**Setup Guide**

**Hadoop, Nifi, Spark, Zeppelin, Flume Kafka**

Table of Contents

[Change Log 2](#_Toc457804116)

[Hadoop 2](#_Toc457804117)

[NiFi 3](#_Toc457804118)

[Spark 4](#_Toc457804119)

[Zeppelin 5](#_Toc457804120)

[Flume Kafka 6](#_Toc457804121)

# Change Log

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Date** | **Revision #** | **Name of Person Making Change** | **Name of Person Requesting Change** | **Description of Change(s) (specified by FDR #)** |
| 8/1/2016 | 1.0 | Amritha Jayanti | Initial Version | Setting up document |
| 8/1/2016 | 1.1 | Drew Patel |  | First Draft |

# Hadoop

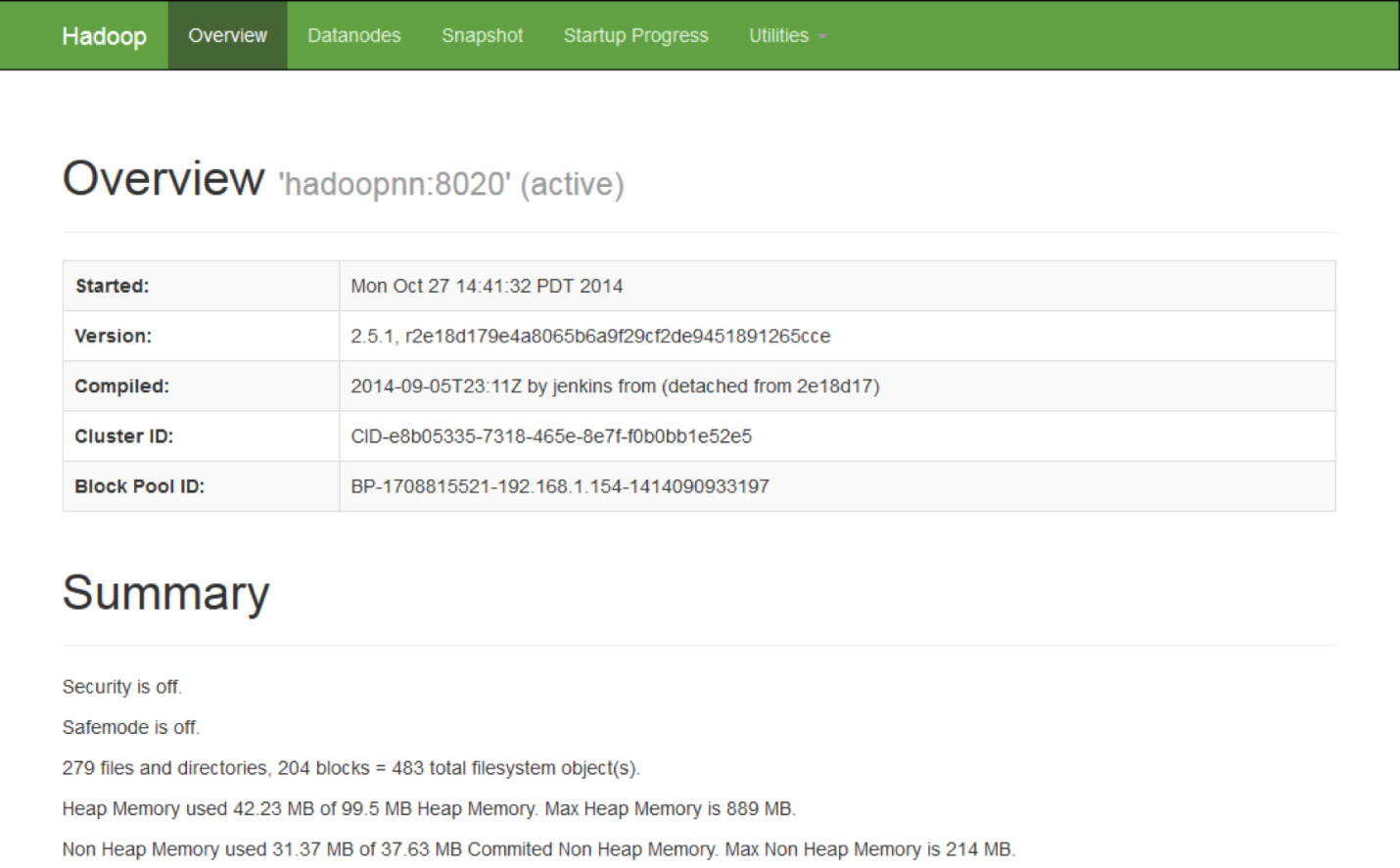
**Prerequisites**: Java 7 or higher must be installed and JAVA\_HOME should be set.

