

Building an Azure Data Analytics Platform

End-to-End



Paul Andrew | Technical Architect in Azure CoE



avanade



mrpaulandrew.tech



@MrPaulAndrew



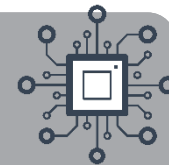
In/MrPaulAndrew



MrPaulAndrew.com



c/MrPaulAndrew



<https://github.com/mrpaulandrew>

CommunityEvents

Demo code, content and slides from various community events.

● C++

[{Event/Location}-{Month}-{Year}](#)

Agenda



1. Design
2. Extract
3. Transform
4. Load

Agenda



1. Design
2. Extract
3. Transform
4. Load

Question:

What is the answer to life, the universe and everything?

Answer:
42



Answer:
It depends!



Question:

What is big data?

Answer:

It depends!



Volume
Velocity
Variety
Veracity
Value

Answer:

Any data that you cannot process
in the time that you have/want
using the technology you have.

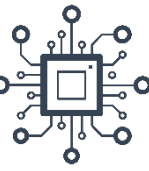


- Buck Woody

@BuckWoodyMSFT



Goal



Data
Sources

Paul's Magic Box -
From the Hogwarts School of Witches & Wizardry

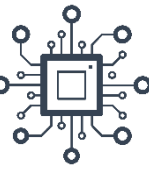
Data
Warehouse



Data
Insights

Data = Information = Knowledge = Power

Goal



Clean
Enrich
Conform
Translate
Transform
Curate
Analyse
Model
Predict
Master



Data
Sources

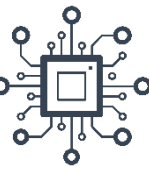


Data
Warehouse

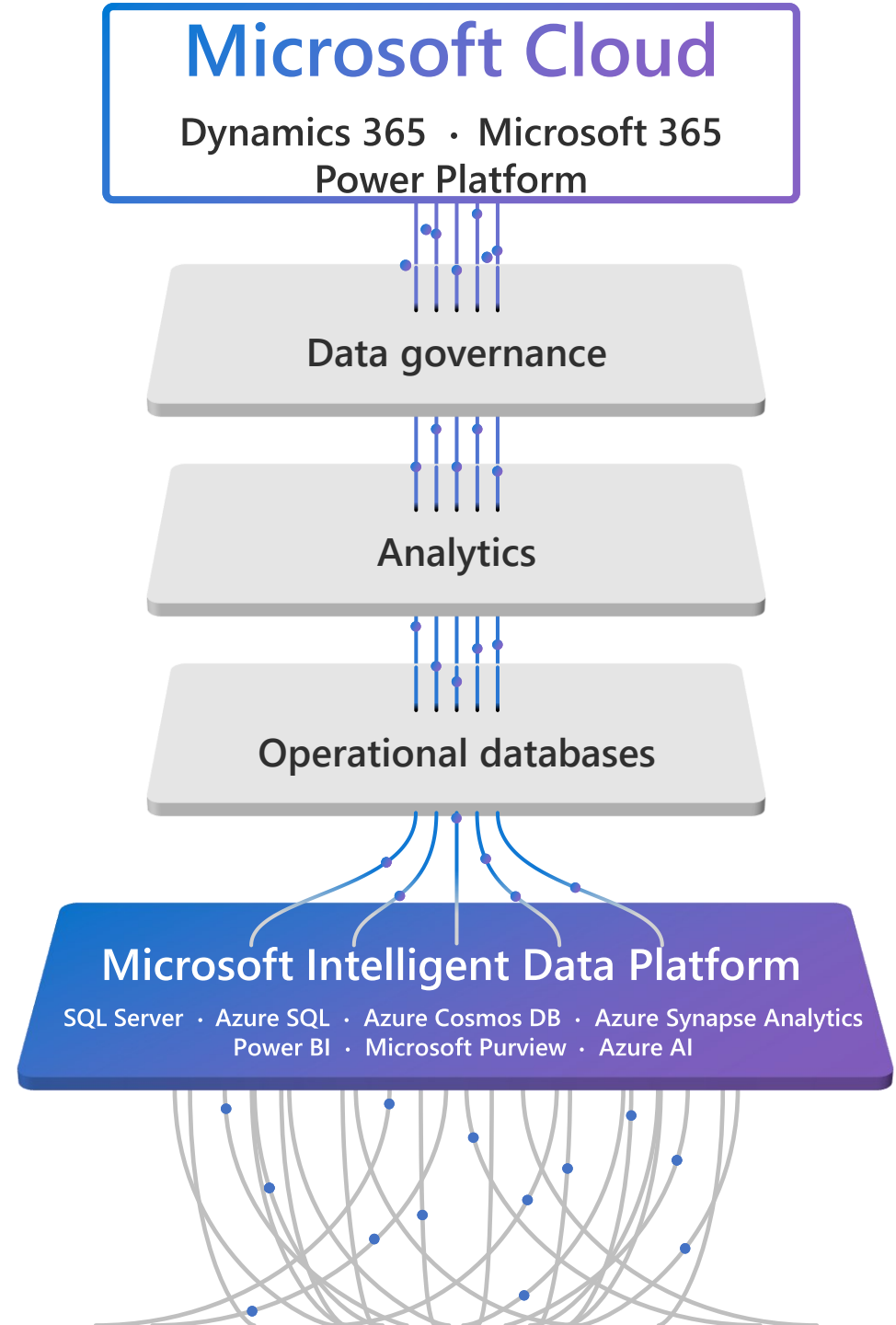


Data
Insights

Paul's Reference Architecture

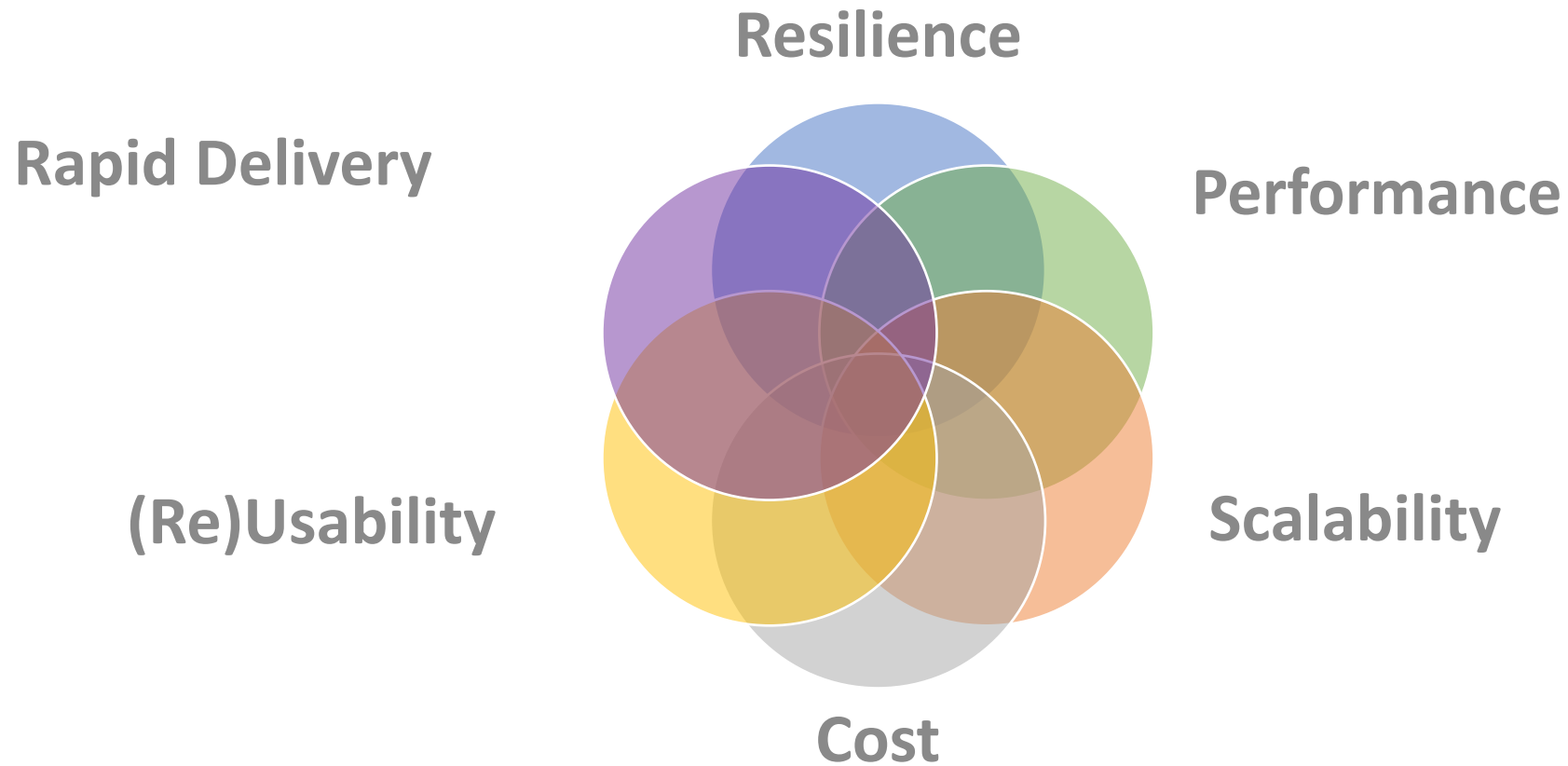


Microsoft's Intelligent Data Platform



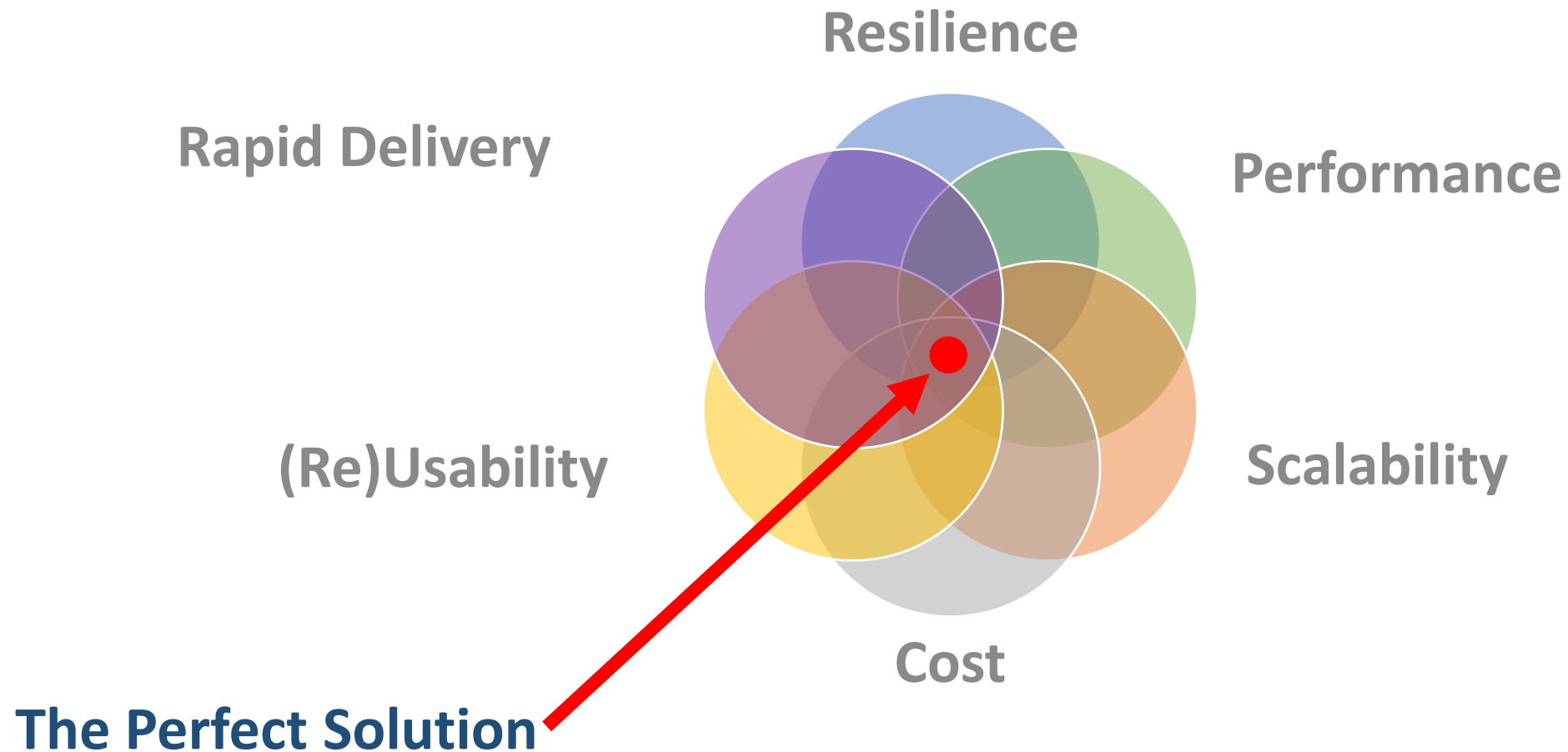


What is your primary design focus?



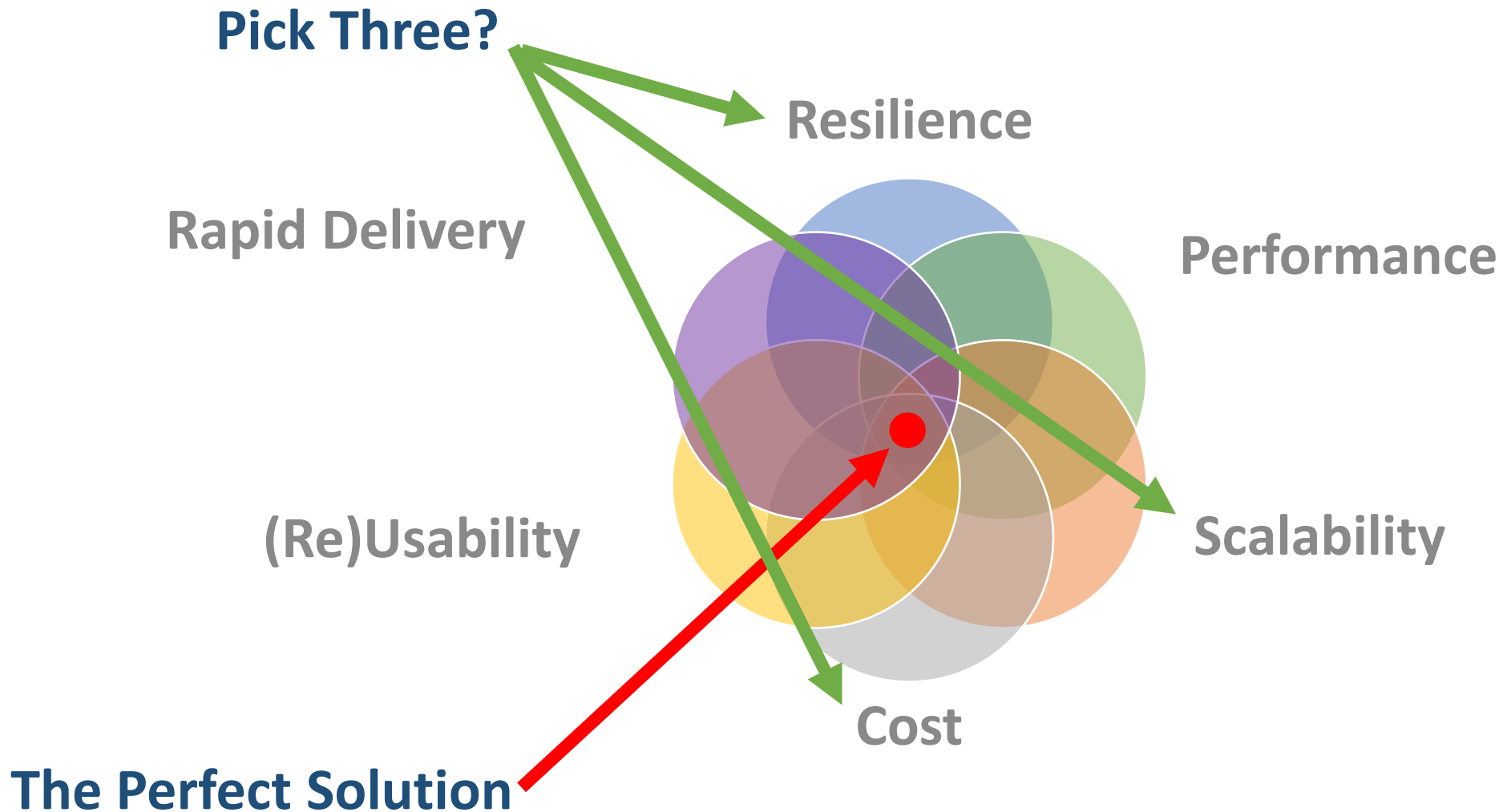


What is your primary design focus?







What is your primary design focus?



Agenda



1.

Design

✓

2.

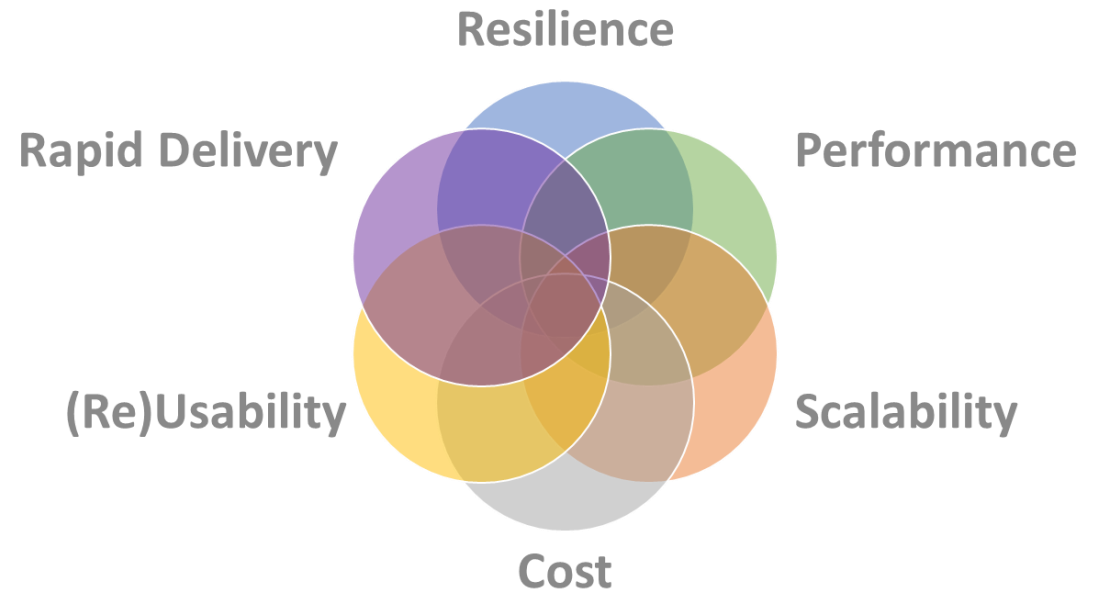
Extract

3.



