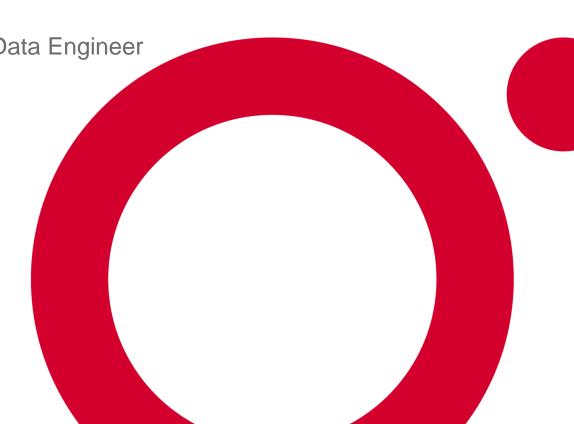
### O'REILLY®

Exam DP-203: Microsoft Azure Data Engineer

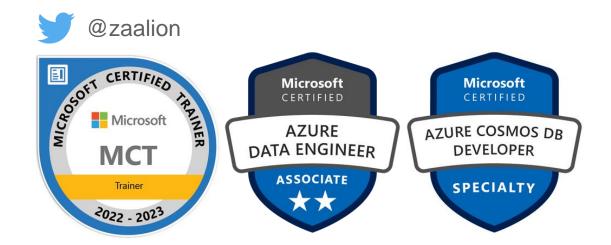
**Associate Crash Course** 

Data Engineering in Microsoft Azure



#### Reza Salehi

**Cloud Consultant** 





### **Course Overview**

#### **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 @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%)

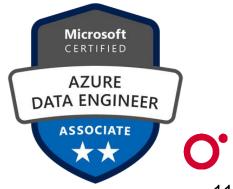


# Design and Implement Data Storage



#### Design and implement data storage (15-20%)

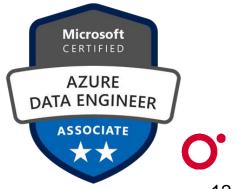
- Implement a partition strategy
- Design and implement the data exploration layer





#### Implement a partition strategy

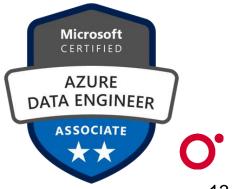
- Implement a partition strategy for files [see <u>1 2 3</u>]
- Implement a partition strategy for analytical workloads [see <u>1 2 3</u>]
- Implement a partition strategy for streaming workloads [see <u>1 2 3</u>]
- Implement a partition strategy for Azure Synapse Analytics [see <u>1</u> <u>2</u>]
- Identify when partitioning is needed in Azure Data Lake Storage Gen2 [see 1]





#### Design and implement the data exploration layer

- Create and execute queries by using a compute solution that leverages SQL serverless and Spark cluster [see <u>1</u> <u>2</u> <u>3</u>]
- Recommend and implement Azure Synapse Analytics database templates [see <u>1</u> <u>2</u>]
- Push new or updated data lineage to Microsoft Purview [see <u>1</u> <u>2</u>]
- Browse and search metadata in Microsoft Purview Data Catalog [see 1 2]



