Microsoft

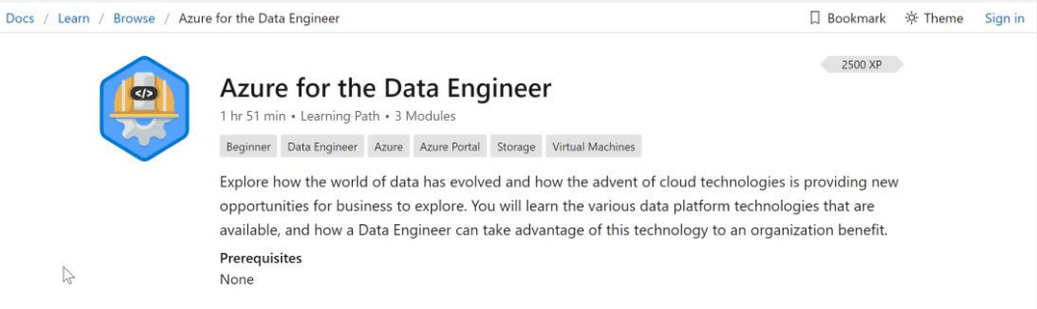
Data Engineering with Azure

Steve Young - Data & AI CSA (Microsoft)
styoung@Microsoft.com
http://5minutebi.com
http://linkedin.com/in/steveyoung/

January 23, 2020

OBJECTIVES

- What is Data Engineering? – Definition & Intro
- Steve's 7 Principles of Data Engineering & Azure
- Walkthrough of Examples in Azure
  - SQL Notebooks
  - Azure Data Factory Sliding Window
  - Data Virtualization – Use Data Where it Resides
- Questions and hopefully answers ☺

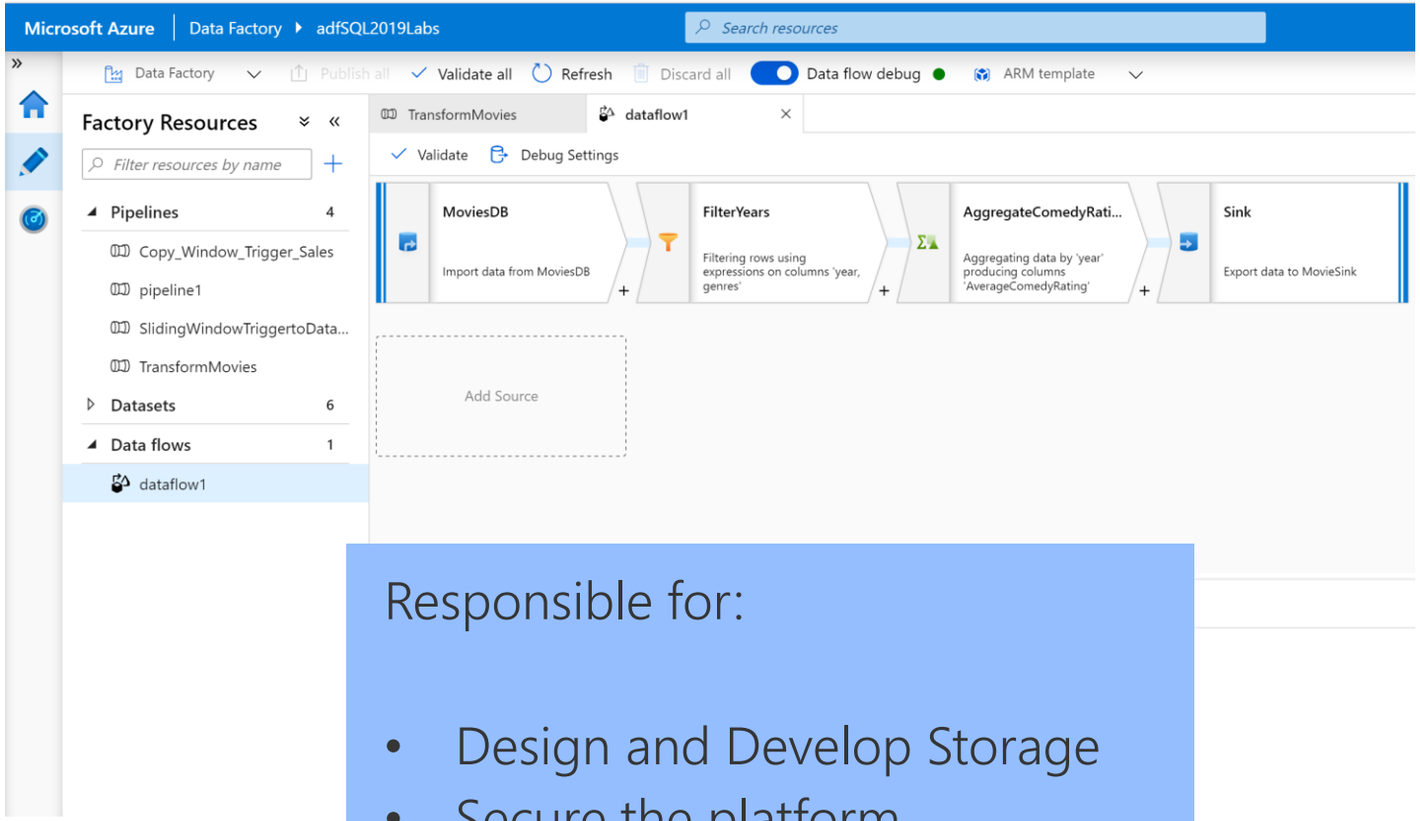


2 Hour MSLearn Course - Azure for the Data Engineer
https://docs.microsoft.com/en-us/learn/paths/azure-for-the-data-engineer/

DEMO

What is Data Engineering

https://docs.microsoft.com/en-us/learn/paths/azure-for-the-data-engineer/



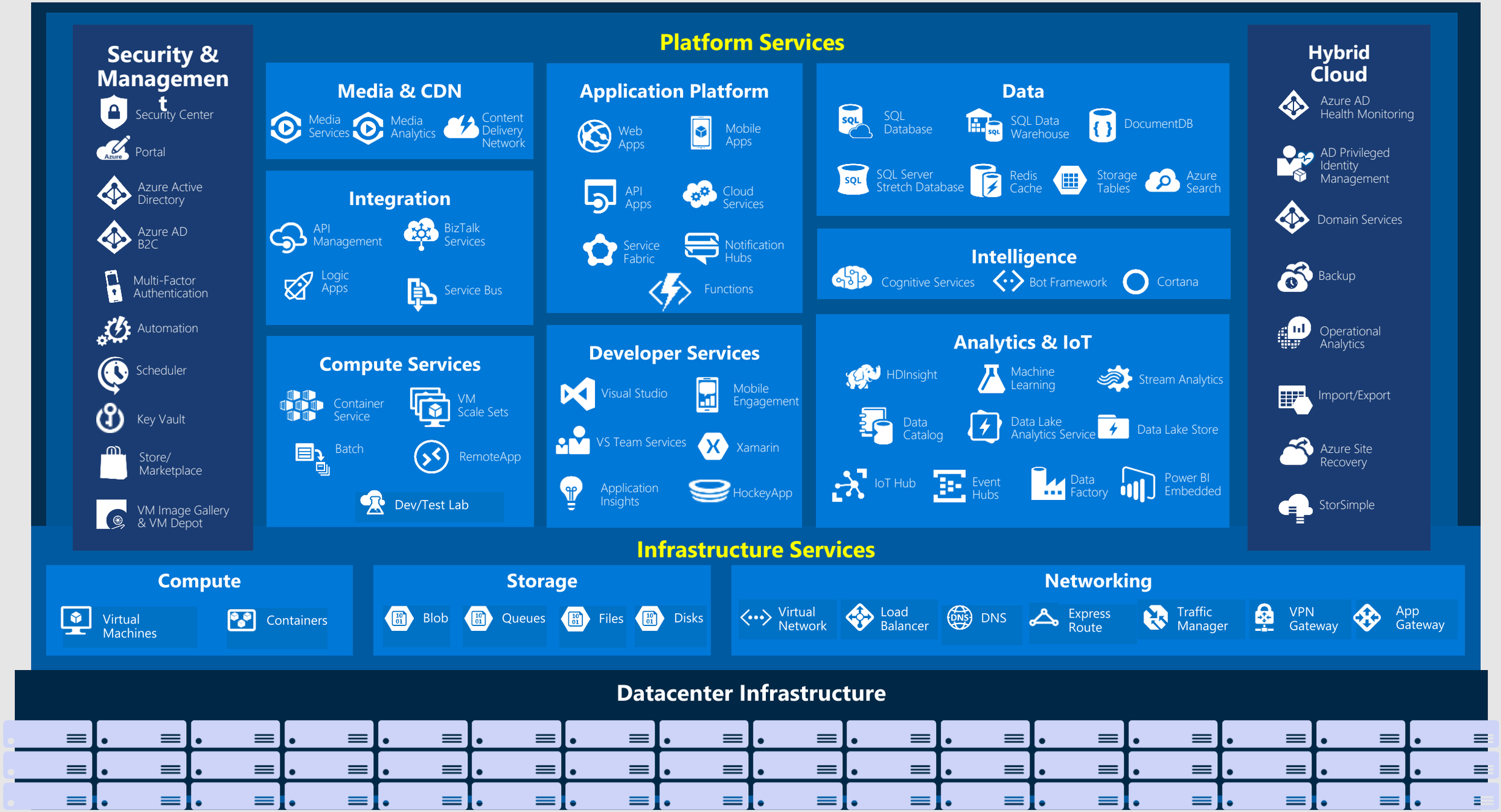
Responsible for:

- Design and Develop Storage
- Secure the platform
- Business continuity
- Move data around
- Monitor and improve process
- Solve Problems

- Data Scien
- Data Engi
- Big Data E
- BI Enginee
- BI Specialist
- Business Analytics Specialist

Scientist
Developer
ct

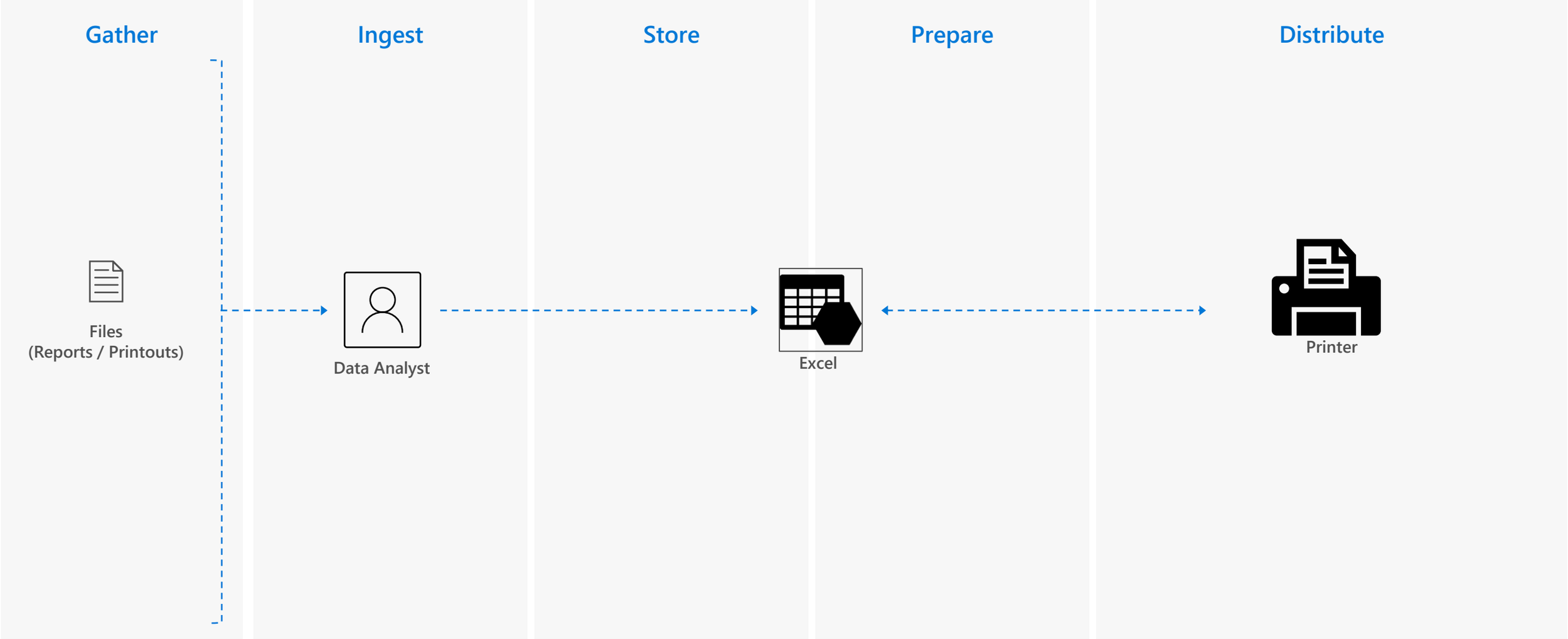
- Analytics Manager


Security & Management
Security Center
Portal
Azure Active Directory
Azure AD B2C
Multi-Factor Authentication
Automation
Scheduler
Key Vault
Store/ Marketplace
VM Image Gallery & VM Depot
Platform Services
Media & CDN
Media Services
Media Analytics
Content Delivery Network
Integration
API Management
BizTalk Services
Logic Apps
Service Bus
Compute Services
Container Service
VM Scale Sets
Batch
RemoteApp
Dev/Test Lab
Application Platform
Web Apps
Mobile Apps
API Apps
Cloud Services
Service Fabric
Notification Hubs
Functions
Developer Services
Visual Studio
Mobile Engagement
VS Team Services
Xamarin
Application Insights
HockeyApp
Data
SQL Database
SQL Data Warehouse
DocumentDB
SQL Server Stretch Database
Redis Cache
Storage Tables
Azure Search
Intelligence
Cognitive Services
Bot Framework
Cortana
Analytics & IoT
HDInsight
Machine Learning
Stream Analytics
Data Catalog
Data Lake Analytics Service
Data Lake Store
IoT Hub
Event Hubs
Data Factory
Power BI Embedded
Hybrid Cloud
Azure AD Health Monitoring
AD Privileged Identity Management
Domain Services
Backup
Operational Analytics
Import/Export
Azure Site Recovery
StorSimple
Infrastructure Services
Compute
Virtual Machines
Containers
Storage
Blob
Queues
Files
Disks
Networking
Virtual Network
Load Balancer
DNS
Express Route
Traffic Manager
VPN Gateway
App Gateway
Datacenter Infrastructure

WHERE DID I START?

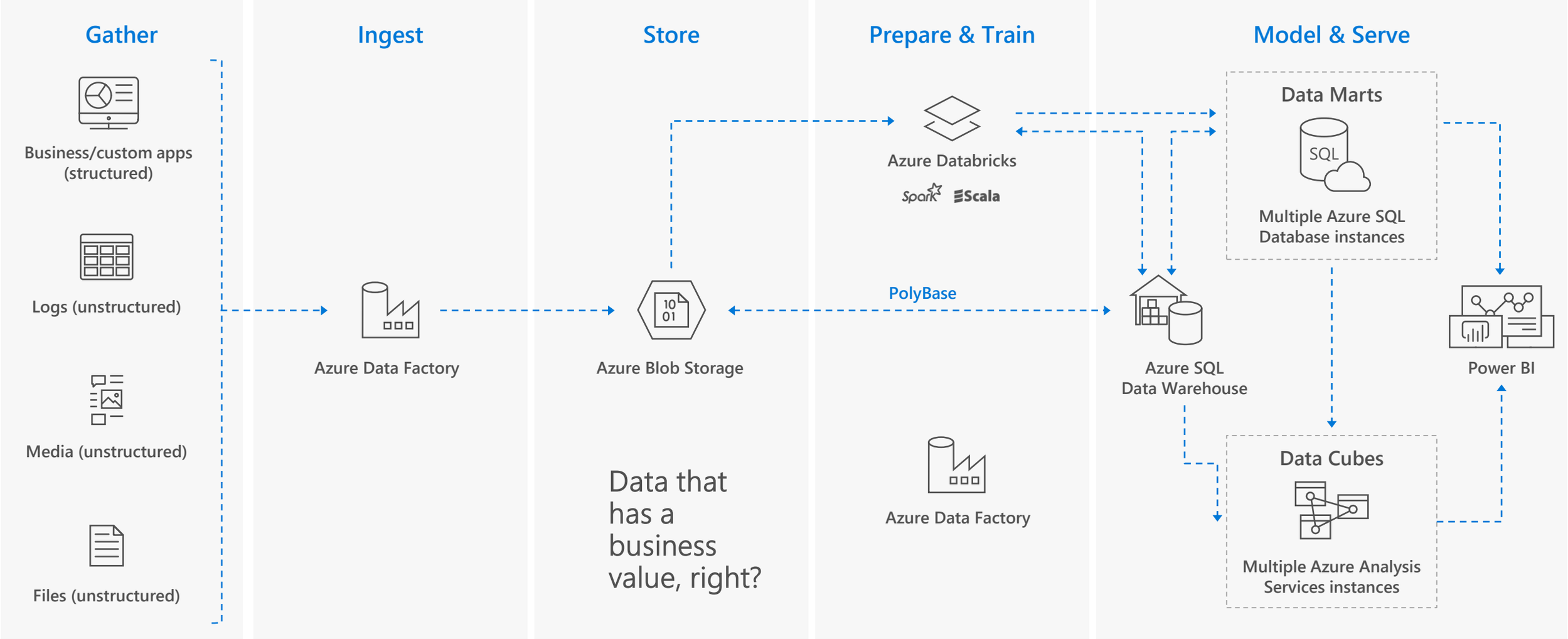
- Story about the Excel
- A Boomer’s Tail

WHERE DID I START? – HAVE WE REALLY CHANGED?



