Modern Analytics Platform

Lambda Architecture in Azure

Simon Whiteley | Adatis

10/03/2017

Microsoft Partner
Microsoft

Gold Data Analytics
Gold Data Platform
Gold Cloud Platform

MVP Microsoft® Most Valuable Professional

adatis

Overview

Cloud BI

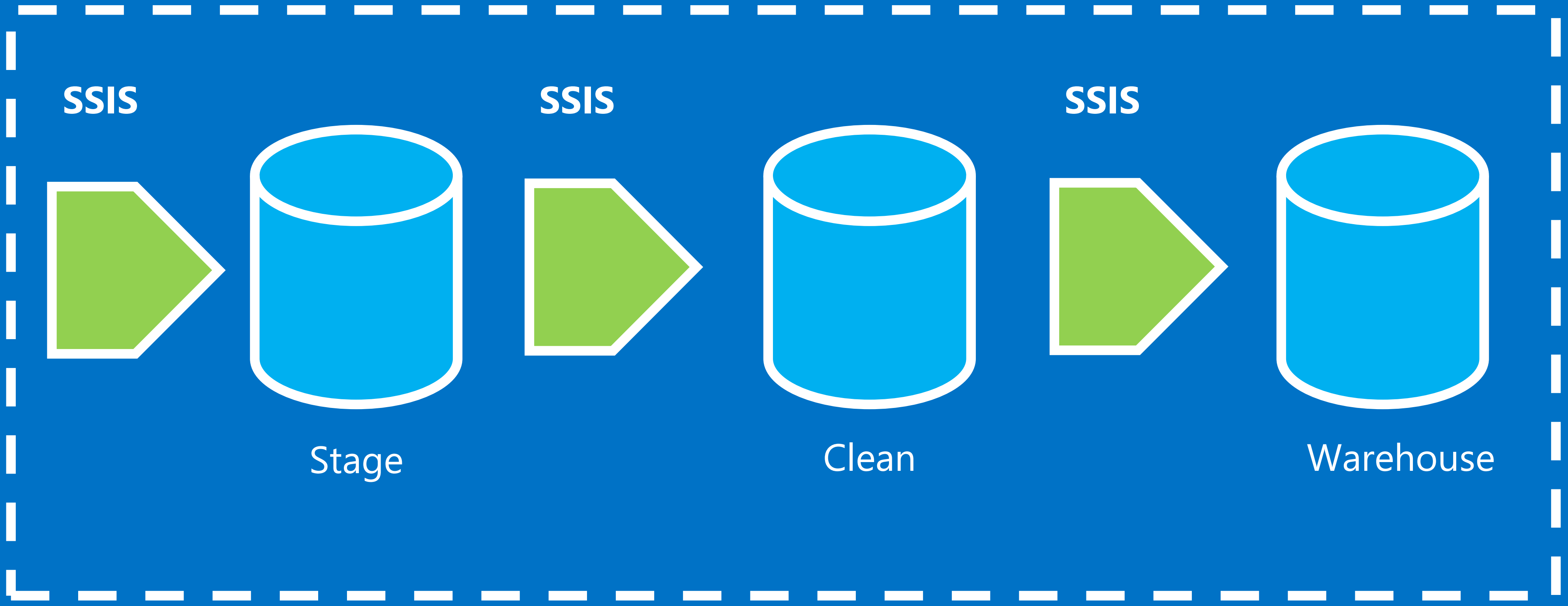
What is Lambda?

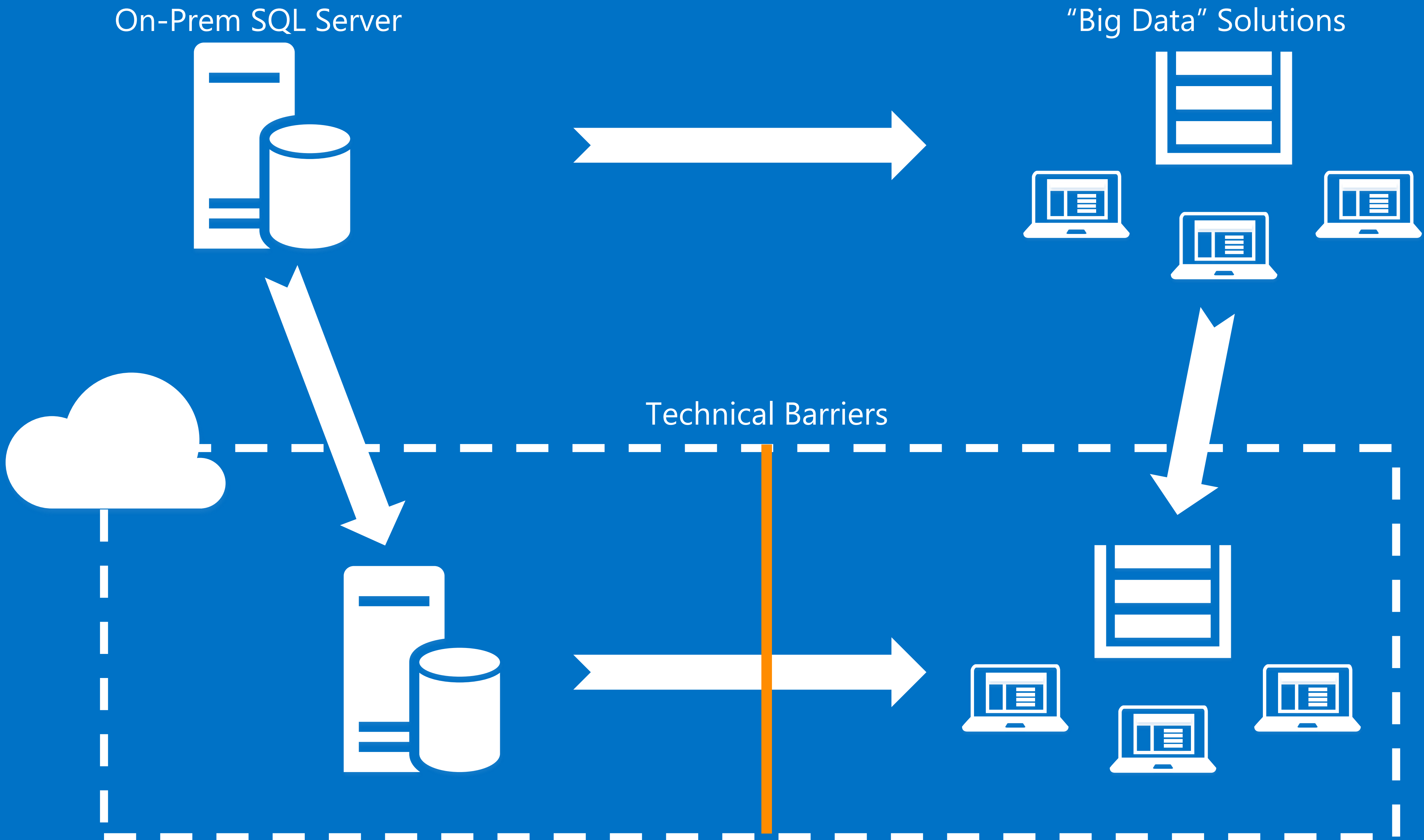
The Native Azure Approach

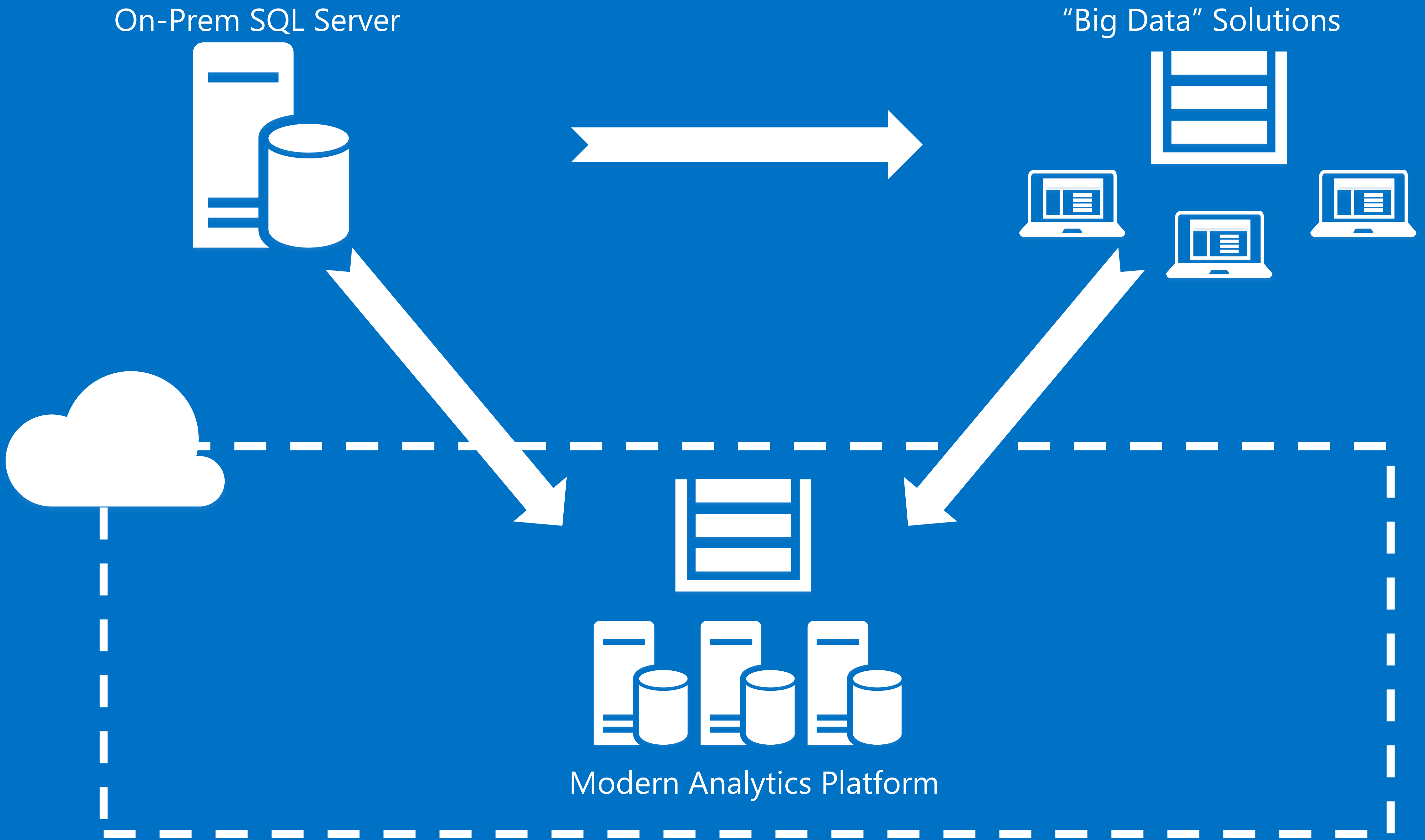
Alternative Models

History

One-Box SQL BI Architecture




On-Prem SQL Server
"Big Data" Solutions
Technical Barriers


On-Prem SQL Server
"Big Data" Solutions
Modern Analytics Platform

My Life Goal:

Never to manage another Server

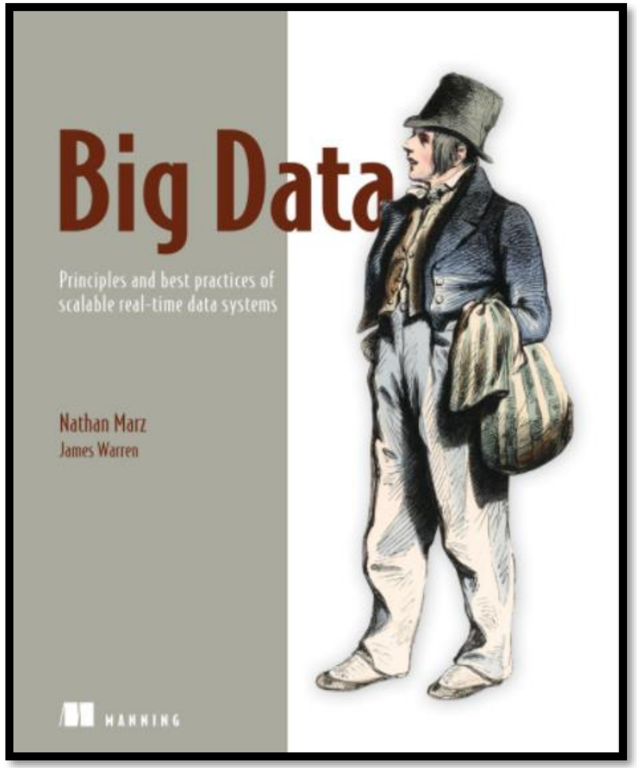
MAP Wish List

- Can Handle Massive Datasets
- Linearly Scalable
- Near Real-Time
- Fault Tolerant
- Low Tech Barrier for SQL Devs

