WHOOPS, THE NUMBERS ARE WRONG!

SCALING DATA QUALITY @ NETFLIX

MICHELLE UFFORD
DATA ENGINEERING & ANALYTICS, NETFLIX
HADOOP SUMMIT 2017

Overview.

The business.

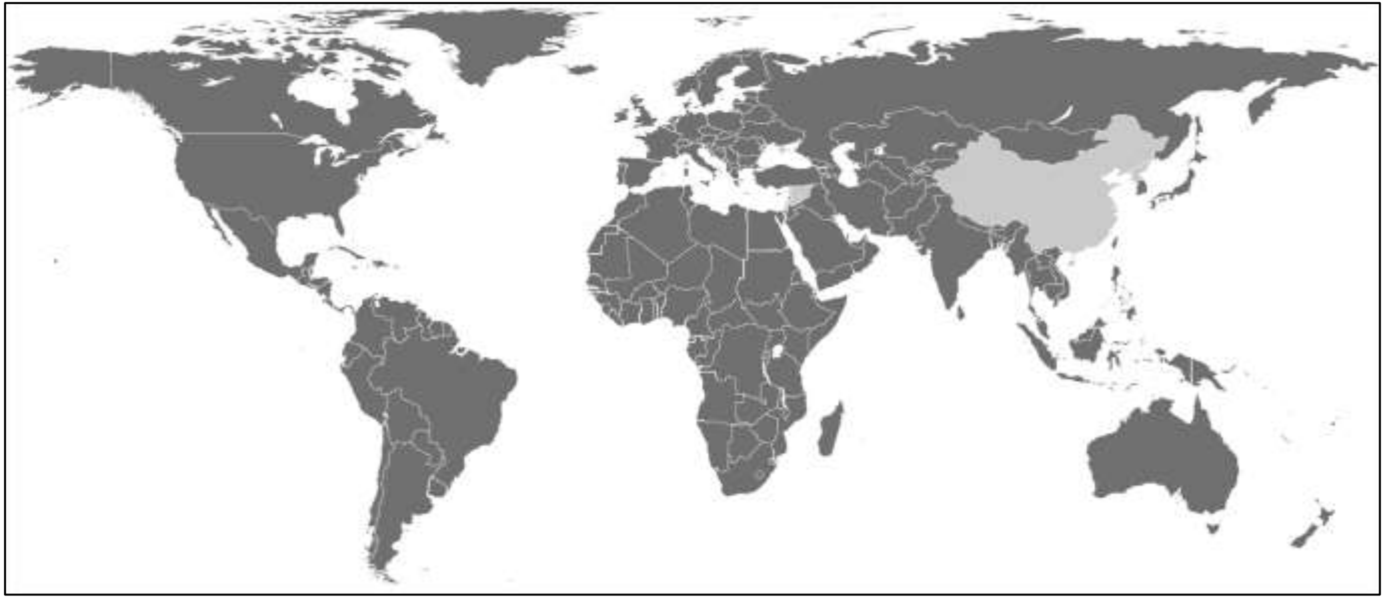
launched
in 1997

100+ million
members

$6 billion
on content

125+ million
hours watched
every. day.

Anytime. Anywhere.*



* Well, *almost* anywhere.

Any device.

NETFLIX ORIGINAL
THE
CROWN
Duty. Family. Power. A young woman rises to rule an empire amid the challenges of the modern age.
PLAY
NETFLIX ORIGINAL
THE CROWN
sense8
CHRIS TUCKER
LIVE

The data.

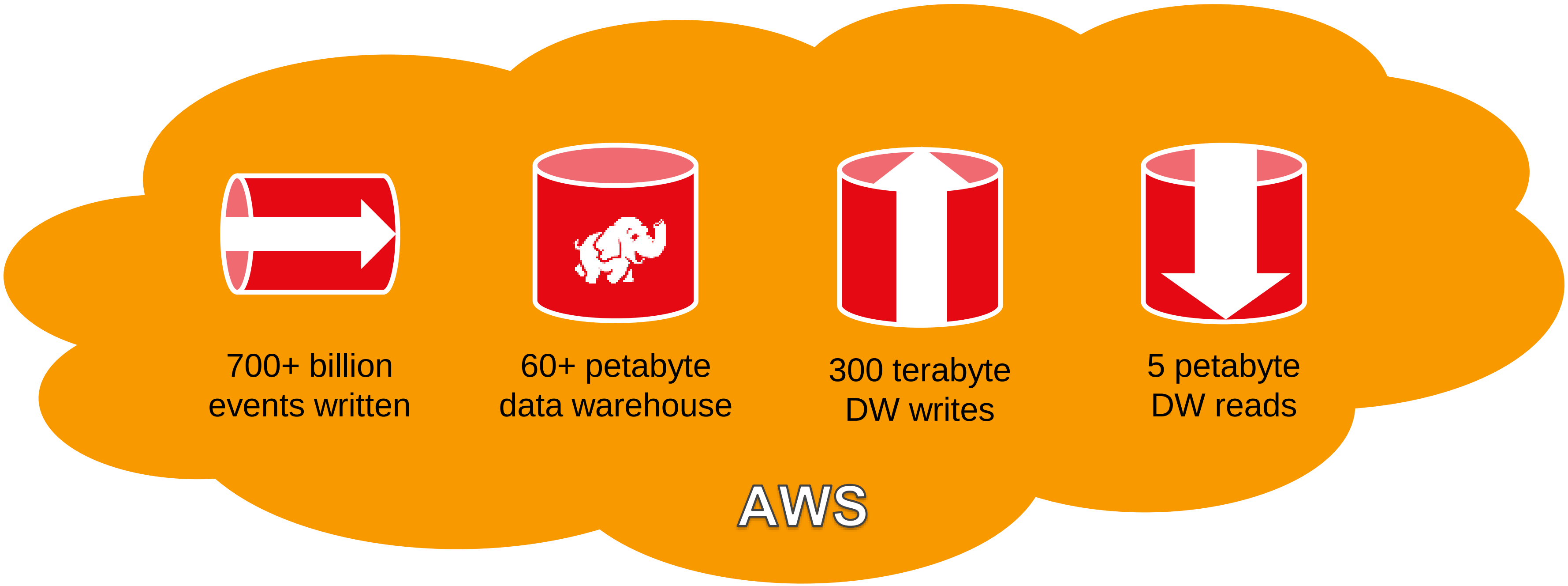
700+ billion
events written

60+ petabyte
data warehouse

300 terabyte
DW writes

5 petabyte
DW reads

The data.



700+ billion events written

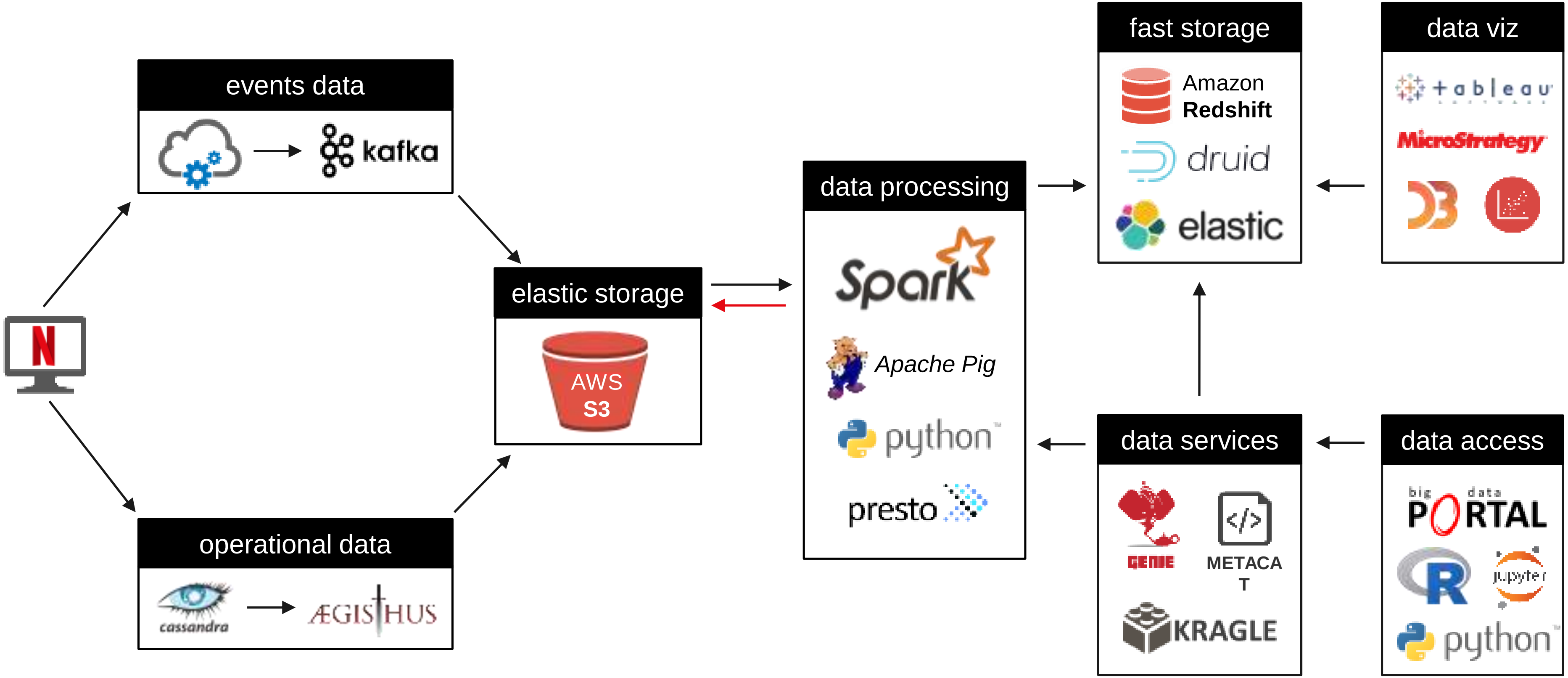
60+ petabyte data warehouse

300 terabyte DW writes

5 petabyte DW reads

AWS

Big Data Platform



Data Quality.

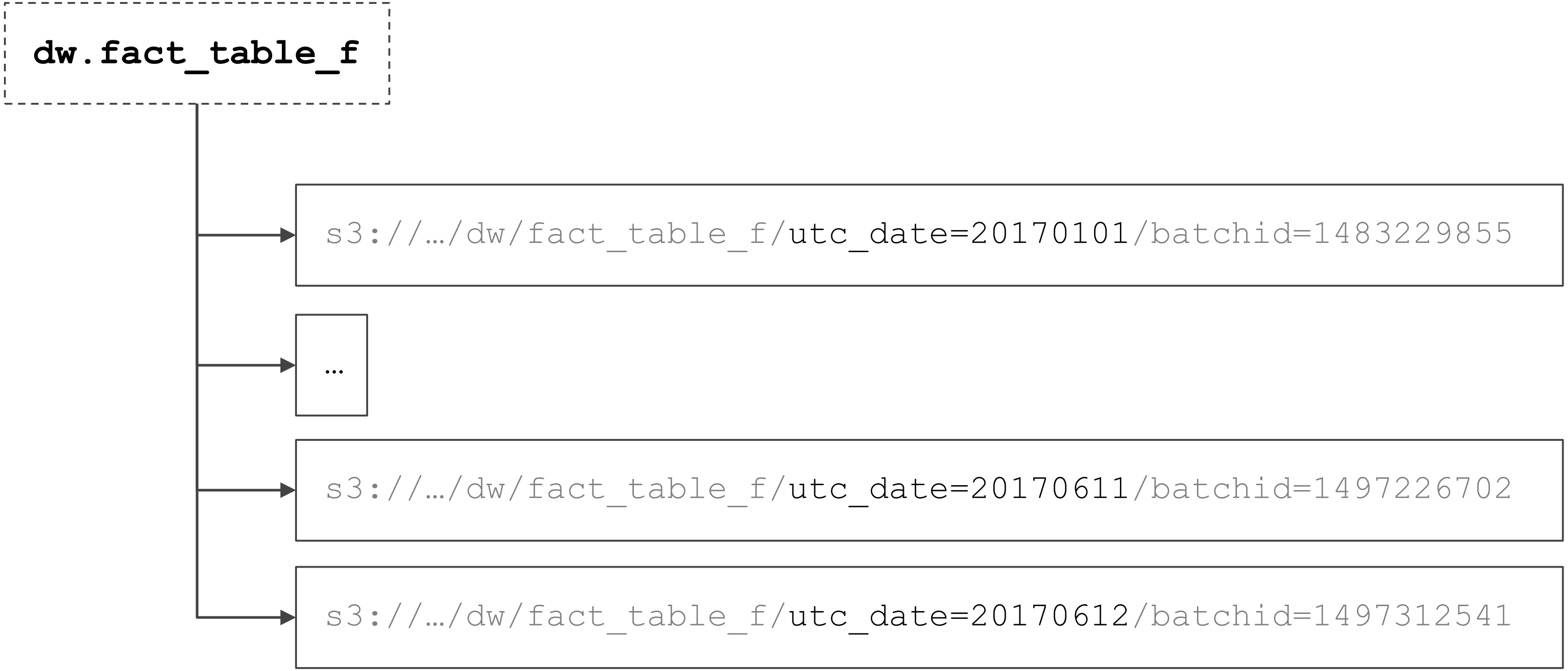
Behind the Scenes.

Metacat.

