INC-1696 Muted Monitor Review: Cloned Initial Investigation

Eric Swanson

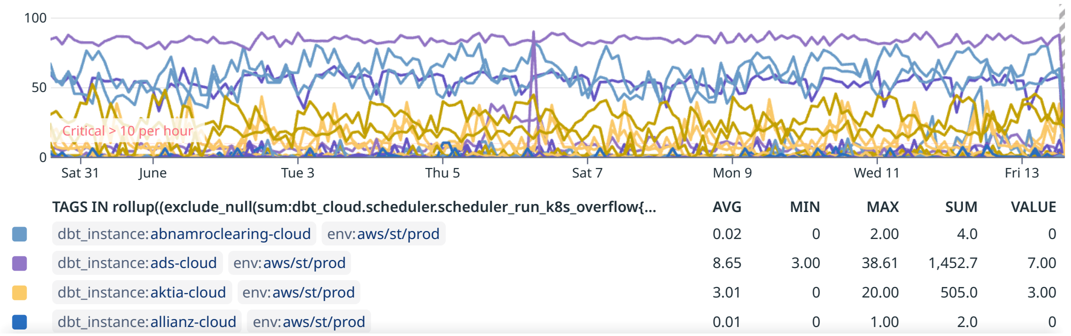
SRE-1890 [SPIKE] Evaluate 15w downtime for removal

- Cloned from INC-1696 Muted Monitor Review
- Investigation begun in thread of #inc-1696-datadog-agent-availability-on-st-instances
- MT NO DATA or Alerting, sorted by Muted Elapsed descending

Orchestration Monitors

- Percentage of overflow expired runs is high
  - noisy. Muted by both of these plus prod-us3-c1
  -
- MT Prod AWS - Scheduler Pod Memory Usage Percent
- MT/MC Production - Percentage of cold runs is high
- MT/MC Production - Percentage of delayed runs is high
- MT/MC Production - Scheduler Slow Update Propagation for Unknown fields

ORC - Percentage of overflow expired runs is high



	TAGS IN rollup((exclude_null(sum:dbt_cloud.scheduler.scheduler_run_k8s_overflow{...	AVG	MIN	MAX	SUM	VALUE
🟦	dbt_instance:abnamroclearing-cloud env:aws/st/prod	0.02	0	2.00	4.0	0
🟪	dbt_instance:ads-cloud env:aws/st/prod	8.65	3.00	38.61	1,452.7	7.00
🟨	dbt_instance:aktia-cloud env:aws/st/prod	3.01	0	20.00	505.0	3.00
🟦	dbt_instance:allianz-cloud env:aws/st/prod	0.01	0	1.00	2.0	0

Sunday 3/23 Investigation

- Focused on a refined filter on muted monitors since 15 weeks ago, managed by terraform
- Initial Query only show muted *at least* this long. We want the opposite.
- Show a list of all tags being filtered by the downtime exclusion.
  - are any of them valid?
  - Invalid query when pasted directly into triggers
  - `group:"dbt_instance:*" AND NOT group:"dbt_instance:prod*"` returns nothing
  - `group:dbt_instance:prod*` (no quotes) returns 55 correct looking triggered anti-patterns.

this works to start with (614 results) -group"dbt_instance:prod*

266 after excluding ExternalSecrets

dbt_instance:prod-us3-c1,env:gcp/mc-mt/prod is matched despite group exclusion; likely invalid syntax. Revisit with team. Example monitor

# MT/MC Production - Cloud CLI Server: Degraded Availability

https://www.notion.so/dbtlabs/Getting-to-Know-You-Datadog-Best-Practices-1bbbb38ebda780108957cfd5ddb70d24?pvs=4#1bfbb38ebda780e4b8f2fc89c10aa305

exclude for a moment anything populated with `env` in the group by
-group:dbt_instance:prod* group:dbt_instance:* muted:true tag:("managed-by:terraform") -tag:("initiative:observability") -tag:(domain:) -*ExternalSecrets -group:env:*

Step back and retry with more recent muted alerts.

Monitor List Search

To further narrow the search, start with the raw list of monitors by tags.