Excel as a service ☺

HUB & SPOKE ARCHITECTURE FOR BI – OUR JOBS ARE MORE COMPLEX

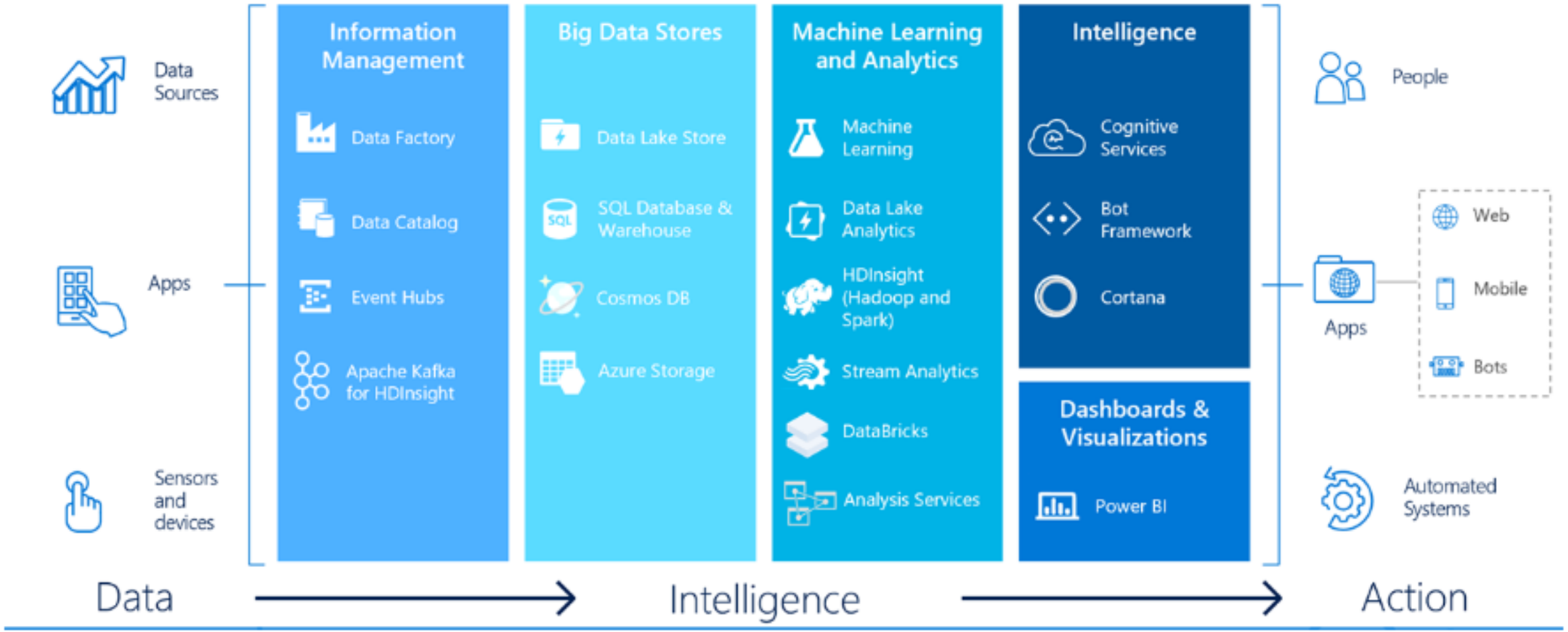


WHAT IS DATA ENGINEERING / A DATA ENGINEER?

- Boils down to being ***responsible*** for preparing data for retention and analysis
- Architect the data flow process
- Architect the technological environment
- Organize data flows
- Translates business requests into technical projects
- Data Model – SQL and No-Sql
- Real Time Data, Real Time Analysis

- Out of all that what is the Key Lesson?

AZURE DATA PLATFORM – OUR CONTEXT



WE NOW HAVE MANY DIFFERENT TOOLS → BYOT!!

DevOps	Nagios, VAGRANT, GRUNT, Xamarin
Management	CHEF, puppet labs, SALTSTACK, mist.io, libcloud, SCALR CLOUD MANAGEMENT
Applications	Joomla!, Drupal, APPRENDA, Pivotal
App frameworks & tools	php, nodeJS, JS, Ruby, APACHE Spark, eclipse, Azure Container Service
Databases & middleware	redis, CLEARDB, cloudera, MySQL, mongoDB, Couchbase, Microsoft SQL Server
Infrastructure	redhat, suse, bitnami, ORACLE LINUX, FreeBSD, docker, OPENSHIFT

WE NOW HAVE MANY DIFFERENT TOOLS → OPEN SOURCE

SQL 2019 – Not Just On-Prem Anymore

DevOps

Management

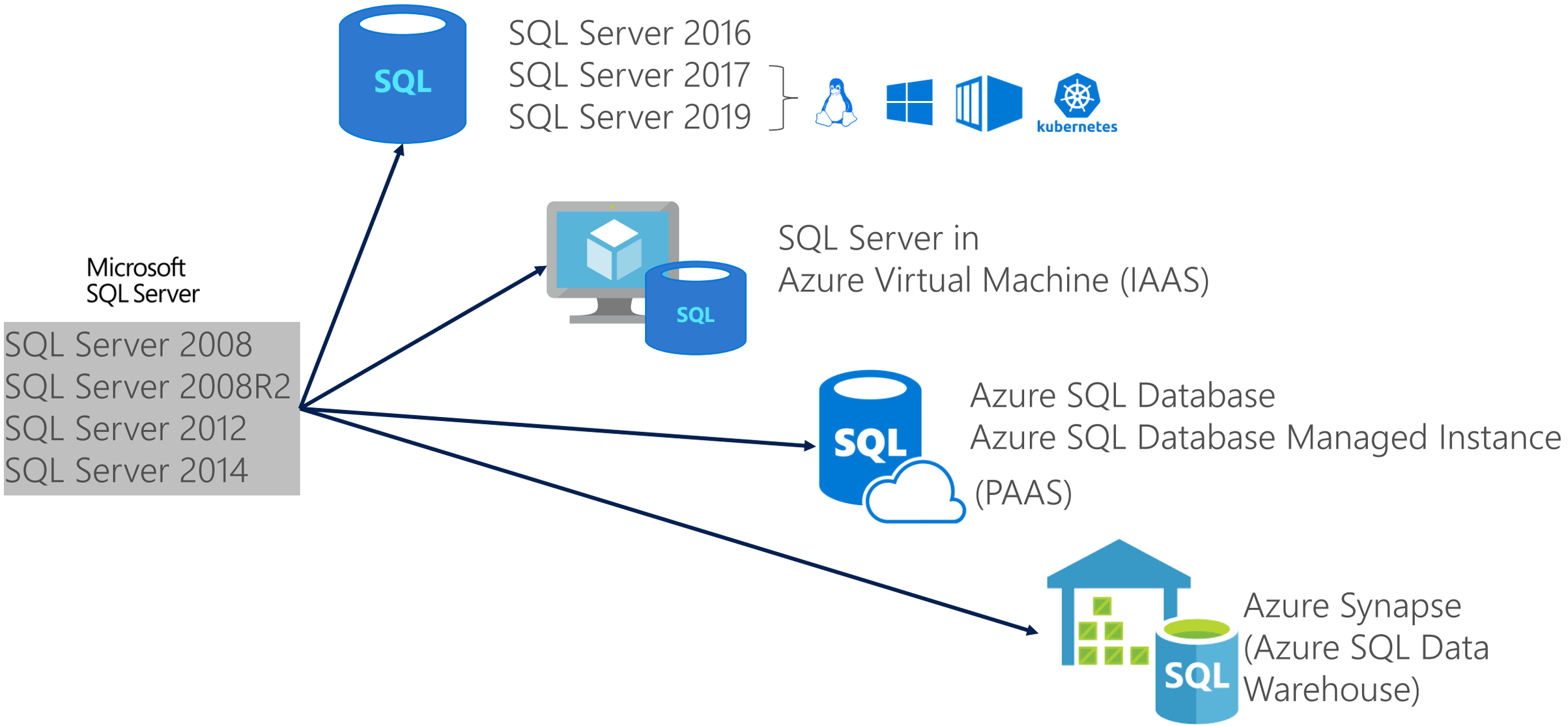
Applications

App frameworks & tools

Databases & middleware

Infrastructure

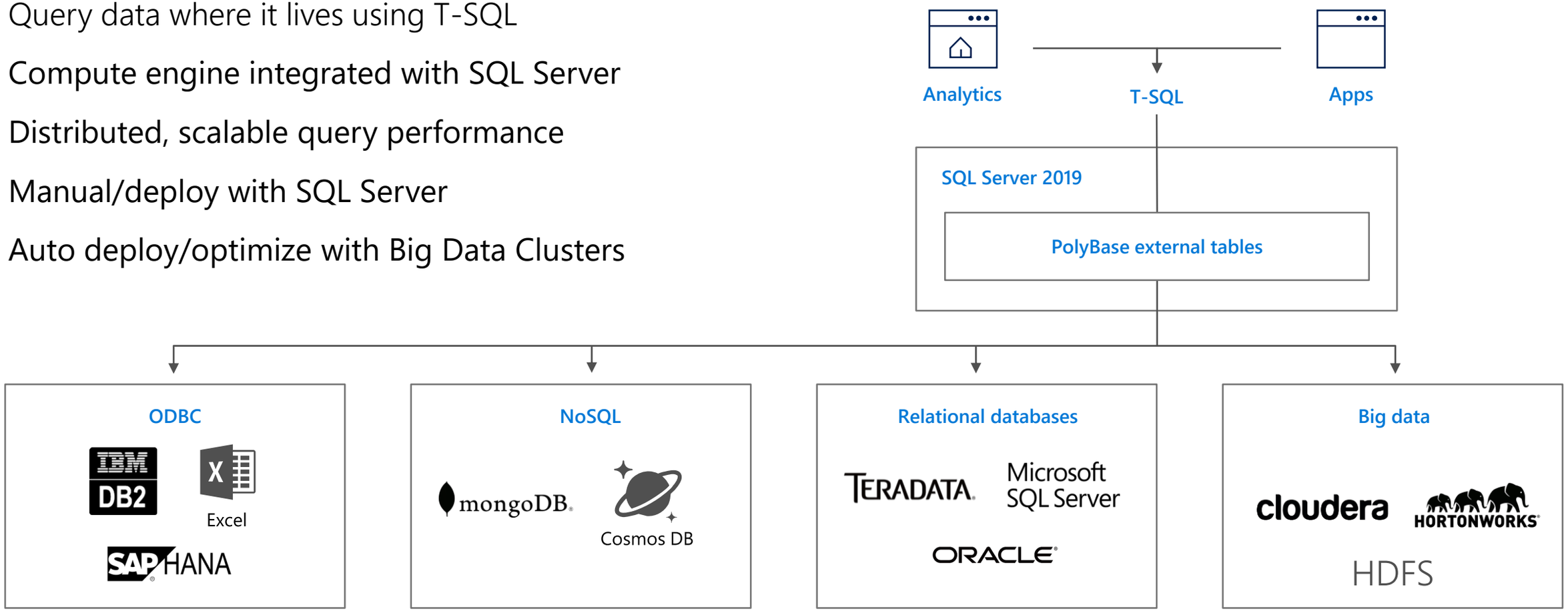
AZURE DATA MODERNIZATION CHOICES



YOU DON'T EVEN HAVE TO MOVE DATA

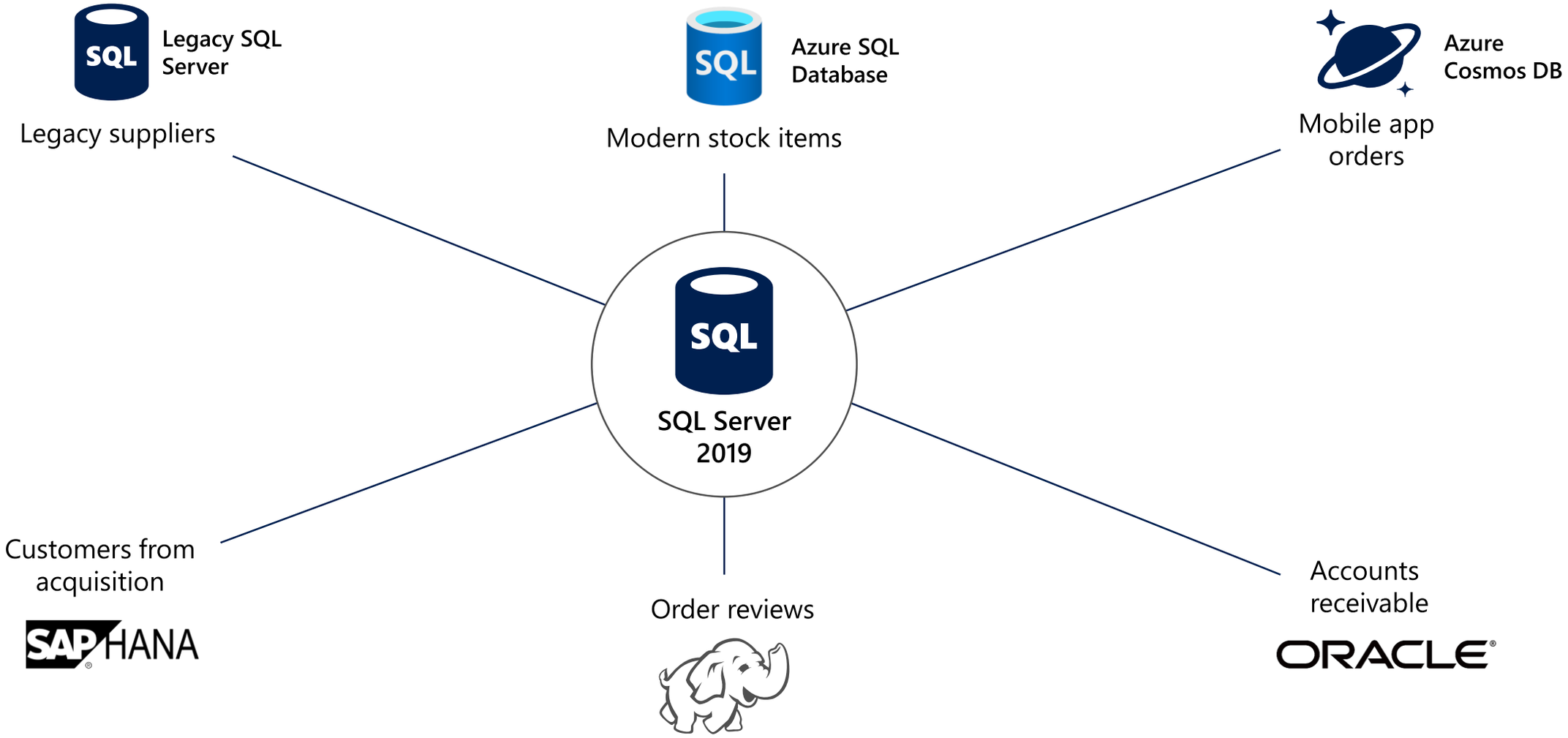
"It's all about Data Virtualization"

- ✓ It all started with T-SQL against Hadoop
- ✓ Query data where it lives using T-SQL
- ✓ Compute engine integrated with SQL Server
- ✓ Distributed, scalable query performance
- ✓ Manual/deploy with SQL Server
- ✓ Auto deploy/optimize with Big Data Clusters



