# DP-900

Microsoft Certified: Azure Data Fundamentals

August 2020

# This course is regularly updated! This might not be the most recent version.

# Get the latest videos at http://sjd.ca/dp900

# Introduction

# Beginning to work with data in the cloud

# DP-900

# Optional certification

# Less expensive than other certifications

# Can lead to Azure DBA or Azure Data Engineer paths

# Azure Data & AI certifications

## Fundamentals

## Role-based

### Associate

### Expert


Microsoft Certified:
Azure Solutions Architect Expert

Microsoft Certified:
Azure Data Scientist Associate

Microsoft Certified:
Azure Data Engineer Associate

Microsoft Certified:
Azure AI Engineer Associate

**Key**

Optional Path

Required Path

# Microsoft Azure focus


# Microsoft Certified: Azure Data Fundamentals

**In response to the coronavirus (COVID-19) situation, Microsoft is implementing several temporary changes to our training and certification program. Learn more.**

Candidates for the Azure Data Fundamentals certification should have foundational knowledge of core data concepts and how they are implemented using Microsoft Azure data services.

This certification is intended for candidates beginning to work with data in the cloud.

Candidates should be familiar with the concepts of relational and non-relational data, and different types of data workloads such as transactional or analytical.

Azure Data Fundamentals can be used to prepare for other Azure role-based certifications like Azure Database Administrator Associate or Azure Data Engineer Associate, but it's not a prerequisite for any of them.

**Note: This certification will be available on or around June 8, 2020.**

**Job role:** Data Engineer, Database Administrator
**Required exams:** DP-900
**Important:** See details
Go to Certification Dashboard

# Core Data Concepts (15-20%)

# Relational Data (25-30%)

# Non-Relational Data (25-30%)

# Data Analytics (25-30%)


SoftwareArchitect
.ca

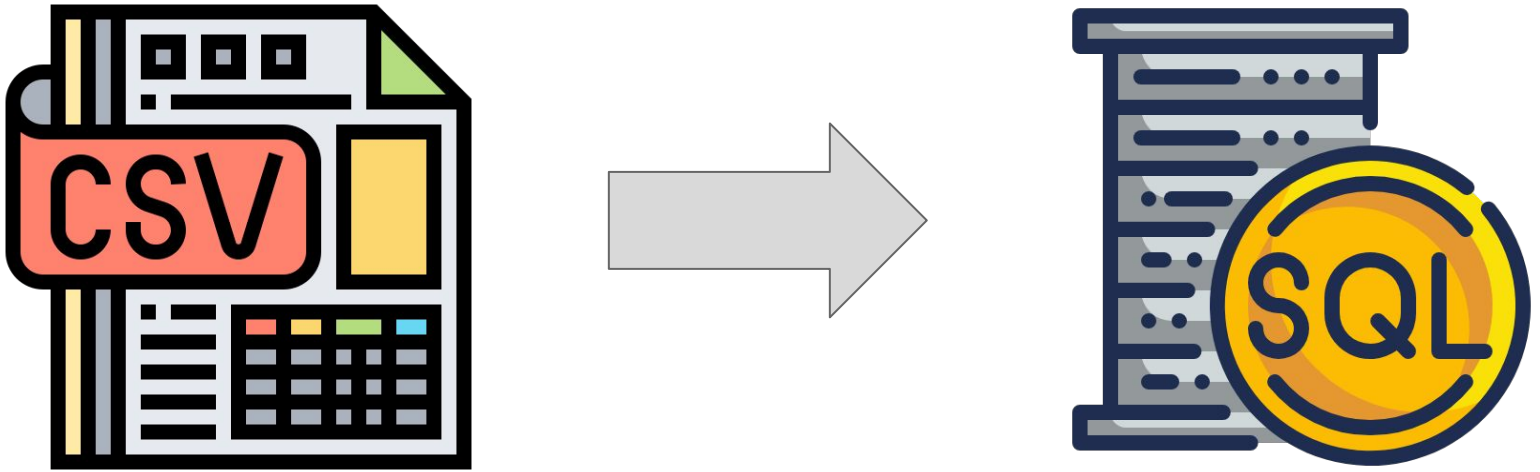
# DP-900

Microsoft Certified: Azure Data Fundamentals

# Core Data Concepts (15-20%)

# Data Workloads

# Batch Data


CSV
SQL

# Examples of Batch Formats

CSV (comma-separated) or TSV (tab-separated) file

JSON or XML

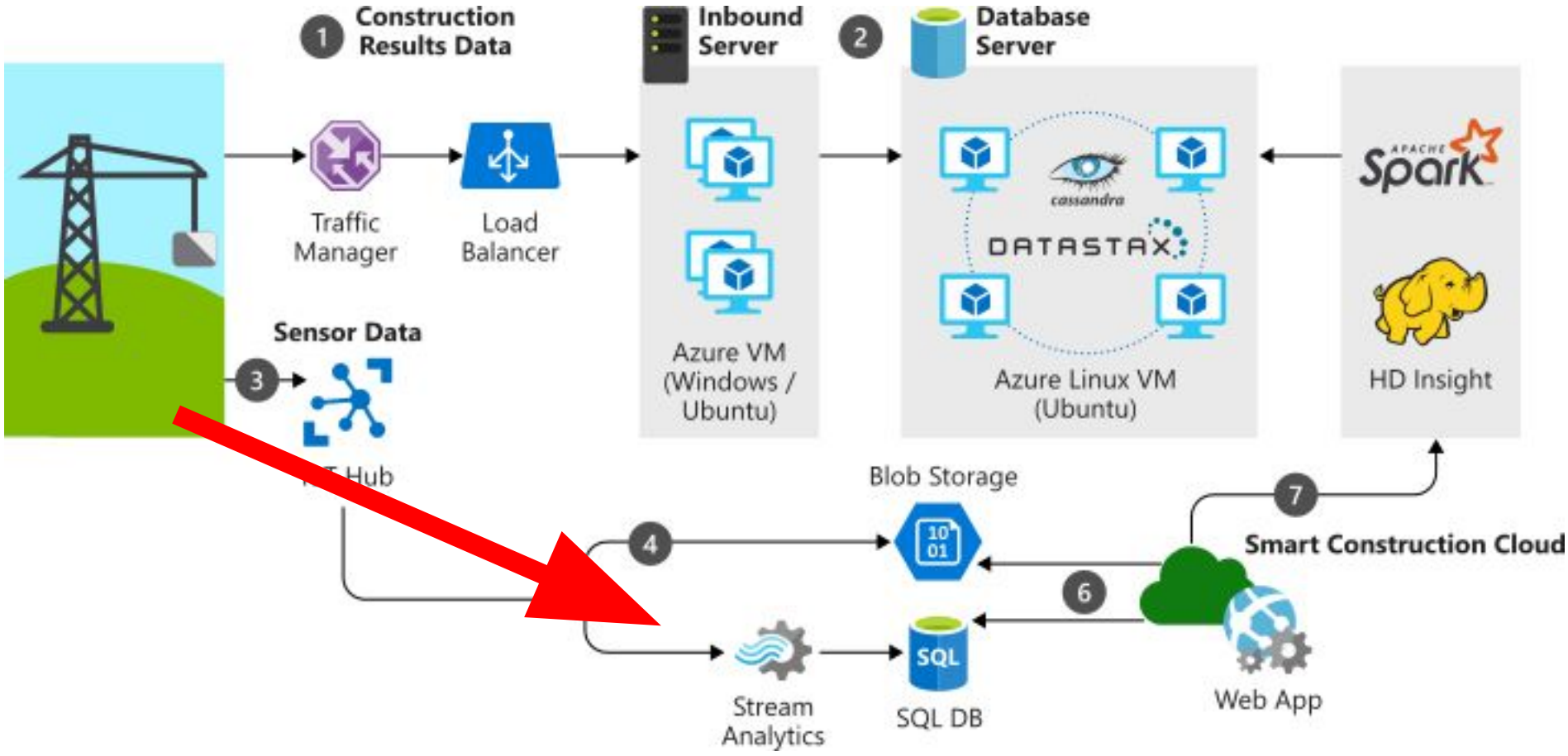
Apache Parquet - more efficient than CSV or JSON

Blob files

Another database

Cache for offline viewing

# Streaming Data


1
Construction Results Data
Inbound Server
2
Database Server
Traffic Manager
Load Balancer
Azure VM (Windows / Ubuntu)
cassandra
DATASTAX
Azure Linux VM (Ubuntu)
APACHE Spark
HD Insight
Sensor Data
3
Hub
Blob Storage
4
10 01
7
Smart Construction Cloud
6
SQL
Stream Analytics
SQL DB
Web App

# Examples of Streaming Data

Event Hub or IoT Hub

Blob storage - Log processing

Apache Kafka

Netflix or YouTube video

This course video

# Characteristics of Relational Data

# Characteristics

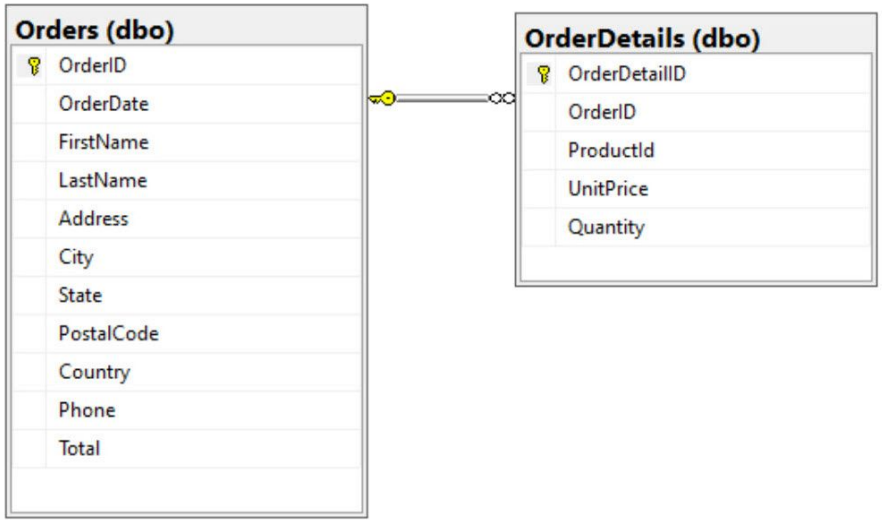
Tables - rows and columns

Views - a query that runs on a table that can be itself queried

Primary Key (PK) uniquely identifies a row

Relationships between the tables using a foreign key - parent-child

Database enforces integrity - referential integrity





```
{
    "OrderId": 1,
    "OrderDate": 1574161910220,
    "FirstName": "John",
    "LastName": "Smith",
    "Address": "10 Street",
    "City": "City",
    "State": "VA",
    "OrderDetails": [
        {
            "UnitPrice": 7.99,
            "OrderDetailId": 2,
            "Quantity": 1,
            "ProductId": 259694,
            "OrderId": 1
        },
        {
            "UnitPrice": 7.99,
            "OrderDetailId": 3,
            "Quantity": 1,
            "ProductId": 295693,
            "OrderId": 1
        }
    ],
    "id": "795c50dc-1a83-11ea-bf07-00163ee85f66",
    "_rid": "VdgtAK23OMANAAAAAAAAAA==",
    "_self": "dbs/VdgtAA==/colls/VdgtAK23OMA=/docs/VdgtAK23OMANAAAAAAAAAA==/",
    "_etag": "\"370017e1-0000-1100-0000-5df770f20000\"",
    "_attachments": "attachments/",
    "_ts": 1576497394
}
```


