An Architects Guide to Delivering Data Insights
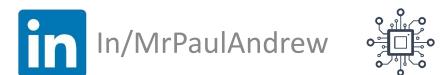
Using the Microsoft Azure Data Platform

Paul Andrew | Principal Consultant & Solution Architect

Microsoft MVP Most Valuable Professional

altius

@MrPaulAndrew   In/MrPaulAndrew   MrPaulAndrew.com

PASS SQLSATURDAY

Thank you
to our
Global
Sponsors

Microsoft Azure

DELL Technologies

IDERA

Quest

SentryOne
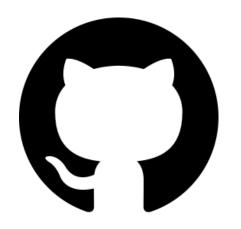
vmware

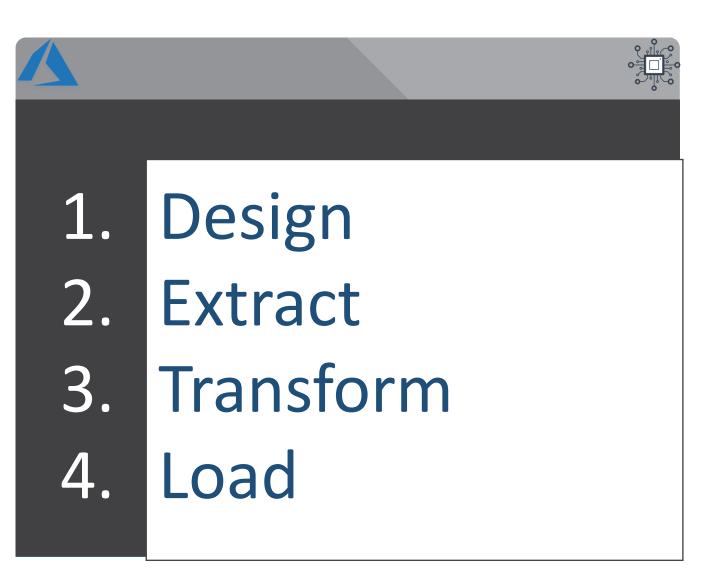# Thank you to our local Partners

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

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 ✔
        using the technology you have.

- Buck Woody

@BuckWoodyMSFT

What is your primary design focus?

- Disaster recovery
- Transaction level restart ability

Resilience

Rapid Delivery

- Metadata driven
- Continuous deployments

Performance

- Complex partitioning
- Large compute clusters

(Re)Usability

- Generic code libraries
- Dataset contracts

Scalability

- Microservices
- Small discrete batches
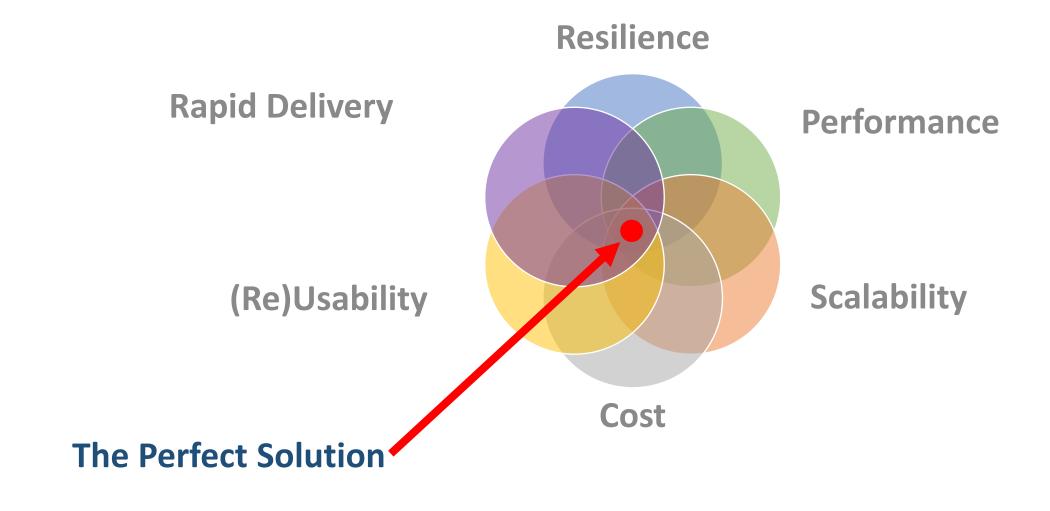
Cost
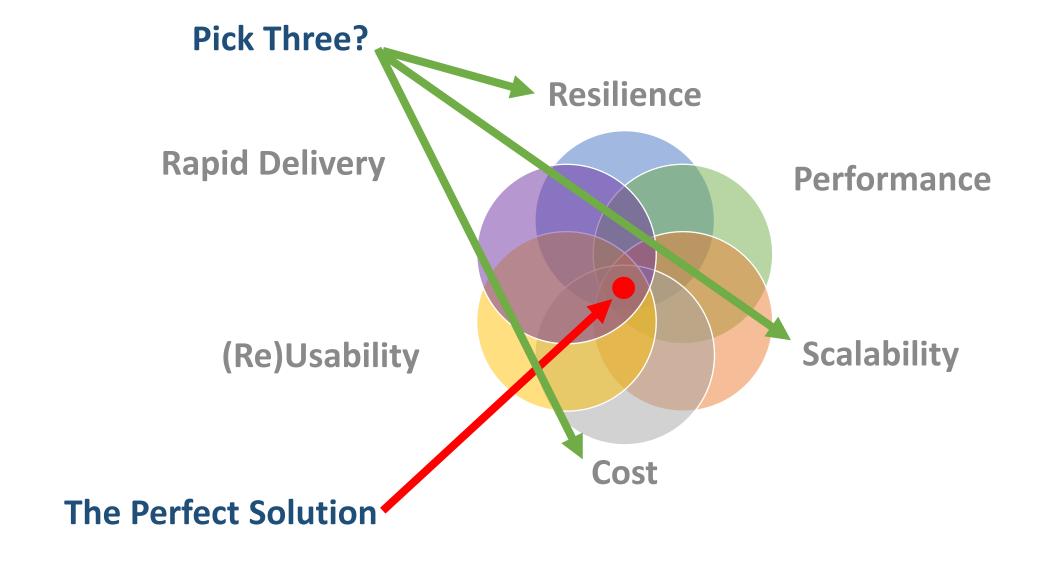
- Minimum resources used
- Careful resource management
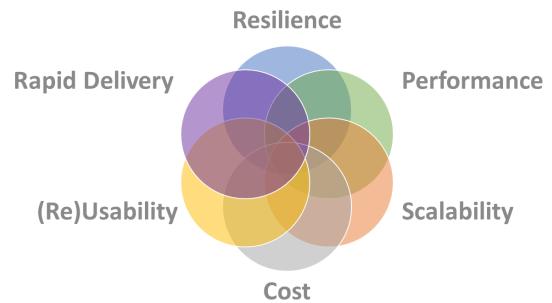
What is your primary design focus?

What is your primary design focus?

What is your primary design focus?

Agenda

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

Resilience

Rapid Delivery

Performance

(Re)Usability

Scalability

Cost

Data Extraction & Ingestion

Data Structure

CSV   TXT   XLS
HTML   TAR   JSON
XML   ZIP
DAT   PNG

Data Source

SAP

Push or Pull

Batch or Stream

Public or Private Transfer

Data Sensitivity

Data Volume

Data Extraction & Ingestion – Architecture 1

Data Structure	Push or Pull	Batch or Stream	Public or Private Transfer	Data Sensitivity	Data Volume

Data Source

Data Extraction & Ingestion – Architecture 1

On Premises | **Azure**
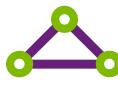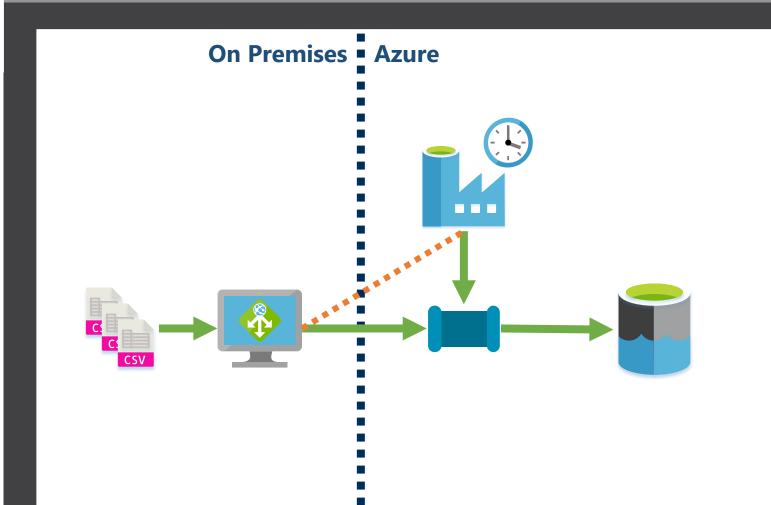
Requirements:
- Flat files
- From local storage
- Pulled from source
- Batch load
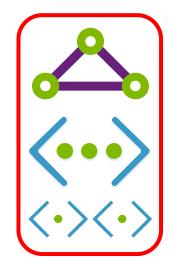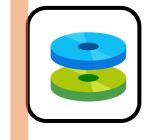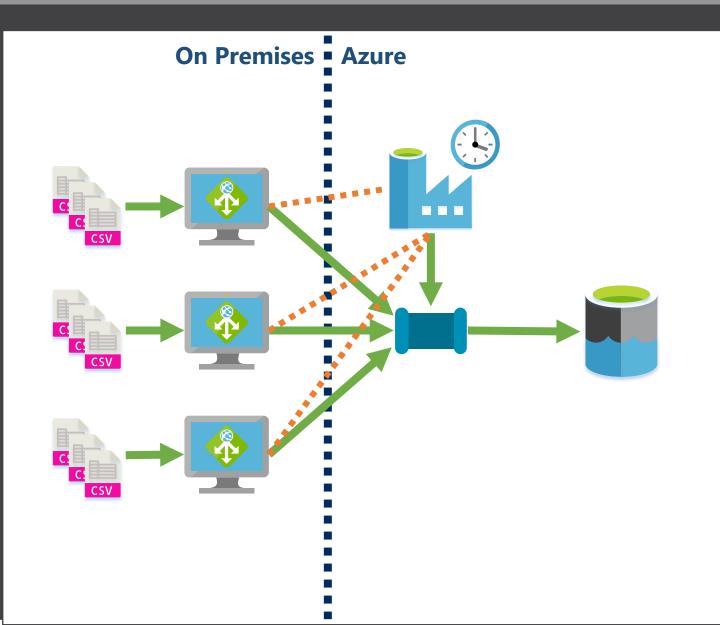- Public connections
- No PII data
- Small data volumes

