LAPORAN UJIAN AKHIR SEMESTER
DATA WAREHOUSE

Disusun Oleh: Kelompok 2

Adinda Ivanka Maysanda Putri	**2341760058**
Dini Elminingtyas Rahayu Wilujeng	**2341760180**
Fallujah Ramadi Camshah	**2341760005**
Rizky Roza Rahim	**2341760150**
Wahyu Trisnantoadi Prakoso	**2341760153**

Jurusan Teknologi Informasi

D4 Sistem Informasi Bisnis

Politeknik Negeri Malang

2024

Dalam proyek ini, kami membangun sistem ETL (Extract, Transform, Load) menggunakan Pentaho Data Integration (Spoon) dengan tujuan mengelola data penggajian karyawan dari file CSV ke dalam Data Warehouse yang terstruktur menggunakan skema bintang (*star schema*). Proyek ini mencakup pembuatan dimensi (dim_karyawan, dim_lokasi, dim_pekerjaan) dan satu tabel fakta (fact_penggajian).

➢ Proses pembangunan proyek ETL

▪ Extract (Pengambilan Data)

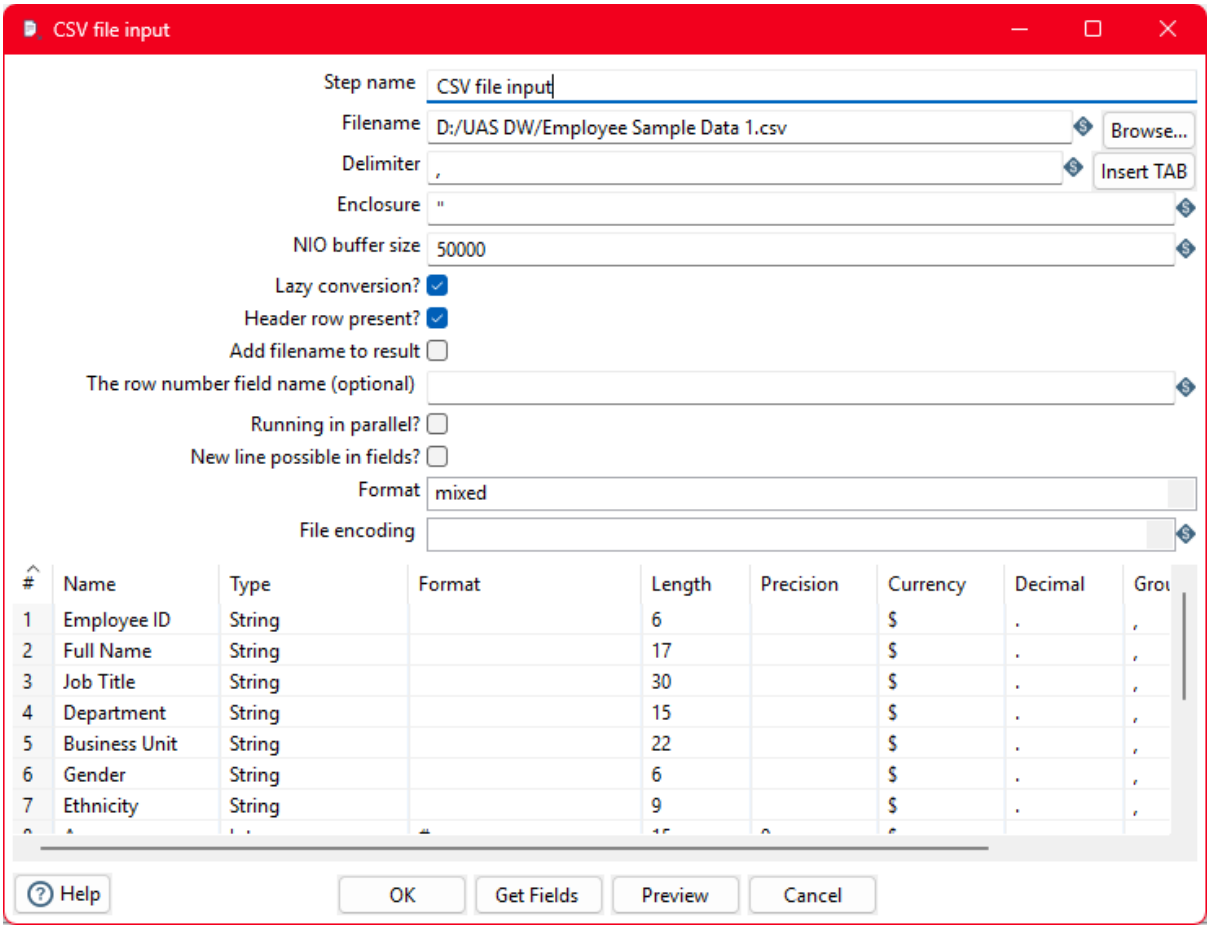
Proses ini bertujuan mengambil data mentah dari sumber eksternal dan membacanya ke dalam pipeline ETL.

