BIG DATA

PySpark - HandOn

K V Subramaniam
Computer Science and Engineering

Overview of lecture

- What is PySpark
- Installation
- PySpark Architecture
- Word count with PySpark and Scala

What is PySpark

Spark Programming

We looked previously at Scala as a language to program Spark

- But Spark also support a python binding

- Write code in python
  - With additional spark transformations/actions
  - Program runs as a Spark job

Interactive and Batch processing

- Supports the pyspark shell for interactive processing
- And regular batch processing jobs can be run by writing pyspark scripts

PySpark configuration

Downloading and Setting up pyspark

- Download and untar the spark tar file from the spark repository
- Pyspark is bundled along with spark
- However, it requires some configuration
  - Setup of proper paths.

Pyspark configuration

- Needs environment variables to be setup
- First modify .bashrc to include

```
export SPARK_HOME = /home/hadoop/spark-2.1.0-bin-hadoop2.7
export PATH = $PATH:/home/hadoop/spark-2.1.0-bin-hadoop2.7/bin
export PYTHONPATH = $SPARK_HOME/python:$SPARK_HOME/python/lib/py4j-0.10.4-src.zip:$PYTHONPATH
export PATH = $SPARK_HOME/python:$PATH
```

Then run so that the env variables are setup

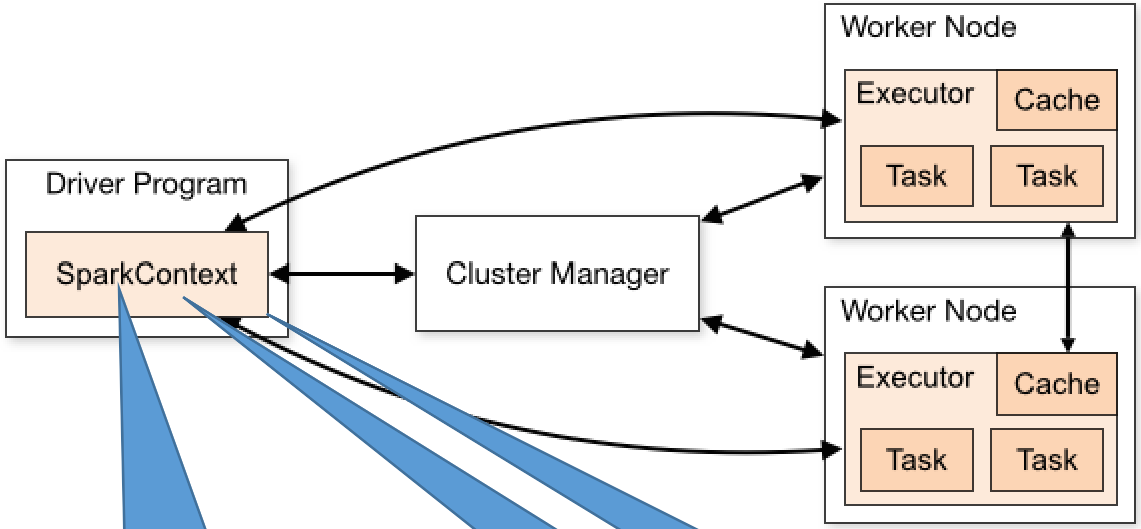
```
# source .bashrc
```

Starting pyspark

- Needs environment variables to be setup
- First modify .bashrc to include

```
./bin/pyspark
```

Recall: The Spark Context



This is the master for your spark application.

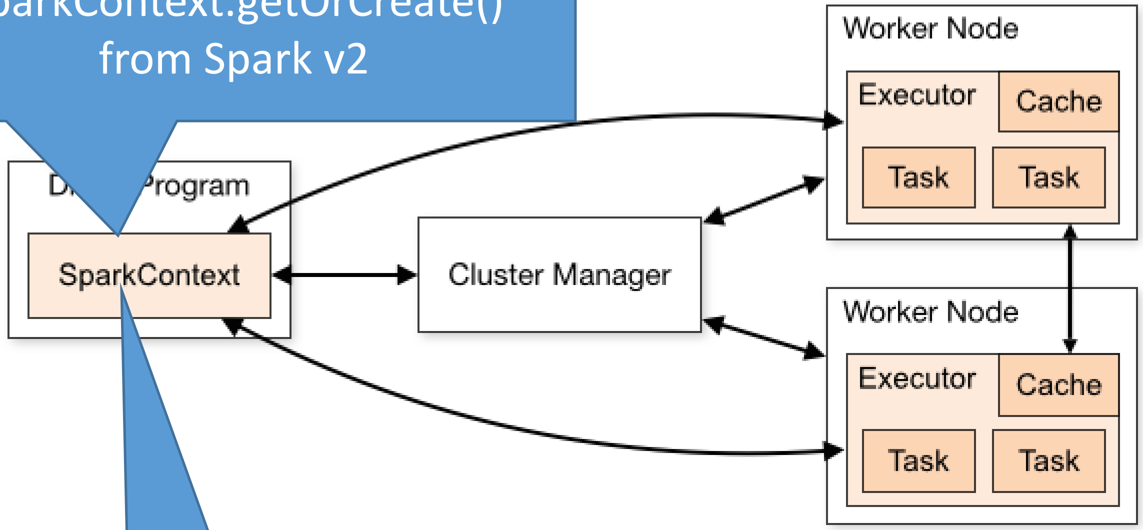
Starting and stopping/cancelling a job.