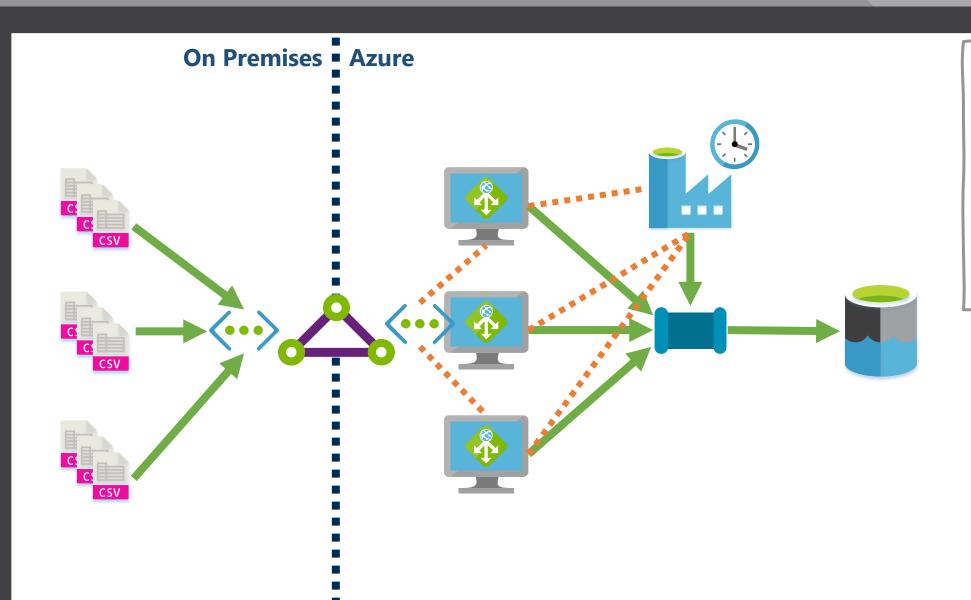# Data Extraction & Ingestion – Architecture 2

On Premises | **Azure**

Requirements:
- Flat files
- From local storage
- Pulled from source
- Batch load
- Public connections
- No PII data
- <u>Large</u> data volumes

Data Extraction & Ingestion – Architecture 3

Data Structure	Push or Pull	Batch or Stream	Public or Private Transfer	Data Sensitivity	Data Volume

On Premises : **Azure**

Requirements:
- Flat files
- From local storage
- Pulled from source
- Batch load
- <u>Private</u> connections
- No PII data
- Large data volumes
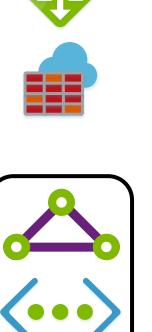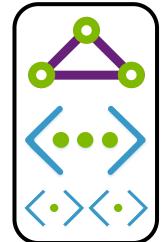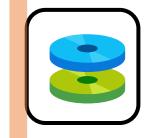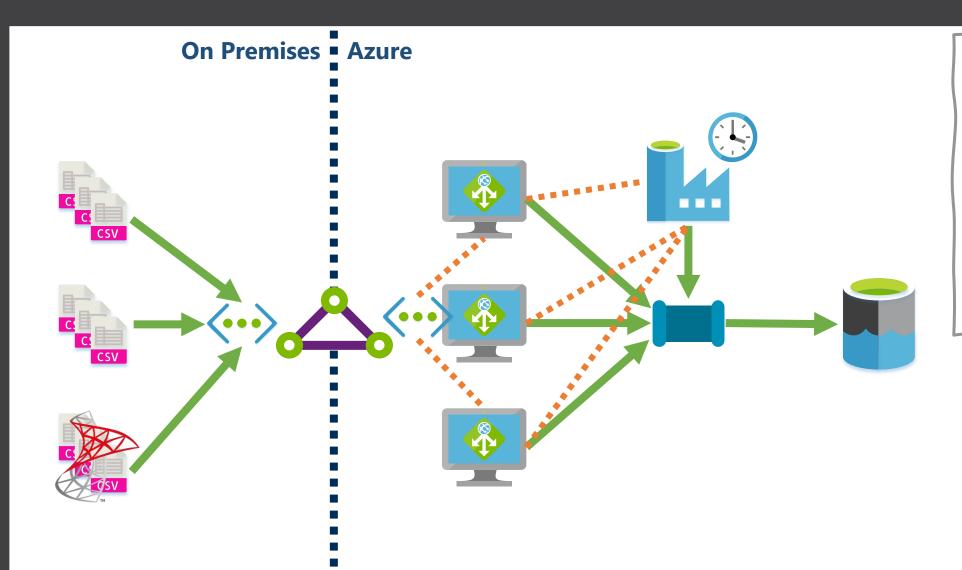
CSV

CSV

On Premises : **Azure**

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

Data Extraction & Ingestion – Architecture 4
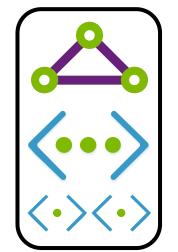
Data Extraction & Ingestion – Architecture 4

On Premises | **Azure**

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

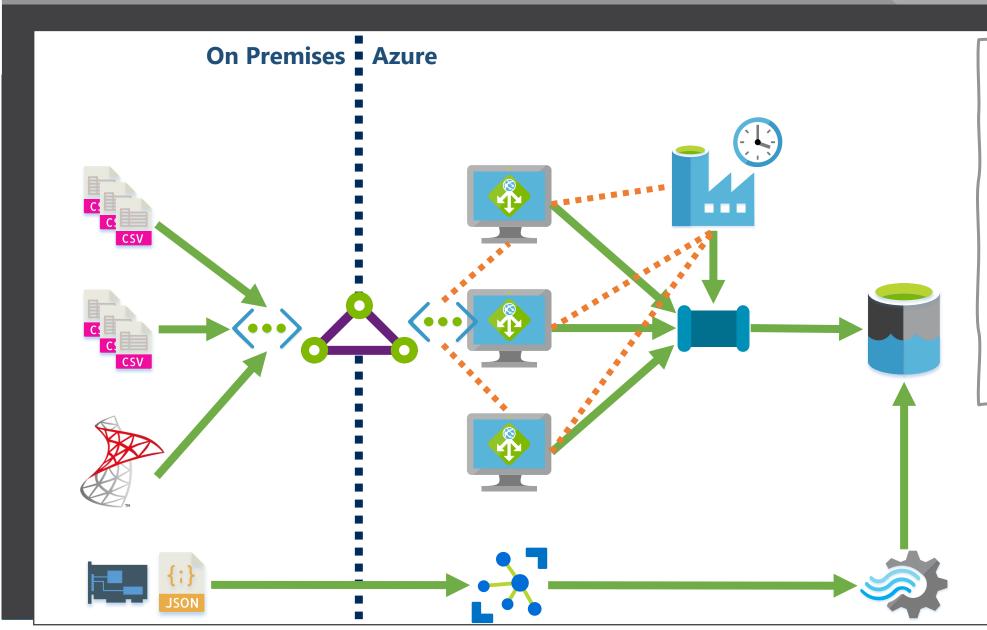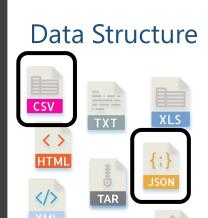# Data Extraction & Ingestion – Architecture 5
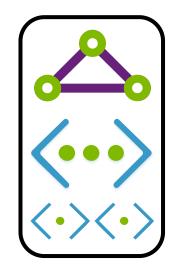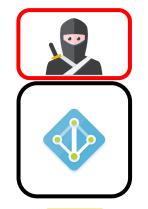
Data Structure

CSV TXT XLS
HTML JSON
XML TAR
DAT PNG ZIP

Data Source

REST SAP

Push or Pull

Batch or Stream

Public or Private Transfer

Data Sensitivity

Data Volume

Data Extraction & Ingestion – Architecture 5

On Premises | **Azure**
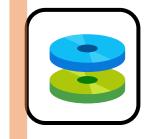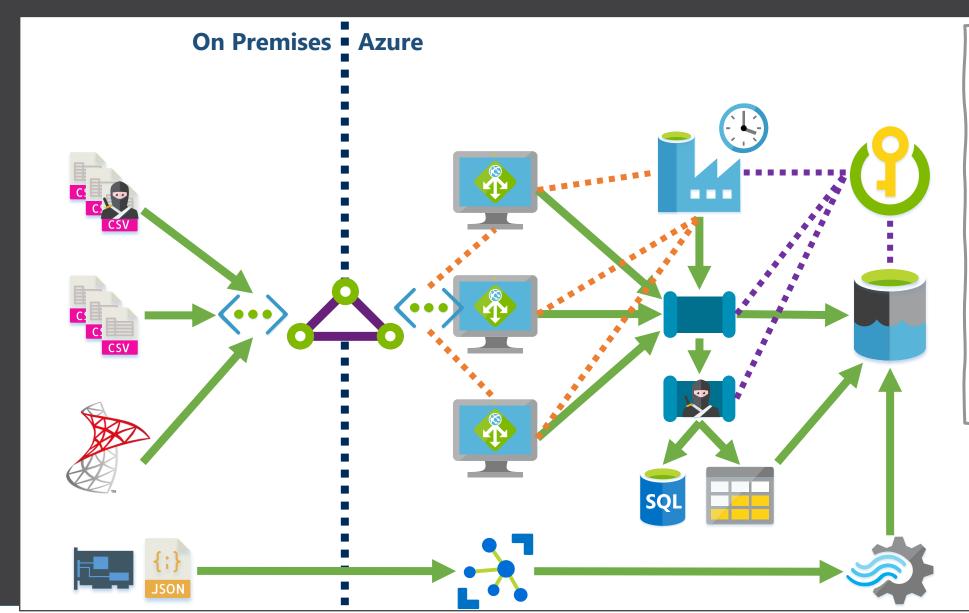
Requirements:
- Flat files & JSON
- From local storage & database tables
- Pulled from source & pushed
- Batch load & streamed
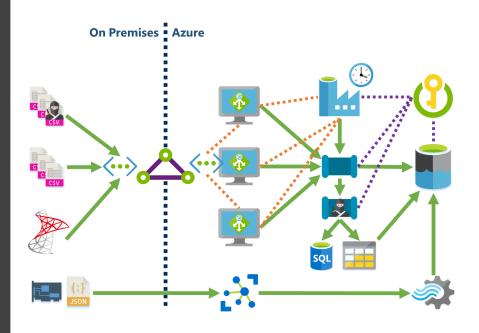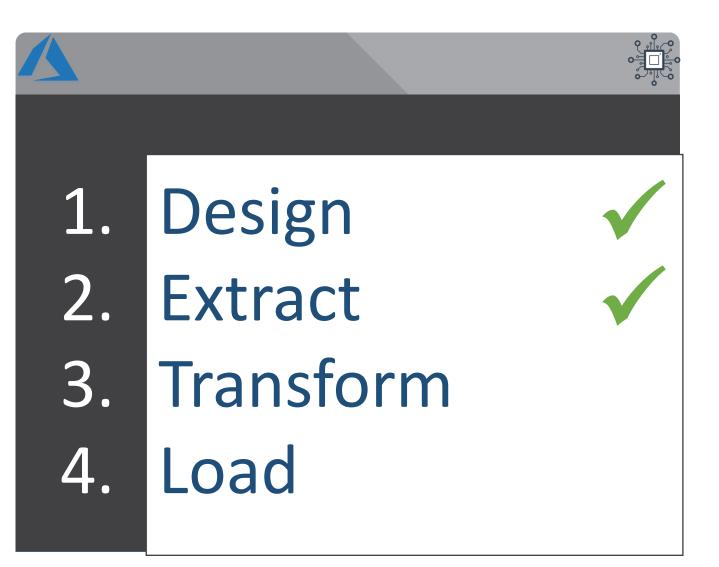- Private connections
- No PII data
- Large data volumes

