

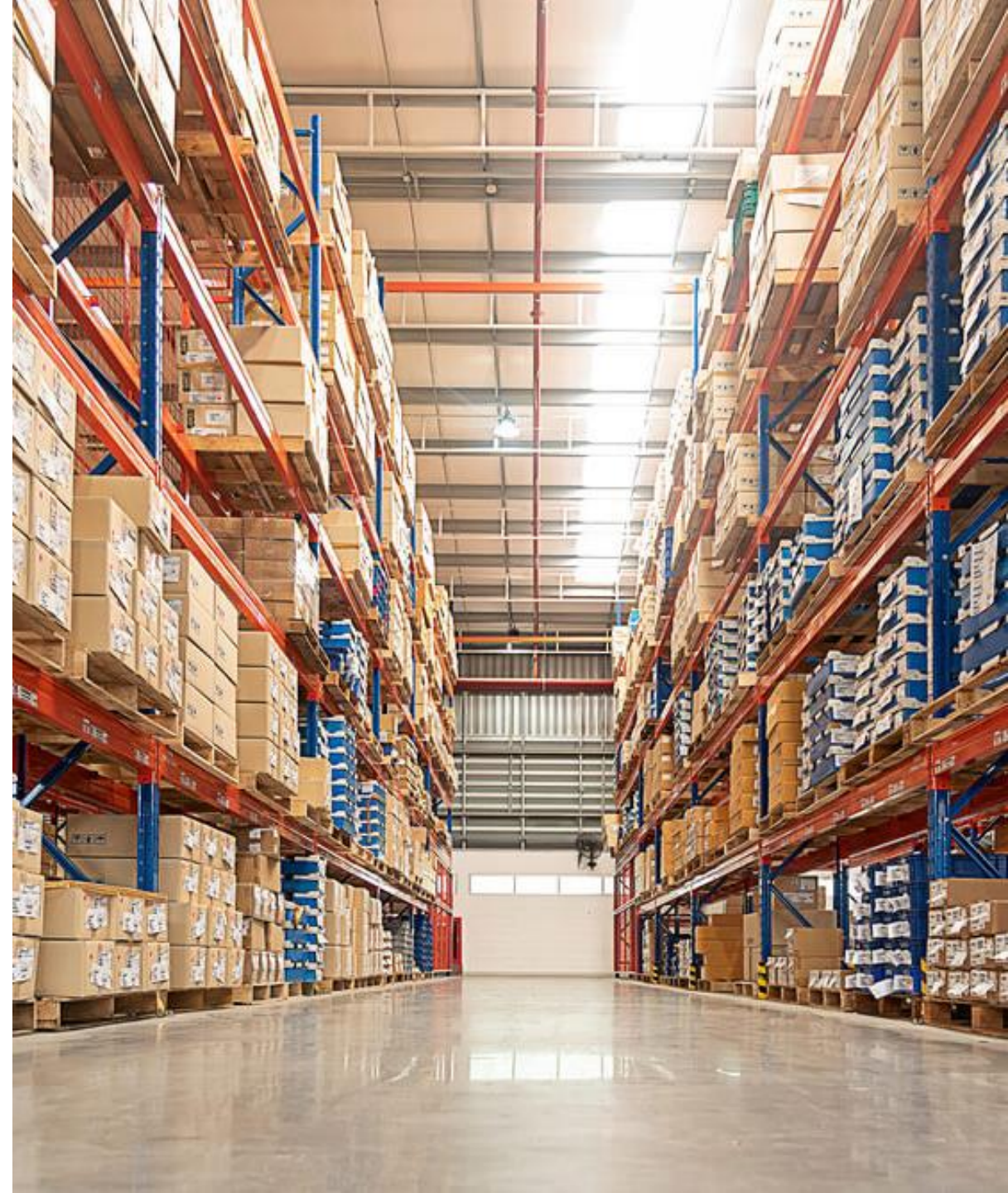


Foreman

Building a tailored data assistant using dbt metadata

Speakers: Luís Ferreira and João Rebelo

Warehouse Foreman



Evolution of Data Warehousing

Traditional Data Warehouse:


- On-premises infrastructure
- High maintenance and operational costs
- Limited scalability