Transform

4.

Load



Agenda



1.

Design

✓

2.

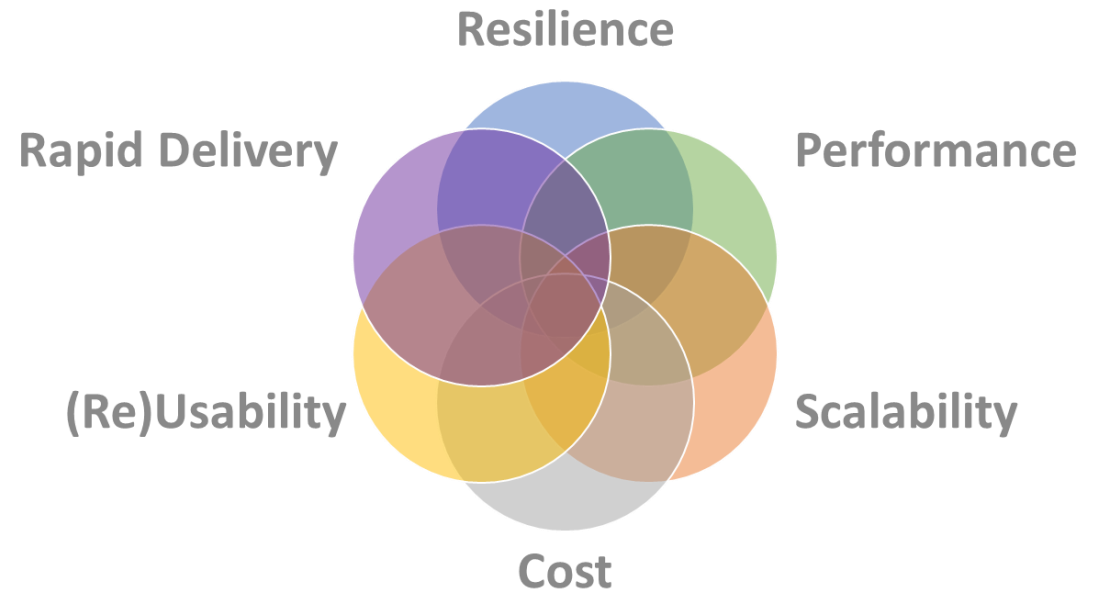
Extract

3.

Transform

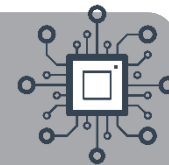
4.

Load





Data Extraction & Ingestion



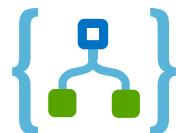
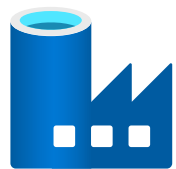
Data Structure



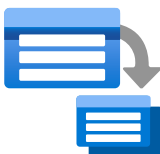
Data Source



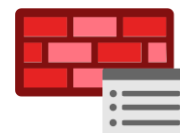
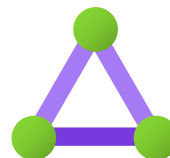
Push or Pull



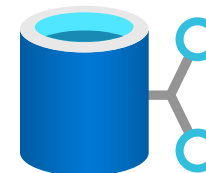
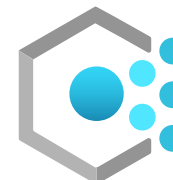
Batch or Speed



Public or Private Transfer



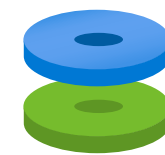
Data Sensitivity



Data Volume



!= Big



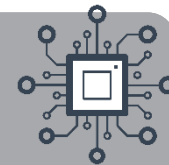
== Big



=> Big



Data Extraction & Ingestion – Spec v1



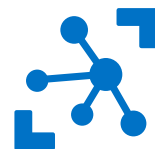
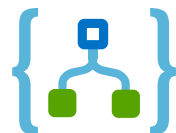
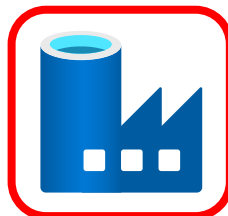
Data Structure



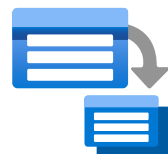
Data Source



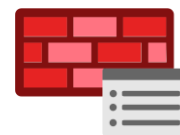
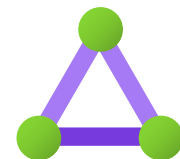
Push or Pull



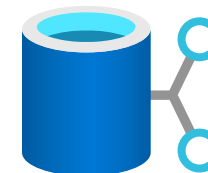
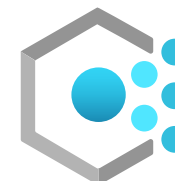
Batch or Speed



Public or Private Transfer



Data Sensitivity

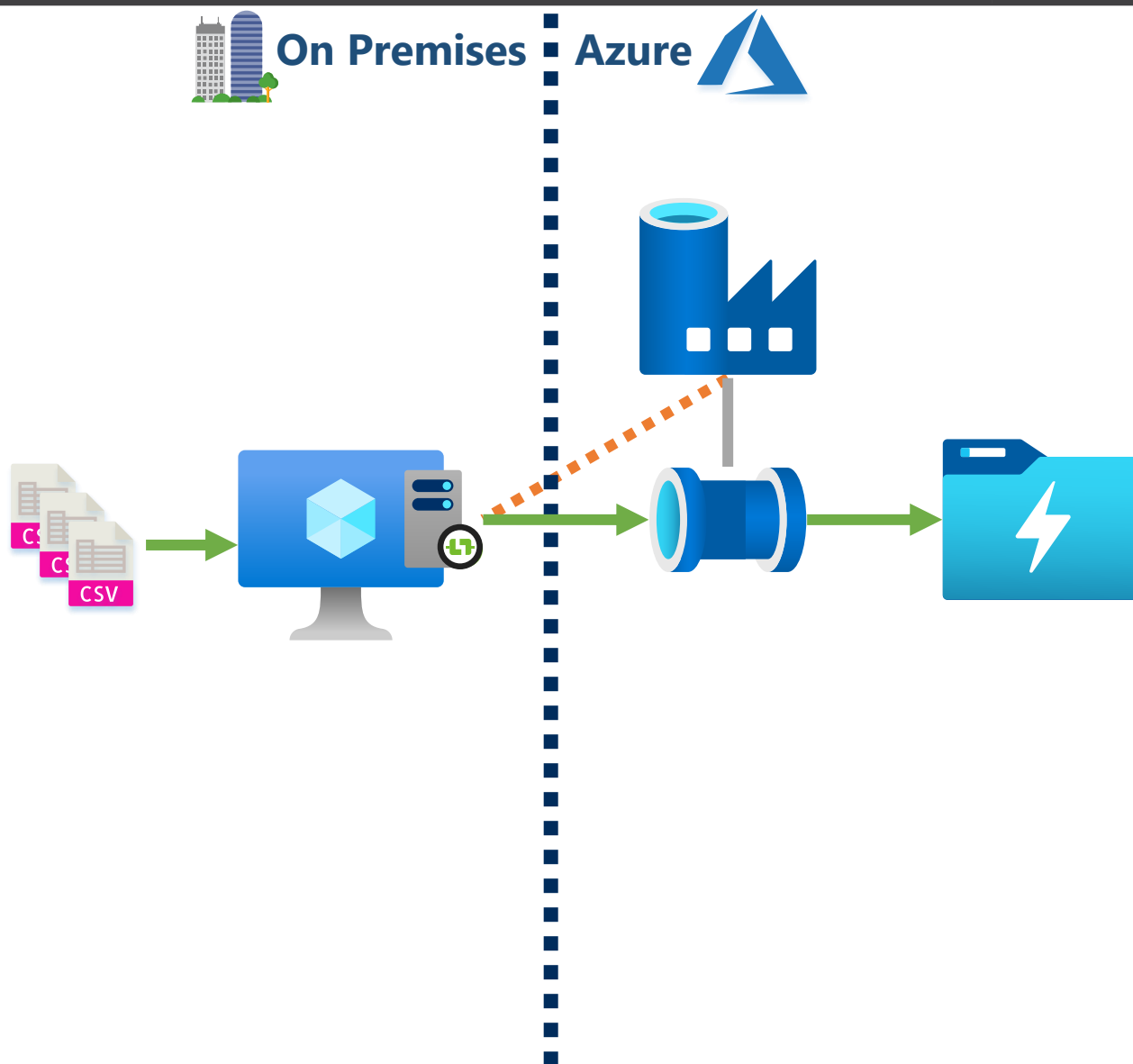
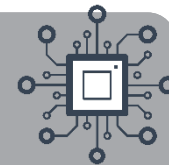


Data Volume





Data Extraction & Ingestion – Solution 1

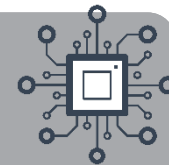


Requirements:

- Flat files
- From local storage
- Pulled from source
- Batch load
- Public connections
- No PII data
- Small data volumes



Data Extraction & Ingestion – Spec v2



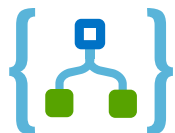
Data Structure



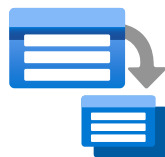
Data Source



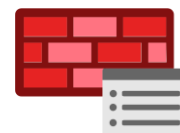
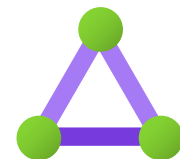
Push or Pull



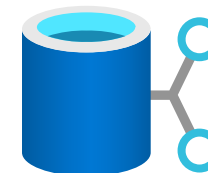
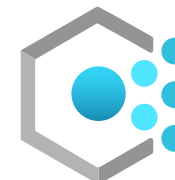
Batch or Speed



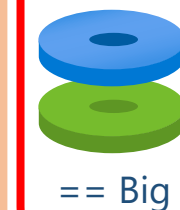
Public or Private Transfer



Data Sensitivity

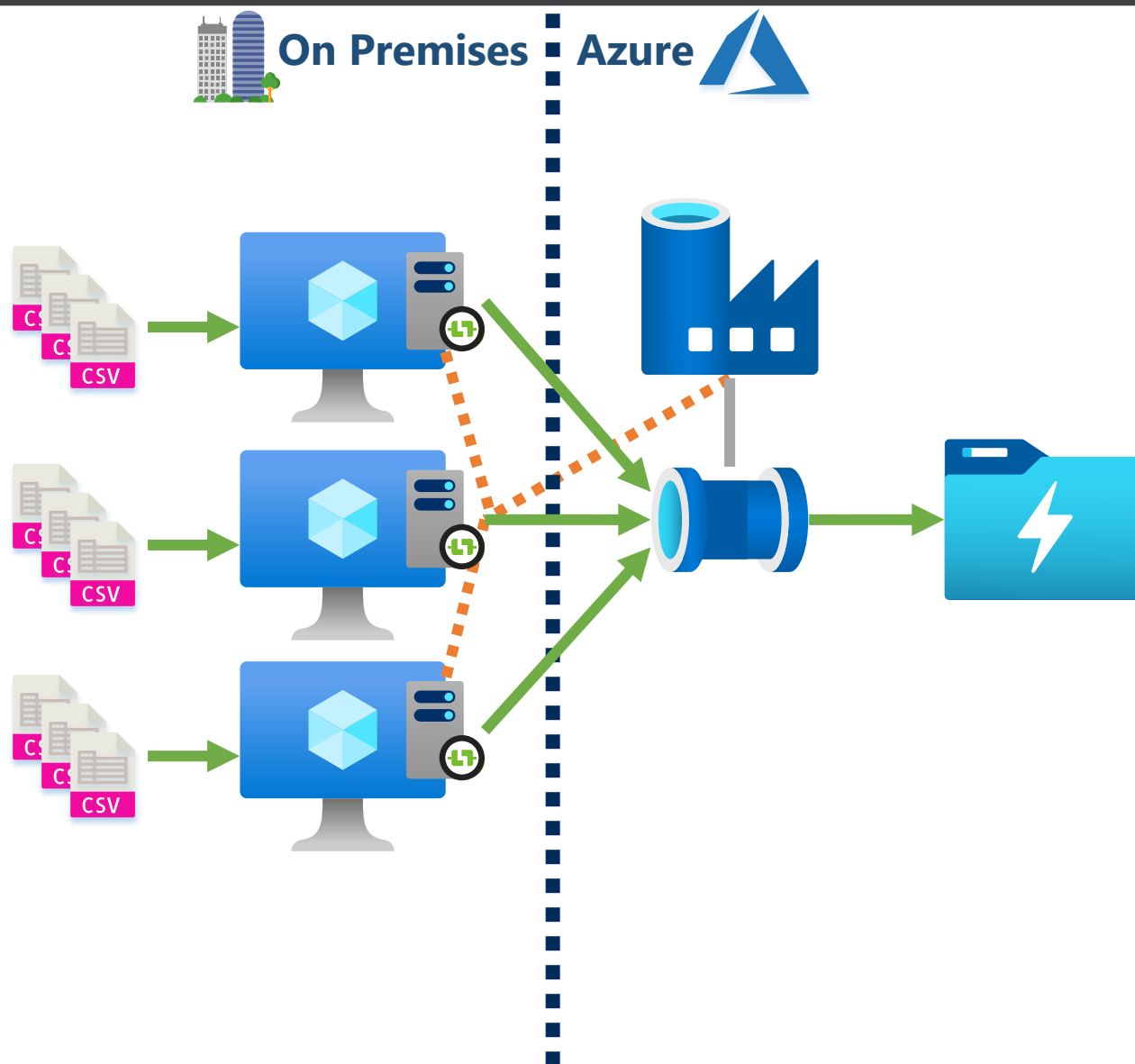
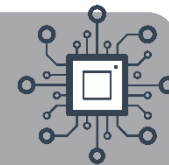


Data Volume





Data Extraction & Ingestion – Solution 2

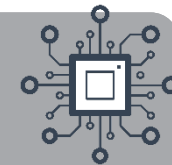


Requirements:

- Flat files
- From local storage
- Pulled from source
- Batch load
- Public connections
- No PII data
- Large data volumes



Data Extraction & Ingestion – Spec v3



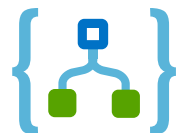
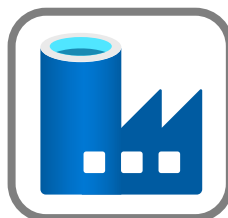
Data Structure



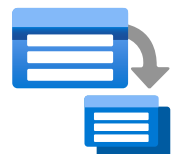
Data Source



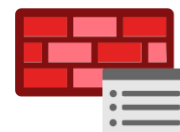
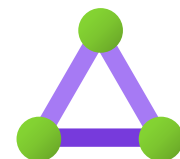
Push or Pull



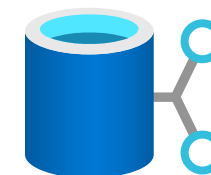
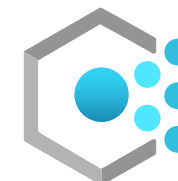
Batch or Speed



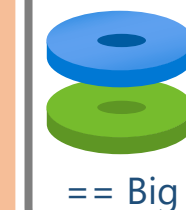
Public or Private Transfer



Data Sensitivity

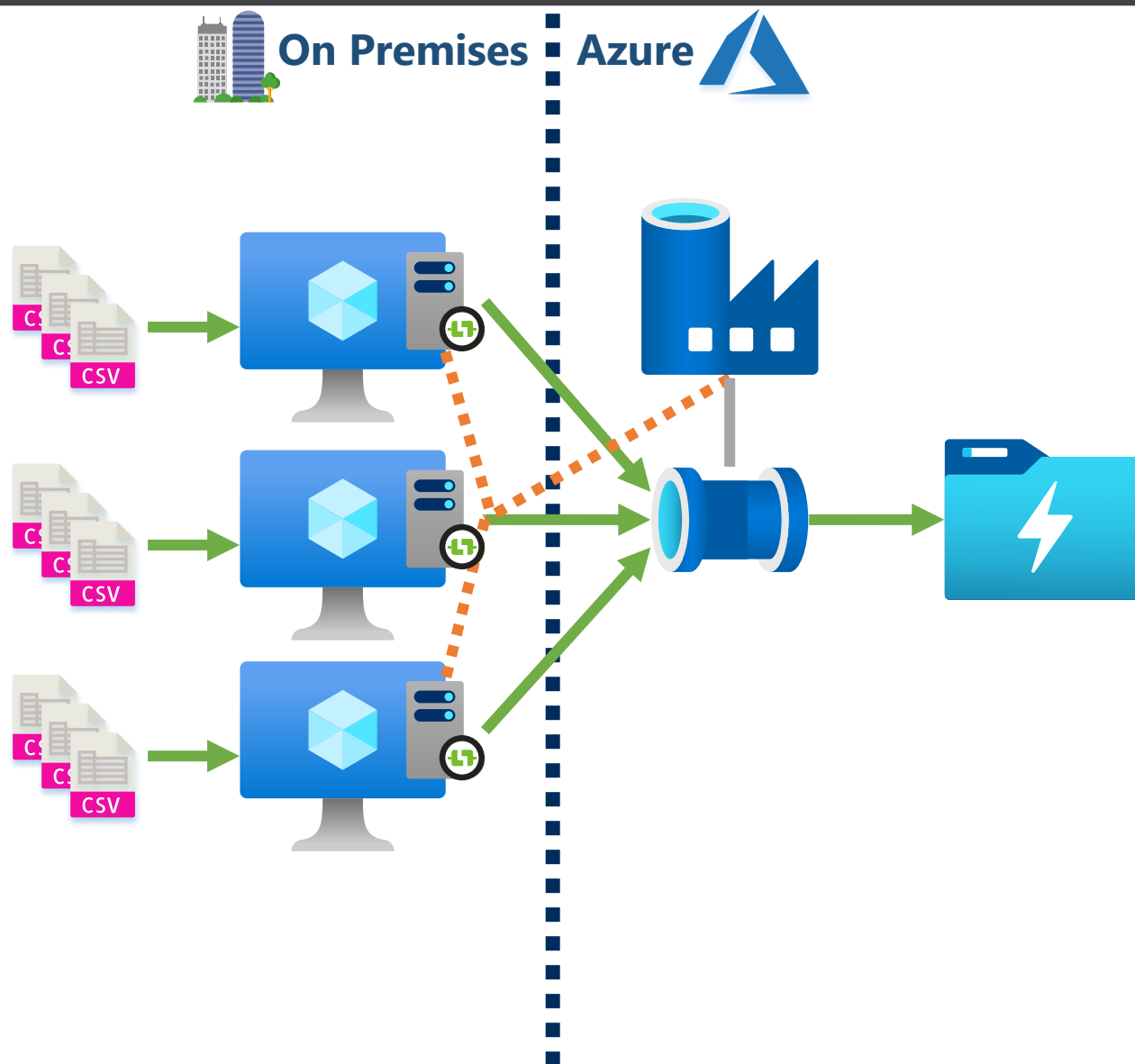
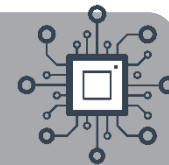


