How to get number of spark clusters running at given time for dataflow

Last updated by | Nelson Gomes | Mar 7, 2023 at 3:37 AM PST

Contents

Description

The methods should give similar count for running clusters. It is expected the # running clusters coming from spark logs to be slightly higher because spark may take some to release a cluster. This information will be useful when customer is reaching the default limit of 200 spark clusters per workspace. Please, use the queries below to confirm if customer indeed has reached the limit.

Method 1: using Data factory logs

```
// ADF kusto logs only: Getting spark clusters till max datetime we have in our logs for a given activity run(
let MyActivityId="cbdb887b-bce4-4e83-8ac1-46194ae58553"; // change here for the activity run you are intereste
let MydateReference=toscalar(cluster('adfneu.kusto.windows.net').database('AzureDataFactory').ActivityRuns
|where activityRunId ==MyActivityId and status !="Succeeded"
|summarize max(PreciseTimeStamp));
let MySubscription=toscalar(cluster('adfneu.kusto.windows.net').database('AzureDataFactory').ActivityRuns
|where activityRunId ==MyActivityId |project subscriptionId| take 1);
let MyADFname=toscalar(cluster('adfneu.kusto.windows.net').database('AzureDataFactory').ActivityRuns
|where activityRunId ==MyActivityId |project dataFactoryName| take 1);
cluster('adfneu.kusto.windows.net').database('AzureDataFactory').ActivityRuns
|where PreciseTimeStamp <=MydateReference and subscriptionId ==MySubscription
|where activityType == "ExecuteDataFlow" and category =="ActivityRuns"
|summarize arg_max(PreciseTimeStamp,*) by activityRunId|where status =="InProgress"|project-rename status1=sta
|join kind=leftouter (
cluster('azuredmprod.kusto.windows.net').database('AzureDataMovement').ExecutionApiCall
|where PreciseTimeStamp <=MydateReference and SubscriptionId =~MySubscription
|where ExecutionType =="executemappingdataflow" //and category =="ActivityRuns"
|summarize arg_max(PreciseTimeStamp,*) by RunId
| where Status =="InProgress") on $left.activityRunId==$right.RunId
|join kind=leftouter(
cluster('adfneu.kusto.windows.net').database('AzureDataFactory').DataflowClusterLogs
|where  Timestamp<=MydateReference
|summarize arg_max(Timestamp,*) by ActivityRunId
) on $left.activityRunId==$right.ActivityRunId
| where Status !="" and IntegrationRuntime !="" and dataFactoryName =~MyADFname //
|summarize RunningClustersTillDateTimeRef=dcount(ClusterTrackingId)  by IntegrationRuntime
|extend ADF=MyADFname,DateTimeRef=MydateReference,ClusterState="Running"
```

Method 2: using Synapse Spark cluster logs

```
//Synapse spark kusto logs only: getting spark clusters till max datetime we have in our logs for a given acti
let MyActivityId="cbdb887b-bce4-4e83-8ac1-46194ae58553";// change here for the activity run you are interested
let MydateReference=toscalar(cluster('adfneu.kusto.windows.net').database('AzureDataFactory').ActivityRuns
|where activityRunId ==MyActivityId and status !="Succeeded"
|summarize max(PreciseTimeStamp));
let MyADFname=toscalar(cluster('adfneu.kusto.windows.net').database('AzureDataFactory').ActivityRuns
|where activityRunId ==MyActivityId |project dataFactoryName| take 1);
 cluster('analytics365prodeu.westeurope.kusto.windows.net').database('Analytics365SparkPROD').ClusterStateChan
 |where PreciseTimeStamp <=MydateReference
 |where Workspace contains MyADFname
 |summarize arg_max(TIMESTAMP,*) by ClusterName
 |where ToClusterState=="Running"  or ToClusterState=="Failed"
 |summarize RunningClustersTillDateTimeRef=count(Specification) by IR=Specification,ToClusterState//by bin(Pre
 |extend ADF=MyADFname,DateTimeRef=MydateReference
 |project-rename ClusterState=ToClusterState
```

Additional information

- **IcM reference:**
- **Ava:**
- **Author:** negome
- **Reviewer:**
- **Keywords:** dataflow clusters, clusters limit, running clusters.

How good have you found this content?