arnoldHicran_ProjectUpdate4

April 19, 2023

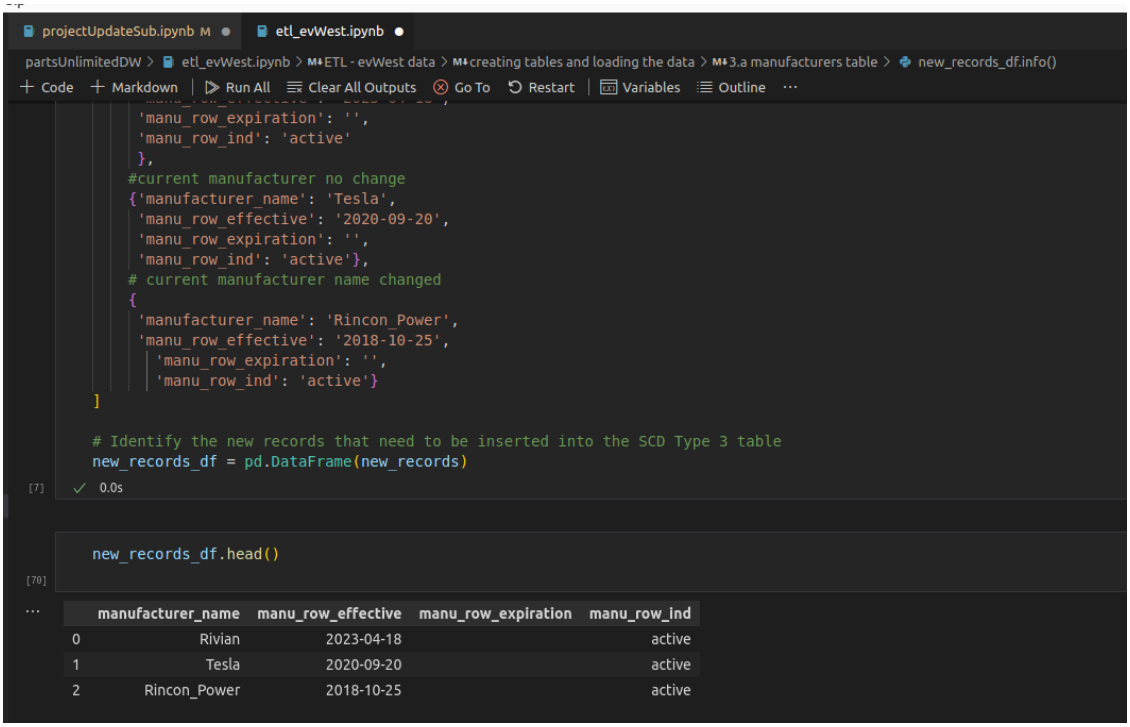
Overview of the Assignment: Let's focus on SCD maintenance of the load process and analysis and presentation.

1 Part 9: Provide updated code and screenshots of loading a delta into the one of the dimensions which is either SCD2 or SCD3. Your code should show how SCD2 or SCD3 is maintained when changes or a brand-new record is detected

I am going to load scd 2 table here is the steps that we need to do 1- check if there is exsisting row 2- if exsist ignore 3- if new add new 4- if there is an adjustment add new one

I faced some issues I noticed in my manufacturer table I have only the manufacturer name, so the only thing will change is the manufacturer name, therefore it does not make sense for me to keep this as scd2, I believe I should switch to scd 3 and change prodocut as scd type 2. Therefore once again I need to adjust the scd types and update scd columns.

