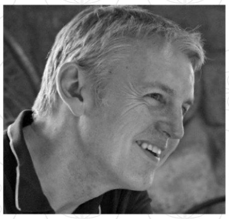
Generative AI in Data Management and Analytics
– A New Era of Assistance, Productivity and Automation

Mike Ferguson
CEO, Intelligent Business Strategies

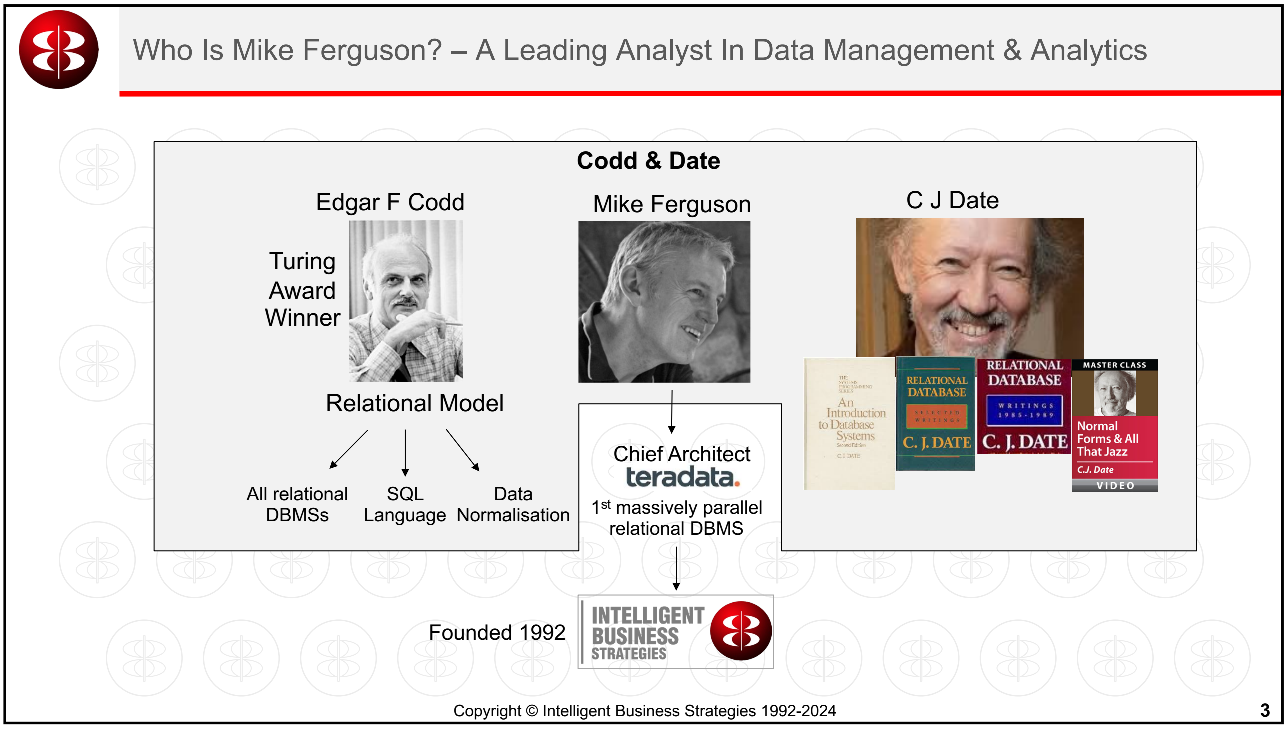
Adept Events Data Warehousing & BI Summit
Utrecht, March 2024



@mikeferguson1

Copyright © 1992-2024 Intelligent Business Strategies Limited. All rights reserved.

All materials pertaining to this presentation are the intellectual property and copyright of Intelligent Business Strategies Ltd and cannot be reproduced or distributed in any form without prior written permission from Intelligent Business Strategies Ltd. In addition, they must not be used by any party for hire and/or to compete against Intelligent Business Strategies Ltd



Who Is Mike Ferguson? – A Leading Analyst In Data Management & Analytics

Codd & Date

Edgar F Codd

Turing Award Winner

Relational Model

All relational DBMSs

SQL Language

Data Normalisation

Mike Ferguson

Chief Architect teradata.

1st massively parallel relational DBMS

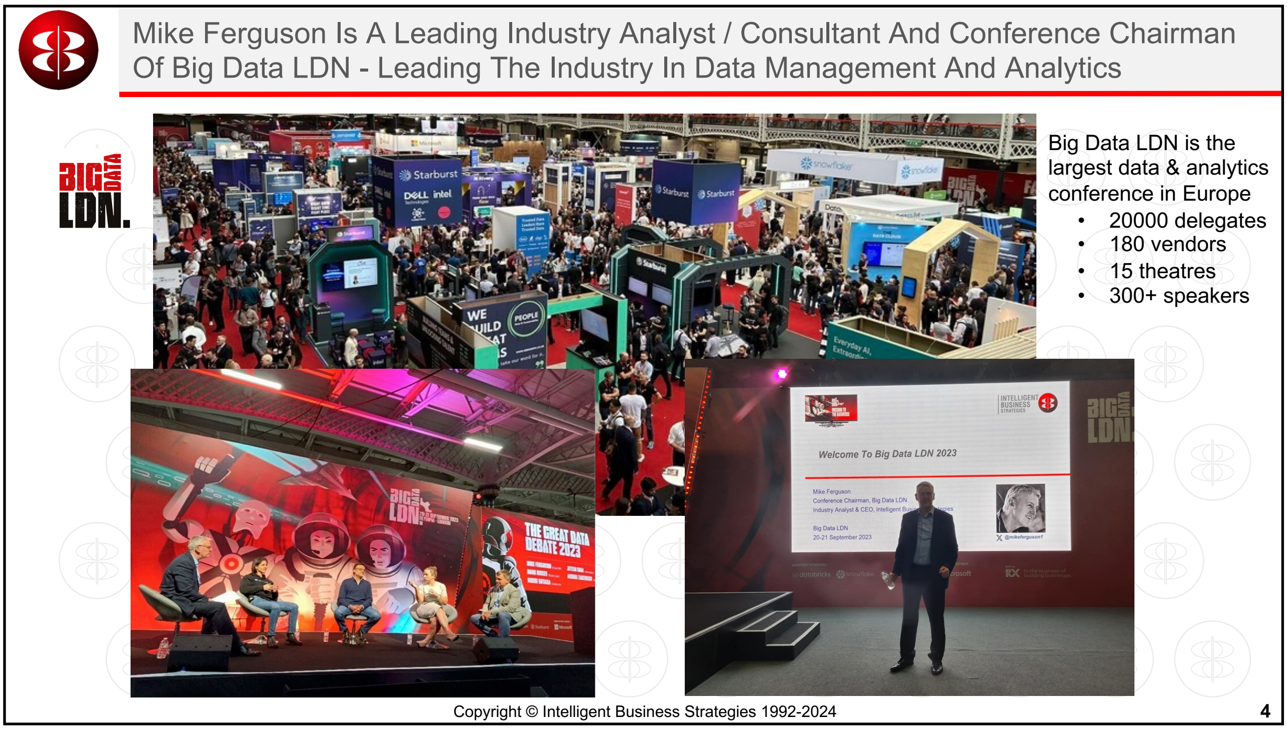
Founded 1992

INTELLIGENT BUSINESS STRATEGIES

C J Date

An Introduction to Database Systems
C.J. DATE
RELATIONAL DATABASE SELECTED WRITINGS
C. J. DATE
RELATIONAL DATABASE WRITINGS 1985-1989
C. J. DATE
MASTER CLASS
Normal Forms & All That Jazz
C.J. Date
VIDEO

Copyright © Intelligent Business Strategies 1992-2024



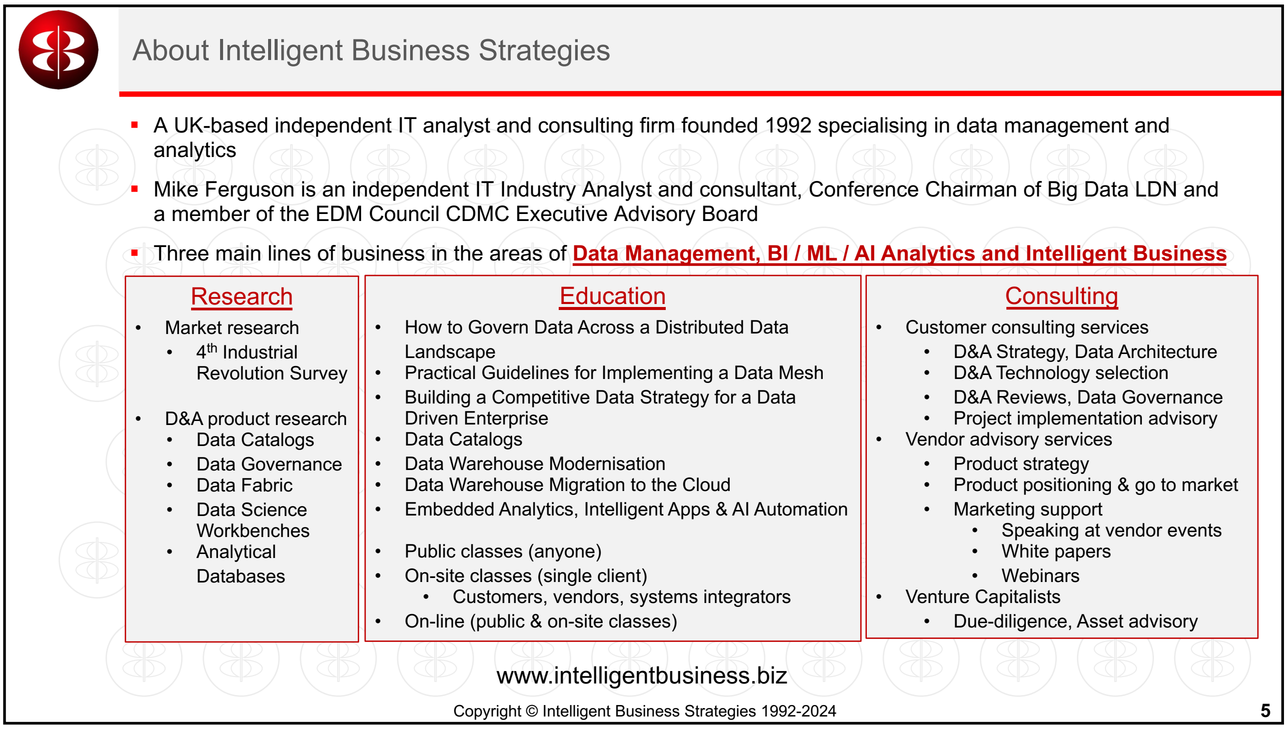
Mike Ferguson Is A Leading Industry Analyst / Consultant And Conference Chairman Of Big Data LDN - Leading The Industry In Data Management And Analytics

Big Data LDN is the largest data & analytics conference in Europe

- 20000 delegates
- 180 vendors
- 15 theatres
- 300+ speakers

BIG DATA LDN.
Starburst
DELL Technologies intel
snowflake
THE GREAT DATA DEBATE 2023
Welcome To Big Data LDN 2023
Mike Ferguson
Conference Chairman, Big Data LDN
Big Data LDN
20-21 September 2023

Copyright © Intelligent Business Strategies 1992-2024



About Intelligent Business Strategies

- A UK-based independent IT analyst and consulting firm founded 1992 specialising in data management and analytics
- Mike Ferguson is an independent IT Industry Analyst and consultant, Conference Chairman of Big Data LDN and a member of the EDM Council CDMC Executive Advisory Board
- Three main lines of business in the areas of **Data Management, BI / ML / AI Analytics and Intelligent Business**

Research

- Market research
    - 4th Industrial Revolution Survey
- D&A product research
    - Data Catalogs
    - Data Governance
    - Data Fabric
    - Data Science Workbenches
    - Analytical Databases

Education

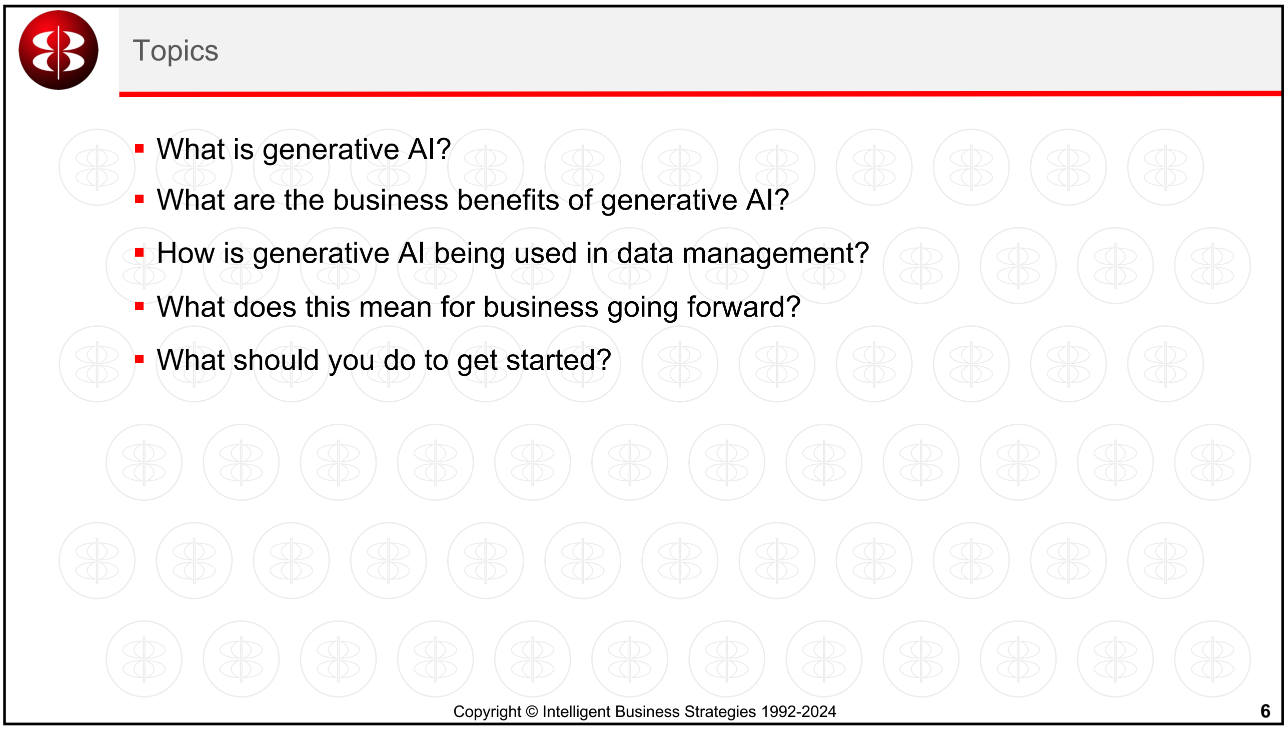
- How to Govern Data Across a Distributed Data Landscape
- Practical Guidelines for Implementing a Data Mesh
- Building a Competitive Data Strategy for a Data Driven Enterprise
- Data Catalogs
- Data Warehouse Modernisation
- Data Warehouse Migration to the Cloud
- Embedded Analytics, Intelligent Apps & AI Automation
- Public classes (anyone)
- On-site classes (single client)
    - Customers, vendors, systems integrators
- On-line (public & on-site classes)

Consulting

- Customer consulting services
    - D&A Strategy, Data Architecture
    - D&A Technology selection
    - D&A Reviews, Data Governance
    - Project implementation advisory
- Vendor advisory services
    - Product strategy
    - Product positioning & go to market
    - Marketing support
        - Speaking at vendor events
        - White papers
        - Webinars
- Venture Capitalists
    - Due-diligence, Asset advisory

www.intelligentbusiness.biz

Copyright © Intelligent Business Strategies 1992-2024



Topics

- What is generative AI?
- What are the business benefits of generative AI?
- How is generative AI being used in data management?
- What does this mean for business going forward?
- What should you do to get started?

Copyright © Intelligent Business Strategies 1992-2024

Copyright © Intelligent Business Strategies 1992-2024

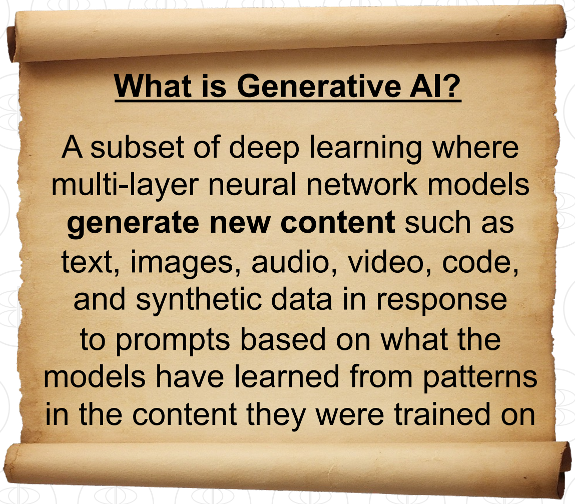
Topics – Where Are We?

- ➢ What is generative AI?
- What are the business benefits of generative AI?
- How is generative AI being used in data management?
- What does this mean for business going forward?
- What should you do to get started?

Top Ten Key Trends In Data Management And Analytics (D&A)

1. Generative AI – innovation in every area of business
2. Hybrid multi-cloud computing is now the norm
3. Architecture modernisation - integration of data warehouses, lakehouses, data lakes & streaming
4. Rationalisation towards a common data & analytics software stack for development & governance
   - Do more with less, integration across tools, accelerated development and shared metadata
5. FinOps - CFOs demand visibility of the full cost of the D&A ecosystem with consumption-based pricing
6. Data governance remains very high priority with AI governance now also on the agenda
   - More sources, distributed data complexity, data quality, security, privacy, usage, observability, sharing, retention
   - Poor data culture and weak data governance are barriers to decentralised development
7. Increasing demand for lower and lower latency data
8. Democratisation and acceleration of data and analytics development
   - E.g., Data Mesh, citizen data engineering, DataOps, autoML, MLOps, CI/CD, code Vs low code / no code
9. Compliant sharing and reuse of data and analytical products in data marketplaces / exchanges
10. Growth in intelligent applications, decision intelligence and AI-Automation

What Is Generative AI?



What is Generative AI?

A subset of deep learning where multi-layer neural network models **generate new content** such as text, images, audio, video, code, and synthetic data in response to prompts based on what the models have learned from patterns in the content they were trained on

- General use cases
  - Text generation
  - Virtual assistant – chat
  - Conversational search
  - Summarisation – text extraction
  - Code generation
  - Synthetic data generation
  - Image generation and classification
  - Video generation
- Benefits
  - Improve customer and employee experience
  - Productivity, automation
  - Ease of use, lower skills bar
  - Democratisation of D&A development

Copyright © Intelligent Business Strategies 1992-2024

Generative AI – What Are Large Language Models (LLMs)?

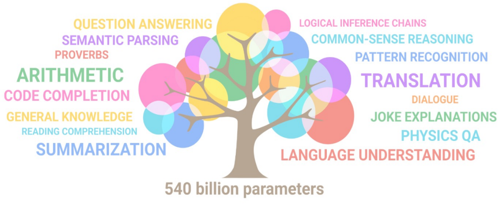
- Neural network models with billions of connections
- Pre-trained using unsupervised learning on very large bodies of public content and code to:
  - Predict the next token / word and generate it
  - Append the token to content generated to date

repeat

- Foundation LLM examples

Vendor	Foundation Models
Amazon	Titan & CodeWhisperer, Bloom
Databricks	Dolly
Google	Codey, LaMDA & PaLM2 • Gemini is the chat interface to PaLM2
Open AI / Microsoft	Generative Pre-Trained Transformer (GPT4) • ChatGPT is a chat interface to GPT

Google Pathways Language Model (PaLM2)



As the scale of the model increases, the performance improves across tasks while also unlocking new capabilities.

