**Azure Synapse**

**Overview**

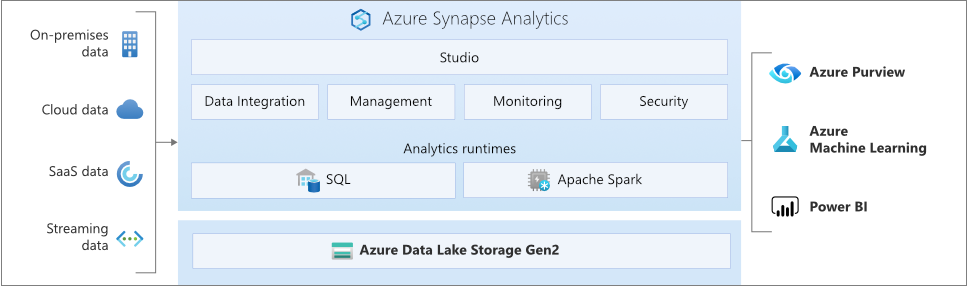
Azure Synapse Analytics is a cloud-based enterprise data warehouse. Synapse analytics service brings together data warehouse of SQL, big data analytics capabilities of Spark and data integration technologies between and from external data sources. This combination provides a unified experience to ingest, explore, prepare, manage, and serve data for immediate BI and machine learning needs. Syanpse platform having controls to manage compute and storage independently at scale.

Azure Synapse contains the same Data Integration engine and experiences as Azure Data Factory, allowing you to create rich at-scale ETL pipelines without leaving Azure Synapse Analytics.

Ingest data from 90+ data sources Code-Free ETL with Data flow activities Orchestrate notebooks, Spark jobs, stored procedures, SQL scripts, and more

**Azure Synapse Architecture**

Synapse SQL leverages a scale out architecture to distribute computational processing of data across multiple nodes. Compute is separate from storage, which enables you to scale compute independently of the data in your system.

[](file:///Users/nnagaraju/sleeve/hcsc/docker-demo/synapse-demo/synapse-architecture.png)

**MPP - Massively Parallel Processing**

Uses many separate CPUs running in parallel to execute a single program and segments communicate using high-speed network between nodes

| **Component** | **Details** |
| --- | --- |
| Control Node | Brain of SQL DW holds metadata, statistics. Determines parallel plan for client query |
| Data Movement Service DMS | Moves data around as needed, distributes sub-query to each compute node |
| Compute Node | Runs Azure SQL Server DB, executes query on its subset of data |

**Data Warehouse Unit**

A DWU represents an abstract, normalized measure of compute resources and performance. Analytic resources are defined as a combination of CPU, memory, and IO and these three resources are bundled into units of compute scale called Data Warehouse Units (DWUs).

**Synapse Storage**

The data in Synapase is sharded into distributions to optimize the performance of the system where sharding pattern to use to distribute the data when a table is defined. The folloiwng are distributions:

| **Hash** | **Round Robin** | **Replicated** |
| --- | --- | --- |
| CREATE TABLE fact\_sales | CREATE TABLE dim\_order | CREATE TABLE dim\_prdct |
| productkey int not null | orderid int not null | productkey int not null |
| ,orderdatekey int not null | ,orderdatekey int not null | ,productcatg varchar not null |
| ,duedatekey int not null | ,duedatekey int not null | ,productdesc varchar not null |
| ,shipdatekey int not null |  |  |
| ) WITH | ) WITH | ) WITH |
| (DISTRIBUTION = HASH(productkey), | (DISTRIBUTION = ROUND\_ROBIN, | (DISTRIBUTION = REPLICATE, |
| CLUSTERED INDEX(orderdatekey), | CLUSTERED COLUMNSTORE INDEX | CLUSTERED INDEX(productcatg) |
| PARTITION (orderdatekey | ); | ); |
| RANGE RIGHT FOR VALUES |  |  |
| (20000101,20201231,))); |  |  |