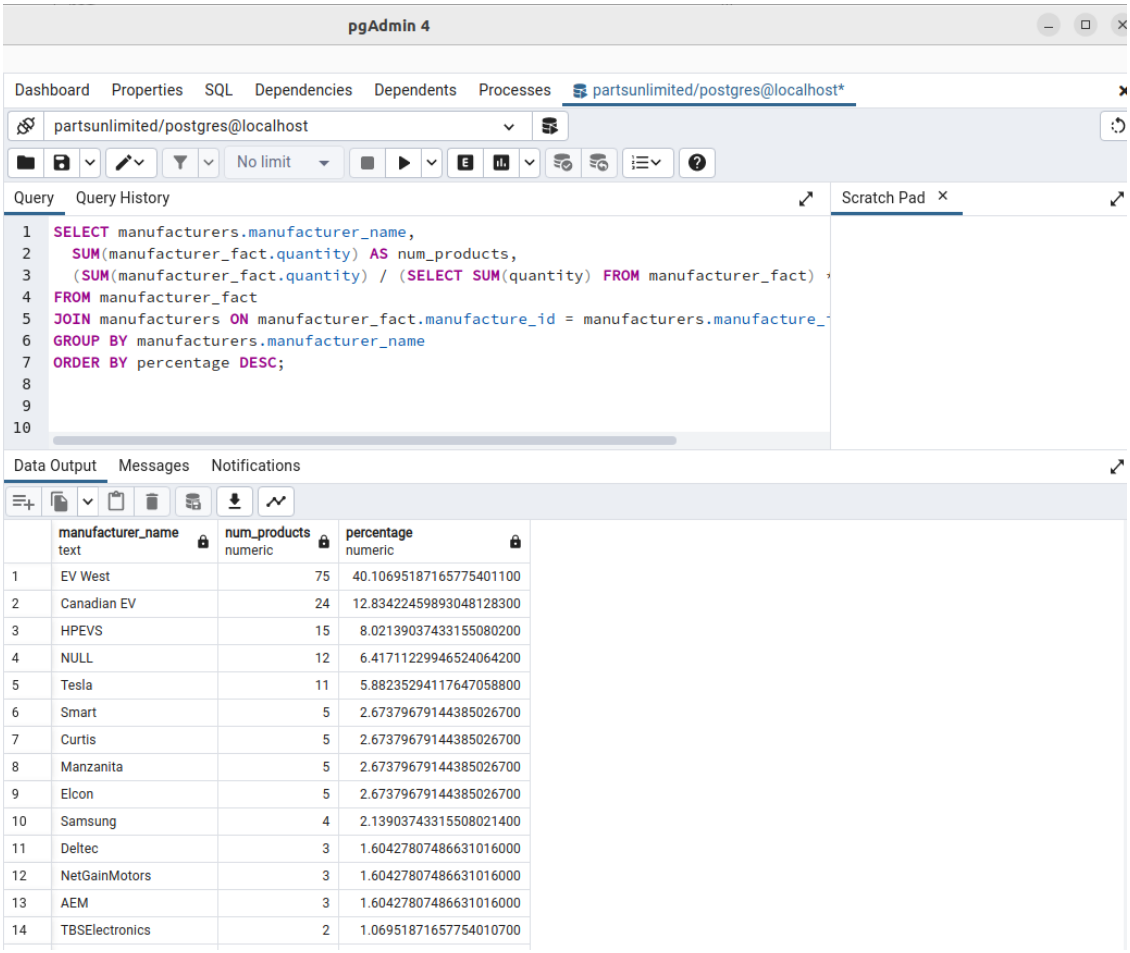
I am facing some design issues in different steps of this project, and I had to go back and fix it, and it is taking more than I anticipated. I fixed the data quality issues from last week, but I realized the incorrect scd type choice in this step. I was planning to load the table as a df and use pandas merge columns to check to see the difference between source and database to determine the new records, but I ran out of time.



2 Part 10: Outline one of the business questions you outlined in part 2 and answer it with an analytical query – it will be important to show complexity within the query (i.e. think about pivots, window functions, inline views, aggregates etc).

What is the current range of EV parts manufacturers and their product offerings that Parts Unlimited is working with, and how can this information be leveraged to optimize their product mix and pricing strategy for increased revenue in the growing EV market? Additionally, how has this range of manufacturers and their product offerings evolved over time, and what trends can be identified for future business planning?

