



Foreman

Building a tailored data assistant using dbt metadata

Luís Ferreira and João Rebelo

Porto, Portugal



The *Data* Warehouse, *Reborn*

🔧 *Started from the bottom...*

- Siloed data
- High maintenance and operational costs
- No DevOps processes
- Poor data governance
- Lack of flexibility

...now we're here 🔧

- Flexible, easily accessible data
- Cloud-based, scalable, performant infrastructure
- Adoption of DevOps & DataOps
- Focus on metadata and "semantic" layers
- Better tooling, open-source initiatives

There's plenty of *'inventory'*. And yet...

"Data practitioners blamed poor data quality on lack of documentation (31%), lack of tooling (27%), and teams not understanding each other (22%)"

- [source](#)



"Only 37.8% of companies have created a data-driven organization. In the same survey, 45.1% of executives reported that they compete on data and analytics, but a majority — 54.9% — stated that they do not."

- [source](#)

"Only 22% of organizations trust their own data, and more than half (56%) of executives don't make managing data a high investment priority."

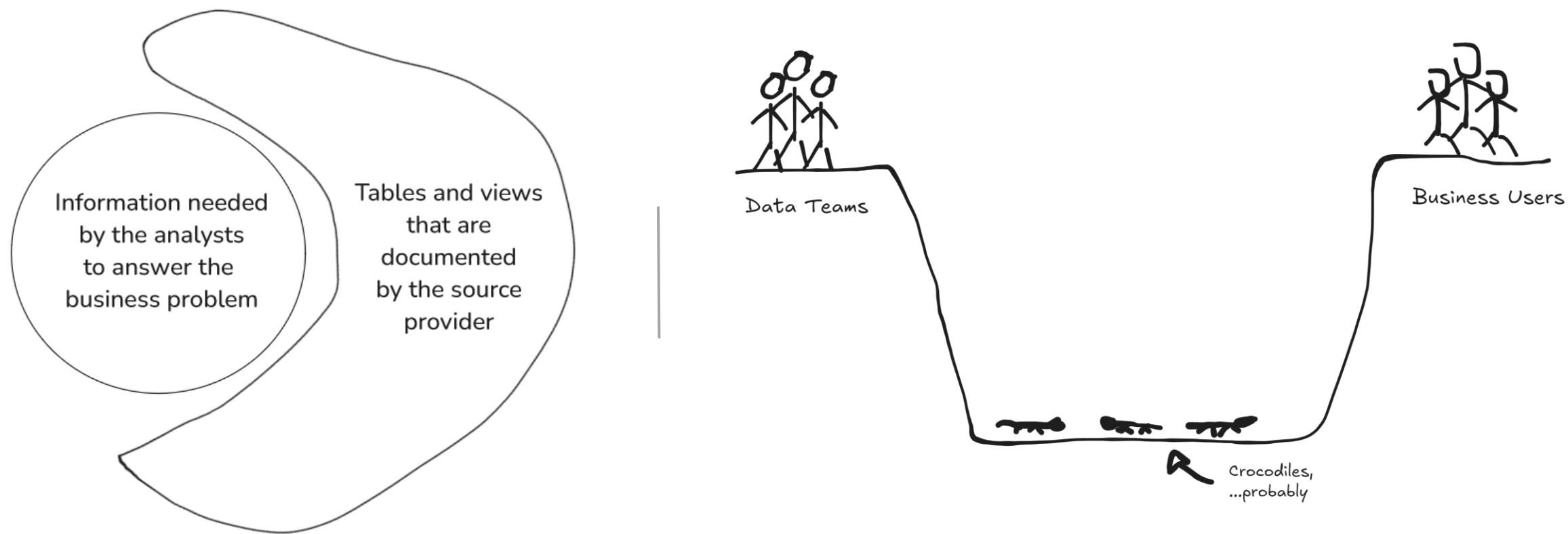
- [source](#)

"Up to 73 Percent of Company Data Goes Unused for Analytics"

- [source](#)

Data Literacy

...or the art of knowing data and how to use data



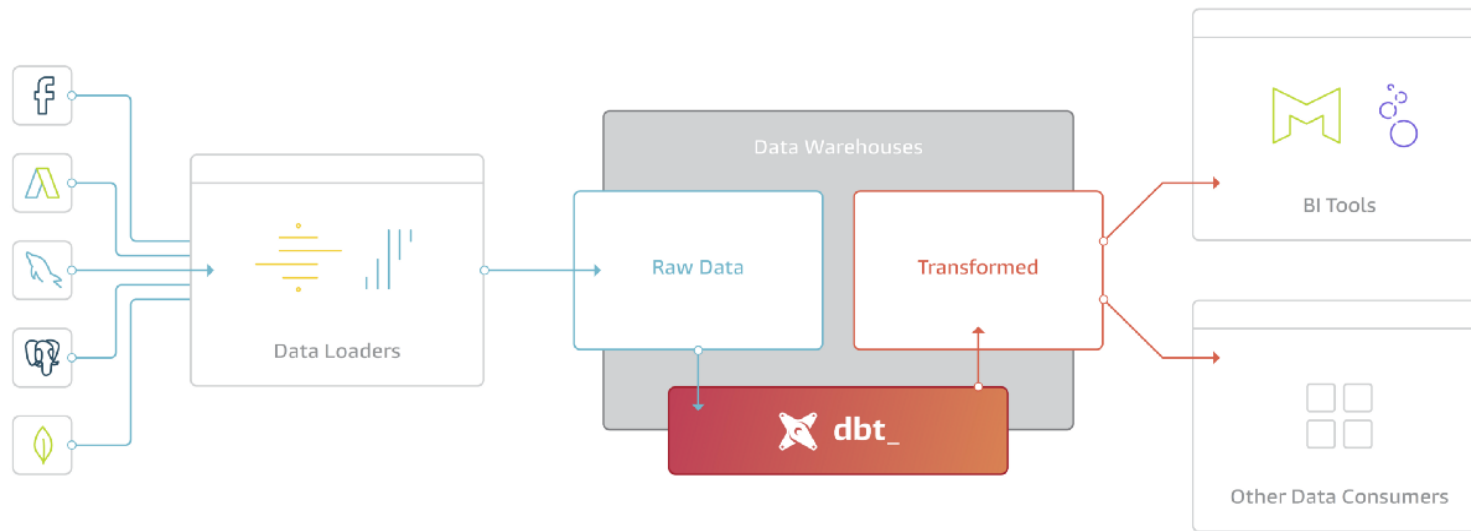
So...what are data
teams doing to help?