Data Extraction & Ingestion – Architecture 6

Data Structure	Push or Pull	Batch or Stream	Public or Private Transfer	Data Sensitivity	Data Volume

Data Source

Data Extraction & Ingestion – Architecture 6

On Premises | **Azure**

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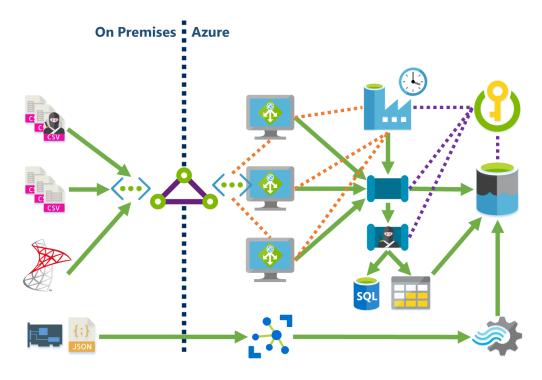- Both PII data & none PII data
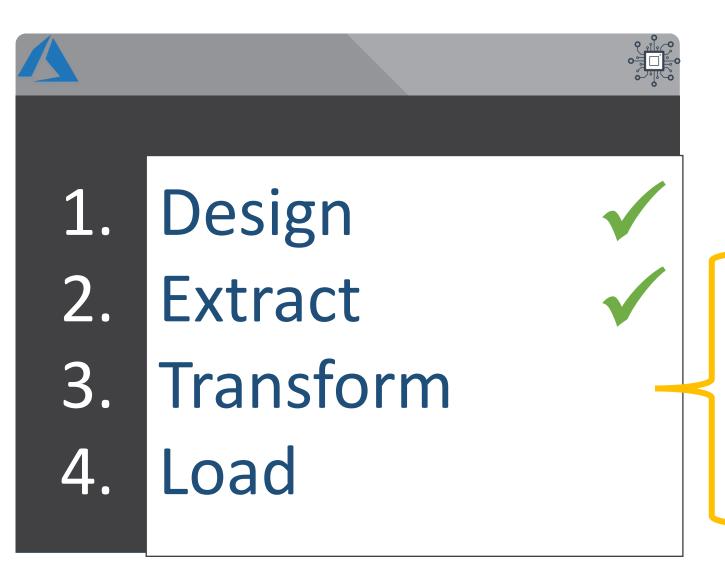- Large data volumes

Extract | Transform | Load

Agenda

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

On Premises | Azure

Agenda

1. Design ✓
2. Extract ✓
3. Transform
4. Load
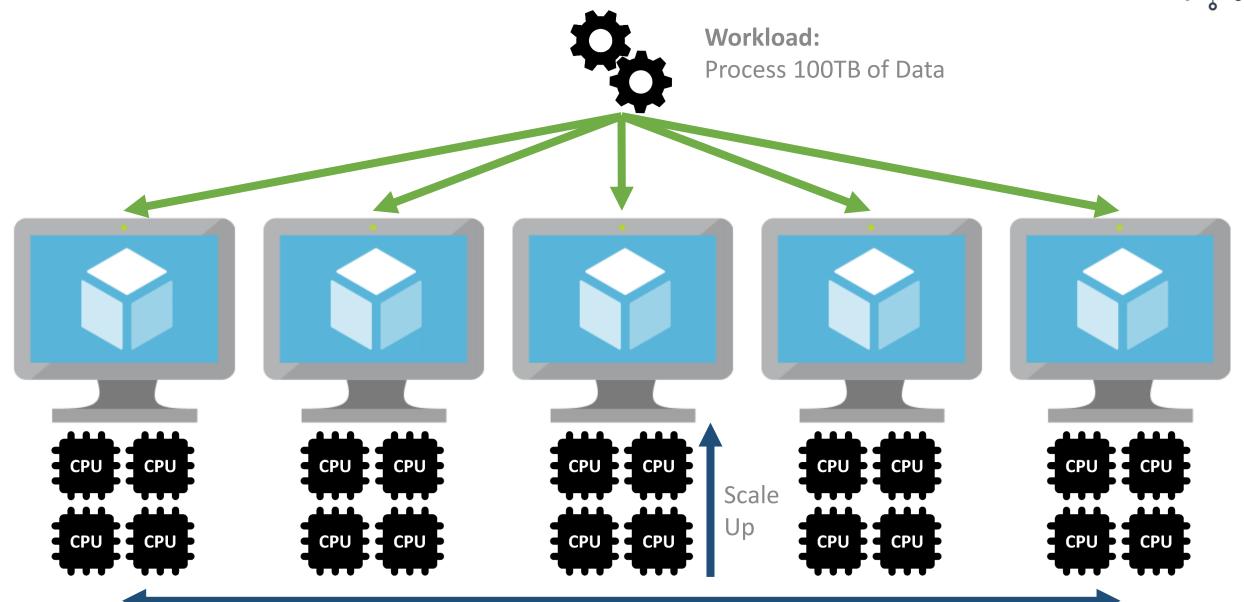
Compute
Storage, Structure
& Data Format
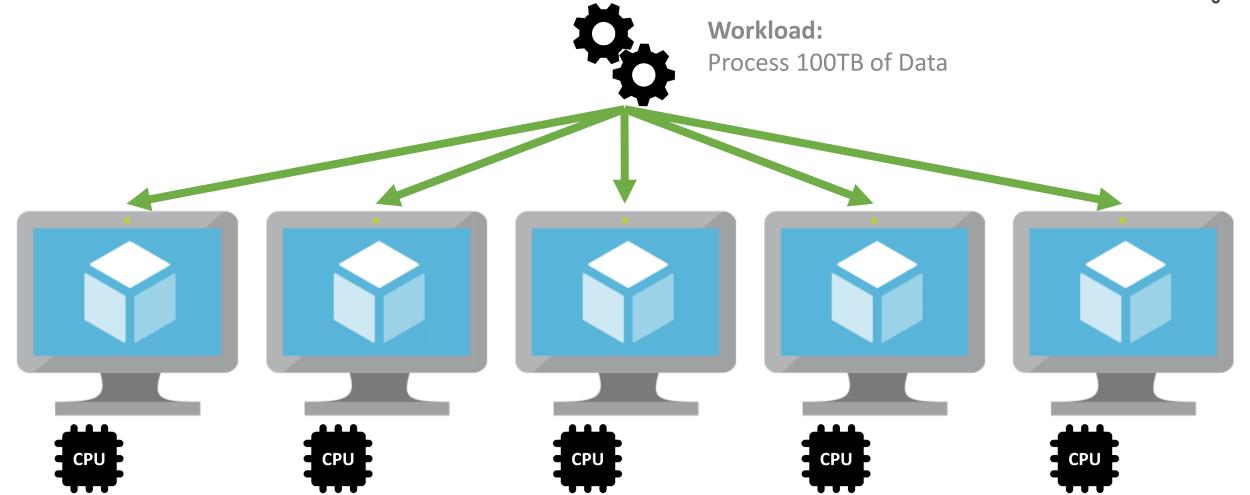
Scaling Up and/or Scaling Out

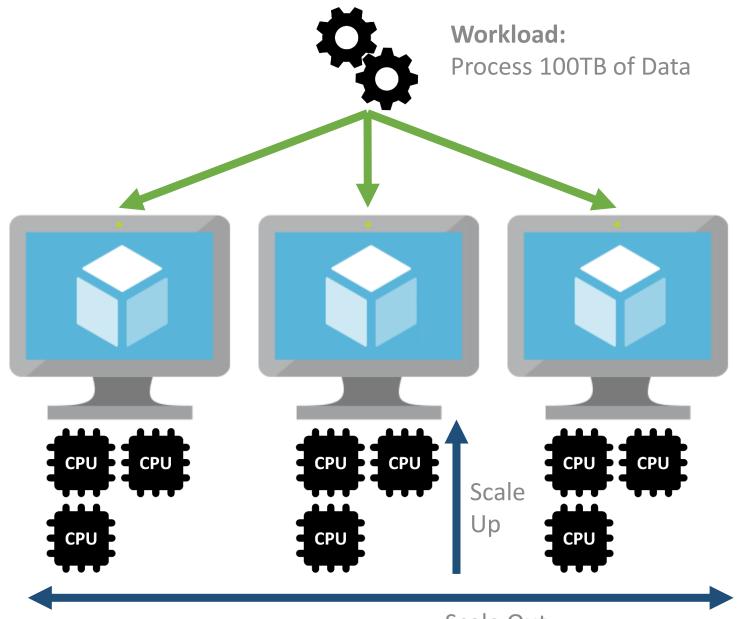**Workload:**
Process 100TB of Data

Scale Out

What Compute Type of Compute?

Workload:
Process 100TB of Data

Platform   **I**nfrastructure

As
A
Service

What Compute Type of Compute?

