Data management second course assignment

May 2021

In this individual course assignment, your task will be to look at different stages of the database development project in case of a hypothetical case company, and perform different tasks that typically occur in a similar situation. Submit one file with your solution to Moodle. Please be aware that the submitted file will be checked in the Urkund system after submission. The deadline is May 31, 23:59.

Background

In the assignment, you will have to solve several tasks that are part of a case company's database development process. The company, health-analytics.ai, provides consulting services focusing on business analytics and business intelligence projects to organizations that use health care and medical data [1]. In the following, you assume the role of a newly hired data analyst. You just graduated from the university and in your studies you have finished several data analytics and management related courses. The company is convinced of your abilities based on your background and job interview, and as your first project they assign you the very important task of assessing and improving their internal data processing systems. In the subsequent tasks, you will analyse and understand different components of the system and propose new ways to data processing in the company.

Step 1: Formulating the business case (10 points)

While the company offers services to clients that utilize state-of-the-art data visualization tools, Artificial Intelligence and machine learning algorithms, health-analytics.ai's internal data processing and management is far from optimal. For the most part, they still rely on traditional file processing systems (individual data silos for clients, consultants, projects etc.), in many cases some consultants store detailed project data locally in Excel, and only input some final aggregated

[1] In order to get some ideas on the market trends in this domain, you may check the following recent reports: `https://www.accenture.com/_acnmedia/PDF-49/Accenture-Health-Artificial-Intelligence.pdf`, `https://www.zdnet.com/article/data-analytics-machine-learning-and-ai-in-healthcare-in-2021/`

information into the organizational information systems. This structure is not ideal for the main goals that were specified by the management for your current organizational data policy development: (i) understanding the clients in great depth (past, existing and prospective customers, i.e. clients who have interacted with health-analytics.ai through any channels), (ii) understanding consultants performance in more depth (what projects they have performed wells), and (iii) automate the tracking of clients and also projects, whenever possible.

Your task in this step is the following:

- Task 1: write a 1-2 page memo to describe the following for the management: (i) what are the problems with the current approach (traditional file processing system), (ii) why a database approach and data-warehousing serves the company's specified goals better, (iii) what data you will most likely need to create a database, and (iv) what are the main steps to implement a database solution (following a generic systems development life cycle approach).

Step 2: Conceptual design (15 points)

After the management have read your memo and gave green light to a database development project, your next task is to develop the conceptual model. After discussing the relevant entities, the business rules and data that can be easily collected and included in a future database, the following relevant entities emerged:

- Consultants: they can perform two main types of tasks. Business consultants are contacted by a business in order to first determine data analytic needs in the client and provide an estimate for the actual services to be performed. Technical consultants perform services according to the specifications developed by the business consultants. Each consultant can be business, technical or both. In the preliminary plan, the following attributes are proposed to be collected: Employee ID, Name, Address (which is composed of Street, City, State, and Zip Code), Phone number, Date Of Birth, Role (multi-valued: technical, business or both), Number of Years in the company, Type of Business [or businesses] the consultant is experienced in, Technical skills (if any).

- Customer: businesses (hospitals, pharmaceutical companies, care and social services, etc.) that have asked for consulting services. Attributes of customer are Customer ID, Company Name, Address (which is composed of Street, City, State, and Zip Code), Contact Name, Contact Title, Contact Phone, Business Type.

- Location: Customers can have multiple locations. Attributes of location are Customer ID, Location ID (which is unique only for each Customer ID), Address (which is composed of Street, City, State, and Zip Code), Phone, and Building Size.

- Service: analytics service is performed for a customer at one or more locations. Before services are performed, an estimate is prepared. Attributes of service are Service ID, Description, Cost

Some additional business rules were also specified:

- In addition to the entities outlined previously, the following information will need to be stored (these may be entities on their own, or attributes of other entities or relationship, the choice is yours when making the final design): (i) Project Estimates, including information about Date, Amount, Services, and Customer, (ii) Services Performed, including information on Date, Amount, Services, Consultant and Customer. In the design, you may also assume that a customer can have many consultants providing many services. As you wish to track both actual services performed as well as services offered, there should be two relationships between customer, service, and consultant, one to show services performed and one to show services offered as part of the estimate (or alternatively you can include this information as an attribute if your design allows for it).

