Intro to Azure and **Data Platforms**

UNIVERSITY OF NORTH CAROLINA CHARLOTTE

DSBA 6190-U90 | Colby T. Ford, Ph.D.

Overview

About Microsoft Azure

Unstructured Data Storage

- Azure Storage
- Azure Data Lake

Structured Data Storage

- Azure SQL DB
- Azure Synapse

Azure

Top Benefits of Cloud Computing

Types of Clouds

Service Offerings

Regions

https://azure.microsoft.com/en-us/overview/what-is-cloud-computing/

Top Benefits of Cloud Computing

Cost

Cloud computing eliminates the capital expense of buying hardware and software and setting up and running on-site datacenters—the racks of servers, the round-the-clock electricity for power and cooling, the IT experts for managing the infrastructure. It adds up fast.

Speed

Most cloud computing services are provided self service and on demand, so even vast amounts of computing resources can be provisioned in minutes, typically with just a few mouse clicks, giving businesses a lot of flexibility and taking the pressure off capacity planning.

Global scale

The benefits of cloud computing services include the ability to scale elastically. In cloud speak, that means delivering the right amount of IT resources—for example, more or less computing power, storage, bandwidth—right when it's needed, and from the right geographic location.

Productivity

On-site datacenters typically require a lot of "racking and stacking"—hardware set up, software patching, and other time-consuming IT management chores. Cloud computing removes the need for many of these tasks, so IT teams can spend time on achieving more important business goals.

Performance

The biggest cloud computing services run on a worldwide network of secure datacenters, which are regularly upgraded to the latest generation of fast and efficient computing hardware. This offers several benefits over a single corporate datacenter, including reduced network latency for applications and greater economies of scale.

Security

Many cloud providers offer a broad set of policies, technologies, and controls that strengthen your security posture overall, helping protect your data, apps, and infrastructure from potential threats.

Types of Clouds

Public Cloud

- Owned and operated by third-party cloud service providers
- Deliver computing resources like servers and storage over the internet

Private Cloud

- Owned exclusively by a single business or organization
- Can be physically located on the company's on-site datacenter
- Could also be hosted by a third-party service provider, but maintained by the organization on a private network
- Most control over security and compliance

Hybrid Cloud

- Somewhere in the middle of public and private
- Bound by technologies that allows data and applications to be shared between them
- Most flexible option as it allows for a broad range of deployment options

Types of Cloud Services

IaaS

- **Infrastructure**
- Most basic form of cloud computing services
- Rent IT infrastructure
- Includes:
  - VMs, Storage, Networks, Operating Systems

PaaS

- **Platform**
- Services that provide on-demand environments for developing, testing, delivering, and managing software applications
- No (less) need to manage underlying infrastructure

SaaS

- **Software**
- Delivery of software applications over the internet
- Usually as an on-demand/subscription basis
- Cloud provider hosts and manages the software application and underlying infrastructure

Serverless

- Overlaps with PaaS
- Focuses on building app functionality without spending time managing the servers and infrastructures required to do so
- Usually highly-scalable and event-driven

Azure Regions

60+ Azure Regions

https://datacenters.microsoft.com/globe/explore/

Reasons to Select a Region

- Speed
  - Pick a region that is the closest to you will help increase the speed of moving data in and out from your resources.
- Cost
  - You will face egress fees if you move data from one region to another. So, it's cheaper for you to have everything nearby.
  - Some services are cheaper in certain regions over others. Using the Azure calculator can help you estimate the cost differences.
- Available Resources
  - Certain services are only available in certain regions. It might not be possible to have everything in the same place. (See Products by Region.)
- Security and Compliance
  - Depending on the scenario, a particular data center might need to be selected for security and compliance reasons. (Such as government or China)

Availability Zones

- Physically separate locations within each Azure region that are tolerant to local failures (earthquakes, floods, fires, etc.)
- Tolerance to failures is achieved because of redundancy and logical isolation of Azure services.
- Azure availability zones are connected by a high-performance network with a round-trip latency of less than 2ms.

Available Service Categories

- Compute
- Developer Tools
- Integration
- Media
- Networking
- Web
- Analytics
- Databases
- Identity
- Management and Governance
- Mobile
- Storage
- AI + Machine Learning
- Containers
- DevOps
- Internet of Things
- Migration
- Security

https://azure.microsoft.com/en-us/services/

Unstructured Data Storage

Azure Storage

- Azure Data Lake Gen 2

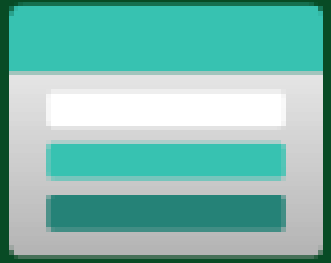
Azure Storage

- Azure Storage provides cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables.

File
Simple, distributed, cross-platform file system
- Lift and shift migration
- Simple and inexpensive
- Move data to cloud with no coding

Disk
Persistent, high-performance disk storage for every workload
- Low latency, high throughput
- Industry-leading, single-instance service-level agreement
- Enterprise-grade durability

Blob
Massively-scalable object storage for unstructured data
- Cost-effective for massive volume
- Tiered storage options
- Single infrastructure with global reach



Archive
Industry leading price point for storing rarely accessed data
- Data automatically encrypted at rest
- Seamless integration with hot and cool storage tiers
- Supported by leading Data Management partners

Data Lake Storage
Secure, massively scalable data lake storage.
- Limitless storage for analytics data
- Optimized for Apache Spark and Hadoop analytics engines
- High-performance file system with support for fine-grained access control lists

https://azure.microsoft.com/en-us/services/storage/?v=18.24

Storage Tiers

Hot

- Hot is pretty much the default tier.
- It provides the best performance for retrieving files quickly and often but carries the highest price tag.
- This is used for files that you plan on consistently accessing, reading, or updating.

