Pipeline Cold Start

This describes how to build an EDSM-style pipeline from scratch and run it for the first time.

Procedure

Setup the Workspace

1. Clone the edsm repo to your laptop.
2. Create an empty private work repo on github.com (your work account).
3. Push the laptop repo to the private work repo (instructions on empty repo webpage).
4. Create a new github folder in Databricks.
   1. Point it at your private work repo.
   2. Allow Databricks to install its app on your work github account. This will grant it access to your repos (or you can choose which ones).
5. On the laptop, use sed, jq and sponge commands to update the JSON files.
6. Commit these changes and push them to the github private work repo.
7. Update the functions/config.py file to reflect your S3 roots for landing and utility.
   1. These are at the bottom of the file.
8. Commit these changes and push them to the github private work repo.
9. Edit the downloader.sh script and update the paths at the top of the script.
10. Commit these changes and push them to the github private work repo.
11. Back in databricks, click the Git icon to open the Git interface.
12. Pull all the changes into the Databricks working copy of the repo.

Setup the Workflow/Job

1. Create a new job.
2. Create a new task called downloader and point to the utilities/downloader notebook.
3. Create a new task called job\_settings and point to the job\_settings notebook.
4. Run the job with special parameters and **only** select the job\_settings task.
5. The job settings will create the schemas, tables and volumes.
6. Now run the job with special parameters and **only** select the downloader task.
7. Go back to the job and create a For Each task and call it bronze\_loop.
   1. For the parameter put {{default-stuff-here.**bronze**}}.
   2. Concurrency = 10.
   3. Task to loop over = 03\_ingest notebook.
      1. Disable task optimization to disable retries.
      2. Set parameter: job\_settings = {{input}}.
      3. Set parameter: color = bronze.
8. Create a For Each task and call it silver\_loop.
   1. For the parameter put {{default-stuff-here.**silver**}}.
   2. Concurrency = 10.
   3. Task to loop over = 03\_ingest notebook.
      1. Disable task optimization to disable retries.
      2. Set parameter: job\_settings = {{input}}.
      3. Set parameter: color = silver.
9. By this time the download should be done. Go back and check on it.
10. If everything downloaded okay, run the whole job.