Building Batch Data Processing Solutions in Microsoft Azure

DEVELOPING BATCH PROCESSING SOLUTIONS WITH AZURE SQL DATA WAREHOUSE

Tim Warner

AUTHOR EVANGELIST, PLURALSIGHT

@TechTrainerTim TechTrainerTim.com

MVP Microsoft® Most Valuable Professional

Microsoft CERTIFIED AZURE SOLUTIONS ARCHITECT EXPERT

Batch Data Processing Course Flow

Developing Batch Processing Solutions with Azure SQL Data Warehouse

Developing Batch Processing Solutions with Azure HDInsight

Developing Batch Processing Solutions with Azure Databricks

Overview

Cover preliminary terminology

Understand Azure SQL Data Warehouse

Perform batch data processing with Azure SQL Data Warehouse

- Azure Data Lake Storage Gen2
- PolyBase
- Azure Data Factory

Exercise Files

What do you want to learn?

Timothy
timothywarner316@gmail.com

Troubleshooting with Microsoft Azure Network Watcher

by Tim Warner

Microsoft now gives you packet-level access to your Windows Server and Linux virtual machines (VMs) running in Azure. You'll learn how to use Network Watcher to troubleshoot network security groups (NSGs), perform packet captures, and much more.

Resume Course | Bookmark | Add to Channel

Table of contents | Description | Transcript | **Exercise files** | Discussion | Learning Check | Recommended

These exercise files are intended to provide you with the assets you need to create a video-based hands-on experience. With the exercise files, you can follow along with the author and re-create the same solution on your computer. We find this to be even more effective than written lab exercises.

Download exercise files

Course author

Tim Warner

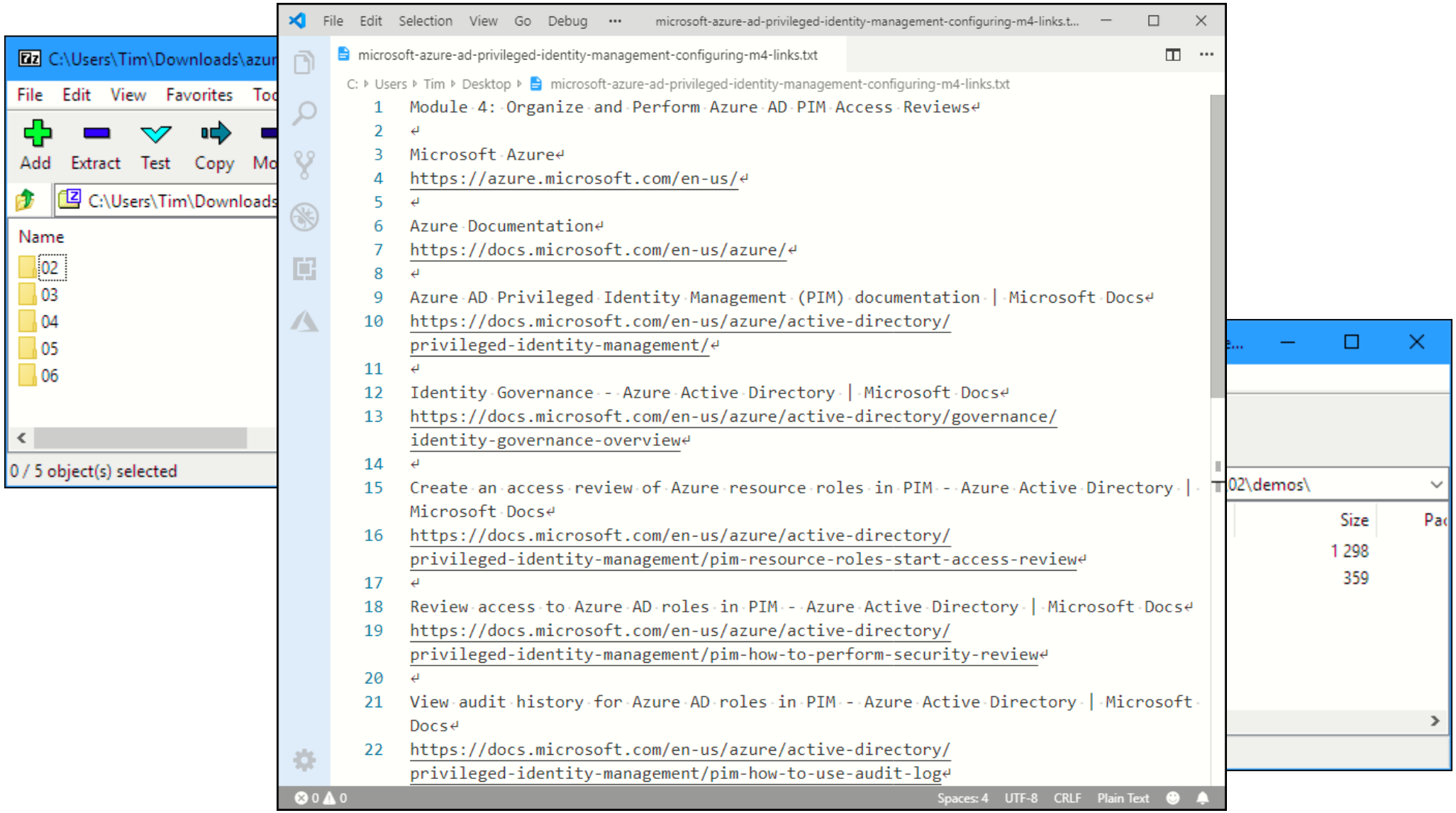
Timothy Warner is a Microsoft Most Valuable Professional (MVP) in Cloud and Datacenter Management who is based in Nashville, TN.