Cool

- This tier is similar to Hot, though I think it goes underused.
- This tier is better for important files that you might need to access, but only every once in a while.
- It carries the same performance as Hot, but at a lower storage cost (and a higher data access cost). This tier is recommended for files that you may only access every month or so.

Archive

- You can think of this tier as the backup version of the other two.
- This tier has a very low storage costs, but it'll take a few hours to access to the data when you want it.
- Archive is recommended for files you access every 6 months or so.

Premium

- Premium isn't really a tier but a different type of storage altogether.
- This provides access to block blobs consistently low latency, which is perfect for high frequency data transactions.
- This carries the highest price tag.

Medallion Architecture

Bronze - Ingesting Raw Data

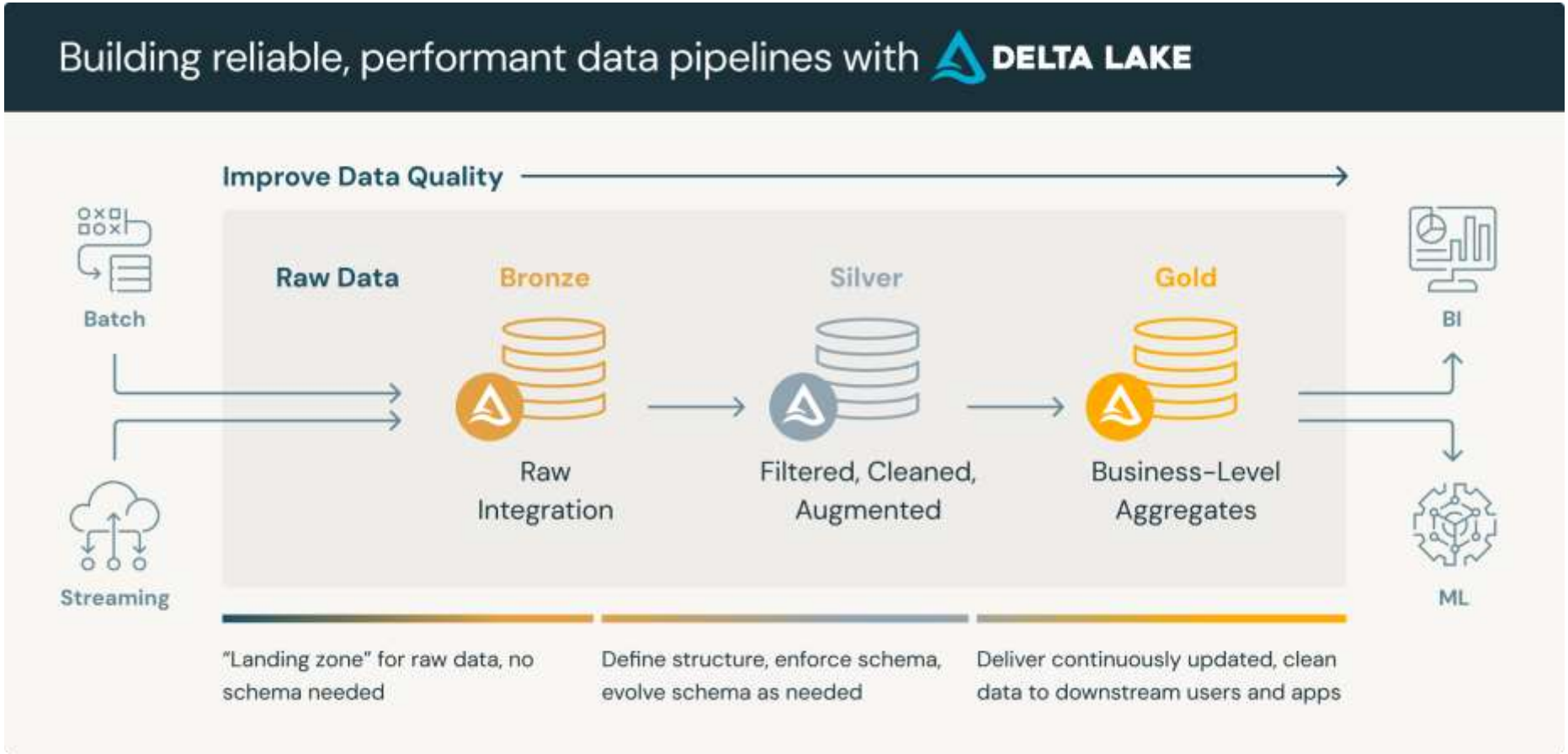
- The bronze layer contains unvalidated data. Data ingested in the bronze layer typically:
  - Maintains the raw state of the data source.
  - Is appended incrementally and grows over time.
  - Can be any combination of streaming and batch transactions.
- Retaining the full, unprocessed history of each dataset in an efficient storage format provides the ability to recreate any state of a given data system.
- Additional metadata (such as source file names or recording the time data was processed) may be added to data on ingest for enhanced discoverability, description of the state of the source dataset, and optimized performance in downstream applications.

Silver - Cleanse, Validate, and Deduplicate

- Recall that while the bronze layer contains the entire data history in a nearly raw state, the silver layer represents a validated, enriched version of our data that can be trusted for downstream analytics.
- While Databricks believes strongly in the lakehouse vision driven by bronze, silver, and gold tables, simply implementing a silver layer efficiently will immediately unlock many of the potential benefits of the lakehouse.
- For any data pipeline, the silver layer may contain more than one table.

