

Configure Hortonworks Sandbox Version 2.0 with Hunk: Splunk Analytics for Hadoop

V2.0

November 4, 2013

Introduction

Summary

This tutorial describes how to connect Hortonworks Sandbox Version 2.0 with Hortonworks Data Platform 2.0 to Hunk™: Splunk Analytics for Hadoop. Hunk offers an integrated platform to rapidly explore, analyze and visualize data that resides natively in Hadoop.

Prerequisites

- [Hortonworks Sandbox 2.0](#) (installed and running)
- Hunk – download a 60 day free trial at <http://www.splunk.com/download/hunk>
- A virtual or physical 64-bit Linux operating system
- Java version 1.6 or later (v1.7 recommended)

Introduction to Hunk Architecture

Hunk is a high performance, scalable software server written in Java, C/C++ and Python. Hunk works with machine data generated by any application, server or device. The Splunk Developer API is accessible via REST or the command line.



After downloading, installing and starting Hunk, you'll find two Hunk Server processes running on your host: splunkd and splunkweb.

- **splunkd** is a distributed C/C++ server that accesses, processes and creates a virtual index from machine data and handles search requests. splunkd supports a command line interface for searching and viewing results.
- **splunkweb** is a Python-based application server providing the Splunk Web user interface. It allows you to search and navigate machine data accessible by Hunk and to manage your Hunk deployment using your browser.

Overview

The following are the main steps to install Hunk, point to your Hortonworks Sandbox, and begin to explore, analyze and visualize data in Hadoop.

1. Install Hunk on 64-bit Linux.
2. Set up the Hunk Search Head.
3. Point Hunk to your Hortonworks Sandbox.
4. Define a Virtual Index to a Data Set in your Sandbox.
5. Use Hunk to Explore, Analyze and Visualize Data in Hadoop.
6. Use Mixed-mode Search.
7. Use the Splunk Developer Platform.

Step 1 – Install Hunk on 64-bit Linux

Download your favorite flavor of Linux and install the Hunk file on this physical or virtual machine. You can install Hunk on 64-bit Linux using RPM, a tar file or DEB install. The following are instructions for each option.

(1) RedHat RPM install

To install the Hunk RPM in the default directory /opt/splunk:

```
rpm -i splunk_package_name.rpm
```

To install Hunk in a different directory, use the --prefix flag:

```
rpm -i --prefix=/opt/new_directory splunk_package_name.rpm
```

If you want to automate your RPM install with kickstart, add the following to your kickstart file:

```
./splunk start --accept-license
```

```
./splunk enable boot-start
```

Note: The second line is optional for the kickstart file.

(2) Tar file install

Expand the tarball into an appropriate directory using the tar command:

```
tar xvfz splunk_package_name.tgz
```

The default install directory is splunk in the current working directory.

To install into /opt/splunk, use the following command:

```
tar xvfz splunk_package_name.tgz -C /opt
```

When you install Hunk with a tarball:

- Some non-GNU versions of tar might not have the -C argument available. In this case, if you want to install in /opt/splunk, either cd to /opt or place the tarball in /opt before running the tar command. This method will work for any accessible directory on your machine's filesystem.
- Hunk does not create the user automatically. If you want Hunk to run as a specific user, you must create the user manually before installing.
- Ensure that the disk partition has enough space to hold the search artifacts.

(3) DEB install

To install the Hunk DEB package:

```
dpkg -i splunk_package_name.deb
```

Note that you can only install the Splunk DEB package in the default location, /opt/splunk.

Step 2 – Set up the Hunk Search Head

In preparation for setting up the Hunk Search Head, you'll want to install Java and the Hadoop Client Libraries.

Hunk requires Java 1.6 or later, (1.7 recommended). To install Java:

```
# yum install java-1.6.0-openjdk
```

Note: Install the full jdk using "yum install java-1.6.0-openjdk-devel" if you want jstack for troubleshooting.

Java Location:

This will install the java binaries in : /usr/lib/jvm/jre-1.6.0

This is a symbolic link to java-1.6.0 -> /etc/alternatives/java_sdk_1.6.0

Keep note of the location of JAVA_HOME and HADOOP_HOME.

