Study Guide

Exam DP-203: Data Engineering on Microsoft Azure

Purpose of this document

This study guide should help you understand what to expect on the exam and includes a summary of the topics the exam might cover and links to additional resources. The information and materials in this document should help you focus your studies as you prepare for the exam.

Useful links	Description
How to earn the certification	Some certifications only require one exam, while others require more. On the details page, you'll find information about what skills are measured and links to registration. Each exam also has its own details page covering exam specifics.
Certification renewal	Once you earn your certification, don't let it expire. When you have an active certification that's expiring within six months, you should renew it—at no cost—by passing a renewal assessment on Microsoft Learn. Remember to renew your certification annually if you want to retain it.
Your Microsoft Learn profile	Connecting your certification profile to Learn brings all your learning activities together. You'll be able to schedule and renew exams, share and print certificates, badges and transcripts, and review your learning statistics inside your Learn profile.
Passing score	All technical exam scores are reported on a scale of 1 to 1,000. A passing score is 700 or greater. As this is a scaled score, it may not equal 70% of the points. A passing score is based on the knowledge and skills needed to demonstrate competence as well as the difficulty of the questions.
Exam sandbox	Are you new to Microsoft certification exams? You can explore the exam environment by visiting our exam sandbox. We created the sandbox as an opportunity for you to experience an exam before you take it. In the sandbox, you can interact with different question types, such as build list, case studies,

Useful links	Description
	and others that you might encounter in the user interface when you take an exam. Additionally, it includes the introductory screens, instructions, and help topics related to the different types of questions that your exam might include. It also includes the non-disclosure agreement that you must accept before you can launch the exam.
Request accommodations	We're committed to ensuring all learners are set up for success. If you use assistive devices, require extra time, or need modification to any part of the exam experience, you can request an accommodation.
Take a practice test	Taking a practice test is a great way to know whether you're ready to take the exam or if you need to study a bit more. Subject-matter experts write the Microsoft Official Practice Tests, which are designed to assess all exam objectives.

Objective domain: skills the exam measures

The English language version of this exam was released on May 4, 2021.

Some exams are localized into other languages, and those are updated approximately eight weeks after the English version is updated. Other available languages are listed in the **Schedule Exam** section of the **Exam Details** webpage. If the exam isn't available in your preferred language, you can request an additional 30 minutes to complete the exam.

> **Note**
> The bullets that follow each of the skills measured are intended to illustrate how we are assessing that skill. Related topics may be covered in the exam.

> **Note**
> Most questions cover features that are general availability (GA). The exam may contain questions on Preview features if those features are commonly used.

Skills measured

- Design and implement data storage (40–45%)
- Design and develop data processing (25–30%)
- Design and implement data security (10–15%)
- Monitor and optimize data storage and data processing (10–15%)

Functional groups

Design and implement data storage (40–45%)

Design a data storage structure

- Design an Azure Data Lake solution
- Recommend file types for storage
- Recommend file types for analytical queries
- Design for efficient querying
- Design for data pruning
- Design a folder structure that represents the levels of data transformation
- Design a distribution strategy
- Design a data archiving solution

Design a partition strategy

- Design a partition strategy for files
- Design a partition strategy for analytical workloads
- Design a partition strategy for efficiency/performance
- Design a partition strategy for Azure Synapse Analytics
- Identify when partitioning is needed in Azure Data Lake Storage Gen2

Design the serving layer

- Design star schemas
- Design slowly changing dimensions
- Design a dimensional hierarchy
- Design a solution for temporal data
- Design for incremental loading
- Design analytical stores
- Design metastores in Azure Synapse Analytics and Azure Databricks

Implement physical data storage structures

- Implement compression
- Implement partitioning
- Implement sharding
- Implement different table geometries with Azure Synapse Analytics pools
- Implement data redundancy
- Implement distributions
- Implement data archiving

Implement logical data structures

- Build a temporal data solution
- Build a slowly changing dimension
- Build a logical folder structure
- Build external tables
- Implement file and folder structures for efficient querying and data pruning

Implement the serving layer

- Deliver data in a relational star
- Deliver data in Parquet files
- Maintain metadata
- Implement a dimensional hierarchy

Design and develop data processing (25–30%)

Ingest and transform data

- Transform data by using Apache Spark
- Transform data by using Transact-SQL
- Transform data by using Data Factory
- Transform data by using Azure Synapse Pipelines
- Transform data by using Stream Analytics
- Cleanse data
- Split data
- Shred JSON
- Encode and decode data
- Configure error handling for the transformation
- Normalize and denormalize values
- Transform data by using Scala
- Perform data exploratory analysis

