Daily task

Thursday, January 11, 2024

10:10 PM

################## Finance PnL #######################

**01-15-2024:**

* To check what's the issue, visiting with Sagar's id. If that pipeline ever run in past (which was developed by us, then do ask bhaskar, what's this new issue). If this is the same pipeline that has not run, ad was developed by us, look into the issue, try debugging (only read-only part - Understanding), and then asking support team to fix it. If you don't understand then put in tredence group that you need help.

* Send in the group (clorox teams) created by akshay, tagging mohan, laurie and arun that we need access to logic app.

* Send in the group (clorox teams) created by akshay, tagging mohan, laurie and arun that we need access to office apps to validate and verify the output generated in VM (AVD).

* Story writing with Ajay, review by krishna.
* Follow up on logic app again.

**01-18-2024:**

* Loop ramesh and LI for access on GIT - RECEIVED ACCESS - DONE
* Ask Doug about my access to snowflake.- RECEIVED ACCESS - DONE

* Ask suprit/Diane for data validation in excel -> Diane is validating the file -> She has given update and want some changes like :

**She told she is expecting values in 2 column which are coming NULL in our export (and also in sql table). Also, she wants date to be formatted in MM-DD-YYYY. Working on it.**

**->** Created group with Ramesh and Suprit, following up on Logic App creation and access

* Update story on data load, in assumption or description. We will use WM01 only, but : [We can use one copy activity with source as EDP table and target as Shiloh].

* IRI pipeline and config check with Adinath team [IRI Product  from Azure Synapse]. Ask him about how to get access to key-vault resource. Also ask for all access related links. -> to follow up with Bhaskar on this.

Connected with Bhaskar (for IRI Product data) and got below updates.

- Changes needs to be moved to prod (was left as network crisis happened)

- SSE tables (DEV env) has data, but Vishwanath is working on testing the connections (which might have not been working fine). Once testing is established, we can test at least the sample export/push to Shiloh databases.

**To again check with him and Vishwanath on this on Monday.**

* Follow-up and gather info about Stat case data in EDP from Debashish
  + Connected with him, and got to know a few table names.
  + Finally Krishna confirmed about a table called **"refined.w\_material\_dim"**. Update this on story.

* Try connecting to databricks (EDP workspace) from WM01 workspace, using credentials provided by Debashish -> Getting error, connecting. Tried creating new cluster, so new cluster-id, still didn't work. -->> Tried the approach of using **AZURE-IR,** and it WORKED.

* Try creating your own access token and then try above point ->
  + Give Akshay a Key vault name, he will help creating a secret for that Access token
  + Just used the **Integration Runtime as "AZURE IR"** and **SECRET** created by Akshay, and TEST CONNECTION worked successfully.

* Go through the whole mail chain, and Ask over mail for Schema detail for pushing data to snowflake. -> Send mail on Monday.

* Ask Akshay/Nithya to create secret names in key-vault for Snowflake password. And test connection on Snowflake database from ADF was SUCCESSFUL - DONE
* Couldn't load snowflake table data to ADLS, getting error in sink side in copy activity. Akshay sugested to get you adls linked service a firewall access on the port which comes in error which testing connection for ADLS (using ON-prem IR).

**02-08-2024:**

* Look into message from stephen, think of which stories to push further in iterations.
* Create a model for data architecture. - RAMESH SAID HE WILL ASSIGN THIS TO LAURIE, AND himself HELP HER.
* Partitioning the data in fact tables needed. - DONE (added in DDL)
* Create landing folder and partition that also - DONE (WE WILL USE [ partitioBy ("year") ] while writing to ADLS
* Run history table and then load incremental for inventory. - in progress
* Comment on story (performance check). - HALF DONE; YET TO COMPLETE THE INVENTORY\_HIST table test

**02-09-2024:**

* Test streaming read-write, with custom schema, taking source as ADLS.
* Ask from Naveen , if above point doesn't work well, for the best possible alternative (as we have to DELSERT/MERGE)
* GOPAL said to check why :
  + WM\_ITEM dim column is not having decimal in it (one column he mentioned in chat).
  + Why in WM01 workspace we don’t have proper data for stat-case table w\_material\_dim\_test.
* Performance check story to be finished
* Create 1 new story (as per RAMESH) for duplicating data then trying DELSERT for performance.
* COMMENT on "DISCOVER CONNECTIVITY" story.