LAMBDA

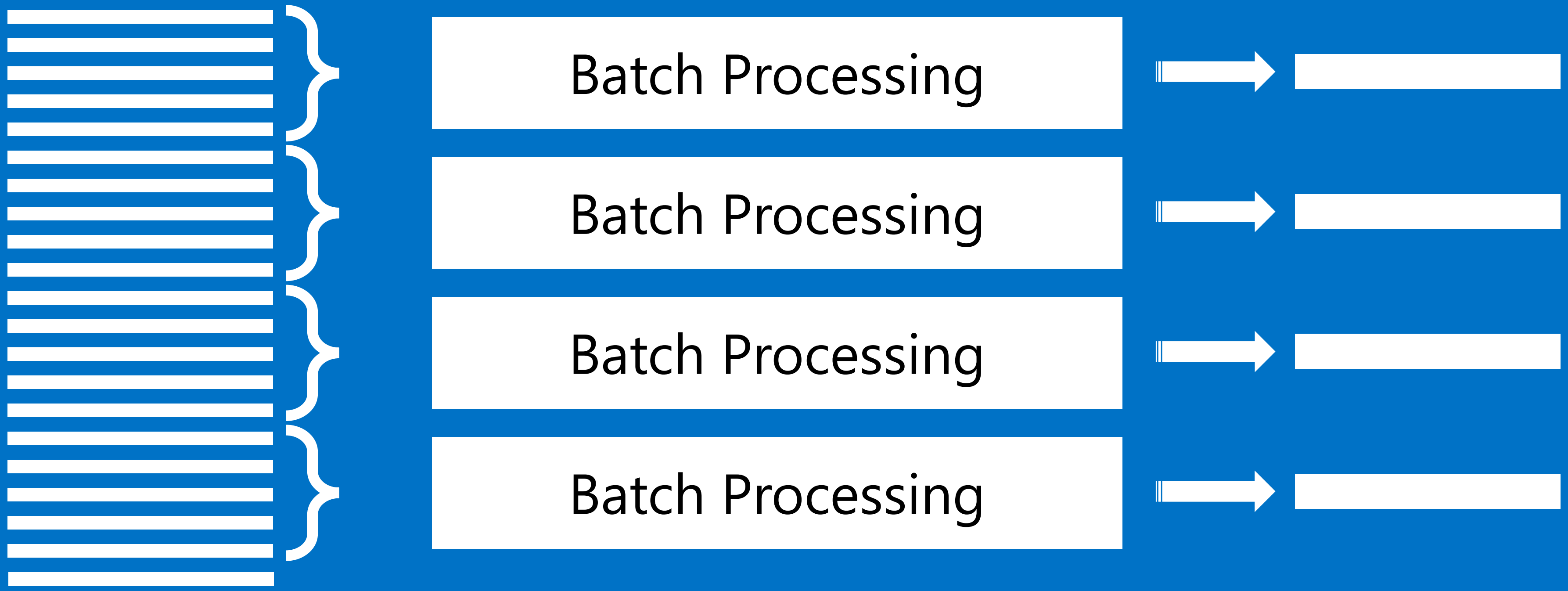
Lambda Architecture

Use Batch and Stream technologies together to balance latency, throughput and fault-tolerance


Big Data
Principles and best practices of scalable real-time data systems
Nathan Marz
James Warren
MANNING

