An Architects Guide to Delivering Data Insights

Using the Microsoft Azure Data Platform

Paul Andrew | Group Manager & Analytics Architect

Microsoft® MVP Most Valuable Professional

avanade

@MrPaulAndrew

In/MrPaulAndrew

MrPaulAndrew.com

https://github.com/mrpaulandrew

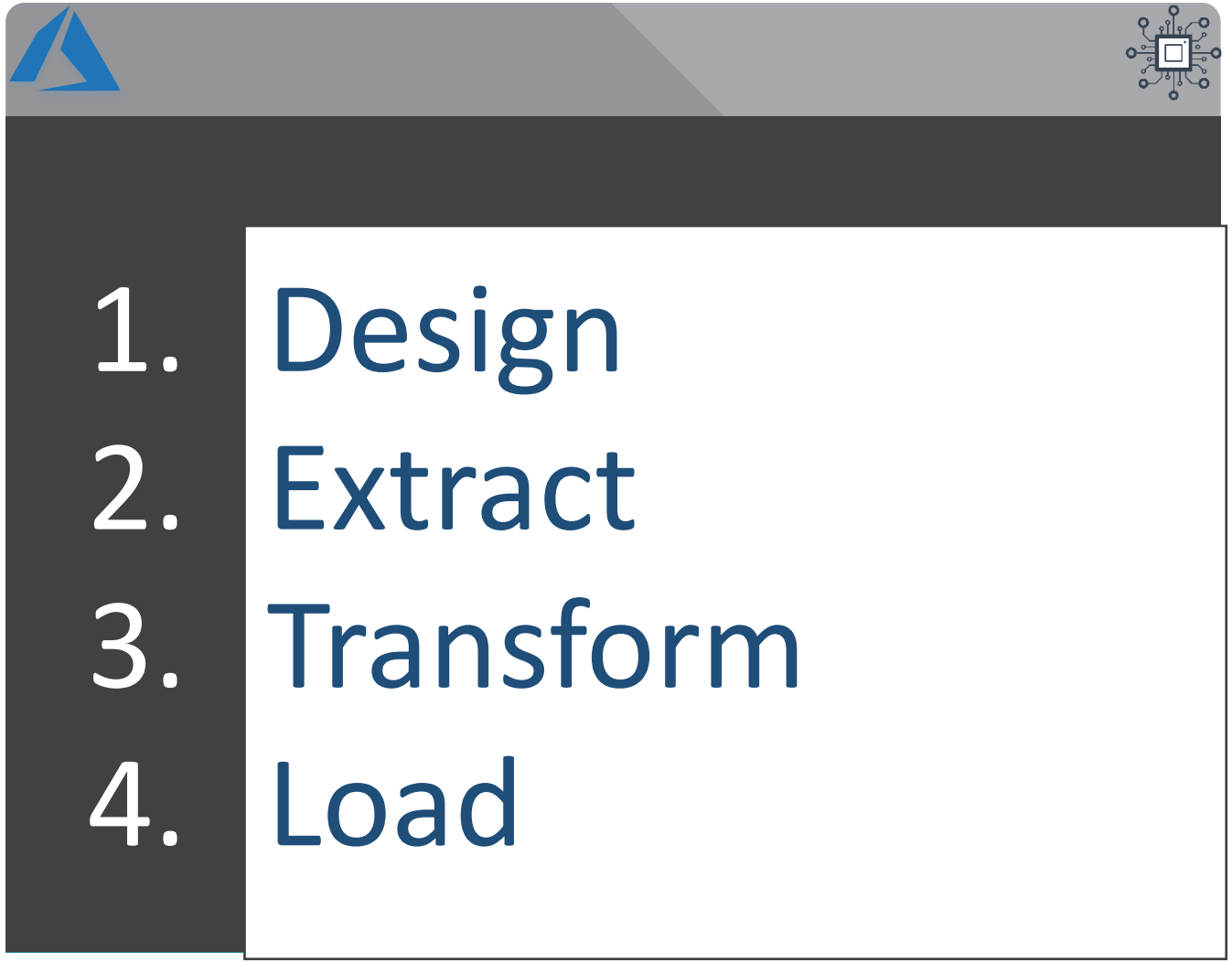
CommunityEvents

Demo code, content and slides from various community events.

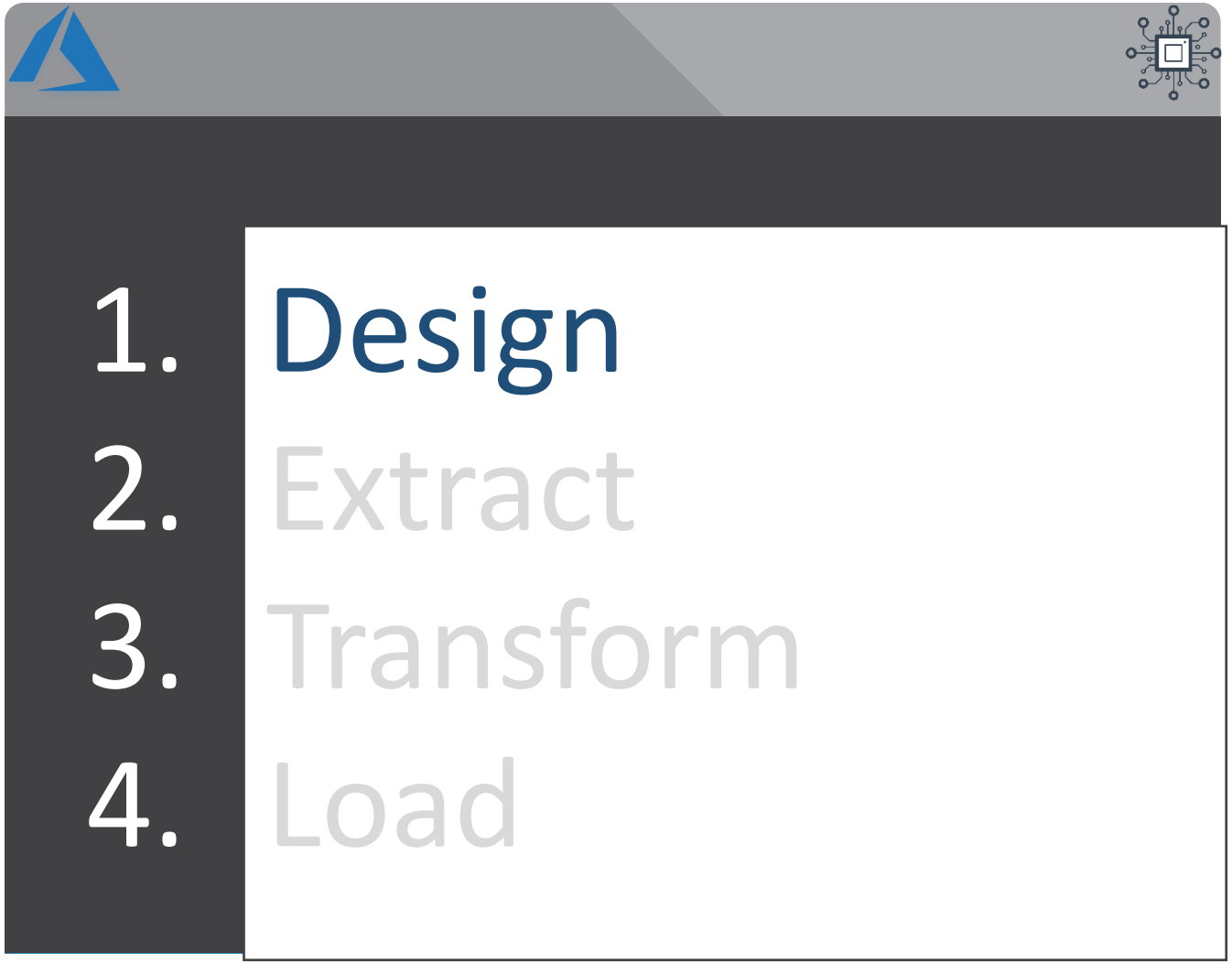
C++

{Event/Location}-{Month}-{Year}

Agenda

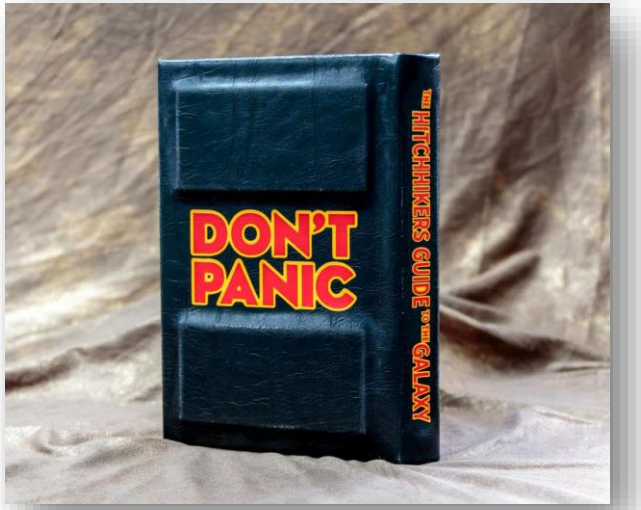


Agenda



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

Answer:
42 ✗



Answer:
It depends! ✓

Question:

What is big data?

Answer:

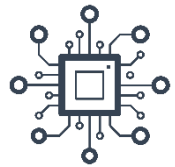
It depends!

Answer:

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

- Buck Woody

@BuckWoodyMSFT



Goal

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

Data Sources

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

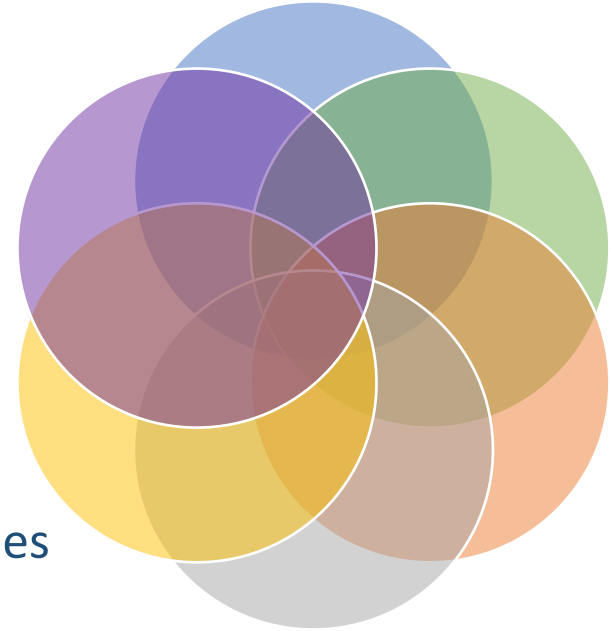
What is your primary design focus?