Next, install the Hadoop client binaries.

Get the Hadoop Client Binaries:

```
# wget http://apache.mesi.com.ar/hadoop/common/hadoop-1.2.0/hadoop-1.2.0.tar.gz
```

Install them in /opt/hadoop

Modify your PATH to include the /opt/hadoop/bin directory:

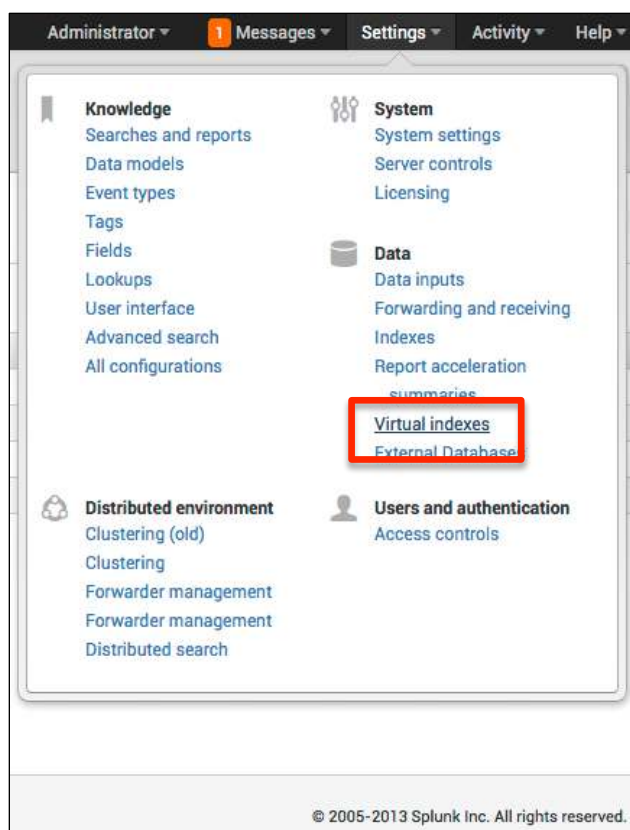
```
[root@sandbox opt] # tar xzvf hadoop-1.2.0.tar.gz -C .
```

Next, start Hunk > /opt/splunk/bin/splunk start

Step 3 – Point Hunk to your Hortonworks Sandbox

In this step, you will point Hunk to the Hadoop cluster running in the Hortonworks Sandbox virtual machine and select your version of MapReduce.

In Hunk, select Settings > Virtual Indexes.



Click the green button New Provider. Title HdpSandbox or the title of your choice.

Below are how the values should look like after you have installed Java and the Hadoop Binaries on the Search Head.

JAVA HOME: /usr/lib/jvm/jre-1.6.0

Hadoop Home: /opt/hadoop

Job tracker: *Leave blank*

Hadoop Version: Choose Hadoop 2.x YARN

File System: hdfs://sandbox:8020

This value is set in the core-site.xml file which is located on the Hortonworks sandbox at /usr/lib/hadoop/conf/core-site.xml under:

```
<property>
<name>fs.default.name</name>
<value>hdfs://sandbox:8020</value>
</property>
```

Provider family

hadoop

Environment variables

Java home

/usr/lib/jvm/jre-1.6.0

Example: /usr/jdk

Hadoop home

/opt/hadoop

Example: /usr/hadoop

Hadoop cluster information

Hadoop version

Hadoop 2.x, (Yarn) ▼

Job tracker

Example: jobtracker.example.com:8021

File system

hdfs://sandbox:8020

Example: hdfs://namenode.example.com:8020

HDFS Working Directory: /user/<user-running-hunk>

This is a scratch space the used by Hunk to store intermediate results and files that it needs to push to the Hadoop nodes.