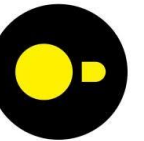
dbt and “Analytics Engineering”

*"**dbt** enables data analysts and engineers to transform their data using the same practices that software engineers use to build applications."*

- [source](#)

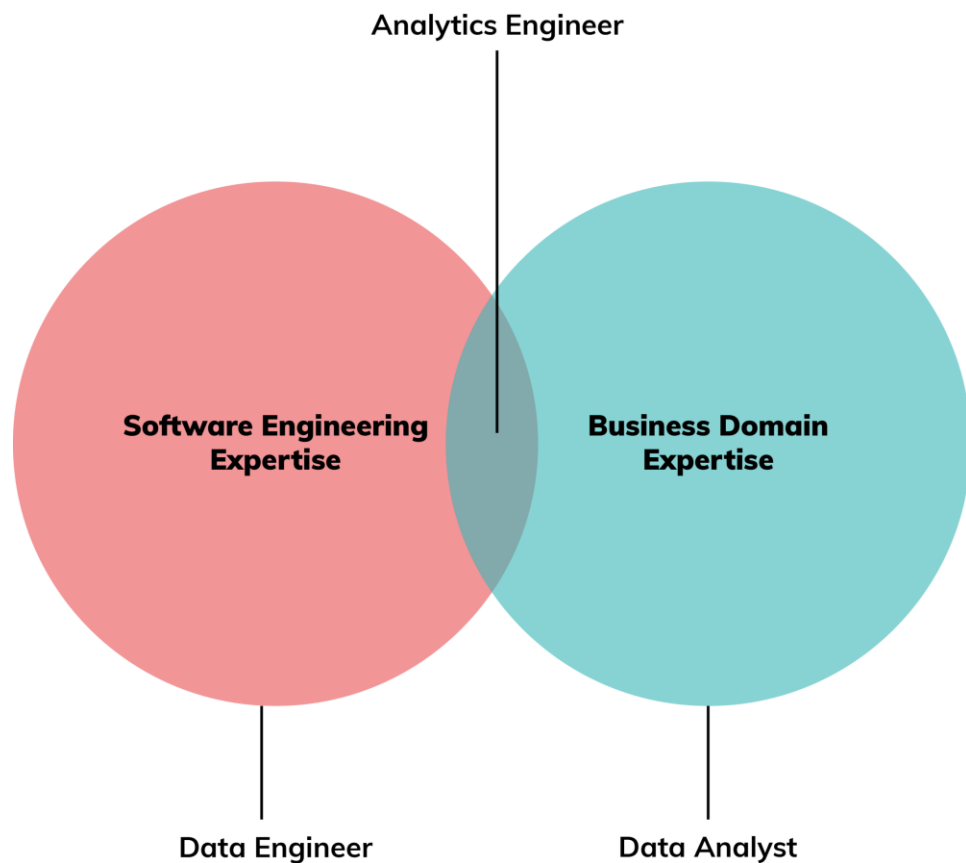


- [source](#)



...and more

dbt and “Analytics Engineering”



Overview - DBT Docs

localhost:8080/index.html#!/model/model.fishtown_internal_analytics.fct_subscription_transactions

Search for models...

dbt

Overview

Project Database

fishtown_internal_analytics

- models
 - admin
 - marts
 - consulting
 - finance
 - mrr
 - fct_subscription_transactions
 - utils
 - transactions
 - product
 - staging
 - util
- quickbooks
- redshift
- snowplow
 - models
 - base
 - identification
 - page_views
 - snowplow_page_views
 - snowplow_web_events
 - snowplow_web_events_internal_fixed
 - snowplow_web_events_scroll_depth
 - snowplow_web_events_time
 - snowplow_web_page_context

analytics.fct_subscription_transactions table

Details Description Columns SQL

Description

Subscription Transactions - Fact Table

About this table

This table contains information about revenue earned by Fishtown Analytics. Transactions are categorized by product, category, and customer.

Data Sources

This table is sourced from:

- Trello boards for consulting clients
- Quickbooks
- Stripe

Columns

TESTS	COLUMN	TYPE
P	id	character varying
F N	customer_id	character varying
N	customer_name	character varying
N	category	character varying
N	date_month	date

Lineage Graph

subscription_transactions_typed

fct_subscription_transactions

This column indicates ...

GenAI for Data Analytics & Data Literacy

Google Cloud Data and AI Trends Report 2024

Trend 1 Trend 2 Trend 3 Trend 4 Trend 5

Why should you care?

It's almost impossible to overstate how significantly gen AI has changed the technological landscape. In the case of business intelligence (BI), as tools become more accessible, even non-technical team members will be able to benefit from these insights; driving productivity and disseminating knowledge faster than ever before. That means **better data literacy** across your organization, **smarter decisions** being made, and ultimately **greater success** in the market.

52% of non-technical users are already using gen AI to draw out insights today.

