

Chukwa Quick Start

=NOTE: This document is updated for Chukwa trunk development instruction; you should probably look at the [Administration Guide](#) for stable release instructions instead.

Purpose

Chukwa is a system for large-scale reliable log collection and processing with Hadoop. The Chukwa design overview discusses the overall architecture of Chukwa. You should read that document before this one. The purpose of this document is to help you install and configure Chukwa.

Pre-requisites

Chukwa should work on any POSIX platform, but GNU/Linux is the only production platform that has been tested extensively. Chukwa has also been used successfully on Mac OS X, which several members of the Chukwa team use for development.

The only absolute software requirements are Java 1.6 or better and Hadoop 0.20.205+. HICC, the Chukwa visualization interface, requires HBase 0.90.4.

The Chukwa cluster management scripts rely on ssh; these scripts, however, are not required if you have some alternate mechanism for starting and stopping daemons.

Installing Chukwa

A minimal Chukwa deployment has three components:

A Hadoop and HBase cluster on which Chukwa will process data (referred to as the Chukwa cluster). A collector process, that writes collected data to HBase. One or more agent processes, that send monitoring data to the collector. The nodes with active agent processes are referred to as the monitored source nodes. In addition, you may wish to run the Chukwa Demux jobs, which parse collected data, or HICC, the Chukwa visualization tool.

<http://people.apache.org/~eyang/docs/chukwa-0.5-arch.png>

Compiling and installing Chukwa

1. To compile Chukwa, just type 'mvn clean package -DskipTests -DHADOOP_CONF_DIR=/path/to/\$HADOOP_CONF_DIR -DHBASE_CONF_DIR=/path/to/\$HBASE_CONF_DIR' in the project root directory.
2. Extract the compiled tar file from target/chukwa-0.x.y.tar.gz to the Chukwa root directory.

Setup Chukwa Cluster

General Hadoop configuration is available at: [Hadoop Configuration](#)

Configure Log4j syslog appender

1. Edit HADOOP_CONF_DIR/log4j.properties, and replace DRFA appender with [SocketAppender](#):

```
log4j.appender.DRFA=org.apache.log4j.net.SocketAppender
log4j.appender.DRFA.RemoteHost=localhost
log4j.appender.DRFA.Port=9096
log4j.appender.DRFA.layout=org.apache.log4j.PatternLayout
log4j.appender.DRFA.layout.ConversionPattern=%d{ISO8601} %p %c: %m%n
```

Save the file.

2. Copy CHUKWA_HOME/hadoop-metrics.properties to HADOOP_CONF_DIR.
3. Copy CHUKWA_HOME/share/chukwa/chukwa-0.5.0-client.jar to HADOOP_HOME/share/hadoop/lib.
4. Copy CHUKWA_HOME/share/chukwa/lib/json-simple-1.1.jar to HADOOP_HOME/share/hadoop/lib.
5. Restart Hadoop Cluster.
6. General HBASE configuration is available at: [HBase Configuration](#)
7. After Hadoop and HBase has been configured properly, run:

```
bin/hbase shell < /path/to/CHUKWA_HOME/conf/hbase.schema
```

This procedure initializes the default Chukwa HBase schema.

Configuring and starting the Collector

1. Edit `etc/chukwa/chukwa-collector-conf.xml` and comment out the default properties for `chukwaCollector.writerClass`, and `chukwaCollector.pipeline`. Uncomment block for `HBaseWriter` parameters, and save.
2. Edit `chukwa-env.sh`. You almost certainly need to set `JAVA_HOME`, `HADOOP_HOME`, `HADOOP_CONF_DIR`, `HBASE_HOME`, and `HBASE_CONF_DIR` at least.
3. In the `chukwa` root directory, run `'bin/chukwa collector'`

Configuring and starting the local agent

1. Verify `etc/chukwa/chukwa-agent-conf.xml` configuration
2. Verify `etc/chukwa/collectors` contains list of collector hostname
3. In the `chukwa` root directory, run `'bin/chukwa agent'`

Starting Adaptors

The local agent speaks a simple text-based protocol, by default over port 9093. Suppose you want Chukwa to monitor system metrics, hadoop metrics, and hadoop logs on the localhost:

1. Telnet to localhost 9093
2. Type [without quotation marks] `"add org.apache.hadoop.chukwa.datacollection.adaptor.sigar.SystemMetrics [SystemMetrics] 60 0"`
3. Type [without quotation marks] `"add [SocketAdaptor] [HadoopMetrics] 9095 0"`
4. Type [without quotation marks] `"add [SocketAdaptor] Hadoop 9096 0"`
5. Type `"list"` – you should see the adaptor you just started, listed as running.
6. Type `"close"` to break the connection.
7. If you don't have telnet, you can get the same effect using the netcat (*nc*) command line tool.

Set Up Cluster Aggregation Script

For data analytics with pig, there are some additional environment setup. Pig does not use the same environment variable name as Hadoop, therefore make sure the following environment are setup correctly:

1. `export PIG_CLASSPATH=$HADOOP_CONF_DIR:$HBASE_CONF_DIR`
2. Setup a cron job for `"pig -Dpig.additional.jars=${HBASE_HOME}/hbase-0.90.4.jar:${PIG_PATH}/pig.jar ${CHUKWA_HOME}/script/pig/ClusterSummary.pig"` to run periodically

Set Up HICC

The Hadoop Infrastructure Care Center (HICC) is the Chukwa web user interface. To set up HICC, do the following:

1. `bin/chukwa hicc`

Data visualization

1. Point web browser to http://localhost:4080/hicc/jsp/graph_explorer.jsp
2. The default user name and password is `"demo"` without quotes.
3. System Metrics collected by Chukwa collector will be browsable through `graph_explorer.jsp` file.