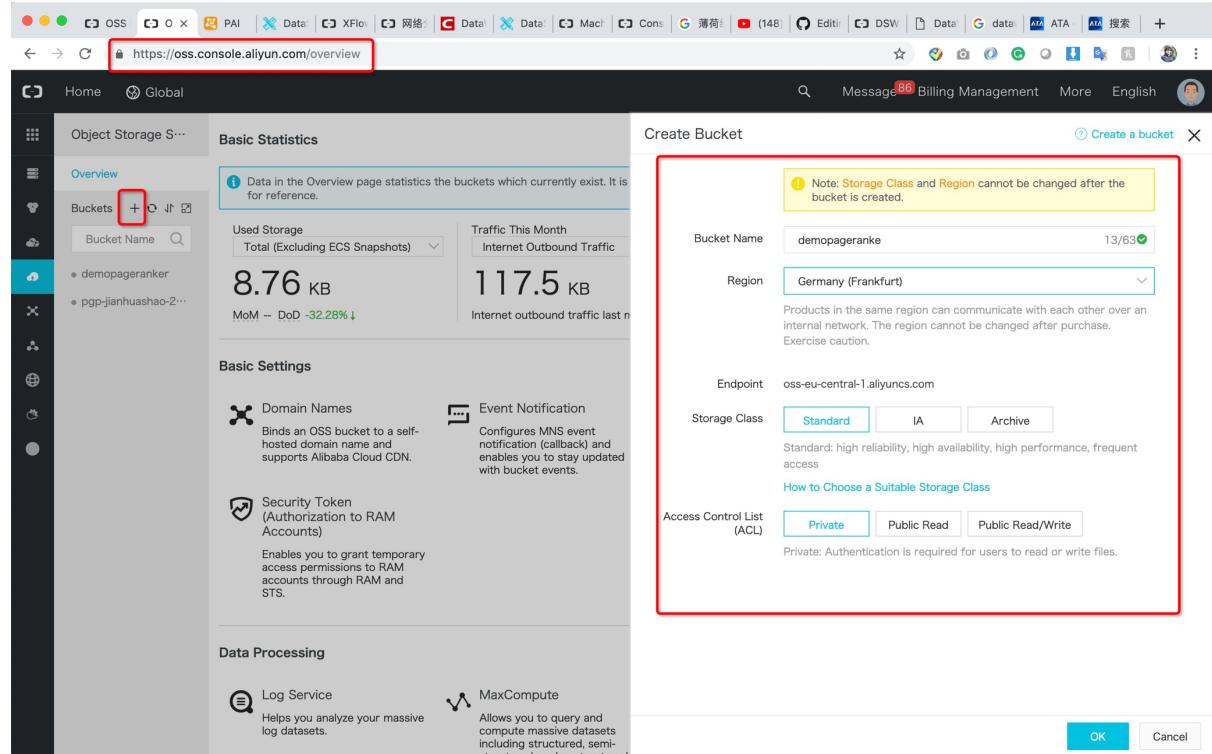


PageRank Demo step-by-step

Jianhua.shao@alibaba-inc.com

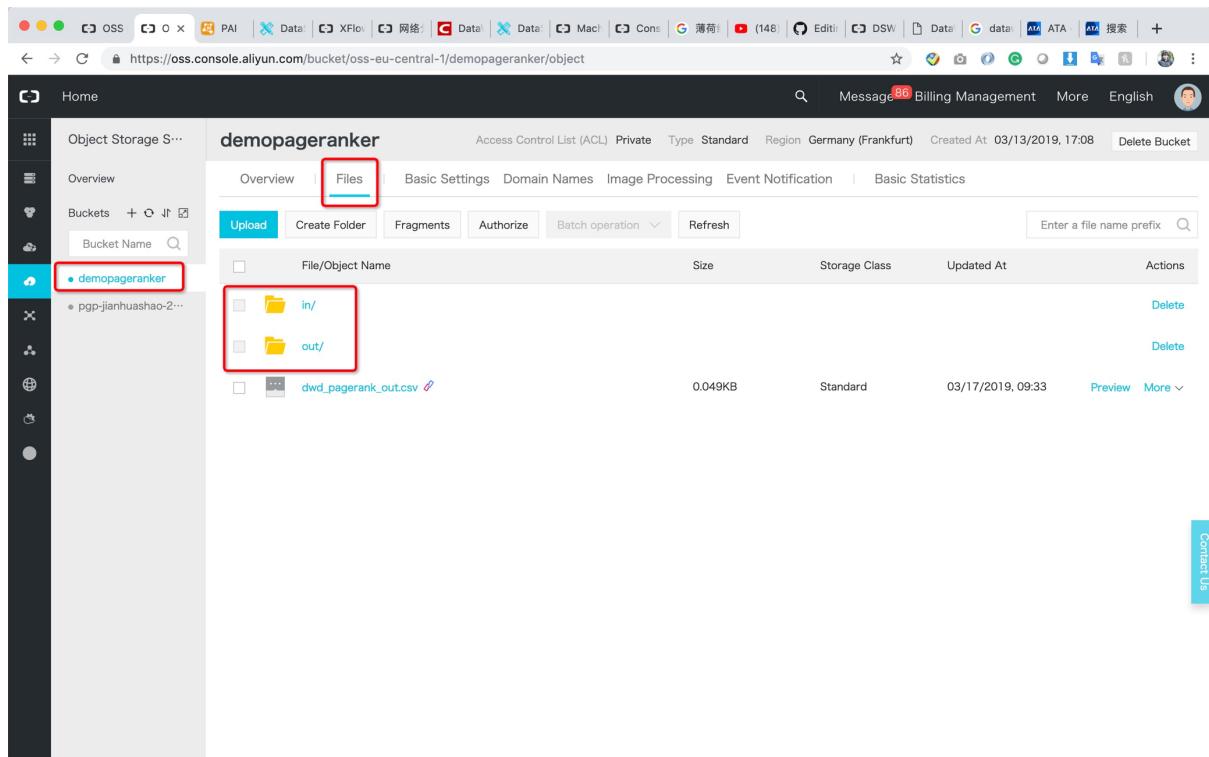
Upload data into OSS (s3 equivalent)
<https://oss.console.aliyun.com/overview>

create a new oss bucket called “demopageranker”.