Resilience

- Disaster recovery
- Transaction level restart ability

Rapid Delivery

- Metadata driven
- Continuous deployments



Performance

- Complex partitioning
- Large compute clusters

(Re)Usability

- Generic code libraries
- Dataset contracts

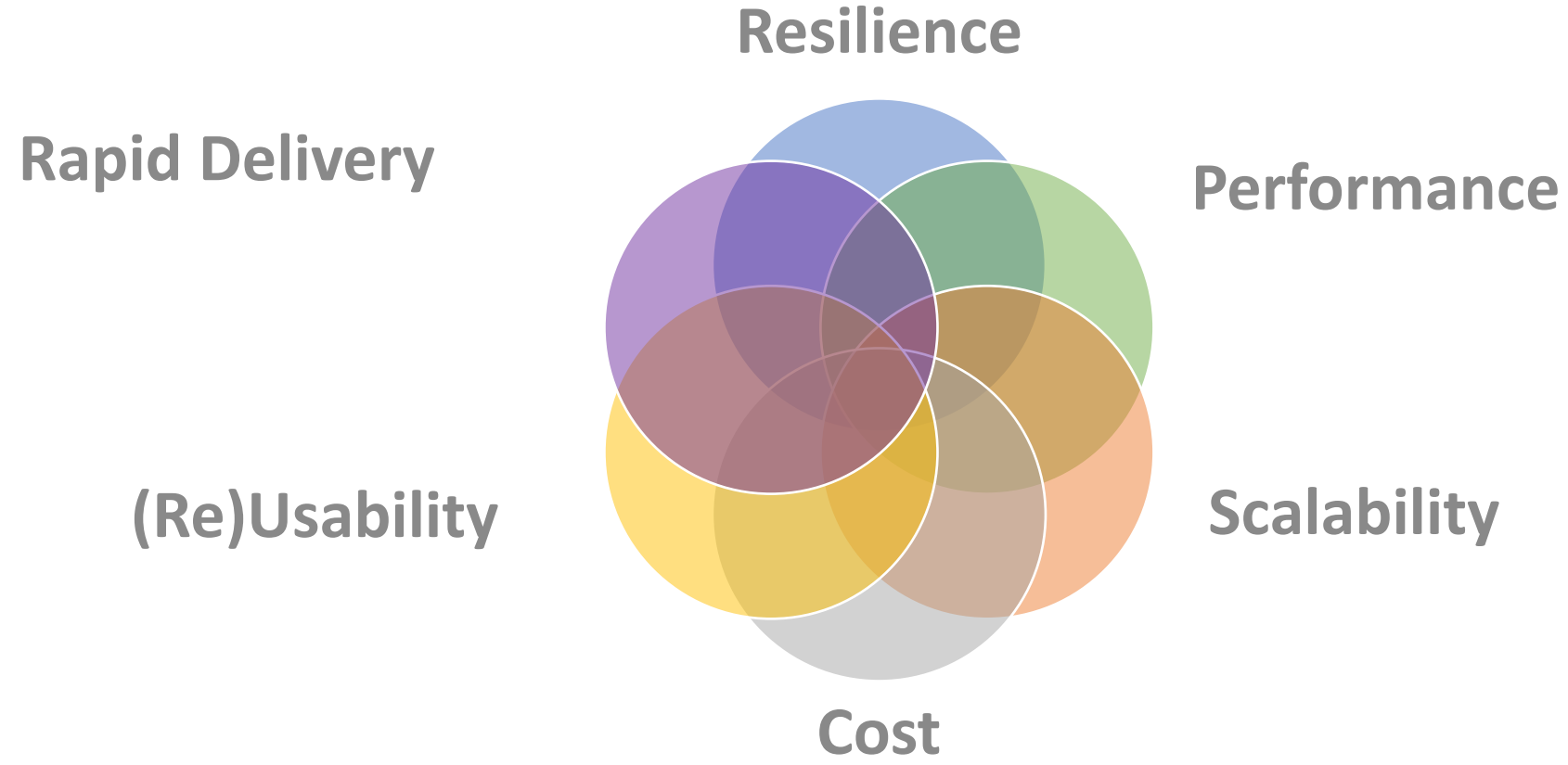
Scalability

- Auto scaling microservices
- Event driven discrete batches

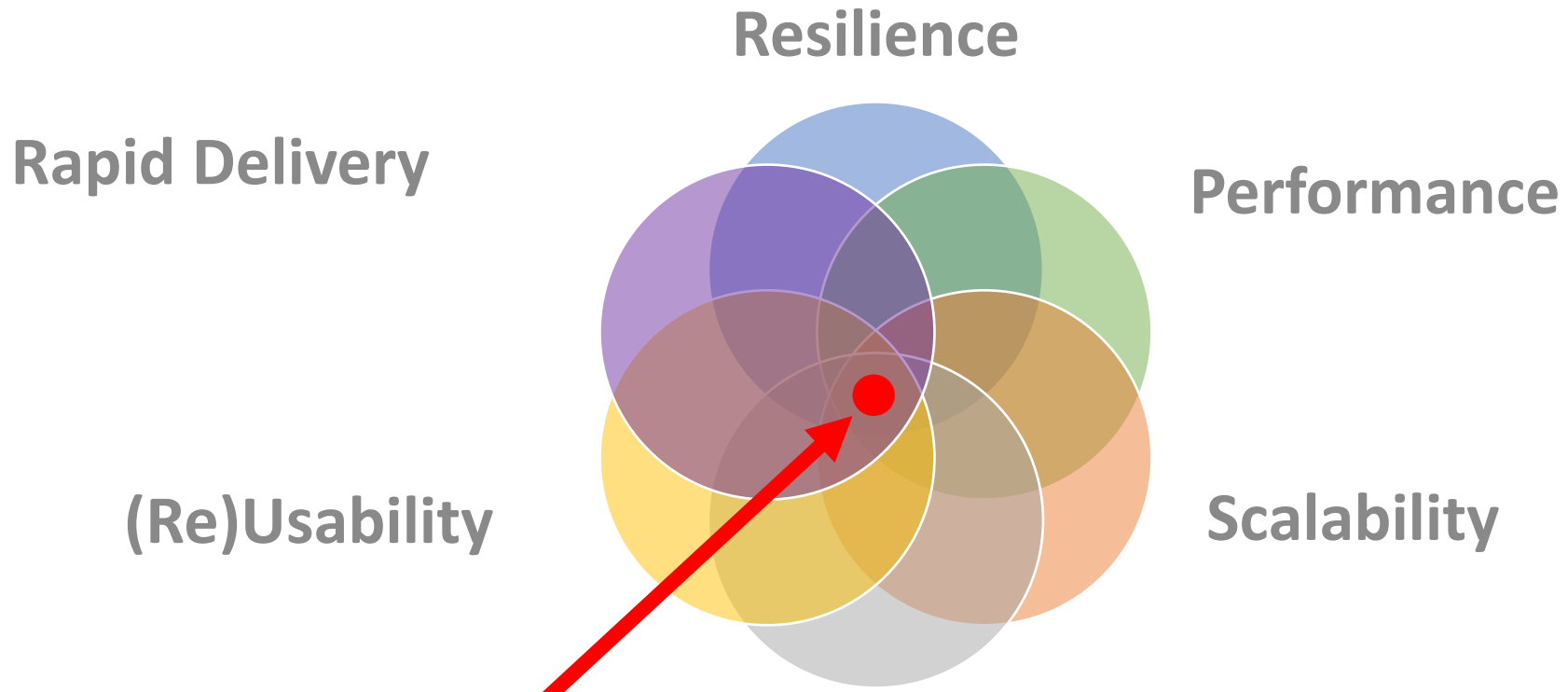
Cost

- Minimum resources used
- Dynamic resource management

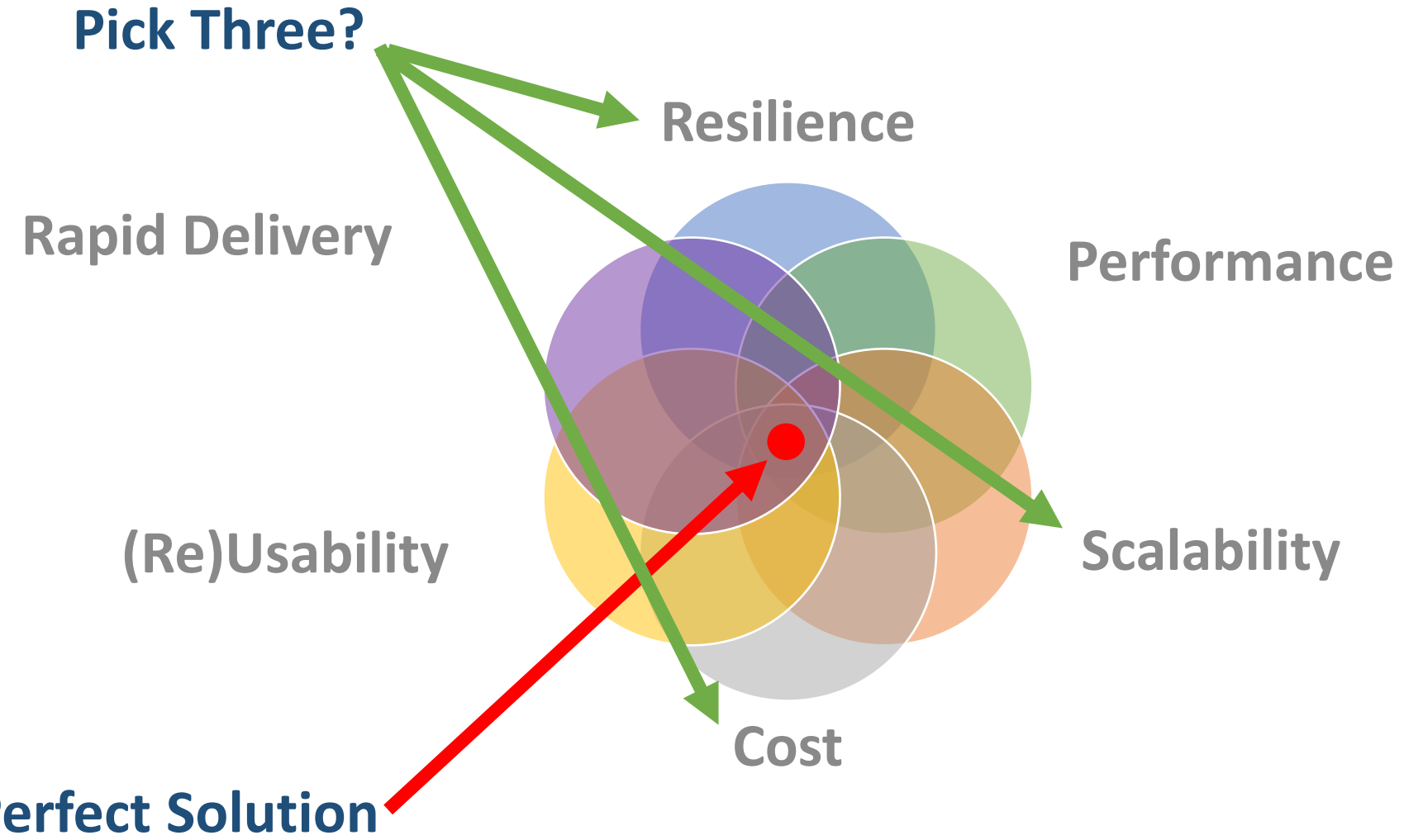
What is your primary design focus?



What is your primary design focus?

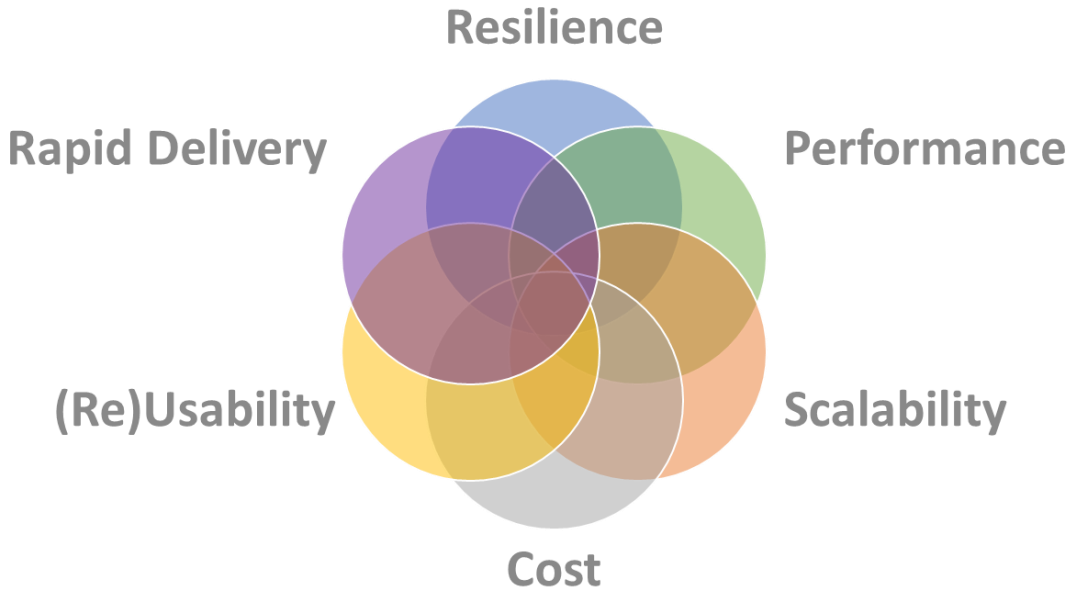


What is your primary design focus?



