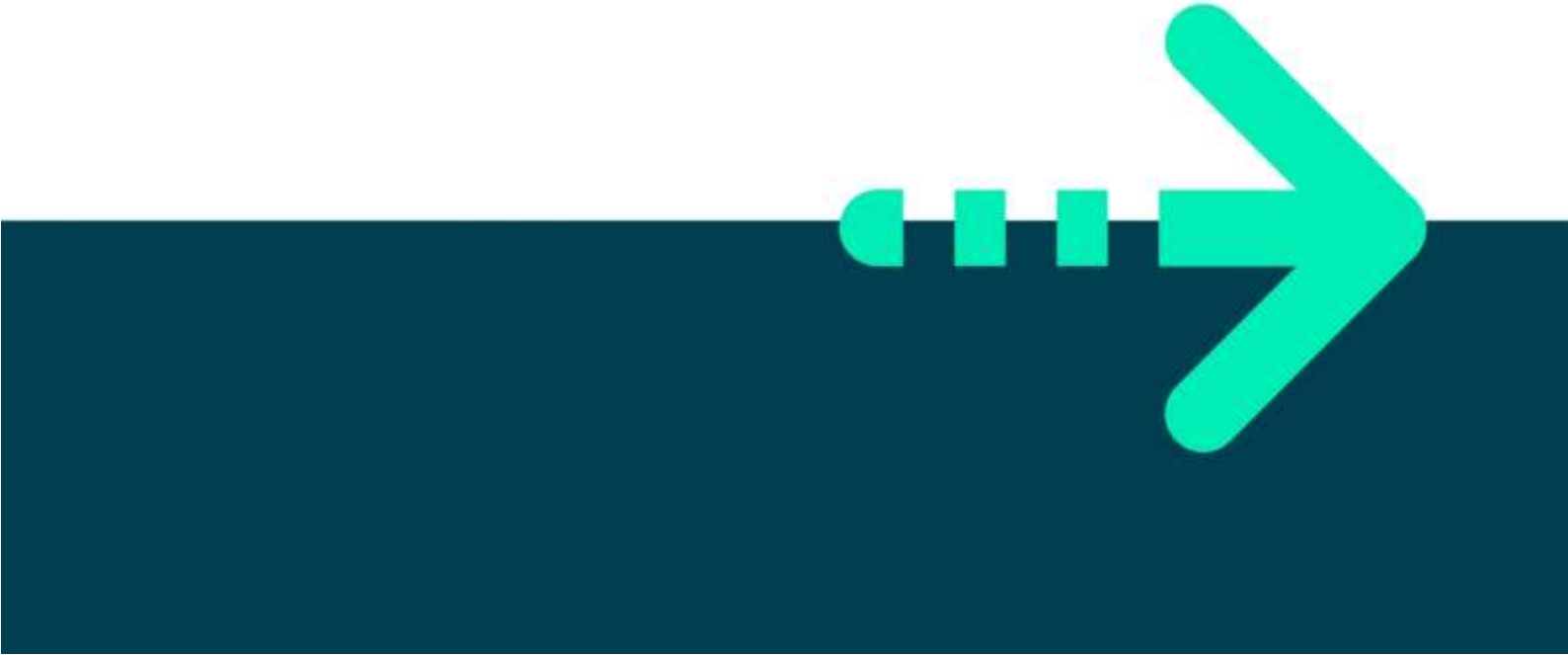
Lloyds Banking Group

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 LLP data into a simple star schema, because as I revealed when we met, this data is currently stored in one large de-normalised table. A denormalised design does not make it practical to mask or anonymise the personally identifiable data. My team will come up with a practical design – the table includes information about our customers, loan schedules and the assets so those are likely to be the 3 main entities, plus a 4th table which holds payments and balance transactions. It's important that the target normalised design should still be efficient to query – as you said we want to minimise GCP costs and we know querying in BQ can be expensive!

