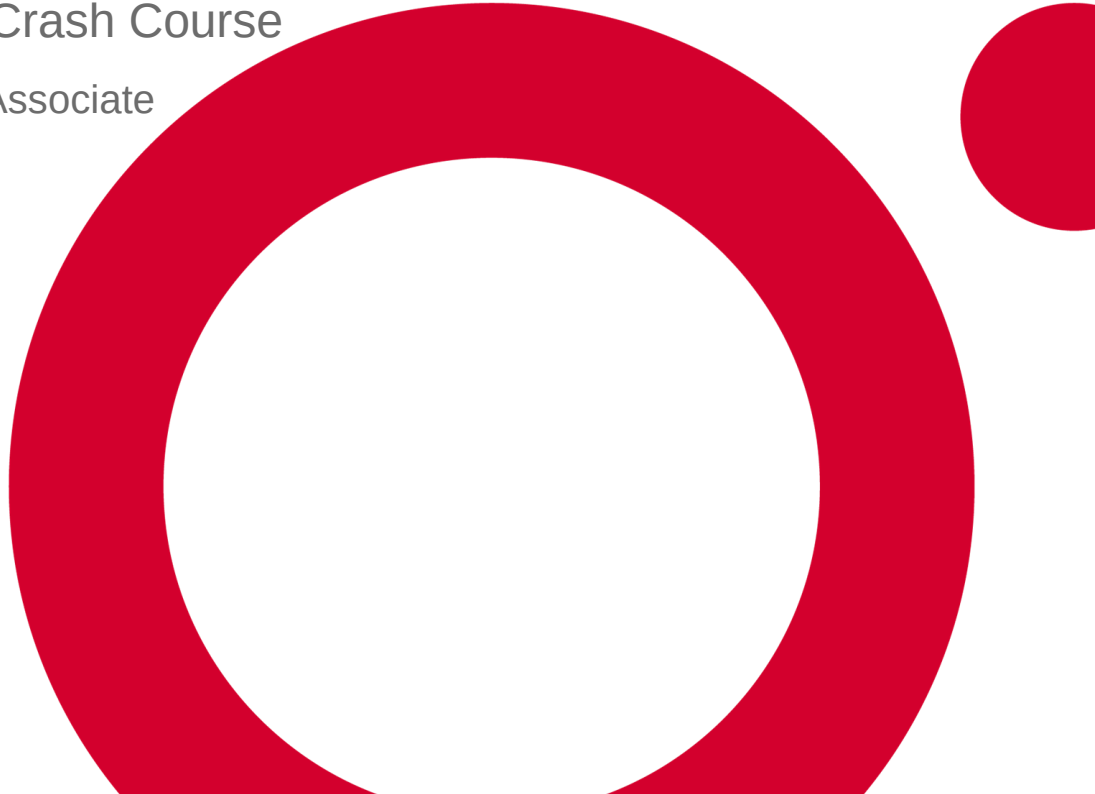
Designing an Azure Data Solution Crash Course

Microsoft Certified: Azure Data Engineer Associate



December/2020

Reza Salehi

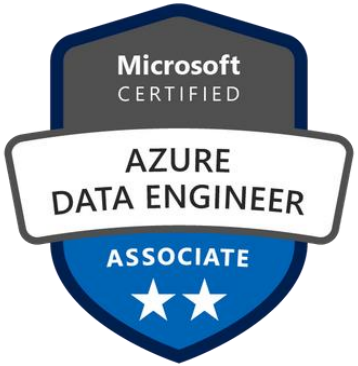
Cloud Consultant and Trainer

O'Reilly Media Instructor, Pluralsight Author

@zaalion

Microsoft CERTIFIED Trainer

2008 - 2018



Course Overview

Questions & Resources

- You can post questions in the QnA panel
- Resources are in the course repository
  - https://github.com/zaalion/oreilly-dp-200-201

- Reach out:
  - Twitter: @zaalion

DP-201 Skills Measured

Exam DP-201: Designing an Azure Data Solution skills

DP-201 Skills Measured

- Skills measured:
  - Design Azure data storage solutions (40-45%)
  - Design data processing solutions (25-30%)
  - Design for data security and compliance (25-30%)

DP-201 Candidate Profile

- Microsoft Azure data engineers
  - Collaborate with business stakeholders to identify and meet the data requirements.
  - To design data solutions that use Azure data services.
  - Design vs. implement

Azure Data Engineers

- Responsible for data-related design tasks
- Include designing Azure data storage solutions
- Use relational and non-relational data stores, batch and real-time data processing solutions
- Data security and compliance solutions

DP-201 Candidates

- Must design data solutions that use:
  - Azure Cosmos DB
  - Azure Synapse Analytics
  - Azure Data Lake Storage, Azure Data Factory, Azure Stream Analytics, Azure Databricks, and Azure Blob storage.

Design Azure Data Storage Solutions

Design Azure Data Storage Solutions

- Recommend an Azure data storage solution based on requirements
- Design non-relational cloud data stores
- Design relational cloud data stores

Big Data

Extremely large data sets which can be analyzed to find patterns and trends.

Data Lake

- A data store
- The data is raw
- The data purpose is completely clear yet
- Heterogeneous data is stored in one place (structured, unstructured, text, Access, Excel, binary, json, AVRO, etc.)
- Data is not strongly typed

Data Warehouse

- Used for reporting and data analysis
- Data is structured
- Data is generally pre-processed (e.g., using Databricks, HDInsight, ML, etc.)
- Optimized for reporting (e.g., not too many joins)
- The data purpose is clear

Recommend an Azure Data Storage Solution Based on Requirements

- Choose the correct data storage solution to meet the technical and business requirements
- Choose the partition distribution type

Choosing the Right Data Storage

- Relational databases
- Document databases
- Key/Value databases
- Graph databases
- Column family databases
- Object storage
- File share
- Data analytics databases
- Search Engine databases
- Time Series databases





Choosing the Right Data Storage

- Store logs / Azure services' output
  - Azure Blob Storage
- Low latency document database
  - Azure Cosmos DB Core API
- Database for social media
  - Azure Cosmos DB Graph API
- Migrating from MongoDB
  - Azure Cosmos MongoDB API

Choosing the Right Data Storage

- Building search around your existing data
  - Azure Cognitive Search
- Fast cache store
  - Azure Cache for Redis (Azure Redis)
- Highly relational data
  - Azure SQL Database
  - Other relational options
- Cheap column database
  - Azure Table Storage

Choosing the Right Data Storage

- Structured data
  - Azure SQL Database, MySQL, PostgreSQL, MariaDB
- Unstructured data
  - Azure Cosmos DB, Azure Table Storage
- Blobs / files
  - Azure Blob Storage, Data Lake Gen 2

Why Partition Your Data?

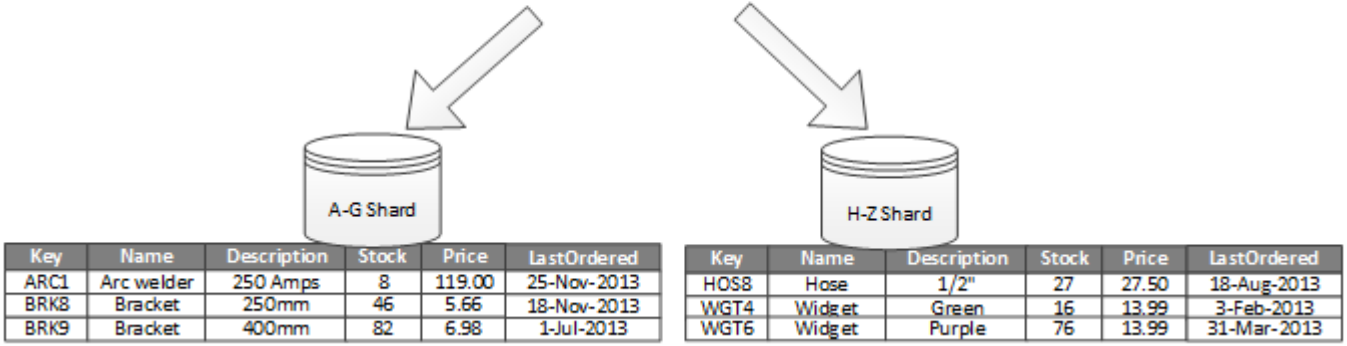
- Data partitioning
  - Improve scalability
  - Improve performance
  - Improve security
  - Provide operational flexibility
  - Match the data store to the pattern of use
  - Improve availability

Choose the Partition Distribution Type

- Data partitioning types
  - Horizontal
  - Vertical
  - Functional

Horizontal Partitioning (Sharding)

Key	Name	Description	Stock	Price	LastOrdered
ARC1	Arc welder	250 Amps	8	119.00	25-Nov-2013
BRK8	Bracket	250mm	46	5.66	18-Nov-2013
BRK9	Bracket	400mm	82	6.98	1-Jul-2013
HOS8	Hose	1/2"	27	27.50	18-Aug-2013
WGT4	Widget	Green	16	13.99	3-Feb-2013
WGT6	Widget	Purple	76	13.99	31-Mar-2013

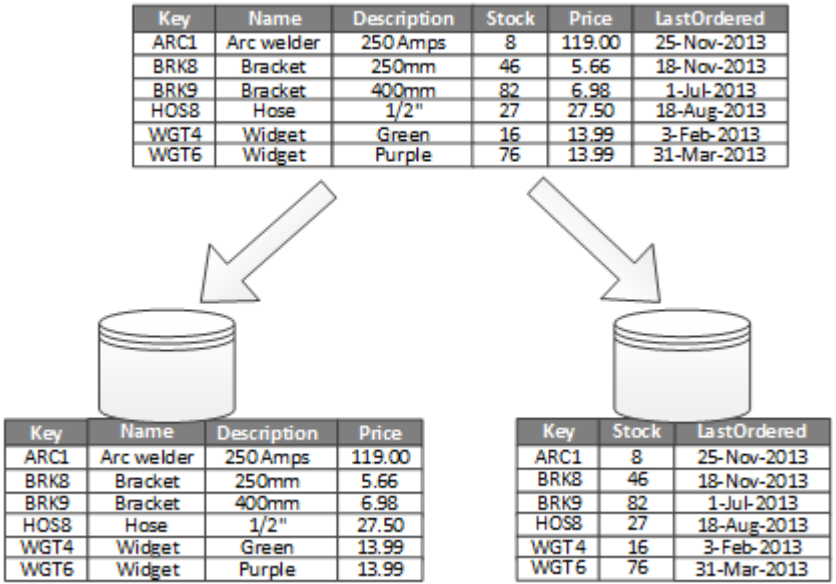


Key	Name	Description	Stock	Price	LastOrdered
ARC1	Arc welder	250 Amps	8	119.00	25-Nov-2013
BRK8	Bracket	250mm	46	5.66	18-Nov-2013
BRK9	Bracket	400mm	82	6.98	1-Jul-2013

Key	Name	Description	Stock	Price	LastOrdered
HOS8	Hose	1/2"	27	27.50	18-Aug-2013
WGT4	Widget	Green	16	13.99	3-Feb-2013
WGT6	Widget	Purple	76	13.99	31-Mar-2013

https://docs.microsoft.com/en-us/azure/architecture/best-practices/data-partitioning#horizontal-partitioning-sharding

Vertical Partitioning



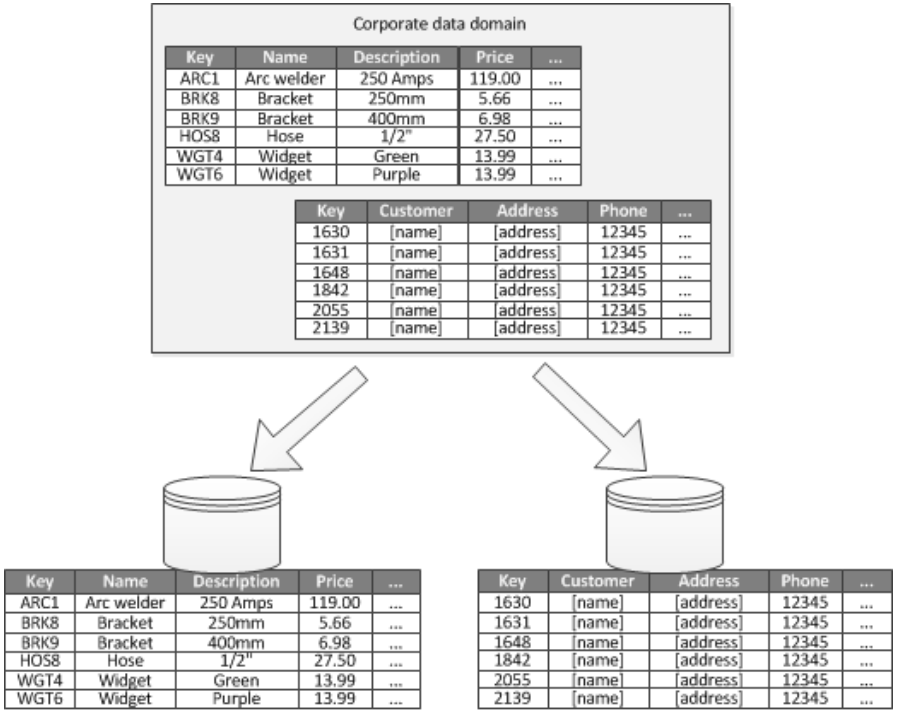
Key	Name	Description	Stock	Price	LastOrdered
ARC1	Arc welder	250 Amps	8	119.00	25-Nov-2013
BRK8	Bracket	250mm	46	5.66	18-Nov-2013
BRK9	Bracket	400mm	82	6.98	1-Jul-2013
HOS8	Hose	1/2"	27	27.50	18-Aug-2013
WGT4	Widget	Green	16	13.99	3-Feb-2013
WGT6	Widget	Purple	76	13.99	31-Mar-2013

