Question-Answering System

# Database Design Document

Version 0.1

02/28/2019

Table of Contents

[1. Overview 1](#__RefHeading___Toc9045_565685251)

[2. Assumptions/Constraints/Risks 2](#__RefHeading___Toc9047_565685251)

[2.1 Assumptions 2](#__RefHeading___Toc9049_565685251)

[2.2 Constraints 2](#__RefHeading___Toc9051_565685251)

[2.3 Risks 2](#__RefHeading___Toc9053_565685251)

[3. Design Decisions 3](#__RefHeading___Toc9055_565685251)

[3.1 Key Factors Influencing Design 3](#__RefHeading___Toc9057_565685251)

[3.2 Functional Design Decisions 3](#__RefHeading___Toc9059_565685251)

[3.3 Database Management System Decisions 3](#__RefHeading___Toc9061_565685251)

[3.4 Security and Privacy Design Decisions 3](#__RefHeading___Toc9063_565685251)

[3.5 Performance and Maintenance Design Decisions 3](#__RefHeading___Toc9065_565685251)

[4. Detailed Database Design 5](#__RefHeading___Toc9067_565685251)

[4.1 Roles and Responsibilities 6](#__RefHeading___Toc9069_565685251)

[4.2 Performance Monitoring and Database Efficiency 6](#__RefHeading___Toc9075_565685251)

[4.2.1 Operational Implications 6](#__RefHeading___Toc9077_565685251)

[4.2.2 Data Transfer Requirements 6](#__RefHeading___Toc9079_565685251)

[4.2.3 Data Formats 6](#__RefHeading___Toc9081_565685251)

[Appendix A: Acronyms 7](#__RefHeading___Toc9083_565685251)

List of Figures

No table of figures entries found.

List of Tables

[Table 1 - Acronyms 9](#Table!1|sequence)

## Overview

Instructions: Briefly introduce the system context and the basic design approach or organization, including dependencies on other systems. Identify if the database will supersede or interface with other databases, and specifically identify them if applicable. Also identify interfaces with other systems to the extent that they significantly impact the database design. Discuss the background to the project, if this will help understand the functionality supported by the database design contained in this document.

In a Question-Answering problem, a question needs to be stored with its respective answers. When a new question arrives, ideally, we want to check if a similar question was already asked before. If so, the question will not be stored as a new one. Instead, the answers for the similar question will be output to the user. This is a NLP (Natural Language Processing) solution for avoiding storing all questions asked in question-and-answer sites such Quora and Stack Overflow. Nevertheless, even using this NLP solution, the set of questions/answers might become very big overtime. Therefore, a Big Data architecture would be the most suitable way of dealing with the increasing amount of data to be stored and accessed in this type of system.

## Assumptions/Constraints/Risks

### Assumptions

Instructions: Describe any assumptions or dependencies regarding the database design for the system. These may concern such issues as: related software or hardware, operating systems, or end-user characteristics.

### Constraints

Instructions: Describe any limitations or constraints that have a significant impact on the database design for the system.

### Risks

Instructions: Describe any risks associated with the database design and proposed mitigation strategies.

## Design Decisions

Instructions: Utilizing the following subsections, describe decisions made that impact the proposed database design. This should include the platform and database management system (DBMS) chosen for the project. Include any other information relevant to the database design decisions (e.g., Data Conversion Plan, Service Level Agreements (SLAs)). The Design Decisions section is written at a higher level than the subsequent Detailed Database Design section, and provides an understanding and rationale for the content in the Detailed Database Design section.

### Key Factors Influencing Design

Instructions: Describe key functional or non-functional requirements that influenced the design. If all such decisions are explicit in the requirements, this section shall so state. Design decisions that respond to requirements designated as critical (e.g., those for performance, availability, security, or privacy) shall be placed in separate subparagraphs.