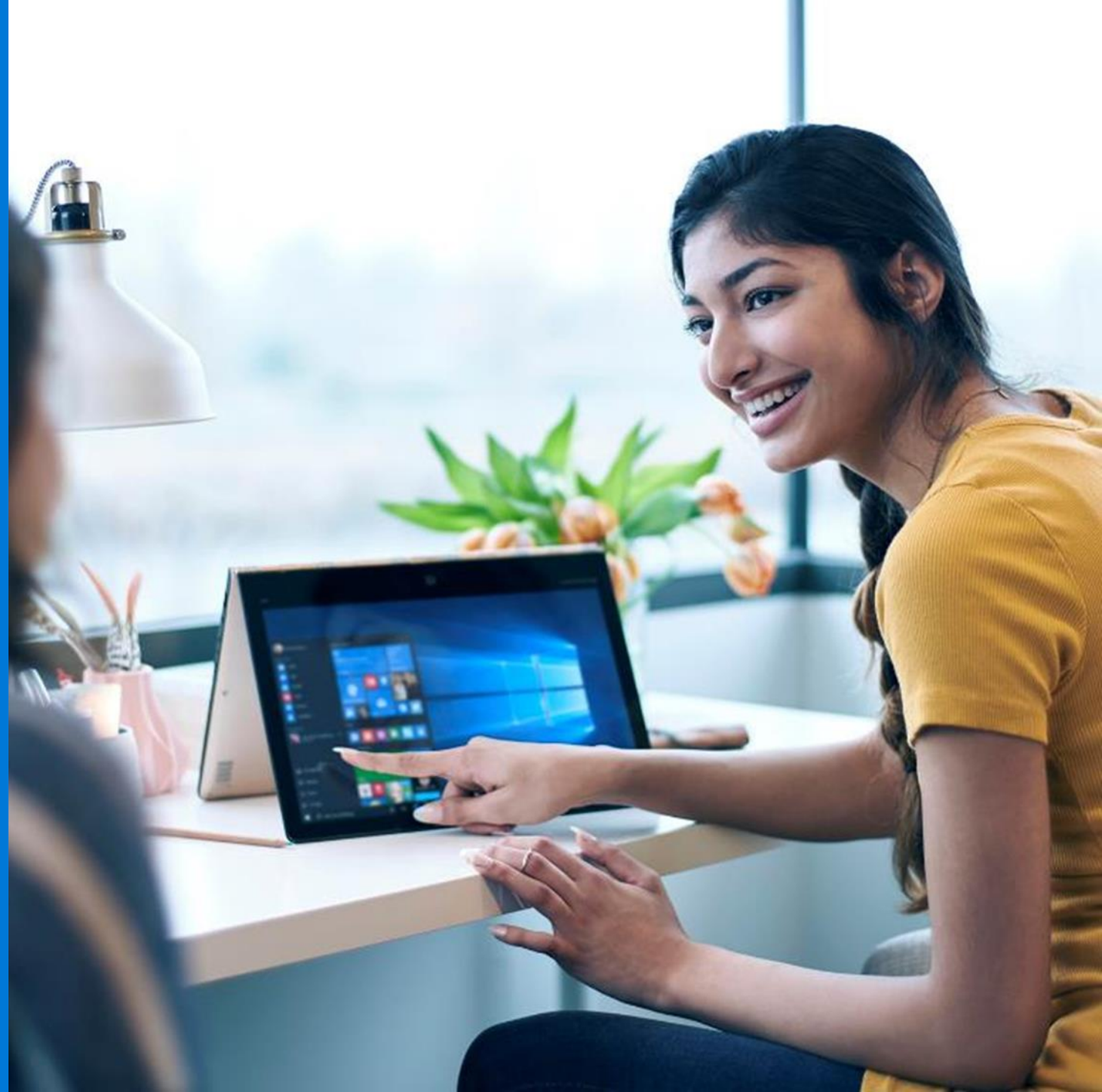
Intelligence over all data

SQL SERVER 2019 SERVES AS THE HUB FOR ALL YOUR DATA



Principles of Data Engineering

(Steve's)



PRINCIPLES ARE ALWAYS A BALANCE

- Need to evolve – things change
- Cost vs Principles
- Am I going to spend 2 weeks automating A 10 MINUTE JOB?
- Am I going to generate 2 TB log files for a process or generate alerts?
- Am I really going to keep everything?
- Use Built in tools when possible
- Do I really need to document this? (Hit, Heck ya)

DATA ENGINEERING DESIGN PRINCIPLES / BEST PRACTICES

1. Worry about security – Avoid the front page of the NY Times
2. The Data Must Always be correct and auditable
3. Document your code and design decisions
4. Design for Failure
5. Design for Process Insights
6. Design for Scale
7. Design for Change

PRINCIPLES - 1 WORRY ABOUT SECURITY

Would You click on this?

Office 365

Your messages couldn't be delivered.

Microsoft found Several Undelivered Messages

Office 365

Action Required

Recipient

Sever Congestion

How to Fix It

Retype the recipient's address, then resend the message - If you're using Outlook, open this non-delivery report message and click **Send Again**. In Outlook on the web, select this message, and then click the " **Send Again**" link located just below the message preview

Send Again

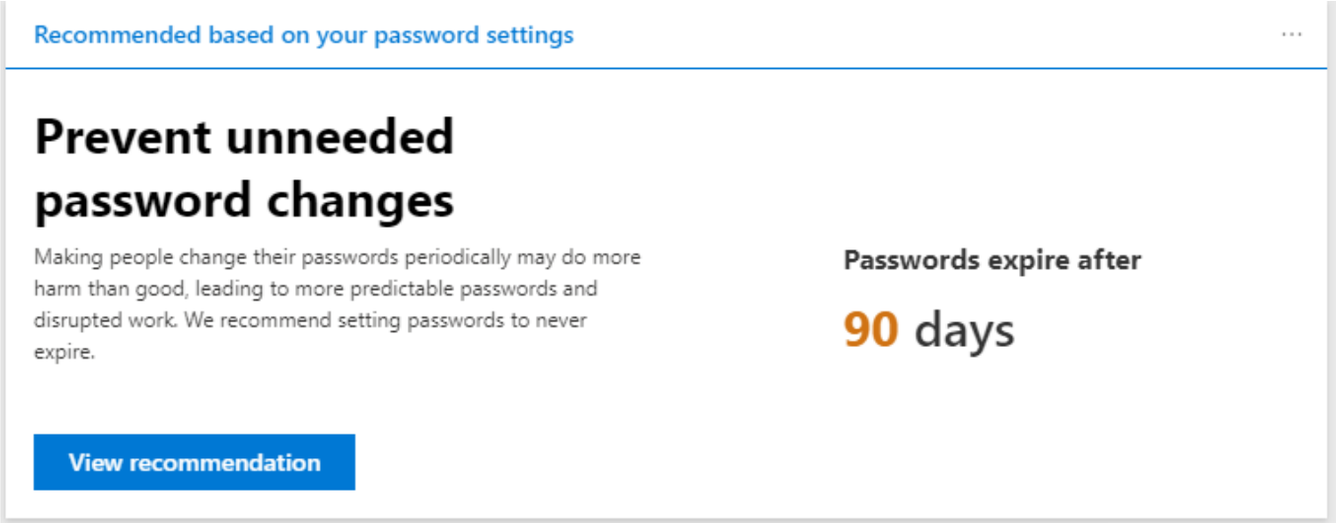
Was this helpful? Send feedback to Microsoft.

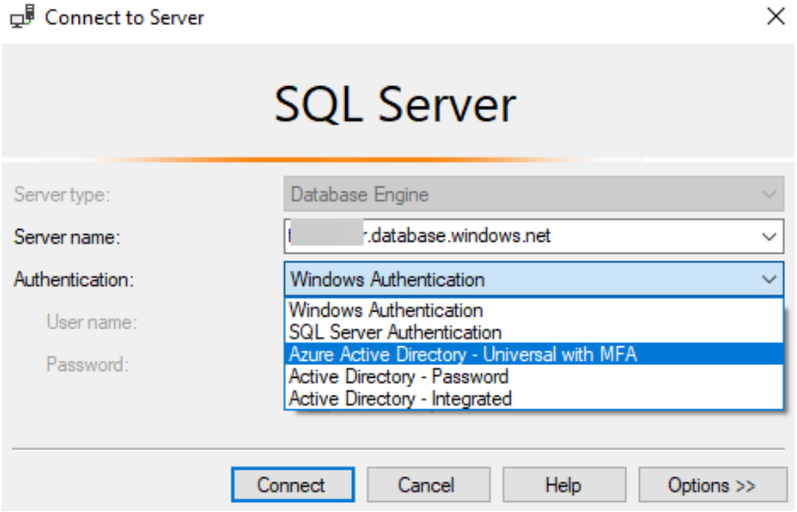
PRINCIPLES - 1 WORRY ABOUT SECURITY

PSA!!

This was interesting.

What is your policy for database passwords?





[O365 - Set the password expiration policy for your organization]
[O365 - Set up multi-factor authentication]
[Configure multi-factor authentication for SQL Server Management Studio and Azure AD]
[Using Multi-factor AAD authentication with Azure SQL Database and Azure SQL Data Warehouse (SSMS support for MFA)]

PRINCIPLES - 1 WORRY ABOUT SECURITY

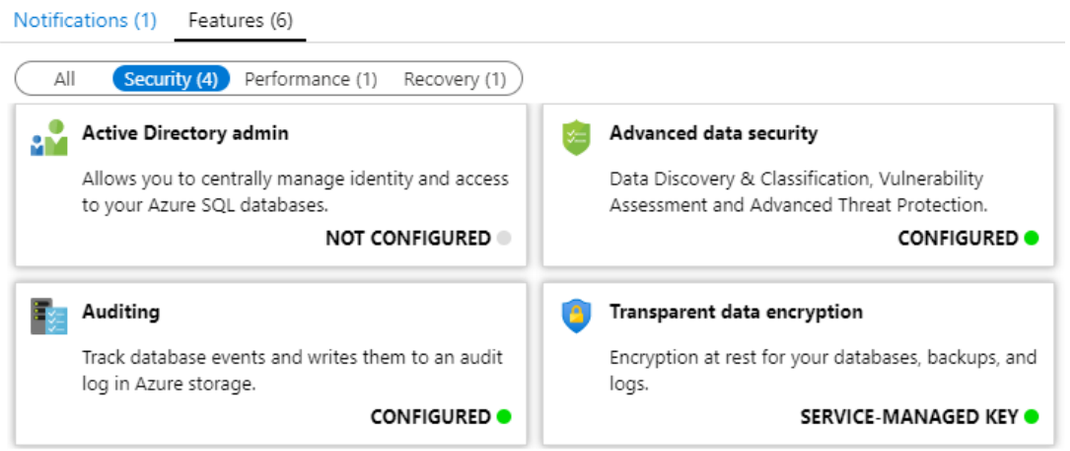
- Avoid the front page of the NY Times
- Your biggest threat is from inside !!

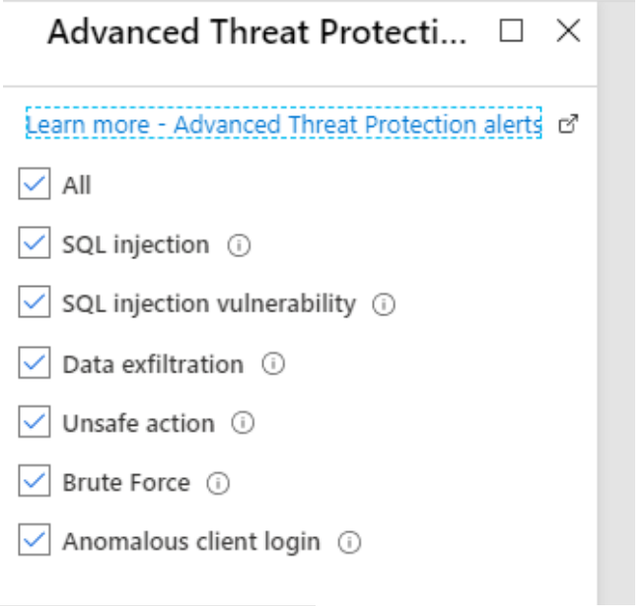
Options

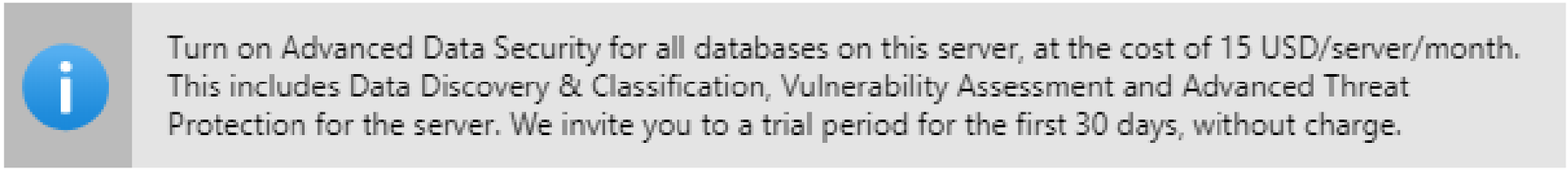
- Involve your security team at start of a project
- Volunteer for internal audits

- Use the systems available

Advanced Data Security Assessment







DEMO

PRINCIPLES - 1 WORRY ABOUT SECURITY – AZURE ASSESSMENT EMAIL

Database	Scan result	Failing checks	Passing checks	Details
AzureDemoDB	Findings	7	43	View results
AdventureworksLT	Findings	6	44	View results

A Vulnerability Assessment scan has completed on your server

Scan results

Databases scanned	Total failing checks	Total passing checks	Risk summary
2	13	87	High risk 3
			Medium risk 3
			Low risk 7

Database	Scan result	Failing checks	Passing checks	Details
AzureDemoDB	Findings	7	43	View results
AdventureworksLT	Findings	6	44	View results

Home > Vulnerability Assessment

Vulnerability Assessment

Scan | Export Scan Results | Scan History | Feedback

Total failing checks: 6

Total passing checks: 44

Risk summary: High Risk 1; Medium 2; Low Risk 3

Last scan time: Sun, 19 Jan 2020 00:38:19 UTC

Learn more: SQL Security Center; Best Practices for SQL Security

Failed (6) | Passed (44)

Filter by ID or security check | Category: All selected | Risk: All selected

ID	SECURITY CHECK	APPLIES TO	CATEGORY	RISK	ADDITIONAL IN...
VA2065	Server-level firewall rules should be tracked and maintained at a strict minimum	master	Surface area reduction	High	Should set an initial ...
VA1143	'dbo' user should not be used for normal service operation	AdventureworksLT	Surface area reduction	Medium	
VA1288	Sensitive data columns should be classified	AdventureworksLT	Data protection	Medium	
VA1145	Feature restrictions should be applied to all principals that are used to access the...	AdventureworksLT	Surface area reduction	Low	
VA2130	Track all users with access to the database	AdventureworksLT	Authentication & Au...	Low	Should set an initial ...
VA2130	Track all users with access to the database	master	Authentication & Au...	Low	Should set an initial ...

