3.1 Conceptual Design

Name: Rajat Nepal
Student ID: 22210602
Assignment: COMP 30770 Final Project

Requirements for Database

We need to answer the following questions in order to outline the requirements for our database (these questions come from lecture slides):

1. What will my Database be used for?
    a. The database will be used to determine the trends of Machine Learning in the last few years and report them back. The users accessing the database will be using it in order to derive summary information that you will report back to stakeholders
2. Who will use my database?
    a. Me and my technical colleagues at the research and development department of Big Four Inc.
3. What are the skills of the people using this database?
    a. The users of this database have sufficient technical skills to perform basic CRUD operations
4. Functional requirements: what operations do we need to perform?
    a. We need to perform basic CRUD operations. We will also need to perform joins, group by, and count queries. Finally, we will need to do data processing to complete a variety of tasks
5. Data requirements?
    a. We need to store relatively small amounts of data (just small website information)
    b. We need our information to be linked as well (each wikipedia article body is related to a reference file)

Since our data will be mostly structured, and have relationships, we should have the following requirements:

1. **Scalability:** As the field of machine learning is rapidly evolving, the database should be able to accommodate large volumes of data and be scalable to meet the increasing demands of the research and development department.

2. **Security:** The database should ensure the security and privacy of the data it holds. It should have proper access controls to prevent unauthorized access to sensitive information, and data encryption should be used to protect data in transit and at rest. (This is beyond the scope of this class, but I thought I would still include it)

3. **Performance:** The database should be designed to ensure optimal performance, with fast query processing times and efficient data retrieval. This is especially important in the field of machine learning, where large datasets are often used and complex queries are run.

4. **Reliability:** The database should be reliable, with mechanisms in place to prevent data loss or corruption. It should also have backup and recovery mechanisms in case of any system failures. (This is again beyond the scope of this class in this context, but I figured it should still be included)

5. **Usability:** The database should be user-friendly, with an intuitive interface that allows the research and development team to easily query and analyze the data.

6. **Maintenance:** The database should be easy to maintain and update, with tools and mechanisms in place to monitor performance, detect errors, and identify areas for improvement. Regular maintenance and updates should be done to keep the database up-to-date with the latest developments in the field of machine learning.

Type of Database: Relational or Document based

Based on the structure of the parsed data, I believe a relational database is most relevant. The entity in this case seems very apparent: Articles. Each entity also has specific attributes, consistent amongst all entities of that type. An Article entity will have a Year, Month, Day, an Article Body, and References. Both relational and document-based databases support basic entities and attributes. However, there are a few advantages of a relational database over a document-based one (at least in this case):

1. All the data is structured. Each Article entity will have the same attributes.
2. We are storing large amounts of text in the Article Body and References attributes. MySQL is able to store this amount of data in an attribute. However, if the wikipedia pages were more than 4 GBs of text (which is not the case), a document-based database may be more useful.
3. While MySQL is not as scalable as MongoDB, at most, there will be 1 update per day to the page (which even that is unlikely). Since we are looking at trends in machine learning over time, day-to-day updates are not needed. With relatively few updates, our data does not need to be highly-scalable.
4. We are performing complex queries, which are much easier in MySQL. Furthermore, many technical staff are going to be users of this database, not just me, the creator. As shown in the slides, relational databases are more widely used than document based, and therefore will have a lower learning curve for our database engineers.
5. There may be relations in the future. Perhaps a specific company makes a lot of progress in the field (Like ChatGPT of OpenAI did with AI). We may want to link our articles to other articles using relationships. This would be much easier than using a document-based database.

For these three reasons, we should use a relational database (MySQL in this case).

MySQL Database

Entity: Article

Attributes:

“articleId”: <ObjectID Type 12-byte Hexadecimal value>,

“year”: <Integer>,

“month”: <Integer>,

“day”: <Integer>,

“articleBody”: <String>,

“references”: <String>

No relationships