Guia de Introdução ao projeto HP

1. Análise De Dados

Temas Importantes De Estudos:

i. Análise de Dados
ii. Aprendizagem de Máquina
iii. Mineração de dados
iv. Processamento de Linguagem Natural (?)
v. Reconhecimento de Padrões
vi. Sistemas de Recomendação
vii. Web Crawling

Curso Rápido de Análise (*Opcional*):

https://lagunita.stanford.edu/courses/HumanitiesSciences/StatLearning/Winter2016/abouthttps://br.udacity.com/course/intro-to-statistics--st101/

Especificamente sobre Spark:

Curso:

iniciante: https://bigdatauniversity.com/courses/what-is-spark/

intermediário: https://bigdatauniversity.com/courses/spark-rdd/

Link Detalhado:

https://www.safaribooksonline.com/library/view/learning-spark/9781449359034/ch04.html

Livro para referência:

https://drive.google.com/file/d/0ByMErZ8giK2vRjRLZzgwMUp6a0E/view?usp=sharing

Leituras importantes:

http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#191990ed69fd

http://www.harlan.harris.name/2011/09/data-science-moores-law-and-moneyball/

http://www.holehouse.org/mlclass/

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

http://www.dataists.com/2010/09/a-taxonomy-of-data-science/

Problemas Básicos:

https://github.com/zegildo/codereviewer/blob/master/1_Find_Optimal_Chopsticks_Length/Data_Analyst_ND_Project0.ipynb

https://github.com/zegildo/codereviewer/blob/master/2_Test_a_Perceptual_Phenomenon/2_Test_a_Perceptual_Phenomenon.ipynb

https://github.com/zegildo/codereviewer/blob/master/getRecession/ZeGildoGetRecessions.ipynb

2. Especificamente sobre o Projeto

Downloads:

HPSA (opt-in): conjunto de serviços para dar suporte a PCs e impressoras de modo a ajudar a mantê-los evitando ou resolvendo problemas por meio das informações do funcionamento do dispositvo. http://www8.hp.com/us/en/campaigns/hpsupportassistant/hpsupport.html

WMI Explorer: Interface para gerenciamento de dispositivos.

https://wmie.codeplex.com/

Open Hardware Monitor GUI: Monitora os sensores de temperatura, a velocidade do ventilador, voltagens e velocidade do clock.

http://openhardwaremonitor.org/

Process Explorer: Substitui de forma mais eficiente o *Task Manager* nativo do Windows. Apresenta processos ativos, seus proprietários e contas e informa quem manipula e utiliza arquivos DLL.

https://technet.microsoft.com/en-us/sysinternals/processexplorer.aspx

HP Performance Advisor: Maximiza a performance da estação de trabalho onde foi instalado. O objetivo é garantir que o dispositivo computacional sempre esteja operando em seu potencial ótimo. Busca-se ajustes a depender da atividade sendo desenvolvida pelo usuário. *(Apenas em máquinas HP ?)*

http://www8.hp.com/us/en/workstations/performance-advisor.html

Disk: Monitora e analisa os dispositivos de armazenamento HDD e SSD. O objetivo é encontrar, testar, diagnosticar e reparar problemas nos *drivers* dos discos rígidos

e SSDs reportando-os *(a quem?)* de modo a evitar degradações na saúde dos dispositivos e eventuais falhas.

Gsmartcontrol (*linux*):

http://gsmartcontrol.sourceforge.net/home/

Hard Disk Sentinel (*win*):

http://www.hdsentinel.com/download.php

Inspirações ao time de Desenvolvimento:

https://www.tableau.com/

http://www.qlik.com/us/

https://www.ibm.com/analytics/watson-analytics/us-en/

http://www8.hp.com/us/en/solutions/touchpoint-manager/details.html

Soluções de Problemas de Instalação do Spark 2.1.0:

Ao instalar o Spark na versão mais atual, é provável que existam alguns problemas de instalação. Certifique-se de que:

1. O Anaconda está instalado: https://www.continuum.io/downloads
    a. Digite '*python*' no terminal para verificar.
2. Faça o download do arquivo winutils.exe:

    https://github.com/steveloughran/winutils/tree/master/hadoop-2.7.1/bin/

3. Dentro da pasta do Spark crie as pastas *winutils/bin/* e insira o arquivo winutils.exe que você acabou de realizar o download.

4. Crie a sua variável de ambiente HADOOP_HOME e direcione para a pasta que contém o arquivo winutils.exe. Neste exemplo a pasta encontra-se em: *C:/User/Embedded/Desktop/HP-softwares/Spark/winutils/*. A Figura 2 apresenta um exemplo de como configurar a sua variável de ambiente.

