Installing Hive

Install the appropriate Hive packages using the appropriate command for your distribution.

OS	Command
RHEL-compatible	`$ sudo yum install <pkg1> <pkg2> ...`
SLES	`$ sudo zypper install <pkg1> <pkg2> ...`
Ubuntu or Debian	`$ sudo apt-get install <pkg1> <pkg2> ...`

The packages are:

- `hive` – base package that provides the complete language and runtime
- `hive-metastore` – provides scripts for running the metastore as a standalone service (optional)
- `hive-server2` – provides scripts for running HiveServer2
- `hive-hbase` - optional; install this package if you want to use Hive with HBase.

Heap Size and Garbage Collection for Hive Components

Hive Component Memory Recommendations

HiveServer2 and the Hive metastore require sufficient memory in order to run correctly. The default heap size of 256 MB for each component is inadequate for production workloads. Consider the following guidelines for sizing the heap for each component, based upon your cluster size.

Number of Concurrent Connections	HiveServer2 Heap Size Minimum Recommendation	Hive Metastore Heap Size Minimum Recommendation
Up to 40 concurrent connections (Cloudera recommends splitting HiveServer2 into multiple instances and load balancing once you start allocating >12 GB to HiveServer2. The objective is to size to reduce impact of Java garbage collection on active processing by the service.	12 GB	12 GB
Up to 20 concurrent connections	6 GB	10 GB
Up to 10 concurrent connections	4 GB	8 GB
Single connection	2 GB	4 GB

Important: These numbers are general guidance only, and may be affected by factors such as number of columns, partitions, complex joins, and client activity among other things. It is important to review and refine through testing based on your anticipated deployment to arrive at best values for your environment.

In addition, the Beeline CLI should use a heap size of at least 2 GB.

The `permGenSize` should be set to 512M for all.

Configuring Heap Size and Garbage Collection for Hive Components

To configure the heap size for HiveServer2 and Hive metastore, set the `-Xmx` parameter in the `HADOOP_OPTS` variable to the desired maximum heap size in the `hive-env.sh` advanced configuration snippet if you use Cloudera Manager or otherwise edit `/etc/hive/hive-env.sh`.

To configure the heap size for the Beeline CLI, set the `HADOOP_HEAPSIZE` environment variable in the `hive-env.sh` advanced configuration snippet if you use Cloudera Manager or otherwise edit `/etc/hive/hive-env.sh` before starting the Beeline CLI.

The following example shows a configuration with the following settings:

- HiveServer2 uses 12 GB heap
- Hive metastore uses 12 GB heap
- Hive clients use 2 GB heap

The settings to change are in bold. All of these lines are commented out (prefixed with a # character) by default. Uncomment the lines by removing the # character.

```
if [ "$SERVICE" = "cli" ]; then
  if [ -z "$DEBUG" ]; then
    export HADOOP_OPTS="$HADOOP_OPTS -XX:NewRatio=12 -Xmx12288m -Xms10m -XX:MaxHeapFreeRatio=40 -XX:MinHeapFreeRatio=15 -XX:+UseParNewGC -XX:-UseGCOverheadLimi
  else
    export HADOOP_OPTS="$HADOOP_OPTS -XX:NewRatio=12 -Xmx12288m -Xms10m -XX:MaxHeapFreeRatio=40 -XX:MinHeapFreeRatio=15 -XX:-UseGCOverheadLimit"
  fi
fi

export HADOOP_HEAPSIZE=2048
```

You can choose whether to use the Concurrent Collector or the New Parallel Collector for garbage collection, by passing `-XX:+UseParNewGC` or `-XX:+UseConcMarkSweepGC` in the `HADOOP_OPTS` lines above, and you can tune the garbage collection overhead limit by setting `-XX:-UseGCOverheadLimit`. To enable the garbage collection overhead limit, remove the setting or change it to `-XX:+UseGCOverheadLimit`.

Configuration for WebHCat

If you want to use WebHCat, you need to set the `PYTHON_CMD` variable in /etc/default/hive-webhcat-server after installing Hive; for example:

```
export PYTHON_CMD=/usr/bin/python
```

Categories: Configuring | Hive | Installing | All Categories

Upgrading Hive
Configuring the Hive Metastore