**02-17-2024:**

* Test deeply, understand and implement Audit log notebook, in pipeline for landing layer at least. Test for all layers, and remove any un-utilized code from the notebook to make it look simpler. Take a copy of that notebook with experimented results in your databricks community.
* Create that lookup refresh-date-COMPARE logic in master pipeline.
* Complete the minor changes, implement audit log and apply a schedule trigger to the EDP-to-Shiloh pipeline once it gets integrated into MASTER PIPELINE.
* Study 10 videos. Prepare notebooks also.
* Ask question about your code and scenario on stack overflow and databricks community edition on - streaming doesn't solve my problem and why cache is helping?

**04-01-2024:**

* Run for-each activity for parallel runs, normal run as it is, but we have to check if it fails in landing/curated whether it fails the whole for-each. So steps will be :-
  + change the config to fail in curated layer.
  + Run and test if it fails there.
  + Since success/failure mail notebook is changed and now the one CAN HANDLE success, failure both by itself, lets test it for curated layer.

**04-16-2024:**

* Test the latest v4 log notebook functionality for all possible layers - Landing, curated,raw,delete,push-data loads.
* Raw layer logging testing for multiple tables at once - whether it is able to update RAW and Curated at same time or not.
* Delete query creation in python (using dbutils.rm), testing it. Creating a notebook which can take parameters for deleting a landing location full of parquet files. Get to know the parquet files count within it, the count of entries they all contain combinedly and be able to log those all info in the log table.
* Create CUSTOM WATERMARK ENTRIES notebook, using merge logic on - retailer, dataset and layer (raw). - CUSTOM NOTEBOOK
* Create update-shiloh-entries-to-Y (from FAILED) notebook which could just take a dynamic parameter of "retailer" and update latest failed entries in shiloh to Y.
* Implement UCT-NOW() to CST() time zone logic in all places at pipeline including :
  + All log-notebook places in pipeline widgets, including one special one inside log-notebook itself
  + All watermark notebook call places
  + All pipeline level activities like - "job start date time", etc
  + One new and special case is of raw layer, where we are using the "current-day" activity.
* Think about Raw Layer execution logic. Like how to check if the entry already exists or not.
* Add three missing parameters in all the logging notebooks of DATA-PUSH pipelines.
* Creating global parameters for all notebook(s)' paths and replacing them in notebook activities.
* Dropping target tables (data and metadata), putting correct names in config as suggested by Li, then updating the config and running Refresh-config pipeline.

Fixes to b done in pipelines:

* Target file, and others count is wrong, fix it
* Add a condition in Raw layer lookup, in case it doesn't find any entry in the watermark, the if condition should go to fail
* Change/create the guid thing in master pipeline using the Notebook, and pass from the notebook output to set-variable.

**Running end-to-end once to check all loads….waiting for it!!!**

**2024-04-20:**

* Clone a feature branch, and create correct cst parameters pipeline.
* Modify the raw layer part in it.

Prod migration docs:

* Key vault (tenant id and all)
* Prod migration steps docx - linked service wala page screenshot left
* Ler config - uploaded
* Devops intake file - to be done
* Config\_prod.csv - to be completed
* Rules\_prod.text - completed and uploaded

**2024-05-13:**

To discuss with Akshay:

* Target refreshed today + Docs updated -> Shall story be updated and moved? Any other docs like they are talking about (adding a new user to a security group) be added in support doc?
* Can we delete the two useless parameters in master (retailer\_handhsake) pipeline, since GP being used now?
* What all enhancement works can be done, like if you have thought any:
  + Automate the update the Shiloh entries to Y just by passing a parameter like module or something. Need more discussion on this to implement.
  + Important scenario: Currently if we have to do re-run from failed activity, we first turn all FAILED flags to Y in handshake then, re-run from failed activity. But it still runs all the active entries and not just one dataset. So needs to be fixed.
* When shall we schedule the call with Adinath and team for the left out notebooks KT?
* Election holiday on 20th.

