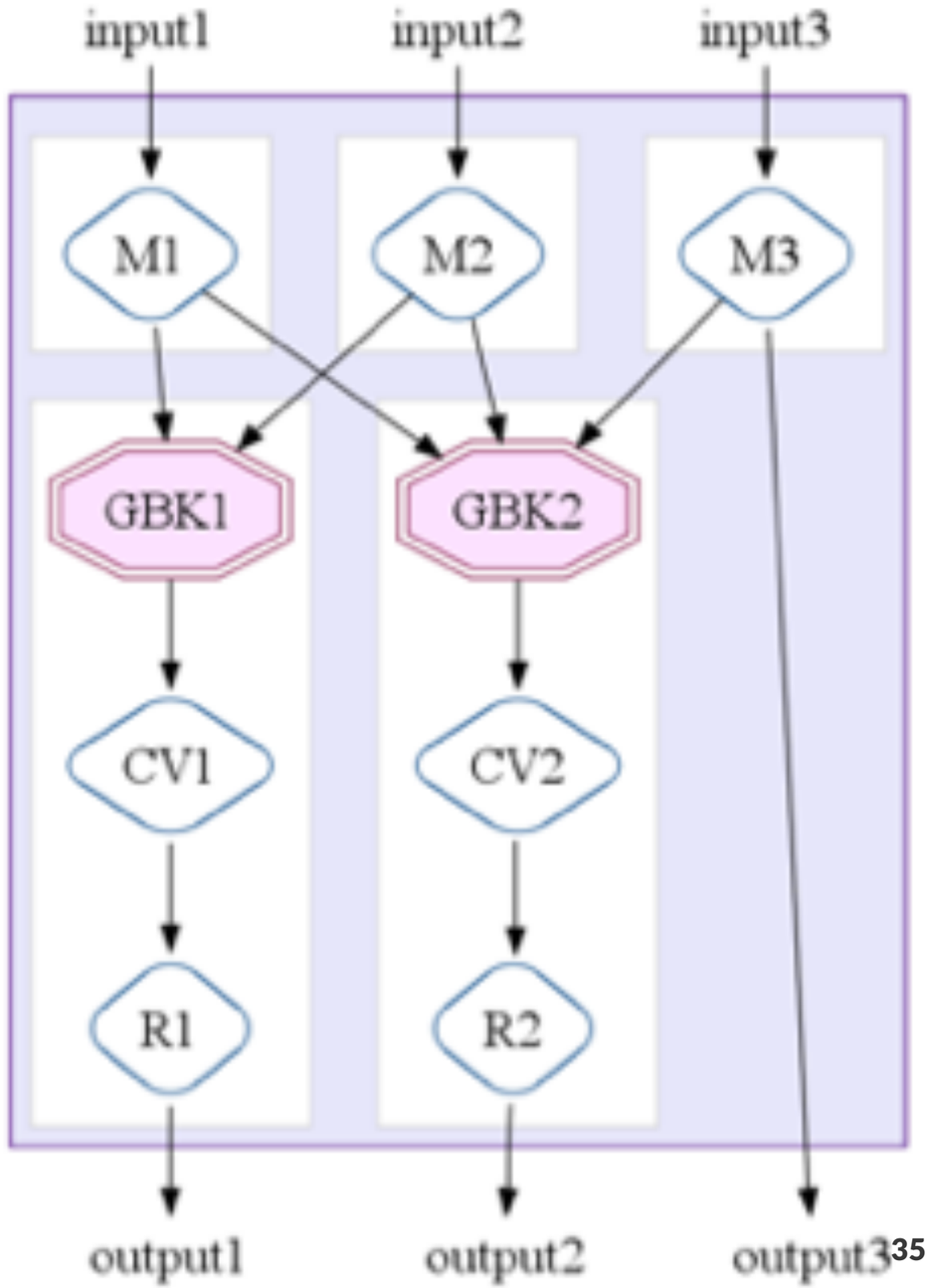
MapShuffleCombineReduce (MSCR)

- Transform combinations of the four primitives into single MapReduce
- Generalizes MapReduce
  - Multiple reducers/combiners



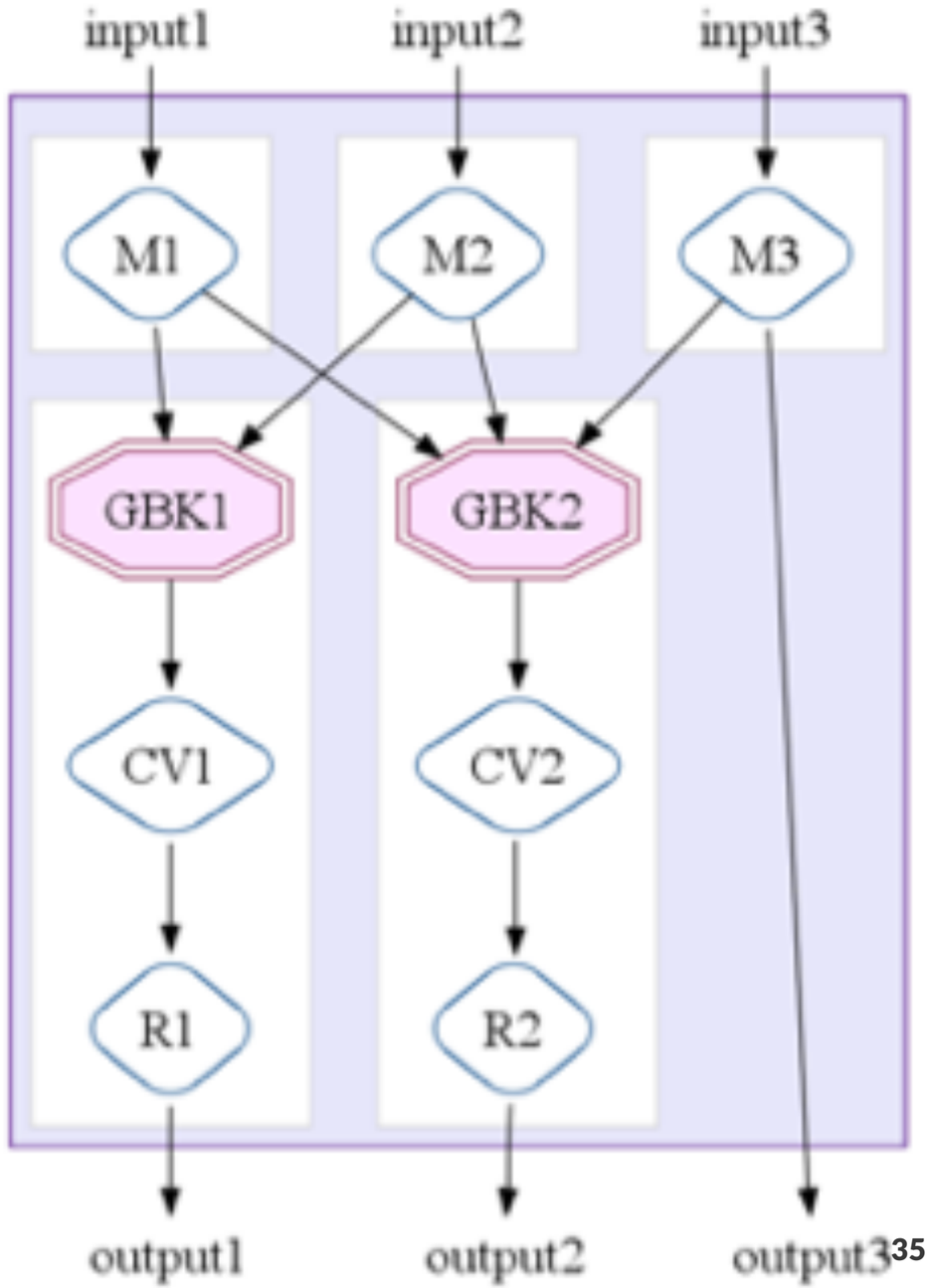
MapShuffleCombineReduce (MSCR)

- Transform combinations of the four primitives into single MapReduce
- Generalizes MapReduce
  - Multiple reducers/combiners
  - Multiple output per reducer



MapShuffleCombineReduce (MSCR)

- Transform combinations of the four primitives into single MapReduce
- Generalizes MapReduce
  - Multiple reducers/combiners
  - Multiple output per reducer
  - Pass-through outputs



Optimizer

Overall Optimizer Strategy

The optimizer performs a series of passes over the execution plan, with the overall goal to produce the fewest, most efficient MSCR operations in the final optimized plan.

Overall Optimizer Strategy

The optimizer performs a series of passes over the execution plan, with the overall goal to produce the fewest, most efficient MSCR operations in the final optimized plan.

- **Sink Flattens**. A Flatten operation can be pushed down through consuming ParallelDo operations by duplicating the ParallelDo before each input to the Flatten.

Overall Optimizer Strategy

The optimizer performs a series of passes over the execution plan, with the overall goal to produce the fewest, most efficient MSCR operations in the final optimized plan.

