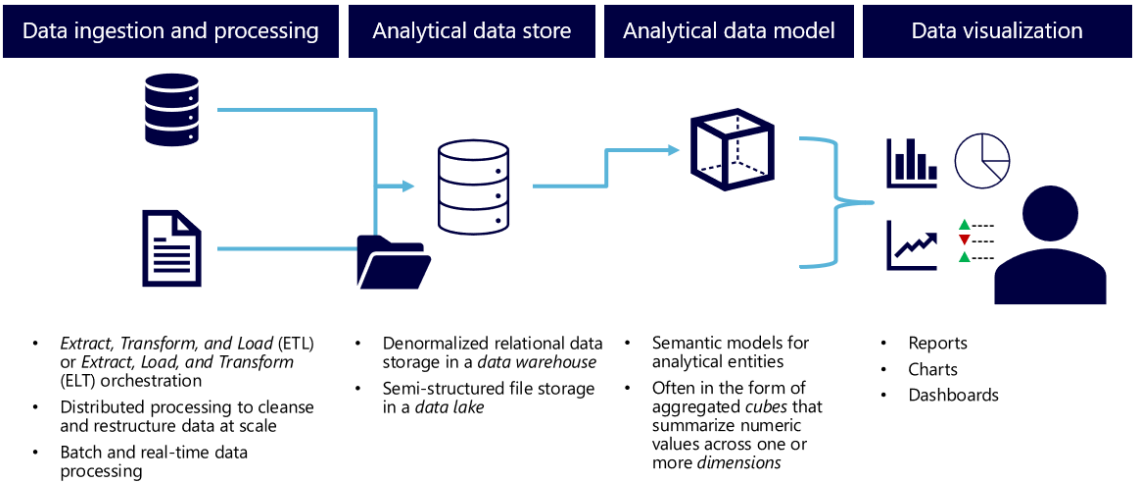
Modern Data Warehousing

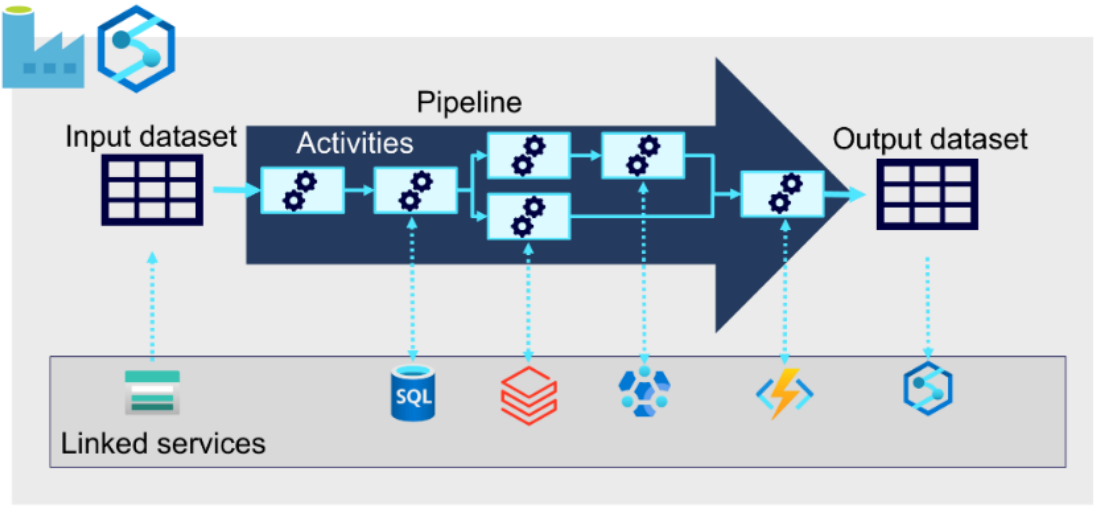
What is modern data warehousing?



Data Ingestion Pipelines:

On Azure, large-scale data ingestion is best implemented by creating *pipelines* that orchestrate ETL processes.

