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

**UNIVERSIDADE ESTÁCIO DE SÁ**

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

**Projeto MapReduce com Java**

**Aluno: Daniel Mendes Guedes**

**Trabalho da disciplina**

**Tecnologias Avançadas**

**Tutor: Prof. Denis Gonçalves Cople**

**Brasília/DF**

**2020**

**PROJETO MAPREDUCE JAVA**

**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

**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 nem tempo suficiente para levantar todas as dúvidas e resolver com o professor, tutor, no período de abertura dos Fóruns. Para mim, além de ser novidade, foi ruim essa metodologia de não ter contato com os professores durante toda a realização do curso.

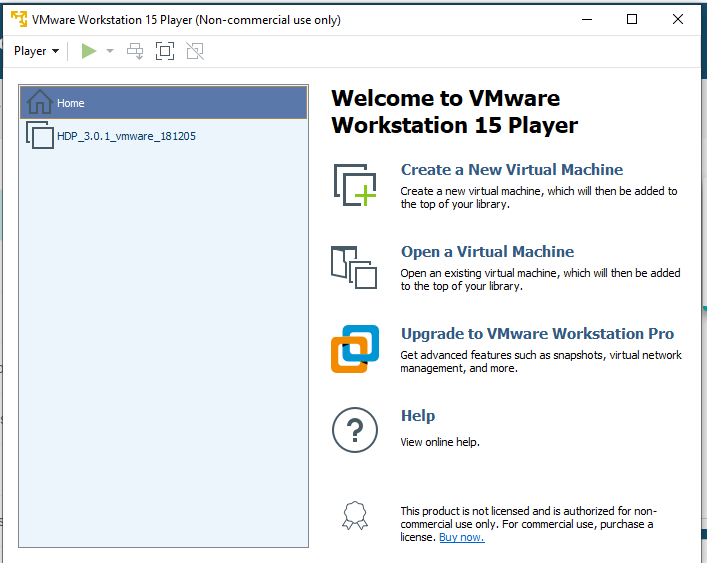
**Dificuldades encontradas**

Após assistir as Videoaulas e rever o conteúdo dos 571 slides da matéria, tive dificuldades para fazer funcionar todas as ferramentas indicadas. Percebi que as versões dos softwares são ultrapassadas e tive dificuldades com usuário e senha de acesso. De todo modo, para fazer a entrega do trabalho obrigatório da Disciplina, irei refazer todos os processos, passo a passo, desde a instalação da Máquina Virtual, da Aula 6 até a execução do Projeto Java na Aula 15 e irei registrar todo o histórico de tentativas e acertos e erros na consecução do projeto e procurando demonstrar o esforço no aprendizado da disciplina.

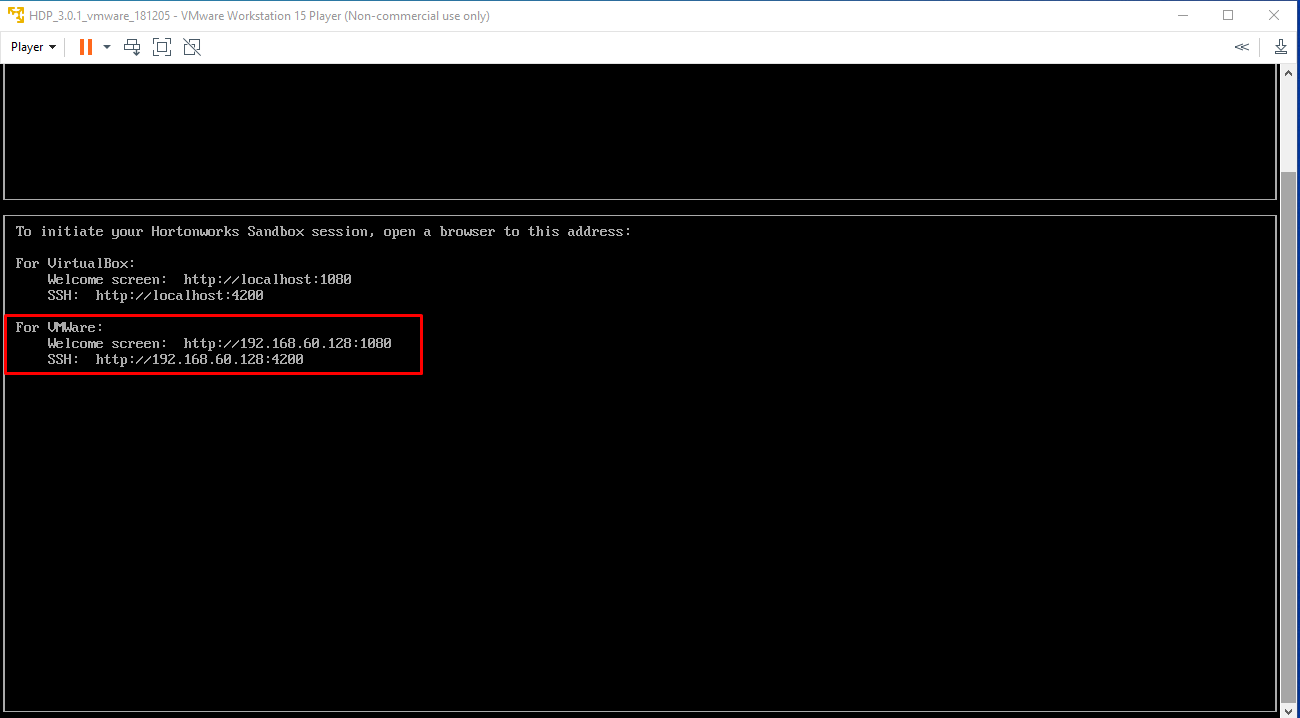
**Atividade proposta no Material da Disciplina**

**Histórico**

Na 3ª tentativa (10/08/2020) consegui realizar a instalação dos Programas VMware Player Pro 15 e a Plataforma de dados Hortonworks Sandbox – HDP\_3.0.1.OVA, versões muito superiores as indicadas no material, entretanto na Aula 6 é informado que pode ser instalado versões mais recentes.



**Imagem: VMware Player 15 instalado com sucesso.**



**Imagem: Instalação ok – Hortonworks Sandbox**

As instalações foram concluídas com sucesso e a Máquina Virtual ficou operante.

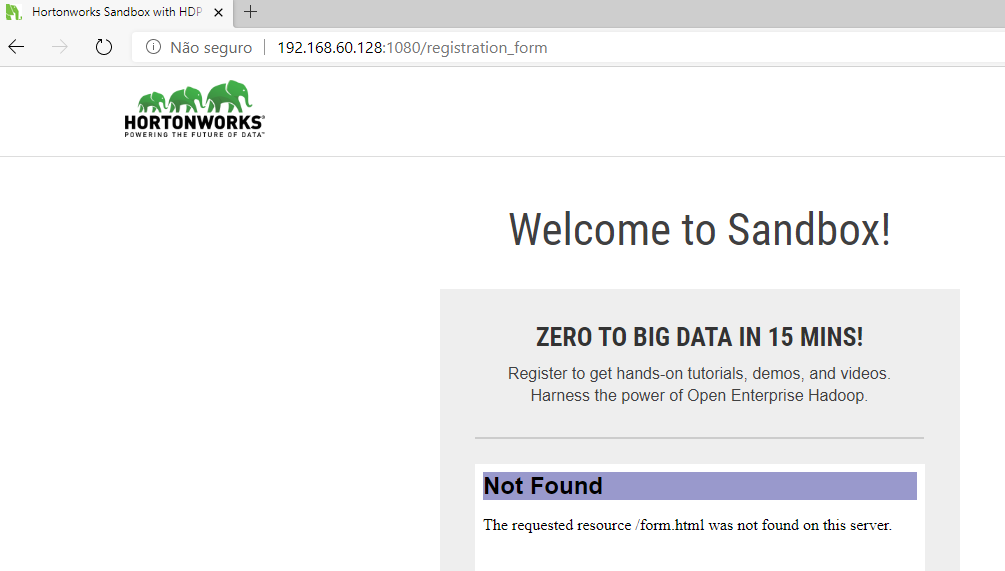
Uma vez que a Maquina Virtual estava funcionando e o Sandbox também estava funcionando e ligado o próximo passo era acessar o formulário de registro de primeiro acesso ao Sandbox.

O endereço informado no material foi: <http://192.168.60.128/registration_form>. Sendo que o meu Número de Ip, conforme Worstation mostra é: .

