Individual Report

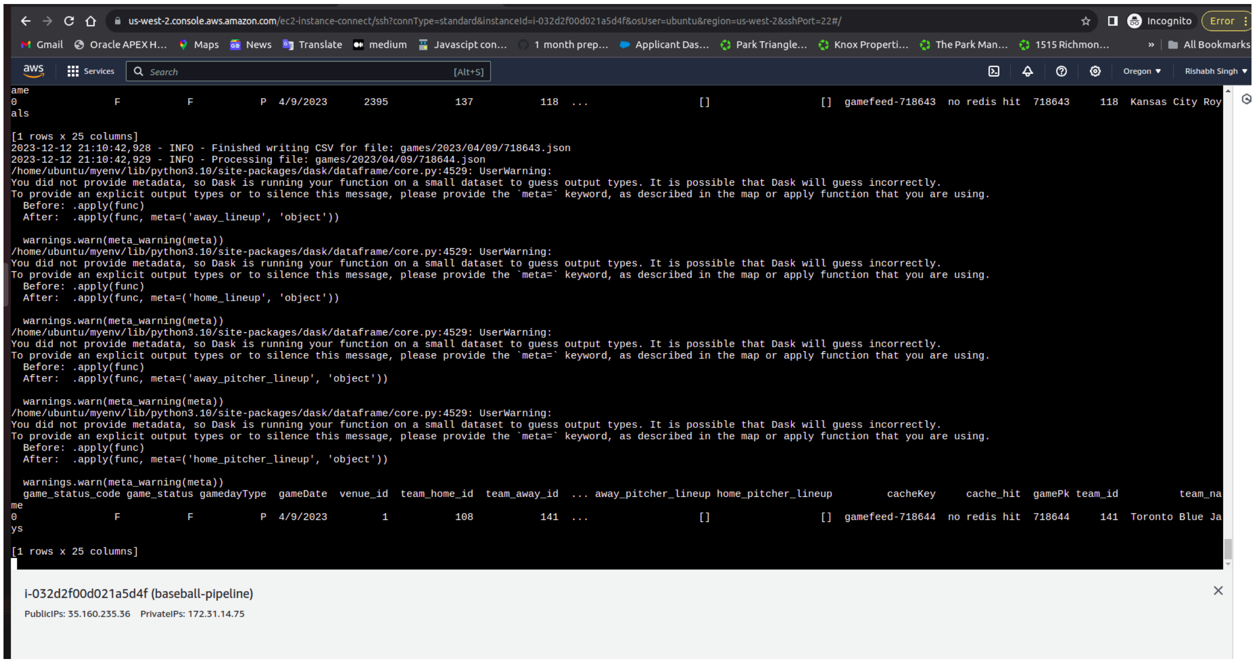
AWS Infrastructure Setup:
In the AWS infrastructure setup, I played a role that complemented our team's efforts. My focus was on configuring EC2 instances and S3 buckets, ensuring they met our project's requirements. This task was done in tandem with the team's strategy on IAM roles and user management. Together, we navigated through security challenges, analyzing access needs and establishing a secure yet user-friendly setup. My involvement was a blend of contributing individual insights and working within the team's framework.

Data Processing with PySpark and Dask:
For data processing, I was involved in integrating PySpark and Dask with our existing processes. My work centered around refining data transformation scripts, making sure they were aligned with the project's scale and complexity. This involved a deep understanding of both the tools and the data, ensuring our methods were efficient and in line with the team's strategy.

The decision to use Dask was a collective one, and I contributed by helping adapt this technology into our workflow. Implementing Dask's parallel processing capabilities helped improve our data handling efficiency, a task that required both individual technical skills and a collaborative approach.

There was analysis done through google colab by reading in the cleaned data for player engagement for fans. Used the python visualization tools for easy understanding of the results and analysis derived

Here is one of the screenshots of the logs:



```
ame
0               F           F          P  4/9/2023      2395          137          118  ...                  []                 []  gamefeed-718643  no redis hit  718643     118  Kansas City Roy
als

[1 rows x 25 columns]
2023-12-12 21:10:42,928 - INFO - Finished writing CSV for file: games/2023/04/09/718643.json
2023-12-12 21:10:42,929 - INFO - Processing file: games/2023/04/09/718644.json
/home/ubuntu/myenv/lib/python3.10/site-packages/dask/dataframe/core.py:4529: UserWarning:
You did not provide metadata, so Dask is running your function on a small dataset to guess output types. It is possible that Dask will guess incorrectly.
To provide an explicit output types or to silence this message, please provide the `meta=` keyword, as described in the map or apply function that you are using.
  Before: .apply(func)
  After:  .apply(func, meta=('away_lineup', 'object'))

  warnings.warn(meta_warning(meta))
/home/ubuntu/myenv/lib/python3.10/site-packages/dask/dataframe/core.py:4529: UserWarning:
You did not provide metadata, so Dask is running your function on a small dataset to guess output types. It is possible that Dask will guess incorrectly.
To provide an explicit output types or to silence this message, please provide the `meta=` keyword, as described in the map or apply function that you are using.
  Before: .apply(func)
  After:  .apply(func, meta=('home_lineup', 'object'))

  warnings.warn(meta_warning(meta))
/home/ubuntu/myenv/lib/python3.10/site-packages/dask/dataframe/core.py:4529: UserWarning:
You did not provide metadata, so Dask is running your function on a small dataset to guess output types. It is possible that Dask will guess incorrectly.
To provide an explicit output types or to silence this message, please provide the `meta=` keyword, as described in the map or apply function that you are using.
  Before: .apply(func)
  After:  .apply(func, meta=('away_pitcher_lineup', 'object'))

  warnings.warn(meta_warning(meta))
/home/ubuntu/myenv/lib/python3.10/site-packages/dask/dataframe/core.py:4529: UserWarning:
You did not provide metadata, so Dask is running your function on a small dataset to guess output types. It is possible that Dask will guess incorrectly.
To provide an explicit output types or to silence this message, please provide the `meta=` keyword, as described in the map or apply function that you are using.
  Before: .apply(func)
  After:  .apply(func, meta=('home_pitcher_lineup', 'object'))

  warnings.warn(meta_warning(meta))
  game_status_code game_status gamedayType  gameDate  venue_id  team_home_id  team_away_id  ... away_pitcher_lineup home_pitcher_lineup         cacheKey     cache_hit  gamePk team_id          team_na
me
0               F           F          P  4/9/2023         1           108           141  ...                  []                  []  gamefeed-718644  no redis hit  718644     141  Toronto Blue Ja
ys

[1 rows x 25 columns]
```

i-032d2f00d021a5d4f (baseball-pipeline)

PublicIPs: 35.160.235.36   PrivateIPs: 172.31.14.75