Nathan Marz & James Warren

PROBLEM

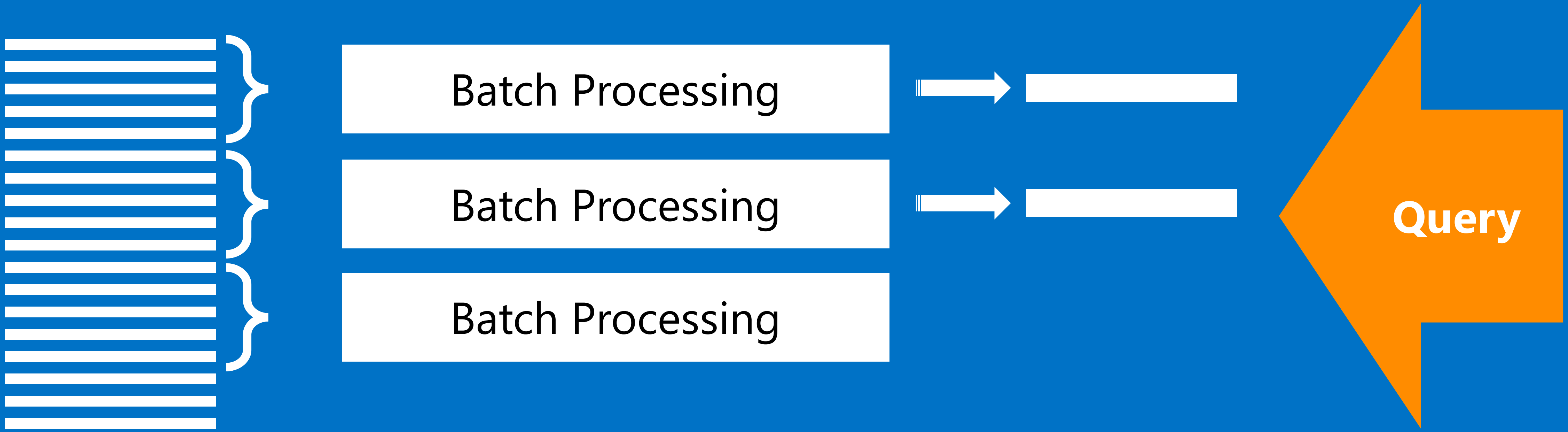


Batch Processing

Batch Processing

Batch Processing

Batch Processing

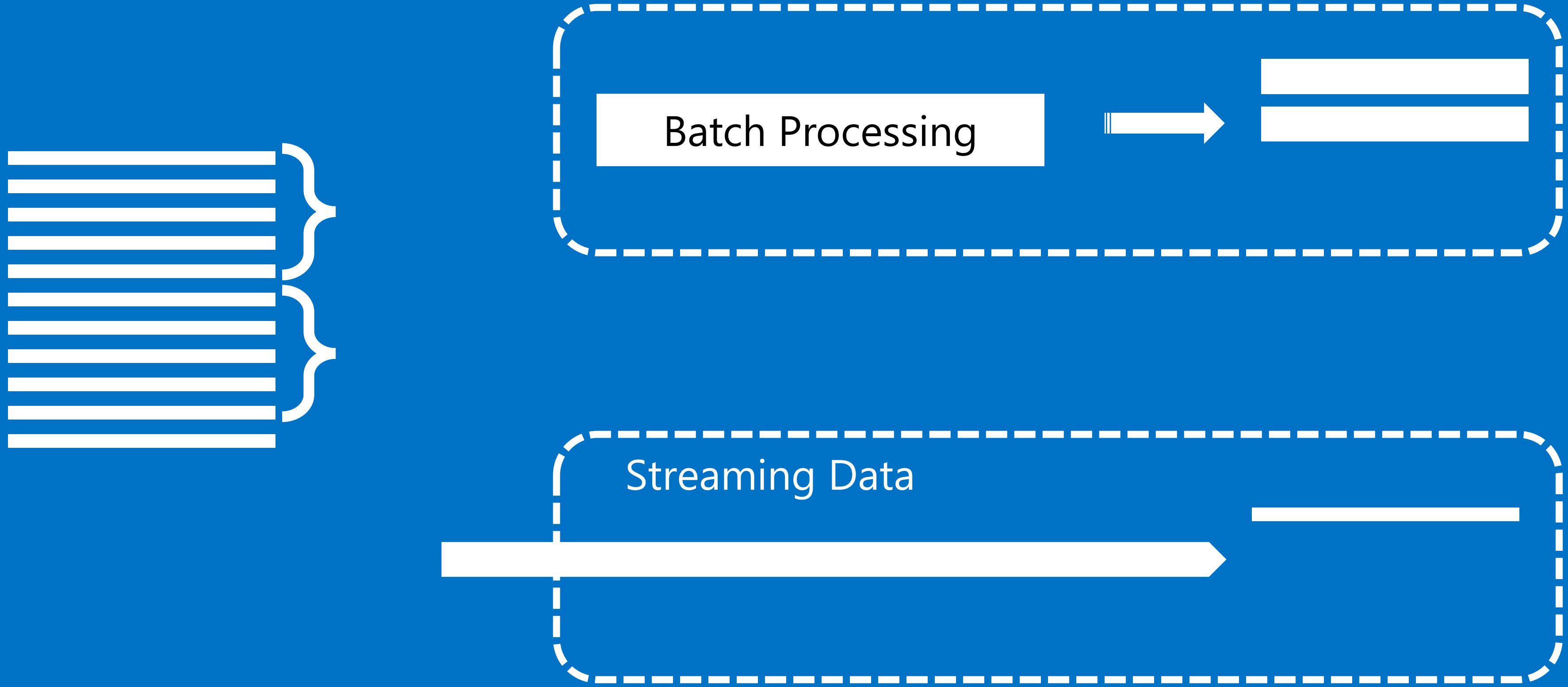

Batch Processing

Batch Processing

Batch Processing

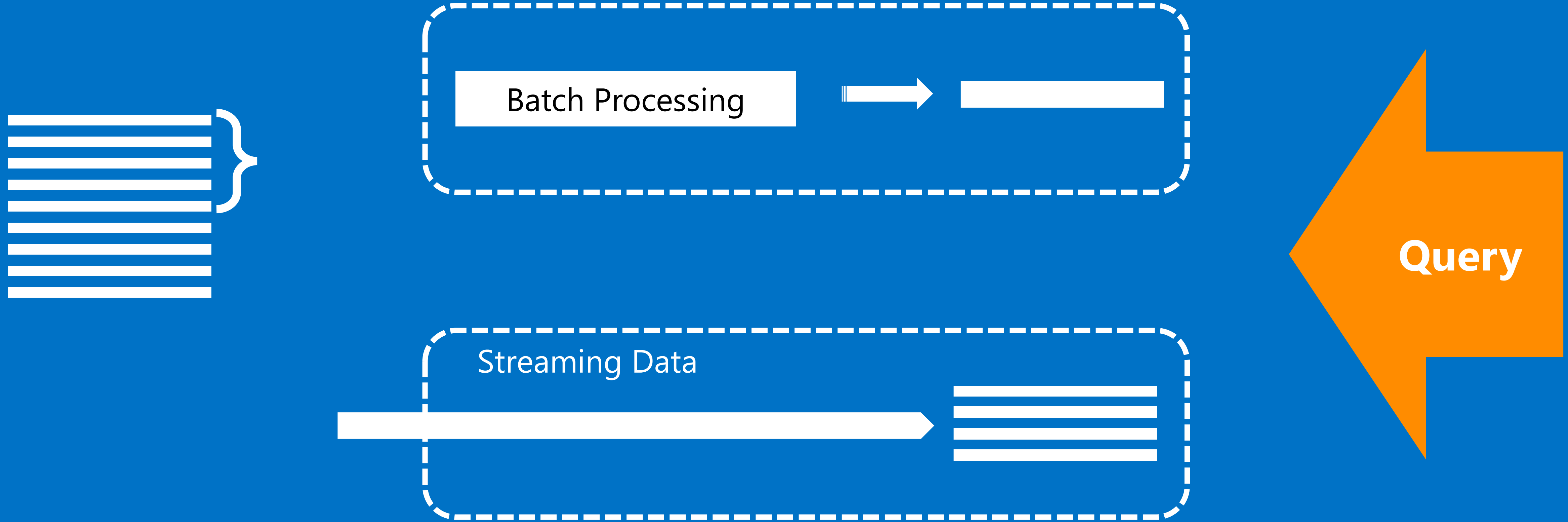
Query

SOLUTION



Batch Processing

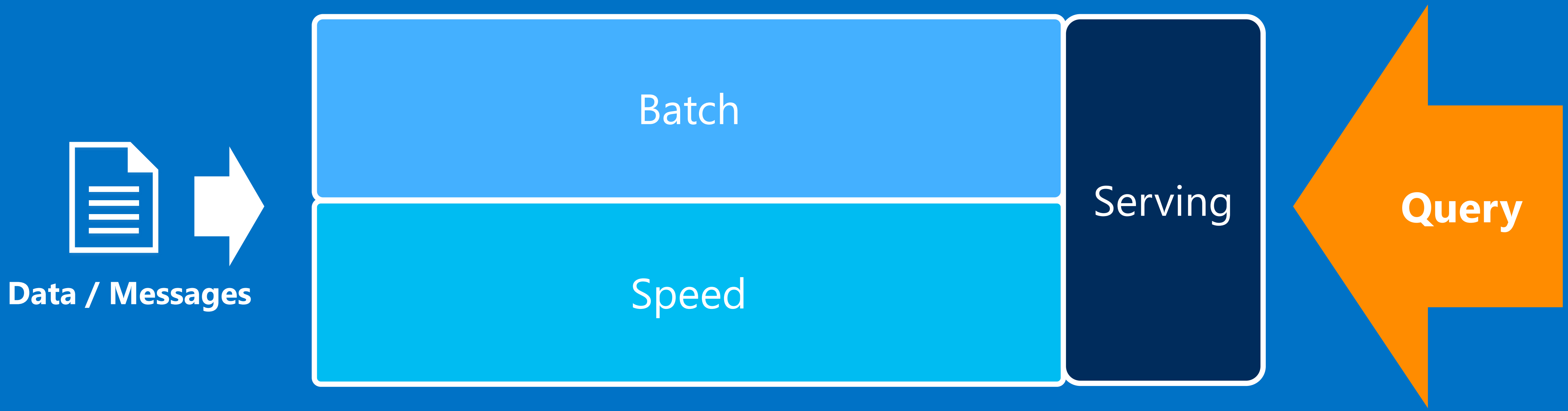
Streaming Data



Batch Processing

Streaming Data

Query



Data / Messages

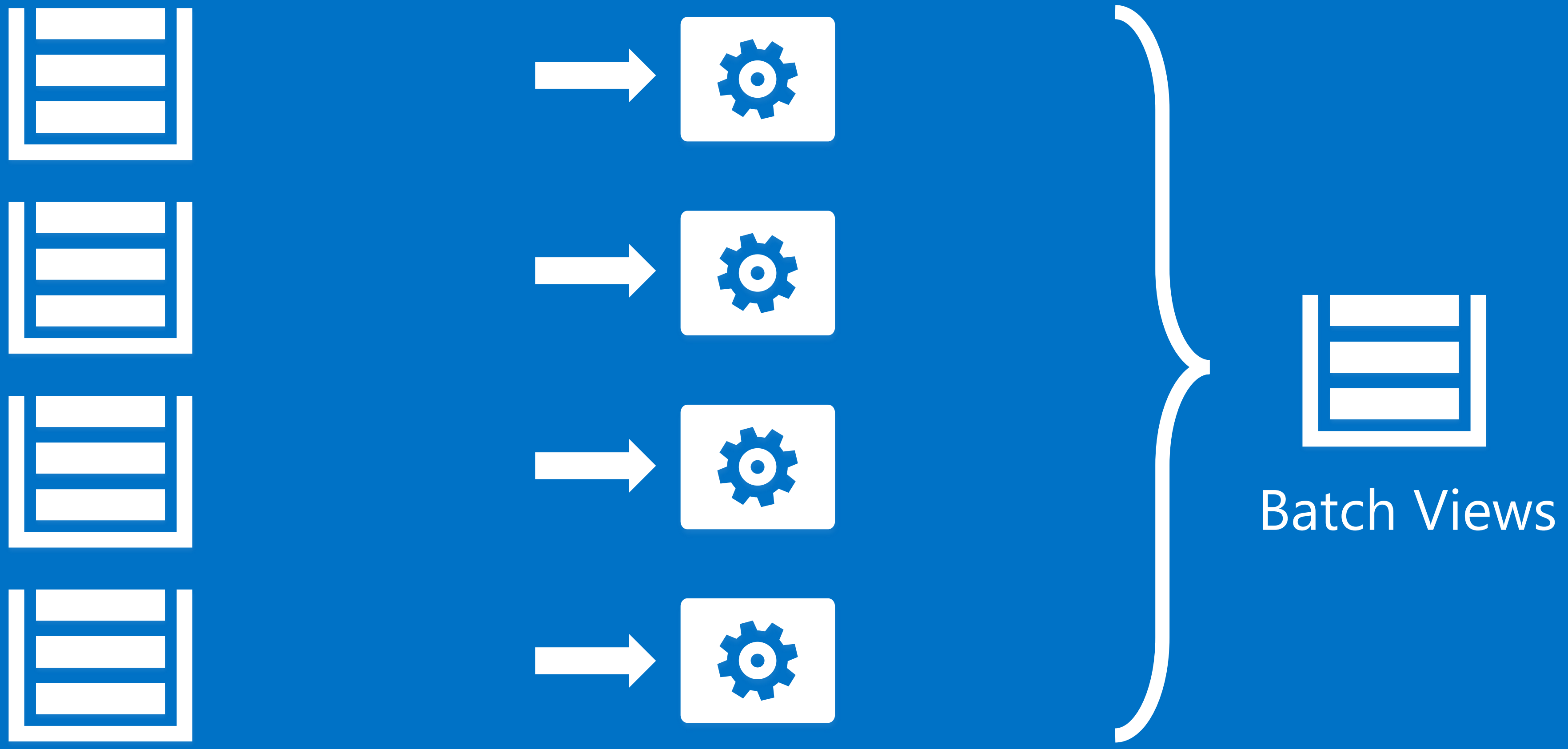
Batch

Speed

Serving

Query

Batch Layer

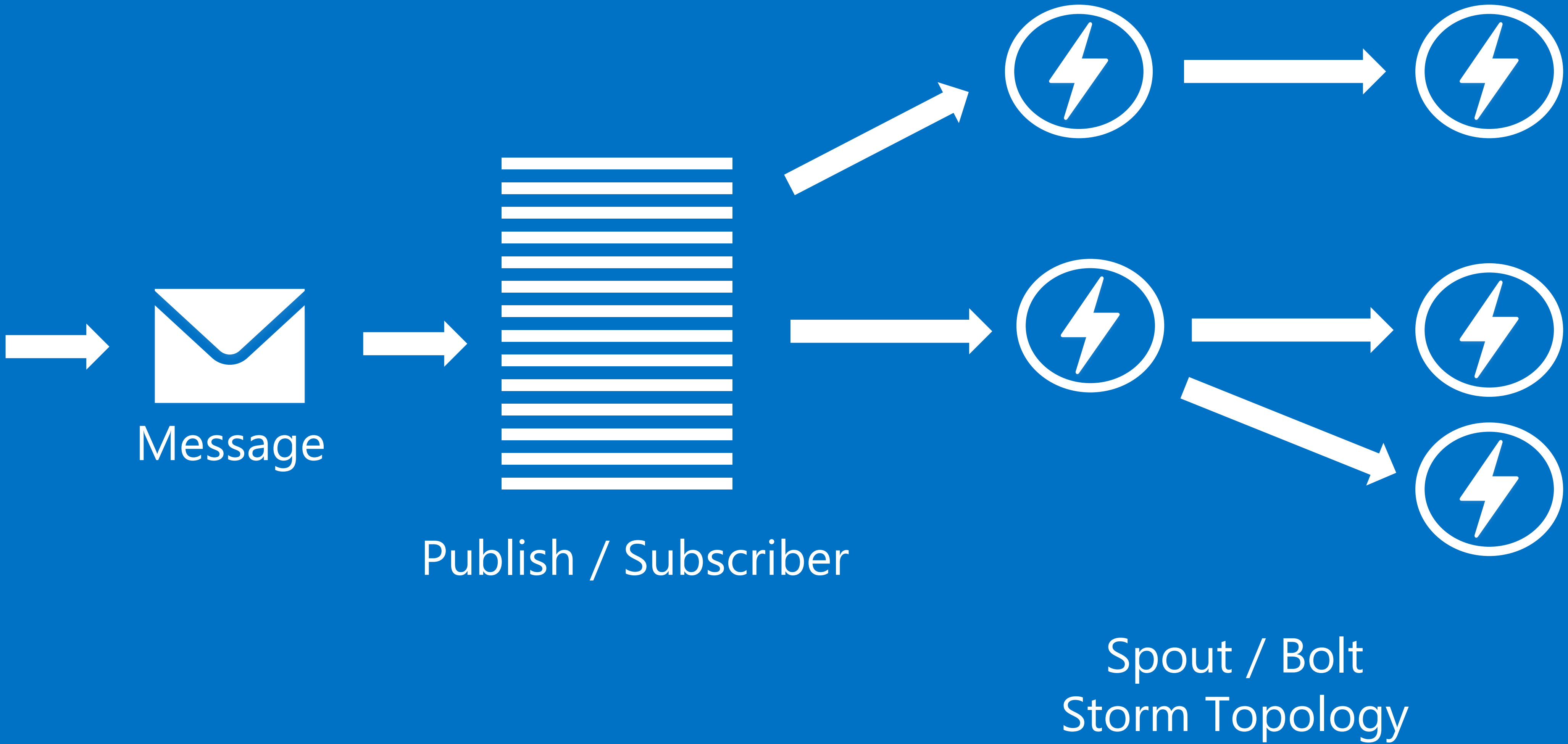


Batch Views

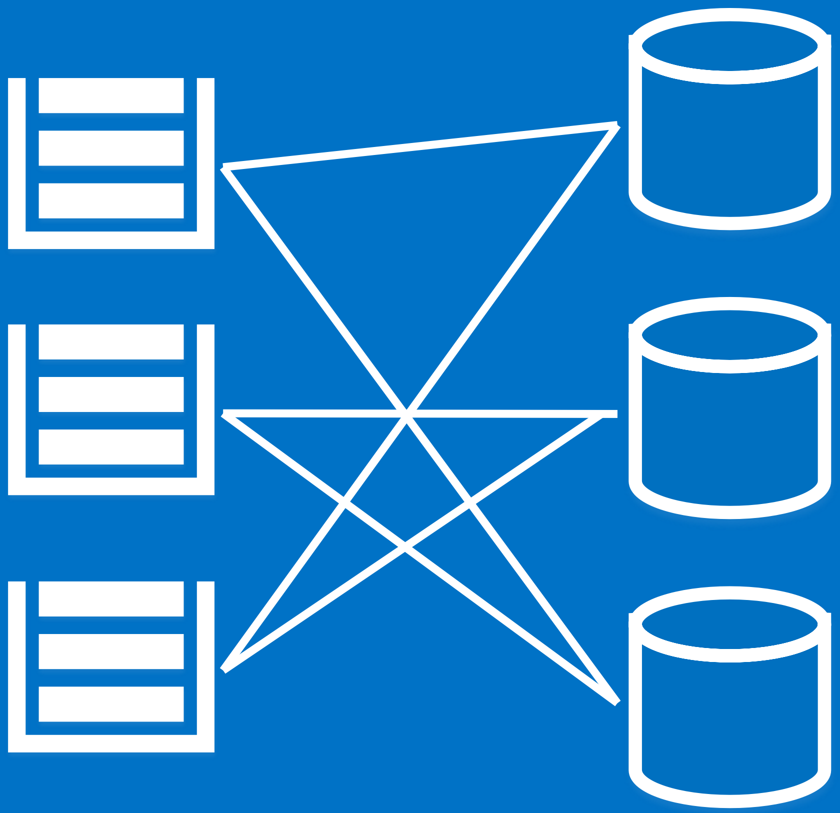
Distributed Storage
(HDFS)

Parallel Processing
(MapReduce)

