User Guide (Version 1.0)

Spatial Data Lab (SDL)

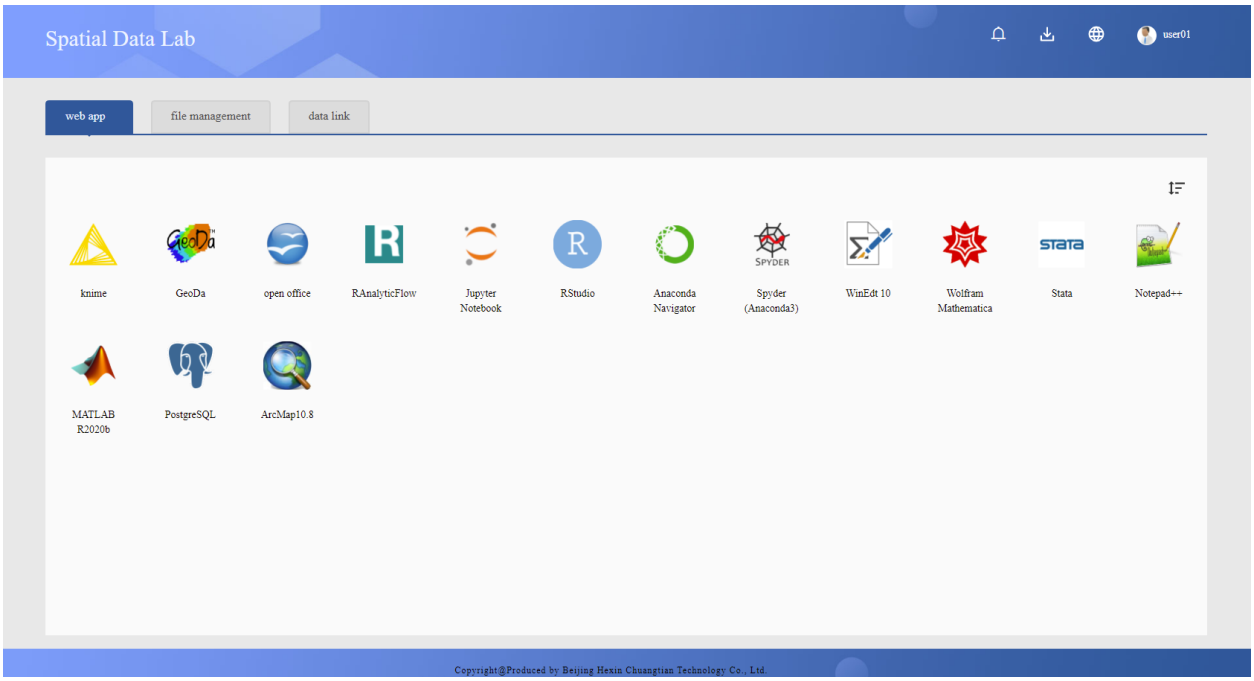
1. User Account Application

1) Submit a user account application via Google Form. Please carefully fill out your **project/research plan** in the form.
2) SDL team will review the applications and contact users after receiving applications.
3) User account will be assigned to users after verification.

2. SDL Platform Overview

1) Log into the platform (http://chinadatalab.org) via a web browser.

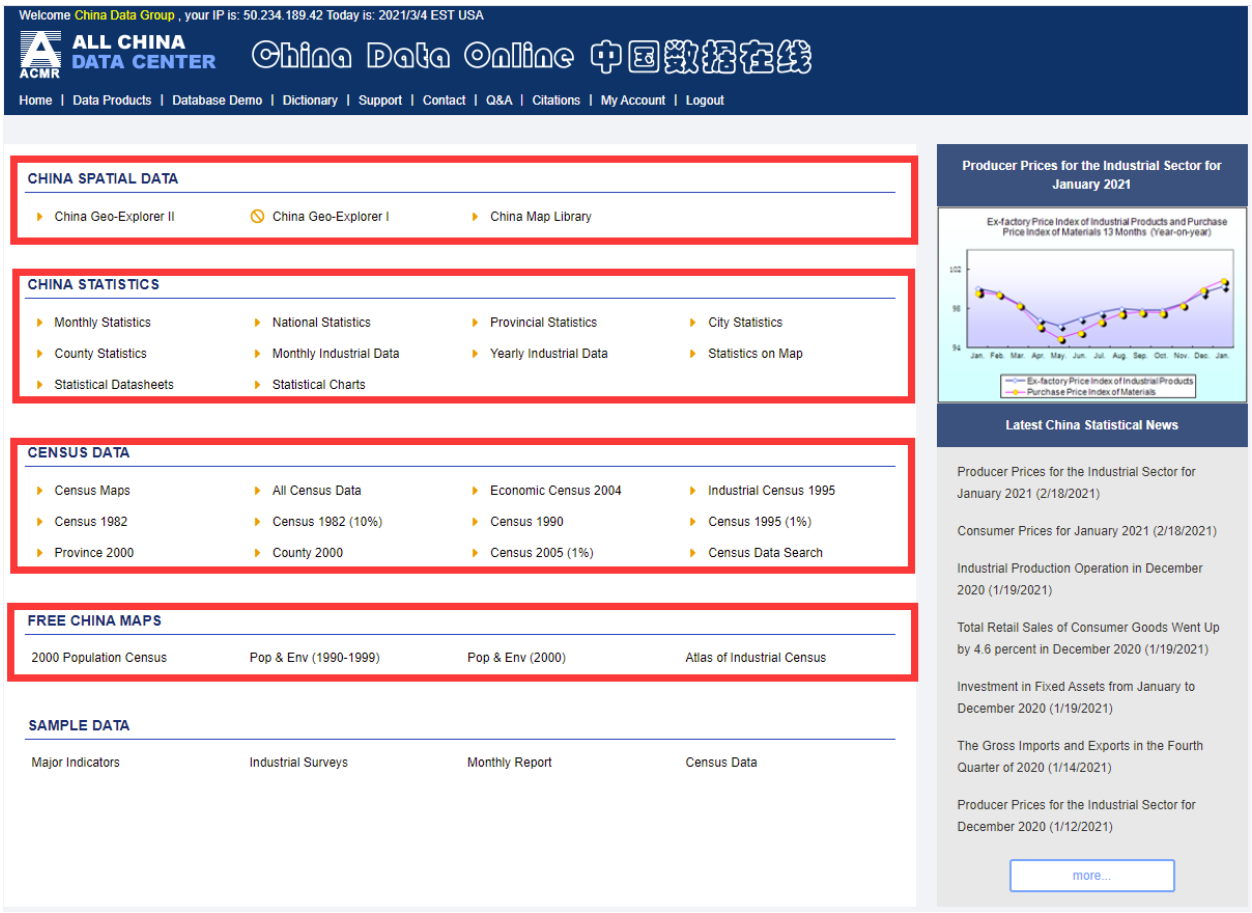
2) Input username and password assigned by SDL team. Once you successfully login to the platform, you will see the interface below.



