Data Copilot
for
Self-Service Analytics

Wen Gong

Deploy

- Data-Copilot
- Ask-AI
- Evaluations
- QA-Results
- KnowledgeBase
- DataBase
- Configure
- Notes
- Acknowledge

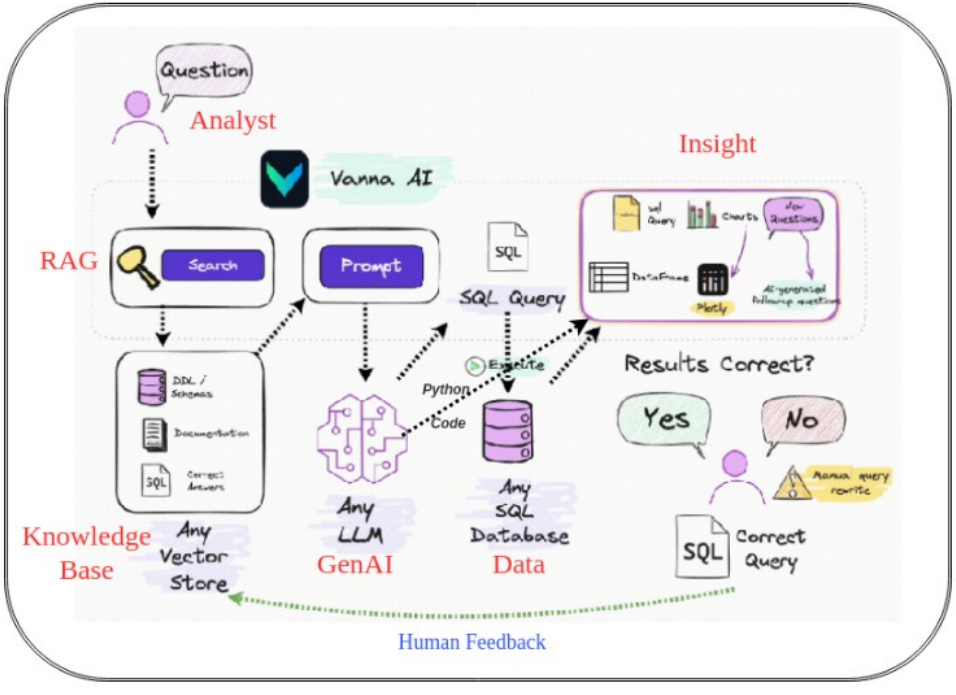
Self-Service Analytics

By streamlining the data-to-insight life-cycle, **Data Copilot** is a game-changer tool for Self-Service Analytics. Built on cutting-edge GenAI models, it empowers data professionals to unlock insights from data faster than ever, therefore allows them to focus on deeper analysis and strategic decision-making.

Key Features

- **Semantic Search**: discover data schema
- **Text-to-SQL**: generate SQL from plain text
- **Data-to-Plot**: generate Python code to visualize data
- **Data Privacy**: achievable by using Ollama and open-source LLM models locally

Architectural Design



Deploy

Data-Copilot

- ❓ Ask-AI
- 💯 Evaluations
- 🚀 QA-Results
- 📚 KnowledgeBase
- 💻 DataBase
- 🛠️ **Configure**
- 📝 Notes
- 💜 Acknowledge

Experiment Setup 🛠️

Data Base

Specify data source:

DB Type: SQLite

DB URL: /home/papagame/projects/1_Biz/vanna/vanna-streamlit/db/chinook.sqlite3

Knowledge Base

Specify vector store:

Vector DB Type: chromadb

GenAI Model

Select LLM model:

GenAI model name

- ○ OpenAI GPT 4
- ◉ OpenAI GPT 3.5 Turbo
- ○ Google Gemini 1.5 Pro
- ○ Anthropic Claude 3.5 Sonnet
- ○ Meta Llama 3 (Open)
- ○ Alibaba QWen 2 (Open)
- ○ Google CodeGemma (Open)
- ○ Google Gemma (Open)
- ○ Mistral (Open)