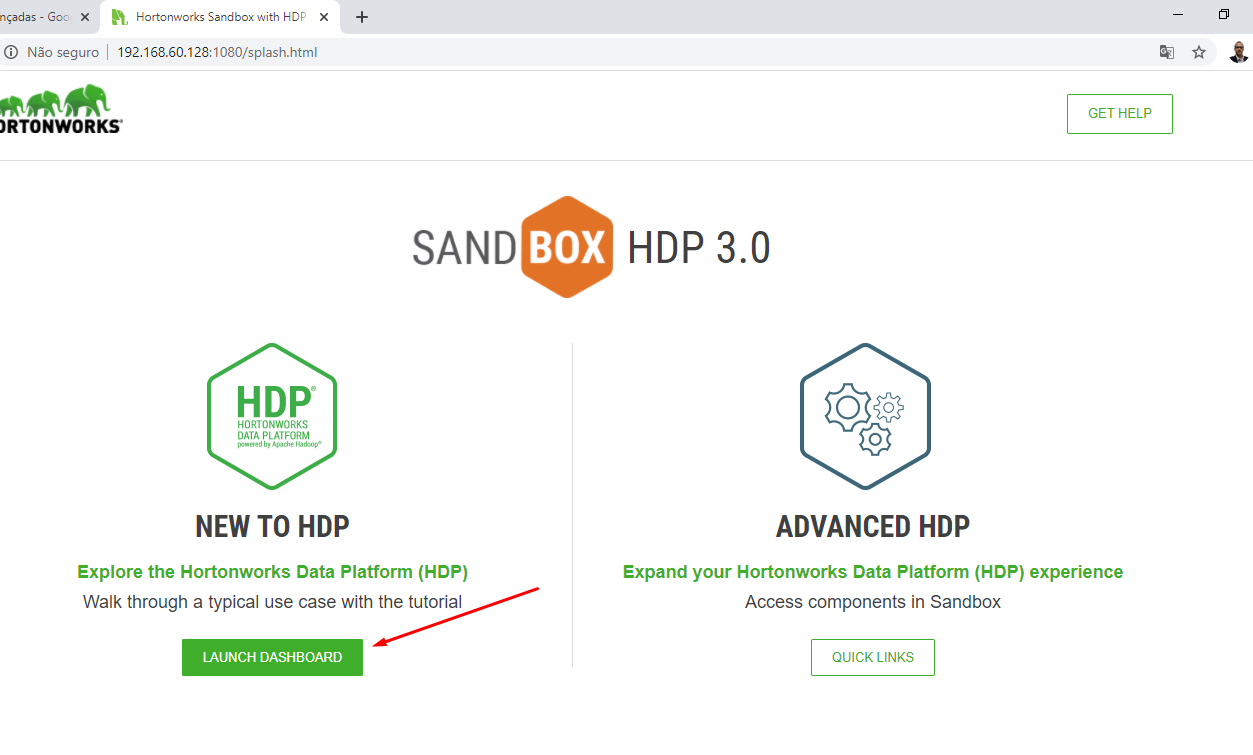
Ao acessar o endereço <http://192.168.60.128/registration_form> o resultado foi:



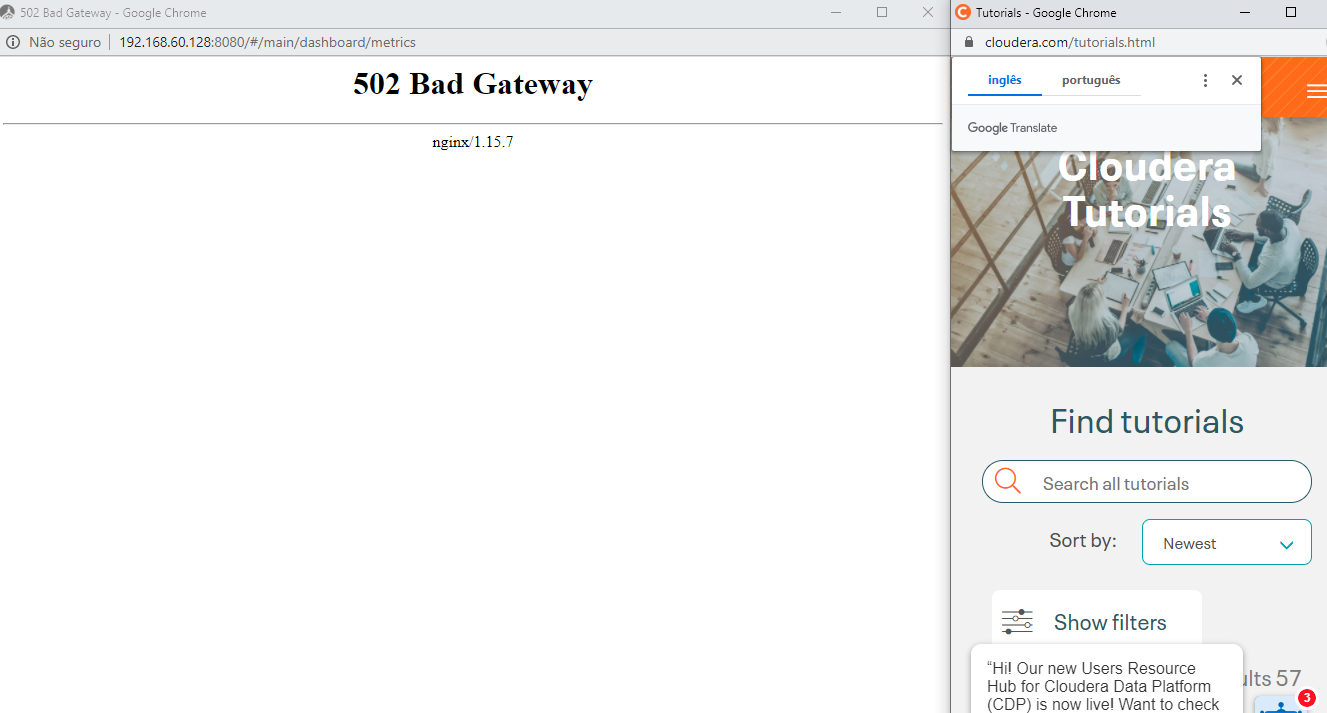
Em seguida tentei o acesso informando a Porta de acesso, no endereço: <http://192.168.60.128:1080/registration_form>, e o resultado solicita que seja feito o registro para acesso aos tutoriais e demos.



Ao tentar o acesso sem informar “*registration\_form”*, apenas a porta de acesso o resultado foi:



Após clicar no botão “*Lanch Dashboard*” o resultado foi:



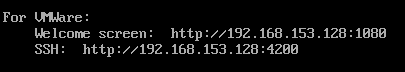
**Dificuldades Encontradas:**

Após instalar a HORTONWORK SANDBOX, verifiquei que a instalação consumiu aproximadamente 83G do meu sistema, que possuía uma Unidade de Armazenamento SSD de 120G com o SO Windows instalado, e um HD-SATA de 500G como auxiliar. O SO Windonws e outros arquivos estavam utilizavam 34G. No total (SandBox + Sistema Nativo) estavam consumindo mais de 117G do SSD. A Falta de memória começou a ser um problema e a máquina virtual começou a apresentar problemas para iniciar, tentei instalar o hadoop no Diretório do HD-500G mas a máquina virtual deu erro.

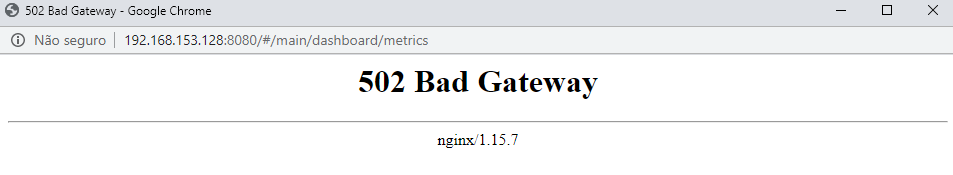
Após 3 dias brigando com os sistemas, excluindo e reinstalado os sistemas e nada de progresso. Encontrei a alternativa de fazer um Clone do dispositivo SSD-120 para o HD-500G. Depois desinstalei o SSD-120, e utilizei apenas o HD de 500G para resolver a questão de memória. Entretanto perdi na velocidade e performance do equipamento.

Em 15/08/2020 reiniciei todos os processos de instalações da VMware, Sandbox e Bitvice, Eclipse, outros aplicativos e programas no HD de 500G.

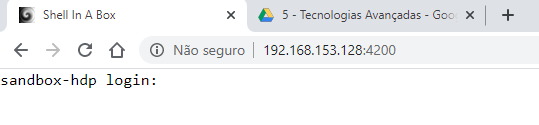
Após a nova instalação da VMware e SandBox, verificou-se que o endereço da máquina foi alterado para:



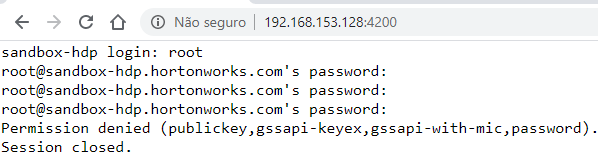
Ao acessar o endereço <http://192.168.153.128:1080> o resultado novamente foi:



