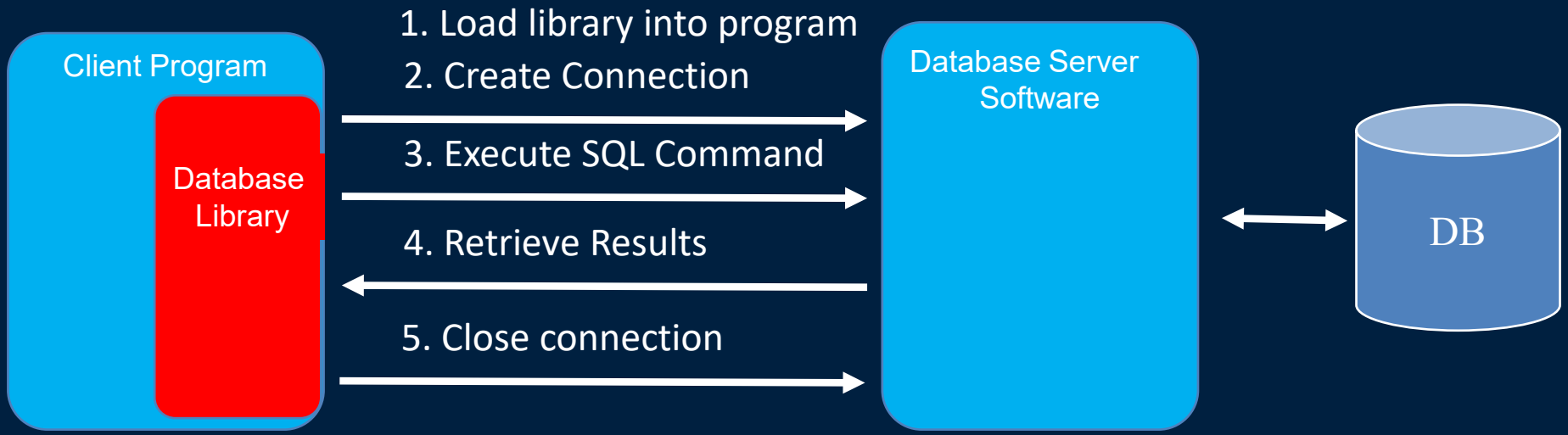
Hosting and Data Wrangling

COSC 304 – Introduction to Database Systems

Review: SQL Programming Architecture



1) Load Library Error:
- Library not found in path
- Wrong library

2) Create Connection Errors:
- Invalid server URL
- Incorrect user/password
- Network issues
- Wrong library

3) Execute SQL Errors:
- Incorrect SQL
- Wrong database/table
- Improper library use

4) Retrieve Results Errors:
- Wrong column name
- Wrong column index
- Off-by-one
- Improper library use

5) Close Connection:
- FORGET TO DO IT!

Database Hosting – Where's the Server?

Database hosting selects the machine where the database software executes. Choices:

- Local machine (`localhost`)
- On-premise physical or virtual machine (`cosc304.ok.ubc.ca`)
- Cloud-based hosting (physical/virtual/container) as a service on platforms such as Amazon, Microsoft Azure, Google, Digital Ocean.

Database host must be accessible over the Internet by the clients that connect to it.

Running a Database on `localhost`

Running a database on your machine (`localhost`) is easy. Steps:

- Download and install database software (MySQL, PostgreSQL, etc.)
- Configure and start the database server software
- *OR:* Run DBMS software virtually using containerization such as Docker.

Advantages:

- Full control over database and install process

Disadvantages:

- May not be easy to connect to by clients depending on machine
- Must take time to install/configure database software

Running a Database On-Premise

Running "On-Premise" is when the database is deployed on a (virtual) machine controlled by the organization.

This is often done for security and for performance.

Advantages:

- Data does not leave organization.
- Potential for higher performance.

Disadvantages:

- Organization responsible for deploying, configuring, maintaining, and securing both hardware and database software.

Cloud Databases

Cloud databases are databases hosted by a service provider that allow for easy setup, administration and scaling.

