[Descrição: Estácio](http://portal.estacio.br/)

**UNIVERSIDADE ESTÁCIO DE SÁ**

**MBA EM CIÊNCIA DE DADOS E BIG DATA ANALYTICS**

**Projeto HDFS**

**Aluno: Daniel Mendes Guedes**

**Trabalho da disciplina**

**Prática e Laboratório I**

**Tutor: Prof. Denis Gonçalves Cople**

**Brasília/DF**

**2020**

**PROJETO HDFS**

**Referência:**

<https://www.vmware.com/go/getworkstation-win> - Donwload VMware Player Pro 15

<https://www.cloudera.com/downloads/hortonworks-sandbox.html> - Donwload Hortonworks Sandbox – HDP\_3.0.1.OVA

[https://www.bitvise.com/ssh-client-download - Donwload Bitvise 8.43](https://www.bitvise.com/ssh-client-download%20-%20Donwload%20Bitvise%208.43)

<https://www.eclipse.org/downloads/> - Donwload – Eclipse IDE 2020 06

<https://download.virtualbox.org/virtualbox/6.1.12/VirtualBox-6.1.12-139181-Win.exe?xd_co_f=b7281a6d-b72b-40c1-9021-ca2f0b5ca2bf> – Donwload VirtualBox

<https://downloads.cloudera.com/demo_vm/virtualbox/cloudera-quickstart-vm-5.13.0-0-virtualbox.zip> - Donwload Cloudera quickstart

<https://www.youtube.com/watch?v=FI5hWp6Hybg&list=PLeFetwYAi-F_l-NP-TUE2MqKeu_haMP79> – Curso Hadoop – Código Fluente

<https://www.youtube.com/watch?v=BYi0uuNZmno&list=PLG1t8jaLbxA_DG_cmlBYgkGW-TZw5DP3B> – Big Data Hadoop Espanôl

<https://sourceforge.net/projects/virtualappliances/files/Linux/CentOS/> - Donwload CentOS-7.0-amd64-minimal.ova

<https://www.udemy.com/course/fundamentos-de-big-data-com-hadoop/> - Curso Big Data com Hadoop: direto ao ponto e foco na prática.

**Prefácio**

Inicialmente ressalto, o caráter precário do desenvolvimento do Trabalho. Devido o fato de ter feito a matricula recente (29/06), não ter tido tempo hábil para contato com todas as 6 disciplinas do Semestre e nem tempo suficiente para levantar todas as dúvidas e resolver com o professor/tutor, no período dos Fóruns.

**Dificuldades encontradas**

Iniciei o estudo da disciplina Prática e Laboratório em 19/08/2020. Após ter passado aproximadamente 15 dias tentando desenvolver o Projeto/Trabalho da Disciplina Anterior: “Tecnologias Avançadas”. Após estudo da apostila e dos slides da disciplina verifiquei que os aplicativos da Disciplina Tecnologias Avançadas, já deveriam estar instaladas para execução da atividades propostas no material da disciplina e consequentemente para realização do projeto/trabalho da Disciplina Prática e laboratório, os sistemas sugeridos e apresentados, em especial o HDFS e Bitvise não funcionaram e apresentaram erro.

Por isso esse trabalho traz algumas tentativas e erros de outras experiências realizadas com base em pesquisas de conteúdo na internet. Ressaltando ainda a proximidade do fim do prazo de entrega dos trabalhos e a necessidade de finalizar dois Trabalhos similares e complementares, cujos softwares e ferramentas indicadas no material didático da disciplina não funcionaram adequadamente.

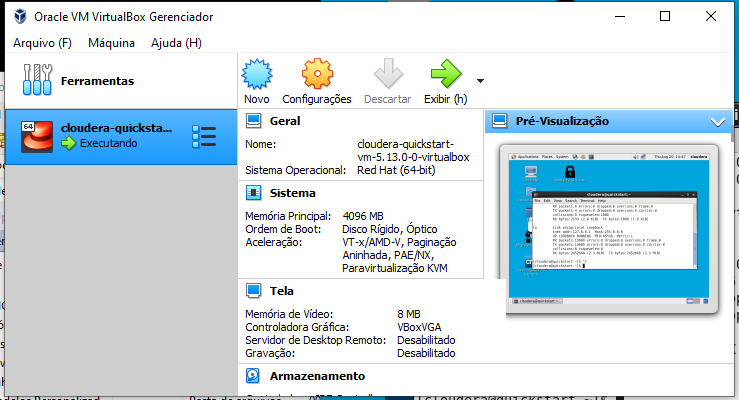
As pesquisas buscaram como palavras chaves Tutorias sobre Hadoop, Map Reduce, Pig, Yarn, Big Data em Geral. Após localizar alguns tutoriais interessantes iniciei a execução dos passos indicados e utilização dos softwares dos vídeos.

Abaixo segue um resumo dos trabalhos desenvolvidos no âmbito da disciplina “Prática e Laboratório” para assimilar seus conteúdos. Novamente destaco que devido o pouco tempo disponível esse foi o trabalho possível a ser entregue.

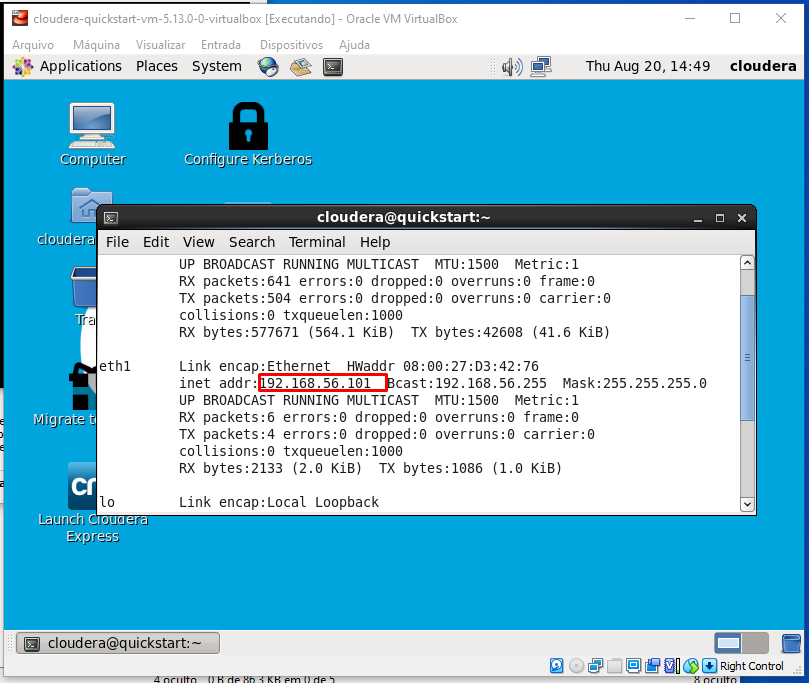
**Exercício 1 – Wordcount MapReduce**

Para realização deste trabalho de MapReduce foi utilizado a Oracle VM Virtual Box, a máquina Virtal Cloudera-quickstart, os programas PuTTY para acesso via SSH e o programa WinSCP para interface gráfica entre os sistemas operacionais.

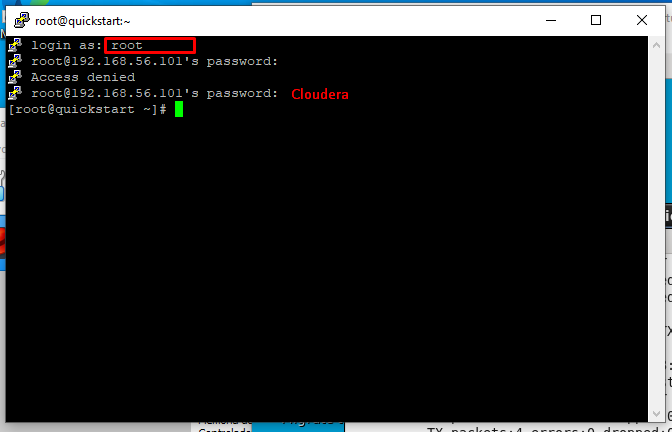
Conforme demonstrado abaixo todas as instalações e configurações indicadas foram realizadas com sucesso e os softwares estão funcionando.



