



# Get started with Microsoft Solutions for Data Engineer & Analyst

Narisorn Limpaswadpaisarn  
Client Technology Lead



# Narisorn Limpaswadpaisarn

## *Microsoft Certified Trainer*

CompTIA Security+ CE, ITIL V4, (ISC)<sup>2</sup> CC

Power Platform Solution Architect

Azure Admin Associated & Azure Security Engineer

M365 Enterprise Admin, M365 Message Admin

MCSE Productivity & Data Management and Analytics

MCSA BI Reporting & Office 365

MCP, MTA Databases

**Microsoft Office Specialist 2016 Master**

Email [narisornl@microsoft.com](mailto:narisornl@microsoft.com)

Facebook <https://facebook.com/narisorn.li>

LinkedIn <https://www.linkedin.com/in/narismadz/>

Instagram @Narismadz

Blog <https://medium.com/echoo365>



**Narisorn Limpaswadpaisarn**, Technology Strategist at Microsoft. Once graduated from SIIT TU and Sasin MBA, he realized that the trend is heavily on Cloud Computing. He has some background on Networks and Electrical Engineering. Experiences in Marketing and Computer Networks, but mostly in Music production. He usually play guitar and MOOC on cloud technologies such as SQL, BI, Offices 365 and Server platforms. His role supports customer and Microsoft partner. He got several certifications and awards including MCT, MCSE, MCP and One Microsoft Award.



# Pre-test questionnaire

<https://tinyurl.com/kupretest>

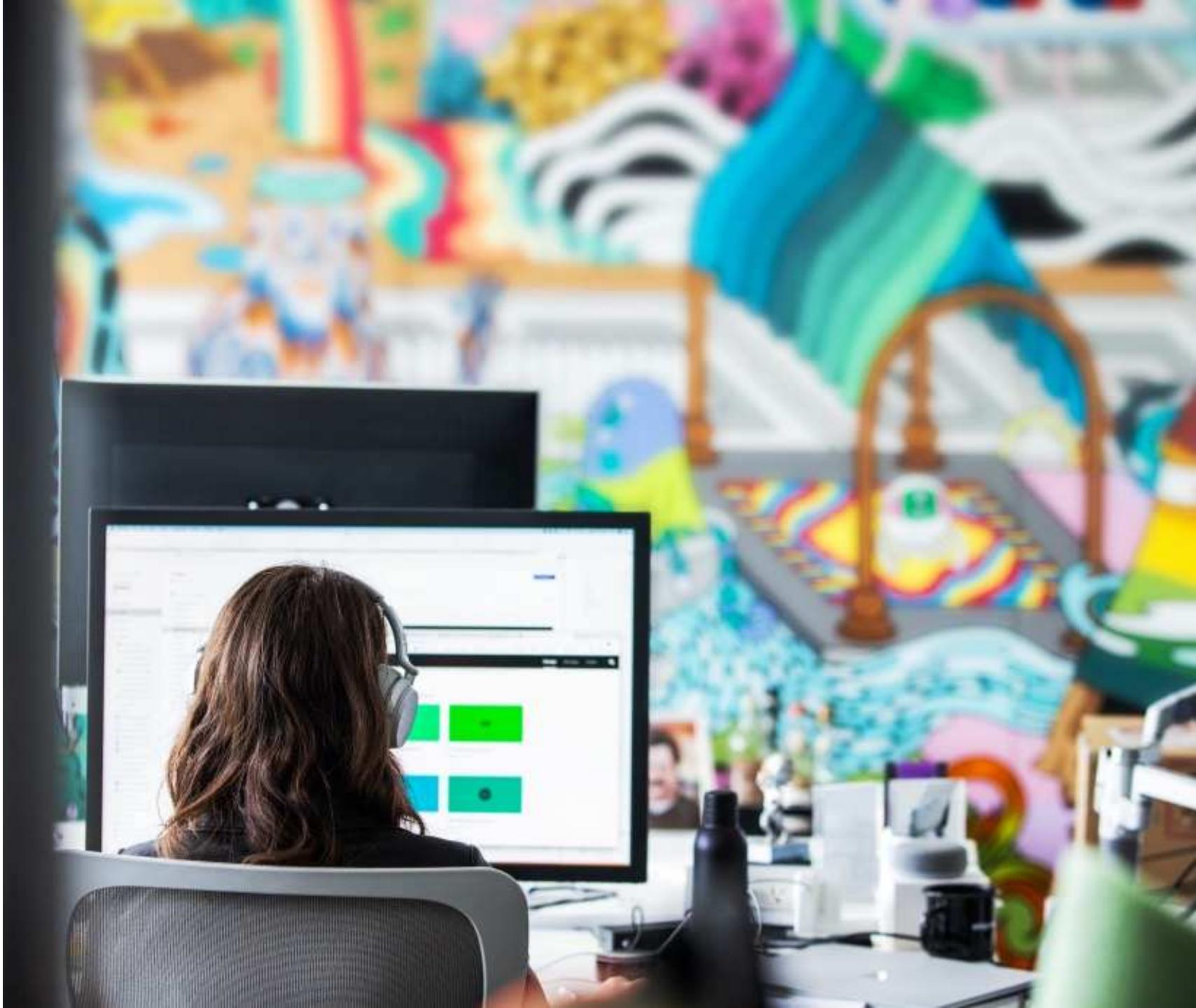
# Slides and Power BI Lab

<https://tinyurl.com/kufabric>

# Data importance and why Data Driven Organization?

**Business Applications Domain**

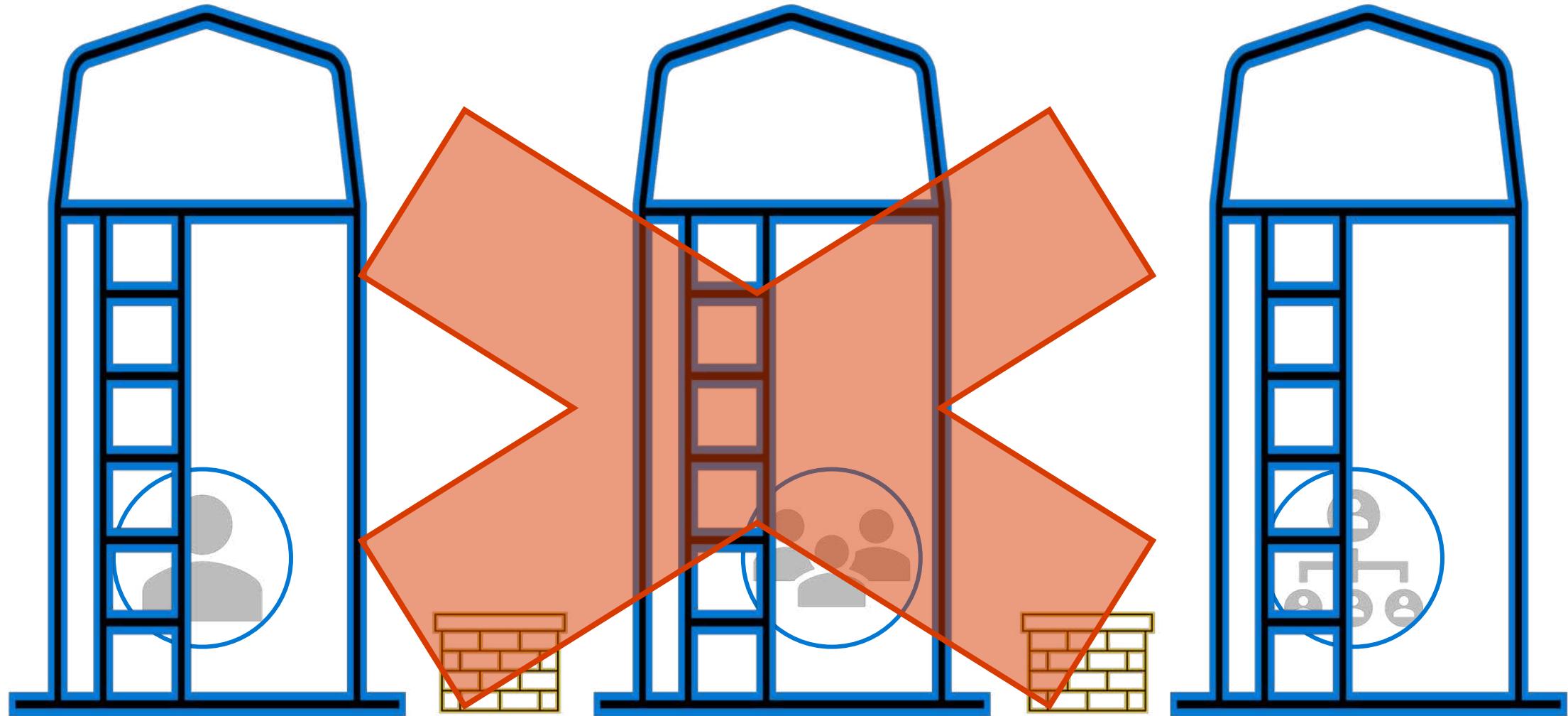
Narisorn Limpaswadpaisarn  
Client Technology Lead



# Today's topic

- Data importance and why Data Driven Organization?
- Cloud fundamentals – Before and Nowadays
- Lab 1: Create SQL Database from Azure Cloud
- Data fundamentals
- Microsoft Power BI (high level)
- Lab 2: Create a basic Power BI Report
- Microsoft Fabric (high level)
- Lab 3: Explore Data Warehouse
- Lab 4: Ingest data with pipeline (Data Factory)
- Lab 5: Explore Data Lakehouse
- Data Governance (high level)

ปัญหาขององค์กรคือข้อมูลกระจัดกระจายไม่รวมศูนย์ (Data Silo), ในมีนโยบายเกี่ยวกับการจัดการข้อมูลที่ดี ในมีระบบสกัดข้อมูล รวมถึงการเข้าถึงข้อมูล



ข้อมูลไม่สามารถขยายผลต่อได้, No single source of truth, ไม่มีการตรวจสอบ model ของข้อมูล  
หรือ มีแต่ไม่ทั่วถึง และไม่สามารถ collaboration ได้

# ผลกระทบของ Data silos & และการที่องค์กรขาดระบบ Centralized Data Platform

**Inconsistent** Definitions, hierarchies, metrics & KPI's ข้อมูลไม่เสมอต้นเสมอปลาย

---

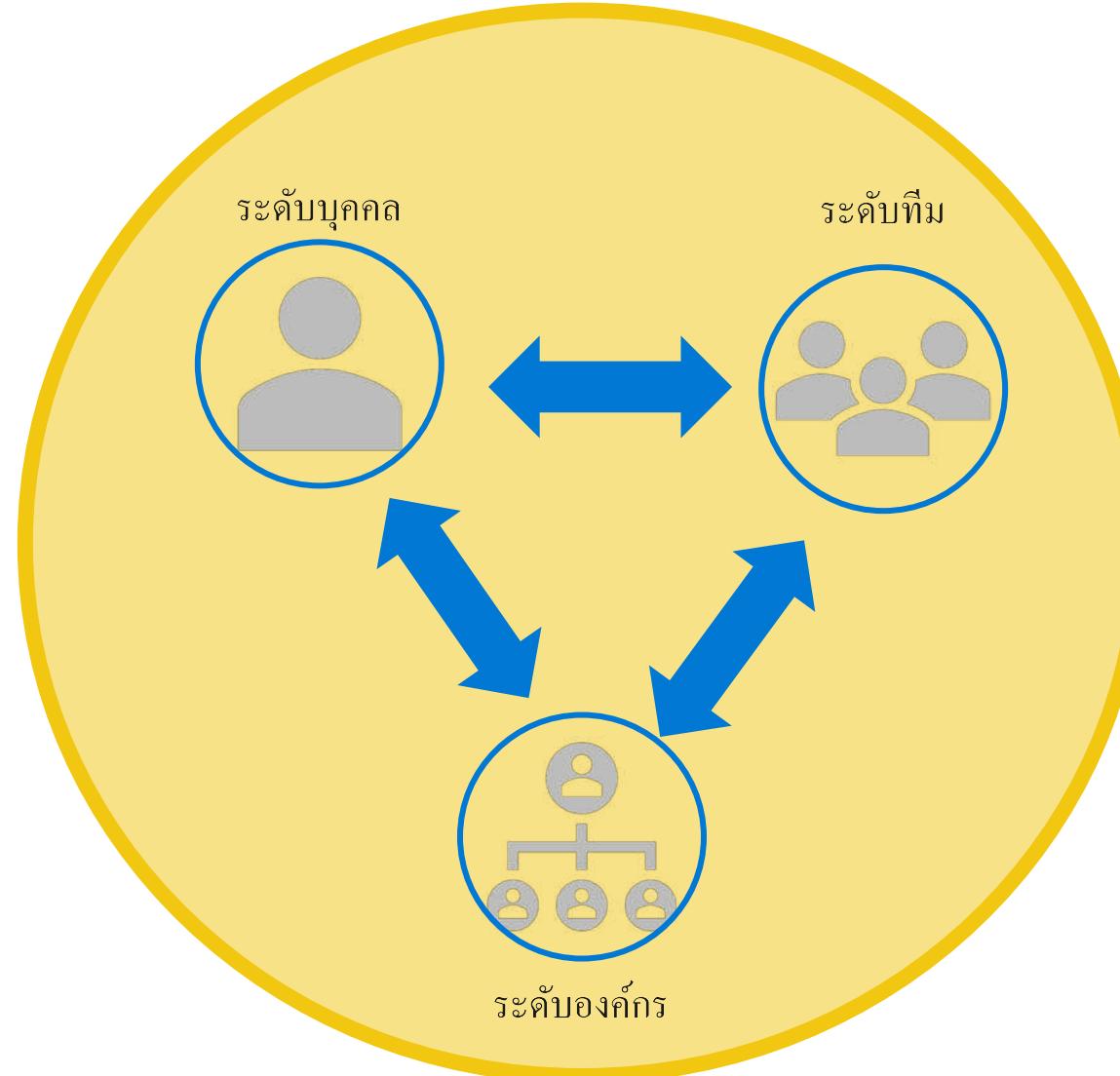
**75%** สำหรับเวลาพนักงานในการรวบรวมข้อมูล และ ตรวจสอบความถูกต้อง

---

**78%** reports ที่สร้างจาก “offline environments” – ข้อมูลไม่ up-to-date

---

การทำระบบการจัดการข้อมูลที่ดี ควบคู่กับการทำ data culture จะทำให้บุคลากรสามารถใช้งานข้อมูลได้อย่างถูกต้อง, มีระบบปลอดภัย,  
และสามารถทำข้อมูลได้ด้วยตนเองแบบ self services ด้วยชุดข้อมูลที่เป็นมาตรฐาน



# Data Management Reasons

Why do we keep all that data still available?

**IDC's CMO of Advisory Services Source:** "[Most collected] data is garbage. IDC's Group Researchers say that some **80% of data collected has no meaning whatsoever**"

**IDG Research Services Source:** "Only 28% of data stored today represents any value to day-to-day business"

**Forbes Source:** "Average cost of Data Management is 3.5% of revenue"

The benefit of keeping everything makes us insure ourselves from fears that we may end up deleting something important.

Understand what you need to do.

# Data Driven Organization

## **Engage your customer (sell more)**

- Develop models to forecast demand and set prices
- Programmatically identify at risk customer (churn) and suggest corrective action
- Personalize customer's experiences based on behavior or segment

## **Empower your employee (do more)**

- Upskilled/Reskilled employees for data skilled, adoption and management (self-service BI etc.)
- Allow employee to define and receives fast alert to effectively manage business changes
- Analyze employee's performance and action for improvement suggestions

## **Optimize your operation (achieve more)**

- Predict equipment downtime, customer's behavior and proactive fix
- Predict operation defining events to prevent cost, downtime or optimized asset utilization

## **Transform your product and Services (grow more)**

- Improve design by capturing the data on use and performance
- Monitor products remotely and drive customer support based on insights
- Transform from product manufacturer to service provider with added value

# Data-driven transformations yield significant benefits

**54%**

Increase in  
revenue performance

**44%**

Faster time to market

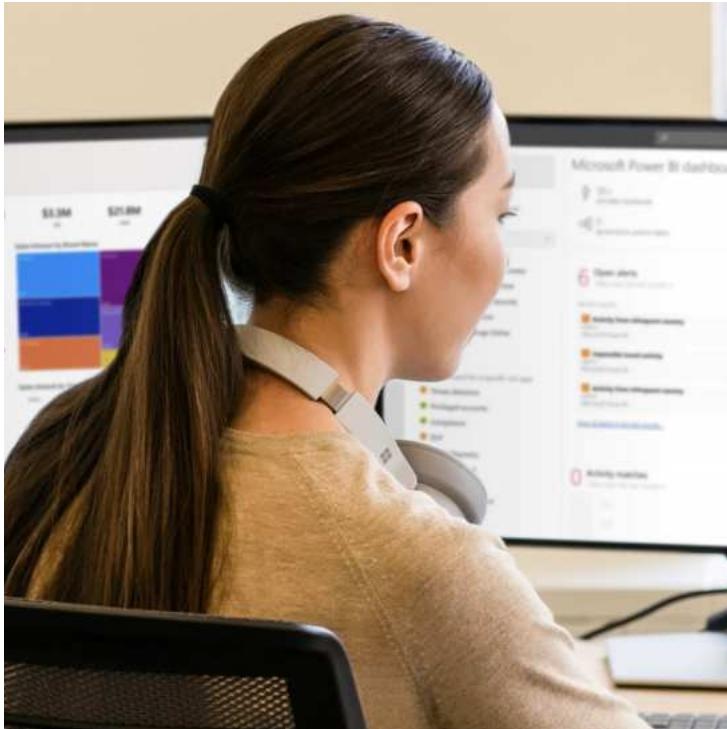
**62%**

Improvement in  
customer satisfaction

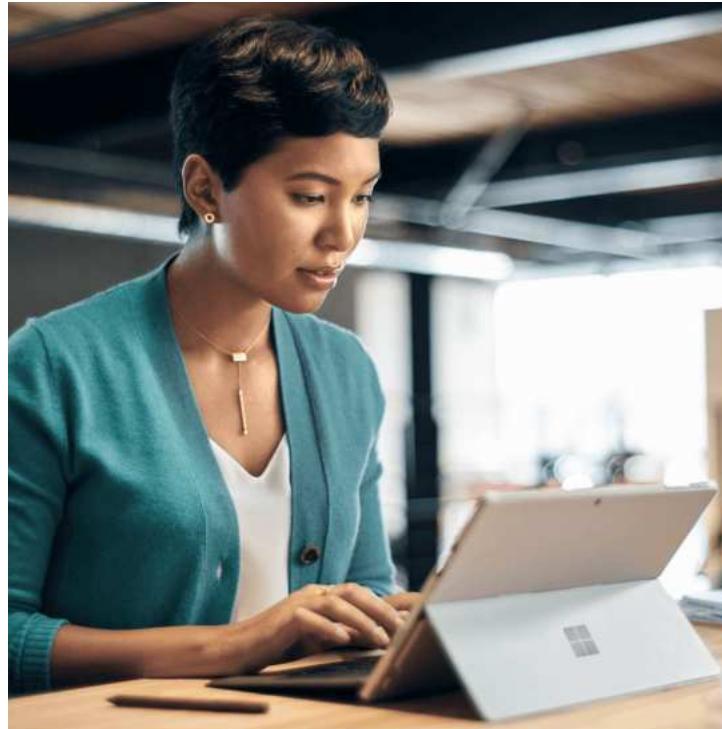
**54%**

Increased profit results

# คุณลักษณะขององค์กรที่มี data culture



ทุกคนทุกตำแหน่งมีความต้องการ และ สามารถหาและ  
วิเคราะห์ข้อมูลเชิงลึกได้



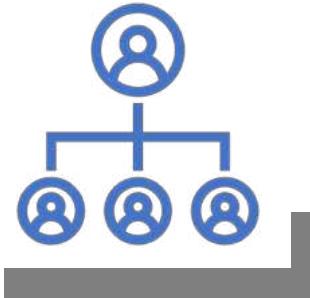
รวมศูนย์ข้อมูลโดยสามารถทำให้เข้าถึงข้อมูลได้โดย  
ปลดภัยและได้ระดับชั้นข้อมูลได้รับการอนุมัติเข้าถึง  
ตรงตามตำแหน่งงาน



ทุกคนทุกตำแหน่งมีการเรียนรู้ต่อนเวลา (**reskill**)

# Data Driven Organization

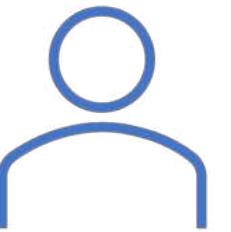
## Enterprise analytics



## IT needs

Governance  
Security  
Scalability  
Visibility  
Sustainability

## Self-services analytics

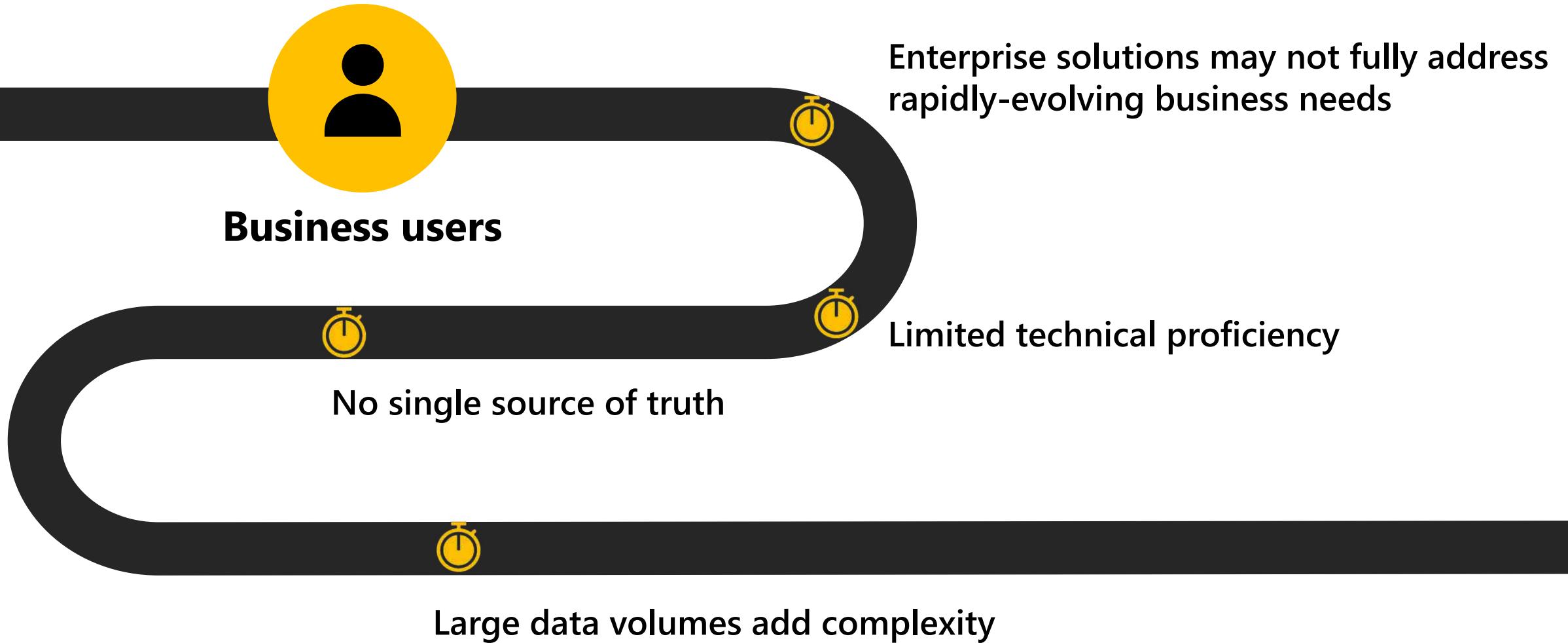


## Business needs

Agility  
Speed  
Flexibility



# Self-Service Analytics Challenges



# Enterprise Analytics Challenges



IT



Limited human resources

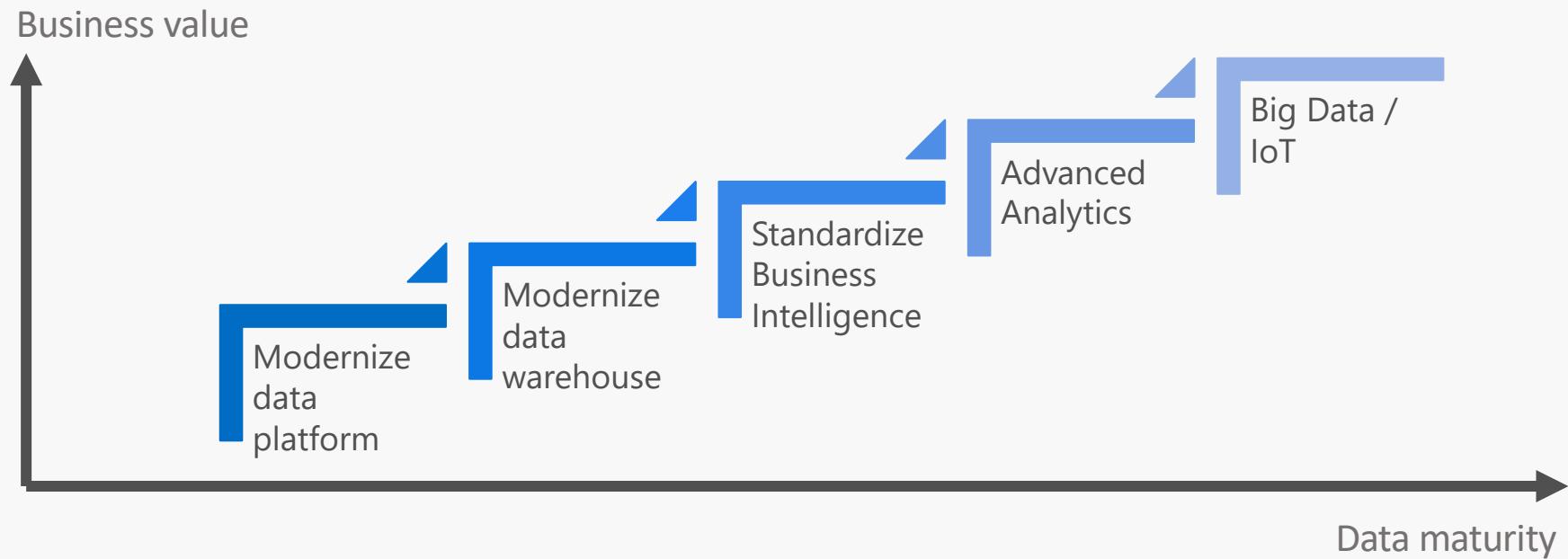
High workloads, long lead times



Large data volumes and data variety add complexity

# What the leaders do

TO BECOME DATA DRIVEN



Note: Mostly, these responsibilities are carried by separate departments

# Legal Reasons

Policy and Regulations drive us into data management, too.

Per Breached Record

**200 EUR**

per Used ID

**200 EUR**

per Document

**200 EUR**

per ID Number

Per Breached Event

**5.000.000 EUR**

**200 EUR**

per eMail

**200 EUR**

per Record

**200 EUR**

per Address

per Breach Event

# ตลาดแรงงานสาย Data และ ความสนใจใน role ของเด็กรุ่นใหม่



LinkedIn search results for "data analyst" in Thailand:

**Data analyst in Thailand** 1,910 results

**Data Analyst**  
UOB - Bangkok, Bangkok City, Thailand (Hybrid)

**Job Responsibilities**

The Business Analytics Analyst is a trainee professional role. Requires a good knowledge of the range of processes, procedures and systems to be used in carrying out assigned tasks and a basic understanding of the underlying concepts and principles upon which the job is based. Good understanding of how the team interacts with others in accomplishing the objectives of the area. Makes evaluative judgements based on the analysis of factual information. They are expected to resolve problems by identifying and selecting solutions through the application of acquired technical experience and will be guided by precedents. Must be able to exchange information in a concise way as well as be sensitive to audience diversity. Limited but direct impact on the business through the quality of the tasks/services provided. Impact of the job holder is restricted to own job.

**Responsibilities:**

- Assist with solutions & implementations
- Performs relatively complex and varied assignments
- Demonstrates technical/administrative expertise
- Applies specialized functional area skills and basic knowledge including underlying professional/technical principles or standards
- Basic knowledge of the organization and its policies required
- Appropriately assess risk when business decisions are made, demonstrating particular consideration for the firm's reputation and safeguarding Citigroup, its clients and assets, by driving compliance with applicable laws, rules and regulations, adhering to Policy, applying sound ethical judgment regarding personal behavior, conduct and business practices, and escalating, managing and reporting control issues with transparency.

