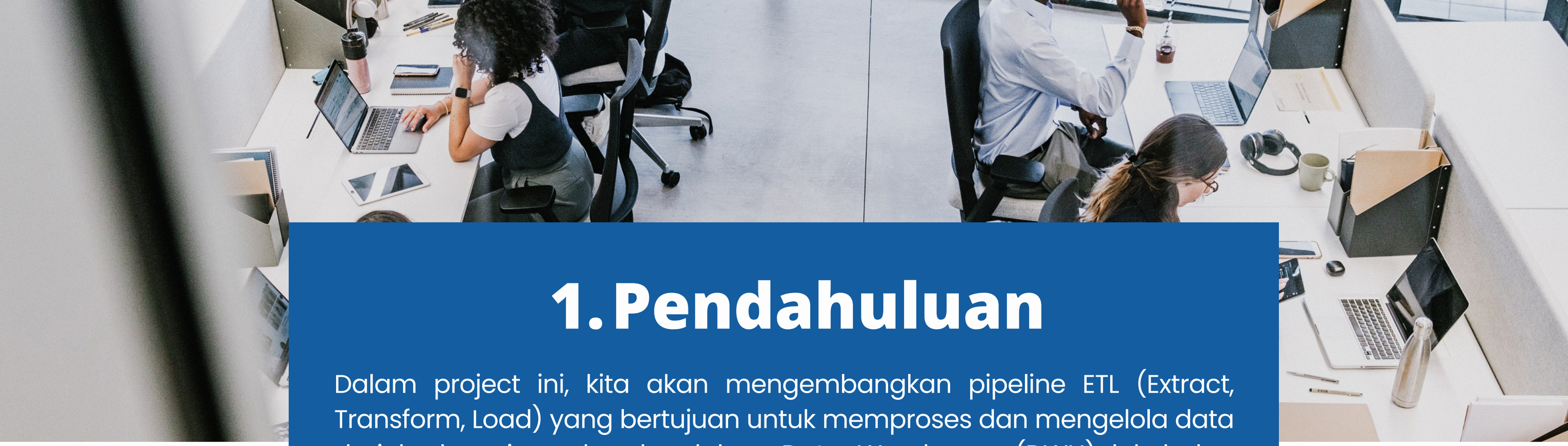
Superstore Integration

Takdir Zulhaq Dessiaming

Overview

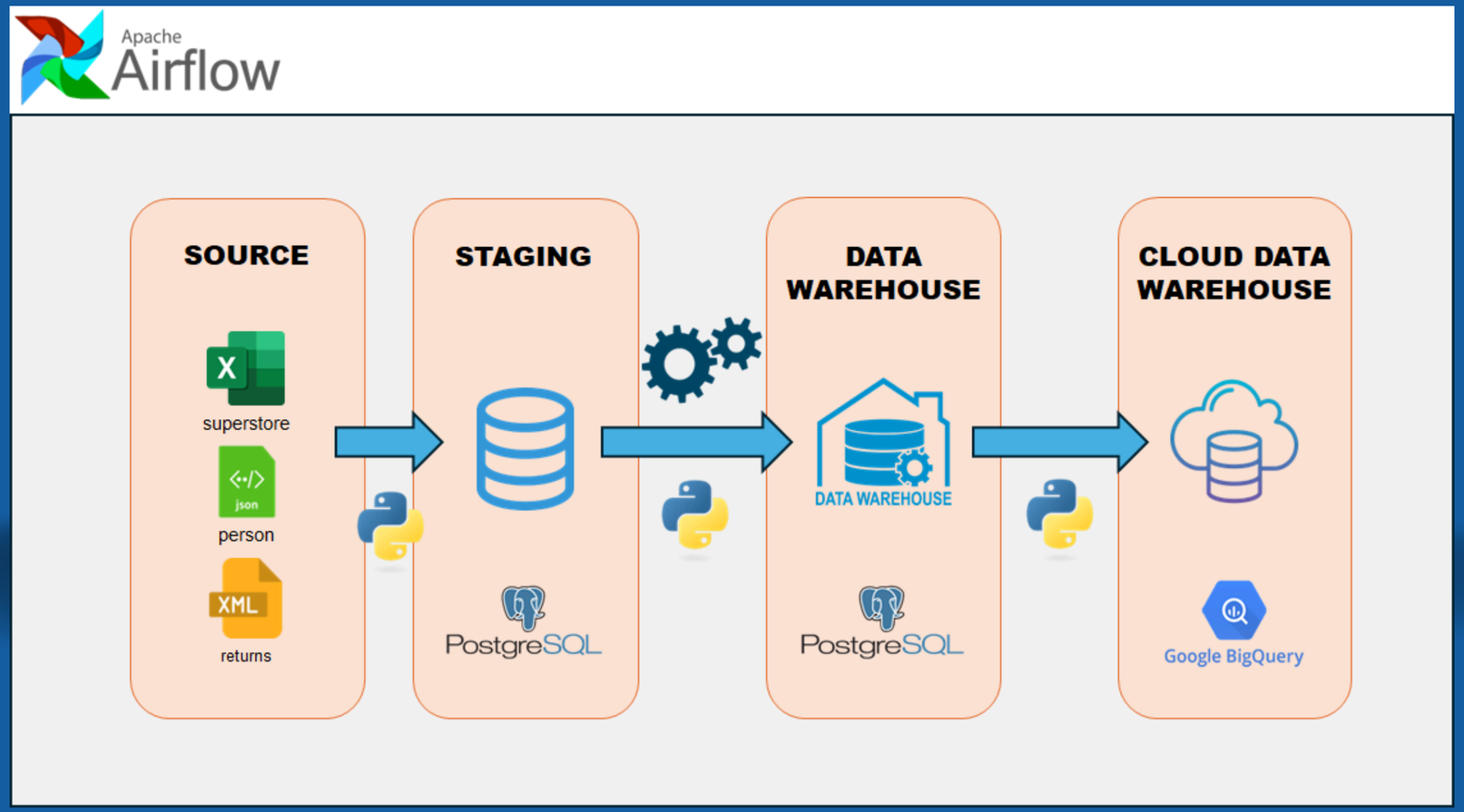


1. Pendahuluan

Dalam project ini, kita akan mengembangkan pipeline ETL (Extract, Transform, Load) yang bertujuan untuk memproses dan mengelola data dari berbagai sumber ke dalam Data Warehouse (DWH) lokal dan kemudian mengunggahnya ke BigQuery sebagai solusi cloud DWH. Proses ini akan dijalankan menggunakan Apache Airflow untuk memastikan alur kerja berjalan dengan efisien dan terjadwal.

