DigitalSkola

BUSINESS INTELLIGENCE AND LINUX

DATA ENGINEER BATCH 9 - WEEK #4

DWI HANDOYO

DigitalSkola

BUSINESS INTELLIGENCE

Business Intelligence (BI) is a technology-driven process for analyzing data and delivering actionable information in forms of reports, summaries, dashboards, graphs, charts etc. in order to help organizations make better data-driven decisions.

Scope of Works in Business Intelligence

Business Intelligence Tools: Power BI, Google Data Studio, Metabase, Looker, Tableau

Data Processing

Data is collected and translated into usable information. In this task, Data Engineer responsibile for activities such as tidy up data, data cleansing, tubular data transformation prior to visualization.

Reporting

In order to display clear insights which is easier to understand, more representative, and more intuitive, a report is made in the form of data visualization on a dashboard. This is an overlapping task between Data Engineer and Data Analyst. Both parties shall work together to come up with best data presentation/visualization. Data Analyst provide mock up of the report, while Data Engineer implement it combined with data source into dashboard report deliverable. Data Engineer responsible for report developent and automation.

Data Ops

Operation on data where Data Analyst or Business Intelligence Engineer is responsible for sustainability of data, ensuring that the visualized data is correct, the value is not wrong, nothing is corrupt or missing. Data integrity and validity shall be ensured.

DigitalSkola

Business Intelligence Workflow

Data Gathering: To collect data by extraction and ingestion data from relevant data sources.

Data Cleansing: To do cleaning on data by excluding corrupt items, data validation, and data standardization (data types, date format eg. YYYY-MM-DD, etc. shall be consistent).

Analysis: The Data Engineer and Data Analyst discuss together on what columns to display or whether to create derivative columns. For example, if there is a birthday column from there, we can create a new column such as the age column, which is obtained from the difference between today and the customer's birthday.

Reporting: The ultimate goal of the reporting process is to generate an automated dashboard which will be connected/combined with ETL for data retrieval. Hence, the data can always be updated so that our company or any relevant divisions in it will always get the latest data.

DigitalSkola

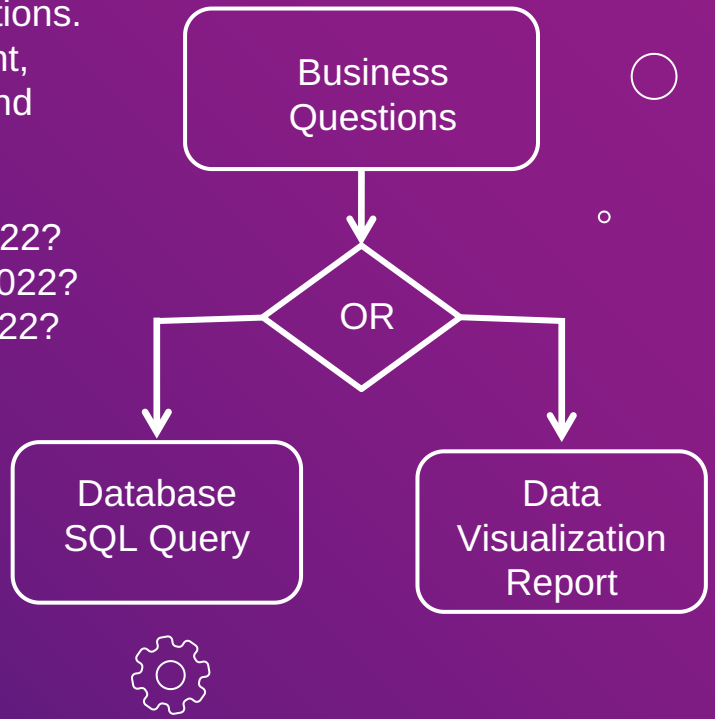
Business Questions

Business Intelligence is useful to respond Bussiness Questions. As an example, the questions raised up by the management, Marketing Division, Marchadising Division, etc. to understand company's sales performance as below:

- How many transactions in first three days of October 2022?
- What was the total sales in first three days of October 2022?
- What are the highest five selling products in October 2022?

Data can be individually retrieved from PostgreSQL database for each questions above by using SQL query via DBeaver interface.

Or another way is by creating visualization of report by using Google Data Studio.



DigitalSkola

Data Source: Sales in October 2022

```
SELECT * FROM prj2_sales;
```

prj2_sales 1

SELECT * FROM prj2_sales | Enter a SQL expression to filter results (use Ctrl+Space)

	id	trx_dt	cust_name	prd_name	qty	total_price
1	1	2022-10-05	Terry Swaine	Downy Softener Passion	3	109,800
2	2	2022-10-01	Donald Kamal	Indomie Mie Instan Kaldu Ayam	10	31,000
3	3	2022-10-03	Marianne Velasco	Chitato Snack Potato Chips Truffle	7	77,700
4	4	2022-10-05	James Hawkins	Chitato Snack Potato Chips Truffle	7	77,700
5	5	2022-10-01	Fred Scott	Harpic Pembersih Kloset	2	59,000
6	6	2022-10-02	Gerald Mosher	Good Mood Lemon	1	5,800
7	7	2022-10-03	Geraldine Ivory	Chitato Snack Potato Chips Salmor	6	68,400
8	8	2022-10-02	Marianne Velasco	Golda Coffe Drink Cappuccino	2	6,000
9	9	2022-10-02	Terry Swaine	Good Mood Stroberi	3	17,400
10	10	2022-10-03	Jose Moczo	Oreo Wafer Vanilla	2	17,200

Using database query each time question come is not practical, in this example, there are three reports to answer questions.

Top 5 product sales in 3 days query (01/10/2022 - 03/10/2022)

```
select prd_name, sum(total_price) as total_sales
from prj2_sales
where trx_dt between '2022-10-01' and '2022-10-03'
group by prd_name
order by total_sales desc
limit 5;
```

Total transaction in 3 days query

```
select count(id) as total_transaction
from prj2_sales
where trx_dt between '2022-10-01' and '2022-10-03';
```

Total sales (price) in 3 days query

```
select sum(total_price) as total_sales
from prj2_sales
where trx_dt between '2022-10-01' and '2022-10-03';
```