Speed Layer



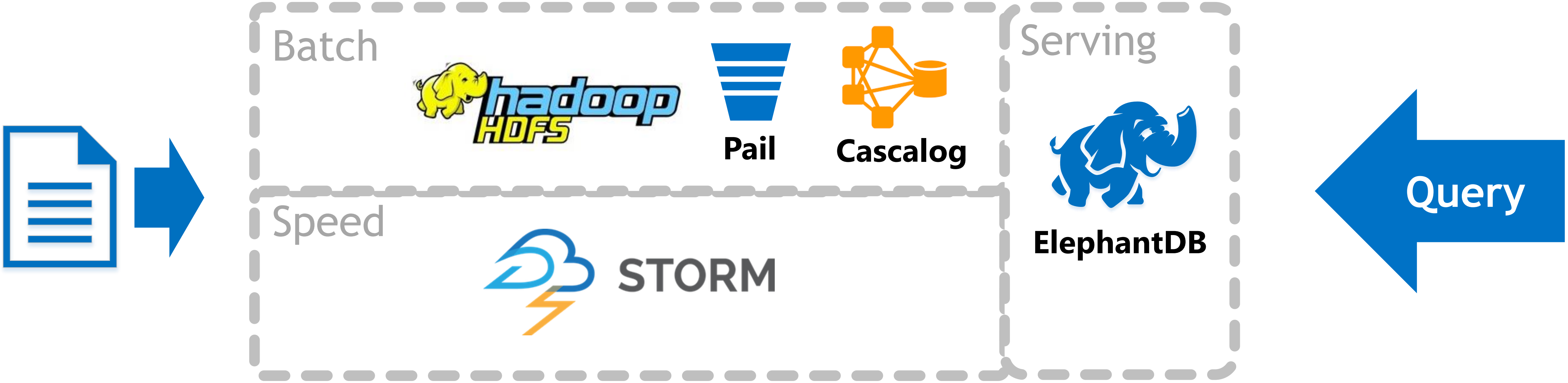
Serving Layer



Sharded Database Layer

Batch Views

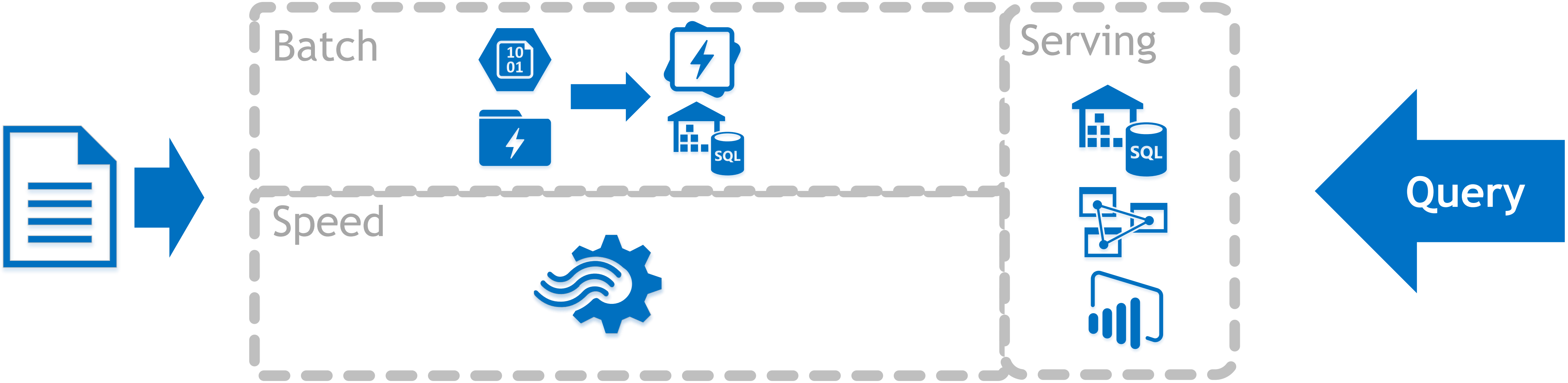
The Marz Lambda Architecture



adatis

The Azure Approach

Applying Lambda to Azure



adatis

Batch - Storage

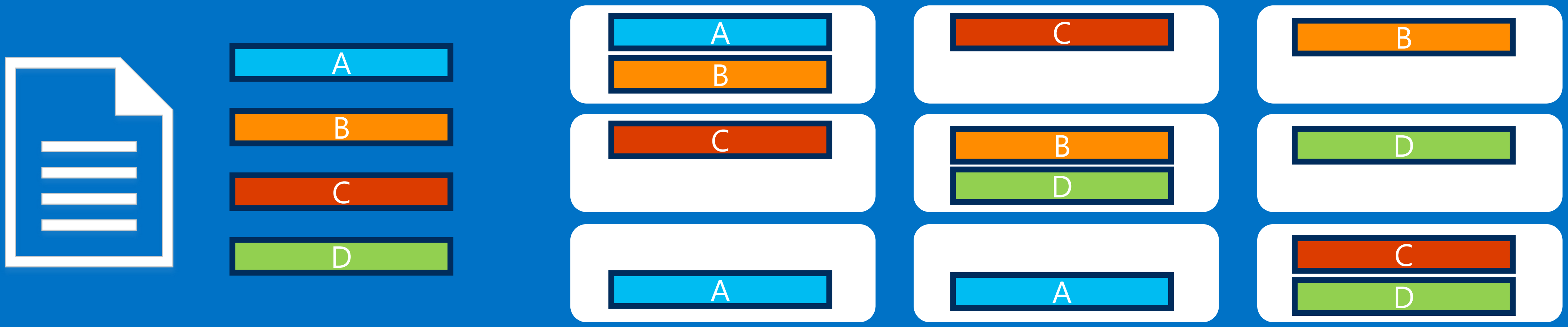
Blob Storage

- HDFS
- Hot/Cold Storage Tiers
- Limited Security
- File Size Limitations
- Widely Compatible / Available

Azure Data Lake Store

- WHDFS
- Single Pricing Model
- AAD-Integrated Security
- No Limitations
- Still Immature

Data Lake Store



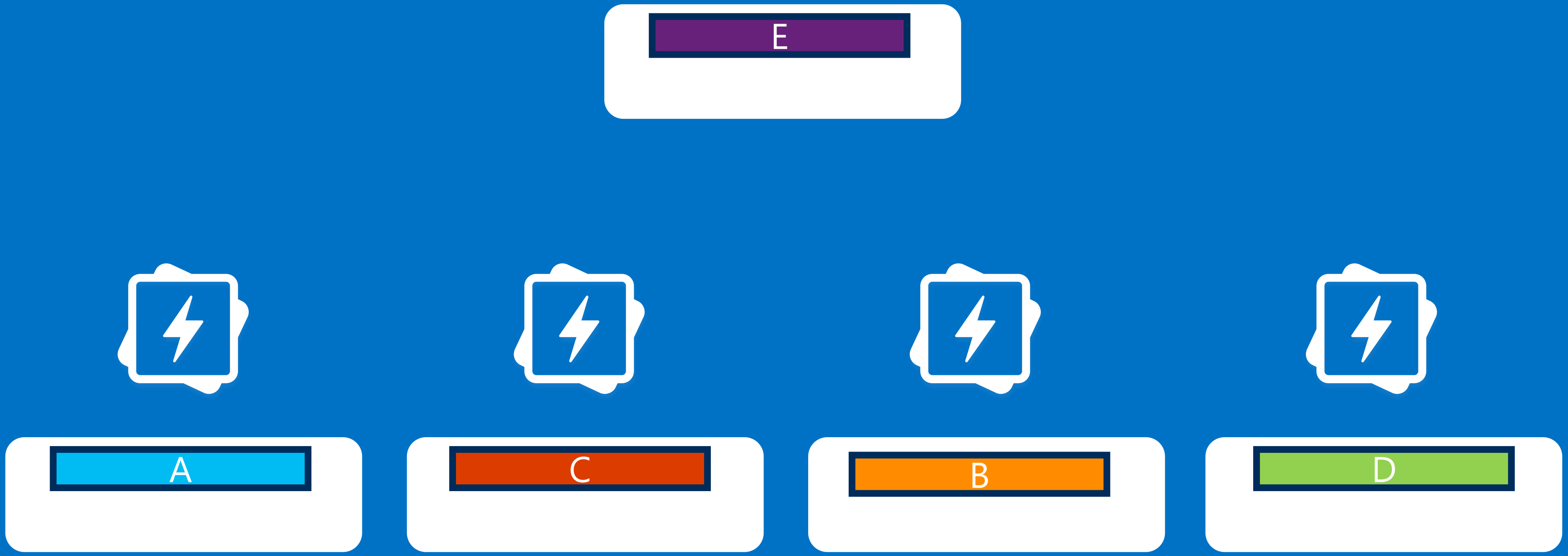
Parallelism

Fault Resilience

Batch - Compute

Azure Data Lake Analytics	Azure SQL DataWarehouse
Pay Per Query / Unit	Pay Per Hour / Node
U-SQL	T-SQL
Outputs Structured/Unstructured	Fully Structured
Uses MapReduce-style processing	Can use MapReduce via Polybase
Batch Mode	Batch or Live Query

Azure Data Lake Analytics



E

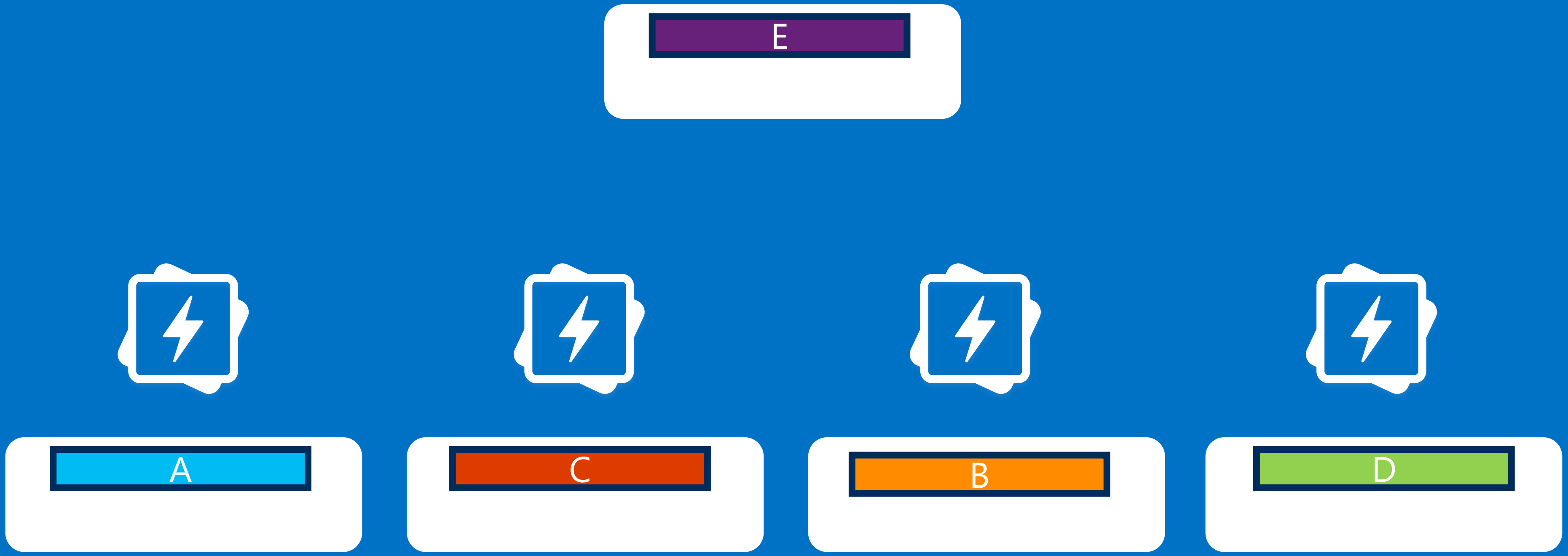
A

C

B

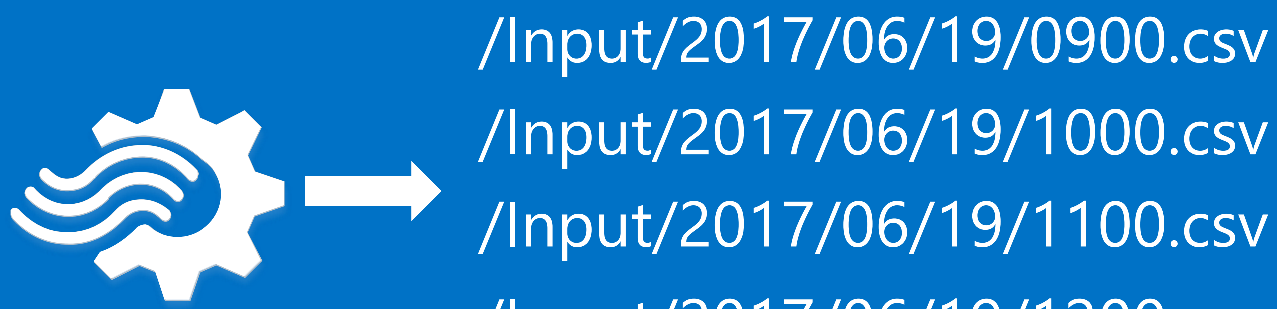
D

Azure Data Lake Analytics



Azure Streaming Analytics

- Only PaaS Native Offering
- Uses SQL Language
- Built-in Azure Integrations
- Can Vertically Partition Files
- Can Write to Multiple Outputs



