

Steps to Run a Spark Streaming Program with Kafka

1. Ensure that VirtualBox is running with the latest version of the VM_Image (see “**How to Set up a Local Hadoop Cluster with a VM Image**”).
2. Ensure that Ambari is enabled (see “**How to Set up a Local Hadoop Cluster with a VM Image**” steps 8-10).
3. Log into [Ambari](#) with username: admin and password: admin and check to see that only the services needed to run Spark/Kafka/Hbase/Zookeeper are started (see “**Services Guide**” in N:/Temp/PCMod_VM_Images/General Documentation). The **easiest way to start/stop services one by one is to click on the “Hosts” tab at the top Ambari tool bar, then click on “sandbox.hortonworks.com” as a host and you will see each service listed.**
4. Enter the VM console mode by hitting Alt+F5, then log in with username: root password: hadoop.
5. Go to Kafka directory using the command:

```
cd /usr/hdp/2.2.4.2-2/kafka
```

6. Start the Kafka process using the command:

```
bin/kafka-server-start.sh config/server.properties
```

Note that while the Kafka process is running it may get out of sync and throw up a **stack trace**. **It will restart on its own, so don't worry about that.**

Open a new shell window (using PuTTY or some other ssh terminal) to continue with the following steps. Use the host address 127.0.0.1 and port 2222, then the same credentials as in step 4 above.

7. Repeat step 5 above in the shell.
8. Create a new topic in Kafka using the command:

```
bin/kafka-topics.sh --create --topic iauto --partitions 1 --replication-factor 1 --zookeeper localhost:2181
```

9. If you are having issues with not enough memory or resources, go to [Ambari](#) and stop unnecessary services as described in step 3 above.
10. Using FileZilla get the jar statefarm-streaming-0.01-jar-with-dependencies.jar from N:\TEMP\FJB7\ folder to /root (home directory for root) directory on your virtual machine. Note that Filezilla can get hung up when trying to copy directly from the N drive to the VM. If that happens you will have to copy and paste the file first to your workstation and FTP them to the VM with Filezilla.
11. Using FileZilla get the contents of logs directory (all files) in N:\TEMP\FJB7\logs\ to /root directory on your virtual machine (see note in step 10 above). If you do not **already have a folder called “logs” in your root directory, then copy the entire “logs” folder over to the VM.**

12. Open the HBase shell by executing the following commands:

1. Switch to HBase user using the command:

```
su hbase
```

2. **If you haven't** started the HBase Master through Ambari, start the HBase master using the command:

```
/usr/hdp/2.2.4.2-2/hbase/bin/hbase-daemon.sh --config  
/usr/hdp/2.2.4.2-2/hbase/conf start master
```

3. **If you haven't started the HBase Region Server through Ambari**, start the HBase Region Server using command:

```
/usr/hdp/2.2.4.2-2/hbase/bin/hbase-daemon.sh --config  
/usr/hdp/2.2.4.2-2/hbase/conf start regionserver
```

4. Go to HBase directory:

```
cd /usr/hdp/2.2.4.2-2/hbase
```

5. Open HBase shell by executing the command:

```
hbase shell
```

6. Create 'pcmoddes' namespace using the command:

```
create_namespace 'pcmoddes'
```

7. Create the table 'sfref' with column family 'n1' using the command:

```
create 'pcmoddes:sfref','n1'
```

8. Create the table 'iauto' with column families 'data', 'ds' using the command:

```
create 'pcmoddes:iauto','data','ds'
```

