1: Introduction to the course

Link: <https://www.youtube.com/watch?v=c7G8sMId7KM>

Notes:

**Databases:** A [**database**](https://en.wikipedia.org/wiki/Database) is a structured repository or collection of data that is stored and retrieved electronically for use in applications. Data can be stored, updated, or deleted from a database.

**Database Management System (DBMS)**: The software used to access the database by the user and application is the database management system. Check out these few links describing a DBMS in more detail.

1. [**Introduction to DBMS**](https://www.geeksforgeeks.org/database-management-system-introduction-set-1/)
2. [**DBMS as per Wikipedia**](https://en.wikipedia.org/wiki/Database#Database_management_system)

2: What is data modeling?

Link: <https://www.youtube.com/watch?v=CAf5CzKJHmc>

Notes:

**Common Questions**

**Why can't everything be stored in a giant Excel spreadsheet?**

* There are limitations to the amount of data that can be stored in an Excel sheet. So, a database helps organize the elements into tables - rows and columns, etc. Also reading and writing operations on a large scale is not possible with an Excel sheet, so it's better to use a database to handle most business functions.

**Does data modeling happen before you create a database, or is it an iterative process?**

* It's definitely an iterative process. Data engineers continually reorganize, restructure, and optimize data models to fit the needs of the organization.

**How is data modeling different from machine learning modeling?**

* Machine learning includes a lot of data wrangling to create the inputs for machine learning models, but data modeling is more about how to structure data to be used by different people within an organization. You can think of data modeling as the process of designing data and making it available to machine learning engineers, data scientists, business analytics, etc., so they can make use of it easily.

**QUIZ QUESTION**

Choose the Correct Order of the Data Modeling Process:

* 

Physical -> Logical -> Conceptual

* 

Logical -> Physical -> Conceptual

* 

Conceptual -> Physical -> Logical

* Conceptual -> Logical -> Physical

SUBMIT

NEXT

3: Why is data modeling important?

Link: <https://www.youtube.com/watch?v=hq0Q86fEd2E>

Notes:

**Key points about Data Modeling**

* **Data Organization:** The organization of the data for your applications is extremely important and makes everyone's life easier.
* **Use cases:** Having a well thought out and organized data model is critical to how that data can later be used. Queries that could have been straightforward and simple might become complicated queries if data modeling isn't well thought out.
* **Starting early:** Thinking and planning ahead will help you be successful. This is not something you want to leave until the last minute.
* **Iterative Process:** Data modeling is not a fixed process. It is iterative as new requirements and data are introduced. Having flexibility will help as new information becomes available.

**Example of Why Data Modeling is Important:**

Let's take an example from Udacity. Here, a Udacity data engineer would help structure the data so it can be used by different people within Udacity for further analysis and also shared with the learner on the website. For instance, when we want to track the students' progress within a Nanodegree program, we want to aggregate data across students and projects within a Nanodegree. In a relational database, this requires the data to be structured in ways that each student's data is tracked across all Nanodegree programs that s/he has ever enrolled in. The data also needs to track the student's progress within each of those Nanodegree programs.

The data model is critical for accurately representing each data object. For instance, a data table would track a student's progress on project submissions, i.e., whether they passed or failed a specific rubric requirement. Furthermore, the data model should ensure that a student's progress is updated and aggregated to provide an indicator of whether the student passed all the rubric requirements and successfully finished the project. Data modeling is critical to track all of these pieces of data so the tables are speaking to each other, updating the tables correctly (e.g., updating a student's progress on a project submission), and meeting defined rules (e.g., project completed when all rubric requirements are passed).

**QUIZ QUESTION**

Who should focus on learning data modeling? Choose one response.

* 

Data Scientists

* 

Data Engineers

* 

Software Engineers

* Everyone who deals with data!

SUBMIT

NEXT

4: Who does this type of work?

Link: <https://www.youtube.com/watch?v=LBqMfr1PniQ>

Notes:

5: Introduction to relational databases

Link: <https://www.youtube.com/watch?v=TAXJtSgGzqA>

Notes:

**QUIZ QUESTION**

What is an RDBMS?

* 

Relateable Database Management System

* Relational Database Management System
* 

Readable Data Management System

* 

Reachable Database Management System

SUBMIT

NEXT

6: Quiz: relational databases

Link:

Notes:

**QUESTION 1 OF 3**

True or False: A column holds multiple tables.

* 

True

* False

SUBMIT

**QUESTION 2 OF 3**

True or False: An attribute is another name for a column.

* True
* 

False

SUBMIT

**QUESTION 3 OF 3**

True or False: A schema is a collection of tables in some database terminology.