Data Volume





Data Extraction & Ingestion – Solution 3

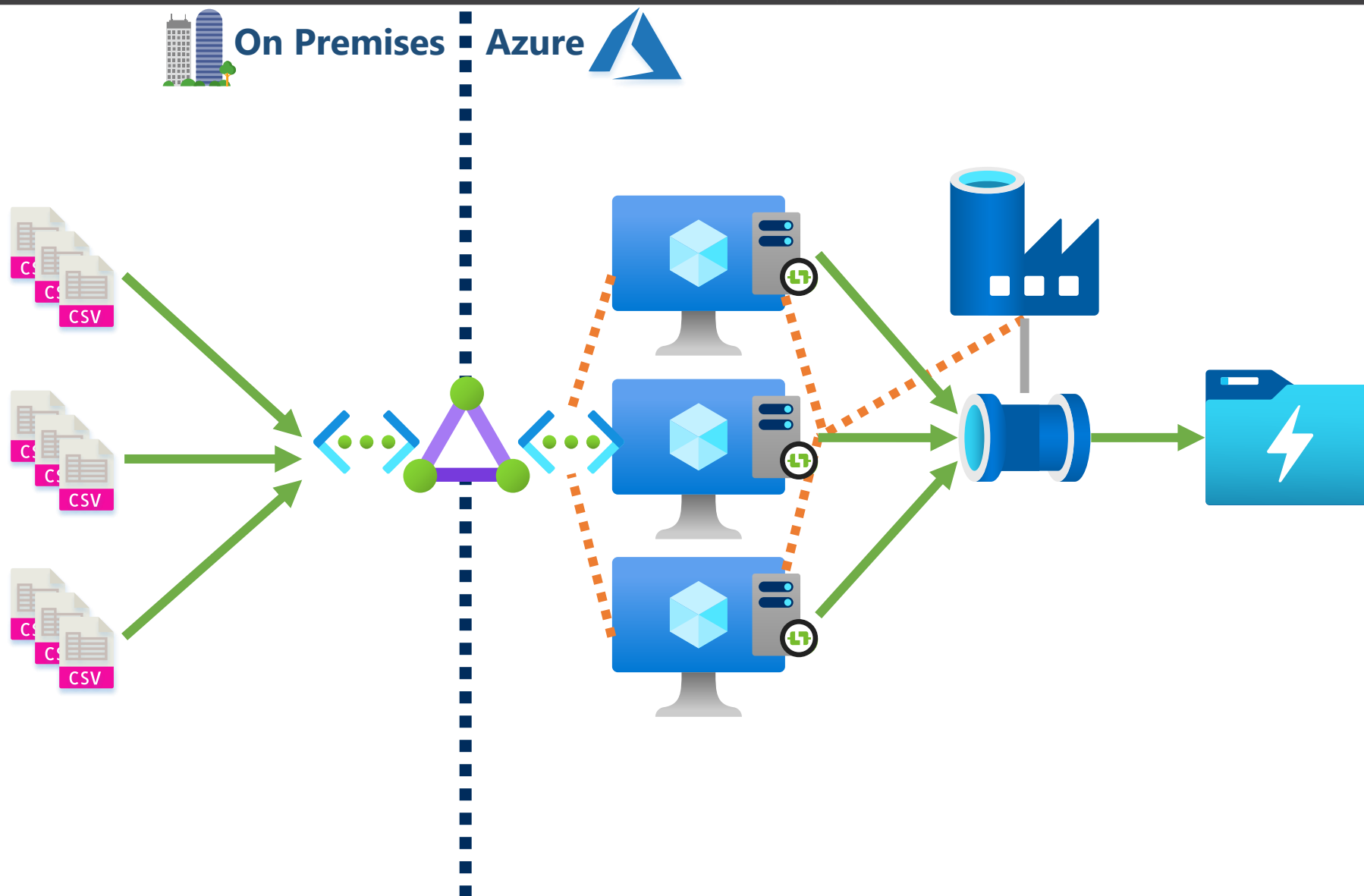
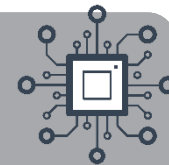


Requirements:

- Flat files
- From local storage
- Pulled from source
- Batch load
- Private connections
- No PII data
- Large data volumes



Data Extraction & Ingestion – Solution 3

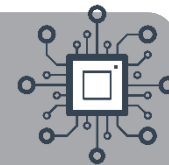


Requirements:

- Flat files
- From local storage
- Pulled from source
- Batch load
- Private connections
- No PII data
- Large data volumes



Data Extraction & Ingestion – Spec v4



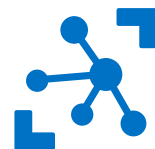
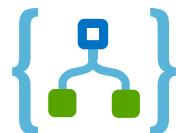
Data Structure



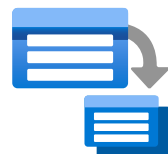
Data Source



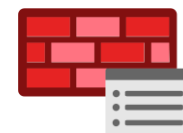
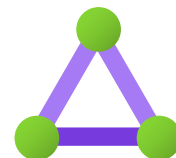
Push or Pull



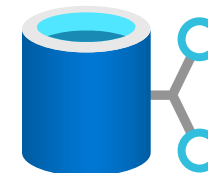
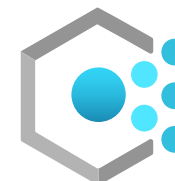
Batch or Speed



Public or Private Transfer



Data Sensitivity

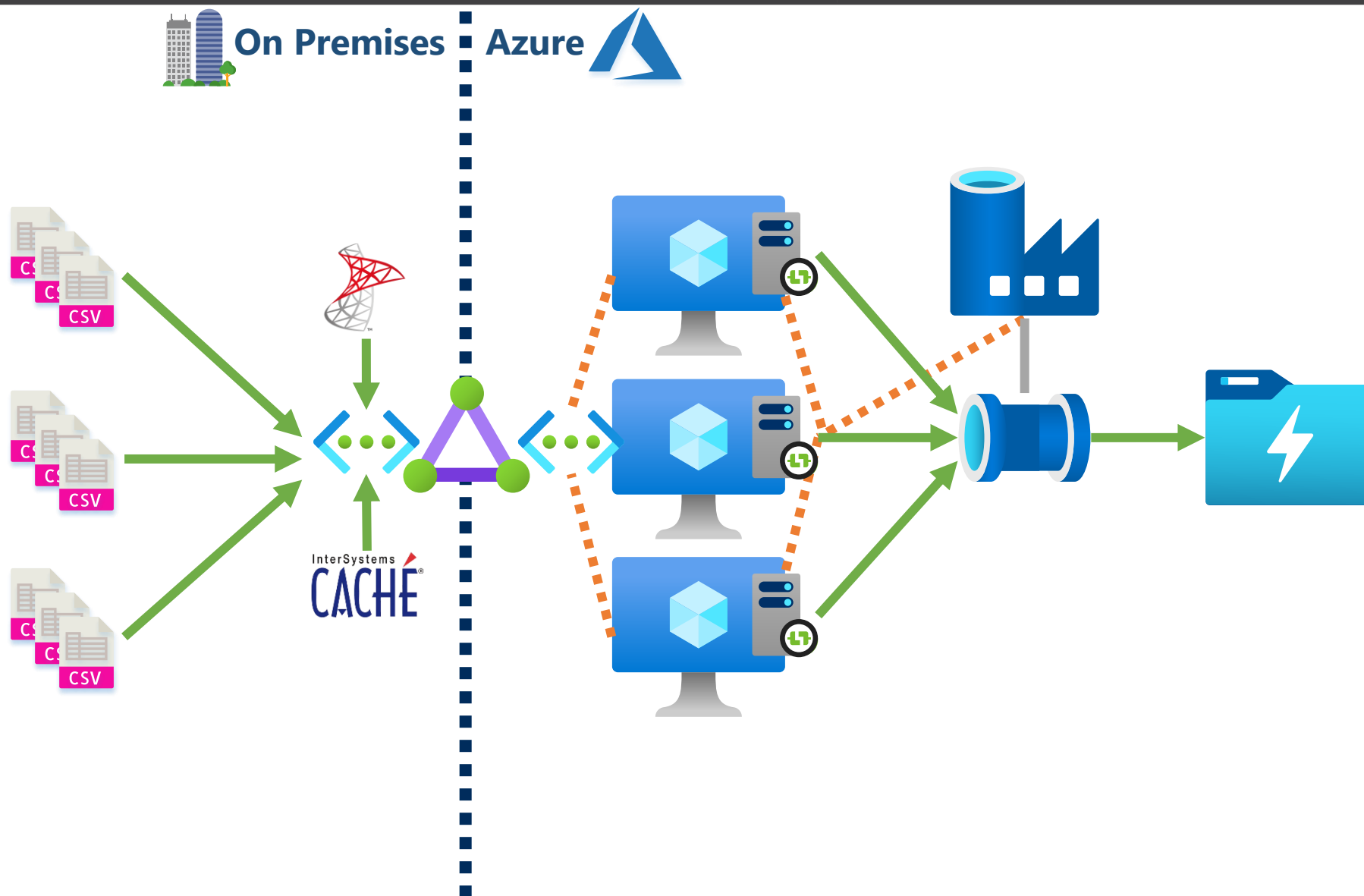
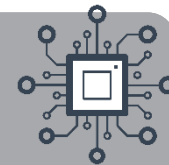


Data Volume





Data Extraction & Ingestion – Solution 4

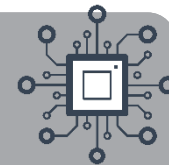


Requirements:

- Flat files
- From local storage & database tables
- Pulled from source
- Batch load
- Private connections
- No PII data
- Large data volumes



Data Extraction & Ingestion – Spec v5



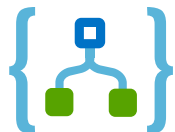
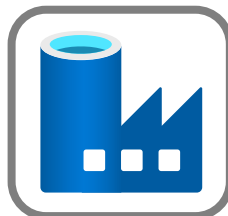
Data Structure



Data Source



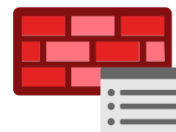
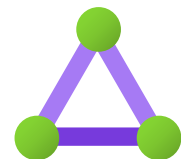
Push or Pull



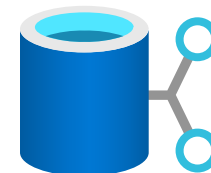
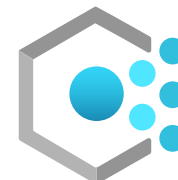
Batch or Speed



Public or Private Transfer



Data Sensitivity

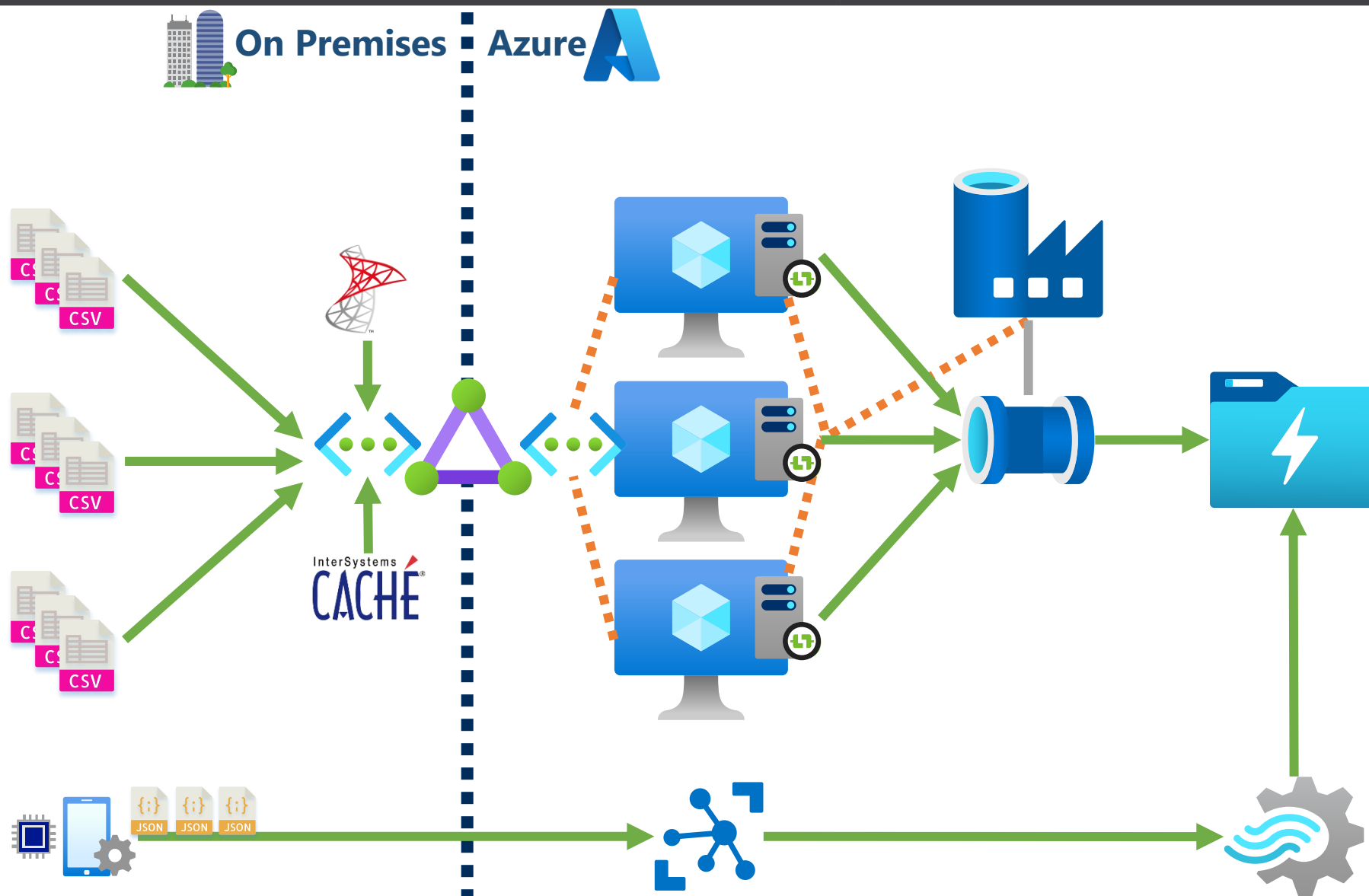
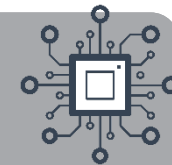


Data Volume





Data Extraction & Ingestion – Solution 5

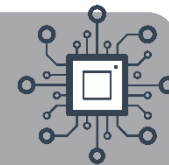


Requirements:

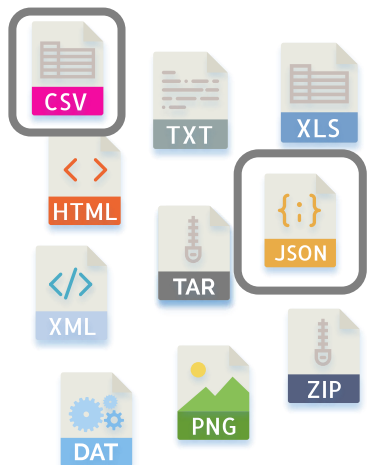
- Flat files & JSON
- From local storage & database tables
- Pulled from source & pushed
- Batch load & streamed
- Private connections
- No PII data
- Large data volumes



Data Extraction & Ingestion – Spec v6



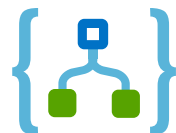
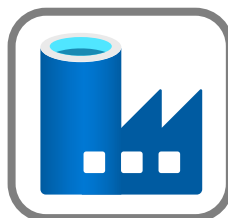
Data Structure



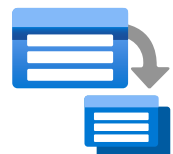
Data Source



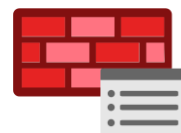
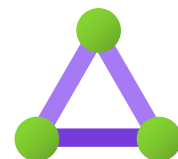
Push or Pull



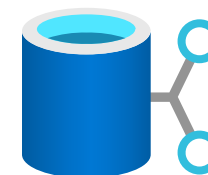
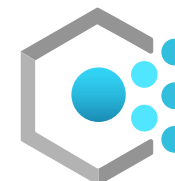
Batch or Speed



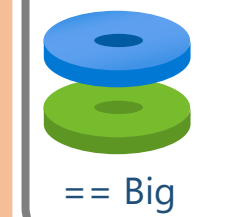
Public or Private Transfer



Data Sensitivity

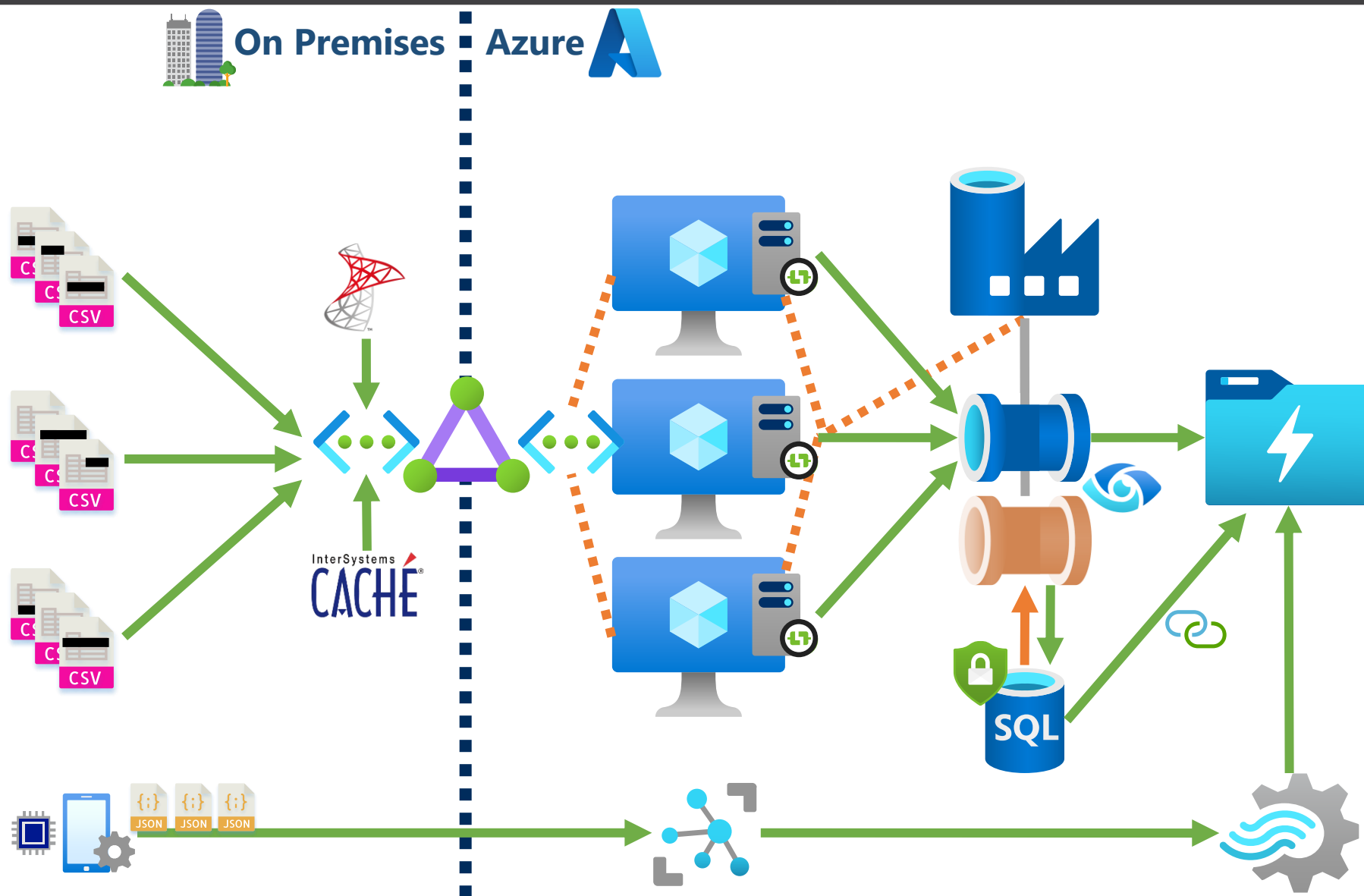
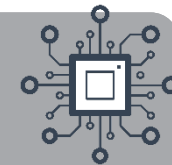


