Understanding Spark: An Overview

Xavier Morera

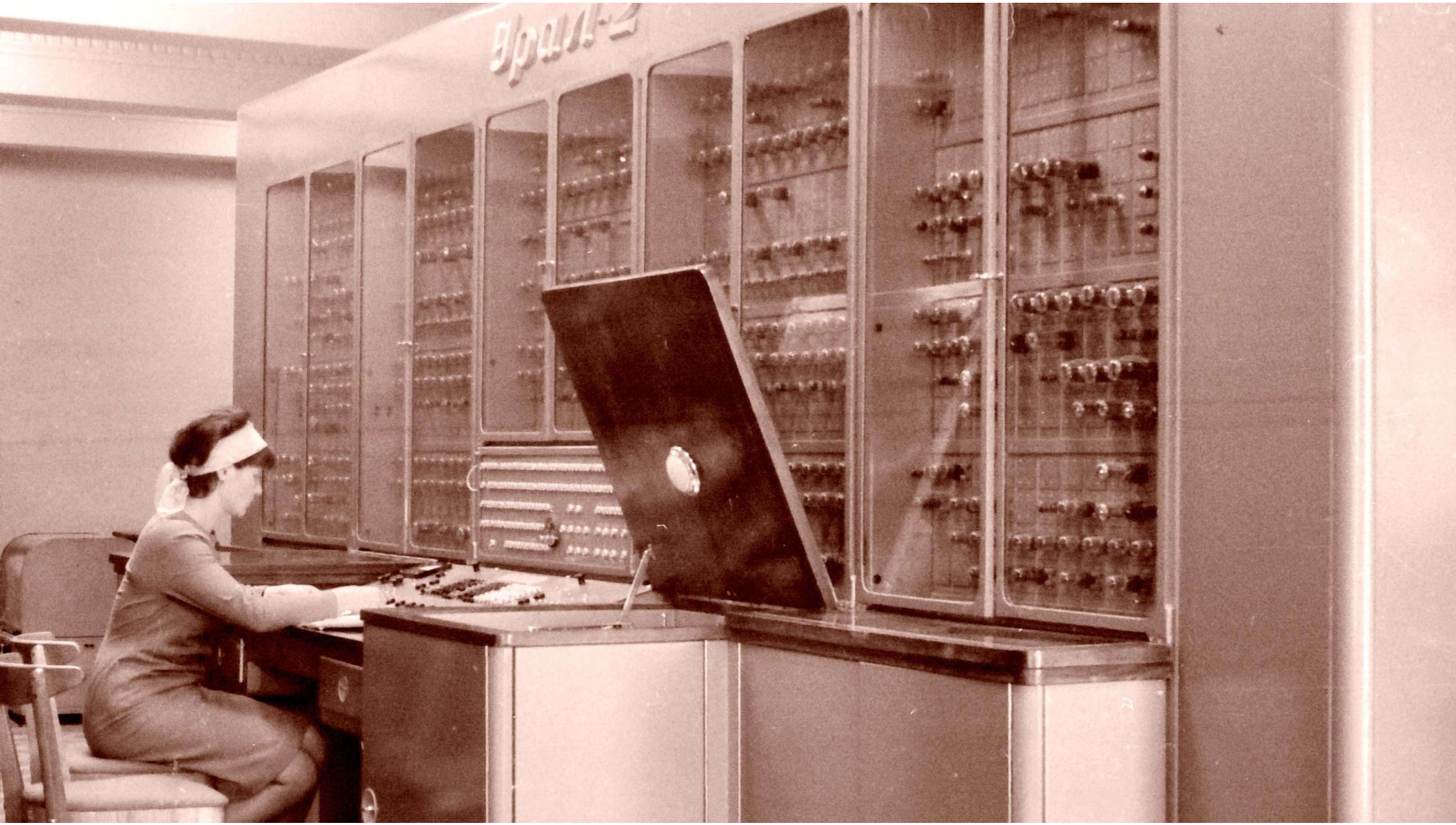
HELPING DEVELOPERS UNDERSTAND SEARCH & BIG DATA

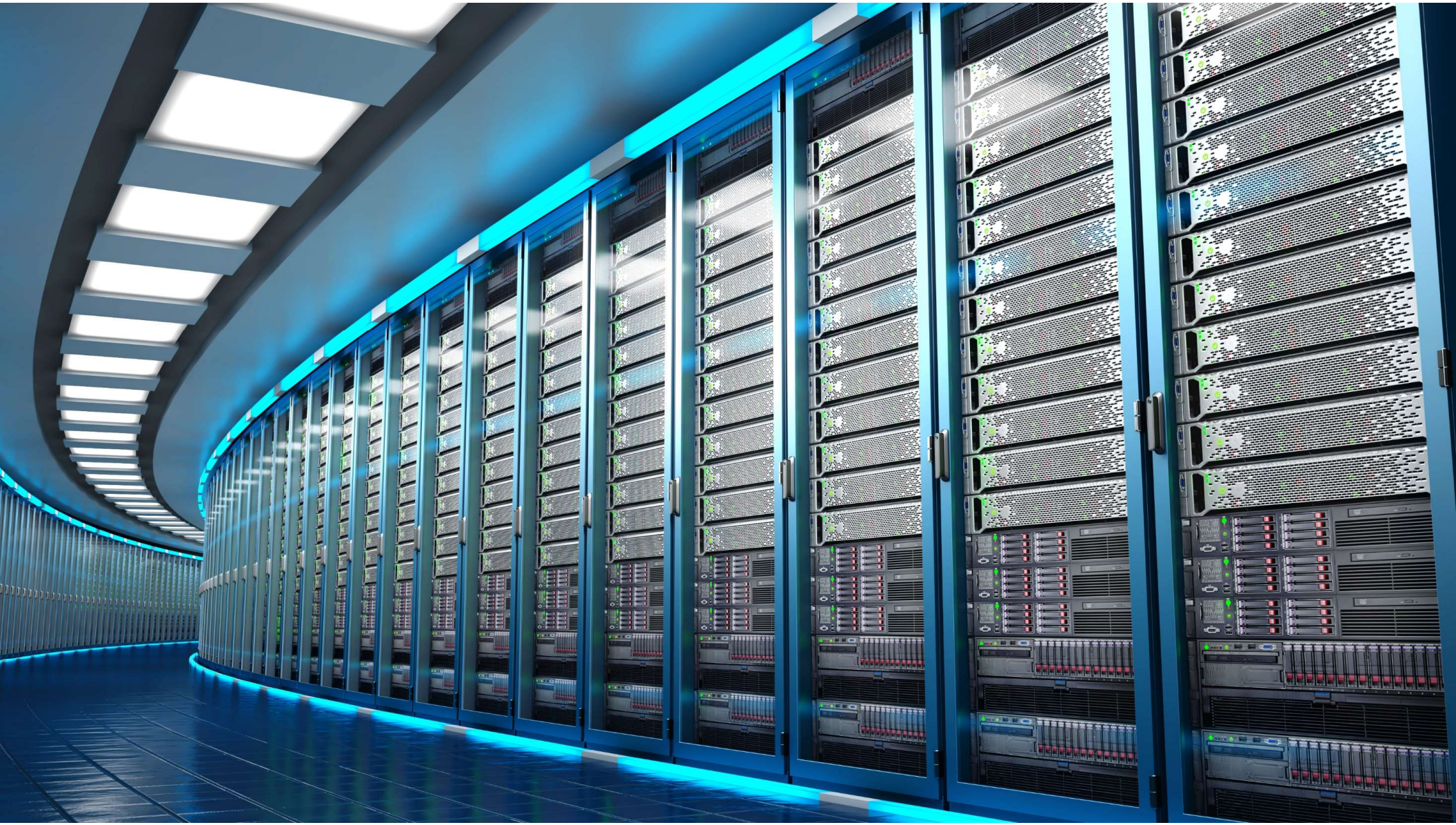
@xmorera www.xaviermorera.com

Understanding

An Overview

Lightning-fast cluster computing


Ural-2



Understanding Spark: An Overview

Lightning-fast cluster computing

StackOverflow Posts

PostId	Tag	User	Title
1	apache-spark	1	How can I use DataFrames in 2.0?
2	apache-spark	2	What is an RDD & Schema RDD
3	sql	1	How do I group by a field?
4	hive	3	Can I use Hive from HUE?

```
> select count(*) from posts
4
```

Counting Number of Records

Easy operation in SQL

Use the count function

```
> select distinct(tag) from posts
apache-spark
sql
hive
```

Which Tags Are Used in the Posts

Another easy operation in SQL

Use the distinct function

StackOverflow Posts

PostId	Tag	User	Title
1	[apache-spark]	1	How can I use DataFrames in 2.0?
2	[apache-spark]	2	What is an RDD & Schema RDD
3	[sql]	1	How do I group by a field?
4	[hive]	3	Can I use Hive from HUE?

```
> Update Posts set Tags = '[apache-spark,sql]' where
PostId=1

1 row affected
```

Fine Grained Transformation

Update 1 record

This is a ‘fine grained’ update

```
> Update posts set tags='[apache-spark,sql'] where postid=1

1 row affected
```

Updating a Tag

Update 1 cell in 1 row in 1 table in 1 database

This is called a fine grained transformation

StackOverflow Posts

PostId	Tag	User	Title
1	[apache-spark]	1	How can I use DataFrames in 2.0?
2	[apache-spark]	2	What is an RDD & Schema RDD
3	[sql]	1	How do I group by a field?
4	[hive]	3	Can I use Hive from HUE?

Caveat #1: Scalability

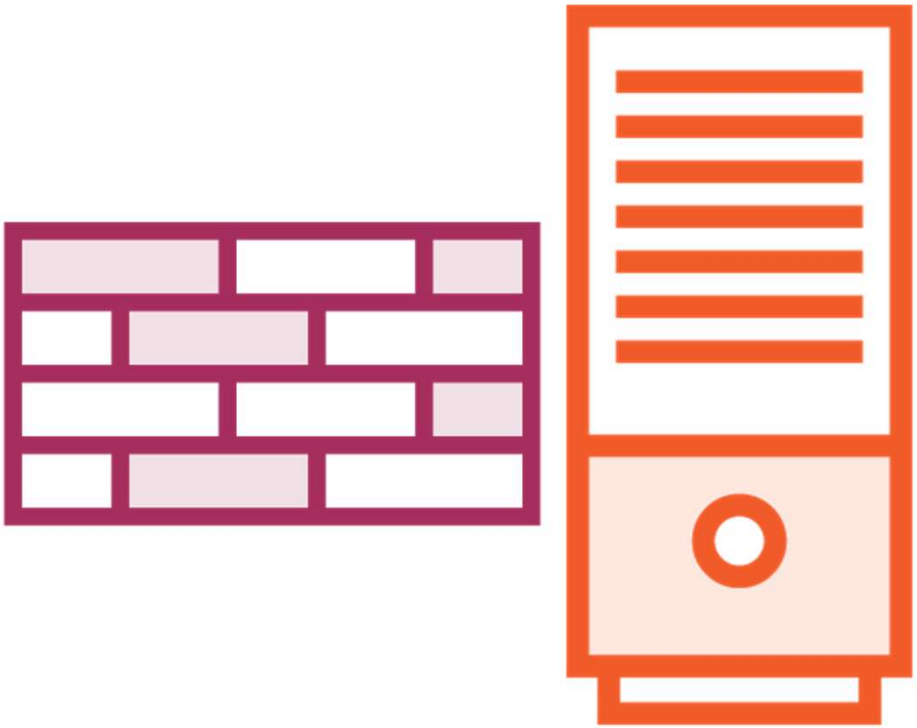
How can I use DataFrames in 2.0?

What is an RDD & Schema RDD

How do I group by a field?

Can I use Hive from HUE?

Caveat #1: Scalability



How can I use DataFrames in 2.0?

What is an RDD & Schema RDD

How do I group by a field?

Can I use Hive from HUE?

Caveat #1: Scalability

How can I use DataFrames in 2.0?

What is an RDD & Schema RDD

How do I group by a field?

Can I use Hive from HUE?

Caveat #2: Transformations

PostId	Tag	User	Title
1	[apache-spark]	1	How can I use DataFrames in 2.0?
2	[apache-spark]	2	What is an RDD & Schema RDD
3	[sql]	1	How do I group by a field?
4	[hive]	3	Can I use Hive from HUE?

Word Count

How can I use DataFrames in 2.0?

What is an RDD & Schema RDD

How do I group by a field?

Can I use Hive from HUE?

(I, 3)

(RDD, 2)

(DataFrames, 1)

(HUE, 1)

...

Search with Typeahead

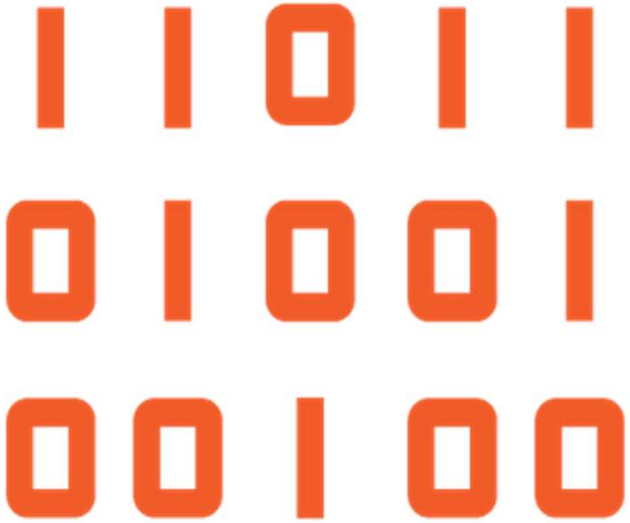
```
using System.Collections.Concurrent;

ConcurrentDictionary<string, int> words;
```

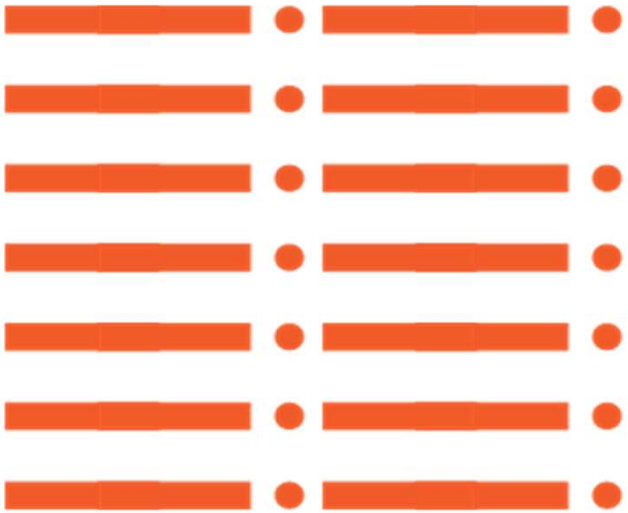
Count Occurrences of Each Word

Not exactly practical nor good for our sanity

Issues on a large dataset

Apache Spark





Word Count

How can I use DataFrames in 2.0

What is an RDD and Schema RDD

How do I group by a field

Can I use Hive from HUE

(I, 3)

(RDD, 2)

(DataFrames, 1)

(HUE, 1)

...

RDD

Resilient Distributed Dataset

How can I use DataFrames in 2.0

What is an RDD and Schema RDD

How do I group by a field

Can I use Hive from HUE

Read Data

Step 1

How can I use DataFrames in 2.0

What is an RDD and Schema RDD

How do I group by a field

Can I use Hive from HUE

Split

Step 2

How can I use DataFrames in 2.0	[How,can,I,use,DataFrames,in,2.0]
What is an RDD and Schema RDD	[What,is,an,RDD,and,Schema,RDD]
How do I group by a field	[How,do,I,group,by,a,field]
Can I use Hive from HUE	[Can,I,use,Hive,from,HUE]

Coarse Grained Transformations

Operations that can be applied to the whole dataset, like map, filter, group, reduce among others

Provides powerful distributed data processing capabilities

FlatMap

Step 3

[How,can,I,use,DataFrames,in,2.0]

[What,is,an,RDD,and,Schema,RDD]

[How,do,I,group,by,a,field]

[Can,I,use,Hive,from,HUE]

**[How,can,I,use,DataFrames,in,2.0,
What,is,an,RDD,and,Schema,RDD,
How,do,I,group,by,a,field,
Can,I,use,Hive,from,HUE]**

Map

Step 4

[How,can,I,use,DataFrames,in,2.0,
What,is,an,RDD,and,Schema,RDD,
How,do,I,group,by,a,field,
Can,I,use,Hive,from,HUE]

(How,1)
(can,1)
(I,1)
(use,1)
(DataFrames,1)
(in,1)
(2.0,1)
(What,1)
(is,1)
(an,1)
(RDD,1)
(and,1)
(Schema,1)
(RDD,1)
(How,1)
(do,1)
(I,1)
(group,1)
(by,1)
(a,1)
(field,1)
(Can,1)
(I,1)
(use,1)
(Hive,1)
(from,1)
(HUE,1)

ReduceByKey

Step 5

(How,1)
(can,1)
(I,1)
(use,1)
(DataFrames,1)
(in,1)
(2.0,1)
(What,1)
(is,1)
(an,1)
(RDD,1)
(and,1)
(Schema,1)

(RDD,1)
(How,1)
(do,1)
(I,1)
(group,1)
(by,1)
(a,1)
(field,1)
(Can,1)
(I,1)
(use,1)
(Hive,1)
(from,1)
(HUE,1)

(and,1)
(a,1)
(use,2)
(group,1)
(I,3)
(is,1)
(DataFrames,1)
(What,1)
(Hive,1)
(from,1)
(do,1)

(HUE,1)
(an,1)
(How,2)
(Schema,1)
(Can,1)
(in,1)
(field,1)
(2.0,1)
(by,1)
(RDD,2)
(can,1)

```
val lines = sc.textFile("file:///se/simple_titles.txt")
val words = lines.flatMap(line => line.split(" "))
val word_for_count = words.map(x => (x,1))
word_for_count.reduceByKey(_ + _).collect()
```

Count Occurrences of Each Word with Spark

Don’t pay too much attention to my commands now

But note how simple it is to apply to a full dataset

```
sc.textFile("file:///se/simple_titles.txt").flatMap(line =>
line.split(" ")).map(x => (x,1)).reduceByKey((x,y) => (x +
y)).collect()
```

Could Even Be Done in a Single Line

Don’t pay too much attention to my commands now

Can 9 Pregnant Women Have a Baby in 1 Month?

