StatGPT

An AI-based SDMX Query Building Assistant

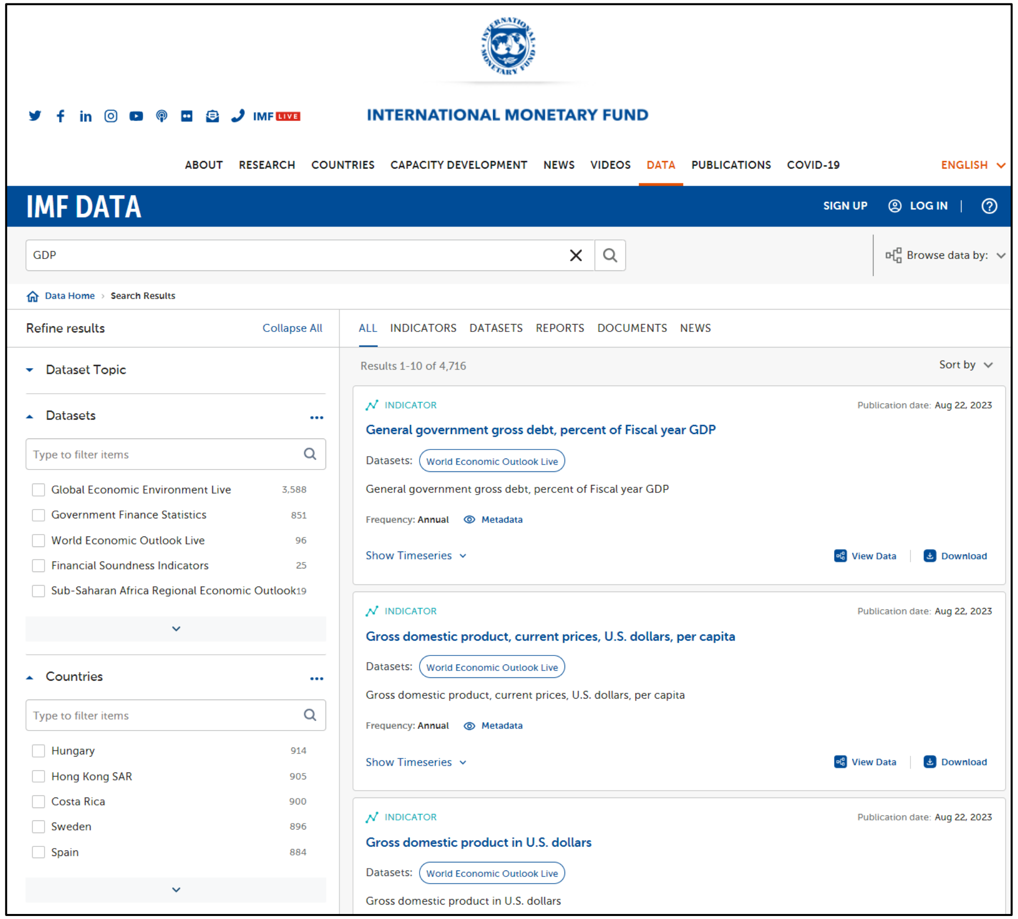
OCTOBER 29, 2023

Jeff Danforth (IMF)
Ilya Gorelik (EPAM)

Outline

- Objectives
- Business capabilities
- Architecture and how it works
- Technical challenges
- Lessons learned
- Demo
- Next steps

StatGPT Objectives



GDP

Results 1-10 of 4,716

Dataset Topic

Datasets

☐ Global Economic Environment Live	3,588
☐ Government Finance Statistics	851
☐ World Economic Outlook Live	96
☐ Financial Soundness Indicators	25
☐ Sub-Saharan Africa Regional Economic Outlook	19

