Microsoft Federal Developer Summit

Building AI Solutions

Ujval Gandhi | Cloud Architect – Microsoft Federal | November 13, 2024

AGENDA

- Why invest in getting your data estate ready before Gen AI
- What is required for this digital transformation?
- How do we go about implementing a data estate strategy?

Why invest in getting your data estate ready?

Data Volume

46%

Increase in average enterprise data volume annually

Data is growing exponentially in most organizations

Data Velocity

30%

Data that will be generated in real time by 2025

Speed at which data (Structured, Semi-Structured and Unstructured) is getting generated is also at an exponential point

Data Variety

67%

Expected growth in data formats from 2021-2026

Variety of data sources has exploded – not just limited to spreadsheets/relational databases

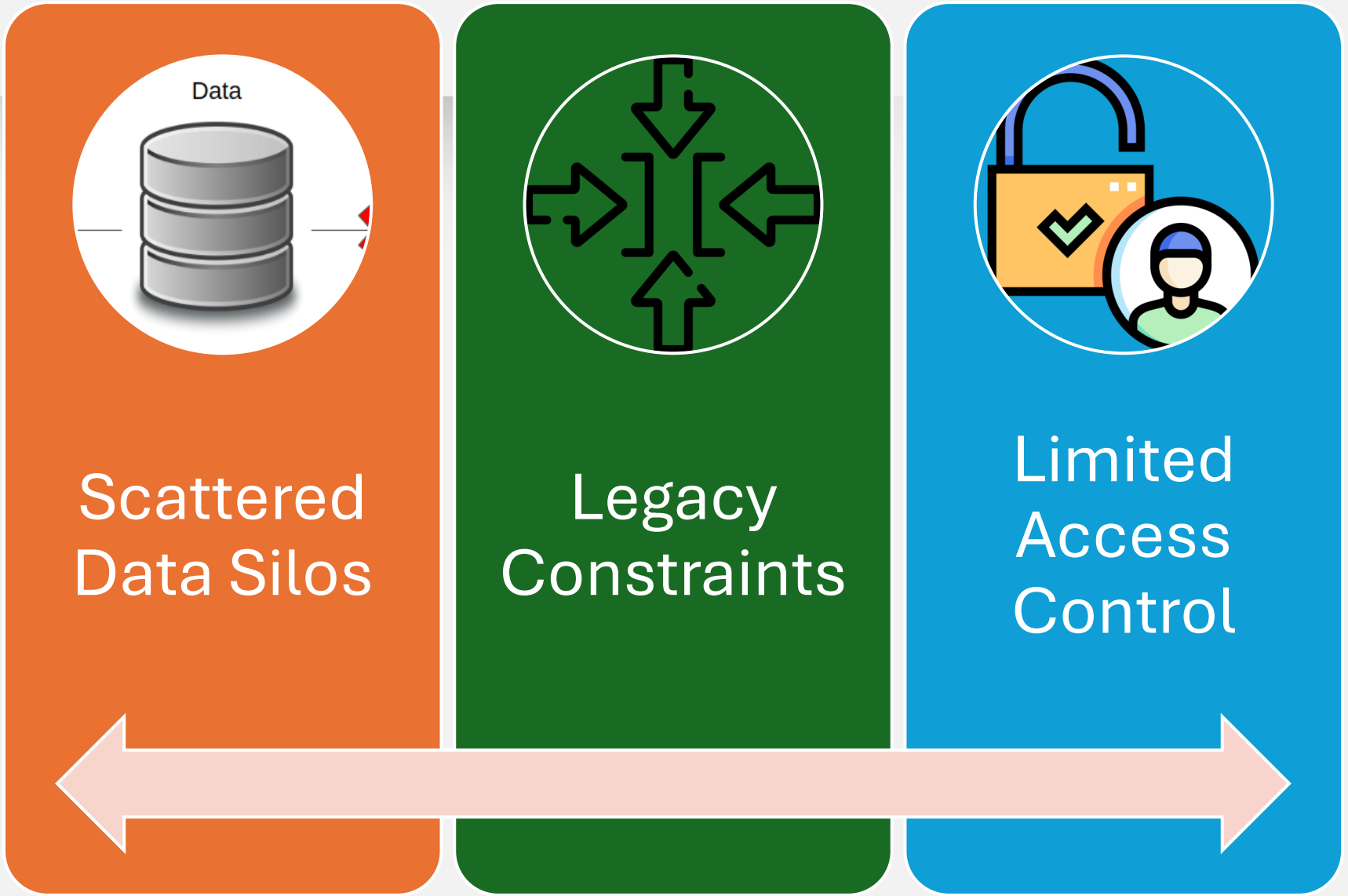
Why invest in getting your data estate ready?

Through 2025,

30%

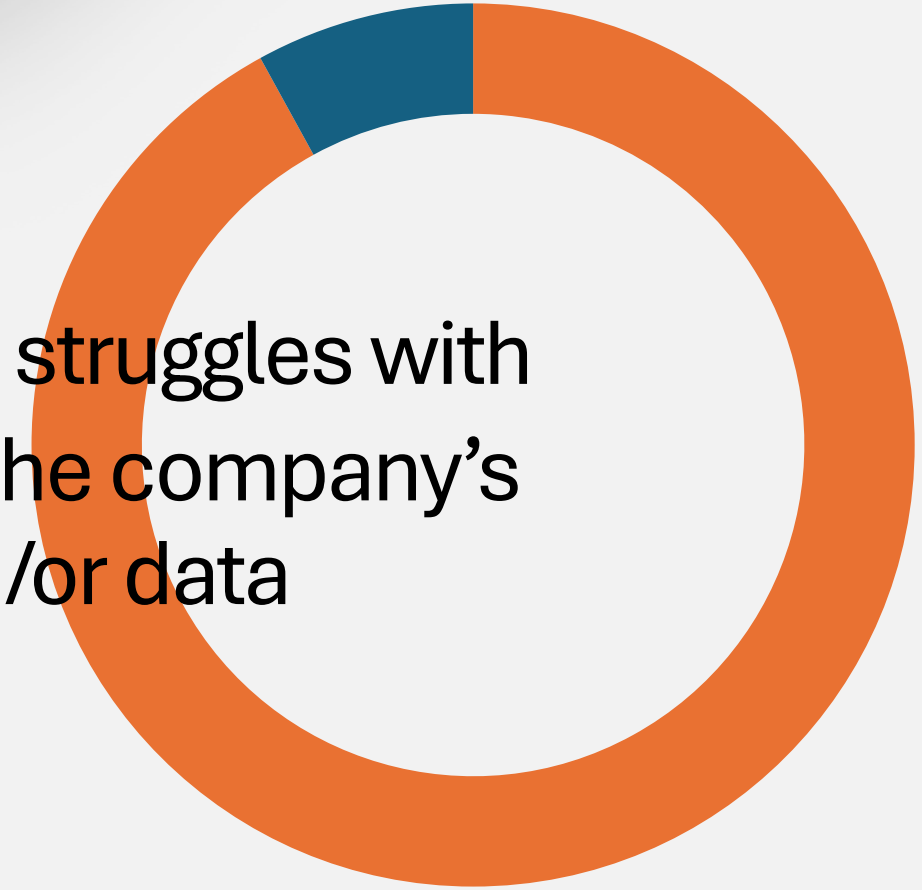
of generative AI (GenAI) projects will be abandoned after proof of concept due to poor data quality, inadequate risk controls, escalating costs or unclear business value.

Why invest in getting your data estate ready?