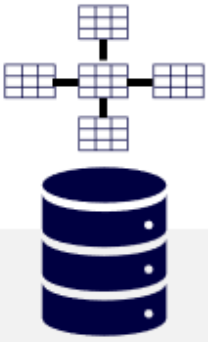
You can create and run pipelines using Azure Data Factory, or you can use the same pipeline engine in Azure Synapse Analytics



Analytical data store:

1. Data Warehouse
2. Data Lake

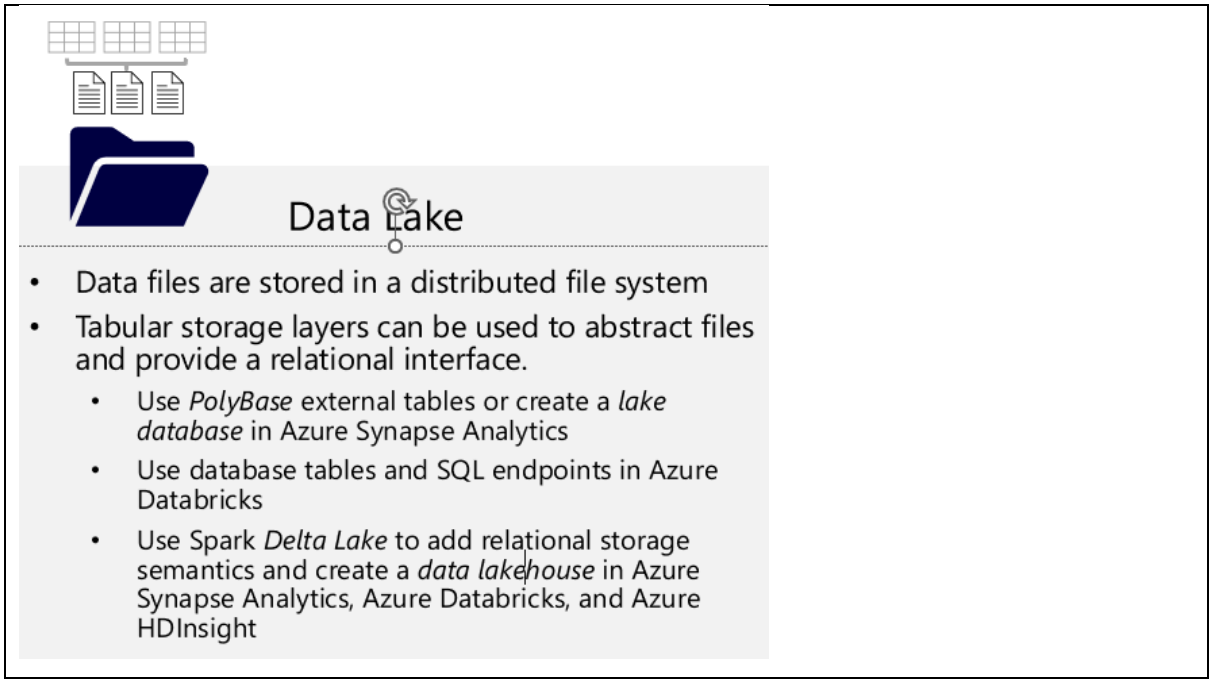
Data Warehouse:



Data Warehouse

- Large-scale relational database store and query engine
- Data is *denormalized* for query optimization
  - Typically, as a *star* or *snowflake* schema of numeric *facts* that can be aggregated by *dimensions*

Data Lake:



Azure Services for Analytical Store:

Azure Synapse Analytics:

Azure Synapse Analytics is an integrated analytics platform, which combines **data warehousing, big data analytics, data integration**, into a single environment.

- Synapse SQL is a distributed query system for T-SQL and offers serverless and dedicated resource models
- Apache Spark for Azure Synapse is used for data preparation, data engineering, ETL, and machine learning.
- Data Integration engine, provides experiences as Azure Data Factory, allowing you to create rich at-scale ETL pipelines without leaving Azure Synapse Analytics.
- Azure Synapse Data Explorer pool, provides native support for log and telemetry analytics
- Synapse Analytics is a great choice when you want to create a single, unified analytics solution on Azure.

