



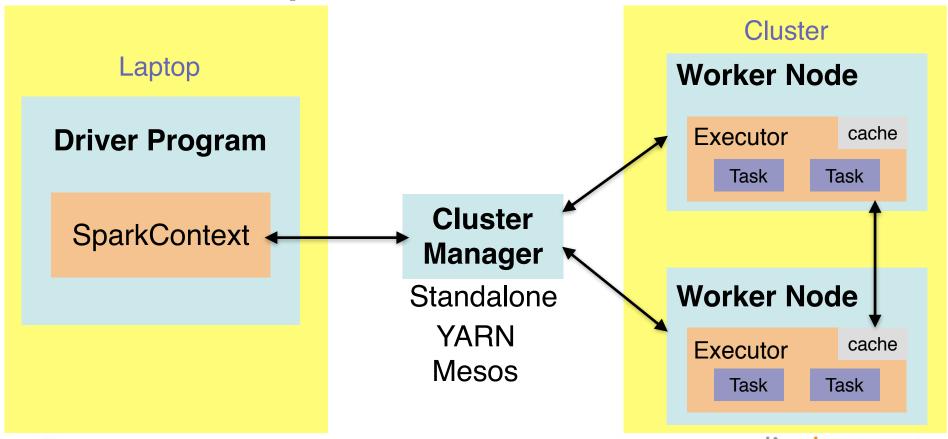
Lesson 4: Spark Internals

4.7 Spark Deployment: Local and Cluster Modes





Spark Execution Context





©2016 Pearson, Inc.

Spark Deployment

Local Mode

Cluster Mode

• Single threaded:

SparkContext('local')

Multi-threaded:

SparkContext('local[4]')

Pseudo-distributed cluster

Standalone

Mesos

YARN

• Amazon EC2





Spark Deployment: Local

Mode	Advantage
Single threaded	sequential execution allows easier debugging of program logic
Multi-threaded	concurrent execution leverages parallelism and allows debugging of coordination
Pseudo-distributed cluster	distributed execution allows debugging of communication and I/O





Standalone

- Packaged with Spark core
- Great if all you need is a dedicated Spark cluster
- Doesn't support integration with any other applications on a cluster.

The Standalone cluster manager also has a high-availability mode that can leverage Apache ZooKeeper to enable standby master nodes.





Mesos

- General purpose cluster and global resource manager (Spark, Hadoop, MPI, Cassandra, etc.)
- Two-level scheduler: enables pluggable scheduler algorithms
- Multiple applications can co-locate (like an operating system for a cluster)





YARN

- Created to scale Hadoop, optimized for Hadoop (stateless batch jobs with long runtimes)
- Monolithic scheduler: manages cluster resources as well as schedules jobs
- Not well suited for long-running, real-time, or stateful/interactive services (like database queries)





EC2

- Launch scripts bundled with Spark
- Elastic and ephemeral cluster
- Sets up:
 - Spark
 - HDFS
 - Hadoop MR





Spark Deployment: Cluster

Mode	Advantage
Standalone	Encapsulated cluster manager isolates complexity
Mesos	global resource manager facilitates multi- tenant and heterogeneous workloads
YARN	Integrates with existing Hadoop cluster and applications
EC2	elastic scalability and ease of setup



