**Introduction to Cloud Data Warehouses**

Building a data warehouse in a cloud environment has its advantages. The power of the cloud comes from connecting computer and storage resources from many physical servers together in order to provide as much or as little resources you need to move, transform, and store data.

You will use several technologies when you’re building data warehouse solutions in a cloud environment.

You'll be working with:

* Database storage technologies for ingesting data as well as making it available to analytics consumers.
* Data pipeline technologies to move data from source to warehouse, as well as between the stages of the Extract, Transform and Load (ETL) processes.
* End-to-end data warehouse solution that provides the ability to manage the various parts of a data warehouse from a single application.

A picture containing text, screenshot, font, diagram

Description automatically generated

**Lesson Outline**

**Lesson Outline**

In this lesson, you'll learn about the technologies involved in building data warehouses in the cloud.

You'll learn more about ETL and ELT, and the differences between loading before transformation and loading after transformation.

In this lesson, you'll also learn about cloud data storage options. Specifically:

* SQL cloud storage
* NoSQL cloud storage

The other technologies involved in building data warehouses in a cloud environment that we'll cover are:

* ETL pipeline services
* and streaming pipeline services

The last part of the lesson is about end-to-end cloud data warehouse solutions offered on cloud computing platforms.

After this lesson, you'll have a better understanding of all of the cloud-based services and tools that comprise a cloud data warehouse solution.

**Expert Perspective: Cloud Data Warehouses**

Data warehouses have been around for a long time but recently they have received significant attention in the data industry in relation to cloud computing. With traditional, on-premise data warehouses, an organization needed to buy physical hardware to fit the pre-defined need. Modern cloud infrastructures allow data warehouses to be scaled on-demand in response to changing needs.

This ability to scale means modern cloud infrastructures can take full advantage of massively parallel processing in a changing data environment. When you run a data warehouse in the cloud, you should design for this flexible scaling and think carefully about data ingestion if your transactional systems reside on-premises.

Some of the key benefits of moving to cloud data warehouses from on-premises include Scalability. This means large amounts of data can be loaded into the data warehouse environment, processed, and stored faster and with relative ease.

Another difference is cloud flexibility. This comes from the ability to add and remove different types of resources from data pipelines as needed. This allows for flexibility in the process as business needs change.

Cost shifting is a result of leveraging cloud database technology to perform the costly and time-consuming transformation of data as the last part of the process rather than doing it earlier as is typical with on-premises data warehouses. By pushing it later, the business can prioritize which transforms they want to complete “just in time” for the business.

**From ETL to ELT**

## ELT in Cloud Environments

Massive parallel processing and distributed file storage capabilities of cloud computing have led to new extract, transform, and load design patterns. Traditional ETL models have worked for decades but the introduction of massive scalability through cloud data warehousing has allowed us to flip the Transform and Load steps.

### ELT originated from ETL.

The letters mean the same thing, but the order is different in an important way. Since the T and L are swapped from TL to LT, this means the most significant difference between ETL and ELT is “Where does the transformation happen?”

* ETL: happens on an intermediate server.
* ELT: happens on the destination server.

This means rather than loading data directly into the final format of the destination data warehouse, data are loaded into the destination as either raw data or staging tables (or sometimes both). Only after loading is transformation performed.

The benefits of doing ELT include:

* Scalability - massive amounts of data can be loaded into the data warehouse environment with relative ease.
* Flexibility - the Transform step takes place using the same tech stack as the data warehouse runs on allowing for more flexibility in the process as business needs change.
* Cost shifting - the Transform step is often the costliest and by doing it last, Data Engineers can perform Just In Time transformations to meet the highest priority business needs first
* Better performance for large datasets
* More flexibility for unstructured (NoSQL) datasets

### Additional Resource

Here's a good Microsoft article on ELT vs. ETL titled [Design ELT Data Loading](https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/design-elt-data-loading).