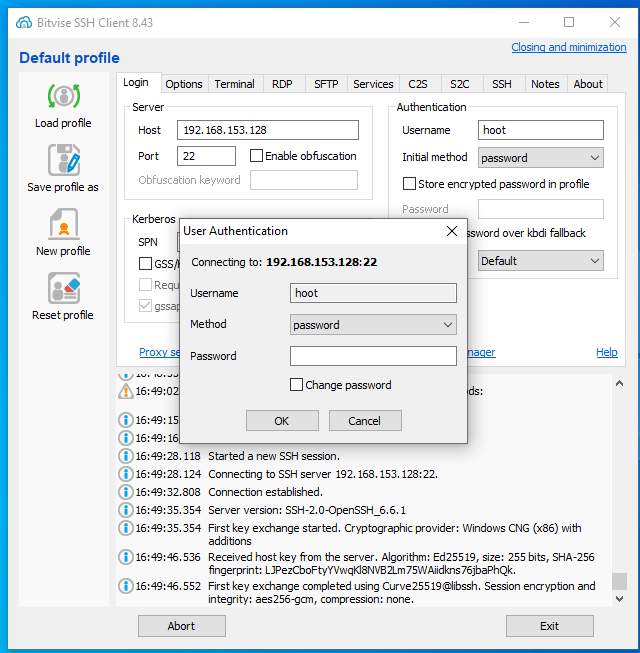
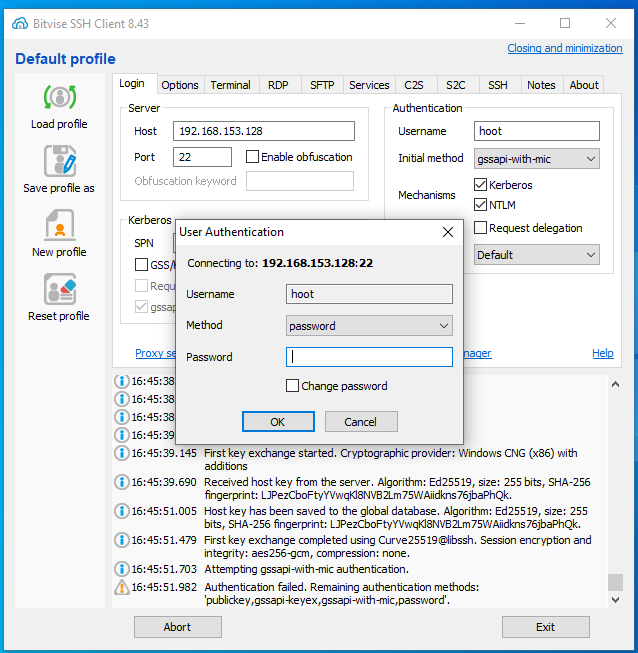
Ao acessar o endereço <http://192.168.153.128:4200> o resultado dessa vez foi:



O Acesso ao terminal Sandbox-hdp, pela Porta 8080 indicou que a máquina estava funcionando e operante. Entretanto não foi possível fazer o acesso nenhum usuário informado no material da disciplina, root e hue, as senhas não conferiram e não encontrei , em minhas pesquisas, a senha correta. Mas, ressalto que o Linux e o Hadoop foram instalados com êxito na Máquina Virtual.

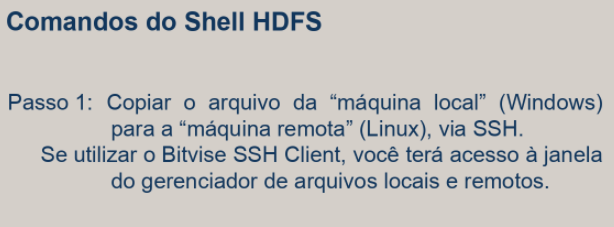


Em continuidade as instalações e orientações da Aula 8 – foi feita a instalação do bitvise com sucesso. Todavia, novamente foi solicitado login e senha que eu não consegui obter nem validar, conforme imagens abaixo:

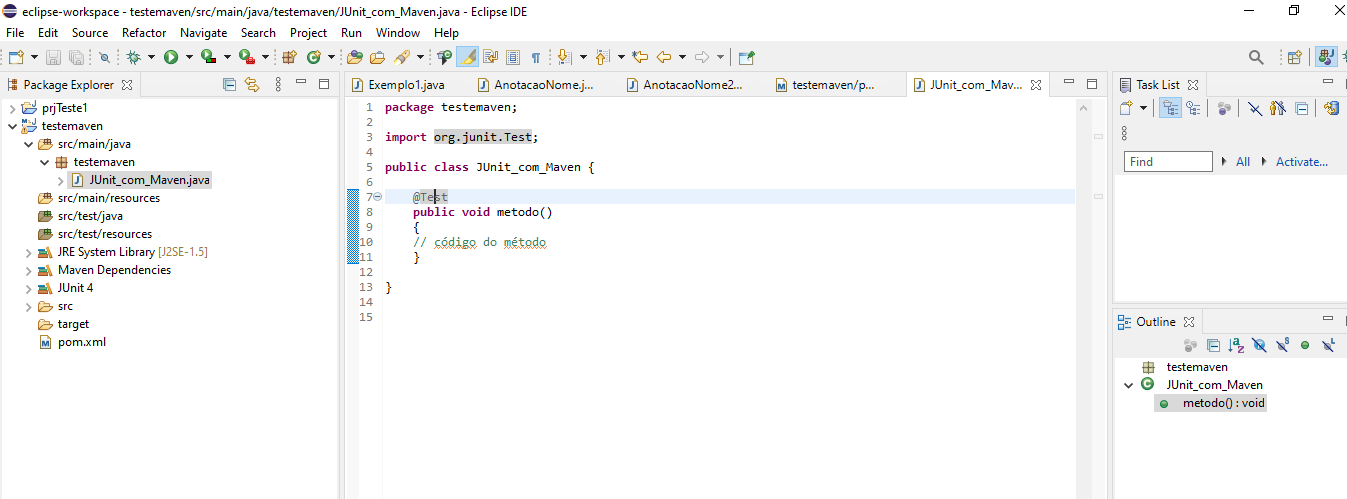


Não foi possível acessar o Bitvise nem sua interface gráfica devido a falta de password do usuário hoot. (Fiz varias pesquisas mas não consegui resolver)

Na sequencia do conteúdo, na Aula 10, é demonstrado como executar comandos do Shell HDFS, entretanto é necessário a utilização do Bitvise para gerenciar os arquivos da máquina local para a máquina virtual, mas eu não consegui iniciar o Bitvise.



Na aula 11 é apresentado a IDE Eclipse para realizar Projetos Java. A IDE foi instalada com sucesso e todas as atividades e ações propostas na Aula 12 foram realizadas também com sucesso, as soluções de dependências e inclusive o Maven.



Na Aula 13 é apresentado o Projeto MapReduce com Java, objeto do Trabalho da Disciplina. As atividades propostas foram realizadas. Entretanto a configuração do JDK apresentou erro. A JDK instalada foi a versão 8, a que estava disponível. Outro erro ocorrido durante os processos foi a obtenção do arquivo 1901.gz., indicado no material da disciplina, mas que eu não consegui baixar. A partir da Aula 14, não foi possível executar as tarefas devido a falta do arquivo 1901.gz e os erros de configurações.

Hoje dia 19/08/2020, os fóruns das Disciplinas estão fechados e não tenho como tirar duvidas ou pegar orientações com o professor sobre o conteúdo da disciplina.

**Outras Atividades Extra Classe Pesquisas na Internet**

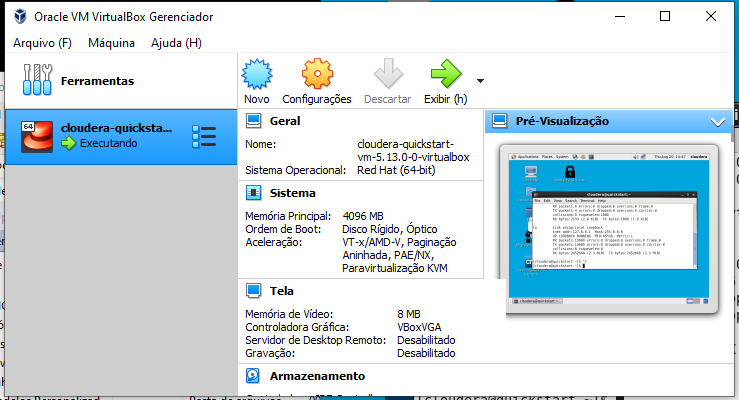
Considerando que não foi possível desenvolver o trabalho com base no material da disciplina, a alternativa foi realizar pesquisas na internet de conteúdos, tutoriais, vídeo aulas e outros materiais complementares de estudo e tentar desenvolver o Projeto Mapreduce Java. Ressaltando ainda a proximidade do fim do prazo de entrega dos trabalhos e a necessidade de desenvolver, ainda, outro Trabalho da Disciplina “Prática e Laboratório”.

Em 20/08/2020, para desenvolver algum Trabalho da Disciplina iniciei pesquisa de Tutorias sobre Projetos Hadoop, Map Reduce, Pig, Yarn, Big Data em Geral e encontrei alguns tutoriais mais atualizados, que indicaram outras ferramentas. A seguir será apresentado 3 experiências com base em cursos e tutorias encontrados

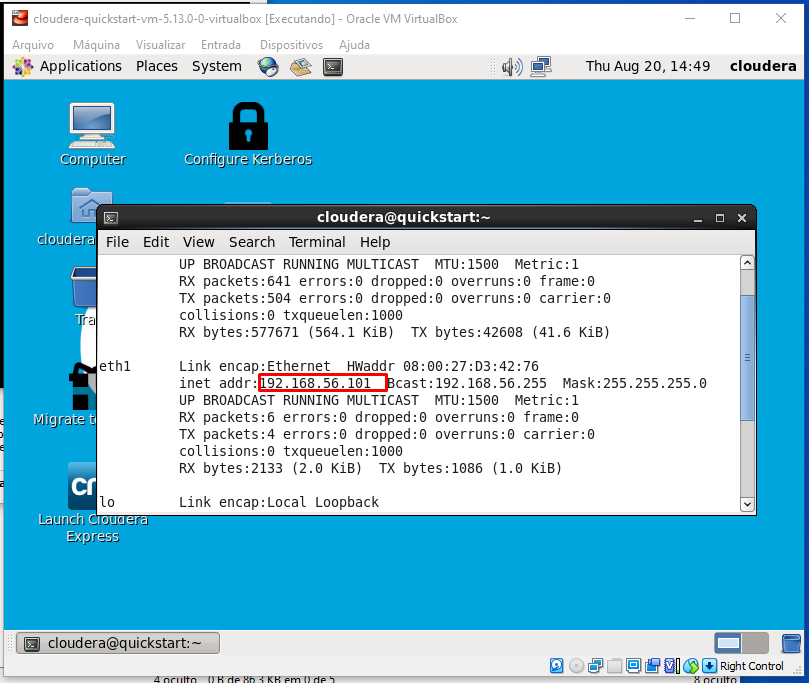
**Experiência 1 – Wordcount MapReduce**

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.

