Internship
Final Presentation

ABACUS digital
29 JULY 2025

ABACUS digital

INTRO

PARINTARA (KAT)
WONGSRISOONTORN

DATA ENGINEER INTERN

Mentor: P' Tae

New York University (NYU)

BS Integrated Design and Media
MS Computer Science

ABACUS digital

A G E N D A

Project Setup & Planning

Building the Solution

Outcomes & Reflections

ABACUS digital

OVERVIEW

Context

Internal metadata is often incomplete or unclear, making it difficult for teams to understand unfamiliar tables and queries

Objective

Build scalable, maintainable tools to automate metadata generation for BigQuery tables and queries using LLMs, schema information, and internal documentation

ABACUS digital

OVERVIEW

Tech Stack

LangChain · OpenAI · BigQuery · Confluence API · Python

Scope

Generate metadata for data-platform-prod.thunder.contract
→ Column & table descriptions

Output

Structured JSON descriptions for column and table metadata

ABACUS digital

A G E N D A

Project Setup & Planning

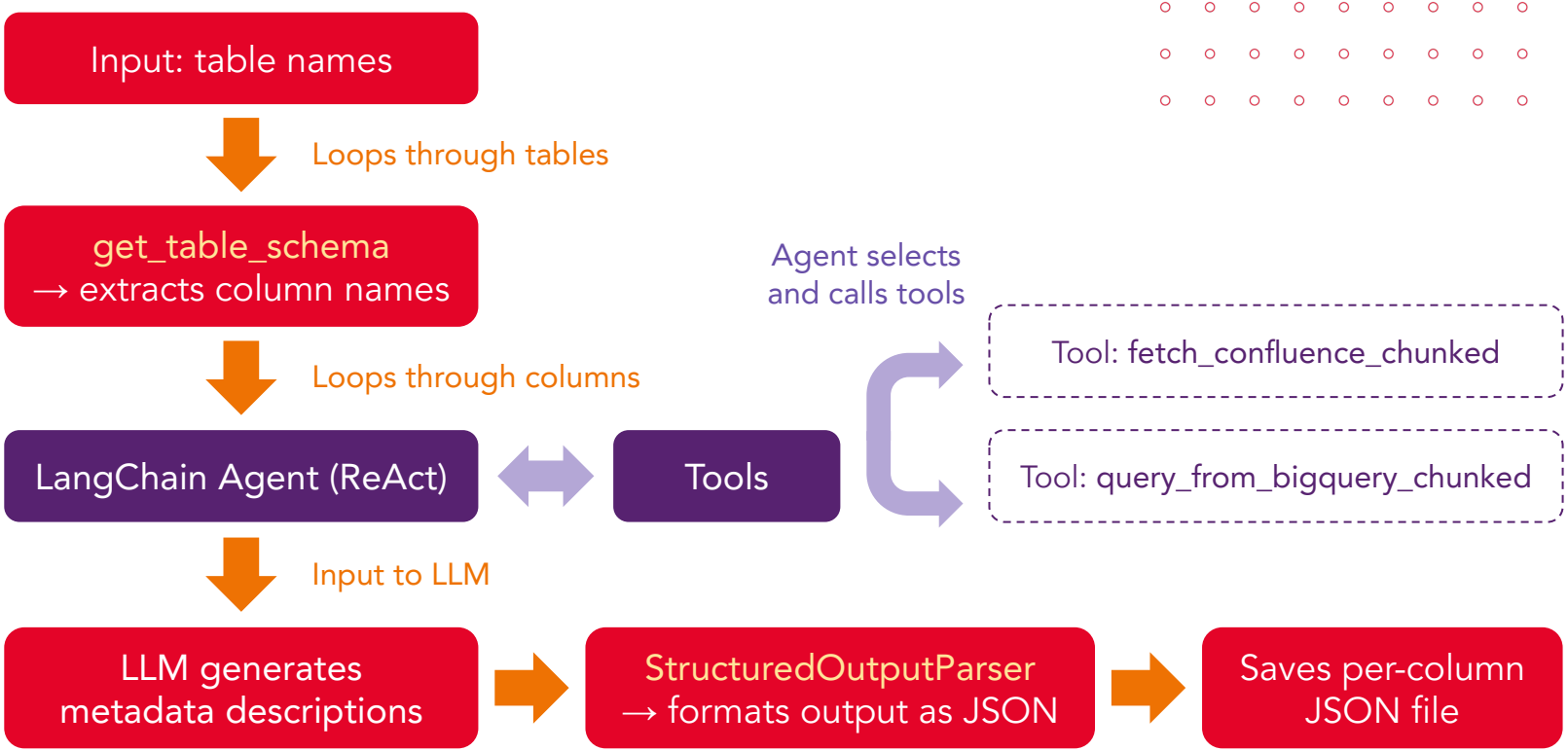
Building the Solution

Outcomes & Reflections

ABACUS digital

[Column Metadata]

