Chapter 1

SQL for Data Preparation

1.1 组合数据

1.1.1 连接的类型

You will learn about three fundamental joins, which are illustrated in Figure 1.1—inner joins, outer joins, and cross joins:

外连接

完全外连接将返回左表和右表中的所有行，无论连接谓词是否匹配。对于满足连接谓词的行，两行将被组合起来，就像内部连接一样。对于不满足的行，两个表中的每一行都将被选择为单独的行，并为另一个表中的列填充 **NULL**。完全外连接是通过使用 **FULL OUTER JOIN** 子句并后跟连接谓词来调用的。

1.1.2 Subqueries

If a query only has one column, you can use a subquery with the **IN** keyword in a **WHERE** clause.

1.1.3 Unions

Please note that there are certain conditions that need to be kept in mind when using **UNION**. Firstly, **UNION** requires the subqueries to have the same number of columns and the same data types for the columns. If they do not, the query will fail to run. Secondly, **UNION** technically may not return all the rows from its subqueries. **UNION**, by default, removes all duplicate rows in the output. If you want to retain the duplicate rows, it is preferable to use the **UNION ALL** keyword.

1.1.4 Common Table Expressions(CTEs)

CTEs are simply a different version of subqueries. CTEs establish temporary tables by using the **WITH** clause. The one advantage of CTEs is that they can be designed to be recursive. **Recursive**

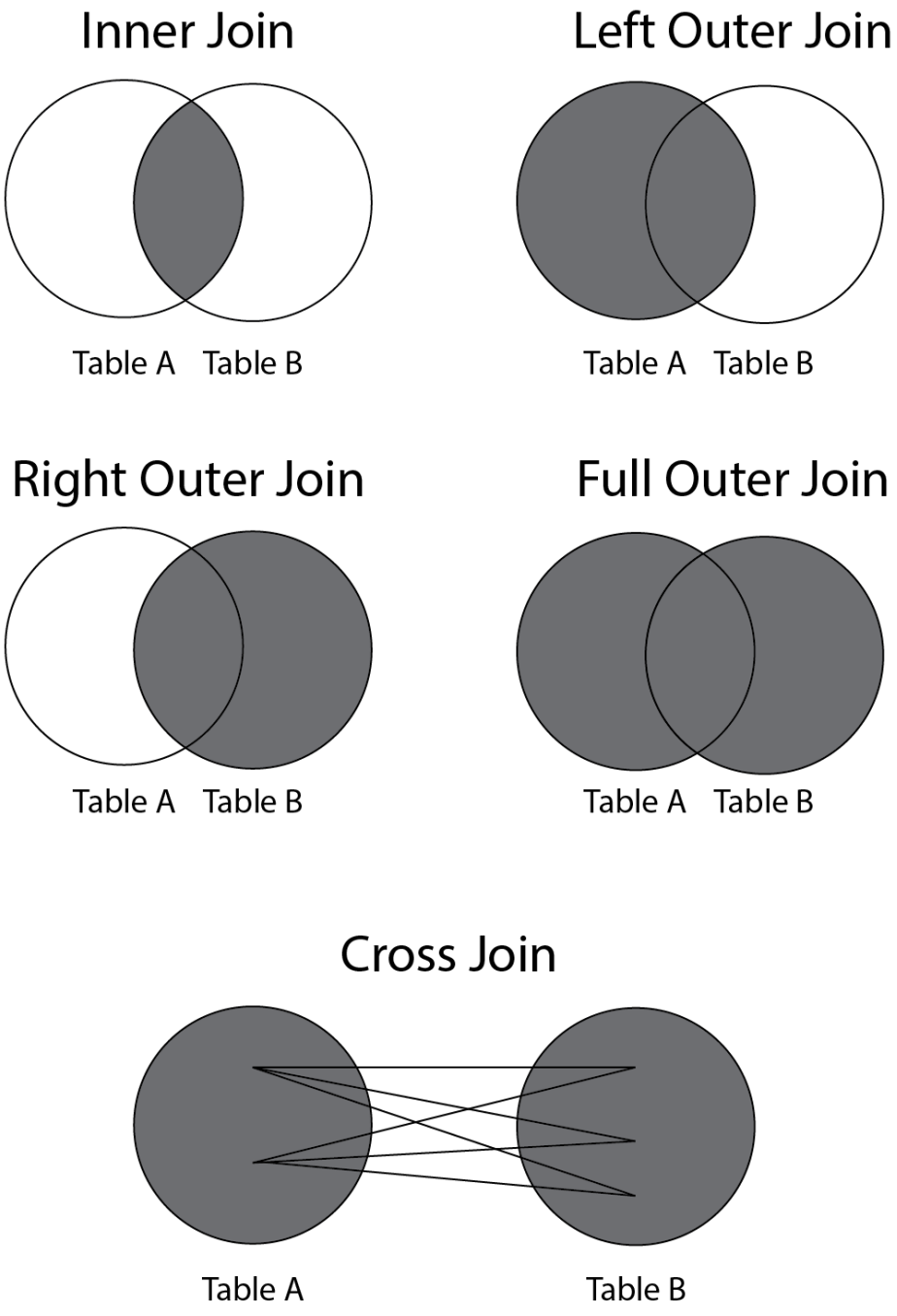

图 1.1: Major types of joins

CTEs can reference themselves. Because of this feature, you can use them to solve problems that other queries cannot.

1.2 Cleaning Data

1.2.1 The CASE WHEN Function

CASE WHEN is a function that allows a query to map various values in a column to other values.

1.2.2 The COALESCE Function

Another common requirement is to replace the NULL values with a standard value. This can be accomplished easily by means of the COALESCE function. COALESCE allows you to list any number of columns and scalar values, and, if the first value in the list is NULL, it will try to fill it in with the second value. The COALESCE function will keep continuing down the list of values until it hits a non-NULL value. If all values in the COALESCE function are NULL, then the function returns NULL.

1.2.3 The NULLIF Function

NULLIF is used as the opposite of COALESCE. While COALESCE is used to convert NULL into a standard value, NULLIF is a two-value function and will return NULL if the first value equals the second value.

1.2.4 The LEAST/GREATEST Functions

Two functions that come in handy for data preparation are the LEAST and GREATEST functions. Each function takes any number of values and returns the least or the greatest of the values, respectively. The simple use of this variable would be to replace the value if it is too high or low.

1.2.5 The Casting Function

To change the data type of a column, you simply need to use the **column::datatype** format, where column is the column name and datatype is the data type you want to change the column to.

Please note that not every data type can be cast to a specific data type.