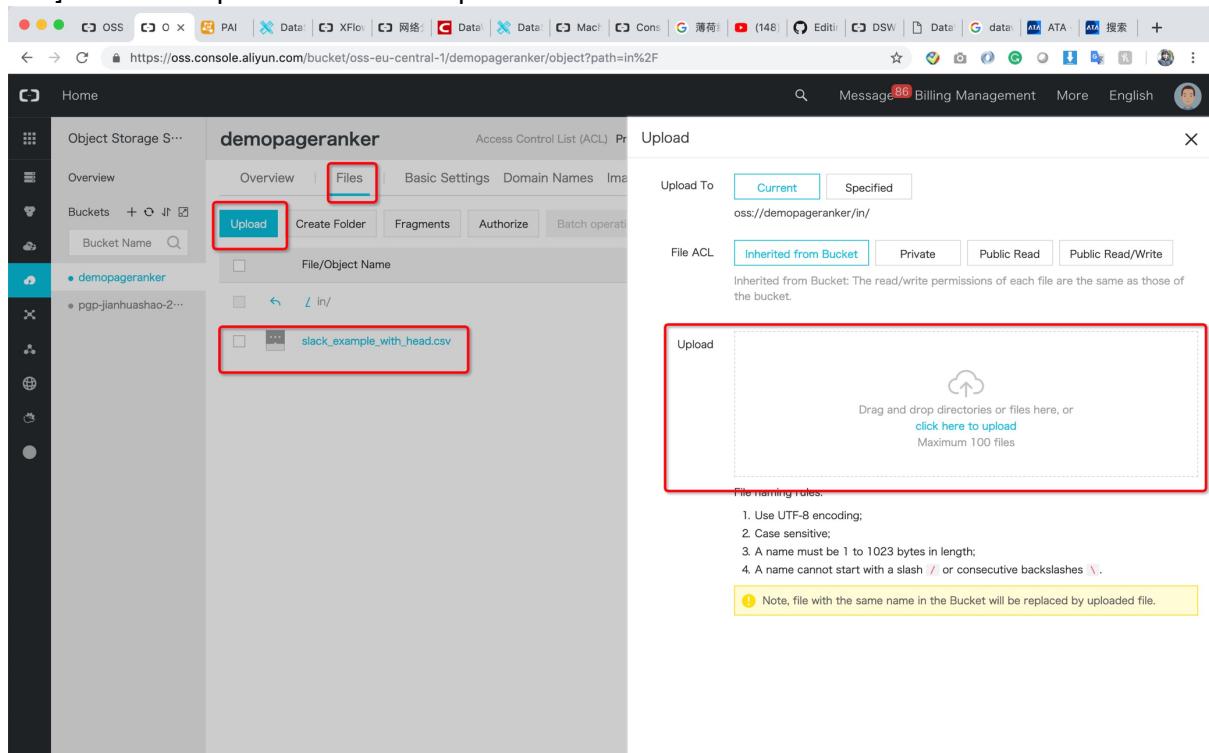


The screenshot shows the AliCloud OSS console interface. On the left, there's a sidebar with various icons. In the center, there's a 'Basic Statistics' section showing 'Used Storage' (8.76 KB) and 'Traffic This Month' (117.5 KB). Below this is a 'Create Bucket' dialog box, which is highlighted with a red border. The dialog contains fields for 'Bucket Name' (set to 'demopageranke'), 'Region' (set to 'Germany (Frankfurt)'), 'Endpoint' (set to 'oss-eu-central-1.aliyuncs.com'), 'Storage Class' (set to 'Standard'), and 'Access Control List (ACL)' (set to 'Private'). There are also tabs for 'IA' and 'Archive'. At the bottom right of the dialog are 'OK' and 'Cancel' buttons.

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://oss.console.aliyun.com/bucket/oss-eu-central-1/demopageranker/overview>

The screenshot shows the Aliyun OSS console interface. On the left, there's a sidebar with various icons. The main area displays the 'demopageranker' bucket details. Under 'Basic Statistics', it shows used storage (8.76 KB), traffic (117.5 KB), requests (168), files (6), and fragments (0). The 'Domain Names' section is highlighted with a red box, listing three access methods: Internet Access, Classic Network Access from ECS (Internal Network), and VPC Network Access from ECS (Internal Network). Each method has its own endpoint and bucket domain name listed. The 'Basic Settings' section shows the bucket is 'Private'.

<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.

