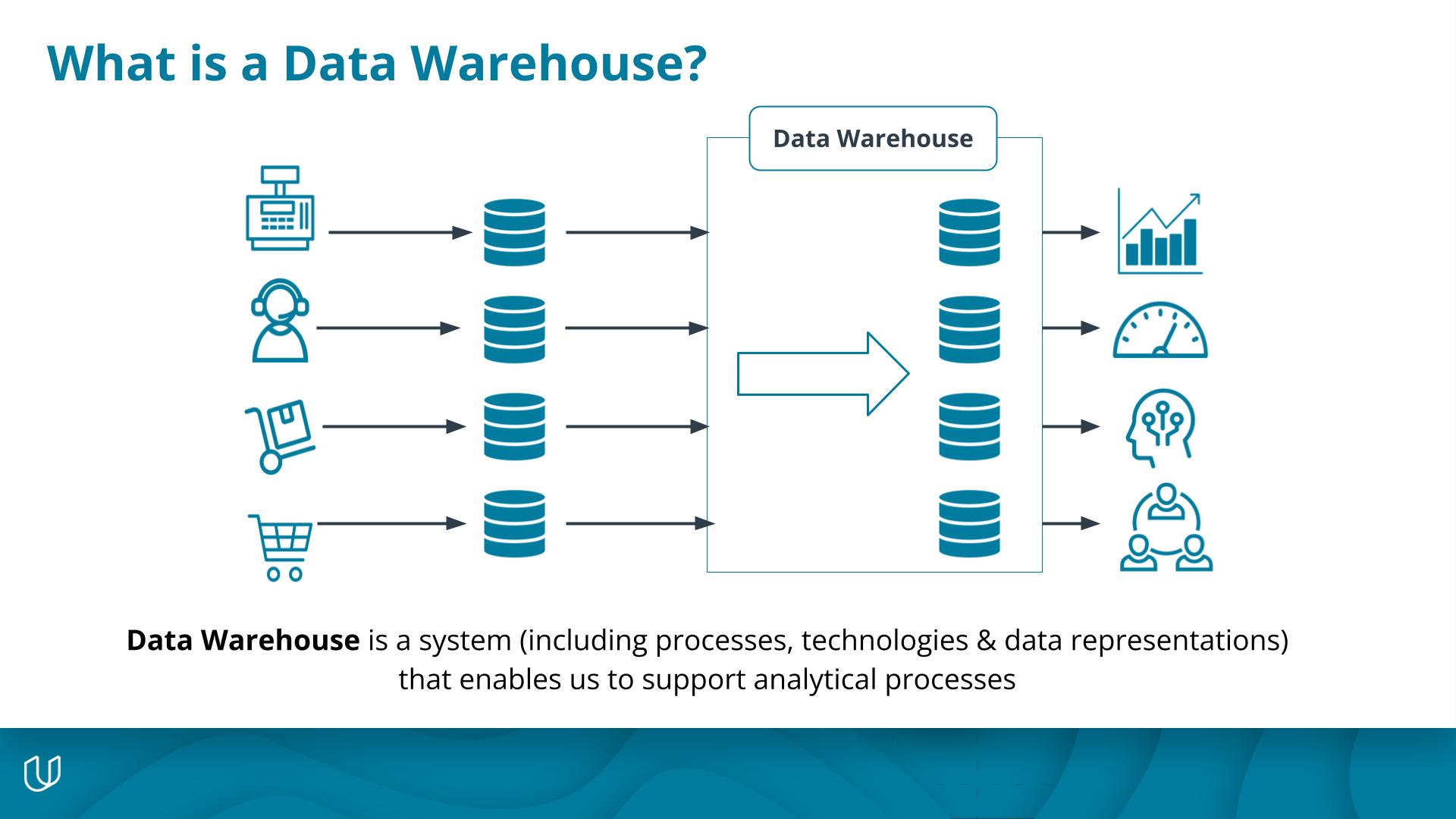
**Introduction to Data Warehouses**

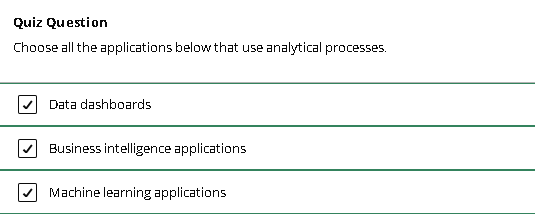
**Introduction to Data Warehouses**

The data warehouse plays a crucial role in the modern enterprise, storing and serving data for data visualization, analytics, and machine learning applications. As a data engineer, you will likely be tasked with designing and building these important data platforms.



