Boosting LLM Accuracy with IdentityRAG

Hendrik Nehnes, Co-founder Tilores

https://www.linkedin.com/in/hendriknehnes/

Tilores

Gartner says

“More Than **80% of Enterprises** Will Have Used **Generative AI** APIs or Deployed Generative AI-Enabled Applications by **2026**”

but

Research from Forrester identifies **data privacy and security** concerns as the top barrier to generative AI adoption.

Sources:

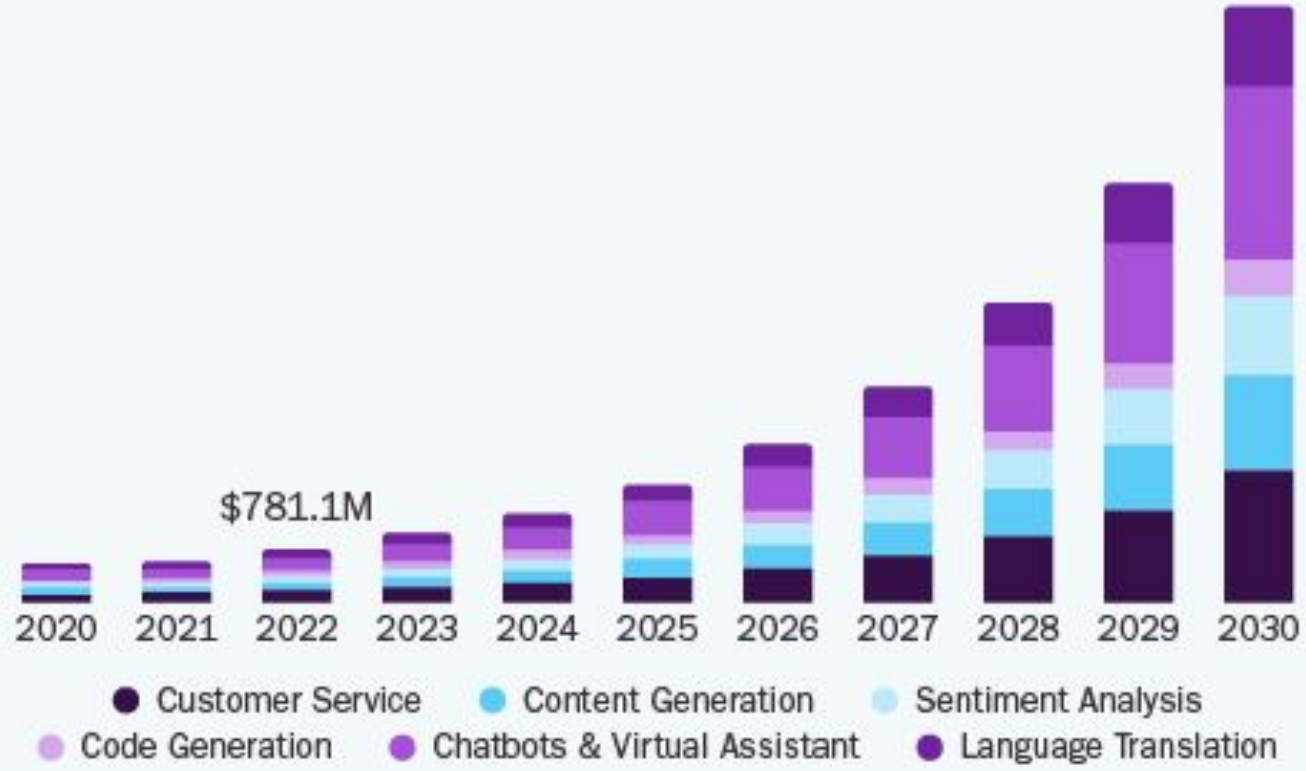
https://www.gartner.com/en/newsroom/press-releases/2023-10-11-gartner-says-more-than-80-percent-of-enterprises-will-have-used-generative-ai-apis-or-deployed-generative-ai-enabled-applications-by-2026

https://www.forrester.com/report/security-and-privacy-concerns-are-the-biggest-barriers-to-adopting/RES180179

Tilores

U.S. Large Language Model Market

Size, by Application, 2020 - 2030 (USD Million)



GRAND VIEW RESEARCH

37.2%

U.S. Market CAGR, 2024 - 2030

Source: www.grandviewresearch.com

https://www.grandviewresearch.com/static/img/research/us-large-language-model-market.png

Tilores

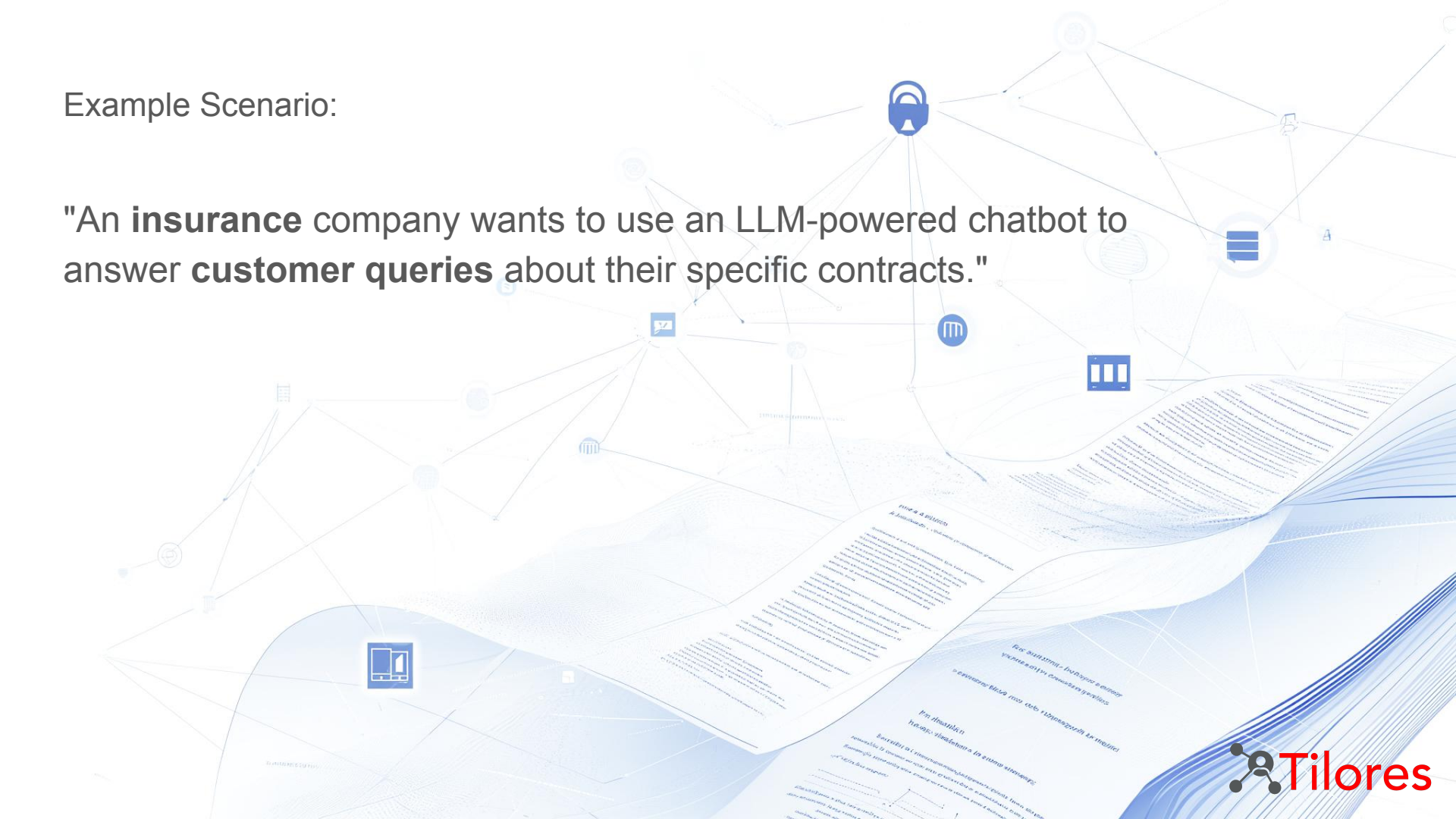
Agenda

- The Challenge: LLMs and Private Data
- Scattered Customer Data in Silos
- Entity Resolution
- Retrieval Augmented Generation (RAG)
- IdentityRAG
- Live Demo: IdentityRAG with Langchain
- Conclusion and Future Implications
- Q&A Session

Tilores

Example Scenario:

"An **insurance** company wants to use an LLM-powered chatbot to answer **customer queries** about their specific contracts."