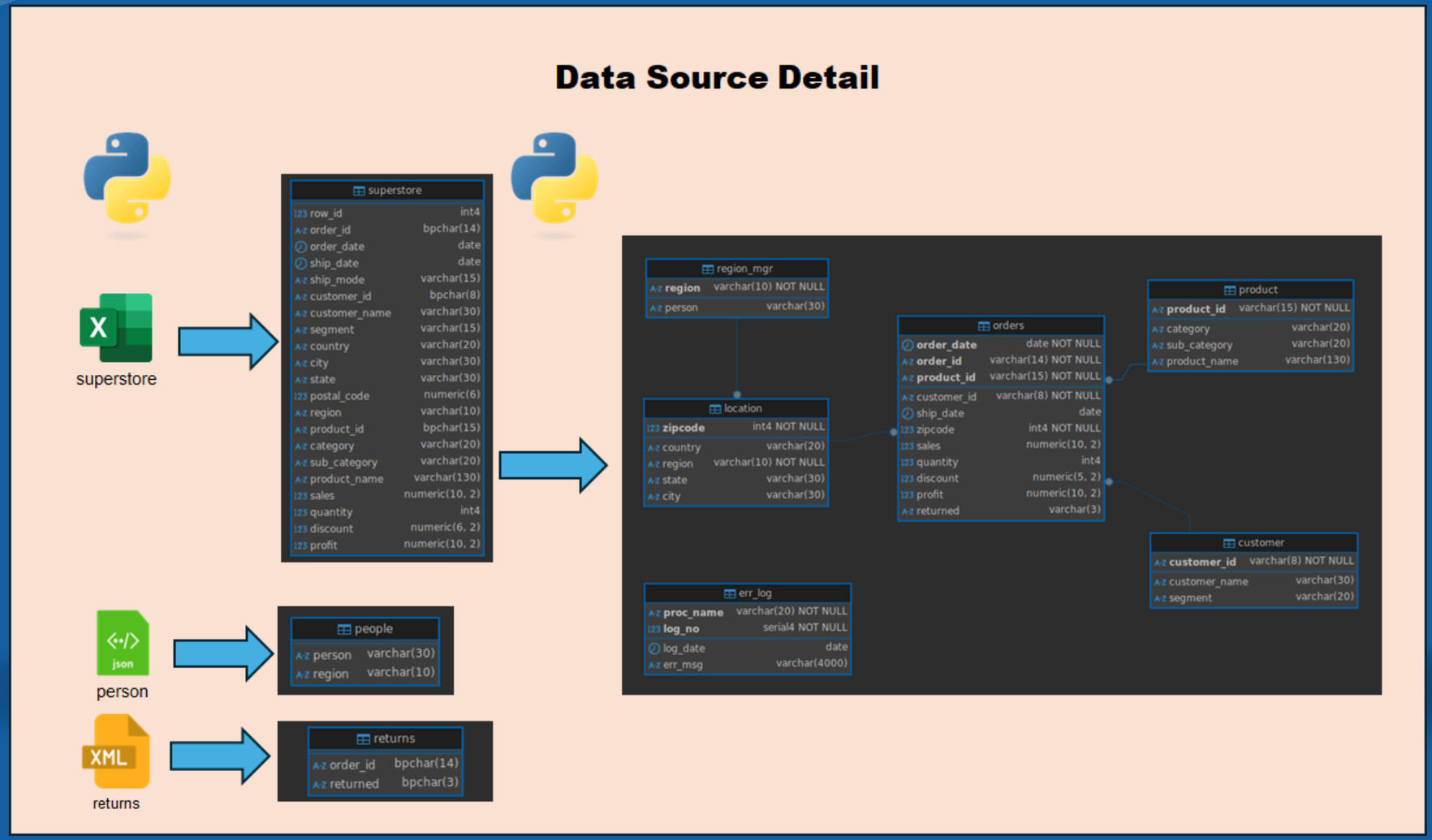
2. Data Flow dan ERD

Data dalam proyek ini bersumber dari berbagai file, seperti Excel (Superstore), JSON (People), dan XML (Returns). Data dari file-file ini akan dimuat ke dalam database staging pada PostgreSQL. Dari staging, data akan diproses menggunakan prosedur ETL yang mengubah dan menyusun data ke dalam format yang sesuai untuk DWH. Selanjutnya, data yang telah disusun dalam DWH akan dipindahkan ke BigQuery untuk kebutuhan analisis.