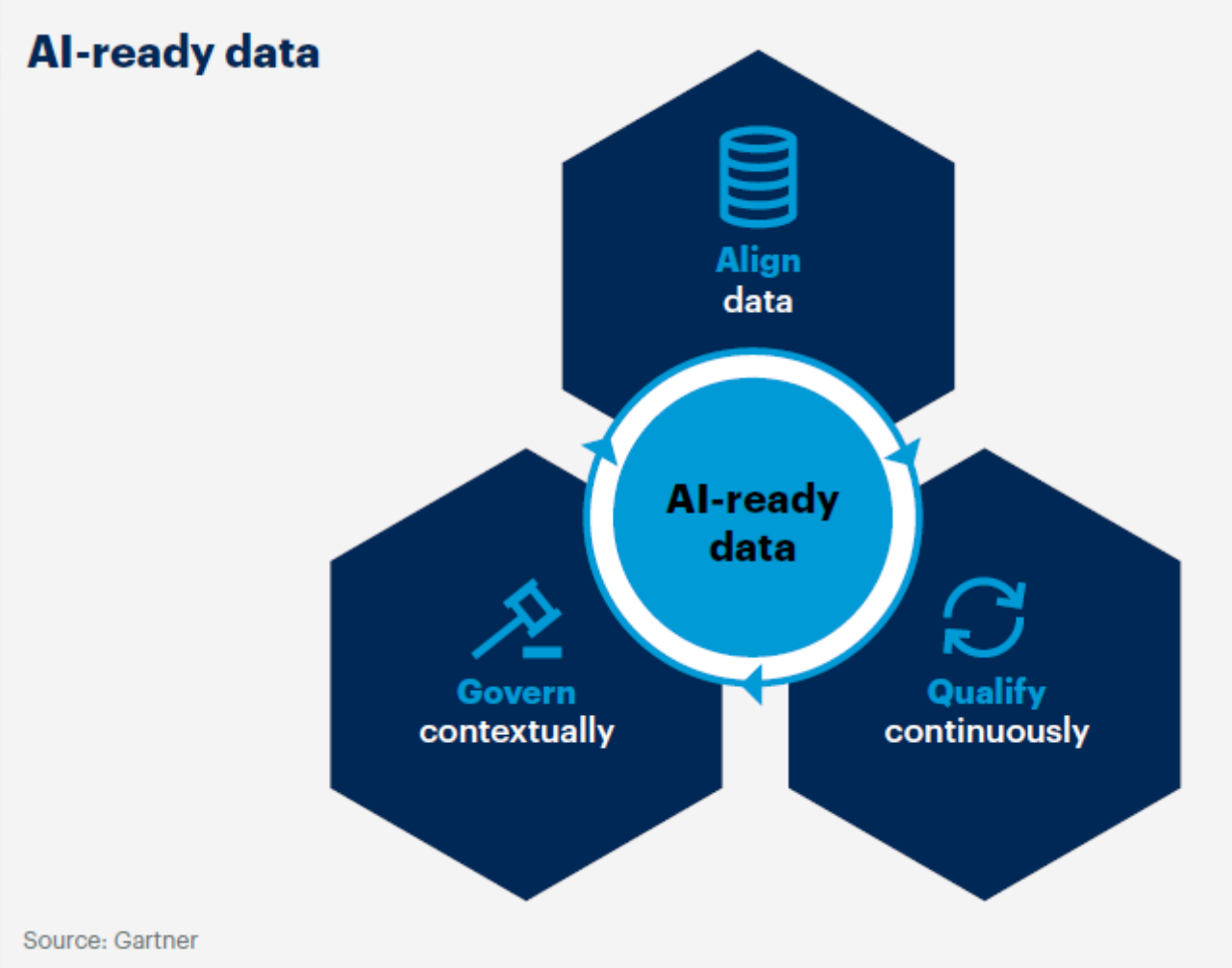


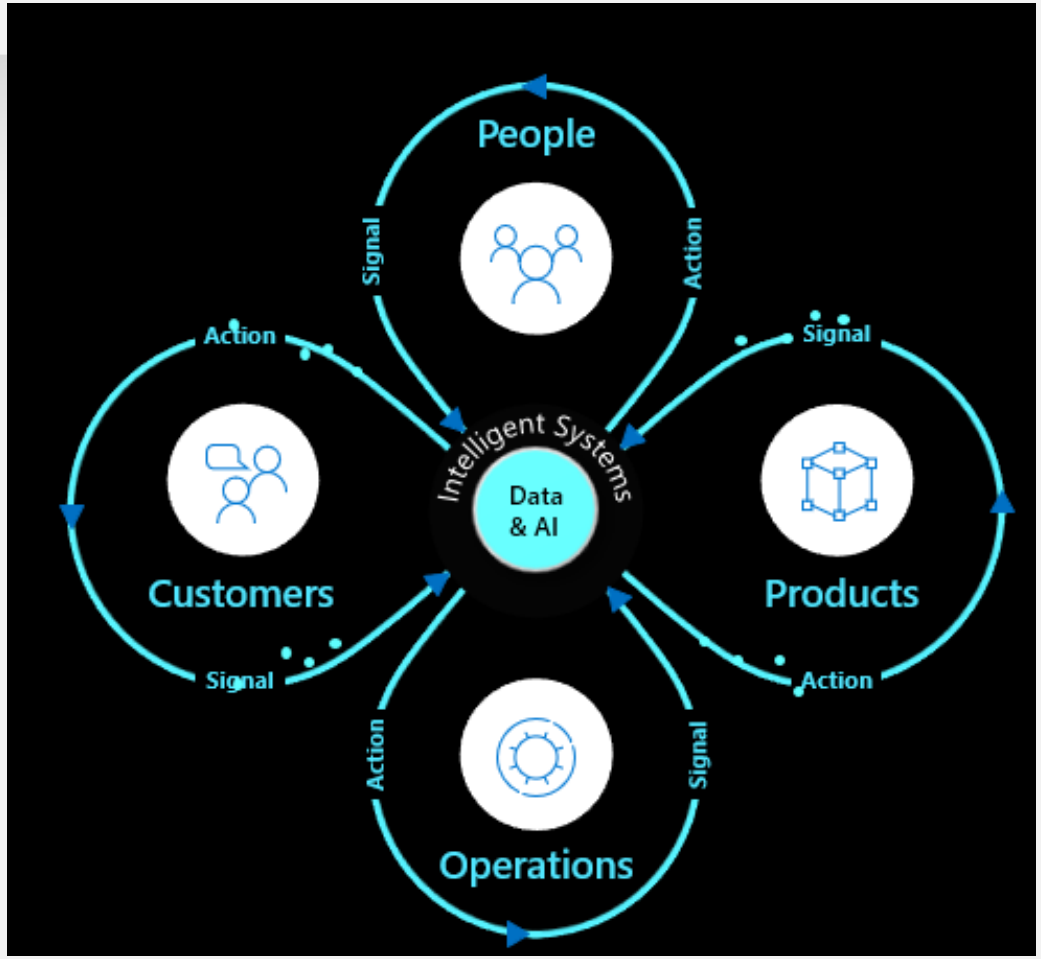
Why invest in getting your data estate ready?

92% of the time an organization struggles with data, it links back to an issue with the company's data strategy, data governance and/or data management.

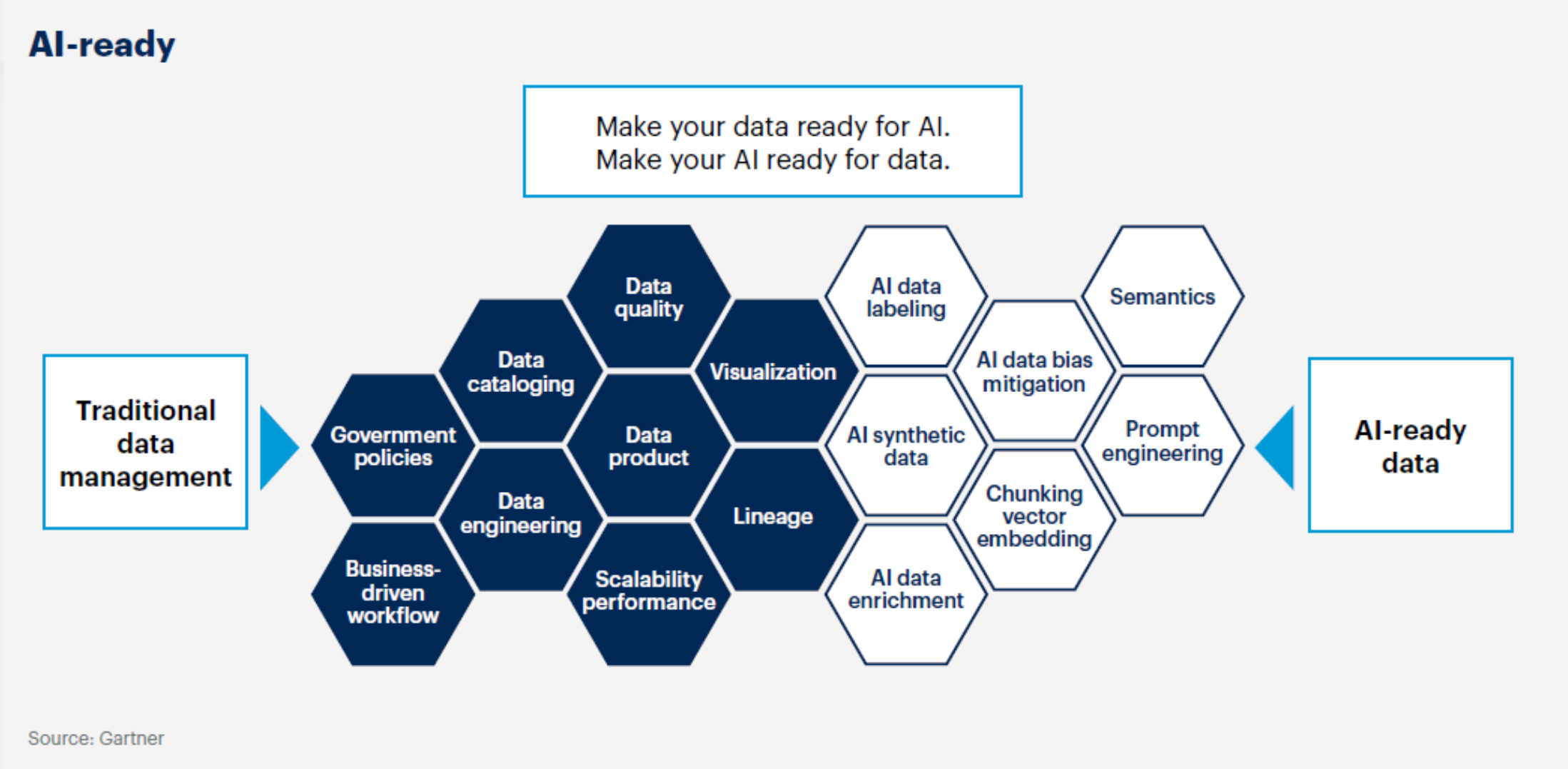


Why invest in getting your data estate ready?