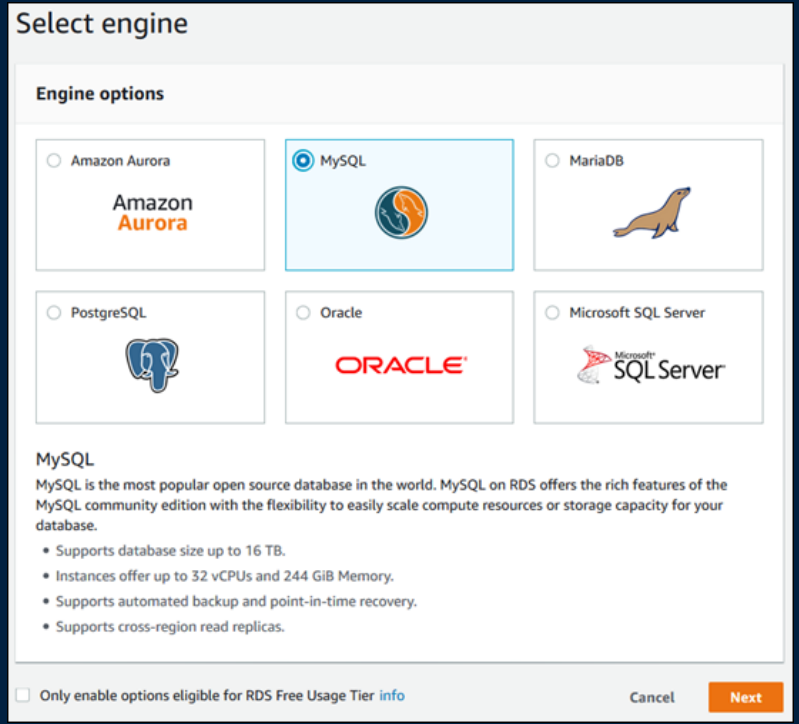
- Database as a service – databases hosted by provider, provide monitoring, backup, fail-over, high-availability, and ability to scale.

Examples: Amazon RDS, Microsoft Azure, Google CloudSQL, Digital Ocean

Ideal for: Quick start without a server, minimal administration, scaling without expertise

Creating MySQL Instance on Amazon

1) Sign in to AWS Management Console.
   - https://console.aws.amazon.com/rds
2) Select database engine/version.
3) Select database instance size (CPU/memory/storage) and user/password configuration.
4) Configure advanced settings (network accessibility).
5) **Verify price** and Create Database.



Reference:
docs.aws.amazon.com/AmazonRDS/latest/UserGuide/CHAP_Tutorials.WebServerDB.CreateDBInstance.html

Creating PostgreSQL on Digital Ocean

Sign in to Digital Ocean.

- https://cloud.digitalocean.com

Option #1: User Managed DB

- Create a Droplet.

- Select droplet size and region.

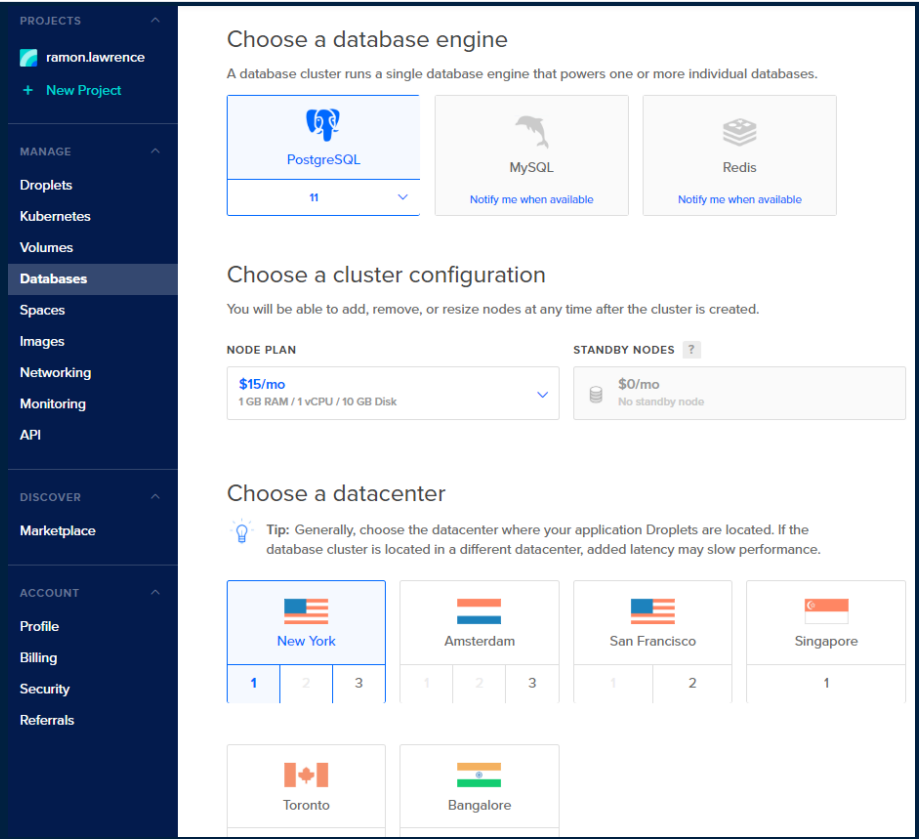
- Login to droplet and install/configure PostgreSQL.

