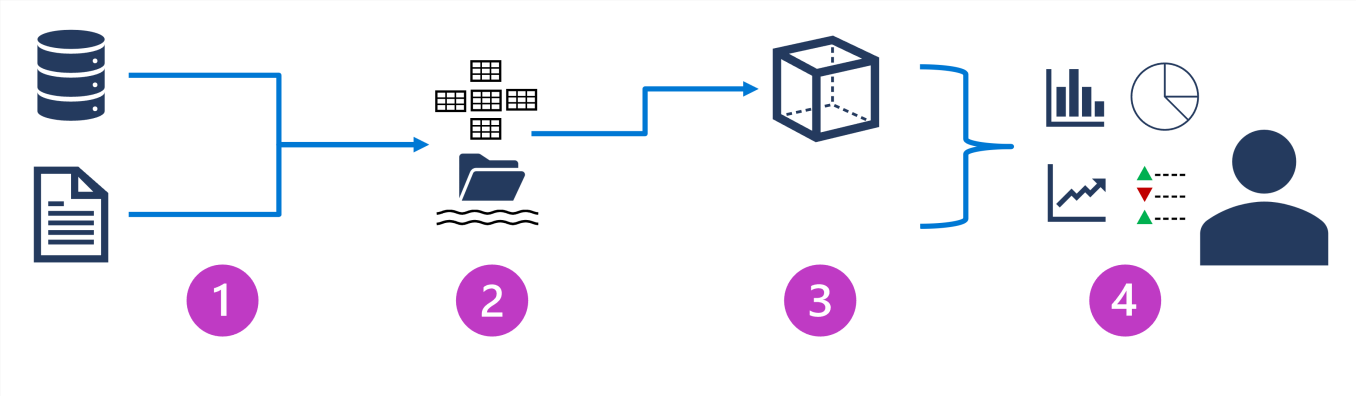
Introduction

- Large-Scale Data Analytics solutions combine conventional data warehousing and data lakehouse techniques to integrate data from files and external sources
- A conventional data warehouse solution typically involves copying data from transactional data stores into a relational db with a schema that's optimized for queryigna nd building multidimensional models
- Data lakehouse solutions on the other hand, are used with large volumes of data in multiple formats, which is batch loaded or captured in real-time streams and stored in a data lake from which distributed processing engines like Apache Spark are used to process it.

Describe Data Warehousing Architecture

- Large-scale data analytics architecture can vary, as the specific technologies used to implement
- In General the following are included



1. Data Ingestion and processing

    - Data from one or more transactional stores, files, or real-time streams, or other sources is loaded into data lake or a relational data warehouse
    - Load operation involves and extract, transform, and load (ETL) or extract, load, transform (ELT) process where the data is cleaned, filtered, and restructured for analysis
    - In ETL processes, the data is transformed before being loaded into an analytical stor
    - In ELT processes, the data is copied into the store, then transformed
    - Either way the result is a structure that is optimized for analytical queries
    - Data processing is often performed by distributed systems that can process high volumes of data in parallel using multi-node clusters
    - Data ingestion includes both batch processing of static data and real-time processing of streaming data

2. Analytical data store

- Data stores for large scale applications include
    - Relational Data Warehouses
    - File-system based data lakes
    - Hybrid architectures
- Combine features of data warehouses and data likes
- Sometimes called data lakehouses or lake databases

3. Analytical data model

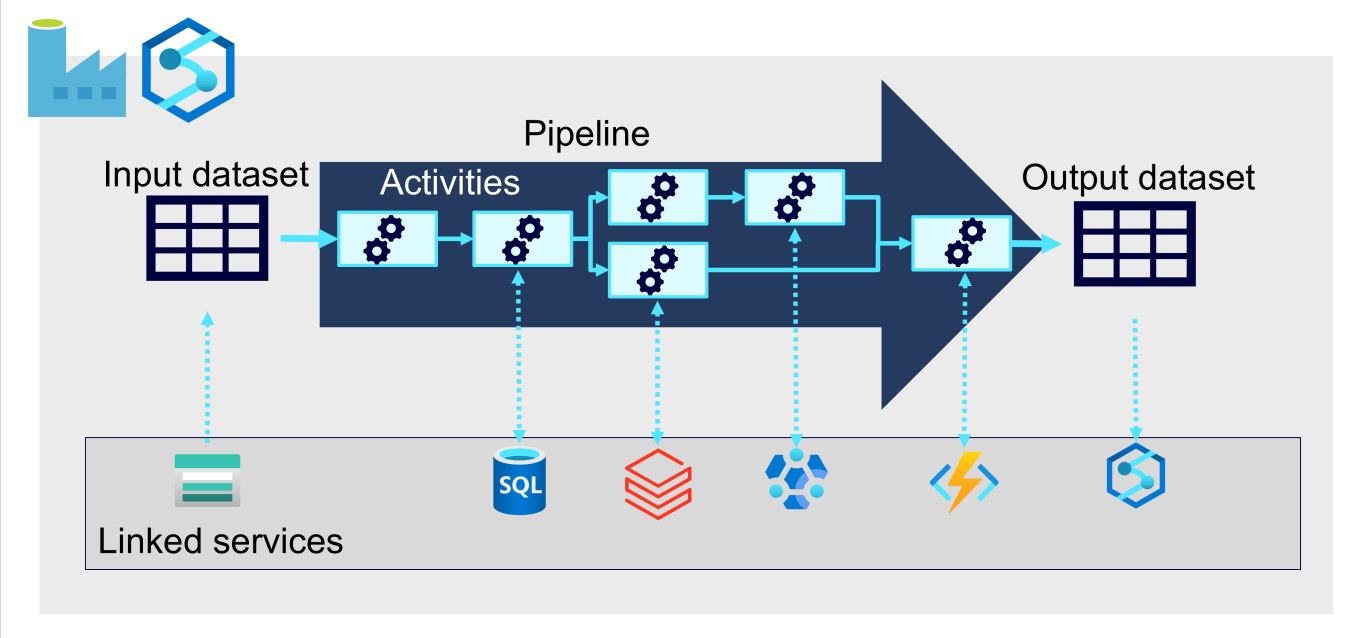
    - In analytical data store its common to create one or more data models that pre-aggregate the data to make it easier to produce reports, dashboards, and interactive visualizations
    - Often data models are described as cubes
    - Numeric values are aggregated across one or more dimensions (for example to determine total sales by product and region)
    - Model encapsulates relationships between data values and dimensional entities to support "drill up/down" analytics