Countries

☐ Hungary	914
☐ Hong Kong SAR	905
☐ Costa Rica	900
☐ Sweden	896
☐ Spain	884

INDICATOR

Publication date: Aug 22, 2023

General government gross debt, percent of Fiscal year GDP

Datasets: World Economic Outlook Live

General government gross debt, percent of Fiscal year GDP

Frequency: Annual

INDICATOR

Publication date: Aug 22, 2023

Gross domestic product, current prices, U.S. dollars, per capita

Datasets: World Economic Outlook Live

Gross domestic product, current prices, U.S. dollars, per capita

Frequency: Annual

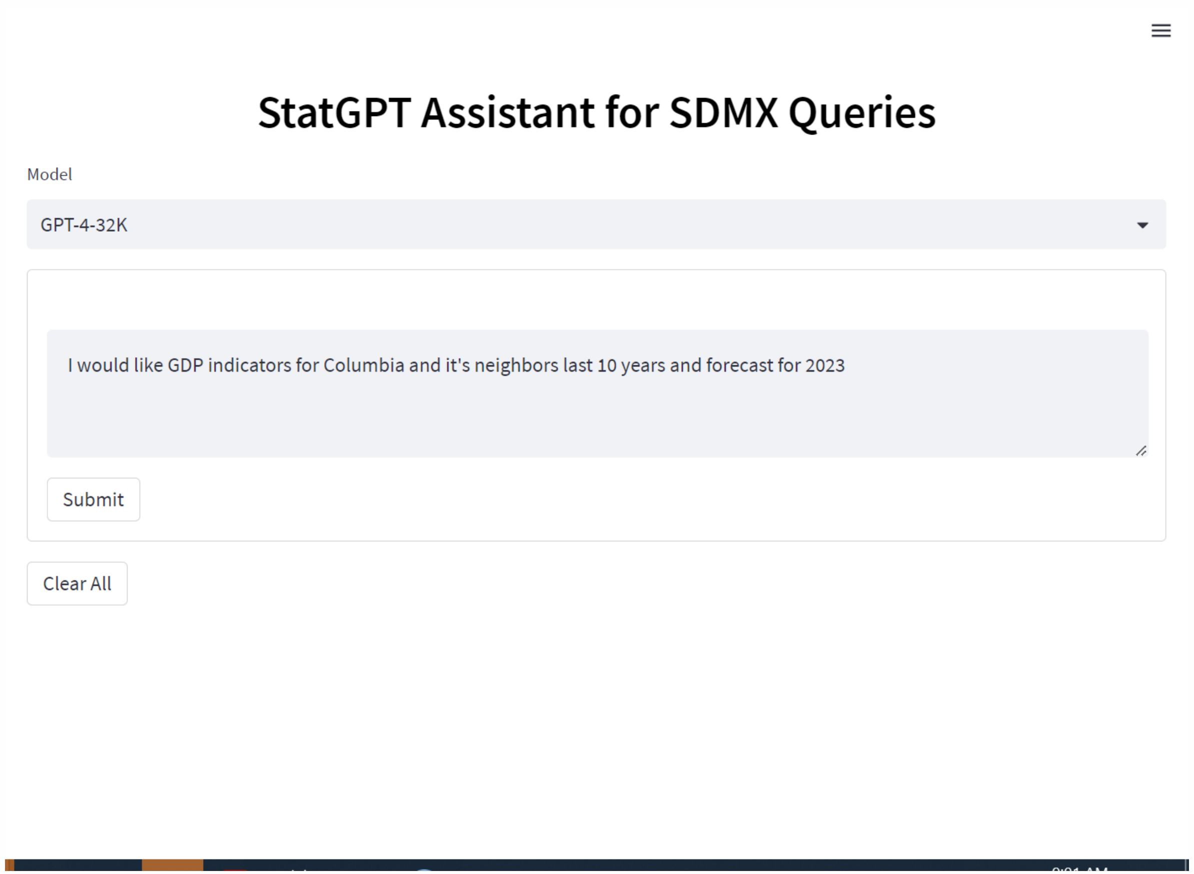
INDICATOR

Publication date: Aug 22, 2023

Gross domestic product in U.S. dollars

Datasets: World Economic Outlook Live

Gross domestic product in U.S. dollars



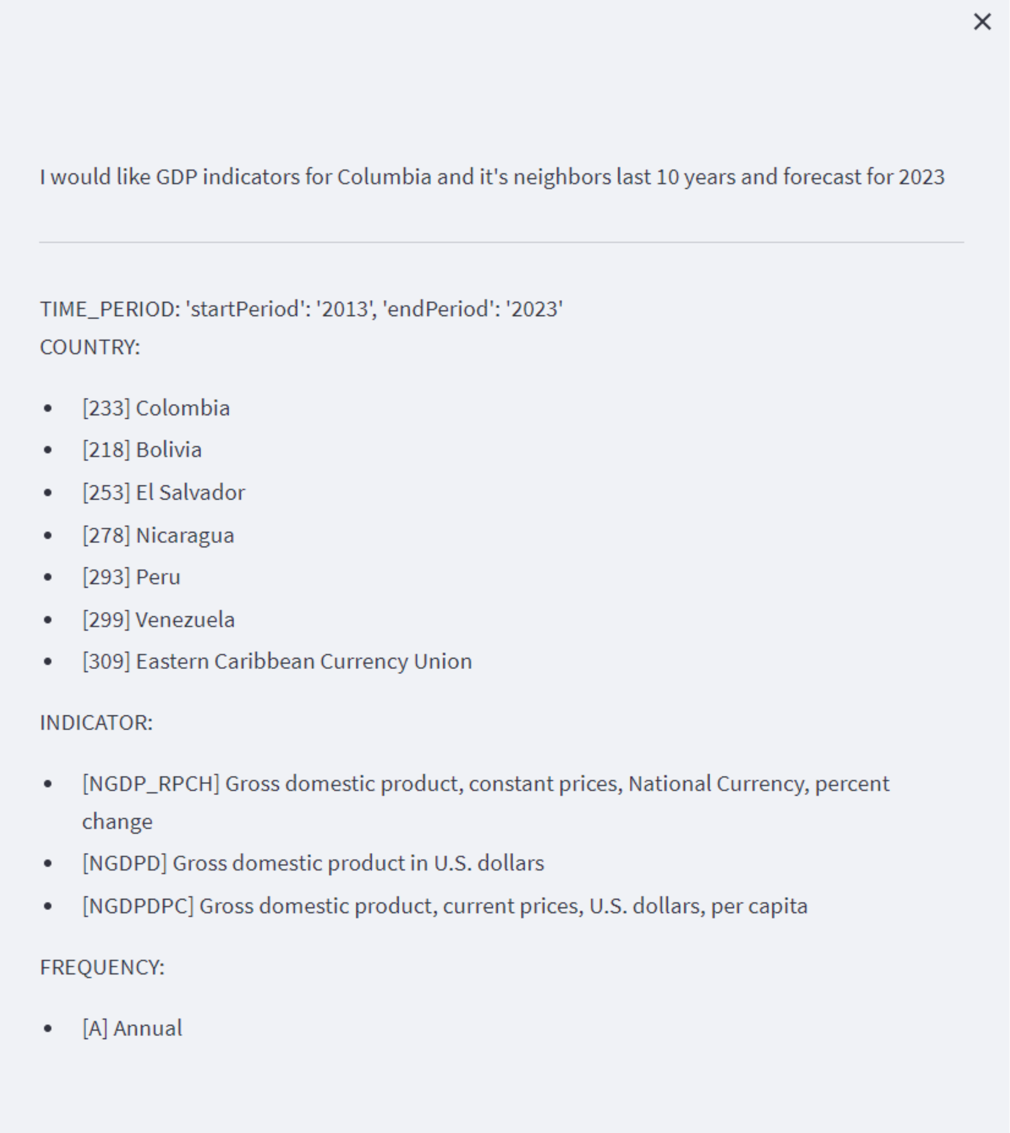
StatGPT Assistant for SDMX Queries

Model: GPT-4-32K

I would like GDP indicators for Columbia and it's neighbors last 10 years and forecast for 2023

