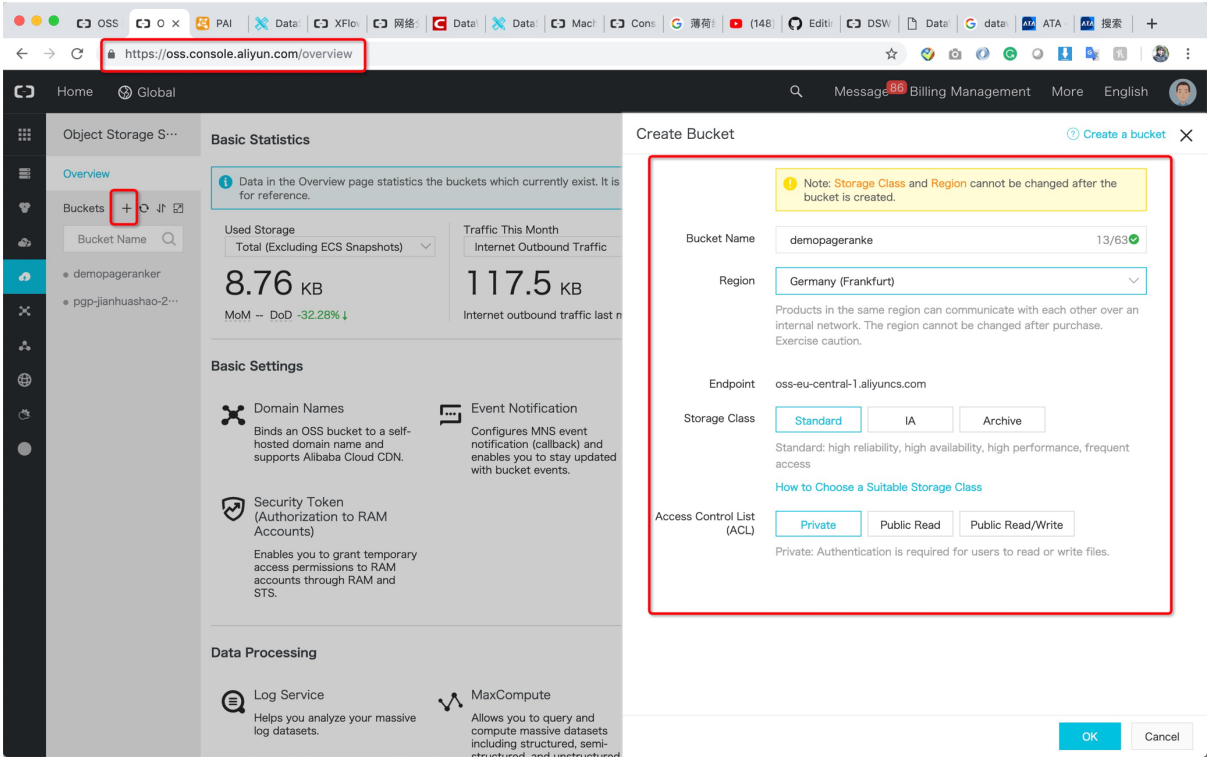
PageRank Demo step-by-step

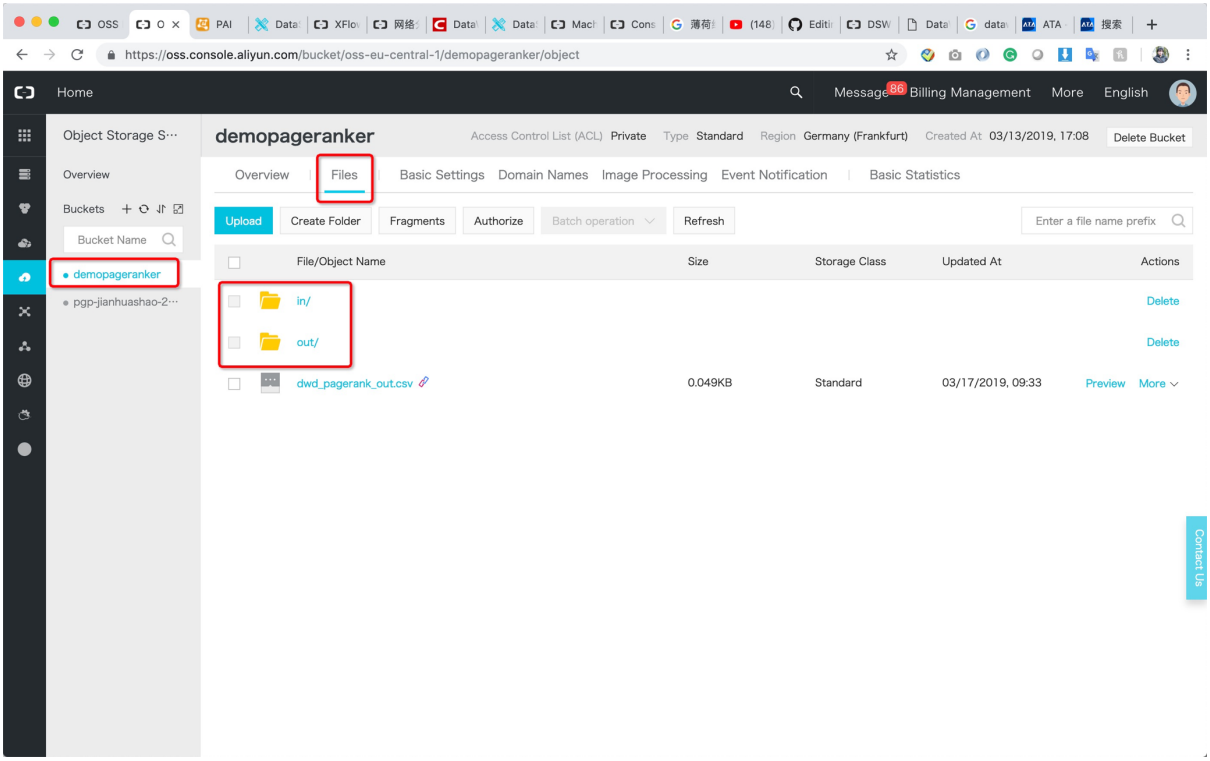
Jianhua.shao@alibaba-inc.com

Upload data into OSS (s3 equivalent)
https://oss.console.aliyun.com/overview

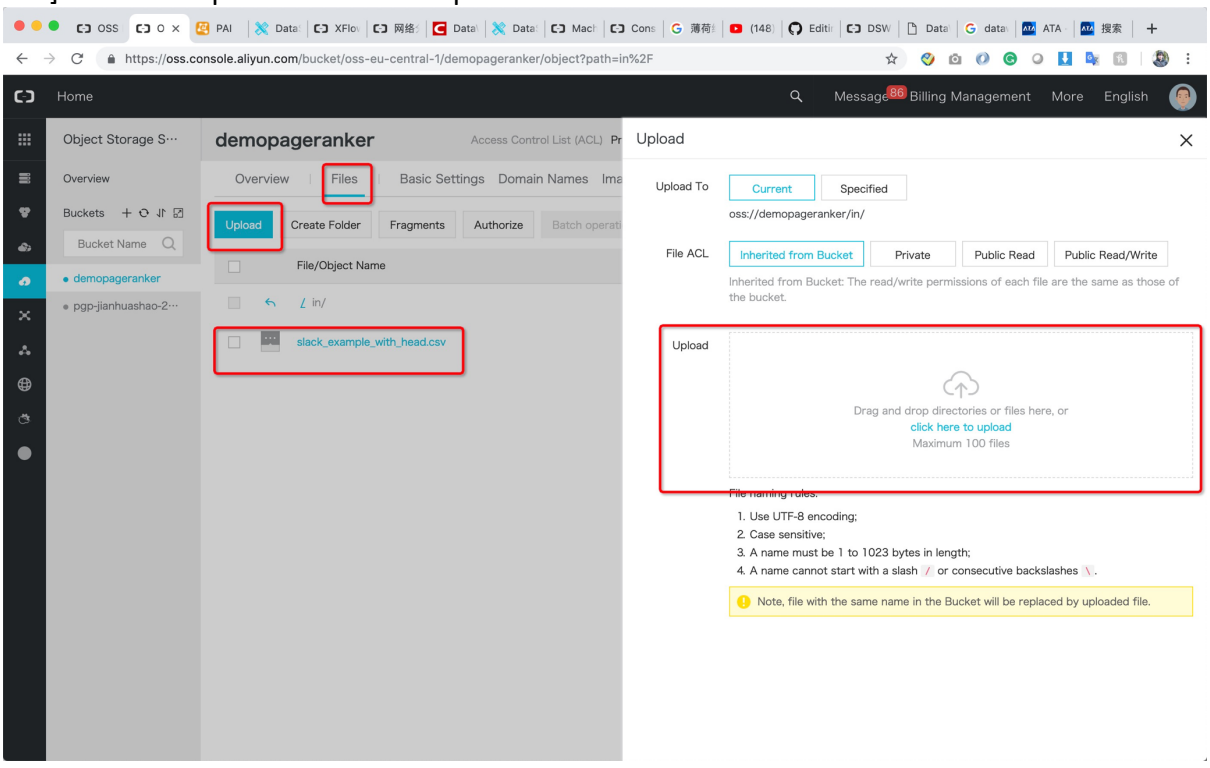
create a new oss bucket called “demopageranker”.