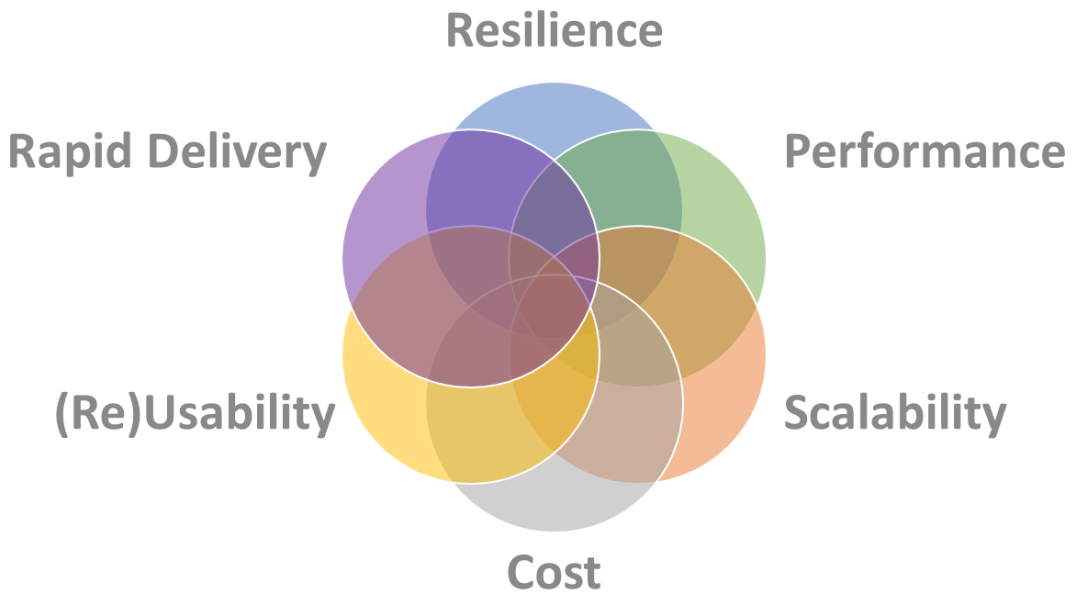
Agenda

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



Agenda

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



Data Extraction & Ingestion

Data Structure

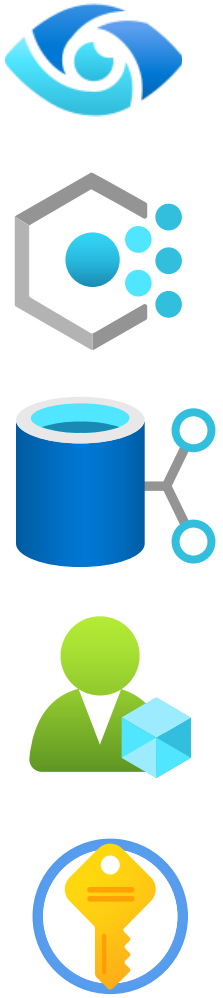
Data Source

Push or Pull

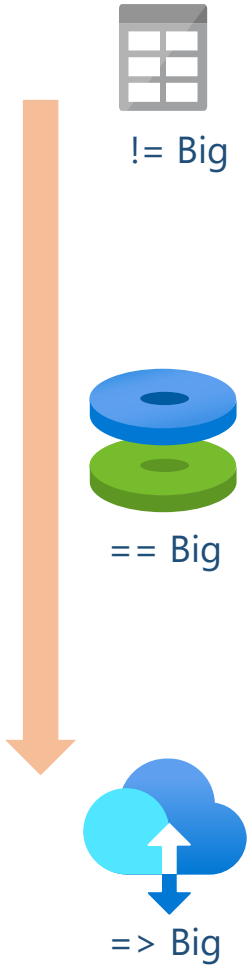
Batch or Speed

Public or Private Transfer

Data Sensitivity



Data Volume



Data Extraction & Ingestion – Spec v1

Data Structure

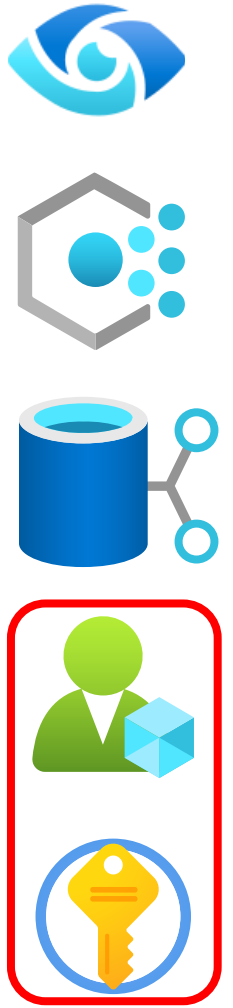
Data Source

Push or Pull

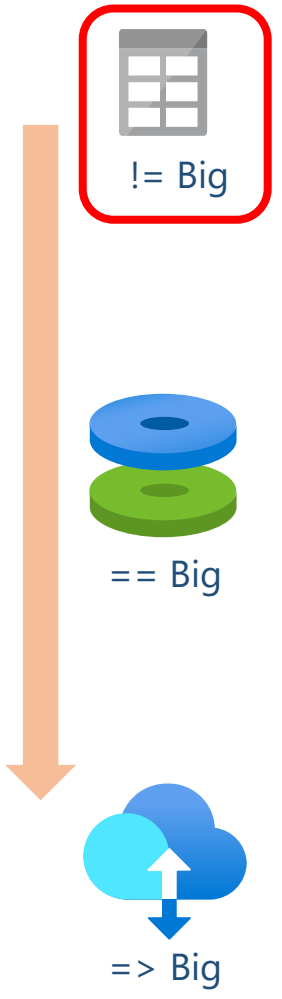
Batch or Speed

Public or Private Transfer

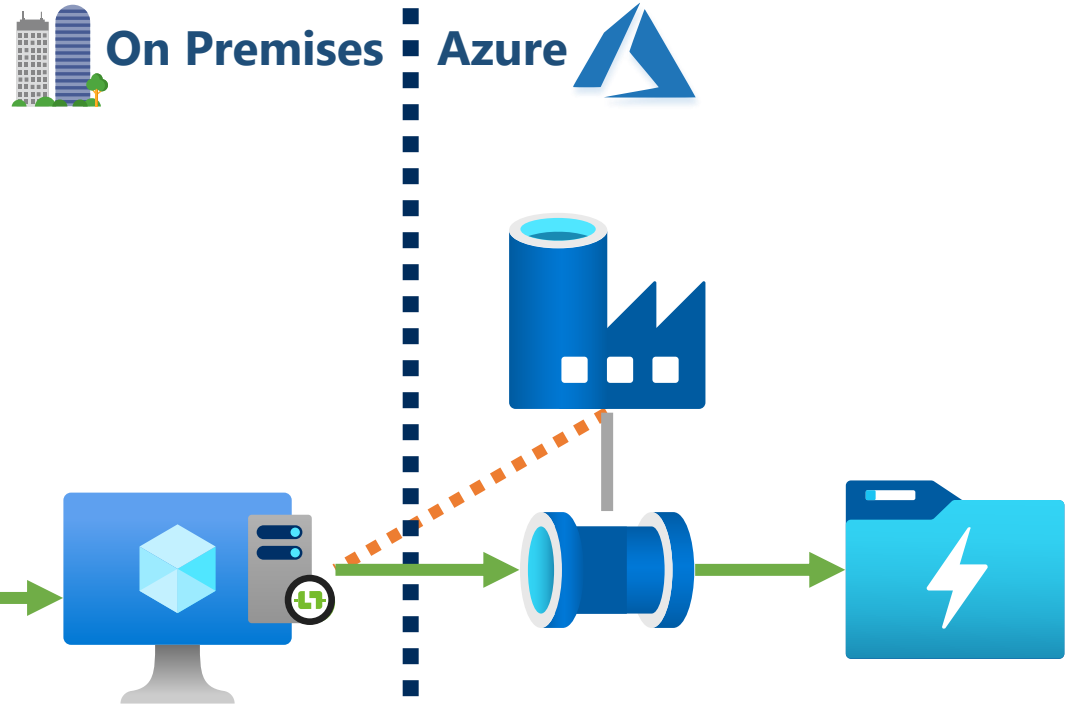
Data Sensitivity



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

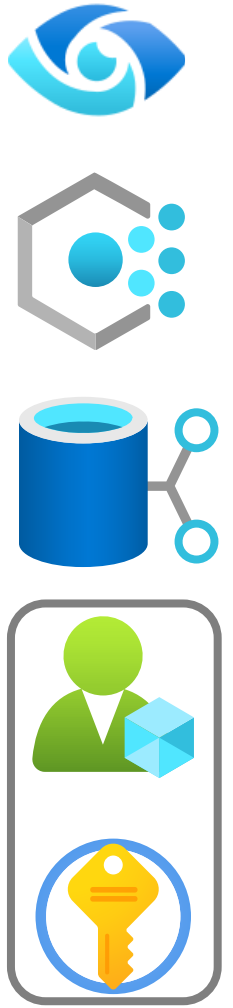
Data Source

Push or Pull

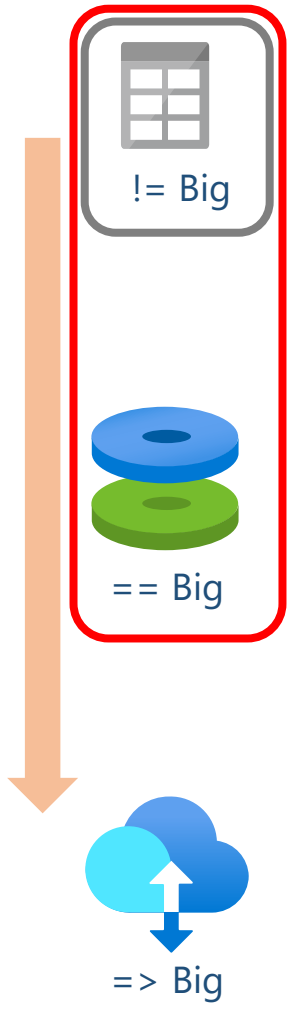
Batch or Speed

Public or Private Transfer

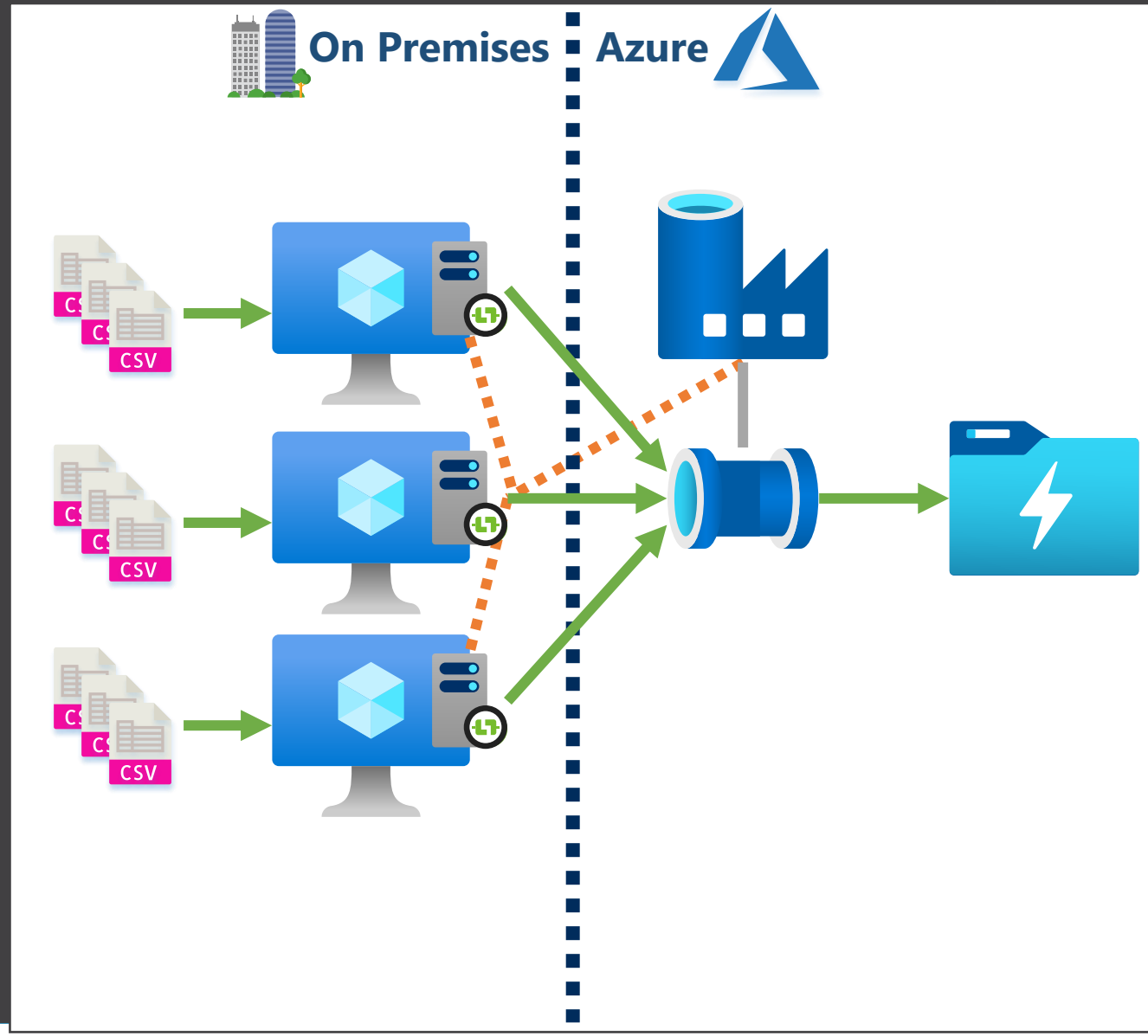
Data Sensitivity



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

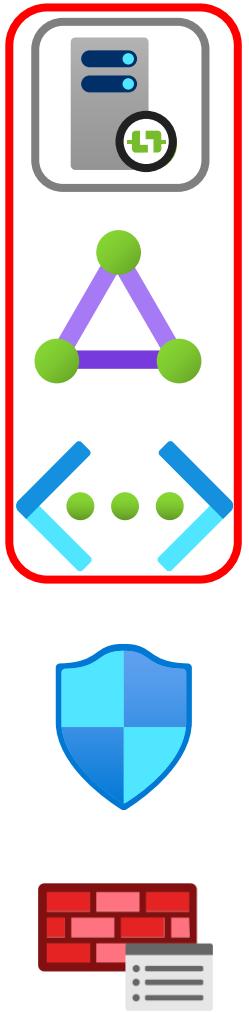
Data Structure

Data Source

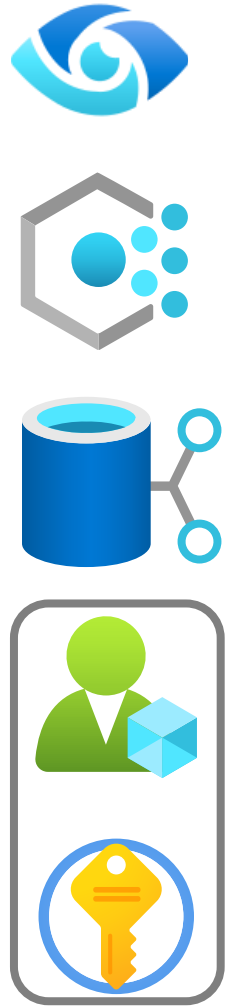
Push or Pull

Batch or Speed

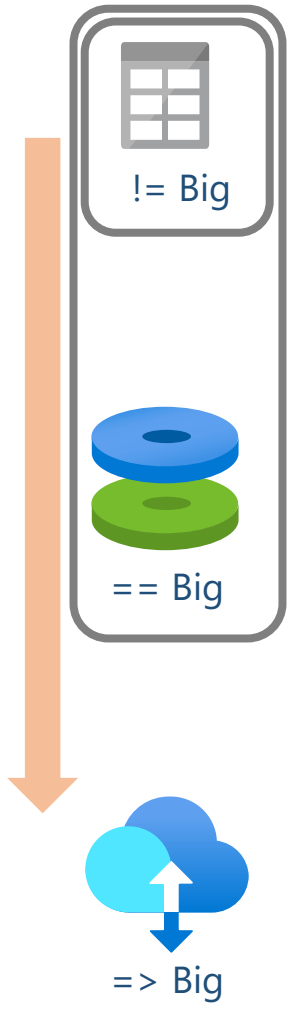
Public or Private Transfer



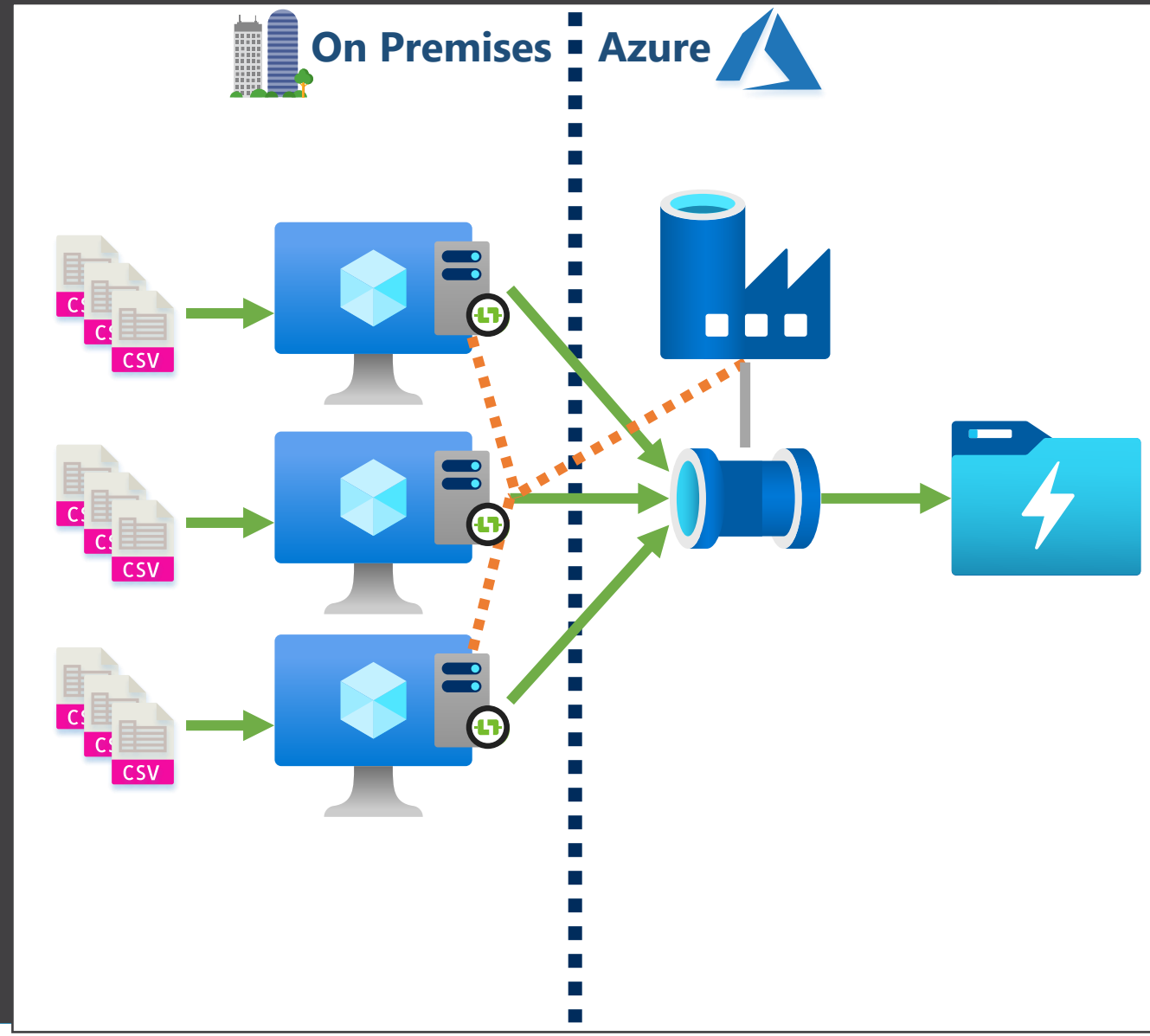
Data Sensitivity



Data Volume



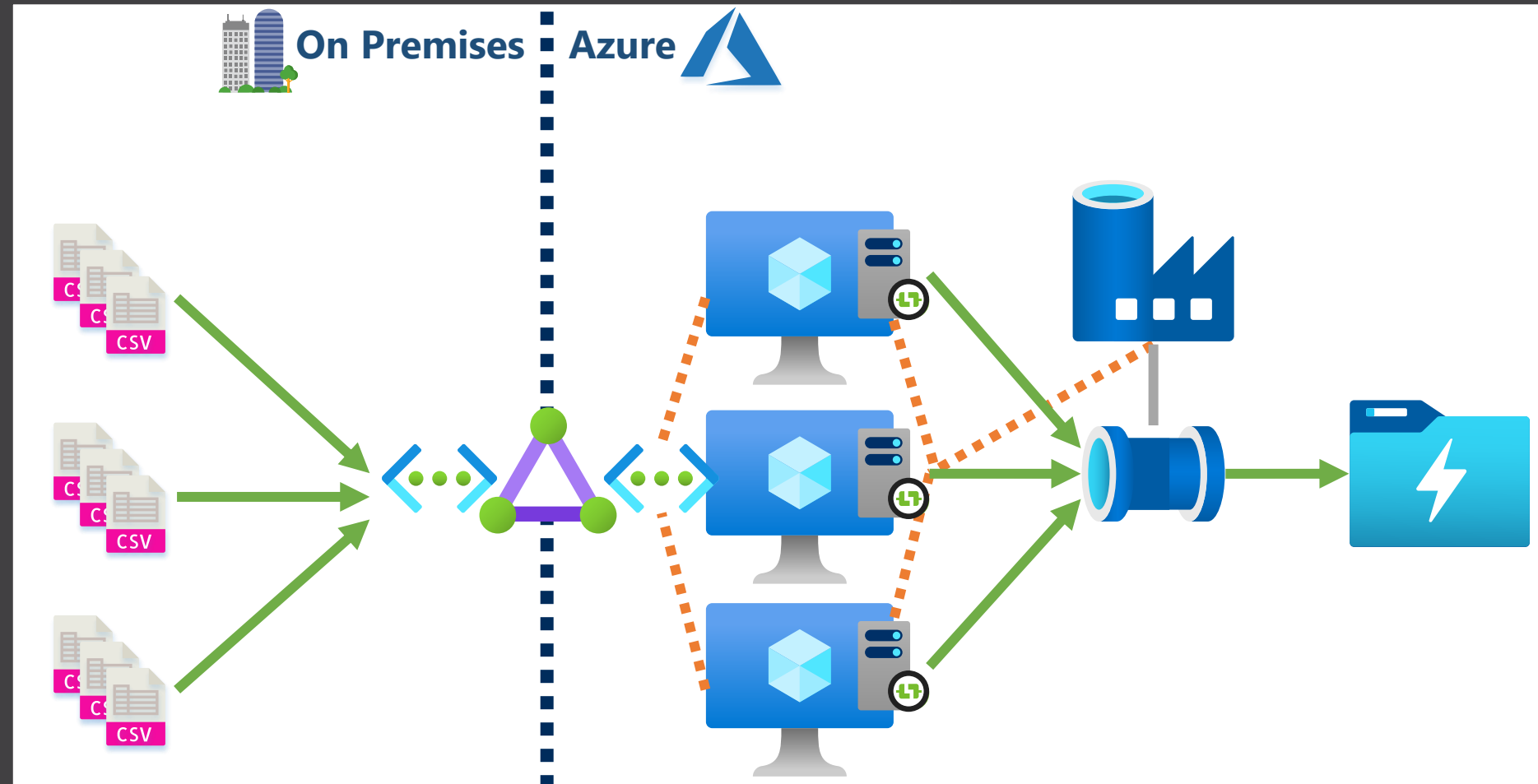
Data Extraction & Ingestion – Solution 3



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

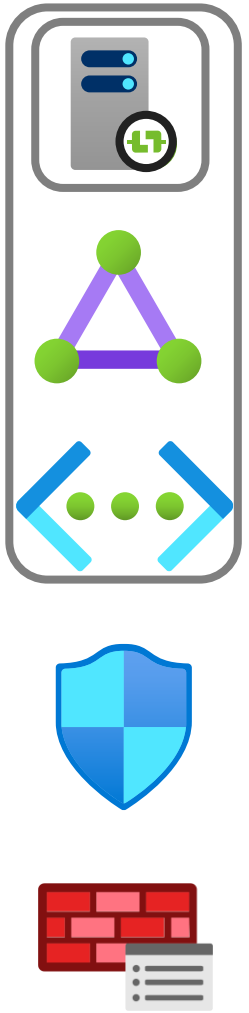
Data Structure

Data Source

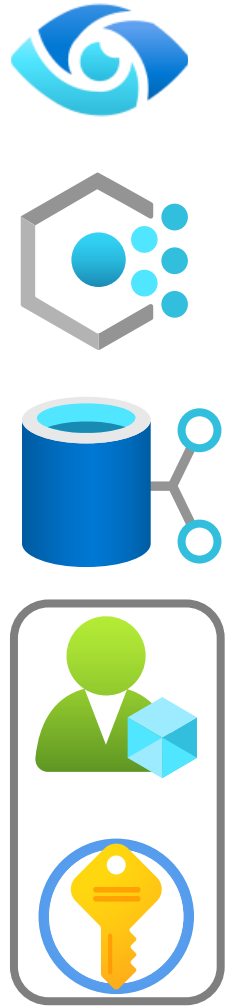
Push or Pull

Batch or Speed

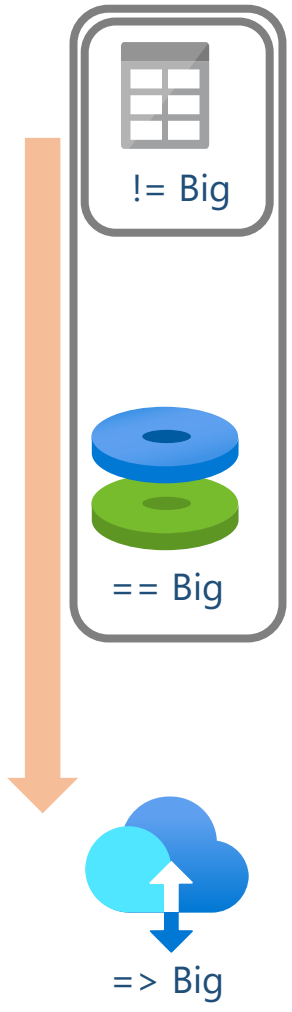
Public or Private Transfer



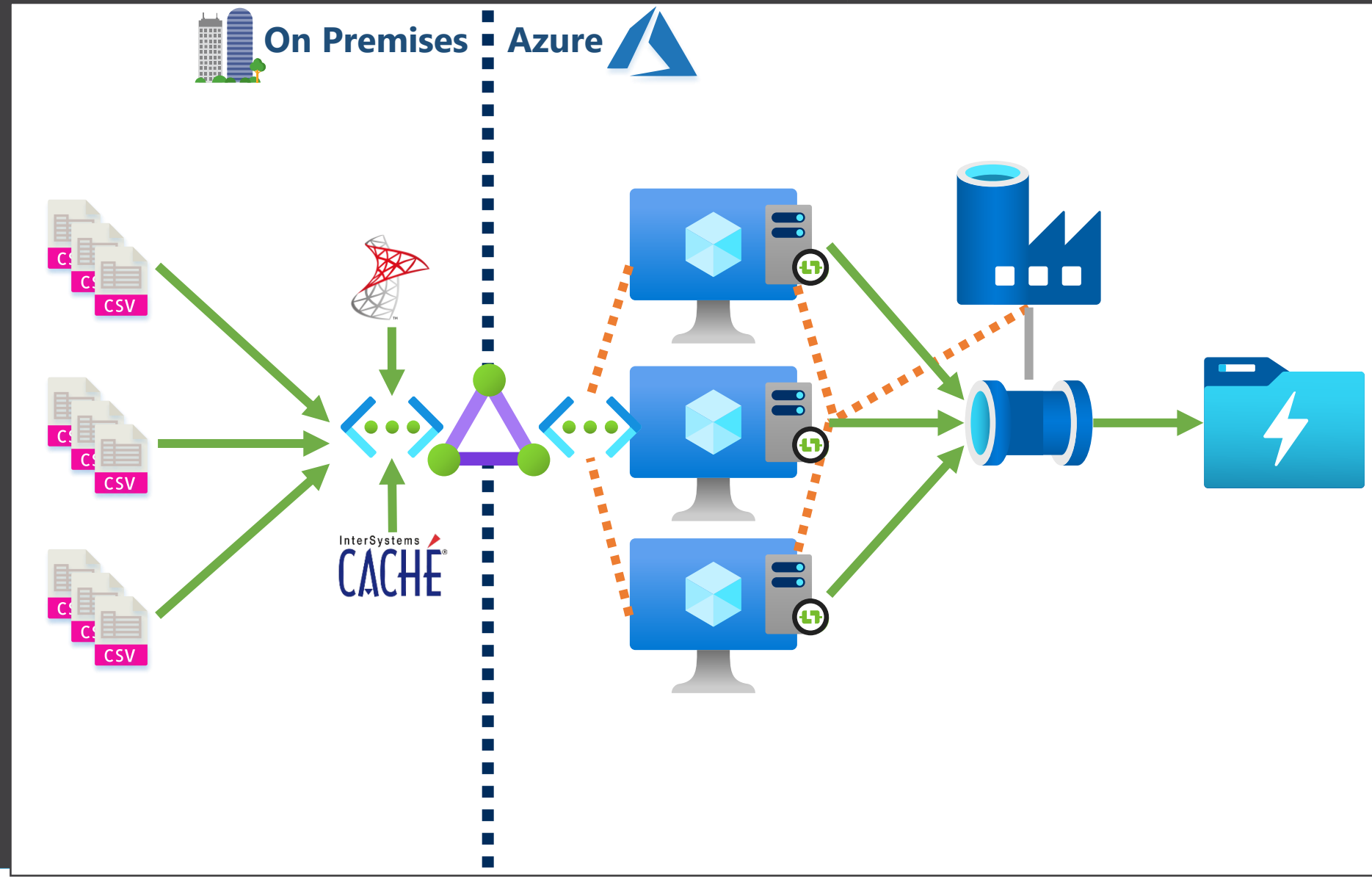
Data Sensitivity



Data Volume



Data Extraction & Ingestion – Solution 4

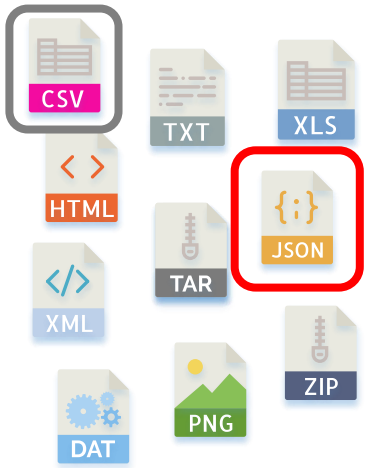


Requirements:

- 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

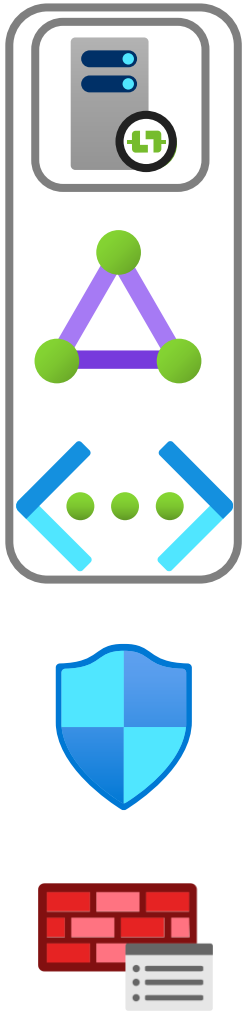


Data Source

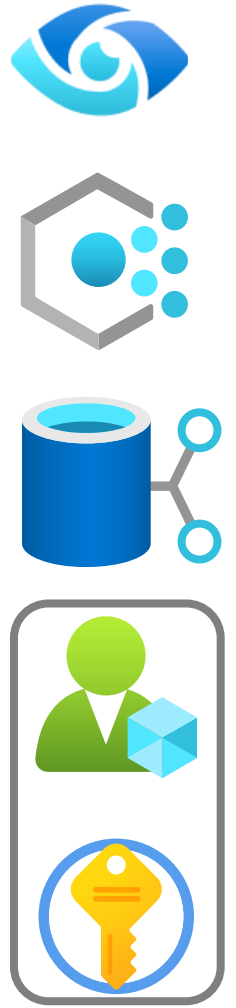
Push or Pull

Batch or Speed

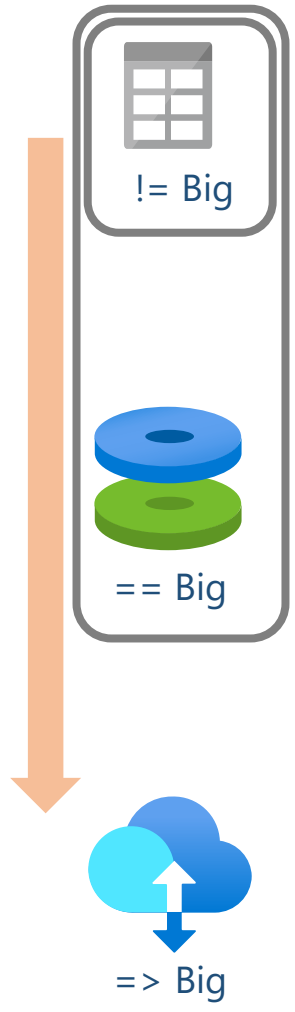
Public or Private Transfer



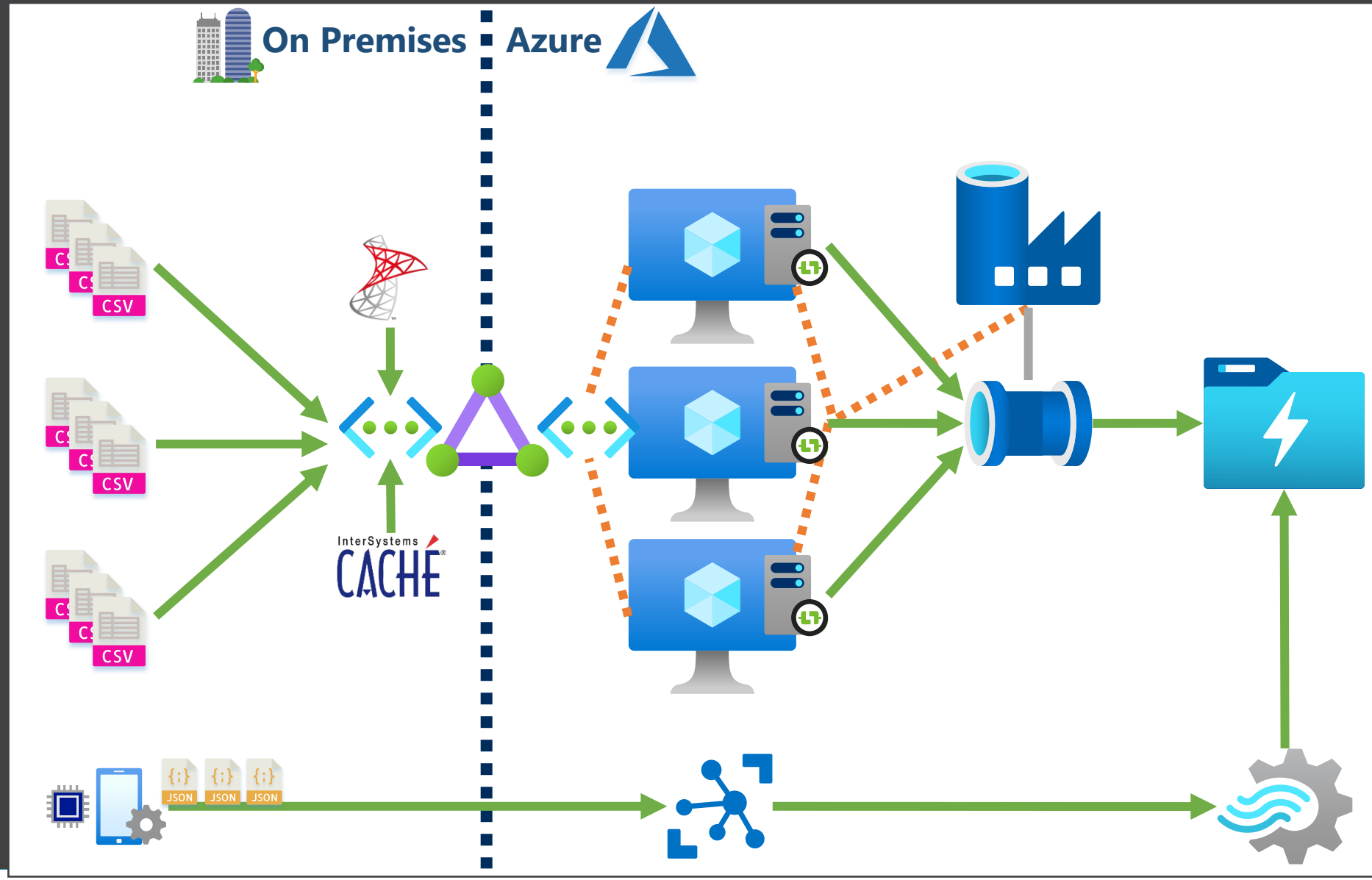
Data Sensitivity



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

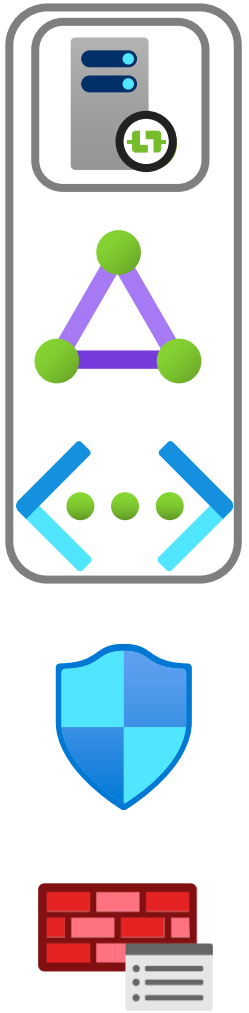
Data Structure

Data Source

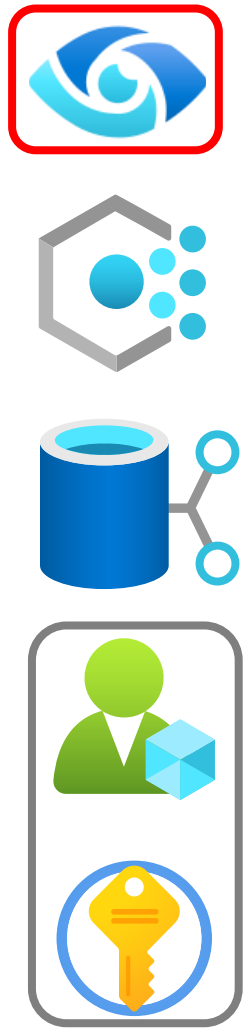
Push or Pull

Batch or Speed

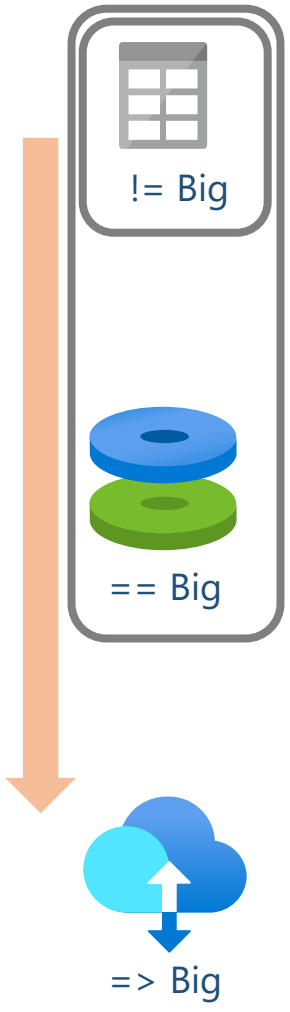
Public or Private Transfer



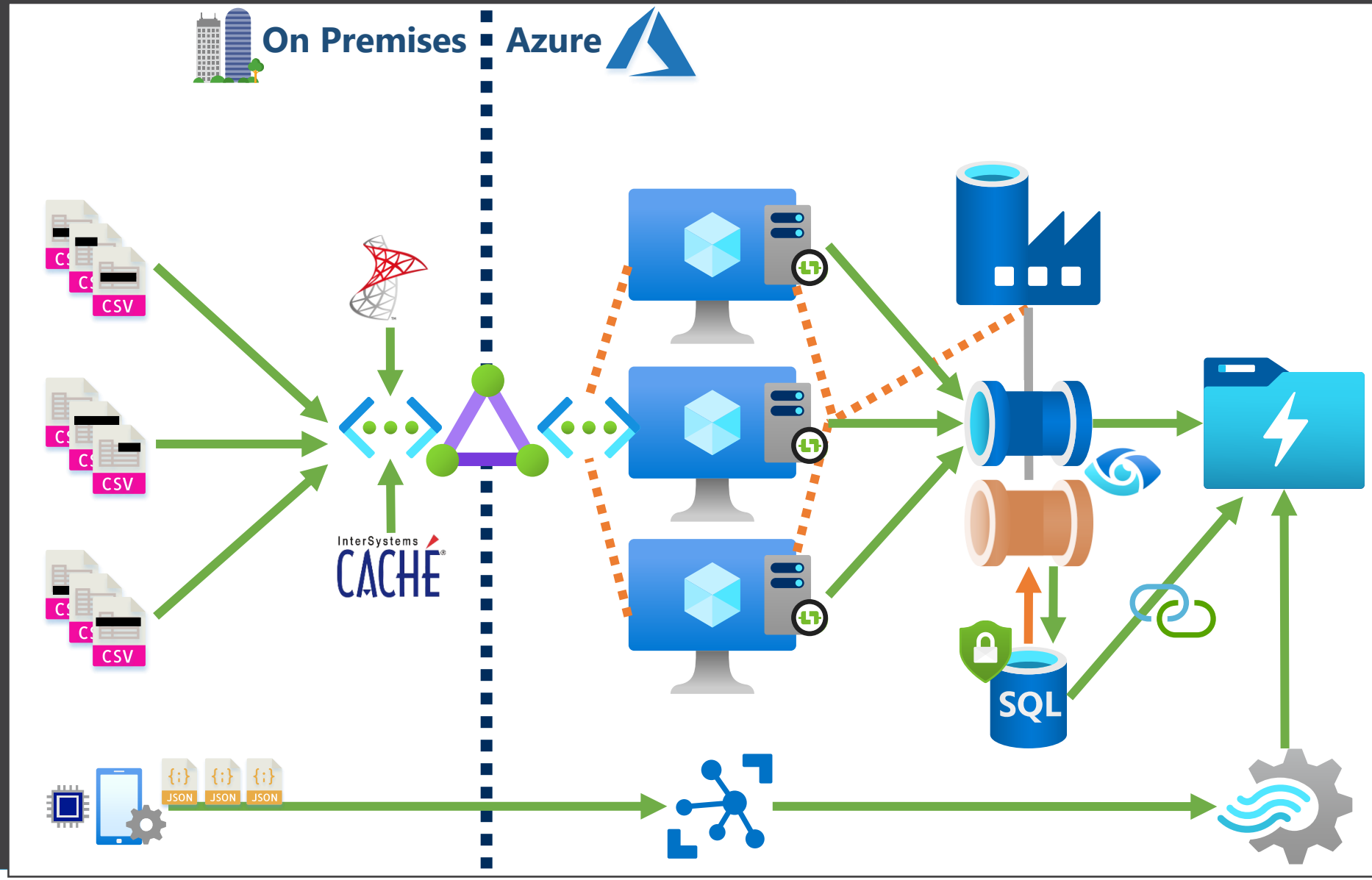
Data Sensitivity



Data Volume



Data Extraction & Ingestion – Solution 6

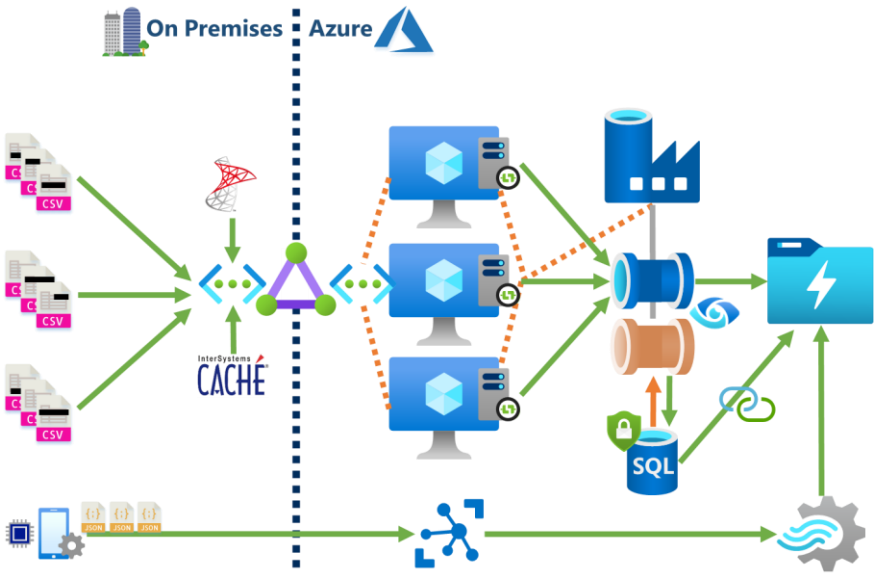


Requirements:

- 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

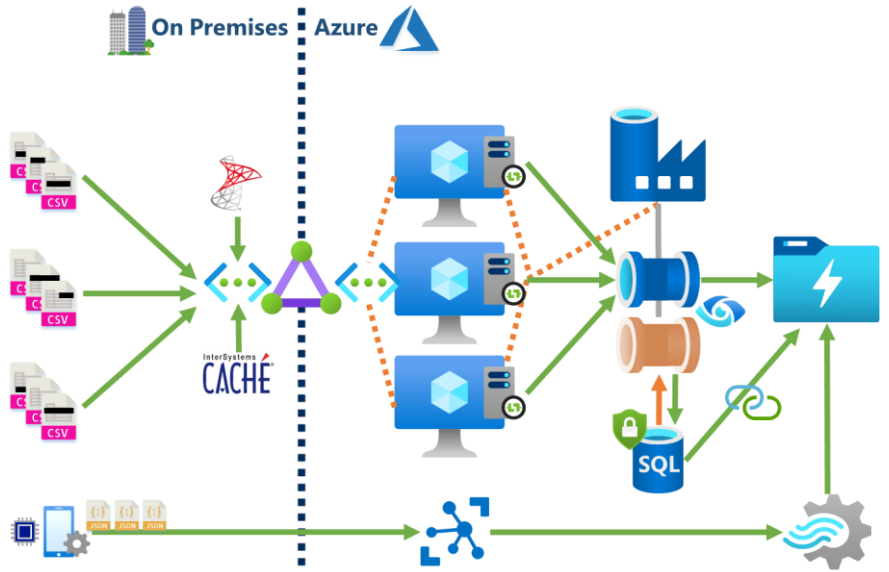
Overall Architecture

Extract | Transform | Load



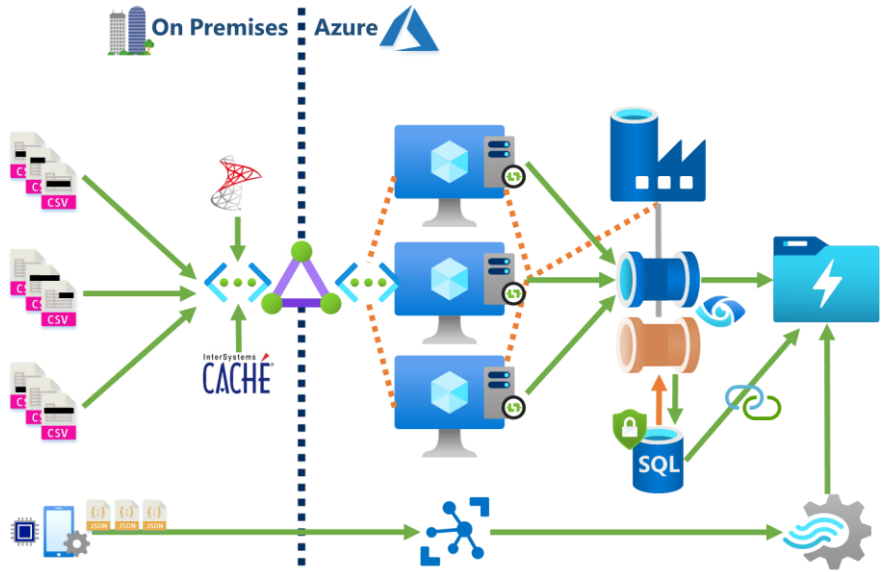
Agenda

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

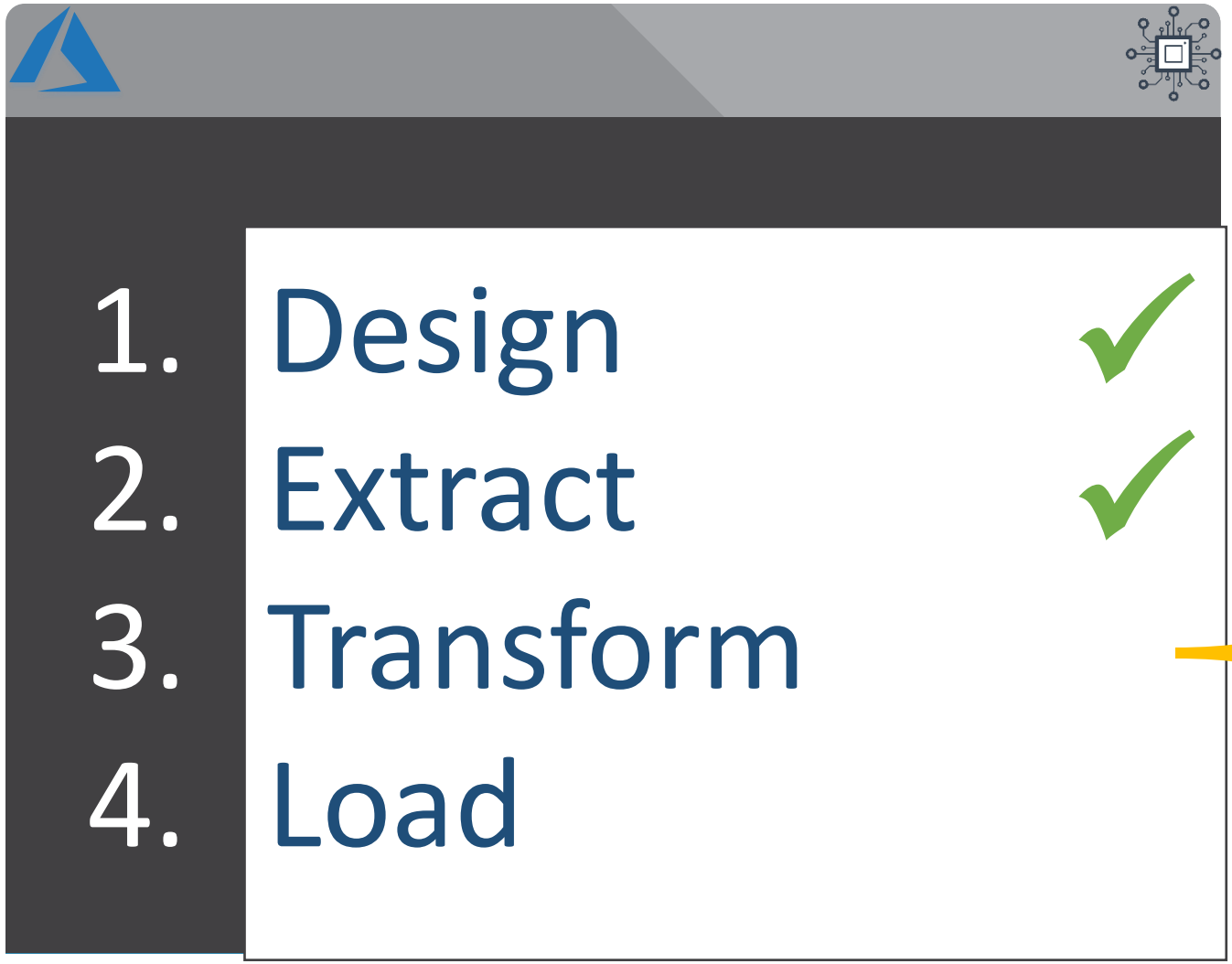


Agenda

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



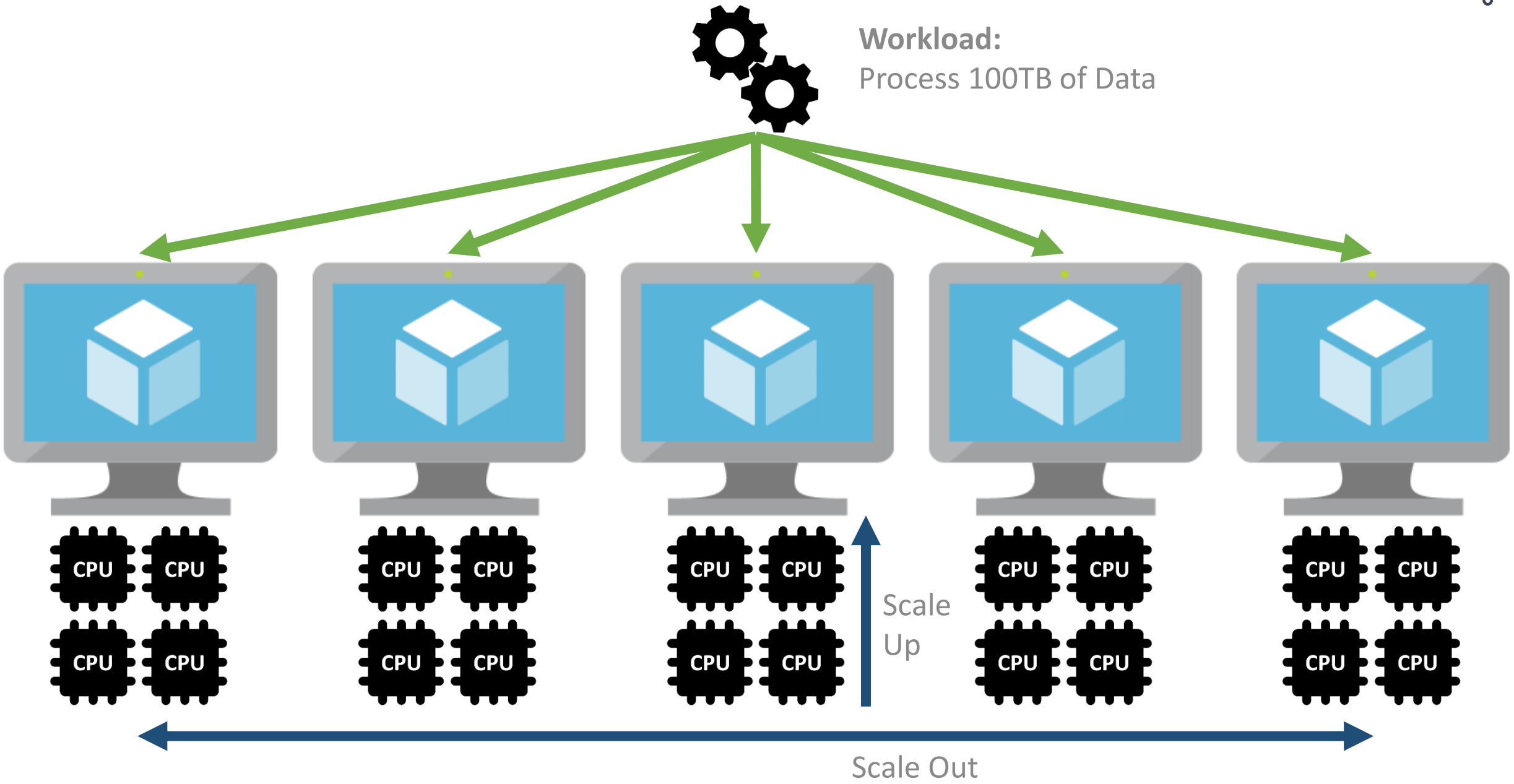
Agenda



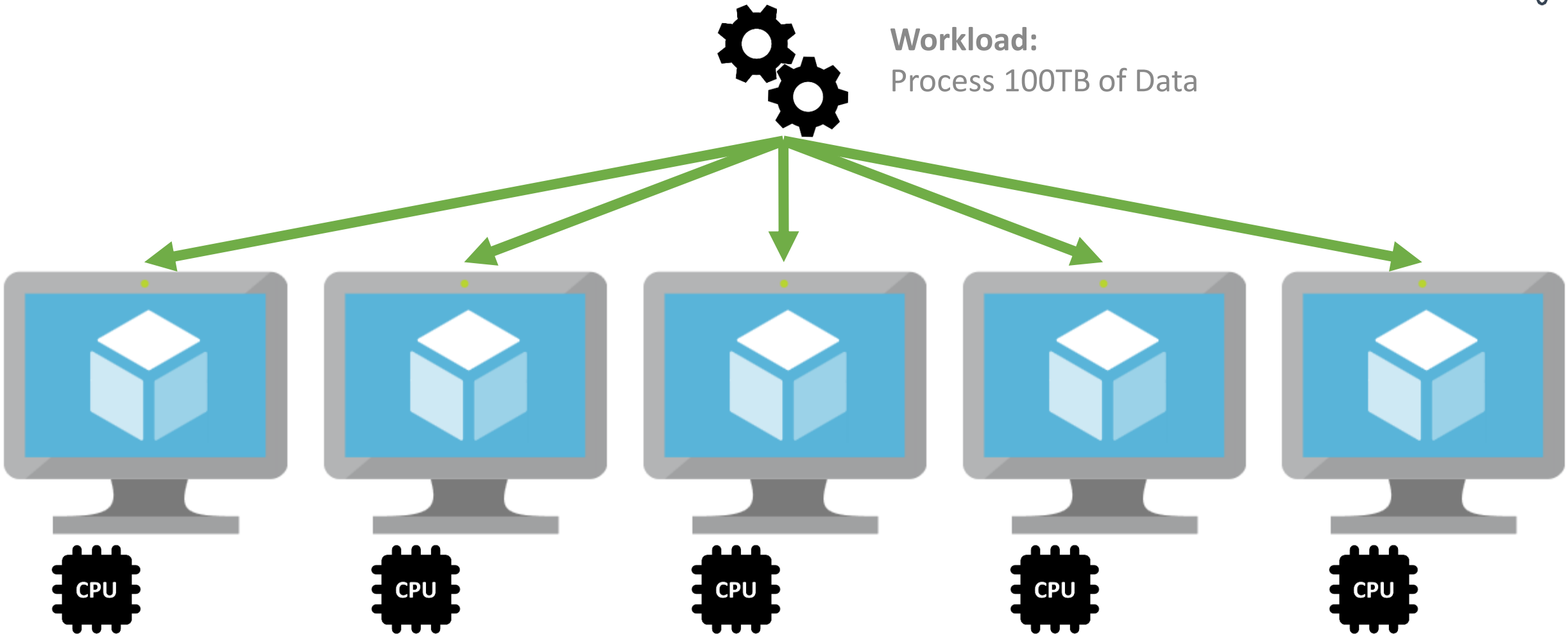
1. Design ✓
2. Extract ✓
3. Transform
4. Load

Compute
Storage, Structure
& Data Format

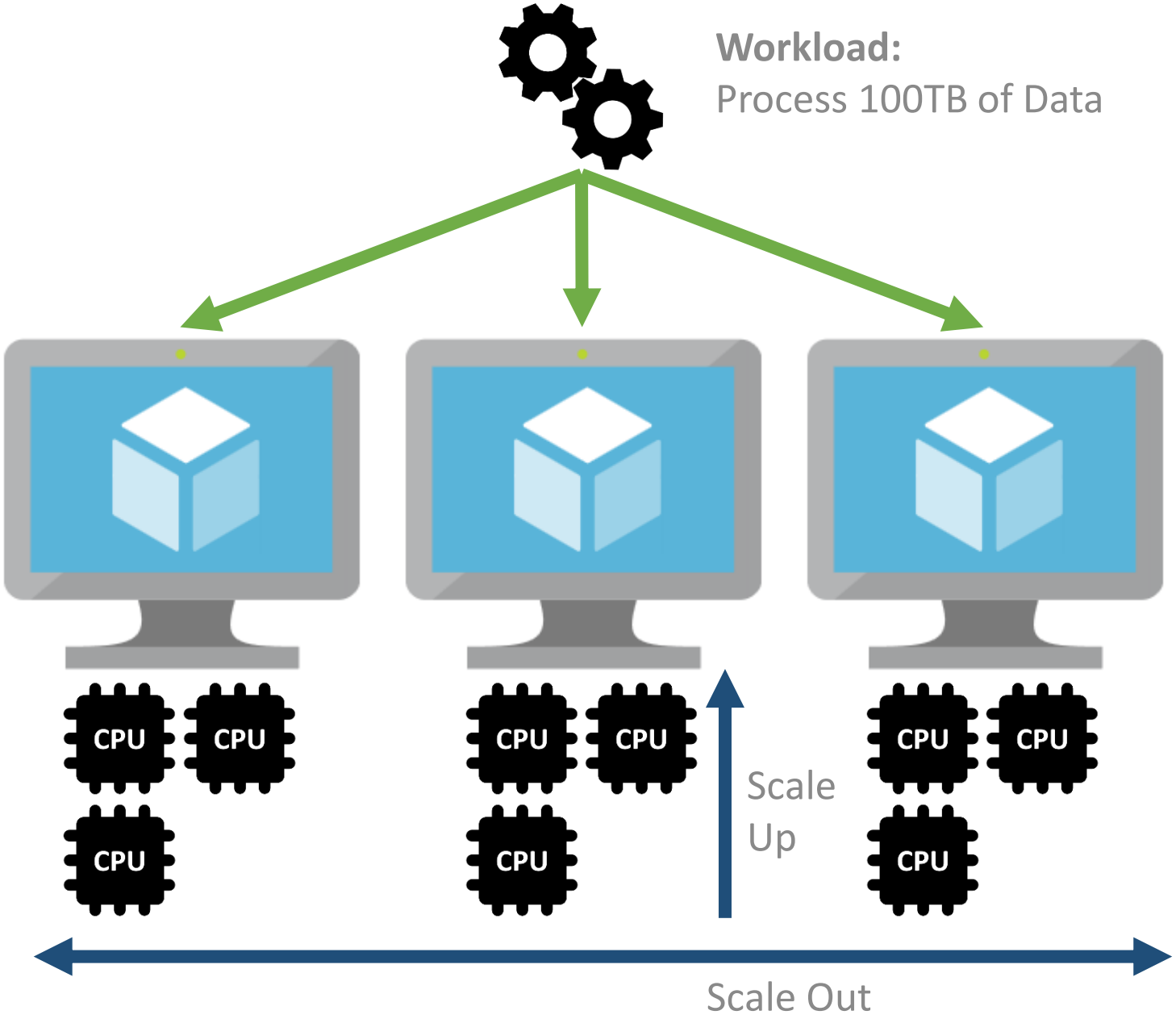
Scaling Up and/or Scaling Out



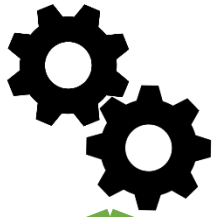
Scaling Up and/or Scaling Out



Scaling Up and/or Scaling Out



What Compute Type of Compute?



