**HW 5. Spark SQL**

**IS 5315**

**What to submit:**  ***Upload one Word document*** to the Canvas Course Web.

**Grading:**

You should provide at least **4 or more advanced-level queries**.

**What to Write in HW Word Document:**

1. Discuss the three topics:
   1. what is Apache Spark
   2. what is Spark SQL
   3. what are the benefits of using Spark SQL
2. Your task is using "world" database. You can use your previous SQL queries but with the DataFrame way as shown below
3. Build **Five Queries**. Each query should have:
   1. A question in plain English
   2. A query in Spark DataFrame
   3. A screenshot of PySpark Shell results

**What to achieve:**

1. **DigitalOcean Web | Dropet | Resize:**

# Click **'Switch OFF'**

# Choose Resize - 2GB 1vCPU

# Click the 'Resize' button at the bottom of screen

# Click **'Switch ON'**

1. **SERVER:**

docker ps -a

docker start $(docker ps -a -q)

# You may run the above command twice

sudo mkdir -p ~/data

sudo chown brand:brand ~/data -R

# http://jupyter-notebook.readthedocs.io/en/latest/public\_server.html#preparing-a-hashed-password

**In [1]: from** **notebook.auth** **import** passwd

**In [2]:** passwd()

Enter password:

Verify password:

**Out[2]:** 'sha1:15570649d606:4b3908abe041b9c6b97ff797073e5aec609060e1'

# spark notebook container with password: bigdataftw

docker run -d --name notebook -p 8888:8888 -v ~/data:/home/jovyan/work --user root -e GRANT\_SUDO=yes jupyter/all-spark-notebook start-notebook.sh --NotebookApp.password='sha1:15570649d606:4b3908abe041b9c6b97ff797073e5aec609060e1'

~~# spark notebook container without toekn~~

~~docker run -d --name notebook -p 8888:8888 -v ~/data:/home/jovyan/work --user root -e GRANT\_SUDO=yes jupyter/all-spark-notebook start-notebook.sh --NotebookApp.token=''~~

# Open bash window to the sparknotebook container

docker exec -it notebook bash

# Download MySQL connector and unzip it

cd work

wget "http://www.java2s.com/Code/JarDownload/mysql/mysql.jar.zip"

unzip mysql.jar.zip

# Run PySpark with mysql connector

/usr/local/spark/bin/pyspark --jars mysql.jar

**After today's class:**

1. **Reconnecting Spark from your server:**

docker exec -it notebook bash

cd work

/usr/local/spark/bin/pyspark --jars mysql.jar

1. **How to restart a container:**

docker restart notebook

**References:**

R1. <https://github.com/jupyter/docker-stacks/tree/master/all-spark-notebook>

R2. <https://www.analyticsvidhya.com/blog/2016/10/spark-dataframe-and-operations/>

Welcome to

\_\_\_\_ \_\_

/ \_\_/\_\_ \_\_\_ \_\_\_\_\_/ /\_\_

\_\ \/ \_ \/ \_ `/ \_\_/ '\_/

/\_\_ / .\_\_/\\_,\_/\_/ /\_/\\_\ version 2.3.0

/\_/

Using Python version 3.6.5 (default, Apr 6 2018 13:39:56)

SparkSession available as 'spark'.

Country = sqlContext.read.format("jdbc").options( url="jdbc:mysql://**YourIP**:3306/world",driver = "com.mysql.jdbc.Driver",dbtable = "country",user="root", password="bigdataftw").load().persist()

Country.columns

CountryLanguage = sqlContext.read.format("jdbc").options( url="jdbc:mysql://**YourIP**:3306/world",driver = "com.mysql.jdbc.Driver",dbtable = "countrylanguage",user="root", password="bigdataftw").load().persist()

CountryLanguage.columns

City = sqlContext.read.format("jdbc").options( url="jdbc:mysql://**YourIP**:3306/world",driver = "com.mysql.jdbc.Driver",dbtable = "city",user="root", password="bigdataftw").load().persist()

City.columns

City.select(City.Name,City.Population).orderBy(City.Population.desc()).show(10,False)

Country.groupBy("Continent").count().show(10)

Country.select(Country.LifeExpectancy,Country.Region,Country.GovernmentForm).orderBy(Country.LifeExpectancy.desc()).distinct().show(50,False)

# joining city and country

city\_lang = City.join(Country, Country.Code == City.CountryCode, "left\_outer")

city\_lang.select(City.Name,City.Population,Country.Name).orderBy(City.Population.desc()).show(10,False)

# joining country and countrylanguage

country\_lang = Country.join(CountryLanguage, CountryLanguage.CountryCode == Country.Code, "left\_outer")

country\_lang.select(CountryLanguage.Percentage, Country.Name,CountryLanguage.Language).filter(CountryLanguage.Percentage > 50).show(20,False)

country\_lang.select(Country.Name,CountryLanguage.IsOfficial,CountryLanguage.Language).where(CountryLanguage.IsOfficial == 'T').where(CountryLanguage.Language == 'English').show(50,False)

## **Assignment Rubric**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Capstone | Milestone Plus | Milestone | Benchmark |
| 10 | 8 | 6 | 4 |
| Explanation of Issues | Issue/problem to be considered critically is stated clearly and described comprehensively, delivering all relevant information necessary for full unerstanding. | Issue/problem to be considered critically is stated, described, and clarified so that understanding is not seriously impeded by omissions. | Issue/problem to be considered critically is stated but description leaves some terms undefined, ambiguities unexplored, boundaries undetermined, and/or backgrounds unknown. | Issue/problem to be considered critically is stated without clarification or description. |
| External Sources | Accesses information using effective, well-designed search strategies and most appropriate information sources. | Accesses information using variety of search strategies and some relevant information sources. Demonstrates ability to refine search. | Accesses information using simple search strategies, retrieves information from limited and similar sources. | Accesses information randomly, retrieves information that lacks relevance and quality. |
| IT Applicability | Implements the solution in a manner that addresses thoroughly and deeply multiple contextual factors of the problem. | Implements the solution in a manner that addresses multiple contextual factors of the problem in a surface manner. | Implements the solution in a manner that addresses the problem but ignores relevant contextual factors. | Implements the solution in a manner that does not directly addresses the problem statement. |
| Organization | Organizational pattern is clearly and consistently observable and is skillful and makes the content of the presentation cohesive. | Organizational pattern is clearly and consistently observable within the presentation. | Organizational pattern is intermittently observable within the presentation. | Organizational pattern is not observable within the presentation. |