### Functional Design Decisions

Instructions: Describe decisions about how the database will behave in meeting its requirements from a user's point of view (i.e., functionality of the database from an application perspective), ignoring internal implementation, and any other decisions affecting further design of the database. Include decisions regarding inputs the database will accept and outputs (displays, reports, messages, responses, etc.) it will need to support, including interfaces with other systems. Describe the general types of processing (sequential versus random for inserts, updates, deletes and queries) required both for data entering the database, and data most frequently accessed. Also include decisions on how databases/data files will appear to the user.

### Database Management System Decisions

Instructions: Describe design decisions regarding the DBMS intended for the initial implementation. Provide the name of the DBMS, the reason for selection, and the type of flexibility built into the database for adapting to changing requirements.

### Security and Privacy Design Decisions

Instructions: Describe design decisions on the levels and types of security and privacy to be offered by the database. General descriptions of classifications of users and their general access rights should be included.

### Performance and Maintenance Design Decisions

Instructions: Describe how performance and availability requirements will be met. Examples include:

* Describe design decisions on database distribution (such as client/server), master database file updates and maintenance, including maintaining consistency, establishing/ reestablishing and maintaining synchronization, enforcing integrity and business rules.
* Describe design decisions to address concurrence issues (e.g., how the data are partitioned or distributed to support multiple applications or competing update functions, if applicable).
* Describe design decisions to support Service Level Agreements (SLAs) for key functions supported by the database.
* Describe design decisions on backup and restoration including data and process distribution strategies, permissible actions during backup and restoration, and special considerations for new or non-standard technologies such as video and sound. Describe the impact this maintenance will have on availability.
* Describe design decisions on data reorganization (i.e., repacking, sorting, table and index maintenance), synchronization, and consistency, including automated disk management and space reclamation considerations, optimizing strategies and considerations, storage and size considerations (e.g., future expansion), and population of the database and capture of legacy data. Describe the impact this maintenance will have on availability.
* Describe design decisions to support purging and/or archiving of data to ensure performance and storage objectives are met. Describe the impact this maintenance will have on availability. Describe any needs to recall archived data back into the database.

## Detailed Database Design

Instructions: Describe the design of all DBMS structure associated with the system. The headings and sub-headings in this section should be structured according to the information to be presented, and may include discussions about or references to the following:

* Conceptual Data Model (CDM)
* Logical Data Model (LDM) and LDM Entity Relationship Diagram (ERD).
* Physical Data Model (PDM) with a description of the DBMS schemas, sub-schemas, records, sets, tables.
* A comprehensive Data Dictionary showing data stores, data element name, type, length, source, constraints, validation rules, maintenance (create, read, update, delete (CRUD) capability), audit and data masking requirements, expected data volumes, life expectancy of the data, information life-cycle management strategy or at least an archiving strategy, outputs, aliases, and description.
* Planned implementation factors (e.g., distribution and synchronization) that impact the design.
* Estimate of the DBMS file size or volume of data per entity.
* Definition of the update frequency of the database tables, views, files, areas, records, sets, and data pages. Also provide an estimate of the number of transactions, if the database is an online transaction-based system.

The detailed database design information can be included as an appendix, such as DDLs, which would be referenced here.

Regarding the database definition, different options have been considered. For instance, a Key-value-based NoSQL would be a good choice if the system allowed only one answer per question. However, a same question can have more than one answer. In a document-based NoSQL (or in a simple file-based system), a question with its respective answers could be grouped in a same data structure (e.g., a json file), under a same key. However, adding new answers to a question (as well as retrieving answers from a question) wouldn't be trivial, since the answers need to be ranked by using a voting scheme. Therefore, a traditional Relational DB (MySQL) has been chosen for our proposed Question-Answering problem.

As for the data properties, despite the choice of using a Relational Data Base, this system can behave well with a BASE Consistency Model. In the following, we define our conceptual (Figure 1) and logical models (Figure 2).