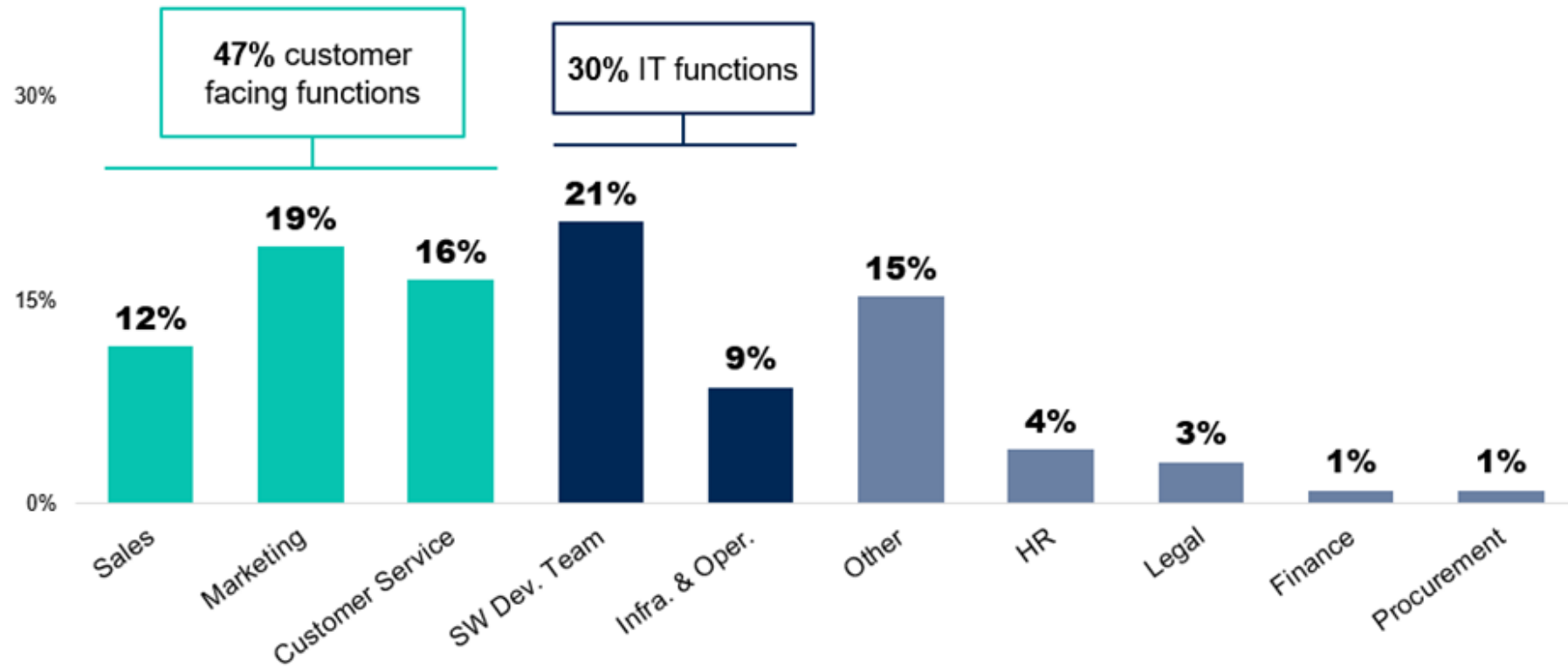
Modern BI tools were already developing ways to bring data to everyone who needed it; reports embedded in the most relevant context for the data, such as account insights appearing in a salesperson's CRM, is an easy example. But those insights have always needed to be carefully curated by an analyst. The end user has always been a step removed from the data. Connecting a large language model to your business data closes that gap. Team members can interact with your data intuitively and conversationally, or create reports and dashboards by simply 'talking' to your data or making a simple search across your business. In fact, many of the organizations surveyed for this report are already putting this into practice.

*“[leads to] **better data literacy** across your organization, **smarter decisions** being made, and ultimately **greater success** in the market.”*

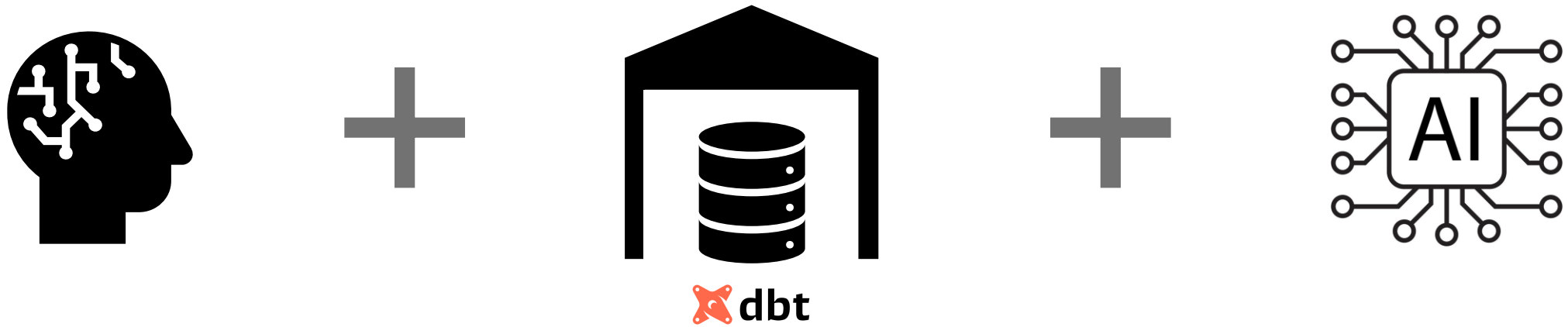
*“52% of **non-technical** users are already using gen AI to draw out insights today.”*

Source: Google's Data And AI Trends Report 2024

GenAI Investment by Business Function



Source: [Gartner](#)



