**SEARCH ENGINE – SYNOPSIS**

The idea behind a search engine is reverse indexing.

Considering the whole of the Internet as a huge dynamic database, it can be considered a 2-column table having URL and the page contents as its respective fields.

Contrary to Indexing where the URL would be used as primary key and the keywords in the webpages as individual records, in reverse indexing, we use keywords as primary keys and URLs as records corresponding to their webpages.

In case of a simple search, this technique as a standalone would be effective. But when the search query consists of multiple keywords, we filter stopwords and proceed with the updated query.

There are four core activities in a search engine: crawling, indexing, ranking and query serving.

**Building blocks of a search engine**

The following components are typical building blocks of a search engine:  
  
**Crawler**

* Select documents to index
* Select parts of the document you want to index
* Regularly update the index and rankings if documents have changed, are added or deleted
* Follow references to other documents
* Deal with duplicates and canonization

**Parser**

* Remove HTML markup and extract text
* Break text into words
* Normalize capitalization, hyphens and umlauts

**Ranking**

* Rank the documents by relevance, time etc.

**Query processing**

* Parse query into terms
* Parse query for phrase and boolean search operators
* Deal with typos
* Create query suggestions
* Implement substring search / instant search

**Search**

* Find documents matching the query and additional filters, sorted relevance
* Layout the search result lists

**Index**

* Store the documents in a data structure which allows scalable, fast and boolean searching

**Caching**  
**Load Balancing**  
**Redundancy**  
**Analytics**

