From 5 PB of logs to thousands of models - Story of a ML pipeline

Fabian Höring

ML related stuff
represents only
10% of your code


200 TB of data every day

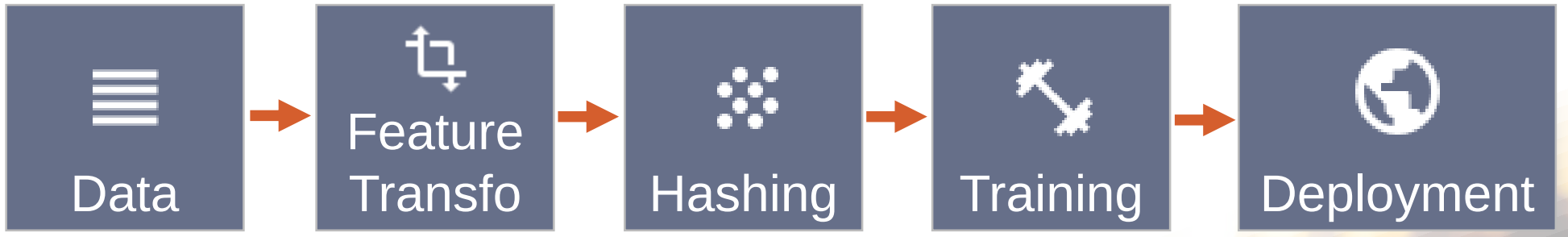
Europe's largest Hadoop Cluster

Given a description of a user and an item history build a model to predict how likely a user will be to click on the ad


YARN
hadoop
HDFS
APACHE
Spark
MLlib

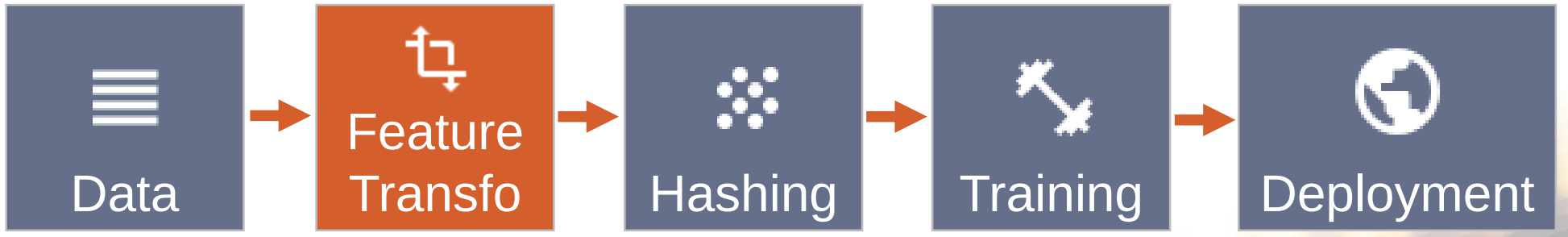

Offline
Online

Offline Learning Workflow for 1 model




Preprocessing
Learning
Data
Feature Transfo
Hashing
Training
Deployment

Feature transformations



“The codepaths that
actually generate input features may differ for training and inference time .. This is sometimes called “training/serving skew” and requires careful monitoring to detect and avoid”

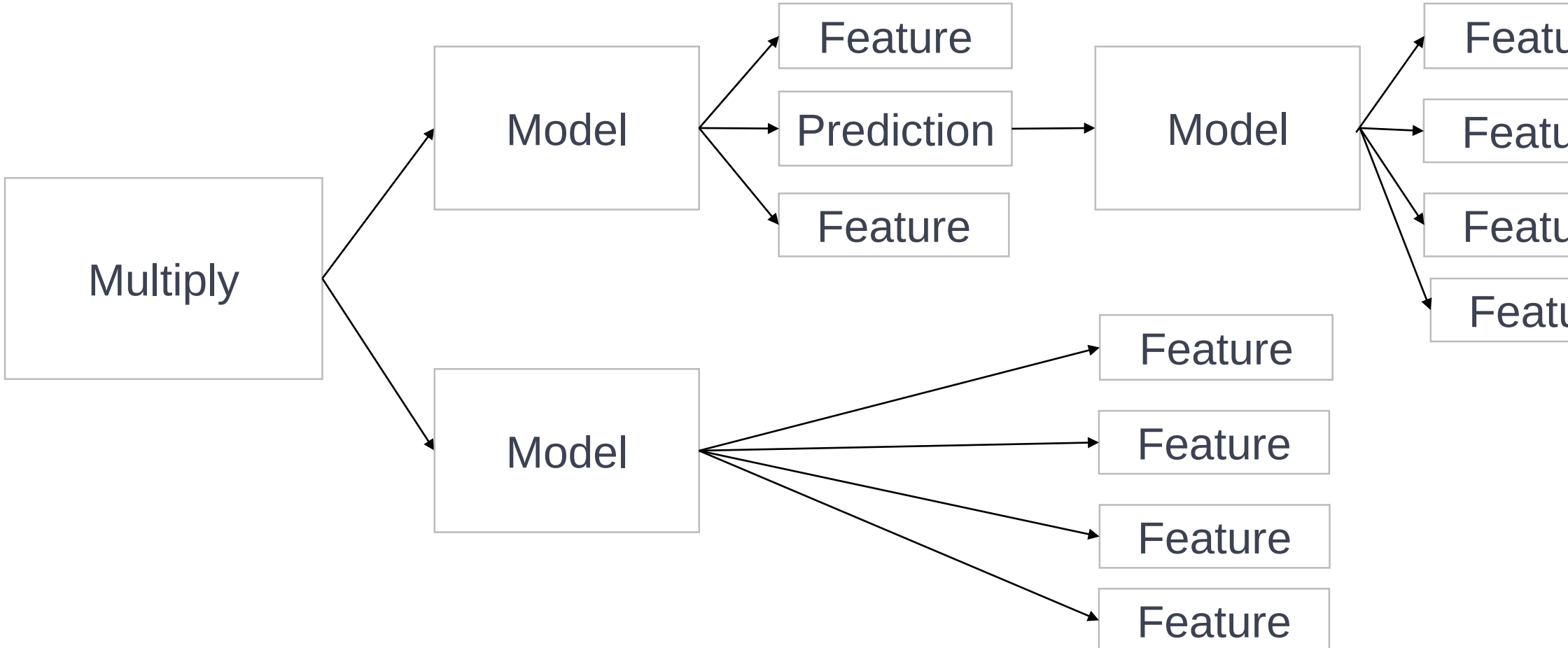
— Google paper ML test score

C# vs JVM

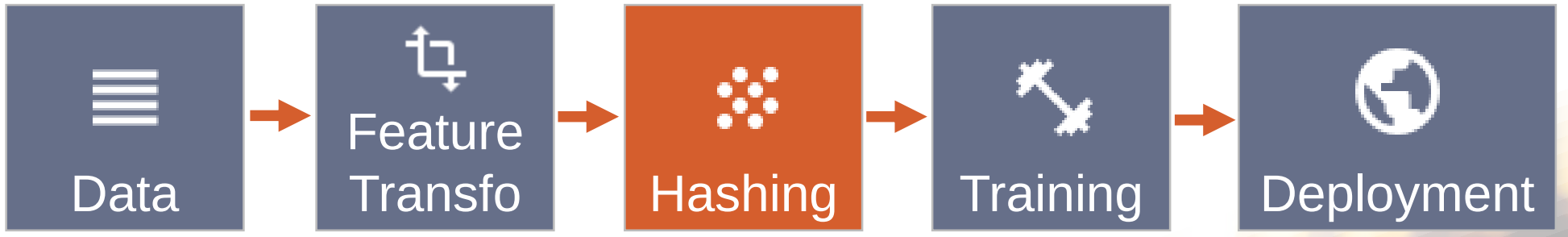
Invoking c# process with Spark pipe function

Feature Transformations

Everything is an ensemble



Hashing the data




Dataset is several TB big


Number of possible modalities
8 Billion

Encoding Categories

Feature User's Country

User's Country
France
Germany
Germany
..
Japan

Training examples

Encoding as Integer

Feature User's Country

User's Country
France = 1
Germany = 2
Germany = 2
..
Japan = 100

Encoding as Integer

Feature User's Country

User's Country
France = 1
Germany = 2
Germany = 2
..
Japan = 100

One hot encoding

Feature User's Country

User's Country
France
Germany
Germany
..
Japan

France	Germany	Japan	…
1	0	0	
0	1	0	
0	1	0	
	..		
0	0	1	

columns = number of distinct modalities

One hot encoding

Feature User's Country

France	Germany	Japan	Brasil	…	…	…	USA
1	0	0	0	0	..	0	0
0	1	0	0	0	..	0	0
0	1	0	0	0	..	0	0
	..						
0	0	1	0	0	..	0	0

190 columns

Hashing trick

Hash1	Hash2	Hash3	Hash4	Hash5	Hash6
1	0	0	1	0	0

France

Chrome Browser

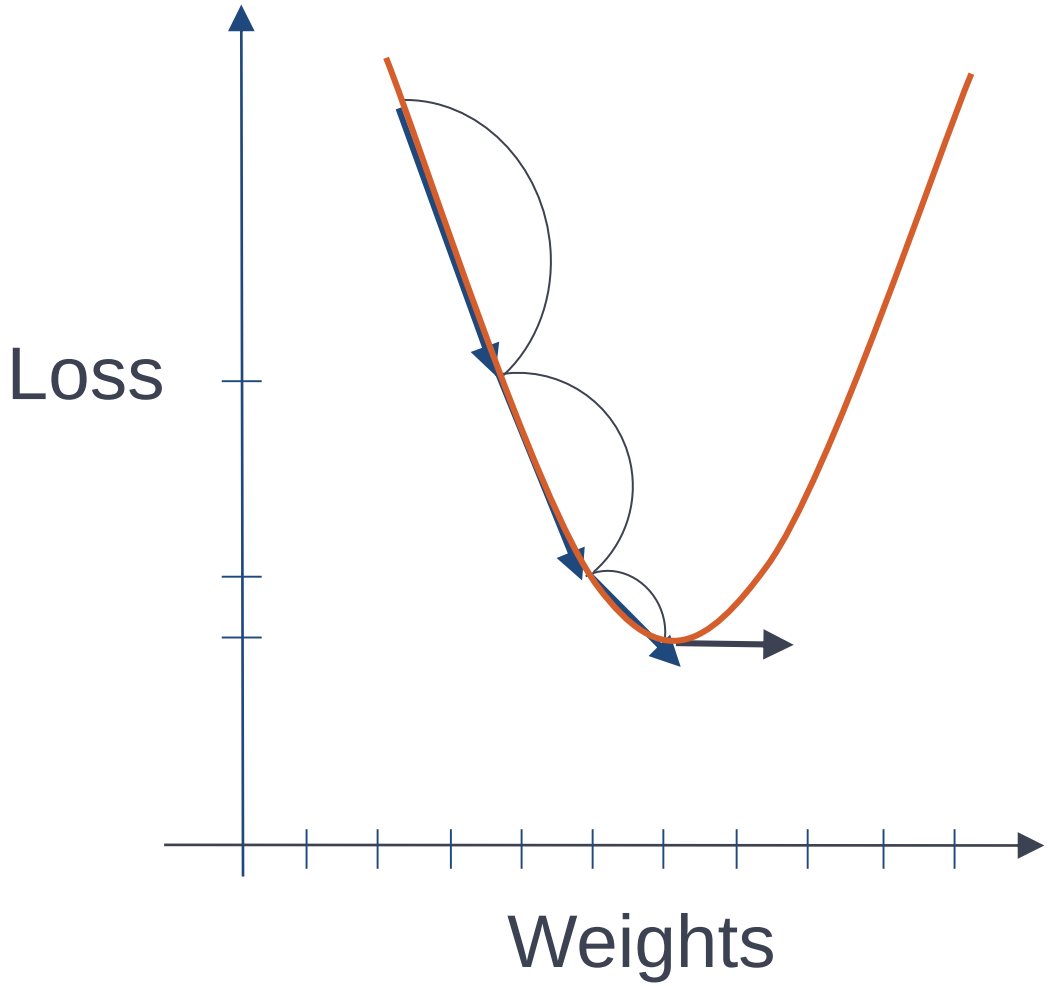
Red Shoes

With the Hashing trick
one feature vector has
68 Mio values
& takes **256 MB** in
memory

