# **Launching Spark on YARN**

确保HADOOP\_CONF\_DIR 或者 YARN\_CONF\_DIR指向包含hadoop集群客户端配置文件的目录。这些配置用于向HDFS写数据以及连接YARN RM。这些配置也会被分发到YARN集群，因此该app使用的所有container都使用相同的配置。如果这些配置引用Java系统属性或者不被YARN管理的环境变量，则应该把它们设置到spark app的配置里。

有两种部署模式可以提交app到YARN，cluster模式下，spark driver运行在application master进程内部，该master进程由YARN管理，client在启动app后可以离开。

Client模式下，driver运行在client进程里，application master仅仅用于向YARN请求资源。

在YARN模式下，RM的地址从hadoop配置文件里获得，因此，--master参数是yarn.

Cluster模式提交spark app:

$ ./bin/spark-submit --class path.to.your.Class --master yarn --deploy-mode cluster [options] <app jar> [app options]

For example:

$ ./bin/spark-submit --class org.apache.spark.examples.SparkPi \

--master yarn \

--deploy-mode cluster \

--driver-memory 4g \

--executor-memory 2g \

--executor-cores 1 \

--queue thequeue \

examples/jars/spark-examples\*.jar \

10

这启动了YARN客户端程序，它启动了默认的application master。然后，SparkPi作为Application Master的子线程运行。Client将周期性的从Application Master拉取状态更新并打印到屏幕。一旦app结束运行，client就会退出。

Yarn-client模式与yarn-cluster模式提交命令类似，只是把cluster改为client即可

使用yarn client模式运行spark-shell:

$ ./bin/spark-shell --master yarn --deploy-mode client

## **添加其他jar**

Cluster模式下，driver与client运行在不同的机器上，因此SparkContext.addJar将不会加载client所在机器的jar。为了让SparkContext.addJar可以获得client机器上的文件，需要使用--jars选项：

$ ./bin/spark-submit --class my.main.Class \

--master yarn \

--deploy-mode cluster \

--jars my-other-jar.jar,my-other-other-jar.jar \

my-main-jar.jar \

app\_arg1 app\_arg2

# 日志

YARN术语里，executors和application masters运行在容器内，YARN在app完成后有两种方式处理容器日志，如果开启日志聚集（yarn.log-aggregation-enable）,容器日志将被拷贝到HDFS并且删除本地机器上的日志，任意节点使用yarn logs命令可访问这些日志

yarn logs -applicationId <app ID>

日志目录在YARN配置文件里：

(yarn.nodemanager.remote-app-log-dir 和yarn.nodemanager.remote-app-log-dir-suffix). Spark Web UI 的Executors 标签页里也可以看

如果日志聚集没有打开，日志就在本地 YARN\_APP\_LOGS\_DIR目录下

To review per-container launch environment, increase yarn.nodemanager.delete.debug-delay-sec to a large value (e.g. 36000), and then access the application cache through yarn.nodemanager.local-dirs on the nodes on which containers are launched. This directory contains the launch script, JARs, and all environment variables used for launching each container. This process is useful for debugging classpath problems in particular. (Note that enabling this requires admin privileges on cluster settings and a restart of all node managers. Thus, this is not applicable to hosted clusters).

对application master或者executors使用自定义log4j配置，可选：

* upload a custom log4j.properties using spark-submit, by adding it to the --files list of files to be uploaded with the application.

危险：同一个节点上executors 和application master往同一个文件写日志

* add -Dlog4j.configuration=<location of configuration file> to spark.driver.extraJavaOptions (for the driver) or spark.executor.extraJavaOptions (for executors). Note that if using a file, the file: protocol should be explicitly provided, and the file needs to exist locally on all the nodes.
* 更新 $SPARK\_CONF\_DIR/log4j.properties file and it will be automatically uploaded along with the other configurations. Note that other 2 options has higher priority than this option if multiple options are specified.

If you need a reference to the proper location to put log files in the YARN so that YARN can properly display and aggregate them, use spark.yarn.app.container.log.dir in your log4j.properties. For example, log4j.appender.file\_appender.File=${spark.yarn.app.container.log.dir}/spark.log. For streaming applications, configuring RollingFileAppender and setting file location to YARN’s log directory will avoid disk overflow caused by large log files, and logs can be accessed using YARN’s log utility.