<https://workbench.data.aliyun.com/consolenew#/>

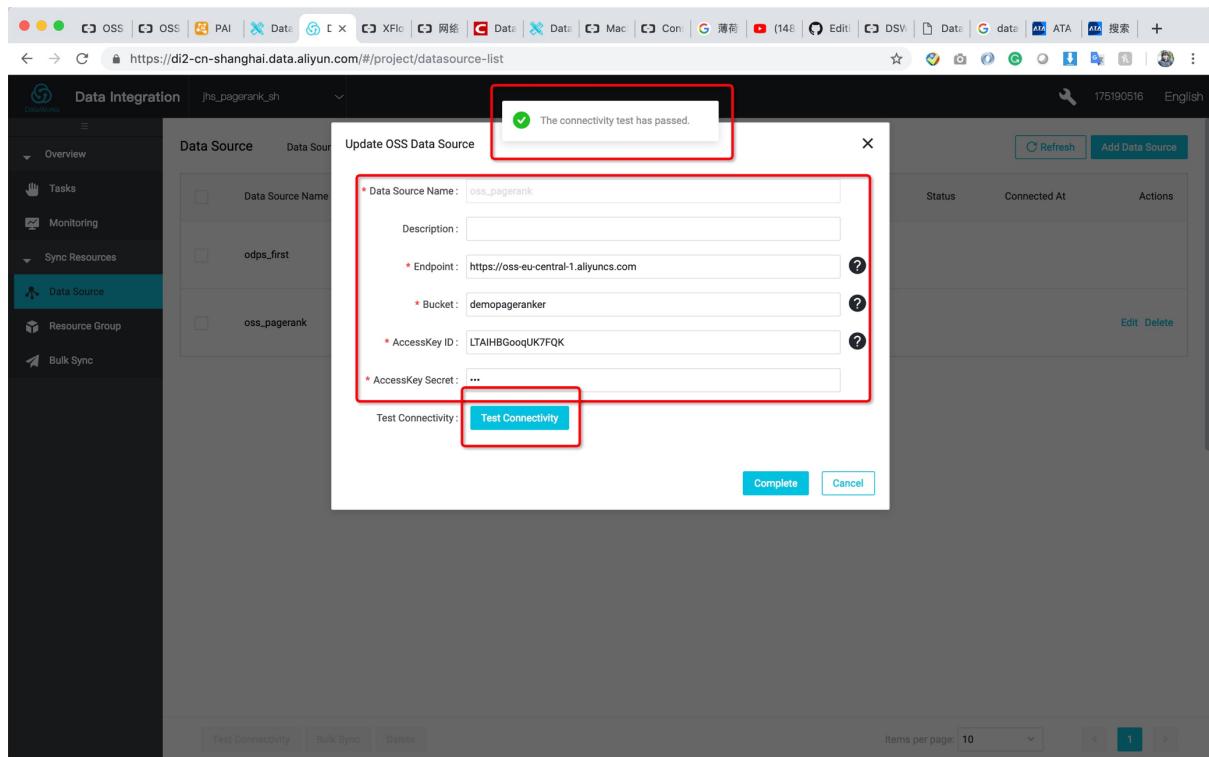
The screenshot shows the DataWorks workspace creation interface. It includes sections for 'Region' (China East 2 selected), 'Compute Engines' (MaxCompute selected, Pay-As-You-Go and Subscription options available), and 'DataWorks Services' (Data Integration selected, Data Analytics, O&M, and Administration available). A 'Create Workspace' button is visible on the left.

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

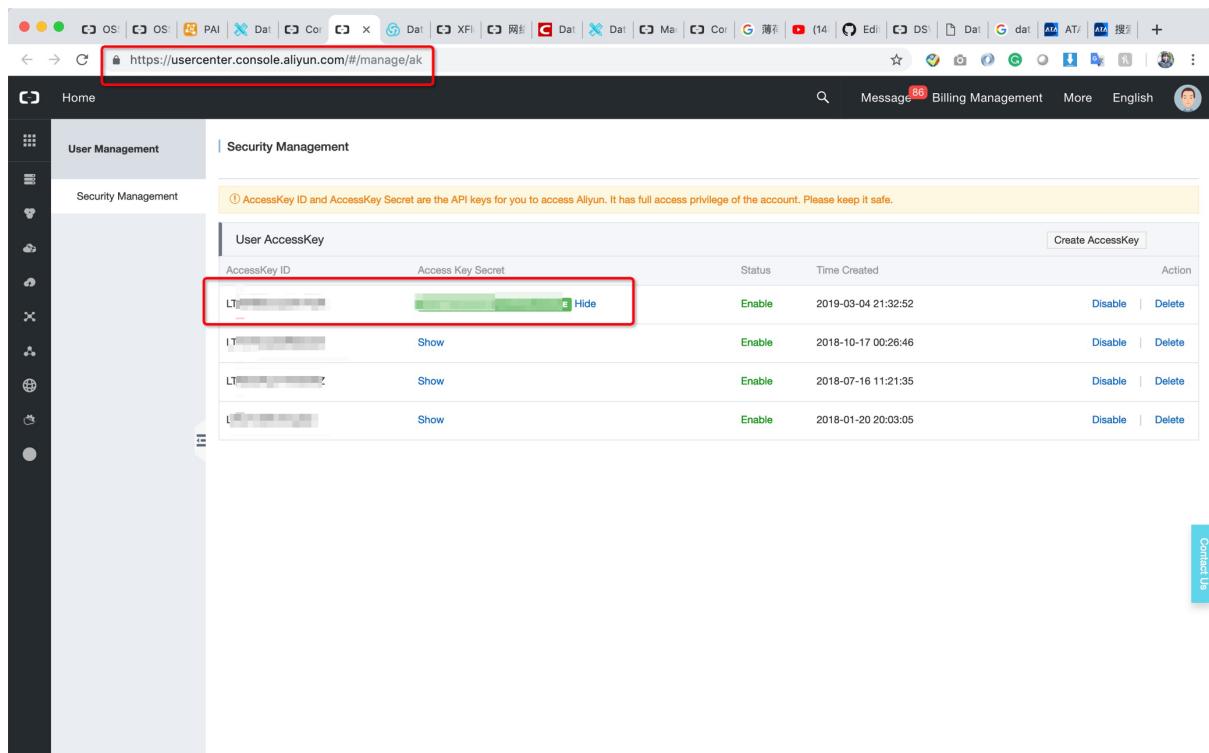
The screenshot shows the Data Integration interface with the URL <https://di2-cn-shanghai.data.aliyun.com/#/project/datasource-list>. The left sidebar has a 'Data Source' item highlighted with a blue box. The main area displays a table of data sources. A red box highlights the 'Add Data Source' button in the top right corner of the table header. The table columns include Data Source Name, Data Source Type, Link Information, Description, Created At, Status, Connected At, and Actions.