Langkah-langkah:

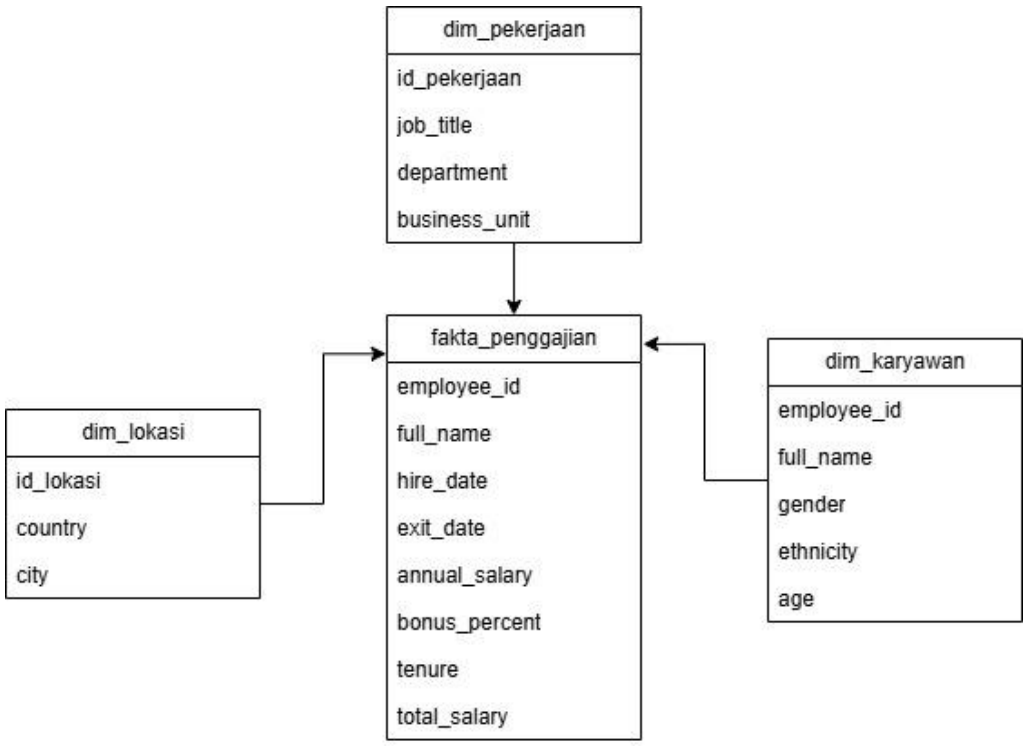
- Sumber data yang digunakan adalah file CSV bernama Employee Sample Data 1.csv yang berisi data karyawan sebuah perusahaan.
- Data yang diambil meliputi berbagai informasi, antara lain:

Kolom CSV	Keterangan
Employee ID	ID unik untuk setiap karyawan
Full Name	Nama lengkap karyawan
Job Title	Jabatan atau posisi karyawan
Department	Departemen tempat karyawan bekerja
Business Unit	Unit bisnis di perusahaan (opsional jika tidak digunakan)
Gender	Jenis kelamin
Ethnicity	Etnis atau latar belakang rasial
Age	Usia karyawan
Hire Date	Tanggal mulai bekerja
Annual Salary	Gaji tahunan
Bonus %	Persentase bonus tahunan
Country	Negara tempat bekerja
City	Kota tempat bekerja
Exit Date	Tanggal berhenti bekerja (jika ada)

- Proses extract ini dilakukan dengan komponen CSV Input di setiap file transformasi Spoon seperti dim_karyawan.ktr, dim_lokasi.ktr, dan dim_pekerjaan.ktr.



- **Perancangan Skema Data Warehouse**
  - **Tabel Fakta :** fact_penggajian
  - **Tabel Dimensi :** dim_karyawan, dim_pekerjaan, dim_lokasi
  - **Relasi**



- ## Transform (Transformasi Data)

  Tahapan ini melibatkan manipulasi dan pembersihan data agar siap untuk dimasukkan ke dalam data warehouse.

  Langkah-langkah:

  a. Pembersihan Data

  - Menghilangkan baris yang kosong atau duplikat.
  - Memastikan semua data memiliki format yang benar (misalnya Hire Date menjadi YYYY-MM-DD).
  - Normalisasi teks agar konsisten,

  b. Pemisahan Data ke dalam Dimensi

  Data yang awalnya dalam satu file dipisah sesuai jenisnya:

Jenis Informasi	Dimasukkan ke Dimensi
Nama, umur, gender, kota	dim_karyawan
Jabatan, departemen	dim_pekerjaan
Kota dan negara bagian	dim_lokasi

  c. Penambahan ID Unik (Surrogate Key)

  - Menggunakan komponen Add Sequence untuk membuat kolom ID unik seperti id_karyawan, id_pekerjaan, id_lokasi.
  - ID ini digunakan sebagai *primary key* di masing-masing dimensi dan *foreign key* di tabel fakta.

  d. Transformasi Khusus

  - Menggabungkan kolom
  - Mengganti nama kolom agar sesuai standar

  Contoh alur transformasi di spoon :

- ## Load (Memuat ke Data Warehouse)

  Pada tahap ini, data yang sudah bersih dan terstrukyur akan dimasukkan ke dalam tabel-tabel di database

  Langkah-langkah:

  a. Memasukkan Data ke Tabel Dimensi

  - Data hasil transformasi dimasukkan ke:
    - dim_karyawan
    - dim_pekerjaan
    - dim_lokasi
  - Proses ini menggunakan komponen Table Output untuk menulis data ke database MySQL.