Option #2: Managed DB

- Select Databases and Click Create a Cluster.

- Determine size and region.

- Record login and URL info.



Database Hosting Question

Question: How many of the following statements are **TRUE**?

1) Hosting a database on-premise is always better than hosting in the cloud.

2) Database as a service is when the database instance is managed for you by the cloud service provider.

3) It is possible to get a virtual machine from a cloud host and install your own database software on it.

4) On-premise hosting is always cheaper than cloud hosting.

A) 0 **B)** 1 **C)** 2 **D)** 3 **E)** 4

Data not in Databases

Databases make it easy and efficient to store and query data, but often people do not use them.

Data not in databases is present in the following forms:

- Text files (CSV, tab separated)
- Structured text files (JSON, XML)
- Microsoft Excel and Microsoft Access
- Custom file formats (text or binary)

How to handle such data?

Everything in a Database?

There are many reasons why data does not reside in a database:

- Cost/time to import it
- Lack of expertise to create/maintain database
- Data has limited long-term value
- Existing format is sufficient for use case

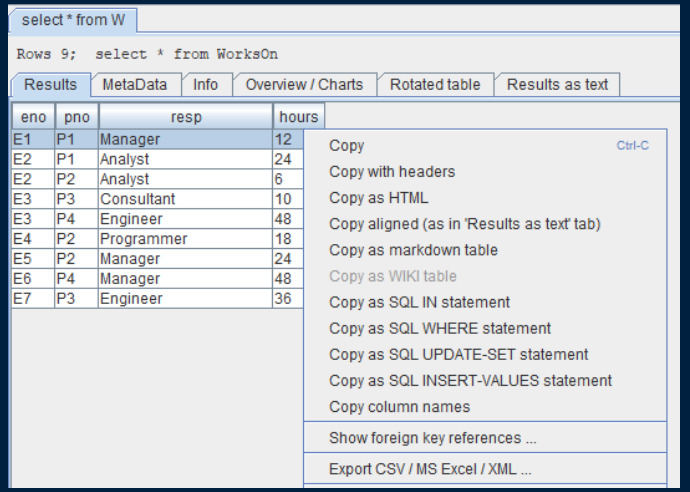
Handling Text Files

Importing text data (CSV files) into a database is so common that most databases have utilities to support it.

- MySQL – `LOAD DATA INFILE "my.csv" INTO TABLE T`
- SQL Server – Import Data task in SQL Server Studio

Exporting from a database into CSV

- MySQL – `SELECT … INTO OUTFILE`
- SQL Server – Export Data task in SQL Server Studio
- SQuirreL – Execute query, right click on result, and select Export CSV…

