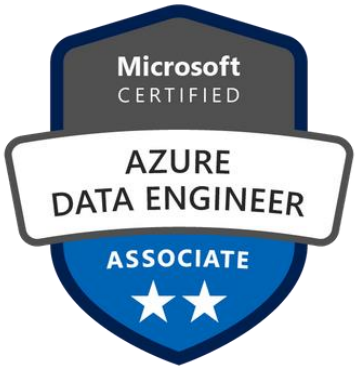
# Exam DP-203: Microsoft Azure Data Engineer Associate Crash Course

## Data Engineering in Microsoft Azure

# Reza Salehi

Cloud Consultant

@zaalion


# Course Overview

# DP-203


Microsoft
CERTIFICATION EXAM
DP-201


Microsoft
CERTIFICATION EXAM
DP-200


Microsoft
CERTIFICATION EXAM
DP-203

# DP-203 Skills Measured

Exam DP-203: Data Engineering on Microsoft Azure

# Questions & Resources

- Post questions in the QnA box
- Resources are in the course repository
  - https://github.com/zaalion/oreilly-dp-203

- Reach out:
  - Twitter: @zaalion

# DP-203 Candidate Profile

- Microsoft Azure data engineers
  - Integrate, transform, and consolidate data from various structured and unstructured data systems ...
  - Into structures that are suitable for building analytics solutions

# DP-203 Candidates

Azure Data Engineers integrate, transform, and consolidate data:

- Knowledge of data processing languages, such as SQL, Python, or Scala
- Understand parallel processing and data architecture patterns.

# DP-203 Skills Measured

Skills measured:

- Design and implement data storage (15-20%)
- Develop data processing (40-45%)
- Secure, monitor, and optimize data storage and data processing (30-35%)

# Data Types


Structured

```
{"widget": {
    "debug": "on",
    "window": {
        "title": "Sample Konfabulator Widget",
        "name": "main_window",
        "width": 500,
        "height": 500
    },
    "image": {
        "src": "Images/Sun.png",
        "name": "sun1",
        "hOffset": 250,
        "vOffset": 250,
        "alignment": "center"
    },
    "text": {
        "data": "Click Here",
        "size": 36,
        "style": "bold",
        "name": "text1",
        "hOffset": 250,
        "vOffset": 100,
        "alignment": "center",
        "onMouseUp": "sun1.opacity = (sun1.opacity / 100) * 90;"
    }
}}
```

Semi-structured


Unstructured

# DP-203 Main Focus (not limited to)

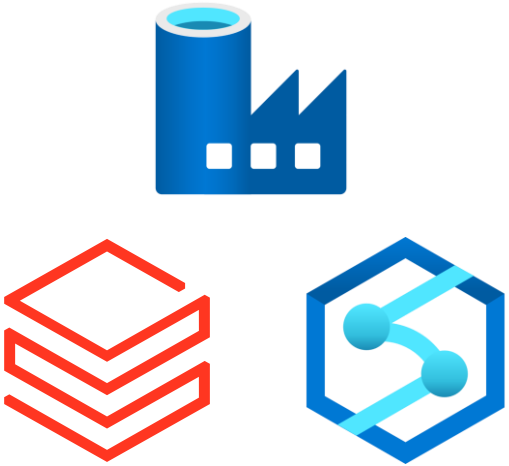
- Azure Data Lake Gen2
- Azure Stream Analytics
- Azure Synapse Analytics
- Azure Data Factory
- Azure Databricks





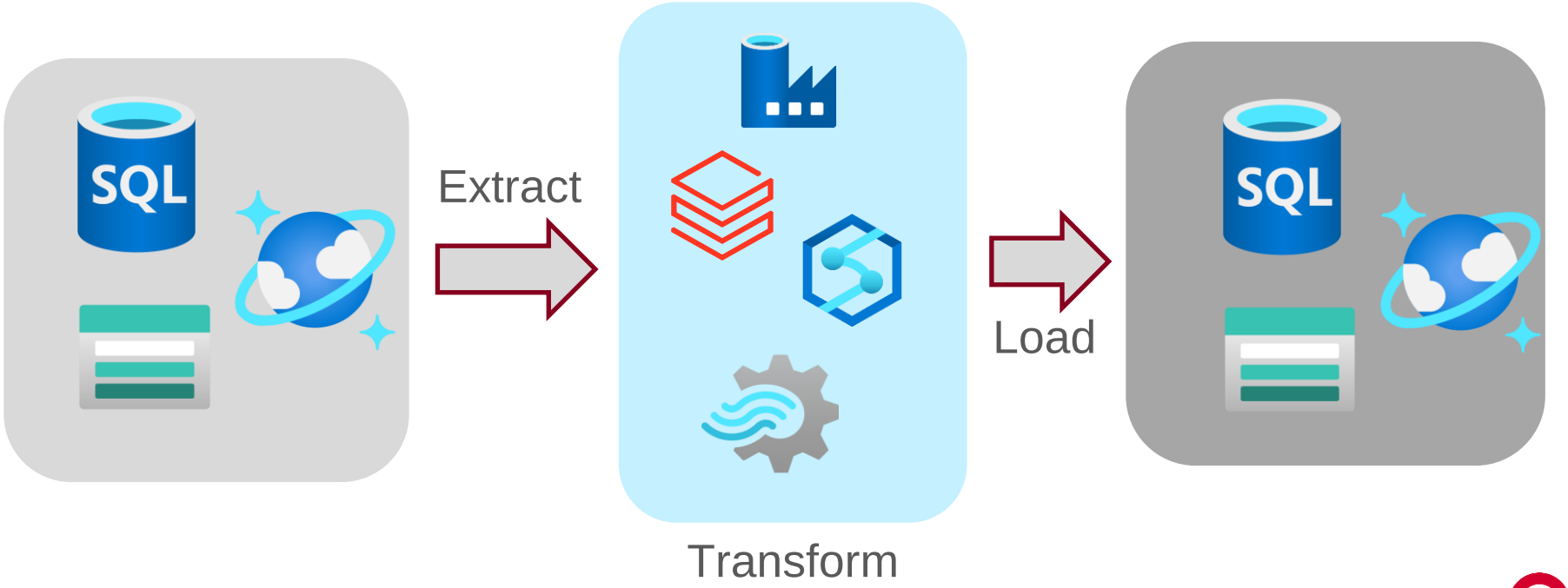
# Data Processing Types

Stream data



Batch data

# ETL/ELT



# Design and Implement Data Storage

# Sharding

- A data store hosted by a single server might be subject to the following limitations:
  - Storage space
  - Computing resources
  - Network bandwidth
  - Geography