**Senior Data Analyst**  
Central Group - Bangkok, Bangkok City, Thailand (Hybrid)

**Data Engineer – Solution Architect**  
Fujitsu (Thailand) Co., Ltd. - Bangkok Metropolitan Area (Hybrid)

**Business Intelligence Analyst**  
Somboon Advance Technology PCL. - Bang Phli, Samut Prakan, Thailand (Hybrid)

# Data professional roles



## Database Administrator

- Database provisioning, configuration and management
- Database security and user access
- Database backups and resiliency
- Database performance monitoring and optimization

## Data Engineer

- Data integration pipelines and ETL processes
- Data cleansing and transformation
- Analytical data store schemas and data loads

## Data Analyst

- Analytical modeling
- Data reporting and summarization
- Data visualization

# Microsoft Mission

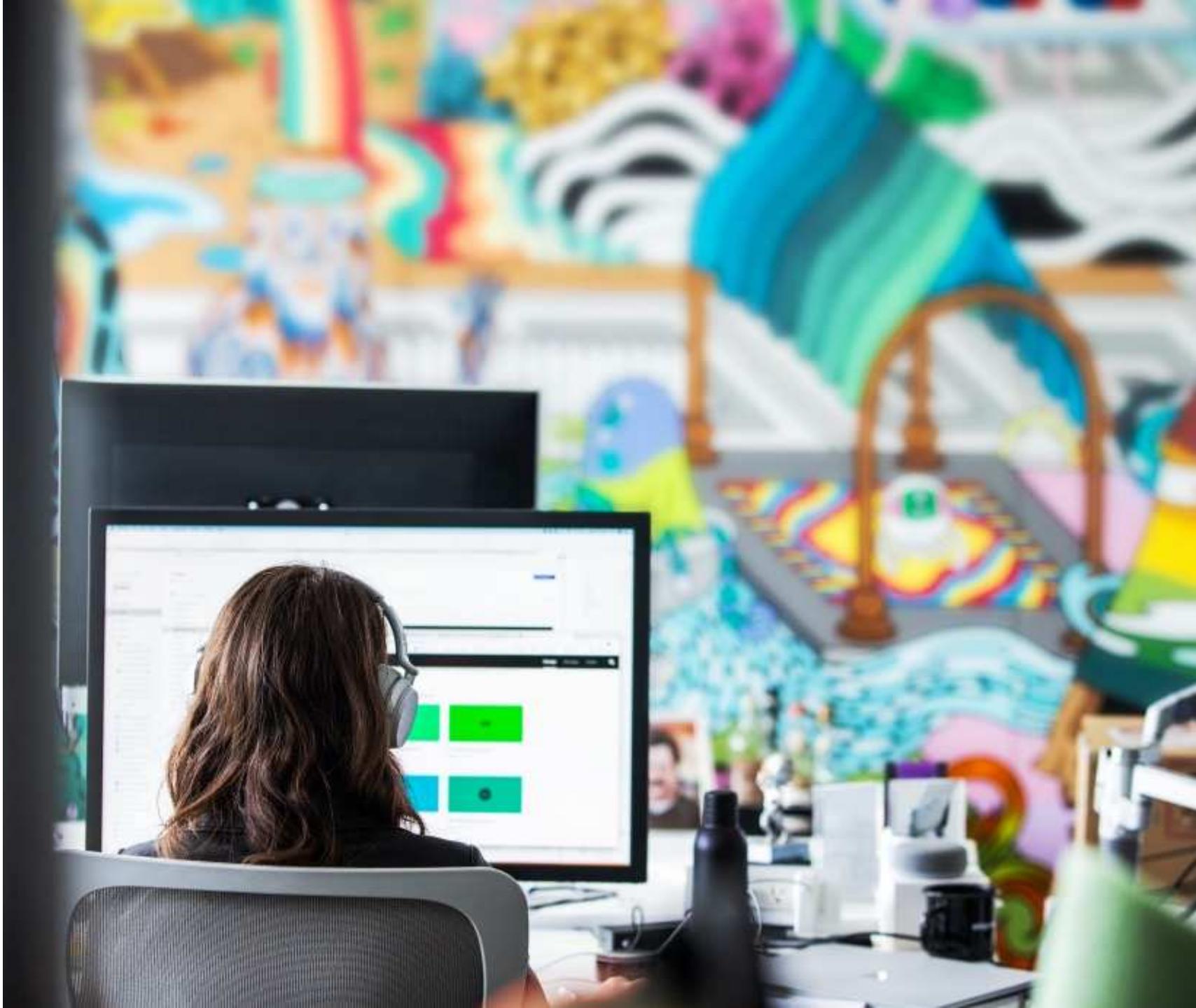
Empower every person and every organization  
on the planet to achieve more.



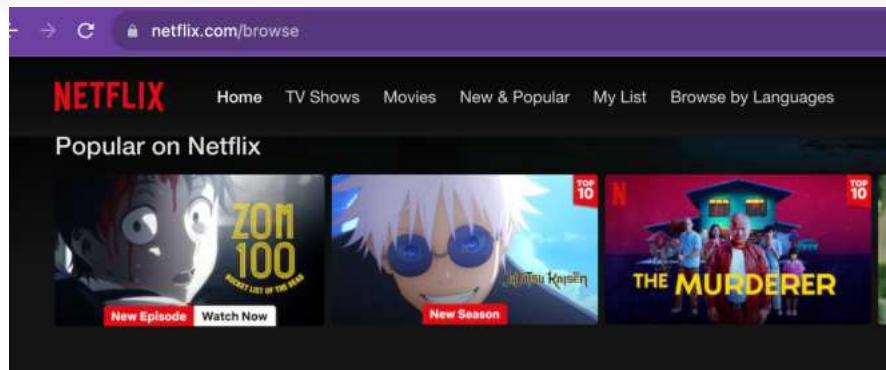
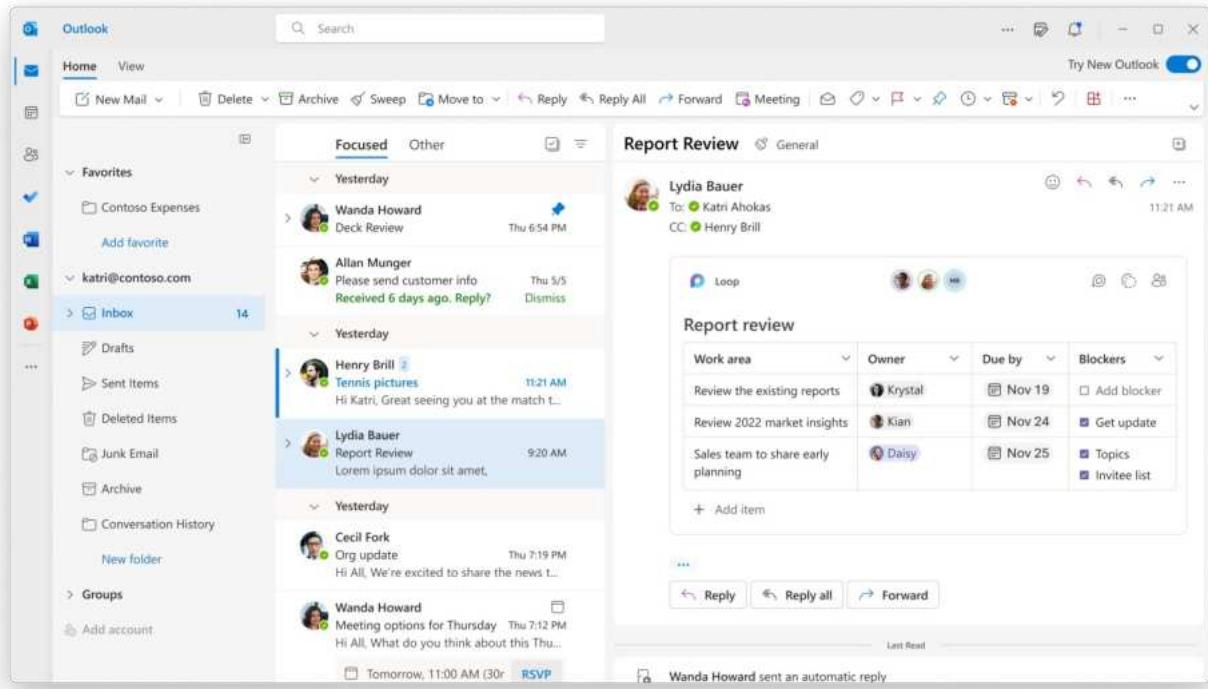
# Cloud fundamentals - Before and Nowadays

**Business Applications Domain**

Narisorn Limpaswadpaisarn  
Client Technology Lead

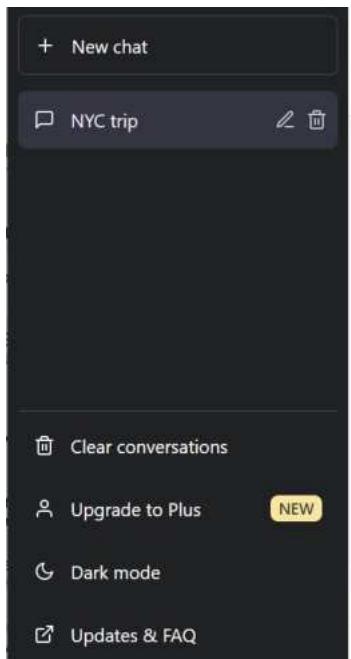


# Server is all around us



# Cloud computing with AIs powered

- ChatGPT and Bing Chat



When is the best time to visit New York City?

New York City is a great destination to visit all year round, with each season offering a unique experience. However, the best time to visit really depends on your preferences and what you plan to do while you're there.

If you're looking for pleasant weather and fewer crowds, then the best time to visit New York City is during the spring (April to June) or fall (September to November). During these months, you can enjoy mild temperatures, colorful foliage, and plenty of outdoor activities.

Summer (July to August) is the busiest time in New York City, with tourists flocking to the city to enjoy the warm weather, outdoor concerts, and festivals. However, it can be hot and I Regenerate response ack accordingly.

Microsoft Bing SEARCH CHAT

narisom@... Protected How do I make an omelette?

✓ Searching for how to make an omelette  
✓ Generating answers for you...

Your personal and company data are protected in this chat.

Here is a simple recipe for making an omelette:

1. Beat two or three eggs in a bowl.
2. Heat a non-stick pan over medium heat.
3. Add butter to the pan and let it melt.
4. Pour the beaten eggs into the pan.
5. Let the eggs cook for about 30 seconds until they start to set.
6. Use a spatula to gently lift the edges of the omelette and let the uncooked eggs flow underneath.
7. Add your desired filling (e.g., cheese, ham, mushrooms) to one side of the omelette.
8. Use a spatula to fold the other side of the omelette over.

Your personal and company data are protected in this chat

New topic Ask me anything...

Explore

# Before cloud – Tower, Mainframe, Blade & Rack server

- Host on customer's premises (on-premise)
- Also called private cloud



# Before cloud – Server room to Data Center

- Server room - one room, the whole floor or entire building (Data Center)
- If company hosted this on local providers, we called local cloud



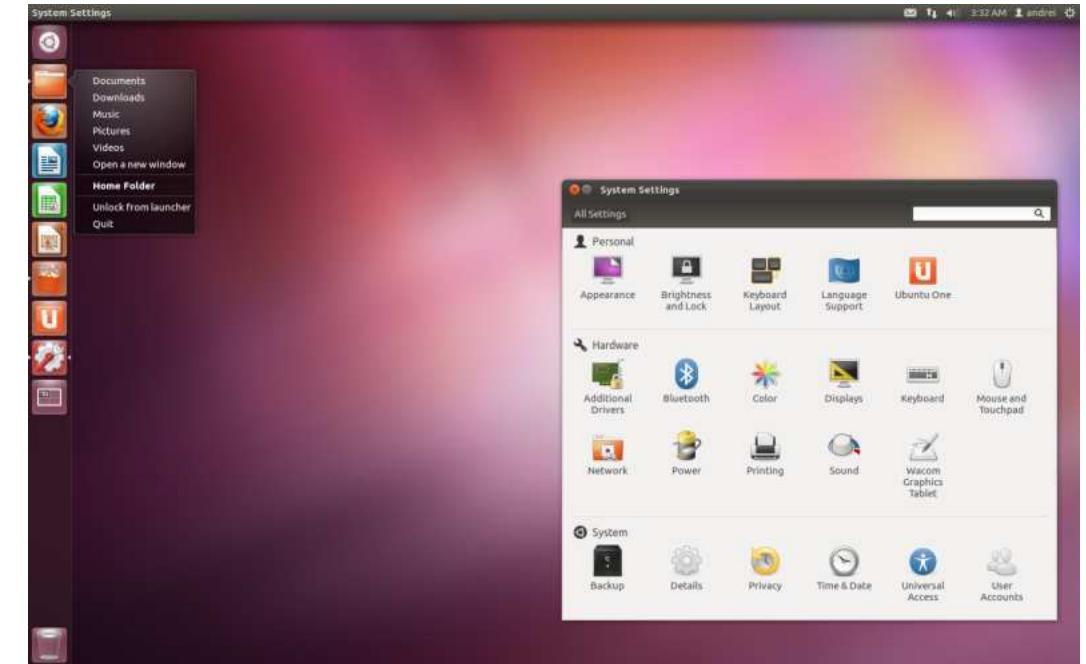
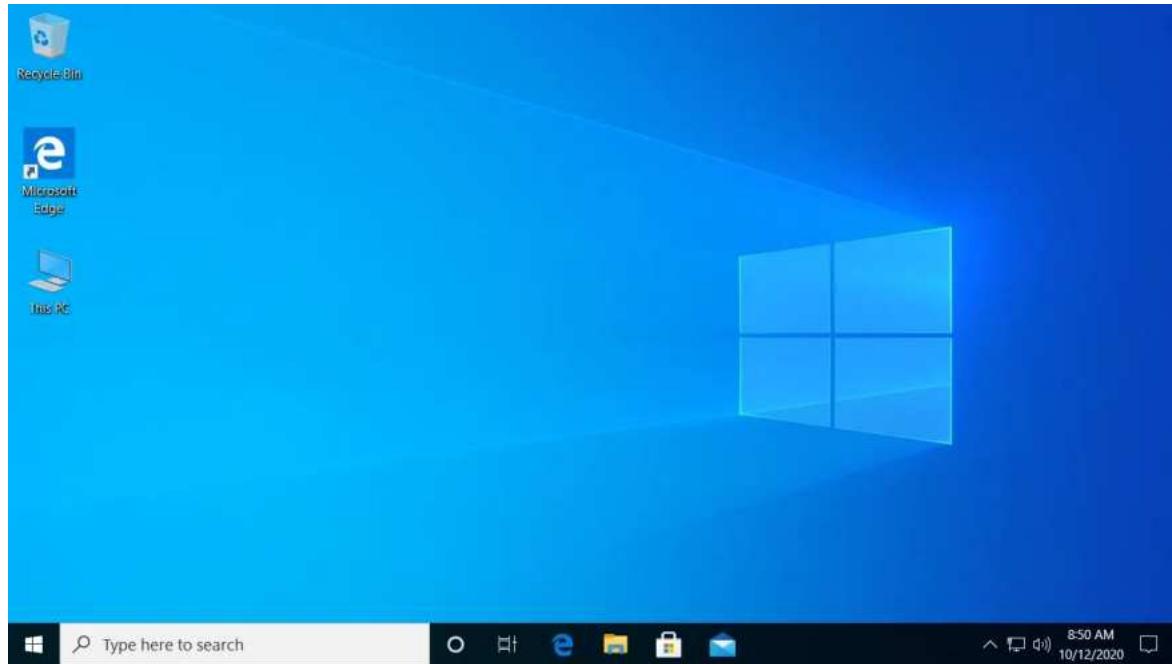
# Before cloud – 24x7 for IT Operations

- Heating, Ventilation and Air Conditioning (HVAC)
- Backup (tape changes etc.)
- Patching & Software update
- Hardware maintenance & replace
- Electrical system and Network wiring
- Physical security (ID card, CCTV etc.)
- and more



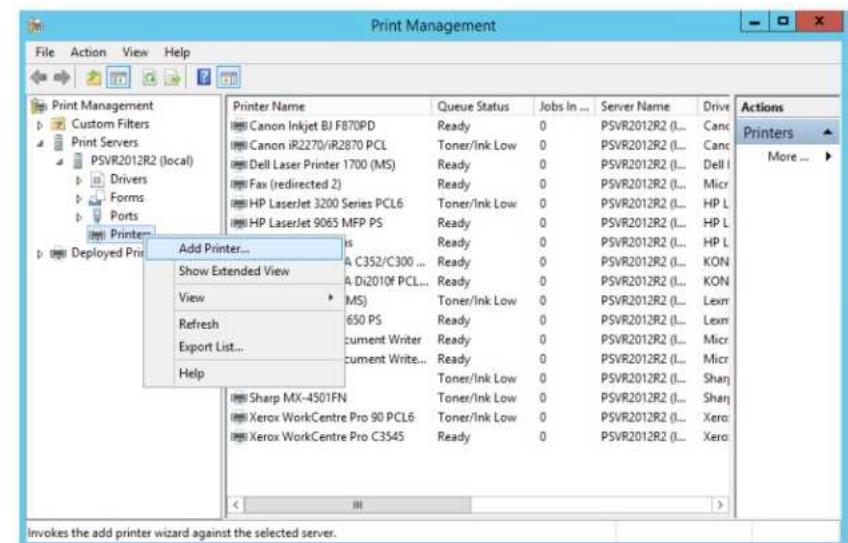
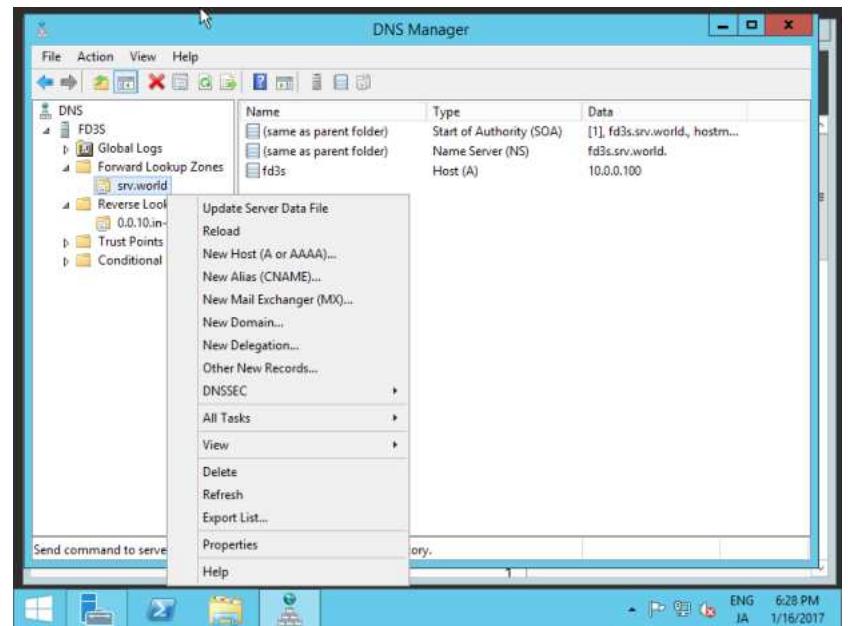
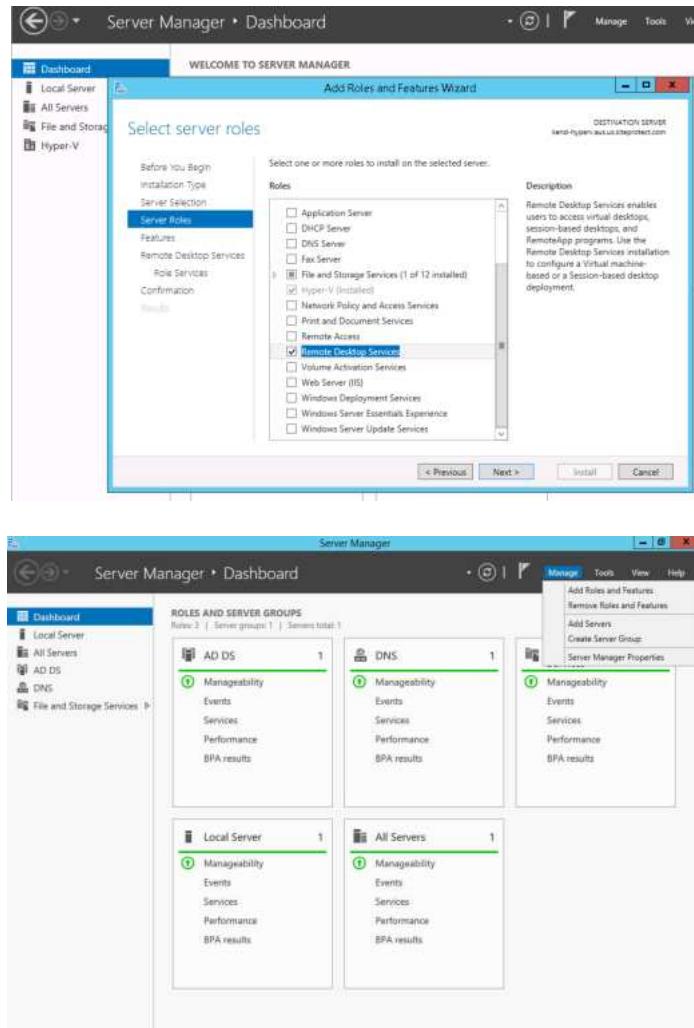
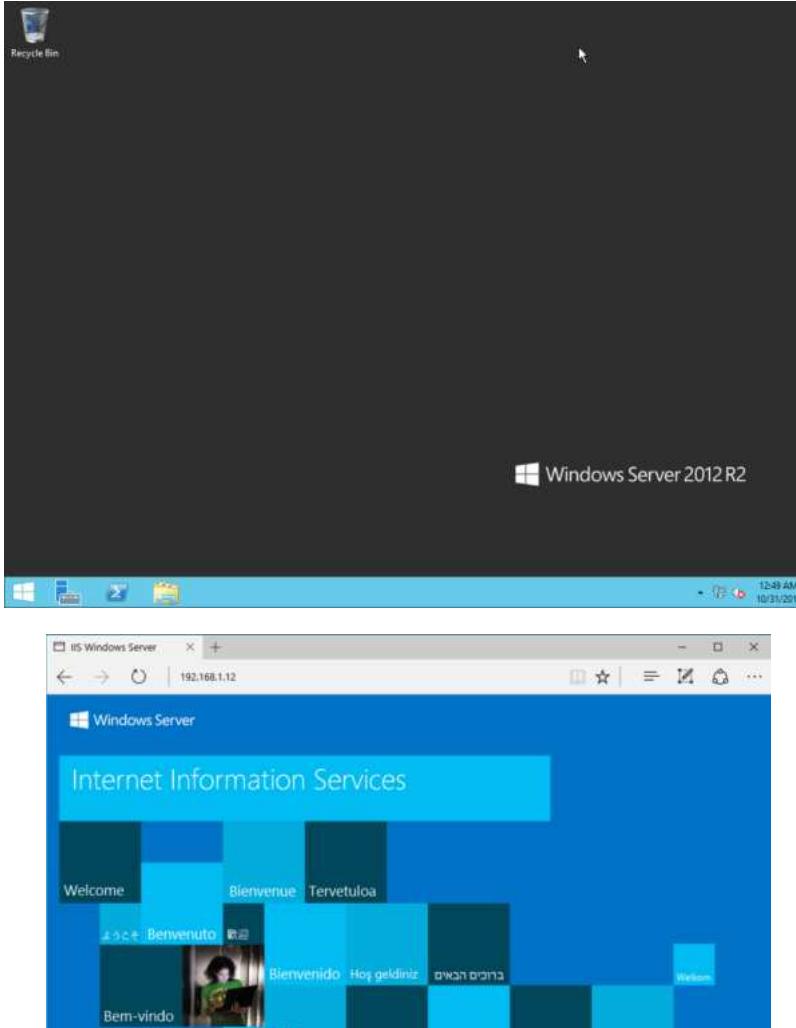
# Before cloud – Client Operating System

- Client OS - Ex. Windows 10 & Ubuntu 12.04



# Before cloud – Server Operating System

- Server OS - Ex. Windows Server 2012 R2



# Before cloud – Client Operating System

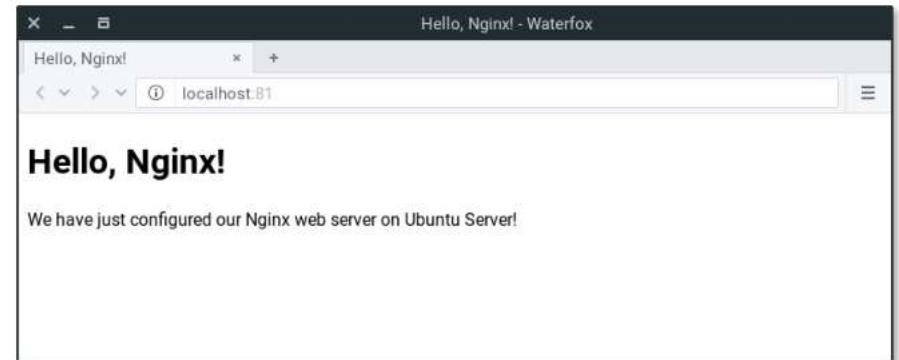
- Server OS - Ex. Ubuntu server

```
root@ip: ~
root@ip's password:
Welcome to Ubuntu 18.04.2 LTS (GNU/Linux 4.15.0-45-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

 * Canonical Livepatch is available for installation.
   - Reduce system reboots and improve kernel security. Activate at:
     https://ubuntu.com/livepatch
New release '20.04.2 LTS' available.
Run 'do-release-upgrade' to upgrade to it.

Last login: Thu Apr 29 14:47:05 2021 from 10.1.149.227
root@ip:~# networkctl status
*           State: routable
              Address: 10.1.149.55 on ens192
                           fe80::250:56ff:fe91:69a8 on ens192
              Gateway: 10.1.149.1 (Palo Alto Networks) on ens192
                  DNS: 10.1.149.10
                           1.1.1.1
Search Domains: cloud.local
root@ip:~#
```



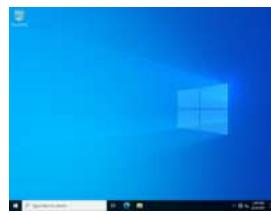
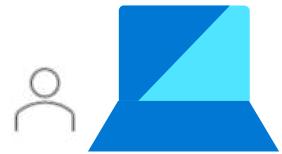
```
test@test-machine:~$ curl -i 127.0.0.1
HTTP/1.1 200 OK
Server: nginx/1.18.0 (Ubuntu)
Date: Thu, 24 Sep 2020 13:03:00 GMT
Content-Type: text/html
Content-Length: 612
Last-Modified: Thu, 24 Sep 2020 10:40:45 GMT
Connection: keep-alive
ETag: "5f6c77ad-264"
Accept-Ranges: bytes

<!DOCTYPE html>
<html>
<head>
<title>Welcome to nginx!</title>
<style>
  body {
    width: 35em;
    margin: 0 auto;
    font-family: Tahoma, Verdana, Arial, sans-serif;
  }
</style>
</head>
<body>
<h1>Welcome to nginx!</h1>
<p>If you see this page, the nginx web server is successfully installed and
working. Further configuration is required.</p>
<p>For online documentation and support please refer to
<a href="http://nginx.org/">nginx.org</a>.<br/>
Commercial support is available at
<a href="http://nginx.com/">nginx.com</a>.</p>
<p><em>Thank you for using nginx.</em></p>
</body>
</html>
```

# Before cloud – Virtualization Technology

- Virtual Machine (VM) – allocating the server's hardware resources
- Running different workloads & utilizing single hardware
- Running software that requires an older version of OS
- Easier when moving to other hosts
- Issue only affect on guest level

Users remote to VM via  
RDP (Windows) or SSH (Linux)



```
Ubuntu@Ubuntu-OptiPlex-5090: ~$ hostnamectl
Static hostname: computer-vm
        Icon name: computer-vm
          Chassis: Desktop
       Machine ID: 193d527dd0e474ad0d14a9df172
         Boot ID: 0091d136e70745800e610d93e714d
  Operating System: Ubuntu 18.10
           Kernel: Linux 4.18.0-19-generic
      Architecture: x86_64
Ubuntu@Ubuntu-OptiPlex-5090: ~$ hostname
Ubuntu
Ubuntu@Ubuntu-OptiPlex-5090: ~$ cat /etc/hostname
Ubuntu
Ubuntu@Ubuntu-OptiPlex-5090: ~|
```



VM Guest 1  
(Windows Server 2022)



VM Guest 2  
(Linux Ubuntu 20.04)



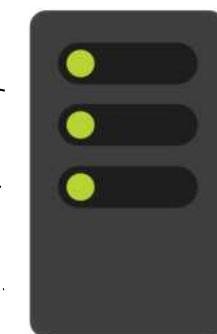
VM Guest 3  
(Windows 10)



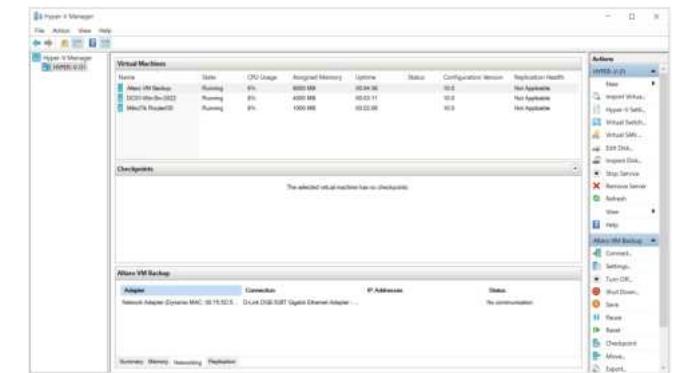
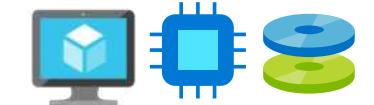
Virtual resources  
vCPU/RAM/Disks



Hardware Server as Host  
Windows Server 2022 with  
Hypervisor role installed



Physical resources  
CPU/RAM/Disks (storage)



Hyper-V manager

# Nowadays – using public hyperscaler cloud

- Small business may use fully public cloud (such as Microsoft Azure, AWS or GCP)
- Hybrid used between on-premise and public cloud (called Hybrid cloud)
- Hybrid used between public cloud provider (called Multi-cloud)

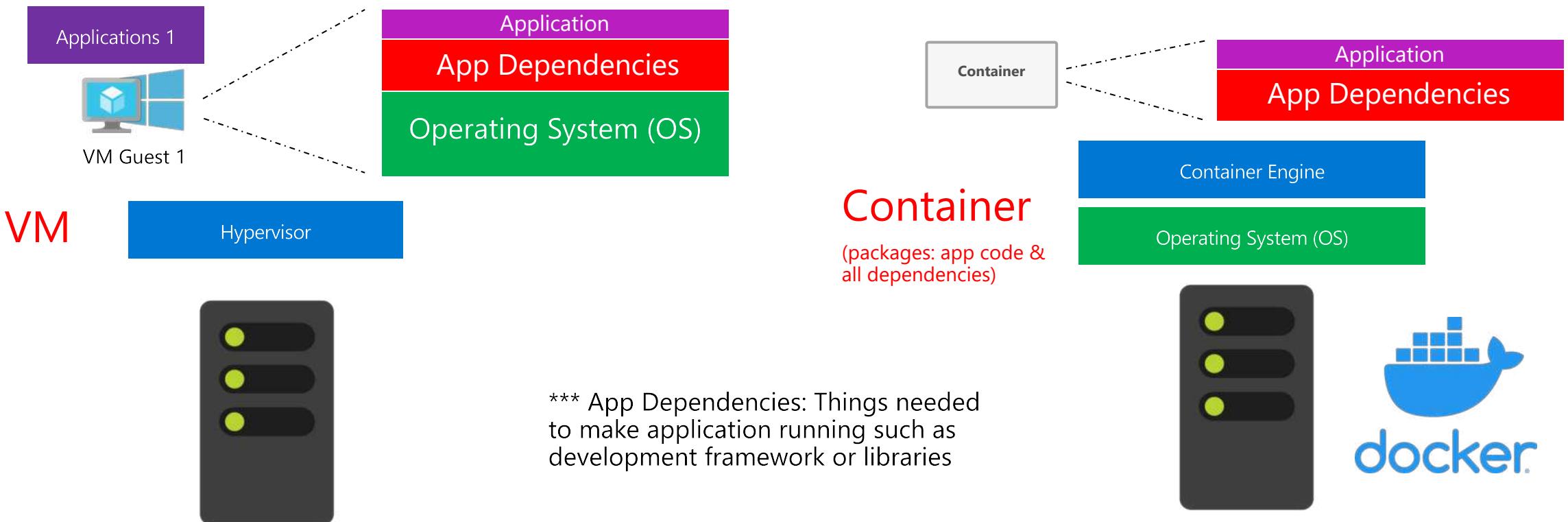
The screenshot shows the Microsoft Azure dashboard under a private dashboard named 'Computer's hope'. The dashboard includes sections for 'All resources', 'Resources' (listing PBI-ReportServer, pbireportserverdiag, Client-Fin1, DC-01, and PBI-ReportServer-vnet), and 'CU-lab-plan' (used by 1 lab). The bottom navigation bar features 'Service Health' and 'Marketplace'.

