Basis Data dan Penelusuran Data

Bakti Siregar, M.Sc

2023-08-26

Contents

Kata Pengantar

Selamat datang dalam modul praktikum mengenai basis data dan penelusuran data. Dalam era digital yang semakin maju, pengelolaan informasi dan akses terhadap data sangatlah penting. Basis data merupakan fondasi utama dalam pengelolaan data yang efisien dan terstruktur, sedangkan penelusuran data memungkinkan kita untuk menggali wawasan berharga dari kumpulan informasi yang tersedia. Dalam modul ini, kita akan menjelajahi konsep-konsep dasar dalam basis data, termasuk jenis-jenis basis data, model data, bahasa kueri, dan praktik terbaik dalam merancang basis data yang optimal. Secara khusus, mudul ini

Selain itu, penelusuran basis data yang menjadi fokus penting adalah menggunakan R Programing dan SQL dalam membuat data analytics system. Penelusuran data melibatkan teknik-teknik dan alat-alat untuk menggali informasi yang berharga dari kumpulan data yang besar dan kompleks. Dengan adanya kemajuan dalam analisis data dan kecerdasan buatan, penelusuran data telah menjadi aspek penting dalam pengambilan keputusan dan inovasi. Penulis berharap bimbingan ini akan memberikan pemahaman yang kokoh tentang basis data dan penelusuran data, serta memberi Anda wawasan yang berguna dalam mengelola data dan mengambil informasi berharga dari sumber daya yang ada. Selamat belajar!

Ringkasan Materi

Adapun isi pembelajaran dalam modul ini adalah sebagai berikut:

- Bab 1
- Bab 2
- Bab 3
- Dst

Penulis

- **Bakti Siregar, M.Sc** adalah Ketua Program Studi di Jurusan Statistika Universitas Matana. Lulusan Magister Matematika Terapan dari National Sun Yat Sen University, Taiwan. Beliau juga merupakan dosen dan konsultan Data Scientist di perusahaan-perusahaan ternama seperti JNE, Samora Group, Pertamina, dan lainnya. Beliau memiliki antusiasme khusus dalam mengajar Big Data Analytics, Machine Learning, Optimisasi, dan Analisis Time Series di bidang keuangan dan investasi. Keahliannya juga terlihat dalam penggunaan bahasa pemrograman Statistik seperti R Studio dan Python. Beliau mengaplikasikan sistem basis data MySQL/NoSQL dalam pembelajaran manajemen data, serta mahir dalam menggunakan tools Big Data seperti Spark dan Hadoop. Beberapa project beliau dapat dilihat di link berikut: Rpubs, Github, Website, dan Kaggle.

Asisten Lab

- **Yonathan Anggraiwan, S.Stat** adalah seorang alumni Statistika yang bersemangat dalam dunia pemrograman dan analisis data. Lahir di Tangerang, minatnya terhadap teknologi dan komputer muncul sejak usia dini. Ia tumbuh dengan rasa ingin tahu yang kuat terhadap bahasa pemrograman, dan ini membawanya menuju dunia analisis data menggunakan bahasa pemrograman R dan Python. Selama menjalankan tugas sebagai asisten lab, Yonathan Anggraiwan berperan dalam membantu mahasiswa dalam memahami konsep-konsep dasar dan kompleks dalam pemrograman R dan Python. Ia memberikan penjelasan yang jelas dan dukungan kepada mahasiswa yang mengalami kesulitan. Selain itu, ia juga terlibat dalam merancang tugas dan ujian praktikum, serta memberikan umpan balik konstruktif kepada para mahasiswa. Dalam perjalanan waktu, Yonathan Anggraiwan mulai mengambil tanggung jawab lebih besar dalam laboratorium. Ia membantu mengembangkan materi pembelajaran tambahan, seperti tutorial online tentang analisis data menggunakan R dan Python. Ia juga aktif dalam berbagai proyek penelitian di bawah bimbingan dosen, yang melibatkan pengolahan data besar untuk analisis statistik dan visualisasi. Dengan semangat yang tinggi, dedikasi, dan keterampilan yang dimilikinya, Yonathan Anggraiwan adalah contoh nyata dari seorang mahasiswa yang berhasil menggabungkan minatnya dalam pemrograman R dan Python dengan peran yang produktif sebagai asisten laboratorium dan kontributor dalam dunia analisis data.

Ucapan Terima Kasih

Saya ingin mengucapkan terima kasih yang tulus kepada semua yang telah mendukung dan berkontribusi dalam perjalanan pembuatan modul "Basis Data dan Penelusuran Data". Modul ini tidak akan mungkin menjadi kenyataan tanpa kerja keras, semangat, dan dukungan yang luar biasa dari berbagai pihak. Terima kasih juga kepada rekan-rekan dan kolega yang telah memberikan masukan, saran, dan diskusi berharga sepanjang perjalanan penulisan modul ini. Kontribusi kalian telah membantu memperkaya isi modul dan menghadirkan sudut pandang yang beragam. Tentu saja,modul ini tidak akan lengkap tanpa rasa terima kasih kepada para peneliti dan praktisi di bidang basis data dan penelusuran data yang telah menciptakan landasan pengetahuan yang menjadi dasar dari modul ini. Pengalaman dan pengetahuan yang kalian bagikan sangat berharga. Saya juga ingin mengucapkan terima kasih kepada keluarga dan teman-teman saya atas dukungan, pengertian, dan dorongan yang tak hentihentinya. Tanpa dukungan kalian, perjalanan menulis modul ini pastinya tidak akan semudah ini.

Akhir kata, semoga modul ini dapat memberikan manfaat dan wawasan baru kepada para pembaca yang ingin mendalami dunia basis data dan penelusuran data. Ucapan terima kasih terakhir saya tujukan untuk semua yang telah berkontribusi, baik secara langsung maupun tidak langsung, dalam menghadirkan modul ini kepada para pembaca.

Masukan & Saran

Semua masukan dan tanggapan Anda sangat berarti bagi kami untuk memperbaiki template ini kedepannya. Bagi para pembaca/pengguna yang ingin menyampaikan masukan dan tanggapan, dipersilahkan melalui kontak dibawak ini!

Email: dsciencelabs@outlook.com

Chapter 1

Pendahuluan

Sejak tahun 1970, **Structured Query Language (SQL)** telah digunakan oleh para programmer untuk membangun dan mengakses **Sistem Basis Data (SBD)**. Banyak sekali perdebatan mengenai cara penyebutan SQL ini, namun pada kenyataannya, kita dapat melafalkannya sebagai "sequel" ataupun "S.Q.L". Mempelajari bahasa pemrograman umum seperti R adalah penting dan akan lebih baik jika memiliki kemampuan SQL dalam bidang pengolahan data.

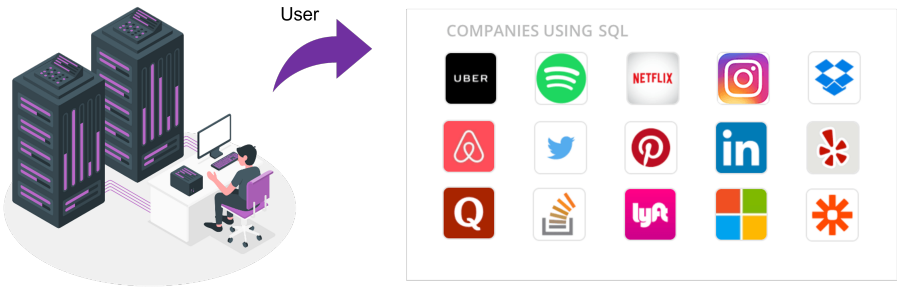


Figure 1.1: Beberapa Perusahaan Besar Pengguna SQL

Banyak perusahaan besar di bidang teknologi menggunakan SQL seperti Uber, Netflix, dan Airbnb. Bahkan dalam perusahaan seperti Facebook, Google dan Amazon, yang telah membuat sendiri **SBD** berkemampuan tinggi, tetap menggunakan SQL untuk melakukan query dan analisis data.

1.1 Apa itu SBD?

Secara umum **SBD** dapat didefinisikan sebagai berikut:

Figure 1.2: Definisi Sistem Basis Data

1.1.1 Komponen SBD

Adapun beberapa komponen dasar yang diperlukan dalam SBD adalah:

1.1.2 Manfaat SBD

Manfaat atau kegunaan penerapan SBD cukup banyak dan cakupannya pun luas dalam mendukung keberadaan lembaga atau organisasi maupun perusahaan, diantaranya:

1.1.3 Definisi SQL vs NoSQL

Sebenarnya perbedaan antara SQL dan NoSQL secara mendasar sudah dapat dijelaskan dari akronimnya.

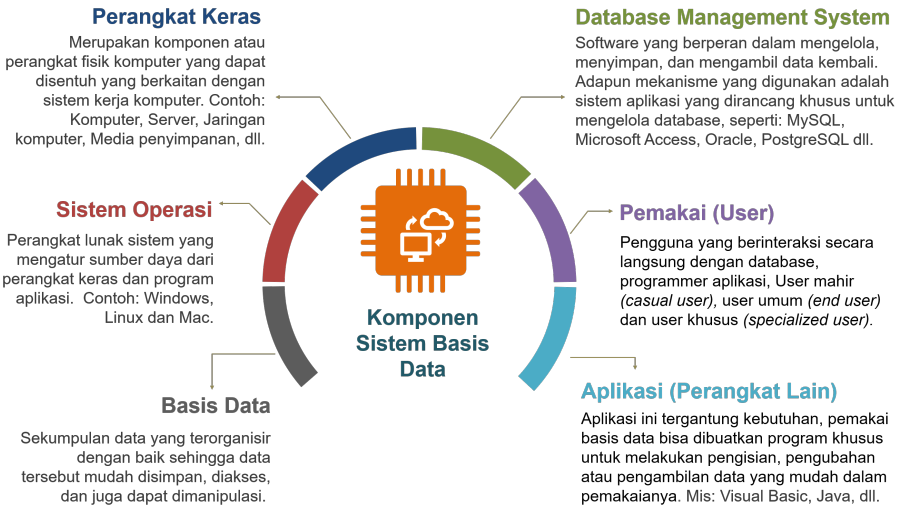


Figure 1.3: Komponen SBD

Figure 1.4: Manfaat SBD

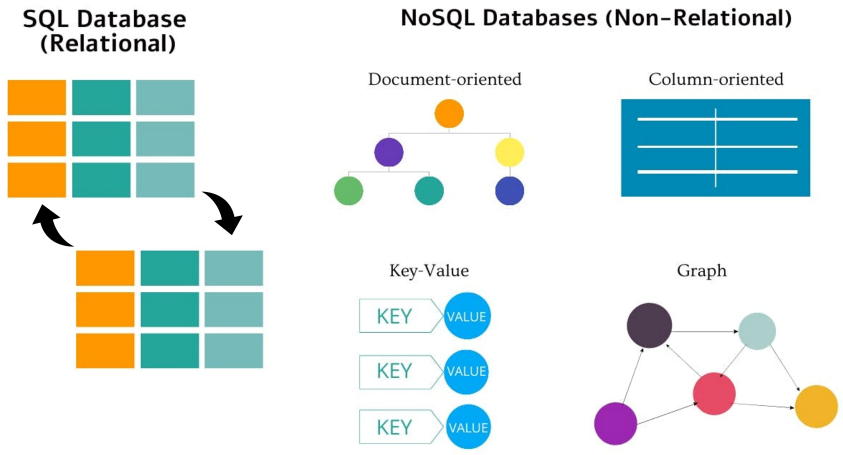


Figure 1.5: SQL vs NoSQL

SQL basis data relasional yang menggunakan 'relasi' (yang biasanya disebut tabel) untuk menyimpan data dan mencocokkan data tersebut dengan memakai karakteristik umum di setiap dataset. Sedangkan, *NoSQL* adalah database yang menggunakan format JSON untuk setiap dokumennya sehingga mudah dibaca dan dimengerti. NoSQL banyak diminati karena memiliki performa yang tinggi dan bersifat non-relasional sehingga dapat memakai berbagai model data.

1.1.4 Perbedaan SQL vs NoSQL

Sebenarnya banyak perbedaan yang dimiliki di antara dua database tersebut tapi inilah perbedaan yang paling mencolok antara SQL dan NoSQL:

1.1.5 Top 7 SQL

Tercatat sampai bulan Februari 2020 ada 334 jenis database menurut db-engines.com. Berikut ini saya merangkum daftar 7 database terpopuler yang menggunakan SQL (Relasional):

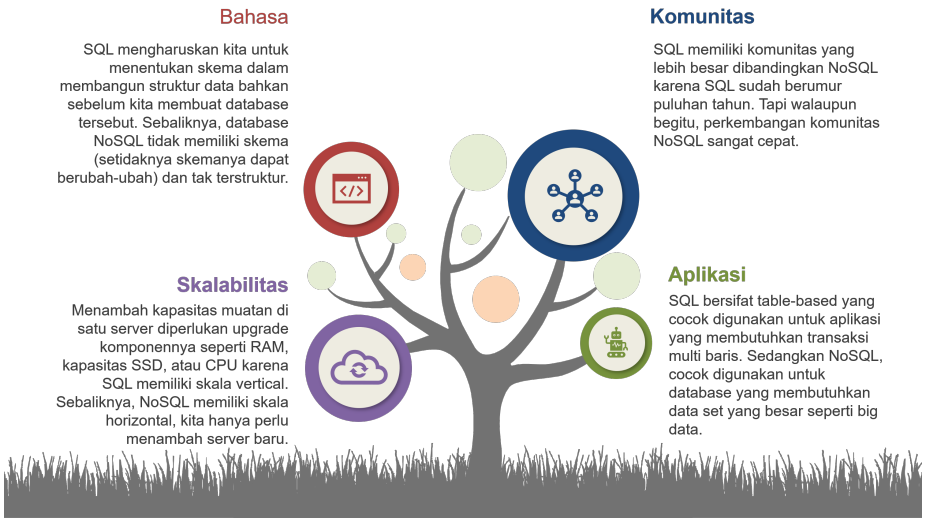


Figure 1.6: Perbedaan SQL vs NoSQL

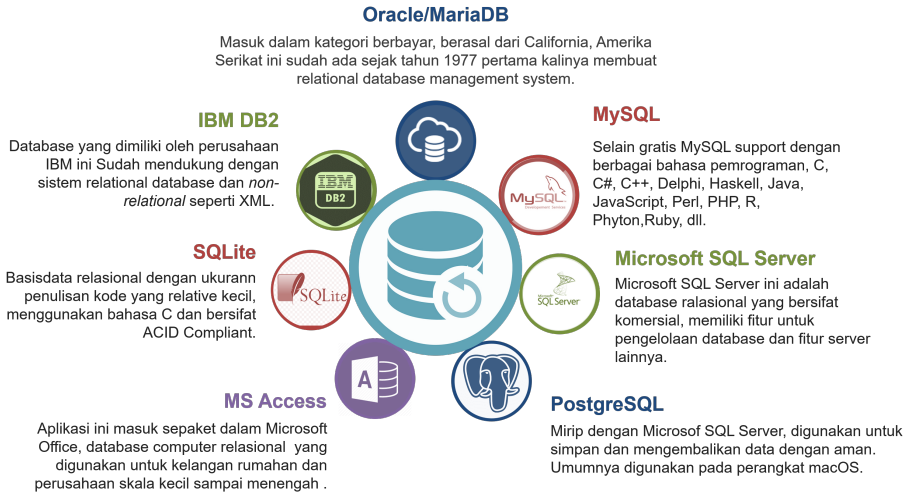


Figure 1.7: Top 7 Perangkat Lunak SQL

1.1.6 Top 8 NoSQL

Kebanyakan basis data NoSQL digunakan dalam dunia aplikasi web waktu nyata (real-time web app). Berikut ini adalah ulasan 8 jenis basis data NoSQL yang paling populer digunakan diseluruh dunia:

Figure 1.8: Top 8 Perangkat Lunak NoSQL

1.2 Mengapa R & SQL?

Menggunakan R dan SQL merupakan kombinasi yang kuat untuk analisis data dan pengelolaan basis data. Keduanya memiliki peran yang berbeda dalam proses analisis dan pengelolaan data. Berikut adalah beberapa alasan mengapa menggunakan R dan SQL bersama:

- **Kekuatan Analisis R**

R adalah bahasa pemrograman yang khusus dirancang untuk analisis statistik dan visualisasi data. R memiliki berbagai paket (packages) yang menawarkan fungsi statistik dan analisis yang kuat, termasuk regresi, pengelompokan, analisis deret waktu, dan banyak lagi. Visualisasi yang dapat dihasilkan dengan R sangat bervariasi, dari grafik sederhana hingga visualisasi interaktif yang kompleks.