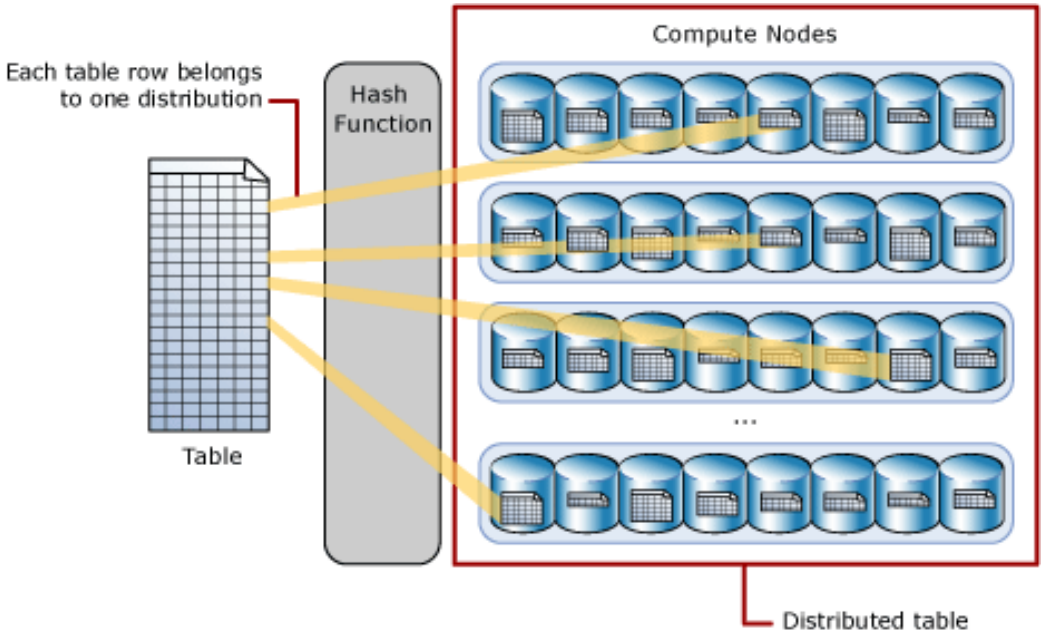
# Sharding

- Solution
  - Divide the data store into horizontal partitions or shards.
  - Each shard has the same schema but holds its own distinct subset of the data.

# Azure Synapse Analytics Shard

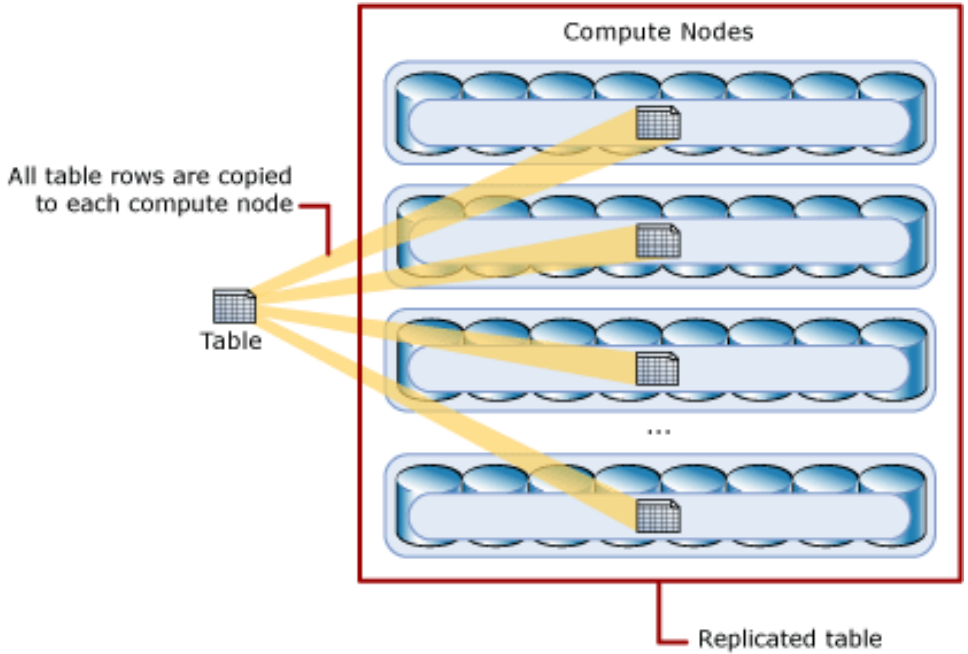
- Azure Synapse Analytics Storage sharding options:
  - Hash-distributed tables
  - Round-robin distributed tables
  - Replicated Tables

# Azure Synapse Distributed Tables (Hash)




See reference

# Azure Synapse Distributed Tables (Replicated)




See reference

# Azure Synapse Distributed Tables (Round Robin)

- The simplest table to create
- Delivers fast performance when used as a staging table for loads
- Distributes data evenly across the table

*See reference*

# Azure Synapse External Tables

- External Tables
  - An external table points to data located in Hadoop, Azure Storage blob, or Azure Data Lake Storage.
  - External tables are used to read data from files or write data to files in Azure Storage.
  - With Synapse SQL, you can use external tables to read external data using dedicated SQL pool or serverless SQL pool.