/Input/2017/06/19/0900.csv

/Input/2017/06/19/1000.csv

/Input/2017/06/19/1100.csv

/Input/2017/06/19/1200.csv

Azure SQL DataWarehouse

- Low Concurrency (32!)
- Direct Query via Polybase
- Huge data capacity

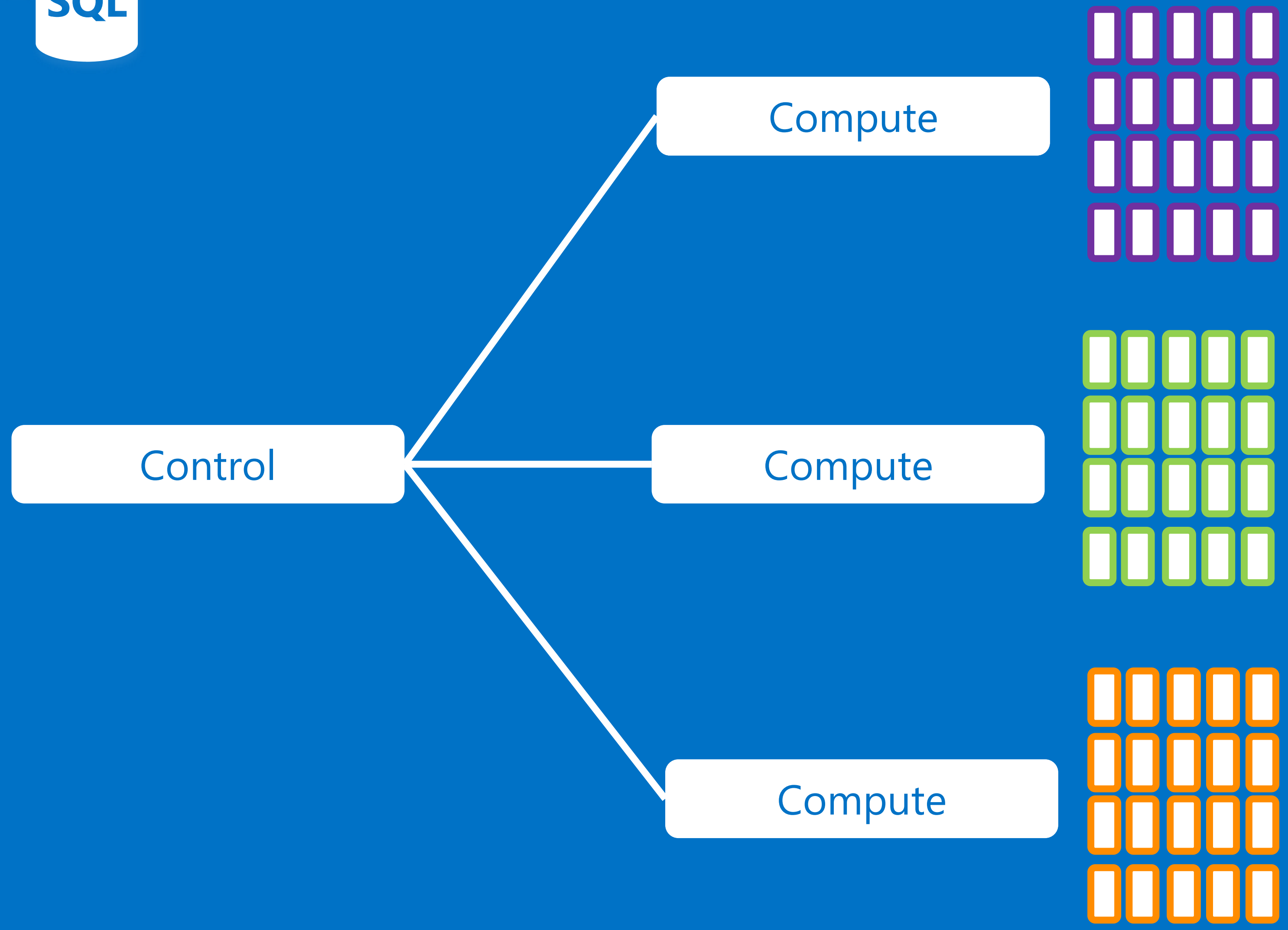
Azure Analysis Services

- High Concurrency
- Scheduled Refresh / Direct over DBs
- Model Size Limits

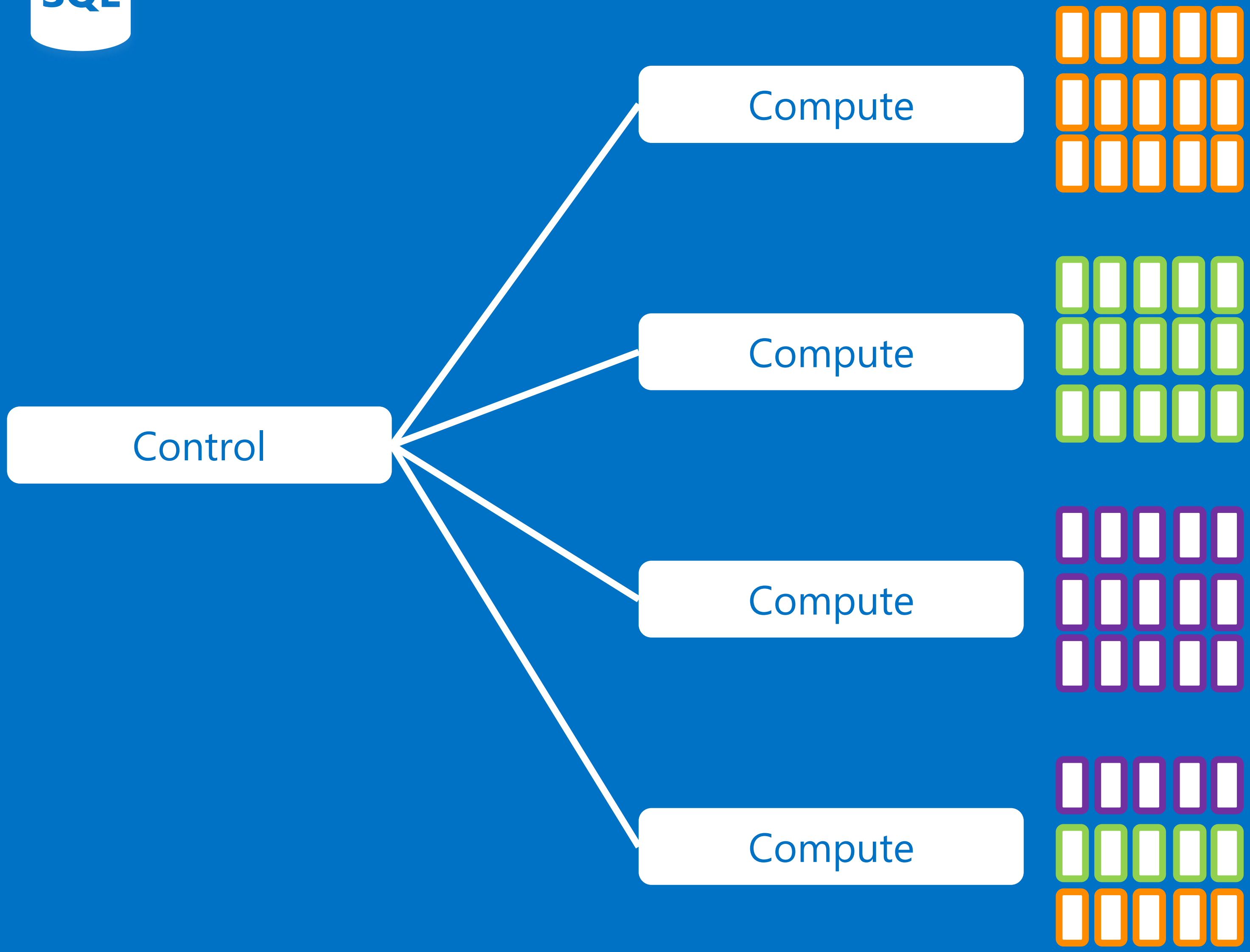
PowerBI

- High Concurrency
- Scheduled Refresh / Direct over DBs
- Model Size Limits

Azure SQL DataWarehouse



Azure SQL DataWarehouse



Compute

Compute

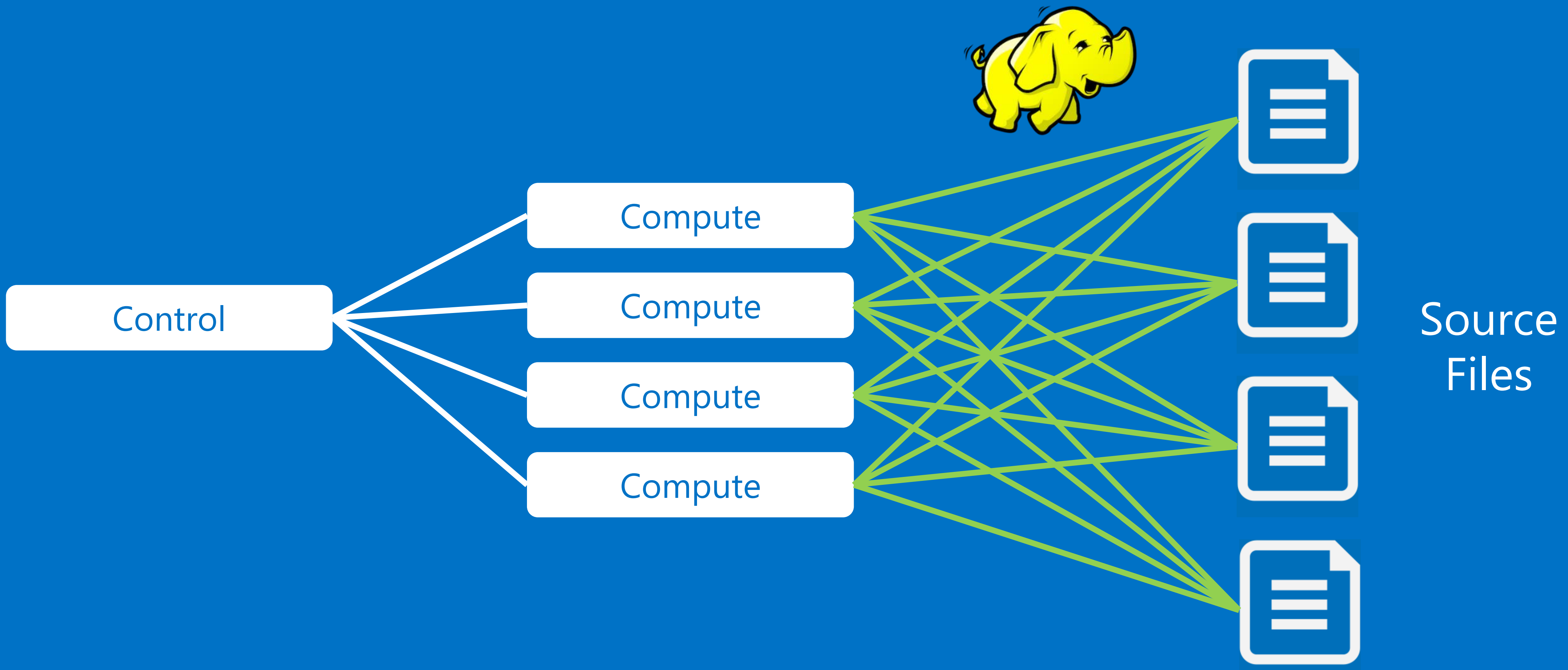
Control

Compute

Compute

