




# The Sequels

Data Analysis and Infrastructure  
DA24STO - Hyper Island



# From sheets to tables

Migrate a spreadsheet data set to a database

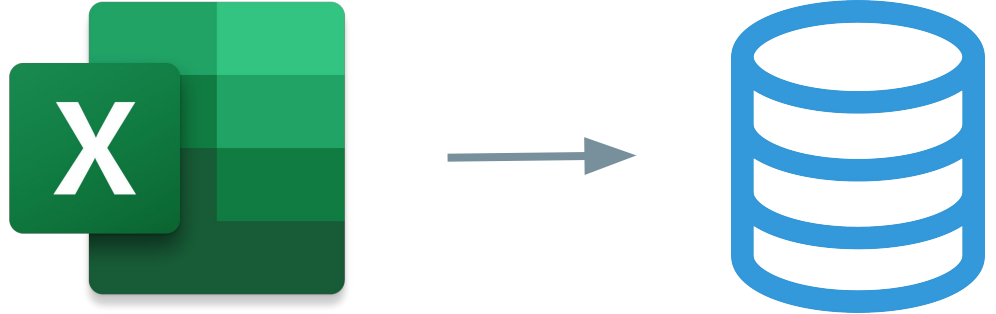
Review the data

Design the tables

Clean the data

Connect the data

Populate the database



# The data

Three given spreadsheets:

- **Classlist**
- **StudentPerformance**
- **MovieChoice**



Classlist Hyper  
Island DA24.xlsx



StudentPerforma  
ncelnln...ment.xlsx



MovieChoiceOfDA  
24Student.xlsx

One created:

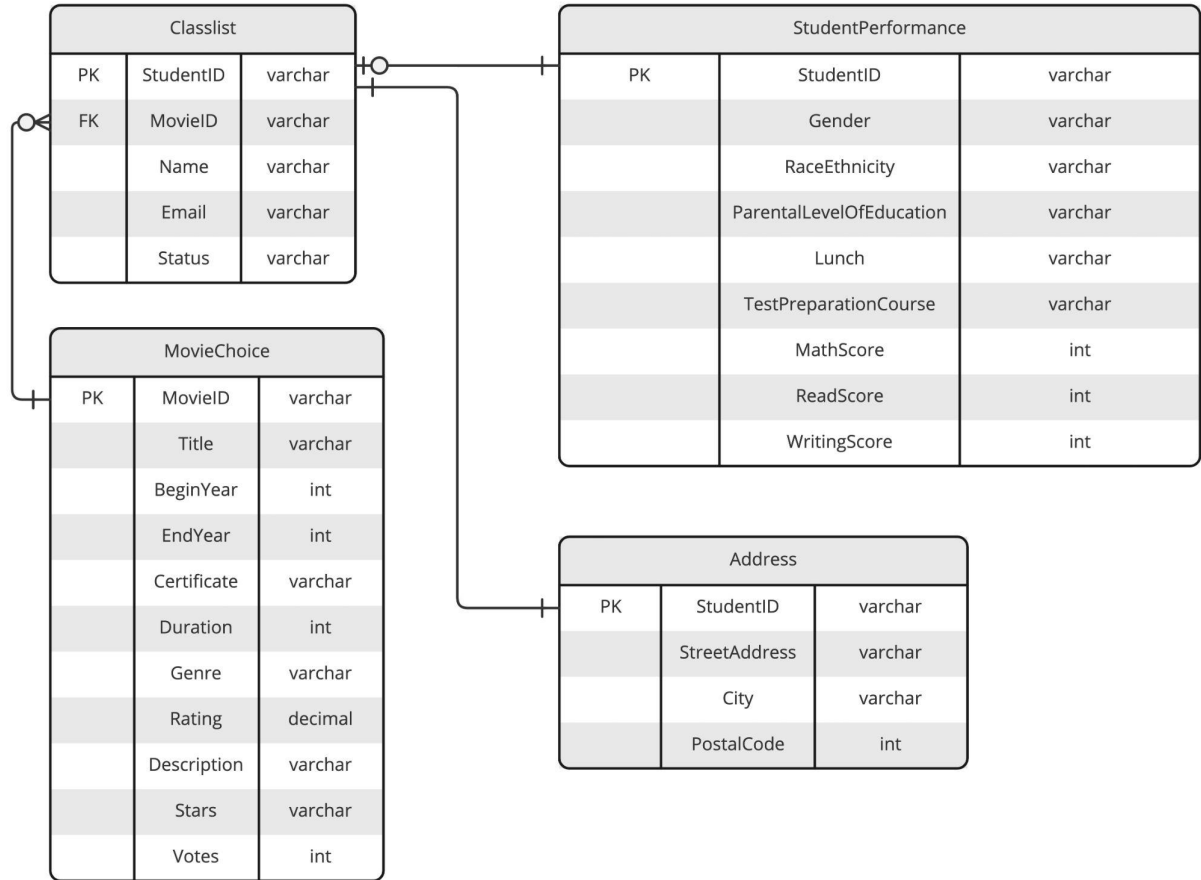
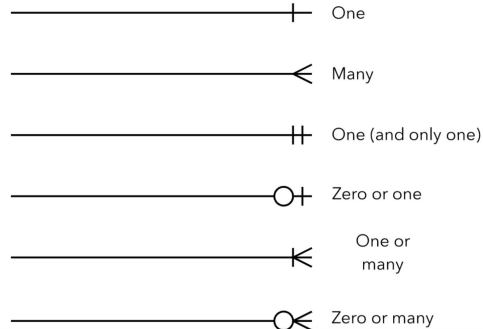
- **Address**

