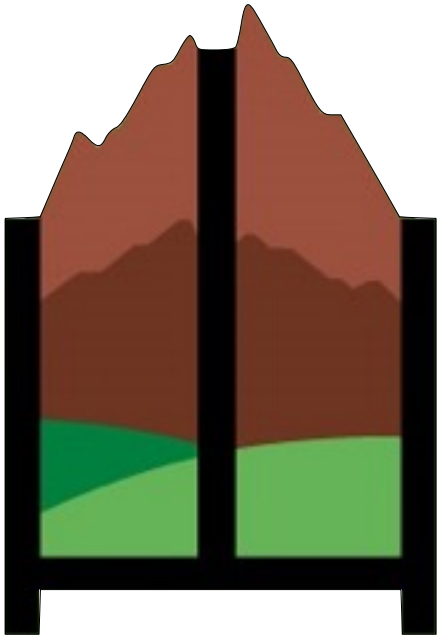


山 SHĀN:

Scalable Search and Web Crawling

Prepared by: Carlos López Roa

November 10, 2016

EXECUTIVE SUMMARY

Objective

To implement a scalable framework for the general task of indexing unstructured documents and enriching them with related documents in the web.

Goals

A. Implement a solution to index unstructured documents
B. Implement a solution to web crawl in the world wide web items related to the unstructured documents
C. Implement an information retrieval solution to correctly answer custom queries in both data sources
D. Provide a scalable architecture based in distributed systems

Solution

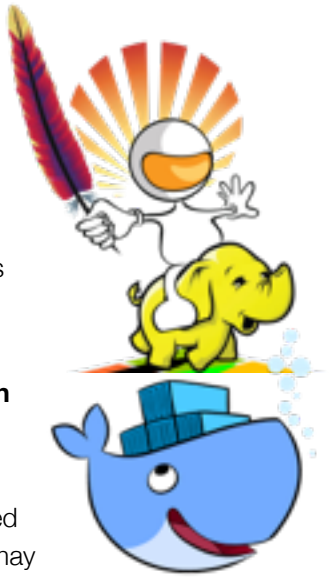
Using the a modular design and taking components from the Apache Software foundation, we may propose to use a Stack composed of:

	Apache Lucene for scalable high performance indexing
	Apache Nutch for highly scalable web crawling
	Apache Solr for blazing-fast search platform
	Apache Hadoop for distributing the work across different nodes
	Apache Web Server for managing url petitions and serve files
	Docker for building shipping and running the different virtual instances

Project Outline

1. Infrastructure Setup
2. Programming of methods
3. Feature enrichment
4. Cloud deployment and testing

Proposed architecture



The proposed stack consists in a container based micro service structure, where instead of a monolithic application in a single stack, a swarm of services will provide independent functions. This helps to provide scalability and resilience.

Hence, a multi instance container based in **docker** will serve as common operating system infrastructure.

- A group of containers will provide Map Reduce support using **Hadoop** installations
- The indexing engines will be served in independent containers running **Solr**
- And web crawling tasks will be done by another group of containers running **Nutch**
- A group of container will respond user petitions using **Apache Web server**

It's important to note that each container can be instantiated on demand and terminated upon the completion of it's task, hence containers with Nutch or Hadoop installations may be instantiated or terminated on request.

Proposed Infrastructure

The development can be done in a personal laptop and the deployment can be done in Cloud infrastructure with low per-hour cost.

A general framework

The proposed framework can be implemented in several case studies. A benchmark case study that is proposed is to index the **Wikipedia** database and enrich it with external links via web crawling. This information will be presented in a Google-like interface with metadata filtering. Performance measures can be analysed.

About the name

Shān (山) is the chinese character for mountain. It can also be composed concatenating the first letter of the components: **S**olr **H**adoop **A**pache **N**utch.