One at a Time

[How,can,I,use,DataFrames,in,2.0]

What is an RDD and Schema RDD

How do I group by a field

Can I use Hive from HUE

Partitioning Data

How can I use DataFrames in 2.0

What is an RDD and Schema RDD

How do I group by a field

Can I use Hive from HUE

RDD

How can I use DataFrames in 2.0

What is an RDD and Schema RDD

How do I group by a field

Can I use Hive from HUE

Resilient
Distributed
Dataset

Parallel Execution

[How,can,I,use,DataFrames,in,2.0]

What is an RDD and Schema RDD

[How,do,I,group,by,a,field]

Can I use Hive from HUE

Closures

Code and variables required to execute a computation in each distributed node over a partition of data

Closures

```
def get_word_for_count(lines)
    words = line.split(' ')
    wc = words.map(lambda x: (x,1))
    return wc
```

How can I use DataFrames in 2.0

What is an RDD and Schema RDD

How do I group by a field

Can I use Hive from HUE

Parallel Execution

How can I use DataFrames in 2.0	[How,can,I,use,DataFrames,in,2.0]
What is an RDD and Schema RDD	[What,is,an,RDD,and,Schema,RDD]
How do I group by a field	[How,do,I,group,by,a,field]
Can I use Hive from HUE	[Can,I,use,Hive,from,HUE]

Parallel Execution

How can I use DataFrames in 2.0

[How,can,I,use,DataFrames,in,2.0]

(How,1) (can,1) (I,1) (use,1)
(DataFrames,1) (in,1) (2.0,1)

Parallel Execution

How can I use DataFrames in 2.0

What is an RDD and Schema RDD

How do I group by a field

Can I use Hive from HUE

Parallel Execution

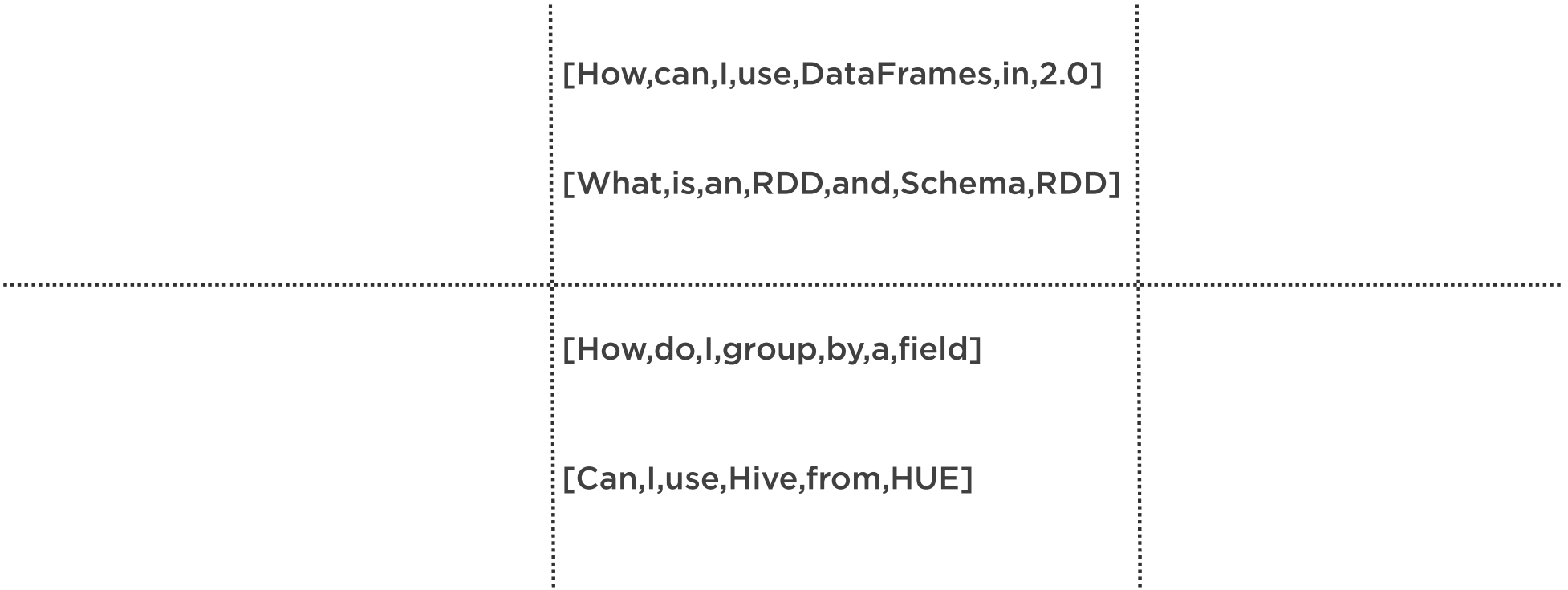
What is an RDD and Schema RDD

[How,can,I,use,DataFrames,in,2.0]

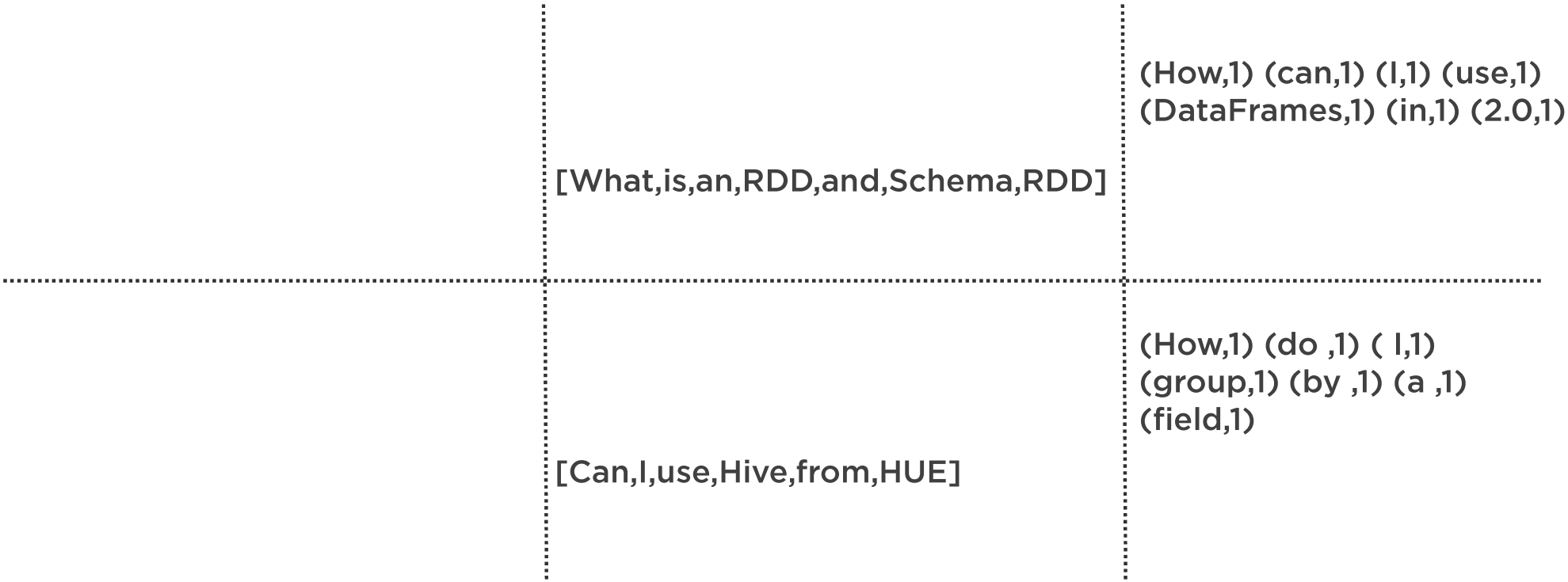
Can I use Hive from HUE

[How,do,I,group,by,a,field]

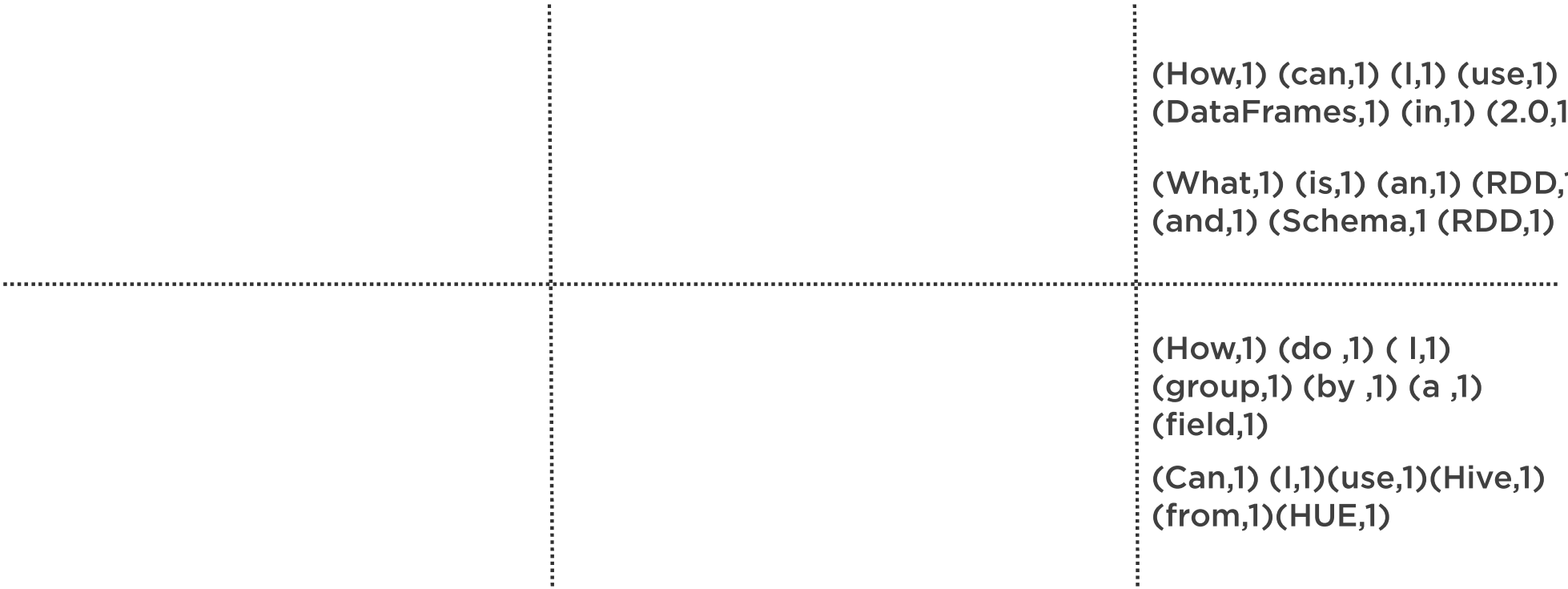
Parallel Execution



Parallel Execution



Parallel Execution



Pipeline

How can I use DataFrames in 2.0

What is an RDD and Schema RDD

How do I group by a field

Can I use Hive from HUE

Pipeline

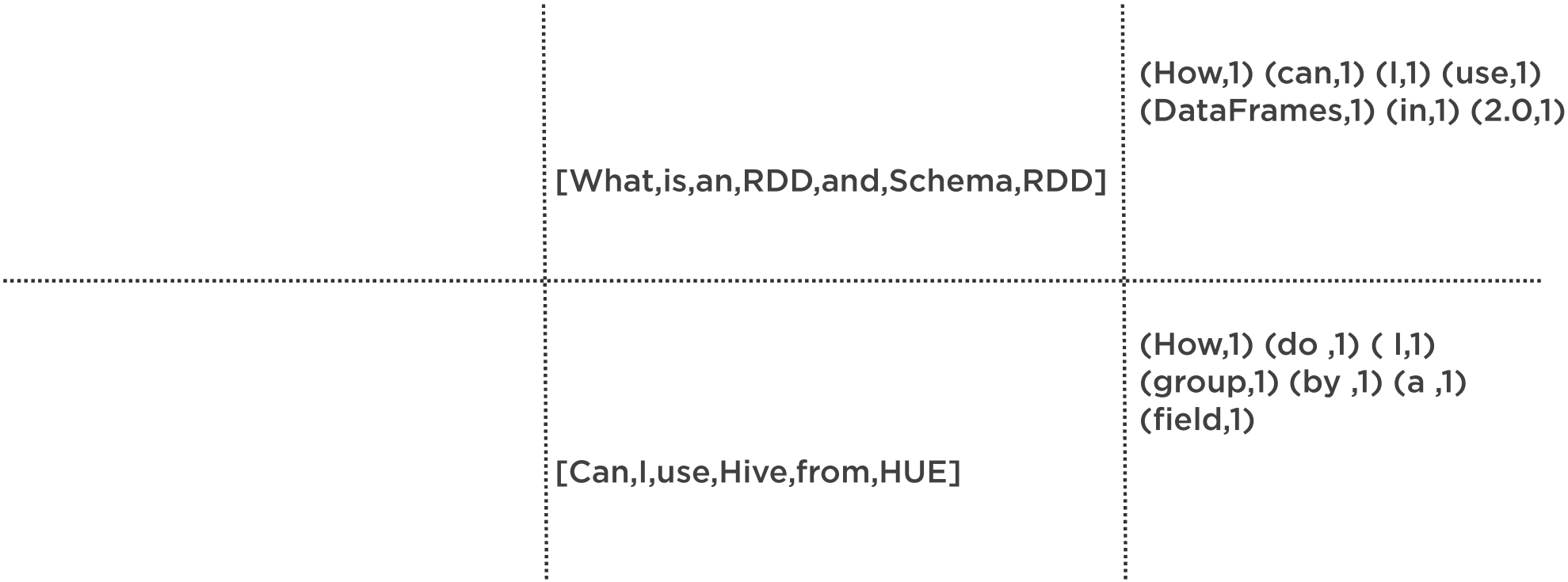
What is an RDD and Schema RDD

[How,can,I,use,DataFrames,in,2.0]

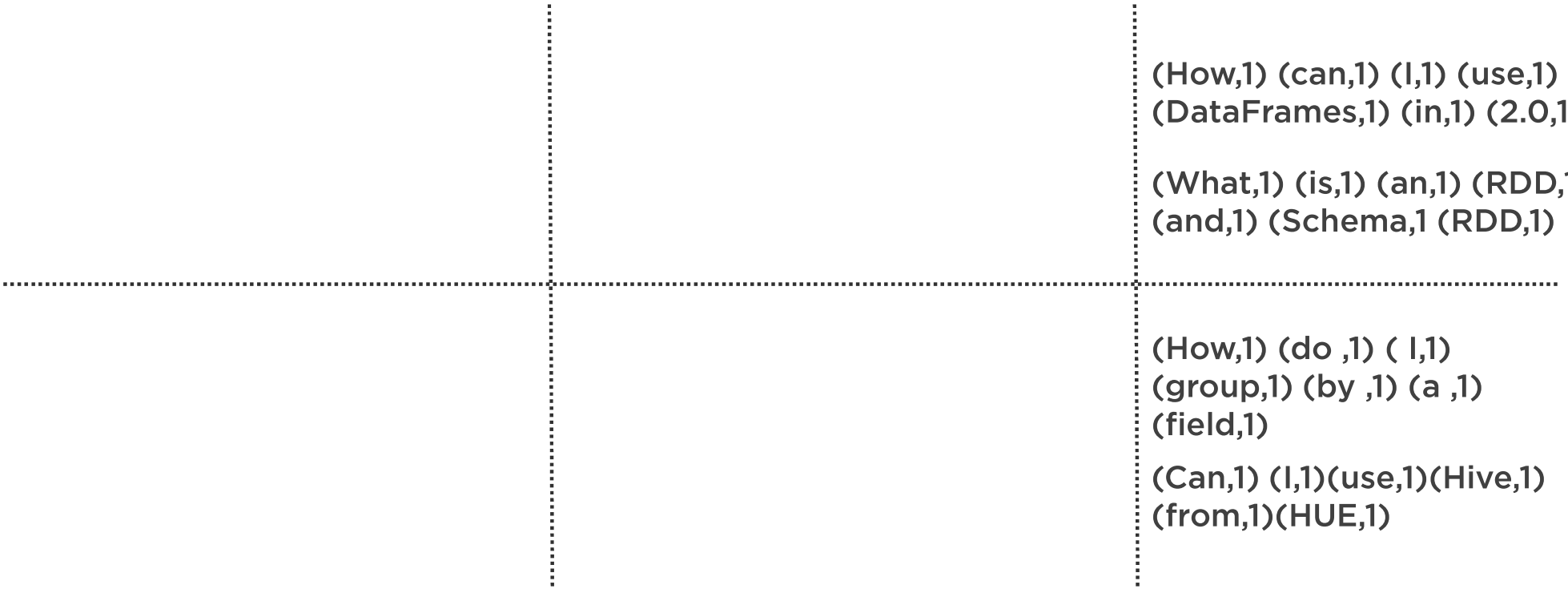
Can I use Hive from HUE

[How,do,I,group,by,a,field]

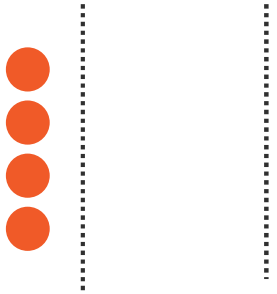
Pipeline


[What,is,an,RDD,and,Schema,RDD]
(How,1) (can,1) (I,1) (use,1)
(DataFrames,1) (in,1) (2.0,1)
[Can,I,use,Hive,from,HUE]
(How,1) (do ,1) (I,1)
(group,1) (by ,1) (a ,1)
(field,1)

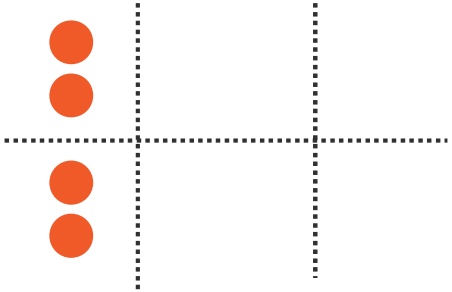
Pipeline


(How,1) (can,1) (I,1) (use,1)
(DataFrames,1) (in,1) (2.0,1)
(What,1) (is,1) (an,1) (RDD,1)
(and,1) (Schema,1 (RDD,1)
(How,1) (do ,1) (I,1)
(group,1) (by ,1) (a ,1)
(field,1)
(Can,1) (I,1)(use,1)(Hive,1)
(from,1)(HUE,1)

