Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

intel

AI and Predictive Analytics in Data-Center Environments

Distributed Computing using Spark

SparkSQL and SparkML

Josep Ll. Berral @BSC

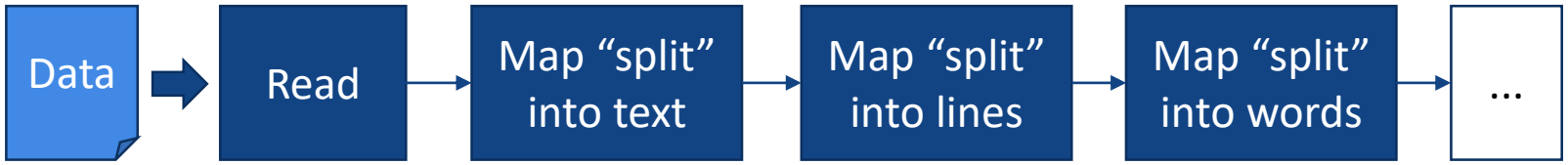
Intel Academic Education Mindshare Initiative for AI

SparkSQL and SparkML

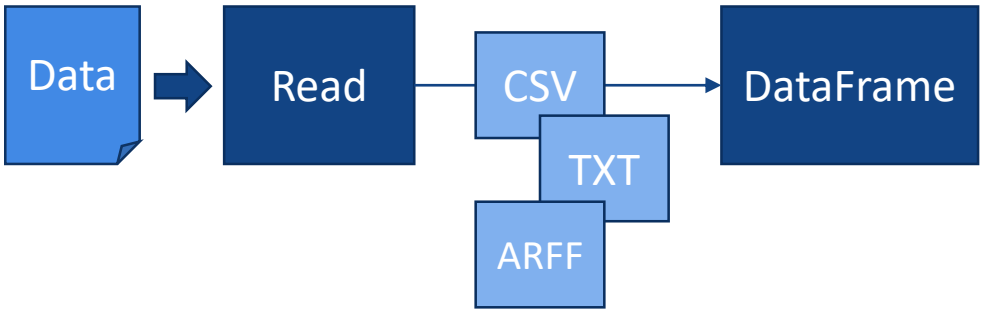
Spark has libraries of Distributed Data Aggregation and Machine Learning

Loading and Selecting Data

- Original versions of Spark
  - Read files and parse to create RDDs

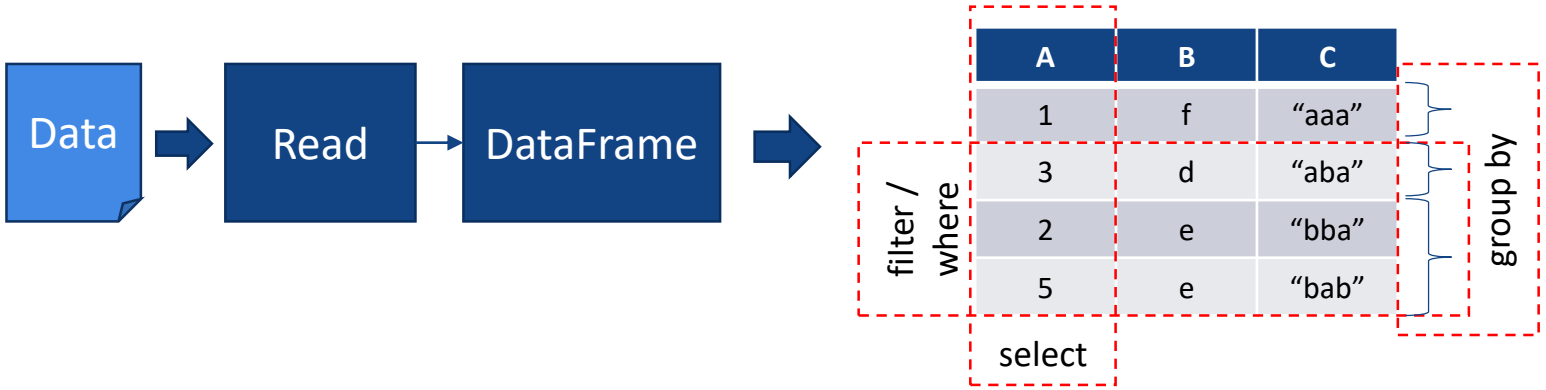


- Structured Data
  - Readers and Writters!



