Universität Göttingen
Department of Computer Science

Julian Kunkel

Exercise Week 3
HPDA / 2021
240 Minutes Total
Discussion in Week 4: 2021-11-15

---

1. The tasks described in this worksheet are part of the formative assessment. They serve the purpose to prepare you for the examination. We will discuss the solutions during the next **interactive session** after they are handed out – while they fit to the lecture of the week they are handed out, they might be discussed in two weeks time due to the bi-weekly exercise schedule.

2. Make sure to plan your time for the whole sheet carefully. The complete exercise should represent approximately three hours of independent study. The time limit indicates how much time you should spend on each task, and not how much time you may actually need; it is important that you engage with the material and not that you complete all tasks perfectly. Feel free to collaborate and team up.

3. The exercises are designed to challenge you and train you further as guided self-study. The time limit might be too ambitious for you; you may team up with colleagues. It is not an issue as long as you manage to at least partially resolve each task within the time budget. If you (and team) are struggling, reach out for help in Teams! You may also share your thoughts on the channel.

4. We recommend that you create a (private) GIT repository where you store your findings and outcomes while processing the exercises. This portfolio of work could be useful in the future.

Contents

Task 1: Relational Data Schema for Wikipedia Articles (120 min) **1**
1.1 Reminder: Each article should store the following properties 1
1.2 Requirements 2
1.3 Word Distribution 2

Task 2: Data-Warehouse Schema for Analytics Data (120 min) **2**

Task 1: Relational Data Schema for Wikipedia Articles (120 min)

In this task, we expand upon the relational database schema from the last sheet - our Wikipedia data.

1.1 Reminder: Each article should store the following properties

- Title
- Text
- Category
- Links to related articles

The following operations should be possible:

- Access article details based on the article's "title"
- Finding related articles (those that link to one another) from a given article
- Retrieving all articles for one category

1.2 Requirements

Create the ER-diagram for the schema.

Create the SQL for the schema, then insert a few samples, and create some sample queries that implement the operators. Discuss the usefulness of indexes for each of the relations.

Please use sqlite3 to store the data and test the queries. For example, you can create the database as follows:
`$ sqlite3 data.db "create table test (TEXT x);"`
Just running `$ sqlite3 data.db` will open an interactive shell to test SQL queries.

1.3 Word Distribution

Add an additional operator, that shows for a given word the distribution of words, i.e., the counts. Create a view from it that allow directly accessing a given articles word distribution.

Hints

- You find the installation instructions for SQLite in `hpda-samples/install/sqlite.sh`
  GitHub repository: JulianKunkel/hpda-samples.git

Portfolio (directory: `3/db`)

`3/db/schema.pdf`	Your normalized database schema in the form of an ER-diagram, including discussions of chosen indexes and keys.
`3/db/operations.sql`	Each operation implemented as SQL statements, including comments.

Task 2: Data-Warehouse Schema for Analytics Data (120 min)

Create a fact based schema (OLAP-Cube). Based on logs of the Wikipedia website the following data is created:

- IP-address
- country of visitor
- city of visitor
- access date
- time spent on the website
- browser's user agent

Build a useful OLAP Cube schema for the data.

Perform the following steps:

1. Discuss which data should be part of the fact table and which should be an attribute of the dimensions. Consider carefully which attribute should be a fact.
2. Create a star schema to map your OLAP-cube to a relational model. Document SQL Commands for creating your relational model.
3. Write an SQL query for retrieving the time spent on the website by users from a certain country within

a specific month. Assume no aggregation within the dimensions has been done, thus, the query must process all individual facts.

4. Implement the schema using `$ sqlite3` and submit some sample queries.

Portfolio (directory: `3/olap`)

`3/olap/olap-schema.txt`	Detailed description of your OLAP schema (facts, dimensions).
`3/olap/rolap.txt`	SQL queries for creating and querying the star schema based on the OLAP model.