Data Volume





Data Extraction & Ingestion – Solution 6

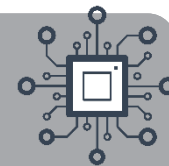


Requirements:

- Flat files & JSON
- From local storage & database tables
- Pulled from source & pushed
- Batch load & streamed
- Private connections
- Both PII & none PII data
- Large data volumes



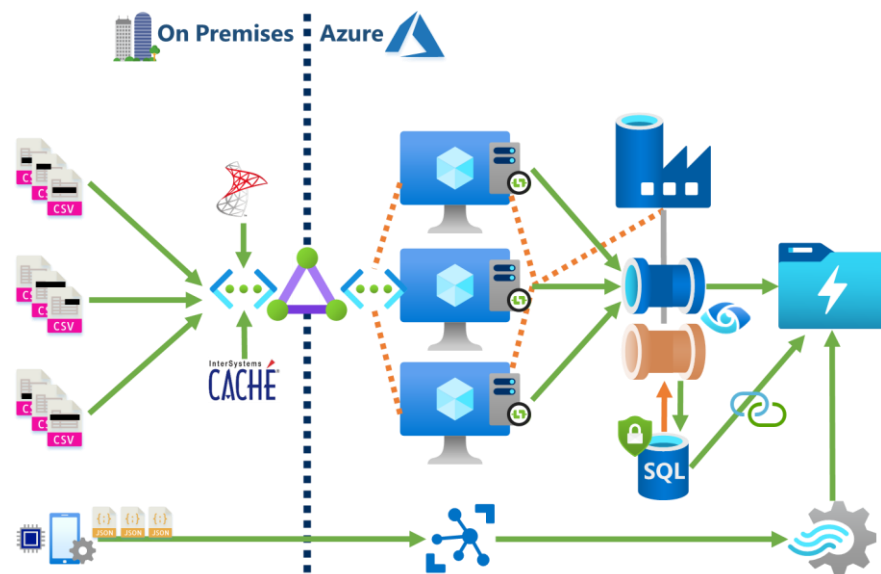
Overall Architecture



Extract

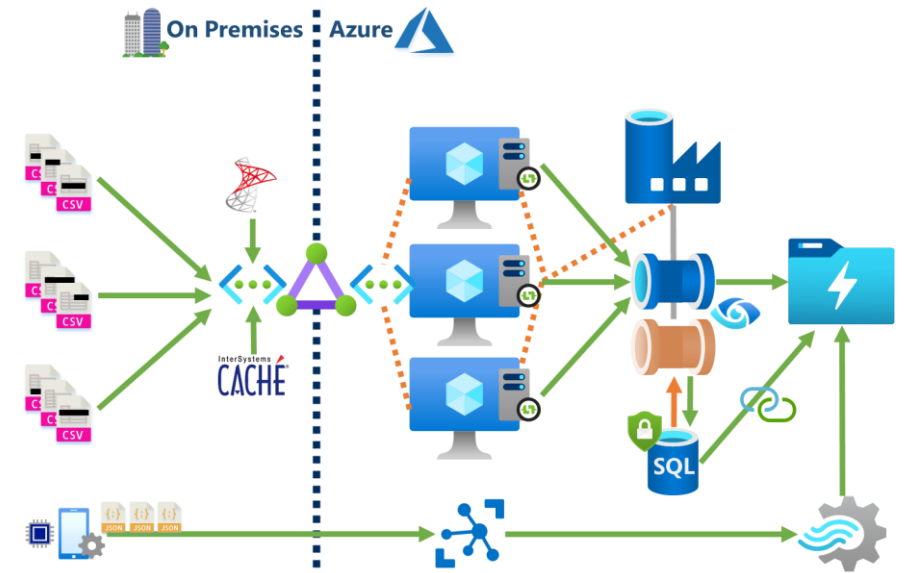
Transform

Load



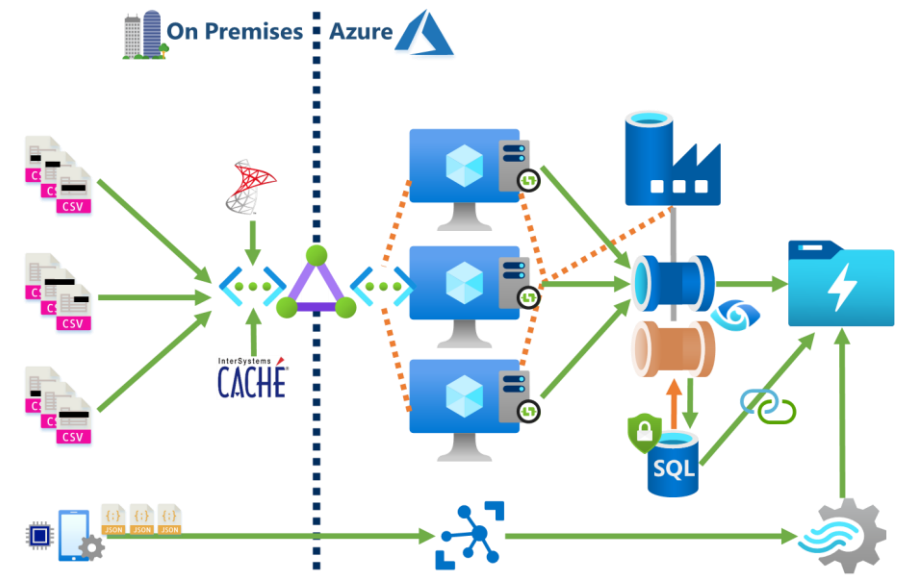
Agenda

1. Design ✓
2. Extract ✓
3. Transform
4. Load




Agenda

1. Design ✓
2. Extract ✓
3. Transform
4. Load



Agenda

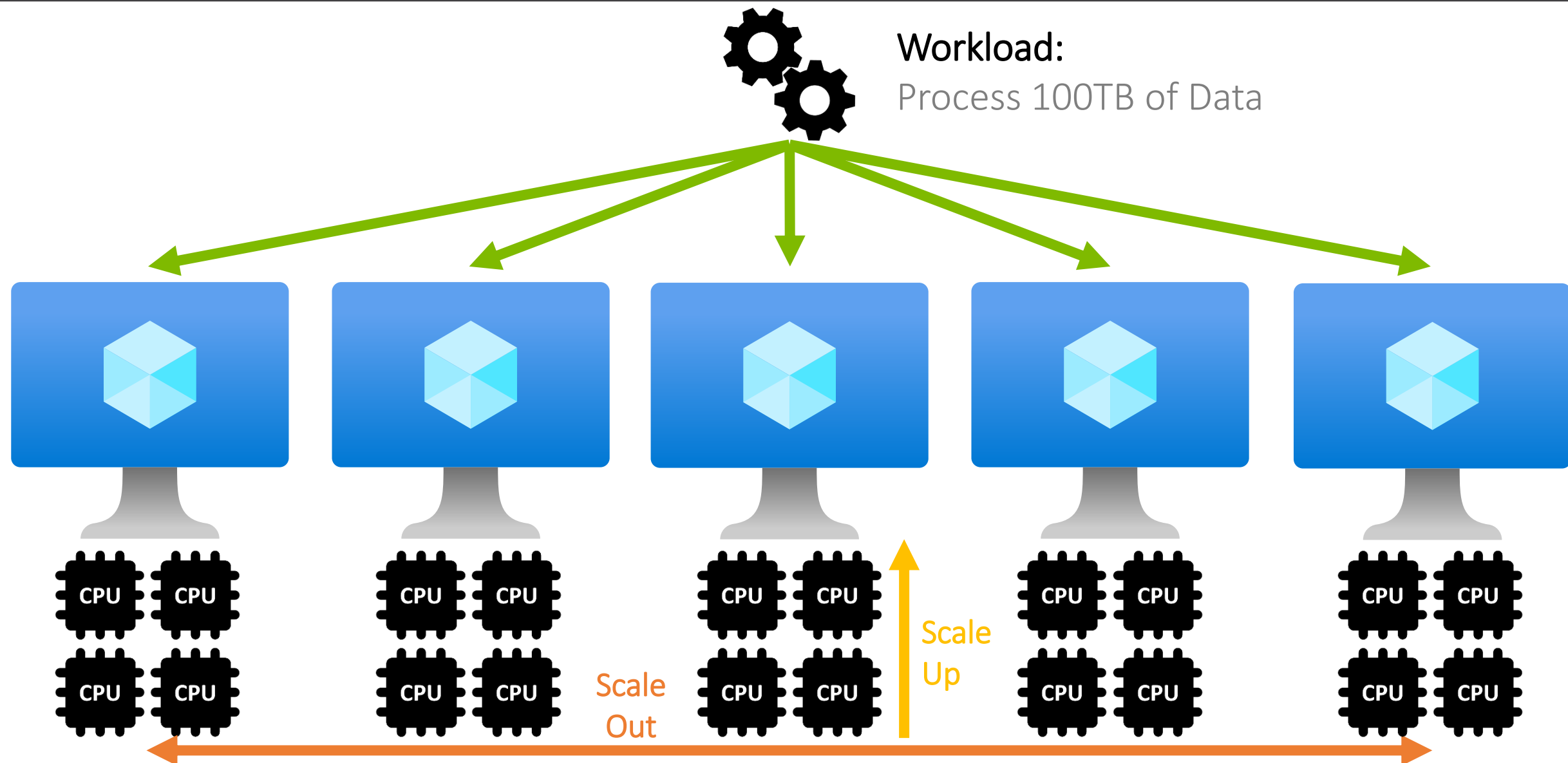
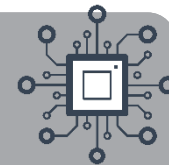


1. Design ✓
2. Extract ✓
3. Transform
4. Load

Compute
Storage, Structure
& Data Format

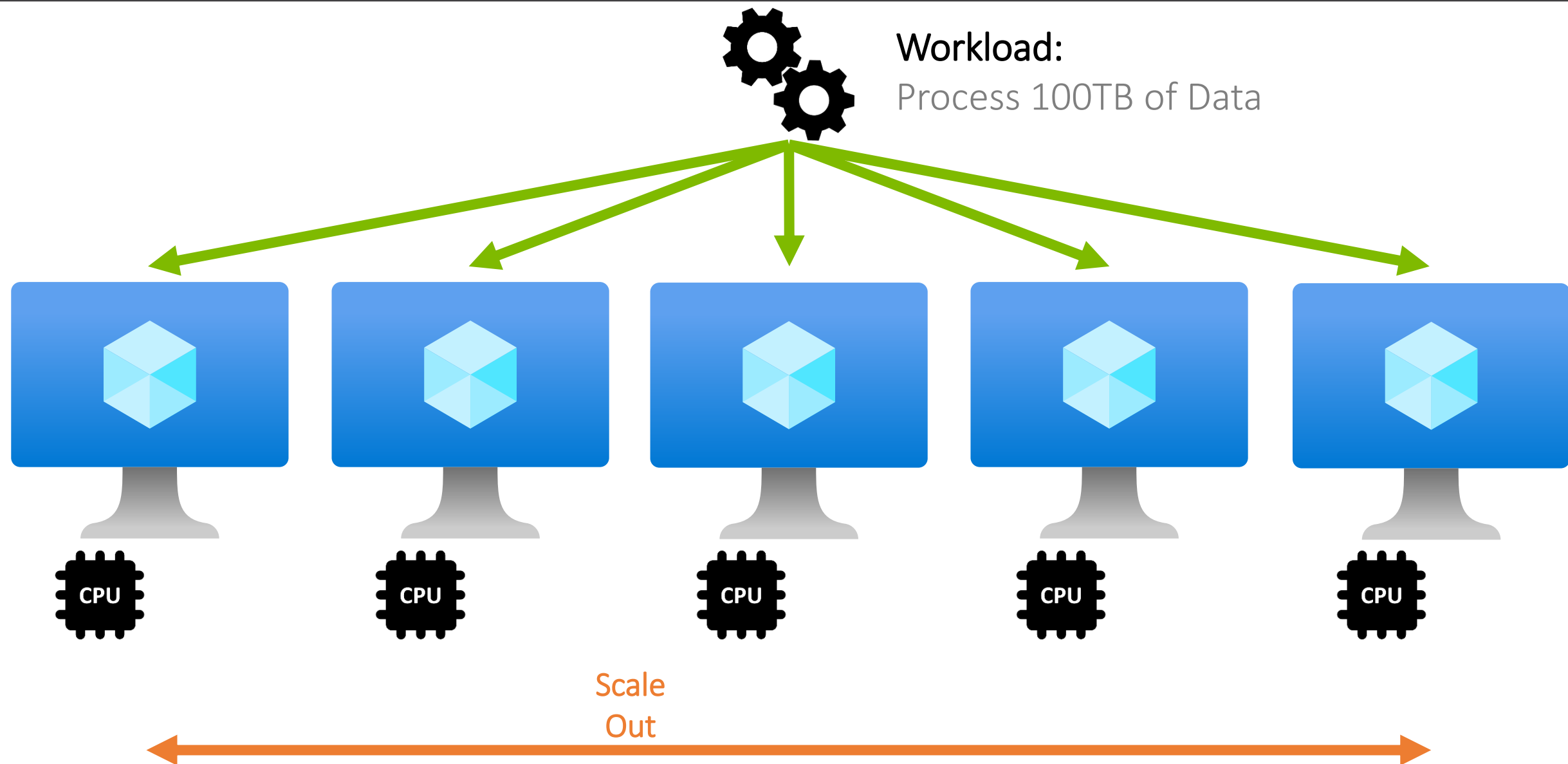
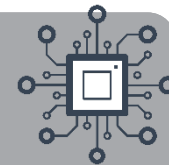


Scaling Up and/or Scaling Out



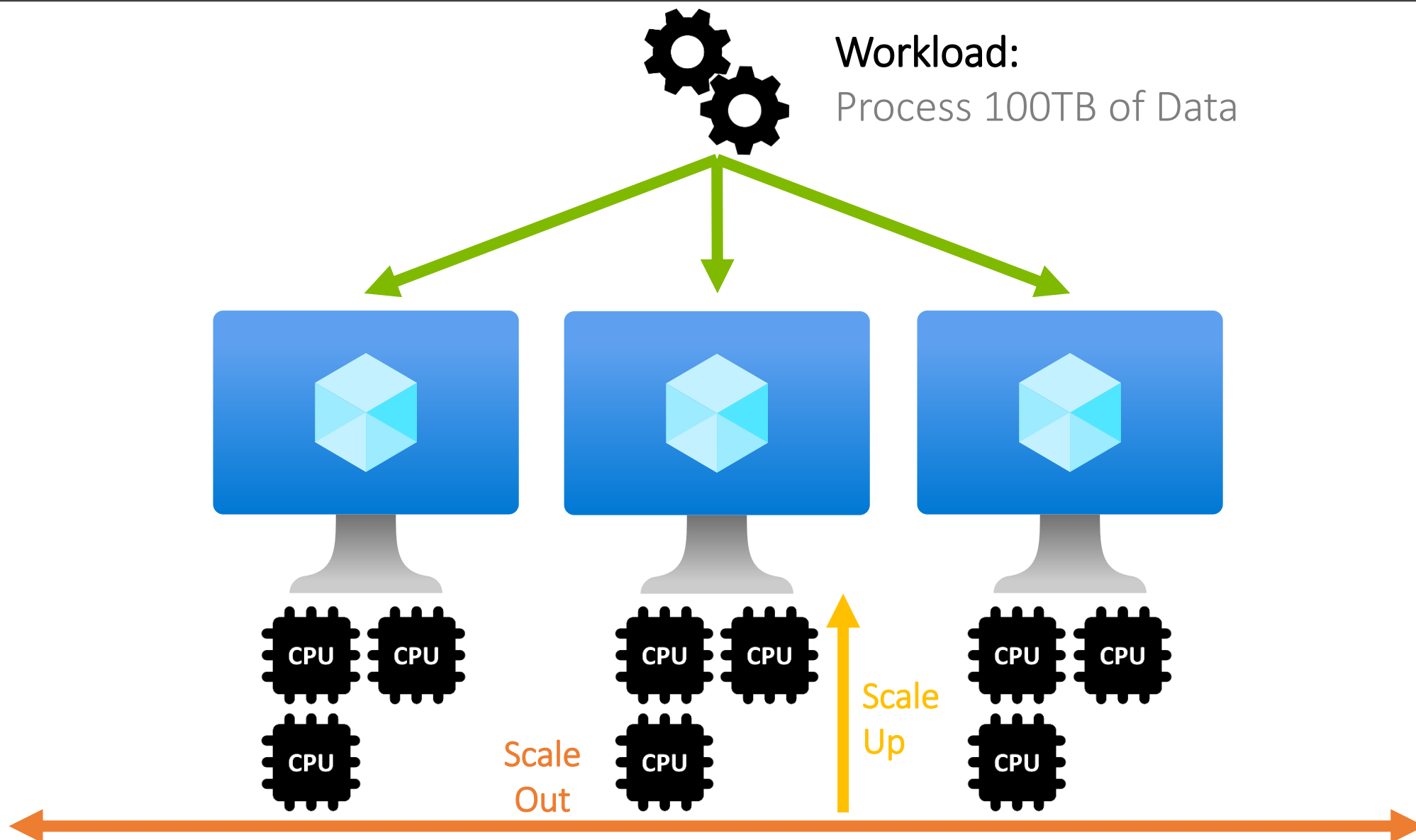
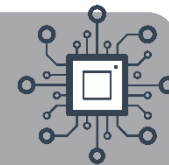


Scaling Up and/or Scaling Out



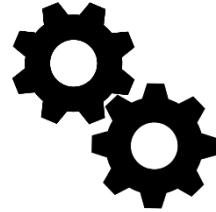


Scaling Up and/or Scaling Out





What Compute Type of Compute?