* True
* 

False

SUBMIT

NEXT

7: When to use a relational database?

Link: <https://www.youtube.com/watch?v=cHErsgQZNlw>

Notes:

**Advantages of Using a Relational Database**

* **Flexibility for writing in SQL queries:** With SQL being the most common database query language.
* **Modeling the data not modeling queries**
* **Ability to do JOINS**
* **Ability to do aggregations and analytics**
* **Secondary Indexes available** : You have the advantage of being able to add another index to help with quick searching.
* **Smaller data volumes:** If you have a smaller data volume (and not big data) you can use a relational database for its simplicity.
* **ACID Transactions**: Allows you to meet a set of properties of database transactions intended to guarantee validity even in the event of errors, power failures, and thus maintain data integrity.
* **Easier to change to business requirements**

8: ACID transactions

Link: <https://www.youtube.com/watch?v=rr35DQljFnc>

Notes:

### ACID Transactions

Properties of database transactions intended to guarantee validity even in the event of errors or power failures.

* **Atomicity:** The whole transaction is processed or nothing is processed. A commonly cited example of an atomic transaction is money transactions between two bank accounts. The transaction of transferring money from one account to the other is made up of two operations. First, you have to withdraw money in one account, and second you have to save the withdrawn money to the second account. An atomic transaction, i.e., when either all operations occur or nothing occurs, keeps the database in a consistent state. This ensures that if either of those two operations (withdrawing money from the 1st account or saving the money to the 2nd account) fail, the money is neither lost nor created. Source [**Wikipedia**](https://en.wikipedia.org/wiki/Atomicity_%28database_systems%29) for a detailed description of this example.
* **Consistency:** Only transactions that abide by constraints and rules are written into the database, otherwise the database keeps the previous state. The data should be correct across all rows and tables. Check out additional information about consistency on [**Wikipedia**](https://en.wikipedia.org/wiki/Consistency_%28database_systems%29).
* **Isolation:** Transactions are processed independently and securely, order does not matter. A low level of isolation enables many users to access the data simultaneously, however this also increases the possibilities of concurrency effects (e.g., dirty reads or lost updates). On the other hand, a high level of isolation reduces these chances of concurrency effects, but also uses more system resources and transactions blocking each other. Source: [**Wikipedia**](https://en.wikipedia.org/wiki/Isolation_%28database_systems%29)
* **Durability:** Completed transactions are saved to database even in cases of system failure. A commonly cited example includes tracking flight seat bookings. So once the flight booking records a confirmed seat booking, the seat remains booked even if a system failure occurs. Source: [**Wikipedia**](https://en.wikipedia.org/wiki/ACID).

We cover ACID transactions in much depth in Lesson 2 of this course.

### QUESTION 1 OF 2

Which of these are benefits of a relational database?

* ACID Transactions
* Ability to do JOINS
* 

Can handle big data

* Easy to change business requirements on the data

SUBMIT

### QUESTION 2 OF 2

Can you JOIN a table with another table on any column?

* Yes, as long there are matching values in the columns
* 

No, only on columns with the same name

* 

Yes, you can join any columns together

* 

No, only on columns in the same table

SUBMIT

NEXT

9: When not touse a relational database?

Link: <https://www.youtube.com/watch?v=yzB2G33pxbQ>

Notes:

**When Not to Use a Relational Database**

* **Have large amounts of data:** Relational Databases are not distributed databases and because of this they can only scale vertically by adding more storage in the machine itself. You are limited by how much you can scale and how much data you can store on one machine. You cannot add more machines like you can in NoSQL databases.
* **Need to be able to store different data type formats:** Relational databases are not designed to handle unstructured data.
* **Need high throughput -- fast reads:** While ACID transactions bring benefits, they also slow down the process of reading and writing data. If you need very fast reads and writes, using a relational database may not suit your needs.
* **Need a flexible schema:** Flexible schema can allow for columns to be added that do not have to be used by every row, saving disk space.
* **Need high availability:** The fact that relational databases are not distributed (and even when they are, they have a coordinator/worker architecture), they have a single point of failure. When that database goes down, a fail-over to a backup system occurs and takes time.
* **Need horizontal scalability:** Horizontal scalability is the ability to add more machines or nodes to a system to increase performance and space for data.

**QUESTION 1 OF 2**

True or False: Relational Databases are traditionally horizontally scalable.

* 

True

* False

SUBMIT

**QUESTION 2 OF 2**

When should you use a Relational Database?