Save

	id	vector_db	llm_vendor	llm_model	llm_api_key	db_type	db_url
0	e3e81113-8e2a-45c3-ac20-fe84abff4d49	chromadb	OpenAI	gpt-3.5-turbo	None	SQLite	/home/papagame/projects/1_Biz/vanna/vanna-streamlit/db/

Deploy

Data-Copilot

- ❓ Ask-AI
- 💯 Evaluations
- 🚀 QA-Results
- 📚 **KnowledgeBase**
- 💻 DataBase
- 🛠️ Configure
- 📝 Notes
- 💜 Acknowledge

KnowledgeBase 📚

Show Training data

Add Training data

DDL script

```
CREATE TABLE IF NOT EXISTS t_person (
    id INT PRIMARY KEY,
    name text,
    email text
```

Prime the vector-store with database and table creation scripts

Add DDL script

Add ALL DDL scripts

Question

Get book counts

SQL query

```
select count(*) from t_book;
```

Prime it with working queries

Add SQL query

Documentation

table "t_book" stores information on book title and author

Prime it with business terms and metadata info

Add Documentation

Remove Training data

Deploy

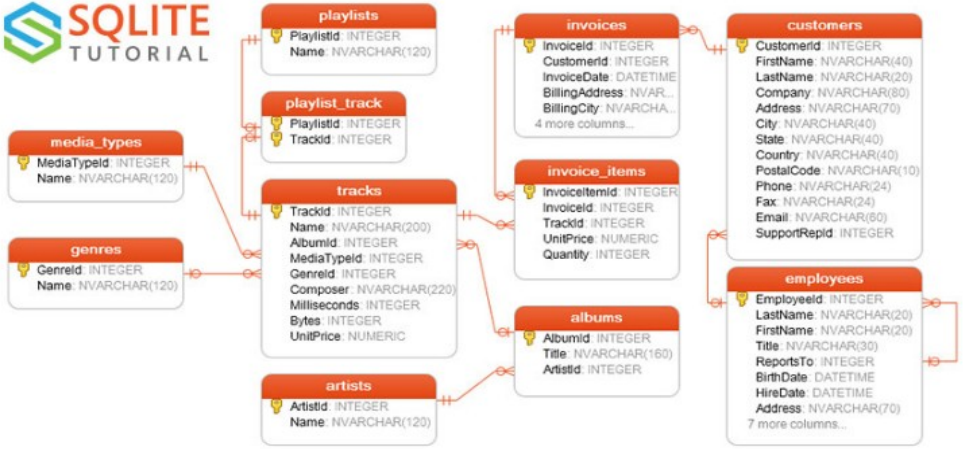
Data-Copilot

- Ask-AI
- Evaluations
- QA-Results
- KnowledgeBase
- DataBase
- Configure
- Notes
- Acknowledge

DataBase

Schema

- SQLite public dataset: Chinook music store



SQL Editor

Table:

customers

Show schema

Schema:

```
CREATE TABLE "customers"
(
  [CustomerId] INTEGER PRIMARY KEY AUTOINCREMENT NOT NULL,
  [FirstName] NVARCHAR(40)  NOT NULL,
  [LastName] NVARCHAR(20)  NOT NULL,
  [Company] NVARCHAR(80),
```

SQL:

```
select * from customers limit 10;
```

SQL Editor

Deploy

Data-Copilot

- Ask-AI
- Evaluations
- QA-Results
- KnowledgeBase
- DataBase
- Configure
- Notes
- Acknowledge

Output Settings:

- [x] Show SQL Query
- [x] Show Dataframe
- [x] Show Python Code
- [x] Show Plotly Chart
- [x] Show Summary
- [ ] Debug

Example prompts:

- List all the tables
- What tables store order information?
- Find top 5 customers by sales
- List all customers from Canada and their email addresses
- Find the top 5 most expensive tracks

Question #1

Ask AI ❓

Find top 5 customers by sales

Question asked

```sql
SELECT c.CustomerId, c.FirstName, c.LastName, SUM(i.Total) AS TotalSales
FROM customers c
JOIN invoices i ON c.CustomerId = i.CustomerId
GROUP BY c.CustomerId
ORDER BY TotalSales DESC
LIMIT 5;
```

SQL generated

	CustomerId	FirstName	LastName	TotalSales
0	6	Helena	Holý	49.62
1	26	Richard	Cunningham	47.62
2	57	Luis	Rojas	46.62
3	45	Ladislav	Kovács	45.62
4	46	Hugh	O'Reilly	45.62

Dataframe returned

```python
import plotly.express as px

if len(df) == 1:
    fig = px.indicators.number(
        value=df['TotalSales'].iloc[0],
        title="Total Sales"
```

Python generated

Ask me a question about your data

Deploy

Data-Copilot

- ❓ Ask-AI
- 💯 Evaluations
- 🚀 QA-Results
- 📚 KnowledgeBase
- 💻 DataBase
- 🛠️ Configure
- 📝 Notes
- 💜 Acknowledge

Output Settings:

- [x] Show SQL Query
- [x] Show Dataframe
- [x] Show Python Code
- [x] Show Plotly Chart
- [x] Show Summary
- [ ] Debug

Example prompts:

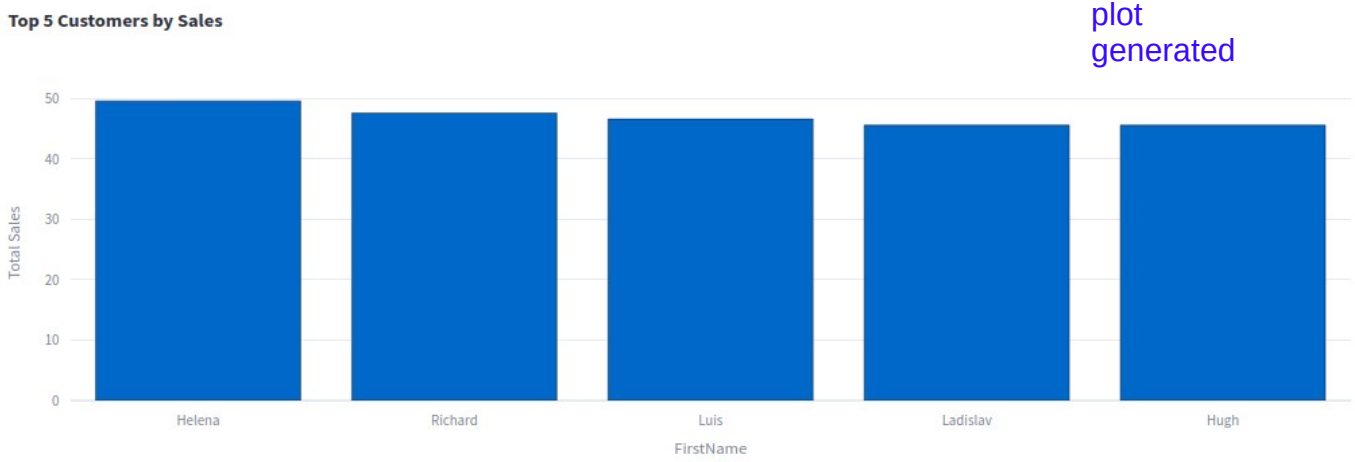
- List all the tables
- What tables store order information?
- Find top 5 customers by sales
- List all customers from Canada and their email addresses
- Find the top 5 most expensive tracks

Question #1

```python
import plotly.express as px

if len(df) == 1:
    fig = px.indicators.number(
        value=df['TotalSales'].iloc[0],
        title="Total Sales"
    )
else:
    fig = px.bar(df, x='FirstName', y='TotalSales', title='Top 5 Customers by Sales', labels={'TotalSales':'Total Sales'})
```

plot generated



summary

The top 5 customers by sales are Helena Holý, Richard Cunningham, Luis Rojas, Ladislav Kovács, and Hugh O'Reilly.

