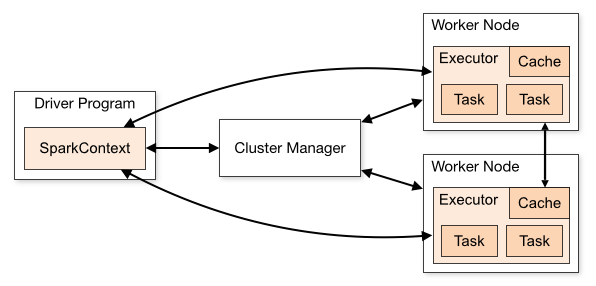
**PySpark Documentation**

Master-Slave architecture

Master 🡪 Driver 🡪 creates a context that is an entry point to your application

Slaves 🡪 Workers 🡪 execute all operations (transformations and actions)



Cluster manager 🡪 standalone, apache mesos, Hadoop YARN, kubernetes, local

use “local” for master() in order to run Spark on your laptop/computer

**Installing Apache PySpark on Ubuntu**

Software needed 🡪 Python, Java, Apache Spark

1. Install python:

sudo apt-get install python3

python3 -version

1. Install pip:

sudo apt-get install python3-pip

1. Install jupyter:

pip3 install jupyter

To make sure not getting error while launching jupyter notebook:

export PATH=$PATH:~/.local/bin

You can launch jupyter notebook by typing:

jupyter notebook

1. Update ubuntu OS:

sudo apt-get update

1. Install java:

sudo apt-get install default-jre

java -version

1. Install scala:

sudo apt-get install scala

scala -version

1. Install py4j:

pip3 install py4j

1. Download Apache Spark:

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

Pindahkan file zip spark-3.2.1-bin-hadoop3.2.tgz ke /home

unzip file:

sudo tar -zxvf spark-3.2.1-bin-hadoop3.2.tgz

1. Connect Apache Spark with Jupyter notebook python by making environment variables:

export SPARK\_HOME='/home/ubuntu/spark-3.2.1-bin-hadoop3.2'

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

1. Launch jupyter notebook from directory ~/spark-3.2.1-bin-hadoop3.2/python:

Open terminal

cd ~/spark-3.2.1-bin-hadoop3.2/python

jupyter notebook

1. If there is error about access denied, then make sure folder is yours, change owner:

sudo chown -R userName:groupName ~/.local/share/jupyter

1. Additional, install pandas:

pip3 install pandas

**Installing Apache PySpark on Windows (CONFUSING!)**

Software needed 🡪 Python, Java, Apache Spark

1. Install python or anaconda package.
2. Install Java:

https://www.oracle.com/java/technologies/downloads/#java8-windows

after installation, set environment variables:

JAVA\_HOME = C:\Program Files\Java\jdk1.8.0\_321

PATH = %PATH%;C:\Program Files\Java\jdk1.8.0\_321\bin

In windows, to make environment variables 🡪 system 🡪 advanced system setting 🡪 advanced 🡪 environment variables 🡪 create new for environment variable HOME, edit add Path for PATH

1. Install apache spark:

https://www.apache.org/dyn/closer.lua/spark/spark-3.2.1/spark-3.2.1-bin-hadoop3.2.tgz

extract the zip folder to C:\apps

after extraction, set:

SPARK\_HOME = C:\apps\spark-3.2.1-bin-hadoop3.2

HADOOP\_HOME = C:\apps\spark-3.2.1-bin-hadoop3.2

PATH=%PATH%;C:\apps\spark-3.2.1-bin-hadoop3.2\bin

1. Download winutils.exe:

https://github.com/cdarlint/winutils/blob/master/hadoop-3.2.2/bin/winutils.exe

1. Install winutils.exe:

https://github.com/cdarlint/winutils/blob/master/hadoop-3.2.2/bin/winutils.exe

1. Run pyspark at cmd:

pyspark

it will run a spark shell terminal

spark web GUI can be open at: http://localhost:4040/