Integration of HDP with Splunk & Hunk

**Table of Contents**

1 Overview 2

2 HDP Integration 4

2.1 Splunk Enterpise Integration with HDP 4

2.2 Splunk Hunk Integration with HDP 4

3 Steps to setup Integration 4

3.1 Setup HDP 2.3 sandbox 5

3.2 Install Splunk within HDP sandbox 6

3.2.1 Get Splunk 6

3.2.2 Install Splunk 7

3.2.3 Setup Splunk Hadoop Connector 9

3.2.4 Setup Hunk HDFS Provider 12

3.2.5 Setup Hunk Virtual Index 15

3.2.6 Search the data via Hunk 16

4 Revision 18

# **Overview**

This Article provides a simple way to setup HDP (Hortonworks Data Platform) Sandbox and Splunk, so that you get a hands on experience on how this integration works. You will find detailed document on both Hortonworks ([Hortonworks tutorial](http://hortonworks.com/hadoop-tutorial/configure-hortonworks-sandbox-version-2-0-with-hunk-splunk-analytics-for-hadoop/)) and Splunk website ([Hunk Tutorial](http://docs.splunk.com/Documentation/Hunk/latest/Hunktutorial/WelcometotheHunktutorial)) on how you can build these integration. For enterprise grade integration setup, definitely use those documents. This article is more focused on setup a simple Hortonoworks Hadoop sandbox and Splunk installed so that you can play around with it.

Before we talk about integrations, here is a quick overview of each of these components.

**HDP** : Architected, developed, and built completely in the open, Hortonworks Data Platform (HDP) provides an enterprise ready data platform that enables organizations to adopt a Modern Data Architecture.

With YARN as its architectural center it provides a data platform for multi-workload data processing across an array of processing methods – from batch through interactive to real-time, supported by key capabilities required of an enterprise data platform — spanning Governance, Security and Operations.

**Splunk:** Splunk offers products that perform real-time and historical search, as well as reports and statistical analysis. The product can index structured or unstructured textual machine-generated data. Search and analytics operations are specified using Search Processing Language, created for managing machine-generated big data. Its scope includes data searching, filtering, modification, manipulation, insertion, and deletion.

Splunk has two primary products that work with Hadoop which are Splunk Enterprise and Hunk.

(Source: https://en.wikipedia.org/wiki/Splunk)

Splunk Enterprise – It provides you operational analytics, around data related to technology infrastructure, security systems and business applications. Basically you ingest you operational data into splunk and it provides you easy way to use GUI using which you can perform operational reporting, dashboarding, analytics, alerting, search etc..

Hunk – Hunk is a Splunk analytics products for data in Hadoop or NoSQL system. Its again a very easy to use GUI tool that can hook to your hadoop or NoSQL system and use the data from those system for performing search, data exploration, reporting, Dashboard creation etc..

# **HDP Integration**

## Splunk Enterpise Integration with HDP

Splunk Enterprise can integrate with HDP via a Hadoop connector that is available as a App within Splunk. Once you have has the hadoop connector setup within splunk for HDP (Hortonworks Data Platform), you can perform three operations

1. Ingest data from Hadoop to Splunk
2. Export data from Splunk to Hadoop
3. Browse the data in hadoop via Splunk GUI interface



## Splunk Hunk Integration with HDP

Hunk can integrate with HDP via a Virtual index which is basically a setting within Hunk that tells where you hadoop cluster is and what directory within HDFS you are trying to access.

In the demo setup I will talk about the steps involved to set this up.

## Splunk App for HadoopOps

Splunk app for HadoopOps can be used to monitor your hadoop cluster for various operational metrics. This might be helpful, if you are planning to have all your enterprise operational intelligence in splunk, including for your hadoop cluster. However if that’s not the case, then HDP (Hortonworks Data Platform) itself has a very robust and powerful cluster management, monitoring and provisioning tool called Ambari.

Here is a link from splunk that provides more details on [Splunk App for HadoopOps](https://splunkbase.splunk.com/app/1173/#/documentation)

## Shuttl

Shuttl is a open source product that can be used for big data archiving and bulk data movement for Splunk. Supports HDFS, NFS, Amazon S3, and Amazon Glacier. However when I checked the github release notes it says “Shuttl development has stalled. There's no known developer working on this project.”.

In any case, here is a document from splunk that talks about how to set this up.

[Shuttl setup for splunk](https://splunkbase.splunk.com/app/1195/)

In rest of this article, I am going to talk about Splunk hadoop connect and Splunk Hunk only, as they are the most prominent components that are used with Hadoop.

