

# Data Core Concepts *CheatSheet*

Cheat sheets, Practice Exams and Flash cards 🖱️ [www.examprom.co/dp-900](http://www.examprom.co/dp-900)

Exam

Pro

**Data** — units of information

**Data Documents** — types of abstract groupings of data

**Data Sets** — unstructured logical grouping of data

**Data Structures** — structured data

- Unstructured — a bunch of loose data that has no organization or possibly relation
  - Flat files — various files that can reside in a file system
- Semi Structured — data that can be browsed or searched (with limitations) eg. CSV, XML, JSON, Parquet
  - XML — markup language that looks like html eg. <hello><world>earth</world></hello>
  - JSON — a text file that is composed of dictionaries and arrays eg. {"hello": ["earth", "mars"]}
  - RCFiles — a storage format designed for MapReduce framework
  - ORC — a columnar data structure, 75% more efficient than RCFiles, limited compatibility, works great well with HIVE
  - AVRO — a row-wise data structure for hadoop systems
  - Parquet — a columnar data-structure that has more support for hadoop systems than ORC
- Structured — data that can be easily browsed or searched eg. tabular data
  - Tabular data — data that is arranged as tables, *think excel spreadsheets*

**Data Types** — how single units of data are intended to be used

**Database Administrator** — **configures and maintains a databases** eg. Azure Data services or SQL server.

**Data Engineer** — **Design and implement data tasks** related to the transfer and storage of **big data**

**Data Analyst** — **Analyzes business data** to reveal important information

**Software as a Service (SaaS)** — A product that is run and managed by the service provider

**Platform as a Service (PaaS)** — Focus on the deployment and management of your apps.

**Infrastructure as a Service (IaaS)** — basic building blocks for cloud IT. Provides access to networking, computers and data storage space.

# Data Core Concepts *CheatSheet*

Cheat sheets, Practice Exams and Flash cards 🖱️ [www.exampro.co/dp-900](http://www.exampro.co/dp-900)

Exam

Pro

- Data Modeling** — an abstract model that organizes elements of data and standardizes how they relate to one another and to real-world entities
- Schema** — a formal language to describe the structure of data used by databases and data stores during the data modeling phase
- Schemaless** — generally used when upfront data modelling can be forgo because the schema is flexible, normally used with NoSQL databases
- Data Integrity** — the maintenance and assurance of, data accuracy and consistency over its entire life-cycle.
- Data Corruption** — the act or state of data not being in the intended state and will result in data loss or misinformation
- Normalized** — A schema design to store non-redundant and consistent data
- Denormalized** — A schema that combines data so that accessing data (querying) is fast
- Extract, Transform and Load (ETL)** — transform data from one data store to another, loads data in an intermediate stage, doesn't work with data lakes
- Extract, Load and Transform (ELT)** — transformations done at the target data store, works with data lakes, more common in cloud services.
- Query** — when a user requests data from a data store by using a query language to return a data result
- Data Source** — A data source is **where data originates from**. analytics and data warehouses tools may be connected to various data sources
- Data consistency** — When data being kept in two different place and **whether the data exactly match** or do not match
- **Strongly Consistent** — Every time you request data (query) you can expect consistent data to be returned within a time
  - **Eventually Consistent** — When you request data you may get back inconsistent data (stale data)
- Synchronization** — continuous stream of data that is synchronized by a timer or clock (guarantee of time)
- Asynchronization** — continuous stream of data separated by start and stop bits (no guarantee of time)
- Data Mining** — The **extraction of patterns and knowledge** from large amounts of data (**not the extraction of data itself**)

# Data Core Concepts *CheatSheet*

Cheat sheets, Practice Exams and Flash cards 🖱️ [www.exampro.co/dp-900](http://www.exampro.co/dp-900)

Exam

Pro

**Data Analytics** — Data analytics is **examining, transforming, and arranging data** so that you can **extract and study useful information**.