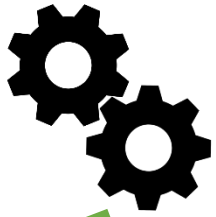
Workload:
Process 100TB of Data

Platform

Infrastructure

As
A
Service

What Compute Type of Compute?



Platform

As
A
Service

Data Transformation – Compute

Data Lake Analytics

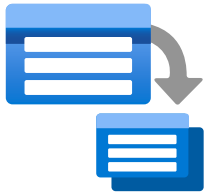
HDInsight

Relational Database

Synapse – SQL Pools or Spark Pools

Databricks

Batch Service



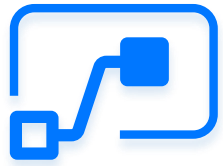
Data Explorer

Automation

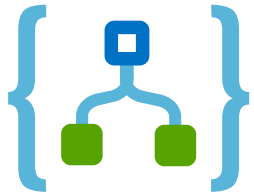
Cosmos

Functions

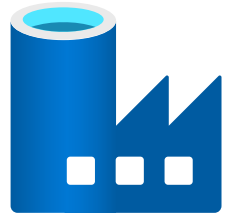
Power BI Data Flows



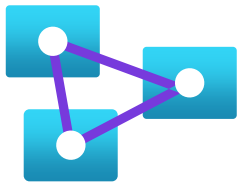
Logic Apps



Data Factory Data Flows



Analysis Services



Data Transformation – Compute

Data Lake Analytics

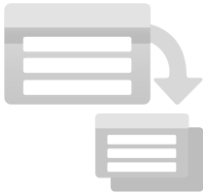
HDInsight

Relational Database

Synapse – SQL Pools or Spark Pools

Databricks

Batch Service



Data Explorer

Automation

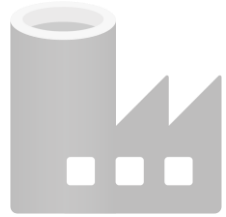
Cosmos

Functions

Power BI Data Flows

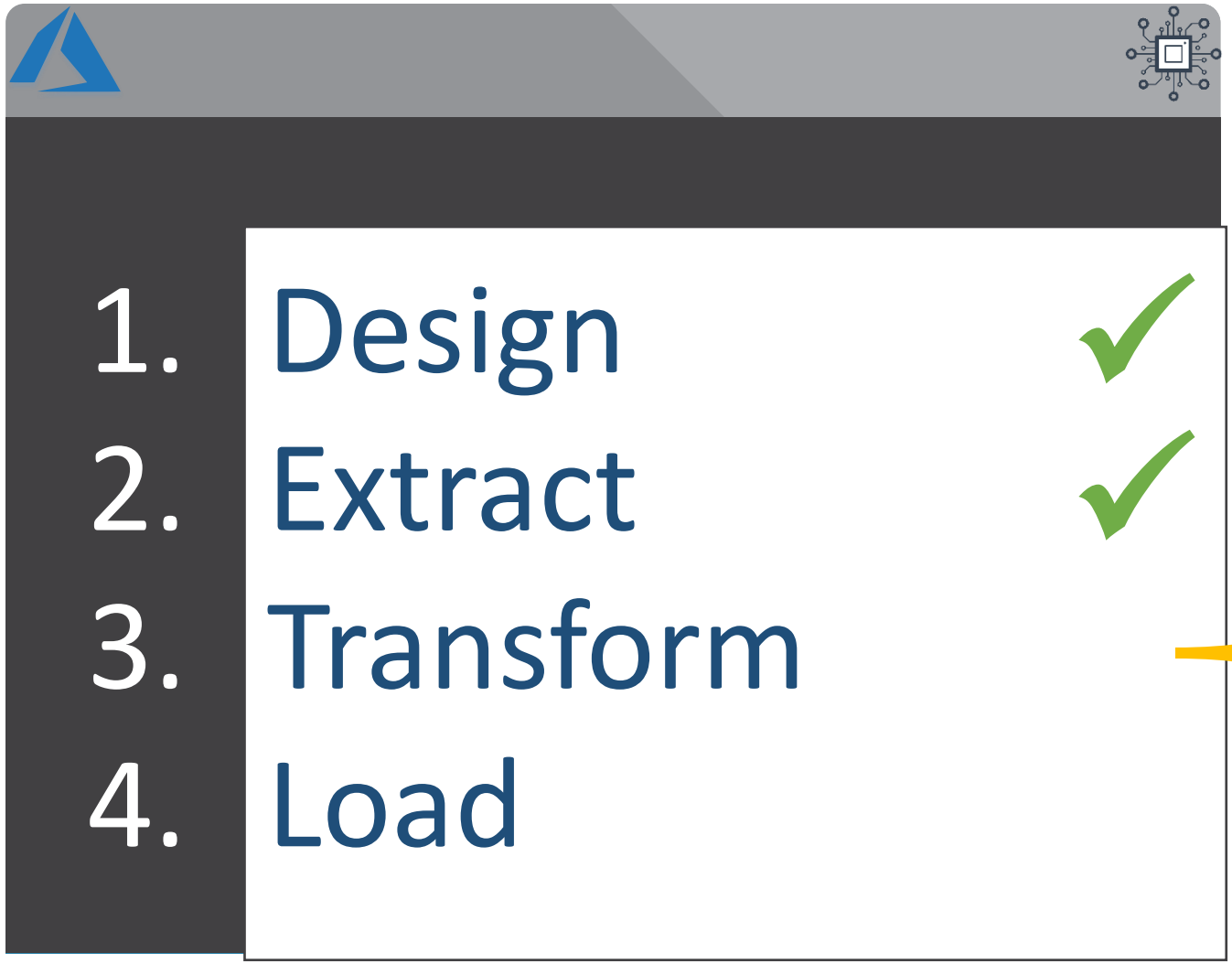
Logic Apps

Data Factory Data Flows



Analysis Services

Agenda



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

Compute ✓
Storage, Structure & Data Format

Data Transformation – Storage & Format

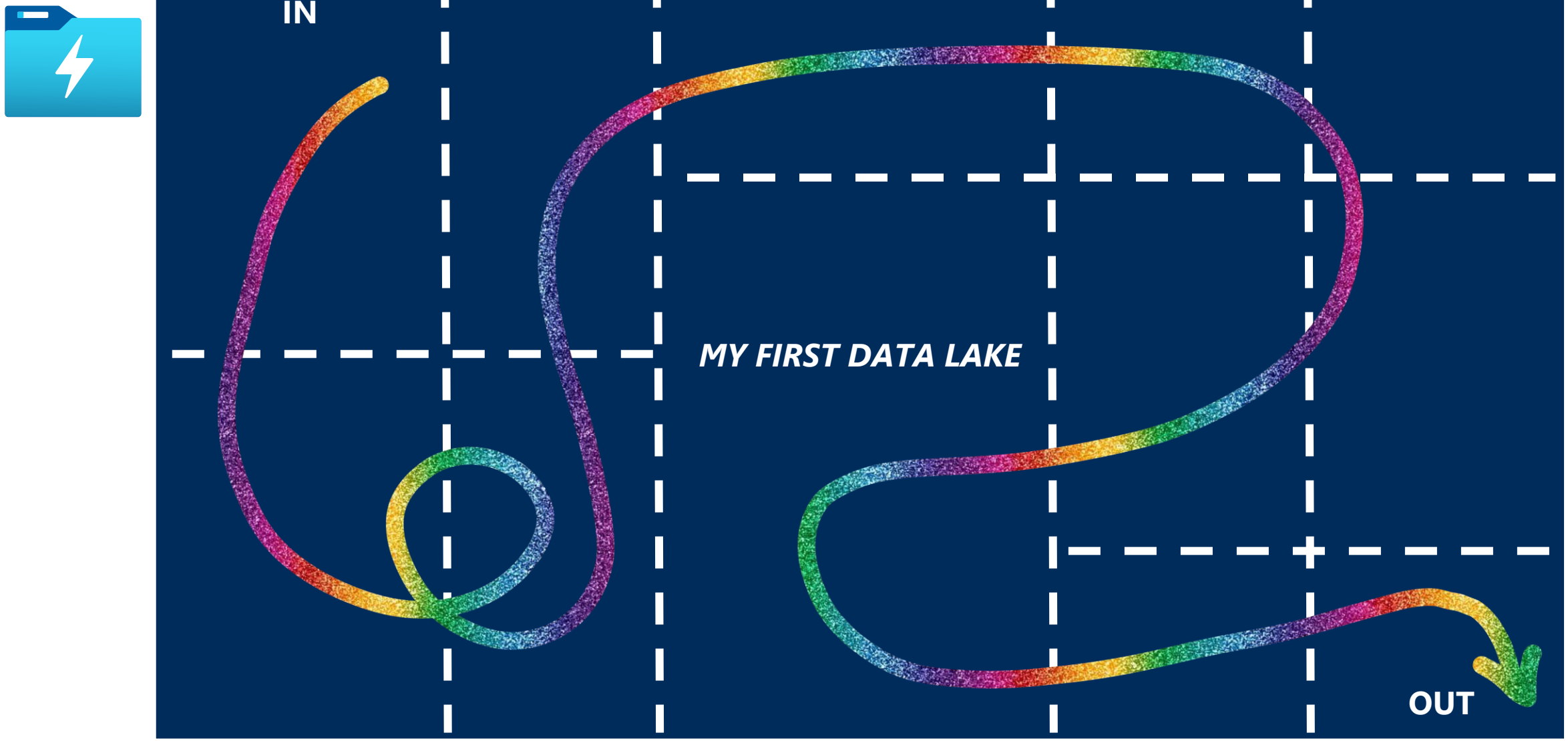
Azure Storage Account

Azure Data Lake Gen2

Hadoop Distributed File System (HDFS)