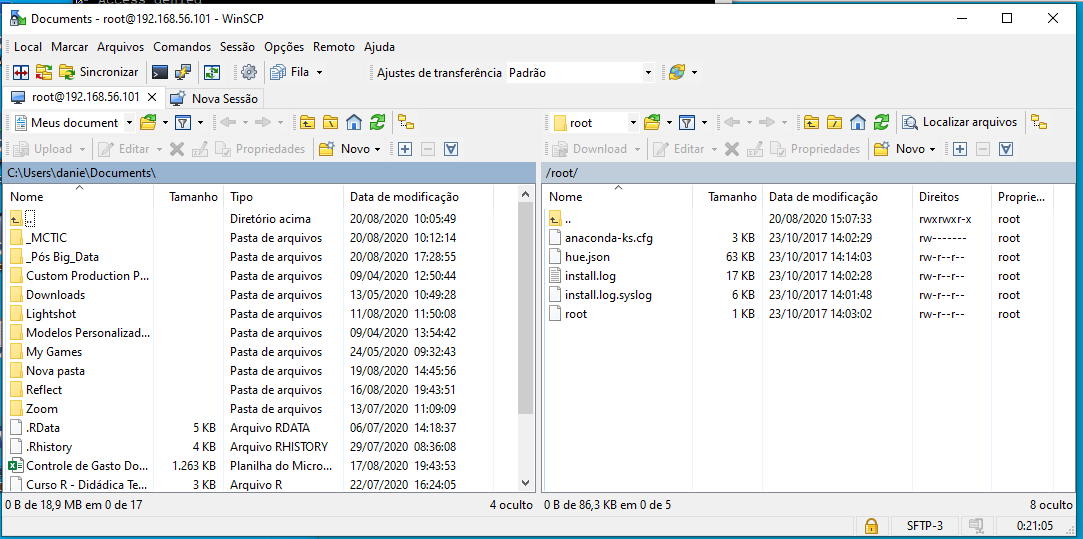
**Oracle VM VirtualBox** – Funcionando



**Cloudera -quickstart-vm 5.13.0** – em Execução



**Software PuTTY –** Acesso SSH OK. Usuário root / Senha: cloudera



**Software WinSCP** – Configurado e acessando os dois sistemas.

Após as instalações com êxito é realizado o desenvolvimento de um Projeto PYSPARK, para a contagem de palavras de um texto. O Spark é uma estrutura de processamento Big Data, Opensource.. PYSPARK é a ligação Python para a Plataforma Spark e a API Spark.

Quando executamos qualquer aplicativo **SPARK**, um programa **Driver** que tem a função **MAIN,** é iniciado e seu **SPARKCONTEXT** é iniciado também.

**Contagem de Palavras com PYSPARK, Passo a passo:**

1. Entrar no PYSPARK abrindo o terminal e utilizar o comando:

**pyspark**

1. Importar as bibliotecas SPARCONTEXT e o SPARKCONF para o PYSPARK com o comando: **from pyspark import SparkContext, SparkConf**
2. Criar Objeto de configuração e definir o nome do app:

**conf = SparkConf().setAppName(“WordCount”)**

**sc = SparkContext.getOrCreate()**

1. Carregar os textos, pelo terminal, para o HDFS, e depois inserir no contentRDD (shakespeare.txt): **Hadoop fs -put Shakespeare.txt/user/estacio/textos/**

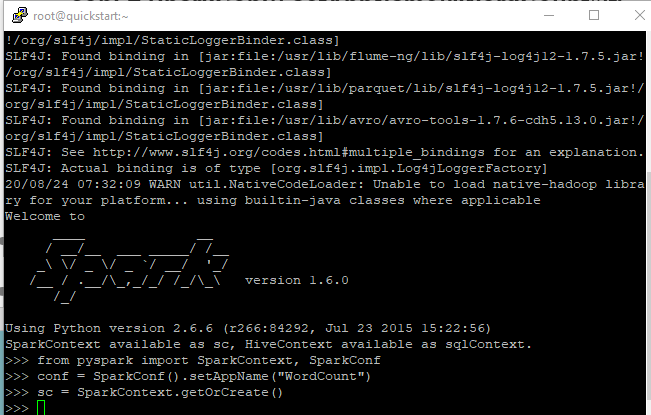
**contentRDD=sc.textFile(“user/estacio/textos/Shakespeare.txt”)**

1. Filtrar Linhas não vazias do arquivo carregado: **filter\_empty\_lines = contentRDD.filter(lambda x: len(x)>0)**
2. Dividir o conteúdo com base no espaço: **words = filter\_empty\_lines.flatMap(lambda x: x.split(‘’)**
3. Conte as palavras: **Wordcount = words.map(lambda x:(x,1)).reduceByKey(lambda x,y:x+y).map(lambda x: (x[1], x[0])).sortByKey(False)**
4. Visualizar o arquivo após o filtro: **For word in wordcount.collect():**

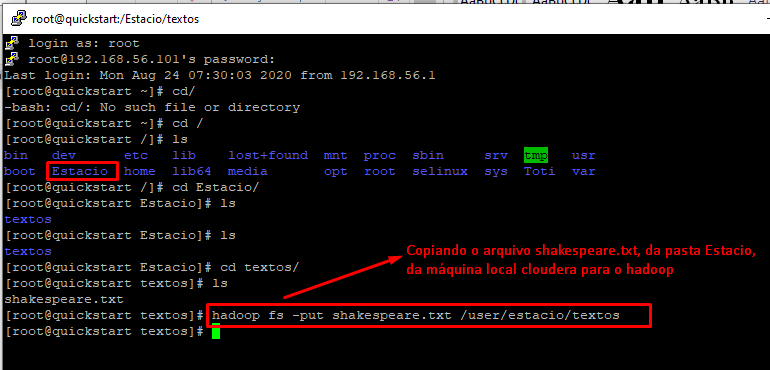
**Print(word)**

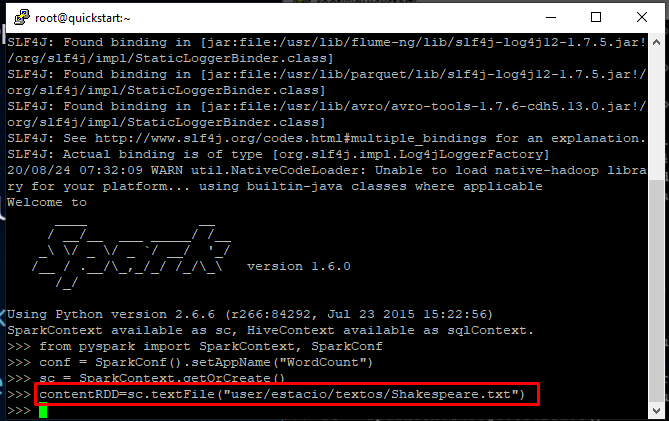
1. Salvar os dados finais no HDFS: **Wordcount.sveAsTextFile(“/user/toti/textos/Wordcont”)**

