**Data Pipelines**

**Defining New Terms**

The video above includes references to a few terms that you may not be familiar with. Below are some definitions that you might find useful.

**Extract Transform Load (ETL) and Extract Load Transform (ELT):**

"ETL is normally a continuous, ongoing process with a well-defined workflow. ETL first extracts data from homogeneous or heterogeneous data sources. Then, data is cleansed, enriched, transformed, and stored either back in the lake or in a data warehouse.

"ELT (Extract, Load, Transform) is a variant of ETL wherein the extracted data is first loaded into the target system. Transformations are performed after the data is loaded into the data warehouse. ELT typically works well when the target system is powerful enough to handle transformations. Analytical databases like Amazon Redshift and Google BigQ."  
*Source:* [Xplenty.com](https://www.xplenty.com/blog/etl-vs-elt/)

This [Quora post](https://www.quora.com/What-is-the-difference-between-the-ETL-and-ELT) is also helpful if you'd like to read more.

**What is S3?**

"Amazon S3 has a simple web services interface that you can use to store and retrieve any amount of data, at any time, from anywhere on the web. It gives any developer access to the same highly scalable, reliable, fast, inexpensive data storage infrastructure that Amazon uses to run its own global network of web sites."  
*Source:* [Amazon Web Services Documentation](https://docs.aws.amazon.com/AmazonS3/latest/dev/Welcome.html).

If you want to learn more, start [here](https://docs.aws.amazon.com/AmazonS3/latest/dev/Welcome.html).

**What is Kafka?**

"Apache Kafka is an **open-source stream-processing software platform** developed by Linkedin and donated to the Apache Software Foundation, written in Scala and Java. The project aims to provide a unified, high-throughput, low-latency platform for handling real-time data feeds. Its storage layer is essentially a massively scalable pub/sub message queue designed as a distributed transaction log, making it highly valuable for enterprise infrastructures to process streaming data."  
*Source:* *Wikipedia*.

If you want to learn more, start [here](https://kafka.apache.org/intro).

**What is RedShift?**

"Amazon Redshift is a fully managed, petabyte-scale data warehouse service in the cloud. You can start with just a few hundred gigabytes of data and scale to a petabyte or more... The first step to create a data warehouse is to launch a set of nodes, called an Amazon Redshift cluster. After you provision your cluster, you can upload your data set and then perform data analysis queries. Regardless of the size of the data set, Amazon Redshift offers fast query performance using the same SQL-based tools and business intelligence applications that you use today.

If you want to learn more, start [here](https://docs.aws.amazon.com/redshift/latest/mgmt/welcome.html).

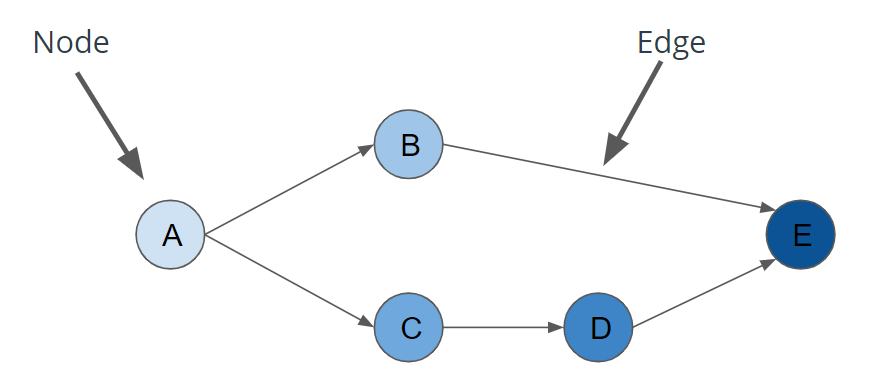
So in other words, S3 is an example of the final data store where data might be loaded (e.g. ETL). While Redshift is an example of a data warehouse product, provided specifically by Amazon.

#### Data Validation

Data Validation is the process of ensuring that data is present, correct & meaningful. Ensuring the quality of your data through automated validation checks is a critical step in building data pipelines at any organization.

**Definitions**

* **Directed Acyclic Graphs (DAGs):** DAGs are a special subset of graphs in which the edges between nodes have a specific direction, and no cycles exist. When we say “no cycles exist” what we mean is the nodes cant create a path back to themselves.
* **Nodes:** A step in the data pipeline process.
* **Edges:** The dependencies or relationships other between nodes.

[[](https://classroom.udacity.com/nanodegrees/nd027/parts/45d1c3b1-d87b-4578-a6d0-7e86bb5fea6c/modules/1fa8af52-682d-440b-a391-16183cb0df08/lessons/af88b28f-ceb2-4239-8262-2a55a3b67bd5/concepts/574d6d41-2b84-40bb-be7e-439c9682c2c6)](https://classroom.udacity.com/nanodegrees/nd027/parts/45d1c3b1-d87b-4578-a6d0-7e86bb5fea6c/modules/1fa8af52-682d-440b-a391-16183cb0df08/lessons/af88b28f-ceb2-4239-8262-2a55a3b67bd5/concepts/574d6d41-2b84-40bb-be7e-439c9682c2c6)

[Diagram of a Directed Acyclic Graph](https://classroom.udacity.com/nanodegrees/nd027/parts/45d1c3b1-d87b-4578-a6d0-7e86bb5fea6c/modules/1fa8af52-682d-440b-a391-16183cb0df08/lessons/af88b28f-ceb2-4239-8262-2a55a3b67bd5/concepts/574d6d41-2b84-40bb-be7e-439c9682c2c6)

**Common Questions**

**Are there real world cases where a data pipeline is not DAG?**

It is possible to model a data pipeline that is not a DAG, meaning that it contains a cycle within the process. However, the vast majority of use cases for data pipelines can be described as a directed acyclic graph (DAG). This makes the code more understandable and maintainable.

**Can we have two different pipelines for the same data and can we merge them back together?**

Yes. It's not uncommon for a data pipeline to take the same dataset, perform two different processes to analyze the it, then merge the results of those two processes back together.

**Tips for Using Airflow's Web Server**

* Use Google Chrome to view the Web Server. Airflow sometimes has issues rendering correctly in Firefox or other browers.
* Make sure you toggle the DAG to On before you try an run it. Otherwise you'll see your DAG running, but it won't ever finish.