Data Analytics Pipeline

Introduction to Data Analysis With Excel and SQL

Data Science Learning Studio

09 January 2022

Hello

Bagoes Suryo Nugroho

Panggil aja **Mas Goes**

Aktivitas Saat Ini

1. **Data Engineer** at Lion Parcel
2. **Strategic Partnership** at DSI
3. **Agile Innovation Facilitator** at CIAS
4. **Individual Data Consultant**

DATA SCIENCE INDONESIA

Data Science Learning Studio

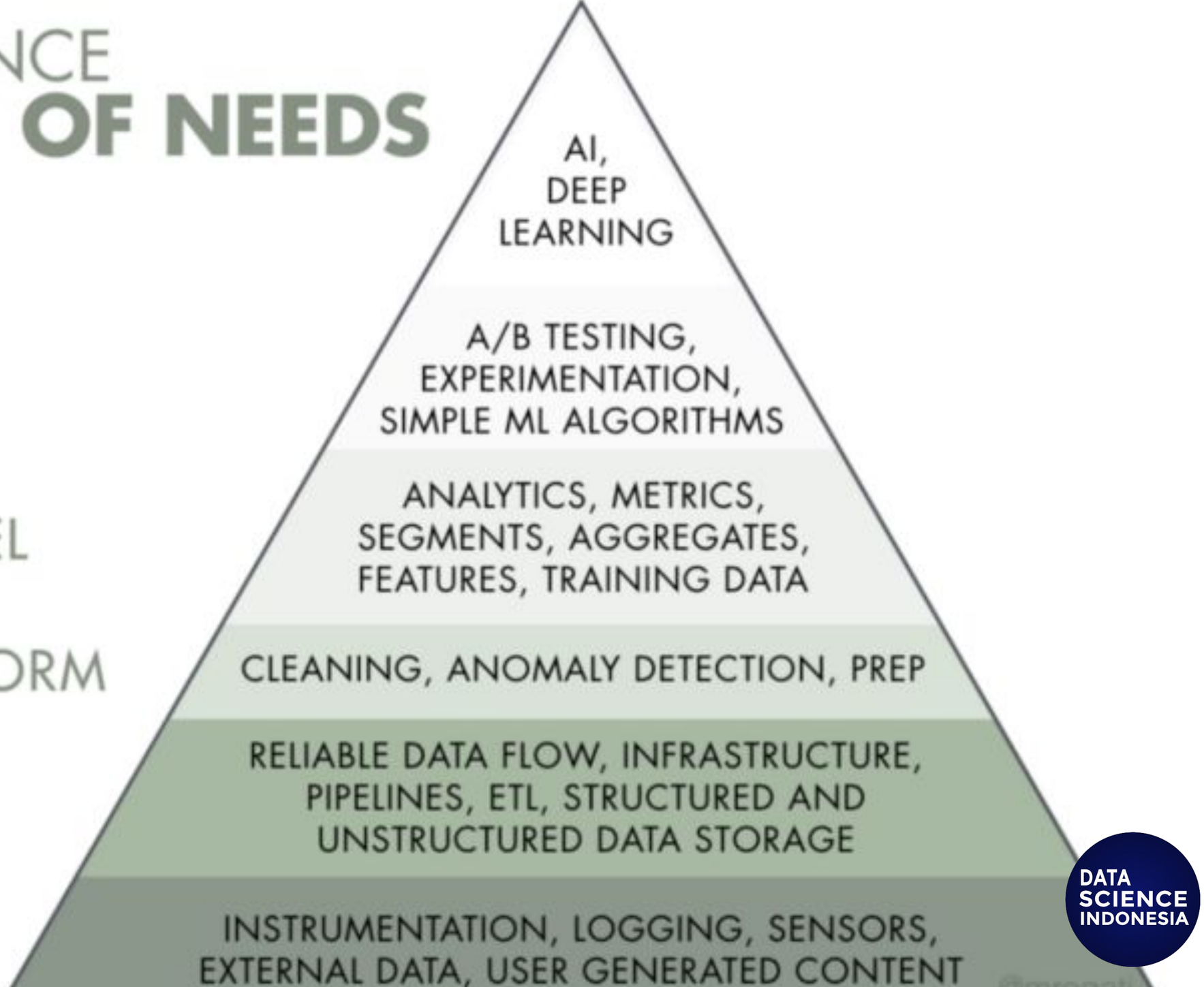
Bab Materi

1. **Introduction** to SQL and Excel
2. **Introduction to Excel** for Analysis
3. **Introduction to SQL** for Analysis
4. **Hands on !**