3. Datasets Introduction

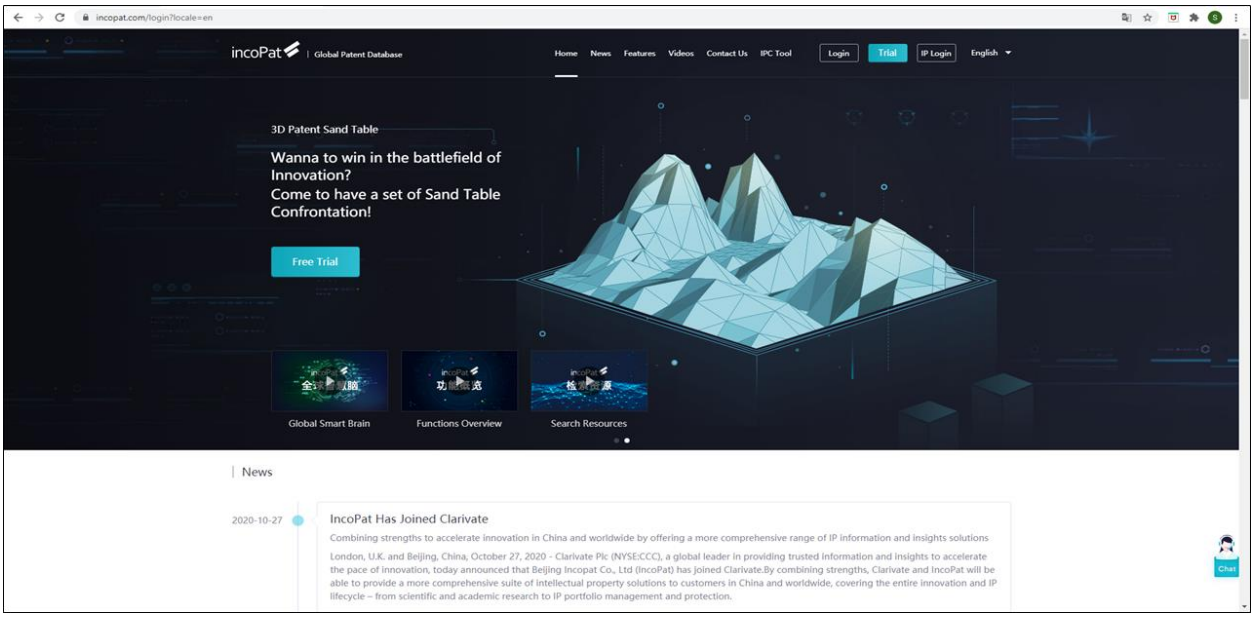
3.1 China Data Online

The China Data Online (http://china-data-online.com) is the primary data source for China studies. It includes (1) China Statistical Databases; (2) China Census Databases; and (3) China Spatial Data Service (China Geo-Explorer). It provides easy access to comprehensive statistics, and Census data of economy and population at national, provincial, city, county, and even township levels. Figure 1 shows the available datasets of China Data Online.



3.2 Patent Data

Collaborated with IncoPa company - A Global Patent Search and Analyze Platform, SDL platform integrates China patent data from 2008 to 2018. It includes over 57 variables, such as title, abstract, patent value, application number, application date, publication date, grant date, publication type, citation, citation-forward, citation number of times, family citation, citation applicant, cpc, applicant province, applicat city, application county, assignee, and industry class. All variables introductions can be found at shared excel on google drive.



SDL platform also includes United States patent data from 1975 to 2016, which are collected from USPTO (United States Patent and Trademark Office). The raw datasets are XML format, so they are analyzed and converted into structured data, saving into the PostSQL database. The US patent data includes more than 60 variables, such as abstracts, applicatiion_country, applicatin_date, application_type, assignee_city, assignee_state, assignee_country, assignee_orgname, assignee_role, citation_doc_country, citation_doc_date, cpc_class, invention_title, inventin_address_city, ipcsr_class, publication_date, and pubication_doc_number.

3.3 COVID-19 Datasets

Our team started to collect COVID-19 related datasets from varieties of datasets since the beginning of the COVID-19 pandemic. The datasets are shared on Harvard Dataverse, which is a meta data sharing platform. The introduction to the datasets are described at China Data Lab website. Users can find other related resources as well. Table 1 below lists all available datasets in Dataverse and more information can be found at the shared data list The shared datasets have been accessed by global users from over 150 countries and downloaded for more than 400,000 times. SDL

platform also makes a copy of the datasets, which are also accessible for platform users.

ID	数据项	Data Sets	Availability	
1	全球疫情数据	Coronavirus cases data	Harvard Dataverse	https://doi.org/10.7910/DVN/L20LOT
2	全美疫情数据	US Cases Data	Harvard Dataverse	https://doi.org/10.7910/DVN/HIDLTK
3	中国疫情数据	China Cases Data	Harvard Dataverse	https://doi.org/10.7910/DVN/MR5IJN
4	中国人口流动数据	China Population mobility data	Harvard Dataverse	https://doi.org/10.7910/DVN/FAEZIO
5	医疗机构数据	Health facilities data	Harvard Dataverse/SDL	https://doi.org/10.7910/DVN/KRSGT3
6	行为轨迹数据	Trace data	SDL	
7	航线航班数据	Flight data	SDL	
8	高铁班次数据	High-speed train data	SDL	
9	全球新闻数据	Global News data	SDL	

10	社交媒体数据	Social media data	SDL/CGA	
11	政策数据	Policy Data	Harvard Dataverse	https://doi.org/10.7910/DVN/OAM2JK
12	气象气候数据	Meteorological data	Harvard Dataverse	https://doi.org/10.7910/DVN/TU0JDP
13	空气质量数据	Air Quality Data	Harvard Dataverse	https://doi.org/10.7910/DVN/XETLSS
14	社会经济数据	Socioeconomic Data	Harvard Dataverse	
15	疫苗数据	US Vaccine Data	Harvard Dataverse	https://doi.org/10.7910/DVN/Y4BQTT
		Global Vaccine Data	Harvard Dataverse	https://doi.org/10.7910/DVN/2M1WLR

3.4 Human Mobility Datasets

Human mobilities play an important role in varieties of applications, especially in the pandemics. Many studies applied human mobilities to explore how the COVID-19 pandemic has greatly impacted the society. We collaborated with Geoinformation and Big Data Research Laboratory (GIBD) at University of South Carolina, which developed ODT (Origin-Destination-Time) Flow Explorer to explore worldwide human mobility at various geographic scales. The human mobility indexes are derived by Geotagged Tweets and SafeGraph. Twitter-derived flows are updated to 12/31/2020, and SafeGraph-derived flows are updated to 02/24/2021. Two new geographic levels including the world first-level subdivision and US census tract (for South Carolina and Texas) are added.

3.5 Meteorological Data in China (1951 ~ 2020)

The meteorological data is provided by https://quotsoft.net/air/ and its time frame is from 1995 to 2020. The variables include EVP (evaporation), GST (ground temperature), PRE (pressure), PRS (air pressure), RHU (relative humidity), SDD (sunshine duration), TEM (air temperature), and WIN (wind speed).

3.6 Other Datasets

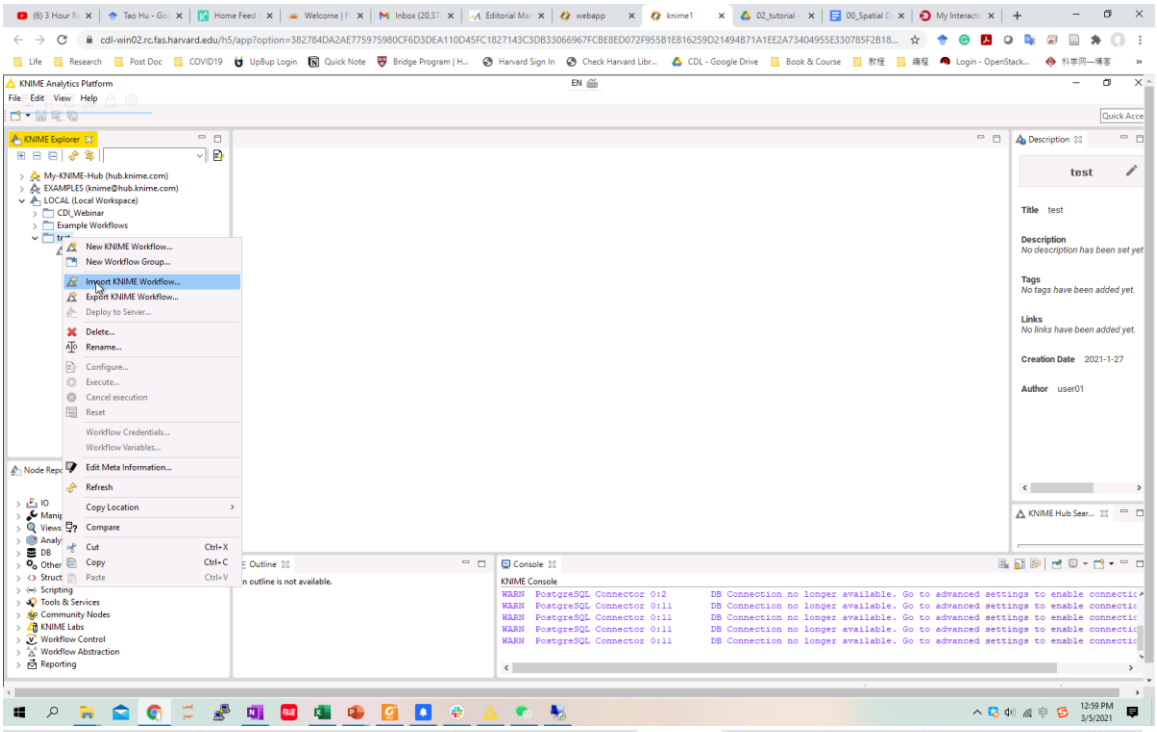
There are some other datasets shared on Spatial Data Lab platform. Please check more details about datasets on shared google doc.

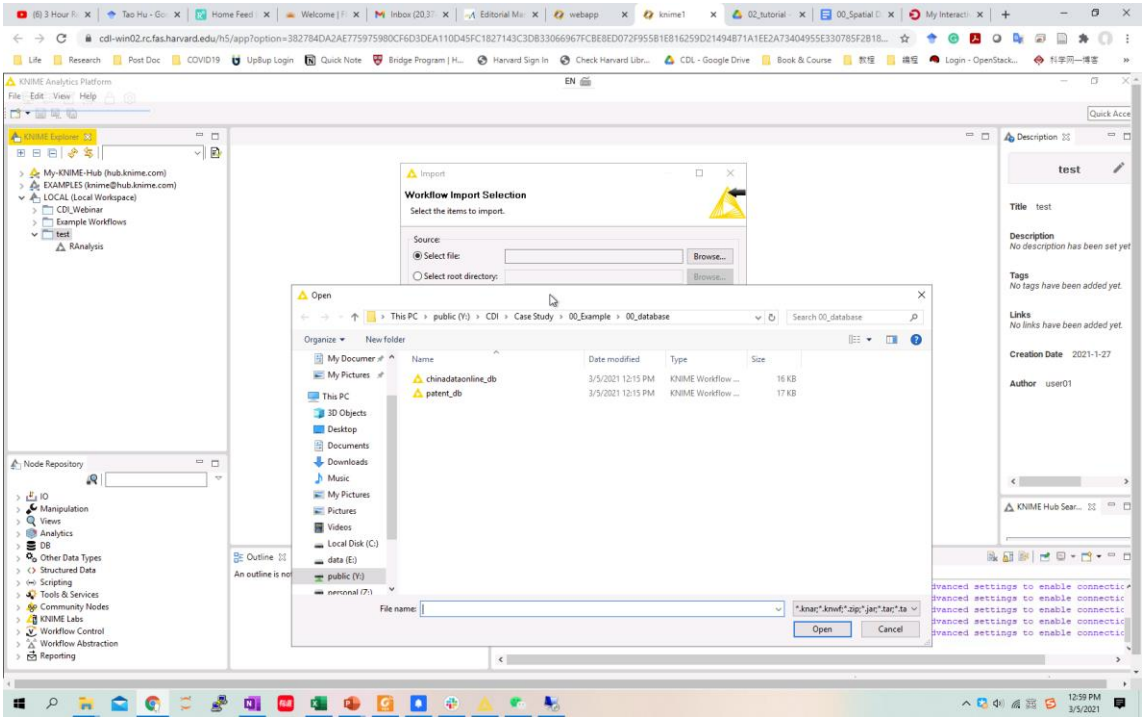
- County-level Air Quality Data
  - PM2.5 (2000–2019)
  - CO2 (1997–2017)
- China Annual Survey of Industrial Firms 工业企业数据库 (1998 ~ 2005, 2007, 2009)
- China City Statistical Yearbook 中国城市统计年鉴 (1993 ~ 2012)

4. Datasets Access

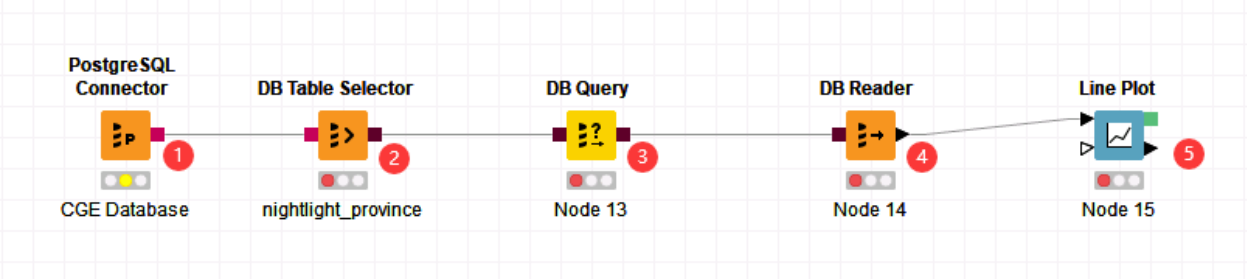
4.1 Data Access to China Data Online

1) Import Workflow from Public Folder **Y:\CDI\Case Study\00_Example\00_database\chinadataonline_db**