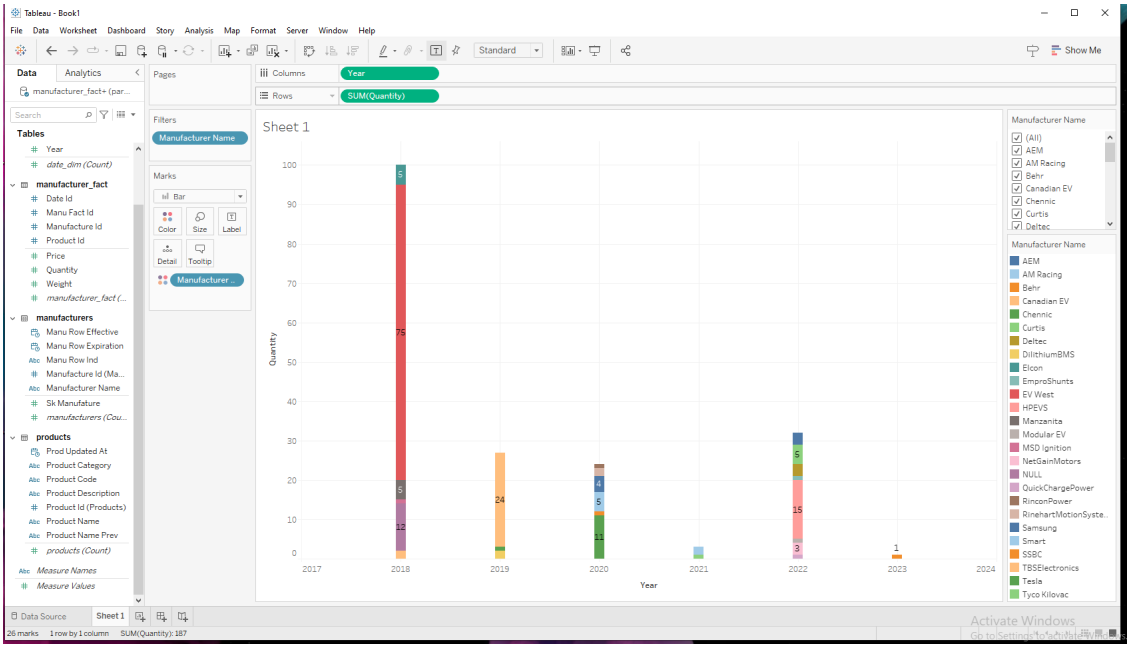
- In this query we are seing the percentage of the each manufacturers. It shows which manufacturers has the monst amount of products. I will add this year column, how many how many manufacturers , how many prodocts to show a trend



	manufacturer_name text	num_products numeric	percentage numeric
1	EV West	75	40.10695187165775401100
2	Canadian EV	24	12.83422459893048128300
3	HPEVS	15	8.02139037433155080200
4	NULL	12	6.41711229946524064200
5	Tesla	11	5.88235294117647058800
6	Smart	5	2.67379679144385026700
7	Curtis	5	2.67379679144385026700
8	Manzanita	5	2.67379679144385026700
9	Elcon	5	2.67379679144385026700
10	Samsung	4	2.13903743315508021400
11	Deltec	3	1.60427807486631016000
12	NetGainMotors	3	1.60427807486631016000
13	AEM	3	1.60427807486631016000
14	TBSElectronics	2	1.06951871657754010700

2.1 Part 11: Outline one of the business questions you outlined in part 2 (can be same as part 10) and answer it with a Tableau or PowerBI visualization. In a single sentence explain why you selected this particular visualization to answer this question.

this query shows the new manufacturers and their and how many new products they have bring to partsUnlimited. This particular visualization you can see the years that there are few new manufacturers and new products and the opposite.



3 Part 1: Are you working on your own or with a partner? If with a partner provide their name. If on your own, just state that this is the case.

- I have adjusted question number 5 with a new question
- I have removed EV car dataset (I could not connect this data set with any other data-sets)
- I have included a new data txt. I noticed that my manufacturer dimentions and are not connected to the rest of the schema, therefore I needed a dimention table to connect them. The location dimention is the only one that I could connect them. However unfortunatelly this data is not available but in my dataset I have only 25 manufacturers, I quicly searched locations of these companies and saved as txt file.
- I have updated ERD schema: I removed the old data set and connected location
- I have adjusted the products table scd type with the scd type 2
- I have added a new time dimension for manufacturers fact table to track products and manufacturer dates

I am working alone in this project

4 Part 2: Determine the project scope

- **In a short paragraph, describe the topic you wish to explore –an update if any**
- **Update the five business questions that your data warehouse will answer.**

Project Description The topic I wish to explore for my data warehouse project is the management of Parts Unlimited's EV parts business. This topic was inspired by "The Unicorn Project," which describes the challenges and opportunities of digital transformation in a large organization. Specifically, I plan to focus on storing and analyzing data related to charging stations, EV products, customer purchases, and geographic location. Parts Unlimited already sells EV parts, and my goal

is to improve the organization's data management, reporting, and analysis capabilities related to this business.

