Big Data Applications

Contents

Introduction

Pre-requisites

- Be comfortable with basic Python. Familiarity with Numpy and Pandas is beneficial but not required.
- Have an environment ready for the several hands-on exercises (e.g., own laptop or cloud resource)

Syllabus

Something broken?

If you encounter any problem or bug in these materials, please remember to add an issue to the course repo, explaining the problem and, potentially, its solution. By doing this, you will improve the instructions for future users. 🎉

Installation Instructions

Introduction

This document contains instructions for installation of the software we'll be using during the course. If you want to follow the training on your own machines, then please complete these instructions.

If you encounter any problem during installation and you manage to solve them, please remember to add an issue, explaining the problem and solution. By doing this you will be helping to improve the instructions for future users! :tada:

What we're installing

- the `Python` programming language (version `3.7` or greater)
- `git` for version control
- your favourite text editor
- `Apache Spark`
- `Jupyter notebooks`

Please ensure that you have a computer or a cloud-based workbench with all of these installed.

Linux

Package Manager

`Linux` users should be able to use their package manager to install all of the software from this page.

However note that if you are running an older `Linux` distribution you may get older versions with different look and features. We target Ubuntu 21.10 for the sake of homogeneity.

> **Tip**
>
> We recommend to upgrade the system and get the latest security packages: `sudo apt-get upgrade`
>
> Please do not forget to restart after the upgrade finishes.

You can test the `Python` version by issuing:

```
python3 --version
```

You should get an output similar to the following:

```
Python 3.9.7
```

Python via package manager

Recents versions of `Ubuntu` come with mostly up to date versions of all needed packages.

The version of `IPython` might be slightly out of date. Thus, you may wish to upgrade this using `pip`.

You should ensure that the following packages are installed using `apt-get`:

- `python3-pip`
- `jupyter`
- `ipython3`

```
sudo apt-update
sudo apt-get install -y gcc make perl python3-pip jupyter ipython3
```

You can test the installation by issuing:

```
pip --version
```

You should get an output similar to the following:

```
pip 20.3.4 from /usr/lib/python3/dist-packages/pip (python 3.9)
```

Editor

You have many different text editors suitable for programming at your fingertips. Here is an opinionated list of editors in case you do not already have a favourite:

- Visual Studio Code
- Atom
- Neovim
- Vim

Apache Spark

We will setup Apache Spark now. You need to open the `downloads` page, and download a spark distribution. We suggest to choose the same options as on the screenshot below. If you see a newer version is available, please feel free to choose it.