- **Sink Flattens**. A Flatten operation can be pushed down through consuming ParallelDo operations by duplicating the ParallelDo before each input to the Flatten.
- **Lift CombineValues** operations. If a CombineValues operation immediately follows a GroupByKey operation the CombineValues is henceforth treated as a normal ParallelDooperation and subject to ParallelDo fusion.

Overall Optimizer Strategy

The optimizer performs a series of passes over the execution plan, with the overall goal to produce the fewest, most efficient MSCR operations in the final optimized plan.

- **Sink Flattens**. A Flatten operation can be pushed down through consuming ParallelDo operations by duplicating the ParallelDo before each input to the Flatten.
- **Lift CombineValues** operations. If a CombineValues operation immediately follows a GroupByKey operation the CombineValues is henceforth treated as a normal ParallelDooperation and subject to ParallelDo fusion.
- **Insert fusion blocks**. If two GroupByKey operations are connected by a producer-consumer chain of one or more ParallelDo operations, the optimizer must choose which ParallelDos should fuse “up” into the output channel of the earlier GroupByKey, and which should fuse “down” into the input channel of the later GroupByKey

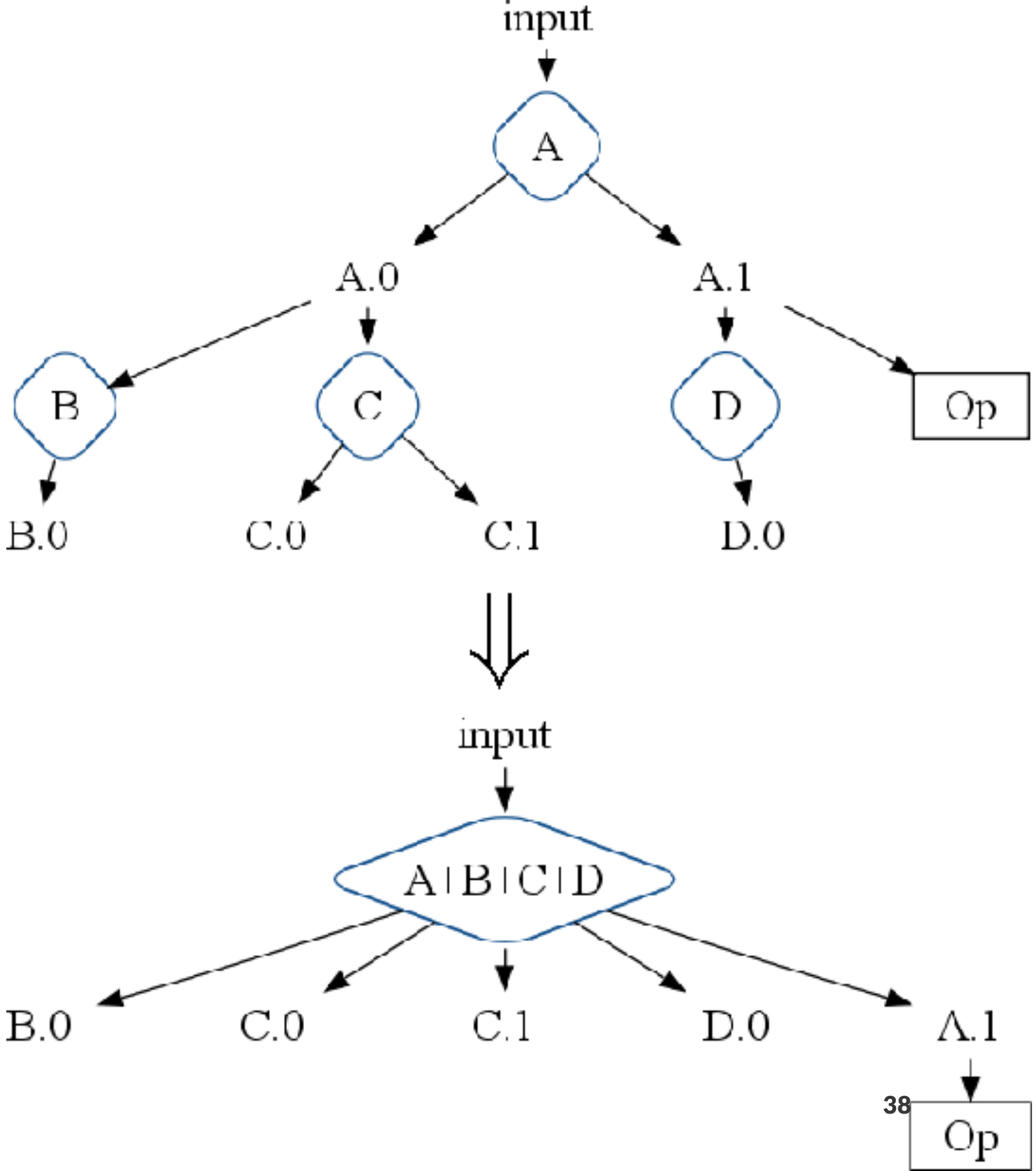
Overall Optimizer Strategy

The optimizer performs a series of passes over the execution plan, with the overall goal to produce the fewest, most efficient MSCR operations in the final optimized plan.

- **Sink Flattens**. A Flatten operation can be pushed down through consuming ParallelDo operations by duplicating the ParallelDo before each input to the Flatten.
- **Lift CombineValues** operations. If a CombineValues operation immediately follows a GroupByKey operation the CombineValues is henceforth treated as a normal ParallelDooperation and subject to ParallelDo fusion.
- **Insert fusion blocks**. If two GroupByKey operations are connected by a producer-consumer chain of one or more ParallelDo operations, the optimizer must choose which ParallelDos should fuse “up” into the output channel of the earlier GroupByKey, and which should fuse “down” into the input channel of the later GroupByKey
- **Fuse ParallelDos**

ParallelDo Fusion

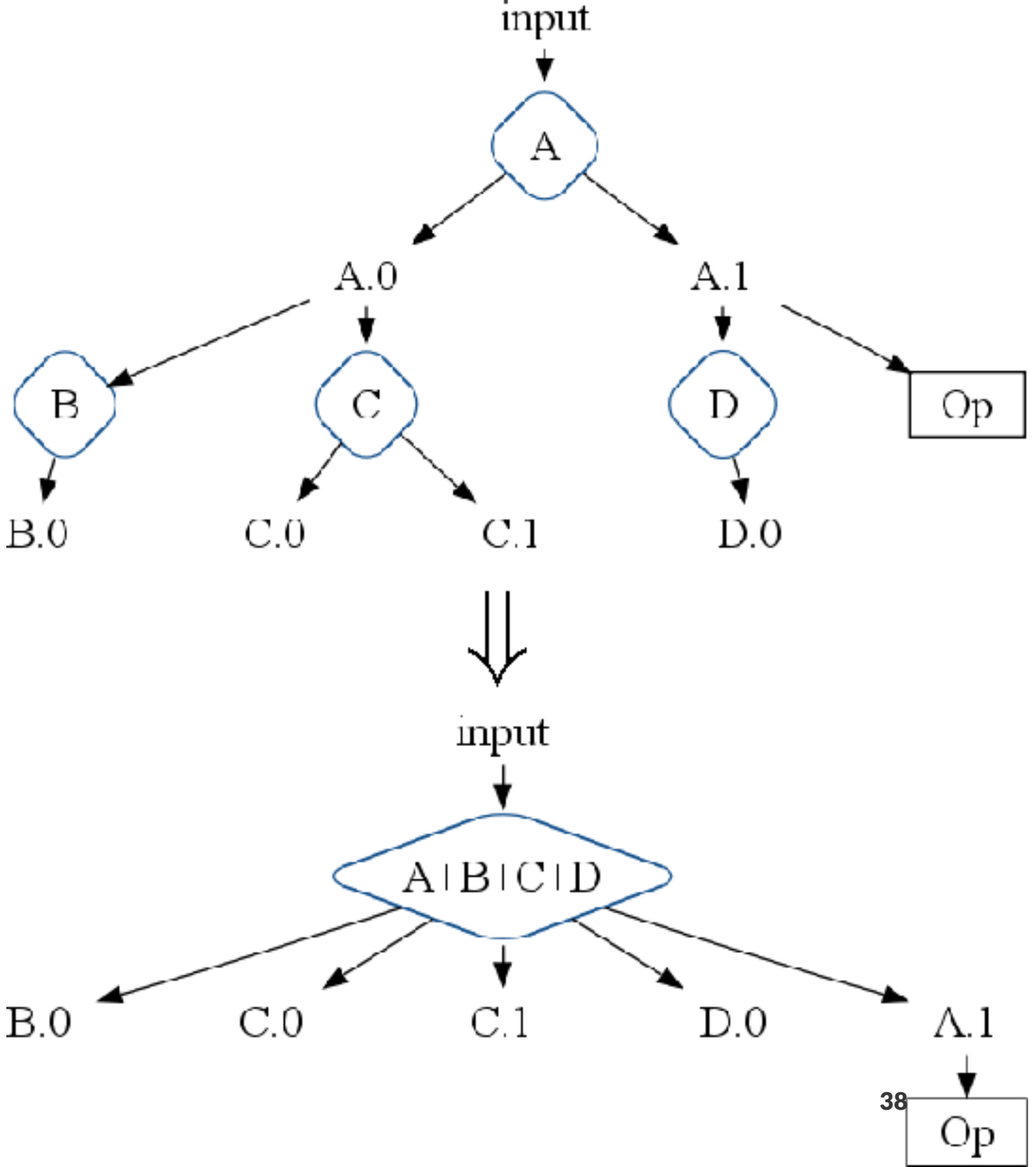
One of the simplest and most intuitive optimizations is ParallelDo producer-consumer fusion.



