Capstone Project Title: Enhance analytics with Google Trends data using AWS Glue, Amazon Athena, and Amazon QuickSight

Objective: Use Google Trends data for a variety of analytical use cases. For example, use it to learn about how your products or brands are faring among targeted audiences. Also use it to monitor competitors and see how well they're performing against your brand.

Solution overview

The solution consists of the following components:

Amazon S3 – The storage layer that stores the list of topics for which Google Trends data has to be gathered. It also stores the results returned by Google Trends.

AWS Glue – The serverless data integration service that calls Google Trends for the list of topics to get the search results, aggregates the data, and loads it to Amazon S3.

Athena – The query engine that allows you to query the data stored in Amazon S3. You can use it for supporting one-time SQL queries on Google Trends data and for building dashboards using tools like QuickSight.

QuickSight – The reporting tool used for building visualizations.

In the following sections, we walk through the steps to set up the environment, download the libraries, create and run the AWS Glue job, and explore the data.

Steps:

1. Create an S3 bucket where you upload the list of movies and TV shows. For this post, we use a Netflix Movies and TV Shows public dataset
2. Create IAM service role that allows AWS Glue to read and write data to the S3 buckets you just created.
3. Create a new QuickSight account with the admin/author role and access granted to Athena and Amazon S3.

Conclusion

Integrating external data sources such as Google Trends via AWS Glue, Athena, and QuickSight can help you enrich your datasets to yield greater insights. We can use it in a data science context when the model is under-fit and requires more relevant data in order to make better predictions. We used movies as an example, but the solution extends to a wide breadth of industries, such as products in a retail context or commodities in a finance context. If the simple inventory histories or the transaction dates are available, you may find little correlation to future demand or prices. But with an integrated data pipeline using external data, new relationships in the dataset make the model more reliable.

In a business context, whether our team wants to test out a machine learning (ML) proof of concept more quickly or have limited access to pertinent data, Google Trends integration is a relatively quick way to enrich your data for the purposes of ML and data insights.

We can also extend this concept to other third-party datasets, such as social media sentiment, as our team's expertise grows and your ML and analytics operations mature. Integrating external datasets such as Google Trends is just one part of the feature and data engineering process, but it's a great place to start and, in our experience, most often leads to better models that businesses can innovate from.