############################%% **Issues in pipeline these days:** %%%%###########################

1. Email notification success/failure sent all objects list in the mail. Although, on that day, Shiloh didn’t provide 2 tables (ending in sales\_D & inventory\_D) in their handshake table. So ideally, our success/failure mail should also tell that this is not present in shiloh or something like that, just the way our summary email calls this out.
2. Target pipeline got failed in watermark update of Raw layer. This occurred while a manual run was going on and raw layer job was forcefully made to run. So since, curated and raw are se to run in parallel, they both entered the watermark update almost at the same time, updating the same partition of the table at the same time, hence failed.

Strange part is, this issue never came up till date.

#############%% **Enhancements in pipeline to-do-list, after target migration:** %%%%#####################

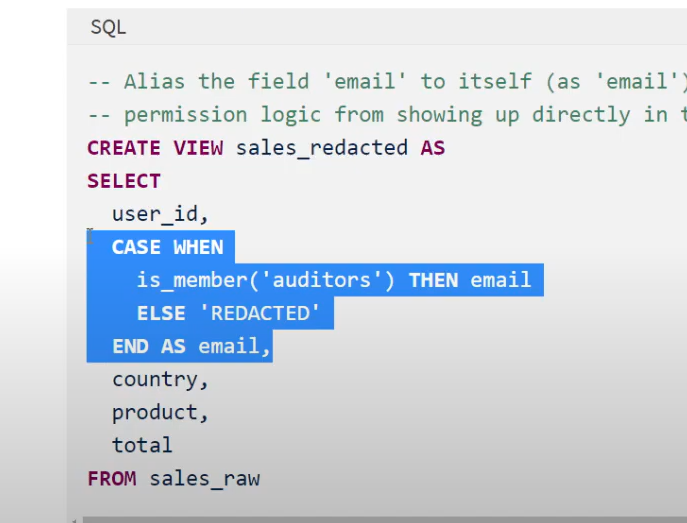
1. Cosmetic changes like restructuring the folders in ADF, naming pipelines according to order of execution, etc.
2. Replacing the GUID notebook activity with a LOOKUP activity.
3. Bringing in the web-activity in place, right after the mail notifications notebook to invoke the logic app using adf.
4. If possible, deep-diving all the pipelines to understand if we could reduce the number of activities or nestings (done using execute activity).
5. Delete the useless parameters, like snowflake schema and table parameter in retailer-level-handshake pipeline

**2024-05-31:**

* Shall we work on cosmetic changes, and a few others mentioned in excel?
* How to bring prod data to dev
* Discuss on story testing.
* **Column addition discussion yesterday's call. :- Create test tables with more columns(and they should have data), and try updating data in current columns, to check if current code can handle it.**
* Two runs started in uat.