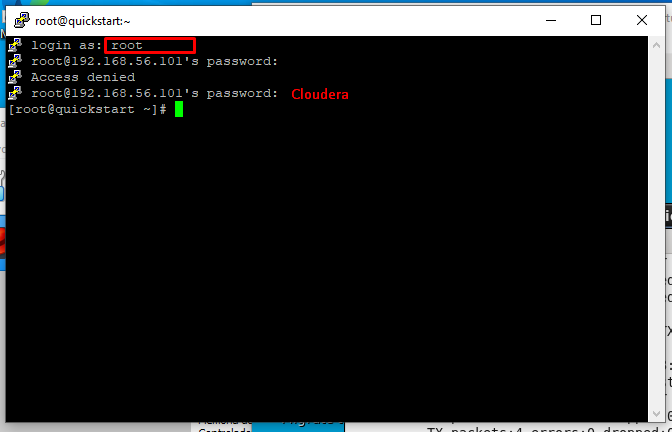
Conforme demonstrado abaixo todas as instalações e configurações indicadas foram realizadas com sucesso e os softwares estão funcionando.



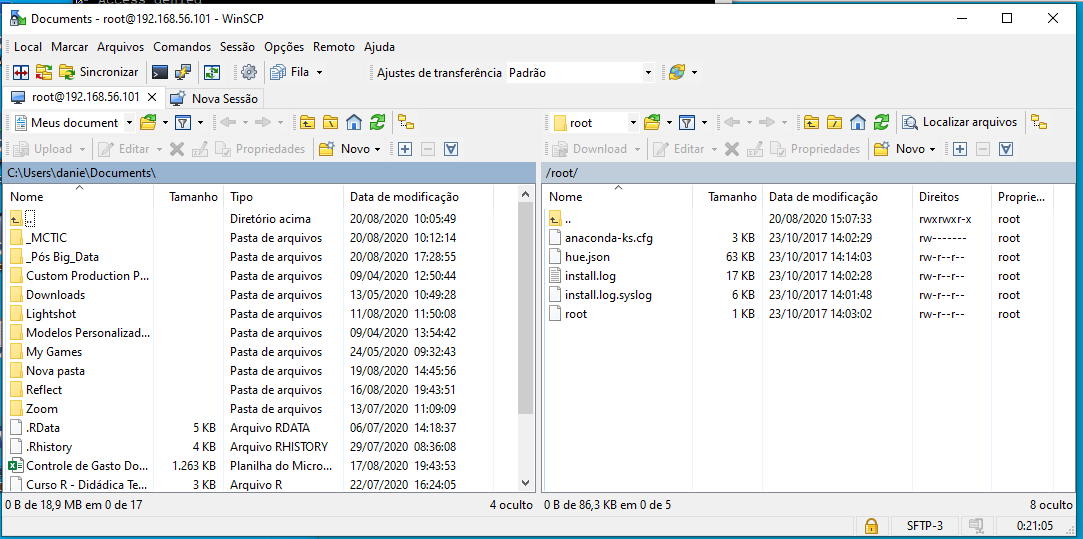
**Oracle VM VirtualBox** – Funcionando



**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/**

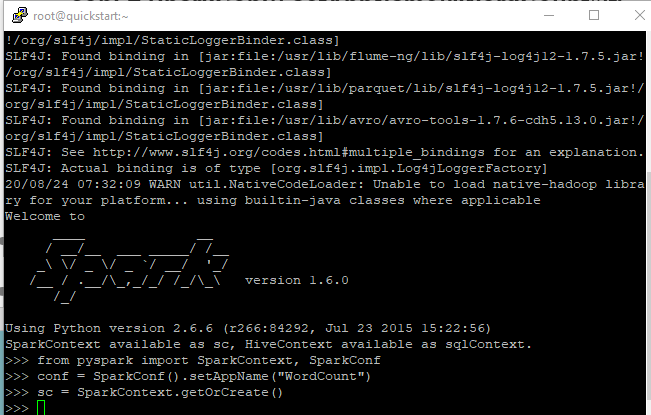
**contentRDD=sc.textFile(“user/estacio/textos/Shakespeare.txt”)**

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():**

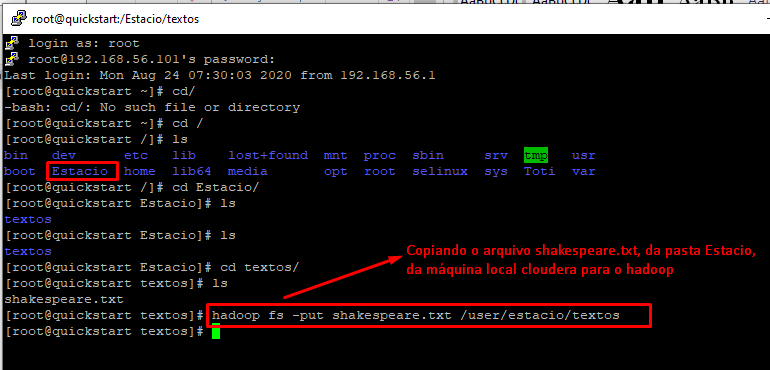
**Print(word)**

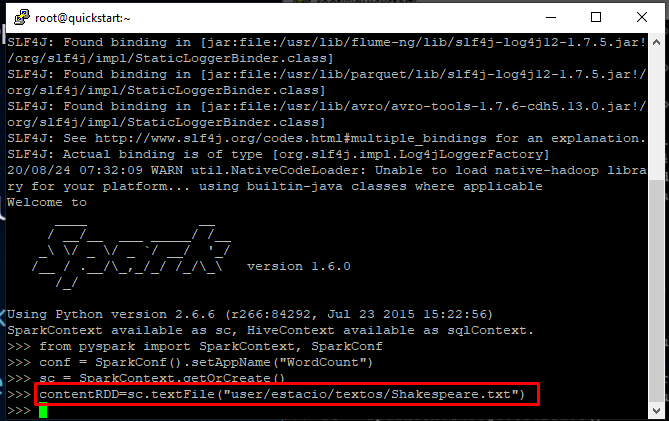
1. Salvar os dados finais no HDFS: **Wordcount.sveAsTextFile(“/user/toti/textos/Wordcont”)**