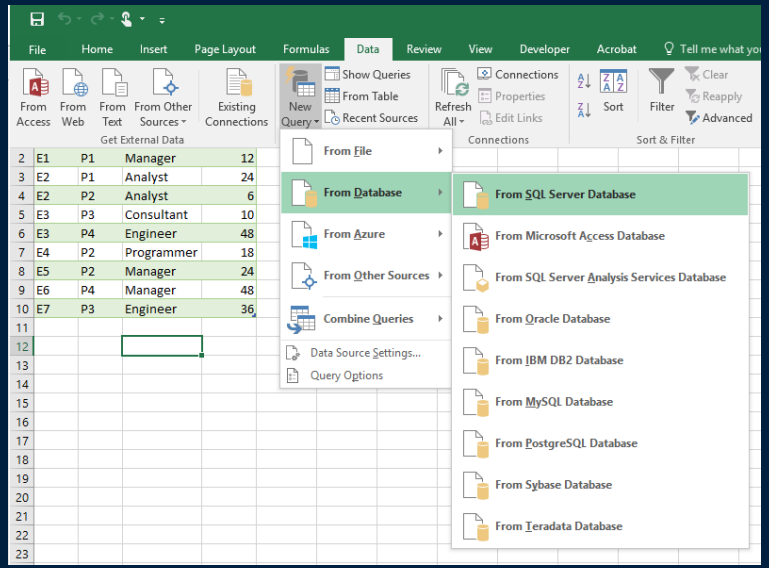


Interfacing with Microsoft Excel

Microsoft Excel is used to store data without using a database. Excel allows importing and exporting data with databases.

Importing data into Excel from a database:

- Data menu, New Query, Select database, write SQL query



Try it!
Open Excel and import data from SQL Server.

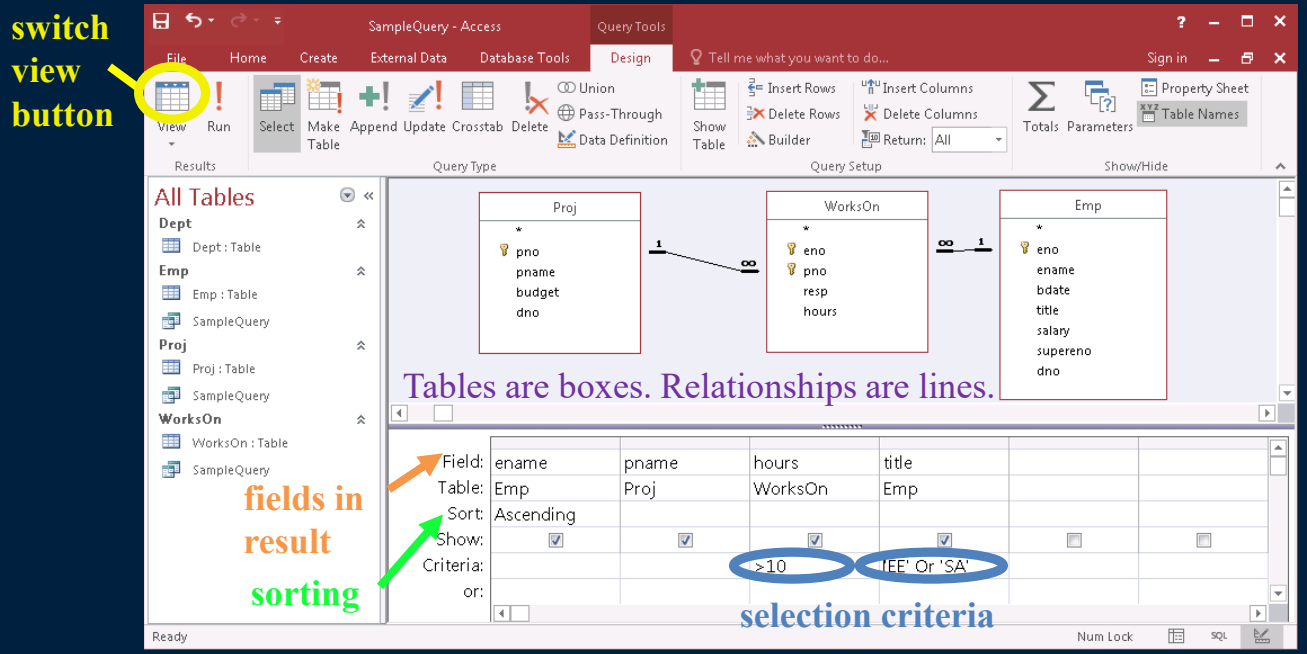
Microsoft Access

Microsoft Access is a file-based database designed for a small number of users.

It supports SQL but most interactions are through graphical interface.

Data can be moved between Access and Excel and other databases easily.

Microsoft Access Query Interface



Reporting and Visualization

Database data is consumed by:

- Programs (Java, Python, R, Web) with embedded visualizations
- Stand-alone reporting tools such as Tableau, QlikView, Google Data Studio, Microsoft Reporting Services/Power BI

Visualization allows people to understand and extract information faster and with more accuracy than displaying text and numbers.

