KPMG

COLUMBIA UNIVERSITY
DATA SCIENCE INSTITUTE

Capstone Project Spring 2024
AI Model Transparency
Final Presentation

David Huang, Numan Khan,
Phillip Kim, Jerry Wang, Tom Yu

Mentors: Ben Carper & Jim Leach

Data Science Institute,
Columbia University

May 1, 2024

Presentation Agenda

- Team Introduction
- Project Motivation
- Project Implementation
  - Why RAG Chatbot?
  - RAG Chatbot Components
  - Synthetic Data Generation
  - Cypher Query Construction
  - Orchestrator Flow
  - Prompt Template
  - Testing and Improvements
  - Live Product Demonstration
- Conclusion & Future Plans
- Reflections

Team Introduction

KPMG

COLUMBIA UNIVERSITY
Data Science Institute

- MS Data Science (MSDS) @ Columbia Engineering
- Capstone is a culmination of skills and knowledge gained, and the final step where MSDS students work on a project sponsored by a DSI industry affiliate.

Phillip Kim
Part-time MSDS
Data Scientist
@ FDIC

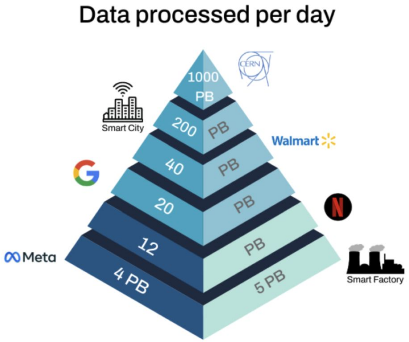
David Huang
Full-time MSDS
December 2024
Graduation

Numan Khan
Part-time MSDS
Software Engineer
@ Amazon

Jerry Wang
Part-time MSDS
Data Analyst
@ EssilorLuxottica

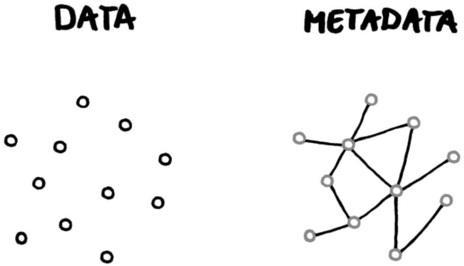
Tom Yu
Full-time MSDS
May 2024
Graduation

Project Motivation



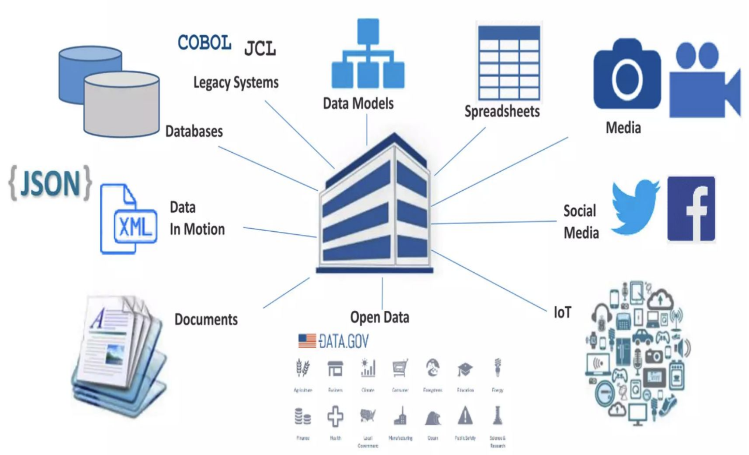
Each day, the world generates roughly 1,000 petabytes = 1mm terabytes of data

Companies employ data analytics and models to generate reports and forecasts

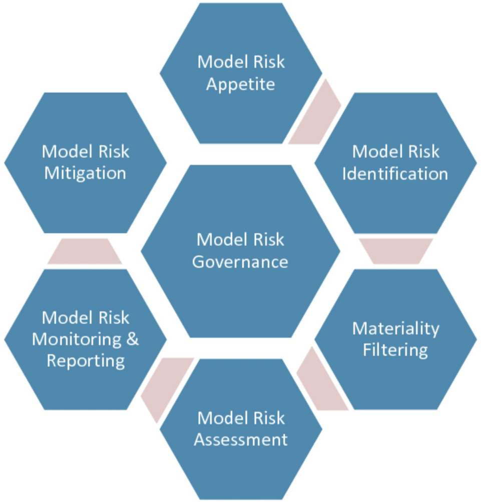


Managing Metadata is Key to AI Transparency!

