

NEWSLETTER

INTEGRASI DATA: SIRI 2

TRANSFORMASI DIGITAL MELALUI INTEGRASI DATA

Apa Itu Integrasi Data?

Model Proses Perniagaan Statistik Generik (GSBPM) mendefinisikan integrasi data di bawah fasa Proses, iaitu merujuk kepada aktiviti menggabungkan data daripada satu atau lebih sumber sebagai sebahagian daripada proses kerja statistik.

Secara umum, integrasi data merujuk kepada proses penyatuan dan penyeragaman data daripada pelbagai sumber untuk menjamin konsistensi dalam capaian dan penyampaian maklumat.

Manfaat Integrasi Data

- Menggabungkan data merentas pelbagai sistem dan agensi secara holistik
- Mengurangkan kewujudan data silo data boleh menyukarkan capaian, pemprosesan dan analisis maklumat
- Meningkatkan keberkesanan pengurusan data melalui pemusatan, penyelarasan dan penyeragaman data
- Menjamin ketepatan dan integriti data bagi menghasilkan laporan yang berkualiti serta menyokong perancangan berasaskan data

Kaedah Pengintegrasian Data

ETL/ELT

- **Extract, Transform, Load (ETL):-** Kaedah ini melibatkan pengekstrakan (*Extract*) data daripada pelbagai sumber, diikuti dengan pengharmonian dan transformasi data (*Transform*) mengikut ketetapan tertentu, sebelum dimuat naik (*Load*) ke destinasi seperti gudang data atau *data lake*.
- **Extract, Load, Transform (ELT):-** Data diekstrak (*Extract*) daripada pelbagai sumber, dimuat naik (*Load*) ke gudang data atau *data lake*, kemudian diharmonikan dan ditransformasikan (*Transform*) mengikut ketetapan tertentu.

API

- **Application Programming Interface:-** Alat yang memudahkan komunikasi antara pelbagai aplikasi atau sistem. Ia berfungsi sebagai "pengantara" yang membolehkan data atau fungsi dari satu sistem diakses oleh sistem lain tanpa perlu memahami secara terperinci sistem tersebut.

CDC

- **Change Data Capture:-** Kaedah untuk mengesan dan merekod perubahan data dalam pangkalan data—seperti tambah, kemas kini atau padam—secara masa nyata, supaya data yang berubah boleh diselaraskan ke sistem lain dengan cepat dan efisien.



Proses Pengintegrasian Data

#1 Mengenal Pasti Sumber Data

1. Mengumpul data dari pelbagai sumber seperti:
 - o Pangkalan data (SQL, NoSQL)
 - o Fail (CSV, Excel, JSON)
 - o Aplikasi (ERP, CRM)
 - o API & Perkhidmatan Web
2. Menentukan struktur dan format data yang berbeza dari pelbagai sumber.



#2 Pemetaan Data (Data Mapping)

Proses menghubungkan medan (*field*) dari sumber berbeza ke format yang telah ditetapkan

Sumber Asal	Sasaran	Rujukan	Transformasi
Nama_Klien	NamaPelanggan		(1:1)
HargaUSD	HargaMYR		Ubah kepada Kadar Pertukaran Mata Wang Malaysia
TarikhDaftar	Tarikh_Keahlian		Tukar format dari "01-Jan-2023" → "2023-01-01"
Jantina	Jantina	lu_jantina	Guna code 01 bagi lelaki Guna code 02 bagi perempuan

#3 Pembersihan Data (Data Cleaning)

a) Pengendalian Missing Data.

1. Buang data yang hilang jika bersesuaian.
2. Gantikan dengan nilai min/median/mod.
3. Pengisian dengan model statistik/*machine learning*.

NAMA	UMUR	PENDAPATAN	TARIKH_LAHIR
Ali	35	1000	03-10-1990
Abu	24	2000	04-06-2001
Ahmad	25	3000	19-06-2000
Ali	35	1000	03-10-1990

NAMA	UMUR	PENDAPATAN	TARIKH_LAHIR
Ali	35	1000	03-10-1990
Abu	24	2000	04-06-2001
Ahmad	25	3000	19-06-2000

NAMA	UMUR	PENDAPATAN	TARIKH_LAHIR
Ali	35	1000	03-10-1990
Abu	24	null	04-06-2001
Ahmad	null	3000	19-06-2000

NAMA	UMUR	PENDAPATAN	TARIKH_LAHIR
Ali	35	1000	03-10-1990
Abu	24	2000	04-06-2001
Ahmad	25	3000	19-06-2000

b) Pengendalian Duplikasi Data.