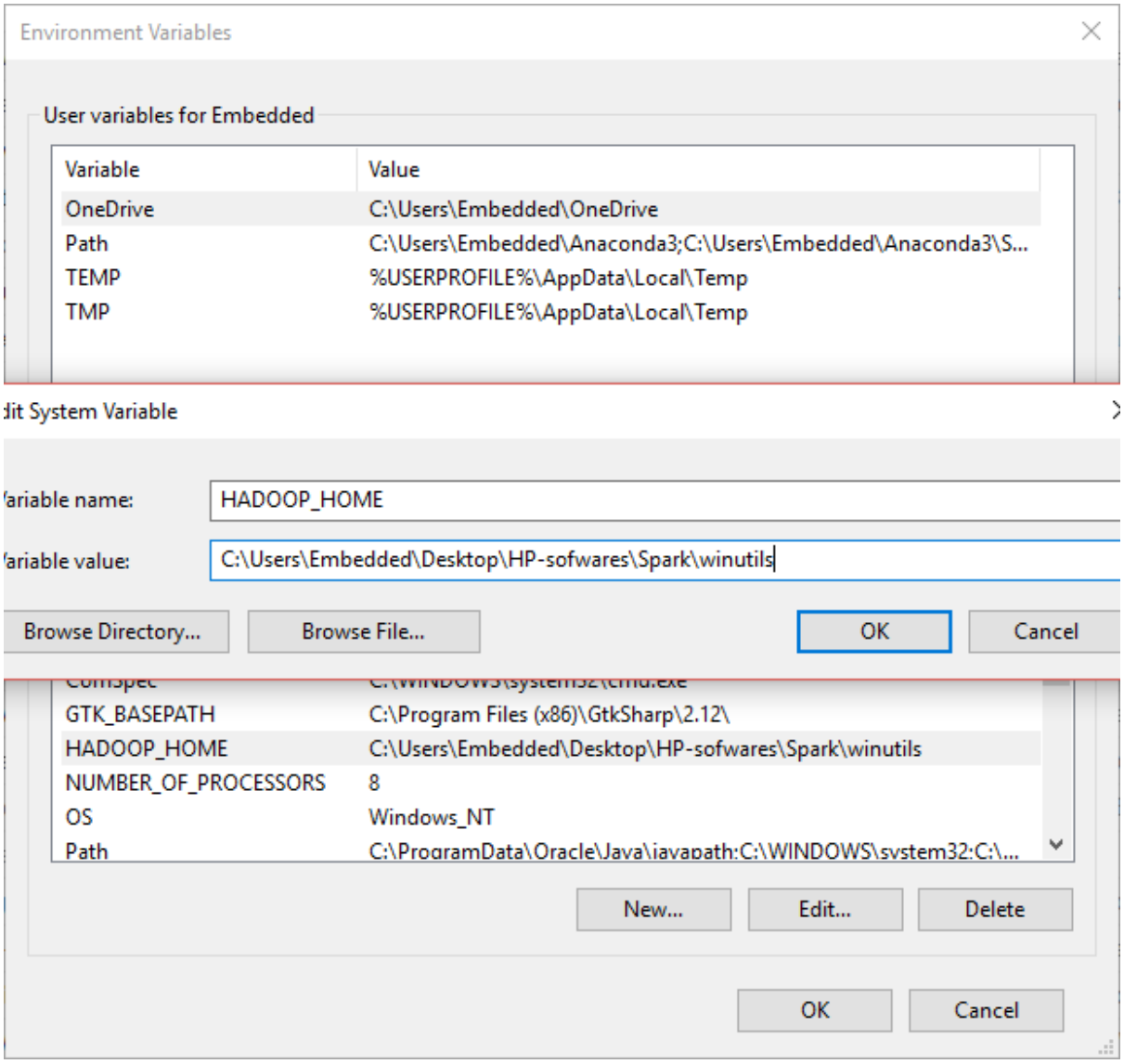

Figura 1. Exemplo de criação da variável de ambiente HADOOP_HOME

5. Execute o Powershell como administrador e execute o seguinte comando:

.\winutils\bin\winutils.exe chmod 777 C:\tmp\hive\

6. Execute o ./bin/pyspark e verifique no terminal se o nome SPARK é impresso.

7. Realize o download dos arquivos **teste.py** e **texto.txt** que encontram-se na pasta **Spark2.1.0-test** presente nos arquivos do grupo **HPGroup.**