Key	Name	Description	Price
ARC1	Arc welder	250 Amps	119.00
BRK8	Bracket	250mm	5.66
BRK9	Bracket	400mm	6.98
HOS8	Hose	1/2"	27.50
WGT4	Widget	Green	13.99
WGT6	Widget	Purple	13.99

Key	Stock	LastOrdered
ARC1	8	25-Nov-2013
BRK8	46	18-Nov-2013
BRK9	82	1-Jul-2013
HOS8	27	18-Aug-2013
WGT4	16	3-Feb-2013
WGT6	76	31-Mar-2013

https://docs.microsoft.com/en-us/azure/architecture/best-practices/data-partitioning#vertical-partitioning

Functional Partitioning

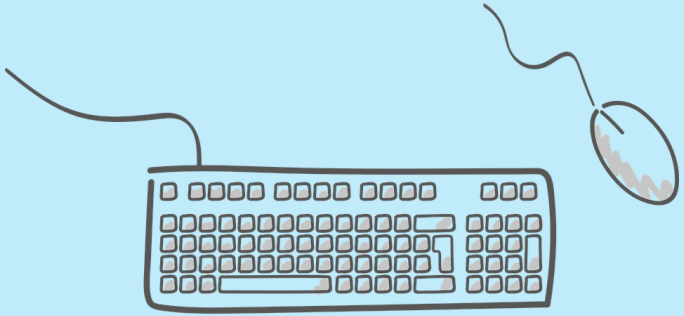


https://docs.microsoft.com/en-us/azure/architecture/best-practices/data-partitioning#functional-partitioning

Data Partitioning

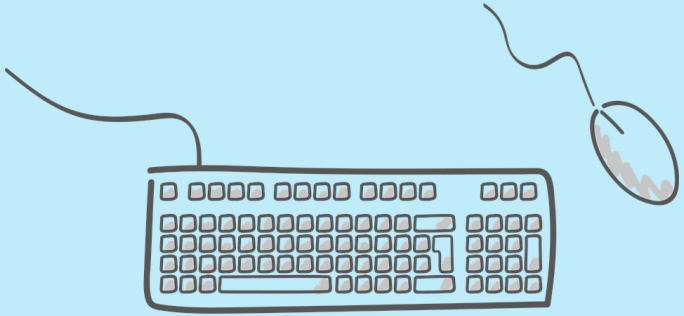
- Azure Cosmos DB
- Azure Table Storage
- Azure Blob Storage
- Azure SQL Database
- Other services

Demo



- Azure Blob Storage
- Azure Data Lake Gen 2
- Azure Table Storage
- Azure File Storage
- Azure Queue Storage

Demo



- Azure SQL Database
- Other relational options (IaaS and PaaS)

Design Non-relational Cloud Data Stores

- Design a solution that uses Cosmos DB, Data Lake Storage Gen2, or Blob storage
- Select the appropriate Cosmos DB API
- Design data distribution and partitions
- Design for scale (including multi-region, latency, and throughput)
- Design a disaster recovery strategy
- Design for high availability

NoSQL Databases

- Data is stored by means other than related tables
- Document databases
- Key-value databases
- Wide-column databases
- Graph databases

Select the Appropriate Cosmos DB API

- Cosmos DB APIs
  - Azure Cosmos DB SQL API
  - Azure Cosmos DB's API for MongoDB
  - Azure Cosmos DB Cassandra API
  - Azure Cosmos DB Gremlin API
  - Azure Cosmos DB Table API
    - vs. Azure Table Storage

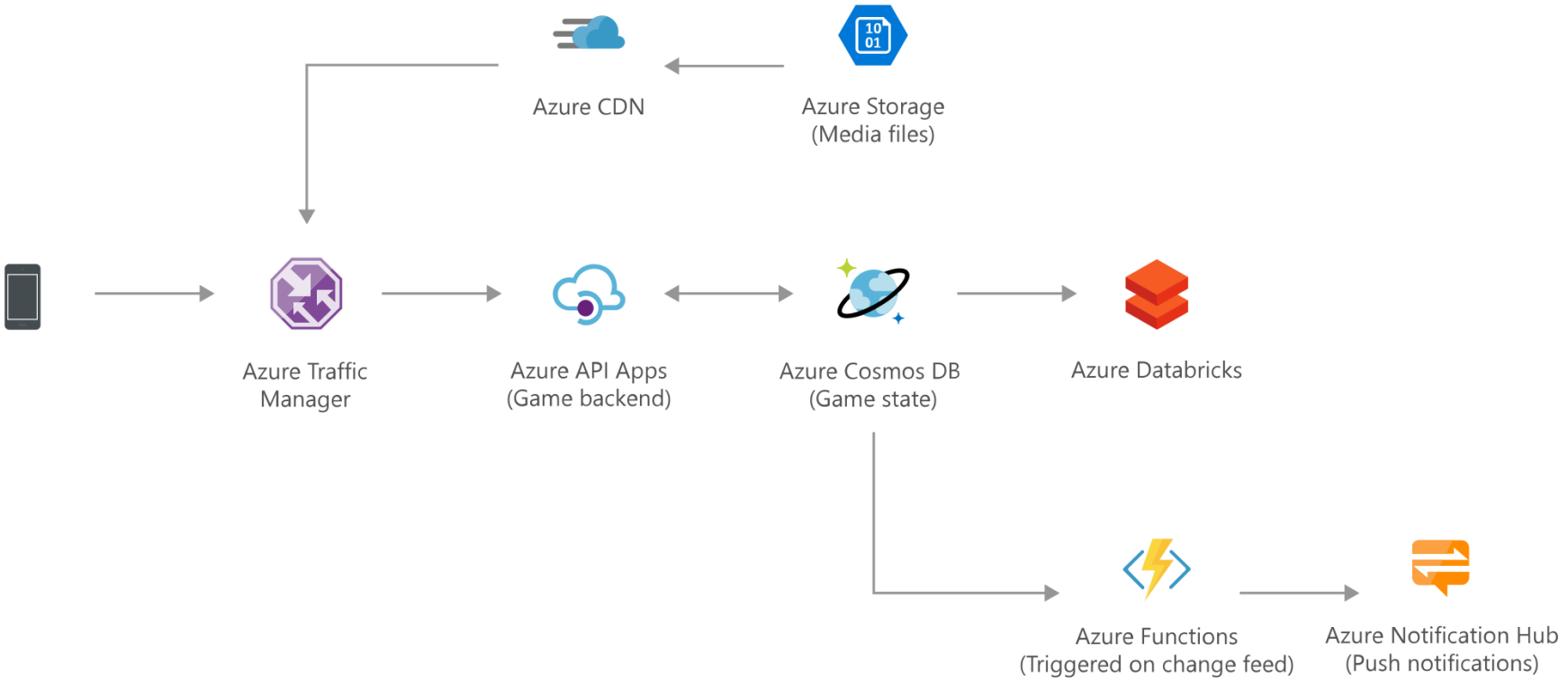
Cosmos DB Data Distribution

- Cosmos DB Data Distribution
  - Azure Cosmos DB multi-homing APIs
  - Consistency levels in Azure Cosmos DB

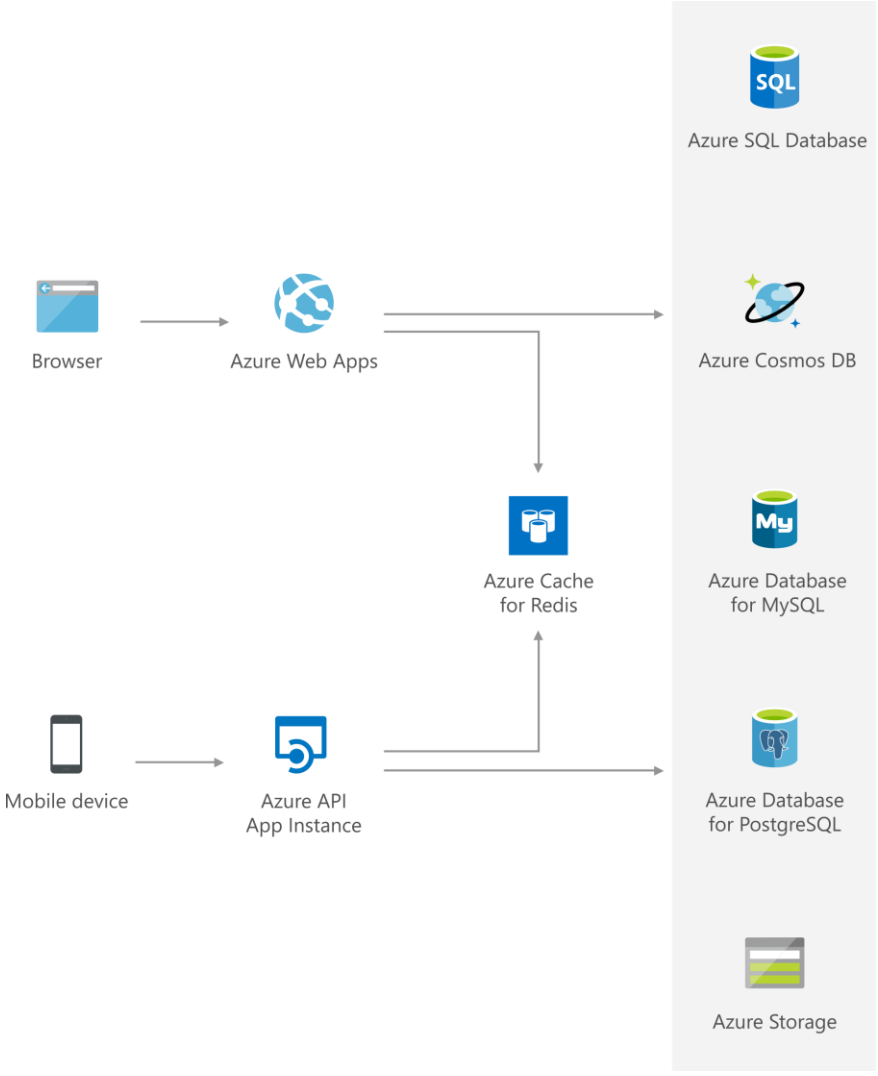
Azure Architectures

- https://docs.microsoft.com/en-us/azure/architecture/browse/#databases

Azure Storage and Cosmos DB



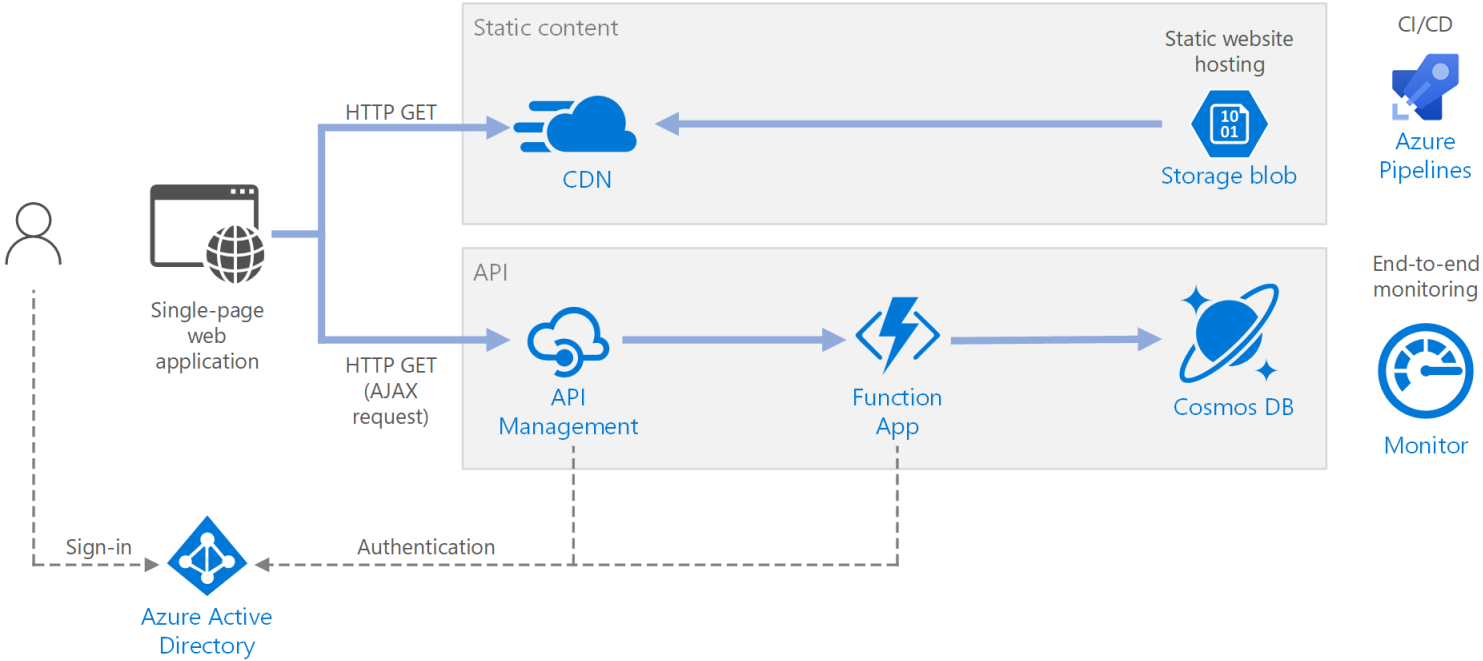
Data Cache


Azure SQL Database
Browser
Azure Web Apps
Azure Cosmos DB
Azure Cache
for Redis
Azure Database
for MySQL
Mobile device
Azure API
App Instance
Azure Database
for PostgreSQL
Azure Storage