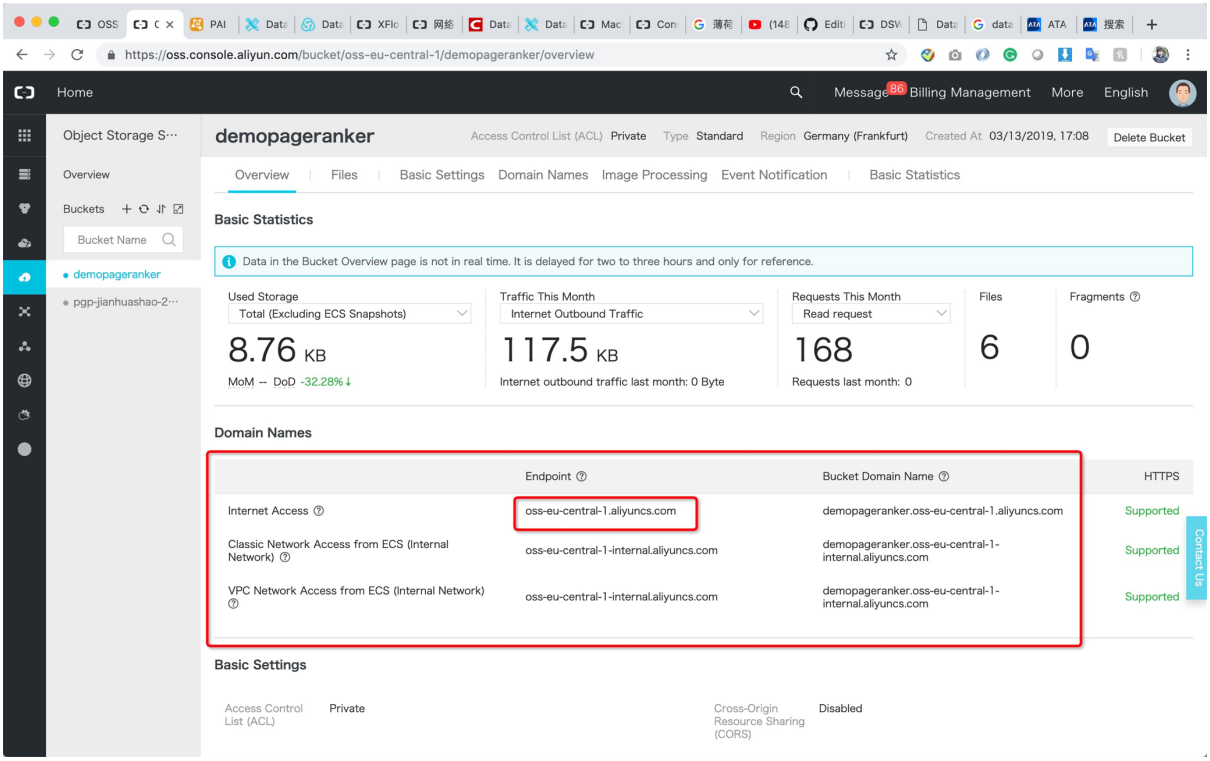
Create a ‘in’ folder to keep raw input data, and a ‘out’ folder for result.



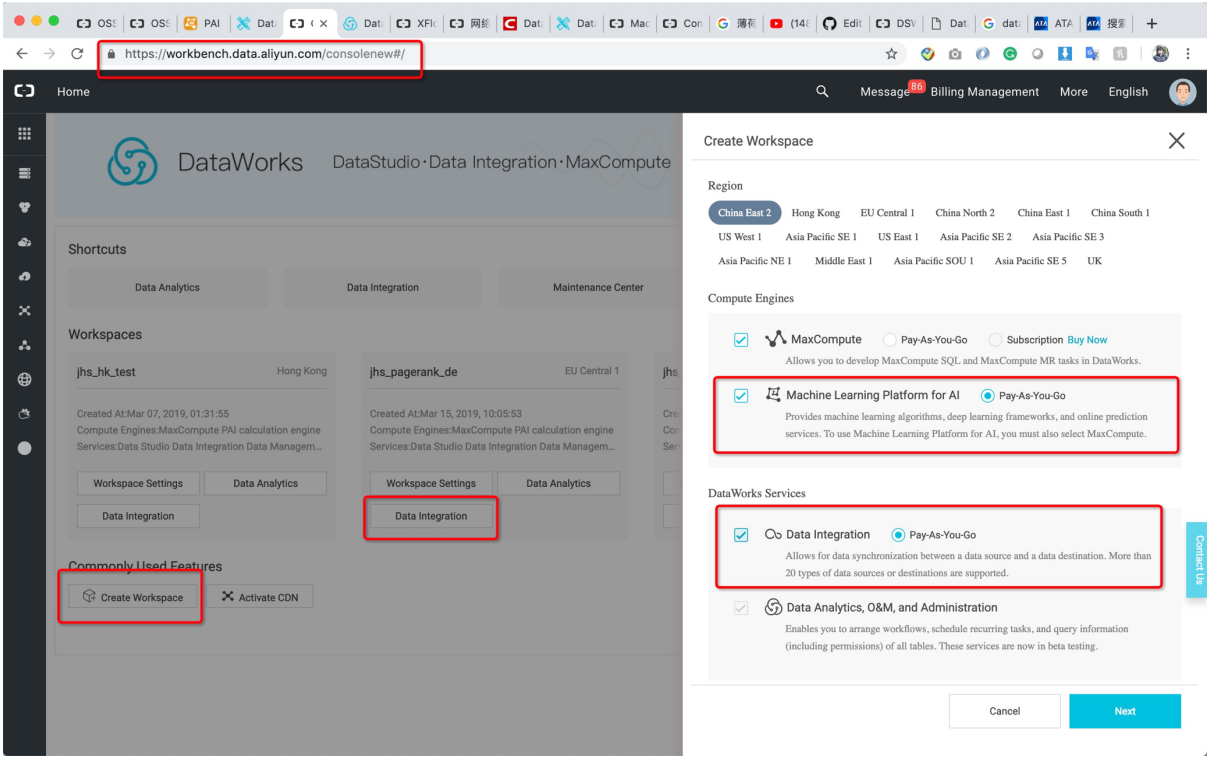
Change filename locally into “slack_example_with_head.csv” [or any other file name you like]. Click on “upload” button to upload the file.



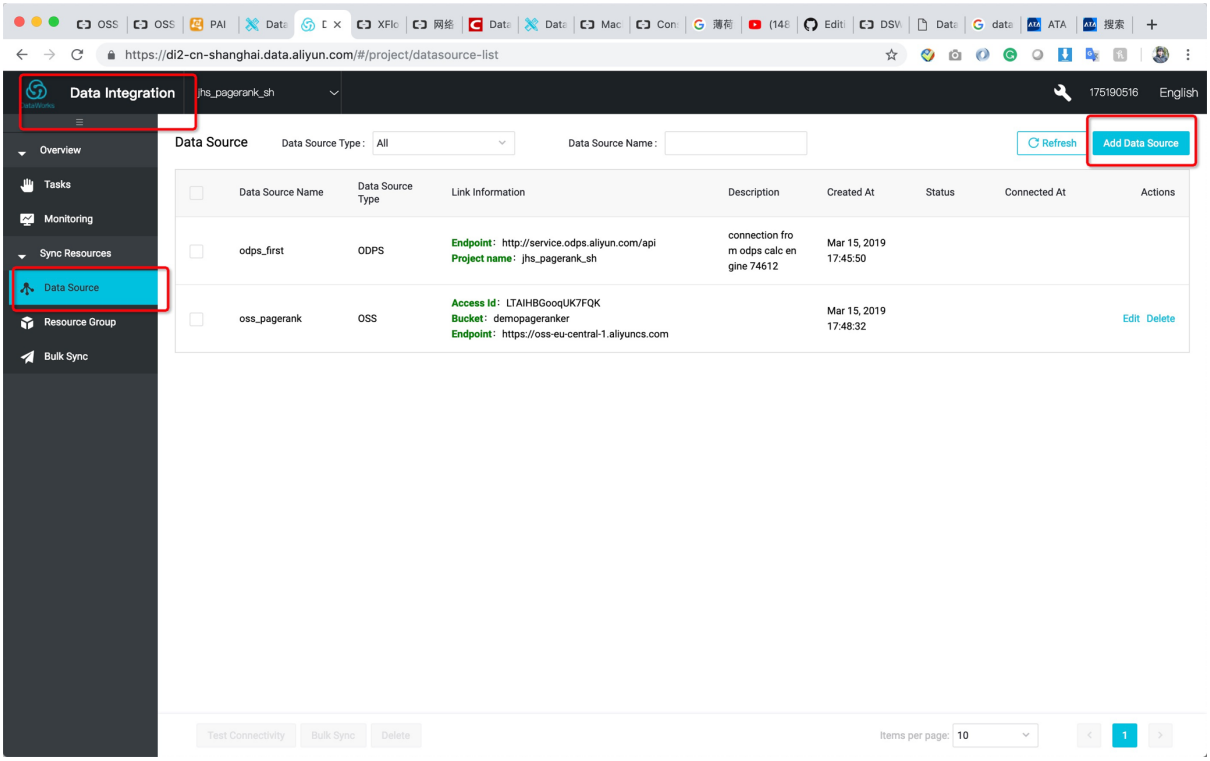
Note down endpoint connection configuration information.

