In this course, you'll see—and use—several datasets. This reading describes these datasets, including their sources and other things you should know. You might want to return to this reading when you have questions about a dataset—though if you just need some understanding of what’s in the dataset, you can use that as an opportunity to practice some of the queries and commands you’ve already learned!

Most of these datasets are preloaded on the VM as tables that can be queried by *all* the available SQL engines (Hive, Impala, MySQL, and PostgreSQL). However, please note the following differences:

* The tables are organized into different databases (named **default**, **fun**, **fly**, **toy**, and **wax**) *only* for Hive and Impala; for MySQL and PostgreSQL, all the tables exist in one database (which is named **mydb** for MySQL and **public** for PostgreSQL).
* The tables in the **fly** database have *not* been loaded into MySQL and PostgreSQL; those are available to query *only* with Hive and Impala.

The data type of each column is specified, but the details of these different data types are beyond the scope of this course. For purposes of this course, you need only know that:

* **TINYINT**, **SMALLINT**, **INT**, and **BIGINT**are all integer numeric data types
* **DECIMAL**, **FLOAT**, and **DOUBLE**are all decimal numeric data types
* **STRING**, **VARCHAR**, and **CHAR**are all character string data types

Please note that these datasets are provided solely for learning purposes. We make no claims or guarantees about the data or its contents, accuracy, completeness, or representativeness. Any trademarks or copyrighted names included in the data are used without permission under the legal doctrine of fair use.

The **default**Database

This database has five tables related to a fictitious, international company. Each table is a tiny table, with a very small number of rows, for demonstration purposes.

Table: **customers**

Description: Fictitious international customers for the company

Number of rows: 4

Columns:

| **Name** | **Data Type** | **Description** | **Sample Value** |
| --- | --- | --- | --- |
| cust\_id | **STRING** | A unique identifier for each customer | **a** |
| name | **STRING** | The customer's name | **Arfa** |
| country | **STRING** | A two-letter code to represent the customer's country | **pk** |

 D​ata:

| **cust\_id** | **name** | **country** |
| --- | --- | --- |
| a​ | Arfa | pk |
| b​ | Brendon | us |
| c​ | Chiyo | ja |
| d​ | Dikembe | ug |

Table: **employees**

Description: Fictitious international employees for the company

Number of rows: 5

Columns:

| **Name** | **Data Type** | **Description** | **Sample Value** |
| --- | --- | --- | --- |
| empl\_id | **INT** | A unique identifier for each employee | **1** |
| first\_name | **STRING** | The employee's first (given) name | **Ambrosio** |
| last\_name | **STRING** | The employee's last name (surname) | **Rojas** |
| salary | **INT** | The employee's annual salary in US dollars | **25784** |
| office\_id | **STRING** | The ID of the office (from the **offices**table) where the employee works | **c** |

D​ata:

| **e​mol\_id** | **first\_name** | **last\_name** | **s​alary** | **o​ffice\_id** |
| --- | --- | --- | --- | --- |
| 1​ | Ambrosio | Rojas | 25784 | c​ |
| 2​ | Val | Snyder | 37506 | e​ |
| 3​ | Virginia | Levitt | 54523 | b​ |
| 4​ | Sabahattin | Tilki | 28060 | a​ |
| 5​ | Lujza | Csizmadia | 39530 | b​ |

Table: **offices**

Description: Locations around the world of the fictitious company's offices

Number of rows: 4

Columns:

| **Name** | **Data Type** | **Description** | **Sample Value** |
| --- | --- | --- | --- |
| office\_id | **STRING** | A unique identifier for each office | **b** |
| city | **STRING** | The city where the office is located | **Chicago** |
| state\_province | **STRING** | The state or province (as appropriate) where the office is located | **Illinois** |
| country | **STRING** | A two-letter code to represent the country where the office is located | **us** |

D​ata:

| **office\_id** | **city** | **state\_province** | **country** |
| --- | --- | --- | --- |
| a​ | Istanbul | Istanbul | tr |
| b​ | Chicago | Illinois | us |
| c​ | Rosario | Santa Fe | ar |
| d​ | Singapore | NULL | sg |

Table: **orders**

Description: Fictitious order information made by the customers in the **customers**table

Number of rows: 5

Columns:

| **Name** | **Data Type** | **Description** | **Sample Value** |
| --- | --- | --- | --- |
| order\_id | **INT** | A unique identifier for each order | **2** |
| cust\_id | **STRING** | The ID for the customer (from the **customer**table) who placed the order | **a** |
| empl\_id | **INT** | The ID for the employee (from the **employees**table) who  took the order | **4** |
| total | **DECIMAL(5,2)** | The order amount in US dollars; negative values are refunds | **28.54** |