**Passos 1 a 3**

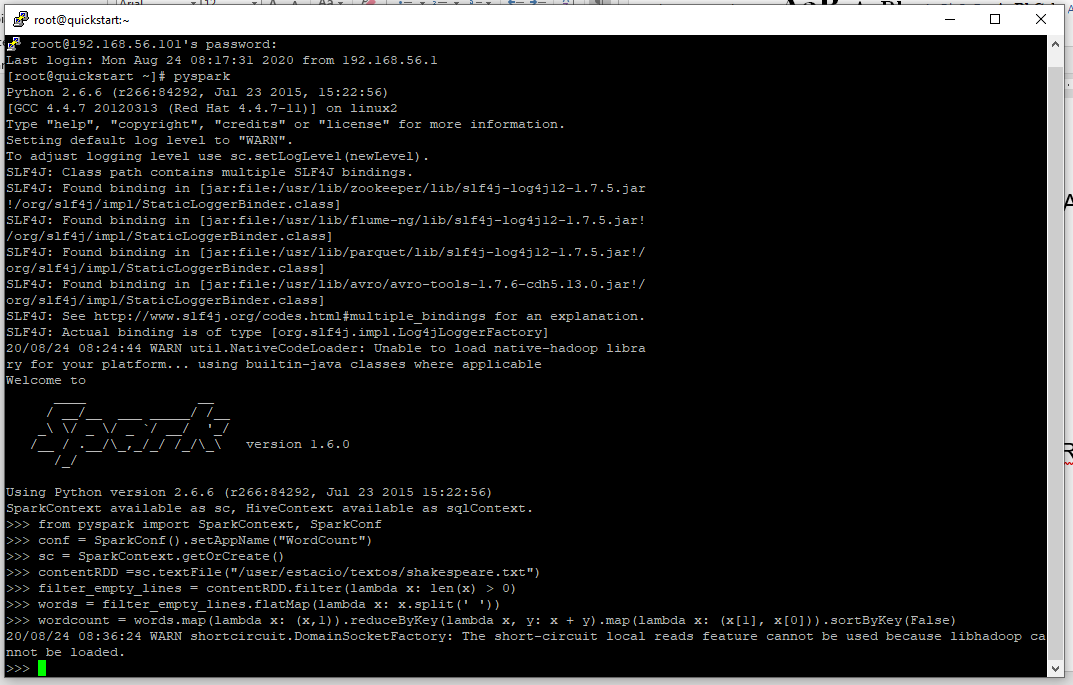


**Passo 4**

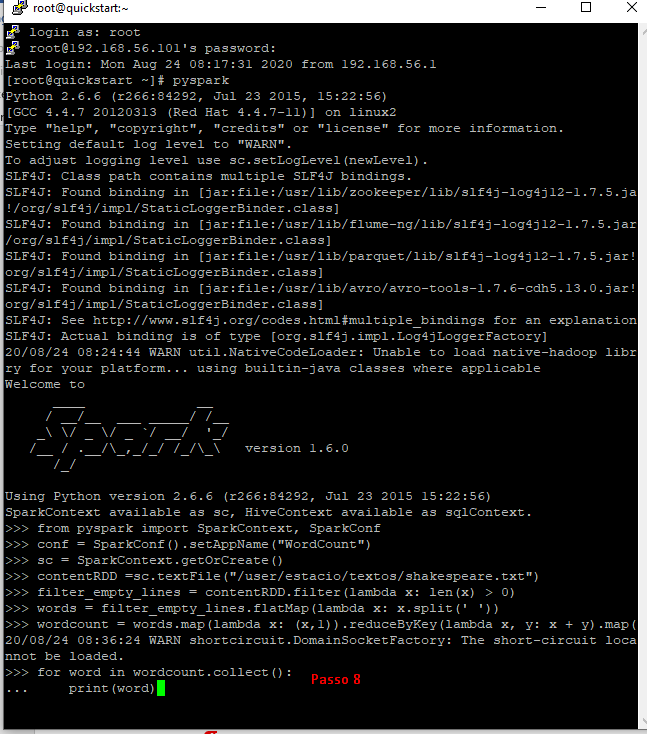




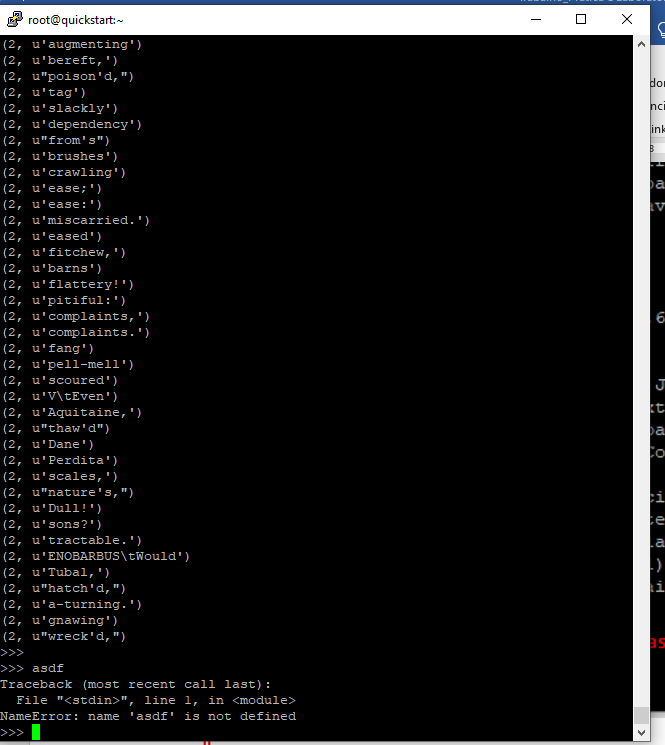
**Passo 5, 6 e 7**



**Passo 8**

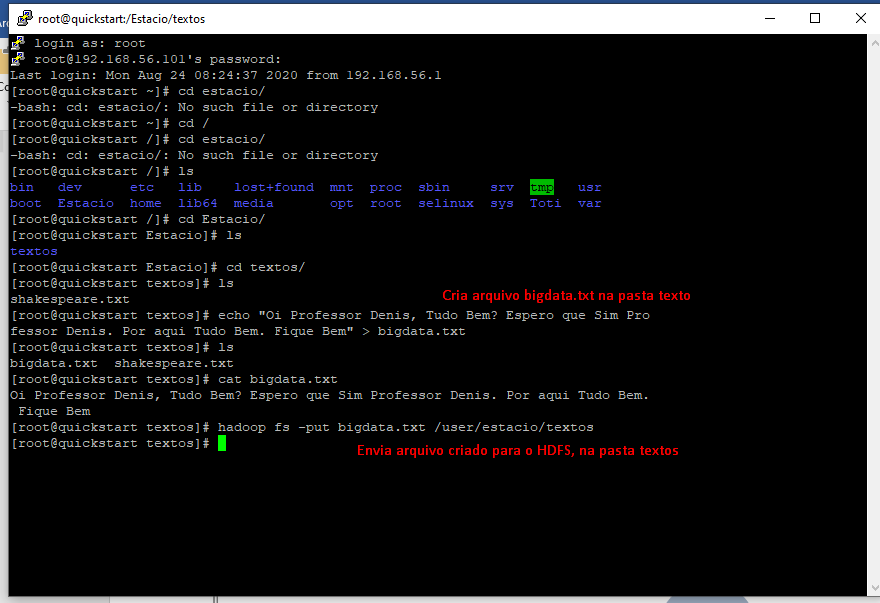


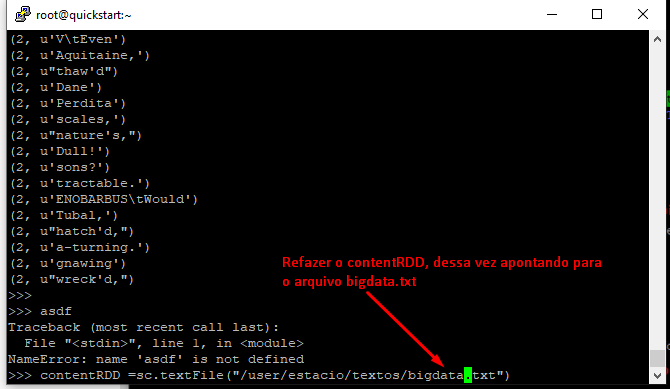
**Resultado da contagem:**

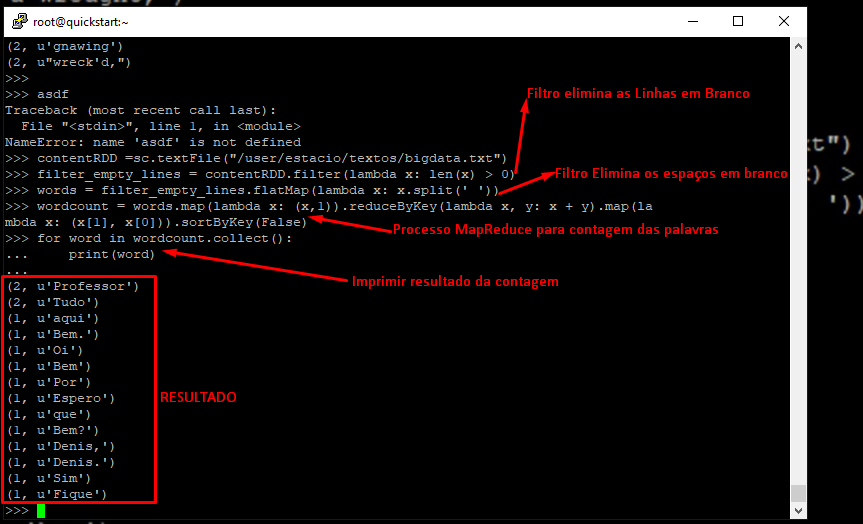


**Outro Exemplo - Wordcount**

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.