The screenshot shows a detailed view of a Microsoft Azure virtual machine named 'PBI-ReportServer'. The 'Overview' tab is selected, displaying information such as Resource group (PBI-ReportServer), Status (Stopped (deallocated)), Location (East US), Subscription (ML-MrgEnvMCAP967116-managed-1), and Tags (Environment: Dev/Test). The 'Essentials' section provides a summary of the VM's configuration, including Operating system (Windows), Size (Standard D4s v3 (4 vcpus, 16 GiB memory)), Public IP address (dynamic), Virtual network/subnet (PBI-ReportServer-vnet/default), DNS name (Not configured), and Health state (Unknown).

Virtualization (VM) on cloud – No hardware required

# Nowadays – Serverless Applications

- Virtualization: full-blown OS (take more storage) need to patched, upgrade and maintained
- Container: more lightweight (less storage) and can run with any hardware with compatibility
- Container engine such as Docker
- Modern Application developer using serverless services instead of using hosting in VM



# Cloud Benefits

High availability

Elasticity (Auto) \*\*\*

Scalability (not auto) \*\*\*

Reliability

Predictability

Security

Governance

Manageability

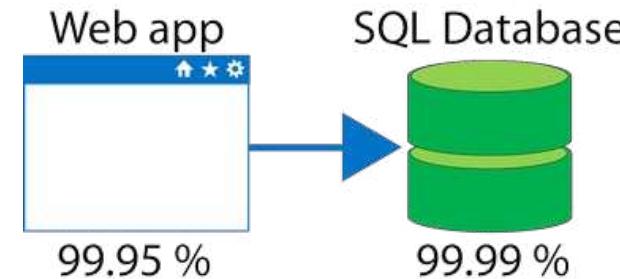
# High Availability (HA)

- **Service Level Agreement (SLA)** - ข้อตกลงระดับการให้บริการ ข้อตกลงดังกล่าวเป็นข้อตกลงที่เกิดขึ้นระหว่างผู้ให้บริการและผู้รับบริการ หรืออาจกล่าวได้ว่า เป็นสัญญาที่ผู้ให้บริการมี ต่อผู้ใช้บริการว่าจะรักษาคุณภาพการให้บริการให้เป็นไปตามที่ระบุไว้ โดยในเอกสารจะต้องระบุข้อตกลงร่วมกันในการให้บริการ ไม่ว่าจะเป็น ขอบเขตการดำเนินงาน ความรับผิดชอบ รวมถึงการรับประกันหากเกิดกรณีต่าง ๆ

SLA: Number of nines

SLA	Downtime/week	Downtime/month	Downtime/year
99%	1.68 hours	7.2 hours	3.65 days
99.9%	10.1 minutes	43.2 minutes	8.76 hours
99.95%	5 minutes	21.6 minutes	4.38 hours
99.99%	1.01 minutes	4.32 minutes	52.56 minutes
99.999%	6 seconds	25.9 seconds	5.26 minutes

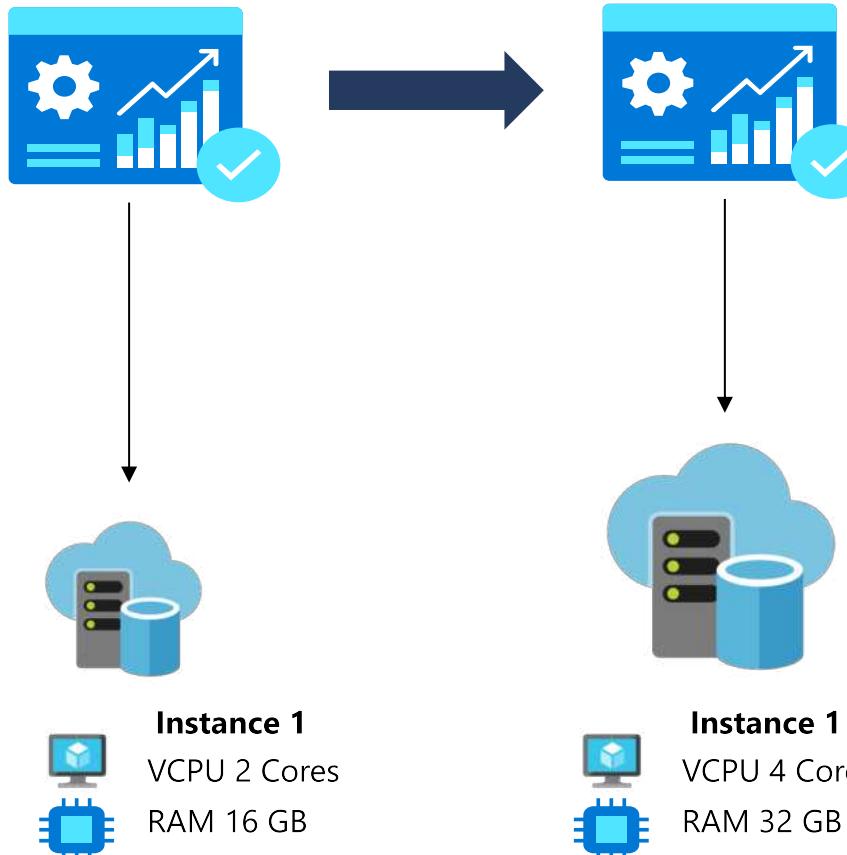
Composite SLAs = combine SLAs & see the possibility of failures



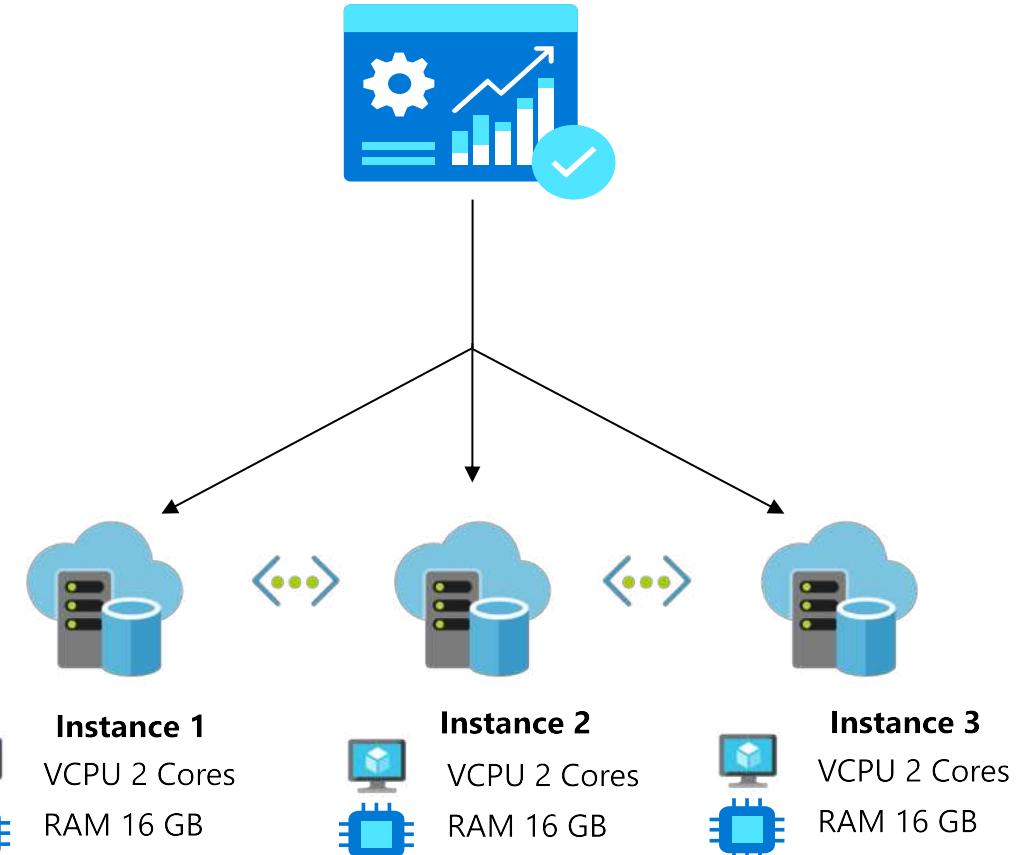
\*\* SLA sometimes mentioned as Service Level Objective (SLO) or Service Level Indicator (SLI)

# Elasticity (Auto) & Scalability (Not-auto)

Scale up (vertical)

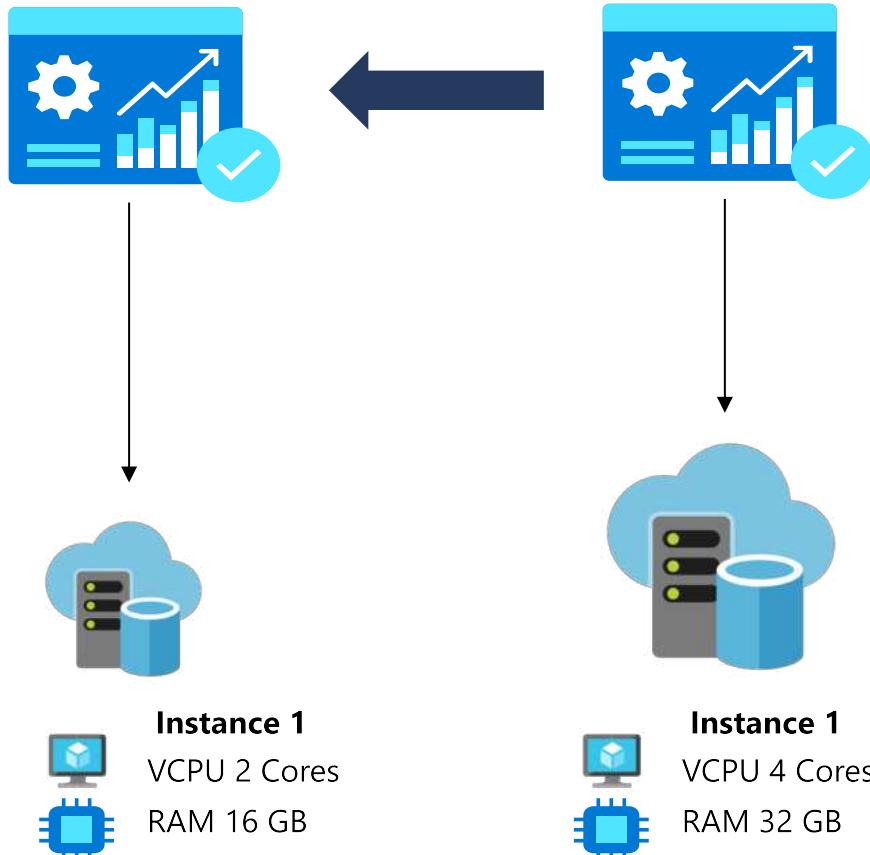


Scale out (horizontal)

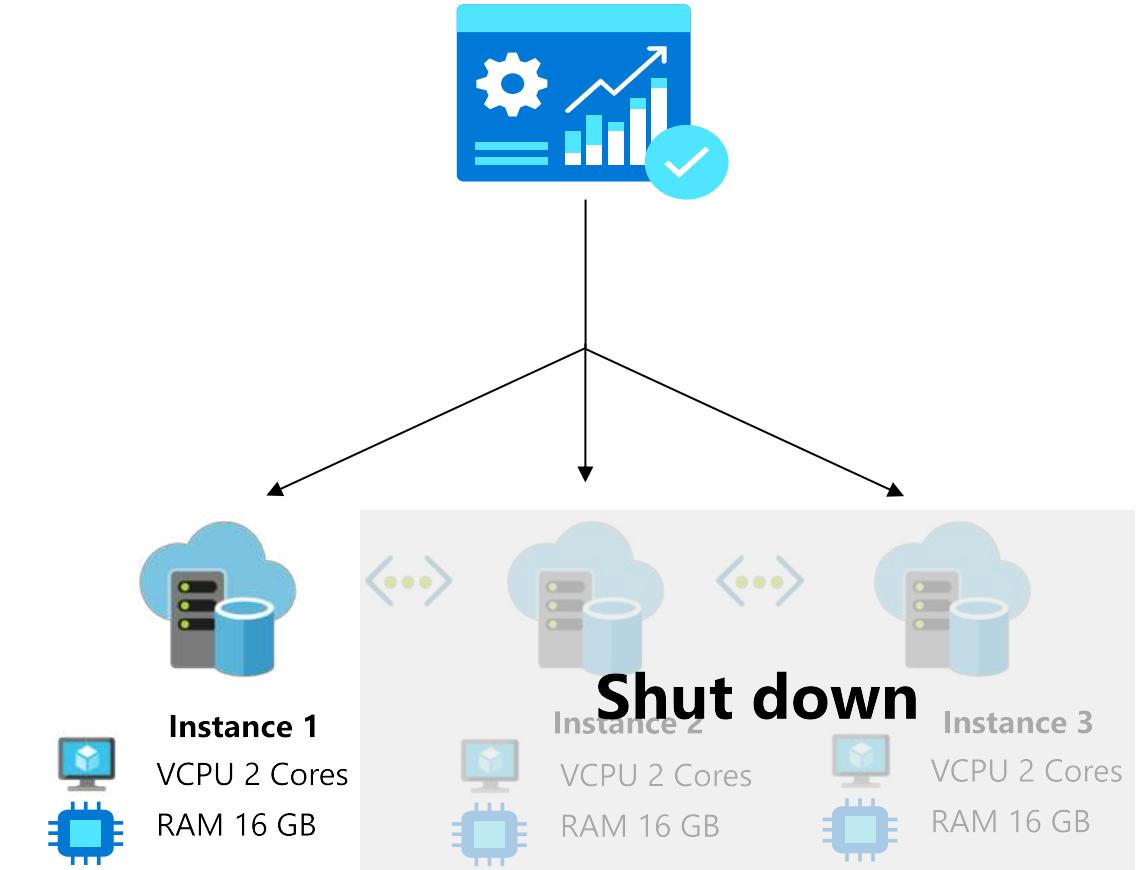


# Elasticity (Auto) & Scalability (Not-auto)

Scale down (vertical)

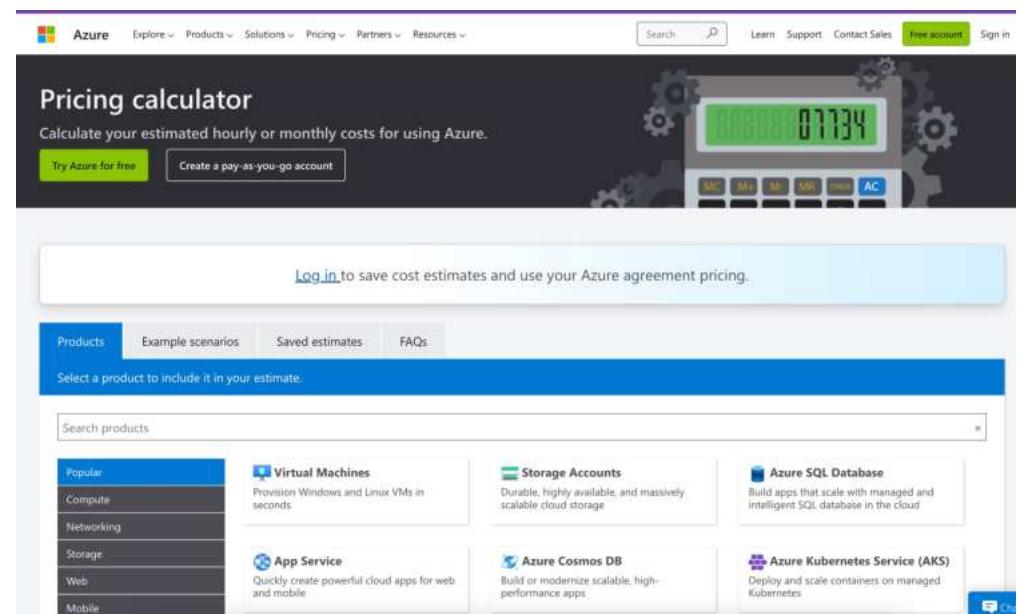
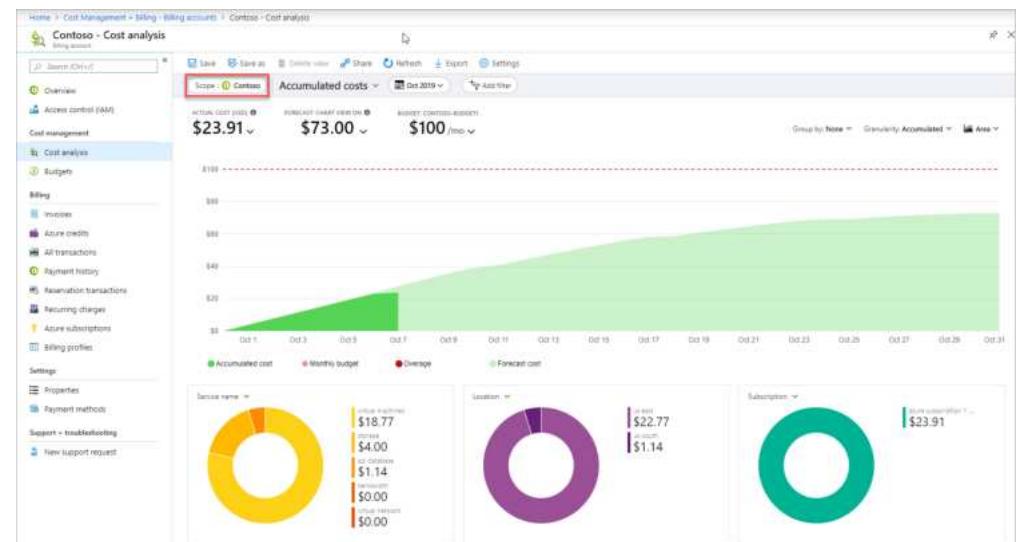
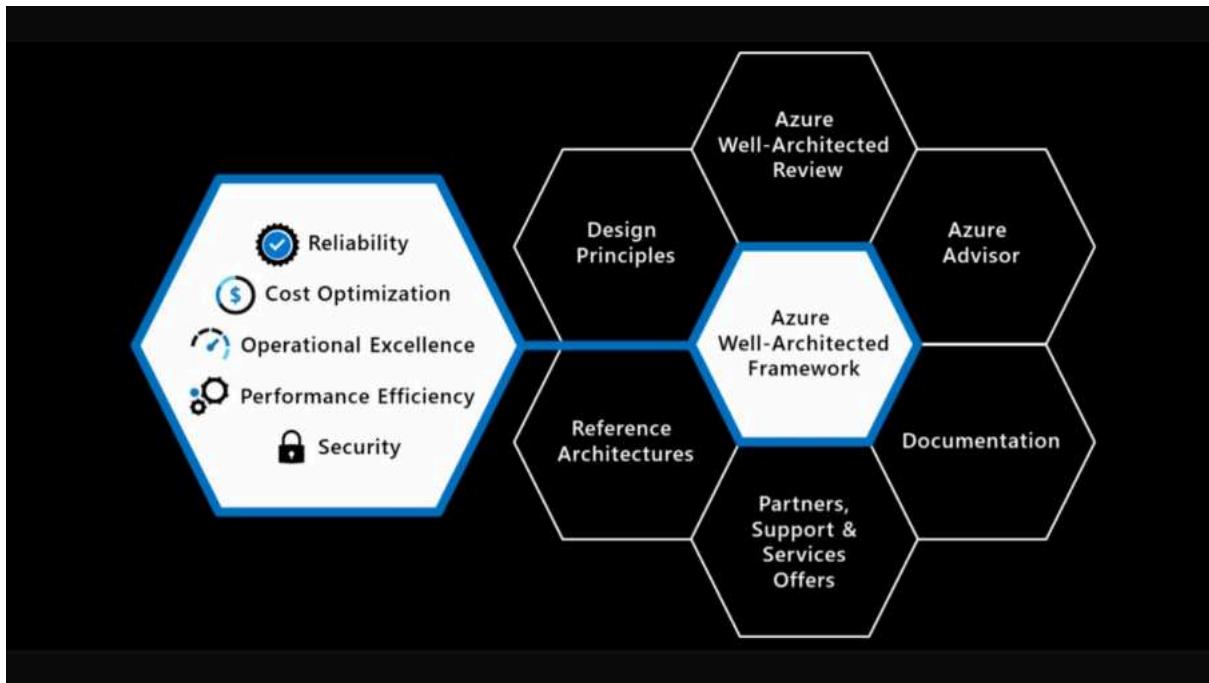


Scale in (horizontal)



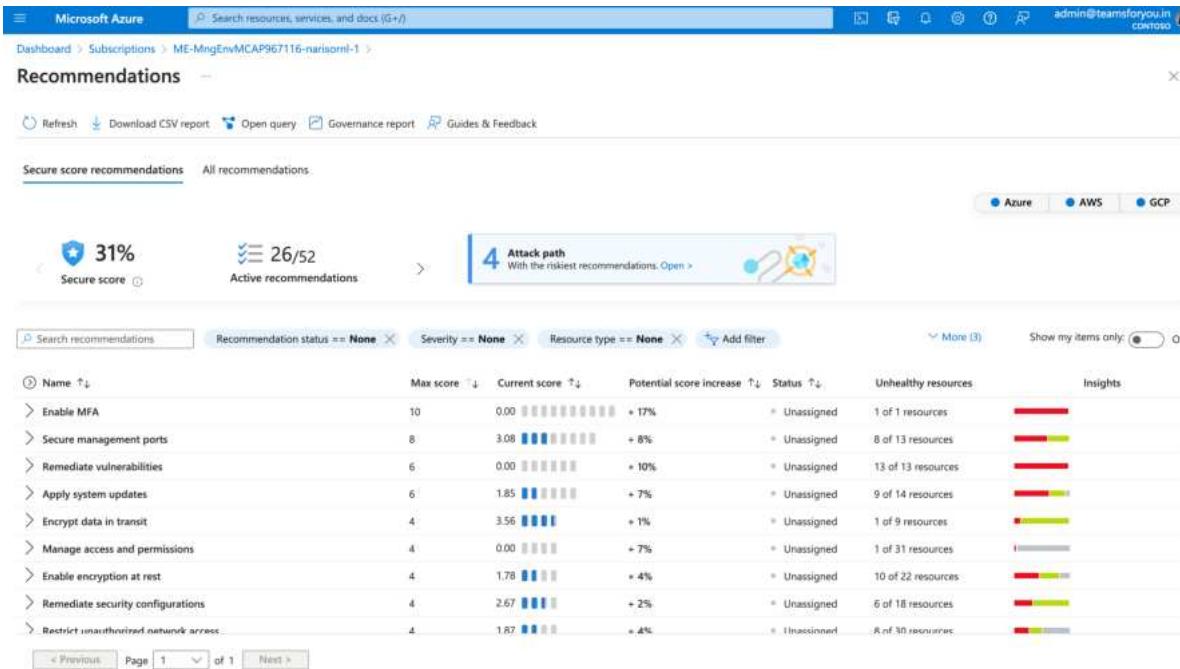
# Reliability & Predictability

- Well-architected – recovered from failure
- Performance & Cost predictability



# Security

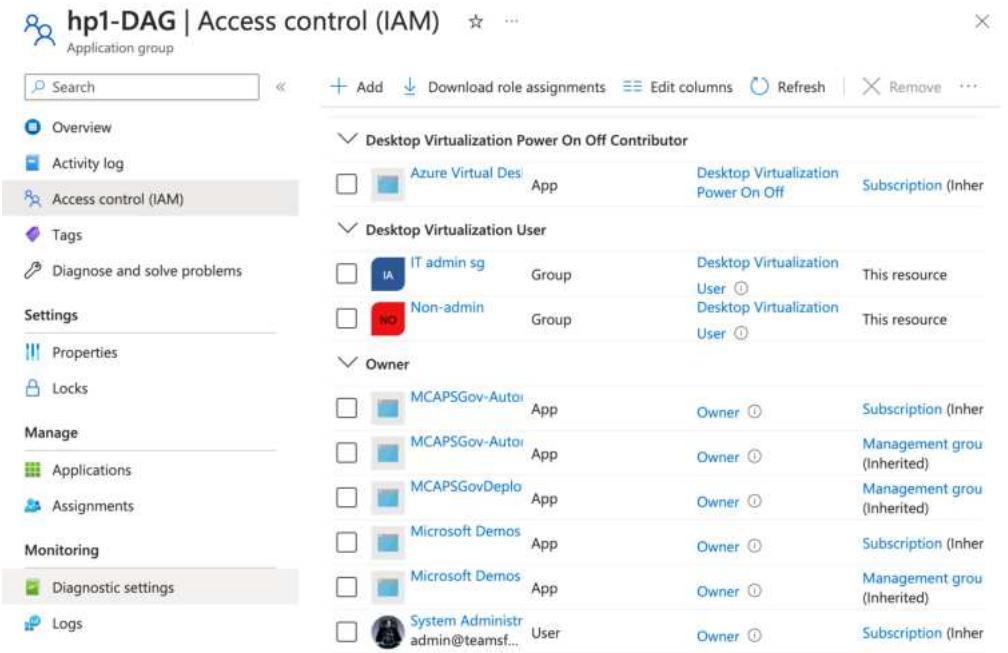
- Security templates and Score
- Role-based Access Control (RBAC)
- Modern Authentication



The screenshot shows the Microsoft Azure Secure Score & Recommendation dashboard. At the top, it displays a secure score of 31% and 26 active recommendations. A search bar and filter options are available. Below, a section titled "Attack path" lists four recommendations with the highest impact. The main area shows a table of recommendations, sorted by name, with columns for Max score, Current score, Potential score increase, Status, and Unhealthy resources. The status column uses color-coded icons to indicate the progress of each recommendation.