Three Query Results

Results 1

select count(id) as total_transac

	total_transaction
1	2,974

Results 1

select sum(total_price) as

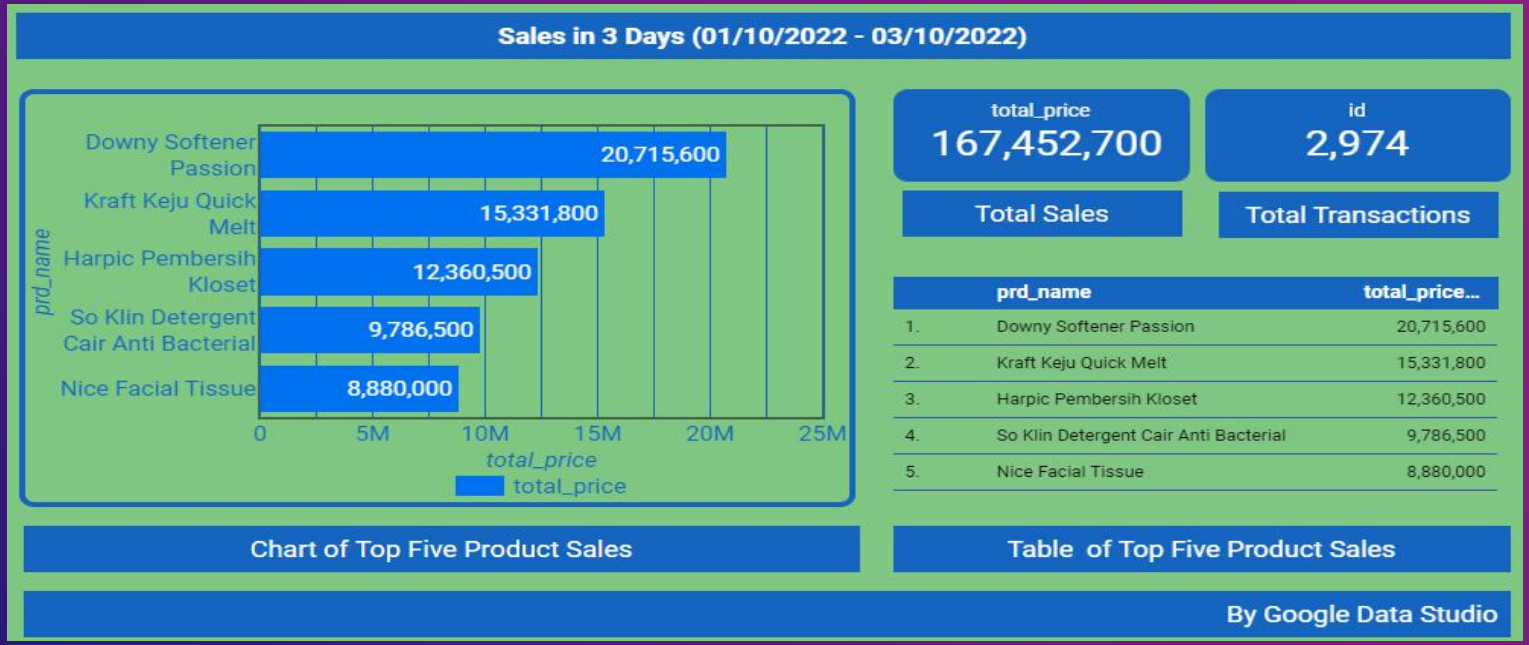
	total_sales
1	167,452,700

prj2_sales 1

select prd_name, sum(total_pric | Enter a SQL expression to filter result

	prd_name	total_sales
1	Downy Softener Passion	20,715,600
2	Kraft Keju Quick Melt	15,331,800
3	Harpic Pembersih Kloset	12,360,500
4	So Klin Detergent Cair Anti Bacterial	9,786,500
5	Nice Facial Tissue	8,880,000

Rather than answering with SQL query, to have more easy to read and intutive report, data visualization is provided. Below is visualization report for answering the three business questions. The report is generated by using Google Data Studio.



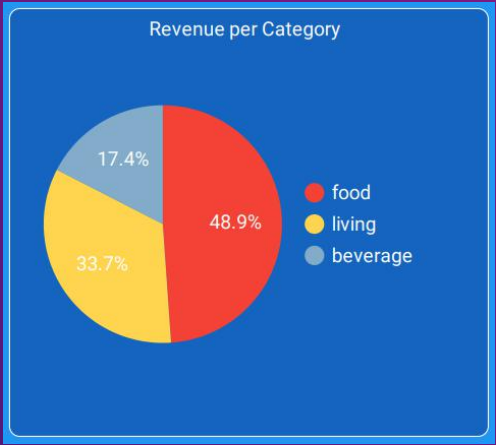
DigitalSkola

Types of Charts in Data Visualization

Scorecard

Scorecard is used to display a value aggregation result.

Pie



Pie chart is used to display contributions from each category (less than 5 categories).

Table

Top 10 Products by Revenue

	Product	Category	Total Revenue (Rp)
1.	Downy Softener Passion	living	35,538,600
2.	Kraft Keju Quick Melt	food	25,118,700
3.	Harpic Pembersih Kloset	living	20,738,500
4.	So Klin Detergent Cair Anti Bacterial	living	16,298,500
5.	Nice Facial Tissue	living	15,628,800
6.	Sari Roti Tawar Kupas	food	13,357,500
7.	Aqua Galon 19L	beverage	13,347,600
8.	Chitato Snack Potato Chips Beef Barbeque	food	10,670,400
9.	Chitato Snack Potato Chips Sourcream Onion	food	10,103,600
10.	Chitato Snack Potato Chips Mi Goreng	food	9,073,500

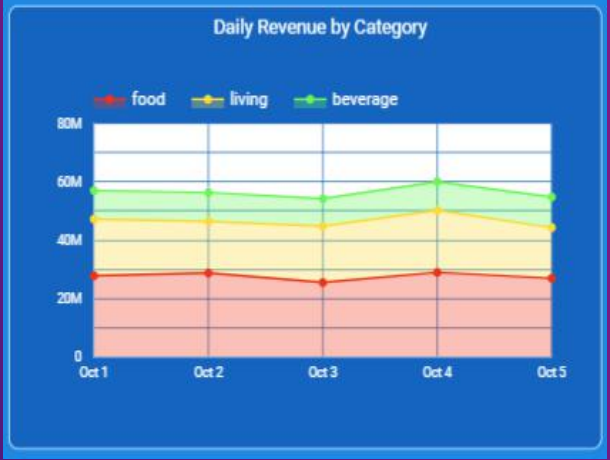
Table is used to display detailed data.

DigitalSkola

Line

Line chart is used to display data and change over time.

Area



Area chart is used to display a value of several categories and their contributions.

Bar

Bar chart is used to display contributions from each category (more than 5 categories or for long category names).

DigitalSkola

Scatter

Scatter chart is used to see the effect of three variables with respect to each other.

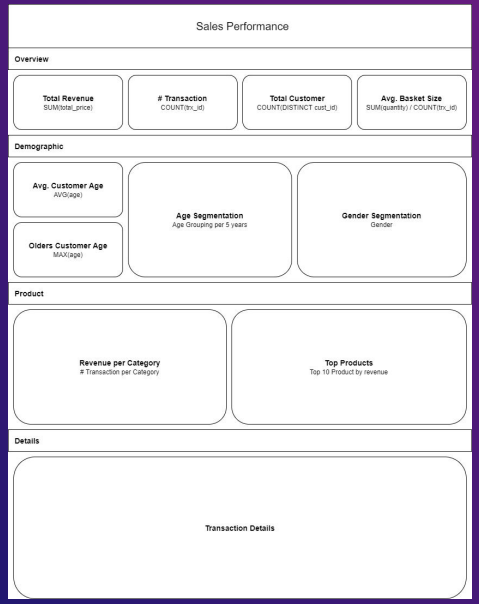
Histogram

Histogram is used for displaying distribution of values.

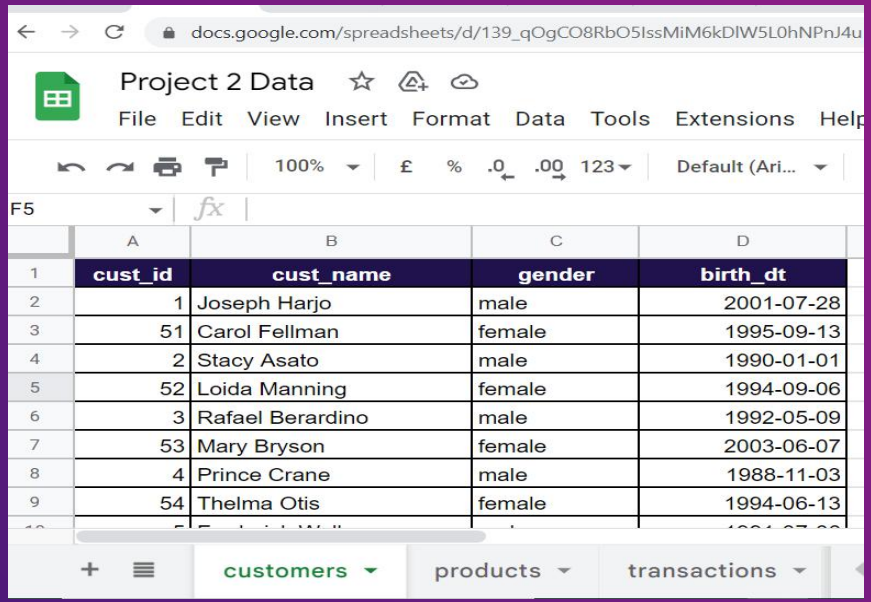
DigitalSkola

PROJECT 2 - GOOGLE DATA STUDIO

This project intention is to create dashboard complying mock up design based on data set given by using Google Data Studio (Looker Studio). The data consists of three tables, ie. customers, products, and transactions, saved in Google Sheet.