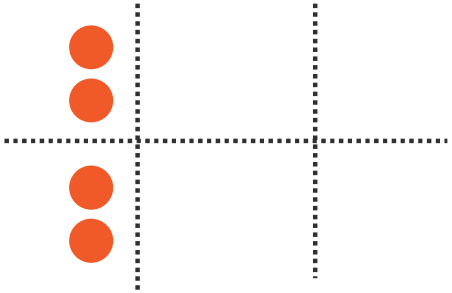
Comparison



One at a time

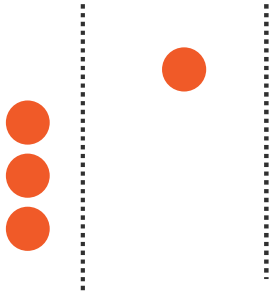


Parallel

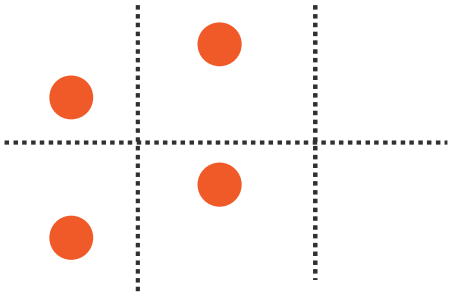


Pipeline

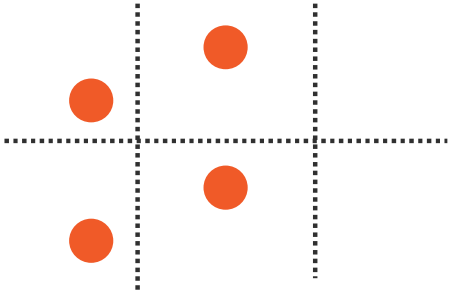
Comparison



One at a time

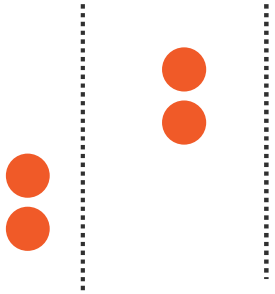


Parallel

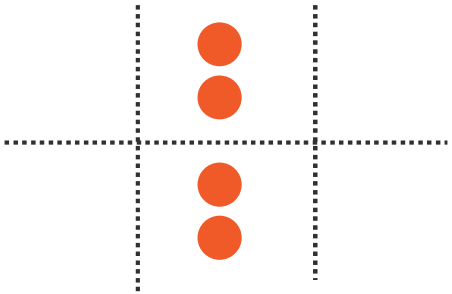


Pipeline

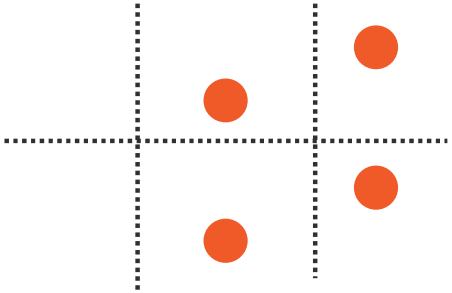
Comparison



One at a time

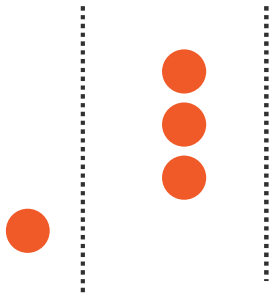


Parallel

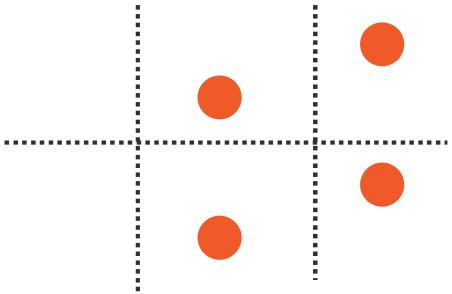


Pipeline

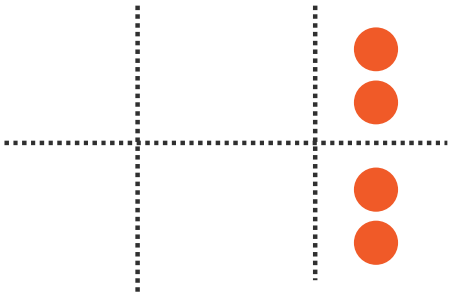
Comparison



One at a time

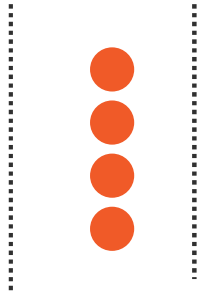


Parallel

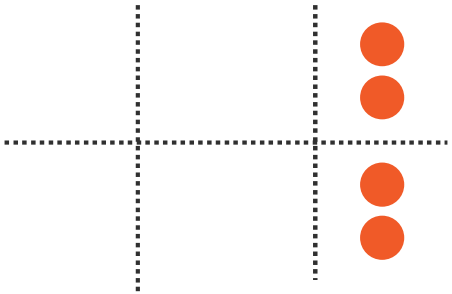


Pipeline

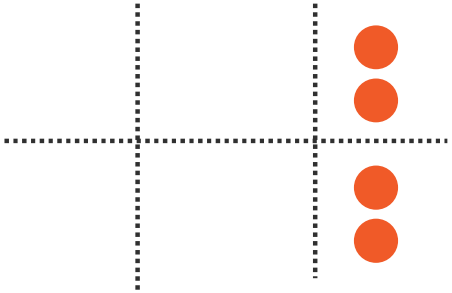
Comparison



One at a time

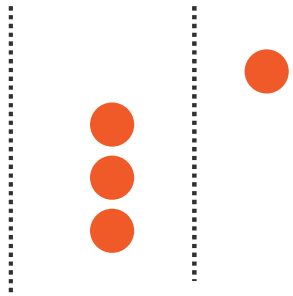


Parallel

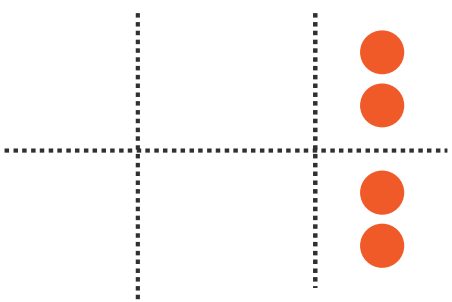


Pipeline

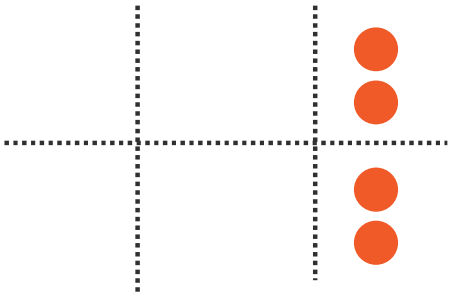
Comparison



One at a time

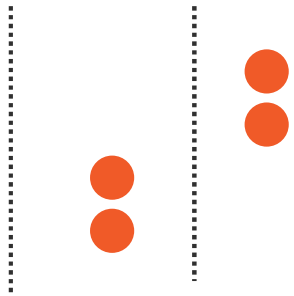


Parallel

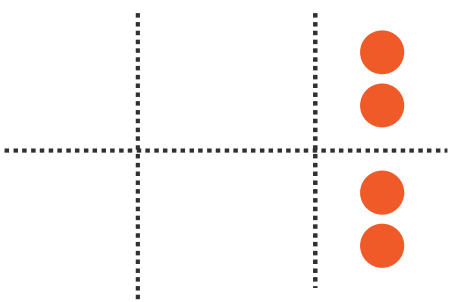


Pipeline

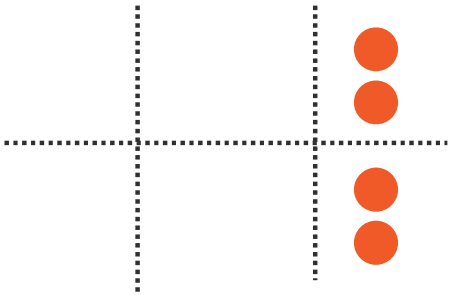
Comparison



One at a time

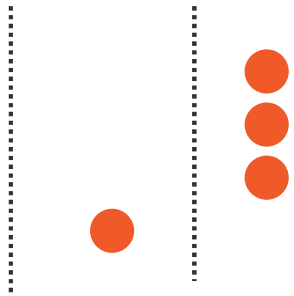


Parallel

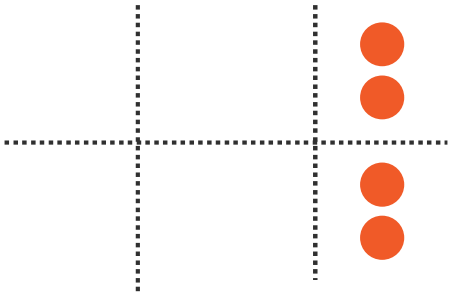


Pipeline

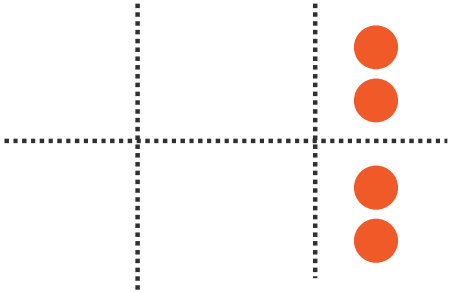
Comparison



One at a time

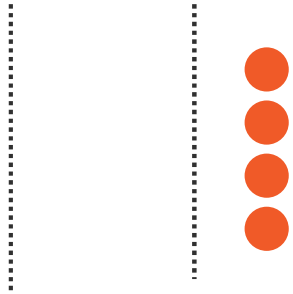


Parallel

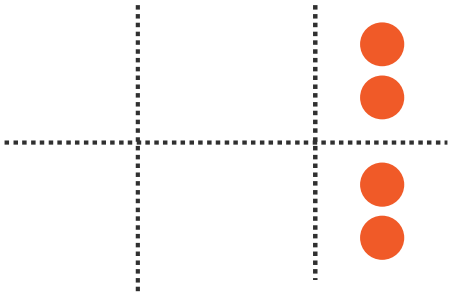


Pipeline

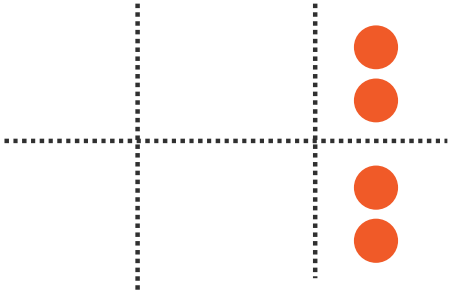
Comparison



One at a time



Parallel



Pipeline

Narrow vs. Wide Transformations

[How,can,I,use,DataFrames,in,2.0]

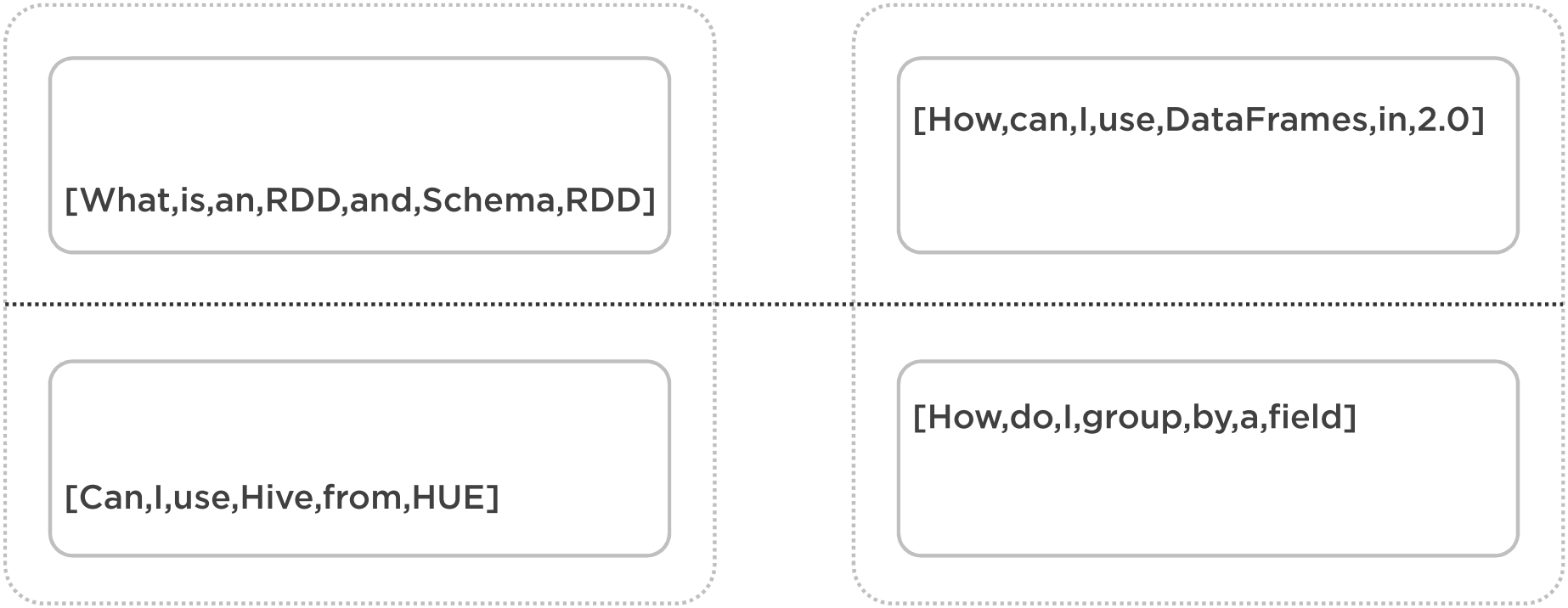
[What,is,an,RDD,and,Schema,RDD]

[How,do,I,group,by,a,field]

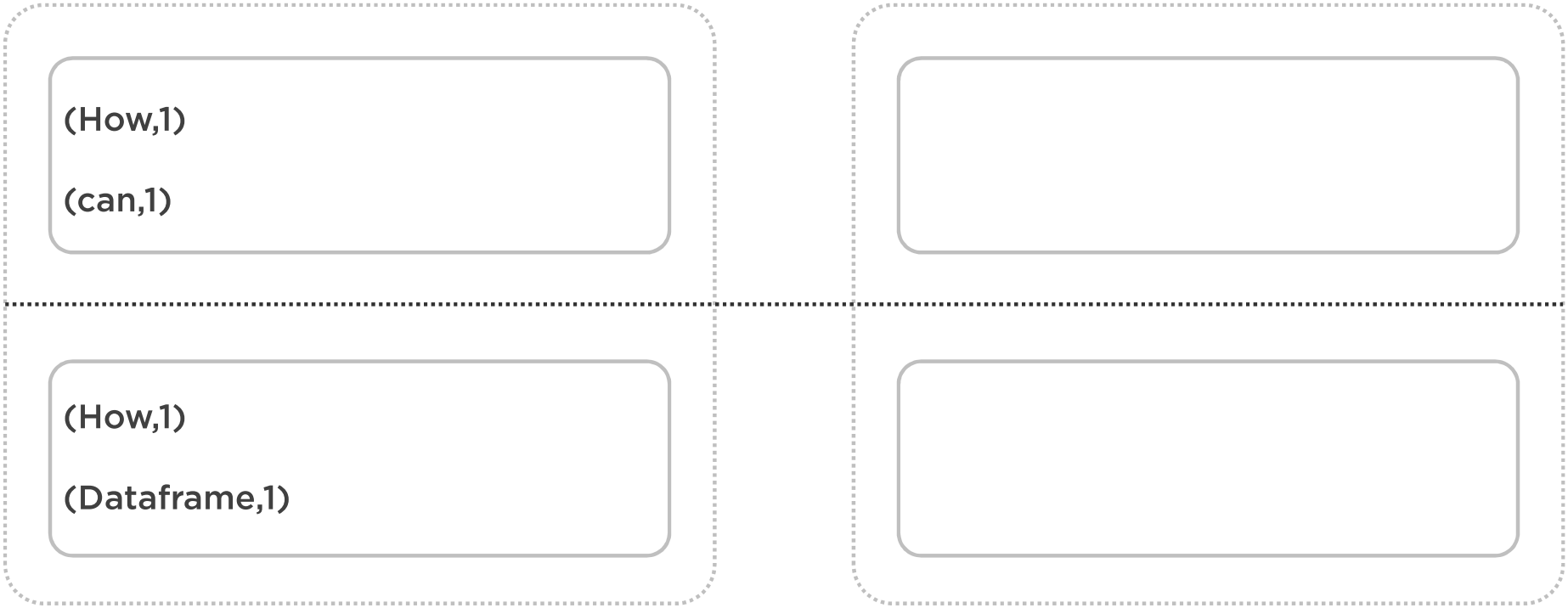
[Can,I,use,Hive,from,HUE]

