**Apache BigTop v1.1 on IBM SoftLayer**

**Ubuntu 14.04 Quick Start**

Amir Sanjay, IBM Big Data Solutions

Bill Phu, IBM Software Developer

Donna Ball, IBM Systems

Maria Ward, IBM Systems

March 23, 2016







 ****

Home

**Table of Content**

Contents

[1. Overview: 3](#_Toc446572424)

[2. Getting Started on SoftLayer 4](#_Toc446572426)

[3. Steps for Preparing the System for the BigTop Install 4](#_Toc446572427)

[4. Installing BigTop Hadoop 5](#_Toc446572428)

[5. Installing and Running the Hadoop Test Script 6](#_Toc446572429)

[6. Installing and Running the Spark Test Script 6](#_Toc446572430)

[7. Zeppelin Tutorial 7](#_Toc446572431)

[8. Zeppelin Stock Intraday Workload 9](#_Toc446572432)

## Overview:

## This document outlines the process of installing Apache Hadoop-Spark BigTop v1.1.0 bundle on an IBM Power SoftLayer Bare Metal system running Ubuntu 14.04. Apache Zeppelin notebook is included in the bundled install script to run an initial benchmark suite.

**Outline of steps:**

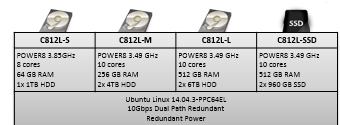
* Client places order:

http://[www.ibm.com/buycloud](http://www.ibm.com/buycloud)

http[://knowledgelayer.softlayer.com](http://www.ibm.com/partnerworld/page/POQ12361USEN)/gettingstarted

* Client receives Welcome package including web portal access with Userid and password
* Client accesses Power SoftLayer system and prepares for Installation
* Client downloads IBM script to download, configure, and install Apache Hadoop - Spark package for Linux on Power
* Client logs onto Zeppelin and runs the pre-configured benchmark

Current Power 8 SoftLayer offerings are bundled in 4 convenient package sizes:



For more information, visit: <http://www.ibm.com/cloud-computing/au/en/start-with-cloud.html>

**System Minimum Requirements for BigTop:**

* + 4 Cores Minimum
  + 32GB of Memory
  + 500GB vSCSI disk
  + Linux Ubuntu 14.04.03 Little Endian
* **Storage**
  + 500GB available disk space per node
* **Networking** 
  + All virtual networking through the Virtual I/O Server

## Getting Started on SoftLayer

**Accessing the IBM Power SoftLayer Bare Metal Server:**

You will receive a Welcome email from SoftLayer with a link to the SoftLayer Control Portal and login information.

**View the very detailed Tutorial Videos that walk a user through first time access to fully configuring and accessing their SoftLayer system:**

**SoftLayer Tutorial Videos:**

[**https://www.youtube.com/watch?v=IaNrq1stI3g**](https://www.youtube.com/watch?v=IaNrq1stI3g)[**os**](http://www.ibm.com/partnerworld/page/POQ12361USEN)

**Follow the steps outlined in the Tutorial Videos to:**

* Logon - first time access
* Quickly navigate through menus to configure SoftLayer BM System
* Understand where to access online tools, including how to re-image the system
* Access the Virtual System

**For more information on SoftLayer setup and configuration, visit:**

http[://knowledgelayer.softlayer.com](http://www.ibm.com/partnerworld/page/POQ12361USEN)/

## Steps for Preparing the System for the BigTop Install

The install\_bigtop.sh script will quickly, in less than 45 minutes, completely install and configure Hadoop, Spark, and Zeppelin (a tool for running benchmarking) automatically.