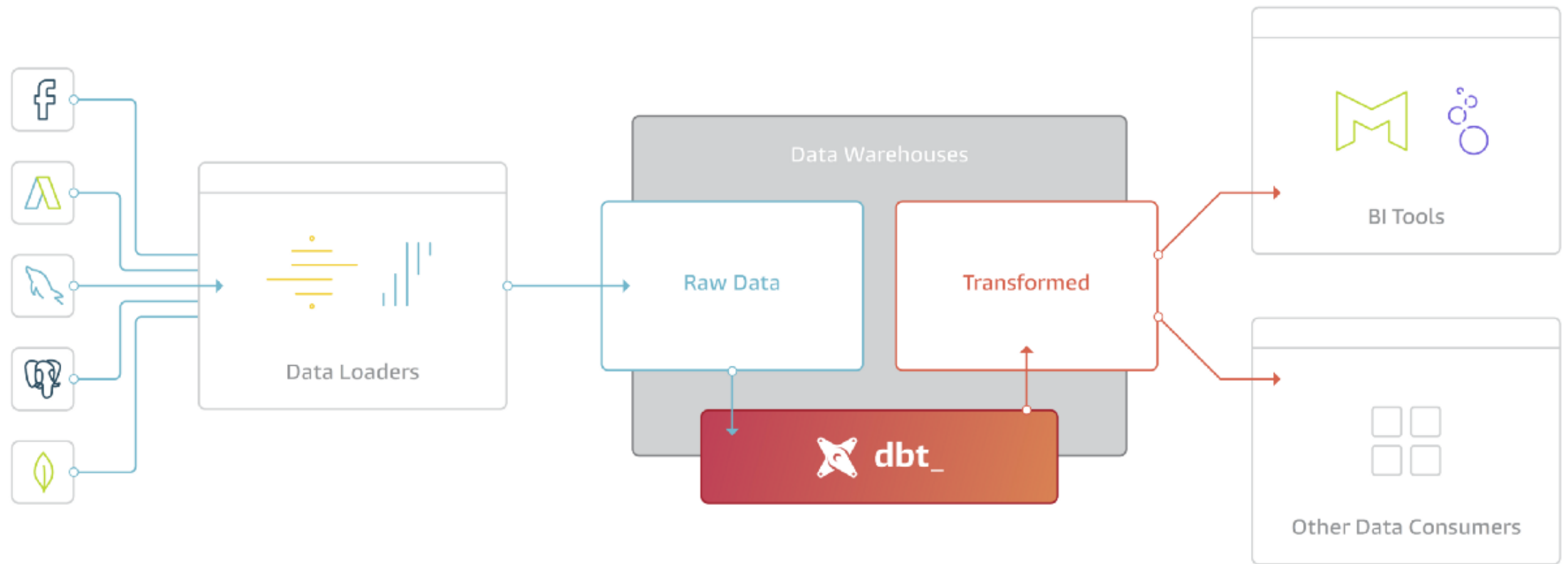
The Cloud Data Warehouse



- Scalable
- Cost Efficient
- Performant
- Integrated

The role of dbt and “Analytics Engineering”

Data Engineering  Analytics Engineering



dbt docs

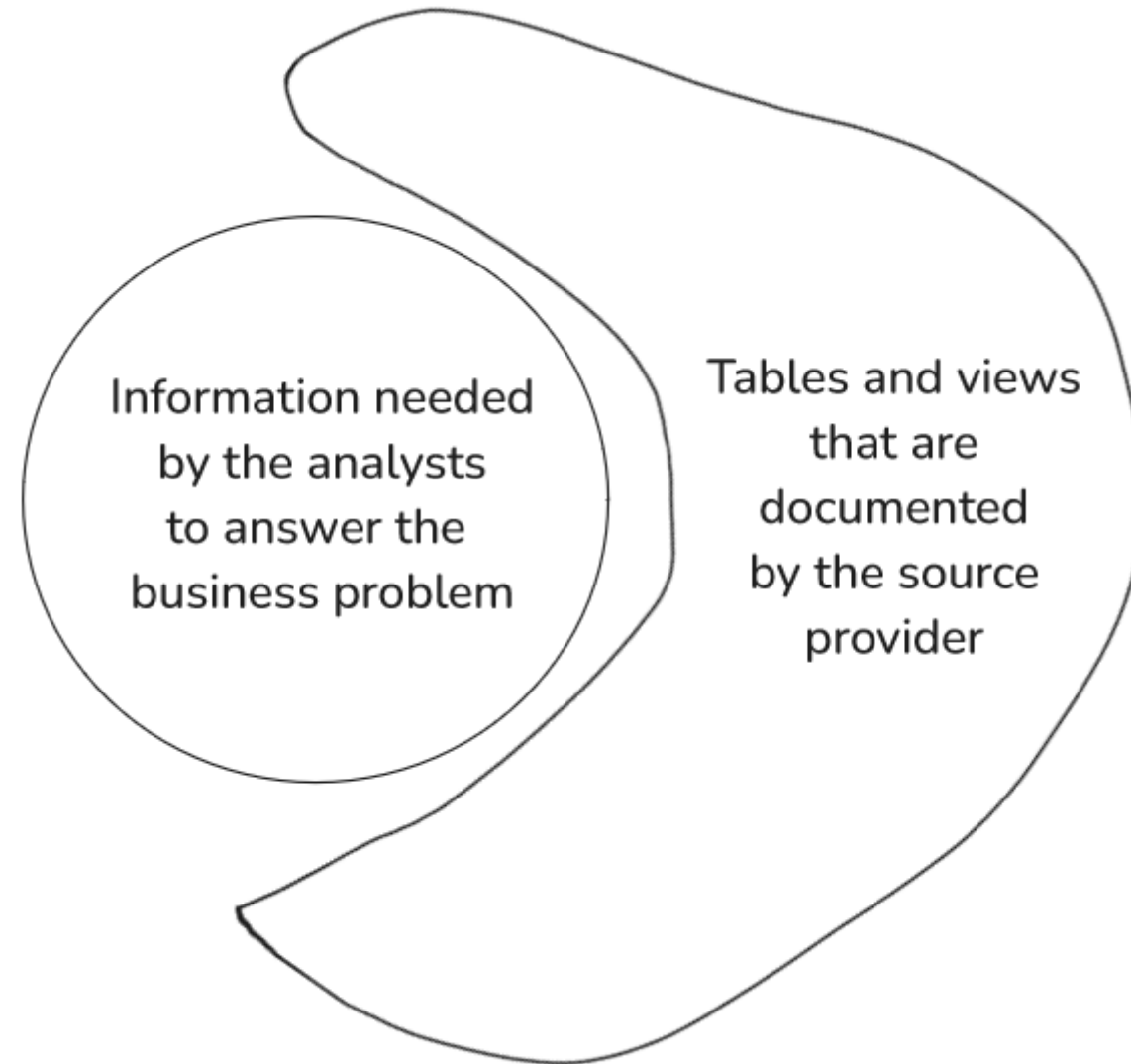
“Good documentation will help downstream consumers discover and understand the datasets you curate for them.”

The screenshot displays the DBT Docs interface for the 'analytics.fct_subscription_transactions' table. The interface is divided into several sections:

- Overview:** Shows the table name and a search bar.
- Project:** A sidebar on the left showing the file tree structure, including folders like 'models', 'admin', 'marts', 'consulting', 'finance', 'mrr', 'utils', 'transactions', 'product', 'staging', 'util', 'quickbooks', 'redshift', 'snowplow', and 'page_views'.
- Description:** A section titled 'Subscription Transactions - Fact Table' providing information about the table's content and data sources.
- Data Sources:** A list of sources including 'Trello boards for consulting clients', 'Quickbooks', and 'Stripe'.
- Columns:** A table listing the columns of the fact table, including 'id', 'customer_id', 'customer_name', 'category', and 'date_month'.
- Lineage Graph:** A dark blue overlay on the right showing the relationship between 'subscription_transactions_typed' and 'fct_subscription_transactions'.