There's no obvious connection to **MovieChoice**, so we created a backstory:  
*During admission, all students were asked what their favourite movie was.*

# The design

## Database schema (entity relationship diagram)

### ERD Cardinality



# Cleaning (aka Excel Magic)



Classlist Hyper  
Island DA24.xlsx



StudentPerformance  
Incentive.xlsx



MovieChoiceOfDA  
24Student.xlsx

**Classlist** and **StudentPerformance** were clean.

**MovieChoice** was a mess:

- Missing data
- Numbers stored as strings
- Duration and Year had additional data

But no movie title was missing, so no line was excluded.

# Cleaning (aka Excel Magic)

**Year**, **Duration**, **Rating** and **Votes** were cleaned with good old Find/Replace.

**Year** was split in two columns, **BeginYear** and **EndYear**, with these formulae:

```
=NUMBERVALUE (LEFT (C2, 4) )
```

```
=IF (ISNUMBER (NUMBERVALUE (RIGHT (C2, 4) ) ) , NUMBERVALUE (RIGHT (C2, 4) ) , "")
```

And then:

```
=IF (L2=0, "", L2)
```

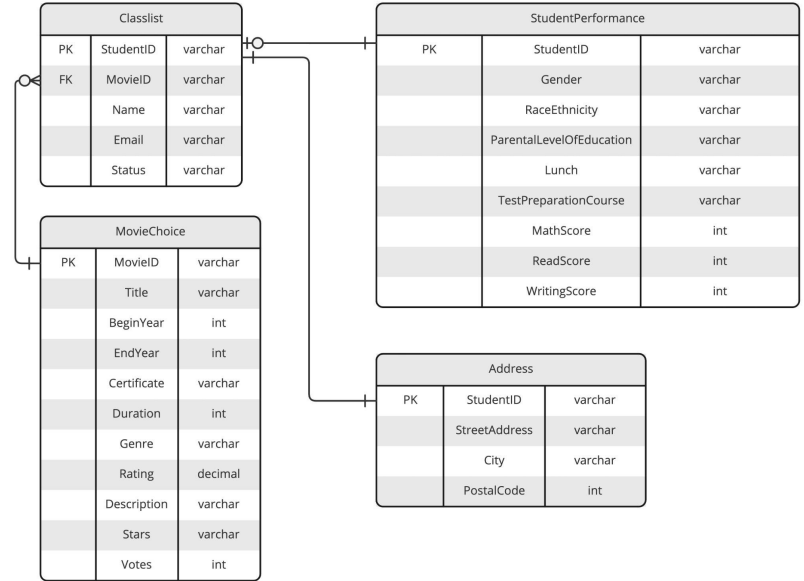
```
=IF (M2=0, "", M2)
```

# Connecting

The tables need a matching ID.

The same GUID (StudentID) as primary key:

- **Classlist**
- **StudentPerformance**
- **Address**



Another GUID (MovieID) to identify the movies:

- **MovieChoice** (primary key)
- **Classlist** (foreign key)

# Connecting

## Why use a GUID?

(a5bd8458-c730-4f1e-831b-8d20281e89f2; 128-bit in size, 32 digits long,  $2^{128}$  possibilities)

- The nature and potential size of the tables.
- Numerical IDs make sense when order and continuity needs to be maintained.
- A GUID or UUID is unlikely to be repeated.
- It can be uniquely created with `NEWID()` on MSSQL and `UUID()` on MySQL.