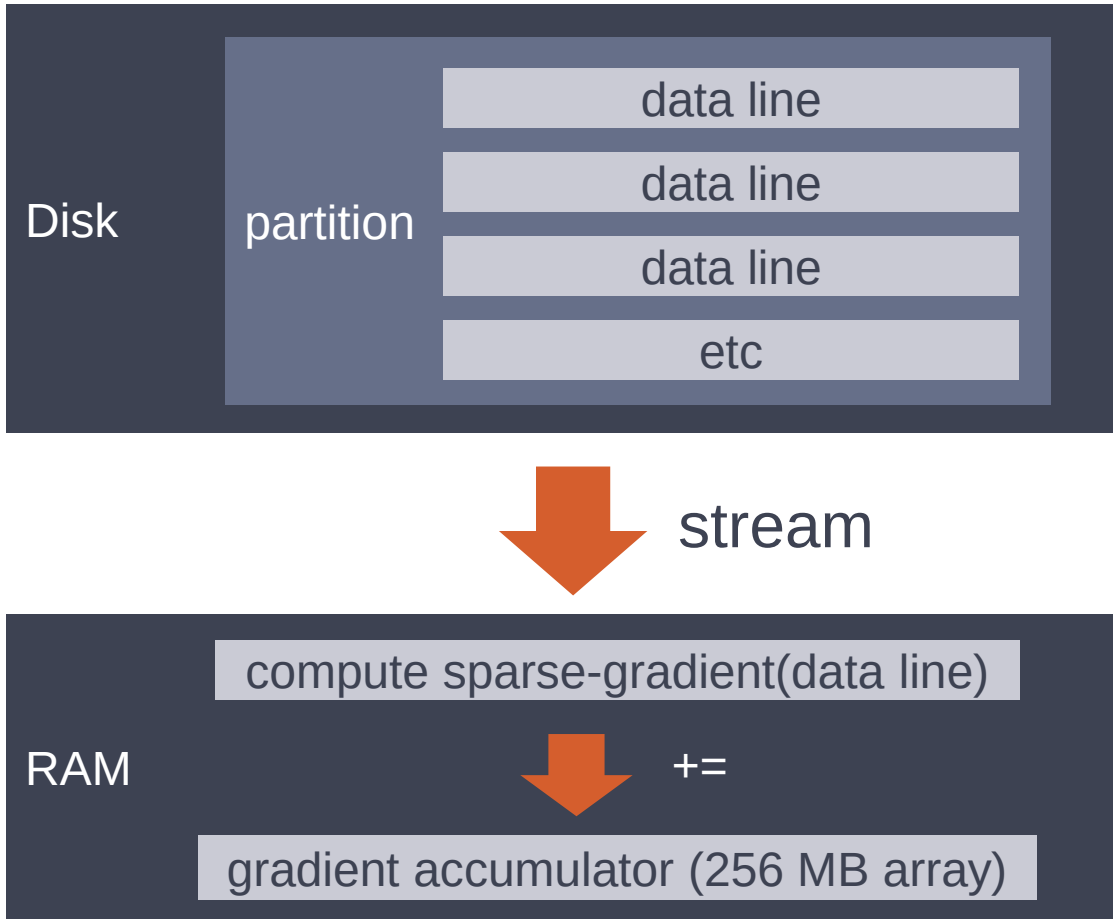

APACHE
Spark
Learning
Data
Feature
Transfo
Hashing
Training
Deployment

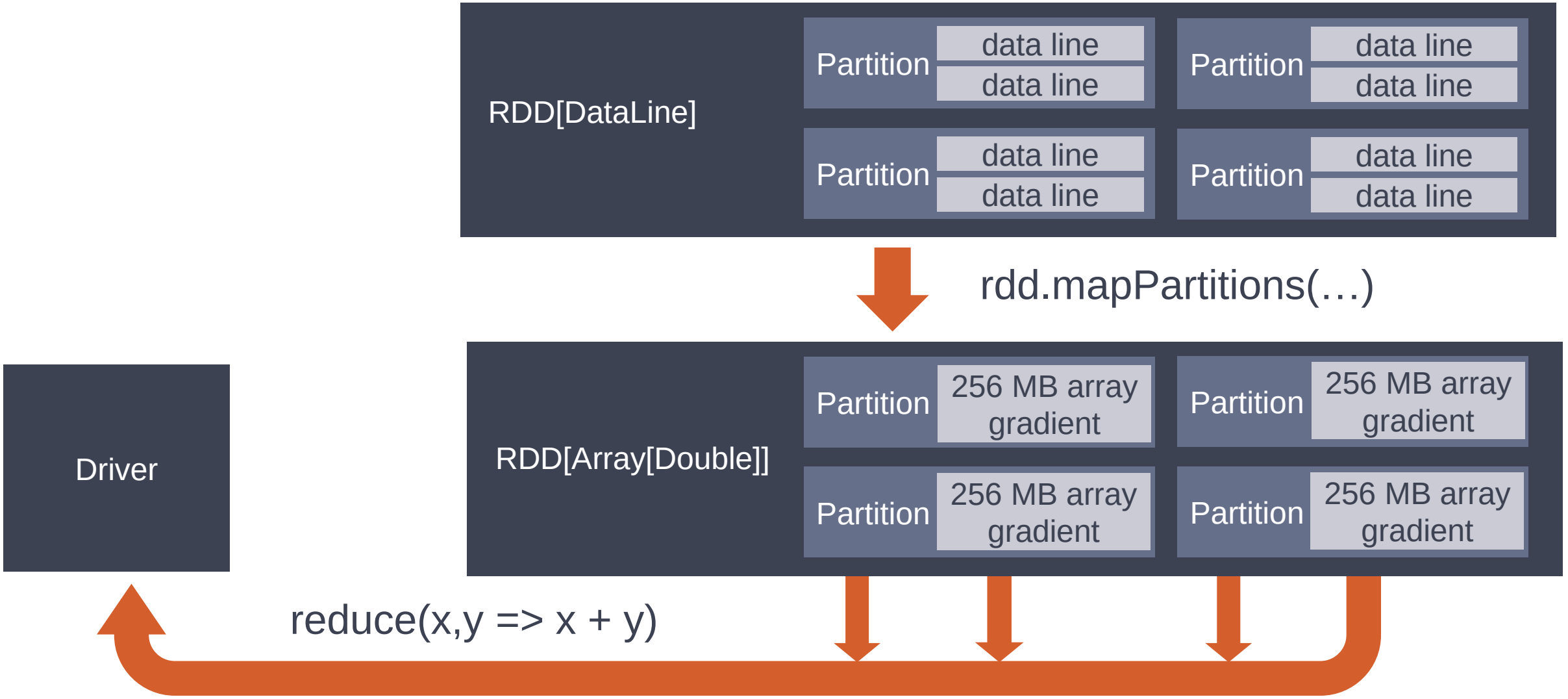
Logistic Regression

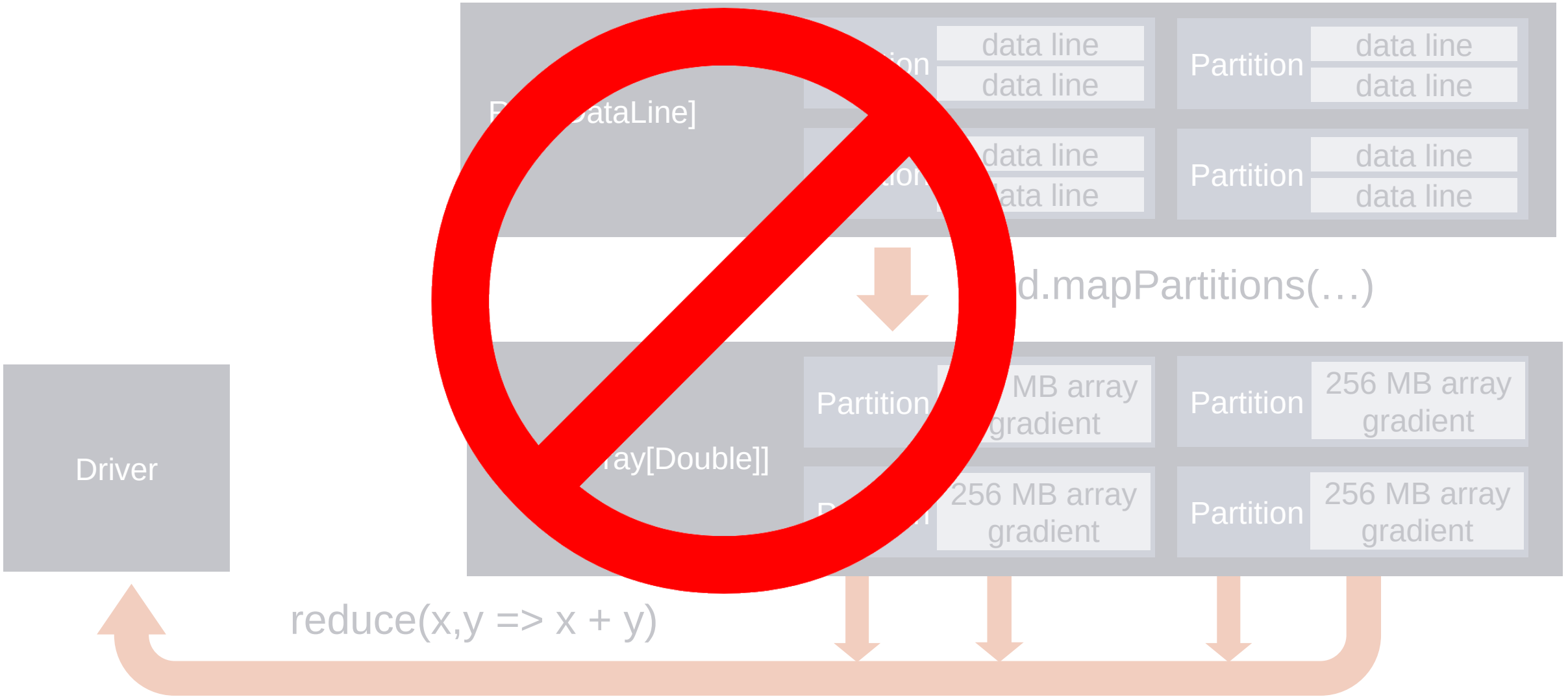
Gradient descent

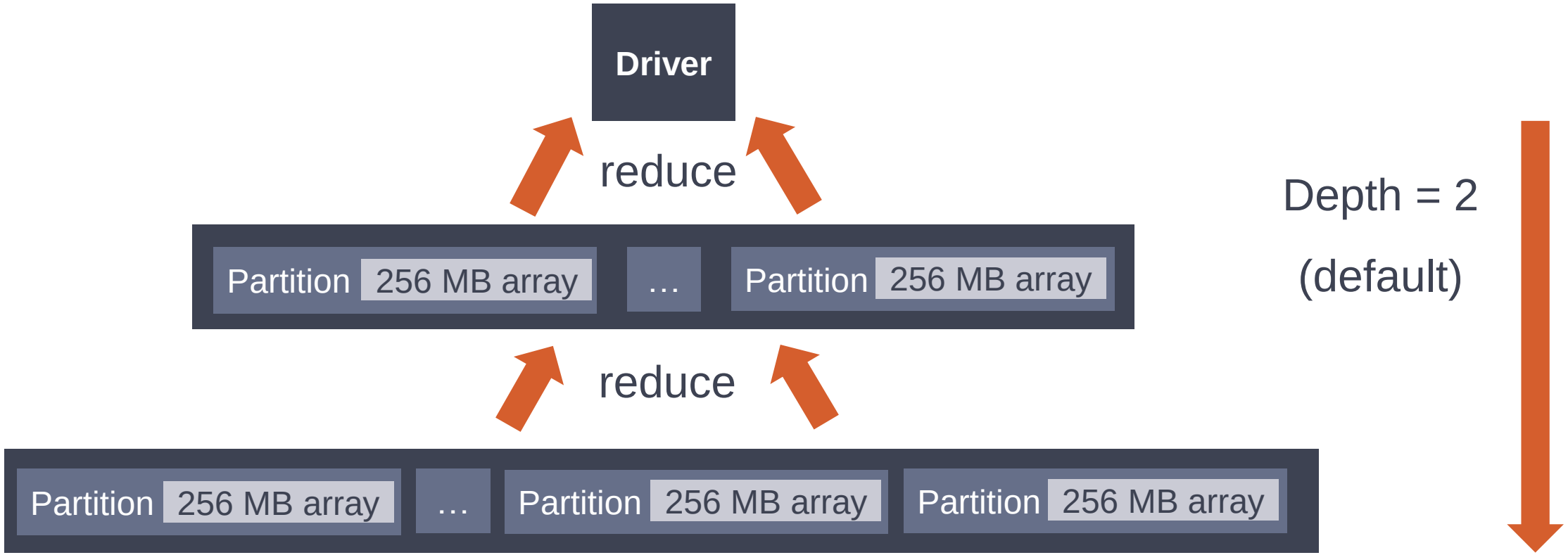


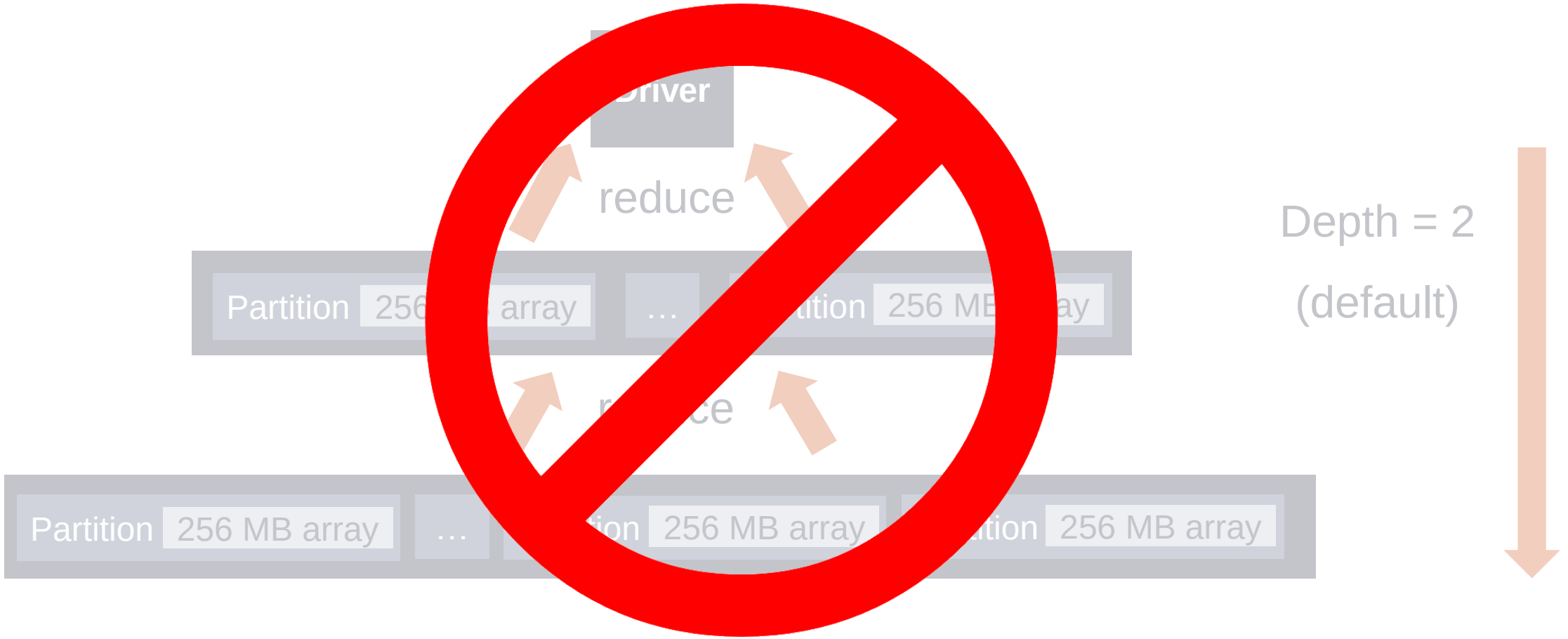
Computation for each partition

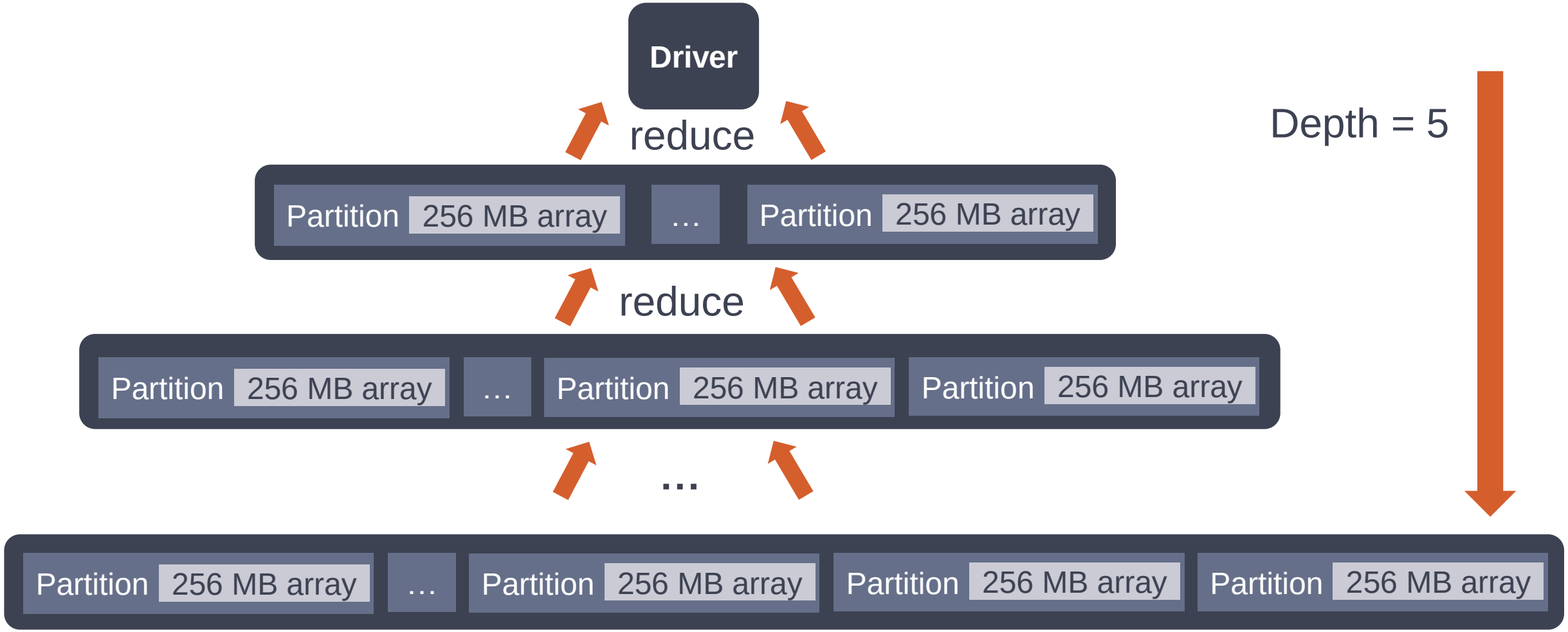


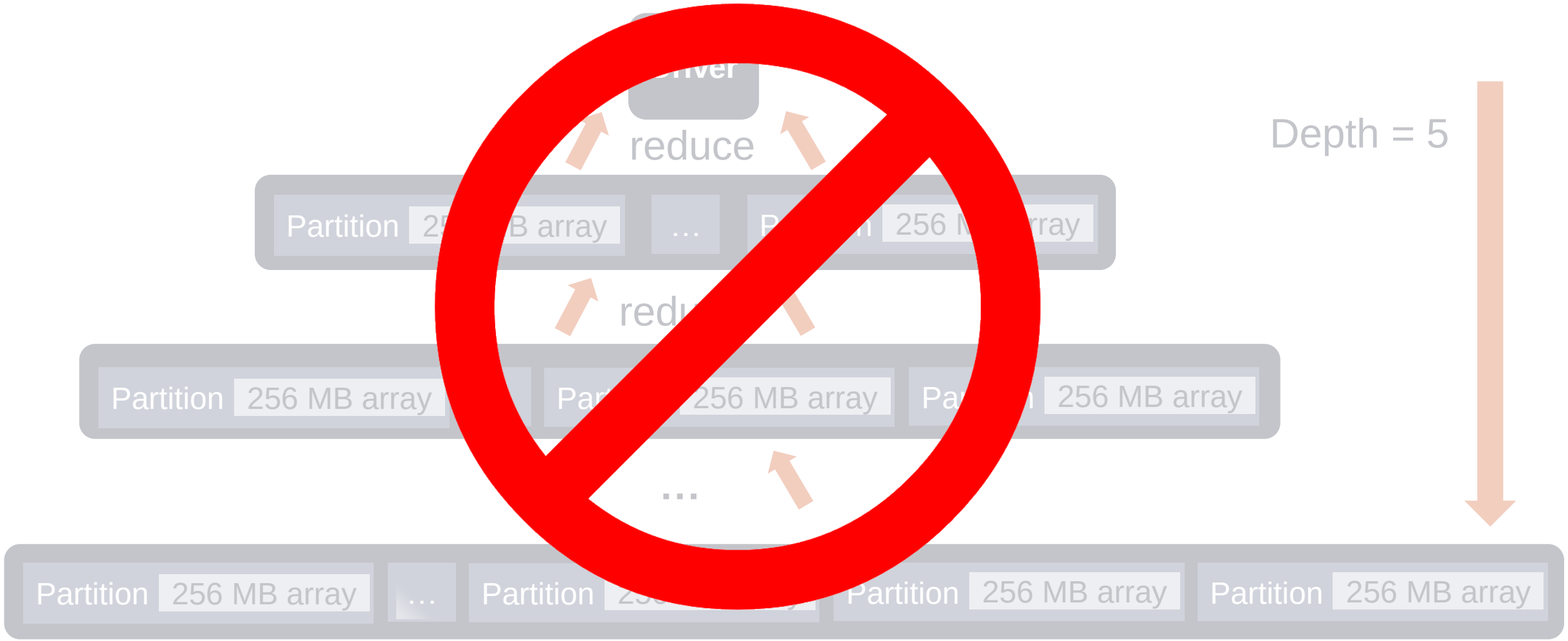

RDD[DataLine]
Partition
data line
data line
Partition