# **Steps to setup Integration**

Here are the high level steps in setup this demo integration

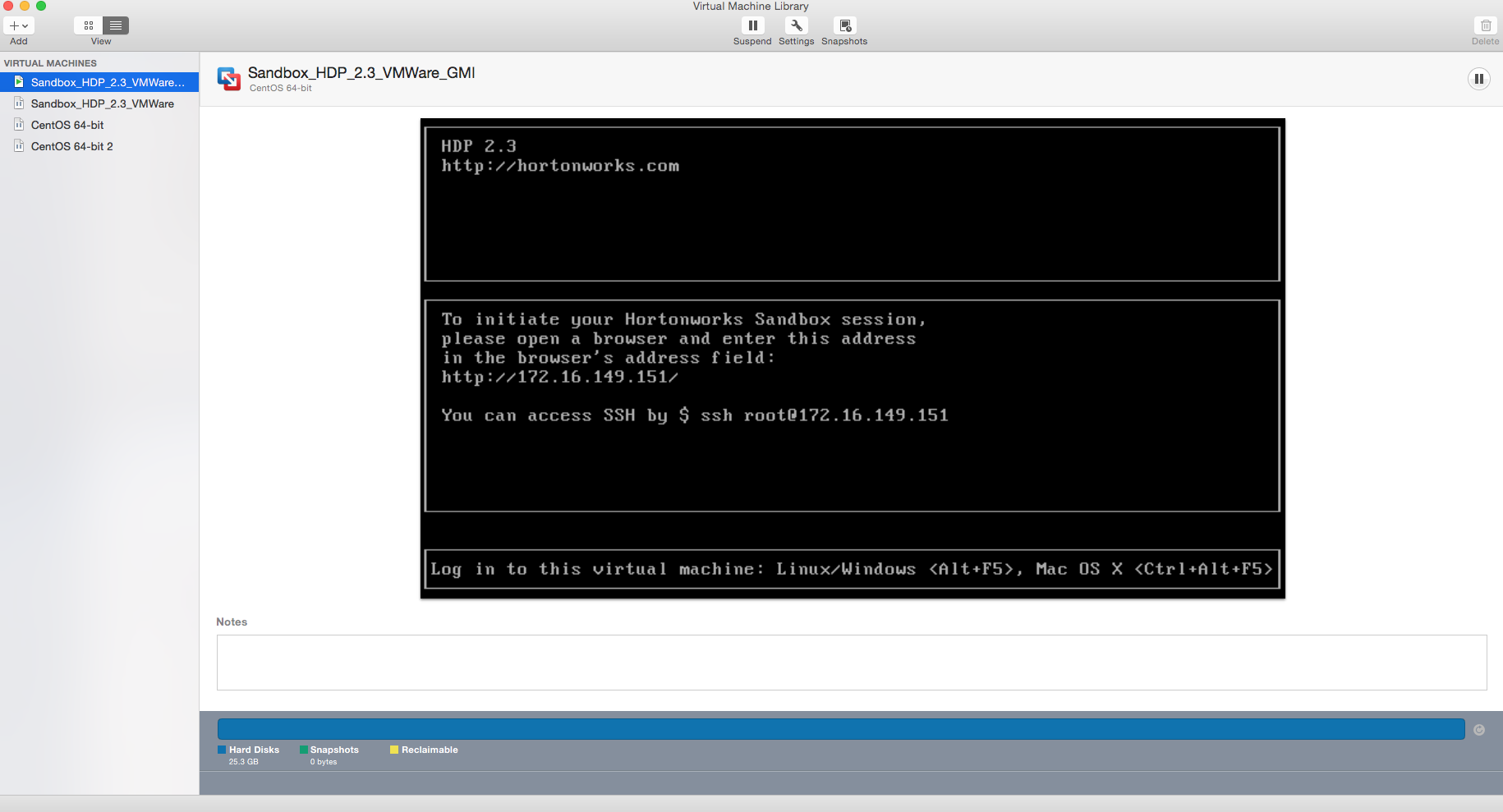
1. Download HDP sandbox and open it via VMWare or Oracle Virtualbox
2. Download Splunk’s Hunk trial software and install it on HDP sandbox
3. Setup a Hadoop Connector for Splunk integration
4. Setup a Virtual Index for Hunk integration

## Setup HDP 2.3 sandbox

Here are the instruction on how you can download HDP sandbox virtual image and open it via VMWare Player or Oracle virtualbox .

<http://hortonworks.com/products/hortonworks-sandbox/#install>

once you are done with the steps mentioned you will have a screen which will look something like this



## Install Splunk within HDP sandbox

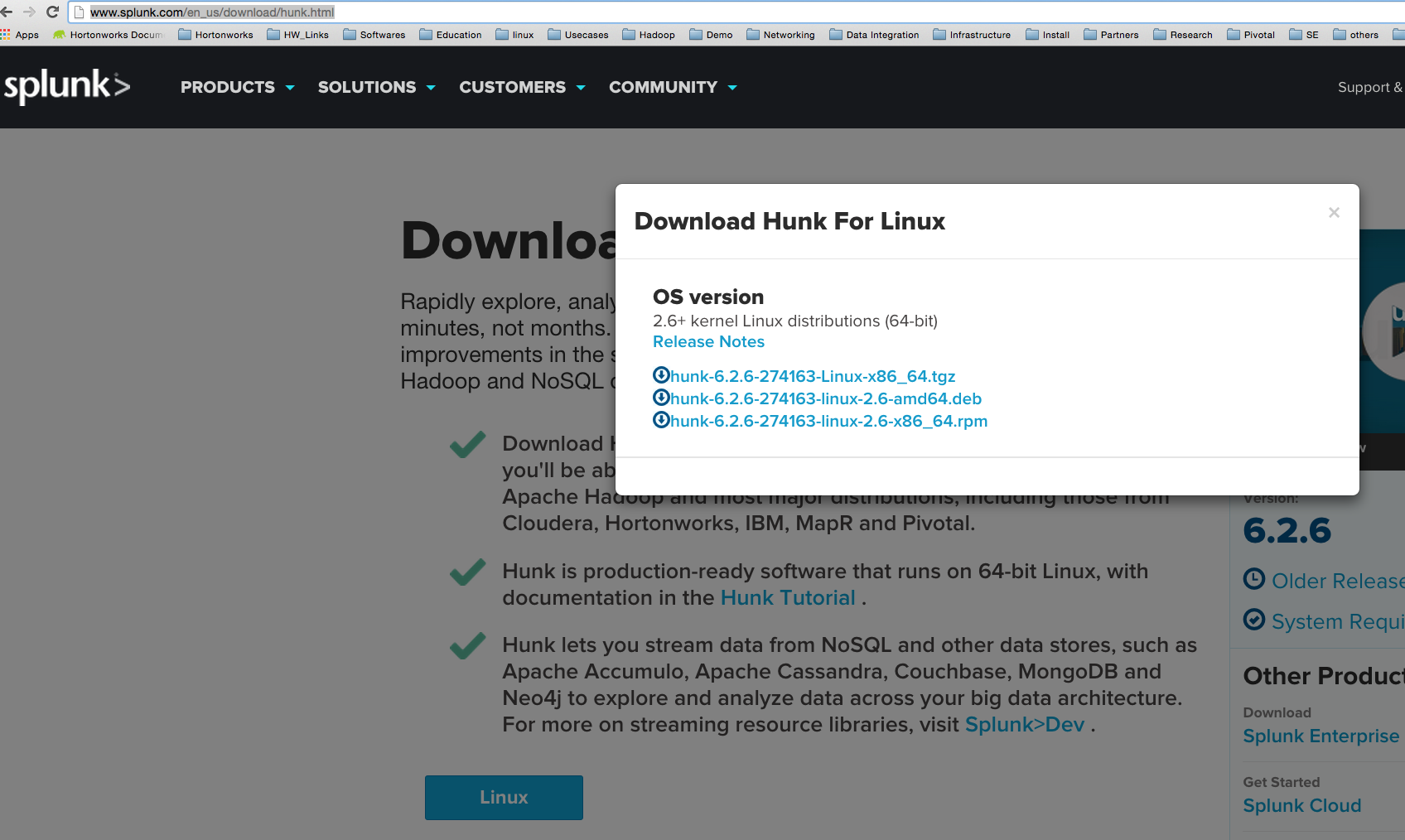
### Get Splunk

Login to Splunk website and you should be able to get hunk trail version for 60 days.

