Microsoft

Microsoft AI Tour

Microsoft

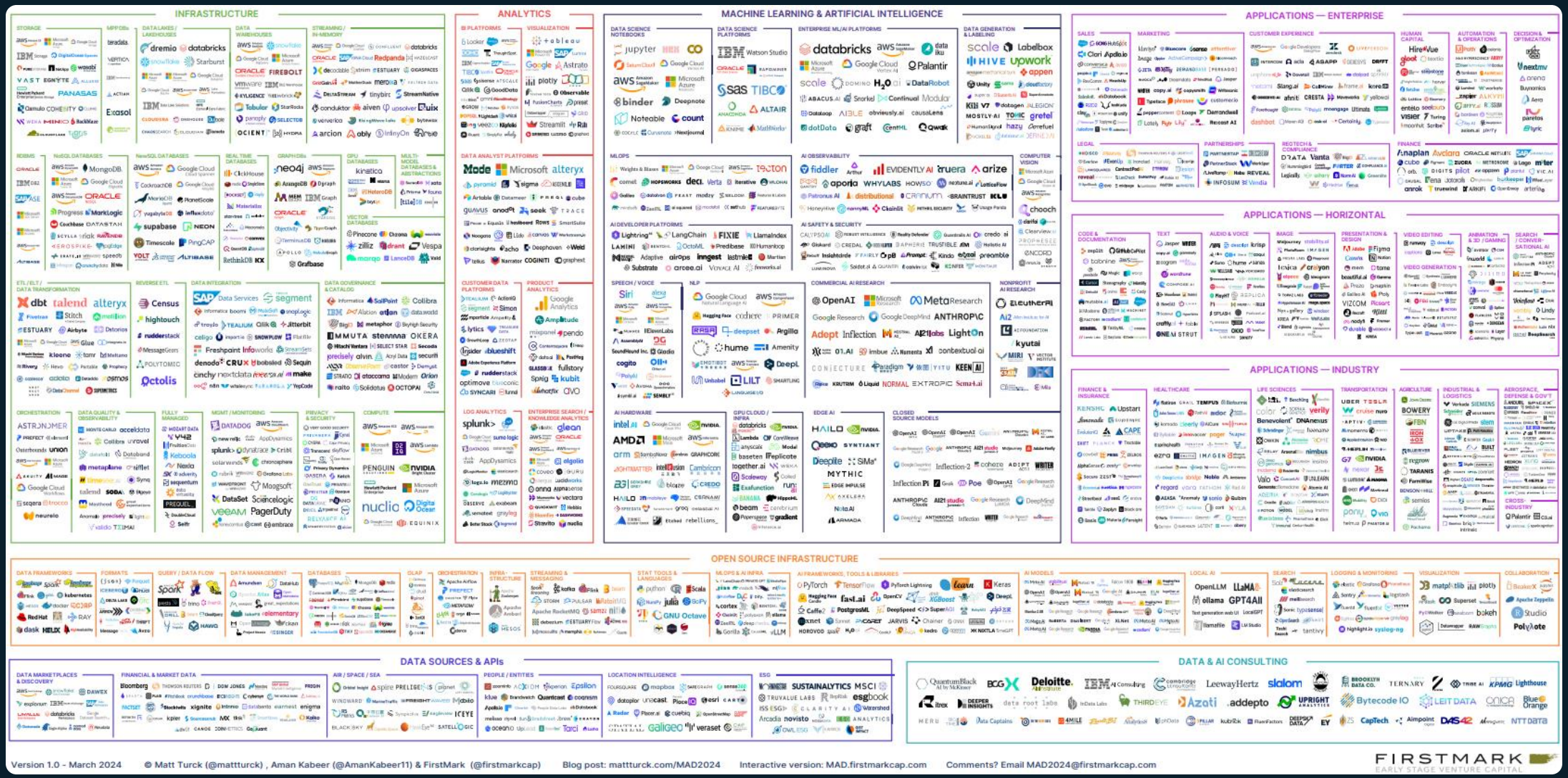
Next gen AI apps with databases at scale anywhere

Microsoft AI Tour

AI is changing the world

Data is the fuel that powers AI

2024 Data and AI Landscape

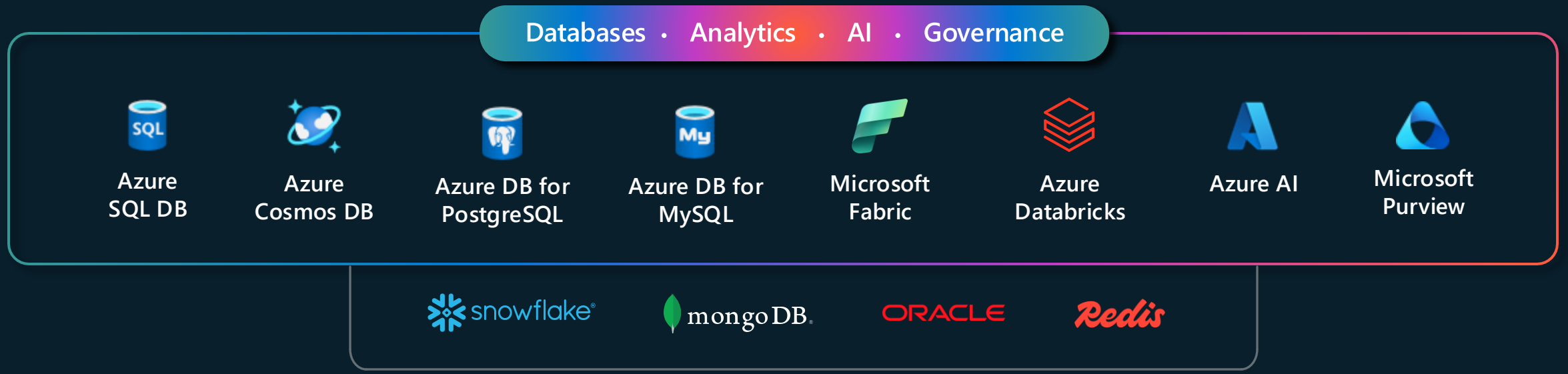


“Unify,

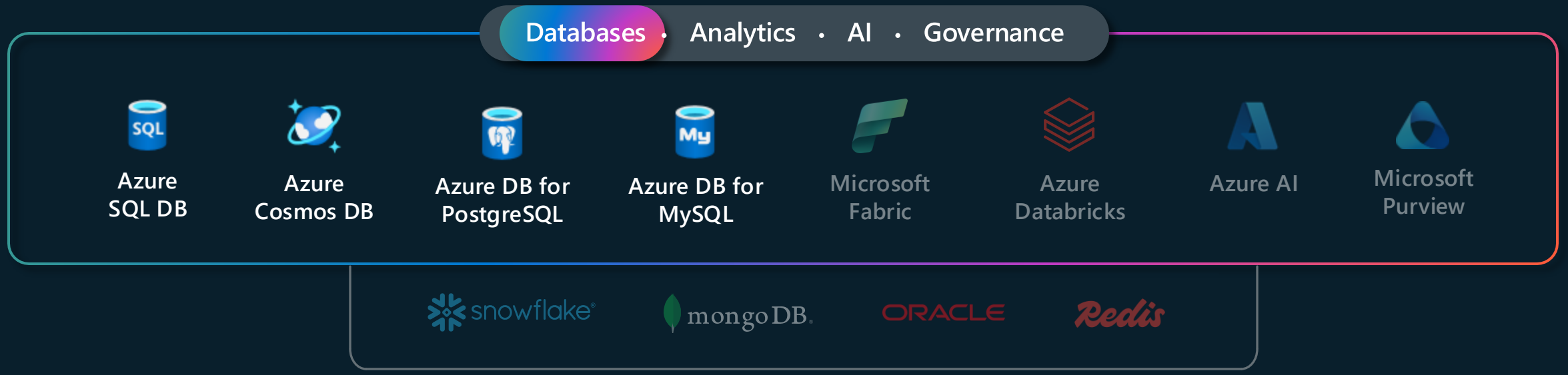
I am the Chief Information Officer and don’t want to be the Chief Integration Officer. Help me unleash AI in my data estate.”