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			Edit Delete
oss_pagerank	OSS	Access Id: LTAIBG0oqUK7FQK Bucket: demopageranker Endpoint: https://oss-eu-central-1.aliyuncs.com		Mar 15, 2019 17:48:32			Edit Delete

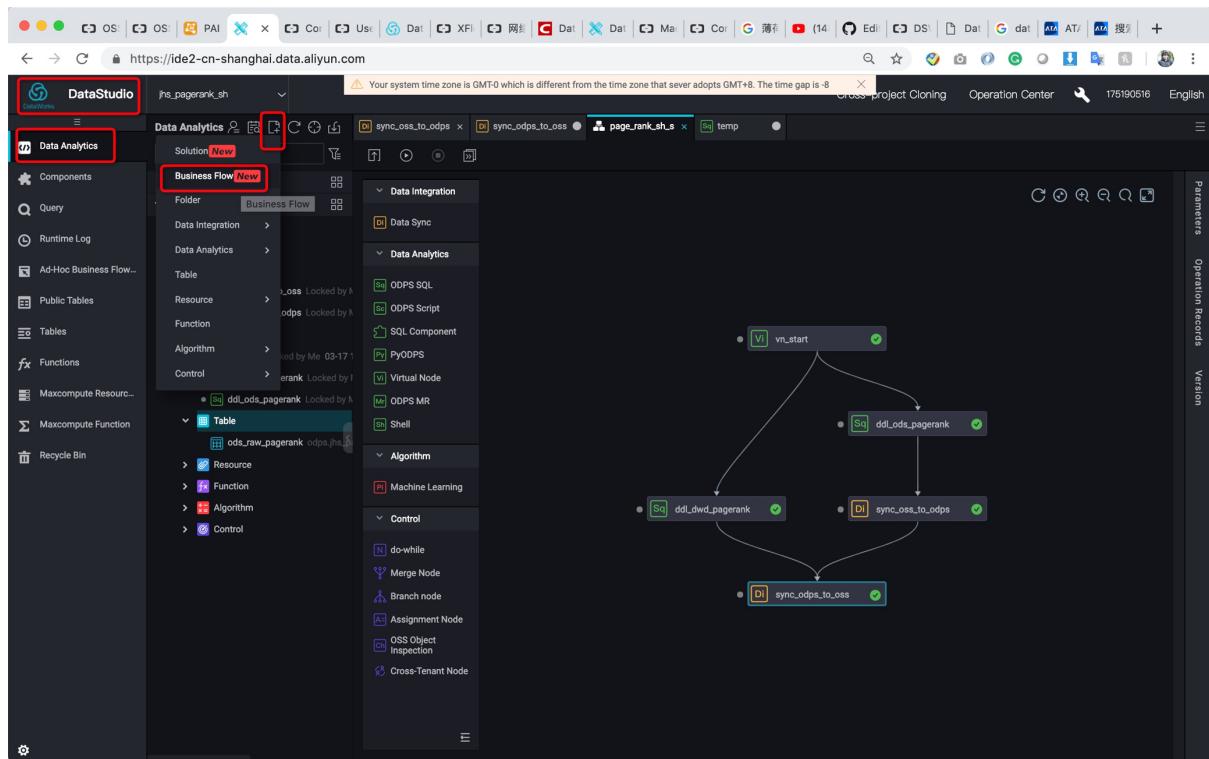
The screenshot shows the 'Add Data Source' dialog box over the main Data Integration interface. The dialog is titled 'Add Data Source' and contains sections for Relational Database, Big data storage, Semi-structured storage, and NoSQL. A red box highlights the 'OSS' icon under the 'Semi-structured storage' section. The 'Relational Database' section includes icons for MySQL, SQL Server, PostgreSQL, Oracle, and DM. The 'Big data storage' section includes icons for DRDS, POLARDB, HybridDB for MySQL, and HybridDB for PostgreSQL. The 'NoSQL' section includes icons for MongoDB, Datahub, AnalyticDB (ADS), and Lightning. The 'Cancel' button is at the bottom right of the dialog.