Workload:

Process 100TB of Data

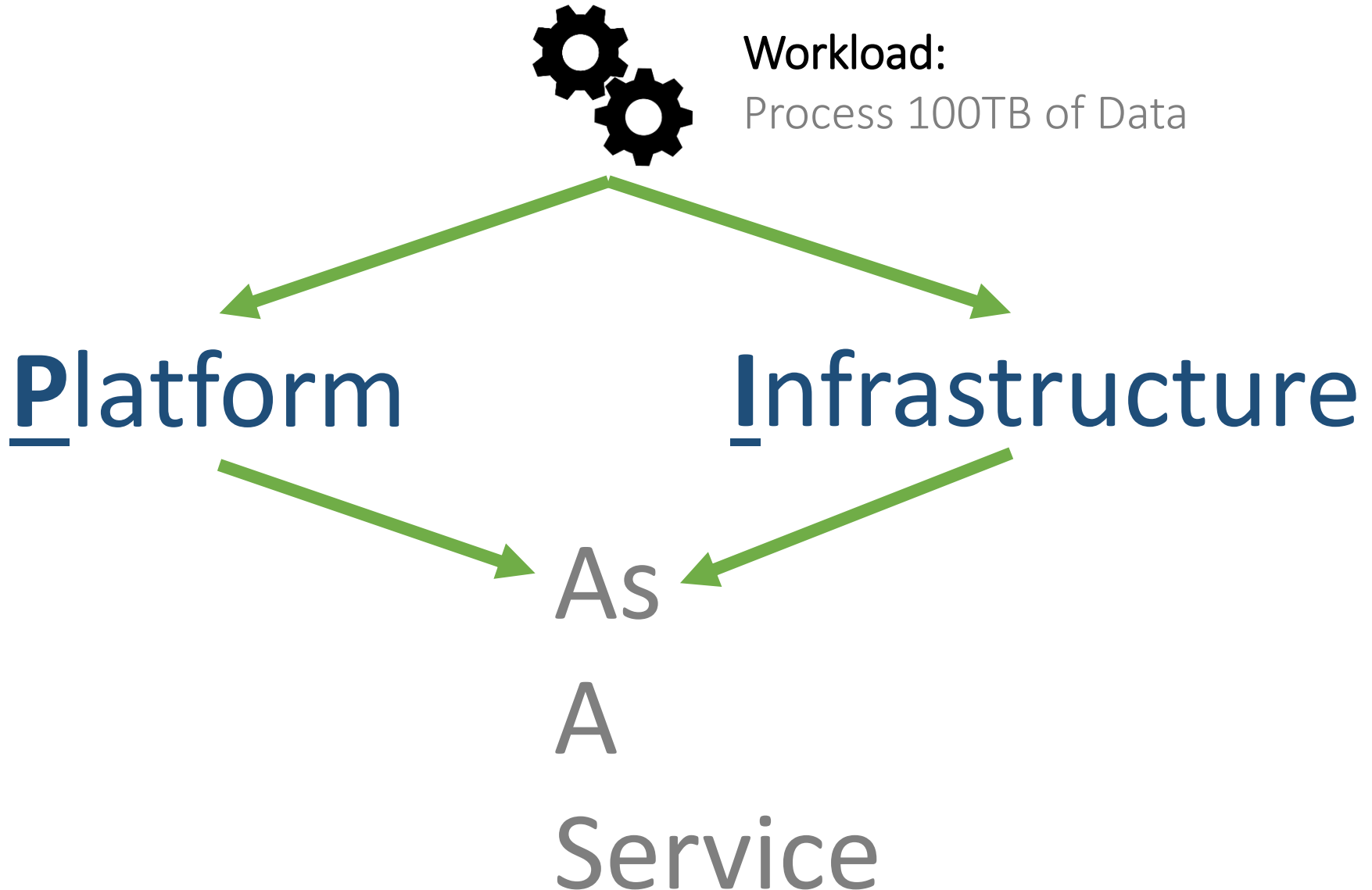
Platform

Infrastructure

As

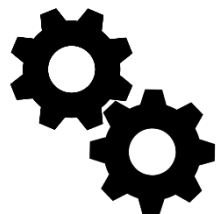
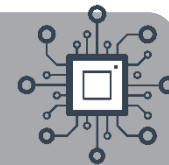
A

Service





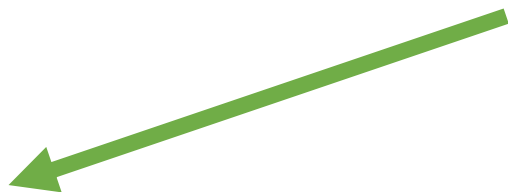
What Compute Type of Compute?



Workload:

Process 100TB of Data

Platform



As

A

Service

IaaS

PaaS

Applications

Applications

Data

Data

Runtime

Runtime

Middleware

Middleware

Operating System

Operating System

Virtualization

Virtualization

Servers

Servers

Storage

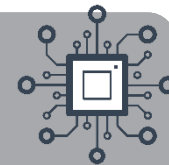
Storage

Networking

Networking



Data Transformation – Compute



Data Lake Analytics



HDInsight



Relational Database



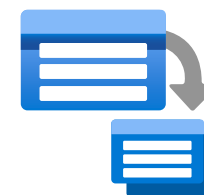
Synapse –
SQL Pools or
Spark Pools



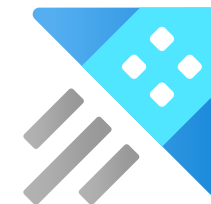
Databricks



Batch Service



Data Explorer



Automation



Cosmos



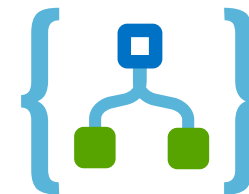
Functions



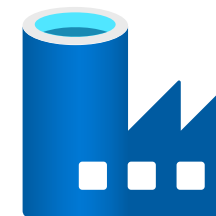
Power BI
Data Flows



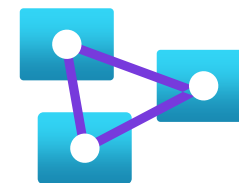
Logic Apps



Data Factory
Data Flows

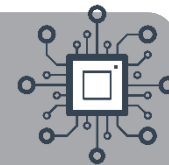


Analysis
Services





Data Transformation – Compute



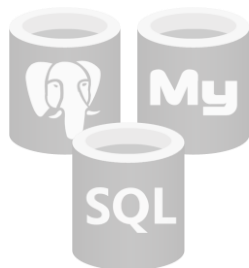
Data Lake Analytics



HDInsight



Relational Database



Synapse –
SQL Pools or
Spark Pools



Databricks



Batch Service



Data Explorer



Automation



Cosmos



Functions



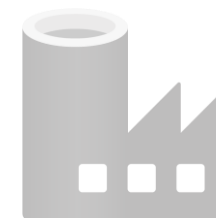
Power BI
Data Flows



Logic Apps



Data Factory
Data Flows

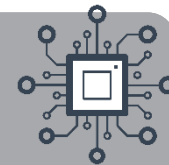


Analysis
Services





Data Transformation – Compute



Data Lake Analytics



HDInsight



Relational Database



Batch Service



Data Explorer



Automation



Cosmos



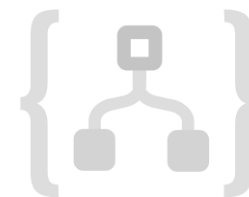
Functions



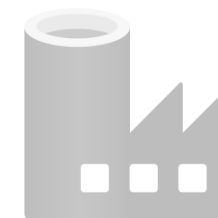
Power BI Data Flows



Logic Apps



Data Factory Data Flows

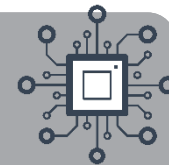


Analysis Services





Data Transformation – Compute



Data Lake
Analytics

HDInsight

Relational
Database



Batch Service

Data Explorer



Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

Languages
العربية
Deutsch
Español
Français

Article **Talk**

Read **Edit** View history

Search Wikipedia

The Lake House (film)

From Wikipedia, the free encyclopedia



This article includes a list of general references, but it remains largely unverified because it lacks sufficient corresponding inline citations. Please help to improve this article by introducing more precise citations. (October 2017) (Learn how and when to remove this template message)

The Lake House is a 2006 American fantasy romantic drama film directed by Alejandro Agresti, starring Keanu Reeves and Sandra Bullock (who had previously appeared together in the box office hit *Speed*). It was written by David Auburn.^[2] A remake of the South Korean motion picture *Il Mare* (2000), it centers on an architect living in 2004 and a doctor living in 2006 who meet via letters left in a mailbox at the lake house where they have lived at separate points in time. They carry on correspondence over two years, remaining separated by their original difference of two years.^[3]

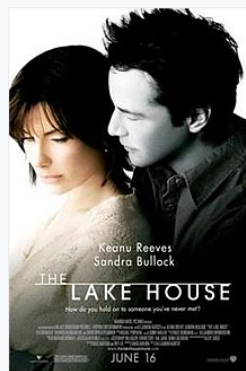
Contents [hide]

- Plot
- Cast
- Production
- Music
- Reception
 - Box office
 - Critical response
 - Home media
 - Awards
- References
- External links

Plot [edit]

In 2006, Dr. Kate Forster (Sandra Bullock) is leaving a lake house that she has been renting in Chicago. Kate leaves a note in the mailbox for the next tenant to forward her mail, adding that the paint-embedded pawprints on the path leading to the house were already there when she arrived.

The Lake House



Theatrical release poster

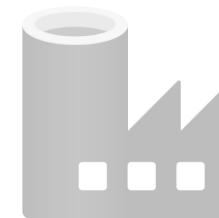
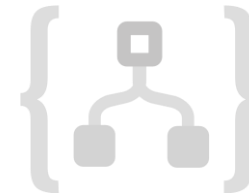
Directed by Alejandro Agresti
Written by David Auburn
Based on *Il Mare*
by Kim Eun-jeong
Kim Mi-yeong
Produced by Doug Davison
Roy Lee
Starring Keanu Reeves

3I
WS


Logic Apps

Data Factory
Data Flows

Analysis
Services



Agenda



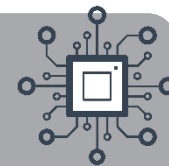
1. Design ✓
2. Extract ✓
3. Transform
4. Load

Compute ✓

Storage, Structure
& Data Format



Data Transformation – Storage & Format



Azure Storage Account



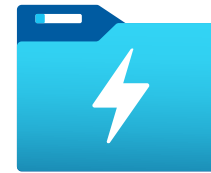
Azure Data Lake Gen2

Hadoop Distributed File System (HDFS)



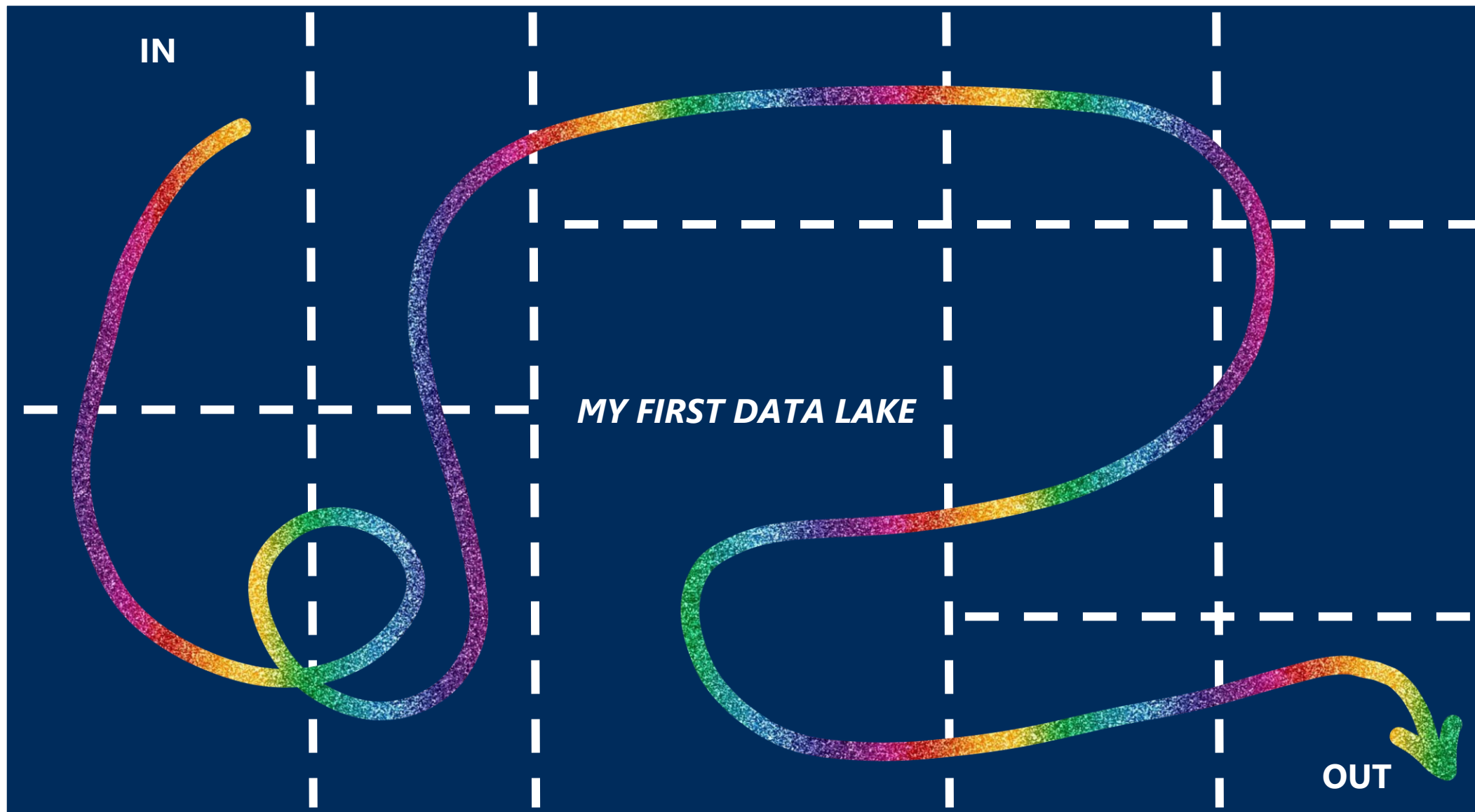
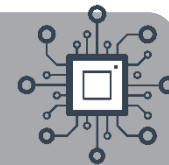


Data Transformation – Storage & Format



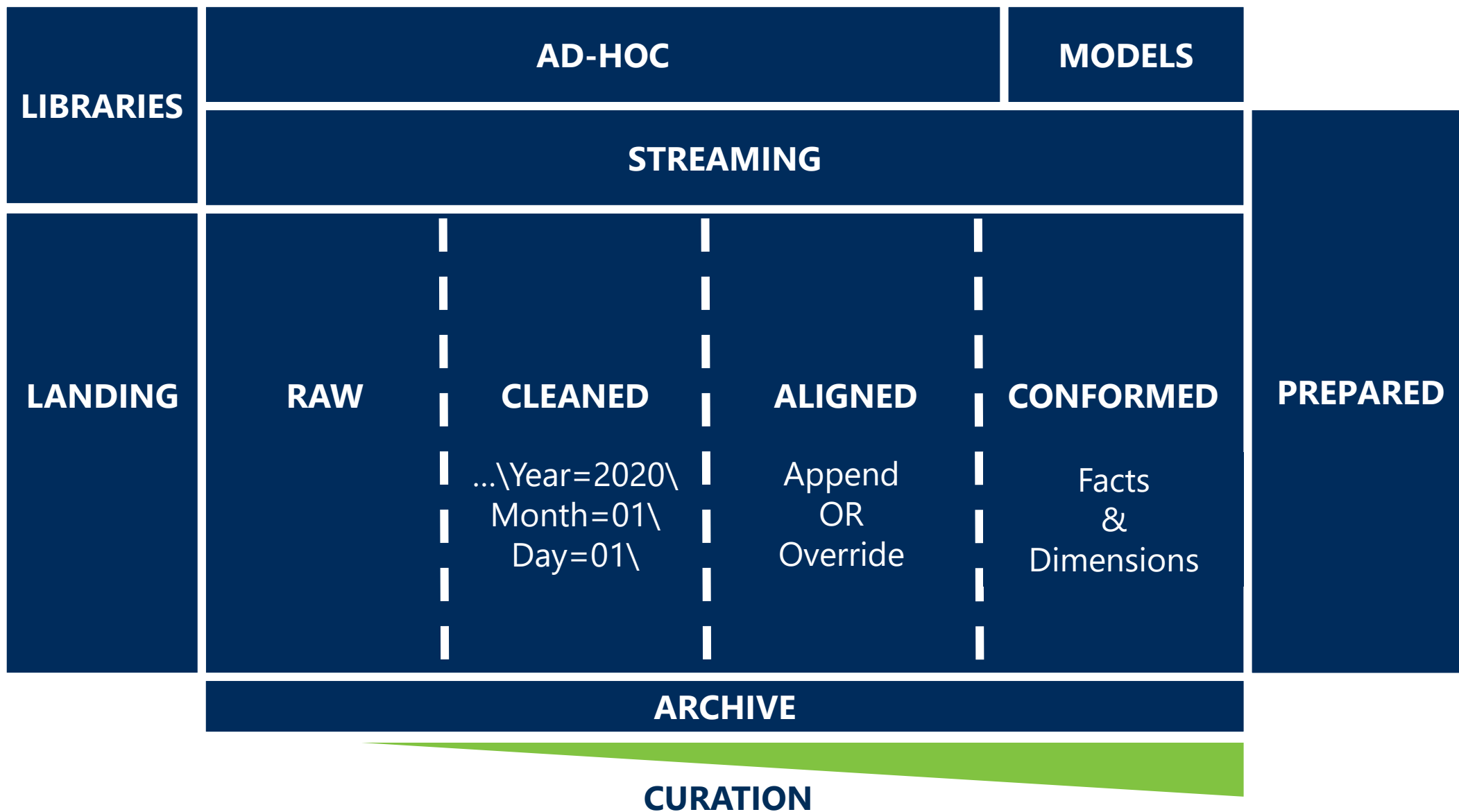
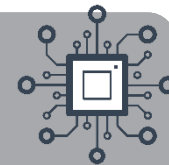


