Overview

Log processing was one of the original purposes of MapReduce. Unfortunately, using Hadoop for MapReduce processing of logs is somewhat troublesome. Logs are generated incrementally across many machines, but Hadoop MapReduce works best on a small number of large files. And HDFS doesn't currently support appends, making it difficult to keep the distributed copy fresh.

Chukwa is a Hadoop subproject devoted to bridging that gap between logs and MapReduce. Chukwa is a scalable distributed monitoring and analysis system, particularly logs from Hadoop and other large systems.

The Chukwa Documentation provides the information you need to get started using Chukwa. You should start with the [Architecture and Design document](#).

If you're trying to set up a Chukwa cluster from scratch, you should read the [Chukwa Administration Guide](#) which shows you how to setup and deploy Chukwa.

If you want to configure the Chukwa agent process, to control what's collected, you should read the [Agent Guide](#). There's also a [Collector Guide](#) describing that part of the pipeline.

And if you want to use collected data, read the [User and Programming Guide](#)

If you have more questions, you can ask on the [Chukwa mailing lists](#).