Here is a link to download it

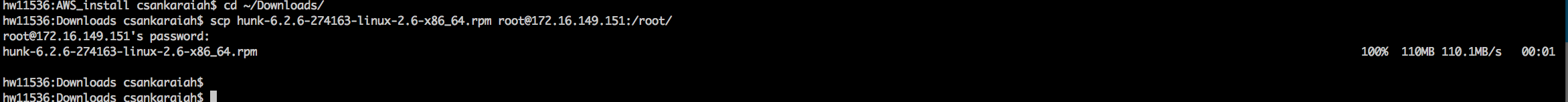
<http://www.splunk.com/en_us/download/hunk.html>

Since I am going to install it in HDP sandbox I am going to use the rpm



Once you have the rpm you can move that your sandbox. Your sandbox IP address is something that you should be able to see in you Virtual box or VMWare. You can use a ftp tool or if you have mac you can use scp command to get the file to sandbox

scp hunk-6.2.6-274163-linux-2.6-x86\_64.rpm [root@172.16.149.151:/root/](mailto:root@172.16.149.151:/root/)



### Install Splunk

Now since you can login to you sandbox via putty (for windows OS) or Terminal (For Mac OS). Here is the command to do that and when asked for password use ‘hadoop’

ssh root@172.16.149.151

once you are logged in you can create a new directory called hunk and move this rpm over to that directory.

mkdir ~/hunk

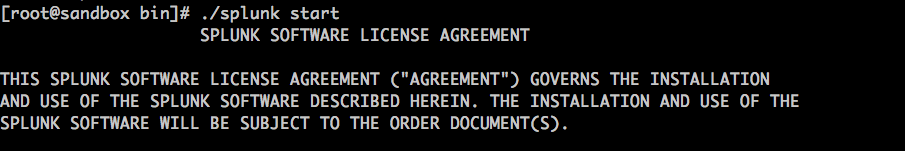
mv ~/ hunk-6.2.6-274163-linux-2.6-x86\_64.rpm ~/hunk/

Now run the rpm to install hunk (it will be installed under default path /opt/hunk)

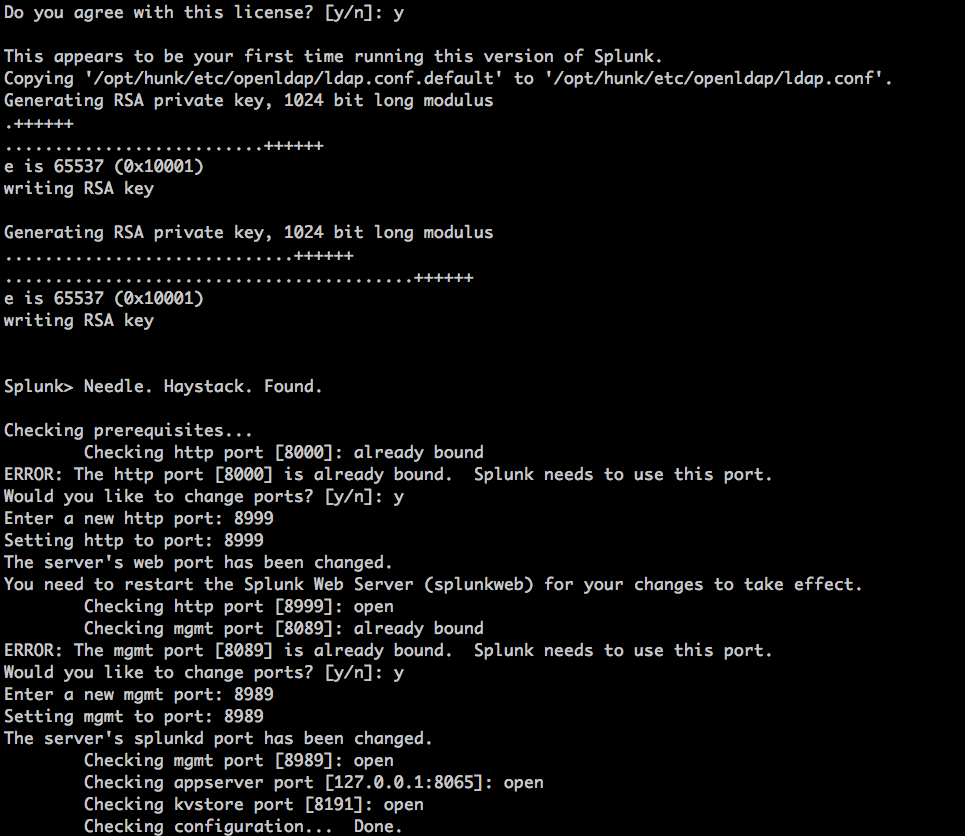
rpm -i hunk-6.2.6-274163-linux-2.6-x86\_64.rpm

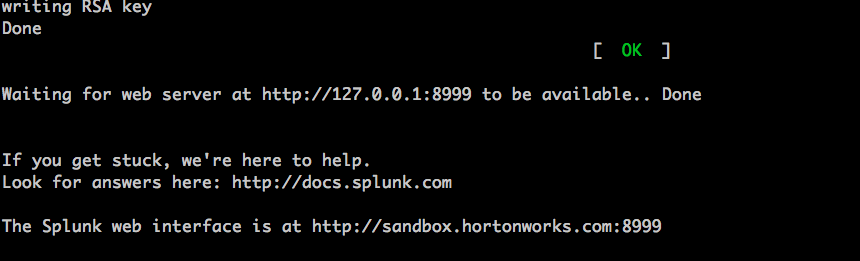
Now go to the directory (/opt/hunk/bin) where splunk is installed and start splunk

./splunk start



You have scroll through the entire page before you can accept the agreement. You will be asked to enter a different port for couple of services that splunk has as some of the existing ports are already used by HDP. I specified port 8999 for spluk and that’s the port that I will be using to connect to splunk.