Job queue: Default

Listed under **Additional settings** are defaults that are changeable.

vix.yarn.resourcemanager.address

sandbox:8050

✕

n.resourcemanager.scheduler.address

sandbox:8030

✕

vix.yarn.resourcemanager.address
sandbox:8050

vix.yarn.resourcemanager.scheduler.address
sandbox:8030.

```
vix.splunk.home.datanode  
/tmp/splunk/$SPLUNK_SERVER_NAME
```

```
vix.splunk.setup.package  
/path/to/splunk/install/bits/on/local/searchhead/splunk-6.0-xxxxxx-Linux-x86_64.tgz
```

Scroll to the bottom of the list of additional settings and click the green button Save. Under the list of Providers you will now see Hortonworks Sandbox. You can reopen the user interface for configurations by clicking on the title Hortonworks Sandbox.

Step 4 – Define a Virtual Index to a Data Set in your Sandbox

You'll want data in your sandbox to be able to search, analyze and visualize with Hunk. You can import data through the HDFS API, Flume, Sqoop or another data ingestion method that connects to the HDFS API. Refer to Hortonworks documentation on how to load data into HDFS.

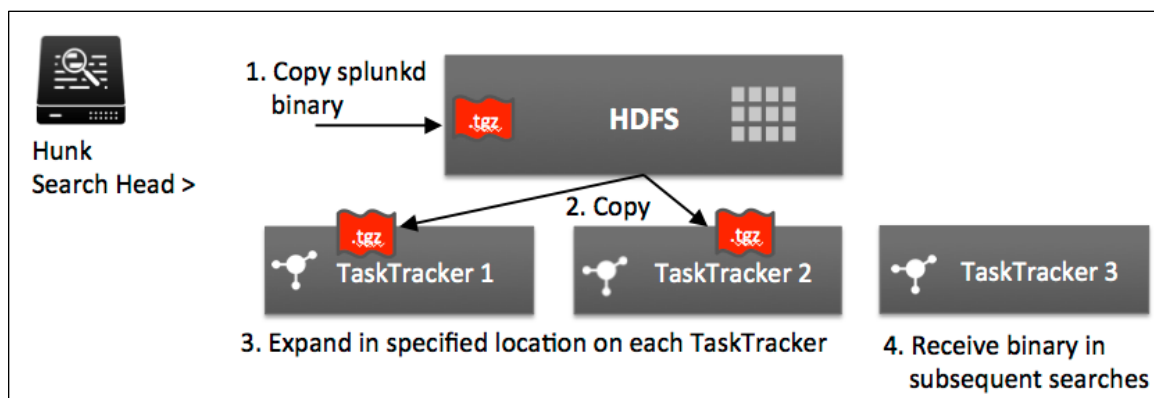
Step 5 – Use Hunk to Explore, Analyze and Visualize Data in Hadoop

If you use Splunk Enterprise you'll be familiar with the Hunk software interface. If you are new to Splunk software, in the upper left-hand corner, select Search & Reporting >. You can now begin to ask and answer questions of your data in Hadoop. Under How to Search, you'll see links for Documentation and Tutorial to help you get started with search, reports, dashboarding and more.

