# Spark on AWS

Taryn Heilman
March 15, 2018

galvanize

# Learning Objectives

- **Understand spark architecture and ways to deploy a spark cluster**
- **Deploy a spark cluster on AWS and use this for machine learning**
- **Monitor your spark cluster through remote UI**

# Review

Why is spark faster than Hadoop MapReduce?

What are the advantages to using spark DataFrames over RDDs? Are there any reasons you would want to use an RDD instead?

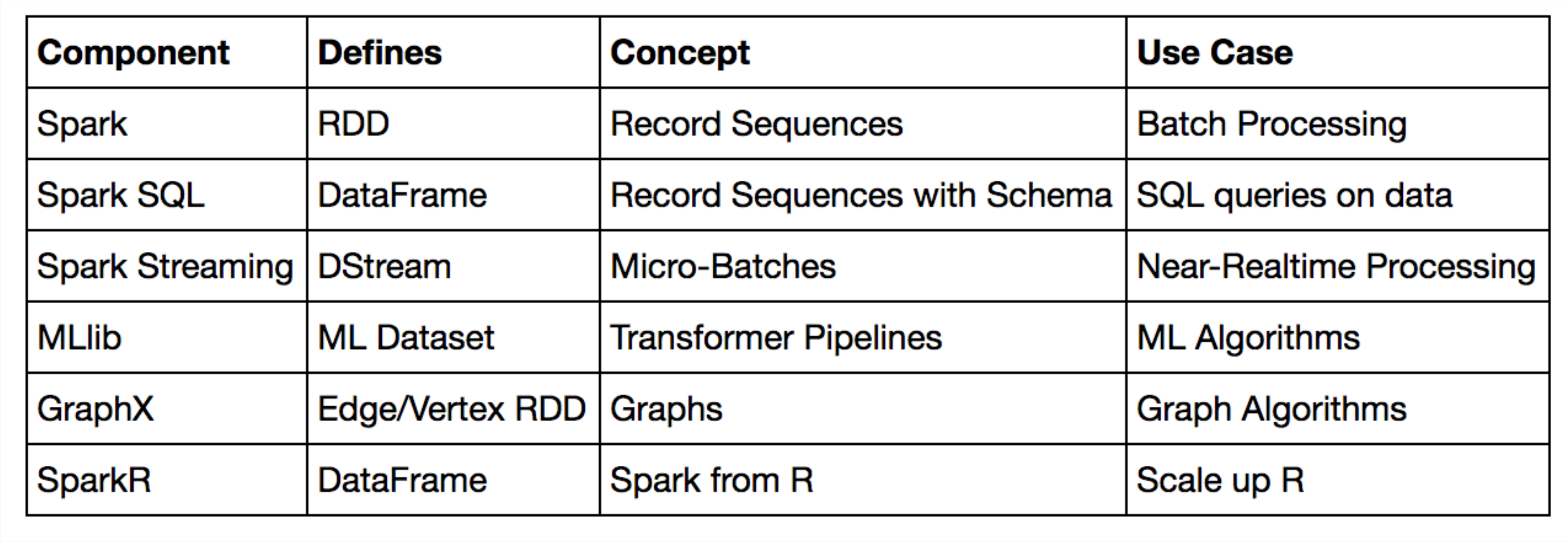
Describe the procedure for implementing a udf in pyspark

What does it mean to partition your data in spark? Why is this a good idea?