Azure Databricks:

- Azure Databricks is an Azure implementation of the popular Databricks platform.
- Databricks is a comprehensive data analytics solution built on Apache Spark.
- It 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.
- Azure Databricks can be choice ,when you want to leverage existing expertise with the platform or if you need to operate in a multi-cloud environment or support a cloud-portable solution

Azure HDInsight:

- **Azure HDInsight** is an Azure service that supports multiple open-source data analytics cluster types.
- This platform is not as user-friendly as Azure Synapse Analytics and Azure Databricks
- It can be a suitable option if your analytics solution relies on multiple open-source frameworks or if you need to migrate an existing on-premises Hadoop-based solution to the cloud.

Demo:

Exercise: Explore Azure Synapse Analytics - Learn | Microsoft Docs

Batch Processing and Stream Processing

Data processing is simply the conversion of raw data to meaningful information through a process.

There are two general ways to process data:

- Batch processing, in which multiple data records are collected and stored before being processed together in a single operation.
- Stream processing, in which a source of data is constantly monitored and processed in real time as new data events occur.

Batch Processing:

- Newly arriving data elements are collected and stored, and the whole group is processed together as a batch.
- You can process data based on a scheduled time interval or it could be triggered when a certain amount of data has arrived, or as the result of some other event.

Use Cases:

Telephone billing application:

- A telecom company runs a monthly batch job to process call data records that include the details of millions of phone calls to calculate charges.

Report generation:

- A manufacturer produces a daily operational report for a production line that is run in a batch window and delivered to managers in the early morning.

Stream Processing:

Each new piece of data is processed when it arrives.

Data is processed as individual units in real-time rather than being processed a batch at a time

Use Cases:

Monitoring Sensor Data

- Used for aircrafts in gaining insights on faults as they occur before coming to a major problem which also results in fewer maintenance delays and improves flight safety.
- Monitoring temperature, pressure etc in oil, gas industry to guarantee the integrity of oil and gas production.

Real-time stock trades

- A financial institution tracks changes in the stock market in real time, computes value-at-risk, and automatically rebalances portfolios based on stock price movements.

Batch Vs Stream Processing

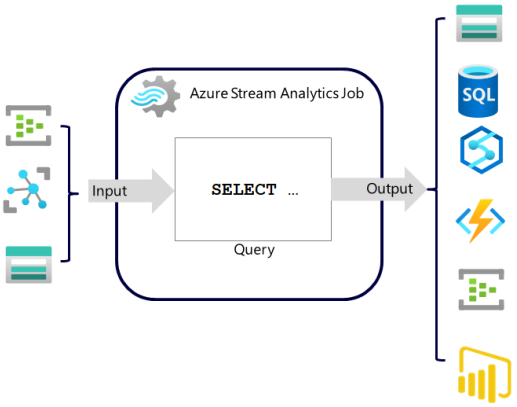
Batch	Stream
Batch processing can process all the data in the dataset	Stream processing typically only has access to the most recent data received, or within a rolling time window.
Batch processing is suitable for handling large datasets efficiently.	Stream processing is intended for individual records or *micro batches* consisting of few records.
The latency for batch processing is typically a few hours.	Stream processing typically occurs immediately, with latency in the order of seconds or milliseconds.
You typically use batch processing for performing complex analytics.	Stream processing is used for simple response functions, aggregates, or calculations such as rolling averages.

Azure Stream Analytics

- Azure Stream Analytics is fully managed, real-time analytics service designed to process fast moving streams of data from multiple sources simultaneously.
- Stream Analytics is used to:
    - Ingest data from an input, such as an Azure event hub, Azure IoT Hub, or Azure Storage blob container.
    - Process the data by using a query to select, project, and aggregate data values.

- Write the results to an output, such as Azure Data Lake Gen 2, Azure SQL Database, Azure Synapse Analytics, Azure Functions, Azure event hub, Microsoft Power BI, or others.