Mock Up Layout Design



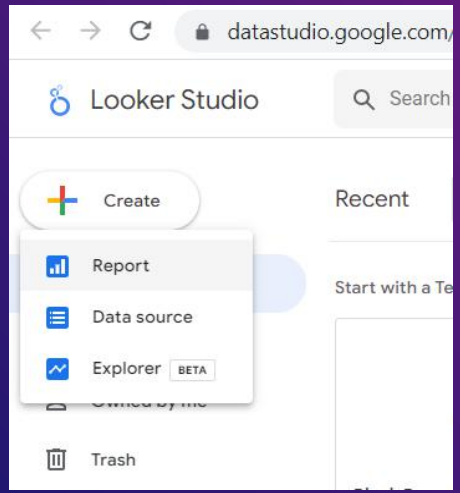
	A	B	C	D
1	cust_id	cust_name	gender	birth_dt
2	1	Joseph Harjo	male	2001-07-28
3	51	Carol Fellman	female	1995-09-13
4	2	Stacy Asato	male	1990-01-01
5	52	Loida Manning	female	1994-09-06
6	3	Rafael Berardino	male	1992-05-09
7	53	Mary Bryson	female	2003-06-07
8	4	Prince Crane	male	1988-11-03
9	54	Thelma Otis	female	1994-06-13

Data Source

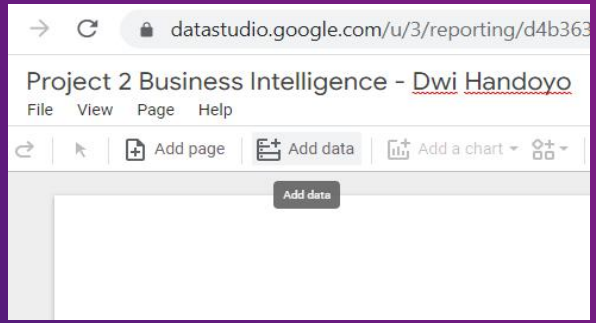
DigitalSkola

Connection with data source.

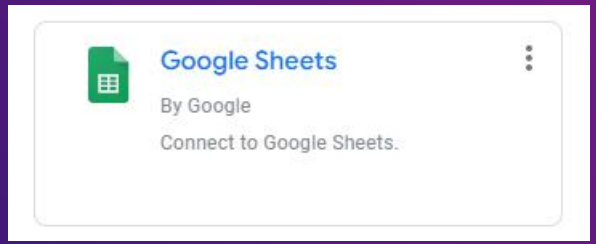
1. Create Report in Google Studio.



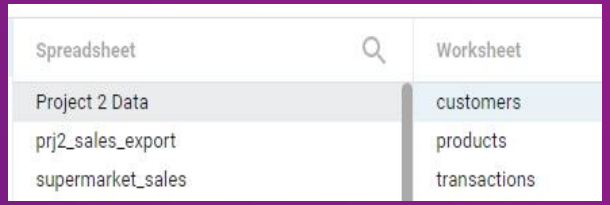
2. Setting Data Source, select Add Data.



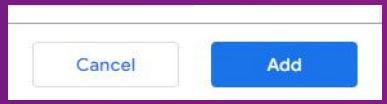
3. Select Google Sheets.

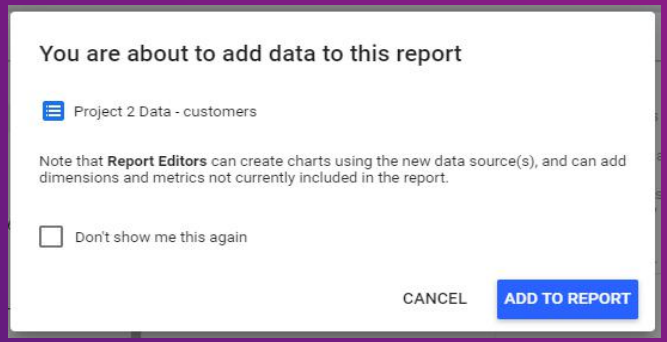


4. Select Project 2 Data, click customers. Repeat for products and transactions tables.



5. Click Add.

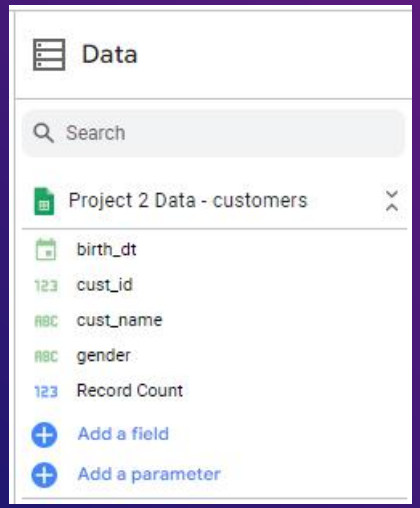




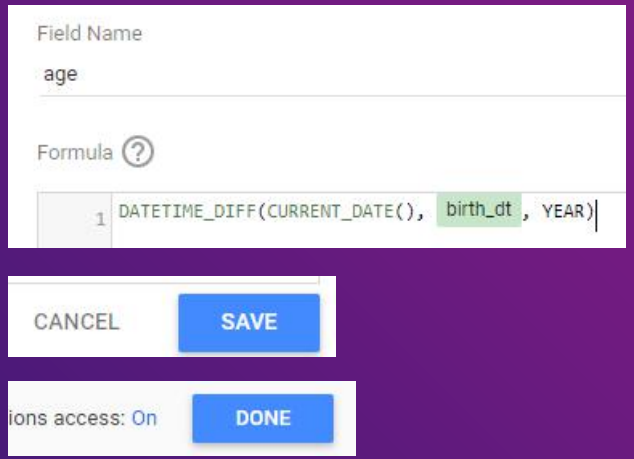
DigitalSkola

Create Derivative Column by Using Function

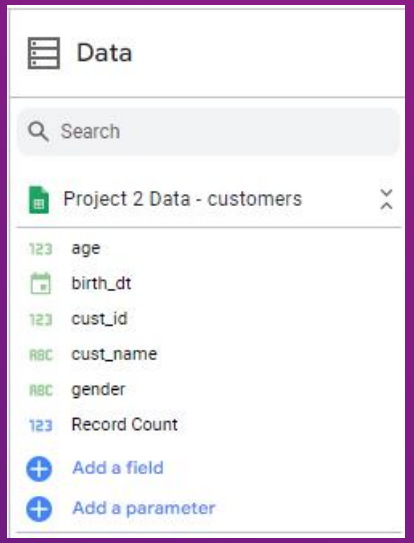
1. Original columns in table customers. Click Add a field on the bottom.



2. Put column/field name, type: age, then type the formula into form, click SAVE, and click DONE.



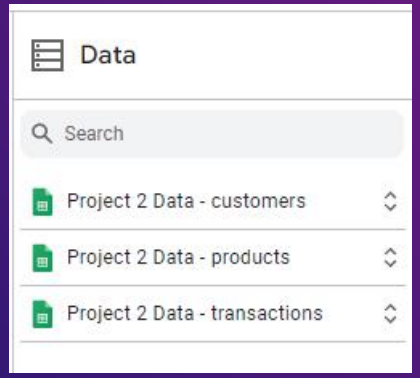
3. New derivative column is added to costumers data