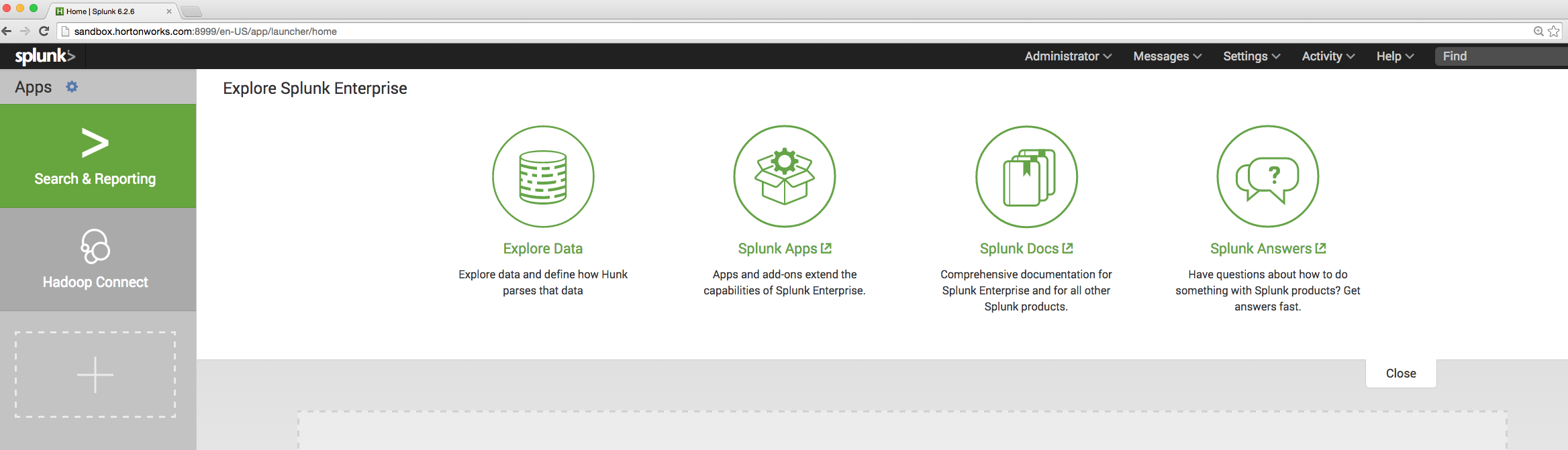




### Setup Splunk Hadoop Connector

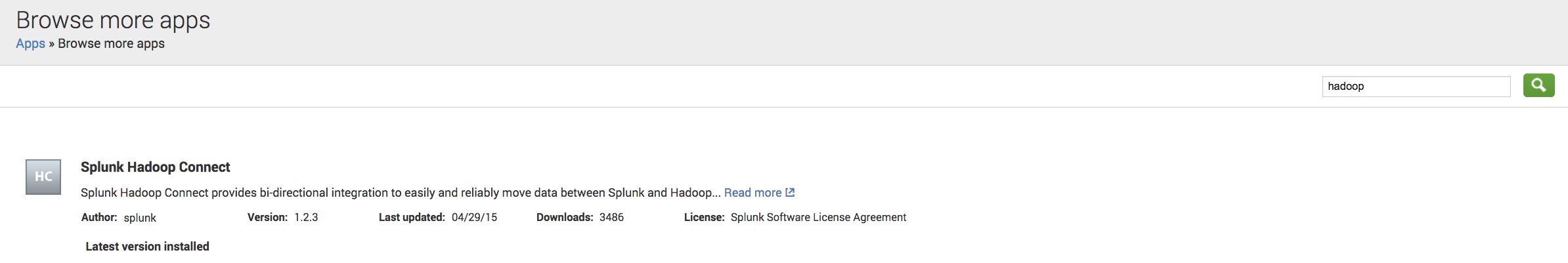
Open up splunk using a browser using the URL IP (or hostname) and port that you have specific. For me it was

<http://sandbox.hortonworks.com:8999/en-US/app/launcher/home>



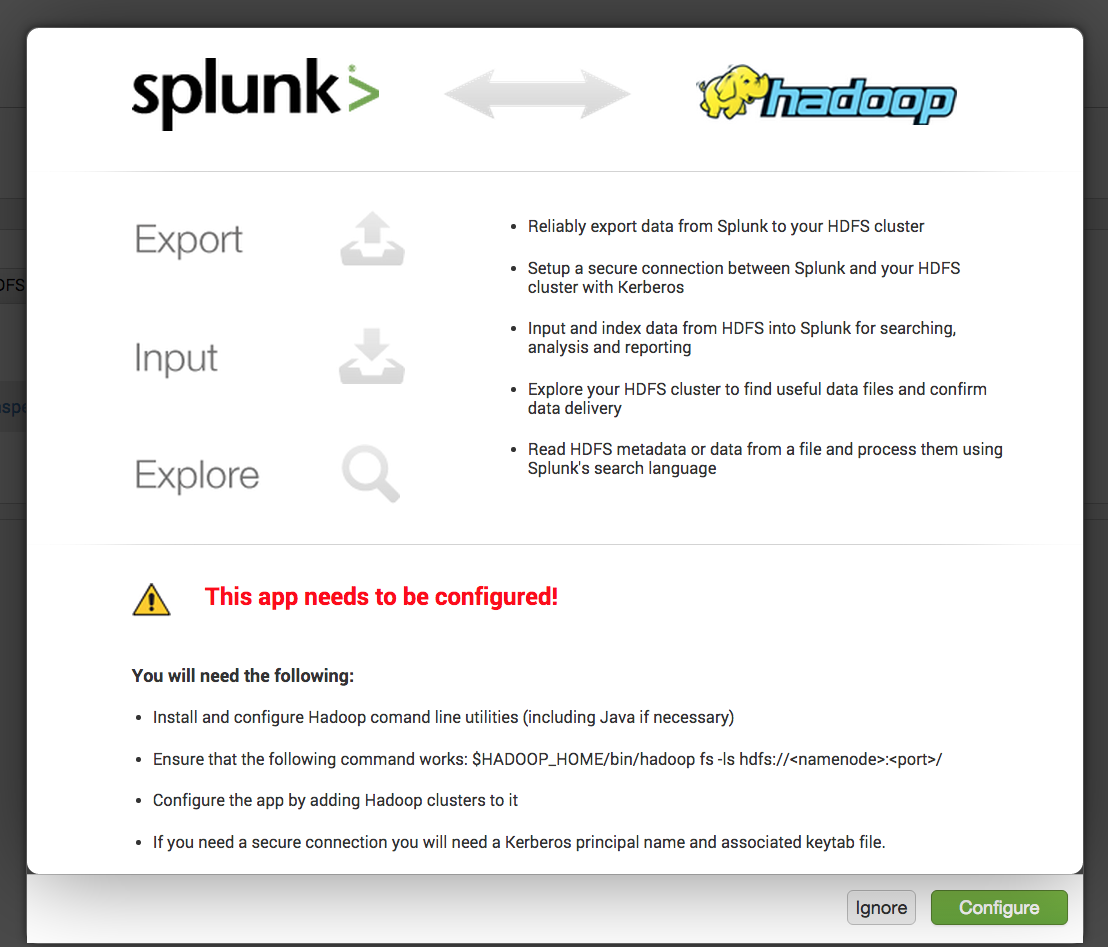
For me it shows Hadoop Connect at the left had explorer under app because I have already added it. You won’t see that the first time you login.