ParallelDo Fusion

One of the simplest and most intuitive optimizations is ParallelDo producer-consumer fusion.

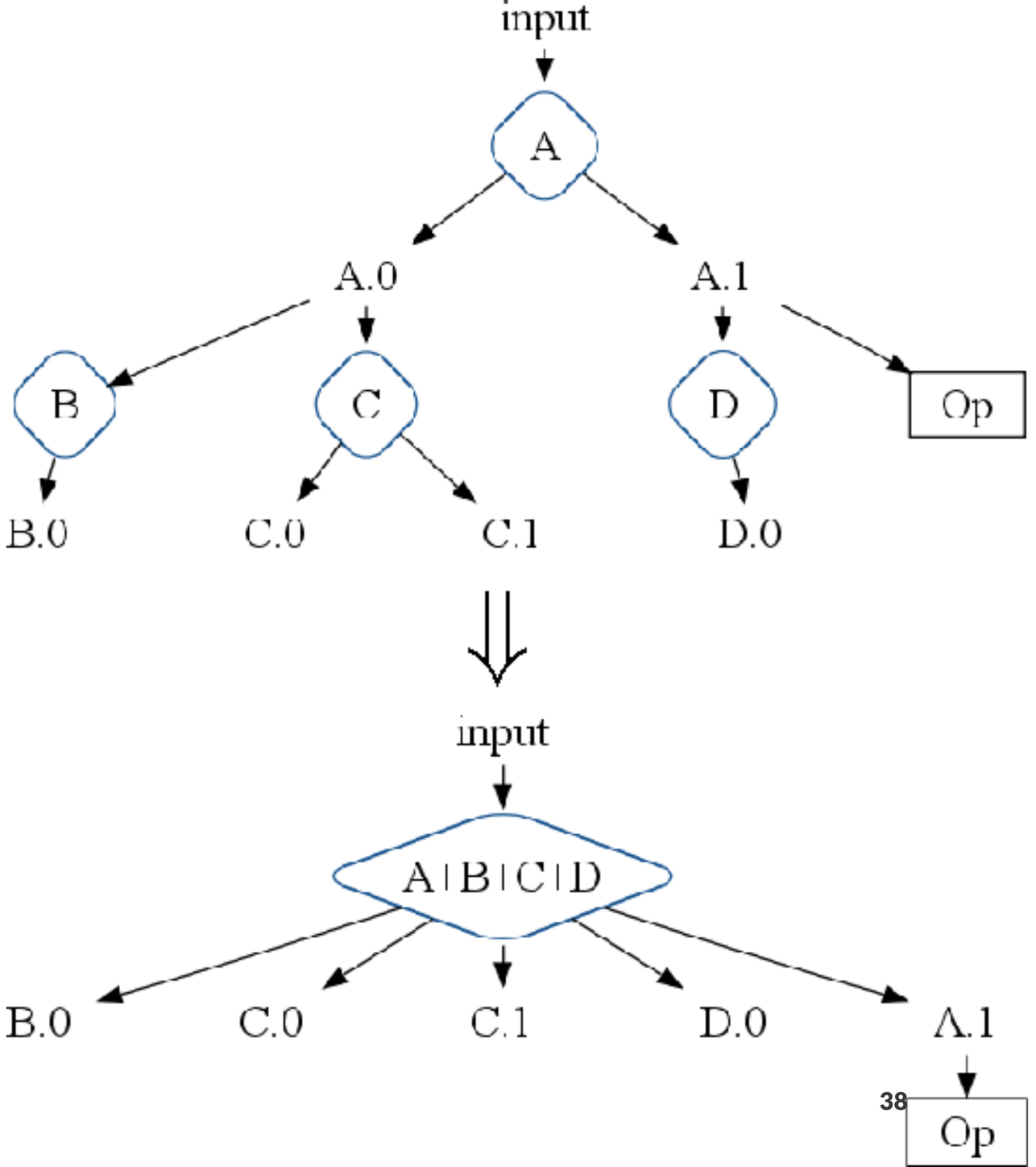
- **function composition**: one ParallelDo operation performs function f, and its result is consumed by another ParallelDo operation that performs function g. The two ParallelDo operations are replaced by a single multi-output ParallelDo that computes both f and g o f.



ParallelDo Fusion

One of the simplest and most intuitive optimizations is ParallelDo producer-consumer fusion.

- **function composition**: one ParallelDo operation performs function f, and its result is consumed by another ParallelDo operation that performs function g. The two ParallelDo operations are replaced by a single multi-output ParallelDo that computes both f and g o f.
- **sibling fusion** when two or more ParallelDo operations read the same input they are fused into a single multi-output ParallelDo operation that computes the results of all the fused operations in a single pass over the input.



Overall Optimizer Strategy

The optimizer performs a series of passes over the execution plan, with the overall goal to produce the fewest, most efficient MSCR operations in the final optimized plan.

Overall Optimizer Strategy

The optimizer performs a series of passes over the execution plan, with the overall goal to produce the fewest, most efficient MSCR operations in the final optimized plan.

- **Sink Flattens**. A Flatten operation can be pushed down through consuming ParallelDo operations by duplicating the ParallelDo before each input to the Flatten.

Overall Optimizer Strategy

The optimizer performs a series of passes over the execution plan, with the overall goal to produce the fewest, most efficient MSCR operations in the final optimized plan.

- **Sink Flattens**. A Flatten operation can be pushed down through consuming ParallelDo operations by duplicating the ParallelDo before each input to the Flatten.
- **Lift CombineValues** operations. If a CombineValues operation immediately follows a GroupByKey operation the CombineValues is henceforth treated as a normal ParallelDooperation and subject to ParallelDo fusion.

Overall Optimizer Strategy

The optimizer performs a series of passes over the execution plan, with the overall goal to produce the fewest, most efficient MSCR operations in the final optimized plan.

- **Sink Flattens**. A Flatten operation can be pushed down through consuming ParallelDo operations by duplicating the ParallelDo before each input to the Flatten.
- **Lift CombineValues** operations. If a CombineValues operation immediately follows a GroupByKey operation the CombineValues is henceforth treated as a normal ParallelDooperation and subject to ParallelDo fusion.
- **Insert fusion blocks**. If two GroupByKey operations are connected by a producer-consumer chain of one or more ParallelDo operations, the optimizer must choose which ParallelDos should fuse “up” into the output channel of the earlier GroupByKey, and which should fuse “down” into the input channel of the later GroupByKey

Overall Optimizer Strategy

The optimizer performs a series of passes over the execution plan, with the overall goal to produce the fewest, most efficient MSCR operations in the final optimized plan.

- **Sink Flattens**. A Flatten operation can be pushed down through consuming ParallelDo operations by duplicating the ParallelDo before each input to the Flatten.
- **Lift CombineValues** operations. If a CombineValues operation immediately follows a GroupByKey operation the CombineValues is henceforth treated as a normal ParallelDooperation and subject to ParallelDo fusion.
- **Insert fusion blocks**. If two GroupByKey operations are connected by a producer-consumer chain of one or more ParallelDo operations, the optimizer must choose which ParallelDos should fuse “up” into the output channel of the earlier GroupByKey, and which should fuse “down” into the input channel of the later GroupByKey
- **Fuse ParallelDos**