- muted:true tag:("managed-by:terraform") -tag:("initiative:observability") tag:(domain:*) -ExternalSecrets
  - returns 96 legacy domain monitors to exclude (recently downtimed by group scope)
- tag:"managed-by:terraform" -tag:("initiative:observability") -tag:(domain:*) -ExternalSecrets notification:webhook-incident-io -muted_elapsed:16w muted_elapsed:14w
  gets us down to 82 monitors muted by this downtime, to explore for overly broad scope.
  - Filtering further on exclude status OK, and only 1 is alerting

Updated query to match impact of second 15w downtime: 80 total

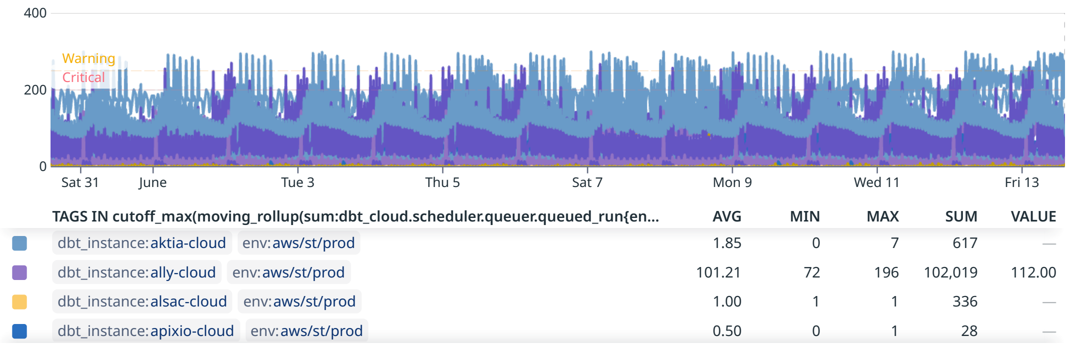
27 triggered in the past 3 days that are not OBS SLO

- `group:(dbt_instance:chrobinson-prod OR dbt_instance:cona-prod OR dbt_instance:bhp-cloud OR dbt_instance:chs-cloud OR dbt_instance:cisco-cloud OR dbt_instance:medtronic-cloud OR dbt_instance:nbfc-cloud OR dbt_instance:nbim-cloud OR dbt_instance:onsemi-cloud OR dbt_instance:rga-cloud OR dbt_instance:siemens-d2go OR dbt_instance:transpower-cloud OR dbt_instance:vfnz-cloud OR dbt_instance:virginmedia-cloud OR dbt_instance:mohnz-prod OR dbt_instance:scentgroup-cloud OR dbt_instance:usaa-stage) -tag:"initiative:observability" triggered:4320`
-

# MT/MC Production - Scheduler queueing fewer Runs than expected

- shows cutover and alerting for metrics not yet rolled off

```
sum:dbt_cloud.scheduler.queuer.queued_run{env:*prod*,!dbt_instance:prod-emea,!dbt_instance:prod-au} by {dbt_instance,env}.as_count().rol
```

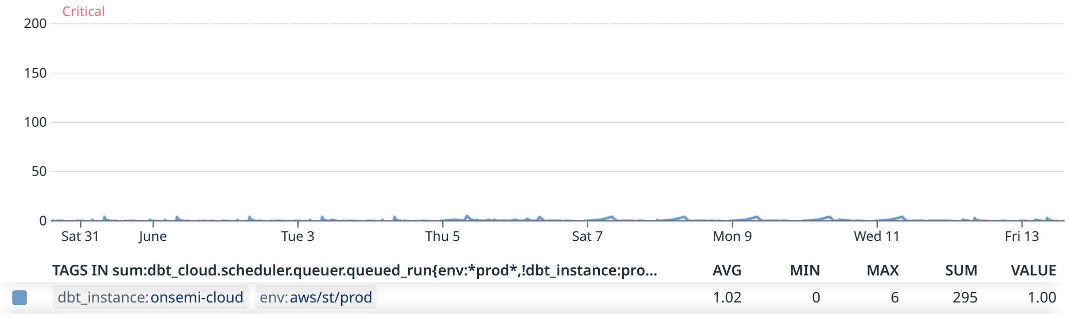


	TAGS IN cutoff_max(moving_rollup(sum:dbt_cloud.scheduler.queuer.queued_run{en...	AVG	MIN	MAX	SUM	VALUE
🟦	dbt_instance:aktia-cloud env:aws/st/prod	1.85	0	7	617	—
🟪	dbt_instance:ally-cloud env:aws/st/prod	101.21	72	196	102,019	112.00
🟨	dbt_instance:alsac-cloud env:aws/st/prod	1.00	1	1	336	—
🟦	dbt_instance:apixio-cloud env:aws/st/prod	0.50	0	1	28	—

another triggered group.

