BDMM- 3rd Project(6)

May 24, 2021

1 Big Data Modeling and Management 2021

1.1 BDMM Third Homework Assignment

The Wide World Importers (WWI) is a wholesales novelty goods importer and distributor operating from the San Francisco bay area. In this assignment we will be working with their database. You can get more information and details about the WWI database can be found in the following link: https://docs.microsoft.com/en-us/sql/samples/wide-world-importers-what-is?view=sql-server-ver15

The focus of the third assignment is modelling. We will use the same data source that was used the previous assignment, the World Wide Importers database, and convert it to a document-based database. To that end, we will be leveraging concepts like data denormalization, indexes, and mongodb design patterns.

More information on the extended datamodel to be found here:
https://docs.microsoft.com/en-us/sql/samples/wide-world-importers-oltp-database-catalog?view=sql-server-ver15

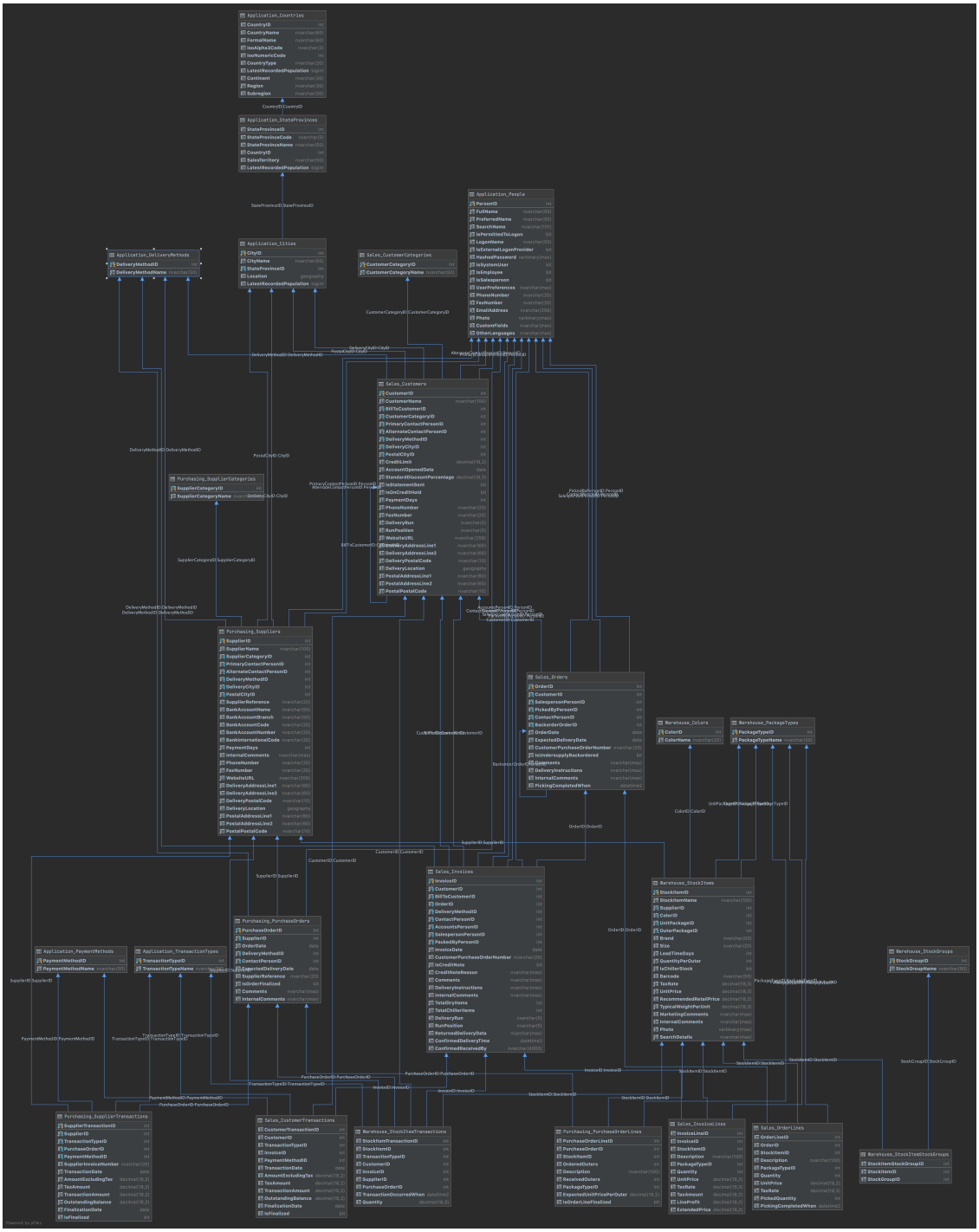
1.2 Problem Description

Your team has just arrived at WWI (a leading company in logitics). Welcome! Even though business is striving, the IT department is going through a bad time. Digitalization was never a priority for the company and now the company operational and analytical requirements is starting to grow beyond the capabilities of their existing data architecture.