   Data Analysis Using SQL and Excel

DATA SCIENCE INDONESIA

THE DATA SCIENCE HIERARCHY OF NEEDS



LEARN/OPTIMIZE

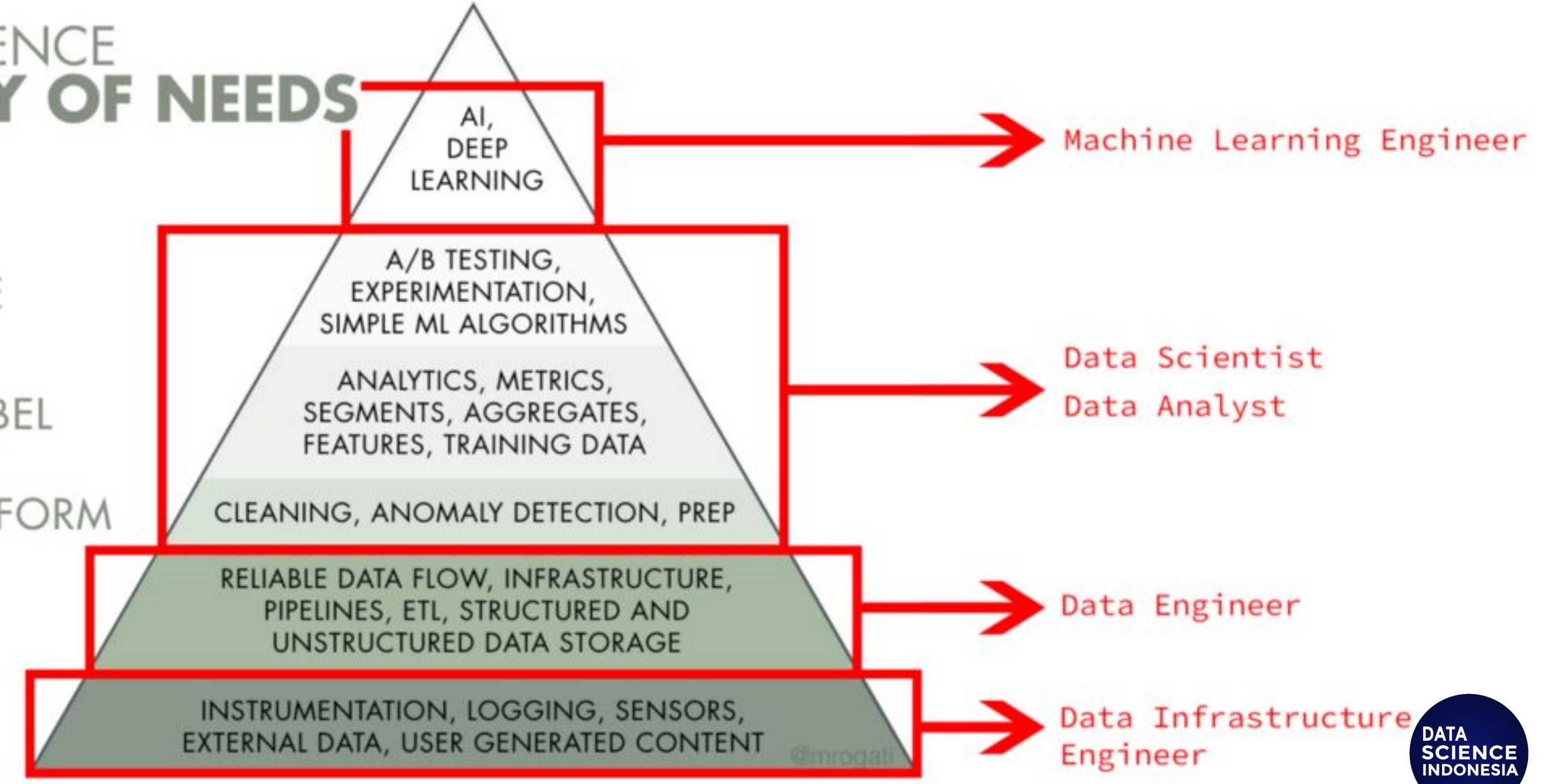
AGGREGATE/LABEL

EXPLORE/TRANSFORM

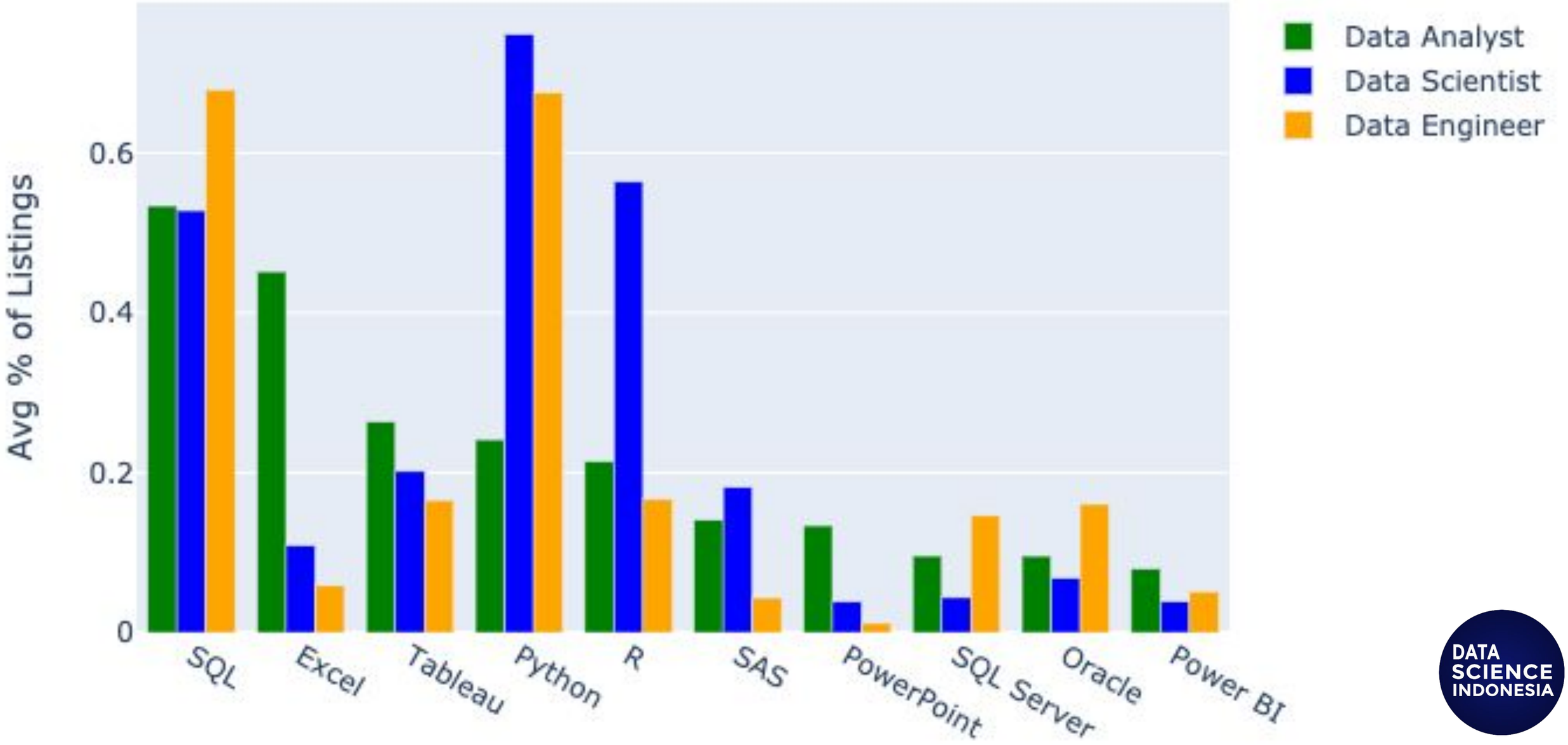
MOVE/STORE

COLLECT

DATA SCIENCE INDONESIA


ENCE
Y OF NEEDS
BEL
FORM
AI,
DEEP
LEARNING
A/B TESTING,
EXPERIMENTATION,
SIMPLE ML ALGORITHMS
ANALYTICS, METRICS,
SEGMENTS, AGGREGATES,
FEATURES, TRAINING DATA