Overall Optimizer Strategy

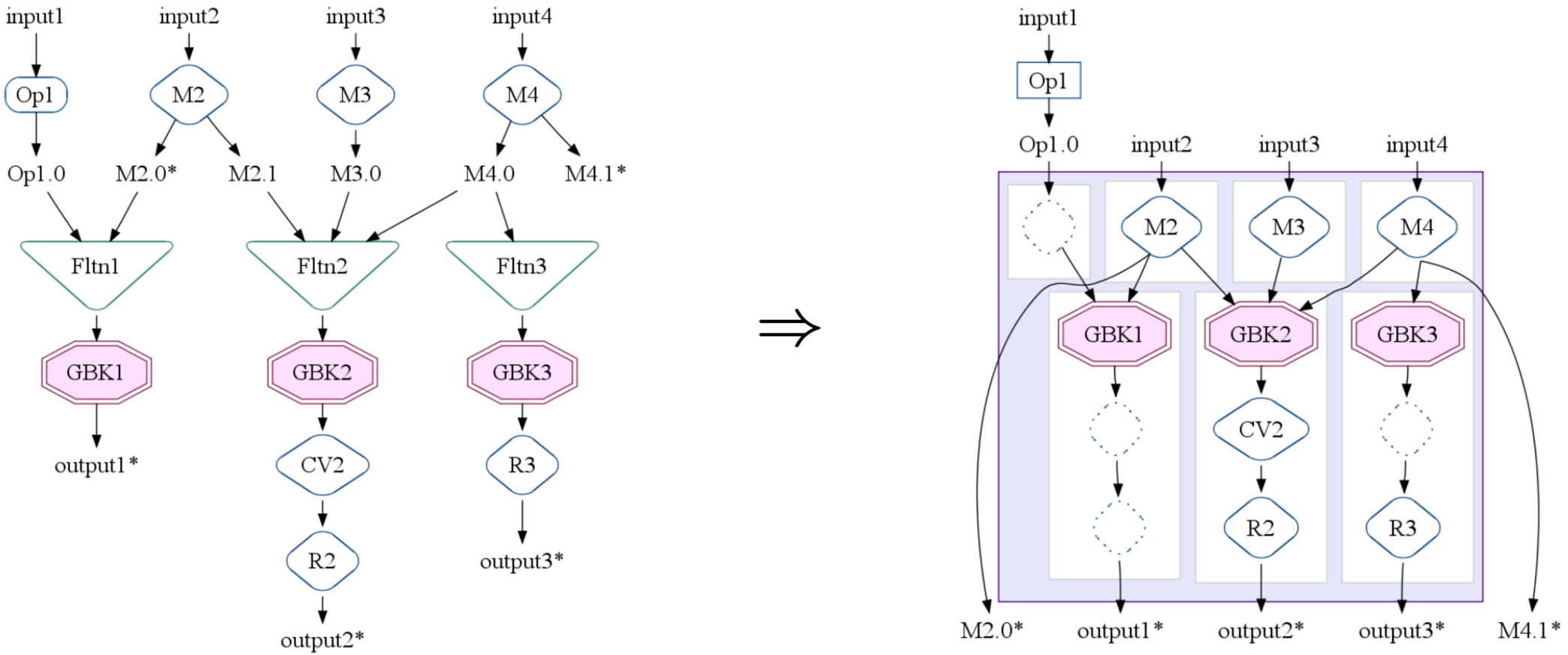
The optimizer performs a series of passes over the execution plan, with the overall goal to produce the fewest, most efficient MSCR operations in the final optimized plan.

- **Sink Flattens**. A Flatten operation can be pushed down through consuming ParallelDo operations by duplicating the ParallelDo before each input to the Flatten.
- **Lift CombineValues** operations. If a CombineValues operation immediately follows a GroupByKey operation the CombineValues is henceforth treated as a normal ParallelDooperation and subject to ParallelDo fusion.
- **Insert fusion blocks**. If two GroupByKey operations are connected by a producer-consumer chain of one or more ParallelDo operations, the optimizer must choose which ParallelDos should fuse “up” into the output channel of the earlier GroupByKey, and which should fuse “down” into the input channel of the later GroupByKey
- **Fuse ParallelDos**
- **Fuse MSCRs**

MSCR Fusion

An MSCR operation is produced from a set of related GroupByKey operations.

In turn, GroupByKey operations are related if they consume the same inputs created by the same (fused) ParallelDo operations.



The resulting MSCR operation has 4 input channels and 5 output channels

Apache Airflow

Apache
Airflow

What Airflow is...

Apache Airflow is a dataflow orchestrator to define data engineering workflows

Airflow allows to parallelize jobs, schedule them appropriately with dependencies and historically reprocess data when needed.

Airflow glues all the data engineering steps, i.e., ingest, transform, store.

What Airflow is not...

A data processing tool. No processing happens in airflow, so there's no ~~need for~~ fault tolerance

Airflow does not provide any built-in workflow versioning

Airflow does not support streaming computation (talk later)[1]

[1] The reason as to why Airflow does not support streaming is that there are no obvious behavior rules that could be set so that the airflow scheduler could deterministically check if it has been completed or not.

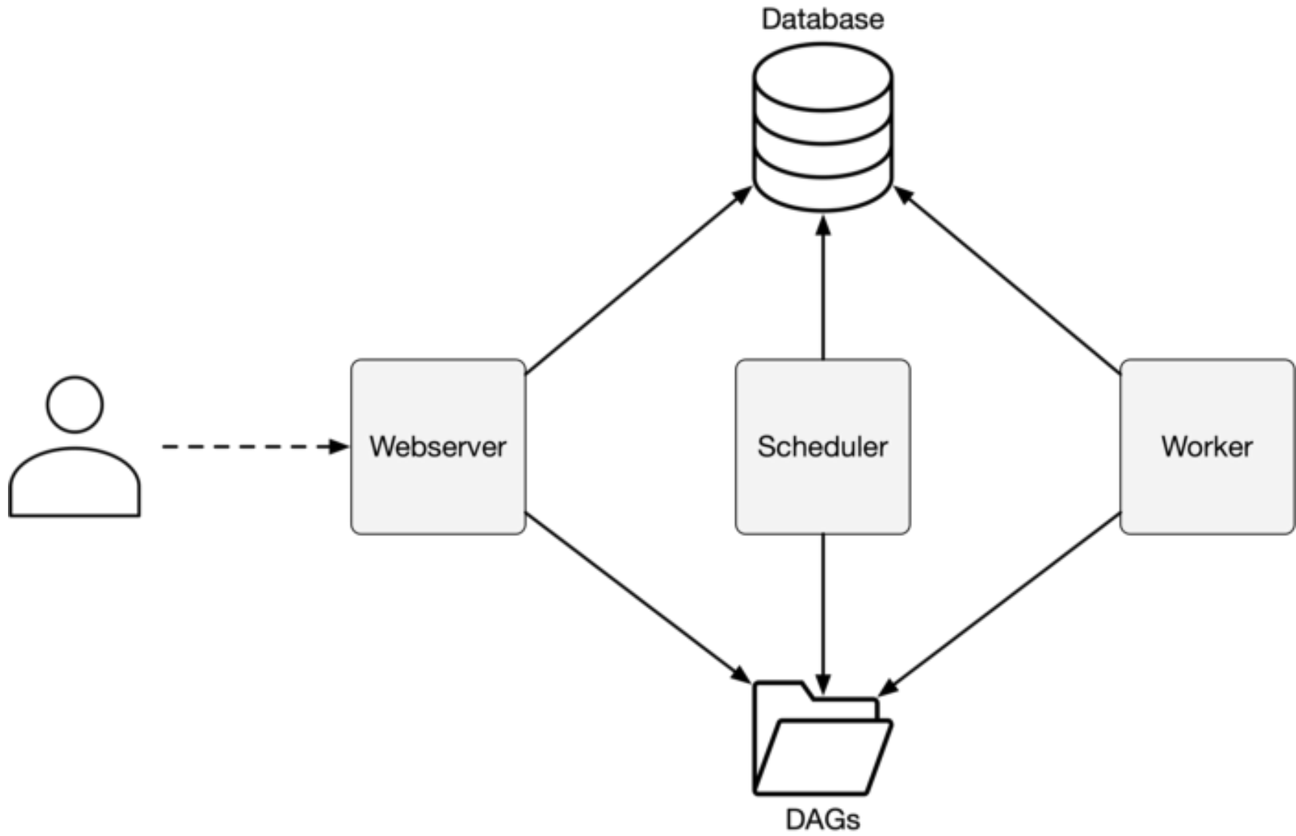
The DAG

Airflow (and almost any orchestrato), organizes the processing in a Directed Acyclic Graph (DAG). Acyclicity is required to reduce ambiguity.

Data engineering tasks are the graph nodes, Edges represent dependency relationships, e.g., if task 1 depends on task 2, then task 2 has to succeed.