Design a Solution That Uses Data Lake Storage Gen2 & Blobs

- https://docs.microsoft.com/en-us/azure/architecture/browse/#storage

Azure Storage



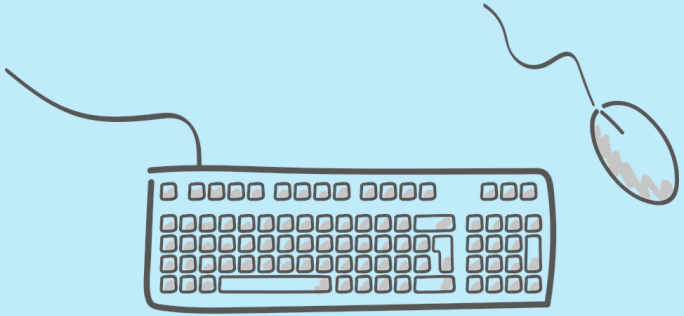
Azure Data Lake Storage Gen 2

- Is built on top of Azure blob storage, simply enable a flag
- Hadoop compatible file system, HDFS (hierarchical namespace, folders)
- Store structured and non-structured data
- The purpose of the stored data is yet to be determined
- Great features of Azure Storage such as lifecycle management and hot/cold tiers
- Used as a data source for services such as Azure Databricks and Azure Synapse

High Availability

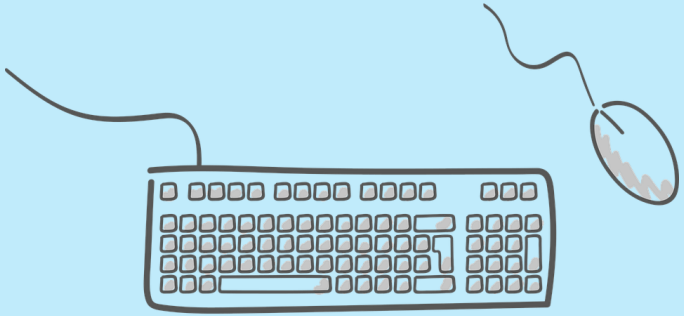
- Azure Storage Account
- Azure Cosmos DB

Demo



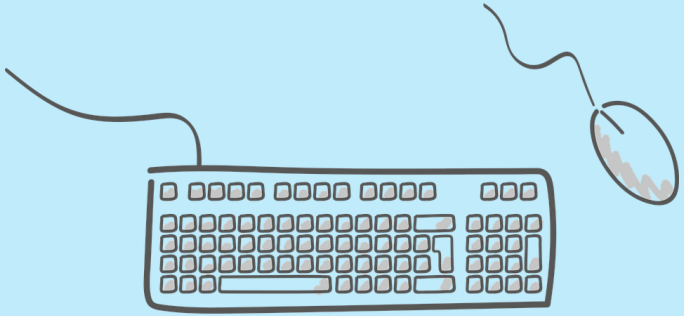
- Azure Cosmos DB
  - Previsioning
  - Multiple APIs
  - Data distribution
  - Multi-region
  - Adjusting throughput

Demo



- Azure Cosmos DB and other Azure services

Demo



- Azure Storage Account and other Azure services

Design Relational Cloud Data Stores

- Design data distribution and partitions
- Design for scale (including latency, and throughput)
- Design a solution that uses Azure Synapse Analytics
- Design a disaster recovery strategy
- Design for high availability

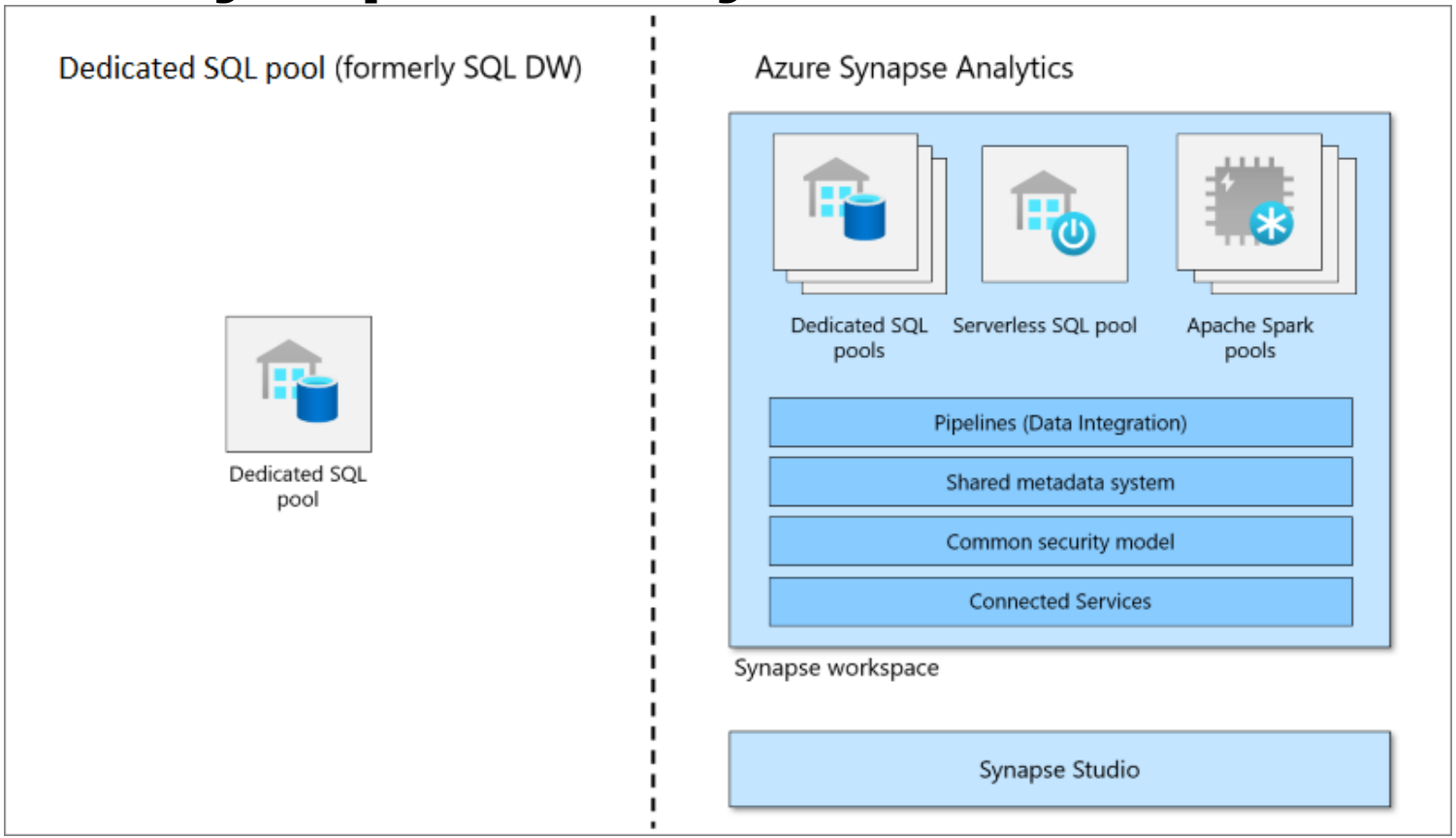
Azure Data Warehouse

- Ability to pause and resume (to save cost)
- Massive parallel processing
- Massive horizontal scaling (vs. vertical in Azure SQL Database)
- OLAP (vs. OLTP in Azure SQL Database)
- PolyBase T-SQL queries

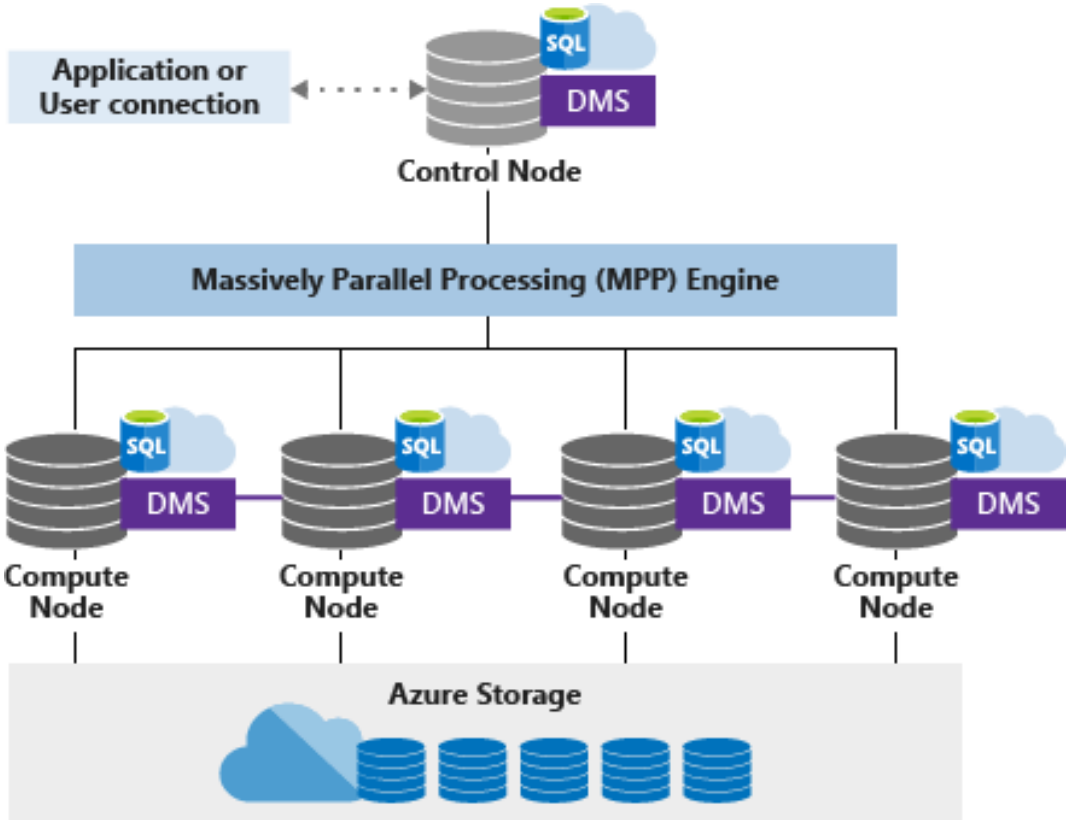
Azure Synapse Analytics

- Components:
  - Synapse SQL: Complete T-SQL based analytics
    - Dedicated SQL pool (pay per DWU provisioned)
    - Serverless SQL pool (pay per TB processed)
  - Spark: Deeply integrated Apache Spark
  - Synapse Pipelines: Hybrid data integration (e.g., Azure Data Factory)
  - Studio: Unified user experience