Interacts with cluster manager like YARN

https://spark.apache.org/docs/latest/cluster-overview.html

When is Spark Context created?

Programmatically : SparkContext.getOrCreate() from Spark v2



Interactively: When you start spark-shell

There is one SparkContext per JVM

https://spark.apache.org/docs/latest/cluster-overview.html

Pyspark architecture

Pyspark architecture

- So if the spark context is maintained per JVM

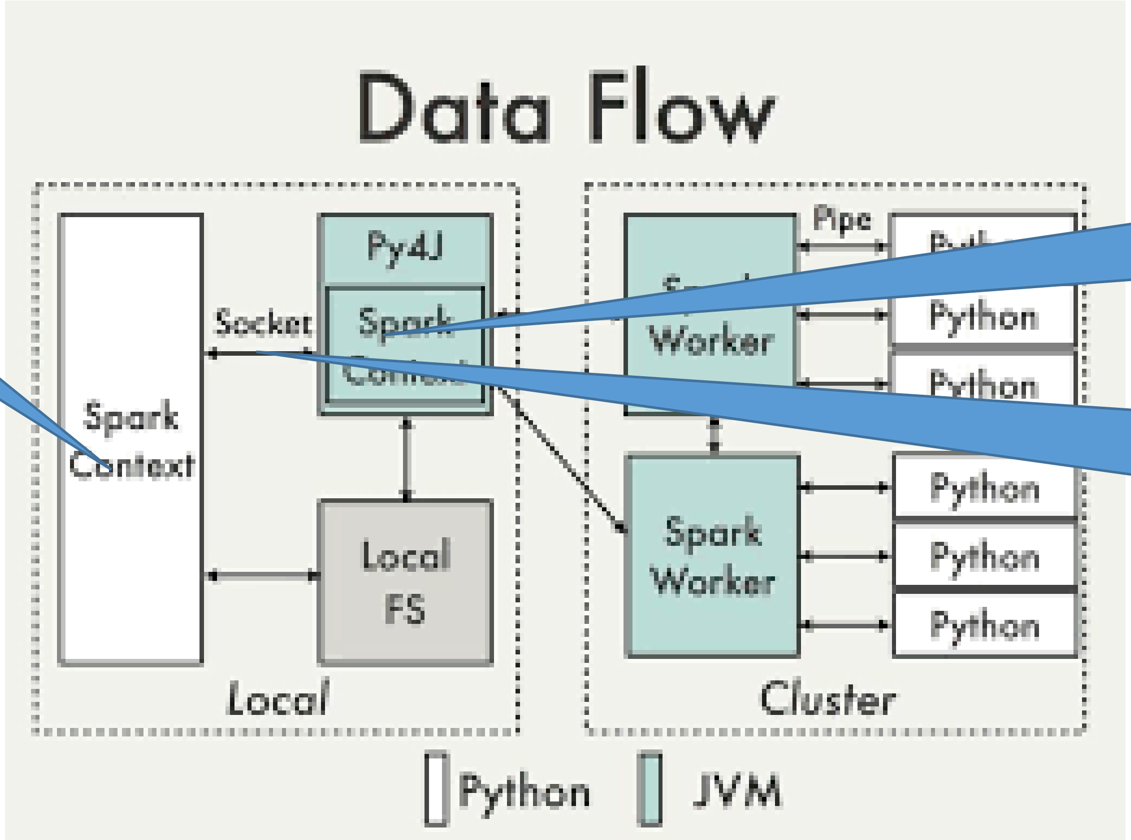
- How does pyspark take care of the SparkContext?
  - Who creates and maintains this?

py4j

- Acts like a bridge between python and java
- Allows python interpreter to access Java objects instantiated within the JVM
- Can invoke methods on the Java objects as if they were python methods

https://www.py4j.org/

Pyspark architecture



Python driver

Jave SparkContext used to drive the application

Data has to be serialized using cPickle serializer

https://cwiki.apache.org/confluence/display/SPARK/PySpark+Internals

PySpark Demo

THANK YOU

K V Subramaniam, Usha Devi

Dept. of Computer Science and Engineering

subramaniamkv@pes.edu

ushadevibg@pes.edu