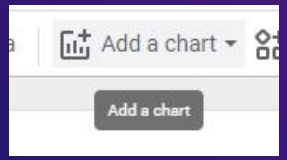
DigitalSkola

Create Combined (Blended) Data Set by Using JOIN

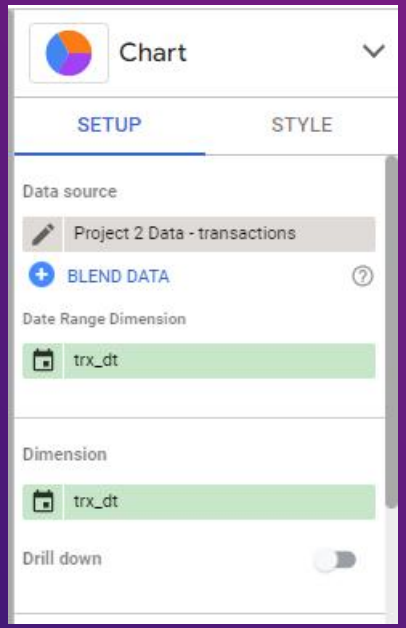
1. Original data sets



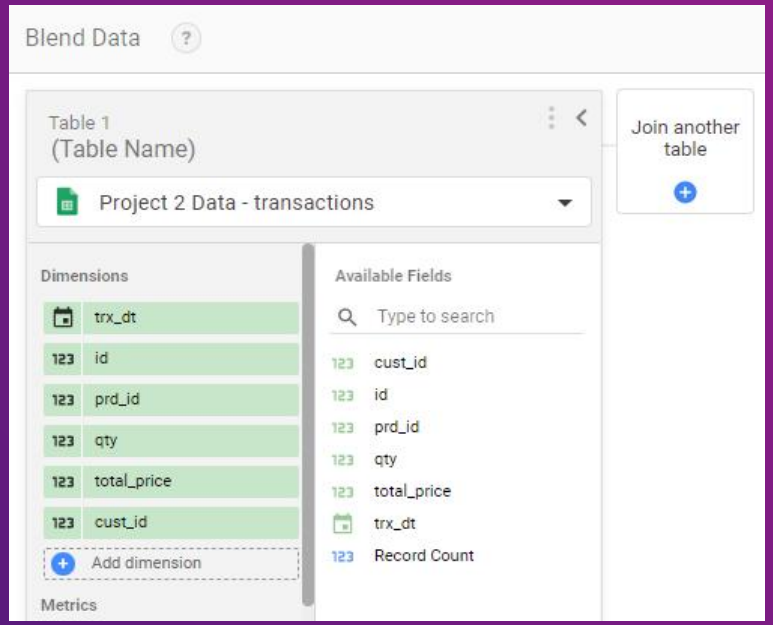
2. Add chart



3. Click BLEND

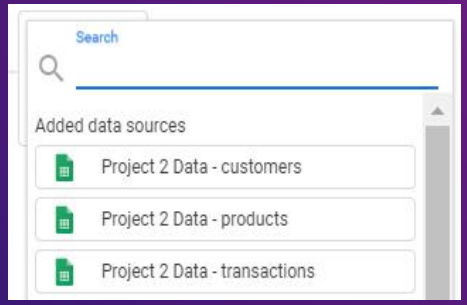


4. Fill in (Table Name) of Table 1 with “trx”, Add dimension from Available Fields, and click Join another table.

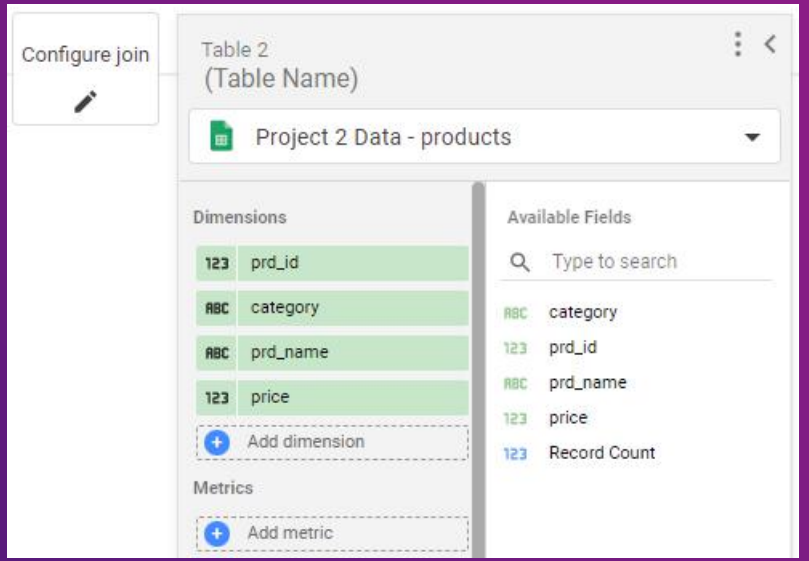


DigitalSkola

5. Select Project 2 Data - products

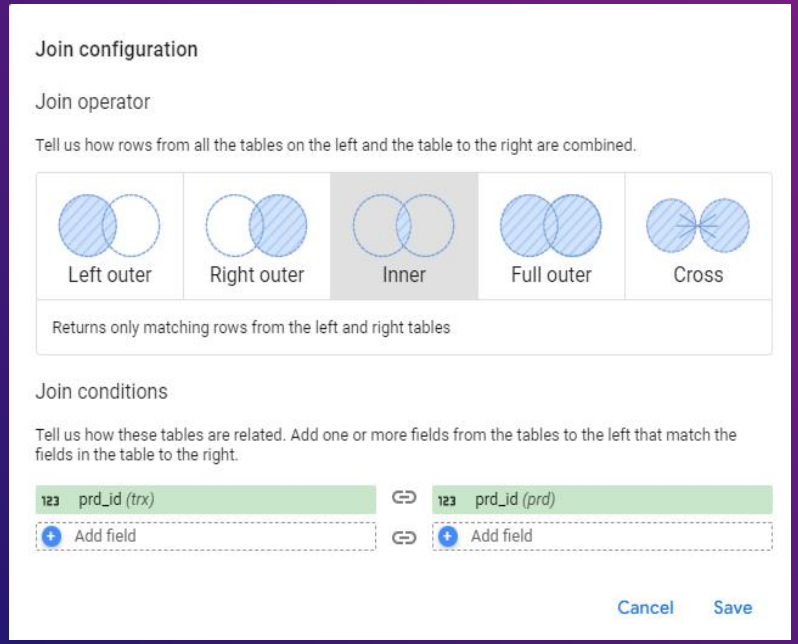


6. Fill in (Table Name) of Table 2 with “prd”, Add dimension from Available Fields, and click Configure join.

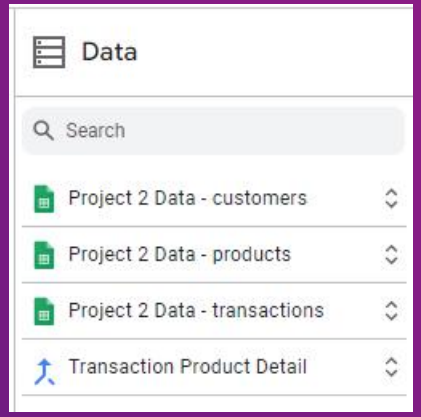


DigitalSkola

7. Select Join operator to Inner, select Join conditions to prd_id, then clicik Save.



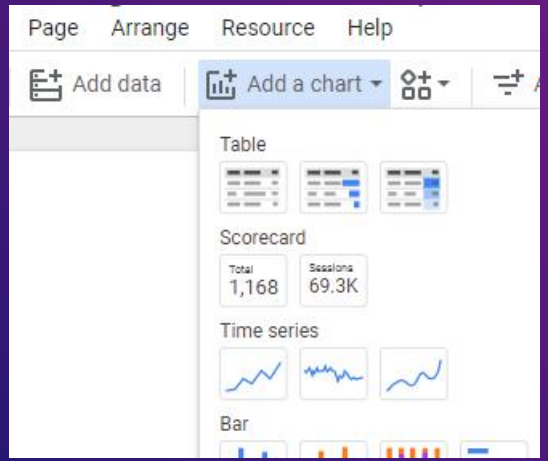
8. Replace Blended Data (1) with “Transaction Product Detail”, click SAVE, then click Close.
9. Finally blended data set has been added at the bottom.