=

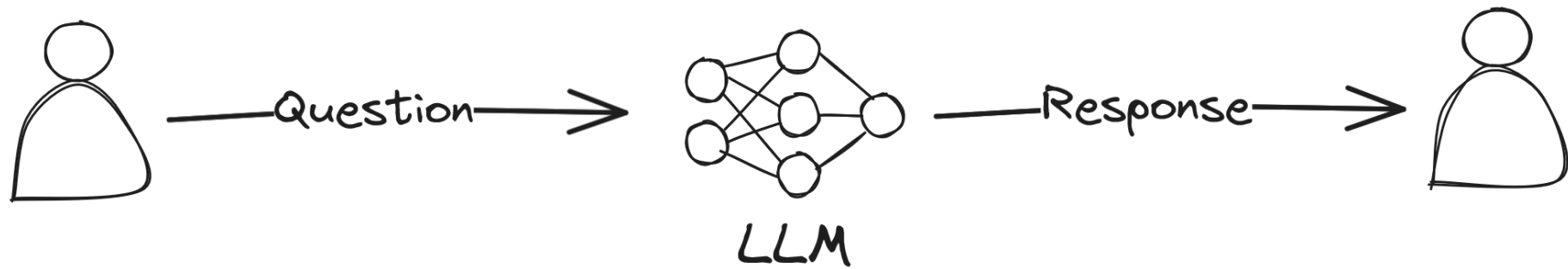
The Data Warehouse



Foreman

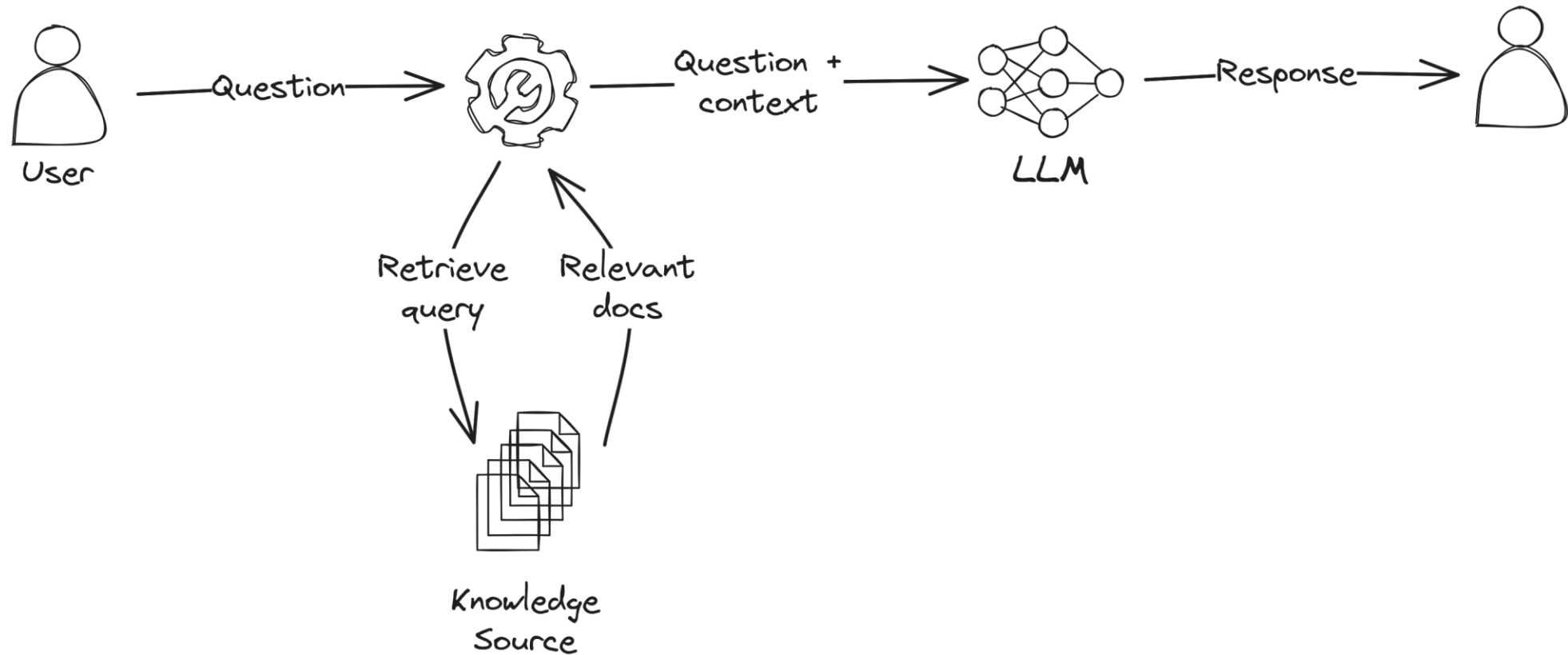
Retrieval Augmented Generation (RAG)

Simple prompt:



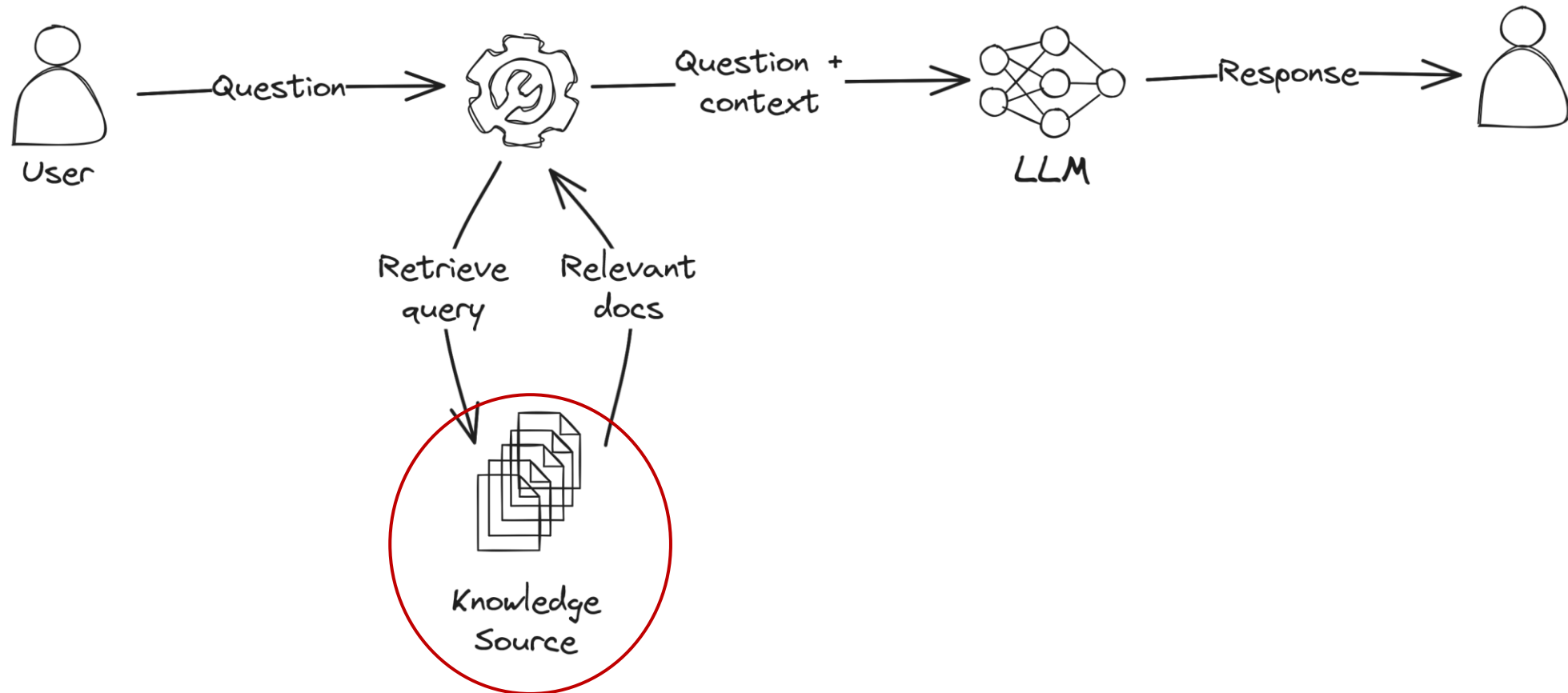
Retrieval Augmented Generation (RAG)

Prompt with RAG:

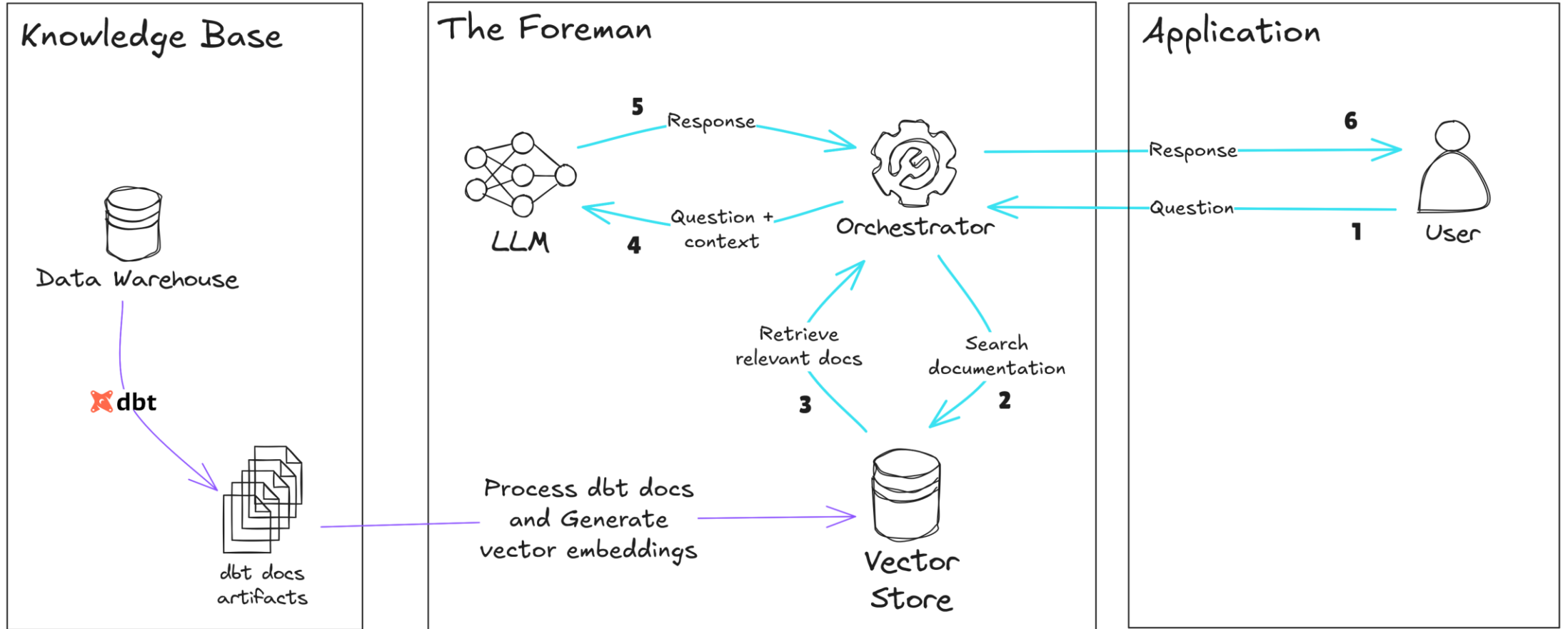


Retrieval Augmented Generation (RAG)