**Database Tables - Indexes**

Clustered Columnstore Indexes Clustered Index Heap Nonclustered Inded

Highest level of compression Best Overa

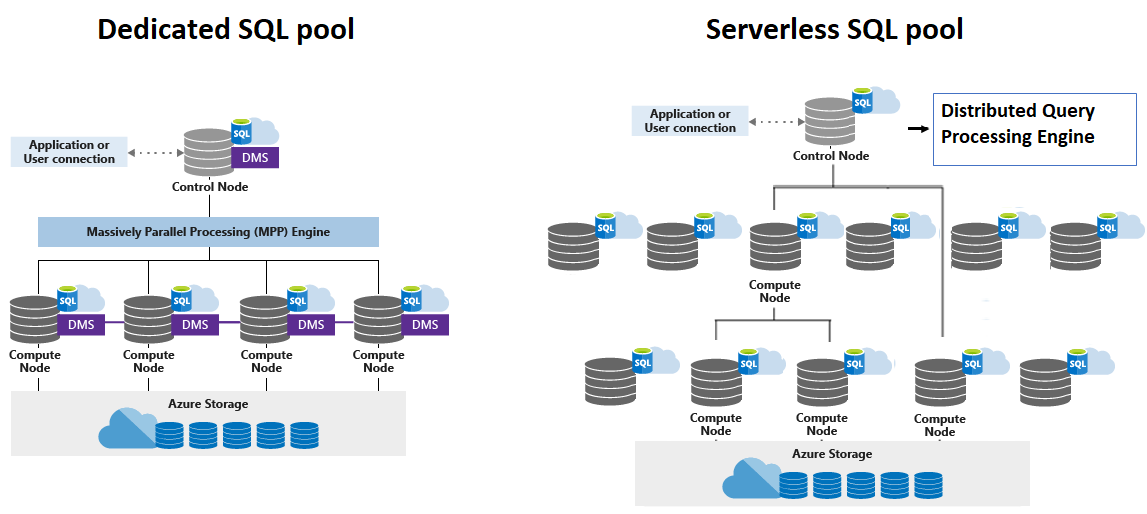
**Database Tables - Views**

Materialized Views Indexed Materialized Views

**Synapse SQL Pool**

Synapse SQL is a distributed query system for T-SQL that enables data warehousing and data virtualization scenarios and extends T-SQL to address streaming and machine learning scenarios.

Synapse SQL offers both serverless and dedicated resource models. For predictable performance and cost, create dedicated SQL pools to reserve processing power for data stored in SQL tables. For unplanned or bursty workloads, use the always-available, serverless SQL endpoint.

[](file:///Users/nnagaraju/sleeve/hcsc/docker-demo/synapse-demo/sql-architecture.png)

CREATE MASTER KEY ENCRYPTION BY PASSWORD = 'S0me!nfo' ;

-- Create a database scoped credential with Azure storage account key as the secret.

CREATE DATABASE SCOPED CREDENTIAL AzureStorageCredential

WITH

IDENTITY = '<my\_account>' ,

SECRET = '<azure\_storage\_account\_key>'

;

-- Create an external data source with CREDENTIAL option.

CREATE EXTERNAL DATA SOURCE MyAzureStorage

WITH

(

LOCATION = 'wasbs://daily@logs.blob.core.windows.net/' ,

CREDENTIAL = AzureStorageCredential ,

TYPE = HADOOP

);

-- Create an external file format

CREATE EXTERNAL FILE FORMAT MyAzureCSVFormat

WITH

(

FORMAT\_TYPE = DELIMITEDTEXT,

FORMAT\_OPTIONS(

FIELD\_TERMINATOR = ',',

FIRST\_ROW = 2)

);

--Create an external table

CREATE EXTERNAL TABLE dbo.FactInternetSalesNew

WITH

(

LOCATION = '/files/Customer', DATA\_SOURCE = MyAzureStorage,

FILE\_FORMAT = MyAzureCSVFormat

);

SELECT TOP 10 \*

FROM

OPENROWSET(

BULK 'puYear=2018/puMonth=9/\*.snappy.parquet',

DATA\_SOURCE = 'YellowTaxi',

FORMAT='PARQUET'

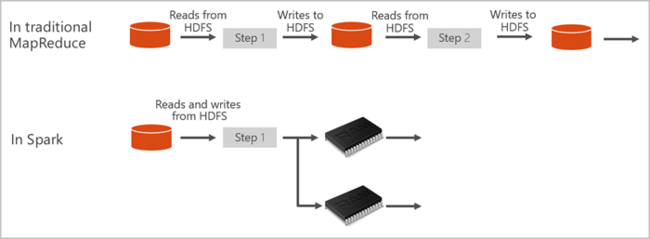
) AS nyc

;