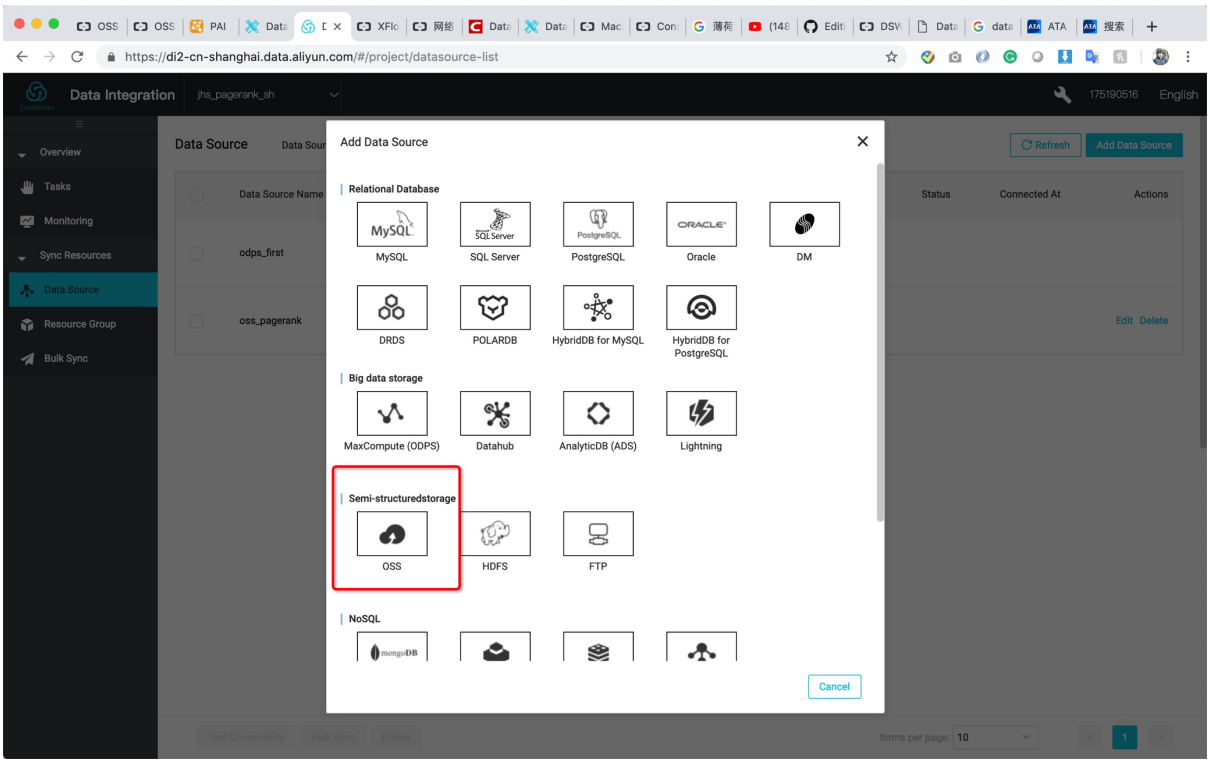


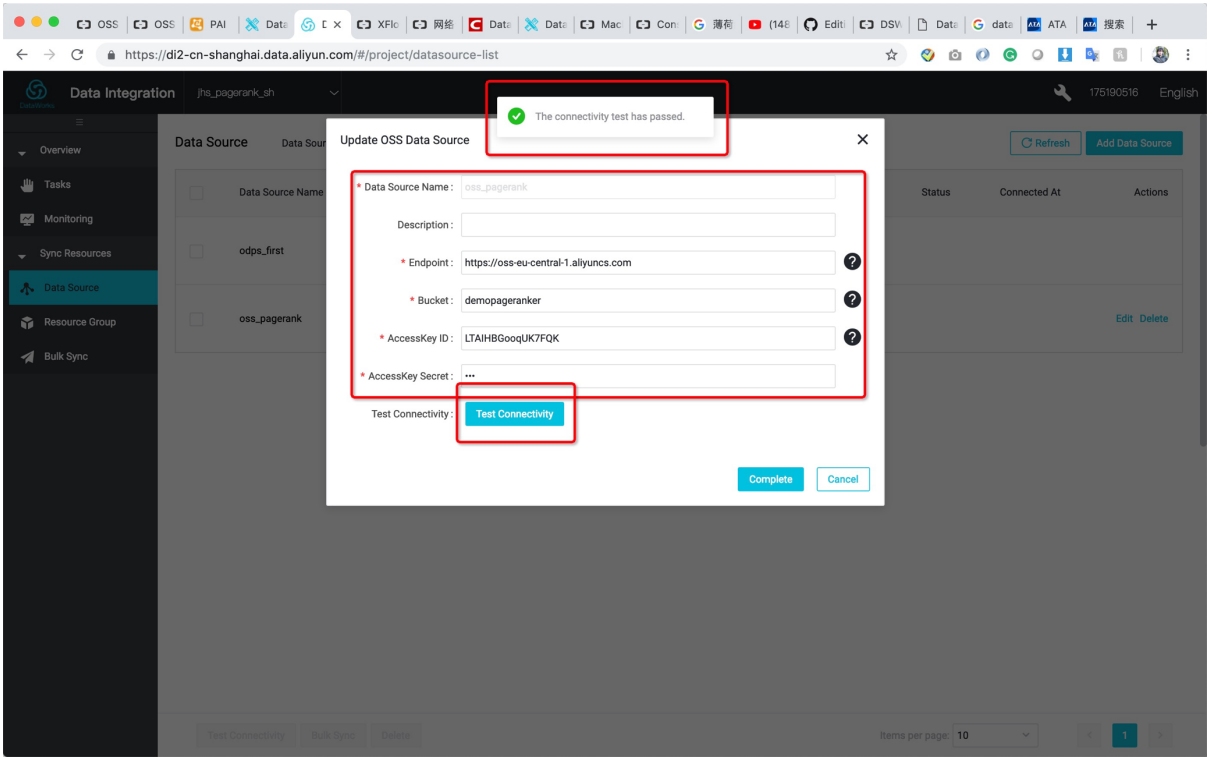
https://workbench.data.aliyun.com/consolenew#/
go to datawork space and create a workspace if you have not. For example, I created workspace called “jhs_pagerank_sh”. Make sure you select to enable PAI and Data Integration.



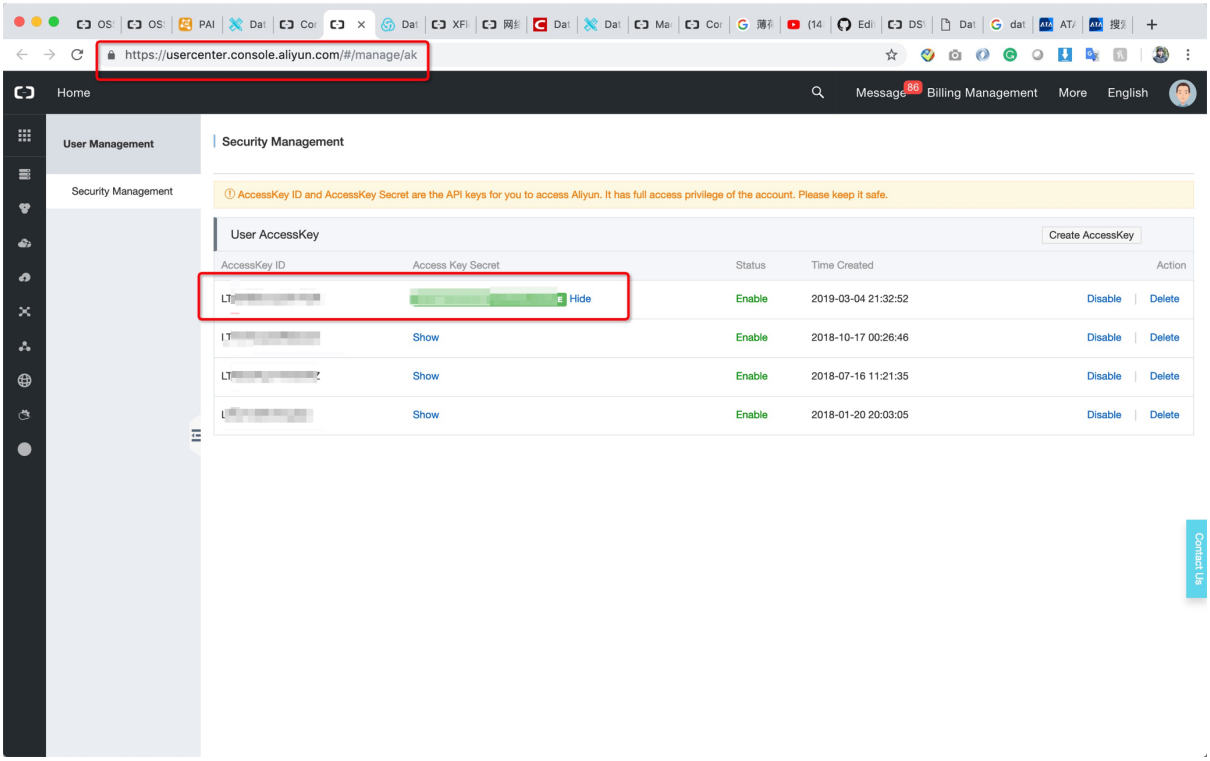
Click on “data integration” button in the workspace and then connect dataworks to oss:


https://di2-cn-shanghai.data.aliyun.com/#/project/datasource-list
Data Integration
jhs_pagerank_sh
175190516
English
Overview
Tasks
Monitoring
Sync Resources
Data Source
Resource Group
Bulk Sync
Data Source
Data Source Type: All
Data Source Name:
Refresh
Add Data Source
Data Source Name
Data Source Type
Link Information
Description
Created At
Status
Connected At
Actions
odps_first
ODPS
Endpoint: http://service.odps.aliyun.com/api
Project name: jhs_pagerank_sh
connection from odps calc engine 74612
Mar 15, 2019 17:45:50
oss_pagerank
OSS
Access Id: LTAIHBGooqUK7FQK
Bucket: demopageranker
Endpoint: https://oss-eu-central-1.aliyuncs.com
Mar 15, 2019 17:48:32
Edit
Delete
Test Connectivity
Bulk Sync
Delete
Items per page: 10
1

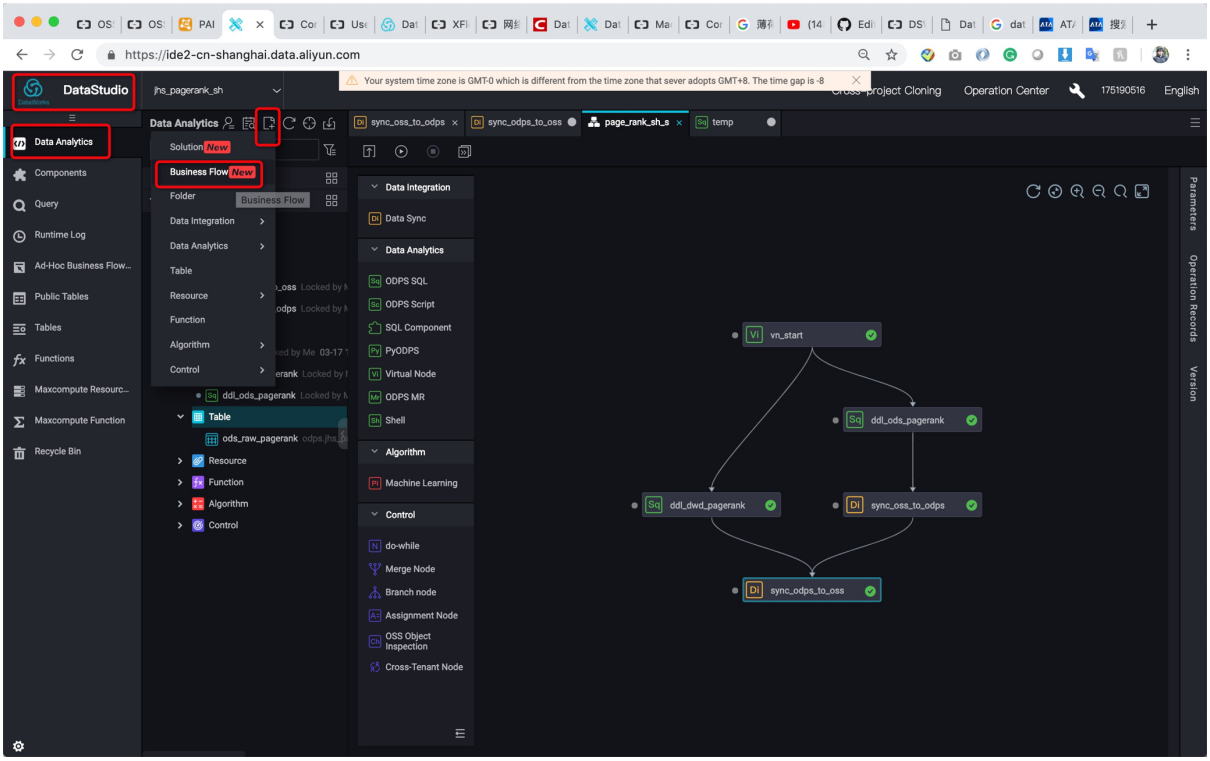

https://di2-cn-shanghai.data.aliyun.com/#/project/datasource-list
Data Integration
jhs_pagerank_sh
175190516
English
Add Data Source
Relational Database
MySQL
SQL Server
PostgreSQL
Oracle
DM
DRDS
POLARDB
HybridDB for MySQL
HybridDB for PostgreSQL
Big data storage
MaxCompute (ODPS)
Datahub
AnalyticDB (ADS)
Lightning
Semi-structuredstorage
OSS
HDFS
FTP
NoSQL
Cancel