CLEANING, ANOMALY DETECTION, PREP
RELIABLE DATA FLOW, INFRASTRUCTURE,
PIPELINES, ETL, STRUCTURED AND
UNSTRUCTURED DATA STORAGE
INSTRUMENTATION, LOGGING, SENSORS,
EXTERNAL DATA, USER GENERATED CONTENT
@mrogati
Machine Learning Engineer
Data Scientist
Data Analyst
Data Engineer
Data Infrastructure
Engineer

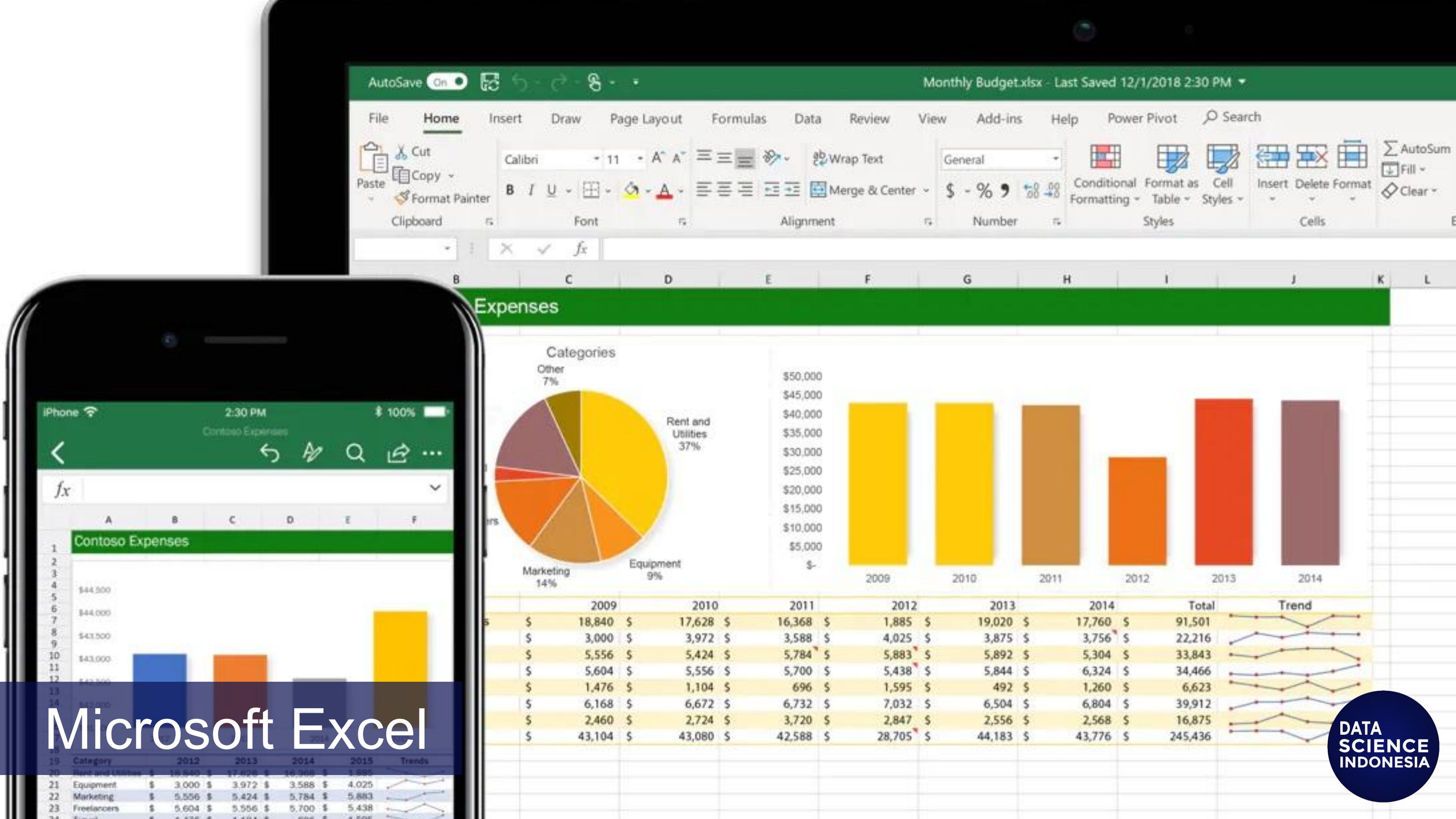
Most Common Keywords for Data Analyst vs. Data Scientist, Data Engineer