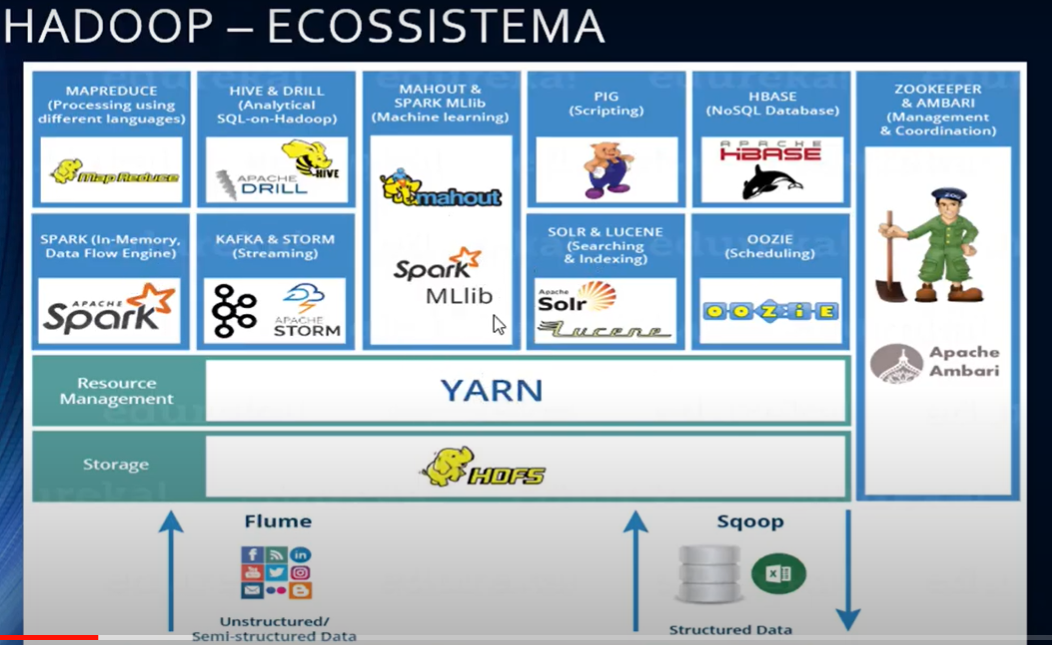


**Experiência 2 – Conexão Flume + Wordcount**

**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 do Flume

*# 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*

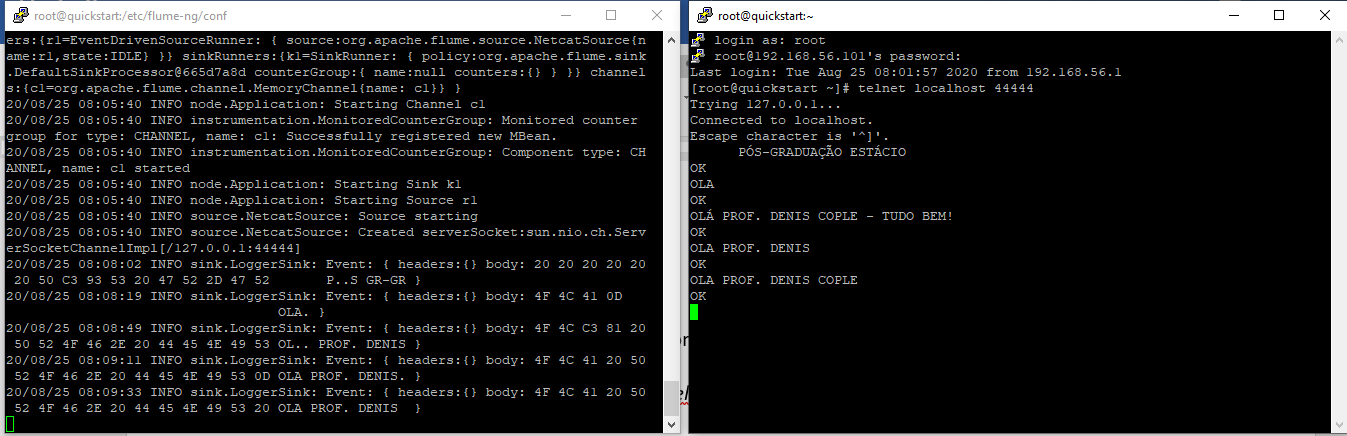
*a1.sources.r1.channels = c1*

*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:

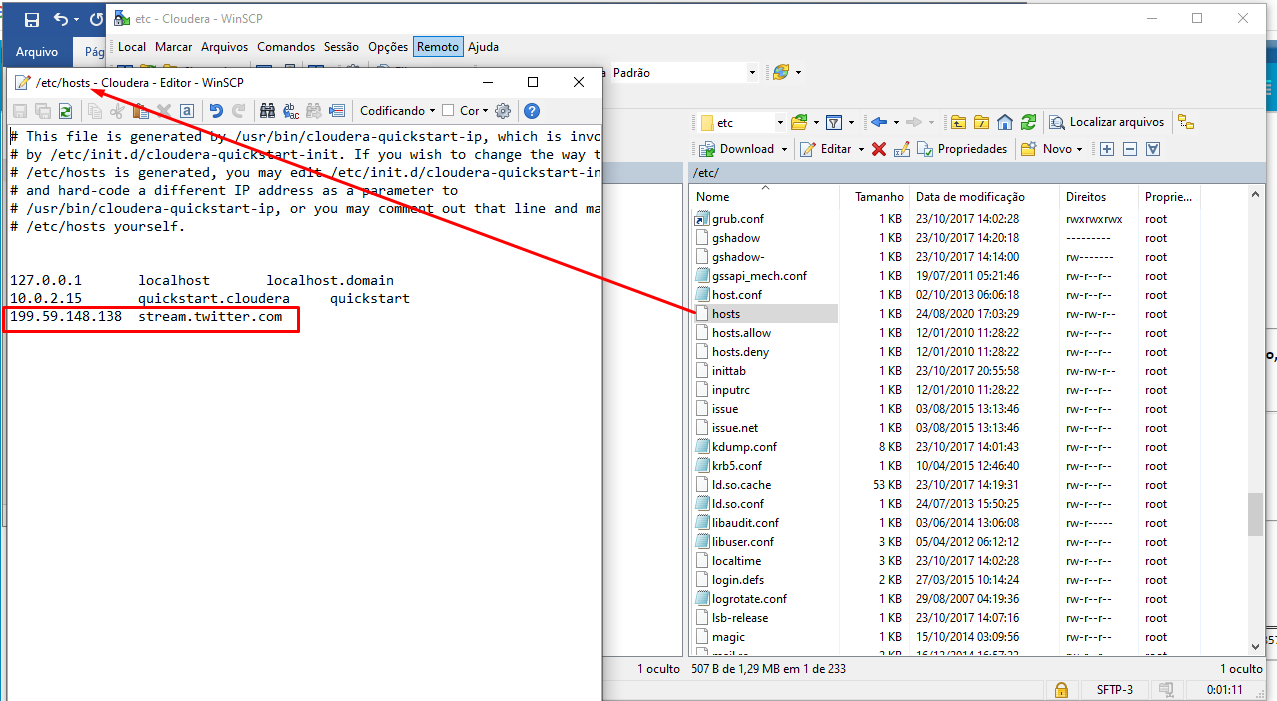
*API key: 6jfpaDKHiwZMXB2auX5lHOVZj*

*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 Quickstart 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*

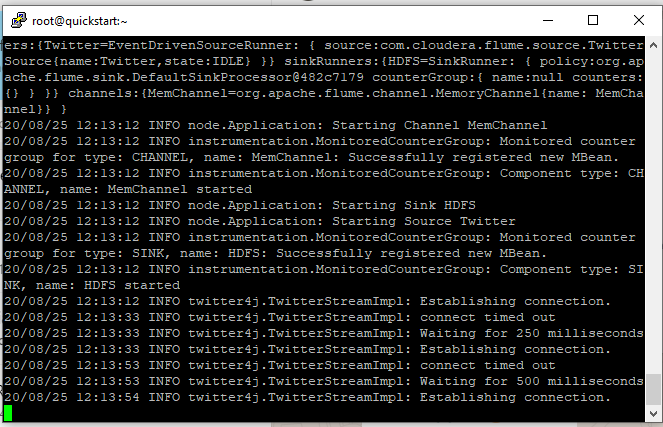
*TwitterAgent.sources.Twitter.channels = MemChannel*

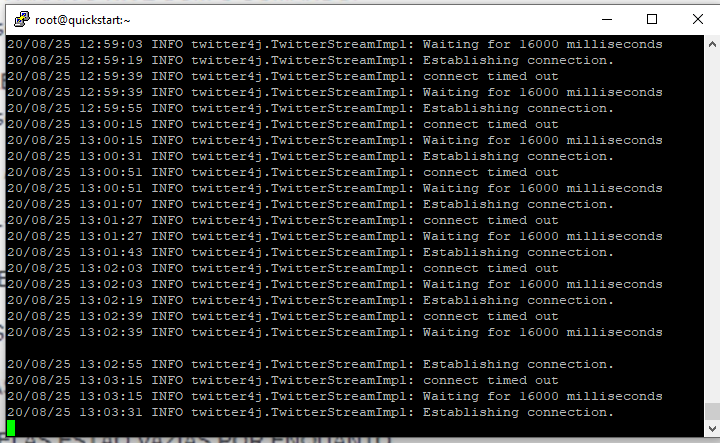
*TwitterAgent.sinks.HDFS.channel = MemChannel*

*#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 ocorreu alguma, talvez capacidade da minha Rede, é não consegue estabilizar a conexão.

Refeito algumas vezes sem sucesso, foi finalizado o exercício 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.

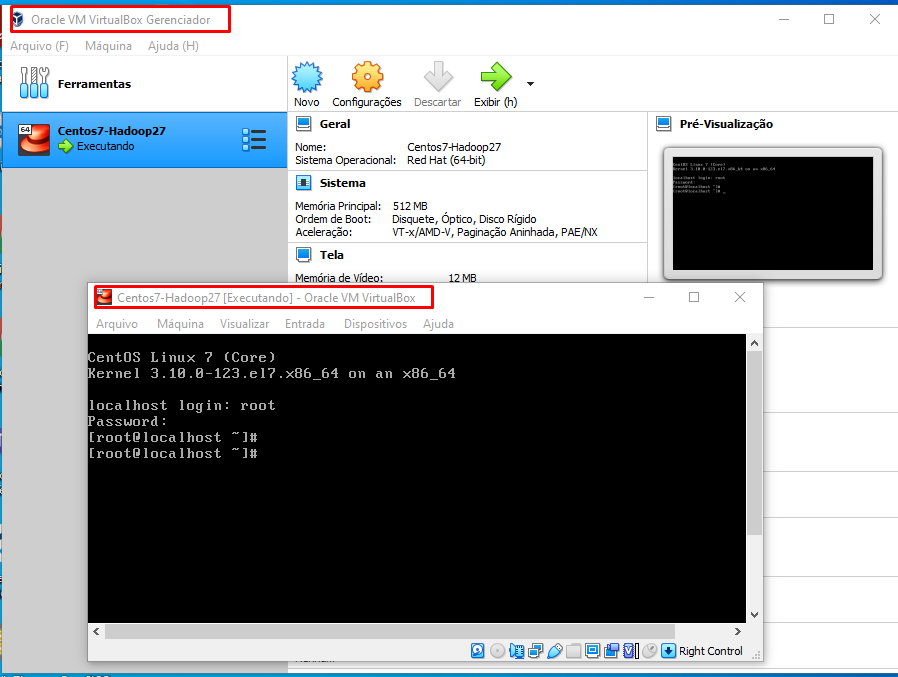
Após refazer as instalações da Vmware e hortonworks Sandbox, e seguir as orientações do material da disciplina, novamente ocorreu erro na interface gráfica e no acesso SSH não foi possível conectar o usuário Root com a senha “**cloudera” .**

