**Overall Description**:

The database utilizes the jdbm1.0 storage mechanism to store data in a (Key, Value) format, where "|" and "," are used as separators within the values ("|" is for different types of data of a record and "," is for numerous records). Also, different databases are stored separately using different record managers. Hence, there will be 8 Data Base/LG Files in the end (6 in phase one, 2 for storing headings are missing in phase one). We will discuss whether to continue this approach (storing data in different record manager and database) or try to combine all the database into one file. Here are the descriptions for each database:

**PageInfo**: This table contains information about each page crawled by the web crawler. It includes fields such as PageID (unique identifier), PageTitle, URL, LastModificationDate, and SizeofPage.

**PageURLMapping**: This table maps URLs to their respective PageIDs. It helps in quickly retrieving the PageID associated with a given URL.

**PageChildMapping**: This table maps child links to their respective parent pages. It allows efficient tracking of child links associated with each parent page.

**InvertedBodyWord**: This table stores the inverted index for words found in the body of pages. It includes fields like WordID (unique identifier for body words), PageID (ID of the page containing the word), Frequency (frequency of the word in the body), and TFIDF (TF-IDF value of the word).

**BodyWordMapping**: This table maps body words to the pages they appear in along with their frequencies. It facilitates the retrieval of words associated with each page and their frequencies.

**InvertedTitleWord**: Similar to InvertedBodyWord, this table stores the inverted index for words found in the title of pages. It includes fields like TitleWordID, PageID, Frequency, and TFIDF.

**TitleWordMapping**: This table maps title words to the pages they appear in. It allows for the efficient retrieval of title words associated with each page.

**WordMapping**: This table maps body word IDs to their respective words. It helps in retrieving the actual word given its ID.

**Database Design**:

We use the true data types in the design stage. However, we implement the key and value as "String" type for easy implementation. Here are the database designs:

**PageInfo**

Description: Contains information about each page.

| Field Name | Data Type | Description |
|---|---|---|
| PageID | Long | Unique key ID for pages |
| PageTitle | String | Title of the page |
| URL | String | URL of the page |
| LastModificationDate | String | Date of the last modification |
| SizeofPage | Long | Size of the page |
| MaxFreq | Int | Maximum Term Frequency of the page |

Data:

| 1 | COMP4321 Project Website\|www.testing.net\|11/03/2024\|96\|20 |
|---|---|

**PageURLMapping**

Description: Maps URL to their respective PageID.

| Field Name | Data Type | Description |
|---|---|---|
| URL | String | URL of the page |
| PageID | Long | Unique key ID for pages |

Data:

| www.testing.net | 1 |
|---|---|

**PageChildMapping**

Description: Maps child links to their respective parent pages.

| Field Name | Data Type | Description |
|---|---|---|
| PageID | Long | ID of the parent page |
| ChildID | Long | ID of the child pages |

Data:

| 1 | 2, 3 |
|---|---|

## InvertedBodyWord

Description: Stores inverted index for words found in the body of pages.

| Field Name | Data Type | Description |
|---|---|---|
| WordID | Long | Unique identifier for body words |
| PageID | Long | ID of the page containing the word |
| Frequency | Long | Frequency of the word in the body |
| TFIDF | Float | TF-IDF value of the word |

Data:

| 1 | 1\|5\|2.1, 2\|4\|1.8 |
|---|---|

## BodyWordMapping

Description: Maps body words to the pages they appear in.

| Field Name | Data Type | Description |
|---|---|---|
| PageID | Long | ID of the page |
| WordID | Long | ID of the body word |
| Frequncy | Integer | Frequency of the word |

Data:

| 1 | 1\|3, 3\|4, 6\|5, |
|---|---|

## InvertedTitleWord

Description: Stores inverted index for words found in the title of pages.

| Field Name | Data Type | Description |
|---|---|---|
| TitleWordID | Long | Unique identifier for title words |
| PageID | Long | ID of the page containing the word |
| Frequency | Long | Frequency of the word in the title |
| TFIDF | Float | TF-IDF value of the word |

Data:

| 3 | 1\|5\|2.1, 2\|4\|1.8 |
|---|---|

## TitleWordMapping

Description: Maps title words to the pages they appear in.

| Field Name | Data Type | Description |
|---|---|---|
| PageID | Long | ID of the page |

| TitleWordID | Long | ID of the title word |
|---|---|---|

Data:

| 1 | 1, 3, 6 |
|---|---|

## WordMapping

Description: Maps body word IDs to their respective words.

| Field Name | Data Type | Description |
|---|---|---|
| Word | String | Body word |
| WordID | Long | ID of the body word |

Data:

| Apple | 1 |
|---|---|