4. Data Visualization

    - Consume data from Analytical models and directly from analytical stores to create reports, dashboards and other visualizations
    - Users who may not be tech professionals might perform self-service data analysis and reporting
    - Visualizations from the data show trends, comparisons, and KPIs and can take the form of reports, graphs, and charts

Explore Data Ingestion Pipelines

- How Data is ingested into an analytical data store from one or more sources



- Large scale ingestion is best implemented by creating pipelines that orchestrate ETL processes
- You can create and run pipelines using Azure Data Factory
- Use similar pipeline engine in Azure Synapse Analytics or Microsoft Fabric if you want to manage all of the components of data analytics solution in a unified workspace
- Pipelines consist of one or more activities that operate on data
- Input dataset provides the source data
- Activities are the flow of data that incrementally manipulates the data until an output dataset is produced
- Pipelines can connect to external data sources to integrate with a wide variety of data services.

Analytical Data Stores

- Two common types of analytical data store
    - Data Warehouse
        - Relational database in which the data is stored in a schema that is optimized for data analytics rather than transactional workloads
        - Commonly the data is transformed into a schema in which numeric values are stored in central fact tables, which are related to one or more dimension tables
        - Represent entities by which the data can be aggregated
        - Fact table might contain sales order data, which can be aggregated by customer, products, store, and time dimensions - Easily find monthly totals sales revenue by product for each store - Star Schema (Often extended into a snowflake schema by adding additional tables related to the dimension tables) - Example, product might be related to Product Categories
        - Data Warehouse is a great choice when you have transactional data that can be organized into a structured schema of tables, and you want to use SQL to query them
    - Data Lakehouse
        - Distributed file system for high performance data access
        - Spark or Hadoop are used to process queries on stored files and return data for reporting and analytics
        - Schema-on-read approach to define tabular schemas on semi-structured data files at the point where the data is read for analysis without applying constraints when its stored
        - Great for supporting a mix of structured, semi-structured, and unstructured data without need for schema enforcement when data is written to the store
- Hybrid approach combines data lakes and data warehouses into data lakehouse or lakebase.
    - Raw data is stored as file and relational storage layer abstracts the underlying files and exposes them as tables.
    - Can be queried using SQL
    - SQL pools in Azure Synapse analytics include Polybase - enables you to define external tables based on files in a data lake (and other sources) and query them using SQL
    - Synapse analytics also supports a Lake database approach in which you can use database templates to define the relational schema of your warehouse, while storing the underlying data in lake storage, separating the storage and compute for your data warehousing solution.
    - Relatively new approach in Spark-based systems, and are enabled through technologies like Delta Lake, which adds relational storage capabilities to Spark
        - Define Tables that enforce schemas and transactional consistency
        - Support batch-loaded and streaming data sources
        - Provide a SQL for API querying

Explore PaaS Solutions

- Three main PaaS services to implement large-scale analytical store
    - Azure Synapse Analytics
        - Unified e2e solution for large scale data analytics
        - SQL server based relational data warehouse with flexibility of data lake and Apache Spark
        - Native support for log and telemetry analytics
        - Built in data pipelines for ingestion and transformation
        - Can be managed through a single user interface (Azure Synapse Studio)
            - Create notebooks in which Spark Code and markdown can be combined

        - Synapse analytics is a great choice when you want to create a single unified analytics solution
    - Azure Databricks
        - Azure implementation of Databricks platform
        - Comprehensive data analytics solution built on Apache Spark
        - Offers native SQL capabilities as well as workload-optimized Spark Clusters for data analytics and data science
        - Databricks provides an interactive user interface through which the system can be managed and data can be explored in interactive notebooks
        - Due to its common use on multiple cloud platforms, you might want to consider using Azure Databricks as your analytical store if you want to use existing expertise with the platform, or if you need to operate in a multicloud environment or support a cloud-portable solution
    - Azure HDInsight
        - Azure service that supports multiple open-source data analytics cluster types
        - Not as user friendly as Azure Synapse Analytics and Data Bricks
        - Suitable option if your analytics solution relies on multiple open-source frameworks or if you need to migrate an existing on-prem Hadoop-based solution to the cloud
- Each of the services can be thought of as an analytical data store
    - Provide schema and interface through which the data can be queried
- Many of the cases data is actually stored in a data lake, and then a service is used to process the ata and run queries
- Some solutions might even combine the use of these services
- ETL ingestion process might copy data into the data lake, and then use one of these services to transform the data, and another to query it.
- Pipeline might use a MapReduce job running in HDInsight or a notebook running Azure Databricks to process a large volume of data in the data ake, and then load into tables in a SQL pool in Azure Synapse analytics

Explore Data Analytics in Azure with Azure Synapse Analytics

- Provision Azure Synapse Analytics Workspace
- Ingest Data
- Use SQL Pool to Analyze Data
- Use Spark Pool to analyze Data

Explore Microsoft Fabric

- Provision Workspace
- Create Lakehouse
- Ingest Data
- Query data in a lakehouse
- Visualize Data in a lakehouse