Every CIO, Every Enterprise

Microsoft Intelligent Data Platform



Microsoft Intelligent Data Platform



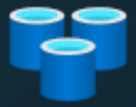
Cloud databases for the era of AI

Azure SQL Database

Azure Database for PostgreSQL

Azure Database for MySQL

Azure Cosmos DB



Azure Cache for Redis

Oracle Database @Azure

MongoDB Atlas

Data is highly variable and both structured and unstructured

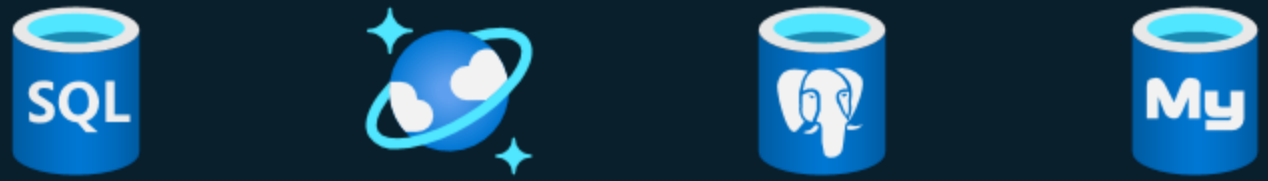
Unpredictable traffic

Globally-distributed

Applications and AI models depend on real-time data

Azure Operational Databases

Comprehensive Portfolio for the era of AI



97%

Fortune 500 Companies are Azure Database Customers

2X

Arc Customer Growth YoY on Azure SQL Server

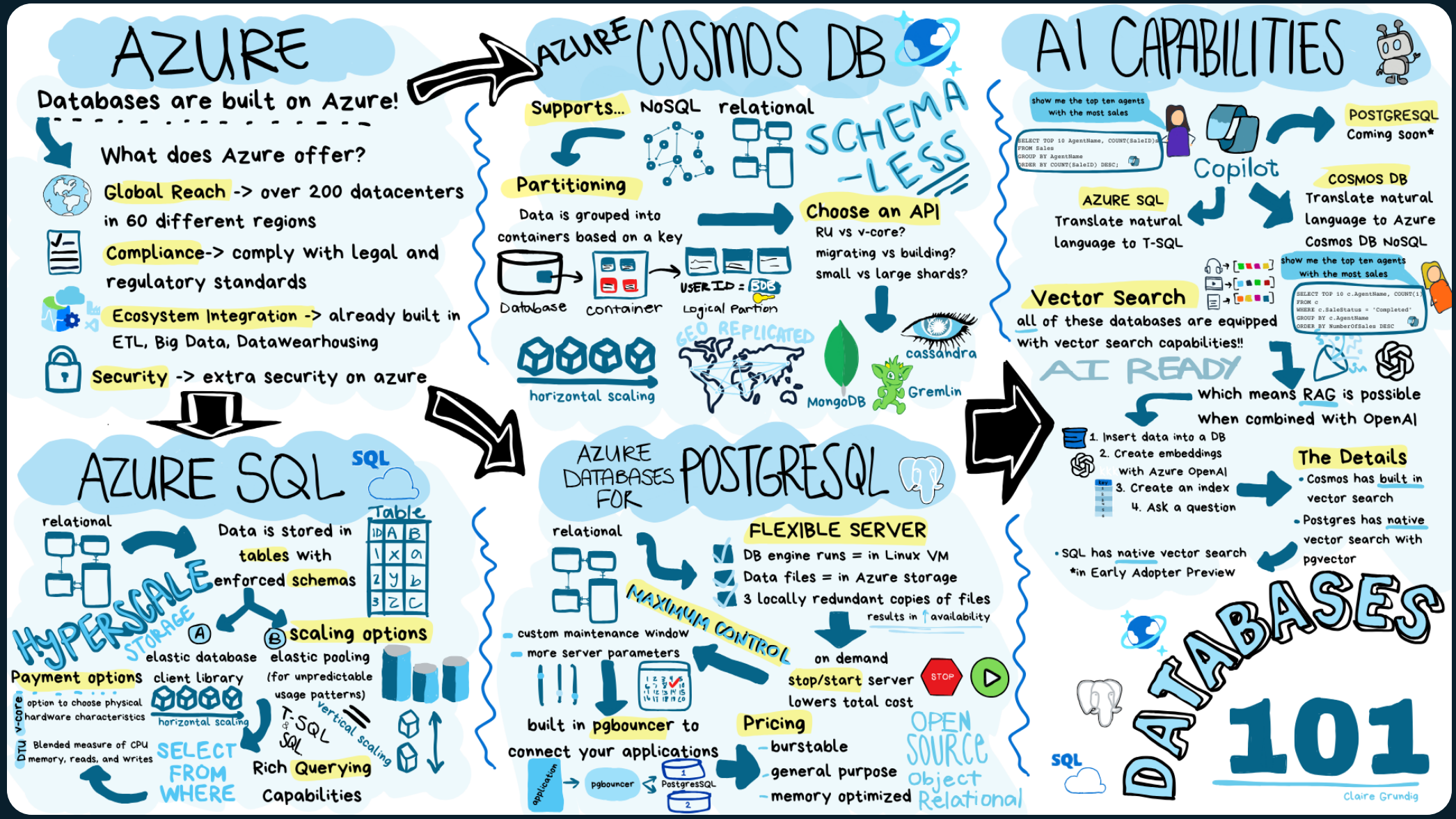
Up to 68%

Azure SQL Hyperscale offers more perf than AWS Aurora per dollar

Up to 2.6X

MySQL transactions/second

Overview of databases and AI



Author: Claire Grundig

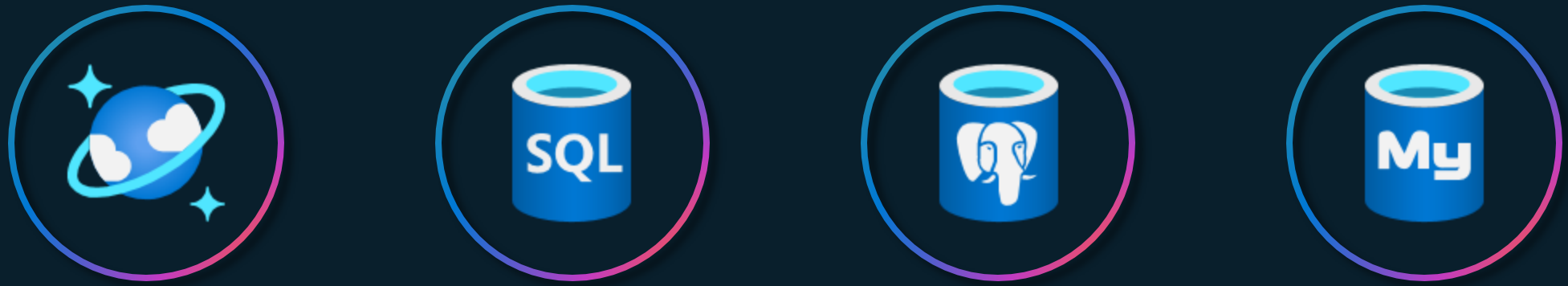
Powering a new era of apps
that run anywhere, at any scale

Intelligent

Integrated

Trusted

A new era of databases: autonomous, and AI-enhanced



Intelligent

AI-powered management and performance at any scale

Integrated

Modern, AI apps on an open, flexible, and common platform

Trusted

A sustainable and reliable platform so your data is always available, safe and compliant

Intelligent

Intelligent

Integrated

Trusted

Intelligent

AI-powered management and performance at any scale

Intelligent

AI-powered management and performance at any scale

Optimize processes, resources, and workloads

Support for any type and size of data

Maintain peak performance and stable workloads

Microsoft Copilot
in Azure for Azure databases

Public preview

Copilot capabilities

Natural language to T-SQL conversion in Azure SQL Database

Natural language to Azure Cosmos DB NoSQL

(Coming Soon) Azure Copilot for PostgreSQL Flexible Server

Natural language to T-SQL

Light-weight Query Editor Experience

Natural Language to Generate Complex SQL Queries

Grounded with your Database Schema

Natural language to Cosmos DB NoSQL

Turn your natural language data questions in NoSQL queries

Integrated into Data Explorer at no charge

(Public Preview)

Copilot in Azure capability for multi-turn conversations

(Coming Soon)

Vector search

Revolutionize Indexing and Retrieval Augmented Generation for LLM Apps

Images

Audio

Video

Graphs

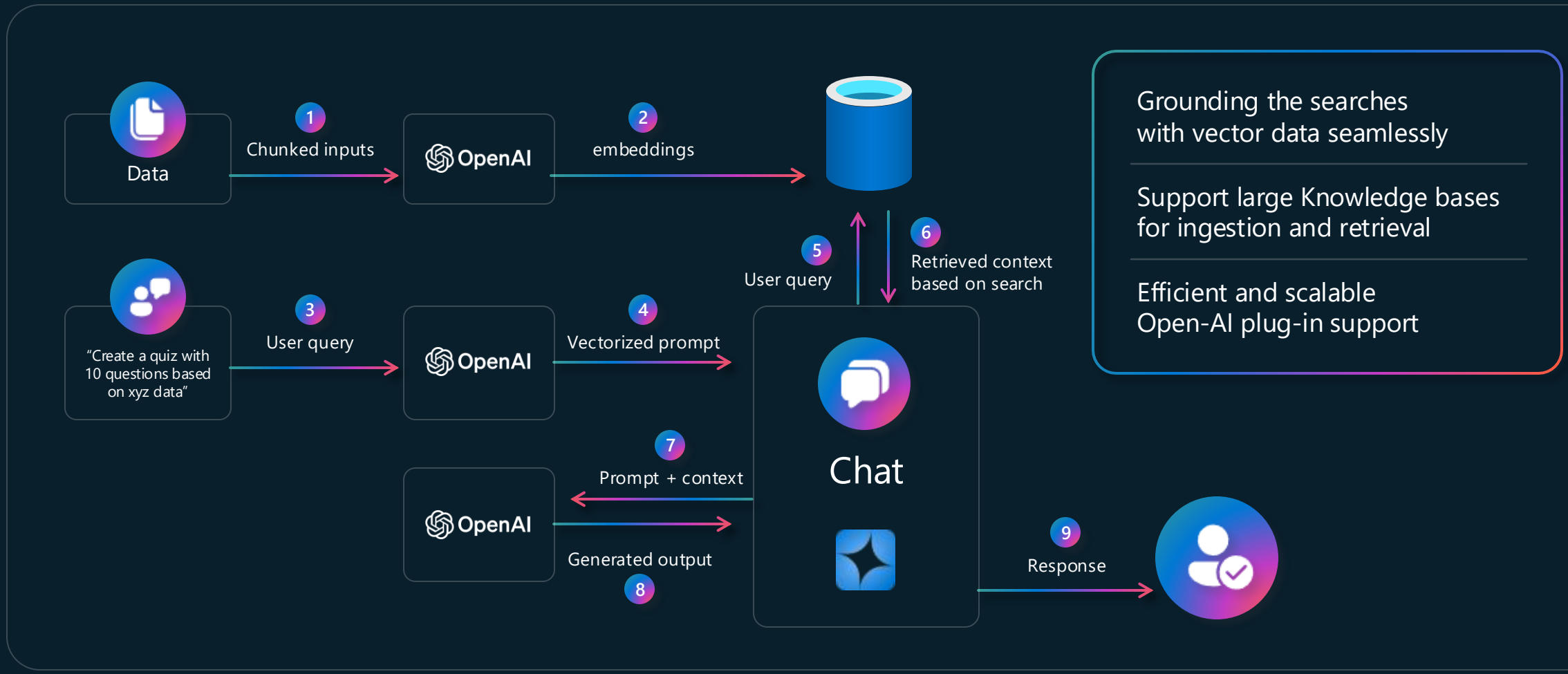
Text

- Leverage data from any data store
- Improve relevancy
- Query across multiple types of data
- Quickly search through large data sets

- Deploy with enterprise-grade security
- Easily scale with changing workloads
- Build retrieval plugins for OpenAI's ChatGPT using Azure OpenAI service

Retrieval Augmented Generation (RAG)

Empower LLMs with Operational Data context



Azure Cosmos DB

A set of highly-scalable & AI-ready databases

Azure Cosmos DB for NoSQL

Serverless or Provisioned Throughput

High elasticity with instant autoscale

Low latency, real-time data transactions

Mission critical reliability (99.999%)

Built-in vector index and search with DiskANN

Integrations with Semantic Kernel and Langchain

Azure Cosmos DB for MongoDB

MongoDB compatible

Provision compute + storage

Store data + vectors together, keep consistent

High reliability (99.995%)

Built-in vector index and search (ft. IVF & HNSW)

Azure Cosmos DB + Generative AI Scenarios

What	Why	When
Semantic Caching	Drastically reduces latency Saves on Token consumption Reduces costs and latency for LLM	Slow moving / static content FAQs, Policies…
Chat History	Conversational context UX improvements LLM optimizations Auditing	A MUST for Chat sessions Improving cost & performance
Retrieval Augmented Generation (RAG)	Personalize LLM on your data Cheaper than fine tuning Faster iteration on new data	Any workload for GenAI apps
Vector + Operational Database	No ETL Consistent data Reduce complexity & costs	Data & vectors together Cosmos DB scale & performance

Vector Search in Azure Cosmos DB for NoSQL

Store data + vectors together

Reduced Complexity & Cost

Transactional Data & Vectors

Optimized for App Developers

Vector Search + Query Filters

Combine with equality, range & spatial filters

Optimize query

Flexible Indexing

Flat, quantized flat, and DiskANN indexing available

Azure Cosmos DB for NoSQL Capabilities

Serverless or provisioned throughput

Built-in multitenancy

Instant & dynamic autoscale

<10ms point-reads

Globally-replicated

Industry-leading 99.999% SLA

The DiskANN Advantage

Creates a graph-based index that stores compressed vectors in memory, and a full-fidelity graph on SSDs. This provides many advantages:

Low latency

- Long range edges help search convergence faster
- Minimize number of hops in graph to reduce disk latency

High accuracy

- Link nodes to directionally diverse neighbors to improve recall
- Search using compressed vectors, re-rank using full vectors

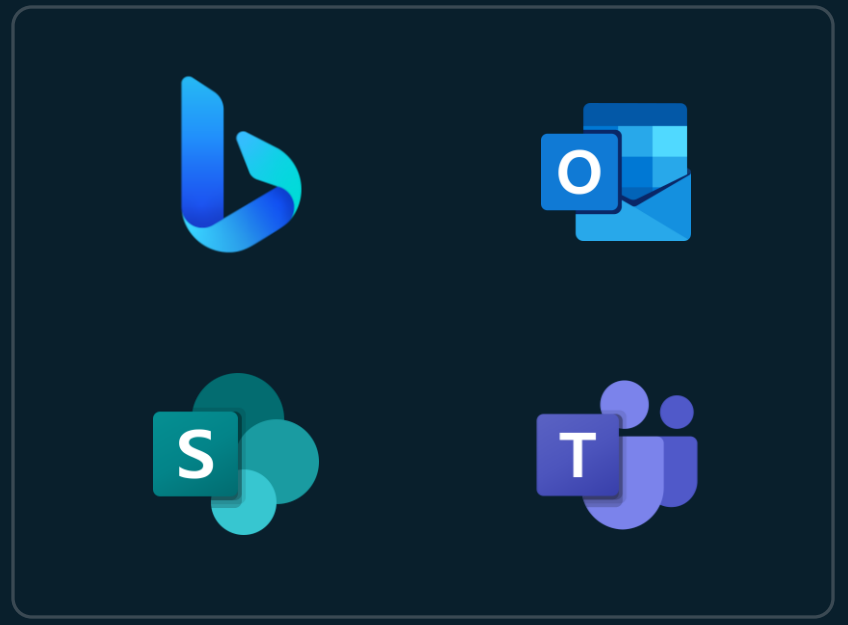
Low cost

- Resource consumption is reduced by storing full-fidelity graph on SSDs rather than in memory

Robust to data changes

- Unlike HNSW, accuracy and performance will not degrade over time with many insertions, deletions, and updates

DiskANN powers semantic search across Microsoft



- Indexes with up to 400B vectors, trillions of points
- Mean query latency <5ms, >90% recall

Demo

Azure Cosmos DB

Resources

Vector Search Announcements
aka.ms/aitour-data-cosmosdb-vectorsearchannouncements

Get Started with Samples
aka.ms/CosmosAISamples

DiskANN Preview Signup
aka.ms/aitour-data-cosmosdb-diskannpreview

Vector Search Documentation
aka.ms/aitour-data-cosmosdb-vectorsearch

Featured Demo
aka.ms/aitour-data-cosmosdb-demo

Try Azure Cosmos DB for Free

No Credit Card Required

6 APIs to choose from

Distributed data across 4 regions

Up to 10 GB of storage

aka.ms/aitour-data-cosmosdb-trycosmosdb

General availability

Unlocking the power of Open AI and pgvector with Azure Database for PostgreSQL

Simplified LLMs with direct use of PostgreSQL data

Azure Database for PostgreSQL:

Native Vector Search

- pgvector extension
- Storing and indexing vectors alongside relational data
- Various indexing & retrieval strategies
- Combine vector queries with metadata filters
- Access control and other relational DB features work with vectors

Generative AI apps

RAG (Retrieval Augmented Generation) apps

Retrieve private data to ground LLM model responses

Recommendation/Semantic Search

Retrieve similar documents by distance between vectors

Hybrid Search

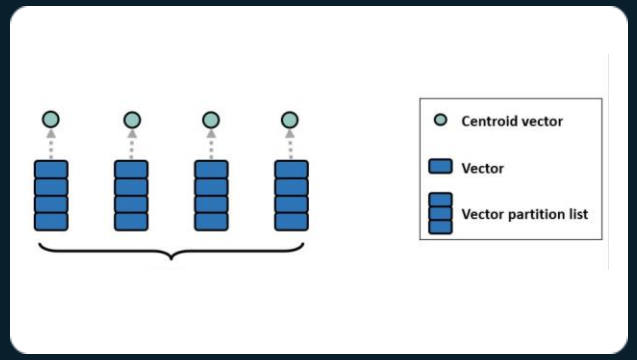
Combine vector search, row filtering, and full-text search

Vector indexes supported today

IVFFlat

Clusters vectors by applying k-means clustering.

