About Apache Nutch

Table of contents

1 Overview

Apache Nutch is an open source web-search software project. Stemming from Apache Lucene, it now builds on Apache Solr adding web-specifics, such as a crawler, a link-graph database and parsing support handled by Apache Tika for HTML and and array other document formats.

Apache Nutch can run on a single machine, but gains a lot of its strength from running in a Hadoop cluster

The system can be enhanced (eg other document formats can be parsed) using a highly flexible, easily extensible and thoroughly maintained plugin infrastructure.

For more information about Apache Nutch, please see the Nutch wiki.