You can get the hadoop connect by clicking on the + sign under the left side explorer (which is basically to add new apps) and then search for hadoop.



You will see a install option over there, click on that. This will also restart splunk and add “Hadoop Connect” to the list of apps.

Once you click on “hadoop Connect”, it will ask for you to configure a cluster



use the configuration option to connect to sandbox.

Here are the three values you need, to connect to the hadoop cluster

HDFS URI: you can get that from HDFS configuration within ambari under parameter fs.defaultFS

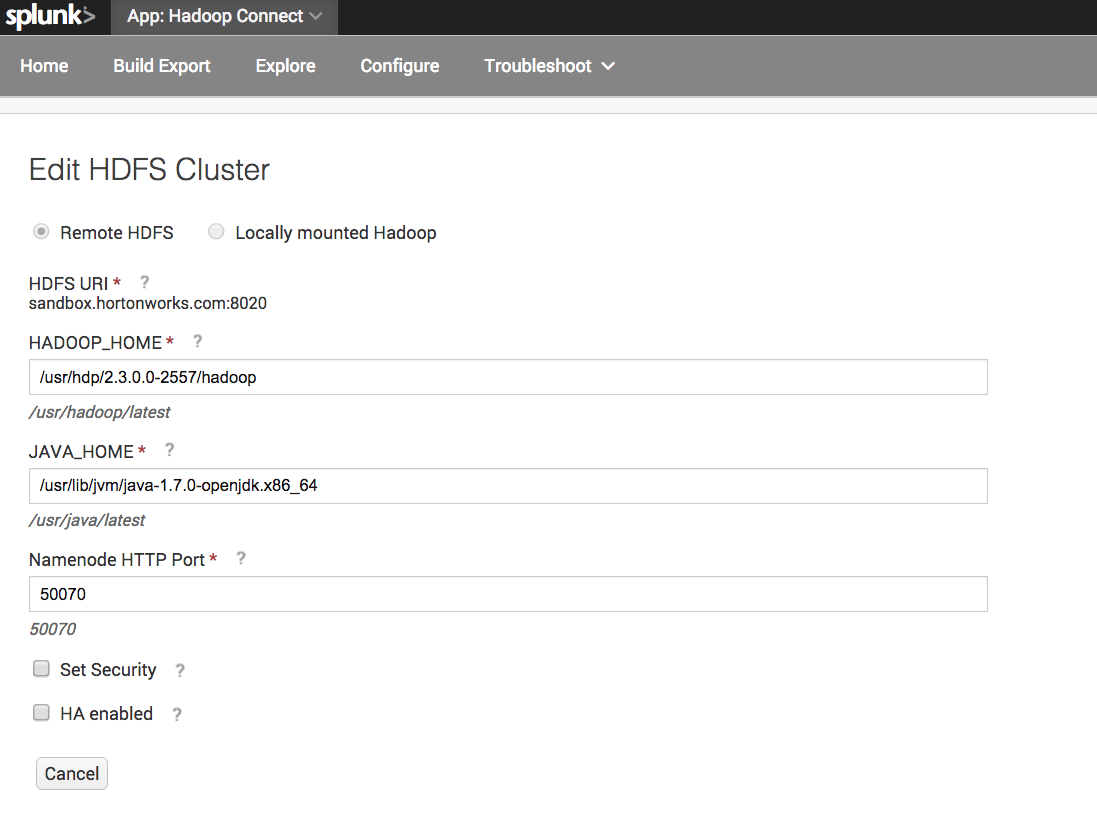
sandbox.hortonworks.com:8020

HADOOP\_HOME: This is basically where you hadoop binaries are available

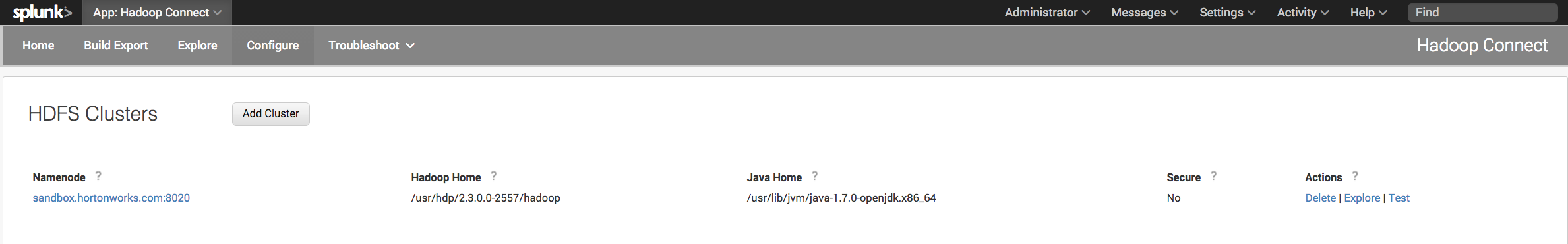
/usr/hdp/2.3.0.0-2557/hadoop

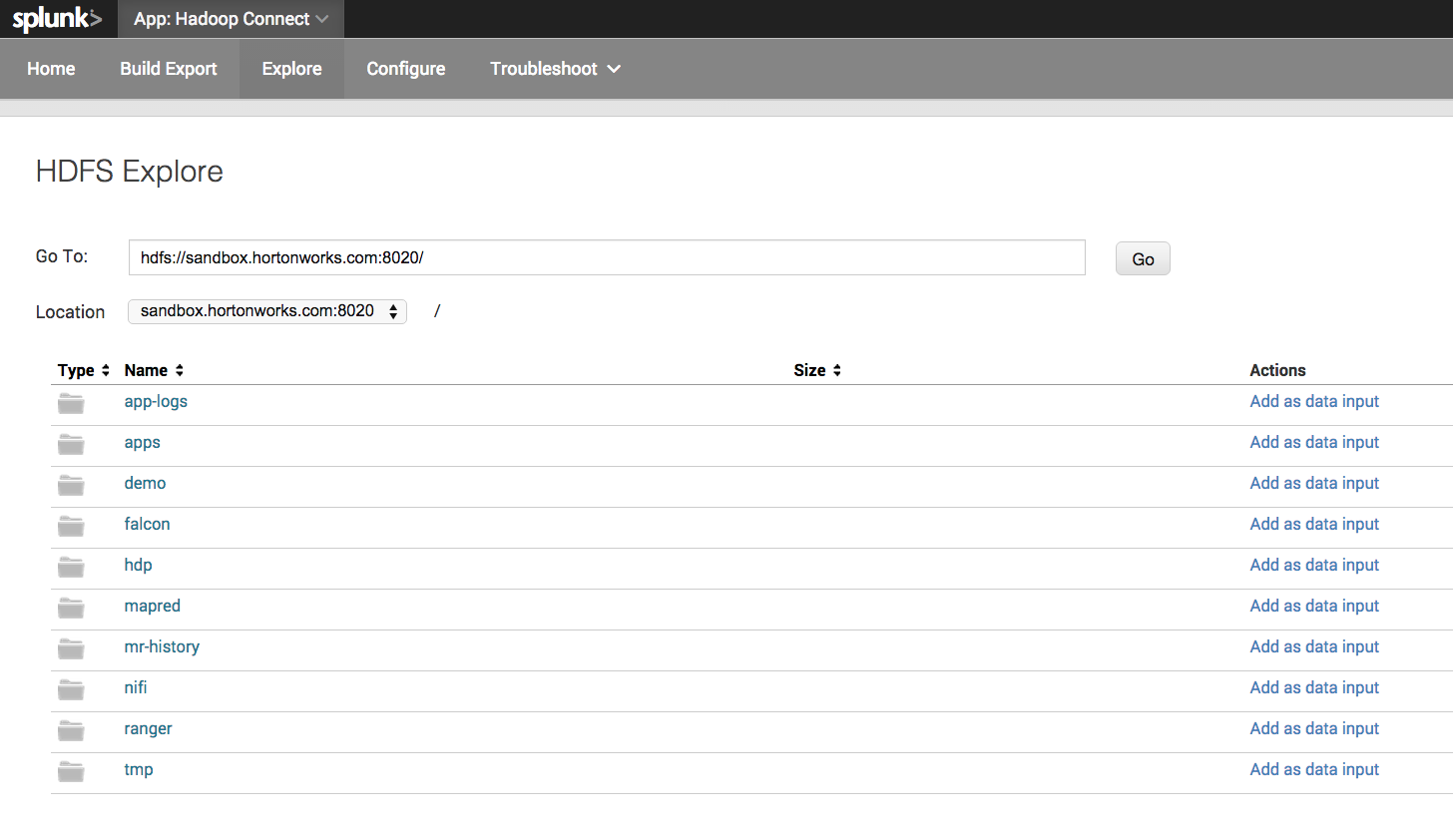
JAVA\_HOME: You Java home path

/usr/lib/jvm/java-1.7.0-openjdk.x86\_64



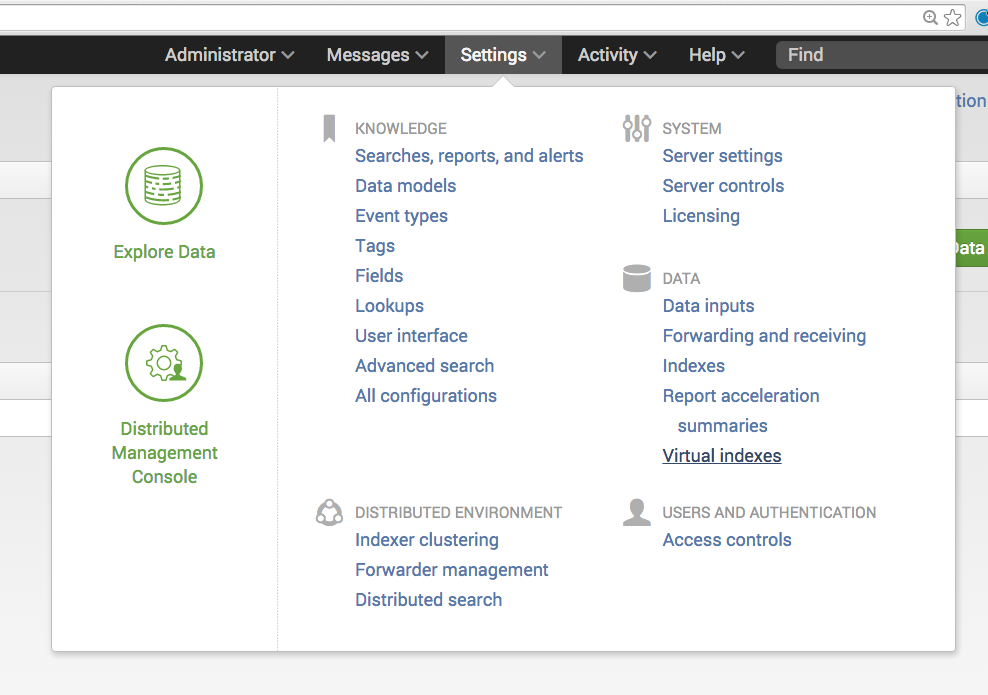
once the connection is setup you should be able to browse the HDFS directories by using the Explore Option under Actions





### Setup Hunk HDFS Provider

Under you home page for Splunk there is a Settings option under which you will see a option to create “Virtual index”. Click on that to configure you HDP connection.



You will be asked to create a new provider under which a virtual Index will be created. In order to create a new provider following are the parameters that you would need to fill

Name : You can give any name for the provider

##### Environment Variables

Java Home: /usr/lib/jvm/java-1.7.0-openjdk.x86\_64

Hadoop Home: /usr/hdp/2.3.0.0-2557/hadoop

##### Hadoop Cluster Information

Hadoop Version: Pick Hadoop 2.x, (YARN)

File System: hdfs://sandbox.hortonworks.com:8020

Resource Scheduler Address: hdfs://sandbox.hortonworks.com:8030

##### Splunk Settings

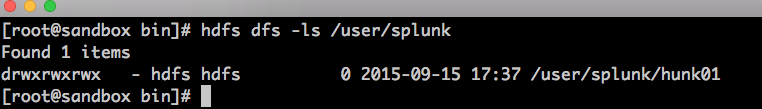
HDFS Working Directory: You have to create a new directory within HDFS which will used by Hunk for its data processing