- **Manipulasi dan Pengelolaan Data dengan SQL**

SQL digunakan untuk mengelola dan mengambil data dari basis data terstruktur. SQL menyediakan cara efisien untuk membuat, mengubah, menghapus, dan memanipulasi data dalam basis data. SQL memiliki fitur untuk menggabungkan data dari berbagai tabel, melakukan agregasi, dan menyaring data.

- **Integrasi Antara R dan SQL**

Banyak perpustakaan R yang mendukung koneksi ke basis data menggunakan SQL. Anda dapat menggunakan perintah SQL dalam skrip R untuk mengambil data dari basis data, memanipulasi data di dalam R, dan kemudian menerapkan analisis statistik menggunakan paket R. Integrasi ini memungkinkan Anda menggabungkan kekuatan analisis statistik R dengan kemampuan pengelolaan data SQL.

- **Skalabilitas dan Efisiensi**

Menggunakan SQL untuk mengambil dan memanipulasi data dalam basis data bisa lebih efisien daripada melakukannya dalam R, terutama untuk dataset besar. SQL memungkinkan query yang dioptimalkan dan penggunaan indeks untuk kinerja yang lebih baik.

- **Data Preprocessing**

Sebelum menerapkan analisis di R, Anda mungkin perlu melakukan prapemrosesan pada data, seperti membersihkan data, menggabungkan tabel, dan mengisi data yang hilang. SQL dapat membantu dalam melakukan tugas-tugas ini.

Jadi, menggunakan R dan SQL bersama memungkinkan Anda menggabungkan kekuatan analisis statistik R dengan kemampuan pengelolaan data SQL. Ini bisa sangat berguna ketika Anda ingin melakukan analisis data yang luas dan kompleks dari berbagai sumber data yang berbeda.

1.3 MySQL vs PostgreSQL

MySQL adalah sistem manajemen basis data relasional yang memungkinkan Anda untuk menyimpan data sebagai tabel dengan baris dan kolom. Sistem ini populer sehingga digunakan di banyak aplikasi web, situs web dinamis, dan sistem tertanam. PostgreSQL adalah sistem manajemen basis data relasional-objek yang menawarkan lebih banyak fitur daripada MySQL. Sistem ini memberi Anda lebih banyak fleksibilitas dalam tipe data, skalabilitas, konkurensi, dan integrasi data.

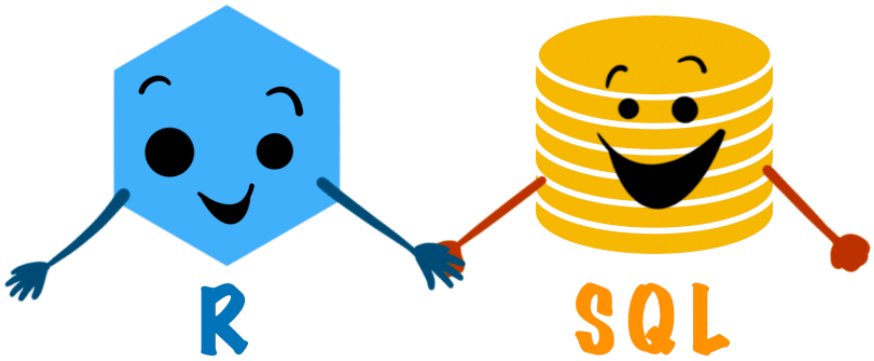


Figure 1.9: R dan SQL [https://irene.rbind.io](https://irene.rbind.io/post/using-sql-in-rstudio/)

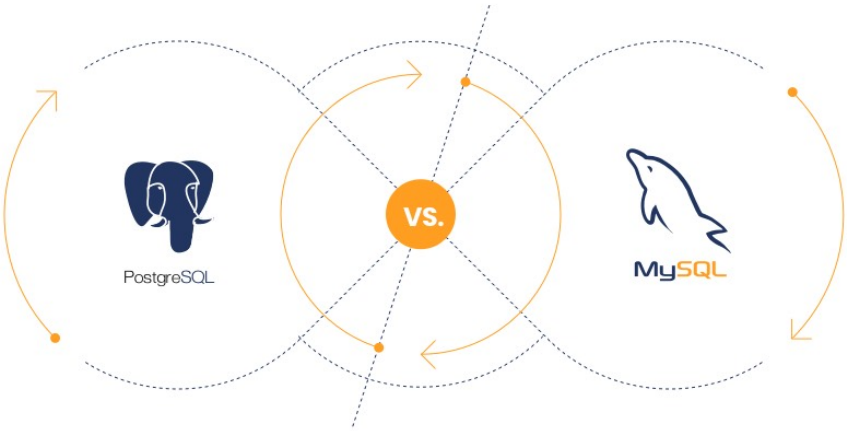


Figure 1.10: MySQL vs PostgreSQL [https://integrio.net/](https://integrio.net/blog/postgresql-vs-mysql)

MySQL dan PostgreSQL, Keduanya menyimpan data di dalam tabel yang terkait satu sama lain melalui nilai kolom umum. Namun keduanya sering dibandingkan karena terdapat beberapa perbedaan. Ingin mengenal lebih dalam? Simak penjelasan di bawah.

1.3.1 Kelebihan

MySQL	PostgreSQL
Integrasi bahasa pemrograman sangat luas;	Support framework website modern seperti Node.js dan Django; Support framework website modern seperti Node.js dan Django;
Aplikasi ringan, tidak membutuhkan spesifikasi hardware yang tinggi;	Dirilis dengan lisensi PostgreSQL sendiri;
Struktur tabel dengan fleksibilitas tinggi;	Bersifat open source dan gratis;
Dibekali banyak administrative tools;	Skala besar, mampu memuat hingga ribuan transaksi data;
Bersifat open source dan gratis (versi basic);	Memiliki banyak fitur yang mumpuni;
Meski open source, MySQL menjamin keamanan tingkat tinggi;	Memiliki banyak fitur yang mumpuni;
Mendukung berbagai variasi Data Type;	Performa sangat baik meski menuntut query yang lebih kompleks;
Dapat digunakan banyak pengguna karena mendukung multi user.	Kecepatan analisis (read-write) sangat cepat; Keamanan yang lebih ketat.

1.3.2 Kekurangan

MySQL	PostgreSQL
Sistem manajemen database kurang cocok untuk aplikasi mobile dan game;	PostgreSQL tidak mendukung semua stack development;
Technical support MySQL dinilai kurang baik;	Meski memiliki integrasi dan skalabilitas tinggi, kecepatan PostgreSQL kalah unggul dibandingkan RDBMS lain;

MySQL	PostgreSQL
Sulit diaplikasikan untuk manajemen database berskala besar.	Sistem kompatibilitas PostgreSQL menuntut pengguna untuk bekerja lebih keras dalam perbaikan dan perawatan.

1.4 Instalasi MySQL (XAMPP)

1.4.1 Download Aplikasi XAMPP

Silakan klik disini untuk mengunduh applikasi XAMPP, pilih salah satu saja sesuai Operating System pada Komputer anda.

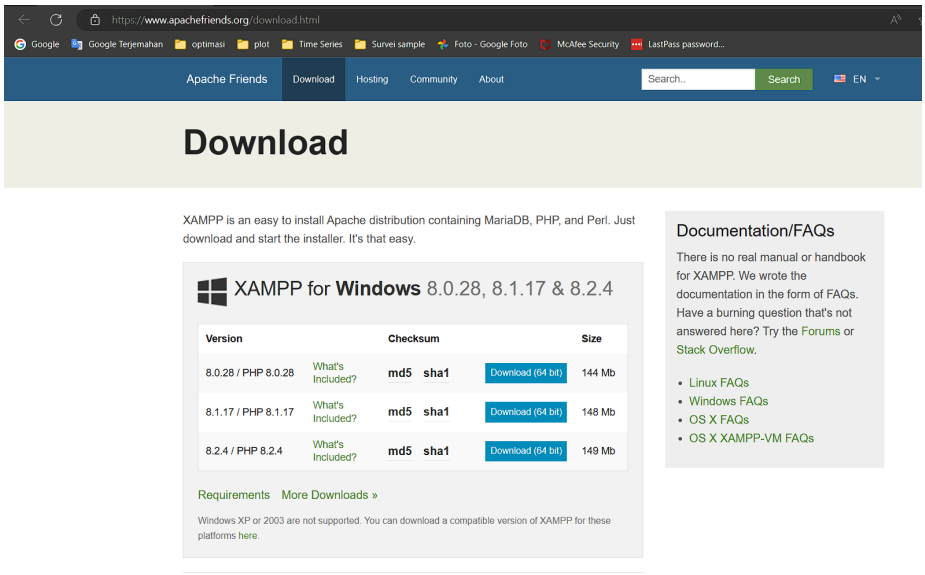


Figure 1.11: Langkah 1, Download XAMPP)

1.4.2 Install Aplikasi

Temukan file XAMPP.exe yang telah anda download, secara default biasanya disimpan di;

Selanjutnya, akan muncul Warning di klik **OK**

selajutunya klik next

Klik next lagi, karena sudah dipilih secara default oleh XAMPP

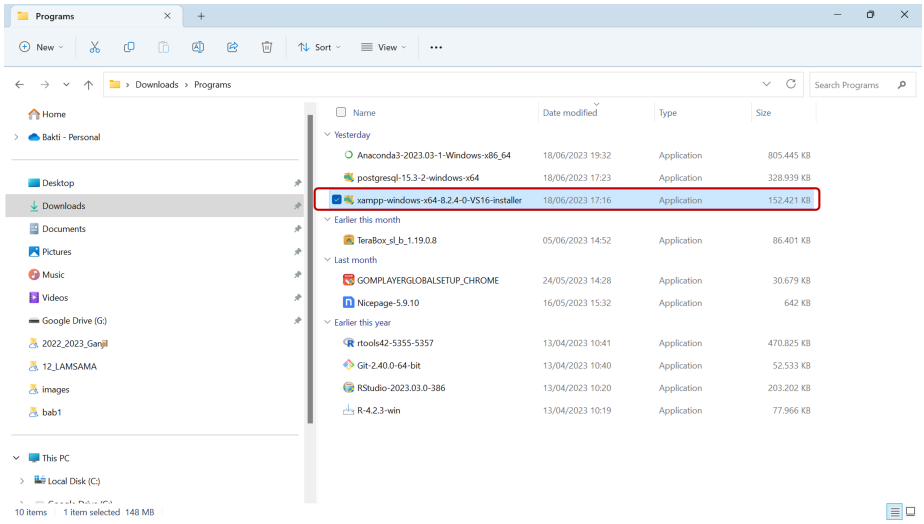


Figure 1.12: Langkah 2, Instalasi XAMPP)

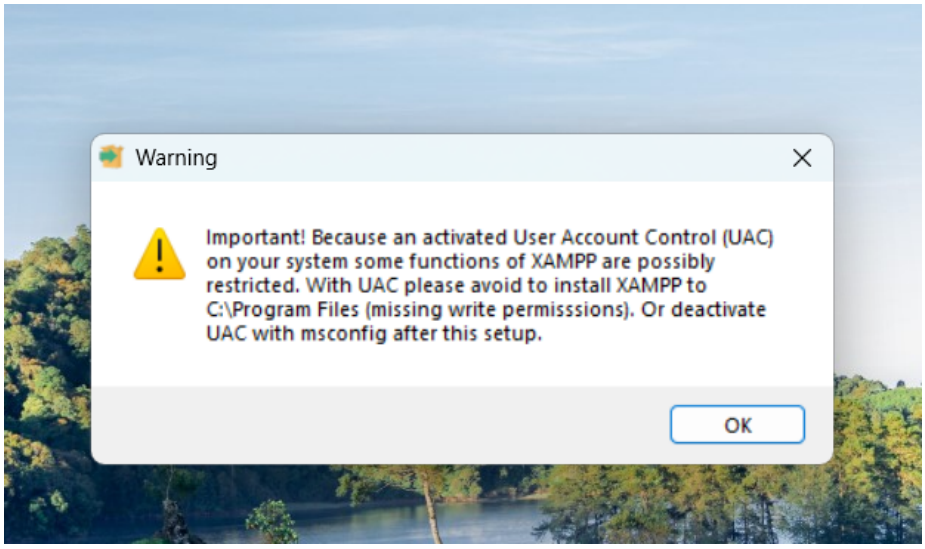


Figure 1.13: Langkah 3, Instalasi XAMPP)