StatGPT Capabilities: Generative

- Generative
  - Generate SDMX Queries
  - Generate SQL Queries
  - Generate REST Queries
  - Generate Code Snippets and Scripts



I would like GDP indicators for Columbia and it's neighbors last 10 years and forecast for 2023

TIME_PERIOD: 'startPeriod': '2013', 'endPeriod': '2023'

COUNTRY:

- [233] Colombia
- [218] Bolivia
- [253] El Salvador
- [278] Nicaragua
- [293] Peru
- [299] Venezuela
- [309] Eastern Caribbean Currency Union

INDICATOR:

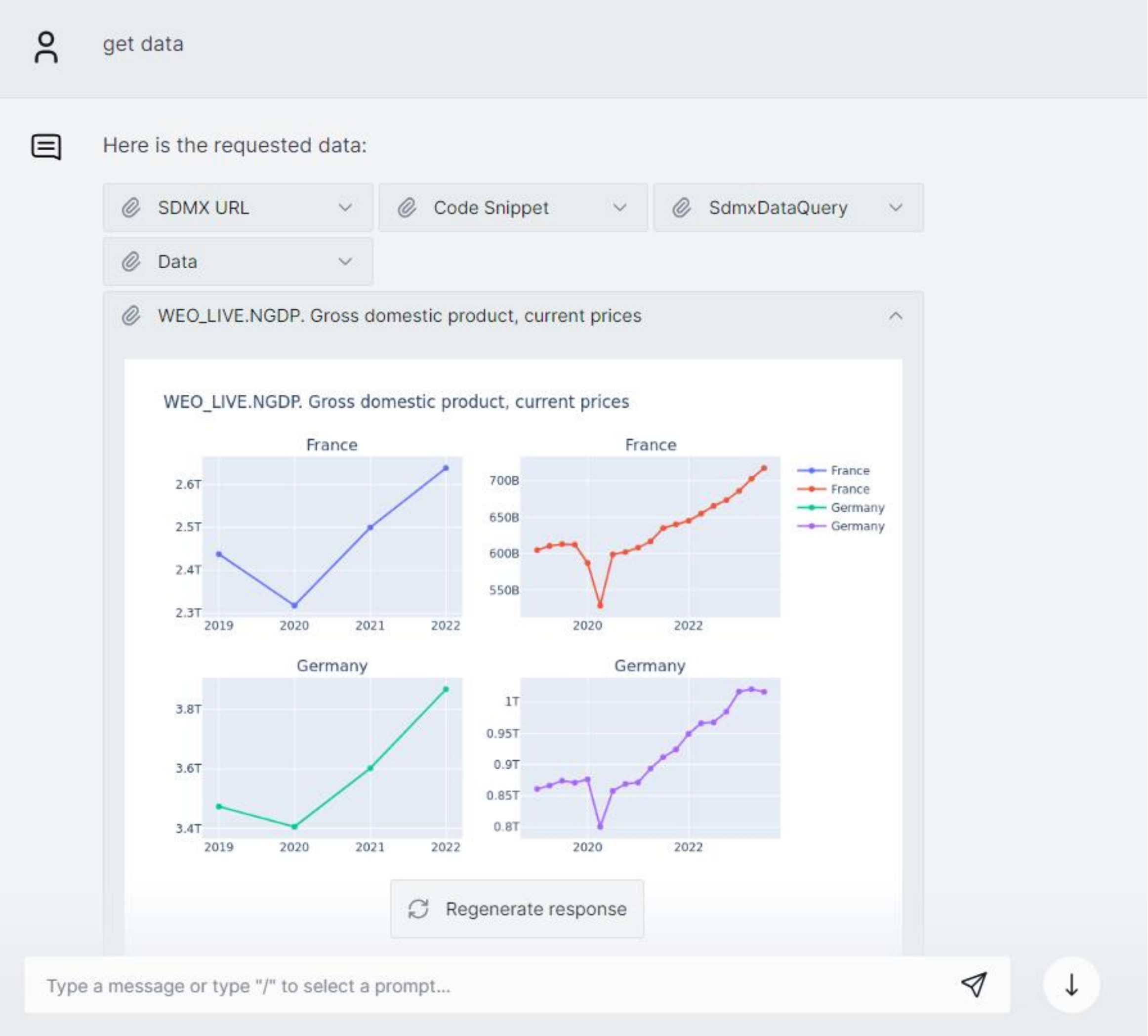
- [NGDP_RPCH] Gross domestic product, constant prices, National Currency, percent change
- [NGDPD] Gross domestic product in U.S. dollars
- [NGDPDPC] Gross domestic product, current prices, U.S. dollars, per capita

FREQUENCY:

- [A] Annual

StatGPT Capabilities: Visualization

- Visualization
  - Render Table Data
  - Render Charts
  - Render Queries
  - Render Code Snippets