Azure Synapse Analytics



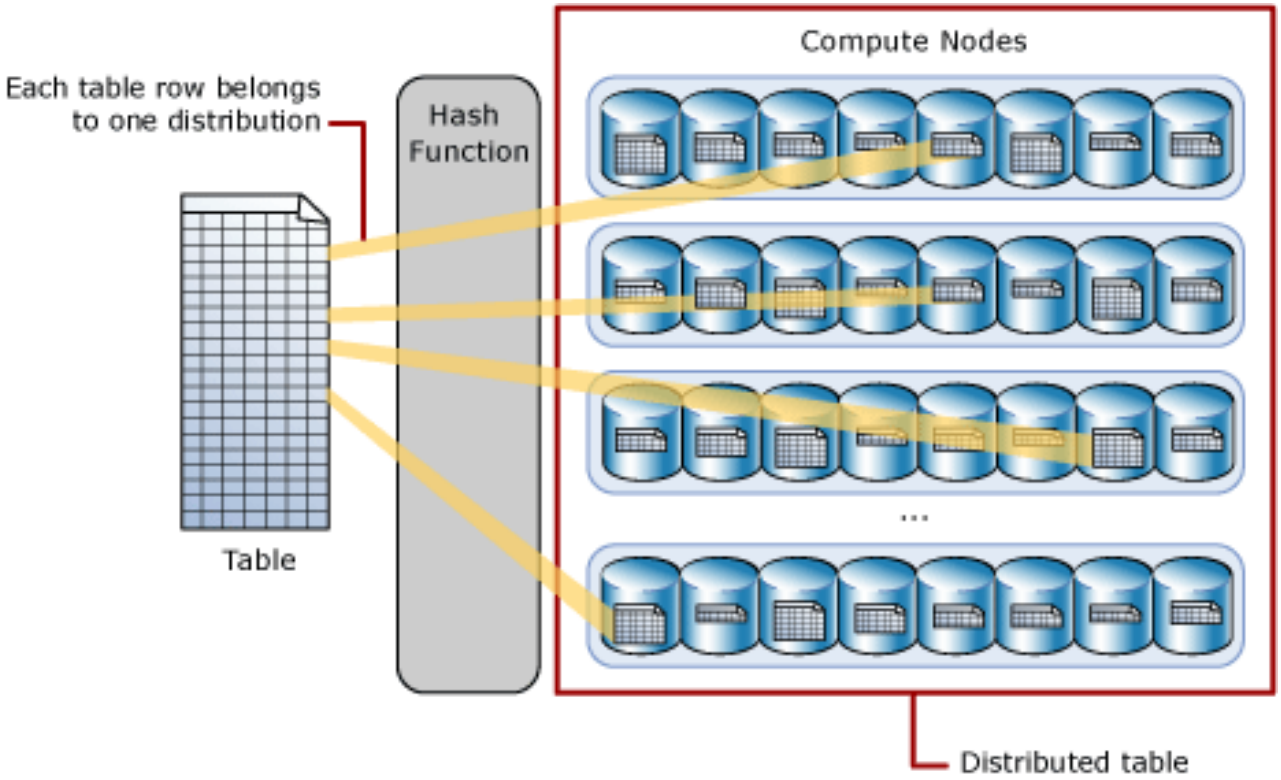
Azure Synapse Analytics



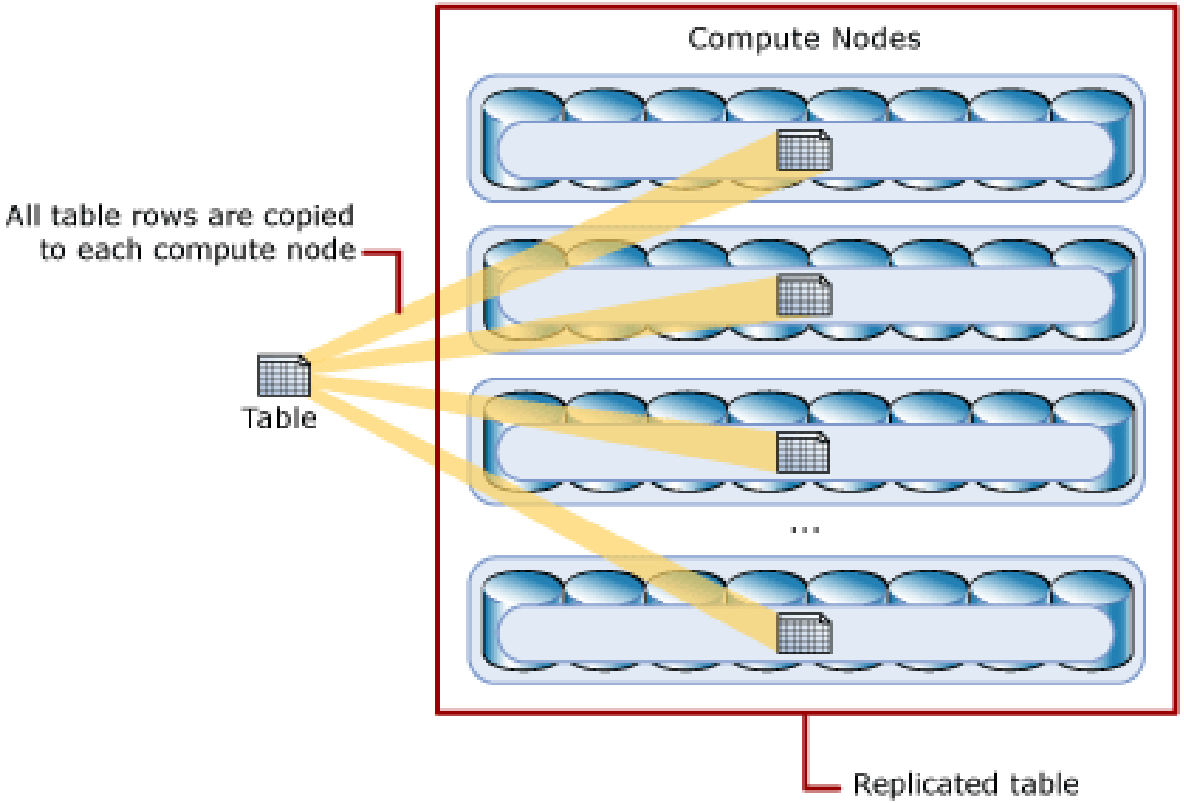
Azure Synapse Analytics Storage

- Azure Synapse Analytics Storage sharding options:
  - Hash-distributed tables
  - Round-robin distributed tables
  - Replicated Tables

Synapse Hash-distributed Tables



Synapse Replicated Tables



Synapse Round-robin Distributed Tables

- Delivers fast performance when used as a staging table for loads
- Distributes data evenly across the table but without any further optimization.
- It is quick to load data into a round-robin table
- Query performance can often be better with hash distributed tables.
- Joins on round-robin tables require reshuffling data, which takes additional time.

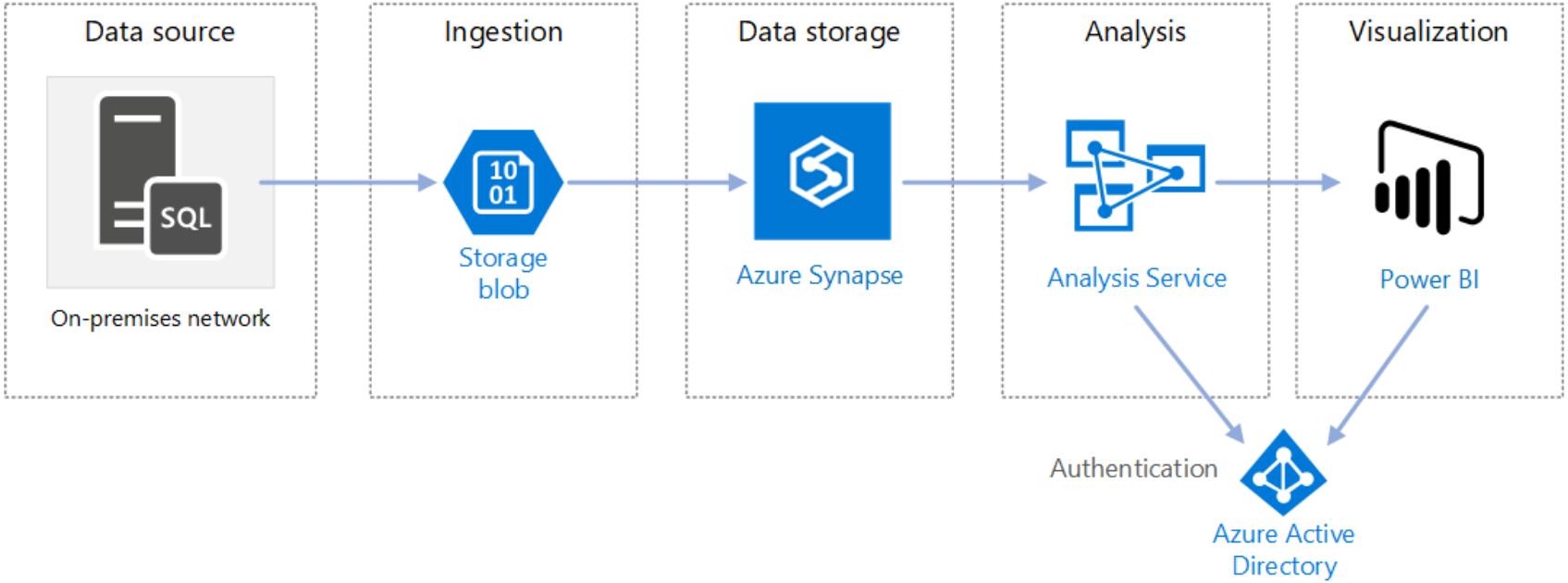
Azure Synapse Analytics Storage

- CREATE TABLE (Azure Synapse Analytics)

Design a Solution That Uses Azure Synapse

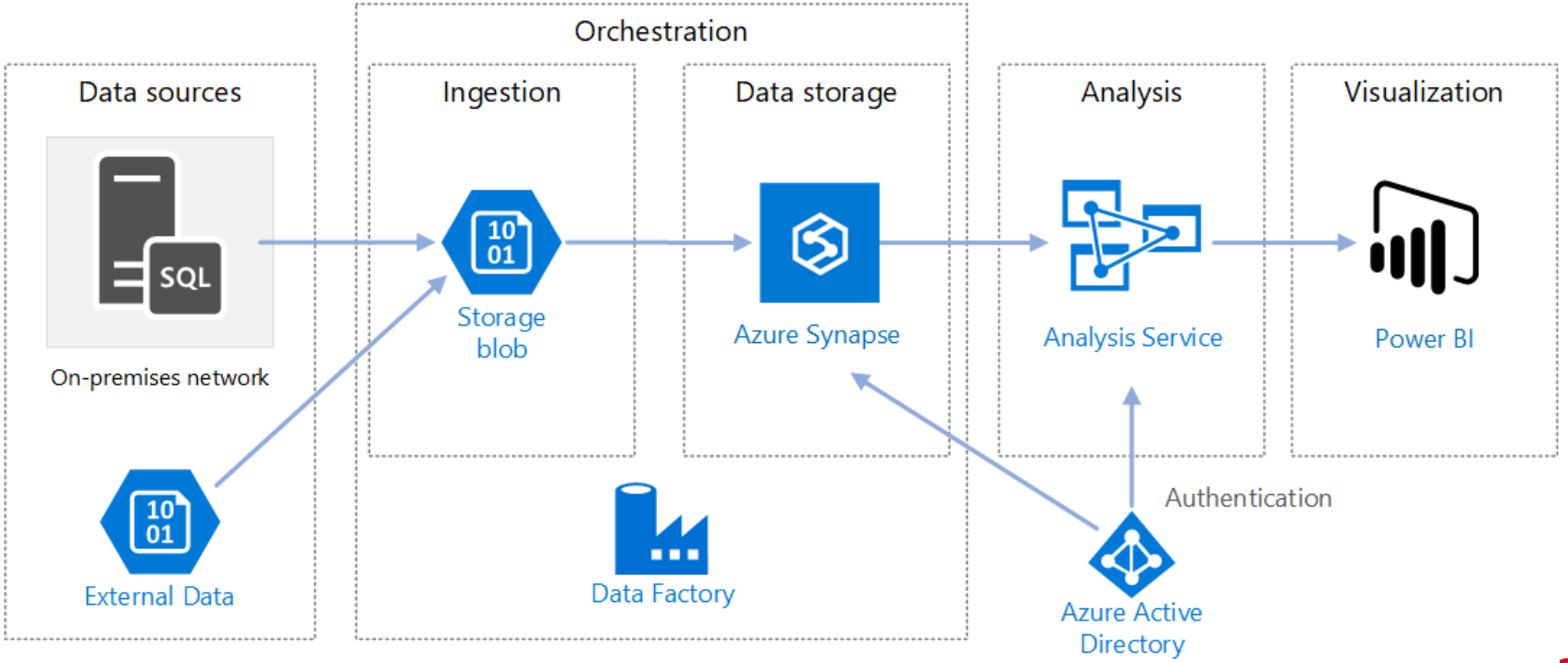
- https://docs.microsoft.com/en-us/azure/architecture/browse/#databases
- https://docs.microsoft.com/en-us/azure/architecture/browse/#integration

Azure Synapse Analytics



https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/data/enterprise-bi-synapse

Azure Synapse Analytics



https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/data/enterprise-bi-adf

Azure Synapse Studio

- Open Synapse Studio

Azure Synapse Ana...

zaalion (Default Directory)

+ Add | Manage view | ...

Filter by name...

Name

syn-demo01 ...

syn-demo01 | Firewalls

Synapse workspace

Search (Ctrl+/)

Save | Discard | + Add client IP

Settings

- SQL Active Directory admin
- Properties
- Locks

Analytics pools

- SQL pools
- Apache Spark pools

Security

- Encryption
- Firewalls
- Managed identities
- Private endpoint connections
- Approved Azure AD tenants
- Azure SQL Auditing

The IPs listed below will have full access to Synapse workspace 'syn-demo01'.

Allow Azure services and resources to access this workspace

ON | OFF

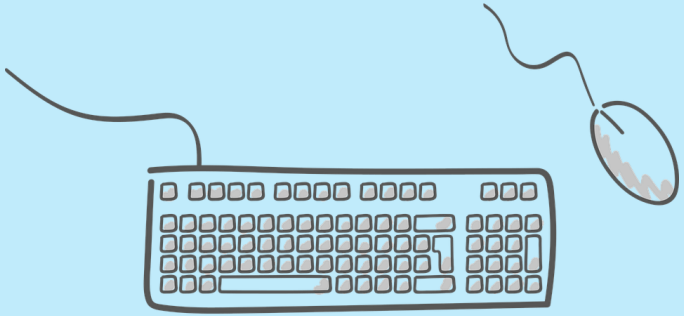
Client IP address 99.230.107.165

Rule name	Start IP	End IP
allowAll	0.0.0.0	255.255.255.255 ...