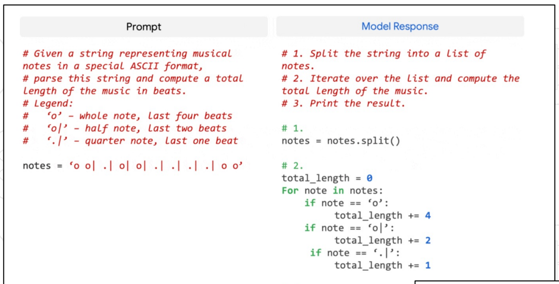
Image source: Google

- Foundation LLMs can be fine-tuned for specific tasks
  - Using supervised learning on smaller labelled data

Copyright © Intelligent Business Strategies 1992-2024

Copyright © Intelligent Business Strategies 1992-2024

Generative AI Example – Code Generation Using LLMs, e.g., PaLM



Prompt

```
# Given a string representing musical
notes in a special ASCII format,
# parse this string and compute a total
length of the music in beats.
# Legend:
#   'o' - whole note, last four beats
#   'o|' - half note, last two beats
#   '.|' - quarter note, last one beat

notes = 'o o| .| o| o| .| .| .| .| o o'
```

Model Response

```
# 1. Split the string into a list of
notes.
# 2. Iterate over the list and compute the
total length of the music.
# 3. Print the result.

# 1.
notes = notes.split()

# 2.
total_length = 0
For note in notes:
    if note == 'o':
        total_length += 4
    if note == 'o|':
        total_length += 2
     if note == '.|':
        total_length += 1

# 3.
print(total_length)
```

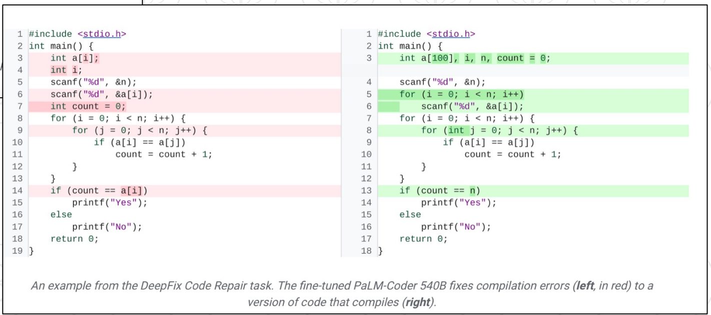
Examples of a fine-tuned PaLM 540B model on text-to-code tasks, such as ***GSM8K-***

code-to-code tasks, such as ***Transcoder****.*

Image source: Google

<ins>Text-to-code</ins>
Automatically generate code from a prompt

PaLM-Coder DeepFix Code Repair
(trained on Python)
Auto debug code so it compiles successfully



```
#include <stdio.h>
int main() {
    int a[i];
    int i;
    scanf("%d", &n);
    scanf("%d", &a[i]);
    int count = 0;
    for (i = 0; i < n; i++) {
        for (j = 0; j < n; j++) {
            if (a[i] == a[j])
                count = count + 1;
        }
    }
    if (count == a[i])
        printf("Yes");
    else
        printf("No");
    return 0;
}
```

```
#include <stdio.h>
int main() {
    int a[100], i, n, count = 0;

    scanf("%d", &n);
    for (i = 0; i < n; i++)
        scanf("%d", &a[i]);
    for (i = 0; i < n; i++) {
        for (int j = 0; j < n; j++) {
            if (a[i] == a[j])
                count = count + 1;
        }
    }
    if (count == n)
        printf("Yes");
    else
        printf("No");
    return 0;
}
```

*An example from the DeepFix Code Repair task. The fine-tuned PaLM-Coder 540B fixes compilation errors (**left**, in red) to a version of code that compiles (**right**).*

Image source: Google

Copyright © Intelligent Business Strategies 1992-2024

Generative AI Concerns / Issues

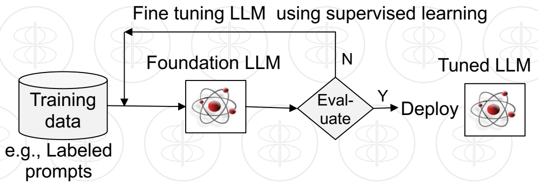
Concern/Issue	More Detail
Hallucinations	• LLM is unreliable because it can make up wrong answers when it doesn't know, e.g., • The wrong calculation or filter on a SQL statement • The wrong transformation • May be caused by a model not being given enough data, cab be reduced by fine tuning
LLMs are trained on very large bodies of public data	• No domain knowledge • No understanding of a company's data • No understanding of a company's metadata, business terms, metrics • No understanding of different user roles or business context
Governance for AI	• Safe use - How to train LLMs on data that is safe to use? • Compliance – use of regulated sensitive data may expose that data and/or violate privacy • Prompt governance - Avoiding sensitive and commercially confidential data in prompts • Keeping a prompt log
Infringement of IP and copyright	• Reputation or legal implications
Ethics	• Ensuring no bias, ensure use that will not cause harm to people or your business
Criminal / Social use	• Deepfake, abuse (e.g., bad code), fraud, phishing…

Copyright © Intelligent Business Strategies 1992-2024

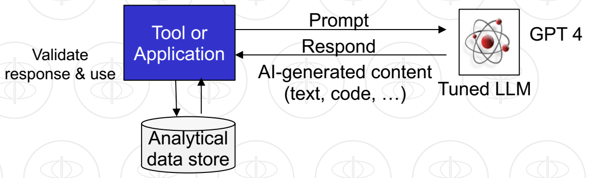
Copyright © Intelligent Business Strategies 1992-2024

How Does Generative AI Work?

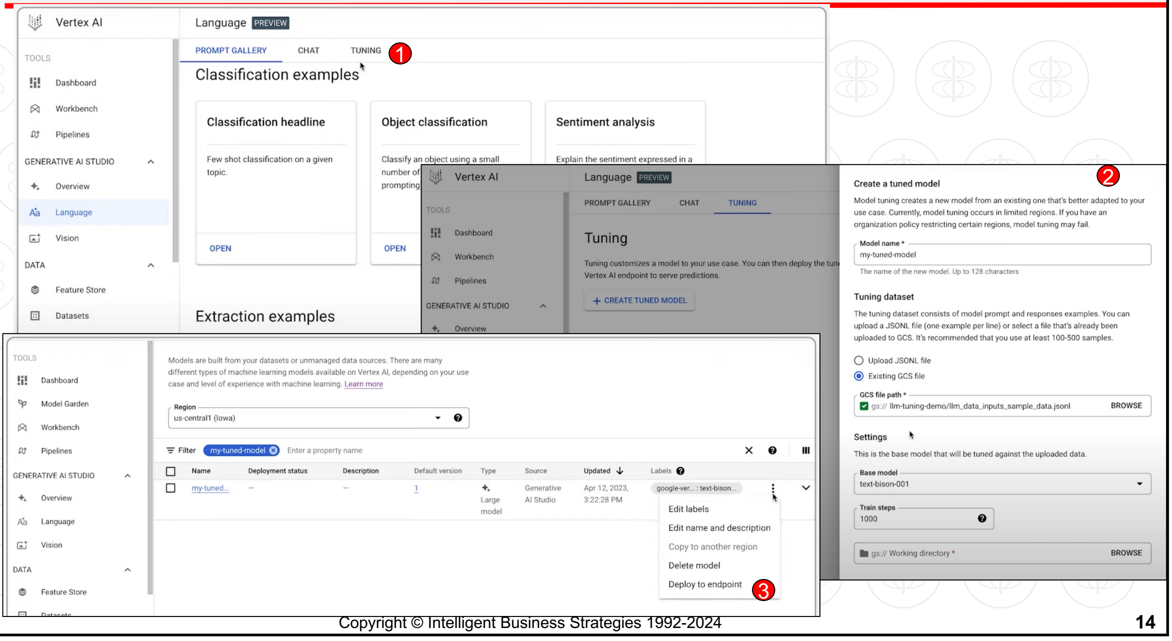
1. Select the Foundation LLM you want to use e.g., Open AI GPT4
2. Use it as is OR fine tune it to create a customised LLM for your needs
   - A combination of supervised learning and re-enforcement learning with human feedback
   - A task which can be minutes to hours



3. Utilise the customised fine-tuned LLM by invoking it from within applications and tools



Tuning A LLM Can Be Done In A Data Science WorkBench _ E.g., Fine Tuning A LLM In Google Vertex AI And Deploying It To An Endpoint



Classification examples

Classification headline

Few shot classification on a given topic.

Object classification

Classify an object using a small number of prompting

Sentiment analysis

Explain the sentiment expressed in a

Extraction examples

Tuning

Tuning customizes a model to your use case. You can then deploy the tun Vertex AI endpoint to serve predictions.

Models are built from your datasets or unmanaged data sources. There are many different types of machine learning models available on Vertex AI, depending on your use case and level of experience with machine learning. Learn more

Region: us-central1 (Iowa)

Filter: my-tuned-model — Enter a property name

☐	Name	Deployment status	Description	Default version	Type	Source	Updated	Labels
☐	my-tuned...	–	–	1	Large model	Generative AI Studio	Apr 12, 2023, 3:22:28 PM	google-ver... : text-bison...

Create a tuned model

Model tuning creates a new model from an existing one that's better adapted to your use case. Currently, model tuning occurs in limited regions. If you have an organization policy restricting certain regions, model tuning may fail.

Model name *: my-tuned-model

The name of the new model. Up to 128 characters

Tuning dataset

The tuning dataset consists of model prompt and responses examples. You can upload a JSONL file (one example per line) or select a file that's already been uploaded to GCS. It's recommended that you use at least 100-500 samples.

○ Upload JSONL file
● Existing GCS file

GCS file path *: gs:// llm-tuning-demo/llm_data_inputs_sample_data.jsonl

Settings

This is the base model that will be tuned against the uploaded data.

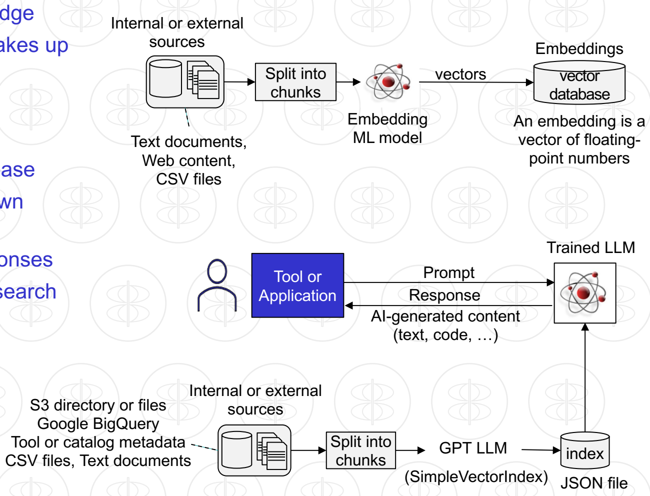
Base model: text-bison-001

Train steps: 1000

gs:// Working directory *

Supplementing And Training Open AI ChatGPT With Your Own Knowledge Base – Creating Embeddings And Indexes

- ChatGPT has been trained on general knowledge
  - It has no or limited domain specific knowledge
  - Hallucinations can occur – when a LLM makes up information
- What are embeddings?
  - They supplement ChatGPT's knowledge base
  - Provide additional information from your own knowledge base that meets your needs
  - Helps provide more relevant, reliable responses
  - Help enable faster retrieval and similarity search
  - Provides long-term memory for LLMs
- Vector indexes
  - Created on vector databases
  - Enable faster search

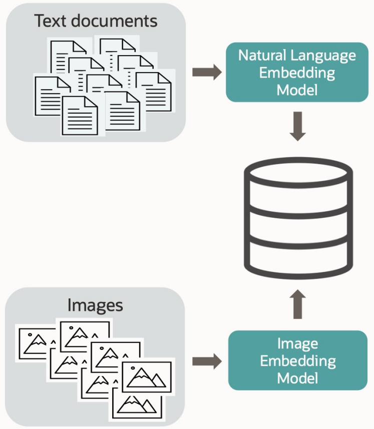


Reference article https://beebom.com/how-train-ai-chatbot-custom-knowledge-base-chatgpt-api/

Copyright © Intelligent Business Strategies 1992-2024

Generating Embeddings In A Vector Database

Creating embeddings addresses the hallucination problem associated with LLM responses
Vectors augment the prompt with enterprise-specific content to produce better responses



Text Vector Table

id	vector	text
1	[0.8, 0.5, 1.6, -2.5, …]	"It was the best of times, it was the worst of times, it was.."
2	[1.1, 0.3, 0.6, -1.3, …]	"It is a truth universally acknowledged, that a single man.."
3	[1.3, 0.1, 0.2, -1.1, …]	"It was a bright cold day in April, and the clocks were striking.."
…	…	…

Image Vector Table

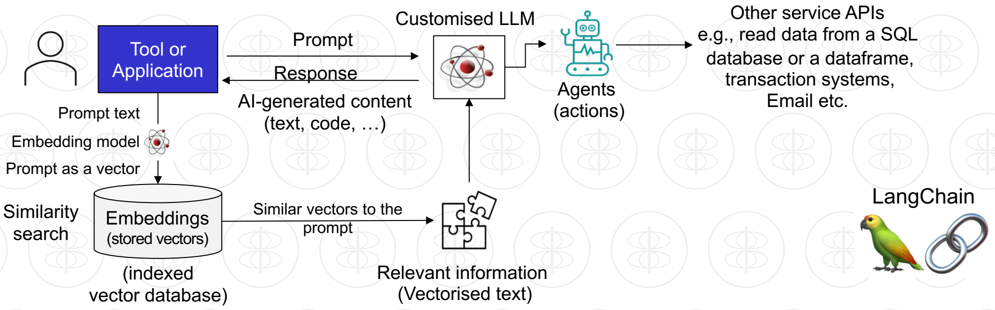
id	vector	Image
1	[0.5, 1.5, 2.6, -1.1, …]	
2	[1.0, 0.9, 1.6, -1.3, …]	
3	[0.6, 1.1, 1.3, -0.9, …]	
…	…	…

Image source and Copyright ©: Oracle - LRN1412 "Enabling Generative AI with AI Vector Search in Oracle Database"

Copyright © Intelligent Business Strategies 1992-2024

LLM API Frameworks Are Also Now Available To Help Make Use Of LLMs In Tools And Applications

- Several new frameworks have emerged to develop LLM powered Gen AI applications and tools
  - E.g., Amazon Bedrock, Dust, LangChain, Steamship…
- LangChain is a framework for developing applications and tools powered by LLMs
  - LangChain is often used inside data management and BI tools



Tool or Application

Prompt

Response

AI-generated content (text, code, …)

Customised LLM

Agents (actions)

Other service APIs e.g., read data from a SQL database or a dataframe, transaction systems, Email etc.

Prompt text

Embedding model

Prompt as a vector

Similarity search

Embeddings (stored vectors)

(indexed vector database)

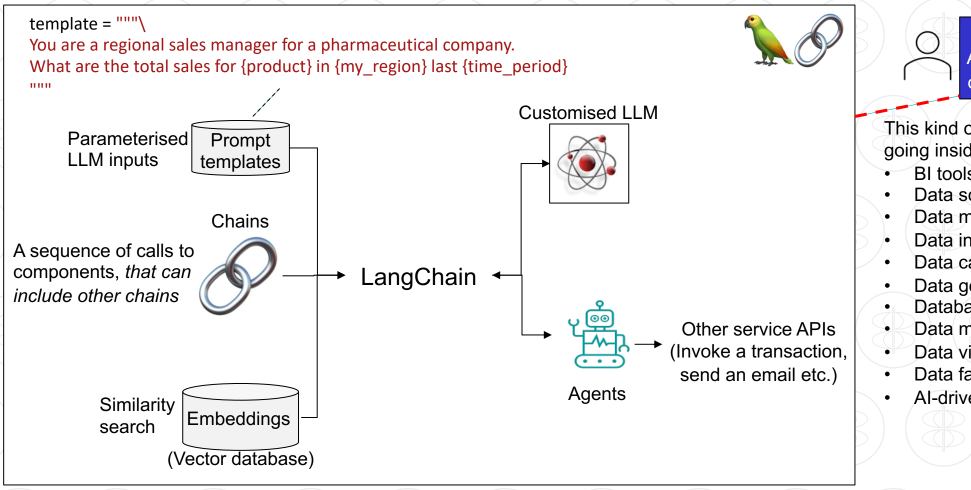
Similar vectors to the prompt

Relevant information (Vectorised text)

LangChain

AI Development Using LLM API Frameworks - LangChain Is Used To Connect LLMs To Your Own Data And Also Use The LLM To Help You Take Actions

```
template = """\
You are a regional sales manager for a pharmaceutical company.
What are the total sales for {product} in {my_region} last {time_period}
"""
```



Parameterised LLM inputs

Prompt templates

Chains

A sequence of calls to components, *that can include other chains*

LangChain

Customised LLM

Agents

Other service APIs (Invoke a transaction, send an email etc.)

Similarity search

Embeddings

(Vector database)

Data & Analytics Tool or Application

This kind of capability is now going inside

- BI tools
- Data science workbenches
- Data modelling tools
- Data integration tools
- Data catalogs
- Data governance tools
- Databases
- Data migration tools
- Data virtualisation
- Data fabric software
- AI-driven applications

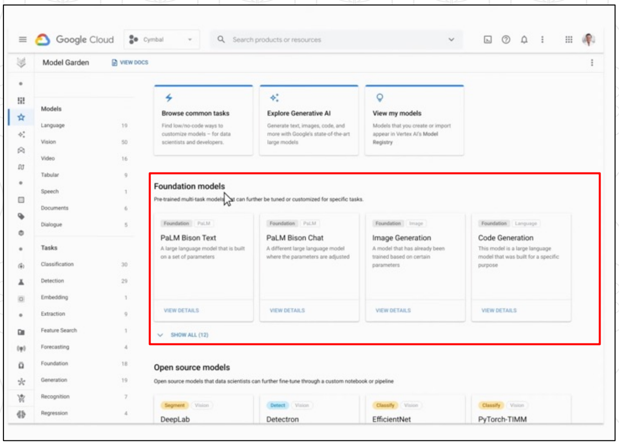
You can have different prompt template libraries for different technical and business use cases
E.g. prompts for data engineering, data science, marketing insights, financial insights, sales insights, HR insights, …

Development Tools Now Exist To Build Generative AI Applications – E.g. Google Vertex AI Search And Conversation (formerly Generative AI App Builder)

Google Vertex AI Search and Vertex AI Conversation

- Brings together Google's foundation AI models and search
- Includes a model garden of Google's foundation and partner AI models
- Build apps without writing any code that make use of generative AI
- Transact during conversation with the bots you build

Conversational AI using LLMs is likely to appear in ALL major applications in the enterprise



Find low/no-code ways to customize models – for data scientists and developers.

Generate text, images, code, and more with Google's state-of-the-art large models

Models that you create or import appear in Vertex AI's Model Registry

Pre-trained multi-task models that can further be tuned or customized for specific tasks.

A large language model that is built on a set of parameters

A different large language model where the parameters are adjusted

A model that has already been trained based on certain parameters

This model is a large language model that was built for a specific purpose

Open source models that data scientists can further fine-tune through a custom notebook or pipeline

Gen Vertex AI Model Garden

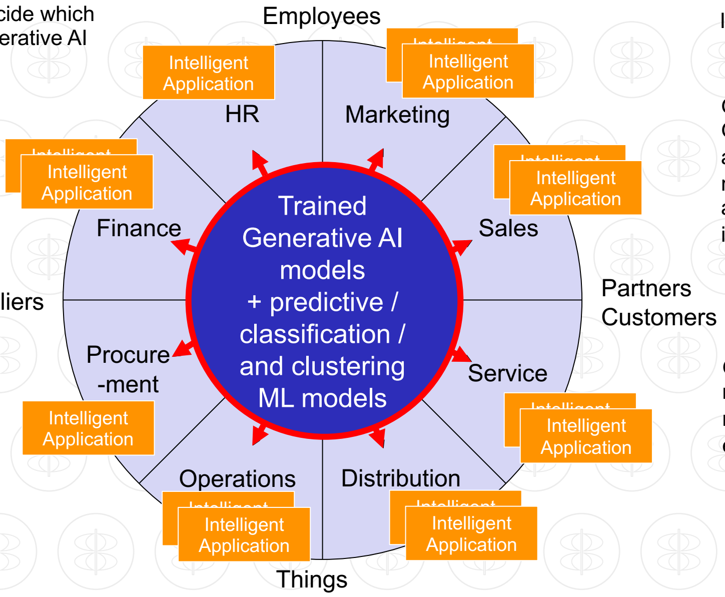
Foundation models can be tuned and tested in Vertex AI

Business Wants Generative AI And ML Models 'Wired Into' Every Application To Improve Productivity, Reduce Costs, And Shorten Time To Action For Better Profitability

Get employees to decide which tasks should use generative AI

AI-Assistant Productivity Automation

Use generative AI to improve ways of working that help yield better returns



Intelligent applications

Governed access to Generative AI LLMs and trained ML models as services across the business is critical to success

Generative AI does not replace other ML models in use in the enterprise

Business Prompt Examples – Getting Things Done With Generative AI In Customer Facing Business Functions

Marketing

- Create a lead gen page for a webinar with a sign-up form
- Create personalised emails to all customer likely to buy our new load product
- Create tweets to promote our new product
- Create a blog introducing our new product

Customer Service

- What is the best reply for this customer?
- Create personalised emails to all customers about our new loan product

Sales

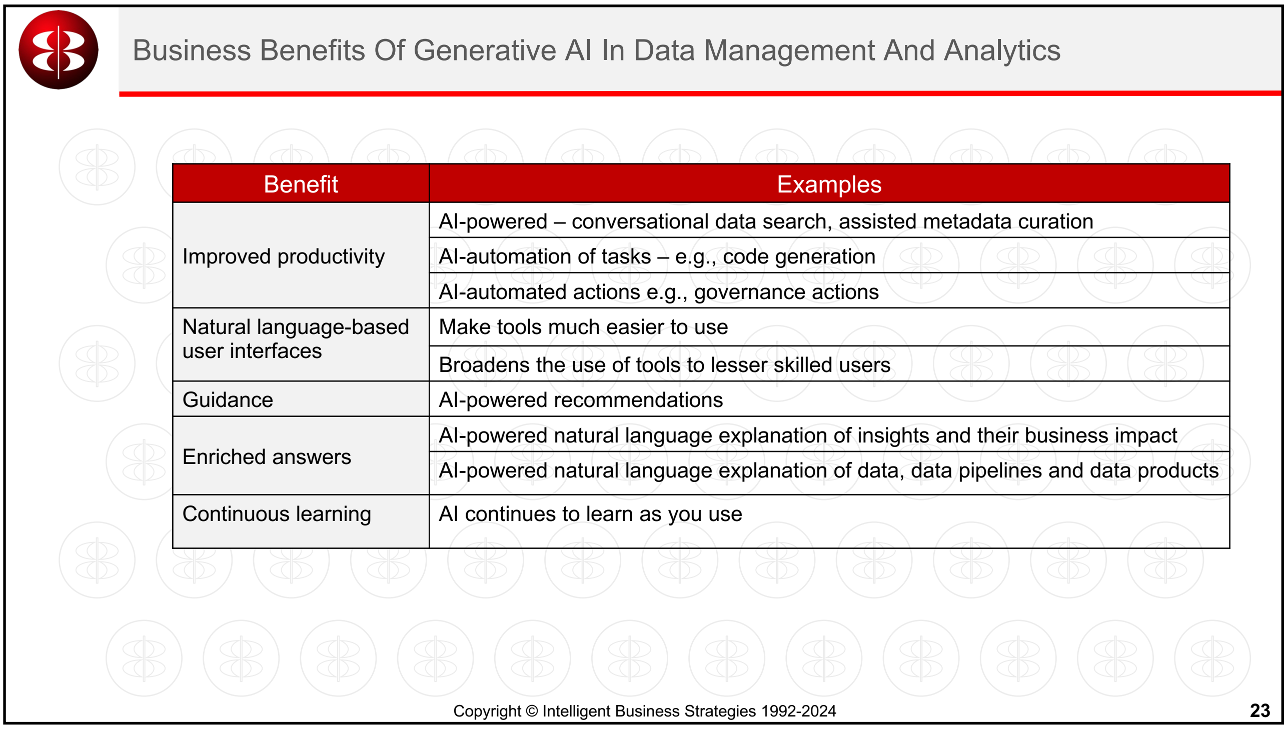
- Who are the top 10 contacts likely to buy the new product?
- Recommend the best contact to go see next?
- Write an informal introduction email
- What insights are available on Acme Corporation?