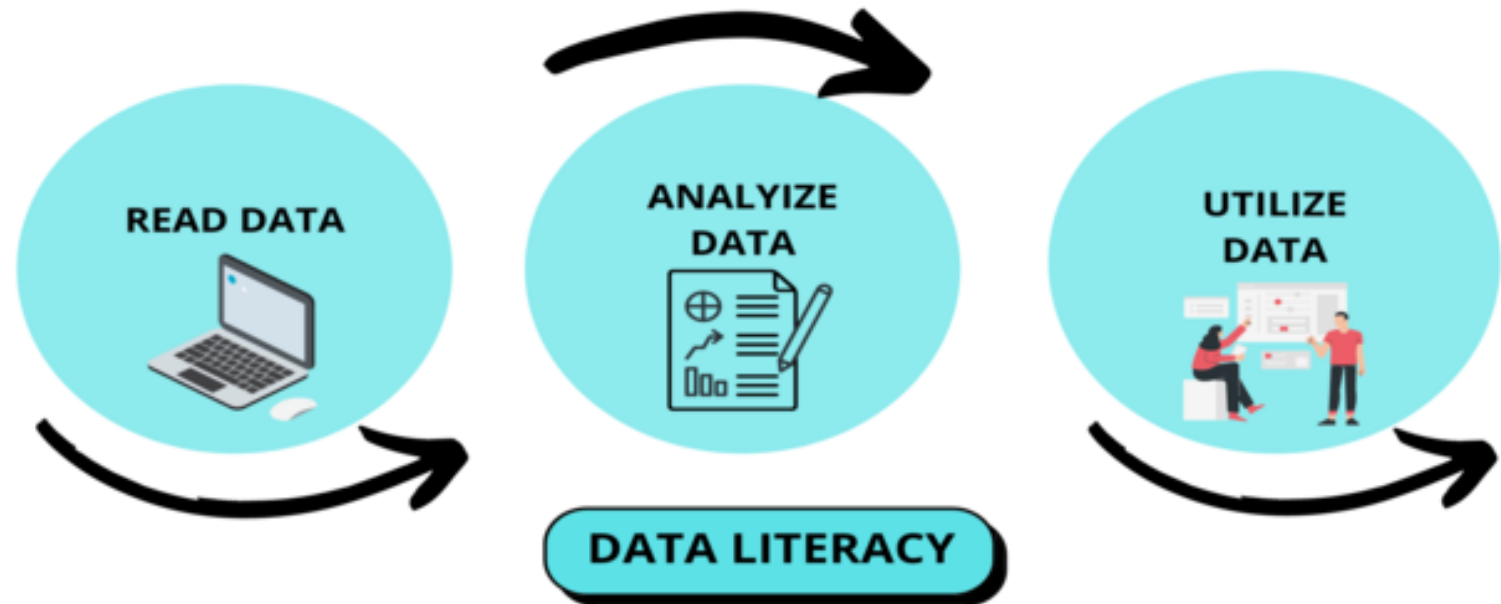
The 'Lineage Graph' shows a flow from 'subscription_transactions_typed' (top) to 'fct_subscription_transactions' (bottom), indicating that the fact table is derived from the typed subscription transactions.

Data Warehouse: Common Problems



Enhancing Data Literacy

“The ability to read, understand, create, and communicate data as information.”



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.

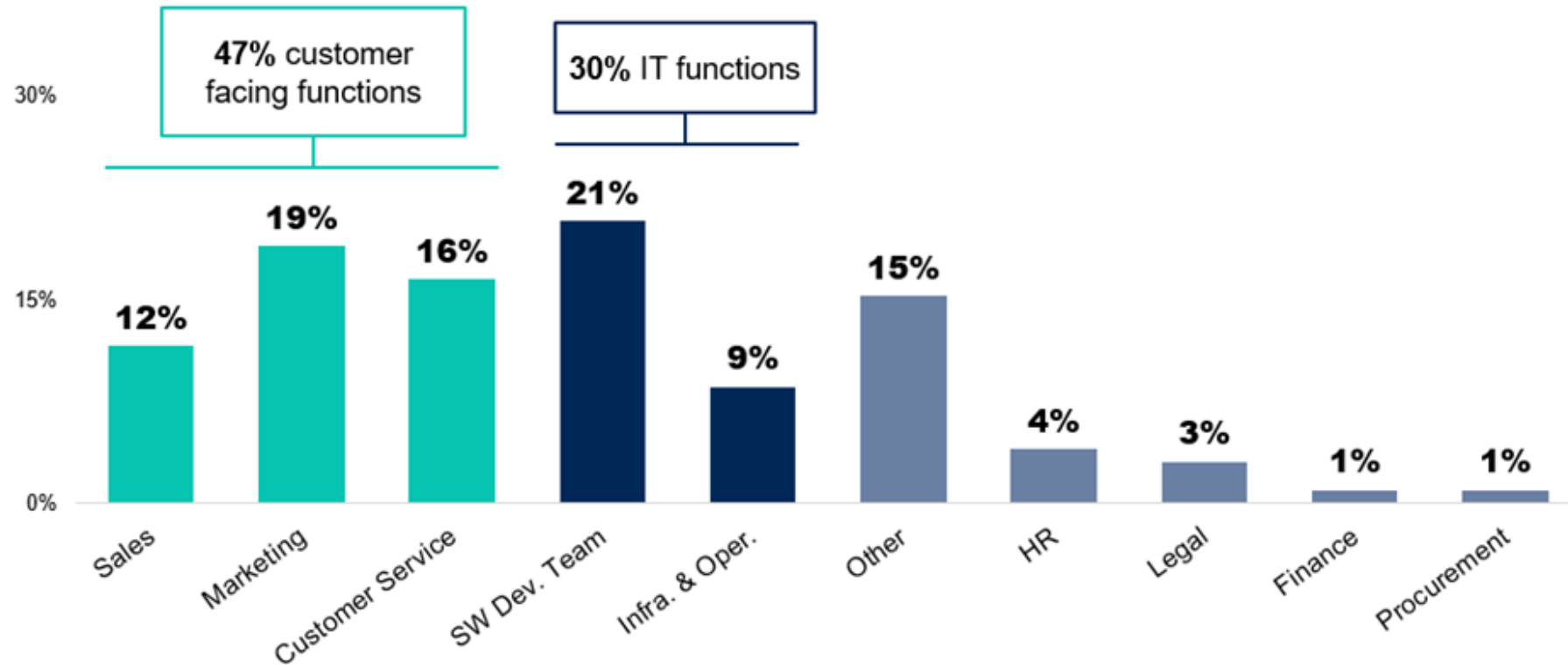
52% of non-technical users are already using gen AI to draw out insights today.