9. Exit out of HBase using `exit` command

10. Switch back to root using `exit` command

13. Verify that spark is running by following commands

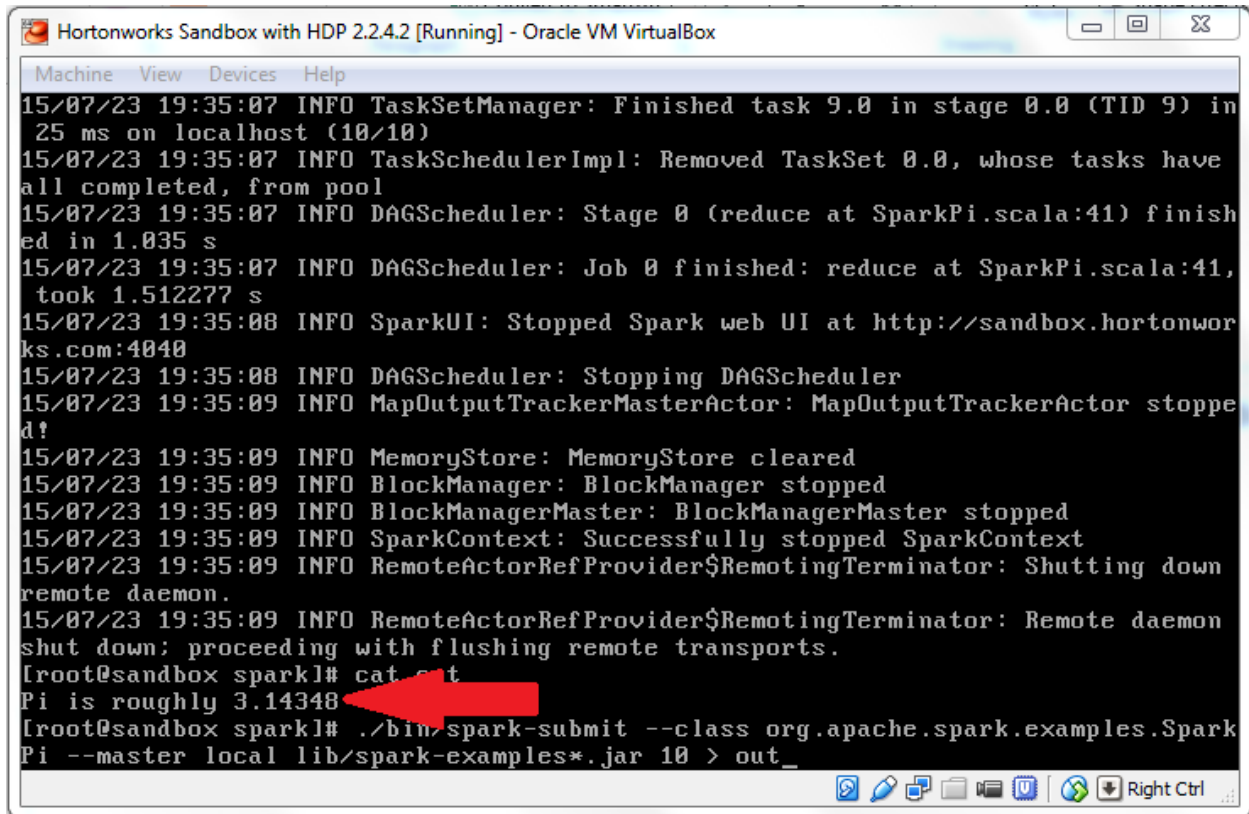
1. Go to spark directory using the command:

```
cd /usr/hdp/2.2.4.2-2/spark
```

2. Execute the following command to run a sample spark program:

```
bin/spark-submit --class org.apache.spark.examples.SparkPi --  
master local lib/spark-examples-1.2.1.2.2.4.2-2-  
hadoop2.6.0.2.2.4.2-2.jar 10 > out
```

If you now type `cat out`, you should see something similar to “Pi is roughly 3.14348”



```
Hortonworks Sandbox with HDP 2.2.4.2 [Running] - Oracle VM VirtualBox  
Machine View Devices Help  
15/07/23 19:35:07 INFO TaskSetManager: Finished task 9.0 in stage 0.0 (TID 9) in  
25 ms on localhost (10/10)  
15/07/23 19:35:07 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have  
all completed, from pool  
15/07/23 19:35:07 INFO DAGScheduler: Stage 0 (reduce at SparkPi.scala:41) finish  
ed in 1.035 s  
15/07/23 19:35:07 INFO DAGScheduler: Job 0 finished: reduce at SparkPi.scala:41,  
took 1.512277 s  
15/07/23 19:35:08 INFO SparkUI: Stopped Spark web UI at http://sandbox.hortonwor  
ks.com:4040  
15/07/23 19:35:08 INFO DAGScheduler: Stopping DAGScheduler  
15/07/23 19:35:09 INFO MapOutputTrackerMasterActor: MapOutputTrackerActor stoppe  
d!  
15/07/23 19:35:09 INFO MemoryStore: MemoryStore cleared  
15/07/23 19:35:09 INFO BlockManager: BlockManager stopped  
15/07/23 19:35:09 INFO BlockManagerMaster: BlockManagerMaster stopped  
15/07/23 19:35:09 INFO SparkContext: Successfully stopped SparkContext  
15/07/23 19:35:09 INFO RemoteActorRefProvider$RemotingTerminator: Shutting down  
remote daemon.  
15/07/23 19:35:09 INFO RemoteActorRefProvider$RemotingTerminator: Remote daemon  
shut down; proceeding with flushing remote transports.  
[root@sandbox spark]# cat out  
Pi is roughly 3.14348  
[root@sandbox spark]# ./bin/spark-submit --class org.apache.spark.examples.Spark  
Pi --master local lib/spark-examples*.jar 10 > out_
```

14. Execute the following command to insert the data into Kafka topic:

```
bin/spark-submit  
--class com.statefarm.streaming.ingest.FileKafkaStuffer  
--master local  
/root/statefarm-streaming-0.01-jar-with-dependencies.jar  
-t iauto  
-k sandbox.hortonworks.com:6667  
-d file:///root/logs/
```

15. Execute the following command to add the data from Kafka to HBase:

```
bin/spark-submit  
--class com.statefarm.streaming.ingest.KafkaIngest  
--master local[2]  
/root/statefarm-streaming-0.01-jar-with-dependencies.jar  
-t iauto  
-z sandbox.hortonworks.com:2181  
-g testgroup  
-x 1
```

```
-c  
com.statefarm.streaming.transformations.iauto.IAutoXMLToHBaseTransformat  
ion  
-f /usr/hdp/2.2.4.2-2/hbase/conf/hbase-site.xml
```