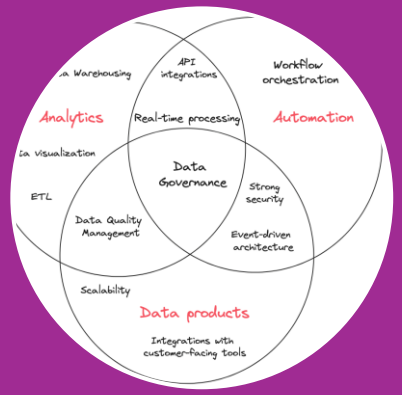




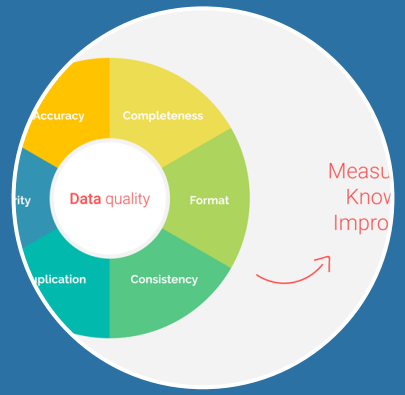
Why invest in getting your data estate ready?



Why invest in getting your data estate ready?



Does your organization's data align with use case requirements?

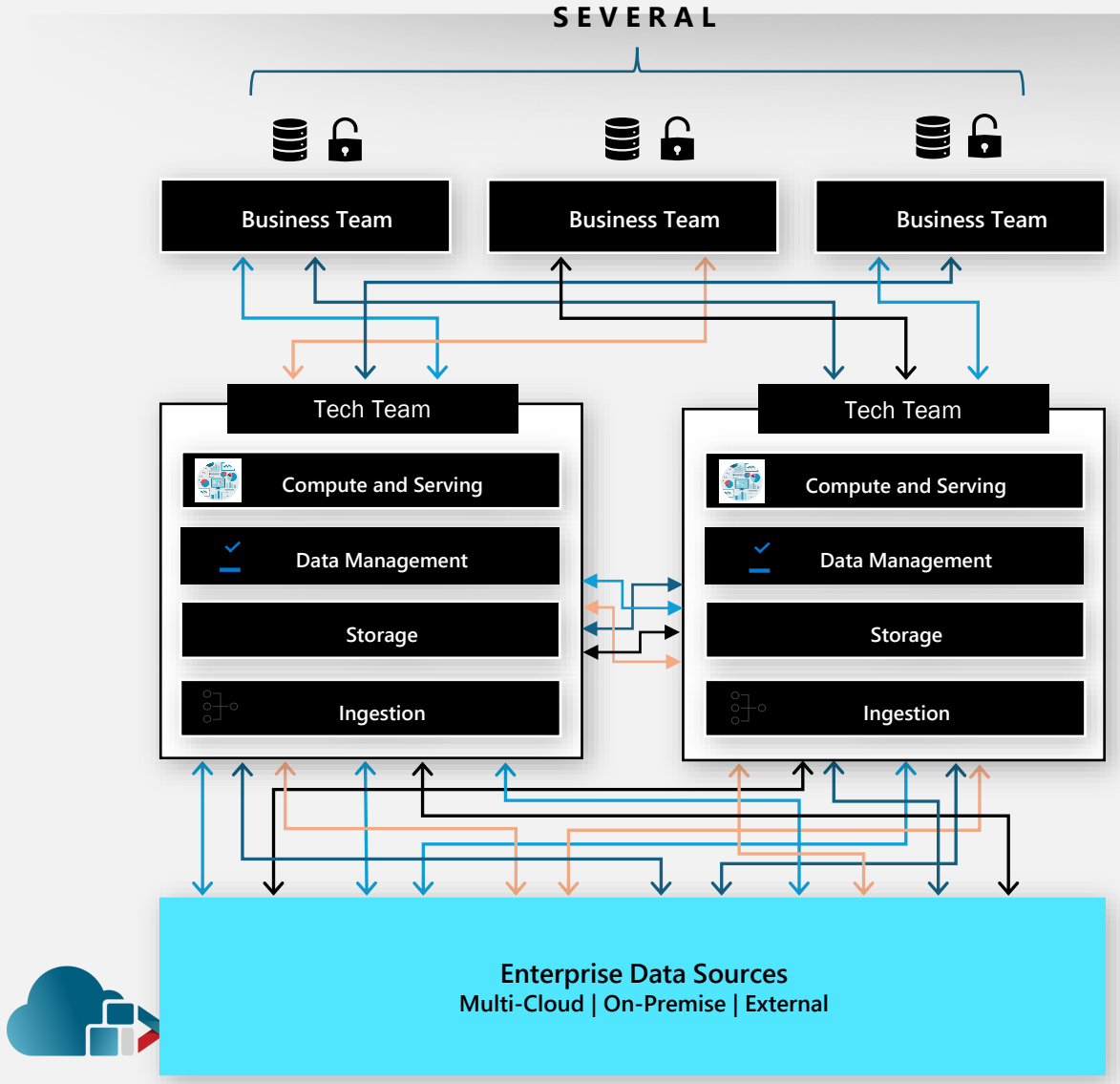


How do you qualify data use to meet AI expected confidence requirements?

How do you govern AI ready data in the context of the use case?

What is required for this digital transformation?

The starting line: a complex, organically evolved data estate



1. Data copies and infrastructure inefficiencies
2. Limited interoperability
3. Data exposure risks

What is required for this digital transformation?

The finish line : powerful insights and a single source of truth

Unifying your hybrid and multi-cloud environments is critical for resilient business transformation and for optimizing the business value that data & analytics can provide

From		To
Fragmented, compartmentalized, and siloed cloud environments	→	An analytics platform that connects to & can analyze all your on-prem, cloud-based, and third-party data sources
High operational costs due to high data storage processing	→	Lakehouse approach that makes it easier and more efficient to store data
Siloed data access issues	→	Data virtualizing in an open Lakehouse to ensure everyone has access to the same data sets
Complex, slow-to-ramp, and lagging analytics solutions	→	Spin up analytics solutions quickly with minimal set-up, deployment, and latency

How do we go about implementing a data estate strategy?

How do we go about implementing a data estate strategy?

Today's data realities

Volume

Variety

Velocity

What data do I have?

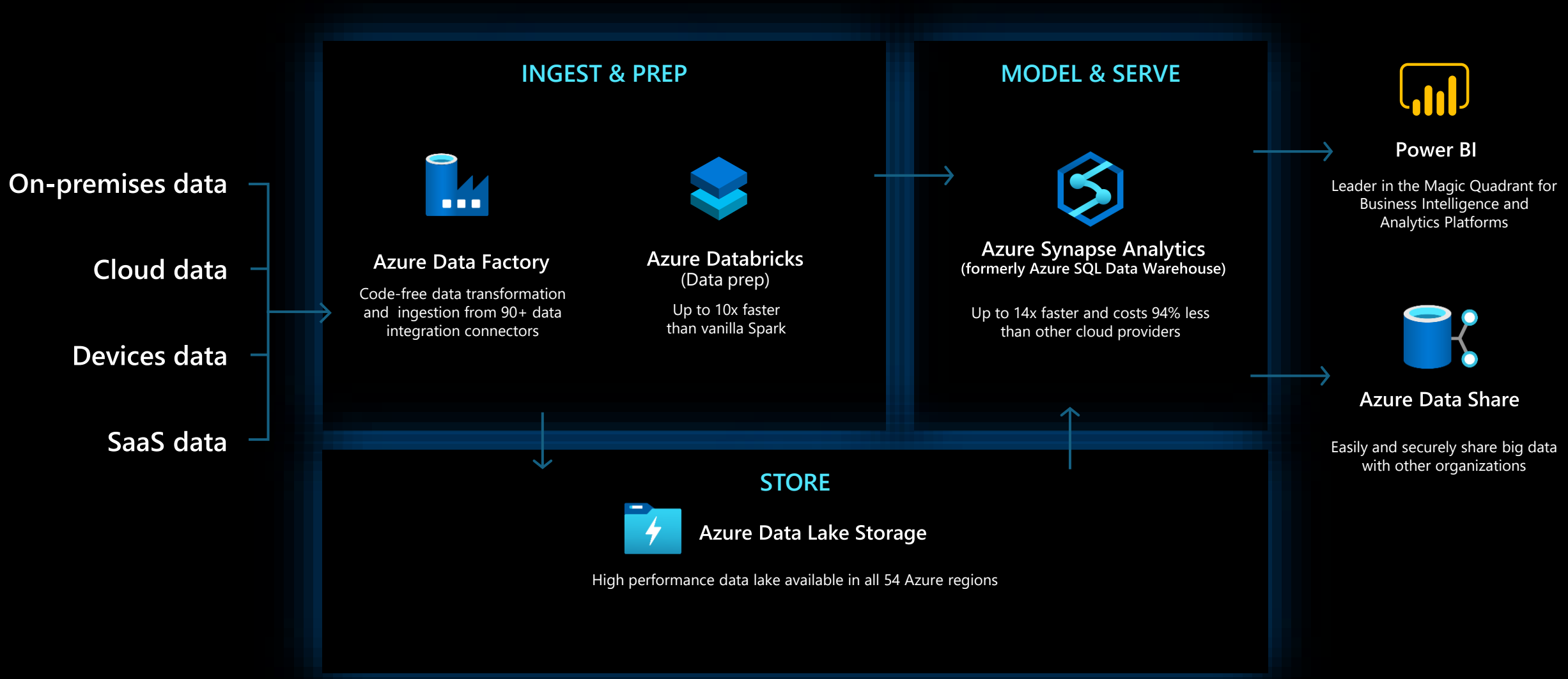
Is it trustworthy?

Can people access the data needed to make the right decisions?

How can I enable faster business insights?

What's my compliance exposure?

