Guía de instalación de Apache Spark

Requisitos recomendados:

VirtualBox 56.1

ISO de Ubuntu

A continuación te explicaré paso a paso como instalar Apache Spark, recomiendo que la instalación l realices dentro de una maquina virtual por dos motivos:

* Actuará como equipo dedicado, separando tus datos privados de los datos que usaremos a lo largo del curso.
* En caso de un error de instalación que resulte fatal, no perderás información personal o tiempo manejando tus respaldos.

Finalmente, aun que nunca es recomendable usar ambientes que modifiquen el estado de tu máquina, y mas si solo será por unas cuantas horas, puedes realizar la instalación en tu equipo personal, pero te recomiendo que hagas un respaldo del estado actual de tu equipo.

Finalmente, la instalación esta pensada para hacerse en un sistema basado en Linux, en el caso particular de este manual será en Ubuntu debido a que es el sabor de Linux mas popular. Sin embargo si gustas usar otra versión de Linux siéntete libre de hacerlo ya que solo deberían cambiar los comandos de descarga.

Sin mas advertencias, ¡manos a la obra!

Lo primero que necesitas hacer es instalar la maquina virtual de java 8, scala y asegurar que tienes instalado python3 y pip3, esto lo hacemos de la siguiente forma:

Agregaremos el repositorio que permite instalar el JDK 8 y actualizamos el sistema para sincronizar los paquetes:

sudo add-apt-respository ppa:openjdk-r/ppa

sudo apt-get -y update

sudo apt-get -y upgrade

Posteriormente instalamos Java 8, Scala, Python3 y pip3

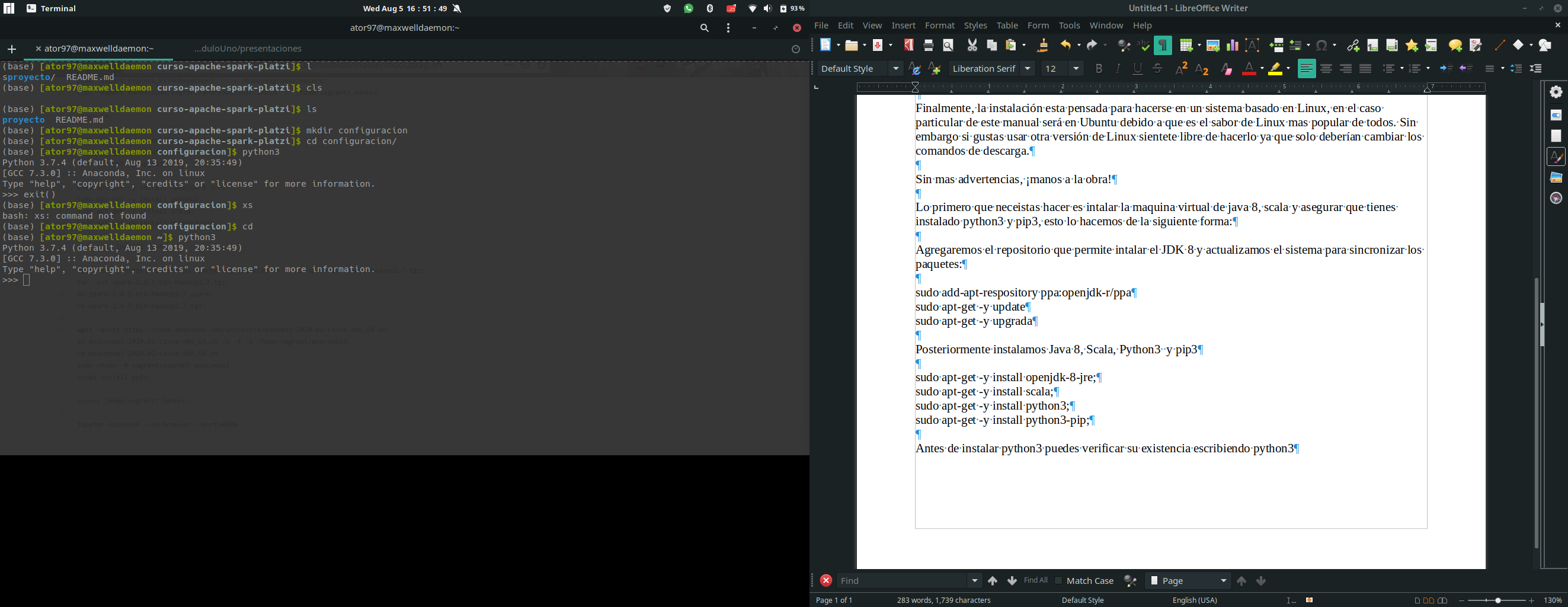
sudo apt-get -y install openjdk-8-jre;

