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 (Hands On)

Josep Ll. Berral @BSC

Intel Academic Education Mindshare Initiative for AI

Hands-On: SparkSQL & SparkML

- SparkSQL
  - Load data
  - Select and aggregate data

- Read DataFrames
- Slice DataFrames
  - Select
  - Filter
  - Group By + Aggregation Operation (Count, Min Max, Avg...)
- Save DataFrames to Files

Hands-On: SparkSQL

- Let's run SPARK (again)!

In this case

pyspark 

```
$ $SPARK_HOME/bin/pyspark
Python 3.6.5 (default, Mar 31 2018, 19:45:04) [GCC] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\ version 2.4.1
      /_/
Using Python version 3.6.5 (default, Mar 31 2018 19:45:04)
SparkSession available as 'spark'.
>>>
```

Summary

- Load, aggregate and store data with SparkSQL