Azure Stream Analytics Job:



Demo:

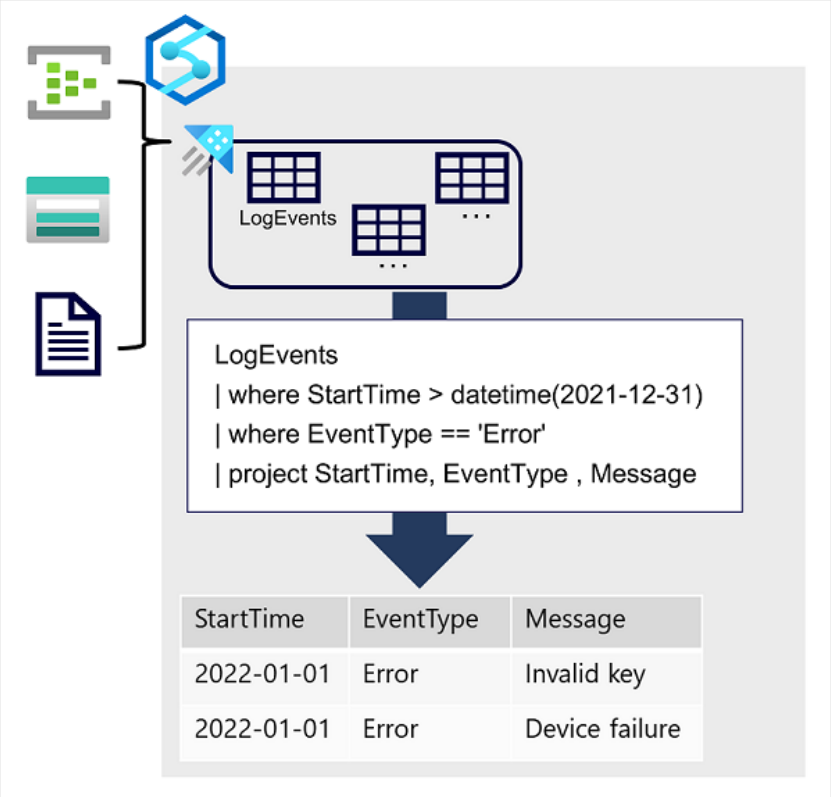
[Exercise: Analyze streaming data - Learn | Microsoft Docs]

Azure Data Explorer

- Azure Data Explorer is a standalone Azure service for efficiently analyzing data.
- Use the service as the output for analyzing large volumes of diverse data from data sources such as websites, applications, IoT devices, and more.
- The service is also encapsulated as a runtime in Azure Synapse Analytics, where it is referred to as Azure Synapse Data Explorer
- Data Explorer supports batching and streaming in near real time to optimize data ingestion.
- The ingested data is stored in tables in a Data Explorer database, where automatic indexing enables high-performance queries.

Azure Data Explorer is a great choice of technology when you need to:

- Capture and analyse real-time or batch data that includes a time-series element; such as log telemetry or values emitted by Internet-of-things (IoT) devices.
- Explore, filter, and aggregate data quickly by using the intuitive and powerful Kusto Query Language (KQL).



Kusto Query Language (KQL):

KQL is a language that is specifically optimized for fast read performance – particularly with telemetry data that includes a timestamp attribute.

Example:

1. Return all the data in LogEvents Table

```
LogEvents
```

2. Get **StartTime**, **EventType**, and **Message** columns from the **LogEvents** table for errors that were recorded after December 31st 2021.

```
LogEvents
| where StartTime > datetime(2021-12-31)
| where EventType == 'Error'
| project StartTime, EventType , Message
```

Demo:

[Exercise: Explore Azure Synapse Data Explorer - Learn | Microsoft Docs]

Example Queries:

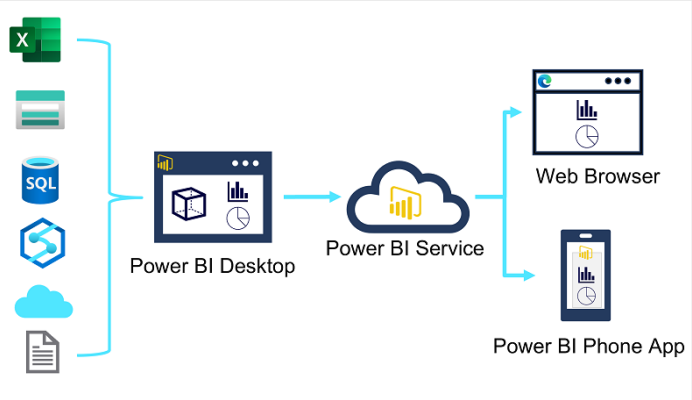
```
devices
```

devices | take 1000
devices | where Device == 'Dev1' | where Time > datetime(2022-01-07)
devices | where Time between (datetime(2022-01-01 00:00:00) .. datetime(2022-07-01 23:59:59)) | summarize AvgVal = avg(Value) by Device | sort by Device asc

Data Visualization with PowerBI

Microsoft Power BI is a suite of tools and services that data analysts can use to build interactive data visualizations for business users to consume.