**Passos 1 a 3**

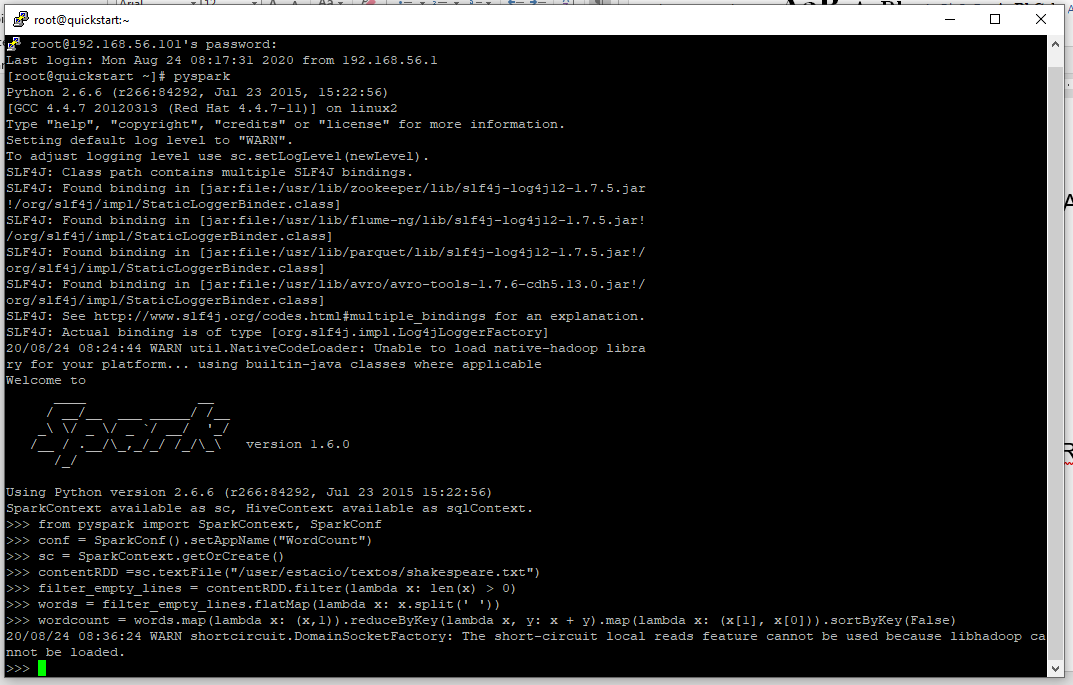


**Passo 4**

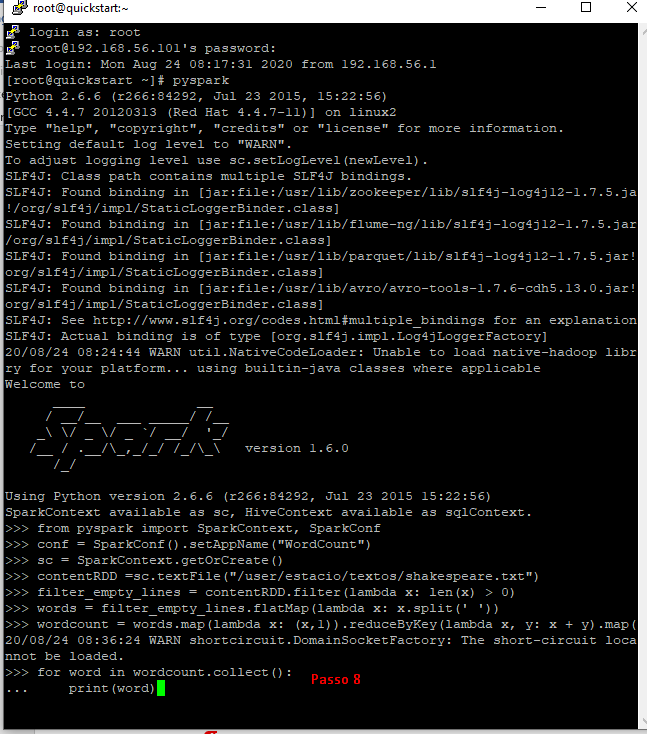




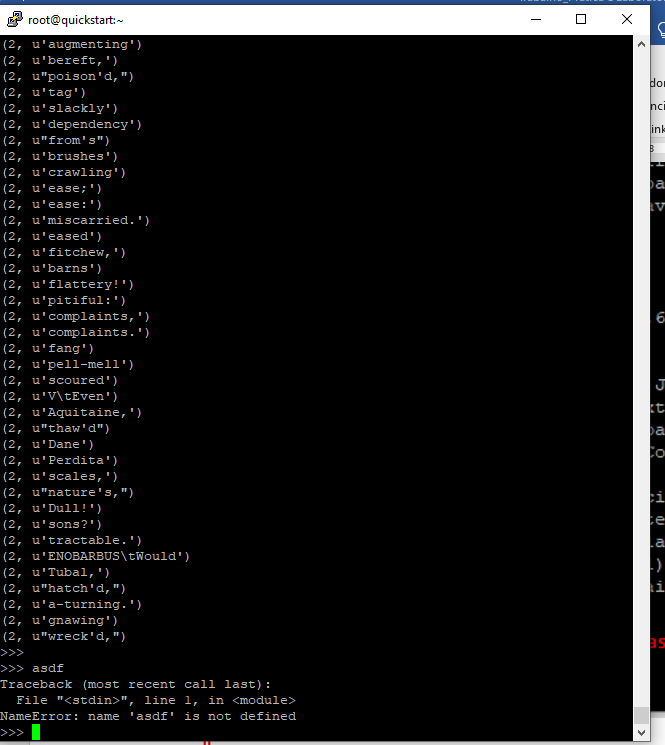
**Passo 5, 6 e 7**



**Passo 8**

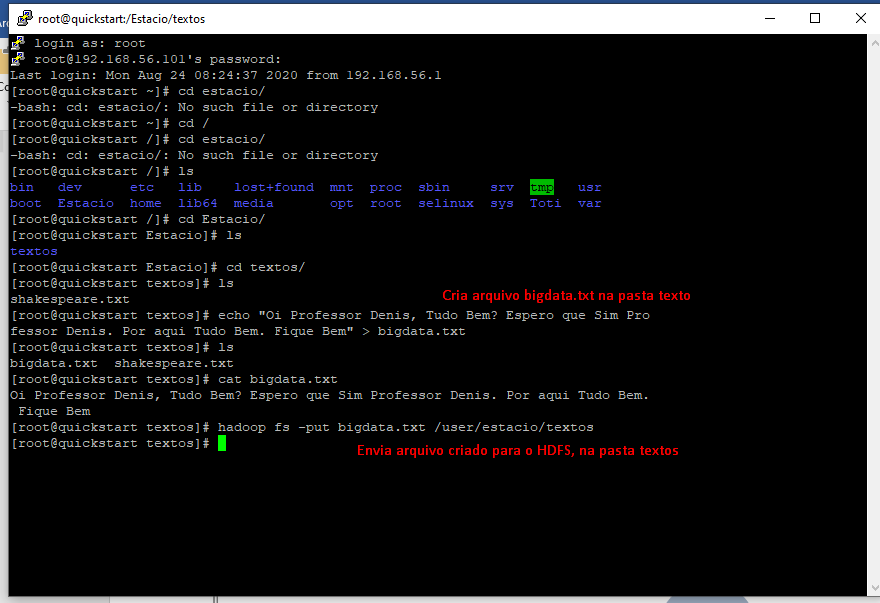


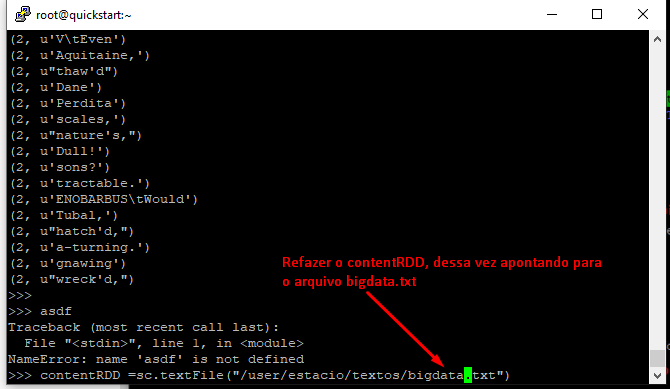
**Resultado da contagem:**

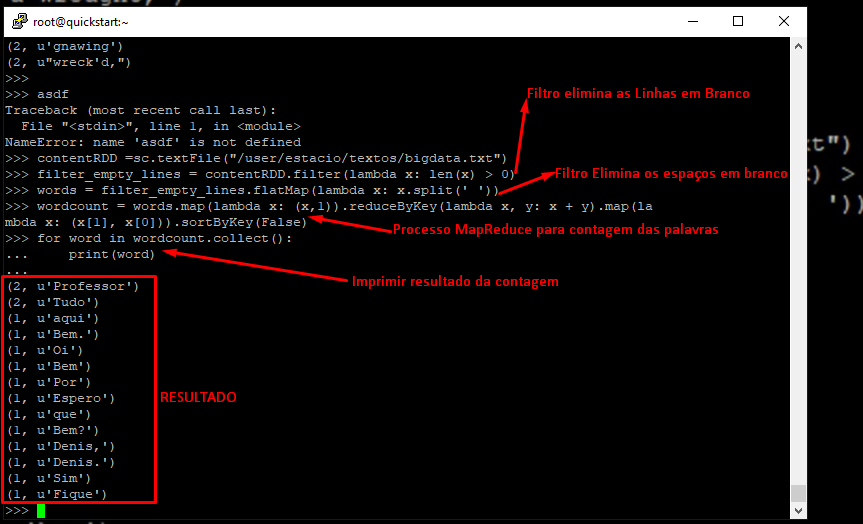


**Outro Exemplo - Wordcount**

Foi realizado ainda a criação de arquivo para servir de exemplo para a contagem de palavras. Criado através do terminal. O objetivo desse exemplo é visualizar melhor o processo de contagem de palavras do texto.



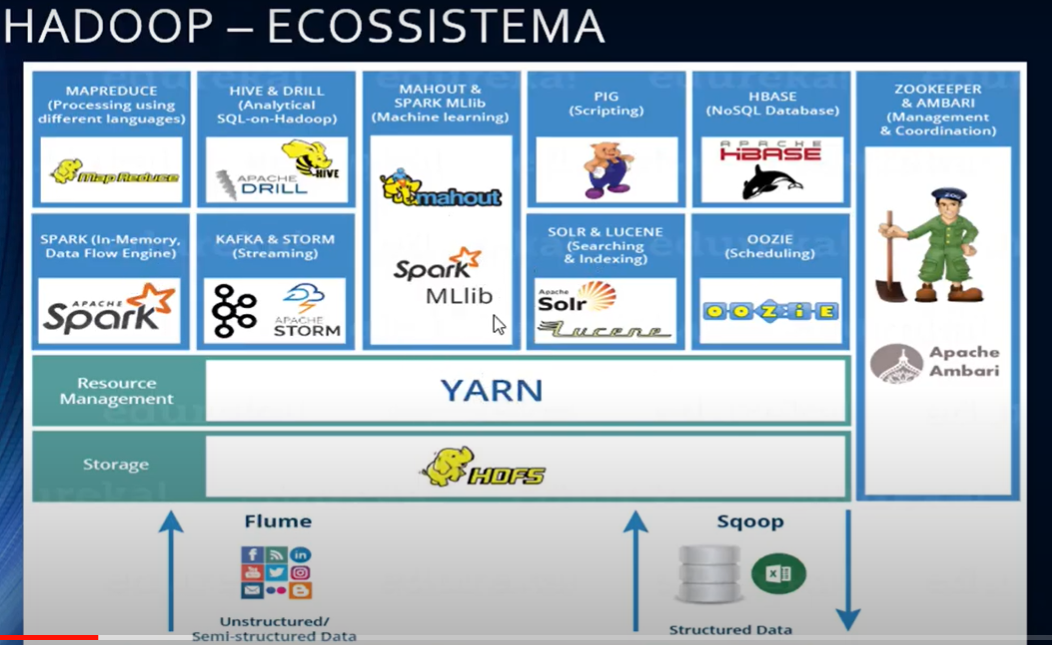




**Exercício 2 - Análise de Sentimento com Flume e Twitter**

Este projeto denominado “Análise de Sentimento” tem o objetivo de utilizar o Twitter como fonte de Dados (Streaming). Será utilizado: O FLUME, que é uma ferramenta do sistema Hadoop, para coletar os dados do Twitter; O HIVE para consultas aos dados e as tabelas e o Spark para desenvolver o Script de análise de sentimentos.

