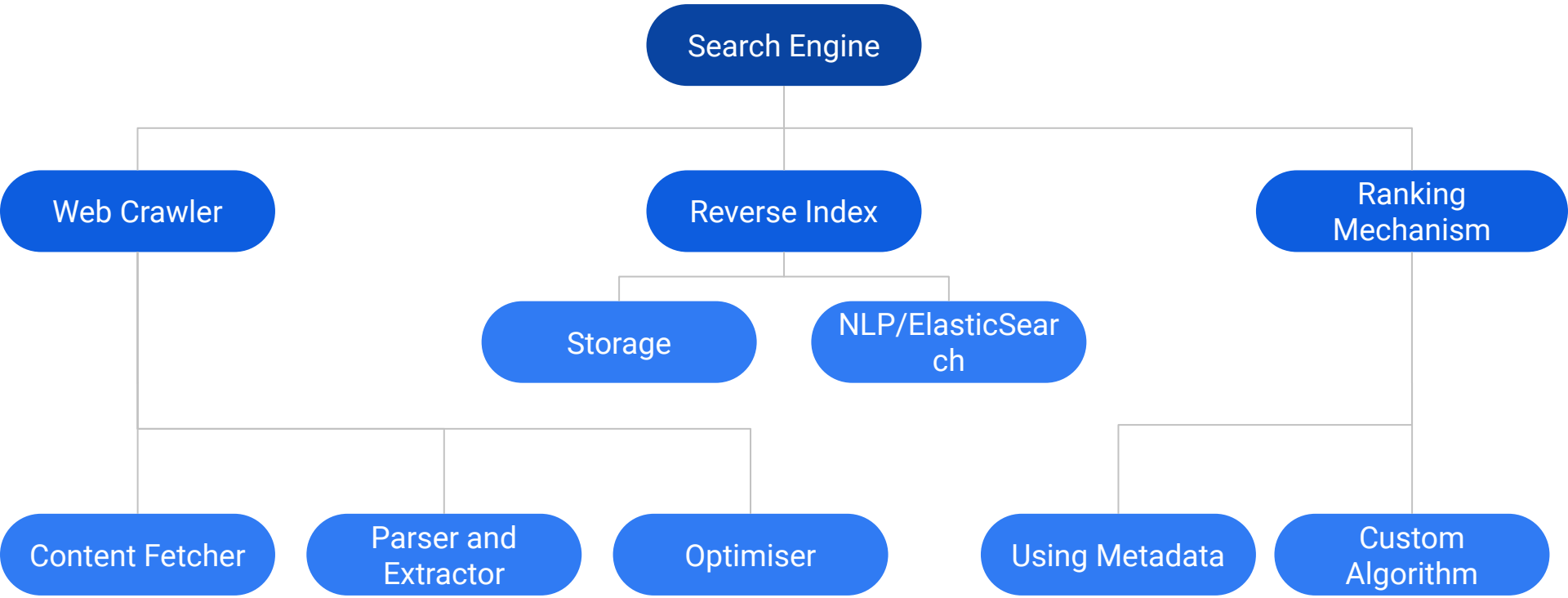
Search Engine

Aliabbas Merchant
Niraj Dhotre
Nehanshu Tripathi


Search Engine
Web Crawler
Reverse Index
Ranking Mechanism
Storage
NLP/ElasticSearch
Content Fetcher
Parser and Extractor
Optimiser
Using Metadata
Custom Algorithm

Web Crawler

1. Content fetcher: A program that takes in a URL and a few parameters and then recursively fetches content of the pages linked to, from that URL
2. Parser and Extractor: To parse the content and to extract the links and content and other data
3. Optimisations: Since we are dealing with data in a huge amount and scale, we need to make our crawler highly efficient, by employing parallel and distributed computing
4. Storage: How to efficiently store the fetched data

Reverse Index

1. Storage: How to efficiently store the extracted and indexed data
2. NLP: To understand the important keywords of the page
3. ElasticSearch: A software that helps to store, search, and analyze at scale.

Ranking Mechanism

Still not sure about this, but one or more of the following can be used:

1. Using meta-data (number of links, etc)
2. Some Custom Algorithm
3. NLP

Proposed Stages and Timeline

Stage	Approximate Start Date	Approximate End Date
Simple Python Crawler, and learning Golang	1st October, 2019	10th October, 2019
Crawler using golang	10th October, 2019	20th October, 2019
Extracting content from the pages	20th October, 2019	10th November, 2019
Optimising the crawler	10th November, 2019	25th November, 2019
Storage of the extracted data	10th December, 2019	25th December, 2019
Ranking mechanism	25th December, 2019	30th January, 2020
Final project	30th January, 2020	10th February, 2020

Resources Required from COE

1. Fast and Reliable Internet connection, for large scale web crawling
2. Parallel computing architecture
3. High speed storage disk
4. Large capacity memory with high access and storage speed

Team Details

1. Aliabbas Merchant - TY, IT
   merchantaliabbas@gmail.com
   9892875512

2. Niraj Dhotre - SY, Electronics
   nirajdhotre123@gmail.com
   91676 14065

3. Nehanshu Tripathi - SY, Electronics
   nehanshu.tripathi17@gmail.com
   77100 42304

Resources and References

1. https://www.quora.com/How-do-you-build-a-search-engine-from-scratch-What%E2%80%99s-the-best-technology-stack-for-this
2. https://www.quora.com/How-can-I-build-a-web-crawler-from-scratch-What-language-or-framework-would-you-recommend
3. http://dl.acm.org/citation.cfm?id=1734789
4. http://link.springer.com/chapter/10.1007/978-3-662-10874-1_7
5. http://dl.acm.org/citation.cfm?id=598684.598733