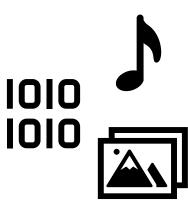
### **Data Types**



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"
  "data": "Click Here",
        "size": 36,
        "style": "bold",
        "name": "text1",
        "hOffset": 250,
        "vOffset": 100,
        "alignment": "center",
        "onMouseUp": "sun1.opacity = (sun1.opacity / 100) * 90;"
}}
```

Semistructured





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**



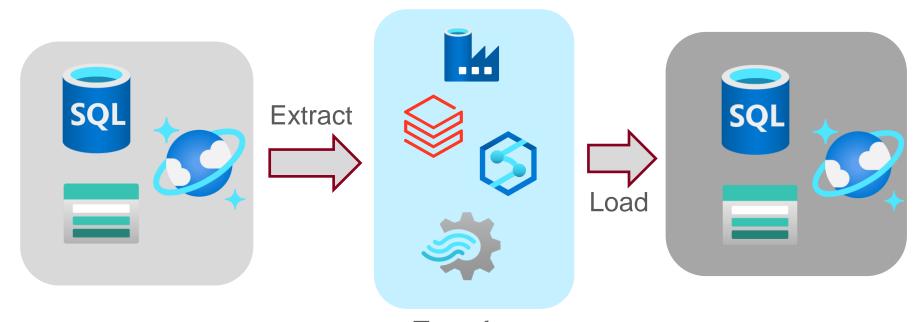
Stream data



Batch data



#### ETL/ELT







#### 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



#### **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.

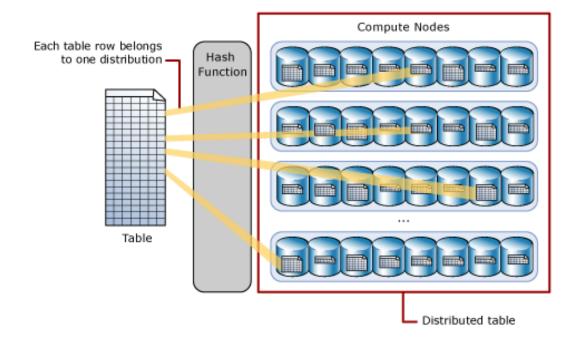


#### **Azure Synapse Analytics Shard**

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

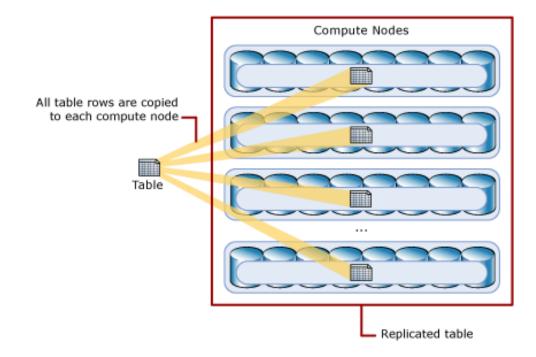


# **Azure Synapse Distributed Tables** (Hash)





# **Azure Synapse Distributed Tables** (Replicated)





# **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



#### **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.

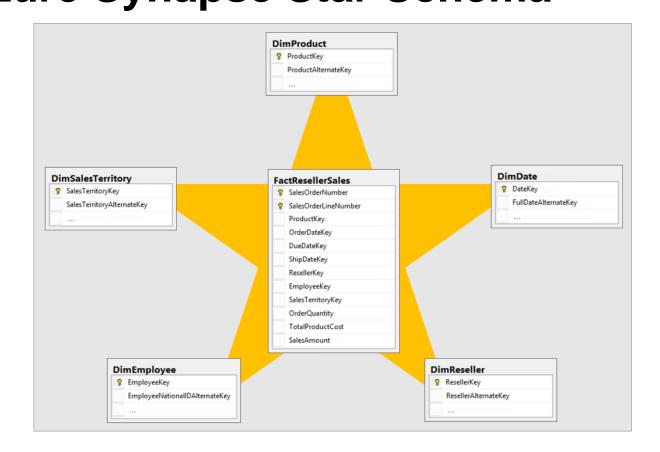


#### **Azure Synapse Star Schema**

- 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

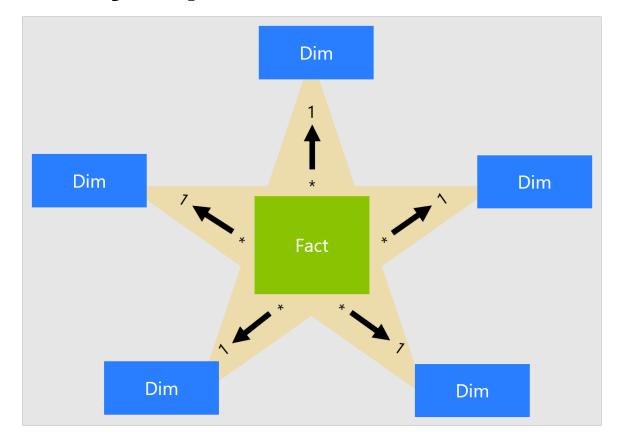


## **Azure Synapse Star Schema**





#### **Azure Synapse Star Schema**





#### **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



#### Type 1 SCD

CustomerID	FirstName	LastName	EmailAddress	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	donna0@aw.com	A Bike Store	2021-03-20	2021-03-20

CustomerID	FirstName	LastName	EmailAddress	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	donna0@aw.com	Bikes, Bikes, Bikes	2021-03-20	2021-03-22



#### Type 3 SCD

CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
<b>3</b>	Donna	Carreras	donna0@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-20

CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	dc3@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-22



#### Type 1 SCD

CustomerID	FirstName	LastName	EmailAddress	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	donna0@aw.com	A Bike Store	2021-03-20	2021-03-20