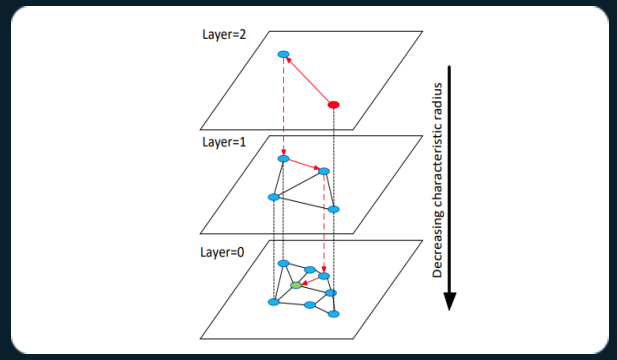
Memory efficient but requires index rebuilds.



HNSW

Builds a multi-layer graph with long and short connections between the vectors.

The graph can be incrementally updated.



DiskANN

Coming soon

Vector compression

Large Vectors
{ D1, D2, D3, D4, D5, …, D99, D100 }

↓

Quantization

↓

Compressed Vectors
{ D1, D2 .., D10 }

Optimized storage

RAM
Compressed vectors

Optimized for minimal SSD reads

SSD
Full vectors + graph

General availability

Azure AI extension in Azure Database for PostgreSQL

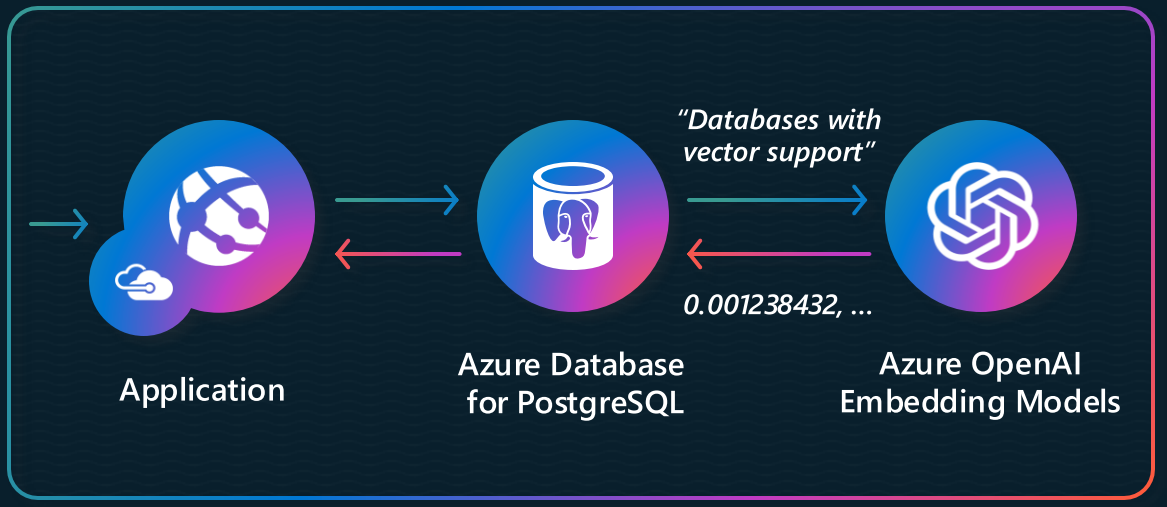
Simplified LLMs with direct use of PostgreSQL data

Vector Generation

Unique Remote + In-Database Embedding Models

Remote Embedding Models

```
SELECT * FROM <table>
ORDER BY
database_description <->
azure_openai.create_embeddings(
'text-embedding-ada-002',
'Databases with vector support')
```



Public preview

In-database embeddings in Azure Database for PostgreSQL

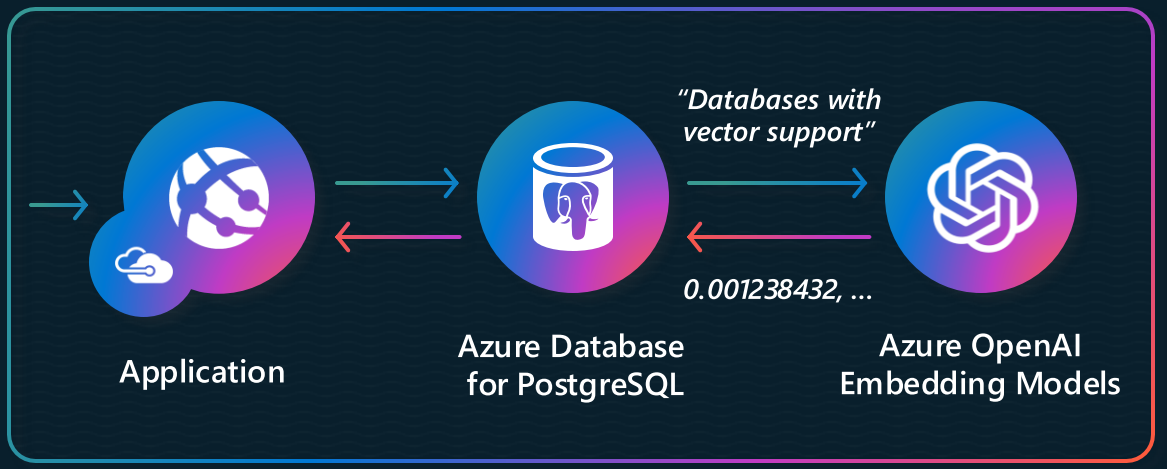
Generate vector embeddings directly in PostgreSQL databases

Vector Generation

Unique Remote + In-Database Embedding Models

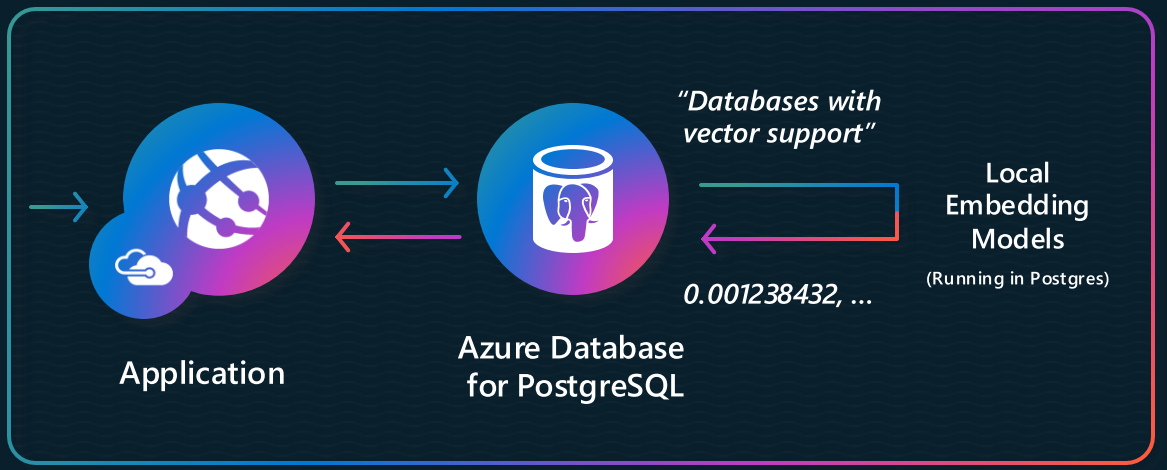
Remote Embedding Models

```
SELECT * FROM <table>
ORDER BY
database_description <->
azure_openai.create_embeddings(
'text-embedding-ada-002',
'Databases with vector support')
```



In-Database Embedding Models (Preview)

```
SELECT * FROM <table>
ORDER BY
recipe_embedding <#>
azure_local_ai.create_embeddings(
'multilingual-e5-small:v1',
'Databases with vector support')
```



