

Links to the videos regarding the DP-900:

John Savill's Technical Training : <https://www.youtube.com/watch?v=LirvmXjZU90>

freeCodeCamp.org : <https://www.youtube.com/watch?v=P3qmqUZJ7I0>

Below you will find my notes for the DP-900. This is by no means a full and complete summary for the DP-900. Its purpose was to list the topics I found most important. I encourage you to use this summary as an addition to your learning.

# DP-900

What is data?

Data can be classified into structured, semi-structured and unstructured.

Structured - tabular data represented by rows and columns. Used for relational databases.

Semi-Structured - information doesn't reside in relational database but still has some structure to it.

Documents held in JSON format,

Also key-value stores and graph databases are considered as semi-structured. It is non-relational database

Unstructured - audio, video, binary data files that don't have a specific structure. Stored in Blob.

# Data processing

Batch processing - newly arriving data elements are collected into a group. The whole group is processed at a future time as Batch.

The process can be scheduled based on scheduled time interval, or triggered when a certain amount of data has arrived, or as a result of some other event.

Stream processing - each piece of data is processed when it arrives. Streaming handles data in real-time.

# Roles in the world of data

Database admin. - manage datasets, assign permissions to users.

Data Engineers - working with data, applying cleaning routines to data, turning data into useful information

Data Analysts - explore and analyze data to create charts and visualizations to enable orgs to make informed decisions

## Tasks and Tools

Admin: Azure Data Studio - GUI for managing different database systems  
SQL Server Management Studio

Engineer - Programming

Analysts - Power BI

# Relational database on-prem vs cloud

	On-premises	Cloud
Personal control of data security	X	
Scalable		X
Hardware maintained		X
Software maintained		X
Low capital expenditure		X
Low operational expenditure	X	

## Non-Relational databases:

Use cases:

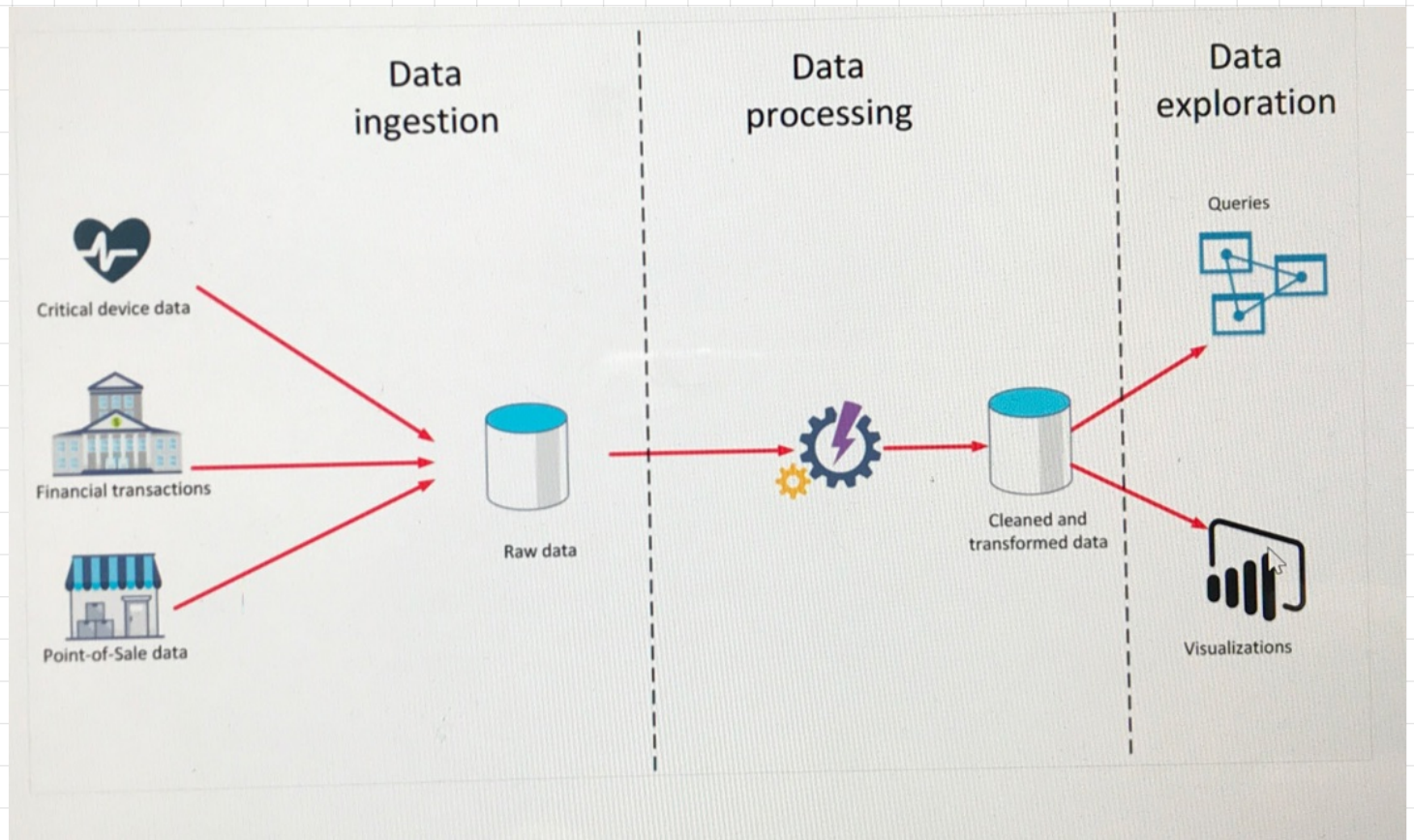
- IoT
- Gaming
- Retail and marketing
- Web and Mobile Apps.

NoSQL databases generally fall into 4 categories

- key value stores
- document databases
- column family databases
- graph databases



# Stages in data analytics.



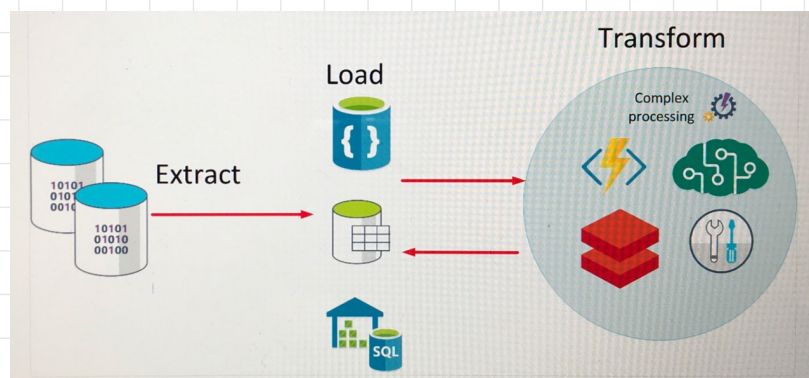
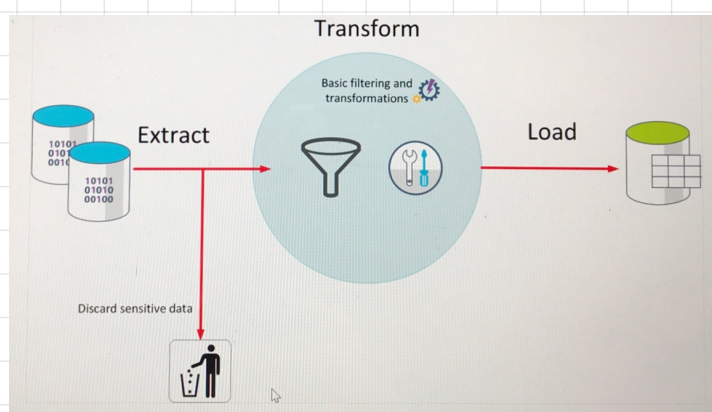
## Data processing approaches

ETL

extract  $\rightarrow$  transform  $\rightarrow$  load

ELT

extract  $\rightarrow$  load  $\rightarrow$  transform



This is a simple table showing the advantages of ETL and ELT in most cases.

	ETL	ELT
Improved data privacy and compliance	X	
Data lake support		X
Does not require specialist skills	X	
Ideal for large volumes of data		X

Data analytics :

Data analytics activity	Purpose
Descriptive analytics	Helps answer questions about what has happened, based on historical data.
Diagnostic analytics	Helps answer questions about why things happened.
Predictive analytics	Helps answer questions about what will happen in the future.
Prescriptive analytics	Helps answer questions about what actions should be taken to achieve a goal or target.
Cognitive analytics	Helps to draw inferences from existing data and patterns



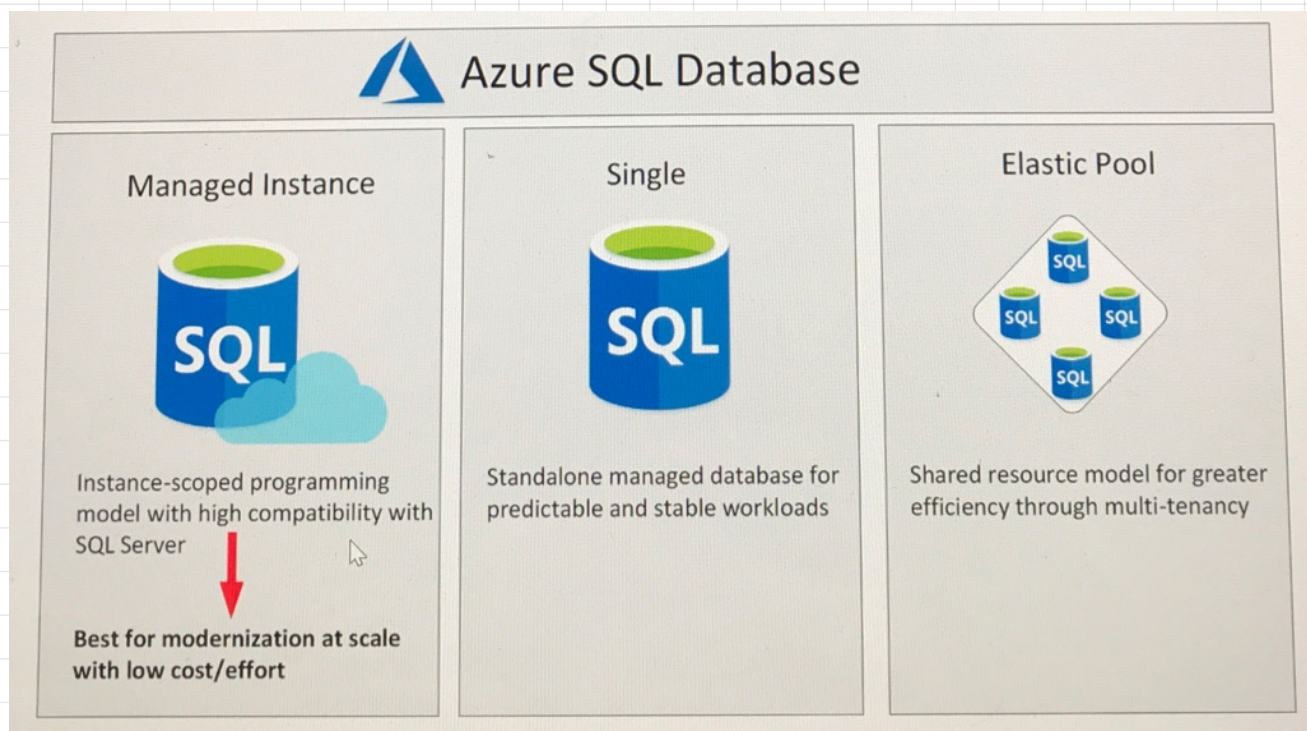
# Azure SQL Database

## Options of databases

- Single

- Elastic pool

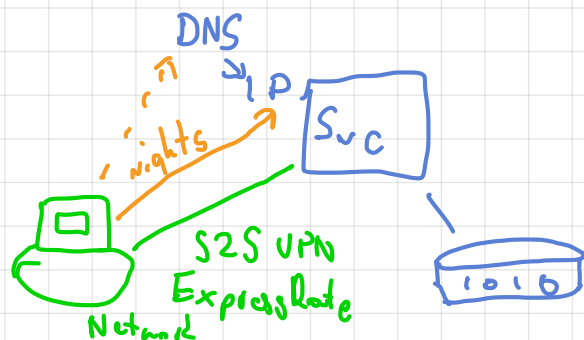
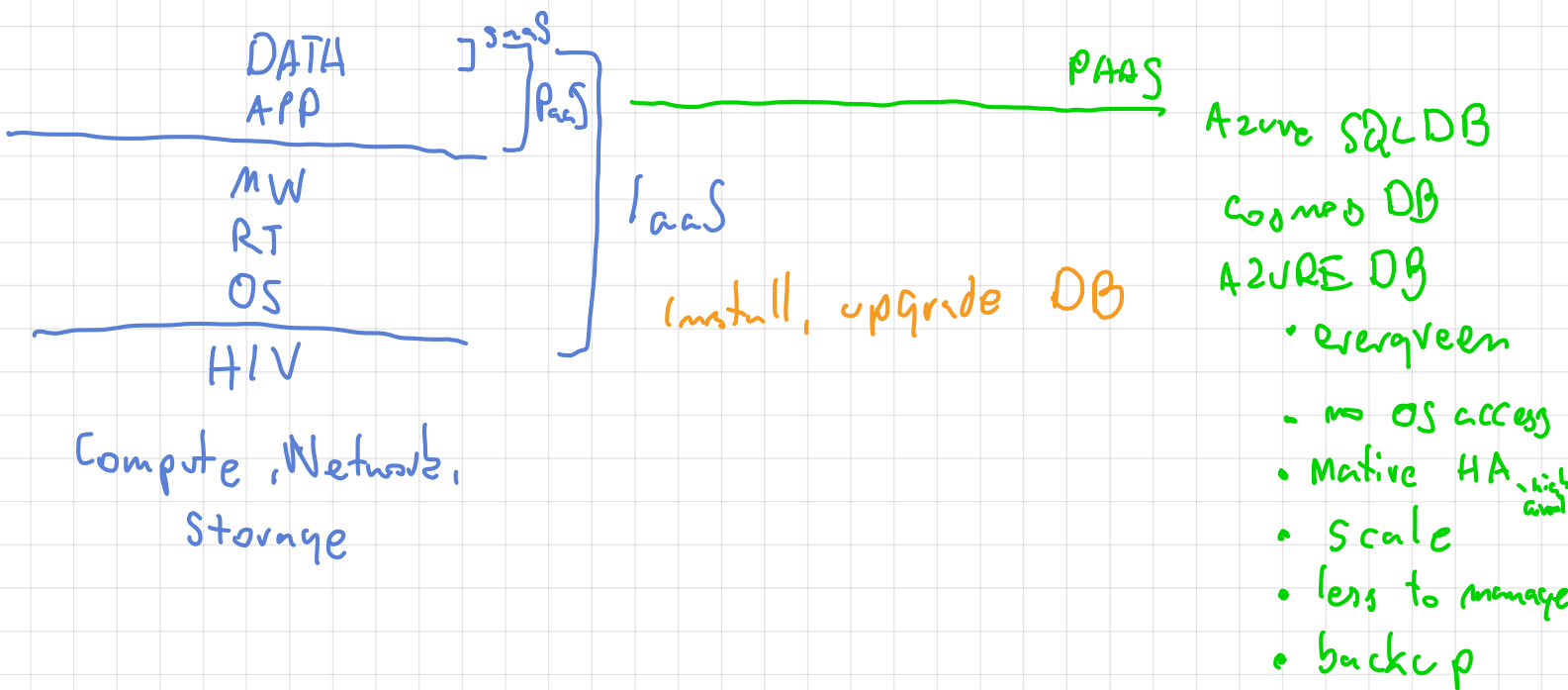
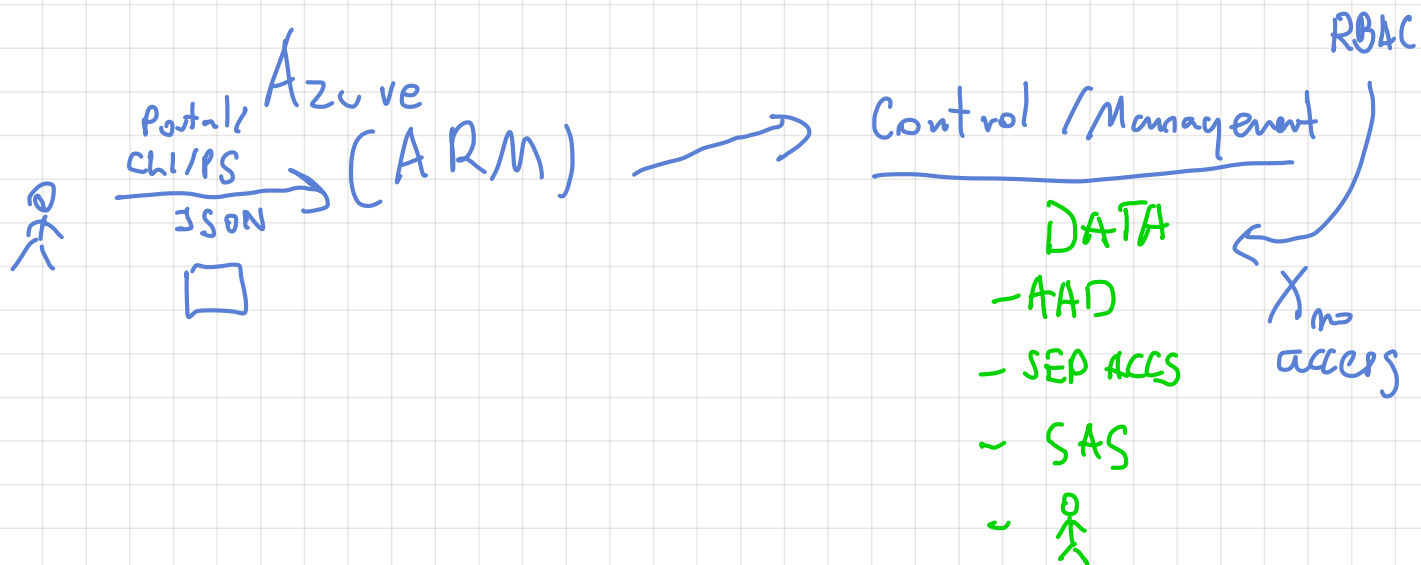
- Managed Instance



