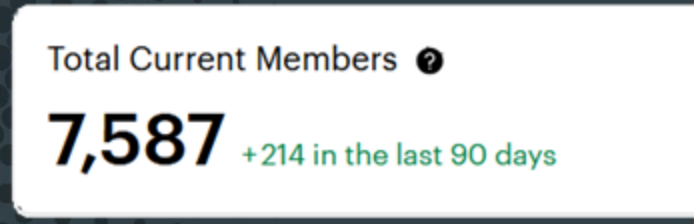
Data
Community

Data Community Poland



- Story of DCPL Association started around year 2008
- Data enthusiasts, but rather on technical side
- Local meetings – look for us on meetup.com
- Conferences (SQLDay, SQLDay Lite, etc.) – send your peers there
- datacommunity.pl & major social media

Smart Data, Smarter Decisions: LLM as the new standard in data management

Hubert Kobierzewski

Bartłomiej Graczyk

Smart Data, Smarter Decisions: LLM as the new standard in data management

???

Hubert Kobierzewski

Bartłomiej Graczyk

The Data Management crisis - real numbers, real impact - The scale of data challenges

Data Community

Time Drain:

80% of data analysts' time spent finding and understanding data

Financial Impact:

Organizations in avg. **lose 12% of revenue** due to poor data quality

Trust Crisis:

50% of data not trusted for decision making**, 70% deemed inaccurate**

Governance Gap:

37% of organizations lack formal data quality management programs

Analytics Failures:

40% of analytics projects fail due to data quality issues

Business Process Impact:

25-30% of business processes affected by poor data quality

Core Data Management Challenges - What keeps us awake

GOVERNANCE & OWNERSHIP GAPS

COMPLIANCE & SECURITY RISKS

QUALITY & TRUST ISSUES

DATA SILOS & ACCESS BARRIERS

Core Data Management Challenges - What keeps us awake

GOVERNANCE & OWNERSHIP GAPS

- Unclear data ownership and accountability
- Inconsistent definitions across teams and departments
- Manual cataloging creates documentation debt

COMPLIANCE & SECURITY RISKS

- 52% struggle with compliance audits
- PII and sensitive data poorly classified
- Manual processes increase security vulnerabilities

QUALITY & TRUST ISSUES

- Data quality score averages only 60/100 across industries
- 90% of quality issues stem from manual data entry errors
- Reactive discovery - problems found after business impact

DATA SILOS & ACCESS BARRIERS

- Departments store data independently (inconsistencies, redundancies)
- SQL remains gatekeeper for 90% of business users
- Critical data locked in legacy systems without modern connectors

The LLM Revolution - Three pillars of transformation

AUTOMATED CATALOGING & DISCOVERY

INTELLIGENT DATA QUALITY

SEMANTIC DATA DISCOVERY

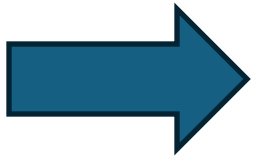
Automated Cataloging - From manual to magical - Real tools solving real problems

Data Community

AUTOMATED CATALOGING & DISCOVERY

Before: The Traditional Approach

- Data Steward manually documents each table (days/weeks per dataset)
- Business glossary creation requires extensive stakeholder interviews
- Data lineage mapping done through spreadsheets and meetings
- Compliance classification performed through manual audits



Now: LLM-Powered Documentation

- **Microsoft Purview:** AI-powered metadata generation and sensitivity labeling
- **Alation ALLIE AI:** Intelligent curation with behavioral learning
- **Collibra AI Governance:** Automated policy suggestions and risk assessment
- **Dataedo+ LLM Integration:** Auto-generated descriptions for tables, columns, and business glossary creation with contextual definitions using GPT-4, GPT-3.5 Turbo, or Azure OpenAI

Intelligent Data Quality - Beyond traditional rules - From reactive rules to proactive intelligence

Data Community

🎯 INTELLIGENT DATA QUALITY

Traditional Approach Limitations:

-- Rigid rules that break when data evolves

```
CHECK (age BETWEEN 0 AND 120)

CHECK (email LIKE '%@%.%')

CHECK (order_date <= CURRENT_DATE)
```

Pre-GenAI tools:

- Anomaly Detection: flagging abnormal values, outliers, missing values (Isolation Forest, SVM, Neural Nets, Random Forest)
- Deduplication/Matching: Finding similar records, fixing typos (K-means, Hierarchical Clustering, Fuzzy Matching)
- Predictive Assessment: Flagging high-risk sources/steps (Decision Trees, Logistic Regression)

