Data Generation

Requirements to Source Systems

You must have at least 2 different source systems (2 datasets – 2 CSV files) for the final project. Your 3NF and Star schemas in the DWH will be filled based on the data from both source systems.

The simplest way to have 2 systems is to find 1st dataset and then generate the data for the 2nd source system based on it.

Pay attention:

1. 2 different source systems should have different attributes
2. The same rows in two sources (if any) must have different IDs

Example of 2 datasets:

Dataset 1

time		distributor_name		product_line	brend_id	brend_name	delivery city	count	price	sales channel
May-23		shop.com		phone	123	samsung s21	Istanbul	2	100	online
Apr-23		hp_shop.com		laptop	45	HP Laptop 17	Wroclaw	1	1000	online
Apr-23		hp_shop.com		laptop	45	HP Laptop 17	Viena	1	1000	online

Dataset 2

time	epmployee	distributor	shop address	product	brend number	brend		count	price	sales channel
Apr-23	Name_1	shop.com	Tbilisi, Rustaveli Avenu	smartphone	21	samsung s21		3	90	offline
Apr-23	Name_2	hp_shop.com	Vilnius, Gedeminas, st	laptop	2	apple MacBook Air 15,3		1	2000	online

How to generate sources:

1st option: you can find 1 dataset and create 2nd based on 1st using **excel**.

You can get any dataset from Google cloud **or other sources**.

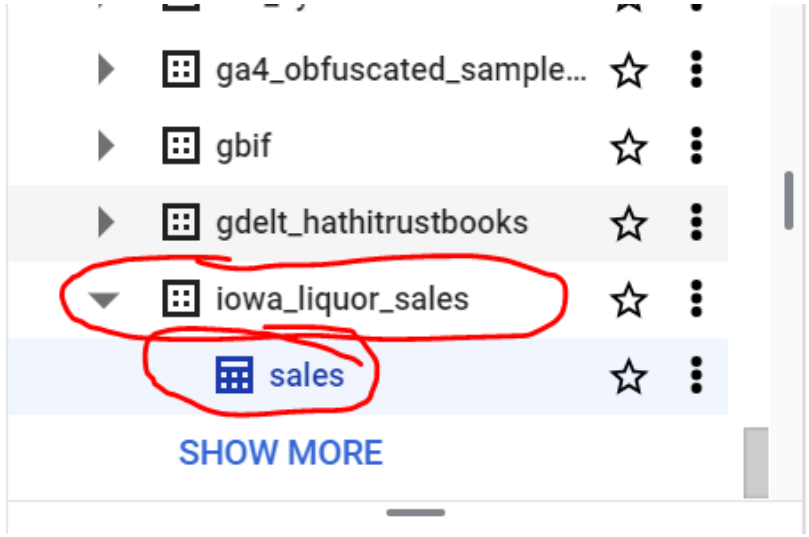
Examples:

1. iowa_liquor_sales – BigQuery – My First Project – Google Cloud console
2. Amazon Sales Dataset (kaggle.com)
3. https://data.nasa.gov/
4. https://github.com/search?q=dataset&type=repositories
5. https://data.gov/
6. https://data.fivethirtyeight.com/

How to generate sources:

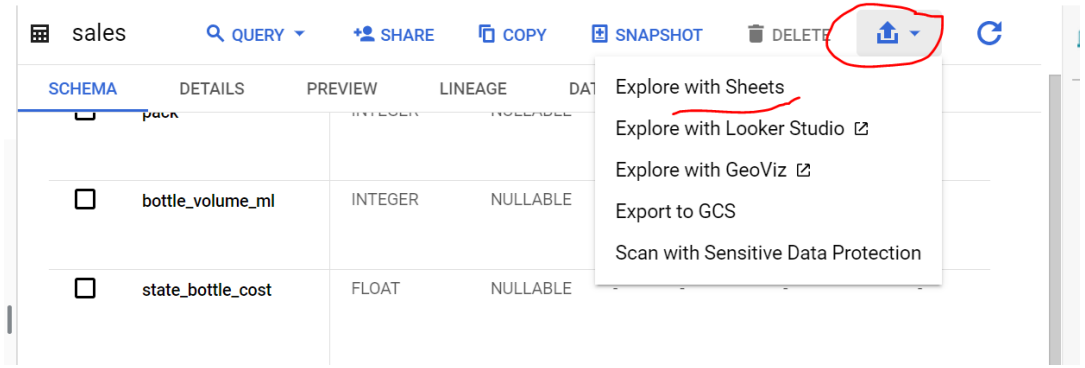
Here is a steps which show you how to download source

1. Choose your source and open SALES transaction table:

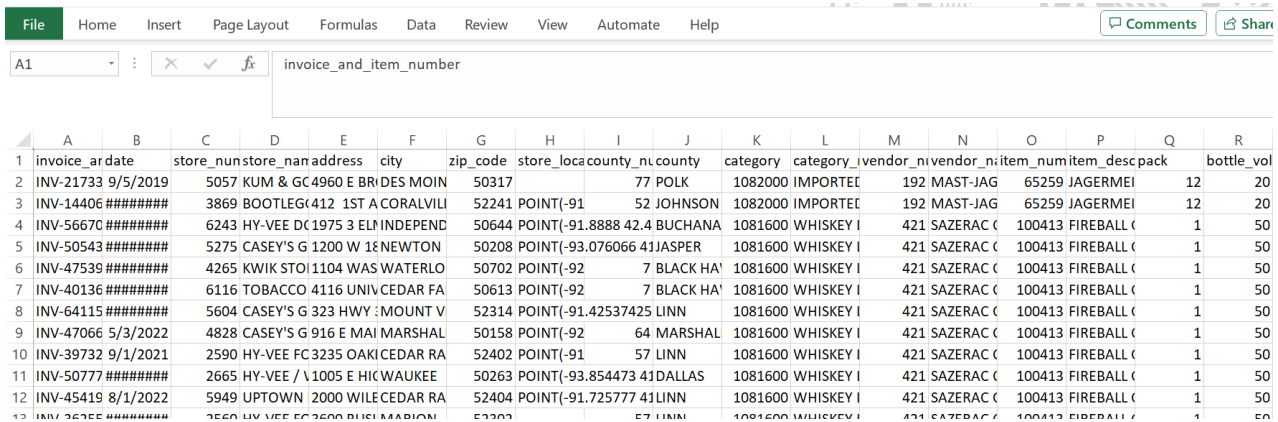


How to generate sources:

2. Open Query section and write “select * from …..”:



3. Save the result in CSV file. Here is an example of CSV file:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	invoice_ar	date	store_nun	store_nan	address	city	zip_code	store_loca	county_nu	county	category	category_	vendor_nu	vendor_na	item_num	item_desc	pack	bottle_vol
2	INV-21733	9/5/2019	5057	KUM & GC	4960 E BR	DES MOIN	50317		77	POLK	1082000	IMPORTEI	192	MAST-JAG	65259	JAGERMEI	12	20
3	INV-14406	########	3869	BOOTLEG(	412 1ST A	CORALVILI	52241	POINT(-91	52	JOHNSON	1082000	IMPORTEI	192	MAST-JAG	65259	JAGERMEI	12	20
4	INV-56670	########	6243	HY-VEE D(	1975 3 ELI	INDEPEND	50644	POINT(-91.8888	42.4	BUCHANA	1081600	WHISKEY I	421	SAZERAC (	100413	FIREBALL (	1	50
5	INV-50543	########	5275	CASEY'S G	1200 W 1E	NEWTON	50208	POINT(-93.076066	41	JASPER	1081600	WHISKEY I	421	SAZERAC (	100413	FIREBALL (	1	50
6	INV-47539	########	4265	KWIK STO	1104 WAS	WATERLO	50702	POINT(-92	7	BLACK HA	1081600	WHISKEY I	421	SAZERAC (	100413	FIREBALL (	1	50
7	INV-40136	########	6116	TOBACCO	4116 UNIV	CEDAR FA	50613	POINT(-92	7	BLACK HA	1081600	WHISKEY I	421	SAZERAC (	100413	FIREBALL (	1	50
8	INV-64115	########	5604	CASEY'S G	323 HWY :	MOUNT V	52314	POINT(-91.42537425		LINN	1081600	WHISKEY I	421	SAZERAC (	100413	FIREBALL (	1	50
9	INV-47066	5/3/2022	4828	CASEY'S G	916 E MAI	MARSHAL	50158	POINT(-92	64	MARSHAL	1081600	WHISKEY I	421	SAZERAC (	100413	FIREBALL (	1	50
10	INV-39732	9/1/2021	2590	HY-VEE FC	3235 OAK	CEDAR RA	52402	POINT(-91	57	LINN	1081600	WHISKEY I	421	SAZERAC (	100413	FIREBALL (	1	50
11	INV-50777	########	2665	HY-VEE / \	1005 E HI(	WAUKEE	50263	POINT(-93.854473	41	DALLAS	1081600	WHISKEY I	421	SAZERAC (	100413	FIREBALL (	1	50
12	INV-45419	8/1/2022	5949	UPTOWN	2000 WILE	CEDAR RA	52404	POINT(-91.725777	41	LINN	1081600	WHISKEY I	421	SAZERAC (	100413	FIREBALL (	1	50

How to generate sources:

4. Now we need to create 2nd source from generated one.

Lets leave 1st source as it is and 2nd source will be generated as ONLINE sales (as on 1st example).

4.1. Change stores to ONLINE:

Before:

	C	D	E	F	G	H	I	J	
	store_number	store_name	address	city	zip_code	store_locati	county_r	county	c
9	5057	KUM & GO #535	4960 E BROADW	DES MOIN	50317		77	POLK	
#	3869	BOOTLEGGIN' B	412 1ST AVE	CORALVILL	52241	POINT(-91.5	52	JOHNSON	
#	6243	HY-VEE DOLLAR	1975 3 ELMS RO	INDEPEND	50644	POINT(-91.8888	42.45	BUCHANA	
#	5275	CASEY'S GENER/	1200 W 18TH ST	NEWTON	50208	POINT(-93.076066	41	JASPER	
#	4265	KWIK STOP 3 / V	1104 WASHING	WATERLO	50702	POINT(-92.3	7	BLACK HA	
#	6116	TOBACCO OUTL	4116 UNIVERSIT	CEDAR FAI	50613	POINT(-92.4	7	BLACK HA	
#	5604	CASEY'S GENER/	323 HWY 30 W	MOUNT VI	52314	POINT(-91.42537425		LINN	
2	4828	CASEY'S GENER/	916 E MAIN ST	MARSHALI	50158	POINT(-92.8	64	MARSHAL	
1	2590	HY-VEE FOOD S	3235 OAKLAND	CEDAR RA	52402	POINT(-91.6	57	LINN	

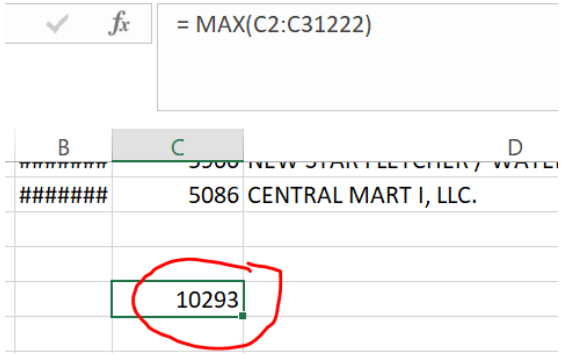
After:

C	D	E	
store_numb	store_name	store_location	c
5057	KUM & GO #535 / DES MOINES	Online	
3869	BOOTLEGGIN' BARZINI'S FIN	Online	
6243	HY-VEE DOLLAR FRESH / INDEPENDENCE	Online	

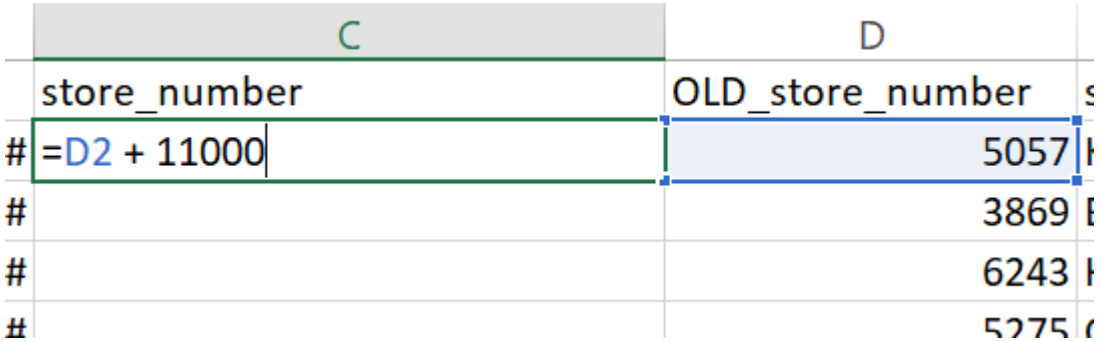
You can see that in Online source we have only 3 attributes: store_number, store_name and store_location as Online shop do not have physical address.

How to generate sources:

4.2. We have a rule: “The same rows in two sources (if any) must have different IDs ”. We should change store_number in our example. It is very easy. We can just check which ID value is MAX:



To make our IDs unique we can add 11000 to each value (as example):



How to generate sources:

Result:

C	D	
store_number	OLD_store_number	s
16057	5057	K
14869	3869	B
17243	6243	H
16275	5275	C
15265	4265	K

“OLD_store_number” column can be deleted

You can (should) do it for all ID numbers.

4.3. We should change store_names to WEB names:

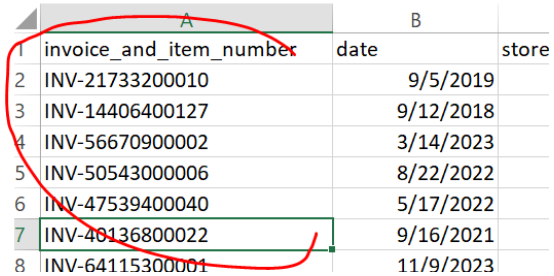
We can just take 10-20 letters from store_name and convert it into web-site:

```
=CONCAT(LEFT(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(E2,"#",""),"/",""),"'",""),"  ",""),20),".com")
```

C	D	E	F	
number	store_name	OLD_store_name	store_location	cat
16057	KUM&GO535DESMOINES.com	KUM & GO #535 / DES MOINES	Online	10
14869	BOOTLEGGINBARZINISFI.com	BOOTLEGGIN' BARZINI'S FIN	Online	10
17243	HY-VEEDOLLARFRESHIND.com	HY-VEE DOLLAR FRESH / INDEPENDENCE	Online	10

How to generate sources:

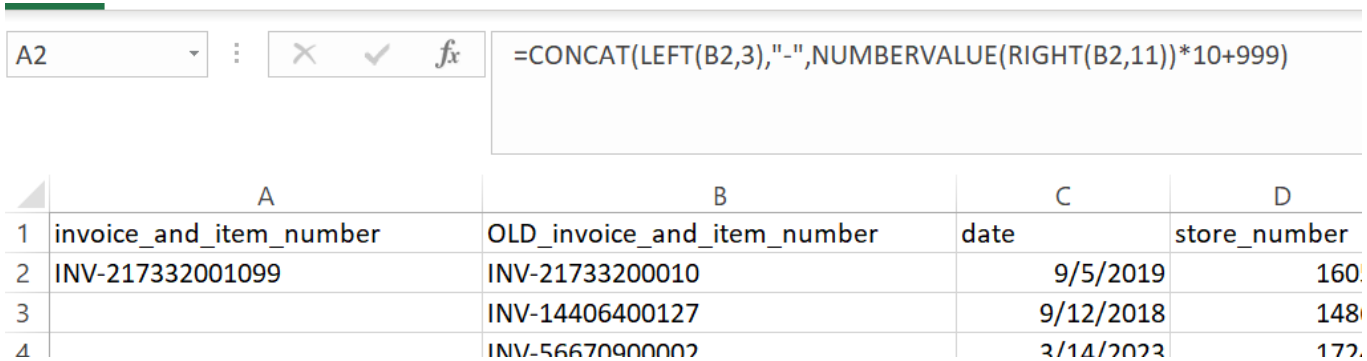
4.3. Last step I will show how to change Natural keys which is a combination of letters and numbers:



	A	B	
1	invoice_and_item_number	date	store
2	INV-21733200010	9/5/2019	
3	INV-14406400127	9/12/2018	
4	INV-56670900002	3/14/2023	
5	INV-50543000006	8/22/2022	
6	INV-47539400040	5/17/2022	
7	INV-40136800022	9/16/2021	
8	INV-64115300001	11/9/2023	

Here is 2 ways:

1. We can generate absolutely new number like we did before.
2. We can use existing, but cut it/multiply/divide/concatenate:



A2

=CONCAT(LEFT(B2,3),"-",NUMBERVALUE(RIGHT(B2,11))*10+999)

	A	B	C	D
1	invoice_and_item_number	OLD_invoice_and_item_number	date	store_number
2	INV-217332001099	INV-21733200010	9/5/2019	160
3		INV-14406400127	9/12/2018	148
4		INV-56670900002	3/14/2023	172

How to generate sources:

Here is a result of our changes (a part of our 2nd source):

	A	B	C	D	E	
1	invoice_and_item_nu	date	store_number	store_name	store_location	ca
2	INV-217332001099	9/5/2019	16057	KUM&GO535DESMOINES.com	Online	1
3	INV-144064002269	9/12/2018	14869	BOOTLEGGINBARZINISFI.com	Online	1
4	INV-566709001019	3/14/2023	17243	HY-VEEDOLLARFRESHIND.com	Online	1
5	INV-505430001059	8/22/2022	16275	CASEYSGENERALSTORE24.com	Online	1
6	INV-475394001399	5/17/2022	15265	KWIKSTOP3WATERLOO.com	Online	1
7	INV-401368001219	9/16/2021	17116	TOBACCOOUTLETPLUS561.com	Online	1
8	INV-641153001009	11/9/2023	16604	CASEYSGENERALSTORE15.com	Online	1
9	INV-470662001069	5/3/2022	15828	CASEYSGENERALSTORE32.com	Online	1
10	INV-397329002899	9/1/2021	13590	HY-VEEFOODSTORE5CEDA.com	Online	1

You can implement same changes using SQL or Python as well as Excel.

How to generate sources:

In next video you can see a full source transformation.

<epam>

Thank you!