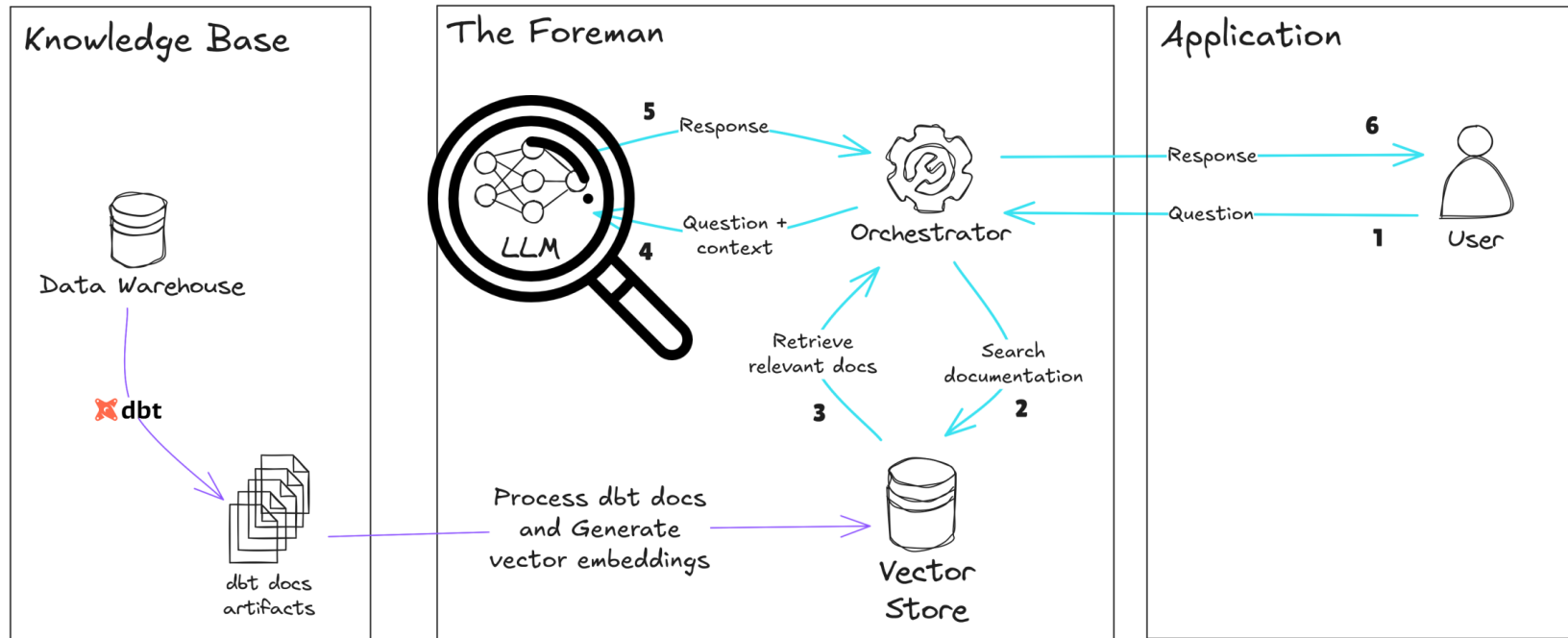
Prompt with RAG:



The 'Foreman' RAG



The 'Foreman' RAG



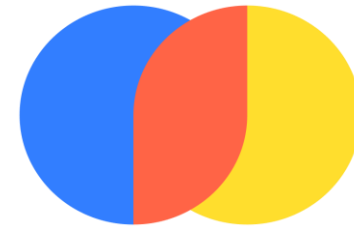
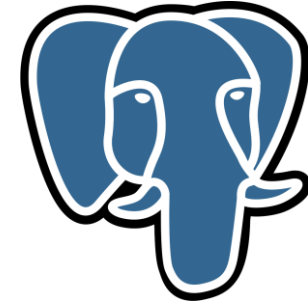
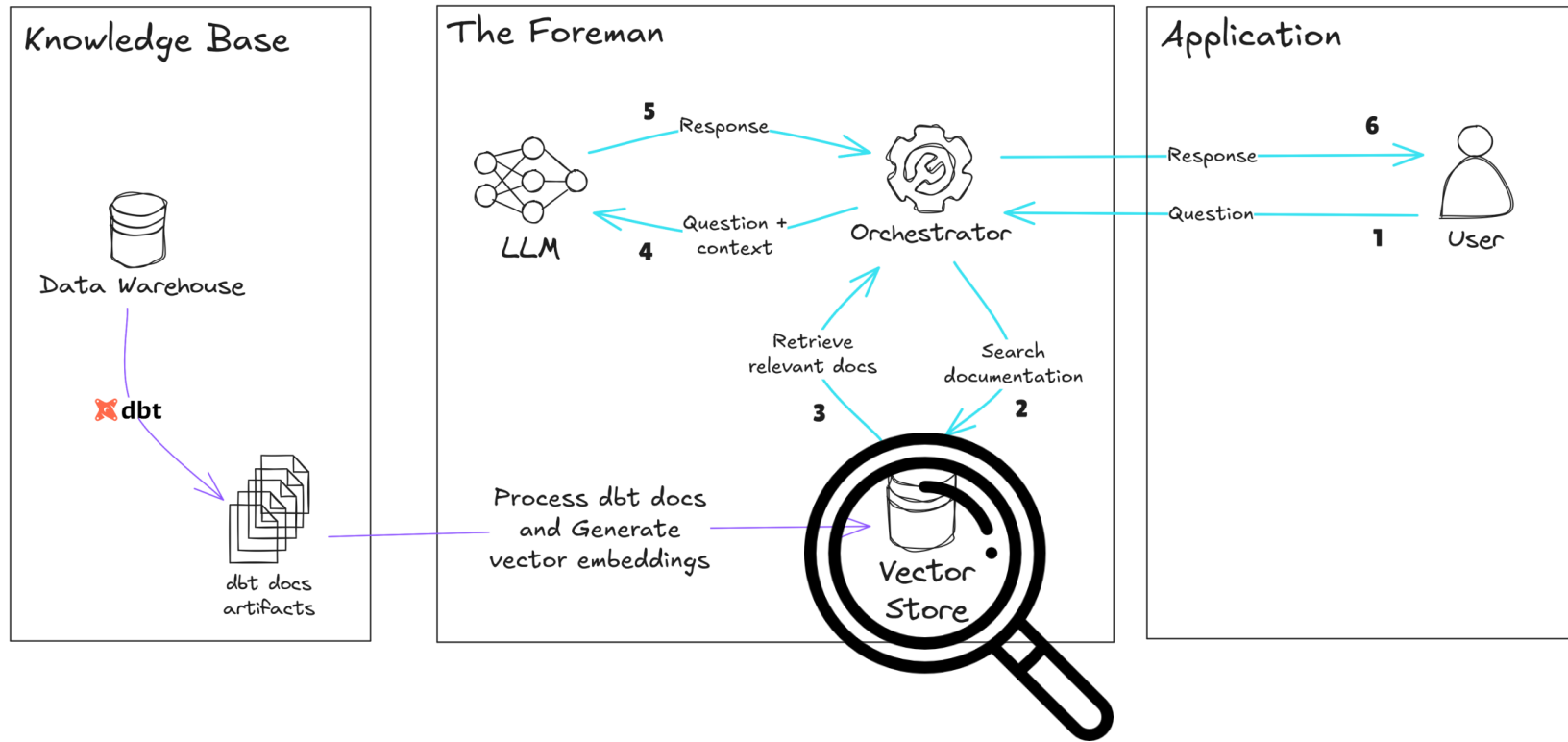
AI