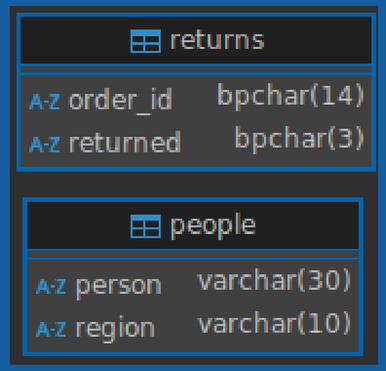
ERD

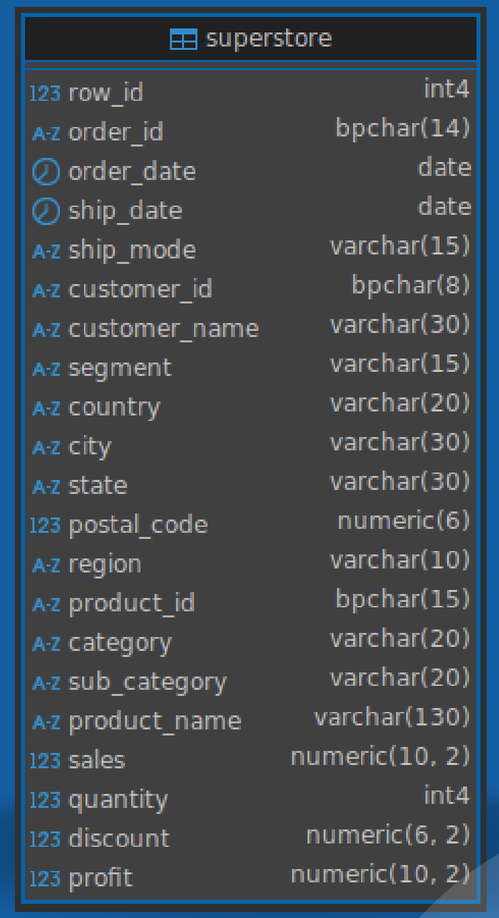
Berikut adalah relasi antar tabel yang menjadi dasar pada pembuatan Data Warehouse



3. Staging

Pada tahap ini, data dari berbagai sumber (Excel, JSON, XML) akan dimuat ke dalam skema staging_superstore di PostgreSQL. Staging digunakan sebagai tempat penyimpanan sementara sebelum data diolah lebih lanjut. Data akan dibersihkan, divalidasi, dan disiapkan untuk transformasi.