Federated metastore & extensible data catalog

Metacat

Federated Metastore



dw.fact_table_f

s3://…/dw/fact_table_f/utc_date=20170101/batchid=1483229855

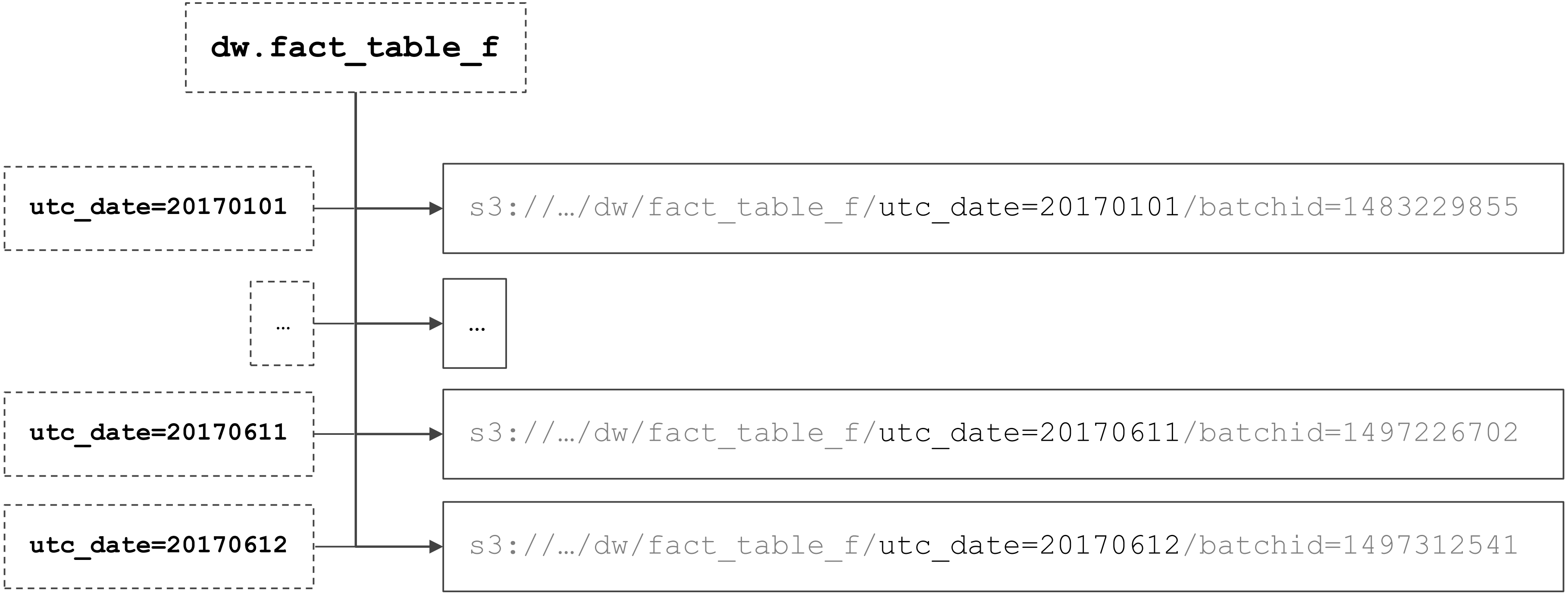
…

s3://…/dw/fact_table_f/utc_date=20170611/batchid=1497226702

s3://…/dw/fact_table_f/utc_date=20170612/batchid=1497312541

Metacat

Federated Metastore



dw.fact_table_f

utc_date=20170101 → s3://…/dw/fact_table_f/utc_date=20170101/batchid=1483229855

… → …

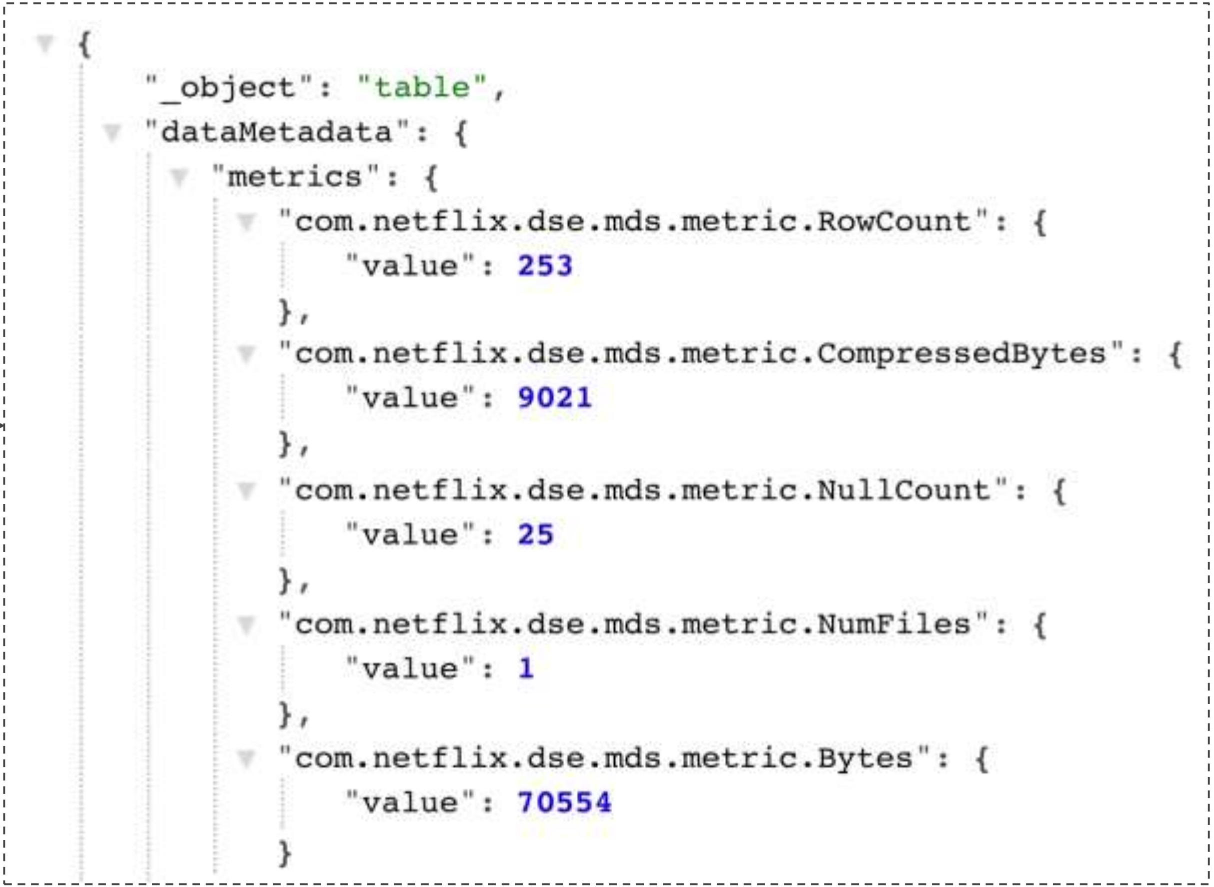
utc_date=20170611 → s3://…/dw/fact_table_f/utc_date=20170611/batchid=1497226702

utc_date=20170612 → s3://…/dw/fact_table_f/utc_date=20170612/batchid=1497312541

Metacat

Federated Metastore

`utc_date=20170101`

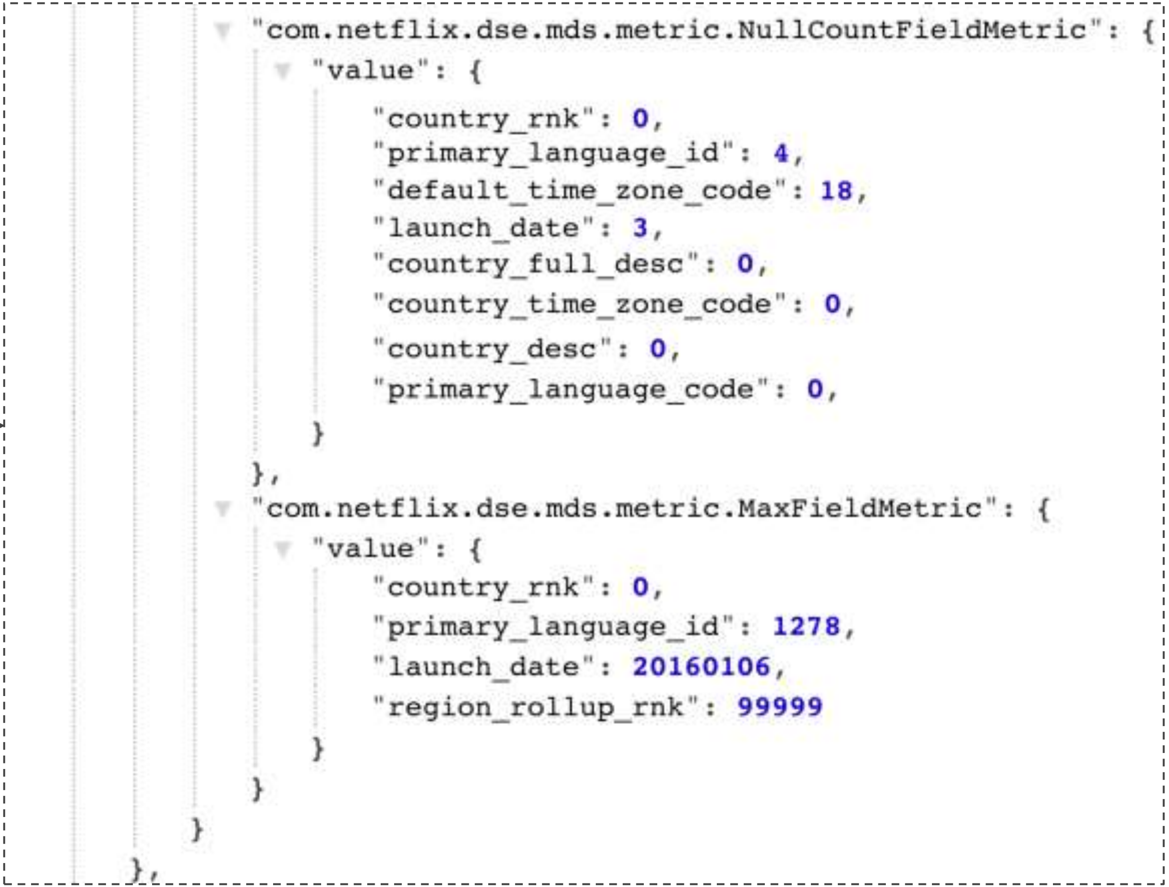


```
{
    "_object": "table",
    "dataMetadata": {
        "metrics": {
            "com.netflix.dse.mds.metric.RowCount": {
                "value": 253
            },
            "com.netflix.dse.mds.metric.CompressedBytes": {
                "value": 9021
            },
            "com.netflix.dse.mds.metric.NullCount": {
                "value": 25
            },
            "com.netflix.dse.mds.metric.NumFiles": {
                "value": 1
            },
            "com.netflix.dse.mds.metric.Bytes": {
                "value": 70554
            }
```

Metacat

Federated Metastore

`utc_date=20170101`



```
        "com.netflix.dse.mds.metric.NullCountFieldMetric": {
            "value": {
                "country_rnk": 0,
                "primary_language_id": 4,
                "default_time_zone_code": 18,
                "launch_date": 3,
                "country_full_desc": 0,
                "country_time_zone_code": 0,
                "country_desc": 0,
                "primary_language_code": 0,
            }
        },
        "com.netflix.dse.mds.metric.MaxFieldMetric": {
            "value": {
                "country_rnk": 0,
                "primary_language_id": 1278,
                "launch_date": 20160106,
                "region_rollup_rnk": 99999
            }
        }
    }
},
```

Metacat

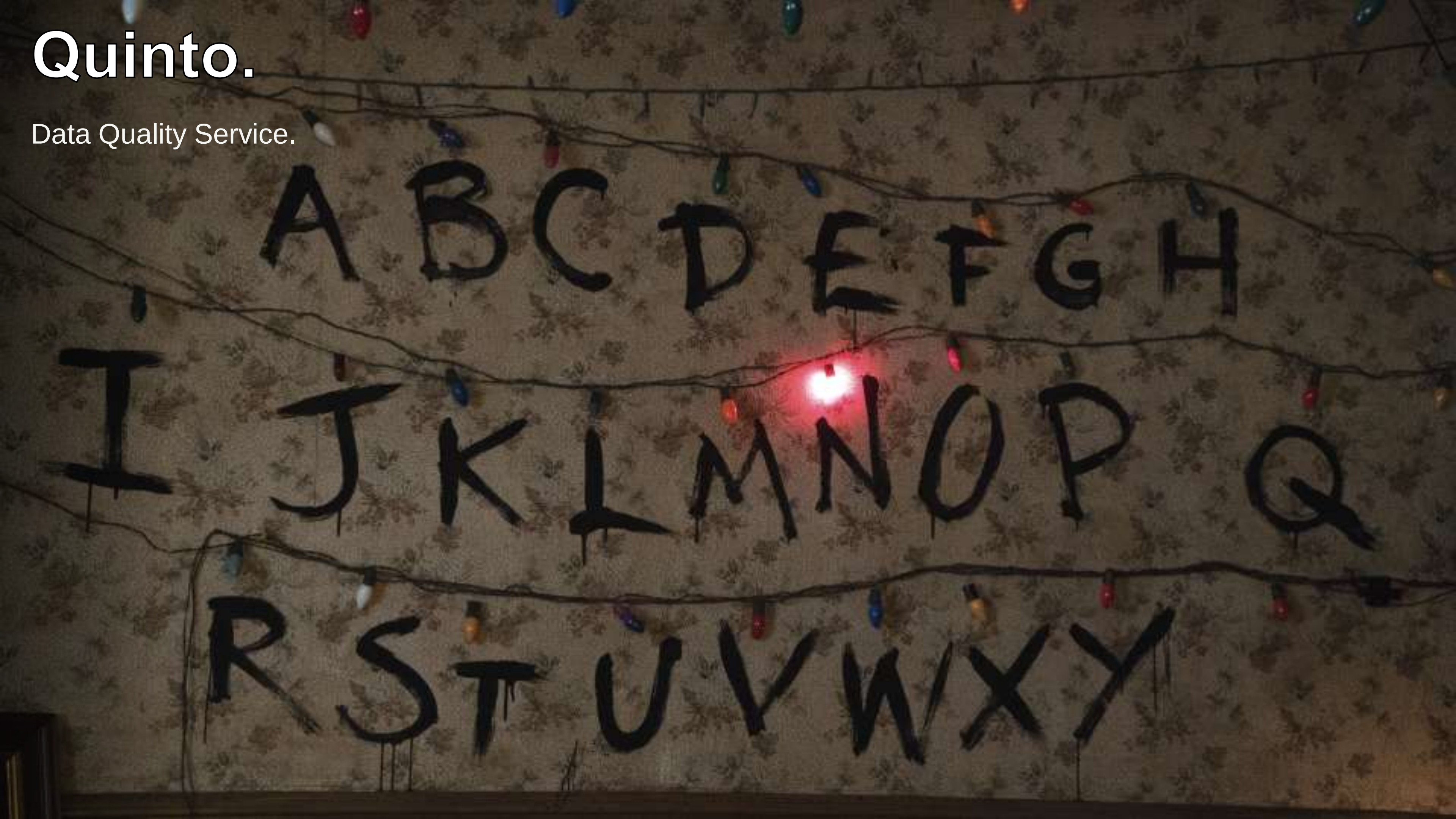
Federated Metastore

Extended table attributes

- primary key(s)
- column types
- lifecycle
- audience
- “valid-thru” timestamp
- … and much more