CustomerID	FirstName	LastName	EmailAddress	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	donna0@aw.com	Bikes, Bikes, Bikes	2021-03-20	2021-03-22



#### Type 6 SCD

SalesRepID	RepSourceld	FirstName	LastName	CurrentRegion	HistoricalRegion	StartDate	EndDate	IsCurrent
1	312	Jun	Cao	Southwest	Southwest	2021-03-20	9999-12-31	True
2	331	Susan	Eaton	Southcentral	Southcentral	2021-03-20	9999-12-31	True

SalesRepID	RepSourceld	FirstName	LastName	CurrentRegion	HistoricalRegion	StartDate	EndDate	IsCurrent
1	312	Jun	Cao	Southwest	Southwest	2021-03-20	9999-12-31	True
2	331	Susan	Eaton	Southeast	Southcentral	2021-03-20	2021-03-21	False
3	331	Susan	Eaton	Southeast	Southeast	2021-03-22	9999-12-31	True



#### **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: <u>First normal form</u>
  - 2NF: <u>Second normal form</u>
  - 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

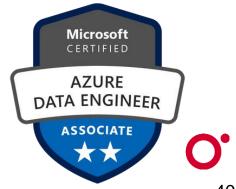


# Develop data processing

#### Develop data processing (40-45%)



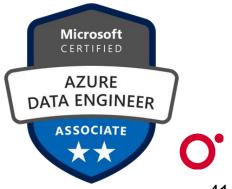
- Ingest and transform data
- Develop a batch processing solution
- Develop a stream processing solution
- Manage batches and pipelines



#### O.

#### Ingest and transform data

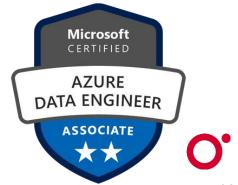
- Design and implement incremental loads [see <u>1</u> <u>2</u>]
- Transform data by using Apache Spark [see 1]
- Transform data by using Transact-SQL (T-SQL) in Azure Synapse Analytics [see <u>1</u> <u>2</u>]
- Ingest and transform data by using Azure Synapse Pipelines or Azure Data Factory [1]
- Transform data by using Azure Stream Analytics [1 2]
- Cleanse data [see 1]
- Handle duplicate data [1 2]
- Avoiding duplicate data by using Azure Stream Analytics Exactly Once Delivery [see 1 2]



#### Ingest and transform data

O.

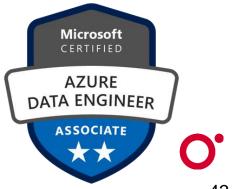
- Handle missing data [see 1]
- Handle late-arriving data [see 1]
- Split data [see 1]
- Shred JSON [see <u>1</u> <u>2</u>]
- Encode and decode data [1]
- Configure error handling for a transformation [1]