A data warehouse is a system including processes, technologies & data representations that enables us to support analytical processes.

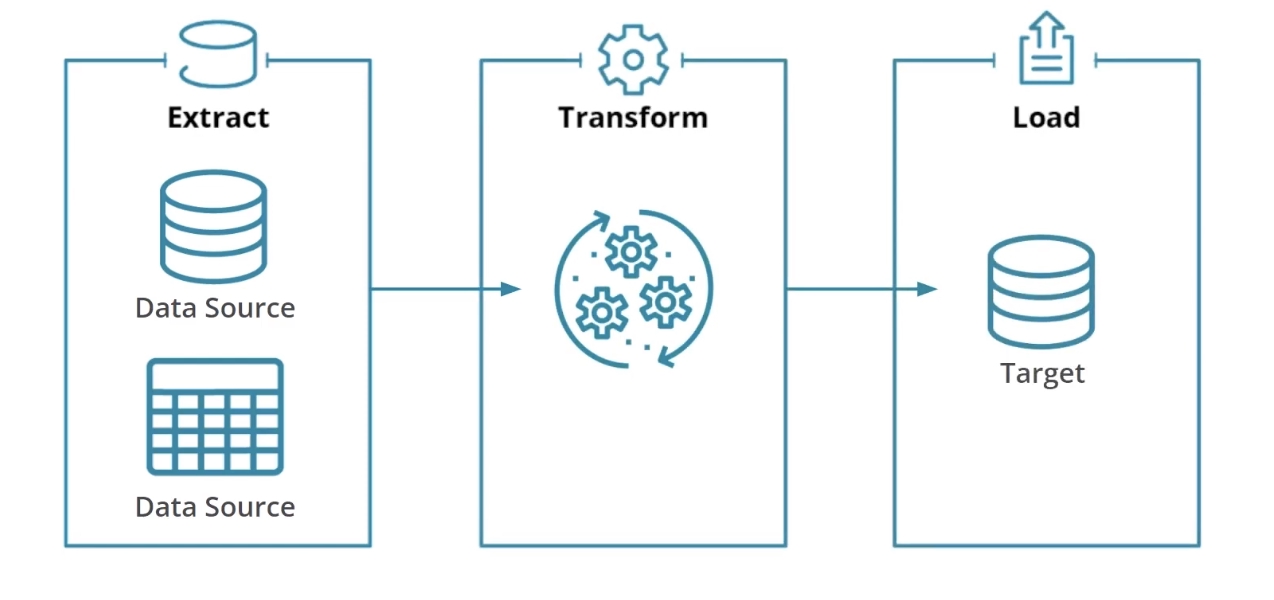
In this course, you'll learn about the purpose of a data warehouse and some techniques and technologies used to build them. By the end of the course, you'll learn how to build a data warehouse solution in the cloud.



# Project Preview

## Project Preview

At the end of the course, you'll use all of the skills you learned to build a data warehouse solution on the AWS cloud. You'll extract data from AWS S3, transform it using SQL and Python, and load it into an AWS Redshift data warehouse.



### Project Description

In this project, you'll act as a data engineer for a music streaming startup called Sparkify.

Sparkify has grown its user base and song database and wants to move its processes, data, and data analytics applications onto the cloud.

Their data resides in S3, in a directory of JSON logs on user activity on the app, and a directory with JSON metadata for the songs in their app. As their data engineer, you will be tasked with building an ETL pipeline that extracts their data from S3, stages them in Redshift, and transforms data into a set of dimensional tables for their analytics team to continue finding insights into what songs their users are listening to.

# Course Overview

## Course Outline

In this course, you'll work through several lessons of concepts with interactive quizzes and exercises.

You'll learn about **data warehouses**, including:

* The business case for data warehouses
* Data warehouse architecture
* Dimensional modeling for data warehouses
* SQL to SQL Extract, transform, and load (ETL) for data warehouses
* OLAP Cubes

Then, we'll cover **cloud data warehouse** concepts like

* Extract, transform, and load vs Extract, load and transform (ELT)
* Managed database services in the cloud
* Cloud storage and ETL pipeline services
* Cloud data warehouse solutions

You'll work with **Amazon Web Services** **(AWS)** data services you'll need to know in order to build AWS data warehouse solutions by:

* Using AWS S3 Storage
* Creating AWS Redshift resources
* Writing Python scripts that interact with AWS data tools using the AWS BOTO3 SDK

Finally, you'll use ETL techniques to build and optimize **AWS Redshift data warehouses**.

