yalo

Data & Insights Challenge

This challenge is designed to assess different skills related to your data management experience. This includes different areas in the world of data, such as: Data Warehousing, Data Management, Data Governance, Big Data, Data Analysis, Data Architecture and a few others.

We will be using the Google Cloud Platform environment. So it is necessary for you to create a free account. The processing of Queries will have no cost. You can consult the following reference links.

"BigQuery: Public datasets hosted in BigQuery provide users with free access of up to 1 TB/month in queries. Queries over 1 TB/mo are subject to query pricing. You can also leverage the BigQuery sandbox to freely access BigQuery, subject to the sandbox's limit."

https://cloud.google.com/bigquery/docs/sandbox
https://cloud.google.com/public-datasets
https://cloud.google.com/bigquery/pricing#free-tier
https://cloud.google.com/bigquery/public-data

The queries and any other material used to reach the results and conclusions must be delivered. Attached in a zip folder or share cloud drive. Important to name the folders with the challenge number.

Challenge 1

We want to update our programming languages and understand what the market is using, that's why we are going to analyze GitHub. In order to do this we are going to use the public data source that BigQuery has. The dataset is "GitHub Activity Data". Our CTO and CDO want to know the following information.

- What is the most popular language?
- What is the repository with the most languages?
- What is the repository with the most copies?
- Who is the most active commit author? In which repository and in what language?
    - Create a timeline with the number of commits the author has made per week based on the committer.time_sec field
- Top 10 repositories that most use "pandas"
- Identify which was the first repository to commit with "panda" and which was the last.

What would be your interpretation of this research and what could I recommend? Creativity is important in presenting this valuable data. (dashboard and executive presentation)