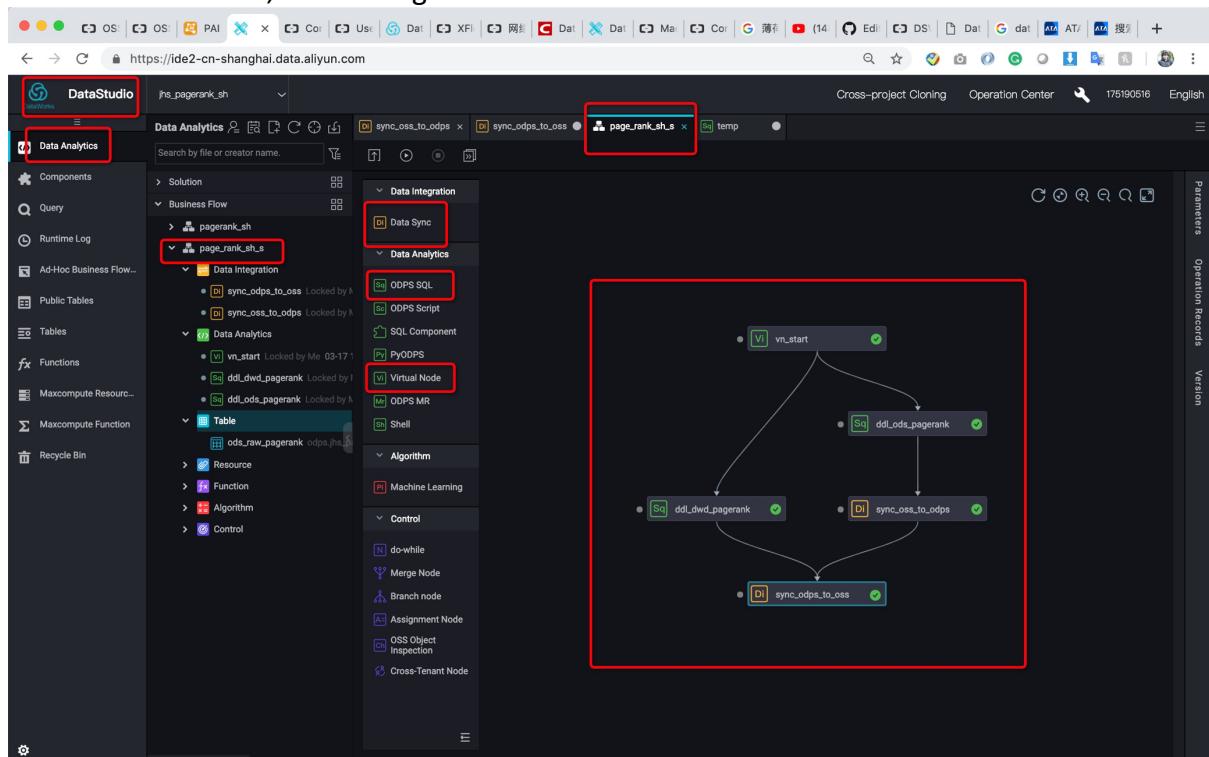
The AccessKey info can be found in : <https://usercenter.console.aliyun.com/#/manage/ak>



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:

The screenshot shows the DataStudio interface for configuring a Data Analytics job. The job name is 'vn_start'. In the 'Dependencies' section, there is one entry: 'jhs_pagerank_sh.root'. The 'Use Root Node' button is checked. The 'Output' section shows two entries: 'jhs_pagerank_sh.500255320_out' and 'jhs_pagerank_sh.vn_start'.

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.

```
--odps sql
--*****+
--author:175198516
--create time:2019-03-17 17:19:10
--*****+
create table if not exists ods_raw_pagerank (
    from_user string,
    to_user string,
    weight double
);
```

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”.

```
--odps sql
-----
--author:175196516
--create time:2019-03-17 17:15:00
-----
create table if not exists dwd_pagerank_out (
    user string,
    weight double
);
```

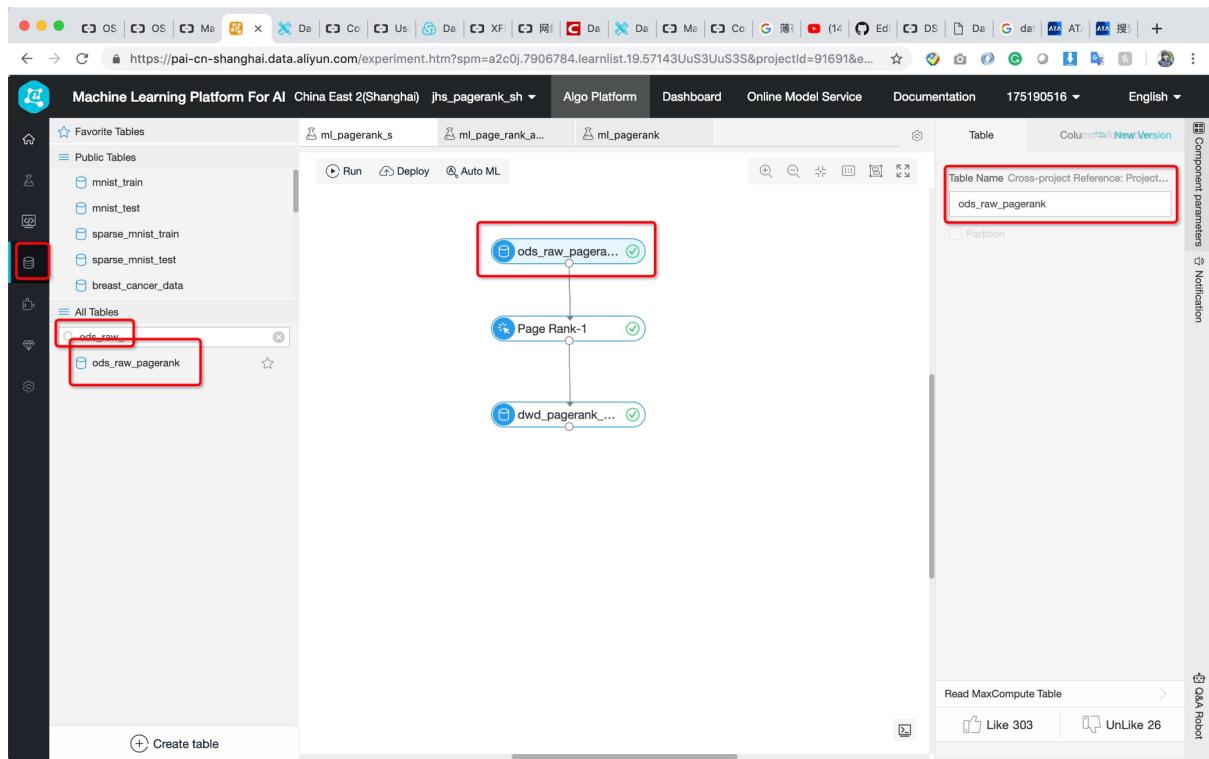
Go to PAI to enter experiment: <https://intl.data.aliyun.com/console/learn>