**2024-06-03:**

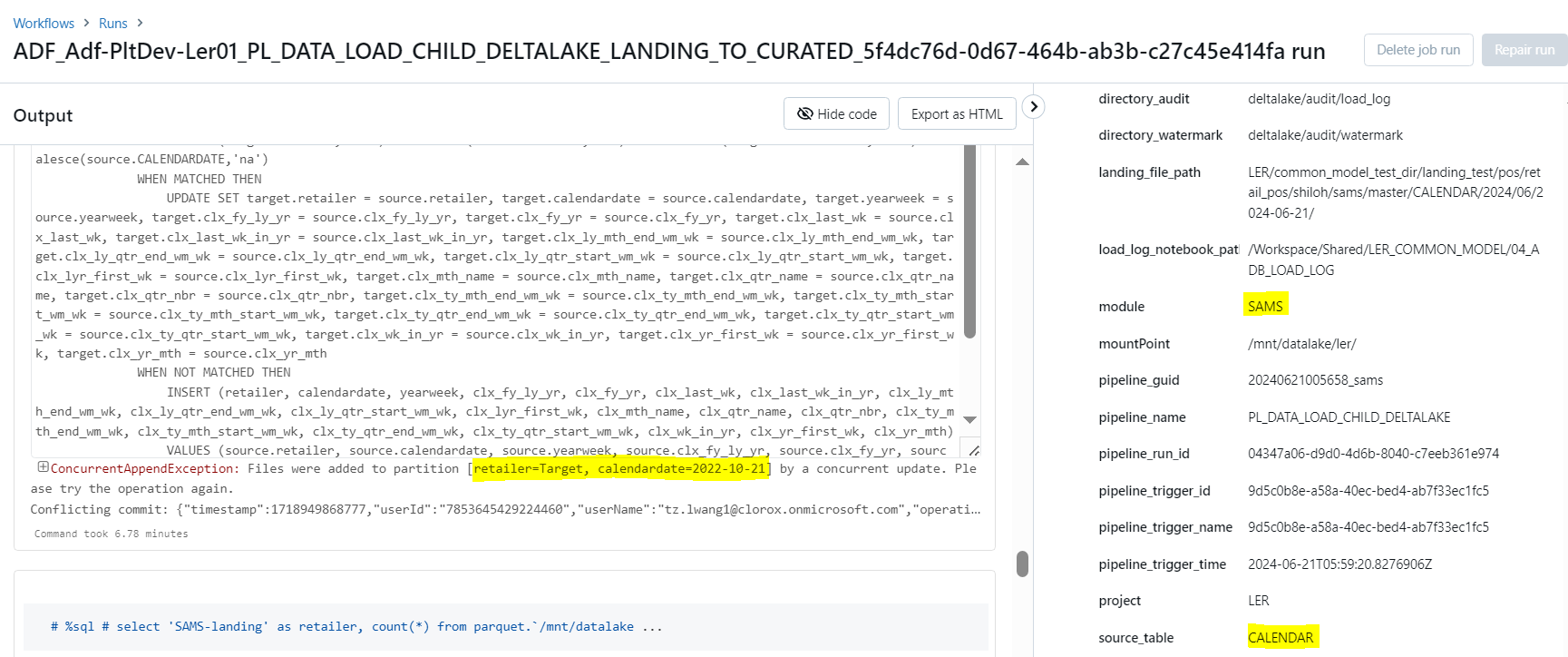
Column level security:



SAMS error :

ConcurrentAppendException: Files were added to partition [retailer=Target, calendardate=2022-10-21] by a concurrent update. Please try the operation again. Conflicting commit: {"timestamp":1718949868777,"userId":"7853645429224460","userName":"tz.lwang1@clorox.onmicrosoft.com","operation":"MERGE","operationParameters":{"predicate":["((coalesce(RETAILER#36875, na) = coalesce(RETAILER#33438, na)) AND (coalesce(cast(CALENDARDATE#36876 as string), na) = coalesce(cast(CALENDARDATE#33510 as string), na)))"],"matchedPredicates":[{"actionType":"update"}],"statsOnLoad":false,"notMatchedBySourcePredicates":[],"notMatchedPredicates":[{"actionType":"insert"}]},"job":{"jobId":"1096702724463269","jobName":"ADF\_Adf-PltDev-Ler01\_PL\_DATA\_LOAD\_CHILD\_DELTALAKE\_LANDING\_TO\_CURATED\_11009104-00e1-4e9d-b53f-4510cf0479a8","runId":"1095657496766746","jobOwnerId":"7853645429224460","triggerType":"manual"},"notebook":{"notebookId":"1190850148517242"},"clusterId":"0831-160602-drily846","readVersion":9,"isolationLevel":"WriteSerializable","isBlindAppend":false,"operationMetrics":{"numTargetRowsCopied":"0","numTargetRowsDeleted":"0","numTargetFilesAdded":"2184","numTargetBytesAdded":"10801931","numTargetBytesRemoved":"10801931","numTargetDeletionVectorsAdded":"0","numTargetRowsMatchedUpdated":"2184","executionTimeMs":"263067","numTargetRowsInserted":"0","numTargetRowsMatchedDeleted":"0","scanTimeMs":"145659","numTargetRowsUpdated":"2184","numOutputRows":"2184","numTargetDeletionVectorsRemoved":"0","numTargetRowsNotMatchedBySourceUpdated":"0","numTargetChangeFilesAdded":"2184","numSourceRows":"2184","numTargetFilesRemoved":"2184","numTargetRowsNotMatchedBySourceDeleted":"0","rewriteTimeMs":"115067"},"tags":{"noRowsCopied":"true","delta.rowTracking.preserved":"false","restoresDeletedRows":"false"},"engineInfo":"Databricks-Runtime/13.3.x-scala2.12","txnId":"72eeb734-c228-4eed-ac8f-42f49805f8b9"} Refer to <https://docs.microsoft.com/azure/databricks/delta/concurrency-control> for more details.