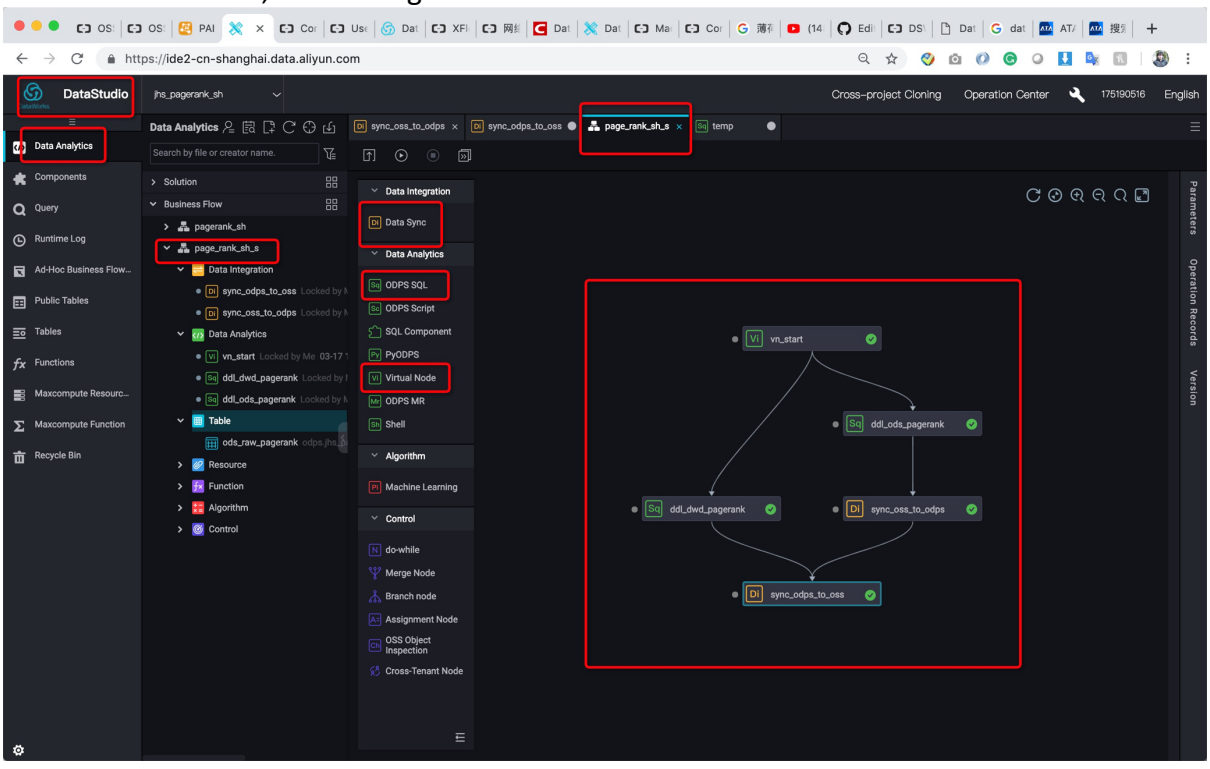
The AccessKey info can be found in : https://usercenter.console.aliyun.com/#/manage/ak



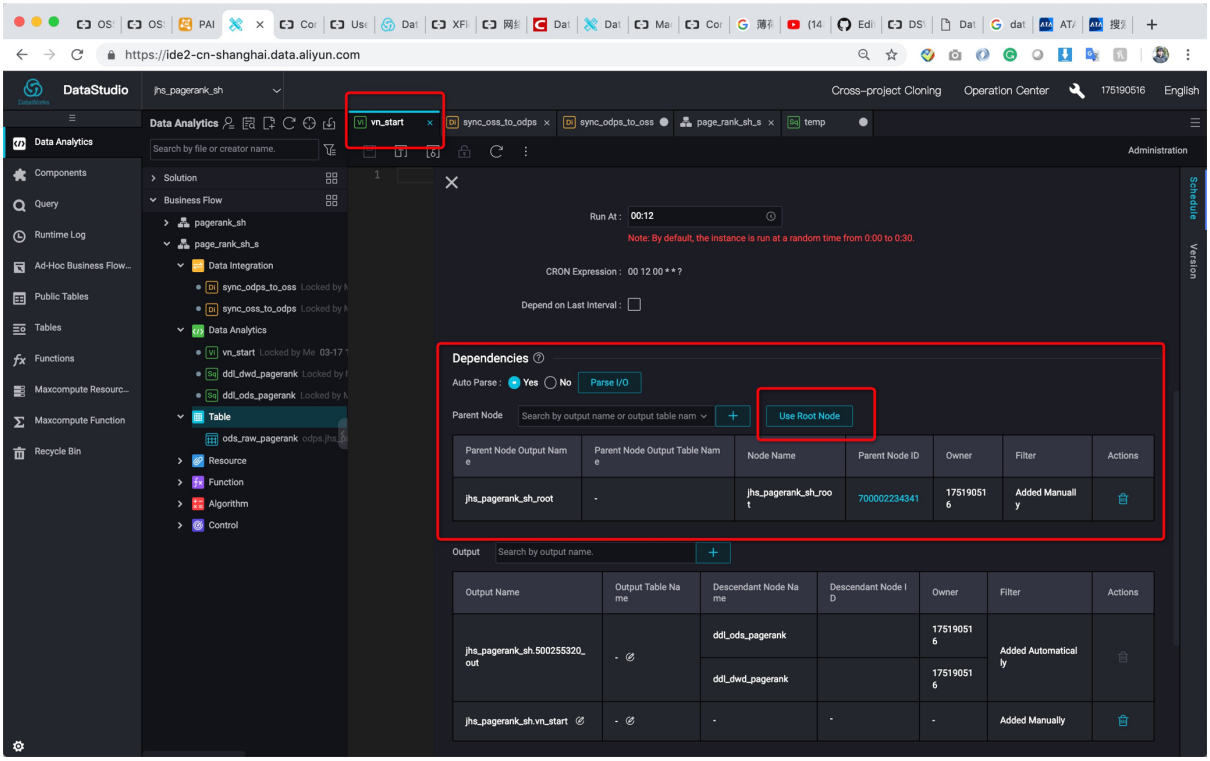
Go to “datastudio” in dataworks to create business flow, such as “page_ran_sh_s”



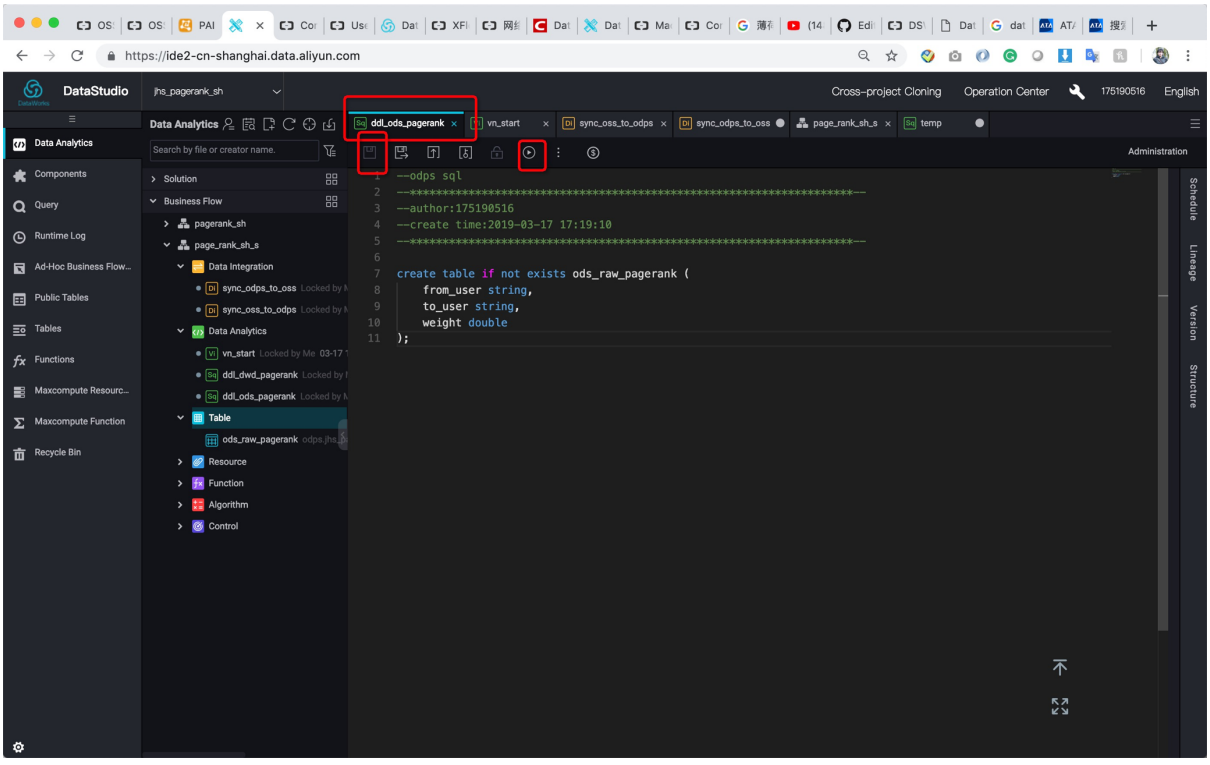
In the business flow, let us design a workflow as bellow:



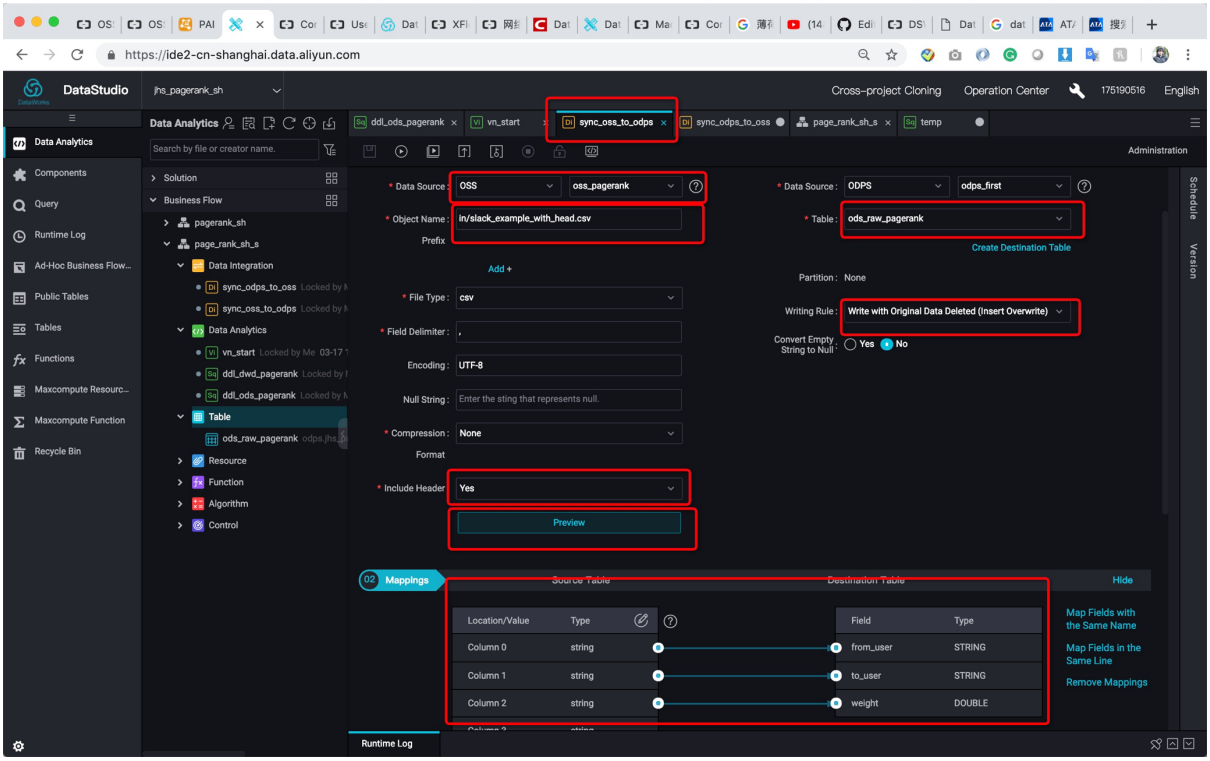
“vn_start” is a virtual node. Make sure you configure to use root node as bellow:



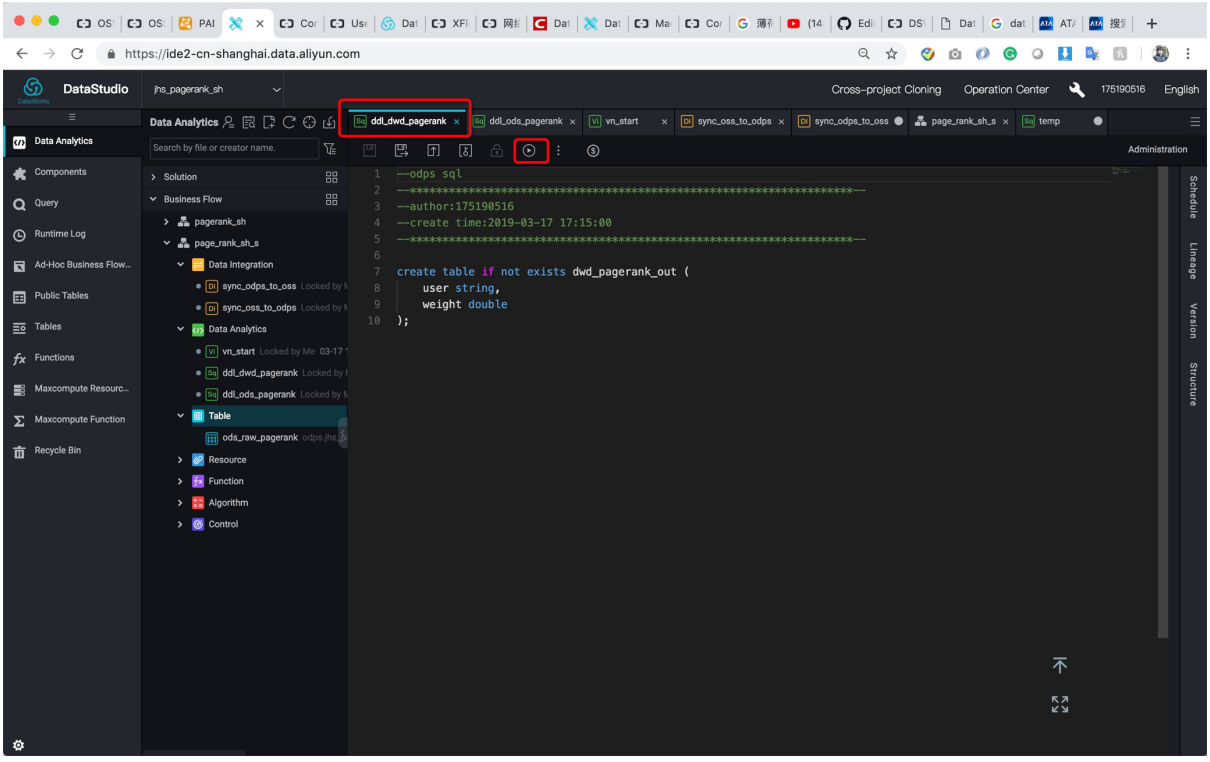
in “ddl_ods_pagerank”, copy paste bellow sql, and then click on “save” button and then click on “run” button. This SQL will ignore partition setting since we only have a very small SQL. This step will create a table called “ods_raw_pagerank” to store data ingested from oss.



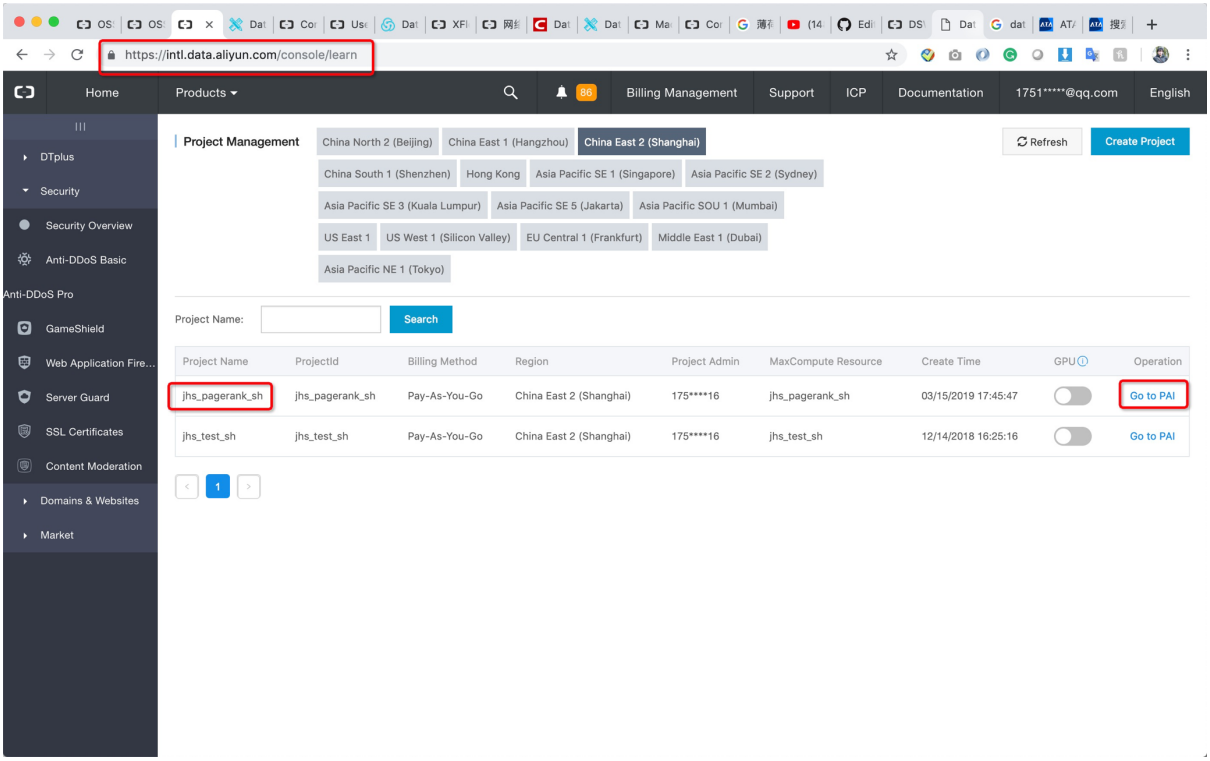
In “sync_oss_to_odps”, we configure to set up data ingestion from oss and store in odps (MaxCompute). Click “run” button to do the first ingestion.