Intelligent Data Quality - Beyond traditional rules - From reactive rules to proactive intelligence

Data Community

🎯 INTELLIGENT DATA QUALITY

LLM-Powered Quality Revolution:

Pattern Learning Without Rules:

- Behavioral Analysis: Learns normal data patterns automatically
- Contextual Anomalies: Identifies business-relevant outliers, not just statistical ones
- Adaptive Thresholds: Adjusts detection sensitivity based on data evolution
- Seasonal Awareness: Understands cyclical patterns in business data

Production-Ready Solutions:

- Snowflake Cortex ML: Built-in anomaly detection with simple SQL interface
- Microsoft Fabric AI Skills / AI Agents: Natural language data validation and quality assessment
- Anomalo Platform: Enterprise-grade automated quality monitoring
- Databricks Lakehouse IQ: AI-powered data observability and quality insights

Intelligent Data Quality - Beyond Traditional Rules.

Data Community

🎯 INTELLIGENT DATA QUALITY

Traditional Approach Limitations:

- Dummy data generation
- Mock data generation
- Anonymized (De-identified) / Masked Data

```
CREATE TABLE Employees (
    EmployeeID INT PRIMARY KEY,
    FullName NVARCHAR(100) MASKED WITH (FUNCTION = 'default()'),
    Email NVARCHAR(100) MASKED WITH (FUNCTION = 'email()'),
    Salary DECIMAL(10, 2) MASKED WITH (FUNCTION = 'random(1000, 5000)')
);
```

- Subsetted Data:
    - TOP (N) CLAUSE
    - WHERE CLAUSE

Synthetic Data Generation:

- Privacy-Safe Testing: Generate realistic data without exposing sensitive information
- Statistical Preservation: Maintain correlations and distributions of original data
- GDPR Compliance: Enable development and testing without privacy risks

Semantic Discovery - breaking the SQL barrier. Natural language access to enterprise data

Data Community

SEMANTIC DATA DISCOVERY

The Democratization Challenge:

- Business analysts spend 60% of time writing/debugging SQL
- Domain experts can't directly access their own data
- IT becomes bottleneck for simple data requests
- Self-service analytics remains elusive for most organizations

AI-Powered Solutions:

Microsoft Fabric + Copilot:

- Find, summarize, and analyze data across multiple reports, semantic models, apps, and data agents
- Business users in Power BI can now chat with data agents without switching context

Databricks AI/BI Genie:

- Natural Language Interface: Chat-based interaction with lakehouse data
- Context Awareness: Understands business terminology and data relationships
- Smart Recommendations: Suggests relevant analyses based on user intent
- Visual Results: Automatic chart generation for discovered insights

The LLM Revolution - Three pillars of transformation

AUTOMATED CATALOGING & DISCOVERY

- Challenge Solved: Manual documentation, unclear ownership
- AI-generated asset descriptions and classifications
- Real-time metadata enrichment and relationship mapping
- Automated PII detection and sensitivity labeling

INTELLIGENT DATA QUALITY

- Challenge Solved: Reactive quality management, manual validation
- Pattern-based anomaly detection without predefined rules
- Synthetic data generation preserving statistical properties
- Proactive alerting with business context and root cause analysis

SEMANTIC DATA DISCOVERY

- Challenge Solved: SQL barriers, data silos, access complexity
- Natural language queries eliminating technical barriers
- Contextual search understanding business intent
- Democratized data access for all organizational roles

Implementation Roadmap - Your path to LLM-powered data management

3-Phase Approach with specific tools

PHASE 1 (1-2 months): Automated Cataloging Foundation
Goal: Prove AI value with documentation automation

Tools & Actions:

- Dataedo Trial: Implement AI auto documentation for 3-5 critical datasets
- Microsoft Purview: Enable AI metadata generation for cloud data assets
- Success Metrics: Time reduction in cataloging, metadata quality scores
- Deliverables:
  ✓ AI-generated documentation for pilot datasets
  ✓ Baseline measurement of manual vs. automated cataloging time
  ✓ Stakeholder feedback on documentation quality and usefulness

Implementation Roadmap - Your path to LLM-powered data management

3-Phase Approach with specific tools

PHASE 2 (3-4 months): Intelligent Quality & Discovery
Goal: Extend AI capabilities to quality monitoring and user access

