horizontal line

GitHub Insights Analytics

**─**

## Problem Statement

**─**

For a comprehensive understanding of GitHub's ecosystem, analyze user profiles, activity patterns, repository characteristics, and programming language preferences. This project aims to uncover key insights to help us better grasp the GitHub community's diverse nature.

## Data Dictionary

**─**

1. hirable: Indicates if the user is looking for a job.
2. public\_repos: Number of public repositories.
3. is\_suspicious: Flags whether an account is suspicious.
4. updated\_at: Timestamp of the latest user information update.
5. id: Unique user ID allocated by GitHub.
6. blog: URL of the user's public blog.
7. followers: Number of followers.
8. location: Public location of the user.
9. type: Identifies if the account is personal, organizational, or a bot.
10. commit\_list: Contains commit operation details.

* commit\_at: Timestamp of code submission.
* generate\_at: Timestamp of author committing code changes.
* committer\_id: ID of the committer.
* author\_id: ID of the author.
* repo\_name: Name of the committed repository.
* repo\_id: ID of the committed repository.
* repo\_description: Description of the committed repository.
* repo\_owner\_id: ID of the owner of the committed repository.

1. bio: Self-description provided by the user.
2. commits: Number of commits made by the user.
3. company: Working unit of the user.
4. public\_gists: Number of public gists.
5. name: Nickname/exhibited name of the user.
6. created\_at: Timestamp of account creation.
7. email: Public email address of the user.
8. following: Number of users following.
9. login: Username/registered name of the user.
10. repo\_list: Contains details of repositories created or forked by the user.

* full\_name: Full name of the repository.
* id: Repository ID allocated by GitHub.
* description: Description of the repository.
* size: Size of the repository.
* license: License associated with the repository.
* stargazers\_count: Number of stars received by the repository.
* fork: Indicates whether the repository is a fork.
* owner\_id: ID of the owner of the repository.
* created\_at: Timestamp of repository creation.
* pushed\_at: Timestamp of the last push operation to the repository.
* updated\_at: Timestamp of the latest change to the repository.
* has\_wiki: Indicates if the repository has a wiki document.
* open\_issues: Number of open issues in the repository.
* language: Programming language used in the repository.
* forks\_count: Number of forks of the repository.
* default\_branch: Default branch of the repository.

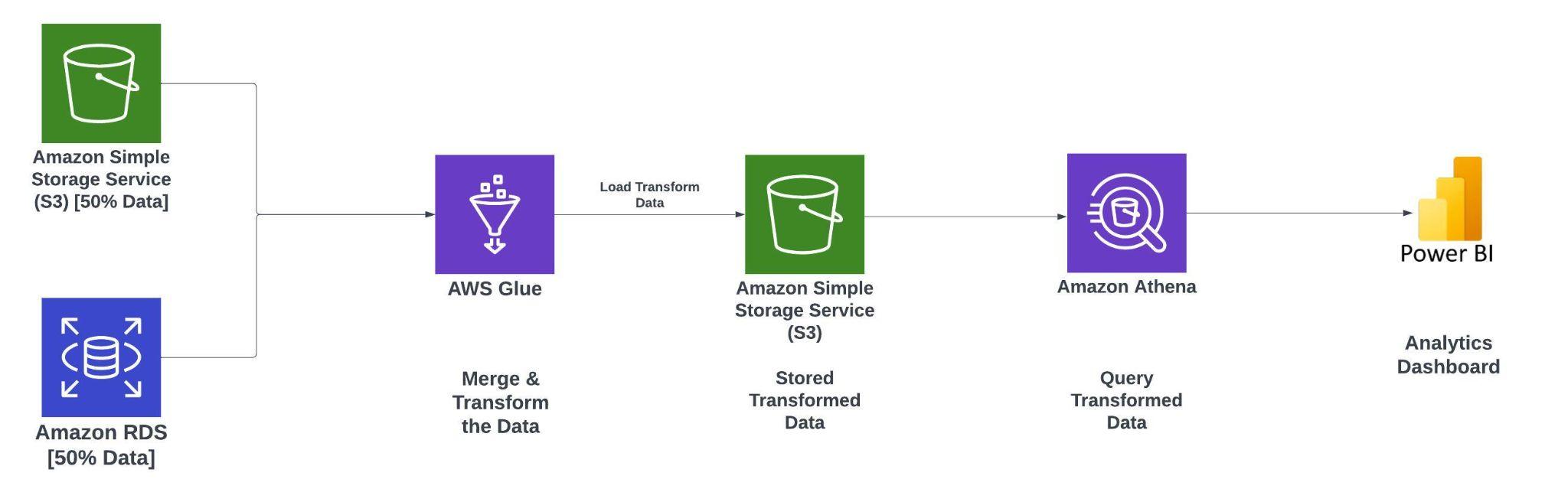
## Technologies Used

**─**

1. AWS: S3, RDS, GLUE, ATHENA, EMR, SPARK
2. POWER BI
3. PYTHON

## Architecture

**─**



## Key Performance Indicators (KPIs)

**─**

1. User Profile Analysis:

- Distribution Ratio: Percentage of personal, organizational, and bot accounts.

- Bot Detection Rate: Percentage of accounts flagged as bots.

2. User Activity Analysis:

- Average Commits: Mean number of commits per user.

- Average Repositories: Mean number of repositories per user.

- Average Followers: Mean number of followers per user.

3. Suspicious Account Detection:

- Suspicious Account Ratio: Percentage of accounts flagged as suspicious.

4. Commit and Repository Analysis:

- Commit Frequency: Average number of commits per repository.

- Repository Fork Ratio: Percentage of repositories that are forks.

- Average Stars: Mean number of stars per repository.

5. User Engagement Analysis:

- Follower-Following Ratio: Ratio of followers to followings for each user.

- Engagement Ratio: Percentage of followers who engage with a user's content.

6. Language and Repository Analysis:

- Language Usage Distribution: Percentage distribution of programming languages used in repositories.

- Top Repositories: Number of repositories with the highest stargazer counts.