02: Introduction to Data Warehouses

# Lesson Introduction

## Lesson Introduction

In this lesson, you'll learn about the business need for data warehouses, their architecture, and data modeling for them.

### Lesson Overview

In the first part of this lesson, we’ll talk about the business needs for data analytics and how these needs gave rise to the creation of data warehouses including:

* The types of data businesses generate and the questions they need to answer with this data.
* How data from transactional systems is not suitable for data analytics
* The **online transactional processes** (OLTP) that generate data from transactional systems
* How they are different from **online analytical processes** (OLAP) for data analytics

In the next part of the lesson, we’ll talk about the design of a data warehouse

* You’ll learn about dimensional modeling
* As well as ETL processes for data warehousing

Then we'll discuss the typical architecture of a data warehouse including:

* Kimball’s architecture and the components of this architecture
* The processes of each data warehouse component
* Storage architecture components for optimization

In the last part of the lesson, we'll cover OLAP cubes.

* We’ll discuss how OLAP cubes are a tool for dimensional analysis
* The most common operations you can perform on them for data analytics

By the end of the lesson, you’ll be able to:

* Describe the business need for data warehouses
* Describe the architecture of data warehouses
* Run ETL processes to support dimensional modeling
* Create an OLAP cube from facts and dimensions

# Introduction to Data Warehouses

## What is a Data Warehouse? A Business Perspective

You are in charge of a retailer’s data infrastructure. Let’s look at some business activities.

* Customers should be able to find goods & make orders
* Inventory Staff should be able to stock, retrieve, and re-order goods
* Delivery Staff should be able to pick up & deliver goods
* HR should be able to assess the performance of sales staff
* Marketing should be able to see the effect of different sales channels
* Management should be able to monitor sales growth

Can one database support these activities? Are all of the above questions of the same nature? Let's take a closer look at details that may affect your data infrastructure.

* Scale
* Complexity
* Tabular data sources
* Performance
* Clear requirements

# Data Warehouse for Business Analytics

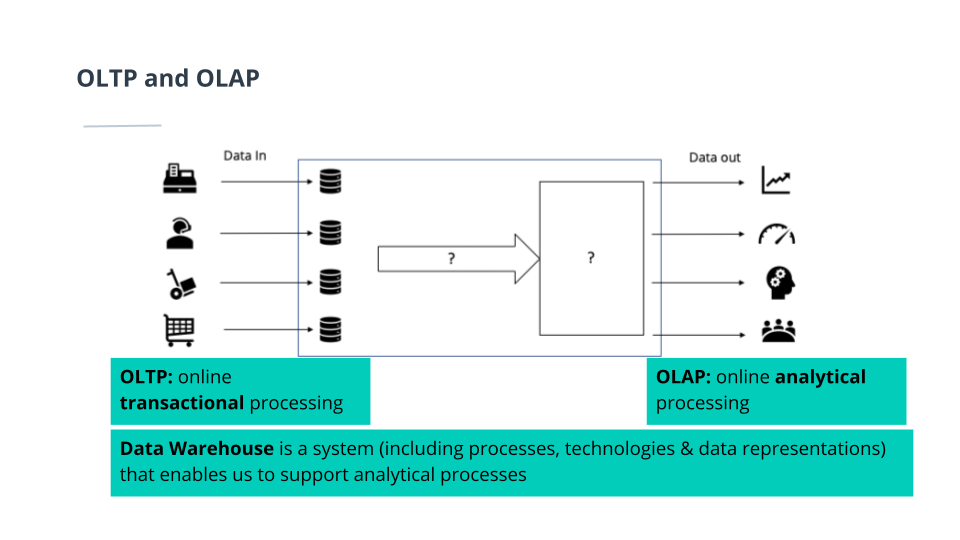
### Operational vs Analytical Business Processes

**Operational processes:** Make it work.

* Find goods & make orders (for customers)
* Stock and find goods (for inventory staff)
* Pick up & deliver goods (for delivery staff)

**Analytical processes:** What is going on?

* Assess the performance of sales staff (for HR)
* See the effect of different sales channels (for marketing)
* Monitor sales growth (for management)