Gold - Curate to Power Analytics

- This gold data is often highly refined and aggregated, containing data that powers analytics, machine learning, and production applications. While all tables in the lakehouse should serve an important purpose, gold tables represent data that has been transformed into knowledge, rather than just information.
- Analysts largely rely on gold tables for their core responsibilities, and data shared with a customer would rarely be stored outside this level.
- Updates to these tables are completed as part of regularly scheduled production workloads, which helps control costs and allows service level agreements (SLAs) for data freshness to be established.
- While the lakehouse doesn't have the same deadlock issues that you may encounter in a enterprise data warehouse, gold tables are often stored in a separate storage container to help avoid cloud limits on data requests.
- In general, because aggregations, joins, and filtering are handled before data is written to the gold layer, users should see low latency query performance on data in gold tables.



https://www.databricks.com/glossary/medallion-architecture

Begin lab…

Structured Data Storage

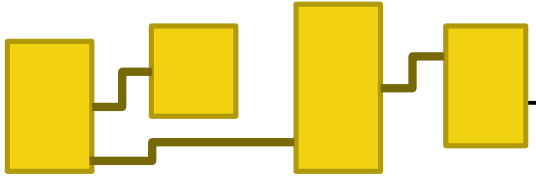
Azure SQL Database

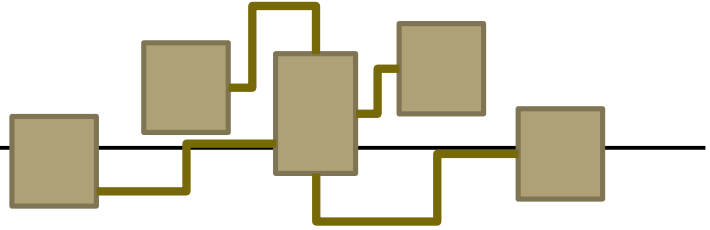
Azure Synapse

Other SQL-based Offerings

Databases vs. Data Warehouses

- Databases:
  - Usually for record-keeping of transactional systems
  - OLTP – Online Transactional Processing
  - Usually for a single application (personnel database, EMR, POS, etc.)
  - Focused on performance
  - Relational Schema
  - Highly normalized tables
- Both:
  - Have multiple tables connected by keys
  - Use the same data types, indices, etc.
  - Can use views
  - Can be run on Azure SQL DB or Azure SQL DW
- Data Warehouses:
  - Usually for keeping organizational historical data
  - OLAP – Online Analytical Processing
  - Usually for multiple systems combined
  - Focused on aggregation
  - Star or Snowflake Schema
  - Dimensions and Facts





Azure SQL Database

- Azure SQL Database is a relational database-as-a-service (DBaaS) based on the latest stable version of Microsoft SQL Server Database Engine. SQL Database is a high-performance, reliable, and secure database you can use to build data-driven applications and websites in the programming language of your choice, without needing to manage infrastructure.

Fully managed
PaaS database that is always running on the latest stable version of SQL Server Database Engine and patched OS with 99.99% availability.

Price/service tiers
Tailor price/performance ratio to your needs with flexible service tiers that span from affordable $5/month to powerful 80-core databases.

Scalability
Easily scale up, scale out, or shard your databases depending on your needs to improve performance of your application.

Single Database
Use the Single database hosted in logical servers for your SaaS aplications and microservices that need a single database with the predictable performances.

Elastic pools
SQL Database elastic pools are a simple, cost-effective solution for managing and scaling multiple databases that have varying and unpredictable usage demands.

Managed Instance
Use the Managed Instance to easily migrate your on-premises databases to the fully managed Azure PaaS database service with minimal or no database code changes.

Platform as a Service
Built-in High-availability, automated backups, and geo-replication, will prevent maintenance operations, infrastructure or hardware failures from stopping your business.

Advanced security
Secure your database with Azure AD authentication, Virtual Networks, Firewalls, Always Encrypted connections. Identify threats and vulnerabilities with built-in security.

Monitoring and tuning
Built-in monitoring and intelligent tuning help you dramatically reduce the costs of running and managing databases and maximizes performance of your application.

https://docs.microsoft.com/en-us/azure/sql-database/

Azure Synapse Analytics

- Synapse Analytics is an enterprise analytics service that includes a data warehousing product along with SQL and Spark runtimes for quickly run complex queries across petabytes of data.

Massive query concurrency

Democratize data across your enterprise.

Integrated data processing

Ingest and query from multiple data types and sources within a single solution.

Quick and easy provisioning

Provision thousands of compute cores in less than five minutes, and scale to a petabyte in hours.

Elastic design

Independently scale for performance or memory with separate compute and storage.

Advanced security

Help protect your data with virtual network service endpoints, advanced threat detection, always-on encryption, auditing, and simplified secure access.

Fully managed infrastructure

Automate infrastructure allocation and workload optimization to focus on data analysis, and use the built-in advisor to optimize your cloud data warehouse.

Strong Ecosystem

Integrate with leading data preparation and visualization vendors and get support from our partners to accelerate time to value.

Powerful SQL engine

Take advantage of Microsoft SQL Server, the industry's top-performing SQL engine, offering comprehensive support for SQL language.