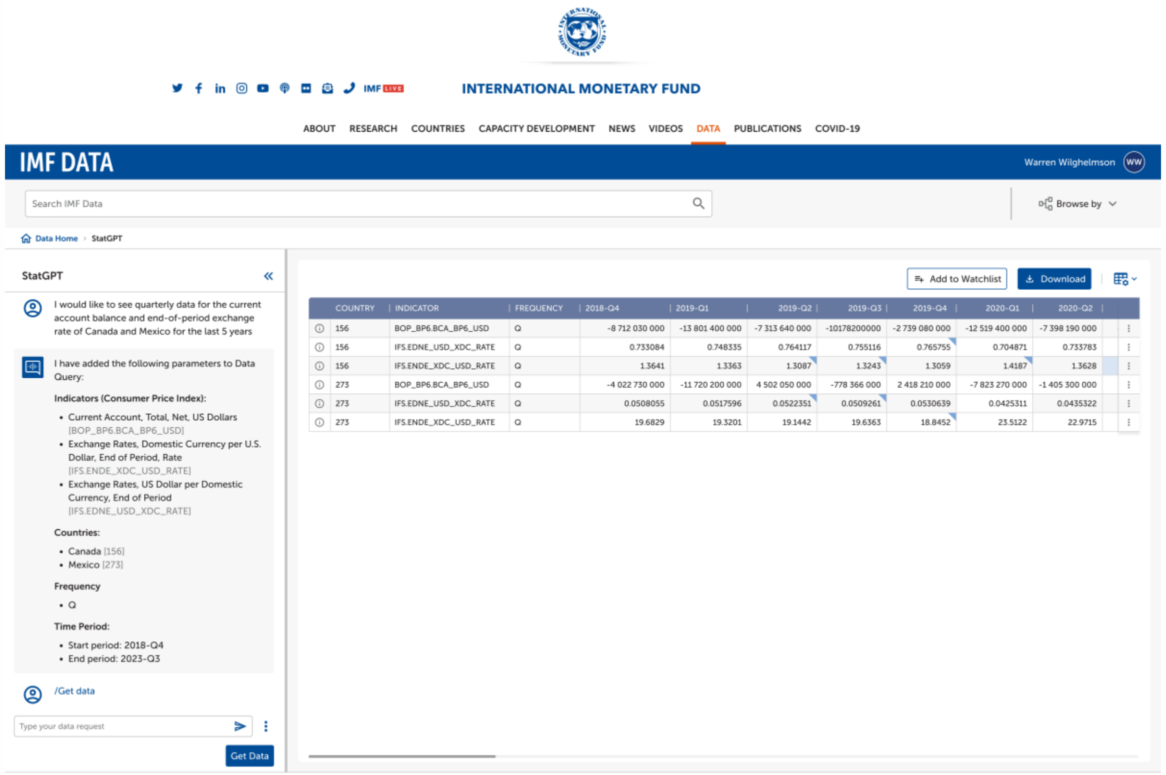


get data

Here is the requested data:

StatGPT Capabilities: Integration

- Integrate natural language query builder into existing dissemination tools



I would like to see quarterly data for the current account balance and end-of-period exchange rate of Canada and Mexico for the last 5 years

I have added the following parameters to Data Query:

Indicators (Consumer Price Index):

- Current Account, Total, Net, US Dollars [BOP_BP6.BCA_BP6_USD]
- Exchange Rates, Domestic Currency per U.S. Dollar, End of Period, Rate [IFS.ENDE_XDC_USD_RATE]
- Exchange Rates, US Dollar per Domestic Currency, End of Period [IFS.EDNE_USD_XDC_RATE]

Countries:

- Canada [156]
- Mexico [273]