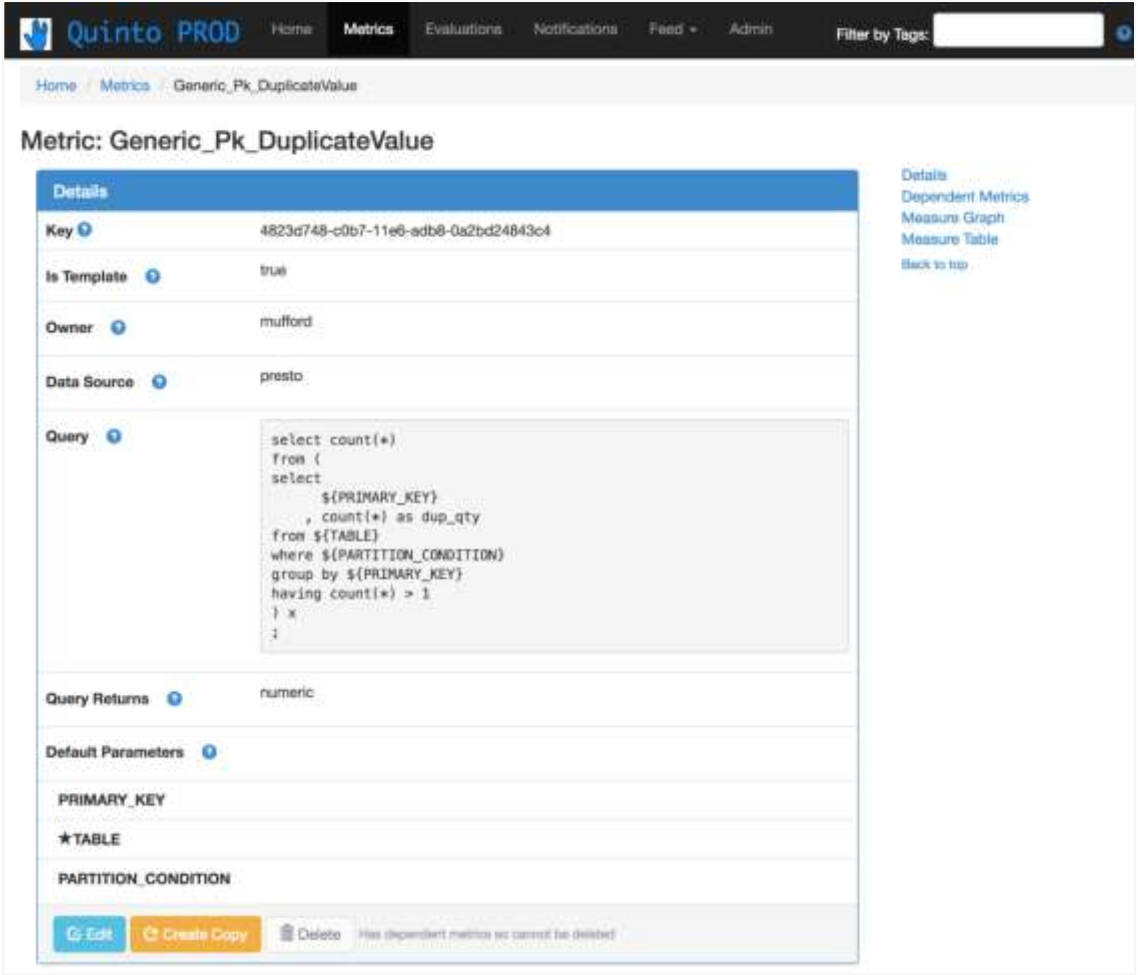
Quinto.

Data Quality Service.



Quinto

Data Quality Service



Metric: Generic_Pk_DuplicateValue

Details

Key	4823d748-c0b7-11e6-adb8-0a2bd24843c4
Is Template	true
Owner	mufford
Data Source	presto
Query	(below)
Query Returns	numeric
Default Parameters	

```
select count(*)
from (
select
      ${PRIMARY_KEY}
    , count(*) as dup_qty
from ${TABLE}
where ${PARTITION_CONDITION}
group by ${PRIMARY_KEY}
having count(*) > 1
) x
;
```

PRIMARY_KEY

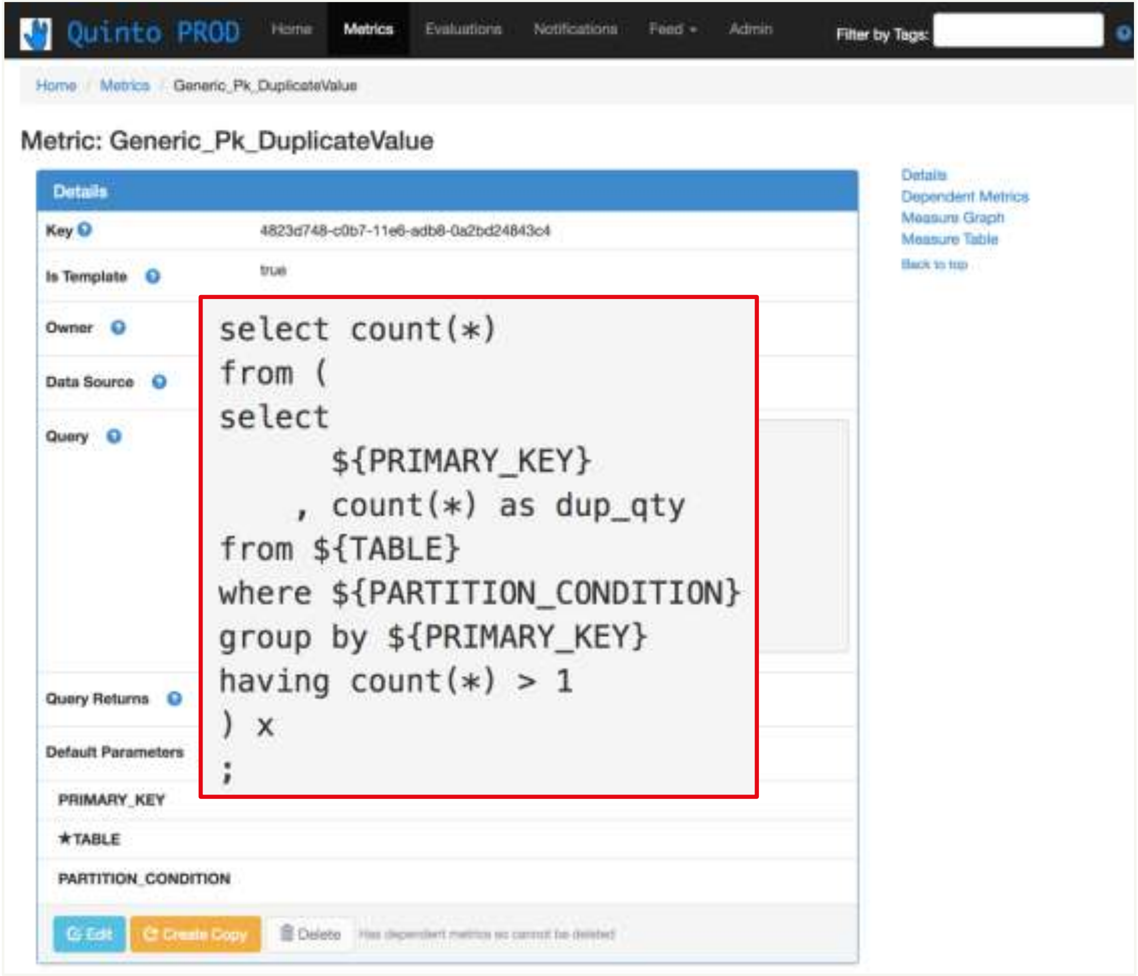
★TABLE

PARTITION_CONDITION

Has dependent metrics so cannot be deleted

Quinto

Data Quality Service



Metric: Generic_Pk_DuplicateValue

Details

Key: 4823d748-c0b7-11e6-adb8-0a2bd24843c4

Is Template: true

Owner

Data Source

Query

```
select count(*)
from (
select
      ${PRIMARY_KEY}
    , count(*) as dup_qty
from ${TABLE}
where ${PARTITION_CONDITION}
group by ${PRIMARY_KEY}
having count(*) > 1
) x
;
```

Query Returns

Default Parameters

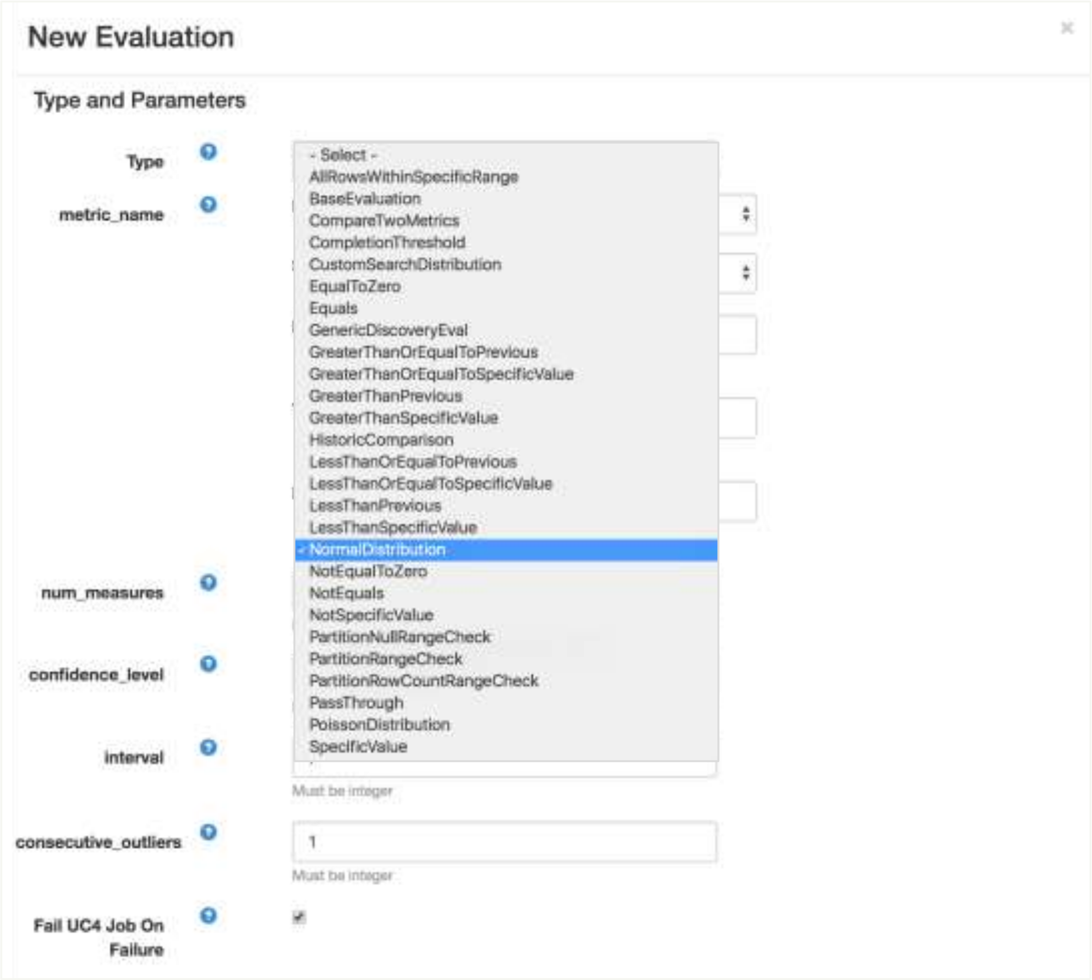
PRIMARY_KEY

★TABLE

PARTITION_CONDITION

Quinto

Data Quality Service



New Evaluation

Type and Parameters

Type: ✓ NormalDistribution

metric_name

num_measures

confidence_level

interval

Must be integer

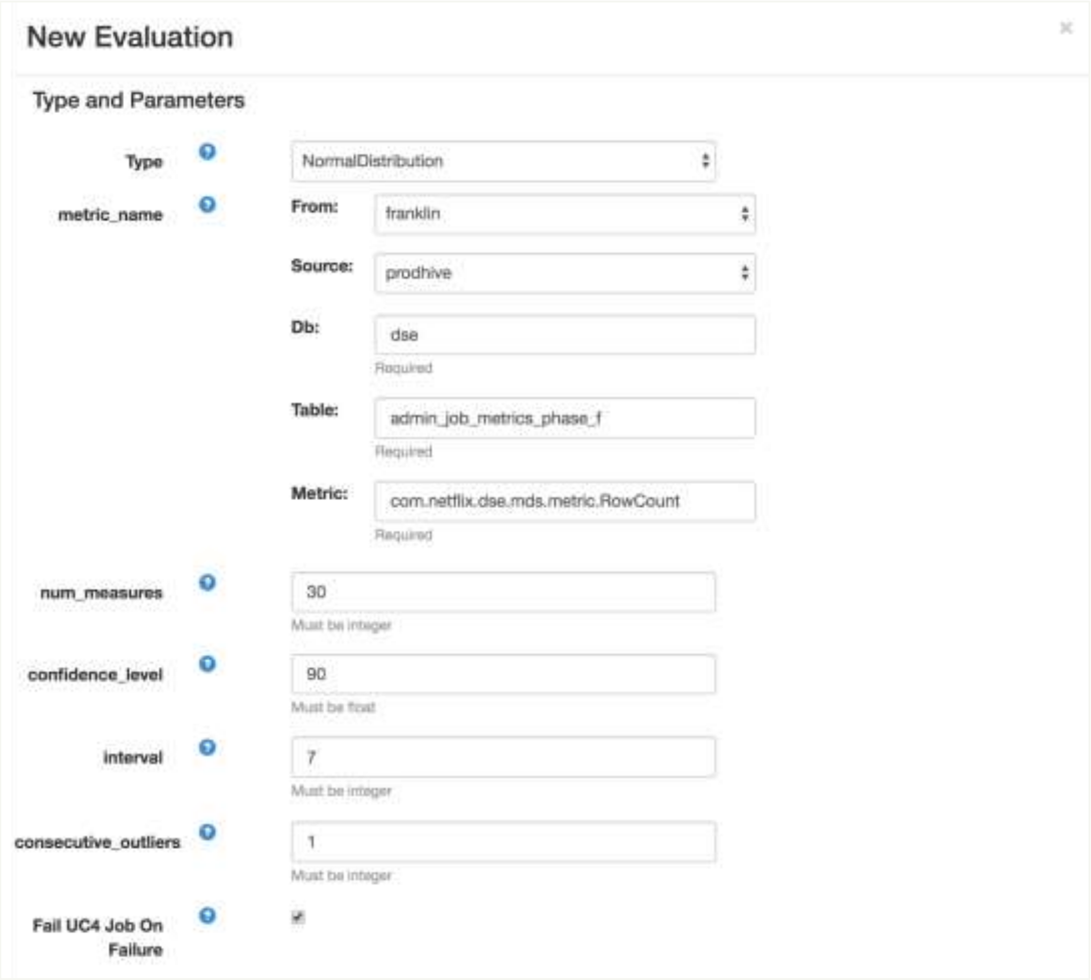
consecutive_outliers: 1

Must be integer

Fail UC4 Job On Failure: ☑

Quinto

Data Quality Service



New Evaluation

×

Type and Parameters

Type: NormalDistribution

metric_name:

From: franklin

Source: prodhive

Db: dse

Required

Table: admin_job_metrics_phase_f

Required

Metric: com.netflix.dse.mds.metric.RowCount

Required

num_measures: 30

Must be integer

confidence_level: 90

Must be float

interval: 7

Must be integer

consecutive_outliers: 1

Must be integer

Fail UC4 Job On Failure: ☑

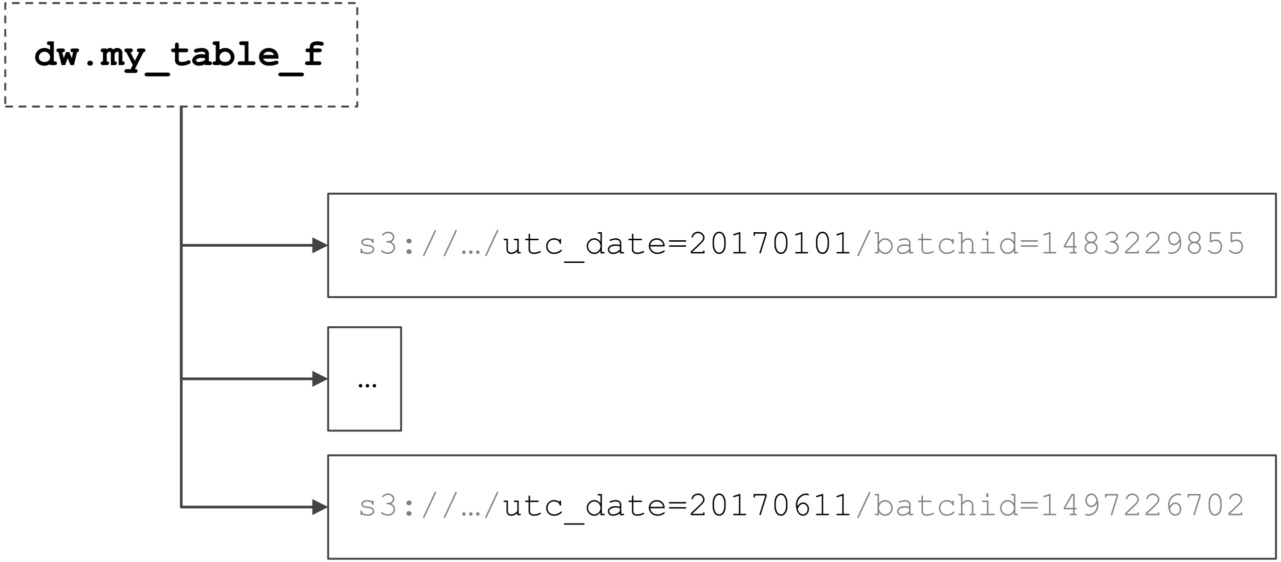
WAP

Write - Audit - Publish

ETL pattern
for high-quality
big data jobs

WAP

ETL Pattern



dw.my_table_f

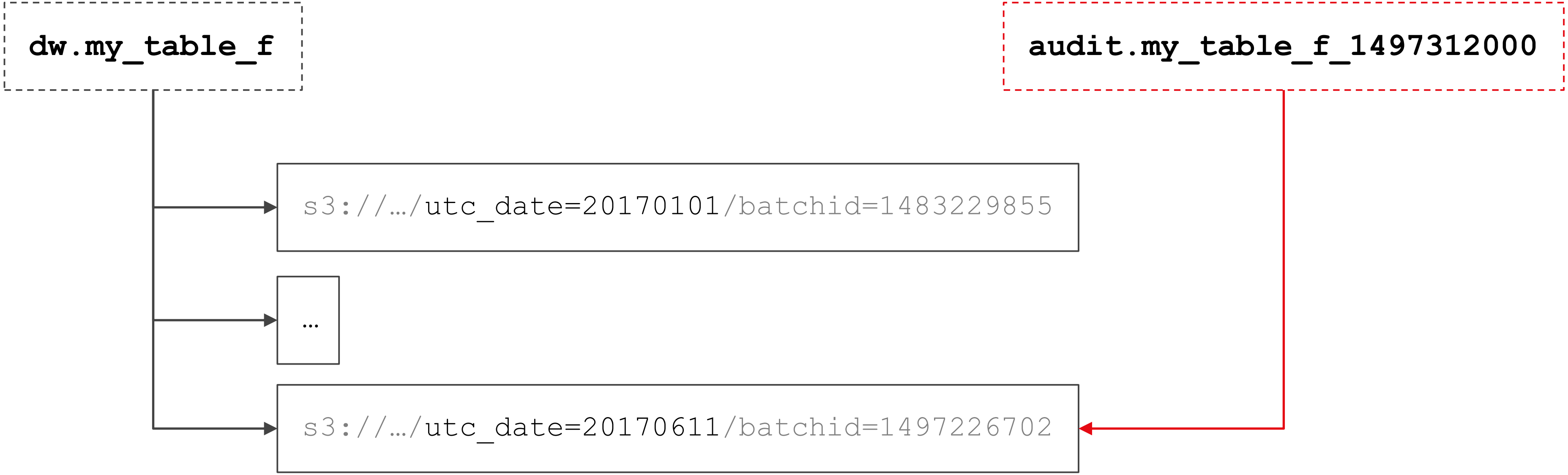
s3://…/utc_date=20170101/batchid=1483229855

…

s3://…/utc_date=20170611/batchid=1497226702

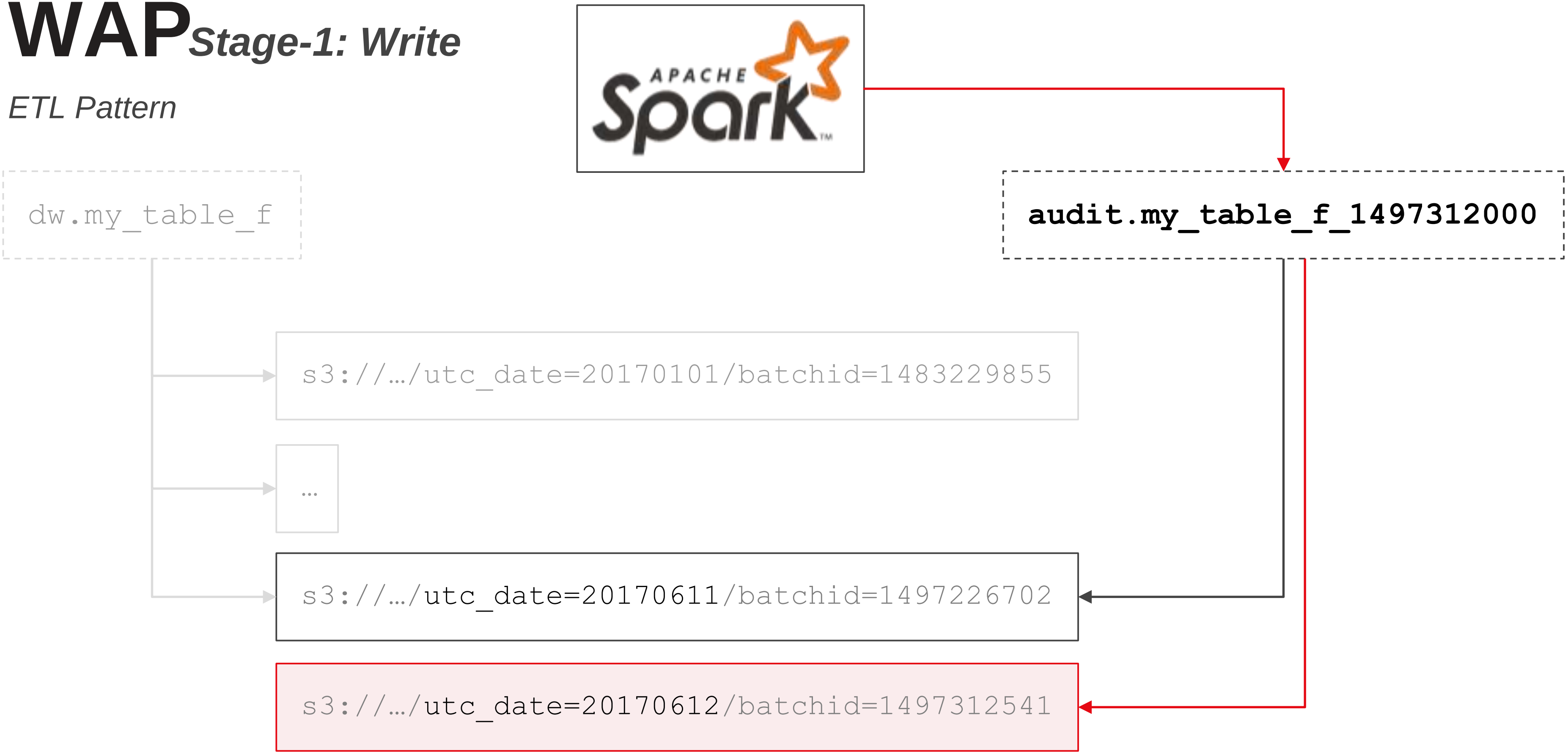
WAP *Stage-0: Prep*

ETL Pattern


dw.my_table_f
audit.my_table_f_1497312000
s3://…/utc_date=20170101/batchid=1483229855
…
s3://…/utc_date=20170611/batchid=1497226702