Data Transformation – Storage & Format



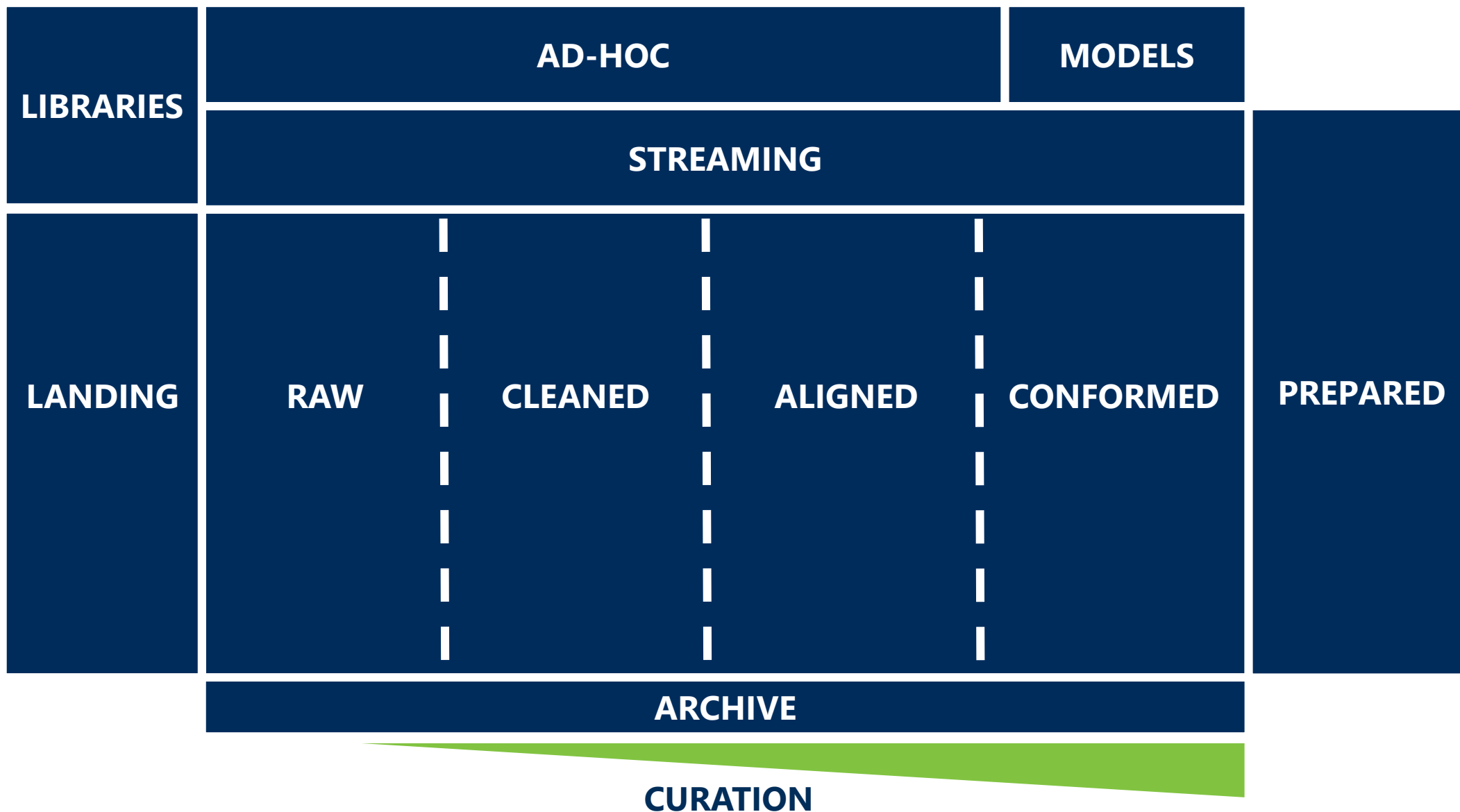
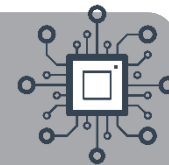


Data Transformation – Storage & Format



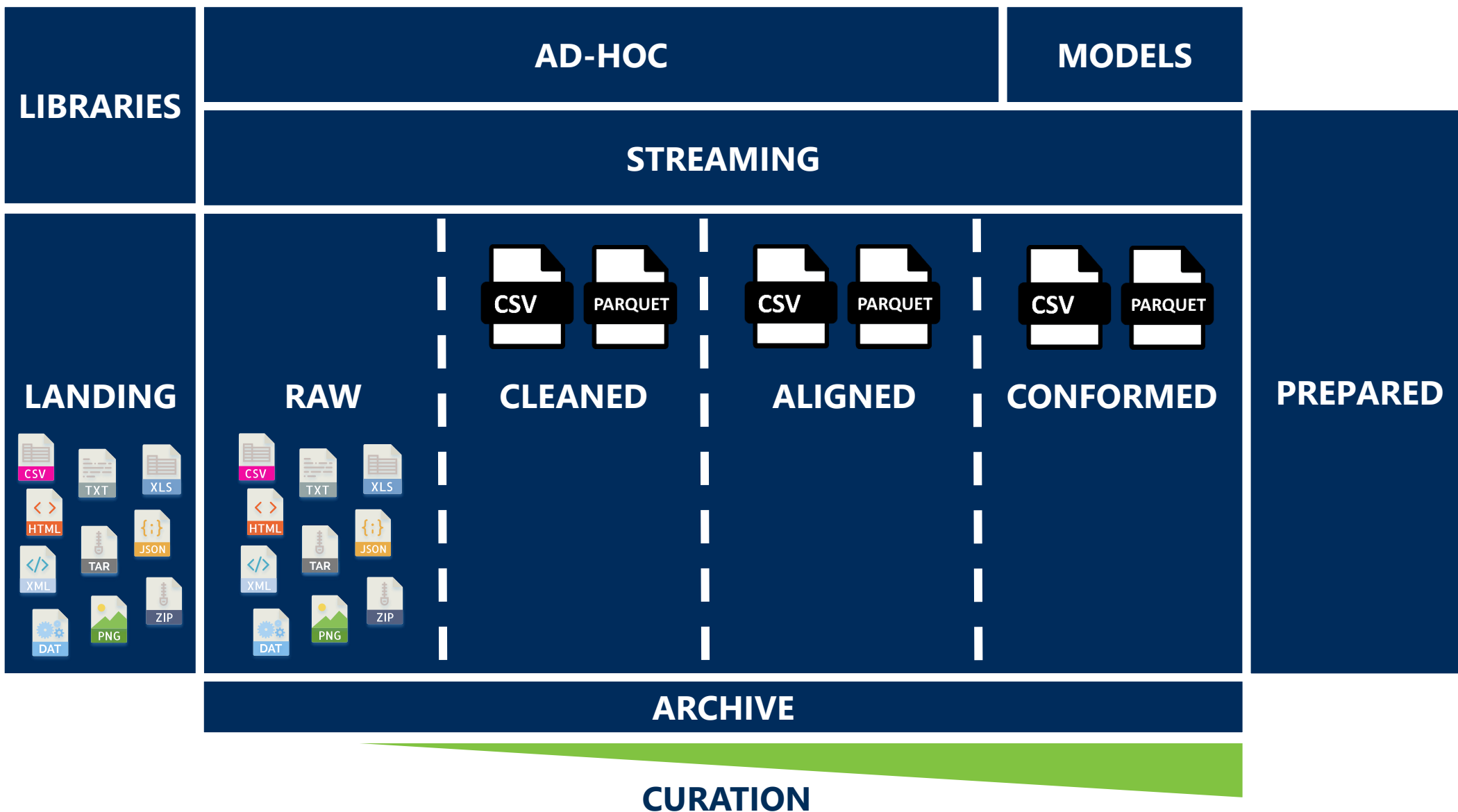
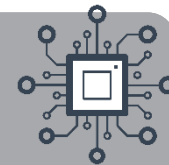


Data Transformation – Storage & Format



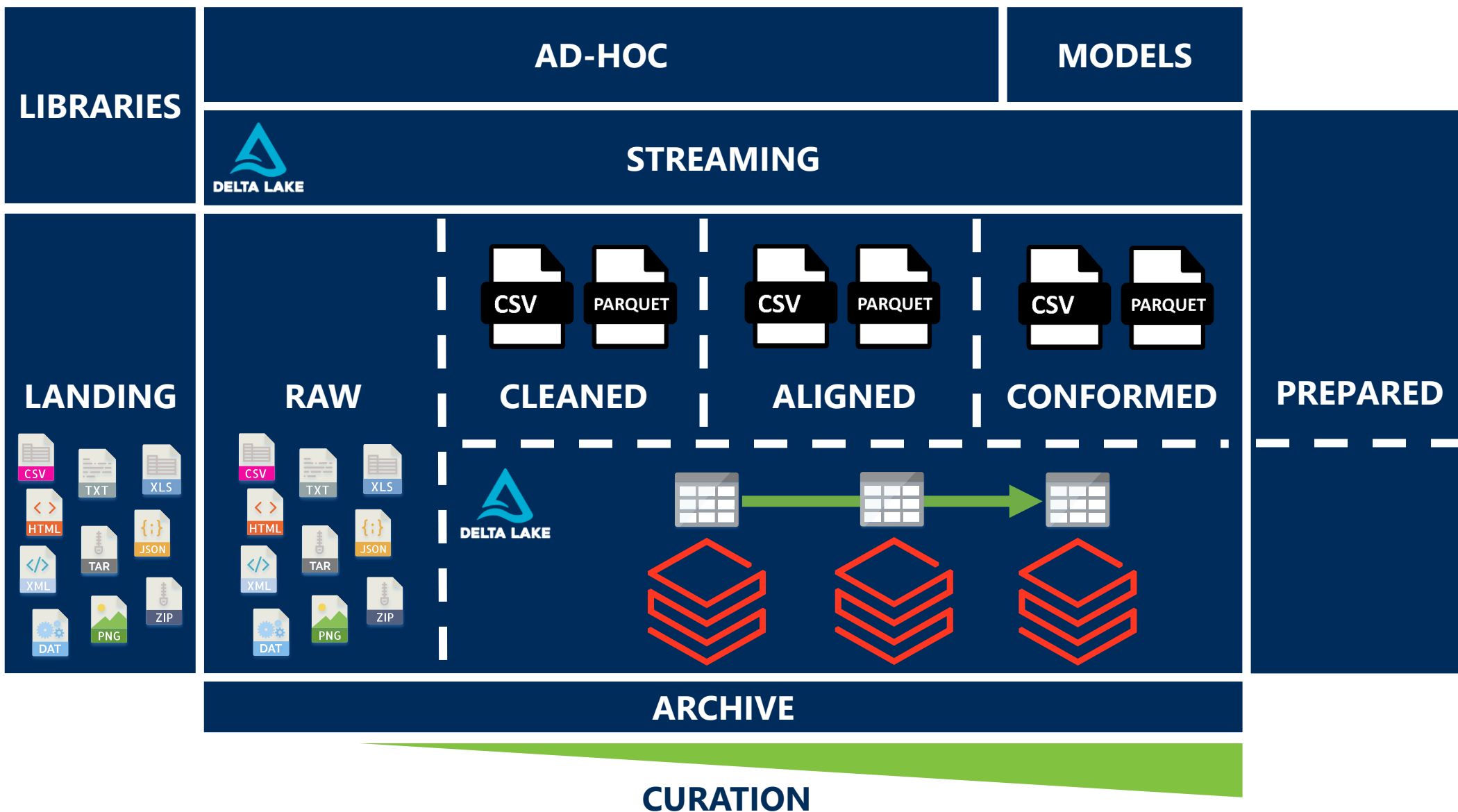
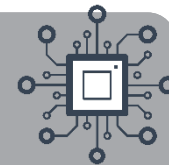


Data Transformation – Storage & Format






Data Transformation – Storage & Format



Agenda



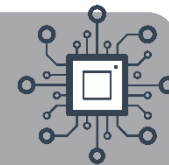
1. Design ✓
2. Extract ✓
3. Transform
4. Load