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.

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

Source: Google's Data And AI Trends Report 2024

GenAI Investment by Business Function



Source: [Gartner](#) (October 2023)

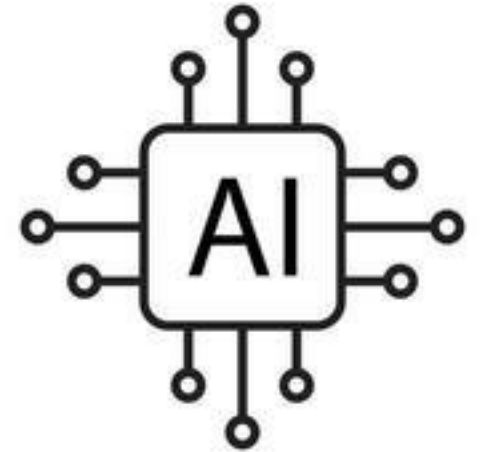
The 'Data Warehouse Foreman'



+

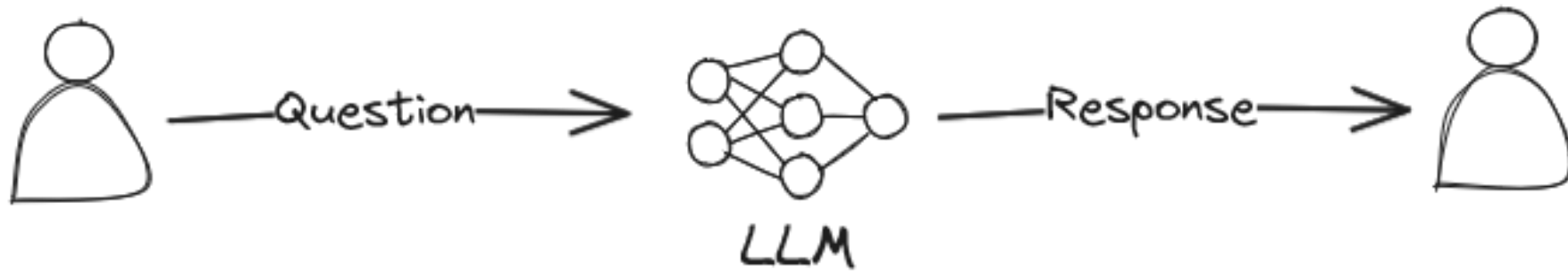


+



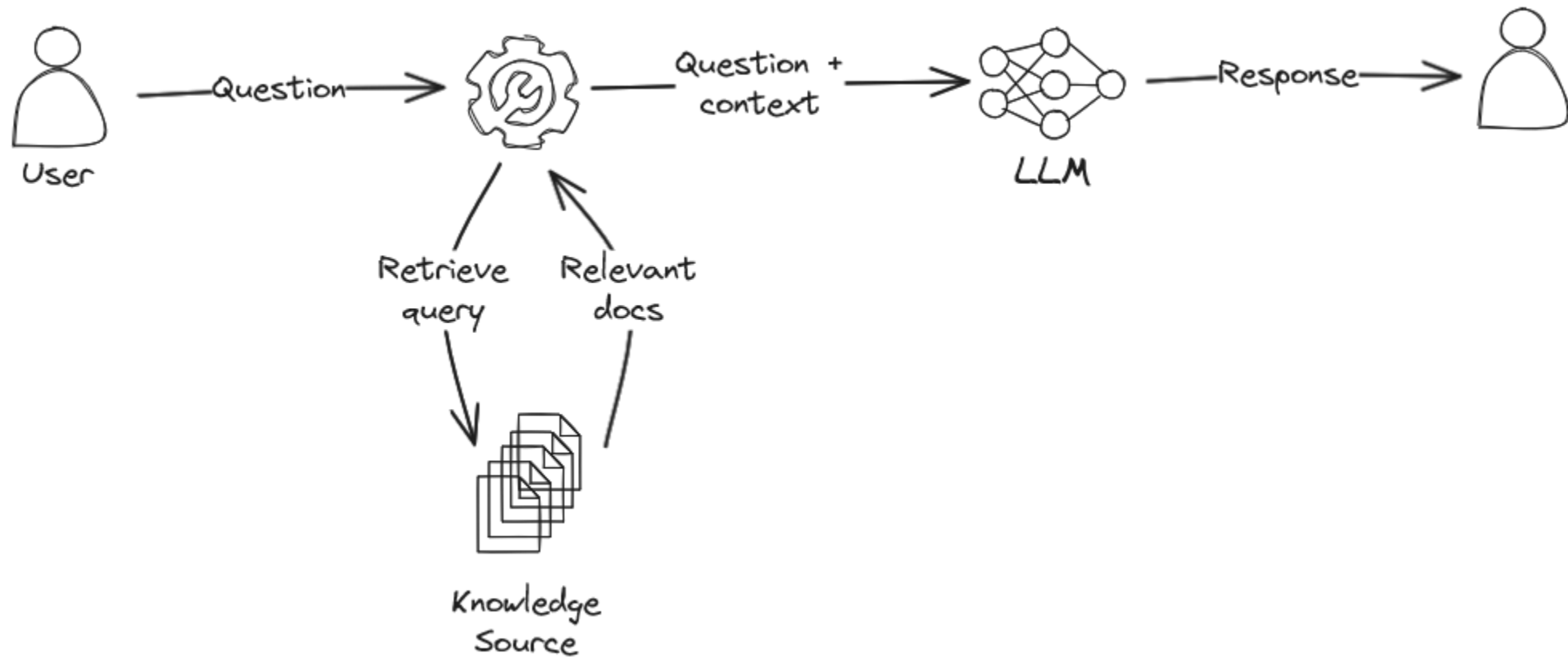
Retrieval Augmented Generation (RAG)

Simple prompt:



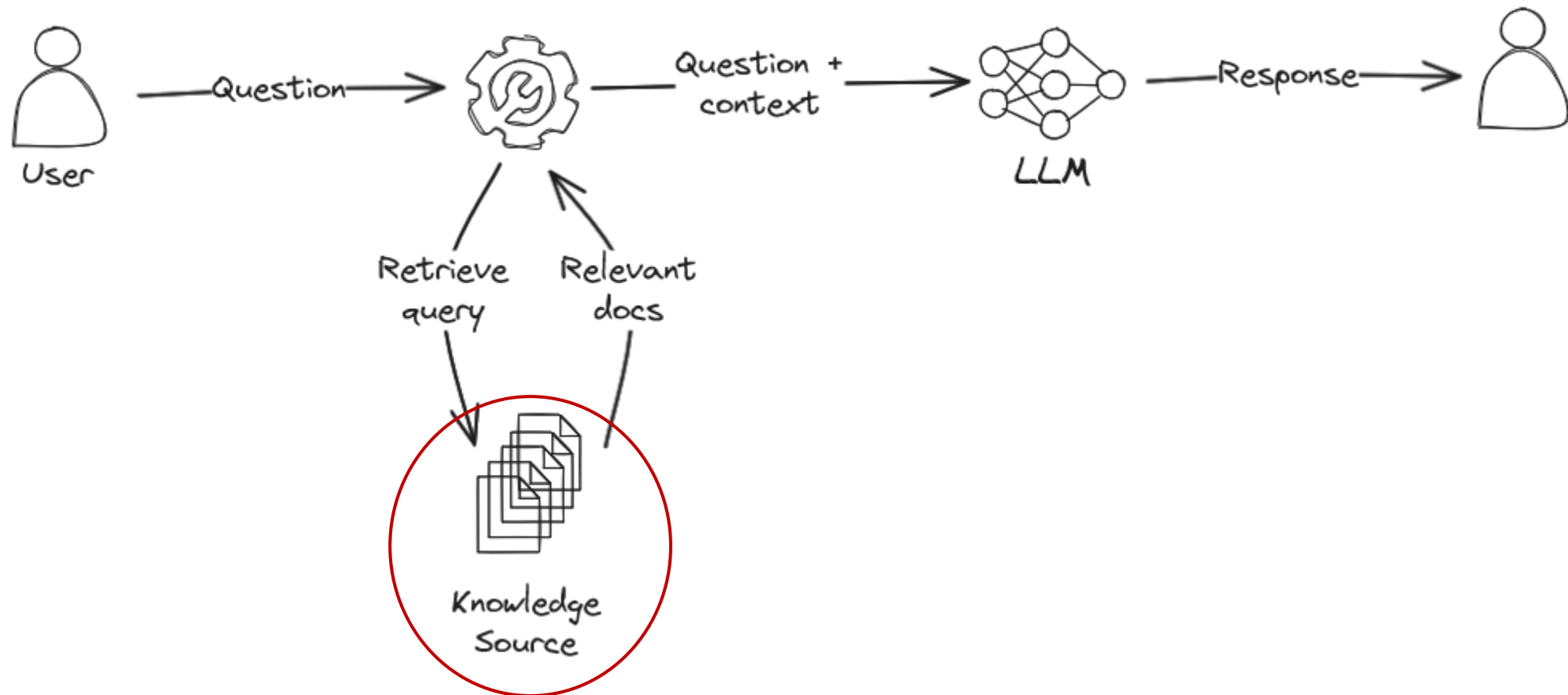
Retrieval Augmented Generation (RAG)

Prompt with RAG:

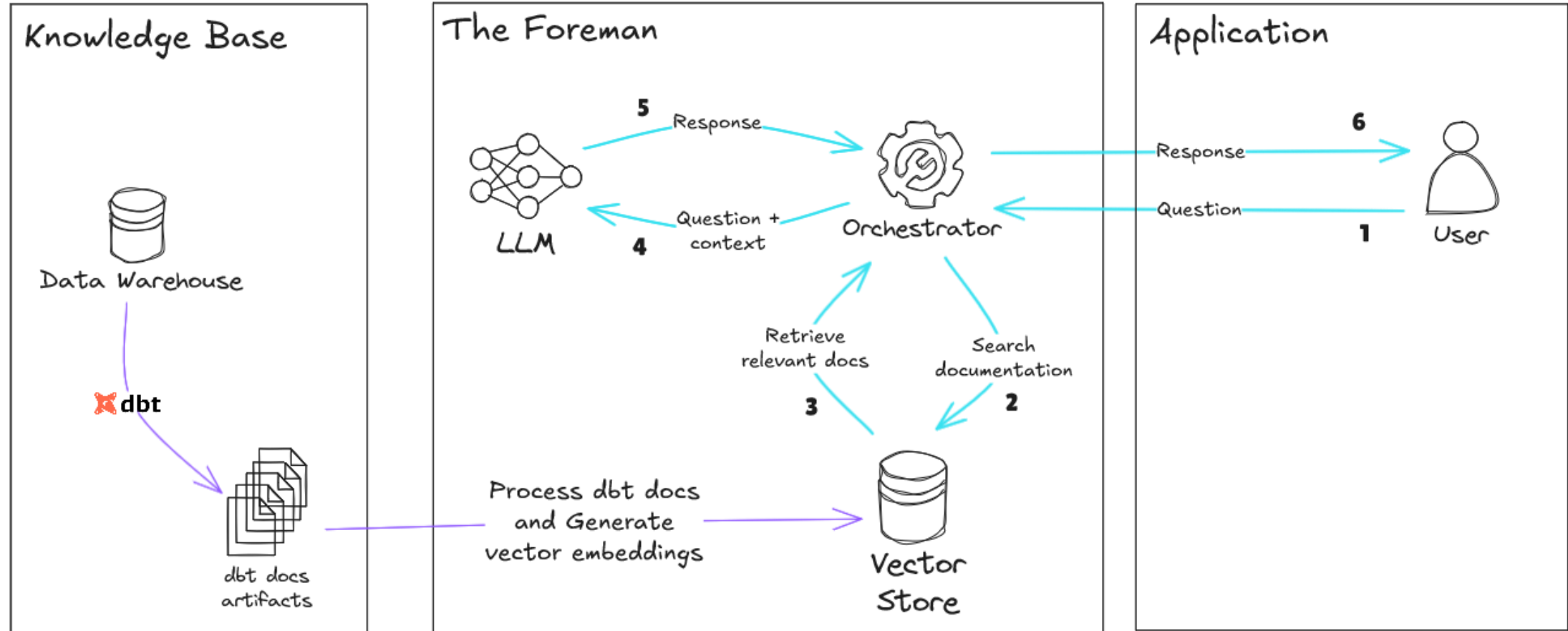


Retrieval Augmented Generation (RAG)