WAP *Stage-1: Write*

ETL Pattern



dw.my_table_f

audit.my_table_f_1497312000

s3://…/utc_date=20170101/batchid=1483229855

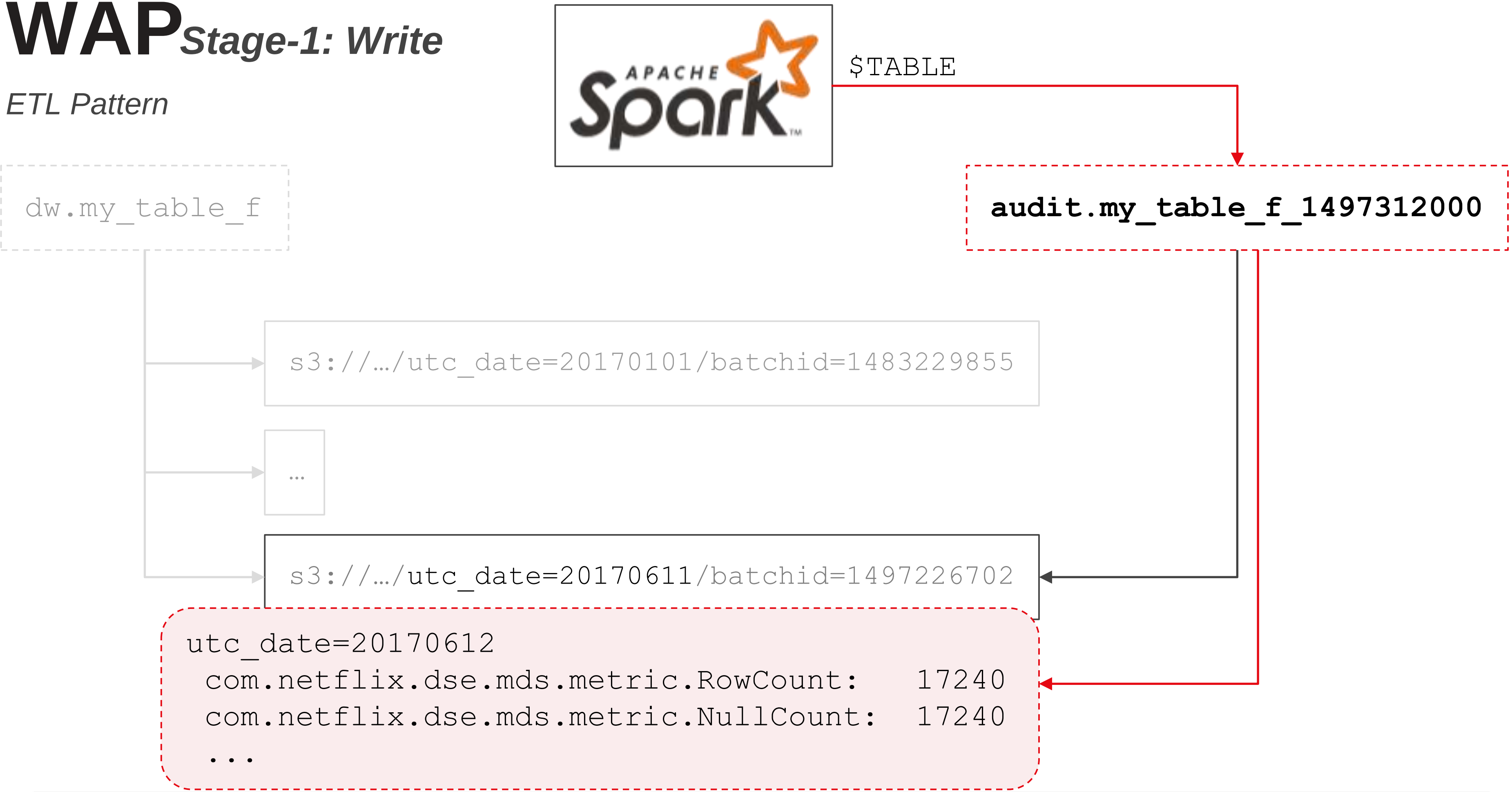
…

s3://…/utc_date=20170611/batchid=1497226702

s3://…/utc_date=20170612/batchid=1497312541

WAP *Stage-1: Write*

ETL Pattern

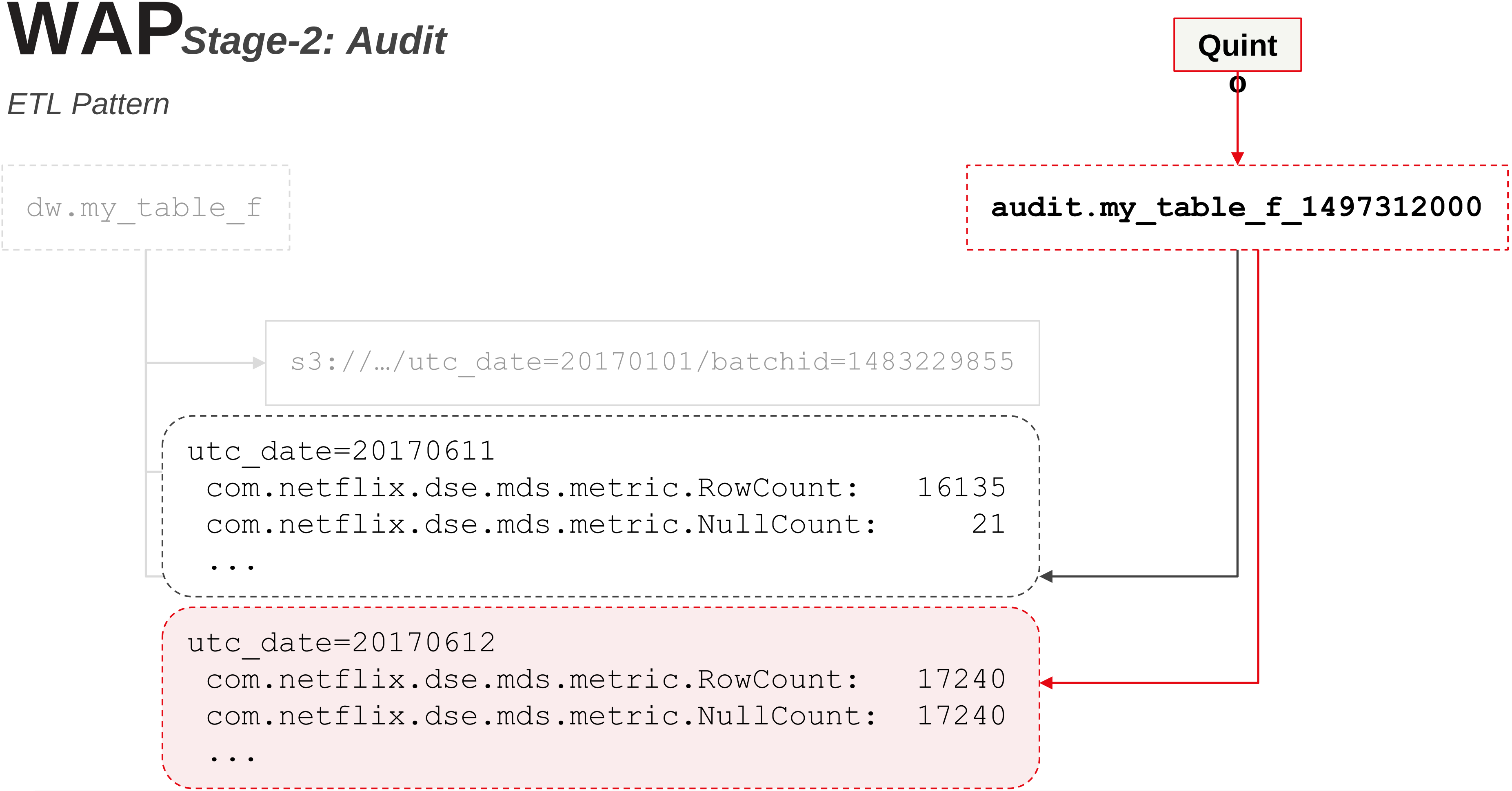


APACHE Spark
$TABLE
dw.my_table_f
audit.my_table_f_1497312000
s3://…/utc_date=20170101/batchid=1483229855
…
s3://…/utc_date=20170611/batchid=1497226702

```
utc_date=20170612
 com.netflix.dse.mds.metric.RowCount:    17240
 com.netflix.dse.mds.metric.NullCount:   17240
 ...
```

WAP *Stage-2: Audit*

ETL Pattern



Quinto

audit.my_table_f_1497312000

dw.my_table_f

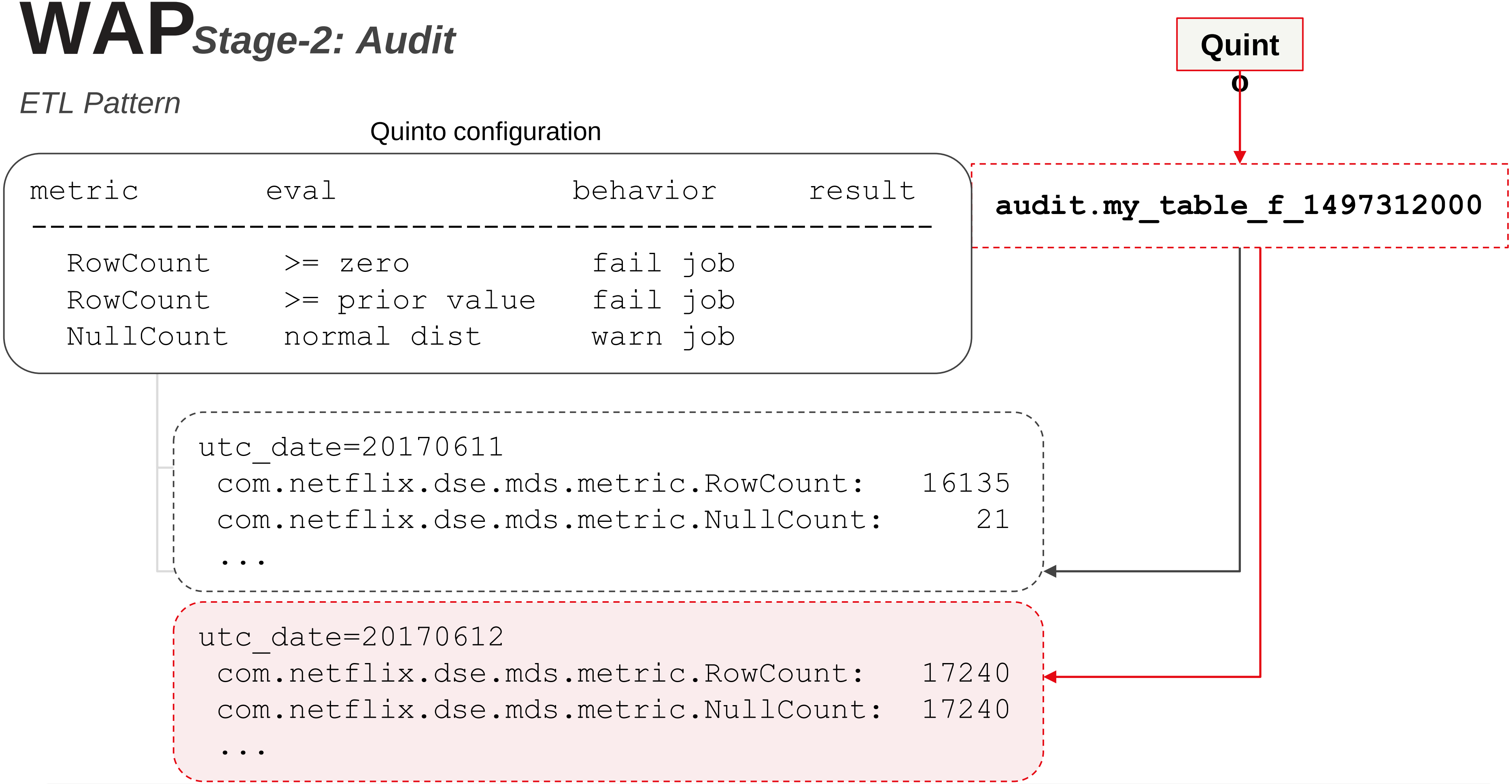
s3://…/utc_date=20170101/batchid=1483229855

```
utc_date=20170611
 com.netflix.dse.mds.metric.RowCount:   16135
 com.netflix.dse.mds.metric.NullCount:     21
 ...
```

```
utc_date=20170612
 com.netflix.dse.mds.metric.RowCount:   17240
 com.netflix.dse.mds.metric.NullCount:  17240
 ...
```

WAP *Stage-2: Audit*

ETL Pattern



Quinto configuration

```
metric          eval                behavior      result
-------------------------------------------------------------
  RowCount     >= zero              fail job
  RowCount     >= prior value       fail job
  NullCount    normal dist          warn job
```

Quint
o

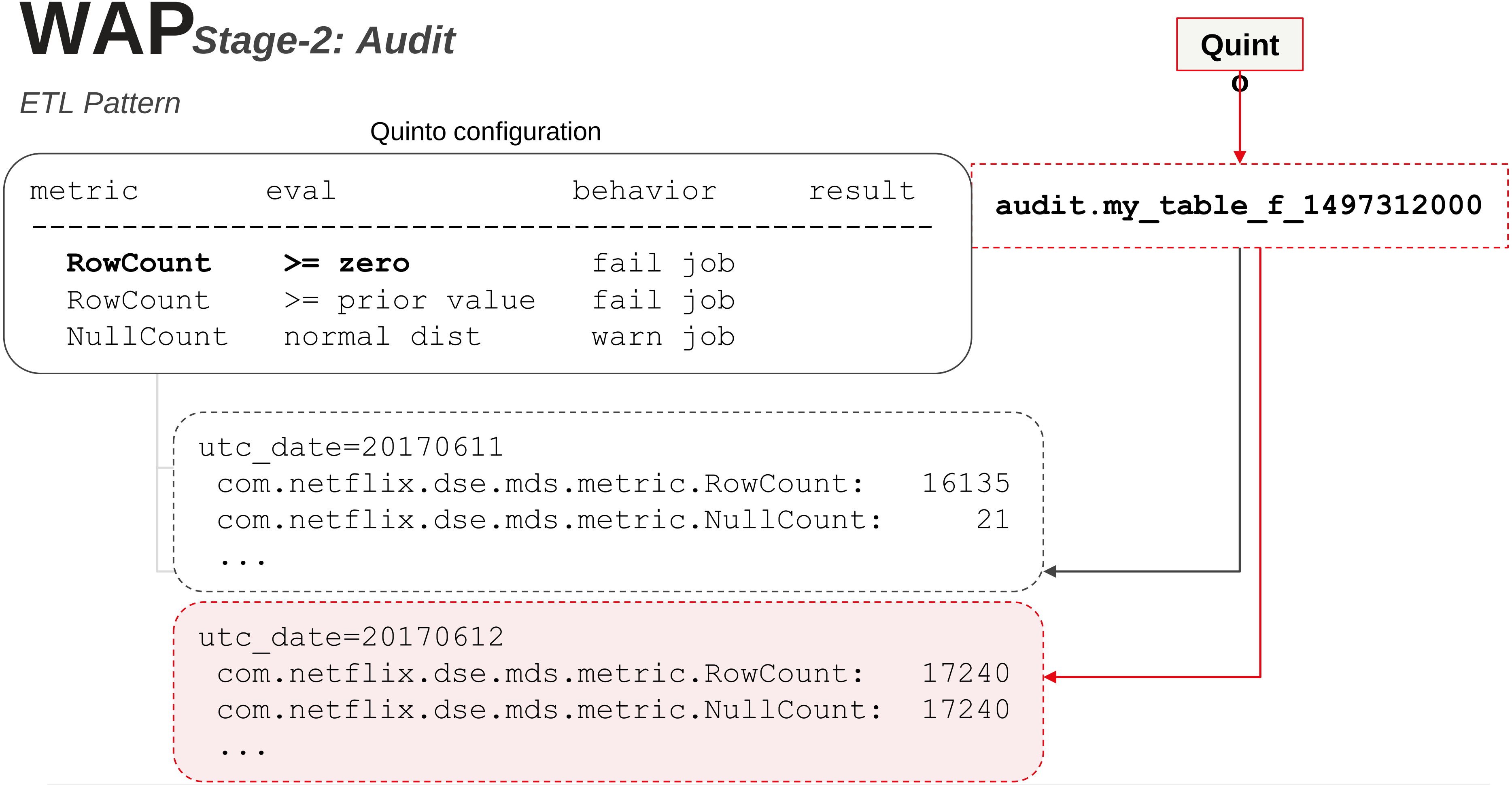
audit.my_table_f_1497312000

```
utc_date=20170611
 com.netflix.dse.mds.metric.RowCount:   16135
 com.netflix.dse.mds.metric.NullCount:     21
 ...
```

```
utc_date=20170612
 com.netflix.dse.mds.metric.RowCount:   17240
 com.netflix.dse.mds.metric.NullCount:  17240
 ...
```

WAP *Stage-2: Audit*

ETL Pattern



Quinto configuration

metric	eval	behavior	result
RowCount	**>= zero**	fail job	
RowCount	>= prior value	fail job	
NullCount	normal dist	warn job	

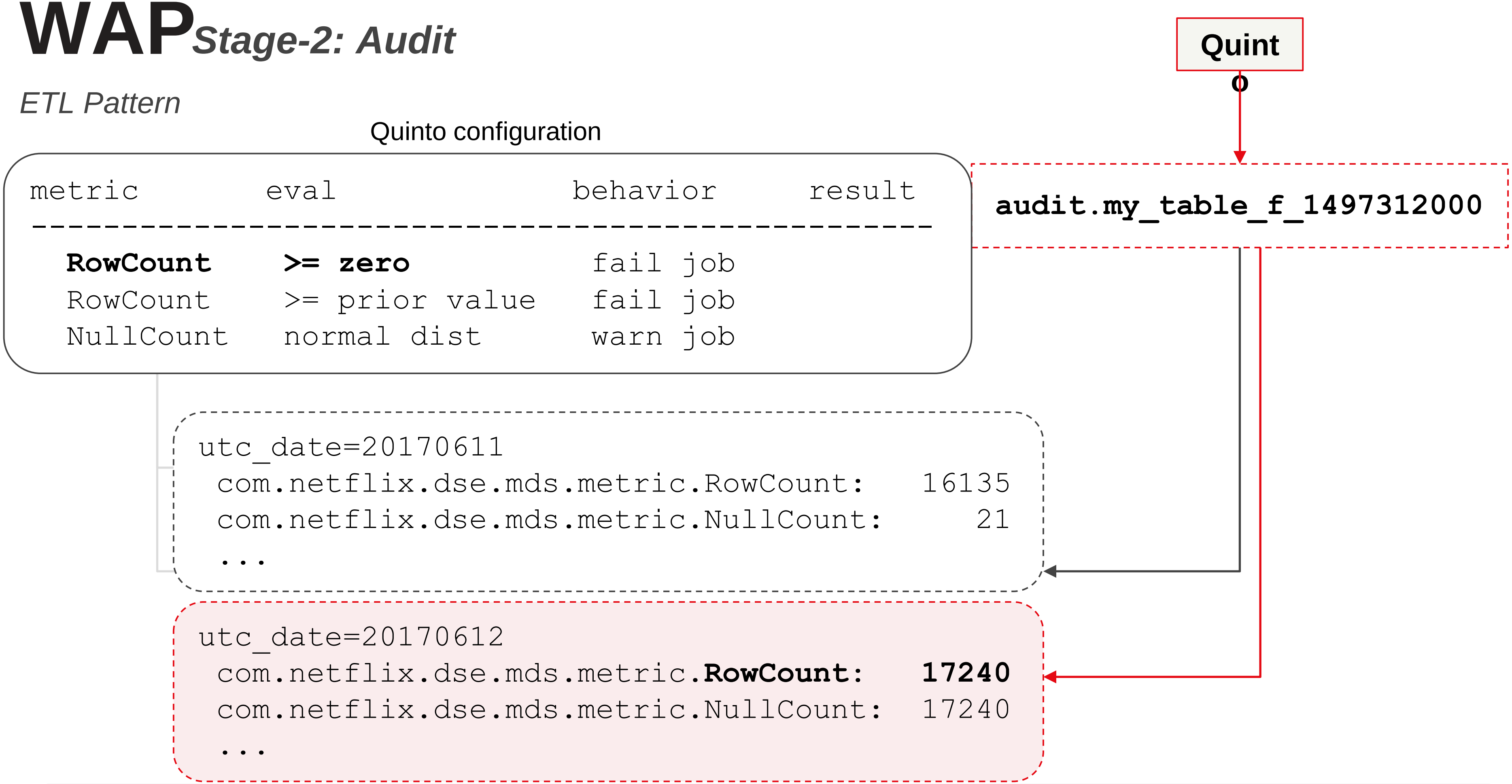
audit.my_table_f_1497312000

```
utc_date=20170611
 com.netflix.dse.mds.metric.RowCount:   16135
 com.netflix.dse.mds.metric.NullCount:     21
 ...
```

```
utc_date=20170612
 com.netflix.dse.mds.metric.RowCount:   17240
 com.netflix.dse.mds.metric.NullCount:  17240
 ...
```