# 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

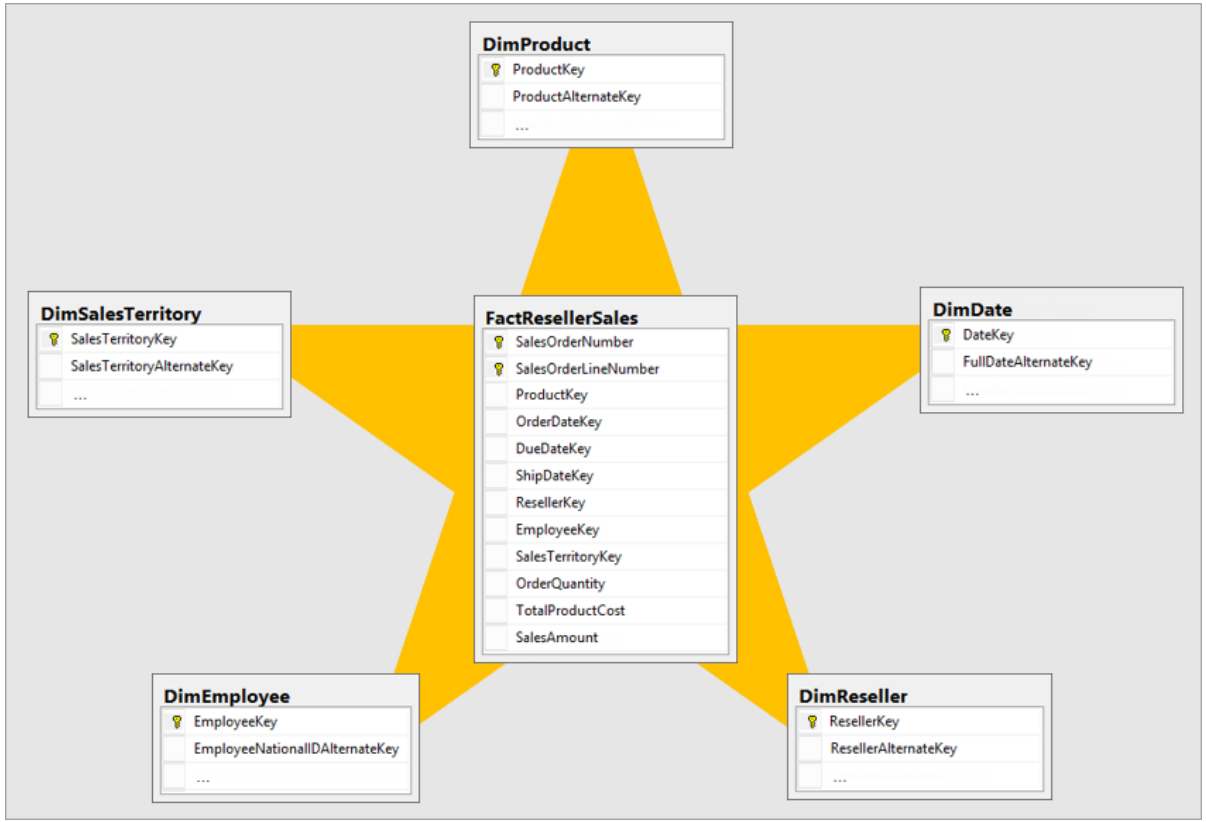
# Choose the Partition Distribution Type

- Data partitioning types
  - Horizontal
  - Vertical
  - Functional

# Azure Synapse Star Schema

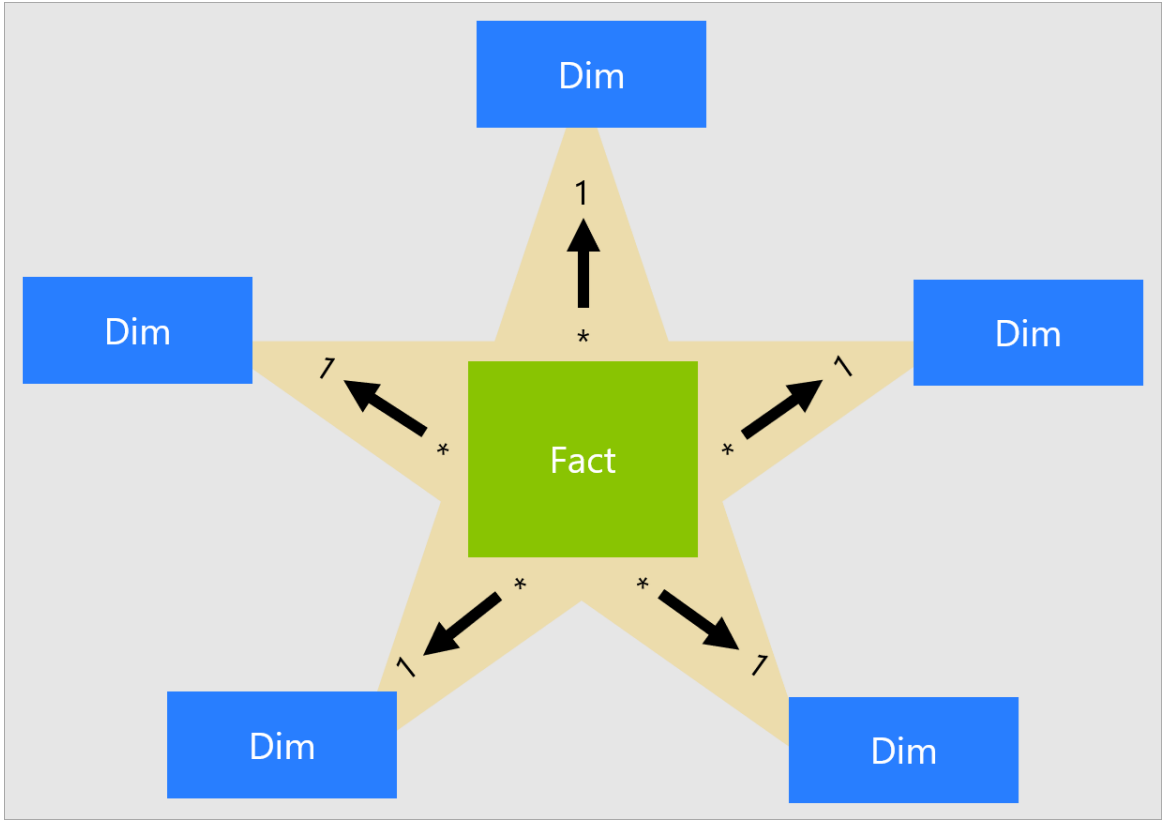
- Star schema
  - A mature modeling approach widely adopted by relational data warehouses. It requires modelers to classify their model tables as either dimension or fact.
    - Dimension tables
    - Fact tables

