Distribution and Instrumentation

Justin Pihony

@JustinPihony | justin-pihony.blogspot.com

Course Overview

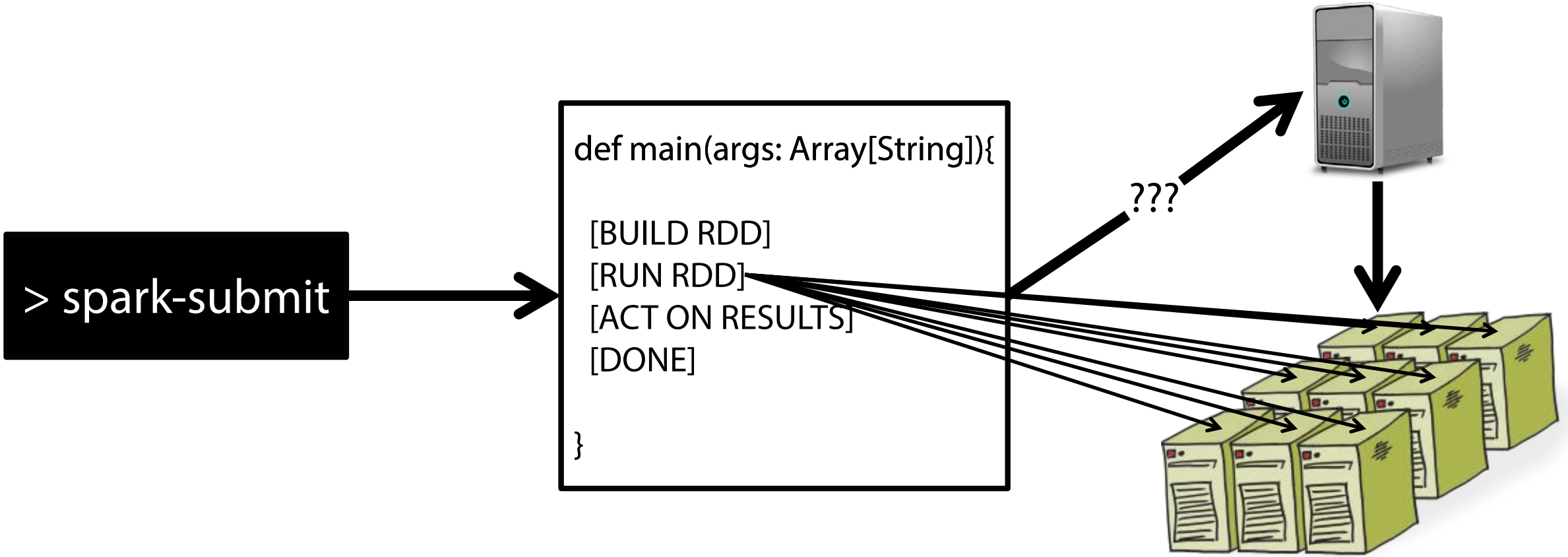
- Basics of Spark
- Core API
- Cluster Managers
- Spark Maintenance
- Libraries
  - SQL
  - Streaming
  - MLlib/GraphX
- Troubleshooting / Optimization
- Future of Spark

Section Overview

- Cluster Managers
- Spark Maintenance
- Cluster Managers
- Spark Maintenance

- Libraries
  - SQL
  - Streaming
  - MLlib/GraphX
- Troubleshooting / Optimization
- Future of Spark

spark-submit



Cluster Managers

Cluster Managers

Cluster Managers

--master yarn-cluster

HADOOP/YARN_CONF_DIR

client/cluster

--num-executors 2

--executor-cores 1

--queues default

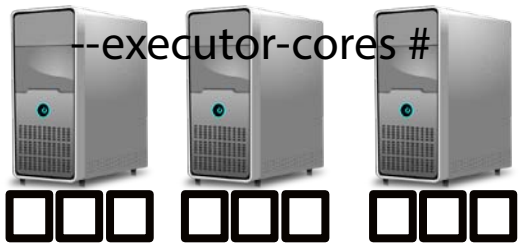
--master spark://[HOST]:7077

client/cluster

spark.deploy.spreadOut=false

--total-executor-cores #

--executor-cores #



--master mesos://[HOST]:5050

client/(cluster)

spark.mesos.coarse=false

--total-executor-cores #

~~spreadOut~~

Spark Standalone

$SPARK_HOME

conf/slaves → [SLAVE_ADDRESS_1]
[SLAVE_ADDRESS_...]
[SLAVE_ADDRESS_N]

conf/spark-env.sh → export SPARK_MASTER_OPTS="-Dspark.deploy.defaultCores=<value>"

```
> ./sbin/start-all.sh
```

SSH

```
> bin/spark-class org.apache.spark.deploy.master.Master
```

```
> bin/spark-class org.apache.spark.deploy.worker.Worker
  spark://[MASTER]:7077
```

https://spark.apache.org/**docs/latest/spark-standalone.html**

Resources

- https://open.mesosphere.com/tutorials/run-spark-on-mesos/
- AWS
  - Pluralsight: Big Data on Amazon Web Services/AWS Developer Fundamentals
  - https://aws.amazon.com
    - /documentation/ec2/
    - /ec2/spot/
    - /elasticmapreduce/details/spark/
- http://bit.ly/1NiEsWS : Understanding your Spark application through visualization
- https://spark.apache.org/docs/latest/configuration.html

Summary

- spark-submit
- Clusters
  - StandAlone
  - YARN
  - Mesos
- Spark on EMR
- Spark UI