# **Spark Overview**

Apache Spark is a fast and general-purpose cluster computing system. It provides high-level APIs in Java, Scala, Python and R, and an optimized engine that supports general execution graphs. It also supports a rich set of higher-level tools including [Spark SQL](http://spark.apache.org/docs/latest/sql-programming-guide.html) for SQL and structured data processing, [MLlib](http://spark.apache.org/docs/latest/mllib-guide.html) for machine learning, [GraphX](http://spark.apache.org/docs/latest/graphx-programming-guide.html) for graph processing, and [Spark Streaming](http://spark.apache.org/docs/latest/streaming-programming-guide.html).

# **Downloading**

Get Spark from the [downloads page](http://spark.apache.org/downloads.html) of the project website. This documentation is for Spark version 1.6.2. Spark uses Hadoop’s client libraries for HDFS and YARN. Downloads are pre-packaged for a handful of popular Hadoop versions. Users can also download a “Hadoop free” binary and run Spark with any Hadoop version [by augmenting Spark’s classpath](http://spark.apache.org/docs/latest/hadoop-provided.html).

If you’d like to build Spark from source, visit [Building Spark](http://spark.apache.org/docs/latest/building-spark.html).

Spark runs on both Windows and UNIX-like systems (e.g. Linux, Mac OS). It’s easy to run locally on one machine — all you need is to have javainstalled on your system PATH, or the JAVA\_HOME environment variable pointing to a Java installation.

Spark runs on Java 7+, Python 2.6+ and R 3.1+. For the Scala API, Spark 1.6.2 uses Scala 2.10. You will need to use a compatible Scala version (2.10.x).

# **Running the Examples and Shell**

Spark comes with several sample programs. Scala, Java, Python and R examples are in the examples/src/main directory. To run one of the Java or Scala sample programs, use bin/run-example <class> [params] in the top-level Spark directory. (Behind the scenes, this invokes the more general [spark-submit script](http://spark.apache.org/docs/latest/submitting-applications.html) for launching applications). For example,

./bin/run-example SparkPi 10

You can also run Spark interactively through a modified version of the Scala shell. This is a great way to learn the framework.

./bin/spark-shell --master local[2]

The --master option specifies the [master URL for a distributed cluster](http://spark.apache.org/docs/latest/submitting-applications.html" \l "master-urls), or local to run locally with one thread, or local[N] to run locally with N threads. You should start by using local for testing. For a full list of options, run Spark shell with the --help option.

Spark also provides a Python API. To run Spark interactively in a Python interpreter, use bin/pyspark:

./bin/pyspark --master local[2]

Example applications are also provided in Python. For example,

./bin/spark-submit examples/src/main/python/pi.py 10

Spark also provides an experimental [R API](http://spark.apache.org/docs/latest/sparkr.html) since 1.4 (only DataFrames APIs included). To run Spark interactively in a R interpreter, usebin/sparkR:

./bin/sparkR --master local[2]

Example applications are also provided in R. For example,

./bin/spark-submit examples/src/main/r/dataframe.R

# **Launching on a Cluster**

The Spark [cluster mode overview](http://spark.apache.org/docs/latest/cluster-overview.html) explains the key concepts in running on a cluster. Spark can run both by itself, or over several existing cluster managers. It currently provides several options for deployment:

* [Amazon EC2](http://spark.apache.org/docs/latest/ec2-scripts.html): our EC2 scripts let you launch a cluster in about 5 minutes
* [Standalone Deploy Mode](http://spark.apache.org/docs/latest/spark-standalone.html): simplest way to deploy Spark on a private cluster
* [Apache Mesos](http://spark.apache.org/docs/latest/running-on-mesos.html)
* [Hadoop YARN](http://spark.apache.org/docs/latest/running-on-yarn.html)

# **Where to Go from Here**

****Programming Guides:****

* [Quick Start](http://spark.apache.org/docs/latest/quick-start.html): a quick introduction to the Spark API; start here!
* [Spark Programming Guide](http://spark.apache.org/docs/latest/programming-guide.html): detailed overview of Spark in all supported languages (Scala, Java, Python, R)
* Modules built on Spark:
  + [Spark Streaming](http://spark.apache.org/docs/latest/streaming-programming-guide.html): processing real-time data streams
  + [Spark SQL, Datasets, and DataFrames](http://spark.apache.org/docs/latest/sql-programming-guide.html): support for structured data and relational queries
  + [MLlib](http://spark.apache.org/docs/latest/mllib-guide.html): built-in machine learning library
  + [GraphX](http://spark.apache.org/docs/latest/graphx-programming-guide.html): Spark’s new API for graph processing