**Download:** [**http://www.apache.org/dyn/closer.cgi/hadoop/common/**](http://www.apache.org/dyn/closer.cgi/hadoop/common/)

**Single Node:**



* Once extracted copy and replace these files in your Hadoop directory. Update hdfs-site.xml with user’s SID.
* Set HADOOP\_Home =<Hadoop\_install\_dir>
* Add <hadoop\_install\_dir>\bin to PATH variable
* Initialize Hadoop file system
  + <hadoop\_install\_dir>\bin\hadoop namenode -format
  + <hadoop\_install\_dir>\sbin\start-dfs.cmd
* Hadoop GUI could be accessed at <http://localhost:50070> to make sure Hadoop was started successfully.



* References:
  + https://www.airpair.com/hadoop/posts/getting-started-with-hadoop

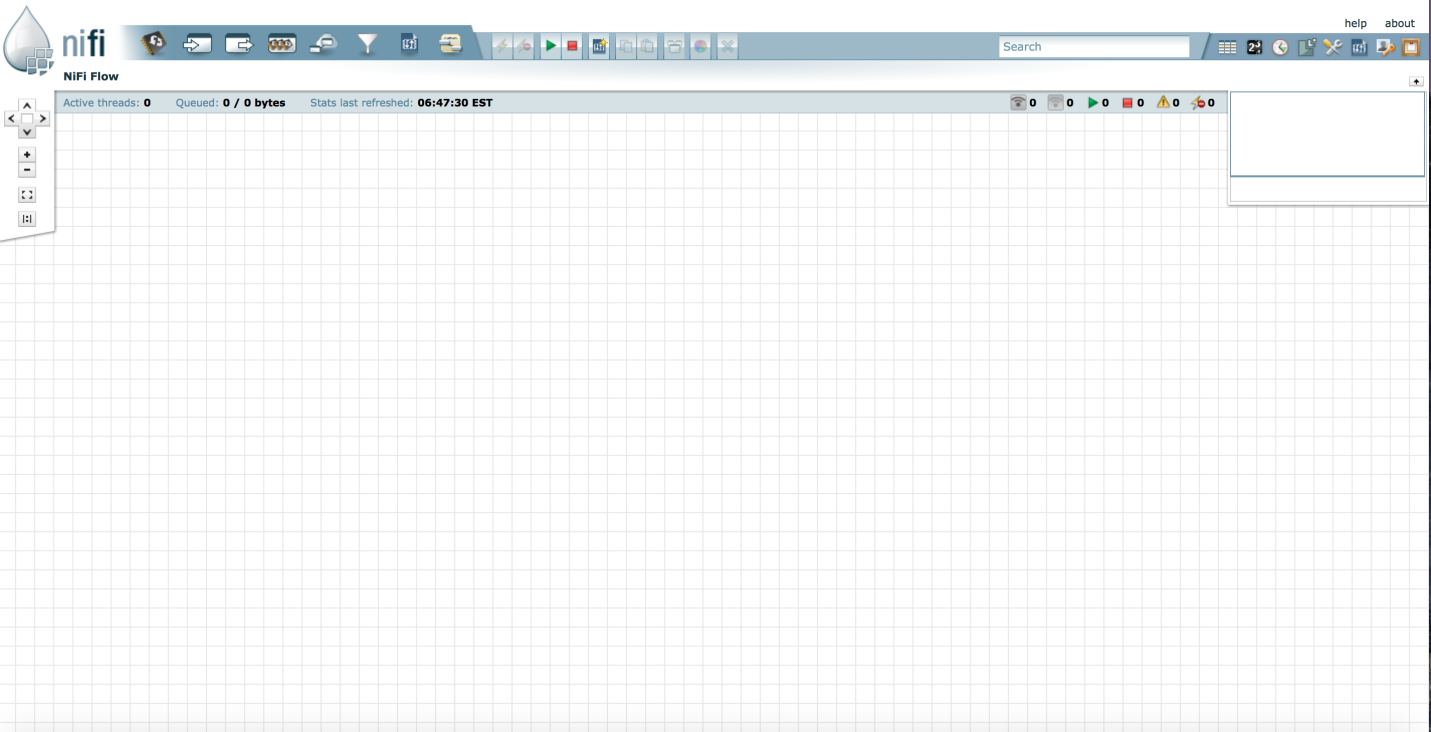
# NiFi

**Prerequisites**: Java 7 or higher must be installed and JAVA\_HOME should be set.

**Download:** NiFi can be downloaded from the [NiFi Downloads Page](http://nifi.apache.org/download.html). There are two packaging options available: a "tarball" that is tailored more to Linux and a zip file that is more applicable for Windows users. Once downloaded simply extract the archive to the location you wish to run the application from.

**Starting NiFi:**

* Once downloaded, For Windows users, navigate to the folder where NiFi was installed. Within this folder is a subfolder named **bin**. Navigate to this subfolder and double-click the **run-nifi.bat** file.
* NiFi is now running in the background, GUI can be accessed at : <http://localhost:8080/nifi>.