1. Kenal pasti duplikasi data
2. Buang data yang mempunyai duplikasi jika bersesuaian.

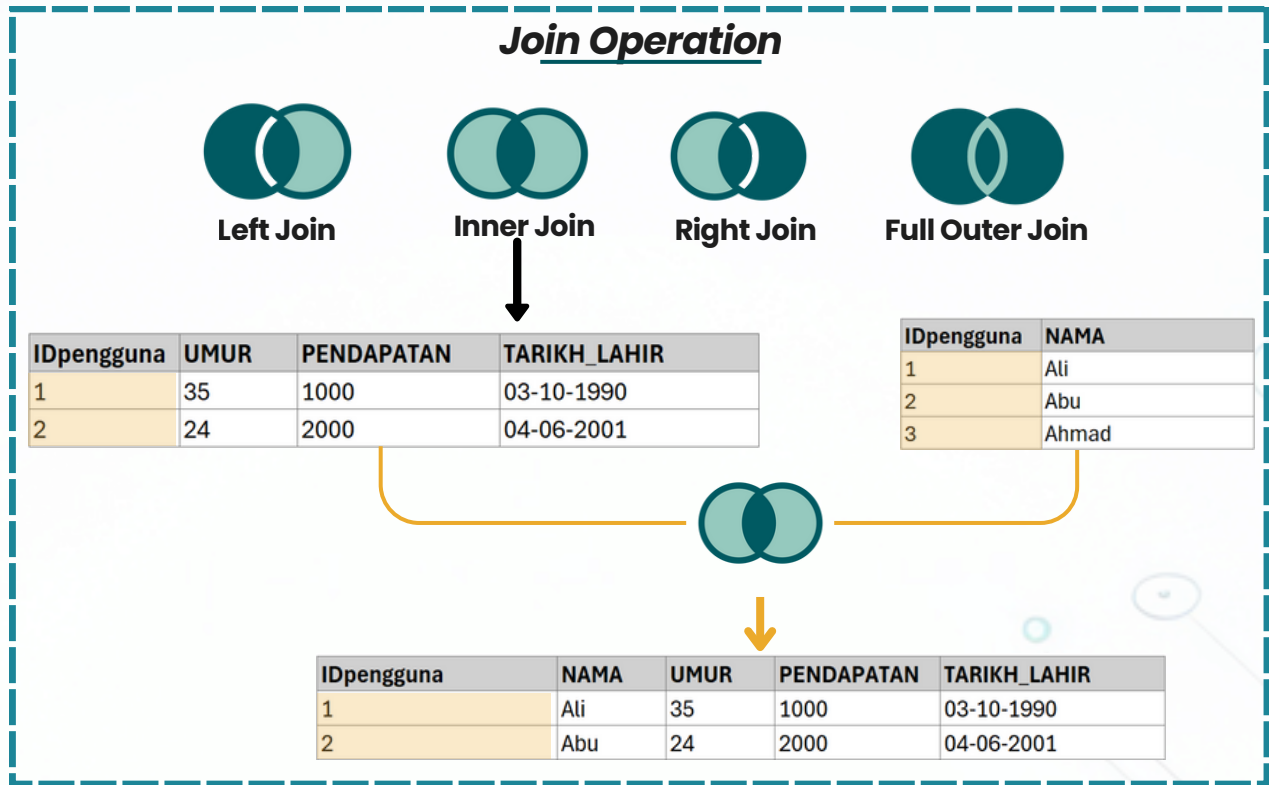


Proses Pengintegrasian Data

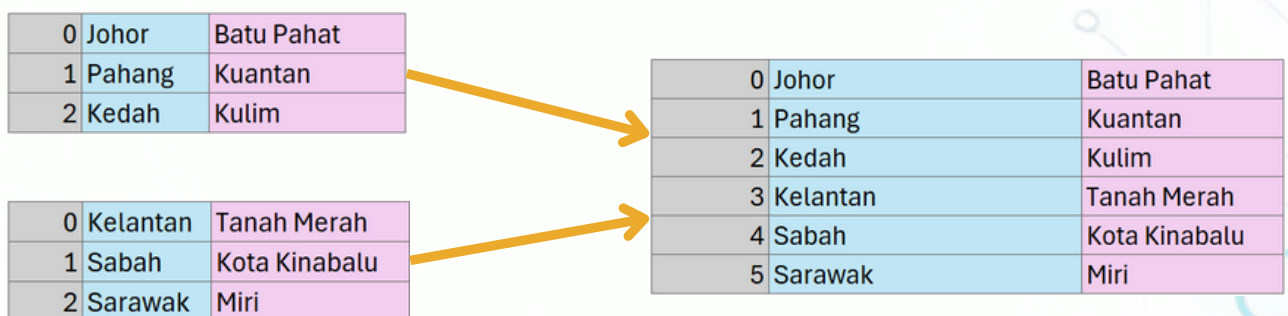
#4 Penggabungan Data (*Data Merging / Data Blending*)

1. Menggabungkan data menggunakan teknik seperti:

- *Join* – menggabungkan jadual berdasarkan ID unik (*key*).
- *Concatenation* – menyambung data secara menegak/mendatar.



