Presentation

on

Fact and Dimension Table

- **Baburam Shrestha**

Fact and Dimension Table

Fact Table:

➢ Contains quantifiable information/data about the business process

➢ Data is loaded at last after all dimension tables are loaded due to foreign key constraints

➢ It contains the foreign key which is primary key in fact table

➢ For star and snowflake schema, it generally lies in the center

Dimension Table:

➢ Acts as a companion to a fact table that provides meaning to it

➢ Data is loaded before loading into the fact table

➢ It contains the primary key which is foreign key in fact table

➢ For star and snowflake schema it generally lies in the edges

Star and Snowflake Schema

Star Schema

- ➢ Has fact and dimension tables
- ➢ Use more space as data redundancy is high
- ➢ Less time to execute queries

Snowflake Schema

- ➢ Fact ,dimension, and sub-dimension table
- ➢ Less space due to less data redundancy
- ➢ More time to execute queries as multiple join operations may be needed

Sample Data: SALES.csv

SALE_ID	ORDER_ID	ORDER_DATE	SHIP_DATE	SHIP_MODE	CUSTOMER_ID	CUSTOMER_NA	SEGMENT	COUNTRY	CITY
1	CA-2017-15...	2017-08-11	2017-11-11	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson
14	CA-2017-16...	2017-05-12	2017-10-12	Standard Cl...	IM-15070	Irene Maddox	Consumer	United States	Seattle
36	CA-2017-117...	2017-08-12	2017-10-12	First Class	GH-14485	Gene Hale	Corporate	United States	Richardson
69	CA-2015-10...	2015-05-12	2015-10-12	Standard Cl...	BS-11590	Brendan Sw...	Corporate	United States	Gilbert
76	US-2018-11...	2018-09-12	2018-11-12	First Class	KB-16600	Ken Brennan	Corporate	United States	Houston
84	CA-2016-14...	2016-03-09	2016-08-09	Standard Cl...	JC-16105	Julie Creight...	Corporate	United States	Durham
89	CA-2017-15...	2017-05-04	2017-10-04	Second Class	GM-14455	Gary Mitchum	Home Office	United States	Houston
96	US-2018-10...	2018-06-11	2018-12-11	Standard Cl...	RB-19705	Roger Barcio	Home Office	United States	Portland
97	CA-2018-16...	2018-09-11	2018-11-11	Second Class	PN-18775	Parhena Nor...	Home Office	United States	New York City
99	CA-2017-14...	2017-06-09	2017-11-09	Standard Cl...	ER-13855	Elpida Ritten...	Corporate	United States	Saint Paul
103	CA-2017-12...	2017-01-12	2017-04-12	Second Class	GZ-14470	Gary Zandus...	Consumer	United States	Rochester

COUNTRY	CITY	STATE	POSTAL_CODE	REGION	PRODUCT_ID	CATEGORY	SUB_CATEGORY	PRODUCT_NAM	SALES
United States	Henderson	Kentucky	42420	South	FUR-BO-100...	Furniture	Bookcases	Bush Somer...	261.96
United States	Seattle	Washington	98103	West	OFF-BI-100...	Office Suppli...	Binders	Fellowes PB...	407.976
United States	Richardson	Texas	75080	Central	TEC-PH-100...	Technology	Phones	GE 30524EE4	1097.544
United States	Gilbert	Arizona	85234	West	TEC-PH-100...	Technology	Phones	netTALK DU...	167.968
United States	Houston	Texas	77041	Central	OFF-BI-100...	Office Suppli...	Binders	Economy Bin...	1.248
United States	Durham	North Carolina	27707	South	OFF-EN-100...	Office Suppli...	Envelopes	Jet-Pak Rec...	200.984
United States	Houston	Texas	77095	Central	OFF-ST-100...	Office Suppli...	Storage	Eldon Porta...	158.368
United States	Portland	Oregon	97206	West	OFF-BI-100...	Office Suppli...	Binders	Flexible Leat...	5.682
United States	New York City	New York	10009	East	FUR-FU-100...	Furniture	Furnishings	9-3/4 Diame...	96.53
United States	Saint Paul	Minnesota	55106	Central	OFF-AP-100...	Office Suppli...	Appliances	Fellowes Ba...	77.88
United States	Rochester	Minnesota	55901	Central	OFF-PA-100...	Office Suppli...	Paper	Universal Pr...	23.92

