- MPP clustered architecture
- OLAP Azure database
- HTAP (hybrid transactional processing) => analyse operational data @ location such as Azure Cosmos DB via Azure Synapse Link

Structure

Concepts

- Analytics engine for large volumes of data
- Ingest external sources/Data Lake/DBMS then transform/aggregate
- Read & process data locally (external sources) => repeatedly query same data
  - Further processing with Azure Analytics Services
- Massively parallel processing (MPP) database/store
  - Control nodes & pool compute resources
  - Control node (brain) = front end, interacts with apps, optimize & co-ordinate parallel queries (batch)
  - Compute node (CPU) = even distribution, co-ordinated by control node, results returned to control node
- SQL (T-SQL) & Spark pools
  - Apache Spark & automated pipelines for MPP

Components

- Components
  1. SQL pools: T-SQL
  2. Spark pool: Apache server cluster for Spark ML/Azure ML
  3. Pipelines: Similar to databricks pipelines (connect to 90+ sources and codeless workflow)
  4. Links: HTAP connect to Cosmos DB (near real-time analytics)
  5. Studio: Web user interface for Data engineering access, tools, create pools/pipelines/links. Manage serverless/provisioned resources/security

Pools

- Spark Pools
  - Spark cluster & notebooks (C#/Python/Scala/Spark)
  - Visualize data (notebooks)
  - Coexist with SQL pools in same instance
  - Many data formats supported
  - Split queries into concurrent parallel tasks
  - In-memory cluster compute
    - Load, cache and requery data => increased performance and decreased memory resources
    - Autoscale while running tasks
  - Used by data engineers for data preparation
    - Machine learning (Apache Spark ML library with Anaconda & Python)
  - Save data to Azure Storage Account/Data Lake

- SQL Pools
  - Distributed T-SQL queies
    - Split data into distributions
    - distribution = basic unit of storage for processing parallel queries on distributed data
  - Move data accross nodes => **Data Movement Service (DMS)**
  - On-demand pools
    - Default, for external files
  - Provisioned pools
    - Ingest/load intio Synapse Analytics
  - Can manually scale (when not running) >= 60 nodes
  - Pause pool & resume (in minutes)
  - Used for complex reporting & data ingestion
    - Azure SQL Database/Data Lake/Storage
  - Query different sources (polybase) => relational/non-relational
    - Polybase => Hadoop/Spark/Azure Blob Storage/Cosmos DB/Oracle/Teradata/MangoDB/SQL Server [not SQL DB]

Links

- Synapse Analytics Link
  - Uses Cosmos DB analytical store
    - Copy of Cosmos DB container as column store (column aggregations)
    - Automatically syncs
  - Analyse Cosmos DB directly (no ETL) via Cosmos DB Analytics Services
  - SQL/Spark pool for near real-time analysis
  - Used for
    - Supply chain analytics/forecast
    - Operational reporting
    - Batch data ingestion/orchestration
    - Real-time in-app personlization
    - IoT maintenance

Compiled by Dr T Oberholster
DP-900 version: March 2021