Data Transformation – Storage & Format

Data Transformation – Storage & Format



Data Transformation – Storage & Format

LIBRARIES

AD-HOC

MODELS

STREAMING

LANDING

RAW

CLEANED

...\Year=2020\
Month=01\
Day=01\

ALIGNED

Append
OR
Override

CONFORMED

Facts
&
Dimensions

PREPARED

ARCHIVE

CURATION

Data Transformation – Storage & Format

LIBRARIES

AD-HOC

MODELS

STREAMING

LANDING

RAW

CLEANED

ALIGNED

CONFORMED

PREPARED

ARCHIVE

CURATION

Data Transformation – Storage & Format

LIBRARIES

AD-HOC

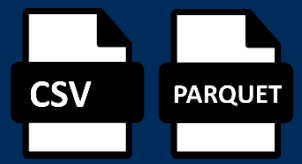
MODELS

STREAMING

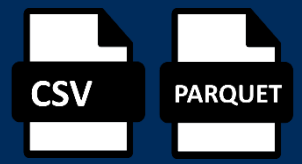
LANDING

RAW

CLEANED



ALIGNED



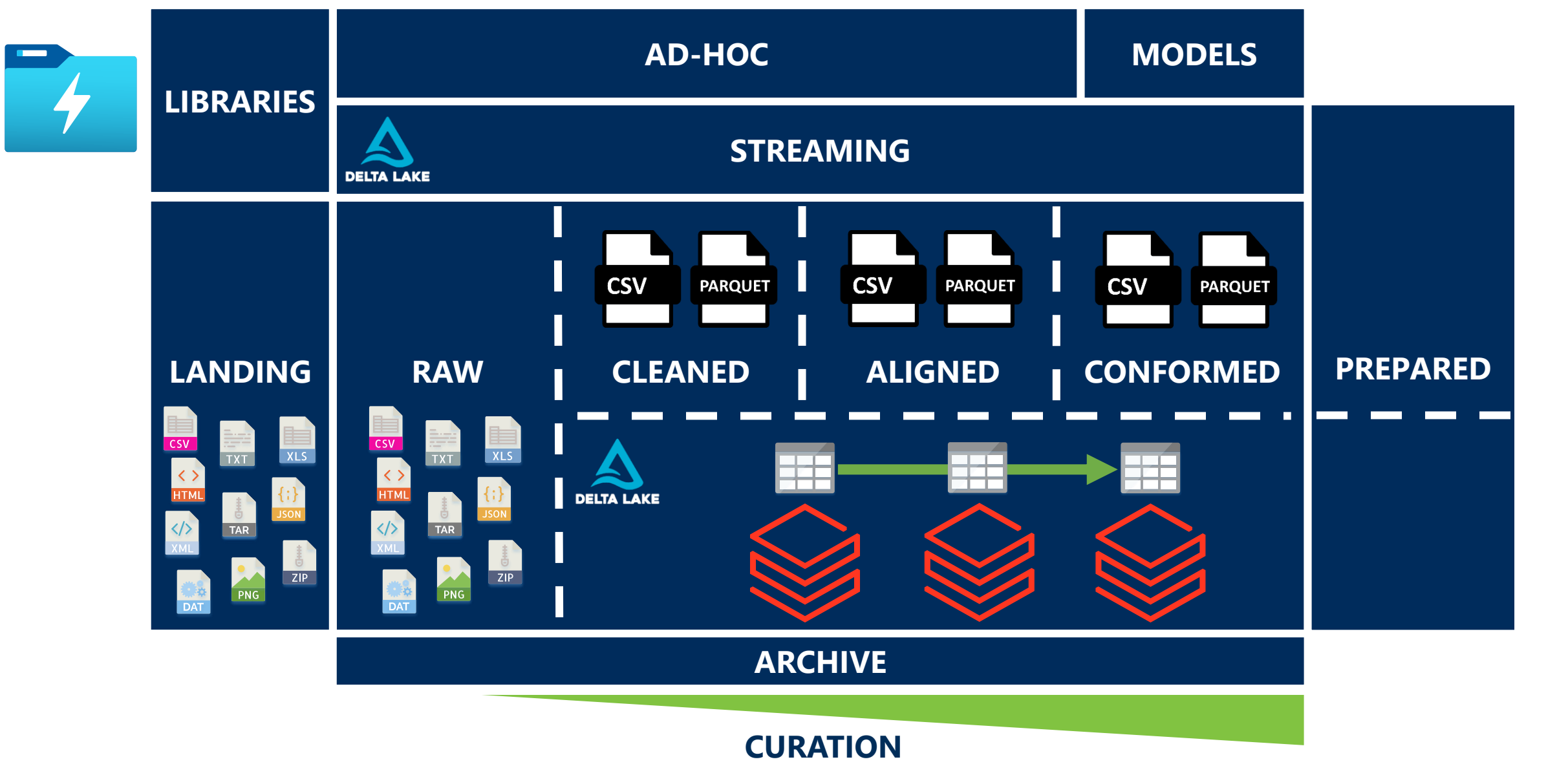
CONFORMED

PREPARED

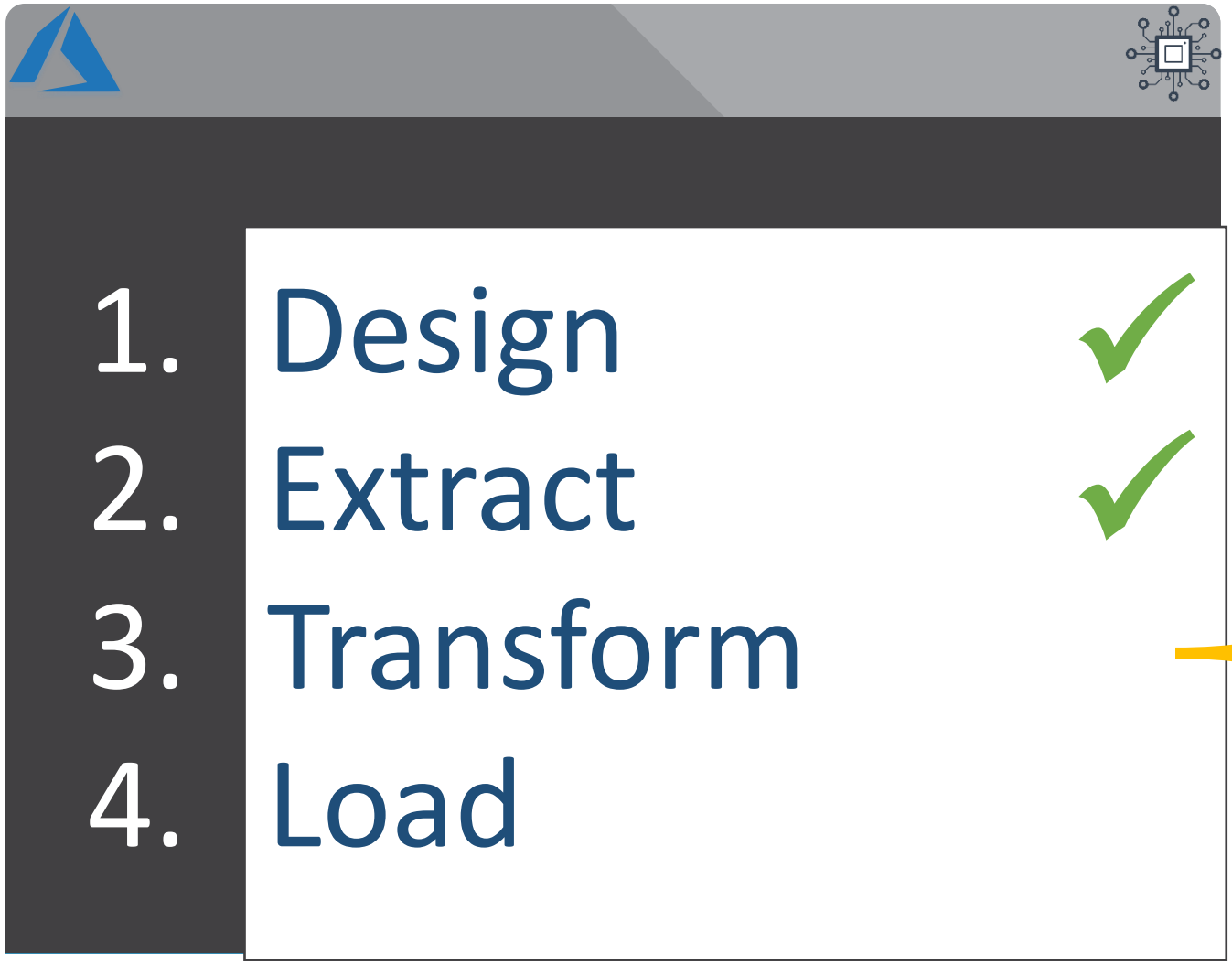
ARCHIVE

CURATION

Data Transformation – Storage & Format



Agenda

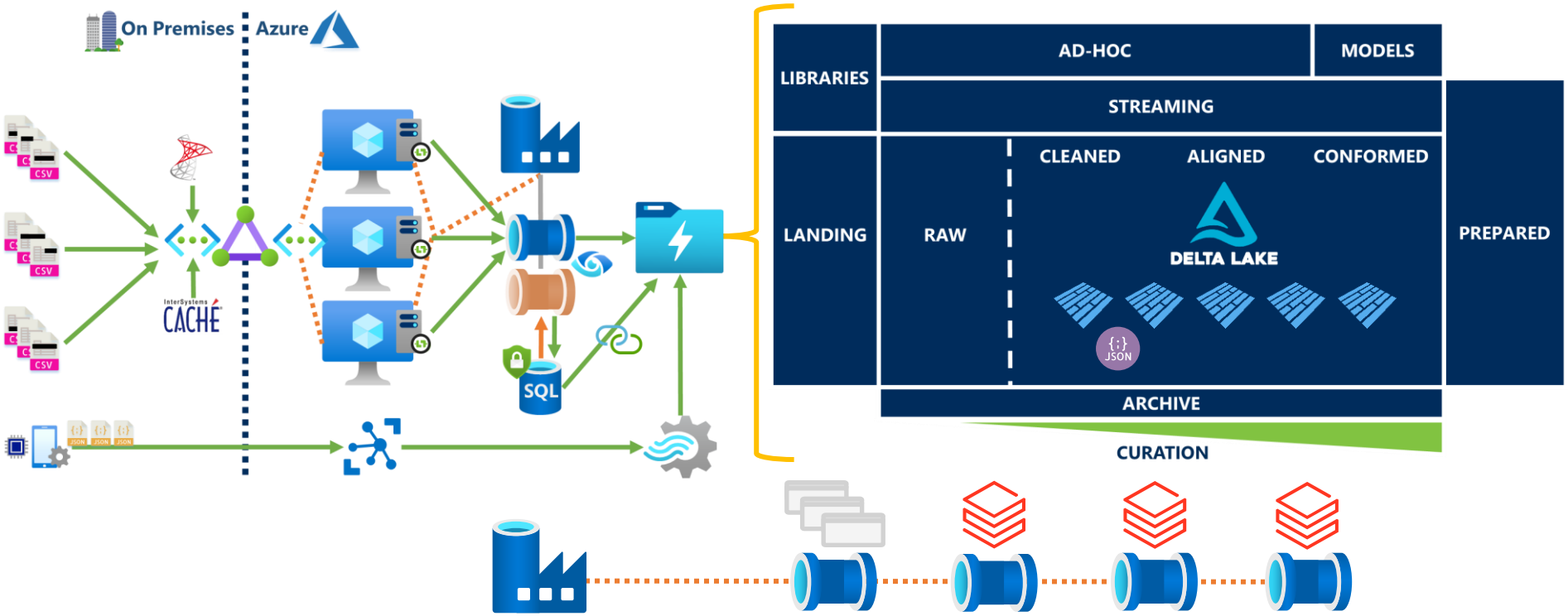


1. Design ✓
2. Extract ✓
3. Transform
4. Load

Compute ✓
Storage, Structure & Data Format ✓

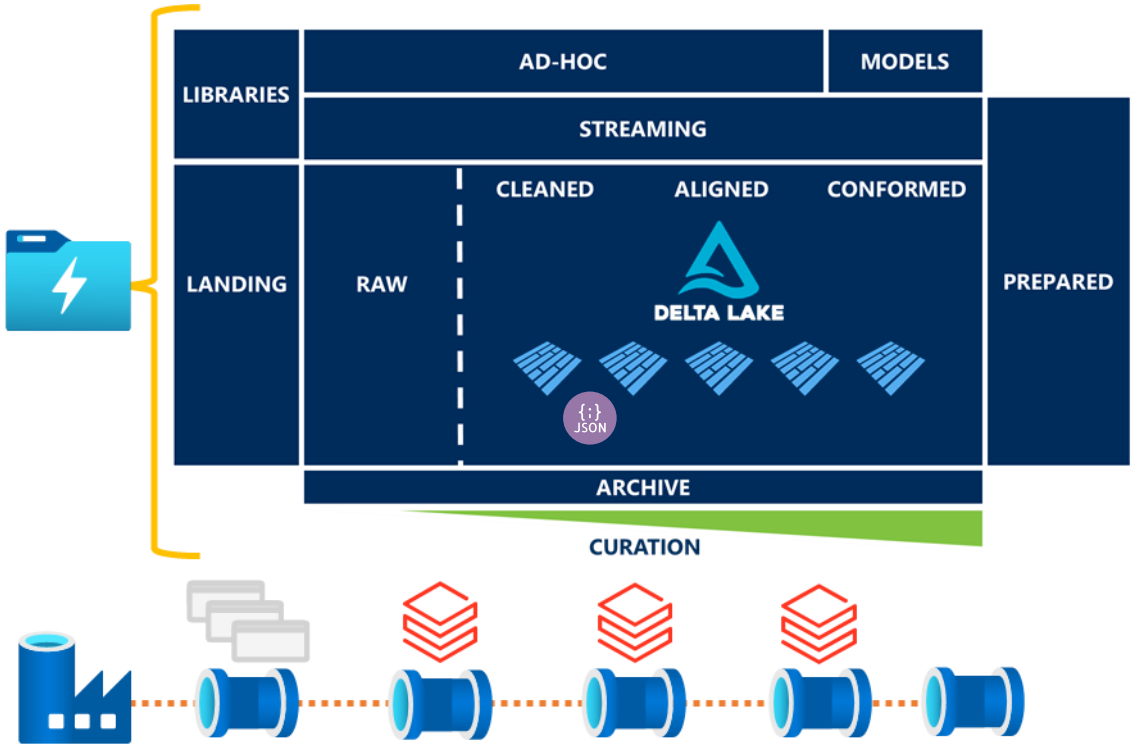
Overall Architecture

Extract | Transform | Load



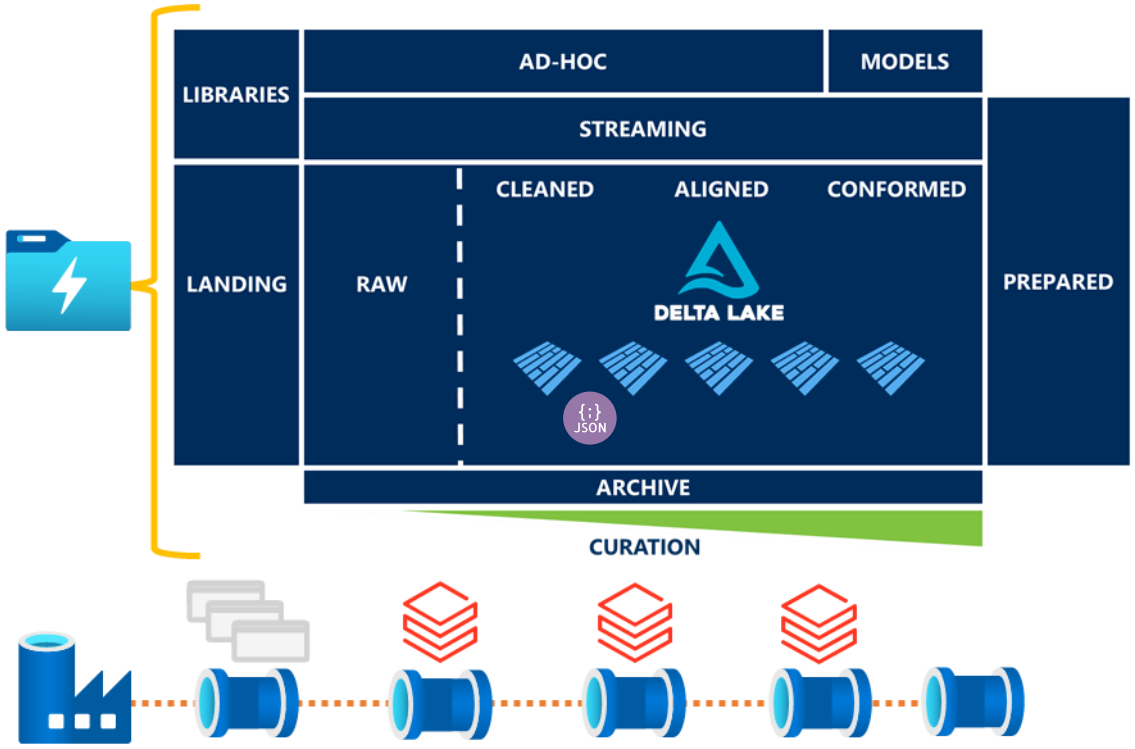
Agenda

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

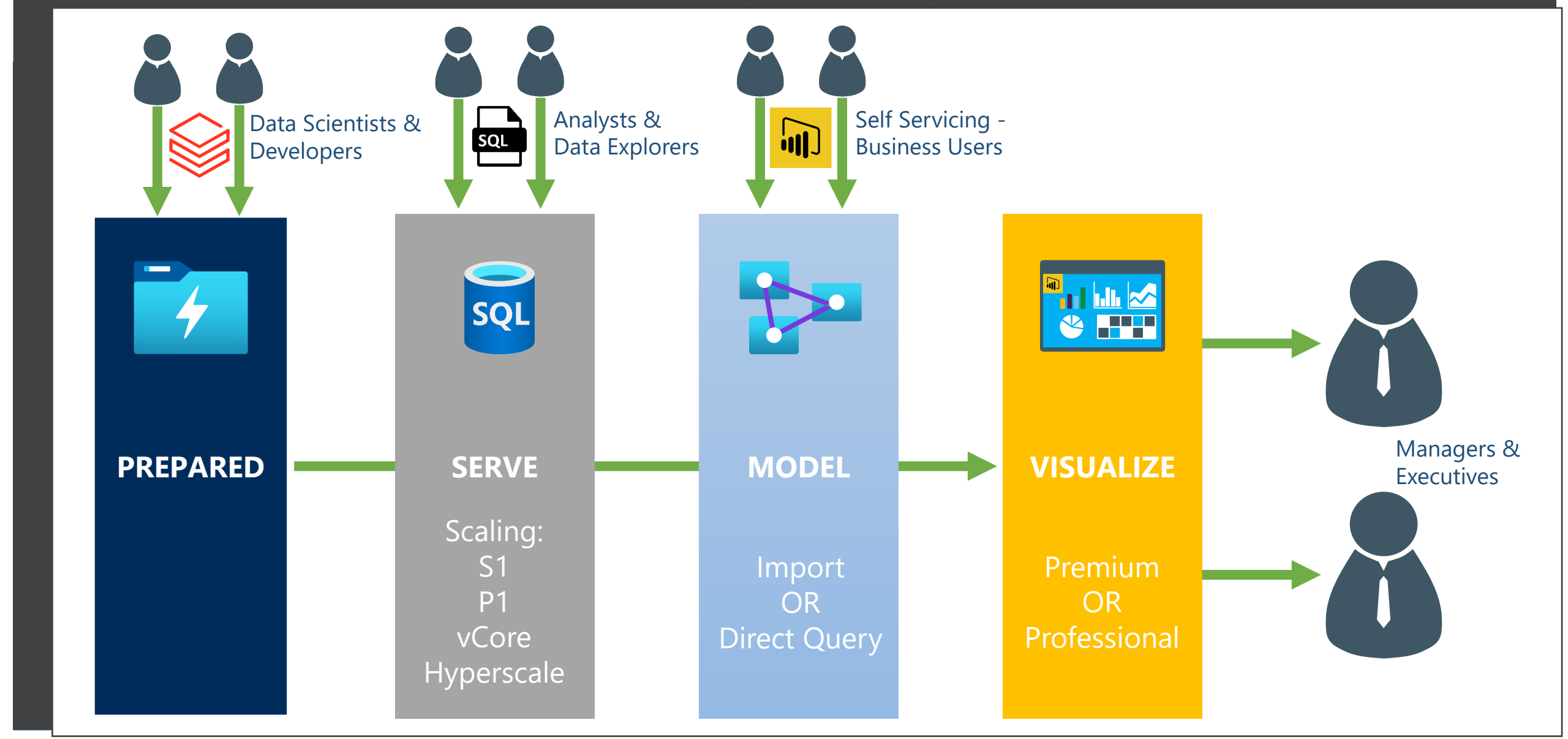


Agenda

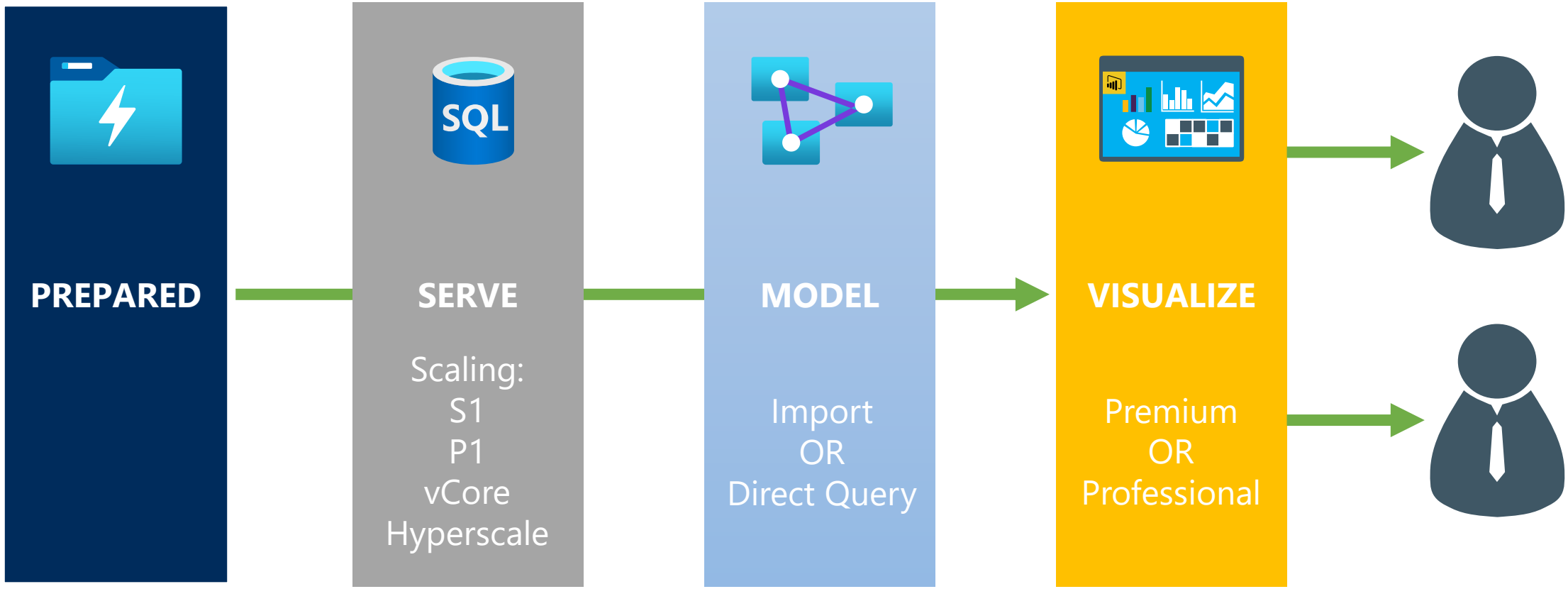
1. Design ✓
2. Extract ✓
3. Transform ✓
4. **Load**