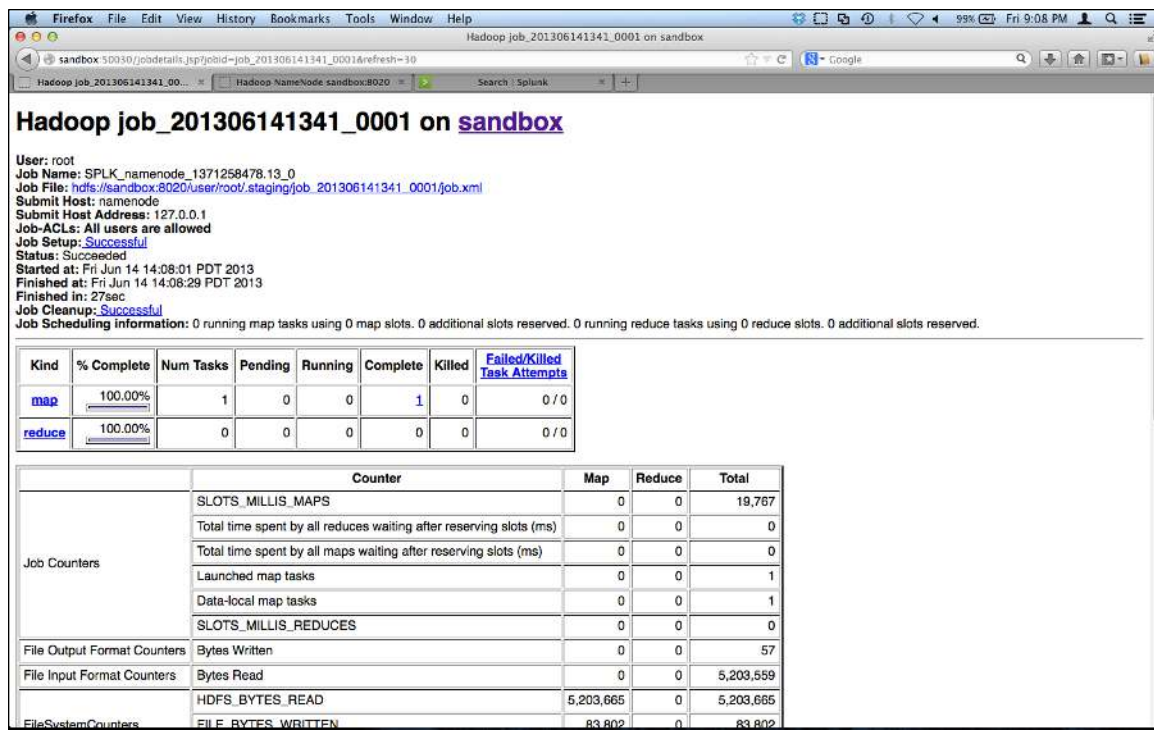
On the first search that spawns a MapReduce job, Hunk installs all the necessary components in the Hadoop nodes.

1. The orchestration process begins when Hunk copies the Hunk binary .tgz file to HDFS.
2. Each TaskTracker (or ApplicationContainer in YARN) fetches the binary.
3. The binary files expand in the specified location on each TaskTracker; the default location is configurable.
4. TaskTrackers not involved in the 1st search will receive the Hunk binary in a subsequent search that involves those TaskTrackers.

This process is one example of why Hunk needs some scratch space in HDFS and in the local file system (TaskTrackers / DataNodes).



Reference <http://sandbox:50070/dfshealth.jsp> or <http://sandbox:50030/jobtracker.jsp> to see the MapReduce Jobs spawned as Splunk reaches out to the nodes in the Hortonworks Sandbox cluster.



Hadoop job_201306141341_0001 on sandbox

User: root
 Job Name: SPLK_namenode_1371258478.13_0
 Job File: hdfs://sandbox:8020/user/root/staging/job_201306141341_0001/job.xml
 Submit Host: namenode
 Submit Host Address: 127.0.0.1
 Job-ACLs: All users are allowed
 Job Setup: Successful
 Status: Succeeded
 Started at: Fri Jun 14 14:08:01 PDT 2013
 Finished at: Fri Jun 14 14:08:29 PDT 2013
 Finished in: 27sec
 Job Cleanup: Successful
 Job Scheduling information: 0 running map tasks using 0 map slots. 0 additional slots reserved. 0 running reduce tasks using 0 reduce slots. 0 additional slots reserved.

Kind	% Complete	Num Tasks	Pending	Running	Complete	Killed	Failed/Killed Task Attempts
map	100.00%	1	0	0	1	0	0 / 0
reduce	100.00%	0	0	0	0	0	0 / 0

	Counter	Map	Reduce	Total
Job Counters	SLOTS_MILLIS_MAPS	0	0	19,767
	Total time spent by all reduces waiting after reserving slots (ms)	0	0	0
	Total time spent by all maps waiting after reserving slots (ms)	0	0	0
	Launched map tasks	0	0	1
	Data-local map tasks	0	0	1
	SLOTS_MILLIS_REDUCES	0	0	0
File Output Format Counters	Bytes Written	0	0	57
File Input Format Counters	Bytes Read	0	0	5,203,665
	HDFS_BYTES_READ	5,203,665	0	5,203,665
FilesystemCounters	FILE_BYTES_WRITTEN	83,802	0	83,802

Hunk applies structure to data at search time and is designed for data exploration across large datasets to preview data and iterate quickly. Unlike Hive or SQL on Hadoop approaches, there is no requirement to understand the data upfront and no brittle schema to maintain or update. You can find patterns and trends across disparate data sets in a “grab bag” Hadoop cluster.