Use generative AI to change customer and employee experiences and simplify complex tasks

- AI powered voice and text conversations with customers
- Build generative AI based personalised marketing into customer experiences
- Help employees get things done quicker

Topics – Where Are We?

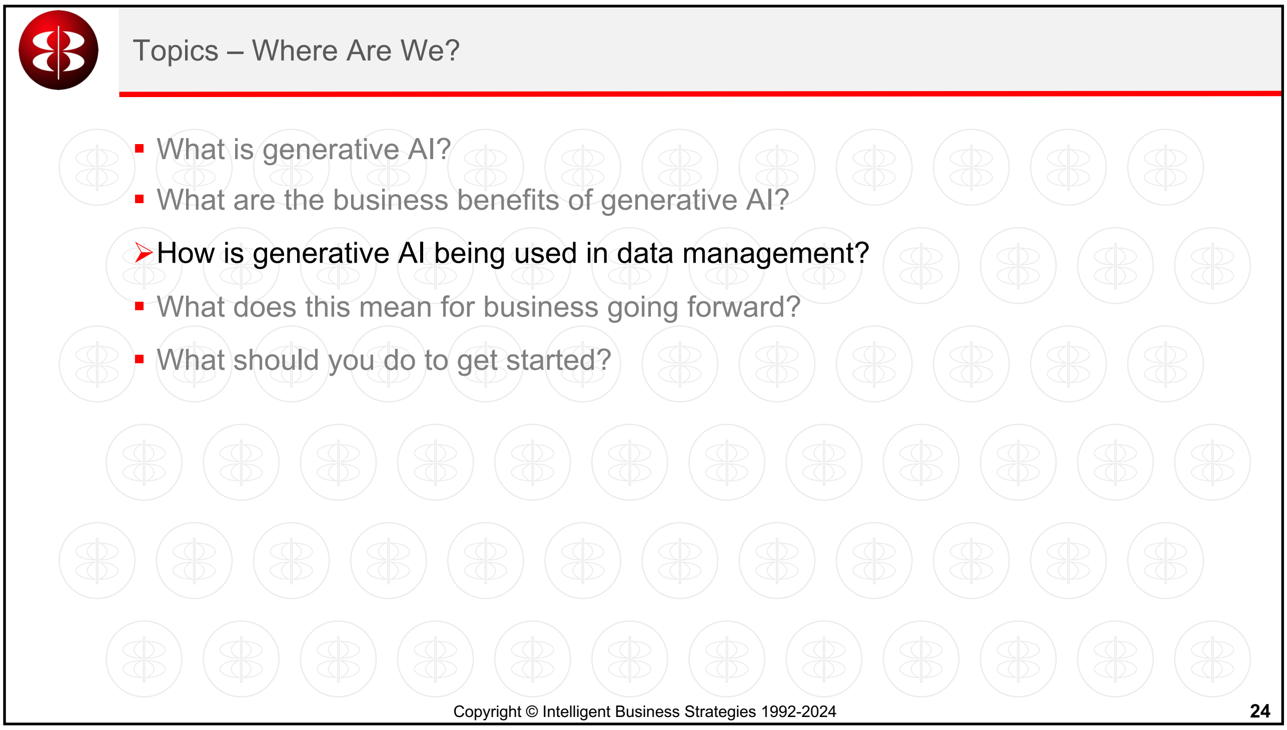
- What is generative AI?
- ➢What are the business benefits of generative AI?
- How is generative AI being used in data management?
- How is generative AI being used in data science and BI
- What does this mean for business going forward?
- What should you do to get started?



Business Benefits Of Generative AI In Data Management And Analytics

Benefit	Examples
Improved productivity	AI-powered – conversational data search, assisted metadata curation
	AI-automation of tasks – e.g., code generation
	AI-automated actions e.g., governance actions
Natural language-based user interfaces	Make tools much easier to use
	Broadens the use of tools to lesser skilled users
Guidance	AI-powered recommendations
Enriched answers	AI-powered natural language explanation of insights and their business impact
	AI-powered natural language explanation of data, data pipelines and data products
Continuous learning	AI continues to learn as you use

Copyright © Intelligent Business Strategies 1992-2024



Topics – Where Are We?

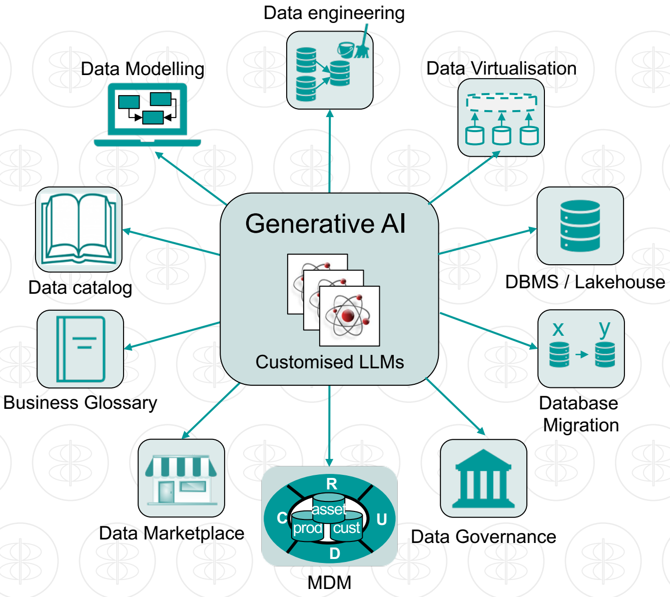
- What is generative AI?
- What are the business benefits of generative AI?
- ➤How is generative AI being used in data management?
- What does this mean for business going forward?
- What should you do to get started?

Copyright © Intelligent Business Strategies 1992-2024

Mega Trends
– Generative AI Has Already Emerged In Data Management And Analytics Software

Data & Analytics Category	Example Data & AI Products Already Supporting Generative AI
Data Catalogs	Atlan
	Collibra
	Data.world
Data Integration	SnapLogic SnapGPT
	Denodo
Data Governance	Informatica CLAIRE GPT (glossaries, policies, rules…) SodaGPT
Database Management Systems	ArangoDB
	Google Duet AI in BigQuery and BigQueryML
	Snowflake, Oracle
	Neo4j, Kinetica
Master Data Management	ViaMedici PIM
Self-service analytics / BI	ThoughtSpot Sage
	Tableau GPT & Einstein CoPilot
Data Science Workbench	Databricks LakehouseIQ & English API, LakehouseAI
	IBM Watsonx.ai
	Google Vertex AI
Decision Intelligence	Aera

Generative AI Has Emerged In Almost Every Aspect Of Data Management



Generative AI

Customised LLMs

Data engineering

Data Modelling

Data Virtualisation

Data catalog

DBMS / Lakehouse

Business Glossary

Database Migration

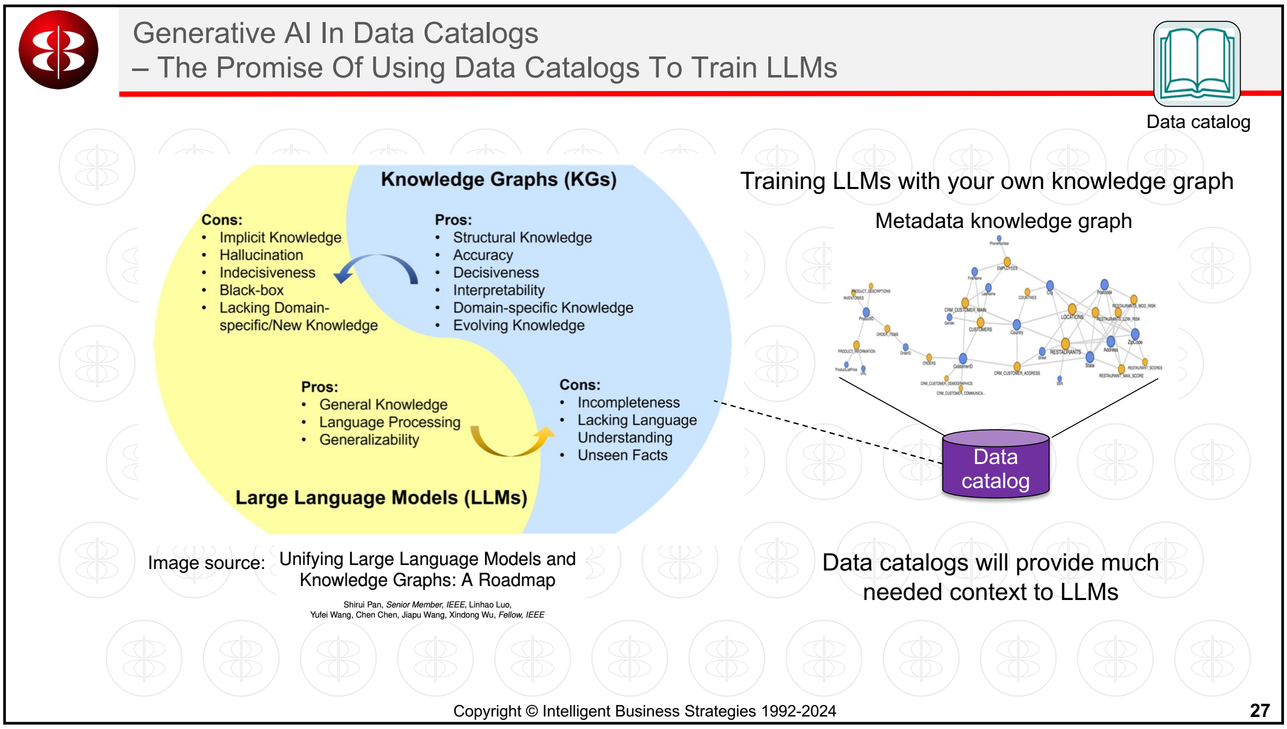
Data Marketplace

MDM

Data Governance

Copyright © Intelligent Business Strategies 1992-2024

Copyright © Intelligent Business Strategies 1992-2024



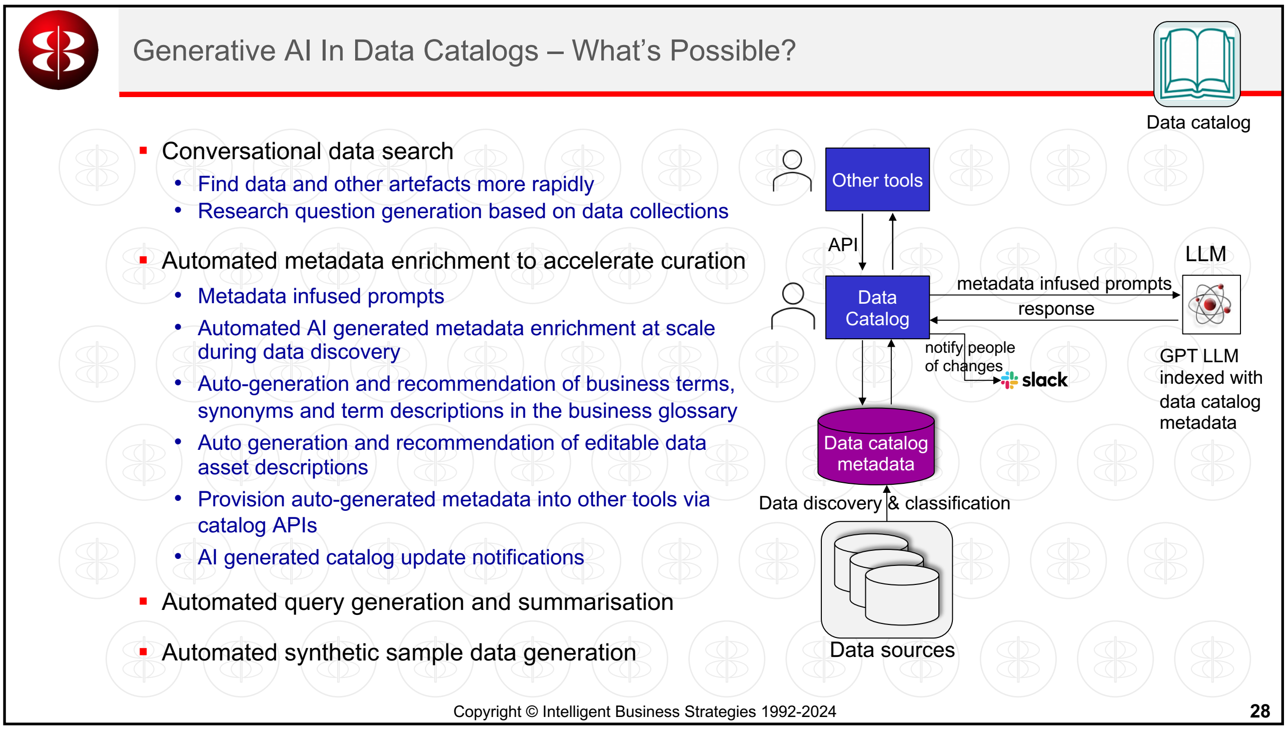
Generative AI In Data Catalogs – The Promise Of Using Data Catalogs To Train LLMs

Image source: Unifying Large Language Models and Knowledge Graphs: A Roadmap

Shirui Pan, *Senior Member, IEEE*, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, Xindong Wu, *Fellow, IEEE*

Training LLMs with your own knowledge graph

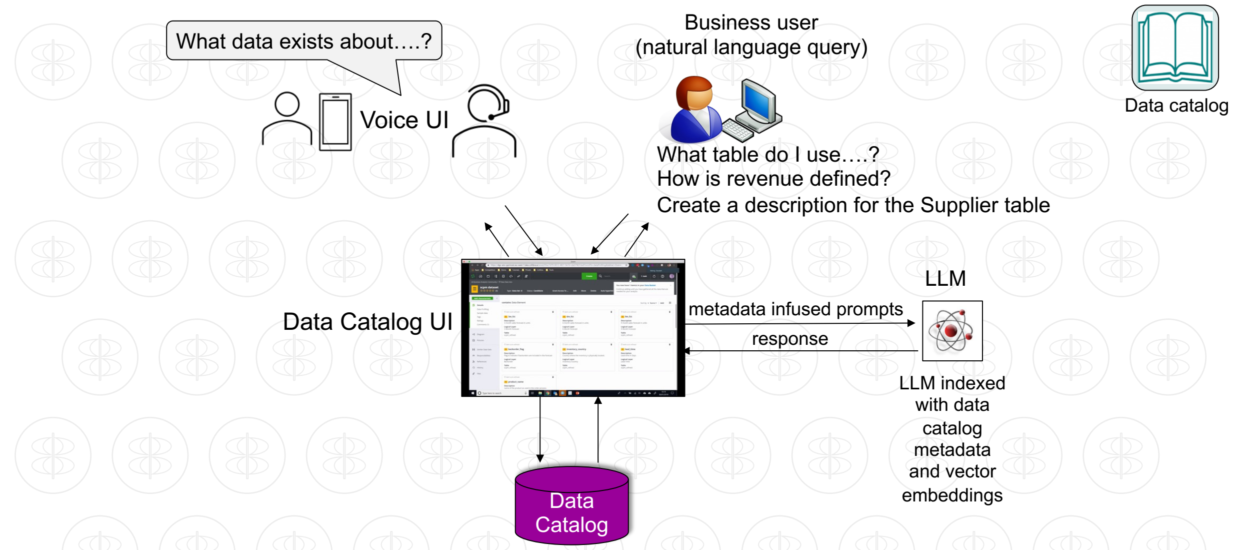
Data catalogs will provide much needed context to LLMs



Generative AI In Data Catalogs – What's Possible?

- Conversational data search
  - Find data and other artefacts more rapidly
  - Research question generation based on data collections
- Automated metadata enrichment to accelerate curation
  - Metadata infused prompts
  - Automated AI generated metadata enrichment at scale during data discovery
  - Auto-generation and recommendation of business terms, synonyms and term descriptions in the business glossary
  - Auto generation and recommendation of editable data asset descriptions
  - Provision auto-generated metadata into other tools via catalog APIs
  - AI generated catalog update notifications
- Automated query generation and summarisation
- Automated synthetic sample data generation

Using Generative AI To Simplifying The Data Catalog User Interface – Conversational Data Search Using Natural Language And Voice To Find Data



What data exists about….?

Voice UI

Business user
(natural language query)

What table do I use….?
How is revenue defined?
Create a description for the Supplier table

Data catalog

Data Catalog UI

metadata infused prompts

response

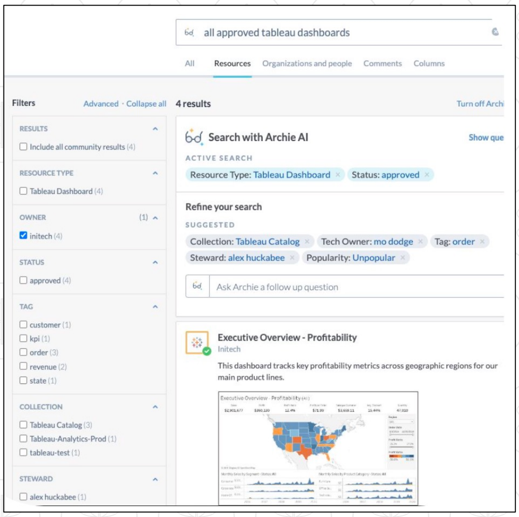
LLM

LLM indexed
with data
catalog
metadata
and vector
embeddings

Data
Catalog

Prompt-Based Search And AI-Assisted Question Generation In Data Catalogs – Product Example: Data.world

AI-Assisted Search



all approved tableau dashboards

All · Resources · Organizations and people · Comments · Columns

Filters · Advanced · Collapse all

4 results · Turn off Archi

RESULTS
☐ Include all community results (4)

RESOURCE TYPE
☐ Tableau Dashboard (4)

OWNER (1)
☑ initech (4)

STATUS
☐ approved (4)

TAG
☐ customer (1)
☐ kpi (1)
☐ order (3)
☐ revenue (2)
☐ state (1)

COLLECTION
☐ Tableau Catalog (3)
☐ Tableau-Analytics-Prod (1)
☐ tableau-test (1)

STEWARD
☐ alex huckabee (1)

Search with Archie AI · Show que

ACTIVE SEARCH
Resource Type: Tableau Dashboard × · Status: approved ×

Refine your search

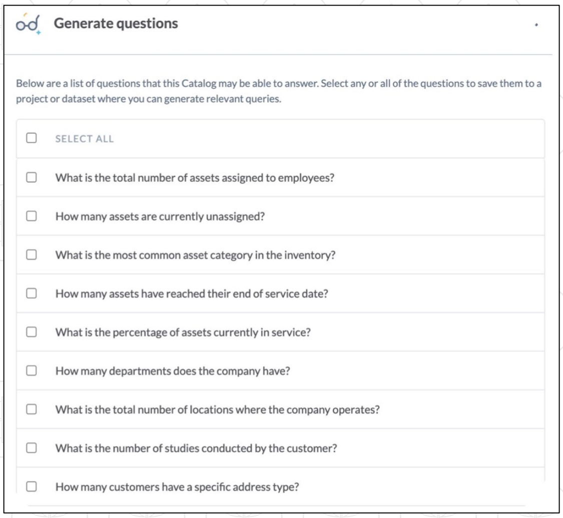
SUGGESTED
Collection: Tableau Catalog × · Tech Owner: mo dodge × · Tag: order ×
Steward: alex huckabee × · Popularity: Unpopular ×

Ask Archie a follow up question

Executive Overview - Profitability
Initech
This dashboard tracks key profitability metrics across geographic regions for our main product lines.

Source: data.world

Automated AI-Assisted Question Generation



Generate questions

Below are a list of questions that this Catalog may be able to answer. Select any or all of the questions to save them to a project or dataset where you can generate relevant queries.

☐ SELECT ALL
☐ What is the total number of assets assigned to employees?
☐ How many assets are currently unassigned?
☐ What is the most common asset category in the inventory?
☐ How many assets have reached their end of service date?
☐ What is the percentage of assets currently in service?
☐ How many departments does the company have?
☐ What is the total number of locations where the company operates?
☐ What is the number of studies conducted by the customer?
☐ How many customers have a specific address type?

data.world now has generative AI bots called Eureka bots

Using Generative AI For Data Catalog Metadata Enrichment – Auto-Generation Of Business Term Descriptions In Atlan Business Glossary

Business Glossary

Data catalog

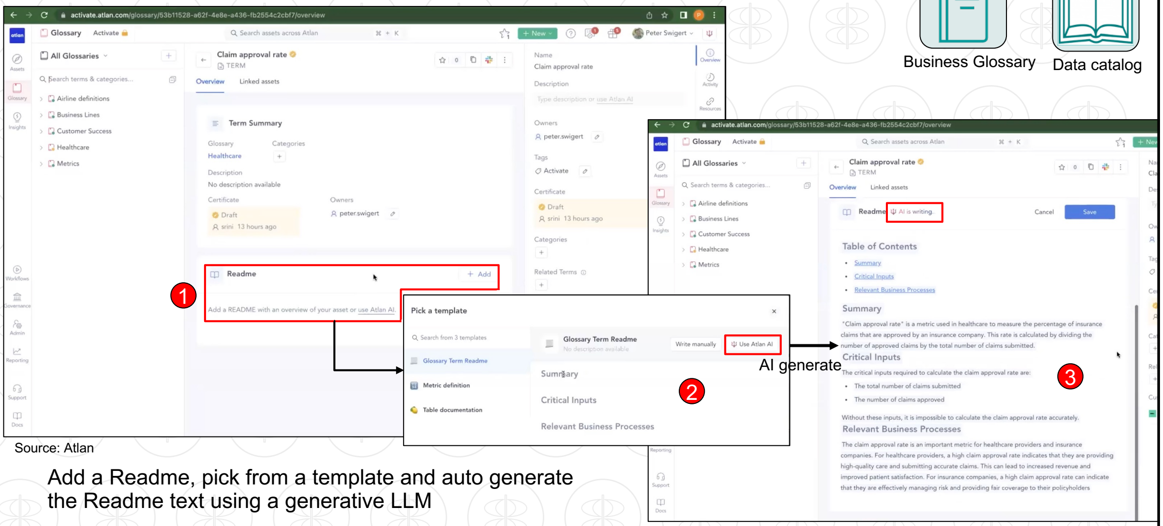


Table of Contents

- Summary
- Critical Inputs
- Relevant Business Processes

Summary

"Claim approval rate" is a metric used in healthcare to measure the percentage of insurance claims that are approved by an insurance company. This rate is calculated by dividing the number of approved claims by the total number of claims submitted.

