IJARSCT

ISSN (Online) 2581-9429

IJARSCT
Impact Factor: 6.252

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Volume 2, Issue 6, June 2022

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.

DOI: 10.48175/568