Frequency

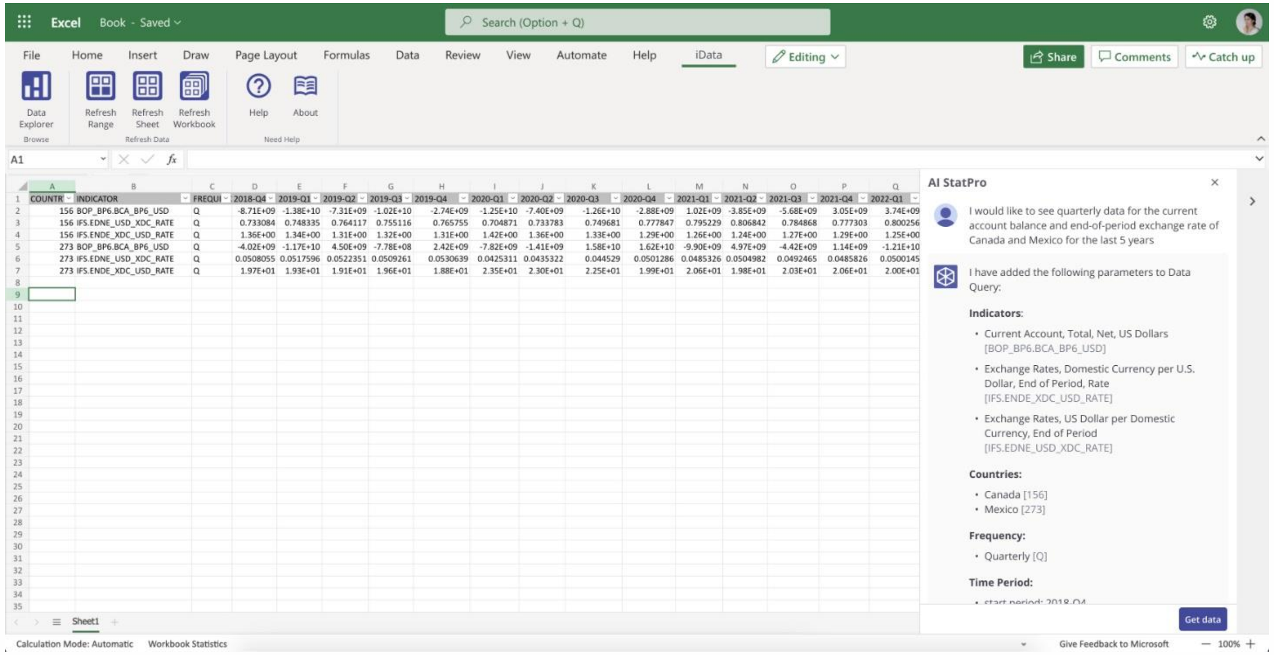
- Q

Time Period:

- Start period: 2018-Q4
- End period: 2023-Q3

/Get data

COUNTRY	INDICATOR	FREQUENCY	2018-Q4	2019-Q1	2019-Q2	2019-Q3	2019-Q4	2020-Q1	2020-Q2
156	BOP_BP6.BCA_BP6_USD	Q	-8 712 030 000	-13 801 400 000	-7 313 640 000	-10178200000	-2 739 080 000	-12 519 400 000	-7 398 190 000
156	IFS.EDNE_USD_XDC_RATE	Q	0.733084	0.748335	0.764117	0.755116	0.765755	0.704871	0.733783
156	IFS.ENDE_XDC_USD_RATE	Q	1.3641	1.3363	1.3087	1.3243	1.3059	1.4187	1.3628
273	BOP_BP6.BCA_BP6_USD	Q	-4 022 730 000	-11 720 200 000	4 502 050 000	-778 366 000	2 418 210 000	-7 823 270 000	-1 405 300 000
273	IFS.EDNE_USD_XDC_RATE	Q	0.0508055	0.0517596	0.0522351	0.0509261	0.0530639	0.0425311	0.0435322
273	IFS.ENDE_XDC_USD_RATE	Q	19.6829	19.3201	19.1442	19.6363	18.8452	23.5122	22.9715



COUNTR	INDICATOR	FREQUI	2018-Q4	2019-Q1	2019-Q2	2019-Q3	2019-Q4	2020-Q1	2020-Q2	2020-Q3	2020-Q4	2021-Q1	2021-Q2	2021-Q3	2021-Q4	2022-Q1
156	BOP_BP6.BCA_BP6_USD	Q	-8.71E+09	-1.38E+10	-7.31E+09	-1.02E+10	-2.74E+09	-1.25E+10	-7.40E+09	-1.26E+10	-2.88E+09	1.02E+09	-3.85E+09	-5.68E+09	3.05E+09	3.74E+09
156	IFS.EDNE_USD_XDC_RATE	Q	0.733084	0.748335	0.764117	0.755116	0.765755	0.704871	0.733783	0.749681	0.777847	0.795229	0.806842	0.784868	0.777303	0.800256
156	IFS.ENDE_XDC_USD_RATE	Q	1.36E+00	1.34E+00	1.31E+00	1.32E+00	1.31E+00	1.42E+00	1.36E+00	1.33E+00	1.29E+00	1.26E+00	1.24E+00	1.27E+00	1.29E+00	1.25E+00
273	BOP_BP6.BCA_BP6_USD	Q	-4.02E+09	-1.17E+10	4.50E+09	-7.78E+08	2.42E+09	-7.82E+09	-1.41E+09	1.58E+10	1.62E+10	-9.90E+09	4.97E+09	-4.42E+09	1.14E+09	-1.21E+10
273	IFS.EDNE_USD_XDC_RATE	Q	0.0508055	0.0517596	0.0522351	0.0509261	0.0530639	0.0425311	0.0435322	0.044529	0.0501286	0.0485326	0.0504982	0.0492465	0.0485826	0.0500145
273	IFS.ENDE_XDC_USD_RATE	Q	1.97E+01	1.93E+01	1.91E+01	1.96E+01	1.88E+01	2.35E+01	2.30E+01	2.25E+01	1.99E+01	2.06E+01	1.98E+01	2.03E+01	2.06E+01	2.00E+01

I would like to see quarterly data for the current account balance and end-of-period exchange rate of Canada and Mexico for the last 5 years

I have added the following parameters to Data Query:

Indicators:

- Current Account, Total, Net, US Dollars [BOP_BP6.BCA_BP6_USD]
- Exchange Rates, Domestic Currency per U.S. Dollar, End of Period, Rate [IFS.ENDE_XDC_USD_RATE]
- Exchange Rates, US Dollar per Domestic Currency, End of Period [IFS.EDNE_USD_XDC_RATE]

Countries:

- Canada [156]
- Mexico [273]

Frequency:

- Quarterly [Q]