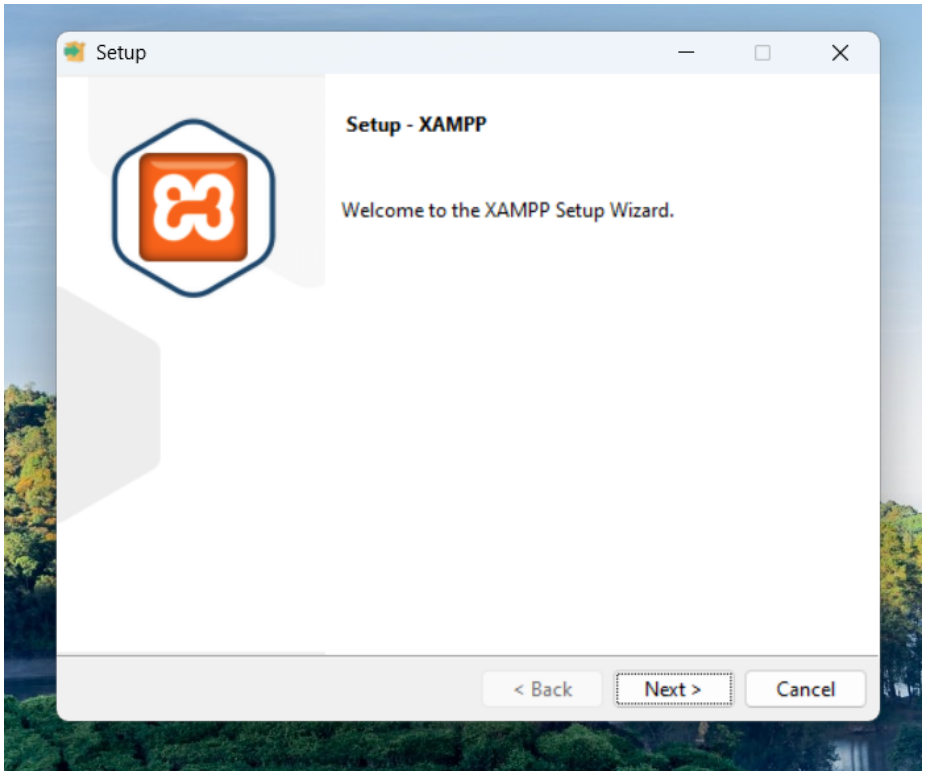


Figure 1.14: Langkah 4: Instalasi XAMPP)

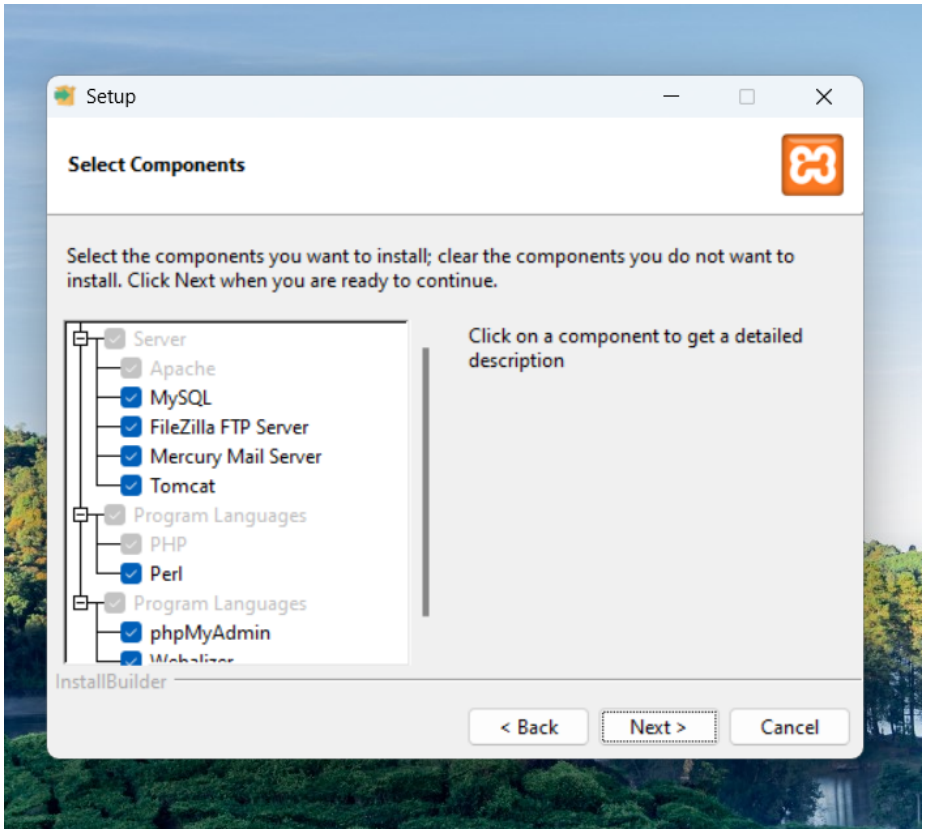


Figure 1.15: Langkah 5, Instalasi XAMPP)

1.4.3 Pilih Folder

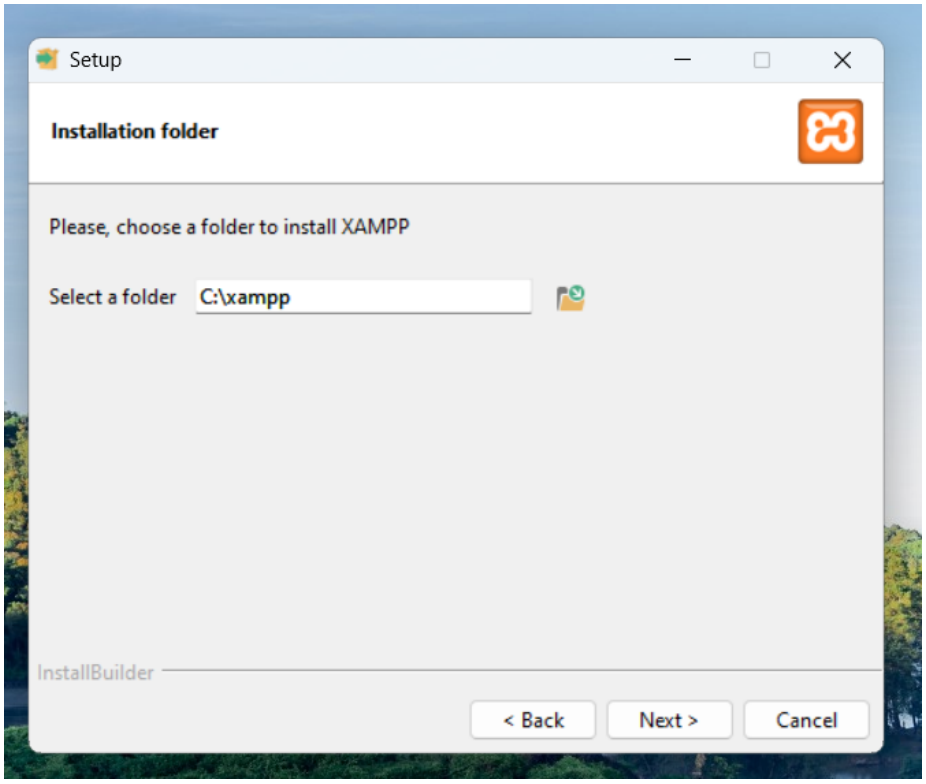


Figure 1.16: Langkah 6, Instalasi XAMPP)

Secara default akan membuat folder baru **C:\xampp**, lalu pilih next.

note: jika anda sudah pernah mendownload aplikasi xampp, perlu di hapus terlebih dahulu file xampp yang lama di file **C:\xampp**

1.4.4 Jalankan proses Instalasi

Tunggu proses instalasi selesai **Biasanya 5-10 menit, tergantung kecepatan komputer anda**.

1.4.5 XAMPP sudah terinstall

Setelah aplikasi terinstall, sudah bisa digunakan.

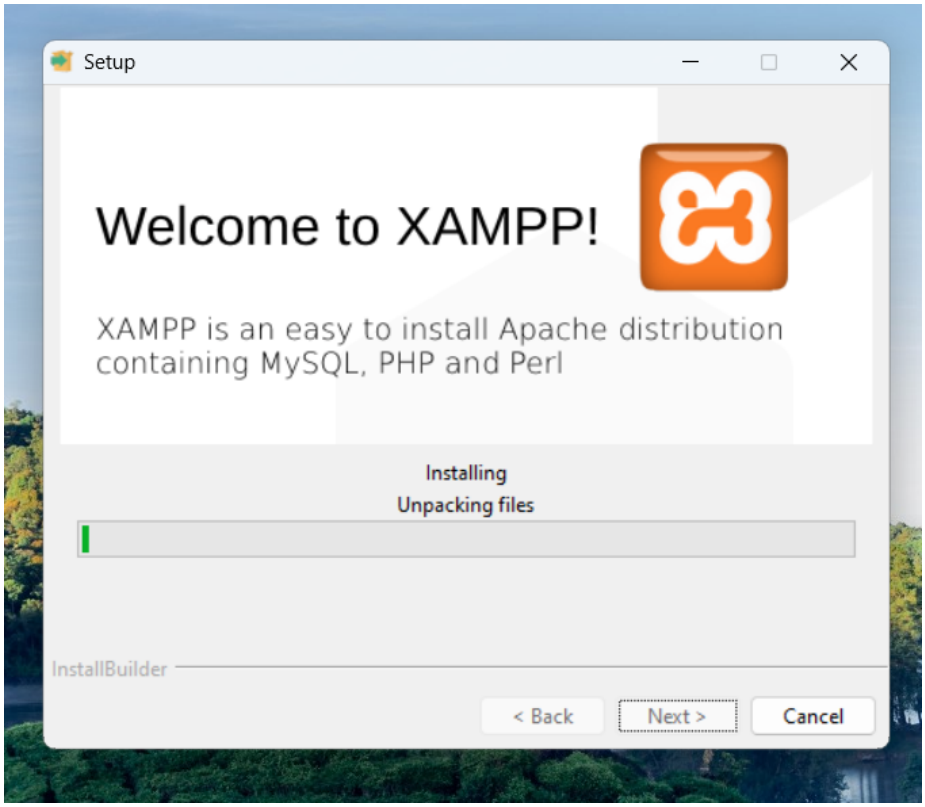


Figure 1.17: Langkah 7, Instalasi XAMPP)

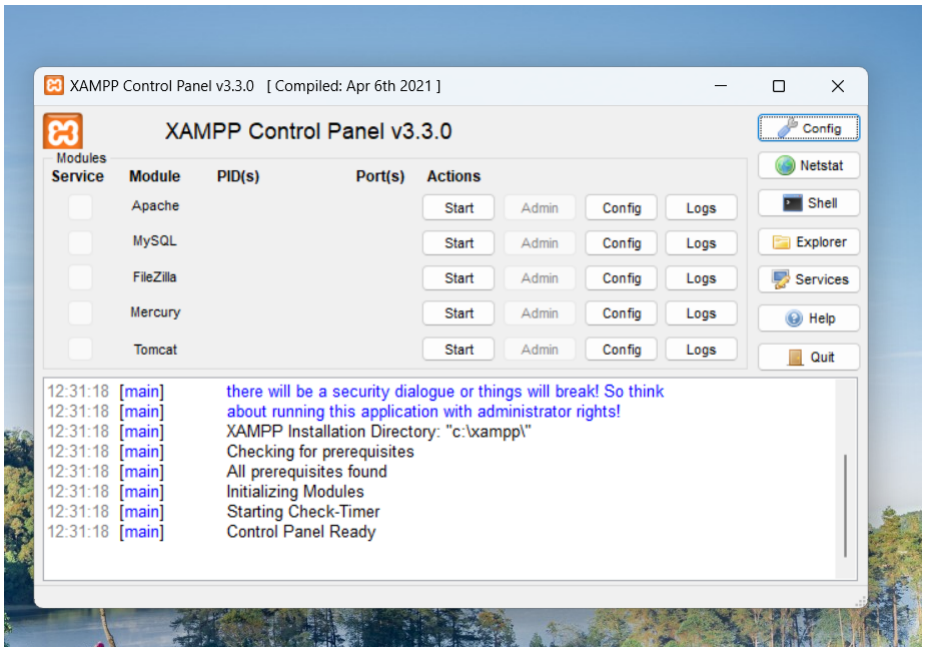


Figure 1.18: Langkah 8, Instalasi XAMPP)

1.4.6 Video Instalasi XAMPP

1.5 Instalasi PostgreSQL

Berikut ini adalah proses langkah demi langkah tentang Cara Menginstal PostgreSQL di Windows:

1.5.1 Buka Browser

Klik https://www.postgresql.org/download and pilih Windows

1.5.2 Cek Option

Klik Download the installer Interactive Installer by EnterpriseDB

1.5.3 Pilih PostgreSQL version

Anda akan diminta untuk memilih versi PostgreSQL dan sistem operasi yang diinginkan. Pilih versi PostgreSQL terbaru dan OS sesuai dengan environment

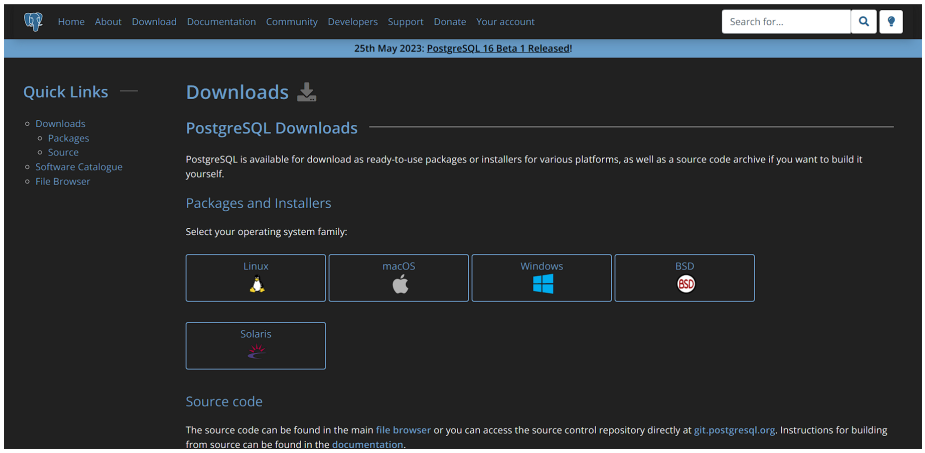


Figure 1.19: Langkah 1, Instalasi PostgreeSQL)