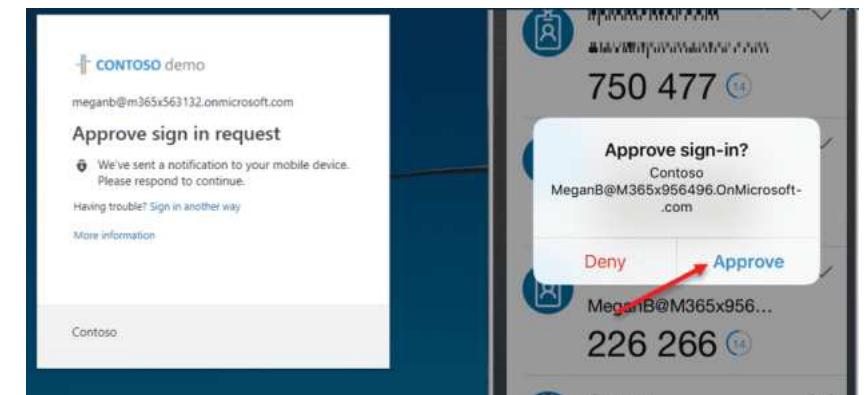
Name	Max score	Current score	Potential score increase	Status	Unhealthy resources
Enable MFA	10	0.00	+ 17%	Unassigned	1 of 1 resources
Secure management ports	8	3.08	+ 8%	Unassigned	8 of 13 resources
Remediate vulnerabilities	6	0.00	+ 10%	Unassigned	13 of 13 resources
Apply system updates	6	1.85	+ 7%	Unassigned	9 of 14 resources
Encrypt data in transit	4	3.56	+ 1%	Unassigned	1 of 9 resources
Manage access and permissions	4	0.00	+ 7%	Unassigned	1 of 31 resources
Enable encryption at rest	4	1.78	+ 4%	Unassigned	10 of 22 resources
Remediate security configurations	4	2.67	+ 2%	Unassigned	6 of 18 resources
Restrict unauthorized network access	4	1.87	+ 4%	Unassigned	8 of 10 resources

Secure Score & Recommendation



The screenshot shows the Microsoft Azure Access control (IAM) interface for the application group "hp1-DAG". The left sidebar includes links for Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Settings, Properties, Locks, Manage, Applications, Assignments, Monitoring, Diagnostic settings, and Logs. The main pane displays a list of role assignments. It shows various roles like "Desktop Virtualization Power On Off Contributor", "Azure Virtual Des... App", "IT admin sg Group", "Non-admin Group", and "Owner". Each entry includes details such as the role name, type (App or Group), owner, and inheritance information.

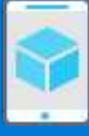
Role-based Access Control (RBAC)



The screenshot shows a Microsoft 365 sign-in request approval interface. It displays a message from "CONTOSO demo" to "meganb@m365x956132.onmicrosoft.com" asking for approval. The user is prompted to respond to a notification sent to their mobile device. A red arrow points to the "Approve" button in the approval dialog box.

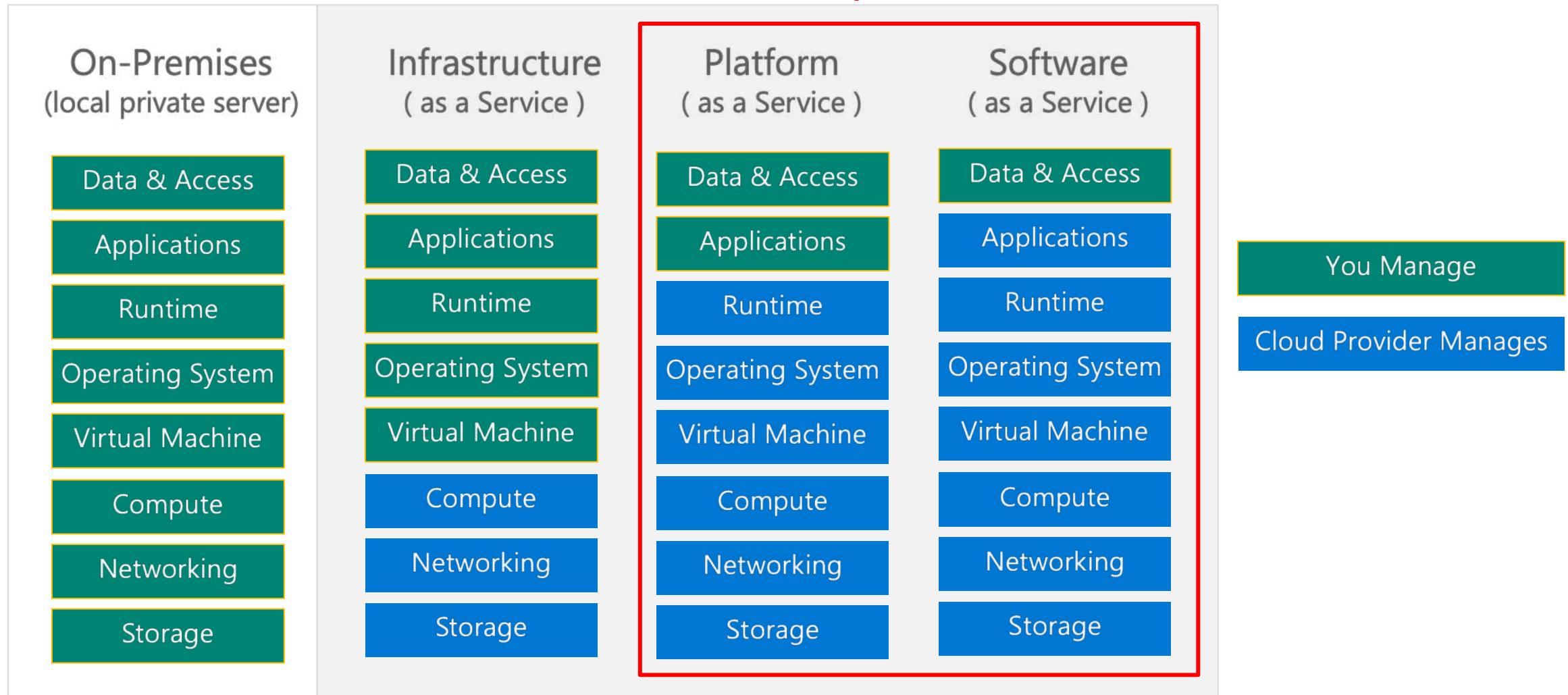
Multi-factor Authentication (MFA)

# Manageability – Cloud resource management console

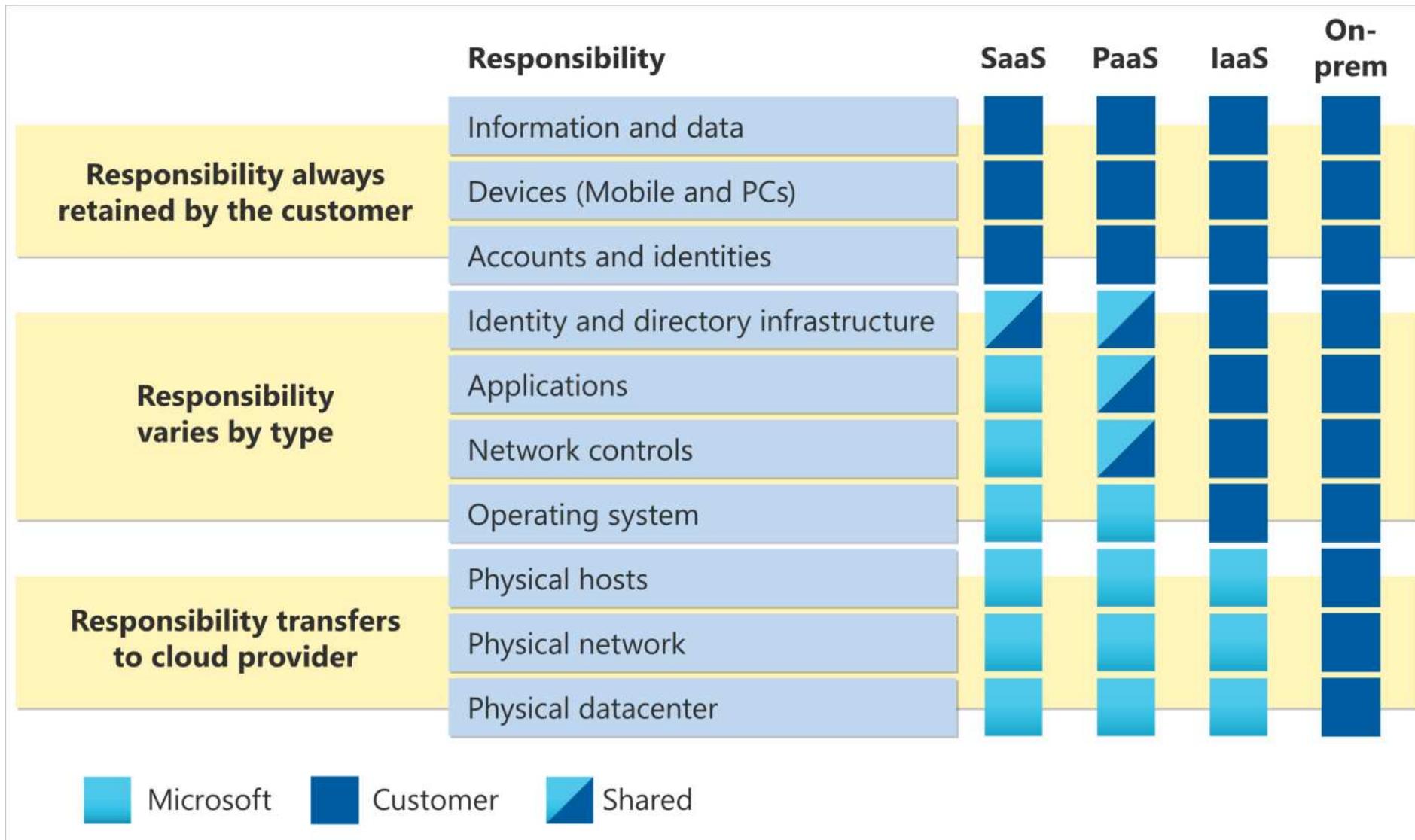
	Azure portal
	Azure PowerShell and Azure Command-Line Interface (CLI)
	Azure Cloud Shell
	Azure mobile app
	Azure REST API

# Discuss shared responsibility model

Today's focused



# Shared responsibility model



# Cloud service comparison

## IaaS

The most flexible cloud service.

You configure and manage the hardware for your application.

## PaaS

Focus on application development.

Platform management is handled by the cloud provider.

## SaaS

Pay-as-you-go pricing model.

Users pay for the software they use on a subscription model.



# Introducing Microsoft Azure

- Cloud Computing Services developed by Microsoft since 2008
- Previous known as Windows Azure and Project Red dog
- IaaS, PaaS and SaaS Offerings
- Seamless integration with other Microsoft Cloud (Office 365 & Dynamics 365)
- Over 200+ services offerings such as AI, Apps, Data Warehouse, Data Lake, VM, VDI, Maps, Remote Rendering, DevOps, Security etc. (Native & 3rd party)
- 90+ Compliance offerings
- 95% of fortune 500 customers
- \$1 billion investment per year for security
- Most Data Center coverage



Select a category:

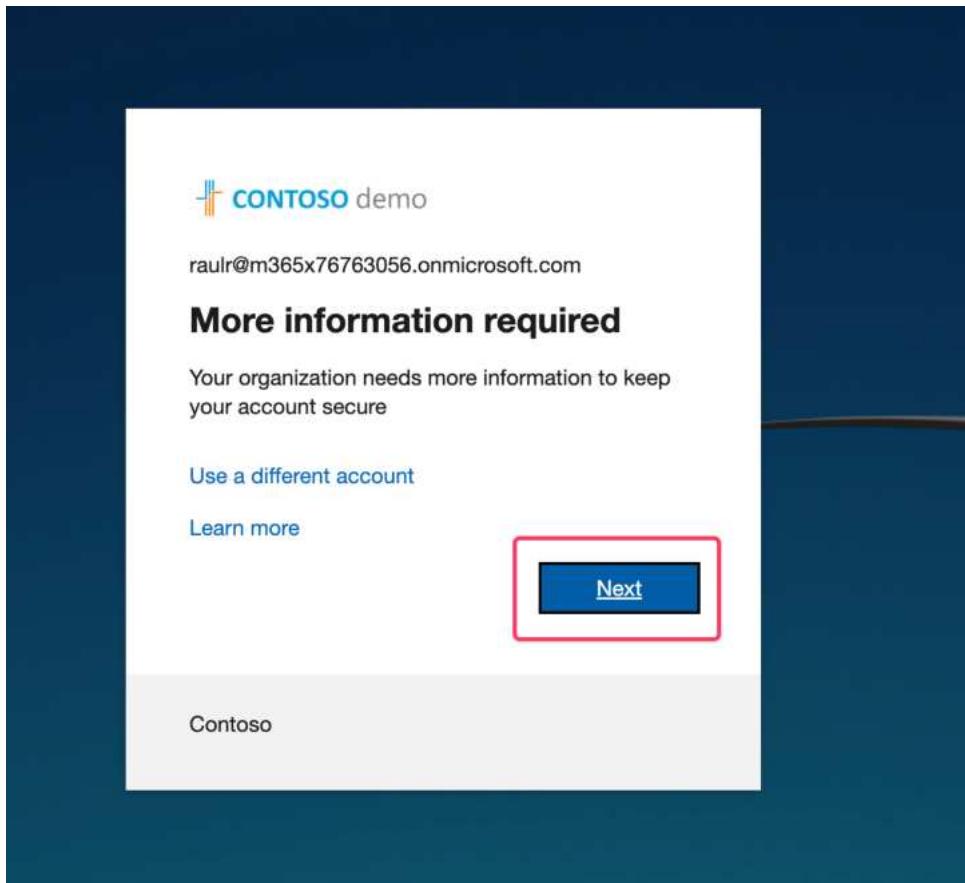
- [AI + machine learning](#)
- [Analytics](#)
- [Compute](#)
- [Containers](#)
- [Databases](#)
- [Developer tools](#)
- [DevOps](#)
- [Hybrid + multicloud](#)
- [Identity](#)
- [Integration](#)
- [Internet of Things](#)
- [Management and governance](#)
- [Media](#)
- [Migration](#)
- [Mixed reality](#)
- [Mobile](#)
- [Networking](#)
- [Security](#)
- [Storage](#)
- [Virtual desktop infrastructure](#)
- [Web](#)

<https://tinyurl.com/kulabusers>

# Lab 1: Create SQL Database from Azure Cloud

<https://learn.microsoft.com/en-us/azure/azure-sql/database/single-database-create-quickstart?view=azuresql&tabs=azure-portal#create-a-single-database>

# You may encounter this (some lab users) ...just click



Keep your account secure

### Phone

You can prove who you are by receiving a code on your phone.

What phone number would you like to use?

United States (+1)  Enter phone number

Receive a code

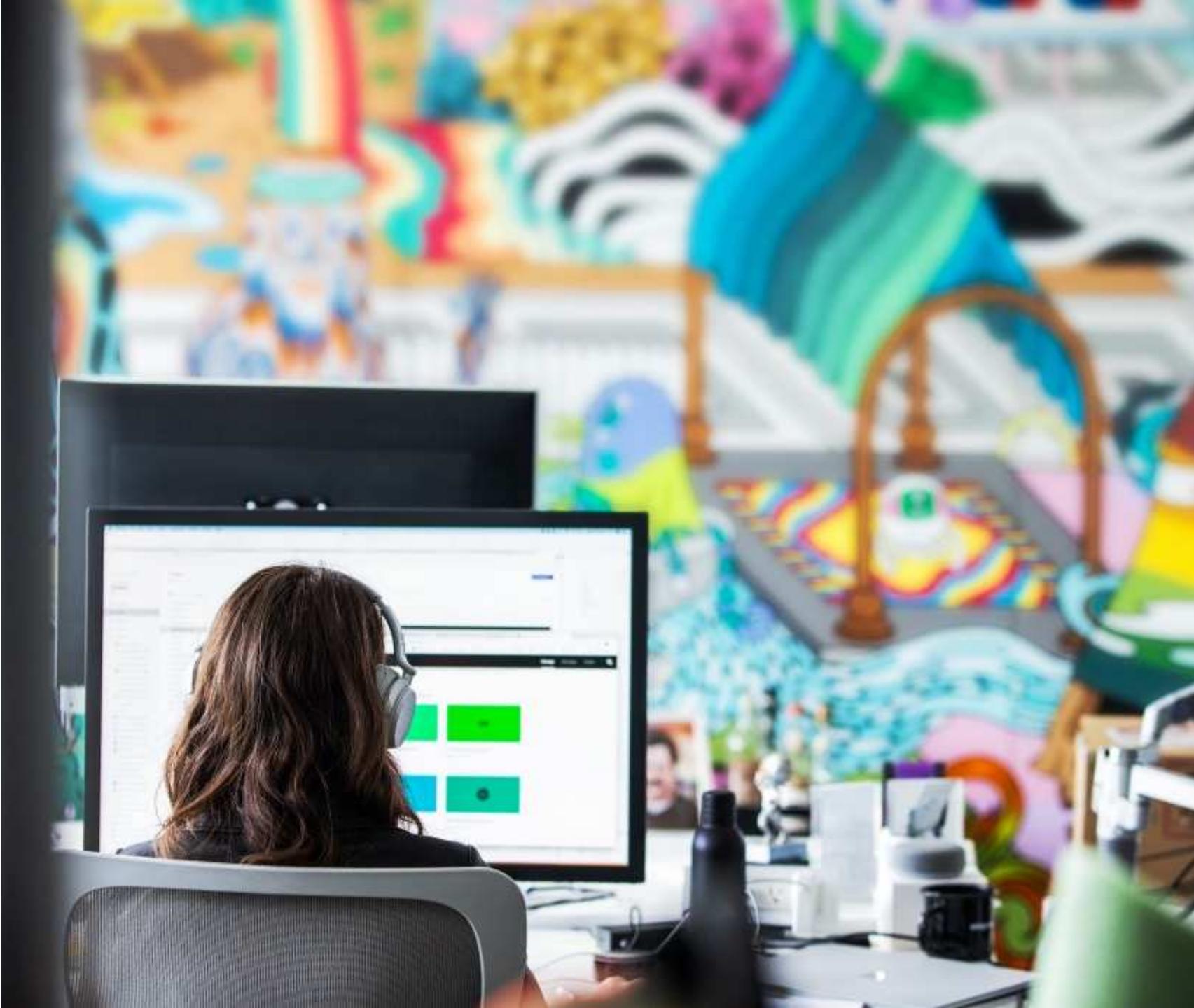
Message and data rates may apply. Choosing Next means that you agree to the [Terms of service](#) and [Privacy and cookies statement](#).

[I want to set up a different method](#)

# Data Fundamentals

Business Applications Domain

Narisorn Limpaswadpaisarn  
Client Technology Lead



# ประเภทของ Data แบ่งตามในแบบนี้

1. รูปแบบลักษณะของตัวมัน - Structure vs Semi-structured vs Unstructured
2. การเก็บข้อมูล (Data stores) – Database vs File Storage
3. Data stores แบบไหนเหมาะสมกับงานแบบไหน? - Transactional vs Analytical
4. รูปแบบการให้ผลและทํากระบวนการ (Data Processing) - Batch vs Streaming

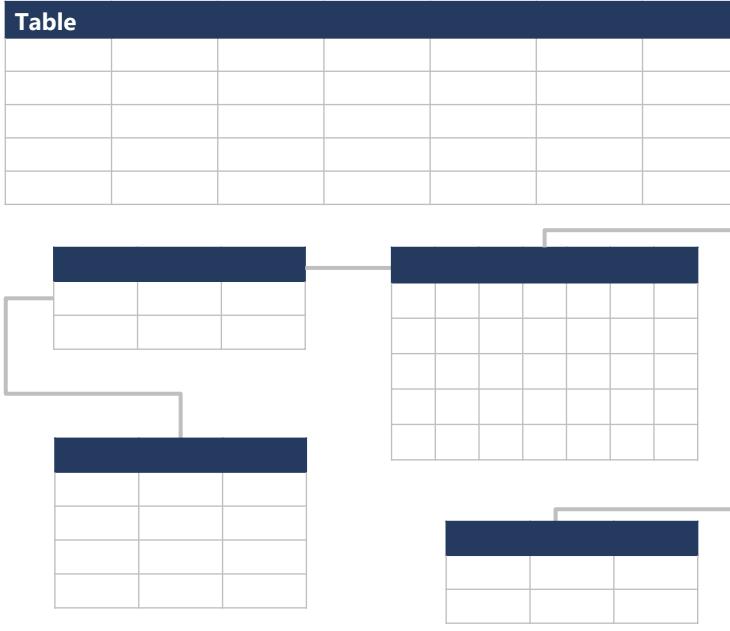
\* Data Ingestion คือ วิธีการได้ Data มา เช่น copy หรือ import นำมา process เลยหรือเก็บไว้ใน data store เพื่อใช้ในคราวหลัง

\*Unstructured บางครั้งก็ถูกเรียกเป็น free-text

# 1. ประเภทของ Data มี 3 รูปแบบลักษณะ

Collection of facts, numbers, descriptions, objects , stored in a structured, semi-structured, unstructured way

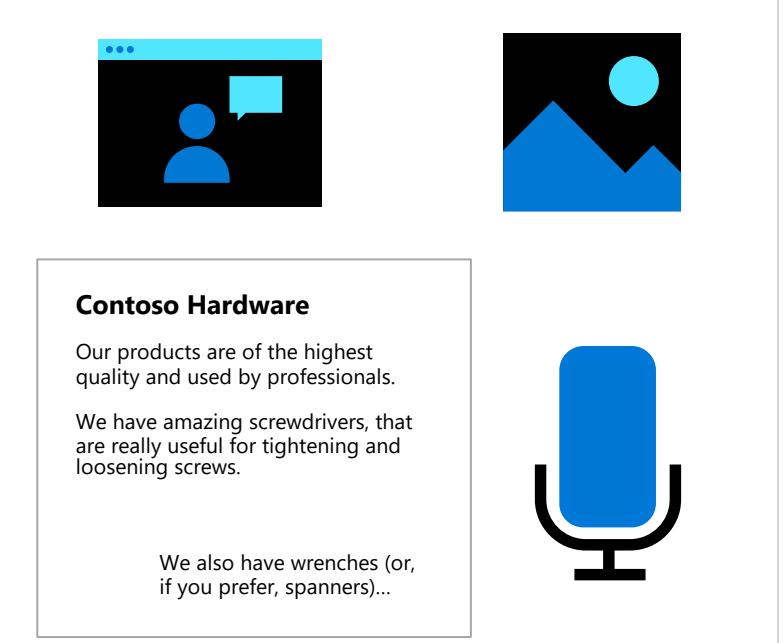
## 1) Structured



## 2) Semi-structured

```
{  
    "firstName": "Joe",  
    "lastName": "Jones",  
    "address":  
    {  
        "streetAddress": "1 Main  
St.",  
        "city": "New York",  
        "state": "NY",  
        "postalCode": "10099"  
    },  
    "contact":  
    [  
        {  
            "type": "home",  
            "number": "555 123-1234"  
        },  
        {  
            "type": "email",  
            "address":  
            "joe@litware.com"  
        }  
    ]  
},  
{  
    "firstName": "Samir",  
    "lastName": "Nadov",  
    "address":  
    {  
        "streetAddress": "123  
Elm Pl.",  
        "unit": "500",  
        "city": "Seattle",  
        "state": "WA",  
        "postalCode": "98999"  
    },  
    "contact":  
    [  
        {  
            "type": "email",  
            "address":  
            "samir@northwind.com"  
        }  
    ]  
}
```

## 3) Unstructured



ตารางที่มี Relation มีการ fix ตัว column

คล้ายๆตารางแต่ไม่ fix ตัว column แต่พอจะคือเป็นรูปแบบตารางได้  
Column ในแต่ละข้อมูล จะมี tag หรือ key ในการจัดแบ่ง  
File เช่น JSON, AVRO, ORC, HTML หรือ Parquet

ข้อมูลหลากหลาย เช่น ประเภท document, log, email, รูปภาพ, เสียง ที่เป็น Binary นำเนื้อ content มาทำในรูปแบบตารางไม่ได้รูปแบบไม่ตายตัว

## 2. ลักษณะที่เก็บ Data (Data Stores)

เราจะจัดเก็บข้อมูล Structured, Semi-structured หรือ Unstructured ไว้ใน Data Stores ซึ่งขอแบ่งเป็น 2 แบบ

### 1. Database – แบ่งเป็นทั้ง

- **Relational Database (RDBMS)** ที่ใช้กันทั่วไป เช่น SQL Server, MySQL, PostgreSQL หรือ Oracle ซึ่งเหล่านี้เป็น Structured
- **Non-relational Database (หรือเรียกว่า NoSQL)** ที่ใช้กันทั่วไป เช่น MongoDB หรือ Cassandra ซึ่งเหล่านี้เป็น Semi-structured

### 2. File storage – image, CSV, JSON, XML, BLOB -> (เก็บไว้ใน File sharing, SharePoint หรือ Data Lake)

\*\*\*\* อธิบายเพิ่มเติม: ตัว NoSQL ก็จะมีแยกชนิดข้างในอีกด้วยว่าเก็บไว้ในลักษณะย่อยแบบไหน

- 1) Key-value
- 2) Document
- 3) Column family
- 4) Graph

Products	
Key	Value
123	"Hammer (\$2.99)"
162	"Screwdriver (\$3.49)"
201	"Wrench (\$4.25)"

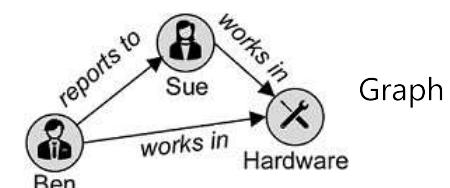
Key-value

Customers	
Key	Document
1	{ "name": "Joe Jones", "email": "joe@litware.com" }
2	{ "name": "Samir Nadoy", "email": "Samir@northwind.com" }

Document

Orders				
Key	Customer		Product	
	Name	Address	Name	Price
1000	Joe Jones	1 Main St.	Hammer	2.99
1001	Samir Nadoy	123 Elm Pl.	Wrench	4.25

Column family



Graph

# 2. ลักษณะที่เก็บ Data (Data Stores)

## Files

### Delimited Text

```
FirstName,LastName,Email  
Joe,Jones,joe@litware.com  
Samir,Nadoy,samir@northwind.com
```

### JavaScript Object Notation (JSON)

```
{  
  "customers":  
  [  
    { "firstName": "Joe", "lastName": "Jones"},  
    { "firstName": "Samir", "lastName": "Nadoy"}  
  ]  
}
```

### Extensible Markup Language (XML)

```
<Customer firstName="Joe" lastName="Jones"/>
```

### Binary Large Object (BLOB)

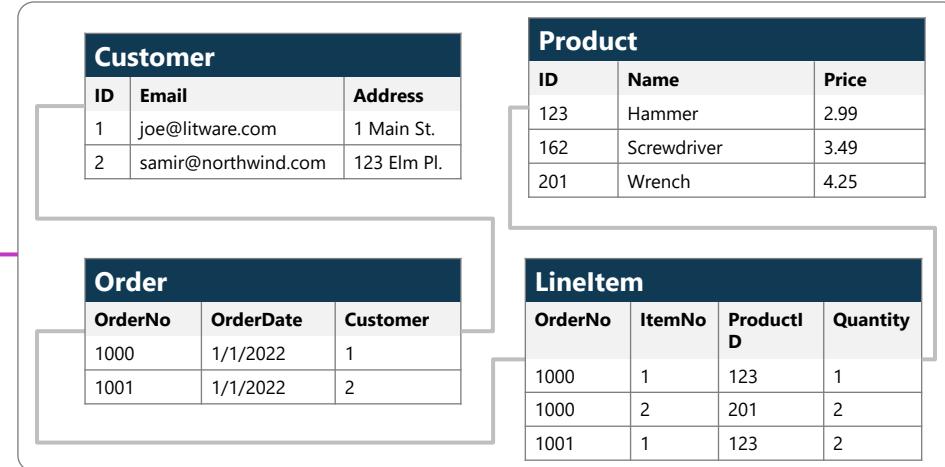
```
10110101101010110010...
```

### Optimized formats:

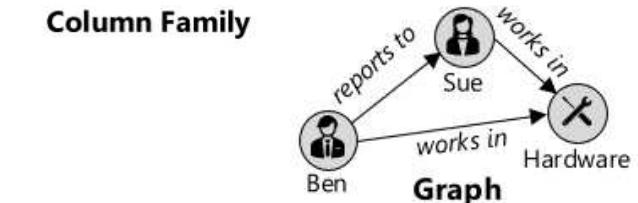
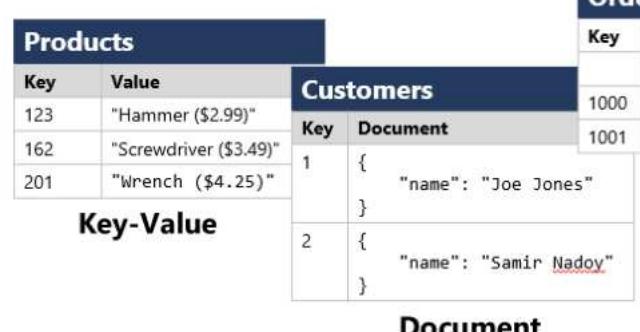
- Avro, ORC, Parquet

