# Installing Impala on MapR

## Installation Overview

Impala is comprised of a set of components that you install and run on a single node or on multiple nodes in a cluster. To run Impala in your cluster, install the required Impala packages on designated nodes. The Impala packages contain the following Impala components:

- Impala daemon
- Impala statestore
- Impala catalog
- Impala binaries

The following table lists the Impala packages and their descriptions:

| Package | Description |
|---|---|
| `mapr-impala` | A package that contains all of the Impala binaries, including the Impala server, impala-shell, statestore, and catalog. |
| `mapr-impala-server` | The role package that installs the Impala daemon role on the node. This package enables Warden to manage the service. It is recommended (not required) that you install the Impala daemon on a node with the MapR fileserver installed. |
| `mapr-impala-statestore` | The role package that installs the Impala statestore role on the node. This package enables Warden to manage the service. |
| `mapr-impala-catalog` | The role package that installs the Impala catalog role on the node. This package enables Warden to manage the service. |

The default Impala memory setting is high, which can result in conflict between Impala and other frameworks running in the cluster. To avoid memory conflicts with Impala, modify the Impala and MapR-FS FileServer settings to ensure that each framework has enough memory to run jobs to completion.  See the installation instructions for more information.

### Prerequisites

To successfully install and run Impala, verify that the system meets the following hardware and software requirements.

| Prerequisite | Requirements |
|---|---|
| Operating System | MapR provides packages for the following 64-bit operating systems:<br>• RedHat 6.x/7<br>• CentOS 6.x/7 |
| MapR Distribution for Hadoop | MapR distribution version 5.0.0 on Hadoop 2.<br><br>Verify that you have added the MapR repository on RedHat or CentOS. You should have the `maprtech.repo` in the directory `/etc/yum.repos.d/` with the following content:<br><br>`[maprtech]`<br>`name=MapR Core Components`<br>`baseurl=http://package.mapr.com/releases/<version>/<operating system>`<br>`enabled=1`<br>`gpgcheck=0`<br>`protect=1`<br><br>`[maprecosystem]`<br>`name=MapR Ecosystem Components`<br>`baseurl=http://package.mapr.com/releases/ecosystem-5.x/<system> (or .../ecosystem/<system>)`<br>`enabled=1`<br>`gpgcheck=0`<br>`protect=1`<br><br>For more information, see Installing MapR Software-Preparing Packages and Repositories. |

| | |
|---|---|
| Hive Metastore | To use Impala for MapR, you must install and configure a Hive metastore. Configure the Hive metastore service and connect to a MySQL database through the service. For more information, see Installing Hive.<br><br>**Note:** Verify that `hive-site.xml` contains the `hive.metastore.uris` setting, and substitute the appropriate host name for `metastore_server_host` on every Impala server node.<br><br>**Example**<br>`<property>`<br>`<name>hive.metastore.uris</name>`<br>`<value>thrift://<metastore_server_host>:9083</value>`<br>`</property>` |
| Hive | Hive 1.2 is required to run Impala 2.2.0. Earlier versions of Hive are not supported. If you have earlier versions of Hive installed, installing the Impala packages uninstalls Hive, except for the configuration logs and process identifiers, and updates the `mapr-hive` package with Hive 1.2. Impala must have access to the same metastore database that Hive uses. For Hive configuration information, see the Hive documentation. |
| HBase | HBase 0.98 is required to run Impala. If you install HBase from MapR's `ecosystem-5.x` or `ecosystem-all` repositories, you currently get HBase 0.98.x, by default. |
| Java | JDK 1.7 or 1.8 |

# Installing Impala

Install the Impala package on nodes in the cluster that you have designated to run Impala. Install the Impala server on every node designated to run `impalad`. Install the `statestore` and `catalog` packages on only one node in the cluster. Install the `impala-shell` on the client machine. You can use the `impala-shell` to connect to an Impala service and run queries from the command line.



It is recommended that you install the statestore and catalog together on a separate machine from the Impala server.

Complete the following steps to install Impala, impala-server, statestore, catalog, and the impala-shell:

1. Install the `mapr-impala` package on all the nodes designated to run Impala:

   To install the package, issue the following command:

   **RedHat and CentOS**

   ```
   $ sudo yum install mapr-impala
   ```

2. In `/opt/mapr/impala/impala-<version>/conf/env.sh`, complete the following steps:
   a. Verify that the statestore address is set to the address where you plan to run the statestore service.
      Example: `IMPALA_STATE_STORE_HOST=<IP address hosting statestore>`
   b. Change the catalog service address to the address where you plan to run the catalog service.
      Example: `CATALOG_SERVICE_HOST=<IP address hosting catalog service>`
   c. Add the `mem_limit` and `num_threads_per_disk` parameters to `IMPALA_SERVER_ARGS` to allocate a specific amount of memory to Impala, and limit the number of threads that each disk processes per impala server daemon. Adding these parameters can alleviate any potential resource conflicts that may occur between Impala and other frameworks running in the cluster.
      Example:

      ```
      export IMPALA_SERVER_ARGS=${IMPALA_SERVER_ARGS:- \
      ```

```
-log_dir=${IMPALA_LOG_DIR} \
-state_store_port=${IMPALA_STATE_STORE_PORT} \
-use_statestore -state_store_host=${IMPALA_STATE_STORE_HOST} \
-catalog_service_host=${CATALOG_SERVICE_HOST} \
-be_port=${IMPALA_BACKEND_PORT} \
-mem_limit=<absolute notation or percentage of physical memory> \
-num_threads_per_disk=<n>}
```

See Additional Impala Configuration Options for more information about these options and other options that you can modify in `env.sh`.



The default maximum heap space allocated to the MapR-FS fileserver should provide enough memory for the MapR-FS fileserver to run concurrently with Impala, however you can modify it if needed. To modify the maximum heap space, navigate to `/opt/mapr/conf/warden.conf`, and change the `service.command.mfs.heapsize.maxpercent parameter.` Issue the following command to restart Warden after you modify the parameter:

```
service mapr-warden restart
```

Refer to warden.conf for more Warden configuration information.

3. Verify that the following property is configured in `hive-site.xml` on all the nodes:

```
<property>
 <name>hive.metastore.uris</name>
 <value>thrift://<metastore_server_host>:9083</value>
</property>
```

4. Install the Impala components:
   a. To install the statestore service, issue the following command:

**RedHat and CentOS**

```
$ sudo yum install mapr-impala-statestore
```

   b. To install the catalog service, issue the following install command:

**RedHat and CentOS**

```
$ sudo yum install mapr-impala-catalog
```

   c. To install the Impala server, issue the following install command:

**RedHat and CentOS**

```
$ sudo yum install mapr-impala-server
```

5. Run `configure.sh` to refresh the node configuration.
   Example: `/opt/mapr/server/configure.sh -R`
6. If the Hive metastore has MapR-SASL enabled, copy `$HIVE_HOME/conf/hive-site.xml` to `$IMPALA_HOME/conf/.` Repeat this step any time `hive-site.xml` is modified.

At this point, the Impala server and statestore should be running. For instructions on how to run a simple Impala query and how to query HBase tables, see Working with Impala.