Azure Analytics

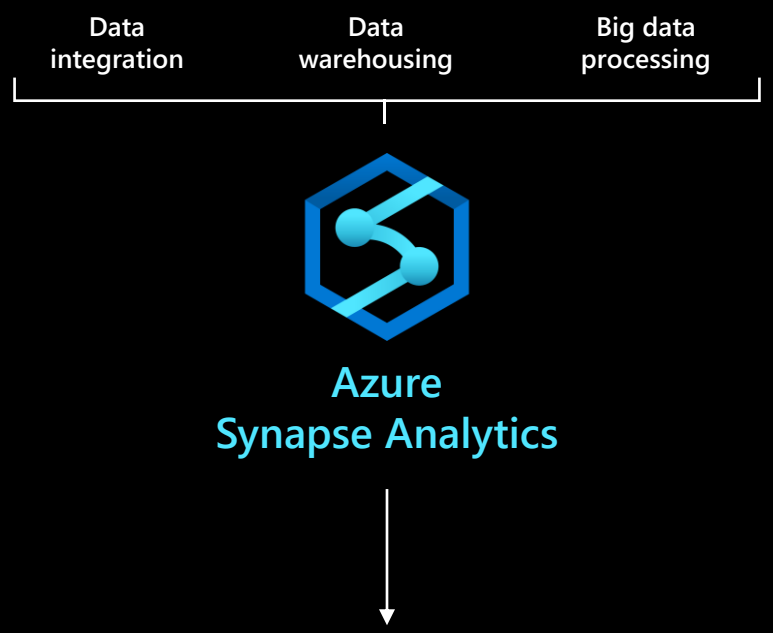


Azure Synapse Analytics

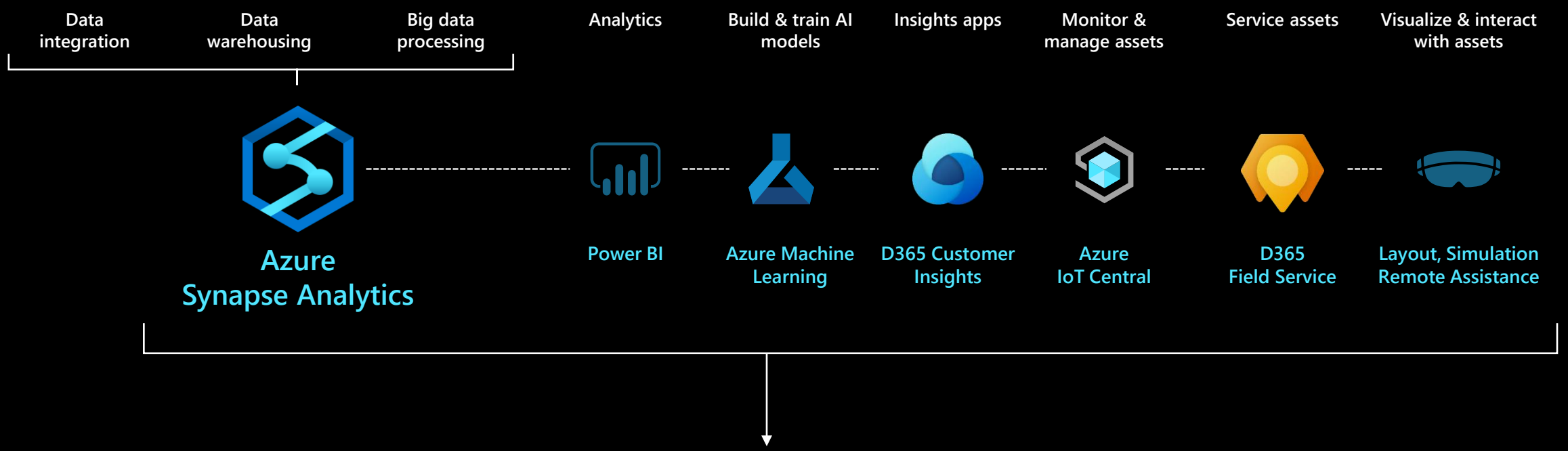
The first unified, cloud native platform for converged analytics

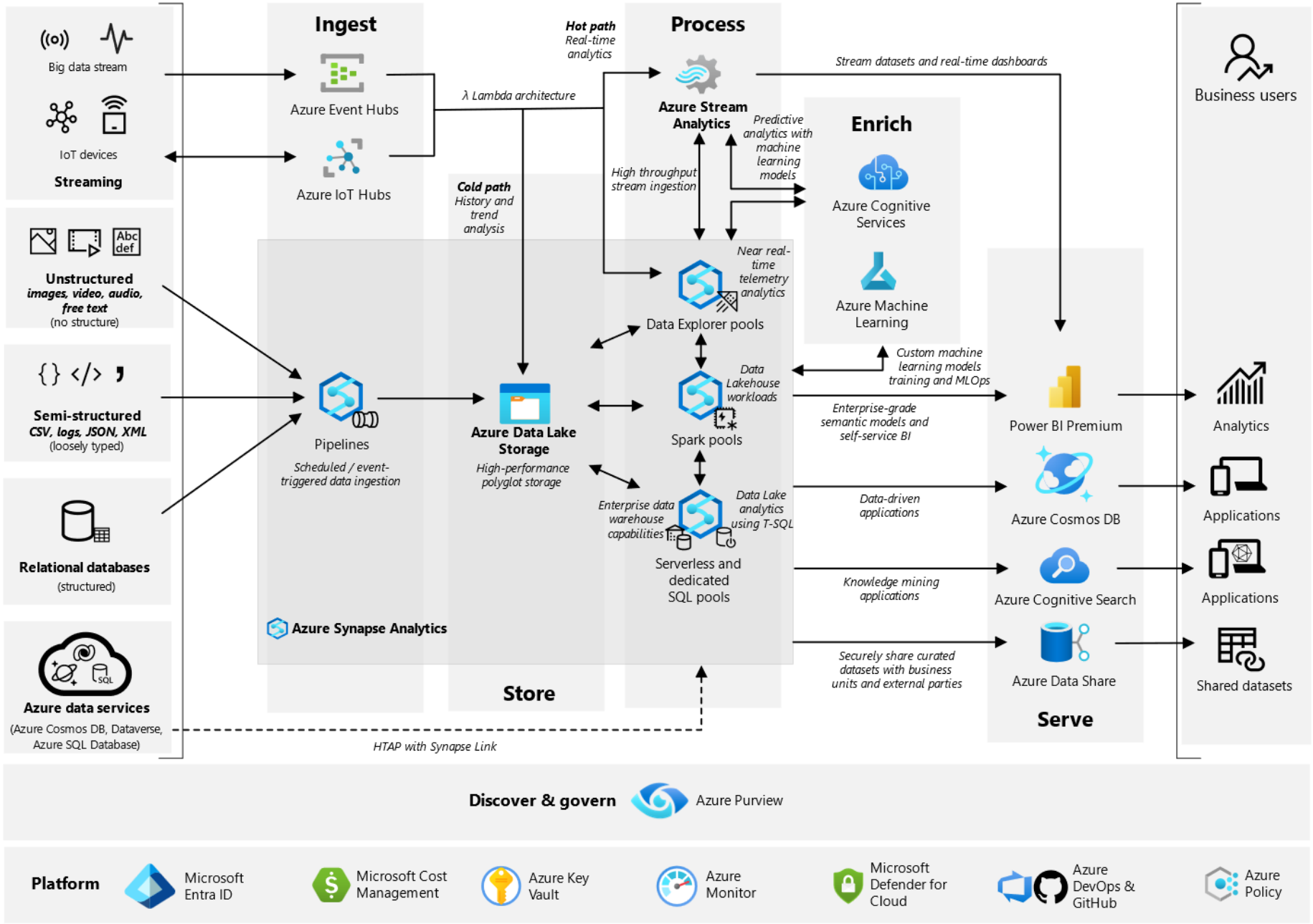
Azure Synapse is the only unified platform for analytics, blending big data, data warehousing, and data integration into a **single cloud native service** for end-to-end analytics at cloud scale.

Create an engine for business-changing insights with seamless ecosystem integration



Create an engine for business-changing insights with seamless ecosystem integration





