



# Data Lineage Databases & SQL

---

Understanding how data lineage is used in databases  
and SQL

R Shane James

Year: 2023

---





# Data Lineage

---



# What is data lineage?

---

Data lineage is the process of keeping track of where data comes from, how it moves through different systems and processes, and where it ends up. It is the history of the data, including where it came from, how it moved, and how it changed over time. It can also include things like timestamps and information about the user.

Data lineage is an essential part of data governance because it helps organizations figure out the quality and origin of their data. It can also help answer questions like, "Where did this information come from?"

- What steps did it take to get there?
-

- Who has used it before?



- What changes has it gone through?



- How did it get cleaned up and changed?



- What does it do?



- What does the data look like right now?



Data lineage is also important for compliance and regulatory purposes. Organizations need to be able to track where the data came from, how it moved, and how it was used to show that they are following regulations like GDPR, HIPAA, and SOX.


Data lineage can be tracked in different ways, such as through code comments, data mapping, data dictionaries, and specialized tools like MANTA, Collibra, Informatica, Talend, etc. They can track the lineage of data in real time and make detailed lineage diagrams to show how data moves and changes. Data catalogs, dashboards, and reports can also be used to show the lineage of data. This makes it easier for users to understand and use the data.

## Let's cover some of the common questions about a database and SQL

### What is a database?

A database is organized information usually kept on a computer system. It is usually set up to make it easy to find, add, and get rid of data. Databases store and manage much information, like customer information, product inventories, financial records, etc.

A database is a powerful way to store and organize information. It is set up to make it easy to find, add, and get rid of information. Databases can store and manage information about customers, products, finances, and more. They can also keep and look at data to help make decisions, make predictions, and run other business tasks. Several programming languages, such as SQL, Java, and Python, can be used to access and change databases.





Databases are used by businesses, government agencies, educational institutions, and non-profit organizations, among other groups and people. They are used to store and organize many data, like information about customers, product inventories, financial records, and more. Professionals such as database administrators, software developers, and data analysts often use databases.

Data lineage is figuring out where data came from and how it got to where it needs to go. It is used to understand how information moves through an organization and find problems or differences. In data lineage, databases store and keep track of the data being tracked. This information can include where the data came from, where it was going, how it was changed, and how long it took to move from its source to its destination. By keeping track of this data, organizations can find problems or differences in the flow of data and fix them.

What is SQL?

SQL (Structured Query Language) is a programming language used to communicate with databases. It creates, updates, deletes, and retrieves data from databases. It is a standard language for relational database management systems, allowing users to work with data in a structured way.



# How is SQL used?

SQL is used to store, manipulate, and retrieve data in databases. It creates, updates, deletes, and recovers data from databases. It also creates database objects such as tables, views, stored procedures and functions, triggers, and indexes. SQL can also be used to control user access to the database and to create database security mechanisms.

What are database views?

Database views are virtual tables created by joining one or more tables. They are typically used to simplify complex queries and to provide a different perspective of the data. Views can also restrict access to specific tables and columns or combine data from multiple tables.

# Examples of the SQL language

```
SELECT * FROM table_name;
```

```
UPDATE table_name SET column_name = new_value;
```

```
DELETE FROM table_name WHERE condition;
```

```
INSERT INTO table_name (column1, column2, ...) VALUES (value1, value2, ...);
```

```
CREATE VIEW view_name AS SELECT column1, column2, ... FROM
table_name WHERE condition;
```

What are stored procedures used for?

Stored procedures are pre-defined SQL statements that are stored in the database. They are used to execute a set of SQL statements repeatedly, such as to perform a complex query or to update multiple tables. Stored procedures can also create user-defined functions, triggers, and views.

Who cares about SQL?

SQL is used by database administrators, software developers, web developers, and anyone who needs to access and manipulate data in a database. It is a powerful and versatile language that can be used to perform a wide variety of tasks.

How is SQL used in data lineage?

SQL is used in data lineage to track the origin and movement of data between systems. It is used to query data sources, identify data transformations, and map data flows between systems. Data lineage can help organizations understand the relationships between data sources and data consumers and ensure that data is being used correctly.

## How do data transformations work in SQL?

Data transformation in SQL typically involves using SQL SELECT statements to extract and transform data from one or more tables in a relational database. Some common SQL operations used for data transformation include:



1. **SELECT** statement: This is used to extract specific columns from a table and can be used to rename, add new columns based on calculations, or filter rows based on specific criteria.
2. **JOIN** statement: This combines data from multiple tables. For example, you can join data from a "customers" table and an "orders" table to create a new table that contains information from both tables.
3. **GROUP BY** statement: This is used to group rows based on one or more columns and can be used to aggregate data, such as calculating a sum or average.
4. **SUBQUERY** statement: This is used to retrieve data from one table based on the data from another table. For example, you can use a subquery to retrieve all orders for a customer with a particular credit rating.
5. **UPDATE** statement: This is used to modify data in a table. For example, you can use an update statement to change the value of a specific column for all rows in a table.
6. **UNION** statement: This combines the results of two or more **SELECT** statements.
7. **CASES** statement: This is used to change the values of a column based on a certain condition; it's like a conditional mapping

All these statements and operations in SQL can be used together to perform complex data transformations. With the help of these operations, it is possible to clean, transform and aggregate data into meaningful and valuable information that can be used for further analysis or reporting.