- Normalize and denormalize data [see <u>1 2 3]</u>
- Perform data exploratory analysis [see 1]





#### Develop a batch processing solution

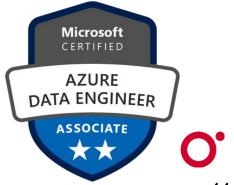
- Develop batch processing solutions by using Azure Data Lake Storage, Azure Databricks, Azure Synapse Analytics, and Azure Data Factory [see 1]
- Use PolyBase to load data to a SQL pool [see 1]
- Implement Azure Synapse Link and query the replicated data [see <u>1</u> <u>2</u>]
- Create data pipelines [1 2 3]
- Scale resources [see <u>1</u> <u>2</u>]
- Configure the batch size [see 1]
- Create tests for data pipelines [see 1]





#### Develop a batch processing solution

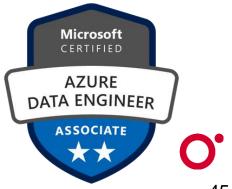
- Integrate Jupyter or Python notebooks into a data pipeline [see 1]
- Upsert data [see <u>1</u> <u>2</u>]
- Revert data to a previous state [see 1]
- Configure exception handling [see 1]
- Configure batch retention [see 1]
- Read from and write to a delta lake [see 1]





#### Develop a stream processing solution

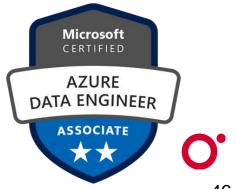
- Create a stream processing solution by using Stream Analytics and Azure Event Hubs [see <u>1</u> <u>2</u>]
- Process data by using Spark structured streaming [see <u>1 2 3</u>]
- Create windowed aggregates [see 1]
- Handle schema drift [see <u>1</u> <u>2</u>]
- Process time series data [see 1]
- Process data across partitions [see <u>1</u> <u>2</u>]
- Process within one partition [see <u>1</u> <u>2</u>]





#### Develop a stream processing solution

- Configure checkpoints and watermarking during processing [see 1]
- Scale resources [see <u>1</u> <u>2</u>]
- Create tests for data pipelines [see 1]
- Optimize pipelines for analytical or transactional purposes [see 1]
- Handle interruptions [see <u>1</u>]
- Configure exception handling [see 1]
- Upsert data [see <u>1</u>]
- Replay archived stream data [see 1]





#### Manage batches and pipelines

- Trigger batches [see 1]
- Handle failed batch loads [see 1]
- Validate batch loads [see 1]
- Manage data pipelines in Azure Data Factory or Azure Synapse Pipelines [see 1]
