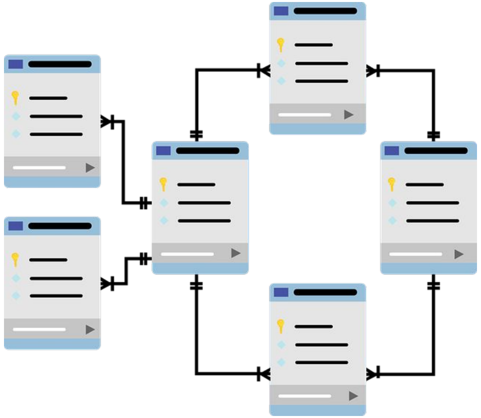
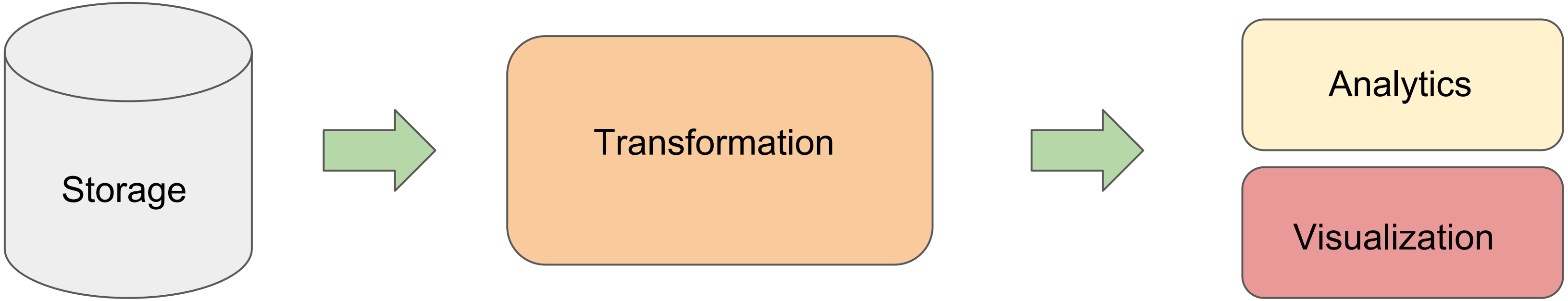
Data Transformations



Data Transformations for Data Science

- Data in source vs Data how you want it
- Data storage and analytics should be “de-coupled”
    - Same data can answer multiple questions
    - What queries should be sped up?
    - What information is needed ASAP vs at set intervals?

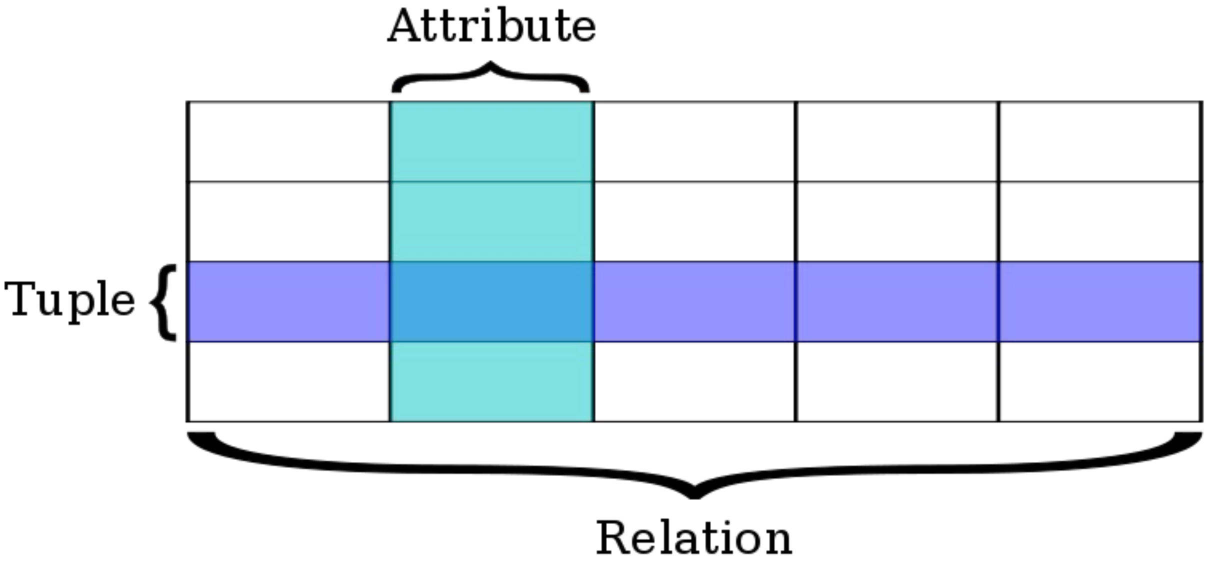


Data Storage

- How can data be stored?
    - Collection of files
        - Versatile
        - Slow for querying
    - Database - Organized by an index
        - Fulfils a more specific purpose
        - Fast querying
        - Requires planning at time of data collection
    - Database - No Index
        - More versatile than with index, still needs up-front planning
        - Quick if in-memory
        - Slower than with an index