Ecossistema Hadoop com as funções do FLUME, HIVE e SPARK.



O Flume é um serviço de ingestão de dados para coletar, agregar e transportar grandes quantidades de fluxo de dados (streaming), como por exemplo: Arquivos de LOG, Eventos, Dados de Redes Sociais, Sensores, Etc, ou seja, várias fontes para um armazenamento centralizado.

Com a utilização da máquina Cloudera Quickstart , instalada na Oracle VirtualBox. O primeiro passo foi a configuração e testes para utilização do FLUME. Para isso foi inserido o arquivo de configuração na pasta /etc/flume-ng/conf, para permitir a conexão entre dois terminais PuTTY.

Script do Arquivo de configuração:

*# Nomeia os componentes do agente*

*a1.sources = r1*

*a1.sinks = k1*

*a1.channels = c1*

*# Descreve / configura a fonte (source)*

*a1.sources.r1.type = netcat*

*a1.sources.r1.bind = localhost*

*a1.sources.r1.port = 44444*

*# Descreve o sink*

*a1.sinks.k1.type = logger*

*# Usa um canal que armazena os eventos na memória*

*a1.channels.c1.type = memory*

*a1.channels.c1.capacity = 1000*

*a1.channels.c1.transactionCapacity = 100*

*# Vincula o source e o sink ao canal*

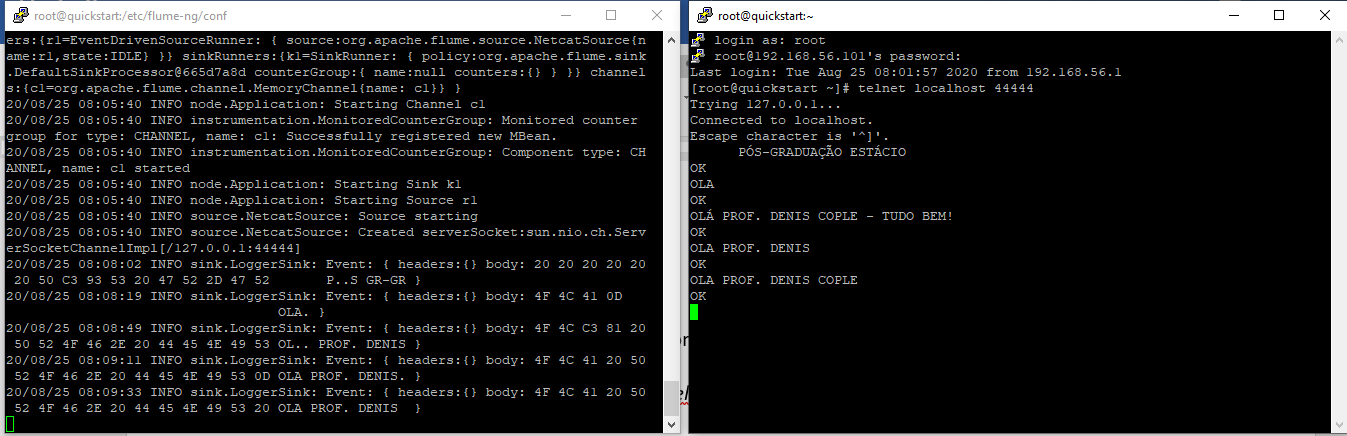
*a1.sources.r1.channels = c1*

*a1.sinks.k1.channel = c1*

Em seguida foi aberto o terminal PuTTY. Com a máquina logada digitamos o comando: *flume-ng agente -n a1 -c conf -f exemplo.conf* .

Em outro Terminal digitamos o comando *telnet localhost4444*

Com a realização das operações acima foi possível verificar a conexão entre os dois terminais.



Após o teste com o FLUME bem sucedido, passou-se a criação de App Twitter, através do endereço <https://developer.twitter.com/> , para a criação do App foi realizado o cadastro da ferramenta e informado o objetivo do acesso que é destinado a criação de Trabalho para a Faculdade.

Quando da criação do APP Twitter foi obtido as seguintes chaves de acesso:

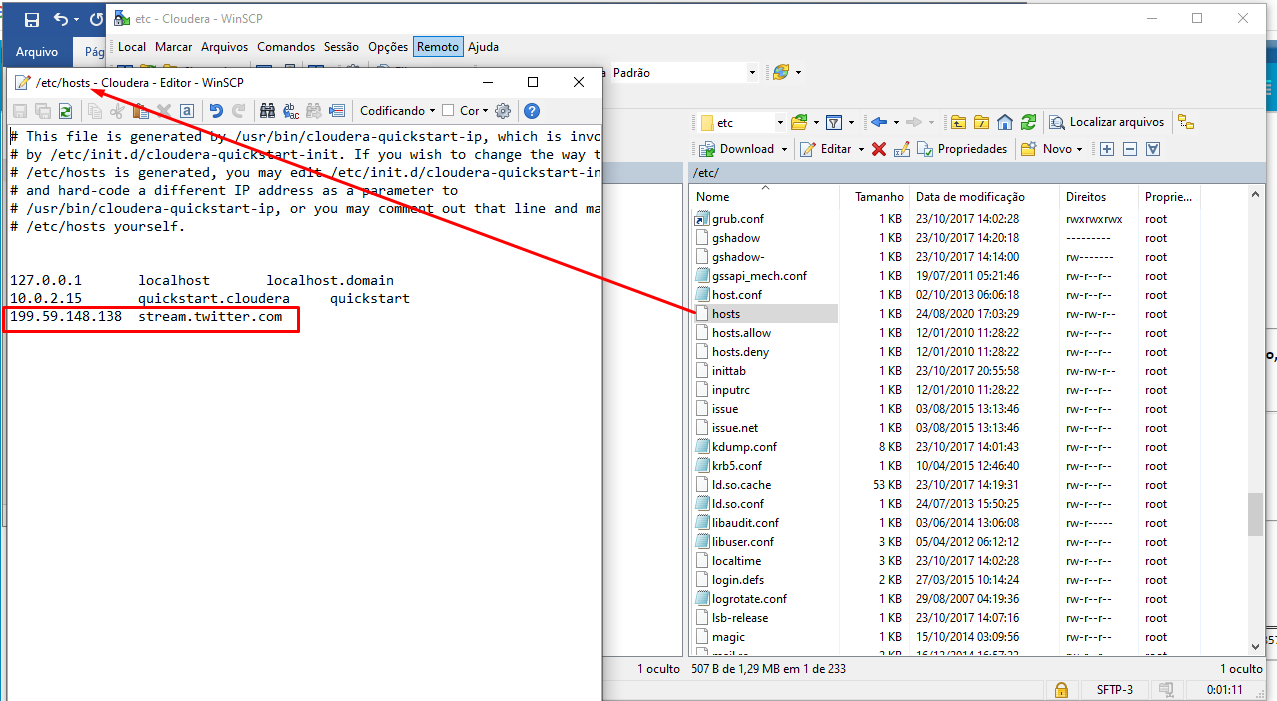
API key: 6jfpaDKHiwZMXB2auX5lHOVZj

API key secret: qeZ8Kdsr2tzQYJBpdnpEIhvrFgj6Fd5y7DEbCU5ZliJksxCqWh

Access token: 1089494852050534400-IrdQqCycXidG4TGlO2BBbh0Od8ShJF

Access token secret: 3zS5z8GKpqdvWZd54rFyNENIpbou85t380DNJ3I1KhXvT

Para liberar o acesso do App Twitter a máquina Cloudera foi necessário alterar o arquivo hosts da máquina Cloudera e inserir o endereço: *199.59.148.138 stream.twitter.com*. *Para isso foi utilizado o* ***WinSCP***



Em seguida foi criado um Diretório no HDFS para salvar os arquivos FLUMEDATA.

Comando: *hadoop fs -mkdir -p /twitteranalytics/*

Foi necessário ainda criar um arquivo de configuração do Agente para o App Twitter e inserir na configuração do FLUME na máquina Cloudera no local: */etc/flume-ng/conf/ .*

**Arquivo de Configuração do Agente para App Twitter:**

*# Naming the components on the current agent.*

*TwitterAgent.sources = Twitter*

*TwitterAgent.channels = MemChannel*

*TwitterAgent.sinks = HDFS*

*# Describing/Configuring the source*

*#TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource*

*TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource*

*TwitterAgent.sources.Twitter.consumerKey = 6jfpaDKHiwZMXB2auX5lHOVZj*

*TwitterAgent.sources.Twitter.consumerSecret = qeZ8Kdsr2tzQYJBpdnpEIhvrFgj6Fd5y7DEbCU5ZliJksxCqWh*

*TwitterAgent.sources.Twitter.accessToken = 1089494852050534400-IrdQqCycXidG4TGlO2BBbh0Od8ShJF*

*TwitterAgent.sources.Twitter.accessTokenSecret = 3zS5z8GKpqdvWZd54rFyNENIpbou85t380DNJ3I1KhXvT*

*TwitterAgent.sources.Twitter.keywords = @IveteSangalo, @ClaudiaLeite, @Jorge&Mateus, @MichelTelo, @GustavoLima, @Anitta, @Ludmilla, @LeoSantana, @TiagoIorc, @Simone&Simaria, @MariliaMendonca @RobertoCarlos, @VitorKley, @CarlinhosBrown*