Narrow Transformation

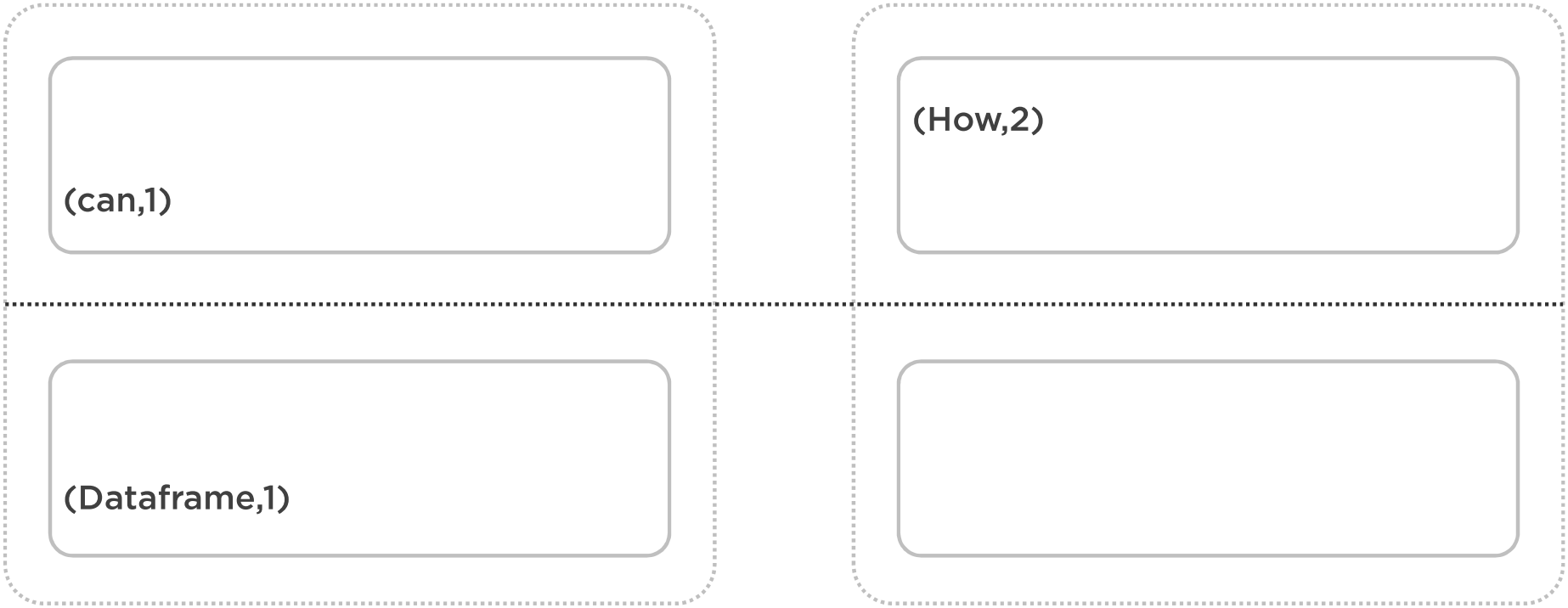
Narrow Transformation



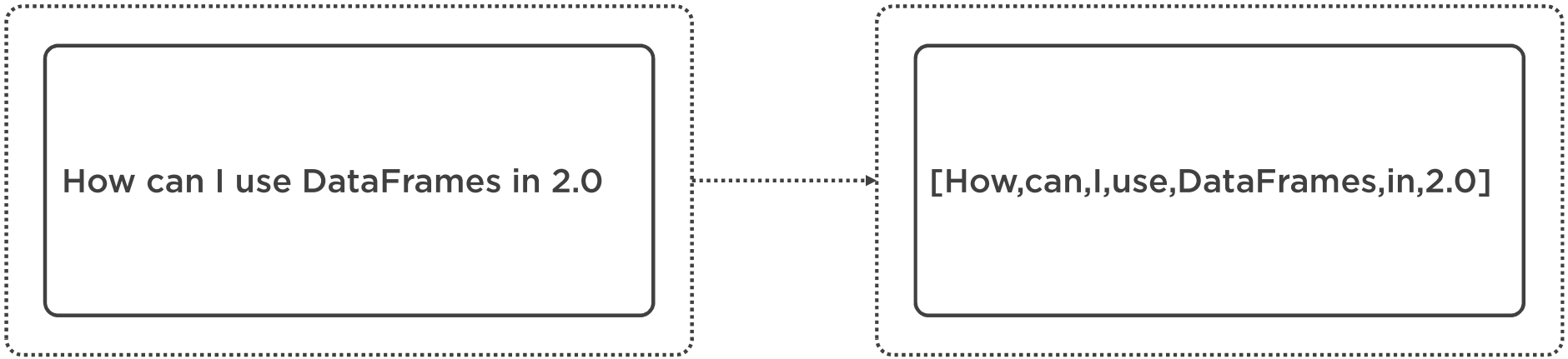
Wide Transformation



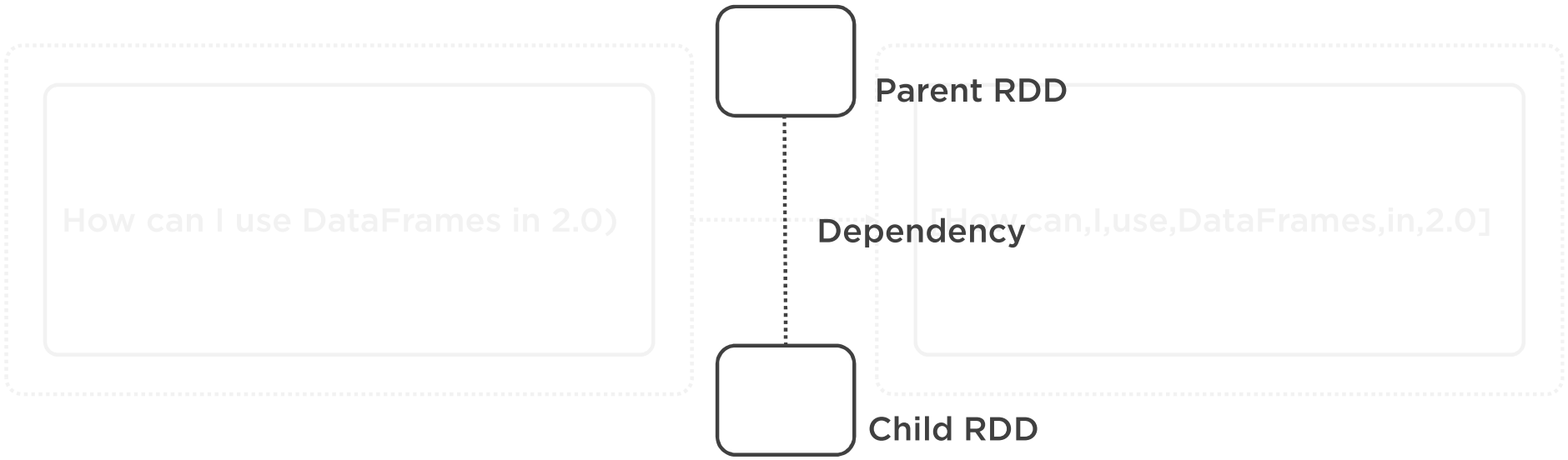
Wide Transformation



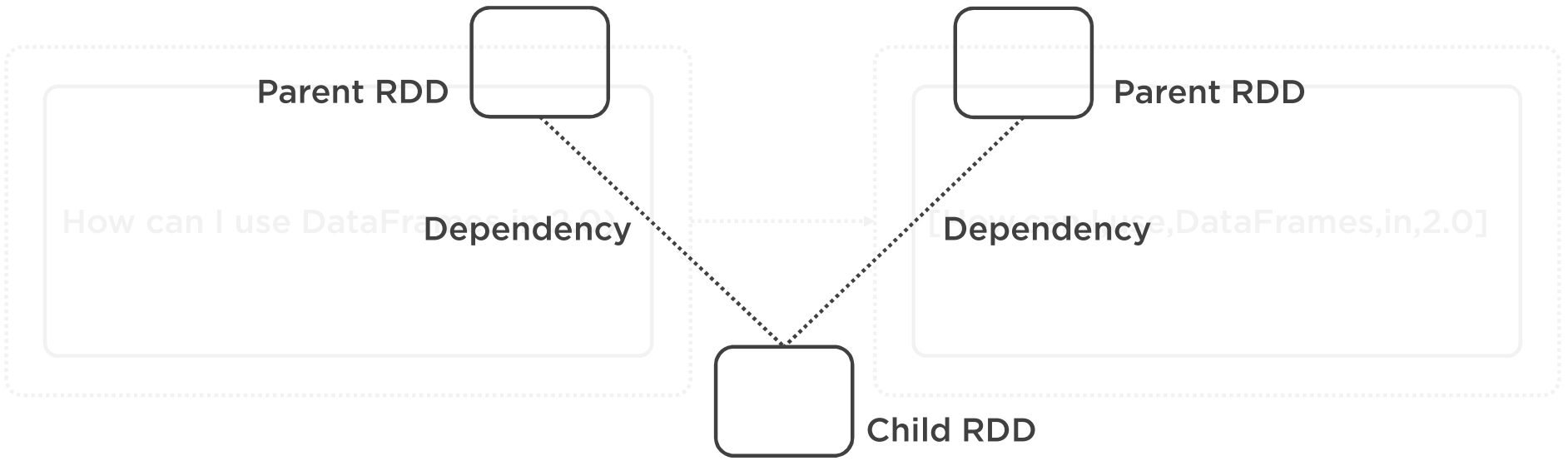
Parent and Child RDD



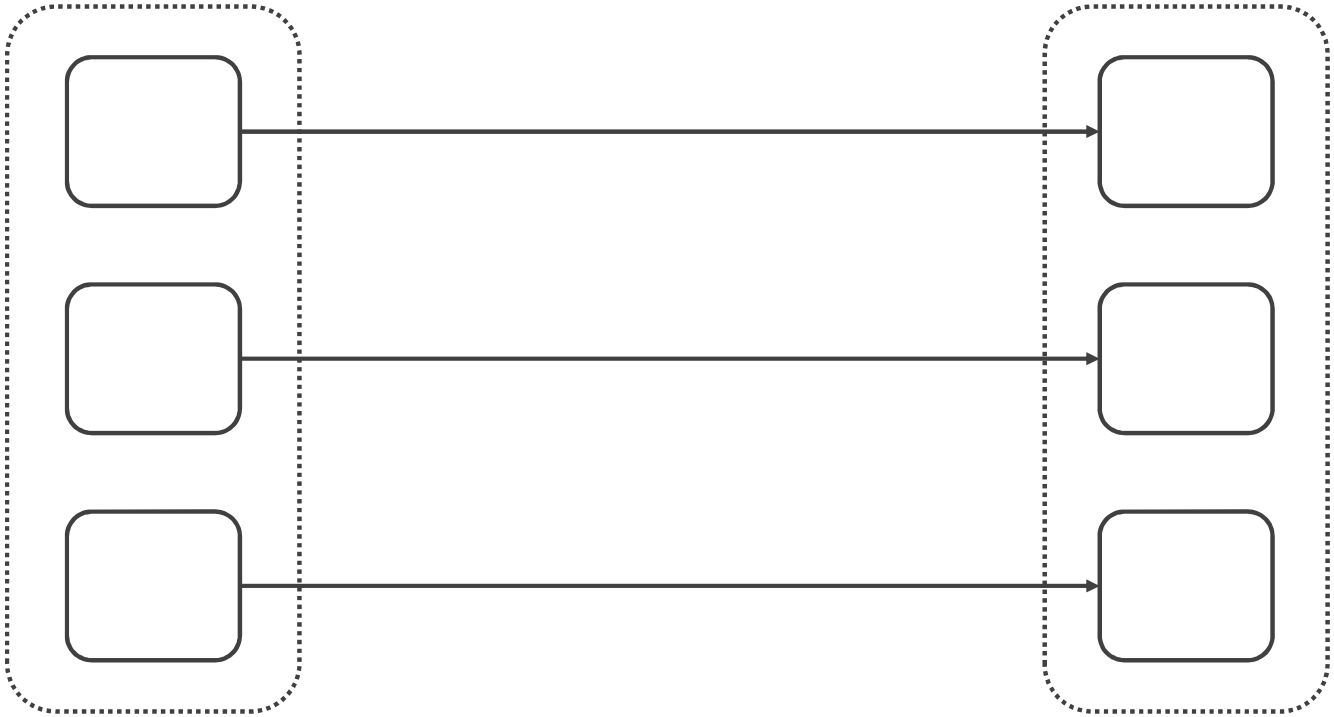
Parent and Child RDD



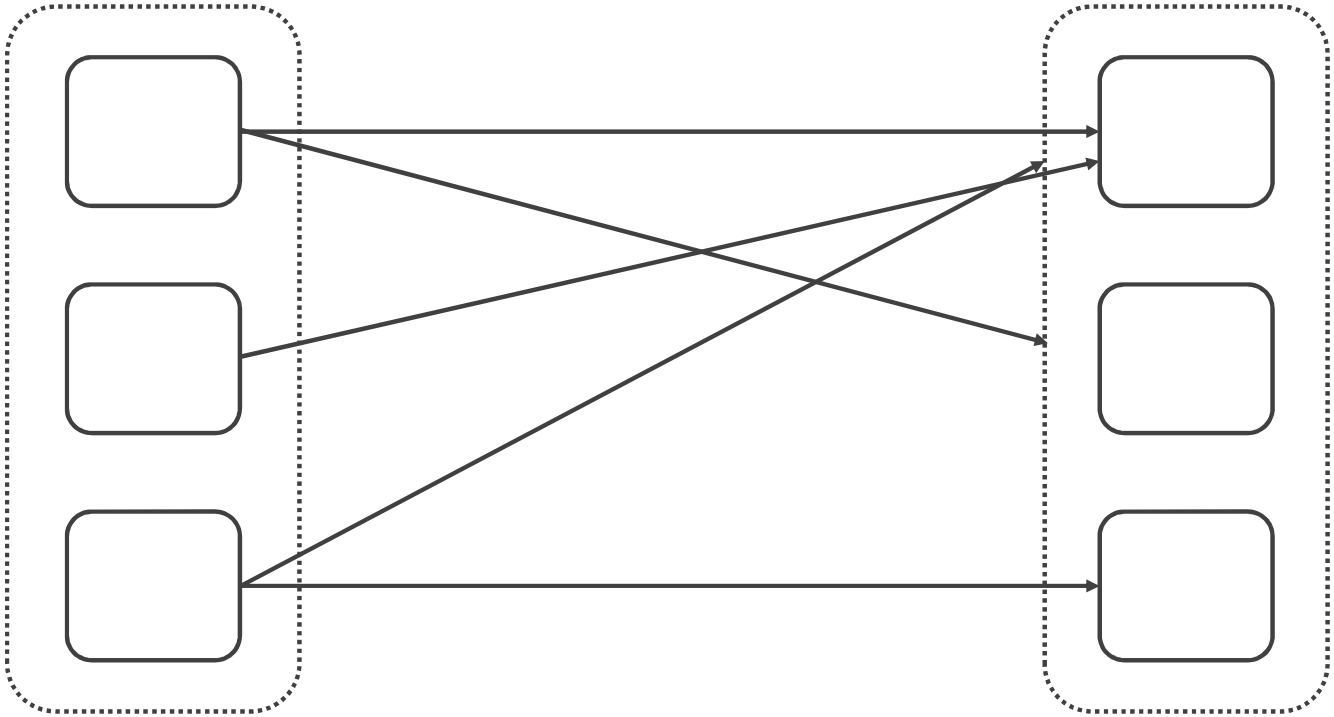
Parent and Child RDD



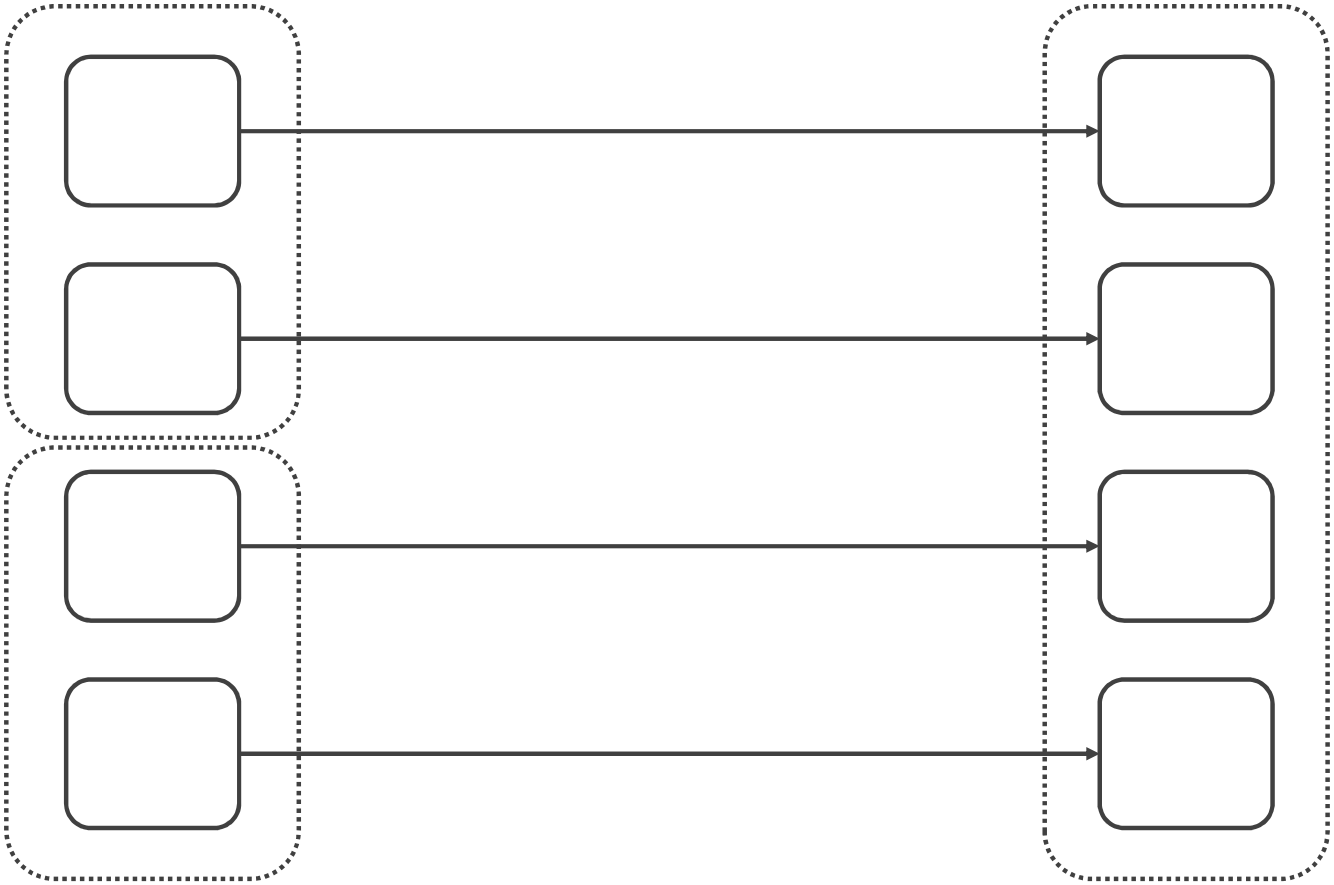
Narrow Transformation



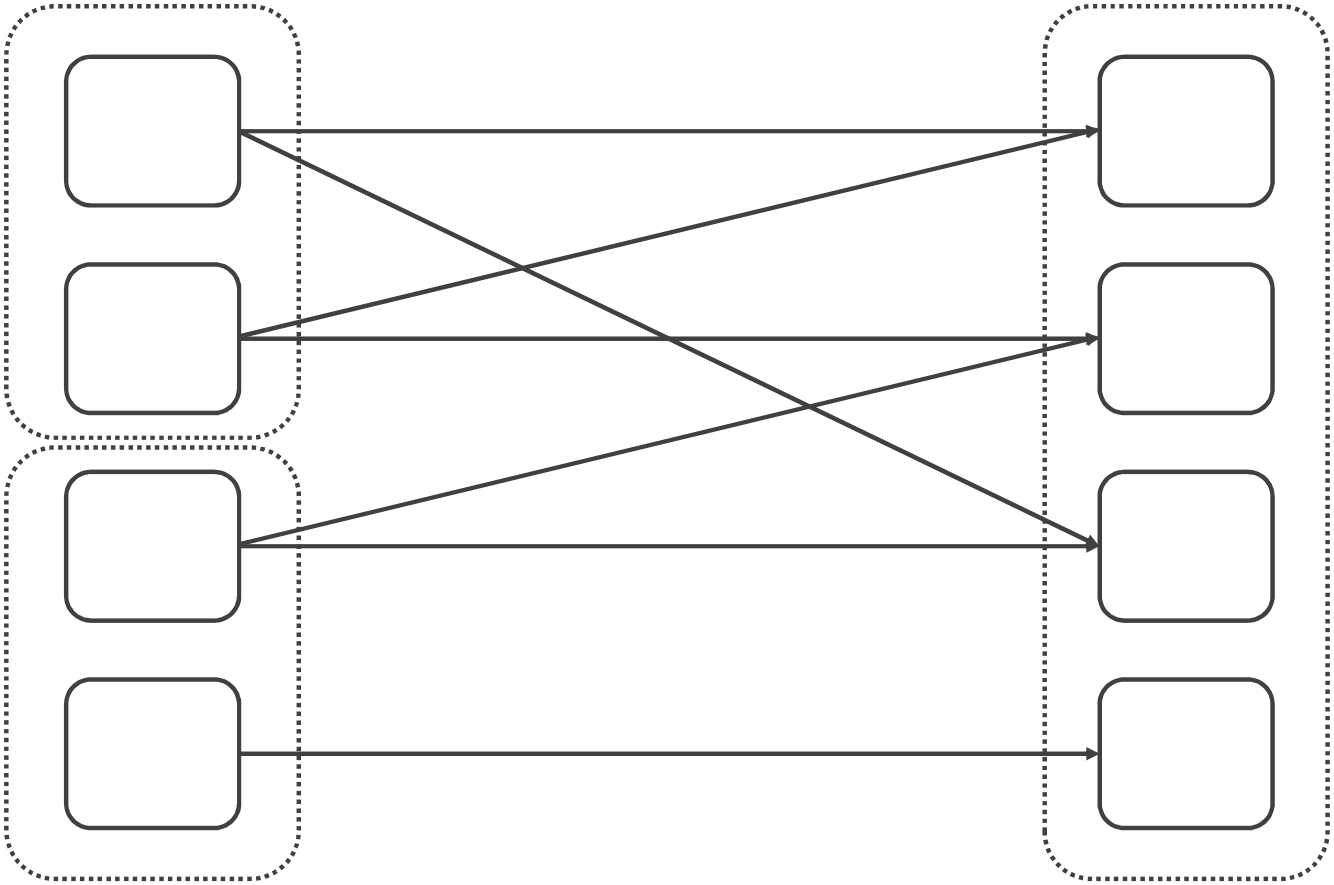
Wide Transformation



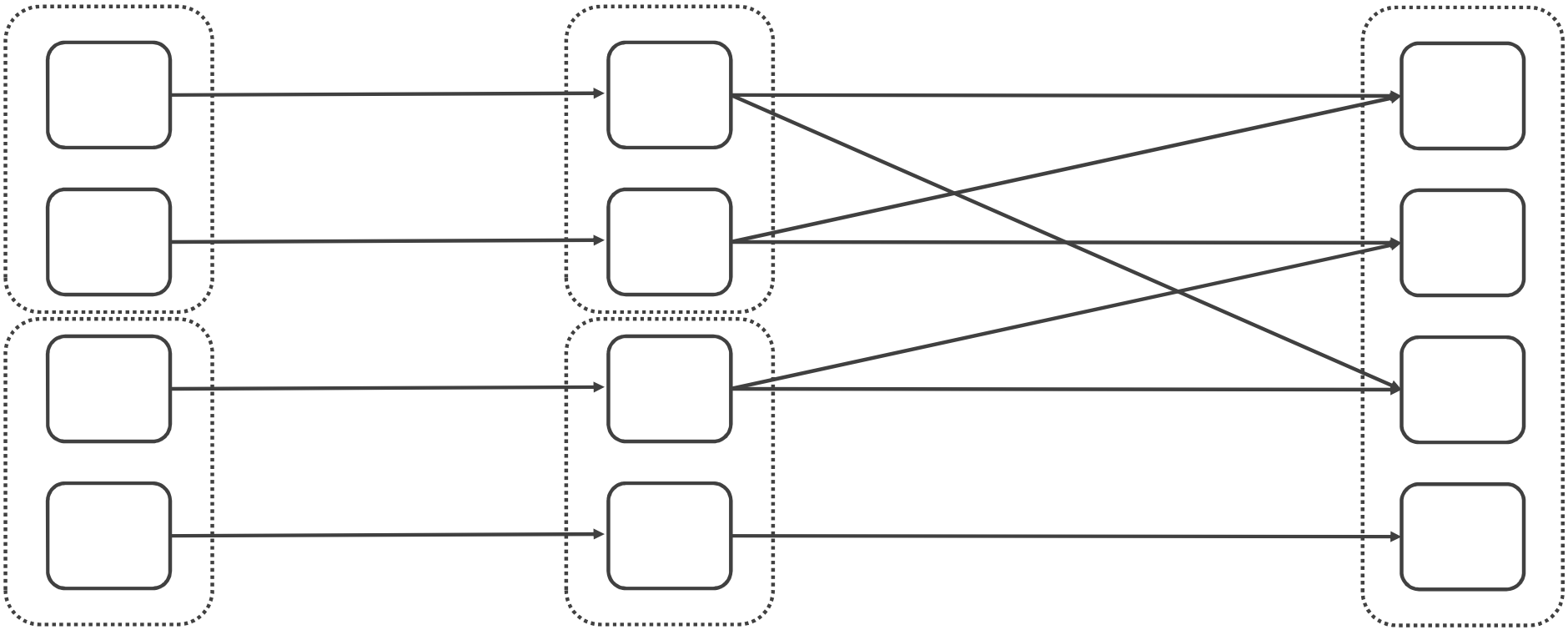
Narrow Transformation



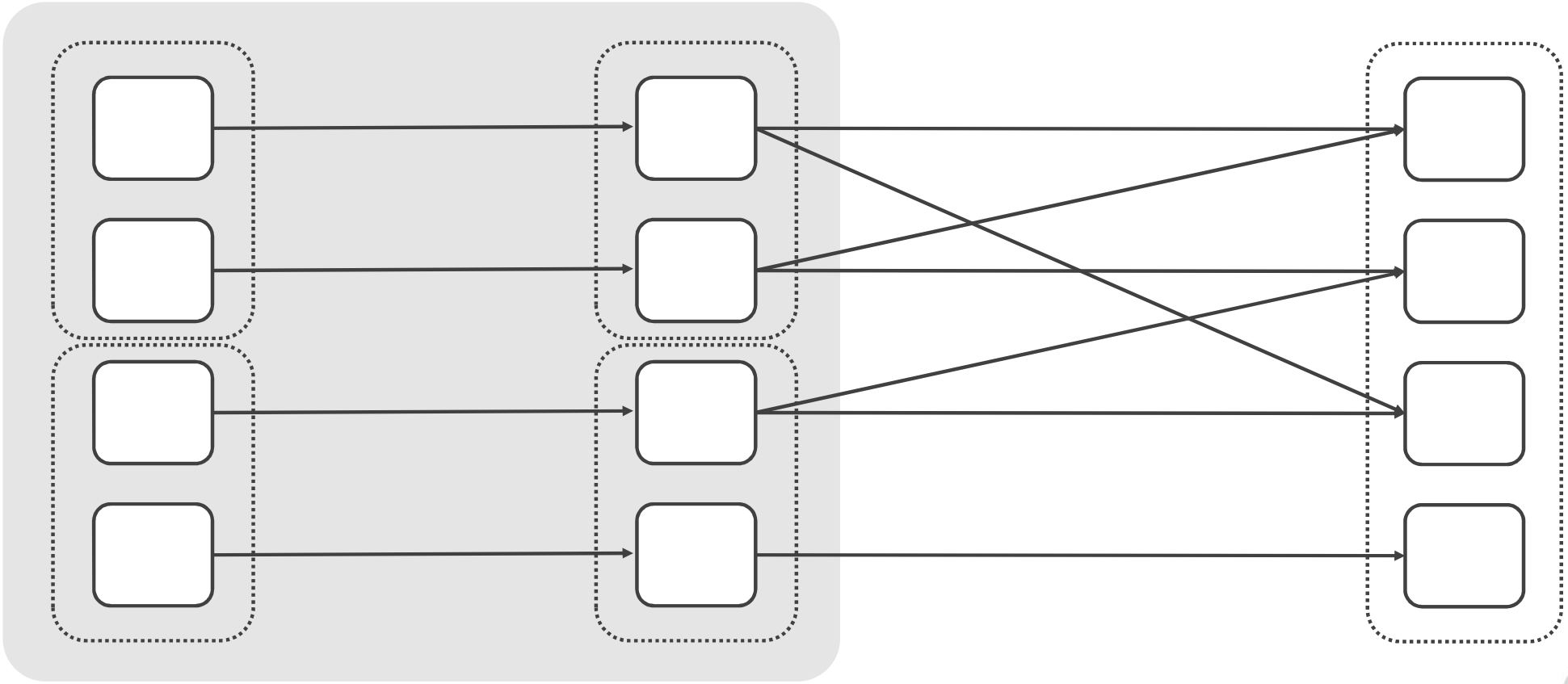
Wide Transformation

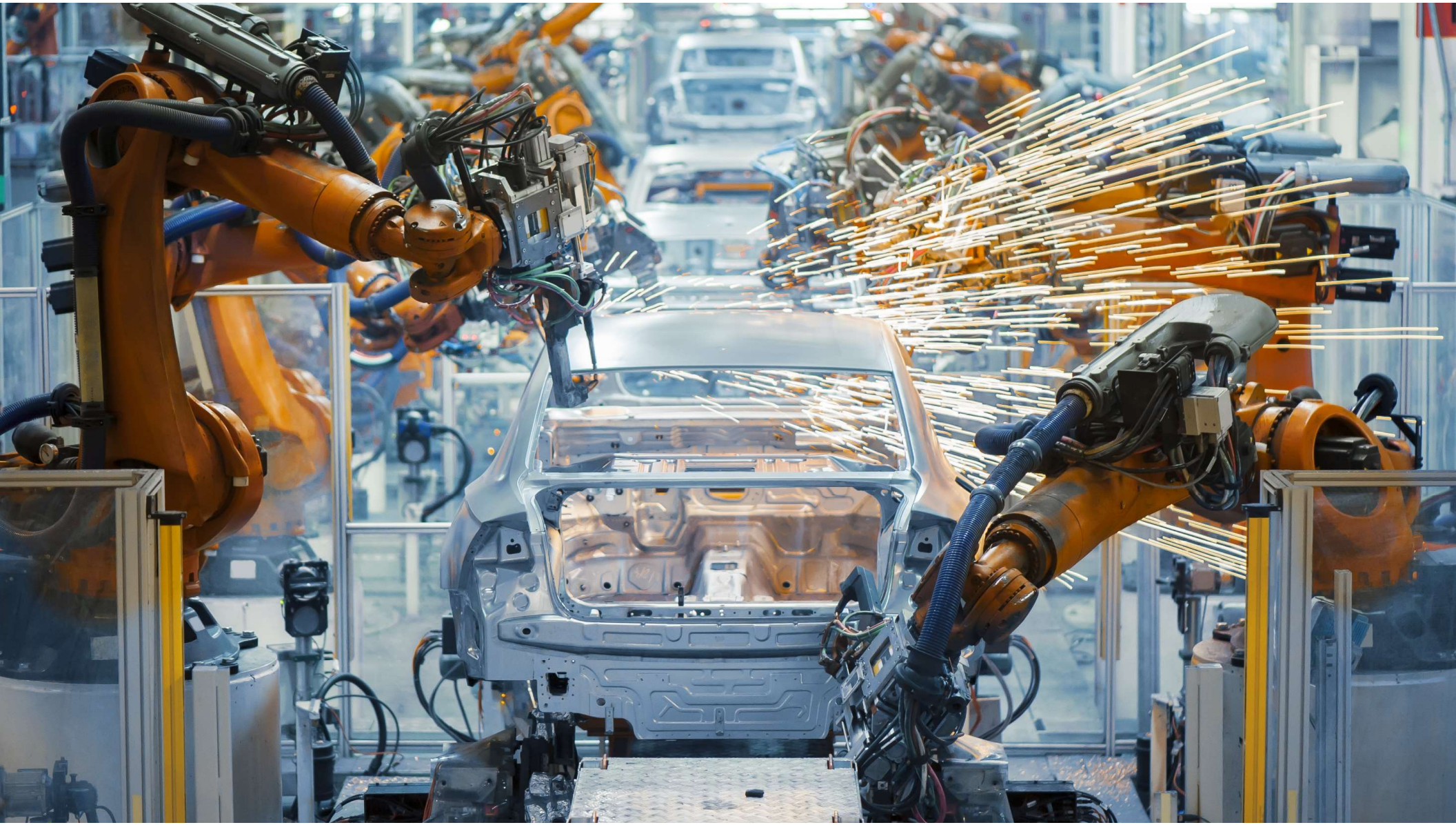


And Why Is This Important?