Course info

Level	Intermediate
Rating	★★★★★
My rating	★★★★★
Duration	2h 12m
Released	31 Oct 2017

Share course

Exercise Files

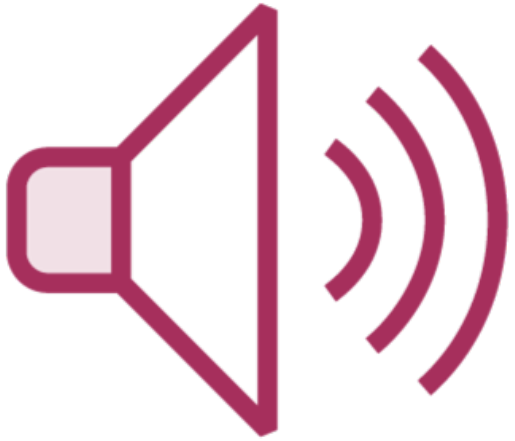


Preliminary Terminology

Big data

[Google] Extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.

The Four V's of Big Data



Volume

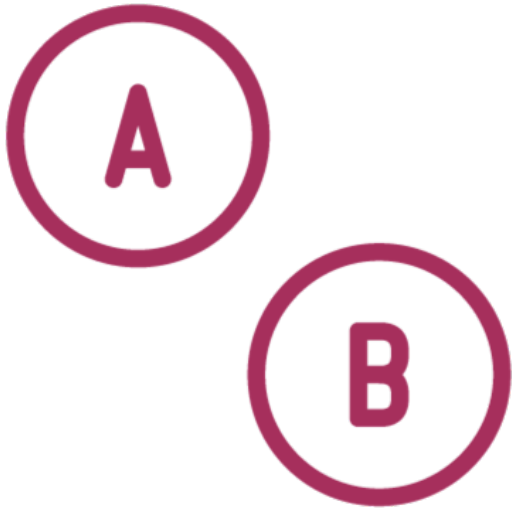
PB, EB

Billions or trillions of records

Velocity

Processing frequency

Latency or real-time?



Variety

Structured, semi-structured, unstructured data

Veracity

Data trustworthiness

Noise, bias

Data Formats

Structured data

- SQL table

Semi-structured data

- JSON, XML

Unstructured data

- CSV
- PNG, EXE (blob)

Data Warehouse

Central repository of integrated data. Data is defined, structured and highly transformed. Operations are performed in a massively parallel way.

Data Lake

Raw data repository whose purpose is not yet determined and is left in-place until needed. Highly accessible and quick to update.

ETL

Extract, Transform, Load. Data is transformed "in flight" between source and destination. Does not scale particularly well.

ELT

Extract, Load, Transform. Data is transformed after it is placed in a data lake. A great fit for the public cloud, given limitless compute and storage resources.

