**Part 2**

**Paper 1:** **Web Crawling**

Web crawling is one of the most important applications of the Internet. Web crawling requests great capacity of storage space and low hardware costs. Web crawling is a process that automatically obtains information from the Web pages through the link relationships between them and expands to the entire Web. This process is mainly done by the Web Crawler which usually consists of spider, controller and original page library. The controller judges and distributes the URLs from the URL database to control the spider crawling the other pages until the URL database is empty.

**Paper 2: NoSQL MongoDB to an SQL**

NoSQL databases are non-relational, hence, very different from SQL databases. This means they are easier to manage and they provide a higher level of flexibility with newer data models. The open source nature of NoSQL databases makes them an appealing solution for smaller organizations with limited budgets. NoSQL database experts often use elastic scalability as a major selling point of NoSQL. NoSQL databases are designed to function on full throttle even with low-cost hardware. SQL databases use structured query language (SQL) for defining and manipulating data. A NoSQL database, has dynamic schema for unstructured data, and data is stored in many ways: it can be column-oriented, document-oriented, graph-based or organized as a KeyValue store. SQL databases are vertically scalable, which means that you can increase the load on a single server by increasing things like CPU, RAM or SSD. NoSQL databases, on the other hand, are horizontally scalable.

**Paper 3: Data Aggregation System**

Data Aggregation System is a system for information retrieval and aggregation from heterogenous sources of relational and non-relational data for the Compact Muon Solenoid experiment on the CERN Large Hadron Collider. Data Aggregation System is designed as an additional layer on top of a heterogenous ecosystem of existing CMS data services. Data Aggregation System assumes that the data in the cache can be recreated from the original data sources at any time, and thus it can function with a limited amount of space by deleting old data, nor does it require backup of this space. DAS is designed as an entirely read-only system, with no ability to write data back to the underlying services, which simplifies the requirements for authentication and validation that this would otherwise require. DAS queries are made in a custom text-based language.