Tools & Actions:

- Snowflake Cortex/Fabric AI Skills: Deploy anomaly detection on key metrics
- Databricks Genie/Fabric Copilot: Enable natural language querying for power users
- Integration: Connect quality alerts with existing monitoring workflows
- Deliverables:
  ✓ Proactive quality monitoring with business-context alerts
  ✓ Natural language interface for 10-15 priority use cases
  ✓ User training and adoption measurement framework

Implementation Roadmap - Your path to LLM-powered data management

3-Phase Approach with specific tools

PHASE 3 (6+ months): Enterprise-Wide AI Governance
Goal: Scale AI capabilities across entire data organization

Tools & Actions:

- Full Platform Integration: Connect all AI tools with existing data infrastructure
- Governance Automation: Implement AI-powered policy enforcement and compliance
- Culture Change: Embed AI-assisted data management in all workflows
- Deliverables:
  ✓ Organization-wide AI-powered data catalog and quality monitoring
  ✓ Self-service analytics capabilities for business users
  ✓ Automated governance workflows with measurable compliance improvements

DEMO

Demo Preview - What You'll See Next

Live in Action – Following every data professional data journey

1. AI-Powered Discovery / Description with Dataedo
2. Intelligent Quality Monitoring
3. Synthetic Data Generation
4. Natural Language Querying with Fabric/Databricks

Demo Preview - What You'll See Next

Live in Action – Following every data professional data journey

1. AI-Powered Discovery / Description with Dataedo
2. Intelligent Quality Monitoring
3. Synthetic Data Generation
4. Natural Language Querying with Fabric/Databricks

Demo Preview - What You'll See Next

Live in Action – Following every data professional data journey

1. AI-Powered Discovery / Description with Dataedo
2. **Intelligent Quality Monitoring**
3. Synthetic Data Generation
4. Natural Language Querying with Fabric/Databricks

Data Quality

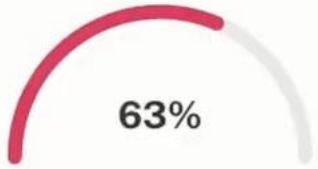
+ Add Moni

Overview | Monitors | Assets

Data Quality Score

Overall health of your assets

How is this calculated?



Week | Month

Data Quality Score Card

Data Quality Health across each metric

How is this calculated?

Data Quality Metric	Score	Trend
Accuracy	72%	
Completeness	62%	
Consistency	0%	
Timeliness	0%	
Validity	80%	
Uniqueness	73%	
Custom	0%	

Monitoring Overview

0.68% of total assets are being monitored

Tables Monitored 1.56%

9/576

Columns Monitored 0.64%

31/4852

BI Reports Monitored 0%

0/425

Failing Monitors

View All

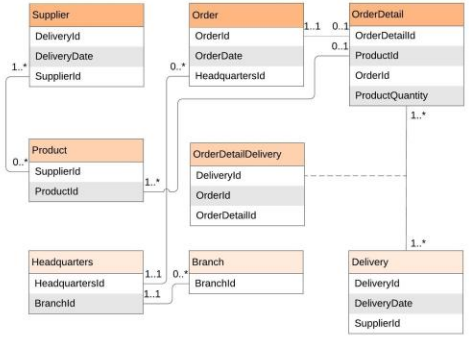
Monitor Name	Assets Monitored	Checks	Status	Failure Rate	Last Run	Next Run
CTE Filters test	1	2	Fail	50.00%	Jun 30, 10:48 PM	Jul 2, 02:48 PM
Mchan test	1	14	Fail	35.71%		Jul 7, 09:08 AM
Test Monitor	1	20	Fail	35.00%	Jun 30, 03:29 AM	Jul 6, 07:29 PM
Test V	5	67	Fail	26.87%	Jul 1, 05:45 PM	Jul 2, 09:45 AM
250624 - MJB Credit	1	8	Fail	25.00%	Jul 1, 12:09 AM	Jul 2, 04:09 PM

Demo Preview - What You'll See Next

Live in Action – Following every data professional data journey

1. AI-Powered Discovery / Description with Dataedo
2. Intelligent Quality Monitoring
3. **Synthetic Data Generation**
4. Natural Language Querying with Fabric/Databricks