High Availability

- Azure SQL Database
- Azure Synapse

Demo



- Azure Synapse

Design Data Processing Solutions

Design Data Processing Solutions

- Design batch processing solutions
- Design real-time processing solutions

Design Batch Processing Solutions

- Design batch processing solutions that use Data Factory and Azure Databricks
- Identify the optimal data ingestion method for a batch processing solution
- Identify where processing should take place, such as at the source, at the destination, or in transit

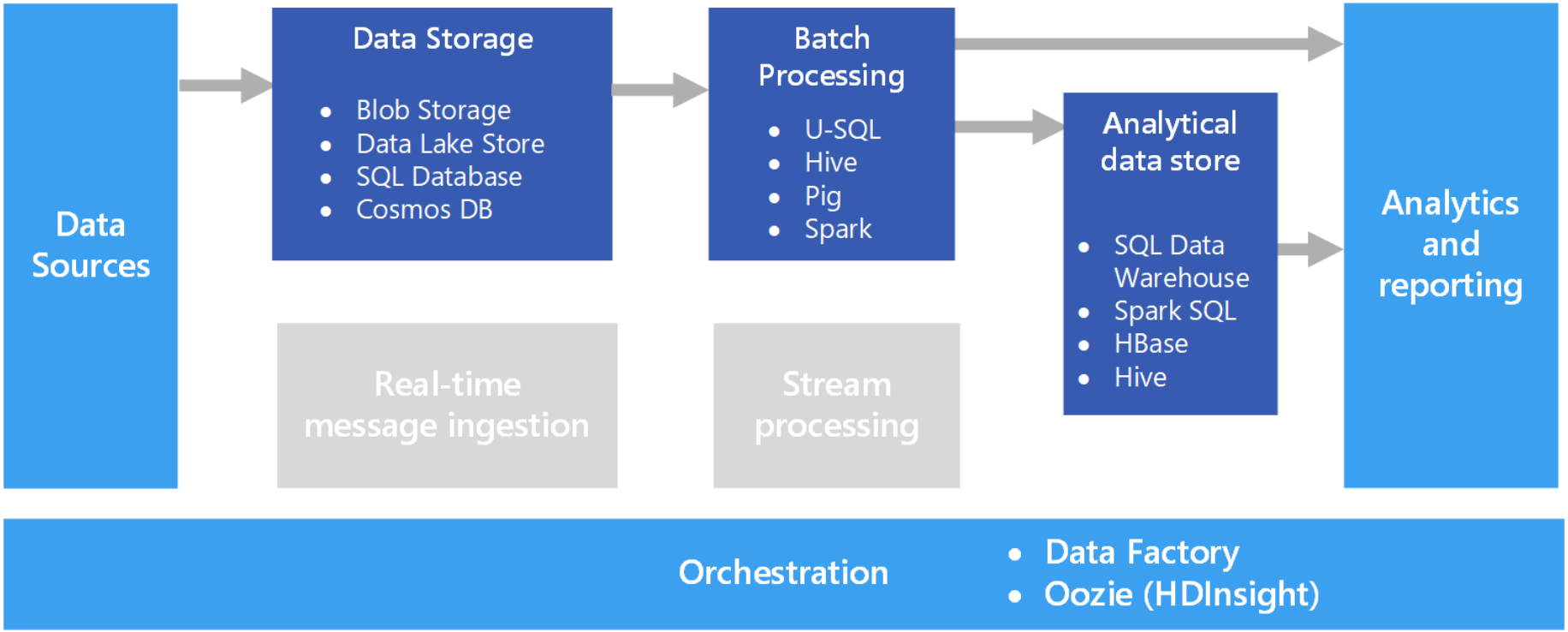
Batch Processing vs. Stream Processing

- Batch processing: working with stored data (e.g. logs, historical data)
- Stream processing: working with live incoming data (e.g. sensors, IoT devices, audit logs)

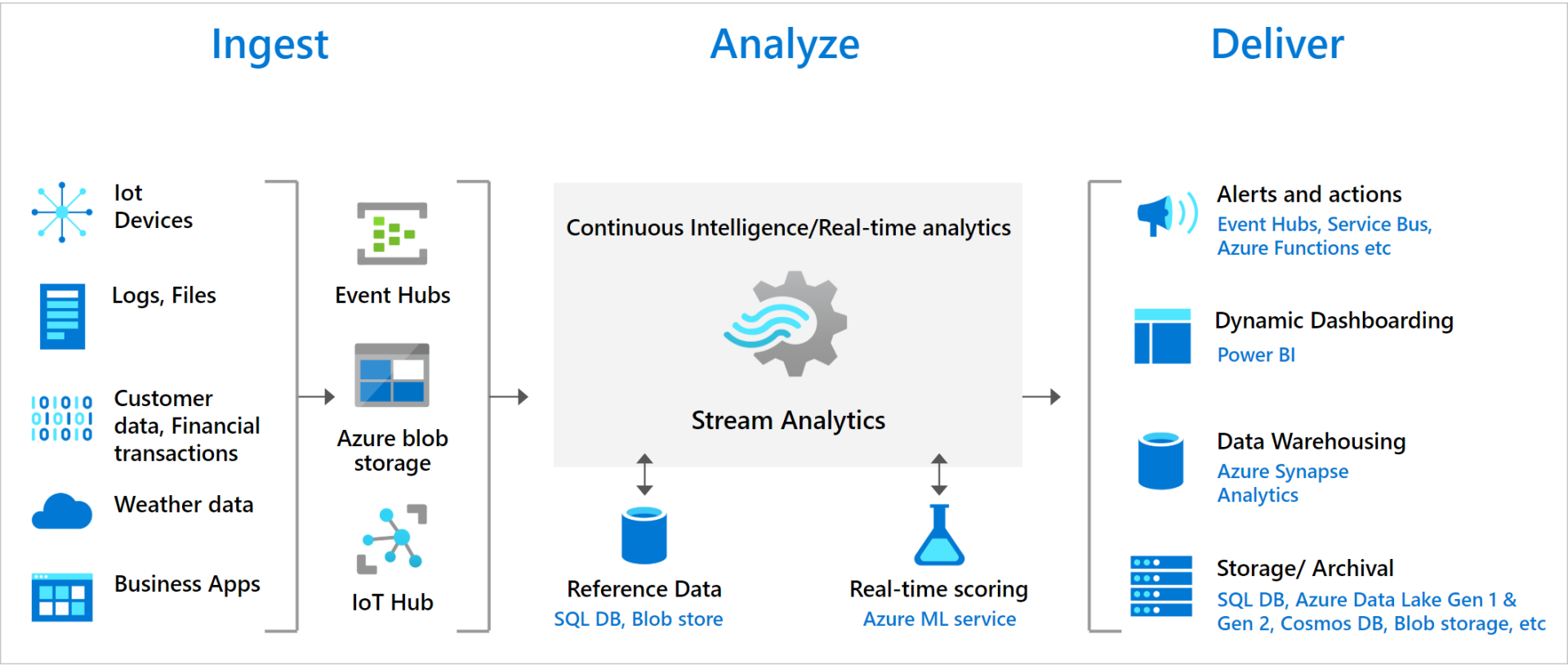
Batch Processing

The process of loading data from a source, processing it and loading the result into a target data store. (ETL & ELT)

Batch Processing



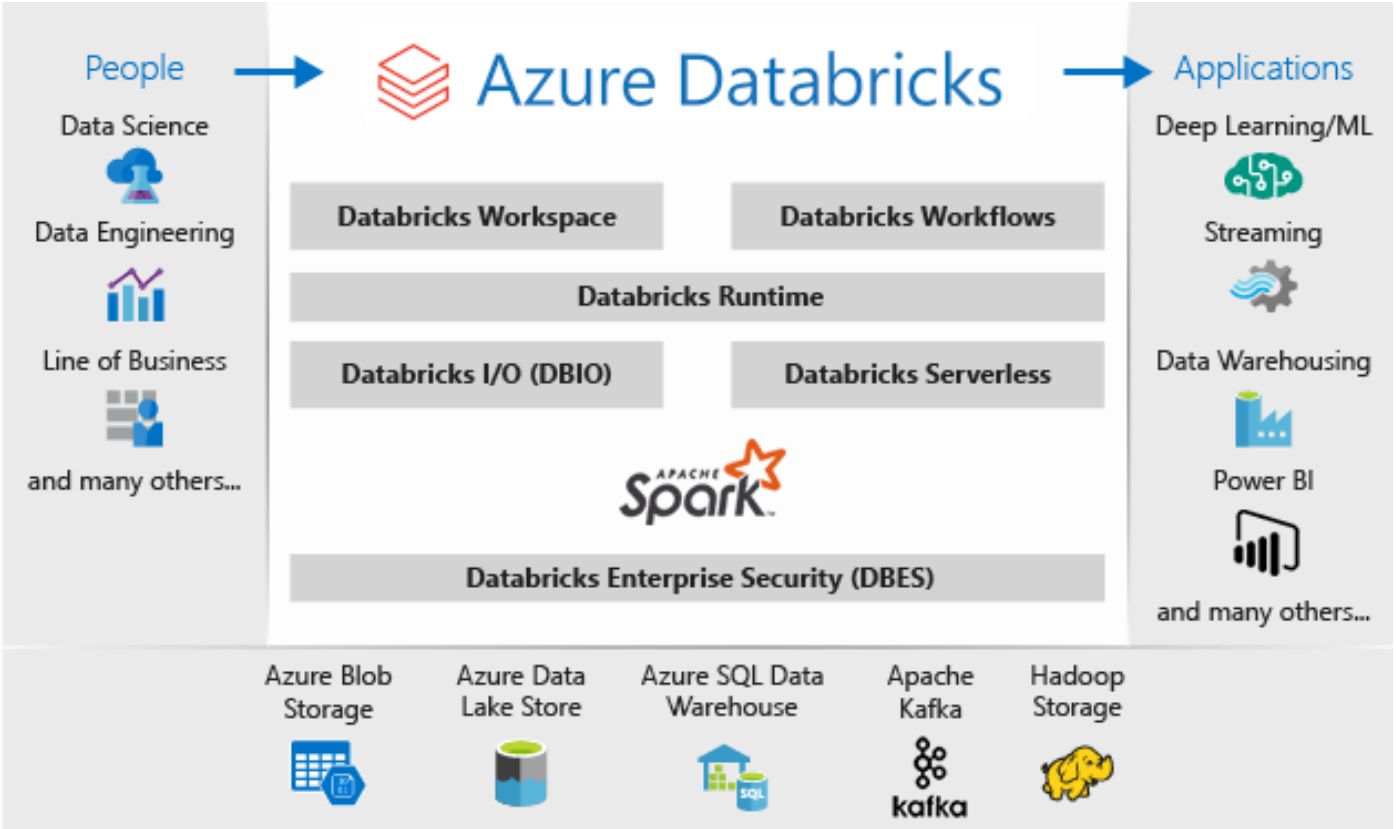
Stream Processing



Databricks

- Cloud-hosted Apache Spark platform.
- Microsoft partnered with Databricks and introduced Azure Databricks
- Azure Databricks is a data analytics platform optimized for the Microsoft Azure cloud services platform.