sudo apt-get -y install scala;

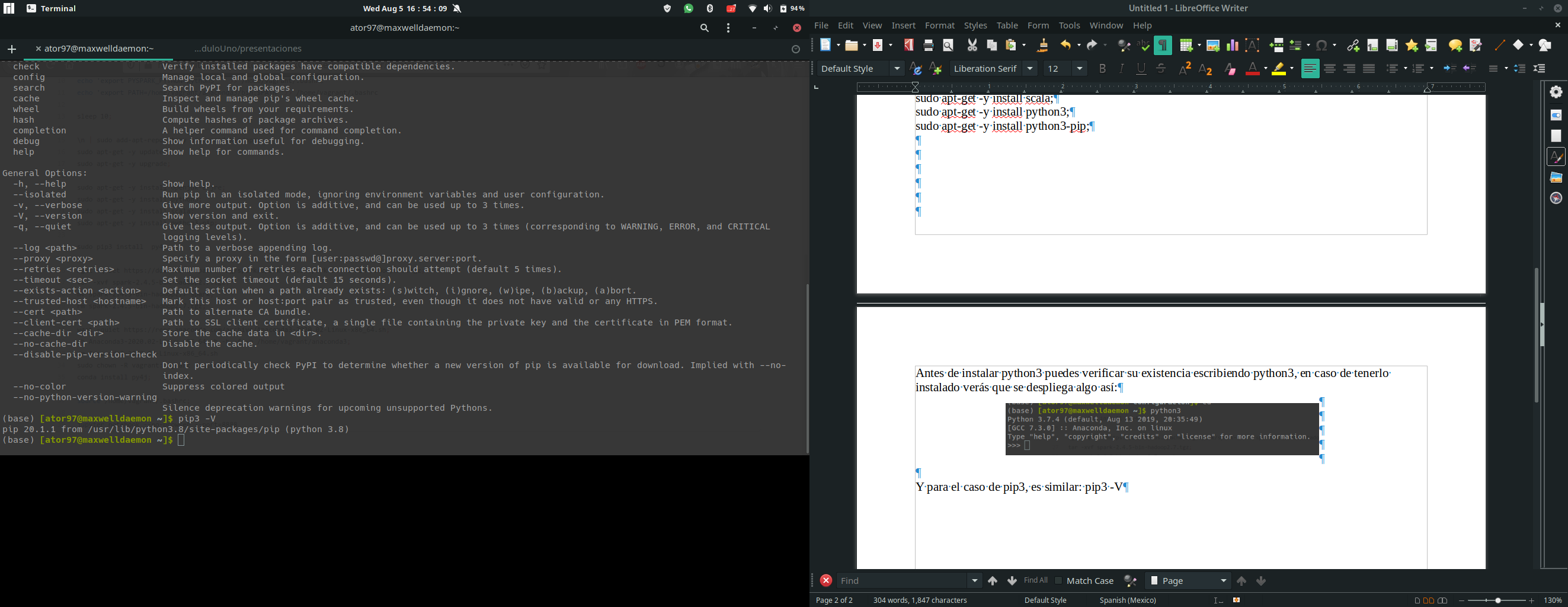
sudo apt-get -y install python3;

sudo apt-get -y install python3-pip;

Antes de instalar python3 puedes verificar su existencia escribiendo python3, en caso de tenerlo instalado verás que se despliega algo así:



Y para el caso de pip3, deberás escribir pip3 -V



En caso de que no los poseas instalados, regresa al paso anterior e instalalos. Los cuatro paquetes que instalarás deberían tardar solo unos segundos.

