5 - Calculate and DW Load

Background story

Caladan is very impressed with the team's work! At last, all of their raw data is in one centralized location and normalized data is available in the ODS. Excitement is growing; they're so close to making sense of it all!

The team will now do the hard work of making a recommendation for policy implemenation. To do so they will need to design a Data Warehouse for serving the data to ML and Reports. The team is also going to need to calculate the effectiveness of each policy on the sample countries by calculating a growth percentage change on a daily basis and aggregate the growth percentage on a weekly basis.

Looking toward the future, if new data becomes available, the marketing team would like to be able to access it within an hour of its creation.

Caladan also wants the team to explore how to efficiently determine whether changes to their solution will work, preferably before deploying it to production. Currently, the process for ensuring the data import solution is correct is to manually analyze the results of a test execution. Considering the time and effort this process requires, and the ever-present chance of human error, they would like the team to automate the process.

Technical details

The team can create a rather simple Star/Snowflake schema or make it more elaborate as they see fit. The calculation should be done as the data is loaded into the Data Warehouse.

Note that the dimensions should all be considered Type 1 dimensions.

Reference: Type 2 dimension

Unit tests avoid using external resources such as remote services or http requests. External resources can be mocked or stubbed as appropriate.

Additional tests (e.g., integration, end to end, synthetic monitoring) would make use of these external resources to ensure that the system components are properly connected. However, this challenge deals only with the unit tests, which assert the more granular behaviors or logic of a single functional component.

Success criteria

- A data warehouse is created to support the marketing team's reporting requirements.
- A process exists to load data from the ODS into the data warehouse which serves the reports.
- While the team is currently working with a static, bulk dataset, the solution should support surfacing any new data in reports within one hour of its ingestion into the data lake.
- Calculations to determine daily and weekly growth percentage are performed inline and annotated for each day and week.
- The team has implemented at least one unit test on the growth calculations to exercise a component of the pipeline.

Tips

- The team will likely need to consider network security. Azure supports a mechanism to allow access from Trusted Microsoft services.

Resources

Ramp Up

Star Schema and Calculations

- Star schema
- Dimension based on a Star Schema Design
- Calculations during Transformation

Unit Testing

- Unit testing
- Continuous Integration & Continuous Delivery with Databricks
    - See *Productionize and write unit test* heading

Choose Your Tools

- What is Azure Synapse Analytics (formerly Azure SQL Data Warehouse)
- What is Azure Analysis Services
- Depending on the team's chosen path, a variety of unit testing tools will be applicable. The team's coach will provide more directed references to ensure the best fit with team's selected technologies.

Dive In

Azure Synapse Analytics (formerly Azure SQL Data Warehouse)

- Loading data with PolyBase
- Accessing Azure Synapse Analytics (formerly Azure SQL Data Warehouse) from Azure Databricks
- Copy data to or from Azure Synapse Analytics (formerly Azure SQL Data Warehouse) by using Azure Data Factory

Azure Analysis Services

- Tabular modeling (1400 compatibility level)

Automating tests for GitHub repositories

- Build and test with Travis CI
- Build and test with CircleCI