Hunk supports almost all of the Splunk Search Processing Language (SPL), excluding Transactions and Localize, which require Splunk Enterprise native indexes.

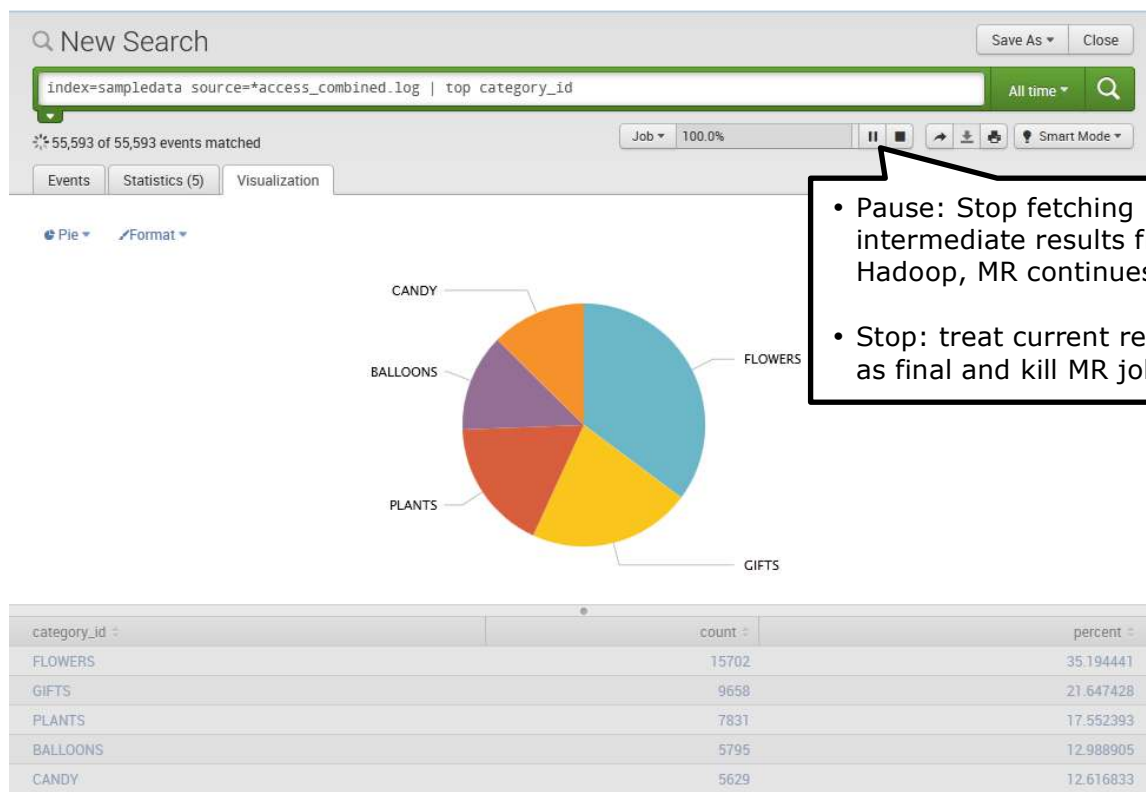
Hunk uses some HDFS space to store binaries, configuration bundles and intermediate search results - the amount depends primarily on the size of the intermediate search results. Between 10 and 20 GBs is common. Hunk also uses DataNode/TaskTracker local temp disk space, at most 5 GBs per DataNode/TaskTracker.

You can continue to use the additional Apache projects and subprojects included with your Hortonworks Sandbox - Hunk requires just MapReduce and HDFS.

Hunk does not manage data ingest. For ingest management, Hadoop system admins can use one of the open source projects for data collection (Flume / Scribe / Chukwa, or Sqoop for relational data), the HDFS API to import or export data, or use Splunk Hadoop Connect for bi-directional data transfer between HDFS and Splunk Enterprise. Hunk works with any compression method supported by HDFS (e.g., gzip, bzip or Izo).