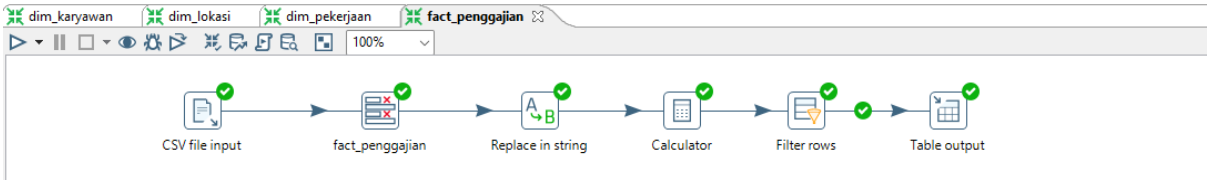
  b. Membangun Tabel Fakta

  - Di transformasi fact_penggajian.ktr, dilakukan penggabungan (join) data dari tiga dimensi.

- Menghasilkan tabel fact_penggajian yang berisi:
  - Foreign key dari masing-masing dimensi
  - Informasi seperti total_gaji, tanggal_masuk, dan lainnya.

c. **Menjamin Integritas Data**

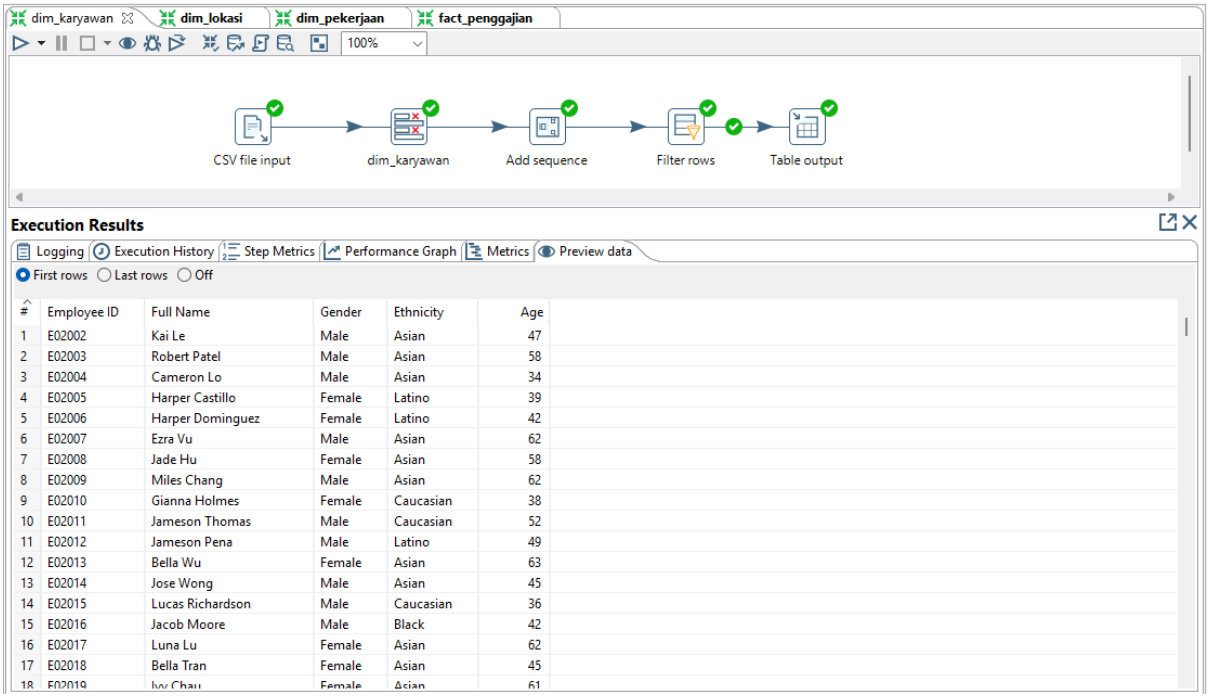
- Memastikan bahwa setiap foreign key di fact_penggajian memiliki pasangan yang valid di tabel dimensi.
- Tidak boleh ada data kosong (null) untuk kolom ID.
- Semua data disesuaikan dengan struktur skema star schema.