Meta

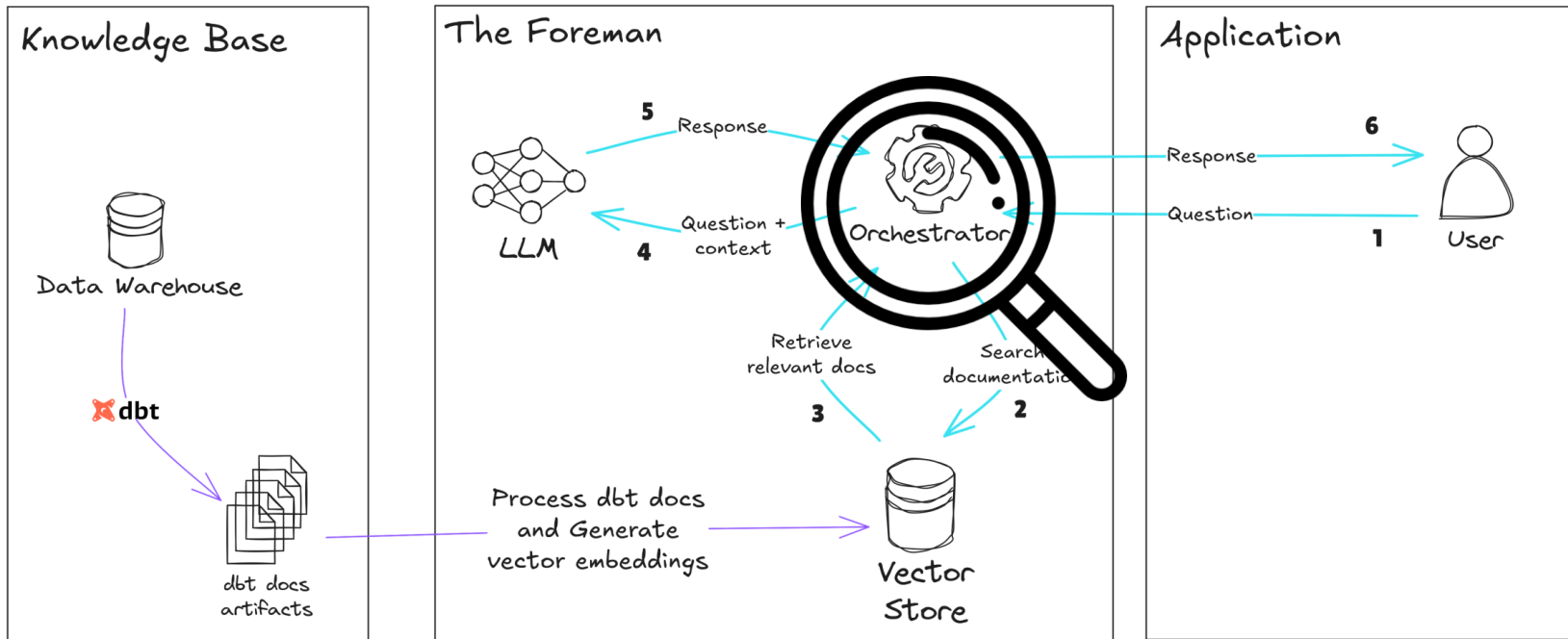


Gemini

The 'Foreman' RAG



The 'Foreman' RAG





Demo


A dbt project for Data Makers Fest 2024




Search for models...


Overview






 Project  Database

 Group

Tables and Views

 dbt

 main

-  attractions
-  partners
-  sessions
-  speakers
-  tutorials

Data Makers Fest 2024

This project contains the data for the Data Makers Fest 2024, under the tagline "the emerging festival for all data practitioners".

About

Data Makers Fest is a festival dedicated to all data makers. For 3 days, we'll gather data professionals and enthusiasts from all fields to learn, network, and make things happen by exploring hot topics in data and AI.

The participants get to embark on a journey into the future of AI and data with dynamic sessions covering cutting-edge topics such as Responsible AI implementation, advancements in Natural Language Processing, and the transformative power of Generative AI.

Topics

The topics the are tackled in the festival, which are also used as tags for the sessions, are listed below.

- Advanced Analytics

Demo

A dbt project for Data Makers Fest 2024



Search for models...

Overview

Project Database Group

Tables and Views

dbt
main
attractions
partners
sessions
speakers
tutorials

speakers view

Details Description Columns Code

Description

Table contains information about the speakers which are presenters of the sessions or tutors in the hands-on tutorials.

Each speaker is uniquely identified by their `id`. The model also tracks their name, the pronouns by which they prefer to be addressed and a short biography.

The `role` column showcases their current professional affiliation including both position (eg. Data Engineer, Data Scientist) and organization (eg. Google, University of Minho). The information is stored as text in the format "{position} at {organization}".