D​ata:

| **order\_id** | **cust\_id** | **empl\_id** | **total** |
| --- | --- | --- | --- |
| 1​ | c​ | 1​ | 2​4.78 |
| 2​ | a​ | 4​ | 2​8.54 |
| 3​ | b​ | 3​ | 4​8.69 |
| 4​ | b​ | 3​ | -​16.39 |
| 5​ | z​ | 2​ | 2​9.92 |

Table: **salary\_grades**

Description: Ranges to categorize employee salary information

Number of rows: 5

Columns:

| **Name** | **Data Type** | **Description** | **Sample Value** |
| --- | --- | --- | --- |
| grade | **TINYINT** | The salary level (1 is lowest) | **1** |
| min\_salary | **INT** | The minimum salary (in US dollars) for the grade level | **10000** |
| max\_salary | **INT** | The maximum salary (in US dollars) for the grade level | **19999** |

 Data:

| **grade** | **min\_salary** | **max\_salary** |
| --- | --- | --- |
| 1​ | 10000 | 19999 |
| 2​ | 20000 | 29999 |
| 3​ | 30000 | 39999 |
| 4​ | 40000 | 49999 |
| 5​ | 50000 | 59999 |

The **fun**Database

This database has four tables related to some popular board games that might or might not be sold in one of two game shops, or related to a standard deck of playing cards (sometimes called “poker cards”). Each table is a tiny table, with a very small number of rows, for demonstration purposes.

Table: **card\_rank**

Description: The ranks of a standard deck of playing cards

Number of rows: 13

Columns:

| **Name** | **Data Type** | **Description** | **Sample Value** |
| --- | --- | --- | --- |
| rank | **STRING** | The rank of the card | **Queen** |
| value | **TINYINT** | The card's usual value | **10** |

 Data:

| **rank** | **value** |
| --- | --- |
| Ace | NULL |
| 2​ | 2​ |
| 3​ | 3​ |
| 4​ | 4​ |
| 5​ | 5​ |
| 6​ | 6​ |
| 7​ | 7​ |
| 8​ | 8​ |
| 9​ | 9​ |
| 10 | 10 |
| Jack | 10 |
| Queen | 10 |
| King | 10 |

Table: **card\_suit**

Description: The suits of a standard deck of playing cards

Number of rows: 4

Columns:

| **Name** | **Data Type** | **Description** | **Sample Value** |
| --- | --- | --- | --- |
| suit | **STRING** | The name of the suit | **Spades** |
| color | **STRING** | The color of the suit | **Black** |

Data:

| **suit** | **color** |
| --- | --- |
| Clubs | Black |
| Diamonds | Red |
| Hearts | Red |
| Spades | Black |

Table: **games**

Description: Information about different board games

Number of rows: 5

Columns:

| **Name** | **Data Type** | **Description** | **Sample Value** |
| --- | --- | --- | --- |
| id | **INT** | A unique identifier for each game | **1** |
| name | **STRING** | The name of the game | **Monopoly** |
| inventor | **STRING** | The person who invented the game | **Elizabeth Magie** |
| year | **STRING** | The year the game was first published | **1903** |
| min\_age | **TINYINT** | The recommended minimum age for players | **8** |
| min\_players | **TINYINT** | The recommended minimum number of players | **2** |
| max\_players | **TINYINT** | The recommended maximum number of players | **6** |
| list\_price | **DECIMAL(5,2)** | The recommended price in US dollars for retail sales | **19.99** |

Data:

| **id** | **name** | **inventor** | **year** | **min\_age** | **min\_players** | **max\_players** | **list\_price** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1​ | Monopoly | Elizabeth Magie | 1903 | 8​ | 2​ | 6​ | 19.99 |
| 2​ | Scrabble | Alfred Moster Butts | 1938 | 8​ | 2​ | 4​ | 17.99 |
| 3​ | Clue | Anthony E. Pratt | 1944 | 8​ | 2​ | 6​ | 9.99 |
| 4​ | Candy Land | Eleanor Abbott | 1948 | 3​ | 2​ | 4​ | 7.99 |
| 5​ | Risk | Albert Lamorisse | 1957 | 10 | 2​ | 5​ | 29.99 |

Notes:

* We assembled this data from various publicly available sources.
* The **list\_price**is the manufacturer's suggested retail price (MSRP), not necessarily the actual price at which a game will be sold.
* Elizabeth Magie is listed as the inventor of *Monopoly*, based on her invention of *The Landlord's Game*in 1903–1904, which was the basis of *Monopoly*. (See the YouTube video Who Really Invented Monopoly?​ for more information on that.)
* The game *Clue*is called *Cluedo*outside of North America.