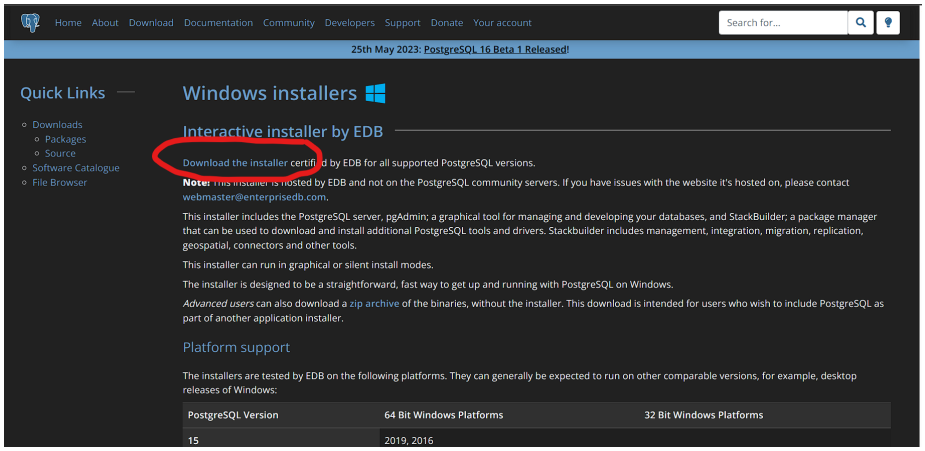


Figure 1.20: Langkah 2, Instalasi PostgreeSQL)

Anda, **klik tombol download.**

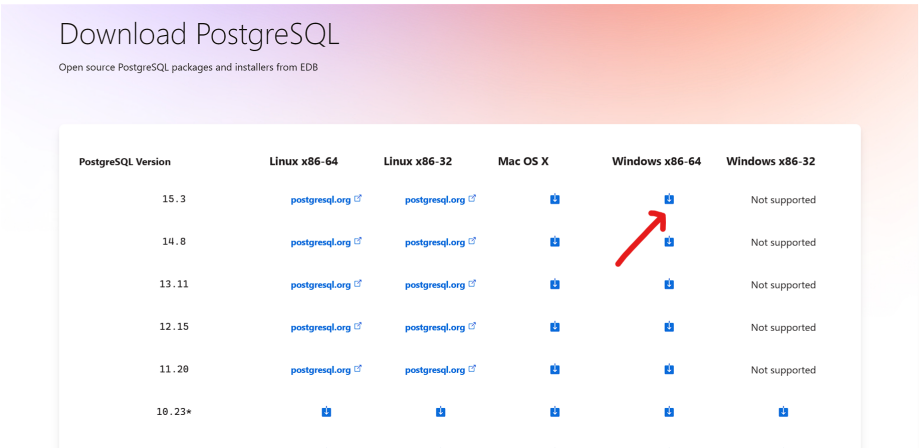


Figure 1.21: Langkah 3, Instalasi PostgreeSQL)

1.5.4 Open exe file

Setelah Anda mengunduh PostgreSQL, buka exe yang telah diunduh dan Klik berikutnya pada layar install welcome screen.

1.5.5 Pilih folder

Ubah direktori Instalasi jika diperlukan, jika tidak, biarkan default, **klik Next.**

1.5.6 Select components

Anda dapat memilih komponen yang ingin Anda instal di sistem Anda. Anda dapat menghapus centang pada Stack Builder (*disarankan ikuti secara default*), **klik Next.**

1.5.7 Check data location

Anda dapat mengubah lokasi data, **Klik Next.**

1.5.8 Masukan Password

Masukkan kata sandi superuser. Catat kata sandi tersebut, **Klik Next.**

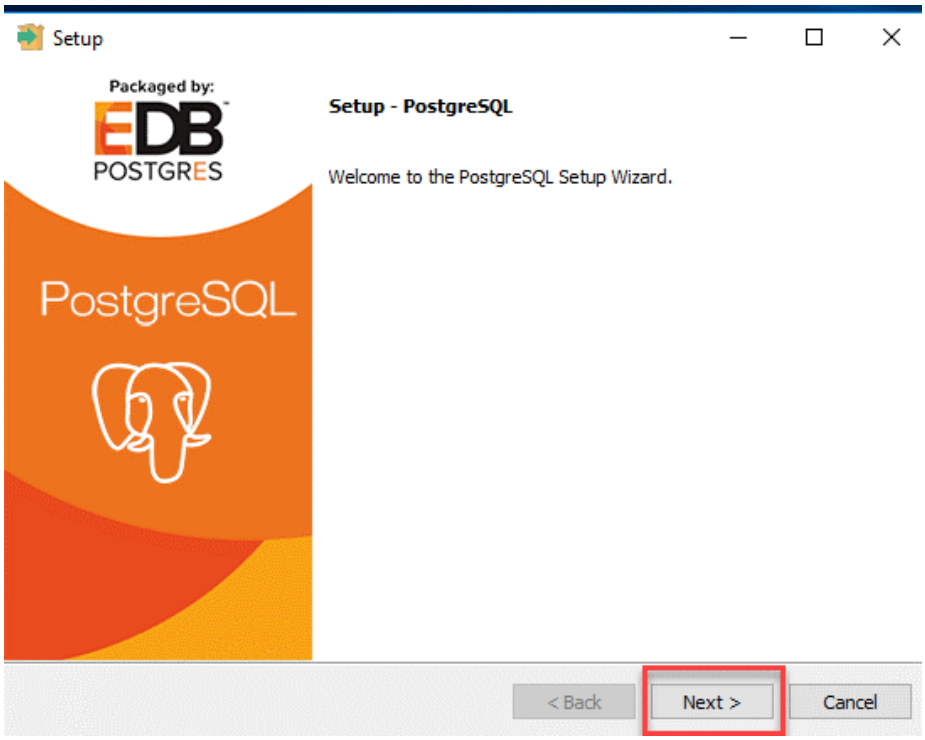


Figure 1.22: Langkah 4, Instalasi PostgreeSQL)

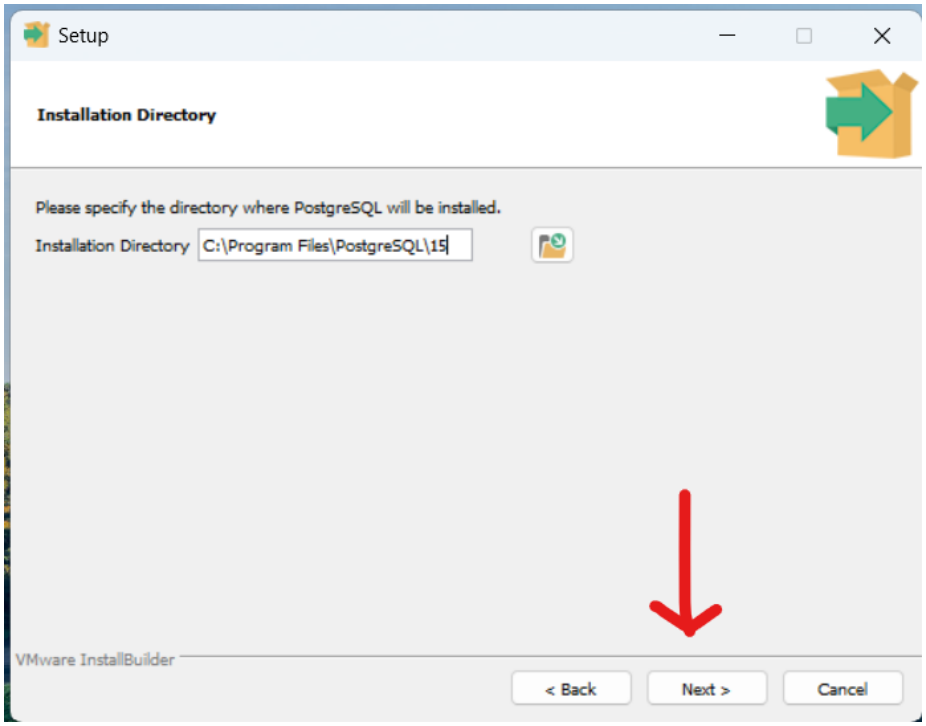


Figure 1.23: Langkah 5, Instalasi PostgreeSQL)

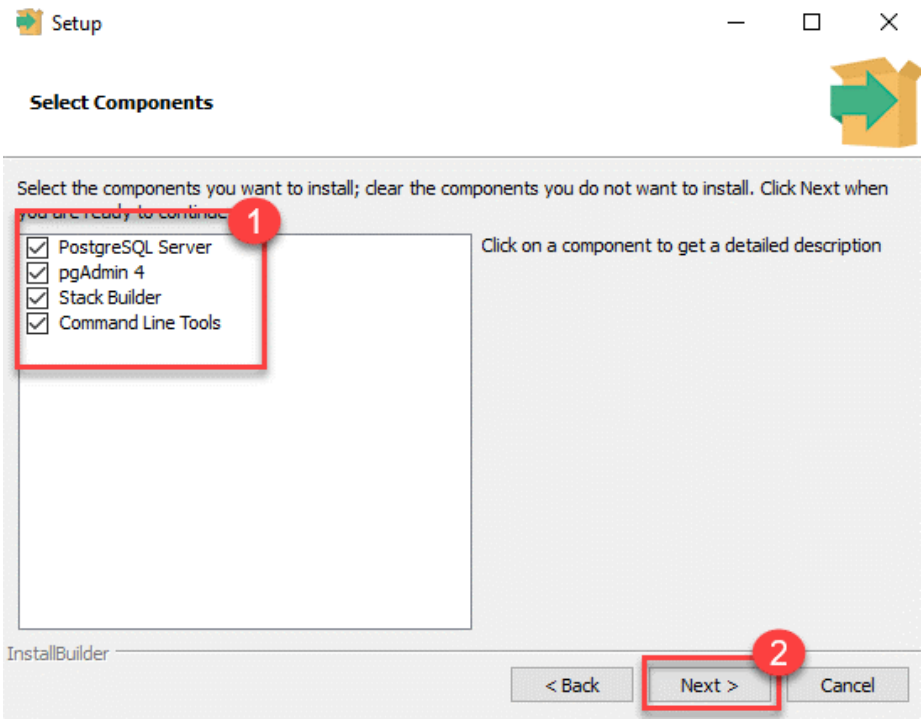


Figure 1.24: Langkah 6, Instalasi PostgreeSQL)

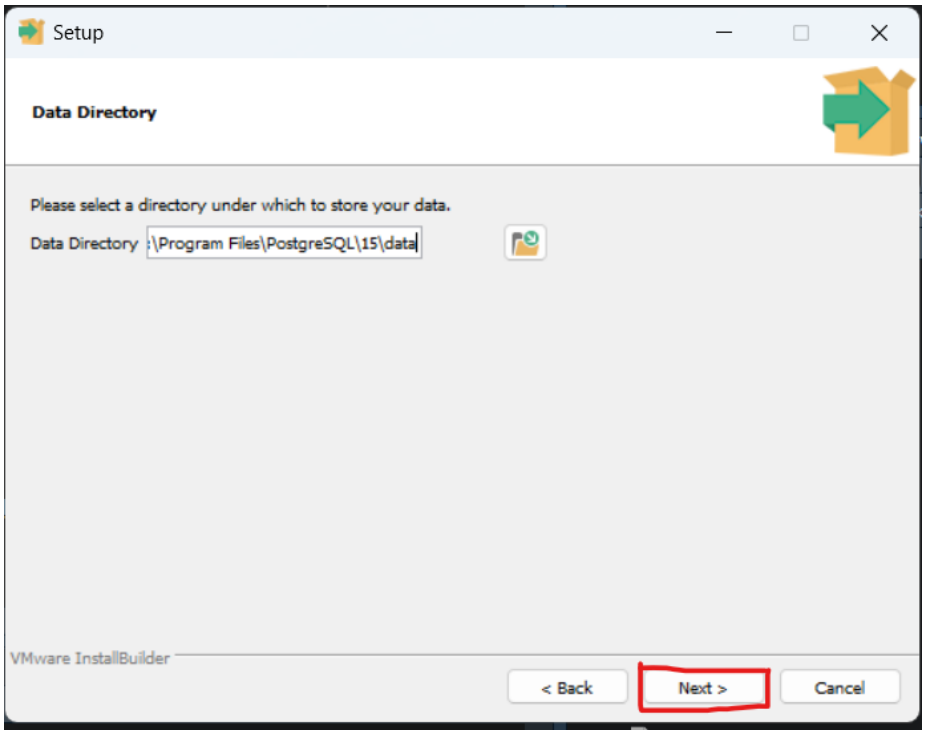


Figure 1.25: Langkah 7, Instalasi PostgreeSQL)

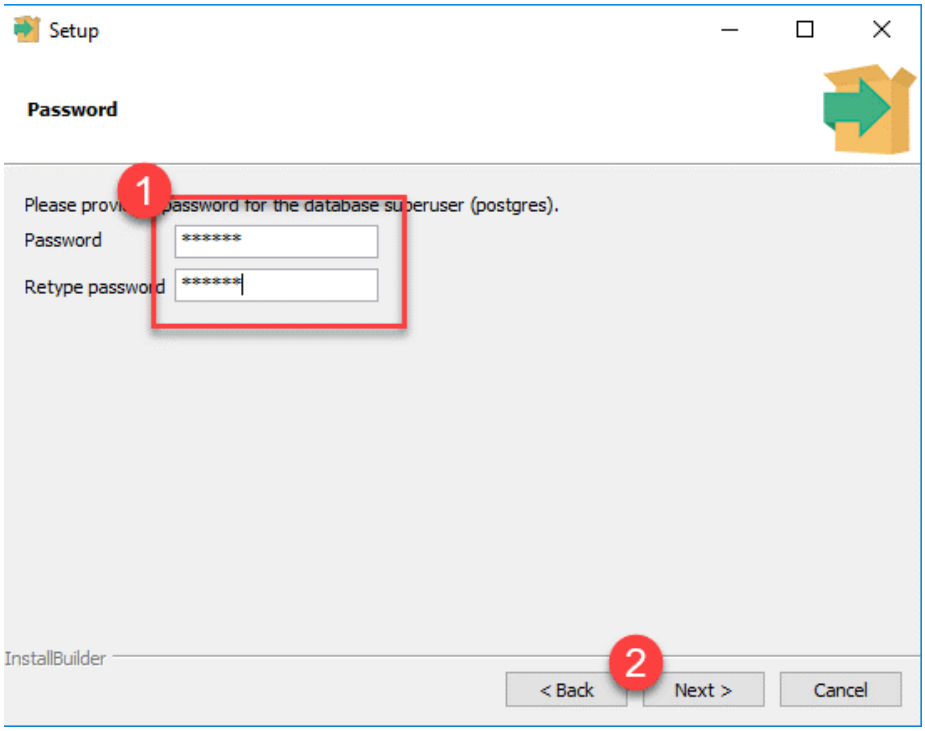


Figure 1.26: Langkah 8, Instalasi PostgreeSQL)