SoftwareArchitect
.ca

# DP-900

Microsoft Certified: Azure Data Fundamentals

# Data Analytics

# Data Visualization

# Exam tip: Microsoft is not going to ask you what chart format is best

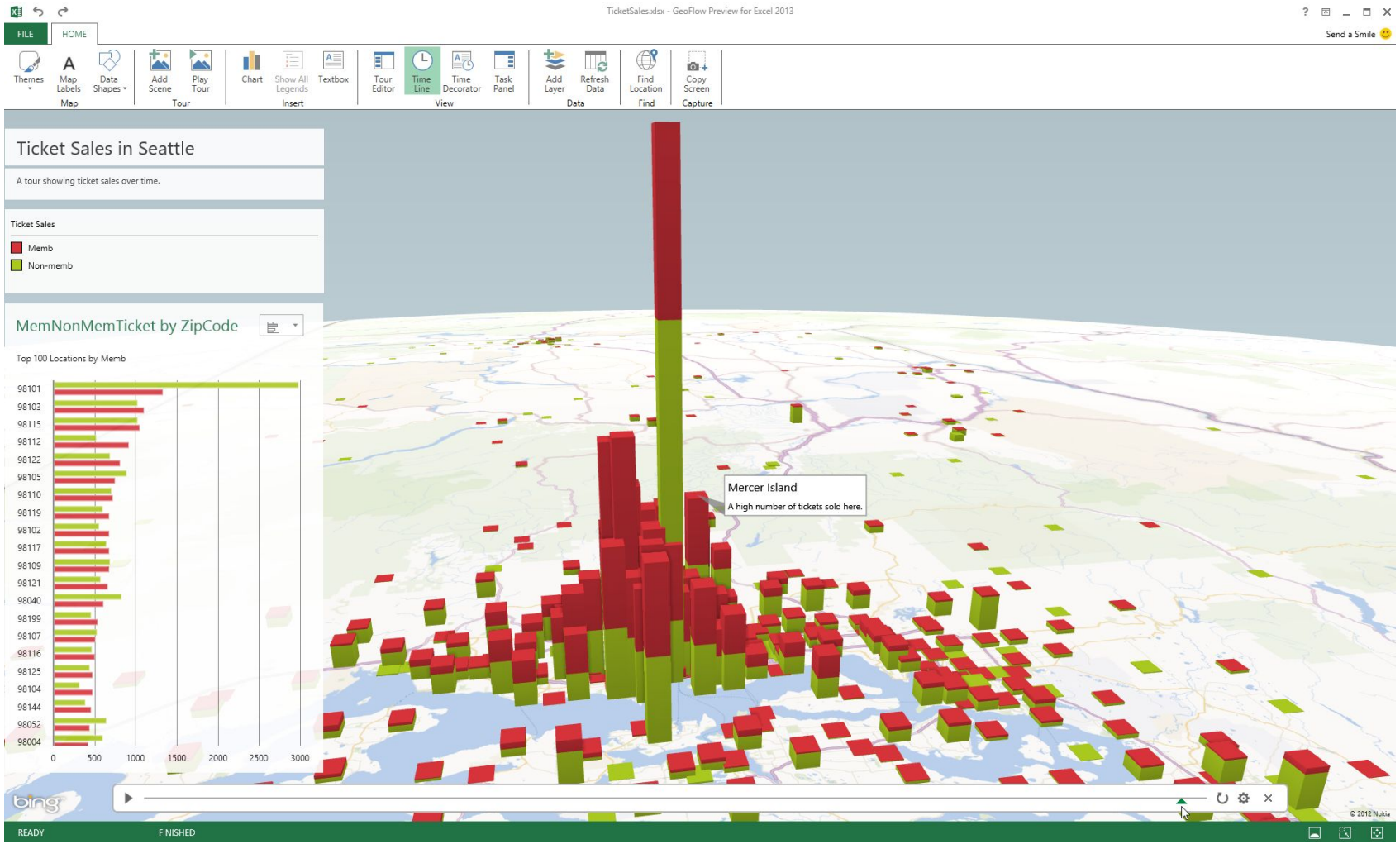
# The Problem of Too Much Data

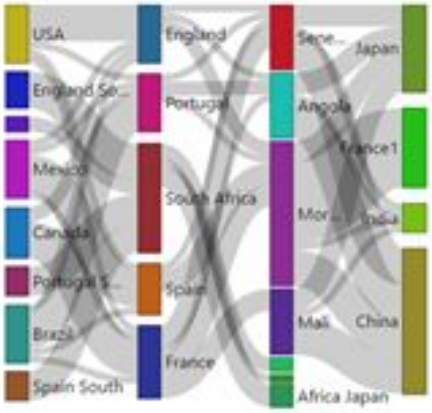
Increasingly easier and cheaper to collect data

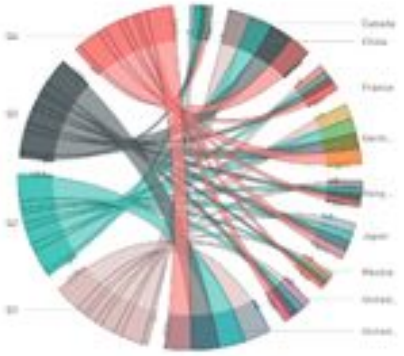
Hard to see trends and insights from raw numbers

Data visualization let's you distill thousands or millions of rows of data into an easily digestible format

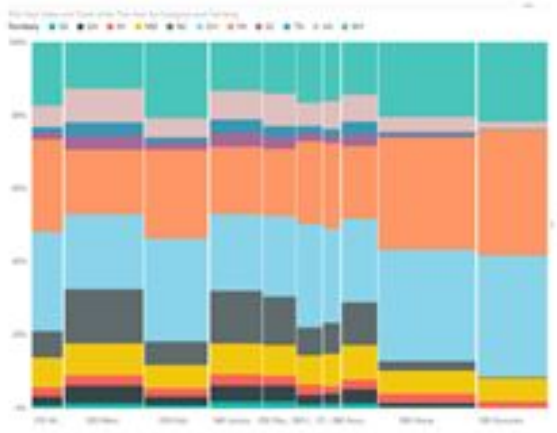
Let's you see relevant trends


TicketSales.xlsx - GeoFlow Preview for Excel 2013
FILE
HOME
Send a Smile
Themes
Map Labels
Data Shapes
Map
Add Scene
Play Tour
Tour
Chart
Show All Legends
Textbox
Insert
Tour Editor
Time Line
Time Decorator
Task Panel
View
Add Layer
Refresh Data
Data
Find Location
Find
Copy Screen
Capture
Ticket Sales in Seattle
A tour showing ticket sales over time.
Ticket Sales
Memb
Non-memb
MemNonMemTicket by ZipCode
Top 100 Locations by Memb
98101
98103
98115
98112
98122
98105
98110
98119
98102
98117
98109
98121
98040
98199
98107
98116
98125
98104
98144
98052
98004
0
500
1000
1500
2000
2500
3000
Mercer Island
A high number of tickets sold here.
bing
© 2012 Nokia
READY
FINISHED


USA
England So...
Mexico
Canada
Portugal S...
Brazil
Spain South
England
Portugal
South Africa
Spain
France
Sene...
Angola
Mor...
Mali
Africa Japan
Japan
France1
India
China



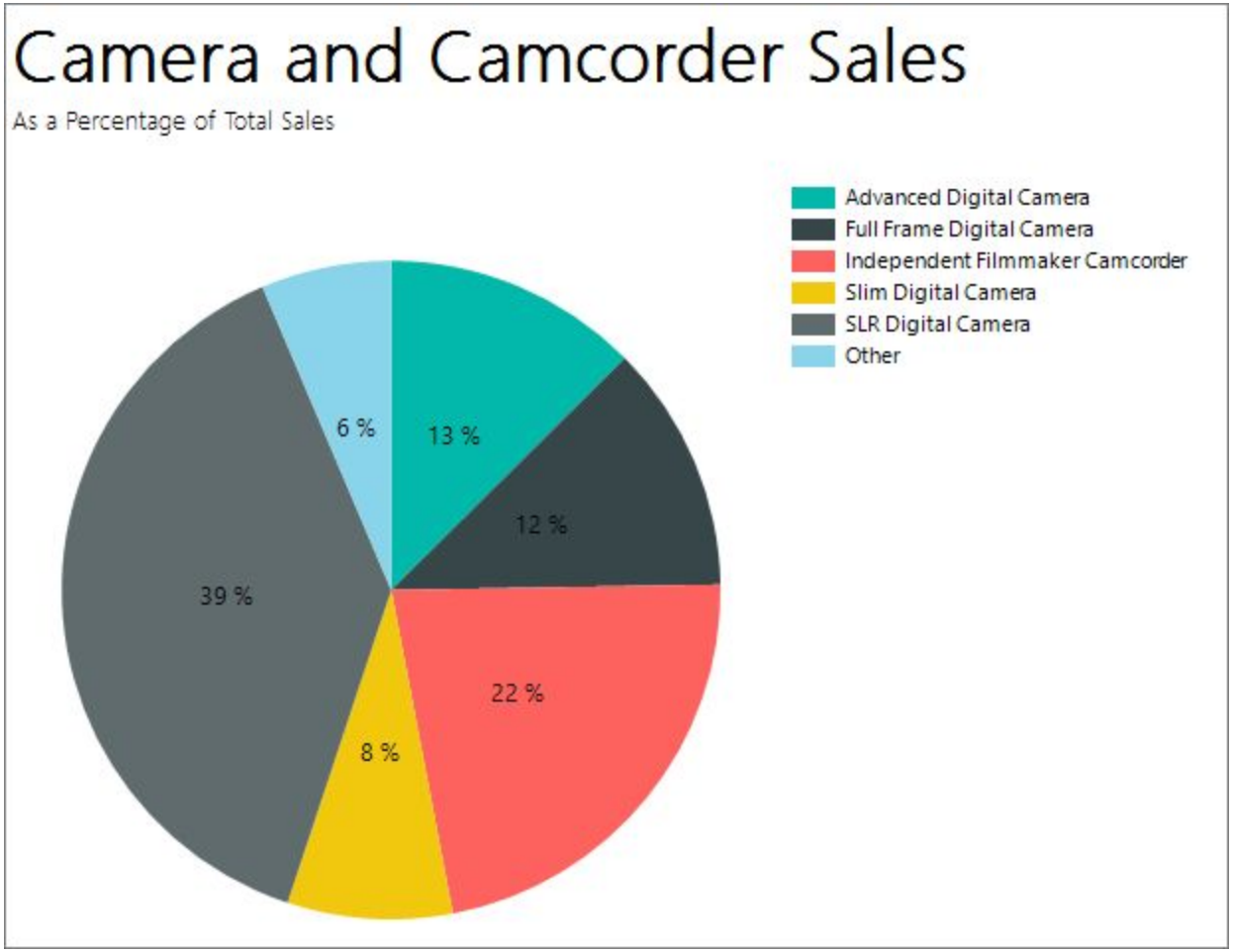

different
control
sites
social
privacy
news
famous
media
digital
friend
fun
use
tools
content
networking
information
like
people
share
easy

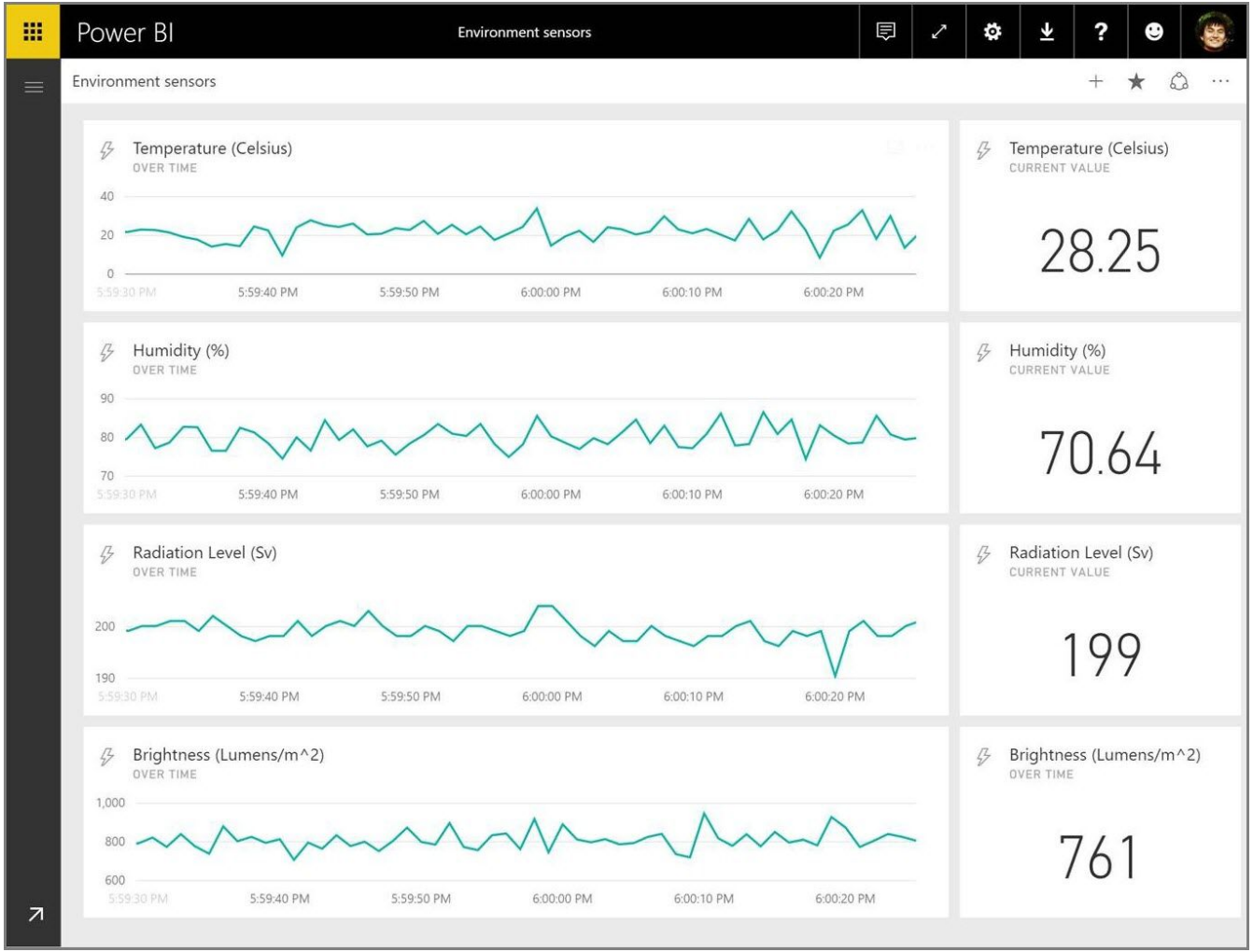


# Chart Types

# Camera and Camcorder Sales

As a Percentage of Total Sales




Power BI
Environment sensors
Environment sensors
Temperature (Celsius)
OVER TIME
40
20
0
5:59:30 PM
5:59:40 PM
5:59:50 PM
6:00:00 PM
6:00:10 PM
6:00:20 PM
Temperature (Celsius)
CURRENT VALUE
28.25
Humidity (%)
OVER TIME
90
80
70
5:59:30 PM
5:59:40 PM
5:59:50 PM
6:00:00 PM
6:00:10 PM
6:00:20 PM
Humidity (%)
CURRENT VALUE
70.64
Radiation Level (Sv)
OVER TIME
200
190
5:59:30 PM
5:59:40 PM
5:59:50 PM
6:00:00 PM
6:00:10 PM
6:00:20 PM
Radiation Level (Sv)
CURRENT VALUE
199
Brightness (Lumens/m^2)
OVER TIME
1,000
800
600
5:59:30 PM
5:59:40 PM
5:59:50 PM
6:00:00 PM
6:00:10 PM
6:00:20 PM
Brightness (Lumens/m^2)
OVER TIME
761