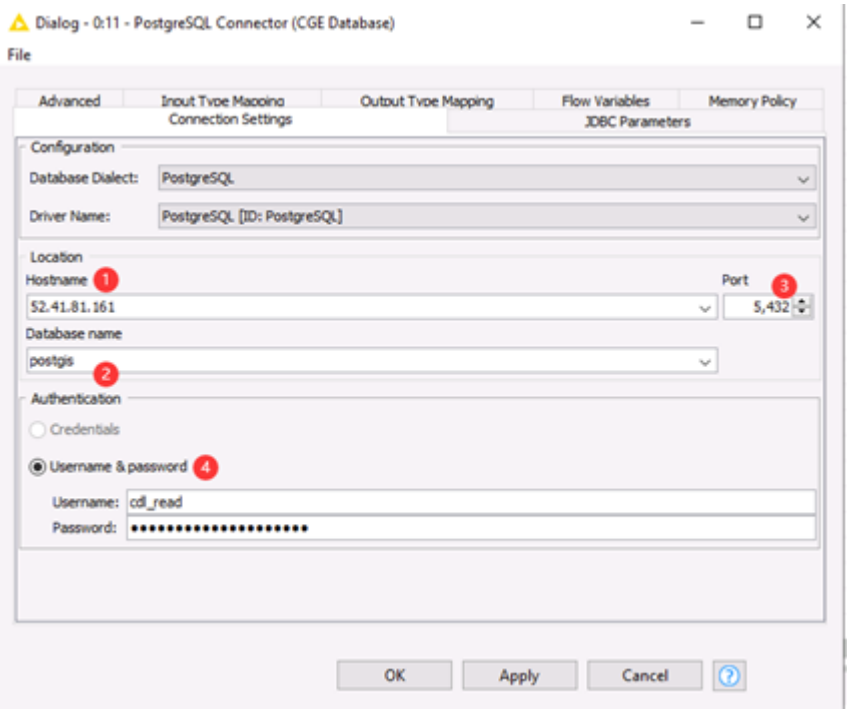




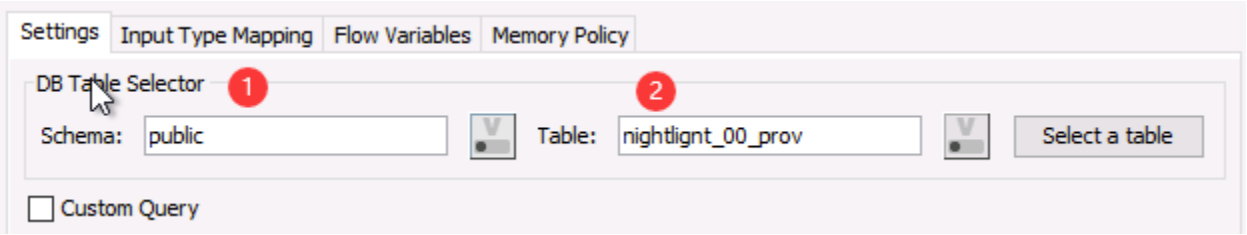
2) The imported workflow is shown as below. There are five components: PostgreSQL database settings; DB Table selector; DB Query; read data from DB; line plot visualization.



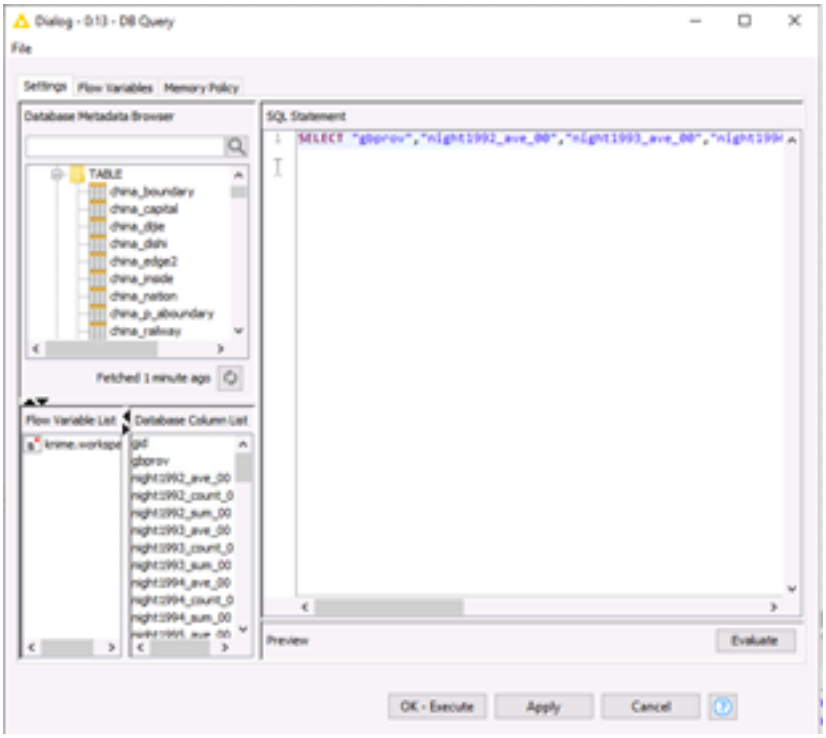
3) For PostgreSQL database settings, apply the default values. You can right click on the node and select ‘configure’ to check the settings.



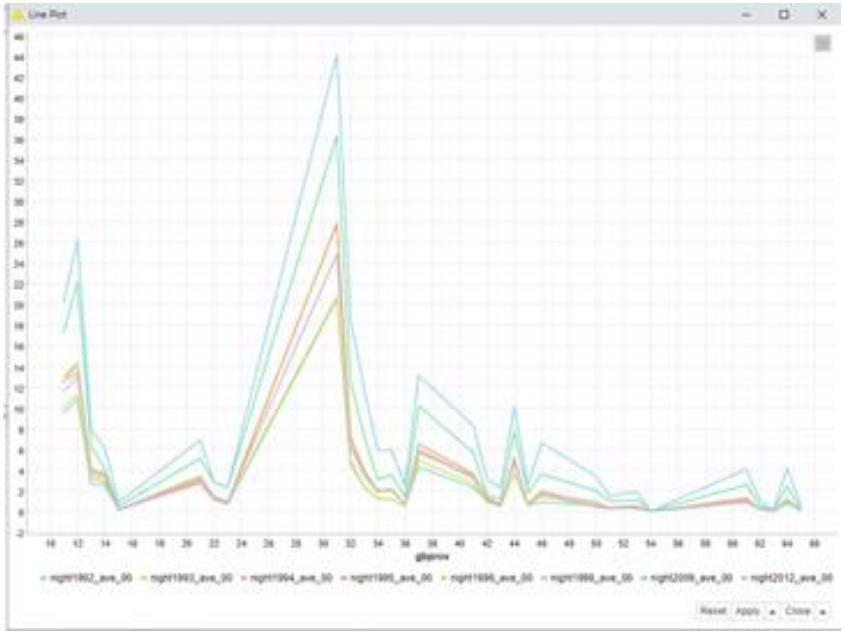
4) Select table in the node ‘DB Table Selector’



5) Create SQL in the node ‘DB Query’, such as “select "gbprov", "night1992_ave_00", "night1993_ave_00" from #table# AS "table""



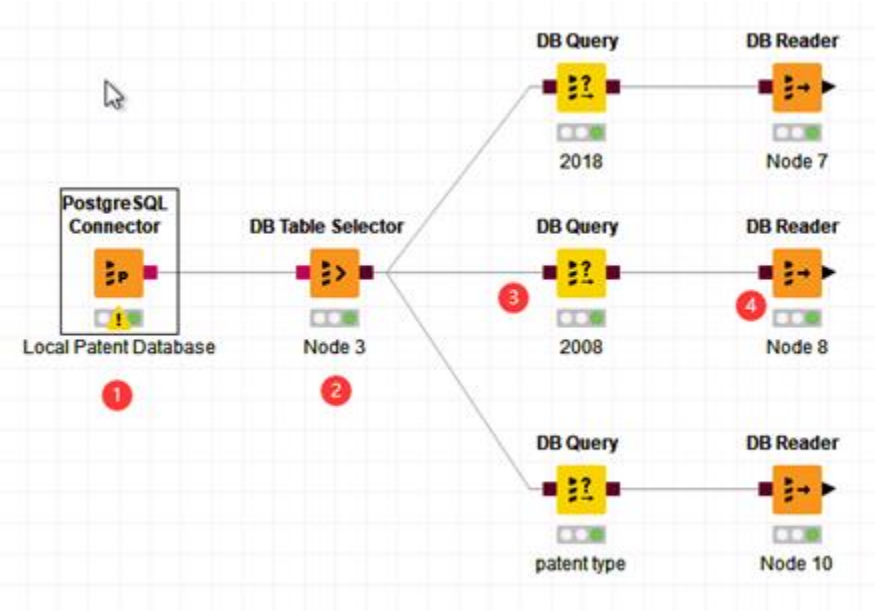
6) Load data by executing node ‘DB Reader’
7) Right click the node and select ‘execute’. After it is completed, right click the node and select ‘Interactive View: Line Plot’.



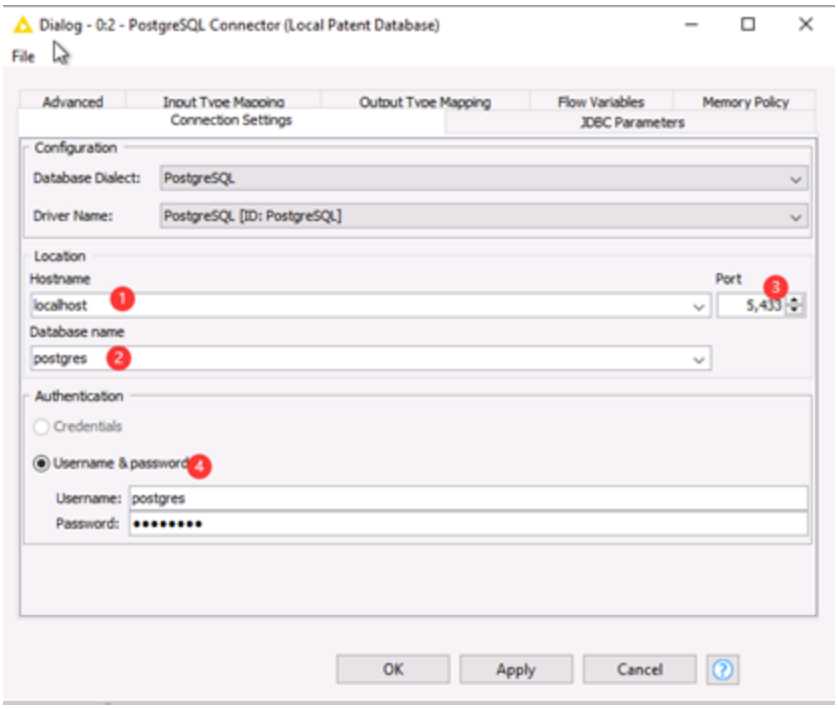
4.2 Data Access to SDL (Patent) Database

1） Import Workflow from Public Folder **Y:\CDI\Case Study\00_Example\00_database\patent_db**

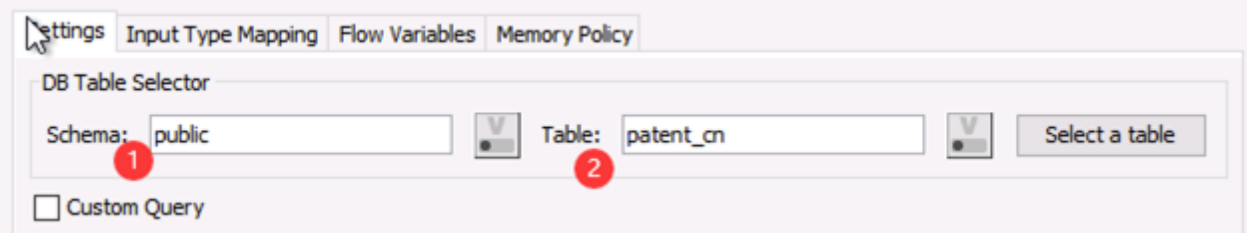
2) The imported workflow is shown as below. There are five components: PostgreSQL database settings; DB Table selector; DB Query; read data from DB to get results.



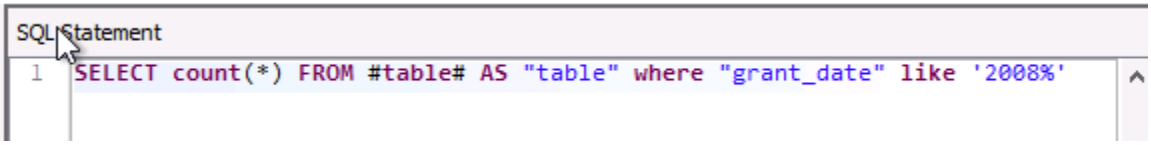
3) For PostgreSQL database settings, apply the default values. You can right click on the node and select ‘configure’ to check the settings.



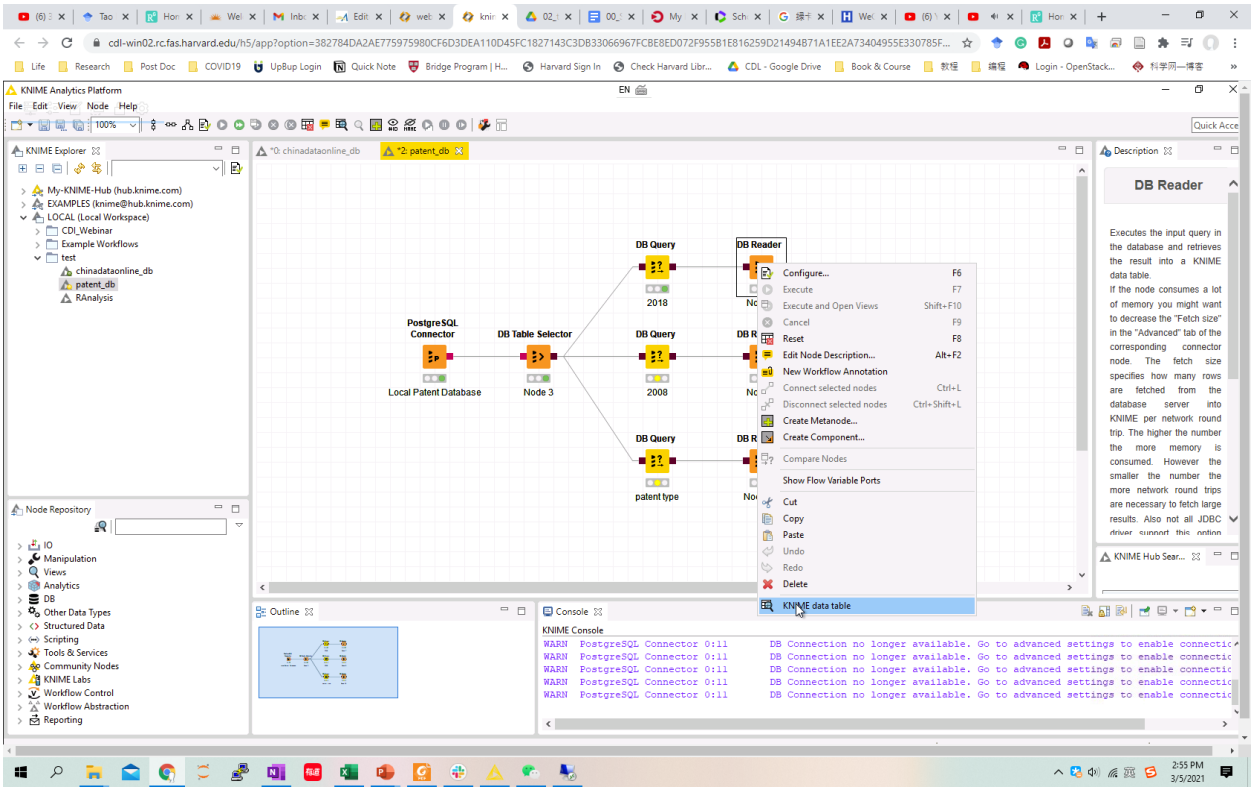
4)Select table in the node ‘DB Table Selector’



5) Create SQL in the node ‘DB Query’ shown as below

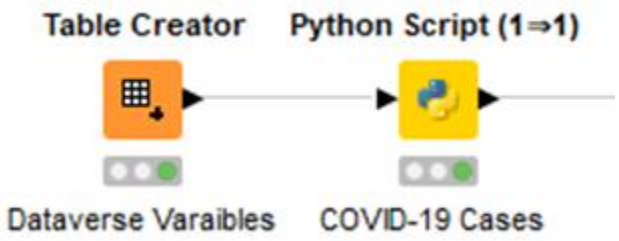


6) Check result from node ‘DB Reader’ by right clicking ‘table’



