Practical Session · Tech Foundations · 17/04/2018

Introduction to Platform Technologies: Apache Spark

Duration: 1.5 hour

PART 1

BBVA's Sales Department is interested in analysing the number and type of credit cards per client so that they can offer customized products to each person.

For this project there are two files which contain client information (clientes.txt) and cards data (tarjetas.txt).

In this exercise you will use Spark through Python or Scala API, whatever you prefer.

First of all, build an Spark Standalone cluster using docker containers. When your Spark Application is ready to run, you will deploy it in that cluster.

Application MUSTS

1. Load both files as RDDs in the Spark Context
2. Count the number of credit cards and show two random items to understand the data structure.
3. Create two classes, one for clients and another for cards. The input will be an string with the attributes splitted by commas (Save those classes in a .py file in the same directory as the Spark app)
    a. Clients attributes: DNI, Name, Address
    b. Cards attributes: DNI, Client name, Card type, Card number
4. Map both RDDs with the corresponding class transformation.
5. Join both RDD by DNI in a RDD called *ClientsCards*. **Hint.** To join datasets, Pair RDDs are needed.
6. Show the dataset values as readable. **Format.** "NAME - CARD TYPE - CARD NUMBER"
7. Group the *ClientsCards* by client, so that this structure is returned: **(key=DNI, value= list of client - card objects)**
8. Extract the data in a friendly way for each client (ensure no client is duplicated): **(key = Client name, value = list of cards objects)**

Resources

- PySpark [Official Doc]
- Scala API in Spark [Official Doc]

PART 2

Let’s dive into complex datasets using Spark SQL. The data for this exercise is contained in **bankcard.csv**. Follow the steps below.

1. Load the file bankcard.csv in the Spark SQL session. To do it successfully, first, you have to specify the correct schema.
    - **DNI: Integer**
    - **Name: String**
    - **Address: String**
    - **CardType: String**
    - **CardNumber: Long**
2. Show the DNI and name of clients whose card type is *Visa*
3. Create a dataframe which contains the card types and its count.
4. Using SQL queries, show the DNI and name of clients (NOT DUPLICATED) whose card type is American Express.

Resources

- Spark SQL for Python [Official Doc]
- Spark SQL for Scala [Official Doc]