WAP*Stage-2: Audit*

ETL Pattern



Quinto configuration

```
metric          eval                behavior        result
--------------------------------------------------------------
  **RowCount      >= zero**             fail job
  RowCount      >= prior value      fail job
  NullCount     normal dist         warn job
```

Quint
o

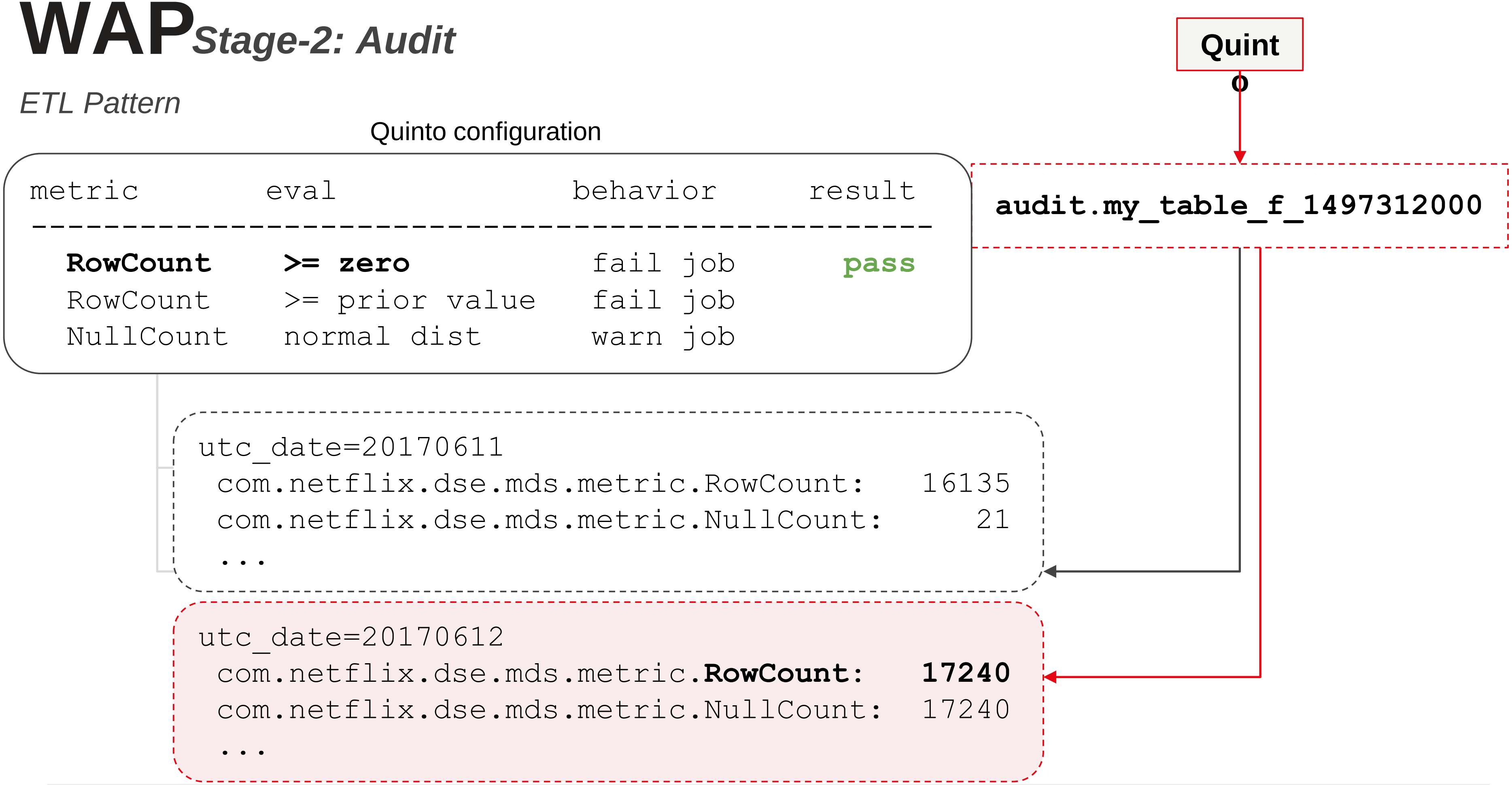
audit.my_table_f_1497312000

```
utc_date=20170611
 com.netflix.dse.mds.metric.RowCount:    16135
 com.netflix.dse.mds.metric.NullCount:      21
 ...
```

```
utc_date=20170612
 com.netflix.dse.mds.metric.RowCount:    17240
 com.netflix.dse.mds.metric.NullCount:   17240
 ...
```

WAP *Stage-2: Audit*

ETL Pattern



Quinto configuration

```
metric          eval                behavior      result
--------------------------------------------------------
  RowCount      >= zero               fail job       pass
  RowCount      >= prior value        fail job
  NullCount     normal dist           warn job
```

Quint
o

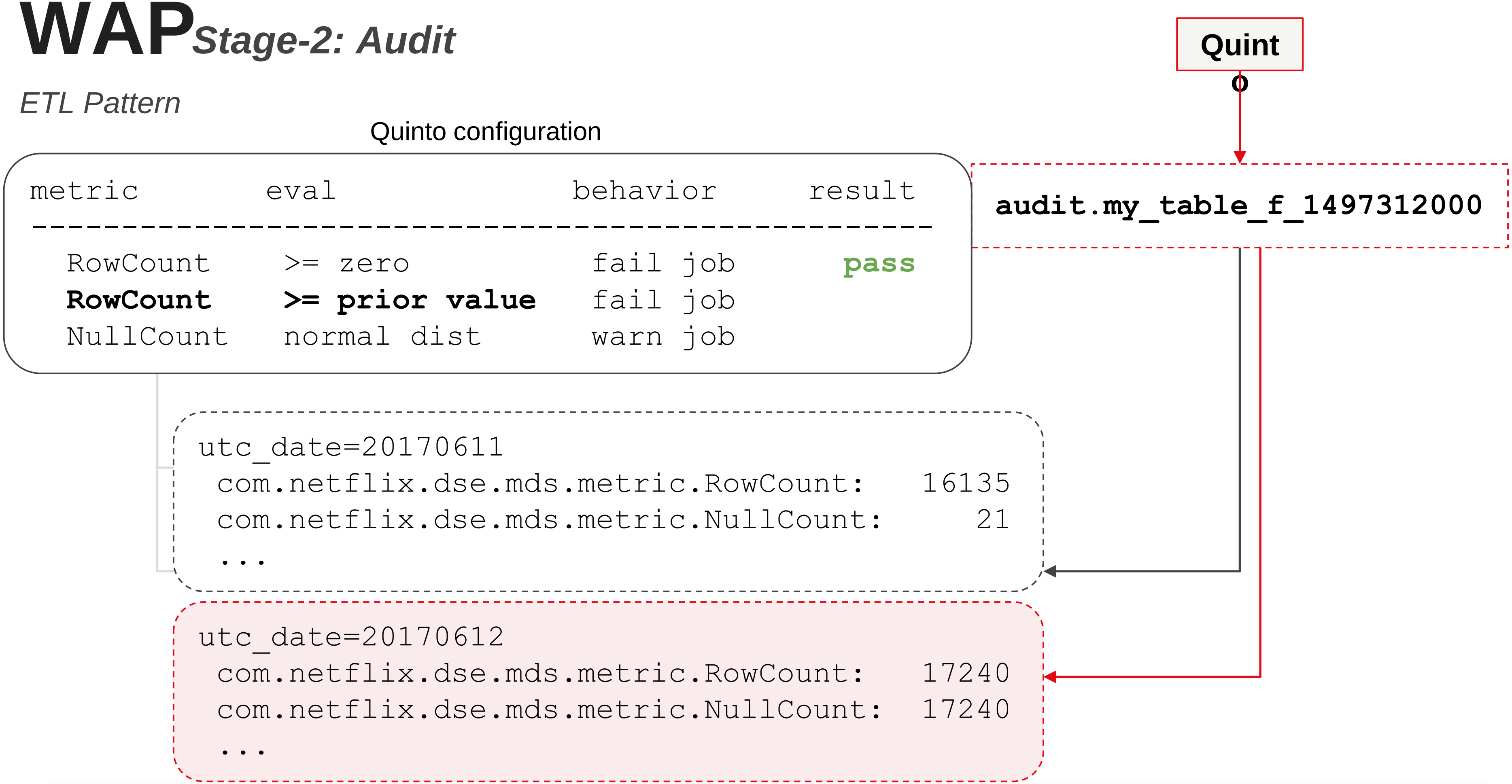
audit.my_table_f_1497312000

```
utc_date=20170611
 com.netflix.dse.mds.metric.RowCount:    16135
 com.netflix.dse.mds.metric.NullCount:      21
 ...
```

```
utc_date=20170612
 com.netflix.dse.mds.metric.RowCount:    17240
 com.netflix.dse.mds.metric.NullCount:   17240
 ...
```

WAP*Stage-2: Audit*

ETL Pattern



Quinto configuration

```
metric         eval                 behavior     result
-----------------------------------------------------------
  RowCount      >= zero              fail job       pass
  RowCount      >= prior value       fail job
  NullCount     normal dist          warn job
```

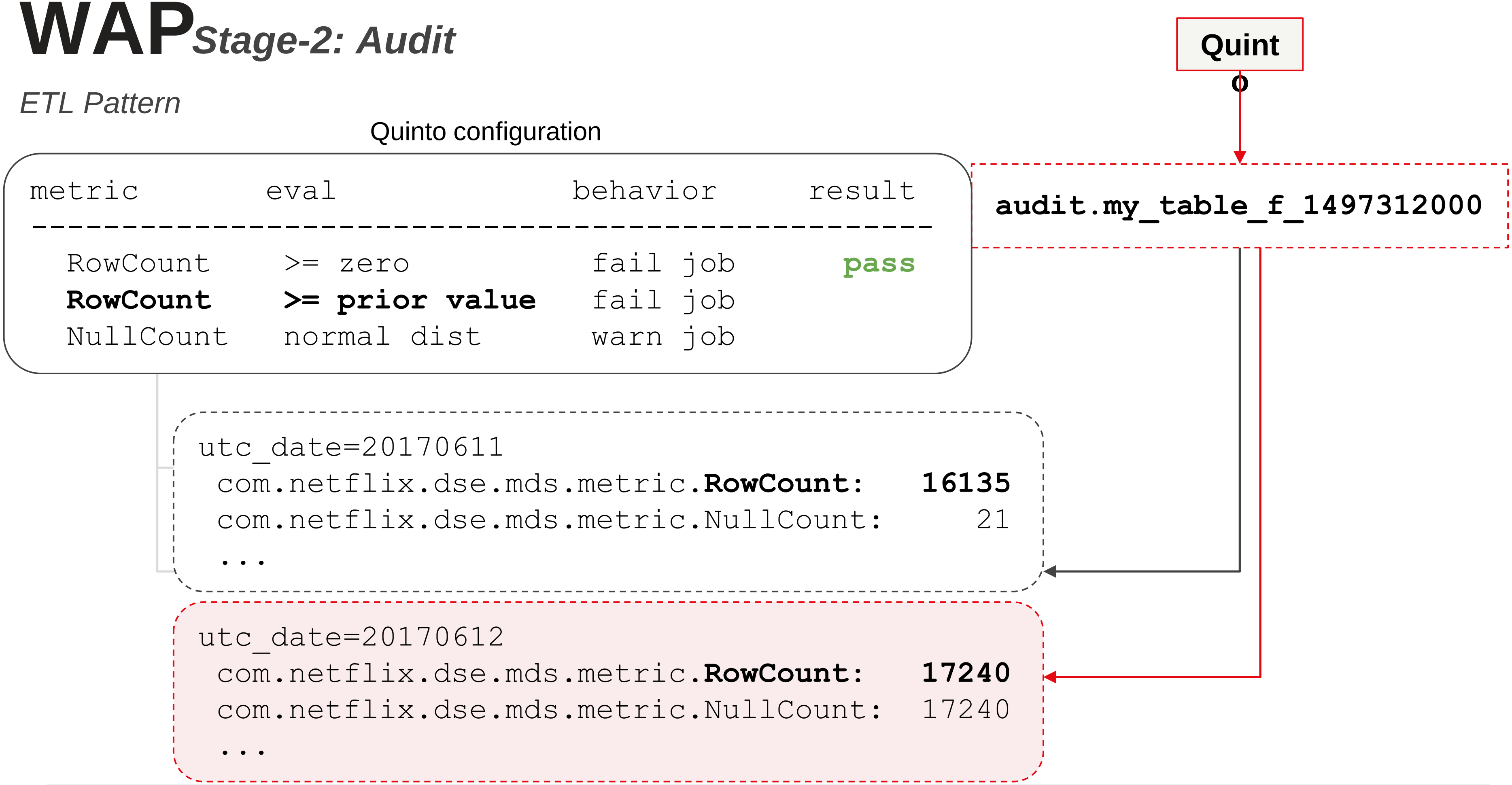
audit.my_table_f_1497312000

```
utc_date=20170611
 com.netflix.dse.mds.metric.RowCount:   16135
 com.netflix.dse.mds.metric.NullCount:     21
 ...
```

```
utc_date=20170612
 com.netflix.dse.mds.metric.RowCount:   17240
 com.netflix.dse.mds.metric.NullCount:  17240
 ...
```