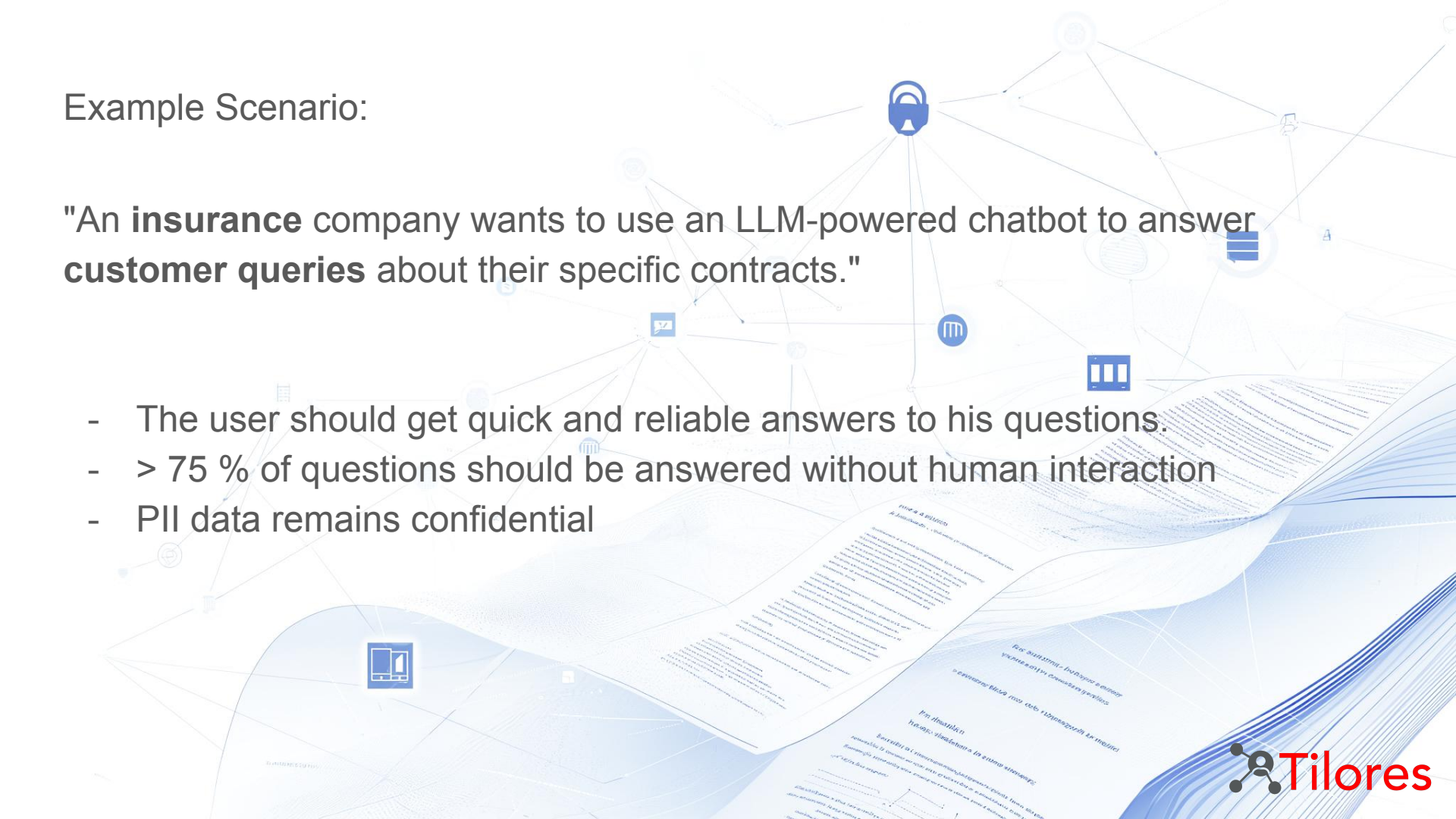


Tilores

Example Scenario:

"An **insurance** company wants to use an LLM-powered chatbot to answer **customer queries** about their specific contracts."

- The user should get quick and reliable answers to his questions.
- > 75 % of questions should be answered without human interaction
- PII data remains confidential



Tilores

LLMs and Private Data

Tilores

LLMs and Private Data

- Lack of specific domain knowledge (general purpose LLMs)

Tilores

LLMs and Private Data

- Lack of specific domain knowledge
- Outdated information

Tilores

LLMs and Private Data

- Lack of specific domain knowledge
- Outdated information
- No access to internal data sources

Tilores

LLMs and Private Data

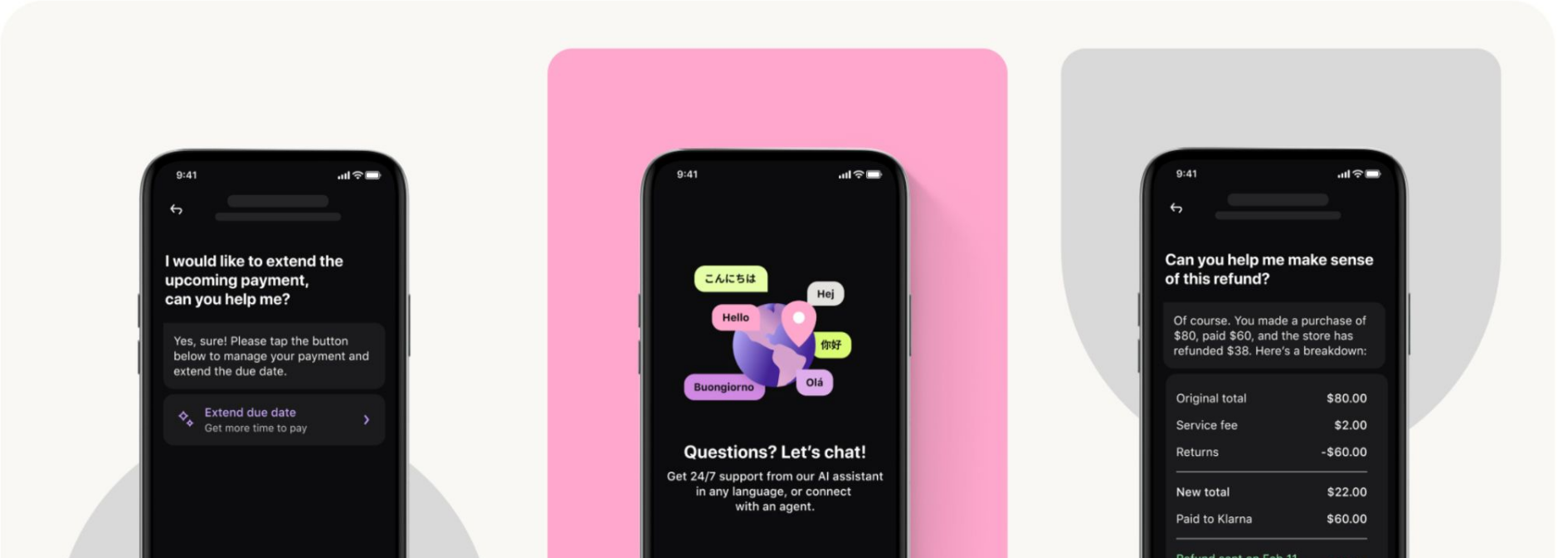
- Lack of specific domain knowledge
- Outdated information
- No access to internal data sources
- Privacy concerns

Tilores

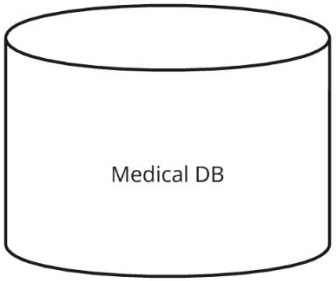
Klarna AI assistant handles two-thirds of customer service chats in its first month

February 27, 2024

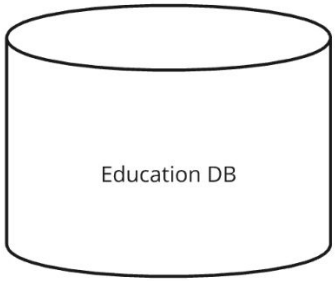
Klarna
Klarna



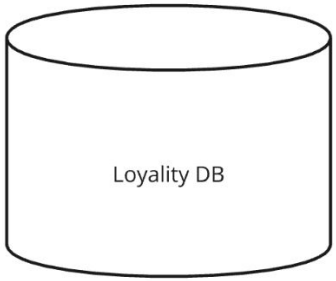
Problem: Scattered Customer Data in Different Silos



