HDFS Connector Introduction

The Hadoop Distributed File system (HDFS) connector will crawl content from any given HDFS Cluster using the WebHDFS http interface.

Features

Some of the features of the HDFS connector include:

- Performs incremental crawling (so that only new/updated documents are indexed)
- Metadata extraction
- Is search engine independent
- Runs from any machine with HTTP access to the given HDFS Namenode
- Filters the crawled documents by paths (including file names) using regex patterns
- Supports Kerberized Clusters by using a delegation token.
- Supports Archive file processing; for more information, visit Archive files processing

WebHDFS Operations

Only two operations are used by this connector:

http://<host>:<port>/webhdfs/v1/<path>?op=OPEN

- Used to fetch the file data to be used to extract its content.

http://<host>:<port>/webhdfs/v1/<path>?op=LISTSTATUS

- Used to scan a directory and get relevant file information like:
    - Last-Modified dates for incremental crawls.
    - Group and Owner