**Quick Notes**

**Designing a Data Schema**

*What Is a Database Schema?*

* Defines how data is organized
* Provides meaning to database
* Physical and logical categories

***Star Schema vs. Snowflake Schema***

**Star Schema**

* Easy to query
* 1 fact table
* 1 dimension table level
* Not normalized

**Snowflake Schema**

* Reduced query performance
* Multiple dimension tables
* Normalized

***Data Warehouse Introduction***

* Repository of Information
* Data that is organized into a schema that describes the layout and type of data. Stored in tables containing no redundant data.

**Cloud Data Warehouse**

* Scalability
* Cost
* Time to market
* Performance

**Relational vs. Non-Relational**

* Relational Databases - SQL
  + Primary and foreign keys
  + ACID (Atomity, Consistency, Isolation, Durability)
* Non-Relational Databases - NoSQL
  + Families

**Data Models and Fact Tables**

* Dimensions: Attributes
* Fact Table: Center of a Schema

**Key Component in Big Data Solutions**

* Ingest : Data Factory
* Store: Blob Storage
* Prep: Databricks, HDInsight
* Mode: SQL Data Warehouse
* Report : Power BI

**Massively Parallel Processing (MPP) Engine**

* Data Warehouse units
* Separation of compute and storage
* Massively parallel processing tool for running complex queries across petabytes of data

**Distributions**

* The law of 60, compute, and control
* Sharding Patterns
  + Hash: Highest query performance for large tables
  + Round Robin: Easiest to create for staging tables
  + Replicate: Fastest query for small tables

***PolyBase Introduction***

**PolyBase**

* Enables queries across relational data stored in SQL DW and non-relational Hadoop data stored in Hadoop Distributed File System (HDFS)

**Allows you to bypass traditional ETL in favor of ELT**

* Perform transformations in the data warehouse
* Better distributes transformation across multiple databases

**Basic Steps**

* Extract the source data into text files.
* Land the data into Azure Blob storage or Azure Data Lake Store.
* Prepare the data for loading.
* Load the data into SQL Data Warehouse staging tables using PolyBase.
* Transform the data.
* Insert the data into production tables.

**Optimize SQL Data Warehouse**

* Have you checked for Advisor recommendations?
* PolyBase
  + Use PolyBase for loading and exporting data
  + Use CREATE TABLE AS SELECT
* Hash-distribute large tables
  + Choose the correct distribution type
* Remember the Rule of 60
  + Don't over-partition
  + Experiment for partitioning to determine workload
  + Maximize throughput by breaking gzip into 60+ files
* Shrink it down
  + Minimize transaction sizes
  + Minimize column sizes
* Troubleshooting Azure SQL Data Warehouse

DWU - Data Warehouse Unit: Measurement used in SQL Data Warehouse