Data Concatenation



#5 Penyimpanan Data (*Data Storing*)

Data yang selesai di proses perlu disimpan dalam sistem secara teratur untuk memudahkan akses dan analisis. Data ini boleh disimpan dalam:

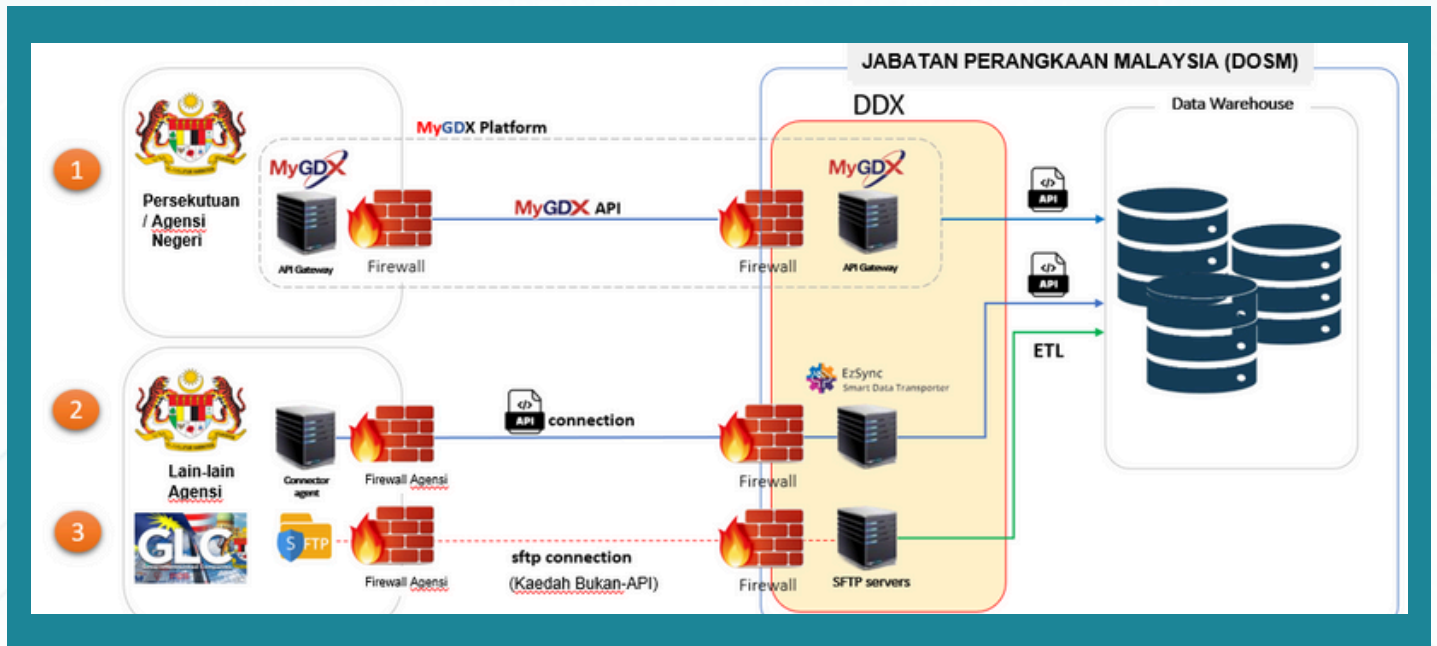
- *Data warehouse* (contoh: Snowflake, BigQuery)
- *Data lake* (contoh: Hadoop, AWS S3)
- Pangkalan data terpusat (contoh: PostgreSQL, MongoDB)



DOSM DATA EXCHANGE (DDX)

Platform perkongsian data **berpusat**.

One Stop Centre untuk integrasi data dengan DOSM.



3 PRINSIP PERKONGSIAN DATA

Agensi Sektor Awam hanya boleh berkongsi data apabila memenuhi prinsip yang berikut:



(PKPA BIL.2 /2021)

Ketelusan

- Perkongsian data berdasarkan keperluan dan persetujuan antara agensi;
- Perkongsian data mesti dilakukan dengan pengetahuan, persetujuan, dokumentasi yang jelas, dan proses yang boleh diaudit.



Keselamatan dan Kerahsiaan

- Perkongsian data terperingkat secara selamat;
- Data mesti dilindungi daripada pencerobohan, kebocoran, atau kehilangan melalui kawalan keselamatan fizikal dan siber.



Prinsip Akauntabiliti

- Perkongsian data menambah nilai kepada perkhidmatan awam, rakyat dan negara.
- Agensi yang berkongsi atau menerima data bertanggungjawab terhadap ketepatan, keselamatan, dan penggunaan data.

