**SWIFT Assignment - Choked Hubs Prioritization**

Submitted by Dwiz Garg

1. Data Visualisation
2. In this assignment I first parsed the json file using sql , the code is given in the sql file itself

Where I made columns like ctime , location, latest\_location, latest\_status ,shipment\_id. After this also exported this to a csv file for easy manoeuvring of the data while executing sql queries. There were approximately 107000 shipments.

1. By examining the file in excel I saw that there exist null values in the location column and also some blanks values in both latest\_location and location column. But I noticed a pattern there and also thought of the relevant reasons which are as follows-
2. NULL – they simply imply that the data is wrongly recorded or there was no appropriate location whose data could be recorded ,and In most of the cases it was the first reason as the ctime for this row and the consecutive row afterward is same and the other row has an appropriate value of location ,hence it is just an error and can be removed.

The number of null valued rows were 3011.

1. Blanks -they simply represent when the shipment is traveling from one place to other – i.e it is on-the-move. Hence it could be deleted appropriately as we will see in the further analysis. Their number in the dataset was 5888 rows out of (671420 total)

Also , there were tuples where status was set to In Transit and the latest\_location and the location fields were both empty so it was just useless for the analysis, hence I deleted those columns also.

2. Analysis

I conducted the Analysis part in two steps:

* There can be warehouses where the shipment is traveling through only one warehouses then gets delivered to buyer
* Other is where it is involving more than one warehouse.

-I have selected the date for the analysis to be 7th Oct 2023 10:AM , I chose this as the maximum ctime I have is on 7oct and also the 10:00 seemed to be right as an regular office time , this is for the analysis of status like ‘Out for Delivery’ and ‘In Transit’.

a) One Warehouse analysis- Let’s dive into this first ,

here what I did was , I created the table raw where I had all the data , then I deleted the data where the location is NULL or blank , then I made a new table where I had the tuples in which location value is distinct to be 1 (the number of tuples were 1210) ,for a given shipment\_id , I also included the columns start\_time and end\_time as min(ctime) and max(ctime) respectively , since only 1 shipment is involved.

Then I filtered the tuples where the latest\_location is blank , and saw where the start\_time and end\_time were same I delted them from the table , as they were just the tuples who had blank rows above and below them symbolizing that the parcel is in-transit , then I added location for all the tuples using primary key shipment\_id.

Now, I introduced the concept of delay (in hours) as follows:

-For Delivered :- it is start\_time to end\_time. As the parcel is already delivered so it just doesn’t make sense to include the date of the current day on which we are analysing the data.

-For In Transit :- It is defined to be end\_time till the current date , as since the parcel is present only at the warehouse but not moving or hasn’t just exited from there as we got info from latest\_location.

-For Picked Up- It is same as the in transit, as the parcel is picked up from a warehouse but still the latest\_location is showing at the warehouse , I personally found this case to be ambiguous or it just doesn’t tell us anything and Is also a minority , like some 20 tuples from 1200 tuples in this dataset.

-For Out for Delivery – I set it as end time to current date as to see how much time has passed since the order has left the warehouse but not delivered.

-For Shipment delayed :- there were not such tuples.

After this is applied the threshold of 48 hours for choked hub prioritization and maintained a column is\_delayed with Boolean values, now then I created a new table with final results where I got average\_delay and occ(occurences) for a particular warehouse , occ is calculated as number of times the warehouses had the is\_delayed column set to 1.

Hence in this way I analysed the warehouses entirely using sql and exported it as csv named ‘onewarehouse.csv’ , for code please check the sql file and I have also included the sql\_output\_history file for authenticity of the commands.

I chose to do One Warehouse analysis first so as to prioritise the single point of failure shipments and also to visualise the data better for understanding of the more than one warehouse analysis.

b)More than one warehouse analysis-

now I have removed those one warehouse shipments from here and , also break them into five tables containing the data with respect to latest\_status as ‘delivered’, ‘intrans’, ‘picked’, ‘delay’, ‘outfordelivery’ so as to do analysis of them in a better way,

Now here I defined a parametric named perecentage\_delayed which is defined as -   
Percentage Delayed=(Number of Delayed Shipments/Total Shipments)×100

Again the delayed shipments are calculated as if a given shipment stays more than 48 hours at a given warehouse,

And using this I assigned the tags for all the different statuses ,

But as we learned in the onewarehouse analysis , the in transit and out for delivery will require special attention for which –

I used the last row of the location column for each shipment id to calculate the delay , i.e for how much time the shipment is at that warehouse and proceeded like I did in the above case , and exported the results based on that computing average delay and the number of occurences , for picked up , I personally couldn’t comprehend about the choked warehouses in this matter , so I simply compiled a percentage\_delayed analysis for this , now for the shipment\_delayed part , since the shipments are delayed there , the hubs there need attention and to further quantify it I used percentage delayed to prioritize them first.

CONCLUSION-

In summary, I meticulously processed shipment data using SQL, ensuring data integrity by addressing null and blank values. The analysis focused on single and multiple warehouse scenarios, introducing a 48-hour delay threshold for prioritization. The One Warehouse analysis produced 'onewarehouse.csv,' detailing average delays and occurrences per warehouse. The More Than One Warehouse analysis further categorized shipments based on statuses, introducing a percentage\_delayed parameter. Special attention was given to In Transit and Out for Delivery, utilizing the last row of the location column. The insights gained offer a practical approach to identifying and prioritizing delayed hubs, providing valuable guidance for optimizing logistics and enhancing overall shipment efficiency.

THANK YOU