Describe the KMeans algorithm in 2-3 sentences

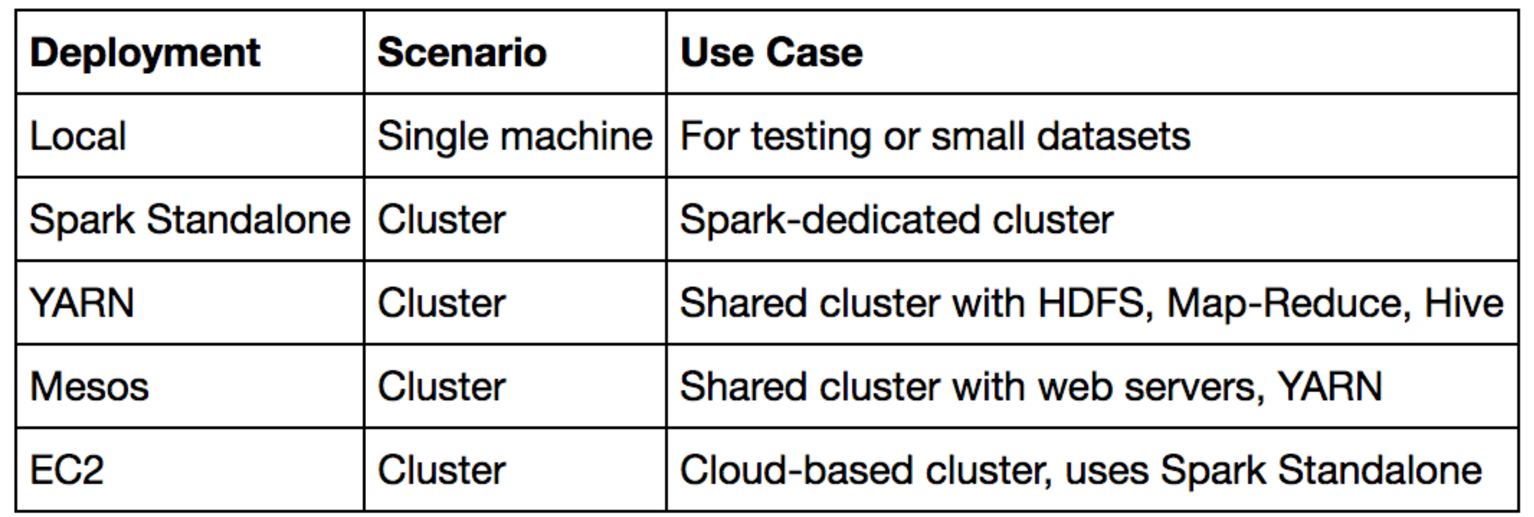
What does TF-IDF stand for? Why do we use this over a simple term frequency matrix?

# Major Spark Components



| Component | Defines | Concept | Use Case |
|---|---|---|---|
| Spark | RDD | Record Sequences | Batch Processing |
| Spark SQL | DataFrame | Record Sequences with Schema | SQL queries on data |
| Spark Streaming | DStream | Micro-Batches | Near-Realtime Processing |
| MLlib | ML Dataset | Transformer Pipelines | ML Algorithms |
| GraphX | Edge/Vertex RDD | Graphs | Graph Algorithms |
| SparkR | DataFrame | Spark from R | Scale up R |

# Ways to deploy Spark



| Deployment | Scenario | Use Case |
|---|---|---|
| Local | Single machine | For testing or small datasets |
| Spark Standalone | Cluster | Spark-dedicated cluster |
| YARN | Cluster | Shared cluster with HDFS, Map-Reduce, Hive |
| Mesos | Cluster | Shared cluster with web servers, YARN |
| EC2 | Cluster | Cloud-based cluster, uses Spark Standalone |

- YARN is the most popular configuration in production environments.
- YARN is used for on-prem while Spark Standalone is used for cloud hosting.
- Spark Standalone is good for demo, proof-of-concept, or testing.
- Spark Standalone is also used on EC2 for cloud-based clusters.

