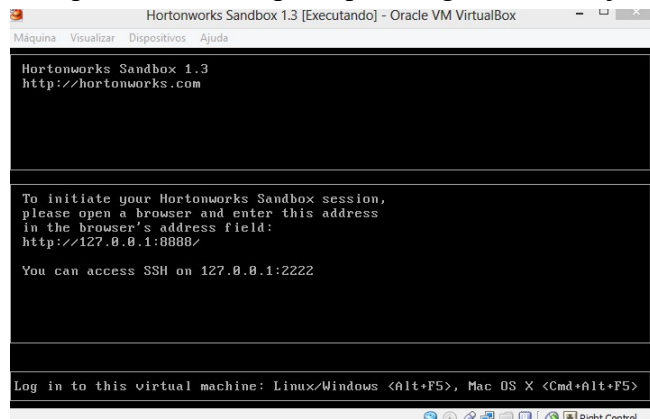


Professor: Cláudio Lúcio
Atividade Verificando processos e RDD's

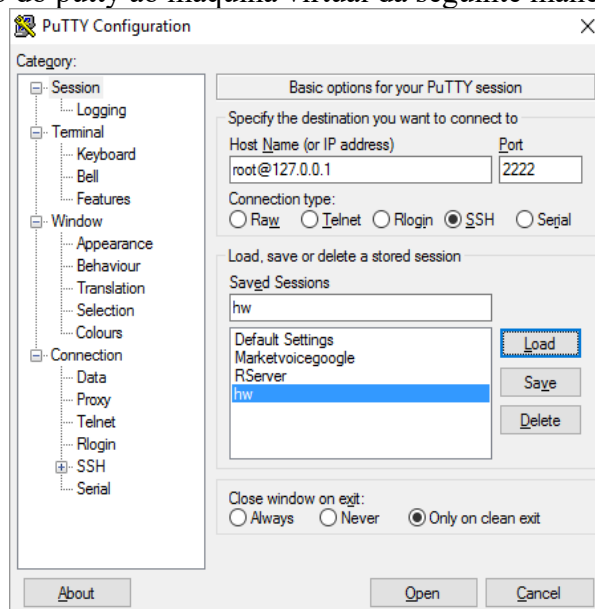
- 1) Para acesso ao Shell do Spark vamos usar a versão do Spark que vem na máquina virtual do HortonWorks:

Esta é a versão 2.3.2 do produto e pode ser obtida em:
<http://hortonworks.com/products/ Hortonworks-Sandbox/>

- 2) Inicialize a máquina virtual. Espere que a seguinte tela seja exibida:



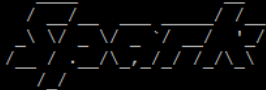
- 3) A recomendação é usar um cliente para acesso SSH ao servidor do spark. Baixe a ferramenta putty.exe;
- 4) Configure o acesso do putty ao máquina virtual da seguinte maneira:



- 5) Clique em “Open” e então digite a senha do Root: hadoop

```
root@sandbox:~  
Using username "root".  
root@127.0.0.1's password:  
Last login: Fri Dec 23 16:56:09 2016 from 10.0.2.2  
[root@sandbox ~]#
```

- 6) Primeiramente vamos interagir com a interface python para o Spark, para tal vamos acessar o bin/pyspark:
- 7) Digite os seguintes comandos:
 - `cd /usr/hdp/2.3.2.0-2950/spark/bin`
 - `pyspark`

```
root@sandbox:/usr/hdp/2.3.2.0-2950/spark/bin  
16/12/26 17:25:39 WARN SparkConf: The configuration key 'spark.yarn.applicationMaster.waitTries' has been deprecated as of Spark 1.3 and may be removed in the future. Please use the new key 'spark.yarn.am.waitTime' instead.  
16/12/26 17:25:39 WARN SparkConf: The configuration key 'spark.yarn.applicationMaster.waitTries' has been deprecated as of Spark 1.3 and may be removed in the future. Please use the new key 'spark.yarn.am.waitTime' instead.  
16/12/26 17:25:39 INFO Executor: Starting executor ID driver on host localhost  
16/12/26 17:25:40 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 53908.  
16/12/26 17:25:40 INFO NettyBlockTransferService: Server created on 53908  
16/12/26 17:25:40 INFO BlockManagerMaster: Trying to register BlockManager  
16/12/26 17:25:40 INFO BlockManagerMasterEndpoint: Registering block manager localhost:53908 with 265.4 MB RAM, BlockManagerId(driver, localhost, 53908)  
16/12/26 17:25:40 INFO BlockManagerMaster: Registered BlockManager  
Welcome to  
 version 1.4.1  
Using Python version 2.6.6 (r266:84292, Jul 23 2015 15:22:56)  
SparkContext available as sc, HiveContext available as sqlContext.  
>>>
```

- 8) Veja que estamos com a versão 1.4.1 do Spark;
- 9) Vamos preparar algumas estruturas e transformá-las em DataFrame:

```
peessoas = []  
peessoas.append({'id':1,'nome':'Bob', 'idade':45,'gen':'M'})  
peessoas.append({'id':2,'nome':'Gloria', 'idade':43,'gen':'F'})  
peessoas.append({'id':4,'nome':'Albert', 'idade':28,'gen':'M'})  
peessoas.append({'id':5,'nome':'Laura', 'idade':33,'gen':'F'})  
peessoas.append({'id':8,'nome':'Simone', 'idade':18,'gen':'T'})  
peessoas.append({'id':12,'nome':'Marta', 'idade':45,'gen':'F'})  
peessoas.append({'id':45,'nome':'Jairo', 'idade':82,'gen':'M'})  
peessoas.append({'id':13,'nome':'Teste', 'idade':38,'gen':'T'})  
peessoasRdd=sc.parallelize(peessoas)  
  
import json  
peessoasValFile = sc.textFile("/user/hue/peessoasval.json")  
peessoasValRdd = pessoasValFile.flatMap(lambda arg : (json.loads(arg)))  
peessoasValRdd.collect()  
  
peessoasDF = pessoasRdd.toDF()  
peessoasValDF = pessoasValRdd.toDF()
```

- 10) Qual a estrutura dos dataframes?

```
peessoasDF.printSchema()  
peessoasValDF.printSchema()
```

11) Verifique o número de linhas de cada dataframe?

```
peessoasDF.count()  
peessoasValDF.count()
```

12) Estatísticas descritivas sobre a tabela de valores das pessoas?

```
peessoasValDF.select('val').describe().show()
```

13) Quantidade de Id's distintos na tabela de valores?

```
peessoasValDF.select('id').distinct().count()
```

14) Geração da tabela cruzada de idade por gênero:

```
peessoasDF.crosstab('idade', 'gen').show()
```

15) Filtros diversos (gênero, idade e valores):

```
peessoasDF.filter(peessoasDF.gen == 'M').show()  
peessoasDF.filter(peessoasDF.idade > 30).show()  
peessoasValDF.filter((peessoasValDF.val > 10) & (peessoasValDF.val < 10000)).show()
```

16) Valor total somado para cada um dos anos (DataFrame Val):

```
peessoasValDFYear = pessoasValDF.withColumn('Ano', pessoasValDF.dat.substr(7,  
11)).groupby('Ano').agg({'val': 'sum'})  
peessoasValDFYear.show()
```