1.5.9 Cek opsi port

Biarkan nomor port menjadi default, **Klik Next.**

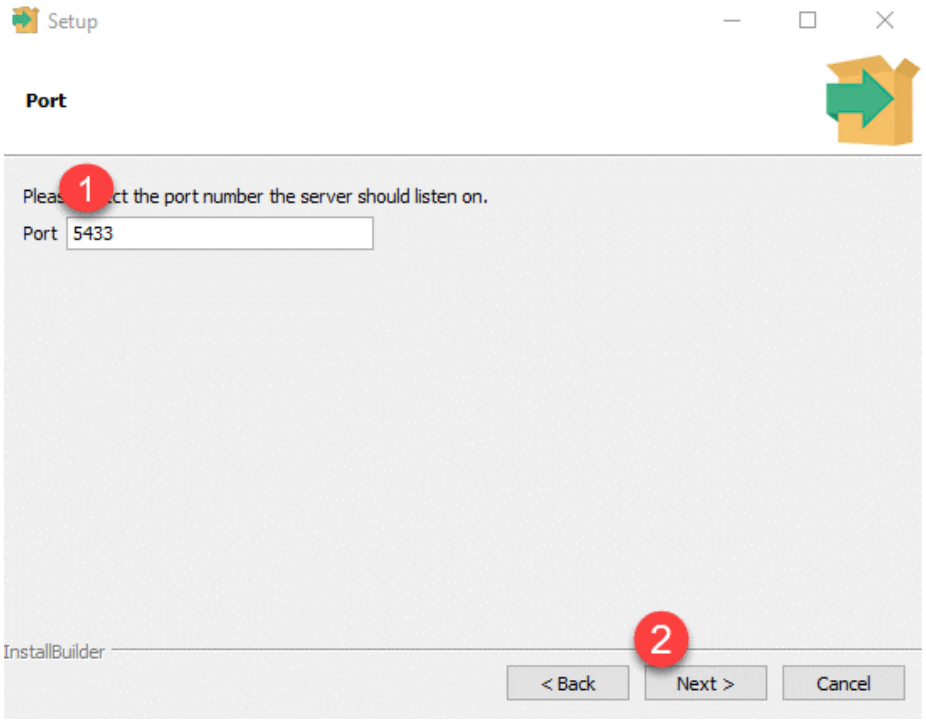


Figure 1.27: Langkah 9, Instalasi PostgreeSQL)

1.5.10 Cek Summary

Periksa pra-penginstalan summary, **Klik Next**

1.5.11 Ready to Install

Klik tombol Next

1.5.12 Check stack builder prompt

Setelah instalasi selesai, Anda akan melihat prompt Stack Builder. Hapus centang pada opsi tersebut. Kita akan menggunakan Stack Builder dalam tutorial selanjutnya, **Klik Finish.**

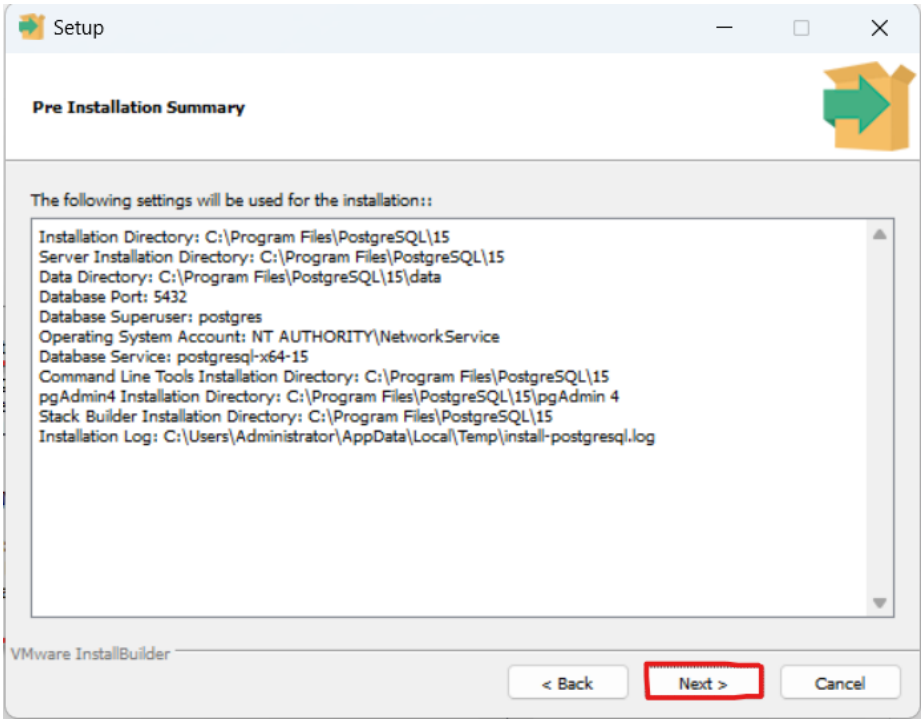


Figure 1.28: Langkah 10, Instalasi PostgreeSQL)

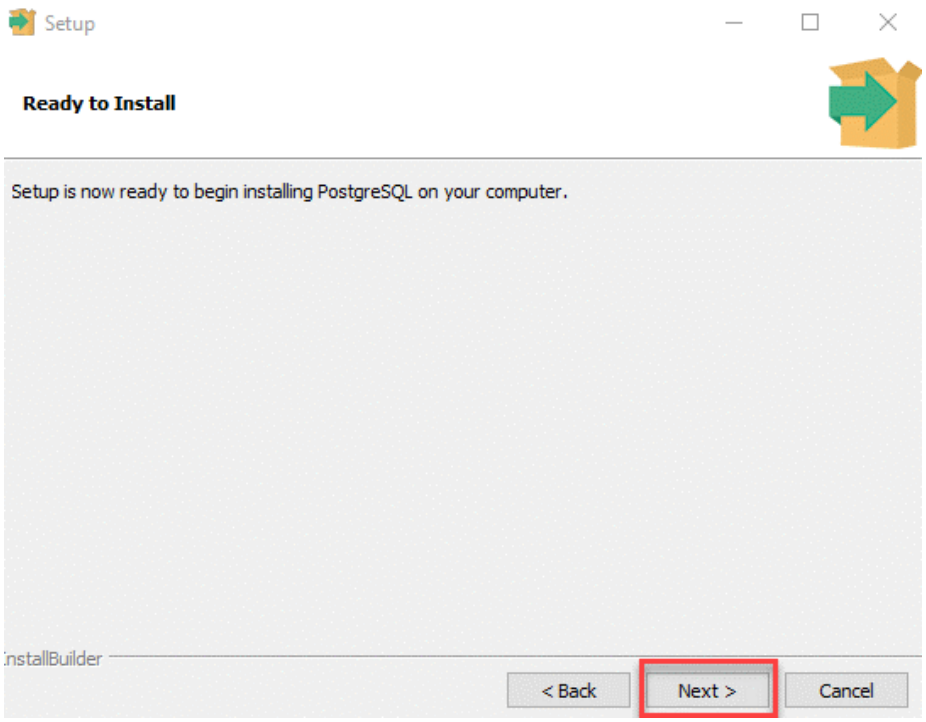


Figure 1.29: Langkah 11, Instalasi PostgreeSQL)

Figure 1.30: Langkah 12, Instalasi PostgreeSQL)

1.5.13 Launch PostgreSQL

Untuk launch PostgreSQL, buka Start Menu dan cari pgAdmin 4

1.5.14 Check pgAdmin

Anda akan melihat beranda pgAdmin

1.5.15 Cari PostgreSQL 15

Klik pada Servers > PostgreSQL 15 di sub sebelah kiri

1.5.16 Enter password

Masukkan kata sandi superuser yang ditetapkan selama instalasi, **Klik OK**

1.5.17 Cek Dashboard

Anda akan melihat Dashboard

1.5.18 Video Instalasi PostgreSQL

1.6 Praktikum

- Tutorial di MySQL (CREATE & Drop Database, Create & Drop Tabel)
- Tutorial di PostgreeeSQL (CREATE & Drop Database, Create & Drop Tabel)

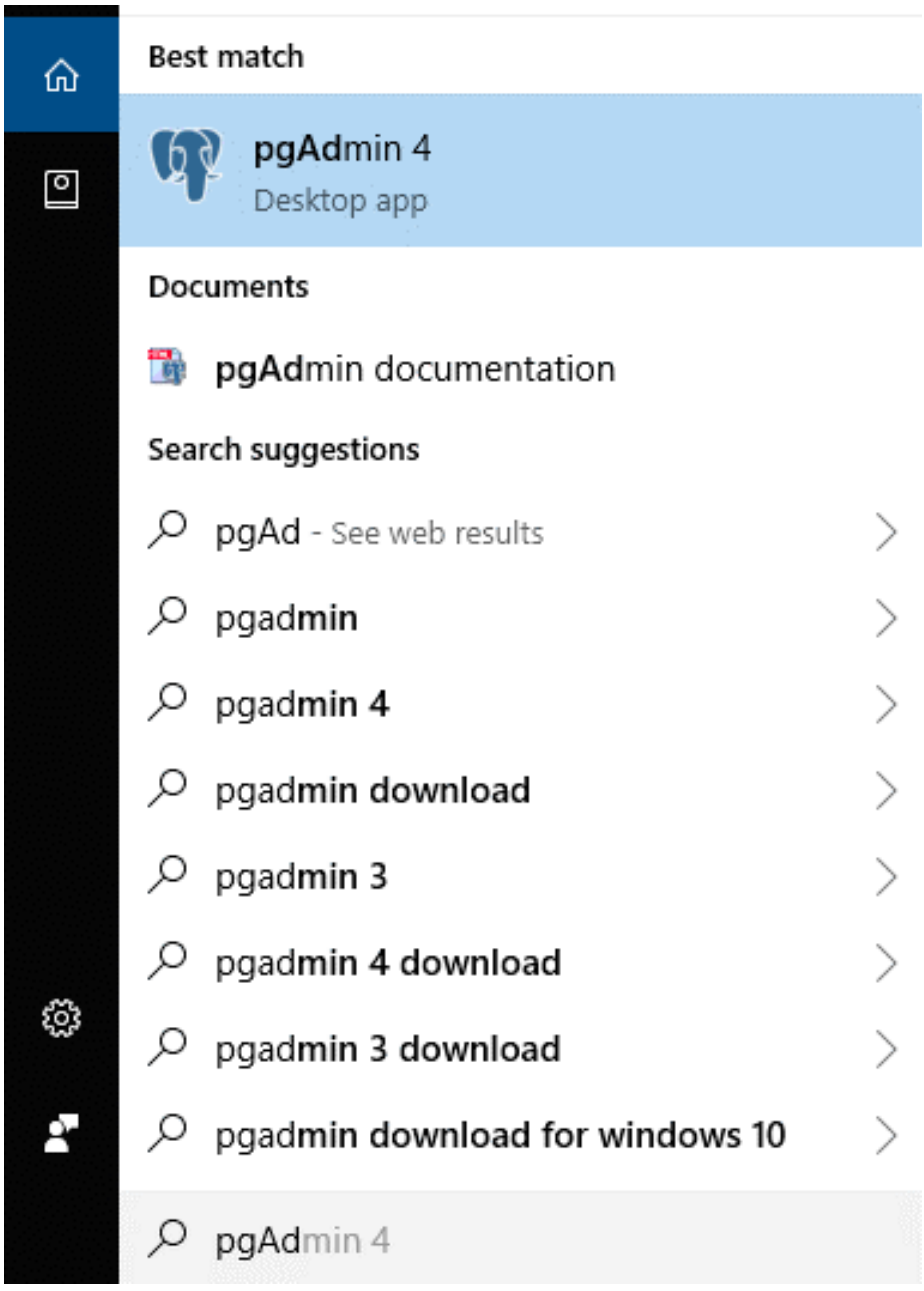


Figure 1.31: Langkah 13, Instalasi PostgreeSQL)

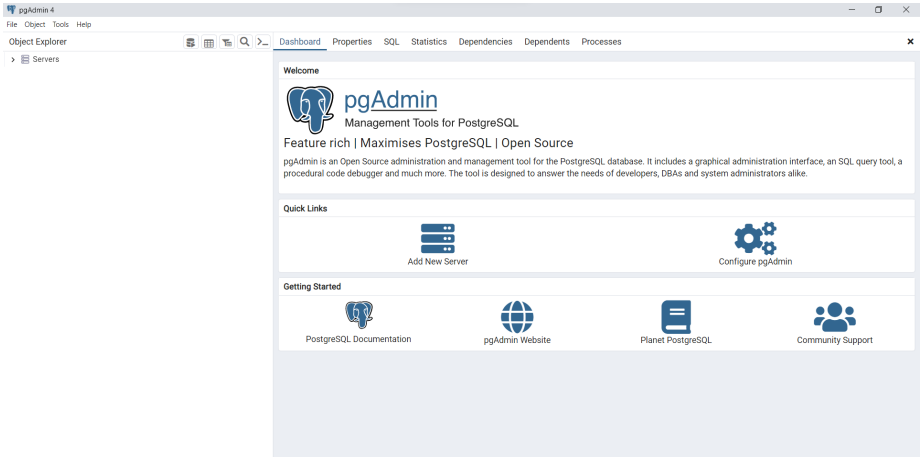


Figure 1.32: Langkah 14, Instalasi PostgreeSQL)

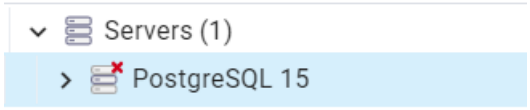


Figure 1.33: Langkah 15, Instalasi PostgreeSQL)

Figure 1.34: Langkah 16, Instalasi PostgreeSQL)

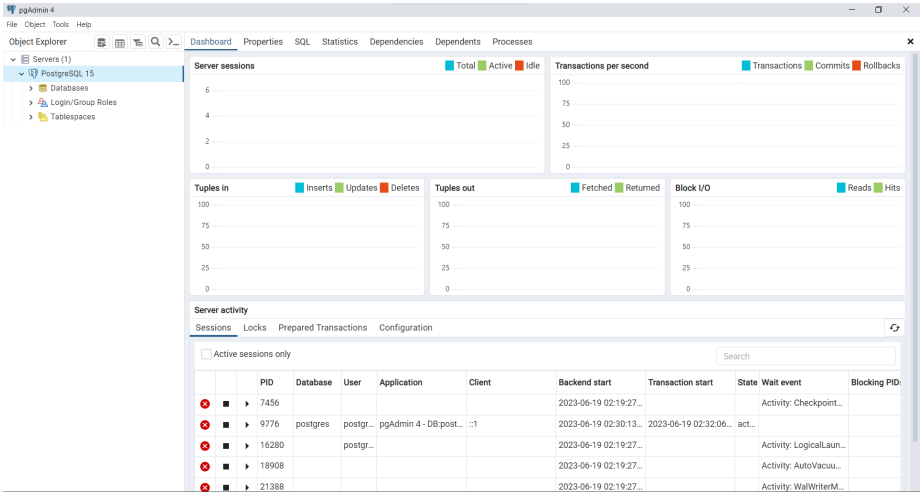


Figure 1.35: Langkah 17, Instalasi PostgreeSQL)

Chapter 2

Connecting R to SQL

2.1 Introduction

A database is a structured set of data. Terminology is a little bit different when working with a database management system compared to working with data in R.