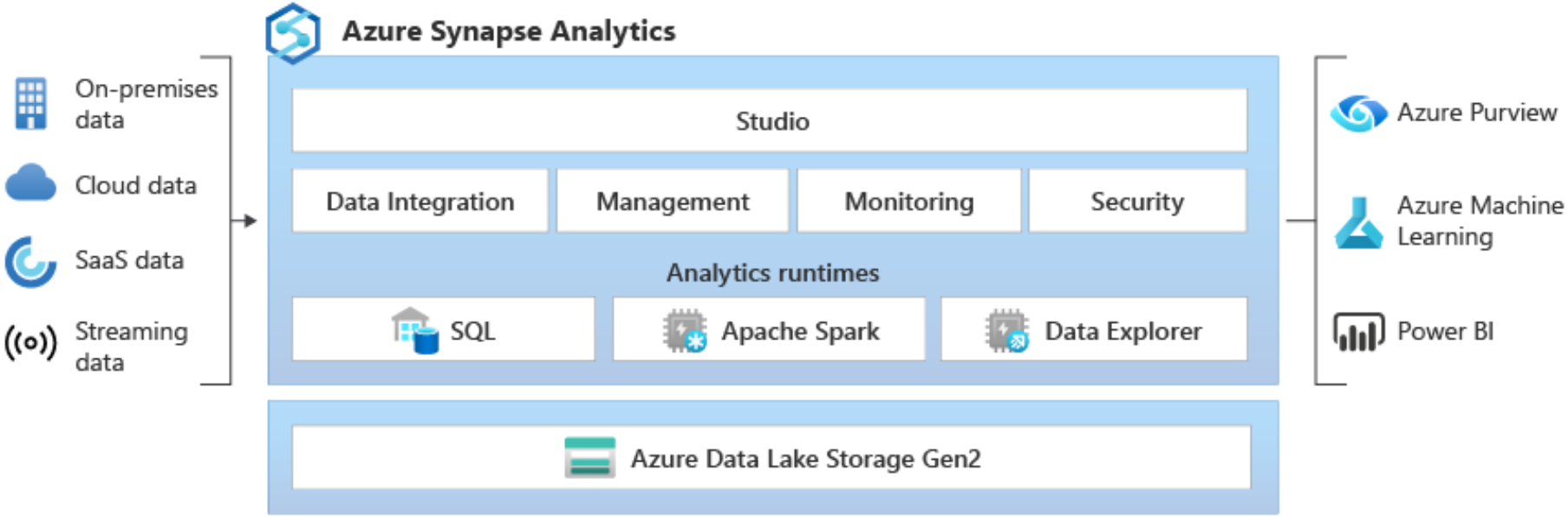
Industry-leading compliance

Help ensure peace of mind with more than 50 government and industry compliance certifications, including HIPAA.

Global availability

Benefit from availability in 40 Azure regions, the most among all cloud-based data warehouse providers.

https://docs.microsoft.com/en-us/azure/sql-data-warehouse/


Azure Synapse Analytics
On-premises data
Cloud data
SaaS data
Streaming data
Studio
Data Integration
Management
Monitoring
Security
Analytics runtimes
SQL
Apache Spark
Data Explorer
Azure Data Lake Storage Gen2
Azure Purview
Azure Machine Learning
Power BI

Pool Party!

Serverless SQL Pools

- The first and, in my opinion, most flexible pool type is the Serverless SQL Pool.
- This pool type allows you to use SQL without having to reserve a certain capacity of compute.
- You're charged based on how much data the pool processes rather than the number of physical nodes that are used.
- I would recommend this pool type for ad hoc querying of data, especially if you want to simply read and query data from your data lake.

Dedicated SQL Pools

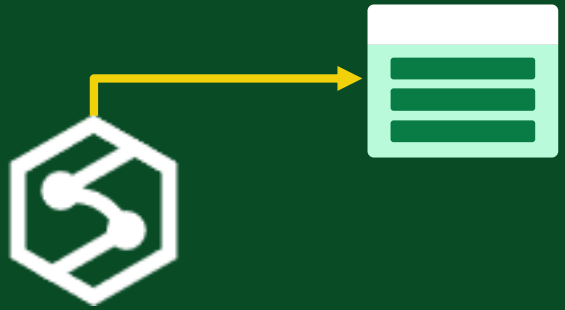
- In contrast to Serverless SQL Pools, Dedicated SQL Pools are provisioned at a specific size, and you pay for them while they are turned on, no matter how much data is being processed.
- This type of pool is ideal for larger, complex data querying or otherwise planned tasks.
  - Why? Because you can turn Dedicated Pools off and avoid charges after you're done. With Dedicated Pools, you pick a specified performance level, which is measured in obscure units called “Data Warehouse Units" or DWUs.
  - The more DWUs, the more CPU cores, memory, and data I/O are available to be utilized.

Serverless Spark Pools

- The last type of pool is a Serverless Spark Pool, which creates a Spark session in the background.
- This allows users to utilize SparkSQL functionality inside of Synapse. The benefit?
  - Highly distributed and scalable processing of data in a Spark cluster, which can speed up some complex data transformation tasks that might otherwise take longer in traditional SQL.
- This pool type is similar to the Serverless SQL Pool in that you are charged based on the amount of data that's processed.
  - This also allows you to install Spark libraries on the pool, which is perfect if you have packages that make certain data processing tasks easier.

What is Polybase?

PolyBase enables your SQL instance to process T-SQL queries that read data from an external data source such as Azure Storage locations as an “EXTERNAL TABLE".



https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-guide?view=sql-server-2017
https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-versioned-feature-summary?view=sql-server-2017

Just because you need to create a data warehouse, you may not need to you use Azure Synapse.
-But Why?

Azure SQL DB vs. Azure Synapse

Capability	Azure SQL Database	Azure Synapse
Data type	Relational	Relational
Active geo-replication	Yes	No
Dynamic Data Masking	Yes	No
Data Encryption at rest	Yes	Yes
Polybase T-SQL queries	No	Yes
Automatic Tuning	Yes	No
Massive Parallel Processing (MPP)	No	Yes
Ability to pause and resume	No	Yes
Max amount of data per database	4TB	1PB
Max concurrent open sessions	30000	1024
Max concurrent queries	6400	128

https://stackify.com/azure-sql-database-vs-warehouse/

https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-service-capacity-limits

But What About Costs?

