Acquire, Install, Config, Run

Acquire

- Open Source Spark
  - http://spark.apache.org/downloads.html
- Databricks Cloud
  - https://databricks.com/product/databricks
- Cloudera Distribution
  - https://www.cloudera.com/content/www/en-us/products/apache-hadoop/apache-spark.html
- Datastax Distribution
  - http://www.datastax.com/products/datastax-enterprise
- MapR Distribution
  - https://www.mapr.com/products/apache-spark
- Hortonworks Distribution
  - http://hortonworks.com/hadoop/spark/

Install

- Runs on Windows and Unix like systems
- Prereqs: Java 1.7+, Python 2.6+, R 3.1+, Scala 2.10 (can be built from source for Scala 2.11)
- Prebuilt Downloads
  - Prebuilt with specific versions of Hadoop
    - Spark reuses Hadoop jars for HDFS and YARN
  - Prebuilt Hadoop-free as well
    - Need to tell Spark where your Hadoop jars are
  - Just unzip the tar ball
- Source code downloads
  - Can be used to build with your own version of Hadoop
  - Can be built with Maven or SBT
  - http://spark.apache.org/docs/latest/building-spark.html
    - Build for your own Hadoop
    - Build with Hive support
    - Build with IDEs (IntelliJ and Eclipse)

Config

- <install_dir>/conf contains config files
  - spark-env.sh
    - Environment variables – Java, Python, R VM locations
    - Spark Standalone settings
  - spark-defaults.conf
    - Default system properties for configuring Spark applications
    - Can be specified at a node level
  - slaves
    - Worker node hostnames
  - log4j.properties
    - Server log configuration

Run

* REPLs
  * <install_dir>/bin
    * spark-shell – Scala REPL
    * pyspark – Python REPL
    * sparkR – R REPL
    * spark-sql – SQL REPL
* Submitting applications
  * <install_dir>/bin/spark-submit