Table: **inventory**

Description: Inventory of board games at two fictitious game shops

Number of rows: 5

Columns:

| **Name** | **Data Type** | **Description** | **Sample Value** |
| --- | --- | --- | --- |
| shop | **STRING** | The name of the shop carrying a particular game | **Dicey** |
| game | **STRING** | The name of the game | **Monopoly** |
| qty | **INT** | How many copies of the game the shop has in stock or in inventory—that is, how many copies of the game are in the shop ready to be sold | **7** |
| aisle | **TINYINT** | The location in the shop where the game can be found | **3** |
| price | **DECIMAL(5,2)** | The sale price of the game in US dollars | **17.99** |

Data:

| **shop** | **game** | **qty** | **aisle** | **price** |
| --- | --- | --- | --- | --- |
| Dicey | Monopoly | 7​ | 3​ | 17.99 |
| Dicey | Clue | 3​ | NULL | 9.99 |
| Board 'Em | Monopoly | 11 | 2​ | 25.00 |
| Board 'Em | Candy Land | 4​ | 2​ | NULL |
| Board 'Em | Risk | 3​ | 1​ | 35.00 |

Notes:

* The **price**in this table is different from **list\_price**in the **games**table—this is the price at which the shop is actually selling the game, which could be greater or less than the MSRP (see the notes for the **games**table).
* The game *Clue*is called *Cluedo*outside of North America.

The **fly**Database

This database has four tables containing real-world data gathered by the United States Department of Transportation. Some of these tables are quite large. We are indebted to Hadley Wickham (at RStudio) for the **nycflights13**R package and to Jeffrey Arnold (at the University of Washington) for the **groundcontrol**R package, both of which helped us to assemble these aviation datasets.

T​hese are larger tables, so we did not provide the full data here.

Table: **airlines**

Description: A mapping of a two-letter code for airline carriers, used by the **flights**table, and the full name of the airline represented by that code

Number of rows: 22

Columns:

| **Name** | **Data Type** | **Description** | **Sample Value** |
| --- | --- | --- | --- |
| carrier | **STRING** | A two-character identifier for the airline carrier | **B6** |
| name | **STRING** | The carrier's full name | **JetBlue Airways** |

Table: **airports**

Description: Information about the airports used in the **flights**table

Number of rows: 1,333

Columns:

| **Name** | **Data Type** | **Description** | **Sample Value** |
| --- | --- | --- | --- |
| faa | **CHAR(3)** | Three-letter FAA (US Federal Aviation Administration) code for the airport | **TYS** |
| name | **STRING** | Full name of the airport | **McGhee Tyson Airport** |
| lat | **DOUBLE** | Latitude of the airport's location | **35.811000819999997** |
| lon | **DOUBLE** | Longitude of the airport's location | **-83.994003300000003** |
| alt | **SMALLINT** | The altitude of the airport | **981** |
| tz | **TINYINT** | The time zone in which the airport is located, represented as an offset from UTC in hours | **-5** |

Table: **flights**

Description: Data on all domestic flights by major US air carriers for the full decade from January 1, 2008 through December 31, 2017

Number of rows: 61,392,822

Columns:

| **Name** | **Data Type** | **Description** | **Sample Value** |
| --- | --- | --- | --- |
| year | **SMALLINT** | The year when the flight departed (formatted as a four-digit number) | **2014** |
| month | **TINYINT** | The month when the flight departed (formatted as a number between 1 and 12) | **9** |
| day | **TINYINT** | The day when the flight departed (formatted as a number between 1 and 31 representing the day of the month) | **16** |
| dep\_time | **SMALLINT** | The actual time when the flight departed its origin airport, in the origin airport’s local time zone, formatted as the one- or two-digit hour (an integer between 0 and 24) followed by the two-digit minute (between 00 and 59) | **548** |
| sched\_dep\_time | **SMALLINT** | The scheduled departure time, in the origin airport’s local time zone, formatted as the one- or two-digit hour (an integer between 0 and 24) followed by the two-digit minute (between 00 and 59) | **600** |
| dep\_delay | **SMALLINT** | The departure delay (difference in minutes between **sched\_dep\_time**and **dep\_time**) | **-12** |
| arr\_time | **SMALLINT** | The actual time when the flight arrived at its destination airport, in the destination airport’s local time zone, formatted as the one- or two-digit hour (an integer between 0 and 24) followed by the two-digit minute (between 00 and 59) | **718** |
| sched\_arr\_time | **SMALLINT** | The scheduled arrival time, in the destination airport’s local time zone, formatted as the one- or two-digit hour (an integer between 0 and 24) followed by the two-digit minute (between 00 and 59) | **728** |
| arr\_delay | **SMALLINT** | The arrival delay (difference in minutes between **sched\_arr\_time**and **arr\_time**) | **-10** |
| carrier | **STRING** | The two-letter code for the airline of the flight | **EV** |
| flight | **SMALLINT** | The flight number for the flight | **4642** |
| tailnum | **STRING** | The tail number of the aircraft used for the flight, a unique identifier for each aircraft | **N26549** |
| origin | **STRING** | The three-letter FAA code for the origin airport from which the flight departed | **TYS** |
| dest | **STRING** | The three-letter FAA code for the destination airport for the flight | **IAD** |
| air\_time | **SMALLINT** | The amount of time (in minutes) that the flight was in the air | **66** |
| distance | **SMALLINT** | The distance (in miles) traveled by the flight | **419** |