**Experiência 3**

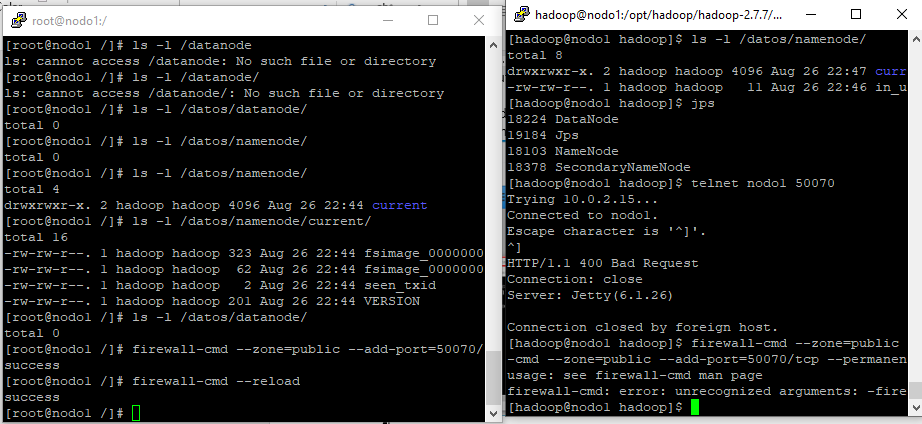
Na busca do conhecimento e aprendizado do conteúdo da disciplina que tem como foco o Ecossistema HADOOP, assisti mais alguns vídeos sobre o assunto, a maioria fala muito das teorias e conceitos envolvidos, mas poucos tem a abordagem prática. Em determinado momento de estudos encontrei a Playlist “Big 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. A seguir é apresentado o passo a passo do exercício realizado 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.



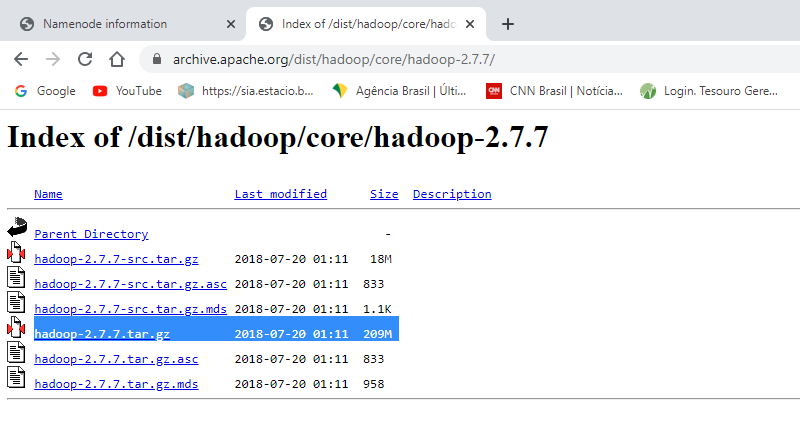
Em seguida foram realizadas as seguintes operações, por ordem:

1. - Configuração do Cliente SSH com a inclusão de uma nova Porta para máquina localhost
2. - Instalação de pacotes hadoop com o comando yum (net-tools; wget; telnet)
3. - Criação de usuário “hadoop” para receber os dados dos processos hadoop
4. - 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

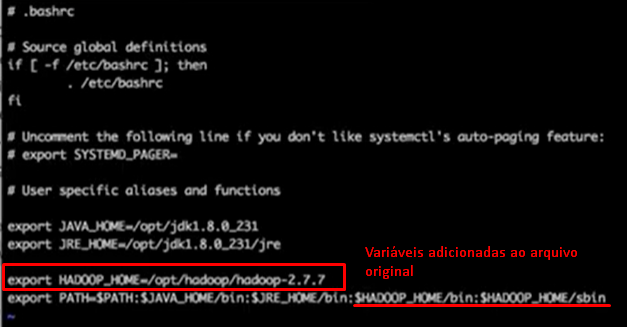




- Em seguida foi feita a descompressão do arquivo através do comando : tar -zxvf hadoop-2.7.7.tar.gz

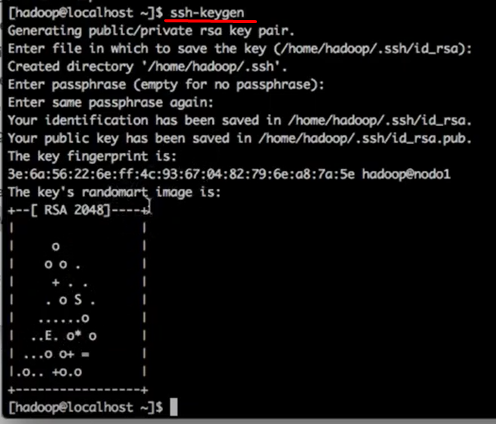
- O próximo passo foi realizar configurações do ambiente Hadoop

- 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*



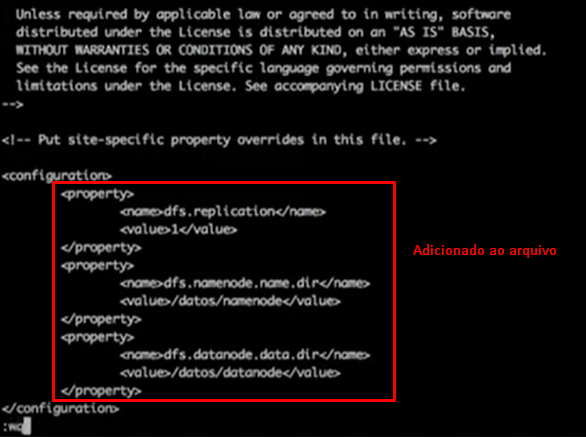
- Desta forma foi estabelecida conexão local

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



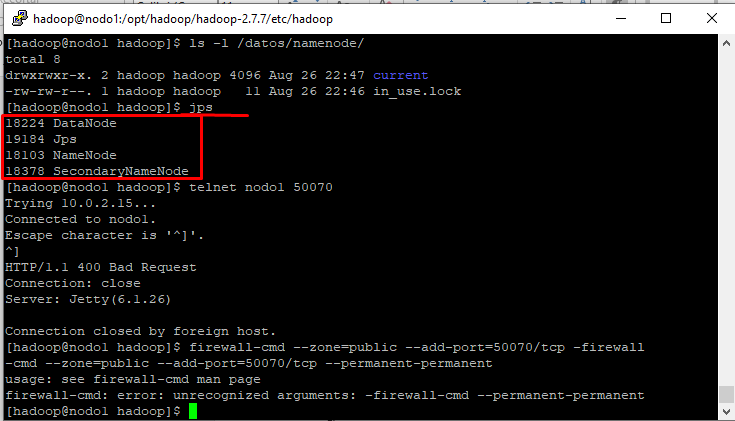
- Foi inserido o comando: firewall-cmd --zone=public --add-port=9000/tcp –permanent , para configurar acesso a nova porta.

- O passo seguinte foi a configuração de alteração do arquivo hdfs-site.xml:

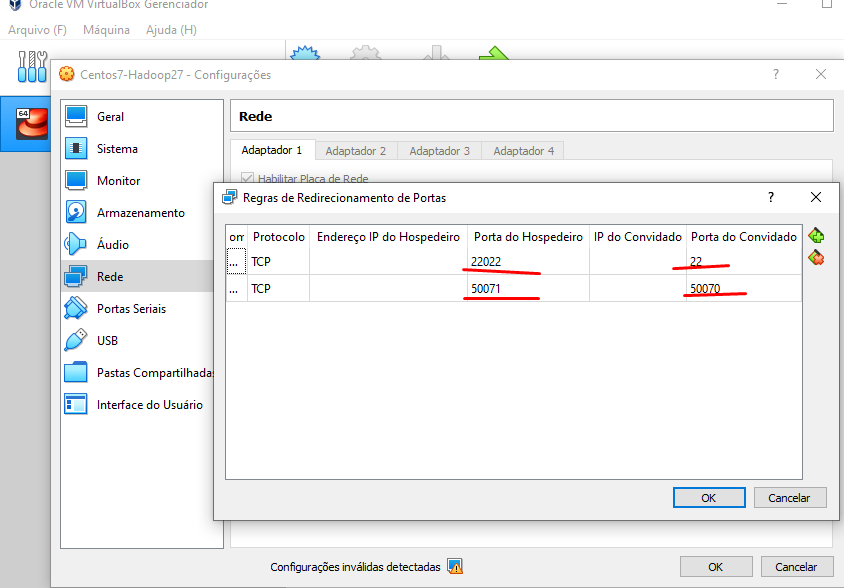


- Em seguida foi realizada a formatação e inicialização do hadoop.