4.3 Data Access to Harvard Dataverse

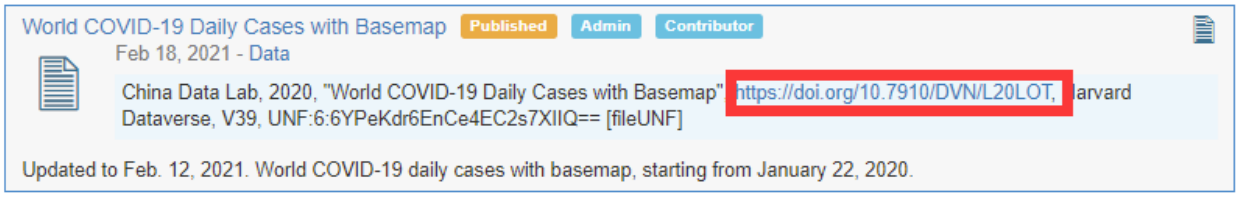
1) Import workflow from public case study folder **Y:\CDI\Case Study\00_Example\01_dataverse\dataverse_api.knwf**



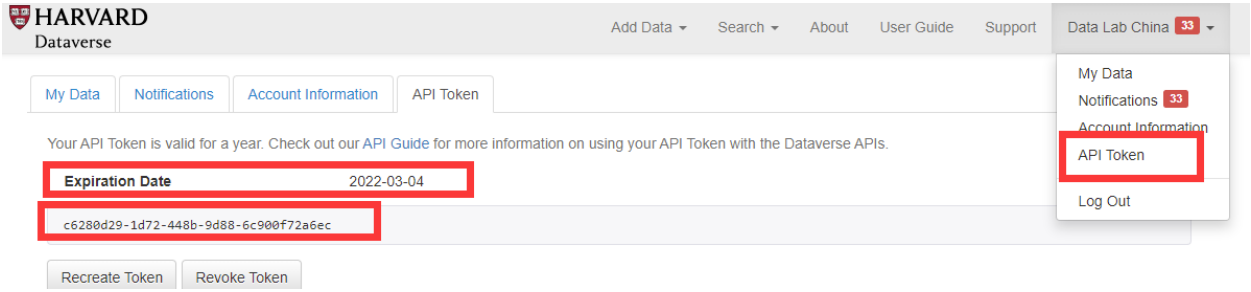
2) Set variables in the 1st node ‘Table Creator’. Right click node and select ‘configure’ and fill out the information in each parameter.

S doi	S api_token	S file_id
7910/DVN/L20LOT	b7918464-48b7-47da-88d8-749a95ffde12	4411772

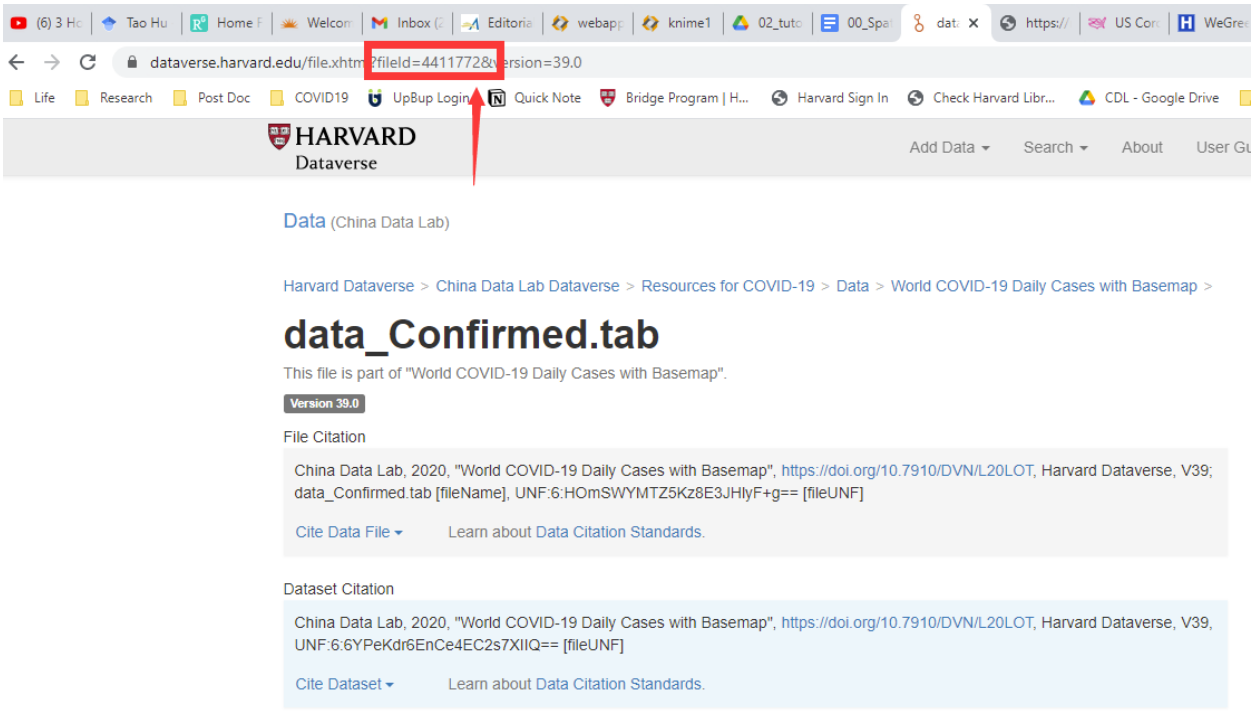
a) **Doi**: dataset doi number shown in the datasets metadata on Harvard Dataverse shown as below. After you go to https://dataverse.harvard.edu/dataverse/2019ncov, datasets will be listed and the DOI number is shown in each dataset.



b) **Api_token**: **c6280d29-1d72-448b-9d88-6c900f72a6ec** is the default value provided by SDL and it will be updated weekly. Users can generate new api_token after logging into Harvard Dataverse. Notice that the Token has an **expiration date**.



c) **File_id**: Go to the data file which you want to use and click the file. In the new web page, you will find the file id in the link. Please follow the steps shown below.



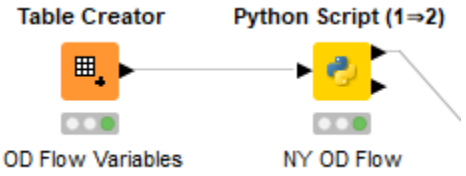
3) Right click the 2nd node ‘Python Script’ and select ‘table’ after it is finished. Users will see the results shown as below.