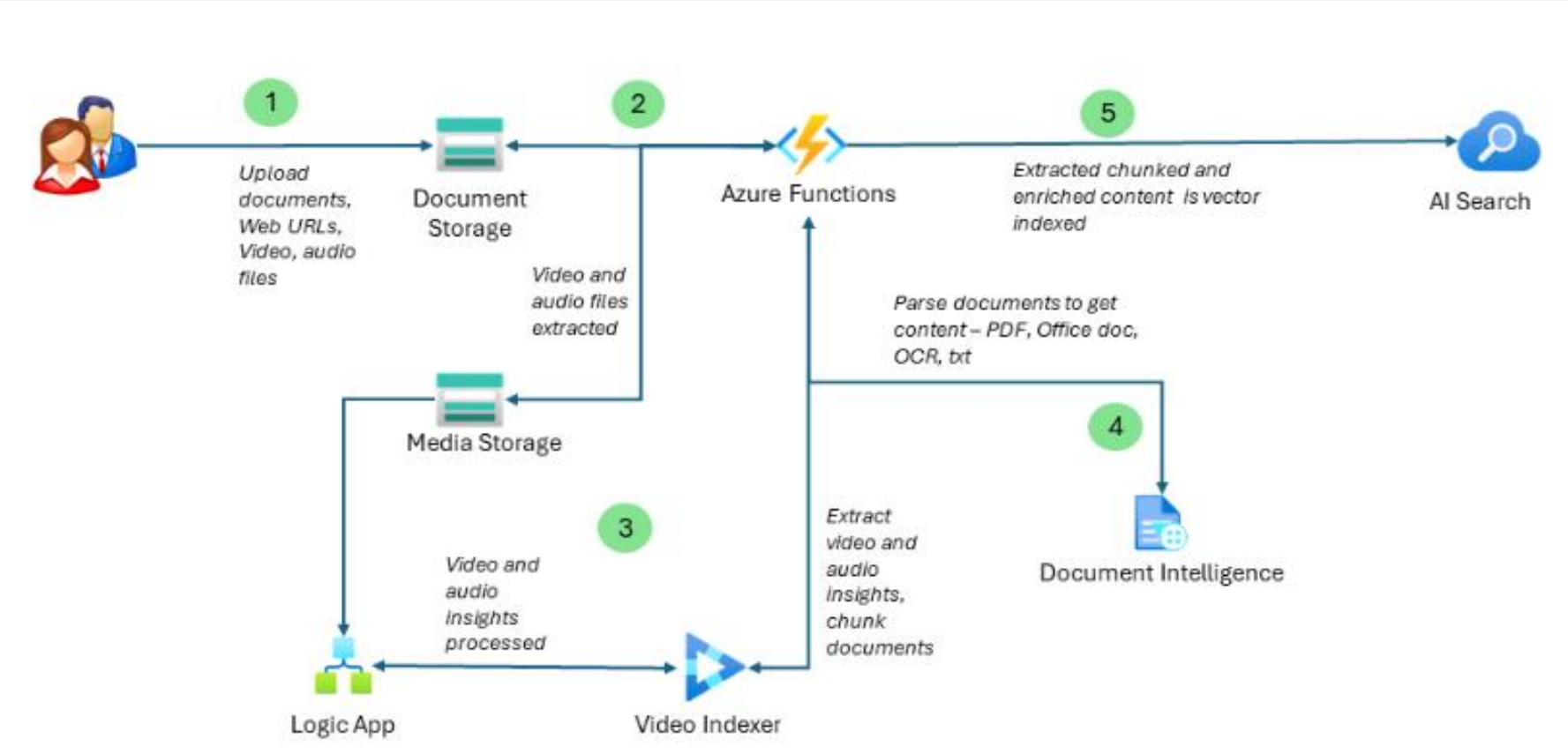
Microsoft Azure

How do we go about implementing a data estate strategy?

Federal Agency had a document repository in a variety of different places – SharePoint Online, SharePoint but hosted on another tenant and external facing documents

They wanted to ingest all the documents in a consolidated ChatBot type of solution

How do we go about implementing a data estate strategy?



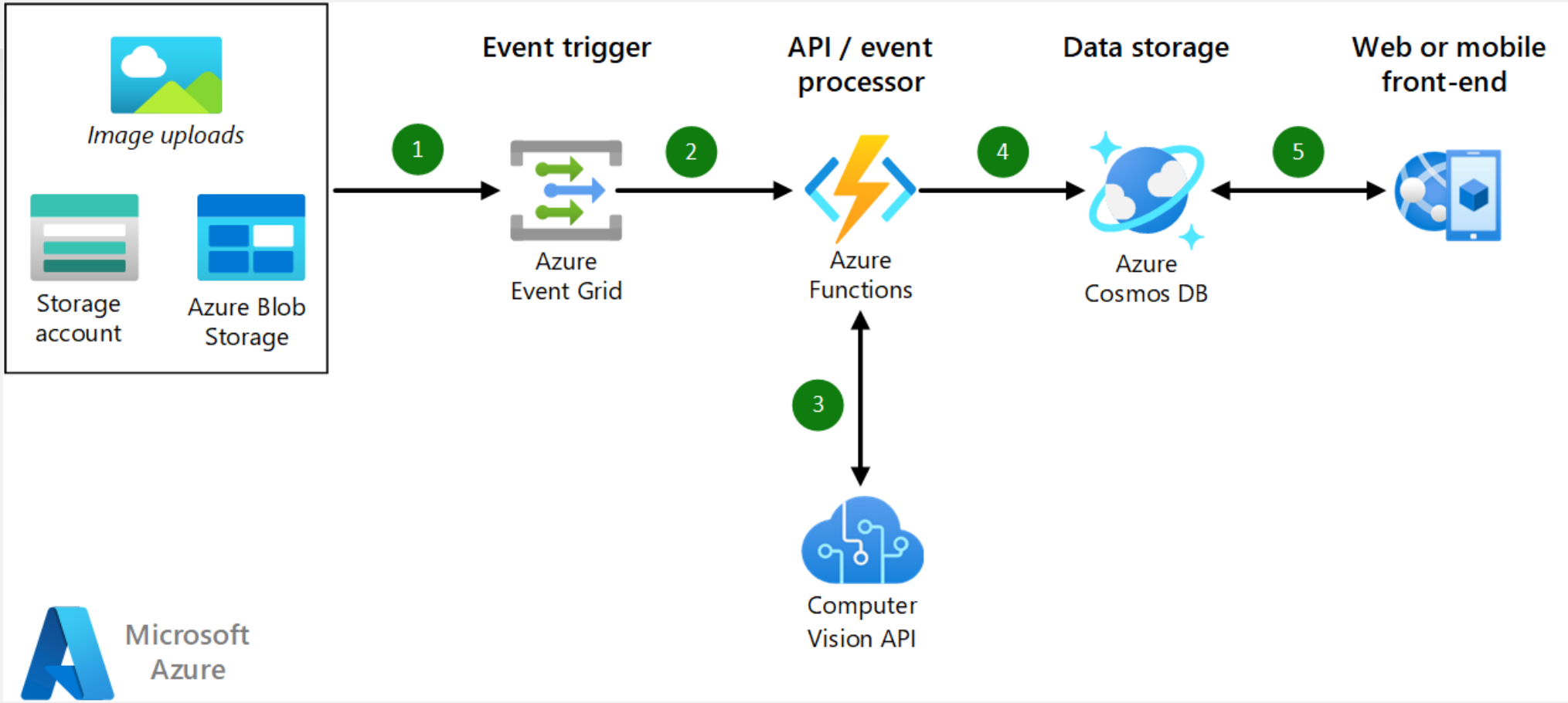
1. Place the URL of the Sharepoint repository in a text file
2. Use of Azure Functions to process the file and copy to the targetcontainer
3. Azure Function will process each file, detect each file type and route to the appropriate underlying service for data augmentation

How do we go about implementing a data estate strategy?

The next phase of discovery uncovered several image files in another repository but they were not in a “searchable” format

Azure Cognitive Services to the rescue !

How do we go about implementing a data estate strategy?



How do we go about implementing a data estate strategy?

Python Copy

```python
# Create a dataframe with the image URLs
df = spark.createDataFrame([
        ("<replace with your file path>/dog.jpg", )
    ], ["image", ])

# Run the Azure AI Vision service. Analyze Image extracts infortmation from/about
analysis = (AnalyzeImage()
    .setLinkedService(ai_service_name)
    .setVisualFeatures(["Categories","Color","Description","Faces","Objects","Tag
    .setOutputCol("analysis_results")
    .setImageUrlCol("image")
    .setErrorCol("error"))

# Show the results of what you wanted to pull out of the images.
display(analysis.transform(df).select("image", "analysis_results.description.tags
```

JSON Copy

```
["dog","outdoor","fence","wooden","small","brown","building","sitting","front","b
```

How do we go about implementing a data estate strategy?

Python Copy

```
df = spark.createDataFrame([
        ("<replace with your file path>/ocr.jpg", )
    ], ["url", ])

ri = (ReadImage()
    .setLinkedService(ai_service_name)
    .setImageUrlCol("url")
    .setOutputCol("ocr"))

display(ri.transform(df))
```

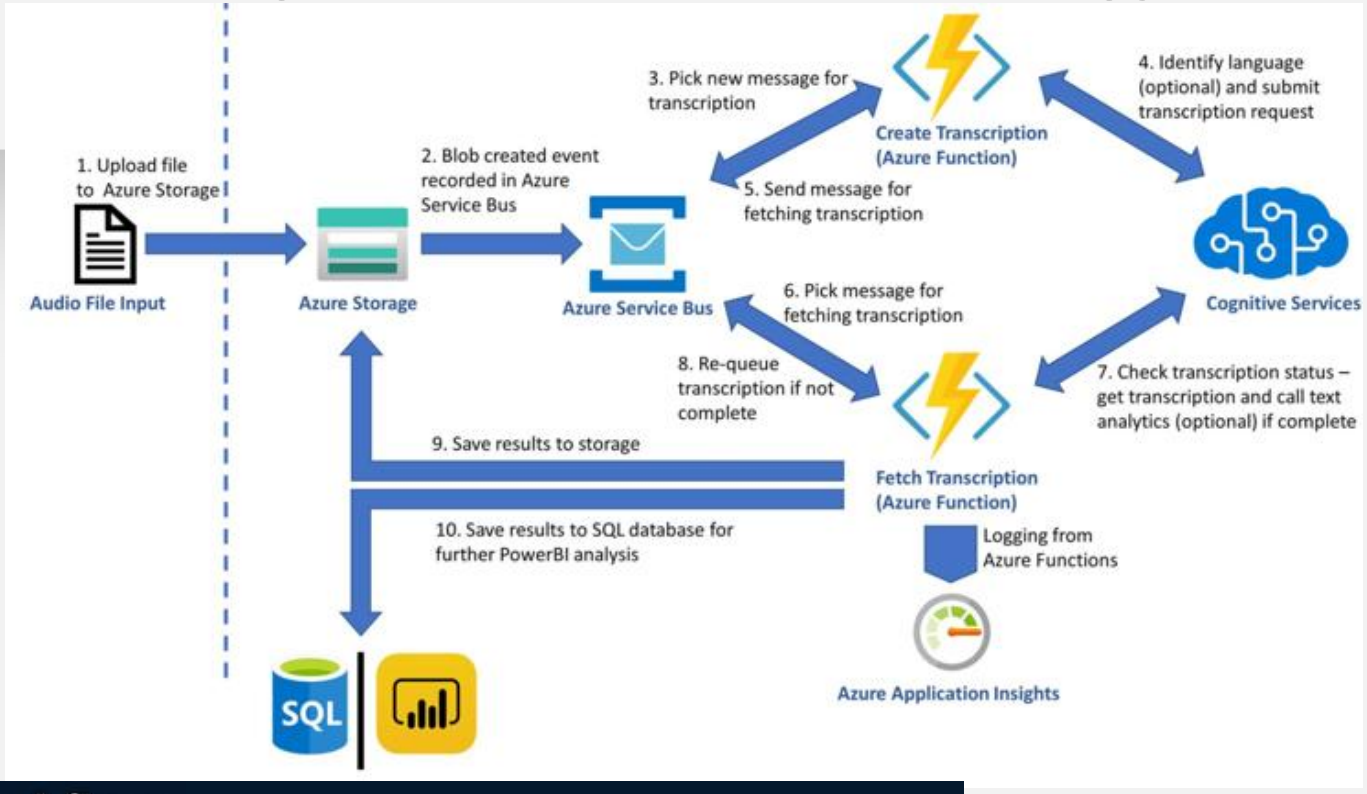
url ↓	ReadImage_ececcae91fa0_er...	ocr
https://mmlspark.bl...	undefined	▼ "{"status":"succeeded","createdDateTime":"2021-10-2..."