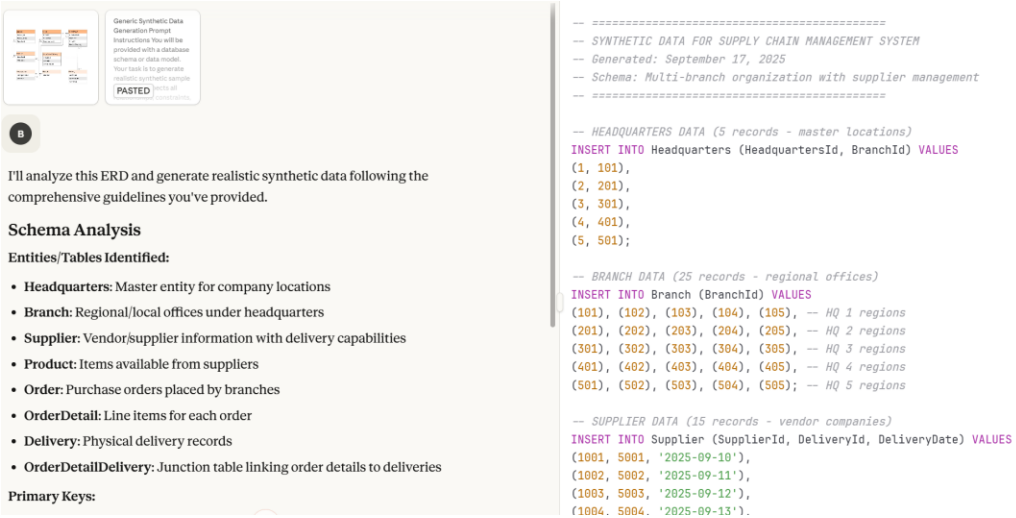
Implementation Roadmap - Your path to LLM-powered data management



Generic Synthetic Data Generation Prompt
Instructions

You will be provided with a database schema or data model. Your task is to generate realistic synthetic sample data that respects all relationships, constraints, and business logic implied by the schema.

Input Requirements
Please provide the database schema in one of the following formats:
Entity Relationship Diagram (ERD)
Database schema description
Table definitions with relationships
SQL DDL statements
Visual diagram or model



Demo Preview - What You'll See Next

Live in Action – Following every data professional data journey

1. AI-Powered Discovery / Description with Dataedo
2. Intelligent Quality Monitoring
3. Synthetic Data Generation
4. Natural Language Querying with Fabric/Databricks

flights

Overview | Sample Data | Details | Permissions | History | Lineage | Insights | Quality

Filter columns...

View relationships | AI generate

Column	Type	Comment	Tags	Column masking rule
Year	int			
Month	int			
DayofMonth	int			
DayOfWeek	int			
DepTime	string			
CRSDepTime	int			
ArrTime	string			
CRSArrTime	int			
UniqueCarrier FK	string			
FlightNum	int			
TailNum	string			
ActualElapsedTime	string			
CRSElapsedTime	int			
AirTime	string			
ArrDelay	int			