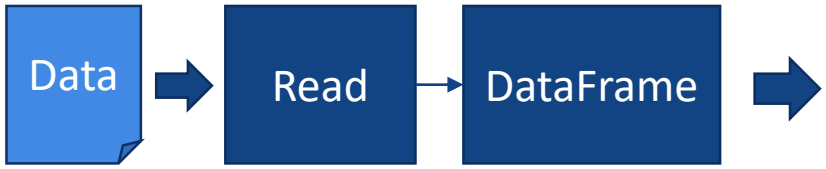
Spark SQL

- Relation operations like with SQL queries
  - DataFrames -> SELECT / GROUP BY / FILTER / ...
  - Operations -> COUNT / SHOW / ...
- ...or just apply SQL queries
  - DataFrames -> SELECT / GROUP BY / WHERE / ...
  - Operations -> COUNT / SHOW / ...



Spark SQL

- Quick SQL functions:
  - SELECT : indicate which features of our DataFrame we want to use
  - FILTER | WHERE : indicate which examples/rows we want to include or exclude
  - COUNT / AVG / ... : operations of aggregation (count, mean, ...)
  - GROUP BY : indicate if we want to aggregate by some criteria
  - ...



A	B	C
1	f	“aaa”
3	d	“aba”
2	e	“bba”
5	e	“bab”

Basics of SparkML

- Machine Learning operations and types

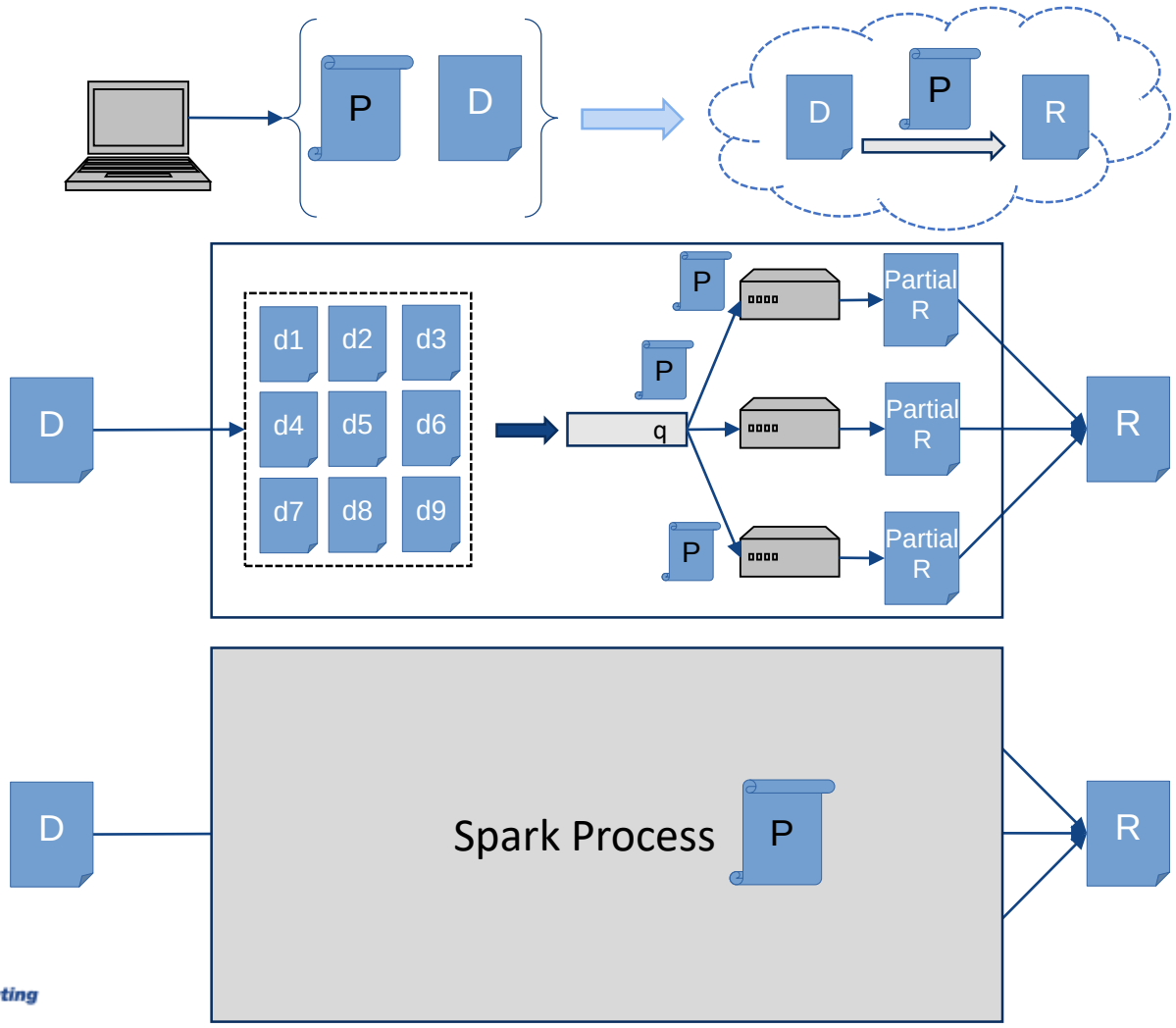
- Distributed ML Algorithms
  - Basic Statistics
  - Summaries, Correlations, Hypothesis Tests...
  - Classification and Prediction
    - E.g. Linear Regression, Support Vector Machines, NNs
  - Clustering
    - E.g. k-means
  - Dimension Reduction
    - E.g. Principal Component Analysis
  - Collaborative Filtering
  - Frequent Pattern Mining
  - ...