1TB of Data

SQL DB

Factors:

- Serverless vs. Provisioned
- Redundancy
- vCores (or DTUs)

Cost Range:

- $150 - $1,000 per Month

Synapse

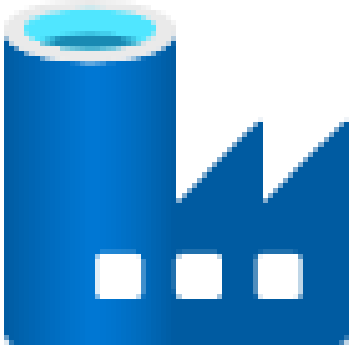
Factors:

- Pool Type
- Pool Usage
- Storage
- DWUs

Cost Range:

- $75 - $10,000 per Month

Azure Data Factory



- Azure Data Factory is a cloud-based, hybrid data integration tool for creating data pipelines at scale. Complete with 70 data sources connectors and a slick UI, think of Data Factory as SSIS in the cloud.

Productive

Build automated data integration solutions with a visual drag-and-drop UI. Move data seamlessly from over 60 sources without writing code.

Hybrid

Build data integration pipelines that span on-premises and cloud. Easily lift your SQL Server Integration Services (SSIS) packages to Azure.

Trusted

Data movement using Azure Data Factory has been certified by HIPAA/HITECH, ISO/IEC 27001, ISO/IEC 27018, and CSA STAR.

Scalable

Build serverless, cloud-based data integration with no infrastructure to manage. Take advantage of elastic capabilities to scale out with your customer growth.

Visual drag-and-drop UI

Maximize productivity by getting pipelines up and running quickly. Use the code-free drag-&-drop interface to build, deploy, monitor, and manage your data integration. Connect this visual tool directly to your Git repository for a seamless development workflow.

Multiple Language Support

Use the visual interface or write your own code in Python, .NET, or ARM to build pipelines using your existing skills. Choose from a wide range of processing services and put them into managed data pipelines to use the best tool for the job, or insert custom code as a processing step in any pipeline.

SSIS package execution in Azure

Easily execute and schedule your SSIS packages in managed execution environment. Gain high availability, scalability, and lower TCO by lifting your SSIS packages to Azure.

Code-free data movement

Improve your TCO with more than 70 natively supported connectors including Azure data services, AWS S3 and Redshift, Google BigQuery, SAP HANA, Oracle, DB2, MongoDB, and many more across multiple global points of presence.

Comprehensive control flow

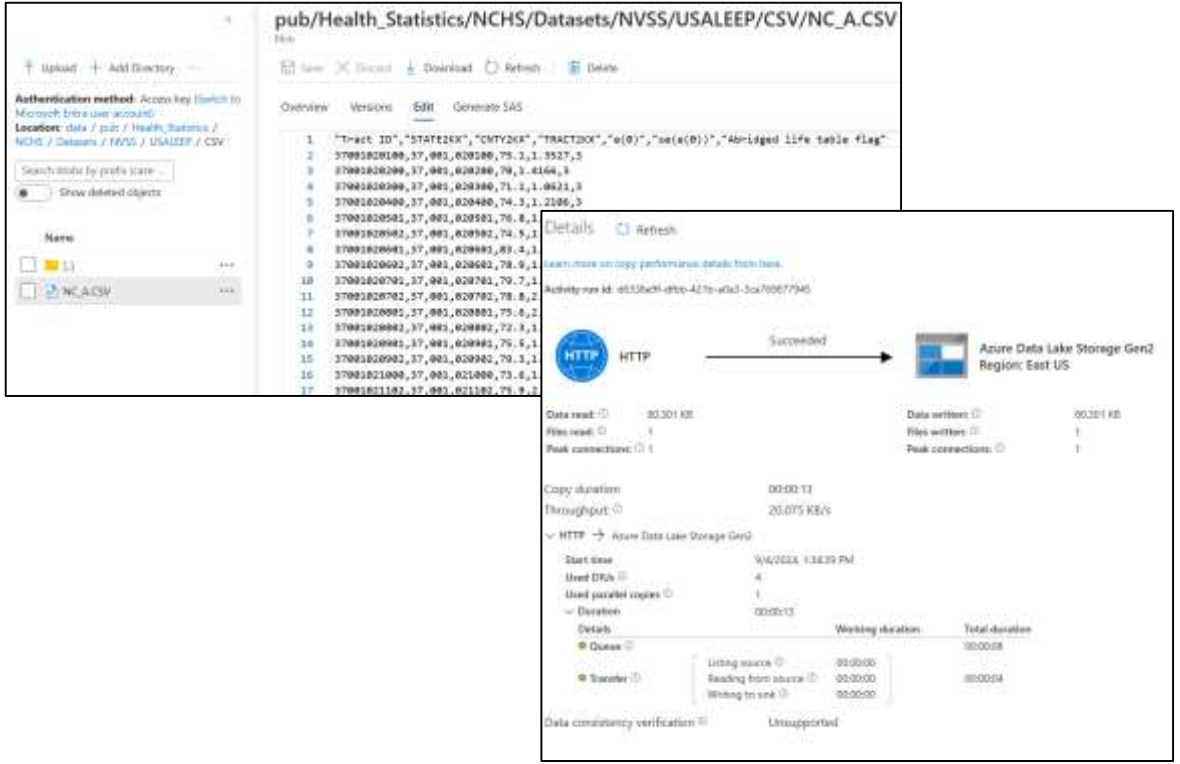
Facilitate looping, branching, conditional constructs, on-demand executions, and flexible scheduling with extensive control flow constructs.