data line
data line
Partition
data line
data line
Partition
data line
data line
rdd.mapPartitions(…)
Driver
RDD[Array[Double]]
Partition
256 MB array gradient
Partition
256 MB array gradient
Partition
256 MB array gradient
Partition
256 MB array gradient
reduce(x,y => x + y)

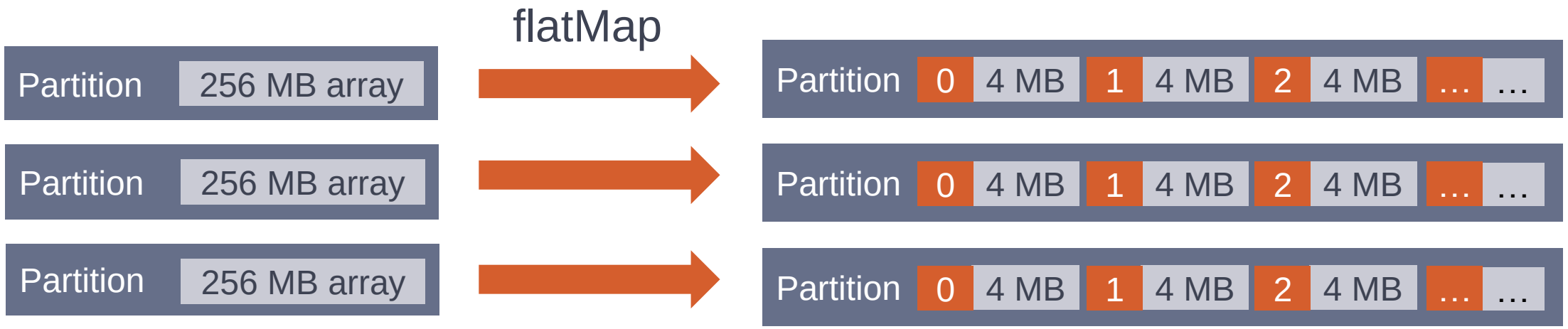

Partition
data line
data line
Partition
data line
data line
Partition
data line
data line
Partition
data line
data line
d.mapPartitions(…)
Partition
MB array
gradient
