CM2606 Data Engineering

Tutorial 02

The aim of this tutorial is to:

1. Identify the components of a data pipeline.
2. Apply that knowledge to design a pipeline for a given scenario.
3. Set up a Hadoop environment in local machines to be used in rest of the tutorials.

Task 1:

Consider a retail supermarket which tries to leverage insights from the data they capture to improve their operations. Organization have installed cctv cameras covering all areas of the store and returning customers are identified at point of sales based on their loyalty IDs. Organization is currently exploring ways to utilize all available data to find out busy hours and predict the no. of customers who would visit the outlet for a given day/hour. Based on this output they are planning to allocate additional staff for the busy hours.

You are required to draw up a data pipeline to cater this requirement.

1. Identify the possible data sources.
2. Draw up the data pipeline you would suggest achieving this requirement.
3. Referring to the design above, briefly explain how the data pipeline supposed to work in achieving the business requirement.

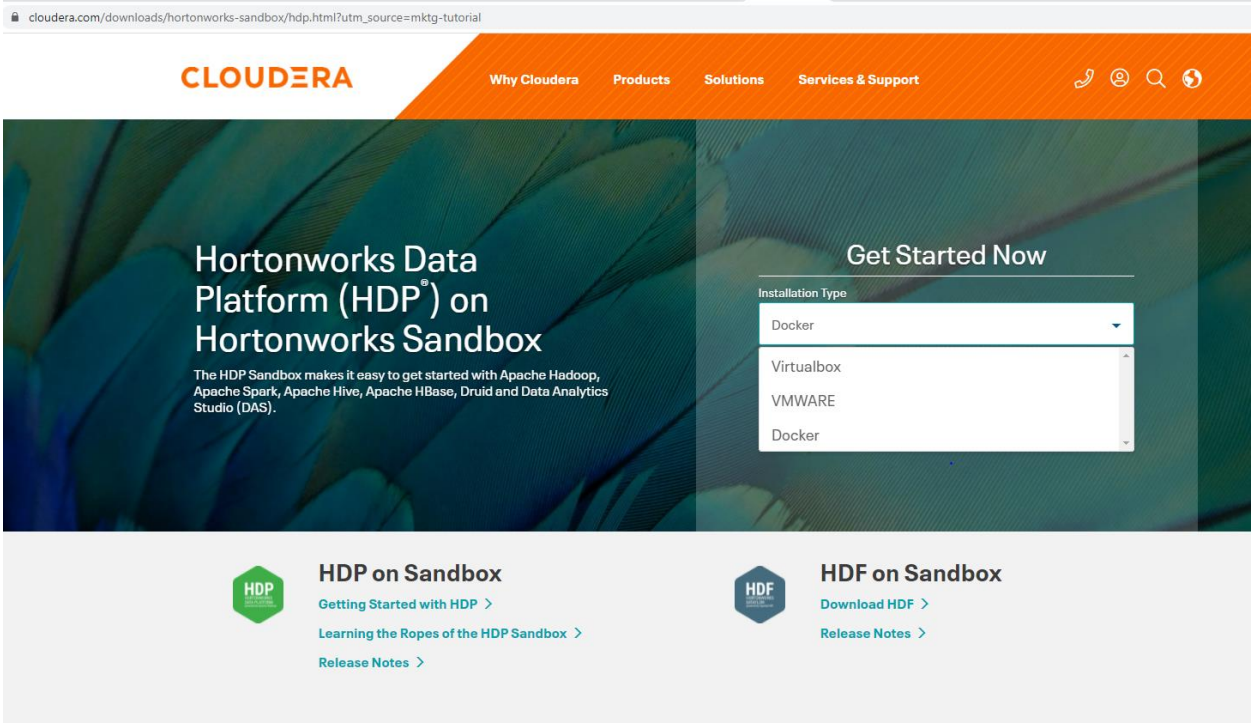
****

Task 2:

Set up Hadoop environment in your local machine and make sure you have access to the following services:

- HDFS
- Hive
- Spark

Step 1: You may download the Cloudera/Hortonworks Sandbox from this location. This tutorial uses the virtual box option (approx. 20GB file to be downloaded) but feel free to use other installation type convenient for you.