DATA SCIENCE INDONESIA

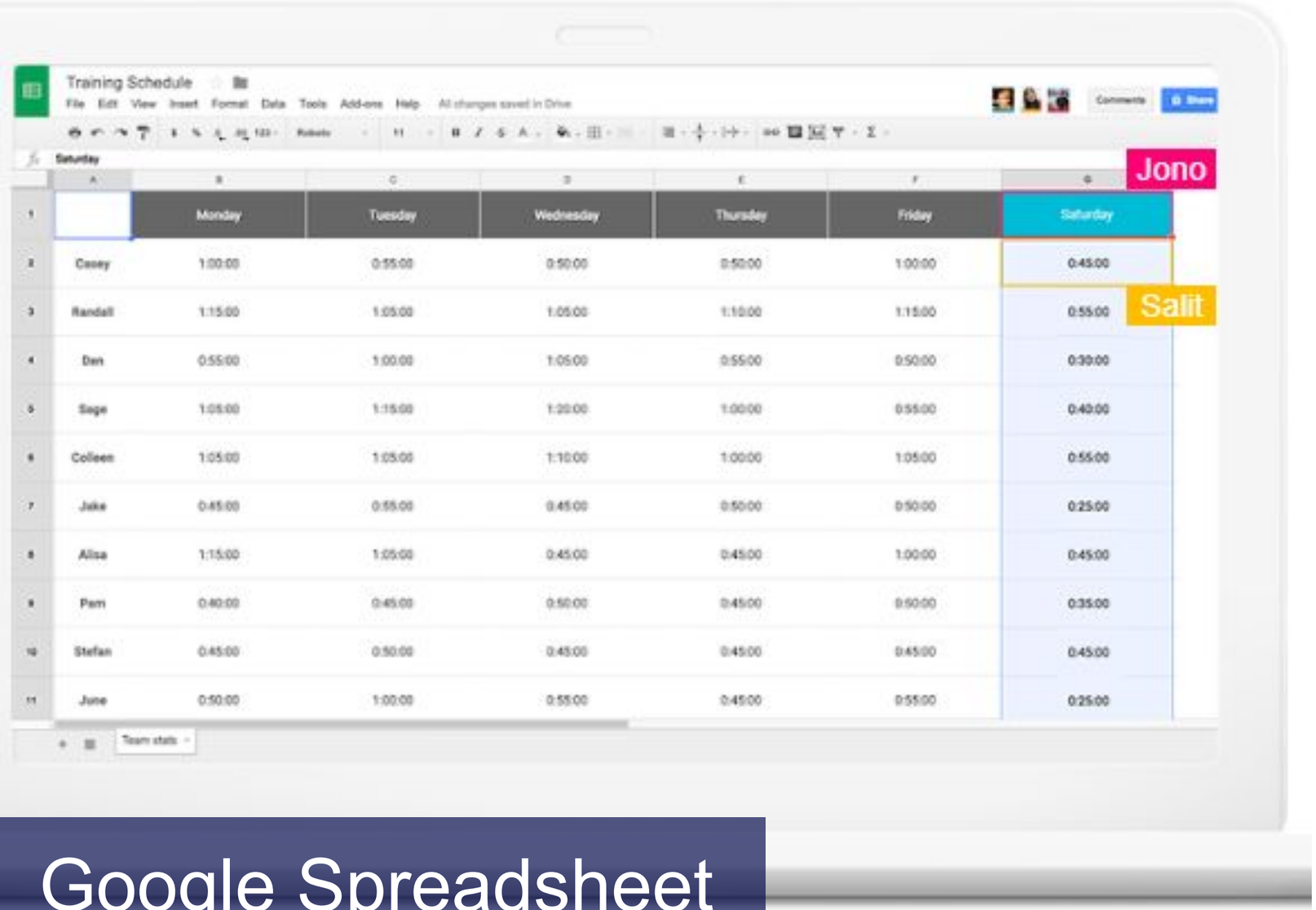
Excel

DATA SCIENCE INDONESIA



Microsoft Excel

DATA SCIENCE INDONESIA



	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Casey	1:00:00	0:55:00	0:50:00	0:50:00	1:00:00	0:45:00
Randall	1:15:00	1:05:00	1:05:00	1:10:00	1:15:00	0:55:00
Dan	0:55:00	1:00:00	1:05:00	0:55:00	0:50:00	0:30:00
Sage	1:05:00	1:15:00	1:20:00	1:00:00	0:55:00	0:40:00
Colleen	1:05:00	1:05:00	1:10:00	1:00:00	1:05:00	0:55:00
Jake	0:45:00	0:55:00	0:45:00	0:50:00	0:50:00	0:25:00
Alisa	1:15:00	1:05:00	0:45:00	0:45:00	1:00:00	0:45:00
Pam	0:40:00	0:45:00	0:50:00	0:45:00	0:50:00	0:35:00
Stefan	0:45:00	0:50:00	0:45:00	0:45:00	0:45:00	0:45:00
June	0:50:00	1:00:00	0:55:00	0:45:00	0:55:00	0:25:00

Google Spreadsheet

Google Spreadsheet

DATA SCIENCE INDONESIA

Microsoft Excel

1. Berbayar dan Offline*
   File disimpan secara offline di local
2. Kurang baik untuk kerjaan Kolaboratif. Tidak bisa sharing sheets.
3. Kemampuan Analisis dan Visualisasinya yang sangat baik karena ada banyak fungsi yang ada

Google Sheets

1. Gratis dan Online.
   File disimpan di Google Drive
2. Lebih baik untuk kerjaan kolaboratif. Dapat sharing sheets.
3. Kemampuan Analisis tidak selengkap Ms Excel dan beberapa Visualisasi dilakukan secara manual

DATA SCIENCE INDONESIA

Format CSV - XLSX

DATA SCIENCE INDONESIA

CSV

1. Hanya menyimpan data tanpa ada formatting, operasi, macro,dll
2. Import Data CSV lebih cepat,mudah dan memakan memori yang sedikit
3. Dapat dibuka di Excel, Notepad, dan Text Editor lainnya

XLSX

1. Dapat menyimpan operasi, macro, tabel, formatting,dll
2. Import Data XLSX lebih lambat,kompleks, dan memakan memori yang banyak
3. Hanya dapat dibuka di Excel

DATA SCIENCE INDONESIA

Invoice 29 Juli .XLSX Saved to Drive

File Edit View Insert Format Data Tools Help Last edit was on September 4, 2020

C15

TOTAL	ANGKA
PENDAPATAN KOTOR	#NAME?
MODAL BERSIH	#NAME?
PROFIT PUSAT	#NAME?
PROFIT MITRA	#NAME?
ONGKIR	#NAME?
PENDAPATAN BERSIH	Rp 1,525,872

Profit tambahan dari tokped

Profit mitra 11650

Modal actual		1300000	
Selisih	#NAME?		
Profit	#NAME?		#NAME?
Ongkir	#NAME?		
Kurir	90000		
	#NAME?		
Bensin	0		
	#NAME?		
	0		
Makan	30000		
Bensin	15000		
Parkir	10000		
	55000		
	#NAME?		
Kelebihan bombay		0	

plastik
bensin
makan
parkir
supir

modal + petty Rp
modal actual Rp

Rp - selisih

Data di Excel?

