## What is Flume?

* Flume is distributed system for collecting log data/unstructured data from many sources, aggregating it, and writing it to HDFS
* It is designed to be reliable and highly available, while providing a simple, flexible, and intuitive programming model based on streaming data flows
* Flume and Chukwa share similar goals and features, however there are notable differences
* Flume maintains a central list of ongoing data flows, stored redundantly in Zookeeper, Chukwa distributes this information more broadly among its services
* Flume adopts a “hop-by-hop” model, while in Chukwa the agents on each machine are responsible for deciding what data to send
* Flume is a streaming data ingest framework
* Using Flume we can write data directly into Solr
* It has a traditional Pubs-Sub architecture
* Used to fitter data on the way to Hadoop

## What is Chukwa?

* Log processing was one of the original purposes of MapReduce. Unfortunately, Hadoop is hard to use for this purpose
* Writing MapReduce jobs to process logs is somewhat tedious and the batch nature of MapReduce makes it difficult to use with logs that are generated incrementally across many machines. Furthermore, HDFS still does not support appending to existing files.
* Chukwa is a Hadoop subproject that bridges that gap between log handling and MapReduce. It provides a scalable distributed system for monitoring and analysis of log-based data. Some of the durability features include agent-side replying of data to recover from errors.