- Schedule data pipelines in Data Factory or Azure Synapse Pipelines [see 1]
- Implement version control for pipeline artifacts [see 1]
- Manage Spark jobs in a pipeline [see 1]



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



- Relational databases
- Document databases
- Key/Value databases
- Graph databases
- Column family databases

- Object storage
- File share
- Data analytics databases
- Search Engine databases
- <u>Time Series databases</u>













- 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



- 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

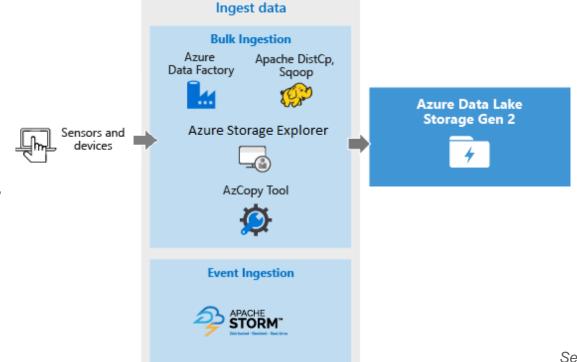


#### **Azure Data Lake Gen2**

- 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 <u>POSIX permissions</u>
  - Cost effective
  - Optimized driver

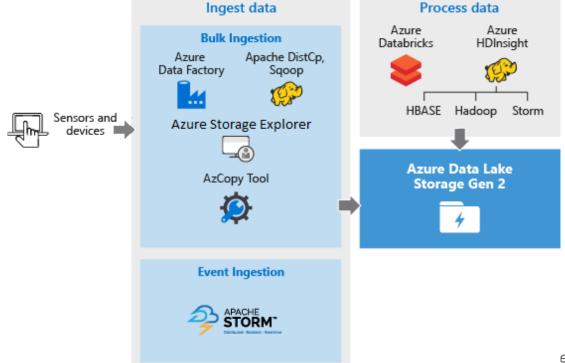


# Data Lake Storage Gen2 for big data requirements



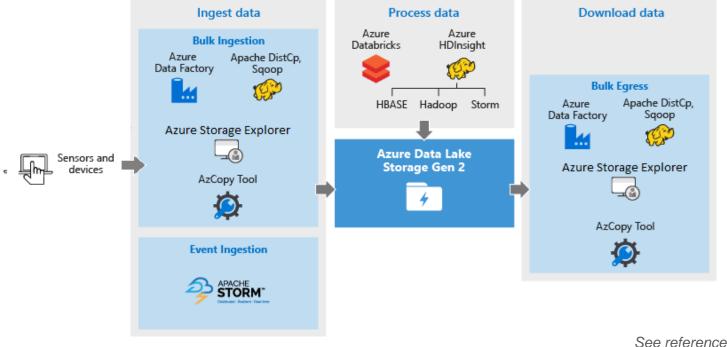


# Data Lake Storage Gen2 for big data requirements





# Data Lake Storage Gen2 for big data requirements





# 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, <u>Access control lists</u> (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.
  - Cold Optimized for storing data that is rarely accessed or modified, but still
    requires fast retrieval. Data in the cold tier should be stored for a minimum of 90
    days. The cold tier has lower storage costs and higher access costs compared to
    the cool tier.
  - 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.



#### Data Lake Storage Gen2 & Blobs

Browse Azure Architectures for Azure Storage



# **Azure Synapse**

Browse Azure Architectures for Azure Synapse



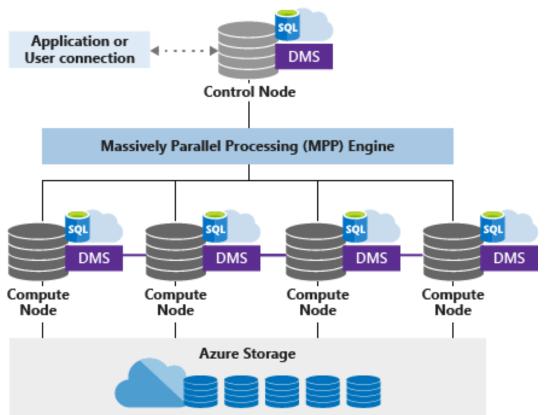
## **Azure Synapse Analytics**

#### 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 <u>Data Factory</u> and <u>Azure Databricks</u>
- 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



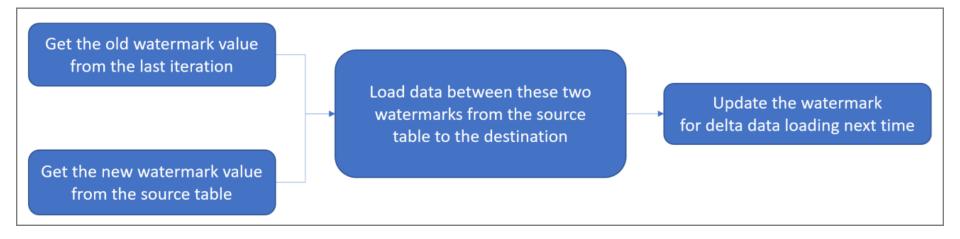
#### **Incrementally Load Data**

#### 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

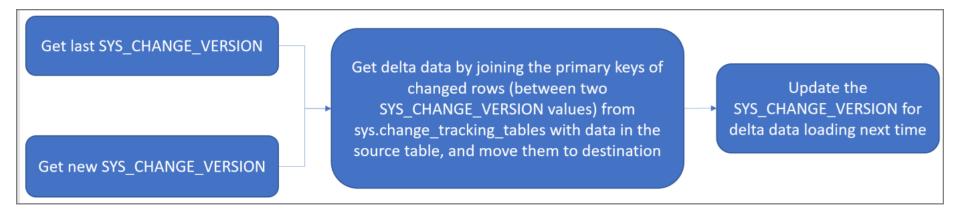