- None of these are paging alerts so far; slack olny

```
sum:dbt_cloud.scheduler.queuer.queued_run{env:*prod*,!dbt_instance:prod-emea,!dbt_instance:prod-au,dbt_instance:onsemi-cloud,env:aws/st/p
```



	TAGS IN sum:dbt_cloud.scheduler.queuer.queued_run{env:*prod*,!dbt_instance:pro...	AVG	MIN	MAX	SUM	VALUE
🟦	dbt_instance:onsemi-cloud env:aws/st/prod	1.02	0	6	295	1.00

Were the thresholds much lower for ST customers on this monitor? just not seeing anywhere near what's expected.

- Evaluate the sum Of the query over the last 10 minutes
- Trigger when the evaluated value is below the threshold for any dbt_instance, env
  - Alert threshold: < 200

dbt_cloud.scheduler.queuer.queued_run

var.scheduler_failing_to_queue_runs

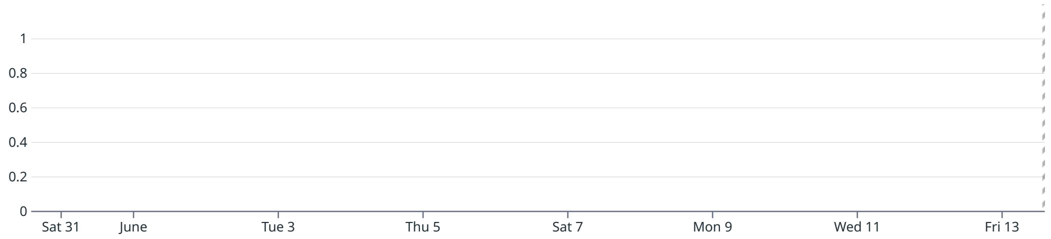
- for deployments of `resource "datadog_monitor" "scheduler_failing_to_queue_runs" {`
- to: `count = contains(var.datadog_account_deploy_list[var.datadog_account_name], "runs") && var.dbt_instance == "prod-us" ? 1 : 0`

:thinking: how is that var set?

MT/MC Production - Background Cleanup Process is Down

- past 2 weeks view of 25 groups at a time, by Triggered.
- Multiple downtimes applied
- exported alerting groups are copied below with rollup (sum in last hour <= 0)
- "We recommend the missing data window be at least 2x the evaluation period above"
- Notify if data is missing for 30 minutes

```
sum:dbt_cloud.background_task.background_cleanup_completed{env:*prod*,dbt_instance:usaa-prod,env:aws/st/prod,env:prod} by {dbt_instance,e
```

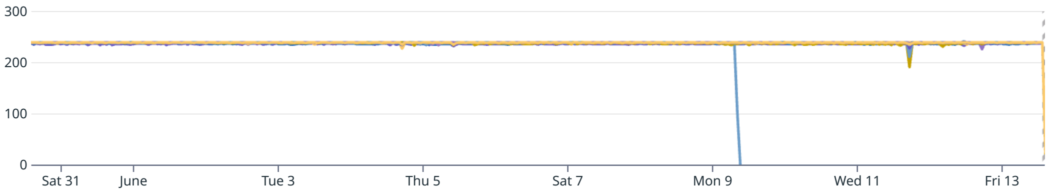


	TAGS	METRIC	AVG	MIN	MAX	SUM	VALUE
🟦	*	default_zero(sum:dbt...	0	0	0	0	0
🟪	*	default_zero(sum:dbt...	0	0	0	0	0

Reduced to original monitor query, with downtime filter added to view impact

- does mute ST metrics, should be removed.

```
sum:dbt_cloud.background_task.background_cleanup_completed{env:*prod*, !dbt_instance:prod*, dbt_instance:* } by {dbt_instance,env}.as_co
```

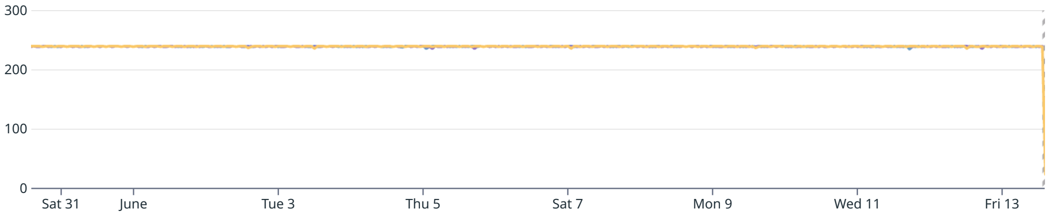


	TAGS IN default_zero(sum:dbt_cloud.background_task.background_cleanup_comple...	AVG	MIN	MAX	SUM	VALUE
🟦	dbt_instance:abnamroclearing-cloud env:aws/st/prod	239	26	241	80.3k	240
🟪	dbt_instance:ads-cloud env:aws/st/prod	239	26	240	80.3k	239
🟨	dbt_instance:aktia-cloud env:aws/st/prod	239	26	240	80.3k	240
🟦	dbt_instance:allianz-cloud env:aws/st/prod	239	26	240	80.3k	240

Cloned and revealed all groups for `dbt_instance:usaa-*`

- exclude `env:prod` and `env:dev` old tag values to see valid data
- :pushpin: check rolloff windows to remove those groups from alerting, then clear downtimes.

```
sum:dbt_cloud.background_task.background_cleanup_completed{!env:prod, !env:dev,dbt_instance:usaa-*} by {dbt_instance,env}.as_count().rol
```



	TAGS IN default_zero(sum:dbt_cloud.background_task.background_cleanup_comple...	AVG	MIN	MAX	SUM	VALUE
🟦	dbt_instance:usaa-dev env:aws/st/prod	239	26	240	80.3k	239
🟪	dbt_instance:usaa-prod env:aws/st/prod	239	26	240	80.2k	240
🟨	dbt_instance:usaa-stage env:aws/st/prod	239	24	240	80.3k	240

Message body

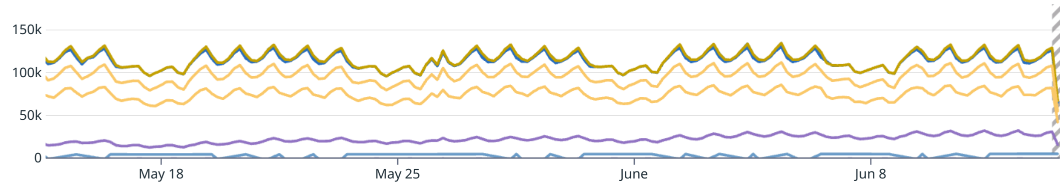
```
The Background Cleanup Process is down or stuck. This can happen because it is falling behind, or because there
is a bad / invalid value in the interface that is causing an unhandled exception. Check the log live tail to
figure it out.

:notion: [Runbook](https://www.notion.so/dbtlabs/Background-Cleanup-Process-is-Down-
f6b60af32f90400bba470d05da25b83d)

{{#is_alert}}
  {{#is_match "env.name" "prod" }}
@webhook-incident-io
@slack-Fishtown_Analytics-dev-orchestration-alerts
{{/is_match}}

{{#is_match "env.name" "staging" }}
@slack-Fishtown_Analytics-dev-orchestration-alerts-staging
{{/is_match}}

{{/is_alert}}

{{#is_no_data}}
  {{#is_match "env.name" "prod" }}
@webhook-incident-io
@slack-Fishtown_Analytics-dev-orchestration-alerts
{{/is_match}}

{{#is_match "env.name" "staging" }}
@slack-Fishtown_Analytics-dev-orchestration-alerts-staging
{{/is_match}}

{{/is_no_data}}

{{#is_alert_recovery}}
  {{#is_match "env.name" "prod" }}
@webhook-incident-io
@slack-Fishtown_Analytics-dev-orchestration-alerts
{{/is_match}}

{{#is_match "env.name" "staging" }}
@slack-Fishtown_Analytics-dev-orchestration-alerts-staging
{{/is_match}}

{{/is_alert_recovery}}

{{#is_no_data_recovery}}
{{#is_match "env.name" "prod" }}
@webhook-incident-io
@slack-Fishtown_Analytics-dev-orchestration-alerts
{{/is_match}}

{{#is_match "env.name" "staging" }}
@slack-Fishtown_Analytics-dev-orchestration-alerts-staging
{{/is_match}}

{{/is_no_data_recovery}}
```