PowerBI Workflow:



Power BI Desktop:

- Bring data from a wide range of data sources, combine and organize the data from these sources in an analytics data model, and create reports that contain interactive visualizations of the data.

Power BI service:

- Publish reports to cloud based service , create new visualizations build dashboards. Share dashboards with others.

Power BI Phone apps and Web browser:

- View and interact with shared dashboards and reports through web browser or on mobile devices by using the Power BI phone app. Available on any device, with native mobile BI apps for Windows, iOS, and Android.

Data Modelling:

- Data Modelling is the process of creating data model which defines the data structure, properties, and relation.
- You can connect to multiple data sources using a relationship.

Fact Table:

- Fact Table is a table which contains Measures which need to be analyzed.

  Example: Sales Amount, sales orders, stock balances, exchange rates, temperatures
- It contains numeric measure column which can be aggregated for analysis and dimension key columns that relate to dimension tables
- Each fact table relates to one or more dimensions tables.

Example:

FactSales
CustomerId
ProductId
TimeId
SalesAmount
SalesQty

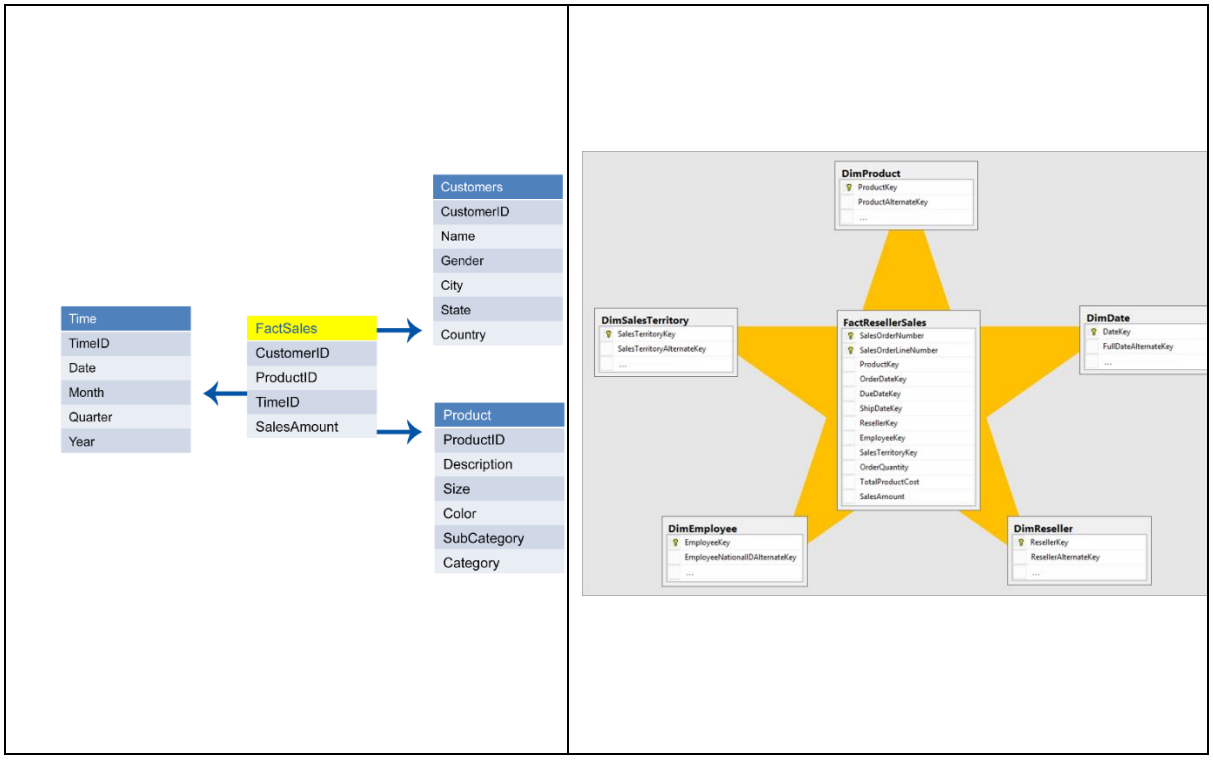
Dimension table:

- It contains the attributes based on which you need to summarize or analyze the data.
- Example: Product,Customer dimension
- It contains textual, descriptive information
- The most consistent table you'll find in a star schema is a date dimension table

Example:

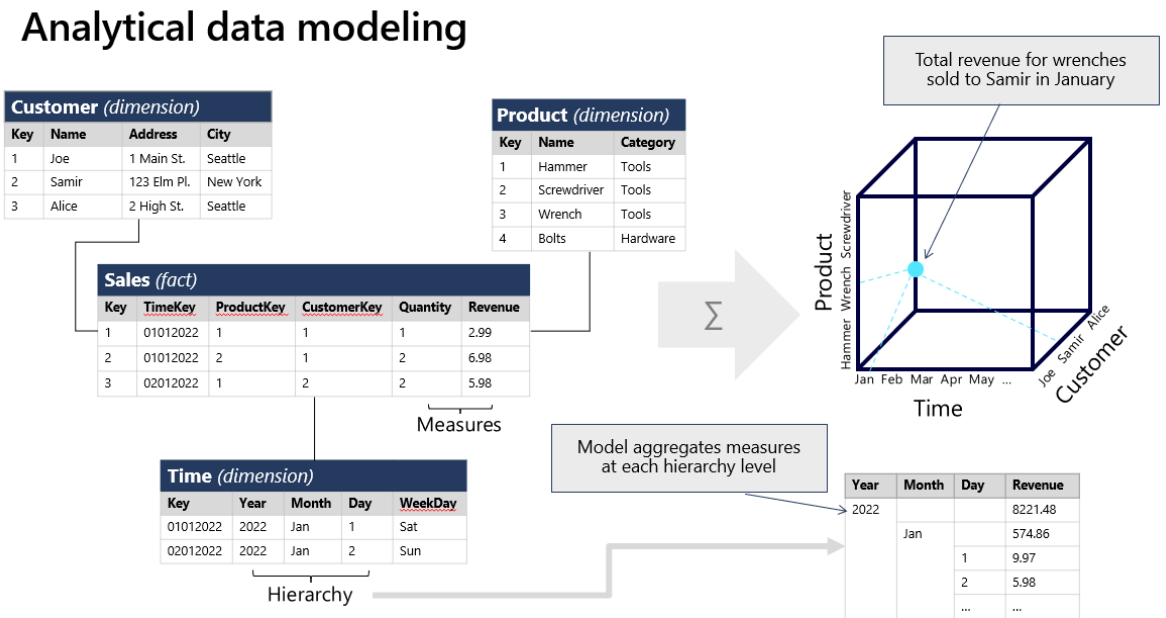
Product
ProductId
ProductName

<u>Dimension Modeling Star Schema Example:</u>



Attribute Hierarchies:

- Attribute *hierarchies* that enable you to quickly *drill-up* or *drill-down* to find aggregated values at different levels in a hierarchical dimension.
- The model can be built with pre-aggregated values for each level of a hierarchy, enabling you to quickly change the scope of your analysis – for example, by viewing total sales by year, and then drilling down to see a more detailed breakdown of total sales by month.



Visuals:

1. **Tables and Text:**

   Tables are useful when numerous related values must be displayed, and individual text values in cards can be a useful way to show important figures or metrics.

2. **Bar and Column Charts:**

   Bar and column charts are a good way to visually compare numeric values for discrete categories.

3. **Line charts:**

   Line charts can also be used to compare categorized values and are useful when you need to examine trends, often over time.

4. **Pie Chart:**

   Pie charts are often used in business reports to visually compare categorized values as proportions of a total.

5. **Scatter plots:**

   Scatter plots are useful when you want to compare two numeric measures and identify a relationship or correlation between them.

6. **Maps:**

   Maps are a great way to visually compare values for different geographic areas or locations.

Note: In Power BI, the visual elements for related data in a report are automatically linked to one another and provide interactivity.

Demo:

Exercise – Visualize data with Power BI - Learn | Microsoft Docs