Project Motivation

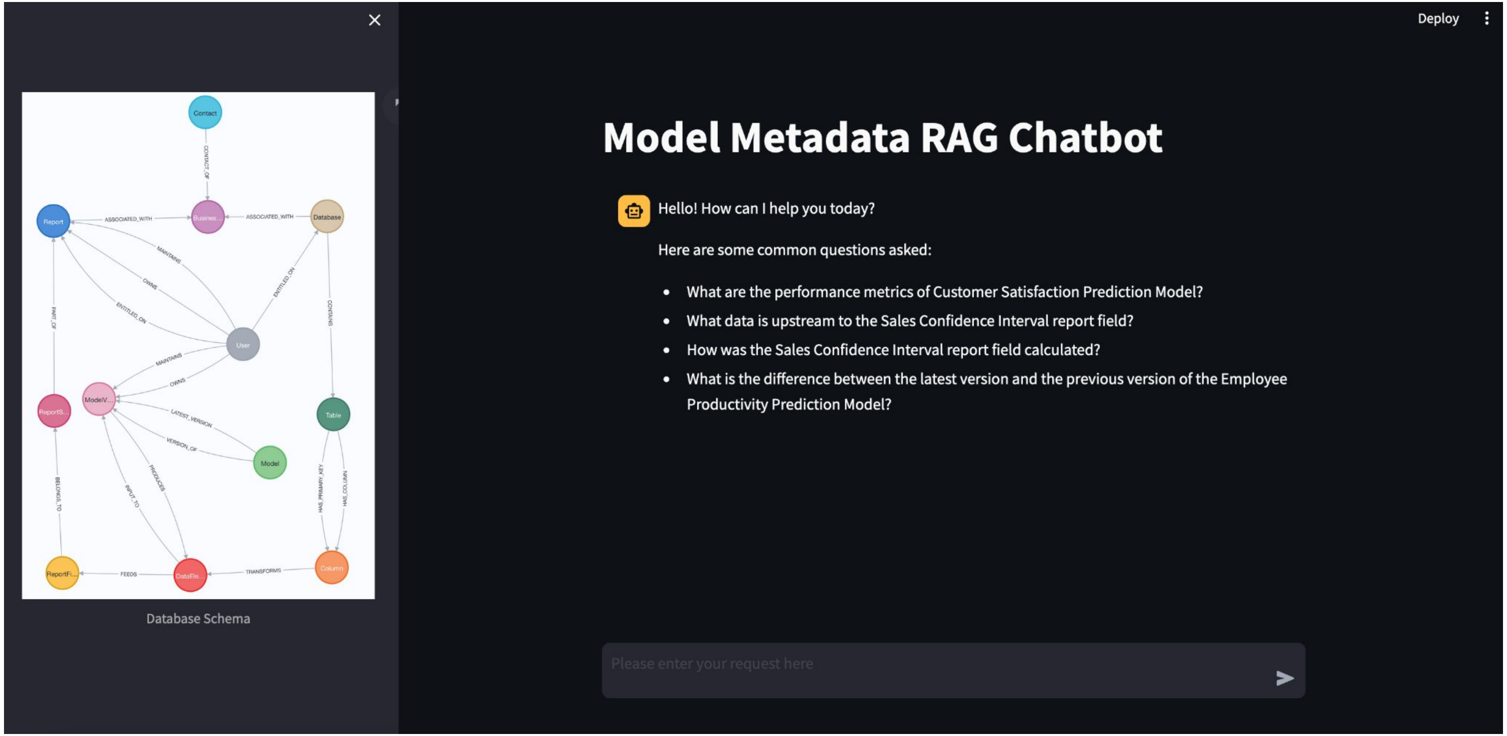


Metadata exists in many sources across and beyond an organization



Our Metadata RAG Chatbot can be an integral component of a holistic Model Risk Management Program!

Streamlit User Interface



Database Schema

Model Metadata RAG Chatbot

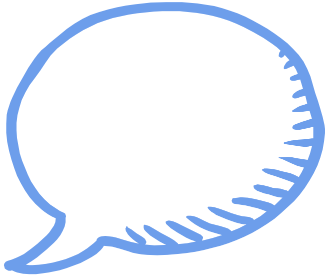
Hello! How can I help you today?

Here are some common questions asked:

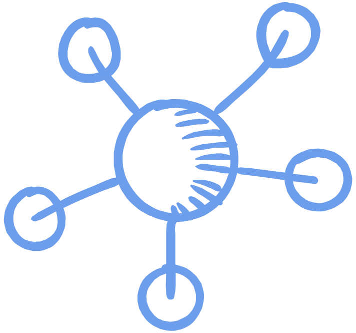
- What are the performance metrics of Customer Satisfaction Prediction Model?
- What data is upstream to the Sales Confidence Interval report field?
- How was the Sales Confidence Interval report field calculated?
- What is the difference between the latest version and the previous version of the Employee Productivity Prediction Model?

Please enter your request here

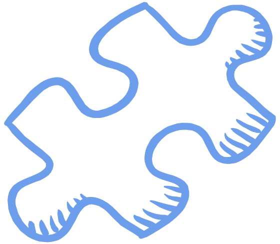
Why RAG Chatbot?