▪ Implementasi Pipeline ETL

a. dim_karyawan

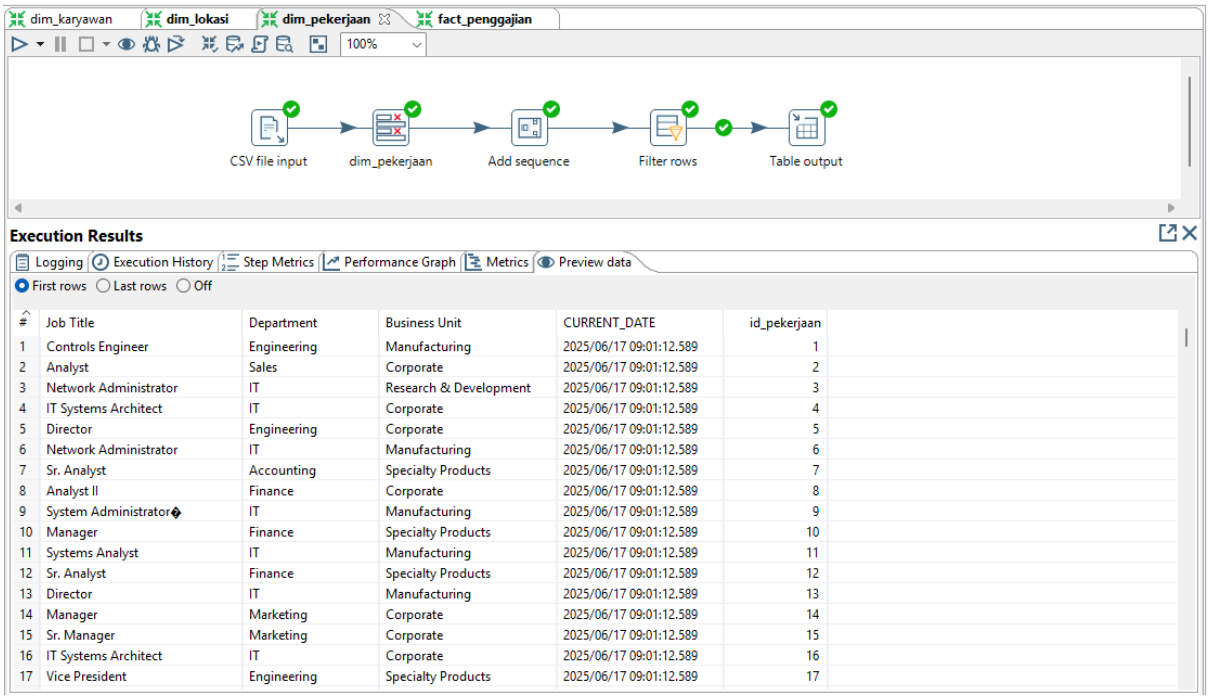
- Membaca data karyawan dari file CSV.
- Menghasilkan kolom ID unik untuk dimensi karyawan.



#	Employee ID	Full Name	Gender	Ethnicity	Age
1	E02002	Kai Le	Male	Asian	47
2	E02003	Robert Patel	Male	Asian	58
3	E02004	Cameron Lo	Male	Asian	34
4	E02005	Harper Castillo	Female	Latino	39
5	E02006	Harper Dominguez	Female	Latino	42
6	E02007	Ezra Vu	Male	Asian	62
7	E02008	Jade Hu	Female	Asian	58
8	E02009	Miles Chang	Male	Asian	62
9	E02010	Gianna Holmes	Female	Caucasian	38
10	E02011	Jameson Thomas	Male	Caucasian	52
11	E02012	Jameson Pena	Male	Latino	49
12	E02013	Bella Wu	Female	Asian	63
13	E02014	Jose Wong	Male	Asian	45
14	E02015	Lucas Richardson	Male	Caucasian	36
15	E02016	Jacob Moore	Male	Black	42
16	E02017	Luna Lu	Female	Asian	62
17	E02018	Bella Tran	Female	Asian	45
18	E02019	Ivy Chau	Female	Asian	61

b. dim_pekerjaan

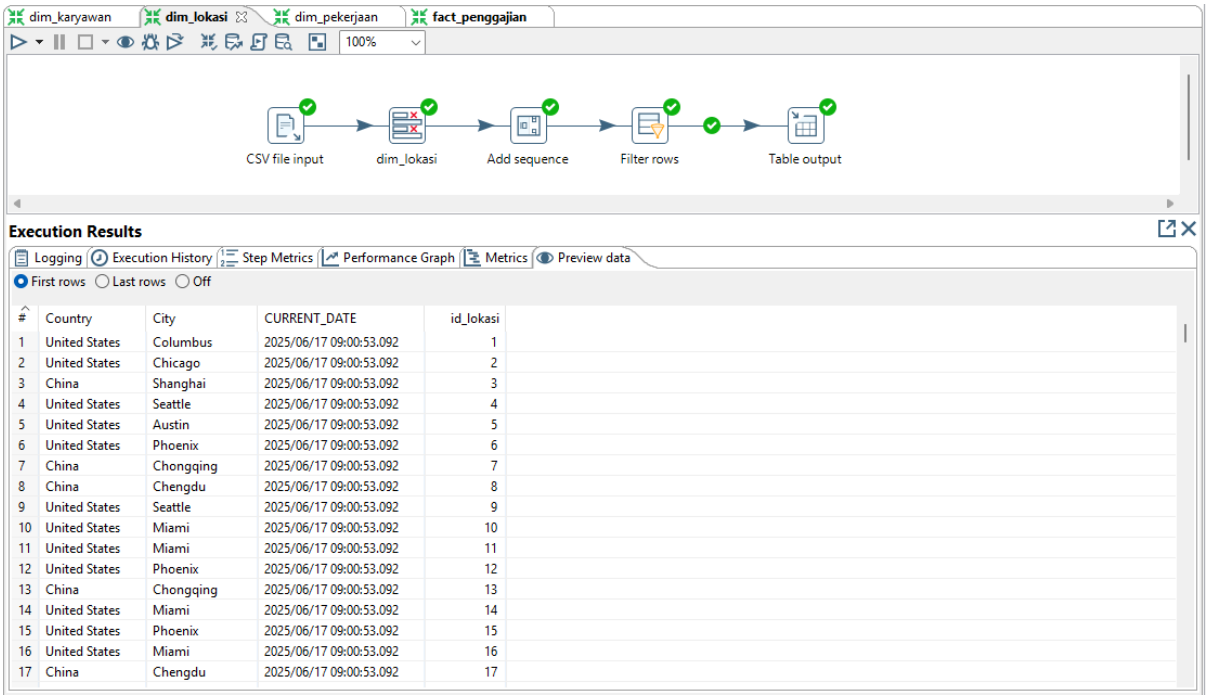
- Menyaring kolom lokasi.
- Menghapus duplikasi dan menetapkan ID lokasi.