Design and develop a batch processing solution

- Develop batch processing solutions by using Data Factory, Data Lake, Spark, Azure Synapse Pipelines, PolyBase, and Azure Databricks
- Create data pipelines
- Design and implement incremental data loads
- Design and develop slowly changing dimensions
- Handle security and compliance requirements
- Scale resources
- Configure the batch size
- Design and create tests for data pipelines

- Integrate Jupyter/Python notebooks into a data pipeline
- Handle duplicate data
- Handle missing data
- Handle late-arriving data
- Upsert data
- Regress to a previous state
- Design and configure exception handling
- Configure batch retention
- Design a batch processing solution
- Debug Spark jobs by using the Spark UI

Design and develop a stream processing solution

- Develop a stream processing solution by using Stream Analytics, Azure Databricks, and Azure Event Hubs
- Process data by using Spark structured streaming
- Monitor for performance and functional regressions
- Design and create windowed aggregates
- Handle schema drift
- Process time series data
- Process across partitions
- Process within one partition
- Configure checkpoints/watermarking during processing
- Scale resources
- Design and create tests for data pipelines
- Optimize pipelines for analytical or transactional purposes
- Handle interruptions
- Design and configure exception handling
- Upsert data
- Replay archived stream data
- Design a stream processing solution

Manage batches and pipelines

- Trigger batches
- Handle failed batch loads
- Validate batch loads
- Manage data pipelines in Data Factory/Synapse Pipelines
- Schedule data pipelines in Data Factory/Synapse Pipelines
- Implement version control for pipeline artifacts
- Manage Spark jobs in a pipeline

Design and implement data security (10–15%)

Design security for data policies and standards

- Design data encryption for data at rest and in transit
- Design a data auditing strategy
- Design a data masking strategy
- Design for data privacy
- Design a data retention policy
- Design to purge data based on business requirements
- Design Azure role-based access control (Azure RBAC) and POSIX-like Access Control List (ACL) for Data Lake Storage Gen2
- Design row-level and column-level security

Implement data security

- Implement data masking
- Encrypt data at rest and in motion
- Implement row-level and column-level security
- Implement Azure RBAC
- Implement POSIX-like ACLs for Data Lake Storage Gen2
- Implement a data retention policy
- Implement a data auditing strategy
- Manage identities, keys, and secrets across different data platform technologies
- Implement secure endpoints (private and public)
- Implement resource tokens in Azure Databricks
- Load a DataFrame with sensitive information
- Write encrypted data to tables or Parquet files
- Manage sensitive information

Monitor and optimize data storage and data processing (10–15%)

Monitor data storage and data processing

- Implement logging used by Azure Monitor
- Configure monitoring services
- Measure performance of data movement
- Monitor and update statistics about data across a system
- Monitor data pipeline performance
- Measure query performance
- Monitor cluster performance
- Understand custom logging options
- Schedule and monitor pipeline tests

- Interpret Azure Monitor metrics and logs
- Interpret a Spark directed acyclic graph (DAG)

Optimize and troubleshoot data storage and data processing

- Compact small files
- Rewrite user-defined functions (UDFs)
- Handle skew in data
- Handle data spill
- Tune shuffle partitions
- Find shuffling in a pipeline
- Optimize resource management
- Tune queries by using indexers
- Tune queries by using cache
- Optimize pipelines for analytical or transactional purposes
- Optimize pipeline for descriptive versus analytical workloads
- Troubleshoot a failed spark job
- Troubleshoot a failed pipeline run

Study Resources

We recommend that you train and get hands-on experience before you take the exam. We offer self-study options and classroom training as well as links to documentation, community sites, and videos.

Study resources	Links to learning and documentation
Get trained	Choose from self-paced learning paths and modules or take an instructor led course
Find documentation	Azure Data Lake Storage Azure Synapse Analytics Azure Databricks Data Factory Azure Stream Analytics Event Hubs Azure Monitor
Ask a question	Microsoft Q&A | Microsoft Docs
Get community support	Analytics on Azure | TechCommunity Azure Synapse Analytics | TechCommunity

Study resources	Links to learning and documentation
Follow Microsoft Learn	Microsoft Learn - Microsoft Tech Community
Find a video	Data Exposed Browse other Microsoft Learn shows