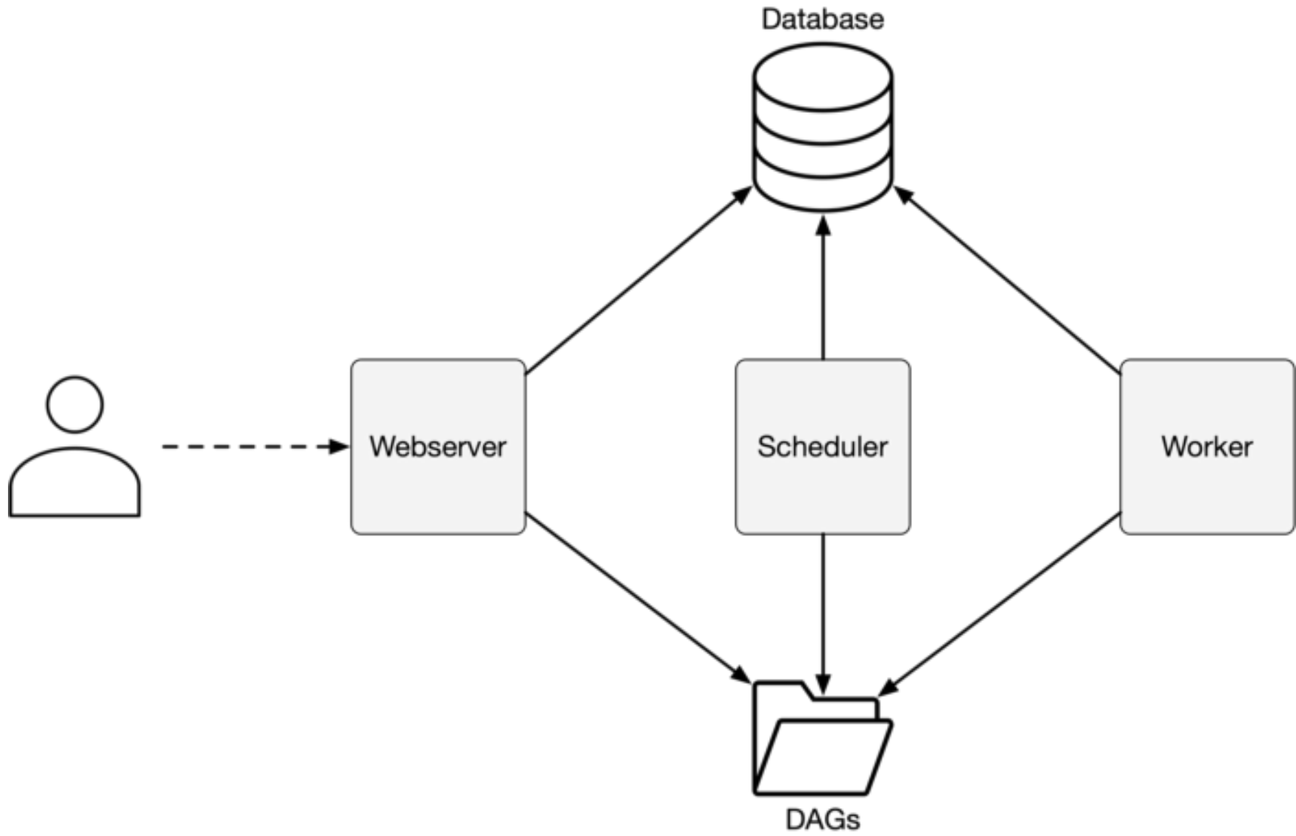
Notably, "success" doesn't have to mean that it has been succesfully completed. It can mean that it hasn't been skipped, for instance.

Airflow architecture



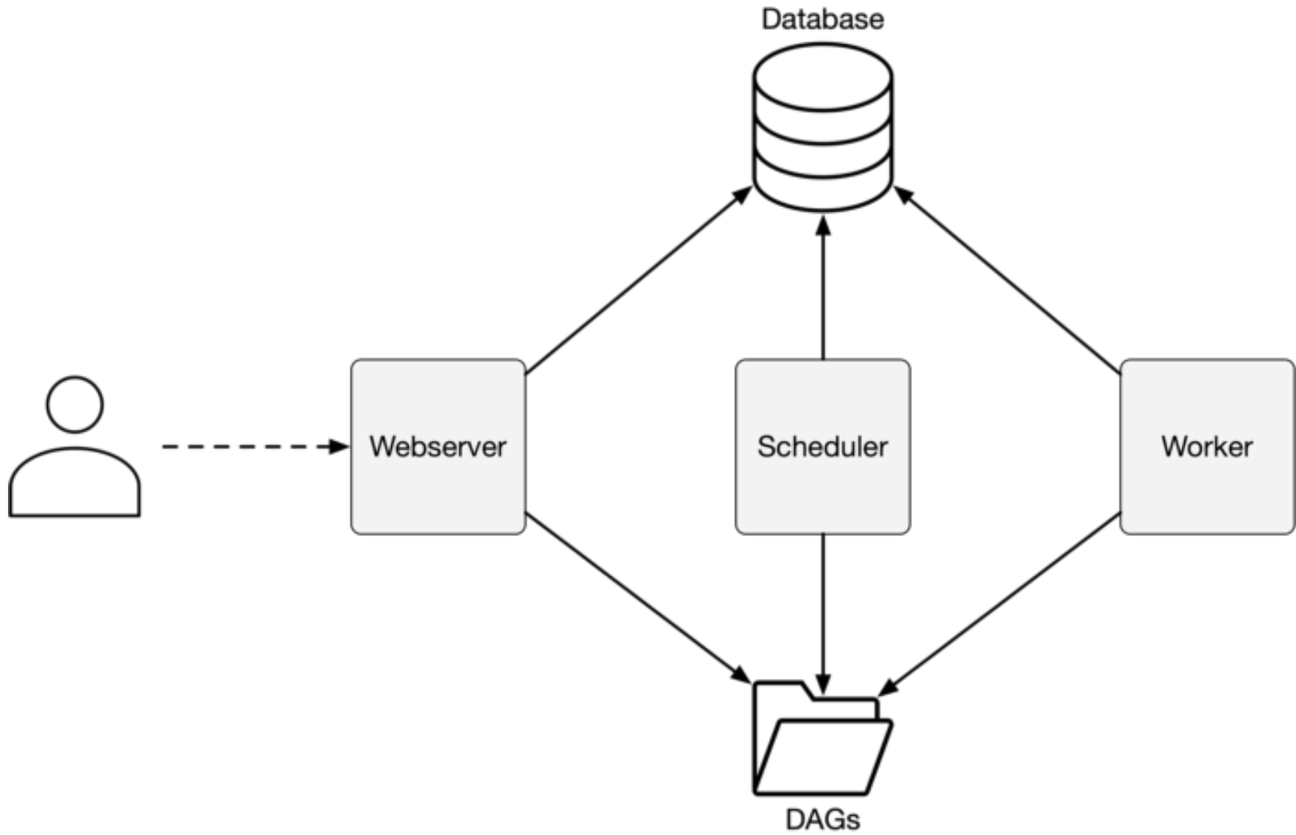
Airflow architecture

- **Web server** - nice GUI with access to the logs and most of Airflow's functionality, such as stopping and clearning DAGs.



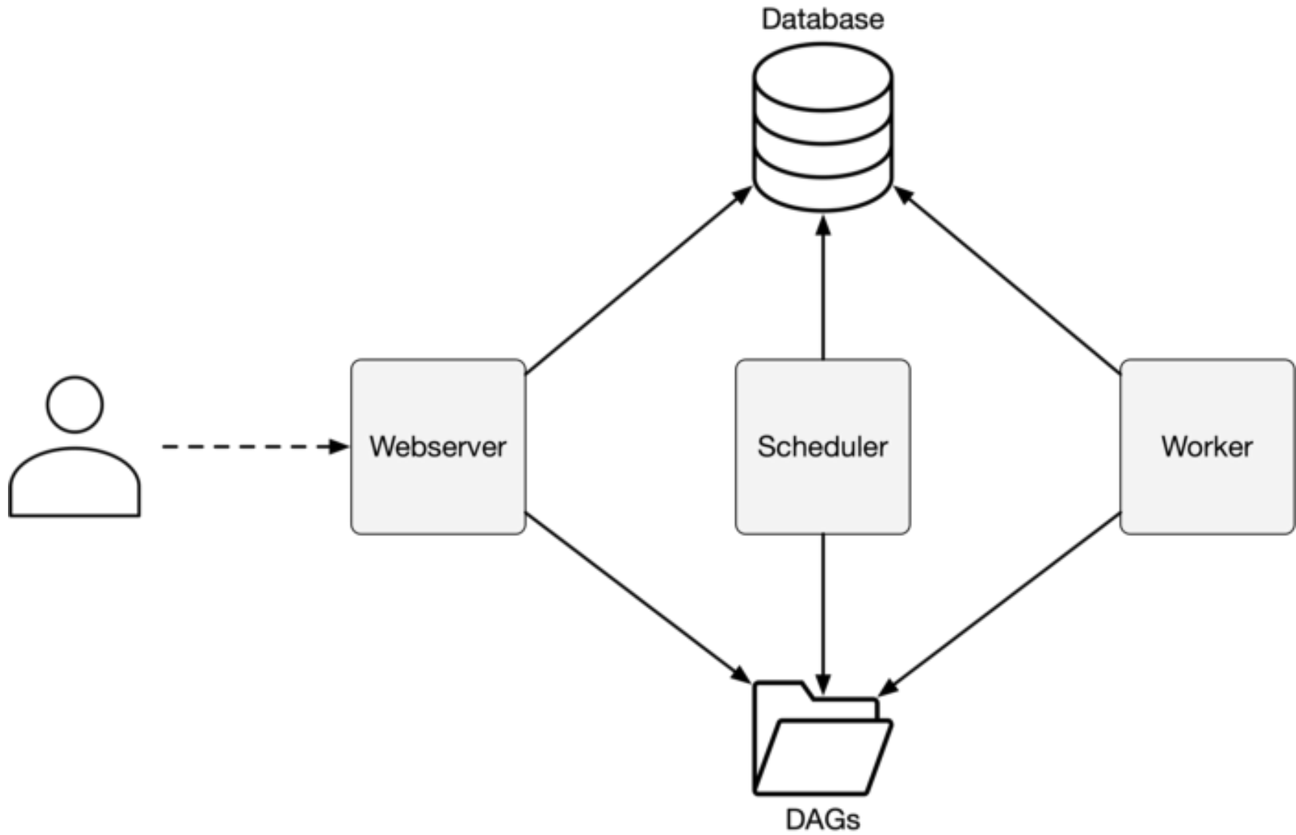
Airflow architecture

- **Web server** - nice GUI with access to the logs and most of Airflow's functionality, such as stopping and clearning DAGs.
- **Scheduler** - Puts the DAGs to action by coordinating the work to be done



Airflow architecture

- **Web server** - nice GUI with access to the logs and most of Airflow's functionality, such as stopping and clearning DAGs.
- **Scheduler** - Puts the DAGs to action by coordinating the work to be done
- **Workers** - do the job assigned by the scheduler.