**Key Performance Indicators** — type of performance measurement that a company or organization to determine **performance over time**

**Descriptive Analytics** (What happened?) — Accurate, comprehensive , live-data and effective visualizations eg. dashboards, reports, KPI, ROI

**Diagnostic Analytics** (Why did it happen?) — drill down to investigate root cause, focused on subset of descriptive analytics dataset

**Predictive Analytics** (What will happen?) — use historical data with statistics and ML to generate trends or predictions

**Predictive Analytics** (What will happen?) — using hybrid data with ML to predict future scenarios that are exploitable

**Cognitive Analytics** (What-if this happens?) — Using ML and NLP to determine what-if scenarios to create plans if they happen

**One Drive** — storage and storage synchronization service for a single user

**SharePoint** — storage and storage synchronization service for an organization

# Azure Synapse and Data Lake *CheatSheet*

Cheat sheets, Practice Exams and Flash cards 📄 [www.examprom.co/dp-900](http://www.examprom.co/dp-900)

Exam

Pro

A data lake is a **centralized data repository for unstructured and semi-structured data**

- A Data Lake is intended to store vast amounts of data
- Data lakes generally use object (blobs) or files as its storage medium.

Azure Data Lake Store (Gen 2)

- Azure Blob storage which has been extended to support big data analytics workloads
- In order to efficiently access data, Data Lake Storage adds a **hierarchical namespace** to Azure Blob Storage
  - ACLs, Throttling Management, Performance Optimizers
- You access the data lake via (Blob) wasbs:// or (File system) abfs://

**Azure Synapse Analytics** — a **data warehouse** and **unified analytics platform**

- Has two underlying transformations engines: SQL Pools and Spark Pools
- Synapse SQL is T-SQL but designed to be distributed
  - SQL Dedicated Pools — reserves compute for processing
  - Serverless Endpoints — on-demand, no guarantee of performance
- Data is stored on Azure Data Lake Store (Gen 2)
- Operations are performed within the Azure Synapse Studio
- PolyBase — enables your SQL Server instance to query data with T-SQL (used to connect many relational data sources)

# Account Storage *CheatSheet*

Cheat sheets, Practice Exams and Flash cards 🖱️ [www.exampro.co/dp-900](http://www.exampro.co/dp-900)

Exam

Pro

**Azure Storage Accounts** — an umbrella service for various forms of managed storage:

- Azure Tables
- Azure Blob Storage
- Azure Files

**Azure Blob Storage** — Object storage that is distributed across many machines.

- Supports 3 types:
  - Blob blobs — store text and binary data, blocks of data that can be managed individually, up to 4.7TiB
  - Append blobs — Optimized for append operations, ideal for logging
  - Page blobs — store random access files up to 8 TB in size.

**Azure Files** is a fully managed **file share** in the cloud.

- To connect to the file share a **network protocol** is used:
  - Server Message Block (SMB)
  - Network File System (NFS)

**Azure Storage Explorer** — a standalone cross-platform app to access various storage formats within Azure Storage accounts

# Power BI *CheatSheet*

Exam

Pro

Cheat sheets, Practice Exams and Flash cards 🖱️ [www.exampromco.co/dp-900](https://www.exampromco.co/dp-900)

---

**Business Intelligence (BI)** — both a data-analysis strategy and **technology** for business info. helps organizations make data-driven decisions

**Power BI Desktop** — A desktop app to design interactive reports from various data sources and can be published to Power BI Service

**Power BI Service** — A web-app to view reports, and create interactive shareable dashboards by pinning various dataset and report visualizations

**Power BI Mobile** — a mobile web-app to view BI reports on the go

**Power BI Report Builder** — windows application build pixel-perfect printable reports (used to build paginated reports)

**Power BI Embedded** — embed Power BI visualizations into web-apps

**Interactive Reports** — Reports in Power BI, drag visualizations, load data from many data sources (Both in Desktop and Service)

**Paginated Reports** — pixel-perfect printable report file format. Tabular data laid out in page format

**Dashboards** — Build sharable dashboards by pinning various Power BI visualizations (a single page report designed for a screen) Only Service

**Dashboard Tiles** — A tile represent a visualization that has been pinned to a dashboard

**Visualizations** — A visualization is a chart or graph that is backed by a dataset.



# Relational Databases *CheatSheet*

Exam

Pro

Cheat sheets, Practice Exams and Flash cards 📄 [www.exampro.co/dp-900](http://www.exampro.co/dp-900)