**install\_bigtop.sh script requirements:**

* User must have super-user privileges

**It is highly recommended** that install\_bigtop.sh be installed on a freshly installed Ubuntu kernel. If the system has already been in use, please run the cleanup.sh script located with the rest of the packages here: **https://github.com/ibmsoe/bigtop/**

**Note:** the install\_bigtop.sh will fail to install if the ~/bigtop/source directory is created

**For a fully documented procedure, starting with SoftLayer configuration through running a sample Zeppelin benchmark, visit: https://github.com/ibmsoe/bigtop/docs/**

* If installing as a non-root user: (Example: bigtop\_user)
  + Add new User:
  + Set the new User password:
  + Login as the new User:
  + Change to the User home directory

$ useradd bigtop\_user –U –G sudo –m

$ passwd bigtop\_user (enter passwords)

$ su bigtop\_user

$ cd ~

* Download the ***install\_bigtop.sh*** script from github:

**For the first-time download:**

**Note: install git first ( sudo apt-get install git )**

git clone <https://github.com/ibmsoe/bigtop>

**For the subsequent updates**: git pull

This will download the following scripts to a directory called bigtop:

* + cleanup.sh
  + install\_bigtop.sh
  + restart-bigtop.sh
  + hadoopTest.sh
  + status.sh
  + sparkTest.sh

**Or** use **wget** command for an individual file download, i.e. download install\_bigtop.sh file

wget <https://raw.githubusercontent.com/ibmsoe/bigtop/master/install_bigtop.sh>

$ git clone <https://github.com/ibmsoe/bigtop>

$ cd bigtop/

~/bigtop$ ls

install\_bigtop.sh restart-bigtop.sh status.sh Stock\_workload.json

cleanup.sh LICENSE.md source stockprices.csv.gz-aa

hadoopTest.sh README.md sparkTest.sh stockprices.csv.gz-ab

* Verify that /etc/hosts has the hostname associated with the private IP if using both private and public IPs
* Increase the default maximum number of open files available: (Example set to 1000000)

Sudo vi /etc/security/limits.conf

add the following lines (this affects all users at next log in):

\*                soft    nofile          1000000

\*                hard    nofile          1000000

* Logout and back in for the new ulimit value to take effect.

## Installing BigTop Hadoop

Downloading and Installing Apache Hadoop – Spark with Zeppelin package using the install\_bigtop.sh script

* Run the ***install\_bigtop.sh*** script:

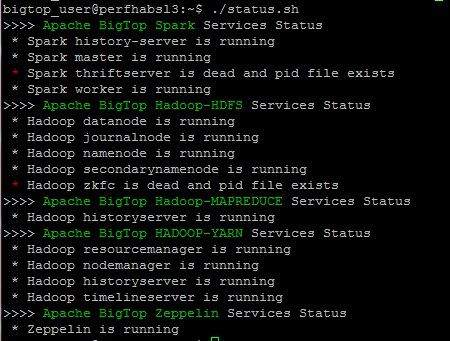
Example: ./install\_bigtop.sh

**Note:** ignore error messages, during the installation process, pertaining to hadoop and yarn not starting. These packages are being installed, however, the system is not configured to allow the packages to start running at this time.

**The install\_bigtop.**sh script will do the following:

* + Installs all dependencies ( Java Open JDK1.8)
  + Download and installs the latest Apache BigTop Hadoop 2.7.1 debian packages, including:
    - Hadoop v 2.7.1
    - Bigtop-groovey v2.4.4
    - Jsvc v1.0.15
    - Tomcat v6.0.36
    - Zookeeper v3.4.6
    - Scala v2.10.4
  + Configures the environment for Hadoop
  + Formats the HDFS
  + Download and installs Apache BigTop Spark 1.5.1
  + Download and installs Zeppelin v0.5.6
  + Starts all configured services on a Single Node
* After the ***install\_bigtop.sh*** script completes, verify that everything is up and running, as expected, using the status.sh script:

Example: $ ./status.sh



**Note:** Spark Thriftserver and Hadoop zkfc are not needed for this benchmark

## Installing and Running the Hadoop Test Script

To verify that the installation was successful, run the following test script:

* Run the provided hadoopTest.sh script to verify that hadoop is working properly

Example: $ ./hadoopTest.sh

## Installing and Running the Spark Test Script

* Run the sparkTest.sh script

Example: $ ./sparkTest.sh

* Verify results

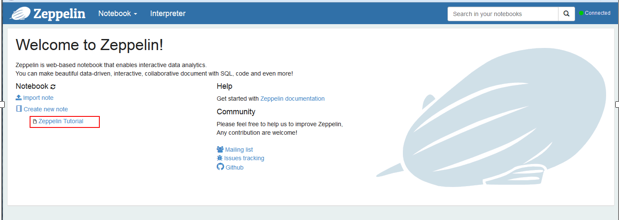
$ sudo ./sparkTest.sh

Pi is roughly 3.1427

## Zeppelin Tutorial

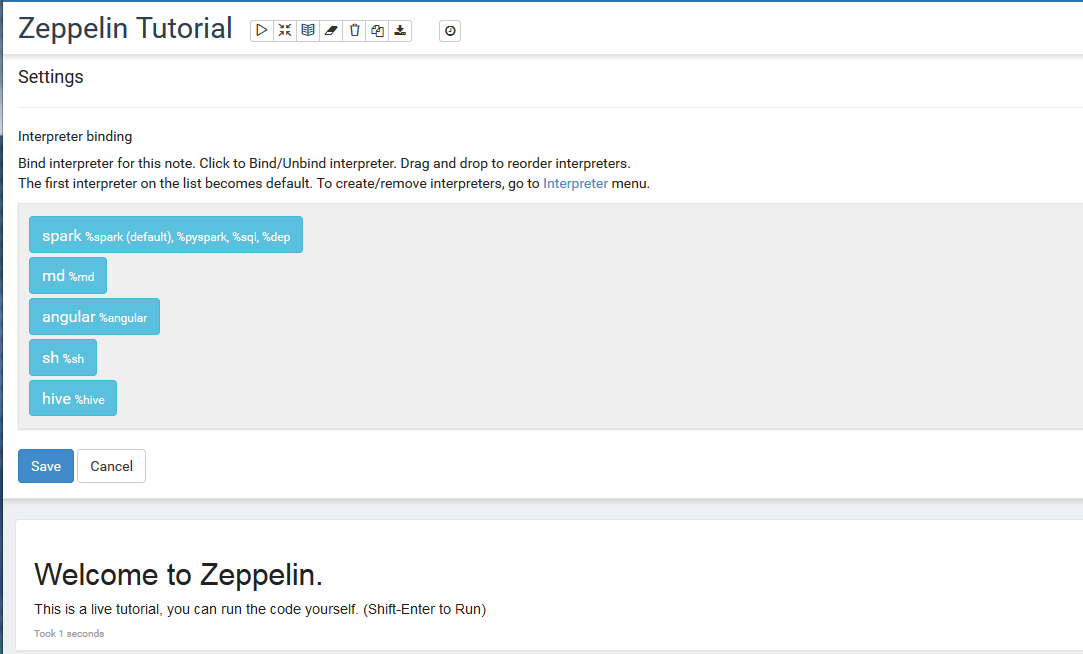
* Login to Zeppelin from your browser:

( Example: http://<Private IP Addr>:8080 )



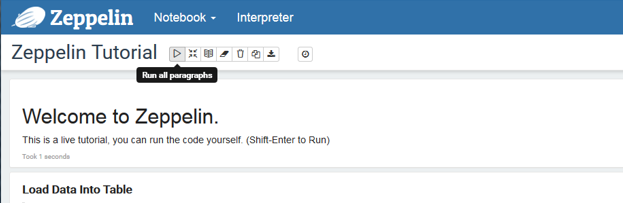
* Run the Tutorial benchmark by Left-clicking on Zeppelin Tutorial

**Note:** If the Interpreter binding section is open with interpreters highlighted in blue ( see image below) Left Click the SAVE button.

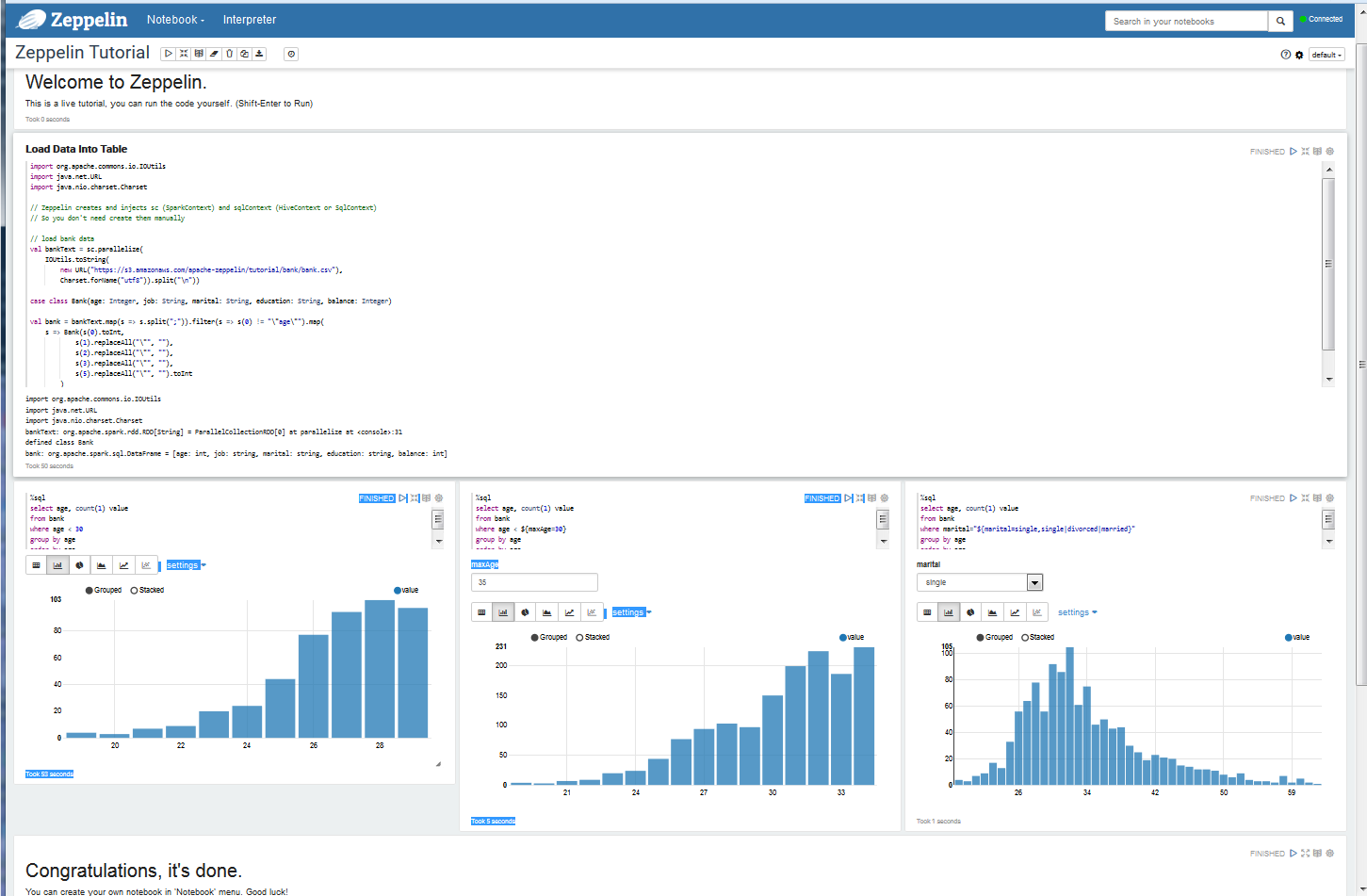


* Then select the *arrowhead* symbol: Run all paragraphs

(Click OK when prompted)



* Scroll down to view results



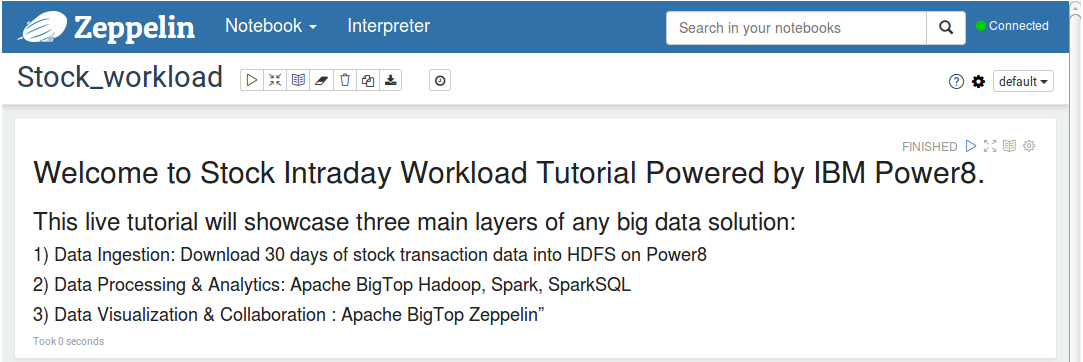
## Zeppelin Sample Stock Intraday Workload

If needed: download the new zeppelin json file:

wget <https://github.com/ibmsoe/bigtop/raw/master/Stock_workload.json>

**Note:** The json file needs to be on the system viewing the Zeppelin Notebook.

* Go to the Zeppelin Welcome page
* Select ***Import note***
* Click on “**Choose a JSON here**” panel
* Browse to select the Stock\_workload.json file
* Run it by selecting the newly added **Stock\_workload.json** notebook on the Welcome page



Before starting the sample Stock\_workload, use the Interpreter tab to optimize the Spark properties according to your system configuration.

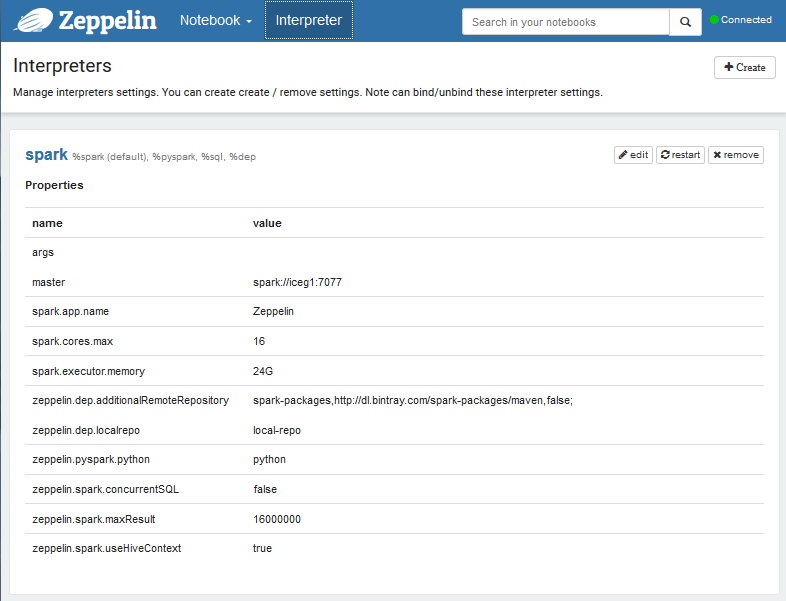
To optimize Spark Interpreter settings:

* Select the ***Interpreter Tab***
* Select ***Edit*** ( see image below)
* Change the values to maximize the Spark properties for your system, specifically consider:
* spark.cores.max
  + Leave the value empty to use all available cpu
* Spark.executor.memory
  + 16G minimum
* Zeppelin.spark.maxResults ( Default is 1000. )
  + 40,000 at a minimum to for good results
* ***Save*** the changes
* Return to the Stock\_workload notebook by using the Notebook pull-down tab and selecting the notebook.

**Note:** Any *Rule of thumbs* given are suggestions.

Example: Best Results for an IBM Power SoftLayer Bare Metal Server of type C812L-L with the 10 Cores and 512G of Memory were obtained when

* spark.cores.max was set to 16
* Spark.executor.memory was set to 100GB
* Zeppelin.spark.maxResults was set to 16000000



**Note:** Starting the sample Stock\_workload by selecting the arrow head to run all the first time will report that all workloads under the Data Ingestion fail. The script appears to be loading data and starting the actual workloads too soon. Subsequent runs will report the Data Ingestion as failing (this is a known bug: it is failing because the data has already been loaded and can be ignored.)

**Note:** When re-running the workload, Data Ingestion will report an Error because the stockprices.csv data has already been copied to /user/zeppelin/. Remove stockprices.csv or ignore the Error.

You can try this same Zeppelin workload on comparable x86 environments and see for yourself the benefits that Linux on Power brings to running Spark.