Based on this information, you need to perform the following tasks:

- Task 1: Considering the main business goals of the data management implementation process, try to assess the information that is provided to you above. Can you think of any other entities, attributes, relationships that could be relevant for the business goals, and the company should be able to collect them in the future?

- Task 2: Create and ER-diagram based on the provided specifications and your suggestions in your first task in this design step. Document the design choices that you make, most importantly what you choose as the identifiers for different entities and how these identifiers are included as attributes of other entities to establish relationships. Furthermore, if for some relationship included in your design, the above specifications do not provide sufficient information on type and cardinality, make your choice and describe your reasoning (i.e. why you specify a relationship as one-to-one, one-to-many, many-to-many, optional or mandatory)

Step 3: Logical design (10 points)

After checking your ER-model proposal, the management is satisfied with the design. The next step is to perform the logical design. While it is clear that for now a relational model and a database management system using relational schema could be sufficient, the management would like to hear your opinion on NoSQL-type of solutions as alternatives. So your first task in this step is:

- Task 1: write a short summary (max 1 page) on the differences between relational and non-relational databases, in what situations you would use

non-relational solutions (in particular about big data), and whether you think in the line of business the company is (and the current business goals of database development, there will be need to transition to NO-SQL solutions in the near future.

As the development is settled for relational design, your main task in this step is the following:

- Task 2: map the ER diagram you developed in the previous step to a relational schema. Be sure to appropriately identify the primary and foreign keys as well as clearly state referential integrity constraints.
- Task 3: assess what normal form your model is in, and if needed, perform the steps to transform it into second normal form at least.

Step 4: Physical database design and performance (10 points)

After the logical design and normalization is completed, the managers are interested in some approaches that can increase the performance of using the database. You mention the concept of denormalization. The managers are very surprised; you just normalized your model, why would you make it worse. Your next task is to explain denormalizaton with some examples:

- Task 1: write a 1 page summary, where you discuss the benefits of denormalization, and try to identify 2-3 examples in the logical model you created above, where you would want to introduce denormalization, and why this denormalization would help in monitoring and analysing data related to business goals (Hint: try to think about queries (no need to formulate them, just consider what they would do) that you would frequently run after the database is in place to address business goals, e.g. profitability analysis of consultant based on some attributes, analysis of revenue from different service types etc.; after you identify these important queries, think about what joins would be performed frequently, and how denormalization would create new tables that would make joins unnecessary)

Additionally to the above task, there is one further thing you need to analyse in relation to physical design: the management asks you to estimate some usage statistics.

- Task 2: you need to perform some basic usage analysis with the following parameters: the company has approx. 200 consultants, there are 100 customers (hospitals, pharmaceutical companies etc.) with each approx. 3-4 locations. In the company, consultants run queries on analysing the status of different provided service types on each of the locations on an hourly basis. How many queries would this mean on a daily basis? Does this problem give you further ideas for denormalization?

Step 5: Data Warehousing and data integration (5 points)

As the last step of the design, you need to consider the data warehousing solution that would be optimal for the business. There are already some data marts in separate parts of the company (sales, finance etc.), now you need to explain to the management the benefits of the Enterprise Data Warehousing (EDW) architecture.

- Task 1: describe in max. 1 page, what are the benefits of EDW over Independent Data Marts. As a note, also consider in your description whether in this line of business and for the specified goals you need real-time data warehousing or not.

As the last task, in order to control data quality, you need to perform a basic check:

- Task 2: as you know from your studies, one main reason for poor data quality is manual entry of information. The management asks you as the final step to assess which of the attributes in your model will most likely require manual entry, i.e. there is no way to automate the collection or even create a simple drop-down with a limited amount of choices (hint: look for attributes that require detailed description)