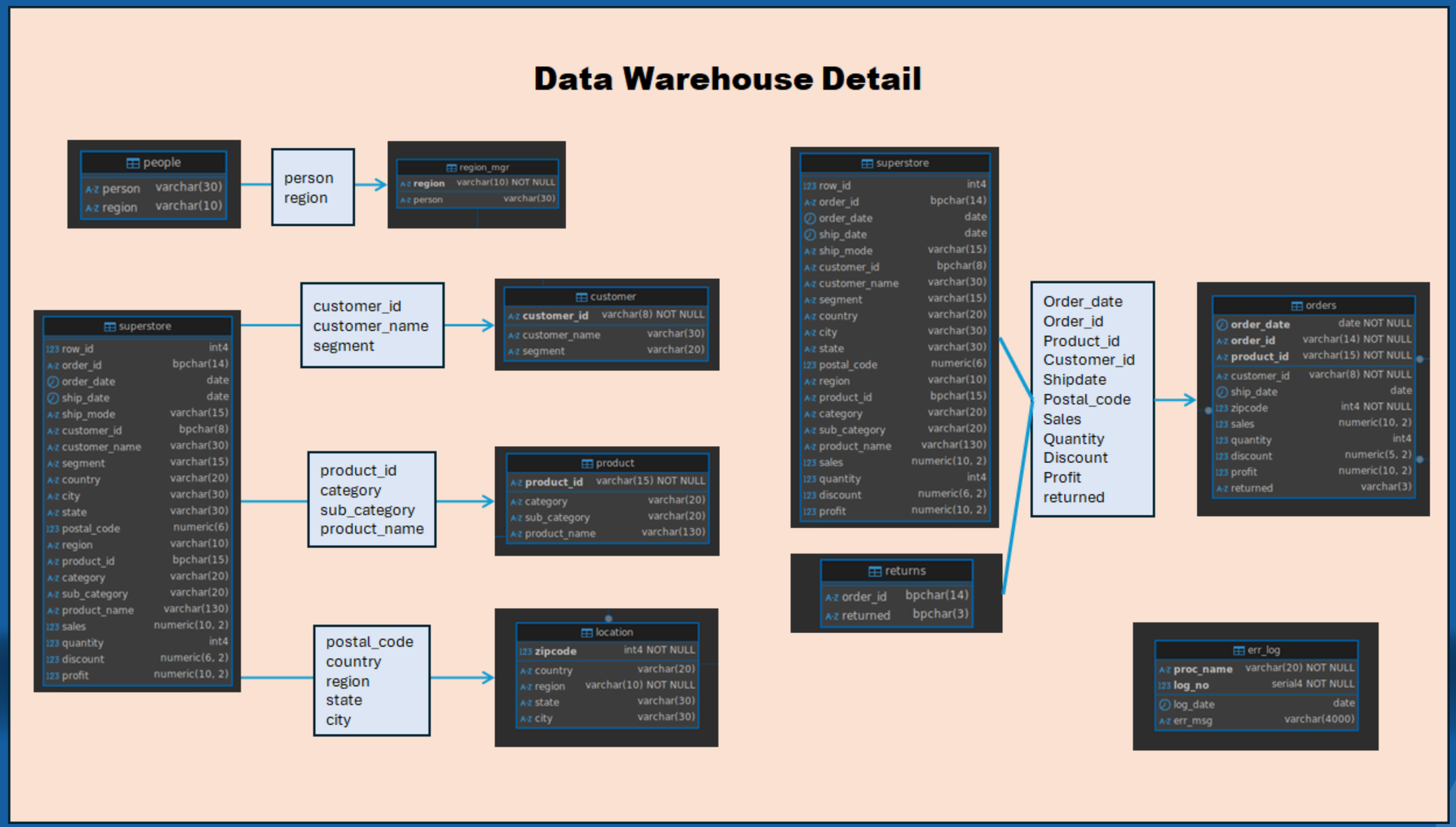
4. Integrasi ke Datawarehouse Local

Setelah data berada di staging, proses ETL akan mentransformasikannya ke dalam skema dwh_superstore. Data akan disusun ke dalam beberapa tabel utama:

- Region Manager: Berisi data manajer regional.
- Location: Menyimpan informasi lokasi.
- Product: Berisi informasi produk.
- Customer: Menyimpan data pelanggan.
- Orders: Berisi data pesanan pelanggan.

Jika ada data yang tidak sesuai aturan atau mengalami error, data tersebut akan dicatat dalam tabel error log untuk ditinjau lebih lanjut.

4. Integrasi ke Datawarehouse Local



5. Hasil Integrasi

Hasil akhir dari proses ini adalah data yang telah terstruktur dengan baik di dalam DWH lokal dan juga terdapat error log untuk tabel Orders yang gagal masuk. Data ini dapat digunakan untuk analisis lebih lanjut. Selain itu, data yang telah diproses akan dipindahkan ke BigQuery untuk analisis berbasis cloud.

Region Mgr

	Person	Region
1	Anna Andreadi	West
2	Chuck Magee	East
3	Kelly Williams	Central
4	Cassandra Bran	South

Location

	ZIPCODE	COUNTRY	REGION	STATE	CITY
1	1,040	United States	East	Massachusetts	Holyoke
2	1,453	United States	East	Massachusetts	Leominster
3	1,752	United States	East	Massachusetts	Marlborough
4	1,810	United States	East	Massachusetts	Andover
5	1,841	United States	East	Massachusetts	Lawrence
6	1,852	United States	East	Massachusetts	Lowell
7	1,915	United States	East	Massachusetts	Beverly

Customer

	CUSTOMER_ID	CUSTOMER_NAME	SEGMENT
1	AA-10315	Alex Avila	Consumer
2	AA-10375	Allen Armold	Consumer
3	AA-10480	Andrew Allen	Consumer
4	AA-10645	Anna Andreadi	Consumer
5	AB-10015	Aaron Bergman	Consumer
6	AB-10060	Adam Bellavance	Home Office
7	AB-10105	Adrian Barton	Consumer

Product

	PRODUCT_ID	CATEGORY	SUB_CATEGORY	PRODUCT_NAME
1	FUR-BO-10000112	Furniture	Bookcases	Bush Birmingham Collection Bookcase, Dark Cherry
2	FUR-BO-10000330	Furniture	Bookcases	Sauder Camden County Barrister Bookcase, Planked (
3	FUR-BO-10000362	Furniture	Bookcases	Sauder Inglewood Library Bookcases
4	FUR-BO-10000468	Furniture	Bookcases	O'Sullivan 2-Shelf Heavy-Duty Bookcases
5	FUR-BO-10000711	Furniture	Bookcases	Hon Metal Bookcases, Gray
6	FUR-BO-10000780	Furniture	Bookcases	O'Sullivan Plantations 2-Door Library in Landvery Oal
7	FUR-BO-10001337	Furniture	Bookcases	O'Sullivan Living Dimensions 2-Shelf Bookcases

5. Hasil Integrasi

Orders

	ORDER_Date	ORDER_ID	PRODUCT_ID	CUSTOMER_ID	SHIP_DATE	ZIPCODE	SALES	QUANTITY	DISCOUNT	PROFIT	RETURNED
1	2011-09-07	CA-2011-100006	TEC-PH-10002075	DK-13375	2011-09-13	10,024	377.97	3	0	109.61	No
2	2011-07-08	CA-2011-100090	FUR-TA-10003715	EB-13705	2011-07-12	94,122	502.49	3	0.2	-87.94	No
3	2011-03-14	CA-2011-100293	OFF-PA-10000176	NF-18475	2011-03-18	32,216	91.06	6	0.2	31.87	No
4	2011-01-29	CA-2011-100328	OFF-BI-10000343	JC-15340	2011-02-04	10,024	3.93	1	0.2	1.33	No
5	2011-04-08	CA-2011-100363	OFF-FA-10000611	JM-15655	2011-04-15	85,301	2.37	2	0.2	0.83	No
6	2011-05-25	CA-2011-100391	OFF-PA-10001471	BW-11065	2011-05-29	10,035	14.62	2	0	6.73	No
7	2011-04-18	CA-2011-100678	OFF-AR-10001868	KM-16720	2011-04-22	77,095	2.69	2	0.2	1.01	No

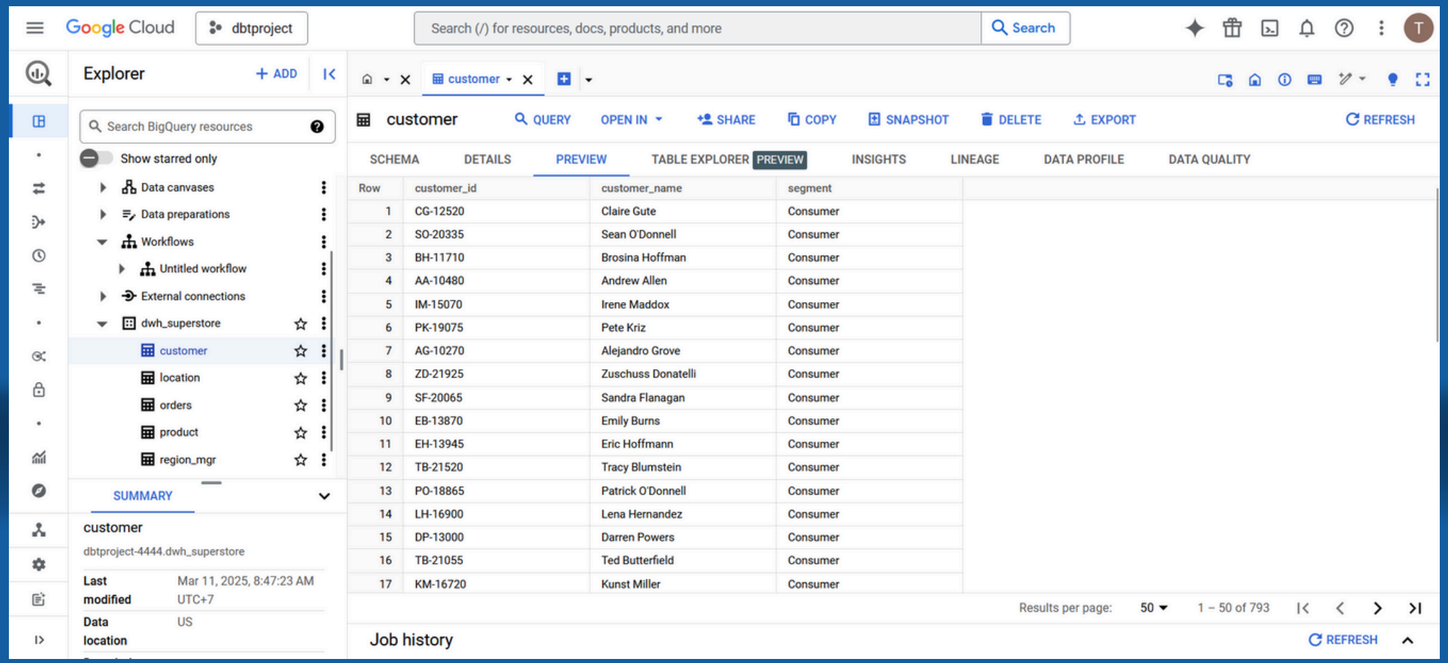
Error_Log

	PROC_NAME	LOG_NO	LOG_DATE	ERR_MSG
1	SYN_ORDERS	26,697	2023-12-16 10:49:46	ORDER_ID=CA-2013-129714 ORDER_DATE=2013-09-02
2	SYN_ORDERS	26,698	2023-12-16 10:49:46	ORDER_ID=US-2013-123750 ORDER_DATE=2013-04-16
3	SYN_ORDERS	26,699	2023-12-16 10:49:46	ORDER_ID=CA-2013-137043 ORDER_DATE=2013-12-24
4	SYN_ORDERS	26,700	2023-12-16 10:49:46	ORDER_ID=CA-2014-152912 ORDER_DATE=2014-11-10
5	SYN_ORDERS	26,701	2023-12-16 10:49:46	ORDER_ID=US-2011-150119 ORDER_DATE=2011-04-23
6	SYN_ORDERS	26,702	2023-12-16 10:49:46	ORDER_ID=CA-2012-103135 ORDER_DATE=2012-07-24
7	SYN_ORDERS	26,703	2023-12-16 10:49:46	ORDER_ID=CA-2014-118017 ORDER_DATE=2014-12-04
8	SYN_ORDERS	26,704	2023-12-16 10:49:46	ORDER_ID=CA-2013-140571 ORDER_DATE=2013-03-16

ERR_MSG
ORDER_ID=CA-2013-129714 ORDER_DATE=2013-09-02 PRODUCT_ID=OFF-PA-10001970 SHIP_DATE=2013-09-04 : unique constraint (orders.XPKORDERS) violated
ORDER_ID=US-2013-123750 ORDER_DATE=2013-04-16 PRODUCT_ID=TEC-AC-10004659 SHIP_DATE=2013-04-22 : unique constraint (orders.XPKORDERS) violated
ORDER_ID=CA-2013-137043 ORDER_DATE=2013-12-24 PRODUCT_ID=FUR-FU-10003664 SHIP_DATE=2013-12-26 : unique constraint (orders.XPKORDERS) violated
ORDER_ID=CA-2014-152912 ORDER_DATE=2014-11-10 PRODUCT_ID=OFF-ST-10003208 SHIP_DATE=2014-11-13 : unique constraint (orders.XPKORDERS) violated
ORDER_ID=US-2011-150119 ORDER_DATE=2011-04-23 PRODUCT_ID=FUR-CH-10002965 SHIP_DATE=2011-04-27 : unique constraint (orders.XPKORDERS) violated
ORDER_ID=CA-2012-103135 ORDER_DATE=2012-07-24 PRODUCT_ID=OFF-BI-10000069 SHIP_DATE=2012-07-28 : unique constraint (orders.XPKORDERS) violated
ORDER_ID=CA-2014-118017 ORDER_DATE=2014-12-04 PRODUCT_ID=TEC-AC-10002006 SHIP_DATE=2014-12-07 : unique constraint (orders.XPKORDERS) violated
ORDER_ID=CA-2013-140571 ORDER_DATE=2013-03-16 PRODUCT_ID=OFF-PA-10001954 SHIP_DATE=2013-03-20 : unique constraint (orders.XPKORDERS) violated

6. Integrasi ke BigQuery

Sebelum data dikirim ke BigQuery, ada beberapa hal yang perlu dipersiapkan di BigQuerynya, yaitu pembuatan akun, project, dataset, dan Service Account untuk autentikasi login. Untuk pembuatan kolom ada beberapa cara yakni bisa membuat langsung atau melalui tools dbt (Data Build Tools) dengan menggunakan csv seed dan query model di dbt. Untuk case ini, saya sebelumnya telah melakukan transformasi menggunakan dbt, sehingga tabel-tabel di dataset sudah terbentuk.



6. Integrasi ke BigQuery

Customer

	customer_id	custome...	segment
1.	ZD-21925	Zuschuss Don...	Consumer
2.	ZC-21910	Zuschuss Carr...	Consumer
3.	YC-21895	Yoseph Carroll	Corporate
4.	YS-21880	Yana Sorensen	Corporate
5.	XP-21865	Xylona Preis	Consumer
6.	WB-21850	William Brown	Consumer
7.	VM-21835	Vivian Mathis	Consumer
8.	VS-21820	Vivek Sundare...	Consumer
9.	VG-21805	Vivek Grady	Corporate
10.	VG-21790	Vivek Gonzalez	Consumer
11.	VW-21775	Victoria Wilson	Corporate

1 - 50 / 793

Location

	zipcode	country	region	state
1.	99301	United States	West	Washington
2.	99207	United States	West	Washington
3.	98661	United States	West	Washington
4.	98632	United States	West	Washington
5.	98502	United States	West	Washington
6.	98270	United States	West	Washington
7.	98226	United States	West	Washington
8.	98208	United States	West	Washington
9.	98198	United States	West	Washington
10.	98115	United States	West	Washington
11.	98105	United States	West	Washington

1 - 50 / 631

Product

	product_id	category	sub_category	product_name
1.	TEC-PH-10004977	Technology	Phones	GE 30524EE4
2.	TEC-PH-10004959	Technology	Phones	Classic Ivory Antique Telephone Z...
3.	TEC-PH-10004924	Technology	Phones	SKILCRAFT Telephone Shoulder R...
4.	TEC-PH-10004922	Technology	Phones	RCA Visys Integrated PBX 8-Line ...
5.	TEC-PH-10004912	Technology	Phones	Cisco SPA112 2 Port Phone Adap...
6.	TEC-PH-10004908	Technology	Phones	Panasonic KX TS3282W Corded p...
7.	TEC-PH-10004897	Technology	Phones	Mediabridge Sport Armband iPho...
8.	TEC-PH-10004896	Technology	Phones	Nokia Lumia 521 (T-Mobile)
9.	TEC-PH-10004875	Technology	Phones	PNY Rapid USB Car Charger - Black
10.	TEC-PH-10004833	Technology	Phones	Macally Suction Cup Mount
11.	TEC-PH-10004830	Technology	Phones	Pyle PRT45 Retro Home Telephone

1 - 50 / 1862

Region Manager

	person	region
1.	Kelly Williams	Central
2.	Chuck Magee	East
3.	Cassandra Brandow	South
4.	Anna Andreadi	West

