## GCP Case Study Day – Data Management & AI in Banking (with Dataplex)

### Overview

This one-day case study is the capstone activity for graduates who have completed GCP training. The focus is on applying concepts such as data lakes, data governance, analytics, and AI in a practical, banking-specific scenario using Google Cloud Platform tools including Dataplex, BigQuery, and Cloud Storage.

### Case Study Title

**“Building a Trusted Data Platform to Enable AI in Banking”**

### Business Context

You are part of the data team at a UK-based retail bank. Senior leadership has approved a new AI initiative to improve customer intelligence. However, the bank’s data is siloed, inconsistently tagged, and lacks a formal governance layer. Your job is to build a governed data platform on GCP to enable downstream AI and BI solutions while ensuring compliance with financial and data protection regulations.

## 🗓 8-Hour Agenda

| Time | Activity | Description |
| --- | --- | --- |
| **09:30–10:00** | **Kick-off & Case Brief** | Instructor explains the scenario, success criteria, expectations, and deliverables. |
| **10:00–11:00** | **Team Planning – Design the Data Platform** | Teams define data lake zones (raw/curated/analytics), governance needs, and propose an architecture using Dataplex. |
| **11:00–11:15** | **Break** |  |
| **11:15–12:30** | **Ingest & Register Data** | Teams load raw CSVs into Cloud Storage and ingest into BigQuery. Register assets and apply metadata using Dataplex. |
| **12:30–13:30** | **Lunch** |  |
| **13:30–14:30** | **Transform & Analyze Data** | Use SQL or BigQuery ML to clean, join, and analyze the data. Produce insights to support a banking AI use case. |
| **14:30–15:15** | **Governance & Risk Discussion** | Teams reflect on tagging strategy, IAM roles, data sensitivity, ethical risks, and regulatory considerations (e.g. GDPR). |
| **15:15–15:30** | **Break** |  |
| **15:30–16:30** | **Team Presentations** | Present architecture, governance strategy, insights, risks, and recommended actions. |
| **16:30–17:00** | **Wrap-Up & Reflection** | Instructor-led group debrief, feedback and Q&A. |

## 🔧 Tools & Services

* Google Cloud Storage
* BigQuery
* Dataplex (zones, catalogs, tags, governance)
* IAM (role-based access)
* Optional: Looker, BigQuery ML, Vertex AI

## 🌟 Learning Objectives

By the end of the day, learners should be able to:

* Organize raw and processed data in GCP using Dataplex
* Apply governance through metadata tagging and access policies
* Design and query a data pipeline that supports AI
* Reflect on data ethics and regulatory risks in financial services
* Communicate a data architecture proposal to business stakeholders

## 📁 Expected Team Deliverables

* Data platform architecture diagram
* List of Dataplex zones and tags
* Sample SQL or BigQuery ML queries
* Summary of key insights from analysis
* Risk/ethics discussion points (e.g. GDPR, explainability)
* Team presentation (5–10 slides)

## 🧹 Sample AI Use Cases – *Choose One per Team*

Each team should select **one** of the following use cases as the basis for their data platform and analysis work. They’ll design the pipeline, apply governance, and generate insights aligned to the chosen use case.

1. **Predicting Customer Churn**
   * Use customer profile and transaction history to flag high-risk churn segments.
   * Targeted retention insights.
2. **Detecting Fraudulent Transactions**
   * Use transaction history to identify anomalies.
   * Tag potentially fraudulent activity for escalation.
3. **Segmenting Customers for Credit Scoring**
   * Use income, account balance, repayment history, and credit score to create segments.
   * Support responsible lending and marketing strategy.
4. **Identifying At-Risk Accounts for Intervention**
   * Combine product usage and support logs to detect customers needing proactive outreach.
   * Often used in regulatory compliance (e.g., vulnerable customers).

### 💡 Why Choose Just One?

* Keeps the team focused on a **clear data question**
* Forces teams to **align design and governance** to a specific purpose
* Reflects real-world practice: data platforms are often built around use cases

## 🗂 Sample Dataset Fields (example CSVs)

* customer\_id
* account\_type
* monthly\_balance
* transactions\_last\_30\_days
* flagged\_fraud
* churned
* credit\_score

## 📜 Mock Client Brief

From: Head of Data Governance, Lloyds Banking Group  
Subject: Enabling AI with Trustworthy Data

“As part of our digital banking strategy, we want to empower our data scientists and analysts to build AI-powered insights. However, this can only happen if our data is well-organized, governed, and trusted. Your team has one day to propose a Dataplex-governed architecture that allows downstream AI solutions to be developed safely and effectively.”

## 🏆 Gold Standard Solution (Instructor Reference Only)

### Data Architecture

* Raw Zone (Dataplex): Contains ingested CSVs from Cloud Storage
* Curated Zone: Transformed BigQuery tables with joins and filters
* Analytics Zone: Final tables used for ML/BI consumption

### Tools Used

* **Cloud Storage**: Hosting source CSVs
* **BigQuery**: Query engine and transformation layer
* **Dataplex**: Cataloging, tagging, and organizing datasets
* **IAM**: Scoped access control per dataset and role
* **BigQuery ML**: For churn prediction (logistic regression model)

### Key Transformations

* Join customers with transactions and support logs
* Calculate aggregated features: total transactions, average spend, complaint count
* Filter to customers active in last 30 days

### Sample Query (Churn Analysis)

SELECT   
 c.customer\_id,  
 c.age,  
 c.credit\_score,  
 c.monthly\_balance,  
 COUNT(t.transaction\_id) AS transaction\_count,  
 AVG(t.amount) AS avg\_transaction\_amount,  
 COUNT(DISTINCT s.log\_id) AS support\_issues,  
 c.is\_churned  
FROM `your\_project.your\_dataset.customers` c  
LEFT JOIN `your\_project.your\_dataset.transactions` t ON c.customer\_id = t.customer\_id  
LEFT JOIN `your\_project.your\_dataset.support\_logs` s ON c.customer\_id = s.customer\_id  
GROUP BY c.customer\_id, c.age, c.credit\_score, c.monthly\_balance, c.is\_churned;

### Governance Features

* Tags: PII, Financial, Transactional, Support
* Access control:
  + Data Scientists: Read access to curated + analytics
  + Governance Team: Full access to tags + metadata
* Zones mapped and enforced via Dataplex policies

### Ethics / Compliance Considerations

* PII tagged with customer\_id
* AI model explainability (BigQuery ML coefficients reviewed)
* GDPR reviewed: only non-sensitive fields shared to analytics zone

### Presentation Outline (Gold Standard)

* Slide 1: Team name and use case
* Slide 2: Architecture diagram
* Slide 3: Dataplex setup (zones, tags, IAM)
* Slide 4: Query outputs and analysis summary
* Slide 5: AI readiness or model demo (if included)
* Slide 6: Risk and compliance reflection

## ✅ Instructor Support Tips

* Monitor team breakout rooms, encourage architectural thinking
* Assist with IAM troubleshooting or BigQuery access issues
* Keep teams on time with check-ins after each phase
* Debrief by tying outcomes back to real-world GCP practices

## 📥 Optional: Add-ons to Prepare in Advance

* Pre-created GCP projects with APIs enabled
* Public GCS bucket with datasets
* Presentation template (Google Slides or PowerPoint)
* Worksheet handouts (optional paper-based design phase)