- **field:** variable or quantity
- **record:** collection of fields
- **table:** collection of records with all the same fields
- **database:** collection of tables

The relationship between R terminology and database terminology is explained below.

R terminology	Database terminology
column	field
row	record
data frame	table
types of columns	table schema
collection of data	frames database

2.2 Connecting R to SQL

Connecting R to SQL databases allows you to leverage the power of R for data analysis while directly interacting with and querying data stored in relational

databases. This connection enables you to retrieve, manipulate, and analyze data using SQL queries within your R environment.

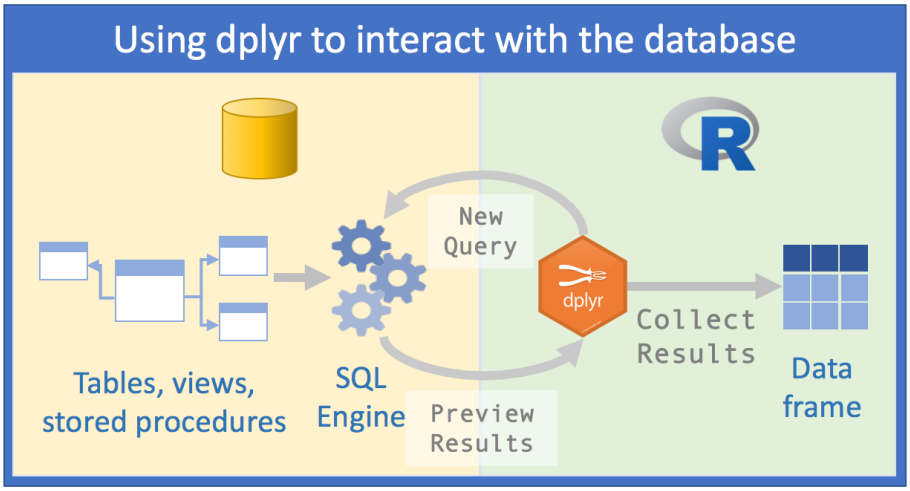


Figure 2.1: Connecting R to SQL [https://rviews.rstudio.com](https://rviews.rstudio.com/2017/05/17/d using-r/)

Here's a step-by-step introduction to connecting R to an SQL database:

2.2.1 Install Required Packages

First, you need to install R packages that facilitate database connections and SQL interactions. The DBI package provides a common interface for database connections, and you'll also need a database-specific package like RMySQL for MySQL, RPostgreSQL for PostgreSQL, or RODBC for ODBC connections.

You can install these packages using the following commands:

```
install.packages(c(
                   "RMariaDB",
                   "RMySQL",
                   "RPostgres"
                   "RSQLite",
                   )
                 )
```

2.2.2 Load Packages

Then, load all these requirement packages:

```
# install.packages("pacman")
pacman::p_load(
               RMariaDB,             # Database Interface and 'MariaDB' Driver
               RMySQL,               # Database Interface and 'RMySQL' Driver
               RPostgres             # Database Interface and 'RPostgres' Driver
               RSQLite,              # Database Interface and 'RSQLite' Driver
          )
```

2.2.3 Establish a Connection

There are many ways to connect your database with R. This article shows you three of the most common ways:

MariaDB

```
MariaDB <- dbConnect(RMariaDB::MariaDB(),
                  user='root',
                  password='',
                  dbname='bakti',
                  host='localhost')
dbListTables(MariaDB)                                # table list on your database
dbExecute(MariaDB,"CREATE DATABASE new_MariaDB")  # Create a new Database
dbExecute(MariaDB,"DROP DATABASE new_MariaDB")    # Drop a Database
```

MySQL

```
MySQL <- dbConnect(MySQL(),
                  user='root',
                  password='',
                  dbname='bakti',
                  host='localhost')
dbListTables(MySQL)                                # table list on your database
dbExecute(MySQL,"CREATE DATABASE new_MySQL")       # Create a new Database
dbExecute(MySQL,"DROP DATABASE new_MySQL")         # Drop a Database
```

Postgres

```
postgres <- dbConnect(RPostgres::Postgres(),
                      user='postgres',
                      password='1234',
                      dbname='postgres',
                      host='localhost')
dbListTables(postgres)                                  # table list on your database
dbExecute(postgres,"CREATE DATABASE new_MySQL")         # Create a new Database
dbExecute(postgres,"DROP DATABASE new_MySQL")           # Drop a Database
```

SQLite

```
RSQLite <- dbConnect(RSQLite::SQLite(), "folder_db/mydb3.sqlite")
dbListTables(RSQLite)                                 # table list on your database
```

- *Notes:* RSQLite will store the database you created in your current working directory.

2.3 Import Data

This section can be ignored if the data (table) that you need is already registered in your database. If not, then it is necessary to import data set according to your available files, download it below:

- Customers.csv
- Categories.csv
- Employees.csv
- OrderDetails.csv

- Orders.csv
- Products.csv
- Shippers.csv
- Suppliers.csv
- RawDatabase.xlsx

2.3.1 CSV Files

When you're working with files in R, such as reading data from a CSV file or saving plots as image files, R needs to know the location of these files. By setting the working directory, you provide a starting point for R to look for and save files.

In R, the setwd() function is used to set the working directory for your R session. The working directory is the folder on your computer where R will look for files and where it will save files unless you specify a different location. Here's why the setwd() function is important and when you might use it:

```
# Set the working directory
setwd("/path/to/your/folder")
```

```
# Now you can read CSV files without specifying the full path
data <- read.csv("file.csv")
```

Replace "/path/to/your/folder" with the actual path to the folder containing your CSV files. Then, you can run the following code!.

```
Customers   <-read.csv("data/Customers.csv")
Categories  <-read.csv("data/Categories.csv")
Employees   <-read.csv("data/Employees.csv")
OrderDetails<-read.csv("data/OrderDetails.csv")
Orders      <-read.csv("data/Orders.csv")
Products    <-read.csv("data/Products.csv")
Shippers    <-read.csv("data/Shippers.csv")
Suppliers   <-read.csv("data/Suppliers.csv")
```

2.3.2 XLSX Files

```
library("readxl")
Customers   <-read_excel("data/RawDatabase.xlsx",sheet=1)
Categories  <-read_excel("data/RawDatabase.xlsx",sheet=2)
Employees   <-read_excel("data/RawDatabase.xlsx",sheet=3)
OrderDetails<-read_excel("data/RawDatabase.xlsx",sheet=4)
Orders      <-read_excel("data/RawDatabase.xlsx",sheet=5)
Products    <-read_excel("data/RawDatabase.xlsx",sheet=6)
Shippers    <-read_excel("data/RawDatabase.xlsx",sheet=7)
Suppliers   <-read_excel("data/RawDatabase.xlsx",sheet=8)
```

2.4 Write Dataframe to Database

The key here is the `dbWriteTable` function which allows us to write an R data frame directly to a database table. The data frame's column names will be used as the database table's fields. In the following example I use `RMariaDB` connection, you can apply another driver as you like.

```
new_con <- dbConnect(MariaDB(),
                     user='root',
                     password='',
                     dbname='new_MariaDB',
                     host='localhost')

dbWriteTable(new_con, "Customers", Customers, append=T)
dbWriteTable(new_con, "Categories", Categories, append=T)
dbWriteTable(new_con, "Employees", Employees, append=T)
dbWriteTable(new_con, "OrderDetails", OrderDetails, append=T)
dbWriteTable(new_con, "Orders", Orders, append=T)
dbWriteTable(new_con, "Products", Products, append=T)
dbWriteTable(new_con, "Shippers", Shippers, append=T)
dbWriteTable(new_con, "Suppliers", Suppliers, append=T)
```

Note: Some important things that must be considered when storing table data are as follows:

- Data Structure adjustments
- Changes Data type (especially, Date and Time)

In this case, we have a problem with the data table `Employees` and `Orders`. When you consider these Table (Employees and Orders), you will find there is no date are written correctly in the database. In order to handle this problem, just type the following code in your R console:

```
dbRemoveTable(new_con, "Orders")
```

```
Orders["OrderDate"] <-as.Date(Orders$OrderDate, format = "%Y-%m-%d")
```

```
dbWriteTable(new_con, "Orders", Orders, append=T)
```

Your Exercise: Do the same thing to update data table `Employees`

2.5 Basic SQL in R

2.5.1 SELECT

The SELECT statement is used to select data from a database.

```
library(DT)
df1<-dbGetQuery(new_con,'SELECT city
                         FROM Customers')
datatable(df1)
```

2.5.2 WHERE

The WHERE clause is used to filter records, extract only those records that fulfill a specified condition.

```
df2<-dbGetQuery(new_con,"SELECT *
                         FROM Customers
                         WHERE Country='Germany'")
datatable(df2)
```

2.5.3 INSERT INTO

If you are adding values for all the columns of the table, you do not need to specify the column names in the SQL query. However, make sure the order of the values is in the same order as the columns in the table. The INSERT INTO syntax would be as follows:

```
dbExecute(new_con,"INSERT INTO Customers(CustomerName,ContactName,Address,City,PostalCode, Countr
                   VALUES('Bakti','Siregar','Jl.Bahagia Selalu','Tangerang','081369','Indonesia')
```

2.5.4 DELETE

The DELETE statement is used to delete existing records in a table.

```
dbExecute(new_con,"DELETE FROM Customers
                   WHERE CustomerName ='Bakti' ")
```

2.5.5 UPDATE

The UPDATE statement is used to modify the existing records in a table.

```
dbExecute(new_con,"UPDATE Customers
                   SET ContactName = 'Alfred Schmidt', City= 'Frankfurt'
                   WHERE CustomerID = 1")
```

2.5.6 Disconnect Database

If you are done with the query process and you don't want to use it anymore, you should disconnect the connection from your database.

```
dbDisconnect(new_con)                                    # disconnect from your da
```

2.6 Your Job

1. Write Dataframe to your Database directory, using the following Engine:

- RMySQL
- RPostgres

- RSQLite

2. Create a 'Basic Queries' tutorial using all engine that you already use at task no 1, (Such as: `SELECT`, `WHERE`, `INSERT INTO`, `DELETE`, and `UPDATE`)!

Chapter 3

Fundamental SQL in R

In the previous section, we learned how to connect R to a Database System (SQL) Such as RMariaDB, RMySQL, and RSQLite. In this section, we continue to cover all that you have to know about fundamental operations in SQL (Here, focus on RMySQL).

3.1 Connecting R to MySQL

Connecting R to MySQL is made very easy with the `RMySQL` package. To connect to a MySQL database simply install the package and load the library.

```
library(RMySQL)
MySQL <- dbConnect(MySQL(),
                   user='root',
                   password='',
                   dbname='mysql',
                   host='localhost')
dbListTables(MySQL)                    # a list of the tables in our connection
```

Note: Open and your XAMPP, click start on Apache and MySQL. Then, make sure you have the admin privilege before creating any database.

3.2 Create DB

If you want to create a new database, then the CREATE DATABASE statement would be as shown below:

```
dbExecute(MySQL, "CREATE DATABASE factory_db")
```

The result show us 1, means that you have succeeded to create a database.

3.3 Drop DB

If you want to delete an existing database, then the DROP DATABASE statement would be as shown below:

```
dbExecute(MySQL, "DROP DATABASE factory_db")
```

The result show us 0, means that you have succeeded to remove (Drop) a database.

3.4 Create Table

Once you have a database, you can continue to create table as shown below:

```
dbExecute(MySQL, "CREATE TABLE Persons(
                  PersonID int,
                  LastName varchar(255),
                  FirstName varchar(255),
                  Address varchar(255),
                  City varchar(255))")
```

3.5 Insert Value

If you are adding values for all the columns of the table, you do not need to specify the column names in the SQL query. However, make sure the order of the values is in the same order as the columns in the table. The INSERT INTO syntax would be as follows:

```
dbExecute(MySQL,"INSERT INTO Persons(PersonID,LastName,FirstName, Address,City)
                  VALUES(1,'Siregar','Bakti', 'Jl.Bahagia','Tangerang')")
```

3.6 Truncate Table

The TRUNCATE TABLE statement is used to delete the data inside a table, but not the table itself.

```
dbExecute(MySQL, "TRUNCATE TABLE Persons")
```

3.7 Drop Table

The DROP TABLE statement is used to drop an existing table in a database.

```
dbExecute(MySQL, "DROP TABLE Persons")
```

3.8 Write Table

The key here is the `dbWriteTable` function which allows us to write an R data frame directly to a database table. The data frame's column names will be used as the database table's fields.

```
Orders      <-read.csv("data/Orders.csv")
dbWriteTable(MySQL, "Orders", Orders, append=T)
```

3.9 Alter Table

The ALTER TABLE statement is used to add, delete, or modify columns in an existing table. The ALTER TABLE statement is also used to add and drop various constraints on an existing table.

3.10 Add Column

To add a column in a table, use the following syntax:

```
dbExecute(MySQL, "ALTER TABLE Orders
                  ADD Email varchar(255)")
```

3.11 Drop Column

To delete a column in a table, use the following syntax (notice that some database systems don't allow deleting a column):

```
dbSendQuery(MySQL, "ALTER TABLE Orders
                    DROP COLUMN Email")
```

3.12 Modify Column

```
dbSendQuery(MySQL," ALTER TABLE Orders
                   MODIFY COLUMN OrderDate date")
```

3.13 Constraints

SQL constraints are used to specify rules for the data in a table. Constraints are used to limit the type of data that can go into a table. This ensures the accuracy and reliability of the data in the table. If there is any violation between the constraint and the data action, the action is aborted.