Business Questions:

Business Question 1
Is there a correlation between the number of EV charging stations in a particular area, the number of EV cars registered in that area, and the time period in which they were registered? And if so, how can we use this information to optimize our expansion strategy and better serve our customers over time?

Parts Unlimited Business Development team is interested in using this information to plan and prioritize their expansion strategy for EV parts and charging station installations. The marketing team can use this information to tailor their marketing campaigns and promotions to specific regions.

Business Question 2
What is the current range of EV parts manufacturers and their product offerings that Parts Unlimited is working with, and how can this information be leveraged to optimize their product mix and pricing strategy for increased revenue in the growing EV market? Additionally, how has this range of manufacturers and their product offerings evolved over time, and what trends can be identified for future business planning?

Purchasing team is interested in this question , they are responsible for sourcing and procuring products for Parts Unlimited, including EV parts from different manufacturers. By understanding the current range of EV parts manufacturers and their product offerings, the Procurement department can make informed decisions about which manufacturers to work with, what products to stock, and how to optimize their product mix and pricing strategy for increased revenue in the growing EV market

Business Question Q3
How does the popularity of different EV models and electric vehicle types vary by geographic region, and how can Parts Unlimited use this information to target their marketing and sales efforts?

The marketing and sales department in Parts Unlimited can use the information to target their efforts more effectively. For example, if a certain geographic region shows a higher preference for a particular EV model or plug type, the marketing and sales team can focus their promotional activities and campaigns in that region to increase sales. They can also use this information to tailor their messaging and product offerings to better meet the needs and preferences of customers in each region.

Business Question Q4
What is the relationship between the location and price of existing EV charging stations over time, and how can this information be used to determine the feasibility of adding new charging stations in the vicinity of Parts Unlimited's stores in partnership with companies like Tesla? Additionally, how can this information be leveraged to increase revenue and customer convenience?

Parts Unlimited is considering providing EV charging stations as a new service. Therefore the Research and Development team can determine if there is a relationship between price and location. This information can provide insights into the cost and demand for EV charging stations in different locations and inform the decision on where to install new stations.

Business Question Q5
What are the top five manufacturers that provides widest range of products to Parts Unlimited,and current year how many ev charging stations located in those locations

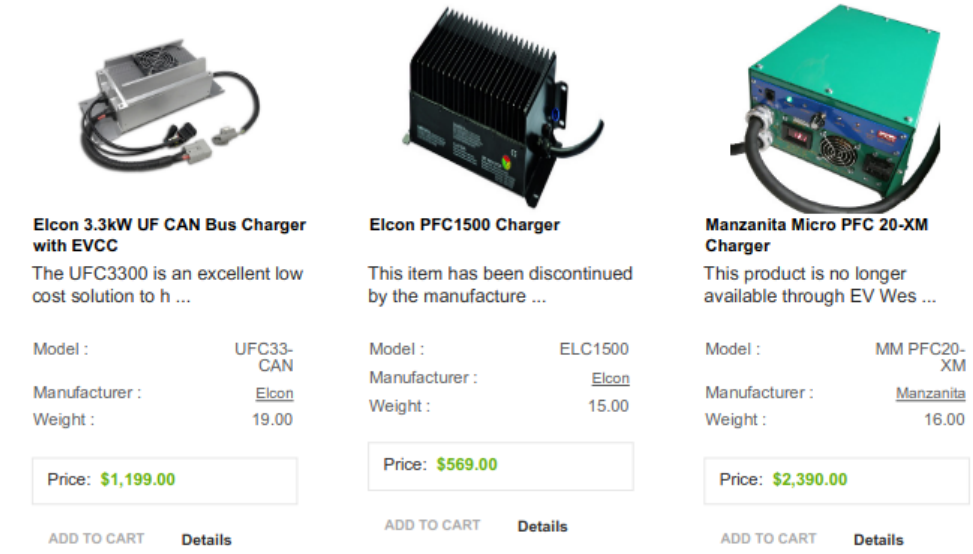
The Supply Chain department at Parts Unlimited can utilize this information to determine the percentage of products supplied by each manufacturer, as well as their respective locations. This can enable the department to better manage logistics and distribution of the products, while also ensuring compliance with any legal requirements related to supply chain operation

5 Part 3: Data Sources

Provide two data sources you will be using, for each data source list the number or columns and rows that are in each data source. Provide a header and first 5 rows from each source. - What is the URL or location of the data? - What information does this data provide that will help answer one or more of the above questions? - Do you see any issues in the data that will require transformation.

