Spark

Cluster Computing & Machine Learning

Philosophy

„In pioneer days they used oxen for heavy pulling, and when one ox couldn’t budge a log, they didn’t try to grow a larger ox. We shouldn’t try for bigger computers, but for more systems of computers.“

— Grace Hopper

What is Spark?

Fast and general-purpose *cluster* computing system

Scales to thousands of nodes

In-Memory processing

Typically runs on top of existing cluster infrastructure (e.g. Hadoop, EMR)

Scala, Java, Python, R

Programming paradigm

Spark 1.x: RDD

Spark 2.x: RDD, DataSet, DataFrame

Programming paradigm

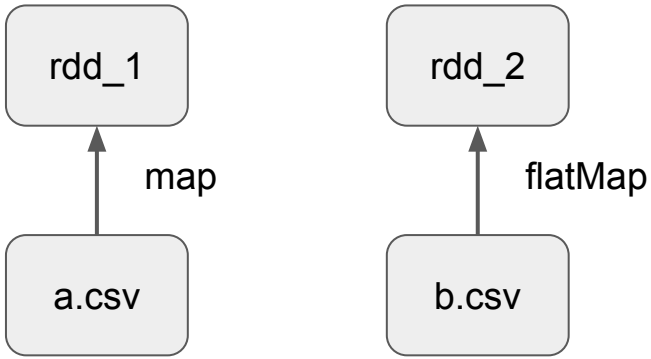
Spark 1.x: RDD

Spark 2.x: RDD, DataSet, DataFrame

Programming paradigm

Spark 1.x: RDD

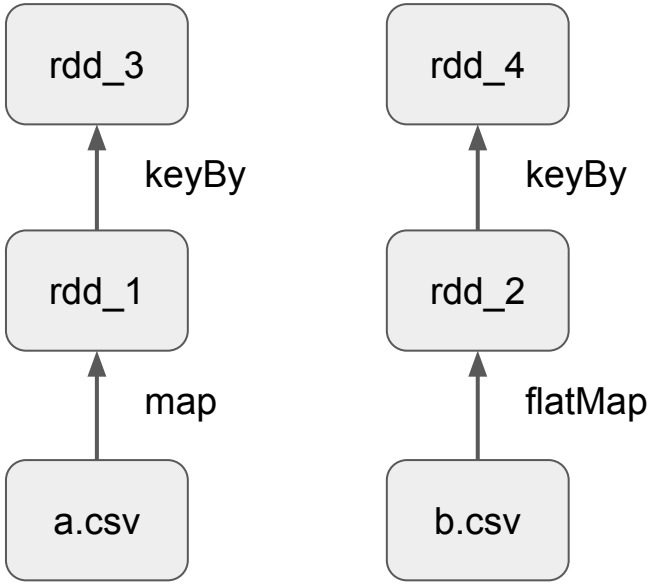
Spark 2.x: RDD, DataSet, DataFrame



Programming paradigm

Spark 1.x: RDD

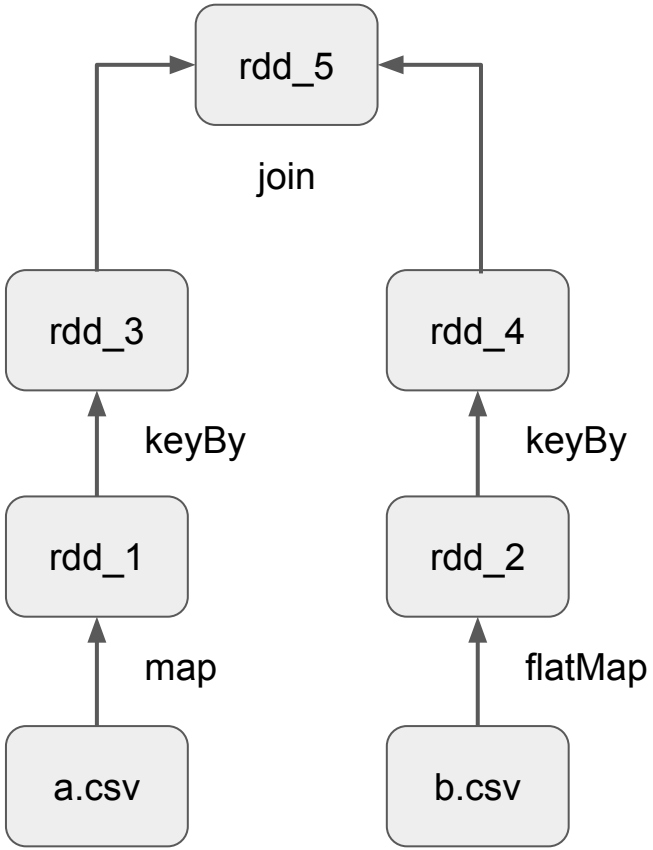
Spark 2.x: RDD, DataSet, DataFrame



Programming paradigm

Spark 1.x: RDD

Spark 2.x: RDD, DataSet, DataFrame



API example

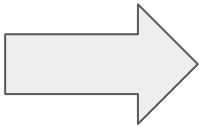
```
val inputRDD = sc.textFile("input.txt")
val wordsRDD = inputRDD.flatMap(line => line.split(" "))
val keyValueRDD = wordsRDD.map(a => (a, 1))
val countRDD = keyValueRDD.reduceByKey((a, b) => a+b)

countRDD.collect()
```

DEMO

the quick brown fox jumps over the lazy dog
the lazy dog is fast asleep
over the lazy dog jumps the quick brown fox

input.txt



is	1
fast	1
lazy	3
dog	3
over	2
...	...