To use a custom metrics.properties for the application master and executors, update the $SPARK\_CONF\_DIR/metrics.properties file. It will automatically be uploaded with other configurations, so you don’t need to specify it manually with --files.

#### **Spark Properties**

|  |  |  |
| --- | --- | --- |
| **Property Name** | **Default** | **Meaning** |
| spark.yarn.am.memory | 512m | Amount of memory to use for the YARN Application Master in client mode, in the same format as JVM memory strings (e.g. 512m, 2g). In cluster mode, use spark.driver.memory instead.  Use lower-case suffixes, e.g. k, m, g, t, and p, for kibi-, mebi-, gibi-, tebi-, and pebibytes, respectively. |
| spark.yarn.am.cores | 1 | Number of cores to use for the YARN Application Master in client mode. In cluster mode, use spark.driver.cores instead. |
| spark.yarn.am.waitTime | 100s | In cluster mode, time for the YARN Application Master to wait for the SparkContext to be initialized. In client mode, time for the YARN Application Master to wait for the driver to connect to it. |
| spark.yarn.submit.file.replication | The default HDFS replication (usually 3) | HDFS replication level for the files uploaded into HDFS for the application. These include things like the Spark jar, the app jar, and any distributed cache files/archives. |
| spark.yarn.stagingDir | Current user's home directory in the filesystem | Staging directory used while submitting applications. |
| spark.yarn.preserve.staging.files | false | Set to true to preserve the staged files (Spark jar, app jar, distributed cache files) at the end of the job rather than delete them. |
| spark.yarn.scheduler.heartbeat.interval-ms | 3000 | The interval in ms in which the Spark application master heartbeats into the YARN ResourceManager. The value is capped at half the value of YARN's configuration for the expiry interval, i.e. yarn.am.liveness-monitor.expiry-interval-ms. |
| spark.yarn.scheduler.initial-allocation.interval | 200ms | The initial interval in which the Spark application master eagerly heartbeats to the YARN ResourceManager when there are pending container allocation requests. It should be no larger than spark.yarn.scheduler.heartbeat.interval-ms. The allocation interval will doubled on successive eager heartbeats if pending containers still exist, untilspark.yarn.scheduler.heartbeat.interval-ms is reached. |
| spark.yarn.max.executor.failures | numExecutors \* 2, with minimum of 3 | The maximum number of executor failures before failing the application. |
| spark.yarn.historyServer.address | (none) | The address of the Spark history server, e.g. host.com:18080. The address should not contain a scheme (http://). Defaults to not being set since the history server is an optional service. This address is given to the YARN ResourceManager when the Spark application finishes to link the application from the ResourceManager UI to the Spark history server UI. For this property, YARN properties can be used as variables, and these are substituted by Spark at runtime. For example, if the Spark history server runs on the same node as the YARN ResourceManager, it can be set to ${hadoopconf-yarn.resourcemanager.hostname}:18080. |
| spark.yarn.dist.archives | (none) | Comma separated list of archives to be extracted into the working directory of each executor. |
| spark.yarn.dist.files | (none) | Comma-separated list of files to be placed in the working directory of each executor. |
| spark.yarn.dist.jars | (none) | Comma-separated list of jars to be placed in the working directory of each executor. |
| spark.yarn.dist.forceDownloadSchemes | (none) | Comma-separated list of schemes for which files will be downloaded to the local disk prior to being added to YARN's distributed cache. For use in cases where the YARN service does not support schemes that are supported by Spark, like http, https and ftp. |
| spark.executor.instances | 2 | The number of executors for static allocation. With spark.dynamicAllocation.enabled, the initial set of executors will be at least this large. |
| spark.yarn.am.memoryOverhead | AM memory \* 0.10, with minimum of 384 | Same as spark.driver.memoryOverhead, but for the YARN Application Master in client mode. |
| spark.yarn.queue | default | The name of the YARN queue to which the application is submitted. |
| spark.yarn.jars | (none) | List of libraries containing Spark code to distribute to YARN containers. By default, Spark on YARN will use Spark jars installed locally, but the Spark jars can also be in a world-readable location on HDFS. This allows YARN to cache it on nodes so that it doesn't need to be distributed each time an application runs. To point to jars on HDFS, for example, set this configuration to hdfs:///some/path. Globs are allowed. |
| spark.yarn.archive | (none) | An archive containing needed Spark jars for distribution to the YARN cache. If set, this configuration replaces spark.yarn.jars and the archive is used in all the application's containers. The archive should contain jar files in its root directory. Like with the previous option, the archive can also be hosted on HDFS to speed up file distribution. |
| spark.yarn.access.hadoopFileSystems | (none) | A comma-separated list of secure Hadoop filesystems your Spark application is going to access. For example, spark.yarn.access.hadoopFileSystems=hdfs://nn1.com:8032,hdfs://nn2.com:8032, webhdfs://nn3.com:50070. The Spark application must have access to the filesystems listed and Kerberos must be properly configured to be able to access them (either in the same realm or in a trusted realm). Spark acquires security tokens for each of the filesystems so that the Spark application can access those remote Hadoop filesystems. spark.yarn.access.namenodes is deprecated, please use this instead. |