* For Linux and OSX users, use a Terminal window to navigate to the directory where NiFi was installed. To run NiFi in the foreground, run **bin/nifi.sh run**. This will leave the application running until the user presses Ctrl-C. To run NiFi is background run **bin/nifi.sh start**.



**Stopping NiFi:**

* Wherever NiFi is running, either in command prompt or in a terminal it could be terminated by pressing **Ctrl-C**.

**Possible Issues:**

* If there is something already running at port 8080, the port number could be changed by going into conf/nifi.properties and editing the nifi.web.http.port value.

**References:**

* **https://nifi.apache.org/docs/nifi-docs/html/getting-started.html**

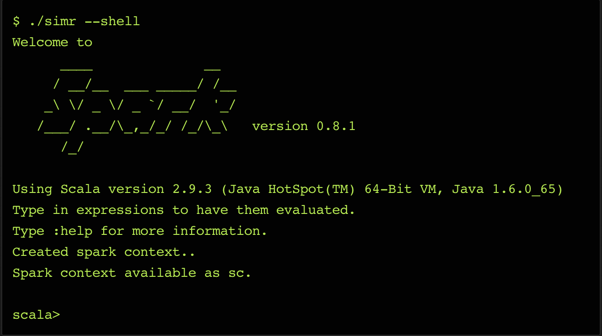
# Spark

**Prerequisites**: Java 7 or higher must be installed and JAVA\_HOME should be set. Hadoop should be downloaded and started.

**Download:** Latest version of spark could be downloaded from [Spark website](http://spark.apache.org/downloads.html).

**Starting Spark:**

* Once the zip folder is downloaded, extract the archive to the directory where you want to run the application from
* Follow the steps to start Hadoop, followed by:
  + <hadoop\_install\_dir>\bin\hadoop fs -mkdir -p /tmp/hive
  + <hadoop\_install\_dir>\bin\hadoop fs -chmod 777 /tmp/hive
  + <hadoop\_install\_dir>\bin\winutils.exe chmod 777 \tmp\hive
* To start you can execute /bin/spark-shell.cmd --master local[2] in windows.
* To start spark in Linux execute : ./bin/spark-shell.
* Once spark is started you should see **scala>** in the command line, ready for use.



**References:**

* <http://stackoverflow.com/questions/25481325/how-to-set-up-spark-on-windows>
* http://spark.apache.org/docs/latest/building-spark.html