*# Describing/Configuring the sink*

*TwitterAgent.sinks.HDFS.type = hdfs*

*TwitterAgent.sinks.HDFS.hdfs.path = /twitteranalytics/incremental*

*TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream*

*TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text*

*TwitterAgent.sinks.HDFS.hdfs.filePrefix = twitter-*

*TwitterAgent.sinks.HDFS.hdfs.rollInterval = 0*

*TwitterAgent.sinks.HDFS.hdfs.rollSize = 524288*

*TwitterAgent.sinks.HDFS.hdfs.rollCount = 0*

*TwitterAgent.sinks.HDFS.hdfs.idleTimeout = 0*

*TwitterAgent.sinks.HDFS.hdfs.batchSize = 100*

*TwitterAgent.sinks.HDFS.hdfs.threadsPoolSize = 2*

*TwitterAgent.sinks.HDFS.hdfs.round = true*

*TwitterAgent.sinks.HDFS.hdfs.roundUnit = hour*

*# Describing/Configuring the channel*

*TwitterAgent.channels.MemChannel.type = memory*

*TwitterAgent.channels.MemChannel.capacity = 10000*

*TwitterAgent.channels.MemChannel.transactionCapacity = 100*

*# Binding the source and sink to the channel*

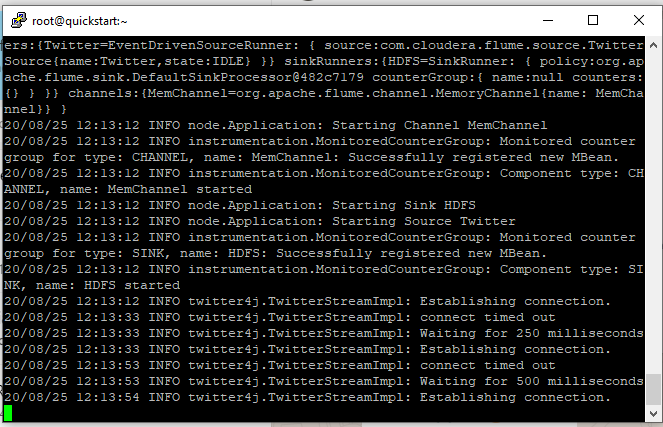
*TwitterAgent.sources.Twitter.channels = MemChannel*

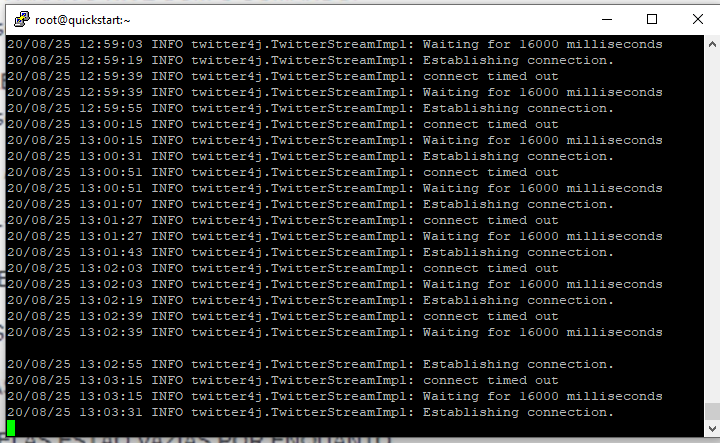
*TwitterAgent.sinks.HDFS.channel = MemChannel*

*#To turn on the agente, in the folder /etc/flume-ng/conf.empty use the command:*

*#flume-ng agent -f flume\_twitter.conf Dflume.root.logger=DEBUG,console -n TwitterAgent*

***[root@quickstart conf]# mkdir -p /usr/lib/flume-ng/plugins.d/twitter-streaming/l***

******



Na execução da conexão apresenta problemas e não é não consegue estabilizar a conexão.

Refeito algumas vezes sem sucesso, finalizo o trabalho de extração de Grande volume de Dados do Twitter para análise de sentimento.

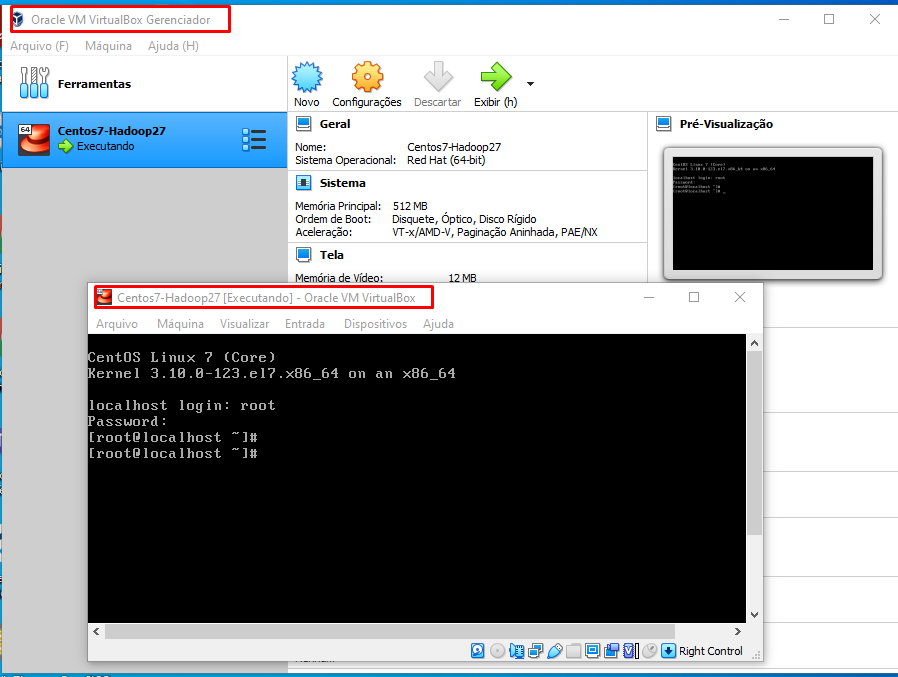
Considerando que durante o Tutorial do Curso de Hadoop , do canal Código Fluente, do youtube, foi apresentado a senha “Cloudera” para usuário hoot da máquina Cloudera quickstart, resolvi fazer novos testes com o conteúdo da Disciplina Tecnologias Avançadas para ver se conseguiria progredir haja vista que o principal problema encontrado até então era a senha de usuário hoot.

Após refazer as instalações da Vmware e hortonworks SAndbox, novamente ocorreu na interface gráfica e no acesso SSH não foi possível conectar com a senha “**cloudera” .**

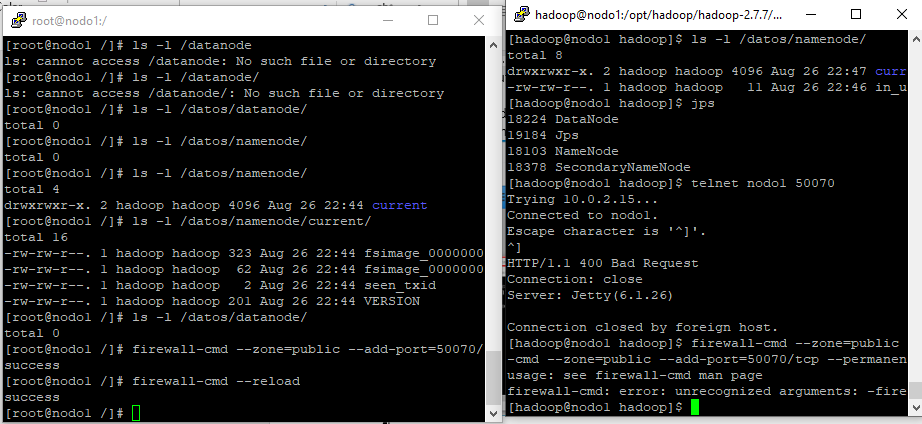
Na busca do conhecimento e aprendizado do conteúdo da disciplina que tem como foco o Ecossistema HADOOP, assisti mais alguns vídeos sobre o assunto, a maioria fala muito das teorias e conceitos envolvidos, mas poucos tem a abordagem prática. Em determinado momento de estudos encontrei a Playlist “Big Data Hadoop Español [CURSO COMPLETO]”, composto de 24 vídeos de conteúdo hands on.

Entendendo que as atividades propostas nos vídeos vão ao encontro do proposito da disciplina e atendem em parte os requisitos do trabalho proposto, apresento a descrição passo a passo das atividades executadas relacionadas ao ambiente Hadoop.

Os Softwares utilizados no Tutorial foram a Oracle VM VirtualBox, a máquina virtual Linux CentOS-7.0-amd64-minimal.ova



Em seguida foi instalado o PuTTY e o WinSCP para as configurações da máquina virtual.



