Capstone Project ETL Report: Group One

Introduction

For this project, we wanted to examine the relationship between a variety of financial indicators to predict certain socioeconomic health and quality of life factors on a global scale. We will do so using data derived from the World Bank Databank for the years 1990-2020.

The data, in its current form, does not match the format we need to carry out the analysis we have planned. First and foremost, the World Bank uses the years as column headers as opposed to actual data values, so this is a major change we need to make within our data. Also, we aim to produce actionable predictive analysis from this data. Our plan for doing so is to use the data from one year to predict the outcomes of the next year. Therefore, we need to shift the outcome values of our data by one year to match the values we are using to predict them.

Data Source

Our main source of data for this project was the World Bank Databank, which we accessed using the World Bank data API. Here is a formal citation of this databank:

The World Bank, *World Bank Open Data*. (1990-2020). Retrieved from https://api.worldbank.org/v2/ on October 29, 2021.

Extraction

The data was called from the World Bank Data API using the standard API call format shown below. This format is adjusted for each indicator using its specific indicator ID.

`https://api.worldbank.org/v2/country/{country}/indicator/{indicator}?format=JSON&Per_page=31&date=1990:2020`

The extracted data is directly called into our Producer for transformation. Specifically, our producer loops through the following country codes and indicator codes, making a unique API call for each unique combination of country and indicator:

Indicator Codes	**Country Codes**
• CM.MKT.TRAD.GD.ZS • FP.CPI.TOTL.ZG • CM.MKT.TRNR • GC.TAX.TOTL.GD.ZS • NY.GDP.PCAP.CD • GC.DOD.TOTL.GD.ZS • SI.POV.GINI • FM.LBL.BMNY.GD.ZS • FI.RES.TOTL.CD	ABWAFGAGOALBANDAREARGARMASMATGAUSAUTAZ EBDIBELBENBFABGDBGRBHRBHSBIHBLRBLZBMUB OLBRABRBBRNBTNBWACAFCANCHECHICHLCHNCIV CMRCODCOGCOLCOMCPVCRICUBCUWCYMCYPCZEDE UDJIDMADNKDOMDZAECUEGYERIESPESTETHFINF JIFRAFROFSMGABGBRGEOGHAGIBGINGMBGNBGNQ GRCGRDGRLGTMGUMGUYHKGHNDHRVHTIHUNIDNIM NINDIRLIRNIRQISLISRITAJAMJORJPNKAZKENK GZKHMKIRKNAKORKWTLAOLBNLBRLBYLCALIELKA LSOLTULUXLVAMACMAFMARMCOMDAMDGMDVMEXMH LMKDMLIMLTMMRMNEMNGMNPMOZMRTMUSMWIMYSN AMNCLNERNGANICNLDNORNPLNRUNZLOMNPAKPAN

• SM.POP.TOTL • SI.DST.10TH.10 • SI.DST.FRST.10 • SI.POV.NAHC • SP.DYN.LE00.IN • SP.DYN.IMRT.IN • SL.UEM.TOTL.ZS • NE.EXP.GNFS.ZS	PERPHLPLWPNGPOLPRIPRKPRTPRYPSEPYFQATRO URUSRWASAUSDNSENSGPSLBSLESLVSMRSOMSRBS SDSTPSURSVKSVNSWESWZSXMSYCSYRTCATCDTGO THATJKTKMTLSTONTTOTUNTURTUVTZAUGAUKRUR YUSAUZBVCTVENVGBVIRVNMVUTWSMXKXYEMZAFZ MBZWE

Each API call returns a list of dictionaries, where each dictionary contains the value of an indicator for one country for one year. For example, the API call

https://api.worldbank.org/v2/country/USA/indicator/SI.POV.GINI?format=JSON&Per_page=31&date=1990:2020

returns a list of dictionaries in which each dictionary contains the GINI coefficient for the United States for one year. Meanwhile, we want to create a list of dictionaries where each dictionary contains the value of *every* indicator for a given country in a particular year. To accomplish this result, we first make API calls for every indicator in a country. Then, we collect the resulting dictionary-lists and loop through them all dictionary-by-dictionary, storing the values for each indicator in a new list of dictionaries organized by year. Finally, we repeat this process for every country in our list of countries.

Transformation

The way the data is called through the API, each country is given its own table, with the first column being the year, and each column after that being the particular financial, economic, or social indicator chosen from the World Bank Database. Thus, the issue of years being column headers instead of data values is accounted for in the process of calling the API.

The next major transformation step is to shift the data so that each row of data has the predictive variables from one year and the response variables from the next year. This is done using the pandas commands 'groupby()' and 'shift().' The command takes the form

```
df.groupby('Group')['Data'].shift(1)
```

where group is replaced by country, 'Data' is whichever dependent variable column we need to shift, and the 1 within the shift command moves the values in each row of that column by one spot.

Load

To load the data into a SQL database, we used a section of code in the consumer. After taking in the messages and sending them to the data lake, we were able to reload them into the consumer as a Pyspark dataframe. We also created a table in our cloud database that matched the data we gathered. We then loaded the connection strings to the database and our table for our group and stored them in variables. After loading the connection strings, we wrote the Pyspark dataframe to the database using the jdbc format. Using the overwrite option, we were able to create an entirely new table for each time

we read the Kafka messages into the database. This will decrease the likelihood of duplicates. After writing the Pyspark dataframe to the database, the database is ready for use.

Code Sample:

```
SQLdf.write.format('jdbc').option("url",
f"jdbc:sqlserver://{server}:1433;databaseName={database};") \
    .mode("overwrite") \
    .option("dbtable", table) \
    .option("user", user) \
    .option("password", password) \
    .option("driver", "com.microsoft.sqlserver.jdbc.SQLServerDriver") \
    .save()
```

Conclusion

Altogether, our ETL process converts a large quantity of freely accessible World Bank data into a format conducive to machine learning. By querying many countries over a thirty-year period with a broad range of financial, economic, and social indicators, we were able to gather a bulky enough base of data to perform machine learning. Next, by transforming the data so that each unique pairing of country and year corresponds to a single dictionary of indicators, we collected the data into a single SQL table. In this state, our data is ready to input into a machine learning algorithm and answer our guiding research questions.