# Azure Synapse Star Schema




*See reference*

# Azure Synapse Star Schema




See reference

# Slowly Changing Dimensions

- Slowly changing dimension
  - Dimensions in data management and data warehousing contain relatively static data about such entities as geographical locations, customers, or products.
  - Data captured by Slowly Changing Dimensions (SCDs) change slowly but unpredictably, rather than according to a regular schedule.
  - See tutorial

# Slowly Changing Dimensions

- Slowly changing dimension types:
  - Type 1 SCD
  - Type 2 SCD
  - Type 3 SCD
  - Type 6 SCD (1+2+3)

# Temporal Data

- Temporal Data
  - A temporal database stores data relating to time instances. It offers temporal data types and stores information relating to past, present and future time.
  - Azure SQL Database

# Database Normalization

- The process of structuring a database in order to reduce data redundancy and improve data integrity.
  - UNF: Unnormalized form
  - 1NF: First normal form
  - 2NF: Second normal form
  - 3NF: Third normal form

# Types of Keys in Data Warehouse

- Primary Key
- Surrogate Key vs. Natural Key (Business key)
- Alternate key (e.g., UNIQUE constraint)
- Foreign Key

# Develop data processing

# Choosing the Right Data Storage

- 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 Cognitive Services output
  - **Azure Blob Storage**
- Low latency document /NoSQL database
  - **Azure Cosmos DB NoSQL API**
- Database to model graphs (e.g., social media )
  - **Azure Cosmos DB Graph API**
- Migrating from MongoDB
  - **Azure Cosmos DB for MongoDB API**

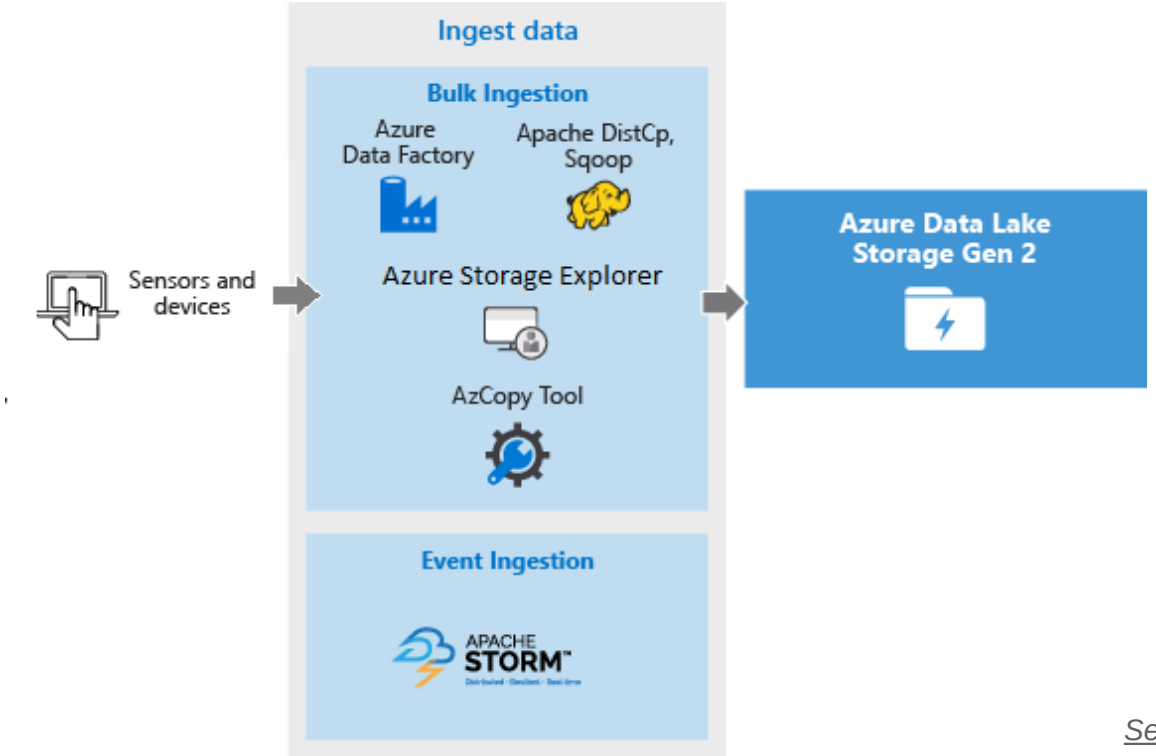
# Choosing the Right Data Storage

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

# Azure Data Lake Gen2

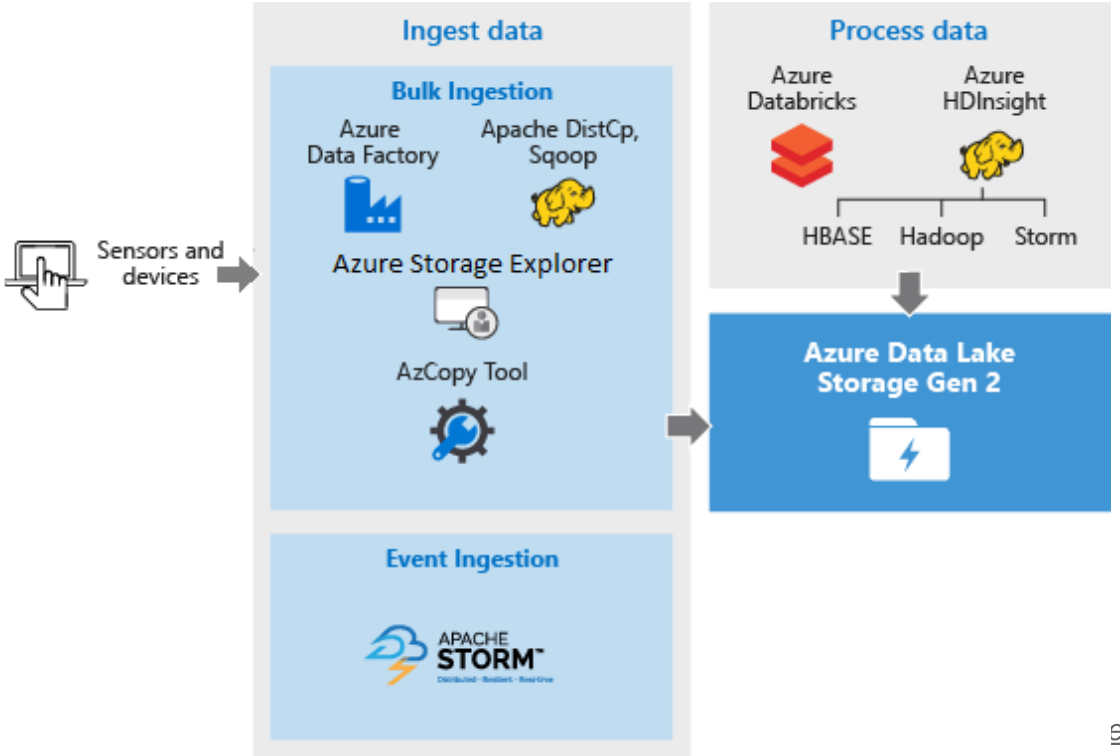
- Azure Data Lake Storage Gen2 is a set of capabilities dedicated to big data analytics, built on Azure Blob storage.
  - Hadoop compatible access
  - A superset of POSIX permissions
  - Cost effective
  - Optimized driver

