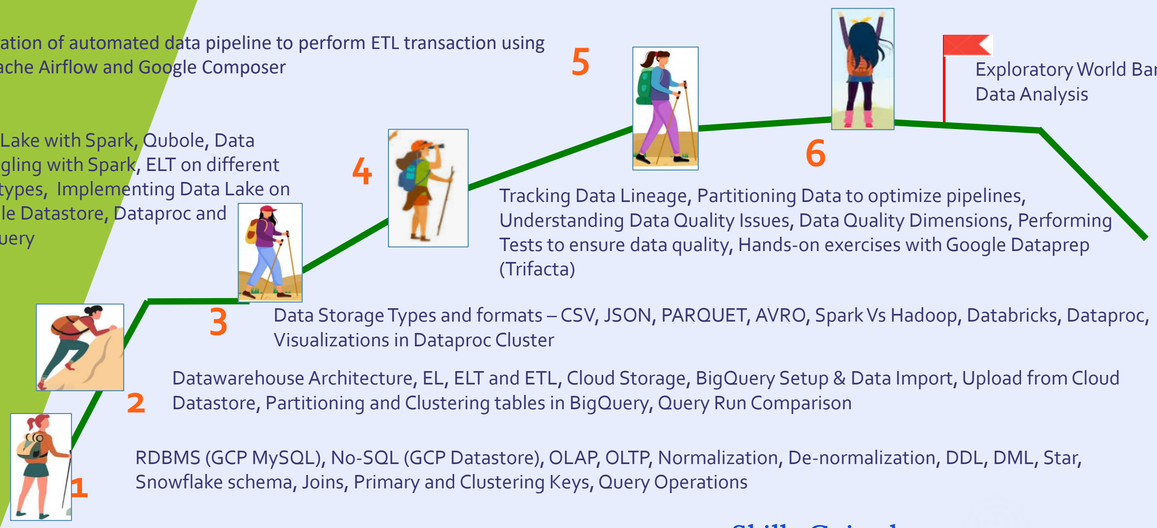
Sadiya Dalvi

Member of Technical Staff @ Acalvio Technologies

About Me

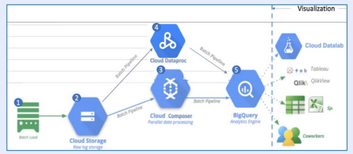
- Self-motivated and goal-oriented professional, offering over 14+ years of experience in software technology
- Bachelor's Degree in Electronic Engineering from the University of Mumbai
- Based out of Bangalore, India with husband and 2 children – a boy (12 yo) and a girl (6 yo)
- Running is my passion. Closely associated to Runner's High – a Bangalore based running group. Additionally, I enjoy reading, cooking and spending quality time with family.
- Reachable at dalvi.sadiya@gmail.com

GHCI Tech Pathways Data Engineering – Learning Journey



Challenge

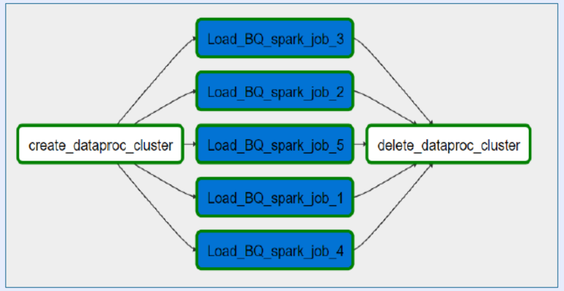
- Create a CI/CD pipeline based on the architecture given below to perform an exploratory data analysis on the World Bank – Global Health Dataset



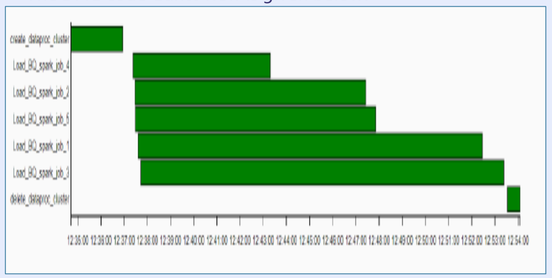
- Optimize the pipeline to be able to handle cost and scaling
- World Bank data consists of demographic and other statistical data related to Population, Employment, Health, GDP, Energy Consumption, etc. for all the countries from the year 1960 to 2018.
- Insights must be derived.

Implementation

- Extracted data from Google dataset as CSV files into Google Storage Bucket.
- Created a CI/CD Pipeline using the Google Cloud Composer and Cloud Build.
- Created a CloudBuild trigger. This trigger is integrated with the Github repository. Whenever a new dag file is checkedin to the Master branch a Cloud Build job is triggered to sync the Airflow Cloud Storage bucket. Once the dag is detected by the composer and the airflow, the pipeline gets triggered.
- The Airflow DAG code consists of the following tasks:
  - Create a Dataproc Cluster
  - Run 5 parallel spark jobs to perform ETL operations
  - The ETL jobs loads the data in Google BigQuery
  - Delete the Dataproc cluster
- The DAG graph view is given below:



- The DAG code Gantt view is given below:



- Create Visualizations using Google Datalab

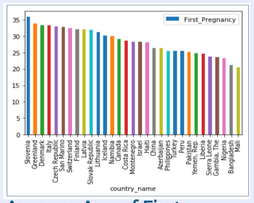
Learnings

- GCP Data analytics technologies simplifies working with huge datasets
- Similar dataset in MYSQL requires a data cluster and data sharding to be used. Additionally, it would need an infrastructure of networked file system that supports large distributed block storage
- Improved query performances by almost 3X times when ran in BigQuery Vs MySQL
- Tables when defined optimally, lead to query responses being under a minute. Performance gains with this approach make exploratory data analysis feasible
- It is a good practice to keep the Dataproc cluster alive only when running the batch load job, to save on running cost.
- The key to the performance magic is the PARTITION and CLUSTER Keys during BigQuery table creation. It is essential to make sure that the fields we want to query upon are in the set of CLUSTER keys

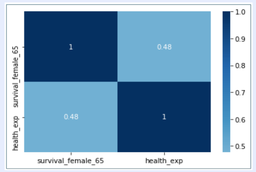
Skills Gained

- Exploring huge data sets for developing important Insights using Google Cloud Technologies
- Setting up dynamic cloud environments such as compute instances, storage, etc. and cleaning them as part of the automated job
- Developing optimized pipelines ensuring minimal usage of resources and error handling capability
- Assumptions and appropriate estimations to make and derive insights when data is missing. For example, Average age of first pregnancy is estimated as the (average age of first marriage + 2 years) since the average age of first pregnancy was missing in the dataset.
- Python, Apache Spark, Pandas, numpy, Pyspark

Results & Visualizations



Average Age of First Pregnancy across the world



Correlation between Health Expenditure and Survival Ages of Females

Several Interesting relationships were found in our data. Few are explained below:

- Positive correlation was observed between the health expenditure and survival age for females
- The rate of infant mortality was low when the age of marriage was high
- The age of first pregnancy was highest in the European countries. It was also observed that richer countries have higher average first pregnancy age. However, this was not true for Saudi Arabia and hence it falls under the outliers.

Conclusion & Way Forward

- After analyzing 2-3 huge datasets, performing etl transactions and running jobs, I am ready to take up the real-world challenge and perform data engineering jobs in the industry
- I am looking at taking up a Data Engineering job to apply all the knowledge I have gained in this journey

together we build tech pathways

Author: Sadiya Dalvi
Feb 2021

ANITA B.ORG GRACE HOPPER CELEBRATION INDIA VIRTUAL