Partition
256 MB array
gradient
Partition
256 MB array
gradient
Partition
256 MB array
gradient
Driver
reduce(x,y => x + y)

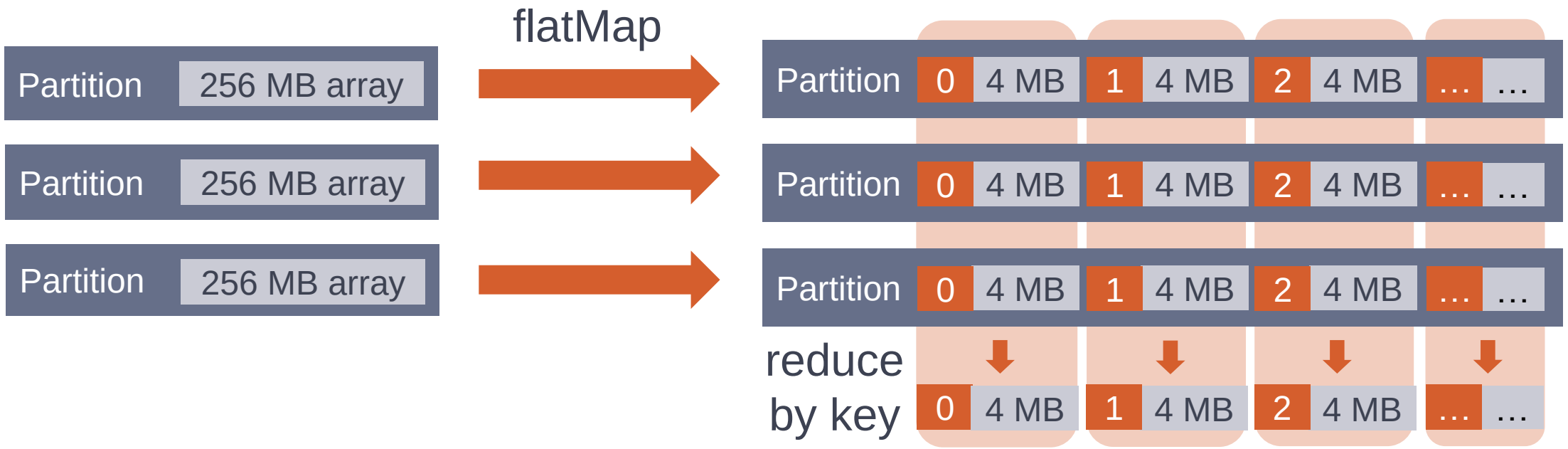

Driver
reduce
Partition 256 MB array
…
Partition 256 MB array
reduce
Partition 256 MB array
…
Partition 256 MB array
Partition 256 MB array
Depth = 2
(default)

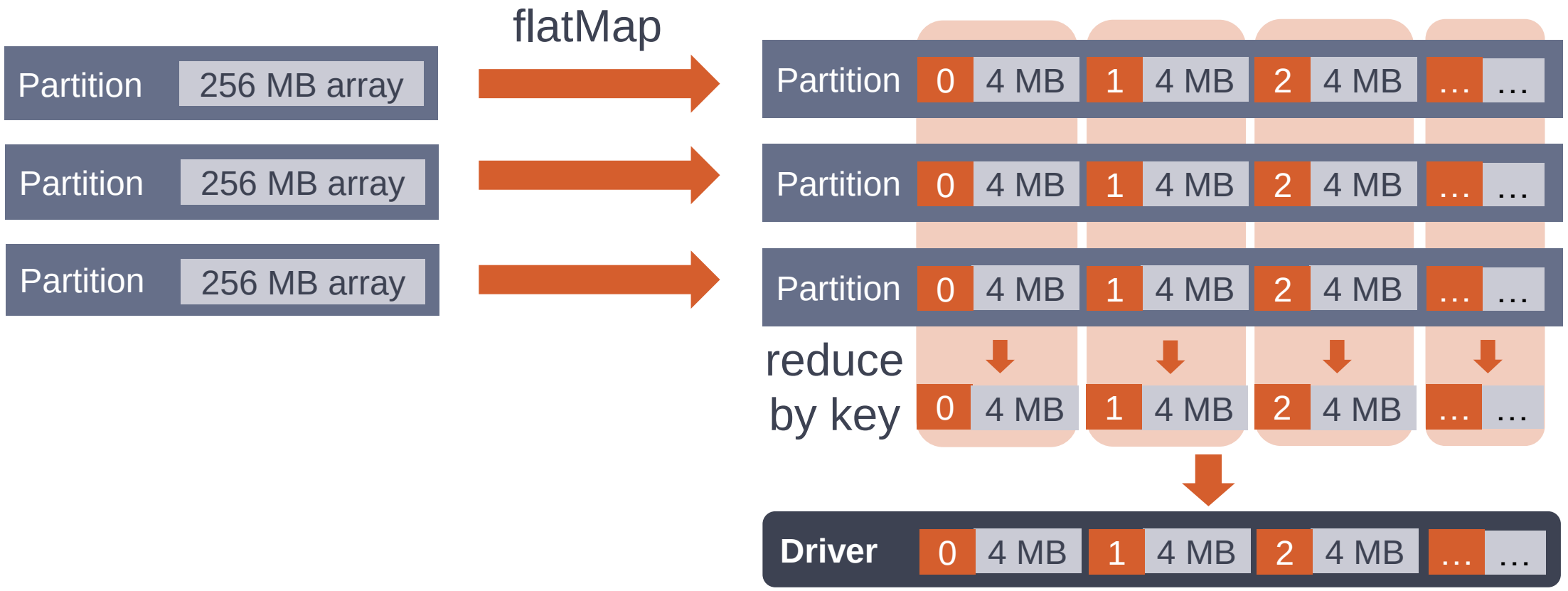

Driver
reduce
Partition 256 array
…
256 MB array
Partition 256 MB array
…
256 MB array
256 MB array
Depth = 2
(default)