TOTAL FINAL | Database | DATABASE HARGA | LIST BELANJA | AWAL | TEMPLATE PUSAT | Dita | Intan

DATA SCIENCE INDONESIA

- Users (7)
  - AHMED
  - ALI
  - FULL_ACCES
  - OSMAN
  - PUBLIC
  - SUHAIB
  - SYSDBA
- Unicorn
- Employee
  - Query Window
  - Tables (12)
    - COUNTRY
    - JOB
    - DEPARTMENT
    - CUSTOMER
    - EMPLOYEE
    - PROJECT
    - SALES
    - EMPLOYEE_PROJECT
    - PROJ_DEPT_BUDGET
    - SALARY_HISTORY
    - SALES2
    - COUNTRY2
  - Generators (2)
    - EMP_NO_GEN
    - CUST_NO_GEN
  - Triggers (4)
  - Views (1)
  - Stored Procedures (11)

Employee: Select first 1000 from COUNTRY | Employee: Select first 1000 from DEPARTMENT

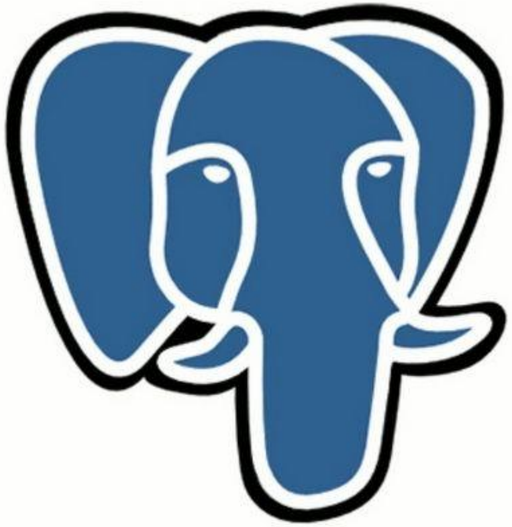
```
1  select first 1000 * from DEPARTMENT;
2
3  select first 1000 * from COUNTRY;
4
```

Standar Query Languange (SQL)

Result # 1 | Result # 2

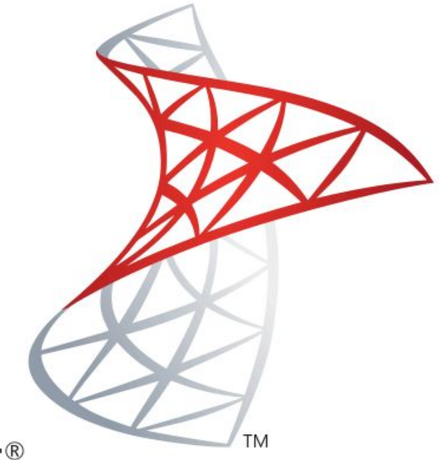
DEPT_NO	DEPARTMENT	HEAD_DEPT	MNGR_NO	BUDGET
000	Corporate Headquarters		105	0.00$
100	Sales and Marketing	000	85	0.00$
600	Engineering	000	2	0.00$
900	Finance	000	46	0.00$
180	Marketing	100		
620	Software Products Div.	600		
621	Software Development	620		