PRINCIPLES – 2 DATA MUST ALWAYS BE CORRECT AND AUDITABLE

- Who has Sign off?
- Sentiment Analysis - Do I worry about missing some records in 5million Twitter messages?
  Vs Financial reporting – Do sales by department match? Balance Sheet Totals

Options

✓ Get summary file from source owner – Auto Compare
✓ Put a Power BI Summary Report Online, make it easy
✓ Alerts etc – Be Proactive
✓ Put Summary file in same place in Azure data lake, PowerBI will update from there

PRINCIPLES – 2 DATA MUST ALWAYS BE CORRECT AND AUDITABLE

5MinuteBI

Employee Dimension Audit Report 4

This reports shows the number of Employee facts and detail that do not have corresponding Employee dimenension members 2

Pay Fact With Missing Employees 3

PayFactID	EmployeeID	NameFull	DateValue	HoursWorked	Per Hour	SalaryCost	Facility Name
100567	-99	Unknown Employee	Wednesday, January 13, 2016	7	1.83	0.00	Unknown
Total				**7**	**1.83**	**0.00**	

Year Month

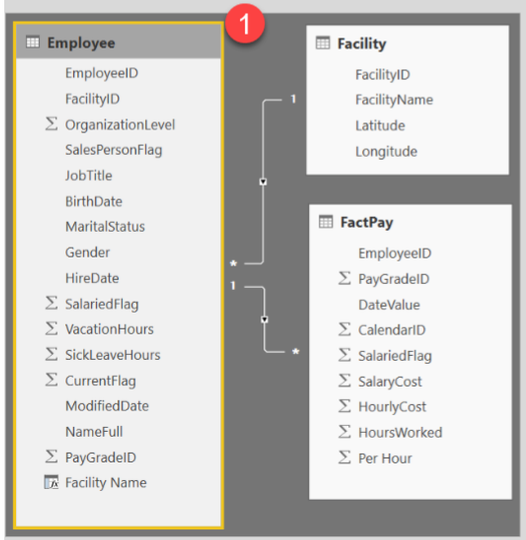
MonthFullYear 6

All

Commentary 5

There have been issues in getting the correct Employees loaded before the facts are built

Employee Dimension Audit Report



DEMO

PRINCIPLES – 3 DOCUMENT YOUR CODE AND DESIGN DECISIONS

- Be kind, Please Rewind – If not for you,
  the person that comes after you.
- (Which might be you ☺)

Options

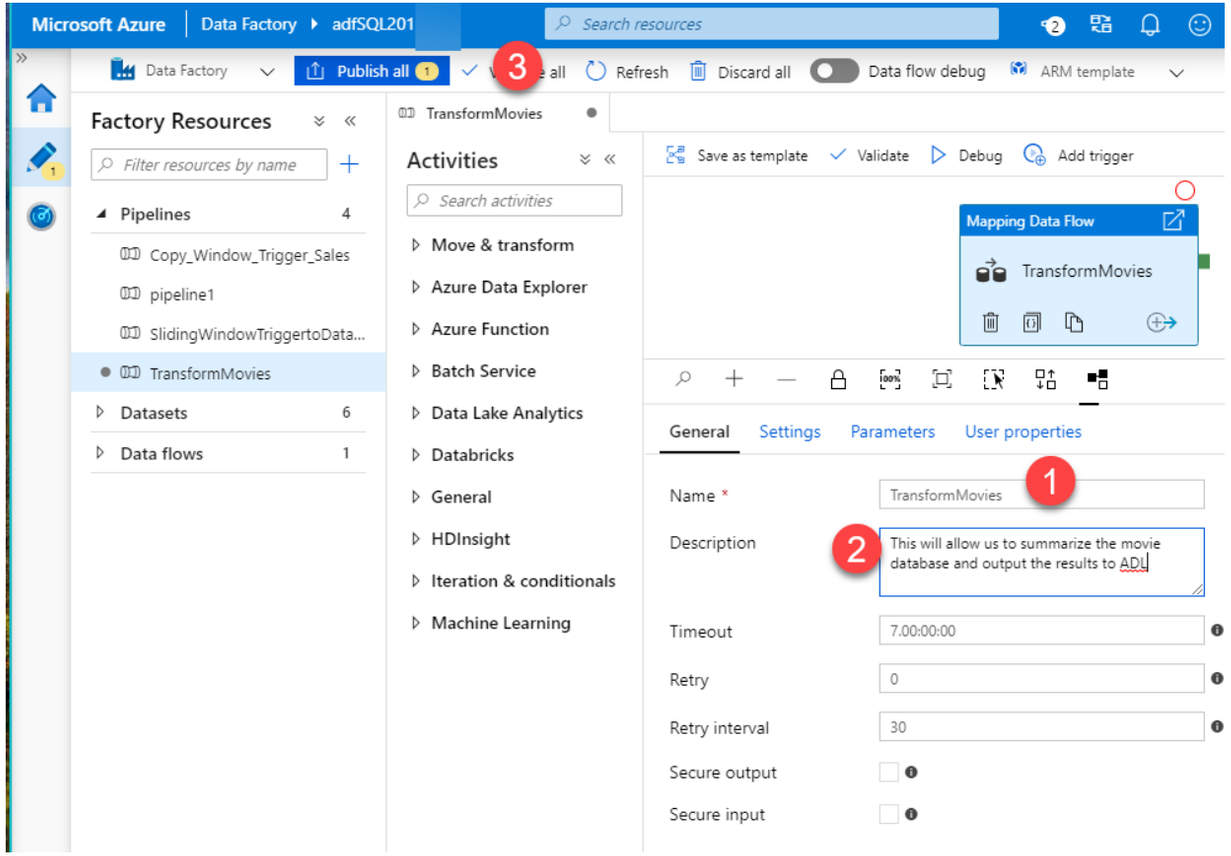
✓ Inline documentation
✓ Build Books
✓ Run Books
✓ Operations Books
✓ On the platform as notes, documentation

Q. Which is better, No Documentation
or of date documentation?

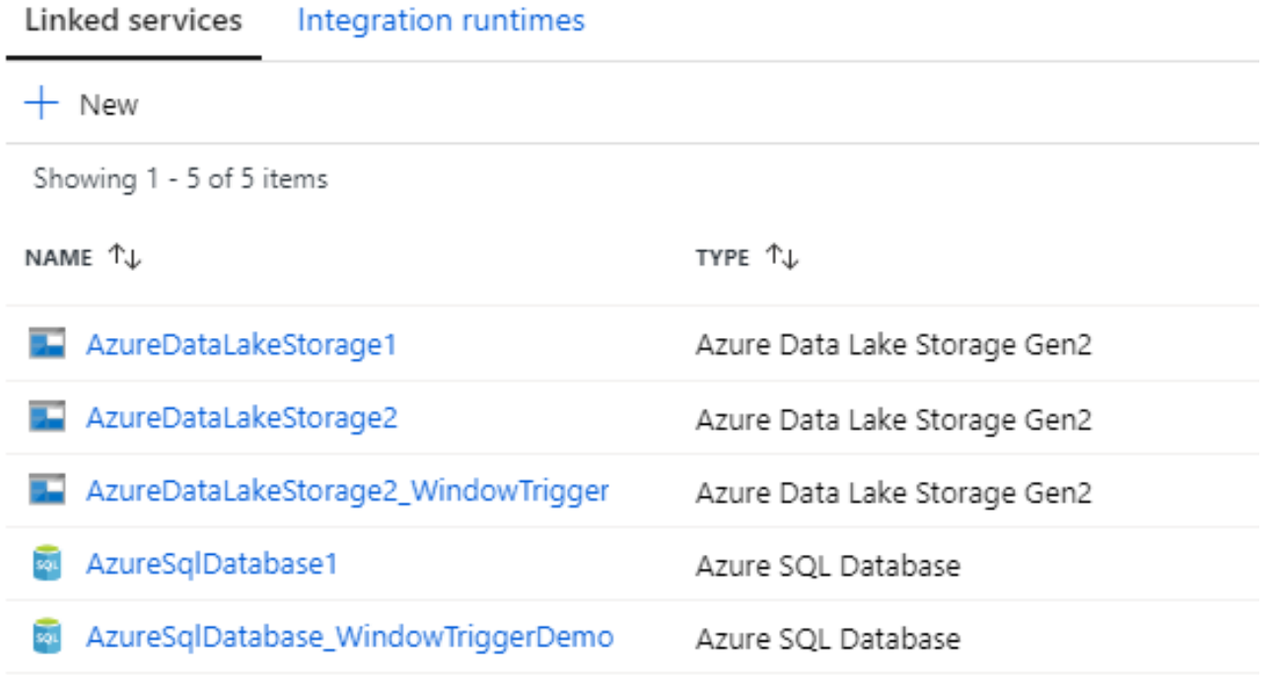
1_CreateLoggingFra...ql - not connected | CreateSQLObjects_S...ql - not connected

```
/*
##  TSQL SCRIPT: 1_CreateLoggingFramework.sql
##
##  Description:  This creates the table and stored procedure for the audit framework
##
##
##
##  Parameters : Varies
##
##  License:
##
##  Repository:
##
##  Documentation on the  package
##
##
##
##
##  Date                    Developer           Action
##  ------------------------------------------------------------------------
##  Jan 23, 2009            Steve Young         Final Version 1
##  Jun 04, 2013            Steve Young         Final Version 2 (SQL 2012)
##  Oct 07, 2017            Steve Young         Changes for SQL Azure & SQL 2016
##
##
##
##
##  TODO:
##  1.
##
##  Testing:
##  ---------------------------------------------------------------
##    1.
##
##  Execute:
##      1.
*/

--  ==================================================================
--              Settings & Declarations
--  ==================================================================

/*    ==Scripting Parameters==



*/

--Comment Out for SQL Azure
USE [ADFTest_AzureImport]
GO


--  ==================================================================
--              Create Audit table
```

PRINCIPLES – 3 DOCUMENT YOUR CODE AND DESIGN DECISIONS



- Which is better?



PRINCIPLES – 4 DESIGN TO SURVIVE FAILURE

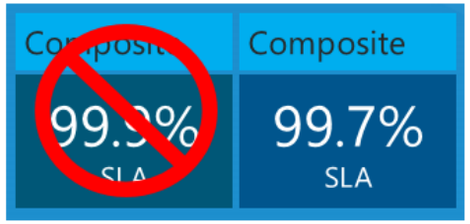
- What happens if your process Fails, stops, is not complete.
- Rollback, Restart,
- What about Disaster Recovery –
- What about the Roll Back?
- Is it a failure or an unexpected result?
- Is your plan stored in the failed region?

- You can have a failure and still meet your SLA.

Availability %	Downtime per year	Downtime per month*	Downtime per week
90% ("one nine")	36.5 days	72 hours	16.8 hours
99% ("two nines")	3.65 days	7.20 hours	1.68 hours
99.9% ("three nines")	8.76 hours	43.2 minutes	10.1 minutes
99.99% ("four nines")	52.56 minutes	4.32 minutes	1.01 minutes
99.999% ("five nines")	5.26 minutes	25.9 seconds	6.05 seconds
99.9999% ("six nines")	31.5 seconds	2.59 seconds	0.605 seconds

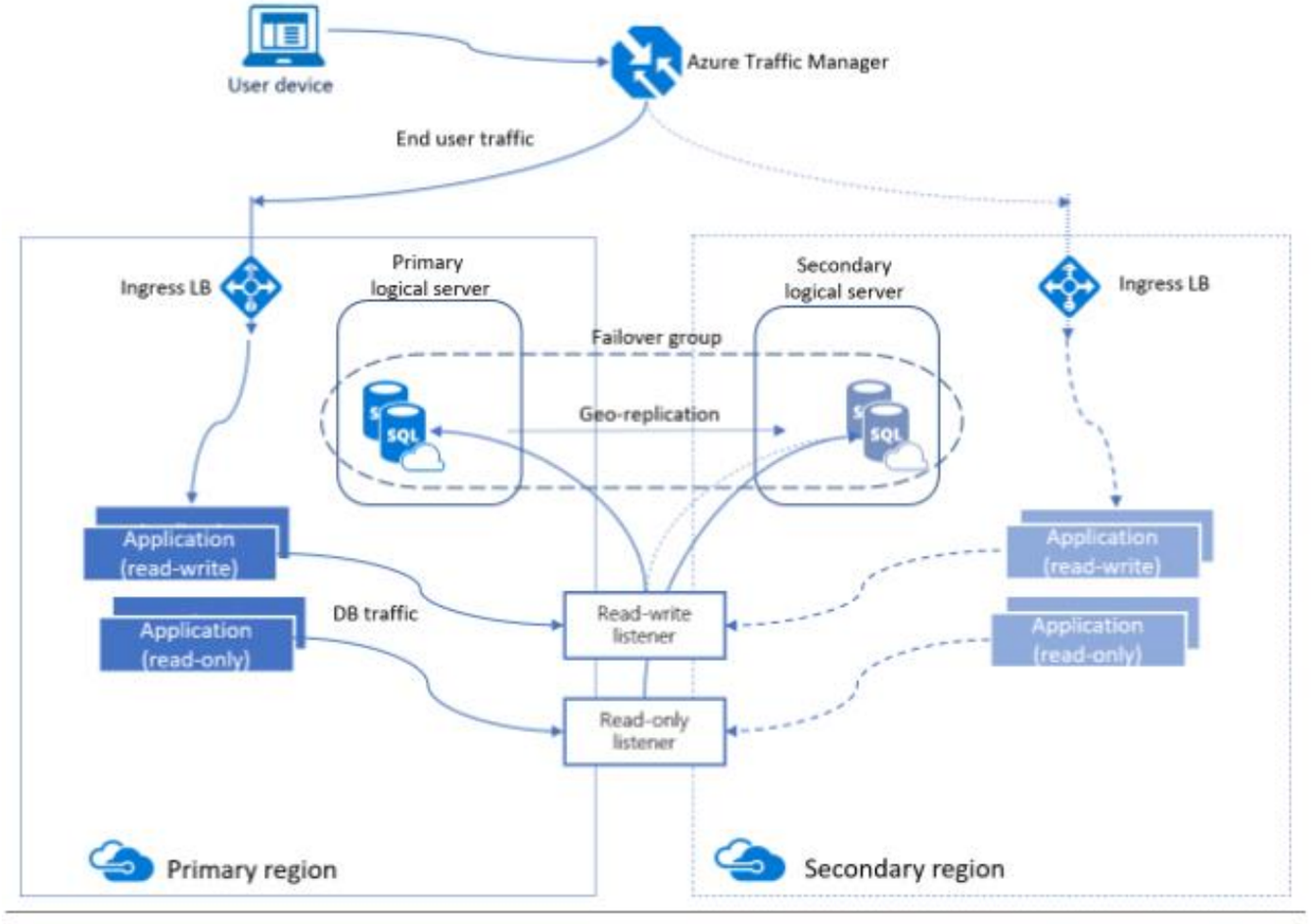
- ✓ Build a prototype – Test it !!
- ✓ Use Source Control, copy to another region !!!!
- ✓ Don't be afraid to redesign
- ✓ How do you communicate?
- ✓ Test Solutions – minimize failure, test failures
- ✓ Do you have Test Data
- ✓ Story boards are useful to talk out and plan options.

- 3 components have 99.9 SLA
- Your composite SLA <> 99.9%



PRINCIPLES – 4 DESIGN TO SURVIVE FAILURE

- The following diagram illustrates a typical configuration of a geo-redundant cloud application using multiple databases and auto-failover group.

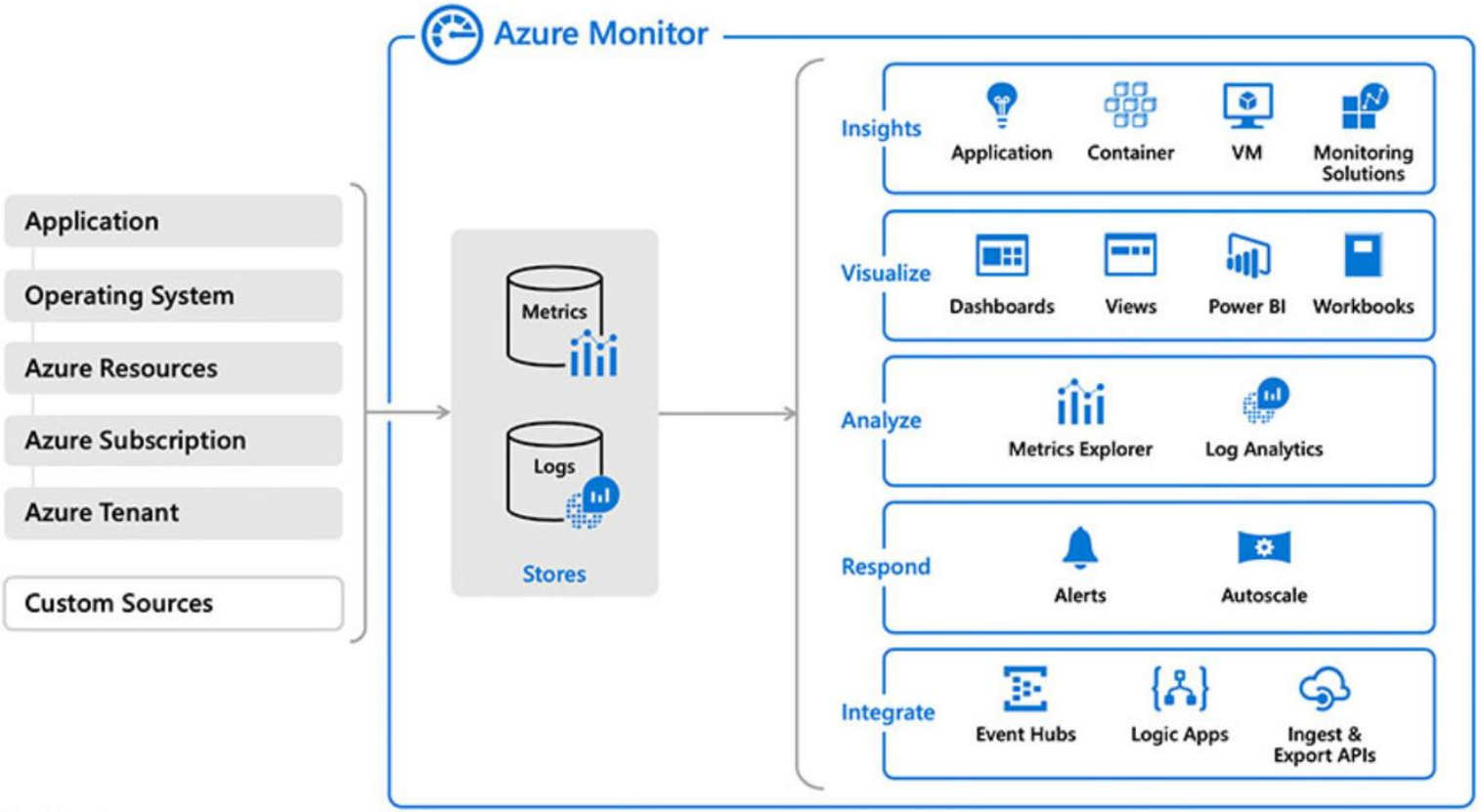


https://docs.microsoft.com/en-us/azure/sql-database/sql-database-auto-failover-group?tabs=azure-powershell