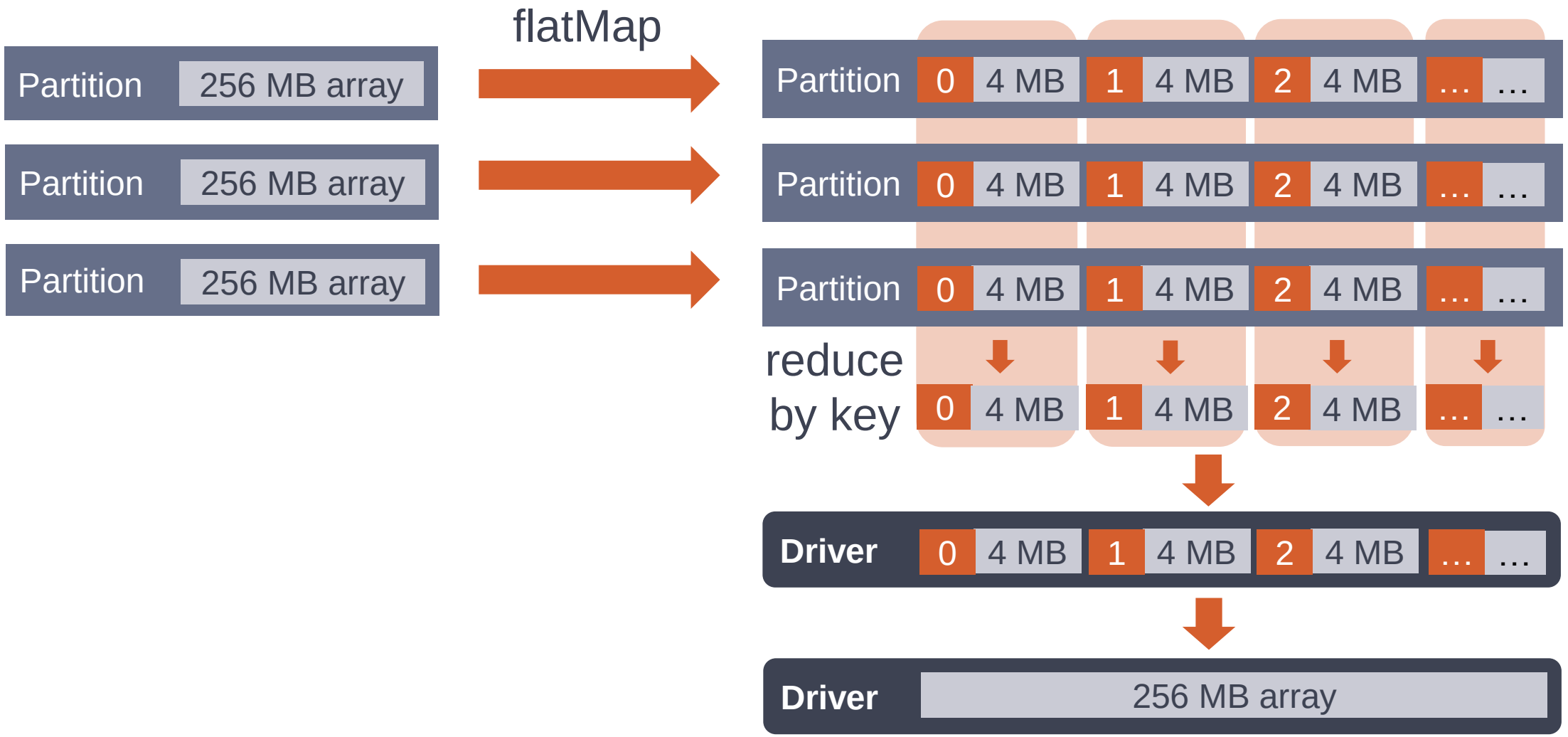

Driver
reduce
Partition 256 MB array
…
Partition 256 MB array
reduce
Partition 256 MB array
…
Partition 256 MB array
Partition 256 MB array
...
Partition 256 MB array
…
Partition 256 MB array
Partition 256 MB array
Partition 256 MB array
Depth = 5


reduce
Partition
…
reduce
Partition 256 MB array
Partition 256 MB array
Partition 256 MB array
…
Depth = 5
Partition 256 MB array
…
Partition
Partition 256 MB array
Partition 256 MB array


flatMap
Partition
256 MB array
Partition
256 MB array
Partition
256 MB array
Partition
0
4 MB
1
4 MB
2
4 MB
…
…
Partition
0
4 MB
1
4 MB
2
4 MB
…
…
Partition
0
4 MB
1
4 MB
2
4 MB
…
…

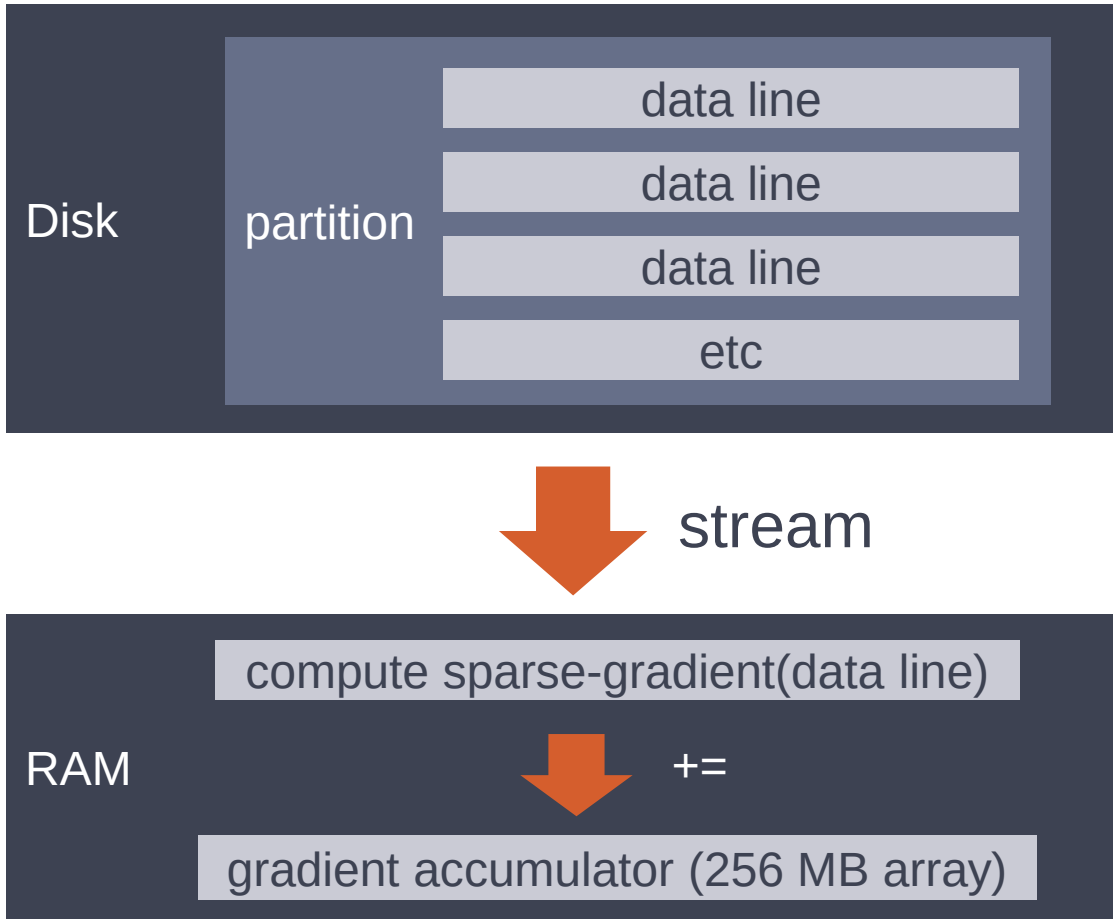

Partition
256 MB array
Partition
256 MB array
Partition
256 MB array
flatMap
Partition
0
4 MB
1
4 MB
2
4 MB
…
…
Partition
0
4 MB
1
4 MB
2
4 MB
…
…
Partition
0
4 MB
1
4 MB
2
4 MB
…
…
reduce
by key
0
4 MB
1
4 MB
2
4 MB
…
…


Partition
256 MB array
Partition
256 MB array
Partition
256 MB array
flatMap
Partition
0
4 MB
1
4 MB
2
4 MB
…
…
Partition
0
4 MB
1
4 MB
2
4 MB
…
…
Partition
0
4 MB
1
4 MB
2
4 MB
…
…
reduce
by key
0
4 MB
1
4 MB
2
4 MB
…
…
Driver
0
4 MB
1
4 MB
2
4 MB
…
…


Partition
256 MB array
Partition
256 MB array
Partition
256 MB array
flatMap
Partition
0
4 MB
1
4 MB
2
4 MB
…
…
Partition
0
4 MB
1
4 MB
2
4 MB
…
…
Partition
0
4 MB
1
4 MB
2
4 MB
…
…
reduce
by key
0
4 MB
1
4 MB
2
4 MB
…
…
Driver
0
4 MB
1
4 MB
2
4 MB
…
…
Driver
256 MB array


Out Of Memory — Java heap space

Computation for each partition



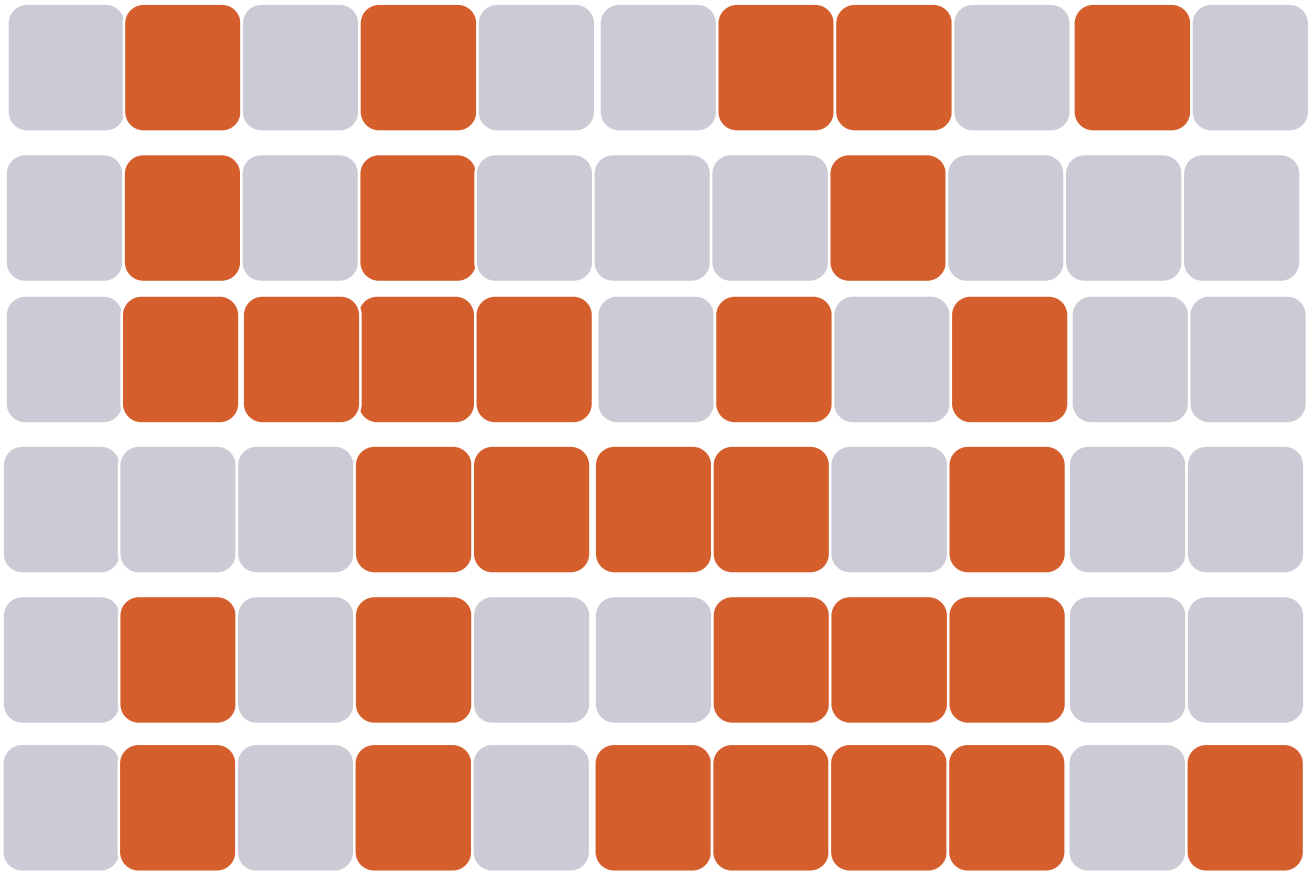
Multiple weeks of investigation

Activated G1 GC logs on Executor

-XX:+PrintGCDetails
-XX:+PrintGCTimeStamps
-XX:+PrintHeapAtGC
-XX:+PrintTenuringDistribution
-XX:+UnlockExperimentalVMOptions
-XX:+UnlockDiagnosticVMOptions
-XX:G1LogLevel=finest
-XX:+G1PrintRegionLivenessInfo

