Project 2: GH Archive ETL Pipeline Retrospective

Group Members: Christopher Lee, Stanley Liu, and Javier Martinez

- **Is there anything your team did (technically, operationally, or otherwise), that you all are especially proud of?**
    - Technically, the partitioning was done well, and we are very proud of that. As a group, we also worked really well together, working on the different parts of our pipelines and communicating often with one another, forming a collective understanding of our code.

- **Is there anything your team did that slowed down your progress or didn't add value to the end result?**
    - Deliberating over which columns we should keep slowed down our process. We had different ideas on what should and shouldn't be kept. This caused multiple ERD drafts and delayed our actual making of the pipeline. Once we did have our ERDs done, we moved through the project at a good pace.

- **Was your initial EDA thorough enough, or did we discover surprises later that we should have caught earlier?**
    - We had a lot of useless events and columns we later dropped - for example, WatchEvent. We also included many useless fields. Many of these useless fields and events were included in our initial ERD. Finally, we incorrectly interpreted the data and created an incorrect ERD based off of that.
    - We also discovered that an actor.id could have multiple actor.logins associated with it.

- **Did your Solution Design Document you crafted during planning match the reality of your solution?**
    - It mostly did, there were a few changes, however. One was that our ERD had to change, after a meeting with an instructor. Another was the need for a more complex schema. Finally, our Gold layer changed much less than we initially anticipated from Silver.

- **What "unique" transformation(s) did your team do in your Bronze -> Silver transformations?**
    - Dealing with multiple logins for an actor id was a unique transformation we did in our cleaning. Formulating a plan so that we could carry out this cleaning was an interesting side quest.

- **Are you happy with the way your partitions turned out in each layer?**
    - Yes, our team is very happy with how the partitions turned out in both our silver and gold layers within ADLS.
    - Within our silver layer, our partitions were parquet files in optimal sizes. Most notably, fact_pull_request and fact_push_event were partitioned as 125-126 MBs parquets, and fact_events was partitioned as 185 MBs parquets. Our other tables were either too small to benefit from partitioning or did not meet the threshold before being partitioned.
    - Within our gold layer, our partitions were parquet files in smaller-than-optimal sizes, notably gold_event_count_count_week as 16-18 MBs parquets and gold_top_repos as 12-22 MBs parquets. We felt that this was fine, as our gold data was only 555 MBs and would not have a big increase in performance from having optimal partitioning.

- **If you were given 2 more additional weeks on this, what would be your next steps (if any)?**
    - Add more necessary columns to various events, and complete the challenge, analyzing languages.
    - Much more aggregations and visualizations, including:
        - PR merge rate by repo/org
        - Most active days of the week, months of the year, or hours of the day
    - Automate the pipeline and add streaming capability on Databricks

- **Alternatively, If you were to redo this project, is there anything you would have changed?**
    - More fine-tuning of our silver layer, as well as enriching the data
    - Processing data in batches or by month instead of all at once

- **What were the biggest blockers you found yourself facing?**
    - Biggest by far - Schema definition for nested JSON, constantly throwing errors because StrucType() was empty
    - Similar to the above - type mismatches in Silver_to_Gold

    - Having to redesign ERD after it was initially designed incorrectly (initial silver ERD)
    - Finding out which partitioning method to use, and which was fastest and easiest to create given serverless architecture

- **What's one thing each team member learned that they didn't know two weeks ago? (provide an answer for each teammate)**
    - **Javier**: Spark's inner workings was an alien concept to me. At first watching the lecture videos I couldn't quite grasp it, but as we progressed through the project and studying for the certification. I grew this deep understanding of Spark. The project itself helped me a lot with connecting the dots as I was able to see how everything came together.
    - **Stanley**: My understanding of how the silver-to-gold pipeline would work was a bit shallow since we skipped it for project 1. However, I consolidated what I learned from lectures by actually building and applying it. The project as a whole was a great overview of what an ETL pipeline could look like in a professional setting and definitely deepened my understanding for future projects.
    - **Chris**: Probably in depth how to actually partition files. I knew how the syntax obviously looks and the concepts behind it - for example, that partition size impacts performance, but the two never really connected until the project. The project was honestly far more useful in actually connecting concepts between in my head and in practice vs actually teaching me anything new.

- **What advice would we give to a team starting this project fresh?**
    - Use a tiny sample
    - Apply schema early and often
    - Perform thorough EDA
    - Your ERDs are your best friend! If you are satisfied with them, then the rest of your project will go smoothly