* Small amounts of data
* Need to be able to do aggregations
* Need ACID transactions
* 

You need to be able to scale out quickly

* Need to be able to join multiple tables

SUBMIT

NEXT

10: What is PostgreSQL?

Link: <https://www.youtube.com/watch?v=5L9O9QK7nLY>

Notes:

11: Demos: Creating a Postgres Table

Link: <https://www.youtube.com/watch?v=QuC6xIj7WiI>

Link2: <https://www.youtube.com/watch?v=OzNizL-Zjt8>

Notes:

### Please note: We have provided a Workspace so you do not need to run Postgres locally on your machine for any of the exercises or projects in this course.

The following information on setting up PostgreSQL on your local machine is completely optional for this course and for users who feel comfortable setting up the environment on their local machine to complete the exercises.

Here is additional Information on how to install and set up Postgres locally in case you want to follow along the demo on your local machine. This [**link**](https://www.codementor.io/engineerapart/getting-started-with-postgresql-on-mac-osx-are8jcopb) provides directions for MacOS. It goes through configuring Postgres, creating users, and creating databases using the psql utility. It will help further explain the Python driver and also help you in running the demos locally.

In addition, here is a short tutorial on psycopg2. This [**link**](https://pynative.com/python-postgresql-tutorial/) gives a good starter tutorial in case you are curious about it.

Here are the two demo files. Feel free to follow along. Just download these and open up the Jupyter Notebook files.

#### **Supporting Materials**

[**L1 Demo\_0\_Creating\_a\_table\_with\_Postgres**](https://video.udacity-data.com/topher/2019/March/5c9e7914_l1-demo-0-creating-a-table-with-postgres/l1-demo-0-creating-a-table-with-postgres.ipynb)

[**L1\_Demo\_1\_Creating\_a\_table\_with\_Postgres**](https://video.udacity-data.com/topher/2019/March/5c9e8bc6_l1-demo-1-creating-a-table-with-postgres/l1-demo-1-creating-a-table-with-postgres.ipynb)

**12: Exercise 1: creating a table with postgres**

**Link:**

**Notes:**

**13: Solution for exercise 1: create a table with postgres**

**Link:** [**https://www.youtube.com/watch?v=2tPN\_pMssiU**](https://www.youtube.com/watch?v=2tPN_pMssiU)

**Notes:**

Here is the solution file. Feel free to follow along. Just download it and open up the Jupyter Notebook file.

#### **Supporting Materials**

[**L1\_Exercise\_1\_Solution\_Creating\_table\_with\_postgres**](https://video.udacity-data.com/topher/2019/March/5c9e8d45_l1-exercise-1-solution-creating-a-table-with-postgres/l1-exercise-1-solution-creating-a-table-with-postgres.ipynb)

**14: NoSQL database**

**Link:** [**https://www.youtube.com/watch?v=0rVgi7NiHFQ**](https://www.youtube.com/watch?v=0rVgi7NiHFQ)

**Notes:**

**QUESTION 1 OF 2**

True or False: A Keyspace in Apache Cassandra is similar to a schema in PostgreSQL

* True
* 

False

SUBMIT

**QUESTION 2 OF 2**

Which of these are examples of non-relational databases?

* 

SQL

* Apache Cassandra
* 

RDBMS

* MongoDB

SUBMIT

NEXT

15: What is Apache Cassandra?

Link: <https://www.youtube.com/watch?v=ku8Ywgkduk8>

Notes:

#### Common Questions:

**What type of companies use Apache Cassandra?**  
All kinds of companies. For example, Uber uses Apache Cassandra for their entire backend. Netflix uses Apache Cassandra to serve all their videos to customers. Good use cases for NoSQL (and more specifically Apache Cassandra) are :

1. Transaction logging (retail, health care)
2. Internet of Things (IoT)
3. Time series data
4. Any workload that is heavy on writes to the database (since Apache Cassandra is optimized for writes).

**Would Apache Cassandra be a hindrance for my analytics work? If yes, why?**  
Yes, if you are trying to do analysis, such as using GROUP BY statements. Since Apache Cassandra requires data modeling based on the query you want, you can't do ad-hoc queries. However you can add clustering columns into your data model and create new tables.

### QUIZ QUESTION

When should you use a NoSQL database?

* Large amounts of data
* 

Need to be able to do aggregations

* Need high availability
* You need to be able to scale out quickly
* 

Need to be able to join multiple tables

SUBMIT

NEXT

16: When to use a NoSql database?