Azure SQL DataWarehouse

PolyBase



Control

Compute

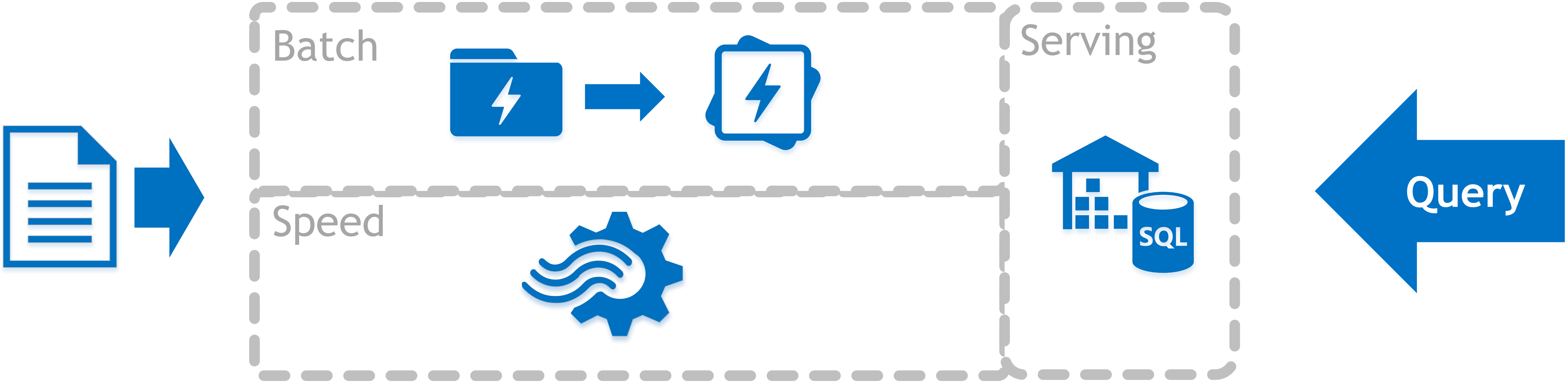
Compute

Compute

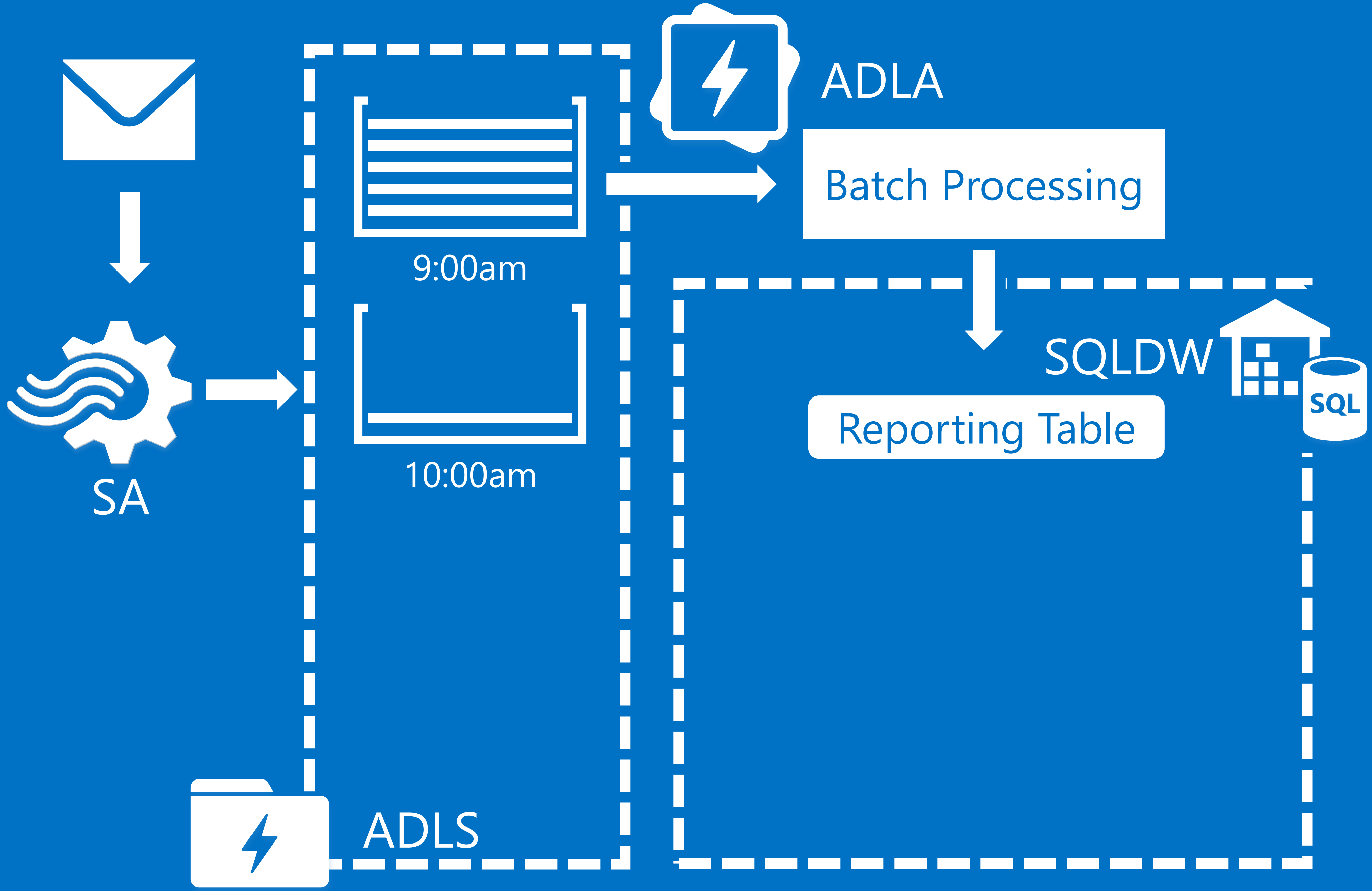
Compute

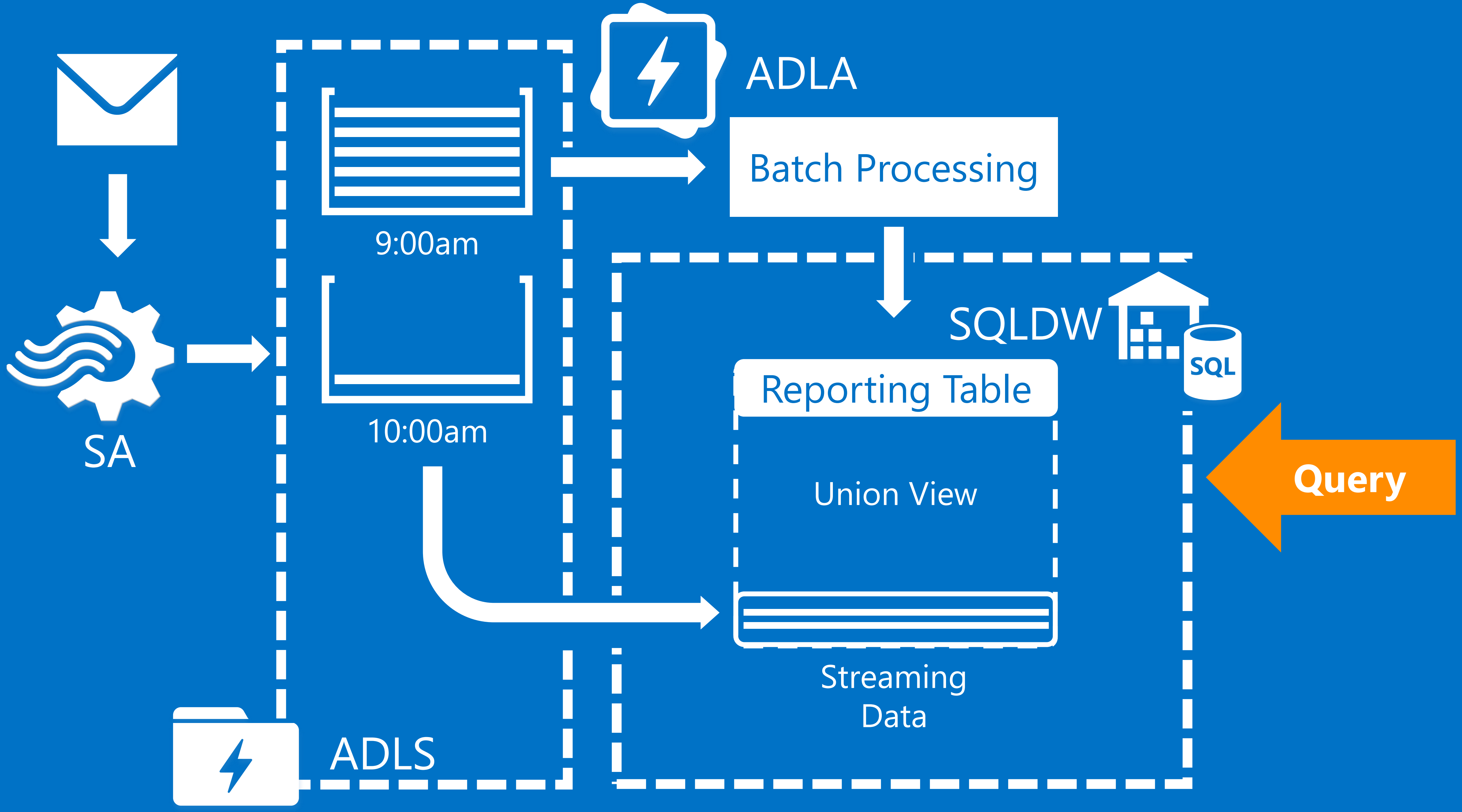
Source Files

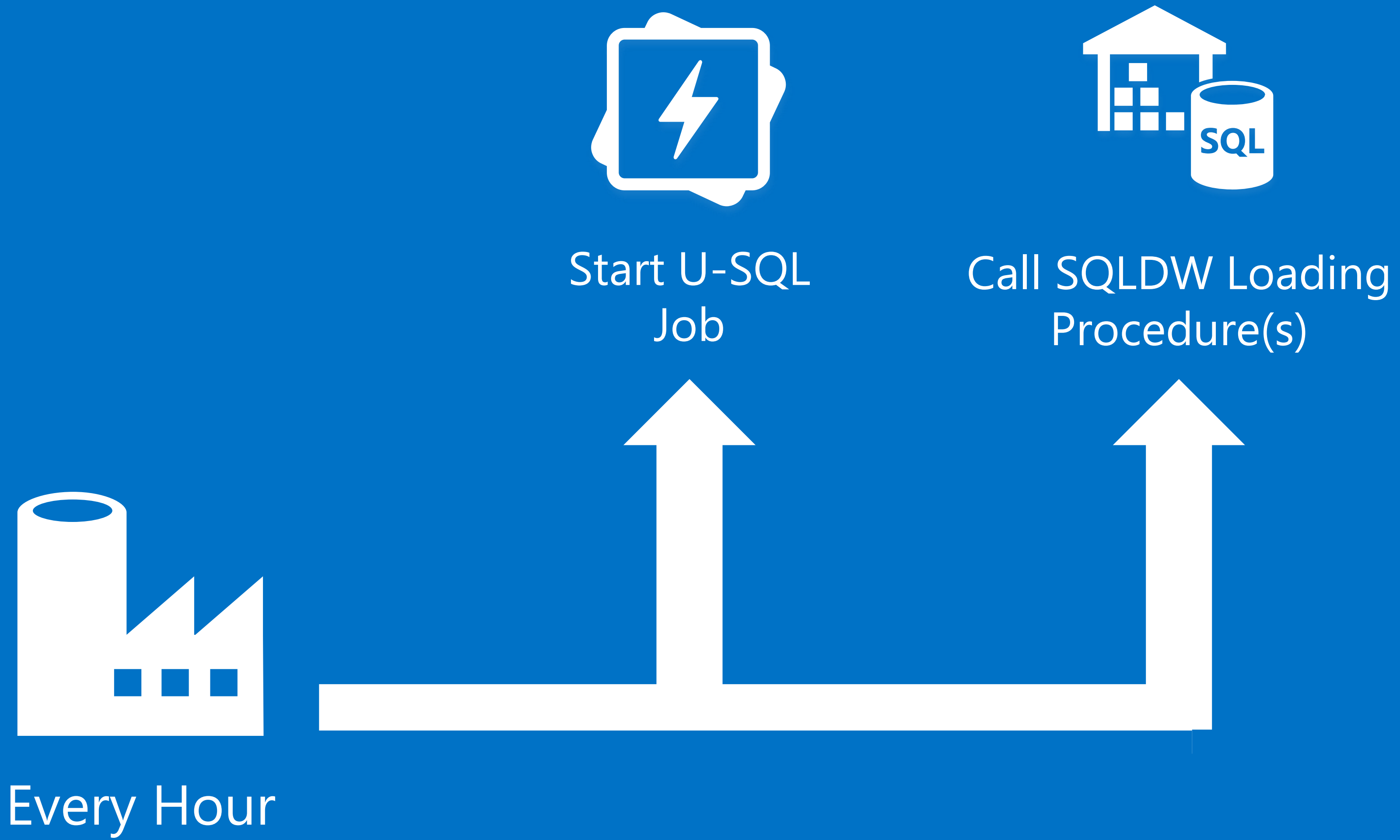
Applying Lambda to Azure



adatis


ADLA
Batch Processing
9:00am
10:00am
SA
SQLDW
SQL
Reporting Table
ADLS


ADLA
Batch Processing
9:00am
10:00am
SA
SQLDW
SQL
Reporting Table
Union View
Streaming Data
Query
ADLS



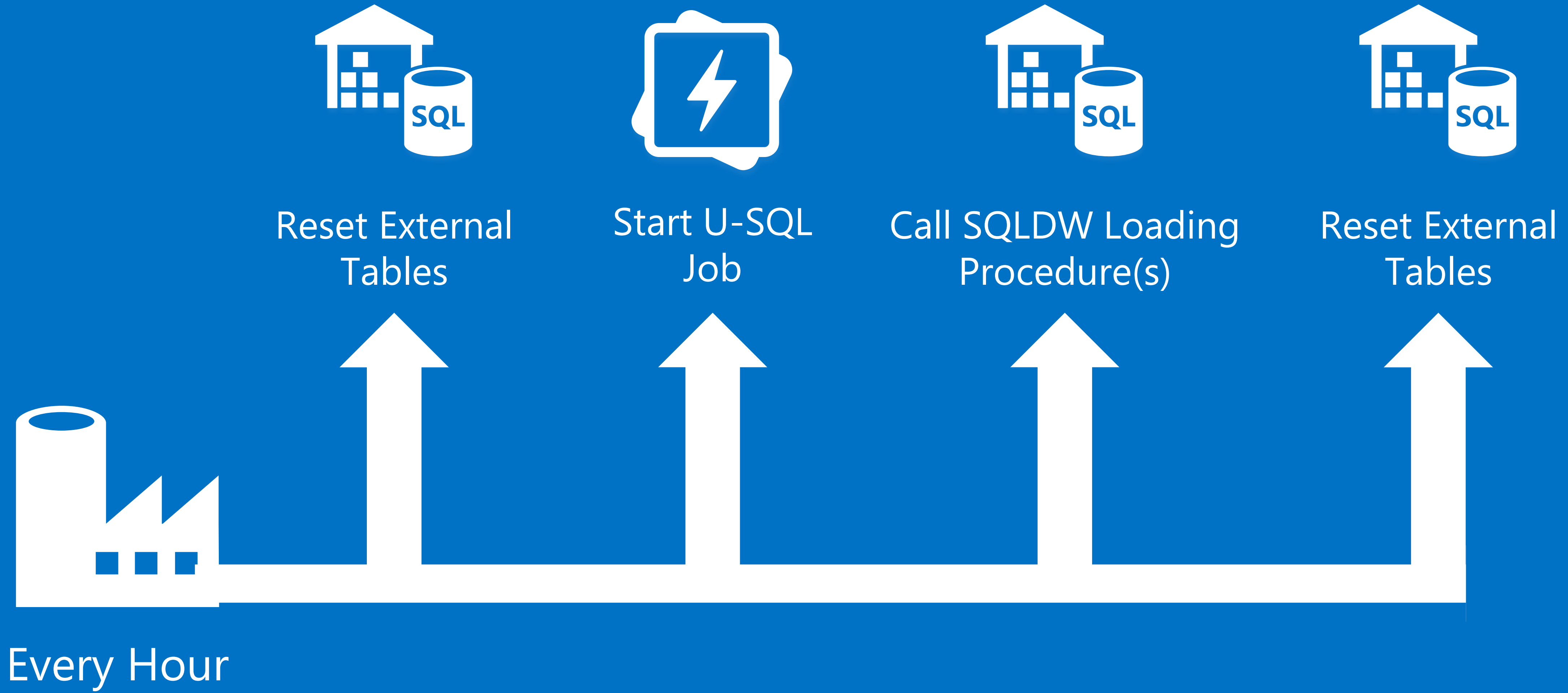
Start U-SQL Job

Call SQLDW Loading Procedure(s)

Every Hour

Azure Data Factory

Demo



Every Hour