WAP*Stage-2: Audit*

ETL Pattern



```
metric          eval                behavior      result
---------------------------------------------------------
  RowCount      >= zero              fail job        pass
  RowCount      >= prior value       fail job
  NullCount     normal dist          warn job
```

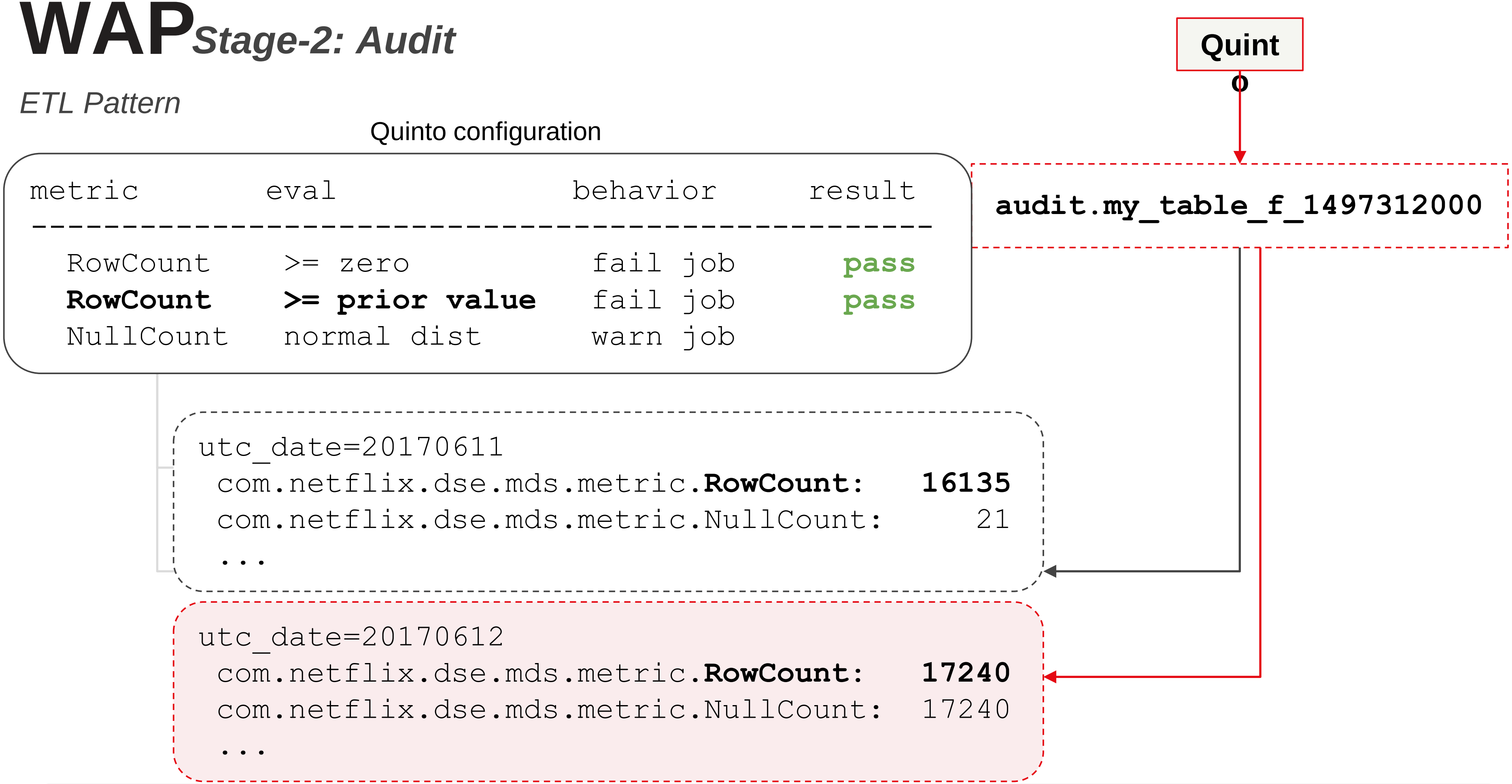
```
audit.my_table_f_1497312000
```

```
utc_date=20170611
 com.netflix.dse.mds.metric.RowCount:    16135
 com.netflix.dse.mds.metric.NullCount:      21
 ...
```

```
utc_date=20170612
 com.netflix.dse.mds.metric.RowCount:    17240
 com.netflix.dse.mds.metric.NullCount:   17240
 ...
```

WAP*Stage-2: Audit*

ETL Pattern



Quinto configuration

```
metric          eval                 behavior     result
--------------------------------------------------------
  RowCount      >= zero               fail job     pass
  RowCount      >= prior value        fail job     pass
  NullCount     normal dist           warn job
```

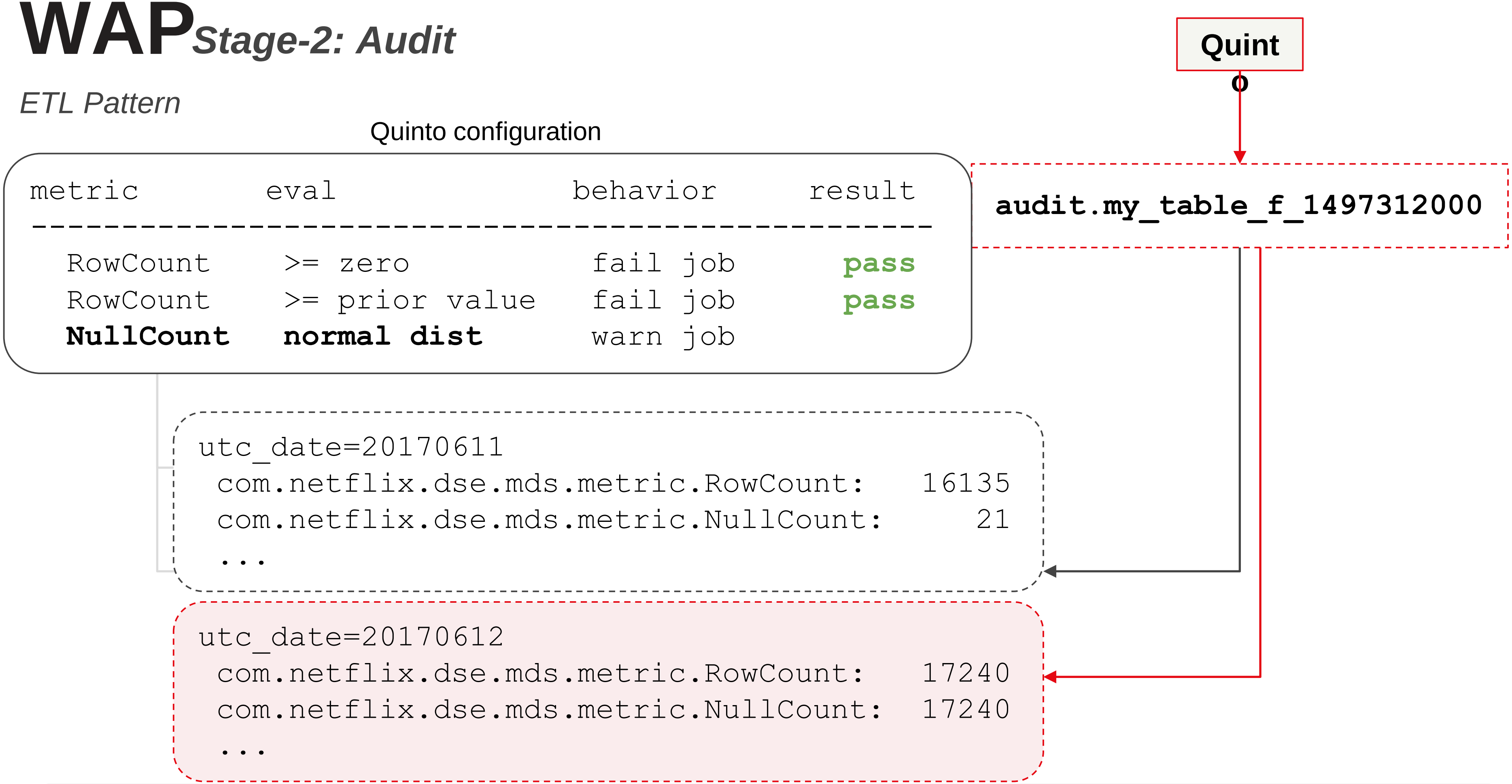
audit.my_table_f_1497312000

```
utc_date=20170611
 com.netflix.dse.mds.metric.RowCount:    16135
 com.netflix.dse.mds.metric.NullCount:      21
 ...
```

```
utc_date=20170612
 com.netflix.dse.mds.metric.RowCount:    17240
 com.netflix.dse.mds.metric.NullCount:   17240
 ...
```

WAP *Stage-2: Audit*

ETL Pattern



Quinto configuration

```
metric         eval               behavior     result
----------------------------------------------------
  RowCount     >= zero             fail job      pass
  RowCount     >= prior value      fail job      pass
  NullCount    normal dist         warn job
```

Quinto

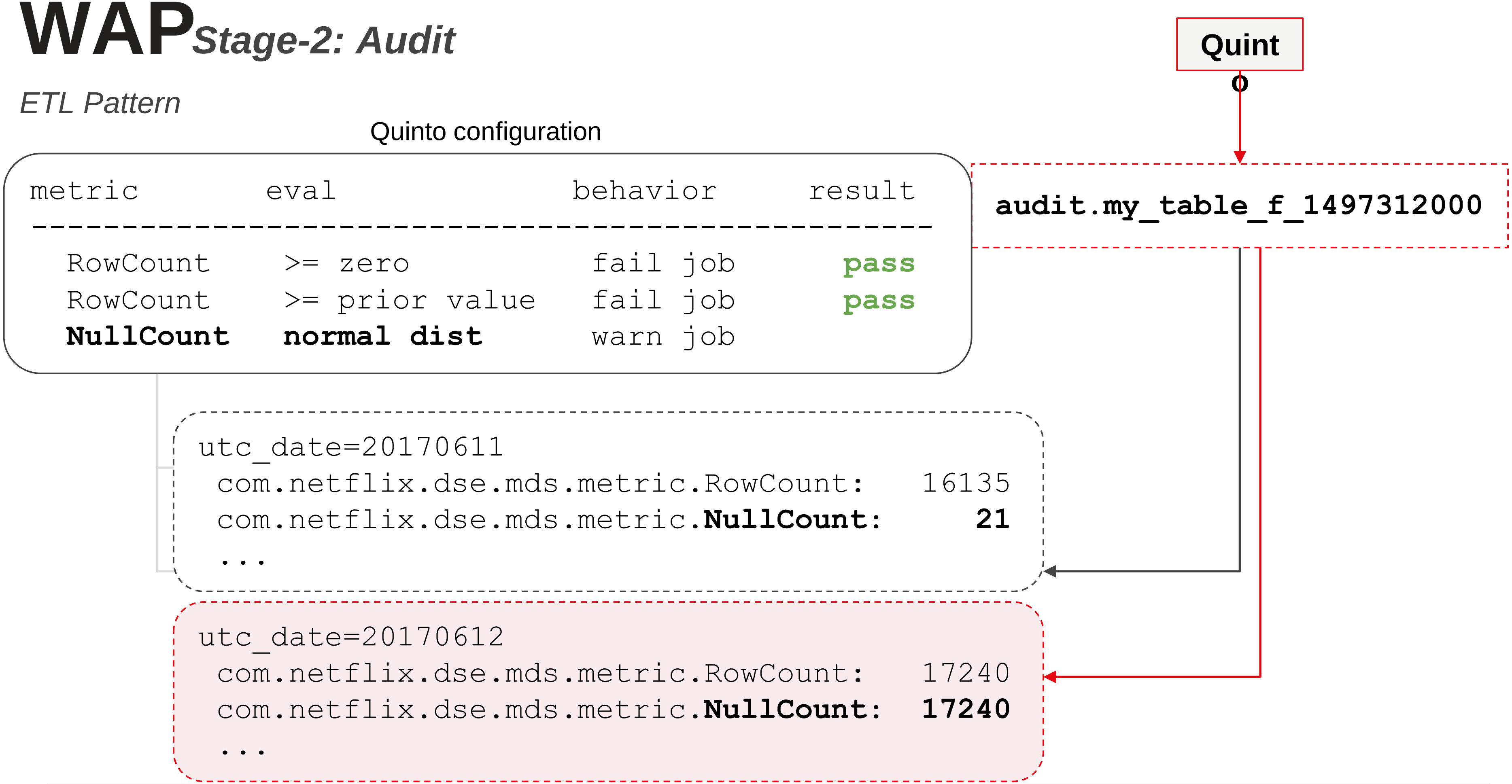
audit.my_table_f_1497312000

```
utc_date=20170611
 com.netflix.dse.mds.metric.RowCount:   16135
 com.netflix.dse.mds.metric.NullCount:     21
 ...
```

```
utc_date=20170612
 com.netflix.dse.mds.metric.RowCount:   17240
 com.netflix.dse.mds.metric.NullCount:  17240
 ...
```

WAP *Stage-2: Audit*

ETL Pattern



Quinto configuration

```
metric         eval                behavior      result
-----------------------------------------------------------
  RowCount     >= zero              fail job         pass
  RowCount     >= prior value       fail job         pass
  NullCount    normal dist          warn job
```