Constraints can be column level or table level. Column level constraints apply to a column, and table level constraints apply to the whole table. The following constraints are commonly used in SQL:

- *NOT NULL:* Ensures that a column cannot have a NULL value
- *UNIQUE:* Ensures that all values in a column are different
- *PRIMARY KEY:* A combination of a NOT NULL and UNIQUE. Uniquely identifies each row in a table
- *FOREIGN KEY:* Uniquely identifies a row/record in another table
- *CHECK:* Ensures that all values in a column satisfies a specific condition
- *DEFAULT:* Sets a default value for a column when no value is specified
- *INDEX:* Used to create and retrieve data from the database very quickly

3.13.1 Not Null

The following SQL ensures that the "ID", "LastName", and "FirstName" columns will NOT accept NULL values when the "Persons_NotNull" table is created:

```
dbSendQuery(MySQL,"CREATE TABLE Person_NotNull (
                   ID int NOT NULL,
                   LastName varchar(255) NOT NULL,
                   FirstName varchar(255) NOT NULL,
                   Age int)")
```

3.13.2 Unique

The following SQL creates a UNIQUE constraint on the "ID" column when the "Persons" table is created:

```
dbSendQuery(MySQL,"CREATE TABLE Persons_Unique (ID int NOT NULL UNIQUE,
                   LastName varchar(255) NOT NULL,
                   FirstName varchar(255) NOT NULL,
                   Age int)")
```

To create a UNIQUE constraint on the "ID" column when the table is already created, use the following SQL:

```
dbSendQuery(MySQL,"ALTER TABLE Persons_Unique
                   ADD UNIQUE (ID)")
```

To define a UNIQUE constraint on multiple columns, use the following SQL syntax:

```
dbSendQuery(MySQL,"ALTER TABLE Persons_Unique
                  ADD CONSTRAINT UNIQUE (ID,LastName)"
```

To drop a UNIQUE constraint, use the following SQL:

```
dbSendQuery(MySQL,"ALTER TABLE Persons_Unique
                  DROP INDEX ID")
```

3.13.3 Primary Key

The PRIMARY KEY constraint uniquely identifies each record in a table. Primary keys must contain UNIQUE values, and cannot contain NULL values. A table can have only ONE primary key; and in the table, this primary key can consist of single or multiple columns (fields).

```
dbSendQuery(MySQL,"CREATE TABLE Persons_PK (ID int NOT NULL PRIMARY KEY,
                   LastName varchar(255) NOT NULL,
                   FirstName varchar(255),
                   Age int)")
```

To allow naming of a PRIMARY KEY constraint, and for defining a PRIMARY KEY constraint on multiple columns, use the following SQL syntax:

```
dbSendQuery(MySQL,"CREATE TABLE Persons_PK (ID int NOT NULL,
                  LastName varchar(255) NOT NULL,
                  FirstName varchar(255),
                  Age int,
                  CONSTRAINT Persons_PK PRIMARY KEY (ID,LastName))")
```

To create a PRIMARY KEY constraint on the "ID" column when the table is already created, use the following SQL:

```
dbSendQuery(MySQL, "ALTER TABLE Persons_PK
                   ADD PRIMARY KEY (ID)")
```

3.13.4 Foreign Key

A FOREIGN KEY is a key used to link two tables together. A FOREIGN KEY is a field (or collection of fields) in one table that refers to the PRIMARY KEY in another table. The table containing the foreign key is called the child table, and the table containing the candidate key is called the referenced or parent table.

Look at the following two tables:

- "Persons" table:

PersonID	LastName	FirstName	Age
1	Xi	Bakti	28
2	Li	Chong	23
3	Gou	Mei	20

- "Orders" table:

OrderID	OrderNumber	PersonID
1	77895	3
2	44678	3

Notice that the "PersonID" column in the "Orders" table points to the "PersonID" column in the "Persons" table.

- The "PersonID" column in the "Persons" table is the PRIMARY KEY in the "Persons" table.
- The "PersonID" column in the "Orders" table is a FOREIGN KEY in the "Orders" table.
- The FOREIGN KEY constraint is used to prevent actions that would destroy links between tables.
- The FOREIGN KEY constraint also prevents invalid data from being inserted into the foreign key column, because it has to be one of the values contained in the table it points to.

To allow naming of a FOREIGN KEY constraint, and for defining a FOREIGN KEY constraint on multiple columns, use the following SQL syntax:

```
dbSendQuery(MySQL,
"CREATE TABLE Orders (OrderID int NOT NULL,
                      OrderNumber int NOT NULL,
                      PersonID int,
                      CONSTRAINT FOREIGN KEY (PersonID))")
```

To allow naming of a FOREIGN KEY constraint, and for defining a FOREIGN KEY constraint on multiple columns, use the following SQL syntax:

```
dbSendQuery(MySQL,
"CREATE TABLE Orders (
    OrderID int NOT NULL,
    OrderNumber int NOT NULL,
    PersonID int,
    PRIMARY KEY (OrderID),
    FOREIGN KEY (PersonID) REFERENCES Persons_pk (PersonID))")
```

To allow naming of a FOREIGN KEY constraint, and for defining a FOREIGN KEY constraint on multiple columns, use the following SQL syntax:

```
dbSendQuery(MySQL,
"ALTER TABLE Orders
ADD CONSTRAINT FK_Person Order
FOREIGN KEY (PersonID) REFERENCES Persons(PersonID)")
```

3.13.5 Check

The CHECK constraint is used to limit the value range that can be placed in a column. If you define a CHECK constraint on a single column it allows only certain values for this column. If you define a CHECK constraint on a table it can limit the values in certain columns based on values in other columns in the row. The following SQL creates a CHECK constraint on the "Age" column when the "Persons" table is created. The CHECK constraint ensures that the age of a person must be 18, or older:

```
dbSendQuery(MySQL,
"CREATE TABLE Persons (ID int NOT NULL,
                       LastName varchar(255) NOT NULL,
                       FirstName varchar(255),
                       Age int,
                       CHECK (Age>=18))")
```

To allow naming of a CHECK constraint, and for defining a CHECK constraint on multiple columns, use the following SQL syntax:

```
dbSendQuery(MySQL,
"CREATE TABLE Persons (
    ID int NOT NULL,
    LastName varchar(255) NOT NULL,
    FirstName varchar(255),
    Age int,
    City varchar(255),
    CONSTRAINT CHK_Person CHECK (Age>=18 AND City='Sandnes'))")
```

To create a CHECK constraint on the "Age" column when the table is already created, use the following SQL:

```
dbSendQuery(MySQL, "ALTER TABLE Persons
                          ADD CHECK (Age>=18)")
```

To allow naming of a CHECK constraint, and for defining a CHECK constraint on multiple columns, use the following SQL syntax:

```
dbSendQuery(MySQL, "ALTER TABLE Persons
                    ADD CONSTRAINT CHK_PersonAge
                    CHECK (Age>=18 AND City='Sandnes')")
```

3.13.6 Default

The DEFAULT constraint is used to provide a default value for a column. The default value will be added to all new records IF no other value is specified. The following SQL sets a DEFAULT value for the "City" column when the "Persons" table is created:

```
dbSendQuery(MySQL,
"CREATE TABLE Persons_default (ID int NOT NULL,
                               LastName varchar(255) NOT NULL,
                               FirstName varchar(255),
                               Age int,
                               City varchar(255) DEFAULT 'Sandnes')")
```

To create a DEFAULT constraint on the "City" column when the table is already created, use the following SQL:

```
dbSendQuery(MySQL,"ALTER TABLE Persons
                         ALTER City SET DEFAULT 'Sandnes'")
```

3.13.7 Index

The CREATE INDEX statement is used to create indexes in tables. Indexes are used to retrieve data from the database more quickly than otherwise. The users cannot see the indexes, they are just used to speed up searches/queries. Creates an index on a table. Duplicate values are allowed:

```
dbSendQuery(MySQL, "CREATE INDEX idx_pname
                    ON Persons (LastName, FirstName)")
```

Note: Updating a table with indexes takes more time than updating a table without (because the indexes also need an update). So, only create indexes on columns that will be frequently searched against.

3.13.8 Auto Increment

Auto-increment allows a unique number to be generated automatically when a new record is inserted into a table. Often this is the primary key field that we would like to be created automatically every time a new record is inserted. The following SQL statement defines the "Personid" column to be an auto-increment primary key field in the "Persons" table:

```
dbSendQuery(MySQL,
"CREATE TABLE Persons_ai (
    Personid int NOT NULL AUTO_INCREMENT,
    LastName varchar(255) NOT NULL,
    FirstName varchar(255),
    Age int,
    PRIMARY KEY (Personid))")
```

3.14 Previewing `.sql` in R

When you open a new `.sql` file in RStudio, it automatically populates the file with the following code:

```
library(RSQLite)
library(dplyr)
library(dbplyr)

conn <- src_memdb() # create a SQLite database in memory
copy_to(conn,
        storms,     # this is a dataset built into dplyr
        overwrite = TRUE)
tbl(conn, sql("SELECT * FROM storms LIMIT 5"))
```

You need to create a `.sql` file with the following code:

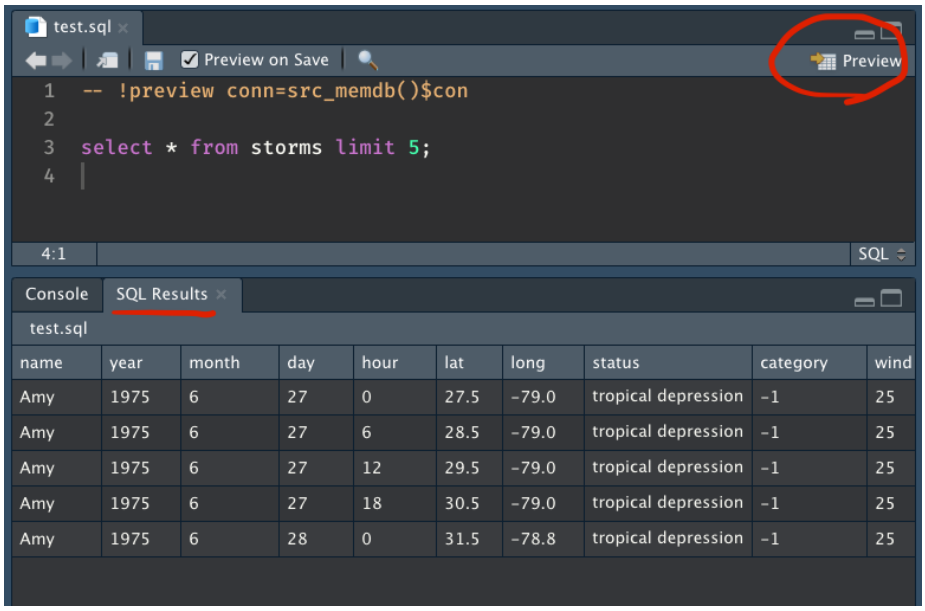
```
-- !preview conn=src_memdb()$con

SELECT * FROM storms LIMIT 5
```

Then, you will the resul like this:

```
library(knitr)
include_graphics("./images/Bab3/sql-file-preview.png")
```

3.15 SQL chunks in RMarkdown

I generally prefer to show RMarkdown output in the console 1 (and it looks like I'm not the only one). This means that when I run code in an .Rmd file, it feels more or less the same as when I run an .R file: the plots show up in the plots pane, code is run in the console, and so on. While you can use SQL chunks with



name	year	month	day	hour	lat	long	status	category	wind
Amy	1975	6	27	0	27.5	-79.0	tropical depression	-1	25
Amy	1975	6	27	6	28.5	-79.0	tropical depression	-1	25
Amy	1975	6	27	12	29.5	-79.0	tropical depression	-1	25
Amy	1975	6	27	18	30.5	-79.0	tropical depression	-1	25
Amy	1975	6	28	0	31.5	-78.8	tropical depression	-1	25

Figure 3.1: Previewing '.sql' in R

this setting, there is NO chunk preview option. You must trust your queries and knit the file to make sure everything runs. You get the syntax highlighting razzle-dazzle but alas– no preview.

It is in this very specific case where inline mode wins big time. SQL previews magically become an option, allowing you to interact with your beautifully colored SQL code.

Chapter 4

Database Normalization in SQL

Normalization is a database design technique that reduces data redundancy and eliminates undesirable characteristics like Insertion, Update and Deletion Anomalies. Normalization rules divides larger tables into smaller tables and links them using relationships. The purpose of Normalisation in SQL is to eliminate redundant (repetitive) data and ensure data is stored logically.

The inventor of the relational model Edgar Codd proposed the theory of normalization of data with the introduction of the First Normal Form, and he continued to extend theory with Second and Third Normal Form. Later he joined Raymond F. Boyce to develop the theory of Boyce-Codd Normal Form.

Here is a list of Normal Forms in SQL:

- 1NF (First Normal Form)
- 2NF (Second Normal Form)
- 3NF (Third Normal Form)
- BCNF (Boyce-Codd Normal Form)
- 4NF (Fourth Normal Form)
- 5NF (Fifth Normal Form)
- 6NF (Sixth Normal Form)

The Theory of Data Normalization in MySQL server is still being developed further. For example, there are discussions even on 6th Normal Form. However, in most practical applications, normalization achieves its best in 3rd Normal Form.

4.1 The Process of Normalization

The process of normalization involves breaking down a large table into smaller, related tables and defining relationships between them. This helps in achieving the following benefits:

- **Elimination of Data Redundancy:** Redundant data can lead to inconsistencies and increased storage requirements. Normalization ensures that each piece of data is stored in only one place, reducing redundancy and promoting consistency.
- **Data Integrity:** Normalization minimizes the chances of inconsistencies and anomalies that may occur when data is duplicated or updated in one place but not in another.
- **Efficient Data Updates:** Since data is stored in smaller, more specific tables, updates are more efficient and require fewer changes.
- **Simpler Queries:** Normalized data allows for more straightforward and efficient querying due to the structured relationships between tables.

The process of database normalization is typically divided into several "normal forms" (often referred to as 1NF, 2NF, 3NF, BCNF, etc.), each with its own set of rules and requirements. These normal forms build on each other, with higher normal forms addressing more complex issues of redundancy and dependency. Here's a brief overview of some common normal forms:

1. **First Normal Form (1NF):**

- Eliminate duplicate columns.
- Create separate tables for related data.
- Define a primary key for each table.

2. **Second Normal Form (2NF):**

- Meet 1NF requirements.
- Remove partial dependencies (attributes dependent on only part of the primary key) by creating
- separate tables.

3. **Third Normal Form (3NF):**

- Meet 2NF requirements.
- Remove transitive dependencies (attributes dependent on non-key attributes) by creating separate tables.

4. **Boyce-Codd Normal Form (BCNF):**

- Meet 3NF requirements.
- Remove overlapping candidate keys by creating separate tables.

Higher normal forms exist beyond these, such as **Fourth Normal Form (4NF)** and **Fifth Normal Form (5NF)**, but they are less commonly encountered and may be more relevant in specific cases of complex data modeling. While normalization offers significant benefits, it's important to strike a balance between normalization and performance. Over-normalization can lead to complex query logic and decreased query performance. Therefore, designing a database often involves considering the nature of the data and the queries that will be performed on it.

4.2 Simple Database Normalization

Let's go through a simple example of database normalization using a hypothetical scenario of an online bookstore. We'll start with an unnormalized table and then progressively normalize it through different normal forms.

Scenario: Consider an unnormalized table that stores information about books, authors, and their publishers.

Unnormalized Table (1NF):

Book ID	Title	Author	Author Birth	Publisher	Year
1	Algorithm	John Smith	1980-05-15	ABC Pub	2000
2	Data Science	Jane Doe	1975-10-20	XYZ Books	2015
3	Database System	John Smith	1980-05-15	ABC Pub	2012

In this unnormalized table, we have duplicate author and publisher information, leading to redundancy. John Smith's information is repeated, and if any of his details change, we need to update multiple rows.

