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

SparkML (Hands On)

Josep Ll. Berral @BSC

Intel Academic Education Mindshare Initiative for AI

Hands-On: SparkML

- SparkML
  - Training models
  - Evaluate models
  - Use models for inference

Hands-On: SparkML

- 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'.
>>>
```

Last remarks for SparkML

- Transforming “tabular” DataFrames to “libsvm” format
  - We use a “Vector Assembler”

```
from pyspark.ml.feature import VectorAssembler
from pyspark.ml.linalg import Vectors

df = spark.read.csv("/home/vagrant/hus/ss13husa.csv", header = True, mode="DROPMALFORMED", inferSchema = True)
slice1 = df.select("SERIALNO","PUMA","DIVISION").limit(10)

assembler = VectorAssembler(inputCols = ["SERIALNO", "PUMA", "DIVISION"], outputCol = "features")
output = assembler.transform(slice1)

output.select("features").show()
```

Summary

- Basic examples of SparkML
  - Train, evaluate and use machine learning models