Lloyds Banking Group

Data Analyst Incubation

Sprint 4 : Big Query and Growth Opportunities

Contents

Background

As you saw in the previous sprint, Quay Bank is embarking on a cloud first collaboration strategy to minimise the risk of data siloisation and potential missed opportunities of information sharing. During this final Sprint you will focus on identifying growth opportunities for the Bank by focusing in on two Quay Bank business portfolios: Luxury loans and Standard loans.

You will soon discover that the recently acquired elite Luxury Loan portfolio offers potential opportunities for Quay Bank's future growth but also creates unique data management challenges for any analyst that works with it.

Email Received (high level requirements)

To: G.Pitt@quaybanking.com
CC: Data Team
Subject: Sharing lux loan data on g cloud

Hi Graham

Thanks for meeting with me today. As I outlined, there are a few steps needed to migrate data for the luxury loans portfolio (LLP) onto our google cloud platform (GCP). While our team absolutely wants to facilitate pragmatic access to this data in Big Query (BQ) for business intelligence and strategic reviews, it is important that we are also meeting our data governance responsibilities and regulatory guidelines.

I will outline my proposed approach to this request:

1) My team will identify and share the relevant data in the luxury loans portfolio. This will be extracted from our on-premises SQL server and migrated to GCP so that it can be analysed with BQ.

2) We will normalise the data into a simple star schema, because as I revealed it is currently stored in one large de-normalised table. My team will come up with a practical design – the table includes information about our customers, loan schedules and the assets. It's important that the target design should still be efficient to query – as you said we want to minimise GCP costs and we know querying can be expensive!

3) We know that a lot of the data we collect on luxury loans is personally identifiable and quite sensitive, so the team will ensure that data sharing is compliant with client privacy and data security regulations (GDPR & CCPA, InfoSec) across all our global data jurisdictions. The data team will work to

strike a balance between anonymisation and utility, so that the data is not ultimately rendered useless for further analysis. They also will ensure that only those colleagues who need access to the data on google cloud will have it.

4) To achieve data governance sign off, all transformational processes applied to LLP data (e.g. anonymisation, masking) and data security protocols (robust access management) will be documented and reviewed at the next board meeting.

5) Once the target data is up in the cloud, my team will also run practice queries in GCP to demonstrate how the data can be used to answer the types of questions we expect colleagues around the business to pursue.

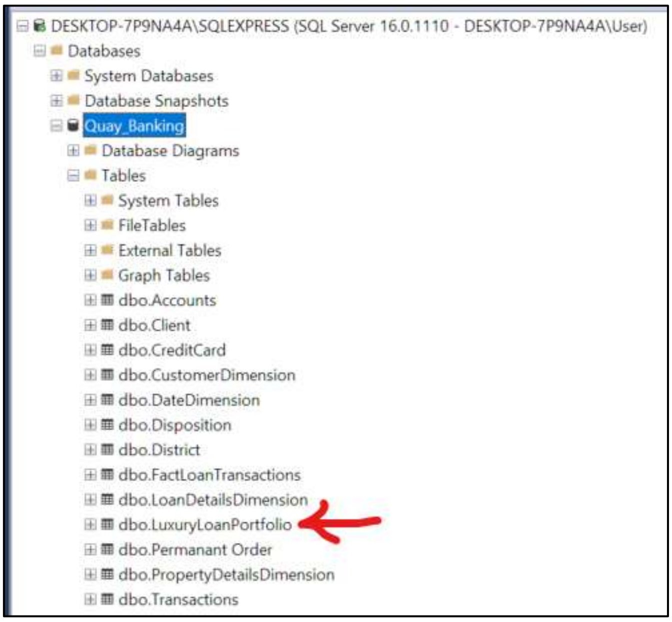
If you have any concerns about this approach, we can meet again to discuss.