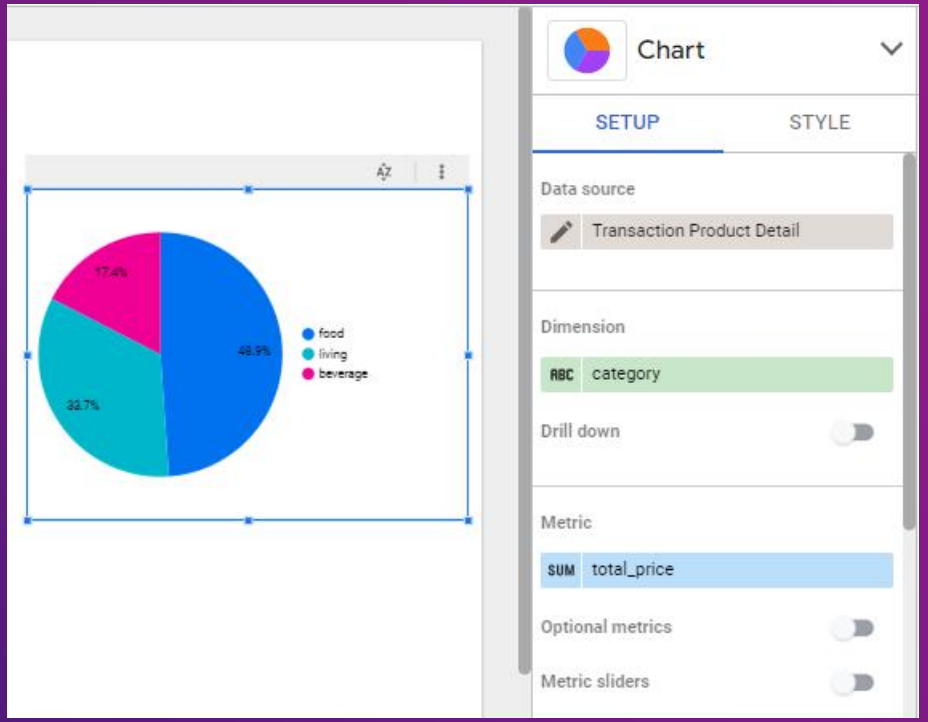
DigitalSkola

Create charts to implement mock up design

1. Click Add Chart, select any required charts, then arrange the position as per mock up.

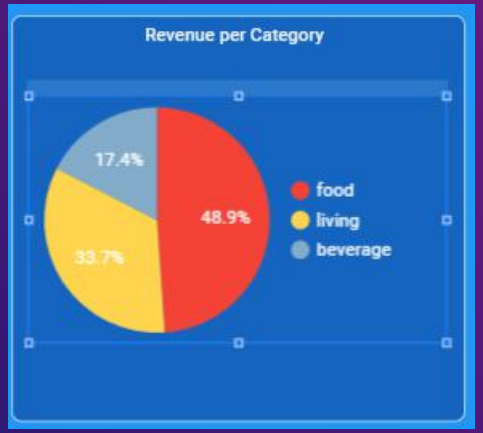


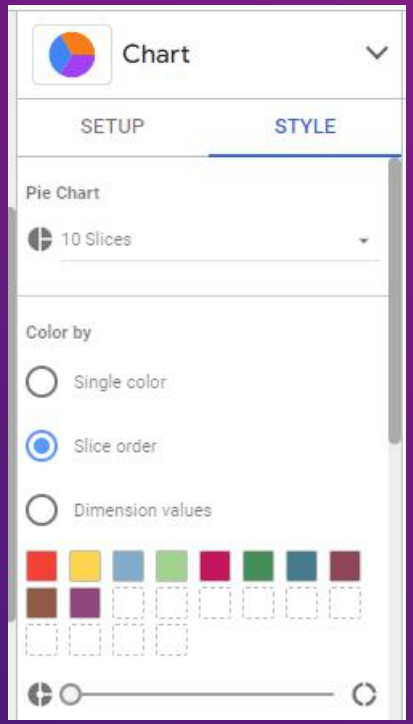
2. At Chart SETUP, select applicable Data source, Dimension, and Metric.

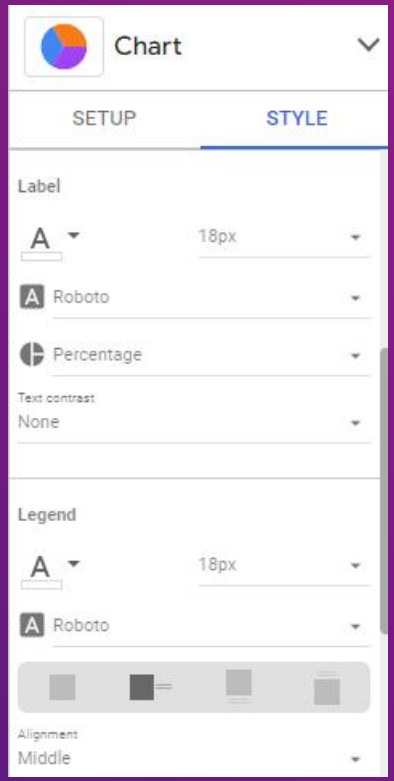


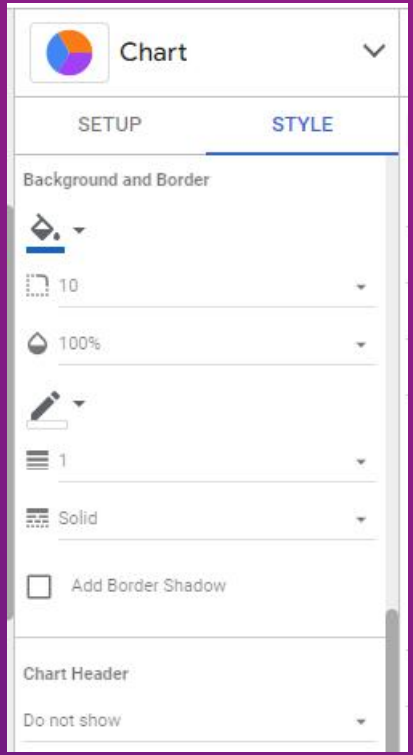
DigitalSkola

3. At Chart STYLE, select applicable color, font, border, etc. as required.







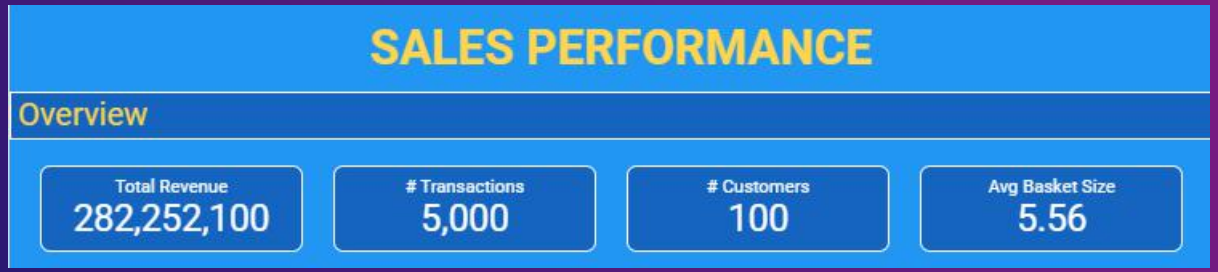


DigitalSkola

Mock Up Design

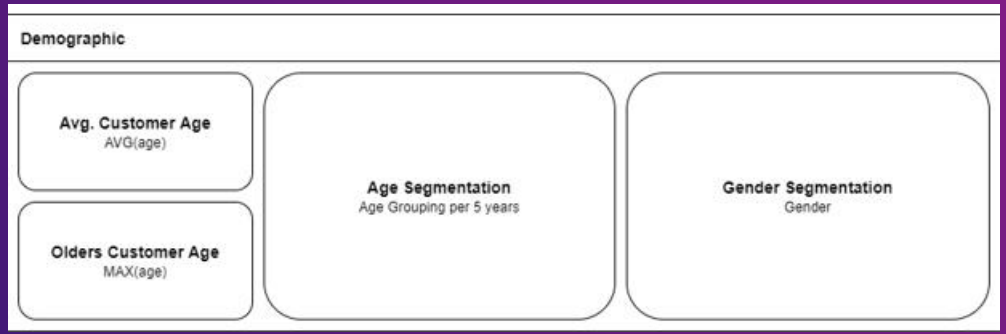
In this Overview part all chart are using Scorecards.