Dataset Name	Columns	Rows	URL
1-Product Info - EV West	5	50	EV West

The data set provides information about EV parts, manufacturer, weight and price information. The data is not available in the PDF version and requires a Python script to extract and format the information.



Dataset Name	Columns	Rows	URL
2-Alternative Fueling Station Locations	65	56,800	Data.gov

Provides information about the current list of charging stations and their locations. There are a lot of missing values that need to be cleaned and lat and long location info needs to be identical with the EV Car Population dataset. Price information is available in a text format and needs to be formatted so it is a number.

EV station Dataset Column Index	Names	EV Car Dataset Column Index	Names	EV station Dataset Column Index	Names	EV station Dataset Column Index	Names
1	Fuel Type Code	21	EV Other Info	41	Access Days Time (French)	61	CNG PSI
2	Station Name	22	EV Network	42	BD Blends (French)	62	CNG Vehicle Class
3	Street Address	23	EV Network Web	43	Groups With Access Code (French)	63	LNG Vehicle Class
4	Intersection Directions	24	Geocode Status	44	Hydrogen Is Retail	64	EV On-Site Renewable Source
5	City	25	Latitude	45	Access Code	65	Restricted Access
6	State	26	Longitude	46	Access Detail Code		
7	ZIP	27	Date Last Confirmed	47	Federal Agency Code		

Index	EV station Dataset Column Names	Index	EV Car Dataset Column Names	Index	EV station Dataset Column Names	Index	EV station Dataset Column Names
8	Plus4	28	ID	48	Facility Type		
9	Station Phone	29	Updated At	49	CNG Dispenser Num		
10	Status Code	30	Owner Type Code	50	CNG On-Site Renewable Source		
11	Expected Date	31	Federal Agency ID	51	CNG Total Compression Capacity		
12	Groups With Access Code	32	Federal Agency Name	52	CNG Storage Capacity		
13	Access Days Time	33	Open Date	53	LNG On-Site Renewable Source		
14	Cards Accepted	34	Hydrogen Status Link	54	E85 Other Ethanol Blends		
15	BD Blends	35	NG Vehicle Class	55	EV Pricing		
16	NG Fill Type Code	36	LPG Primary	56	EV Pricing (French)		
17	NG PSI	37	E85 Blender Pump	57	LPG Nozzle Types		
18	EV Level1 EVSE Num	38	EV Connector Types	58	Hydrogen Pressures		
19	EV Level2 EVSE Num	39	Country	59	Hydrogen Standards		
20	EV DC Fast Count	40	Intersection Directions (French)	60	CNG Fill Type Code		

[1]:
```
import warnings
warnings.simplefilter(action='ignore', category=Warning)
```

[2]:
```
import  pandas as pd
pd.set_option("display.max_columns",10 )
```

[8]:
```
import  pandas as pd
ev_station = pd.read_csv('data/alt_fuel_stations.csv')
ev_station.head(5)
```

[8]:
```
  Fuel Type Code                                    Station Name  \
0            CNG          Spire - Montgomery Operations Center
1            CNG                                PS Energy - Atlanta
```

```
2            CNG       Metropolitan Atlanta Rapid Transit Authority
3            CNG                                  United Parcel Service
4            CNG  Clean Energy - Texas Department of Transportation

         Street Address                                Intersection Directions  \
0      2951 Chestnut St                                                    NaN
1      340 Whitehall St  From I-7585 N, exit 91 to Central Ave, left on…
2   2424 Piedmont Rd NE                                                    NaN
3  270 Marvin Miller Dr                                                    NaN
4   7721A Washington St  I-10, Washington Ave exit, 1.5 blocks to the s…

          City  …    CNG PSI CNG Vehicle Class  LNG Vehicle Class  \
0  Montgomery  …       3600                MD                 NaN
1     Atlanta  …       3600                MD                 NaN
2     Atlanta  …       3000                LD                 NaN
3     Atlanta  …       3600                HD                 NaN
4     Houston  …  3000 3600                MD                 NaN

  EV On-Site Renewable Source Restricted Access
0                         NaN               NaN
1                         NaN               NaN
2                         NaN               NaN
3                         NaN               NaN
4                         NaN               NaN

[5 rows x 65 columns]
```