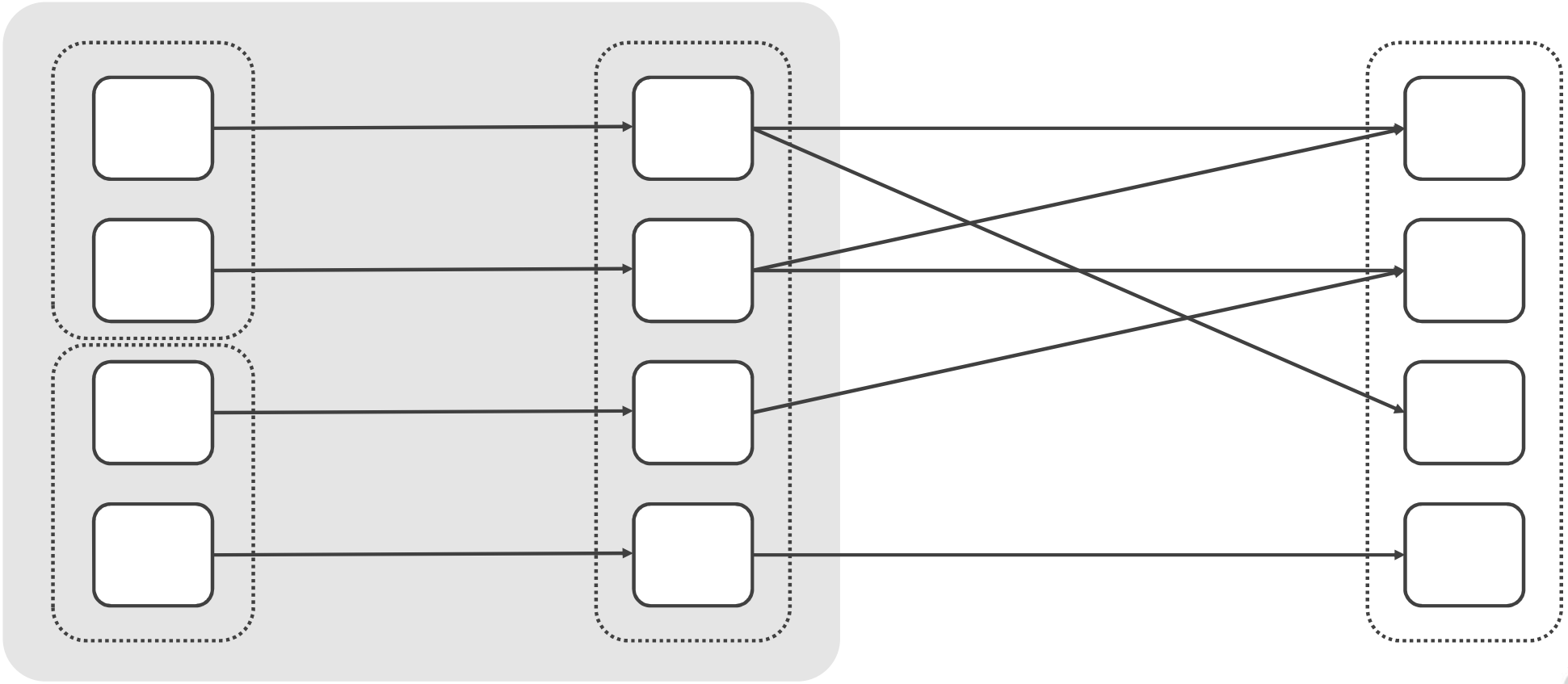


Because Pipelines and Stages





Lineage

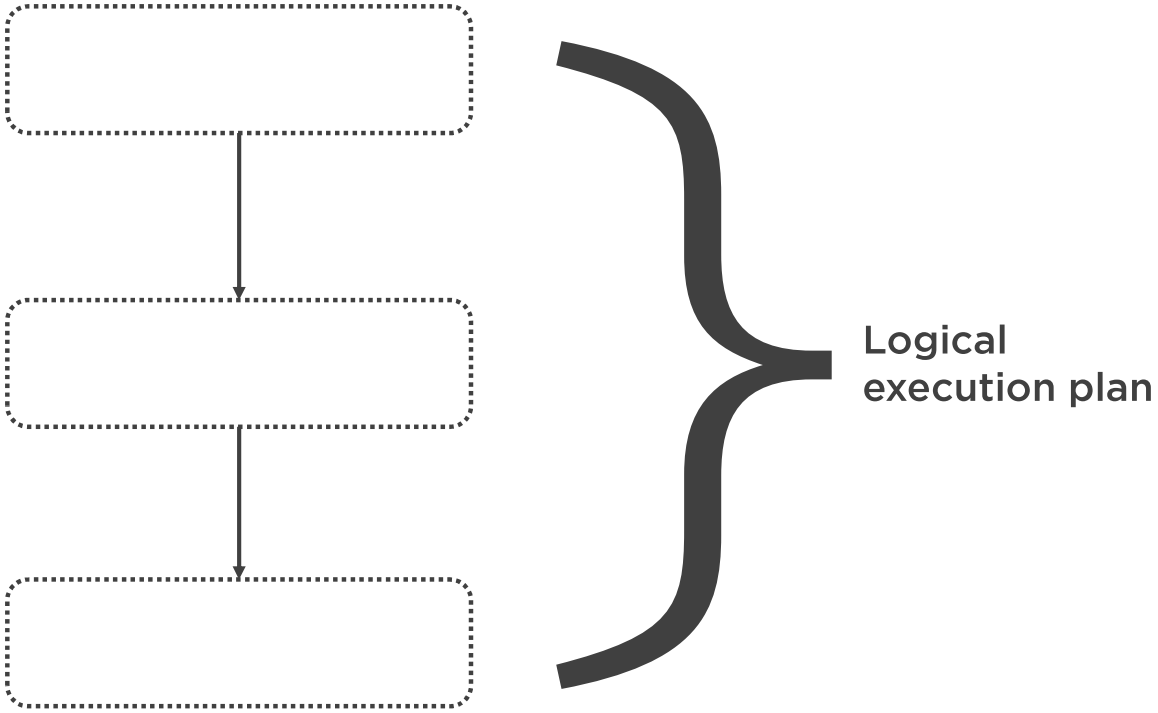


RDD Lineage

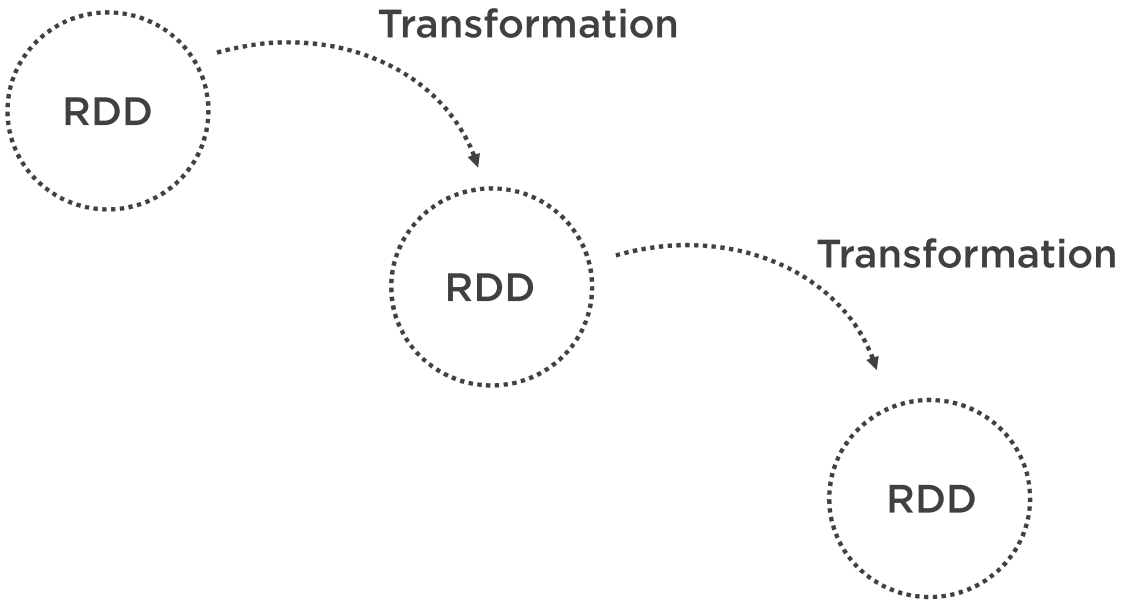
Graph of transformation operations required to execute when an action is called

RDD operator graph or RDD dependency graph

Lineage



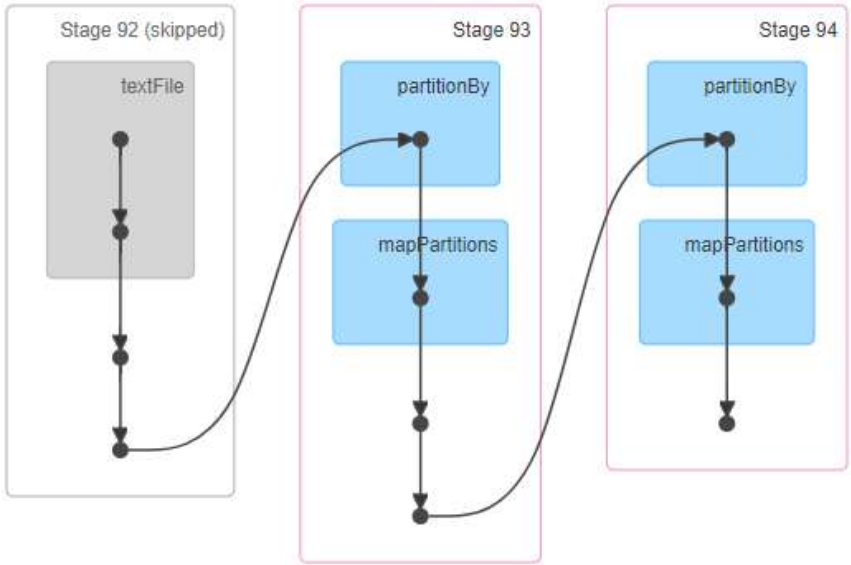
Directed Acyclic Graph or DAG



Details for Job 62

Status: SUCCEEDED
Completed Stages: 2
Skipped Stages: 1

▸ Event Timeline
▾ DAG Visualization



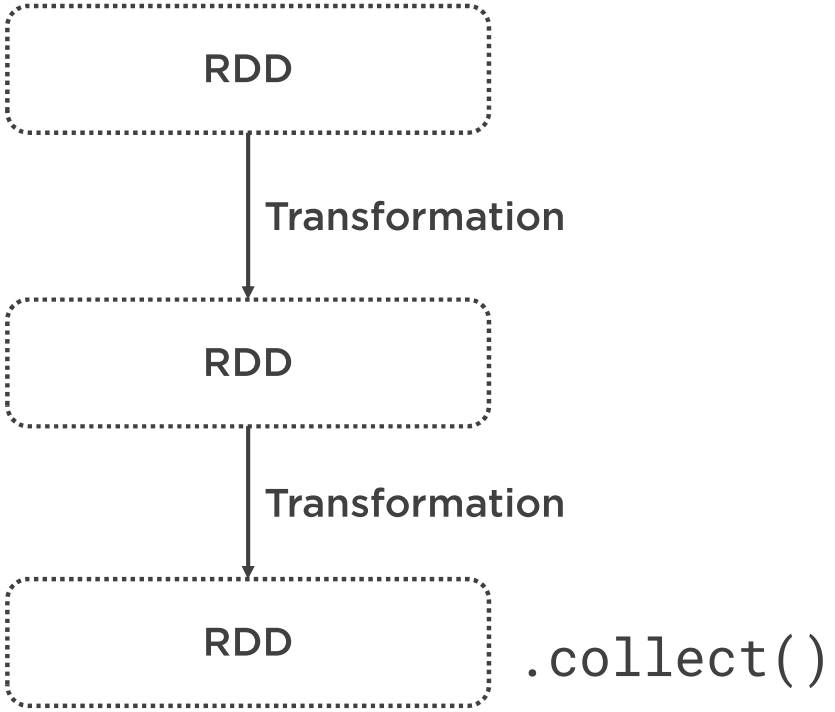
```
val lines = sc.textFile("file:///se/simple_titles.txt")
val words = lines.flatMap(line => line.split(" "))
val word_for_count = words.map(x => (x,1))
word_for_count.reduceByKey(_ + _).collect()
```