 Notes:

* This data does contain errors and omissions. (This is real-world data; there are bound to be some erroneous and missing values!)
* The time columns (such as **dep\_time**and **sched\_arr\_time**) use a 24-hour time clock and are provided using local time to the airport (departures for the origin airport and arrivals for the destination airport), so 1335 at BOS (Boston) is 1:35 p.m. in the Eastern time zone, and 803 at SFO (San Francisco) is 8:03 a.m. in the Pacific time zone. Both **arr\_time**and **dep\_time**range from 1 to 2400, while **sched\_arr\_time**is from 0 to 2400 and **sched\_dep\_time**is from 0 to 2359.

Table: **planes**

Description: Information about various aircraft, which might or might not be included in the **flights**table

Number of rows: 453,361

Columns:

| **Name** | **Data Type** | **Description** | **Sample Value** |
| --- | --- | --- | --- |
| tailnum | **STRING** | The tail number of the aircraft used for the flight, a unique identifier for each aircraft | **N26549** |
| year | **INT** | The year the aircraft was manufactured | **2002** |
| type | **STRING** | The type of aircraft | **Fixed wing multi engine** |
| manufacturer | **STRING** | The name of the manufacturer of the aircraft | **EMBRAER** |
| model | **STRING** | The manufacturer's model designation of the aircraft | **EMB-145LR** |
| engines | **INT** | The number of engines that the aircraft has | **2** |
| seats | **INT** | The number of seats on the aircraft | **55** |
| engine | **STRING** | The type of engine used by the aircraft | **Turbo-fan** |

The **toy**Database

This database has two tables containing data about a few children’s toys and toy makers. Each table is a tiny table for demonstration purposes.

Table: **makers**

Description: Information about companies that make certain toys

Number of rows: 3

Columns:

| **Name** | **Data Type** | **Description** | **Sample Value** |
| --- | --- | --- | --- |
| id | **INT** | A unique identifier for each company | **105** |
| name | **STRING** | The company's name | **Hasbro** |
| city | **STRING** | The city where the company's headquarters is located | **Pawtucket, RI** |

Data:

| **id** | **name** | **city** |
| --- | --- | --- |
| 105 | Hasbro | P​awtucket, RI |
| 106 | Ohio Art Company | Bryan, OH |
| 107 | Mattel | Segundo, CA |

Table: **toys**

Description: Information about toys

Number of rows: 3

Columns:

| **Name** | **Data Type** | **Description** | **Sample Value** |
| --- | --- | --- | --- |
| id | **INT** | A unique identifier for each toy | **21** |
| name | **STRING** | The name of the toy | **Lite-Brite** |
| price | **DECIMAL(5,2)** | Retail price for the toy in US dollars | **14.47** |
| maker\_id | **INT** | The ID of the company that makes the toy (used in the **makers** table) | **105** |

D​ata:

| **id** | **name** | **price** | **maker\_id** |
| --- | --- | --- | --- |
| 21 | Lite-Brite | 14.47 | 105 |
| 22 | Mr. Potato Head | 11.50 | 105 |
| 23 | Etch A Sketch | 29.99 | 106 |

Notes:

* We assembled this data from various publicly available sources.
* The **price**column is an actual retail price for the toy, but it might not be the manufacturer's suggested retail price.

The **wax**Database

This database has one table, which gives information about crayon colors. T​his is a large table, so we did not provide the full data here.

Table: **crayons**

Description: Information about colors available for Crayola crayons

Number of rows: 120

Columns:

| **Name** | **Data Type** | **Description** | **Sample Value** |
| --- | --- | --- | --- |
| color | **VARCHAR(25)** | The name of the color | **Chestnut** |
| hex | **CHAR(6)** | A hex code that approximates the color | **BC5D58** |
| red | **SMALLINT** | The red component of the RGB code that approximates the color | **188** |
| green | **SMALLINT** | The green component of the RGB code that approximates the color | **92** |
| blue | **SMALLINT** | The blue component of the RGB code that approximates the color | **88** |
| pack | **TINYINT** | The number of crayons in the *smallest*pack that includes that color | **32** |