Ask me a question about your data

Deploy

Data-Copilot

- ❓ Ask-AI
- 💯 Evaluations
- 🚀 QA-Results
- 📚 KnowledgeBase
- 💻 DataBase
- 🛠️ Configure
- 📝 Notes
- 💜 Acknowledge

Output Settings:

- [x] Show SQL Query
- [x] Show Dataframe
- [x] Show Python Code
- [x] Show Plotly Chart
- [x] Show Summary
- [ ] Debug

Example prompts:

- List all the tables
- What tables store order information?
- Find top 5 customers by sales
- List all customers from Canada and their email addresses
- Find the top 5 most expensive tracks

Ask AI ❓

Question #2

List all the tables

```
SELECT name
FROM sqlite_master
WHERE type = 'table';
```

	name
0	albums
1	sqlite_sequence
2	artists
3	customers
4	employees
5	genres
6	invoices
7	invoice_items
8	media_types
9	playlists

The list of all tables in the database includes tables such as albums, artists, customers, employees, genres, invoices, and others.

Ask me a question about your data

Deploy

Question #3

Data-Copilot

- ❓ Ask-AI
- 💯 Evaluations
- 🚀 QA-Results
- 📚 KnowledgeBase
- 💻 DataBase
- 🛠️ Configure
- 📝 Notes
- 💜 Acknowledge

Output Settings:

- [x] Show SQL Query
- [x] Show Dataframe
- [x] Show Python Code
- [x] Show Plotly Chart
- [x] Show Summary
- [ ] Debug

Example prompts:

- List all the tables
- What tables store order information?
- Find top 5 customers by sales
- List all customers from Canada and their email addresses
- Find the top 5 most expensive tracks

Ask AI ❓

What tables store order information?

Invalid SQL: The "invoices" and "invoice_items" tables store customer's order information.

Ask me a question about your data

Deploy

Data-Copilot

- Ask-AI
- Evaluations
- QA-Results
- KnowledgeBase
- DataBase
- Configure
- Notes
- Acknowledge

Output Settings:

- [x] Show SQL Query
- [x] Show Dataframe
- [x] Show Python Code
- [x] Show Plotly Chart
- [x] Show Summary
- [ ] Debug

Example prompts:

- List all the tables
- What tables store order information?
- Find top 5 customers by sales
- List all customers from Canada and their email addresses
- Find the top 5 most expensive tracks

Ask AI ❓

Question #4

List all customers from Canada and their email addresses

```sql
SELECT FirstName, LastName, Email
FROM customers
WHERE Country = 'Canada';
```

	FirstName	LastName	Email
0	François	Tremblay	ftremblay@gmail.com
1	Mark	Philips	mphilips12@shaw.ca
2	Jennifer	Peterson	jenniferp@rogers.ca
3	Robert	Brown	robbrown@shaw.ca
4	Edward	Francis	edfrancis@yachoo.ca
5	Martha	Silk	marthasilk@gmail.com
6	Aaron	Mitchell	aaronmitchell@yahoo.ca
7	Ellie	Sullivan	ellie.sullivan@shaw.ca

The data includes a list of customers from Canada along with their email addresses.