PRINCIPLES – 5 DESIGN FOR PROCESS INSIGHTS

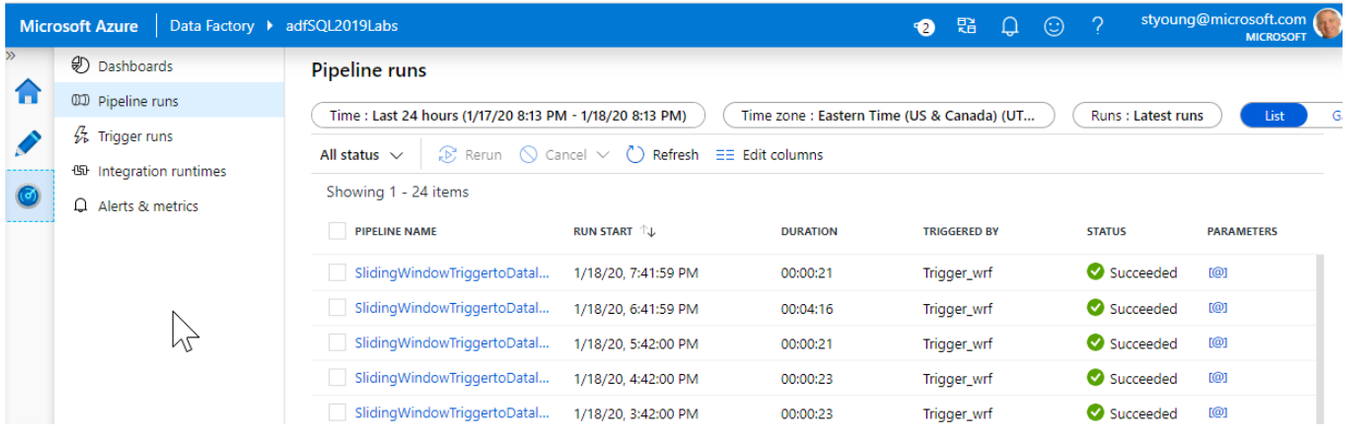
- Users complain the process takes too long!!!!
- More than 1 user needs to know what happened

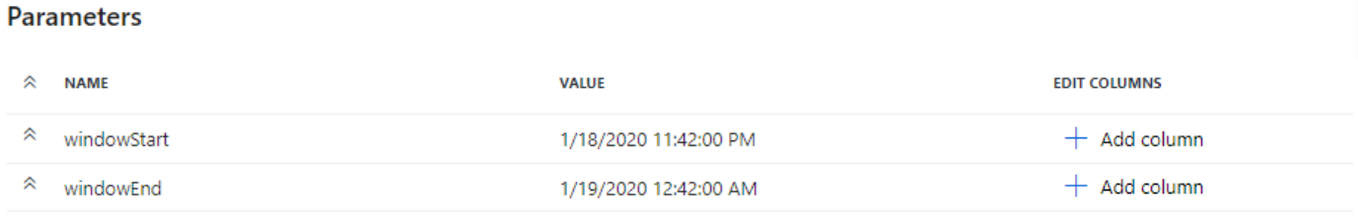
✓ Azure Monitor
✓ Azure Log Analytics
✓ Azure Data Factory Monitor
✓ Custom logging and reporting

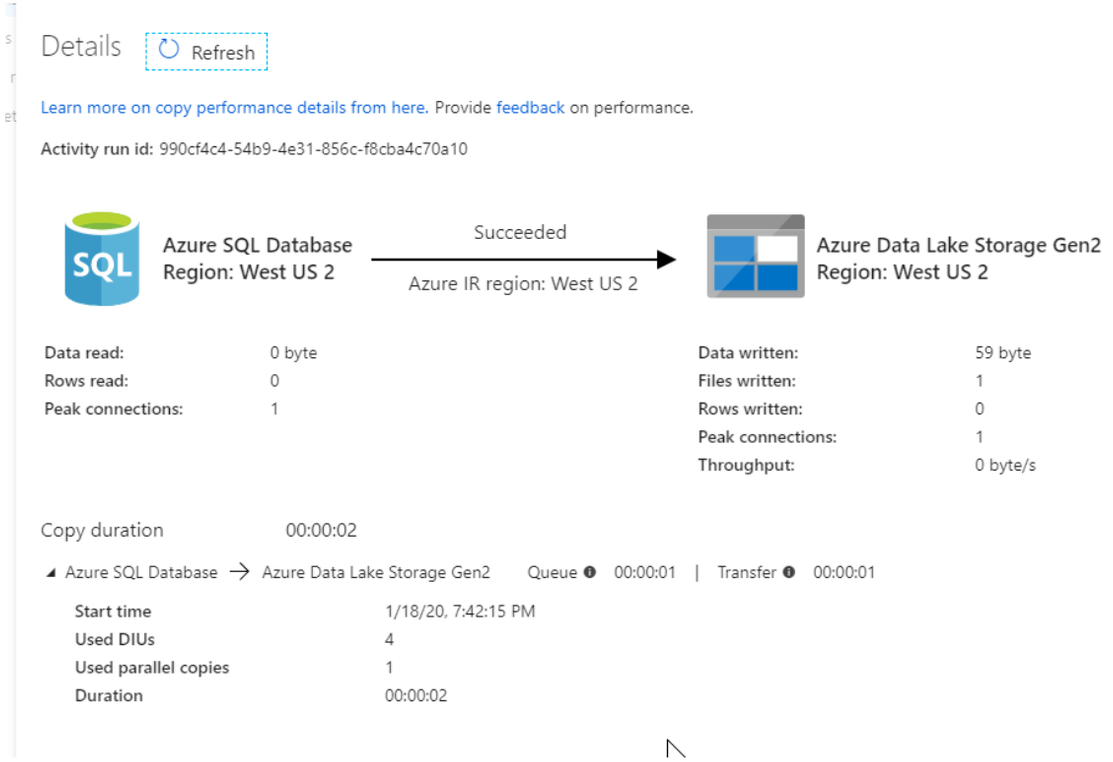


PRINCIPLES – 5 DESIGN FOR PROCESS INSIGHTS - WALKTHROUGH

- Azure Data Factory monitor.







DEMO

PRINCIPLES – 5 DESIGN FOR PROCESS INSIGHTS

- **Azure Log Analytics** - Data Factory stores pipeline-run data for only 45 days. Use Monitor if you want to keep that data for a longer time. With Monitor, you can route diagnostic logs for analysis. You can also keep them in a storage account so that you have factory information for your chosen duration.

- Save your diagnostic logs to a storage account for auditing or manual inspection. You can use the diagnostic settings to specify the retention time in days.
- Stream the logs to Azure Event Hubs. The logs become input to a partner service or to a custom analytics solution like Power BI.
- Analyze the logs with Log Analytics.

https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor

PRINCIPLES – 5 DESIGN FOR PROCESS INSIGHTS

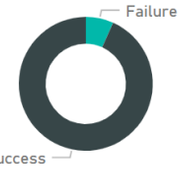
Custom monitoring

System Alerts and Audit notifications

Current Success...
28

Current Failures
2

Latest Run Summary



Batch # All | Success All | RunStage All | PkgName All

Current Failures Details

MB#	PkgName	ExecStartDT	AuditID	SQLErrorMessage	RunStage	Notes
38	usp_ErrorPackage_DEV_LOAD_LeapKnee_1YR_From_Blob_storage	11/14/2017 4:26:30 PM	41	Cannot bulk load because the file "LeapKnee1_1YR.csv" could not be opened. Operating system error code 1117(The request could not be performed because of an I/O device error.).	Package Start	This is a test batch entry with table LeapKnee1_1YR.csv
20	usp_Test_Failure_DEV_LOAD_LeapKneeBaseline_From_Blob_storage	11/8/2017 9:02:30 PM	1		Package Start	This is a test batch entry with table LeapKnee1_Baseline.csv

Package Run Details

Batch # All | Success All | RunStage All | PkgName All | ExecStartDT 11/8/2017 11/30/2017

Details

Success	MB#	AuditID	PkgName	RunStage	ExecStartDT	Notes	SQLErrorMessage	CommandTxt
N	95	163	[usp_Stage_To_Prod_LeapKnee_1YR]	Begin Procedure	11/30/2017 12:52:54 PM	Stage to Production		
N	95	165	[usp_Stage_To_Prod_LeapKnee_1YR]	Begin Transforms	11/30/2017 12:52:54 PM	Stage to Production		
N	95	164	[usp_Stage_To_Prod_LeapKnee_1YR]	Insert Success	11/30/2017 12:52:54 PM	Stage to Production		
Y	95	166	[usp_Stage_To_Prod_LeapKnee_1YR]	Process Complete	11/30/2017 12:52:54 PM	Stage to Production		

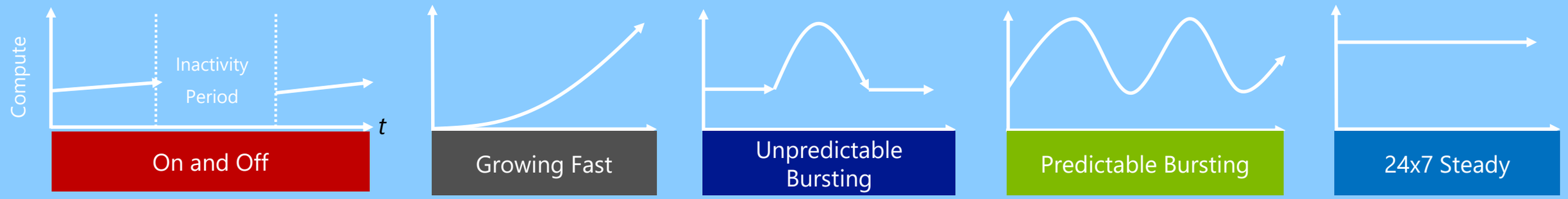
DEMO

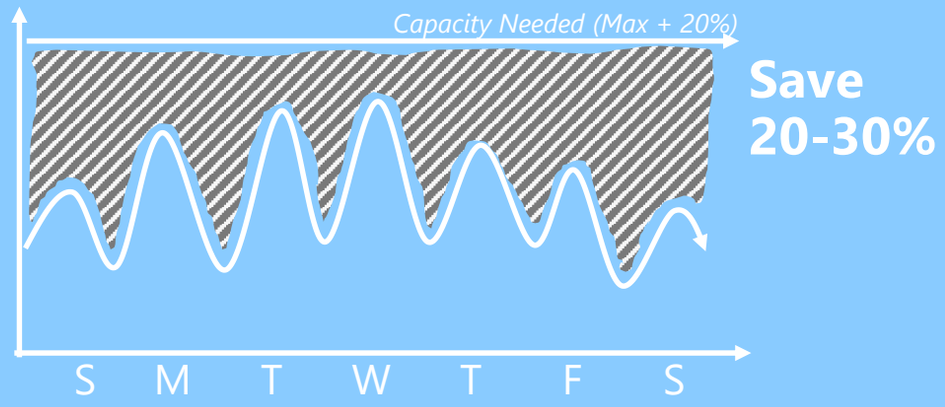
PRINCIPLES – 6 DESIGN FOR SCALE

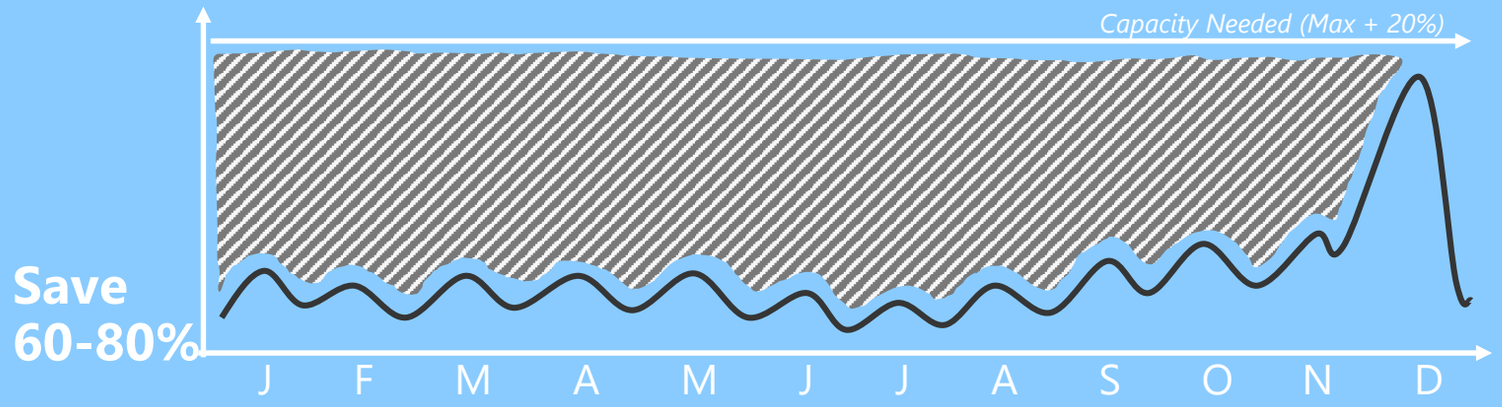
- Incremental Load / Historical Load
- Archive data? How much history do you need to keep.

✓ Azure data Lake and Polybase comes in handy
✓ Scale SQL up and down

Saving cost – IF you are careful !!





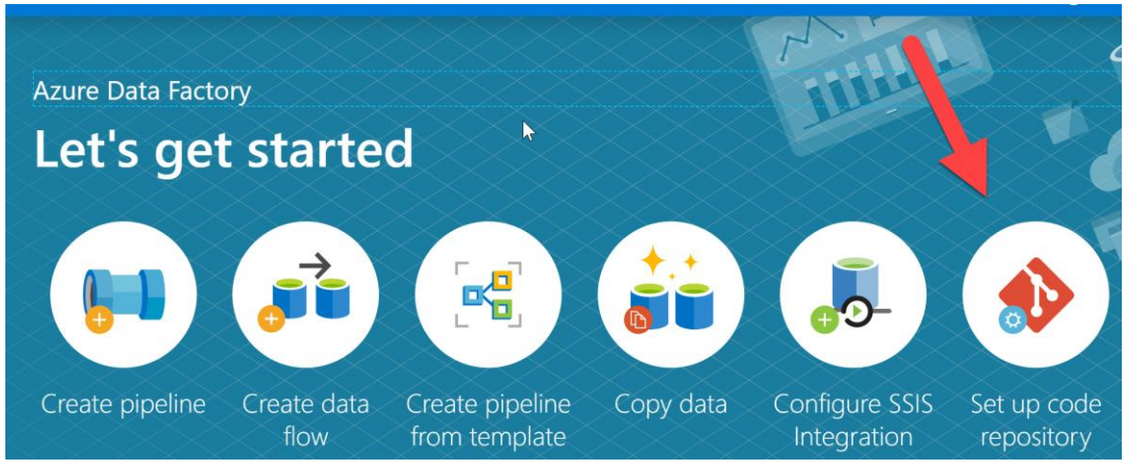


On and Off (30%)	Growing Fast (15%)	Unpredictable Bursting (25%)	Predictable Bursting (20%)	24x7 Steady (10%)

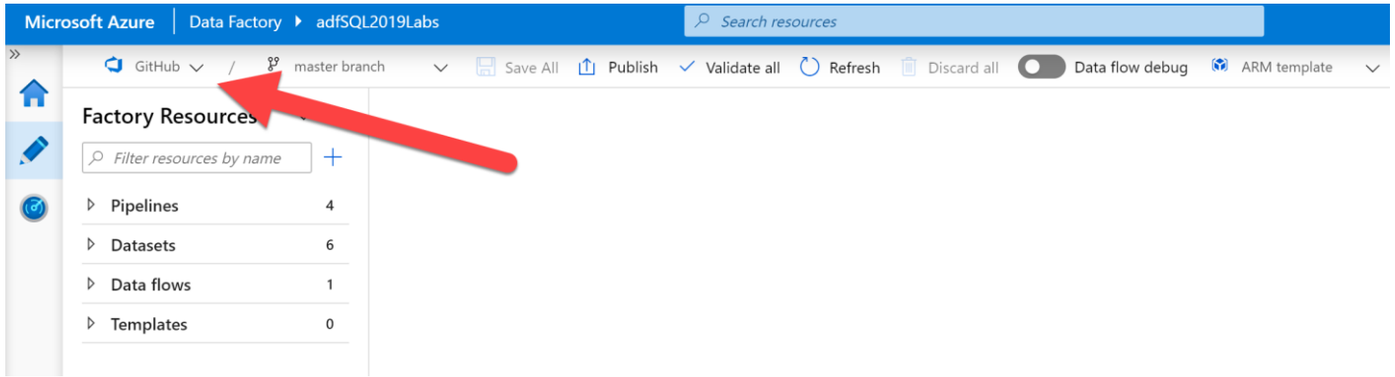
Your Application Portfolio – What Does it Look Like..?

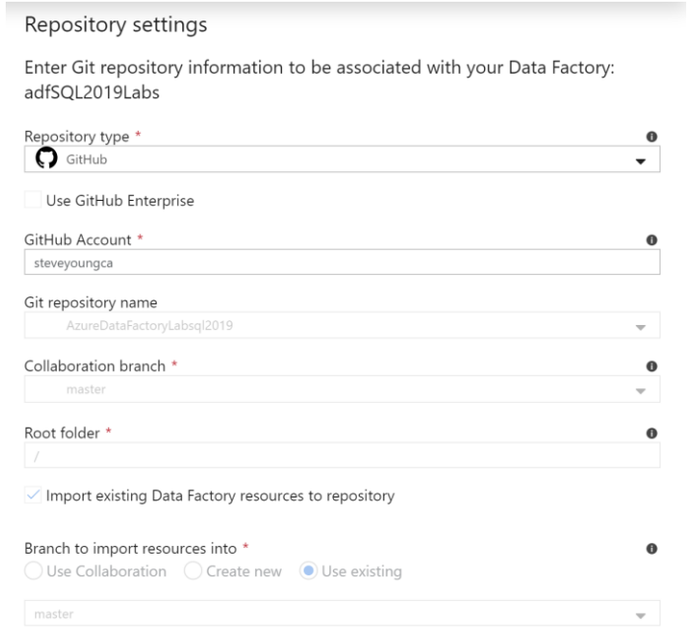
PRINCIPLE – 7 DESIGN FOR CHANGE

- Everything Changes
- Speed and Control are not exclusive
- Be able to roll back, no matter how difficult



✓ Don't box yourself in
✓ You can also use GITHUB for the source control etc.
✓ Used instead of Azure Data Factory

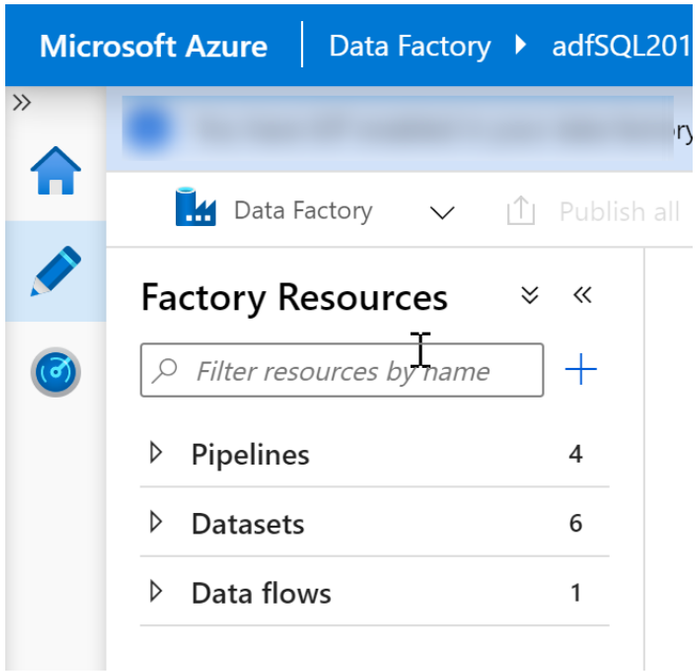


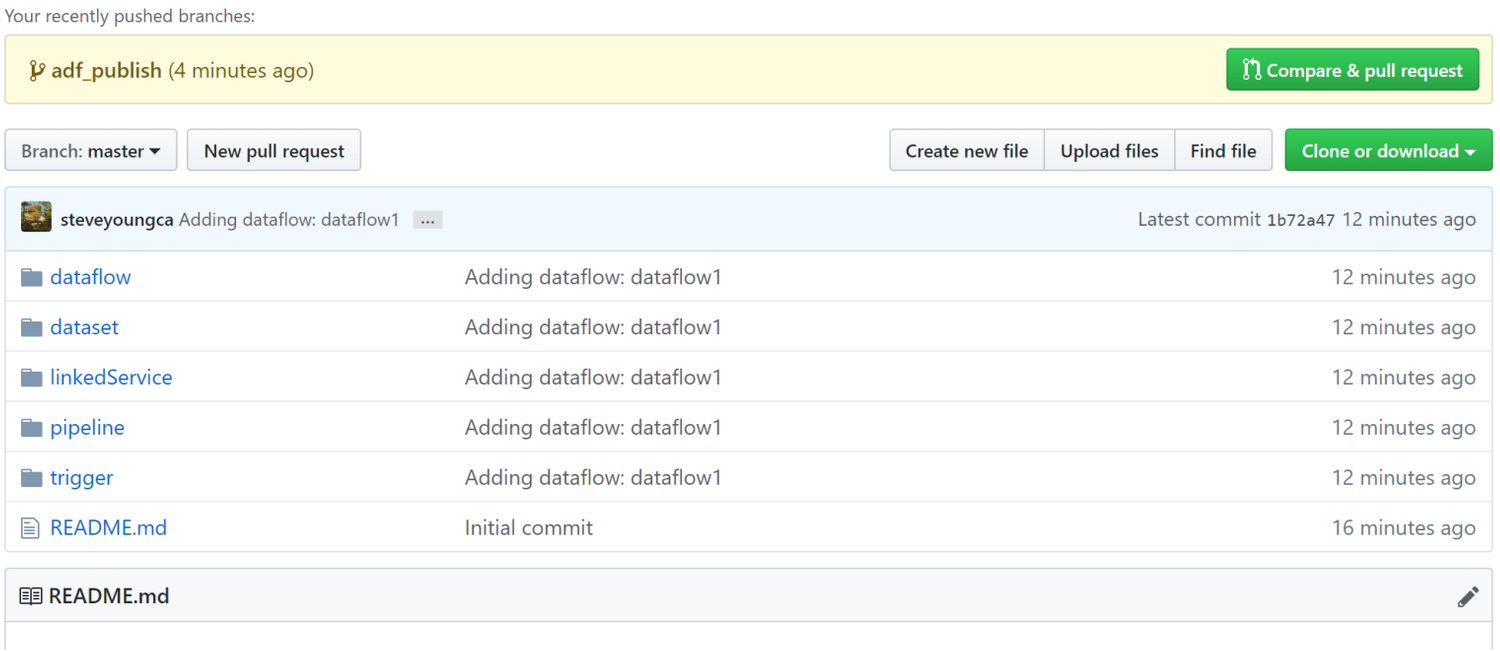


1TB Restore Story

PRINCIPLE – 7 DESIGN FOR CHANGE

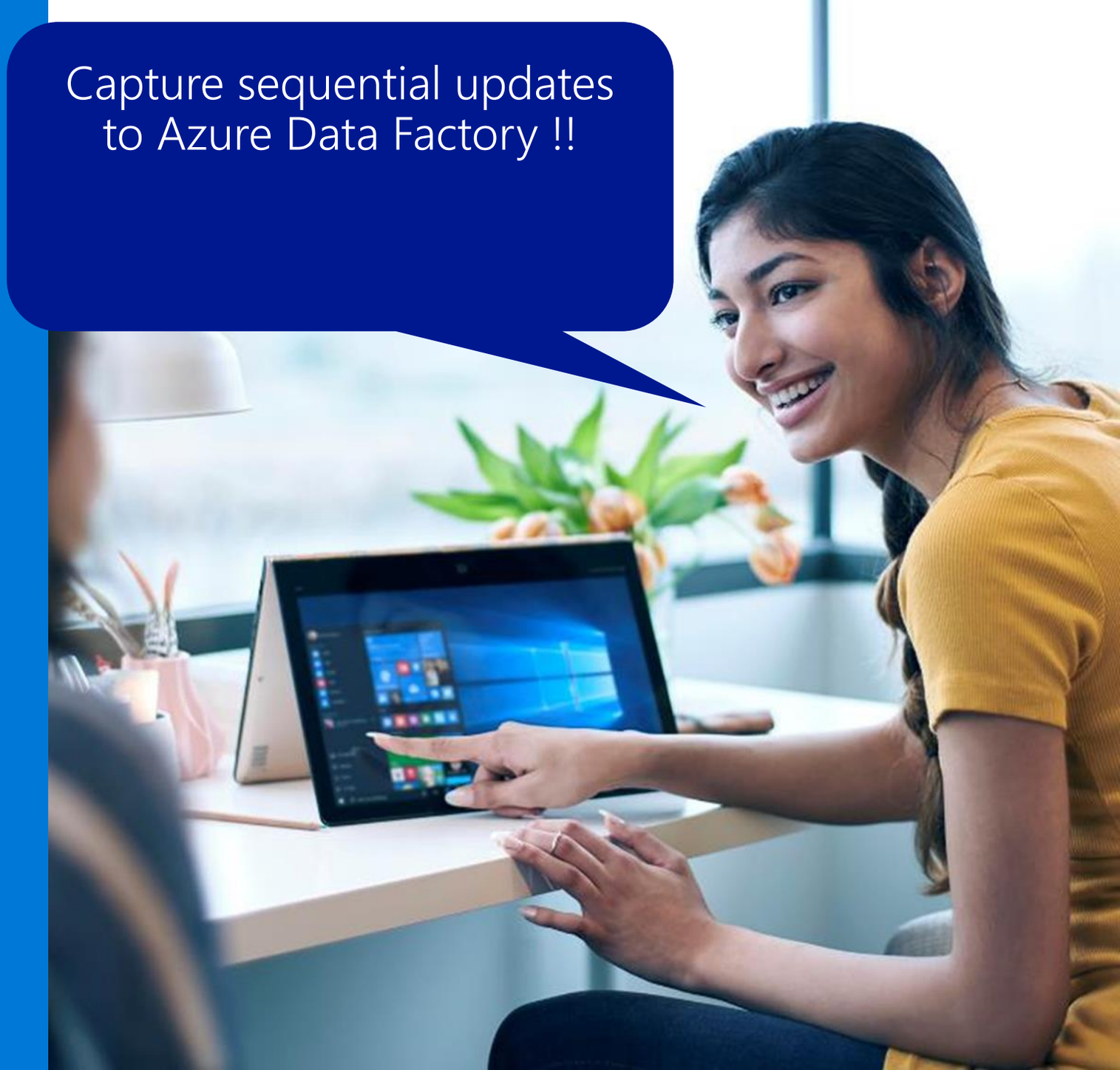
✓ Azure Data Factory vs GitHub





https://github.com/steveyoungca/AzureDataFactoryLabsql2019

Demo 1
Azure Data Factory Sliding Window



CLOUD SCALE ANALYTICS

Ingest

Azure Data Factory

Copy data from over 75 data sources

Drag-n-drop UI

Execute your SQL Server Integration Services packages in Azure

Store

Azure Data Lake Storage

Industry leading performance

POSIX-compliant, fine grained access control

Tiered multi-modal storage

Prep

Azure Databricks

Fast, easy, collaboration Apache® Spark-based™ analytics platform

Native integration with Azure services

Enterprise-grade Azure security

Polybase

Serve

Azure SQL Data Warehouse

Unlimited scale & performance

Flexibility to fit your needs

Enterprise-grade DW at the click of a button

Reporting

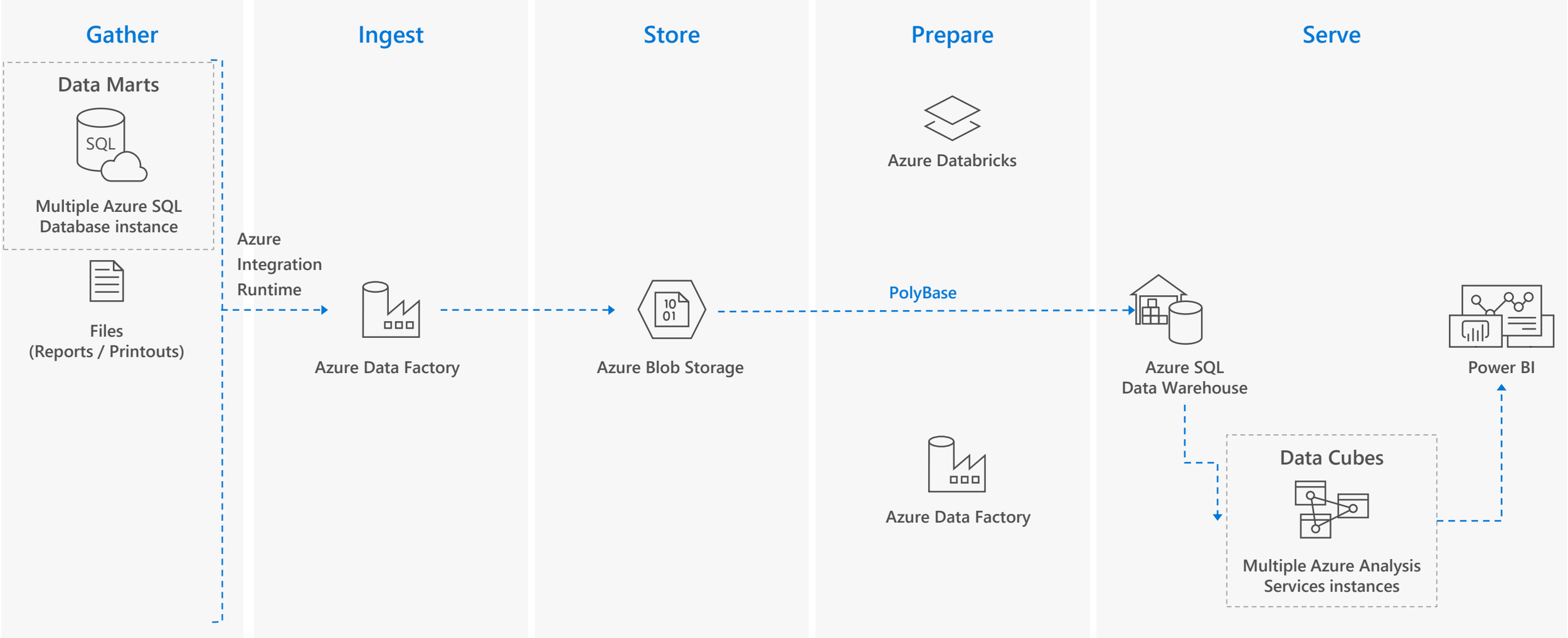
Power BI

Get started in seconds

Ask questions, get answers

Access intelligence anywhere

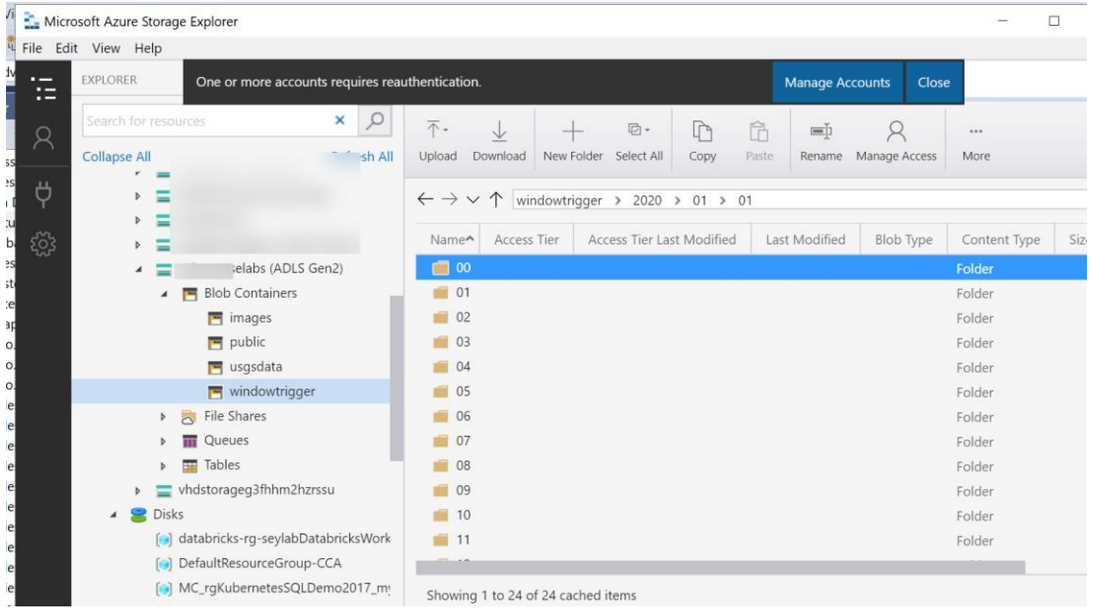
WHERE DID I START?

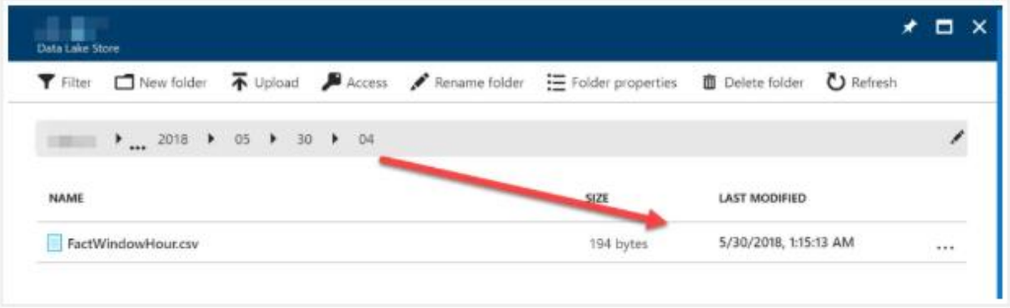


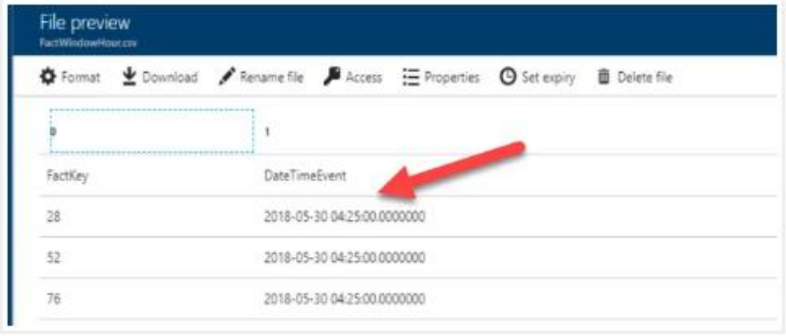
DEMO 1

Results | Messages

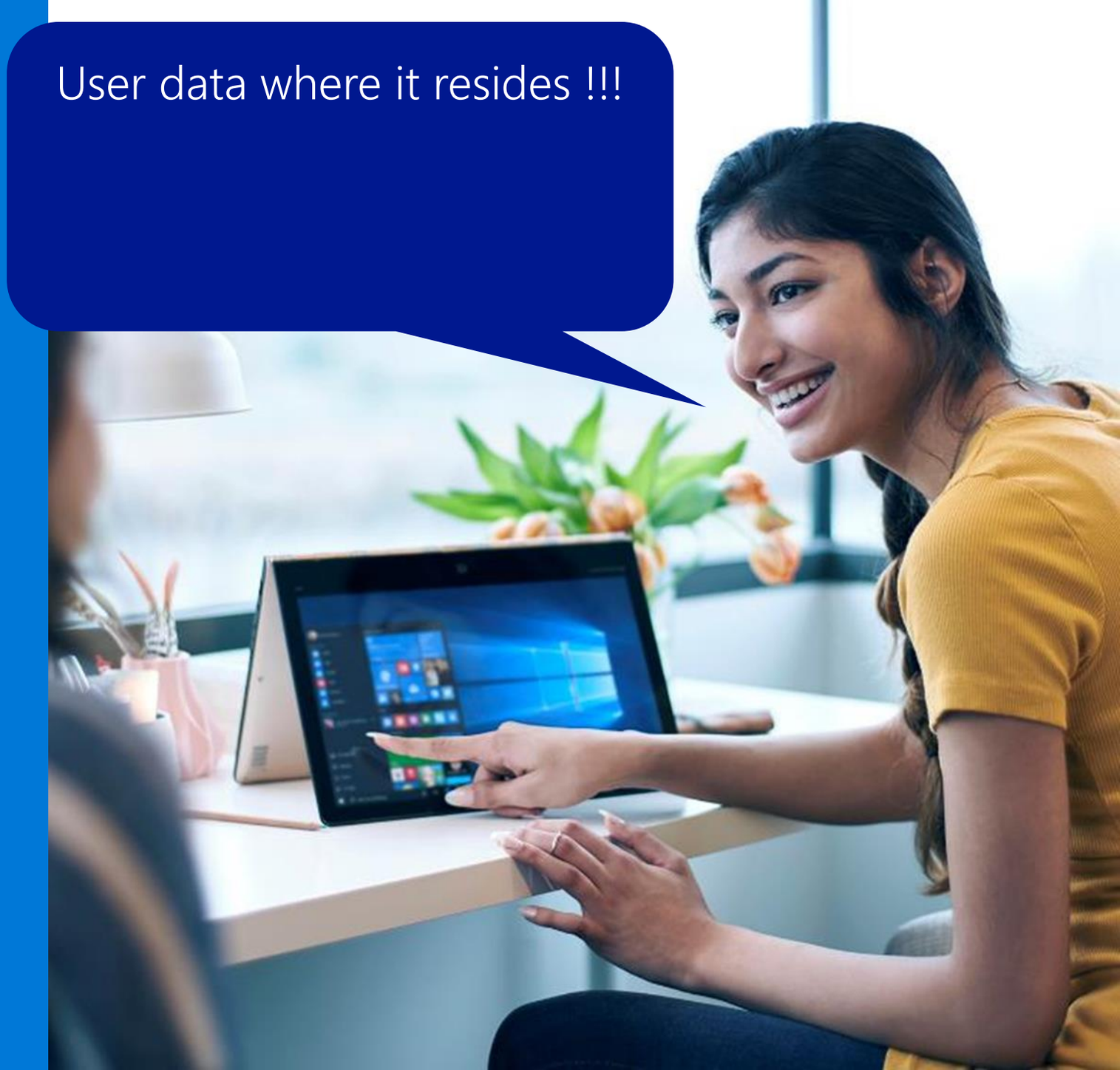
	FactKey	DateTimeEvent	Dimension1Key	Dimension2Key	MyNotes
1	1	2019-12-06 20:38:38.833	1	2	FirstLoad1
2	2	2019-12-06 20:38:38.843	1	2	FirstLoad1
3	3	2019-12-06 20:38:39.893	1	2	FirstLoad1
4	4	2019-12-06 20:38:39.900	1	2	FirstLoad1
5	5	2019-12-06 20:38:40.797	1	2	FirstLoad1
6	6	2019-12-06 20:38:40.800	1	2	FirstLoad1
7	7	2019-12-06 21:10:17.890	1	2	FirstLoad1
8	8	2019-12-06 21:10:17.903	1	2	FirstLoad1
9	9	2019-12-06 21:10:18.890	1	2	FirstLoad1
10	63	2020-01-12 19:05:08.530	1	2	FirstLoad1
11	64	2020-01-12 19:05:08.537	1	2	FirstLoad1







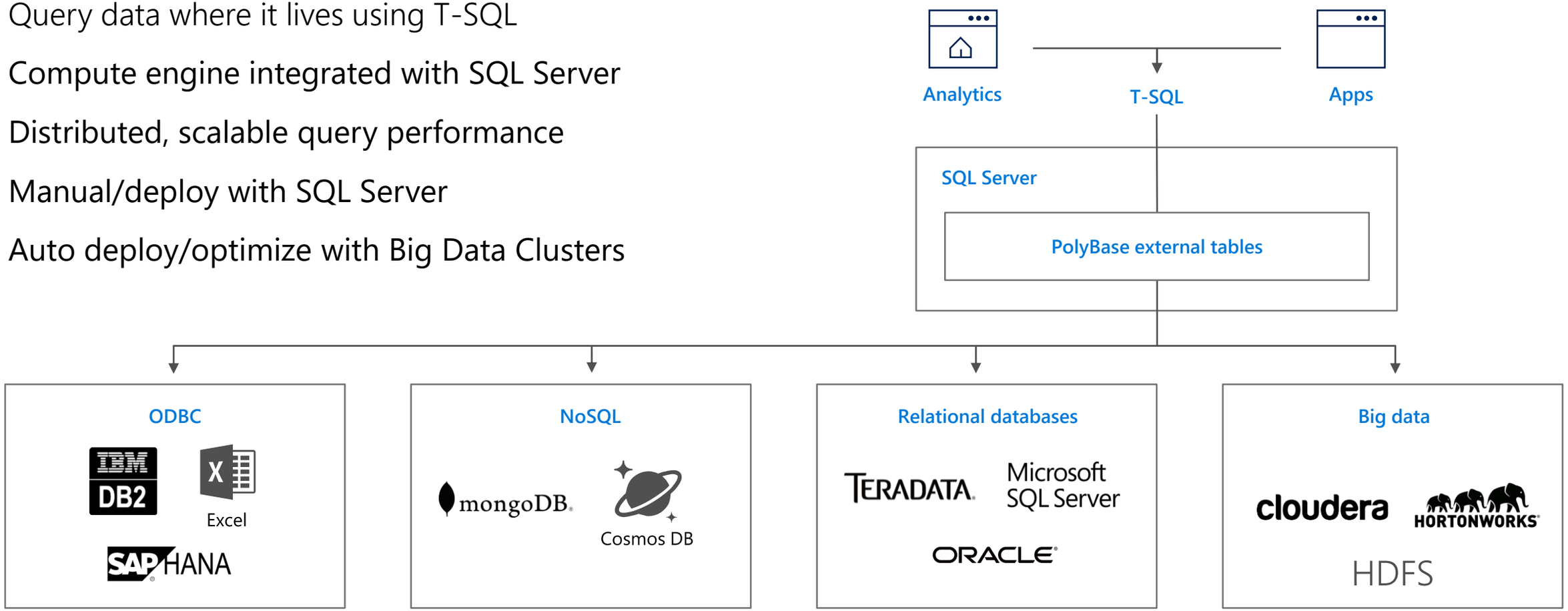
Demo 2
Data Virtualization



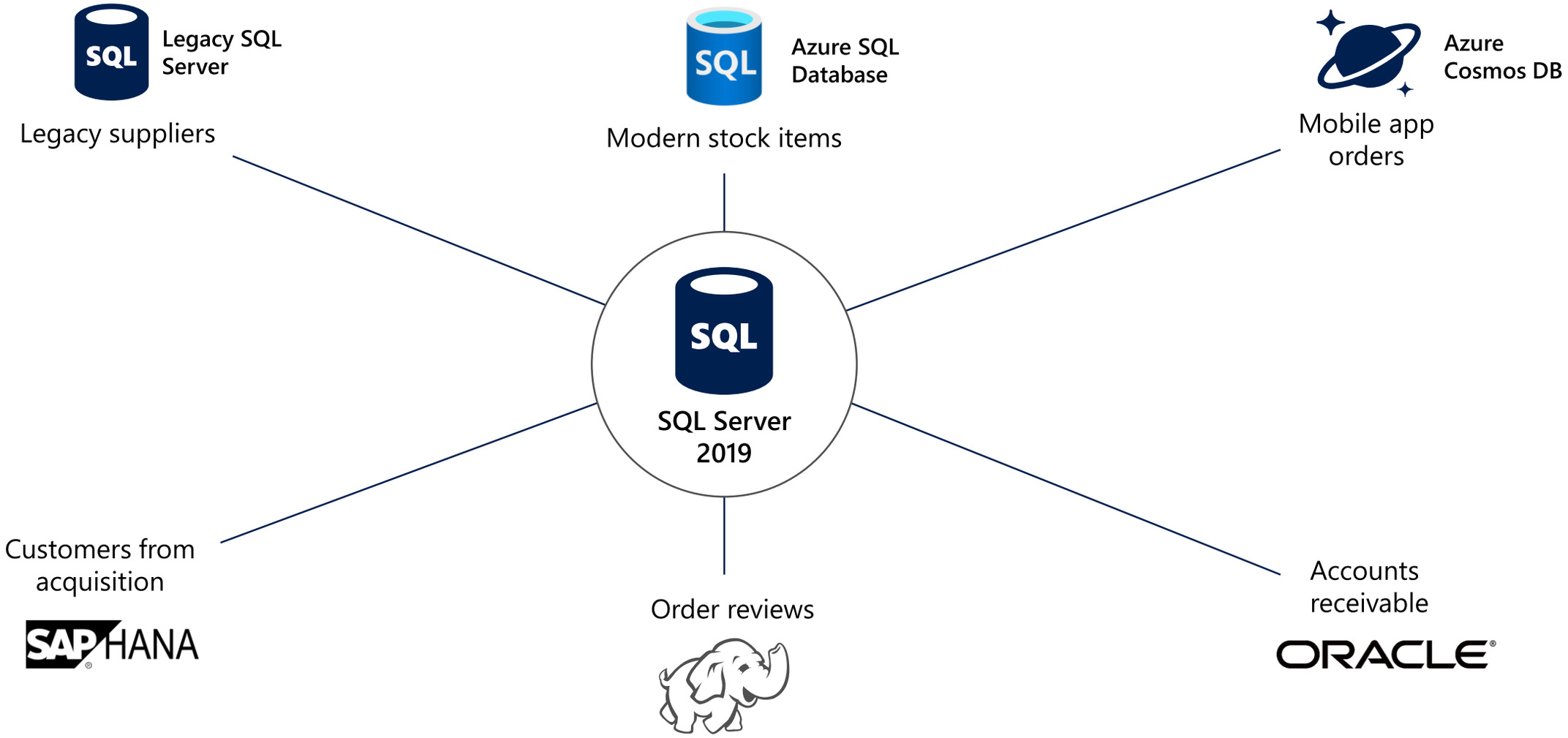
YOU DON'T EVEN HAVE TO MOVE DATA

"It's all about Data Virtualization"