First Normal Form (1NF):

To achieve 1NF, we break the table into smaller tables and remove duplicate data. We create separate tables for authors and publishers.

Authors Table:

Author ID	Author	Author Birth
1	John Smith	1980-05-15
2	Jane Doe	1975-10-20

Publishers Table:

Publisher ID	Publisher
1	ABC Pub
2	XYZ Books

Books Table (1NF):

Book ID	Title	Author ID	Publisher ID	Year
1	Algorithm	1	1	2000
2	Data Science	2	2	2015
3	Database System	1	1	2012

We've eliminated redundancy by referencing author and publisher IDs in the books table.

Second Normal Form (2NF):

To achieve 2NF, we identify partial dependencies and create a separate table for author information.

Authors Table (2NF):

Author ID	Author	Author Birth
1	John Smith	1980-05-15
2	Jane Doe	1975-10-20

Books Table (2NF):

Book ID	Title	Author ID	Publisher ID	Year
1	Algorithm	1	1	2000
2	Data Science	2	2	2015
3	Database System	1	1	2012

No changes are required in the Books table for 2NF since there were no partial dependencies.

Third Normal Form (3NF):

To achieve 3NF, we identify transitive dependencies and create a separate table for publisher information.

Publishers Table (3NF):

Table 4.9: First Normal Form (1NF)

BookID	Title	Author	Genre	Publisher	PublicationYear
1	Book A	Author X	Fiction	dsciencelabs	2021
2	Book B	Author Y	Mystery	Matana	2022
3	Book B	Author X	Romance	dsciencelabs	2023

Publisher ID	Publisher
1	ABC Pub
2	XYZ Books

Books Table (3NF):

Book ID	Title	Author ID	Publisher ID	Year
1	Algorithm	1	1	2000
2	Data Science	2	2	2015
3	Database System	1	1	2012

No changes are required in the Books table for 3NF since there were no transitive dependencies. The result is a normalized database structure that eliminates redundancy and ensures data integrity.

Please note that the above example is simplified for demonstration purposes. In real-world scenarios, databases can have more complex structures and relationships, which may require deeper levels of normalization to achieve higher normal forms like BCNF or 4NF.

4.3 Your Job

Consider a hypothetical database for an online bookstore. We'll start with a denormalized table and then go through the normalization process step by step.

Suppose we have a single table called Books with the following columns:

Your job is the following statements:

1. Display Database Normalization Process
2. Create Database to your PC After Normalization Process using R and SQL

Chapter 5

Join Table in SQL

A SQL join is a Structured Query Language (SQL) instruction to combine data from two sets of data (i.e. two tables). Before we dive into the details of a SQL join, let's briefly discuss what SQL is, and why someone would want to perform a SQL join.

SQL is a special-purpose programming language designed for managing information in a relational database management system (RDBMS). The word relational here is key; it specifies that the database management system is organized in such a way that there are clear relations defined between different sets of data. Typically, you need to extract, transform, and load data into your RDBMS before you're able to manage it using SQL, which you can accomplish by using a tool like Stitch.

5.1 Relational Database

Imagine you're running a store and would like to record information about your customers and their orders. By using a relational database, you can save this information as two tables that represent two distinct entities: customers and orders .

5.1.1 Table Customers

```
library(DT)
costomers<-read.csv("data/customers.csv")
```

```
datatable(head(costomers, 5),
          caption = htmltools::tags$caption(
            style = 'caption-side: bottom; text-align: center;',
            htmltools::em('Table 1: customers.')),
          options = list(dom = 't'))
```

Table 1, informs about each customer is stored in its own row, with columns specifying different bits of information, including their first name, last name, and email address. Additionally, we associate a unique customer number, or primary key, with each customer record.

5.1.2 Table Orders

```
orders<-read.csv("data/orders.csv")
datatable(head(orders,5),
          caption = htmltools::tags$caption(
            style = 'caption-side: bottom; text-align: center;',
            htmltools::em('Table 2: orders.')),
          options = list(dom = 't'))
```

Again, Table 2 are contains information about a specific order. Each order has its own unique identification key `order_id` for this table – assigned to it as well.

5.2 Relational Model

You've probably noticed that these two examples share similar information. You can see these simple relations diagrammed below:

Note that the orders table contains two keys: one for the order and one for the customer who placed that order. In scenarios when there are multiple keys in a table, the key that refers to the entity being described in that table is called the primary key *(PK)* and other key is called a foreign key *(FK)*.

In our example, `order_id` is a primary key in the orders table, while `customer_id` is both a primary key in the customers table and a foreign key in the orders table. Primary and foreign keys are essential to describing relations between the tables, and in performing SQL joins.

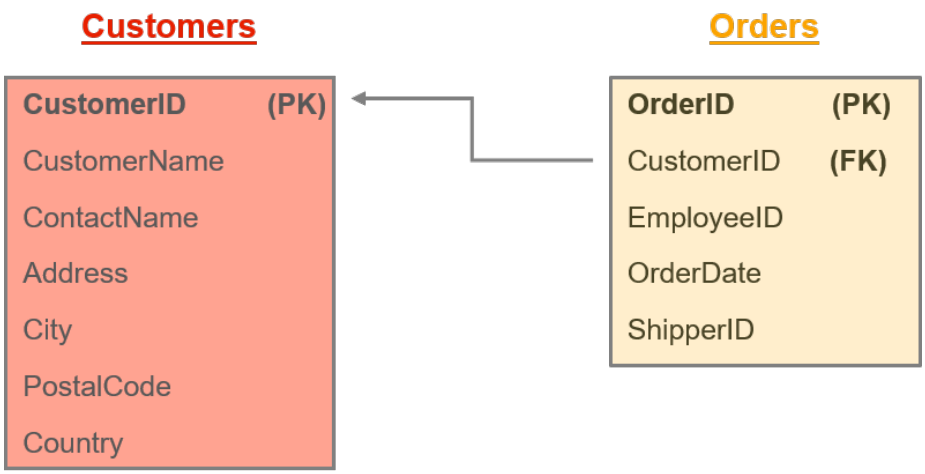


Figure 5.1: Relational Table

5.3 Factory Database

To make you more convenient about all the data tables that we will use in this section. Here, I summarize the following SQL relational for database `factory_db`:

Note: Don't forget to consider the data structure of your database (all table)

5.4 Basic SQL Join Types

There are four basic types of SQL joins: inner, left, right, and full. The easiest and most intuitive way to explain the difference between these four types is by using a Venn diagram, which shows all possible logical relations between data sets.

Again, it's important to stress that before you can begin using any join type, you'll need to extract the data and load it into an RDBMS like Amazon Redshift, where you can query tables from multiple sources. You build that process manually, or you can use an ETL service like Stitch, which automates that process for you.

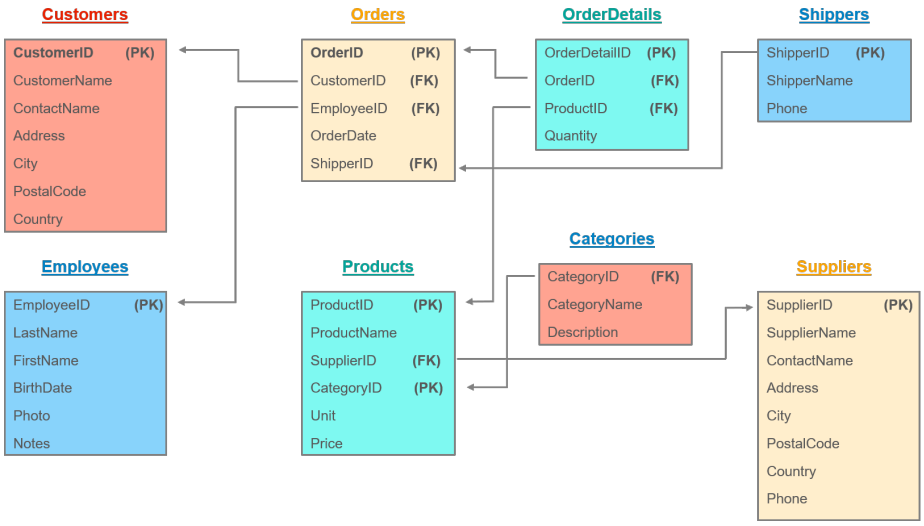


Figure 5.2: Relational Table of Factory Database

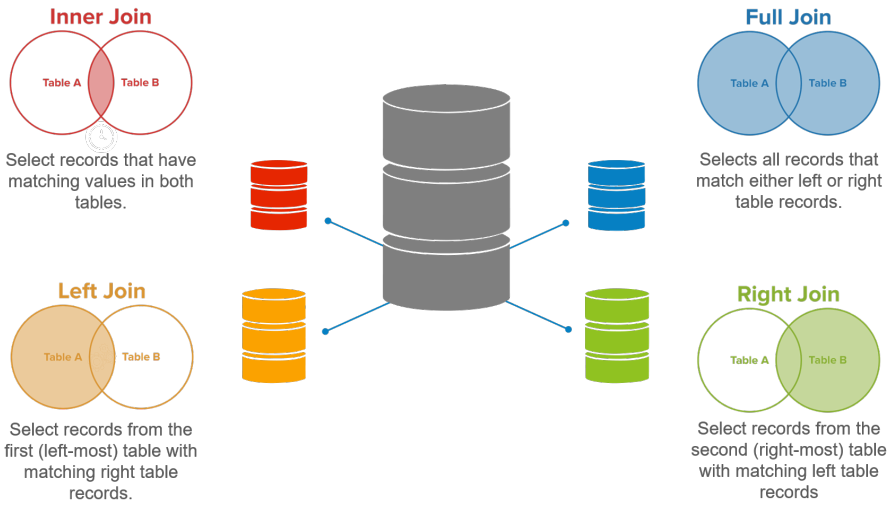


Figure 5.3: Basic Join Table

5.5 Connect to MySQL

Reading data from `MySQL` into R workspace, it requires two R libraries, `RMySQL` and `DBI`. The connection data should not be embedded in analysis code. Separate the connection code in another script. The script should set up the connection and save it into the workspace.

The saved connection is accessible by its name in the analysis code. In the dbConnect function, you need to replace dbname, username, pwd, dbserver and port with the actual values of your remote database.

```
# set up the connection and save it into the workspace
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
library(RMySQL)
library(DBI)
bakti <- dbConnect(RMySQL::MySQL(),
                   dbname='factory_db',
                   username='root',
                   password='',
                   host='localhost',
                   port=3306)
knitr::opts_chunk$set(connection = "bakti")   # set up the connection
```

After set up the connection and save it into the workspace. Next, we can run SQL in a code chunk of type sql. By setting the connection in the code chuck and adding the option output.var, the resulting table from the SQL is written into a variable in R.

```
'''{sql connection=bakti, output.var="report_model_by_make"}
      Your SQL code Here
'''
```

5.6 Inner Join

Let's say we wanted to get a list of those customers who placed an order and the details of the order they placed. This would be a perfect fit for an inner join, since an inner join returns records at the intersection of the two tables.

```
SELECT OrderID, CustomerName
  FROM Orders O
    INNER JOIN Customers C
      ON O.CustomerID = C.CustomerID
```

```
library(DT)
datatable(Inner1,
          caption = htmltools::tags$caption(
            style = 'caption-side: bottom; text-align: center;',
            htmltools::em('Table 3: SQL Inner Join Two Tables.')))
```

The following SQL statement selects all orders with customer and shipper information:

```
SELECT *
  FROM ((Orders O
    INNER JOIN Customers C
      ON O.CustomerID = C.CustomerID)
    INNER JOIN Shippers S
      ON O.ShipperID = S.ShipperID)
```

```
datatable(Inner2,
          caption = htmltools::tags$caption(
            style = 'caption-side: bottom; text-align: center;',
            htmltools::em('Table 4: SQL Inner Join Three Tables.')),
          extensions = 'FixedColumns',
          options = list(scrollX = TRUE, fixedColumns = TRUE)
          )
```

5.7 Left Join

If we wanted to simply append information about orders to our customers table, regardless of whether a customer placed an order or not, we would use a left join. A left join returns all records from table A and any matching records from table B. The result is NULL from the right side, if there is no match.

```
SELECT CustomerName, OrderID
  FROM Customers C
    LEFT JOIN Orders O
      ON C.CustomerID = O.CustomerID
        ORDER BY C.CustomerName
```

```
datatable(Left,
          caption = htmltools::tags$caption(
            style = 'caption-side: bottom; text-align: center;',
            htmltools::em('Table 5: SQL Left Join Two Tables.')))
```

5.8 Right Join

The following SQL statement will return all employees, and any orders they might have placed. The result is NULL from the left side, when there is no match.

```
SELECT OrderID, LastName, FirstName
  FROM Orders O
    RIGHT JOIN Employees E
      ON O.EmployeeID = E.EmployeeID
        ORDER BY O.OrderID
```

```
datatable(Right,
          caption = htmltools::tags$caption(
            style = 'caption-side: bottom; text-align: center;',
            htmltools::em('Table 6: SQL Right Join Two Tables.')))
```

5.9 Full Join

The FULL OUTER JOIN keyword returns all records when there is a match in left (table1) or right (table2) table records. FULL OUTER JOIN, FULL JOIN, and JOIN (MariaDB) are the same. The following SQL statement selects all customers, and all orders:

```
SELECT CustomerName, OrderID
  FROM Customers C
    JOIN Orders O
      ON C.CustomerID=O.CustomerID
        ORDER BY C.CustomerName
```

```
datatable(Full,
          caption = htmltools::tags$caption(
            style = 'caption-side: bottom; text-align: center;',
            htmltools::em('Table 7: SQL Full Join Two Tables.')))
```

5.10 Self JOIN

A self JOIN is a regular join, but the table is joined with itself. The following SQL statement matches customers that are from the same city:

```
SELECT A.CustomerName AS CustomerName1,
       B.CustomerName AS CustomerName2,
       A.City
  FROM Customers A,
       Customers B
    WHERE A.CustomerID <> B.CustomerID
      AND A.City = B.City
        ORDER BY A.City
```

```
datatable(Self,
          caption = htmltools::tags$caption(
            style = 'caption-side: bottom; text-align: center;',
            htmltools::em('Table 8: SQL Self Join Two Tables.')))
```

After finishing the work with the database, close the connection.

```
DBI::dbDisconnect(bakti)
```

5.11 Your Job

1. Apply Left join and Right join to returns all records from table Orders and any matching records from table Suppliers.
2. Choose the correct JOIN clause to select all records from the two tables (Orders and Suppliers) where there is a match in both tables.
3. Choose the correct JOIN clause to select all the records from the Suppliers table plus all the matches in the Orders table.

Chapter 6

Referensi