Sample Dataset (SALES.csv)

Dataset link:
https://www.kaggle.com/datasets

-This data set is about order details of certain sales stores
-It contains 18 fields and more than 2000 rows.

For this data, it is possible to create create fact and dimension table using the star schema as:
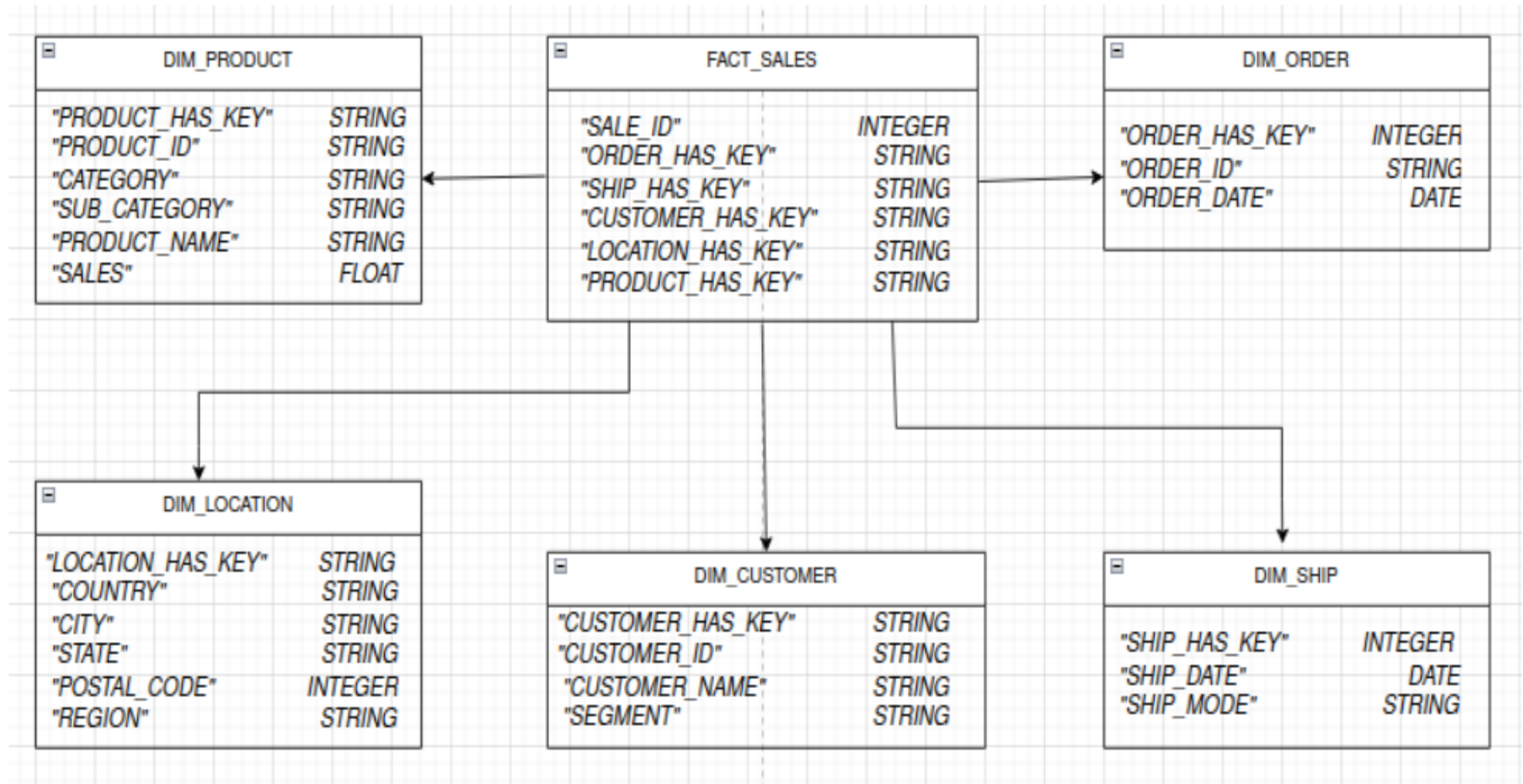
Fact Table:
- FACT_SALES

Dimension Tables:
- ORDER_DIM,
- SHIP_DIM,
- CUSTOMER_DIM,
- LOCATION_DIM and
- PRODUCT_DIM

Why did I choose Star Schema?
- Sales data set has simple data that can be normalized with fact and dimension table, and
- For the fact access

SALES	
"SALE_ID"	INTEGER
"ORDER_ID"	STRING
"ORDER_DATE"	DATE
"SHIP_DATE"	DATE
"SHIP_MODE"	STRING
"CUSTOMER_ID"	STRING
"CUSTOMER_NAME"	STRING
"SEGMENT"	STRING
"COUNTRY"	STRING
"CITY"	STRING
"STATE"	STRING
"POSTAL_CODE"	INTEGER
"REGION"	STRING
"PRODUCT_ID"	STRING
"CATEGORY"	STRING
"SUB_CATEGORY"	STRING
"PRODUCT_NAME"	STRING
"SALES"	FLOAT

Fact and Dimension Table

DIM_PRODUCT

"PRODUCT_HAS_KEY"	STRING
"PRODUCT_ID"	STRING
"CATEGORY"	STRING
"SUB_CATEGORY"	STRING
"PRODUCT_NAME"	STRING
"SALES"	FLOAT

FACT_SALES

"SALE_ID"	INTEGER
"ORDER_HAS_KEY"	STRING
"SHIP_HAS_KEY"	STRING
"CUSTOMER_HAS_KEY"	STRING
"LOCATION_HAS_KEY"	STRING
"PRODUCT_HAS_KEY"	STRING

DIM_ORDER

"ORDER_HAS_KEY"	INTEGER
"ORDER_ID"	STRING
"ORDER_DATE"	DATE

DIM_LOCATION

"LOCATION_HAS_KEY"	STRING
"COUNTRY"	STRING
"CITY"	STRING
"STATE"	STRING
"POSTAL_CODE"	INTEGER
"REGION"	STRING

DIM_CUSTOMER

"CUSTOMER_HAS_KEY"	STRING
"CUSTOMER_ID"	STRING
"CUSTOMER_NAME"	STRING
"SEGMENT"	STRING

DIM_SHIP

"SHIP_HAS_KEY"	INTEGER
"SHIP_DATE"	DATE
"SHIP_MODE"	STRING

Name	Type	Remarks
SALES_WAREHOUSE	Warehouse	
SALES_DATABASE	Database	
SALES_SCHEMA	Schema	To store the sales table
FACT_DIM_SCHEMA	Schema	To store the fact and dimension table to target schema
SALE_STAGE	Stage	To store the raw dataset(SALES.csv) temporarily
SALES	Table	Columns: "SALE_ID" , "ORDER_ID", "ORDER_DATE", "SHIP_DATE", "SHIP_MODE","CUSTOMER_ID", "CUSTOMER_NAME" , "SEGMENT" ,"COUNTRY","CITY" ,"STATE", "POSTAL_CODE", "REGION", "PRODUCT_ID", "CATEGORY", "SUB_CATEGORY", "PRODUCT_NAME", "SALES"
FACT_SALES	Fact Table	Columns: "SALE_ID" ,"ORDER_HAS_KEY","SHIP_HAS_KEY","CUSTOMER_HAS_KEY", "LOCATION_HAS_KEY","PRODUCT_HAS_KEY"
DIM_ORDER	Dimension Table	Columns: "ORDER_HAS_KEY","ORDER_ID", "ORDER_DATE"
DIM_SHIP	Dimension Table	Columns: "SHIP_HAS_KEY",, "SHIP_DATE", "SHIP_MODE"
DIM_LOCATION	Dimension Table	Columns: "CUSTOMER_HAS_KEY","CUSTOMER_ID", "CUSTOMER_NAME" , "SEGMENT"
DIM_CUSTOMER	Dimension Table	Columns: "LOCATION_HAS_KEY","COUNTRY","CITY" ,"STATE", "POSTAL_CODE", "REGION",
DIM_PRODUCT	Dimension Table	Columns: "PRODUCT_HAS_KEY","PRODUCT_ID", "CATEGORY", "SUB_CATEGORY", "PRODUCT_NAME", "SALES"

COPY_FACT_SALES()	Procedure	To make the target fact table and insert the data from sales table
COPY_DIM_ORDER()	Procedure	To make the target dimension table and insert the data from sales table
COPY_DIM_SHIP()	Procedure	To make the target dimension table and insert the data from sales table
COPY_DIM_CUSTOMER()	Procedure	To make the target dimension table and insert the data from sales table
COPY_DIM_LOCATION()	Procedure	To make the target dimension table and insert the data from sales table
COPY_DIM_PRODUCT()	Procedure	To make the target dimension table and insert the data from sales table
CSV_FILE_FORMAT	File Format	Csv file format

Work Flow Chart

SALES_WAREHOUSE

SALES_DATABSSE

SALES_STAGE

SALES_SCHEMA

STORED PROCEDURE

FACT_DIM_SCHEMA

SALES.csv.gz

SALES

CSV_FILE_FORMAT

COPY_FACT_SALES()

FACT_SALES

COPY_DIM_ORDER()

DIM_ORDER

COPY_DIM_SHIP()

DIM_SHIP

COPY_DIM_CUSTOMER()

DIM_CUSTOMER

COPY_DIM_LOCATION()

DIM_LOCATION

COPY_DIM_PRODUCT()