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

usaleep | Neighborhood Life Expectancy Project

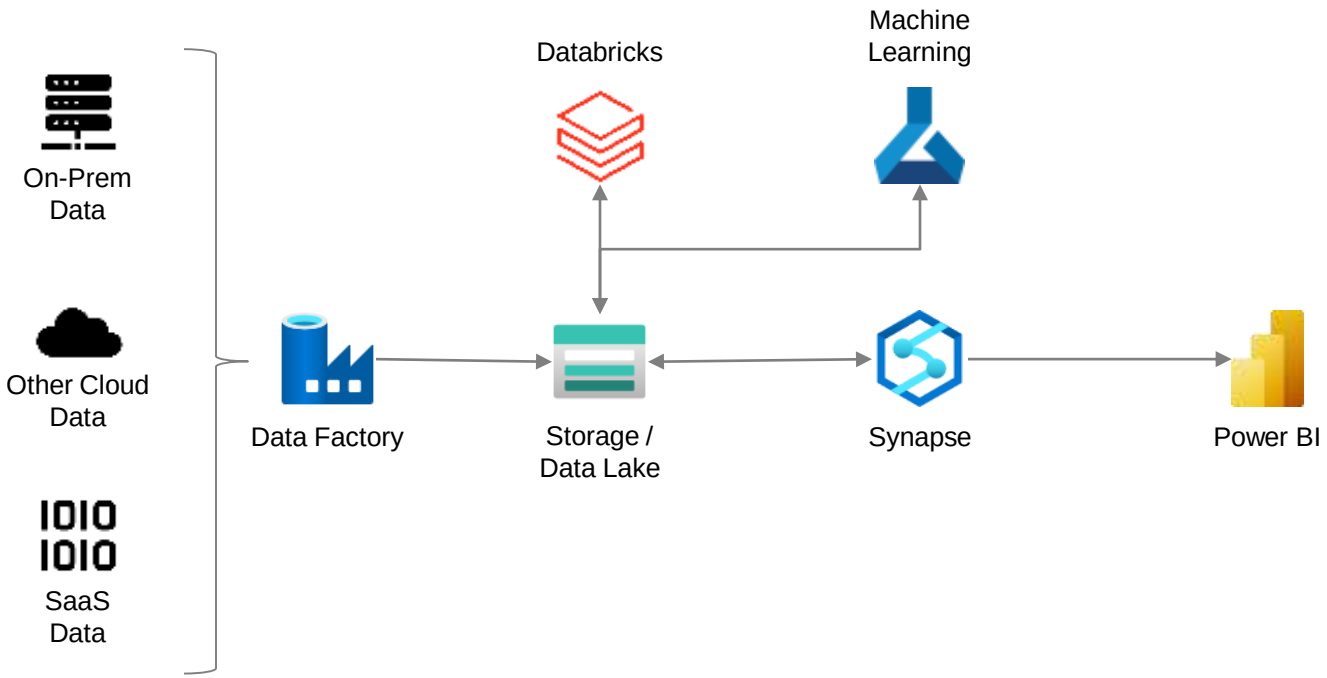
CDC Data Transfer Example

- U.S. Small-area Life Expectancy Estimates Project – USALEEP Example Data: https://www.cdc.gov/nchs/nvss/usaleep/usaleep.html
  - ftp.cdc.gov -/pub/Health_Statistics/NCHS/Datasets/NVSS/USALEEP/CSV/



Begin lab…

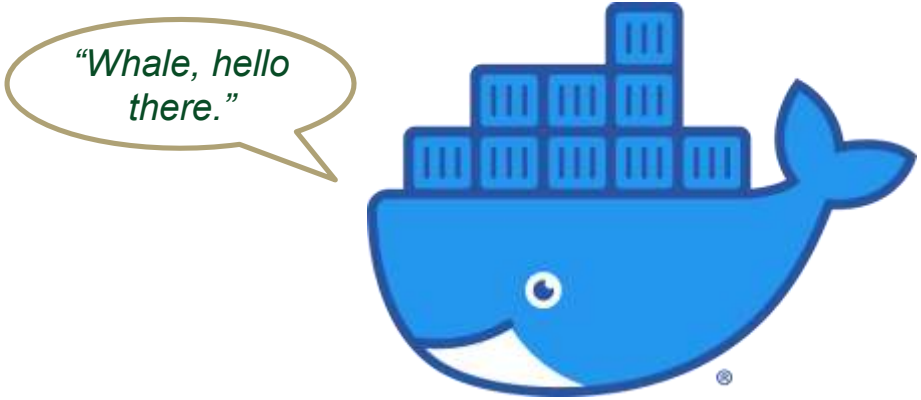
Example Architecture



https://docs.microsoft.com/en-us/azure/architecture/data-guide/

Containerization

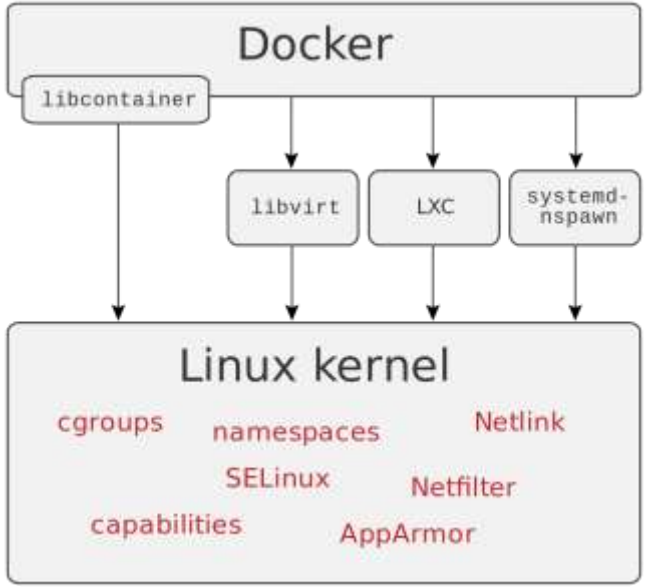
—



Docker

- Support for Linux and Windows Server containers (and now WebAssembly).
- Flexibility to support microservices and traditional app workloads.
- Integrated graphical user interface-based management and operation.
- Granular role-based access control, LDAP, and Azure Active Directory integration.
- Connection to custom networking and volumes (data storage)

- Think of Docker is a trimmed down VM image with sets of packages and settings that are configured for reuse.