Workflows

Workflows are written entirely in Python

The first thing to do is to instance the DAG, declaring a dictionary with default arguments such as *start date*, *concurrency* and *schedule interval*.

```
default_args_dict = {
    'start_date': airflow.utils.dates.days_ago(0),
    'concurrency': 1,
    'schedule_interval': None,
    'retries': 1,
    'retry_delay': datetime.timedelta(minutes=5),
}

first_dag = DAG(
    dag_id='first_dag',
    default_args=default_args_dict,
    catchup=False,
)
```

Idiosyncrasies

If you thought it would **start** at the moment of **start***date, you thought wrong~~, haha~~. The first task will actually start on start*date + schedule_interval.

In case you tried to run a DAG, for the first time, with a start date of let's say, 1 month ago, and a schedule interval of 5 minutes, then airflow would generate 8640 runs, which is the amount of 5 minute intervals within a month.

In case you didn't put `catchup=False` you will be in for a treat next time you start it, since airflow will re-run the DAG for n schedule intervals between when you turned it off and the moment of turning it on again.

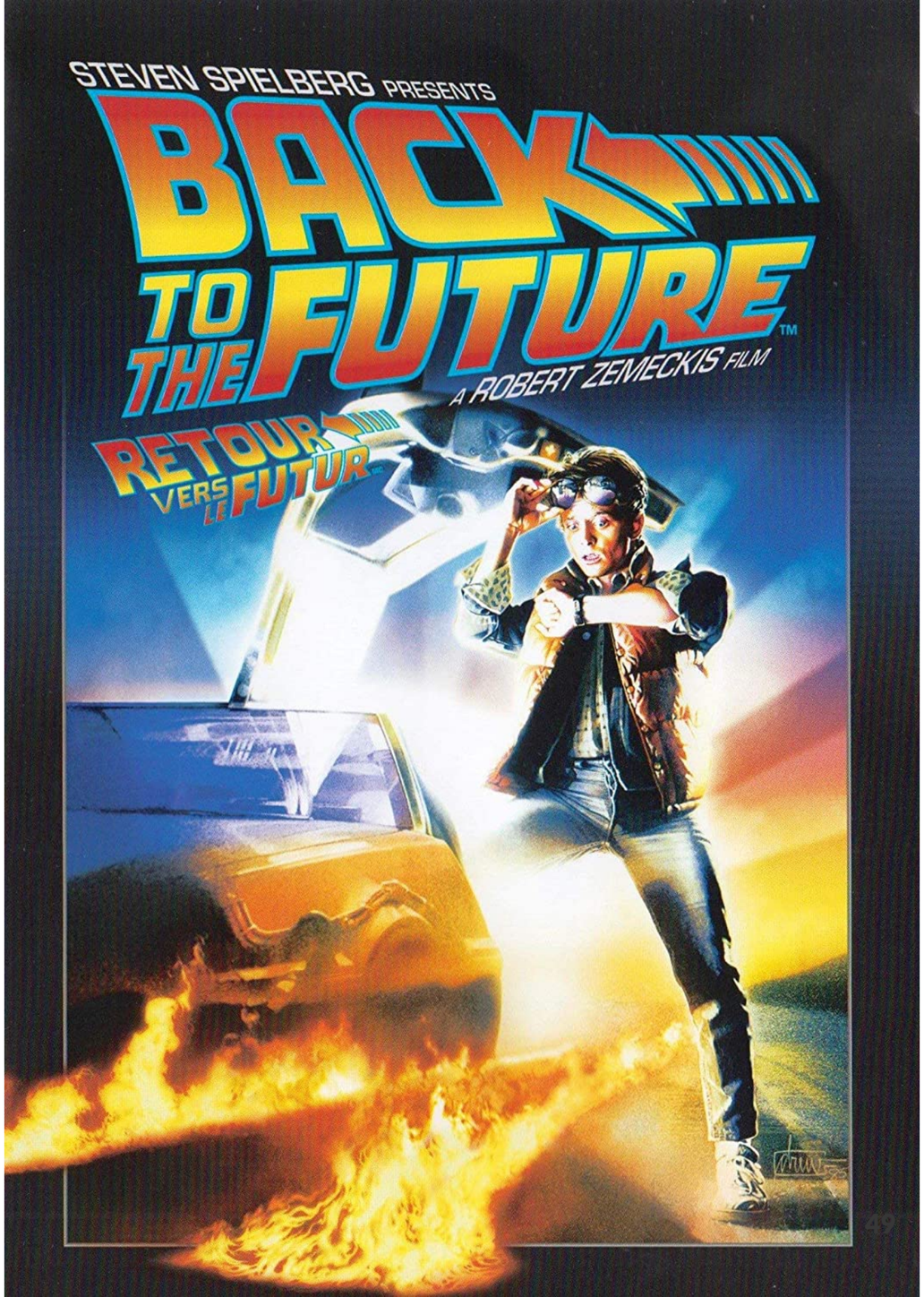
Backfilling

Often, we might desire to revisit the historical trends and movements. In such cases, we would need to compute metric and dimensions in the past,

Airflow has a recipe for Backfilling. You can call the command

```
backfill -s START_DATE -e
END_DATE dag_id
```

from the cli, and it will rerun any DAG with the specified dates, and that way you can very effortlessly satisfy your tech lead's needs.



Operators

DAGs describe **how** to run a data pipeline, operators describe **what** to do in a data pipeline.

Operators trigger data transformations, which corresponds to the **T**ransform step

Operators: Sensors, Operators, and Transfers

Operators: Sensors, Operators, and Transfers

- **Sensors**: waits for a certain time, external file, or upstream data source

Operators: Sensors, Operators, and Transfers

- **Sensors**: waits for a certain time, external file, or upstream data source
- **Operators**: triggers a certain action (e.g. run a bash command, execute a python function, or execute a Hive query, etc)

Operators: Sensors, Operators, and Transfers

- **Sensors**: waits for a certain time, external file, or upstream data source
- **Operators**: triggers a certain action (e.g. run a bash command, execute a python function, or execute a Hive query, etc)
- **Transfers**: moves data from one location to another

DummyOperator (1/2)

The simplest, yet not useless, operator that there is, is the dummy:

```
task_one = DummyOperator(
    task_id='get_spreadsheet',
    dag=first_dag,
    depends_on_past=False,
)
```

Operators are declared similarly as dags, and only require two arguments at first:

DummyOperator (1/2)

The simplest, yet not useless, operator that there is, is the dummy:

```
task_one = DummyOperator(
    task_id='get_spreadsheet',
    dag=first_dag,
    depends_on_past=False,
)
```

Operators are declared similarly as dags, and only require two arguments at first:

- `task_id` = unique id for the operator

DummyOperator (1/2)

The simplest, yet not useless, operator that there is, is the dummy:

```
task_one = DummyOperator(
    task_id='get_spreadsheet',
    dag=first_dag,
    depends_on_past=False,
)
```

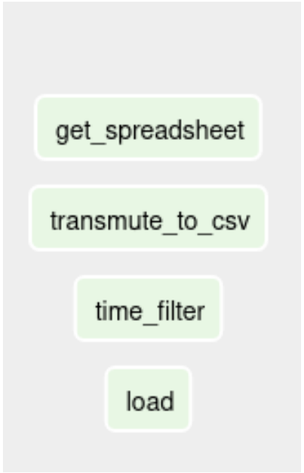
Operators are declared similarly as dags, and only require two arguments at first:

- `task_id` = unique id for the operator
- `dag` = which dag does does this operator belong to? Because there can be sub-dags, this needs to be declared explicitly.

DummyOperator (2/2)

In practice, the dummy operator does nothing, but it can be used to do two things:

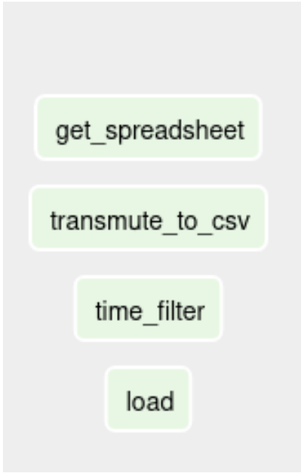
More often than not we would like to visualize an ETL process in the least time-wasting manner as possible.



DummyOperator (2/2)

In practice, the dummy operator does nothing, but it can be used to do two things:

- sketching a dag out



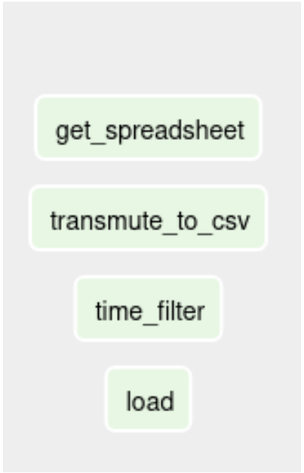
More often than not we would like to visualize an ETL process in the least time-wasting manner as possible.