Em seguida foram realizadas as seguintes operações, por ordem:

- Configuração do Cliente SSH com a inclusão de uma nova Porta para máquina localhost

- Instalação de pacotes hadoop com o comando yum (net-tools; wget; telnet)

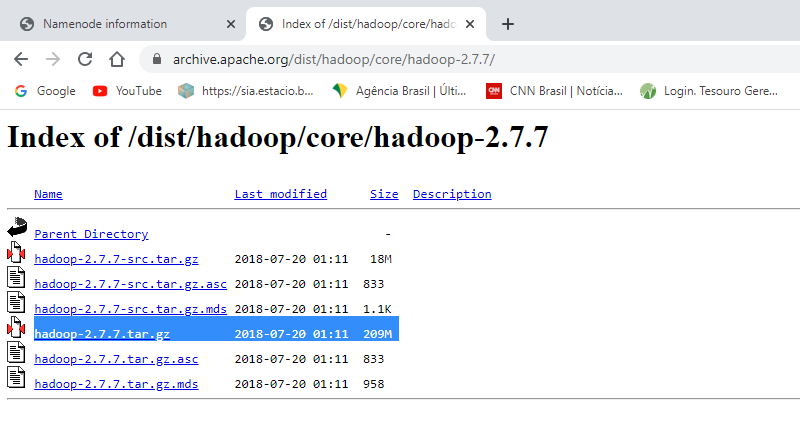
- Criação de usuário “hadoop” para receber os dados dos processos hadoop

- O Próximo passo foi a instalação do JDK 8 direto na máquina CentOS7: A instalação foi feita pelo PuTTY através do comando alternatives:



- Após as instalações do JAVA foi realizada a configuração do JAVA\_HOME.

- Depois foi a vez de baixar a versão hadoop-2.7.7 para máquina CentOS07, através do link do arquivo

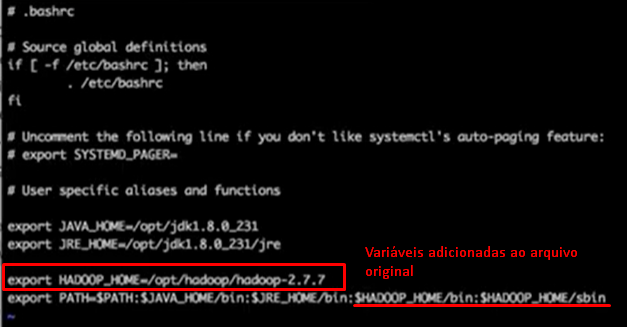




- Em seguida foi feita a descompressão do arquivo através do comando : tar -zxvf hadoop-2.7.7.tar.gz

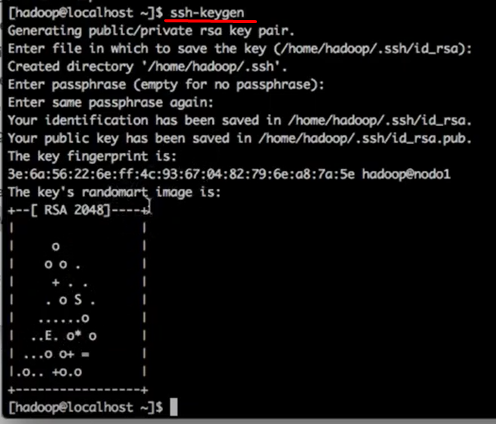
- O próximo passo foi realizar configurações do ambiente Hadoop

- A primeira alteração foi a inclusão de variáveis no arquivo .bashrc



- em seguida foi realizada instalação do ambiente de cluster para o Hadoop trabalhar.

- Foi criada uma sessão pública no ambiente para acesso livre entre os nodes . Comando: ssh-keygen



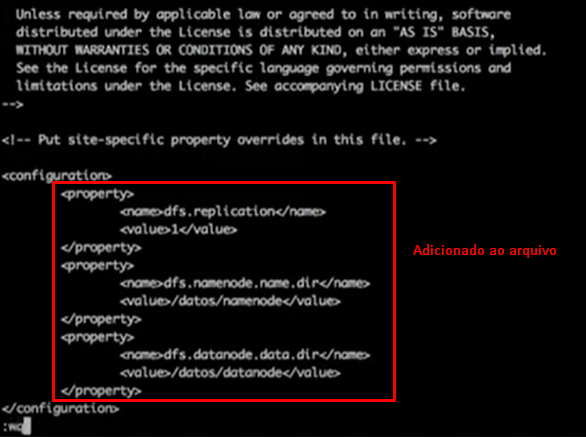
- Desta forma foi estabelecida conexão local

- A Próxima configuração foi a inclusão de variáveis ao arquivo core-site.xml



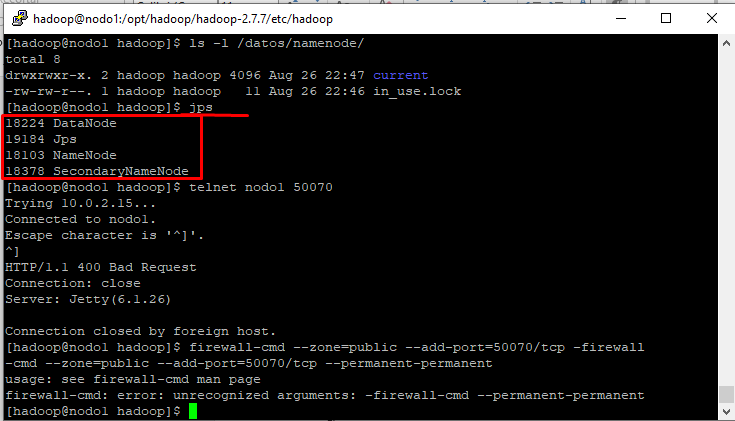
- em seguida foi inserido o comando: firewall-cmd --zone=public --add-port=9000/tcp –permanent , para configurar acesso a nova porta.

- A configuração seguinte foi a alteração do arquivo hdfs-site.xml:

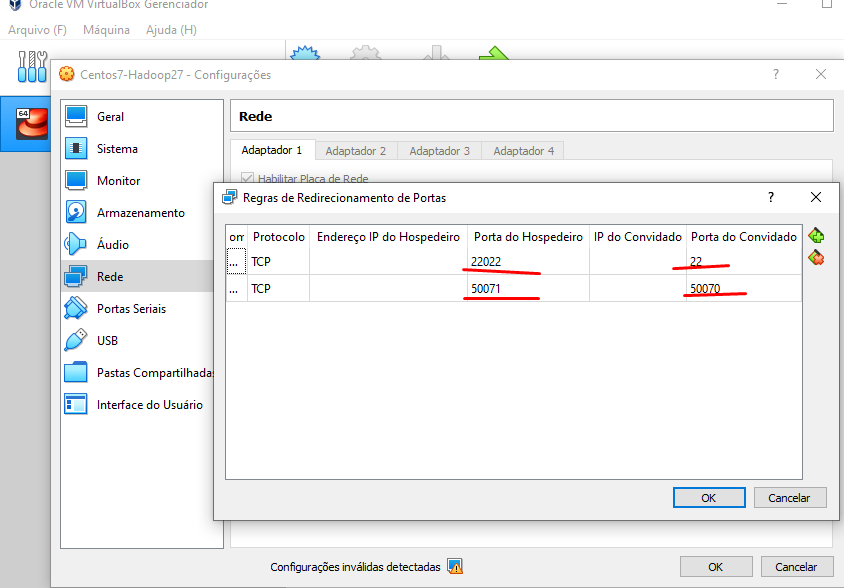


- Em seguida foi realizada a formatação e inicialização do hadoop.

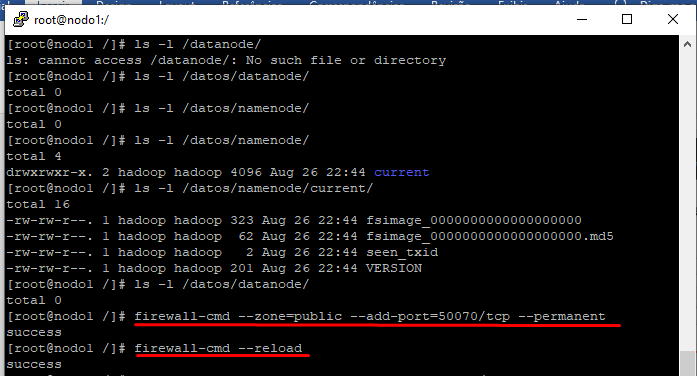
- O Ambiente e de cluster para trabalhos Hadoop Java é verificada conforme comando *jps*



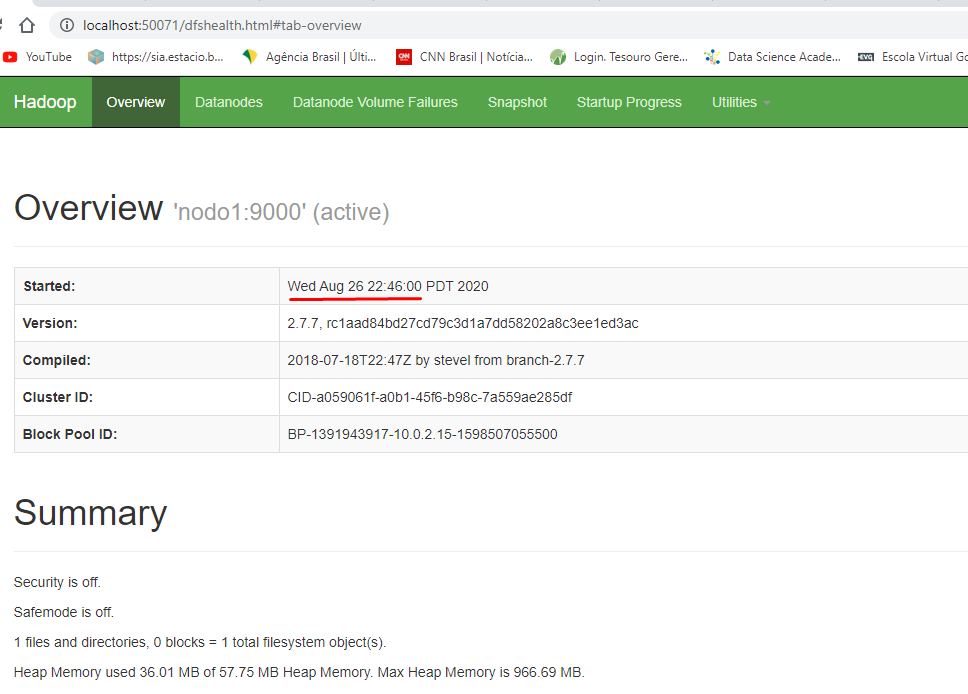
- Adicionalmente, para resolver problema de conexão foi necessário configurar a Porta direto na máquina virtual, conforme abaixo :



- Em seguida foi aberta a Porta 50070 no HADOOP.