| spark.yarn.appMasterEnv.[EnvironmentVariableName] | (none) | Add the environment variable specified by EnvironmentVariableName to the Application Master process launched on YARN. The user can specify multiple of these and to set multiple environment variables. In cluster mode this controls the environment of the Spark driver and in client mode it only controls the environment of the executor launcher. |
| spark.yarn.containerLauncherMaxThreads | 25 | The maximum number of threads to use in the YARN Application Master for launching executor containers. |
| spark.yarn.am.extraJavaOptions | (none) | A string of extra JVM options to pass to the YARN Application Master in client mode. In cluster mode, use spark.driver.extraJavaOptions instead. Note that it is illegal to set maximum heap size (-Xmx) settings with this option. Maximum heap size settings can be set with spark.yarn.am.memory |
| spark.yarn.am.extraLibraryPath | (none) | Set a special library path to use when launching the YARN Application Master in client mode. |
| spark.yarn.maxAppAttempts | yarn.resourcemanager.am.max-attempts in YARN | The maximum number of attempts that will be made to submit the application. It should be no larger than the global number of max attempts in the YARN configuration. |
| spark.yarn.am.attemptFailuresValidityInterval | (none) | Defines the validity interval for AM failure tracking. If the AM has been running for at least the defined interval, the AM failure count will be reset. This feature is not enabled if not configured. |
| spark.yarn.executor.failuresValidityInterval | (none) | Defines the validity interval for executor failure tracking. Executor failures which are older than the validity interval will be ignored. |
| spark.yarn.submit.waitAppCompletion | true | In YARN cluster mode, controls whether the client waits to exit until the application completes. If set to true, the client process will stay alive reporting the application's status. Otherwise, the client process will exit after submission. |
| spark.yarn.am.nodeLabelExpression | (none) | A YARN node label expression that restricts the set of nodes AM will be scheduled on. Only versions of YARN greater than or equal to 2.6 support node label expressions, so when running against earlier versions, this property will be ignored. |
| spark.yarn.executor.nodeLabelExpression | (none) | A YARN node label expression that restricts the set of nodes executors will be scheduled on. Only versions of YARN greater than or equal to 2.6 support node label expressions, so when running against earlier versions, this property will be ignored. |
| spark.yarn.tags | (none) | Comma-separated list of strings to pass through as YARN application tags appearing in YARN ApplicationReports, which can be used for filtering when querying YARN apps. |
| spark.yarn.keytab | (none) | The full path to the file that contains the keytab for the principal specified above. This keytab will be copied to the node running the YARN Application Master via the Secure Distributed Cache, for renewing the login tickets and the delegation tokens periodically. (Works also with the "local" master) |
| spark.yarn.principal | (none) | Principal to be used to login to KDC, while running on secure HDFS. (Works also with the "local" master) |
| spark.yarn.kerberos.relogin.period | 1m | How often to check whether the kerberos TGT should be renewed. This should be set to a value that is shorter than the TGT renewal period (or the TGT lifetime if TGT renewal is not enabled). The default value should be enough for most deployments. |
| spark.yarn.config.gatewayPath | (none) | A path that is valid on the gateway host (the host where a Spark application is started) but may differ for paths for the same resource in other nodes in the cluster. Coupled withspark.yarn.config.replacementPath, this is used to support clusters with heterogeneous configurations, so that Spark can correctly launch remote processes.  The replacement path normally will contain a reference to some environment variable exported by YARN (and, thus, visible to Spark containers).  For example, if the gateway node has Hadoop libraries installed on /disk1/hadoop, and the location of the Hadoop install is exported by YARN as the HADOOP\_HOME environment variable, setting this value to /disk1/hadoop and the replacement path to $HADOOP\_HOMEwill make sure that paths used to launch remote processes properly reference the local YARN configuration. |
| spark.yarn.config.replacementPath | (none) | See spark.yarn.config.gatewayPath. |
| spark.security.credentials.${service}.enabled | true | Controls whether to obtain credentials for services when security is enabled. By default, credentials for all supported services are retrieved when those services are configured, but it's possible to disable that behavior if it somehow conflicts with the application being run. For further details please see [Running in a Secure Cluster](running-on-yarn.html#running-in-a-secure-cluster) |
| spark.yarn.rolledLog.includePattern | (none) | Java Regex to filter the log files which match the defined include pattern and those log files will be aggregated in a rolling fashion. This will be used with YARN's rolling log aggregation, to enable this feature in YARN side yarn.nodemanager.log-aggregation.roll-monitoring-interval-seconds should be configured in yarn-site.xml. This feature can only be used with Hadoop 2.6.4+. The Spark log4j appender needs be changed to use FileAppender or another appender that can handle the files being removed while it is running. Based on the file name configured in the log4j configuration (like spark.log), the user should set the regex (spark\*) to include all the log files that need to be aggregated. |
| spark.yarn.rolledLog.excludePattern | (none) | Java Regex to filter the log files which match the defined exclude pattern and those log files will not be aggregated in a rolling fashion. If the log file name matches both the include and the exclude pattern, this file will be excluded eventually. |