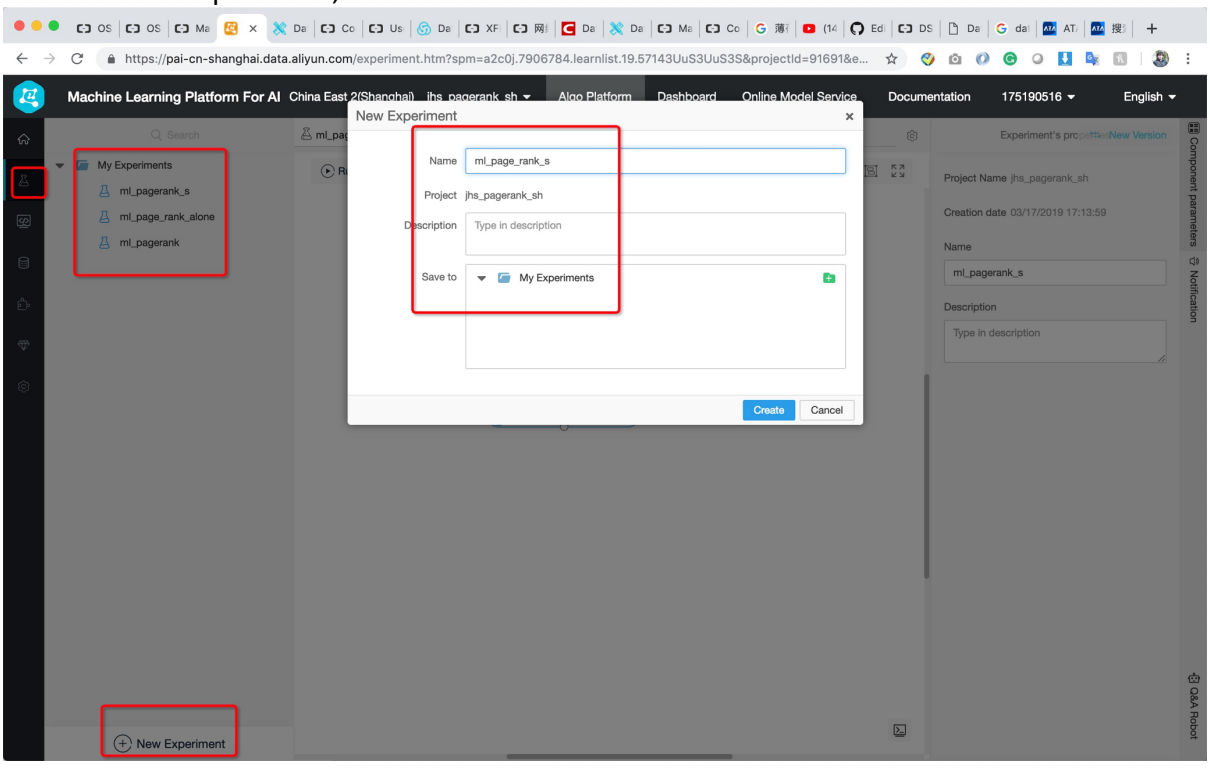
In “ddl_dwd_pagerank”, paste bellow sql to create a placeholder to store result from PAI. The table is called “dwd_pagerank_out”.



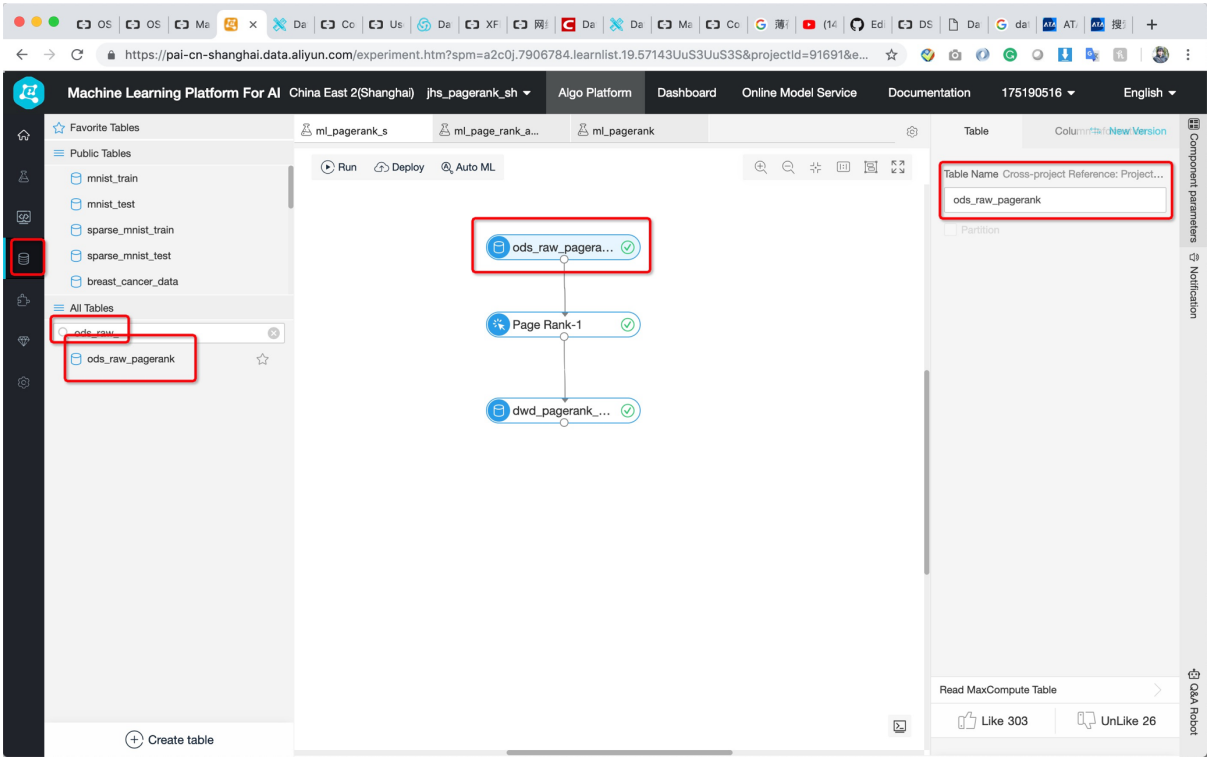
Go to PAI to enter experiment: https://intl.data.aliyun.com/console/learn



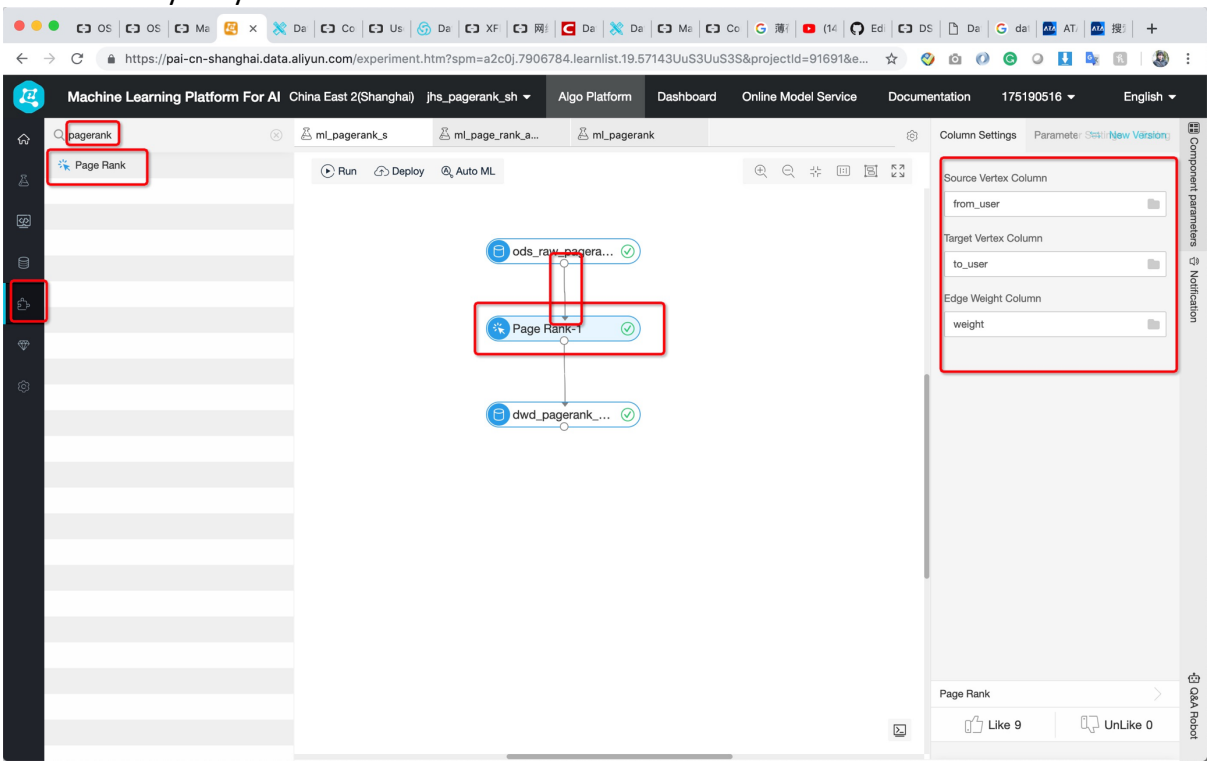
Create a new experiment, called



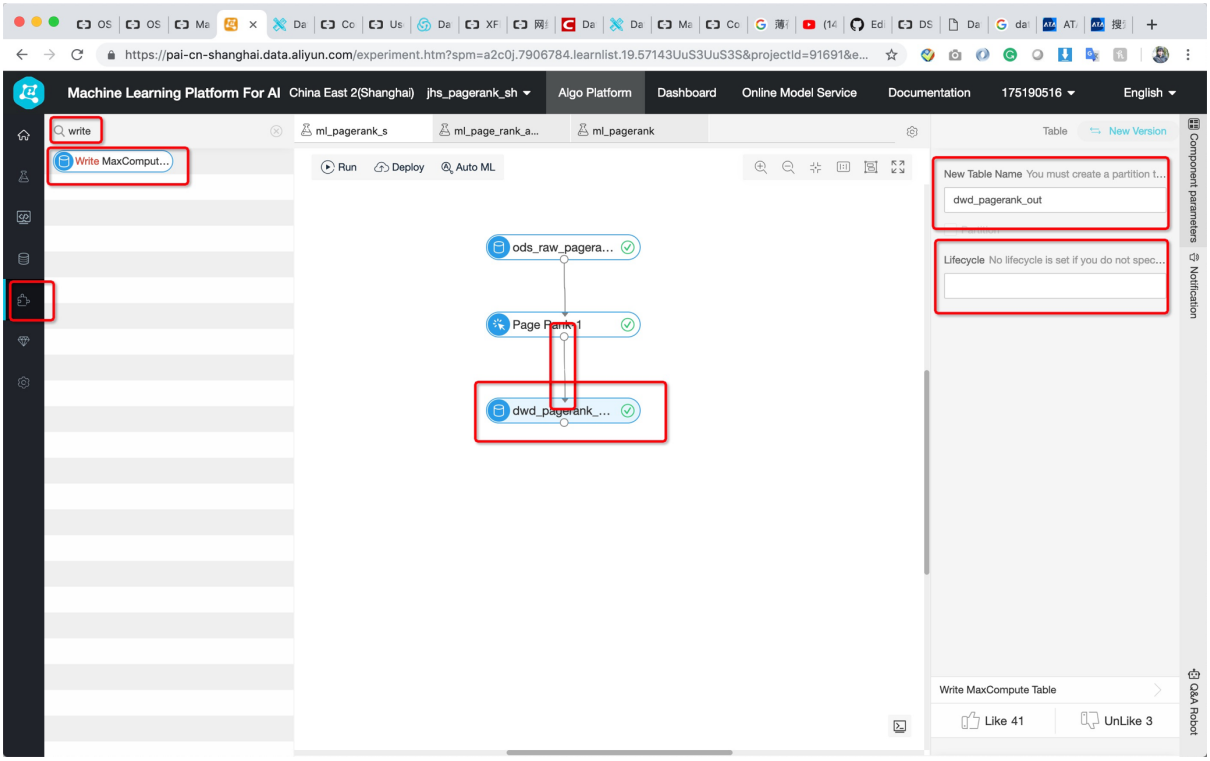
Go to “data source”, and then search for “ods_raw_pagerank” which is what we have ingested, simply just drag and drop it into the PAI canvas.