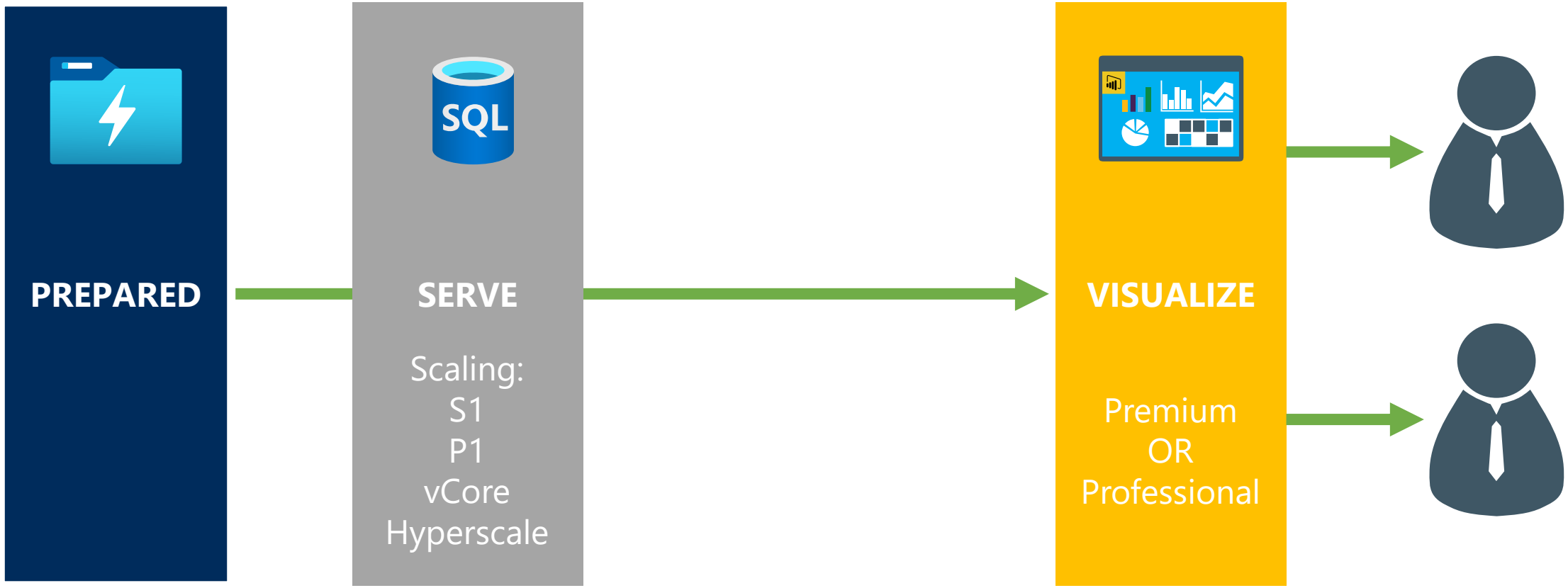
Loading & Consuming Data



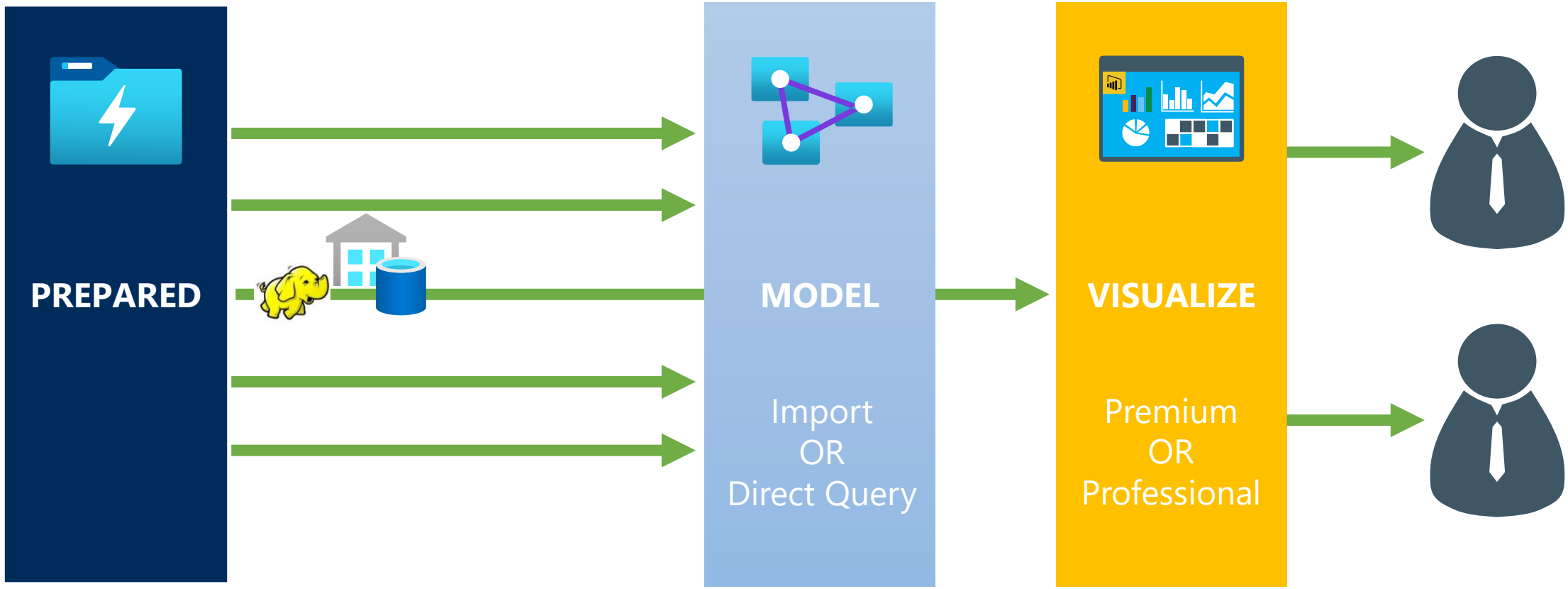
Loading & Consuming Data



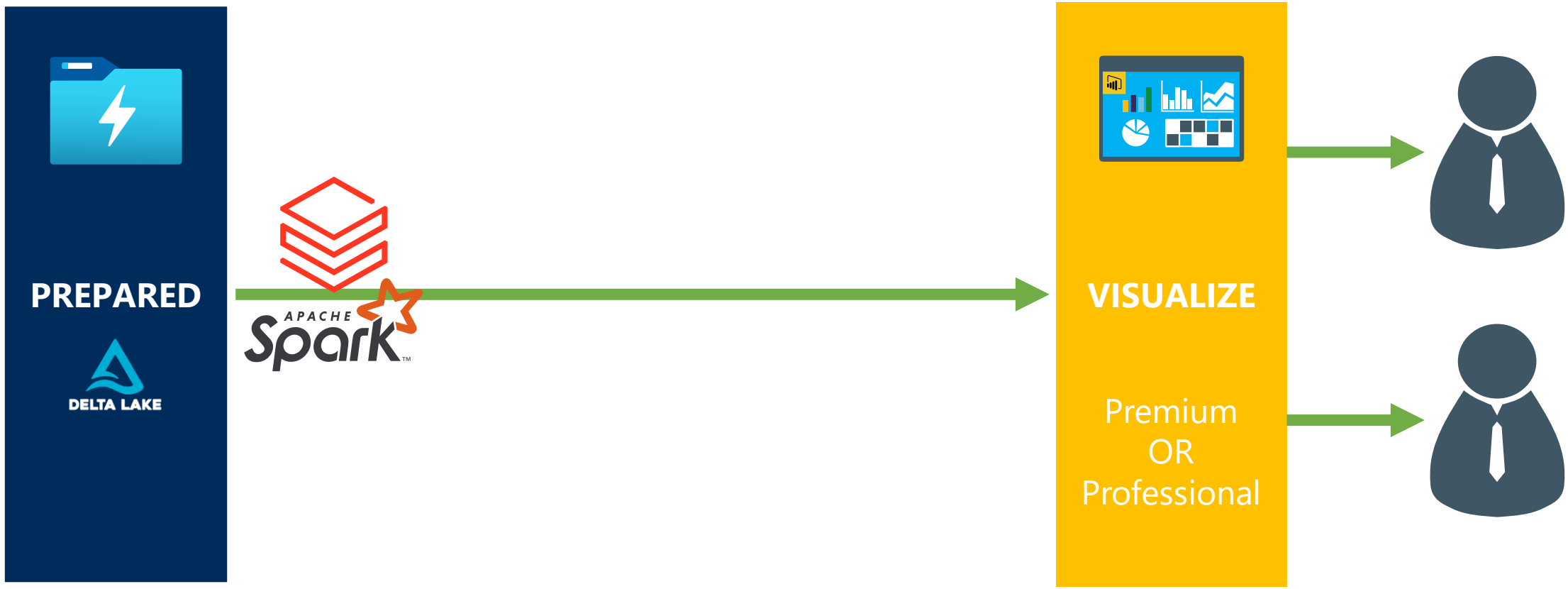
Loading & Consuming Data



Loading & Consuming Data

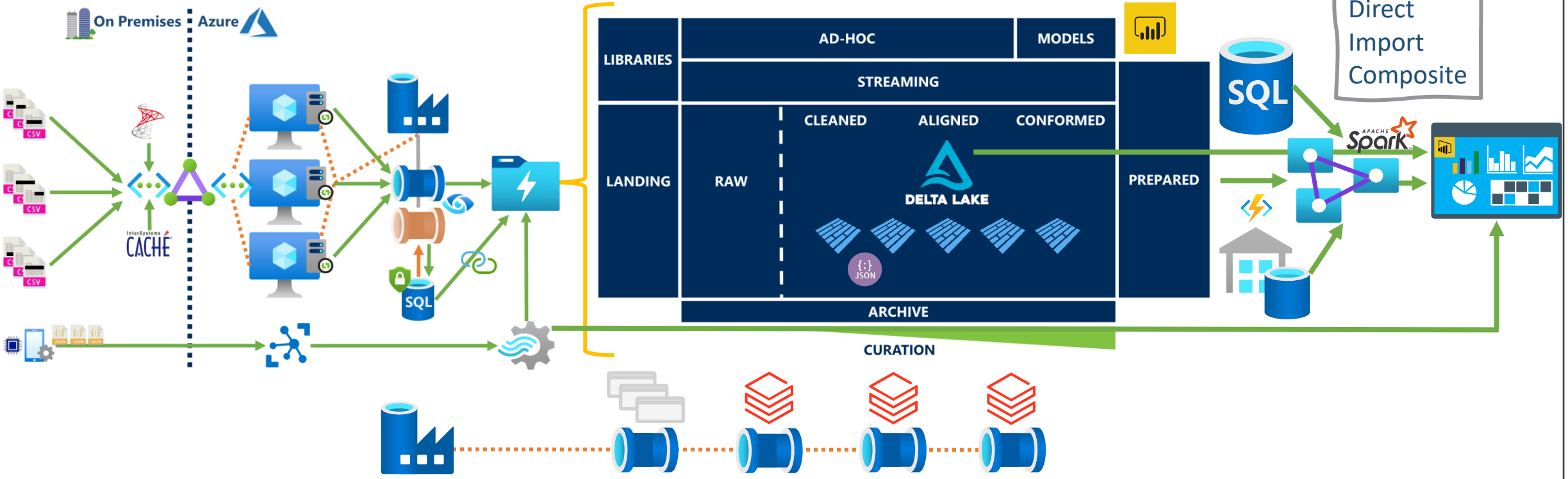


Loading & Consuming Data

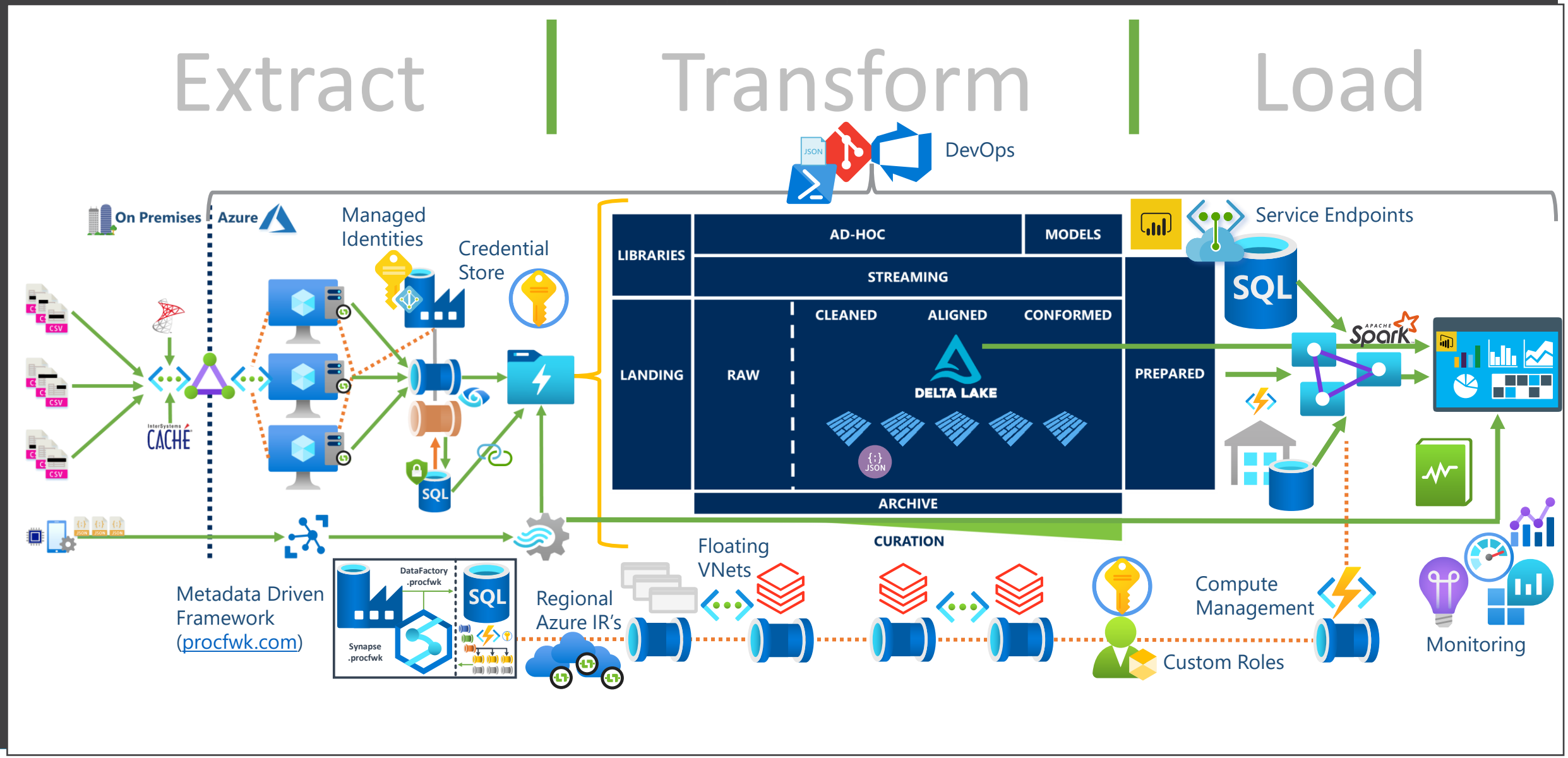


Overall Architecture

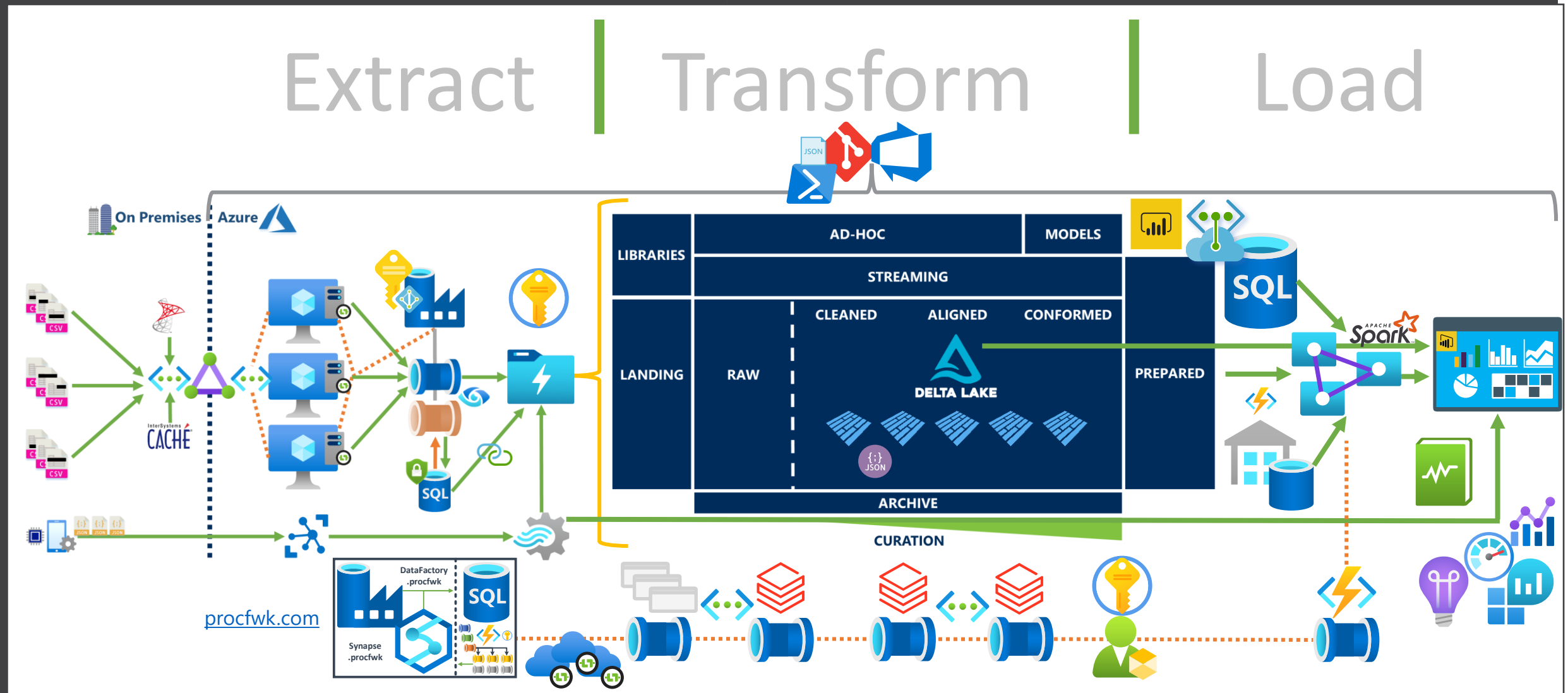
Extract | Transform | Load



Overall Architecture



Overall Architecture



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

Thank you for listening…

Paul Andrew

Microsoft® MVP Most Valuable Professional

avanade

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

Twitter: @mrpaulandrew
LinkedIn: In/mrpaulandrew

GitHub: github.com/mrpaulandrew