Information Retrieval

Unstructured data: { Web blogs 
  { Web pages 
    { Documents 
      Email 
  
Search for specific keywords: how to order the retrieved documents according to priority of documents by ranking the documents.

Google & Web Crawler: find all words in each web document and build an inverted index similar to index of a textbook where each web document is a page is assigned a unique integer ID (Document ID).

Google Search Engine searched the inverted index for all & user given keywords and identifies the documents. These documents need to be listed according to their importance (to their document rank).

Page Ranking Method:

Page Rank of a Document: directly proportional to the
  - the no. of external references from other documents to this document
  - the Page Rank of those referring documents.
Page Rank of a document inversely proportional to the number of references made by this document to other documents.

Google Data cluster:
- provides fault tolerance at hardware level (storage devices, processors, switches)
- connectivity (fallback connections)