S Countr...	I 20200122	I 20200123	I 20200124
Afghanistan	0	0	0
Albania	0	0	0
Algeria	0	0	0
Andorra	0	0	0
Angola	0	0	0
Antigua and...	0	0	0
Argentina	0	0	0
Armenia	0	0	0
Australia	0	0	0
Austria	0	0	0
Azerbaijan	0	0	0
Bahamas	0	0	0
Bahrain	0	0	0
Bangladesh	0	0	0
Barbados	0	0	0
Belarus	0	0	0
Belgium	0	0	0
Belize	0	0	0
Benin	0	0	0
Bhutan	0	0	0
Bolivia	0	0	0

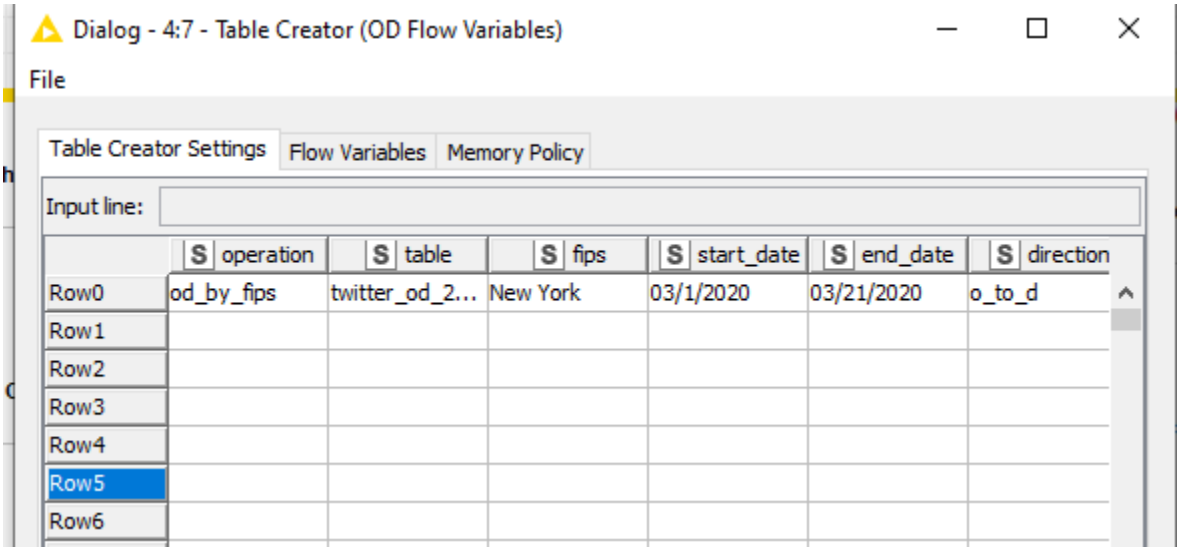
4.4 Data Access via Rest API

4.4.1 Human Mobility Data Access

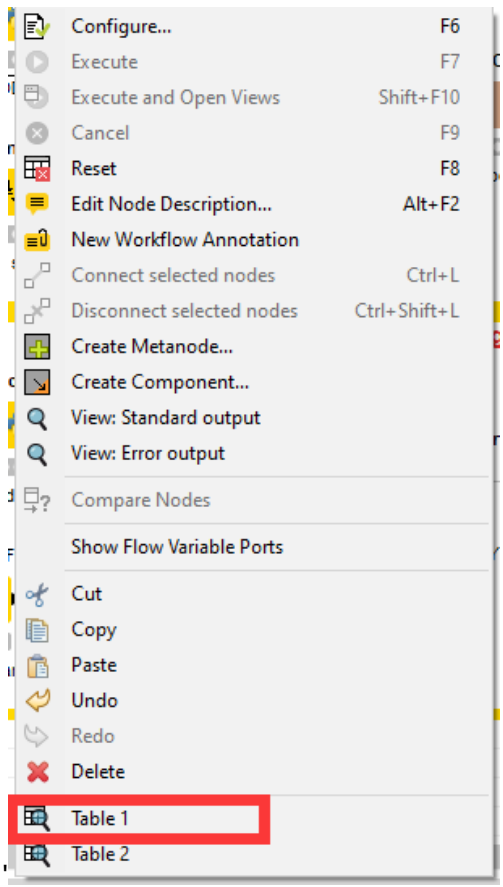
1) Import workflow from public case study folder **Y:\CDI\Case Study\00_Example\02_Rest_API\geotweets_test.knwf**

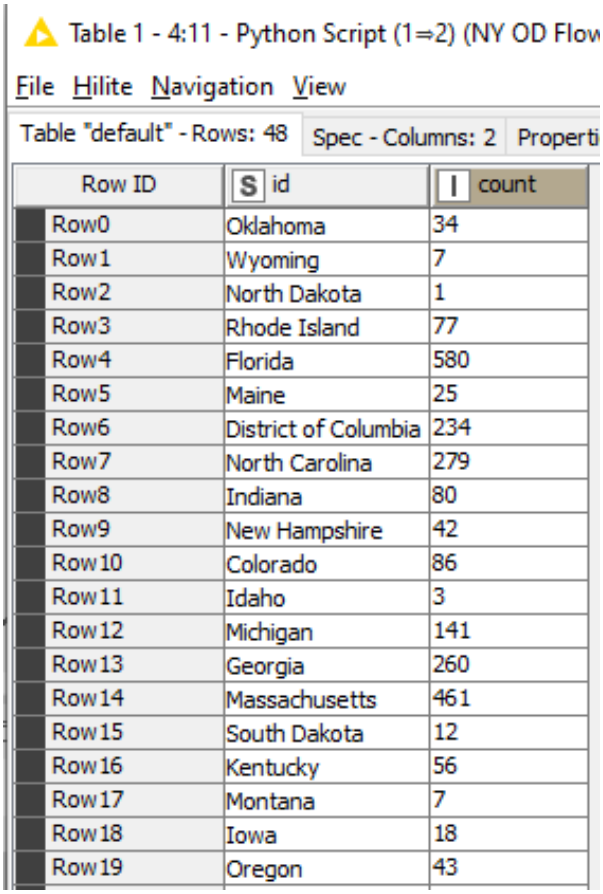


2) In the 1st node ‘Table Creator’, fill out values for each variable. Detailed description for the variables can be found at shared google doc.



3) Right click the node ‘Python Script’ and click Table 1. The data will be shown as below.

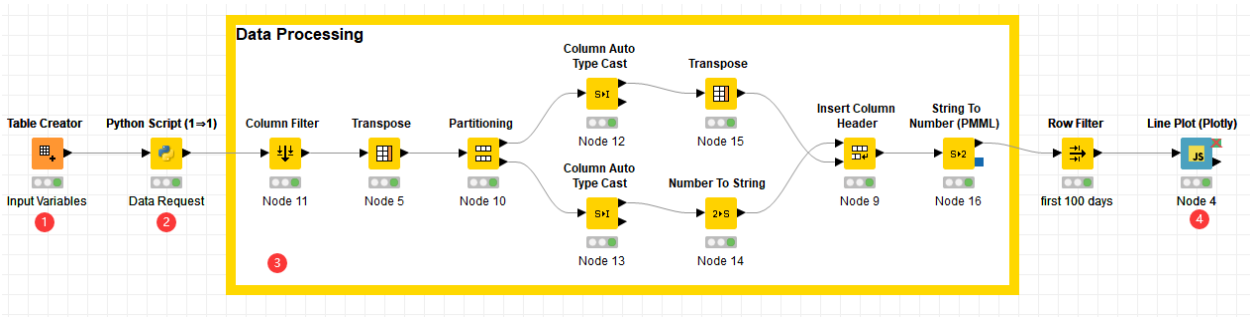




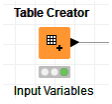
Row ID	id	count
Row0	Oklahoma	34
Row1	Wyoming	7
Row2	North Dakota	1
Row3	Rhode Island	77
Row4	Florida	580
Row5	Maine	25
Row6	District of Columbia	234
Row7	North Carolina	279
Row8	Indiana	80
Row9	New Hampshire	42
Row10	Colorado	86
Row11	Idaho	3
Row12	Michigan	141
Row13	Georgia	260
Row14	Massachusetts	461
Row15	South Dakota	12
Row16	Kentucky	56
Row17	Montana	7
Row18	Iowa	18
Row19	Oregon	43