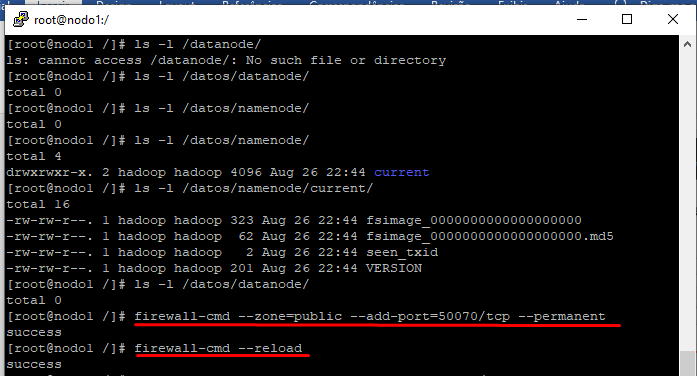
- O Ambiente e de cluster para trabalhos Hadoop Java é verificada conforme comando jps



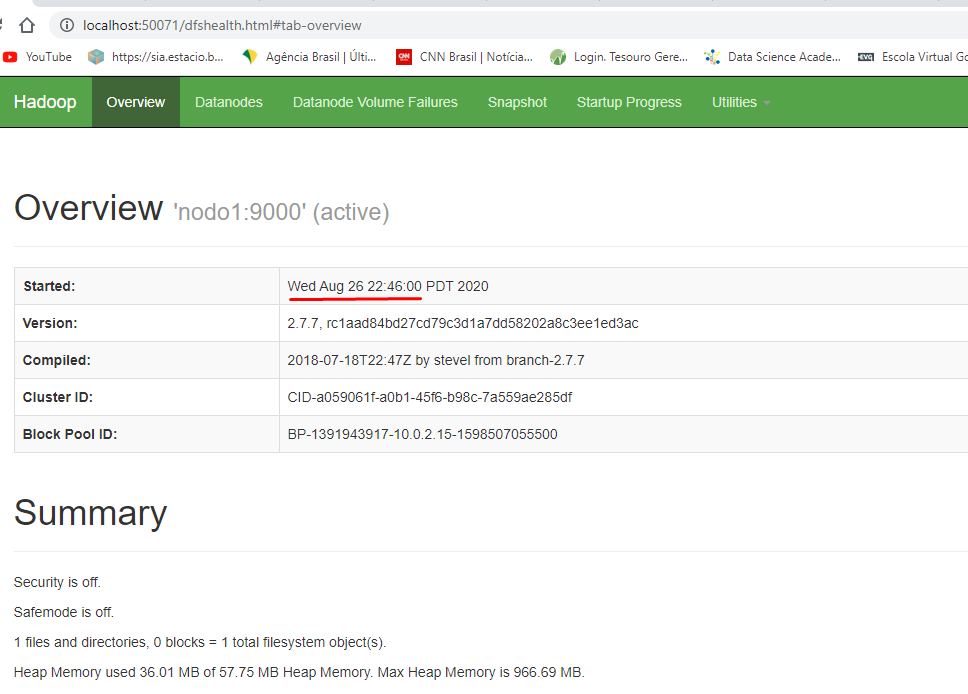
- Adicionalmente, para resolver problema de conexão foi necessário configurar a Porta direto na máquina virtual, conforme abaixo :



- Em seguida foi aberta a Porta 50070 no HADOOP.



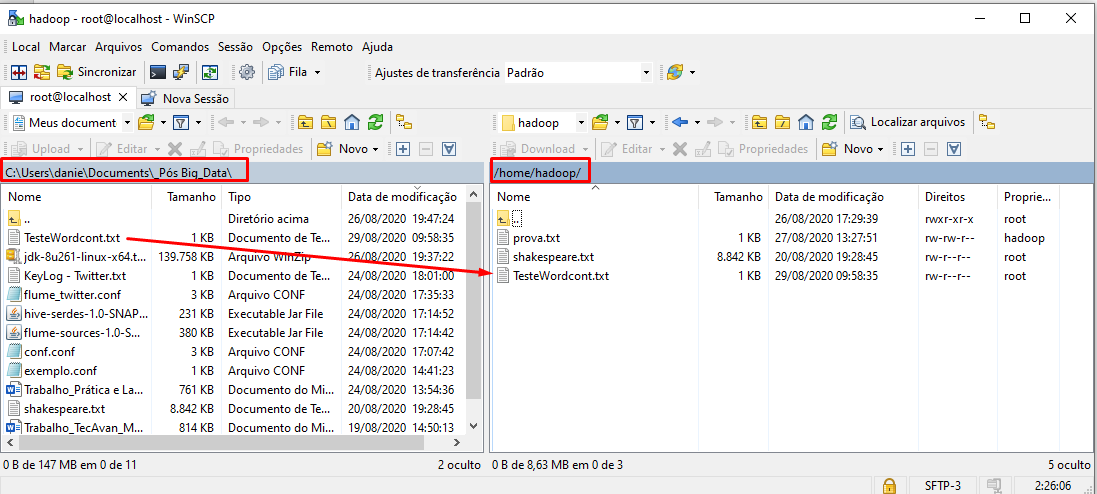
- É possível ser feita a prova da conexão do hadoop com com a nova porta através do browse, digitando o endereço: localhost:50071 :



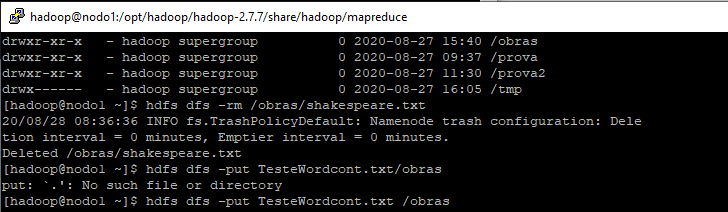
**Executando MapReduce – JAVA – wordcount**

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*

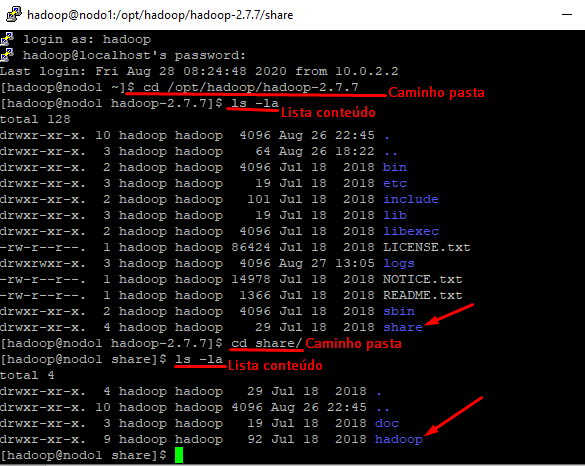
Foi utilizado WinSCP para copiar o arquivo da pasta local para o pasta criada dentro do haddop na máquina virtual CentOS7.

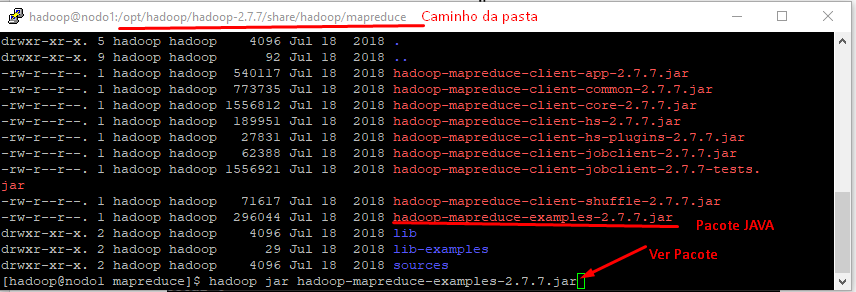


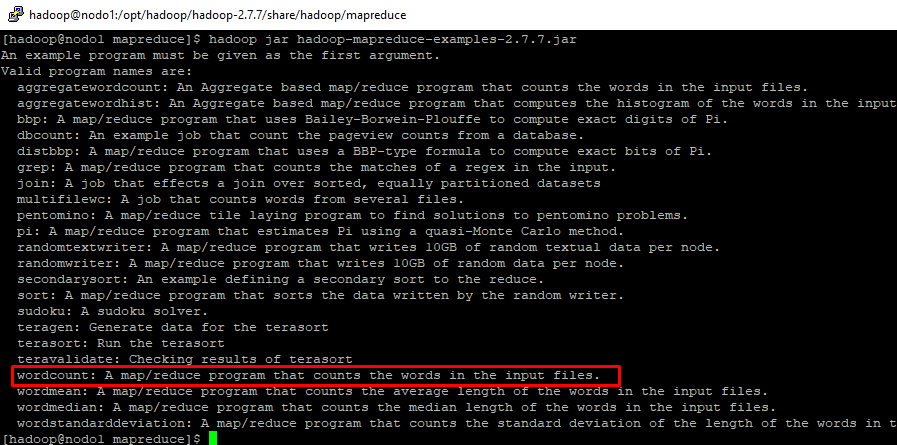
Em seguida foi feita a transferência do arquivo do diretório hadoop para uma pasta (obras) criada em *hdfs dfs* :



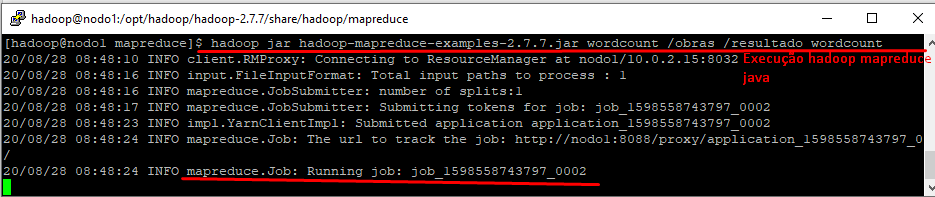
Para visualizar a função Java – wordcount, é possível abrir seu diretório conforma demonstrado abaixo:







**Execução MapReduce Java - WordCount**



# 

**Endereço para obtenção dos dados:**

* <https://github.com/Dmguedes/Trabalho-Big-Data/blob/master/TecAvan_MapReduce>