AI Services integrated into Azure Postgres

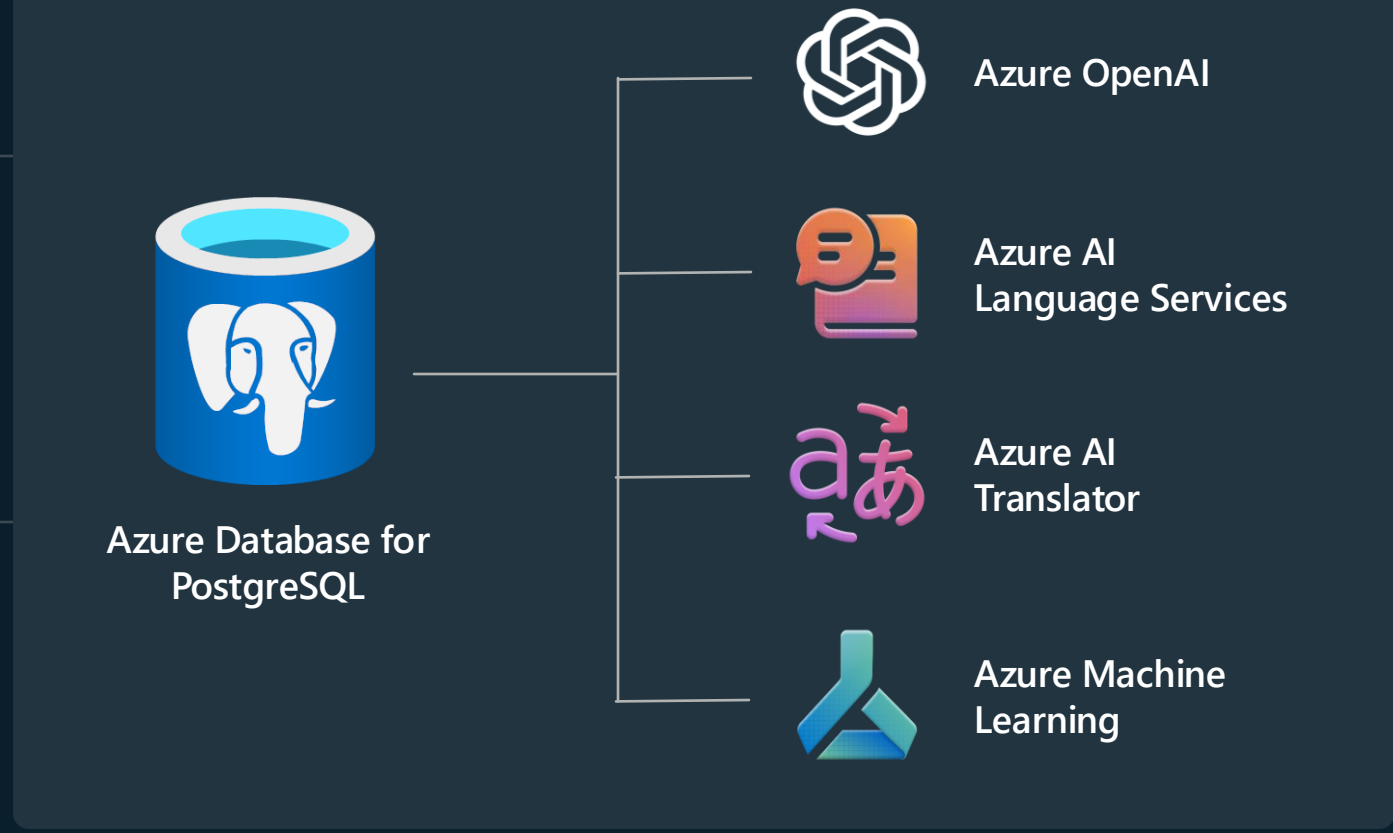
Make remote calls directly from PostgreSQL

azure_ai extension

Exceptional simplicity out of the box

- Azure OpenAI
- Azure AI Language Services
- Azure AI Translator
- Azure Machine Learning

Enables developers to rapidly adopt new AI capabilities in their solution without complex re-architecture or refactoring



Public preview

Index Recommendations in Azure Database for PostgreSQL

Automatic optimization for any PostgreSQL application

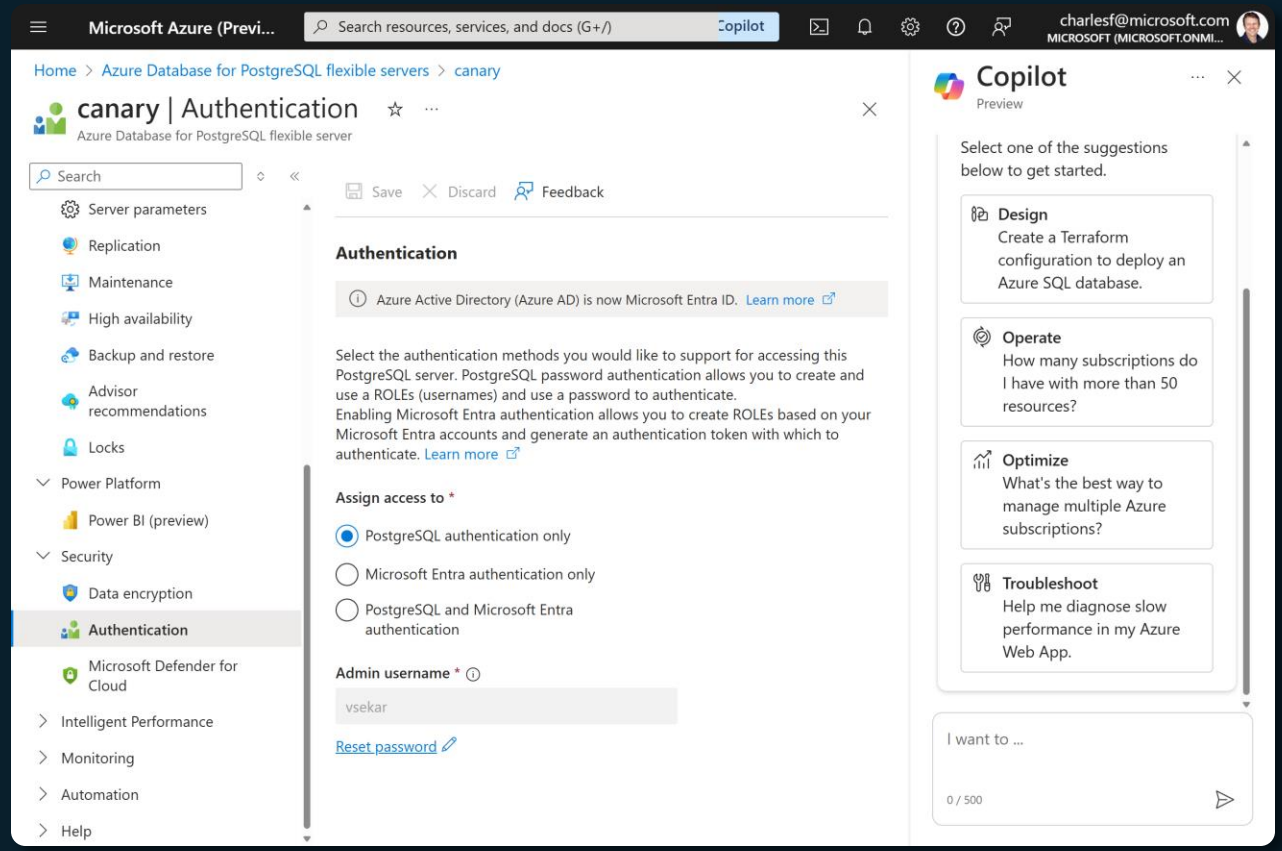
Coming soon

Azure Copilot for PostgreSQL Flexible Server

Provides a chat-based interface for querying or debugging the state of a Postgres Flexible Server

Contextual awareness of the individual server provides the most relevant responses on how to troubleshoot vs. generic documentation

Read-only scenario initially, followed by actions where the Copilot can apply the recommended changes to the server itself with user approval



Demo

Azure Database for PostgreSQL

Resources

Build AI Apps with Azure PostgreSQL
aka.ms/aitour-data-postgresql-build-ai-apps
aka.ms/aitour-data-postgresql-pgvector
aka.ms/aitour-data-postgresql-azure-ai

AI Demos
aka.ms/aitour-data-postgresql-demo
aka.ms/aitour-data-postgresql-product-demo

Azure Database for PostgreSQL
aka.ms/aitour-data-postgresql
aka.ms/aitour-data-postgresql-learn

PostgreSQL Migration Service
aka.ms/aitour-data-postgresql-migration

Try Azure Database for PostgreSQL for free

Azure Database for PostgreSQL flexible server for **free for 12 months with monthly limits:**

- **750 hours** of **Burstable B1MS** instance
- **32 GB** storage / **32 GB** backup storage

aka.ms/trypostgresql

Integrated

Intelligent

Integrated

Trusted

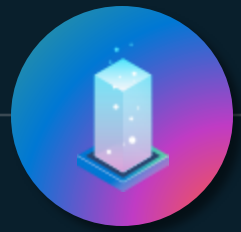
Integrated

Modern, AI apps on an open, flexible, and common platform

Integrated

Modern, AI apps on an open, flexible, and common platform

Open-source fully managed versions of MySQL, PostgreSQL, and Oracle as well as API compatibility with MongoDB, Cassandra, Gremlin, and Redis