Dashboard Report Google Studio



DigitalSkola

Mock Up Design

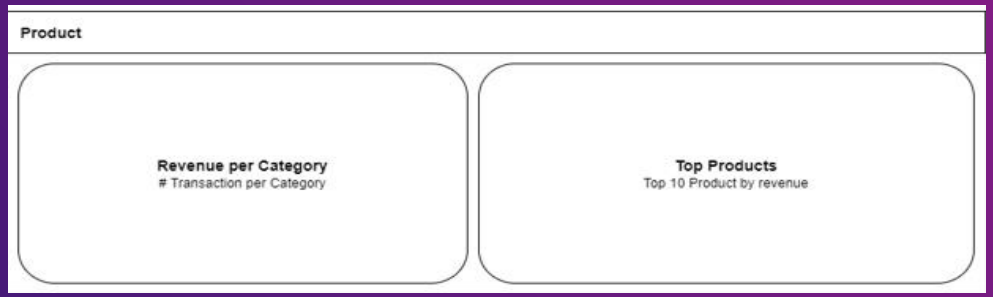


For Demographic, applicable charts are Scorecards, Bar, and Pie.

Dashboard Report Google Studio

DigitalSkola

Mock Up Design



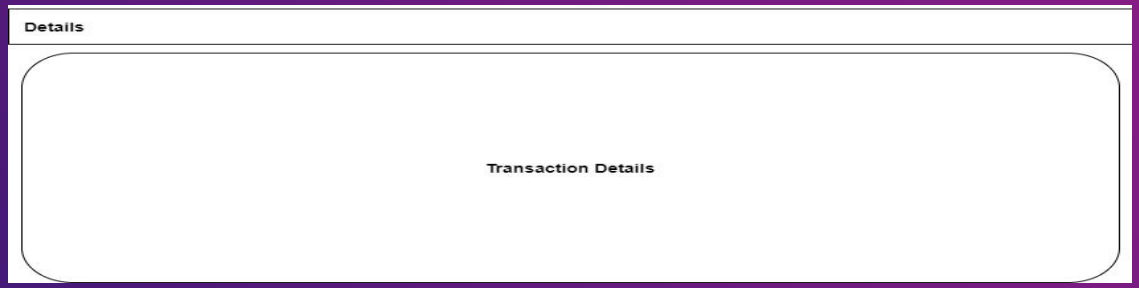
For Product section, applicable charts are Pie and Table.

Dashboard Report Google Studio

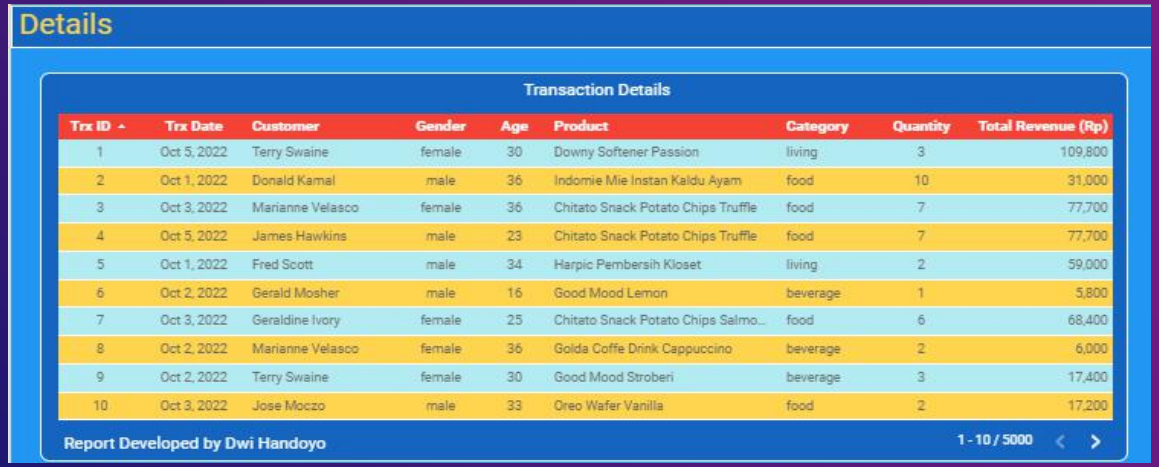
	Product	Category	Total Revenue (Rp)
1.	Downy Softener Passion	living	35,538,600
2.	Kraft Keju Quick Melt	food	25,118,700
3.	Harpic Pembersih Kloset	living	20,738,500
4.	So Klin Detergent Cair Anti Bacterial	living	16,298,500
5.	Nice Facial Tissue	living	15,628,800
6.	Sari Roti Tawar Kupas	food	13,357,500
7.	Aqua Galon 19L	beverage	13,347,600
8.	Chitato Snack Potato Chips Beef Barbeque	food	10,670,400
9.	Chitato Snack Potato Chips Sourcream Onion	food	10,103,600
10.	Chitato Snack Potato Chips Mi Goreng	food	9,073,500

DigitalSkola

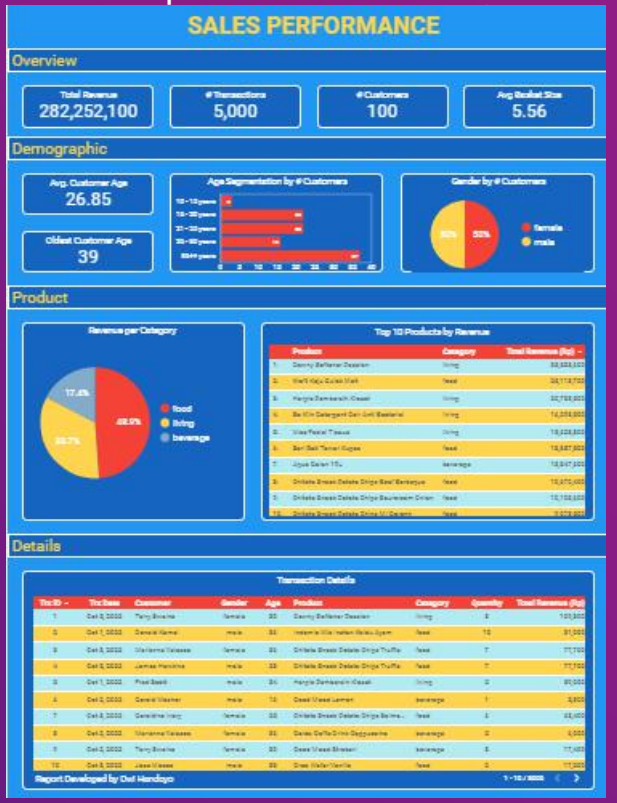
Mock Up Design



Dashboard Report Google Studio



Final Report



LINUX AND SYSTEM ADMINISTRATION

Linux is a free, open source version of Unix. It is an operating system alternative to MS Windows and MacOS. Linux is a multi-users and multi-tasks operating system, where many people can run multiple different applications on single computer at the same time.

Comparison list between Linux and Unix:

UNIX	LINUX
AT&T Bell Labs team led by Ken Thompson released Unix operating system in 1960	Linus Torvalds built Linux operating system at University of Helsinki in 1991
Source code is proprietary	Source code is public
Shell default is Bourne Shell	Shell default is BASH
Longer waiting time to get proper patch to fix bugs	Threat detection and solution is very fast
Versions: HP-UX, MacOS, Oracle Solaris, etc.	Distro: Redhat, Ubuntu, OpenSuse, etc.

Linux Distro

A distro or distribution is an operating system created from collection of software based on Linux kernel.

Linux Kernel

The Linux Kernel is the main component of the Linux Operating System. The kernel is the main interface between computer hardware and the system or application software, in charge of managing resource usage efficiently.

Logging System

The system log provides a lot of diagnostic information about our computer. Starting from activities in the kernel to user actions are logged by the Linux system log, allowing us to see almost all the actions performed on the server. A dedicated directory to keep logs is /var/logs. This directory contains logs from OS, services and applications running on the system.

Linux Installation

1. MacOS Environment

MacOS operating system is a version of Unix, the terminal in MacOS is already Unix system, so there is no need to install Linux separately. Various Linux commands can bed run on MacOS terminal directly.

2. Virtual Machine (VM)

Local VM enables the posibility of running a guest operating system (Linux distro) on top of another operating system (Windows). A virtual machine is a digital version of a physical computer. From the guest's viewpoint, it appears to be running on its very own PC. It's actually running in a VM, which is a limited virtual sub-system of local Personal Computer. The two most popular desktop VMs are VMware Workstation and Oracle VirtualBox.