# **重要tips**

* 在cluster模式，spark executors和spark driver使用的本地目录是配置在YARN（yarn.nodemanager.local-dirs）里的，如果用户指定了spark.local.dir,则配置被忽略。在client模式，spark executors使用YARN中配置的,而spark driver使用spark.local.dir配置的目录。
* --files和--archives支持使用#指定文件名，例如，--files localtest.txt#appSees.txt，这将把本地文件appSees.txt上传到HDFS，该文件由”appSees.txt”引用，app使用”appSees.txt”应用这个文件。
* --jars选项允许在cluster模式下使用SparkContext.addJar正常加载本地文件。

# **Running in a Secure Cluster**

As covered in [security](http://spark.apache.org/docs/latest/security.html), Kerberos is used in a secure Hadoop cluster to authenticate principals associated with services and clients. This allows clients to make requests of these authenticated services; the services to grant rights to the authenticated principals.

Hadoop services issue *hadoop tokens* to grant access to the services and data. Clients must first acquire tokens for the services they will access and pass them along with their application as it is launched in the YARN cluster.

For a Spark application to interact with any of the Hadoop filesystem (for example hdfs, webhdfs, etc), HBase and Hive, it must acquire the relevant tokens using the Kerberos credentials of the user launching the application —that is, the principal whose identity will become that of the launched Spark application.

This is normally done at launch time: in a secure cluster Spark will automatically obtain a token for the cluster’s default Hadoop filesystem, and potentially for HBase and Hive.

An HBase token will be obtained if HBase is in on classpath, the HBase configuration declares the application is secure (i.e. hbase-site.xml sets hbase.security.authentication to kerberos), and spark.security.credentials.hbase.enabled is not set to false.

Similarly, a Hive token will be obtained if Hive is on the classpath, its configuration includes a URI of the metadata store in "hive.metastore.uris, and spark.security.credentials.hive.enabled is not set to false.

If an application needs to interact with other secure Hadoop filesystems, then the tokens needed to access these clusters must be explicitly requested at launch time. This is done by listing them in the spark.yarn.access.hadoopFileSystems property.

spark.yarn.access.hadoopFileSystems hdfs://ireland.example.org:8020/,webhdfs://frankfurt.example.org:50070/

Spark supports integrating with other security-aware services through Java Services mechanism (see java.util.ServiceLoader). To do that, implementations of org.apache.spark.deploy.yarn.security.ServiceCredentialProvider should be available to Spark by listing their names in the corresponding file in the jar’s META-INF/services directory. These plug-ins can be disabled by setting spark.security.credentials.{service}.enabled to false, where {service} is the name of credential provider.