Humongous Object

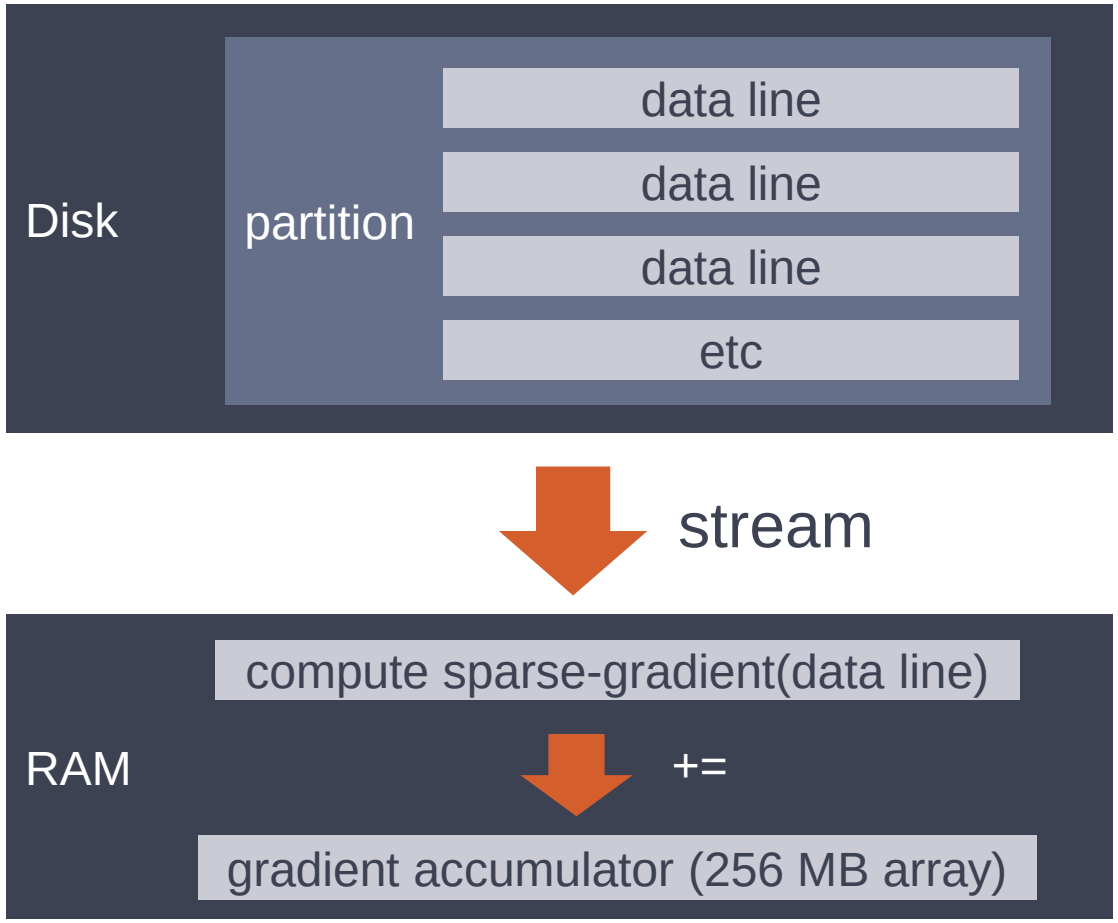
Fragmented Heap



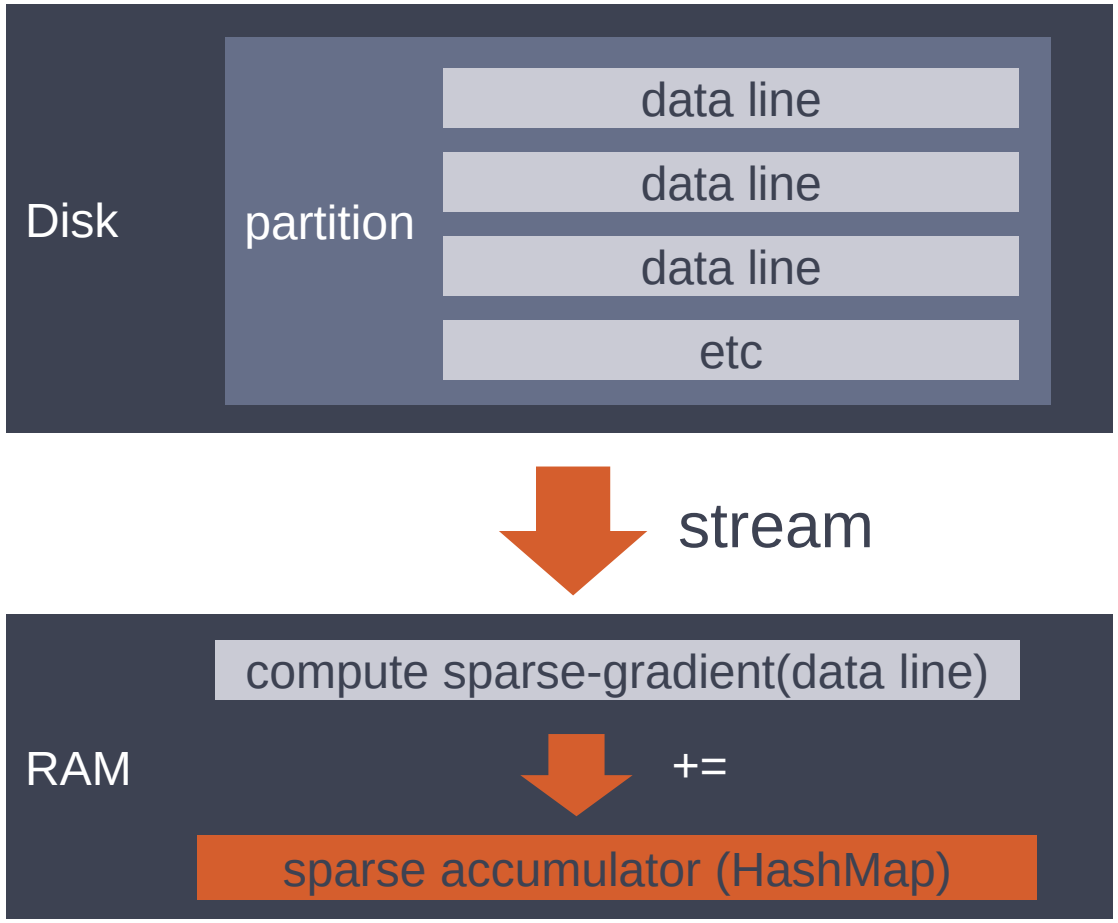
Taken

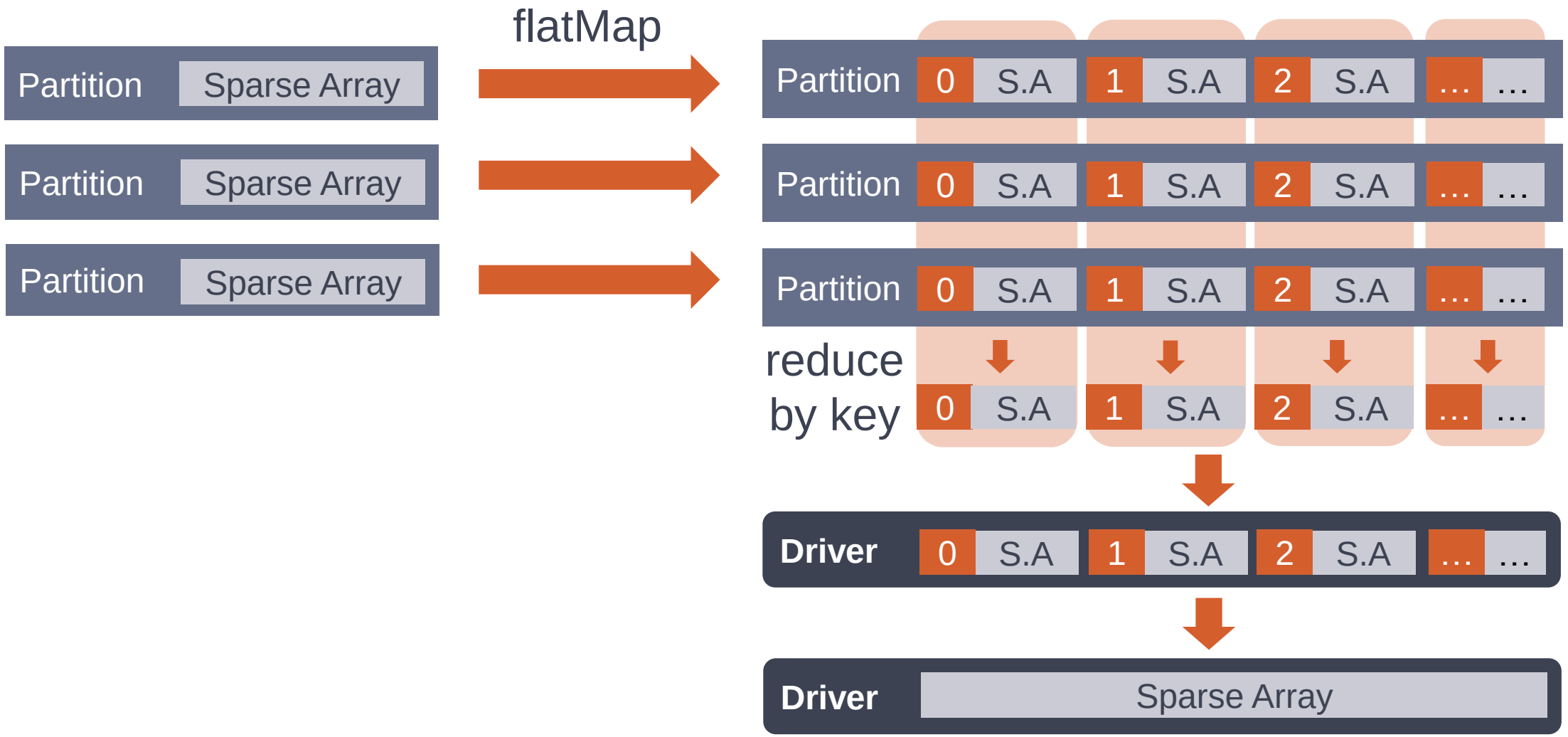
Free

Computation for each partition



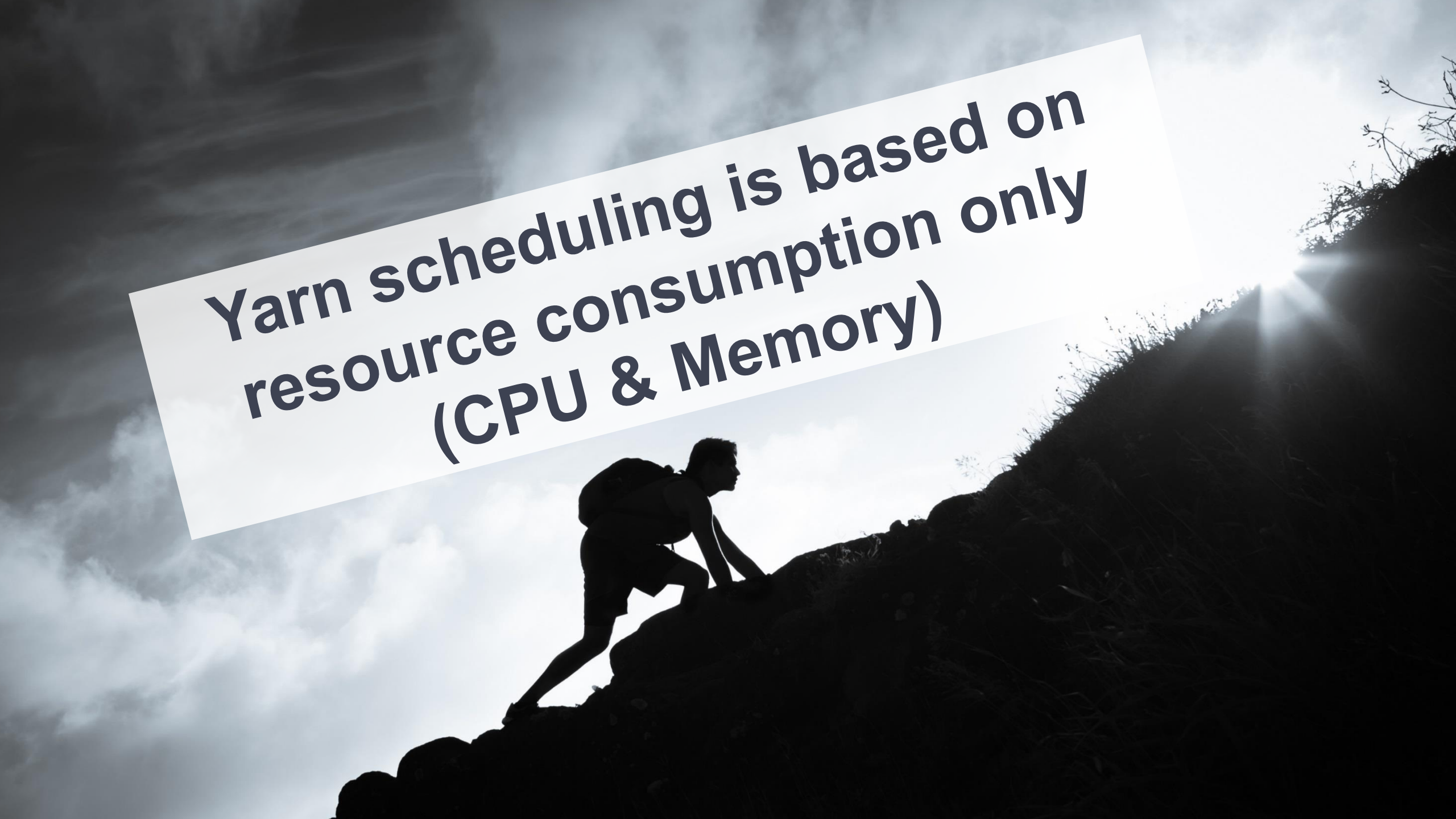
Computation for each partition

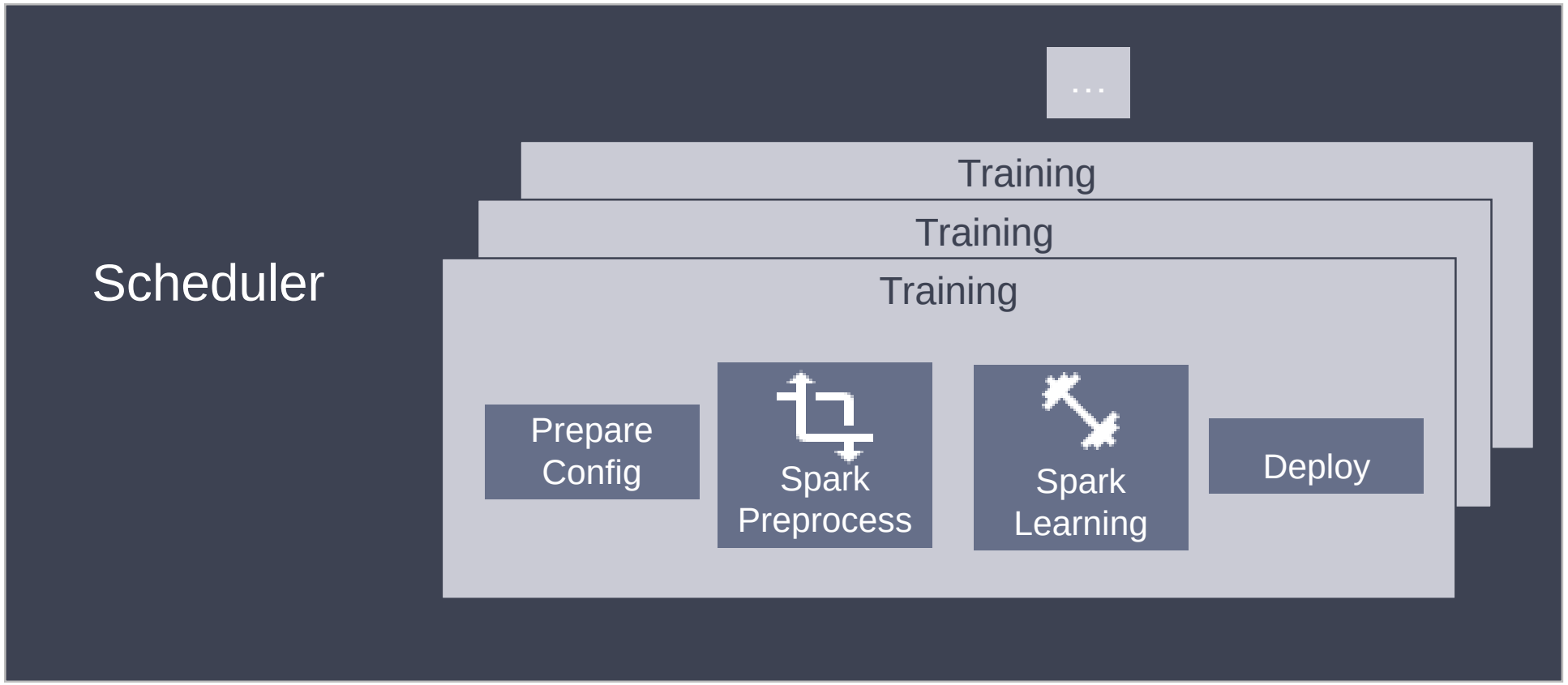