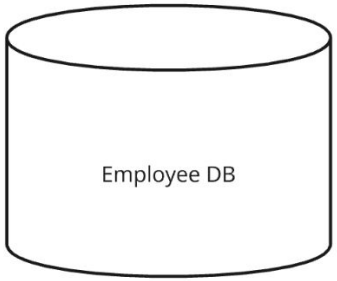
Alexander Thompson-Wilson
22/07/1985
789 Maple Road
M5V 2T6 Toronto



Alexander Wilson
1985-07-22
45 Pitt Street
2000 Sydney
wilson.alex@university.edu.au



Alex Wilson
22-07-1985
45 Pitt Street
2000 Sydney
0285559876



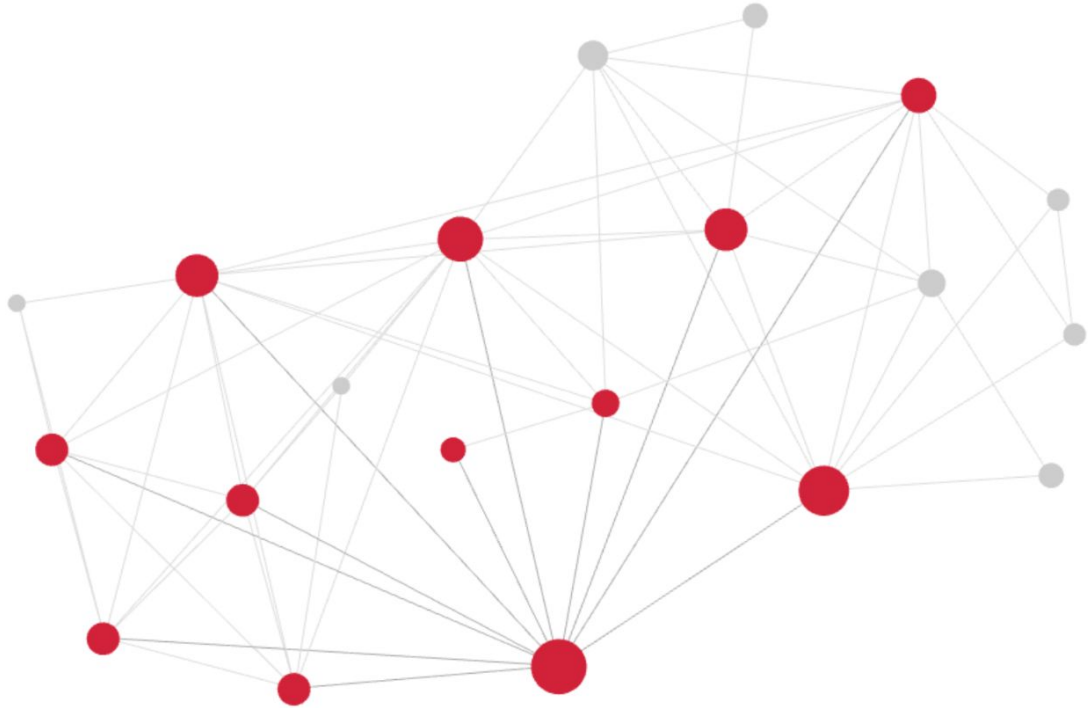
Alex Thompson
1985-07-22
123 Main Street
10001 New York
a.thompson@company.com
2125551234

Tilores

Problem: Scattered Customer Data in Different Silos

- **Fragmented View:** Incomplete customer profiles across systems

Tilores



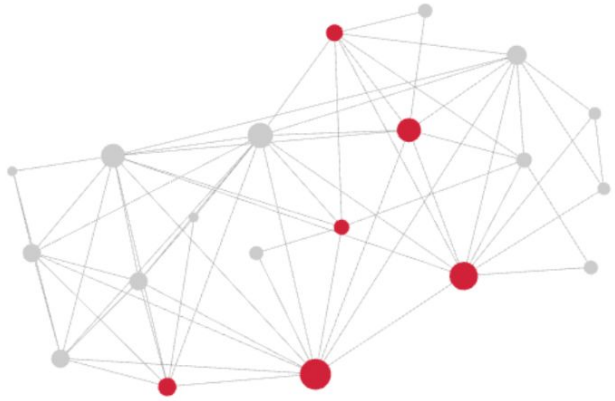
Id	Address Line	Birthday	City	Dob	Email	First Name	Last Name	Phone	Postal Code	Source
aa001001-0021-4000-a000-000000000021	23 George Street		Sydney	1985-07-22	alex.wilson@email.com.au	Alexander	Wilson	0285551234	2000	CUSTOMER
aa001001-0001-4000-a000-000000000001	123 Main St		New York	1985-07-22	alex.thompson@email.com	Alexander	Thompson	2125551234	10001	CUSTOMER

Tilores

Problem: Scattered Customer Data in Different Silos

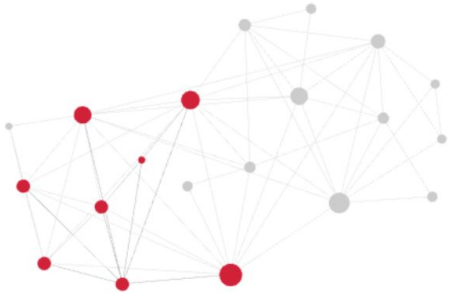
- **Fragmented View:** Incomplete customer profiles across systems
- **Inconsistent Information:** Conflicting data between departments

Tilores



Id	Address Line	Birthday	City	Dob	Email	First Name	Last Name	Phone	Postal Code	Source
aa001001-0021-4000-a000-000000000021	23 George Street		Sydney	1985-07-22	alex.wilson@email.com.au	Alexander	Wilson	0285551234	2000	CUSTOMER
aa001001-0016-4000-a000-000000000016	101 Queen Street		Toronto	1985-07-22		Alexander	Wilson		M5H 2M9	GOVERNMENT
aa001001-0001-4000-a000-000000000001	123 Main St		New York	1985-07-22	alex.thompson@email.com	Alexander	Thompson	2125551234	10001	CUSTOMER
aa001001-0022-4000-a000-000000000022	23 George St		Sydney	1985-07-22	a.wilson@company.com.au	Alex	Wilson	0285551234	2000	EMPLOYEE
aa001001-0011-4000-a000-000000000011	789 Maple Road		Toronto	1985-07-22	alex.thompsonwilson@email.ca	Alexander	Thompson-Wilson	4165551234	M5V 2T6	CUSTOMER
aa001001-0012-4000-a000-000000000012	789 Maple Rd		Toronto	1985-07-22	a.thompsonwilson@company.ca	Alex	Thompson-Wilson	4165551234	M5V 2T6	EMPLOYEE

This would be **> 10** customer profiles in most companies!



Id	Address Line	Birthday	City	Dob	Email	First Name	Last Name	Phone	Postal Code	Source
aa001001-0021-4000-a000-000000000021	23 George Street		Sydney	1985-07-22	alex.wilson@email.com.au	Alexander	Wilson	0285551234	2000	CUSTOMER
aa001001-0001-4000-a000-000000000001	123 Main St		New York	1985-07-22	alex.thompson@email.com	Alexander	Thompson	2125551234	10001	CUSTOMER
aa001001-0016-4000-a000-000000000016	101 Queen Street		Toronto	1985-07-22		Alexander	Wilson		M5H 2M9	GOVERNMENT
aa001001-0030-4000-a000-000000000030	45 Pitt Street		Sydney	1985-07-22	wilson.alex@university.edu.au	Alexander	Wilson		2000	EDUCATION
aa001001-0020-4000-a000-000000000020	101 Queen Street		Toronto	1985-07-22	wilson.alex@university.ca	Alexander	Wilson		M5H 2M9	EDUCATION
aa001001-0026-4000-a000-000000000026	45 Pitt Street		Sydney	1985-07-22		Alexander	Wilson		2000	GOVERNMENT
aa001001-0018-4000-a000-000000000018	101 Queen Street	22-07-1985	Toronto			Alex	Wilson	4165559876	M5H 2M9	LOYALTY
aa001001-0017-4000-a000-000000000017	101 Queen St		Toronto	1985-07-22	a.wilson@finance.ca	Alexander	Wilson	4165559876	M5H 2M9	FINANCIAL
aa001001-0002-4000-a000-000000000002	123 Main Street		New York	1985-07-22	a.thompson@company.com	Alex	Thompson	2125551234	10001	EMPLOYEE
aa001001-0006-4000-a000-000000000006	456 Oak Avenue		San Francisco	1985-07-22		Alexander	Thompson		94102	GOVERNMENT

Problem: Scattered Customer Data in Different Silos

- **Fragmented View:** Incomplete customer profiles across systems
- **Inconsistent Information:** Conflicting data between departments
- **Compliance Risks:** Difficulty in maintaining data accuracy and privacy

Tilores

Compliance Risks: Difficulty in maintaining data accuracy and privacy

Art. 15 GDPR

Right of access by the data subject

1. The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information:

   (a) the purposes of the processing;

   (b) the categories of personal data concerned;

   (c) the recipients or categories of recipient to whom the personal data have been or will be

Tilores

Compliance Risks: Difficulty in maintaining data accuracy and privacy

- **Data Subject Access Request**: Complete response within 30 days.
- **Data Deletion Request:** Data has to be deleted from all systems.

  (Upstream, downstream, SQL, Vector, Graph… DB)
- **Compliance Risks:** Difficulty in maintaining data accuracy and privacy

  (what happens, if someone used OpenAI?).
- **Costs:** If not properly implemented a lot of manpower is needed.

Tilores

Problem: Scattered Customer Data in Different Silos

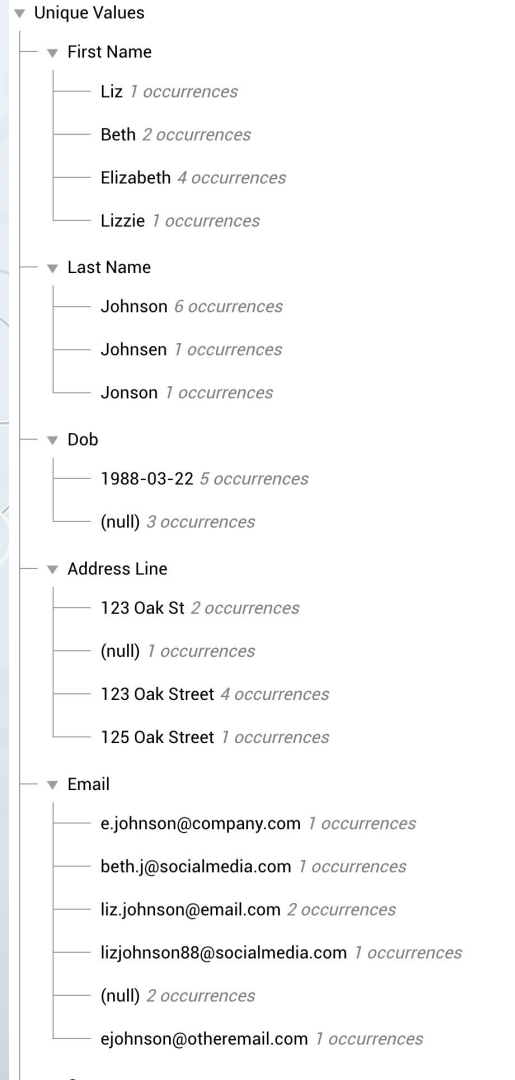
- **Fragmented View:** Incomplete customer profiles across systems
- **Inconsistent Information:** Conflicting data between departments
- **Compliance Risks:** Difficulty in maintaining data accuracy and privacy
- **Analytics Challenges:** Inaccurate insights due to data discrepancies

Tilores

Problem: Scattered Customer Data in Different Silos

- **Analytics Challenges:**

  Inaccurate insights due to data discrepancies



Problem: Scattered Customer Data in Different Silos

- **Fragmented View:** Incomplete customer profiles across systems
- **Inconsistent Information:** Conflicting data between departments
- **Compliance Risks:** Difficulty in maintaining data accuracy and privacy
- **Analytics Challenges:** Inaccurate insights due to data discrepancies

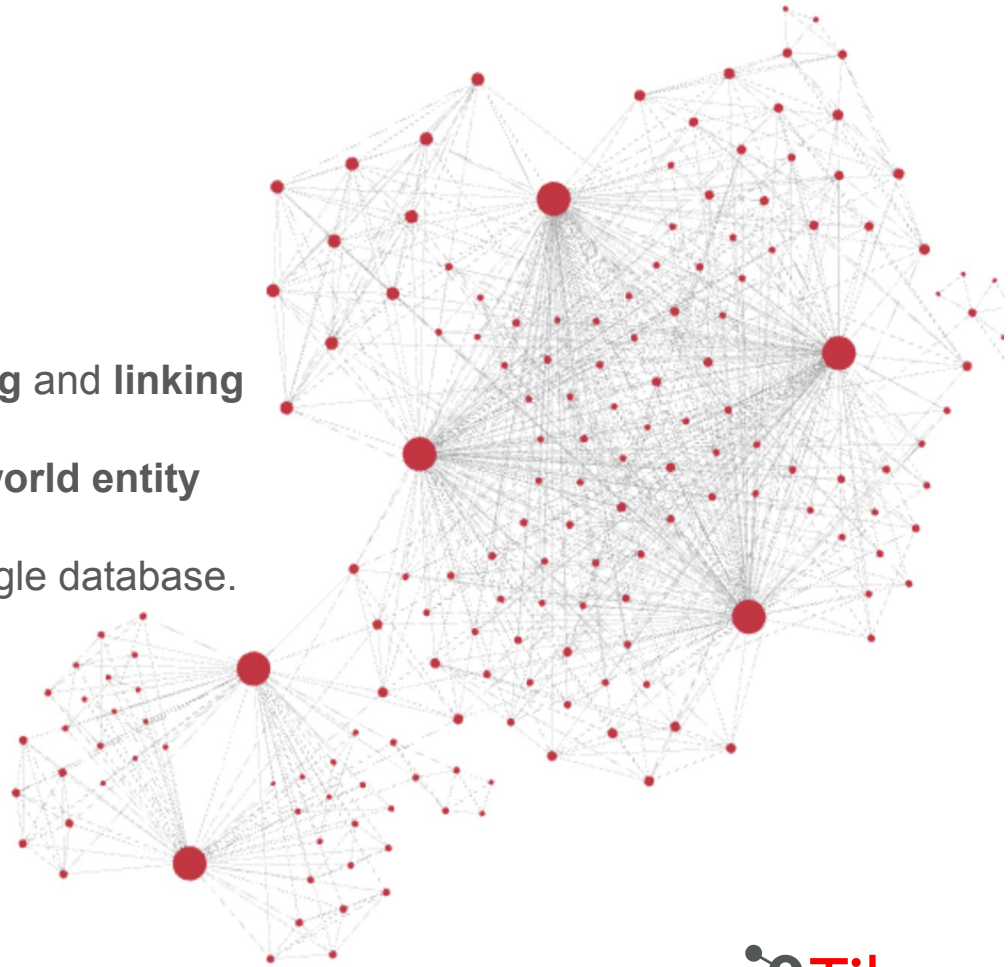
Solution?

Tilores


ENTITY
RESOLUTION
The solution for scattered customer data issues
Tilores

What is Entity Resolution

Entity resolution is the process of **identifying** and **linking** different representations of the **same real-world entity** across various **data sources** or within a single database.



https://medium.com/p/7ccf3af6b2f6

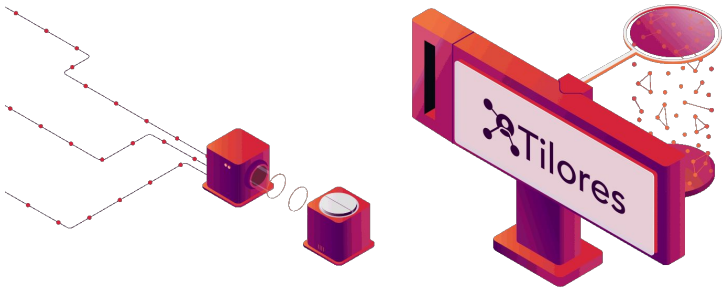
Tilores

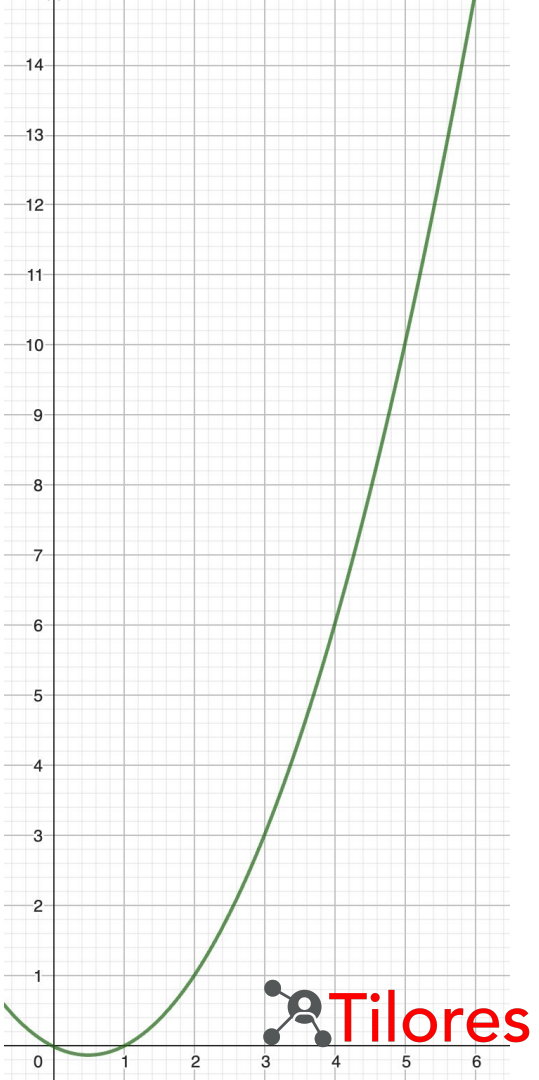
Why is this so complex?

n*(n-1)/2 possible unique pairs

1 mio records = 499.999.500.000 possible combinations

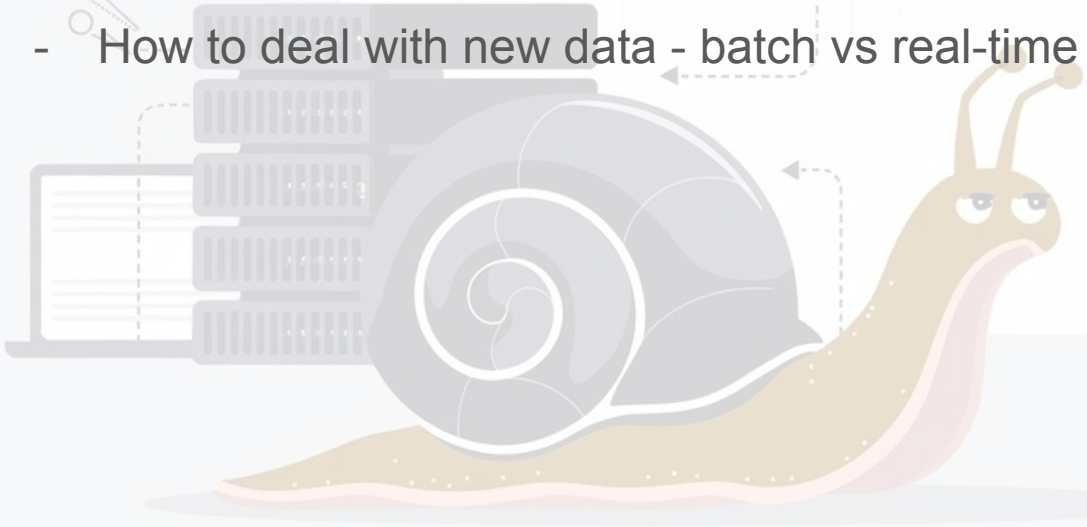
- Scale
- Speed
- Cost of Operations





Scale

- Number of combinations slows down the system
- Over 1 mio records most system have response times of several seconds
- How to deal with new data - batch vs real-time

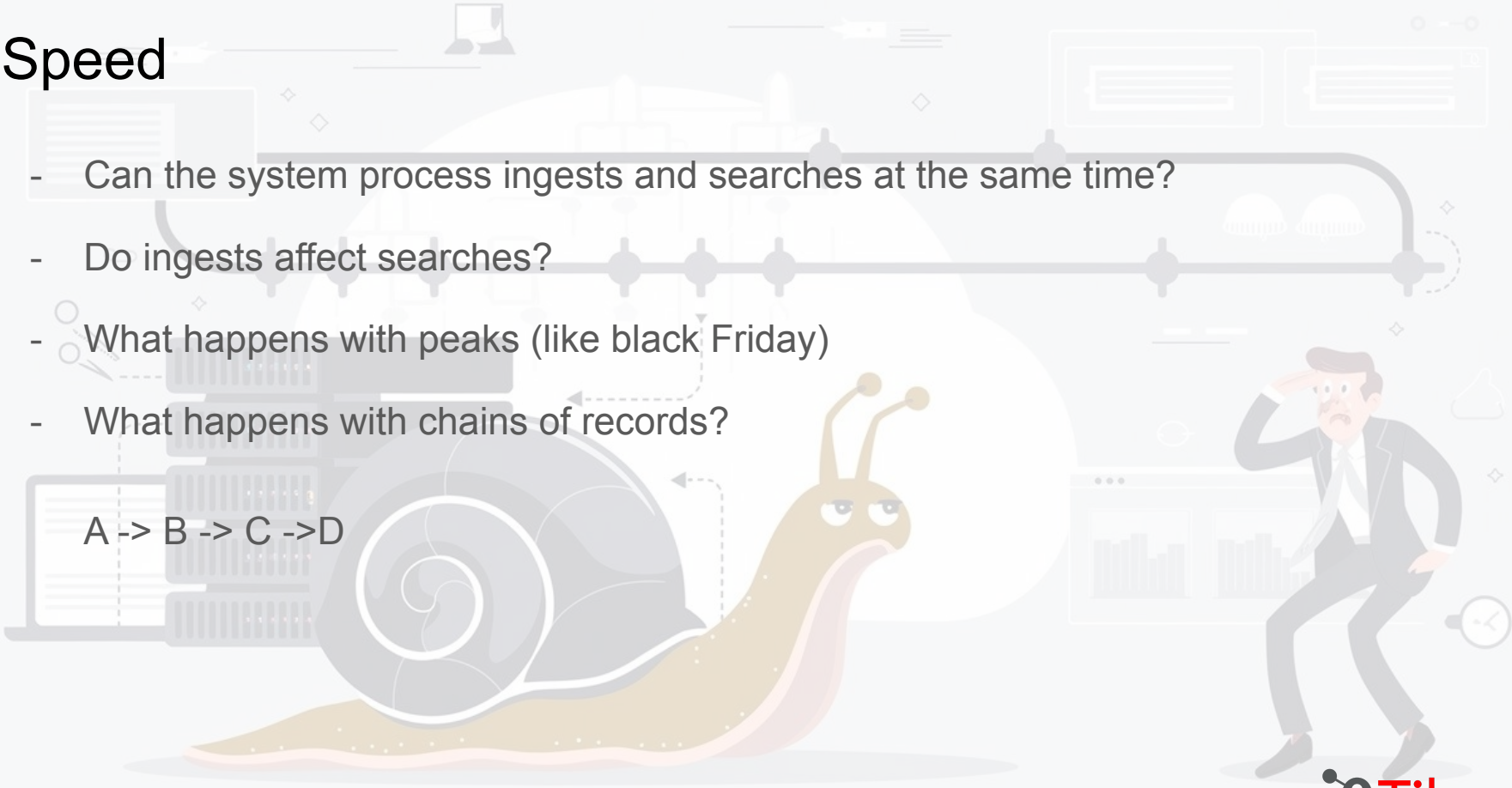


Tilores

Speed

- Can the system process ingests and searches at the same time?
- Do ingests affect searches?
- What happens with peaks (like black Friday)
- What happens with chains of records?