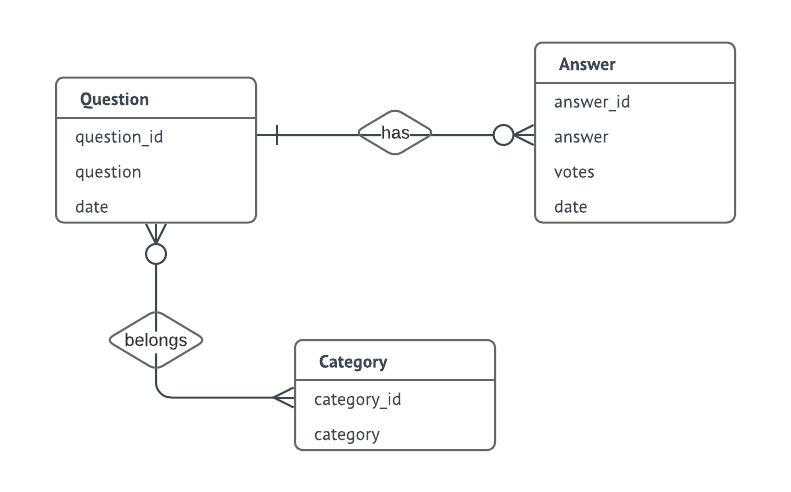


Figure 1 – Conceptual Model

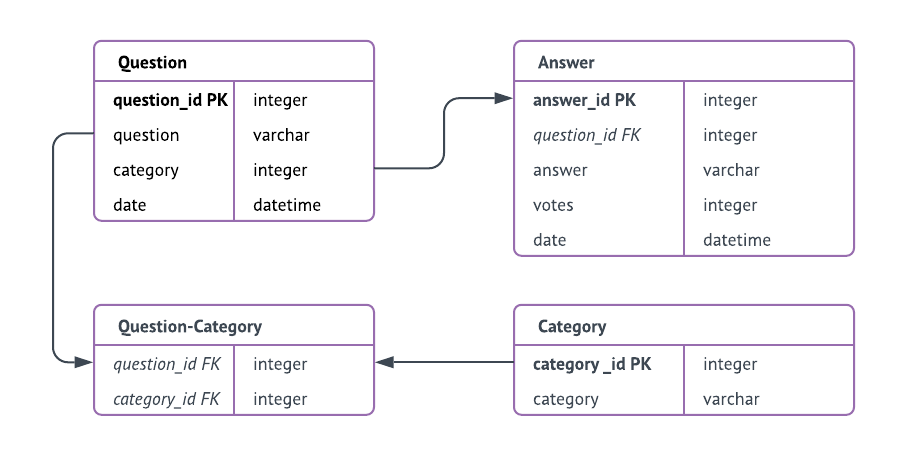


Figure 2 – Logical Model, where all attributes are mandatory (not null)

As for the DDLs, please refer to the SQL script:

<https://github.com/luana-be/CEB1250_repo/blob/master/QUESTION-ANSWERING.sql>

### Roles and Responsibilities

Instructions: Identify the organizations and personnel responsible for the following database administrative functions: database administrator, system administrator, and security administrator. Describe specific administration skill requirements applicable to the database.

### Performance Monitoring and Database Efficiency

#### Operational Implications

Instructions: Describe operational implications of data transfer, refresh and update scenarios and expected windows.

#### Data Transfer Requirements

Instructions: Describe data transfer requirements to and from the software, including data content, format, sequence, volume/frequency and any conversion issues.

#### Data Formats

Instructions: Describe formats of data for both the sending and receiving systems, including the data item names, codes, or abbreviations that are to be interchanged, as well as any units of measure/conversion issues.

Appendix A: Acronyms

Instructions: Provide a list of acronyms and associated literal translations used within the document. List the acronyms in alphabetical order using a tabular format as depicted below.

Table 1 - Acronyms

| Acronym | Literal Translation |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |