Analisa Rating Berdasarkan Divisi, Departemen, dan Kelas

Pada projek yang saya lakukan, dataset berasal dari Kaggle.com [1]. Dataset ini berjudul **Women's E-Commerce Clothing Reviews** yang memiliki 23485 baris dan 11 kolom. Pada deskripsi yang tertera 11 kolom tersebut antara lain [1].

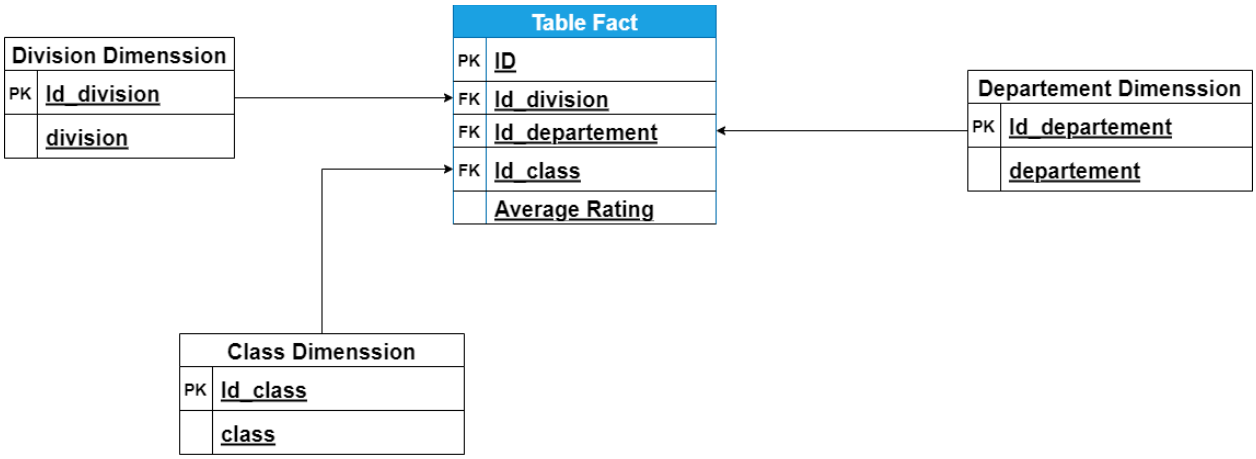
1. Clothing ID: Kolom yang berisi ID review dengan tipe integer
2. Age: Umur reviewer dengan integer
3. Title: Judul review yang diberikan dengan tipe string
4. Rating: Positive integer ordinal yang memiliki peringkat 1 (paling buruk) hingga 5 (paling baik)
5. Recommend IND: Kolom yang berisi nilai binary dimana nilai 1 berarti direkomendasi dan 0 berarti tidak direkomendasikan
6. Positive Feedback Count: Jumlah pelanggan yang memberikan nilai positif.
7. Division Name: Bertipe kategori untuk produksi
8. Department Name: Bertipe kategori untuk nama departemen produk
9. Class Name: Bertipe kategori untuk nama kelas produk.

Proses analisa menggunakan software Pentaho Data Integration (PDI) dengan menggunakan prinsip OLAP dengan terdapat tabel fakta dan tabel dimensi. Dimensi model yang digunakan adalah Star Schema sebagai visualisasi gambaran tabel fakta dan tabel dimensi dengan melakukan normalisasi 2NF. Star schema atau skema bintang adalah struktur logical yang memiliki tabel fakta yang terdiri atas data faktual sebagai hasil akhir analisis yang digunakan dengan dikelilingi tabel dimensi [2]. Normalisasi 2NF sendiri merupakan tahapan ketika normalisasi data sudah dianggap data normal (1NF) tanpa adanya atribut yang memiliki nilai ganda atau baris rangkap [2].

Permasalahan: Mengetahui rata-rata rating yang didapatkan berdasarkan Divisi, Departemen dan Kelas.

1. Star Schema

Permasalahan di atas dapat dilakukan pembuatan star schema yang menghasilkan star schema sebagai berikut.



Star schema diatas sudah dilakukan 2 NF dikarenakan nilai sudah bergantung dengan primary key nya. Seperti contoh Dimensi divisi yang bersifat FK di table Fact yang tergantung pada PK di dimensinya sendiri.

2. Extract

Proses extract disini dilakukan mendownload data dari Kaggle dan dilakukan import pada database. Database yang digunakan ialah PostgreSQL dengan nama database women_e_commerce.

a. Pembuatan Table

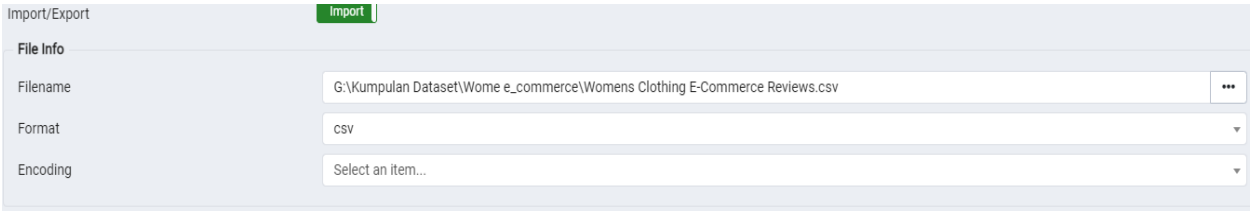
Pembuatan table dilakukan karena PostgreSQL tidak dapat menerima jika langsung dilakukan import pada dataset. Berikut query pembuatan tabel dengan nama ‘review’.

```
CREATE TABLE IF NOT EXISTS public.review
(
    "number" integer NOT NULL,
    clothing_id integer,
    "Age" integer,
    title "char",
    review_text "char",
    "Rating" integer,
    recommend boolean,
    division_name "char",
    department "char",
    class_name "char",
    PRIMARY KEY ("number")
```

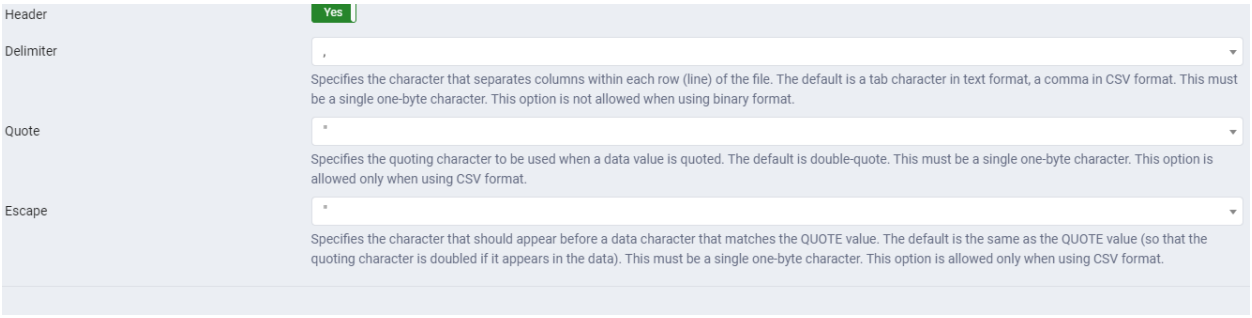
);

b. Import data

Setelah dilakukan pembuatan tabel dengan tipe kolom yang telah ditentukan, saatnya dilakukan proses import.



Dataset yang dimiliki bernama **Womens Clothing E-Commerce Reviews.csv** dengan format .csv. Encoding disini tidak dipilih dikarenakan encoding dilakukan jika bersifat **.csv UTF-8** atau yang lain.



Header dirubah menjadi Yes untuk menjadikan kolom sebagai baris pertama dan pemisah antara kolom (delimiter) menggunakan ‘,’ (koma) dan Quote serta Escape menggunakan ‘ ” ’ (petik dua). Petik dua digunakan dikarenakan terdapat baris yang bersifat teks. Berikut gambar data dengan LIMIT 5.

```
3  SELECT * FROM public.review LIMIT 5;
```

Data Output

	number [PK] integer	clothing_id integer	Age integer	title "char" (1)	review_text text
1	0	767	33	[null]	Absolutely wonderful - silky and sexy and comfortable
2	1	1080	34	[null]	Love this dress! it's sooo pretty. i happened to find it in a store, and i'm glad i did bc i never would have ordered it online bc it's pet
3	2	1077	60	S	I had such high hopes for this dress and really wanted it to work for me. i initially ordered the petite small (my usual size) but i four
4	3	1049	50	M	I love, love, love this jumpsuit. it's fun, flirty, and fabulous! every time i wear it, i get nothing but great compliments!
5	4	847	47	F	This shirt is very flattering to all due to the adjustable front tie. it is the perfect length to wear with leggings and it is sleeveless so

3. Transform dan Load

a. Tabel Dimensi

Proses transform yang dilakukan antra lain replace null ‘with unknown’, mengambil nilai unik dan menambahkan primary key pada setiap hasil tabel yang akan di load ke database. Sebelumnya mari kita berapa total nilai null pada tiap kolom.

```
SELECT COUNT(*) FROM public.review WHERE "division_name" IS NULL;

SELECT COUNT(*) FROM public.review WHERE "department" IS NULL;

SELECT COUNT(*) FROM public.review WHERE "class_name" IS NULL;
```

Hasil query diatas menghasilkan total 14 nilai null pada tiap-tiap tabel dan jika dilakukan pada fitur Pentaho, didapatkan seperti berikut.

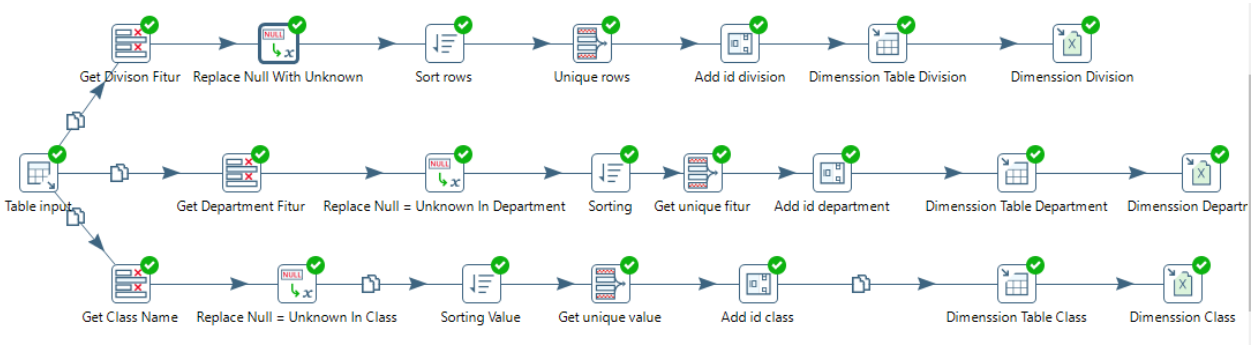
	count bigint
1	14

#	division
1	<null>
2	<null>
3	<null>
4	<null>
5	<null>
6	<null>
7	<null>
8	<null>
9	<null>
10	<null>
11	<null>
12	<null>
13	<null>
14	<null>
15	General
16	General
17	General

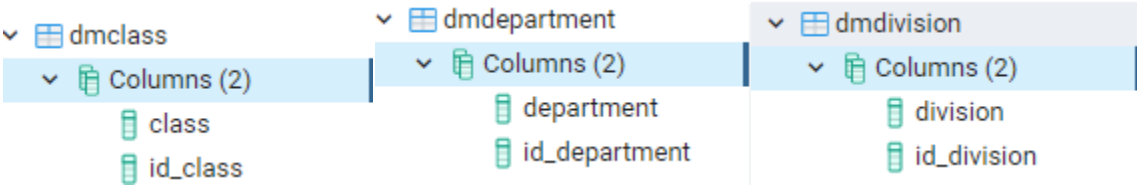
Hasil diatas didapatkan dengan melakukan sorting pada nilai. Jika dilihat menggunakan SQL, hasil yang didaptkan seperti berikut.

division_name character (100)	department character (100)	class_name character (100)
[null]	[null]	[null]
[null]	[null]	[null]
[null]	[null]	[null]
[null]	[null]	[null]
[null]	[null]	[null]
[null]	[null]	[null]

Dapat dinyatakan bahwa nilai null yang tersebar pada 3 kolom diatas saling berkaitan. Sehingga, pada kasus ini saya mengganti nilai null dengan “Unknown”. Setelah penggantian nilai null akan dilakukan sorting untuk mempersiapkan pengambilan nilai unik dan pertambahan id sebagai primary key. Rangkaian ETL selengkapnya seperti berikut.

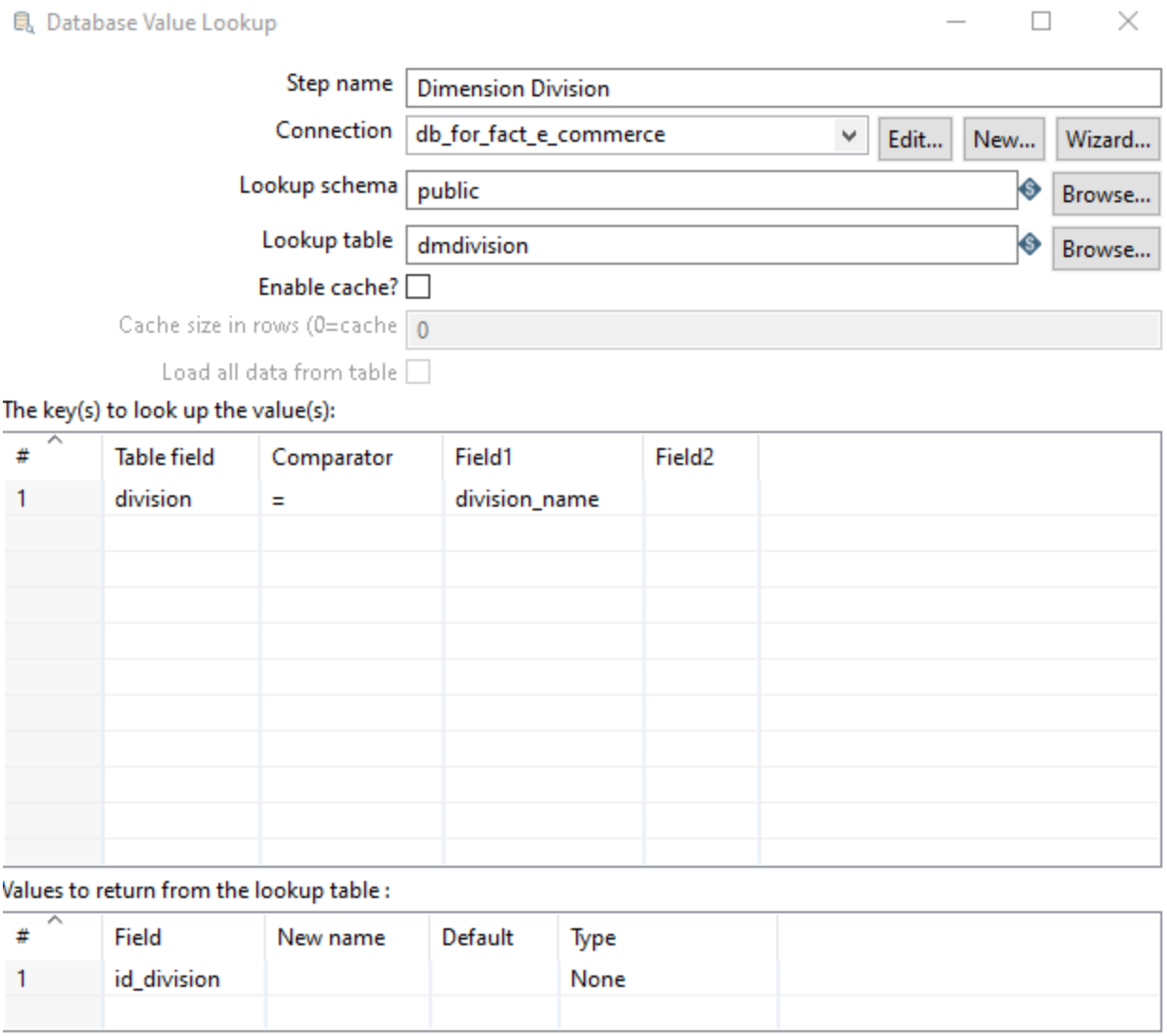


Rangkaian di atas menghasilkan tabel dimensi yang sebelumnya telah divisualikan dengan star schema diatas. Hasil di atas di load pada database yang sama dan diextract kedalam file excel (.xlsx).



b. Tabel Fakta

Hasil dimensi diatas akan siap digunakan untuk menghasilkan tabel fakta. Rangkaian tabel fakta yang dilakukan seperti proses join, lebih tepatnya dengan INNER JOIN. Fitur pada pentaho untuk itu dilakukan menggunakan “Database Lookup” yang memanfaatkan value yang sama untuk pada tabel data real dengan hasil tabel dimensi.



1. Lookup schema -> digunakan untuk mengetahui status tabel dan di database yang dibuat, status tabel yang digunakan adalah public.
2. Lookup table -> digunakan untuk melakukan proses INNER join.
3. Table field -> digunakan untuk tabel yang berisi data real
4. Field1 -> digunakan untuk mengambil hasil dimensi division
5. Comparator -> digunakan untuk membandingkan nilai antar kedua tabel
6. Values to return from the lookup table -> digunakan untuk mengembailkan nilai yang diinginkan. Tabel fakta menyimpan id dari tabel dimensi, sehingga yang dikembalikan adalah id dari dimensi tersebut

Namun karena tabel data real masih memiliki null, maka diperlukan pengubah nilai menjadi ‘Unknown’. Fungsi yang digunakan adalah COALESCE. Fungsi tersebut menggubah nilai pada tabel yang diinputkan dengan value yang diinginkan (‘Unknown’) dan dilakukan measure (Average) dari tabel Rating. Berikut query yang digunakan.

```
SELECT
	COALESCE(division_name,'Unknown') AS division_name,
	COALESCE(department,'Unknown') AS department,
	COALESCE(class_name, 'Unknown') AS class_name,
	AVG("Rating")
FROM review
GROUP BY 1,2,3
```

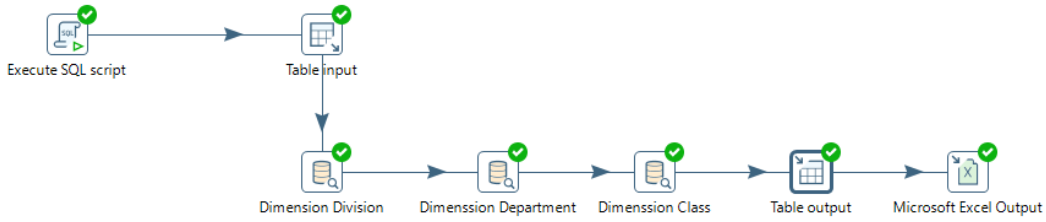
Proses untuk 2 dimensi lainnya sama seperti dimensi division. Hasil dari tabel fakta seperti berikut.

class_name	avg	id_division	id_department	id_class
Dresses	4.1630026809651475	1	2	4
Sweaters	4.2348484848484848	2	5	18
Pants	4.2263473053892216	1	1	14
Knits	4.2083333333333333	2	5	9
Swim	4.1971428571428571	3	3	19
Fine gauge	4.3325301204819277	2	5	5
Blouses	4.1419969894631209	1	5	1
Dresses	4.1332560834298957	2	2	4
Jackets	4.3517915309446254	2	4	7
Skirts	4.2703488372093023	2	1	16
Lounge	4.240343347639485	2	3	12
Unknown	5	4	7	21
Pants	4.325497287522604	2	1	14
Casual bottoms	4.5	1	1	2
Skirts	4.2312811980033278	1	1	16
Layering	4.3767123287671233	3	3	10
Outerwear	4.125	2	4	13
Jeans	4.3916666666666667	2	1	8
Lounge	4.3318777292576419	3	3	12
Legwear	4.2787878787878788	3	3	11
Jackets	4.2518891687657431	1	4	7
Trend	3.8229166666666667	1	6	20
Outerwear	4.2217741935483871	1	4	13
Trend	3.7826086956521739	2	6	20
Shorts	4.2555205047318612	1	1	15
Sweaters	4.1466666666666667	1	5	18
Fine gauge	4.2175182481751825	1	5	5

Hasil tersebut akan difilter kolom mana saja yang digunakan. Filter yang digunakan seperti berikut yang berdasarkan star schema yang digunakan.

#	Table field	Stream field
1	id_division	id_division
2	id_departm...	id_department
3	id_class	id_class
4	avg	avg

Rangkaian lengkap tabel fakta dan hasilnya seperti berikut.



facttable

Columns (4)

id_division

id_department

id_class

avg

Data Output

	id_division integer	id_department integer	id_class integer	avg double precision
1	1	2	4	4.163002680965148
2	2	5	18	4.234848484848484
3	1	1	14	4.226347305389222
4	2	5	9	4.208333333333333
5	3	3	19	4.1971428571428575