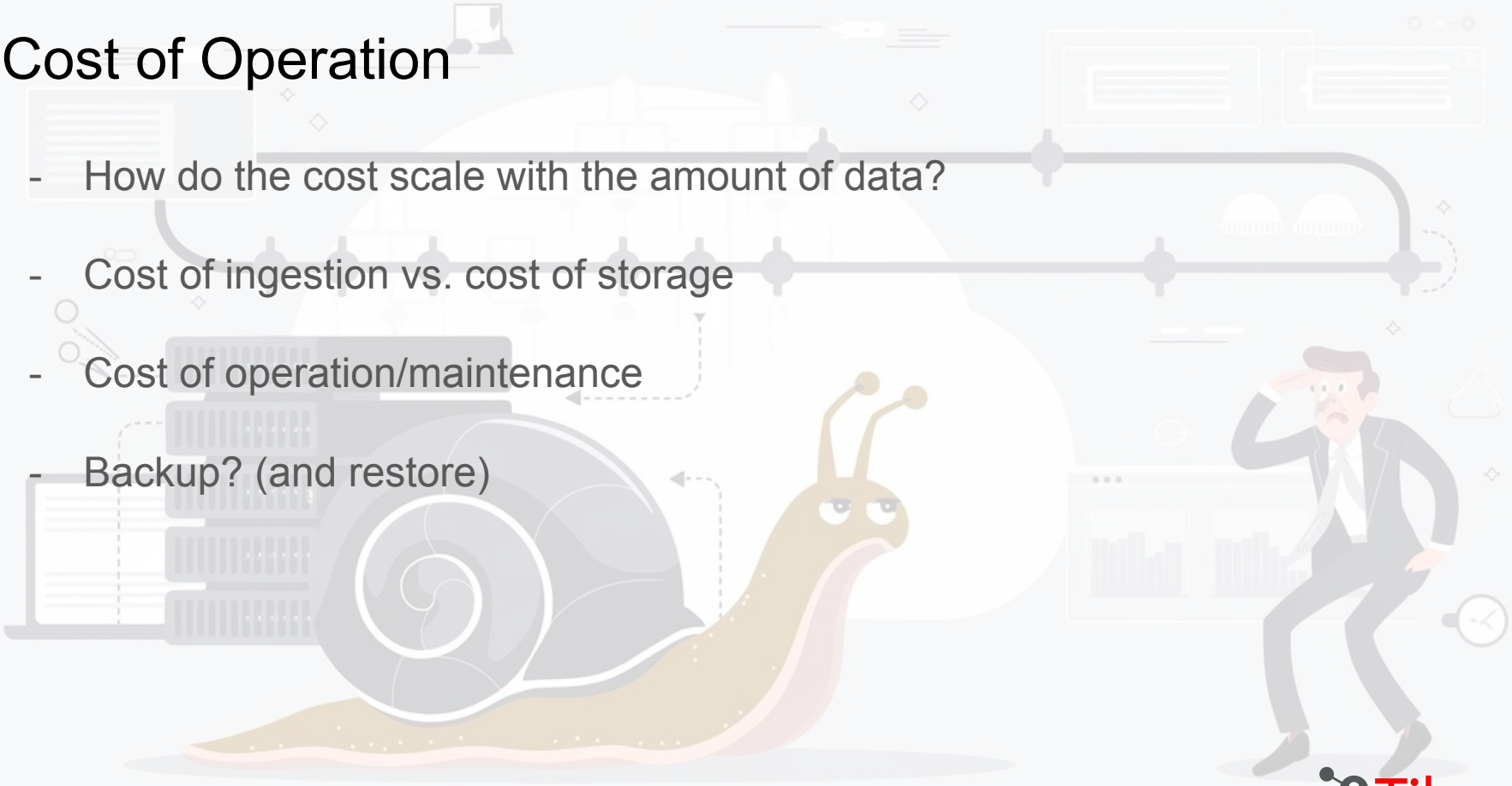
  A -> B -> C ->D



Tilores

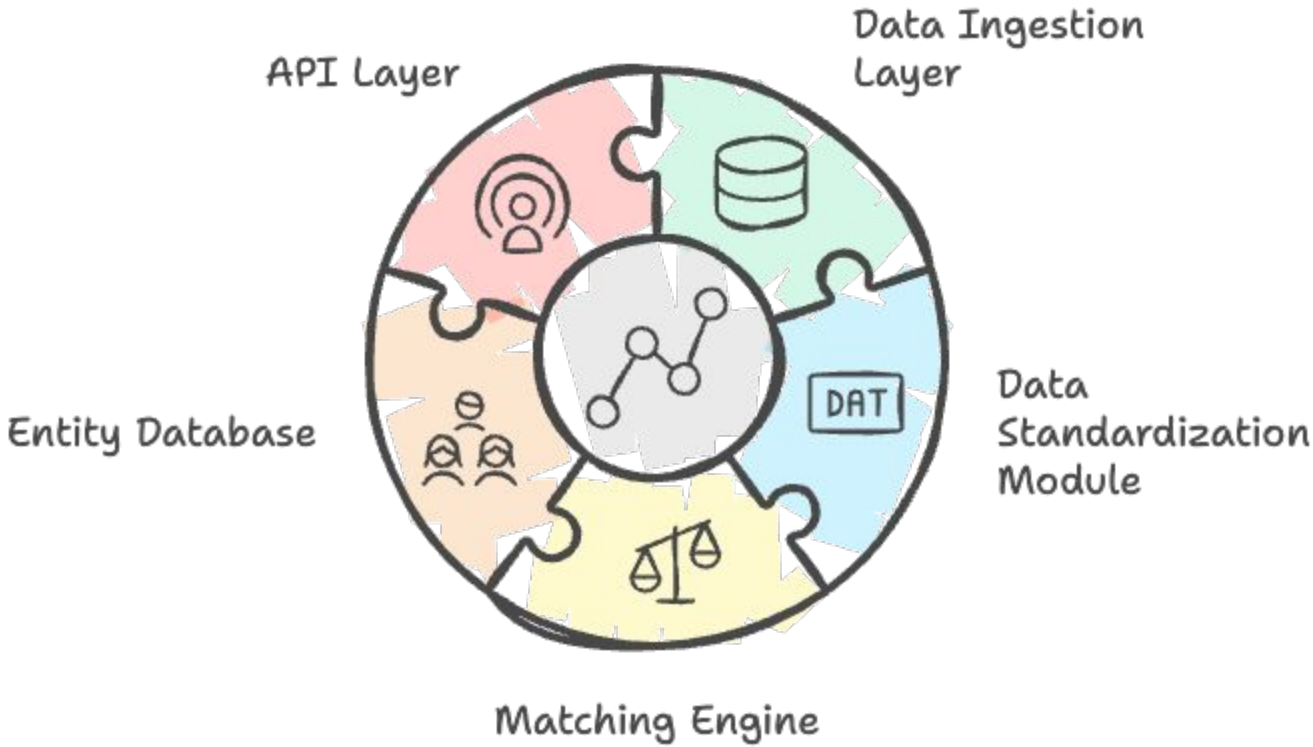
Cost of Operation

- How do the cost scale with the amount of data?
- Cost of ingestion vs. cost of storage
- Cost of operation/maintenance
- Backup? (and restore)