Distributed ML

- Spark takes advantage of splitting data in subsets
  - Subsets are distributed and processed for models
  - Partial Models are aggregated into a general model
  - Such methodology is not as fitted as centralized approaches...
  - ... But at least can be processed
  - ... Also, we could discuss how huge datasets could bring to statistically significant sampled subsets
- ML process relies on a Map/Reduce strategy
  - ...but it is invisible for us!

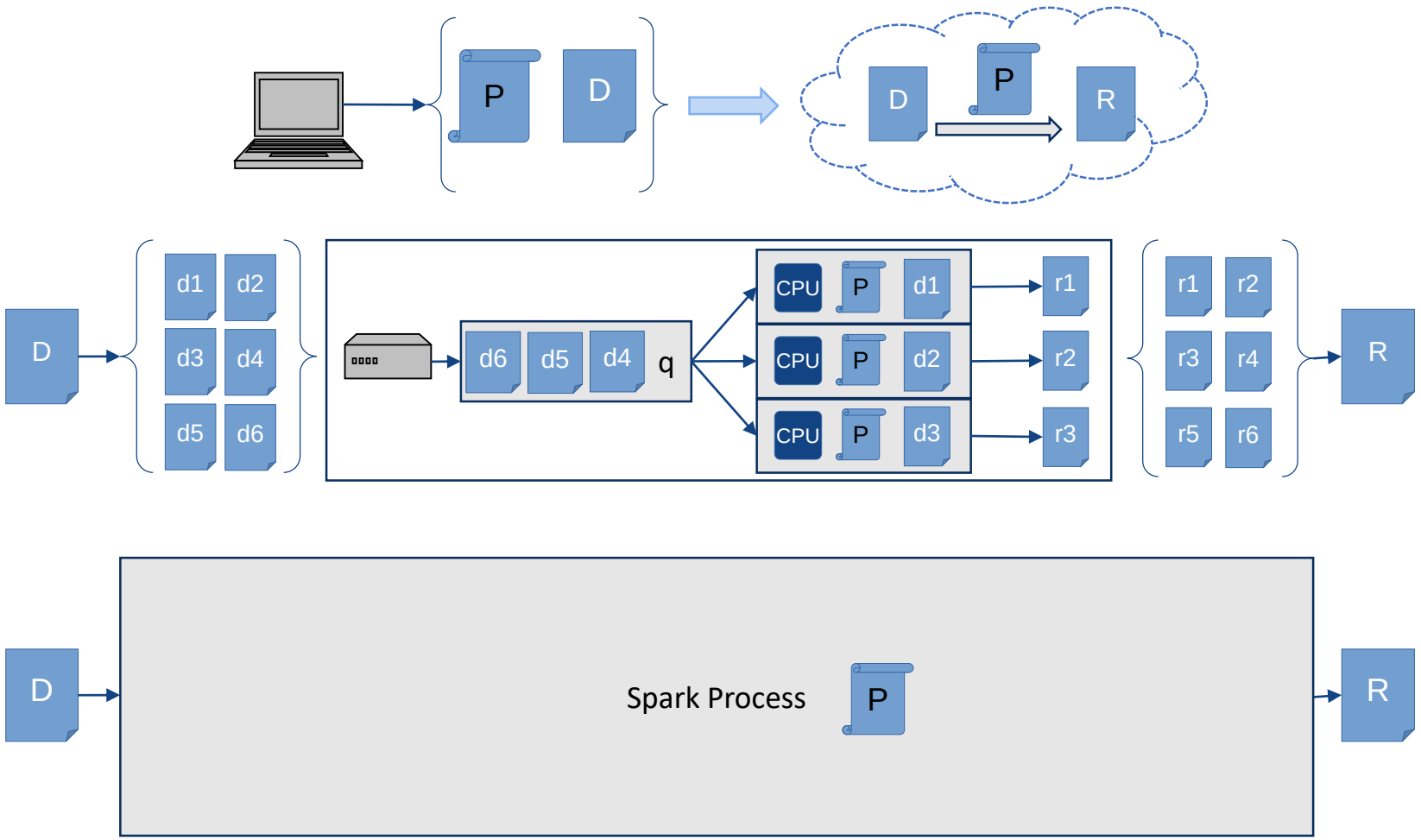
SparkML

- Distributed ML Algoritms



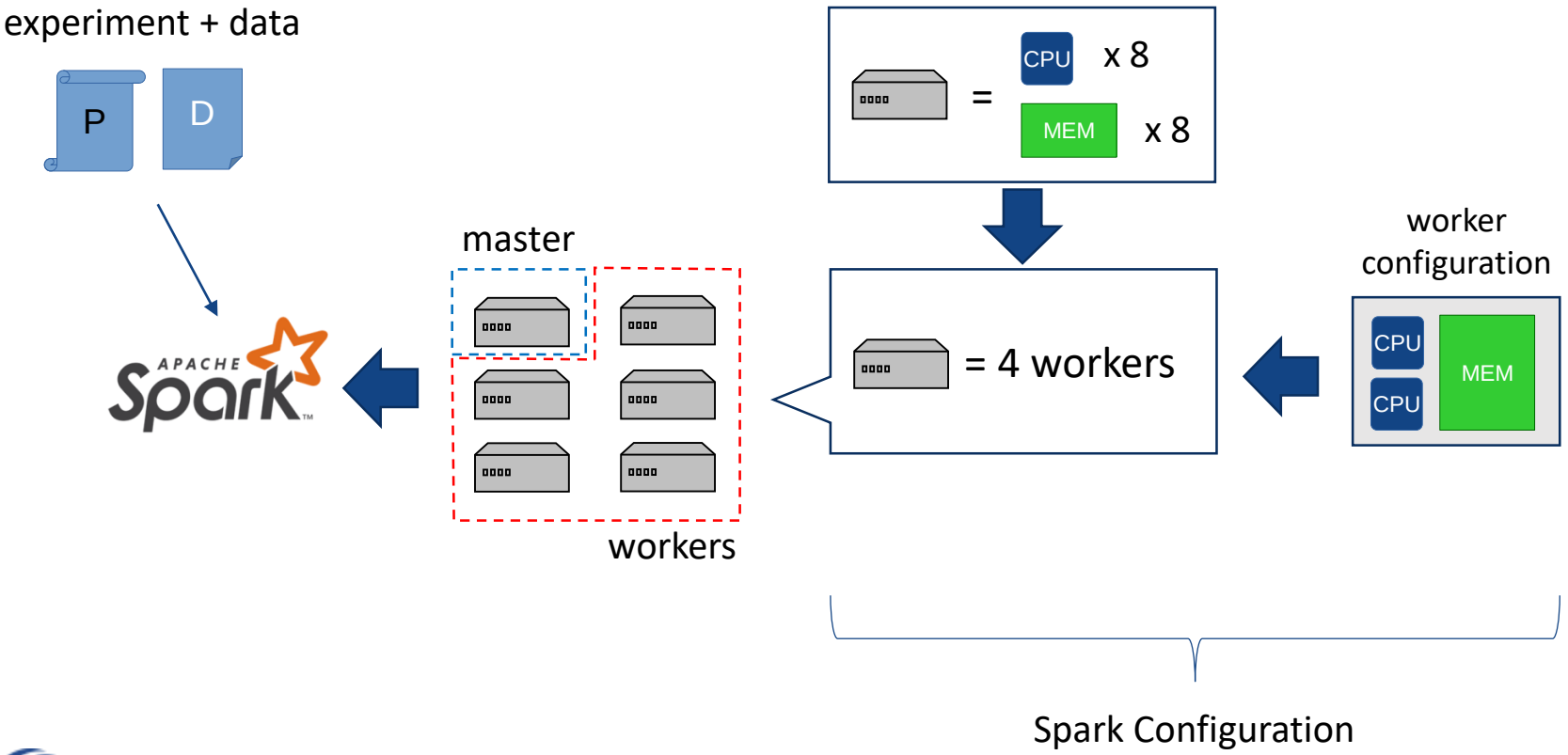
SparkML

- Distributed ML Algoritms



Resources!

- Indicate Spark which resources it can use



SparkML

- Machine Learning standard functions:
  - "builders" for the desired algorithm
    - "new ModellingAlgorithm"
  - "fit" functions
    - Enter data, exit model
  - "evaluate" functions
    - Enter model and data, exit error metrics
  - "predict" functions
    - Enter model + new data, exit predictions

Define model & hyper-parameters

Training the model

Validating the model

Inference & Usage

Summary

- SparkSQL
  - We can slice, select group, and perform standard analytics over our data, in a “relational” way
- SparkML
  - We can train, evaluate and use models with our data
  - ... using distributed computing
  - ... transparent to the user
- Next Hands-On:
  - We’ll do some experiments on Spark with ML