$1.03M
Germany
$2.42M
Canada
$1.25M
France
$1.48M
Great Britain
Sales for Top 5 Categories
Sales Amount
Units
$9M
$8M
$7M
$6M
$5M
$4M
$3M
$2M
$1M
$0M
22K
20K
18K
16K
14K
12K
10K
8K
6K
4K
2K
0K
21165
8239
5766
5498
3933
3520
1029
2000
$8.4M
$4.3M
$2.9M
$2.0M
Computers
Home Appliances
TV and Video
Cameras and camcorders
Cell phones
Audio
Music, Movies and Audio Books
Games and Toys
Net Satisfaction
1
3
Average Sale Amount by Class
$1.75M
$9.38M
$7.78M
Units by Class & Brand
Class
Deluxe
Economy
Regular
Avg. RePurch
2.2
2.4
2.6
2.8
Avg. NSAT
VISUALIZATIONS
FIELDS
Search
Page Information
Page Size
Page Background
Color
Transparency
25 %
Overview.png
Image Fit
Fit
Revert to default
Wallpaper
Color
Transparency
0 %
+ Add Image
Revert to default
PowerBI.tips

http://sql.bi/visual-reference

# THE VISUALS REFERENCE

— FOR MICROSOFT POWER BI —

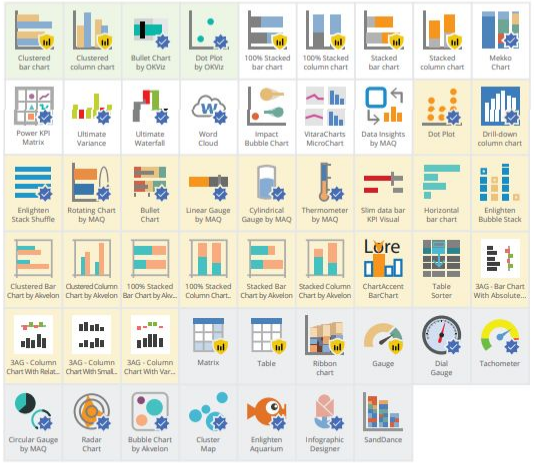
sqlbi OKViz

SEP. 2018

http://sql.bi/visual-reference

## COMPARISON

To compare the magnitude of measures

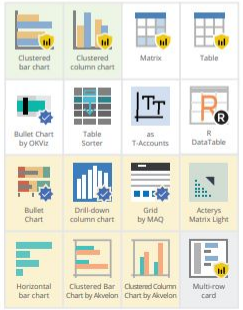



## CHANGE OVER TIME

To display the changing trend of measures


## RANKING

To rank measures in an order

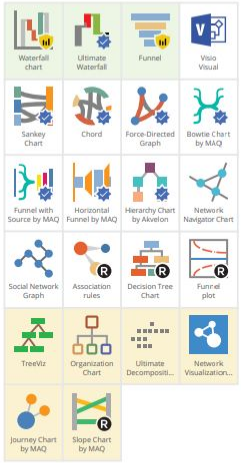



## SPATIAL

To display measures over spatial maps

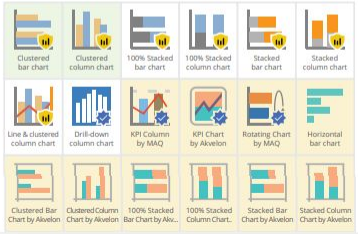

## FLOW

To display a flow or dynamic relations




## PART-TO-WHOLE

To identify the parts making up a measure total

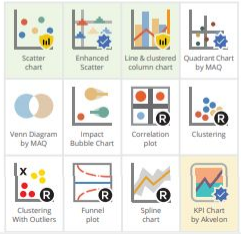



## DISTRIBUTION

To display the distribution of values

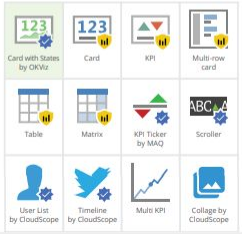

## CORRELATION

To show correlations between measures




## SINGLE

To present single values




## FILTER

To control report filters


## NARRATIVE

To tell a story with data

# Analytics Techniques

# Exam tip: Pay particular attention to this

# 5 Types of Analytics

- Descriptive
- Diagnostic
- Predictive
- Prescriptive
- Cognitive

# Descriptive Analytics

What happened?

- Today’s sales
- Service appointments booked
- Net margins

Summary of existing data records

“Hindsight”

# Diagnostic Analytics

Why did it happen?

Drill down into the data

- Change to sales per location
- Comparing one region against another over time
- Customers who had their problems solved on the first visit vs customers who had to come back for 2 or more visits

# Predictive Analytics

What will likely happen in the future?

Based on past trends

- Projected sales reports
- Azure Portal predicts your bill at the end of the month
- Weather prediction

# Prescriptive Analytics

What should I do?

Advise on best approach for maximum success

- Google Maps navigation
- Recommendations - If you liked this movie, you might like that one
- Search Engine Optimization tools

# Cognitive Analytics

Artificial intelligence and machine learning

Analyzing data to come up with a “model” of how the world works

Makes predictions based on that model

Learn and improve over time

- Reading Twitter to determine brand sentiment
- Self-driving car

# ELT and ETL

# ELT and ETL

A general name for the style in which you take data (from outside a system) and get it into a system ready for use by the system

Very rarely is data from outside 100% formatted exactly how you need it

**Extraction** is how you get the data from outside

**Loading** is the action of getting that data into a database

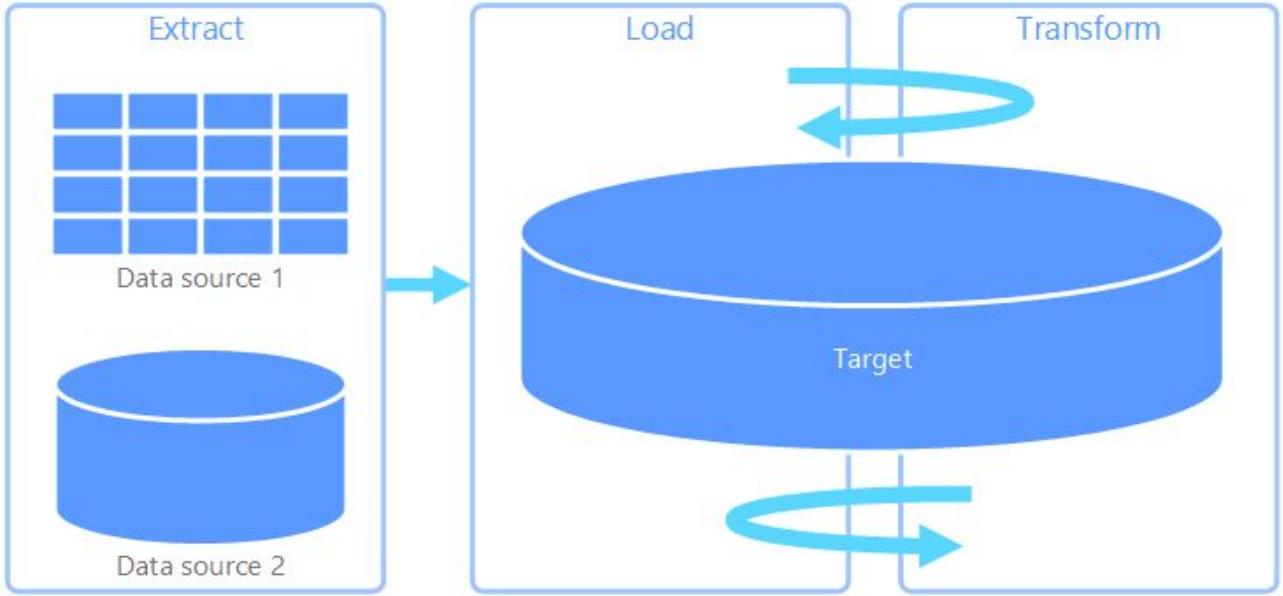
Some **Transformation** might be required to modify columns type, format, order, and even generate new data (i.e. fname + ‘ ‘ + lname = full_name)