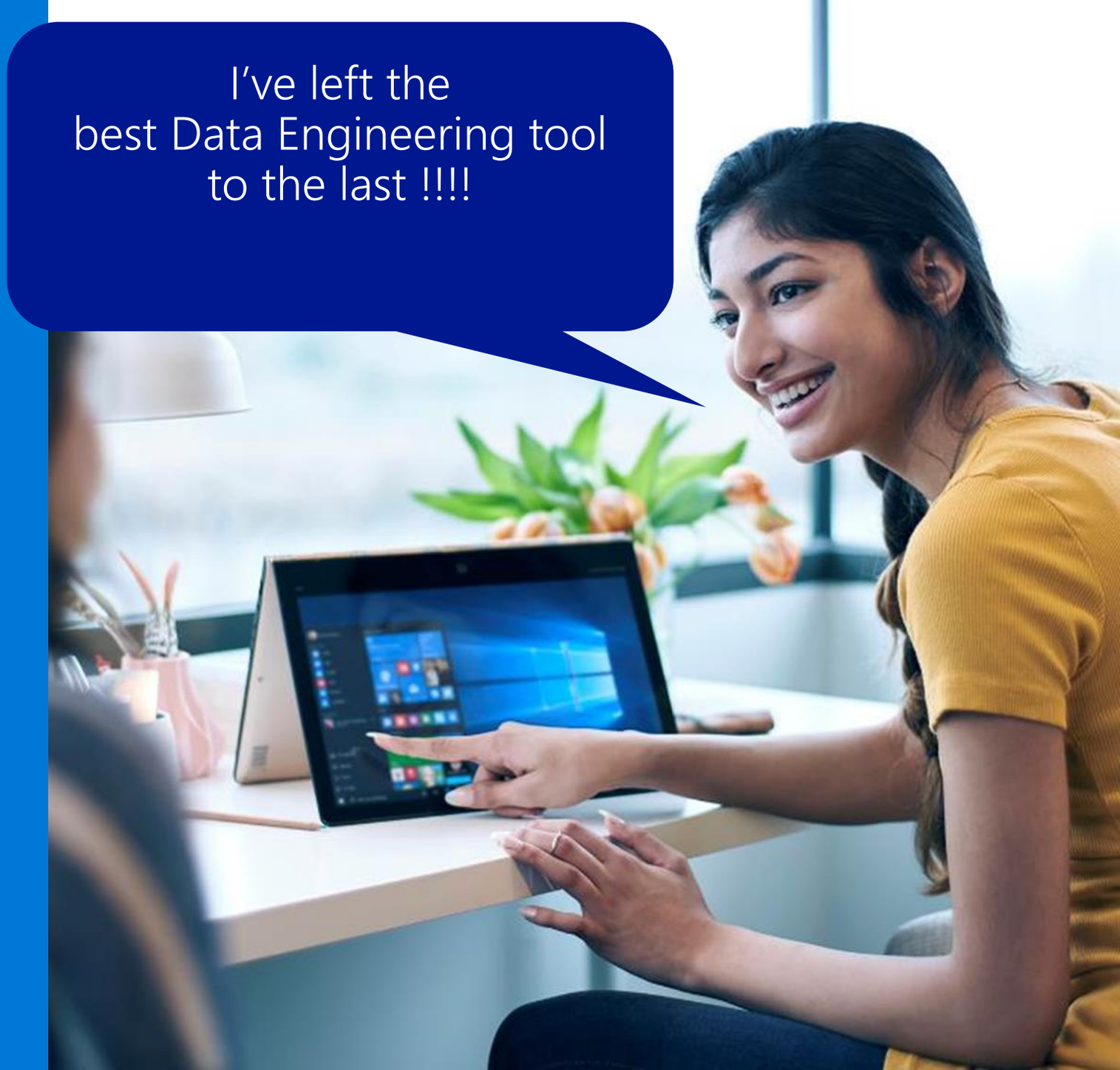
- ✓ It all started with T-SQL against Hadoop
- ✓ Query data where it lives using T-SQL
- ✓ Compute engine integrated with SQL Server
- ✓ Distributed, scalable query performance
- ✓ Manual/deploy with SQL Server
- ✓ Auto deploy/optimize with Big Data Clusters



SQL SERVER 2019 SERVES AS THE HUB FOR ALL YOUR DATA



Demo 3
Azure Data Studio Notebooks



AZURE DATA STUDIO

1. Cross Platform – Windows, macOS, Linux

https://docs.microsoft.com/en-us/sql/azure-data-studio/download-azure-data-studio?view=sql-server-ver15

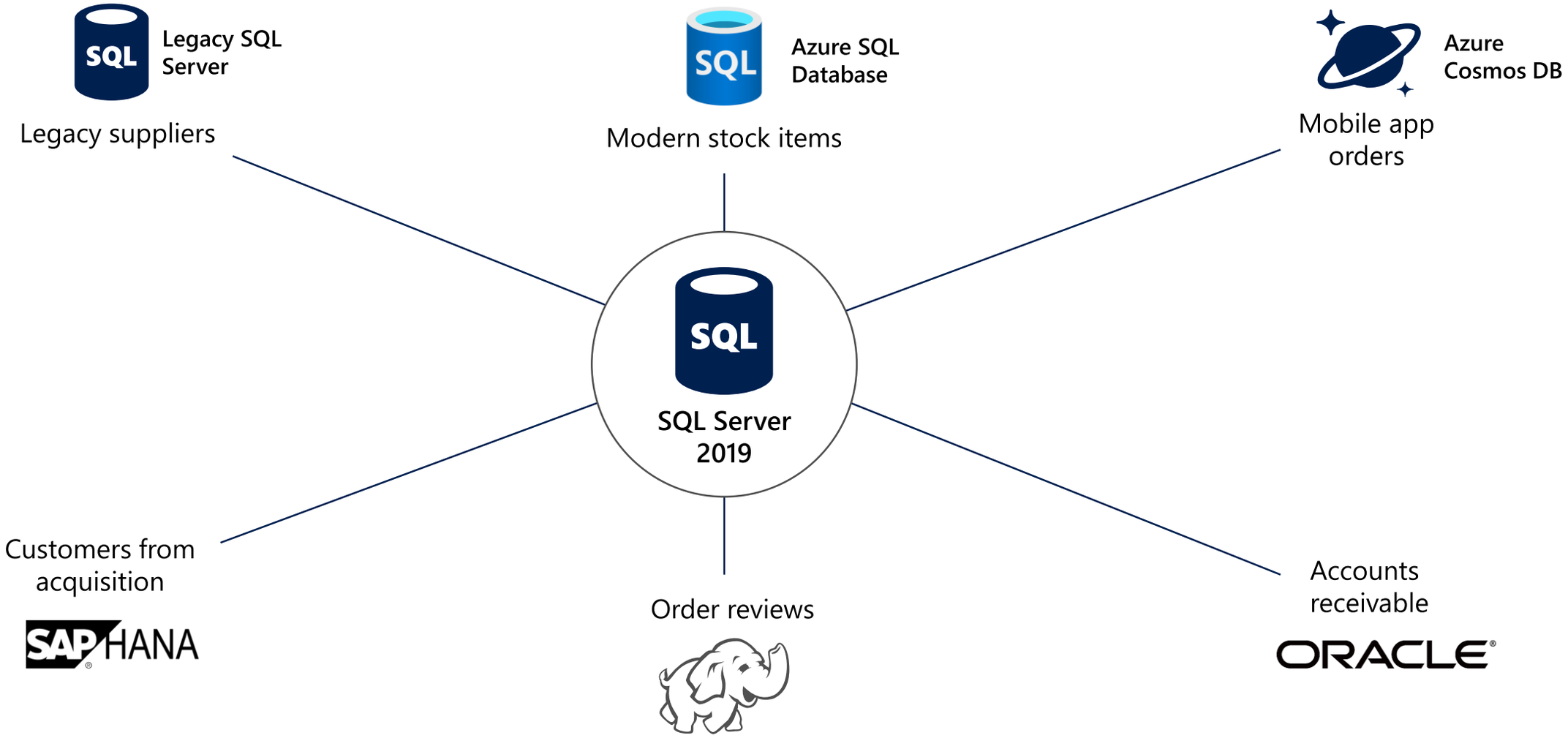
AZURE DATA STUDIO

References

Quickstart: Use Azure Data Studio to connect and query Azure SQL database -
https://docs.microsoft.com/en-us/sql/azure-data-studio/quickstart-sql-database?view=sql-server-ver15

https://docs.microsoft.com/en-us/sql/azure-data-studio/download-azure-data-studio?view=sql-server-ver15

SQL SERVER 2019 SERVES AS THE HUB FOR ALL YOUR DATA



Azure and Data Estate Data View

DATA SOURCES

Structured

- OLTP
- MOBILE
- ERP
- LOB

Unstructured

- Graph
- Social
- IoT
- Media

Blob

DATA MANAGEMENT

Data virtualization

ON-PREM | CLOUD

Operational data

- SQL Server
- Azure SQL DB
- SQL in VM
- Cosmos DB
- Managed DB for MySQL & PostgreSQL

Data warehousing

- SQL Server Data Warehouse
- Azure SQL Data Warehouse

Big data processing

- Apache Hadoop
- Azure HDInsight
- Databricks
- Azure Data Lake

Data integration

DATA INSIGHTS

Business intelligence

- Power BI
- SQL Server Reporting Services and Mobile BI

AI & Data Science

- SQL Analysis Services and R/Pythion built-in
- Azure Databricks Azure ML
- Azure Analysis Services
- Azure Cognitive Services

SECURITY

Least vulnerable data platform, with more certifications than any other cloud provider

FLEXIBILITY

Any language, any platform, anywhere

.NET JAVA Azure 3rd

Microsoft Azure BI & Analytics services portfolio

Devices	Device Connectivity	Storage	Analytics	Presentation & Action
	IoT Hub	SQL Database	Machine Learning	App Service
	Event Hub	Table/Blob Storage	Stream Analytics	Power BI
	Service Bus	DocumentDB	HDInsight	Notification Hubs
	External Data Sources	3rd party Databases	Data Factory	Mobile Services
			Data Lake	BizTalk Services

TOP 10 FEATURES IN SQL SERVER 2019

1. Harness the power of big data

Big data clusters with scalable compute and storage composed of SQL Server, Spark, and HDFS. Cache data in scale-out data marts.

2. Bring AI to your workloads

A complete AI platform to train and operationalize models in SQL Server ML Services or Spark ML using Azure Data Studio notebooks.

3. Eliminate the need for data movement

Data virtualization allows queries across relational and non-relational data without movement or replication.

4. Explore and interact with visual data

Visual data exploration and interactive analysis using SQL Server BI tools and Power BI Report Server.

5. Run real-time analytics on operational data

In-memory technologies for analytics on operational data using HTAP. Higher concurrency and scale through persistent memory.

6. Automatically tune SQL Server

Intelligent Query Processing improves scaling of queries and Automatic Plan Correction resolves performance problems.

7. Reduce database maintenance and increase business uptime

Greater uptime with more online indexing operations. Now run Always On availability groups on containers using Kubernetes.

8. Boost security and protect data in use

SQL Server enables layers of security including protection of computations in Always Encrypted secure enclaves.

9. Track compliance with sophisticated resources

Data Discovery & Classification labeling for GDPR and Vulnerability Assessment tool to track compliance.

10. Optimize with choice and flexibility

Support for your choice of Windows, Linux, and containers. Run Java code on SQL Server and store and analyze graph data.

SQL SERVER 2019

Industry-leading performance and security, with intelligence over all your data

Intelligence over any data

AI and Machine Learning over all data with the power of SQL and Apache Spark

Choice of platform and language

T-SQL
Java
C/C++

PHP
Node.js
C#/VB.NET

Python
Ruby

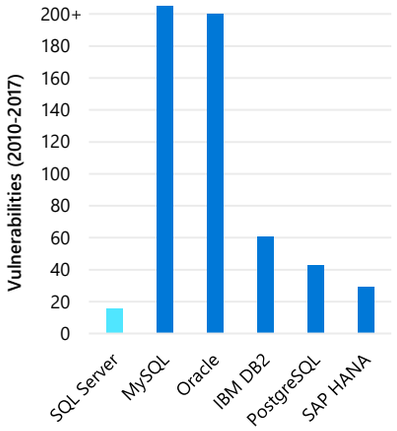
Industry-leading performance

#1 OLTP performance[1]

#1 DW performance on 1TB[2], 10TB[3], and 30TB[4]

Intelligent Query Processing

Most secure over the last 8 years[5]



Insights in minutes and rich reports

The best of Power BI and SQL Server Reporting Services with Power BI Report Server

---

Private cloud

In-memory across all workloads

Most consistent data platform

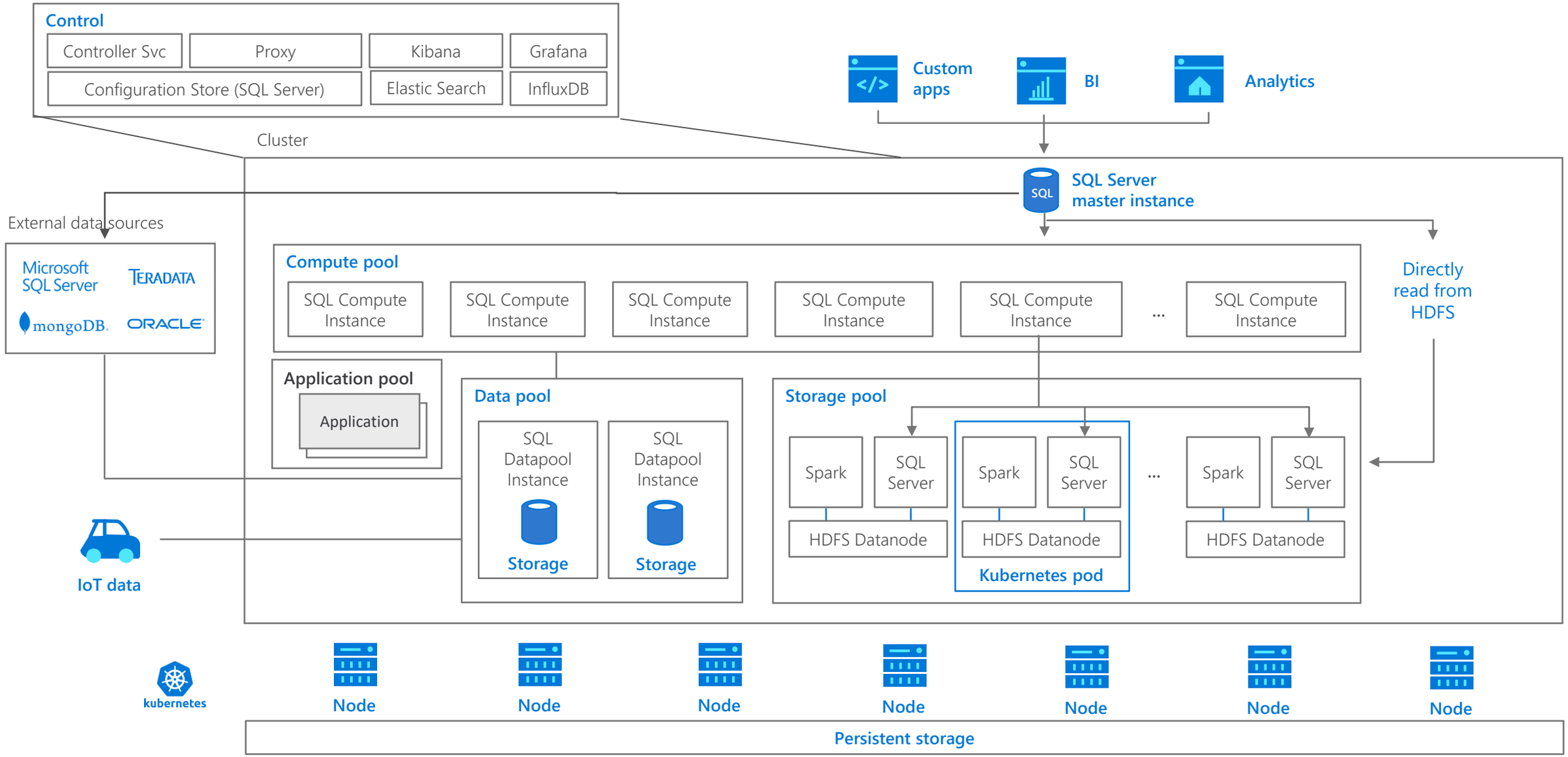
1/10th the cost of Oracle

Public cloud

All TPC Claims as of 1/19/2018.

[1] http://www.tpc.org/4081; [2] http://www.tpc.org/3331; [3] http://www.tpc.org/3326; [4] http://www.tpc.org/3321; [5] National Institute of Standards and Technology Comprehensive Vulnerability Database

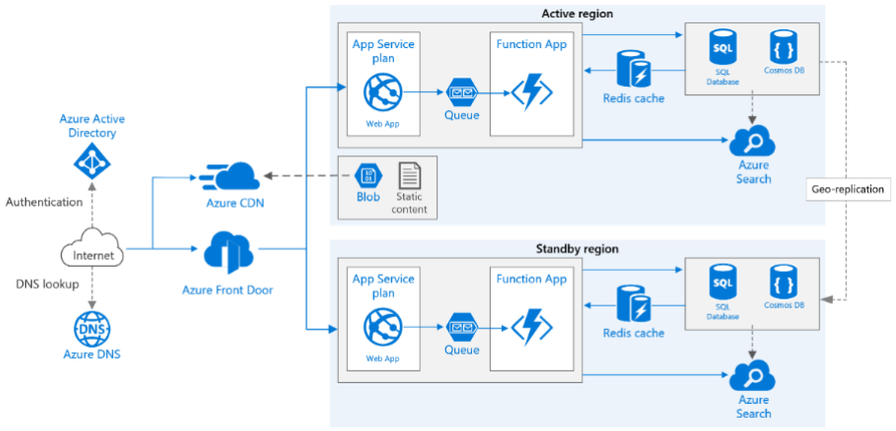
BIG DATA CLUSTERS DEPLOYED ON KUBERNETES, OPENSHIFT, AKS



PRINCIPLES -

- This reference architecture shows how to run an Azure App Service application in multiple regions to achieve high availability.

https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/app-service-web-app/multi-region



Attribution

Some pictures used are from http://www.Pixels.com/about - Pexels provides high quality and completely free stock photos licensed under the Pexels license.