- Structured Query Language (SQL)** — designed to **access and maintain data for a relational database management system (RDBMS)**
- Online Transaction Processing (OLTP)** — frequent and short queries for transactional information eg. Databases
- Online Analytical Processing (OLAP)** — complex queries for large databases to produce reports and analytics eg. Data Warehouses
- MySQL** — MySQL is a **pure relational database (RDBMS)** it is easy to setup and use, most popular open-source relational database
- MariaDB** — MariaDB is an fork of MySQL
- Postgres** — Postgres is an object-relational database (ORDBMS), it is more advanced and well liked among developers
- Read Replicas** — a duplicate of your database kept in-sync with the main to help to reduce reads on your primary databases
- Azure SQL** — An umbrella service for different offerings of MS SQL databases hosting services
- **SQL VMs** — for lift-and-shift when you want OS access and control, or you need to bring-your-own-license (BYOL) for Azure Hybrid Benefit
  - **Managed SQL** — for lift-and-shift when you the broadest amount of compatibility with SQL versions
    - you can use Azure Arc to run this service on-premise
    - gives you many of the benefit of a fully-managed databases
  - **SQL Databases** — Fully managed SQL databases
    - Run a single server
    - Run as a database (collection of servers)
    - Run in an Elastic Pool (databases of different sizes residing on one server to save costs)
- Connection Policy**
- Three modes:
    - Default — choose Proxy or Redirect initially depending on if the server is within or outside the Azure Network
    - Proxy — outside the Azure network, proxied through a gateway
      - listen on port 1443 when connecting via Proxy mode through a gateway outside the Azure Network
    - Redirect — redirected within the Azure Network

# T-SQL *CheatSheet*

Cheat sheets, Practice Exams and Flash cards 🖱️ [www.exampro.co/dp-900](https://www.exampro.co/dp-900)

Exam

Pro

**Transact-SQL (T-SQL)** is a set of programming extensions from Sybase and Microsoft that add several features to the Structured Query Language (SQL).

For Microsoft SQL Server there are five groups of SQL Commands:

- Data **Definition** Language (DDL)
  - used to define the database schema
- Data **Query** Language (DQL)
  - used for performing queries on the data
- Data **Manipulation** Language (DML)
  - manipulation of data in the database
- Data **Control** Language (DCL)
  - rights, permissions and other controls of the database
- **Transaction Control** Language (TCL)
  - transactions within the database



# Database Security *CheatSheet*

Cheat sheets, Practice Exams and Flash cards 🖱️ [www.examprom.co/dp-900](https://www.examprom.co/dp-900)

Exam

Pro

## MS SQL Database Authentication

Two modes when setting up MS SQL server (remoted into Windows Machine):

- Windows Authentication mode — enables Windows Authentication and disables SQL Server Authentication
- Mixed mode — enables both Windows Authentication and SQL Server Authentication
- Windows Authentication (recommended) — authenticate via windows users
- SQL Server Authentication — username and password, connect from anywhere

## Network Connectivity

- Public Endpoint — reachable outside the Azure Network over the internet (use server firewall for protection)
- Private Endpoint — only reachable within the Azure Network (use Azure PrivateLinks to keep traffic within Azure Network)

**Azure Defender SQL** — a unified package for advanced SQL security capabilities for **Vulnerability Assessment** and **Advanced Threat Protection**

**Server Firewall Rules** — an internal firewall that resides on the database server, All connections are **rejected by default** to database

**Always Encrypted** — a feature that encrypts columns in an Azure SQL Database or SQL Server

Role-Based-Access-Control (RBAC) for databases:

**SQL DB Contributor** — manage SQL databases, but not access to them, can't manage their security-related policies or their parent SQL servers

**SQL Managed Instance Contributor** — manage SQL Managed Instances and required network configuration, can't give access to others

**SQL Security Manager** — manage the security-related policies of SQL servers and databases, but not access to SQL servers

**SQL Server Contributor** — manage SQL servers and databases, but not access to them SQL servers

**Transparent Data Encryption (TDE)** — **encrypts data-at-rest** for Microsoft Databases

**Dynamic Data Masking** — you can choose your database columns to that will be masked (obscured) for specific users

**Azure Private Links** — allows you to **establish secure connections** between Azure resources so traffic **remains within the Azure Network**