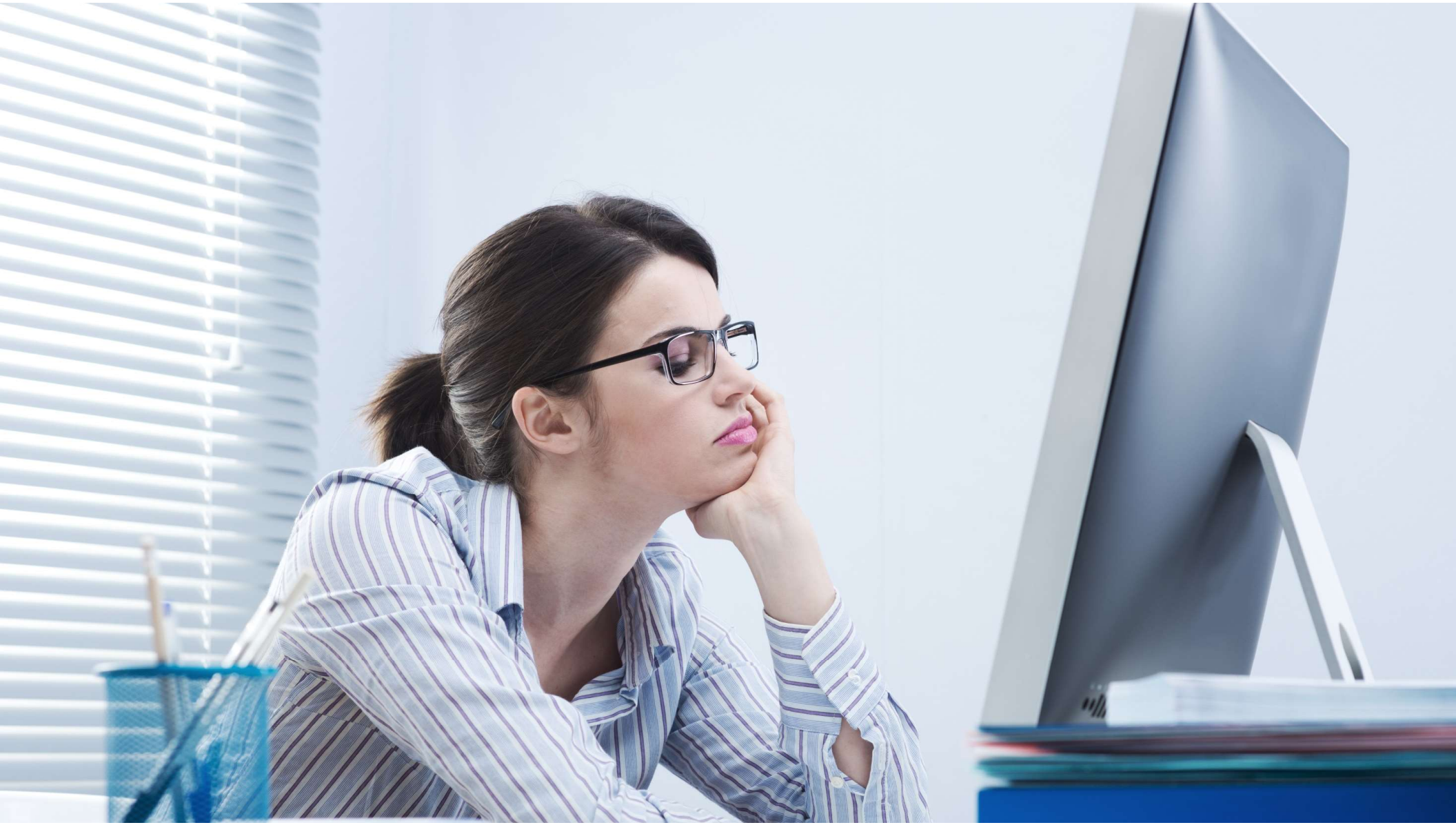
Logical Execution Plan

Starts with the data input

Ends with the Action

Spark is Lazy





RDD

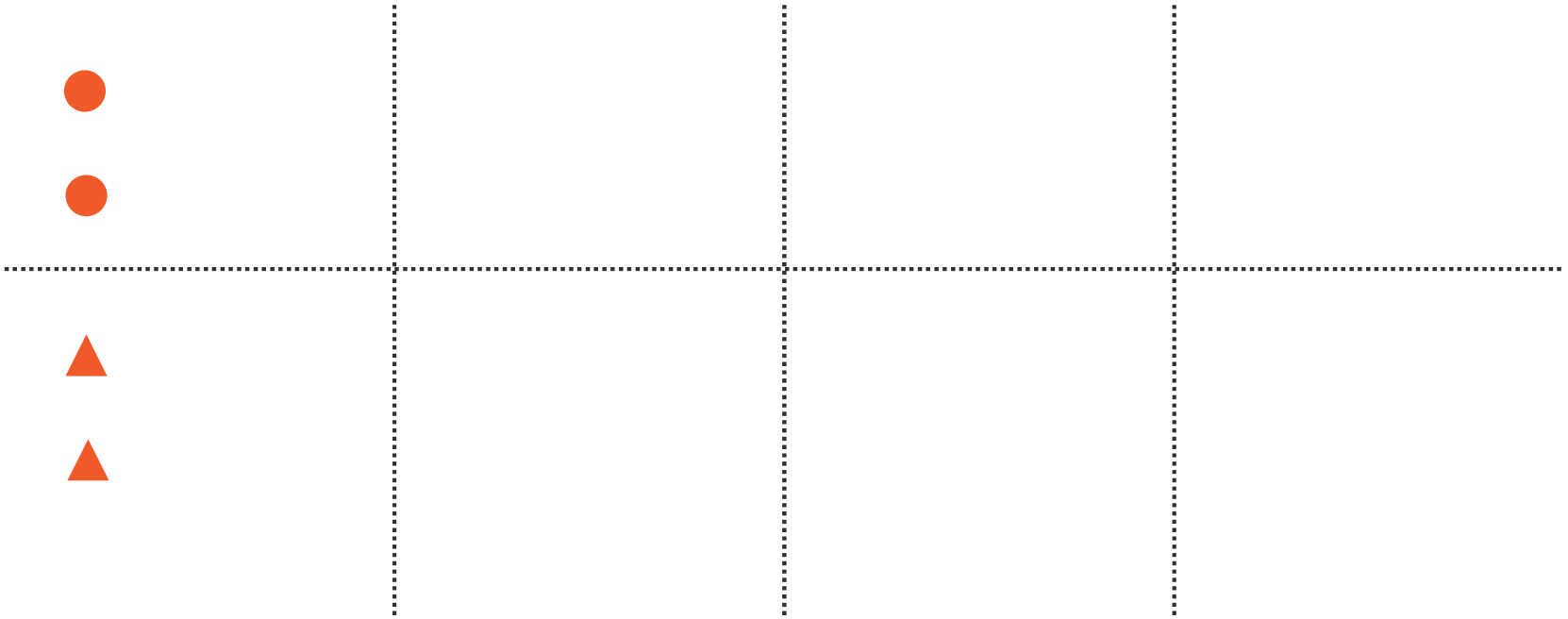
How can I use DataFrames in 2.0

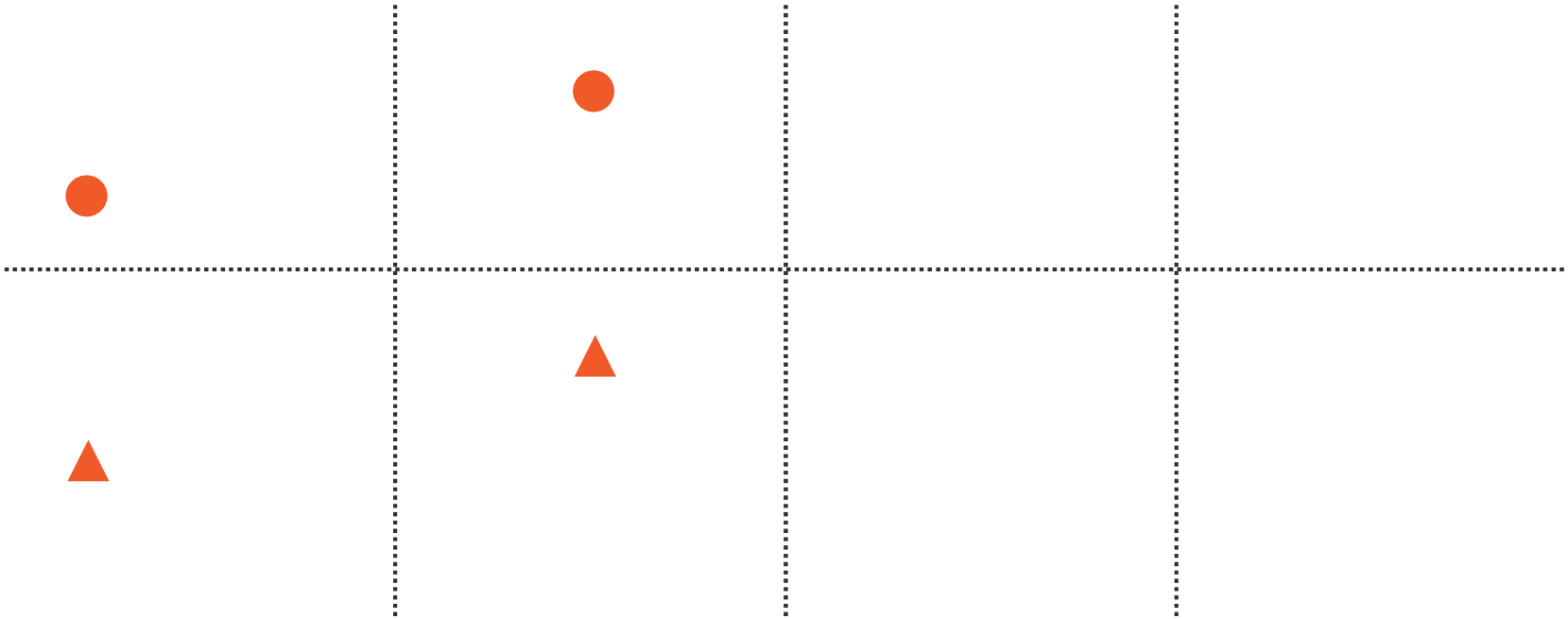
What is an RDD and Schema RDD

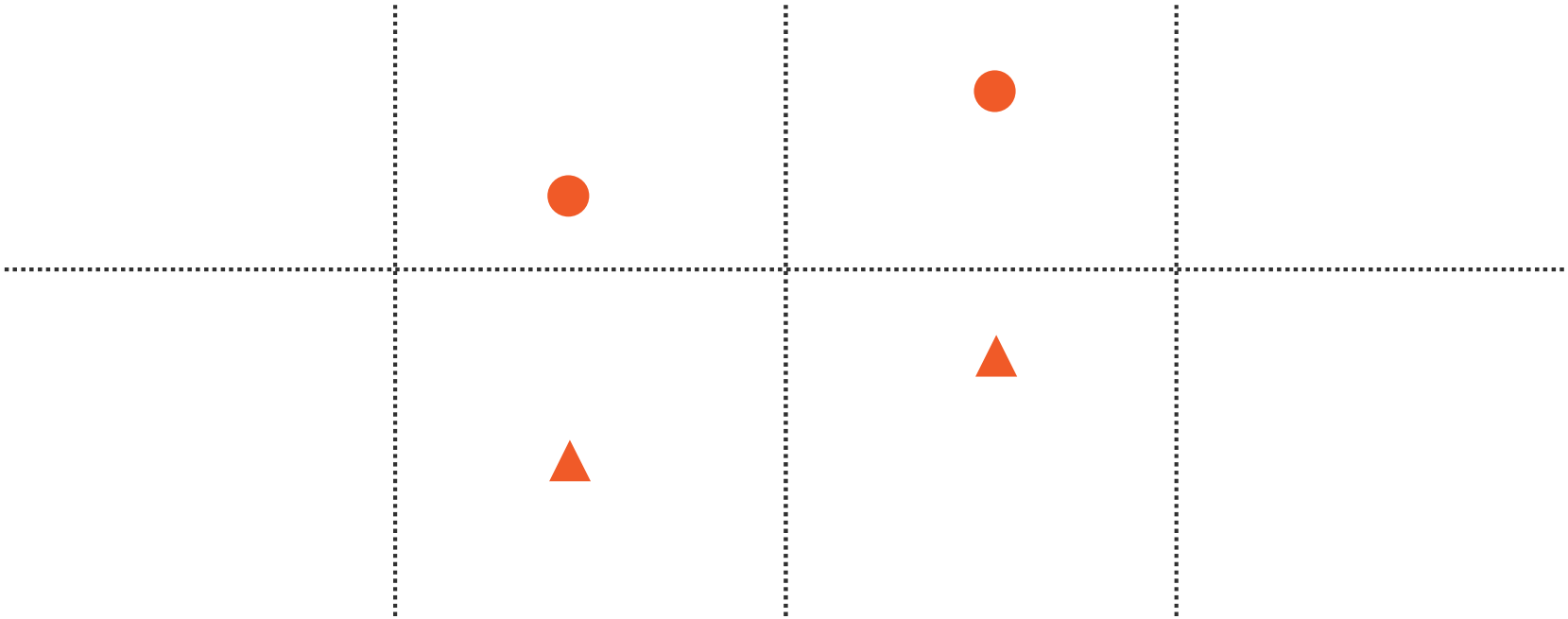
How do I group by a field

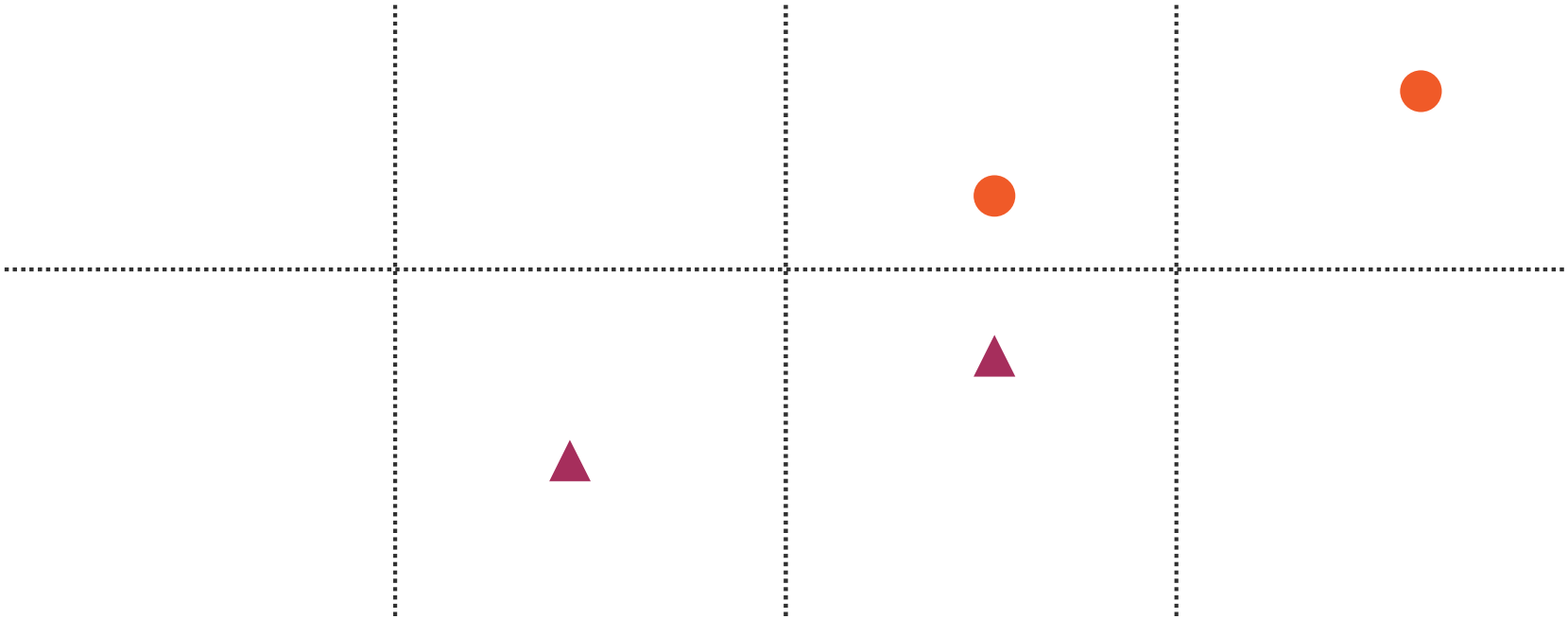
Can I use Hive from HUE

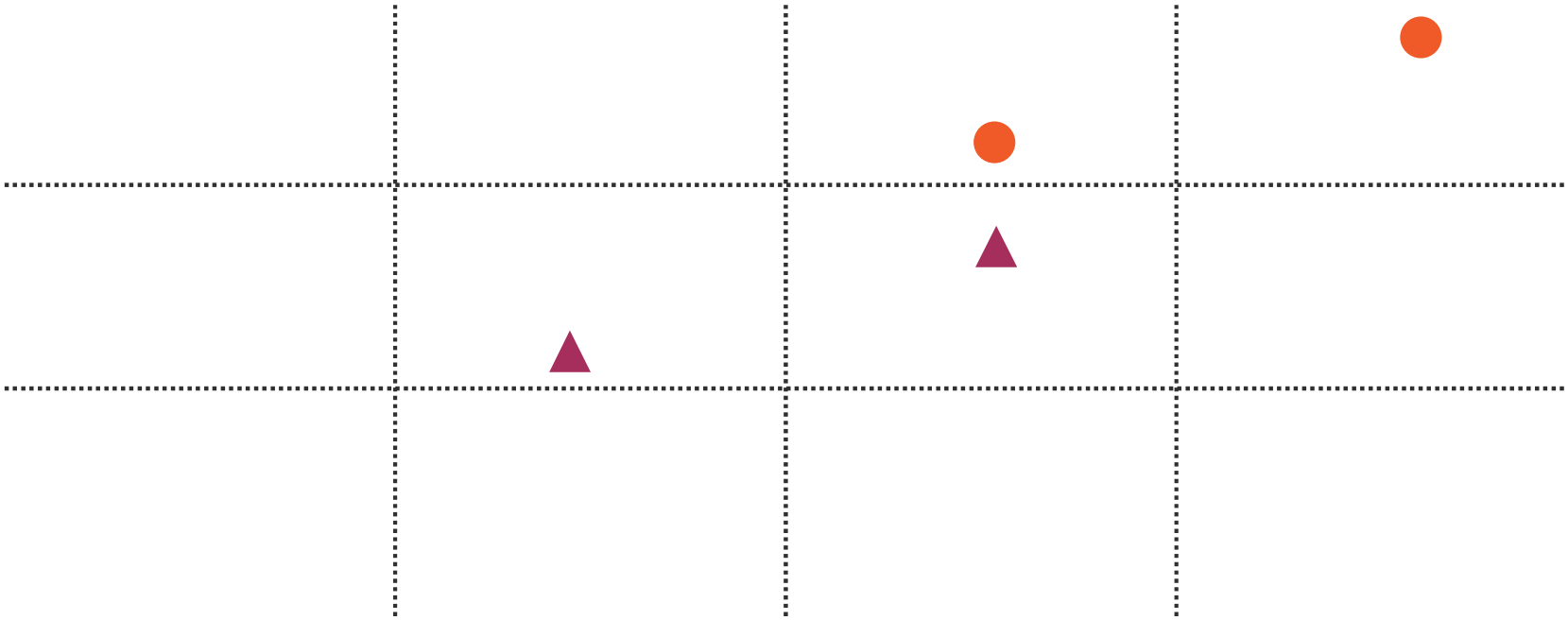
Resilient
Distributed
Dataset

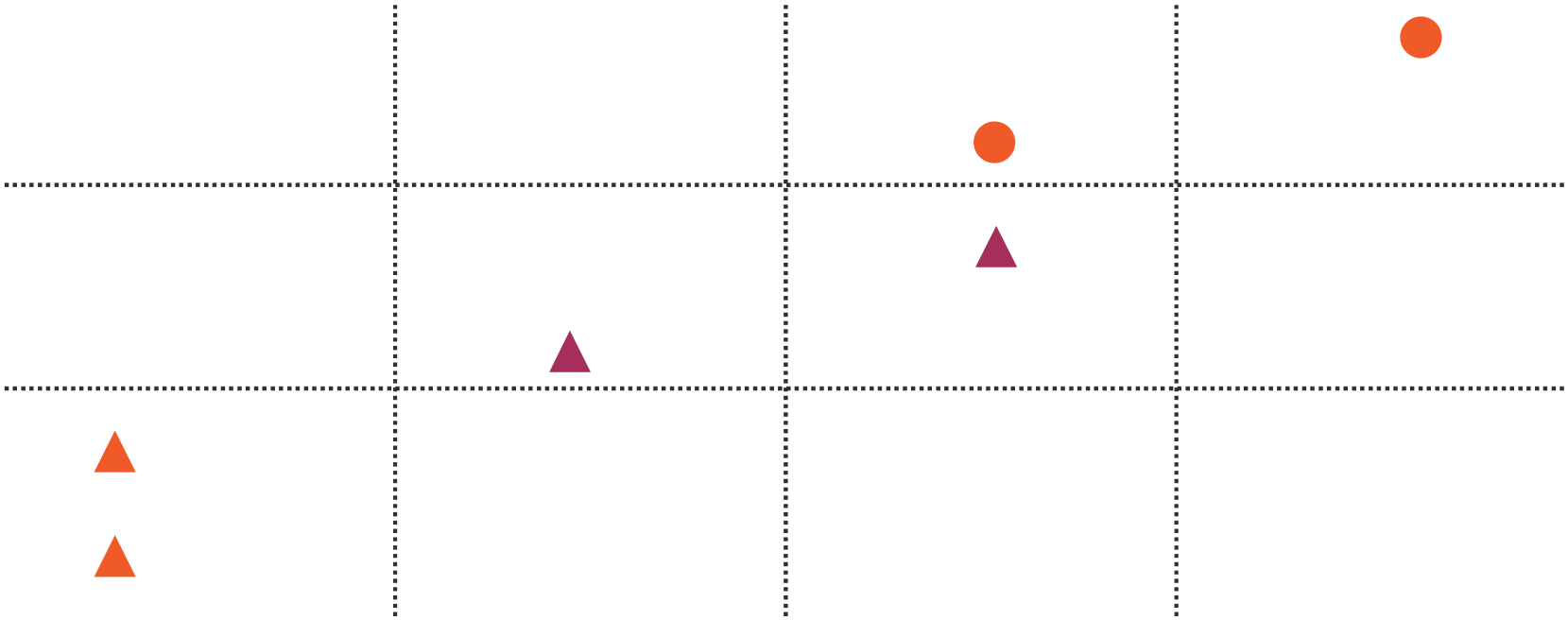


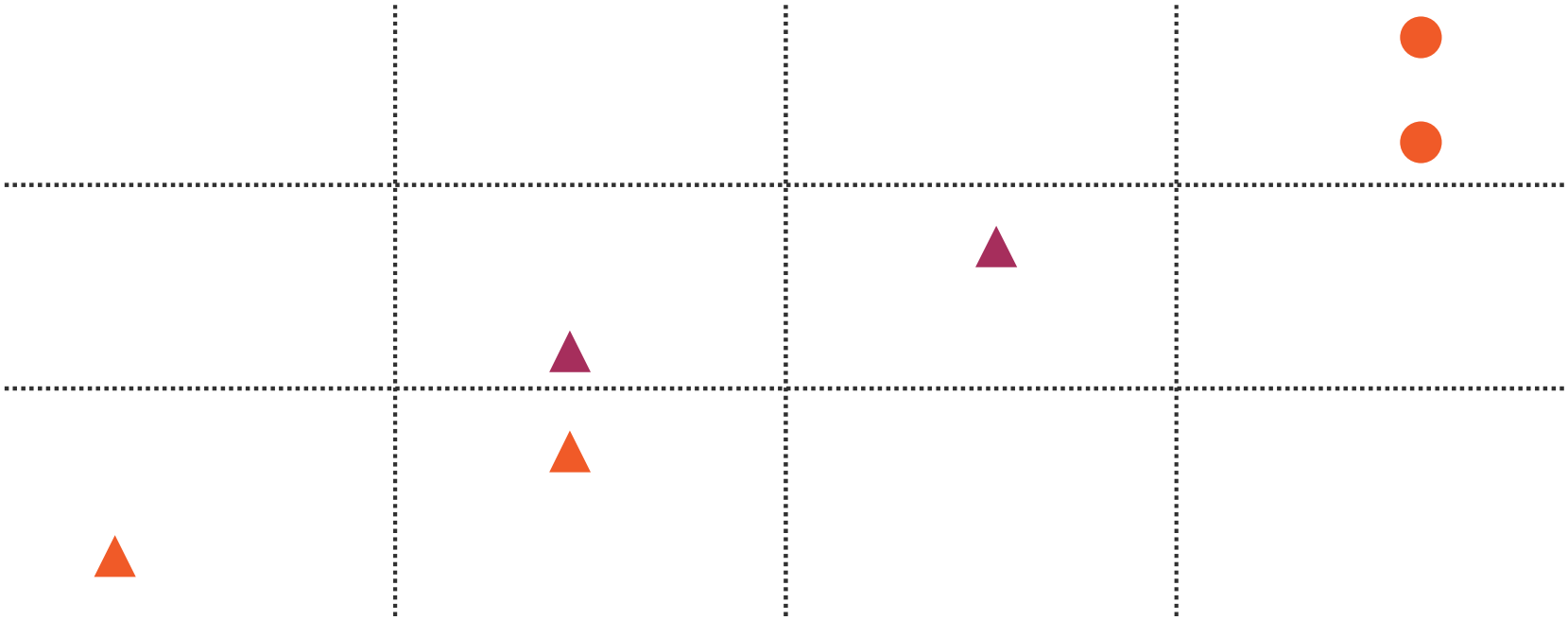


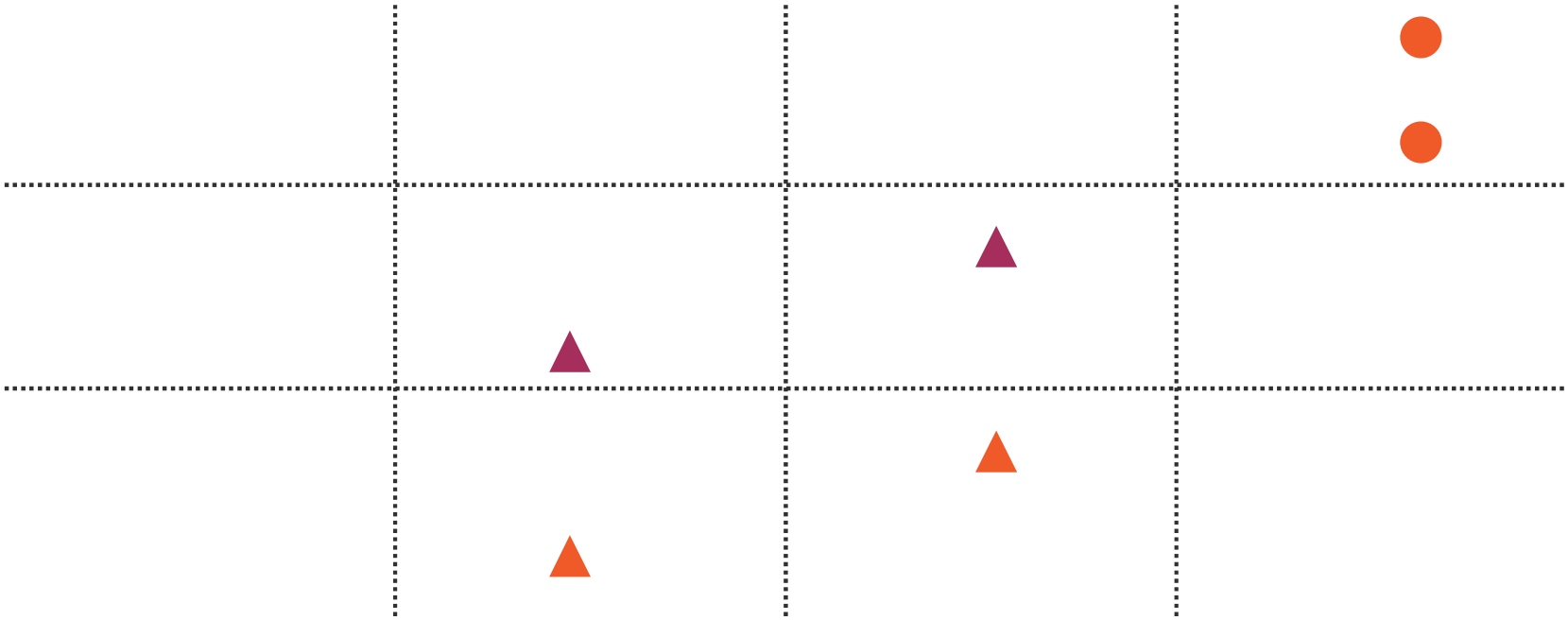


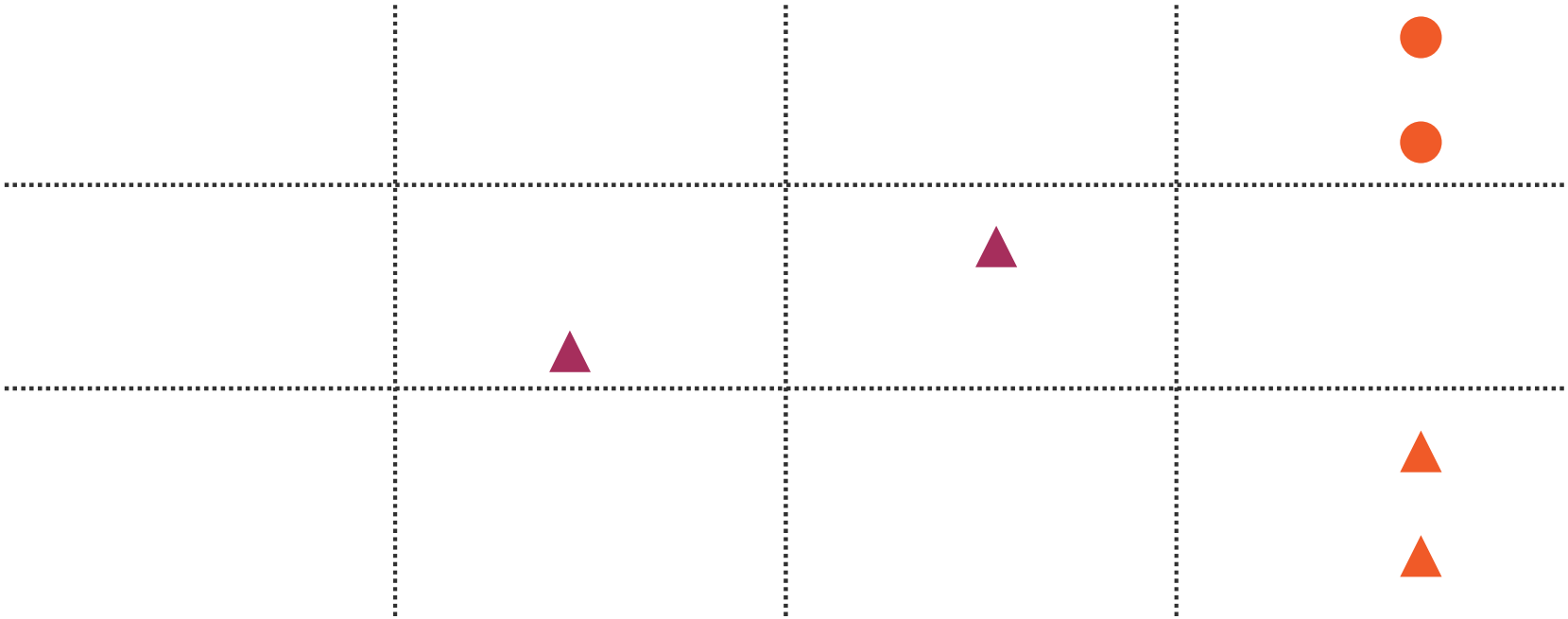


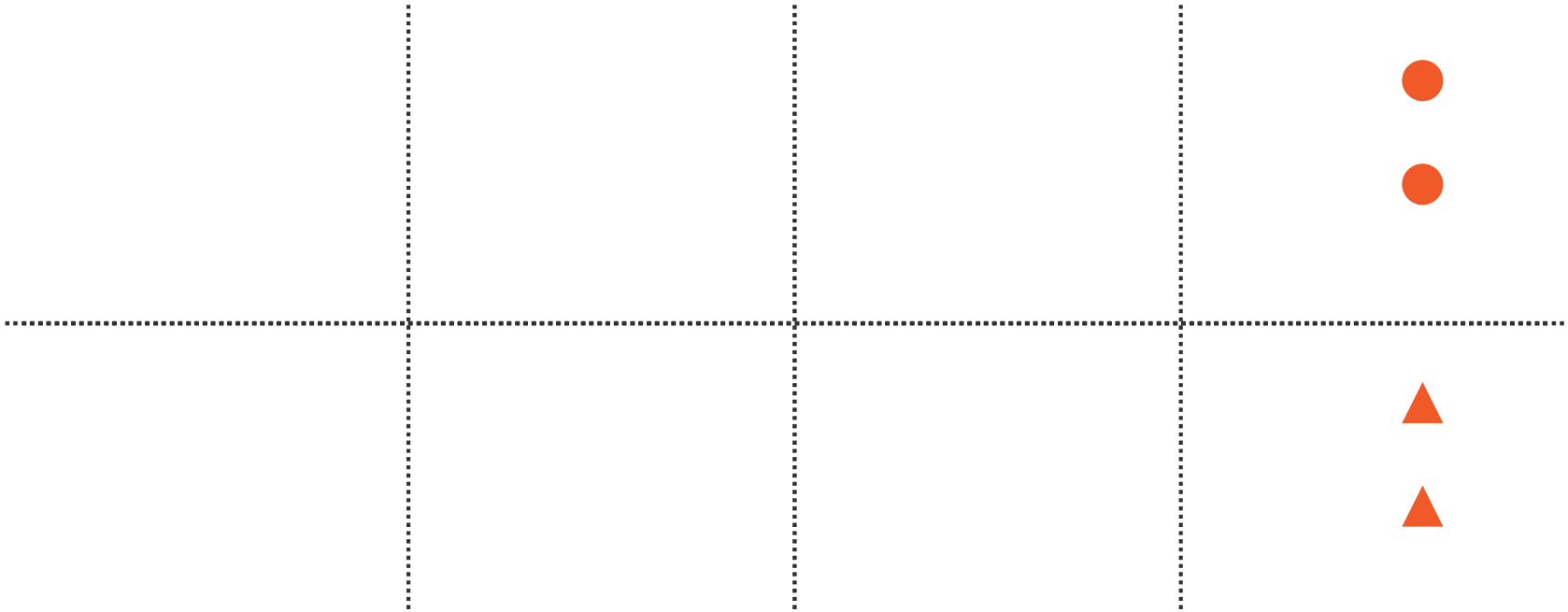


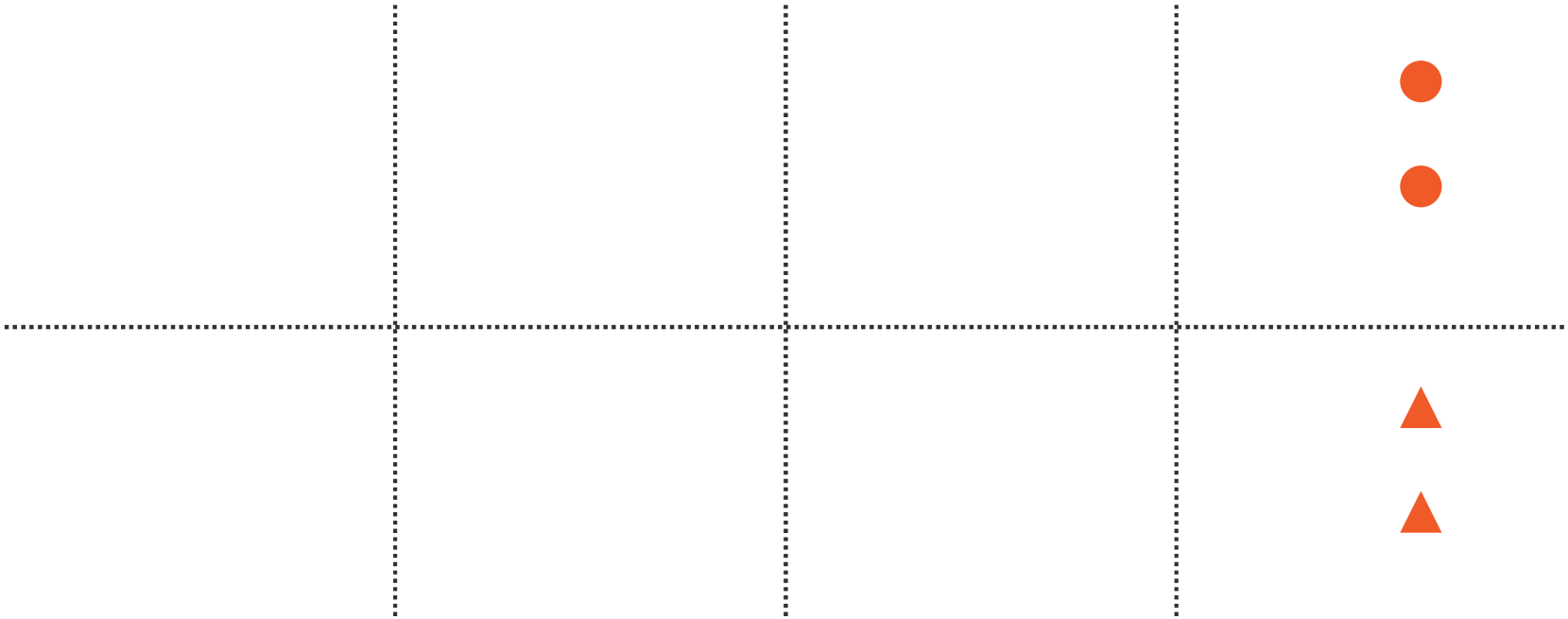




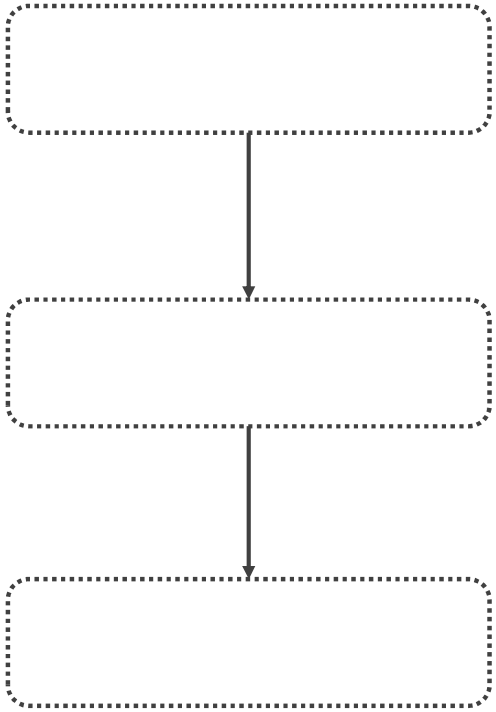




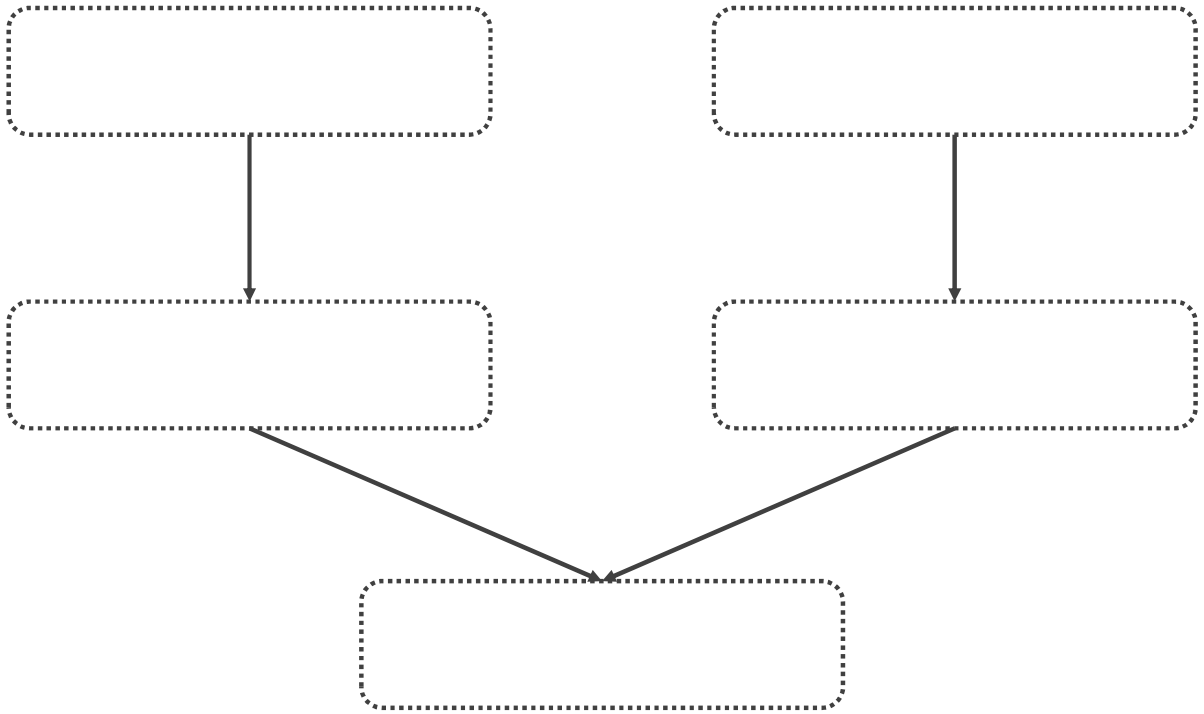




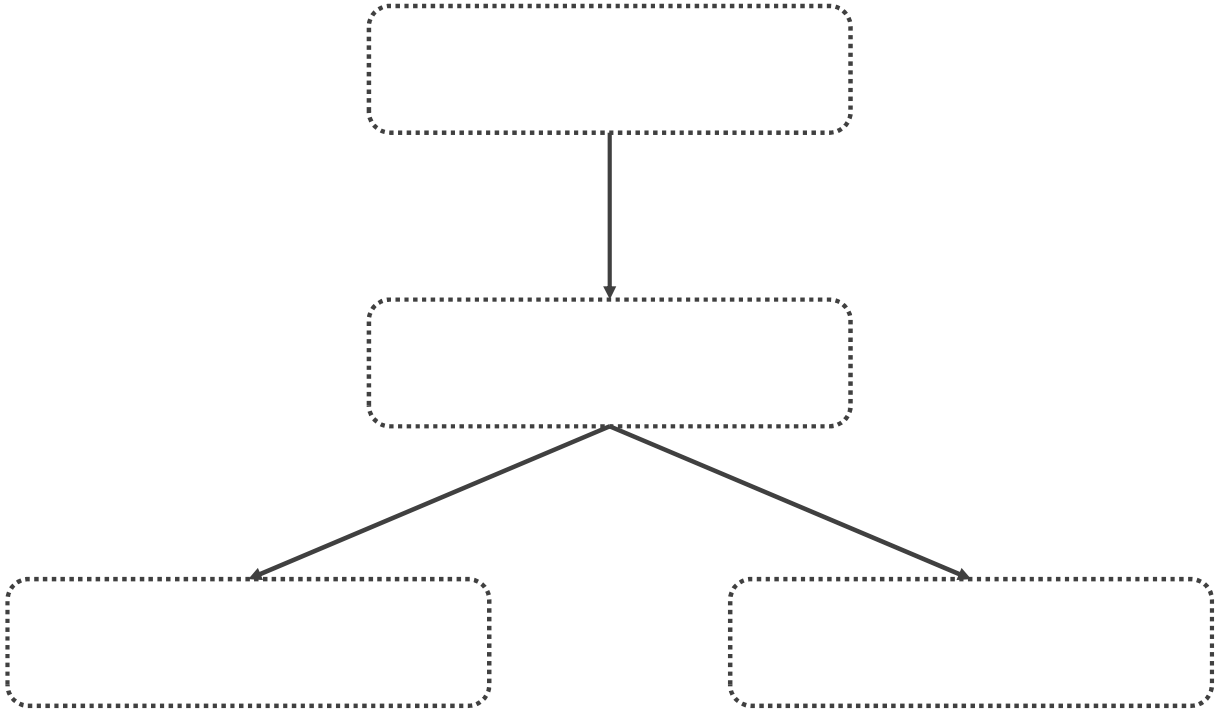
Lineage, Lazy Execution & DAGs

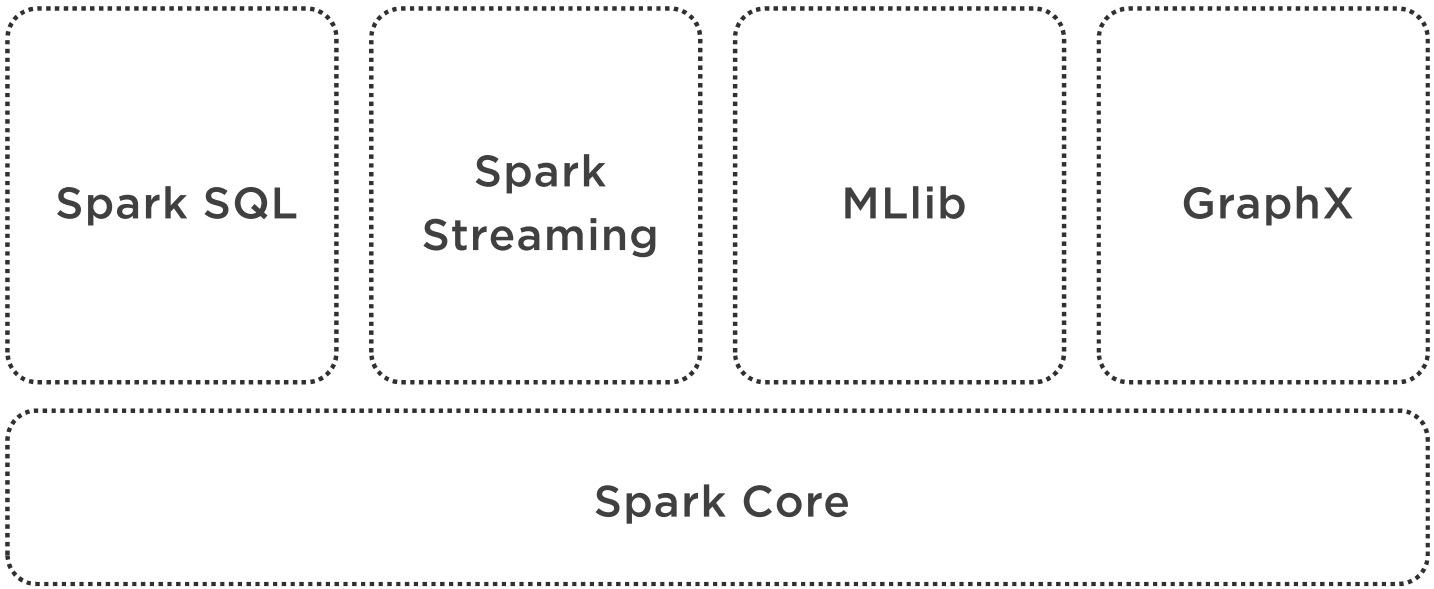


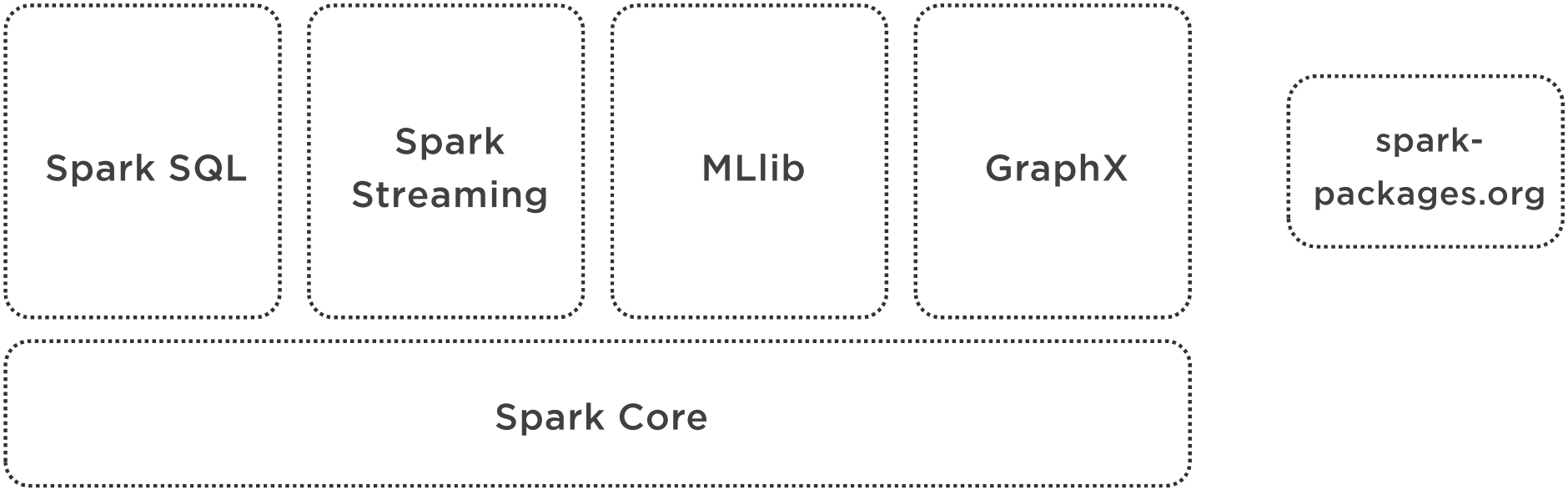