Columns

COLUMN	TYPE	DESCRIPTION	CONSTRAINTS	DATA TESTS	MORE?
id	integer	The identification number of the speaker			>
name	string	The name of the speaker			>
pronouns	string	How the speaker prefers to be addressed			>
role	string	The current position of the speaker			>
bio	string	Short text description of the speaker's background and i...			>



Generic LLM

#1



What is Data Makers Fest?



Data Makers Fest is an event that celebrates data science and analytics, bringing together professionals, enthusiasts, and experts to share knowledge, innovations, and best practices in the field.



Generic LLM

#2



Where is it hosted?



The location of Data Makers Fest can vary each year, so it's best to check the event's official website or announcements for the current venue.



Generic LLM

#3



Tell me one of the cities where it has been hosted



One of the cities where Data Makers Fest has been hosted is San Francisco.



Generic LLM

#4



What types of sessions does it have?



Data Makers Fest typically features a variety of sessions, including keynote speeches, workshops, panel discussions, and hands-on labs focused on data science, analytics, machine learning, and related technologies.



Foreman

powered by dbt docs

#1



What is Data Makers Fest?



Data Makers Fest is a festival dedicated to all data makers. For 3 days, data professionals and enthusiasts from all fields gather to learn, network, and explore hot topics in data and AI. The event includes dynamic sessions on topics such as Responsible AI implementation, advancements in Natural Language Processing, and the transformative power of Generative AI. The festival is held at Alfândega do Porto in Porto.



Foreman

powered by dbt docs

#2



What types of sessions does it have?



Data Makers Fest features two types of sessions: short and long. Additionally, there are hands-on tutorials, which are part of the same agenda but are tracked separately in a different model named `tutorials`.



Foreman

powered by dbt docs

#3



Where are the tutorials taking place?



The tutorials for Data Makers Fest 2024 will be held at Porto Business School on the 23rd of September.

Foreman can also help with analytics by generating SQL

```
{
  "metadata": {
    "dbt_schema_version": "https://schemas.getdbt.com/dbt/manifest/v12.json",
    "dbt_version": "1.8.6",
    "generated_at": "2024-09-11T12:03:25.139090Z",
    "invocation_id": "8f0f1daa-e299-4dea-841d-df3e5cd0263c",
    "env": {},
    "project_name": "warehouse",
    "project_id": "372d30dd2849813ef674855253900679",
    "user_id": "322a1275-475e-40c0-9d87-2716db252f59",
    "send_anonymous_usage_stats": true,
    "adapter_type": "duckdb",
    "nodes": {
      "model.warehouse.partners": {
        "database": "dbt",
        "schema": "main",
        "name": "partners",
        "resource_type": "model",
        "package_name": "warehouse",
        "path": "partners/partners.sql",
        "original_file_path": "models/partners/partners.sql",
        "unique_id": "model.warehouse.partners",
        "fqdn": ["warehouse", "partners", "partners"],
        "alias": "partners",
        "checksum": {
          "name": "sha256",
          "checksum": "4bcfedaa2cfd3a2c54558993b0895486a484d3fbdf7bfb79c128da4ccbde9767"
        },
        "config": {
          "enabled": true,
          "alias": null,
          "schema": null,
          "database": null,
          "tags": [],
          "meta": {},
          "group": null,
          "materialized": "view",
          "incremental_strategy": null,
          "persist_docs": {},
          "post-hook": [],
          "pre-hook": [],
          "quoting": {},
          "column_types": {},
          "full_refresh": null,
          "unique_key": null,
          "on_schema_change": "ignore",
          "on_configuration_change": "apply",
          "grants": {},
          "packages": [],
          "docs": {
            "show": true,
            "node_color": null
          },
          "contract": {
            "enforced": false,
            "alias_types": true,
            "access": "protected"
          }
        }
      }
    }
  }
}
```

Intelligent parsing
of dbt docs metadata

+

```
sql_prompt = """
You are a friendly AI assistant,
helping humans with their questions regarding Data Makers Fest.
The answer should be in SQL format.
Just provide the SQL query, not the result.

Take into account the following information about the event:
Event overview: {event}
Data warehouse documentation: {docs}
"""
```

Prompt template for
Text-to-SQL scenario



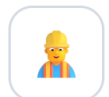
Foreman

powered by dbt docs and Text-to-SQL instructions

#1



How many sessions are there in total?



```
SELECT COUNT(*) AS total_sessions  
FROM sessions;
```



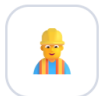
Foreman

powered by dbt docs and Text-to-SQL instructions

#2



How many partners of each type are there?



```
SELECT tier, COUNT(*) AS partner_count  
FROM partners  
GROUP BY tier;
```



Foreman

powered by dbt docs and Text-to-SQL instructions

#3



Are there any speakers with more than one participation?



```
SELECT speaker_id, COUNT(*) as participation_count
FROM (
  SELECT speaker_id FROM sessions
  UNION ALL
  SELECT speaker_id FROM tutorials
) AS all_participations
GROUP BY speaker_id
HAVING COUNT(*) > 1;
```



Foreman

powered by dbt docs and Text-to-SQL instructions

#4



What is the number of organizations that have speakers in the event?



```
SELECT COUNT(DISTINCT SUBSTRING_INDEX(role, ' at ', -1)) AS number_of_organizations
FROM speakers;
```


Lessons Learned

- *Once you get past demo datasets, the performance drops”^{1,2,3}*
- Consistency in semantics is very important²
- Data & ML Engineers can achieve more by working together

[1: Text2SQL is Not Enough: Unifying AI and Databases with TAG \(arxiv.org\)](#)

[2: Snowflake Cortex Analyst: Behind the Scenes](#)

[3: A Case Study: Text-to-SQL Failures on Enterprise Data - Numbers Station](#)

Key Takeaways

- Invest in good documentation and dissemination of domain knowledge
- Use existing technologies to make GenAI more helpful without breaking the bank



DATA MAKERS *fest*

Foreman: Building a tailored data assistant using dbt metadata

Porto, 24/09/2024



Codebase
[luisferreira97/foreman](#)

