Formação Engenheiro de Dados

Estudo de Caso

PROF. FERNANDO AMARAL
www.datascientist.com.br

Contagem de Palavras

- Criar Diretório
- Criar uma aplicação PySpark de streaming para monitorar este diretório
- Rodar a aplicação
- Colar arquivo pesquisabike.txt na pasta
- No console onde a aplicação rodou observar a contagem das palavras

Aplicação

```
from pyspark import SparkContext
from pyspark.streaming import StreamingContext
sc = SparkContext("local[2]", "Contagem")
ssc = StreamingContext(sc, 10)
pesquisa = ssc.textFileStream("file:///home/cloudera/spark/")
contagem = pesquisa.flatMap(lambda palavra: palavra.split(" "))
contagem = contagem.map(lambda pal: (pal, 1))
contagem = contagem.reduceByKey(lambda a, b: a + b)
contagem.pprint()
ssc.start()
ssc.awaitTermination()
```