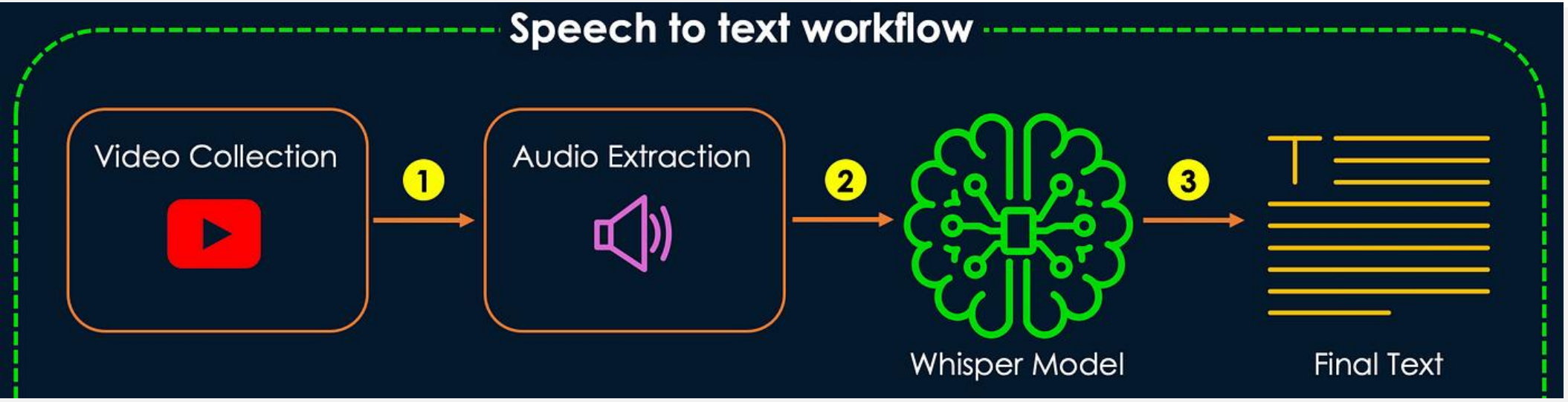
status: ""succeeded""
createdDateTime: ""2021-10-21T10:39:29Z""
lastUpdatedDateTime: ""2021-10-21T10:39:30Z""
▼ analyzeResult: "{"version":"3.0.0","readResults":[{"angle":20.4048..."
version: ""3.0.0""
▼ readResults: "[{"angle":20.4048,"page":1,"height":1008,"unit":"p..."
▼ 0: "{"angle":20.4048,"page":1,"height":1008,"unit":"pi..."
angle: "20.4048"
page: "1"
height: "1008"
unit: ""pixel""
▼ lines: "[{"boundingBox":[182,511,652,677,609,796,140,630],..."
▼ 0: "{"boundingBox":[182,511,652,677,609,796,140,630],"..."
▶ boundingBox: "[182,511,652,677,609,796,140,630]"
text: ""CLOSED""
▶ words: "[{"boundingBox":[181,512,652,681,601,796,140,630],..."
▼ 1: "{"boundingBox":[146,643,603,813,593,841,136,671],"..."
▶ boundingBox: "[146,643,603,813,593,841,136,671]"
text: ""WHEN ONE DOOR CLOSES, ANOTHER""
▶ words: "[{"boundingBox":[146,644,220,672,210,699,136,671],..."
▼ 2: "{"boundingBox":[75,676,644,887,633,919,65,706],"te..."
▶ boundingBox: "[75,676,644,887,633,919,65,706]"
text: ""OPENS. ALL YOU HAVE TO DO IS WALK IN""
▶ words: "[{"boundingBox":[75,677,161,710,152,739,65,706],"t..."
width: "756"

How do we go about implementing a data estate strategy?

The same discovery phase around images also uncovered an audio repository

You have a choice of using an AI Service – Azure Speech to Text to handle Transcription OR use the new Azure OpenAI Whisper model to handle language detection and transcription