Saturday 3/22 Investigation: No concrete findings

- MT/MC Production - No codex env loaders are starting
  - NO DATA for the past month.
  - Notifies when NO DATA for the past 10 minutes 😬
  - Exported query below shows `env` is not a tag value, but `environment` is 🤔
  - Check list of expected iteration through monitored tag values against reality and ask team to update if needed.

```
sum:codex.loader.started{codex_consumer_type:*} by {codex_consumer_type}.as_count()
```

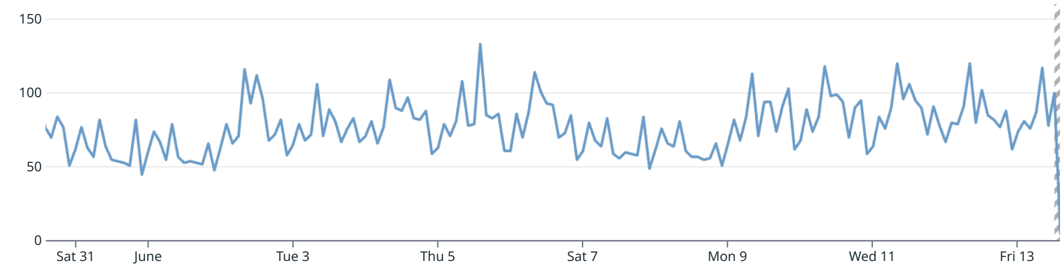
1mo Past 1 Month



	codex_consumer_type in sum:codex.loader.started{codex_consumer_type:*}.as_cou...	AVG	MIN	MAX	SUM	VALUE
🟦	billing	115.9k	67.5k	134.7k	20.86M	119.5k
🟪	catalog_metadata_consumer	12.88	1	35	554	—
🟨	cll	74.9k	42.8k	86.2k	13.48M	78.6k
🟦	environment	114.2k	65.7k	130.5k	20.56M	118.4k

- # MT/MC Production - No codex cll loaders are starting
  - similar to above, but NO DATA only for new ST instances; resolve those alerts and should have them drop off
  - specific valid instances do flap a lot; this monitor query needs to be adjusted or replaced with a different type.
    - those instances currently permamuted due to low traffic volume

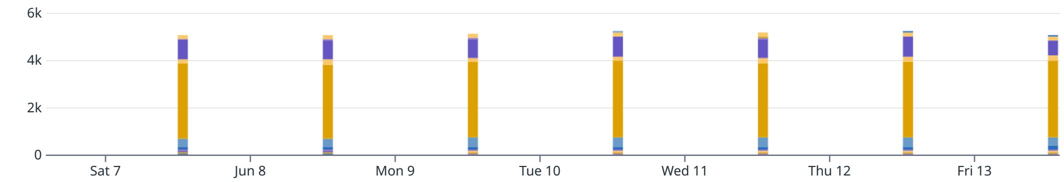
```
sum:codex.loader.started{codex_consumer_type:cll,!dbt_instance:prod-au,!dbt_instance:prod-us-c2,dbt_instance:prod-us2-c1} by {dbt_instanc
```



	dbt_instance in sum:codex.loader.started{codex_consumer_type:cll,!dbt_instance:pr...	AVG	MIN	MAX	SUM	VALUE
🟦	prod-us2-c1	76.8	5	133	12.9k	71.0

- # MT/MC Production - No codex model-query-history loaders are succeeding
  - misnamed; tag has underscored, not hyphens 🤦‍♂️
  - no recent data... converted lines to bars and confirmed this is once a day.

```
sum:codex.loader.success{codex_consumer_type:model-query-history,!dbt_instance:prod-au,!dbt_instance:prod-us-c2} by {dbt_instance}.as_co
```