Ask me a question about your data

Deploy

Data-Copilot

- Ask-AI
- Evaluations
- QA-Results
- KnowledgeBase
- DataBase
- Configure
- Notes
- Acknowledge

Evaluation 💯

Summary

Results by asking 24 questions on Chinook dataset using 9 LLM models

- Closed: gpt-4, gpt-3.5, claude-3.5-sonnet, gemini-1.5-pro
- Open: llama3, qwen2, codegemma, gemma, mistral

9 LLM models

24 questions

Great results

	J	K	L	M	N	O	P	Q	R	S	T
1	Question	gpt-4	gpt-3.5	Claude-3.5-sonnet	Gemini-1.-5-pro	llama3	qwen2	codeg	gemm	mistra	aya
5	what are the top 5 countries that customers come from?	pass	pass	pass	pass	pass	pass	pass	pass	failed	pass
6	List all albums and their corresponding artist names	pass	pass	pass	failed	pass	pass	pass	pass	pass	
7	Find all tracks with a name containing "What" (case-insensitive)	pass	pass	pass	pass	pass	pass	pass	failed	pass	
8	Get the total number of invoices for each customer	pass	pass	pass	pass	pass	pass	pass	pass	pass	
9	Find the total number of invoices per country:	pass	pass	pass	pass	pass	pass	pass	pass	pass	
10	List all invoices with a total exceeding $10:	pass	pass	pass	pass	pass	pass	pass	pass	pass	
11	Find all invoices since 2010 and the total amount invoiced:	pass	pass	failed	pass	failed	pass	pass	pass	pass	
12	List all employees and their reporting manager's name (if any)	pass	pass	pass	pass	pass	pass	pass	pass	failed	
13	Get the average invoice total for each customer	pass	pass	pass	pass	pass	pass	pass	pass	pass	
14	Find the top 5 most expensive tracks (based on unit price)	pass	pass	pass	pass	pass	pass	pass	pass	failed	
15	List all genres and the number of tracks in each genre:	pass	pass	pass	pass	pass	pass	pass	pass	pass	
16	Get all genres that do not have any tracks associated with them:	pass	pass	pass	pass	pass	pass	pass	pass	pass	
17	List all customers who have not placed any orders:	pass	pass	pass	pass	pass	pass	failed	pass	pass	
18	There are 3 tables: artists, albums and tracks, where albums and artists are linked by ArtistId, albums and tracks are linked by AlbumId. Can you find the top 10 most popular artists based on the number of tracks	pass	pass	pass	pass	pass	pass	pass	pass	failed	
19	List all customers from Canada and their email addresses:	pass	pass	pass	pass	pass	pass	failed	failed	pass	
20	Find the customer with the most invoices	pass	pass	pass	pass	pass	pass	pass	pass	pass	
21	Find the customer who bought the most albums in total quantity (across all invoices)	pass	pass	pass	pass	pass	failed	pass	pass	failed	
22	Hint: album quantity is found in invoice_items, Find the top 5 customers who bought the most albums in total quantity (across all invoices):	pass	pass	pass	pass	pass	pass	pass	pass	pass	
23	Find the top 5 customers who spent the most money overall, Hint: order total can be found on invoices table, calculation using invoice_items detail table is unnecessary	pass	pass	pass	pass	pass	pass	failed	failed	pass	
24	Get all playlists containing at least 10 tracks and the total duration of those tracks:	pass	pass	pass	pass	pass	pass	pass	pass	failed	
25	Identify artists who have albums with tracks appearing in multiple genres:	pass	pass	pass	pass	pass	failed	pass	failed	pass	
26											
27	Success-rate (%)	100.00	100.00	95.83	95.83	95.83	91.67	87.50	83.33	70.83	

Deploy

Data-Copilot

- ❓ Ask-AI
- 💯 Evaluations
- 🚀 QA-Results
- 📚 KnowledgeBase
- 💻 DataBase
- 🛠️ Configure
- 📝 Notes
- 💜 **Acknowledge**

Acknowledgement 💜

- Vanna.ai → RAG App framework
- Streamlit → Web UI
- RAG
- Ollama → Serve LLM model locally
- SQLite → Database
- SQL Assistant
- Emoji Cheatsheet