Dataset Name	Columns	Rows	URL
3- EV Population Data	17	121,978	Data.gov

This dataset shows EV cars that are currently registered through Washington State Department of Licensing (DOL). Provides information about EV cars, car types, and registered locations. Location fields have to match with the format of the EV Charging dataset. There are some missing fields. Currently, I do not see any other issues related to this dataset.

Index	Electric_Vehicle_Population
1	VIN (1-10)
2	County
3	City
4	State
5	Postal Code
6	Model Year
7	Make
8	Model
9	Electric Vehicle Type

Index	Electric_Vehicle_Population
10	Clean Alternative Fuel Vehicle (CAFV) Eligibility
11	Electric Range
12	Base MSRP
13	Legislative District
14	DOL Vehicle ID
15	Vehicle Location
16	Electric Utility
17	2020 Census Tract

[62]:
```
ev_pop = pd.read_csv('data/Electric_Vehicle_Population_Data.csv')
ev_pop.head(5)
```

[62]:
```
    VIN (1-10)    County      City State  Postal Code  …  \
0  5YJ3E1EB2J   Suffolk   Suffolk    VA      23435.0  …
1  5YJ3E1ECXL    Yakima    Yakima    WA      98908.0  …
2  WA1LAAGE7M    Yakima    Yakima    WA      98908.0  …
3  5YJ3E1EA1K  Danville  Danville    VA      24541.0  …
4  1FADP5CU9E   Norfolk   Norfolk    VA      23518.0  …

   Legislative District DOL Vehicle ID             Vehicle Location  \
0                   NaN      476647986    POINT (-76.42443 36.8752)
1                  14.0      103490145  POINT (-120.56916 46.58514)
2                  14.0      144941534  POINT (-120.56916 46.58514)
3                   NaN      168513922    POINT (-79.4172 36.58598)
4                   NaN      150749378   POINT (-76.21549 36.92478)

  Electric Utility 2020 Census Tract
0              NaN      5.180008e+10
1       PACIFICORP      5.307700e+10
2       PACIFICORP      5.307700e+10
3              NaN      5.159000e+10
4              NaN      5.171001e+10

[5 rows x 17 columns]
```

Dataset Name	Columns	Rows	URL
4- Manufacture Location	6	25	extracted by online search

Provides information about EV cars. This dataset is clean, currently I do not see any issues.

Index	Manufacture Location
1	company_name
2	latitude

Index	Manufacture Location
3	longitude
4	city
5	state
6	country

6 Part 4: Dimensions - Review the data and the business questions from part 2.

- **What fields (attributes) are in the data that will be used for the dimensions.**
- **Determine the dimension tables. There should be at least two non-date dimensions and one date dimension for each fact table.**
- **At least one (non-date) dimension in your design should have a hierarchy.**
- **What are the attributes that will be tracked via slowly changing dimensions?**
- **What attributes within the dimensions will need transformation before they are loaded into the dimension, for example it could be to build consistency or any other issues? This is where for example you might build case statements in your code to handle various scenarios. Two to three examples showing some sample data and what you think the transformation will be during your ETL would be helpful here.**

1-Table Name : ev-car-population Table Attributes: PK DOL Vehicle-Id SK Car-Pop-Id FK location-id FK EV-Charg-Stat-Rec-Date Make Model Model-Year Electric-Vehicle-Type Status-Flag Status-DeAct-TimeStamp SCD TypeInfo: SCD type 2. if a car no longer exists we can check the status and check deactivation date, we can track the record date. and sk help us track the history SCD Tracked Attributes : SK CAR-POP-ID Status-Flag Status-DeAct-TimeStamp Transform Needed Attributes: Loc-ID:Location Id Not exist this will be latitude and longitude concatenation EV-Charge-State-Rec-Date, The data creation date mentioned in the source but not included in the csv file, I will need to insert this data Status-Flag: I will need to insert this date based on other columns info Status-DeAct-Timestamp: I will need to insert this date based on other columns info