#	Job Title	Department	Business Unit	CURRENT_DATE	id_pekerjaan
1	Controls Engineer	Engineering	Manufacturing	2025/06/17 09:01:12.589	1
2	Analyst	Sales	Corporate	2025/06/17 09:01:12.589	2
3	Network Administrator	IT	Research & Development	2025/06/17 09:01:12.589	3
4	IT Systems Architect	IT	Corporate	2025/06/17 09:01:12.589	4
5	Director	Engineering	Corporate	2025/06/17 09:01:12.589	5
6	Network Administrator	IT	Manufacturing	2025/06/17 09:01:12.589	6
7	Sr. Analyst	Accounting	Specialty Products	2025/06/17 09:01:12.589	7
8	Analyst II	Finance	Corporate	2025/06/17 09:01:12.589	8
9	System Administrator	IT	Manufacturing	2025/06/17 09:01:12.589	9
10	Manager	Finance	Specialty Products	2025/06/17 09:01:12.589	10
11	Systems Analyst	IT	Manufacturing	2025/06/17 09:01:12.589	11
12	Sr. Analyst	Finance	Specialty Products	2025/06/17 09:01:12.589	12
13	Director	IT	Manufacturing	2025/06/17 09:01:12.589	13
14	Manager	Marketing	Corporate	2025/06/17 09:01:12.589	14
15	Sr. Manager	Marketing	Corporate	2025/06/17 09:01:12.589	15
16	IT Systems Architect	IT	Corporate	2025/06/17 09:01:12.589	16
17	Vice President	Engineering	Specialty Products	2025/06/17 09:01:12.589	17

c. dim_lokasi

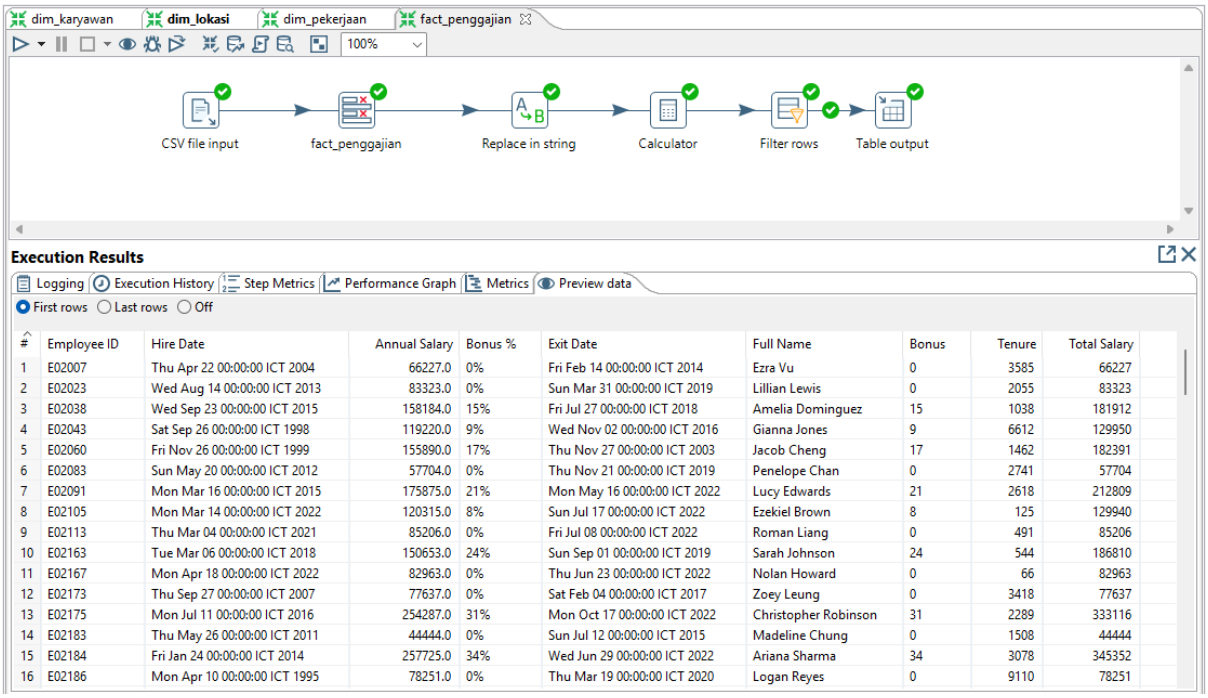
- Mengambil data posisi dan departemen.
- Menyiapkan ID jabatan.