Compute ✓

Storage, Structure
& Data Format ✓



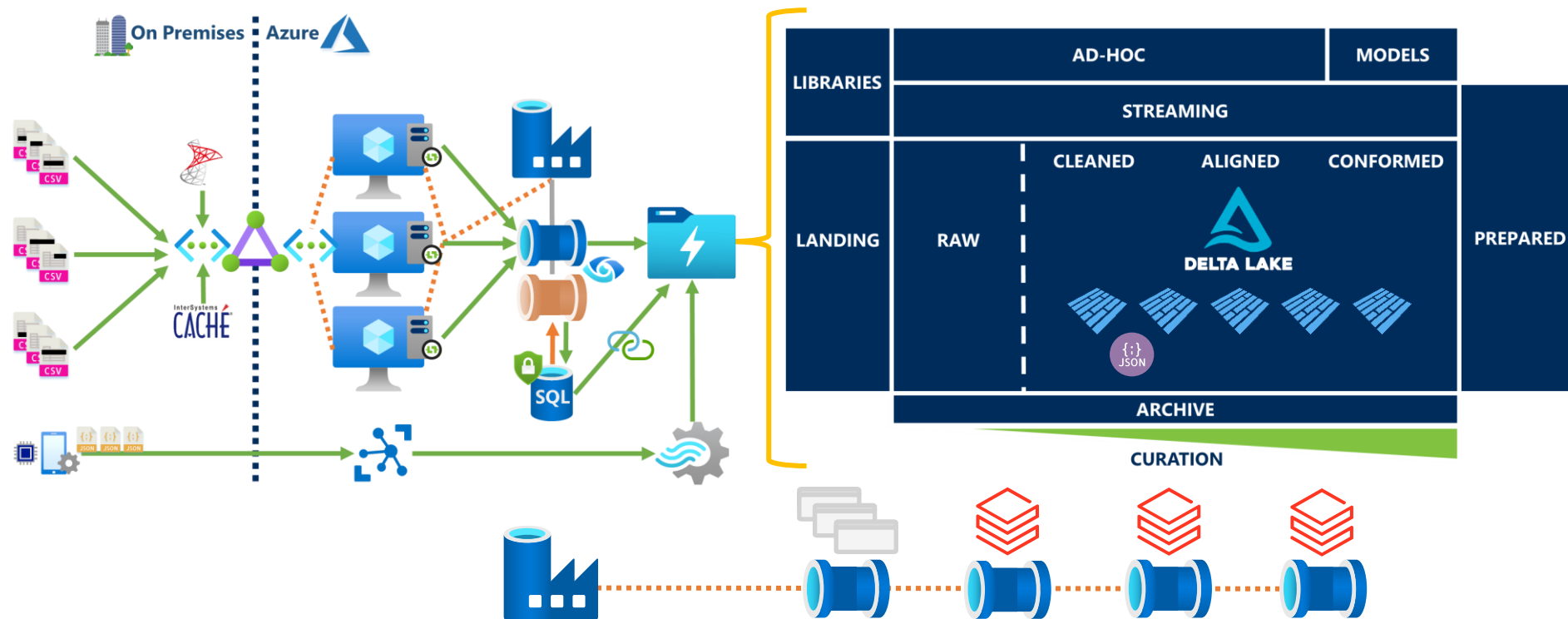
Overall Architecture



Extract

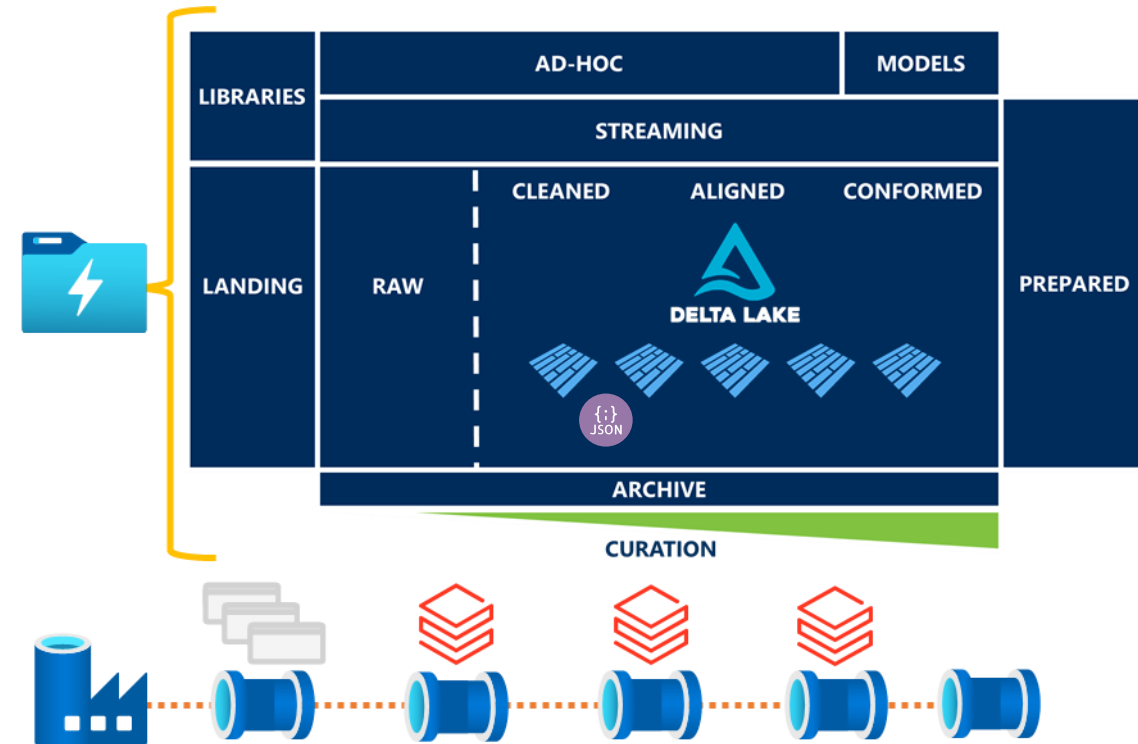
Transform

Load



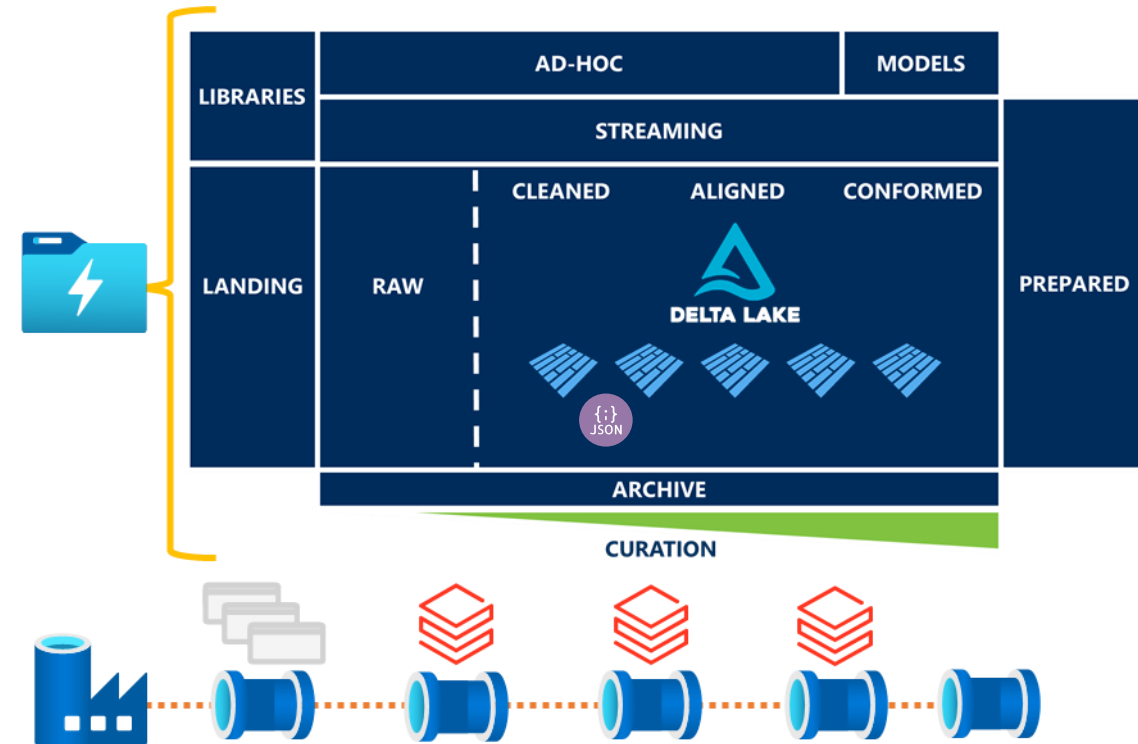
Agenda

1. Design ✓
2. Extract ✓
3. Transform ✓
4. Load



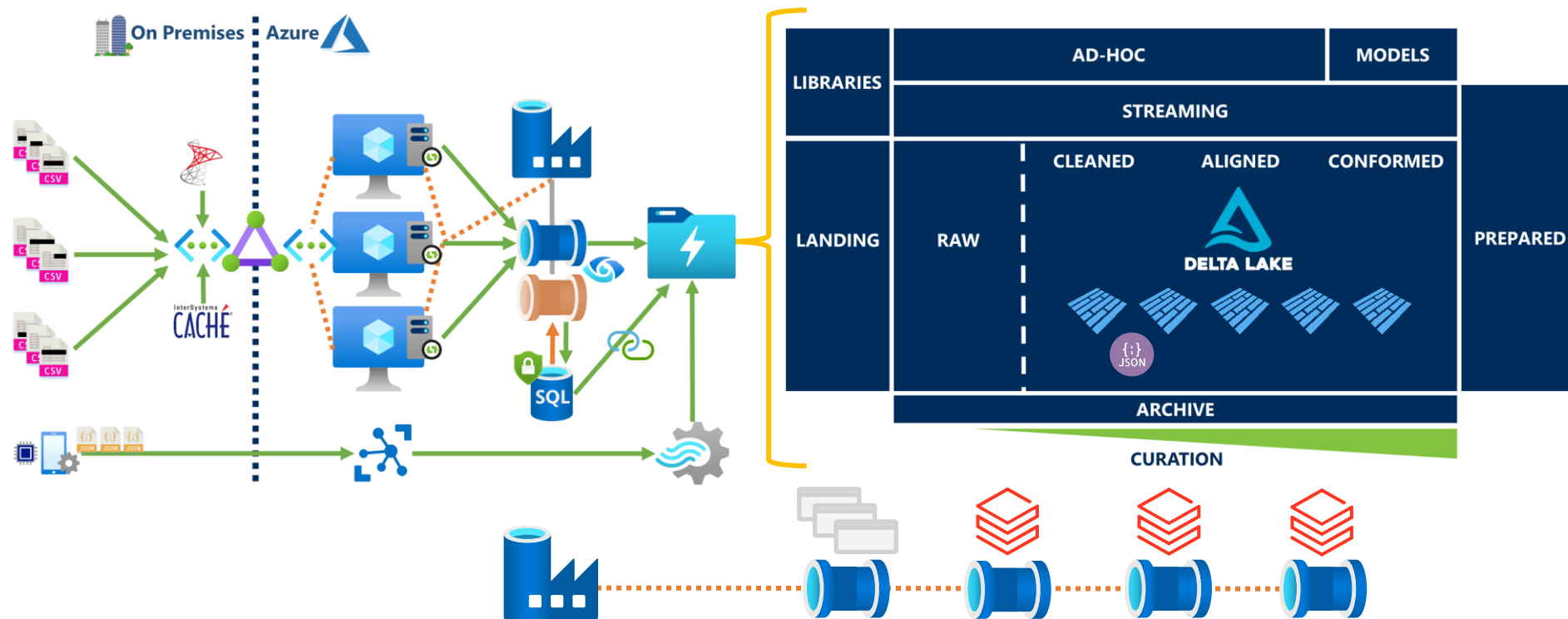
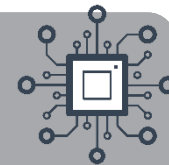
Agenda

1. Design ✓
2. Extract ✓
3. Transform ✓
4. Load



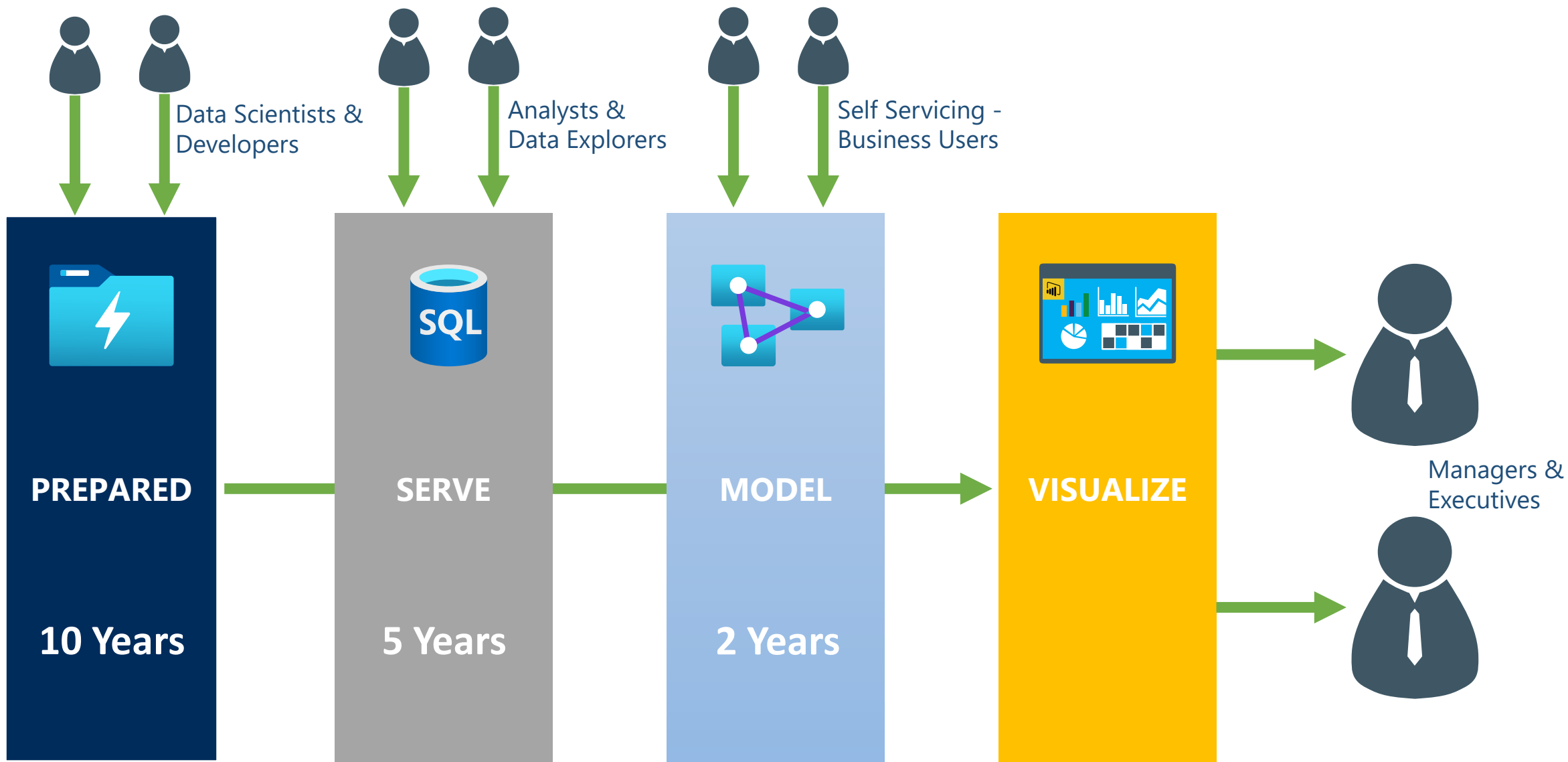
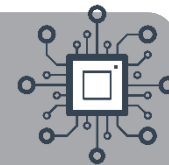


Overall Architecture



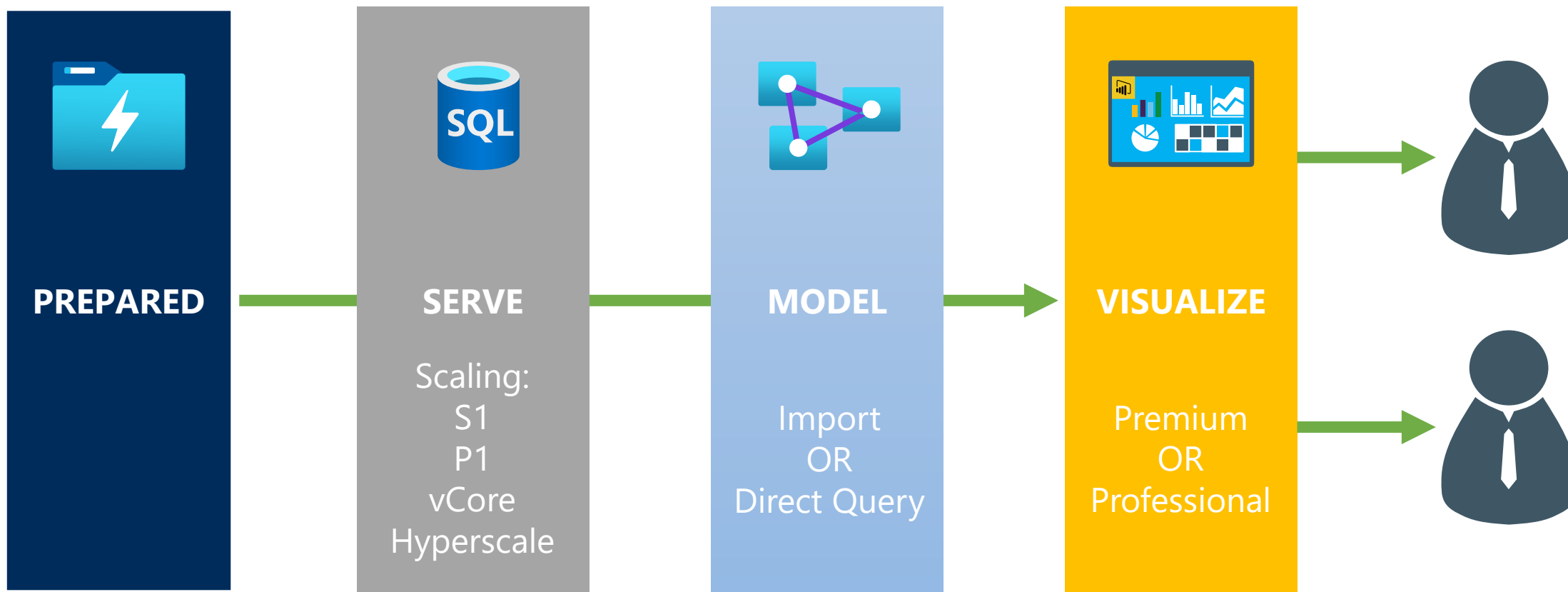
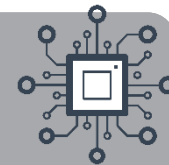


Loading & Consuming Data



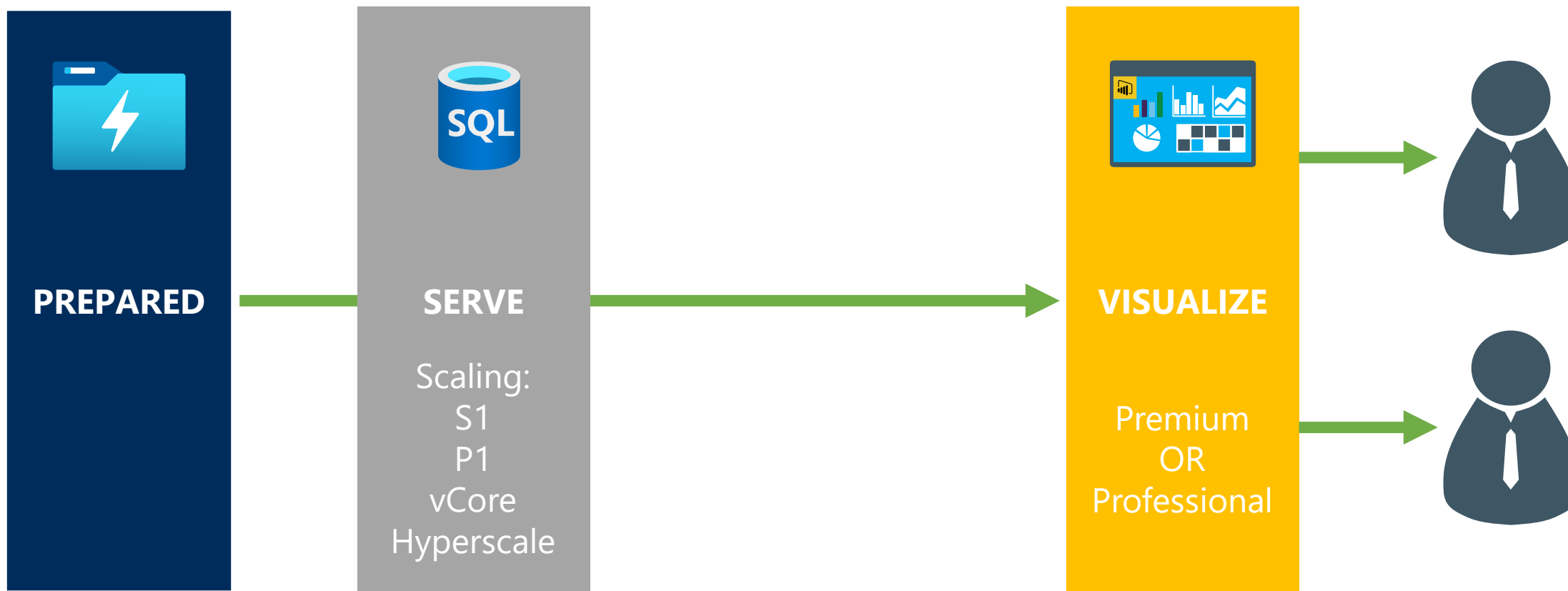
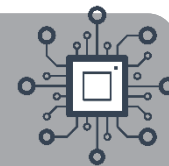


Loading & Consuming Data



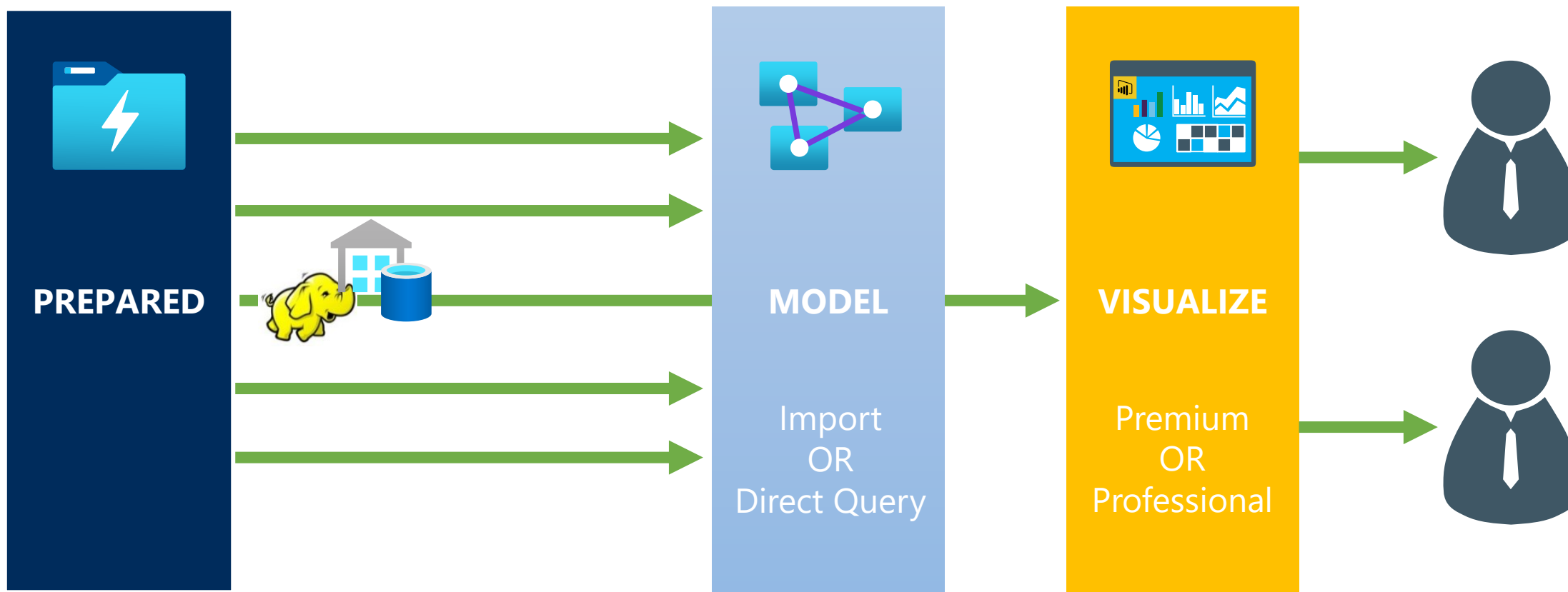
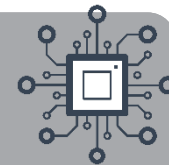