All reporting/visualization software interact with the database in essentially the same way: create a connection, build a SQL statement to execute, retrieve results, and visualize.

The difference is how the various tools automate or hide the complexities of connecting and building SQL statements.

Introduction to Tableau

Tableau (http://www.tableau.com/) was founded in 2003 as a spin-off from Stanford University by Chris Stolte, Christian Chabot and Pat Hanrahan.

- 2019 revenue was about $1.5 billion with over 4000 employees. Bought by SalesForce in 2019.
- The goal of Tableau is "to help people see and understand their data." - Christian Chabot, Tableau CEO

Tableau makes is very easy to construct visualizations from data. It connects to many data sources including: Excel/Access, text files, relational databases.

- Queries are generally built through the graphical interface.

Tableau Workspace Items

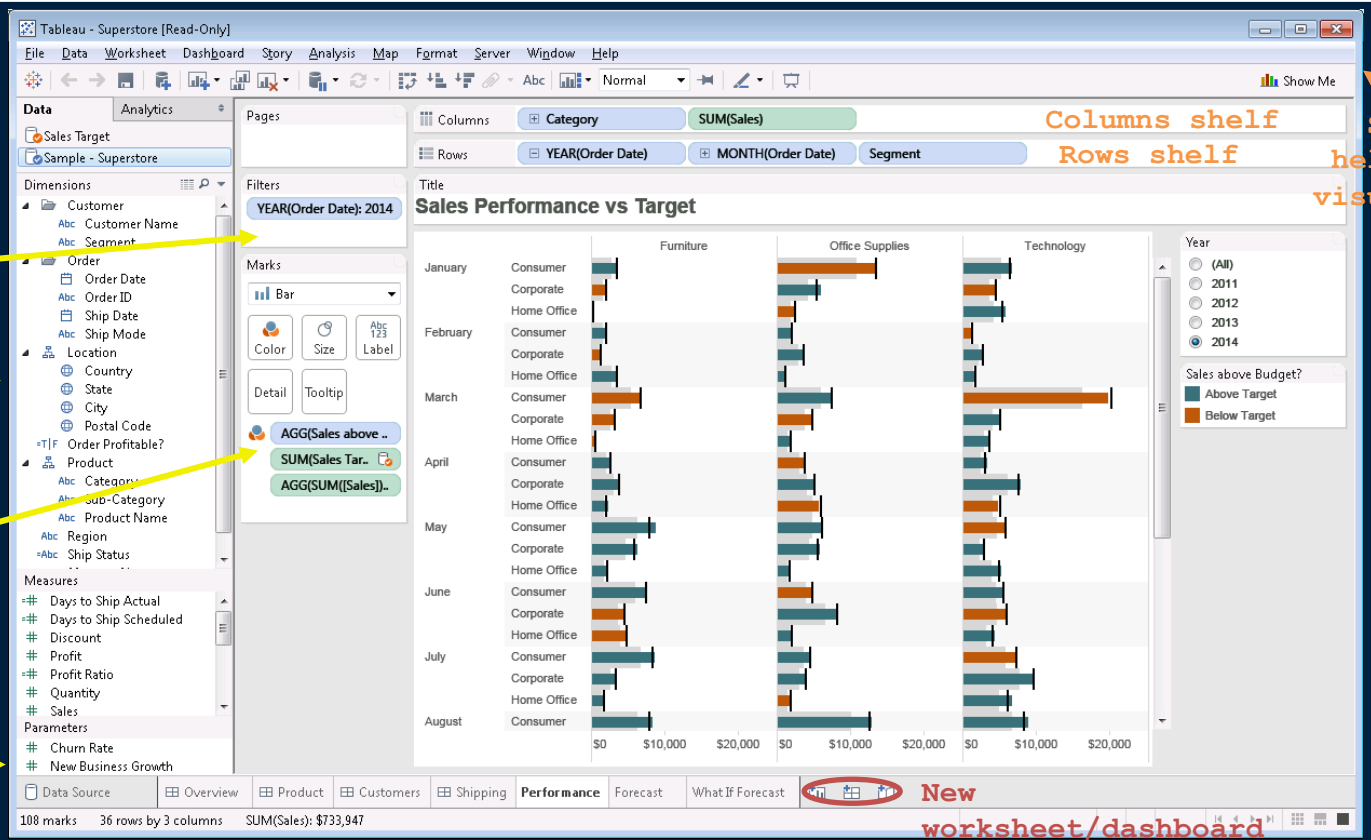
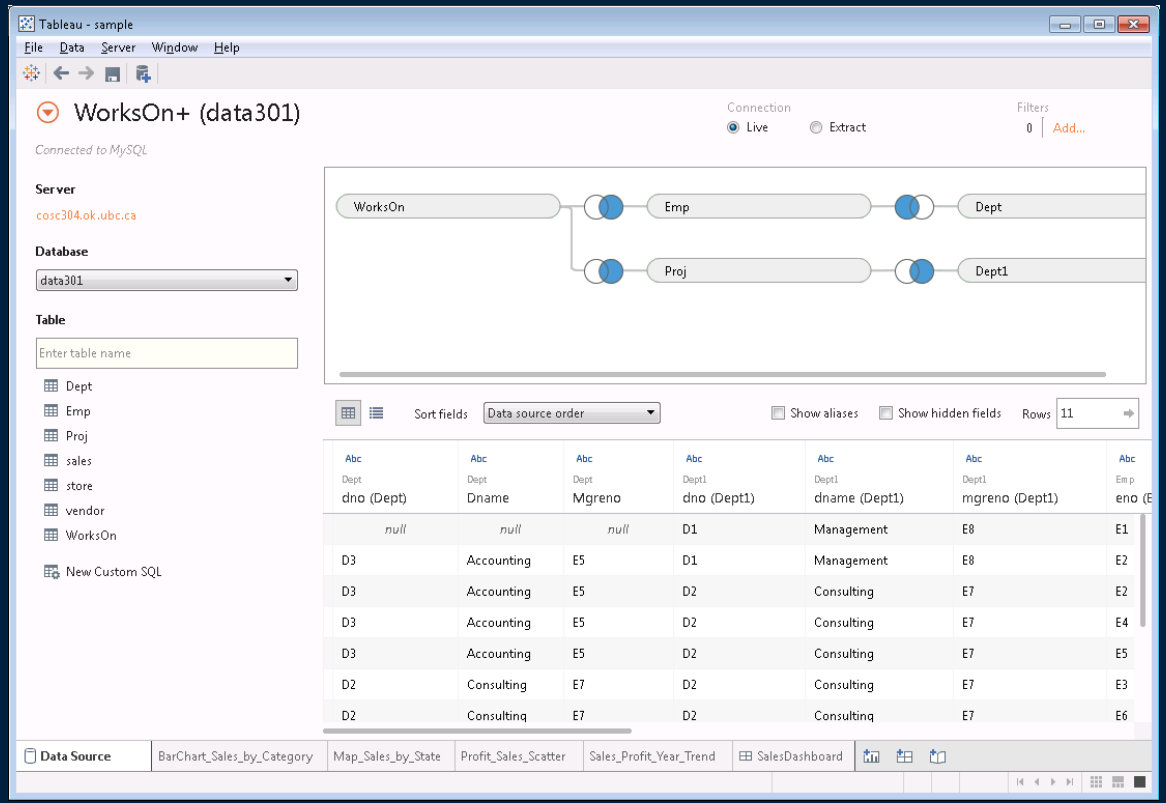
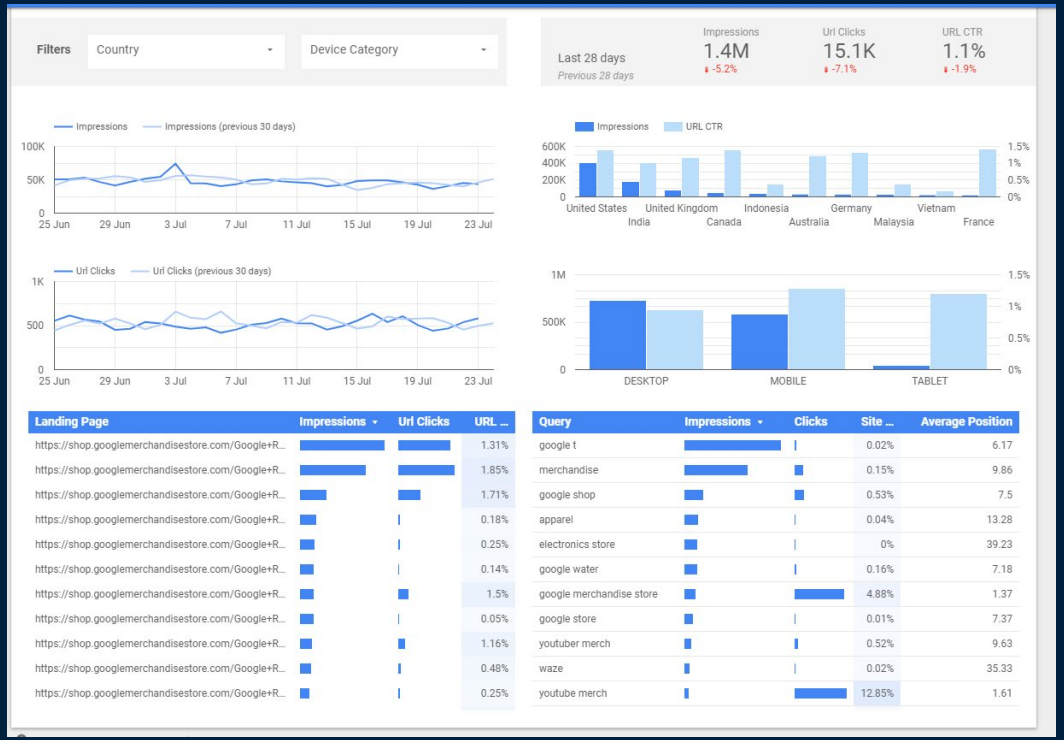