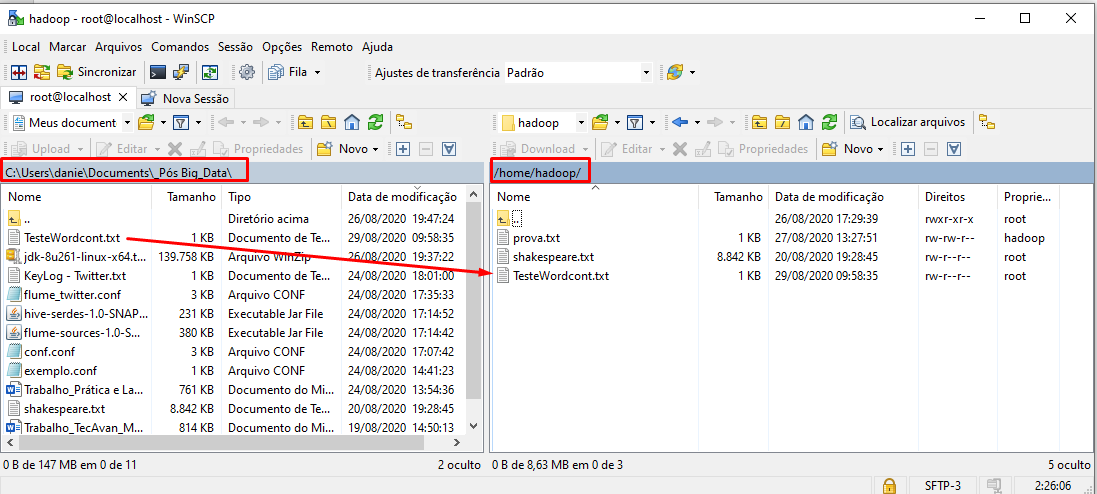
- É possível ser feita a prova da conexão do hadoop com com a nova porta através do browse, digitando o endereço: localhost:50071 :



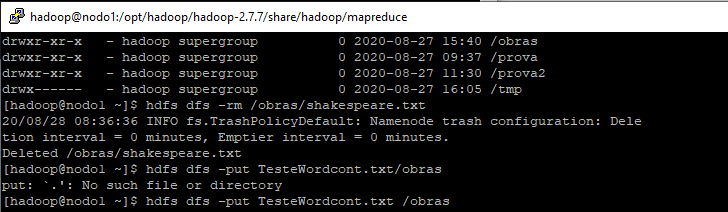
**Executando MapReduce – JAVA – wordcount**

Para execução do MapReduce e contagem de palavras com a utilização do JAVA , inicialmente copiamos o arquivo que queremos contar as palavras para uma pasta criada dentro do *hadoop hdfs dfs*

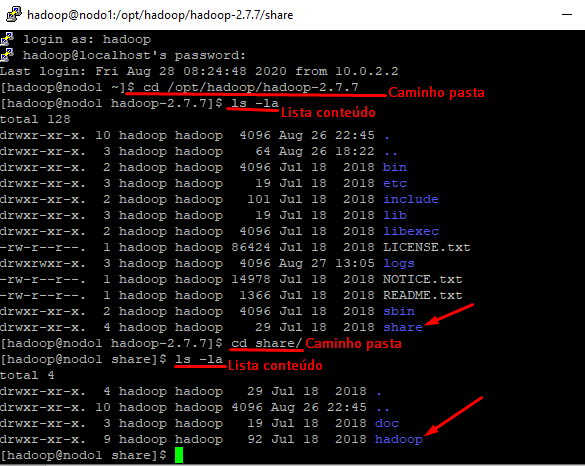
Foi utilizado WinSCP para copiar o arquivo da pasta local para o pasta criada dentro do haddop na máquina virtual CentOS7.

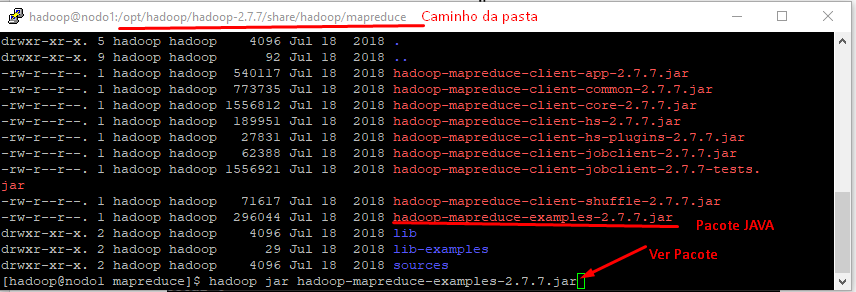


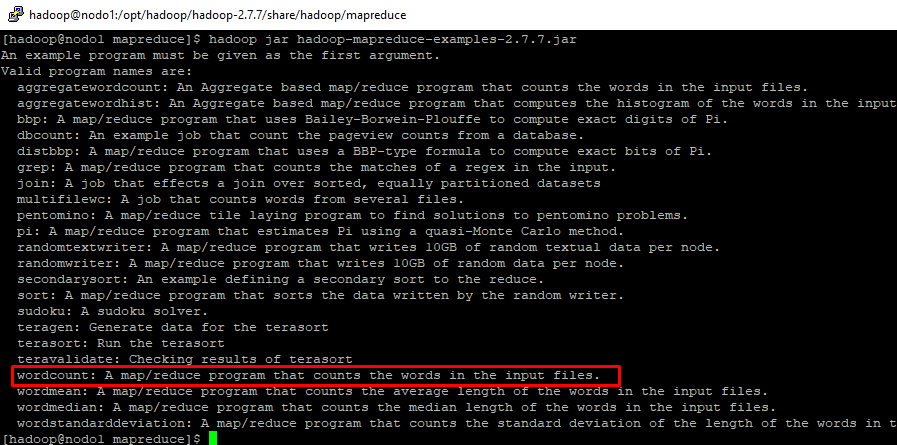
Em seguida foi feita a transferência do arquivo do diretório hadoop para uma pasta (obras) criada em *hdfs dfs* :



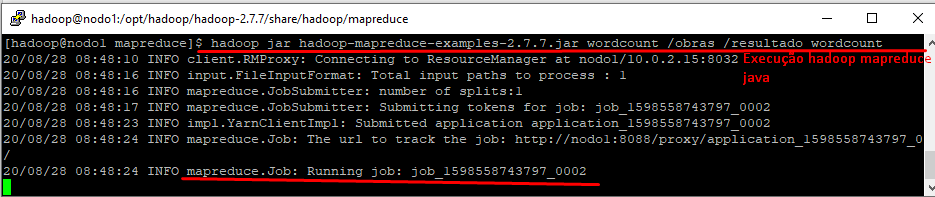
Para visualizar a função Java – wordcount, é possível abrir seu diretório conforma demonstrado abaixo:







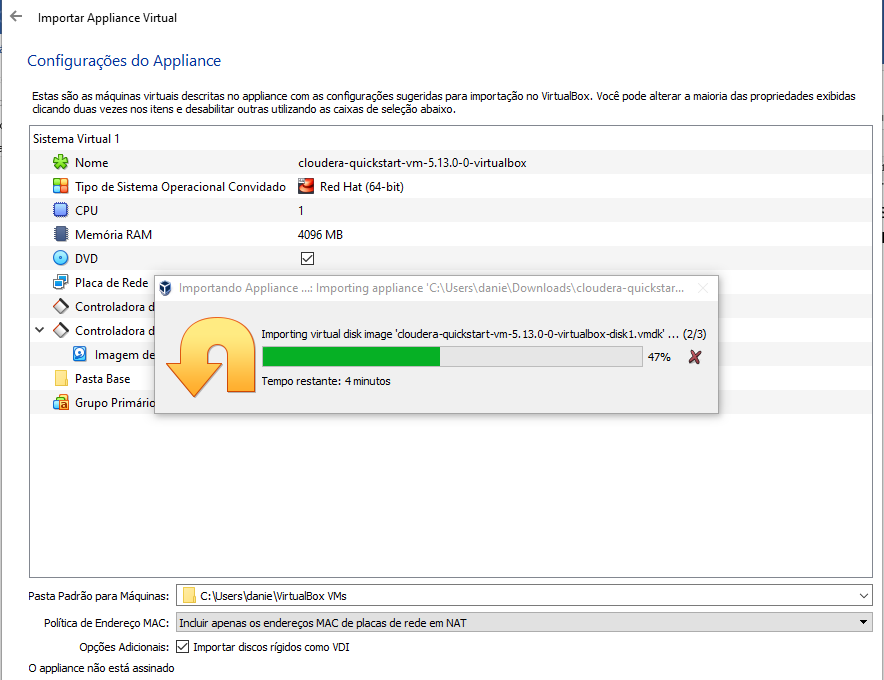
Execução MapReduce Java - WordCount

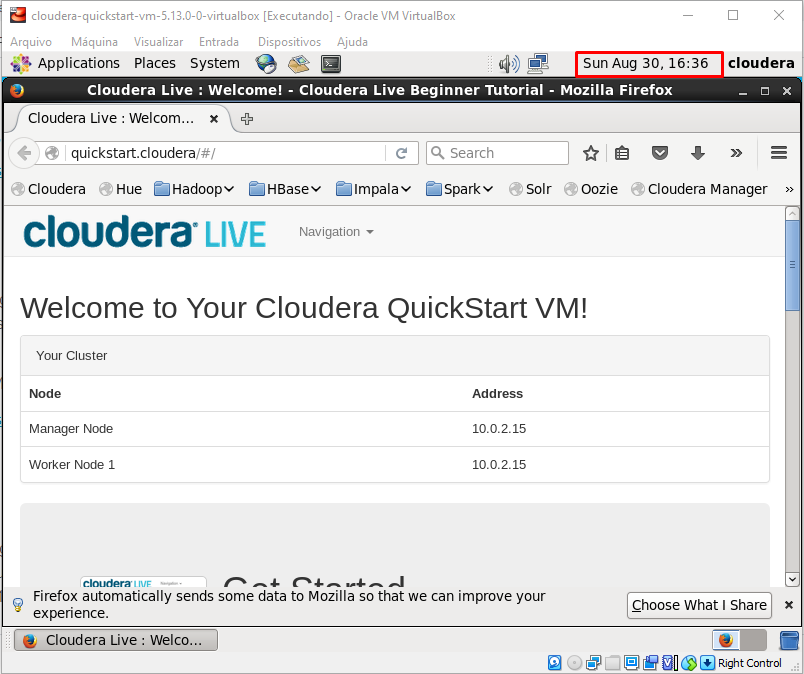


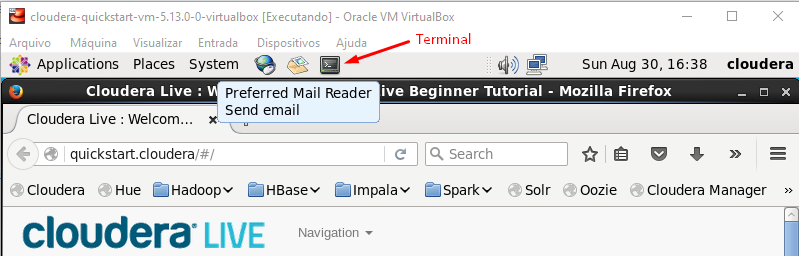
# 

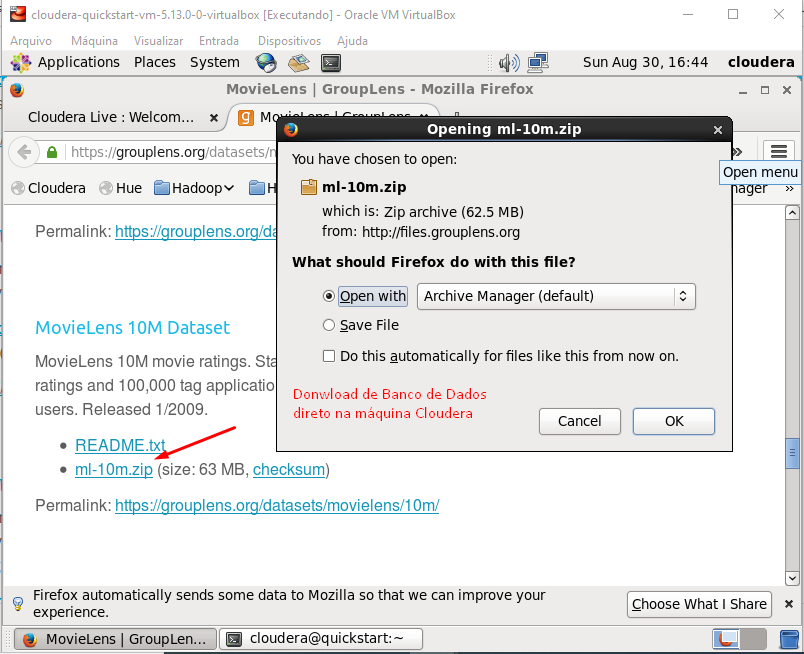
**Exercício 3**

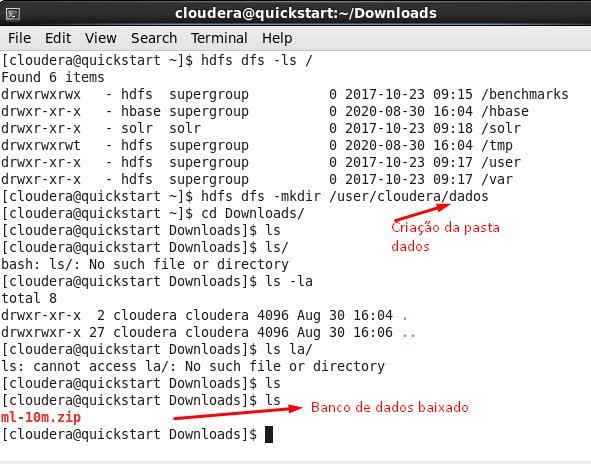
Para realização das atividades do Curso Udemy – Big Data com Hadoop: direto ao ponto e foco na prática. Foi utilizado a máquina Virtual Cloudera Quickstar VM. Instalada na Oracle VM Virtual Box.

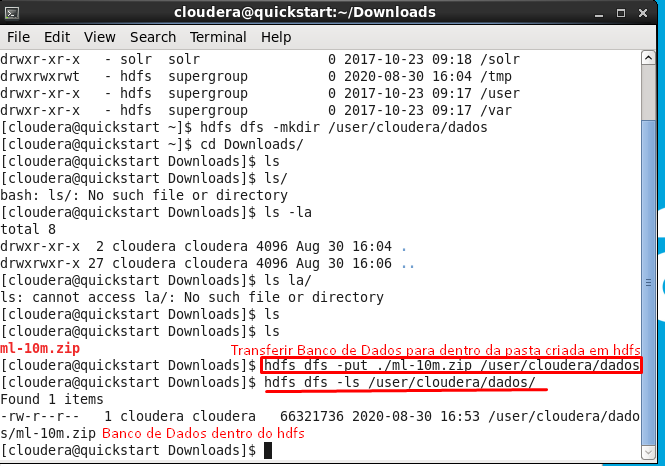


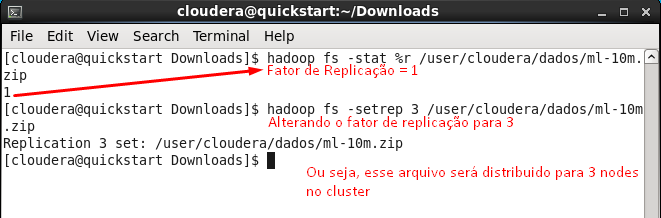


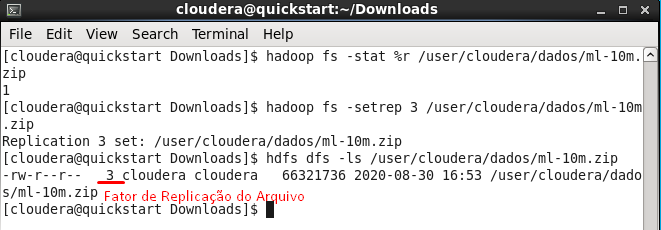


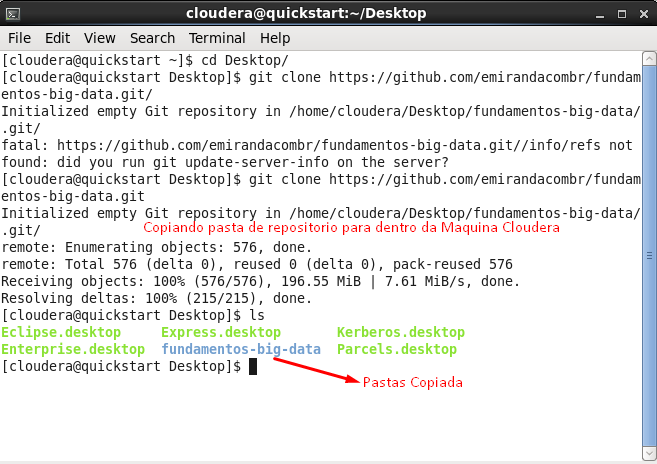












Endereço para obtenção dos dados:

<https://github.com/Dmguedes/Trabalho-Big-Data/blob/master/Pratica_e_Laboratorio.md>