## **Configuring the External Shuffle Service**

To start the Spark Shuffle Service on each NodeManager in your YARN cluster, follow these instructions:

1. Build Spark with the [YARN profile](http://spark.apache.org/docs/latest/building-spark.html). Skip this step if you are using a pre-packaged distribution.
2. Locate the spark-<version>-yarn-shuffle.jar. This should be under $SPARK\_HOME/common/network-yarn/target/scala-<version> if you are building Spark yourself, and under yarn if you are using a distribution.
3. Add this jar to the classpath of all NodeManagers in your cluster.
4. In the yarn-site.xml on each node, add spark\_shuffle to yarn.nodemanager.aux-services, then set yarn.nodemanager.aux-services.spark\_shuffle.class to org.apache.spark.network.yarn.YarnShuffleService.
5. Increase NodeManager's heap size by setting YARN\_HEAPSIZE (1000 by default) in etc/hadoop/yarn-env.sh to avoid garbage collection issues during shuffle.
6. Restart all NodeManagers in your cluster.

The following extra configuration options are available when the shuffle service is running on YARN:

|  |  |  |
| --- | --- | --- |
| **Property Name** | **Default** | **Meaning** |
| spark.yarn.shuffle.stopOnFailure | false | Whether to stop the NodeManager when there's a failure in the Spark Shuffle Service's initialization. This prevents application failures caused by running containers on NodeManagers where the Spark Shuffle Service is not running. |

## **Launching your application with Apache Oozie**

Apache Oozie can launch Spark applications as part of a workflow. In a secure cluster, the launched application will need the relevant tokens to access the cluster’s services. If Spark is launched with a keytab, this is automatic. However, if Spark is to be launched without a keytab, the responsibility for setting up security must be handed over to Oozie.

The details of configuring Oozie for secure clusters and obtaining credentials for a job can be found on the [Oozie web site](http://oozie.apache.org/) in the “Authentication” section of the specific release’s documentation.

For Spark applications, the Oozie workflow must be set up for Oozie to request all tokens which the application needs, including:

* The YARN resource manager.
* The local Hadoop filesystem.
* Any remote Hadoop filesystems used as a source or destination of I/O.
* Hive —if used.
* HBase —if used.
* The YARN timeline server, if the application interacts with this.

To avoid Spark attempting —and then failing— to obtain Hive, HBase and remote HDFS tokens, the Spark configuration must be set to disable token collection for the services.

The Spark configuration must include the lines:

spark.security.credentials.hive.enabled false

spark.security.credentials.hbase.enabled false

The configuration option spark.yarn.access.hadoopFileSystems must be unset.

## **Troubleshooting Kerberos**

Debugging Hadoop/Kerberos problems can be “difficult”. One useful technique is to enable extra logging of Kerberos operations in Hadoop by setting the HADOOP\_JAAS\_DEBUG environment variable.

export HADOOP\_JAAS\_DEBUG=true

The JDK classes can be configured to enable extra logging of their Kerberos and SPNEGO/REST authentication via the system properties sun.security.krb5.debug and sun.security.spnego.debug=true

-Dsun.security.krb5.debug=true -Dsun.security.spnego.debug=true

All these options can be enabled in the Application Master:

spark.yarn.appMasterEnv.HADOOP\_JAAS\_DEBUG true

spark.yarn.am.extraJavaOptions -Dsun.security.krb5.debug=true -Dsun.security.spnego.debug=true

Finally, if the log level for org.apache.spark.deploy.yarn.Client is set to DEBUG, the log will include a list of all tokens obtained, and their expiry details

## **Using the Spark History Server to replace the Spark Web UI**

It is possible to use the Spark History Server application page as the tracking URL for running applications when the application UI is disabled. This may be desirable on secure clusters, or to reduce the memory usage of the Spark driver. To set up tracking through the Spark History Server, do the following:

* On the application side, set spark.yarn.historyServer.allowTracking=true in Spark’s configuration. This will tell Spark to use the history server’s URL as the tracking URL if the application’s UI is disabled.
* On the Spark History Server, add org.apache.spark.deploy.yarn.YarnProxyRedirectFilter to the list of filters in the spark.ui.filtersconfiguration.

Be aware that the history server information may not be up-to-date with the application’s state.