Batch vs. Stream Processing

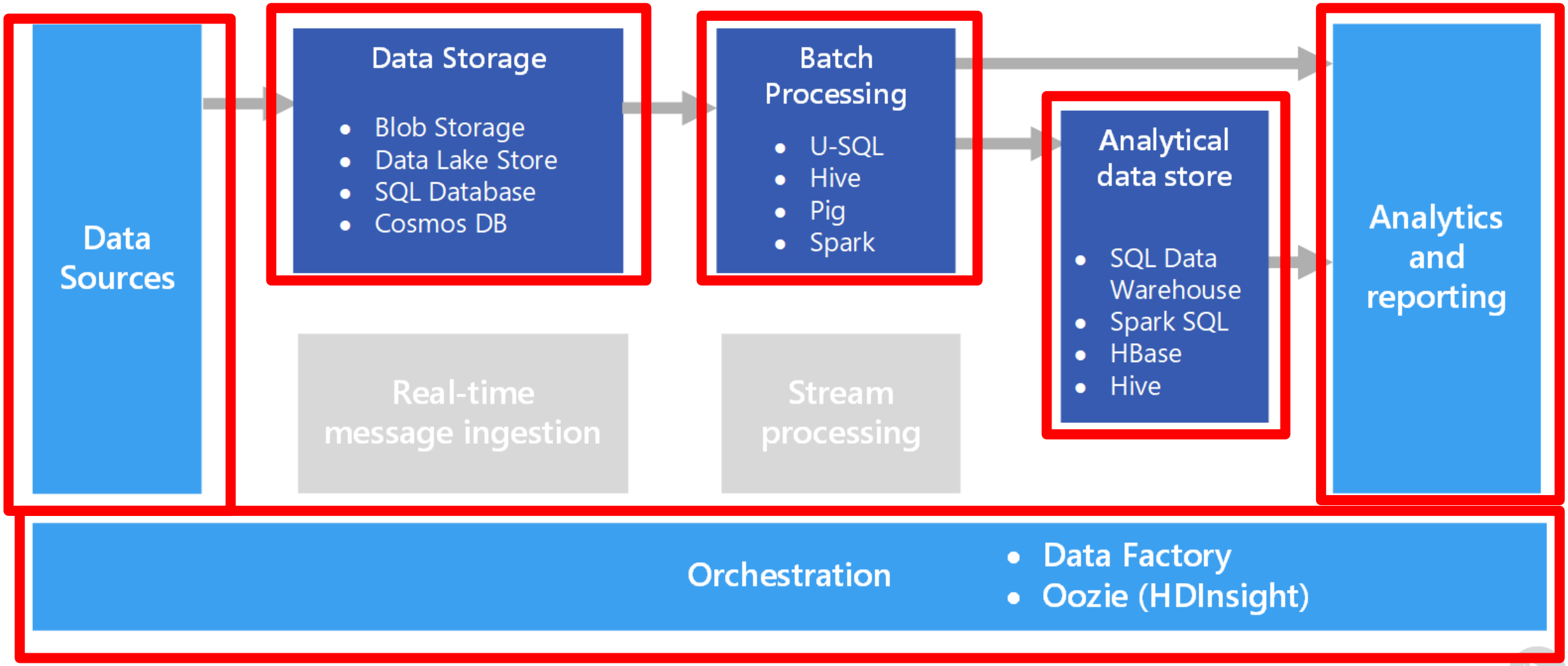
Batch processing: Analyze previously stored data

Stream processing: Analyze incoming data in real time

Batch Data Processing Characteristics

- Long-running batch jobs
- Filter, aggregate, and prepare date for analysis
- Read source files from scalable storage
- Process the data
- Write output to scalable storage
- Scaled-out computation