Infrastructure

As
A
Service

Database
Engine

Reporting
Services

Integration
Services

Data Quality
Services

Analysis
Services

Master Data
Services

Microsoft®
SQL Server®
™

SQL Agent

2019
Big Data Cluster

Microsoft®
SQL Server®

Microsoft® SQL Server®

2019

Big Data Cluster

Data Transformation – Compute

Microsoft SQL Server
2019
Big Data Cluster

Azure SQL Database (SQLDB)

Azure SQL Data Warehouse (SQLDW)

Compute

T-SQL
Compute Node

D-SQL
Control Node

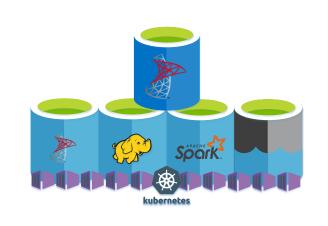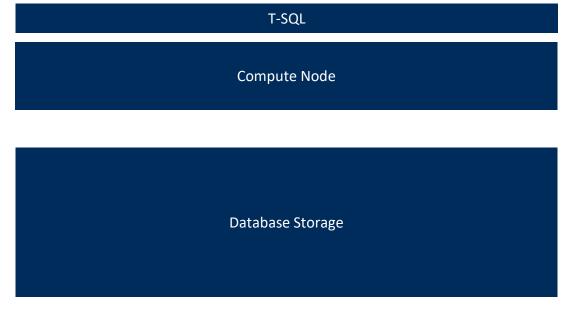| Compute Node	Compute Node	Compute Node	Compute Node

Data

Database Storage

Scale Up Compute

Scale Out Compute & Distributed Storage

Data Transformation – Compute

Microsoft SQL Server 2019 Big Data Cluster

Azure SQL Database (SQLDB)

Azure SQL Data Warehouse (SQLDW)

Compute

T-SQL				
Compute Node	Read Only	Read Only	Read Only	Read Only

D-SQL
Control Node

Compute Node	Compute Node	Compute Node	Compute Node

Data

Database Storage

Scale Up Compute

Scale Out Compute & Distributed Storage

Data Transformation – Compute

Microsoft® SQL Server® 2019 Big Data Cluster

T-SQL			
Compute Node			

D-SQL			
Control Node			
Compute Node	Compute Node	Compute Node	Compute Node

Database Storage

Data Transformation – Compute

Microsoft® SQL Server® 2019
Big Data Cluster

T-SQL	Python/Scala

Control — Master Instance

Compute

Compute Node	Compute Node	Compute Node	Compute Node

Data & Storage

Data Pool
(Storage Shards)

Storage Pool
(SQL Engine, HDFS, Spark)

Linux Containers in Kubernetes

What Compute Type of Compute?

Platform

As
A
Service

Data Transformation – Compute
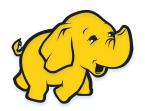
Data Lake Analytics

HDInsight

SQL Database

SQL Data Warehouse (Synapse)

Databricks

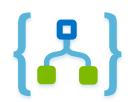Custom Activity Batch Service

SSIS Packages

Automation

Cosmos

Functions

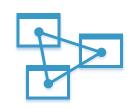Power BI Data Flows

Logic Apps

Data Factory Data Flows

Analysis Services

Data Transformation – Compute
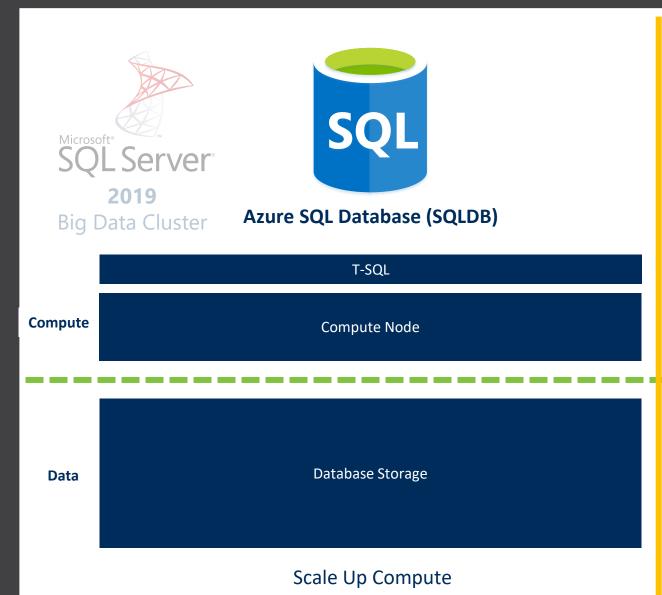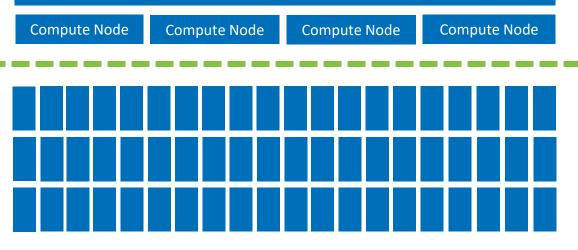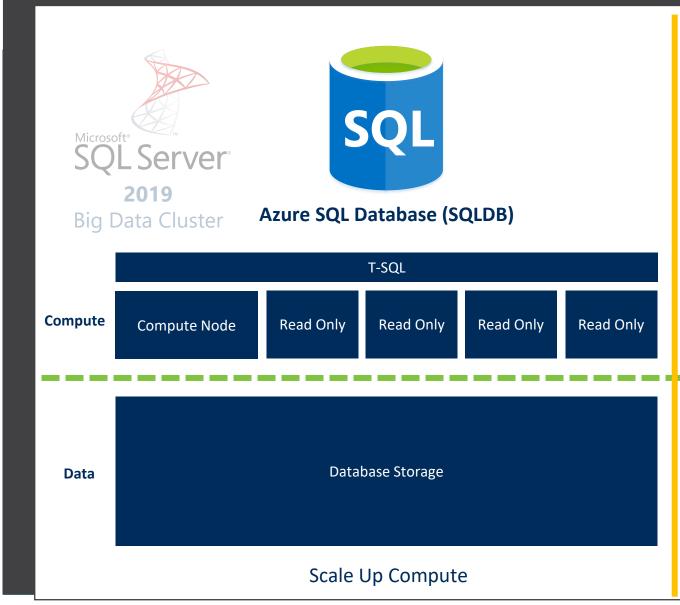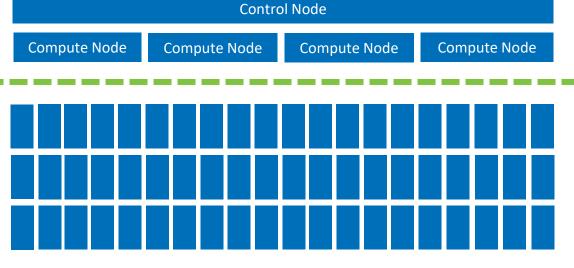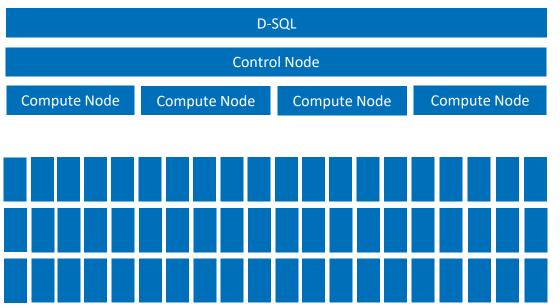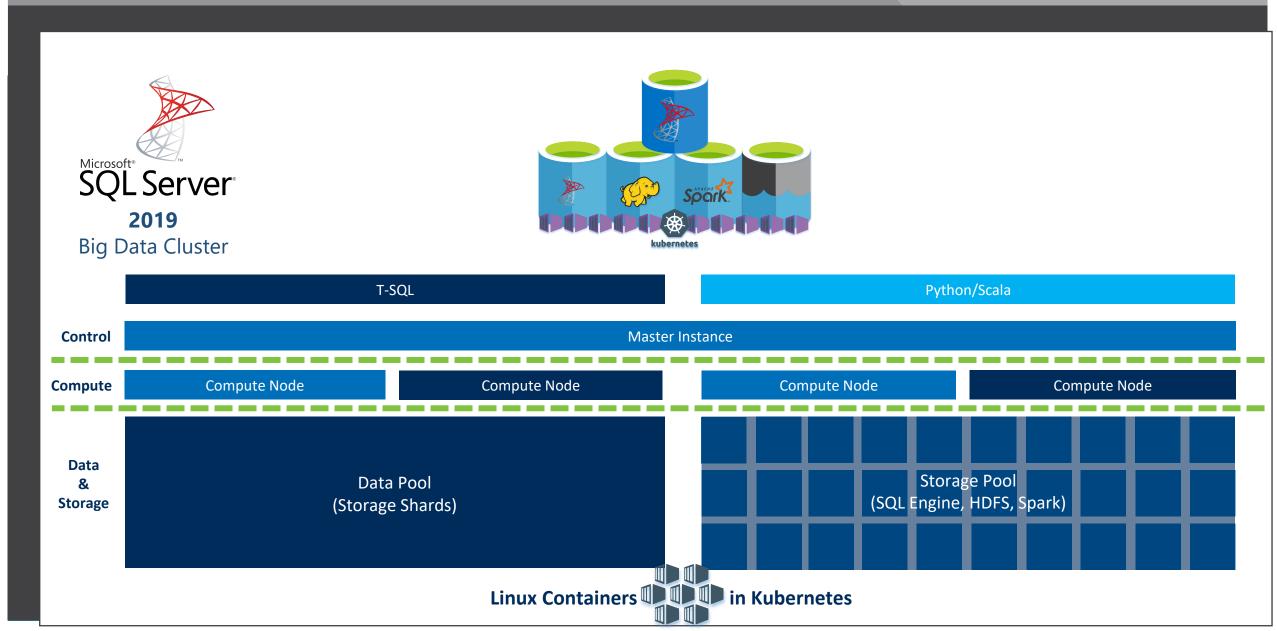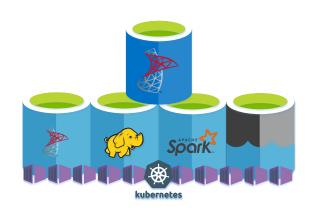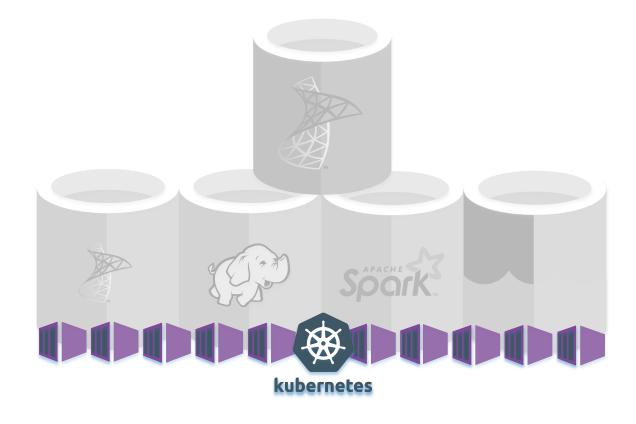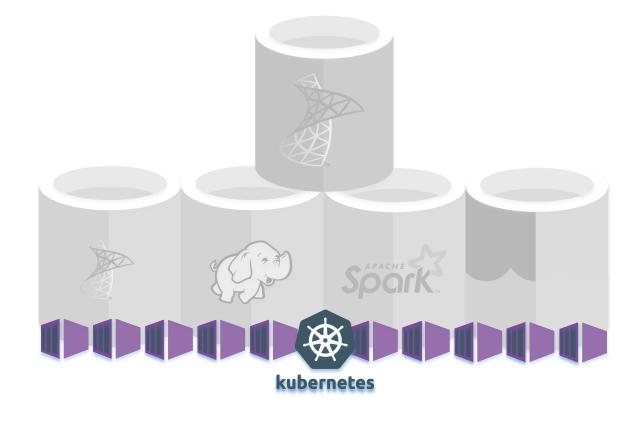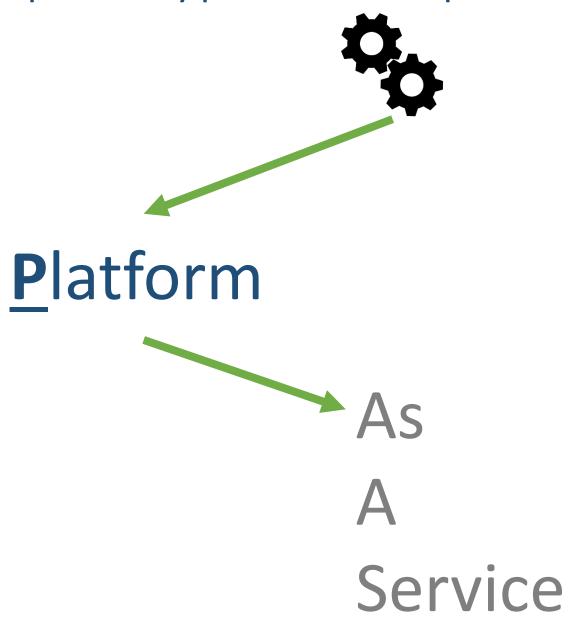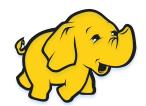
Data Lake Analytics