DATA SCIENCE INDONESIA



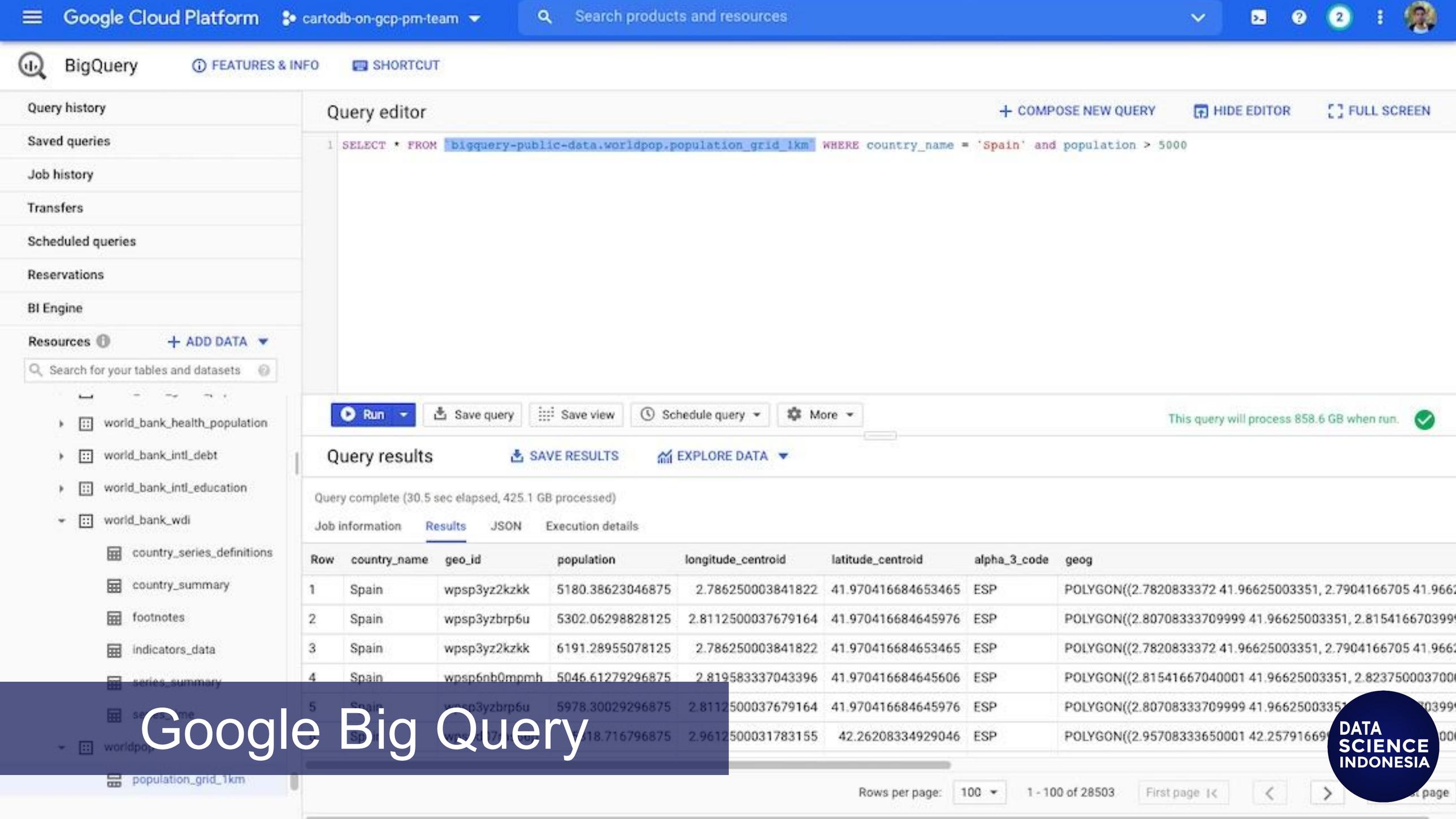
PostgreSQL

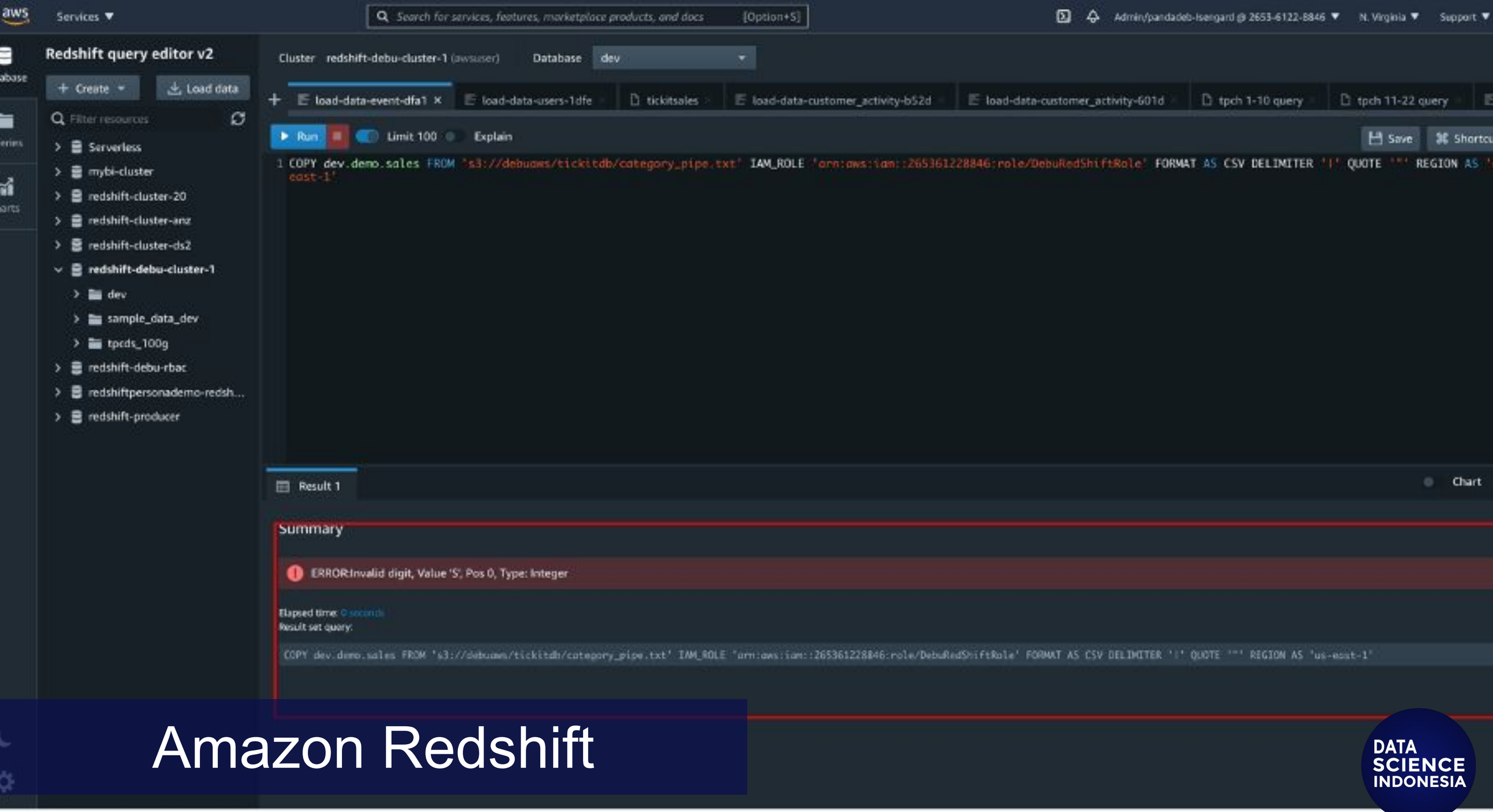
Jenis Database SQL



Microsoft® SQL Server™

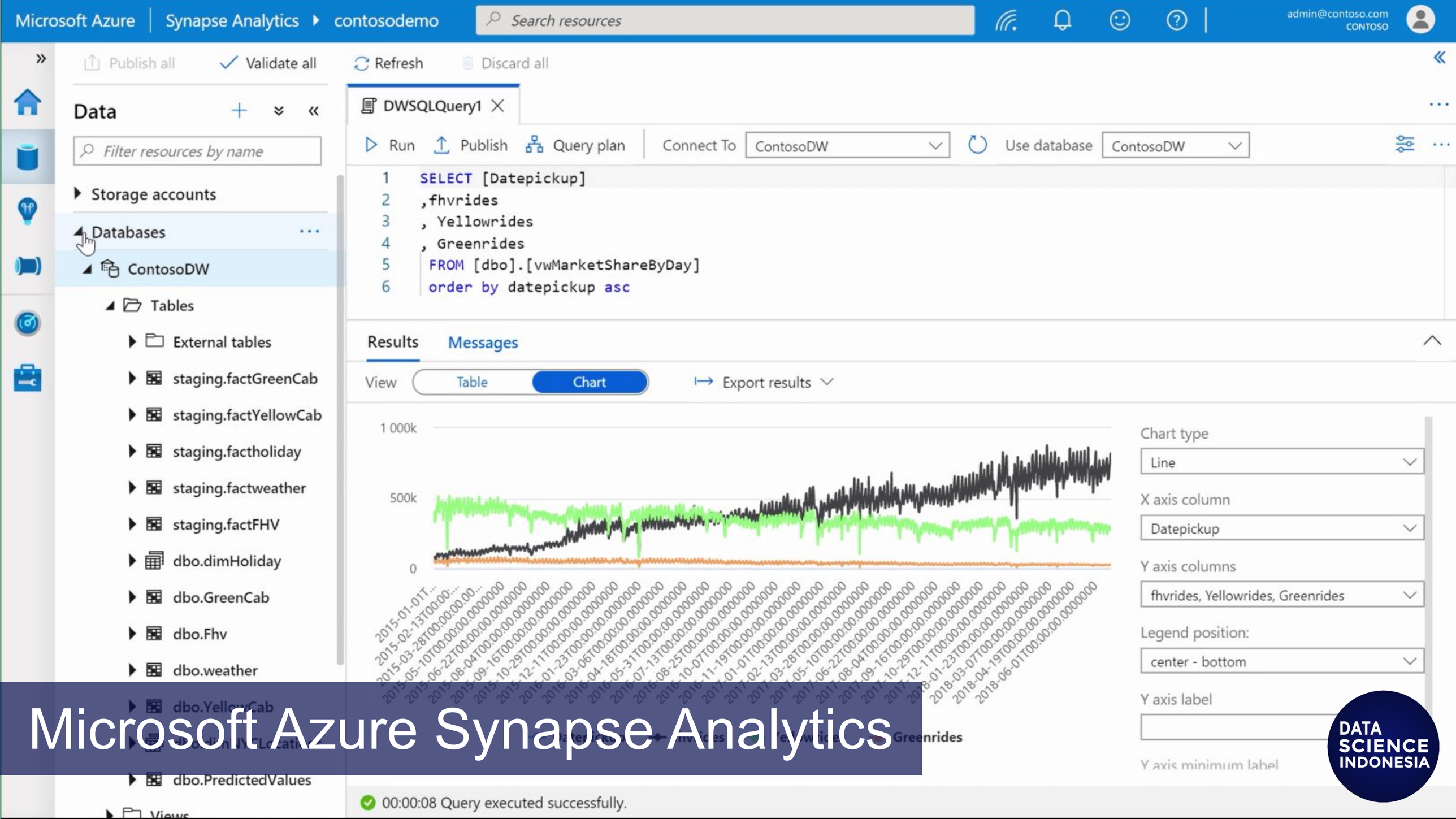
DATA SCIENCE INDONESIA


Google Cloud Platform
cartodb-on-gcp-pm-team
Search products and resources
BigQuery
FEATURES & INFO
SHORTCUT
Query history
Saved queries
Job history
Transfers
Scheduled queries
Reservations
BI Engine
Resources
ADD DATA
Search for your tables and datasets
world_bank_health_population
world_bank_intl_debt
world_bank_intl_education
world_bank_wdi
country_series_definitions
country_summary
footnotes
indicators_data
series_summary
population_grid_1km
Query editor
COMPOSE NEW QUERY
HIDE EDITOR
FULL SCREEN
1 SELECT * FROM `bigquery-public-data.worldpop.population_grid_1km` WHERE country_name = 'Spain' and population > 5000
Run
Save query
Save view
Schedule query
More
This query will process 858.6 GB when run.
Query results
SAVE RESULTS
EXPLORE DATA