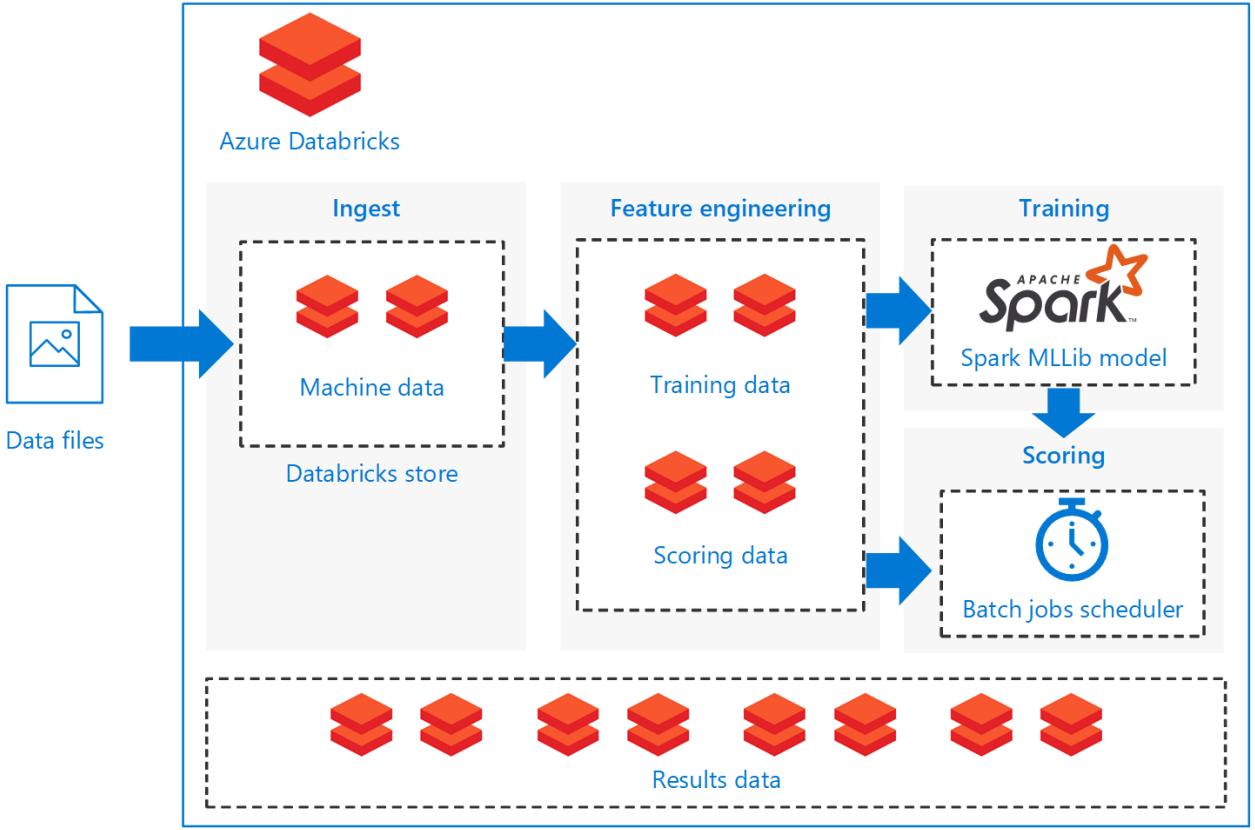
Azure Databricks



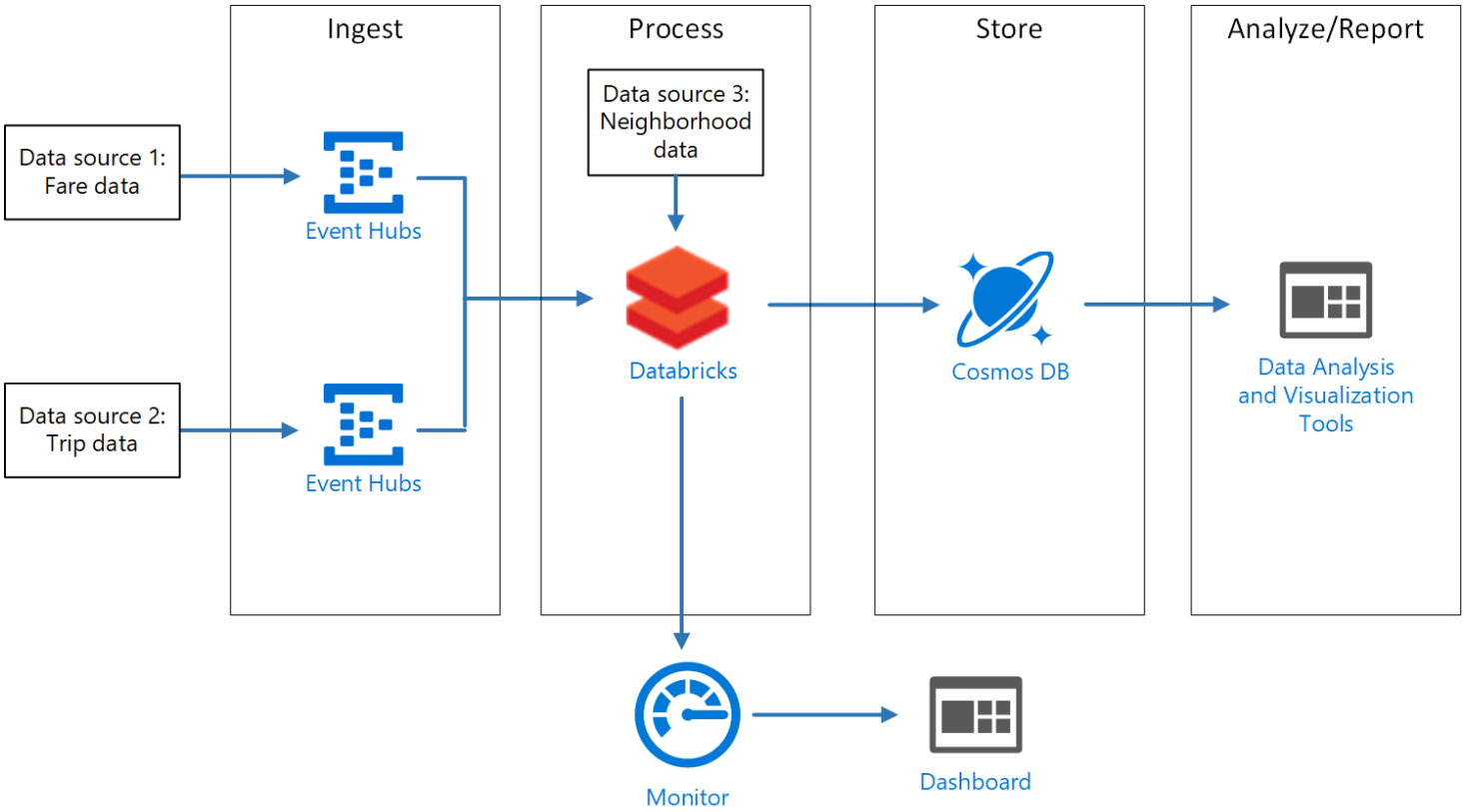
Design a Solution That Uses Azure Databricks

- https://docs.microsoft.com/en-us/azure/architecture/browse/#ai--machine-learning

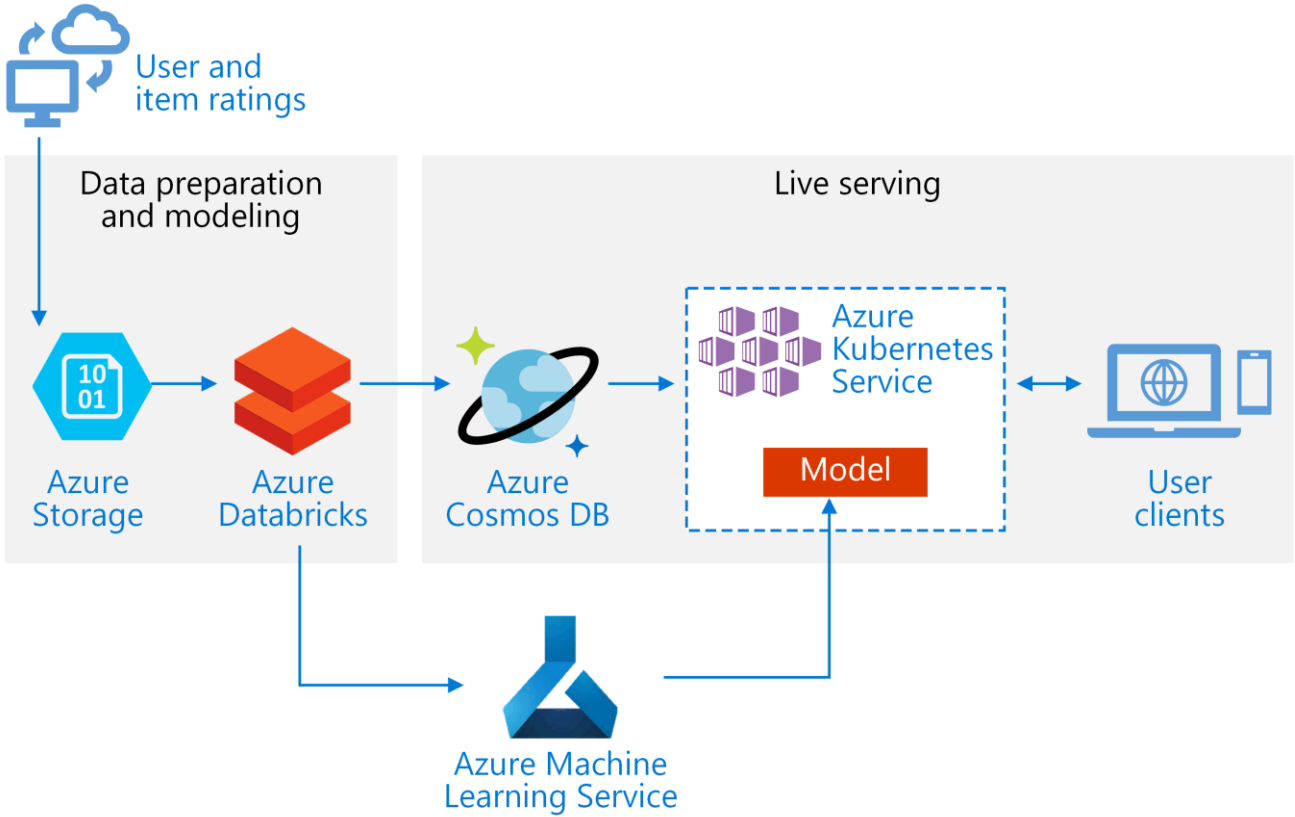
Azure Databricks



Azure Databricks



Azure Databricks



Azure Synapse Ana...

zaalion (Default Directory)

+ Add | Manage view | ...

Filter by name...

Name

syn-demo01 ...

syn-demo01 | Firewalls

Synapse workspace

Search (Ctrl+/)

Save | Discard | + Add client IP

Settings

- SQL Active Directory admin
- Properties
- Locks

Analytics pools

- SQL pools
- Apache Spark pools

Security

- Encryption
- Firewalls
- Managed identities
- Private endpoint connections
- Approved Azure AD tenants
- Azure SQL Auditing

The IPs listed below will have full access to Synapse workspace 'syn-demo01'.

Allow Azure services and resources to access this workspace

ON | OFF

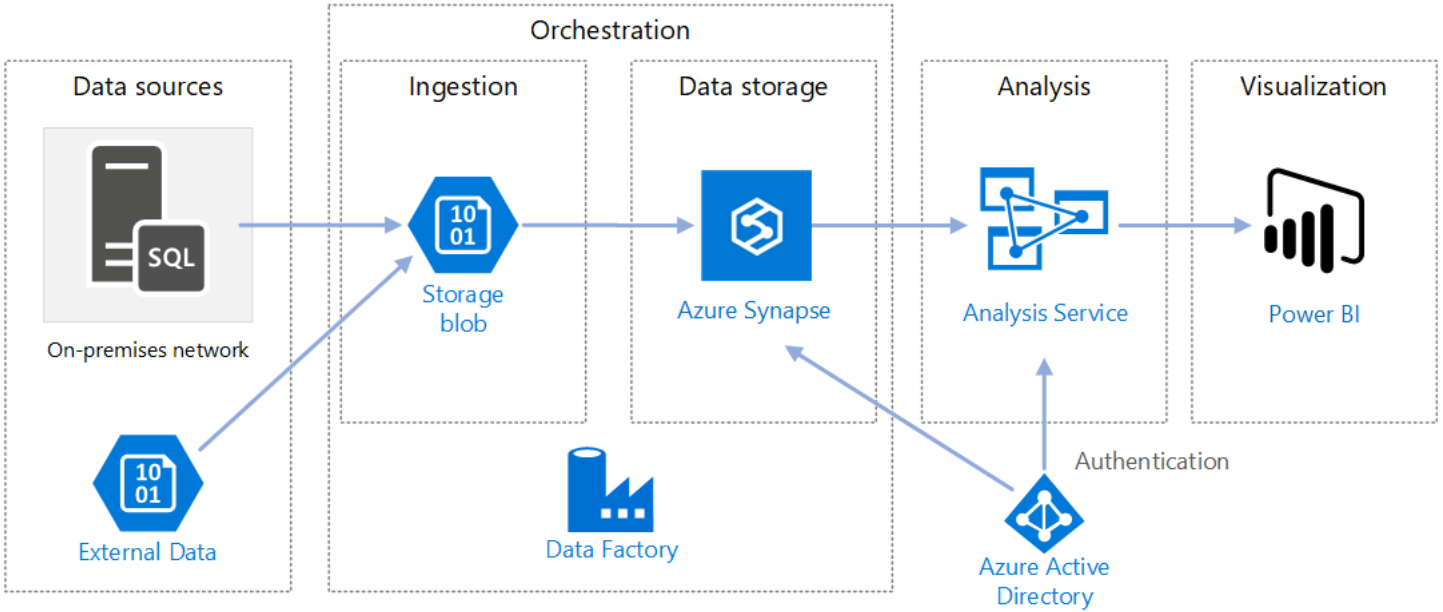
Client IP address 99.230.107.165

Rule name	Start IP	End IP
allowAll	0.0.0.0	255.255.255.255 ...