DIM_PRODUCT
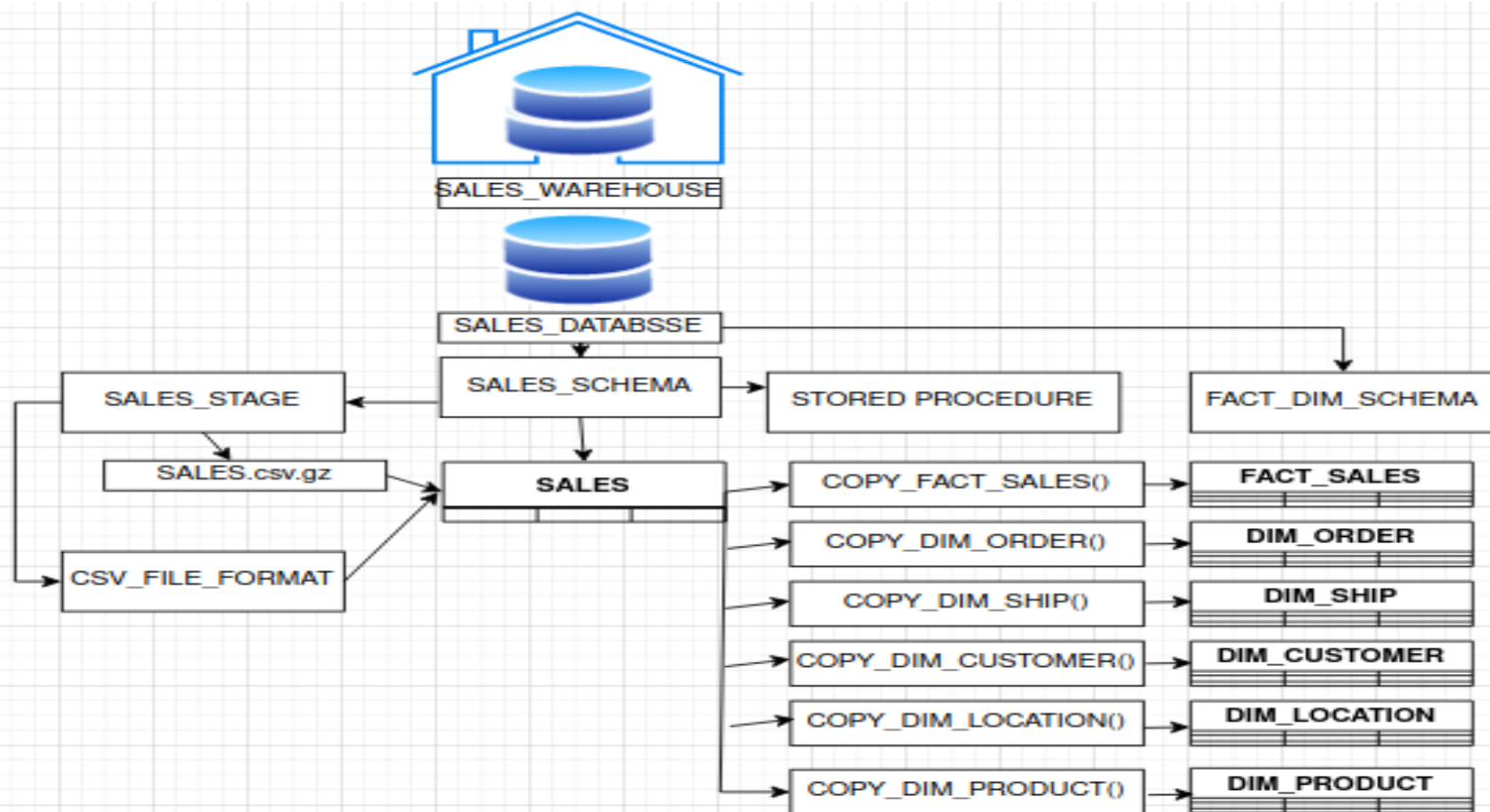
DATA FROM FACT_SALES TABLE

SALE_ID	ORDER_HAS_KEY	SHIP_HAS_KEY	CUSTOMER_HAS_KEY	LOCATION_HAS_KEY	PRODUCT_HAS_KEY
1	-3167119526233458530	-4847909140316461210	4276261313029538929	3771190386958874132	14033436625530933978
14	-7550465111006103039	-8344712918650322306	6901003581497509693	4890559700903138221	-8856566870948700400
36	-6976892719817715820	2157506248436069439	-7786783779261884382	-7312717150105753501	18491987975354387776
69	4198188682122815243	-4837678753108482897	-2167329333978851334	-4493725974168302790	-9819404079896366587
76	3904808208657385001	-762238745111245991	-6338985466377324839	-7186855032880248468	-2650941558079039525
84	-4986266459459473402	2977896322095346702	4044925526464553191	-3413615522219311362	3328523468741916194
89	2283009200683985904	8601452806769915934	-1534896954822540851	6190122100539369043	-6152233193102534447
96	6092324598220716871	7280760766699576989	-7129026584076646642	-8321011656027551075	4291672650359804393