Query complete (30.5 sec elapsed, 425.1 GB processed)
Job information
Results
JSON
Execution details
Row country_name geo_id population longitude_centroid latitude_centroid alpha_3_code geog
1 Spain wpsp3yz2kzkk 5180.38623046875 2.786250003841822 41.970416684653465 ESP POLYGON((2.7820833372 41.96625003351, 2.7904166705 41.966
2 Spain wpsp3yzbrp6u 5302.06298828125 2.8112500037679164 41.970416684645976 ESP POLYGON((2.80708333709999 41.96625003351, 2.8154166703999
3 Spain wpsp3yz2kzkk 6191.28955078125 2.786250003841822 41.970416684653465 ESP POLYGON((2.7820833372 41.96625003351, 2.7904166705 41.966
4 Spain wpsp6nb0mpmh 5046.61279296875 2.819583337043396 41.970416684645606 ESP POLYGON((2.81541667040001 41.96625003351, 2.8237500037000
5 Spain wpsp3yzbrp6u 5978.30029296875 2.8112500037679164 41.970416684645976 ESP POLYGON((2.80708333709999 41.96625003351
Spain 2.9612500031783155 42.26208334929046 ESP POLYGON((2.95708333650001 42.25791669
Rows per page: 100
1 - 100 of 28503
First page
Google Big Query
DATA SCIENCE INDONESIA