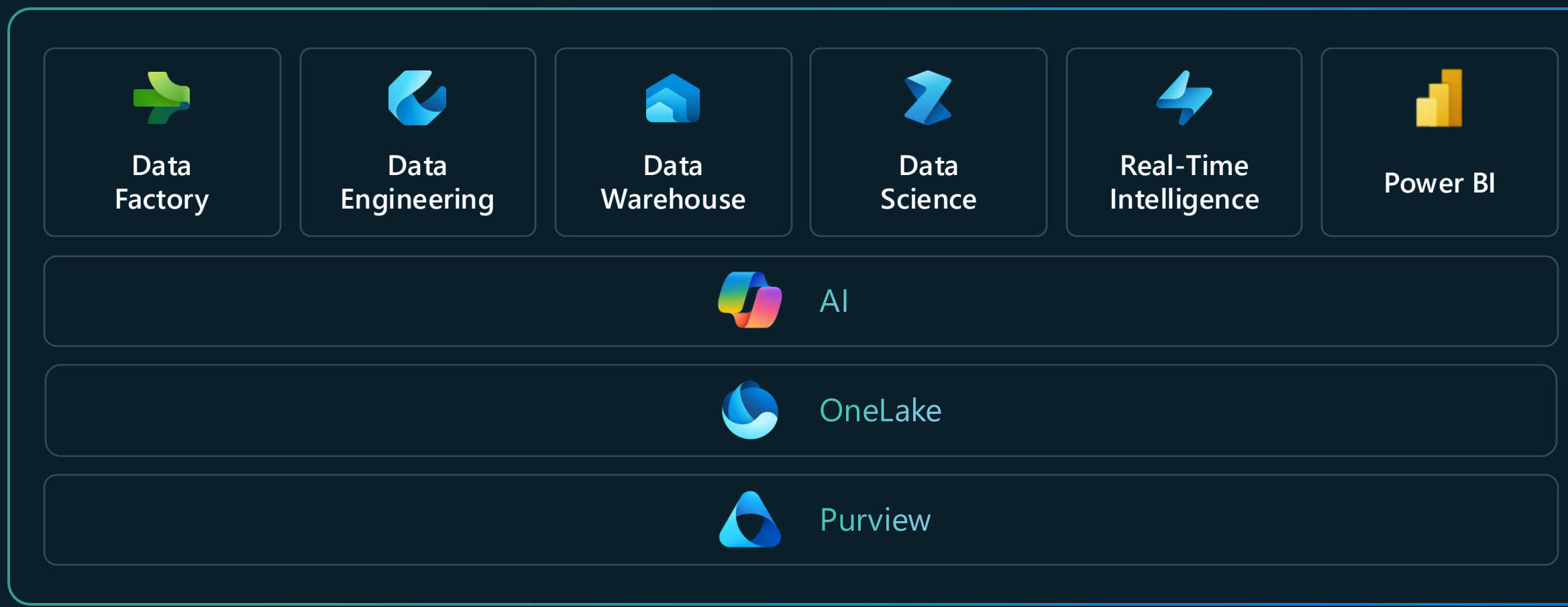


Access and ingest data continuously and seamlessly, in near real-time from any database into Microsoft Fabric for fast analytics

Integration with Azure OpenAI Service simplifies app development with prebuilt and curated generative AI models

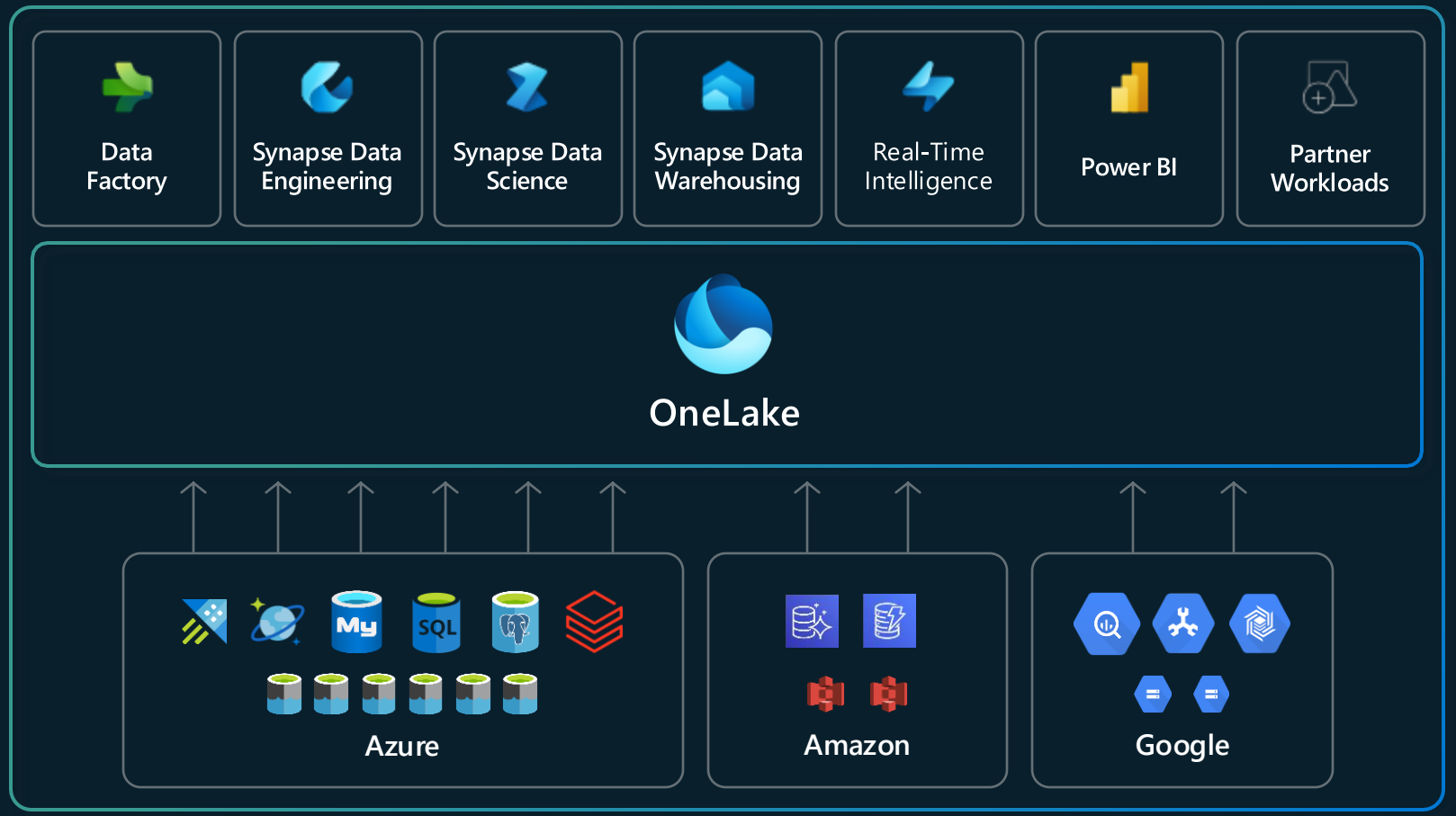
Microsoft Fabric

The unified data platform for the era of AI



Unifying data in OneLake

Mirroring of External Databases



Factionless linking of external databases, with full replicas created with a couple of clicks

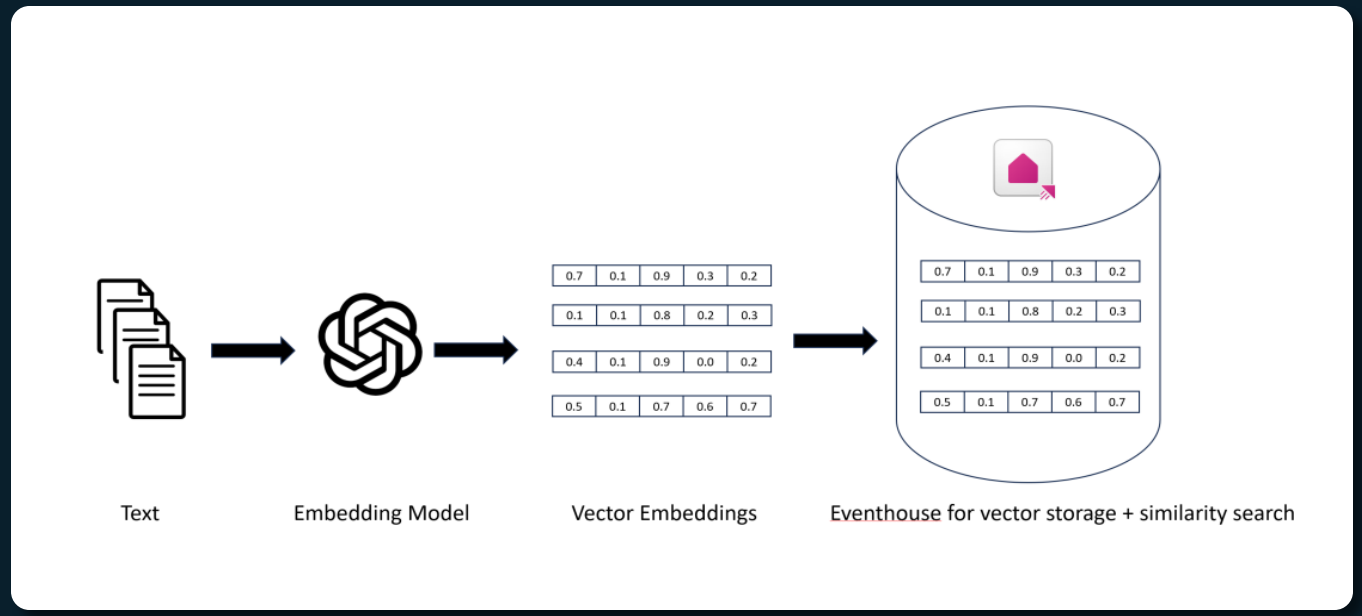
Available for both multi-cloud and on-premises databases

Real time updates of the replicas using the CDC feeds of the database

Data is stored in Delta Parquet tables, with all Fabric services instantly available

Eventhouse as a Vector Database

Vector Databases store and manage data in the form of vectors.