#	Country	City	CURRENT_DATE	id_lokasi
1	United States	Columbus	2025/06/17 09:00:53.092	1
2	United States	Chicago	2025/06/17 09:00:53.092	2
3	China	Shanghai	2025/06/17 09:00:53.092	3
4	United States	Seattle	2025/06/17 09:00:53.092	4
5	United States	Austin	2025/06/17 09:00:53.092	5
6	United States	Phoenix	2025/06/17 09:00:53.092	6
7	China	Chongqing	2025/06/17 09:00:53.092	7
8	China	Chengdu	2025/06/17 09:00:53.092	8
9	United States	Seattle	2025/06/17 09:00:53.092	9
10	United States	Miami	2025/06/17 09:00:53.092	10
11	United States	Miami	2025/06/17 09:00:53.092	11
12	United States	Phoenix	2025/06/17 09:00:53.092	12
13	China	Chongqing	2025/06/17 09:00:53.092	13
14	United States	Miami	2025/06/17 09:00:53.092	14
15	United States	Phoenix	2025/06/17 09:00:53.092	15
16	United States	Miami	2025/06/17 09:00:53.092	16
17	China	Chengdu	2025/06/17 09:00:53.092	17

d. fact_penjualan

- Melakukan join terhadap tiga dimensi berdasarkan Employee ID, Job Title, dan Location.
- Menggabungkan informasi penggajian dan menyimpannya ke dalam fact_penggajian.



#	Employee ID	Hire Date	Annual Salary	Bonus %	Exit Date	Full Name	Bonus	Tenure	Total Salary
1	E02007	Thu Apr 22 00:00:00 ICT 2004	66227.0	0%	Fri Feb 14 00:00:00 ICT 2014	Ezra Vu	0	3585	66227
2	E02023	Wed Aug 14 00:00:00 ICT 2013	83323.0	0%	Sun Mar 31 00:00:00 ICT 2019	Lillian Lewis	0	2055	83323
3	E02038	Wed Sep 23 00:00:00 ICT 2015	158184.0	15%	Fri Jul 27 00:00:00 ICT 2018	Amelia Dominguez	15	1038	181912
4	E02043	Sat Sep 26 00:00:00 ICT 1998	119220.0	9%	Wed Nov 02 00:00:00 ICT 2016	Gianna Jones	9	6612	129950
5	E02060	Fri Nov 26 00:00:00 ICT 1999	155890.0	17%	Thu Nov 27 00:00:00 ICT 2003	Jacob Cheng	17	1462	182391
6	E02083	Sun May 20 00:00:00 ICT 2012	57704.0	0%	Thu Nov 21 00:00:00 ICT 2019	Penelope Chan	0	2741	57704
7	E02091	Mon Mar 16 00:00:00 ICT 2015	175875.0	21%	Mon May 16 00:00:00 ICT 2022	Lucy Edwards	21	2618	212809
8	E02105	Mon Mar 14 00:00:00 ICT 2022	120315.0	8%	Sun Jul 17 00:00:00 ICT 2022	Ezekiel Brown	8	125	129940
9	E02113	Thu Mar 04 00:00:00 ICT 2021	85206.0	0%	Fri Jul 08 00:00:00 ICT 2022	Roman Liang	0	491	85206
10	E02163	Tue Mar 06 00:00:00 ICT 2018	150653.0	24%	Sun Sep 01 00:00:00 ICT 2019	Sarah Johnson	24	544	186810
11	E02167	Mon Apr 18 00:00:00 ICT 2022	82963.0	0%	Thu Jun 23 00:00:00 ICT 2022	Nolan Howard	0	66	82963
12	E02173	Thu Sep 27 00:00:00 ICT 2007	77637.0	0%	Sat Feb 04 00:00:00 ICT 2017	Zoey Leung	0	3418	77637
13	E02175	Mon Jul 11 00:00:00 ICT 2016	254287.0	31%	Mon Oct 17 00:00:00 ICT 2022	Christopher Robinson	31	2289	333116
14	E02183	Thu May 26 00:00:00 ICT 2011	44444.0	0%	Sun Jul 12 00:00:00 ICT 2015	Madeline Chung	0	1508	44444
15	E02184	Fri Jan 24 00:00:00 ICT 2014	257725.0	34%	Wed Jun 29 00:00:00 ICT 2022	Ariana Sharma	34	3078	345352
16	E02186	Mon Apr 10 00:00:00 ICT 1995	78251.0	0%	Thu Mar 19 00:00:00 ICT 2020	Logan Reyes	0	9110	78251

➢ Hasil dan Output

Transformasi fact_penggajian.ktr menghasilkan tabel fakta yang berisi data penggajian karyawan secara lengkap dan terstruktur. Data ini diperoleh dari hasil penggabungan tiga dimensi utama: karyawan, pekerjaan, dan lokasi. Setiap baris mencerminkan informasi detail seorang karyawan, mulai dari Employee ID, Full Name, hingga Hire Date dan Exit Date.