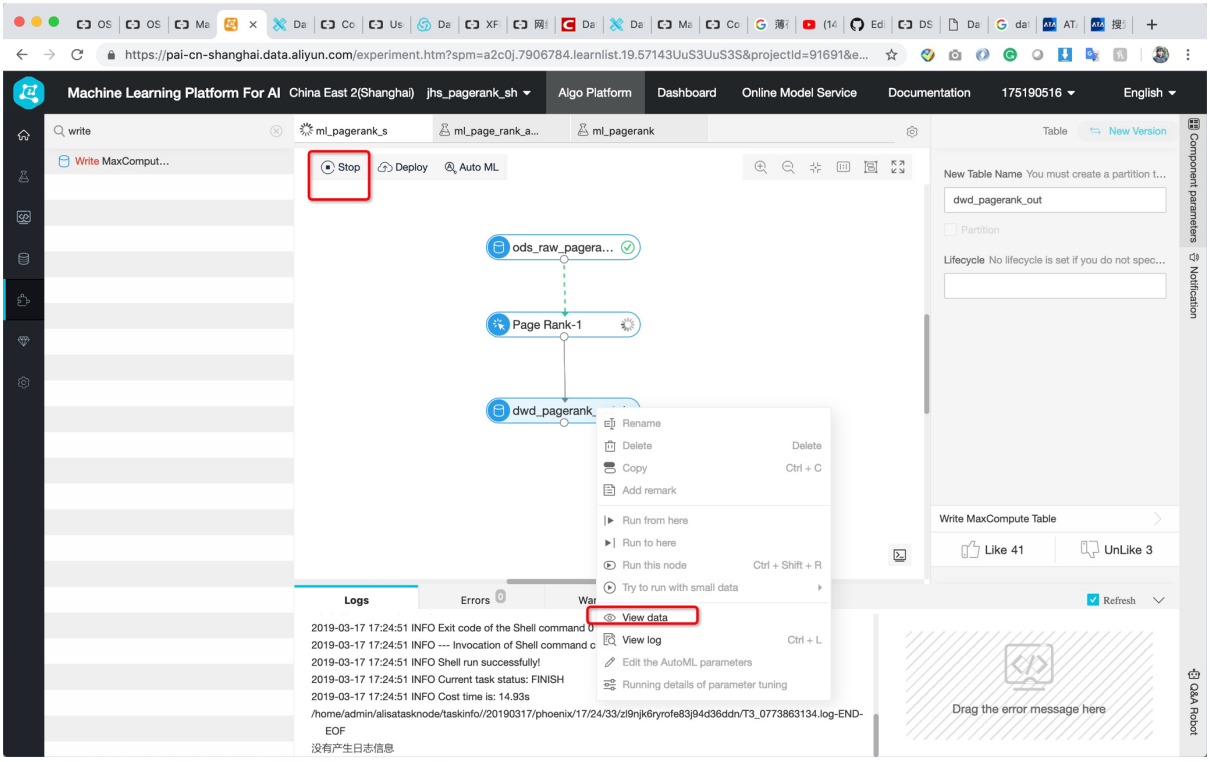
Go to “component” and then search for “pagerank”, you can simply just drag-drop into canvas, and connect “ods_raw_pagerank_xxx” component to “page rank_xxx” component, you can then configure “page rank” component, these filed should be able to pop up automatically for you to select.



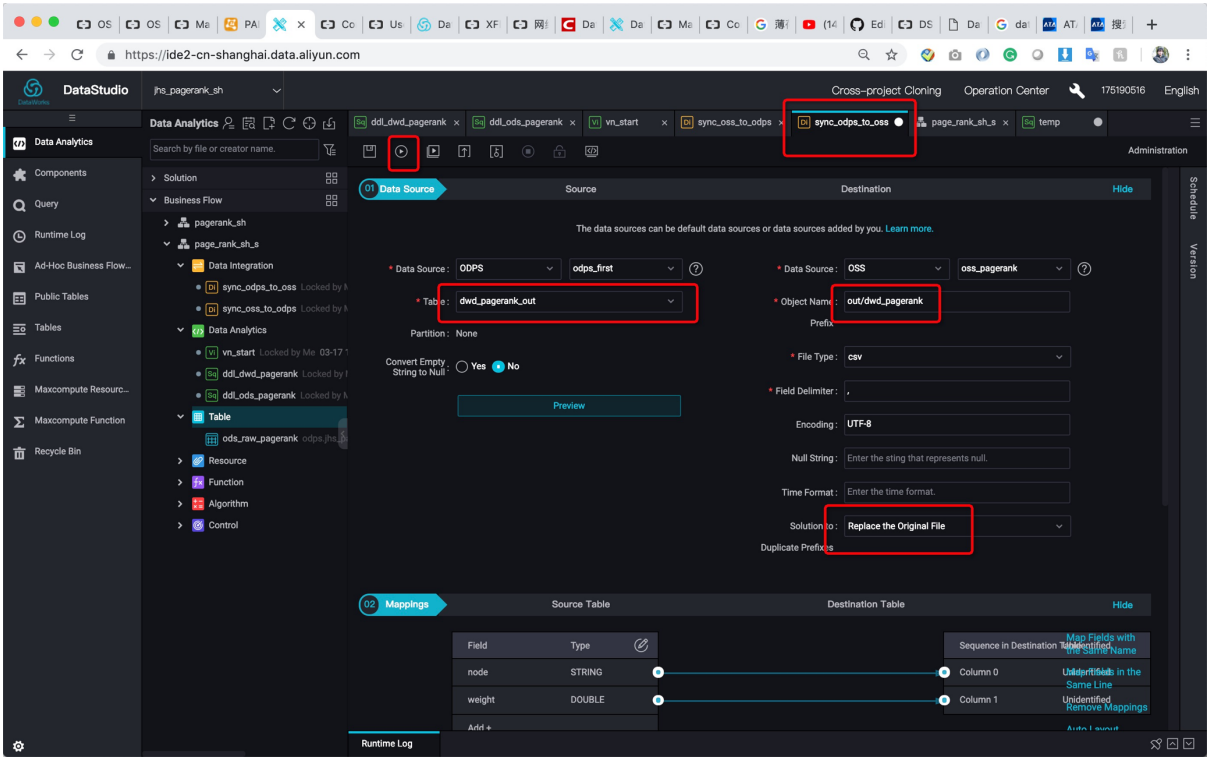
Go to “component” and then search for “write”, find “write maxcopute table” component and then drag-drop into canvas and type the table name we just created before to store result. Change lifecycle to empty to avoid table recycle.



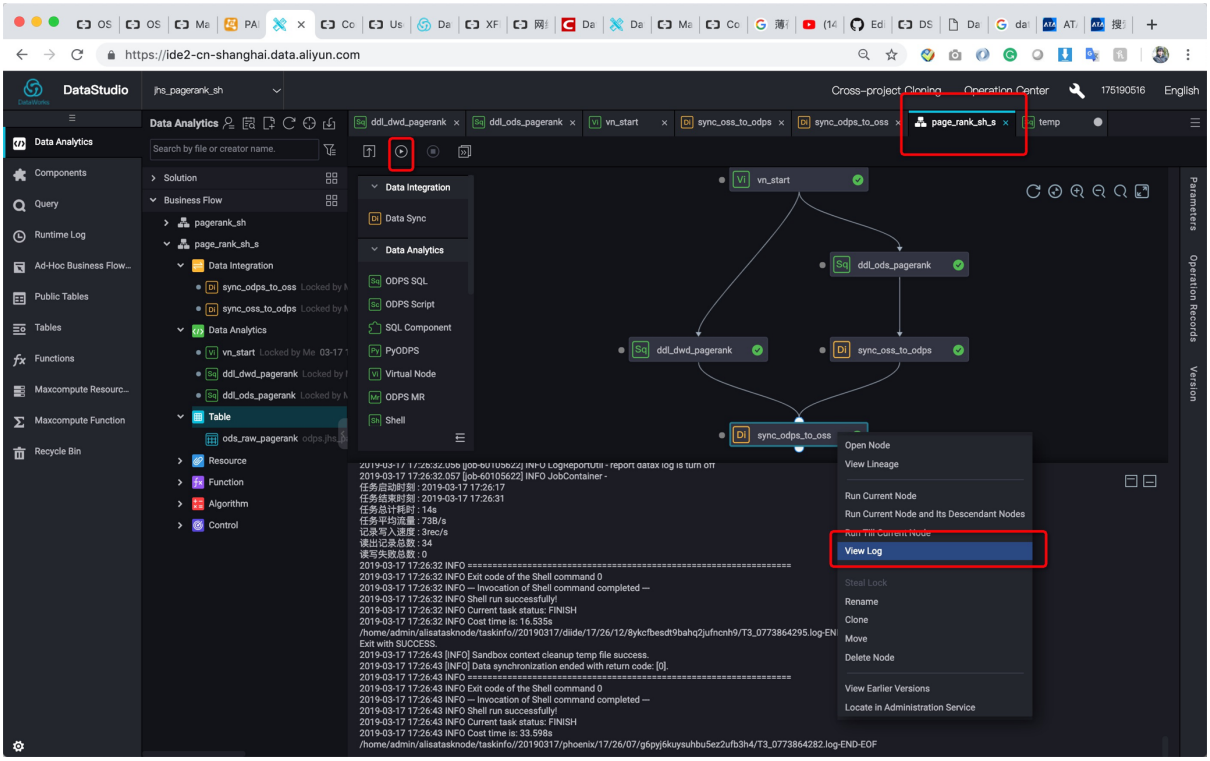
Now, you can start to “run” the pipeline. Once it finish, you can view data or log to validate if it has run.