DummyOperator (2/2)

In practice, the dummy operator does nothing, but it can be used to do two things:

- sketching a dag out
- trigger rule sorcery

More often than not we would like to visualize an ETL process in the least time-wasting manner as possible.



Non Dummy Operators

Most of airflow's operators function in a similar way.

Non Dummy Operators

Most of airflow's operators function in a similar way.

- `PostgreSQL` allows you to communicate with postgres instances.

Non Dummy Operators

Most of airflow's operators function in a similar way.

- `PostgreSQL` allows you to communicate with postgres instances.
- `BashOperator` allows you to run shell scripts

Non Dummy Operators

Most of airflow's operators function in a similar way.

- `PostgreSQL` allows you to communicate with postgres instances.
- `BashOperator` allows you to run shell scripts
- `PythonOperator` allows you to run Python code

BashOperator

Let us define a non-dummy operator, that runs a bash command:

```
task_one = BashOperator(
    task_id='get_spreadsheet',
    dag=first_dag,
    bash_command="curl http://www.gerbode.net/spreadsheet.xlsx --output /usr/local/airflow/data/{{ds_nodash}}.xlsx",
    trigger_rule='all_success',
    depends_on_past=False

)
```

BashOperator

Let us define a non-dummy operator, that runs a bash command:

```
task_one = BashOperator(
    task_id='get_spreadsheet',
    dag=first_dag,
    bash_command="curl http://www.gerbode.net/spreadsheet.xlsx --output /usr/local/airflow/data/{{ds_nodash}}.xlsx",
    trigger_rule='all_success',
    depends_on_past=False

)
```

- `bash_command`- in case you know bash, you have already figured it out that this works almost exactly as typing `bash -c "some command"`

BashOperator

Let us define a non-dummy operator, that runs a bash command:

```
task_one = BashOperator(
    task_id='get_spreadsheet',
    dag=first_dag,
    bash_command="curl http://www.gerbode.net/spreadsheet.xlsx --output /usr/local/airflow/data/{{ds_nodash}}.xlsx",
    trigger_rule='all_success',
    depends_on_past=False

)
```

- `bash_command`- in case you know bash, you have already figured it out that this works almost exactly as typing `bash -c "some command"`
- `depends_on_past=False` means that if it failed during the previous dag run then it won't run this time. This is important because, as we've seen, airflow can have overlapping dag runs.

Python Operator

The last part we have to go over is both the briefest and most important, the Python Operator:

Very similar to the bash operator, having `task_id`, `dag`, `trigger_rule` and `depends_on_past`, as pretty much any other operator.

Python Operator takes a function as an input `python_callable`, that is where the python part of the python operator lies at. The callable's arguments are given by the `op_kwargs` dictionary, which also allows you to template these arguments.

```
task_one = PythonOperator(
    task_id='get_spreadsheet',
    dag=second_dag,
    python_callable=_get_spreadsheet,
    op_kwargs={
        "output_folder": "/usr/local/airflow/data",
        "epoch": "{{ execution_date.int_timestamp }}",
        "url": "http://www.gerbode.net/spreadsheet.xlsx"
    },
    trigger_rule='all_success',
    depends_on_past=False,
)
```

Task Relationships

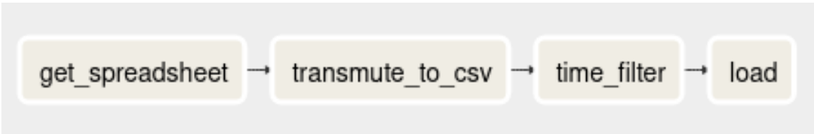
The directed relationships are given by the edges, and are very easily defined as follows, written at the end of the DAG file:

```
node_1 >> node_2 >> node_3 >> node_4
```

Looks like

Example

We want to show that the ETL process starts with `get_spreadsheet`, then its format is transmuted to a proper csv on `transmute_to_csv`, which is then filtered by time, `time_filter` to be loaded somewhere, `load`.




```
get_spreadsheet >> transmute_to_csv >> time_filter >> load
```

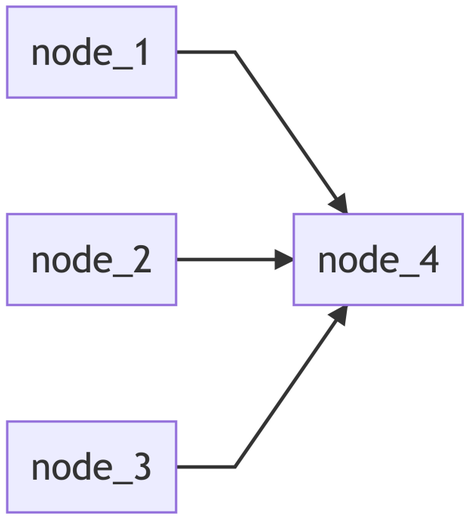
Will yield the following sequential configuration:

N-Ary Relationships

N-Ary are encoded as lists, so you can put to code that many nodes have an edge to one node as follows:

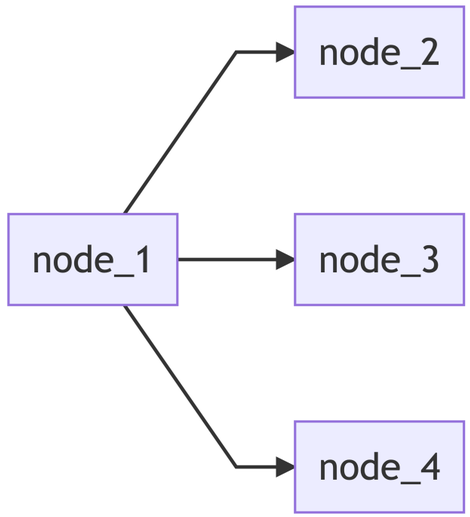
many-to-one

```
[node_1, node_2, node_3] >> node_4
```



one-to-many

```
node_1 >> [node_2, node_3, node_4]
```



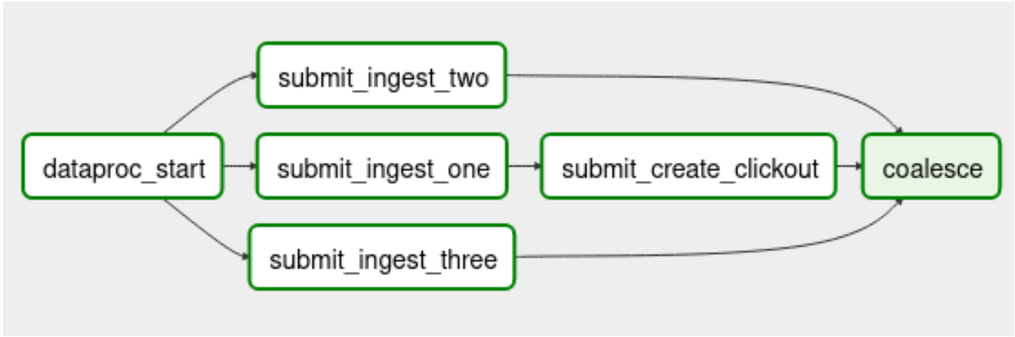
Example

Let's say that you have to create an ETL process entirely done in the cloud, and we define the following tasks:

We could define that with the following edge configuration:

```
dataproc_start >>
    [dataproc_ingest_one, dataproc_ingest_two, dataproc_ingest_three]
dataproc_ingest_one
    >> dataproc_create_clickout

[dataproc_create_clickout,
    dataproc_ingest_two,
        dataproc_ingest_three] >> end
```



Example

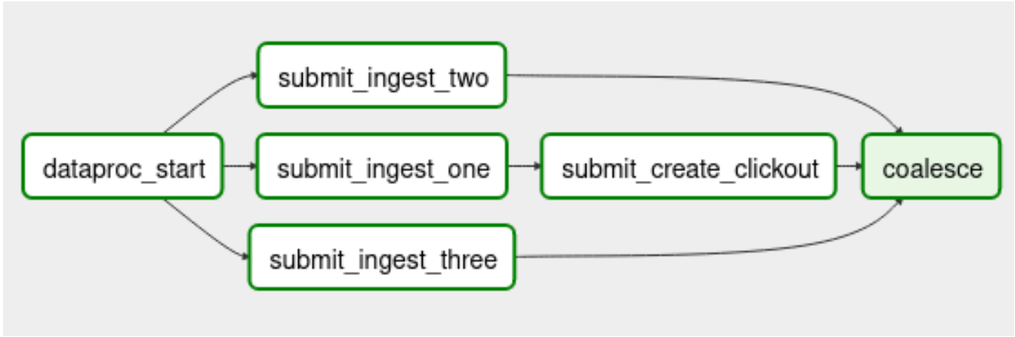
Let's say that you have to create an ETL process entirely done in the cloud, and we define the following tasks:

- `dataproc_start` - we use the google CreateDataProc operator in order to start a dataproc cluster that will be used to run pyspark scripts.