—

Una vez que tienes instalados los paquetes. Deberás instalar py4j, este módulo permite traducir código de Python a código máquina de Java, esto es debido a que aun que vamos a trabajar con Python a lo largo del curso, Spark solo es capaz de comprender código máquina de Java, por lo que requerimos un traductor para poder interactuar desde Python con Spark:

sudo pip3 install py4j;

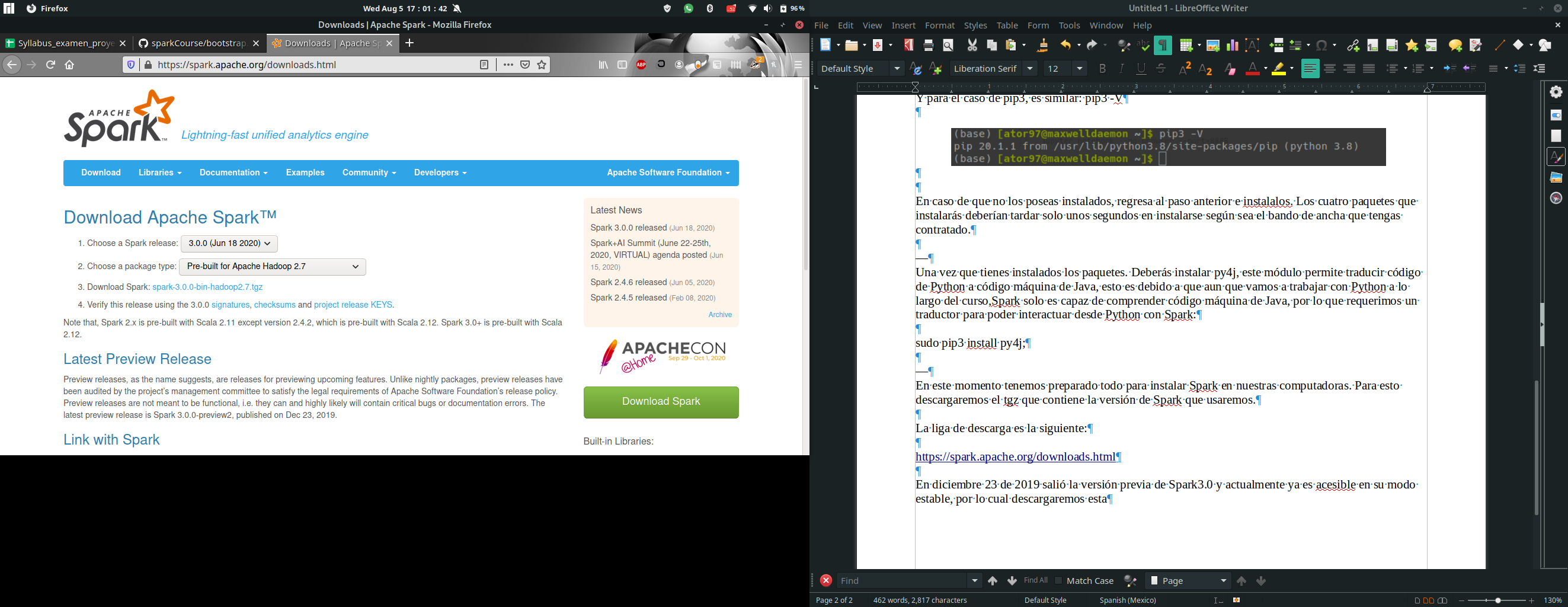
—

En este momento tenemos preparado todo para instalar Spark en nuestras computadoras. Para esto descargaremos el tgz que contiene la versión de Spark que usaremos.