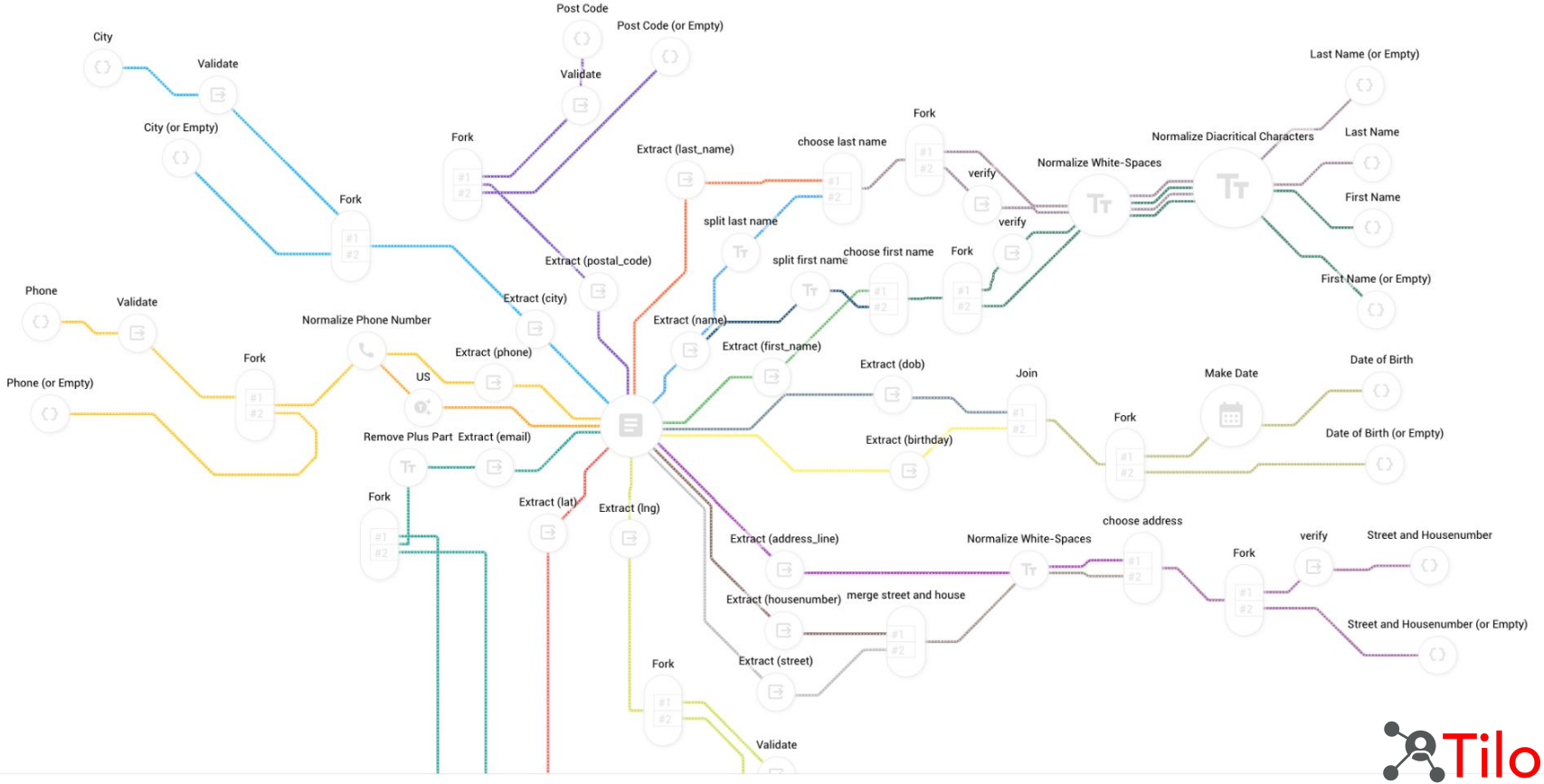
Tilores

Components of an Entity Resolution System



Tilores

Data Standardization/Normalization



Tilores

API Layer - GraphQL API

GraphQL

Operations schema

```
query {
  entity(input: { id: "07ac3e7d-23f6-40e3-8958-896178a6d52f" }) {
    entity {
      goldenRecord: recordInsights {
        firstName: frequencyDistribution(
          field: "first_name"
          top: 1
          direction: DESC
        ) {
          value
        }
        lastName: frequencyDistribution(
          field: "last_name"
          top: 1
          direction: DESC
        ) {
          value
        }
        address: newest(field: "receivedDate") {
          address_line
          postal_code
          city
          phone
        }
        emails: valuesDistinct(field: "email")
      }
    }
  }
```

Query Variables

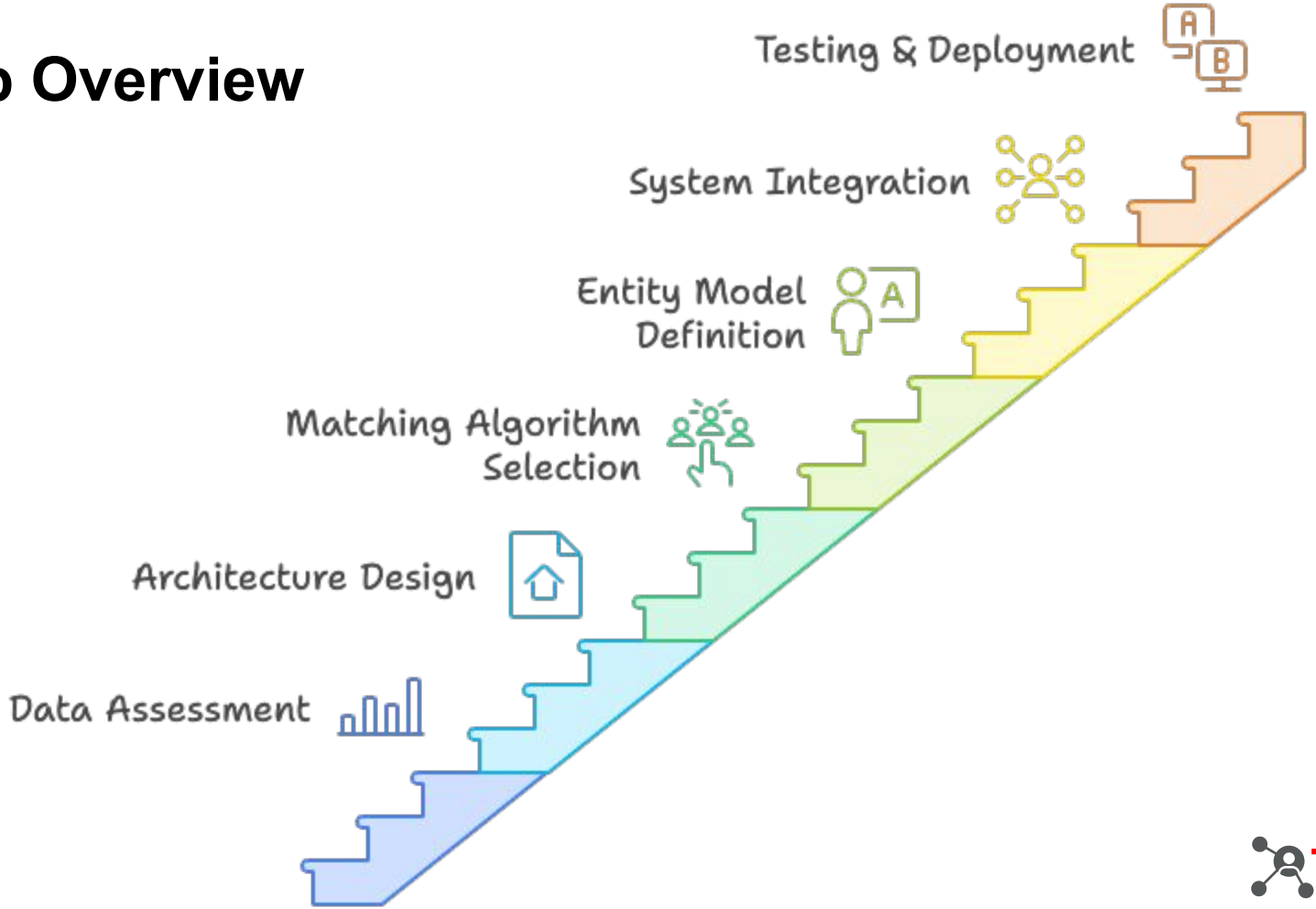
```
1
```

Preview


```
{
  "data": {
    "entity": {
      "entity": {
        "goldenRecord": {
          "firstName": [
            {
              "value": "Sophie"
            }
          ],
          "lastName": [
            {
              "value": "Müller"
            }
          ],
          "address": {
            "address_line": "Via del Corso 12",
            "postal_code": "00187",
            "city": "Roma",
            "phone": "0698765432"
          },
          "emails": [
            "mueller.sophie@uni-frankfurt.de",
            "s.muller@company.co.uk",
            "sophie.muller@email.co.uk",
            "mueller.sophie@uni-berlin.de",
            "sofi90@social.it",
            "sophie.muller@newemail.it",
            "muller.s@subservice.co.uk",
            "s.muller@company.it",
            "s.muller@newcompany.it",
            "s.muller@finance.co.uk",
            "sophie.mueller@email.de",
            "s.mueller@finance.de",
            "s.mueller@newcompany.de",
            "muller.s@subservice.it",
            "muller.s@newsubservice.it",
            "mueller.s@newsubservice.de",
```