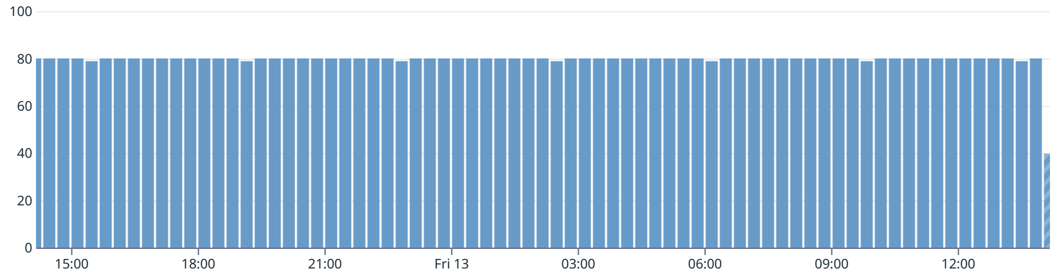


	dbt_instance in sum:codex.loader.success{codex_consumer_type:model_query_histo...	AVG	MIN	MAX	SUM	VALUE
🟦	bhp-cloud	18.00	18	18	126	—
🟦	biahealthnz-prd	9.00	9	9	63	—
🟪	cisco-cloud	1.00	1	1	7	—
🟪	cona-prod	1.00	1	1	7	—

- # MT/MC Production - Background Cleanup Process is Down
- Notify if data is missing for more than 30 minutes
  - Resolve old ST instance alerts and should drop them & return to green

```
sum:dbt_cloud.background_task.background_cleanup_completed{dbt_instance:tinder-cloud} by {dbt_instance}.as_count()
```

1d Past 1 Day

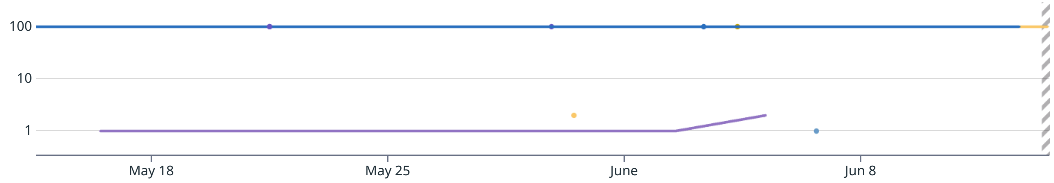


	dbt_instance in default_zero(sum:dbt_cloud.background_task.background_cleanup_...	AVG	MIN	MAX	SUM	VALUE
🟦	tinder-cloud	79.3	40	80	5.71k	80.0

- https://dbtlabsmt.datadoghq.com/monitors/67632680
- https://dbtlabsmt.datadoghq.com/monitors/71053047
- https://dbtlabsmt.datadoghq.com/monitors/65173499
- https://dbtlabsmt.datadoghq.com/monitors/114021527
  - needed default zero (follow up with pr)

- # MT/MC Production - Social Authentication Failure Rate
  - Does NOT notify on no data.
  - stopped emitting metrics on 11/22 after consistent data from prod-us-c3
  - Consider an slo style out of total hits, or default zero replacing fill zero

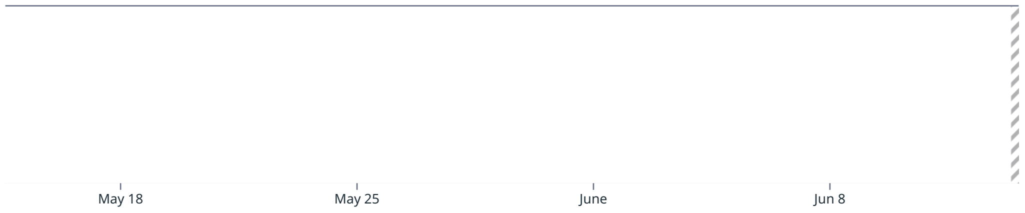
```
sum:trace.django.request.errors{resource_name:post_complete/_str:backend,!dbt_instance:dev} by {dbt_instance,service}.fill(zero), sum:tra
```