Login to HDP sandbox and use this command

su hdfs

hdfs dfs -mkdir /user/splunk/hunk01/

hdfs dfs –chmod 777 /user/splunk/hunk01/



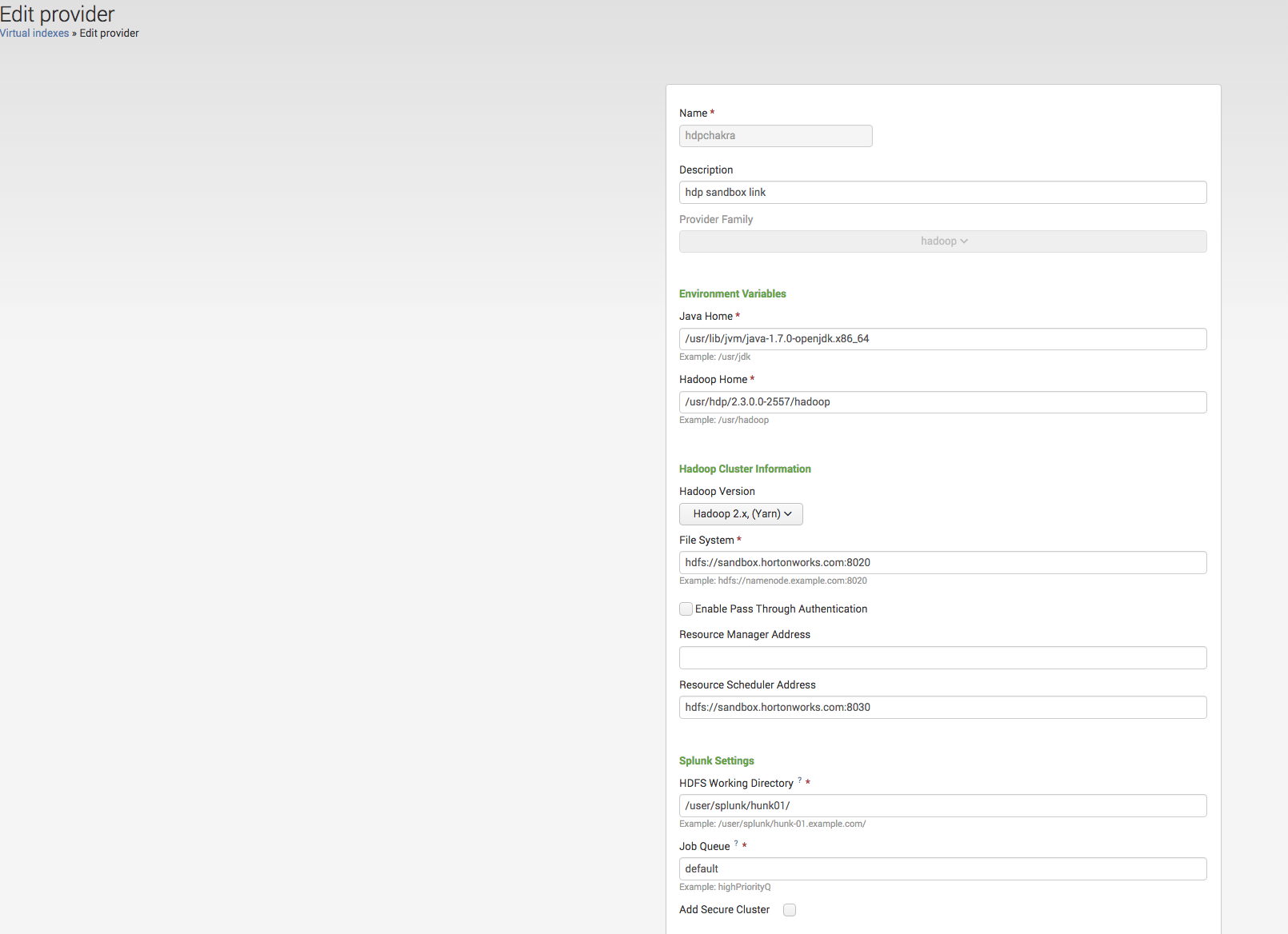
##### Additional Setting

There are three more setting that need to be added which you can add using “New Setting” option at the bottom

vix.mapreduce.framework.name: yarn

vix.yarn.resourcemanager.address: hdfs://sandbox.hortonworks.com:8050

vix.yarn.resourcemanager.scheduler.address: hdfs://sandbox.hortonworks.com:8030

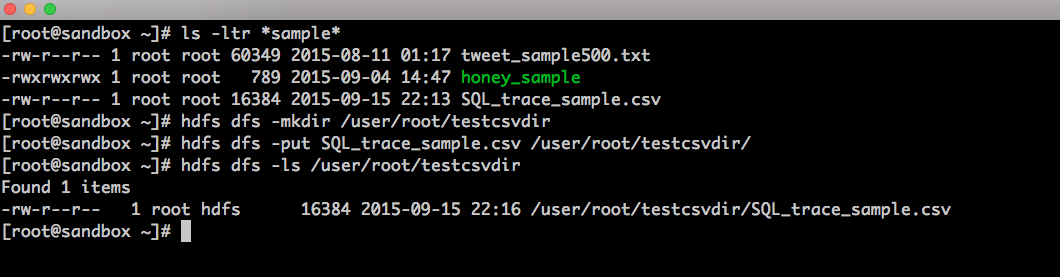


### Setup Hunk Virtual Index

We need to upload some data into hdfs so that it can create a Virtual index on that data in Hunk. I have a sample file named SQL\_trace\_sample.csv in sandbox under ~/ so moved that to a new directory within hdfs using following commands

hdfs dfs -mkdir /user/root/testcsvdir  
hdfs dfs -put SQL\_trace\_sample.csv /user/root/testcsvdir/

