**Apache Spark - PySpark Guide**

Stephen Blystone ([smb032100@utdallas.edu](mailto:smb032100@utdallas.edu))

Document Version 1.0

Last Updated: Mar 27, 2018

Table of Contents

[**Revision History** 3](#_Toc509962514)

[**Installing Apache Spark and PySpark** 3](#_Toc509962515)

[**Streaming Applications** 3](#_Toc509962516)

# **Revision History**

|  |  |  |
| --- | --- | --- |
| **Version** | **Date** | **Changes** |
| 1.0 | 3/27/2018 | Initial Version |

# **Installing Apache Spark and PySpark**

1. Download Apache Spark 2.3.0 to your VM. To choose your version you can go to <https://spark.apache.org/downloads.html>. Below is the mirror it gave me:

wget <http://mirrors.ocf.berkeley.edu/apache/spark/spark-2.3.0/spark-2.3.0-bin-hadoop2.7.tgz>

tar -xzvf spark-2.3.0-bin-hadoop2.7.tgz

cd ./spark-2.3.0-bin-hadoop2.7/

1. To launch standalone Master Server:

./sbin/start-master.sh

**NOTE**: You can view the log that is written to check if there are any errors. The log file is displayed when you run the above command.

* 1. Launch the Spark Master web GUI on 10.28.34.14:8080
     1. IP address is the address of the VM.
     2. Record the “URL: value” for launching workers.
        1. 
        2. My VM name is “Application”. Yours may be different.

1. Launch a Worker by running the following command:

./sbin/start-slave.sh spark://Application:7077

The highlighted portion is what you recorded above on the Spark Master web GUI.

**NOTE**: You can view the log that is written to check if there are any errors. The log file is displayed when you run the above command.

1. For more details: <https://spark.apache.org/docs/latest/spark-standalone.html>

# **Streaming Applications**

1. <https://spark.apache.org/docs/latest/streaming-programming-guide.html>
2. Asdf