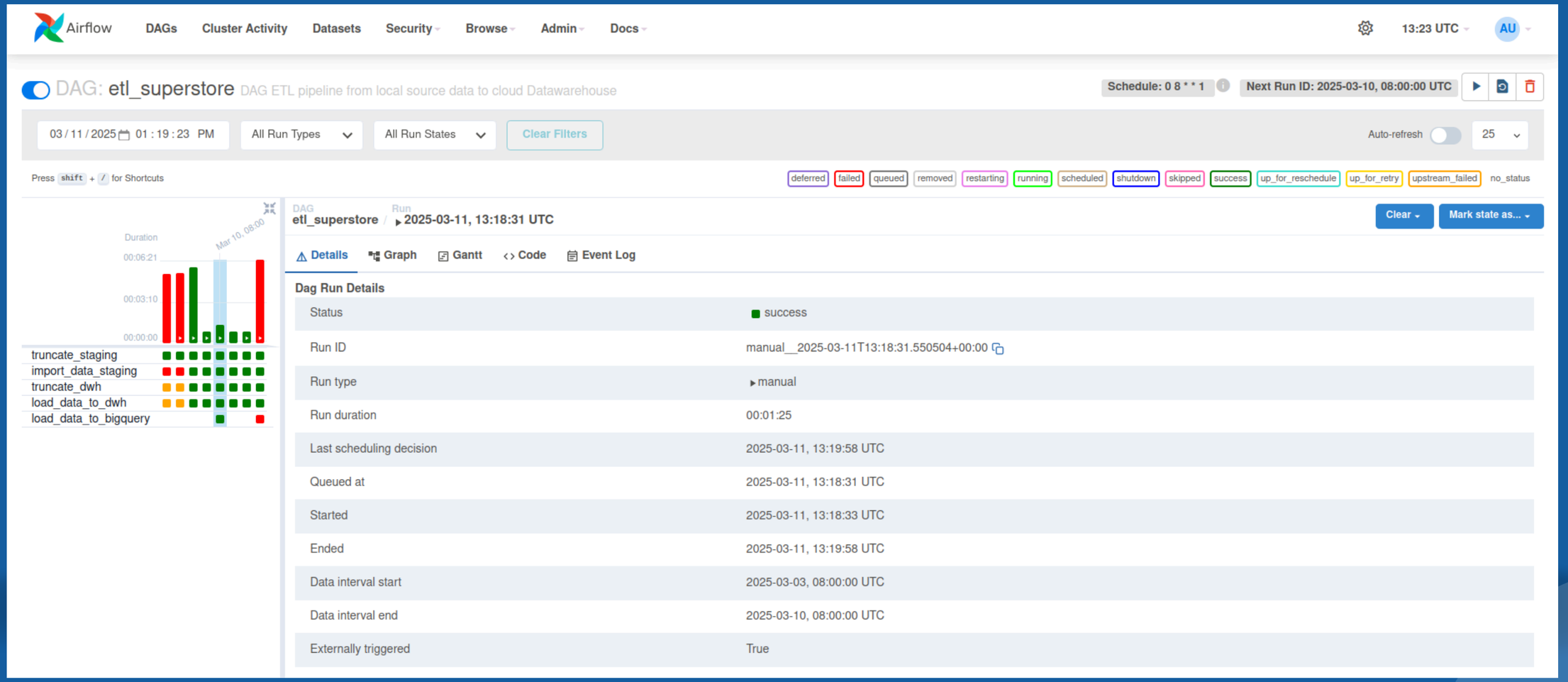
1 - 4 / 4

Orders

	order_date	order_id	product_id	customer_id	ship_date	zipcode	sales	quantity	discount	profit	returned
1.	Dec 31, 2014	CA-2014-115427	OFF-BI-10004632	EB-13975	Jan 4, 2015	94533	20.72	2	0.2	6.47	Yes
2.	Dec 31, 2014	CA-2014-143259	FUR-BO-10003441	PO-18865	Jan 4, 2015	10009	323.14	4	0.2	12.12	No
3.	Dec 31, 2014	CA-2014-115427	OFF-BI-10002103	EB-13975	Jan 4, 2015	94533	13.9	2	0.2	4.52	Yes
4.	Dec 31, 2014	CA-2014-156720	OFF-FA-10003472	JM-15580	Jan 4, 2015	80538	3.02	3	0.2	-0.6	No
5.	Dec 31, 2014	CA-2014-126221	OFF-AP-10002457	CC-12430	Jan 6, 2015	47201	209.3	2	0	56.51	No
6.	Dec 31, 2014	CA-2014-143259	TEC-PH-10004774	PO-18865	Jan 4, 2015	10009	90.93	7	0	2.73	No
7.	Dec 31, 2014	CA-2014-143259	OFF-BI-10003684	PO-18865	Jan 4, 2015	10009	52.78	3	0.2	19.79	No
8.	Dec 30, 2014	CA-2014-146626	FUR-FU-10002501	BP-11185	Jan 6, 2015	92804	101.12	8	0	37.41	No
9.	Dec 30, 2014	US-2014-158526	FUR-CH-10001270	KH-16360	Jan 2, 2015	40214	258.75	3	0	77.62	No
10.	Dec 30, 2014	US-2014-158526	FUR-CH-10002602	KH-16360	Jan 2, 2015	40214	1207.84	8	0	314.04	No
11.	Dec 30, 2014	CA-2014-130631	FUR-FU-10004093	BS-11755	Jan 3, 2015	98026	68.46	2	0	20.54	Yes

1 - 50 / 9986

Proses running script oleh Airflow



TERIMA KASIH

GITHUB