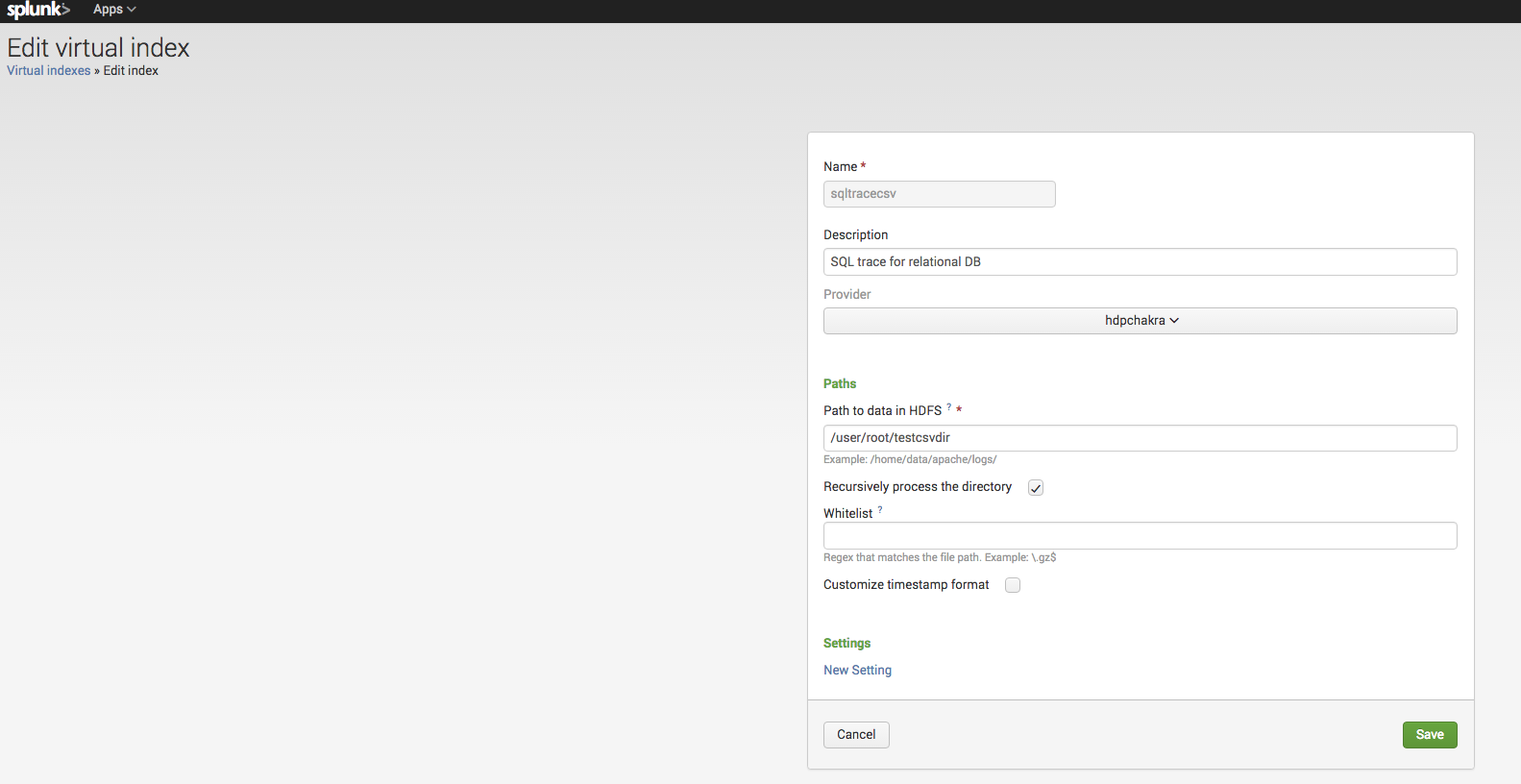
hdfs dfs -ls /user/root/testcsvdir



Now you can create a new Virtual index for it in Hunk. Here are the parameters for which you need to enter the values

Name: needs to be all lowercase

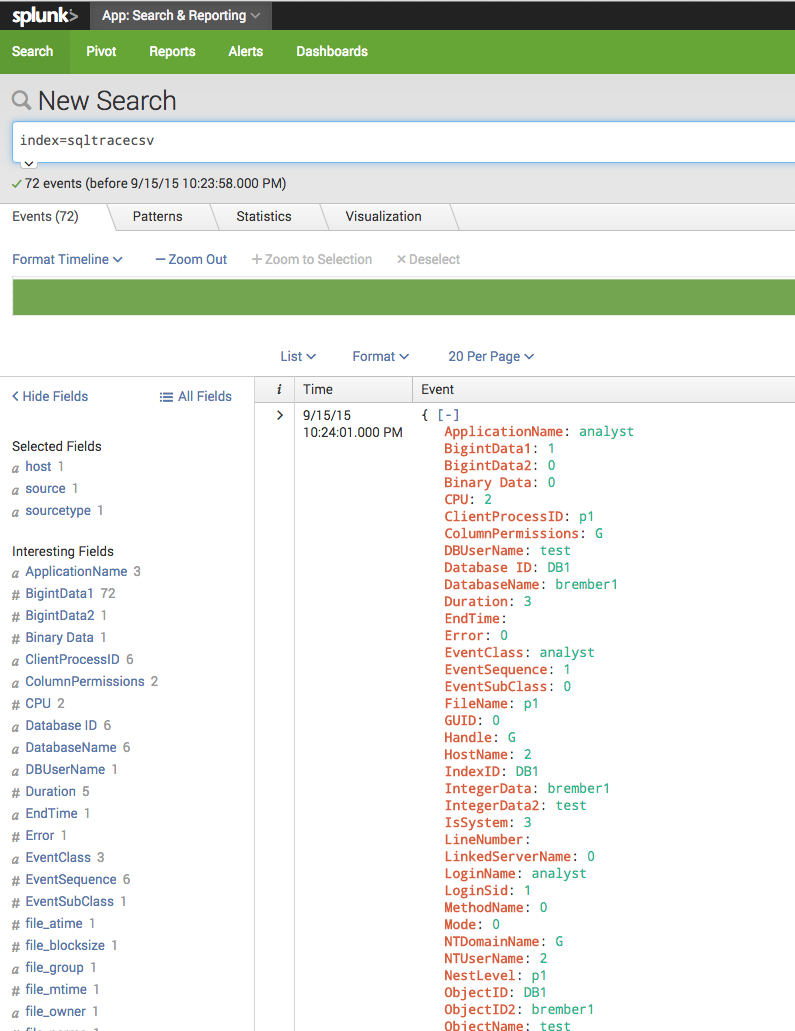
Path to data in HDFS: /user/root/testcsvdir



### Search the data via Hunk

Now as we have virtual index we can search the data in hdfs via Hunk. You can do that by click Search under the action column.





Now you are all set to use the power of Hunk to analyze this data.

# **Revision**

|  |  |  |
| --- | --- | --- |
| Date | Name | Description |
| 09/15/2015 | Chakra Sankaraiah | Initial draft of the document. |
|  |  |  |
|  |  |  |