no need to learn query language

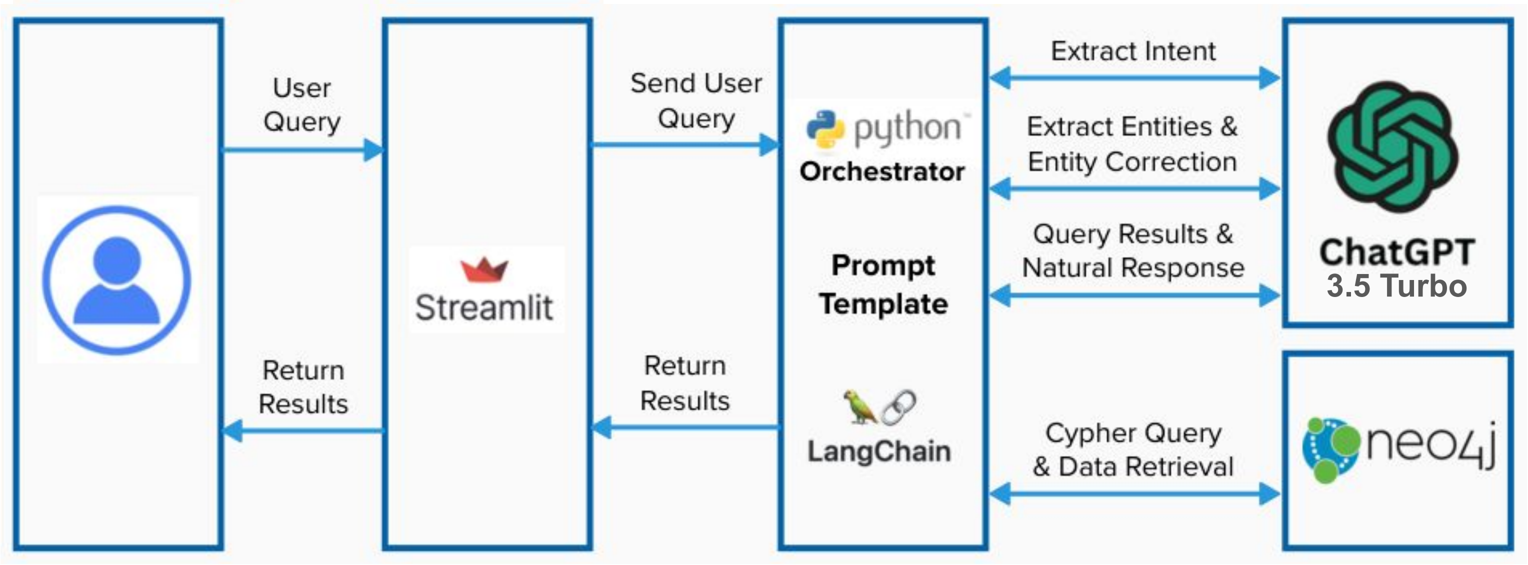


store data as a network

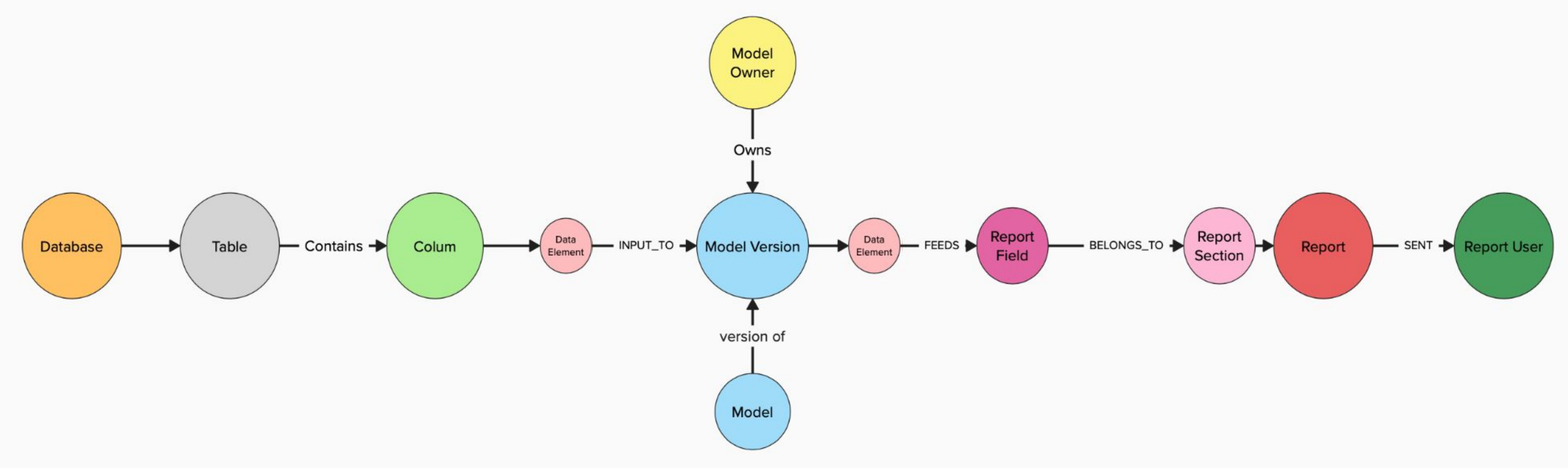


high level of flexibility

RAG Chatbot Components

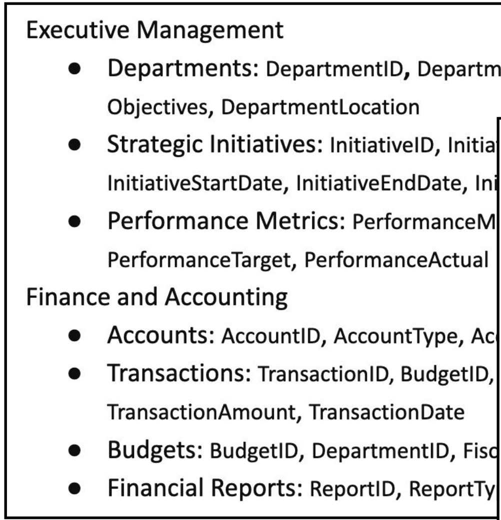


Synthetic Data Generation



Initial Data Schema Design

Synthetic Data Generation

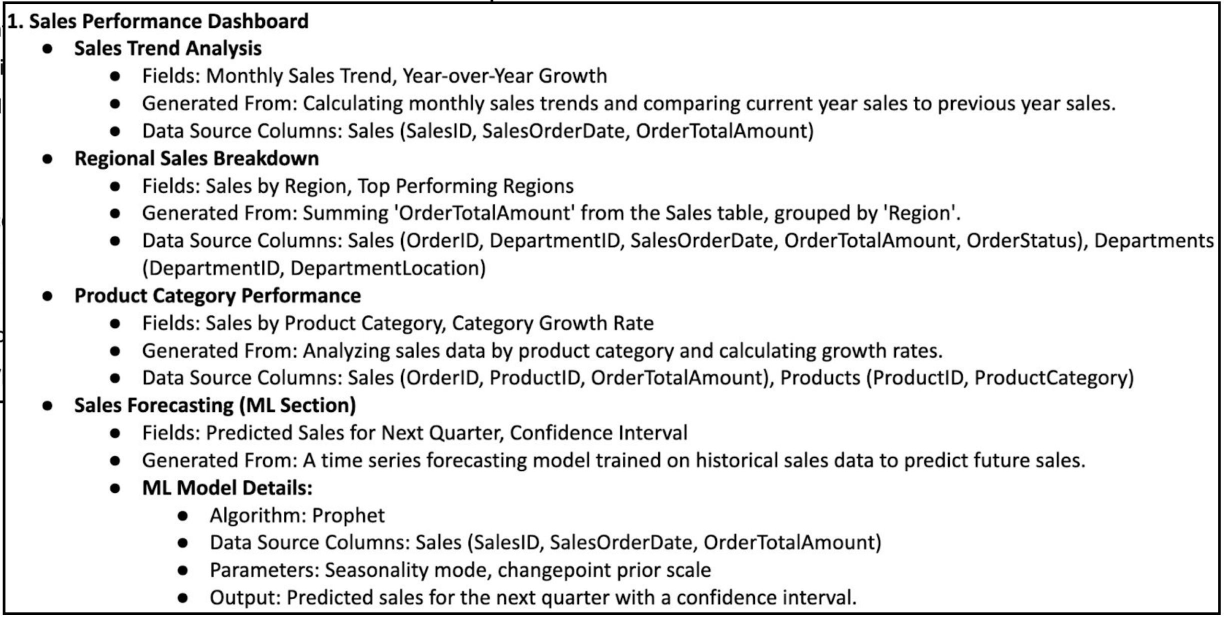


Executive Management

- Departments: DepartmentID, DepartmentName, ManagerID, DepartmentBudget, Objectives, DepartmentLocation
- Strategic Initiatives: InitiativeID, Initia
  InitiativeStartDate, InitiativeEndDate, Ini
- Performance Metrics: PerformanceM
  PerformanceTarget, PerformanceActual

Finance and Accounting

- Accounts: AccountID, AccountType, Ac
- Transactions: TransactionID, BudgetID,
  TransactionAmount, TransactionDate
- Budgets: BudgetID, DepartmentID, Fisc
- Financial Reports: ReportID, ReportTy

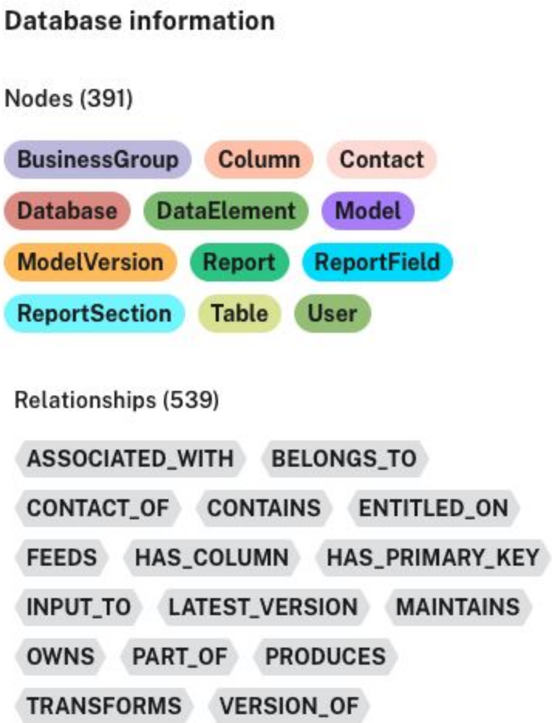


1. Sales Performance Dashboard

- **Sales Trend Analysis**
  - Fields: Monthly Sales Trend, Year-over-Year Growth
  - Generated From: Calculating monthly sales trends and comparing current year sales to previous year sales.
  - Data Source Columns: Sales (SalesID, SalesOrderDate, OrderTotalAmount)
- **Regional Sales Breakdown**
  - Fields: Sales by Region, Top Performing Regions
  - Generated From: Summing 'OrderTotalAmount' from the Sales table, grouped by 'Region'.
  - Data Source Columns: Sales (OrderID, DepartmentID, SalesOrderDate, OrderTotalAmount, OrderStatus), Departments (DepartmentID, DepartmentLocation)
- **Product Category Performance**
  - Fields: Sales by Product Category, Category Growth Rate
  - Generated From: Analyzing sales data by product category and calculating growth rates.
  - Data Source Columns: Sales (OrderID, ProductID, OrderTotalAmount), Products (ProductID, ProductCategory)
- **Sales Forecasting (ML Section)**
  - Fields: Predicted Sales for Next Quarter, Confidence Interval
  - Generated From: A time series forecasting model trained on historical sales data to predict future sales.
  - **ML Model Details:**
    - Algorithm: Prophet
    - Data Source Columns: Sales (SalesID, SalesOrderDate, OrderTotalAmount)
    - Parameters: Seasonality mode, changepoint prior scale
    - Output: Predicted sales for the next quarter with a confidence interval.