Amazon Redshift

DATA SCIENCE INDONESIA



Microsoft Azure Synapse Analytics

DATA SCIENCE INDONESIA

```
// MongoDB Playground
// To disable this template go to Settings | MongoDB | Use Default Template For Playground.

// Select the database to use.
use('test');

// Insert a few documents into the sales collection.
db.sales.insertMany([
  { '_id' : 1, 'item' : 'abc', 'price' : 10, 'quantity' : 2, 'date' : new Date('2014-03-01T08:00:00Z') },
  { '_id' : 2, 'item' : 'jkl', 'price' : 20, 'quantity' : 1, 'date' : new Date('2014-03-01T09:00:00Z') },
  { '_id' : 3, 'item' : 'xyz', 'price' : 5, 'quantity' : 10, 'date' : new Date('2014-03-15T09:00:00Z') },
  { '_id' : 4, 'item' : 'xyz', 'price' : 5, 'quantity' :  20, 'date' : new Date('2014-04-04T11:21:39.736Z') },
  { '_id' : 5, 'item' : 'abc', 'price' : 10, 'quantity' : 10, 'date' : new Date('2014-04-04T21:23:13.331Z') },
  { '_id' : 6, 'item' : 'def', 'price' : 7.5, 'quantity': 5, 'date' : new Date('2015-06-04T05:08:13Z') },
  { '_id' : 7, 'item' : 'def', 'price' : 7.5, 'quantity': 10, 'date' : new Date('2015-09-10T08:43:00Z') },
  { '_id' : 8, 'item' : 'abc', 'price' : 10, 'quantity' : 5, 'date' : new Date('2016-02-06T20:20:13Z') },
]);

// Run an aggregation to view total sales for each product in 2014.
const aggregation = [
  { $match: { d: { $gte: new Date('2014-01-01'), $lt: new Date('2015-01-01') } } },
  { $group: { _ date                                                    e', '$quantity' ] } } } }
];
db.sales.aggregate(aggregation);
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL

```
[
  {
    _id: 'abc',
```

NoSQL - MongoDB

DATA SCIENCE INDONESIA

Query Editor Query History Scratch Pad

```
1  SELECT
2      id, stockmarket, stockname, stockholder, stocktime, investoriban
3  FROM sqlshack_regex
4  WHERE investoriban ~ '^.*$'
```

Data Output Explain Messages Notifications

	id [PK] integer	stockmarket text	stockname text	stockholder text	stocktime timestamp without time zone	investoriban text
1	667	NASDAQ	Popular, Inc.	Rochette Guilfoyle	2018-05-27 16:58:37	PL19 0214 4300 8029 3204 5272 3724
2	908	NYSE	Assured Guaran...	Sammy Luety	2020-11-11 03:10:30	PK66 LIKZ DIYK NYGZ FUGM M4JJ
3	1239	NYSE	GasLog Partner...	Latisha Fumagall	2017-12-07 08:47:56	FR17 9602 2435 91IZ JZ1T E7BY F69
4	4848	NYSE	State Street Cor...	Rock Godbert	2019-08-05 13:17:30	NL42 VRGQ 7633 4289 04
5	1872	NYSE	CBS Corporation	Demetri Tynemo...	2018-02-28 21:41:46	HU15 1293 4607 8941 9706 4939 5676
6	2492	NYSE	Care.com, Inc.	Janeen Semonin	2017-11-04 07:32:33	KZ03 527J RFOL ASMP 4VLD
7	2681	NYSE	Aegon NV	Ingaborg Pinchin	2018-05-27 19:28:21	FR65 7904 0244 56CF KM7D RSLA S32
8	2917	NYSE	Innovative Indu...	Merv Brass	2020-03-24 10:06:25	LI87 4802 4WE0 MQMK NKWT C
			an Hom...	Fidelity Novic	2019-10-19 06:36:20	CZ34 6278 9268 8166 5597 5039
			EPUBLI...	Gayelord Allin	2019-03-01 16:40:43	BA28 2853 1449 4039 5790
			Bank	Bartram Marple	2020-02-18 23:26:21	KZ05 596R 810K QGZ9 CKOR
12	4942	NYSE	Brookfield DTL...	Portia MacAndie	2019-06-07 08:39:31	LV11 YREW P54N TSDQ 4DB1 Q

Penggunaan SQL

DATA SCIENCE INDONESIA

Ringkasan
Excel vs SQL

SQL

- SQL adalah **bahasa**
- Data disimpan di dalam **database**
- Untuk real project dengan data besar, SQL **berbayar**
- Ukuran data yang bisa disimpan bisa di*upgrade* sesuai kebutuhan

Excel

- Excel adalah **program**
- Data disimpan di **lokal** atau **cloud** (jika menggunakan Google Sheet)
- Untuk real project dengan data besar, Excel dapat **gratis** (Google Spreadsheet)
- Batas maksimum baris data sampai 1,2 juta

DATA SCIENCE INDONESIA

Kenapa memakai Excel vs SQL ?

SQL

- Jumlah data yang ada besar dan bergerak cepat
- SQL dibuat khusus untuk Data
- Lebih memungkinkan melaksanakan **Collaborative Work**
- Tingkat keamanan yang lebih baik

Excel

- Jumlah data yang sedkit dan relatif lambat
- Hampir seluruh orang familiar dengan Excel
- Multifungsi (Penyimpanan Data, Analisis Data, dan melakukan hal lain diluar data)
- Mudah dipelajari

DATA SCIENCE INDONESIA

Yuk Belajar **Excel**

Formula	Operasi Excel yang menggunakan rumus bawaan excel
Equation	Operasi Excel yang menggunakan cell-cell untuk menghitung suatu nilai
Vlookup	**Vertical Lookup,** mencari suatu nilai pada tabel berdasarkan acuan nilai tertentu
Pivot Table	Resume atau ringkasan singkat dari data yang kita miliki
Chart	Membuat visualisasi data yang telah kita miliki ke dalam grafik

DATA SCIENCE INDONESIA

DATA SCIENCE INDONESIA

Excel
Hands On !

Data Analytics using Excel

DSLS

DATA SCIENCE INDONESIA

DATA SCIENCE INDONESIA

Yuk Belajar SQL

Operasi SQL

- Filtering
- Ordering
- Alias
- Aggregate
- Groupby
- Subquery
- Join
- Conditional

Database SQL

- Create
- Read
- Update
- Delete

DATA SCIENCE INDONESIA

DATA SCIENCE INDONESIA

SQL
Hands On !

Data Analytics using SQL

DSLS

DATA SCIENCE INDONESIA

THANK YOU

BSN

DATA SCIENCE INDONESIA

For more information,
let’s talk!

Contact person:

Bagoes
bagoessuryo19@gmail.com

linkedin.com/in/bagoes-suryo-nugroho/