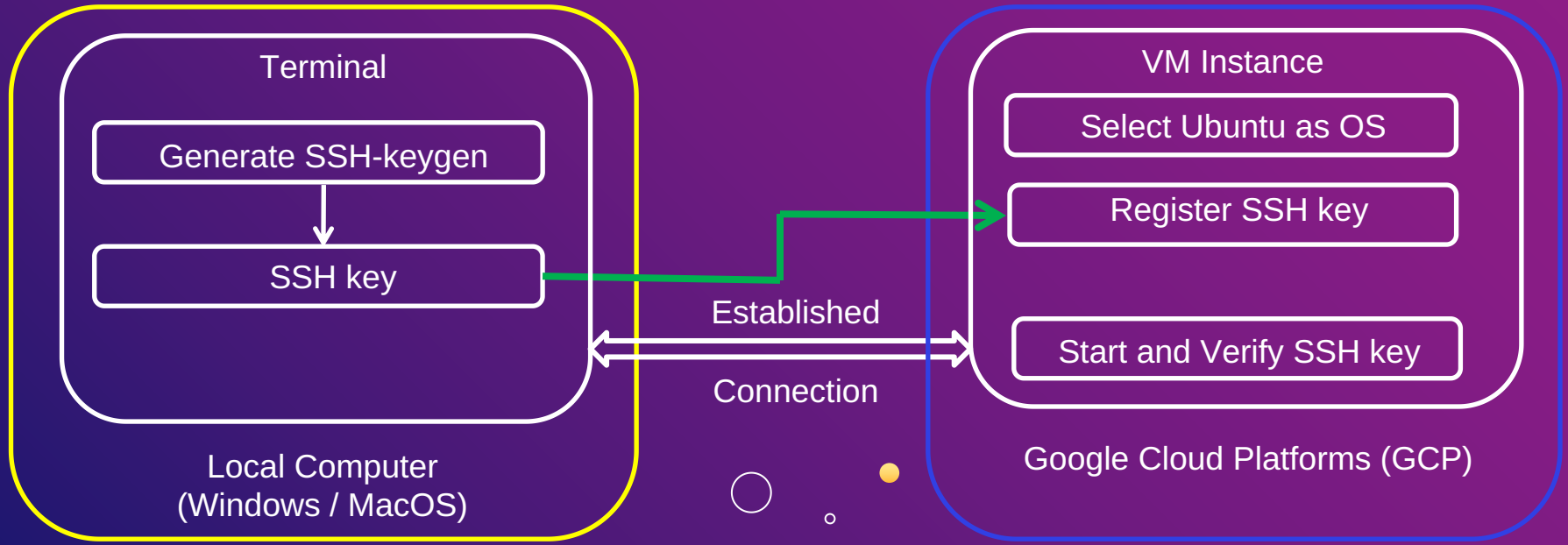
3. Dual Booting

Installing multiple operating systems, for example Linux and Windows on a single computer, and being able to choose which one to boot. This means both operating systes can not run at the same time.

DigitalSkola

4. Virtual Machine on Cloud

No need to install Linux at local, just direct connect to cloud service with SSH key and create VM Instance at cloud such as Google Cloud Platforms (GCP) to host Linux OS.



SSH stands for "Secure Shell." The SSH protocol is designed as a secure alternative to remote shell protocol. It uses client-server concept, where the client and the server communicates over a secure channel, which is by using SSH Keys.

Generate SSH Keys using the following command in command prompt (terminal) of local/client computer.

```
ssh-keygen -t rsa -f ~/.ssh/compute_engine -C INSERT_USERNAME_HERE -b 2048
or
ssh-keygen -t rsa -C INSERT_USERNAME_HERE -b 2048
```

The SSH key used for accessing remote servers on Compute Engine GCP (Google Cloud Platforms). The SSH Keys listed above must be entered into the settings in the GCP so that remote access can be done from a local computer. After it is entered, use the following command to access the remote server.

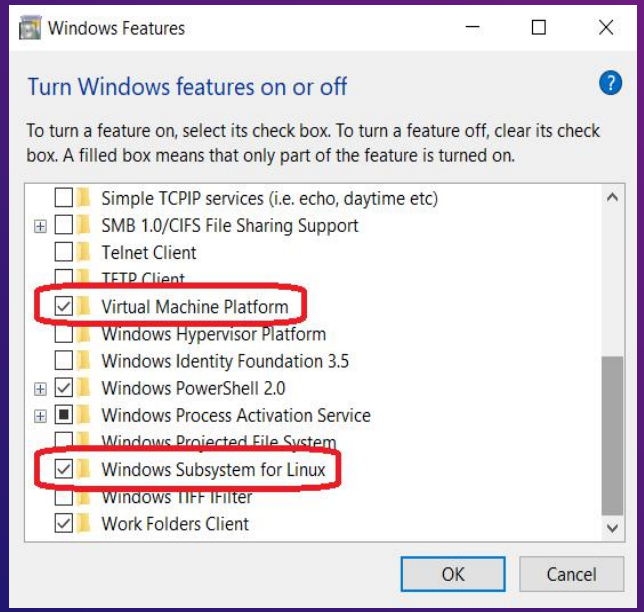
```
ssh -i ~/.ssh/compute_engine supposeradiktas@34.101.43.214
```

The path, ~/.ssh/compute_engine can be modified if necessary, according to the path of folder where we store the ssh keys in our computer, and make sure to use the appropriate username. The IP used is external IP that might be changed whenever the machine is turned off or stopped.

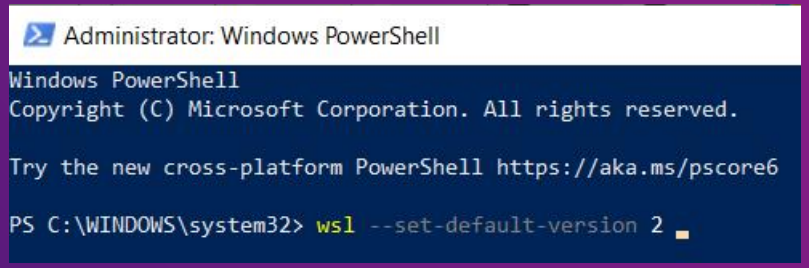
5. Windows Environment

Use Windows Subsystem for Linux (WSL) to lets users run a Linux environment directly on Windowsvita terminal (PowerShell or command prompt). It's a full Linux OS running inside Windows so we can use the same apps and files seamlessly.

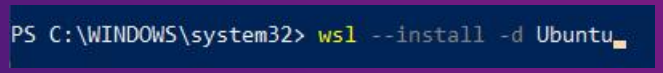
Select Virtual Machine Platform and Windows Subsystem for Linux in Windows Features.



Set WSL version 2 (WSL2) by typing command below in Windows PowerShell as Administrator.



Install Linux Ubuntu as per below command.



Linux Directories

/ – Root Directory

All Linux files and directories are located under 'root' ('/').

/bin – Binaries

The '/bin' directory contains the executable files of various shell commands such as ls, cps, cds, etc.

/dev – Device Files

This dir usually contains device-related files, such as /dev/null, /dev/zero, /dev/random.

/etc – Configuration Files

The main configuration of the system, such as config, password & system admin.

/usr – User Binaries and Program Data

Consists of executable files, libraries, and program resources. Usually access is limited. Example files: /usr/bin, /usr/lib, /usr/share.

/home – User Personal Data
Personal data of each user

/lib – Shared Libraries
Libraries are code used by file binaries in program execution

/sbin – System Binaries
Same as '/bin' but only root/sudo/admin user can access

/tmp – Temporary Files
Save temporary files that will be deleted when the system is restarted

/var – Variable Data Files
Save logs, user tracking, cache created by the program

Linux Basic Commands

To run Linux commands, it is highly recommended to use Command-Line Interface (CLI). We can use terminal such as command prompt or PowerShell at local as interface by typing “wsl”.

Many Linux commands are very useful for us to do our job as a data engineer. These commands can help us when interacting with the cloud server, deploy, update codes, install or update software and also package modules.