Batch Processing Workflow

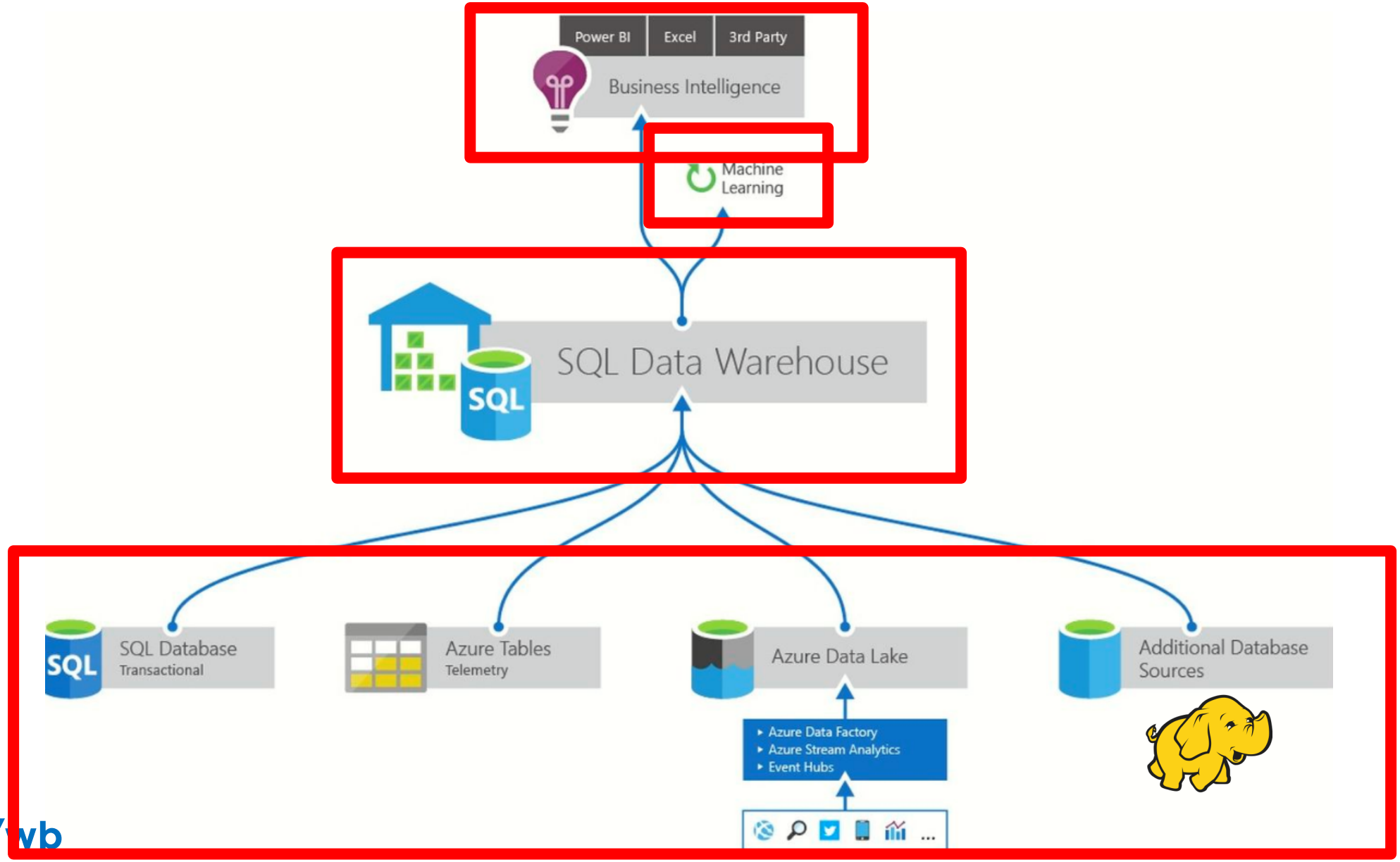


About Azure SQL Data Warehouse

Product Comparison

Azure SQL Database	Azure SQL Data Warehouse
OLTP/CRUD	OLAP/querying and reporting
SMP	MMP
Vertical scale	Horizontal scale
No PolyBase	Can pause the virtual server to save costs
	PolyBase

Data Architecture



Demo

1

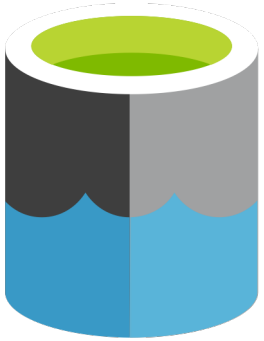
Create an Azure SQL Data Warehouse

Customize firewall rules

Connect via SSMS and Visual Studio

Data Inflows and Outflows with Azure SQL Data Warehouse

Azure Data Lake Storage Gen2



Excellent platform for big data analytics

Best of both worlds:

- Low cost and flexibility of Azure blob storage
- HDFS compatibility and file system semantics from Azure Data Lake Storage Gen1