Dockerfile

- Contains layers of instructions for configuring the system and installing libraries and files.

- Specifies a base layer, which is useful for “picking up where someone left off”

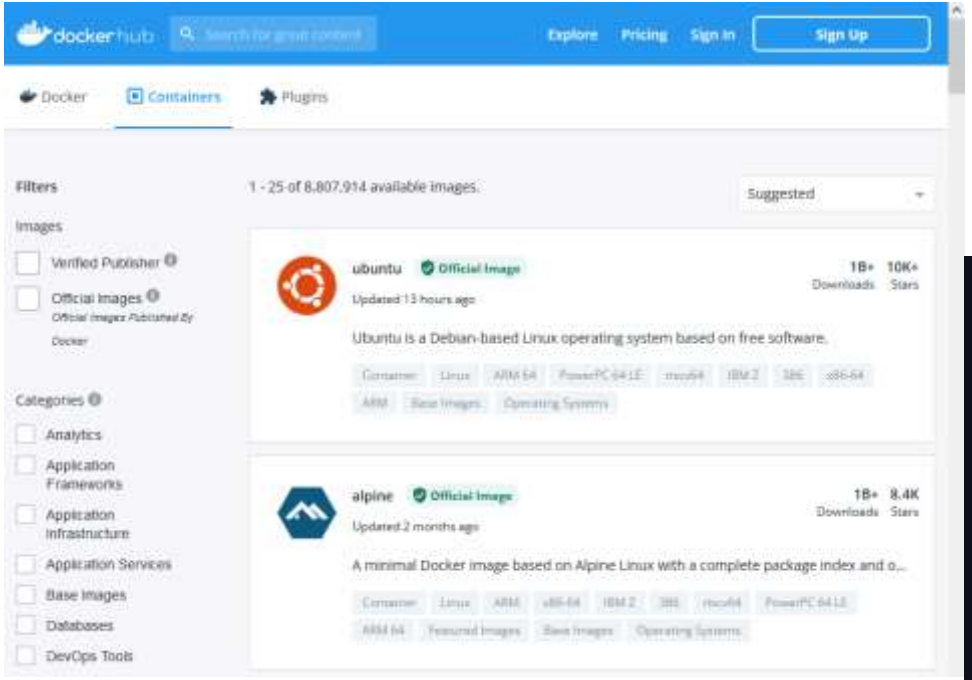
18 lines (13 sloc) | 546 Bytes

```r
FROM rocker/r-ver:4.1.2

LABEL org.opencontainers.image.licenses="GPL-2.0-or-later" \
      org.opencontainers.image.source="https://github.com/rocker-org/rocker-versioned2" \
      org.opencontainers.image.vendor="Rocker Project" \
      org.opencontainers.image.authors="Carl Boettiger <cboettig@ropensci.org>"

ENV S6_VERSION=v2.1.0.2
ENV RSTUDIO_VERSION=2021.09.2+382
ENV DEFAULT_USER=rstudio
ENV PATH=/usr/lib/rstudio-server/bin:$PATH

RUN /rocker_scripts/install_rstudio.sh
RUN /rocker_scripts/install_pandoc.sh

EXPOSE 8787

CMD ["/init"]
```

40 lines (34 sloc) | 1.08 KB

```dockerfile
FROM nvidia/cuda:11.3.1-base-ubuntu20.04

# Install some basic utilities
RUN apt-get update && apt-get install -y \
    curl \
    ca-certificates \
    sudo \
    git \
    bzip2 \
    libx11-6 \
 && rm -rf /var/lib/apt/lists/*

# Create a working directory
RUN mkdir /app
WORKDIR /app

# Create a non-root user and switch to it
RUN adduser --disabled-password --gecos '' --shell /bin/bash user \
 && chown -R user:user /app
RUN echo "user ALL=(ALL) NOPASSWD:ALL" > /etc/sudoers.d/90-user
USER user

# All users can use /home/user as their home directory
ENV HOME=/home/user
RUN chmod 777 /home/user

# Set up the Conda environment
ENV CONDA_AUTO_UPDATE_CONDA=false \
    PATH=/home/user/miniconda/bin:$PATH
COPY environment.yml /app/environment.yml
RUN curl -sLo ~/miniconda.sh https://repo.continuum.io/miniconda/Miniconda3-py39_4.10.3-Linux-x86_64.sh \
 && chmod +x ~/miniconda.sh \
 && ~/miniconda.sh -b -p ~/miniconda \
 && rm ~/miniconda.sh \
 && conda env update -n base -f /app/environment.yml \
 && rm /app/environment.yml \
 && conda clean -ya

# Set the default command to python3
CMD ["python3"]
```