Linux Command	Description
cd <directory_name>	Move to a specific directory
ls -al	Displays all files in detail on the directory
pwd	Displays the current working directory
mkdir <directory_name>	Create a new directory

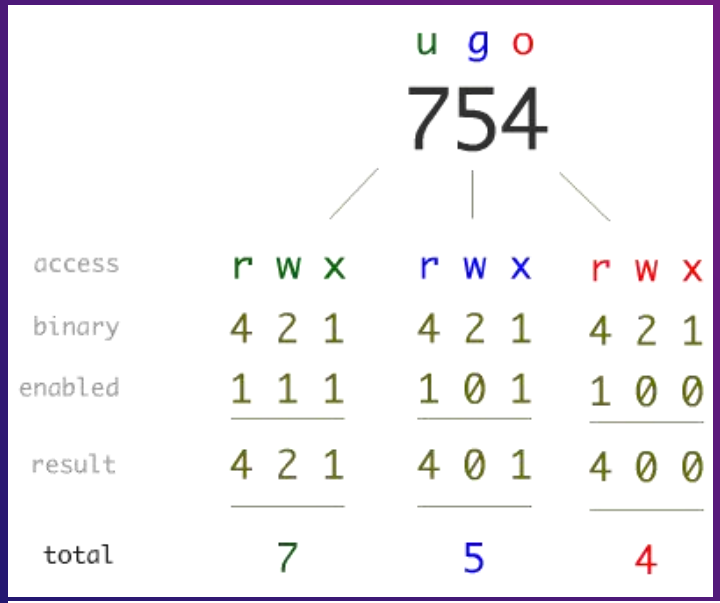
DigitalSkola

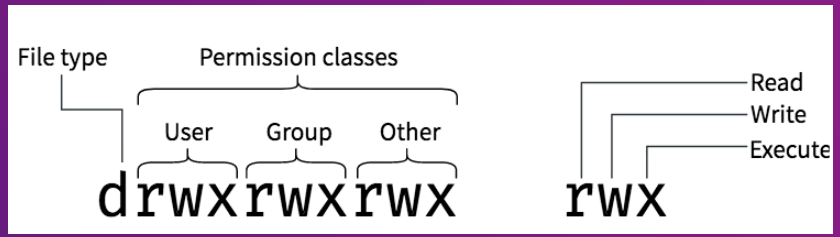
Linux Command	Description
touch <filename>	Create a new file.
rm <filename>	Delete a file.
rm -rf <directory name>	Recursively delete directories.
cp <filename 1> <filename 2>	Copying files, can also be used for copy files to a different directory.
mv <directory 1/filename 1> <directory 2>	Move files to another directory.
cat <filename>	Displays the contents in the file.
history	View all command history.
grep <pattern name> <filename>	Searches for the matching pattern in the file.
find /home/jojo -name 'prefix*'	Search for files in /home/jojo with certain prefix
sudo (SuperUser Do)	used to perform tasks that require admin access
chmod	To change access permission files or directories
chown	To change or transfer ownership of files or directories

DigitalSkola

File Access Permission

In a Linux system, basically there are two levels of users, namely superuser and other users. Type of permission given are 'r' (user can read file), 'w' (user can write/modify file), and 'x' (user can execute file). If 'rwx' replaced by '-', no permission is given.





	4	2	1	
0	-	-	-	no permissions
1	-	-	x	only execute
2	-	w	-	only write
3	-	w	x	write and execute
4	r	-	-	only read
5	r	-	x	read and execute
6	r	w	-	read and write
7	r	w	x	read, write and execute

File type code:
'd' for directory
'-' for file

DigitalSkola

Users, Groups, and Others

User – permission only applies to the owner of the file or directory.
Group – permissions only apply to groups that have been assigned to a file or directory.
Other - permissions apply to all other users on the system.

Create files then give permissions with chmod command

```
handoyo@LAPTOP-Q5ACFH2M:~/baru$ cd ..
handoyo@LAPTOP-Q5ACFH2M:~ $ touch tes1.txt
handoyo@LAPTOP-Q5ACFH2M:~ $ touch tes2.txt
handoyo@LAPTOP-Q5ACFH2M:~ $ touch tes3.txt
handoyo@LAPTOP-Q5ACFH2M:~ $ touch tes4.txt
handoyo@LAPTOP-Q5ACFH2M:~ $ chmod 777 baru
handoyo@LAPTOP-Q5ACFH2M:~ $ chmod 775 tes1.txt
handoyo@LAPTOP-Q5ACFH2M:~ $ chmod 755 tes2.txt
handoyo@LAPTOP-Q5ACFH2M:~ $ chmod 664 tes3.txt
handoyo@LAPTOP-Q5ACFH2M:~ $ chmod 644 tes4.txt
handoyo@LAPTOP-Q5ACFH2M:~ $
```

Check permissions with ls -l command

```
handoyo@LAPTOP-Q5ACFH2M:~ $ ls -l
total 4
drwxrwxrwx 2 handoyo handoyo 4096 Oct 10 14:42 baru
-rwxrwxr-x 1 handoyo handoyo    0 Oct 15 11:35 tes1.txt
-rwxr-xr-x 1 handoyo handoyo    0 Oct 15 11:35 tes2.txt
-rw-rw-r-- 1 handoyo handoyo    0 Oct 15 11:37 tes3.txt
-rw-r--r-- 1 handoyo handoyo    0 Oct 15 11:38 tes4.txt
handoyo@LAPTOP-Q5ACFH2M:~ $
```

Crontab

Crontab is a list of commands that we want to run regularly following a certain schedule. Crontab stands for "cron table" because it uses the cron scheduler to run the program. Cron itself is named after "chronos," the Greek word for time. Cron is a system process that will automatically perform tasks according to a predetermined schedule set. This schedule is called crontab, which is also the name of the program used to edit that schedule.

Create Crontab - An Example

Create a new directory named linux
~$ mkdir && cd linux

Create executable file .sh
~$ touch hello_world.sh

Go to file
~$ vim hello_world.sh

Press the “i” key on the keyboard to insert
~$ #!/bin/bash
~$ echo `date` - "Hello World" >> /path as per local directory/linux/hello_world.txt

The above command will create a file in the local directory called hello_world.txt which contains date and time when the script was executed and the text Hello World.

Then press ESC, then press :wq to save the file. Set schedule of cron

~$ sudo crontab -e
Enter your computer password, sudo means we will execute it as superuser/admin after that schedule cron with the following cmd

~$ * * * * * /path as per local directory/linux/hello_world.sh
Then press ESC, then press :wq to save the file. Check if cron is installed successfully

~$ sudo crontab -l

Cron will be executed every minute.

DigitalSkola

Shell Script Programming

Shell scripts are computer programs designed to be run by Linux shell, a command-line interpreter.

```
handoyo@LAPTOP-Q5ACFH2M: $ PS1="$ "; export PS1
handoyo@LAPTOP-Q5ACFH2M: $ echo '#!/bin/sh' > my-script.sh
handoyo@LAPTOP-Q5ACFH2M: $ echo 'echo Hello World' >> my-script.sh
handoyo@LAPTOP-Q5ACFH2M: $ chmod 755 my-script.sh
handoyo@LAPTOP-Q5ACFH2M: $ ./my-script.sh
Hello World
```

Accessing Values

```
handoyo@LAPTOP-Q5ACFH2M: $ #!/bin/sh
handoyo@LAPTOP-Q5ACFH2M: $ NAME='Godam'
handoyo@LAPTOP-Q5ACFH2M: $ echo $NAME
Godam
```

Command-Line Arguments

```
handoyo@LAPTOP-Q5ACFH2M: $ #!/bin/sh
handoyo@LAPTOP-Q5ACFH2M: $ echo "File Name: $0"
File Name: -bash
handoyo@LAPTOP-Q5ACFH2M: $ echo "First Parameter: $1"
First Parameter:
handoyo@LAPTOP-Q5ACFH2M: $ echo "Second Parameter: $2"
Second Parameter:
```

Basic Operations

```
handoyo@LAPTOP-Q5ACFH2M: $ #!/bin/sh
handoyo@LAPTOP-Q5ACFH2M: $ val='epr 2 + 2'
handoyo@LAPTOP-Q5ACFH2M: $ echo "Total value : $val"
Total value : epr 2 + 2
```

THANK YOU