8. Altere o interior do arquivo **teste.py** passando a direção correta do arquivo **texto.txt**.

9. Execute o comando a seguir atento à direção do arquivo **teste.py**:

.\spark-submit C:\...\teste.py

10. A parte final do arquivo deve assemelhar-se à Figura 2. Perceba que é impresso **"Lines with a: 9", lines with b: 8"** e o resultado é correto dado que o arquivo **texto.txt** possui exatamente 17 linhas.

```
17/01/19 14:39:00 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated size 6.7 KB, free 366.0 MB)
17/01/19 14:39:00 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 4.3 KB, free 366.0 MB)
17/01/19 14:39:00 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on 10.100.100.159:51325 (size: 4.3 KB, free: 366.3 MB)
17/01/19 14:39:00 INFO SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:996
17/01/19 14:39:00 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (PythonRDD[3] at count at C:/Users/Embedded/Desktop/teste.py:7)
17/01/19 14:39:00 INFO TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
17/01/19 14:39:00 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, executor driver, partition 0, PROCESS_LOCAL, 6104 bytes)
17/01/19 14:39:00 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
17/01/19 14:39:00 INFO BlockManager: Found block rdd_1_0 locally
17/01/19 14:39:01 INFO PythonRunner: Times: total = 398, boot = 396, init = 2, finish = 0
17/01/19 14:39:01 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 1641 bytes result sent to driver
17/01/19 14:39:01 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 424 ms on localhost (executor driver) (1/1)
17/01/19 14:39:01 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
17/01/19 14:39:01 INFO DAGScheduler: ResultStage 1 (count at C:/Users/Embedded/Desktop/teste.py:7) finished in 0,425 s
17/01/19 14:39:01 INFO DAGScheduler: Job 1 finished: count at C:/Users/Embedded/Desktop/teste.py:7, took 0,457664 s
Lines with a: 9, lines with b: 8
17/01/19 14:39:02 INFO SparkContext: Invoking stop() from shutdown hook
17/01/19 14:39:02 INFO SparkUI: Stopped Spark web UI at http://10.100.100.159:4040
17/01/19 14:39:02 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
17/01/19 14:39:02 INFO MemoryStore: MemoryStore cleared
17/01/19 14:39:02 INFO BlockManager: BlockManager stopped
17/01/19 14:39:02 INFO BlockManagerMaster: BlockManagerMaster stopped
17/01/19 14:39:02 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
17/01/19 14:39:02 INFO SparkContext: Successfully stopped SparkContext
17/01/19 14:39:02 INFO ShutdownHookManager: Shutdown hook called
17/01/19 14:39:02 INFO ShutdownHookManager: Deleting directory C:\Users\Embedded\AppData\Local\Temp\spark-ba4bf810-8eaf-4727-b171-6602585db514\pys
park-ee43b4e4-ffcc-495c-a753-7817f67998dc
17/01/19 14:39:02 INFO ShutdownHookManager: Deleting directory C:\Users\Embedded\AppData\Local\Temp\spark-ba4bf810-8eaf-4727-b171-6602585db514
```

Figura 2. Exemplo de saída da execução do Spark2.1.0

Acessar as informações da HP

1. No Windows Explorer selecione o check-box **hidden items**
2. Os arquivos necessários ao processamento encontram-se no caminho:

C:/ProgramData/Hewlett-Packard/HP Active Health

A Figura 3 apresenta um exemplo de conteúdos da pasta HP Active Health.

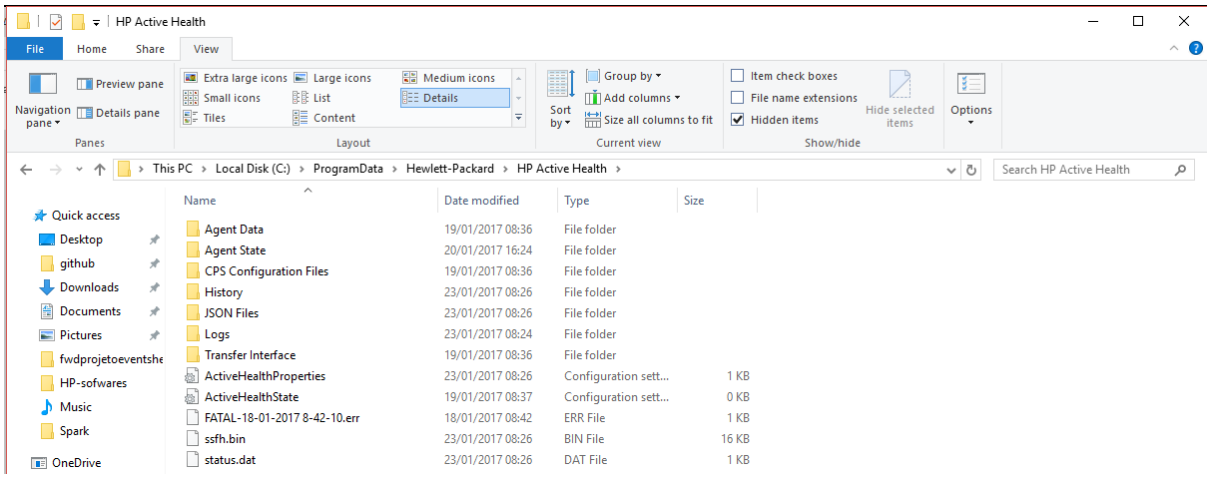

Figura 3. Exemplo de configuração e conteúdo da pasta HP Active Health

Sobre o Databricks

- *Introdução:*

  https://docs.databricks.com/user-guide/getting-started.html#welcome-to-databricks

- *Scientists:*

  https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/346304/2168141618055194/484361/latest.html

- *Videos - Summit 2016:*

https://www.youtube.com/watch?v=OheiUI_uXwo&list=PL-x35fyliRwhDv3g1dae8v2F6-_bzBfGK

- *Códigos:*

https://virtustex.sharepoint.com/sites/hpgroup/Documentos%20Compartilhados/Spark-DataScience-TD1

- *Artigos:*

  https://databricks.com/blog/category/engineering

Sobre o estilo de programação

https://google.github.io/styleguide/pyguide.html