A VM da Cloudera utilizada nestes exercícios de Spark está disponível em:

<http://training.cloudera.com/cloudera/VMs/Cloudera-Training-Spark-Developer-VM-5.0a-vmware.zip>

**EXERCÍCIO 5 - SPARK**

Neste exercício será realizado uma análise de logs de um servidor web, assim como a análise dos dados das contas dos usuários, utilizando pares de chave-valor de RDDs.

1. Abra um terminal no Linux ( - na barra superior) e execute o script:

~/scripts/sparkdev/training\_setup\_sparkdev.sh

1. Execute o seguinte comando para visualizar o conteúdo do log do web server:

head /home/training/training\_materials/sparkdev/data/weblogs/2014-03-15.log



1. Execute o seguinte comando para visualizar o conteúdo do arquivo accounts:

head /home/training/training\_materials/sparkdev/data/accounts.csv

****

1. Abra o pyspark:

pyspark



1. Crie um RDD baseado nos arquivos de log:

logsweb=sc.textFile("file:/home/training/training\_materials/sparkdev/data/weblogs/\*")

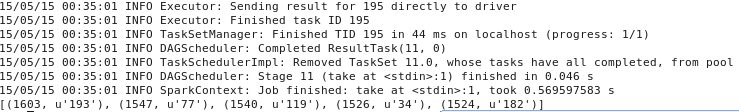
****

1. Realizar o map em cada requisição (linha) para criar um par de chave (userid) e valor (1), depois some a quantidade de requisições para cada um:

requisicaousuario = logsweb.map(lambda linha: linha.split()).map(lambda palavras: (palavras[2],1)).reduceByKey(lambda conta1,conta2: conta1 + conta2)

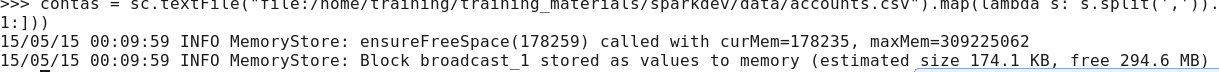
1. Exiba os registros dos 5 usuários com as maiores contagens de requisições:

requisicaousuario.map(lambda par:(par[1],par[0])).sortByKey(False).take(5)

****

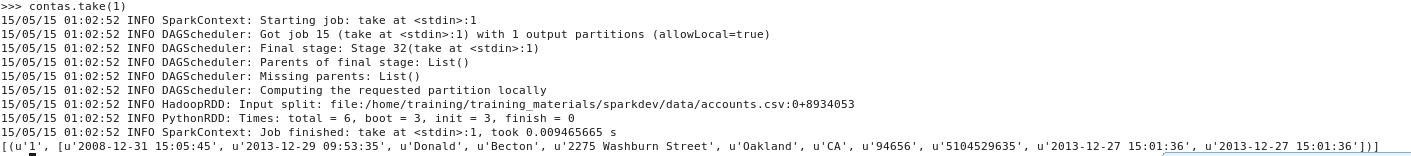
1. Realize o map da planilha com as contas dos usuários, considerando que o primeiro campo é a chave (userid) e o resto da linha é o valor (detalhes sobre o cliente como data de criação, primeiro e último nome):

contas = sc.textFile("file:/home/training/training\_materials/sparkdev/data/accounts.csv").map(lambda s: s.split(',')).map(lambda conta: (conta[0],conta[1:]))



1. Pegue o primeiro registro para entender como ficou o par de chave-valor:

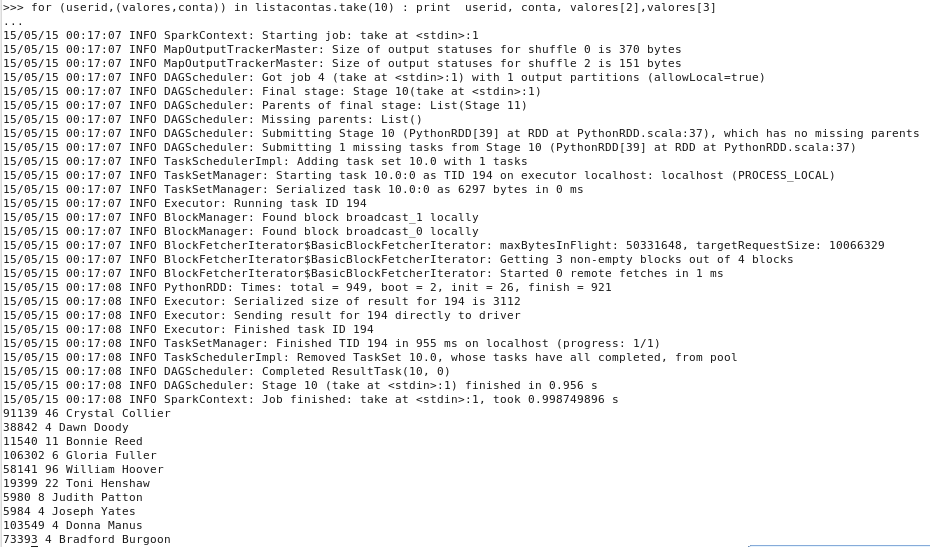
contas.take(1)



1. Realize o join das contas com as quantidades de requisições por userid, juntando ambos em uma lista de valores:

listacontas = contas.join(requisicaousuario)

for (userid,(valores,conta)) in listacontas.take(10) : print userid, conta, valores[2],valores[3]



**EXERCÍCIO 4 (EXTRA)**

1. Verifique se de fato o userid “91139” possui 46 requisições no dataset requisicaousuario.
2. Repita a mesma validação para o userid “38842” com apenas 4 requisições.
3. Realize a contagem no dataset logsweb para ambos userid “91139” e “38842”, validando se a quantidade é a mesma.