Loading & Consuming Data



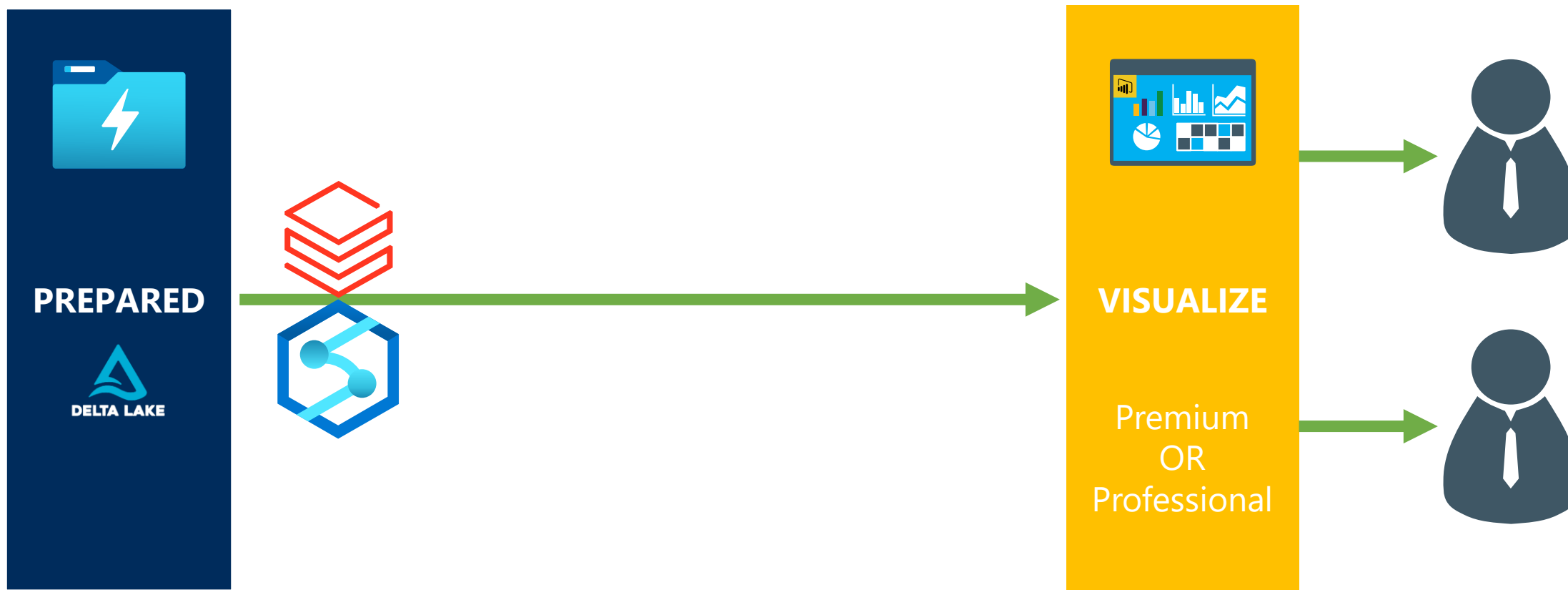
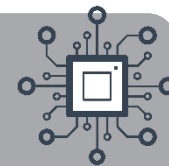


Loading & Consuming Data



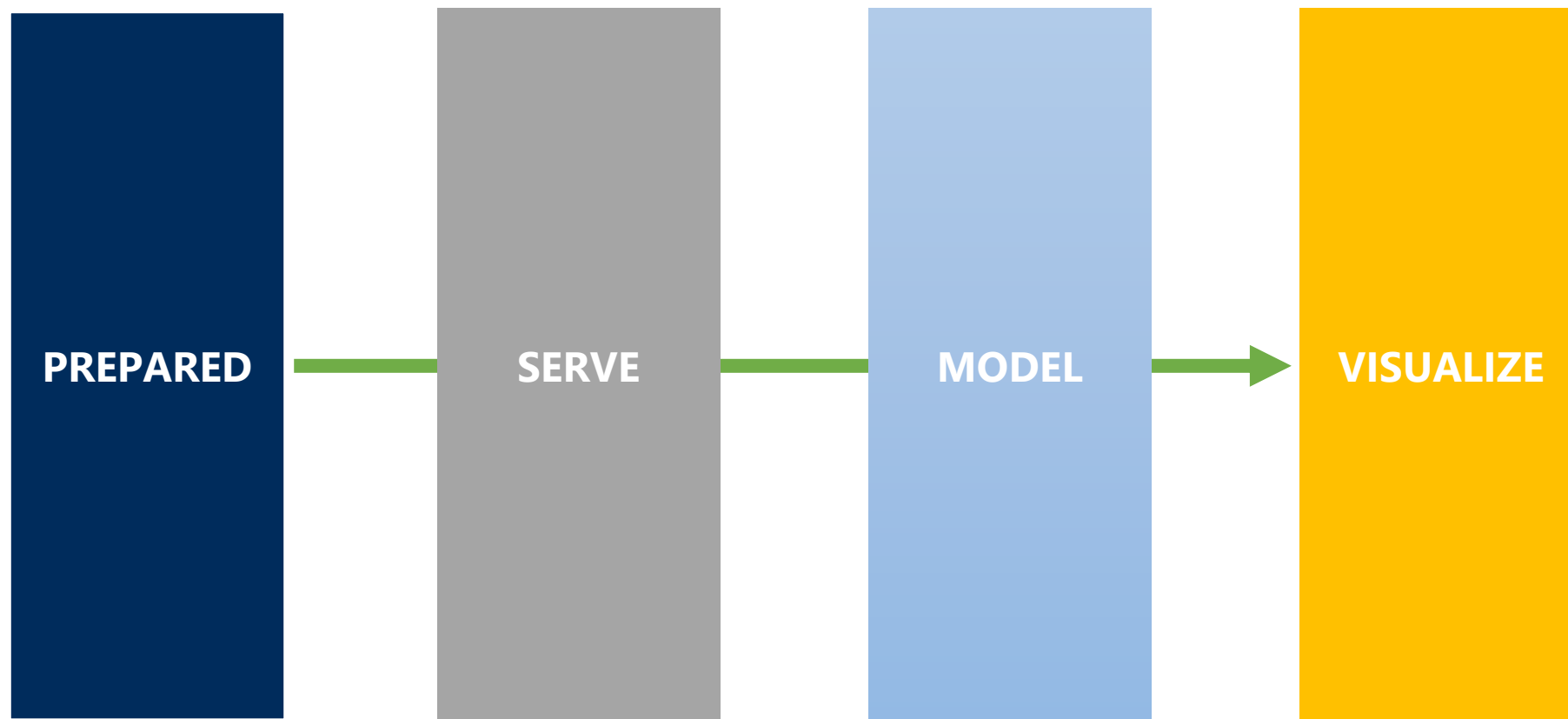
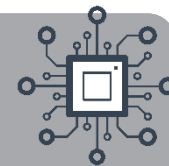


Loading & Consuming Data



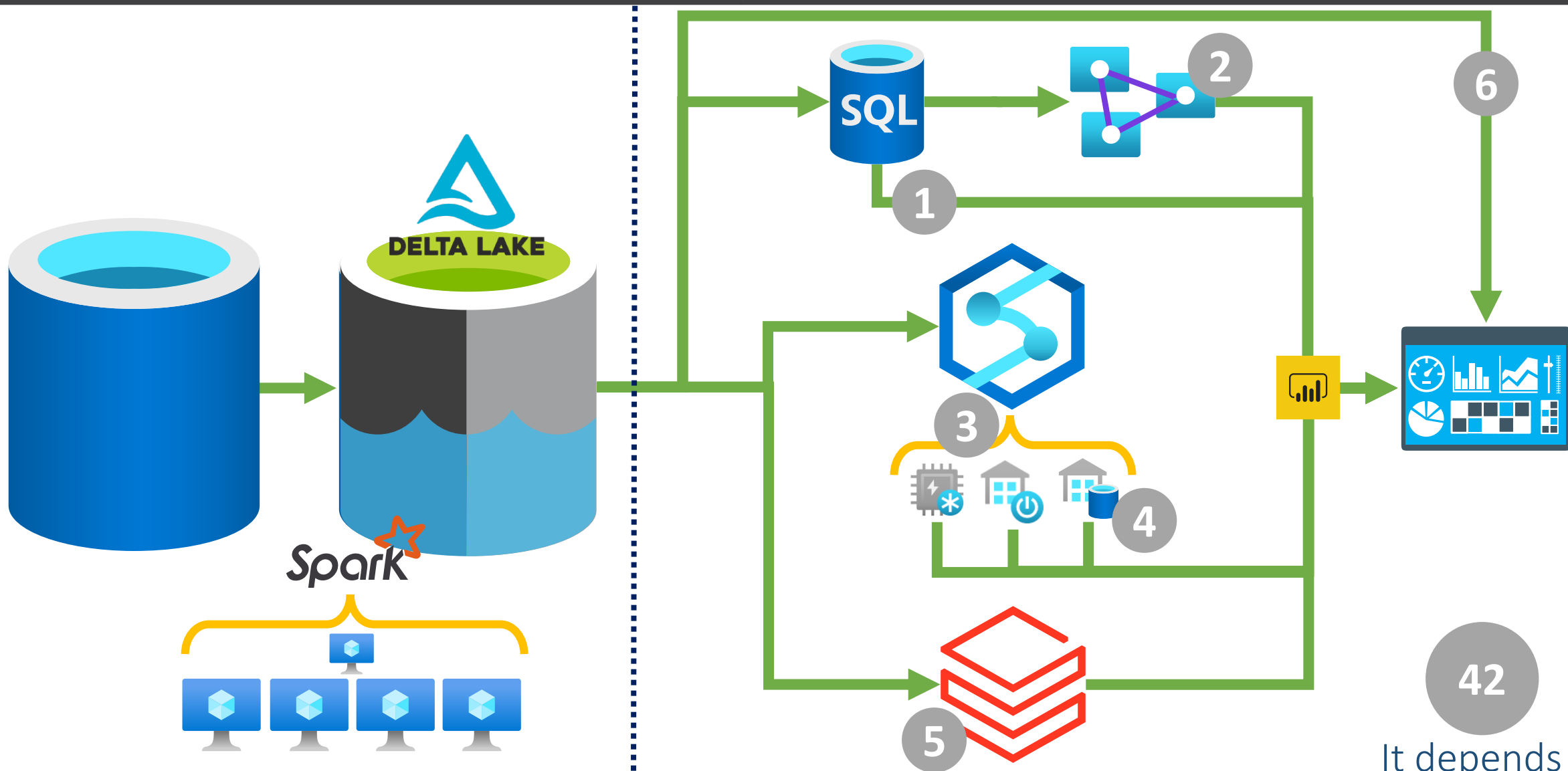
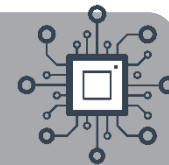


Consuming Our Lake House in Azure



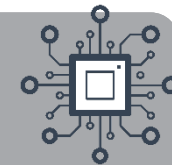


Consuming Our Lake House in Azure





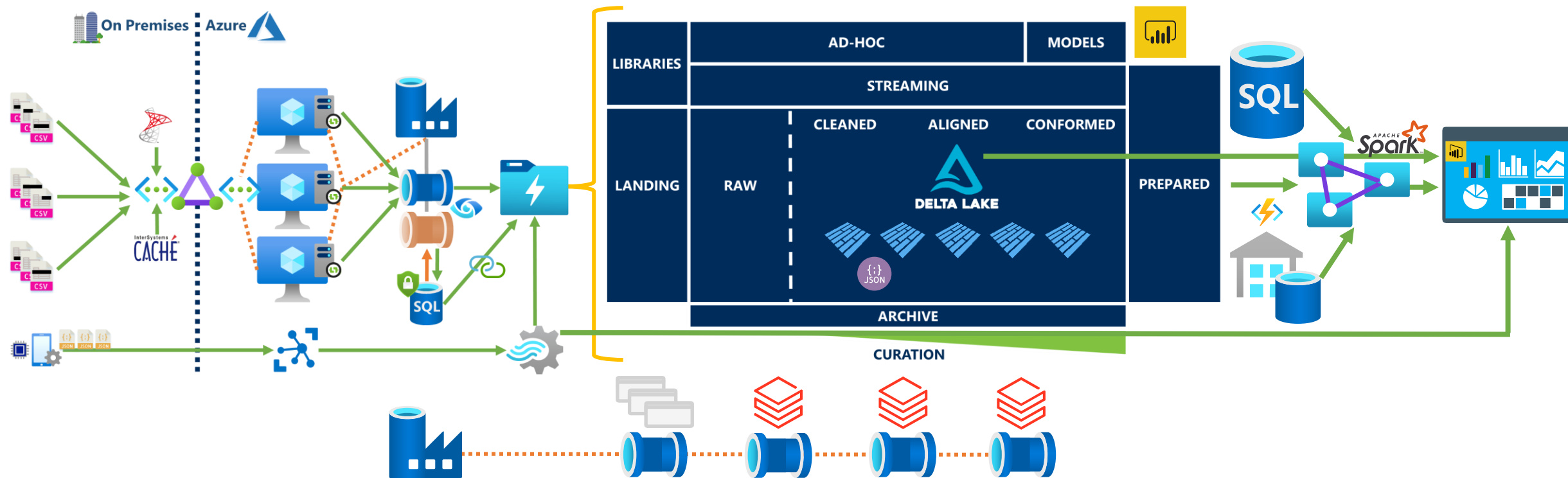
Overall Architecture



Extract

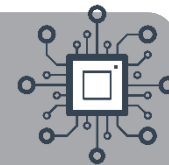
Transform

Load





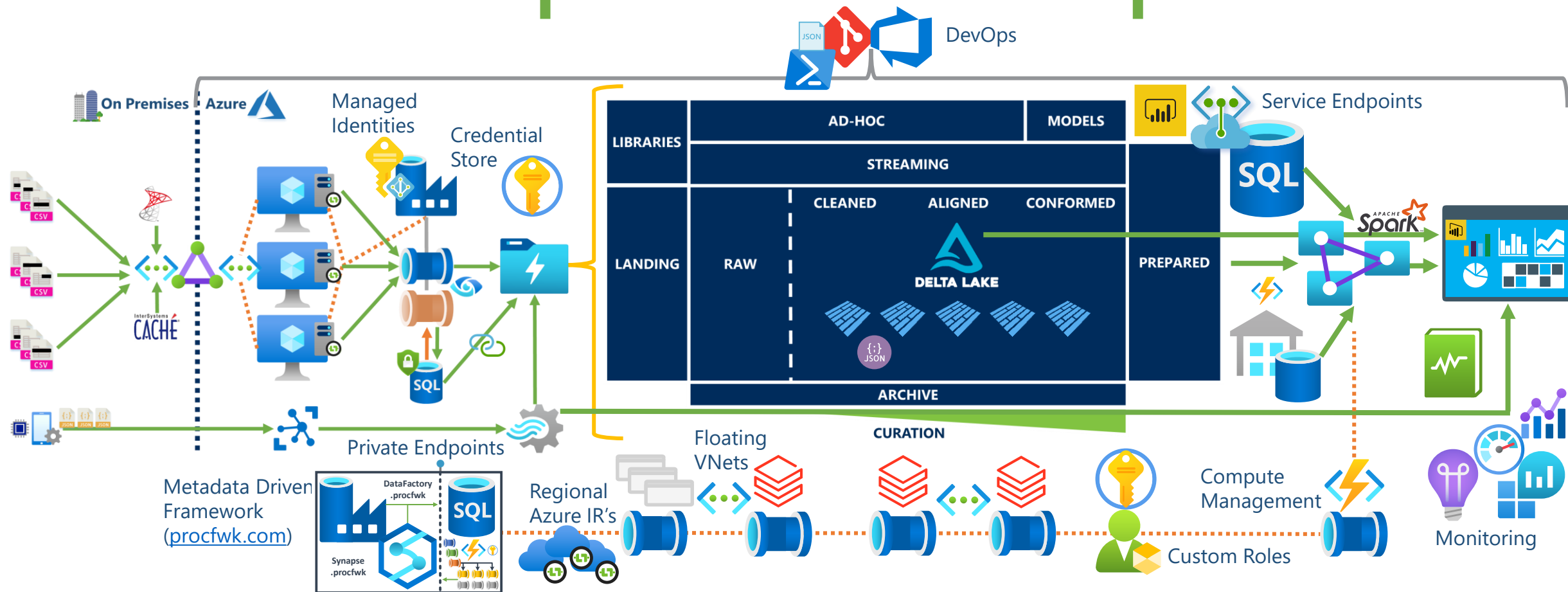
Overall Architecture



Extract

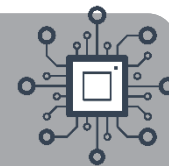
Transform

Load





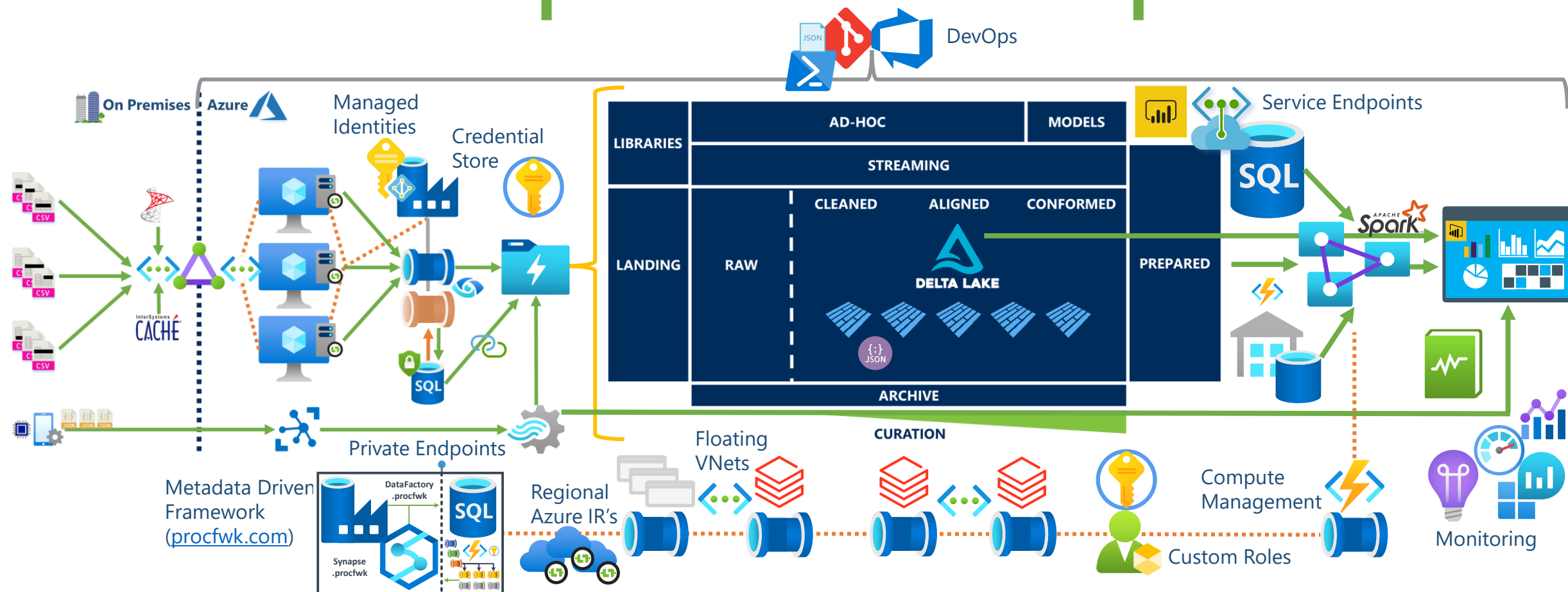
Overall Architecture



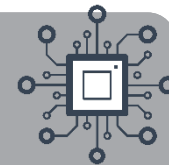
Extract

Transform

Load



Q: Should we build our data platform solution like this?... A: It depends!



Thank you for listening...

Paul Andrew



avanade



mrpaulandrew.tech

Blog: mrpaulandrew.com
YouTube: c/mrpaulandrew
Email: paul@mrpaulandrew.com

Twitter: @mrpaulandrew
LinkedIn: In/mrpaulandrew

GitHub: github.com/mrpaulandrew

Session
Feedback