Tableau Connecting to MySQL



Google Looker Studio

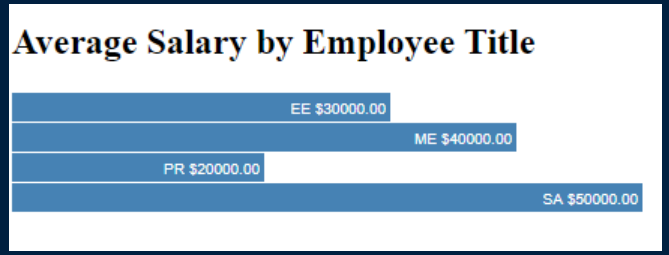
Google Looker Studio is a free, cloud-based data analysis and reporting software. https://lookerstudio.google.com



D3

D3 is a JavaScript visualization library that can retrieve and display data from the database.

- Since JavaScript is executing on the client, a REST API or AJAX call to a server to get data is required.



Steps:

- Client retrieves HTML page containing D3 code.
- D3 requests data from server using a URL.
- Server has code in PHP/JSP/Node.js to return data in JSON format.
- D3 renders data on chart in browser.

Visualization and Reporting Question

Question: How many of the following statements are **TRUE**?

1) In Tableau most database interactions are through a graphical interface.

2) Tableau allows users to write their own Custom SQL query.

3) Google Looker Studio is a cloud-based visualization solution.

4) D3 is written in Java.

5) D3 connects to a database directly to retrieve data.

A) 0 **B)** 1 **C)** 2 **D)** 3 **E)** 4

Conclusion

Database hosting selects the machine where the database software executes.

- Hosting can be on a local machine, on a machine within the organization, or on some cloud service.

Handling text files with databases is common for both import and export of data.

- Databases have specific features for handling CSV/text import and export so often code does not need to be written.

Visualization and reporting is a key use case for database data. All libraries require making a connection and executing a query to get results. Differences are on the user interface and how database details are hidden from the user.

Objectives

- Define database hosting.
- Compare and contrast benefits/challenges with hosting on local machine, on premise, and on a cloud server.
- List some text file formats.
- Be aware of database features and tools that make it easy to import and export data from databases.
- Understand how Excel can both import and export database data.
- Appreciate that Microsoft Access is a simple, file-based database that makes it easy for users to build database and queries through a graphical interface.
- List some different visualization and reporting software and explain their role in a database application architecture.


UBC
THE UNIVERSITY OF BRITISH COLUMBIA