## Databases

### Relational



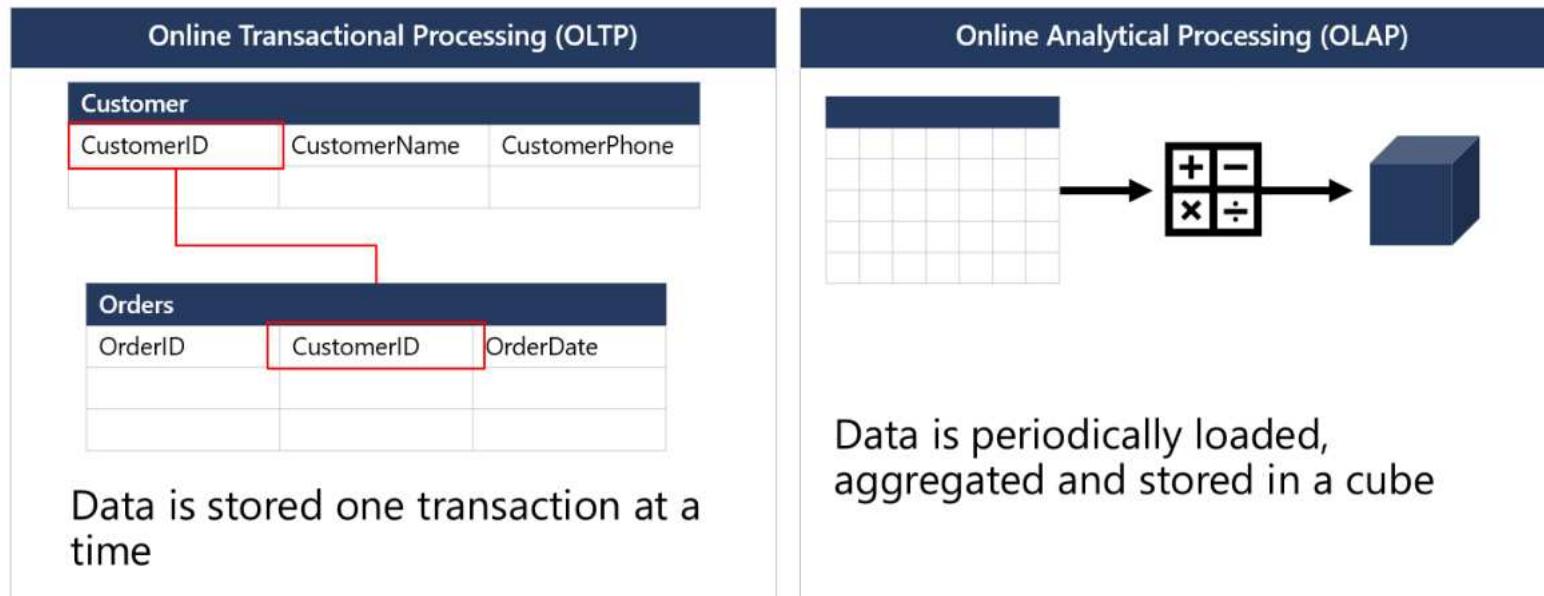
### Non-Relational



### 3. Data Stores – Transactional vs Analytical

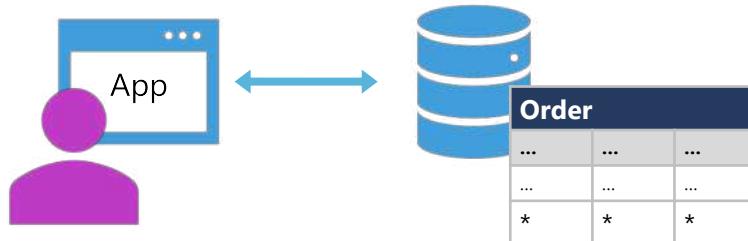
ใน Database โดยเฉพาะ Relational Database ก็จะมีการเก็บที่จะ optimized ได้ 2 นูน

- 1) **Transactional (OLTP)** – ก็หมายความว่า application มาต่อเพื่อบันทึกหรือเปลี่ยนแปลงข้อมูลในนูน **operation** เช่น สั่ง order สินค้า ซึ่งก็คือ **write date** ไปเก็บข้อมูลการสั่งลงไปที่ Database ที่ต่อกับ application
- 2) **Analytical (OLAP)** – เป็นข้อมูลที่หาผลรวม (Aggregation) เพื่อให้เหมาะสมในการเอาไปทำ **report** ที่การตอบคำถามแบบ drill-down ได้ เช่นหาข้อมูลสรุปยอดขายทั้งหมด โดย **Department**, **Product** หรือ **Location** ตามช่วงเวลาของเดือนต่างๆ แบบนี้การ query ก็มักจะใช้เรียกว่า **Data Warehouse** ซึ่งเป็น relational database ที่เหมาะสมกับการ **query** ด้าน **read data**



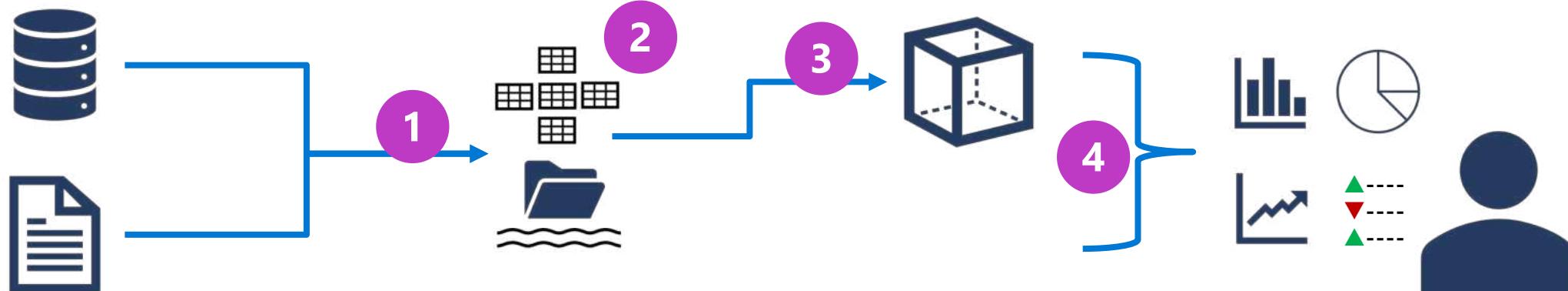
# 3. Data Stores – Transactional vs Analytical

## Transactional

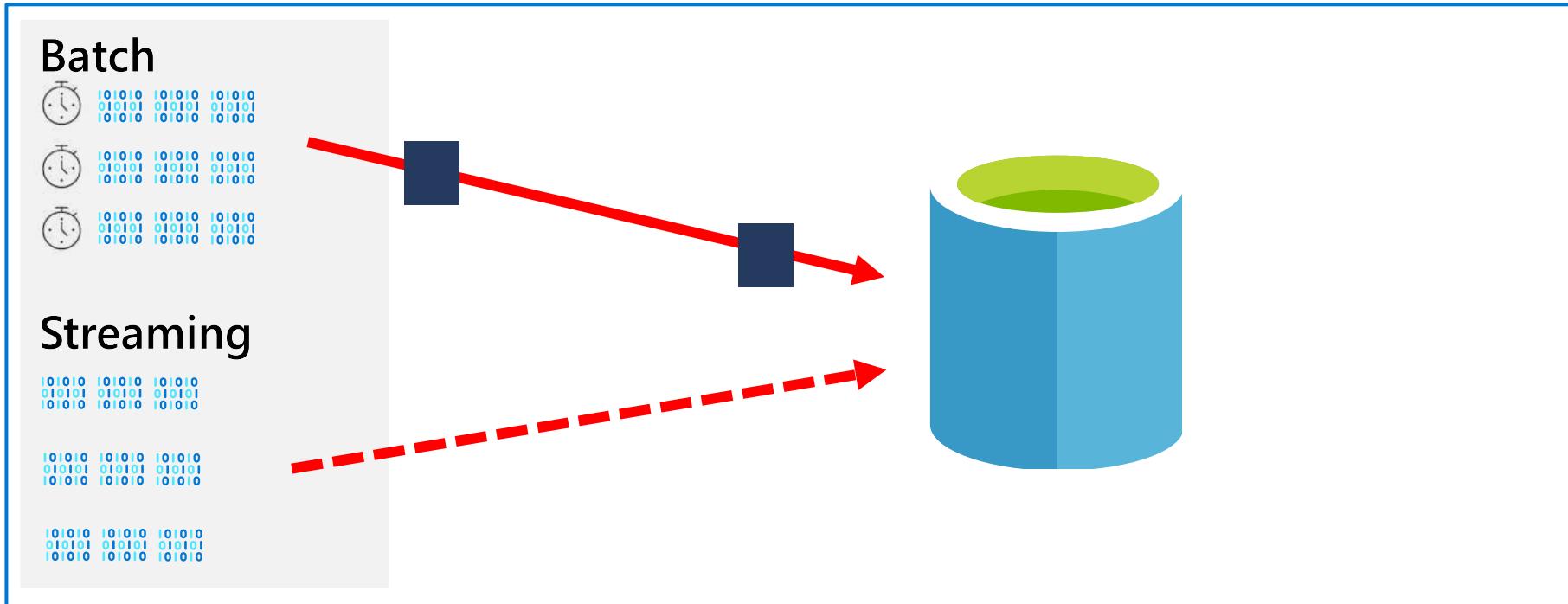


- 1 Operational data is extracted, transformed, and loaded (ETL) into a *data lake* for analysis
- 2 Data is loaded into a schema of tables - typically in a Spark-based *data lakehouse* with tabular abstractions over files in the data lake, or a data warehouse with a fully relational SQL engine
- 3 Data in tables may be aggregated and loaded into an online analytical processing (OLAP) model, or cube
- 4 The files in the data lake, relational tables, and analytical model can be queried to produce *reports* and *dashboards*

## Analytical



## 4. ประเภทการไหลของข้อมูล (Data Processing) – Batch vs Streaming



1) **Batch (Cold path)** ข้อมูลจะเก็บเป็น group ในแต่ละช่วงเวลา หรือตาม amount (ที่คนสร้างระบบกำหนด) เอาไว้ process ทีหลัง เช่น โหลดข้อมูลตารางจาก Excel, ข้อมูล Payroll หรือ Orders ใน database ของเดือนที่จะมาวิเคราะห์ โดยสามารถวิเคราะห์ข้อมูลที่ซับซ้อนได้

2) **Streaming (Hot path)** ก็คือข้อมูลจะไหลเข้ามาเรื่อยๆ ออกแบบ (near) real time เป็นหน่วย sec/millisecond. แล้ว process ได้เลย (การวิเคราะห์ไม่ควรซับซ้อนมาก เช่นหา ผนรวม หรือ rolling averages) เช่น ข้อมูลจาก Sensors, หุ้นเทรด, Social media feed หรือ transaction ที่จะจับ fraud detection

# ลักษณะ Relational (แบบ Structured)

Customers		
CustomerID	CustomerName	CustomerPhone
100	Muisto Linna	XXX-XXX-XXXX
101	Noam Maoz	XXX-XXX-XXXX
102	Vanja Matkovic	XXX-XXX-XXXX
103	Qamar Mounir	XXX-XXX-XXXX
104	Zhenis Omar	XXX-XXX-XXXX
105	Claude Paulet	XXX-XXX-XXXX
106	Alex Pettersen	XXX-XXX-XXXX
107	Francis Ribeiro	XXX-XXX-XXXX

Data is stored in a table

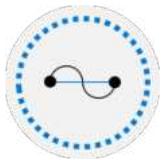
Table consists of rows and columns

All rows have same # of columns

Each column is defined by a datatype

และก็สามารถมี Relation เชื่อมไปที่ Table อื่นๆ เช่น Customer กับ Order เชื่อมกันที่ column เซ็นต์ CustomerID ที่ในมีทั้ง 2 Tables

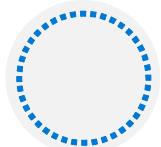
# Use cases ของ Relational database



## IoT:

Although typically considered for non-relational, the data from IoT devices could be structured and consistent

---



## Online transaction processing:

For example, order systems that perform many small transactional updates

---



## Data warehousing:

Large amounts of data can be imported from multiple sources and structured to enable high-performance queries

IoT บางกรณีเป็น non-relational บางกรณี Relational

# ลักษณะของ non-relational data (แบบ semi-structured)

## Entities

ยังพอเห็นโครงสร้างแต่รูปแบบไม่ได้ fix ตายตัวแบบ 100%

```
## Customer 1 ID: 1
Name: Mark Hanson
Telephone: [ Home: 1-999-9999999, Business: 1-888-8888888, Cell: 1-777- 7777777 ]
Address: [ Home: 121 Main Street, Some City, NY, 10110,
           Business: 87 Big Building, Some City, NY, 10111 ]
## Customer 2 ID: 2
Title: Mr
Name: Jeff Hay
Telephone: [ Home: 0044-1999-333333, Mobile: 0044-17545-444444 ]
Address: [ UK: 86 High Street, Some Town, A County, GL8888, UK,
           US: 777 7th Street, Another City, CA, 90111 ]
```

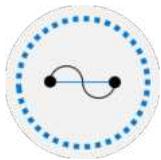
## Non-relational collections can have:

Multiple entities in the same collection or container with different fields

Have a different, non-tabular schema

Are often defined by labeling each field with the name it represents

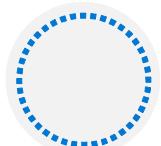
# Use cases ของ non-relational database (NoSQL)



## IoT and Telematics:

Often require to ingest large amounts of data in frequent burst of activity, data is either semi structured or structured, often requires real time processing

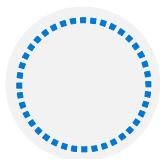
---



## Retail and Marketing:

Common scenarios for globally distributed data, document storage

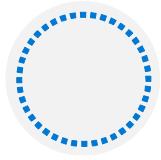
---



## Gaming:

In-game stats, social media integration, leaderboards, low-latency applications

---



## Web and Mobile:

Commonly used with web click analytics, modern applications including bots

---

\*\* IoT บางกรณีเป็น non-relational บางกรณี Relational

# ขั้นตอนการสกัดข้อมูล Extract Transform Load (ETL)

Data sources ต้นทาง เช่น  
files, OLTP databases



**On-premises data**  
SQL Server, Oracle,  
fileshares, SAP



**Cloud data**  
Azure, AWS, GCP



**SaaS data**  
Salesforce, Dynamics

Ingest &  
Clean data  
(Transform)

**Data ingestion**



Load เก็บข้อมูลใน  
Data Warehouse  
หรือ Data Lake

**Data storage**



Aggregate ข้อมูล  
และ Load ลง OLAP  
model เพิ่มเติม

**Data processing**



Report  
Visualization



เนื่องจากข้อมูลอาจมาจาก  
หลายที่มาจากระบบต่างๆ  
อาจจะมี column ที่ไม่ได้  
ใช้, format ไม่ได้, ต้องใช้  
การ lookup หรือ ต้อง  
join table ระหว่างระบบ  
ทำให้ต้องทำการ clean  
เสียก่อน

ETL จะได้โครงสร้างเหมาๆ กับการ read ทำ report  
เพาะ Data มีการหาผลรวม (Aggregation)

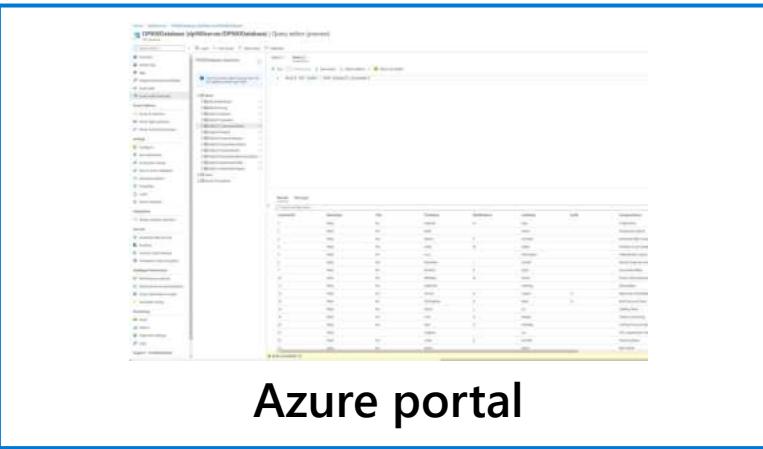
\*ระบบสมัยใหม่เนื่องจากมี Compute ที่ดีขึ้น สามารถ load ก่อนแล้ว transform ก่อนแล้ว load กลับไป ELT

# Big Data Analytics

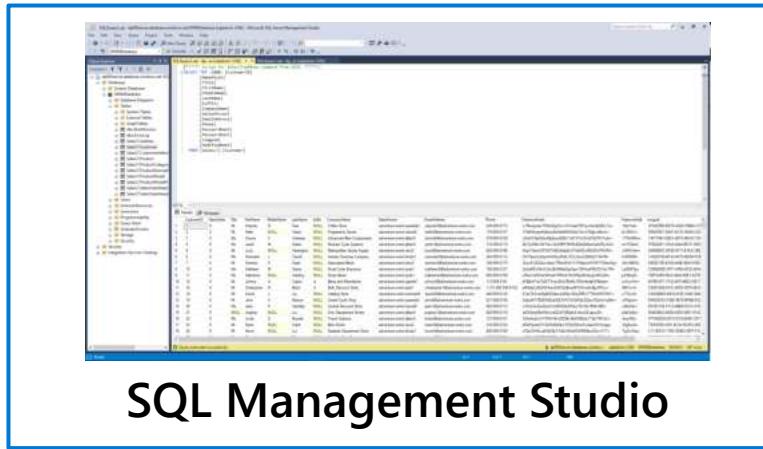
คุณลักษณะของระบบที่จะรองรับการทำ Big Data Analytics

- High volume
- High velocity (streaming)
- High variance (Structured, Semi-Structured or Unstructured)

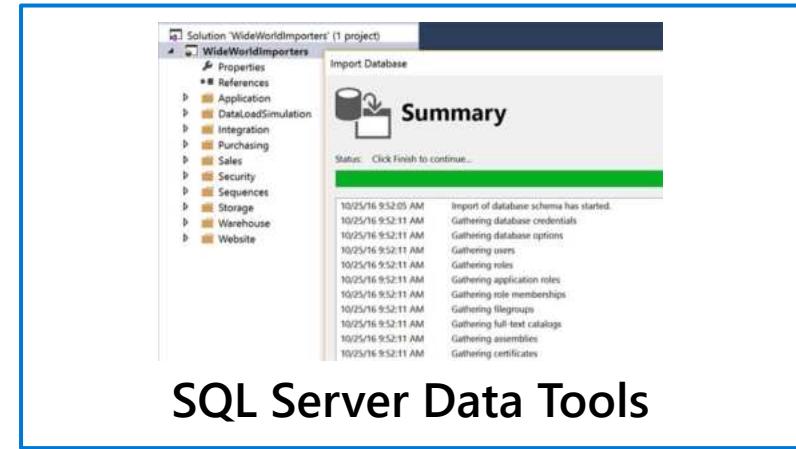
# Query tools



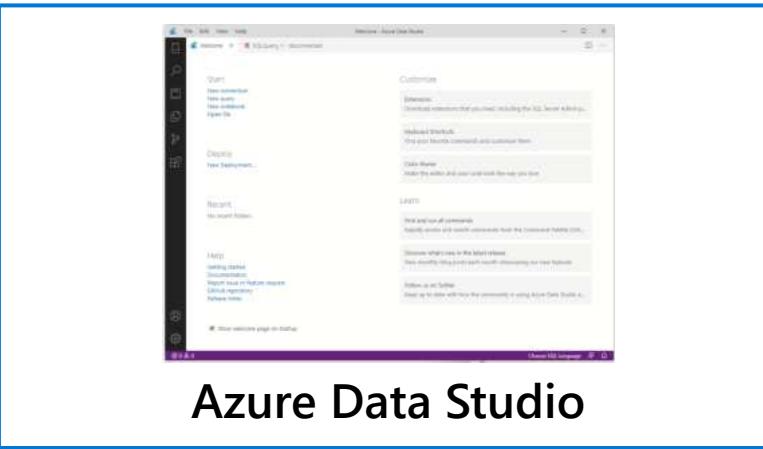
## Azure portal



# SQL Management Studio



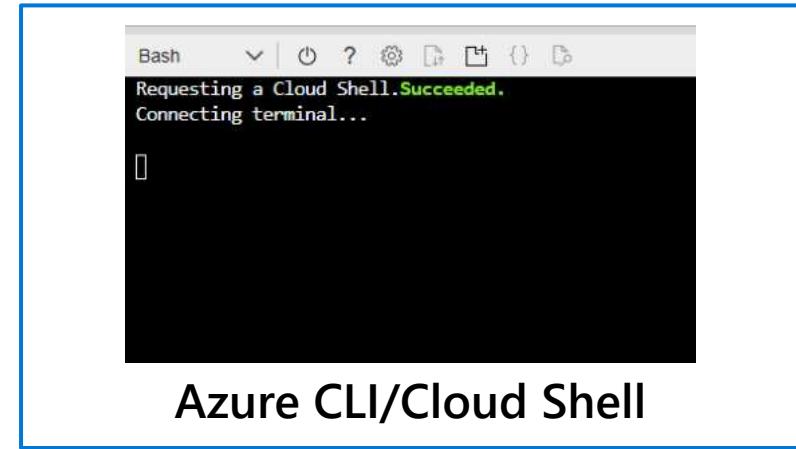
# SQL Server Data Tools



# Azure Data Studio



# SQLCMD

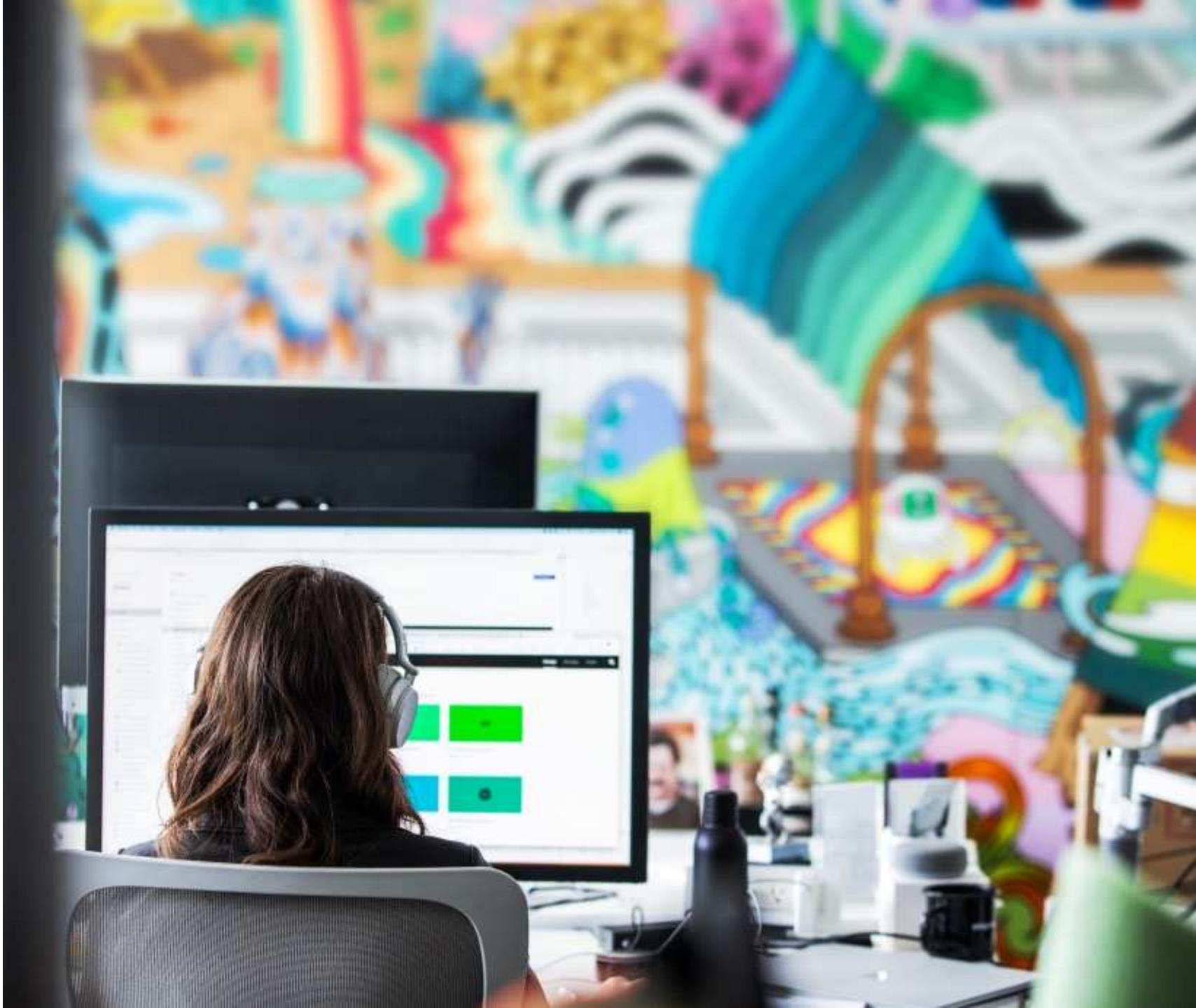


# Azure CLI/Cloud Shell

# Microsoft Power BI (high level)

**Business Applications Domain**

Narisorn Limpaswadpaisarn  
Client Technology Lead



# Power BI for every function



End users



Analysts



BI Professionals



Data Scientists

Natural Language exploration

Explore influencers

Capabilities:

Quick Insights

Q&A

Explorations with Curie



Insights

Out of box AI in PQ

One-click AI transformations

Capabilities:

Sentiment analysis

Key phrase extraction

OCR, Text Translations



Enrichment

Transforms & Visualizations

Custom Python models

Languages:

R

Python

Generate ML models in clicks

Turnkey model deployments

Capabilities:

Predictions, Classifications,  
Forecasting, Clustering,  
Recommendations



Auto ML

PQ integration for Azure ML

Other Azure hosted models

Integration:

Azure ML

Azure Frameworks



Azure ML

# Microsoft business intelligence & analytics

Trusted platform

Powerful self-service

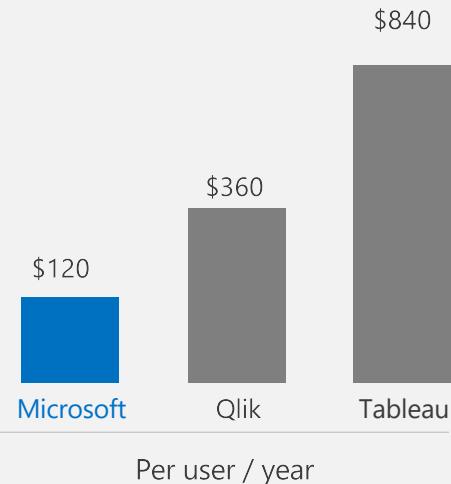
Faster insights

Industry leader



Gartner Magic Quadrant for Business Intelligence & Analytics Platforms, March 2022

At a fraction of the cost



5 seconds to get started,  
5 minutes to WOW!



## The most complete BI solution



The above graphics were published by Gartner, Inc. as part of a larger research document and should be evaluated in the context of the entire document. The Gartner document is available upon request from Microsoft. Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.