The screenshot shows the Aliyun Data Console interface. The URL in the address bar is https://intl.data.aliyun.com/console/learn. On the left sidebar, there are several sections: Home, Products (with Project Management selected), Security Overview, Anti-DDoS Basic, Anti-DDoS Pro (GameShield, Web Application Firewall, Server Guard, SSL Certificates, Content Moderation), Domains & Websites, and Market. The main content area is titled 'Project Management' and shows a table of projects. One project, 'jhs_pagerank_sh', is highlighted with a red box around its row. The 'Operation' column for this project contains a blue button labeled 'Go to PAI', which is also highlighted with a red box.

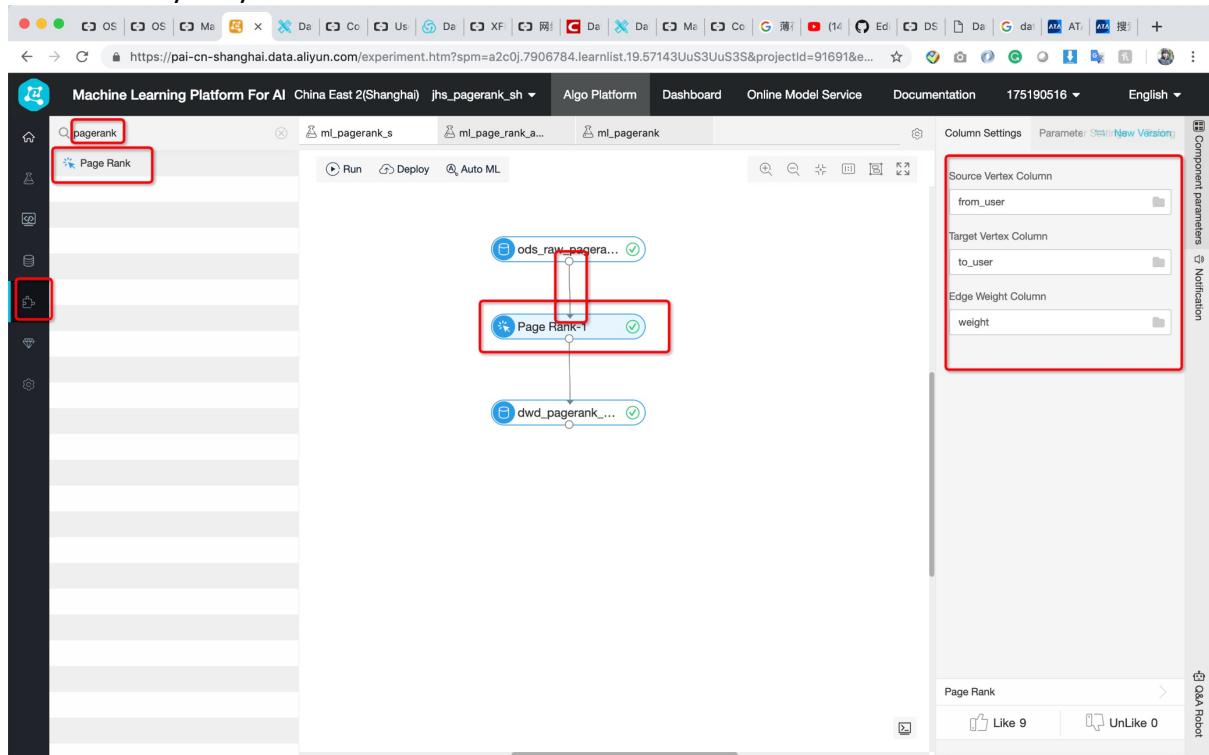
Create a new experiment, called

The screenshot shows the PAI interface. The title bar says 'Machine Learning Platform For AI China East 2(Shanghai) jhs_pagerank_sh'. The left sidebar has a 'My Experiments' section with three entries: 'ml_pagerank_s', 'ml_page_rank_alone', and 'ml_pagerank'. In the center, a 'New Experiment' dialog box is open, containing fields for 'Name' (set to 'ml_page_rank_s'), 'Project' (set to 'jhs_pagerank_sh'), and 'Description' (empty). Below these is a 'Save to' dropdown set to 'My Experiments'. At the bottom right of the dialog is a 'Create' button. At the bottom left of the page, there is a red box highlighting a '+ New Experiment' button.