Lineage, Lazy Execution & DAGs

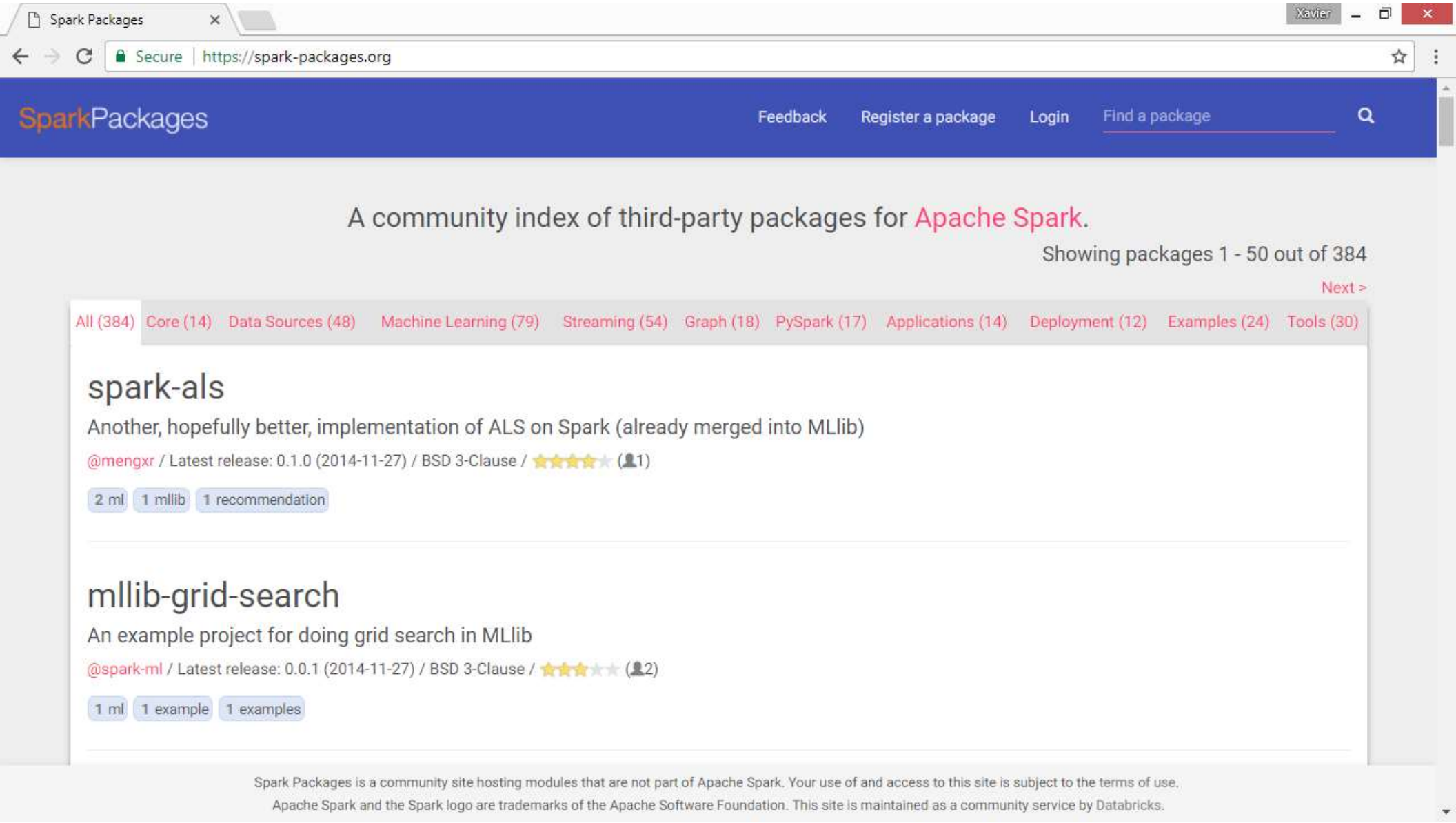


Lineage, Lazy Execution & DAGs




Spark SQL
Spark Streaming
MLlib
GraphX
Spark Core


Spark SQL
Spark Streaming
MLlib
GraphX
spark-packages.org
Spark Core


Spark Packages
Xavier
Secure | https://spark-packages.org
SparkPackages
Feedback
Register a package
Login
Find a package
A community index of third-party packages for Apache Spark.
Showing packages 1 - 50 out of 384
Next >
All (384) Core (14) Data Sources (48) Machine Learning (79) Streaming (54) Graph (18) PySpark (17) Applications (14) Deployment (12) Examples (24) Tools (30)
spark-als
Another, hopefully better, implementation of ALS on Spark (already merged into MLlib)
@mengxr / Latest release: 0.1.0 (2014-11-27) / BSD 3-Clause / (1)
2 ml 1 mllib 1 recommendation
mllib-grid-search
An example project for doing grid search in MLlib
@spark-ml / Latest release: 0.0.1 (2014-11-27) / BSD 3-Clause / (2)
1 ml 1 example 1 examples