****API Docs:****

* [Spark Scala API (Scaladoc)](http://spark.apache.org/docs/latest/api/scala/index.html" \l "org.apache.spark.package)
* [Spark Java API (Javadoc)](http://spark.apache.org/docs/latest/api/java/index.html)
* [Spark Python API (Sphinx)](http://spark.apache.org/docs/latest/api/python/index.html)
* [Spark R API (Roxygen2)](http://spark.apache.org/docs/latest/api/R/index.html)

****Deployment Guides:****

* [Cluster Overview](http://spark.apache.org/docs/latest/cluster-overview.html): overview of concepts and components when running on a cluster
* [Submitting Applications](http://spark.apache.org/docs/latest/submitting-applications.html): packaging and deploying applications
* Deployment modes:
  + [Amazon EC2](http://spark.apache.org/docs/latest/ec2-scripts.html): scripts that let you launch a cluster on EC2 in about 5 minutes
  + [Standalone Deploy Mode](http://spark.apache.org/docs/latest/spark-standalone.html): launch a standalone cluster quickly without a third-party cluster manager
  + [Mesos](http://spark.apache.org/docs/latest/running-on-mesos.html): deploy a private cluster using [Apache Mesos](http://mesos.apache.org/)
  + [YARN](http://spark.apache.org/docs/latest/running-on-yarn.html): deploy Spark on top of Hadoop NextGen (YARN)

****Other Documents:****

* [Configuration](http://spark.apache.org/docs/latest/configuration.html): customize Spark via its configuration system
* [Monitoring](http://spark.apache.org/docs/latest/monitoring.html): track the behavior of your applications
* [Tuning Guide](http://spark.apache.org/docs/latest/tuning.html): best practices to optimize performance and memory use
* [Job Scheduling](http://spark.apache.org/docs/latest/job-scheduling.html): scheduling resources across and within Spark applications
* [Security](http://spark.apache.org/docs/latest/security.html): Spark security support
* [Hardware Provisioning](http://spark.apache.org/docs/latest/hardware-provisioning.html): recommendations for cluster hardware
* Integration with other storage systems:
  + [OpenStack Swift](http://spark.apache.org/docs/latest/storage-openstack-swift.html)
* [Building Spark](http://spark.apache.org/docs/latest/building-spark.html): build Spark using the Maven system
* [Contributing to Spark](https://cwiki.apache.org/confluence/display/SPARK/Contributing+to+Spark)
* [Supplemental Projects](https://cwiki.apache.org/confluence/display/SPARK/Supplemental+Spark+Projects): related third party Spark projects

****External Resources:****

* [Spark Homepage](http://spark.apache.org/)
* [Spark Wiki](https://cwiki.apache.org/confluence/display/SPARK)
* [Spark Community](http://spark.apache.org/community.html) resources, including local meetups
* [StackOverflow tag apache-spark](http://stackoverflow.com/questions/tagged/apache-spark)
* [Mailing Lists](http://spark.apache.org/mailing-lists.html): ask questions about Spark here
* [AMP Camps](http://ampcamp.berkeley.edu/): a series of training camps at UC Berkeley that featured talks and exercises about Spark, Spark Streaming, Mesos, and more.[Videos](http://ampcamp.berkeley.edu/3/), [slides](http://ampcamp.berkeley.edu/3/) and [exercises](http://ampcamp.berkeley.edu/3/exercises/) are available online for free.
* [Code Examples](http://spark.apache.org/examples.html): more are also available in the examples subfolder of Spark ([Scala](https://github.com/apache/spark/tree/master/examples/src/main/scala/org/apache/spark/examples), [Java](https://github.com/apache/spark/tree/master/examples/src/main/java/org/apache/spark/examples), [Python](https://github.com/apache/spark/tree/master/examples/src/main/python), [R](https://github.com/apache/spark/tree/master/examples/src/main/r))