U-SQL to be deprecated?

HDInsight

Manual cluster configuration.

SQL Database

Only scales up. Hyperscale tier.

SQL Data Warehouse (Synapse)

Scales out. Is Synapse ready?

Databricks

Learn Scala. As open as Spark?

Custom Activity Batch Service

Build your own! Scales up.

SSIS Packages

IaaS to run. Scales up.

Automation

Short amounts of compute.

Cosmos

Transactional data. Slow with batch data.

Functions
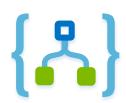
Build your own! Durable?

Power BI Data Flows

Not enterprise grade.

Logic Apps
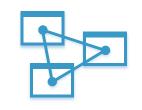
Event based actions & data movements.

Data Factory Data Flows

Currently limited transformations.
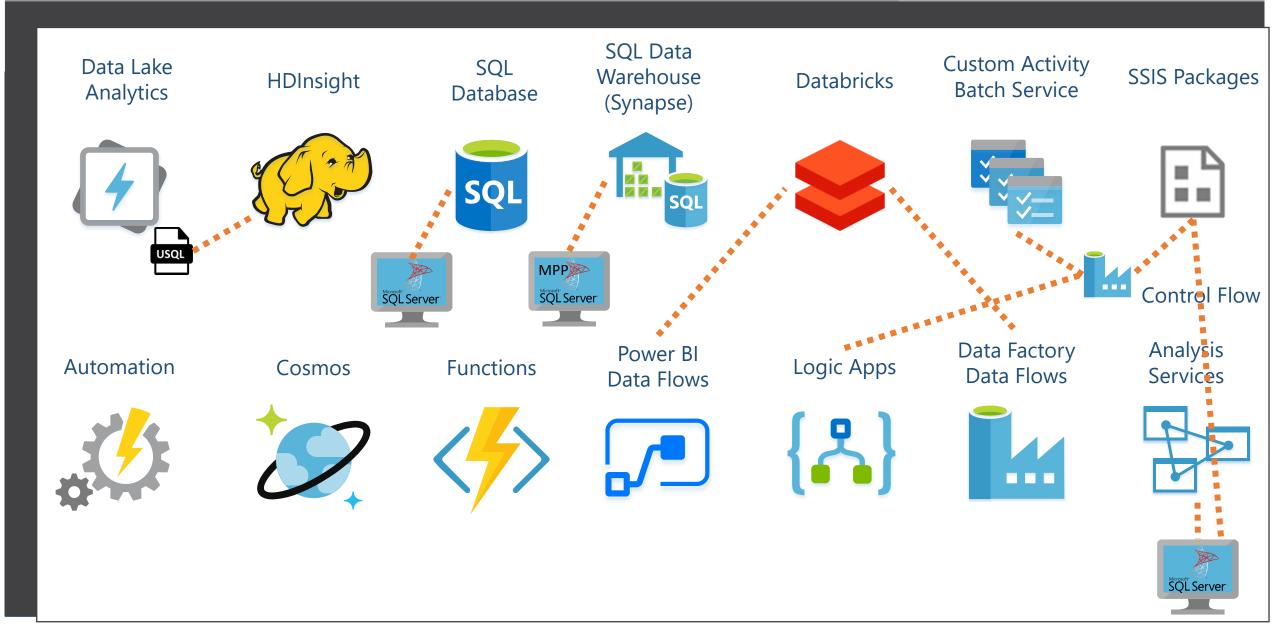
Analysis Services

Calculation optimised. Not for engineering.

Data Transformation – Compute

Data Lake Analytics

HDInsight

SQL Database

SQL Data Warehouse (Synapse)

Databricks

Custom Activity Batch Service

SSIS Packages

Automation
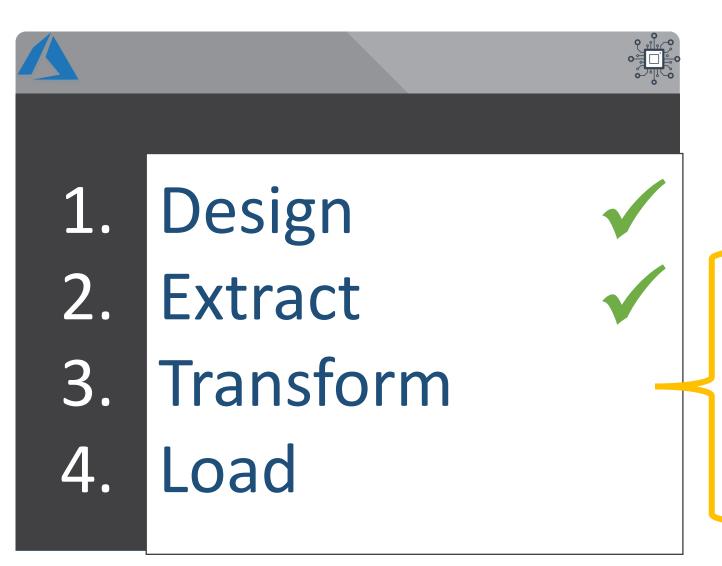
Cosmos

Functions

Power BI Data Flows

Logic Apps

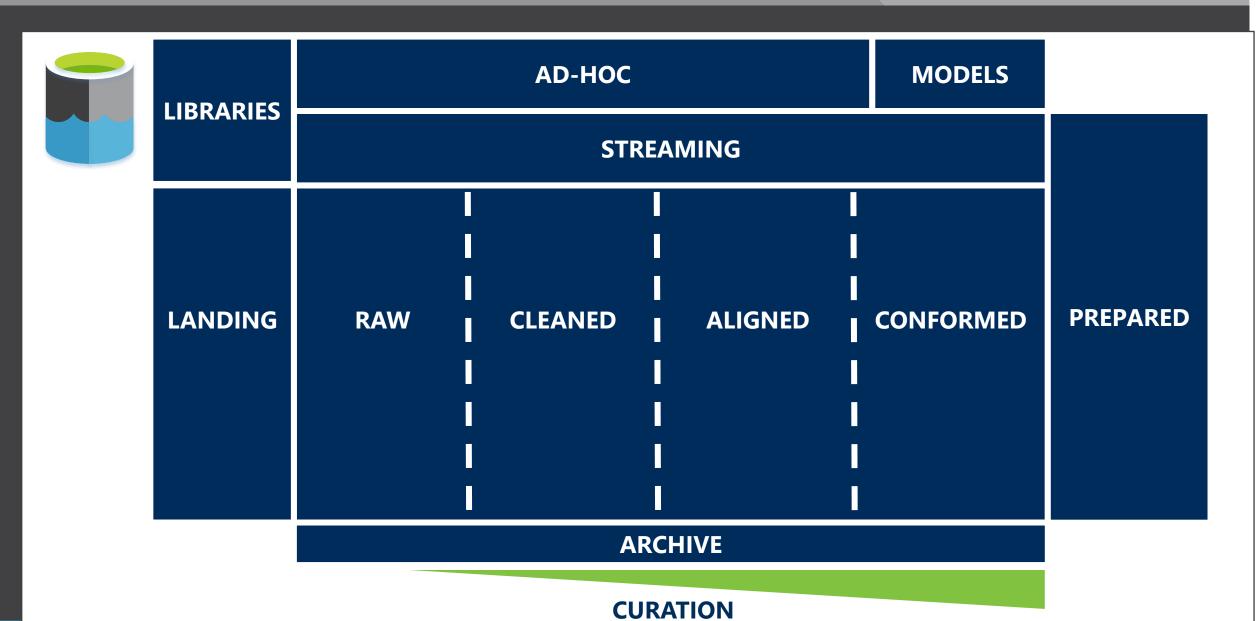Data Factory Data Flows

Analysis Services

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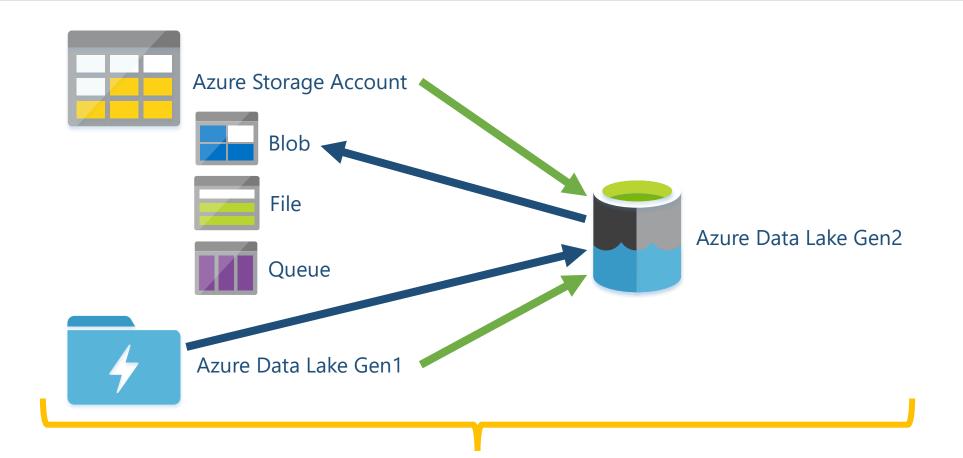
1. Design ✓
2. Extract ✓
3. Transform
4. Load