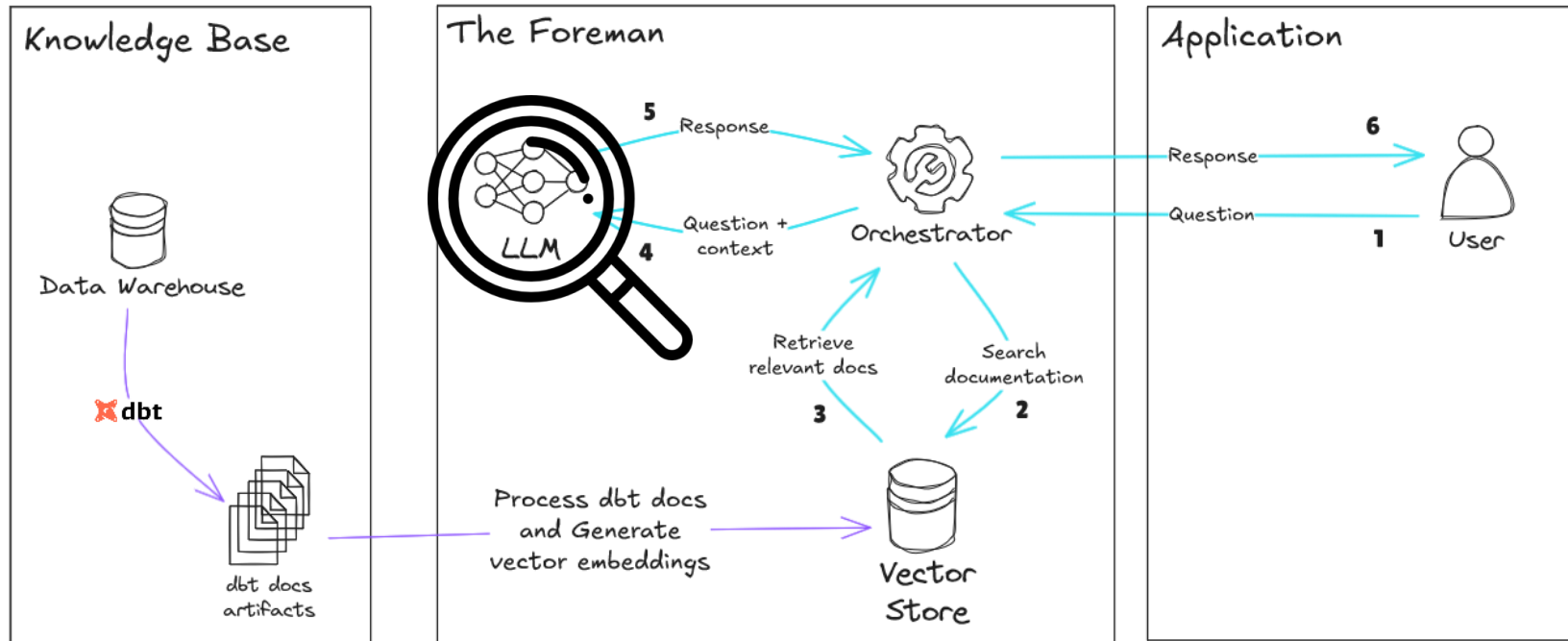
Prompt with RAG:



The 'Foreman' RAG



The 'Foreman' RAG



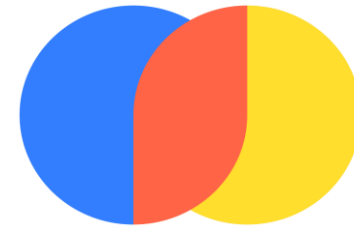
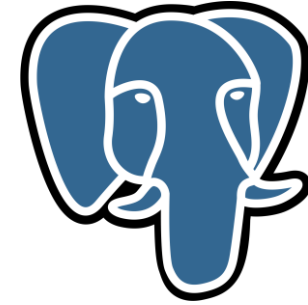
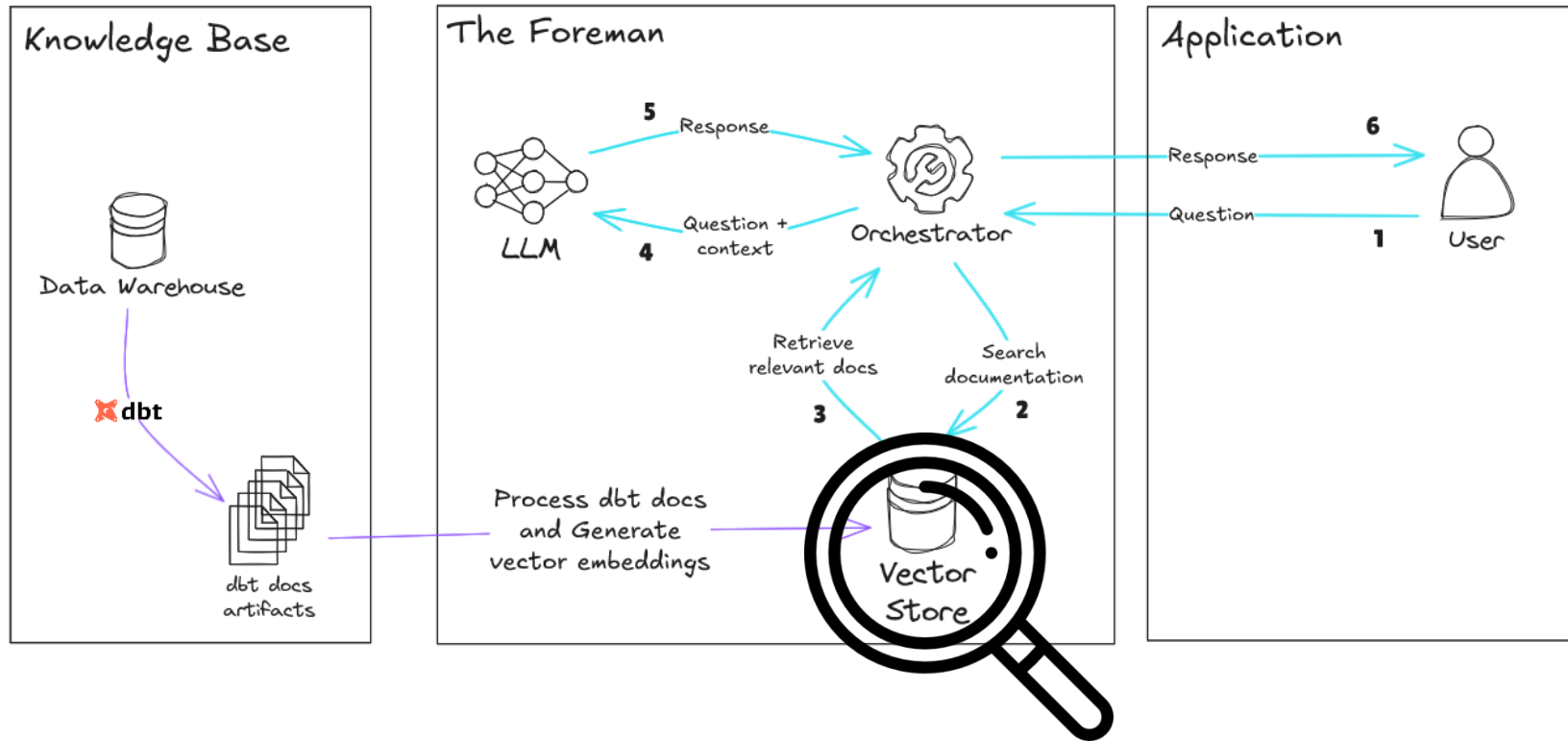
AI

Meta

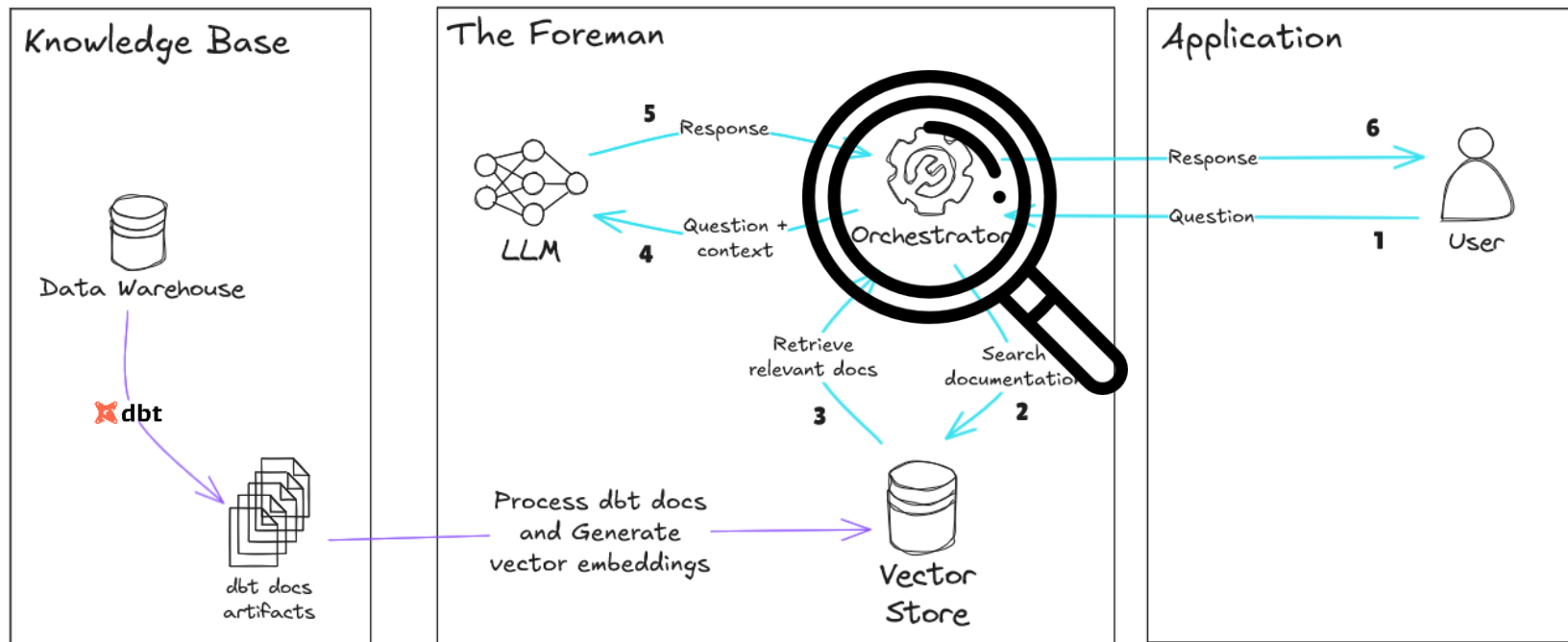


Gemini

The 'Foreman' RAG



The 'Foreman' RAG



Example: Data Makers Fest

Models:

- Attractions
- Partners
- Sessions
- Speakers
- Tutorials

```
# Name your project! Project names should contain only lowercase characters
# and underscores. A good package name should reflect your organization's
# name or the intended use of these models
name: 'warehouse'
version: '1.0.0'

# This setting configures which "profile" dbt uses for this project.
profile: 'default'

# These configurations specify where dbt should look for different types of files.
# The 'model-paths' config, for example, states that models in this project can be
# found in the "models/" directory. You probably won't need to change these!
model-paths: ["models"]
analysis-paths: ["analyses"]
test-paths: ["tests"]
seed-paths: ["seeds"]
macro-paths: ["macros"]
snapshot-paths: ["snapshots"]

clean-targets:           # directories to be removed by `dbt clean`
- "target"
- "dbt_packages"

# Configuring models
# Full documentation: https://docs.getdbt.com/docs/configuring-models

# In this example config, we tell dbt to build all models in the example/
# directory as views. These settings can be overridden in the individual model
# files using the `{% config(...) %}` macro.
models:
  analytics:
    # Config indicated by + and applies to all files under models/example/
    example:
      +materialized: view
```

dbt_project.yml

```
✓ warehouse
  > analyses
  > logs
  > macros
  ✓ models
    ✓ attractions
      attractions.md
      attractions.sql
      schema.yml
    ✓ partners
      partners.sql
      patners.md
      schema.yml
    ✓ sessions
      schema.yml
      sessions.md
      sessions.sql
    ✓ speakers
      schema.yml
      speakers.md
      speakers.sql
    ✓ tutorials
      schema.yml
      tutorials.md
      tutorials.sql
      overview.md
  > seeds
  > snapshots
  > target
  > tests
    .user.yml
    dbt_project.yml
    dbt.duckdb
    profiles.yml
    README.md
```

Folder structure

Questions with 'base' LLM

#1

Question: What is Data Makers Fest?

Answer: 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.

Questions with 'base' LLM

#1

Question: What is Data Makers Fest?

Answer: 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.

#2

Question: Where is it hosted?

Answer: 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.

Questions with 'base' LLM

#1

Question: What is Data Makers Fest?

Answer: 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.

#2

Question: Where is it hosted?

Answer: 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.

#3

Question: Tell me one of the cities where it has been hosted

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

Questions with 'base' LLM

#1

Question: What is Data Makers Fest?

Answer: 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.

#2

Question: Where is it hosted?

Answer: 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.

#3

Question: Tell me one of the cities where it has been hosted

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

#4

Question: What types of sessions does it have?

Answer: 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.

Questions with 'Foreman' using dbt docs

#1

Question: What is Data Makers Fest?

Answer: Data Makers Fest is a festival dedicated to all data practitioners. It spans 3 days and gathers data professionals and enthusiasts from various fields to learn, network, and explore hot topics in data and AI. The event features dynamic sessions on cutting-edge 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.

Questions with 'Foreman' using dbt docs

#1

Question: What is Data Makers Fest?

Answer: Data Makers Fest is a festival dedicated to all data practitioners. It spans 3 days and gathers data professionals and enthusiasts from various fields to learn, network, and explore hot topics in data and AI. The event features dynamic sessions on cutting-edge 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.

#2

Question: What types of sessions does it have?

Answer: Data Makers Fest features two types of sessions: short and long presentations. Additionally, there are Hands-On Tutorials, which are part of the same agenda but are tracked separately.

Questions with 'Foreman' using dbt docs

#1

Question: What is Data Makers Fest?

Answer: Data Makers Fest is a festival dedicated to all data practitioners. It spans 3 days and gathers data professionals and enthusiasts from various fields to learn, network, and explore hot topics in data and AI. The event features dynamic sessions on cutting-edge 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.

#2

Question: What types of sessions does it have?

Answer: Data Makers Fest features two types of sessions: short and long presentations. Additionally, there are Hands-On Tutorials, which are part of the same agenda but are tracked separately.

#3

Question: Where are the tutorials taking place?

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

‘Foreman’ with Text-to-SQL instructions

dbt model for the
‘Sessions’ table

```
warehouse > models > sessions > schema.yml > # version
qxz105a, yesterday | 1 author (qxz105a)
1 version: 2 qxz105a, 3 days ago • First draft of warehouse, dbt project
2
3 models:
4 - name: sessions
5   description: "{{ doc('sessions') }}"
6   columns:
7     - name: id
8       type: integer
9       description: "The identification number of the session"
10
11     - name: speaker_id
12       type: integer
13       description: "The primary key for this table"
14
15     - name: title
16       type: string
17       description: "The title of the session"
18
19     - name: summary
20       type: string
21       description: "Short text description of the session content and details"
22
23     - name: type
24       type: string
25       description: "The type of session"
26
27     - name: location
28       type: string
29       description: "The name of the room in the venue where the session takes place"
30
31     - name: date
32       type: date
33       description: "The date of the session in format YYYY-MM-DD"
34
35     - name: timeslot
36       type: string
37       description: "The time slot of the session in format HH:MM-HH:MM"
38
39     - name: duration
40       type: integer
41       description: "The duration, in minutes, of the session"
42
```

'Foreman' with Text-to-SQL instructions

#1

```
Question: How many sessions are there in total?
```

```
Answer:
```

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


'Foreman' with Text-to-SQL instructions

#1

Question: How many sessions are there in total?

Answer:

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

#2

Question: How many partners of each type are there?

Answer:

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

'Foreman' with Text-to-SQL instructions

#3

Question: Are there any speakers with more than one participation?

Answer:

```
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;
```

Road to production: Bitter lessons

“Once you get past demo datasets, the performance drops”

- Databricks: “Text2SQL is Not Enough”¹
- Snowflake: “Need for a semantic model”²
- Numbers Station AI (Text-to-SQL start-up): “Text-to-SQL Failures on Enterprise Data”³

1: arxiv.org/html/2408.14717v1

2: snowflake.com/engineering-blog/snowflake-cortex-analyst-behind-the-scenes/

3: numbersstation.ai/a-case-study-text-to-sql-failures-on-enterprise-data/

Embracing the Foreman

“Goods sit in the warehouse until information moves them.”

C. J. Cherryh (science-fiction writer)





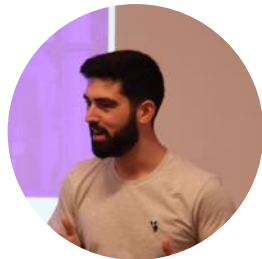
DATA
MAKERS
fest



luisferreira97



luisferreira97/foreman



joaocrebelo

Thank you!