Design a Solution That Uses Azure Data Factory

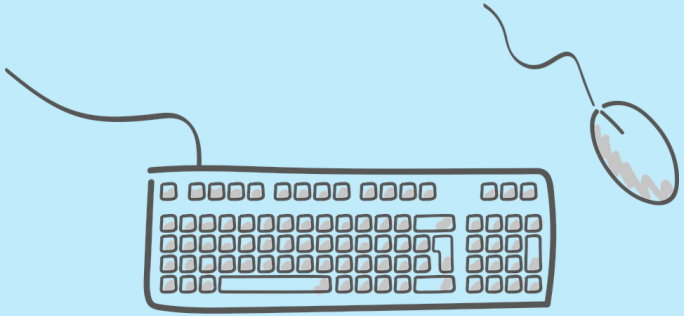
- https://docs.microsoft.com/en-us/azure/architecture/browse/#analytics

Azure Data Factory



https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/data/enterprise-bi-adf

Demo

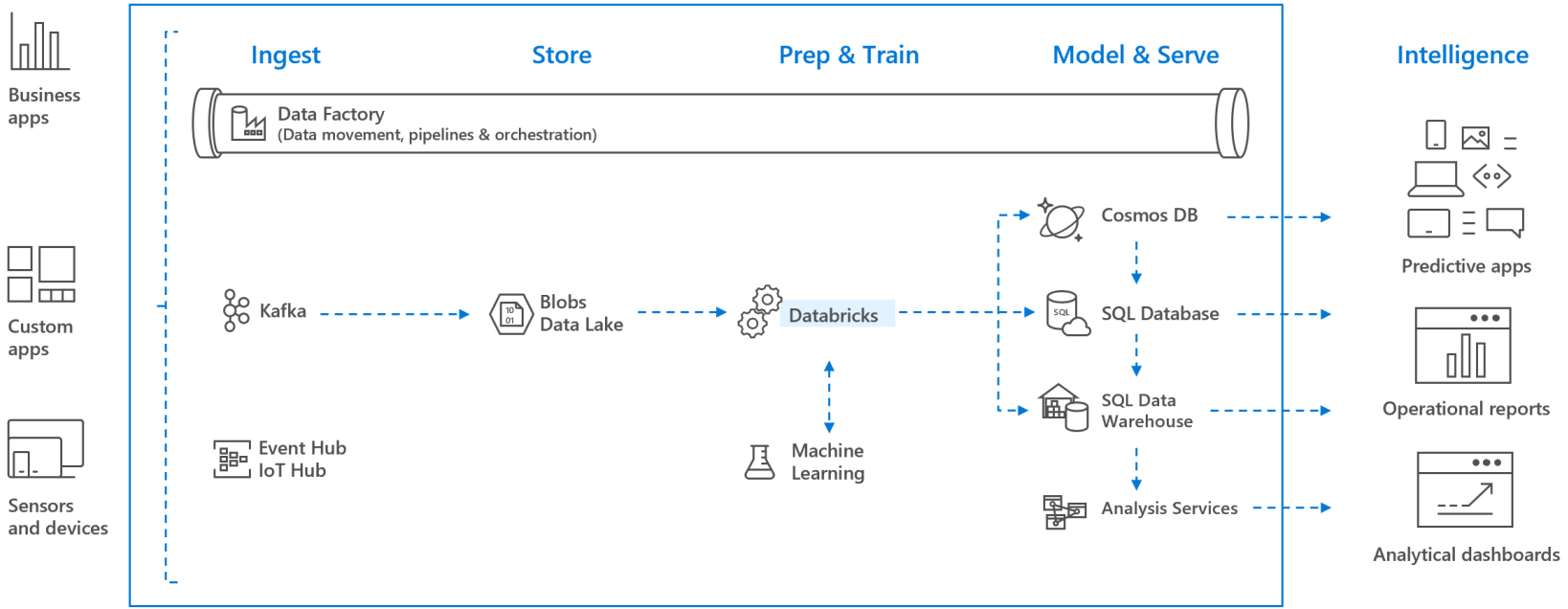


- Azure Data Factory

Design real-time processing solutions

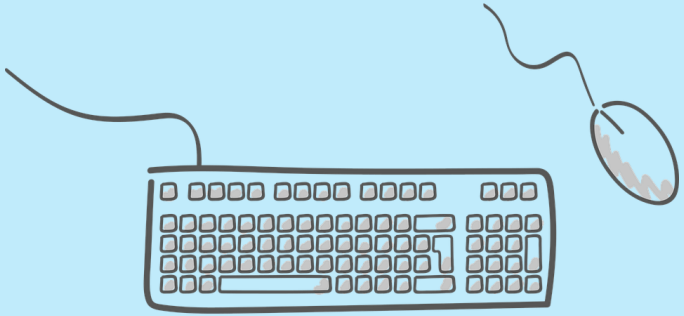
- Design for real-time processing by using Stream Analytics and Azure Databricks
- Design and provision compute resources

Azure Databricks Workspace



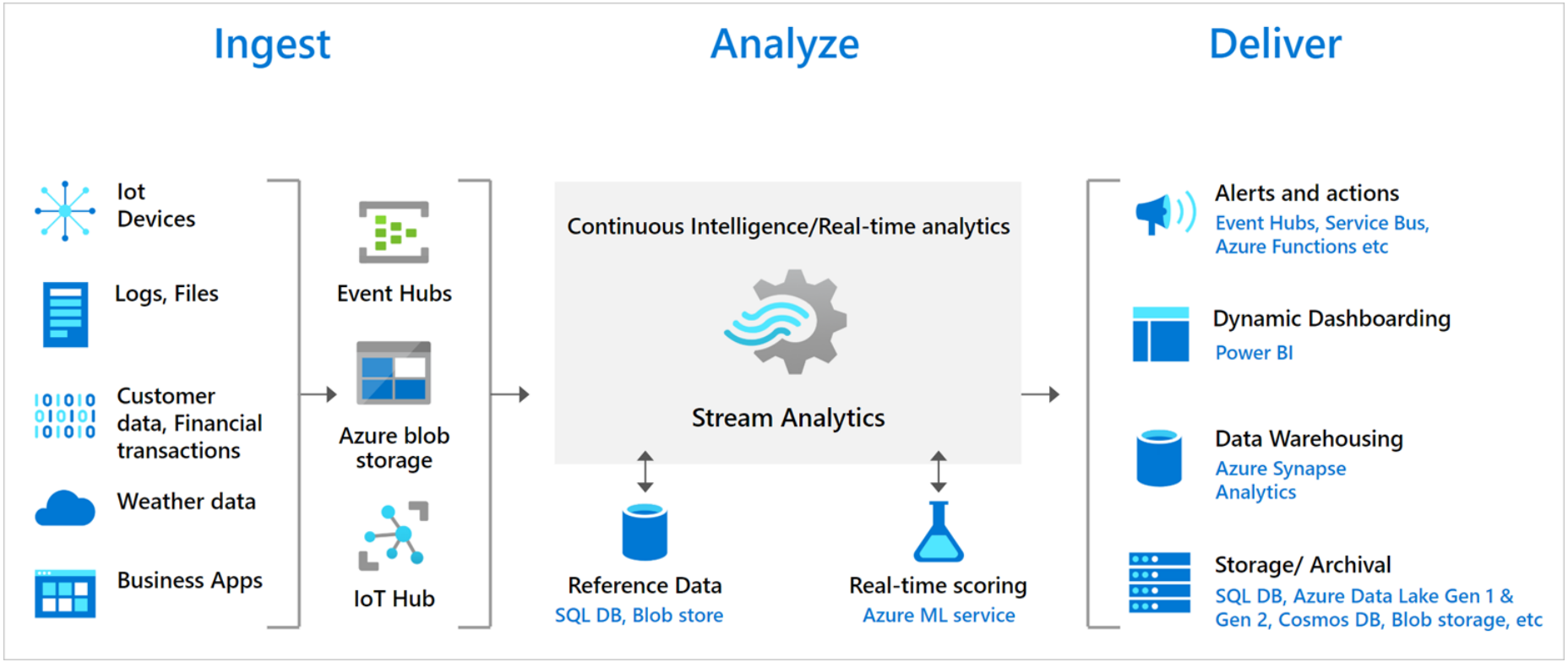
https://docs.microsoft.com/en-us/azure/databricks/scenarios/what-is-azure-databricks-ws

Demo



- Azure Databricks

Azure Stream Analytics



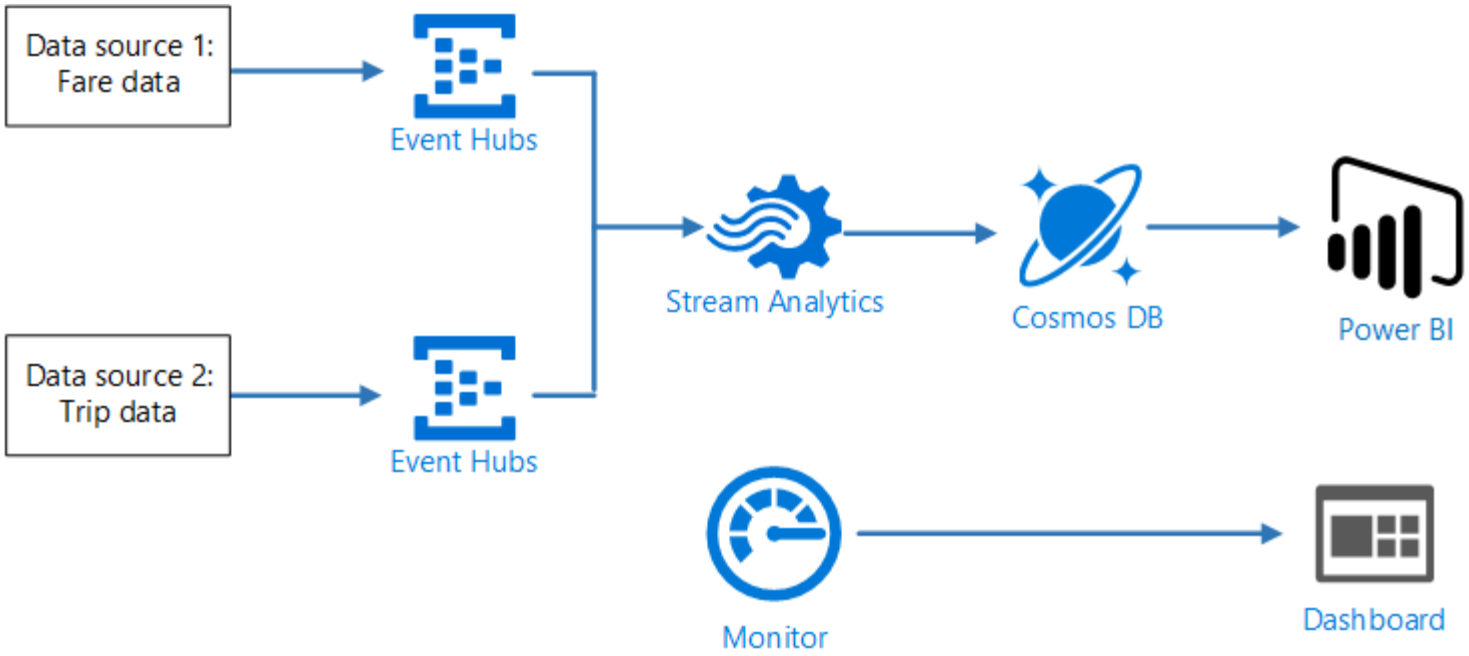
Develop Streaming Solutions

- Azure Stream Analytics
  - Ingest and process real-time data
    - Ingest from *IoT Hub*, *Event Hubs* and *Blob Storage*
    - Process using a *SQL*-like language
    - Output to several services such as *Event Hubs, Power BI, Logic Apps*, etc.

