ETL

1. Batching

Data jalan berdasarkan schedule yang ditetapkan

Untuk data-data yang tidak perlu real time

1. Streaming

Proses data secara terus menerus secara real time

Ribet mulai dari deployment sampai ke monitoring

Jauh lebih mahal

Jika ingin mendeteksi penipuan

ETL

Transform dilakukan di sisi pipeline. Contoh menggunakan python

Bergantung pada tool etlnya, karena di code base

Workernya atau eksekutornya harus kuat untuk mentransform

Waktu kapan data ada di load lebih lama

ELT

Karena masuk dulu ke dwh baru di transform, maka transformnya menggunakan query atau sql

Transform dikakukan di dwh

Dwh yang harus kuat, setelah ditransform hasilnya juga ditaro di dwh

Waktu lebih cepat

Jauh lebih fleksibel

EXTRACTION

Proses pengambilan/penarikan data

INGESTION MODE

1. FULL

Ngambil semua data lalu di overide ke target / dwh

1. INCREMENTAL

Saat extract sebagian data dan dia disimpan ke target tanpa overide

Bisa menggunakan id auto increment integer

Dwh

1.

2.

3

Max 3

Select \* from src where id > 3

Src

1

2

3

4

5

Atau timestamp (created at | update at)

Dwh

1.

2.

3

Max 3

Select \* from src where id > 3 update\_at between jam 23 and jam 00 {job start\_date -1} and

Src

1

2

3

4

5

SCD

1. Tipe 1

Jika ada perubahan data maka data tersebut akan di replace

src

Id | nama | updated\_at

1 | galuh | 1 januari

2 | priska | 2 januari

3 | lies | 4 januari

Dwh

Id | nama | updated\_at

1 | galuh | 1 januari

2 | priska | 2 januari

3 | lies | 3 januari (akan mereplace jadi 4 januari)

Kelebihan tipe nya terjaga, source dan dwh teteap berjumlah sama datanya

Tidak bisa menganalisis pergantian data, behaviour data change

1. Tipe 2 -- > nambah

dwh

Id | nama | updated\_at | is\_active

1 | galuh | 1 januari |

2 | priska | 2 januari

3 | lies | 3 januari

4 | surya | 4 januari

src

Id | nama | updated\_at

1 | galuh | 1 januari

2 | priska | 2 januari

4 | surya | 4 januari

Ada meta data’

Select \* from dwh

Qualify row\_number() over

Bisa ngelacak ganti nama jamberapa

Sizenya besar banget

1. Tipe 3

Penggabungan kelebihna tipe 1 dan tipe 2

src

Id | nama | updated\_at

1 | galuh | 1 januari

2 | priska | 2 januari

3 | lies | 3 januari KALAU INI JADI surya 4 januari

dwh

Id | nama | updated\_at | prev\_nama | prev\_update\_at

1 | galuh | 1 januari | NULL

2 | priska | 2 januari |

3 | surya | 4 januari | lies | 3 januari

Kelemahan, preview historical data terbatas, hanya 1 perubahan terakhir

STAGING AREA

Extract transform dan load adalah job yang terpisah.

Cara dapaten data dari ekstrak, caranya ekstrak nyimpen data dulu ke datalake

Transform ke load ke datalake dulu

Biasanya saat diextract ditambah meta data

src

Id | nama | updated\_at

1 | galuh | 1 januari

2 | priska | 2 januari

4 | surya | 4 januari

dwh

Id | nama | updated\_at | extract\_at

1 | galuh | 1 januari | 10 jan

2 | priska | 2 januari

3 | lies | 3 januari | 11 jan

4 | surya | 4 januari

./raw/10 jan.parquet

./transformed/10 jan.parquet

HADOOP

Banyak komputer yang bekerja sama untuk menyelesaikan pekerjaan

SPARK

Melakukan prosesing data bersama2

Untuk transformasi

KAFKA

Massage broker

Untuk data streaming

AIRFLOW

Job2 yang saling terkoneksi

DBT

Tool untuk trasformasi data, data yang di dwh

Adanya leneage data