4.4.2 GitHub Data Access

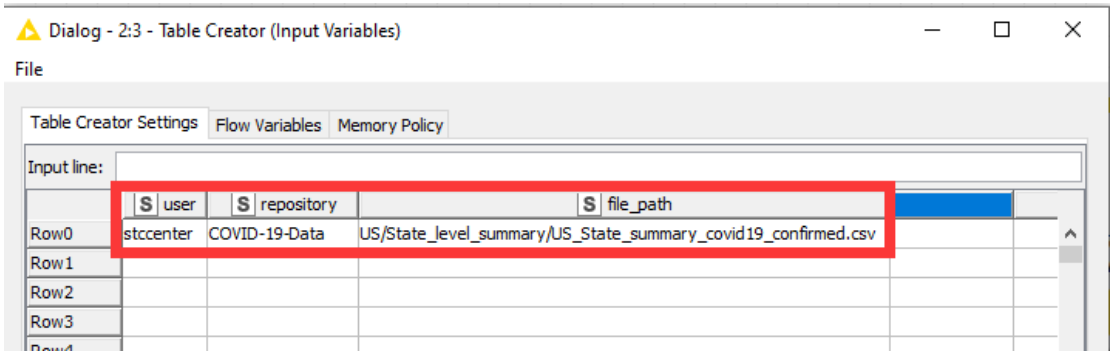
GitHub, is a provider of Internet hosting for software development and version control using Git. It offers the distributed version control and source code/data management functionality of Git. There are many open COVID-19 datasets shared on GitHub, such as STCCenter COVID-19 data repository. Thus, we build a workflow to make it easy to access the datasets in GitHub. The figure below demonstrates the workflow which requests data from STCCenter's Github repository and visualizes the data by line plot.

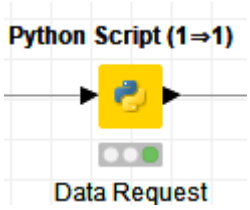


1) Import workflow from public case study folder Y:\CDI\Case Study\00_Example\02_Rest_API\github_test.knwf .

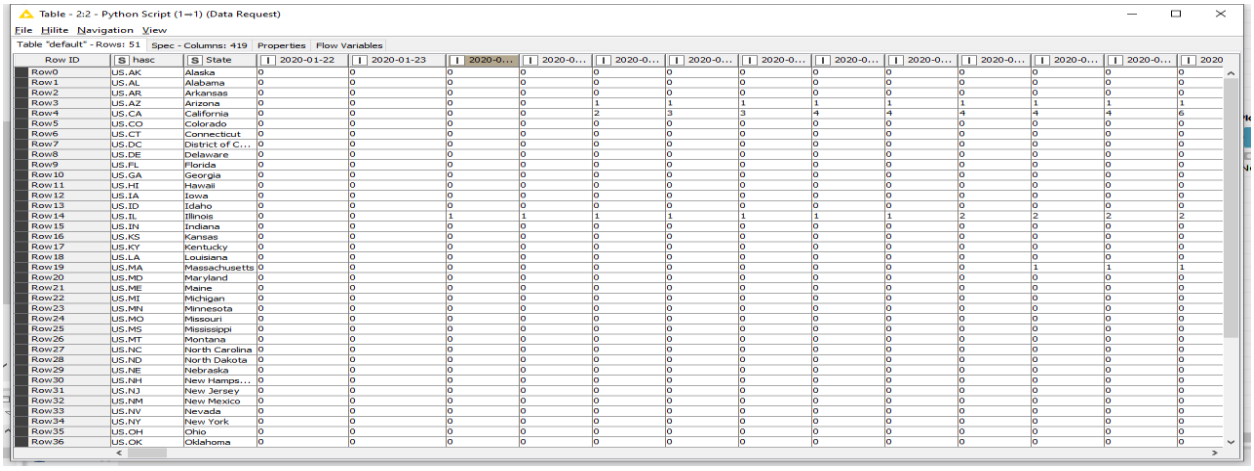
2) Set variables in the first node 

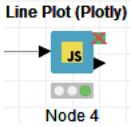
    a) User: GitHub username. The default value is **stccenter**.
    b) Repository: repository name is the user's github. The default value is COVID-19-Data, which is published by stccenter.
    c) File_path: the path in the COVID-19-Data repository. The default value is US/State_level_summary/US_State_summary_covid19_confirmed.csv



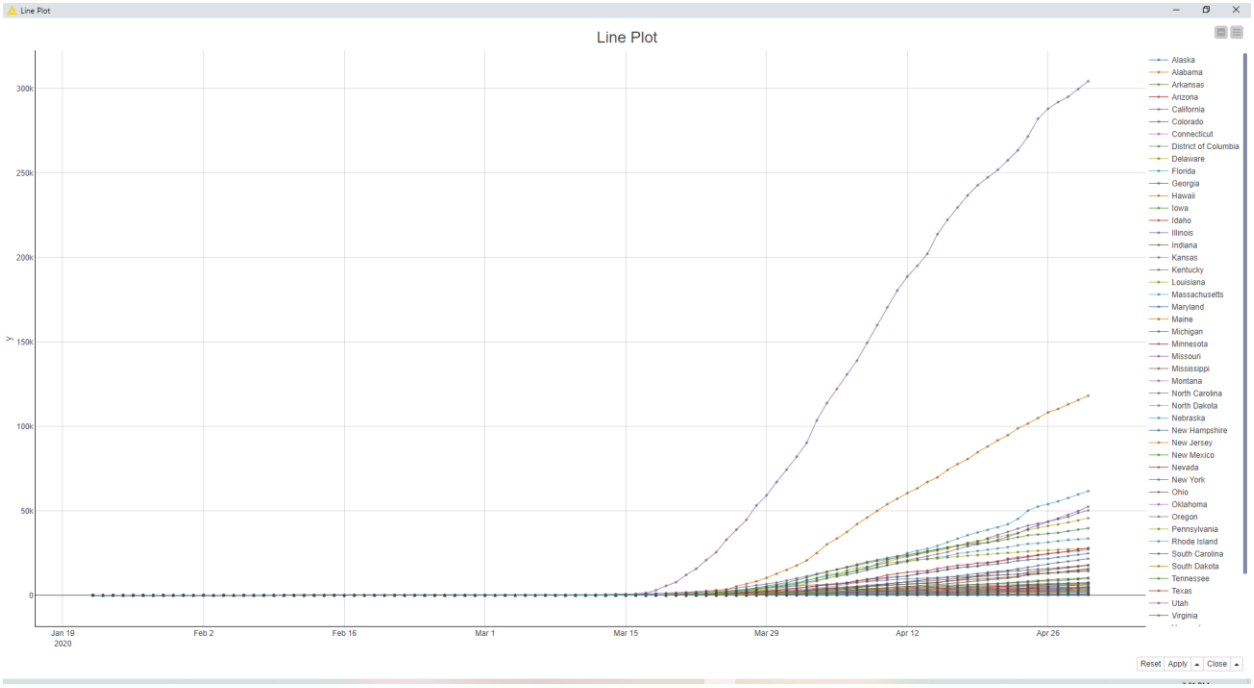


3) Check the output in the 2nd node Data Request by right click and select table. The output from the table is shown as below



4) Click run button  in the toolbar and check the visualization from Node 4 

. The figure below displays the visualization result.



1. Appendix

[1] Spatial Data Lab Data List.
https://docs.google.com/spreadsheets/d/1RO-hy__7gm2bSEKeoWj9MrKTJ8RDlVyWCvZeBag6FVI/edit#gid=0
[2] STCCenter COVID-19 Data on GitHub.
https://github.com/stccenter/COVID-19-Data
[3] Harvard Dataverse COVID-19 data sources.
https://dataverse.harvard.edu/dataverse/2019ncov
[4] Human movement ODT (Origin-Destination-Time) Flow Explorer.
http://gis.cas.sc.edu/GeoAnalytics/od.html