ChatGPT 4 Generated Database Tables & Reports

Synthetic Data Generation



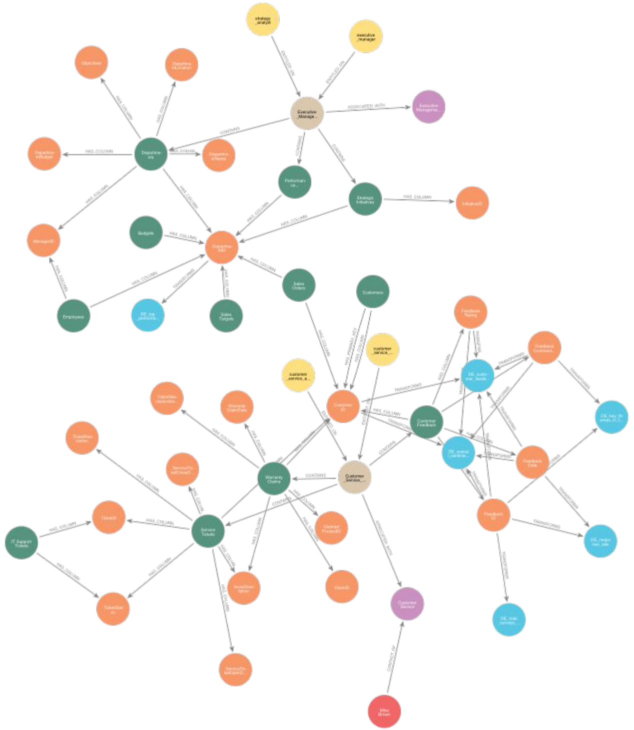
Database information

Nodes (391)

BusinessGroup Column Contact Database DataElement Model ModelVersion Report ReportField ReportSection Table User

Relationships (539)

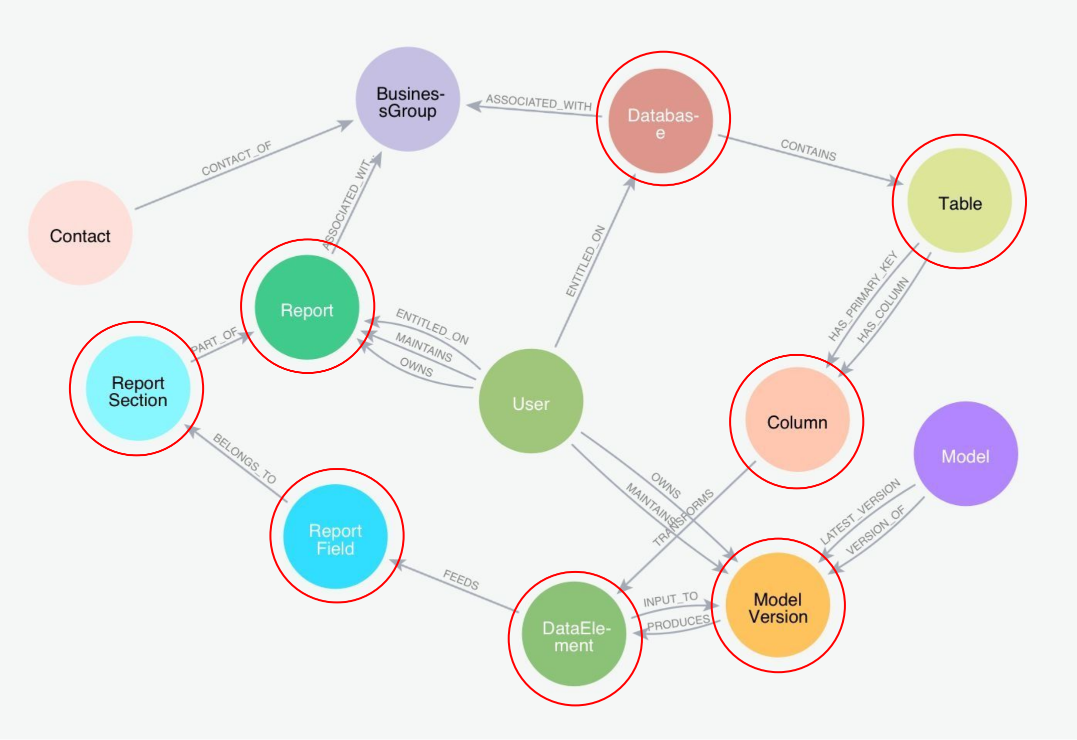
ASSOCIATED_WITH BELONGS_TO CONTACT_OF CONTAINS ENTITLED_ON FEEDS HAS_COLUMN HAS_PRIMARY_KEY INPUT_TO LATEST_VERSION MAINTAINS OWNS PART_OF PRODUCES TRANSFORMS VERSION_OF



Synthetic Data Generation

KPMG

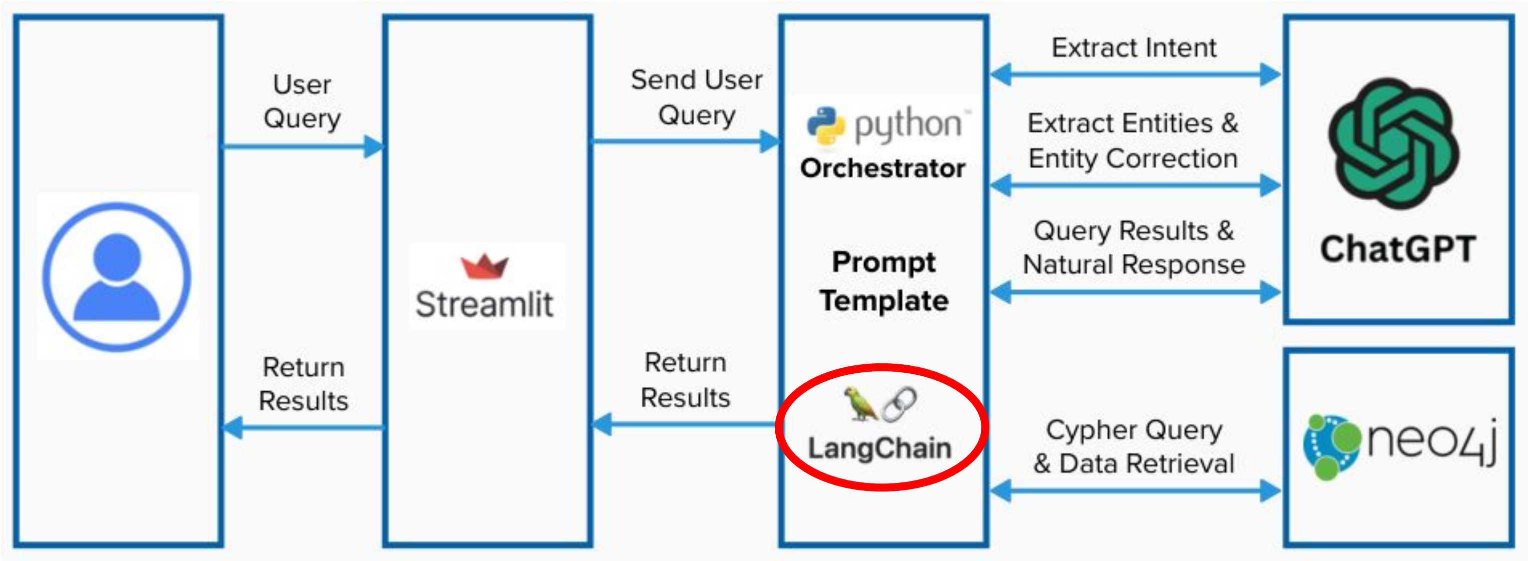
COLUMBIA UNIVERSITY
Data Science Institute



