**Assignment 1 (Part 2)**

**Application of NoSQL Database in Web Crawling**

**Introduction:**

Web crawling is the process of collecting, filtering, and storing information from the internet in order to build a search engine. As a result, the database needed for this must have a huge storage capacity and a cheap hardware cost. Scalability, performance, and availability are the primary goals. When dealing with huge volumes of data, traditional relational databases store data in 2D tables, which is inefficient. Many firms utilize their own databases, such as Bigtable from Google. Furthermore, the paper describes about the comparison between the relational i.e., SQL and non-relational i.e. No-SQL by incorporating them into a solution for a specific meteorological BBS data collection system.

**Web Crawling System:**

A web crawler is a process that searches the internet for material, examines it, and stores it in databases. It works by widening the connective interaction between webpages to the full page. It makes use of a spider to scan web pages, this process usually takes some time so that multiple spiders can process simultaneously to improve the crawl rate and save URLs in URL databases, a controller to operate the spider by deciding which URLs to read from URL databases, and a page library to record page data.

**Meteorological BBS Information Collection System:**

Meteorological BBS information collecting system filters and gathers postings from typical meteorological BBSs on the Internet, such as clud.weather.com.cn, www.cmabbs.com, and others. It gives a professional meteorological search engine database. The postings are recorded in text files, which are subsequently preserved in databases. The URL and title of the post appear first, followed by "#floor," which indicates the beginning of the first level, "postby," which indicates the person who posted this floor, and the time and content of this floor. "#floor postby time content" will be repeated if there are additional floors. The post's structure includes a set URL and title, as well as unfixed floors. Because each post has a varied number of storeys, their constructions change. The database structure design must be both practical for storing material from txt documents and simple to query.

The system retrieves all of the post's content from the database based on the ID, including all floors.