	TAGS	METRIC	AVG	MIN	MAX	SUM	VALUE
🟦	dbt_instance:prod-us service:dbt-cloud-app	sum:trace.django.req...	1.00 errs	1 errs	1 errs	1 errs	—
🟪	dbt_instance:prod-us-c3 service:dbt-cloud-app	sum:trace.django.req...	1.20 errs	1 errs	2 errs	6 errs	—
🟨	dbt_instance:staging-us-c1 service:dbt-cloud-app	sum:trace.django.req...	2.00 errs	2 errs	2 errs	2 errs	—
🟦	dbt_instance:abnamroclearing-cloud	100 * 1 - (sum:trace.d...	100	100	100	100	—

- # MT/MC Production CELL - API Gateway Replica Availability
- https://dbtlabsmt.datadoghq.com/monitors/122124246
- https://dbtlabsmt.datadoghq.com/monitors/103240000
  - notify if NO DATA
    - is default_zero wrapper disrespecting this setting?
    - *should* there be data in here?

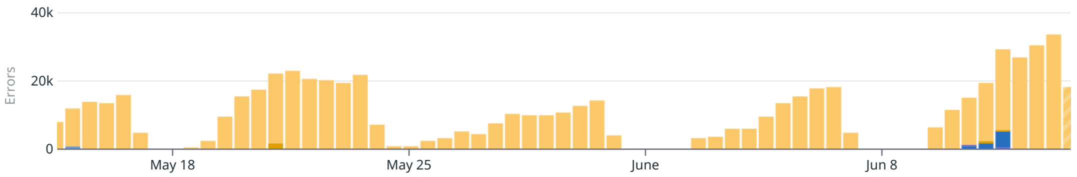
```
avg:kubernetes_state.deployment.replicas{kube_deployment:api-gateway} by {dbt_instance}, avg:kubernetes_state.deployment.replicas_desire
```



	TAGS	METRIC	AVG	MIN	MAX	SUM	VALUE
🟦	*	default_zero(avg:kub...	0	0	0	0	0
🟪	*	default_zero(avg:kub...	0	0	0	0	0
🟨	*	100 * (default_zero(a...	—	—	—	0	—

- [# MT/MC Production - Cloud CLI Server: Elevated Number of 5xx Errors]
    - default zero added
    - another: do we expect data?
        - very intermittent population of data

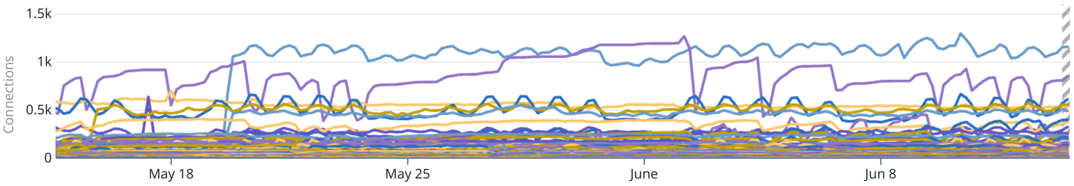
```
sum:trace.fastapi.request.errors.by_http_status{service:dbt-cloud-cli-server,http.status_class:5xx} by {dbt_instance}.as_count()   1m
```



	dbt_instance in sum:trace.fastapi.request.errors.by_http_status{service:dbt-clou...	AVG	MIN	MAX	SUM	VALUE
🟦	abnamroclearing-cloud	11.00 errs	11 errs	11 errs	11 errs	—
🟦	ads-cloud	21.00 errs	21 errs	21 errs	21 errs	—
🟪	ally-cloud	1.17 errs	1 errs	2 errs	7 errs	—
🟪	arrowelectronics-cloud	2.67 errs	1 errs	13 errs	24 errs	—

- # MT/MC Production - [db.m5.2xlarge] Database is Running out of Connections
- https://dbtlabsmt.datadoghq.com/monitors/133223600
- https://dbtlabsmt.datadoghq.com/monitors/133240305
- https://dbtlabsmt.datadoghq.com/monitors/134575594
-
    - NO DATA no dbinstanceclass matching that value.
        - review existing tags with ipa and confirm
        - one is double-tagged
        - one (+?) is null