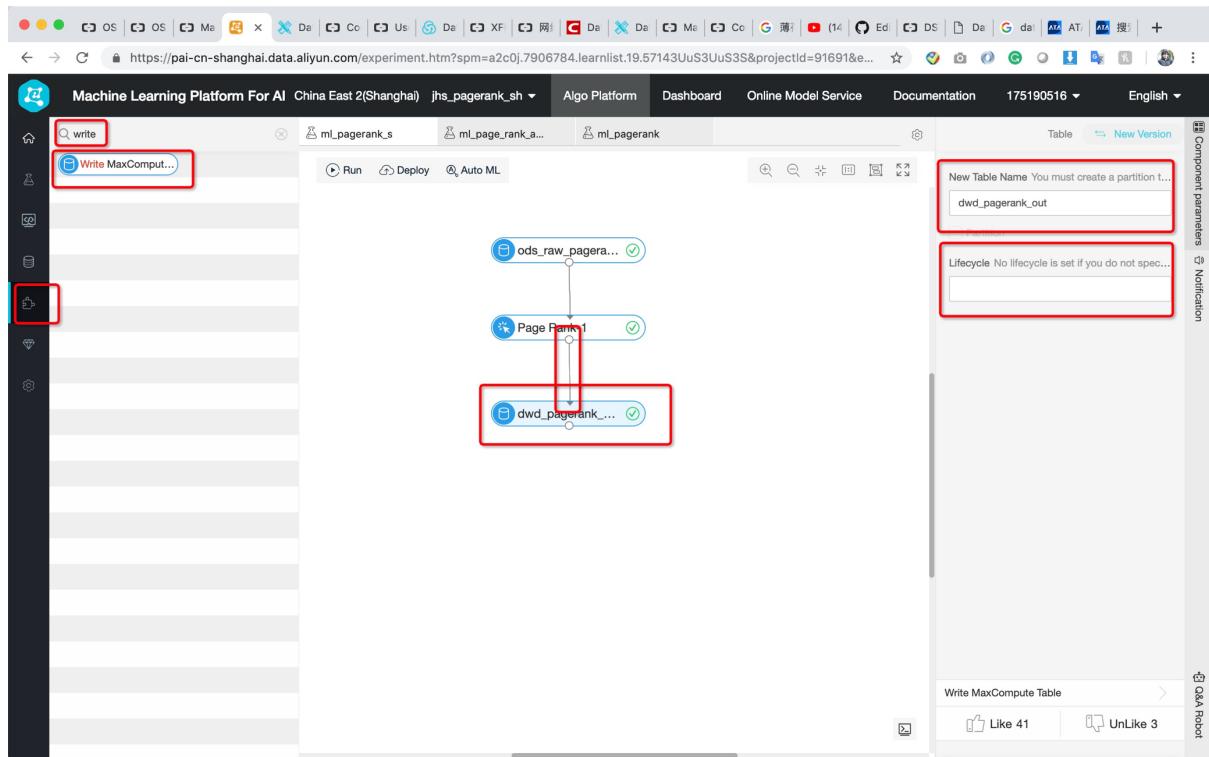
Go to “data source”, and then search for “ods_raw_pageRank” which is what we have ingested, simply just drag and drop it into the PAI canvas.



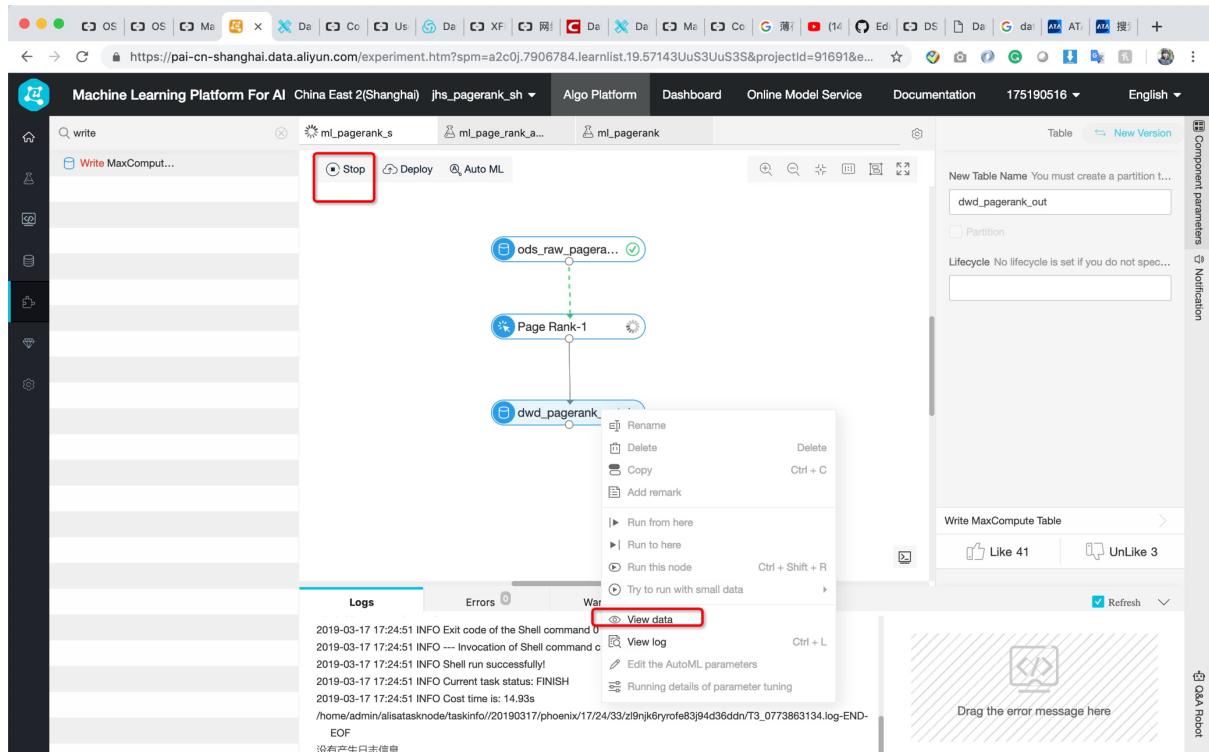
Go to “component” and then search for “pagerank”, you can simply just drag-drop into canvas, and connect “ods_raw_pagerank_xxx” component to “page rank_xxx” component, you can then configure “page rank” component, these filed should be able to pop up automatically for you to select.



