Formação Engenheiro de Dados

Estudo de Caso

PROF. FERNANDO AMARAL
www.datascientist.com.br

Estudo de Caso

- 1º Parte, operações básicas, ações e transformações

```
pyspark
numeros = sc.parallelize([1,2,3,4,5,6,7,8,9,10])

numeros.count()

filtro = numeros.filter(lambda filtro: filtro > 2)
```

Estudo de Caso

- 2º Parte, operações chave-valor

```
compras = sc.parallelize([(1,200),(2,300),(3,120),(4,250),(5,78)])

soma = compras.mapValues(lambda soma: soma + 1)
agrupa = compras.groupByKey().mapValues(list)
```

Estudo de Caso

- 3º Caso: Word Count, usando mesmo arquivo do caso Hadoop

- pesquisa = sc.textFile("file:///home/cloudera/Download/pesquisa.txt")

Estudo de Caso

- 4º Caso: Analisar dados do Hive (DW)
- Copiar arquivo de configuração

```
ls /usr/lib/hive/conf/hive-site.xml
cat /usr/lib/hive/conf/hive-site.xml
sudo cp /usr/lib/hive/conf/hive-site.xml /usr/lib/spark/conf/
```

Estudo de Caso

- Criar contexto HiveContext

```
from pyspark.sql import HiveContext
contexto= HiveContext(sc)
banco = contexto.table("ed.des_vendas")
banco.show()

banco.registerTempTable("des_vendas")
contexto.sql("select * from des_vendas ").show()
contexto.sql("select sum(valortotal) from des_vendas").show()
```

Data Frame

```
vendas  = contexto.sql("select * from des_vendas")
vendas.show()
vendas.show(100)
vendas.printSchema()
vendas.select('estado','status').show()
vendas.select('estado','status').distinct().show(30)
vendas.filter(vendas.estado=='RS').show()
vendas.filter(vendas.estado=='RS').count()
```