Data Warehouse is a system that enables us to support analytical processes

A close-up of a diagram

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a quiz

Description automatically generated

# Data Warehouse Architecture

## Data Warehouse Architecture

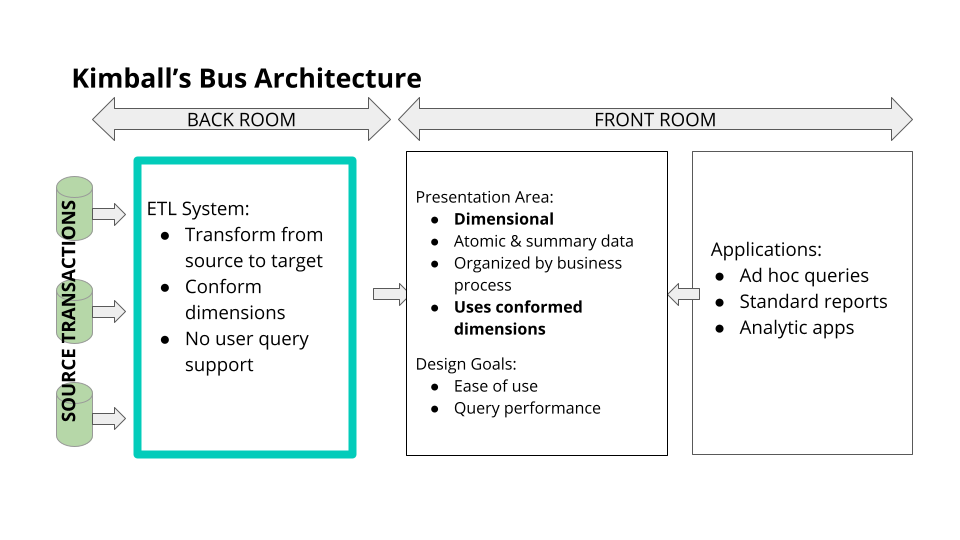
A data warehouse is a copy of transaction data specifically structured for query and analysis. - Kimball

A data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions. - Inmon

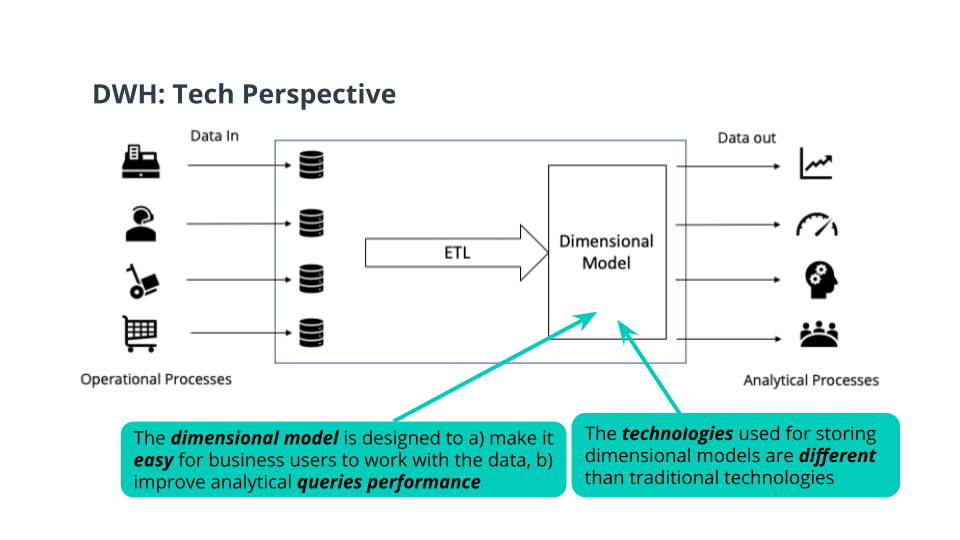
A data warehouse is a system that retrieves and consolidates data periodically from the source systems into a dimensional or normalized data store. It usually keeps years of history and is queried for business intelligence or other analytical activities. It is typically updated in batches, not every time a transaction happens in the source system. - Rainard

### Data Warehouse: Technical Perspective

Extract the data from the source systems used for operations, transform the data, and load it into a dimensional model



Kimball's Bus architecture



The Dimensional Model of a Data Warehouse

Business-user-facing application are needed, with clear visuals - Business Intelligence (BI) apps

