**Specs and Work flow for Forecasting PM2.5 using big data analysis**

* **Prerequisites**

We need to have ArcGIS Enterprise licensed and configured with the linux based ArcGIS GeoAnalytics server and python3.6 has to be installed in the same machine.

* **Introduction**

This notebook demonstrates the geoanalytics server to forecast hourly PM2.5 given the historic time series data for more than one time-dependent variable. The most common factors in the weather environment used in this analysis are PM2.5, PM10, wind speed, wind direction, and relative humidity.

* Querying and summarizing your data using SQL
* Turning analysis workflows into pipelines of GeoAnalytics tools
* Modeling data with included machine learning libraries

The forecasted result will then be displayed on a dashboard using the recently introduced arcgis.dashboard module in ArcGIS API for Python.

* **Necessary imports**

Import all packages and libraries which are required for this analysis

* **Connect to your ArcGIS Enterprise organization**

Connect the ArcGIS Enterprise portal with the suitable credentials.

* **Ensure your GIS supports GeoAnalytics**

After connecting to the Enterprise organization, we need to ensure an ArcGIS Enterprise GIS is set up with a licensed GeoAnalytics server.

* **Prepare data**

Create a big data file share: To register a file share or an HDFS, we need to format datasets as subfolders within a single parent folder and register the parent folder. This parent folder becomes a datastore, and each subfolder becomes a dataset.

* **Get data for analysis**

Search for big data file shares:

Adding a big data file share to the Geoanalytics server adds a corresponding big data file share item on the portal. We can search for these types of items using the item\_type parameter.

* **Uncover patterns in data**

Describe data: The describe\_dataset method provides an overview of big data. By default, the tool outputs a table layer containing calculated field statistics and a dictionary outlining geometry and time settings for the input layer.

Commonly used methods of measurement: The 'Method Name' attribute contains information about the type of instrument used for measurement. 'Parameter Name' attribute tells about the name or description assigned in AQS to the parameter measured by the monitor. For more details, read

Average PM 2.5 value by county: This function filters the data by rows that give information about PM2.5 pollutant. To find the average PM2.5 value of each county, we will us join\_features tool.

* **Prepare time series data**

We have observed that the data is spread across US and comes from multiple stations. So we will create a dataframe that contains data points from one station. Additionally, for the purpose of sample, we will only use 2017 and 2018 data to train our model and foreacst on 2019 data.

* **Predict PM2.5 using Facebook's Prophet model**

This function uses pyspark pandas UDF function fb-prophet model to forecast pm2.5 hourly value for the month of 2019 January.

* **Visualize result on Dashboard**

The arcgis.apps module includes Dashboard submodule to create dashboards programmatically. The dashboard submodule contains classes for different widgets which can be configured and be used to publish a dashboard.

* **Conclusion**

Finally we have predicted and created interactive dashboard for the PM2.5 using Facebook’s Prophet Model.