2-Table Name : EV Charging Station Table Attributes: PK Id SK Charging_Stat_ID fk location-ID Location-ID EV-Charg-Stat-Rec-Date Station Name Updated At Date Last Confirmed Updated At SCD TypeInfo: SCD type 2. if a charging station has been updated. Updated at field allow us to track we also created sk to track history. SCD Tracked Attributes : SK Charging_Stat_ID Updated-At Transform Needed Attributes: We need to create surrogate key

3-Table Name : Location Dimension Table We can implement a hierarchy in this table in between country,state,city Table Attributes: PK location-ID City State ZIP Latitude Longitude Latitude-prev Longitude-prev SCD TypeInfo: SCD type 3. We have prev attributes for lat and long SCD Tracked Attributes : SK Charging_Stat_ID Updated-At

4-Table Name : manufacturer PK Manufacture-ID SK sk-manufacture manufacturer-name manu-status-flag(scd) Vendor_since (scd) manu-updated-at(scd) SCD Tracked Attributes : vendor since (I am not sure about this one, I will do a research on this one to see if I will need to track in the time dimension instead) manu-updated-at(scd) manu-status-flag(scd) SCD TypeInfo: SCD

type 2. SCD Tracked Attributes : SK Charging_Stat_ID manu-updated-at-At Transform Needed Attributes: we are missing vendor since and updated at we need to add those data also transform data from pdf format

5- Table Name :Product Table Attributes: pk product-ID SK SK-product product-name product-description Product-category product-status-flag(scd) product-updated-at(scd) SCD Tracked Attributes : product-status-flag(scd) product-updated-at(scd) Transform Needed Attributes: we need to bring all the values from the pdf and make sure that they are in the correct format.

6-Table Name : manufacturer-product-time PK man-product-time-id man-product-release-date man-product-discontinuation_date

SCD Tracked Attributes : None SCD TypeInfo: SCD type 1. SCD Tracked Attributes : None Transform Needed Attributes: CWe need to populate this fields

7 Part 5: Facts – Review the data and the business questions from step 1.

What measurements are in the data that will be used for the fact tables? What measures will you be calculating (i.e. using an aggregate function, or some other transformation – recall as an example some of the aggregation you did in assignment 1A)

EV Charging Station Usage: Calculating number of cars and stations to measure, it can also calculate average price per location etc manufacture-facts: This table is keeping track of manufacturers and manufacturer products that the company sells. The measurements are weight,price

Table Name :1-ev-car-population
Table Attributes:
—
PK DOL Vehicle-Id
SK Car-Pop-Id
FK location-id
FK EV-Charg-Stat-Rec-Date
Make
Model
Model-Year
Electric-Vehicle-Type
Status-Flag
Status-DeAct-TimeStamp

SCD TypeInfo:
SCD type 2. if a car no longer exists we can check the status and check deactivation date, we can track the record date. and sk help us track the history

SCD Tracked Attributes :
SK CAR-POP-ID
Status-Flag
Status-DeAct-TimeStamp

Transform Needed Attributes:
Loc-ID:Location Id Not exist this will be latitude and longitude concatenation
EV-Charge-State-Rec-Date, The data creation date mentioned in the source but not included in the csv file, I will need to insert this data
Status-Flag: I will need to insert this date based on other columns info
Status-DeAct-Timestamp: I will need to insert this date based on other columns info

8 Part 7: Provide code and screenshots of loading your data into staging/data frame

9 Part 8: Provide code and screenshots of transforming the data. Perhaps you are adjusting for consistency of data or calculating aggregates.

10 Part 9: Provide code and screenshots of loading the data into the two dimensions and the fact. At this time, you do not need to worry about maintenance of slowly changing dimensions, the focus is on the initial data load. If you are loading into SCD2 or SCD3, make sure to show the SCD maintenance attributes populated.

For part 7, 8 9: This week I worked on extracting data from EV West pdf documents. Extract : First I have extract all the pdfs and saved as a csv files and later I combined them as a big csv Transform : I have transformed data so that it has all the attributes. Load : I have created three dimention table and one fact table for this dataset. I am only missing a location dimention to complete this fact table