**A screenshot of a computer

Description automatically generated with medium confidence**

ELT has several advantages such as better performance for large datasets, more flexibility for unstructured datasets, and better scalability in the cloud.

A close-up of a black text

Description automatically generated with low confidence

Since the business wants to prioritize some of the transforms, it's probably best to use an ELT process. This will allow you to do the transforms as the last part of the process.

# Solution: ELT

Given the scenario outlined, probably the most appropriate process would be an ELT. The bullets below help articulate why this is the best choice.

* ELT will provide cost shifting that allows the team to prioritize some of the tasks before others.
* ELT will provide scalability for the large amounts of data being transformed every day
* ELT will also allow your data engineering team to quickly and easily ingest the large amounts of data from the transactional systems.

**Cloud Managed SQL Storage**

Data Warehouses in the cloud leverage many of the same SQL style, relational databases that are used for OLTP systems.

* Oracle
* Microsoft SQL Server
* PostgreSQL
* MySQL
* MariaDB

The major cloud providers provide all of these databases as managed databases meaning the user doesn't have to manage the hardware resources to gain optimal performance.

* Microsoft Azure
  + Azure SQL Database (MS SQL Server)
  + Azure Database for MySQL
  + Azure Database for MariaDB
  + Azure Database for PostgreSQL
* GCP
  + Cloud SQL (MySQL, PostgreSQL, and MS SQL Server)
* AWS
  + Amazon RDS (MySQL, PostgreSQL, MariaDB, Oracle, MS SQL Server)

**Cloud Managed NoSQL Storage**

ELT makes it easier to use many NoSQL database management systems in Data Warehousing scenarios. These databases come in many flavors such as:

Key value

Document

Column oriented

Graph

Time series

Each of the major cloud providers offers a variety of managed NoSQL databases:

Azure - CosmosDB

Gremlin - graph database

MongoDB - document

Cassandra - column oriented

GCP

Big Table - column oriented

Firestore - document

MongoDB Atlas - document

AWS

DynamoDB - Key value

DocumentDB - document

Keyspaces = column oriented

Neptune - graph

Time stream - time series

A screenshot of a computer

Description automatically generated with medium confidence

**Cloud ETL Pipeline Services**

ETL / ELT processes rely on data pipelines often built using cloud-based tools.

Major Cloud providers

* Azure Data Factory
* AWS Glue
* GCP Dataflow

In addition to these tools, a large number of companies offer cloud-based tools for solving ETL / ELT challenges. Some of the major tool providers in this space are:

* Informatica
* Talend
* Xplenty
* Matillion

One advantage of doing ELT over ETL is the ability to load large amounts of data quickly. One excellent example of this is ingesting streaming data. In modern architectures, this streaming data is often coming from Internet of Things (IoT) devices; however, it could be coming from more traditional sources such as server or transaction logs.

Each of the major cloud platforms has offering for ingesting large amounts of streaming data:

* Azure - Streaming Analytics
* AWS - Kinesis
* GCP - Dataflow

**A screenshot of a quizzes

Description automatically generated with medium confidence**

**Cloud Data Warehouse Solutions**

The last piece of the puzzle for Cloud Data Warehousing is the data warehouse itself. Modern cloud data warehouse solutions seamlessly combine elements from Cloud Storage and Cloud Pipelines with powerful analytics capabilities. Each of the three major cloud providers has its own flavor of Cloud Data Warehouse that works seamlessly with its other cloud data engineering offerings.

* Azure Synapse
* Amazon Redshift
* GCP Big Query

A screenshot of a quiz

Description automatically generated with medium confidence

**Lesson Review**

In this lesson, we covered the components of data warehouse solutions built in the cloud. When you’re working with cloud data warehouses, you’ll have to make decisions about:

* Using ETL and ELT techniques
* Using Relational and NoSQL databases in the cloud
* The pipelines that move data through the data warehouse in ETL and ELT processes

And recognize these components in end-to-end data warehouse solutions provided by cloud vendors.