

KattaIntegration

Introduction

Katta integration with Solr allows Hadoop indexing into shards, which are replicated to N nodes/servers of a Solr cluster. This is useful for large Solr clusters that require failover, replication and the ability to provision shards dynamically. Katta uses Zookeeper to coordinate the creation and deployment of shards to Solr servers.

See <http://issues.apache.org/jira/browse/SOLR-1395>

See <http://sourceforge.net/projects/katta/>

See <http://hadoop.apache.org/zookeeper>

Features

- Uses Hadoop RPC which is implemented with non-blocking (NIO) sockets underneath. This should scale better than the current HTTP approach when there are hundreds of nodes because HTTP can create unnecessary overhead.
- All current distributed Solr requests function properly with no changes
- Incremental indexing may be accomplished by creating new shards and deploying them into the Katta cluster. The alternative method is to update a shard deployed on a Solr server (using the Solr normal XML over HTTP interface). On commit, the newly updated shard would be uploaded back into the Katta cluster, and the old version of the shard removed.
- Solr Katta has built in failover