# Data Lake Storage Gen2 for big data requirements



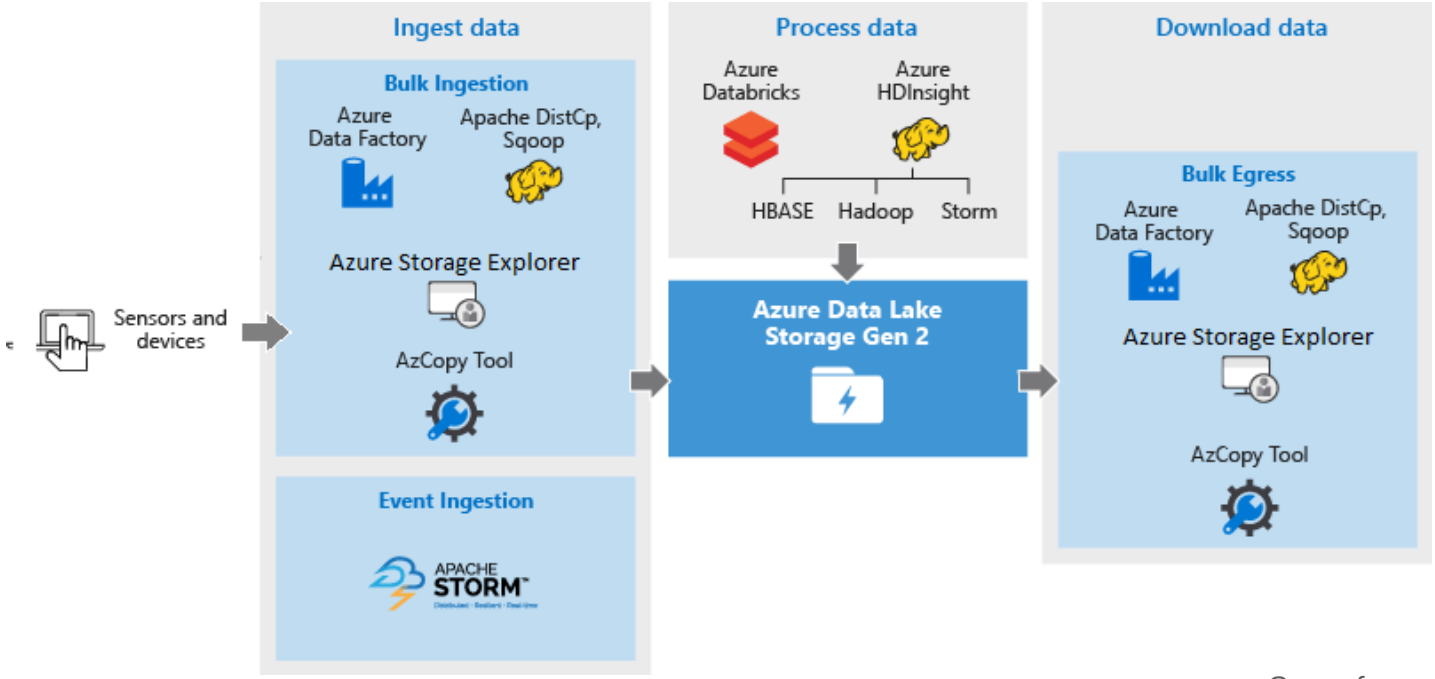

*See reference*

# Data Lake Storage Gen2 for big data requirements




ee reference

# Data Lake Storage Gen2 for big data requirements




*See reference*

# File Types for Storage (Data Lake)

- Avro format
- Binary format
- Delimited text format
- Excel format
- JSON format
- ORC format
- Parquet format
- XML format

# File Types for Storage (Data Lake)

- AVRO is a row-based storage format whereas PARQUET is a columnar based storage format.
- The Optimized Row Columnar (ORC) file format provides a highly efficient way to store Apache Hive data.

# Data Lake Access Control Model

- Data Lake Storage Gen2 supports the following authorization mechanisms:
  - **Shared Key authorization**
  - **Shared access signature (SAS) authorization**
  - **AAD, Role-based access control (Azure RBAC)**
  - **AAD, Access control lists (ACL)**

# Data Lake Archiving

- Access tiers for Azure Blob Storage
  - **Hot** - Optimized for storing data that is accessed frequently.
  - **Cool** - Optimized for storing data that is infrequently accessed and stored for at least 30 days.
  - **Archive** - Optimized for storing data that is rarely accessed and stored for at least 180 days with flexible latency requirements, on the order of hours.

