

RAG BOT for Skipton Building Society

WHAT is it?

WHY now?

HOW is it done?

ATTRIBUTES

LIMITATIONS

WHAT IS RAG?

RAG - Retrieval Augmented Generation



Group results for the year ended 31 December

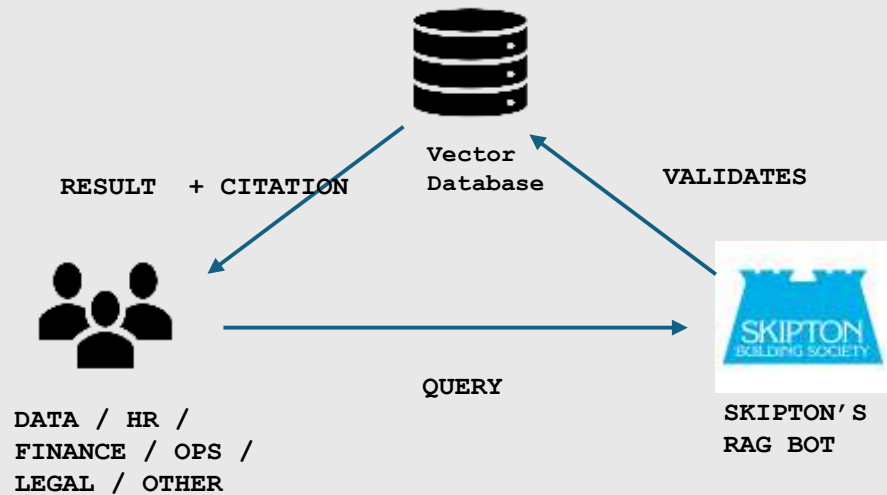
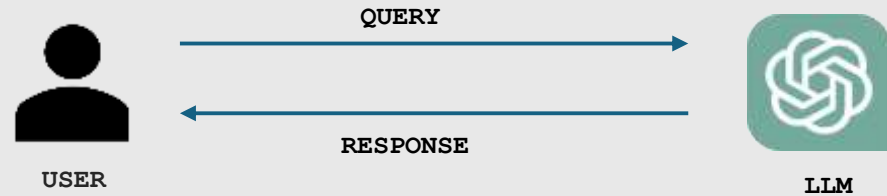
	2023 £m	2022 £m
Net interest receivable	542.8	424.4
Other income and charges	986.3	1,081.8
Fair value gains	23.5	1.3
Profit on disposal of subsidiary undertakings	-	0.1
Administrative expenses and provisions	(1,223.6)	(1,190.1)
Impairment credits / (losses)	2.4	(18.7)
Profit for the year before taxation	333.4	298.8
Taxation	(78.8)	(67.8)
Non-controlling interests	0.2	0.2
Profit for the financial year attributable to members	254.8	231.2

Image Courtesy - Skipton Group- Financial Statement '23

WHY SHOULD WE USE IT?



