Workspace Workload Analysis

Disclaimer

This application is **not** part of the Snowflake Service and is governed by the terms in LICENSE.txt, unless expressly agreed to in writing. You use this application at your own risk, and Snowflake has no obligation to support your use of this application.

License

Overview

The following notebook Workspace Workload Analysis allows you to analyze information about your Databricks workspace workload without giving direct access to your system.

How it works

- This notebook uses standard Databricks python API to gather minimal clusters, jobs, tasks, and runs information.
- It is required to be executed by the workspace owner.
- The information is stored in PySpark dataframes and a few visualizations are provided for your perusal.
- If you decide to share the information, At the end of the notebook the data frames are converted to CSV, and zipped. You can then press the **download zip** button to get the file.

Requirements

Permissions

The workspace admin account will required the following permissions:

- Personal Access Tokens
- Workspace visibility Control
- Cluster Visibility Control
- Job Visibility Control
- DBS File Browser

Workflows/Jobs

To be able to make an estimation, we need each jobs:

- Scheduling must be configured
- At least one successful execution in the last 60 days.
- If configured in staging or production mirror, it must include
  - Same machine configuration
  - Complete dataset

Steps

1. Open your workspace and create a new notebook.
2. Open File menu and click import.
3. Select URL and paste ***workspace_estimator.html***
4. Follow notebook instructions.

1. Install dependency with pip.
2. Update configuration according to your cluster (host_url, and token). ***We Advise against using the token directly in the notebook. Please store it in a secret scope, using the CLI***. For more details Authentication

Sample Data

Clusters

cluster_id	cluster_name	driver_node_type_id	node_type_id	state	cluster_source	runtime_engine	spark_version	workload	num_workers	autoscale_min	autoscale_max	calc_workers	cluster_memory_mb	cluster_cores	cluster_gpus
0123-345678-g4bimt0u	photon Shared Compute Cluster	Standard_D4as_v5	Standard_DS3_v2	RUNNING	UI	PHOTON	11.3.x-scala2.12	STANDARD	1	1	2	1	30720	8	0
1234-567890-abcdef0h	Another Cluster	Standard_DS3_v2	Standard_DS3_v2	TERMINATED	UI	STANDARD	10.4.x-cpu-ml-scala2.12	ML	1	0	0	1	28672	8	0
0234-567890-abcdef0h	demo	Standard_DS3_v2	Standard_DS3_v2	TERMINATED	UI	STANDARD	11.3.x-scala2.12	STANDARD	0	0	0	0	14336	4	0

Jobs

job_id	name	num_tasks	schedule	status
123456778000000	different_driver_and_worker_job	1		
123456778000001	sample_job	2	32 34 14 ? * MON-TUE	PAUSED
35	automl	1	57 23 19 * * ?	PAUSED

Jobs Tasks

job_id	name	num_tasks	schedule	status
123456778000004	Python_job	1	0 17 10 * * ?	PAUSED
123456778000000	sample_job	2	32 34 14 ? * MON-TUE	PAUSED
35	automl	1	57 23 19 * * ?	PAUSED

All Jobs Clusters

job_id	job_cluster_key	spark_version	driver_node_type_id	node_type_id	num_workers	runtime_engine	cluster_source
123456778000006	different_driver_and_worker_cluster	11.3.x-scala2.12	Standard_DS4_v2	Standard_DS3_v2	1	STANDARD	JOB
123456778000008	Job_cluster	11.3.x-scala2.12	Standard_DS3_v2	Standard_DS3_v2	8	STANDARD	JOB
123456778000008		12.2.x-scala2.12	Standard_DS4_v2	Standard_DS4_v2	2	STANDARD	UI

Job Runs

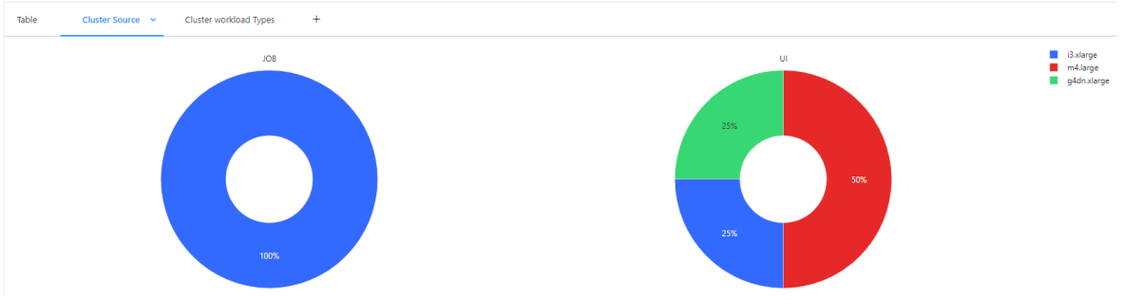
run_id	job_id	result_state	life_cycle_state	start_time	end_time	run_duration_sec	setup_duration_sec	execution_duration_sec	cleanup_duration_sec	date	trigger
987654	123456778000006	SUCCESS	TERMINATED	1679343600648	1679344048022	447.374	0	0	0	2023-03-20	PERIODIC
987653	123456778000001	SUCCESS	TERMINATED	1674185037643	1674185582464	544	254	290	0	2023-01-20	PERIODIC

Job Duration

job_id	name	num_drivers	num_workers	driver_node_type_id	node_type_id	runtime_engine	cluster_source	schedule	duration_min	success_rate
123456778000009	multi-cluster-job_with_jobs_cluster	1	8	Standard_DS3_v2	Standard_DS3_v2	STANDARD	JOB		11	1
123456778000009	different_driver_and_worker_job	1	1	Standard_DS4_v2	Standard_DS3_v2	STANDARD	JOB		5	1
123456778000011	sample_job	1	8	Standard_DS3_v2	Standard_DS3_v2	STANDARD	JOB	32 34 14 ? * MON-TUE	18	0.3333333333333333

Examples of Visualizations

Clusters



Top Time-consuming jobs