3) Because a lot of the data we collect on luxury loans is personally identifiable and quite sensitive, my team will ensure that our approach to data sharing is compliant with client privacy and data security regulations (GDPR & CCPA, InfoSec, ISO27001) across all our global data jurisdictions. The data team will find a balance between anonymisation and utility, so that the data is not ultimately rendered useless for further analysis. They also will ensure that only colleagues who need access to the data via google cloud will see it.

4) in line with Quay data governance, 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 shared may be used to answer the types of questions we expect colleagues around the business to pursue.

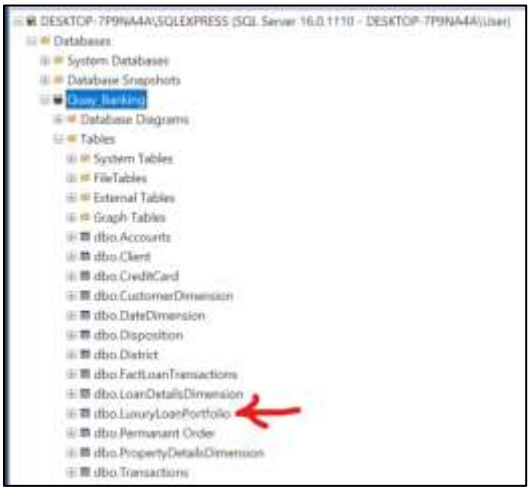
If you have any concerns about this approach, we can meet again to discuss.

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 the table has not been (and cannot be) integrated into the existing operational schema.



Note that the columns shown indicate the table contains what would commonly be considered personally identifiable and highly sensitive data. The nature of the customer relationships in the LLP demands a high degree of tactful treatment. This data is governed by enhanced data privacy and information security regulations.

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 at least 3 separate dimension tables containing relevant (i.e., useful for data analysis) attributes extracted from the LLP of:

- The Customer (strongly identifiable information)

- The Loan
- The Asset (potentially identifiable information)

A fact table should be constructed to contain payment information and the current loan balance. The latter is a current position snapshot for the sample, which we expect to update and expand in the future with an ETL/ pipeline procedure. (The live pipeline will be developed in a later phase of this project; the focus now is on the sample data set)

As this is a sample, a static date of 31/12/2019 can be attached to the current balance position to provide a timeline or metadata for the snapshot.

This step can be completed either:

Locally, prior to the migration to GCP

OR

During data migration (ETL/pipeline)

OR

Inside GCP before access to the data is granted

2. Sensitive Data Handling

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 to GCP must be obscured. Practically this will ensure that the human being who took out the loan and their asset cannot be identified by colleagues during analysis.

Possible techniques for sensitive data handling include:

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

This step can be completed either:

Locally, prior to the migration to GCP

OR

During data migration (ETL/pipeline)

OR

Inside GCP before access to the data is granted

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

It is important that the approach selected is documented, including the tagging/identification of sensitive attributes and a description of any data handling steps (but no examples of actual data), so that this procedure can be signed off by the Data Governance Board.

3. Migration of Data to GCP (ETL/ Pipeline)

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

This step must be completed either:

During data migration (ETL/pipeline)

It is important that the approach selected is documented, including a suitable data flow diagram (but no examples of actual data), so that this procedure can be signed off by the Data Governance Board.

4. GCP Access Control

It is necessary to ensure that the LLP data inside GCP is secured by appropriate access controls. This typically includes leveraging the 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 already exist, to ensure IAM access configuration is working as expected.

This step must be completed:

Inside GCP before access to the data is granted

5. Data Analysis Queries through GCP BQ

To demonstrate to Quay Bank colleagues that the LLP data is accessible via GCP BQ, identify at least 8 sample business questions that can be answered with it. Assemble the BQ syntax solving those queries, alongside a sample of the (anonymised, redacted) data results shown, in a slide deck or pdf document for your colleagues at the Bank to learn from.

Here are some examples of such questions:

- 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?