# Cosmos DB

- Browse Azure Architectures for Cosmos DB

# Data Lake Storage Gen2 & Blobs

- Browse Azure Architectures for Azure Storage

# Select the Appropriate Cosmos DB API

- Cosmos DB APIs
  - API for NoSQL
  - API for MongoDB
  - API for PostgreSQL
  - API for Apache Cassandra
  - API for Apache Gremlin
  - API for Table
    - vs. Azure Table Storage

# Cosmos DB Data Distribution

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

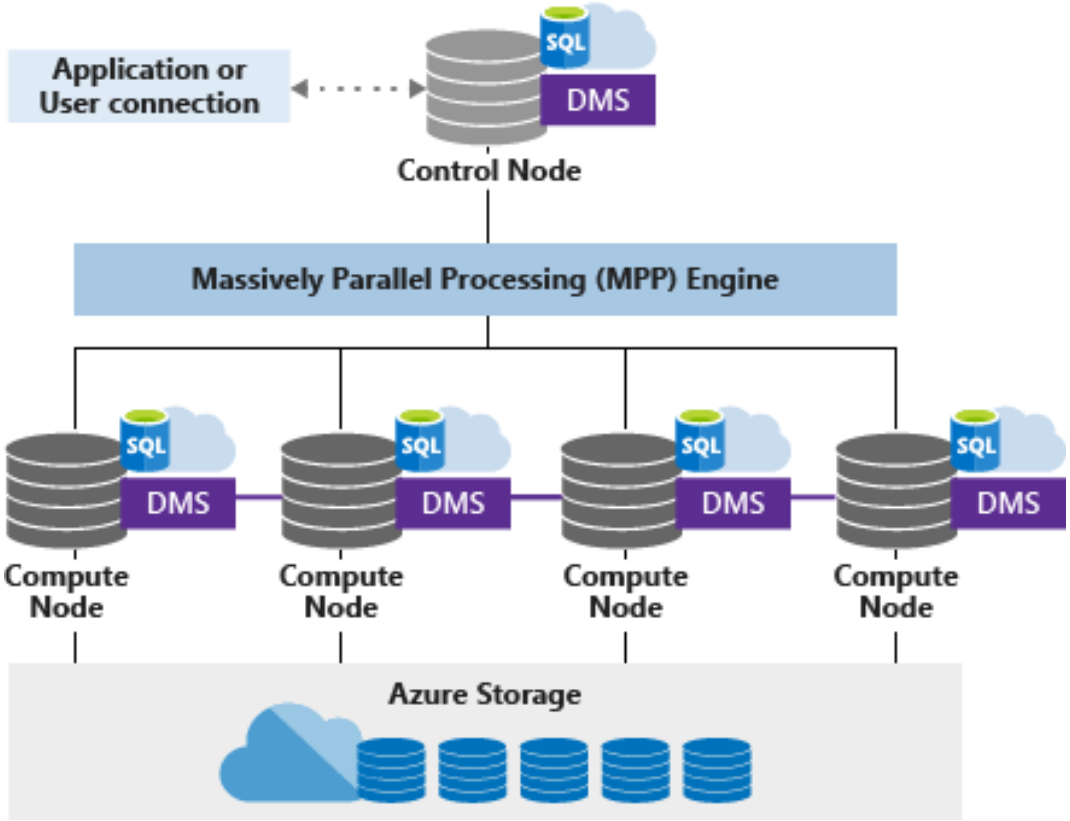
# Azure Synapse

- Browse Azure Architectures for Azure Synapse

# Azure Synapse Analytics

- Components:
  - Synapse SQL: Complete T-SQL based analytics – Generally Available
    - Dedicated SQL pool (pay per DWU provisioned)
    - Serverless SQL pool (pay per TB processed)
  - Spark: Deeply integrated Apache Spark
  - Synapse Pipelines: Hybrid data integration
  - Studio: Unified user experience

# Azure Synapse Analytics



# 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

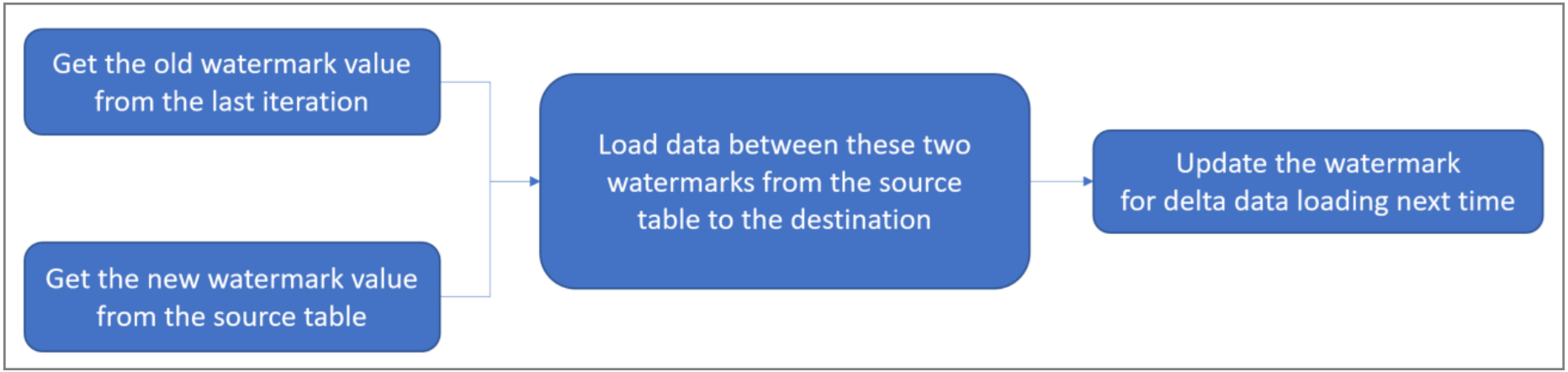
# Backup and Restore in Azure Synapse

- Data warehouse snapshot
  - Creates a restore point you can leverage to recover or copy your data warehouse to a previous state
  - Snapshots are a built-in feature that creates restore points