Architecture

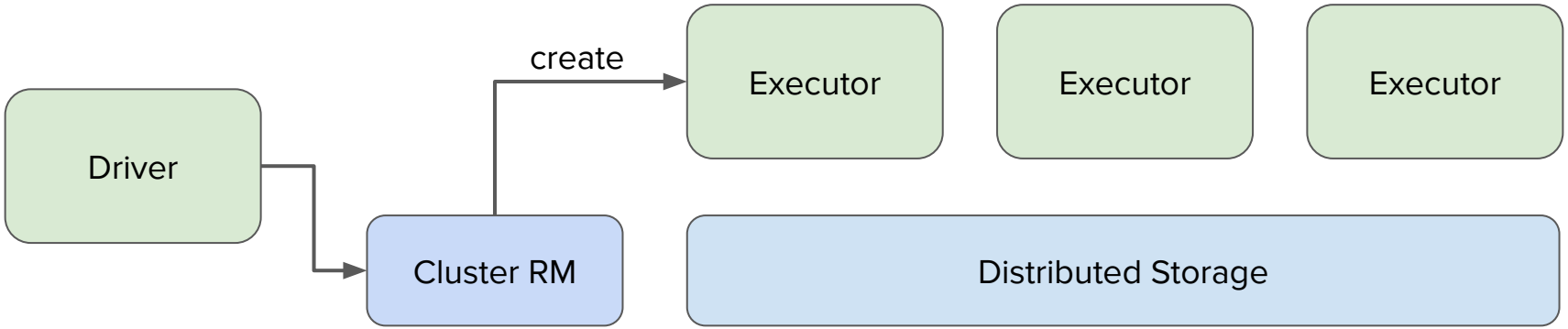
Driver

Cluster RM

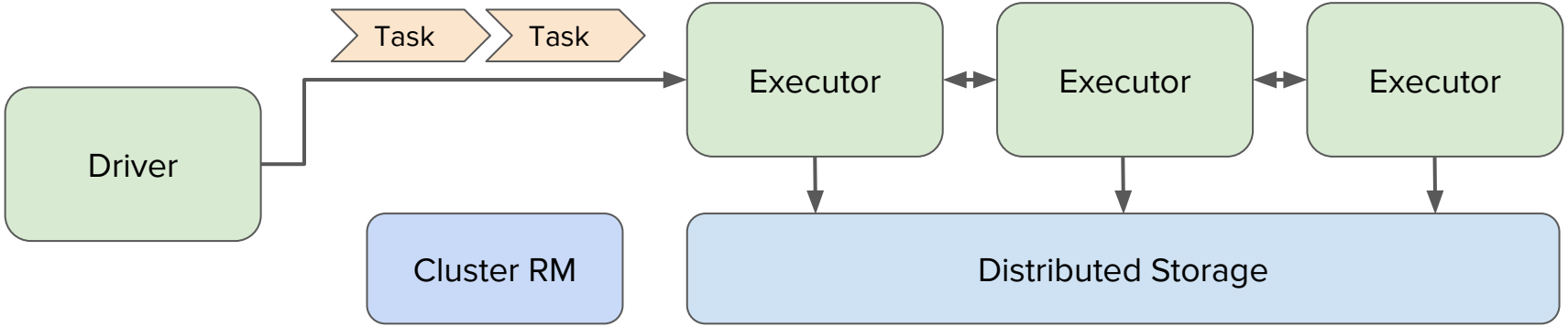
Distributed Storage

Architecture

Architecture



Architecture



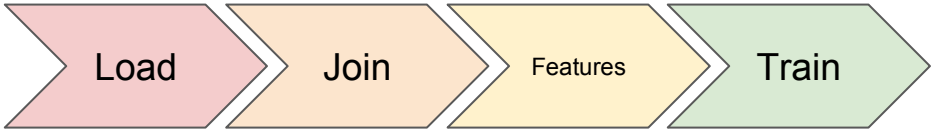
Spark MLlib - Machine Learning with Spark

ML Algorithms: common learning algorithms such as classification, regression, clustering, and collaborative filtering
Featurization: feature extraction, transformation, dimensionality reduction, and selection
Pipelines: tools for constructing, evaluating, and tuning ML Pipelines
Persistence: saving and load algorithms, models, and Pipelines
Utilities: linear algebra, statistics, data handling, etc.

Example: Nettebad - predict number of visitors

Pipeline:

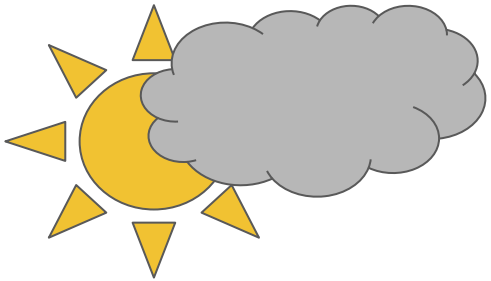
- Load pool and weather data
- Join data sets
- Extract features
- Train model with cross validation



Use model:

- Make predictions
- Save model

DEMO



predict

2005-03-20	915
2005-03-22	1482
2005-03-24	1790
2005-03-26	1471
2005-03-28	1942
...	...