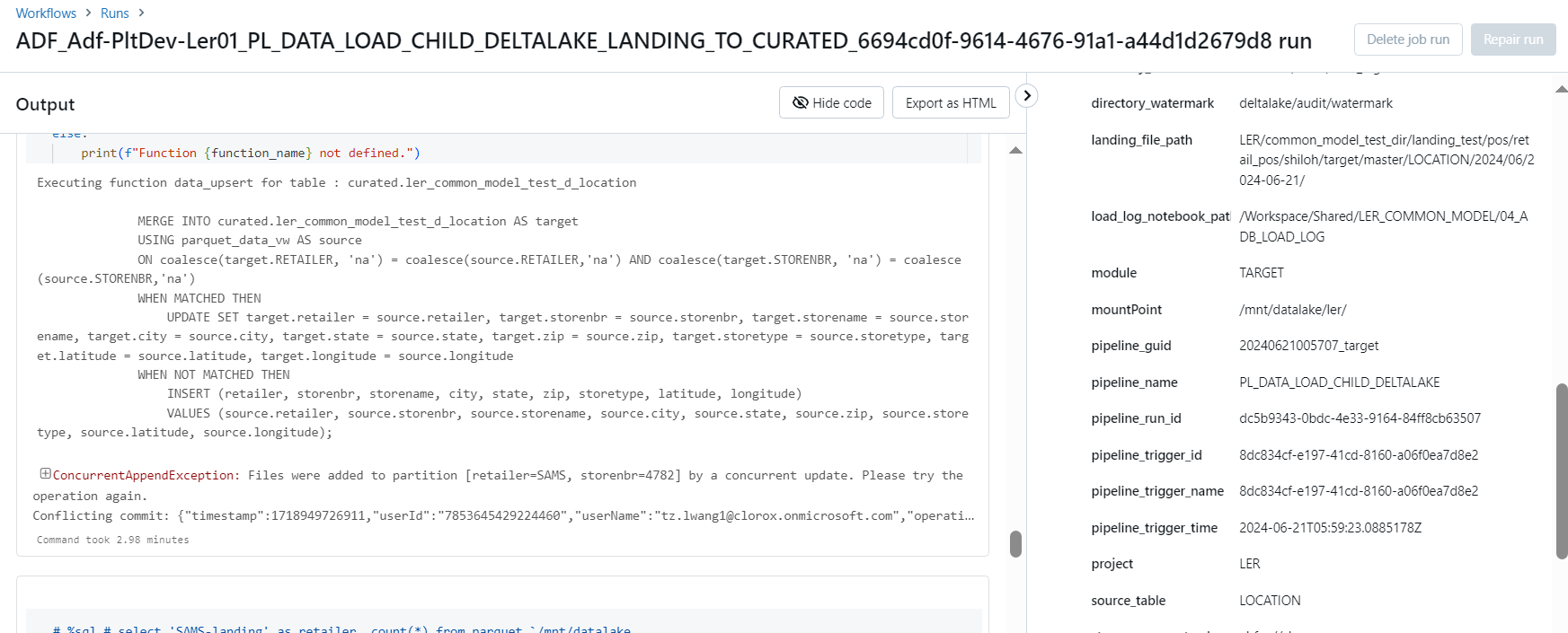
*From <*[*https://adb-1414755063160892.12.azuredatabricks.net/jobs/1067646516598199/runs/997676502147883?o=1414755063160892*](https://adb-1414755063160892.12.azuredatabricks.net/jobs/1067646516598199/runs/997676502147883?o=1414755063160892)*>*