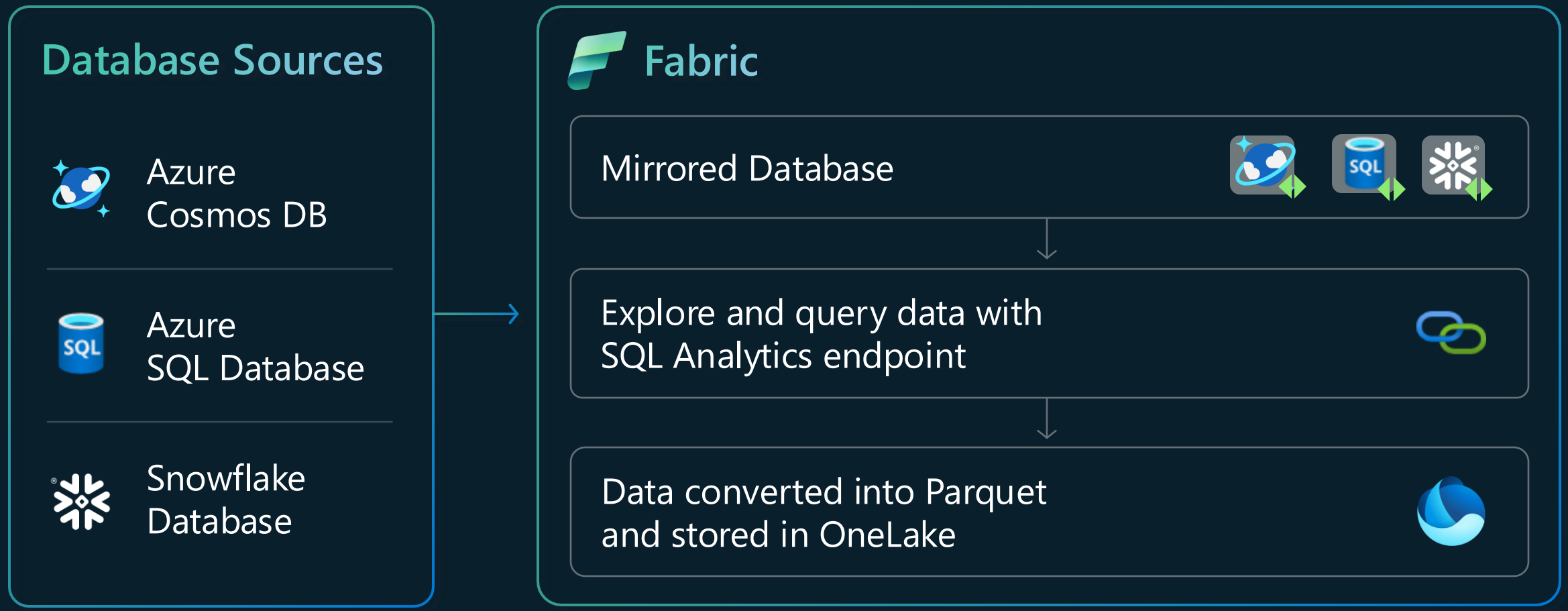


Fully integrated within Microsoft Fabric

- Embed data
- Store vectors
- Embed a query
- Query vectors
- Direct access from notebooks

Mirrored Databases

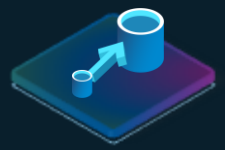
Near real-time incremental replication and inserts/updates/deletes



Azure SQL Database

A foundation to innovate with AI

Multiple ways to start infusing AI into your apps



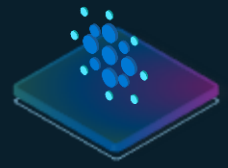
Hyperscale grows with you as your apps grow

Flexible, cloud native architecture allows storage to grow as needed

Nearly instantaneous backups and fast database restores in minutes

Native JSON for modern app development

Easily validate JSON documents and/or convert SQL data to JSON using JSON constructors



Azure OpenAI, Vectors, Azure AI Search

Implement RAG-patterns, and Hybrid Search with Azure AI Search and Azure Open AI Service

Start testing and developing for free

Free offer provides the first 100,000 vCore seconds, 32 GB of data and 32 GB of backup storage free per month for the lifetime of the subscription

Azure SQL Database and AI Scenarios

Prompt Engineering and Retrieval Augmented Generation (RAG)

Integrate SQL with Azure AI Search to build RAG application patterns

Index with Azure AI Search

Vector Search with Azure OpenAI

SDKs to use SQL as data source Small Language Models (SLM)



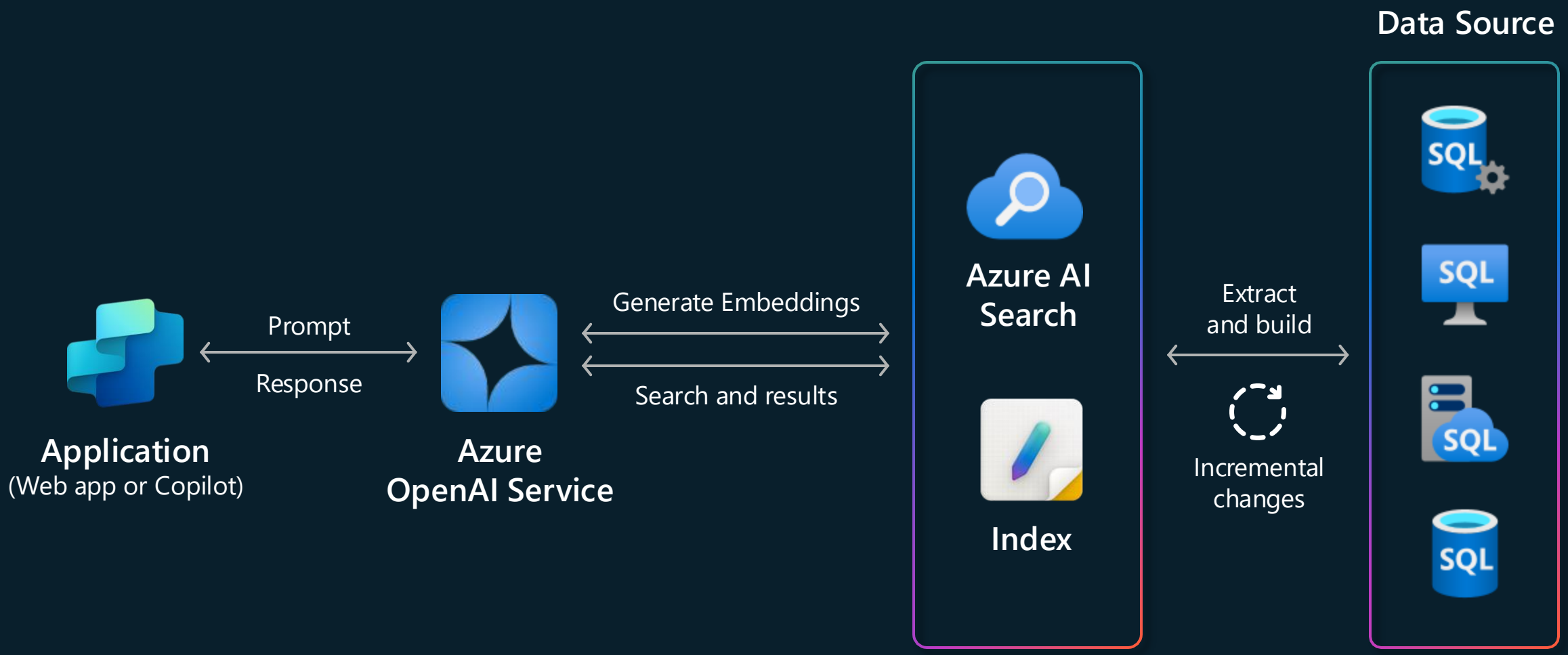

Natural Language to generate SQL queries

Orchestrate GenAI apps

Use your data in SQL for RAG applications Hybrid Search


Store vectors in SQL Invoke Models using REST API

Get smarter with your SQL data



Private preview

Azure SQL connector

Vector Store

Semantic Kernel

Entity Framework Core

General availability

Data API Builder

Provides modern REST and GraphQL endpoints to your Azure Databases

Small Language Models (SLM)

Phi-3 is a family of small, open models developed by Microsoft Research.