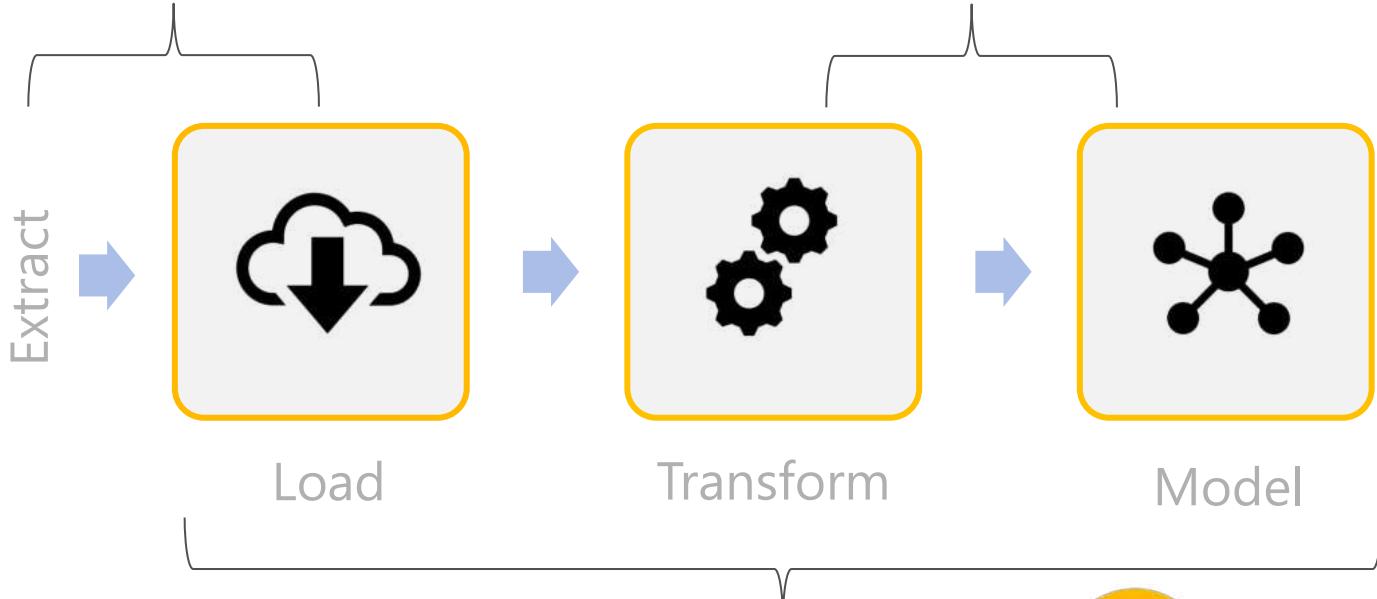
# Power BI - Self Service Development Workflow



Viewers



## Get Data (Power BI)

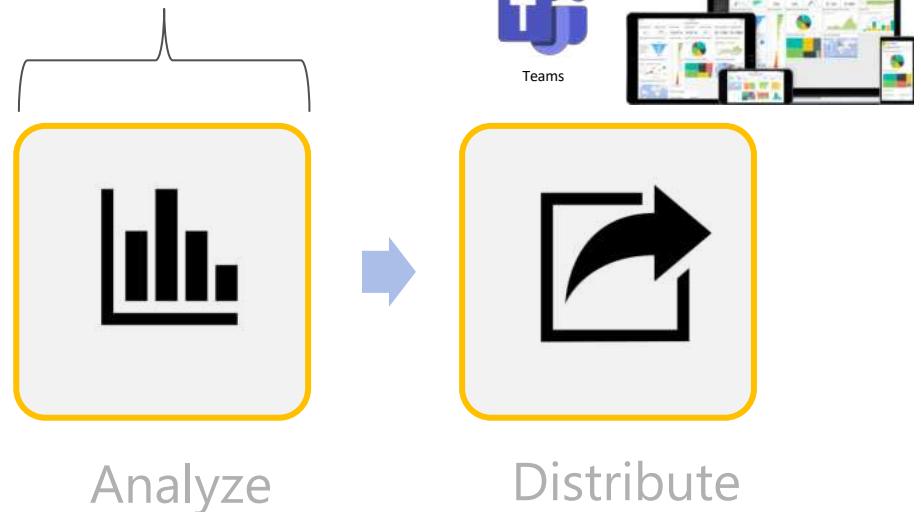


Load

Transform

Model

## Report creation



Analyze

Distribute



Report maker  
(Analyst)

Semantic Model\*

Power BI Desktop



Report maker  
(Analyst)

Power BI Service

(in Microsoft Fabric)



(in Microsoft Fabric)

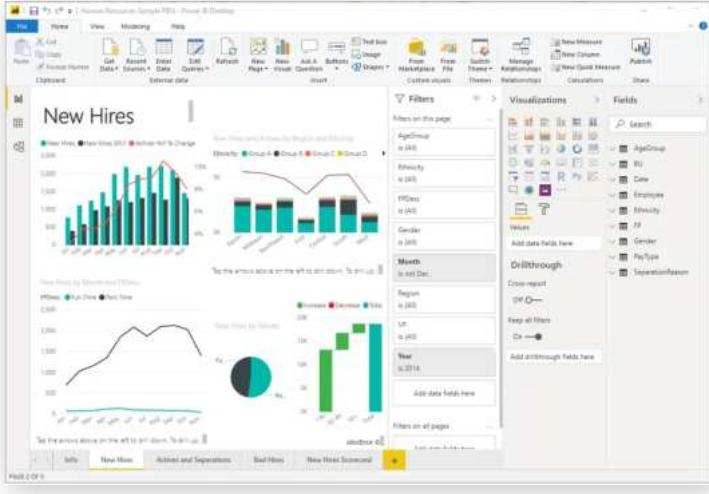
\*Previously known as Dataset

# An overview of Power BI

## Power BI Desktop



Report maker  
(Analyst)

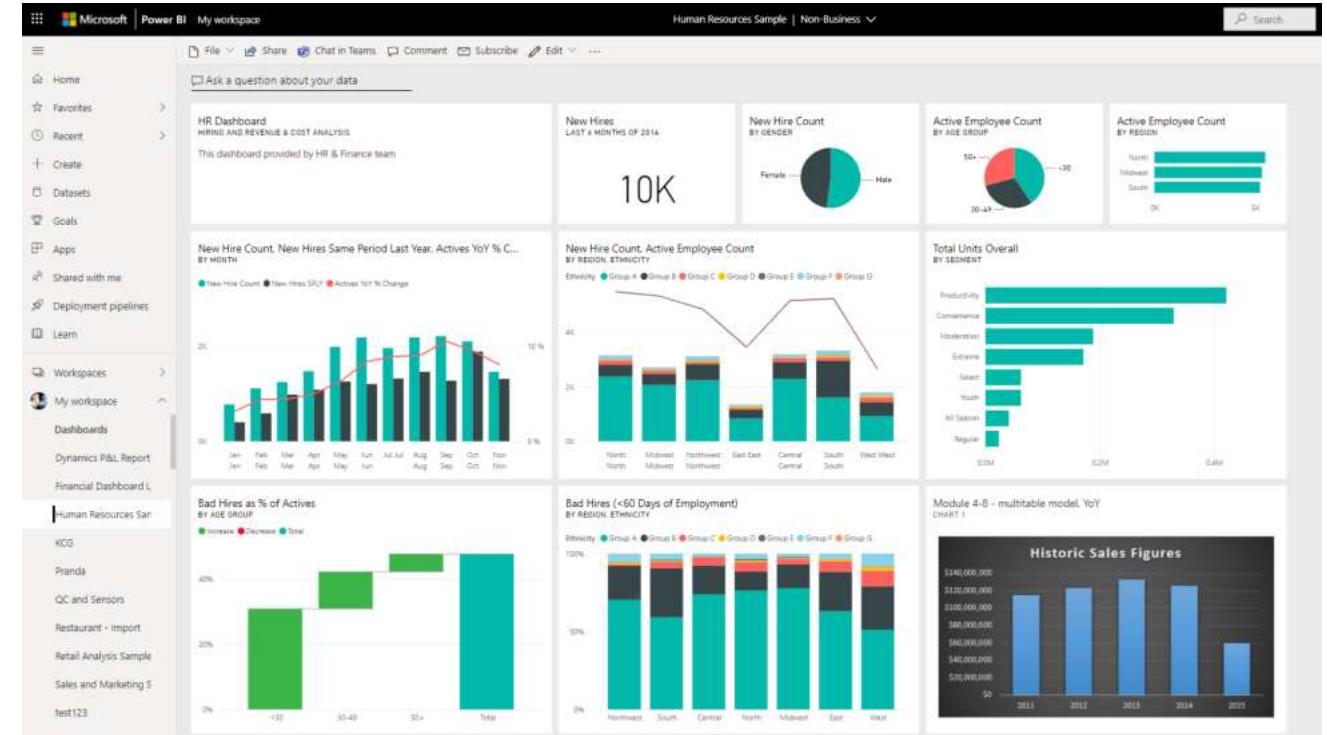


Viewers (on mobile)



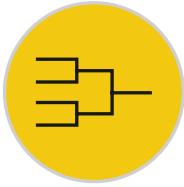
## Power BI Mobile

## Power BI Service (Cloud)



Viewers (on web)

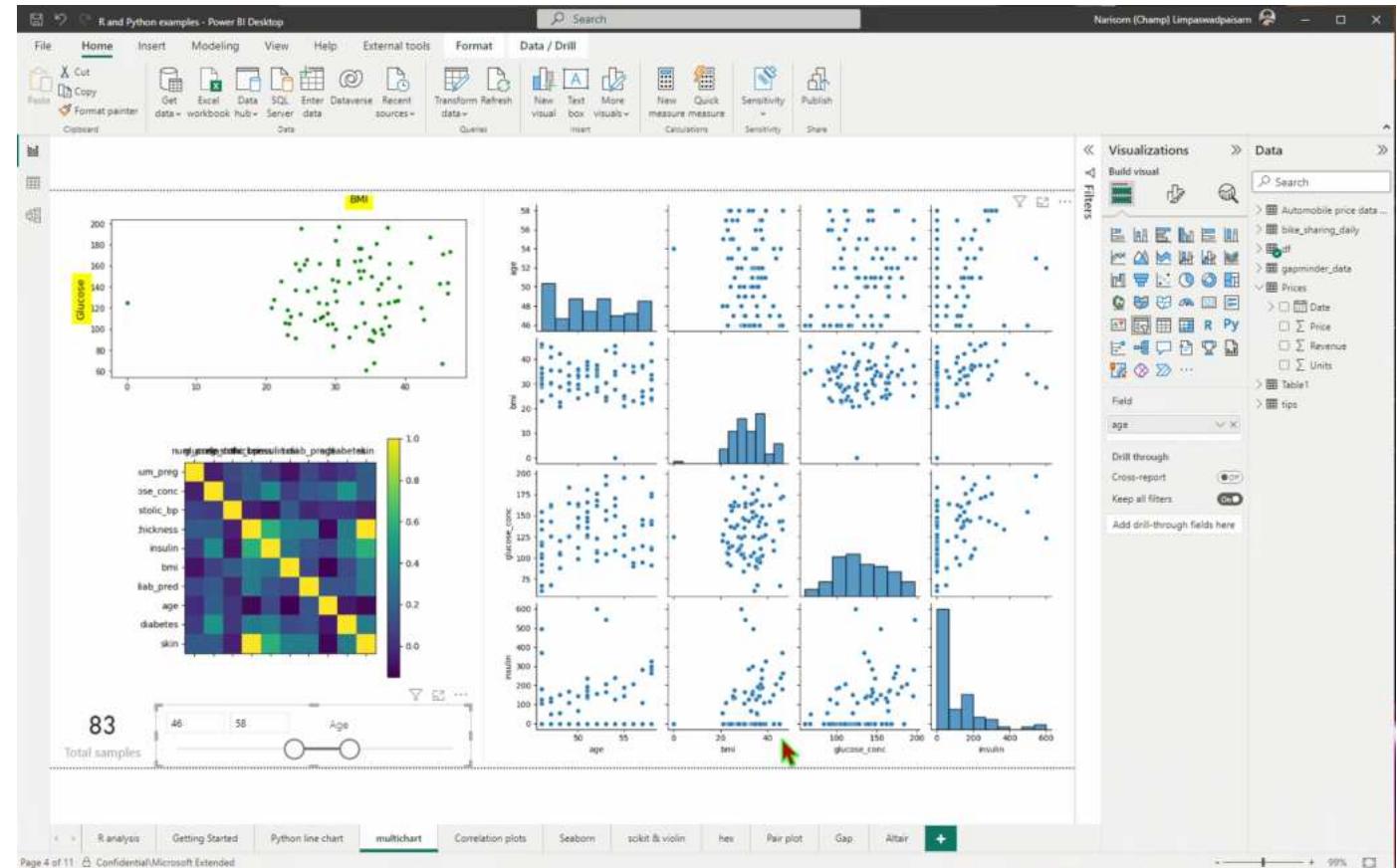
# Easy for data scientists and analysts to collaborate



## Python visualization

Use your python scripts in Power BI

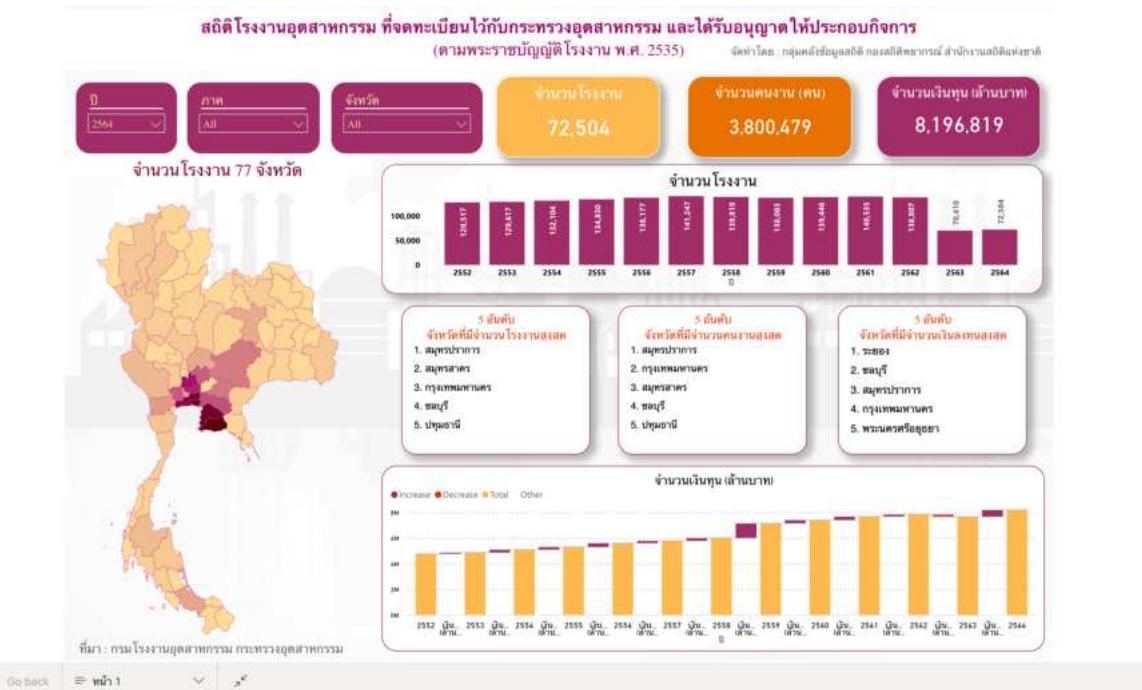
Python visualization behaves like any other visualization such than slicing on data triggers the script to rerun and the visual is refreshed



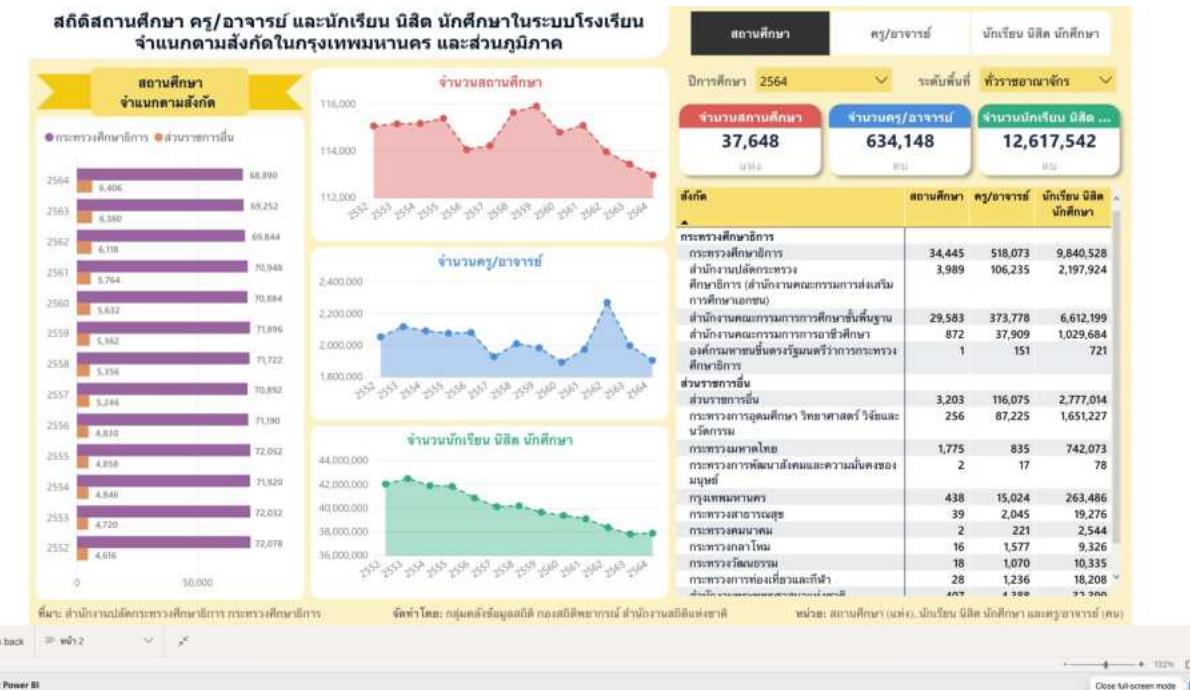
# Power BI กับตัวอย่างการใช้ร่วมกับ Public data

สามารถนำ Power BI ไปใช้ในเว็บไซต์ที่เป็นสาธารณะได้

<http://ittdashboard.nso.go.th>



[http://ittdashboard.nso.go.th/preview.php?id\\_project=54](http://ittdashboard.nso.go.th/preview.php?id_project=54)



[http://ittdashboard.nso.go.th/preview.php?id\\_project=79](http://ittdashboard.nso.go.th/preview.php?id_project=79)

# Power BI กับตัวอย่างการใช้ร่วมกับ Public data

หมวด Application



<https://mohpromt.moph.go.th/mpc/mp-pf/statistics/>

The screenshot shows a Microsoft Power BI dashboard titled "mohpromt.moph.go.th/mpc/mp-pf/statistics/". The dashboard features several data visualizations and tables. At the top right, there is a search bar and navigation links. The main area includes a header with the MoH Promt logo and a "Digital Signature" section showing counts for Portal, HIS, and Station. Below this are two charts: "เขตเมืองที่ออก Digital Signature" (a map of Thailand with numbers 118, 53, 58, 77, 88, 83, 96, 93, 75, 91, 87, 34) and "จำนวน Digital Signature ที่ออกในแต่ละเขต" (a bar chart with values 118, 53, 58, 77, 88, 83, 96, 93, 75, 91, 87, 34). The bottom half of the dashboard is a large table with columns for "เขตเมือง", "จำนวน Digital Signature", "แพทย์ผู้ออก Digital Signature (MOPH Certificate)", "หน่วยงานที่ออก Digital Signature", "แพทย์ผู้ออก Digital Signature (Station)", "จำนวนหน้าจอคอมพิวเตอร์ Portal (台)", "จำนวนหน้าจอคอมพิวเตอร์ HIS (台)", and "จำนวนหน้าจอคอมพิวเตอร์ HIS (台)". The table lists 13 districts with their respective statistics.

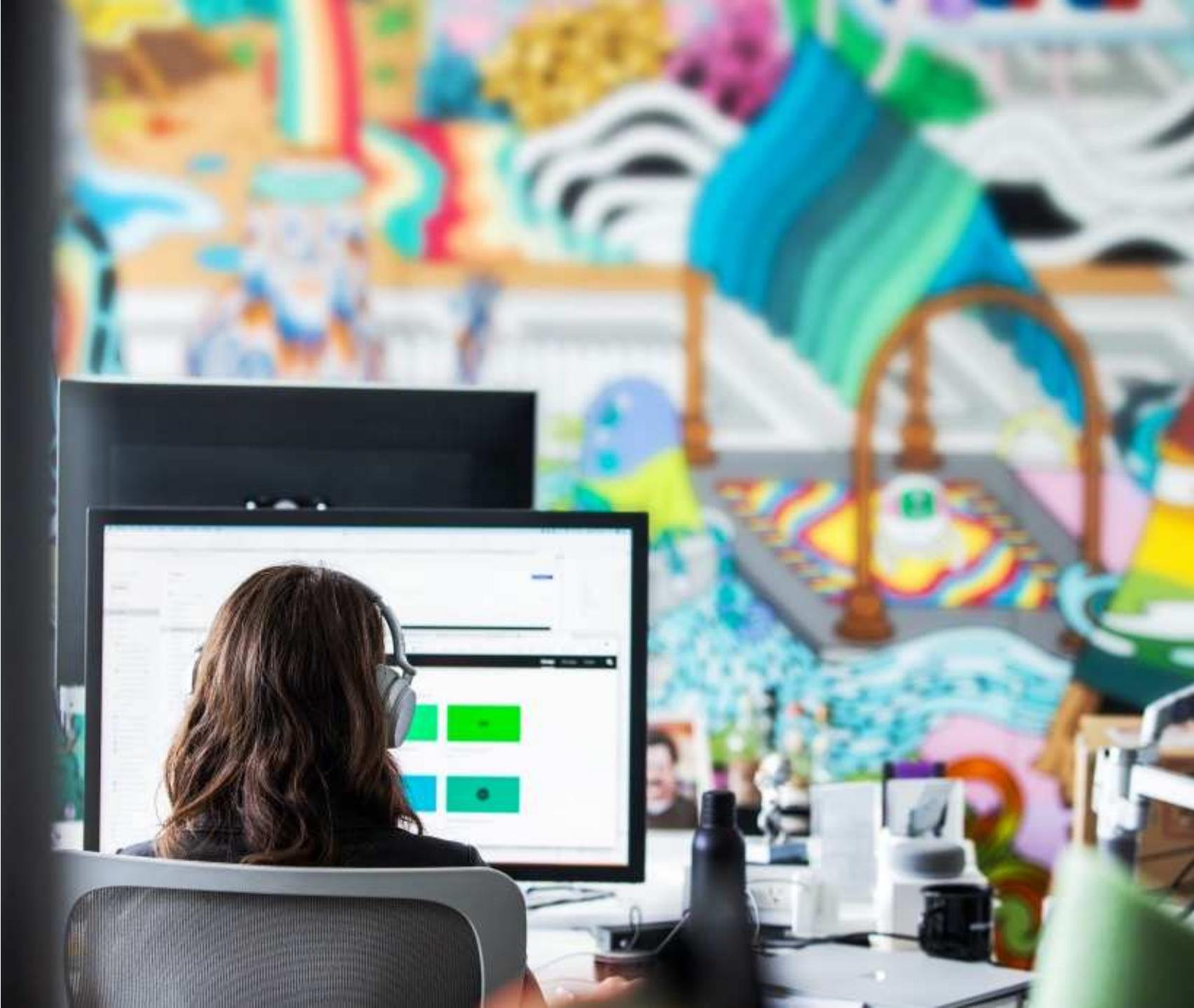
# Lab 2: Create a basic Power BI Report

<https://tinyurl.com/kufabric>

# Microsoft Fabric (high level)

**Business Applications Domain**

Narisorn Limpaswadpaisarn  
Client Technology Lead



# Microsoft cloud services for data

## Operational data workloads



### Azure SQL

- Family of SQL Server based relational database services



### Azure Database for open-source

- Maria DB, MySQL, PostgreSQL



### Azure Cosmos DB

- Highly scalable non-relational database system



### Azure Storage

- File, blob, and table storage
- Hierarchical namespace for data lake storage

## Analytical data workloads

### Software-as-a-Service (SaaS)



#### Microsoft Fabric

Unified, SaaS based analytics platform based on open and governed lakehouse:

- Data ingestion and ETL
- Data Lakehouse
- Data Warehouse
- Data Science and ML
- Realtime Analytics
- Data visualization
- Data governance and management

Today's focused

### Platform-as-a-Service (PaaS)



#### Azure Synapse Analytics

- Integrated solution for data analytics in Azure
- Pipelines, Apache Spark, SQL, Data Explorer



#### Azure Databricks

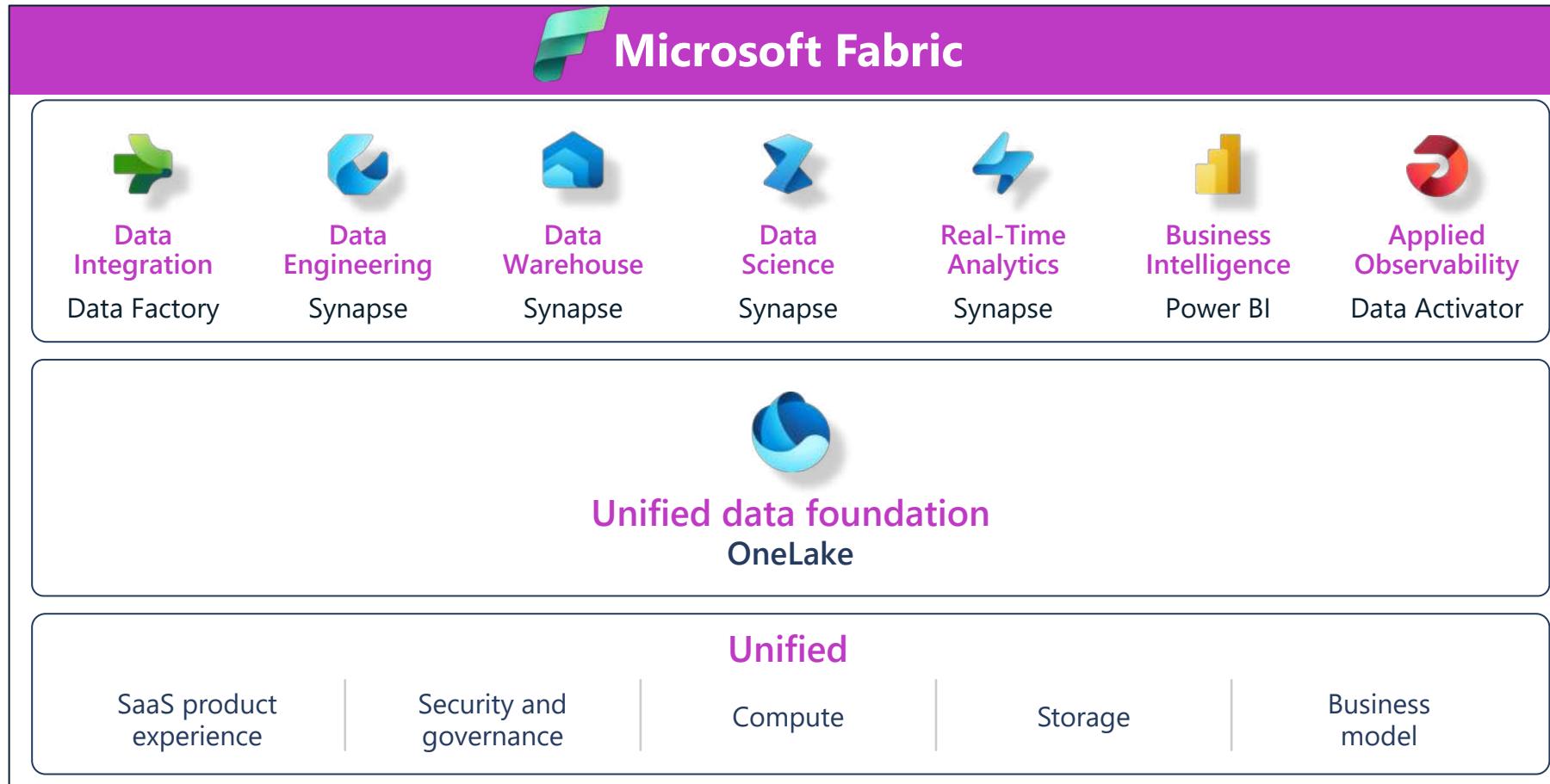
- Apache Spark analytics and data processing



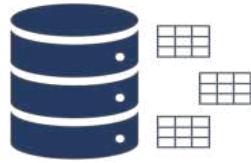
#### Azure HDInsight

- Apache open-source platform

# SaaS data analytics with Microsoft Fabric



# Data processing in large-scale analytics



## Relational Database

- Well established model for relational data storage and processing
- Comprehensive SQL language support for querying and data manipulation



## Apache Spark

- Open-Source platform for scalable, distributed data processing
- Multi-language data processing code (Python, Scala, Java, SQL, ...)

# Star schema (& Snowflake)

Normally for transactional DB, data is **normalized** to reduce duplication.