# Incrementally Load Data

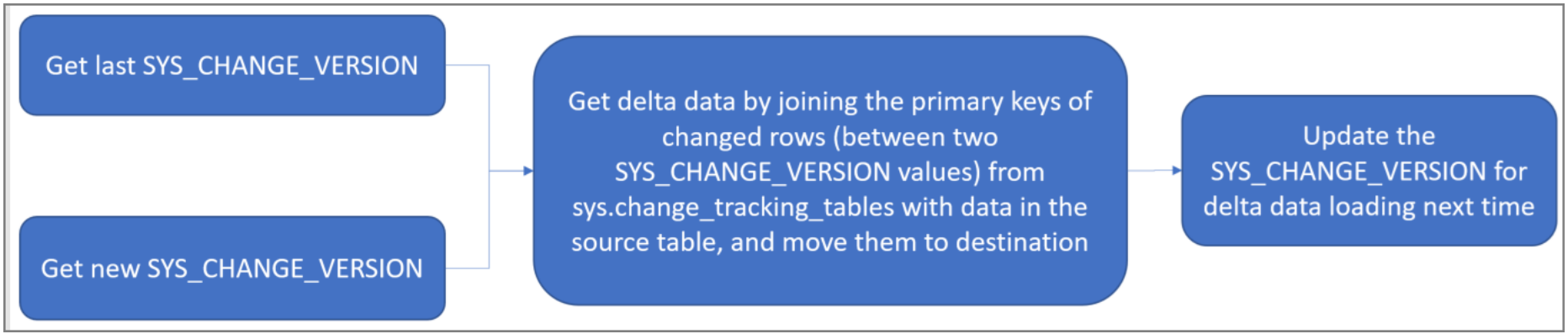
- Methods
  - Delta data loading from database by using a watermark
  - Delta data loading from SQL DB by using the Change Tracking technology
  - Loading new and changed files only by using *LastModifiedDate*
  - Loading new files only by using time partitioned folder or file name

# Using a watermark




See reference

# Using Change Tracking




*See reference*

# Azure Data Factory

- Browse Azure Architectures for Data Factory

# Azure Data Factory

- Pipelines
- Activities

# Transform Data using Azure Data Factory

- Azure SQL Database
- Spark activity

# Source control in Azure Data Factory

- To provide a better authoring experience, Azure Data Factory allows you to configure a Git repository with either Azure Repos or GitHub.

# Azure Data Factory Error Handling

- Handle SQL truncation error
- Troubleshoot Azure Data Factory UX Issues
- Monitor and Alert Data Factory by using Azure Monitor

# Real-time Processing Solutions

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

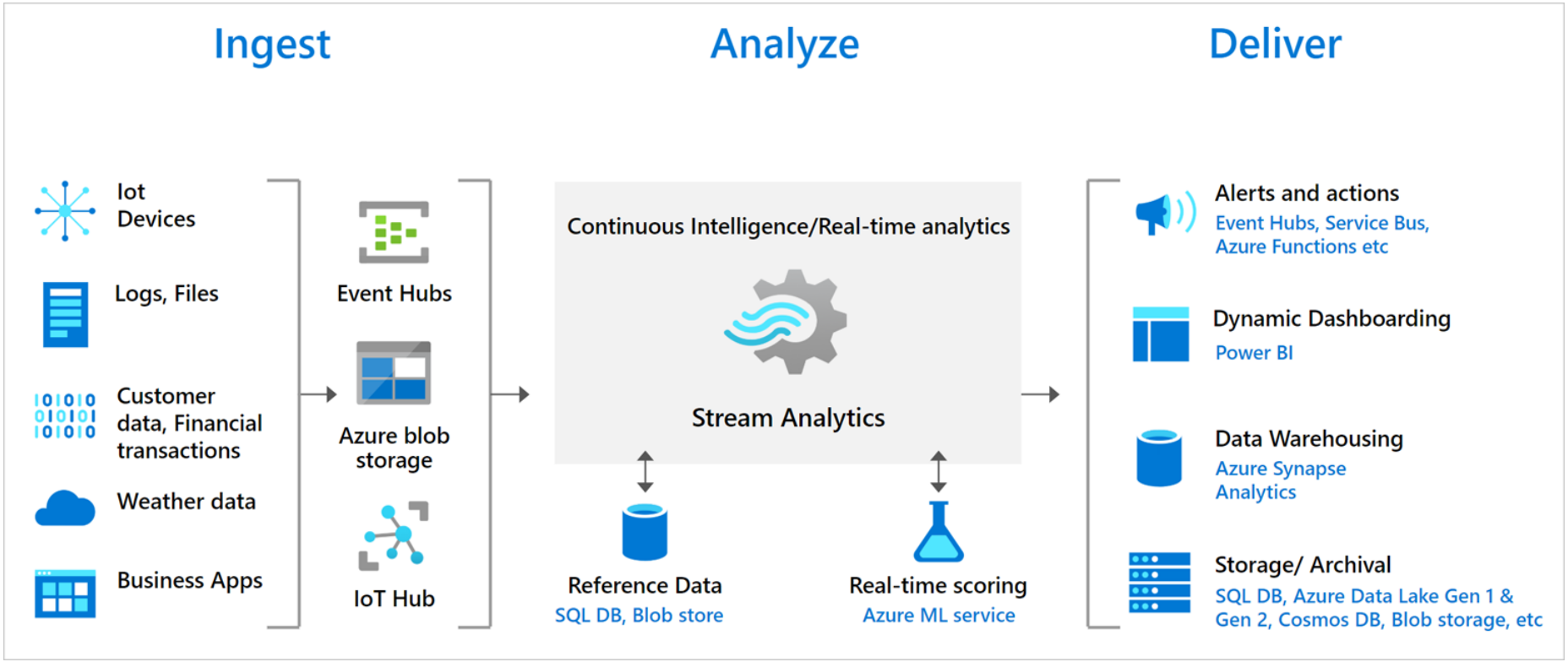
# Azure Stream Analytics

- Browse Azure Architectures for 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.

# Azure Stream Analytics



# Stream Analytics Windowing Functions

- Window types
  - Tumbling
  - Hopping
  - Sliding
  - Session
  - Snapshot

# Stream Analytics Input Types

- Stream input
- Reference input

# Time Handling in Azure Stream Analytics

- Time handling, late arriving data
- Event ordering policies
- Out of order and late-arriving events

# Azure Databricks

- Browse Azure Architectures for Azure Databricks

# Azure Databricks Clusters

- An Azure Databricks cluster is a set of computation resources and configurations on which you run data engineering, data science, and data analytics workloads, such as production ETL pipelines, streaming analytics, ad-hoc analytics, and machine learning.

