Spark Recap

Apache Spark

- **Spark** is a **PROCESSING** framework, not a STORAGE system
- **Spark** is faster than Hadoop
- **Spark** is one of the most used distributed data processing systems in both industry and research

Spark Core

- **SparkContext** is the entry point to Spark Core
- A SparkContext instance *is* a Spark application
- A SparkContext allows the user to the create/read/load an RDD

- **RDD:** Resilient Distributed Dataset, it is a collection of records spread over one or many partitions
- **Resilient:** i.e., fault-tolerant, able to recompute missing or damaged partitions due to node failures
- **Distributed:** with data residing on multiple nodes in a cluster
- **Dataset:** is a collection of primitive values (strings, integers, …) or values of values (tuples, arrays, or other objects)

Spark Core

- **Operations**: transformations, and actions
- **Transformation:** operations that return another RDD
  (map, flatMap, filter)
- **Actions:** operations that trigger computation and return values
  (count, collect)
- **Lazy computation:** the data inside RDD is not available or transformed until an action is executed that triggers the execution

Spark SQL

- **SparkSession** is the entry point to Spark SQL and Spark in general, since from it we are able to create/access the SparkContext
- A SparkSession allows for creating a DataFrame from an RDD, accessing the Spark SQL services, executing SQL queries, access the DataFrameReader interface to load a dataset of the format of your choice
- **DataFrame:** evolution of RDD for tabular data, easier to access a field and to save as output
- DataFrames are distributed through multiple nodes in the same way an RDD is

Spark ML

- Two main operations: **fit** and **transform**
- At the beginning we have a ML algorithm (RandomForestClassifier, Kmeans, …)

- With **fit** we use the training dataset to obtain a ML **model**

- With **transform** we apply the model on the test dataset

Extra Tip

- You can read a file or even an entire folder
- It loads in an RDD or a DataFrame (it depends on what you did) the content of each file in the folder

Contacts

For any problem, send a mail to

daniele.foroni@unitn.it