# ELT

E - Extract

L - Load

T - Transform

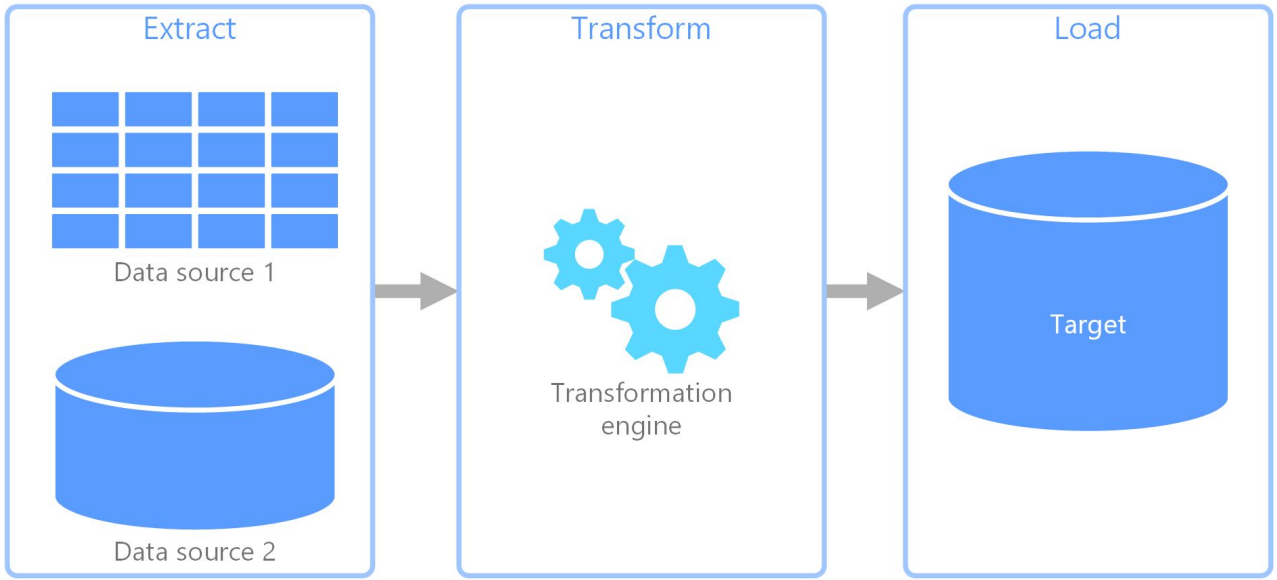


# ETL

E - Extract

T - Transform

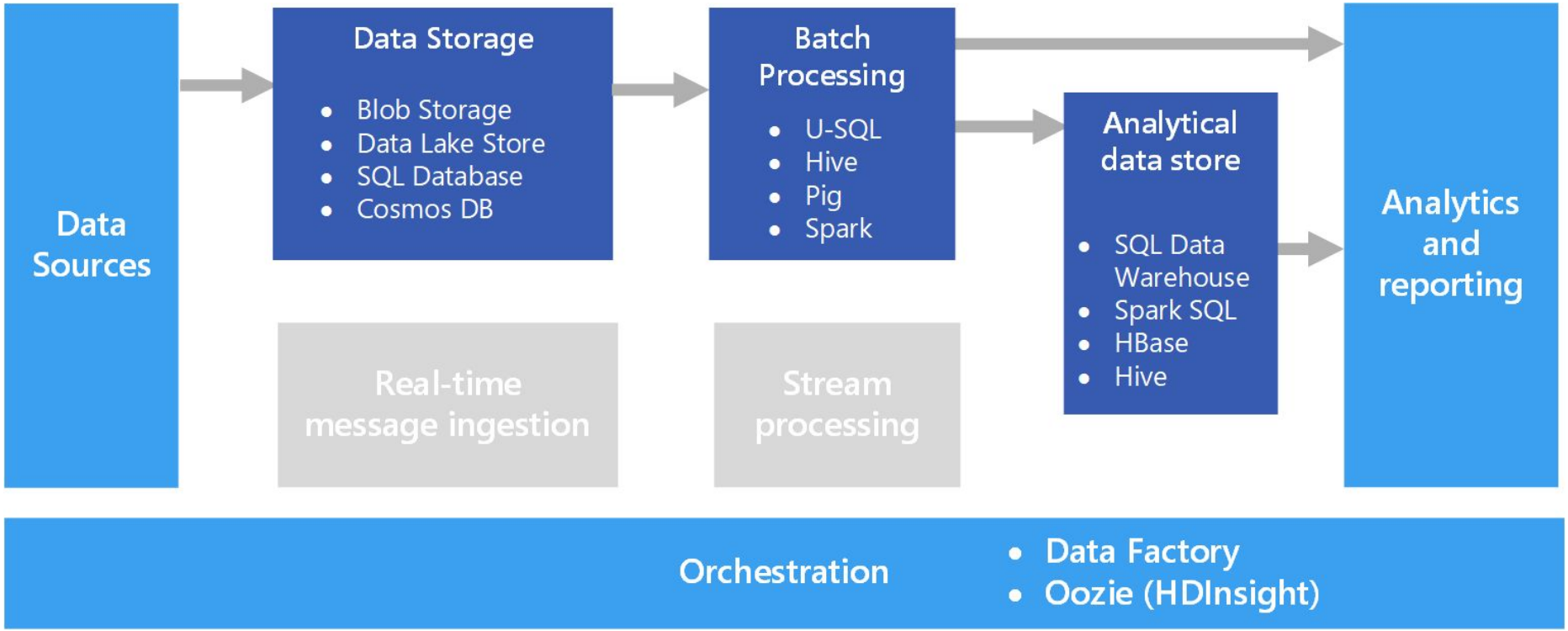
L - Load

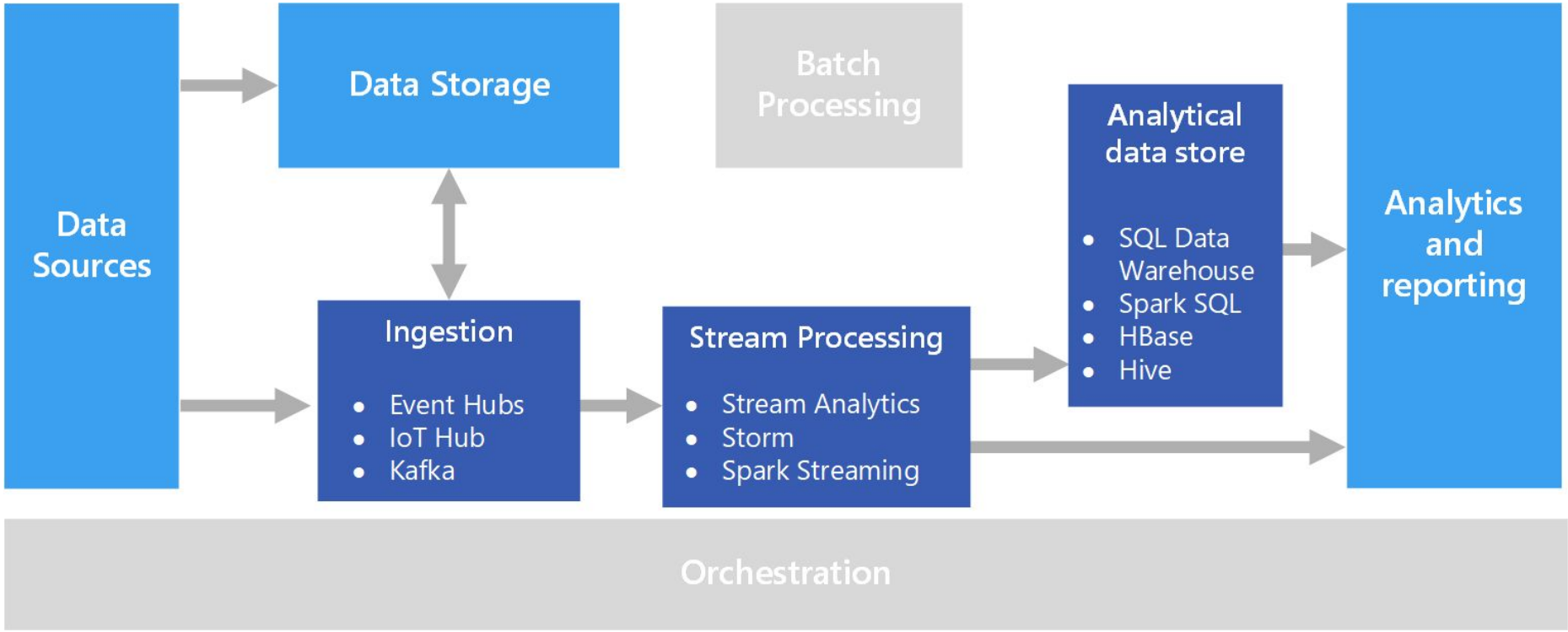


# Data Processing

# Two types...

- Batch processing
- Stream processing


Data Sources
Data Storage
Blob Storage
Data Lake Store
SQL Database
Cosmos DB
Batch Processing
U-SQL
Hive
Pig
Spark
Analytical data store
SQL Data Warehouse
Spark SQL
HBase
Hive
Analytics and reporting
Real-time message ingestion
Stream processing
Orchestration
Data Factory
Oozie (HDInsight)


Data Sources
Data Storage
Batch Processing
Analytical data store
SQL Data Warehouse
Spark SQL
HBase
Hive
Analytics and reporting
Ingestion
Event Hubs
IoT Hub
Kafka
Stream Processing
Stream Analytics
Storm
Spark Streaming
Orchestration


SoftwareArchitect
.ca

# DP-900

Microsoft Certified: Azure Data Fundamentals

# Relational Data (25-30%)

# Azure Relational DBs

# Relational DB Services

- SQL Server in a VM
- SQL Managed Instance
- Azure SQL Database
- Azure Database for MySQL, PostgreSQL, or MariaDB

# SQL Server in a VM

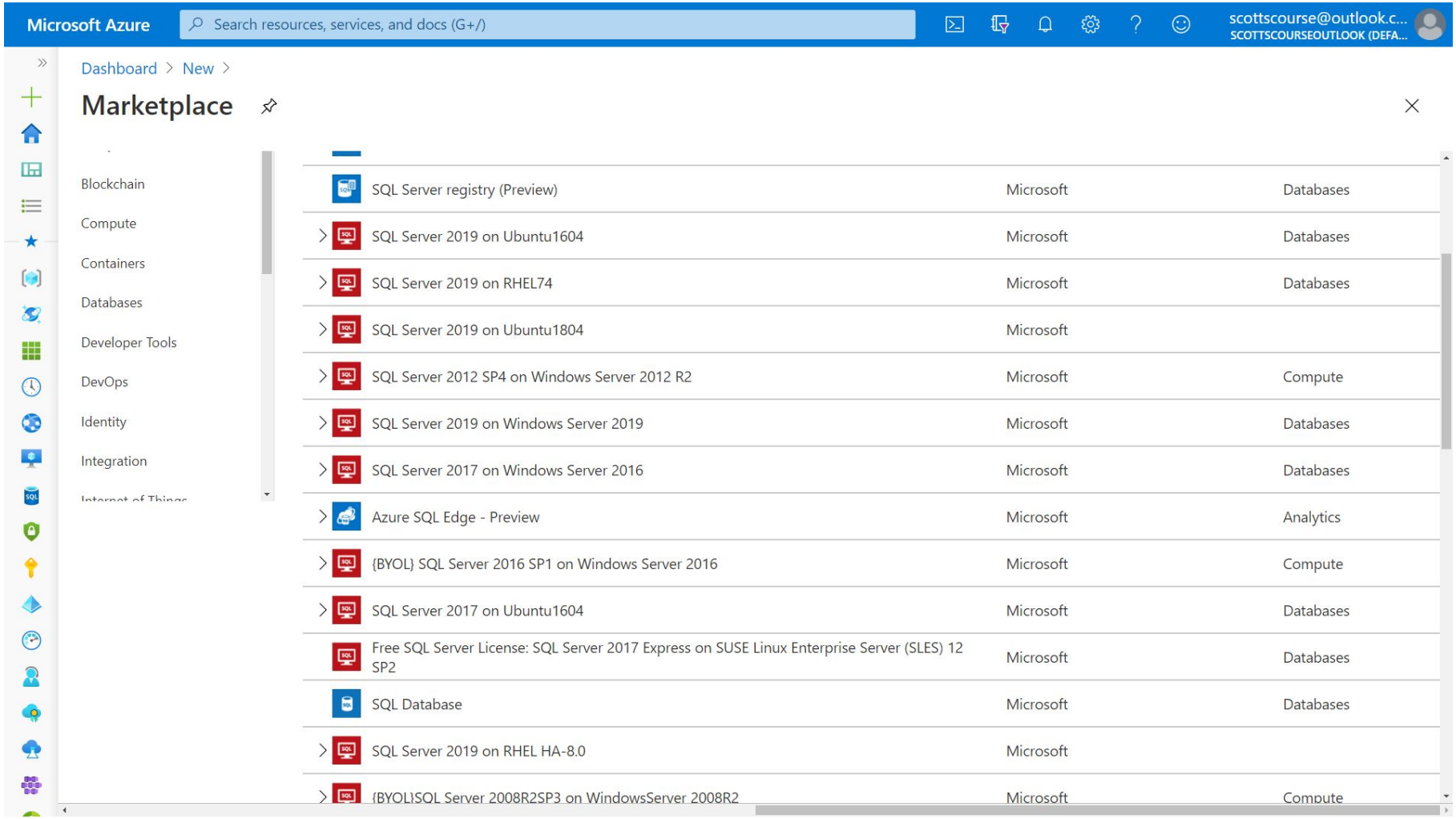
Guaranteed compatibility to SQL Server on premises

You manage everything - OS upgrades, software upgrades, backups, replication

Push the maximum performance out of the CPU through expert tweaks

No limitations (i.e. can support data above 4TB)

Pay for the server and licensing, not per DB


Microsoft Azure
Search resources, services, and docs (G+/)
scottscourse@outlook.c...
SCOTTSCOURSEOUTLOOK (DEFA...
Dashboard > New >
Marketplace
Blockchain
Compute
Containers
Databases
Developer Tools
DevOps
Identity
Integration
SQL Server registry (Preview) Microsoft Databases
SQL Server 2019 on Ubuntu1604 Microsoft Databases
SQL Server 2019 on RHEL74 Microsoft Databases
SQL Server 2019 on Ubuntu1804 Microsoft
SQL Server 2012 SP4 on Windows Server 2012 R2 Microsoft Compute
SQL Server 2019 on Windows Server 2019 Microsoft Databases
SQL Server 2017 on Windows Server 2016 Microsoft Databases
Azure SQL Edge - Preview Microsoft Analytics
{BYOL} SQL Server 2016 SP1 on Windows Server 2016 Microsoft Compute
SQL Server 2017 on Ubuntu1604 Microsoft Databases
Free SQL Server License: SQL Server 2017 Express on SUSE Linux Enterprise Server (SLES) 12 SP2 Microsoft Databases
SQL Database Microsoft Databases
SQL Server 2019 on RHEL HA-8.0 Microsoft
{BYOL}SQL Server 2008R2SP3 on WindowsServer 2008R2 Microsoft Compute

# SQL Managed Instance

Close to 100% compatibility to SQL Server on premises

Fully-managed service

4 to 80 vCores

32 GB to 8 TB storage

# Azure SQL Managed Instance

Microsoft

## Azure SQL Managed Instance

Save for later

Microsoft

Create

Overview Plans

SQL Database Managed Instance is a deployment option in Azure SQL Database that is highly compatible with SQL Server, providing out-of-the-box support for most SQL Server features and accompanying tools and services. Managed Instance also provides native virtual network (VNET) support for an isolated, highly-secure environment to run your applications. Now you can combine the rich SQL Server programming surface area in the cloud with the operational and financial benefits of an intelligent, fully-managed database service, making Managed Instance the best PaaS destination for your SQL Server workloads.

Useful Links

Documentation

Pricing details

Migrate with Azure Database Migration Service

Deploy Managed Instance inside a new virtual network

# Azure SQL Database

Close to 100% compatibility to SQL Server on premises

Lots of options for provisioned and serverless databases

Pay for performance or pay for hardware

2 to 80 vCores

5 GB and 4 TB storage

Starting at $5 per month

# Azure SQL

Microsoft


## Azure SQL

Save for later

Microsoft

Create

Overview Plans

Azure SQL allows you to create and manage your SQL Server resources from a single view, ranging from fully managed PaaS databases to IaaS virtual machines with direct OS and database engine access. All deployment options enable you to bring your on-premises licenses to Azure using Azure Hybrid Benefit.

**Databases**
Single databases are optimized for modern application development of new cloud-born applications. Databases provide a fully managed SQL experience with extensive and easy to use manageability features.
**Includes**: single databases, elastic pools, and database servers

**Managed instances**
Managed instances provide the PaaS benefits of SQL databases with added capabilities that were previously only available in SQL virtual machines. This includes a native virtual network and near 100% compatibility with on-premises SQL Server.
**Includes**: single instances, instance pools

**SQL virtual machines**
SQL virtual machines offer an IaaS architecture with extensive control over SQL Server and the underlying OS. Deployments include a management resource that focuses on SQL configuration and enables license updates with no server downtime.
**Includes**: 60+ available images combining SQL Server 2008-2019 and a variety of available OS and license types

Useful Links
Documentation
Product information

# Azure Database

Managed versions of MySQL, PostgreSQL, and MariaDB

Azure manages the servers and software upgrades

Compatible with your legacy systems

# Azure Database for MySQL

Microsoft

## Azure Database for MySQL

Save for later

Microsoft

Create

Overview Plans

Azure Database for MySQL is a MySQL database service built on Microsoft's scalable cloud infrastructure for application developers. Leverage your existing open-source MySQL skills and tools and scale on-the-fly without downtime to efficiently deliver existing and new applications with reduced operational overhead. Built-in features maximize performance, availability, and security. Azure Database for MySQL empowers developers to focus on application innovation instead of database management tasks.