Go to “component” and then search for “write”, find “write maxcompute 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.

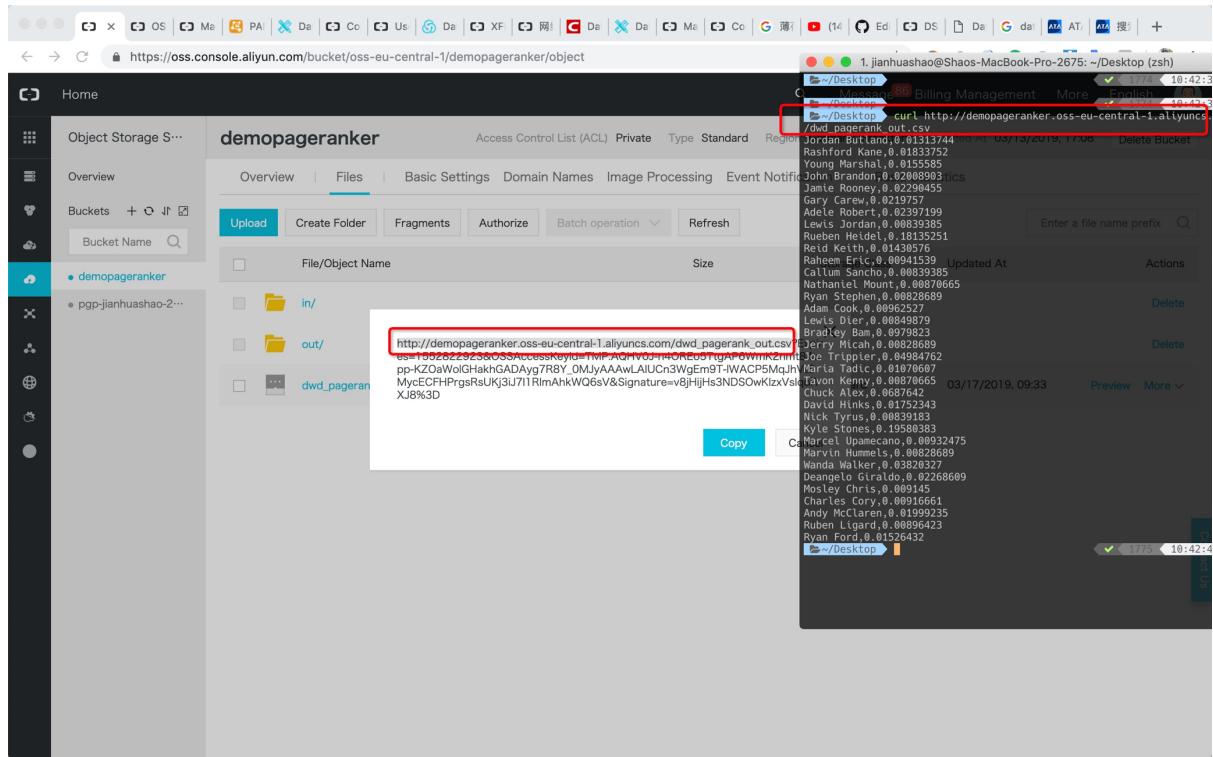
You can see the result file appear in /out/ folder. You can also setup “soft link” to have a customised file name.

The screenshot shows the Aliyun OSS console interface. On the left, there's a sidebar with various icons and a tree view showing a bucket named 'demopageranker'. The main area is titled 'demopageranker' and has tabs for 'Overview', 'Files', 'Basic Settings', etc. Under 'Files', there's a table listing objects. One object, 'dwd_pagerank_out.csv', is selected and highlighted with a red box. A context menu is open over this file, with several options highlighted with red boxes: 'Set soft link', 'Set ACL', 'Download', 'Copy File URL', and 'Delete'.

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

The screenshot shows the Aliyun OSS console interface. On the left, there's a sidebar with various icons and a tree view showing a bucket named 'demopageranker'. The main area is titled 'demopageranker' and has tabs for 'Overview', 'Files', 'Basic Settings', etc. Under 'Files', there's a table listing objects. One object, 'dwd_pagerank_out.csv', is selected and highlighted with a red box. A tooltip is displayed over this file, showing the source file address: 'out/dwd_pagerank_4ed4422c44cb451689d3fafaf50dd288'. The 'More' button in the file list is also highlighted with a red box.

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