# Creating/Populating tables

Create CSV files.

When Excel runs out of magic:

- CSV files with semicolons, instead of commas
- Weird results when importing comma-separated numbers (used anywhere aside from the USA)

Just open the spreadsheets on Google Sheets and export them to CSV files.

Import into a clean database on MSSQL using the Import Wizard.

Watch out for the data types and primary keys.

# Queries

-- Their favourite movies?

USE TheSequels

SELECT Name, Title AS Movie

FROM Classlist

INNER JOIN MovieChoice

ON Classlist.MovieID

= MovieChoice.MovieID

ORDER BY Name;

	Name	Movie
1	Ali Akeel	Coco Before Chanel
2	Alva Salkert	The Flash
3	Ameer Abdulal	Call Me Chihiro
4	Amirhossein Zarabadi Pour	Mi soledad tiene alas
5	Anna Högman	Estocolmo
6	Ansar Yasmin	Larva Island
7	Carl Rickmon	Workin' Moms
8	Cláudio Rosa	Cuckoo Song
9	Eleni Kontou	The Queen and the Conqueror
10	Emy Hildeman	Daybreak
11	Eva Mayer	Blood Brothers: Malcolm X & Muhammad Ali
12	Fereshteh Mohammadi	The Tale of Nokdu
13	Guilherme Donati Bracco	Kong: King of the Apes
14	Hessam Ghaniyi	The Music of Strangers: Yo-Yo Ma and the ...
15	Idil Omar	Antiracist Baby
16	Jacob Einebrant Aspaas	Untitled Netflix/Chronicles of Narnia Series

# Queries

-- Where they live?

USE TheSequels

```
SELECT Name,  
        StreetAddress,  
        City,  
        PostalCode  
FROM Classlist  
INNER JOIN Address  
ON Classlist.StudentID  
    = Address.StudentID  
ORDER BY Name;
```

	Name	StreetAddress	City	PostalCode
1	Ali Akeel	Alsteråvägen 28	Venjan	79040
2	Alva Salkert	Alingsåsvägen 15	Hedared	50546
3	Ameer Abdulal	Grängsbo Löten 7	Junosuando	98062
4	Amirhossein Z...	Orrspelsv 37	Blåviksjön	92030
5	Anna Högman	Knektvägen 18	Luleå	95149
6	Ansar Yasmin	Söråsele 3	Hällaström	91063
7	Carl Rickmon	Figgeberg Gårdeb...	Edsele	88041
8	Cláudio Rosa	Backsippestigen 30	Mantorp	59020
9	Eleni Kontou	Mjölkalånga 78	Hofors	81300
10	Emy Hildeman	Östantorp Vinö 73	Lammhult	36030
11	Eva Mayer	Orrspelsv 8	Kirstineberg	92040
12	Fereshteh Mo...	Norr fjäll 54	Orrefors	38040
13	Guilherme Do...	Södra Kroksdal 87	Töreboda	54538
14	Hessam Ghaniyi	Eggelstad 3	Bureå	93015
15	Idil Omar	Villagatan 24	Trehörningsjö	89054
16	Jacob Einebra...	Käbbatorp Locketo...	Smedjebacken	77779

# Queries

	Name	Genre	ParentalLevelOfEducation
1	Cláudio Rosa	Drama, Horror, Mystery	high school
2	Kinza Shehzad	Comedy, Drama, Romance	high school
3	Lovisa Boberg	Documentary, Short	high school
4	Mohammad Reza Salarkarimi	Drama, Romance	high school
5	Rufta Zeray	Action, Comedy, Drama	high school

-- Genre of the movie chosen by the students whose parents only finished high school.

```
SELECT Name, Genre, ParentalLevelOfEducation
FROM Classlist, MovieChoice, StudentPerformance
WHERE Classlist.StudentID = StudentPerformance.StudentID
      AND Classlist.MovieID = MovieChoice.MovieID
      AND ParentalLevelOfEducation = 'high school'
ORDER BY Name;
```

# Queries

```
-- Which students have a score of 80 or more in Math?
```

```
USE TheSequels
```

```
SELECT Name, MathScore
```

```
FROM Classlist
```

```
INNER JOIN StudentPerformance
```

```
ON Classlist.StudentID
```

```
    = StudentPerformance.StudentID
```

```
    AND MathScore > 80
```

```
ORDER BY Name;
```

	Name	MathScore
1	Alva Salkert	83
2	Jonathan Devli	83
3	Milton Strandberg	89
4	Pamela Maunes Gumera	81
5	Shahzil Athar	82

EOF

Thanks!