Relational Databases

- Relational Model
  - Relations = Tables



SQL - Structured Query Language

- This is not an SQL course!
- SQL is a standard
  - Implementations can differ - the language is fairly consistent
  - Will encounter it in jobs often
  - Can transform data
  - Can pull data (query)
- Mostly meant for relational databases
  - Some systems have adapted it for types of storage
- https://www.w3schools.com/sql/

Select

```
SELECT column_1, column_2, ...
FROM table_name;
```

```
SELECT *
FROM table_name;
```

- Stored in “Result Set”
- Returns all rows

Distinct

```
SELECT DISTINCT column_1, column_2, ...
FROM table_name;
```

```
SELECT DISTINCT *
FROM table_name;
```

- Removes duplicates

Conditions - Where Statement

```
SELECT column1, column2, ...
FROM table_name
WHERE condition;
```

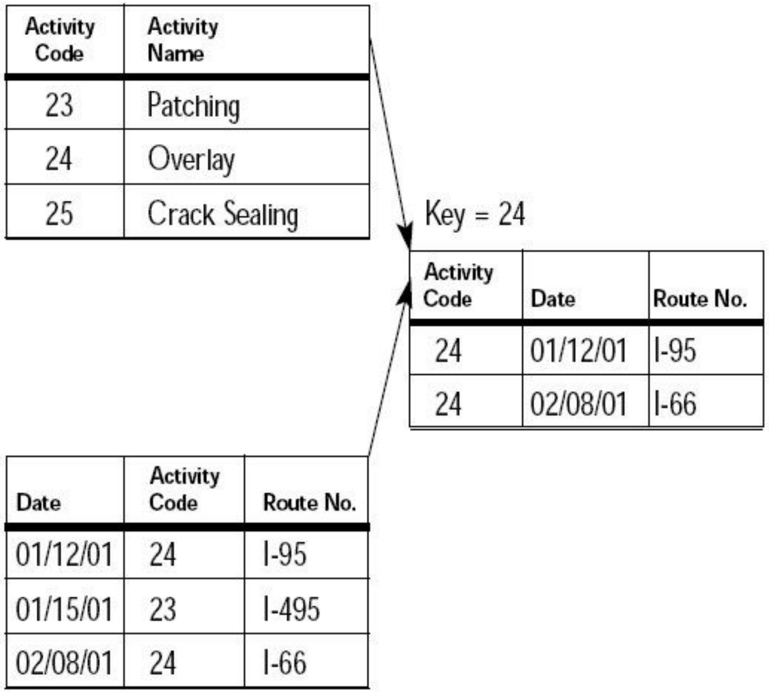
- Introduces a constraint to a query
- “Filters” results
- E.g. SELECT name from student_table WHERE class_enrolled=DS-CERT.

Where - Conditions

=	Equal
<> or !=	Not equal
>	Greater than
<	Less than
>=	Greater than or equal
<=	Less than or equal
BETWEEN	Between a range (inclusive)
LIKE	Patterns - https://www.w3schools.com/sql/sql_like.asp
IN	List of possible values

Integrating between tables

Relational Model



Activity Code	Activity Name
23	Patching
24	Overlay
25	Crack Sealing

Key = 24

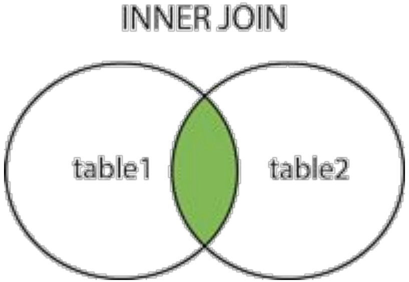
Activity Code	Date	Route No.
24	01/12/01	I-95
24	02/08/01	I-66