Tilores

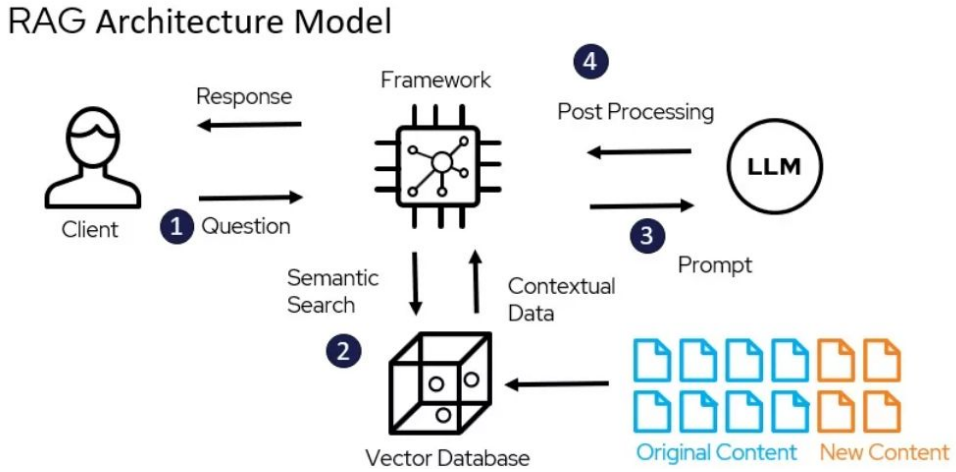
Setup Overview



Tilores

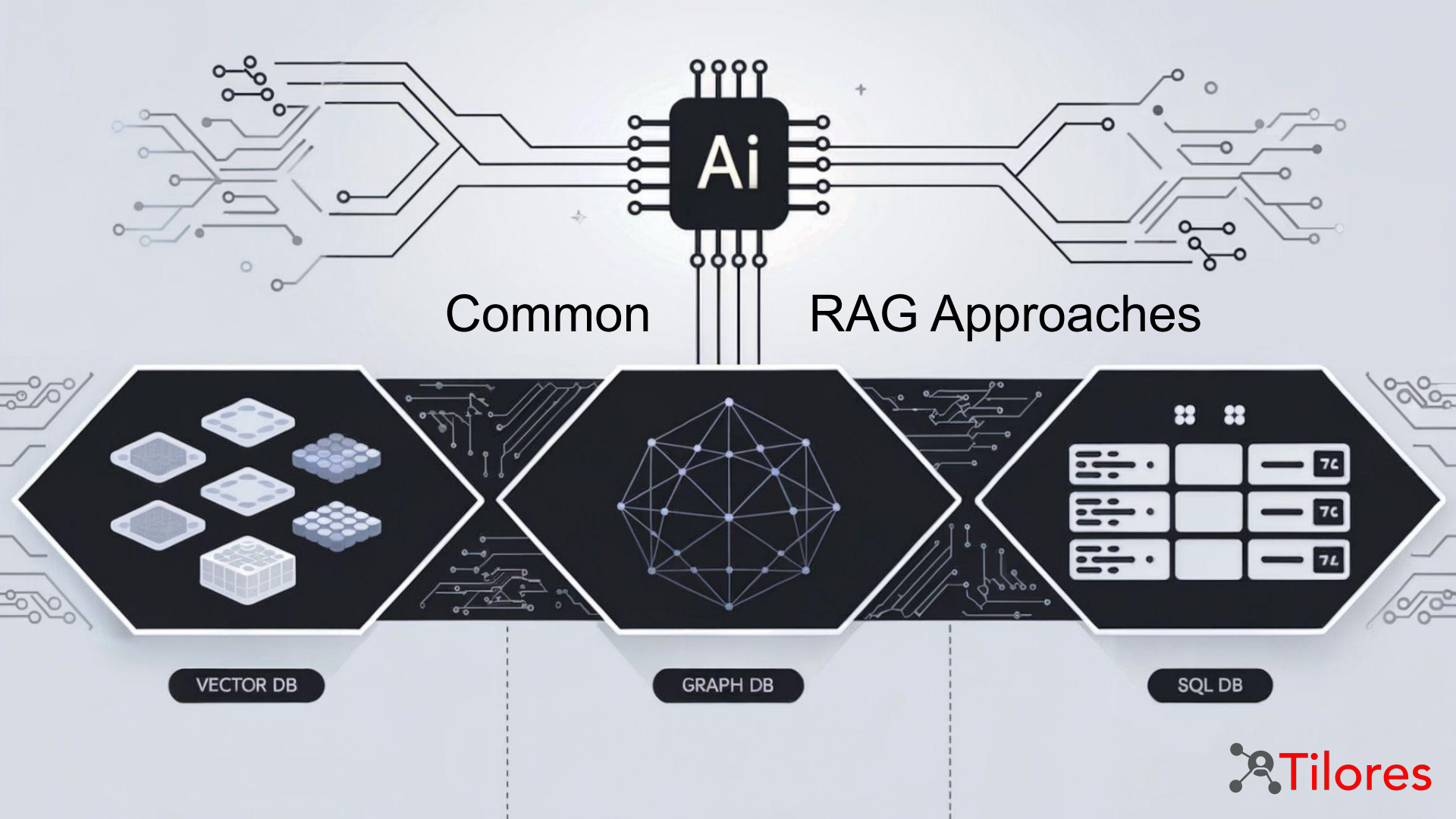
Retrieval Augmented Generation

- Definition: "RAG is a technique that combines the power of large language models with the ability to retrieve and use specific, up-to-date information."
- Key Components:
  - Large Language Model
  - Knowledge Base (e.g., policy documents, claim guidelines, customer data)
  - Retrieval System (often using vector databases)



https://www.linkedin.com/pulse/advanced-rag-techniques-overview-yugank-aman-t4kkf/

Tilores


Ai
Common
RAG Approaches
VECTOR DB
GRAPH DB
SQL DB
Tilores

Usual RAG DBs are not optimized for customer data

Customer data is complex

Each entity/person is made up from different attributes

- Name, Address, DoB, E-Mail, Phone, …

The data usually is not clean, contains typos and errors

Persons move, get married, etc. thus attributes change

New data is constantly added and can change the edges between records

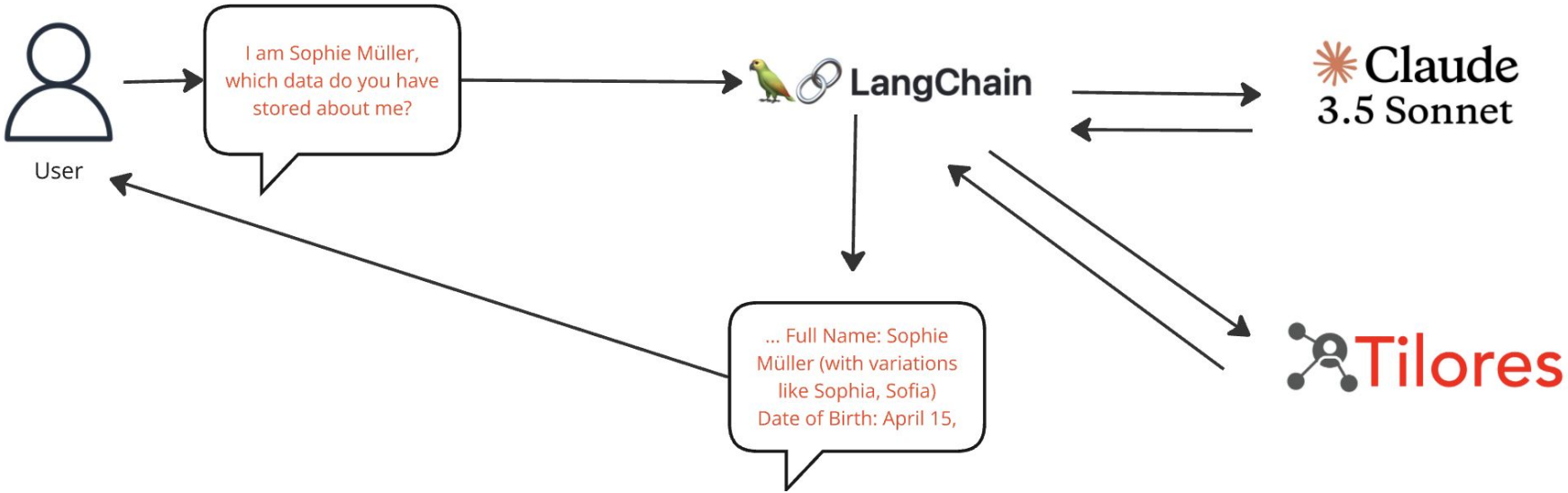
Tilores

IdentityRAG

- Is an advanced RAG system that integrates real-time entity resolution to provide LLMs with accurate, unified data about specific entities like people or companies.
- Uses the single source of truth for all PII data - the entity resolution system.
- Helps your system to stay GDPR compliant.
- Can be combined with Vector RAG and GraphRAG using Agentic RAG.

Tilores

Live Demo: IdentityRAG with Langchain



Tilores

in Chat Demo Reado...

nboard
ch Entities
ytics
st Data
age Instance
ta Schema
eate Rule
atching
Rules
Matchers
Extract / Transform
Tokenization
Text Comparison
tegration
GraphQL API
Snowflake
Webhook
dvanced Settings
umentation

hendrik.nehnes@tilotech

Search In Natural Language | By Attributes | Via File

Search Parameters

These are the search parameters according to your data schema. For fields that are not required, you should still provide the data that complies with your search rules. For complex data we recommend you to switch from the visual editor to the JSON editor.

VISUAL EDITOR | JSON EDITOR

First Name
Sophie

Last Name
Müller

+ NAME

Dob
1990-04-15

+ BIRTHDAY
+ ADDRESS LINE
+ STREET
+ HOUSENUMBER
+ POSTAL CODE
+ ZIP
+ CITY
+ PHONE
+ EMAIL
+ LAT
+ LNG
+ RECEIVED DATE
+ CONTRACT URL
+ MARKETING PDF

Conclusion and Future Implications

- Identity RAG empowers an LLM with private information
- Agentic RAG helps the LLM to understand which tool to use
- Based on the used LLM the outcomes vary (Claude vs ChatGPT)
- Data is easily accessible for every user without the need to understand an API
- The combination of LLMs and specialized systems will make data from different sources accessible for every user
- IdentityRAG creates a unified view across different systems

Tilores

Best Practices

List of key considerations and best practices

- Use the right technology for each use case
- Use the best model for the use case
- Combine different technologies for best results
- Stay on top of your data. Data compliance is key when dealing with PII data.

Tips for successful implementation

- Entity Resolution is a challenging discipline. Use professional vendors instead of struggling with scaling and operations of such a system.

Tilores


QUESTIONS?

Hendrik Nehnes, Co-founder Tilores

https://www.linkedin.com/in/hendriknehnes/

https://medium.com/@hendrik.nehnes

Producthunt launch tomorrow

Tilores