Step 6 – Use Mixed-mode Search

Hunk starts streaming and reporting modes concurrently. The streaming mode transfers data from HDFS to the Hunk search head for immediate processing resulting in quick result previews. Hunk continues streaming data until the reporting (MapReduce) results start to become available. This allows you to search interactively by stopping and refining your searches.

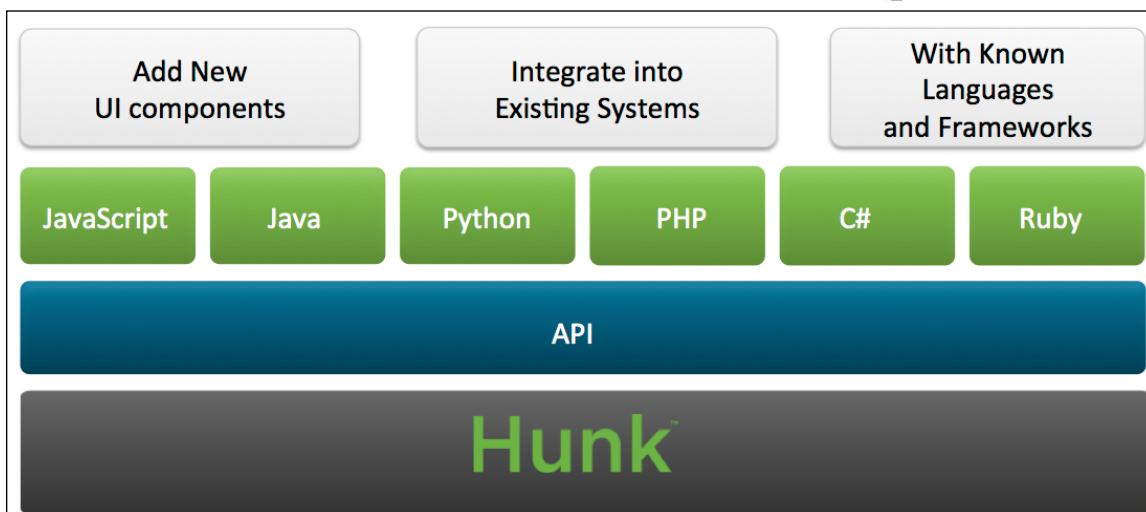


- Pause: Stop fetching intermediate results from Hadoop, MR continues
- Stop: treat current results as final and kill MR job

Lastly, we'll explore how to further extend Hunk's capabilities with the Splunk Developer Platform.

Step 7 – Use the Splunk Developer Platform

Splunk Enterprise and Hunk offer a powerful developer platform with familiar tools you can use to build big data enterprise apps on top of data in Hadoop. Use the tools and frameworks that your developers already know. Integrate Hunk charts, dashboards and query results into other applications. Create workflows that trigger an action in an external system or use REST endpoints.



Feedback

We're interested in your feedback on this tutorial. Please take this [short survey](#).

About Splunk

Splunk Inc. (NASDAQ: SPLK) provides the engine for machine data™. Splunk® software collects, indexes and harnesses the machine-generated [big data](#) coming from the websites, applications, servers, networks, sensors and mobile devices that power business. Splunk software enables organizations to monitor, search, analyze, visualize and act on massive streams of real-time and historical machine data. More than 6,000 enterprises, universities, government agencies and service providers in over 90 countries use Splunk Enterprise to gain [Operational Intelligence](#) that deepens business and customer understanding, improves service and uptime, reduces cost and mitigates cybersecurity risk. To learn more, please visit www.splunk.com/company.

About Hortonworks

Hortonworks develops, distributes and supports the only 100-percent open source distribution of Apache Hadoop explicitly architected, built and tested for enterprise grade deployments. Developed by the original architects, builders and operators of Hadoop, Hortonworks stewards the core and delivers the critical services required by the enterprise to reliably and effectively run Hadoop at scale. Our distribution, Hortonworks Data Platform, provides an open and stable foundation for enterprises and a growing ecosystem to build and deploy big data solutions. Hortonworks also provides unmatched technical support, training and certification programs. For more information, visit www.hortonworks.com. The Hortonworks Sandbox can be found at: www.hortonworks.com/sandbox.