La liga de descarga es la siguiente:

https://spark.apache.org/downloads.html

En diciembre 23 de 2019 salió la versión previa de Spark3.0 y actualmente ya es acesible en su modo estable, por lo cual descargaremos esta:



Spark al ser un sistema distribuido orientado a trabajar en cluster puede ser instalado de varias formas:

Standalone

con Yarn (Hadoop)

con Mesos

Nosotros usaremos a lo largo del curso la versión Standalone. Aparte de simplificar la instalación, esto será suficiente para el contenido que veremos a lo largo del curso ya que los conceptos que absorverás son los mismos pero en la practica profesional solo contarás con mas capacidad de computo y tareas distribuidas distribuidas de forma real.

Para evitar hacer la descarga con interfaz grafica, podemos hacer lo siguiente:

wget -quiet <https://www.apache.org/dyn/closer.lua/spark/spark-3.0.0/spark-3.0.0-bin-hadoop2.7.tgz>

tar xvf spark3.0.0-bin-hadoop2-7.tgz;

mkdir ~/spark

mv spark-3.0.0-bin-hadoop2.7 ~/spark

rm spark-23.0.0 -bin-hadoop2.7.tgz;

Con esto descargaremos, descomprimiremos y pondemos en una ruta deseada el archivo, para fines de este tutorial, el paradero será en la carperta llamada spark dentro de la carpeta home de nuestro usuario.

Debido a que usaremos anaconda a lo largo del curso, realizaremos una instalación parecida :

wget –quiet <https://repo.anaconda.com/archive/Anaconda3-2020.02-Linux-x86_64.sh>;

sh Anaconda3-2020.02-Linux-x86\_64.sh -b -f -p ~/anaconda3;

rm Anaconda3-2020.02-Linux-x86\_64.sh

sudo chown -R **userName**:**userGroup** anaconda3

NOTA: No olvides cambiar el userName y userGroup por los datos que posees en tu equipo personal.

Del mismo modo instalaremos en anaconda py4j

conda install py4j

Para finalizar. Le avisaremos al sistema operativo donde ubicar Spark y Anaconda modificando el archivo ~/.bashrc, el cual posee las configuraciones de nuestro usuario, abreglo con tu editor de texto favorito y al final del archivo agrega las siguientes líneas de codigo:

export JAVA\_HOME="/usr/lib/jvm/java-8-openjdk-amd64"

export PATH=$JAVA\_HOME/bin:$PATH

export SPARK\_HOME="~/spark"

export PATH=$SPARK\_HOME:$PATH

export PYTHONPATH=$SPARK\_HOME/python:$PYTHONPATH

export PYSPARK\_DRIVER\_PYTHON="jupyter"

export PYSPARK\_DRIVER\_PYTHON\_OPTS="notebook"

export PYSPARK\_PYTHON=python3

export PATH=/home/vagrant/anaconda3/bin:$PATH

Como tal vez sepas, los cambios realizados en el archivo ~/.bashrc no se ejecutan al momento por lo cual deberás recargarlo con el siguiente comando:

source ~/.barshc;

Con esto tienes instalado todo lo que necesitas para llevar este curso.

NOTA: Para poder iniciar jupyter usa el siguiente comando:

jupyter notebook --no-browser –port=8888

Te mostrará una url a la cual cuando le des click te llevará a una pestaña de Juptyter

En el caso que desees automatizar la instalación, puedes usar la plantilla que te dejo acontinuación:

https://github.com/terranigmark/curso-apache-spark-platzi

Para automatizarlo, deberás instalar vagrant y virtual box. En el readme del github podrás ver como lanzar la maquina virutal paso a paso y replicarla donde sea