## 1 User Information

### Data Description

The corona-out-3.csv file primarily contains data related to tweets about the Coronavirus. The data reflects two types of actions, tweets and retweets. It also includes information about the users who performed these actions, such as User ID, User Name, User Screen Name, User Location, User Website URL, and User Description, among others. In this search application, user information is mainly used for searching and determining user interface states. Most of the information is redundant and can be dropped directly.

### 1.2 MySQL

In this section, MySQL is employed to store user information.

MySQL is an open-source relational database management system (RDBMS) that has gained widespread popularity due to its performance, stability, and ease of use. MySQL is commonly used for a variety of applications, including web development, data warehousing, e-commerce, and more.

Benefits of using MySQL to store user information:

* MySQL is a relational database management system that organizes data in tables with rows and columns. This structure is well-suited for handling user information, which consists of relationships with other data entities.
* MySQL can handle a large amount of data and scale to accommodate the growth of user information over time. It can support millions of records while maintaining efficient query performance.
* MySQL offers built-in mechanisms, such as primary keys, foreign keys, and transactions, to ensure data consistency and integrity.

### 1.3 Schema

Based on the search application’s features, the four properties below is chosen from user imformation:

|  |  |  |
| --- | --- | --- |
| ***id\_str* (Primary key)** | String | The user's unique identifier. In the context of Twitter's database, the user ID uniquely represents a user. |
| ***name*** | String | The user's real name. This field is not unique within Twitter's database, meaning multiple users may have the same name. |
| ***screen\_name*** | String | The user's username. It is a unique and recognizable string used to represent the user on Twitter. No two users can have the same username. |
| ***protected*** | Boolean | Describe whether the account is protected or not. When the “protected” value is set to “True”, only approved followers can view the user’s tweets and retweets. When the value is set to “False”, it means that the user’s tweets and retweets are visible to everyone. |

The first three: *id*, *name* and *screen\_name* are used in searching users. And *protected* is used in identifying user interface states.

### 1.3 Optimization

Optimizing data storage formats to enhance performance.

#### 1.3.1 Dropping irrelevant fields

Removing fields that are not required for the search application to reduce storage space and minimize the amount of data when processing search queries.

As shown above, only four fields are retained in this application.

#### 1.3.2 Indexing

Creating indexes on tables based on query requirements can significantly increase query speed. It is important to carefully analyze the most common query patterns and identify the columns that would benefit most from indexing.

#### 1.3.3 Trade-off

Optimizing data storage comes with certain trade-offs:

* Storage space: Indexes consume additional storage space. As the number of indexes increases, so does the storage requirement for maintaining those indexes.
* Insertion/Update performance: When inserting or updating data, the presence of indexes might decrease performance, as the index structures need to be updated as well. The more indexes there are, the more overhead is involved in keeping them up to date during data modifications.
* it is important to carefully evaluate which fields are truly irrelevant before dropping them.