IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, June 2022
Impact Factor: 6.252

Web Crawling and Indexing using Apache Nutch and Elastic Search

Vipul Sharma
UG Student, Department of Information Technology
Dronacharya College of Engineering, Gurgaon, Haryana, India

I. INTRODUCTION

Apache Nutch is a highly extensible and scalable open source web crawler software project. the project comprises two codebases, namely:

Nutch 1.x: Nutch 1.x is an active branch, well matured, production ready crawler. This branch enables fine grained configuration, this branch does not have good datastorage options as compared to 2.x branch. But it is more stable, efficient and fast .

Nutch 2.x (INACTIVE): Nutch 2.x is an inactive alternative taking direct inspiration from 1.x, but it differs in one key area; datastorage options, it provides several data stores by using Apache Gora. We can implement Nosql database,Mysql,Hbase,etc. No more releases or bug fixes are anticipated for this branch.

Apache Nutch is an open-source Web search engine that aims to index the World Wide Web as effectively as commercial search engines. Nutch provides facilities for fetching,parsing, indexing, urlfilters for custom implementations. It has a highly modular architecture, which means developers can create plug-ins for audio-video-image parsing, data retrieval, sitemap crawling. Nutch is implemented in java and it's code is open source , thus we can run nutch on many operating systems and hardware configurations. Nutch is configurable and plugin friendly, it supports elastic search and solr indexing platforms. We can run it using command line interface which is the preferred way, also we can configure Eclipse if we need a friendly user interface,but it gets a bit complicated when exceptions arise.

II. NUTCH ARCHITECTURE

The core architecture of apache nutch includes following components:

Configuring Nutch Parameters: In this step we configure nutch according to our crawling requirements, we set up the core structure, add plugins like elastic search and kibana, host id, proxy, url filters, etc

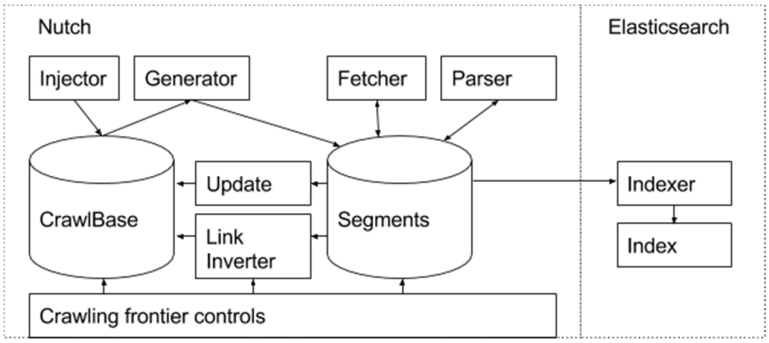
Injecting URLS: A text file named seeds.txt is created, all the urls to be crawled are added in each line.After that this file is injected to the crawlbase database for url injection,urlfiltering.

Generating Segments: Based on the base urls found in seeds.txt, segments are created, for each child node urls, another segment is created. Each segment contains the data crawled from the particular node url. Generated segments are then updated in crawlbase.

Fetching: Urls generated are fetched and data is inserted in segments.

Parsing: Fetched urls are parsed and crawlbase is updated..

Indexing: All the crawled data is indexed to the indexer used, i.e.. elastic search. After that we can also perform operations on elastic search to get the better search results.



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, June 2022
Impact Factor: 6.252

2.1 Crawling

Crawling is the process that navigates and retrieves the all the information in web pages, depending upon the search configurations. Crawling is a complex process. The most efficient way to crawl a particular website is to follow the sitemap rules, many web sites are difficult to crawl, but following the sitemap we can crawl all the links, sublinks. There are several bots, crawlers that are blocked by the website host.

The performance of crawling is limited by the bandwidth of the network between the host system (crawler) and the world wide web traffic. A single crawling cycle consists of injecting urls, generating segments, fetching, parsing those for links, then updating the crawldb.

2.2 Indexing

Indexing is an efficient way of organizing, retrieving, and viewing the data. It is a way to optimize the performance of a database by minimizing the number of disk accesses required when a query is processed. It is a data structure technique which is used to quickly locate and access the data in a database. We need two fields to search in an index, “key” and “value”.We can use either elastic search or Apache solr for indexing, both are good for indexing.

2.2.1 Elastic Search

Elastic search is a distributed search and analytics engine used for Indexing of data , performing analytical operations. It provides fast search responses and analytical suggestions.It is a collection of co-related documents. Data is stored in Json format that has ‘key’, ‘value’ pair. Elastic search uses inverted index data structure which can list each unique word that appears in a document and can identify all of the documents , in which the word occurs.

During the process, Elasticsearch stores documents and builds an inverted index to make the data searchable in real-time.

2.2.2 Kibana

Kibana is a tool for elastic search that is used for data visualization and management. It provides real time histograms, graphical view, pie charts, maps, etc. Kibana also includes advanced applications such as Canvas, which allows users to create custom dynamic infographics based on their data, and Elastic Maps for visualizing geospatial data