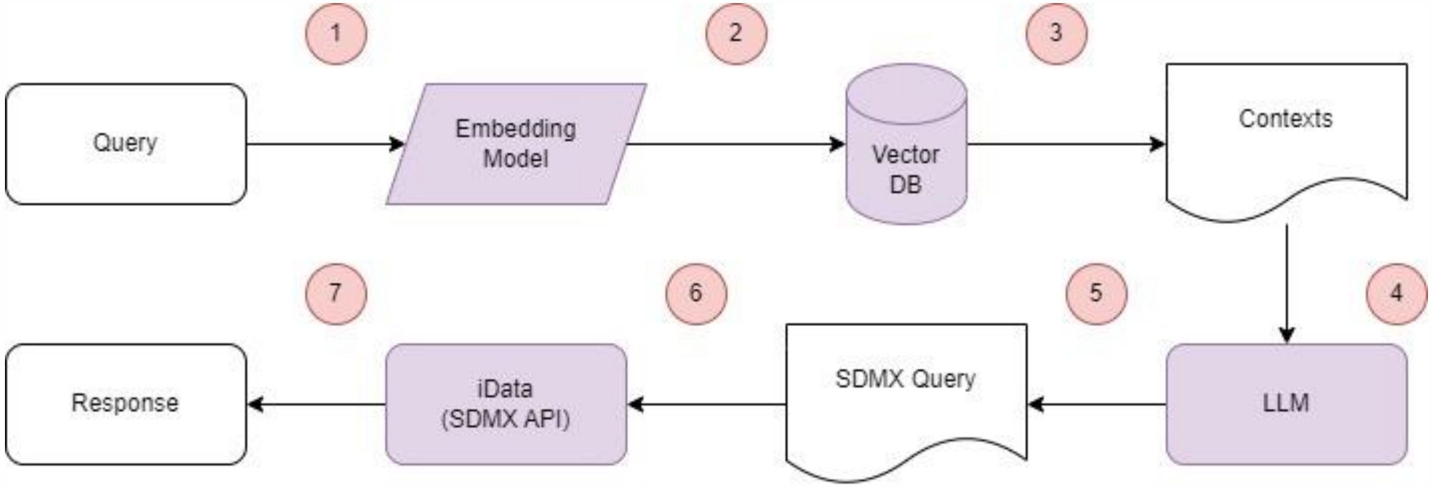
Time Period:

- start period: 2018-Q4

- Visualize data using existing interface
- Analyze data using existing Excel add-in

How it works

1. Process the query through embedding model to semantically represent original request as a query vector
2. Pass the vector to the vector DB that contains vectors for all indicators, countries and groups known to the system
3. Retrieve top-k relevant contexts measured by cosine distance between the query and embedded items in the knowledge base
4. Add relevant contexts to original user query and send it to LLM for parsing
5. Build SDMX query based on the response from LLM
6. Send SDMX query to SDMX API to get the data
7. Generate response using the retrieved data



Technical Challenges

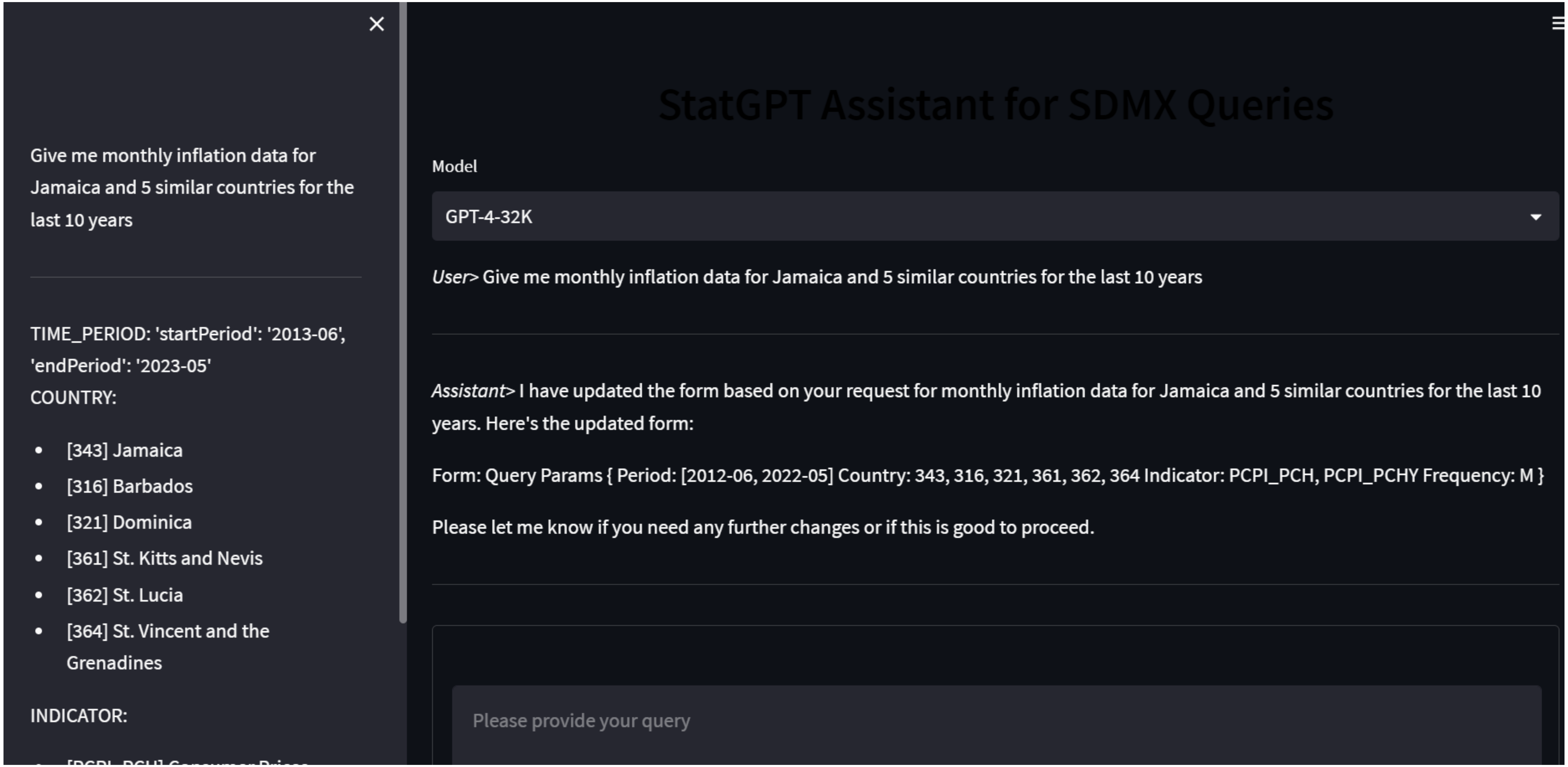
- Embeddings distance not always gives all relevant items, specifically for longer prompts
- Inclusion of multiple indicators and countries creates a large context for LLM (affects cost and speed of execution)
- Queries like “Give me data for France and all its neighbors” require a pre-flight LLM request to retrieve an actual list of neighbors
- Keeping only necessary and sufficient context during the conversation with LLM is not trivial
- Indicator prioritization (key indicators) is not supported by embeddings and must be implemented separately

