Private Preview : User Guide for Hybrid Pool

Overview

On 1st Dec 2020, we will start the Private Preview of Hybrid Pools. Hybrid Pools allows customers to create clusters and select different Databricks pools for driver and workers. It provides a way to support driver vs. worker heterogeneity, and therefore, considerably increases the flexibility of Databricks pools. Some typical examples supported:

- It would allow on-demand driver and spot workers on the cluster
- It would allow a beefier driver and smaller sized workers on the cluster
- It would allow worker and drivers coming from different instance families

Setup and spinning up Hybrid Pool Clusters/ Jobs

Add new field in to current clusters/jobs API

(AWS API docs, Azure API docs)

driver_instance_pool_id	STRING	The optional ID of the instance pool for the driver of the cluster belongs. The pool cluster uses the instance pool with id(instance_pool_id) if the driver pool is not assigned.

** The driver_instance_pool_id is not a final API field. It's for private preview.*

Example of Hybrid Pool Cluster

```
{
    "num_workers": 1,
    "cluster_name": "test-hybrid-create",
    "spark_version": "7.2.x-scala2.12",
    "spark_conf": {},
    "aws_attributes": {
    },
    "ssh_public_keys": [],
    "custom_tags": {},
    "spark_env_vars": {
        "PYSPARK_PYTHON": "/databricks/python3/bin/python3"
    },
    "autotermination_minutes": 120,
    "init_scripts": [],
    "instance_pool_id": "1109-172550-mimic2-pool-worker",
    "driver_instance_pool_id": "1109-172516-retch1-pool-driver"
}
```

By this setting, a cluster with 1 worker will be created. The driver of the cluster will come from the driver pool with id "1109-172516-retch1-pool-driver" and the worker will come from the worker pool with id "1109-172550-mimic2-pool-worker".

Example of Hybrid Pool Job

```
{
  "name": "test-job-hybrid",
  "new_cluster": {
    "num_workers": 1,
    "spark_version": "7.2.x-scala2.12",
    "spark_conf": {},
    "aws_attributes": {
    },
    "ssh_public_keys": [],
    "custom_tags": {},
    "spark_env_vars": {
        "PYSPARK_PYTHON": "/databricks/python3/bin/python3"
    },
    "init_scripts": [],
    "instance_pool_id": "1109-172550-mimic2-pool-worker",
    "driver_instance_pool_id": "1109-172516-retch1-pool-driver"
  },
  "libraries": [
    {
      "jar": "dbfs:/my-jar.jar"
    },
    {
      "maven": {
        "coordinates": "org.jsoup:jsoup:1.7.2"
      }
    }
  ],
  "timeout_seconds": 3600,
  "max_retries": 1,
  "spark_jar_task": {
    "main_class_name": "com.databricks.ComputeModels"
  }
}
```

By this setting, a job with 1 worker will be created. The driver of the cluster will come from the driver pool with id "1109-172516-retch1-pool-driver" and the worker will come from the worker pool with id "1109-172550-mimic2-pool-worker". When the job starts running, the cluster will be created with the driver and workers from different pools.

Assumptions & Limitations

- During private preview, cluster creation / edit from hybrid pools is only supported via API. But once the cluster is created, it can be viewed in the UI. *→ Cluster create / edit UI will be supported at the end of Q4.*
- Tag propagation to Clusters: if a cluster's driver and workers are from different pools. Only the worker pool custom tags will be propagated as cluster tags. However,

Databricks will automatically tag the cluster with the DriverInstancePoolId for identification purposes. (Note that tag propagation to VMs will remain unchanged from today's behavior - driver VM gets driver pool tags; worker VMs get worker pool tags) → *Future improvements on this limitation will be based on customer feedback during preview.*

- The feature is currently in private preview with no SLA guarantee. Therefore, we do not recommend customers to use the feature for production workloads until fully tested.