But for data warehouse, data is **de-normalized** to reduce number of joins required to query

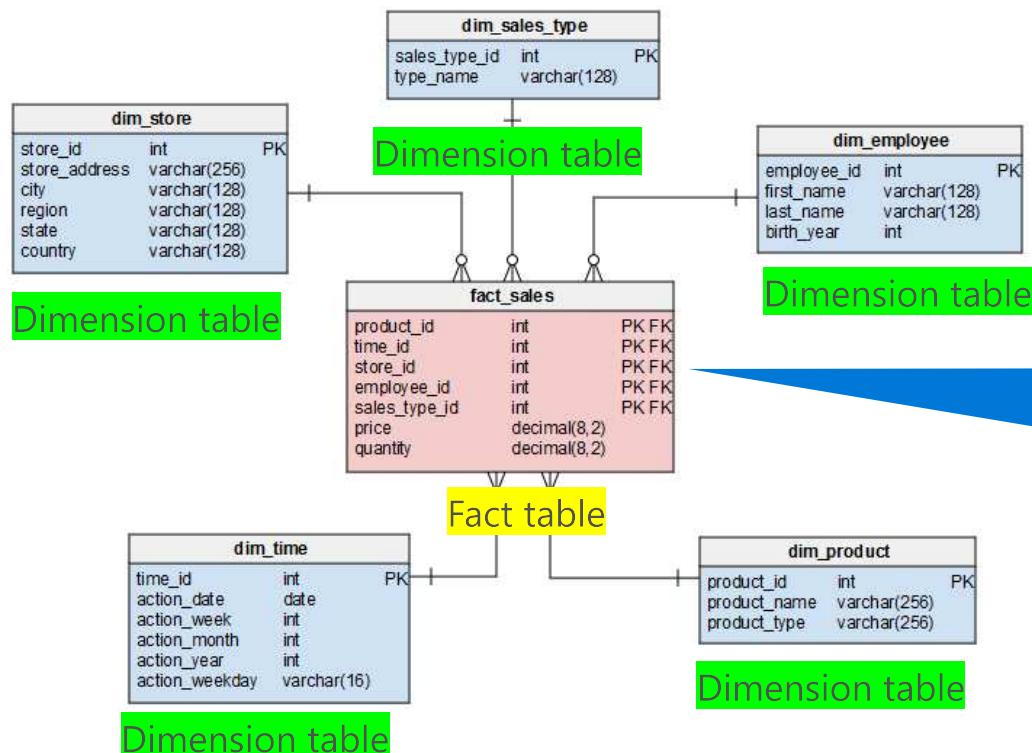
- Avoid wide flat tables, use Star Schema for OLAP (warehouse)
- Fact table: Often created from DB view, **measures** are kept for business questions
- Dimension table: **Attributes** are kept for explain values in Fact table
- If there is another one-to-many table linked from dimension table, we called **snowflake schema**

Fact table can be such as

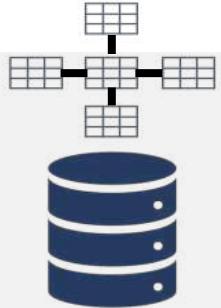
- Sales order
- Stock balances
- Exchange rates
- Temp readings

\*\* It's make sense to have some normalization applied as snowflake such as

- dim\_geography linked to both dim\_customer and dim\_store
- dim\_supplier added to dim\_product



# Analytical data store architectures



## Data Warehouse

- Data is stored in a relational database and queried using a SQL query engine
- Tables are *denormalized* for query optimization
  - Typically as a star or snowflake schema of numeric *facts* that can be aggregated by *dimensions*



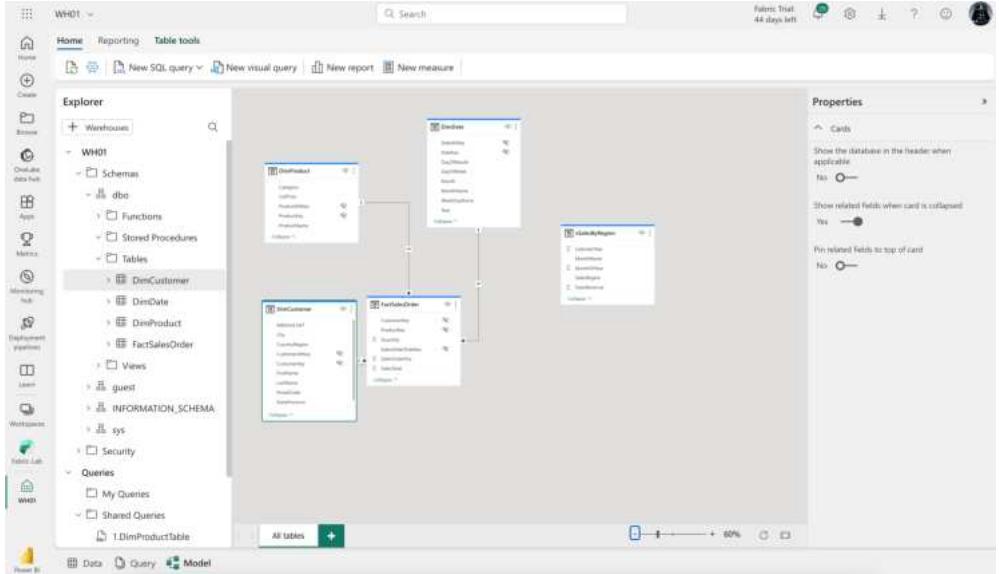
## Data Lakehouse

- Data files are stored in a distributed file system (a *data lake*) and typically processed using Apache Spark
- Metadata is used to define tables that provide a relational SQL interface to the file data
  - Commonly, a *delta lake* format is used to provide transactional database functionality

Both can be achieved with Microsoft Fabric  
Warehouse & Lakehouse

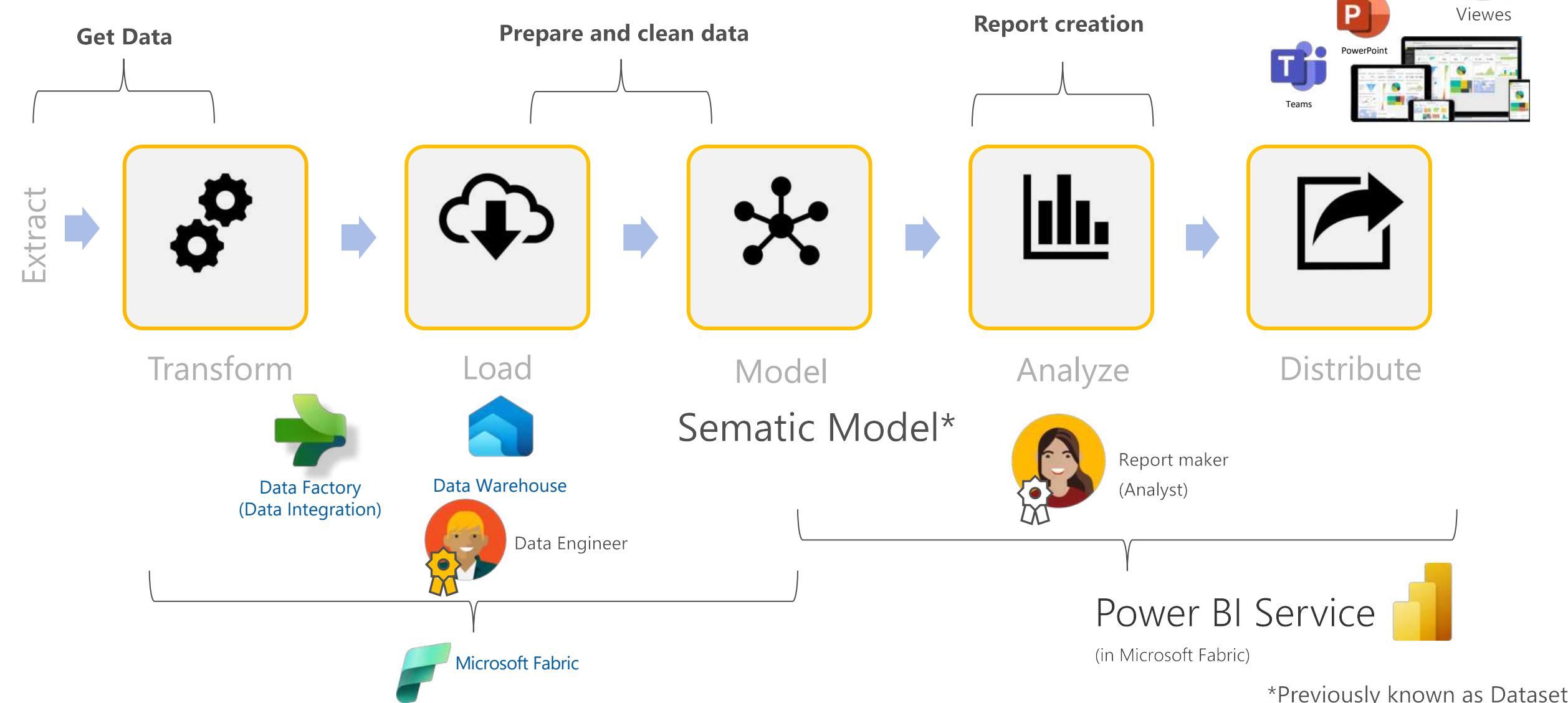
# Fabric Warehouse

This screenshot shows the Microsoft Fabric Data preview interface. The left sidebar includes sections for Home, Reporting, Table tools, Create, Browse, Overlake data hub, App, Metrics, Monitoring hub, Deployment pipelines, Learn, Workspaces, Fabric Lab, WH01, and Power BI. The main area displays a table titled "Data preview" with 417 rows. The columns include CustomerKey, CustomerAltKey, Title, FirstName, LastName, AddressLine1, City, StateProvince, and CountryRegion. The table lists various customers from different locations like Van Nuys, California, and Lewiston, Idaho.



This screenshot shows the Microsoft Fabric Query interface. The top navigation bar includes Home, Reporting, Table tools, and a search bar. The left sidebar has sections for Home, Reporting, Table tools, Create, Browse, Overlake data hub, App, Metrics, Monitoring hub, Deployment pipelines, Learn, Workspaces, Fabric Lab, WH01, and Power BI. The central area is a query editor with a code pane containing T-SQL queries. The code is annotated with notes such as "Note that the attributes in the time dimension enable you to aggregate!" and "Run the modified query and review the results, which now include sales!". The queries involve joins between DimCustomer, DimDate, DimProduct, FactSalesOrder, and FactSalesOrder tables.

# Lakehouse Development Workflow



# Lab 3: Explore Data Warehouse

<https://microsoftlearning.github.io/mslearn-fabric/Instructions/Labs/06-data-warehouse.html>



# Fabric Lakehouse

This screenshot shows the Microsoft Fabric Data Explorer interface. On the left, the sidebar includes Home, Create, Browse, Apps, Metrics, Monitoring hub, Deployment pipelines, Learn, Workspaces, and a specific section for champLH1. The main area displays the 'salesorders' table with columns: SalesOrderID, SalesOrderLineID, OrderDate, CustomerID, Email, Item, Quantity, UnitPrice, and Tax. The table contains 18 rows of sample data. A message at the top states: "A SQL endpoint for SQL querying and a default dataset for reporting were created and will be updated with any tables added to the lakehouse."

SalesOrderID	SalesOrderLineID	OrderDate	CustomerID	Email	Item	Quantity	UnitPrice	Tax
5051555	7	6/23/2021 1...	Chloe Garcia	chloe27@ad...	Patch Kit/B...	1	2.29	0.1832
5054042	7	8/9/2021 1...	Logan Collins	logan29@ad...	Half-Finger...	1	24.49	1.9932
5054784	7	8/22/2021 ...	Autumn Li	autumn30@ad...	All-Purpose...	1	139	12.72
5058572	7	10/25/2021 ...	Cesar Soto	cesar9@ad...	Short-Sleev...	1	53.99	4.3192
5058845	7	10/30/2021 ...	Peter Shi	peter10@ad...	Sport-100...	1	34.99	2.7992
5058845	8	10/30/2021 ...	Peter Shi	peter10@ad...	Long-Sleev...	1	49.99	3.9982
5060333	7	11/16/2021 ...	Jason Mitchell	jason40@ad...	Sport-100...	1	34.99	2.7992
5061412	7	12/3/2021 ...	Nathaniel C.	nathaniel5@ad...	Short-Sleev...	1	53.99	4.3192
5062984	7	12/29/2021 ...	Miguel Sanchez	miguel72@ad...	Racing Soc...	1	8.99	0.7192
5051555	6	6/23/2021 ...	Chloe Garcia	chloe27@ad...	Mountain B...	1	9.99	0.7992
5052058	6	7/4/2021 1...	Elijah Ross	elijah7@ad...	Short-Sleev...	1	53.99	4.3192
5053255	6	7/26/2021 ...	Edward Taylor	edward31@ad...	Short-Sleev...	1	53.99	4.3192
5053852	6	8/6/2021 1...	Maria Reed	maria4@ad...	Hydration P...	1	54.99	4.3992
5054042	6	8/9/2021 1...	Logan Collins	logan29@ad...	Short-Sleev...	1	53.99	4.3192
5054377	6	8/15/2021 ...	Ashlee Xu	ashlee12@ad...	Water Bottl...	1	4.99	0.3992
5054784	6	8/22/2021 ...	Autumn Li	autumn30@ad...	Patch Kit/B...	1	2.29	0.1832
5054784	1	8/22/2021 ...	Autumn Li	autumn30@ad...	Patch Kit/B...	1	2.29	0.1832

This screenshot shows the Microsoft Fabric Data Explorer interface, similar to the one above but with a different view. The sidebar includes Home, Create, Browse, Apps, Metrics, Monitoring hub, Deployment pipelines, Learn, Workspaces, and a specific section for champLH1. The main area displays the 'salesorders (file view)' with a table showing file details. A message at the top states: "A SQL endpoint for SQL querying and a default dataset for reporting were created and will be updated with any tables added to the lakehouse."

Name	Date modified	Type	Size
_delta_log	6/2/2023 2:37:17 PM	Folder	2 items
part-00000-4ccaf5c3-b182-4833-bec5-1e619aafac9...	6/2/2023 2:37:18 PM	PARQUET	413 KB
part-00001-d354995e-8aa8-40e7-8d89-bcdaf30f21...	6/2/2023 2:37:17 PM	PARQUET	80 KB
part-00002-9ece1406-4a89-4b51-a165-984d0971b...	6/2/2023 2:37:17 PM	PARQUET	37 KB

# Jupyter Notebook



- The Jupyter Notebook is the original web application for creating and sharing computational documents. It offers a simple, streamlined, document-centric experience.
- Jupyter support many programming language including Python, Markup, HTML, R or Scala
- Interactive output for learning or experiment (Ex. Query result, HTML, Image or video)
- Library & Big data integration for data science and engineering : Apache Spark, pandas, scikit-learn etc.
- In Fabric, we have Fabric notebook as Jupyter notebook implementation



Track Machine Learning experiments and models

A machine learning model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data. Once you have trained the model, you can use it to reason over data that it hasn't seen before, and make predictions about that data.

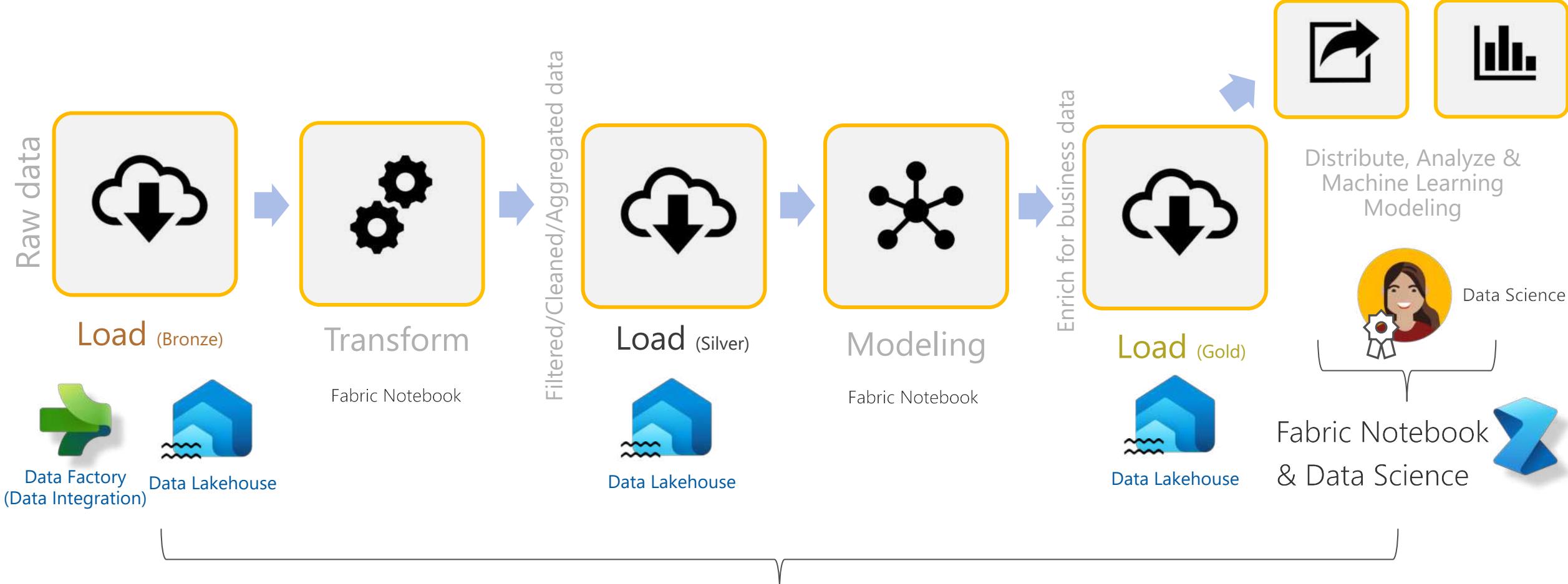
In this notebook, you will learn the basic steps to run an experiment, add a model version to track run metrics and parameters and register a model.

```
1 import mlflow
2
3 # Set given experiment as the active experiment. If an experiment with this name does not exist, a new experiment will be created.
4 mlflow.set_experiment("your-experiment-name")
5
```

PySpark (Python) ▾

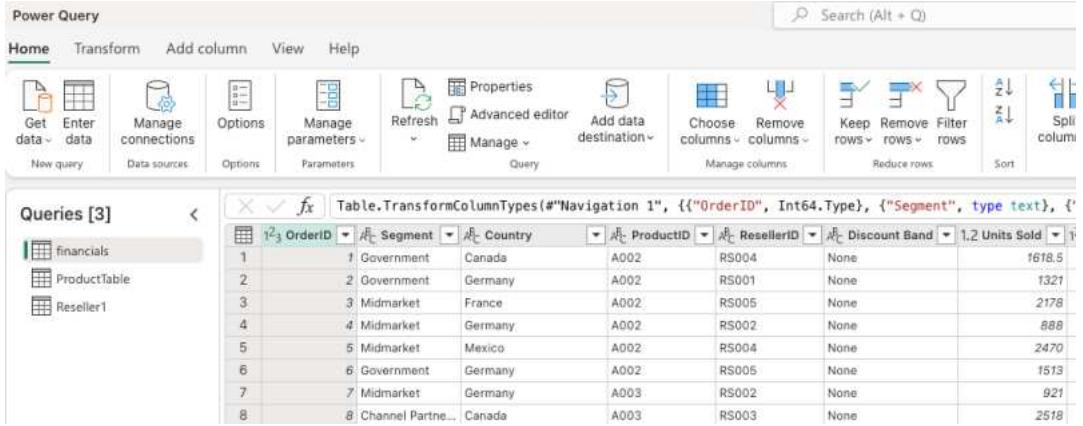
```
1 import mlflow
2 import numpy as np
3 from sklearn.linear_model import LogisticRegression
4 from mlflow.models.signature import infer_signature
5
6 # Start your training job with 'start_run()'
7 with mlflow.start_run() as run:
8
9     lr = LogisticRegression()
10    X = np.array([-2, -1, 0, 1, 2, 3]).reshape(-1, 1)
11    y = np.array([0, 0, 1, 1, 1, 0])
12    lr.fit(X, y)
13    score = lr.score(X, y)
```

# Medallion Lakehouse Development Workflow



\*Previously known as Dataset

# Tool for data ingestion & transformation

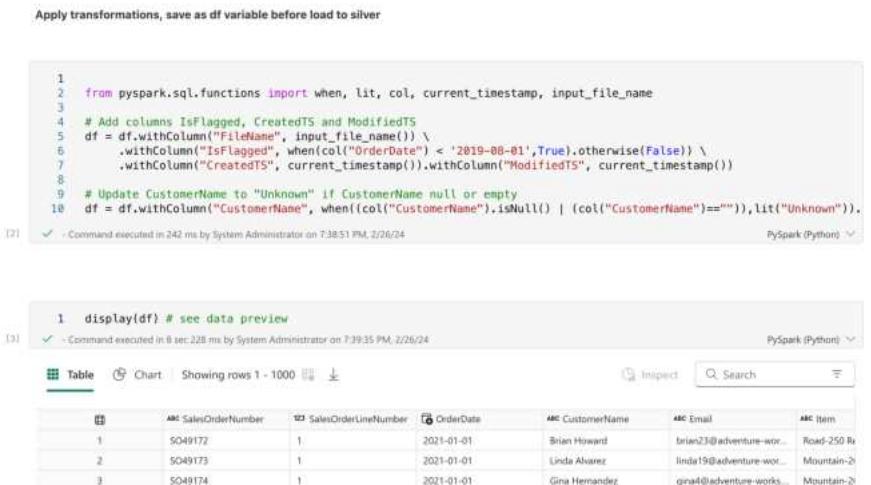


The screenshot shows the Microsoft Power Query interface. The top navigation bar includes Home, Transform, Add column, View, Help, and a search bar. Below the bar are various icons for managing data sources, properties, and transformations. The main area displays a table titled "Table.TransformColumnTypes(#"Navigation 1", {"#OrderID": Int64.Type}, {"Segment", type text}, {"#"). The table contains 8 rows of data with columns: OrderID, Segment, Country, ProductID, ResellerID, Discount Band, and Units Sold. A sidebar on the left lists "Queries [3]" with entries: Financials, ProductTable, and Reseller1.

Dataflow with Power Query engine  
(usually for Warehouse)



Data Factory for orchestration  
(both Warehouse & Lakehouse)

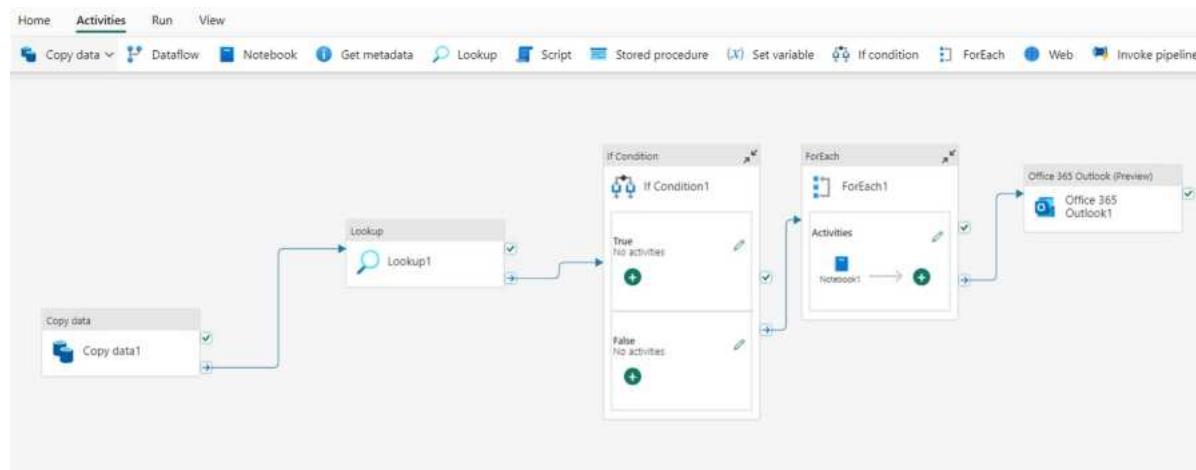


The screenshot shows a Jupyter Notebook cell with PySpark code. The code imports pyspark.sql.functions, defines a DataFrame df, adds columns IsFlagged, CreatedTS, and ModifiedTS, and updates the CustomerName to "Unknown" if it's null or empty. Below the code is a data preview table with columns: SalesOrderNumber, SalesOrderLineNumber, OrderDate, CustomerName, Email, and Item. The table shows three rows of sample data.

```
from pyspark.sql.functions import when, lit, col, current_timestamp, input_file_name
# Add columns IsFlagged, CreatedTS and ModifiedTS
df = df.withColumn("FileName", input_file_name())
       .withColumn("IsFlagged", when(col("OrderDate") < '2019-08-01', True).otherwise(False))
       .withColumn("CreatedTS", current_timestamp())
       .withColumn("ModifiedTS", current_timestamp())
# Update CustomerName to "Unknown" if CustomerName null or empty
df = df.withColumn("CustomerName", when((col("CustomerName").isNull()) | (col("CustomerName")=="")), lit("Unknown"))

display(df) # see data preview
```

Notebook with Spark engine  
(usually for Lakehouse)



# Lab 4: Ingest data with pipeline (Data factory)

<https://microsoftlearning.github.io/mslearn-fabric/Instructions/Labs/04-ingest-pipeline.html>



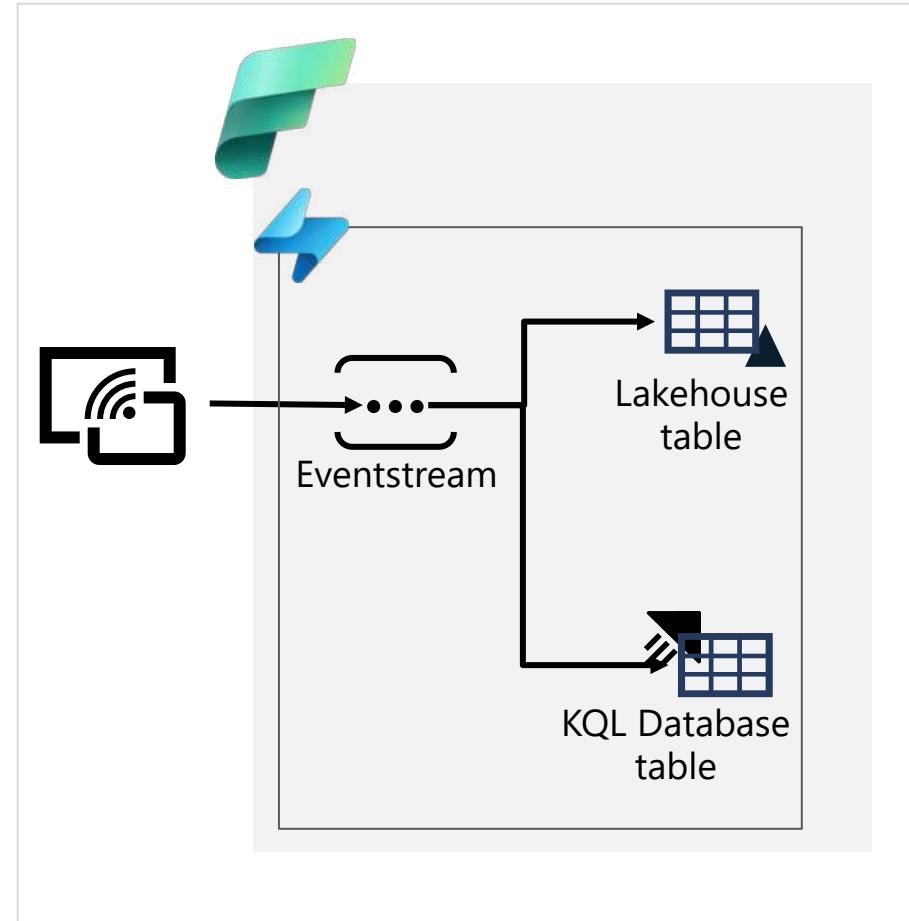
# Lab 5: Explore Data Lakehouse

<https://microsoftlearning.github.io/mslearn-fabric/Instructions/Labs/02-analyze-spark.html>