Critical Inputs

The critical inputs required to calculate the claim approval rate are:

- The total number of claims submitted
- The number of claims approved

Without these inputs, it is impossible to calculate the claim approval rate accurately.

Relevant Business Processes

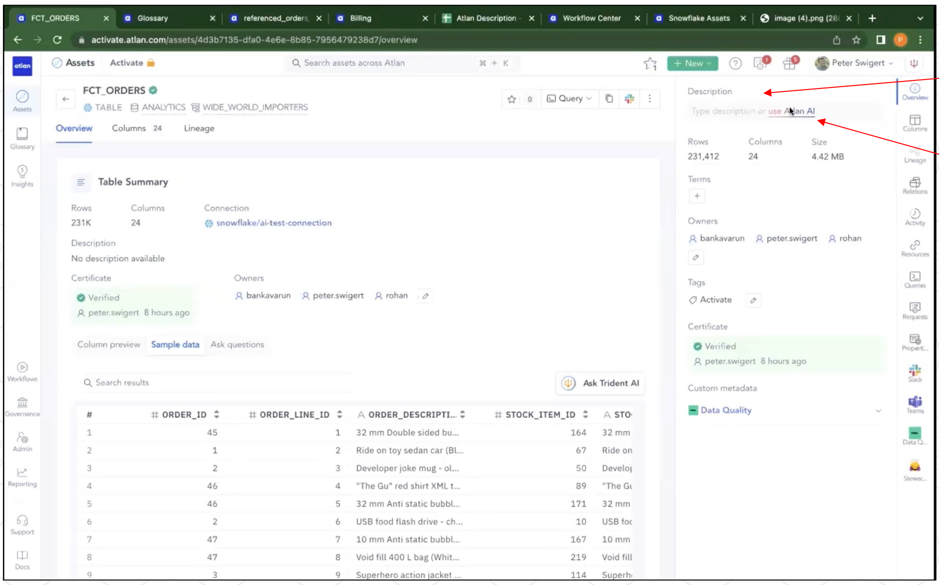
The claim approval rate is an important metric for healthcare providers and insurance companies. For healthcare providers, a high claim approval rate indicates that they are providing high-quality care and submitting accurate claims. This can lead to increased revenue and improved patient satisfaction. For insurance companies, a high claim approval rate can indicate that they are effectively managing risk and providing fair coverage to their policyholders

Source: Atlan

Add a Readme, pick from a template and auto generate the Readme text using a generative LLM

Copyright © Intelligent Business Strategies 1992-2024

Generative AI In Data Catalogs – Automated Metadata Enrichment Using Atlan AI To Generate A Description Of An Asset



Rows 231K

Columns 24

Connection snowflake/ai-test-connection

Description

No description available

Rows 231,412

Columns 24

Size 4.42 MB

#	ORDER_ID	ORDER_LINE_ID	ORDER_DESCRIPTI...	STOCK_ITEM_ID	STO
1	45	1	32 mm Double sided bu...	164	32 mm
2	1	2	Ride on toy sedan car (Bl...	67	Ride on
3	2	3	Developer joke mug - ol...	50	Develop
4	46	4	"The Gu" red shirt XML t...	89	"The Gu
5	46	5	32 mm Anti static bubbl...	171	32 mm
6	2	6	USB food flash drive - ch...	10	USB foo
7	47	7	10 mm Anti static bubbl...	167	10 mm
8	47	8	Void fill 400 L bag (Whit...	219	Void fill
9	3	9	Superhero action jacket ...	114	Superh

FCT_ORDERS table currently has no description

Request Atlan AI to generate one

Data catalog

Source: Atlan

Copyright © Intelligent Business Strategies 1992-2024

Atlan Generated Contextual Description – E.g. The Description Includes The Database The Table Is In, The Schema

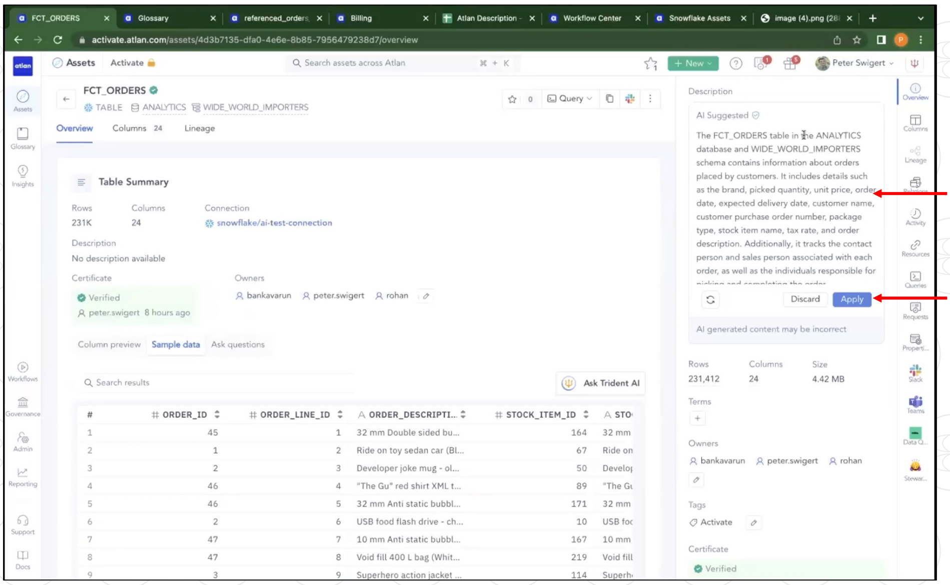


Table Summary

Rows: 231K
Columns: 24
Connection: snowflake/ai-test-connection

Description: No description available

Certificate: Verified — peter.swigert 8 hours ago

Owners: bankavarun, peter.swigert, rohan

#	ORDER_ID	ORDER_LINE_ID	ORDER_DESCRIPTI...	STOCK_ITEM_ID	STO
1	45	1	32 mm Double sided bu...	164	32 mm
2	1	2	Ride on toy sedan car (Bl...	67	Ride on
3	2	3	Developer joke mug - ol...	50	Develo
4	46	4	"The Gu" red shirt XML t...	89	"The Gu
5	46	5	32 mm Anti static bubbl...	171	32 mm
6	2	6	USB food flash drive - ch...	10	USB foo
7	47	7	10 mm Anti static bubbl...	167	10 mm
8	47	8	Void fill 400 L bag (Whit...	219	Void fill
9	3	9	Superhero action jacket ...	114	Superh

Description

AI Suggested

The FCT_ORDERS table in the ANALYTICS database and WIDE_WORLD_IMPORTERS schema contains information about orders placed by customers. It includes details such as the brand, picked quantity, unit price, order date, expected delivery date, customer name, customer purchase order number, package type, stock item name, tax rate, and order description. Additionally, it tracks the contact person and sales person associated with each order, as well as the individuals responsible for picking and completing the order.

AI generated content may be incorrect

Rows: 231,412
Columns: 24
Size: 4.42 MB

Terms

Owners: bankavarun, peter.swigert, rohan

Tags: Activate

Certificate: Verified

Data catalog

Generated description

User can accept this by clicking "Apply"

The user can also edit the description before accepting it

Source: Atlan

Atlan AI Activity Shows A Log Of The Changes Made By A User Using Atlan AI

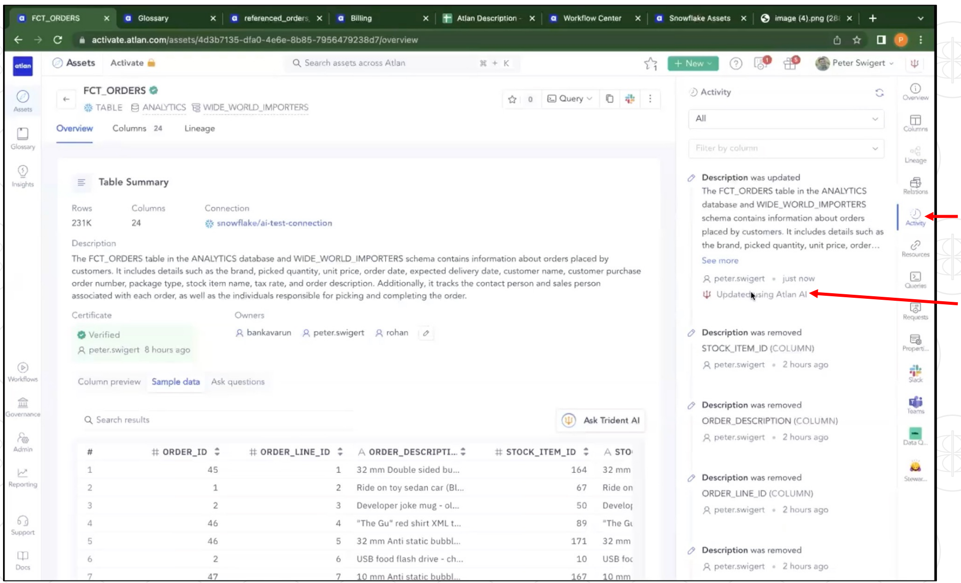


Table Summary

Rows: 231K
Columns: 24
Connection: snowflake/ai-test-connection

Description: The FCT_ORDERS table in the ANALYTICS database and WIDE_WORLD_IMPORTERS schema contains information about orders placed by customers. It includes details such as the brand, picked quantity, unit price, order date, expected delivery date, customer name, customer purchase order number, package type, stock item name, tax rate, and order description. Additionally, it tracks the contact person and sales person associated with each order, as well as the individuals responsible for picking and completing the order.

Certificate: Verified — peter.swigert 8 hours ago

Owners: bankavarun, peter.swigert, rohan

#	ORDER_ID	ORDER_LINE_ID	ORDER_DESCRIPTI...	STOCK_ITEM_ID	STO
1	45	1	32 mm Double sided bu...	164	32 mm
2	1	2	Ride on toy sedan car (Bl...	67	Ride on
3	2	3	Developer joke mug - ol...	50	Develo
4	46	4	"The Gu" red shirt XML t...	89	"The Gu
5	46	5	32 mm Anti static bubbl...	171	32 mm
6	2	6	USB food flash drive - ch...	10	USB foo
7	47	7	10 mm Anti static bubbl...	167	10 mm

Description was updated
The FCT_ORDERS table in the ANALYTICS database and WIDE_WORLD_IMPORTERS schema contains information about orders placed by customers. It includes details such as the brand, picked quantity, unit price, order…
See more
peter.swigert • just now
Updated using Atlan AI

Description was removed
STOCK_ITEM_ID (COLUMN)
peter.swigert • 2 hours ago

Description was removed
ORDER_DESCRIPTION (COLUMN)
peter.swigert • 2 hours ago

Description was removed
ORDER_LINE_ID (COLUMN)
peter.swigert • 2 hours ago

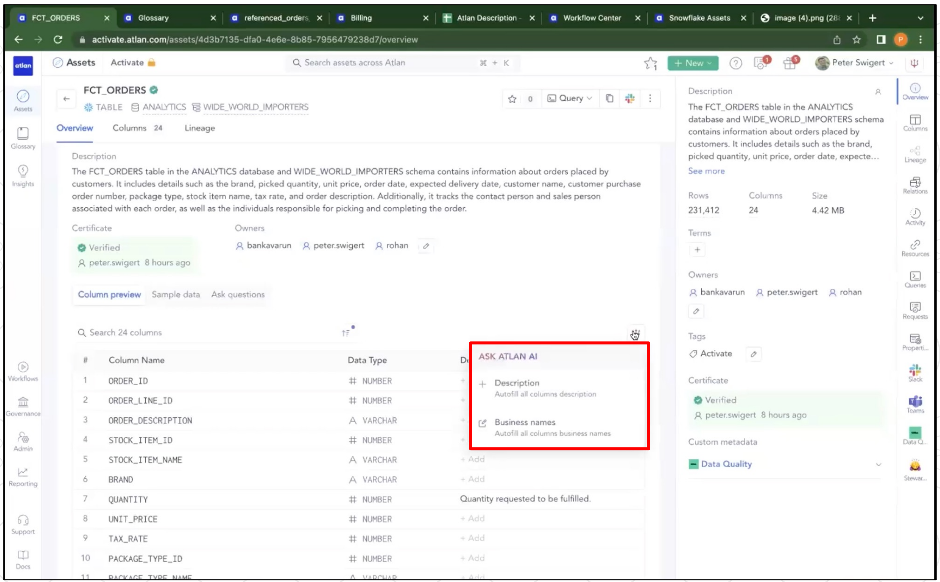
Description was removed
peter.swigert • 2 hours ago

Data catalog

'Atlan keeps an activity' log that shows the description was updated by a user using Atlan AI

Source: Atlan

Generative AI And Data Catalogs - Automated Metadata Enrichment Using Atlan AI To Auto-Generate Column Descriptions And Business Data Names



FCT_ORDERS
TABLE · ANALYTICS · WIDE_WORLD_IMPORTERS

Overview · Columns 24 · Lineage

Description
The FCT_ORDERS table in the ANALYTICS database and WIDE_WORLD_IMPORTERS schema contains information about orders placed by customers. It includes details such as the brand, picked quantity, unit price, order date, expected delivery date, customer name, customer purchase order number, package type, stock item name, tax rate, and order description. Additionally, it tracks the contact person and sales person associated with each order, as well as the individuals responsible for picking and completing the order.

Certificate
Verified
peter.swigert 8 hours ago

Owners
bankavarun · peter.swigert · rohan

Column preview · Sample data · Ask questions

Search 24 columns

#	Column Name	Data Type	De
1	ORDER_ID	NUMBER	+
2	ORDER_LINE_ID	NUMBER	+
3	ORDER_DESCRIPTION	VARCHAR	+
4	STOCK_ITEM_ID	NUMBER	+
5	STOCK_ITEM_NAME	VARCHAR	+ Add
6	BRAND	VARCHAR	+ Add
7	QUANTITY	NUMBER	Quantity requested to be fulfilled.
8	UNIT_PRICE	NUMBER	+ Add
9	TAX_RATE	NUMBER	+ Add
10	PACKAGE_TYPE_ID	NUMBER	+ Add

ASK ATLAN AI
- Description
  Autofill all columns description
- Business names
  Autofill all columns business names

Description
The FCT_ORDERS table in the ANALYTICS database and WIDE_WORLD_IMPORTERS schema contains information about orders placed by customers. It includes details such as the brand, picked quantity, unit price, order date, expecte...
See more

Rows	Columns	Size
231,412	24	4.42 MB

Terms

Owners
bankavarun · peter.swigert · rohan

Tags
Activate

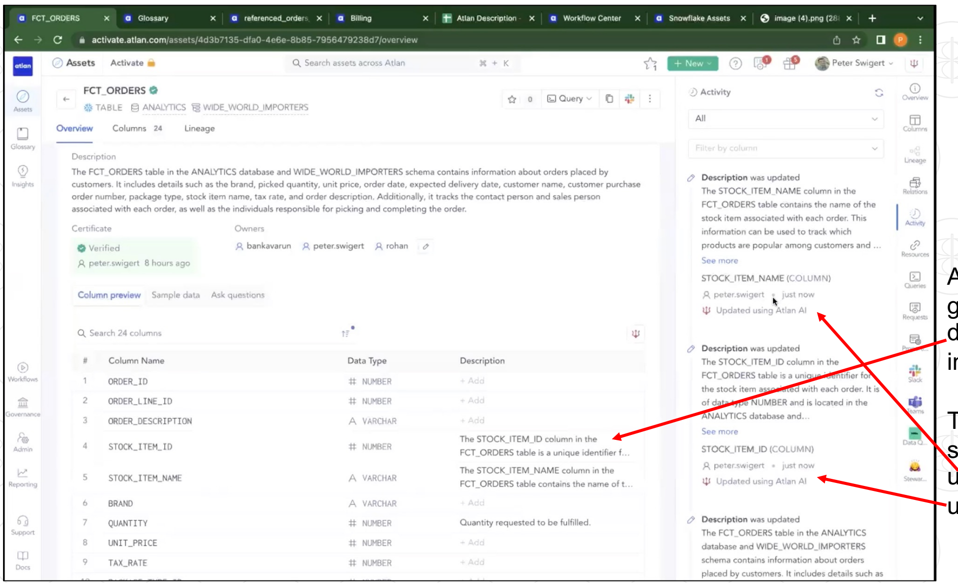
Certificate
Verified
peter.swigert 8 hours ago

Custom metadata
Data Quality

Data catalog

Source: Atlan

Atlan AI Activity Shows A Log Of The Column Description Changes Made By A User Using Atlan AI



FCT_ORDERS
TABLE · ANALYTICS · WIDE_WORLD_IMPORTERS

Overview · Columns 24 · Lineage

Description
The FCT_ORDERS table in the ANALYTICS database and WIDE_WORLD_IMPORTERS schema contains information about orders placed by customers. It includes details such as the brand, picked quantity, unit price, order date, expected delivery date, customer name, customer purchase order number, package type, stock item name, tax rate, and order description. Additionally, it tracks the contact person and sales person associated with each order, as well as the individuals responsible for picking and completing the order.

Certificate
Verified
peter.swigert 8 hours ago

Owners
bankavarun · peter.swigert · rohan

Column preview · Sample data · Ask questions

Search 24 columns

#	Column Name	Data Type	Description
1	ORDER_ID	NUMBER	+ Add
2	ORDER_LINE_ID	NUMBER	+ Add
3	ORDER_DESCRIPTION	VARCHAR	+ Add
4	STOCK_ITEM_ID	NUMBER	The STOCK_ITEM_ID column in the FCT_ORDERS table is a unique identifier f...
5	STOCK_ITEM_NAME	VARCHAR	The STOCK_ITEM_NAME column in the FCT_ORDERS table contains the name of t...
6	BRAND	VARCHAR	+ Add
7	QUANTITY	NUMBER	Quantity requested to be fulfilled.
8	UNIT_PRICE	NUMBER	+ Add
9	TAX_RATE	NUMBER	+ Add

Activity
All
Filter by column

Description was updated
The STOCK_ITEM_NAME column in the FCT_ORDERS table contains the name of the stock item associated with each order. This information can be used to track which products are popular among customers and ...
See more
STOCK_ITEM_NAME (COLUMN)
peter.swigert • just now
Updated using Atlan AI

Description was updated
The STOCK_ITEM_ID column in the FCT_ORDERS table is a unique identifier for the stock item associated with each order. It is of data type NUMBER and is located in the ANALYTICS database and...
See more
STOCK_ITEM_ID (COLUMN)
peter.swigert • just now
Updated using Atlan AI

Description was updated
The FCT_ORDERS table in the ANALYTICS database and WIDE_WORLD_IMPORTERS schema contains information about orders placed by customers. It includes details such as

Data catalog

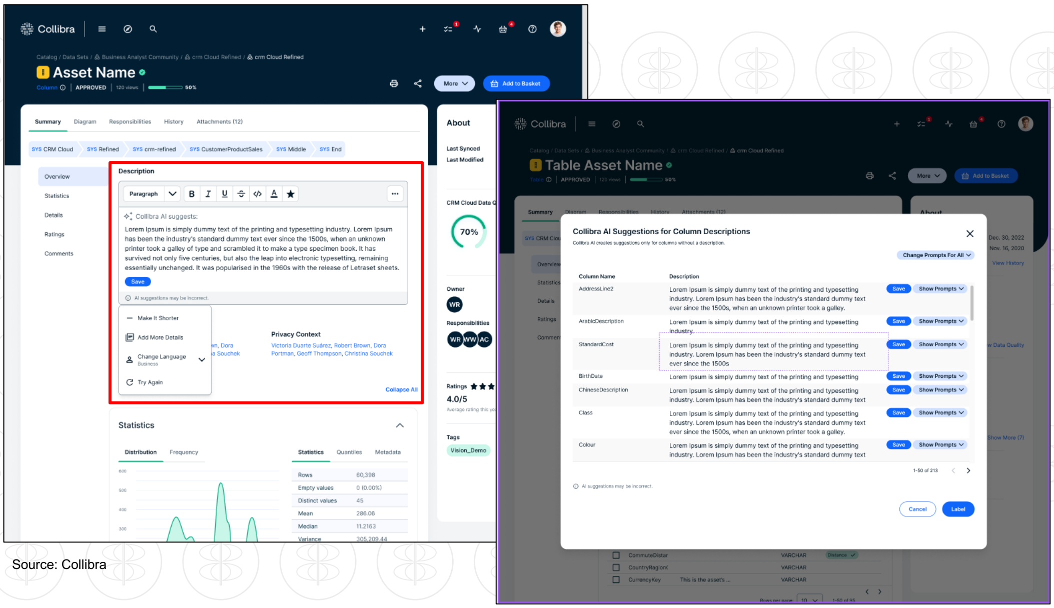
Accepted AI-generated column descriptions are immediately visible

The activity log shows they were updated by a user using Atlan AI

Source: Atlan

Generative AI In Data Catalogs - Automated Metadata Enrichment Using AI-Generated Column Descriptions In Collibra

Data catalog



Collibra AI suggests:

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets.