Compute ✓
Storage, Structure & Data Format

IN

MY FIRST DATA LAKE

OUT

Data Transformation – Storage & Format

LIBRARIES	AD-HOC				MODELS
	STREAMING				
LANDING	RAW	CLEANED	ALIGNED	CONFORMED	PREPARED
		...\Year=2020\ Month=01\ Day=01\	Append OR Override	Facts & Dimensions	

ARCHIVE

CURATION

Data Transformation – Storage & Format

Data Transformation – Storage & Format

Agenda

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

Compute ✓
Storage, Structure
& Data Format ✓

Agenda

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

Loading & Consuming Data

PREPARED

SERVE

Scaling:
S1
P1
vCore
Hyperscale

MODEL

Import
OR
Direct Query

VISUALIZE

Premium
OR
Professional

Loading & Consuming Data

Loading & Consuming Data

PREPARED

MODEL

Import

VISUALIZE

PREPARED

DELTA LAKE

VISUALIZE
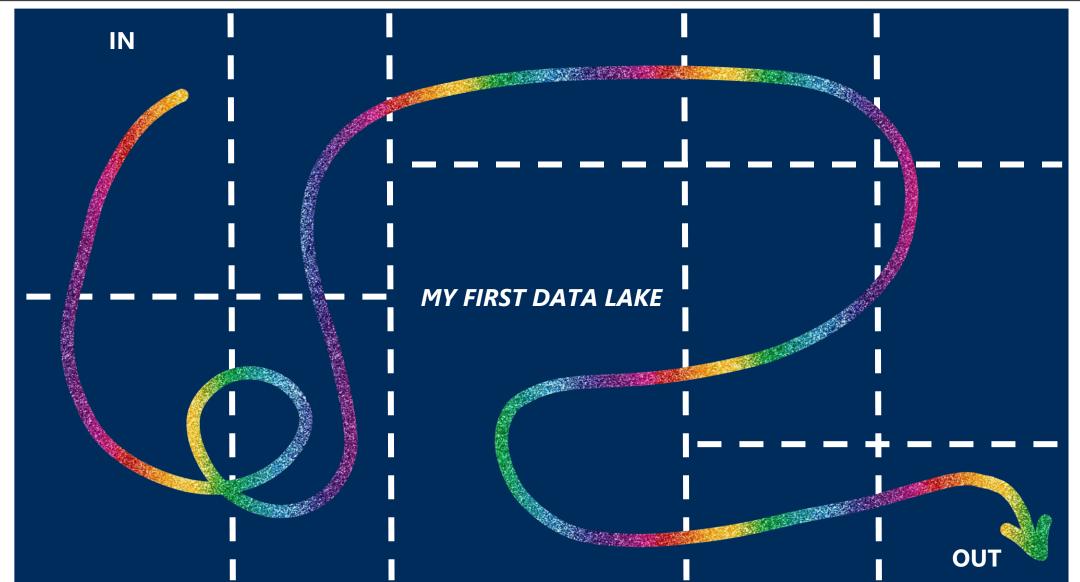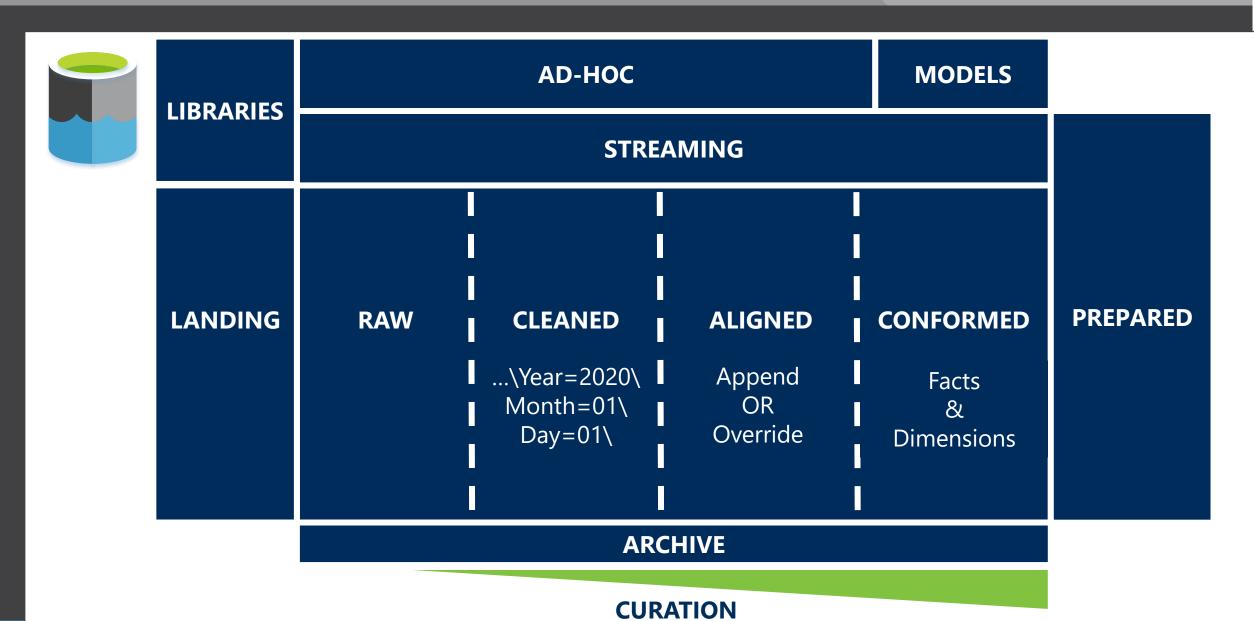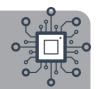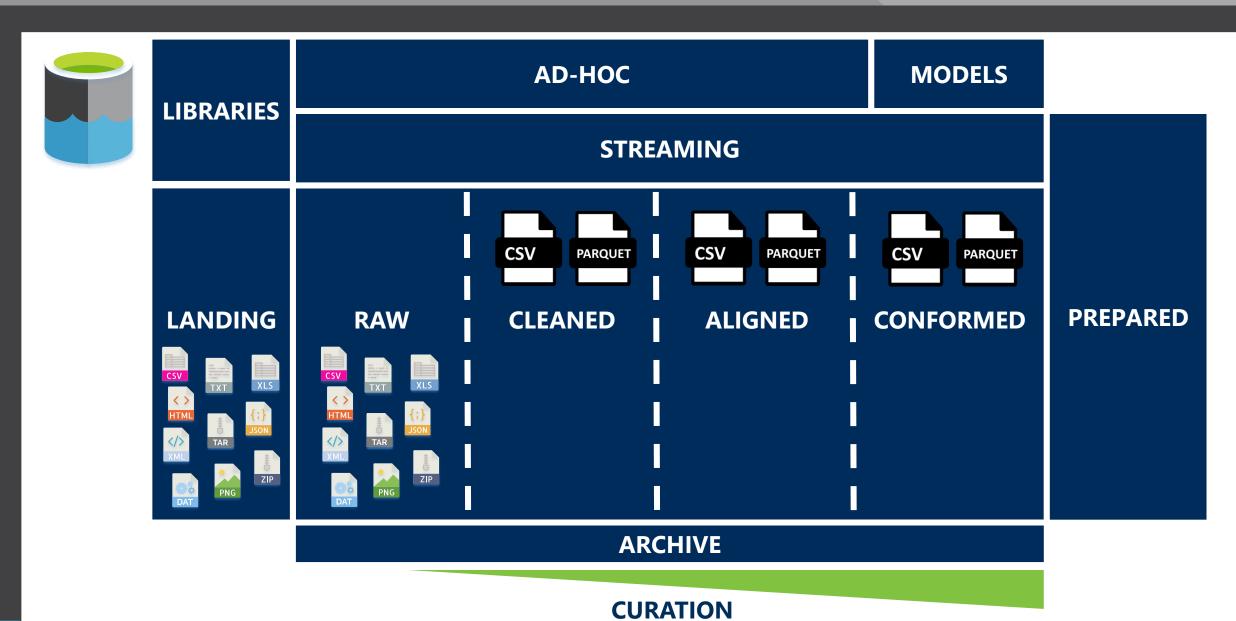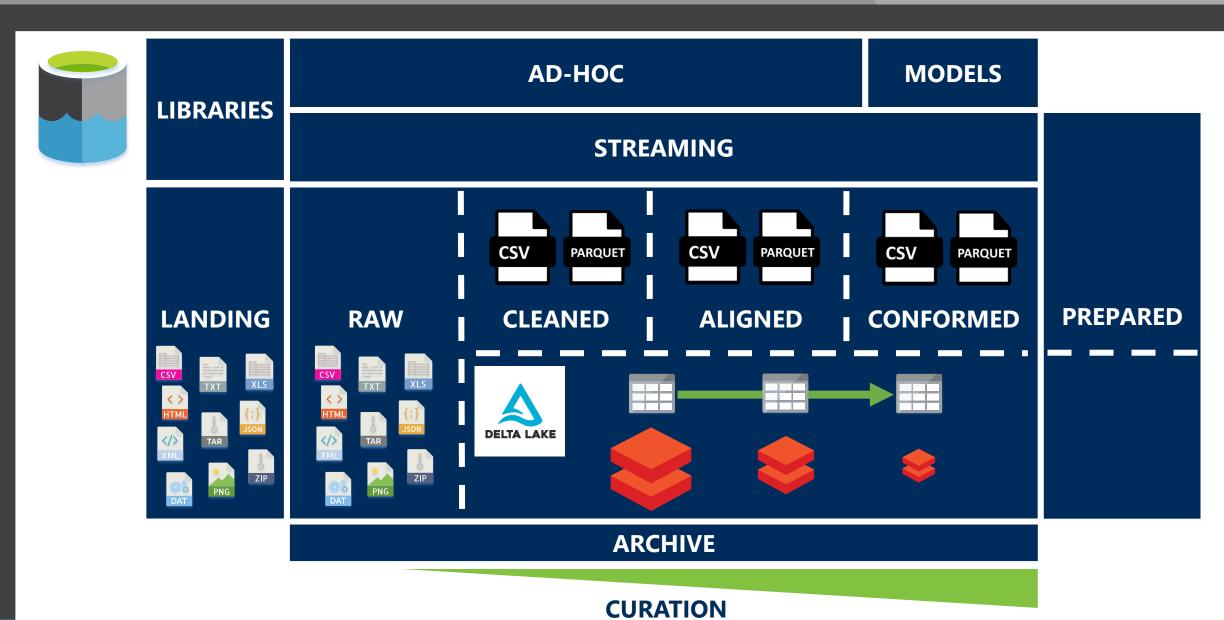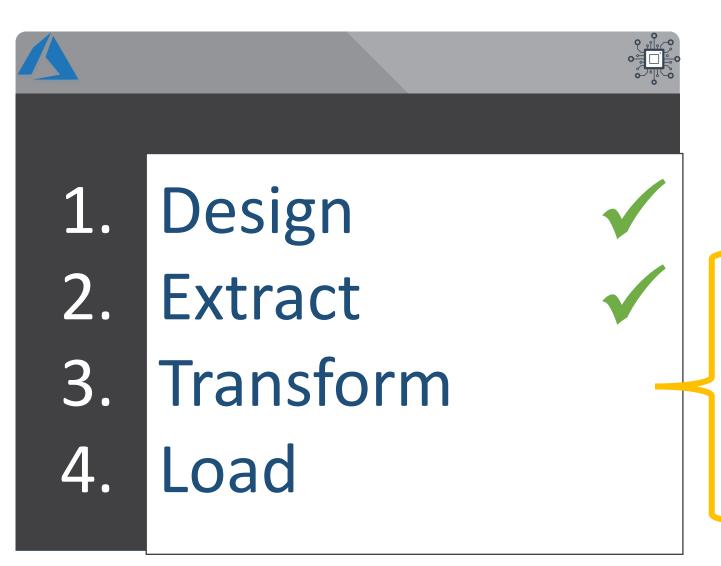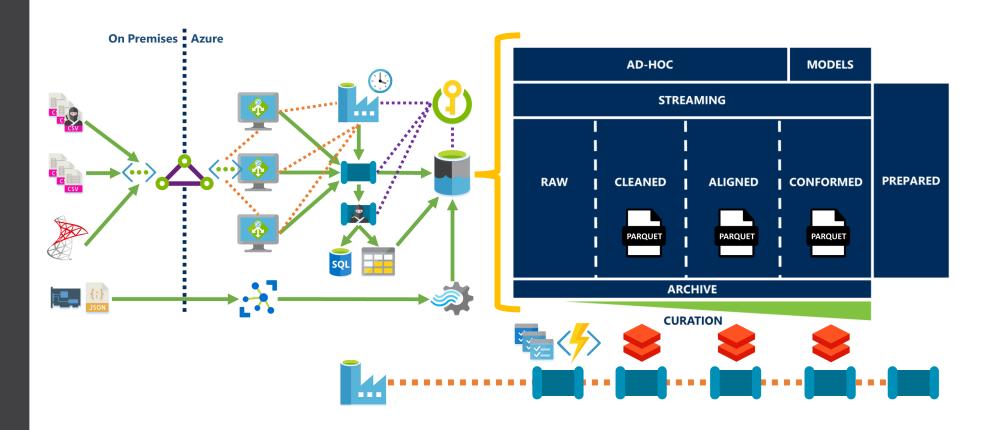