About this table

Owner

Data source Del

Popularity

Last updated 6 minu

Size 28.3M

Tags

Add tags

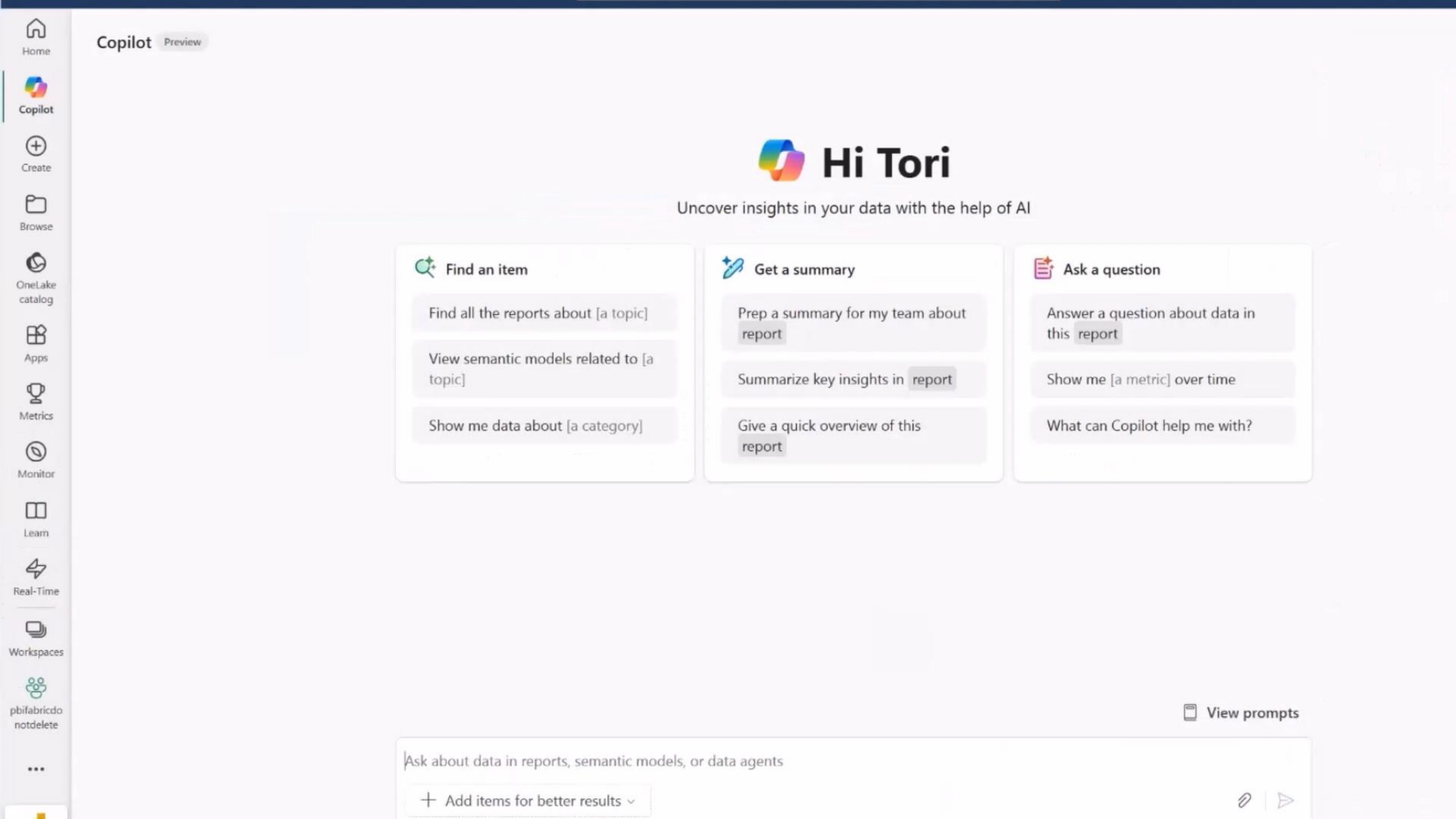
Row filter

Add filter

AI Suggested Descript

The 'flights' table contains h
airline industry. It includes d
times, carriers, flight numbe
data can be used for various
scheduling, performance an
efficiency improvement. It c
flight delays and identifying
specific flights or carriers.

Accept | Edit


Copilot
Preview
Home
Copilot
Create
Browse
OneLake catalog
Apps
Metrics
Monitor
Learn
Real-Time
Workspaces
pbifabricdo notdelete
Hi Tori
Uncover insights in your data with the help of AI
Find an item
Find all the reports about [a topic]
View semantic models related to [a topic]
Show me data about [a category]
Get a summary
Prep a summary for my team about report
Summarize key insights in report
Give a quick overview of this report
Ask a question
Answer a question about data in this report
Show me [a metric] over time
What can Copilot help me with?
View prompts
Ask about data in reports, semantic models, or data agents
Add items for better results

Prep data for AI

- Get started
- Simplify the data schema
- Verified answers
- Add AI instructions

Prep this model to be AI-ready

Improve Copilot insights about this semantic model in a few simple steps. Publishing to the Power BI service will apply your changes to all Copilot chats, but updates may take a moment. Learn more

Simplify the data schema

Improve response accuracy by deselecting fields Copilot doesn't need to analyze.

Verified answers

Save predefined responses for your most critical business topics.

Add AI instructions

Help Copilot understand industry terms, business priorities, and important data fields.