AI suggestions may be incorrect.

Statistics	
Rows	60,398
Empty values	0 (0.00%)
Distinct values	45
Mean	286.06
Median	11.2163
Variance	305,209.44

Collibra AI Suggestions for Column Descriptions

Collibra AI creates suggestions only for columns without a description.

Column Name	Description
AddressLine2	Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley.
ArabicDescription	Lorem Ipsum is simply dummy text of the printing and typesetting industry.
StandardCost	Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s
BirthDate	Lorem Ipsum is simply dummy text of the printing and typesetting
ChineseDescription	Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text
Class	Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley.
Colour	Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text

1-50 of 213

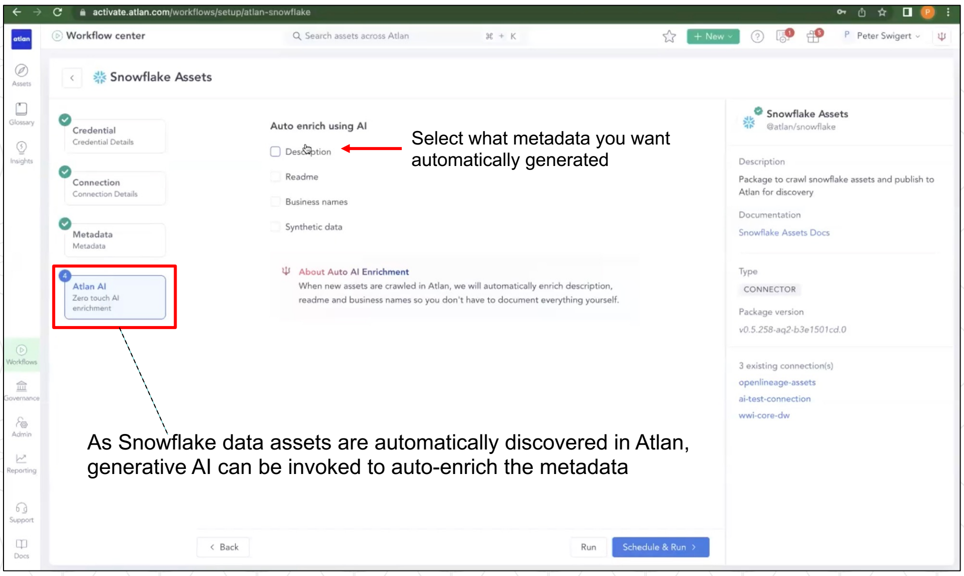
AI suggestions may be incorrect.

Source: Collibra

Copyright © Intelligent Business Strategies 1992-2024

Generative AI In Data Catalogs – Automatic AI-Generated Metadata Enrichment At Scale During Source Data Discovery In Atlan

Data catalog



Auto enrich using AI

☐ Description

☐ Readme

☐ Business names

☐ Synthetic data

Select what metadata you want automatically generated

About Auto AI Enrichment

When new assets are crawled in Atlan, we will automatically enrich description, readme and business names so you don't have to document everything yourself.

Snowflake Assets
@atlan/snowflake

Description
Package to crawl snowflake assets and publish to Atlan for discovery

Documentation
Snowflake Assets Docs

Type
CONNECTOR

Package version
v0.5.258-aq2-b3e1501cd.0

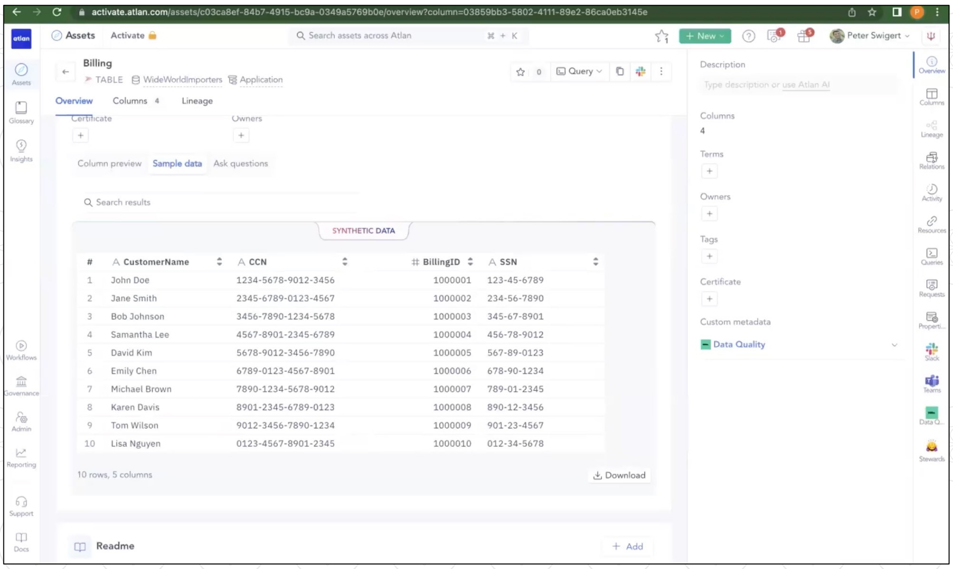
3 existing connection(s)
openlineage-assets
ai-test-connection
wwi-core-dw

As Snowflake data assets are automatically discovered in Atlan, generative AI can be invoked to auto-enrich the metadata

Source: Atlan

Copyright © Intelligent Business Strategies 1992-2024

Synthetic Sample Data Generation In Atlan Data Catalog Using Generative AI



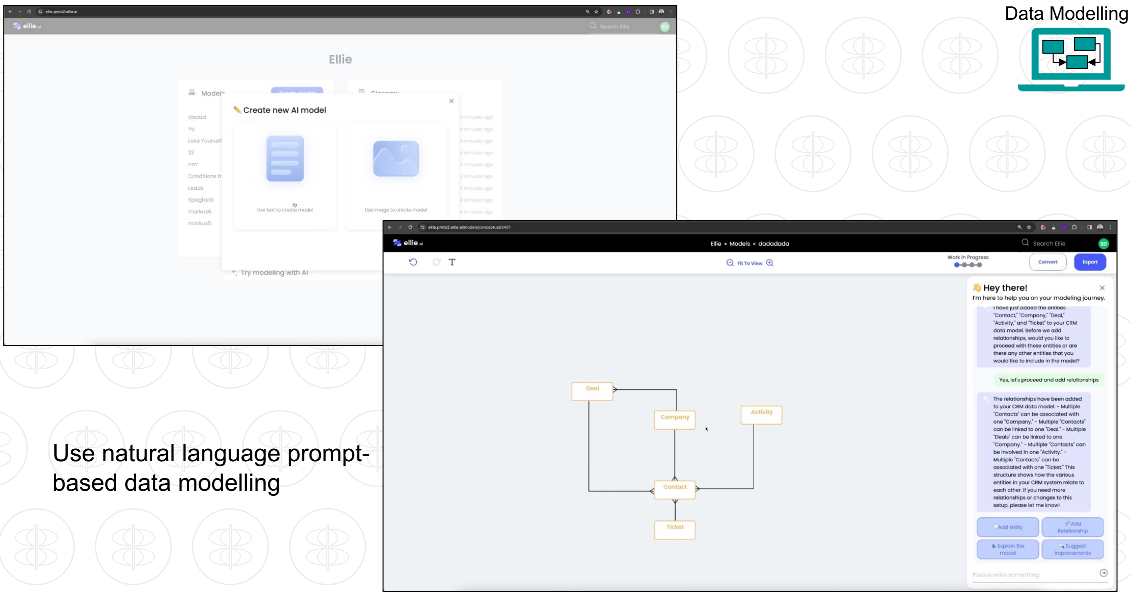
#	CustomerName	CCN	BillingID	SSN
1	John Doe	1234-5678-9012-3456	1000001	123-45-6789
2	Jane Smith	2345-6789-0123-4567	1000002	234-56-7890
3	Bob Johnson	3456-7890-1234-5678	1000003	345-67-8901
4	Samantha Lee	4567-8901-2345-6789	1000004	456-78-9012
5	David Kim	5678-9012-3456-7890	1000005	567-89-0123
6	Emily Chen	6789-0123-4567-8901	1000006	678-90-1234
7	Michael Brown	7890-1234-5678-9012	1000007	789-01-2345
8	Karen Davis	8901-2345-6789-0123	1000008	890-12-3456
9	Tom Wilson	9012-3456-7890-1234	1000009	901-23-4567
10	Lisa Nguyen	0123-4567-8901-2345	1000010	012-34-5678

Data catalog

Source: Atlan

Generative AI And Data Modelling - Ellie.ai Is a Popular Data Modelling Tool That Can Use Generative AI to Help Build Data Models

Data Modelling



Hey there!
I'm here to help you on your modeling journey.

I have just added the entities "Contact," "Company," "Deal," "Activity," and "Ticket" to your CRM data model. Before we add relationships, would you like to proceed with these entities or are there any other entities that you would like to include in the model?

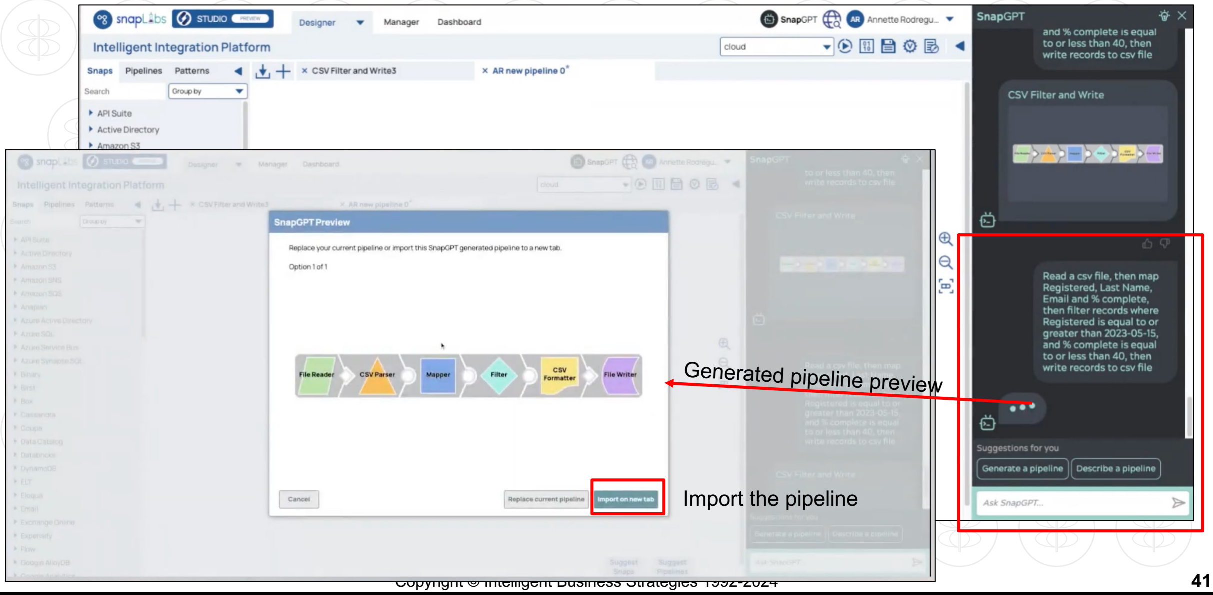
Yes, let's proceed and add relationships

The relationships have been added to your CRM data model: - Multiple "Contacts" can be associated with one "Company." - Multiple "Contacts" can be linked to one "Deal." - Multiple "Deals" can be linked to one "Company." - Multiple "Contacts" can be involved in one "Activity." - Multiple "Contacts" can be associated with one "Ticket." This structure shows how the various entities in your CRM system relate to each other. If you need more relationships or changes to this setup, please let me know!

Use natural language prompt-based data modelling

Generative AI In Data Engineering – SnapLogic SnapGPT Prompt Based Data Engineering

Data engineering



SnapGPT

and % complete is equal to or less than 40, then write records to csv file

SnapGPT Preview

Replace your current pipeline or import this SnapGPT generated pipeline to a new tab.

Option 1 of 1

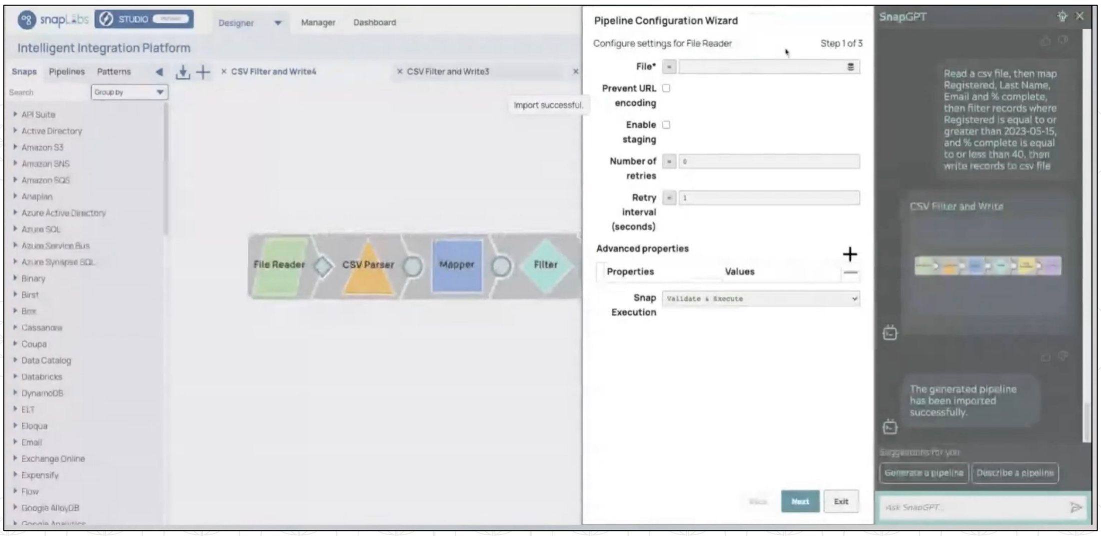
Read a csv file, then map Registered, Last Name, Email and % complete, then filter records where Registered is equal to or greater than 2023-05-15, and % complete is equal to or less than 40, then write records to csv file

Generated pipeline preview

Import the pipeline

Generative AI Prompt Based Data Engineering – SnapLogic SnapGPT Pipeline Configuration Wizard

Data engineering



Import successful.

Pipeline Configuration Wizard

Configure settings for File Reader

Step 1 of 3

File*

Prevent URL encoding ☐

Enable staging ☐

Number of retries: 0

Retry interval (seconds): 1

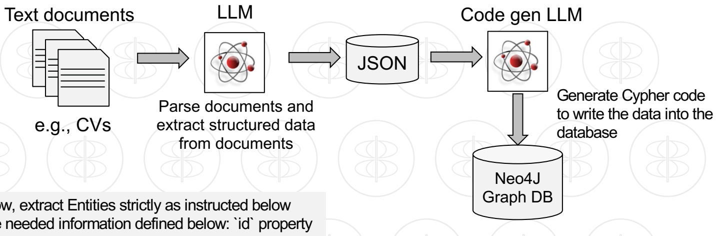
Snap Execution: Validate & Execute

Read a csv file, then map Registered, Last Name, Email and % complete, then filter records where Registered is equal to or greater than 2023-05-15, and % complete is equal to or less than 40, then write records to csv file

The generated pipeline has been imported successfully.

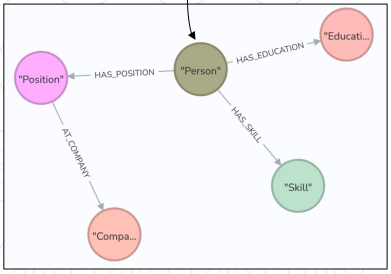
Source: SnapLogic

Prompt Based Data Engineering Example
– Parsing Text, Extracting Data And Populating A Graph Database Using Gen AI


Text documents
e.g., CVs
LLM
Parse documents and extract structured data from documents
JSON
Code gen LLM
Generate Cypher code to write the data into the database
Neo4J Graph DB

Prompt to extract data from text for the Person node

```
person_prompt_tpl="""From the Resume text for a job aspirant below, extract Entities strictly as instructed below
1.     First, look for the Person Entity type in the text and extract the needed information defined below: `id` property
       of each entity must be alphanumeric and must be unique among the entities. You will be referring this property
       to define the relationship between entities. NEVER create new entity types that aren't mentioned below.
       Document must be summarized and stored inside Person entity under `description` property
        Entity Types:
        label:'Person',id:string,role:string,description:string //Person Node
2. Description property should be a crisp text summary and MUST NOT be more than 100 characters
3. If you cannot find any information on the entities & relationships above, it is okay to return empty value. DO NOT
create fictious data
4. Do NOT create duplicate entities
5. Restrict yourself to extract only Person information. No Position, Company, Education or Skill information should
be focussed.
6. NEVER Impute missing values Example Output JSON: {"entities": [{"label":"Person","id":"person1","role":"Prompt
Developer","description":"Prompt Developer with more than 30 years of LLM experience"}]}

Question: Now, extract the Person for the text below –
$ctext
Answer:
"""
```



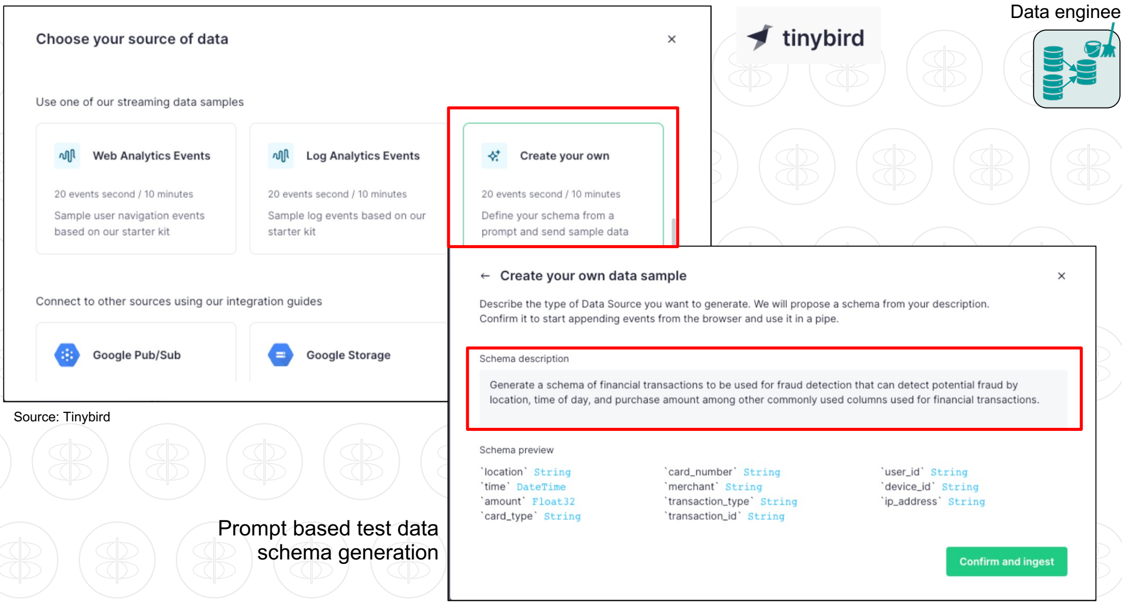
Source: https://github.com/neo4j-partners/neo4j-generative-ai-google-cloud

Copyright © Intelligent Business Strategies 1992-2024

Generative AI In Data Engineering - Synthetic Data Generation Using LLMs
e.g., Tinybird Mockingbird – Generate A Schema And Synthetic Streaming Test Data

tinybird

Data engineering



Choose your source of data

Use one of our streaming data samples

Web Analytics Events
20 events second / 10 minutes
Sample user navigation events based on our starter kit

Log Analytics Events
20 events second / 10 minutes
Sample log events based on our starter kit

Create your own
20 events second / 10 minutes
Define your schema from a prompt and send sample data

Connect to other sources using our integration guides

Google Pub/Sub

Google Storage

← Create your own data sample

Describe the type of Data Source you want to generate. We will propose a schema from your description. Confirm it to start appending events from the browser and use it in a pipe.

Schema description

Generate a schema of financial transactions to be used for fraud detection that can detect potential fraud by location, time of day, and purchase amount among other commonly used columns used for financial transactions.

Schema preview

```
`location` String        `card_number` String        `user_id` String
`time` DateTime          `merchant` String           `device_id` String
`amount` Float32         `transaction_type` String   `ip_address` String
`card_type` String       `transaction_id` String
```

Source: Tinybird

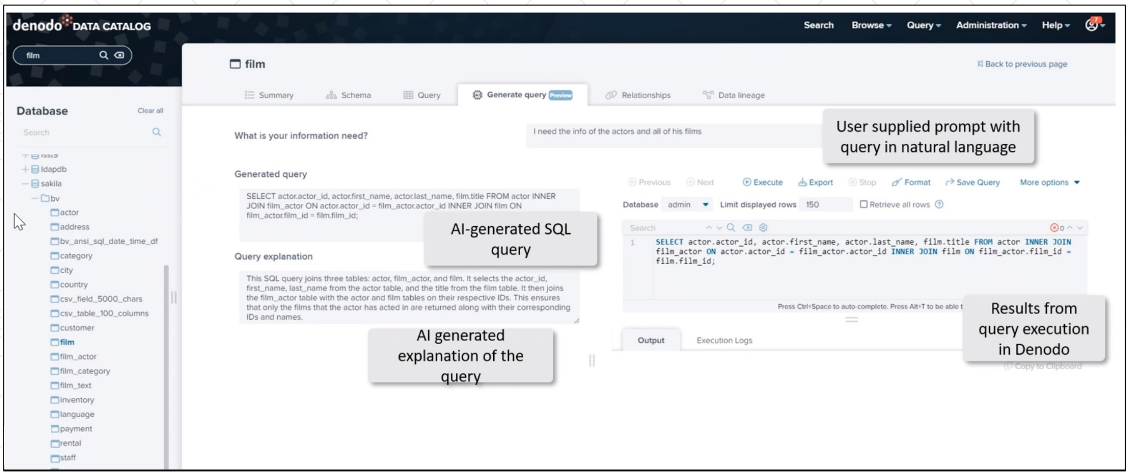
Prompt based test data schema generation

Copyright © Intelligent Business Strategies 1992-2024

Source: Tinybird

Copyright © Intelligent Business Strategies 1992-2024

Generative AI In Data Virtualisation – Product Example: Denodo NLP Queries And LLMs



film

What is your information need?