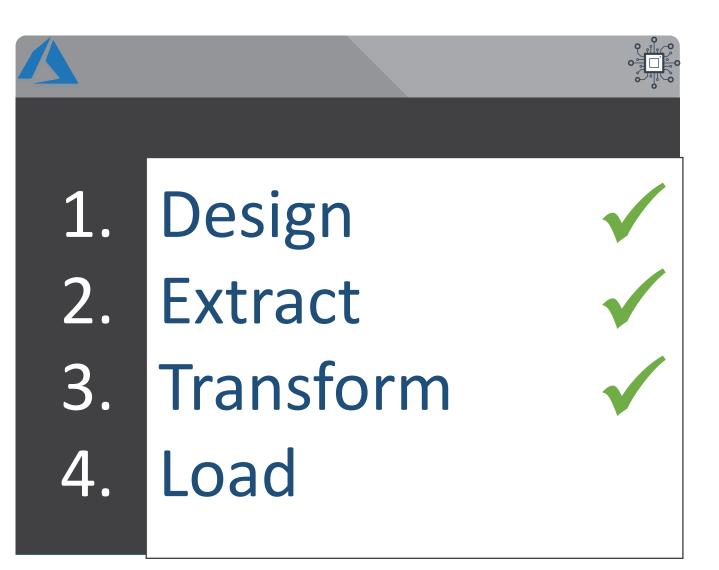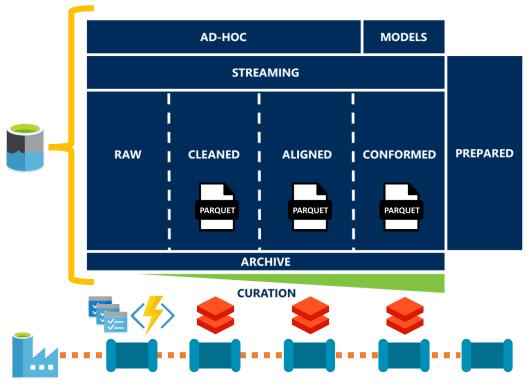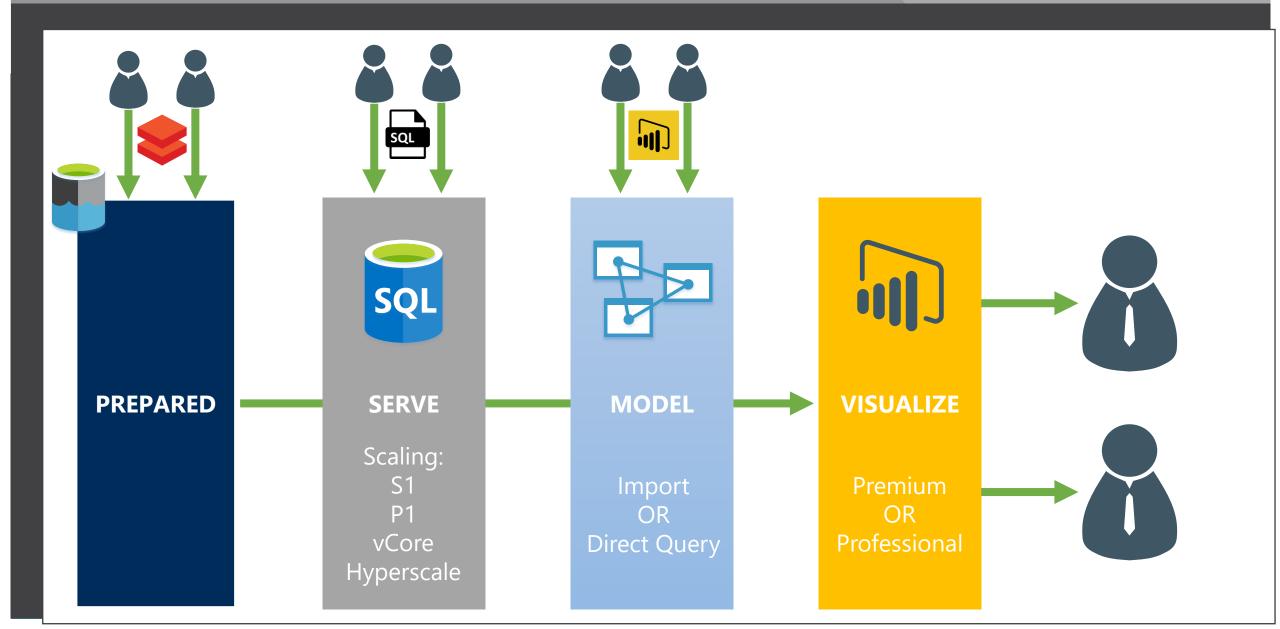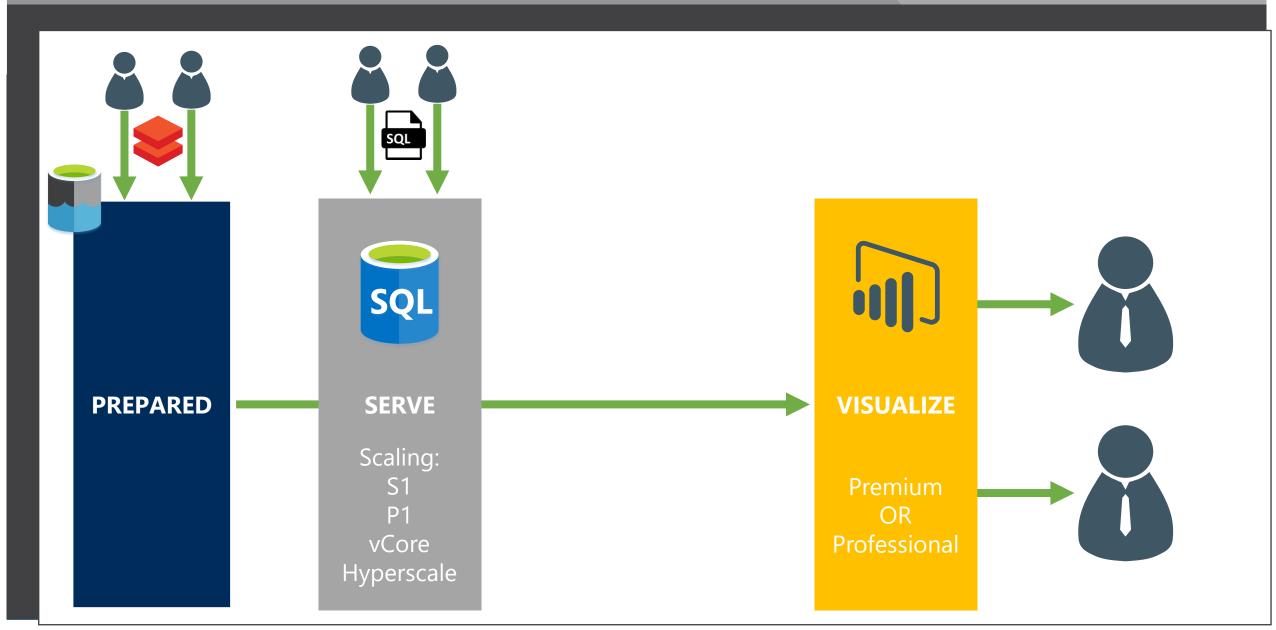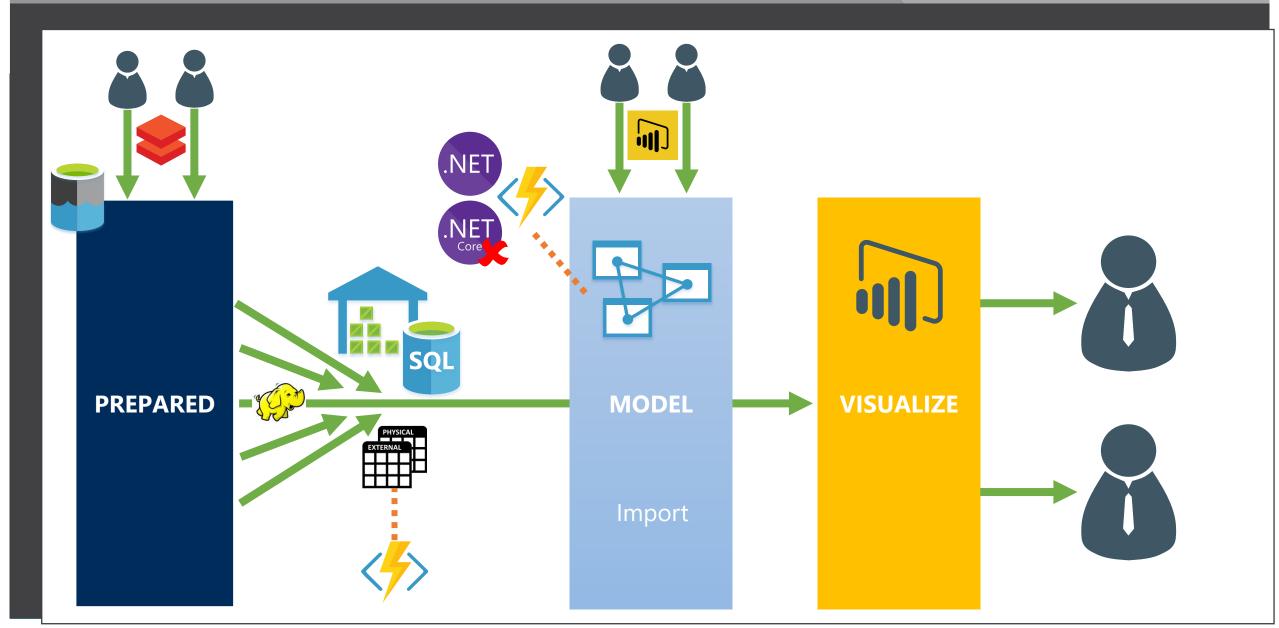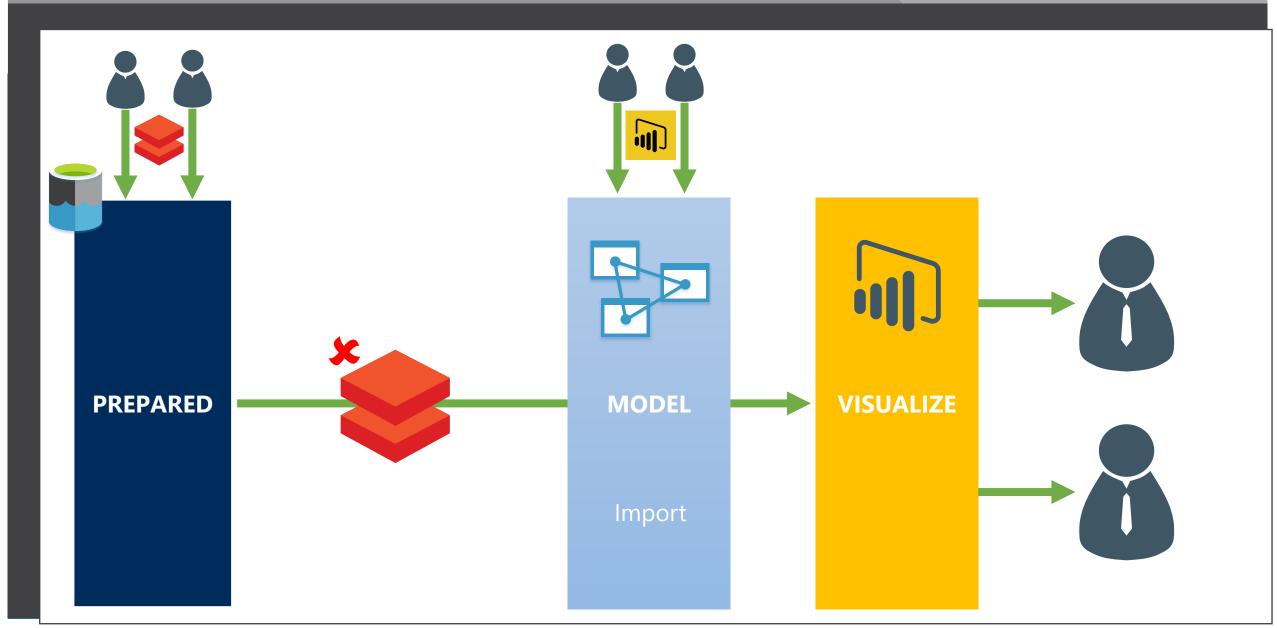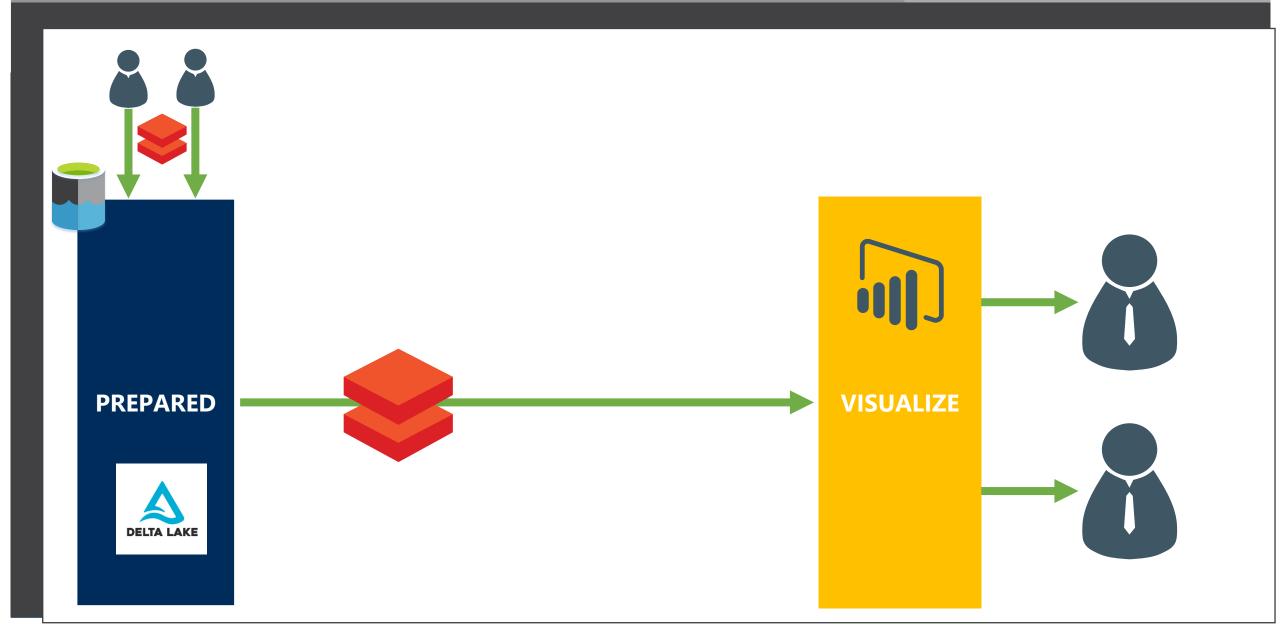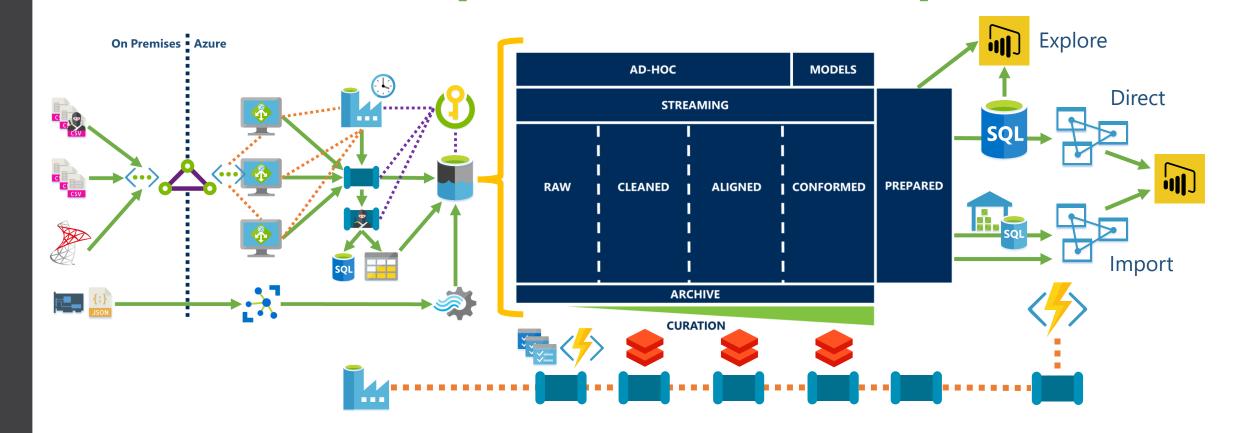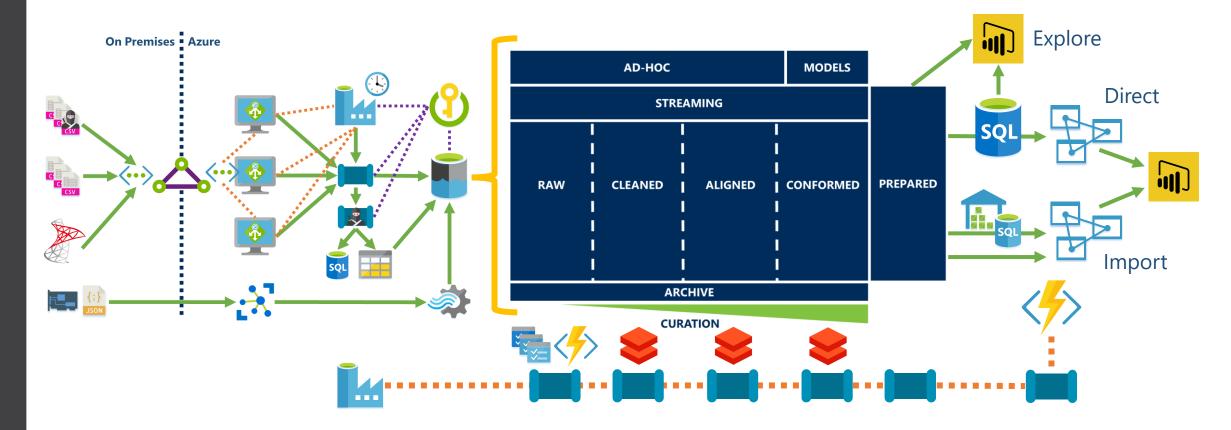
Thank you for listening...

Paul Andrew

Blog: mrpaulandrew.com
Email: paul@mrpaulandrew.com

Twitter: @mrpaulandrew
LinkedIn: In/mrpaulandrew

GitHub: github.com/mrpaulandrew