The Phi-3 family consists of four models:

Phi-3 Mini

Phi-3 Small

Phi-3 Medium

Phi-3 Vision

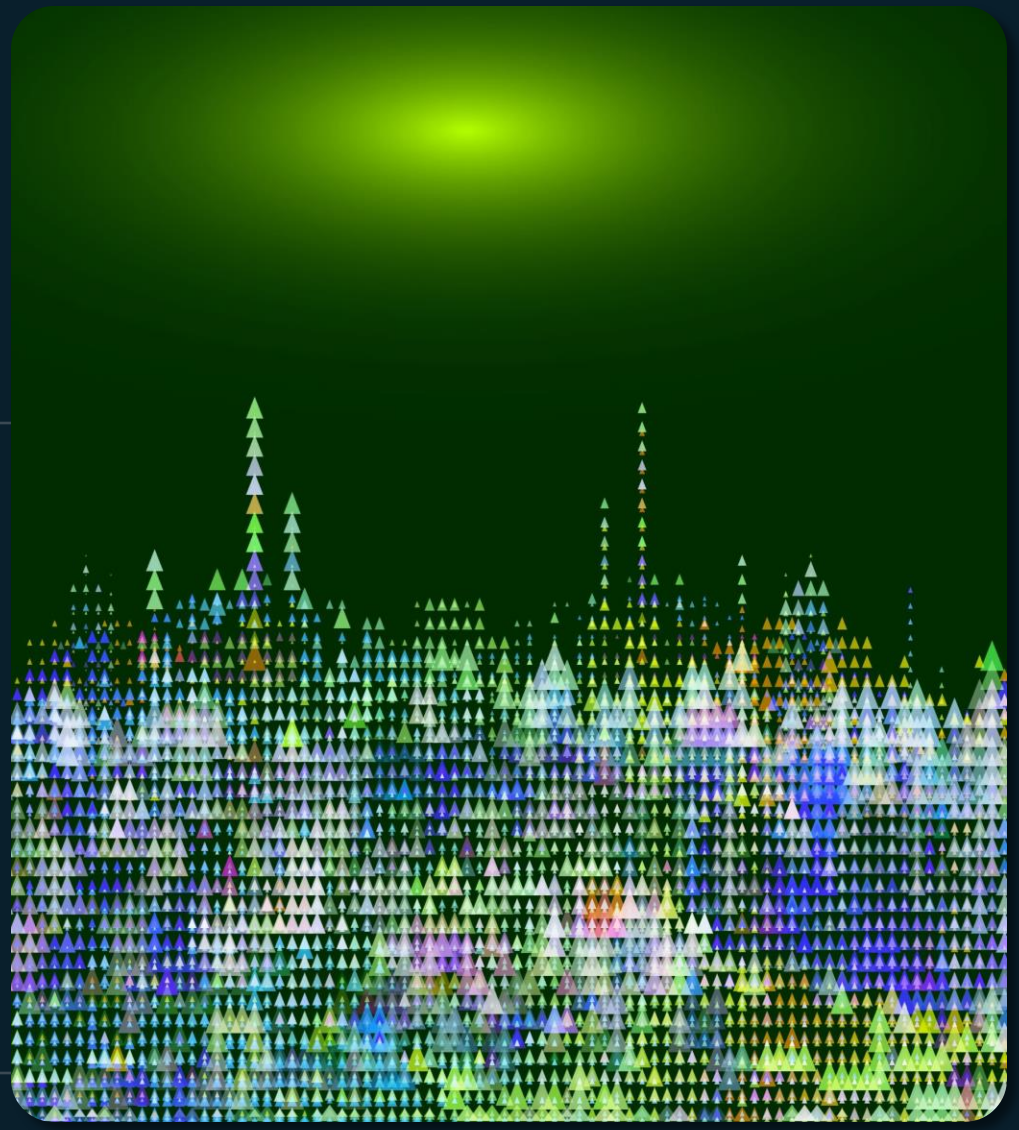
Considerations for SLMs:

RAG vs. Fine tuning

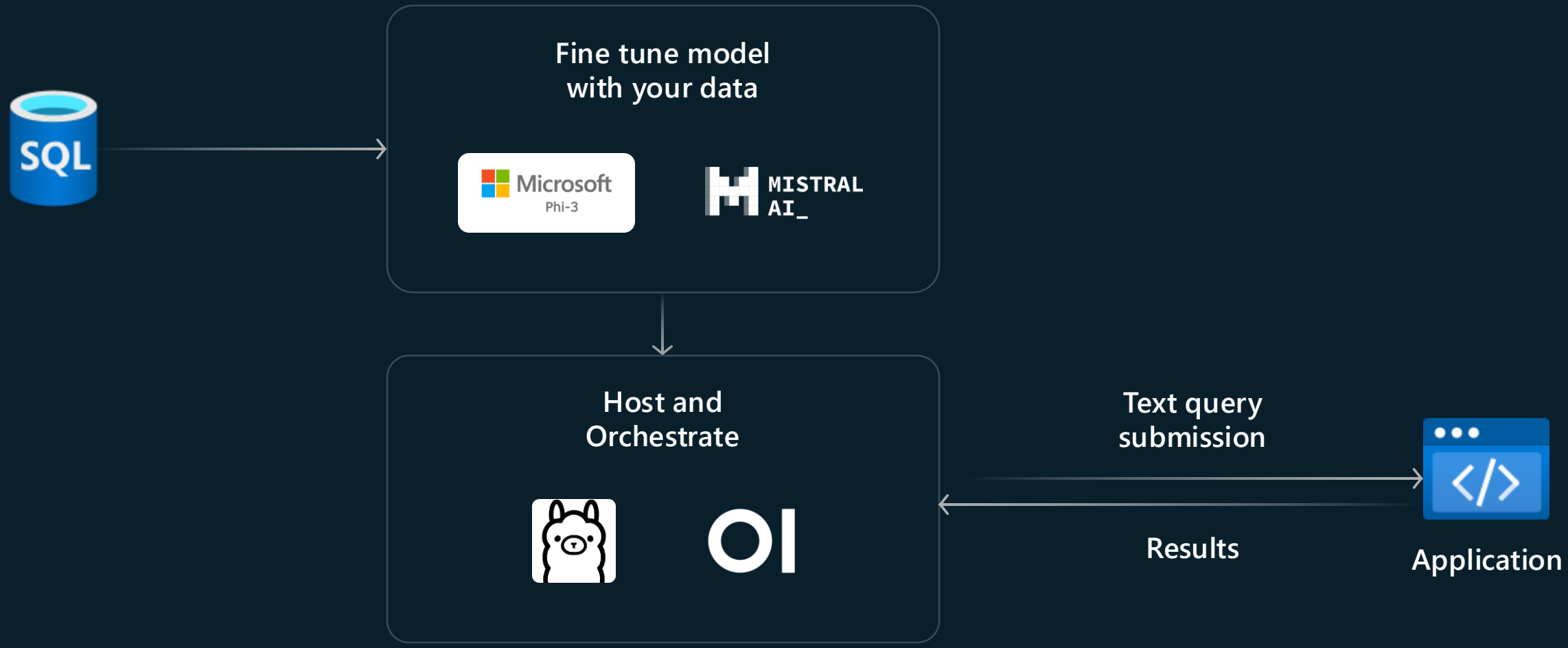
Latency

Total cost of ownership (TCO)

Optimizations



App Architecture with SLMs



Demo

Azure SQL Database

Resources

SQL AI Docs
aka.ms/aitour-data-azuresql-aidocs

SQL AI samples
aka.ms/aitour-data-azuresql-aisamples

Videos from the Azure SQL team
aka.ms/aitour-data-azuresql-dataexposed
aka.ms/aitour-data-azuresql-essentials

Vector private preview signup
aka.ms/aitour-data-azuresql-vector-eap

Try Azure SQL Database free of charge

One Azure SQL Database per Azure subscription with 100,000 vCore seconds compute every month

32 GB data storage + 32 GB backup storage

aka.ms/aitour-data-azuresql-freedboffer

Trusted

Intelligent

Integrated

Trusted

Trusted

A sustainable and reliable platform so your data is always available, safe, and compliant

Trusted

A sustainable and reliable platform so your data is always available, safe, and compliant

Stronger security posture with built-in controls like data masking, enterprise-grade encryption, and intelligent threat detection

Highest financially-backed availability SLAs in the industry: up to 99.999% availability

"Achieve high availability with Azure Cosmos DB," Microsoft, 2020.

More certifications than any other public cloud provider, including ISO 27001, HIPAA, FedRAMP, SOC 1, SOC 2, and many international specifications

Public preview

Cross-region disaster recovery in Azure Cosmos DB for MongoDB vCore

Ensure uptime with data replicas across separate regions

Session resources

- Next steps
- Resources
- Related sessions

Find all of this and more
on the session details page.

aka.ms/aitour/databases