Useful Links

Documentation

Landing Page

Pricing Details

# Relational Data Structures

# How is data stored?

Relational data is stored in “tables”

Tables have rows and columns

Like an Excel spreadsheet

The columns of a table are defined in advance, called a schema

# Tables

Typically, they are intended to store a single “type” of data

“Employees”, “Orders”, “Products”, “OrderDetails”

Best practice is for every table to have a “primary key”, also called an index

This is often an “ID” field when no other suitable column exists

# Example - Employees

**EmployeeID** - int (PRIMARY KEY)

**FirstName** - string

**LastName** - string

**PhoneNumber** - string (NULL)

**DepartmentID** - int (FOREIGN KEY) ← Another table called Departments

**ManagerID** - int (FOREIGN KEY) ← Points back to the Employees table (recursive)

**DOB** - string (UNIQUE KEY)

# Example - Employees

**EmployeeID** - int (PRIMARY KEY)

**FirstName** - string

**LastName** - string

**PhoneNumber** - string (NULL)

**DepartmentID** - int (FOREIGN KEY)

**ManagerID** - int (FOREIGN KEY)

**DOB - string (UNIQUE KEY) ← date of birth cannot be unique**

# Indexes

Indexes are used to improve the performance of queries

The primary key is by default an index

The data of the table is physically stored by primary key sorted order

But you can define other indexes if it’s common to query on them

It might be common to query on ManagerID, and that could be an index

# Views

A view is like a table - you can run a query on it

But the data it returns is based on another table

It can be a simplified view of a table, or a more complex one

# Manager View

```
SELECT Emp.EmployeeID, Emp.FirstName, Emp.LastName,
    Manager.EmployeeID, Manager.FirstName, Manager.LastName

FROM Employees Emp

INNER JOIN Employees Manager ON
    Manager.EmployeeID = Emp.ManagerID
```


SoftwareArchitect
.ca

# DP-900

Microsoft Certified: Azure Data Fundamentals

# Azure Relational DBs

# IaaS, PaaS and SaaS

# SQL Server in a VM

# Marketplace

- Containers
- Databases
- Developer Tools
- DevOps
- Identity
- Integration
- Internet of Things
- IT & Management Tools
- Media
- Mixed Reality
- Networking
- Security
- Software as a Service (SaaS)
- Storage
- Web

### SQL Server 2019 on Ubuntu1604
Microsoft
SQL Server 2019 images on Ubuntu 16.04

### SQL Server 2019 on Ubuntu1804
Microsoft
SQL Server 2019 images on Ubuntu 18.04

### SQL Server 2019 on Windows Server 2019
Microsoft
SQL Server 2019 images on Windows Server 2019

### SQL Server 2012 SP4 on Windows Server 2012 R2
Microsoft
SQL Server 2012 Service Pack 4

### SQL Server 2019 on RHEL74
Microsoft
SQL Server 2019 images on Red Hat Enterprise Linux 7.4

### SQL Server 2017 on Windows Server 2016
Microsoft
SQL Server 2017 images on Windows Server 2016

### SharePoint Server 2016
Microsoft
Deploy a SharePoint Server 2016 farm in Azure with the click of a button.

### {BYOL} SQL Server 2016 SP1 on Windows Server 2016
Microsoft
Database platform for intelligent, mission-critical applications (Bring Your Own License)

# SQL Server in a VM

Virtual Machine options - Windows or Linux

SQL Server options - 2008, 2012, 2014, 2017, 2019

Tiers - Free Developer, Web, Standard, Enterprise

# Advantages

Guaranteed to act identically to your on premises DB

No retraining, no retooling

Just change the connection string

# Azure SQL Database

# SQL Database

Microsoft

## SQL Database

Save for later

Microsoft

Create

Overview | Plans

SQL Database is a cloud database service built for application developers that lets you scale on-the-fly without downtime and efficiently deliver your applications. Built-in advisors quickly learn your application's unique characteristics and dynamically adapt to maximize performance, reliability, and data protection.

Use this template to create a new database in the SQL Database service. You can create the database on a new logical server or on a logical server that already exists in your subscription.

Useful Links

Documentation

Service Overview

Solutions you can deliver

Pricing Details

# Azure SQL Database

Uses SQL Server engine underneath

Designed to work in the cloud

Easy to grow to larger plans as your needs grow without downtime

Additional tools like advisors to tune your DB

# Types

Single Database - allocate resources to a specific database

Elastic Database - allocate resources to a group of databases to share (pool)

# Advantages

Mostly compatible with SQL Server

You can scale easily

Modern, cloud approach

# Azure Synapse Analytics (SQL DW)

Azure Synapse Analytics (formerly SQL DW)
Microsoft

# Azure Synapse Analytics (formerly SQL DW)

Save for later

Microsoft

Create

Overview | Plans

Azure Synapse Analytics is a limitless analytics service that brings together enterprise data warehousing and Big Data analytics.

It gives you the freedom to query data on your terms, using either serverless on-demand or provisioned resources-at scale. Azure Synapse brings these two worlds together with a unified experience to ingest, prepare, manage, and serve data for immediate BI and machine learning needs.

Simply put, Azure Synapse is Azure SQL Data Warehouse evolved. We have taken the same industry leading data warehouse to a whole new level of performance and capabilities. Businesses can continue running their existing data warehouse workloads in production today with Azure Synapse and will automatically benefit from the new capabilities which are in preview.

Useful Links
Documentation
Service Overview
Pricing Details

# Synapse Analytics

Used to be called SQL Data Warehouse (SQL DW)

Evolution past DW

Serverless on-demand or provisioned

A database designed for reporting and analytics

# Azure Database for PostgreSQL, MySQL, MariaDB

# Other Azure Database Options

Azure is adding other database engines, as a service

PostgreSQL, MySQL and MariaDB

Instead of you hosting those and maintaining them

Allows you to maintain your existing solution and migrate to the cloud

# SQL Managed Instance

# SQL Managed Instance

Aims to bridge the gap between Azure SQL Database and SQL Server in a VM

Let Microsoft manage the server and the database engine software

Compatible with SQL Server

Not a common option, in my view


SoftwareArchitect
.ca

# DP-900

Microsoft Certified: Azure Data Fundamentals

# Managing Relational Data

# Provisioning and Deployment

# Deployment Options

Create through the Azure Portal

# ARM Templates

# Data Security

# Connectivity

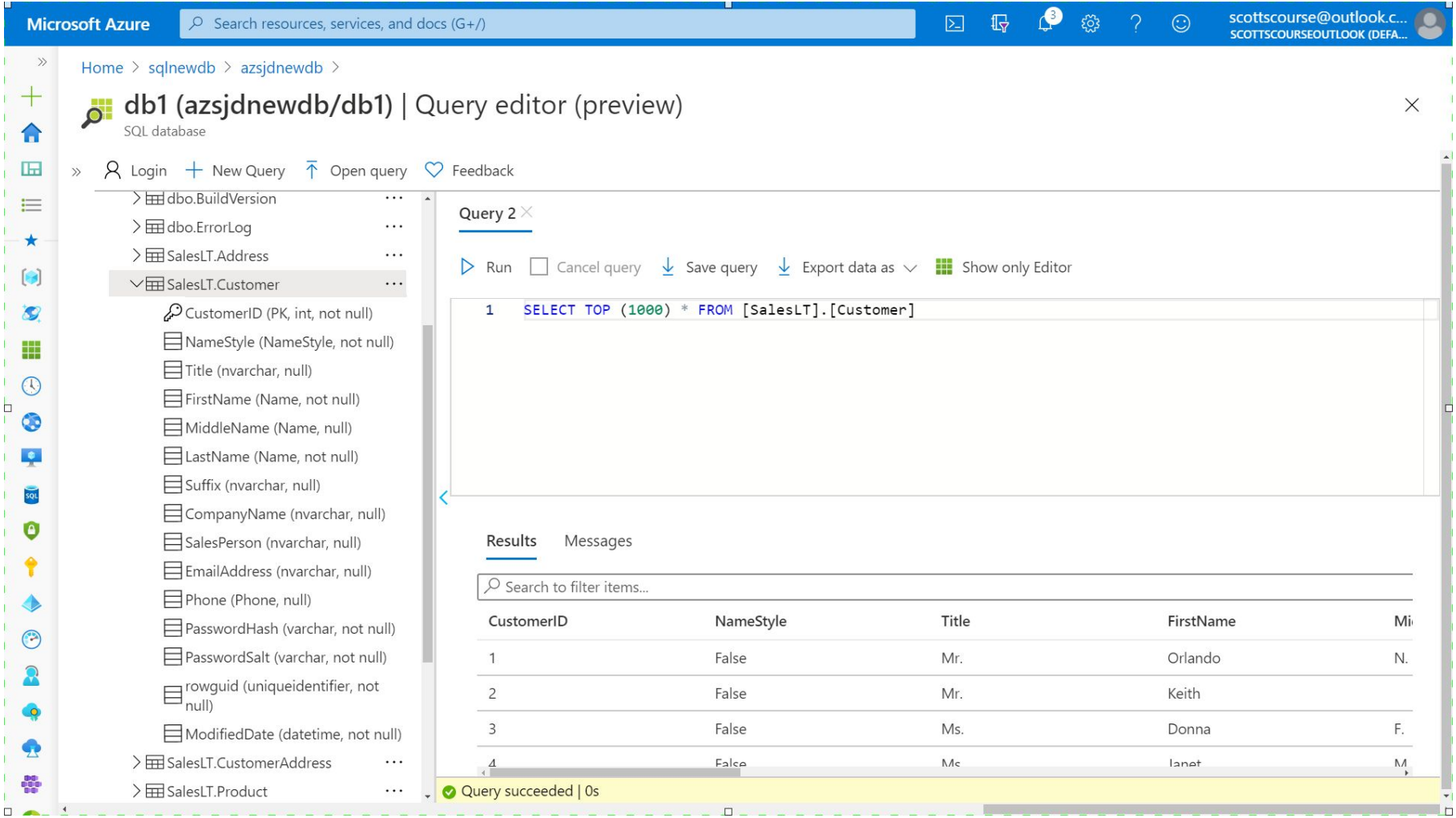
# Query Tools

# Query Editor (Portal)

Query editor built into the portal

Access from the SQL Database database module

Execute queries


Microsoft Azure
Search resources, services, and docs (G+/)
scottscourse@outlook.c...
SCOTTSCOURSEOUTLOOK (DEFA...
Home > sqlnewdb > azsjdnewdb >
db1 (azsjdnewdb/db1) | Query editor (preview)
SQL database
Login
New Query
Open query
Feedback
dbo.BuildVersion
dbo.ErrorLog
SalesLT.Address
SalesLT.Customer
CustomerID (PK, int, not null)
NameStyle (NameStyle, not null)
Title (nvarchar, null)
FirstName (Name, not null)
MiddleName (Name, null)
LastName (Name, not null)
Suffix (nvarchar, null)
CompanyName (nvarchar, null)
SalesPerson (nvarchar, null)
EmailAddress (nvarchar, null)
Phone (Phone, null)
PasswordHash (varchar, not null)
PasswordSalt (varchar, not null)
rowguid (uniqueidentifier, not null)
ModifiedDate (datetime, not null)
SalesLT.CustomerAddress
SalesLT.Product
Query 2
Run
Cancel query
Save query
Export data as
Show only Editor
1 SELECT TOP (1000) * FROM [SalesLT].[Customer]
Results
Messages
Search to filter items...
CustomerID
NameStyle
Title
FirstName
1
False
Mr.
Orlando
2
False
Mr.
Keith
3
False
Ms.
Donna
Query succeeded | 0s

# Azure Data Studio

Cross-platform software to work with SQL Server relational data

- Windows, Mac OS, and Linux

Intellisense, code snippets, source control, integrated terminal

Charting of query results

Open source, free

# SQL Server Management Studio (SSMS)

The “OG” of SQL Server management

Supports database administration tasks

- User management, vulnerability assessment, security features
- Performance tuning advisors
- Import and export of DACPAC files

# sqlcmd

Command line utility

Execute T-SQL statements, stored procs, and script files

Windows and Linux options (RHEL, Ubuntu, SUSE)

MacOS in preview

Can run in a docker container


SoftwareArchitect
.ca

# DP-900

Microsoft Certified: Azure Data Fundamentals

# SQL

# DDL vs DML

# Data Definition Language (DDL)

The data management SQL commands around database structures

CREATE, ALTER, DROP, TRUNCATE, COMMENT, RENAME

# Data Manipulation Language (DML)

The SQL commands around data manipulation

SELECT, INSERT, UPDATE, DELETE, LOCK TABLE

# Data Control Language (DCL)

Granting rights and permissions to others

GRANT and REVOKE

# Querying Other Relational DBs

# Structured Query Language (SQL)

First developed as SEQUEL in the early 1970s

Had to change to SQL due to an existing trademark

SQL Standard first formalized in 1986 as SQL-86, now up to SQL:2019

# Vendors!

Each database engine has its own version of SQL that extends the standard for its own purpose, usually to add programming elements (stored procs!)