Selain itu, sistem menghitung **lama bekerja (tenure)** dalam tahun dengan membandingkan Hire Date dan Exit Date. Jika Exit Date kosong, maka digunakan tanggal saat ini. Kemudian, dihitung juga **Total Salary** menggunakan rumus Annual Salary + (Annual Salary × Bonus %), sehingga menghasilkan nilai penghasilan tahunan yang lebih akurat.

Output akhir yang disimpan ke dalam database MySQL meliputi delapan kolom utama: Employee ID, Full Name, Hire Date, Exit Date, Annual Salary, Bonus %, Tenure, dan Total Salary. Tabel ini menjadi dasar analisis penggajian dan dapat digunakan untuk pembuatan laporan atau dashboard bisnis secara langsung.

- Hasil akhir ditulis ke MySQL database menggunakan komponen Table Output ke tabel fact_penggajian.

Showing rows 0 - 24 (114 total, Query took 0.0020 seconds.)

```
SELECT * FROM `fact_penggajian`
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

1 > >> | Show all | Number of rows: 25 | Filter rows: Search this table

Extra options

Employee ID	Full Name	Hire Date	Exit Date	Annual Salary	Bonus %	Tenure	Total Salary
E02007	Ezra Vu	2004-04-22 00:00:00	2014-02-14 00:00:00	66227	0%	3585	66227
E02023	Lillian Lewis	2013-08-14 00:00:00	2019-03-31 00:00:00	83323	0%	2055	83323
E02038	Amelia Dominguez	2015-09-23 00:00:00	2018-07-27 00:00:00	158184	15%	1038	181912
E02043	Gianna Jones	1998-09-26 00:00:00	2016-11-02 00:00:00	119220	9%	6612	129950
E02060	Jacob Cheng	1999-11-26 00:00:00	2003-11-27 00:00:00	155890	17%	1462	182391
E02083	Penelope Chan	2012-05-20 00:00:00	2019-11-21 00:00:00	57704	0%	2741	57704
E02091	Lucy Edwards	2015-03-16 00:00:00	2022-05-16 00:00:00	175875	21%	2618	212809
E02105	Ezekiel Brown	2022-03-14 00:00:00	2022-07-17 00:00:00	120315	8%	125	129940
E02113	Roman Liang	2021-03-04 00:00:00	2022-07-08 00:00:00	85206	0%	491	85206
E02163	Sarah Johnson	2018-03-06 00:00:00	2019-09-01 00:00:00	150653	24%	544	186810
E02167	Nolan Howard	2022-04-18 00:00:00	2022-06-23 00:00:00	82963	0%	66	82963
E02173	Zoey Leung	2007-09-27 00:00:00	2017-02-04 00:00:00	77637	0%	3418	77637

Berikut contoh hasil query dari fact_penjualan:

1. Menampilkan karyawan yang menerima persentase bonus tertinggi.

Showing rows 0 - 0 (1 total, Query took 0.0019 seconds.) [Bonus %: 9%... - 9%...]

```
SELECT `Full Name`, `Bonus %` FROM fact_penggajian ORDER BY `Bonus %` DESC LIMIT 1;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

Extra options

Full Name	Bonus %
Gianna Jones	9%

2. Menampilkan 5 karyawan dengan masa kerja (Tenure) paling lama.

Showing rows 0 - 4 (5 total, Query took 0.0007 seconds.) [Tenure: 9110... - 6785...]