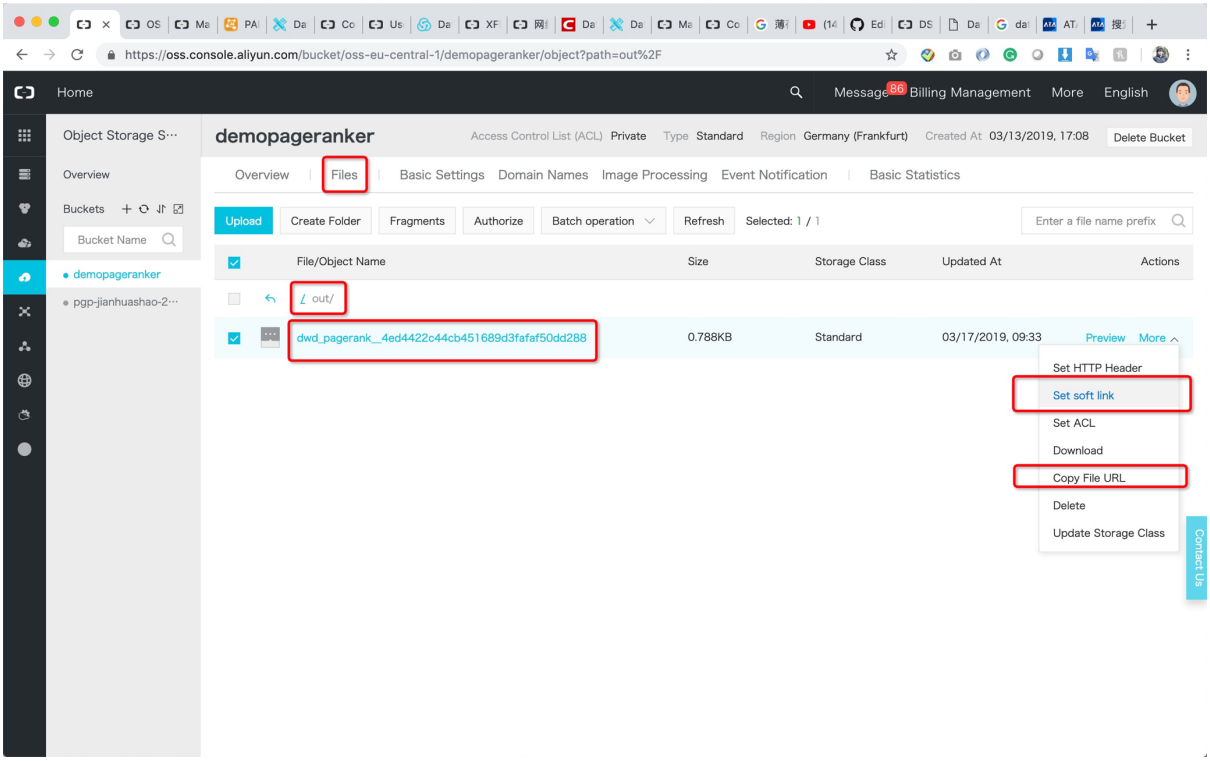
Go back to dataworks, and do last step to sync data into oss for further consumption. And then run it.



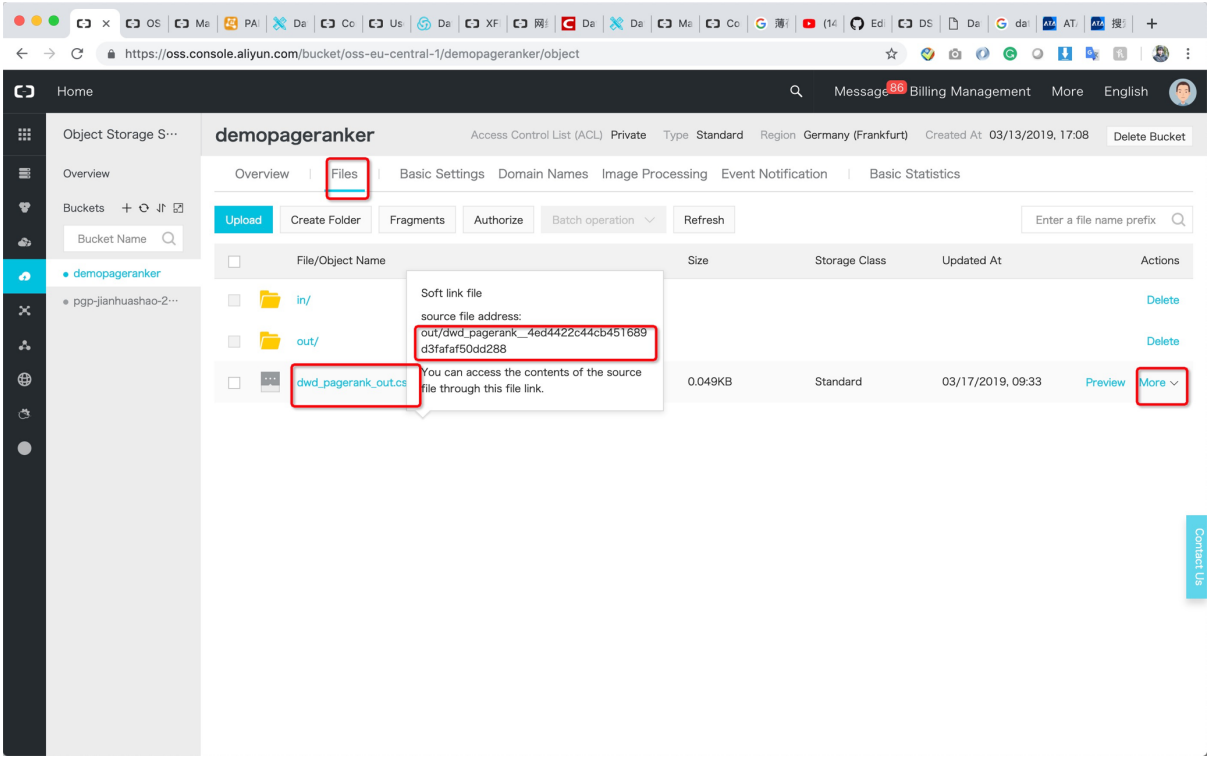
You can do back to the canvas. Run all process in one go.



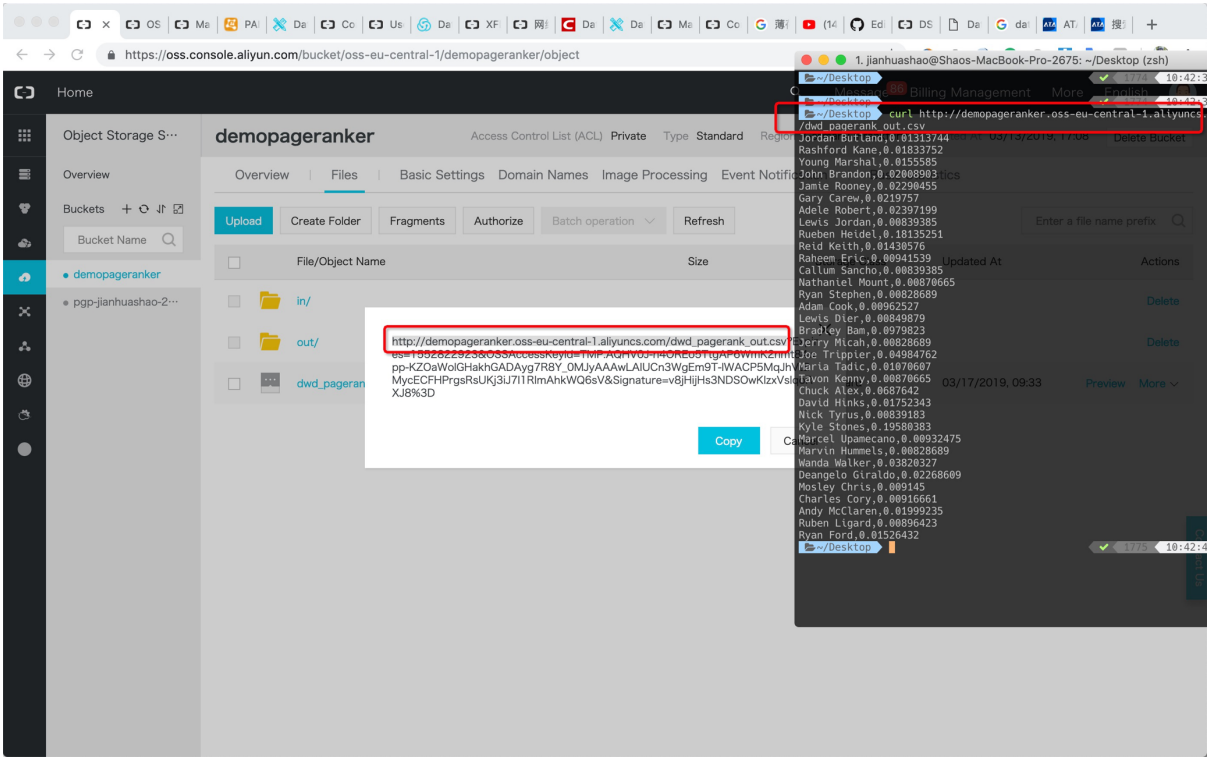
You can see the result file appear in /out/ folder. You can also setup “soft link” to have a customised file name.



For example, I have set a soft link as “dwd_pagerank_out.csv”.



You should be able to access to the result in your terminal or with http request.



The benefit of using dataworks is that once you setup the workflow, you can result it when you change the input file , etc.

The demo result file can be access in: https://singularitynet.iohttp://demopageranker.oss-eu-central-1.aliyuncs.com/dwd_pagerank_out.csv