Prep data for AI

- Get started
- Simplify the data schema
- Verified answers
- Add AI instructions

Simplify the data schema (preview)

Improve response accuracy by deselecting fields Copilot doesn't need to analyze. This creates a dedicated AI data schema saved with your semantic model. Learn more

- [ ] Association
- [x] Calendar
- [ ] Customer
  - [x] Age
  - [x] Amount
  - [x] Category
  - [x] Gender
  - [ ] ID
  - [x] Issue
  - [x] Price Range
  - [x] Product
  - [x] Promotion
  - [x] Segment
  - [x] Store
  - [x] Type
  - [x] Unit
- [x] Details
- [x] Issues and Promotions
- [x] Product
- [ ] Sales
- [ ] STable
- [x] Store
- [ ] Tooltip Info
- [ ] Tooltip Info2

Apply Close

Prep data for AI

- Get started
- Simplify the data schema
- Verified answers
- Add AI instructions

Verified answers (preview)

← All verified answers

Phrases connected to verified answers

Add phrases people might use when asking about the model data, and Copilot will respond with this visual. Test your phrases out in the chat before publishing.

Enter a phrase | Add

Copilot suggestions

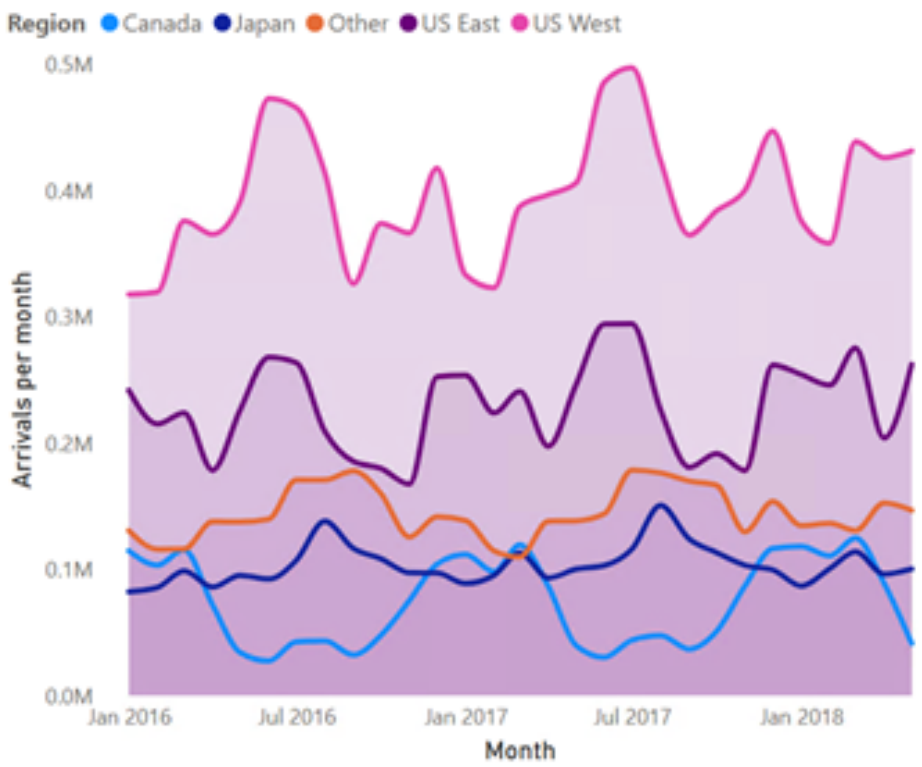
- What are the monthly arrival trends?
- How do arrivals vary by month?
- Which month has the highest arrivals?

Refresh

Copilot uses AI. Always review content for mistakes. Read terms

Visual

The visual below will appear in Copilot's responses to phrases you add. Any changes to the report won't impact this visual.



Apply | Close

Prep data for AI

- Get started
- Simplify the data schema
- Verified answers
- Add AI instructions

Verified answers (preview)

Filter by keyword

Phrases connected to verified answers

- why do folks visit hawaii? 2 phrases
  - breakdown of trip reason
- what are the monthly arrival trends? 3 phrases
  - monthly arrivals
  - visitor trends over time
- Expenditures per month for all visitor countries 3 phrases
  - Expenditures across visitor countries
  - Expenditures across regions
- Give me a breakdown of user spending over time 3 phrases
- Summary of spending 2 phrases

Apply Close

Prep data for AI

- Get started
- Simplify the data schema
- Verified answers
- Add AI instructions

Add AI instructions (preview)

Provide context on important data fields, business goals, and industry terms to improve Copilot insights about this semantic model. View some examples

Busy season is April – November.
ASOT should always reference the Top Sales field.
Our company specializes in financial analytics—frame insights around risk assessment and market trends.

189/5000 characters

Apply Close

Data
Community