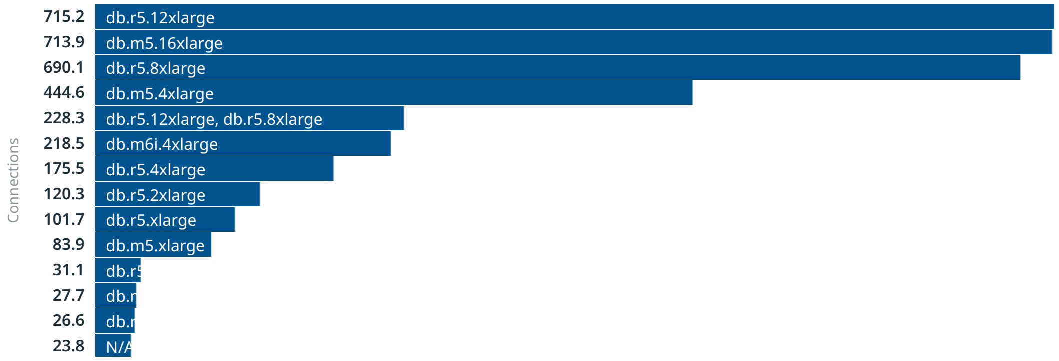
```
avg:aws.rds.database_connections{dbinstanceclass:db.m5.2xlarge} by {dbt_instance,dbinstanceidentifier}, avg:aws.rds.database_connections
```



	TAGS IN default_zero(avg:aws.rds.database_connections...	AVG	MIN	MAX	SUM	VALUE
🟦	dbinstanceidentifier:dbt-cloud-production-read-replica	924.27 conns	8.32 conns	1,294.43 conns	166,369 conns	1,089.36 conns
🟪	dbinstanceidentifier:aurora-codex-postges-2	862.20 conns	394.35 conns	1,267.83 conns	155,197 conns	1,004.00 conns
🟨	dbinstanceidentifier:aurora-codex-postges-1	559.97 conns	515.61 conns	707.13 conns	100,794 conns	556.85 conns
🟦	dbinstanceidentifier:dbt-cloud-production	510.25 conns	381.24 conns	671.99 conns	91,845 conns	502.01 conns

```
avg:aws.rds.database_connections{*} by {dbinstanceclass}
```

1mo Past 1 Month



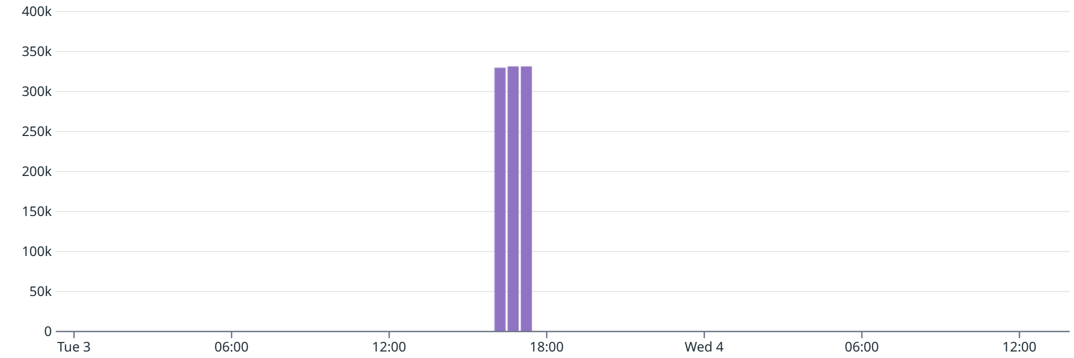
https://dbtlabsmt.datadoghq.com/monitors/156860309

- notify if missing data
- under sre improvements already.

- # Cloud IDE Availability SLO Monitor
  - Brief impact when ST instances were being routed to MT, but continues to appear in dbt_instance list due to SLO formula and default_zero calculating no impact / 100%
  - Manually created monitor. Further scope to terraform managed

trace.nginx.request.errors.by_http_status, trace.nginx.request.hits, kubernet...

2d Dec 2 - Dec 4


400k
350k
300k
250k
200k
150k
100k
50k
Tue 3
06:00
12:00
18:00
Wed 4
06:00
12:00