### ETL: A Closer Look

Extracting:

* Transfer data to the warehouse

Transforming:

* Integrates many sources together.
* Possibly cleansing inconsistencies, duplication, missing values, etc...
* Possibly producing diagnostic metadata

Loading:

* Structuring and loading the data into the dimensional data model

# ETL and Dimensional Modeling

## ETL and Dimensional Modeling for Data Warehouse

### Dimensional Model Review

**Goals of the Star Schema**

* Easy to understand
* Fast analytical query performance

**Fact Tables**

* Record business events, like an order, a phone call, a book review
* Fact tables columns record events recorded in quantifiable metrics like quantity of an item, duration of a call, a book rating

**Dimension Tables**

* Record the context of the business events, e.g. who, what, where, why, etc..
* Dimension tables columns contain attributes like the store at which an item is purchased or the customer who made the call, etc.

A close-up of a white paper

Description automatically generated

### ETL: A Closer Look

**Extracting:**

* Transfer data to the warehouse

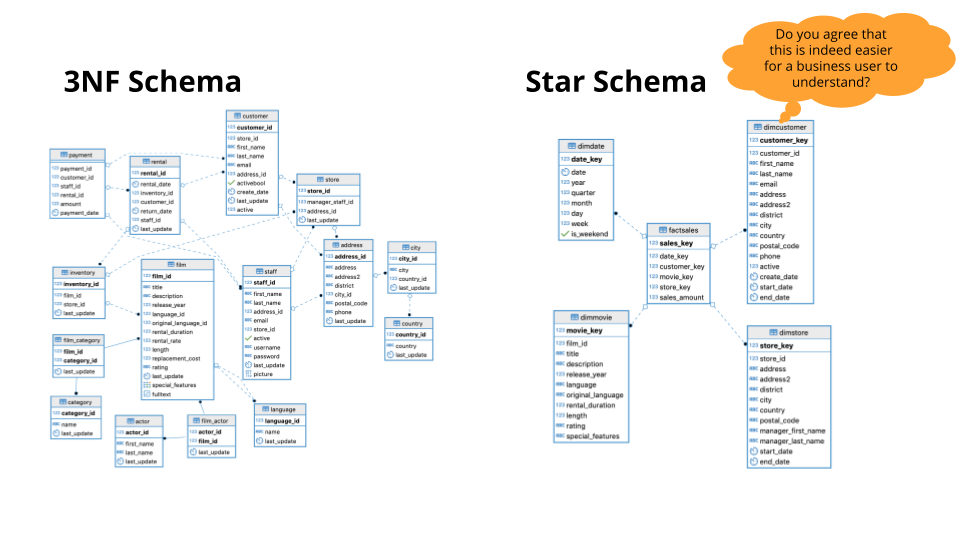
**Transforming:**

* Integrates many sources together
* Possibly cleansing: inconsistencies, duplication, missing values, etc..
* Possibly producing diagnostic metadata

**Loading:**

* Structuring and loading the data into the dimensional data model

### Example: The DVD Rentals Sample Database



Third Normal Form Schema VS Star Schema

### Naive Extract Transform and Load (ETL): From Third Normal Form to ETL

**Extract**

* Query the 3NF DB

**Transform**

* Join tables together
* Change types
* Add new columns

**Load**

* Insert into facts & dimension tables

# Exercise: Exploratory Data Analysis

## Exercise: Exploring a Dataset

In this exercise, you'll learn how to connect to a PostgreSQL database running in the workspace, and explore it to determine the design, scope, size, and type of data, as well as do some inital data analysis.

For this demo and the following exercises, we'll be using a public domain database design and dataset called [Sakila](https://dev.mysql.com/doc/sakila/en/). We'll also be using iPython to write SQL queries in Python.

How to use iPython to write SQL in Python:

* load ipython-sql: %load\_ext sql
* To execute SQL queries, prepend your SQL statements with either of these options:
* %sql

For a single line SQL query:

* Use $ to access a python variable

%%sql

* For a multi-line SQL query
* You will **NOT** be able to access a python variable using $

On the workspace below, you can practice exploring the data yourself:

1. Create and Connect to the Sakila/Pagila database
2. Run SQL queries to establish the size of the data
3. Run SQL queries to find out when events in the database occur
4. Create a SQL query to display the number of addresses by district in the address table. Limit the table to the top 10 districts.