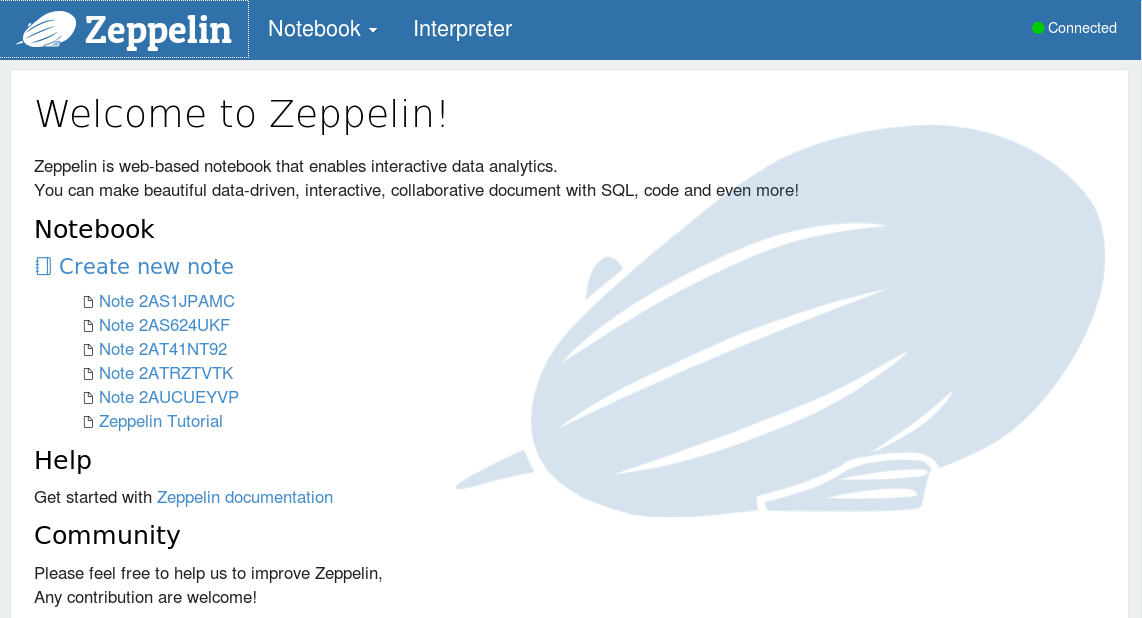
# Zeppelin

**Prerequisites**: Java 7 or higher must be installed and JAVA\_HOME should be set. Hadoop should be downloaded and started.

**Download:** Latest version of spark could be downloaded from [Zeppelin website](https://zeppelin.apache.org/download.html).

**Starting Zepplin:**

* Once you download the prebuilt binary package, move the archive to directory from where you would like to run the application.
* To start Zepplin, execute:
  + **Linux**: bin/zeppelin-daemon.sh start
  + **Windows:**  bin\zeppelin.cmd
* After successful start, visit http://localhost:8080 with your web browser to access zeppelin.



**Stopping Zeppelin:**

* To stop zeppelin execute in Linux : bin/zeppelin-daemon.sh stop
* To stop zeppelin in windows press **Ctrl-C.**

# Flume

**Prerequisites**:

* Java Runtime Environment - Java 1.6 or later (Java 1.7 Recommended)
* Memory - Sufficient memory for configurations used by sources, channels or sinks
* Disk Space - Sufficient disk space for configurations used by channels or sinks
* Directory Permissions - Read/Write permissions for directories used by agent

**Download:** Latest version of spark could be downloaded from [Flume website](https://flume.apache.org/download.html).

**Starting Flume:**

* Once you download the prebuilt binary package, move the archive to directory from where you would like to run the application.
* In order to start a flume agent execute:
  + bin\flume-ng agent -name MyAgent -f conf/MyAgent.properties -property "flume.root.logger=INFO,LOGFILE,console;flume.log.file=MyLog.log;hadoop.home.dir=C:/winutils".
* An agent takes in a config files for which a template can be found under **conf** folder.

**References:**

* https://flume.apache.org/FlumeUserGuide.html

# Kafka

**Prerequisites**:

* Java Runtime Environment - Java 1.6 or later (Java 1.7 Recommended)

**Download:** Latest version of spark could be downloaded from [Apache Kafka website](http://kafka.apache.org/downloads.html)

**Starting Kafka:**

* Kafka uses ZooKeeper so you need to first start a ZooKeeper server if you don't already have one.
  + bin/windows/zookeeper-server-start.bat config/zookeeper.properties
* Now start the Kafka server.
  + bin/windows/kafka-server-start.bat config/server.properties
* Kafka streams messages on a topic, in order to create a topic execute:
  + bin/kafka-topics.bat --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic test
* To see that topic is running:
  + bin/kafka-topics.bat --list --zookeeper localhost:2181
* To send messages on the topic, you need a producer. To start a producer execute:
  + bin/windows/kafka-console-producer.bat --broker-list localhost:9092 --topic test
* To check if the flow is working, Kafka allows you to run a consumer that will dump out the messages to standard output:
  + bin/windows/kafka-console-consumer.bat --zookeeper localhost:2181 --topic test --from-beginning

**References:**

* **http://kafka.apache.org/documentation.html#quickstart**