W O R K F L O W



ABACUS digital

PROMPT

You are a metadata assistant that generates useful descriptions for database columns

Your task is to generate a **{meta_key}** for the column **{column_name}** in the BigQuery table **{table}**

ABACUS digital

PROMPT

Follow this process:

1. Use the `query_from_bigquery_chunked` tool to examine real data from the table

Only proceed to call the `fetch_confluence_chunked` tool to retrieve internal documentation if the query results are missing, unclear, or insufficient to complete your task

ABACUS digital

PROMPT

2. Based on the information you have gathered,
write a meaningful and informative **{meta_key}**
for the column

3. Finally, return your result in the **exact JSON format** below
**Important: Output only the JSON object — no additional text,
no explanations, no markdown formatting**

{format_instructions}

ABACUS digital

[Agent Tool Execution]

W O R K F L O W

LangChain Agent (ReAct)

Tool: query_from_bigquery_chunked

sql_query_checker
→ validates SQL

query_from_bigquery

chunk_and_summarize_text

Tool: fetch_confluence_chunked

get_confluence
→ retrieves Confluence content

chunk_and_summarize_text

ABACUS digital

[LangChain Agent]

D E C I S I O N

```
> Entering new AgentExecutor chain...
I need to gather data from the BigQuery table `data-platform-prod.thunder.contract` to understand the `contract_id` column. I will use the `query_from_bigquery_chunked` tool
 from this table.

Action: query_from_bigquery_chunked
Action Input: "SELECT contract_id FROM `data-platform-prod.thunder.contract` LIMIT 100"
```

```
The **contract_id** column contains unique numerical identifiers assigned to contracts. These IDs are primarily large integers, typically starting with "100" followed by six digits, though som
smaller numeric values also appear. Each contract_id serves as a distinct reference key for individual contracts within the dataset.I have gathered sufficient information about the `contract_i
 column from the query results. Now, I will formulate a concise summary and a detailed explanation for the `long_description`.

```json
{
        "short_description": "The contract_id is a unique identifier for each contract in the dataset, primarily represented as large integers.",
        "long_description": "The contract_id column in the data-platform-prod.thunder.contract table serves as a unique identifier for each contract recorded in the dataset. This identifier is
rucial for distinguishing between different contracts and ensuring data integrity. The values in this column are primarily large integers, typically formatted to start with '100' followed by s
 additional digits, although there are instances of smaller numeric values. Each contract_id is assigned at the time of contract creation and remains constant throughout the contract's lifecyc
. This consistency allows for reliable tracking and referencing of contracts across various systems and reports. The uniqueness of the contract_id ensures that no two contracts can share the s
e identifier, which is essential for maintaining accurate records and facilitating efficient data retrieval and analysis."
}
```
```

Thought: The task is to generate a long description for the column contract_id in the specified BigQuery table. The provided information is already sufficient to create the required JSON outpu
Final Answer: {
        "short_description": "The contract_id is a unique identifier for each contract in the dataset, primarily represented as large integers.",
        "long_description": "The contract_id column in the data-platform-prod.thunder.contract table serves as a unique identifier for each contract recorded in the dataset. This identifier is
rucial for distinguishing between different contracts and ensuring data integrity. The values in this column are primarily large integers, typically formatted to start with '100' followed by s
 additional digits, although there are instances of smaller numeric values. Each contract_id is assigned at the time of contract creation and remains constant throughout the contract's lifecyc
. This consistency allows for reliable tracking and referencing of contracts across various systems and reports. The uniqueness of the contract_id ensures that no two contracts can share the s
e identifier, which is essential for maintaining accurate records and facilitating efficient data retrieval and analysis."
}
```

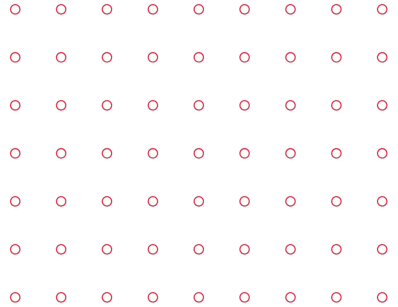
> Finished chain.
```

Sample Execution of Agent Decision Making and Tool Selection

ABACUS digital

[Table Metadata]