We could define that with the following edge configuration:

```
dataproc_start >>
    [dataproc_ingest_one, dataproc_ingest_two, dataproc_ingest_three]
dataproc_ingest_one
    >> dataproc_create_clickout

[dataproc_create_clickout,
    dataproc_ingest_two,
        dataproc_ingest_three] >> end
```



Example

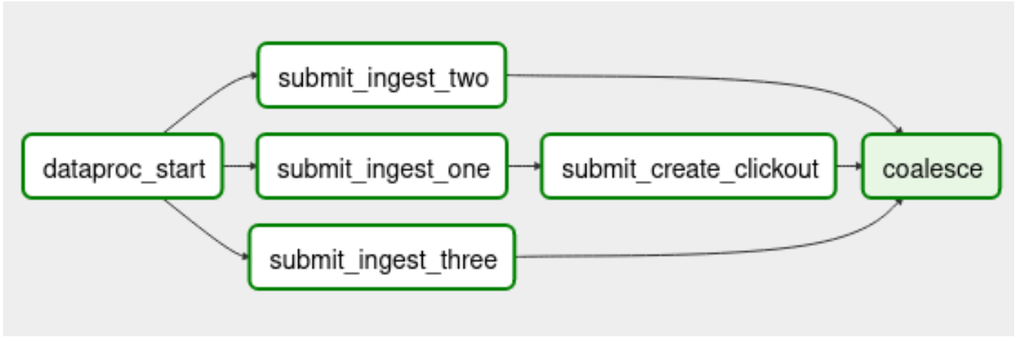
Let's say that you have to create an ETL process entirely done in the cloud, and we define the following tasks:

- `dataproc_start` - we use the google CreateDataProc operator in order to start a dataproc cluster that will be used to run pyspark scripts.
- `submit_ingest_one` - next up we create a data processing step to submit a pyspark job that will ingest data from somewhere, and will also be processed further. However, we want to make it so that if this fails, then whatever comes next fails as well.

We could define that with the following edge configuration:

```
dataproc_start >>
    [dataproc_ingest_one, dataproc_ingest_two, dataproc_ingest_three]
dataproc_ingest_one
    >> dataproc_create_clickout

[dataproc_create_clickout,
    dataproc_ingest_two,
        dataproc_ingest_three] >> end
```



Example

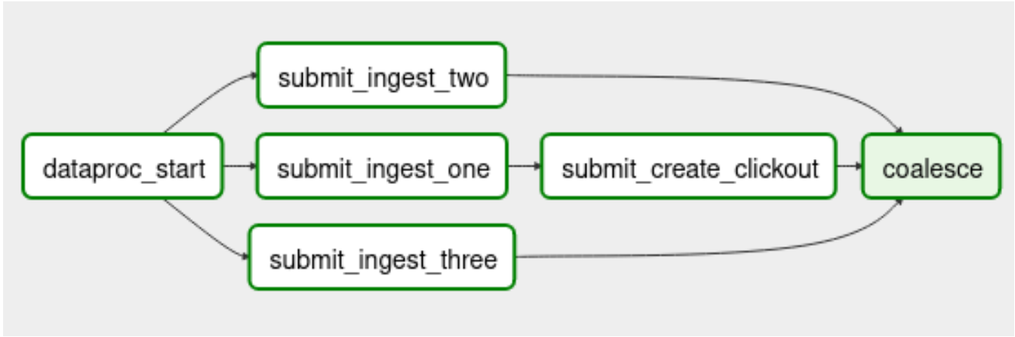
Let's say that you have to create an ETL process entirely done in the cloud, and we define the following tasks:

- `dataproc_start` - we use the google CreateDataProc operator in order to start a dataproc cluster that will be used to run pyspark scripts.
- `submit_ingest_one` - next up we create a data processing step to submit a pyspark job that will ingest data from somewhere, and will also be processed further. However, we want to make it so that if this fails, then whatever comes next fails as well.
- `submit_ingest_two, submit_ingest_three` - the same as `submit_ingest_one` except no other process depends on it

We could define that with the following edge configuration:

```
dataproc_start >>
    [dataproc_ingest_one, dataproc_ingest_two, dataproc_ingest_three]
dataproc_ingest_one
    >> dataproc_create_clickout

[dataproc_create_clickout,
    dataproc_ingest_two,
        dataproc_ingest_three] >> end
```



Example

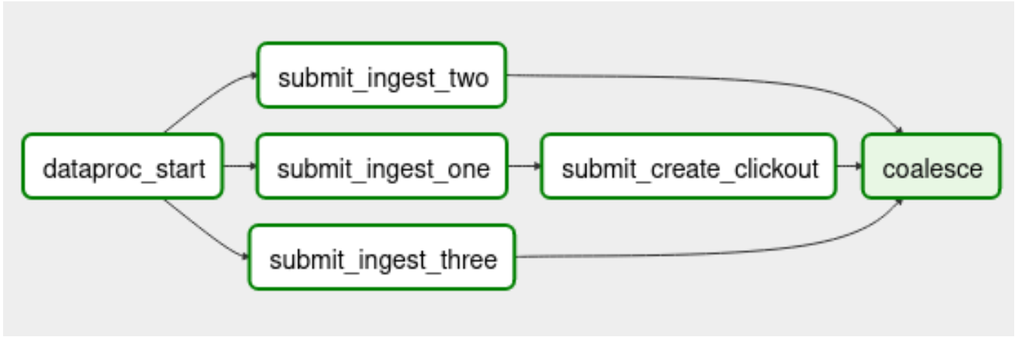
Let's say that you have to create an ETL process entirely done in the cloud, and we define the following tasks:

- `dataproc_start` - we use the google CreateDataProc operator in order to start a dataproc cluster that will be used to run pyspark scripts.
- `submit_ingest_one` - next up we create a data processing step to submit a pyspark job that will ingest data from somewhere, and will also be processed further. However, we want to make it so that if this fails, then whatever comes next fails as well.
- `submit_ingest_two, submit_ingest_three` - the same as `submit_ingest_one` except no other process depends on it
- `coalesce` - after all tasks have finished their execution, we would like to turn off the cluster.

We could define that with the following edge configuration:

```
dataproc_start >>
    [dataproc_ingest_one, dataproc_ingest_two, dataproc_ingest_three]
dataproc_ingest_one
    >> dataproc_create_clickout

[dataproc_create_clickout,
    dataproc_ingest_two,
        dataproc_ingest_three] >> end
```



Trigger Rules

Trigger rules are, in airflow-speech, a name for how should the node at the end of the edge behave when some exit status comes out of its parents.

Trigger rules can be added to any operator with the argument `trigger_rule`, and there are quite a couple for you to choose from, out of which the following are easily exemplified:

Trigger Rules

Trigger rules are, in airflow-speech, a name for how should the node at the end of the edge behave when some exit status comes out of its parents.

Trigger rules can be added to any operator with the argument `trigger_rule`, and there are quite a couple for you to choose from, out of which the following are easily exemplified:

- `all_success` - use this when you want everything to work in the most obvious way as possible, and whatever you are doing happens to be very linear.

Trigger Rules

Trigger rules are, in airflow-speech, a name for how should the node at the end of the edge behave when some exit status comes out of its parents.

Trigger rules can be added to any operator with the argument `trigger_rule`, and there are quite a couple for you to choose from, out of which the following are easily exemplified:

- `all_success` - use this when you want everything to work in the most obvious way as possible, and whatever you are doing happens to be very linear.
- `all_done` - this is very often used whenever there's a cluster start or end job. For instance, let's say that irrespective of the parent tasks status, given that they have finished their execution, you would like this task to run anyways. It's commonly used to shut down a cluster.

Trigger Rules

Trigger rules are, in airflow-speech, a name for how should the node at the end of the edge behave when some exit status comes out of its parents.

Trigger rules can be added to any operator with the argument `trigger_rule`, and there are quite a couple for you to choose from, out of which the following are easily exemplified:

- `all_success` - use this when you want everything to work in the most obvious way as possible, and whatever you are doing happens to be very linear.
- `all_done` - this is very often used whenever there's a cluster start or end job. For instance, let's say that irrespective of the parent tasks status, given that they have finished their execution, you would like this task to run anyways. It's commonly used to shut down a cluster.
- `none_failed` - is most often than not paired with the branching operator. Maybe you will need to send the data through a different path in the dag and thus it is acceptable to skip a chunk of it. If you want your dag execution to continue, you have to make it so that the parent tasks were either skipped or successfully completed.

Trigger Rules

Trigger rules are, in airflow-speech, a name for how should the node at the end of the edge behave when some exit status comes out of its parents.

Trigger rules can be added to any operator with the argument `trigger_rule`, and there are quite a couple for you to choose from, out of which the following are easily exemplified:

- `all_success` - use this when you want everything to work in the most obvious way as possible, and whatever you are doing happens to be very linear.
- `all_done` - this is very often used whenever there's a cluster start or end job. For instance, let's say that irrespective of the parent tasks status, given that they have finished their execution, you would like this task to run anyways. It's commonly used to shut down a cluster.
- `none_failed` - is most often than not paired with the branching operator. Maybe you will need to send the data through a different path in the dag and thus it is acceptable to skip a chunk of it. If you want your dag execution to continue, you have to make it so that the parent tasks were either skipped or successfully completed.
- `all_failed` - ~~pair this one up with a callback so you can notify yourself of getting fired~~