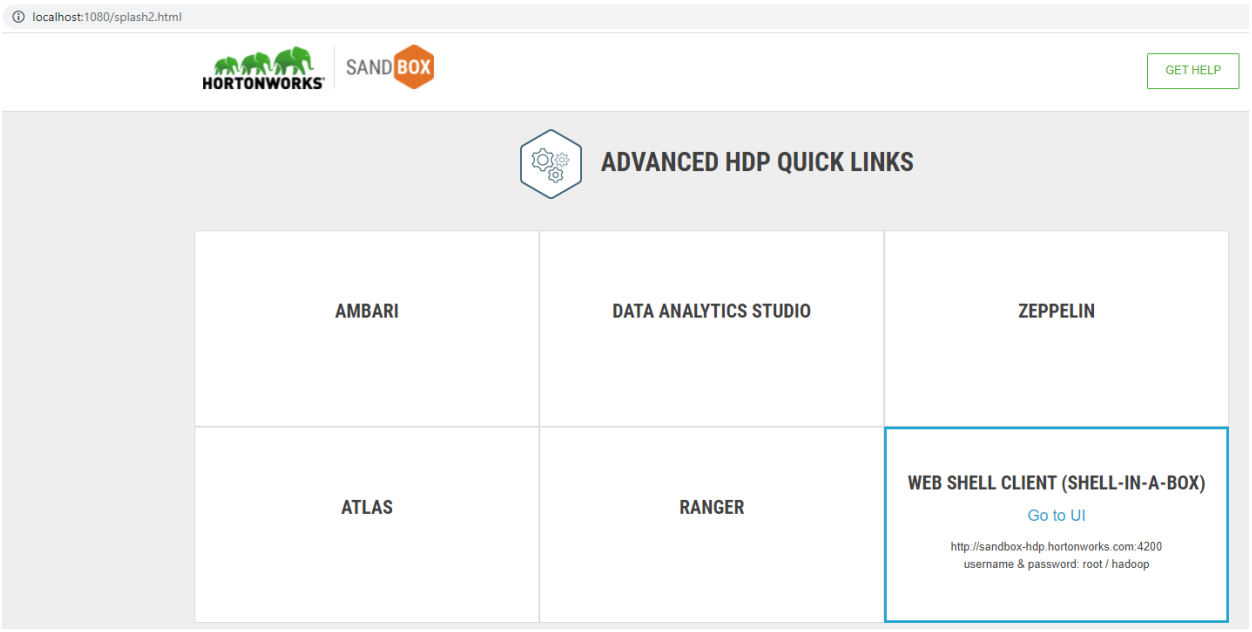


You can also follow any other approach that is convenient for you, such as:

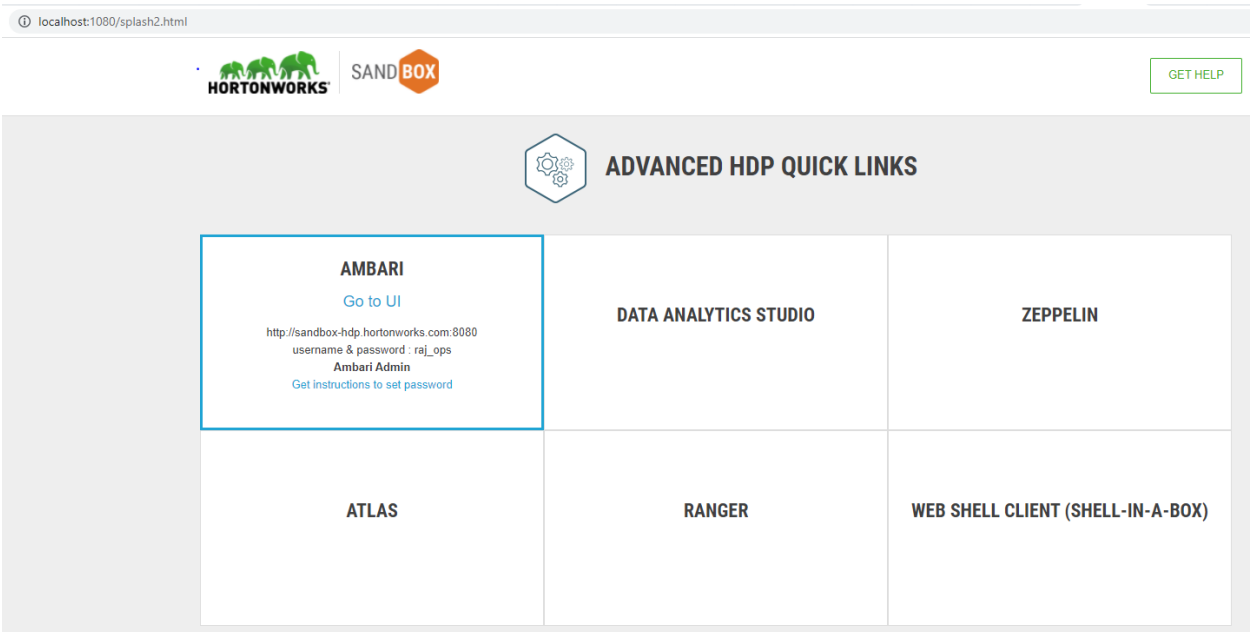
- Cloudera QuickStart Docker Image
- Use Hadoop bin file as explained in this video. You might need to install Hive on top of this.