An Overview on HOW to build



Attributes

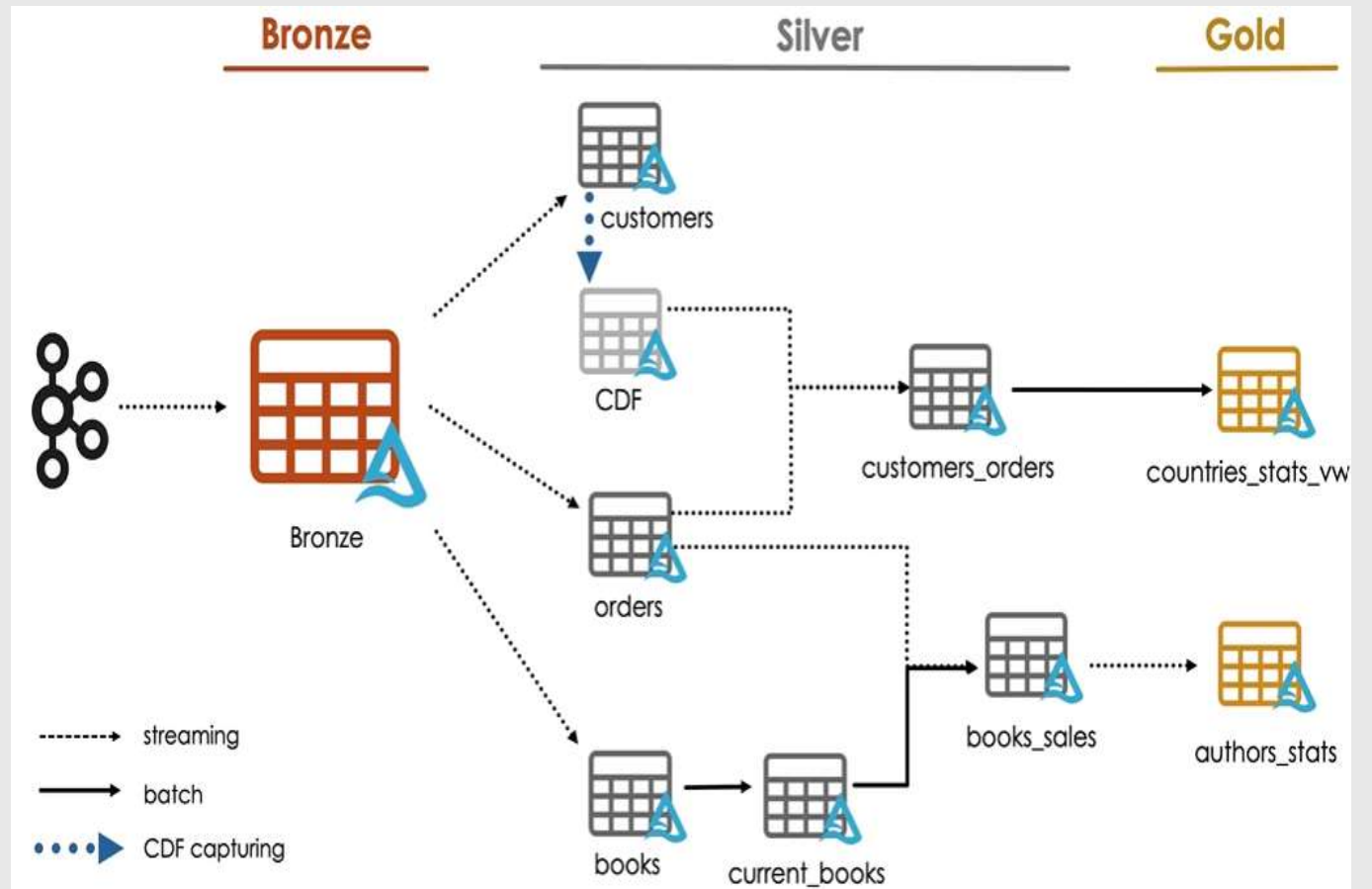
Scalability
Maintainability
Interoperability

Limitations

Reliability
Maintainability
Governance Layer?

DATABRICKS PROJECT USING KAFKA EVENTS

1. Data Modelling
2. Data Processing
3. Performance Improvements
4. Databricks jobs
5. Security and Governance
6. Testing & Deployment
7. Monitoring & Logging



Eg: Executor
Memory =
10GB

Spark Memory Management

Spark.executor.memory

Spark.memory.
fraction
= 0.6

Spark.memory.
storage.fraction = 0.5

For simplicity
1GB = 1000MB
4GB - 300MB = 3.7GB

Spark Executor Container

On Heap Memory

spark.executor.memory

Execution Memory

Joins, Shuffles,
Sorting, GroupBy

6GB

Storage Memory

RDD, Dataframe
Caching, Broadcast
Variables

3GB

User Memory

User Defined DS:
Variables, Objects

4GB

Reserved Memory

300 MB

Off Heap Memory

spark.memory.offHeap.enabled
spark.memory.offHeap.size

Spark.memory.offHeap.size=0

Overhead

spark.executor.memoryOverhead
max(384MB
OR
10%
spark.executor.memory)

Spark.executor.memoryOverhead

Data Modelling

Multiplex Bronze
Table: Many-to-One
Mapping.

Streaming from
multiplex Bronze

Quality Enforcement

Streaming
Deduplication

SCD- TYPE 2

Data Processing

Change Data Capture

Processing CDF Feed

Stream-
Stream/Static Joins

Materialised Gold
Tables

Performance Improvement

Partitioning DLT

Auto Optimize

SECURITY

Dynamic Views
Propagating
Deletes

TESTING

Relative
Imports
Data Pipeline
Testing

MONITORING

Managing
Clusters

FUTURE MODEL

- Write event messages to az adls gen2 > delta using `'databricks_copy_into macro'`
- Run Dbt productions jobs in databricks workflows.
- Run `dbt_project_evaluator` to ensure the project meets best practices, and add `SQLFluff` to standardize SQL in dbt projects



- Cost and Time efficient
- Better Reliability and increased Data Observability