Link: <https://www.youtube.com/watch?v=jjwYgUEPSUc>

Notes:

**When to use a NoSQL Database**

* **Need to be able to store different data type formats**: NoSQL was also created to handle different data configurations: structured, semi-structured, and unstructured data. JSON, XML documents can all be handled easily with NoSQL.
* **Large amounts of data**: Relational Databases are not distributed databases and because of this they can only scale vertically by adding more storage in the machine itself. NoSQL databases were created to be able to be horizontally scalable. The more servers/systems you add to the database the more data that can be hosted with high availability and low latency (fast reads and writes).
* **Need horizontal scalability**: Horizontal scalability is the ability to add more machines or nodes to a system to increase performance and space for data
* **Need high throughput**: While ACID transactions bring benefits they also slow down the process of reading and writing data. If you need very fast reads and writes using a relational database may not suit your needs.
* **Need a flexible schema**: Flexible schema can allow for columns to be added that do not have to be used by every row, saving disk space.
* **Need high availability**: Relational databases have a single point of failure. When that database goes down, a failover to a backup system must happen and takes time.

17: When not to use a NoSql Database?

Link: <https://www.youtube.com/watch?v=fT_USRXmtlo>

Notes:

### When NOT to use a NoSQL Database?

* **When you have a small dataset**: NoSQL databases were made for big datasets not small datasets and while it works it wasn’t created for that.
* **When you need ACID Transactions**: If you need a consistent database with ACID transactions, then most NoSQL databases will not be able to serve this need. NoSQL database are eventually consistent and do not provide ACID transactions. However, there are exceptions to it. Some non-relational databases like MongoDB can support ACID transactions.
* **When you need the ability to do JOINS across tables**: NoSQL does not allow the ability to do JOINS. This is not allowed as this will result in full table scans.
* **If you want to be able to do aggregations and analytics**
* **If you have changing business requirements**: Ad-hoc queries are possible but difficult as the data model was done to fix particular queries
* **If your queries are not available and you need the flexibility**: You need your queries in advance. If those are not available or you will need to be able to have flexibility on how you query your data you might need to stick with a relational database

### Caveats to NoSQL and ACID Transactions

There are some NoSQL databases that offer some form of ACID transaction. As of v4.0, MongoDB added multi-document ACID transactions within a single replica set. With their later version, v4.2, they have added multi-document ACID transactions in a sharded/partitioned deployment.

* Check out this documentation from MongoDB on [**multi-document ACID transactions**](https://www.mongodb.com/collateral/mongodb-multi-document-acid-transactions)
* Here is another link documenting [**MongoDB's ability to handle ACID transactions**](https://www.mongodb.com/blog/post/mongodb-multi-document-acid-transactions-general-availability)

Another example of a NoSQL database supporting ACID transactions is MarkLogic.

* Check out this link from their [**blog**](https://www.marklogic.com/blog/how-marklogic-supports-acid-transactions/) that offers ACID transactions.

18: Demo 2: Creating table with Cassandra

Link1: <https://www.youtube.com/watch?v=u-GlcpAndmY>

Link2: <https://www.youtube.com/watch?v=QFYO0SqZMuw>

Notes:

### Please note: We have provided a Workspace so you do not need to run Apache Cassandra locally on your machine for any of the Exercises or Projects in this Course.

The following information on setting up Cassandra on your local machine is completely optional for this course and for users who feel comfortable setting up the environment on their local machine to complete the exercises. Please note, Apache Cassandra is easier to install on MacOS than a Windows machine.

**Installing Apache Cassandra to run locally on your machine:**  
[**Cassandra Documentation**](http://cassandra.apache.org/doc/latest/getting_started/installing.html)

Again, if you want to follow along, here is the demo notebook showcased in the video above.

#### **Supporting Materials**

[**L1\_Demo\_2\_Creating\_table\_with\_Cassandra**](https://video.udacity-data.com/topher/2019/March/5c9e8ea2_l1-demo-2-creating-a-table-with-apache-cassandra/l1-demo-2-creating-a-table-with-apache-cassandra.ipynb)

19: Exercise 2: create table with Cassandra

Link:

Notes:

20: Solution for exercise 2: create table with Cassandra

Link: <https://www.youtube.com/watch?v=9m4_k_luAqI>

Notes:

Here is the link to the solution notebook.

#### **Supporting Materials**

[**L1\_Exercise\_2\_Solution**](https://video.udacity-data.com/topher/2019/March/5c9e961b_l1-exercise-2-solution-creating-a-table-with-cassandra/l1-exercise-2-solution-creating-a-table-with-cassandra.ipynb)

21: Conclusion

Link: <https://www.youtube.com/watch?v=_II7WYiFslk>

Notes: