Fundamentos de Big Data

Demo 1
Hello World Spark

Centro Universitário 7 Setembro - Uni7

Especialização em Ciência de Dados com Big Data, BI e Data Analytics

Prof. Manoel Ribeiro

Pré Requisitos para Windows

★ HADOOP 3.0.1
   ○ http://hadoop.apache.org/releases.html
   ○ binary 3.0.1
   ○ http://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-3.0.1/hadoop-3.0.1.tar.gz
   ○ Abrir hadoop-3.0.1.tar.gz
   ○ Abrir pasta hadoop-3.0.1
   ○ Extrair conteúdo desta pasta para **c:\bigdata\hadoop**
★ WINUTILS para Haddop 3.0.0 ou superior
   ○ https://github.com/steveloughran/winutils/tree/master/hadoop-3.0.0
   ○ download **winutils.exe** para c:\bigdata\hadoop\bin

Pré Requisitos para Windows

★ SPARK 2.3.0
   ○ https://spark.apache.org/downloads.html
   ○ Pre-built for Apache Hadoop 2.7 and later
   ○ Direct Download
   ○ Abrir spark-2.3.0-bin-hadoop2.7.tgz
   ○ Abrir pasta spark-2.3.0-bin-hadoop2.7
   ○ Extrair conteúdo desta pasta para **c:\bigdata\spark**
★ JAVA JRE
   ○ java -version
      ■ Deve ser 7 ou 8 (9 ainda não funciona)
      ■ Instalar customizado para **c:\bigdata\java**

Setup

```
>set HADOOP_HOME=C:\bigdata\hadoop
>set SPARK_HOME=c:\bigdata\spark
>cd %HADOOP_HOME%\bin
>hdfs namenode -format
c:\bigdata\hadoop\bin>winutils ls c:\tmp\
drwxrwxrwx 1 LSBD\manoel.ribeiro LSBD\Domain Users 0 Oct
3 2017 c:\tmp\hive

c:\bigdata\hadoop\bin>winutils chmod 777 c:\tmp\
```

Iniciando

```
>cd %SPARK_HOME%\bin
c:\bigdata\spark\bin>pyspark
Welcome to

      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.2.0
      /_/

Using Python version 3.4.2 (v3.4.2:ab2c023a9432, Oct  6 2014
22:15:05)
SparkSession available as 'spark'.
>>>
```

Hello World - Spark Hive!

```
>>> data=[('Iphone8', 5000),('Pixel2', 4000),
('GalaxyS8',3000), ('MotoZ', 2500)]

>>> df=spark.createDataFrame(data,
('smartphone','valor'))
>>> df.printSchema()
root
 |-- smartphone: string (nullable = true)
 |-- valor: long (nullable = true)
```

Fim