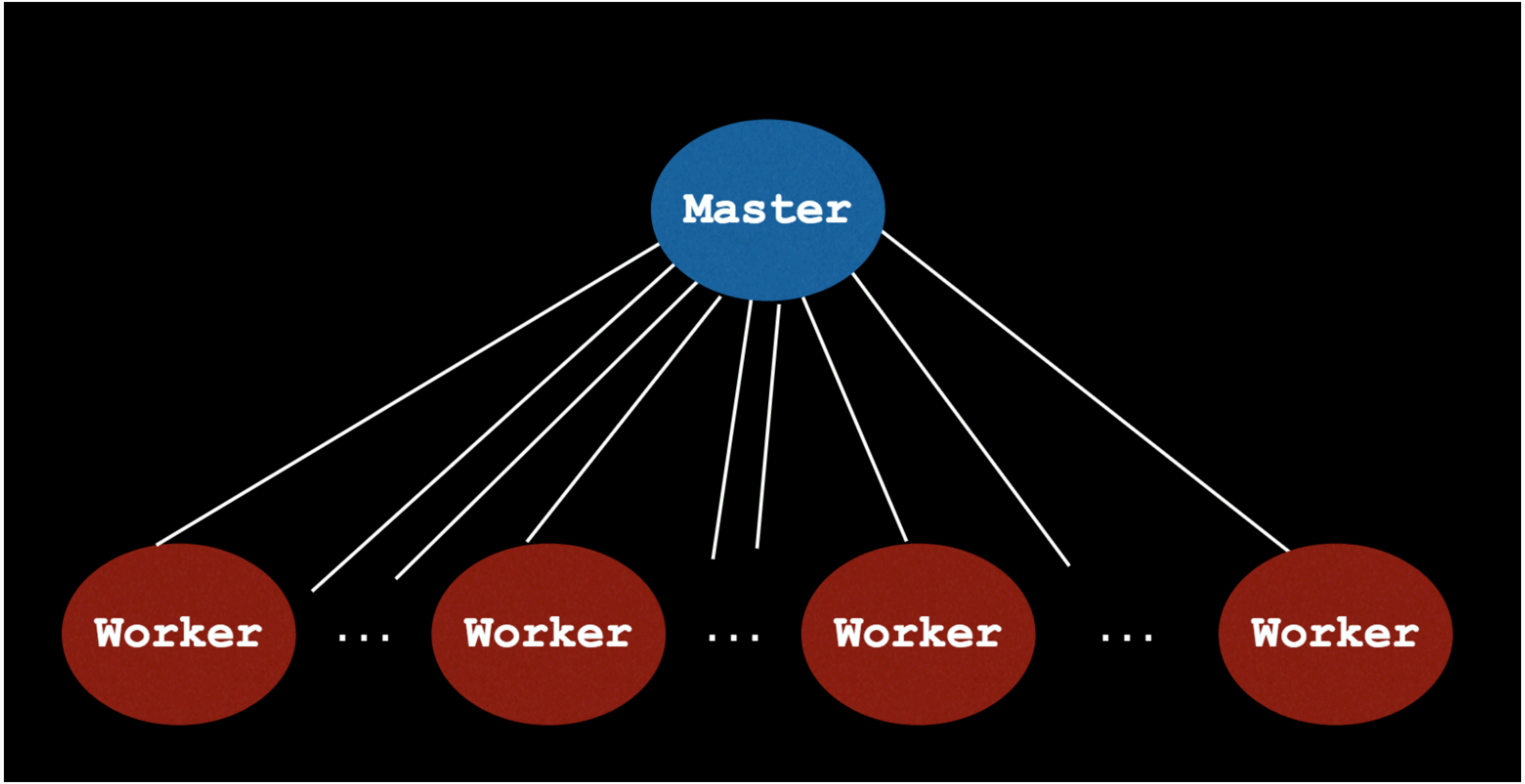
# Discussion

Is there an advantage to using spark on a local machine (to do some process, as opposed to doing the process in native python, e.g.)?

What are some disadvantages to using spark with small datasets? 💬

# Spark on AWS

Create a cluster on amazon EMR with a Hadoop file system, using YARN



# Cluster actions

- Launch
  - Start a new cluster with specified settings

- Terminate
  - Permanently stops the cluster. You will not be charged for any further compute or storage, and you will not be able to use this cluster again

- Stop
  - Stops the cluster without terminating.
  - Quicker to restart than to launch a new one.
  - You will be charged for storage for a stopped cluster (much less than compute time, though)

- Restart
  - Restart a stopped cluster.

# Cluster settings

You'll need to choose number of cores and allocate memory to driver and executors

Driver is the master node, executors are workers/slaves. General rule of thumb is that they should have the same amount of memory.

If you don't select enough cores, your job will be slow. If you don't select enough memory, your JVM will crash. (Note: the .toPandas() function is exceptionally memory intensive. Use sparingly!)

# Spark Application UI

View metrics through the UI to find out:

- How many worker nodes do I have?
- How many executors do I have?
- How many partitions is my data split into?
- Why is it taking so long?

The application UI serves up metrics through a web UI on port 4040 on the machine your driver is running on.

Try it now: http://localhost:4040.

# Spark Application UI

Live demo

# Using tmux

Live demo….

# Recap: Learning Objectives

- **Understand spark architecture and ways to deploy a spark cluster**
- **Deploy a spark cluster on AWS and use this for machine learning**
- **Monitor your spark cluster through remote UI**

# Questions?

Warning: this assignment is a little long! The first part involves you setting up your own spark cluster on AWS using several different methods, will need to follow the directions exactly for this to work. Afterwards, you will use spark ML to do some machine learning on a dataset.

*NOTE - I have made some changes to this repo to update to the latest version of spark. There are minor so you don’t necessarily need to re-download, but you should view branch spark-2.2 on the web to see the updates