# Using a watermark





# **Using Change Tracking**





# **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 <u>configure a</u>
 <u>Git repository</u> 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 <u>Stream Analytics</u> and <u>Azure Databricks</u>
- Design and <u>provision compute resources</u>



#### **Azure Stream Analytics**

Browse Azure Architectures for Azure Stream Analytics

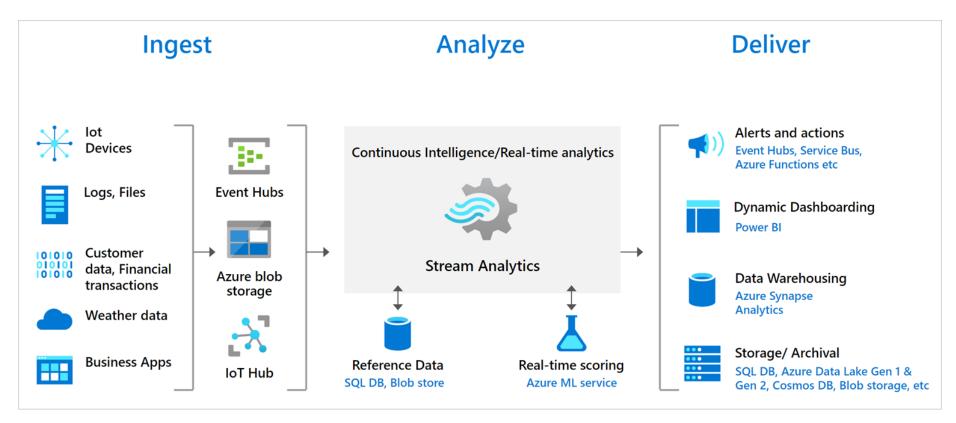


#### **Develop Streaming Solutions**

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

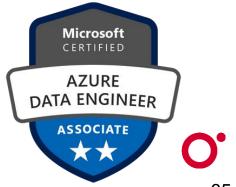


# Secure, monitor, and optimize data storage and data processing



## Secure, monitor, and optimize data storage and data processing (30–35%)

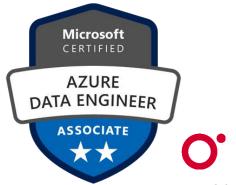
- Implement data security
- Monitor data storage and data processing
- Optimize and troubleshoot data storage and data processing





#### Implement data security

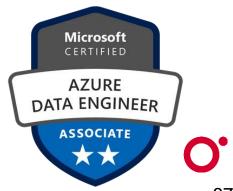
- Implement data masking [see 1]
- Encrypt data at rest and in motion [see 1]
- Implement row-level and column-level security [see <u>1 2 3</u>]
- Implement Azure role-based access control (RBAC) [see 1]
- Implement POSIX-like access control lists (ACLs) for Data Lake Storage Gen2 [see 1]
- Implement a data retention policy [see 1]
- Implement secure endpoints (private and public) [see <u>1 2</u>]
- Implement resource tokens in Azure Databricks [see 1]
- Load a DataFrame with sensitive information [see 1]
- Write encrypted data to tables or Parquet files [see 1 2]
- Manage sensitive information [see <u>1</u>]





#### Monitor data storage and data processing

- Implement logging used by Azure Monitor [see <u>1</u> <u>2</u>]
- Configure monitoring services [see 1]
- Monitor stream processing [see 1]
- Measure performance of data movement [see 1]
- Monitor and update statistics about data across a system [see 1]
- Monitor data pipeline performance [see 1]
- Measure query performance [see <u>1</u> <u>2</u>]
- Schedule and monitor pipeline tests [see 1]
- Interpret Azure Monitor metrics and logs [see 1]
- Implement a pipeline alert strategy [see 1]