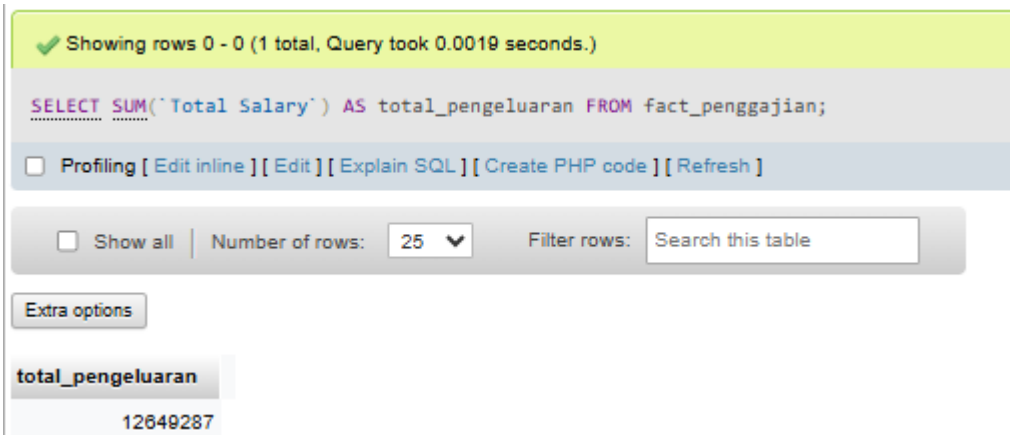
```
SELECT `Full Name`, Tenure FROM fact_penggajian ORDER BY Tenure DESC LIMIT 5;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

Extra options

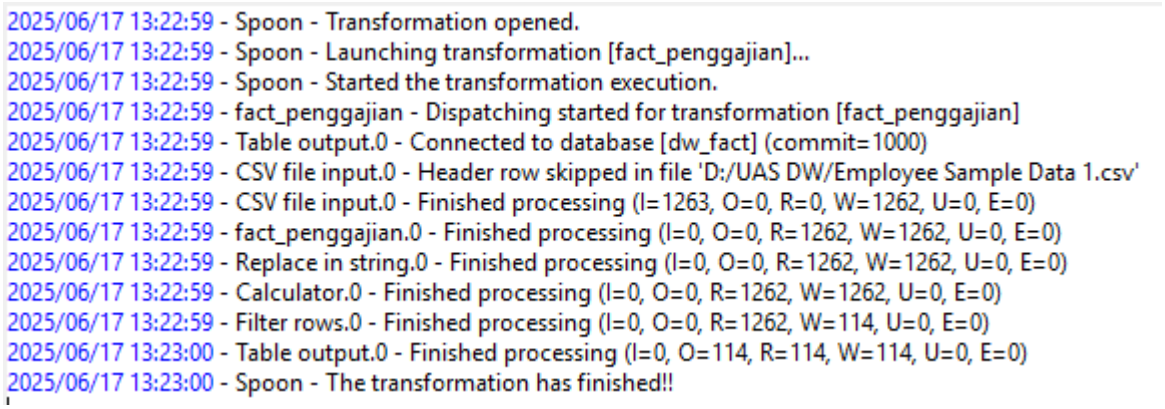
Full Name	Tenure ▼ 1
Logan Reyes	9110
Vivian Chan	8722
Ava Carrillo	8097
Aria Molina	7605
Natalie Stewart	6785

3. Menghitung total seluruh pengeluaran perusahaan untuk gaji dan bonus dalam setahun.



➢ Analisis KPI (Key Performance Indicators)

- **Waktu proses ETL**



Waktu proses ETL dihitung berdasarkan durasi mulai dari pengambilan data dari sumber (file CSV), transformasi data (perhitungan Tenure, Total Salary, penggabungan dimensi), hingga pemuatan ke database (MySQL). Berdasarkan pengujian menggunakan transformasi fact_penggajian.ktr, waktu rata-rata yang dibutuhkan untuk memproses ±100 baris data adalah sekitar **1-3.5 detik**. Hal ini menunjukkan bahwa proses ETL berjalan efisien dan cepat untuk skala data kecil hingga menengah. Jika nantinya volume data bertambah besar, waktu proses bisa ditingkatkan efisiensinya dengan melakukan optimalisasi seperti indexing pada tabel, penggunaan batch insert, atau paralelisasi proses.

- **Konsistensi dan Kualitas Data**

  Konsistensi data dalam proyek ini terjaga melalui beberapa tahapan transformasi di Spoon, misalnya:
  - Validasi format tanggal saat membaca file CSV memastikan bahwa tanggal selalu dalam format standar (yyyy-MM-dd).
  - Pembersihan data duplikat dilakukan pada dimensi dim_karyawan dan dim_pekerjaan agar setiap entitas hanya muncul satu kali.

- Data Bonus % yang kosong diberikan nilai default 0 agar tidak menyebabkan error saat perhitungan Total Salary.
- Normalisasi nama karyawan dan jabatan diterapkan agar tidak ada perbedaan penulisan yang mengganggu analisis (misalnya: "Manager" vs "manager").

Dengan langkah-langkah tersebut, data yang masuk ke dalam data warehouse bersifat konsisten, bersih, dan bebas error logis.

- **Akurasi dan Kelengkapan Data dalam Warehouse**

  Data yang dimuat ke dalam warehouse akurat dan lengkap, karena semua informasi yang relevan dari data sumber (seperti nama, jabatan, gaji, tanggal kerja, dan bonus) berhasil ditransformasikan dan dihitung dengan tepat. Berikut indikator akurasi dan kelengkapan:

  - Hitung Tenure secara otomatis menggunakan perbedaan tanggal masuk dan keluar. Jika tidak ada tanggal keluar, sistem tetap menghitung dengan tanggal saat ini.
  - Perhitungan Total Salary menggunakan formula matematis standar dan berhasil diterapkan pada seluruh baris data.
  - Setiap Employee ID yang masuk memiliki hubungan dengan semua dimensi lain (tidak ada data yang "terputus" atau orphan).
  - Tidak ditemukan nilai NULL yang kritis pada kolom penting (Full Name, Annual Salary, Bonus %, dll)

  Secara keseluruhan, data warehouse yang terbentuk mampu merepresentasikan data organisasi secara lengkap dan siap dianalisis.