Target error :

ConcurrentAppendException: Files were added to partition [retailer=SAMS, storenbr=4782] by a concurrent update. Please try the operation again. Conflicting commit: {"timestamp":1718949726911,"userId":"7853645429224460","userName":"tz.lwang1@clorox.onmicrosoft.com","operation":"MERGE","operationParameters":{"predicate":["((coalesce(RETAILER#53695, na) = coalesce(RETAILER#53551, na)) AND (coalesce(STORENBR#53696, na) = coalesce(STORENBR#53561, na)))"],"matchedPredicates":[{"actionType":"update"}],"statsOnLoad":false,"notMatchedBySourcePredicates":[],"notMatchedPredicates":[{"actionType":"insert"}]},"job":{"jobId":"1068247766681626","jobName":"ADF\_Adf-PltDev-Ler01\_PL\_DATA\_LOAD\_CHILD\_DELTALAKE\_LANDING\_TO\_CURATED\_77508bbd-1ff7-45b5-8d3e-c2cf26d6b940","runId":"166829694696330","jobOwnerId":"7853645429224460","triggerType":"manual"},"notebook":{"notebookId":"1190850148517242"},"clusterId":"0831-160602-drily846","readVersion":10,"isolationLevel":"WriteSerializable","isBlindAppend":false,"operationMetrics":{"numTargetRowsCopied":"0","numTargetRowsDeleted":"0","numTargetFilesAdded":"786","numTargetBytesAdded":"1701885","numTargetBytesRemoved":"1701885","numTargetDeletionVectorsAdded":"0","numTargetRowsMatchedUpdated":"786","executionTimeMs":"70781","numTargetRowsInserted":"0","numTargetRowsMatchedDeleted":"0","scanTimeMs":"6908","numTargetRowsUpdated":"786","numOutputRows":"786","numTargetDeletionVectorsRemoved":"0","numTargetRowsNotMatchedBySourceUpdated":"0","numTargetChangeFilesAdded":"786","numSourceRows":"786","numTargetFilesRemoved":"786","numTargetRowsNotMatchedBySourceDeleted":"0","rewriteTimeMs":"63478"},"tags":{"noRowsCopied":"true","delta.rowTracking.preserved":"false","restoresDeletedRows":"false"},"engineInfo":"Databricks-Runtime/13.3.x-scala2.12","txnId":"45eb23eb-08e4-40ad-b6bc-c729021ba397"} Refer to <https://docs.microsoft.com/azure/databricks/delta/concurrency-control> for more details.

*From <*[*https://adb-1414755063160892.12.azuredatabricks.net/jobs/690900856658914/runs/680778232486710?o=1414755063160892*](https://adb-1414755063160892.12.azuredatabricks.net/jobs/690900856658914/runs/680778232486710?o=1414755063160892)*>*



What I have to ask Li:

1. Ask for new high size databricks cluster to perform a merge command on the big data of walmart fact table.
2. Also performing salting on the dataset to see how it behaves after salting is done.
3. Merge into github using databricks repos.
4. Understand and talk to her on the security part.
5. Adf for performing/learning a few data mapping flow pipelines.