Design a Solution That Uses Azure Stream Analytics

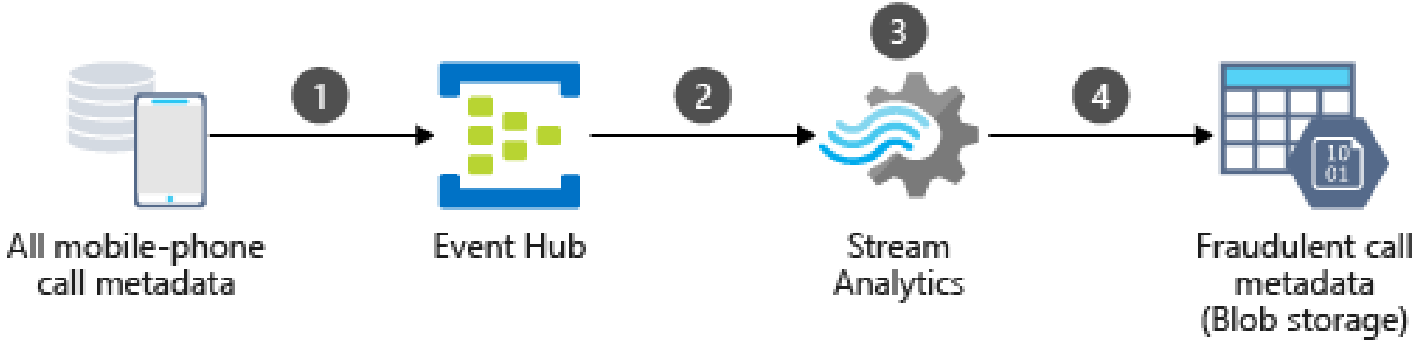
- https://docs.microsoft.com/en-us/azure/architecture/browse/#analytics
    - https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/data/stream-processing-stream-analytics

Azure Stream Analytics



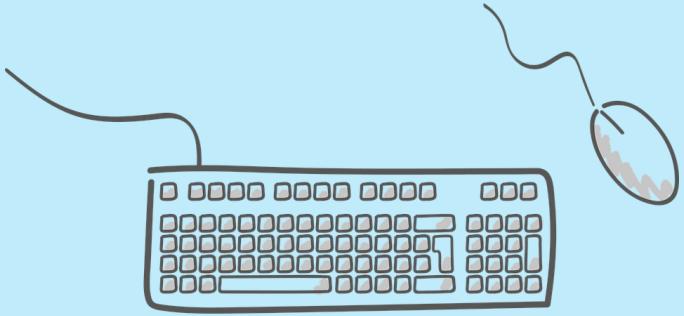
https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/data/stream-processing-stream-analytics

Azure Stream Analytics



https://docs.microsoft.com/en-us/azure/architecture/example-scenario/data/fraud-detection

Demo



- Azure Stream Analytics

Design for Data Security and Compliance

Design for Data Security and Compliance

- Design security for source data access
- Design security for data policies and standards

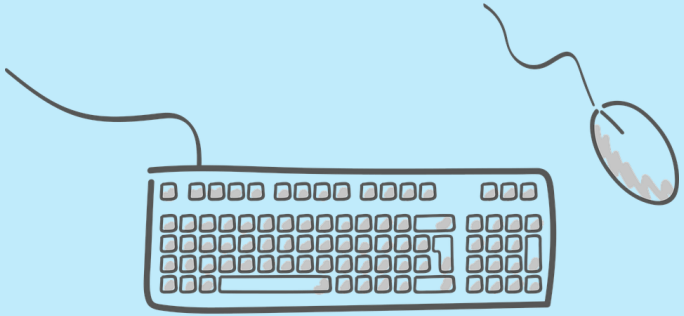
Design Security for Source Data Access

- Plan for secure endpoints (private/public)
- Choose the appropriate authentication mechanism, such as access keys, shared access, signatures (SAS), and Azure Active Directory (Azure AD)

Plan for Secure Endpoints

- Secure endpoints:
  - Azure Cosmos DB
  - Azure Storage Account
  - Azure Synapse Analytics
  - Azure Data Factory
  - Azure Databricks

Demo



- Securing endpoints
  - Azure Storage Account
  - Azure Cosmos DB
  - Azure Synapse Analytics

Design Security for Data Policies and Standards

- Design data encryption for data at rest and in transit
- Design for data auditing and data masking
- Design for data privacy and data classification
- Design a data retention policy
- Plan an archiving strategy
- Plan to purge data based on business requirements

Design Data Encryption for Data at Rest and in Transit

- Data encryption:
  - Azure Cosmos DB
  - Azure Storage Account
  - Azure Synapse Analytics

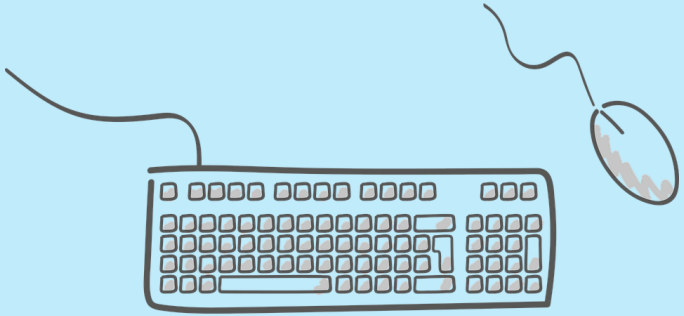
Azure SQL Database Security

- Transparent Data Encryption (TDE)
- Always Encrypted
- Dynamic Data Masking

Azure Policy

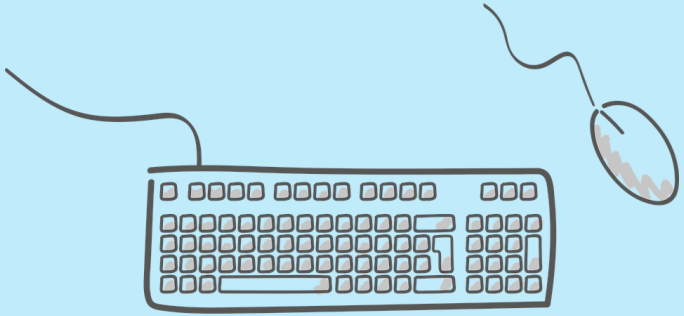
- Azure Policy helps to enforce organizational standards and to assess compliance at-scale.

Demo



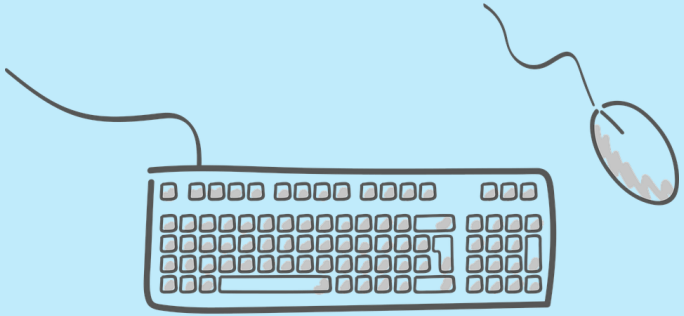
- Azure Storage Account data lifecycle

Demo



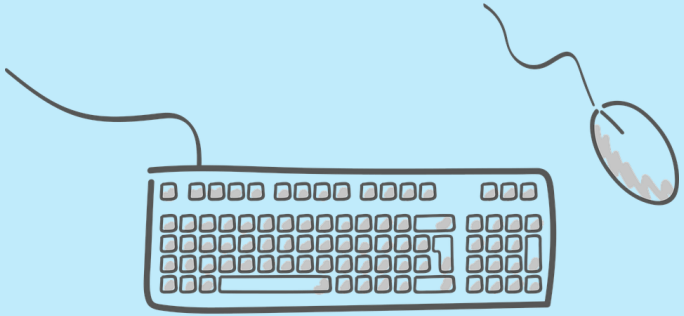
- Azure Synapse data masking

Demo



- Azure Synapse data classification

Demo



- Azure Policy for data services

The Exam

Questions in DP-201

- Multiple choice
- Drag and drop
- Scenario based
- There will be hands-on labs

DP-201

- Exam DP-201 :

  https://docs.microsoft.com/en-us/learn/certifications/exams/dp-201

- Skills measured :

  https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE3VRMb

Candidates for this exam are Microsoft Azure data engineers who collaborate with business stakeholders to identify and meet the data requirements to design data solutions that use Azure data services.

Azure data engineers are responsible for data-related design tasks that include designing Azure data storage solutions that use relational and non-relational data stores, batch and real-time data processing solutions, and data security and compliance solutions.

Candidates for this exam must design data solutions that use the following Azure services: Azure Cosmos DB, Azure Synapse Analytics, Azure Data Lake Storage, Azure Data Factory, Azure Stream Analytics, Azure Databricks, and Azure Blob storage.

Part of the requirements for: Microsoft Certified: Azure Data Engineer Associate
Related exams: 1 related exam
Important: See details
Go to Certification Dashboard

Schedule exam

Exam DP-201: Designing an Azure Data Solution

United States

Languages: English, Japanese, Chinese (Simplified), Korean
Retirement date: none

This exam measures your ability to accomplish the following technical tasks: design Azure data storage solutions; design data processing solutions; and design for data security and compliance.

$165 USD*

Price based on the country in which the exam is proctored.

Schedule exam >

Official practice test for Designing an Azure Data Solution
All objectives of the exam are covered in depth so you'll be ready for any question on the exam.

Verify exam discount eligibility

For Microsoft employees

Microsoft employees are eligible for discounted exams. The discount will be reflected at the end of the checkout process. For MOS exams at Certiport, please request a voucher through the Microsoft Employee Voucher Portal.

To verify you are a Microsoft employee, link your Microsoft work account (alias@microsoft.com).

Link account

For Microsoft event attendees

If you recently attended a Microsoft event, you may be eligible for a discounted Microsoft Certification exam. To check eligibility, select an event you attended and verify the account used to register for the event.
Terms and Conditions apply.

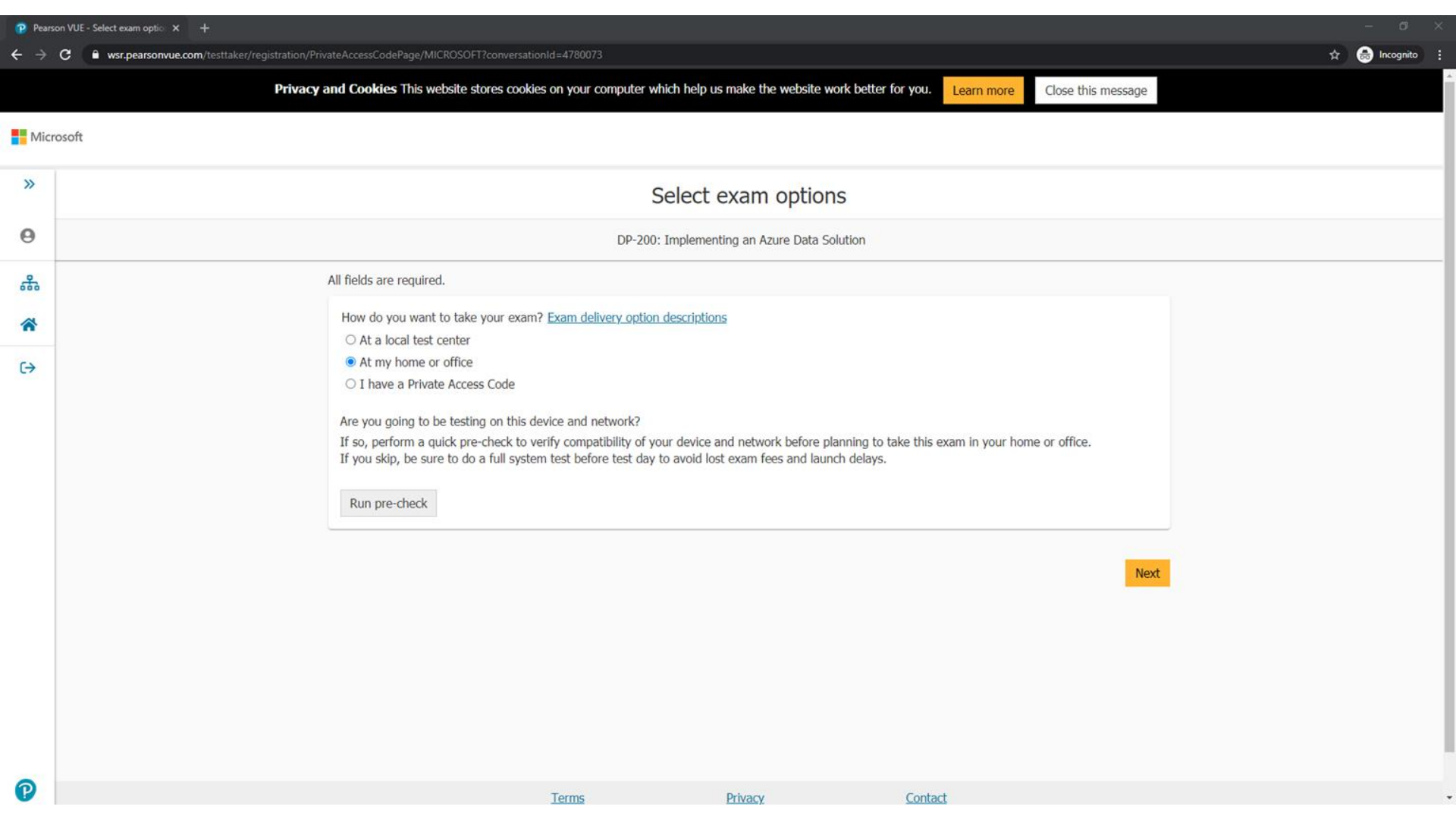
Microsoft Ignite 2019, Orlando

Verify account

Continue scheduling exam

Proceed to the Pearson VUE website to complete the exam scheduling process.

Go to Pearson VUE


Pearson VUE - Select exam optio
wsr.pearsonvue.com/testtaker/registration/PrivateAccessCodePage/MICROSOFT?conversationId=4780073
Incognito
Privacy and Cookies This website stores cookies on your computer which help us make the website work better for you.
Learn more
Close this message
Microsoft
Select exam options
DP-200: Implementing an Azure Data Solution
All fields are required.
How do you want to take your exam? Exam delivery option descriptions
At a local test center
At my home or office
I have a Private Access Code
Are you going to be testing on this device and network?
If so, perform a quick pre-check to verify compatibility of your device and network before planning to take this exam in your home or office.
If you skip, be sure to do a full system test before test day to avoid lost exam fees and launch delays.
Run pre-check
Next
Terms
Privacy
Contact

Pearson VUE - Select exam optic | System Check

systemcheck.proctorcam.com/systemcheck?candidateId=247036059

Incognito

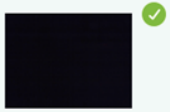
Pearson | VUE

System check - Checking your requirements

Microphone

Default - Microphone (SI

Internet speed



Webcam

Integrated Webcam (0c4

Next

Course Repository

https://github.com/zaalion/oreilly-dp-200-201

Q&A

Thank you!

Reza Salehi

@zaalion