MySQL	MariaDB	PostgreSQL
<ul style="list-style-type: none"><li>• Very popular</li><li>• Available as free Community edition or paid-for, and more functional, Standard and Enterprise editions</li><li>• Azure Database for MySQL is based on the free Community edition, but adds high availability and scalability</li></ul>	<ul style="list-style-type: none"><li>• Compatible with Oracle Database</li><li>• Built-in support for temporal data</li></ul>	<ul style="list-style-type: none"><li>• Can store both relational and non-relational data</li><li>• Can store geometric data</li><li>• Extensible</li></ul>



# Summary:



## Structured Relational DB

P					
K					
E					
V					

Table

Schema  
Normalize

## Semi-Structured Loose or no schema

JSON  
XML } → Document

key  
Partition key - even distribution

Logical partitions  
↓  
PHYSICAL Partitions



## Unstructured

blob Az Data Lake Storage

- images
- video
- documents

## Graph

key-value

Time-Series



x3 copies

- Blob = block  
- append  
- page

unstructured

- Tables

Key: Value

- Queue

Messages

- Files

File share

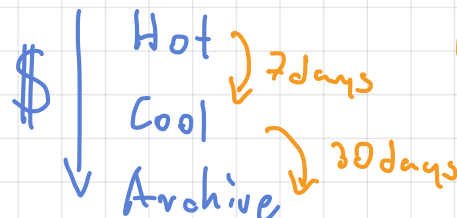
Redundancy

LRS - 3 copies in DC

ZRS - 3 copies over 3 AZ's

GRS - 3 LRS → <sup>async</sup> pair 3x

Performance



Lifecycle  
MGMT

OLTP - online transaction processing

- high volume of small transactions
- fast access
- normalize

SQL Server

Azure DB

- PostgreSQL

- MariaDB

- MySQL

OLAP - online analytical processing

- Large volume
- Data Warehousing
- Analytics

Synapse

SQL Server

PaaS - Azure SQL DB

IaaS - SQL Server → VM

Relational  
Database

T-SQL

D	ata
M	anipulation
L	anguage

Data change  
INSERT  
UPDATE  
DELETE  
SELECT


NO-SQL

Not only SQL

Cosmos DB

Multi-Region Write / AT Create  
/ AFTER

variable consistency

STRONG  EVENTUAL

Prod vs Non-prod - Portal (Azure)

API

SQL, MongoDB → Document

Cassandra → Column

Table → Key : Value

Gremlin → Graph

Request Units

Provisioned / Autoscale  $\xrightarrow{\text{Assign}}$  Database  
↓  
Containers



# Data Warehouse

Synapse

Compute / Storage  
PAUSE

Data Factory

Databricks

Tooling:

SQL Server Management Studio SSMS

- Deep MGMT, Query, HA, SQL, Store

Azure Data Studio

- Data Queries, Visualizations

Power BI

Visualization

## Data In ?

### Batch

- data collected
- processed on interval
- large volume
- latency

ETL - extract, transform, load  
old days when memory was expensive.

ELT - extract, load, transform

Transform:

- Mapping - Data Factory
- complex analysis - #OfInsight - Databricks

### Stream

- process as arrives
- real time
- scale!

Orchestration of ELT, ETL

Azure Data Factory

source -> sink

Trigger  
- scheduled  
- manual  
- Query  
-> Pipeline

CRM → ADLS<sub>gen2</sub> → Databricks → CosmosDB → Power BI

↙ ↗  
example

## Analysis

Descriptive - what happened

Diagnostic - why it happened

Predictive - what will happen (based on history)

Prescriptive - what should I do to achieve outcome

Cognitive - conclusion / knowledge

UI  
↓  
Power BI  
Visual

Data mining: extraction of patterns and knowledge from large amounts of data (not the extraction of data itself)

Data mining can be divided into 6 phases:

1. Business understanding - what does the business need
2. Data understanding - what data do we have, what do we need
3. Data preparation - how do we organize the data for modeling
4. Modeling - what modeling techniques should we apply
5. Evaluation - which data model best meets the business objective
6. Deployment - How do people access the data

Data mining methods:

clustering

classification

regression

sequential

association rules

outlier detection

prediction

Data wrangling - process of transforming and mapping data from "raw" data form to another format.

Steps :

1) Discovery

6) Publishing

2) Structuring

3) Cleaning

4) Enriching

5) Validating

# Azure Synapse Analytics -

data warehouse and unified analytics platform

Dedicated SQL pool - query service over the data in your data warehouse

Serverless SQL pool - query service over data in the data lake

Data lake : is a centralized data repository for unstructured and semi-structured data.

Polybase is data virtualization feature for SQL Server