WWI data is spread accross different systems. Namely, an old SQL database, data extracted through an API, and data stored in CSV files. Currently, the costs to develop the necessary queries to collect data to answer questions asked by the different departments are too high. Management concluded it is the right time to revise and revamp the data architecture, in order to speed up operations.

In that context, your team was tasked with merging all the company data into a single and coherent Mongo database. It is expected that, with your solution, WWI will have a better understanding of their business and that the different departments will be able to obtain efficiently the answers they desperatly need.

The WWI team shared with you an ERD of their current datamodel:



Addtionally, the WWI team asked you the deliver the following outputs in **10 days**: - Understand and model the database.
- Migrate all data to the database - Answer the questions.
- Submit the results by following the instructions.
- Prepare a short oral presentation to explain your design choices and the results you obtained.

With these deliveries, you will have created a prototype and allows the management to decide whether MongoDB is a good solution that meets their requirements.

1.2.1 Design Requirements

You have been informed that the WWI has the following query requirements to the database.

The web team needs:
- From which state province are our suppliers from?
- From which state province are the customers who have a higher credit limit?

The warehouse group needs:
- To know which items get ordered together the most?
- Which items get ordered the most in bulk (bigger amounts)?
- Which customers have delivery addresses under 10km of distance?

The CFO:
- Would like to know the monthly order count?
- Would like to know the average monthly sales prices?
- Would like to know the yearly expenditures with suppliers (per supplier name)?

Partnerships:
- Would like to know what's the most common payment type?
- Which supplier of `Novelty Goods Supplier` as the most transactions?

The marketing team:
- Want to make an appreciation post and needs the name of the sales person with the most invoices in 2013 (person who's customers brought the most money)?

Transform the SQL tables, API results and CSV files provided in the annex with this file and model a database following mongo's best practises.

Write MongoDB queries to awnser the above mentioned queries

Take advantage of database indexes to improve your query speeds

1.2.2 Deliverables

1. Notebook with all DB creation operations and CRUD operations;
2. Second notebook with all required queries to anwser the above metioned queries;

1.2.3 Data Source Materials

For the development of this assignment you will have access to the RDBMS/SQL database hosting the original WWI database. To connect to the database use the following credentials:

```
host:rhea.isegi.unl.pt
user:wwi-read-only-user
pass:jGp2GCqrss6nfTEu5ZawhW3mksLsQYQb
database:WWI

# !pip install mysql-connector-python
import mysql.connector
mydb = mysql.connector.connect(host={host}, user={user}, database={database}, port=3306, passw
mycursor = mydb.cursor()
```

```
mycursor.execute('SHOW TABLES;')
print(f"Tables: {mycursor.fetchall()}")
mycursor.execute('DESCRIBE Purchasing_PurchaseOrderLines;')
print(f"Purchasing_PurchaseOrderLines schema: {mycursor.fetchall()}")
```

Additionally you have access to the following documents.

CSV with Warehouse Data
https://liveeduisegiunl-my.sharepoint.com/:f:/g/personal/fpinheiro_novaims_unl_pt/Eh8Mj-m6r4dOt84tPDGUnhUBd5oMC0CJKAeyJm3urNB-8g?e=JuPMuW

API with Application data
http://rhea.isegi.unl.pt:8080/

1.3 Additional Information

Groups This is a group activity. Students should form groups of at least 4 and at most 5. We will use the current defined groups that have been established during the previous assignments, and that are identified on Moodle.

MongoDB database access Each group will have access to its own mongodb instance. Each group will receive an email with their access credentials. You will use the database to store your results.

Connection details will have the following template:

```
Host: rhea.isegi.unl.pt:27017
Username: {groups_username}
Password: {groups_password}
```

Which then can be used as follows:

```
client = MongoClient(f"{protocol}://{user}:{password}@{host}:{port}/")
```

Submission Deadline The submission must contain both notebooks and their results, also indicate the name of the database that you created. Upload the notebook on moodle before **23:59 of May 30th**

Evaluation The third homework assignment counts 20% towards your final mark of the curricular unit. The assignment will be scored from 0 to 20. Your final task will be to present the owner of the company your database proposal and how would it make everyone satisfied.

Each group submission will be evaluated on two components: 1. correctness of results; 2. simplicity of the solution;

50% - Database design
50% - Query results
* 25% - Correctness of queries
* 25% - Right results

Please note that all code delivered in this assignment will go through plagiarism automated checks. Groups high similarity levels in their code will undergo investigation.

Presentations

Presentations will be held between the 2nd and 3rd of June and you need to sign up your group in this calendly link: https://calendly.com/aestevao/presentations(Please try to avoid empty windows)