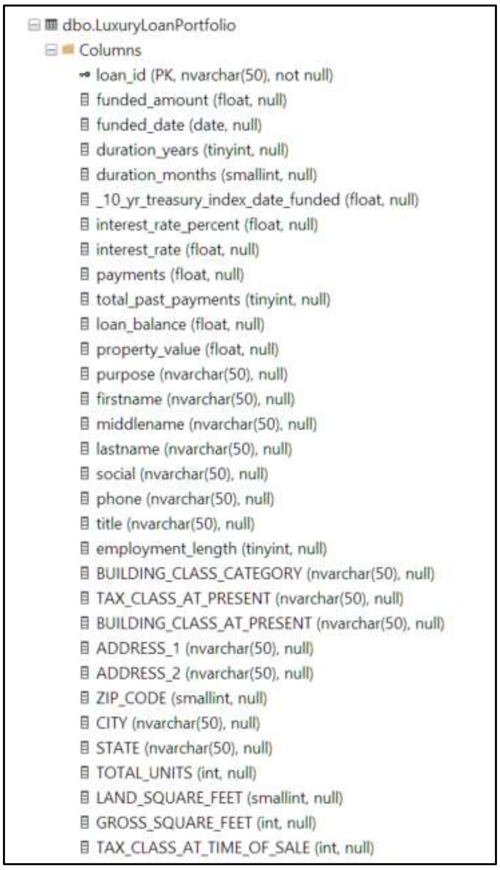
Best Regards,

Lauren Hill
Data Analyst Team Lead
Quay Banking Ltd

About the Data Source

A small sample of the data under review is stored in the Quay Bank SQL database, within a single table named [LuxuryLoanPortfolio]. This business was acquired independently by Quay Bank and thus the table has not been and cannot be integrated into the existing operating SQL schema.





```
dbo.LuxuryLoanPortfolio
  Columns
    loan_id (PK, nvarchar(50), not null)
    funded_amount (float, null)
    funded_date (date, null)
    duration_years (tinyint, null)
    duration_months (smallint, null)
    _10_yr_treasury_index_date_funded (float, null)
    interest_rate_percent (float, null)
    interest_rate (float, null)
    payments (float, null)
    total_past_payments (tinyint, null)
    loan_balance (float, null)
    property_value (float, null)
    purpose (nvarchar(50), null)
    firstname (nvarchar(50), null)
    middlename (nvarchar(50), null)
    lastname (nvarchar(50), null)
    social (nvarchar(50), null)
    phone (nvarchar(50), null)
    title (nvarchar(50), null)
    employment_length (tinyint, null)
    BUILDING_CLASS_CATEGORY (nvarchar(50), null)
    TAX_CLASS_AT_PRESENT (nvarchar(50), null)
    BUILDING_CLASS_AT_PRESENT (nvarchar(50), null)
    ADDRESS_1 (nvarchar(50), null)
    ADDRESS_2 (nvarchar(50), null)
    ZIP_CODE (smallint, null)
    CITY (nvarchar(50), null)
    STATE (nvarchar(50), null)
    TOTAL_UNITS (int, null)
    LAND_SQUARE_FEET (smallint, null)
    GROSS_SQUARE_FEET (int, null)
    TAX_CLASS_AT_TIME_OF_SALE (int, null)
```

Note that the columns shown above indicate the table contains what would commonly be considered personally identifiable and highly sensitive data. The nature of the customer relationships in this loan portfolio and the size of the loans taken demands a high degree of tactful treatment. This data is governed by enhanced data privacy and information security regulations, including GDPR and CCPA.

Assignment Detail

The following deliverables should be completed during this Sprint:

1. Target LLP database in GCP

A normalized database schema, built out from the LLP. This is expected to include 3 separate dimension tables containing relevant (i.e., useful for data analysis) attributes extracted from the LLP of:

- The Customer (sensitive information)
- The Loan
- The Asset (sensitive information)

A fact table should be constructed to contain payment information and the current loan balance. This is a current position snapshot of the sample which we should expect to update and expand in the future with an ETL/ pipeline procedure but that will occur in a later phase of this project. We can include today's date with the current balance position.

2. Data Handling (offline, during transfer or inside GCP)

For this data to be shared with other teams, any personally identifiable, confidential or individually sensitive data attributes that you are migrating from the LLP must be obscured. Practically this will ensure that the human being who took the loan and their asset cannot be identified during analysis.

This step can be completed either: locally, prior to the migration to GCP; during data migration (pipeline); or when the data has been uploaded to the cloud.

Possible techniques include:

- Generalization
- Aggregation
- Pseudonymization or Masking
- Anonymization
- Synthetic data
- Data perturbation
- Data swapping

Important! Ensure that the approach chosen does not excessively reduce data utility.

It is important that the chosen approach has been documented, including the tagging/identification of sensitive attributes, a suitable data flow diagram and a description of the data handling approach (but no examples of actual data), so that the process can be signed off by the Data Governance Board.

3. Deployment to GCP with access control

The transfer of LLP data from our local SQL servers to the GCP platform must follow a simple and secure methodology. This could include data ingestion via any of:

- Cloud Storage
- Pub/Sub
- DataFlow
- Cloud Data Fusion
- BQ DTS
- DataStream for BQ
- Data clean rooms in BQ

It is necessary to ensure that the LLP data inside GCP is secured by appropriate access controls. This typically includes leveraging Identity and Access Management (IAM) pre-defined roles and assigning users/ user groups for any relevant GCP services (Cloud Storage, CloudSQL and Big Query). It is recommended that test users and groups are created in the GCP project if none exist, to ensure IAM is working as expected.

4. Data Analysis Queries through GCP BQ

To demonstrate to Quay Bank colleagues that the LLP data is accessible via GCP BQ, identify at least 6 sample business questions that can be answered from the LLP data. Assemble the BQ syntax of those queries alongside a sample of the (anonymised) results, in a slide deck or pdf document for your colleagues at the Bank to learn from.

These questions could look like the following examples:

- How many (and what proportion of) unique customers have taken loans for the purposes of investing in commercial property, compared to the number of customers borrowing for residential and other assets?

- What are the average, minimum and maximum loan to asset value for this portfolio?

- What types or classification of professions are observed in the LLP and how do these relate to the types of assets funded?