# Azure Tables and CosmosDB CheatSheet

Cheat sheets, Practice Exams and Flash cards 🖱️ [www.exampopro.co/dp-900](http://www.exampopro.co/dp-900)



**Azure Tables** — a key / value data store

- can be hosted on Account Storage, its designed for a single region and single table
- can be hosted on CosmosDB, its designed for scale across multiple regions

**CosmosDB** — A fully-managed NoSQL service that supports multiple NoSQL engines called APIs

- Core SQL API (default) — a document database, you can use SQL to query documents
- Graph API — a graph databases, you can use Gremlin to traverse the nodes and edges
- MongoDB API — a mongodb database (document database)
- Tables API — Azure Tables Key/Value

Apache TinkerPop — an open-source framework to have an agnostic way to talk to many graph databases

- Gremlin — graph traversal language to traverse nodes and edges

MongoDB — an open-source document database

- Binary JSON (BSON) — An storage and compute optimized version of JSON, introduces new data types

CosmosDB Explorer — a web-ui to view cosmos databases

# Hadoop *CheatSheet*

Cheat sheets, Practice Exams and Flash cards 🖱️ [www.examprom.co/dp-900](http://www.examprom.co/dp-900)

Exam

Pro

Apache Hadoop — open-source framework for **distributed processing of large data sets**

- **Hadoop Distributed File System (HDFS)** — a resilient and redundant file storage distributed on clusters of common hardware
- **Hadoop MapReduce** — writes apps that can process multi-terabyte data in-parallel on large clusters of common hardware
- **Hbase** — a distributed, scalable, big data store
- **YARN** — manages resources, nodes, containers and performs scheduling
- **HIVE** — used for generating reports using an **SQL** language
- **PIG** — A high-level **scripting** language to write complex data transformations

**Apache Spark** — can perform is 100x faster in memory and 10x faster than disk than Hadoop, supports ETLs, Streaming and ML flows

**Apache Kafka** — a streaming pipeline and analytics service

**HDInsights** — is managed service to run popular open-source analytics service. It is fully-managed hadoop system

# Apache Spark and DataBricks *CheatSheet*

Cheat sheets, Practice Exams and Flash cards 📄 [www.exampro.co/dp-900](http://www.exampro.co/dp-900)

Exam

Pro

Apache Spark —an open-source **unified analytics engine** for **big data and machine learning**

- 100x faster in memory than hadoop
- 10x faster than disk than hadoop
- perform ELT (batch), streaming and ML workloads
- The Apache ecosystem is composed of:
  - **Spark Core** — The underlying engine and API.
  - **Spark SQL** — Use SQL and also a new data structure called DataFrame to work with data
  - **Spark Streaming** — ingest data from many streaming services
  - **GraphX** — distributed graph-processing framework
  - **Machine Learning Library (MLib)** — a distributed machine-learning framework
- **Resilient Distributed Dataset (RDD)** is a domain specific language (DSL) to execute various parallel operations on an Apache Spark cluster.

Databricks is a software company specializing in **providing fully managed Apache Spark clusters**

Azure Databricks is a **partnership between Microsoft and Databricks** to offer the **Databricks Platform within the Azure Portal** running on Azure compute services

- Azure Databricks offers two environments:
  - **Azure Databricks Workspace** —DataBrick Platform with integrations to **Azure data-related services** for building big data pipelines.
  - **Azure Databricks SQL Analytics** — run query your data lake

# ETL and SQL Tools *CheatSheet*

Cheat sheets, Practice Exams and Flash cards 🖱️ [www.exampro.co/dp-900](http://www.exampro.co/dp-900)

Exam

Pro

Azure Data Factory is a managed service for **ETL, ELT and data integration**

- Create **data-driven workflows** for orchestrating **data movement** and **transforming data** at scale
- Build ELT pipelines visually without writing any code via a web-interface

SQL Server Integration Services (SSIS) — a platform for building **enterprise-level data integration and data transformations** solutions

- a low-code tool for building ELT pipelines, very similar to Azure Data Factory but existed 15 years prior.
- Integrates with Azure Data Factory

Azure Data Studio — An IDE similar Visual Studio Code, that is cross-platform and works with SQL and non-relational data, has many extensions.

SQL Server Management Studio (SSMS) — an IDE for **managing any SQL infrastructure** that only works for Windows. More mature than Data Studio