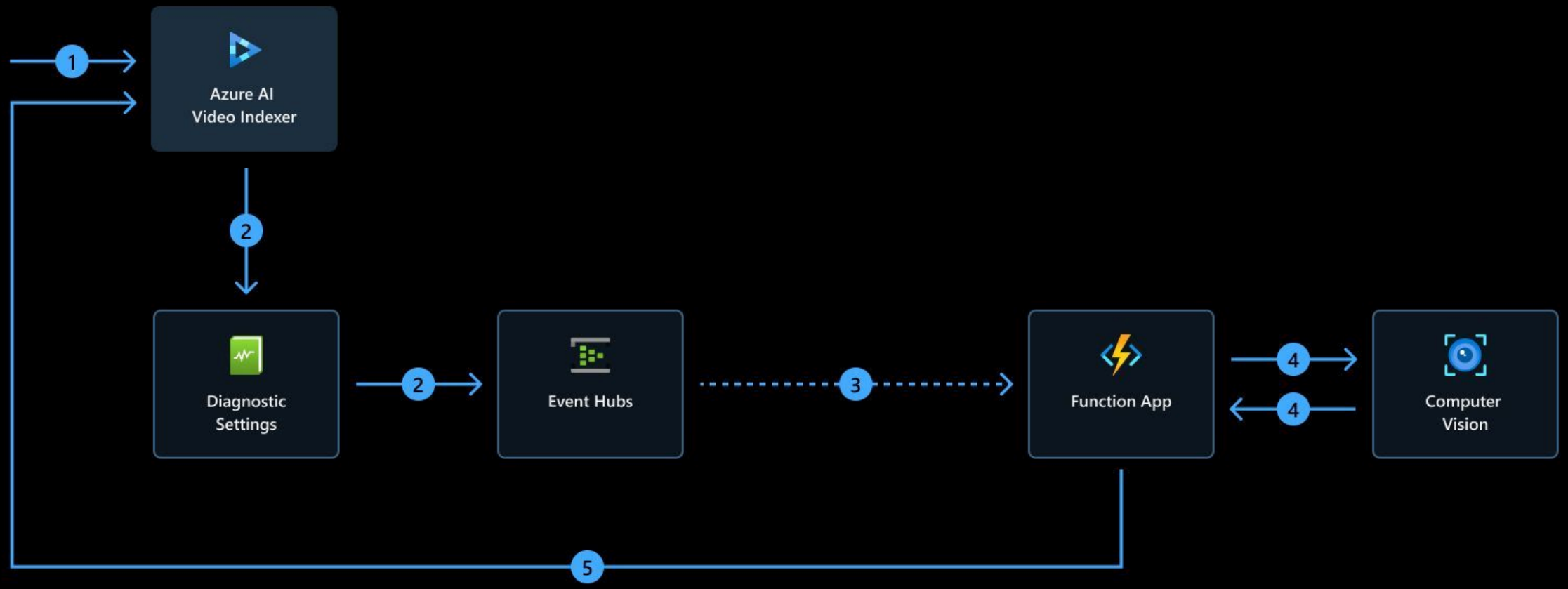




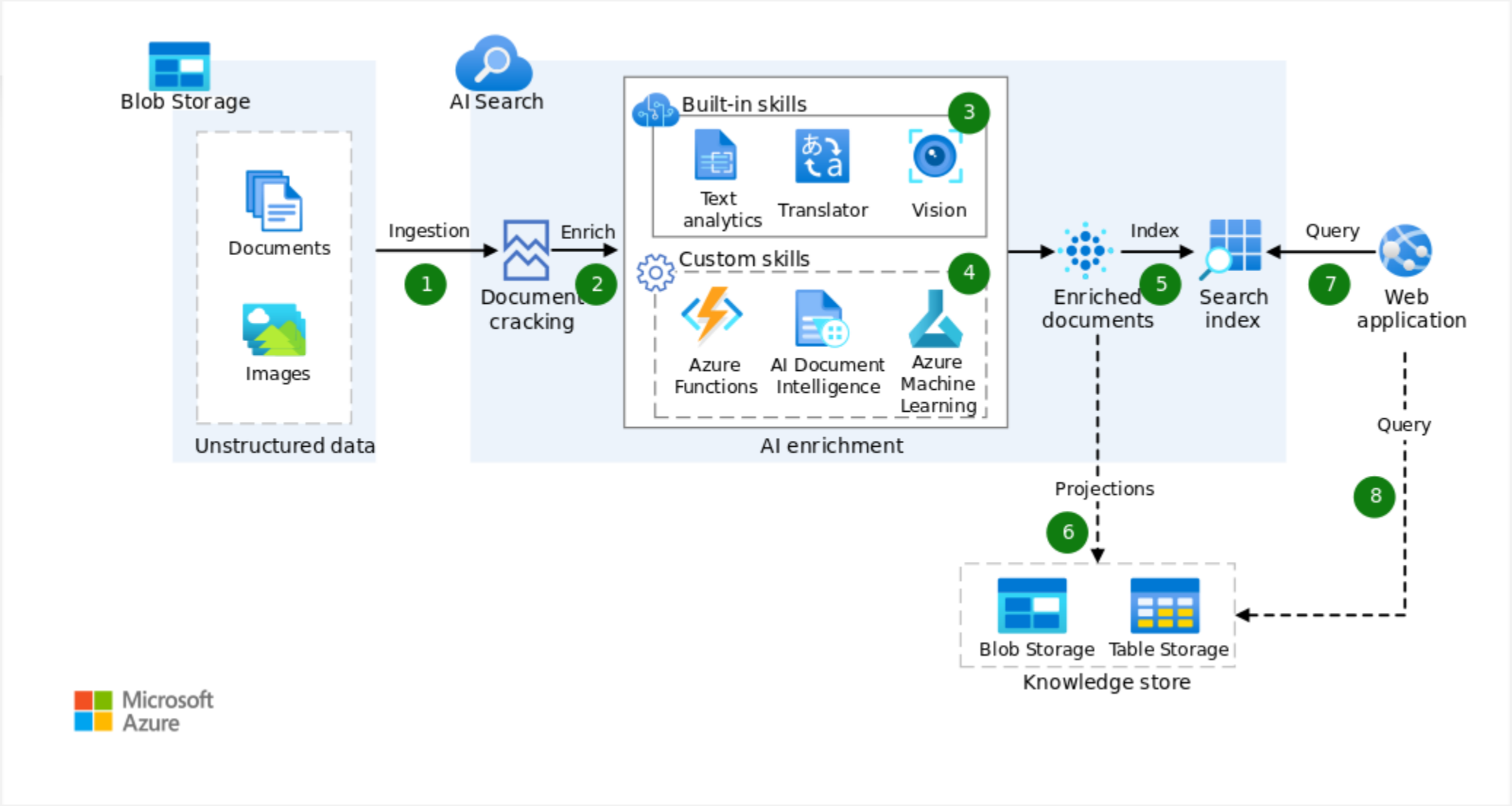
How do we go about implementing a data estate strategy?

Another business unit brought forth the requirement around videos – their human team needed a way to quickly parse the videos so they could only focus on the most immediate call to action instead of watching the entire video stream

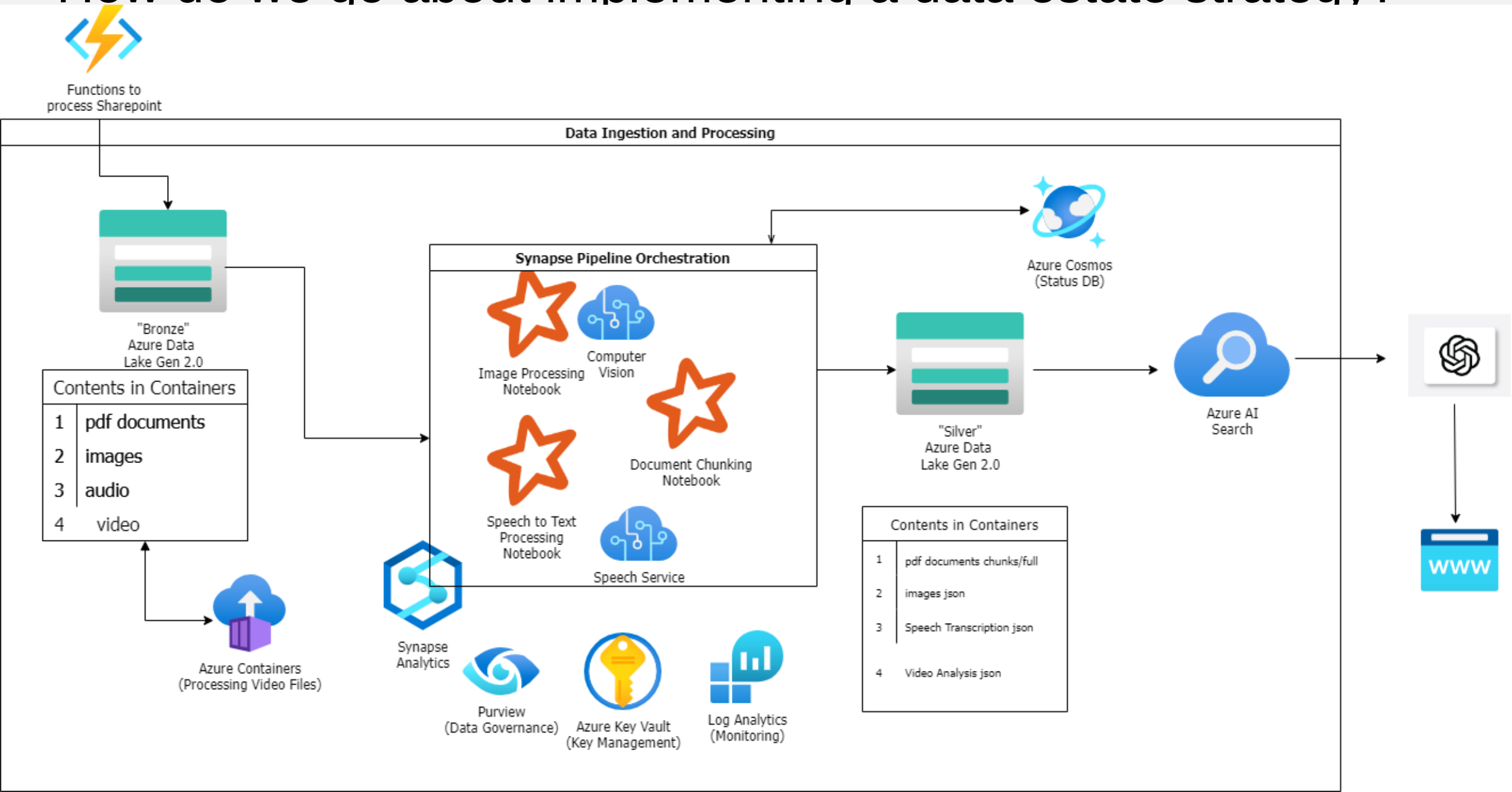
How do we go about implementing a data estate strategy?



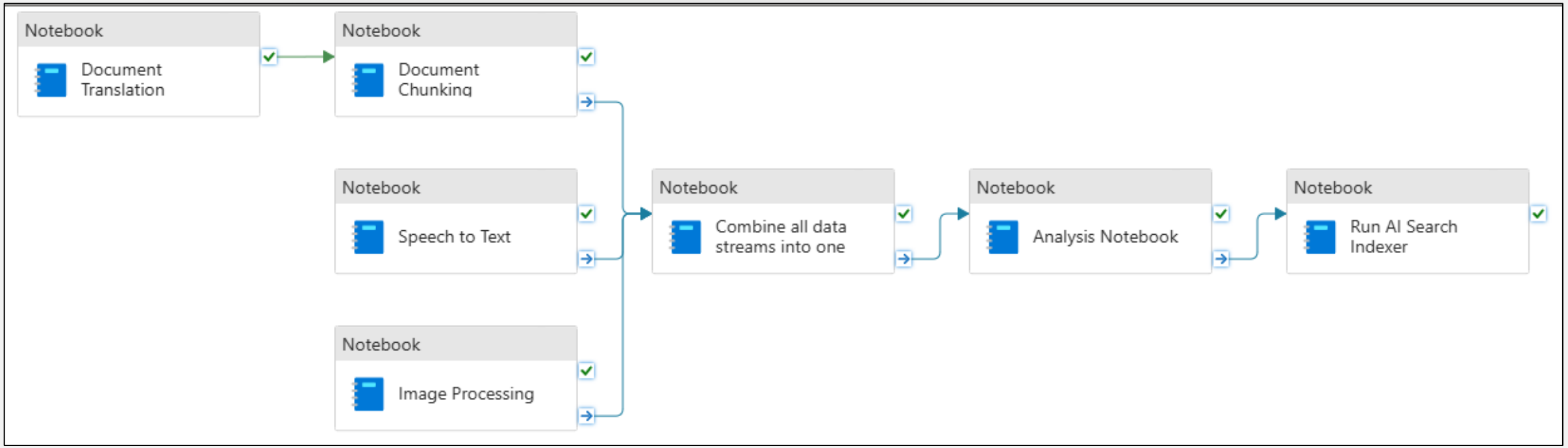
How do we go about implementing a data estate strategy?