Graph DB Schema in Neo4j

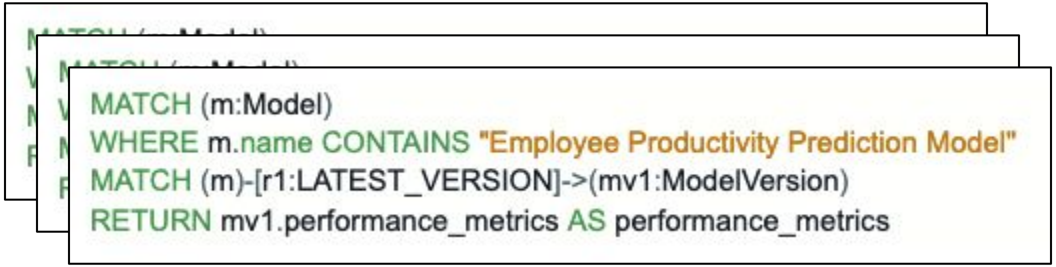
Cypher Query Construction

- Langchain helps incorporate LLMs into an application



Cypher Query Construction

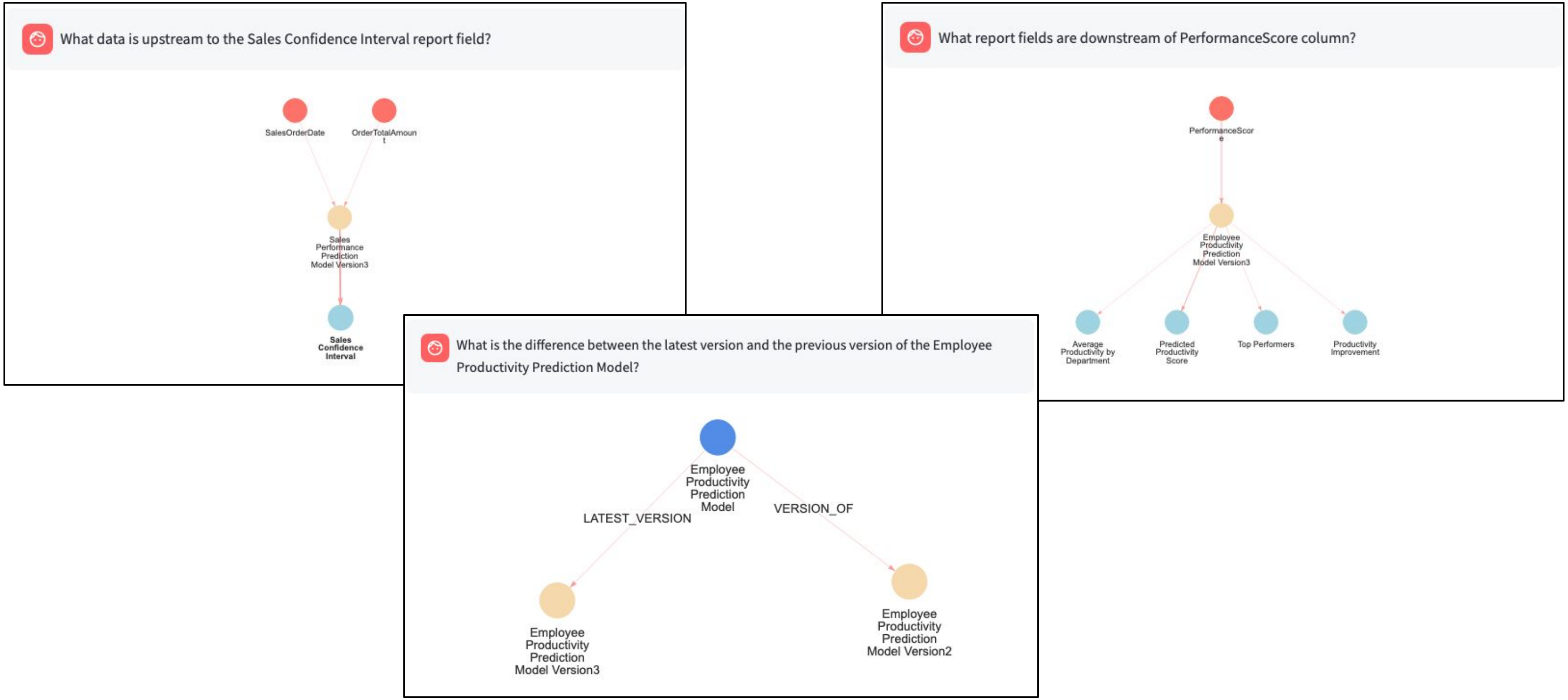
Common Questions	Uncommon Questions	Other Questions
• How does the number in the reportfield come from?	• Which users have access to the IT_Database?	• What is the fastest land animal?



```
MATCH (m:Model)
WHERE m.name CONTAINS "Employee Productivity Prediction Model"
MATCH (m)-[r1:LATEST_VERSION]->(mv1:ModelVersion)
RETURN mv1.performance_metrics AS performance_metrics
```

Cypher Query Templates

Streamlit Integration (Agraph)



What data is upstream to the Sales Confidence Interval report field?