I need the info of the actors and all of his films

Generated query

SELECT actor.actor_id, actor.first_name, actor.last_name, film.title FROM actor INNER JOIN film_actor ON actor.actor_id = film_actor.actor_id INNER JOIN film ON film_actor.film_id = film.film_id;

Query explanation

This SQL query joins three tables: actor, film_actor, and film. It selects the actor_id, first_name, last_name from the actor table, and the title from the film table. It then joins the film_actor table with the actor and film tables on their respective IDs. This ensures that only the films that the actor has acted in are returned along with their corresponding IDs and names.

Database: admin
Limit displayed rows: 150
☐ Retrieve all rows

```
SELECT actor.actor_id, actor.first_name, actor.last_name, film.title FROM actor INNER JOIN
film_actor ON actor.actor_id = film_actor.actor_id INNER JOIN film ON film_actor.film_id =
film.film_id;
```

Press Ctrl+Space to auto-complete. Press Alt+T to be able t

User supplied prompt with query in natural language

AI-generated SQL query

AI generated explanation of the query

Results from query execution in Denodo

Source: Denodo

Generative AI In The Database – SQL Generation Using LLMs In Oracle Autonomous Database

DBMS

Enabled using the new **DBMS_CLOUD_AI** PL/SQL package

New AI keyword

```
SELECT AI
What are our top 10 streamed movies that were released after 2018
```

Automated SQL query generation pipeline

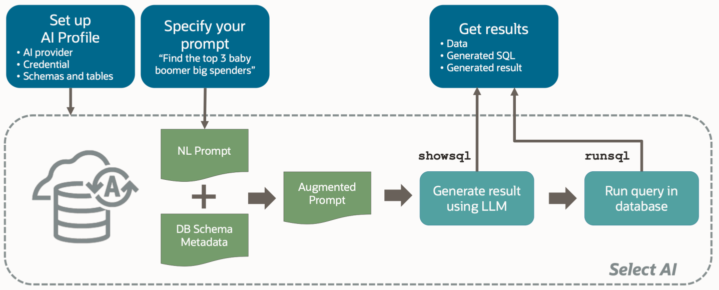
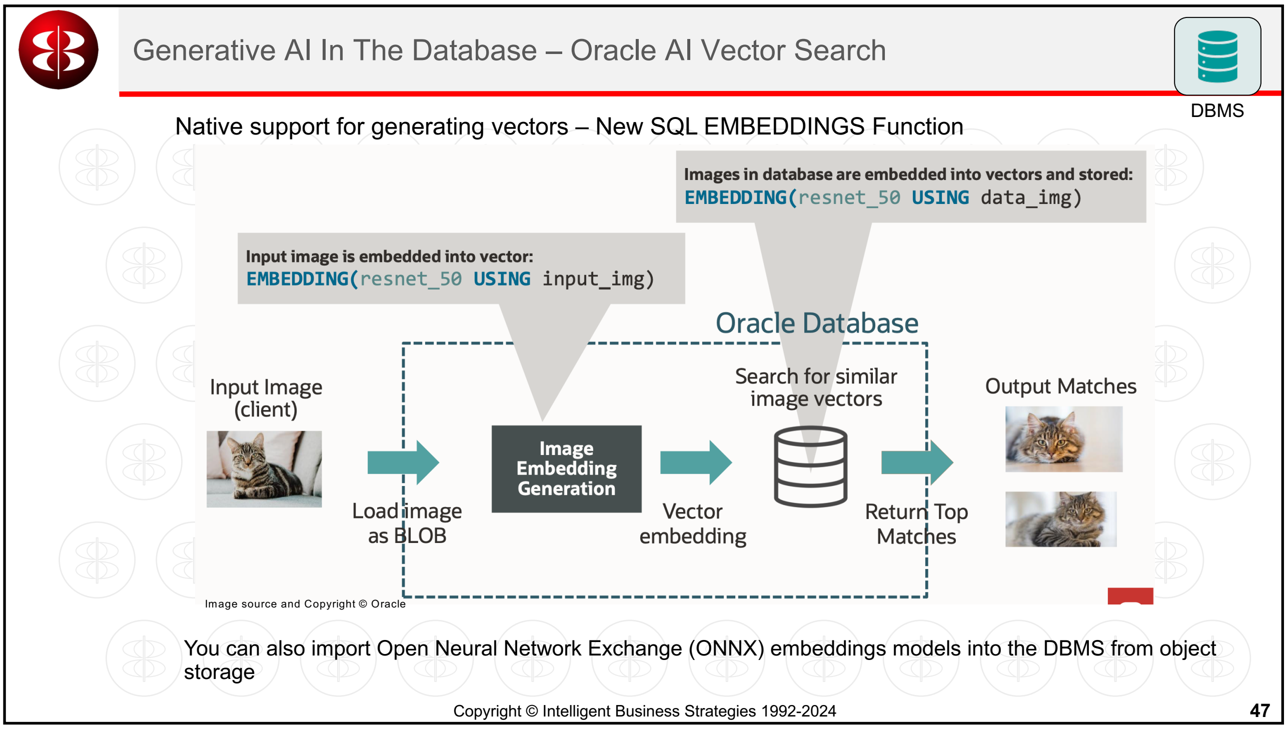


Image source and copyright © Oracle

Actions
runsql - return the SQL result set (default)
showsql – return the generated query
narrate – return a conversational result
chat - general AI chat

Use generative AI models in combination with your database data
You can also combine relational and vector data in the same query



Generative AI In The Database – Oracle AI Vector Search

Native support for generating vectors – New SQL EMBEDDINGS Function

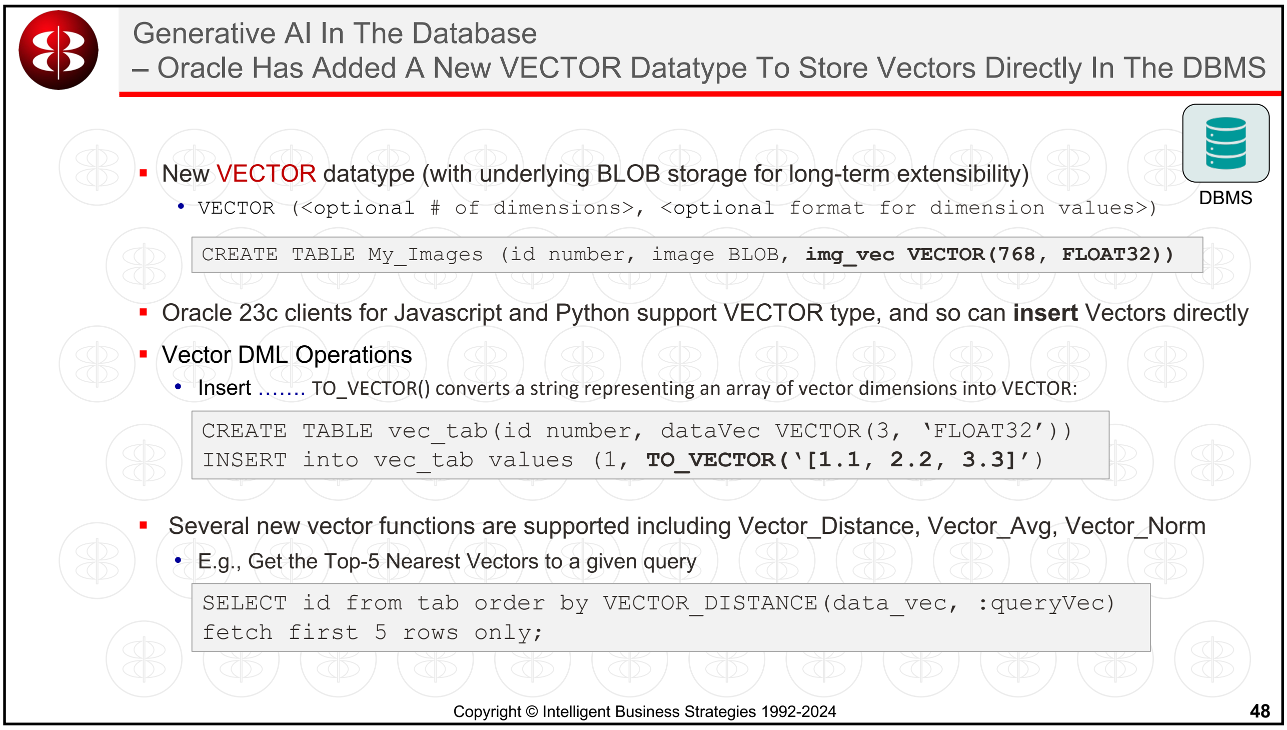
Images in database are embedded into vectors and stored:
`EMBEDDING(resnet_50 USING data_img)`

Input image is embedded into vector:
`EMBEDDING(resnet_50 USING input_img)`

Image source and Copyright © Oracle

You can also import Open Neural Network Exchange (ONNX) embeddings models into the DBMS from object storage

Copyright © Intelligent Business Strategies 1992-2024 47



Generative AI In The Database – Oracle Has Added A New VECTOR Datatype To Store Vectors Directly In The DBMS

- New VECTOR datatype (with underlying BLOB storage for long-term extensibility)
  - `VECTOR (<optional # of dimensions>, <optional format for dimension values>)`

```
CREATE TABLE My_Images (id number, image BLOB, img_vec VECTOR(768, FLOAT32))
```

- Oracle 23c clients for Javascript and Python support VECTOR type, and so can **insert** Vectors directly
- Vector DML Operations
  - Insert ……. TO_VECTOR() converts a string representing an array of vector dimensions into VECTOR:

```
CREATE TABLE vec_tab(id number, dataVec VECTOR(3, ‘FLOAT32’))
INSERT into vec_tab values (1, TO_VECTOR(‘[1.1, 2.2, 3.3]’)
```

- Several new vector functions are supported including Vector_Distance, Vector_Avg, Vector_Norm
  - E.g., Get the Top-5 Nearest Vectors to a given query

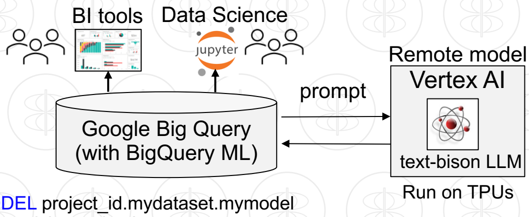
```
SELECT id from tab order by VECTOR_DISTANCE(data_vec, :queryVec)
fetch first 5 rows only;
```

Copyright © Intelligent Business Strategies 1992-2024 48

Generative AI In The Database – Google Duet AI In BigQuery AND BigQuery ML Which Has Been Extended To Use LLMs In SQL Queries

DBMS

Duet AI in Big Query enables you to:
- **Generate** a SQL query
- **Complete** a SQL query
- **Explain** a SQL query



Uses
- Classification
- Sentiment Analysis
- Entity extraction
- Extractive Question Answering
- Summarization
- Re-writing text in a different style
- Ad copy generation
- Concept ideation

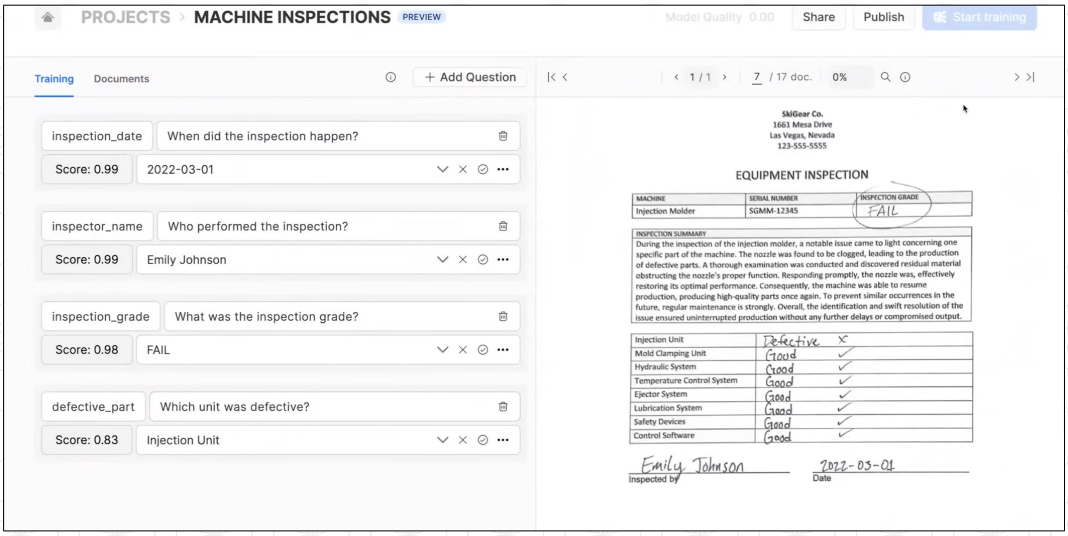
```
CREATE MODEL project_id.mydataset.mymodel
REMOTE WITH CONNECTION `myproject.us.test_connection`
OPTIONS(REMOTE_SERVICE_TYPE="CLOUD_AI_LARGE_LANGUAGE_MODEL_V1")

SELECT * FROM ML.GENERATE_TEXT( MODEL mydataset.llm_model TABLE mydataset.prompt_table,
STRUCT( 0.2 AS temperature, 75 AS max_output_tokens, 0.3 AS top_p, 15 AS top_k, TRUE AS flatten_json_output));
```

- Provides prompt data from a table column that's named prompt
- Returns a shorter generated text response
- Returns a more probable generated text response
- Flattens the JSON response into separate columns

Copyright © Intelligent Business Strategies 1992-2024

Generative AI In The Database – Use Natural Language Queries On Unstructured Documents Via Snowflake Document AI

DBMS



inspection_date	When did the inspection happen?	
Score: 0.99	2022-03-01	
inspector_name	Who performed the inspection?	
Score: 0.99	Emily Johnson	
inspection_grade	What was the inspection grade?	
Score: 0.98	FAIL	
defective_part	Which unit was defective?	
Score: 0.83	Injection Unit	

SkiGear Co.
1661 Mesa Drive
Las Vegas, Nevada
123-555-5555

EQUIPMENT INSPECTION

MACHINE	SERIAL NUMBER	INSPECTION GRADE
Injection Molder	SGMM-12345	FAIL

INSPECTION SUMMARY

During the inspection of the injection molder, a notable issue came to light concerning one specific part of the machine. The nozzle was found to be clogged, leading to the production of defective parts. A thorough examination was conducted and discovered residual material obstructing the nozzle's proper function. Responding promptly, the nozzle was, effectively restoring its optimal performance. Consequently, the machine was able to resume production, producing high-quality parts once again. To prevent similar occurrences in the future, regular maintenance is strongly. Overall, the identification and swift resolution of the issue ensured uninterrupted production without any further delays or compromised output.

Injection Unit	Defective	✗
Mold Clamping Unit	Good	✓
Hydraulic System	Good	✓
Temperature Control System	Good	✓
Ejector System	Good	✓
Lubrication System	Good	✓
Safety Devices	Good	✓
Control Software	Good	✓

Emily Johnson
Inspected by

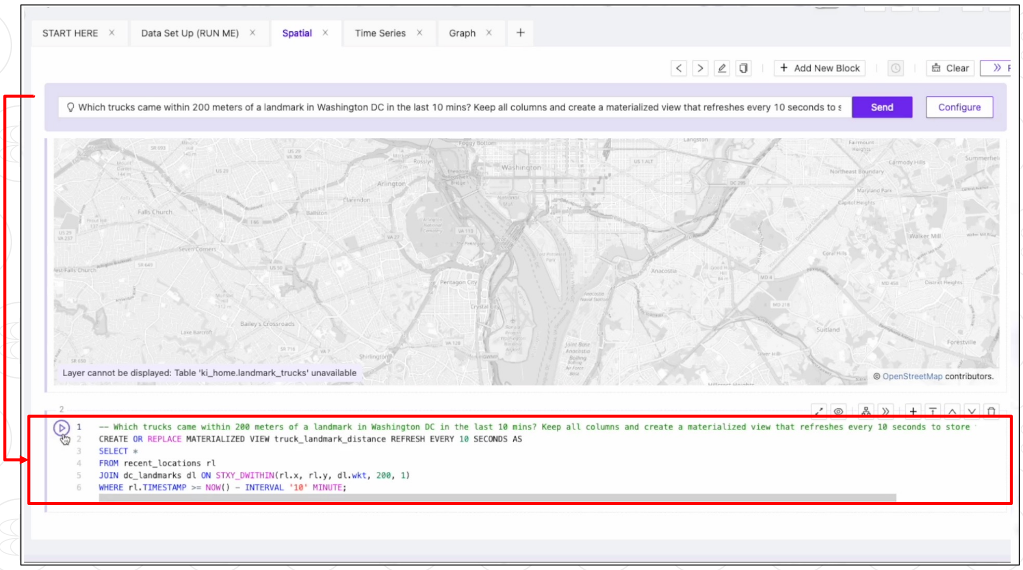
2022-03-01
Date

Source: Snowflake

Copyright © Intelligent Business Strategies 1992-2024

Copyright © Intelligent Business Strategies 1992-2024

Generative AI In The DBMS – Natural Language Queries In Kinetica DBMS Become AI-Generated Geospatial SQL Queries



Which trucks came within 200 meters of a landmark in Washington DC in the last 10 mins? Keep all columns and create a materialized view that refreshes every 10 seconds to s

Layer cannot be displayed: Table 'ki_home.landmark_trucks' unavailable

```
-- Which trucks came within 200 meters of a landmark in Washington DC in the last 10 mins? Keep all columns and create a materialized view that refreshes every 10 seconds to store
CREATE OR REPLACE MATERIALIZED VIEW truck_landmark_distance REFRESH EVERY 10 SECONDS AS
SELECT *
FROM recent_locations rl
JOIN dc_landmarks dl ON STXY_DWITHIN(rl.x, rl.y, dl.wkt, 200, 1)
WHERE rl.TIMESTAMP >= NOW() - INTERVAL '10' MINUTE;
```

DBMS

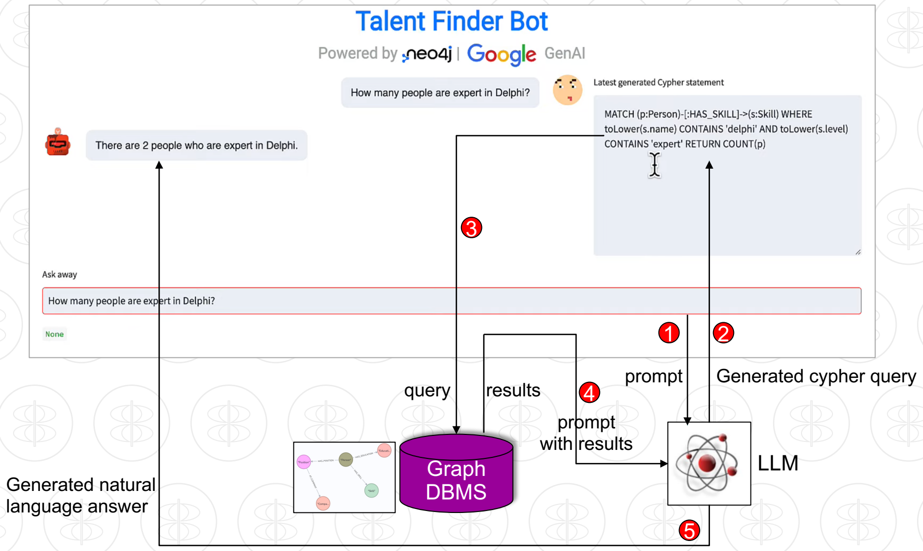
Note that LLMs can run in Kinetica because this is a MPP SQL DBMS running on GPUs

Source: Kinetica

Copyright © Intelligent Business Strategies 1992-2024

Generative AI And Graph Databases – Natural Language Queries And Natural Language Responses In Neo4j - Generates Cypher Queries From Text

DBMS



How many people are expert in Delphi?

There are 2 people who are expert in Delphi.

```
MATCH (p:Person)-[:HAS_SKILL]->(s:Skill) WHERE
toLower(s.name) CONTAINS 'delphi' AND toLower(s.level)
CONTAINS 'expert' RETURN COUNT(p)
```

How many people are expert in Delphi?