# Real-time analytics in Microsoft Fabric

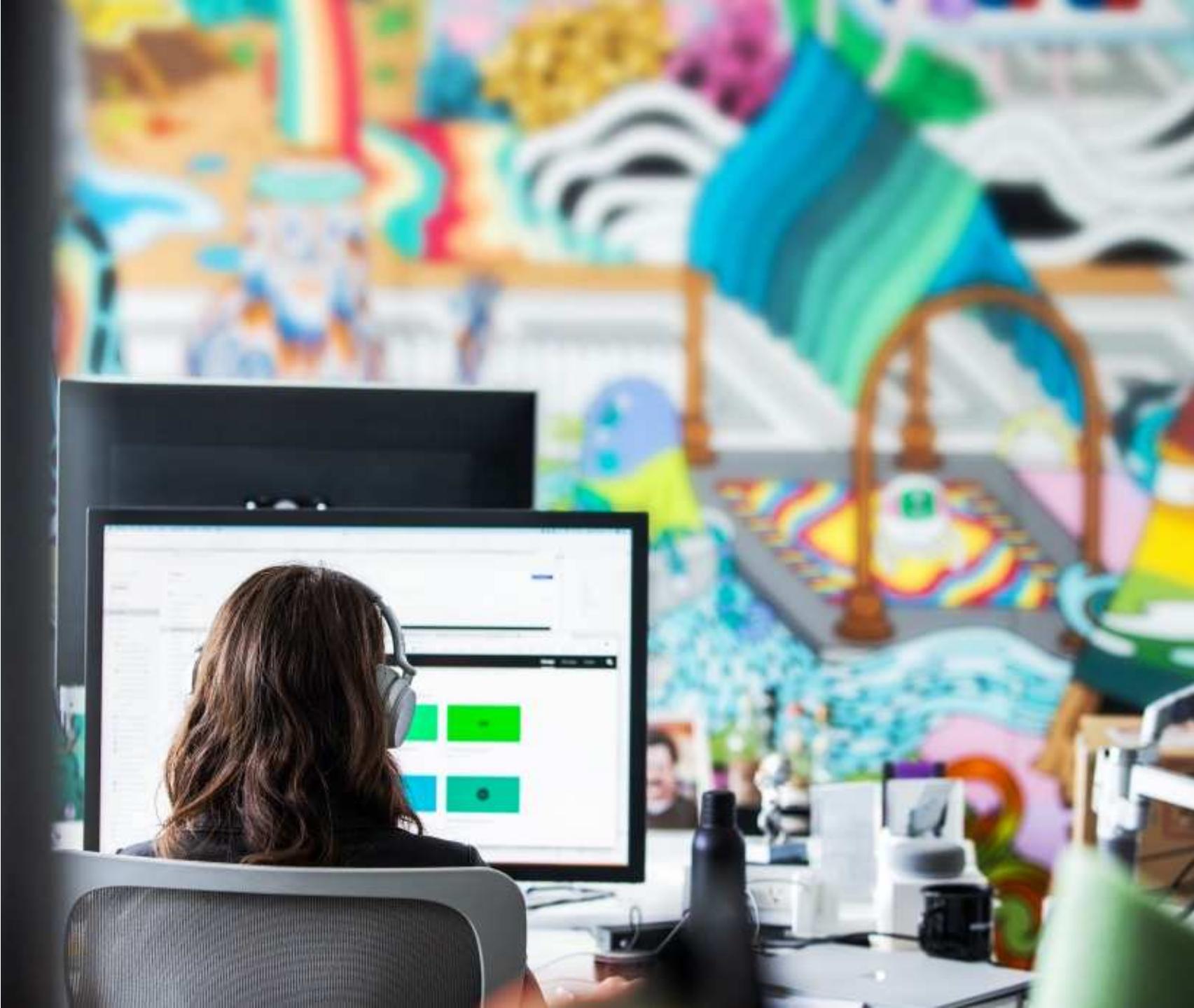
- Support for continuous data ingestion from multiple sources
- Capture streaming data in an **eventstream**
- Write real-time data to a table in a Lakehouse or a KQL database
- Query real-time data using SQL or KQL
- Build real-time visualizations



# Data Governance (high level)

**Business Applications Domain**

Narisorn Limpaswadpaisarn  
Client Technology Lead



# What is Data Governance



**Data Governance**

EP. 1

**Data Governance คืออะไร ?**

# Today's data realities

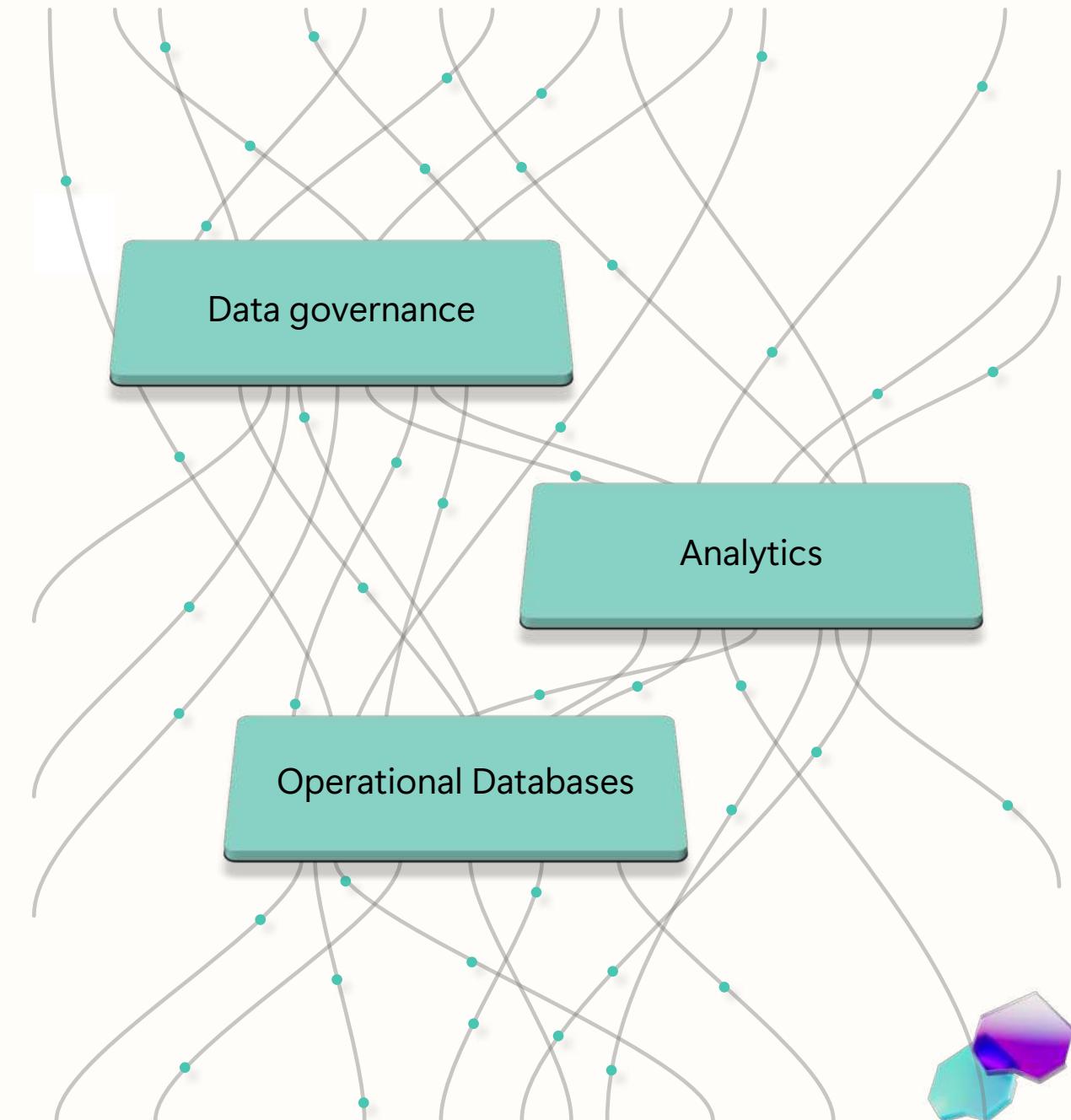
What **data** do I have?

Is it **trustworthy**?

Can people **access the data needed** to make the right decisions?

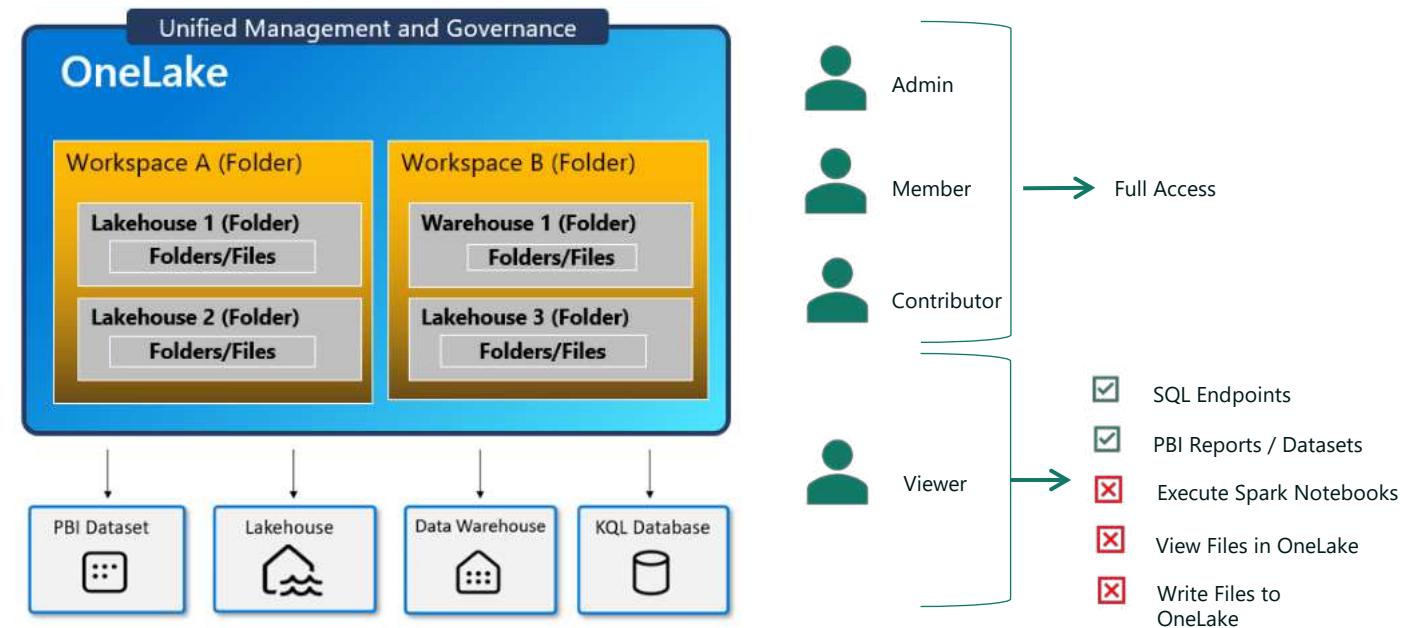
How can I enable faster **business insights**?

What's my **compliance** exposure?



# Workspace and item security in Fabric

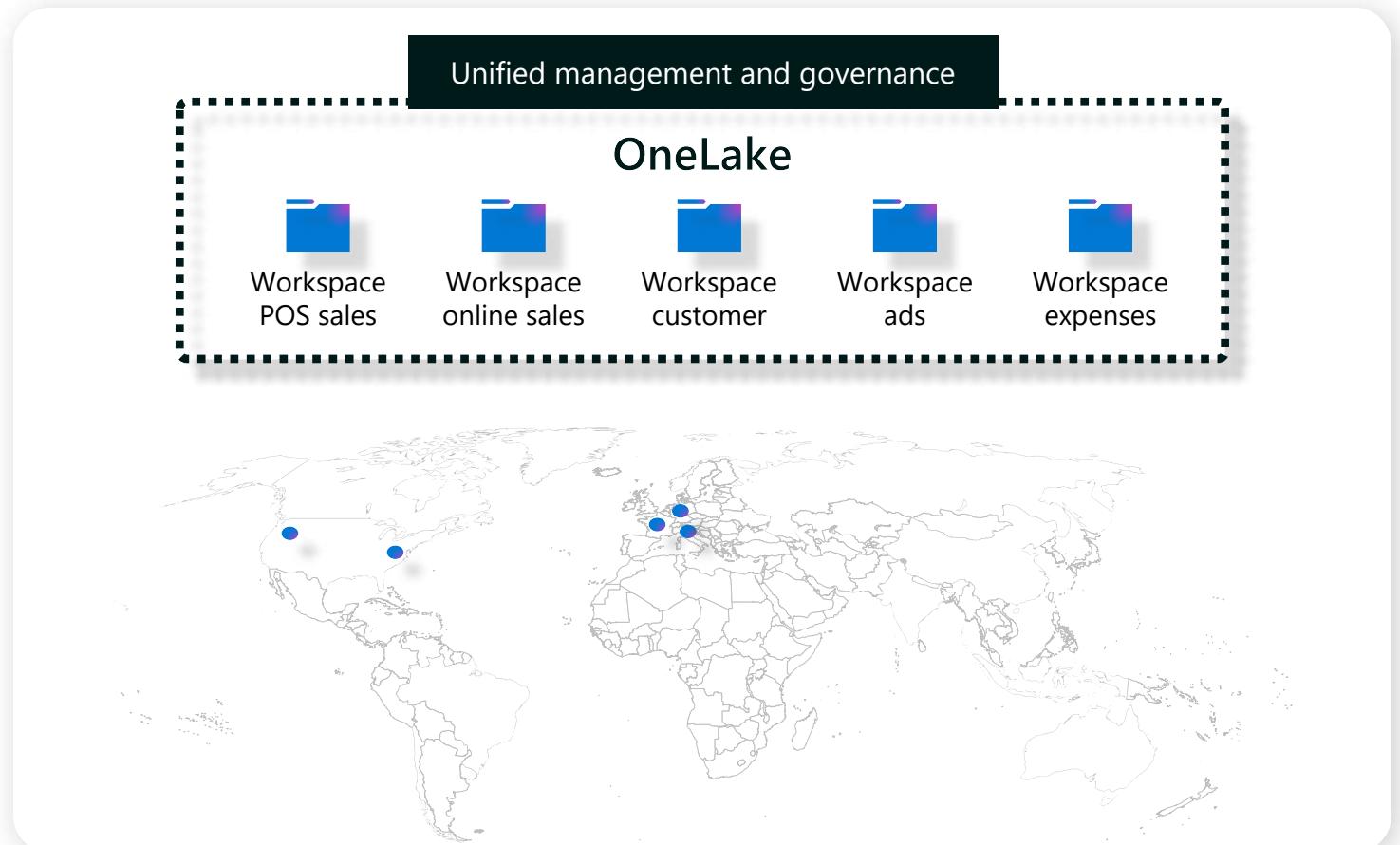
- Easily manage security in the workspace level (hosting all Power BI, Lakehouse etc.) and data associate using workspace with workspace role
- Assign permission directly to report, warehouse or lakehouse



# Data residency in Fabric

With the largest global footprint, Fabric multi-geo capacities allows control over content storage location in one of **54 data centers world-wide**

Workspaces can reside in **different regions** around the world while still being part of the same data lake.



# Information Protection (for data classification) in Fabric

champLH1 | Public

Name: champLH1

Location: ChampFabric1

Sensitivity: Public

Non-Business

General

Confidential

Highly Confidential

Learn more

partitioned\_data

products

transformed\_data

Highly Confidential Microsoft Extended  
Data is classified and protected. Microsoft Full Time Employees (FTE) and non employees can view, edit and reply.

Microsoft FTE

Any User (No Protection)

data.xlsx Highly Confidential

Page Layout Formulas Data Review View Automate

Territory

	B	C	D
State or Province	Revenue Won	Qualified Pipeline	
Total	\$7,087,045	\$10,642,117	
Total	\$5,224,507	\$8,939,484	
Total	\$4,042,481	\$6,092,445	
Total	\$2,335,420	\$3,822,838	
MA	\$1,074,432	\$1,763,270	

Time Intelligence - Power BI Desktop

File Home Insert Modeling View Optimize Help External tools

FY2020 Mountain-200 Bike Stock (UnitsBalance)

Product	2019 Jul	2019 Aug	2019 Sep	2019 Oct	2019 Nov	2019 Dec	2020 Jan	2020 Feb	2020 Mar	2020 Apr	2020 May	2020 Jun	Total
Mountain-200 Black, 38	4,884	4,649	4,784	4,873	4,605	4,943	5,208	4,872	5,208	5,040	5,208	2,520	56,794
Mountain-200 Black, 42	5,248	5,175	5,246	5,380	5,200	5,415	5,177	4,943	5,177	5,010	5,177	2,505	59,555
Mountain-200 Black, 46	5,522	5,617	5,287	5,419	5,273	5,663	5,890	5,510	5,890	5,700	5,890	2,850	64,512
Mountain-200 Silver, 38	5,412	5,190	5,114	5,245	5,220	5,400	5,890	5,510	5,890	5,700	5,890	2,850	63,317
Mountain-200 Silver, 42	5,252	5,262	5,043	5,269	5,072	5,208	5,022	4,698	5,022	4,860	5,022	2,430	58,157
Mountain-200 Silver, 46	5,004	5,019	4,960	5,012	4,847	5,231	5,053	4,727	5,053	4,898	5,053	2,445	57,294
Total	31,323	30,912	30,434	31,197	30,217	31,866	32,240	30,160	32,240	31,200	32,240	15,600	359,629

FY2020 Mountain-200 Bike Stock (Stock on Hand measure)

Product	2019 Jul	2019 Aug	2019 Sep	2019 Oct	2019 Nov	2019 Dec	2020 Jan	2020 Feb	2020 Mar	2020 Apr	2020 May	2020 Jun	Total
Mountain-200 Black, 38	151	171	99	172	30	168	168	168	168	168	168	168	1,168
Mountain-200 Black, 42	165	186	116	176	76	167	167	167	167	167	167	167	1,167
Mountain-200 Black, 46	182	184	131	172	111	190	190	190	190	190	190	190	1,190

You should now hide the inventory table UnitsBalance column. It will prevent report authors from inappropriately summarizing snapshot unit balances.

No records in 2020 Jun since it's doesn't happen

Sensitivity: Non-Business, Public, General, Confidential, Highly Confidential

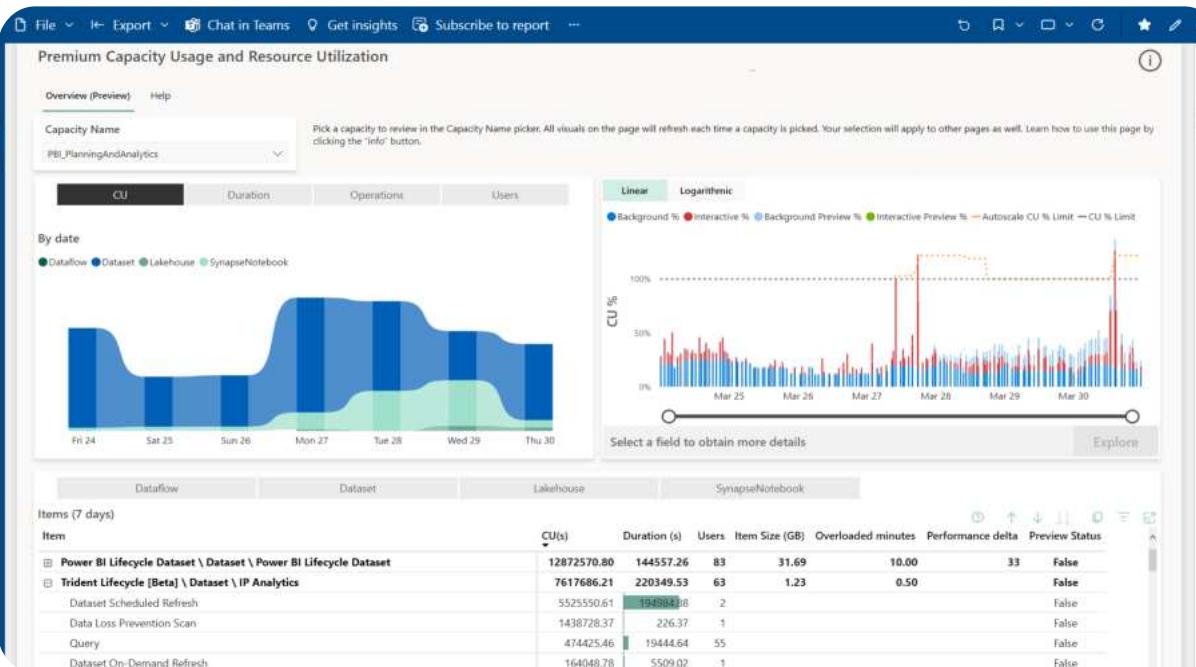
Visualizations: Build visual, Microsoft FTE, Microsoft Extended, Any User (No Protection)

Filters on all pages

Add data fields here

# Monitoring & Auditing

- Monitor compute usage for each capacity (Power BI report style)
- Monitor who doing what in Fabric tenant for legal investigation

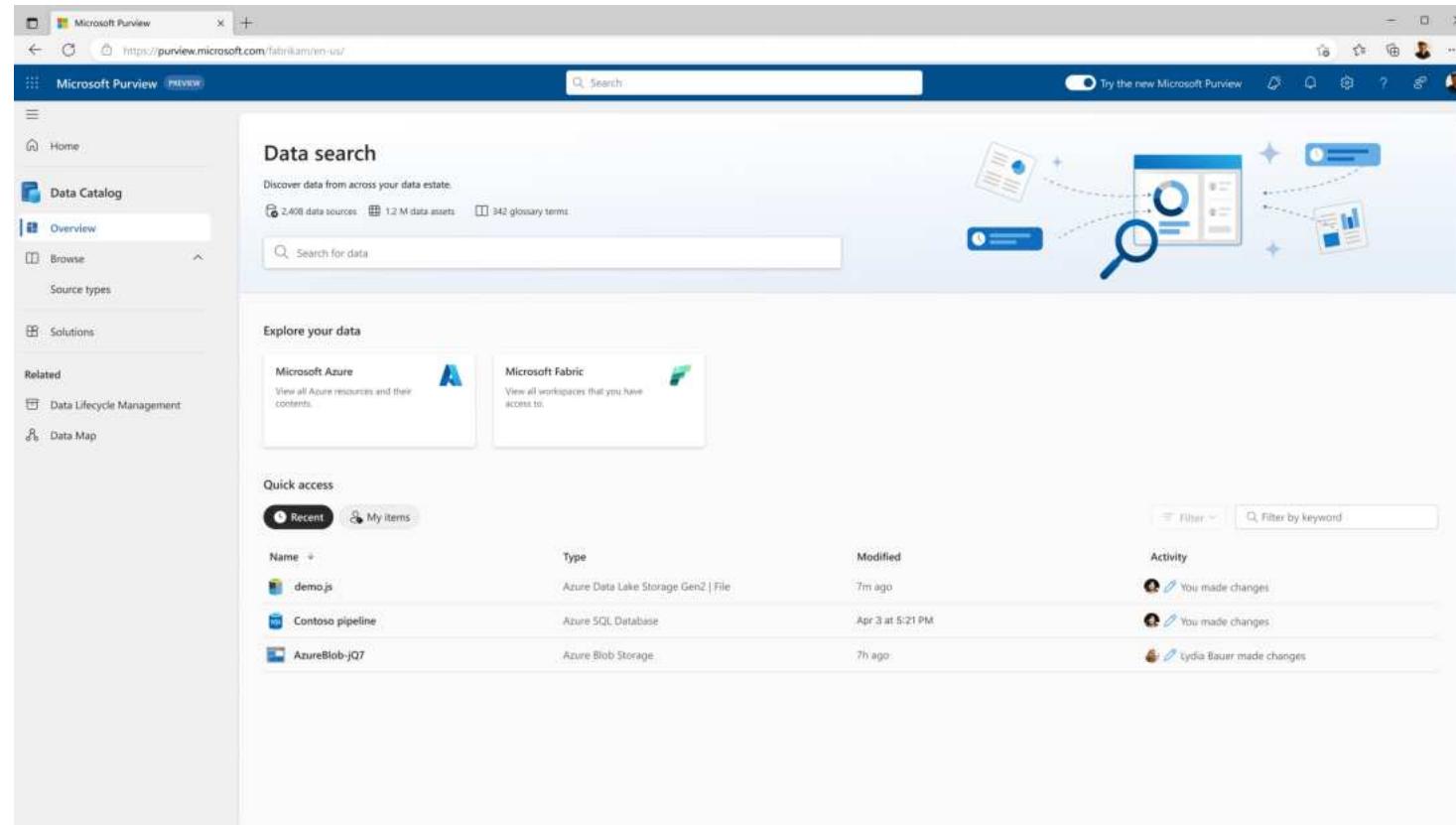


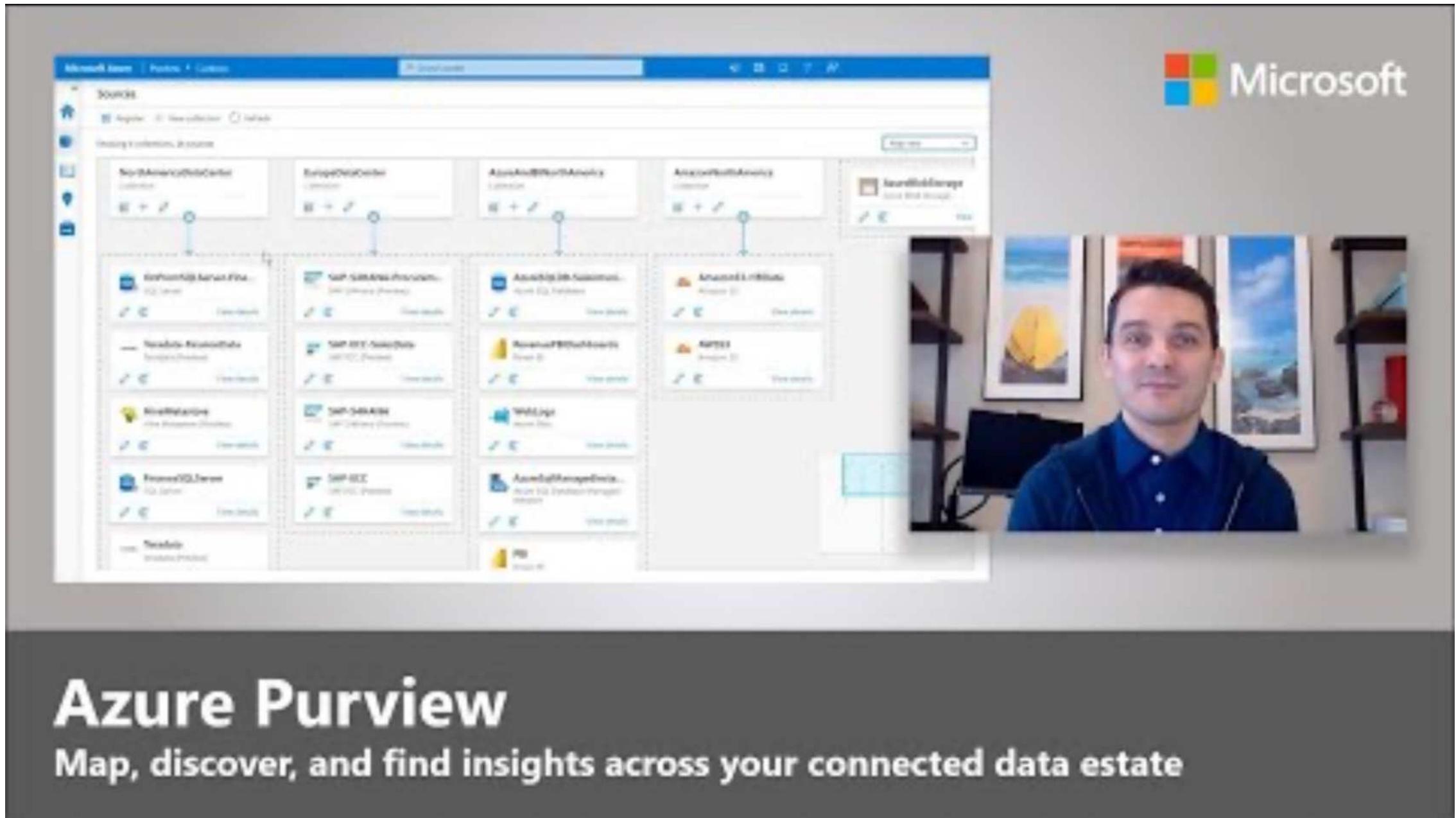
The screenshot shows the Microsoft Purview Audit interface. It displays a list of completed searches (22), active searches (2), and active unfiltered searches (1). It includes filters for Date and time range, Record type, Keyword search, and Workload, along with a table of audit logs.

Search name	Job status	Progress ...	Search ti...	Total results	Creation time ...	Search performed by
Mar 23 - Apr 24 createshortcut,createwarehouse,cre...	Queued	1s	0	0	24 Apr 2023 08:35	adminuser@pbidaily.onmicrosoft.com
Apr 23 - Apr 24	Queued	39s	0	0	24 Apr 2023 08:34	adminuser@pbidaily.onmicrosoft.com
Apr 24 - Apr 24 AdminUser	Completed	100%	6m, 32s	1463	24 Apr 2023 04:17	adminuser@pbidaily.onmicrosoft.com
Apr 24 - Apr 24 AdminUser	Completed	100%	7m, 16s	1424	24 Apr 2023 03:11	adminuser@pbidaily.onmicrosoft.com
Apr 19 - Apr 20 PowerBI	Completed	100%	3m, 13s	6216	20 Apr 2023 11:22	adminuser@pbidaily.onmicrosoft.com

# Data catalog and mapping in Fabric

- Auto-Discover what data (and Owner) is available across the data estate with the data catalog
- Search and explore data assets including Fabric data with technical properties, sensitive data classification, and data lineage





# Certifications in Fabric

- Fabric supports a wide range of compliance standard including GDPR, ISO 27001, 27701, 27017, 27018, HIPPA and etc.



# **Post-test questionnaire**

<https://tinyurl.com/kuposttest>