Reset External Tables

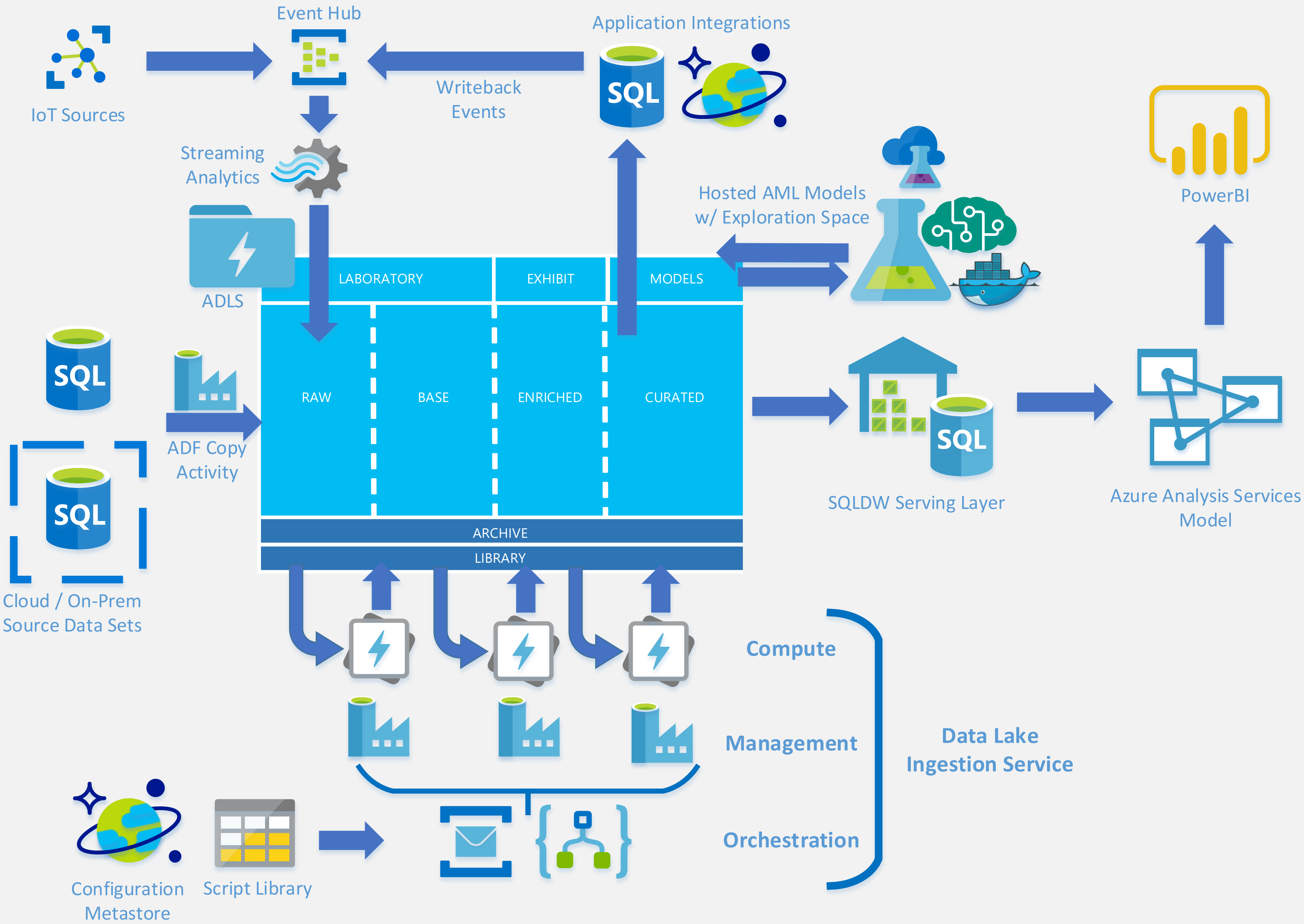
Start U-SQL Job

Call SQLDW Loading Procedure(s)

Reset External Tables

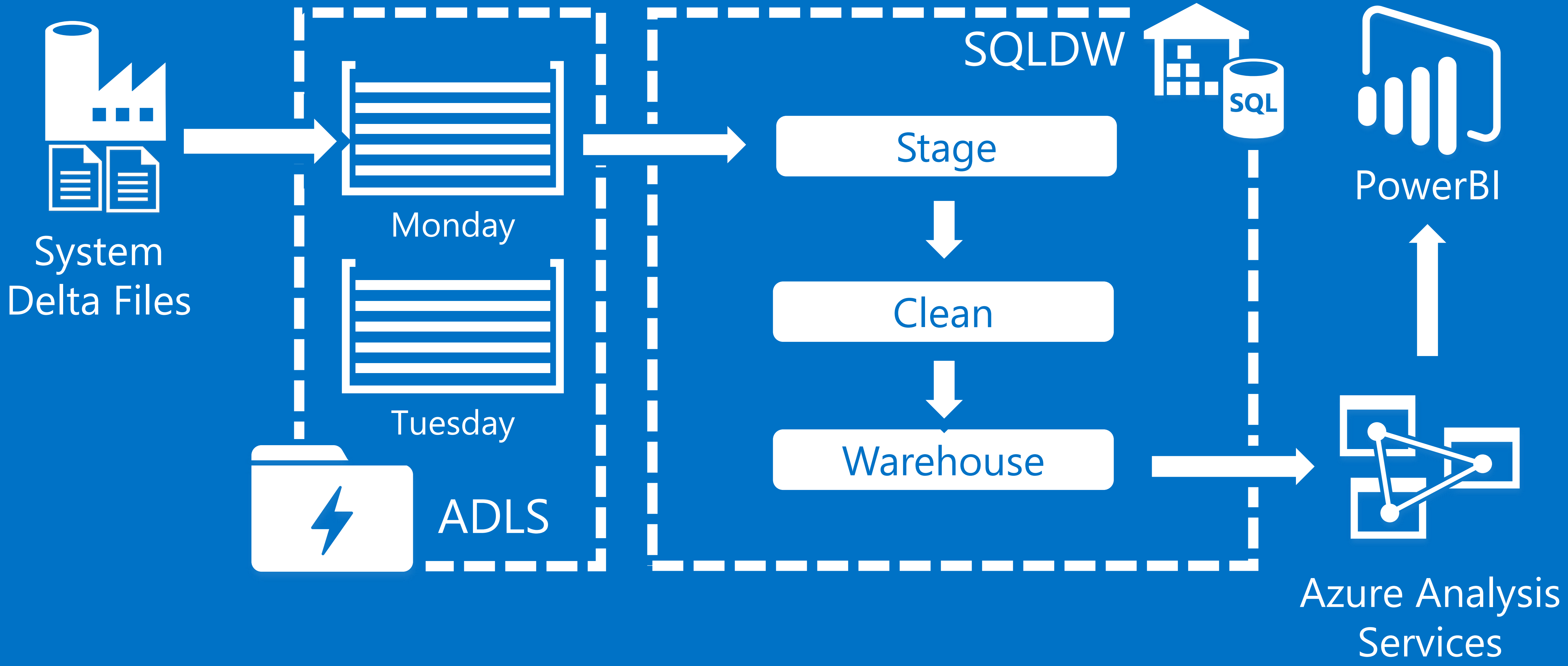
Azure Data Factory

Conclusions & Variations


Event Hub
Application Integrations
IoT Sources
Writeback Events
Streaming Analytics
Hosted AML Models w/ Exploration Space
PowerBI
ADLS
LABORATORY
EXHIBIT
MODELS
RAW
BASE
ENRICHED
CURATED
ARCHIVE
LIBRARY
ADF Copy Activity
SQLDW Serving Layer
Azure Analysis Services Model
Cloud / On-Prem Source Data Sets
Compute
Management
Orchestration
Data Lake Ingestion Service
Configuration Metastore
Script Library