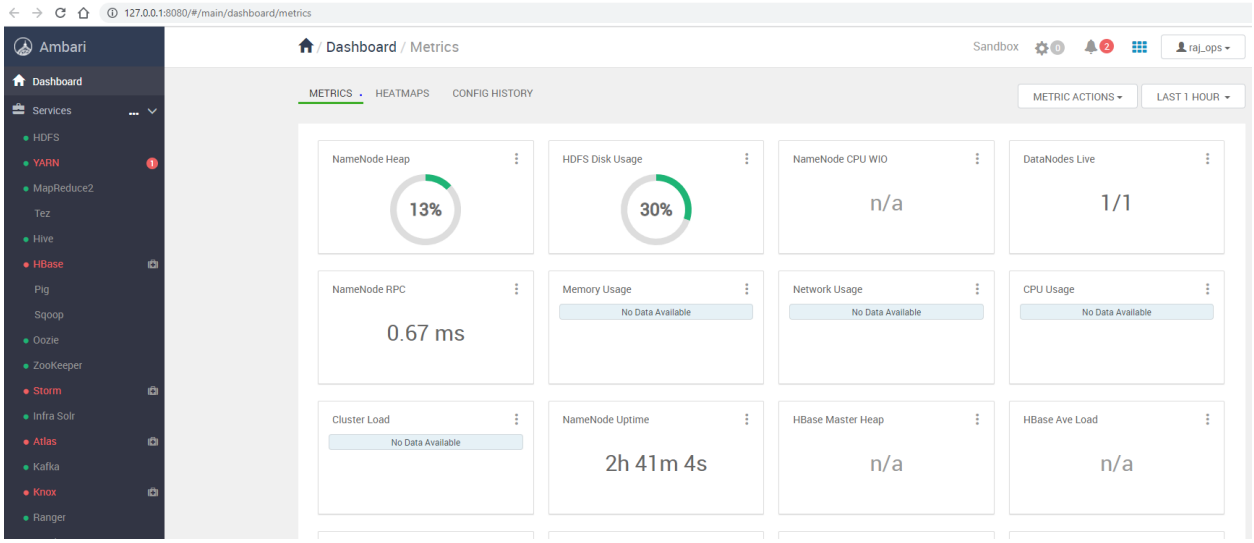
Step 2: If you chose to download the virtual box in step 1 you can follow this guide and this video to install the same. 4GB RAM and 2 cores should be enough to start with.

Step 3: After following the above guide to launch home page of the sand box (http://localhost:1080), open the Web-Shell Client using the mentioned root password. After login for the first time, you will be asked to reset the root password.



Step 4: You could also explore the entire Hadoop eco system using Ambari Dashboard.




127.0.0.1:8080/#/main/dashboard/metrics
Ambari
Dashboard
Services
HDFS
YARN
MapReduce2
Tez
Hive
HBase
Pig
Sqoop
Oozie
ZooKeeper
Storm
Infra Solr
Atlas
Kafka
Knox
Ranger
Dashboard / Metrics
Sandbox
raj_ops
METRICS
HEATMAPS
CONFIG HISTORY
METRIC ACTIONS
LAST 1 HOUR
NameNode Heap
13%
HDFS Disk Usage
30%
NameNode CPU WIO
n/a
DataNodes Live
1/1
NameNode RPC
0.67 ms
Memory Usage
No Data Available
Network Usage
No Data Available
CPU Usage
No Data Available
Cluster Load
No Data Available
NameNode Uptime
2h 41m 4s
HBase Master Heap
n/a
HBase Ave Load
n/a