# Azure Databricks ETL Data

- Using Scala
  - Scala

# Secure, monitor, and optimize data storage and data processing

# Data Security

- 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

# 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

# Data Encryption for Data at Rest and in Transit

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

# Azure compliance documentation

- Azure compliance

# Monitor Data storage and data processing

- Implement logging used by Azure Monitor
- Measure performance of data movement
- Monitor data pipeline performance
- Query Performance Insight for Azure SQL Database
- Monitor cluster performance in Azure HDInsight
- Use Azure Monitor with your Azure Synapse Analytics workspace
- Monitoring Azure Databricks

# Monitor Data storage and data processing

- Collect custom logs with Log Analytics agent in Azure Monitor
- Azure Monitor Metrics overview
- Data spill, data breach
- GDPR Breach Notification
- Azure and Dynamics 365 breach notification under the GDPR

# Monitor Data storage and data processing

- Skewness
- Choose a distribution column with data that distributes evenly
- Determine if the table has data skew
- Troubleshoot performance bottlenecks in Azure Databricks
- Automatic tuning in Azure SQL Database and Azure SQL Managed Instance
- Automatic tuning
- Performance tuning with result set caching
- Known issues for Apache Spark cluster on HDInsight
- Troubleshoot Azure Data Factory

# The Exam

# Questions in DP-203

- 40-60 questions in beta (watch the time!)
- Questions
  - Multiple choice
  - Drag and drop
  - Scenario based
- There is no hands-on lab (might change)
- No negative marking

# DP-203

- Exam DP-203: Data Engineering on Microsoft Azure
- Skills measured

# DP-203 Main Focus (not limited to)

- Azure Data Lake Gen2
- Azure Stream Analytics
- Azure Synapse Analytics
- Azure Data Factory
- Azure Databricks





processing languages, such as SQL, Python, or Scala, and they need to understand parallel processing and data architecture patterns.

**Part of the requirements for:** Microsoft Certified: Azure Data Engineer Associate
**Related exams:** none
**Important:** See details
Go to Certification Dashboard

## Schedule exam

### Exam DP-203: Data Engineering on Microsoft Azure

United States

**Languages:** English
**Retirement date:** none

This exam measures your ability to accomplish the following technical tasks: design and implement data storage; design and develop data processing; design and implement data security; and monitor and optimize data storage and data processing.

**$165 USD***

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

Schedule exam >

## Skills measured

# 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.

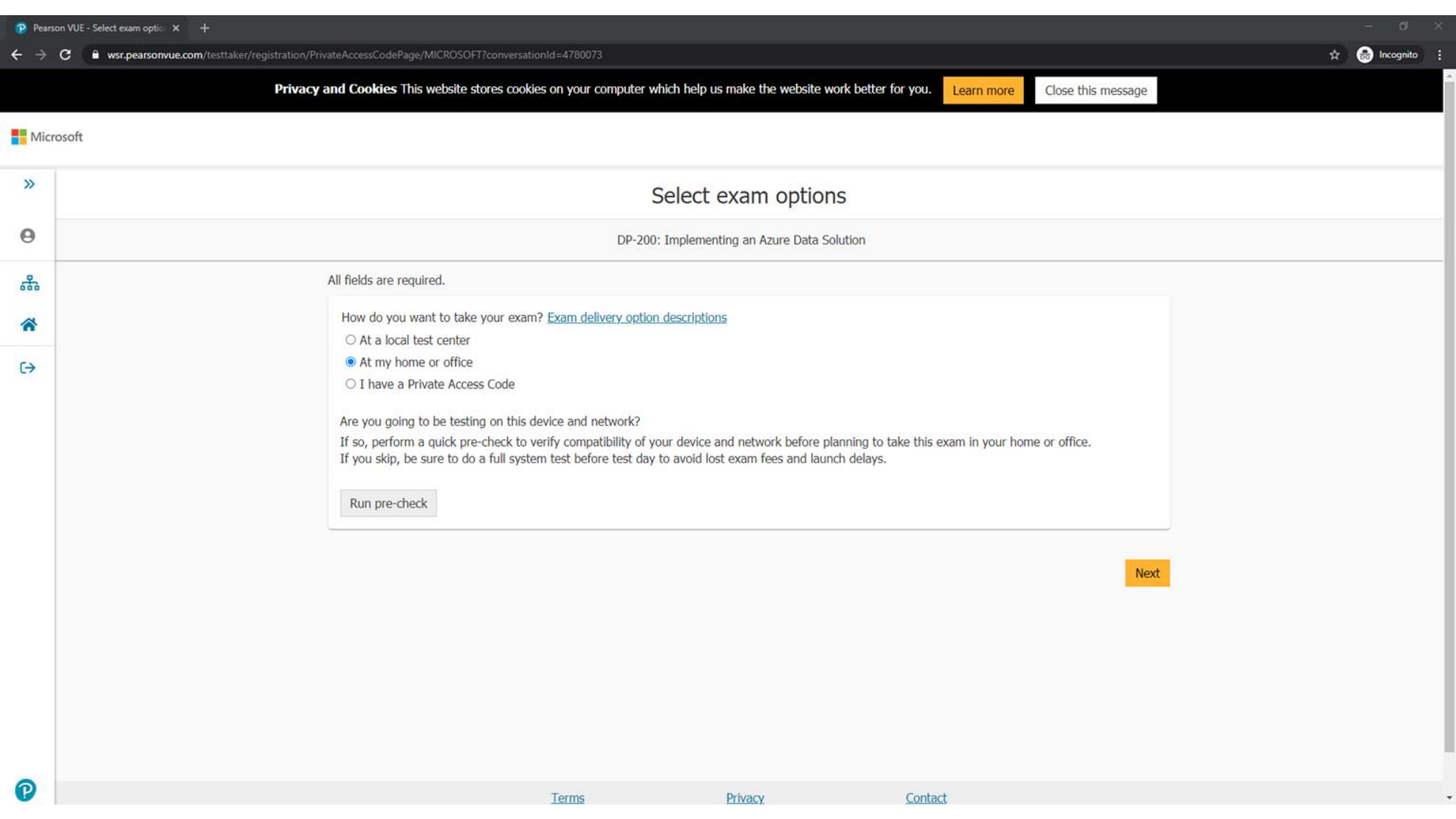
Microsoft Ignite 2019, Orlando

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

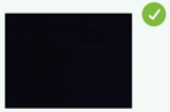
Pearson | VUE

## System check - Checking your requirements

Microphone

Default - Microphone (SI

Internet speed



Webcam

Integrated Webcam (0c4

Next

# Course Repository

https://github.com/zaalion/oreilly-dp-203

O'REILLY®

# Thank you!

**Reza Salehi**

@zaalion