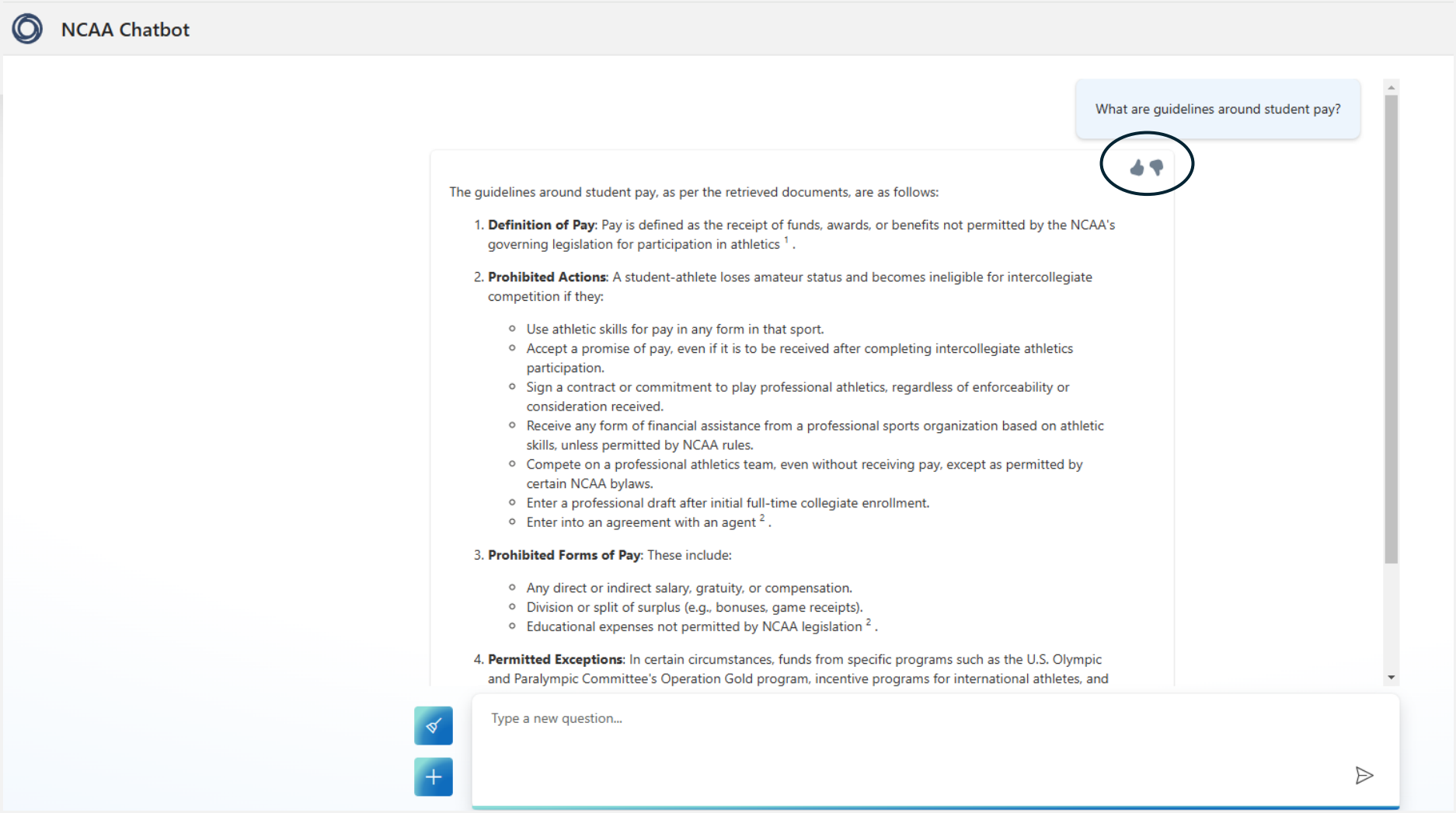
How do we go about implementing a data estate strategy?



How do we go about implementing a data estate strategy?



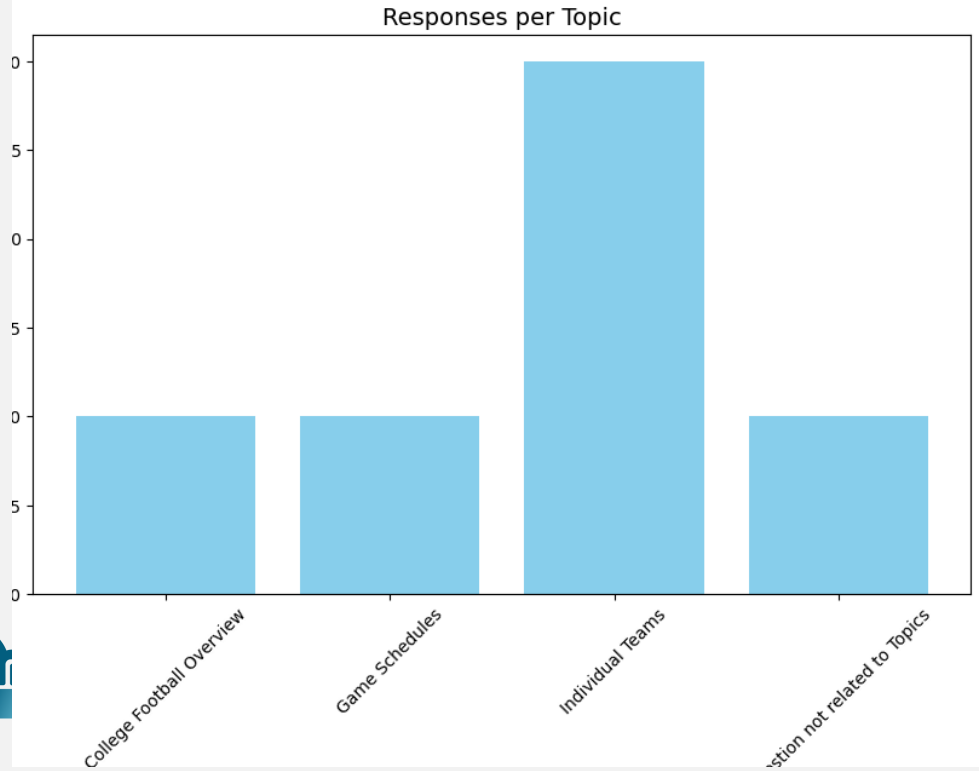
How do we go about implementing a data estate strategy?

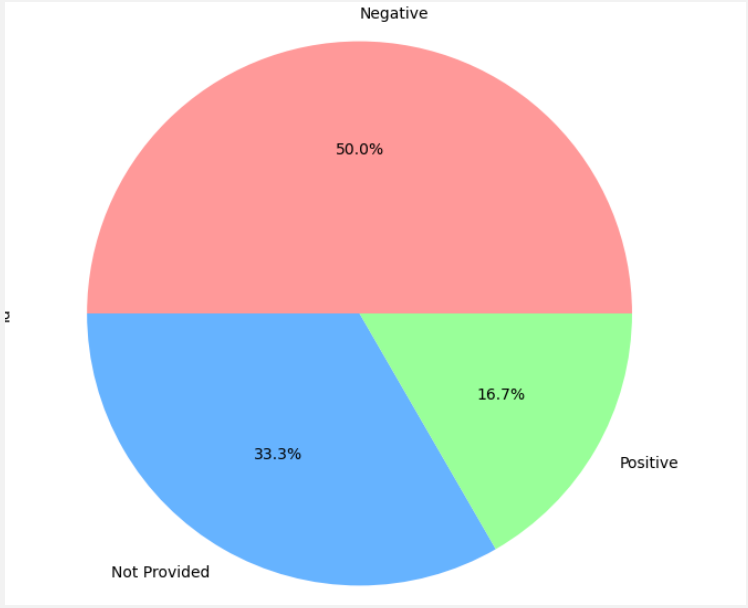


How do we go about implementing a data estate strategy?

View | Table | Chart | Export results

updatedAt	role	feedback	id	conversationId	createdAt	content
2024-10-03T00:30:10.8159	user		a104dba1-fc97-460a-b07d-8b28...	dce626d7-2f34-473c-860e-c789c...	2024-10-03T00:30:10.815886	What can you te
2024-10-03T00:30:17.183498	tool		207d585f-5a3c-4a13-a82e-b292...	dce626d7-2f34-473c-860e-c789c...	2024-10-03T00:30:17.183491	{"citations": [{"cc
2024-10-03T00:30:17.238214	assistant	positive	0aaf2323-effe-4d14-8749-42ac9...	dce626d7-2f34-473c-860e-c789c...	2024-10-03T00:30:17.238206	The retrieved do
2024-10-03T00:30:29.499908	user		7e1c7854-ac0b-4844-be6e-d987...	dce626d7-2f34-473c-860e-c789c...	2024-10-03T00:30:29.499901	What other colle
2024-10-03T00:30:32.13003	tool		dae46c2f-5c62-4e90-b255-5e80...	dce626d7-2f34-473c-860e-c789c...	2024-10-03T00:30:32.130019	{"citations": [{"cc
2024-10-03T00:30:44.192871	user		5201a9e6-be25-492e-8d0e-b43d...	dce626d7-2f34-473c-860e-c789c...	2024-10-03T00:30:44.192863	What about Cler
2024-10-03T00:30:50.950664	tool		d6dca7c6-cf7d-49ab-8798-3f955...	dce626d7-2f34-473c-860e-c789c...	2024-10-03T00:30:50.950654	{"citations": [{"cc





THANKS

Feedback

Do you want us to follow up after the event? Do you have feedback?

https://aka.ms/summit/feedback

CONTACT

Ujval Gandhi |

ujvalgandhi@microsoft.com |

Microsoft Federal Developer Summit
Building AI Solutions