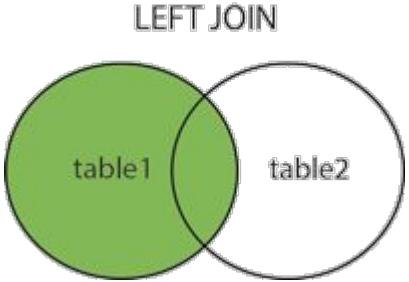
Date	Activity Code	Route No.
01/12/01	24	I-95
01/15/01	23	I-495
02/08/01	24	I-66

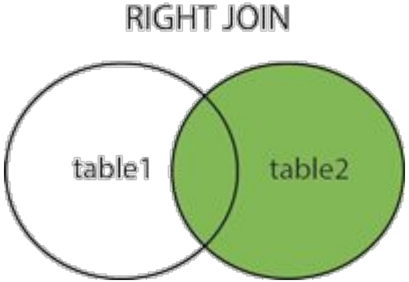
Joins

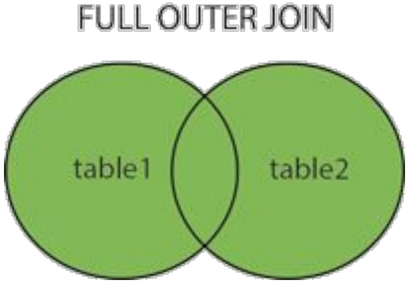
Images from: https://www.w3schools.com/sql/sql_join.asp

```
Table_1
<TYPE> JOIN Table_2 ON Table_1.column=Table2.column;
```

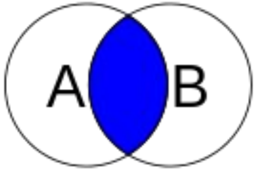




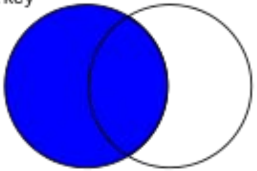




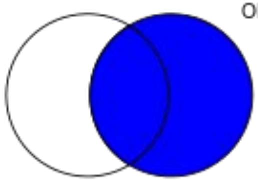
```sql
SELECT <fields>
FROM TableA A
INNER JOIN TableB B
ON A.key = B.key
```



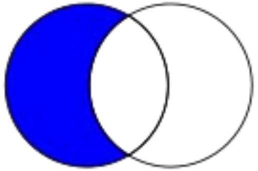
```sql
SELECT <fields>
FROM TableA A
LEFT JOIN TableB B
ON A.key = B.key
```



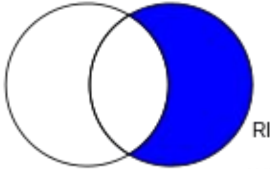
```sql
SELECT <fields>
FROM TableA A
RIGHT JOIN TableB B
ON A.key = B.key
```



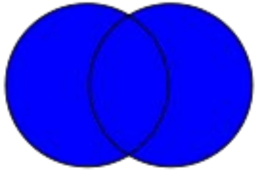
SQL JOINS



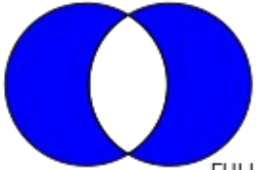
```sql
SELECT <fields>
FROM TableA A
LEFT JOIN TableB B
ON A.key = B.key
WHERE B.key IS NULL
```



```sql
SELECT <fields>
FROM TableA A
RIGHT JOIN TableB B
ON A.key = B.key
WHERE A.key IS NULL
```



```sql
SELECT <fields>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.key = B.key
```



```sql
SELECT <fields>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.key = B.key
WHERE A.key IS NULL
OR B.key IS NULL
```

Group and Aggregate

- Collect data into groups and then perform an operation
- E.g. Get average grade per student
  - Student is the group
  - Average is the operation

Group By Statement

```
SELECT column_1, column_2, ….
FROM table_name
WHERE condition
GROUP BY column_name(s)
ORDER BY column_name(s);
```

https://www.w3schools.com/sql/sql_groupby.asp

- Columns in select must be used in GROUP BY
- Alternatively you can choose aggregation operations

Aggregation Operations

- Get average grade per student?

```
SELECT student_id, avg(grade)
FROM students
GROUP BY student_id
```

- (COUNT, MAX, MIN, SUM, AVG)

SQL for data science

- Often we query to get data into one form and then transform in another
- Often use SQL to get data into a “flat” representation
- We will move on to Pandas in Python3 for exploring, analyzing and transforming data

Notebooks

http://jupyter.org/

https://colab.research.google.com

Integrates with Pandas, matplotlib, sklearn, etc.