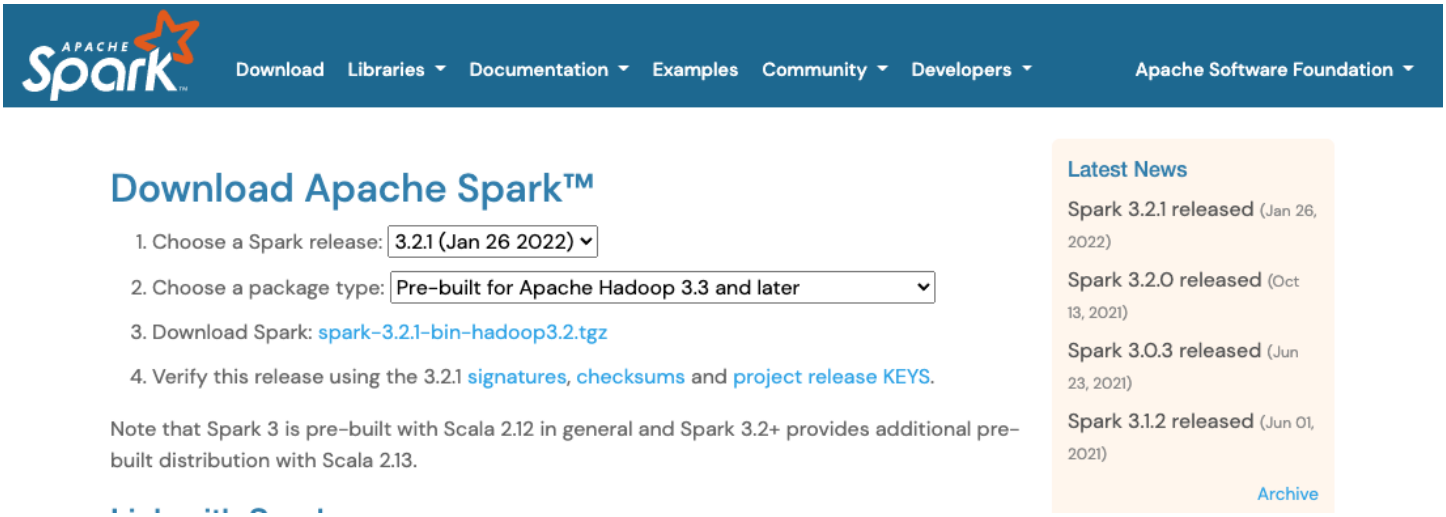
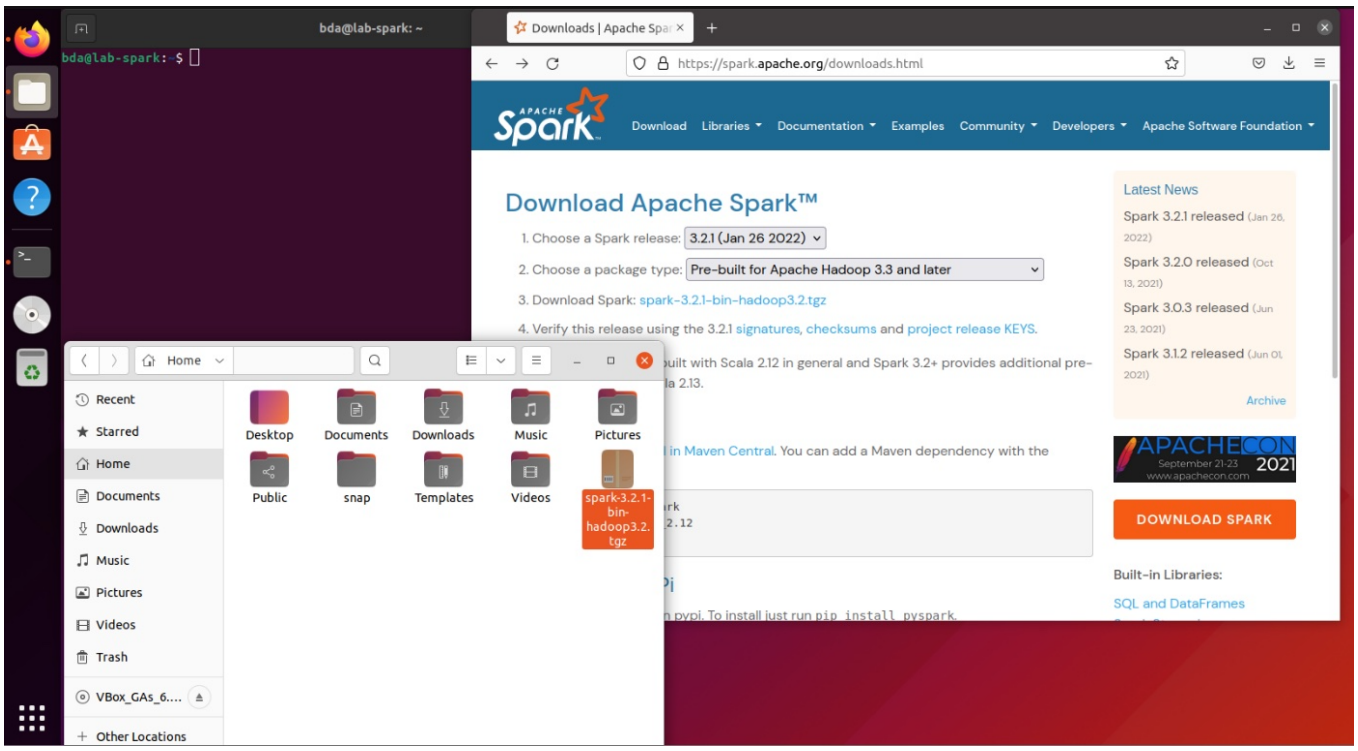


Image Link

> **Tip**
>
> We want the package in the right location, so open the file explorer and place it into your home folder.
>
> 
>

Then go to your command line and issue the following to unzip the downloaded file:

```
sudo tar -xzvf spark-3.2.1-bin-hadoop3.2.tgz
```

> **Attention**
>
> The above command assumes you downloaded version 3.2.1 with hadoop 3.2 binaries. Please ammend it as needed.

Now what we need to do is telling `Python` how to find Spark:

```
echo "export SPARK_HOME=~/spark-3.2.1-bin-hadoop3.2" >> ~/.bashrc
echo "export PATH=$SPARK_HOME:$PATH" >> ~/.bashrc
echo "export PYTHONPATH=$SPARK_HOME/python:$PYTHONPATH" >> ~/.bashrc
echo "export PYSPARK_DRIVER-PYTHON=jupyter" >> ~/.bashrc
echo "export PYSPARK_DRIVER_PYTHON_OPTS=notebook" >> ~/.bashrc
echo "export PYSPARK_PYTHON=python3" >> ~/.bashrc
```

Important

Please do not forget to reload your interactive shell session. [1]

---

[1] https://www.delftstack.com/howto/linux/reload-bashrc/

By Carlos Montemuiño