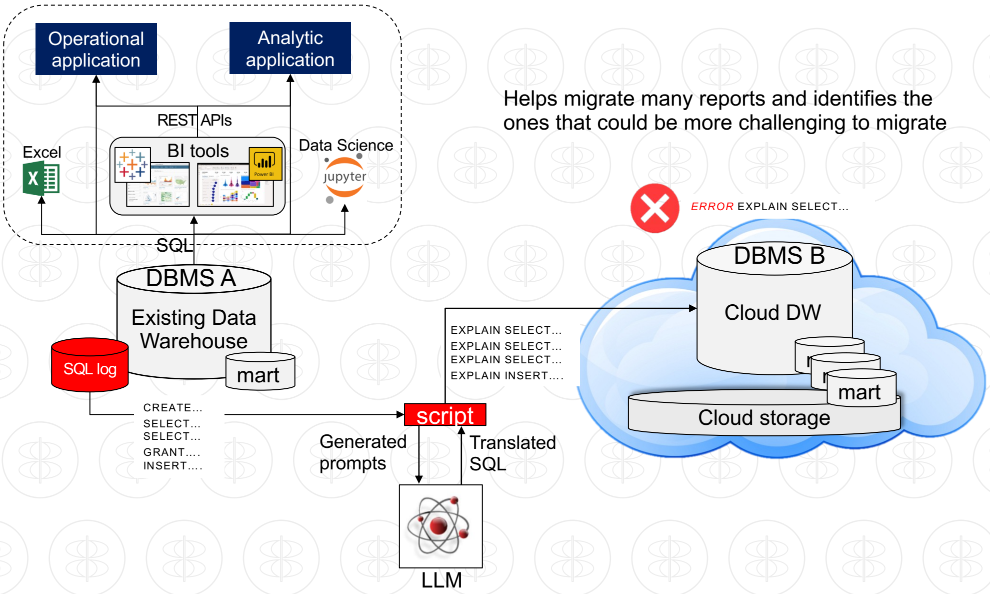
Copyright © Intelligent Business Strategies 1992-2024

Copyright © Intelligent Business Strategies 1992-2024

Generative AI In DW Migration – Automated SQL Translation To Your New DBMS

Database Migration

Helps migrate many reports and identifies the ones that could be more challenging to migrate



Copyright © Intelligent Business Strategies 1992-2024

Generative AI In Data Governance

Data Governance

- Provide a generative AI-Assistant to data stewards

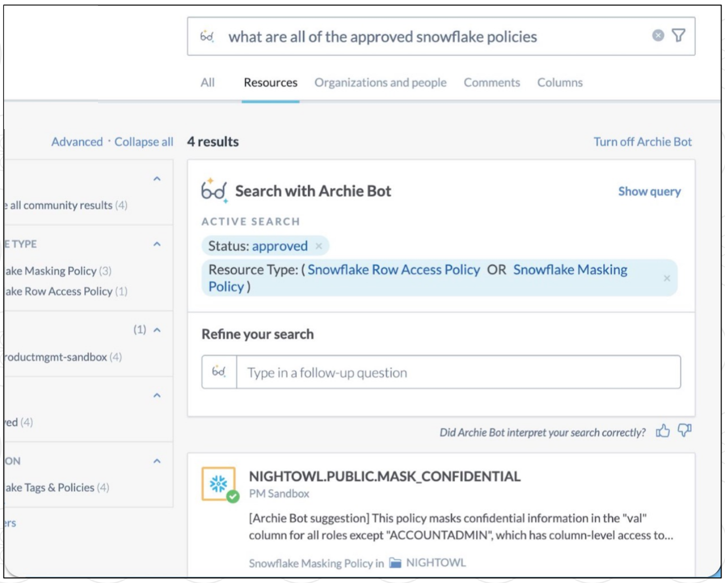
Data Governance Discipline	Auto AI-Generation Of
Data Quality	Data quality validation check rules
	Data quality cleansing rules and code
	Synthetic data to replace missing values in data to improve data completeness
Master Data Management	Data matching rules
	Data survivorship rules
	Master data descriptions
	Product information e.g., digital assets - catalogs, product brief, marketing content
Data privacy	Data privacy policies
	Synthetic data that excludes PII data to avoid risks
Data access security	Data access security policies
	Data loss prevention policies
Data sharing	Data sharing policies
Data retention	Data retention policies

Copyright © Intelligent Business Strategies 1992-2024

Copyright © Intelligent Business Strategies 1992-2024

Generative AI In Data Governance – Prompt Based Discovery Of Policies In Source Systems Scanned By The Catalog In Data.world

Data Governance



4 results

Search with Archie Bot

ACTIVE SEARCH

Status: approved

Resource Type: (Snowflake Row Access Policy OR Snowflake Masking Policy)

Refine your search

Did Archie Bot interpret your search correctly?

NIGHTOWL.PUBLIC.MASK_CONFIDENTIAL

PM Sandbox

[Archie Bot suggestion] This policy masks confidential information in the "val" column for all roles except "ACCOUNTADMIN", which has column-level access to...

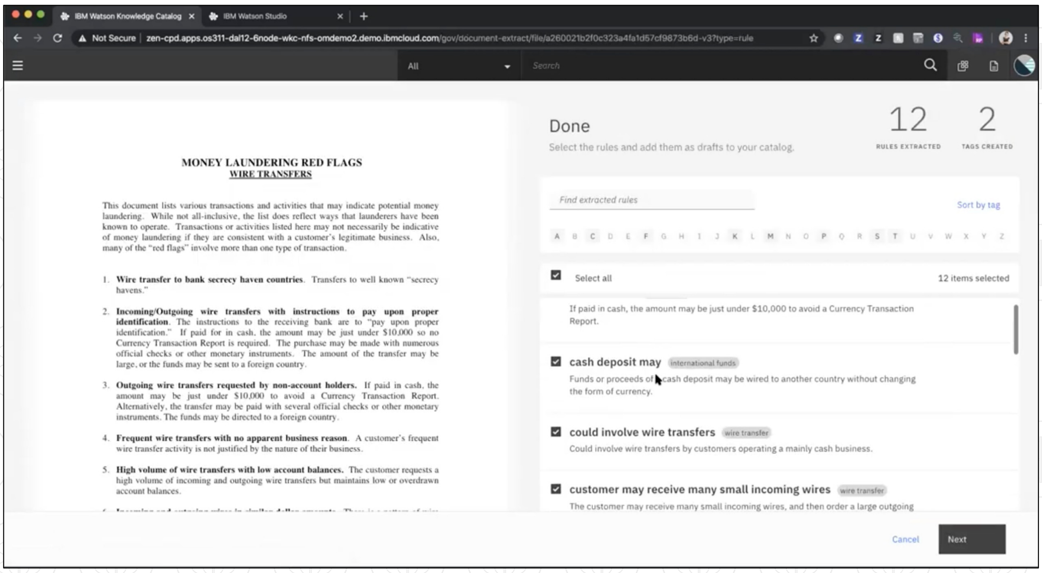
Snowflake Masking Policy in NIGHTOWL

Source: Data.world

Copyright © Intelligent Business Strategies 1992-2024

Generative AI In Data Governance – It Is Also Possible To Automatically Extract And Generate Data Governance Rules From Documents, E.g., IBM Watson Knowledge Catalog

Data Governance


IBM Watson Knowledge Catalog
IBM Watson Studio
Not Secure | zen-cpd.apps.os311-dal12-6node-wkc-nfs-omdemo2.demo.ibmcloud.com/gov/document-extract/file/a260021b2f0c323a4fa1d57cf9873b6d-v3?type=rule
All
Search
Find extracted rules
Sort by tag
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
Cancel
Next

MONEY LAUNDERING RED FLAGS
WIRE TRANSFERS

This document lists various transactions and activities that may indicate potential money laundering. While not all-inclusive, the list does reflect ways that launderers have been known to operate. Transactions or activities listed here may not necessarily be indicative of money laundering if they are consistent with a customer's legitimate business. Also, many of the "red flags" involve more than one type of transaction.

1. **Wire transfer to bank secrecy haven countries**. Transfers to well known "secrecy havens."
2. **Incoming/Outgoing wire transfers with instructions to pay upon proper identification**. The instructions to the receiving bank are to "pay upon proper identification." If paid for in cash, the amount may be just under $10,000 so no Currency Transaction Report is required. The purchase may be made with numerous official checks or other monetary instruments. The amount of the transfer may be large, or the funds may be sent to a foreign country.
3. **Outgoing wire transfers requested by non-account holders.** If paid in cash, the amount may be just under $10,000 to avoid a Currency Transaction Report. Alternatively, the transfer may be paid with several official checks or other monetary instruments. The funds may be directed to a foreign country.
4. **Frequent wire transfers with no apparent business reason**. A customer's frequent wire transfer activity is not justified by the nature of their business.
5. **High volume of wire transfers with low account balances.** The customer requests a high volume of incoming and outgoing wire transfers but maintains low or overdrawn account balances.

Done

Select the rules and add them as drafts to your catalog.

12 RULES EXTRACTED

2 TAGS CREATED

☑ Select all — 12 items selected

If paid in cash, the amount may be just under $10,000 to avoid a Currency Transaction Report.

☑ **cash deposit may** international funds
Funds or proceeds of cash deposit may be wired to another country without changing the form of currency.

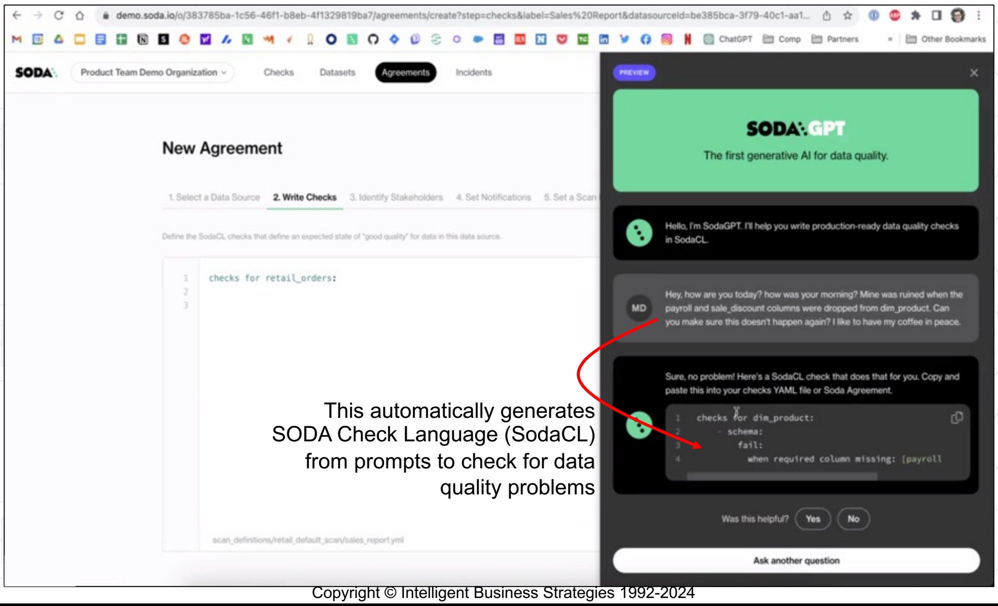
☑ **could involve wire transfers** wire transfer
Could involve wire transfers by customers operating a mainly cash business.

☑ **customer may receive many small incoming wires** wire transfer
The customer may receive many small incoming wires, and then order a large outgoing

Copyright © Intelligent Business Strategies 1992-2024

Copyright © Intelligent Business Strategies 1992-2024

Generative AI In Data Governance – SodaGPT Lets Business Analysts Create Data Quality Checks In Natural Language That Can Be Shared With Data Engineers To Fix Data



demo.soda.io/o/383785ba-1c56-46f1-b8eb-4f1329819ba7/agreements/create?step=checks&label=Sales%20Report&datasourceId=be385bca-3f79-40c1-aa1...

New Agreement

Define the SodaCL checks that define an expected state of "good quality" for data in this data source.

```
checks for retail_orders:
```

scan_definitions/retail_default_scan/sales_report.yml

The first generative AI for data quality.

Hello, I'm SodaGPT. I'll help you write production-ready data quality checks in SodaCL.

MD: Hey, how are you today? how was your morning? Mine was ruined when the payroll and sale_discount columns were dropped from dim_product. Can you make sure this doesn't happen again? I like to have my coffee in peace.

Sure, no problem! Here's a SodaCL check that does that for you. Copy and paste this into your checks YAML file or Soda Agreement.

```
checks for dim_product:
  - schema:
      fail:
        when required column missing: [payroll
```

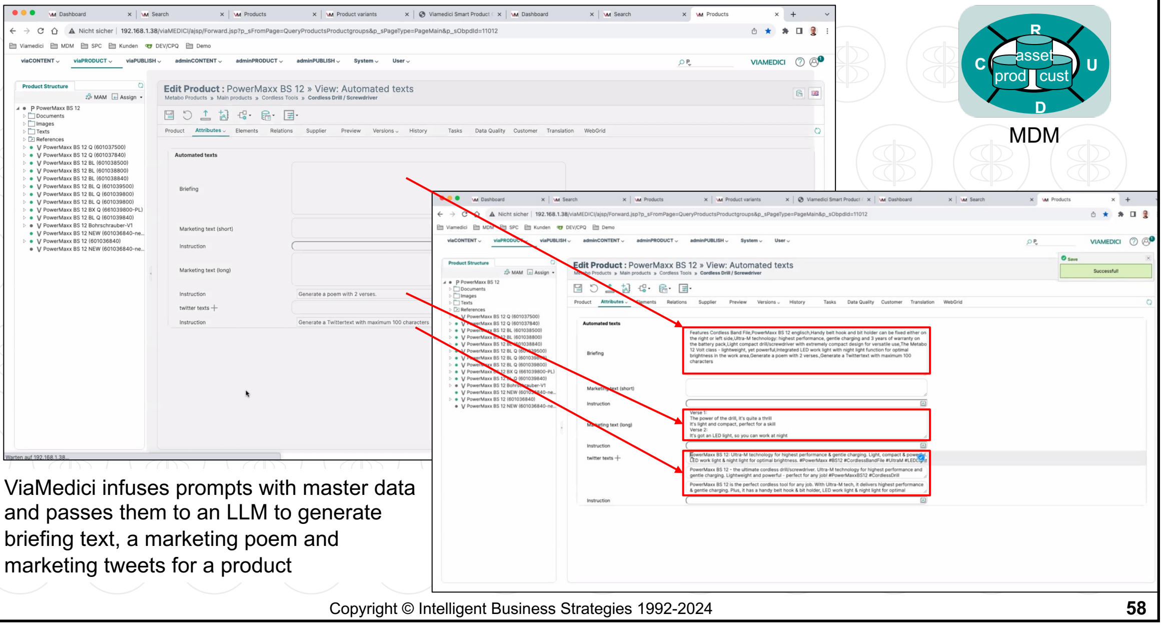
Was this helpful? Yes No

This automatically generates SODA Check Language (SodaCL) from prompts to check for data quality problems

Data Governance

SodaCL is a low code abstract language built on SQL and Spark

Generative AI In Master Data Management (MDM) – Product Example: ViaMedici Product MDM & PIM AI-Powered Text Generation



192.168.1.38/viaMEDICI/ajsp/Forward.jsp?p_sFromPage=QueryProductsProductgroups&p_sPageType=PageMain&p_sObpdId=11012

Edit Product : PowerMaxx BS 12 » View: Automated texts

Metabo Products » Main products » Cordless Tools » Cordless Drill / Screwdriver

Briefing: Features Cordless Band File,PowerMaxx BS 12 englisch,Handy belt hook and bit holder can be fixed either on the right or left side,Ultra-M technology: highest performance, gentle charging and 3 years of warranty on the battery pack,Light compact drill/screwdriver with extremely compact design for versatile use,The Metabo 12 Volt class - lightweight, yet powerful,Integrated LED work light with night light function for optimal brightness in the work area,Generate a poem with 2 verses.,Generate a Twittertext with maximum 100 characters

Marketing text (short):

Instruction:

Marketing text (long):
Verse 1:
The power of the drill, it's quite a thrill
It's light and compact, perfect for a skill
Verse 2:
It's got an LED light, so you can work at night

Instruction: Generate a poem with 2 verses.

twitter texts:
PowerMaxx BS 12: Ultra-M technology for highest performance & gentle charging. Light, compact & powerful. LED work light & night light for optimal brightness. #PowerMaxx #BS12 #CordlessBandFile #UltraM #LEDLight
PowerMaxx BS 12 - the ultimate cordless drill/screwdriver. Ultra-M technology for highest performance and gentle charging. Lightweight and powerful - perfect for any job! #PowerMaxxBS12 #CordlessDrill
PowerMaxx BS 12 is the perfect cordless tool for any job. With Ultra-M tech, it delivers highest performance & gentle charging. Plus, it has a handy belt hook & bit holder, LED work light & night light for optimal

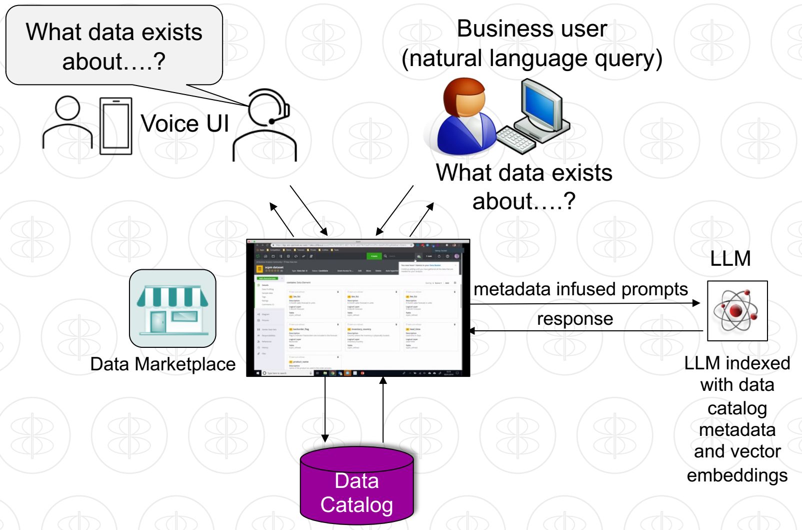
Instruction: Generate a Twittertext with maximum 100 characters

ViaMedici infuses prompts with master data and passes them to an LLM to generate briefing text, a marketing poem and marketing tweets for a product

MDM

Generative AI In Data Marketplaces

Data Marketplace



What data exists about….?

Voice UI

Business user (natural language query)

What data exists about….?

Data Marketplace

metadata infused prompts

response

LLM

LLM indexed with data catalog metadata and vector embeddings

Data Catalog

Copyright © Intelligent Business Strategies 1992-2024

Topics – Where Are We?

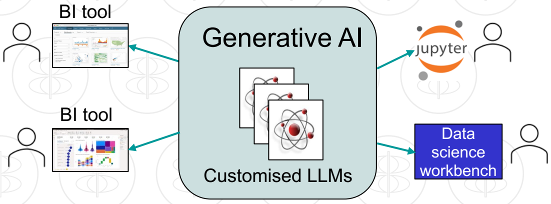
- What is generative AI?
- What are the business benefits of generative AI in data management and analytics?
- How is generative AI being used in data management?
- ➢How is generative AI being used in data science and BI
- What does this mean for business going forward?
- What should you do to get started?

Copyright © Intelligent Business Strategies 1992-2024

Copyright © Intelligent Business Strategies 1992-2024

Generative AI And Business Intelligence And Data Science

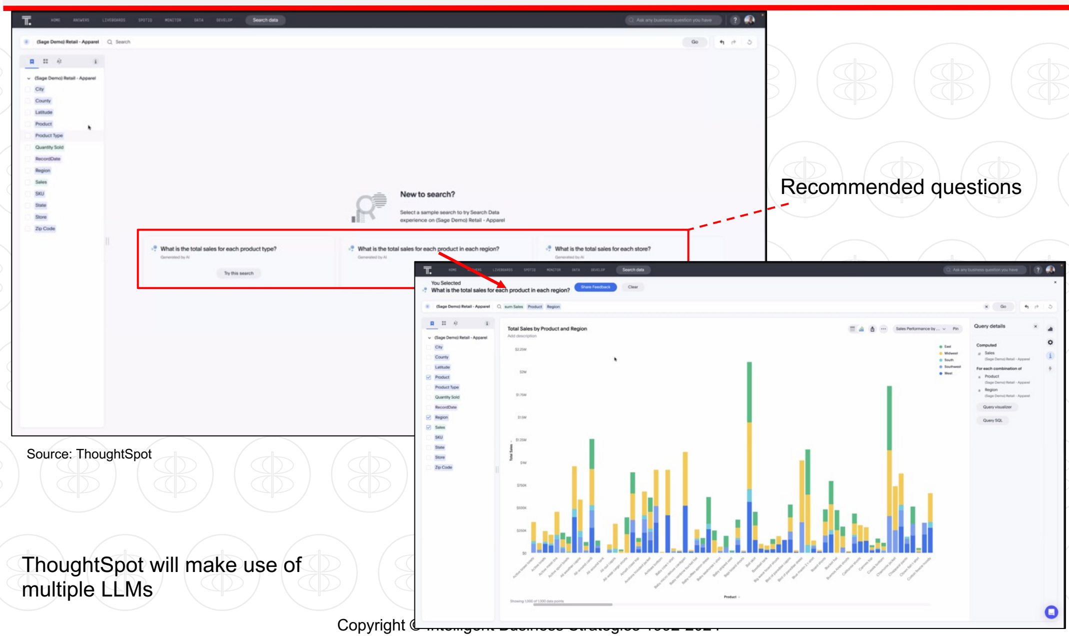
- Gen-AI in business intelligence
  - AI recommended questions from auto-analysis of your data to quick start query & analysis on your data
  - AI-generated queries from voice
  - AI-generated queries from natural language text questions
  - AI generated answers
  - Auto-predict and generate charts from your data
  - Natural language generation to explain insights
  - Metadata enrichment, e.g., generate synonyms for data



- Gen-AI in data science
  - AI-generated code
  - Prompt based querying of data
  - Foundation models and model tuning
  - Building generative AI applications

Copyright © Intelligent Business Strategies 1992-2024

Generative AI In Business Intelligence – Natural Language Questions – E.g., ThoughtSpot Sage Auto Analyses The Data First And Recommends Questions



New to search?

Select a sample search to try Search Data experience on (Sage Demo) Retail - Apparel

What is the total sales for each product type?
Generated by AI

What is the total sales for each product in each region?
Generated by AI

What is the total sales for each store?
Generated by AI

You Selected
What is the total sales for each product in each region?