Azure Data Lake Storage Gen2

Microsoft Azure Storage Explorer

File Edit View Preview Help

EXPLORER

Search for resources

Collapse All Refresh All

- Data Lake Storage Gen1 (Preview)
- Storage Accounts
- Pluralsight (timothywarner316@gmail.com)
  - Data Lake Storage Gen1 (Preview)
    - adlsgen1tw
    - psdatalake10
  - Storage Accounts
    - pslakegen2 (ADLS Gen2)
      - Blob Containers
        - hdinsight
        - hditestcluster1-fs
        - sales-reports
        - testclusterhdi1-2019-07-31t19-28-37-090z
      - File Shares
      - Queues
      - Tables

Actions Properties

URL	https://pslakegen2.dfs.core.windows.net/hditestcluster1
Type	Blob Container (ADLS Gen2)
HNS Enabled	true
DFS Endpoint	https://pslakegen2.dfs.core.windows.net/

hditestcluster1-fs

Upload Download New Folder Select All Rename Manage Access Properties Delete Refresh

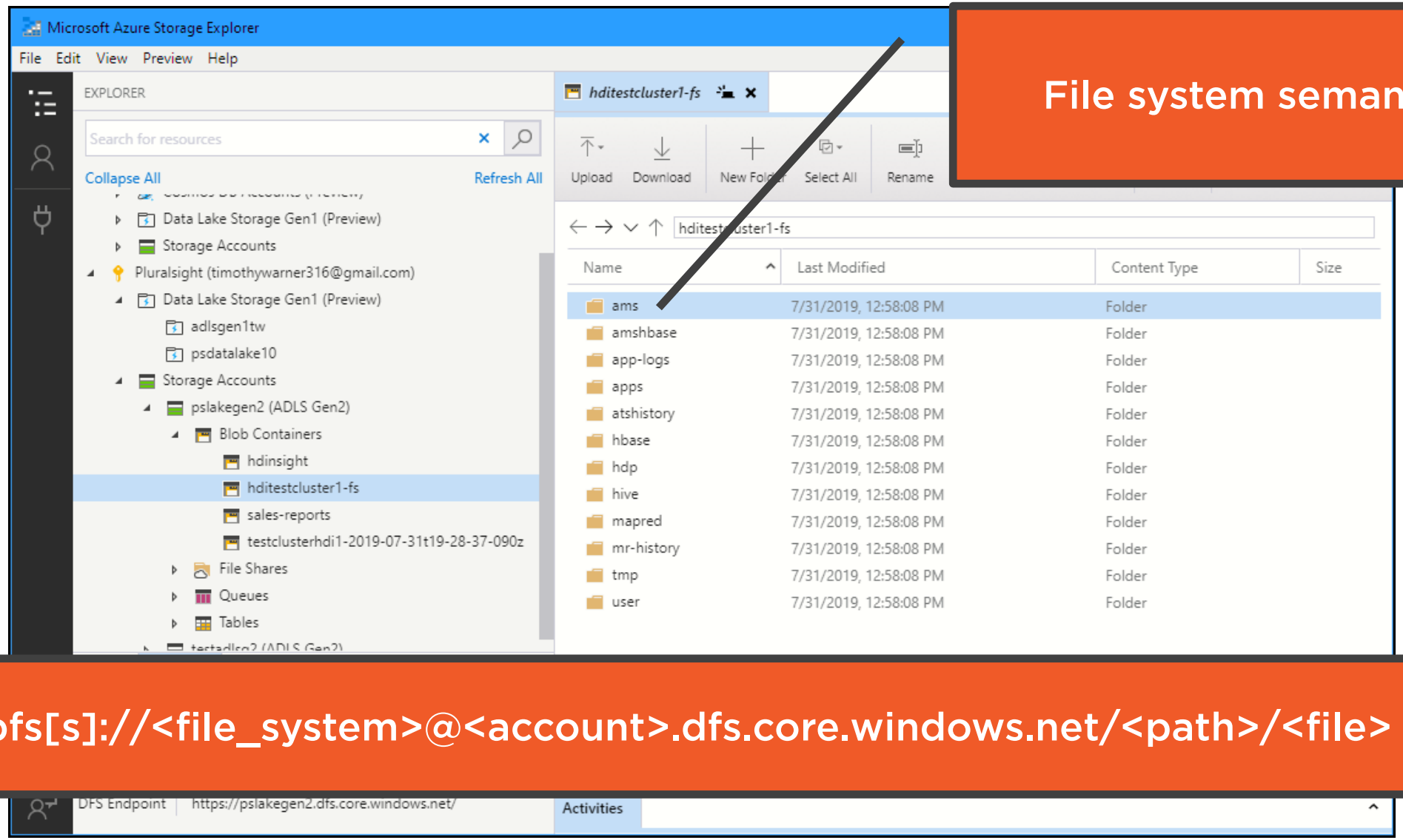
hditestcluster1-fs

Name	Last Modified	Content Type	Size
ams	7/31/2019, 12:58:08 PM	Folder	
amshbase	7/31/2019, 12:58:08 PM	Folder	
app-logs	7/31/2019, 12:58:08 PM	Folder	
apps	7/31/2019, 12:58:08 PM	Folder	
atshistory	7/31/2019, 12:58:08 PM	Folder	
hbase	7/31/2019, 12:58:08 PM	Folder	
hdp	7/31/2019, 12:58:08 PM	Folder	
hive	7/31/2019, 12:58:08 PM	Folder	
mapred	7/31/2019, 12:58:08 PM	Folder	
mr-history	7/31/2019, 12:58:08 PM	Folder	
tmp	7/31/2019, 12:58:08 PM	Folder	
user	7/31/2019, 12:58:08 PM	Folder	

Showing 1 to 12 of 12 cached items

Activities

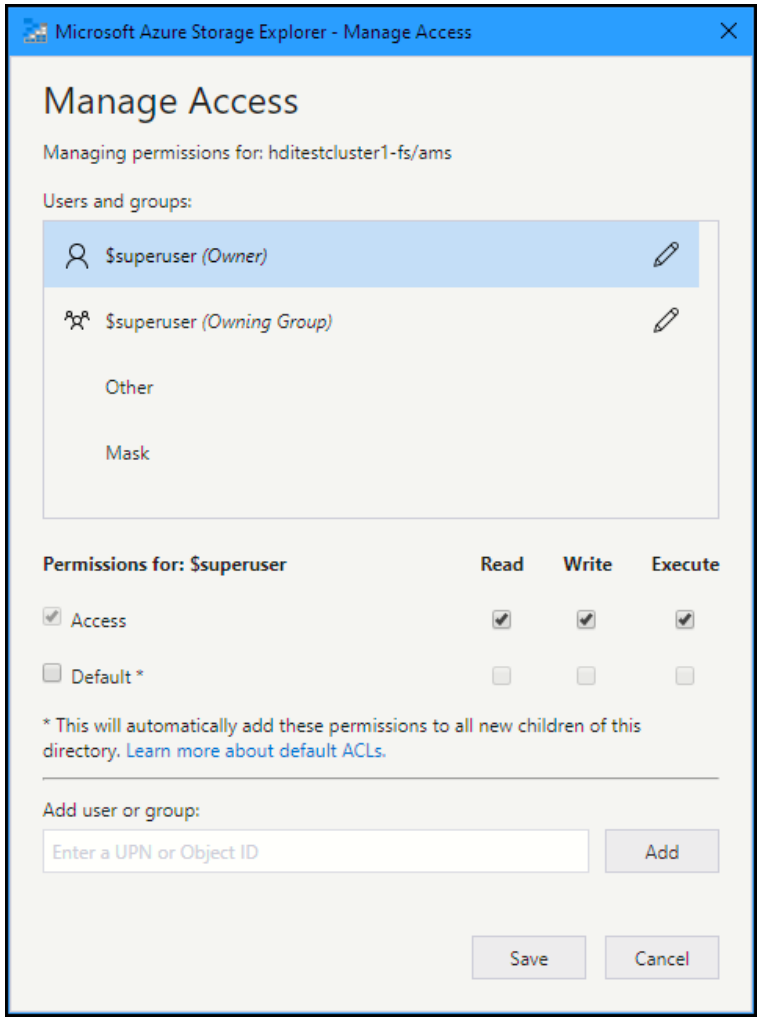
Azure Data Lake Analytics Gen2



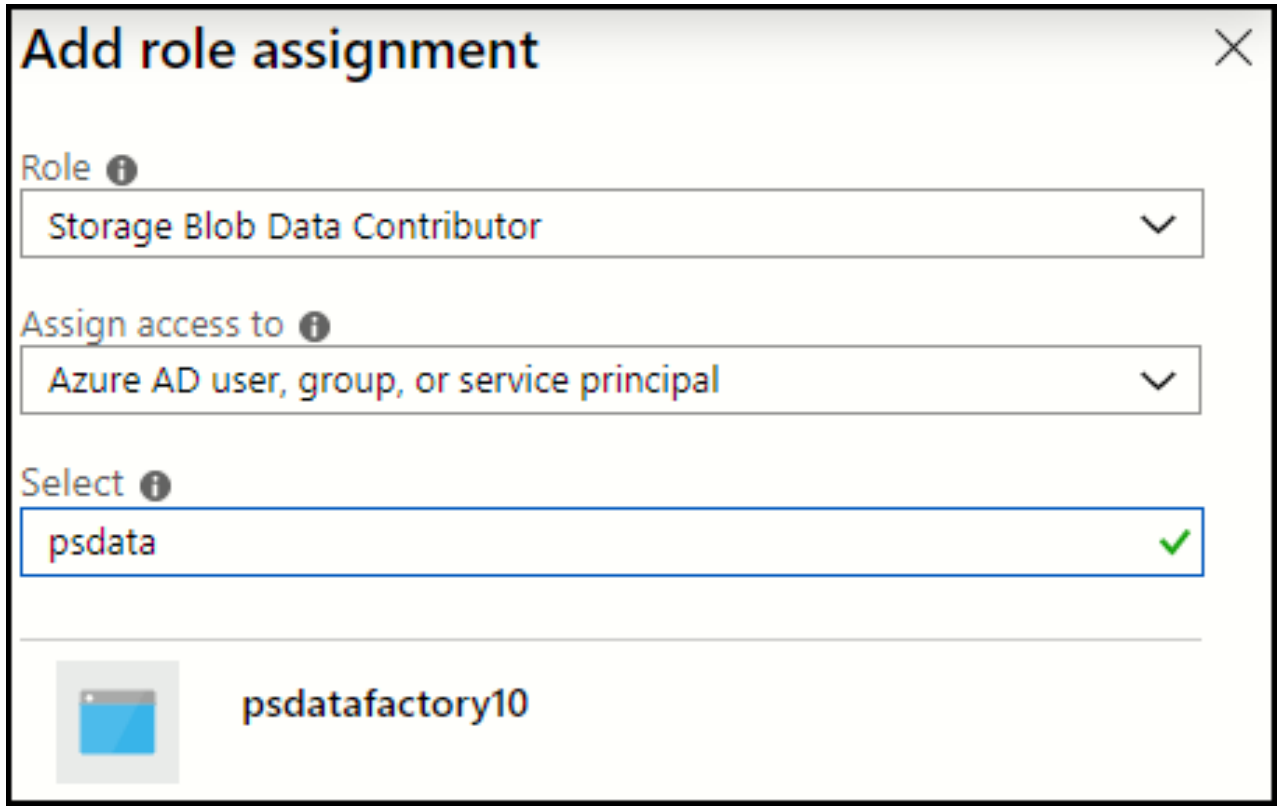
File system semantics

abfs[s]://<file_system>@<account>.dfs.core.windows.net/<path>/<file>

Azure Data Lake Analytics Gen2



ACL/POSIX permissions



Role-based access control (RBAC)

PolyBase

- Access external data using T-SQL
- Azure blob storage
- Azure Data Lake Storage (Hadoop)
- Query the external data
- Load or export

Demo

2

Blob storage into Data Warehouse

- Query
- Data load

Azure Data Factory

Code-free data integration solution

"Cloud-based SSIS"

Build hybrid ETL and ELT pipelines with a visual design surface

Over 80 pre-built connectors to different data sources

Demo

3

Data warehouse into Data Lake Storage

Azure Data Factory

For Further Learning

Azure SQL Data Warehouse: First Look
(Warner Chaves)
Excellent supplement for beginners

Plan for Data Warehousing with Microsoft Azure
(John Savill)
An IT operations perspective

Summary

ASDW covers the relational/structured data model well

What about data engineers who want to use native Hadoop big data analysis tools?

Next module: Developing Batch Processing Solutions with Azure HDInsight