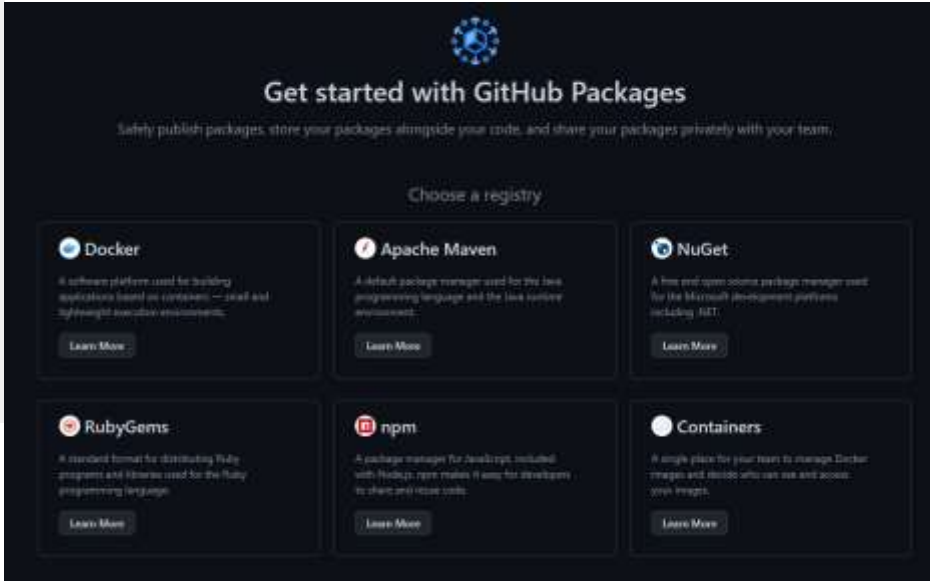
Container Storage



DockerHub

Azure Container Registry

GitHub Packages



Container Usage

- Deploy containerized applications to Azure for quick and scalable use.
- Pick the Runtime Stack (Node, .NET, Python, etc.) and OS (Linux or Windows)
- Pick the size of machine (RAM, CPUs, GPUs)

Azure Web Apps

Azure Container Instances

Azure Kubernetes Service

Container Orchestration

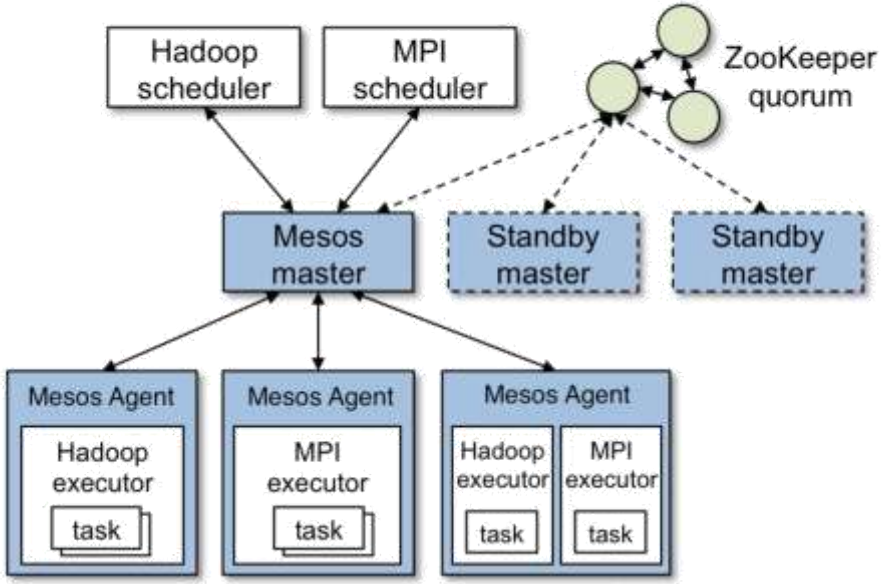
- Once containers are created, use container orchestration to organize, coordinate, and schedule their use.

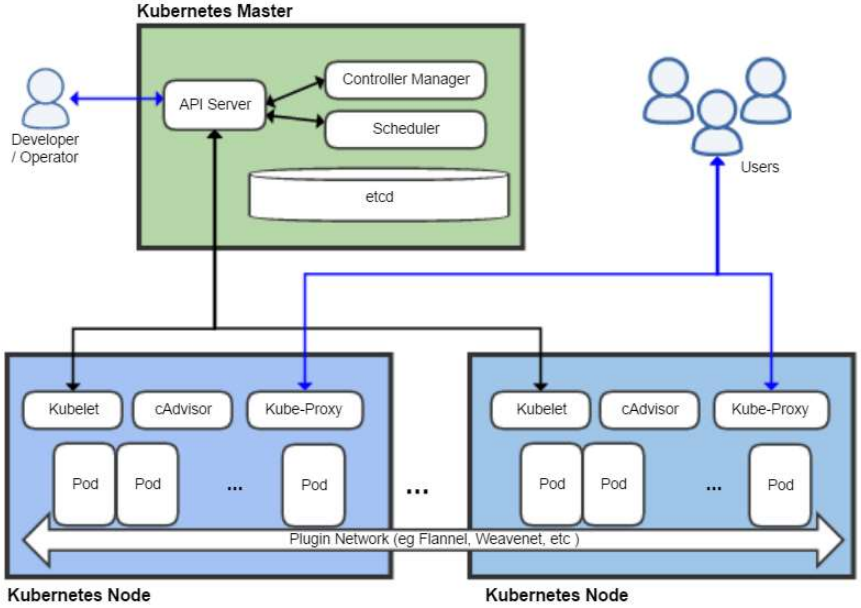
- Useful for scaling containers and making use of distributed environments

- Other capabilities:
  - Security management
  - API Serving
  - Resource Monitoring
  - Load Balancing

https://azure.microsoft.com/en-us/resources/videos/azure-kubernetes-service-overview/

Mesos Architecture





Kubernetes + Queue-Based Architecture