DATA FROM DIM_ORDER TABLE

ORDER_HAS_KEY	ORDER_ID	ORDER_DATE
-3167119526233458530	CA-2017-152156	2017-08-11
-7550465111006103039	CA-2017-161389	2017-05-12
-6976892719817715820	CA-2017-117590	2017-08-12
4198188682122815243	CA-2015-106376	2015-05-12
3904808208657385001	US-2018-118038	2018-09-12
-4986266459459473402	CA-2016-149734	2016-03-09
2283009200683985904	CA-2017-159695	2017-05-04
6092324598220716871	US-2018-109484	2018-06-11
-6438232367784042250	CA-2018-161018	2018-09-11

DATA FROM DIM_SHIP TABLE

SHIP_HAS_KEY	SHIP_DATE	SHIP_MODE
-4847909140316461210	2017-11-11	Second Class
-8344471291865032306	2017-10-12	Standard Class
2157506248436069439	2017-10-12	First Class
-4837678753108482897	2015-10-12	Standard Class
-762238745111245991	2018-11-12	First Class
2977896322095346702	2016-08-09	Standard Class
8601452806769915934	2017-10-04	Second Class
7280760766699576989	2018-12-11	Standard Class
-5169073561908891185	2018-11-11	Second Class

DATA FROM DIM_CUSTOMER TABLE

CUSTOMER_HAS_KEY	CUSTOMER_ID	CUSTOMER_NAME	SEGMENT
42762613130229538929	CG-12520	Claire Gute	Consumer
6901003581497509693	IM-15070	Irene Maddox	Consumer
-77867837779261884382	GH-14485	Gene Hale	Corporate
-21673293333978851334	BS-11590	Brendan Sweed	Corporate
-63389854663377324839	KB-16600	Ken Brennan	Corporate
40449255526464553191	JC-16105	Julie Creighton	Corporate
-15348969548225240851	GM-14455	Gary Mitchum	Home Office

DATA FROM DIM_LOCATION TABLE

LOCATION_HAS_KEY	COUNTRY	CITY	STATE	POSTAL_CODE	REGION
37711903869958874132	United States	Henderson	Kentucky	42420	South
48905597009031382211	United States	Seattle	Washington	98103	West
-7312717150105753501	United States	Richardson	Texas	75080	Central
-4493725974168302790	United States	Gilbert	Arizona	85234	West
-7186855032880248468	United States	Houston	Texas	77041	Central
-341361552221931362	United States	Durham	North Carolina	27707	South
6190122100539369043	United States	Houston	Texas	77095	Central
-8321011656027551075	United States	Portland	Oregon	97206	West
6662592650660234844	United States	New York City	New York	10009	East
5885409604060017159	United States	Saint Paul	Minnesota	55106	Central

DATA FROM DIM_PRODUCT TABLE

PRODUCT_HAS_KEY	PRODUCT_ID	CATEGORY	SUB_CATEGORY	PRODUCT_NAME	SALES
1403343662530933978	FUR-BO-10001798	Furniture	Bookcases	Bush Somerset Collection Book...	261.96
-8856566870948700400	OFF-BI-10003656	Office Supplies	Binders	Fellowes PB200 Plastic Comb B...	407.976
1849198797535438776	TEC-PH-10004977	Technology	Phones	GE 30524EE4	1097.544
-981940407989636587	TEC-PH-10002726	Technology	Phones	netTALK DUO VoIP Telephone S...	167.968
-2650941558079039525	OFF-BI-10004182	Office Supplies	Binders	Economy Binders	1.248
3328523468741916194	OFF-EN-10000927	Office Supplies	Envelopes	Jet-Pak Recycled Peel 'N' Seal ...	200.984
-6152233193102534447	OFF-ST-10003442	Office Supplies	Storage	Eldon Portable Mobile Manager	158.368
4291672650359804393	OFF-BI-10004738	Office Supplies	Binders	Flexible Leather- Look Classic ...	5.682
7202632461898611291	FUR-FU-10000629	Furniture	Furnishings	9-3/4 Diameter Round Wall Clock	96.53
8892074227930352992	OFF-AP-10000358	Office Supplies	Appliances	Fellowes Basic Home/Office Se...	77.88