SalesOrderDate
OrderTotalAmount
Sales Performance Prediction Model Version3
Sales Confidence Interval

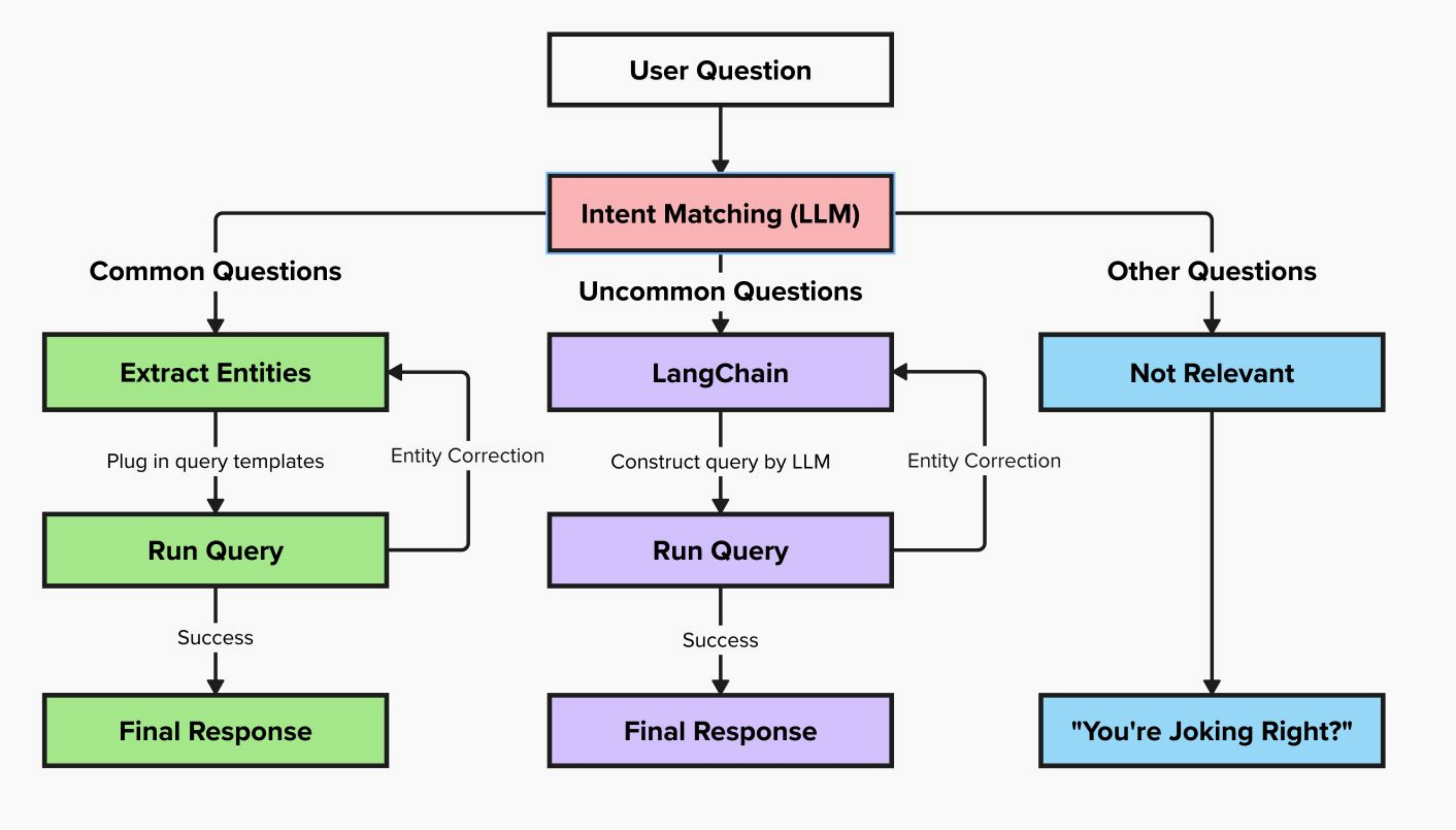
What report fields are downstream of PerformanceScore column?

PerformanceScore
Employee Productivity Prediction Model Version3
Average Productivity by Department
Predicted Productivity Score
Top Performers
Productivity Improvement

What is the difference between the latest version and the previous version of the Employee Productivity Prediction Model?

Employee Productivity Prediction Model
LATEST_VERSION
VERSION_OF
Employee Productivity Prediction Model Version3
Employee Productivity Prediction Model Version2

Orchestrator Flow



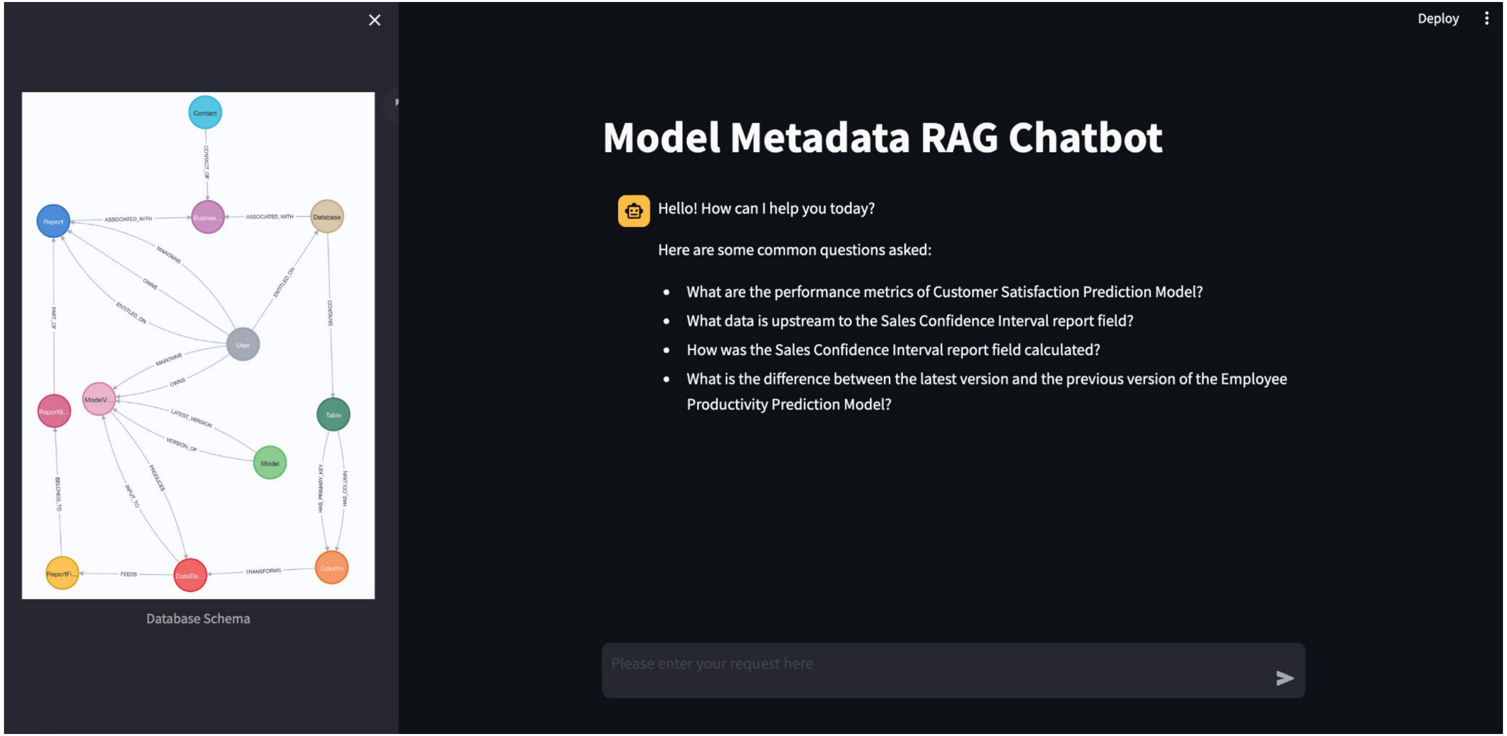
Prompt Engineering

Template 1: Determine user request intent based on examples

Template 2: Given database schema and user question, extract parameter from the question