Spark Packages is a community site hosting modules that are not part of Apache Spark. Your use of and access to this site is subject to the terms of use.
Apache Spark and the Spark logo are trademarks of the Apache Software Foundation. This site is maintained as a community service by Databricks.

Takeaway

Brief history of what led to Spark

This data looks like a SQL table

- Fine grained operations
- Easy with SQL

And then the Word Count

- Hello World of Big Data
- Useful for a typeahead

Takeaway

Word Count not so simple

- In SQL
- Other higher level languages

A few lines in Spark

- Scales really well

Takeaway

How Spark works

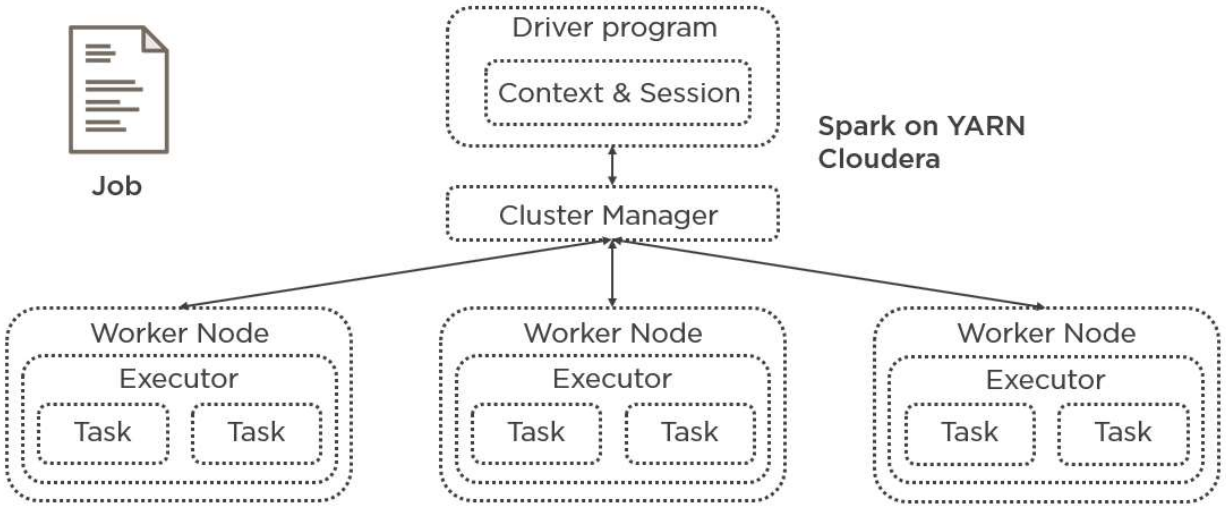
- Parallelism, partitioning, pipelining, DAG, lazy evaluation, fault tolerance, ...
- RDD

Spark's libraries

Time to get technical with Spark

- Spark's Architecture

Takeaway


Job
Driver program
Context & Session
Cluster Manager
Spark on YARN
Cloudera
Worker Node
Executor
Task
Task
Worker Node
Executor
Task
Task
Worker Node
Executor
Task
Task