Lessons Learned

- Concurrency is limited to capabilities of LLM (LLM is a bottleneck)
- LLM load-balancing is crucial for robust implementation (multiple accounts might be required)
- GPT-4 is an optimum model for the task (GPT-4-32K is slower and more expensive, GPT-3.5 is not that precise)
- Exposing the tool to internal users is much easier than making it available to public (cost and performance considerations)

DEMO

StatGPT



Give me monthly inflation data for Jamaica and 5 similar countries for the last 10 years

TIME_PERIOD: 'startPeriod': '2013-06', 'endPeriod': '2023-05'
COUNTRY:

- [343] Jamaica
- [316] Barbados
- [321] Dominica
- [361] St. Kitts and Nevis
- [362] St. Lucia
- [364] St. Vincent and the Grenadines

INDICATOR:

- [PCPI_PCH] Consumer Prices

StatGPT Assistant for SDMX Queries

Model: GPT-4-32K

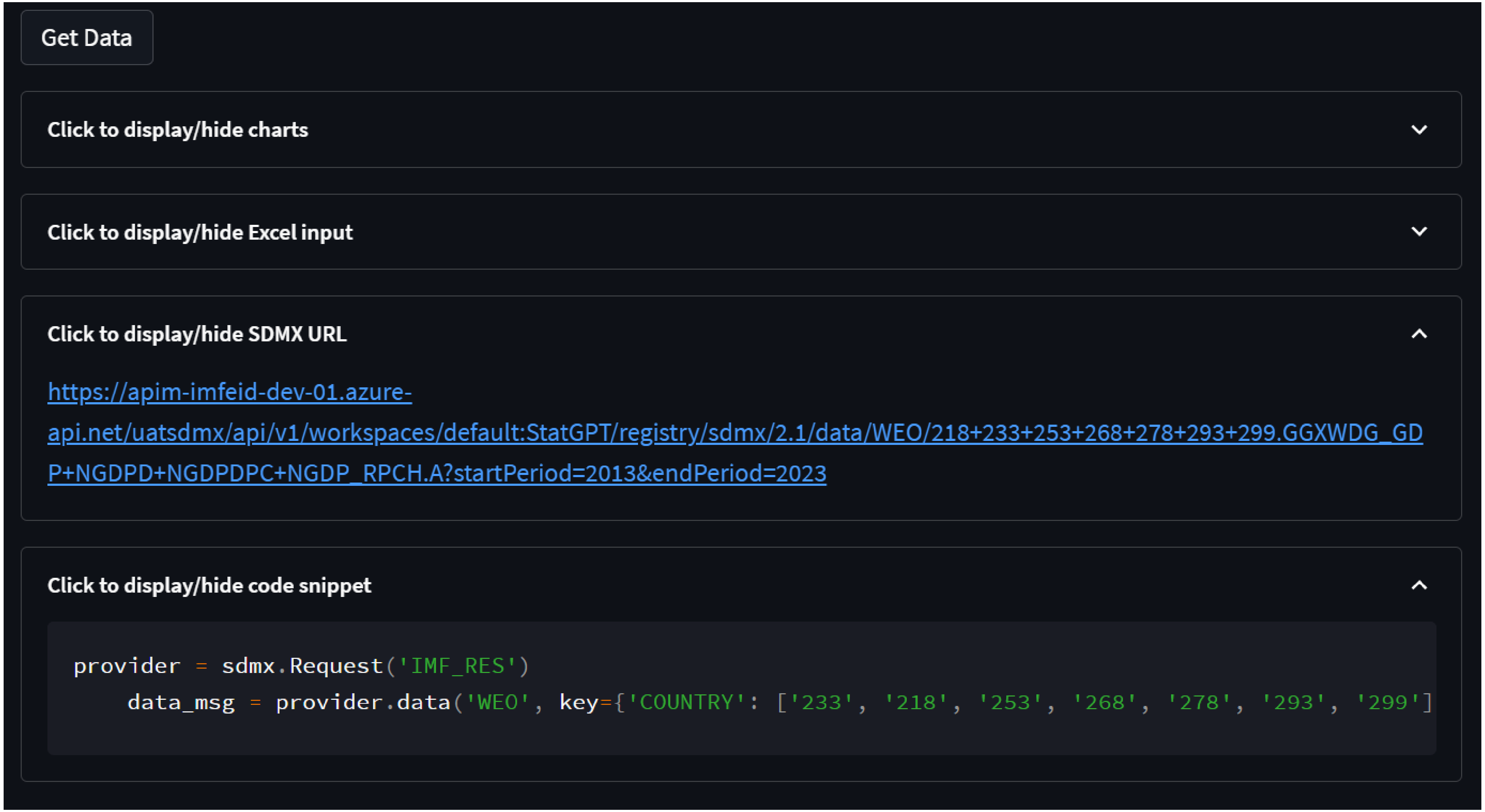
User> Give me monthly inflation data for Jamaica and 5 similar countries for the last 10 years

Assistant> I have updated the form based on your request for monthly inflation data for Jamaica and 5 similar countries for the last 10 years. Here's the updated form:

Form: Query Params { Period: [2012-06, 2022-05] Country: 343, 316, 321, 361, 362, 364 Indicator: PCPI_PCH, PCPI_PCHY Frequency: M }

Please let me know if you need any further changes or if this is good to proceed.

