

Lesson 4: Spark Internals

4.7 Spark Deployment: Local and Cluster Modes

Spark Execution Context

Laptop

Driver Program

SparkContext

Cluster Manager

Standalone
YARN
Mesos

Cluster

Worker Node

Executor

cache

Task

Task

Worker Node

Executor

cache

Task

Task



Spark Deployment

Local Mode

- Single threaded:
`SparkContext('local')`
- Multi-threaded:
`SparkContext('local[4]')`
- Pseudo-distributed cluster

Cluster Mode

- Standalone
- Mesos
- YARN
- Amazon EC2



Spark Deployment: Local

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



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

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