W O R K F L O W



Input: table names

Loops through tables

LangChain Agent (ReAct)

get_column_descriptions
→ reads column metadata

Loops through columns

get_table_schema
→ extracts column names

Looks up column descriptions

LLM generates metadata descriptions

StructuredOutputParser
→ formats output as JSON

Saves table-level JSON file

ABACUS digital

A G E N D A

Project Setup & Planning

Building the Solution

Outcomes & Reflections

ABACUS digital

[Column Metadata]

O U T P U T

Sample JSON output for one of 249 columns
in the data-platform-prod.thunder.contract table

ABACUS digital

[Table Metadata]

O U T P U T

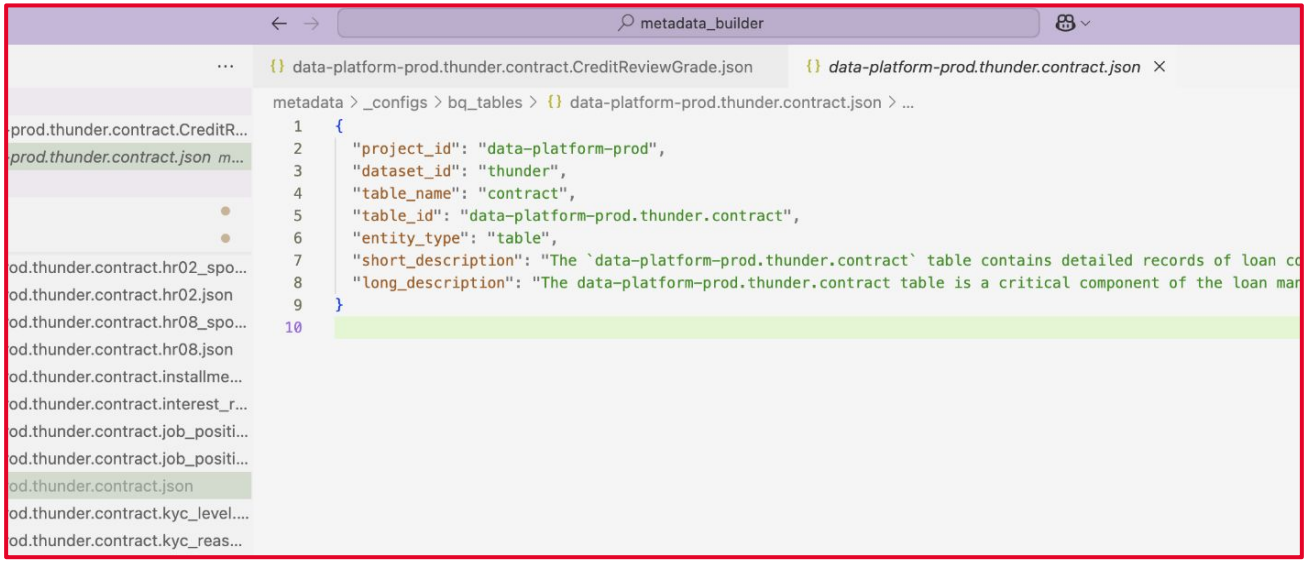


Table-level JSON output for the
data-platform-prod.thunder.contract table

ABACUS digital

CHALLENGES

Token Limit

Issue: Large Confluence pages and queries exceeded model limits

Solution: Implemented chunk_and_summarize_text to reduce content size before sending to LLM

SQL Format Error

Issue: Agent sometimes generated invalid queries

Solution: Wrapped SQL tool in try-except and looped until a valid response was returned

ABACUS digital

CHALLENGES

Parsing Failure

Issue: LLM output did not always match the expected JSON format

Solution: Refined the prompt and used StructuredOutputParser with fallback handling

Rate Limiting

Issue: Too many requests triggered API limits

Solution: Paused after every 50 columns for 60 seconds to avoid exceeding limits

ABACUS digital

T A K E A W A Y S

Gained exposure to the data engineering industry and practical tools used in the field

Discovered how data, AI, and engineering intersect in real-world and future innovation

Learned about fintech and how technology supports money lending and digital finance

Enjoyed working on hands-on projects that go beyond academic concepts

Inspired by the smart, collaborative people and the open culture of a modern company

ABACUS digital

Internal Tools Shape Company Culture ”

ABACUS digital

Special Thanks to the Data Engineering (DE) Team

NATTH BEJRABURNIN
Chief Data & AI Officer

VARUNYA THAVORNUN
Head of Data Engineering

WISAROOT LERTTHAWEEDECH
Senior Data Engineer

THAYANEE RUENNARK
Data Engineer

WARIS PUNNAHITANOND
Data Engineer

PONTHAI KLINSAMPAS
Senior Data Engineer

ABACUS digital

Q & A

ABACUS digital

ABACUS digital

THANK YOU

ABACUS digital