StatGPT


Get Data
Click to display/hide charts
Click to display/hide Excel input
Click to display/hide SDMX URL

https://apim-imfeid-dev-01.azure-api.net/uatsdmx/api/v1/workspaces/default:StatGPT/registry/sdmx/2.1/data/WEO/218+233+253+268+278+293+299.GGXWDG_GDP+NGDPD+NGDPDPC+NGDP_RPCH.A?startPeriod=2013&endPeriod=2023

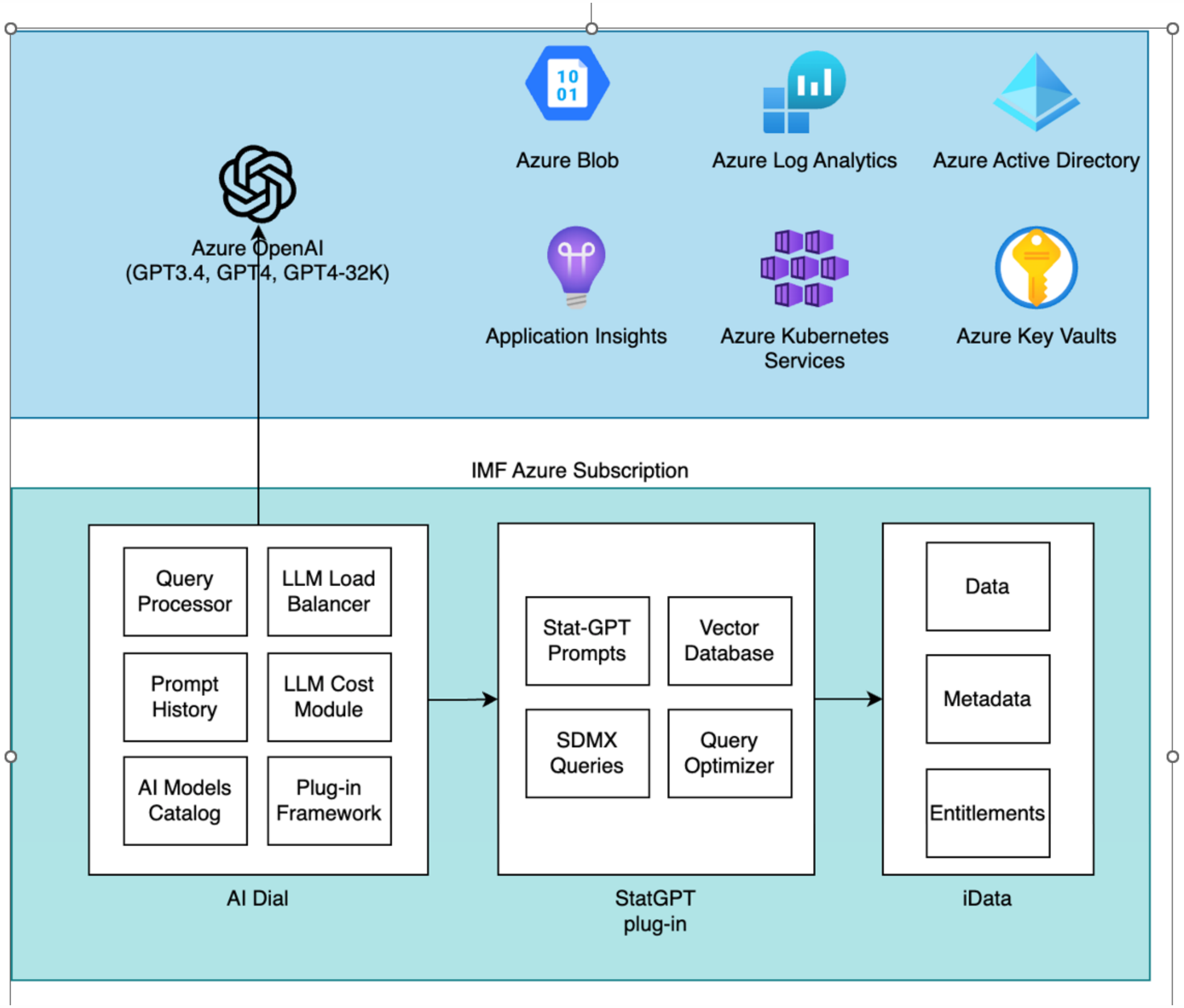
```
provider = sdmx.Request('IMF_RES')
    data_msg = provider.data('WEO', key={'COUNTRY': ['233', '218', '253', '268', '278', '293', '299']
```

Next Steps

- More training of the model with Fund users
- Implementing within IMF IT architecture
- Data transformations
- International collaboration?

QUESTIONS?

Architecture



- Integrates into existing IMF architecture
- Explanation of what AI Dial is?