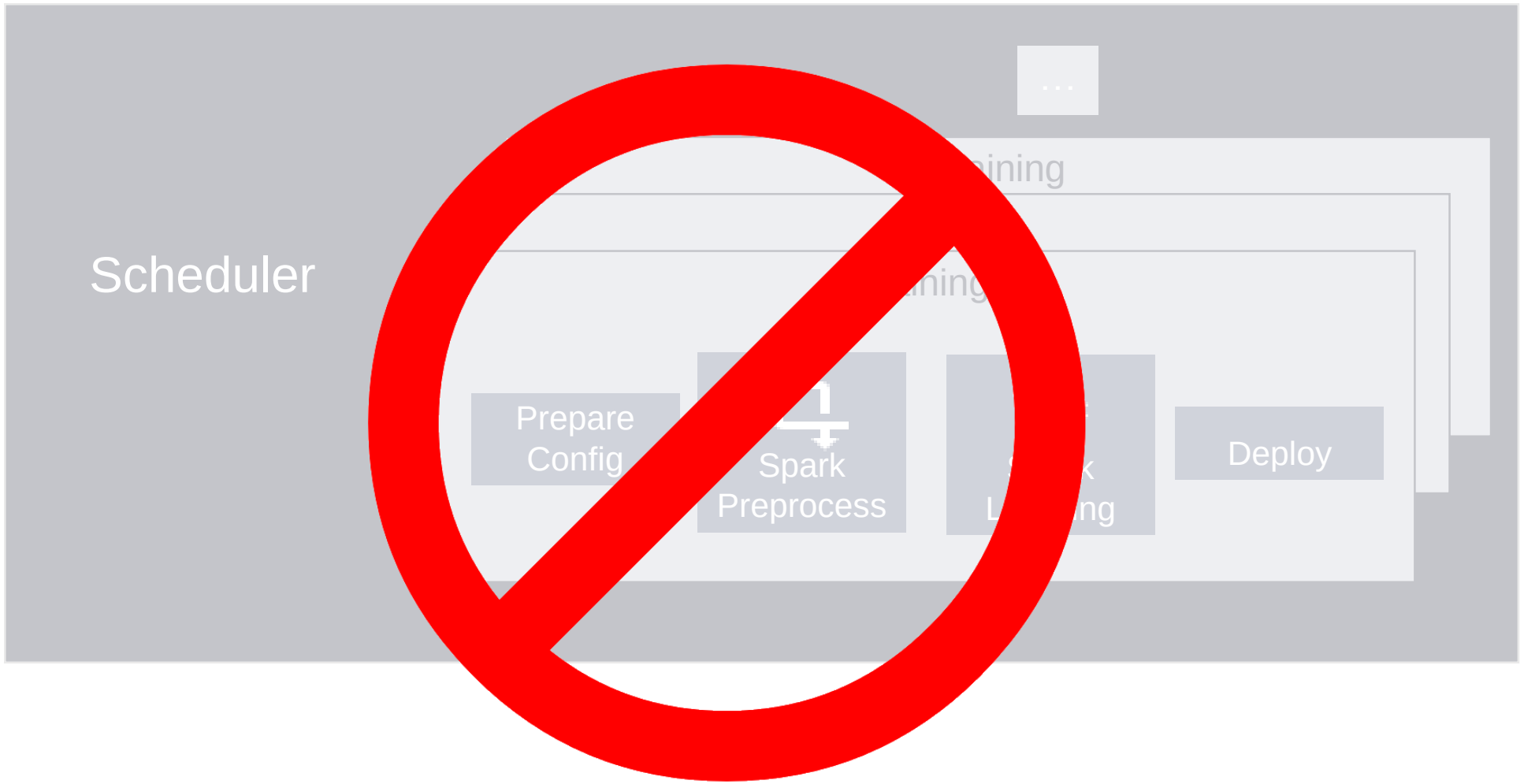

Partition
Sparse Array
Partition
Sparse Array
Partition
Sparse Array
flatMap
Partition
0
S.A
1
S.A
2
S.A
…
…
Partition
0
S.A
1
S.A
2
S.A
…
…
Partition
0
S.A
1
S.A
2
S.A
…
…
reduce
by key
0
S.A
1
S.A
2
S.A
…
…
Driver
0
S.A
1
S.A
2
S.A
…
…
Driver
Sparse Array