adatis

Relational Batch Systems



System Delta Files

Monday

Tuesday

ADLS

SQLDW

Stage

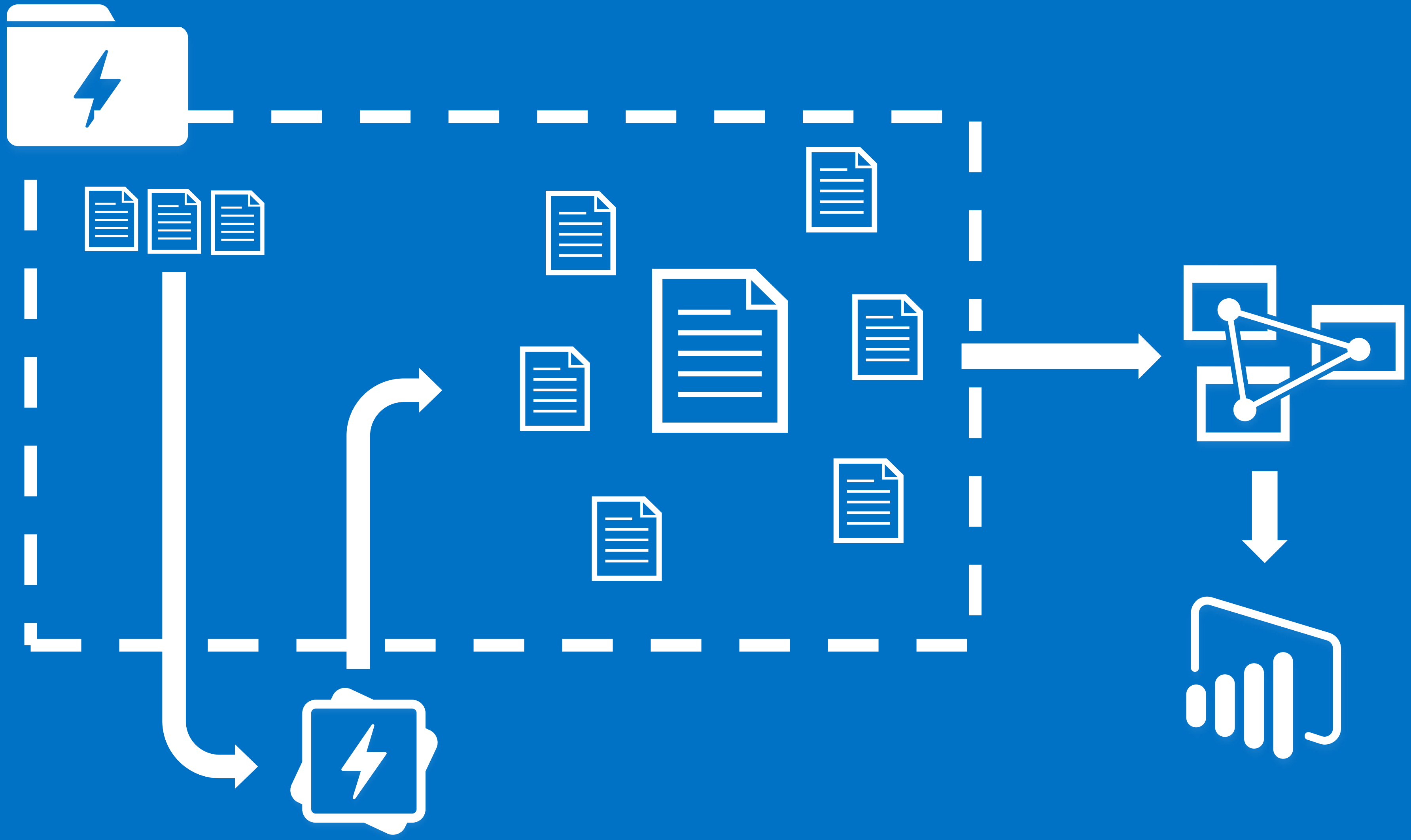
Clean

Warehouse

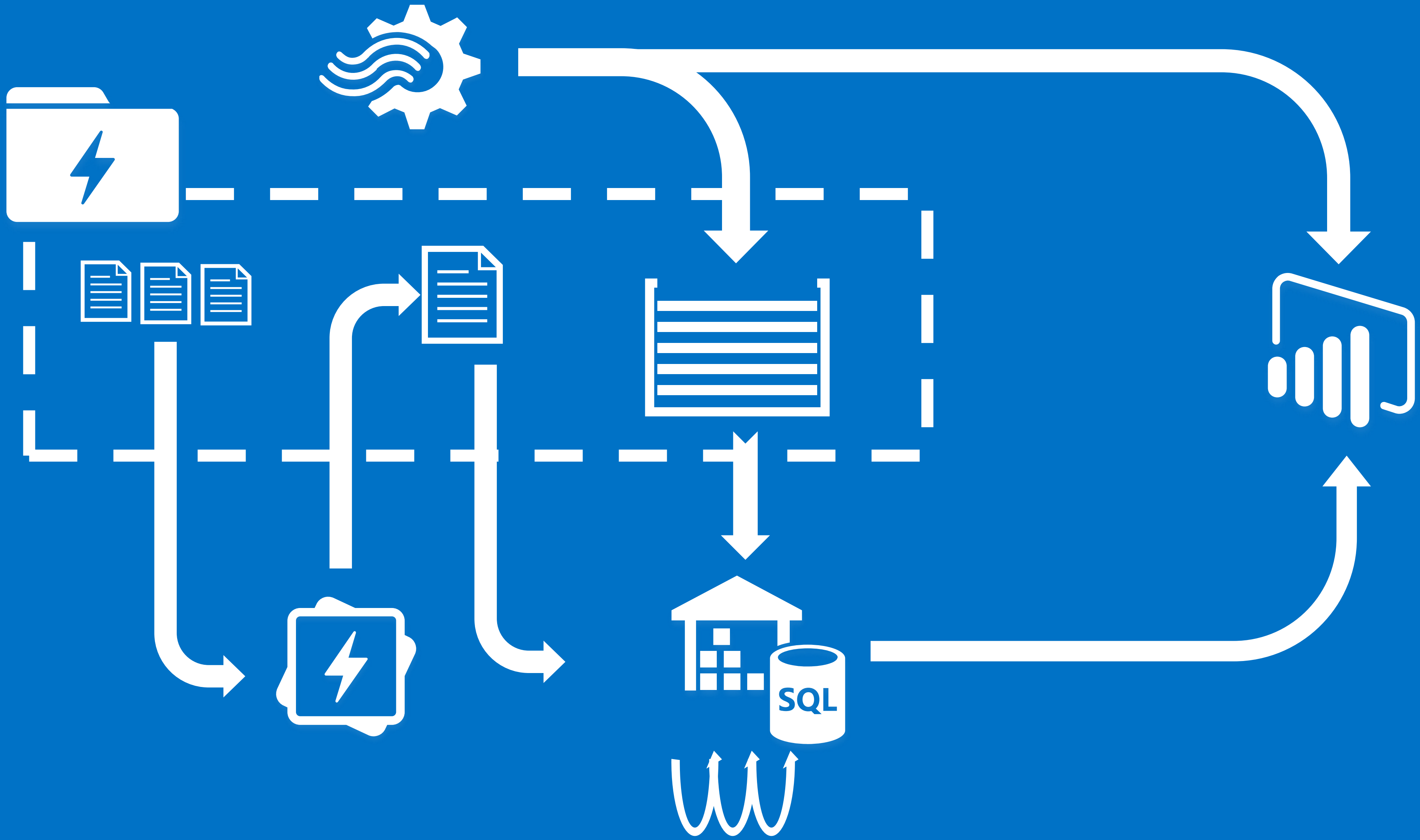
PowerBI

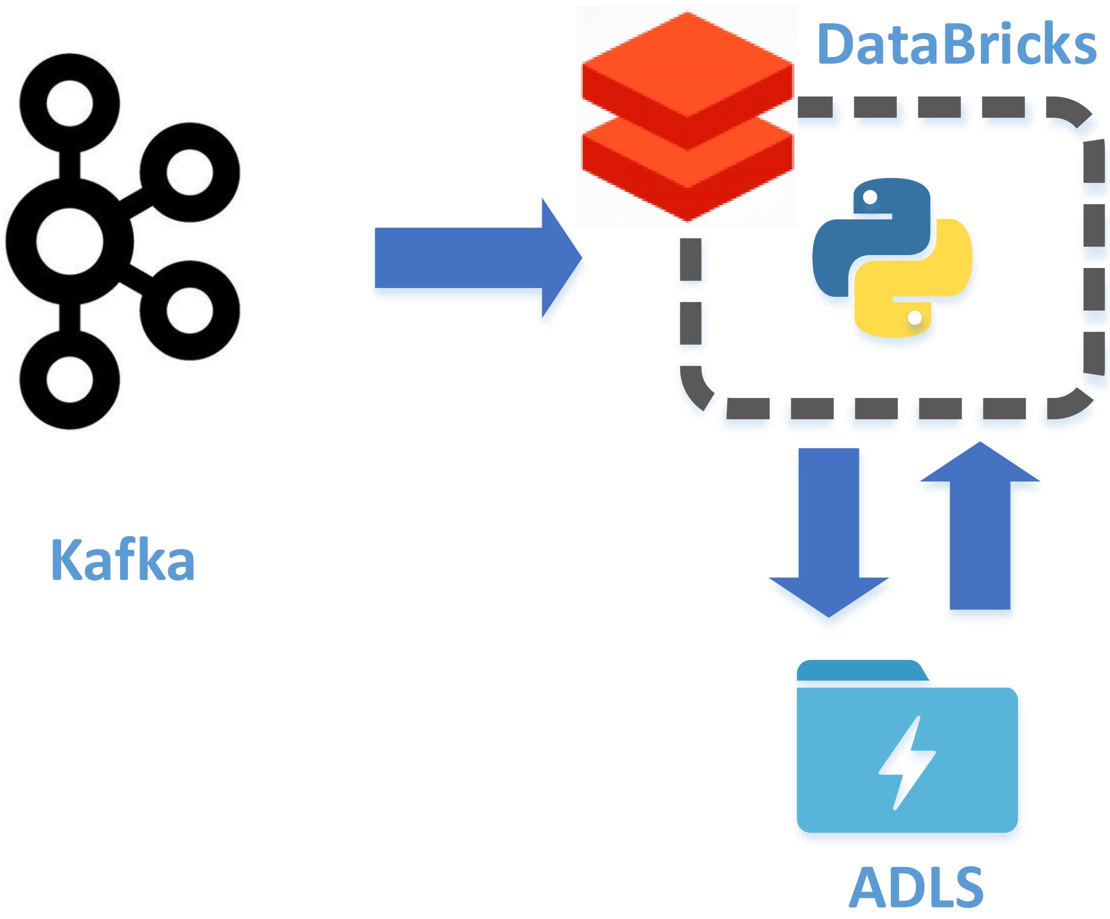
Azure Analysis Services

Who Needs a Database?



Parallel Streaming





adatis

Thanks for Listening

Simon Whiteley

@MrSiWhiteley

adatis

http://blogs.adatis.co.uk