

# HadoopIndexing

- [What is Solr + Hadoop Indexing?](#)
- [Architecture](#)
  - [CSVIndexer](#)
  - [SolrDocumentConverter](#)
  - [Heartbeater](#)
  - [Number of Reduce Tasks](#)
  - [Zipped Shards](#)
  - [Indexing](#)
- [Solr + Hadoop Example](#)

## What is Solr + Hadoop Indexing?

SOLR-1301 implements a contrib module that enables indexing using Hadoop. This methodology is useful in cases where indexing to N master servers simply consumes too much time. This necessarily happens with systems that have terabytes of indexes. In this larger than average use case, simple schema changes would overwhelm the master servers, whereas creating shards in parallel (the best case being a Hadoop node per shard), the overall process will take minutes to hours, instead of days.

## Architecture

### CSVIndexer

CSVIndexer is provided as an example, though for your own application, you will need to create your own CSVIndexer like class. CSVIndexer extends `org.apache.hadoop.conf.Configured` to be instantiated via the Hadoop command line. Your CSVIndexer like class will set your custom mapper and set the output format as [SolrOutputFormat](#). Use the [SolrDocumentConverter.setSolrDocumentConverter](#) to set your custom [SolrDocumentConverter](#).

### SolrDocumentConverter

Implement this class to convert an object from a proprietary format into a [SolrInputDocument](#) that may be indexed into Solr.

### Heartbeater

Hadoop will try to kill a running task if it doesn't receive a periodic heartbeat. This is why in a background thread the [HeartBeater](#) class continuously notifies Hadoop that the current task is still executing. This is necessary with Solr and Lucene because some tasks such as an optimizing a large shard can take several minutes to several hours.

### Number of Reduce Tasks

The number of reduce tasks is a positive integer that will define the number of shards created. Setting the number of reduce tasks is important because you will need to choose a number that creates shards that are not too large and not too small.

### Zipped Shards

It is recommended that shards in HDFS are stored in zip format. To turn this setting on, use [SolrOutputFormat.setOutputZipFormat](#) on the Hadoop [JobConf](#).

### Indexing

An [EmbeddedSolrServer](#) is created, pointed at a local filesystem core (not an HDFS filesystem due to inherent limitations making it not compatible with creating Lucene indexes). Once all Solr documents are written to the embedded core, the index is optimized, then zipped directly to HDFS.

## Solr + Hadoop Example