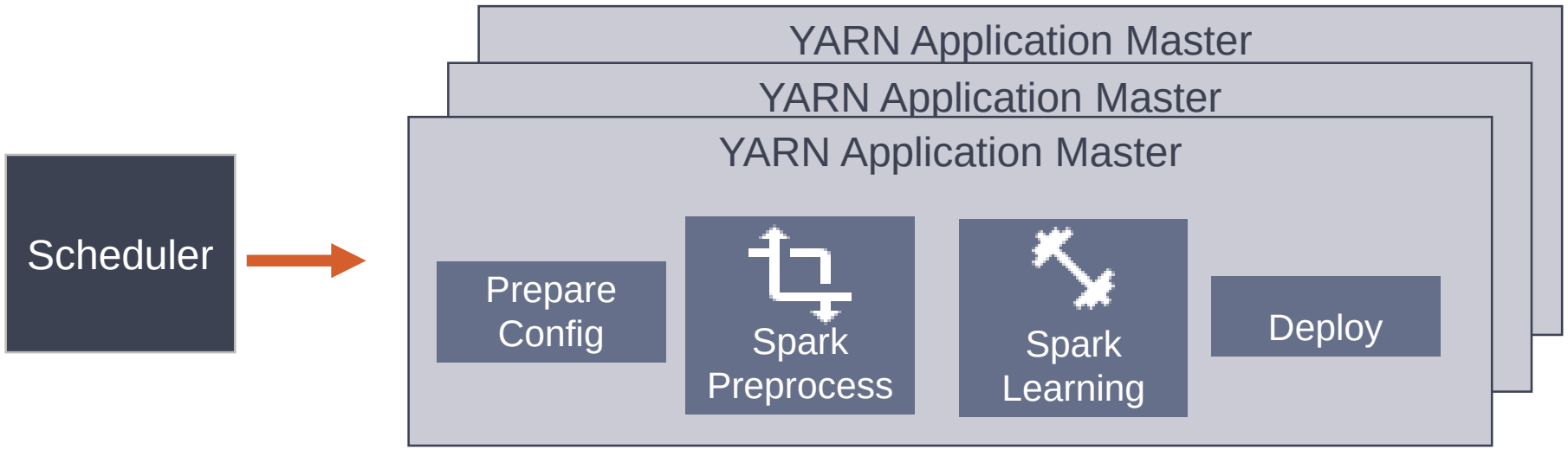
Learning 1000s of models

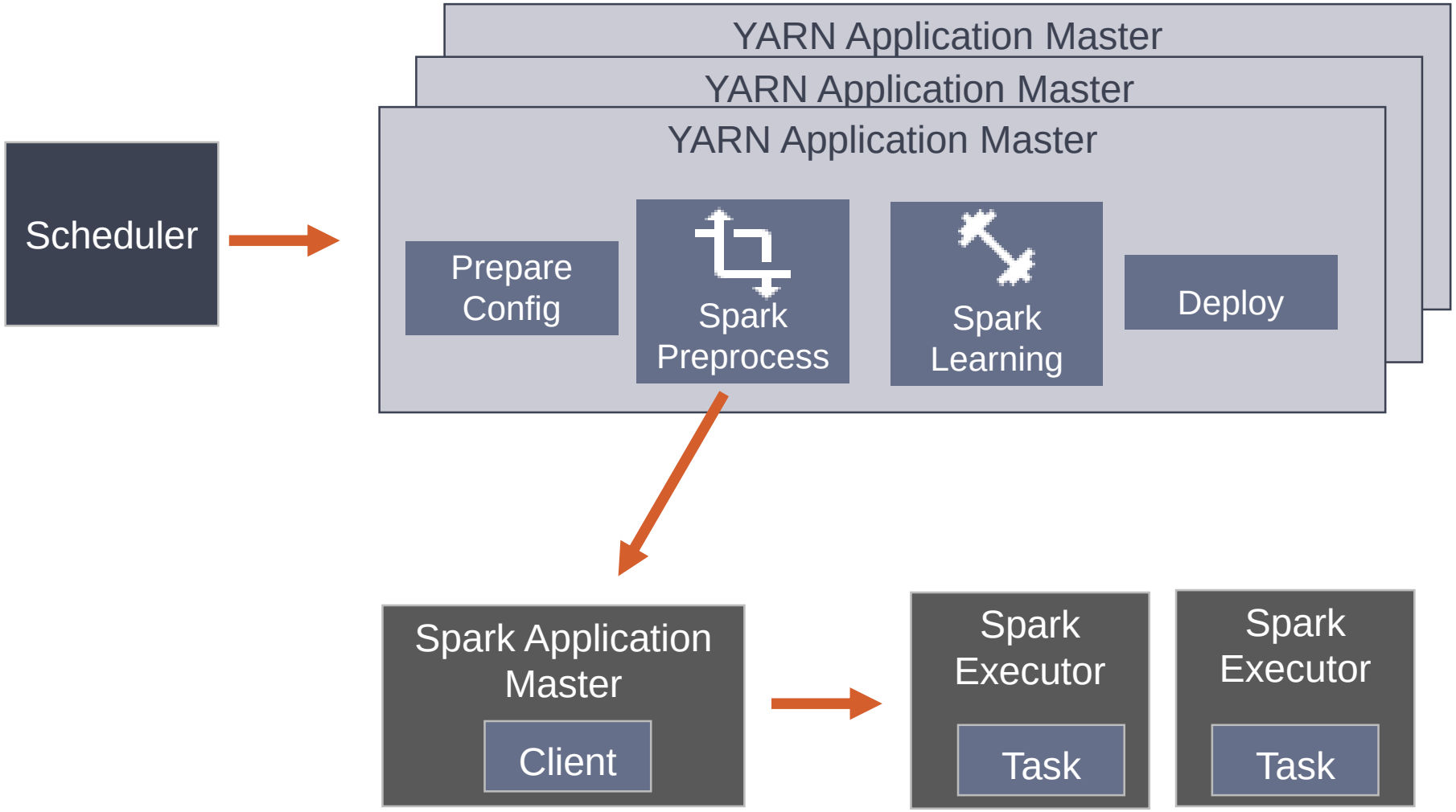

Several 1000s of models must be learned every day


Yarn scheduling is based on
resource consumption only
(CPU & Memory)

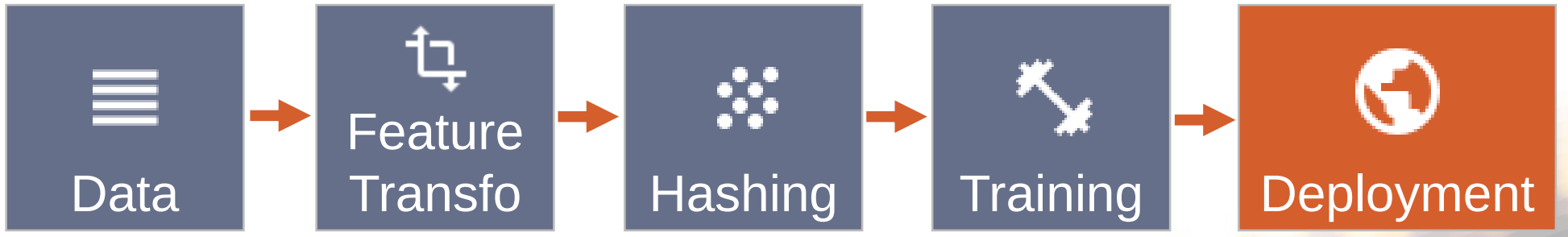

Scheduler
...
Training
Training
Training
Prepare Config
Spark Preprocess
Spark Learning
Deploy


Scheduler
...
Prepare Config
Spark Preprocess
Deploy


Scheduler
YARN Application Master
YARN Application Master
YARN Application Master
Prepare Config
Spark Preprocess
Spark Learning
Deploy


Scheduler
YARN Application Master
YARN Application Master
YARN Application Master
Prepare Config
Spark Preprocess
Spark Learning
Deploy
Spark Application Master
Client
Spark Executor
Task
Spark Executor
Task

Model deployment

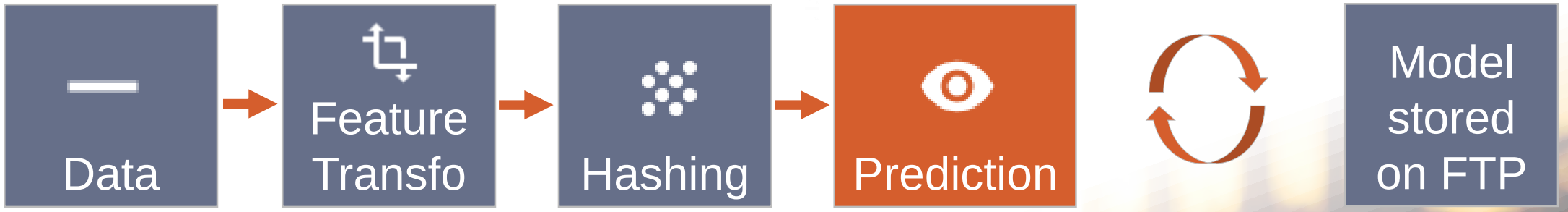



PA4

Online prediction


Prediction must be made in less then 100µs

Online Prediction Workflow



10 Million

predictions per second

Monitoring

Finatra Prometheus Graphite

Questions

f.horing@criteo.com
@f_hoering