### Optimize and troubleshoot data storage and data processing

- Compact small files [see 1]
- Handle skew in data [see 1]
- Handle data spill [see <u>1</u> <u>2</u>]
- Optimize resource management [see 1]
- Tune queries by using indexers [see <u>1</u>]
- Tune queries by using cache [see 1]
- Troubleshoot a failed Spark job [see <u>1</u> <u>2</u>]
- Troubleshoot a failed pipeline run, including activities executed in external services [see 1]



#### Plan for Secure Endpoints

- Secure endpoints:
  - Azure Storage Account
  - Azure Synapse Analytics
  - Azure Data Factory
  - Azure Databricks
  - Azure Stream Analytics
  - Azure Event Hubs



## Data Encryption for Data at Rest and in Transit

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



#### Azure compliance documentation

Azure compliance



#### The Exam

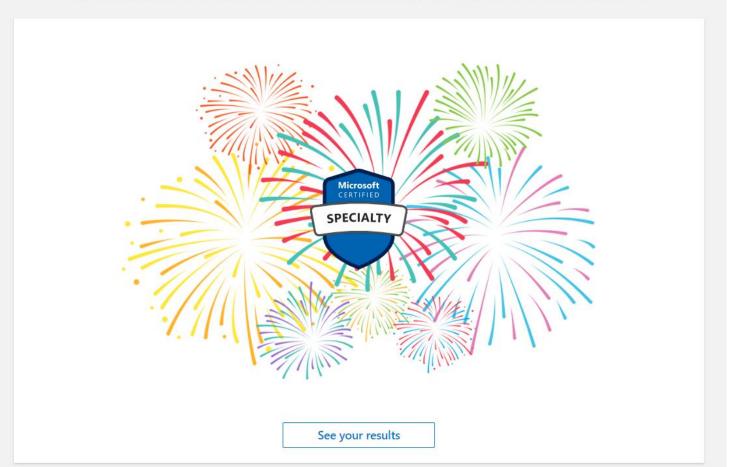
#### **Questions in DP-203**

- 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



#### Congratulations, you passed!

You've renewed your Microsoft Certified: Azure Cosmos DB Developer Specialty and have extended it by one year.





#### **DP-203**

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



#### **DP-203**

#### (i) Important

The English language version of this exam was updated on November 2, 2023. Review the study guide linked in the "Tip" box for details about the latest changes. If a localized version of this exam is available, it will be updated approximately eight weeks after this date. While Microsoft makes every effort to update localized versions as noted, there may be times when the localized versions of this exam are not updated on this schedule.

Passing score: 700. Learn more about exam scores.



#### **DP-203 Main Focus**

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



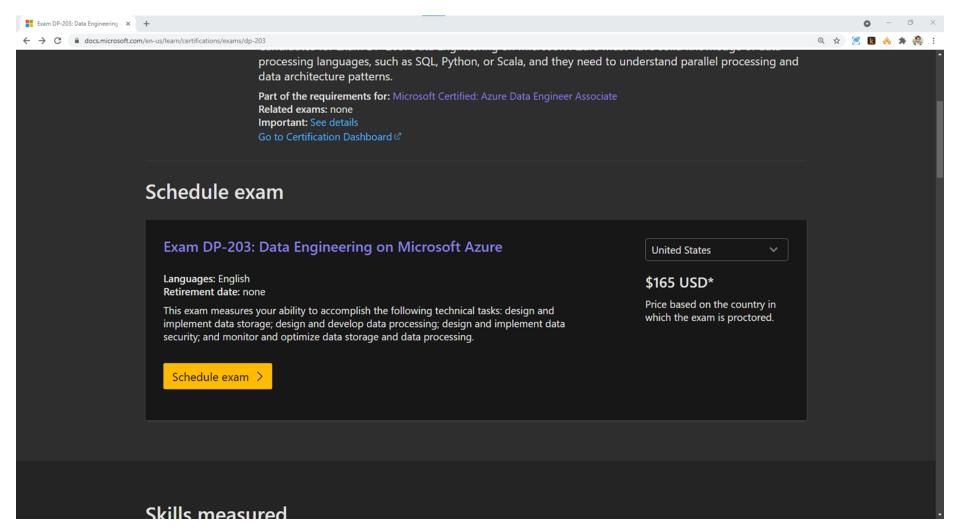














Contact us Privacy & Cookies Terms of use Trademarks Accommodations

♣ Incognito :





#### **Course Repository**

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



## O'REILLY® Thank you!

Reza Salehi

@zaalion