**Synapse Spark Pool**

Apache Spark is a parallel processing framework that supports in-memory processing to boost the performance of big-data analytic applications with use cases data engineering and machine learning. Azure Synapse makes it easy to create and configure a serverless Apache Spark pool. The benefits of Spark Pool in Synapse Analytics to include

| **Benefit** | **Details** |
| --- | --- |
| Speed and efficiency | Spark instances start in 2 minutes for <60 nodes and approx 5 minutes for >60 nodes |
| Ease of creation | Create using Azure portal, PowerShell, or the Synapse Analytics .NET SDK |
| Ease of use | Custom notebook derived from Nteract, can be used for interactive data processing and visualization |
| REST APIs | Apache Livy, a REST API-based Spark job server to remotely submit and monitor jobs |
| ADLS Generation 2 | Can use Azure Data Lake Storage Generation 2 as well as BLOB storage |
| IDEs | Syanapse IDE plugin for IntelliJ IDEA to create and submit applications to a Spark pool |
| Pre-loaded Anaconda libraries | 200 libraries for machine learning, data analysis, visualization |
| Scalability | pools scale by adding or removing nodes as needed. |

[](file:///Users/nnagaraju/sleeve/hcsc/docker-demo/synapse-demo/map-reduce-vs-spark.png)

**Read Synapse Table using Spark**

val scalaDataFrame = spark.read.sqlanalytics("mySQLPoolDatabase.dbo.mySQLPoolTable")

scalaDataFrame.createOrReplaceTempView( "mydataframetable" )

val df = spark.sql("SELECT \* FROM mydataframetable")

df.show

**Dynamic Management Views - DMVs**

These are queries that return information about model objects, server operations and server health.

* Benefits
  + all open, closed sessions
  + count sessions by user
  + count completed queries by user
  + all active, complete queries
  + longest running queries
  + memory consumption

**Azure Synapse Studio**

A single place for Data Engineers, Data Scientists, and IT Pros to collaborate enterprise analytics. Studio is divided into Activity hubs to build analtics solution.

| **Activity** | **Details** |
| --- | --- |
| Overview Hub | A dashboard with important links to tasks, artifacts and documentation, allows to view recently opened artifacts |
| Data Hub | Explore the data in workspace as well as linked storage accounts, |
| Data Hub - Storage Accounts | ADLS2 files, view data, POSIX ACLs. Analyze the data either T-SQL or Spark Notebook. Script from multiple selected files. |
| Data Hub - Databases | Explore different databases available in workspace - SQL pool, SQL On-Demand, Spark. Generate SQL Scripts from table objects, Spark code into notebook to load table into a dataframe |
| Data Hub - Datasets | A linked service defined dataset, can be used as source or sink or both |
| Develop Hub | Provides development experience to query, analyze, model data. Multiple languages, switch between notebnooks and scripts |
| Develop Hub - SQL Scripts | Run SQL scripts either on SQL or On-Demand Pools. View results chart or table |
| Develop Hub - Notebook | Multiple languages, temporary tables can be used across languages |
| Develop Hub - Dataflow | code-free. A visual way of specifying how to transform data |
| Develop Hub - Power BI | Create Power BI reports, realtime update from Synapse, publish reports |
| Orchestrate Hub | Provides ability to create pipelines to ingest, transform and load with connectors |
| Monitor Hub | Provides ability to monitor orchestration, activities and resources |
| Monitor Hub - Orchestration | Monitor progress and status of pipeline execution |
| Monitor Hub - Spark | Monitor Spark pools, Spark applications for the progress and status |
| Manage Hub | Provides ability to manage Linked Services, Orchestration and Security |
| Manage - Linked Services | Defines connection information to connect to external resources 90+ connectors |
| Manage - Access Control | Provides access control management on artifcats and pools |
| Manage - Triggers | Defines a process unit that determines when a pipeline executes |

**Developer Tools**

There are multiple options available based on the platform.

* Azure Cloud Service (Portal)
* Visual Studio - SSDT databse projects (Windows)
* Azure Data Studio (Windows/Linux/macOS)
* SQL Server Management Studio (Windows)
* Visual Studio Code (Windows/Linux/macOS)

**Continous Integration and Delivery (CI/CD)**

Continuous Integration (CI) is the process of automating the build and testing of code every time a team member commits changes to version control. Continuous Deployment (CD) is the process to build, test, configure, and deploy from multiple testing or staging environments to a production environment.

In an Azure Synapse Analytics workspace, continuous integration and delivery (CI/CD) moves all entities from one environment (development, test, production) to another. To promote your workspace to another workspace, there are two parts. First, use an Azure Resource Manager template (ARM template) to create or update workspace resources (pools and workspace). Then, migrate artifacts (SQL scripts, notebook, Spark job definition, pipelines, datasets, data flows, and so on) with Azure Synapse Analytics CI/CD tools in Azure DevOps.

Azure DevOps release pipeline to automate the deployment of an Azure Synapse workspace to multiple environments.

**Snapshots and Restores**

Copy the data warehouse state automatically throught the day and are available for 7 days.