Total Sales by Product and Region

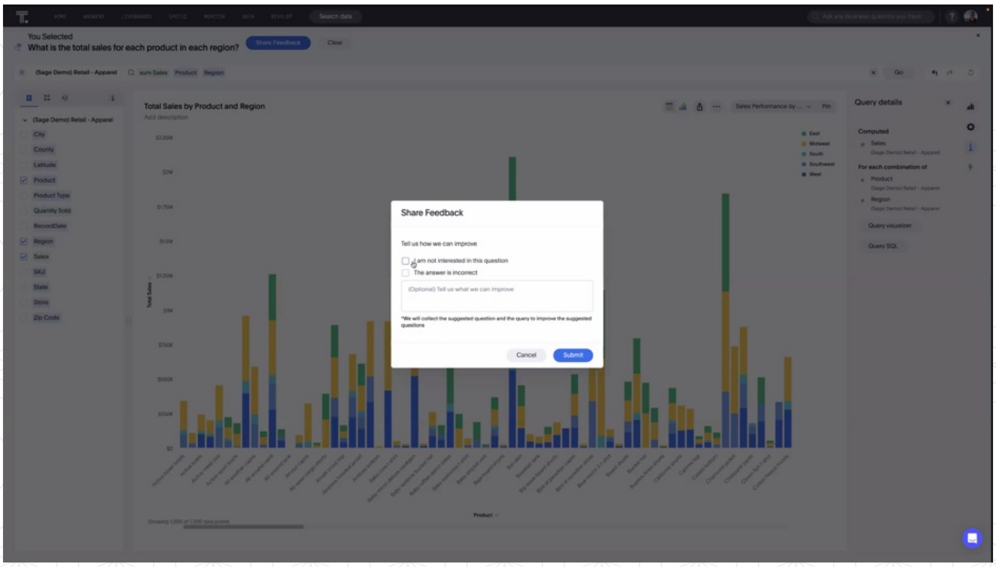
Recommended questions

Source: ThoughtSpot

ThoughtSpot will make use of multiple LLMs

Copyright ©

Generative AI In BI Tools - ThoughtSpot Sage Lets You Provide Feedback To Enable The AI Engine To Improve Question Recommendations It Creates



You Selected
What is the total sales for each product in each region?

Share Feedback

Tell us how we can improve

☐ I am not interested in this question
☐ The answer is incorrect

(Optional) Tell us what we can improve

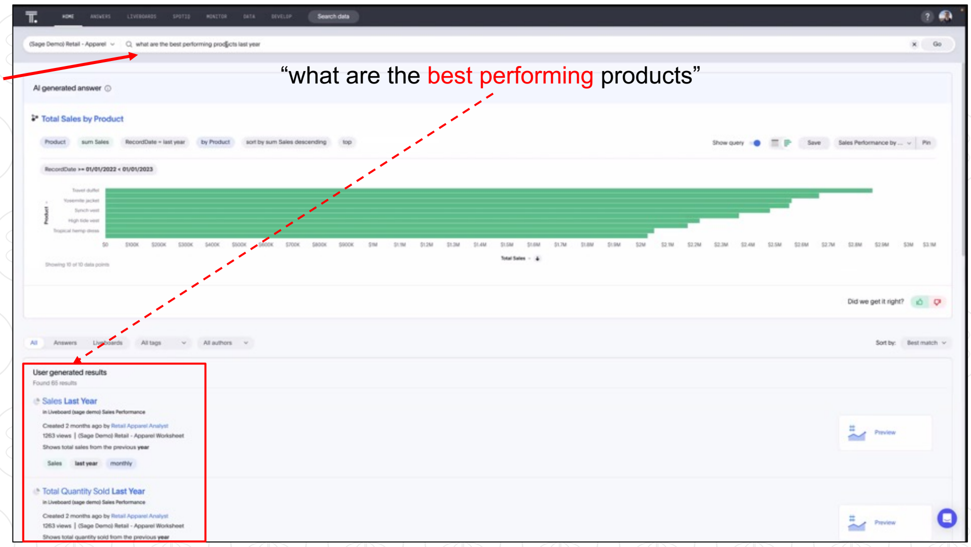
*We will collect the suggested question and the query to improve the suggested questions

Source: ThoughtSpot

You can say you are not interested in a question so that it learns to get better at generating questions you are interested in

Copyright © Intelligent Business Strategies 1992-2024

Generative AI In Business Intelligence – ThoughtSpot Sage Generates Search Queries From Natural Language Text Questions While Dealing With Ambiguity



what are the best performing products last year

AI generated answer

Total Sales by Product

User generated results
Found 65 results

Sales Last Year
in Liveboard (sage demo) Sales Performance
Created 2 months ago by Retail Apparel Analyst
1263 views | (Sage Demo) Retail - Apparel Worksheet
Shows total sales from the previous year

Total Quantity Sold Last Year
in Liveboard (sage demo) Sales Performance
Created 2 months ago by Retail Apparel Analyst
1263 views | (Sage Demo) Retail - Apparel Worksheet
Shows total quantity sold from the previous year

Natural language text question

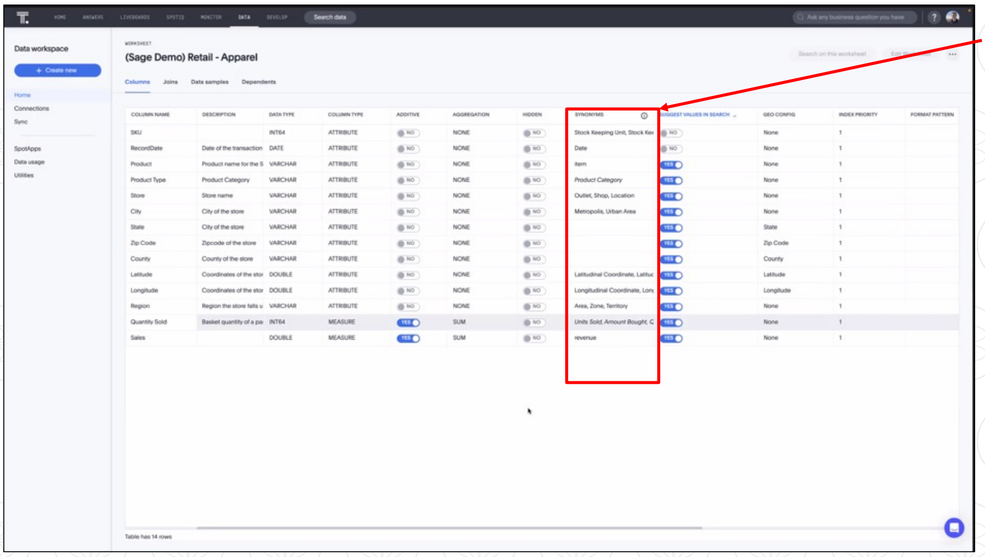
“what are the best performing products”

“best performing” is ambiguous so ThoughtSpot Sage generates intelligence for a number of options

Source: ThoughtSpot

Copyright © Intelligent Business Strategies 1992-2024

Generative AI In Business Intelligence – GPT Generates Synonyms So ThoughtSpot Sage Search Engine Understands Synonyms For The Same Data



WORKSHEET

(Sage Demo) Retail - Apparel

COLUMN NAME	DESCRIPTION	DATA TYPE	COLUMN TYPE	ADDITIVE	AGGREGATION	HIDDEN	SYNONYMS	SUGGEST VALUES IN SEARCH	GEO CONFIG	INDEX PRIORITY	FORMAT PATTERN
SKU		INT64	ATTRIBUTE	NO	NONE	NO	Stock Keeping Unit, Stock Ke	NO	None	1	
RecordDate	Date of the transaction	DATE	ATTRIBUTE	NO	NONE	NO	Date	NO	None	1	
Product	Product name for the S	VARCHAR	ATTRIBUTE	NO	NONE	NO	item	YES	None	1	
Product Type	Product Category	VARCHAR	ATTRIBUTE	NO	NONE	NO	Product Category	YES	None	1	
Store	Store name	VARCHAR	ATTRIBUTE	NO	NONE	NO	Outlet, Shop, Location	YES	None	1	
City	City of the store	VARCHAR	ATTRIBUTE	NO	NONE	NO	Metropolis, Urban Area	YES	None	1	
State	City of the store	VARCHAR	ATTRIBUTE	NO	NONE	NO		YES	State	1	
Zip Code	Zipcode of the store	VARCHAR	ATTRIBUTE	NO	NONE	NO		YES	Zip Code	1	
County	County of the store	VARCHAR	ATTRIBUTE	NO	NONE	NO		YES	County	1	
Latitude	Coordinates of the stor	DOUBLE	ATTRIBUTE	NO	NONE	NO	Latitudinal Coordinate, Latitu	YES	Latitude	1	
Longitude	Coordinates of the stor	DOUBLE	ATTRIBUTE	NO	NONE	NO	Longitudinal Coordinate, Lon	YES	Longitude	1	
Region	Region the store falls u	VARCHAR	ATTRIBUTE	NO	NONE	NO	Area, Zone, Territory	YES	None	1	
Quantity Sold	Basket quantity of a pa	INT64	MEASURE	YES	SUM	NO	Units Sold, Amount Bought, C	YES	None	1	
Sales		DOUBLE	MEASURE	YES	SUM	NO	revenue	YES	None	1	

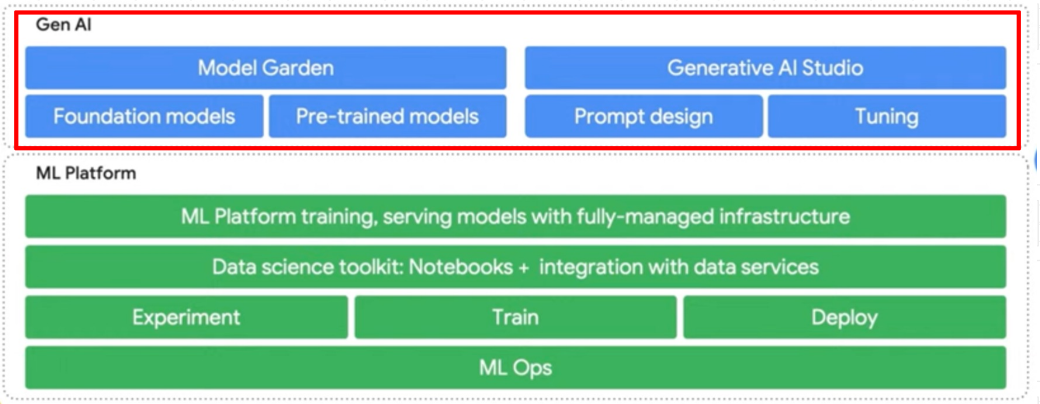
Table has 14 rows

ThoughtSpot Sage queries the schema and then uses GPT to recommend synonyms to the terms in the schema

This allows users to query using synonyms and still get the same answer

Source: ThoughtSpot

Generative AI In Data Science – You Can Now Build Custom LLMs In Data Science Workbenches – E.g., Google Vertex AI Now Includes A Generative AI Studio



Gen AI

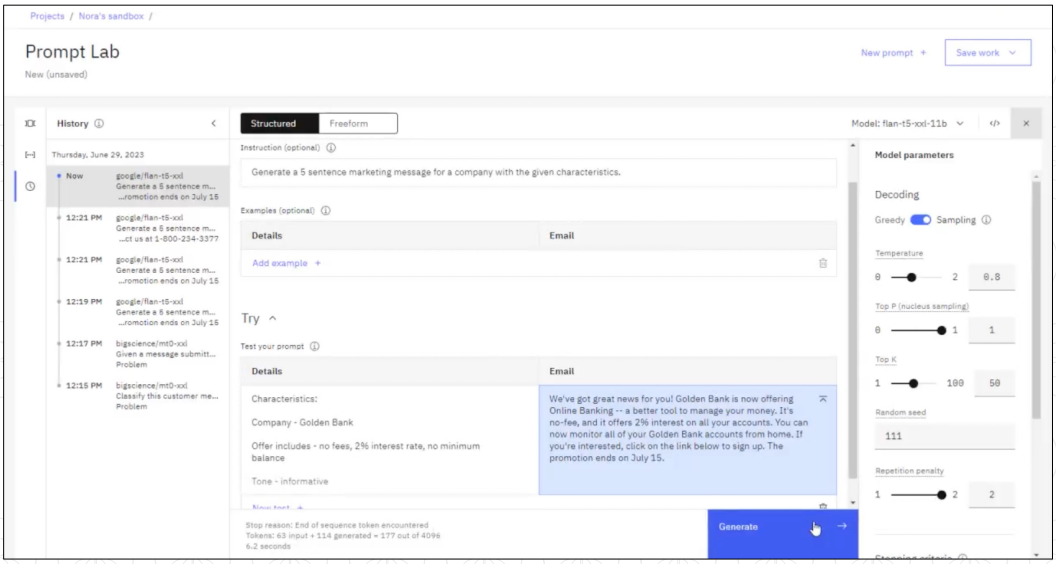
Model Garden
Foundation models
Pre-trained models
Generative AI Studio
Prompt design
Tuning

ML Platform

ML Platform training, serving models with fully-managed infrastructure
Data science toolkit: Notebooks + integration with data services
Experiment
Train
Deploy
ML Ops

Source: Google

Generative AI In Data Science – IBM Watsonx.ai Data Science Workbench Allow You To Build, And Tune LLMs And Make Use Of A Prompt Lab To Work With LLMs



Thursday, June 29, 2023

- Now google/flan-t5-xxl Generate a 5 sentence m... ...romotion ends on July 15
- 12:21 PM google/flan-t5-xxl Generate a 5 sentence m... ...ct us at 1-800-234-3377
- 12:21 PM google/flan-t5-xxl Generate a 5 sentence m... ...romotion ends on July 15
- 12:19 PM google/flan-t5-xxl Generate a 5 sentence m... ...romotion ends on July 15
- 12:17 PM bigscience/mt0-xxl Given a message submitt... Problem
- 12:15 PM bigscience/mt0-xxl Classify this customer me... Problem

Instruction (optional)

Generate a 5 sentence marketing message for a company with the given characteristics.

Examples (optional)

Details	Email

Test your prompt

Details	Email
Characteristics: Company - Golden Bank Offer includes - no fees, 2% interest rate, no minimum balance Tone - informative	We've got great news for you! Golden Bank is now offering Online Banking -- a better tool to manage your money. It's no-fee, and it offers 2% interest on all your accounts. You can now monitor all of your Golden Bank accounts from home. If you're interested, click on the link below to sign up. The promotion ends on July 15.

Stop reason: End of sequence token encountered
Tokens: 63 input + 114 generated = 177 out of 4096
6.2 seconds

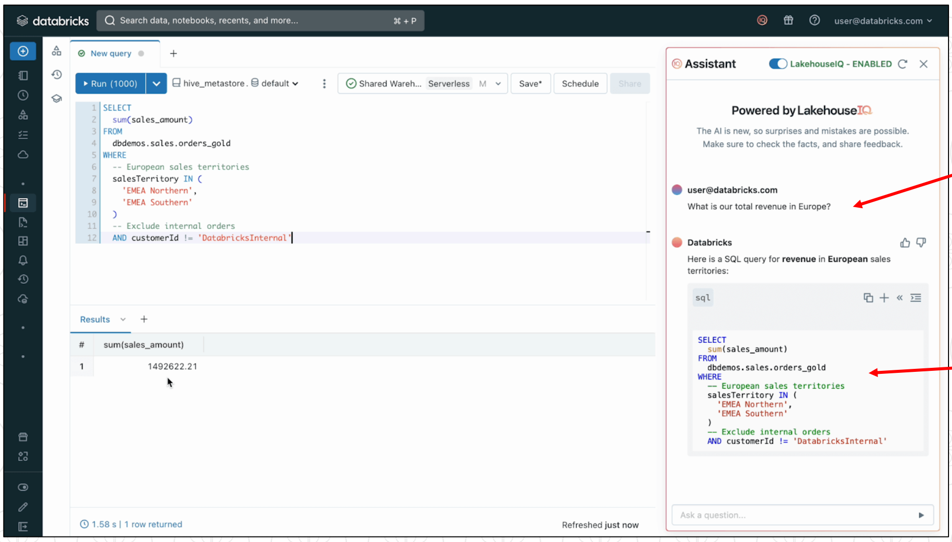
Model parameters

Decoding: Greedy / Sampling (Sampling selected)

Temperature: 0 – 2, 0.8
Top P (nucleus sampling): 0 – 1, 1
Top K: 1 – 100, 50
Random seed: 111
Repetition penalty: 1 – 2, 2

Source: IBM

Generative AI In Data Science
– Databricks LakehouseIQ Prompt Based Artificially Intelligent Query Generation



```
SELECT
  sum(sales_amount)
FROM
  dbdemos.sales.orders_gold
WHERE
  -- European sales territories
  salesTerritory IN (
    'EMEA Northern',
    'EMEA Southern'
  )
  -- Exclude internal orders
  AND customerId != 'DatabricksInternal'
```

#	sum(sales_amount)
1	1492622.21

1.58 s | 1 row returned

Refreshed just now

Powered by LakehouseIQ

The AI is new, so surprises and mistakes are possible. Make sure to check the facts, and share feedback.

user@databricks.com
What is our total revenue in Europe?

Databricks
Here is a SQL query for **revenue** in **European** sales territories:

```sql
SELECT
  sum(sales_amount)
FROM
  dbdemos.sales.orders_gold
WHERE
  -- European sales territories
  salesTerritory IN (
    'EMEA Northern',
    'EMEA Southern'
  )
  -- Exclude internal orders
  AND customerId != 'DatabricksInternal'
```

Lakehouse

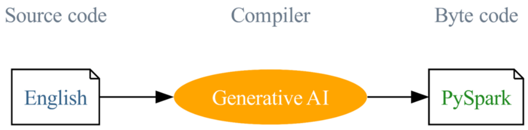
Prompt

LakehouseIQ knows the company has two European sales regions and so automatically adds them to the query.

It has also learned from other queries, dashboards and notebooks that used this dataset and so automatically adds a filter to exclude internal usage

Source: Databricks

Generative AI In Data Science – Databricks English SDK



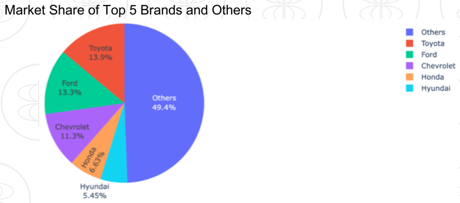
Generative AI's expert knowledge about Spark is built into the English SDK

Data ingestion

```
spark_ai = SparkAI() auto_df = spark_ai.create_df("2022 UK national auto sales by brand")
```

Dataframes

```
spark_ai.activate()  # Activate the df.ai methods
auto_df.ai.plot("pie chart for UK sales market shares, show the top 5 brands and the sum of others")
```


Market Share of Top 5 Brands and Others

```
auto_top_growth_df=auto_df.ai.transform("top brand with the highest growth")
auto_top_growth_df.show()
```

brand	uk_sales_2022	sales_change_vs_2021
Toyota	134726	327

Topics – Where Are We?

- What is generative AI?
- What are the business benefits of generative AI?
- How is generative AI being used in data management?
- ➢What does this mean for business going forward?
- What should you do to get started?

What Does Generative AI In Data And Analytics Mean For Business?

Data Producers	Data Consumers
Conversational data search in data catalogs	Conversational data search in data marketplaces
Automated metadata enrichment at scale	Automatic exploration of data, data pipelines and lineage
Automated physical schema generation	Natural language queries
Prompt-based data engineering	AI-generated natural language answers
Synthetic data generation	Auto generated virtual views on data
AI-assisted generation of data quality validation rules, master data matching rules and data governance policies	
AI-assisted generation of digital product information	
AI-assisted generation of product information in MDM (Digital web pages, brochures, marketing content)	

- Generative AI in data and analytics:
  - Lowers the skills bar and broadens inclusion
  - Accelerates development
  - Explains business insights and the business impact
  - Shortens the time to value
  - Shortens the time to act

Everyone will get a self-learning AI-Assistant!

Copyright © Intelligent Business Strategies 1992-2024

Topics – Where Are We?

- What is generative AI?
- What are the business benefits of generative AI?
- How is generative AI being used in data management?
- What does this mean for business going forward?
- ➤ What should you do to get started?

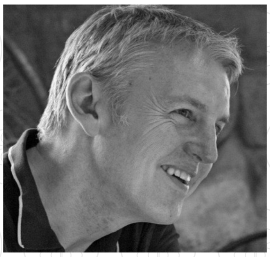
Copyright © Intelligent Business Strategies 1992-2024

Copyright © Intelligent Business Strategies 1992-2024

Conclusion - Getting Started

- Generative AI is rapidly finding its way into many different data management, data governance, BI and data science tools with more to come
- Fine tune foundation LLMs to understand your data and your metadata
- Exploit generative AI in data management to:
  - Improve data catalogs through conversational data search and automated data curation
  - Accelerate and start automating data governance tasks
  - Increase the number of citizen data engineers
  - Accelerate data engineering to produce data products more rapidly
  - Accelerate the build of master data
  - Auto generate product information and marketing content from product master data

About Mike Ferguson



www www.intelligentbusiness.biz

✉ mferguson@intelligentbusiness.biz

X @mikeferguson1

☎ (+44) 1625 520700

Mike Ferguson is Managing Director of Intelligent Business Strategies Limited. As an independent IT industry analyst and consultant, he specialises in BI / analytics and data management. With over 40 years of IT experience, Mike has consulted for dozens of companies on BI/Analytics, data strategy, technology selection, enterprise architecture, and data management. Mike is also conference chairman of Big Data LDN, the largest data and analytics conference in Europe and a member of the EDM Council CDMC Executive Advisory Board. He has spoken at events all over the world and written numerous articles. Formerly he was a principal and co-founder of Codd and Date – the inventors of the Relational Model (which caused the birth of relational databases and the SQL language), and Chief Architect at Teradata on the Teradata DBMS He teaches popular master classes in Data Strategy, Data Warehouse Modernisation, Practical Guidelines for Implementing a Data Mesh, Big Data, How to Govern Data across a Distributed Data Landscape, Machine Learning and Advanced Analytics, and Embedded Analytics, Intelligent Apps and AI Automation

Thank You!