Most vendor version of SQL are not 100% compatible with the standard

SQL Server uses T-SQL, Oracle uses PL/SQL, MySQL uses SQL/PSM, PostgreSQL uses PL/pgSQL

They are not often compatible - an Oracle PL/SQL command would not work unaltered on SQL Server


SoftwareArchitect
.ca

# DP-900

Microsoft Certified: Azure Data Fundamentals

# Non-Relational Data (25-30%)

# Non-Relational Data

# What is a Non-Relational Database?

Technically: any database that is not based on tables, rows and columns

Often, the way the data is stored better supports the application's use

# Not only SQL (NoSQL)

# The internet has changed the way we package data to send around

Facebook - 1994
Google - 1998
Twitter - 2006

HTML - 1991
XML - 1998
JSON - 2005+

# There are 200 billion tweets per year

# Imagine trying to run a search query on a table with 1 trillion rows...

# Non-relational databases are optimized for different uses


SoftwareArchitect
.ca

# DP-900

Microsoft Certified: Azure Data Fundamentals

# Data types

# Document Data - Cosmos DB Core SQL

| Key | Document |
|---|---|
| 1001 | { "CustomerID": 99, "OrderItems": [ { "ProductID": 2010, "Quantity": 2, "Cost": 520 }, { "ProductID": 4365, "Quantity": 1, "Cost": 18 }], "OrderDate": "04/01/2017" } |
| 1002 | { "CustomerID": 220, "OrderItems": [ { "ProductID": 1285, "Quantity": 1, "Cost": 120 }], "OrderDate": "05/08/2017" } |

*** Used to be named DocumentDB

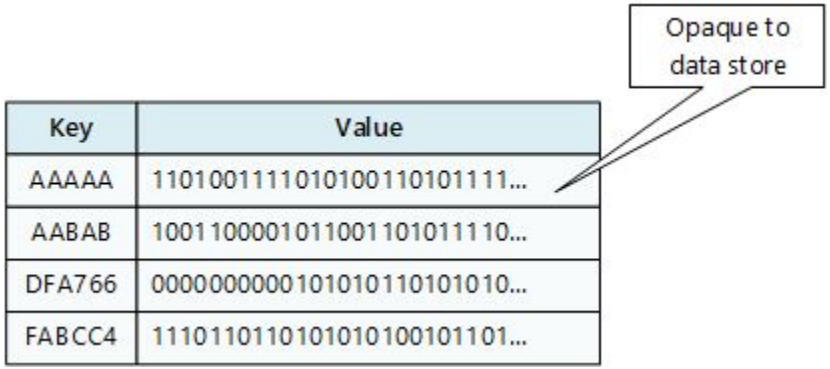
# Column-Family Data - Cosmos DB Cassandra API

| CustomerID | Column Family: Identity |
|---|---|
| 001 | **First name:** Mu Bae<br>**Last name:** Min |
| 002 | **First name:** Francisco<br>**Last name:** Vila Nova<br>**Suffix:** Jr. |
| 003 | **First name:** Lena<br>**Last name:** Adamcyz<br>**Title:** Dr. |

| CustomerID | Column Family: Contact Info |
|---|---|
| 001 | **Phone number:** 555-0100<br>**Email:** someone@example.com |
| 002 | **Email:** vilanova@contoso.com |
| 003 | **Phone number:** 555-0120 |

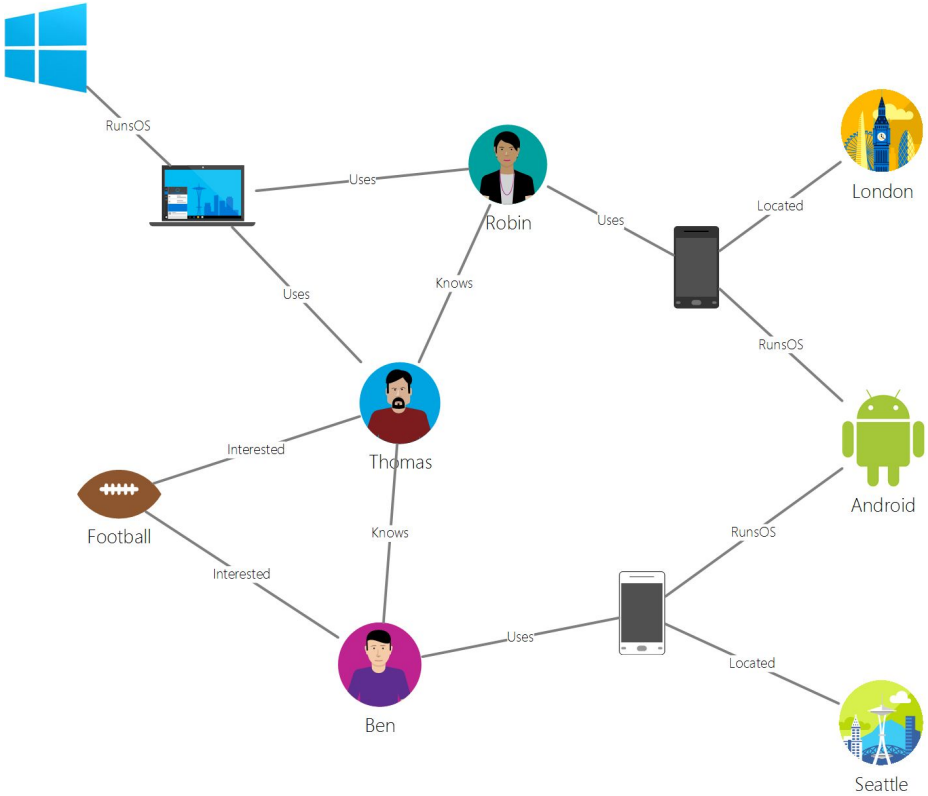
*** Also HBase in HDInsight

# Key-Value Data - Cosmos DB Table API




*** Also Table Storage, Redis Cache

# Graph Data - Cosmos DB Graph API



# Time Series Data - Time Series Insights

| timestamp | deviceid | value |
| --- | --- | --- |
| 2017-01-05T08:00:00.123 | 1 | 90.0 |
| 2017-01-05T08:00:01.225 | 2 | 75.0 |
| 2017-01-05T08:01:01.525 | 2 | 78.0 |

*** Also OpenTSDB with HBase on HDInsight

# Object Data - Blob Storage

| path | blob | metadata |
| --- | --- | --- |
| /delays/2017/06/01/flights.cscv | 0XAABBCCDDEEF... | {created: 2017-06-02} |
| /delays/2017/06/02/flights.cscv | 0XAADDCCDDEEF... | {created: 2017-06-03} |
| /delays/2017/06/03/flights.cscv | 0XAEBBDEDDEEF... | {created: 2017-06-03} |

*** Also Data Lake Storage, File Storage

# Azure Search

| id | search-document |
| --- | --- |
| 233358 | {"name": "Pacific Crest National Scenic Trail", "county": "San Diego", "elevation":1294, "location": {"type": "Point", "coordinates": [–120.802102,49.00021]}} |
| 801970 | {"name": "Lewis and Clark National Historic Trail", "county": "Richland", "elevation":584, "location": {"type": "Point", "coordinates": [–104.8546903,48.1264084]}} |
| 1144102 | {"name": "Intake Trail", "county": "Umatilla", "elevation":1076, "location": {"type": "Point", "coordinates": [–118.0468873,45.9981939]}} |


SoftwareArchitect
.ca

# DP-900

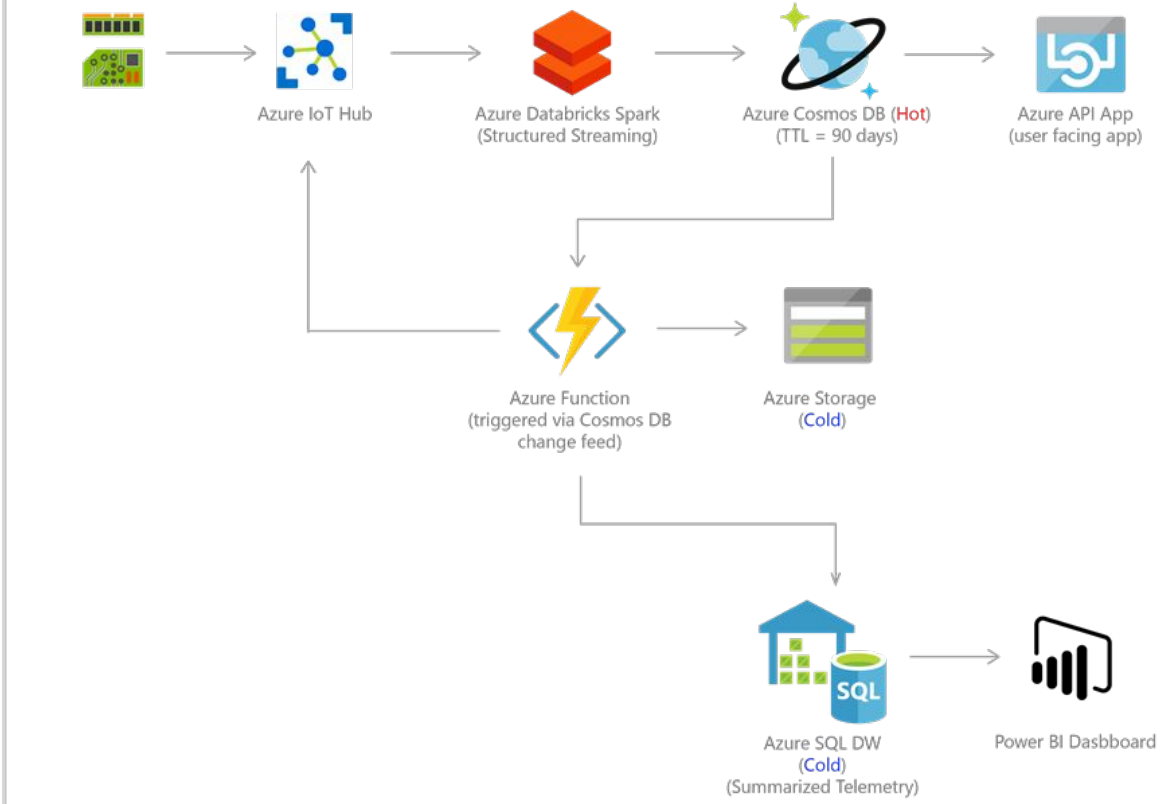
Microsoft Certified: Azure Data Fundamentals

# How to choose?

# One Ring to Rule Them All...

There is no single data store that is best for all use cases

Sometimes you use more than one type


Azure IoT Hub
Azure Databricks Spark
(Structured Streaming)
Azure Cosmos DB (Hot)
(TTL = 90 days)
Azure API App
(user facing app)
Azure Function
(triggered via Cosmos DB
change feed)
Azure Storage
(Cold)
SQL
Azure SQL DW
(Cold)
(Summarized Telemetry)
Power BI Dasbboard

# Benefits of Relational Databases

- Normalization reduces duplication, reduces mistakes
- Data relationships drive data insights
- Data integrity
- Many of the established products are decades old, bug free
- Business intelligence tools like Power BI makes the data accessible
- Database enforces constraints, fixed schema

# Benefits of NoSQL Databases

- Optimized for specific data types
- Can be designed for performance at internet scales
- Might not require massive hardware, reduced expense
- Open source


SoftwareArchitect
.ca

# DP-900

Microsoft Certified: Azure Data Fundamentals

# Azure Non-Relational DBs

# Azure Non-Relational Options

# Cosmos DB

- Enterprise-grade non-relational database
- Supports many data models (graph, document, table, column-family)
- Compatible with established APIs (Cassandra, MongoDB, Gremlin, Etcd)
- Select the data consistency you need
- Easy to scale worldwide
- Sub-10 ms latency
- Service Level Agreements for throughput, latency, availability and consistency

# Table Storage

- Azure Storage account
- 5 PB maximum limit
- $0.045 per GB - cheapest data storage option on Azure
- Plus pay for operations - $0.00036 per 10,000 transactions
- Service Level Agreement shockingly poor (10 seconds for a query)

# Blob Storage

- Azure Storage account
- 5 PB maximum limit
- $0.0208 per GB - cheapest data storage option on Azure
- Supports premium, hot, cool, archive access tiers
- Supports “reserved capacity” - ~$0.014 per GB for 100TB/3 years
- Supports blob indexing
- Plus pay for operations - $0.00036 per 10,000 transactions
- Service Level Agreement shockingly poor (2 seconds per MB)

# File Storage

- Azure Storage account
- 5 PB maximum limit
- $0.06 per GB - cheapest data storage option on Azure
- Standard and premium options
- Service Level Agreement shockingly poor (2 seconds per MB)


SoftwareArchitect
.ca

# DP-900

Microsoft Certified: Azure Data Fundamentals

# Management

# Provisioning and Deployment

# ARM Templates

# Data Security

# Connectivity

# Management Tools


SoftwareArchitect
.ca

# DP-900

Microsoft Certified: Azure Data Fundamentals

# Data Analytics (25-30%)

# Analytics Workloads

# Types of Workloads

There are three main (relational) database workloads:

- As a place to store business transactions as they occur (OLTP)
- As a place to hold data for complex analysis (OLAP)
- As a centralized repository for data from different sources (data warehouse)

# OLTP - Online Transaction Processing

- Most business applications require a place to store and retrieve data
- As business transactions occur, they are recorded to the database
- Existing rows can be updated
- Data can be retrieved by SQL queries
- Optimized for general use

# Traits of OLTP

- Database normalization
- Schema heavily enforced, data integrity
- Strong consistency
- Heavy writes, moderate reads
- Updateable
- Data size MBs to TBs

# Azure OLTP

- Azure SQL Database
- SQL Server in a VM
- Azure Database for MySQL
- Azure Database for PostgreSQL

# OLAP - Online Analytical Processing

- Data stored in a transactional database was not designed for complex analysis
- Data stored in a transactional database can change at any time
- Running complex reports can slow down a transactional database
- Can take time to prepare the data for analysis
- Cubes, dimensions, measures

# Traits of OLAP

- No locking
- No updates
- Heavy reads, read-only
- Multi-dimensional indexing
- Data size GBs

# Azure OLAP

- SQL Server with Columnstore indexes
- Azure Analysis Services
- SQL Server Analysis Services

# Data Warehousing

- Central repository of data from one or more different sources
- Current and historical data used for reporting and analysis
- Can rename or reformat columns to make it easier for users to create reports
- Users can run reports without affecting the day-to-day business data systems

# Azure Data Warehousing

Symmetric Multiprocessing (SMP):

- Azure SQL Database
- SQL Server in a VM

Massively Parallel Processing (MPP):

- Azure Synapse Analytics (SQL DW)
- Apache Hive on HDInsight
- Interactive Query (Hive LLAP) on HDInsight

# When to Use a DW

When queries are long running or affect day-to-day operations

When data needs further processing (ETL or ELT) before it can be analyzed

When you want to remove historical data from your day-to-day systems (archiving)

When you need to integrate data from several sources

When users are confused by the data structures, table names or column names when building reports in PowerBI

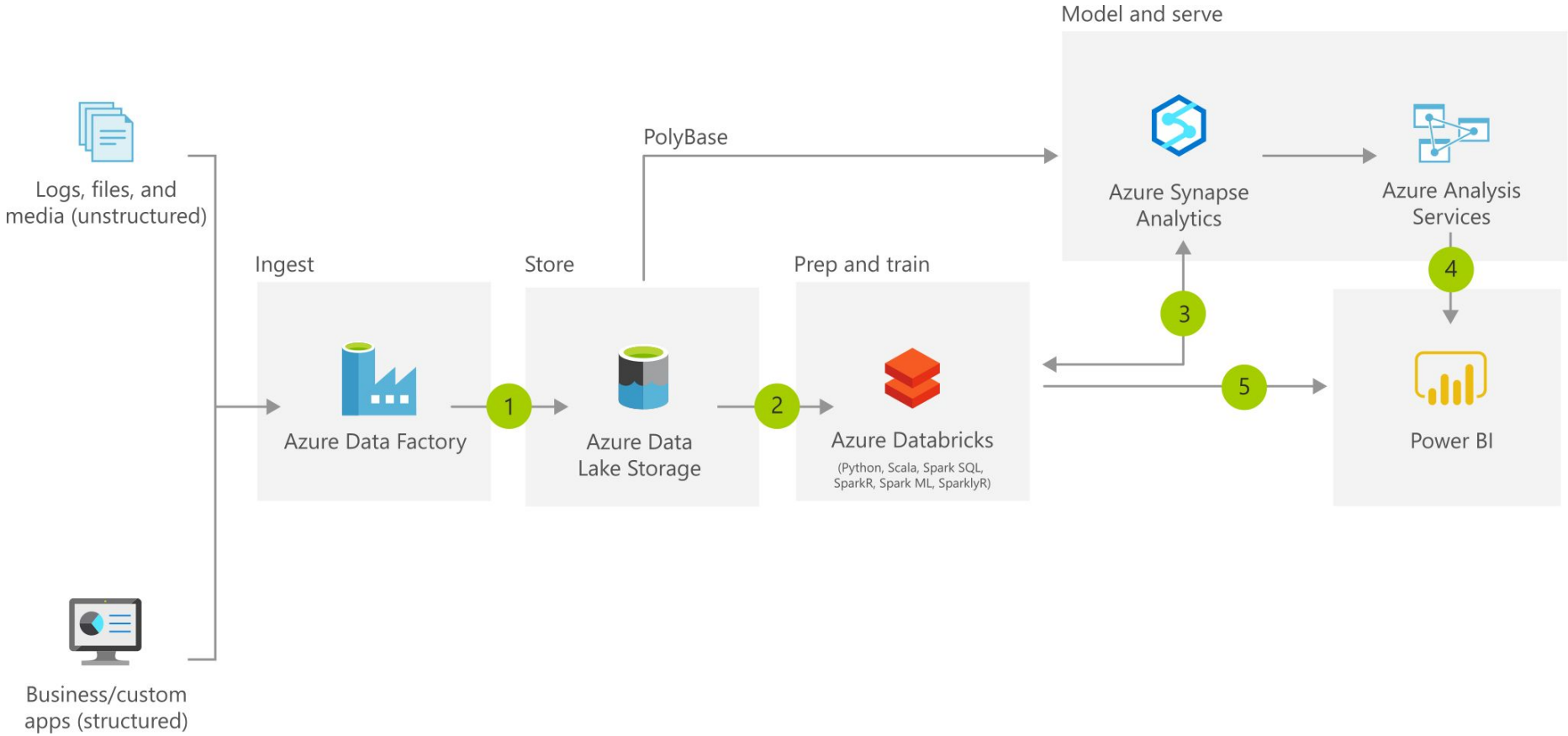

SoftwareArchitect
.ca

# DP-900

Microsoft Certified: Azure Data Fundamentals

# Data Warehouse

# Describe the components of a modern data warehouse


Model and serve
Logs, files, and media (unstructured)
PolyBase
Azure Synapse Analytics
Azure Analysis Services
Ingest
Store
Prep and train
4
3
Azure Data Factory
1
Azure Data Lake Storage
2
Azure Databricks
(Python, Scala, Spark SQL, SparkR, Spark ML, SparklyR)
5
Power BI
Business/custom apps (structured)

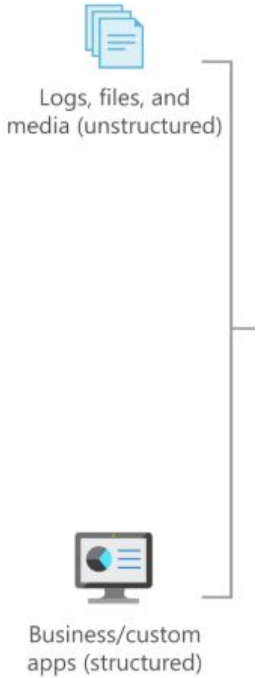
# Data Sources

All data comes from somewhere else

One or more data sources

It can be structured data - existing SQL databases

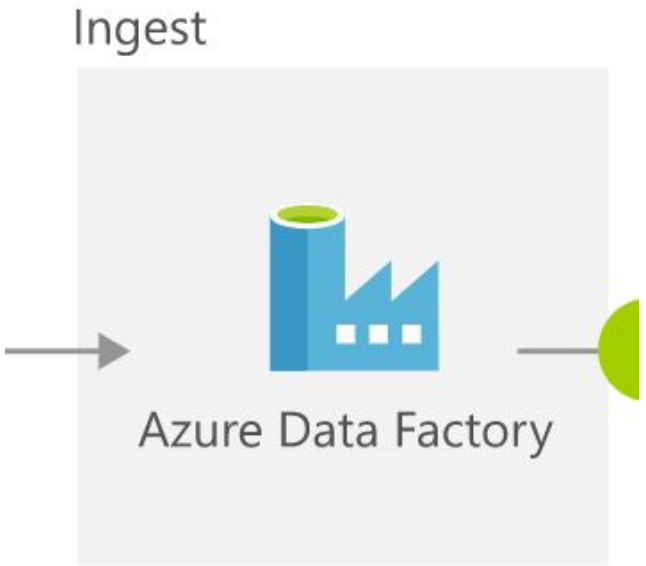
It can be unstructured data - CSVs, JSON, log files



# Azure Data Factory

**Azure Data Factory** is a hybrid data integration service that allows you to create, schedule and orchestrate your ETL / ELT workflows

Azure Data Factory moves the data from outside of Azure to inside of Azure

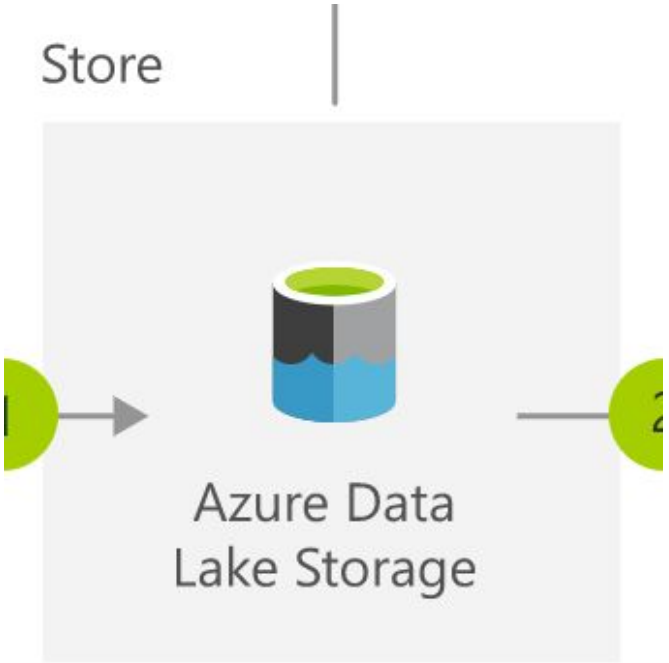


# Azure Blob Storage / Data Lake

**Azure Blob Storage** is a massively scalable object storage for any type of unstructured data - images, videos, audio, documents and other

**Azure Data Lake** is a type of blob storage that is designed to handle even larger amounts of data

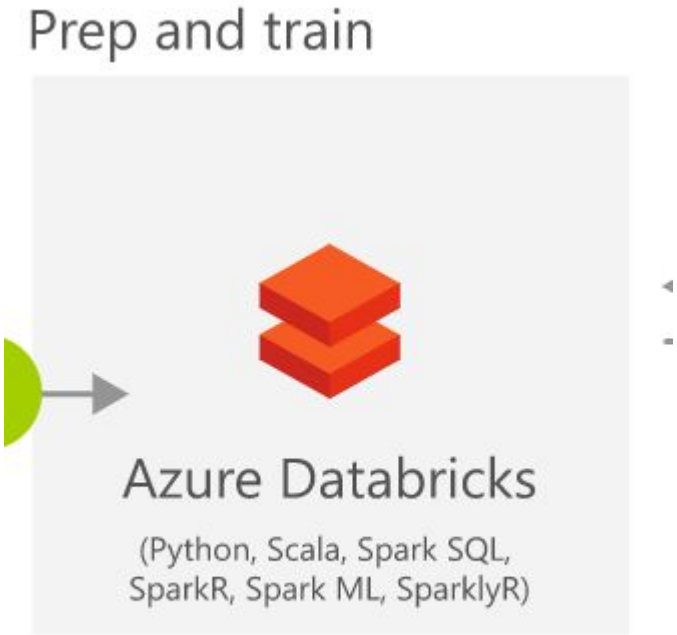
The unprocessed data is stored here



# Azure Databricks

**Azure Databricks** is a fast, easy and collaborative Apache Spark-based analytics platform

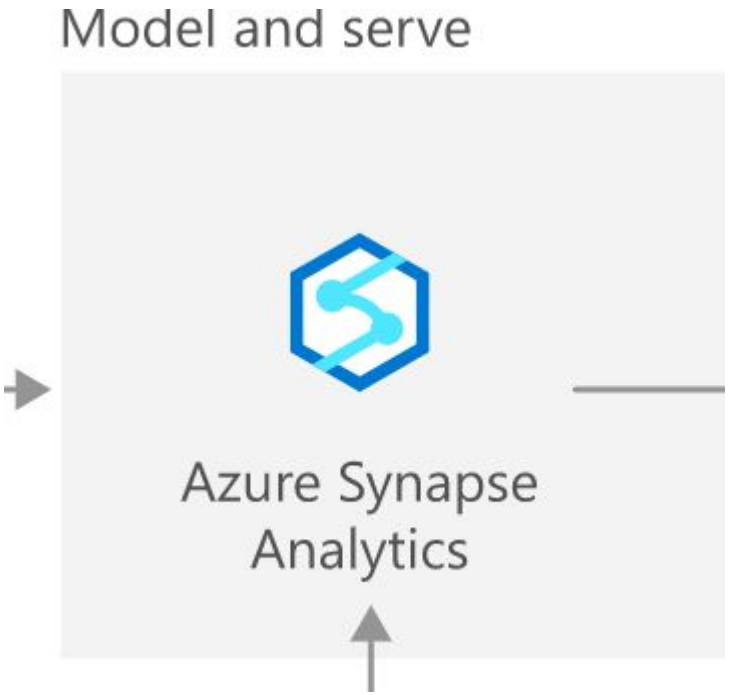
Databricks allows you to manipulate data at large scales, by linking to one data source (like Azure Data Lake), modifying the data, and storing it in another data source (like Azure Synapse Analytics / SQL DW)



# Azure Synapse Analytics

**Azure Synapse Analytics** is the fast, flexible and trusted cloud data warehouse that lets you scale, compute and store elastically and independently, with a massively parallel processing architecture

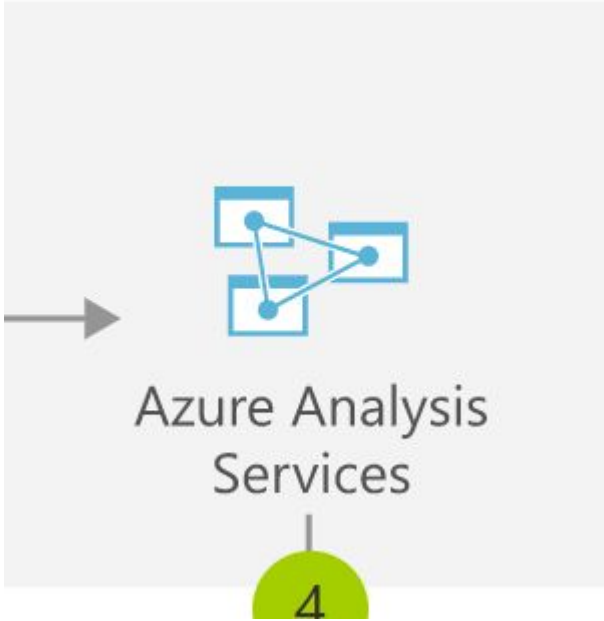
Data is optimized for read-only queries



# Azure Analysis Services

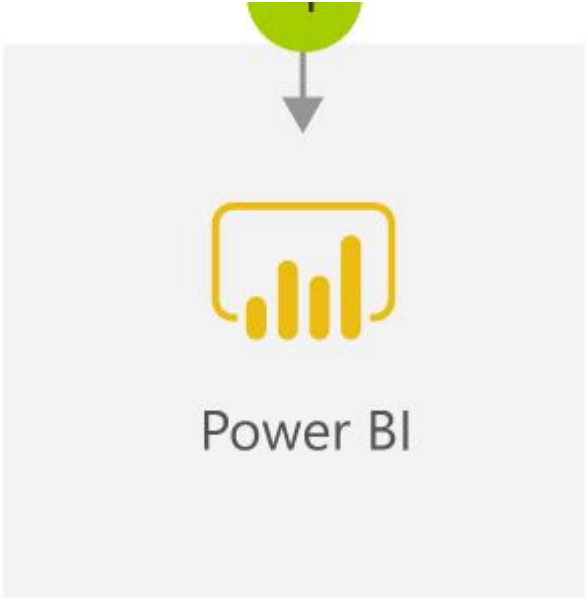
**Azure Analysis Services** is an enterprise grade analytics as a service that lets you govern, deploy, test, and deliver your BI solution with confidence

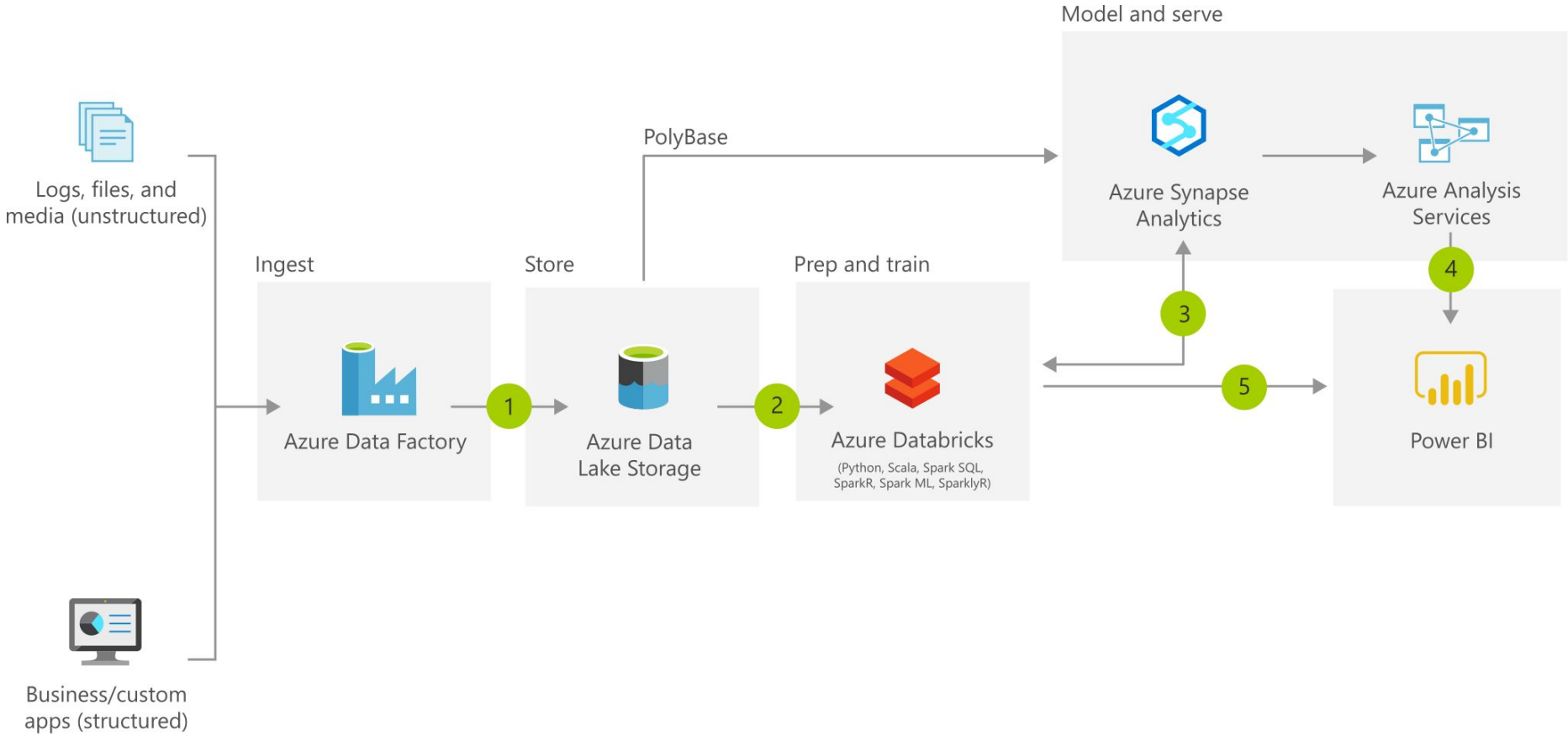
Data is optimized for complex queries with cubes and dimensions



# Power BI

**Power BI** is a suite of business intelligence tools that make is easier for users to look at data, analyze it, and create reports. They can then publish them to the organization to consume on the web or on mobile devices.




Model and serve
Logs, files, and media (unstructured)
PolyBase
Azure Synapse Analytics
Azure Analysis Services
Ingest
Store
Prep and train
4
3
Azure Data Factory
1
Azure Data Lake Storage
2
Azure Databricks
(Python, Scala, Spark SQL, SparkR, Spark ML, SparklyR)
5
Power BI
Business/custom apps (structured)


SoftwareArchitect
.ca

# DP-900

Microsoft Certified: Azure Data Fundamentals

# Data Ingestion and Processing

# The job of “Data Engineer”

# Data Engineer

Azure data engineers are responsible for data-related implementation tasks that include:

- **provisioning** data storage services;
- **ingesting** streaming and batch data;
- **transforming** data;
- implementing **security** requirements;
- implementing **data retention** policies;
- identifying **performance** bottlenecks; and
- accessing **external data** sources.

# Move on to ingesting and transforming data

# Azure Data Factory

# Data Factory = Data Orchestration

Brings data from external sources…

… through a series of transformations

… into an end data store

# Data Factory Features

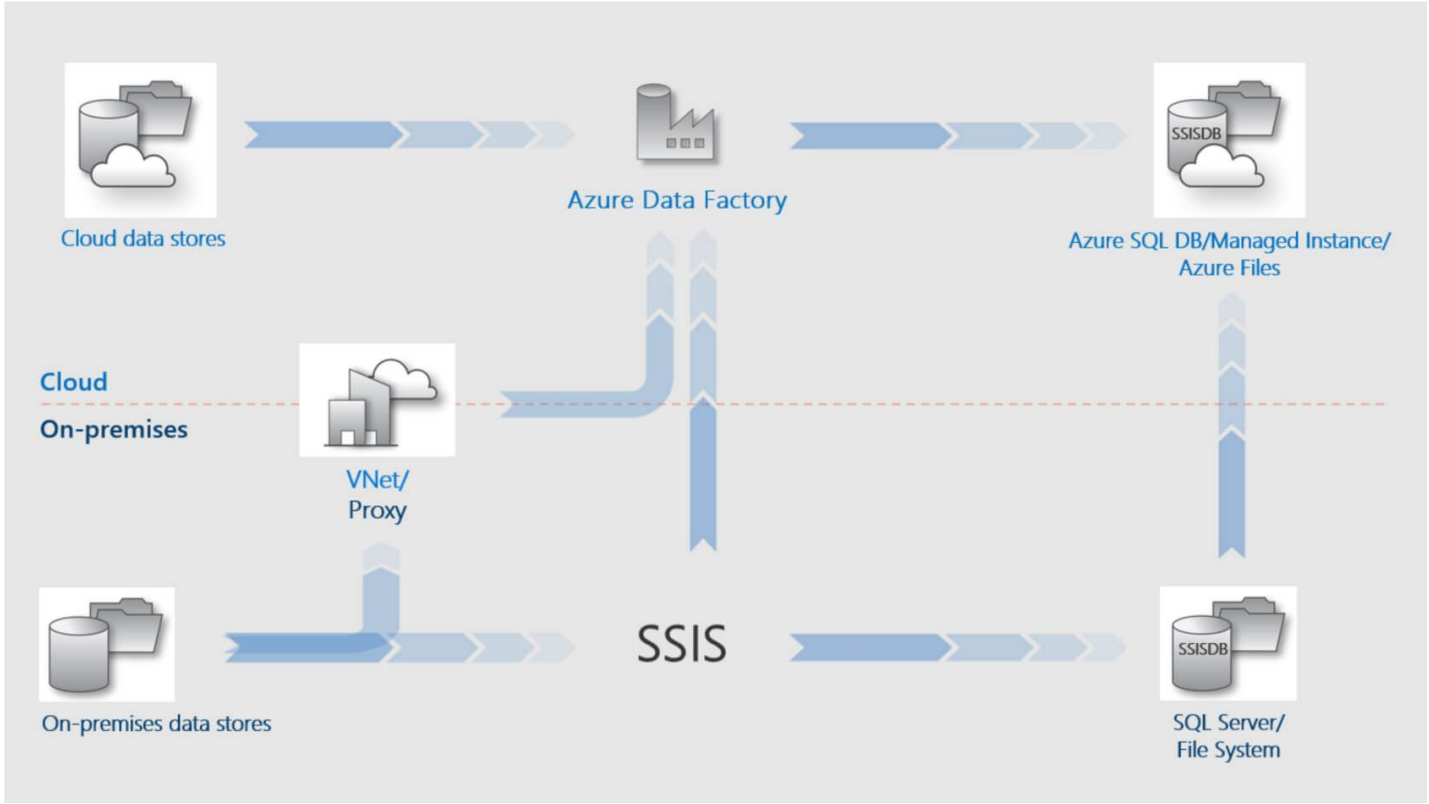
Orchestration platform - **data workflows** from A to B

Data transformation

Create and schedule jobs

Cloud version of ***SQL Server Integration Services (SSIS)***

# Host SSIS Packages in the Cloud



# Data Factory Activities

**Activities** define actions to perform on your data.

- Data movement activities
- Data transformation activities
- Control activities

# Data Factory Pipelines

**Pipeline** - A logical grouping of activities to perform some task

A data factory can contain multiple pipelines

Pipelines can perform their tasks sequentially (serial) or perform several tasks in parallel

# Pipeline Triggers

**Pipeline Run** - an instance of a pipeline execution

Pipelines can be run manually, or started using a trigger

**Trigger** - an event that causes a pipeline to run:

- Scheduled trigger
- Tumbling window
- Event-based

# Manual Pipeline Run

Just as it sounds

You manually start the pipeline to run by executing it

# Scheduled Trigger

Runs at a predetermined recurring schedule

- Every Monday at 8:00 AM
- Every Monday at 8:00 AM and Friday at 5:00 PM
- Every Hour

You can even set start and end times on a trigger, so they don’t fire after a certain date

# Tumbling Window

It’s a type of scheduled trigger that has some interesting properties

You set it up to run at a predetermined interval (i.e. hourly)

It can be set up to run “in the past”

Good for when the pipeline is designed to process data by the time period specified

*“Tumbling windows are a series of fixed-sized, non-overlapping, and contiguous time intervals.”*


SoftwareArchitect
.ca

# DP-900

Microsoft Certified: Azure Data Fundamentals

# Data Visualization

# Exam tip: know the names of the various Power BI components

# Paginated reports

# Paginated Reports

Designed to be printed and shared

Formatted to fit well on a page

Contains all the data, even if it spans multiple pages

Examples: Invoice, Sales Detail Report, Profit/Loss Statement

# PowerBI Feature

SQL Server Reporting Services (SSRS) capability within Power BI

**PowerBI Report Builder** creates paginated reports

Standalone tool separate from Power BI

Reports get published to **Power BI Service**

# Interactive reports

# Interactive

Reports designed to be viewed on screen

Can click for more details, drill down on the data

Report is “interactive”

Very visual, not a dump of data rows

Make use of “hover”

Report can change design / layout if based on user action

# Power BI Feature

Uses Power BI Server installed on premises

Part of Power BI Premium

# Dashboards

# What are Dashboards?

Typically a mixture of chart types on a single page

All the relevant data, at a glance

Click to go to the individual report

Helps identify anomalies visually

# Power BI Content Workflow

# Typical workflow in Power BI

Connect to the data source that has the data

Pull what you need into the in-memory data model

Edit, transform the data as you require

Build reports using Power BI Desktop

Share the report


SoftwareArchitect
.ca

# DP-900

Microsoft Certified: Azure Data Fundamentals

# Thank you!


SoftwareArchitect
.ca