Template 3: Return the final human readable response

Product Demo



Database Schema

Model Metadata RAG Chatbot

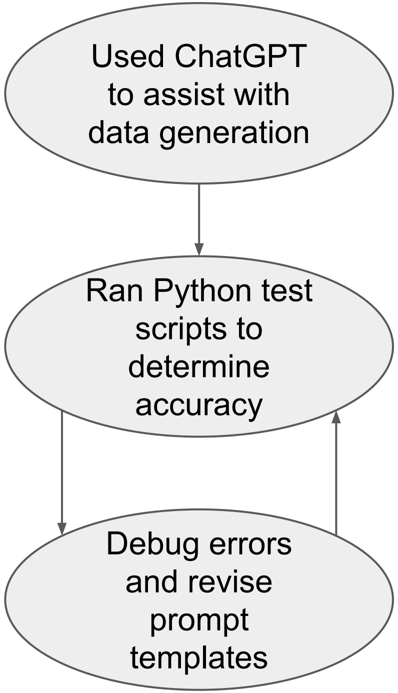
Hello! How can I help you today?

Here are some common questions asked:

- What are the performance metrics of Customer Satisfaction Prediction Model?
- What data is upstream to the Sales Confidence Interval report field?
- How was the Sales Confidence Interval report field calculated?
- What is the difference between the latest version and the previous version of the Employee Productivity Prediction Model?

Please enter your request here

Testing and Learnings


Used ChatGPT to assist with data generation
Ran Python test scripts to determine accuracy
Debug errors and revise prompt templates

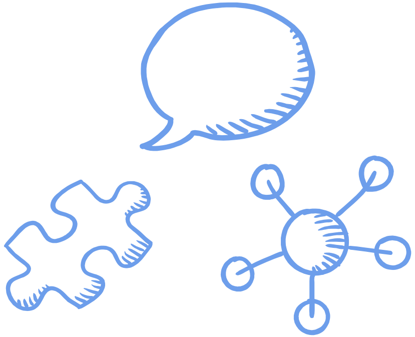
	Common	Uncommon	None
Intent Matching	93%	78%	100%
Parameter Extraction	97%	N/A	N/A
Chatbot Response	99%	60%	N/A

Testing and Learnings

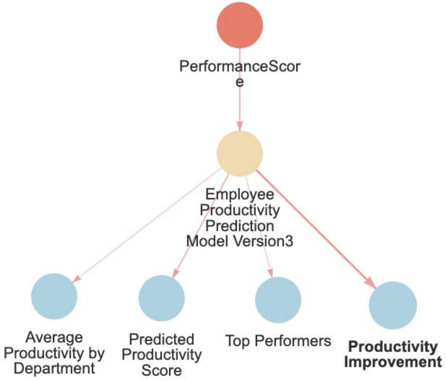
Common Workflow	Uncommon Workflow	General Testing
Added more context in template Cypher queries to guide the LLM when generating the final response	Provided the database schema, the LLM generates incorrect Cypher queries	Difficulty validating chatbot responses

Summary

Metadata is essential



Why RAG Chatbot?



Metadata Management

Future Plans

Large Language Models	User Experience	Other
● Improve intent matching's latency and save costs by using Aurelio Lab's semantic router ● Experiment using different LLMs and their configurations	● Add follow-up questions for when the orchestrator fails to fetch data ● Support multiple parameter requests	● Experiment using one template for both intent matching and parameter extraction ● Hosting Streamlit application

Thank You and QA

Product Demonstration

- Demonstration of our Model Metadata RAG chatbot

Project Implementation

*The prompt template is the key to ensuring the LLM responds in a predictable manner. The template is divided into two separate tasks of **intent matching** and **entity extraction**.*

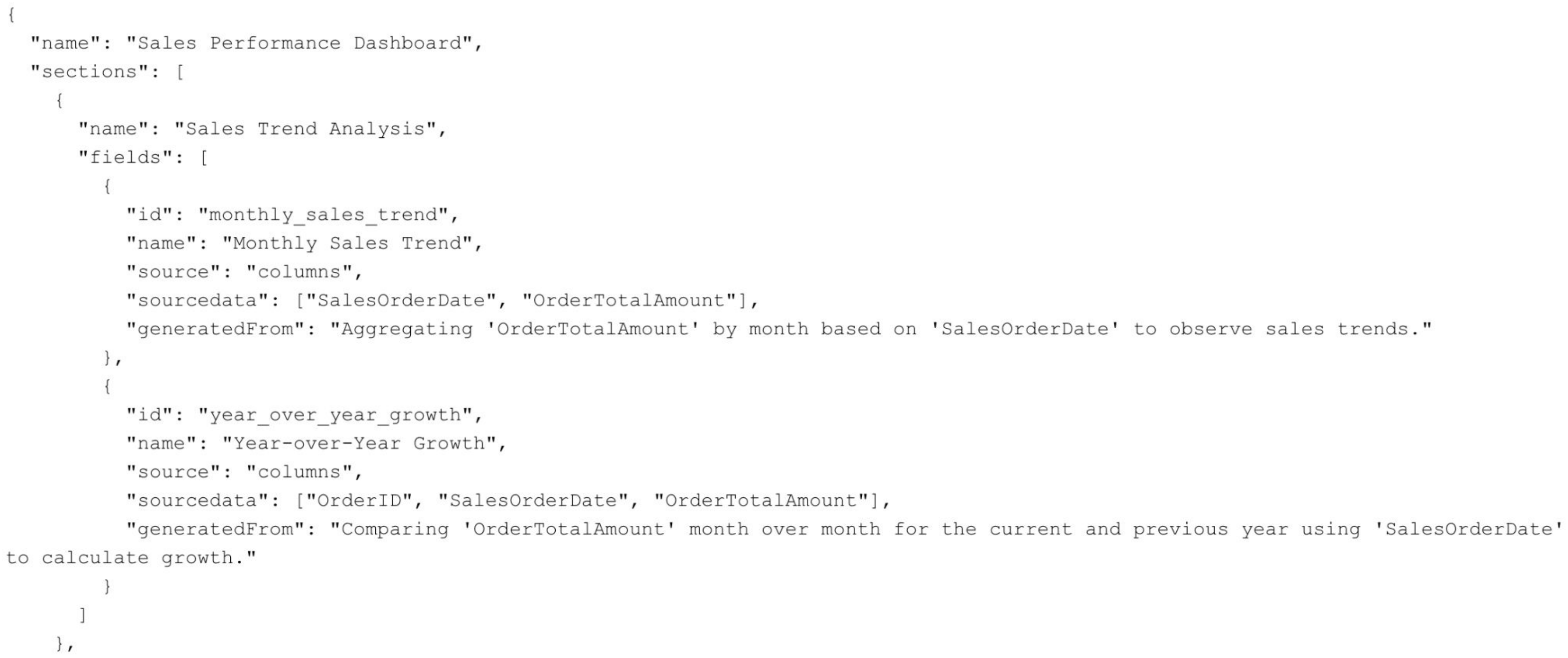
Task 1: Determine user request intent based on the following examples

- Common Questions:
  - What report fields are downstream of a specific column?
  - What are the performance metrics of a specific model?
- Example:
  - Question: What are the performance metrics of Customer Satisfaction Prediction Model?
  - Answer: [COMMON,2]

Task 2: Given a Neo4j schema and a question, extract the single parameter from the question and its data type

- Example:
  - Question: What data is upstream to the Sales Confidence Interval report field?
  - Return [Sales Confidence Interval,ReportField]

Synthetic Data Generation



```
{
  "name": "Sales Performance Dashboard",
  "sections": [
    {
      "name": "Sales Trend Analysis",
      "fields": [
        {
          "id": "monthly_sales_trend",
          "name": "Monthly Sales Trend",
          "source": "columns",
          "sourcedata": ["SalesOrderDate", "OrderTotalAmount"],
          "generatedFrom": "Aggregating 'OrderTotalAmount' by month based on 'SalesOrderDate' to observe sales trends."
        },
        {
          "id": "year_over_year_growth",
          "name": "Year-over-Year Growth",
          "source": "columns",
          "sourcedata": ["OrderID", "SalesOrderDate", "OrderTotalAmount"],
          "generatedFrom": "Comparing 'OrderTotalAmount' month over month for the current and previous year using 'SalesOrderDate'
to calculate growth."
        }
      ]
    },
```

ChatGPT 4 Generated JSON Files

Cypher Query Construction

- Langchain is slow
- Classify the questions
  - Common questions
    - Upstream: how does the number in the reportfield come from?
    - Downstream: If changing a column in the tableA of databaseB, how many reportfield would be affected?
    - model performance
    - modelversion difference
  - Uncommon questions
    - Which users have access to the IT_Database and what are their roles?
  - Irrelevant questions
- Create Cypher Query for common questions

```
MATCH (rf:ReportField {name: "Top Expense
Categories"})
OPTIONAL MATCH
(rf)<-[:FEEDS]-(de1:DataElement)<-[:TRANSFORMS]-(co
l1:Column)-[r1]-(t1:Table)
WITH rf, de1, collect(DISTINCT col1.name) AS cols1
OPTIONAL MATCH
(rf)<-[:FEEDS]-(de2_1:DataElement)<-[:PRODUCES]-(mv
:ModelVersion)<-[:INPUT_TO]-(de2_2:DataElement)<-[:T
RANSFORMS]-(col2:Column)-[r2]-(t2:Table)
WITH rf, de1, cols1, de2_1, collect(DISTINCT col2.name)
AS cols2, mv, collect(DISTINCT de2_2.name) AS de2_2s
WITH
rf,
COALESCE(de1.name, de2_1.name) AS de,
(cols1 + cols2) AS cols,
mv,
de2_2s
RETURN {
ReportField: rf.name,
DataElement_FeedReportField: de,
ModelVersion: mv.name,
DataElement_ModelInput: de2_2s,
Column: cols
} AS result
```

RAG Chatbot Components

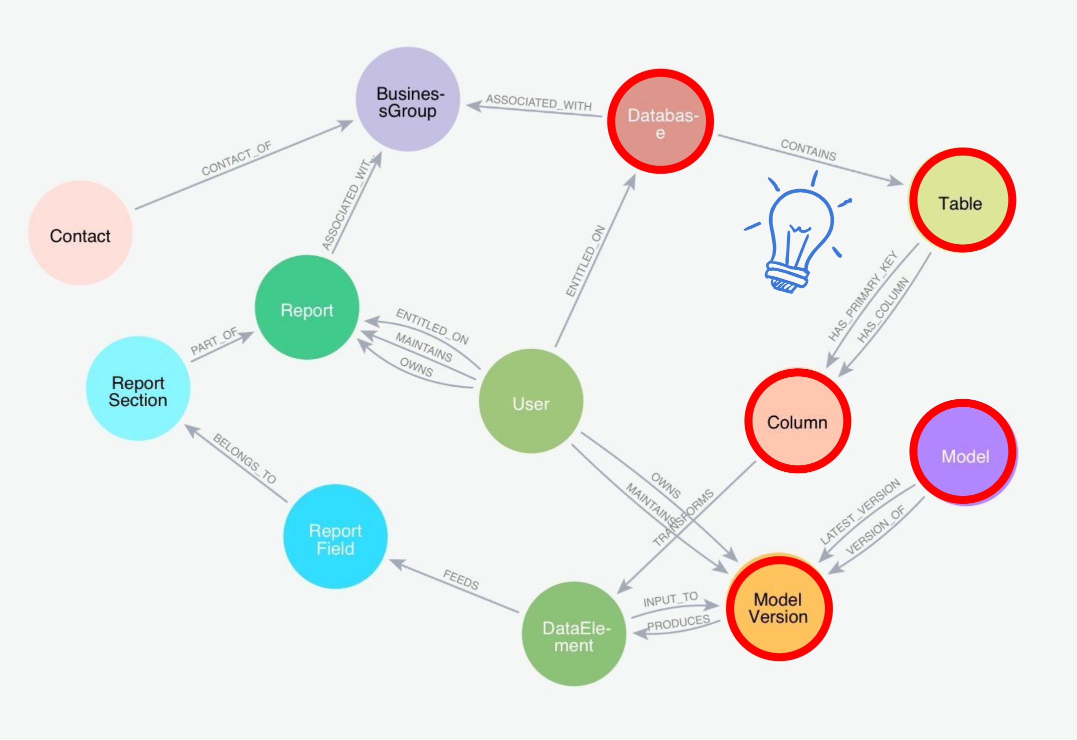
Streamlit

CHATGPT

neo4j

python™

Synthetic Data Generation



Graph DB Schema in Neo4j

Extra resources

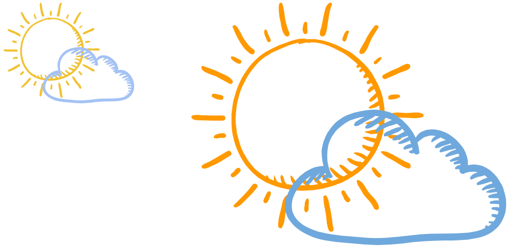
SlidesCarnival icons are editable shapes.

This means that you can:

- Resize them without losing quality.
- Change fill color and opacity.

Isn't that nice? :)

Examples:



Diagrams and infographics