Quint
o

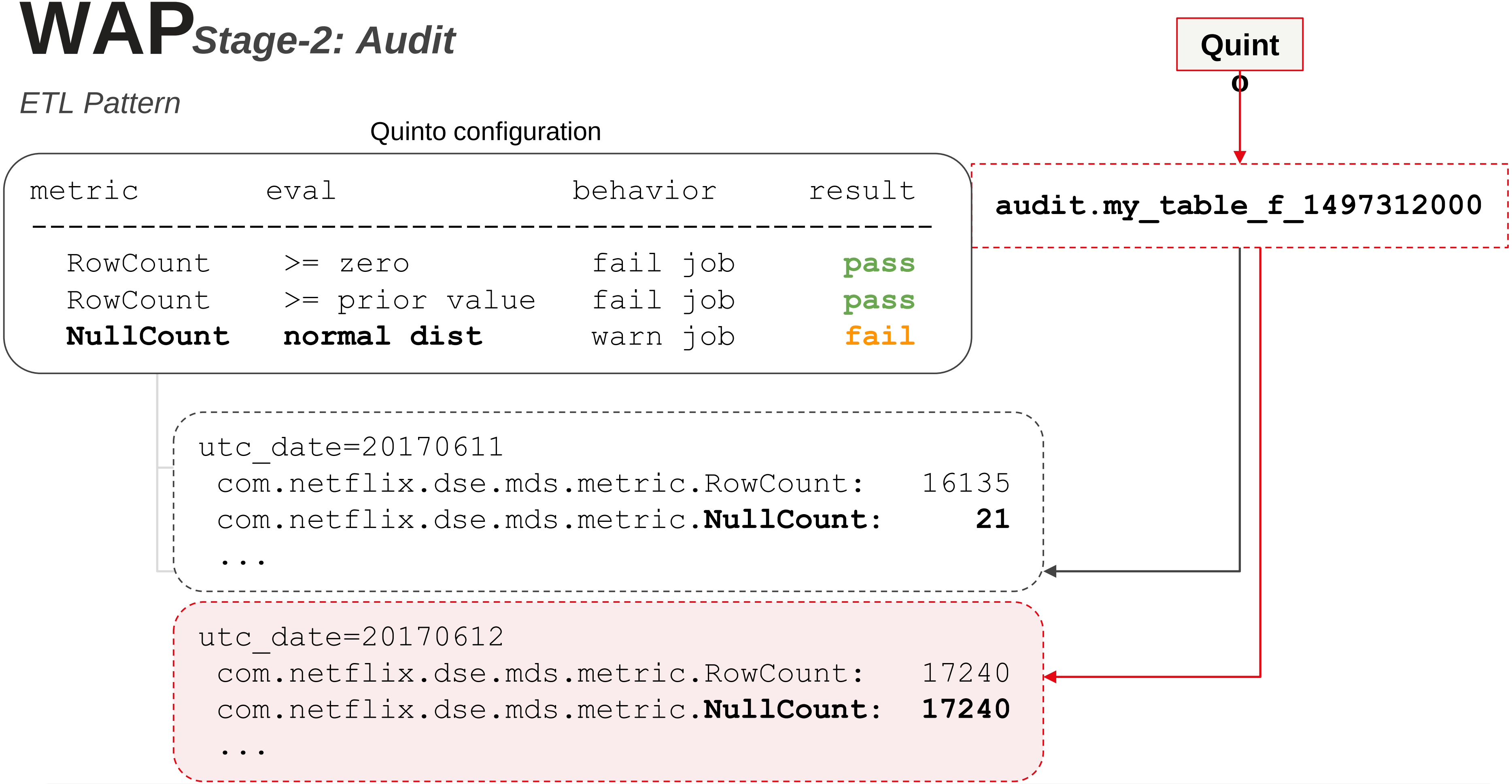
audit.my_table_f_1497312000

```
utc_date=20170611
 com.netflix.dse.mds.metric.RowCount:    16135
 com.netflix.dse.mds.metric.NullCount:      21
 ...
```

```
utc_date=20170612
 com.netflix.dse.mds.metric.RowCount:    17240
 com.netflix.dse.mds.metric.NullCount:   17240
 ...
```

WAP *Stage-2: Audit*

ETL Pattern



Quint
o

```
audit.my_table_f_1497312000
```

Quinto configuration

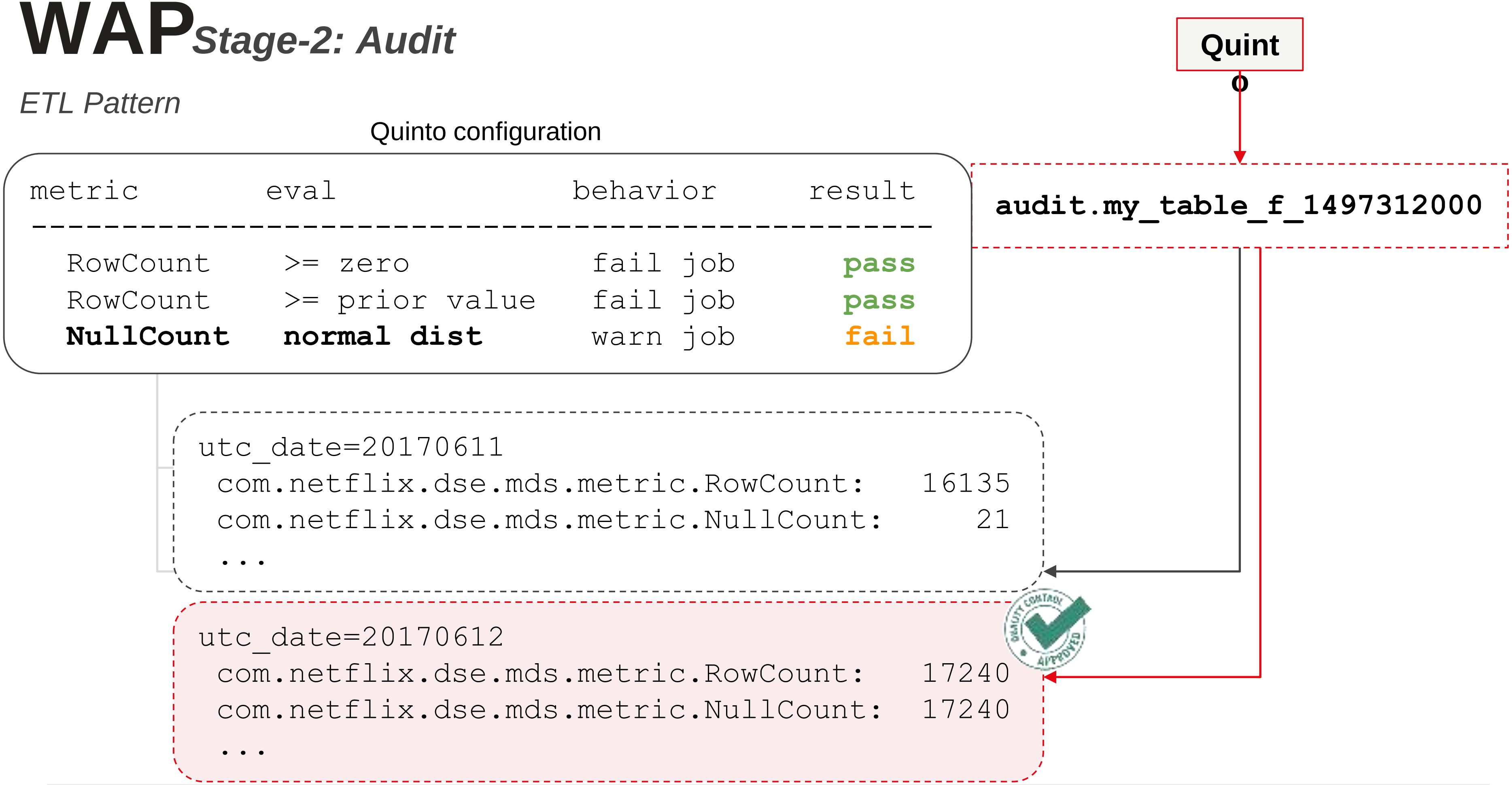
```
metric         eval               behavior     result
----------------------------------------------------
  RowCount     >= zero             fail job      pass
  RowCount     >= prior value      fail job      pass
  NullCount    normal dist         warn job      fail
```

```
utc_date=20170611
 com.netflix.dse.mds.metric.RowCount:    16135
 com.netflix.dse.mds.metric.NullCount:      21
 ...
```

```
utc_date=20170612
 com.netflix.dse.mds.metric.RowCount:    17240
 com.netflix.dse.mds.metric.NullCount:   17240
 ...
```

WAP *Stage-2: Audit*

ETL Pattern



Quinto configuration

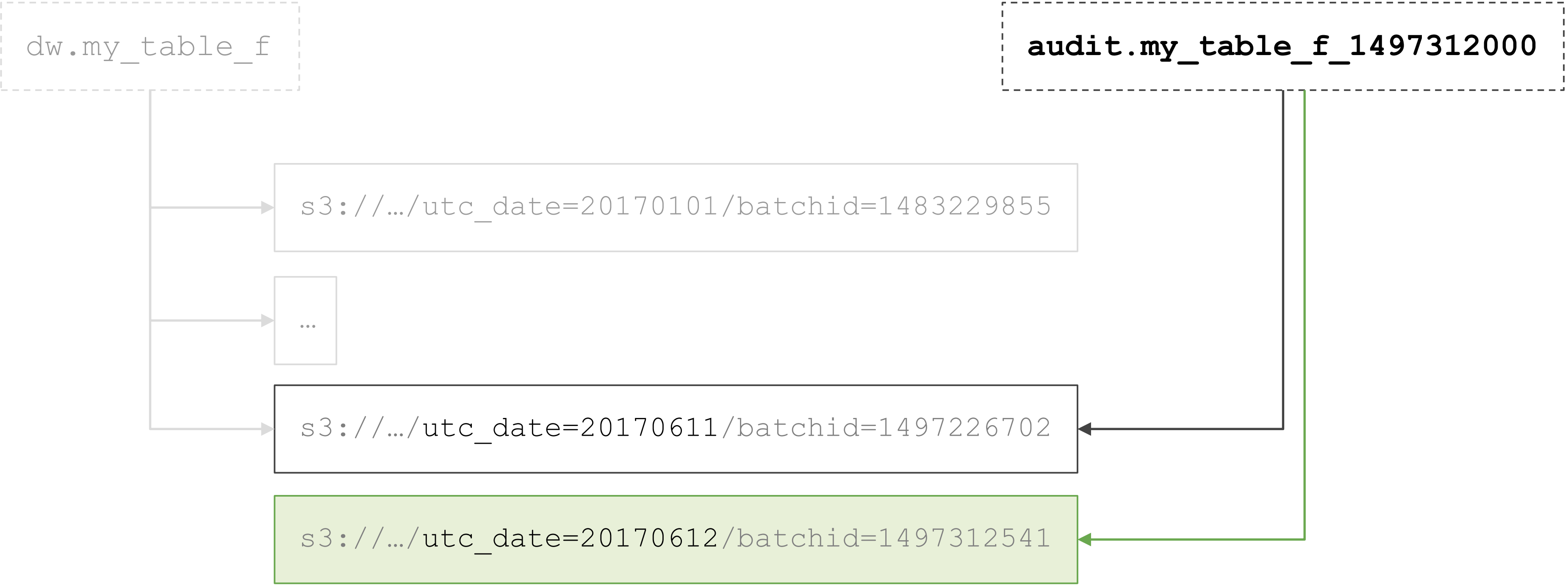
```
metric         eval              behavior     result
---------------------------------------------------
  RowCount     >= zero            fail job      pass
  RowCount     >= prior value     fail job      pass
  NullCount    normal dist        warn job      fail
```

```
utc_date=20170611
 com.netflix.dse.mds.metric.RowCount:   16135
 com.netflix.dse.mds.metric.NullCount:     21
 ...
```

```
utc_date=20170612
 com.netflix.dse.mds.metric.RowCount:   17240
 com.netflix.dse.mds.metric.NullCount:  17240
 ...
```

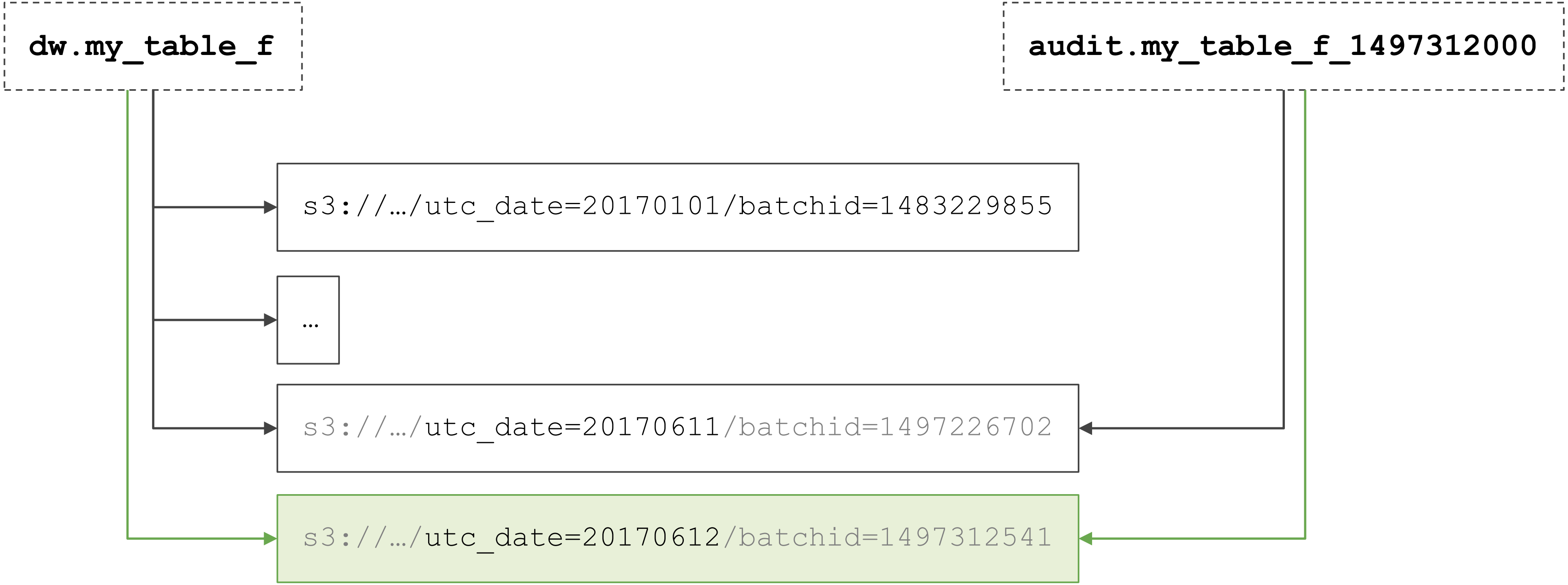
WAP *Stage-3: Publish*

ETL Pattern


dw.my_table_f
audit.my_table_f_1497312000
s3://…/utc_date=20170101/batchid=1483229855
…
s3://…/utc_date=20170611/batchid=1497226702
s3://…/utc_date=20170612/batchid=1497312541

WAP*Stage-3: Publish*

ETL Pattern



dw.my_table_f

audit.my_table_f_1497312000

s3://…/utc_date=20170101/batchid=1483229855

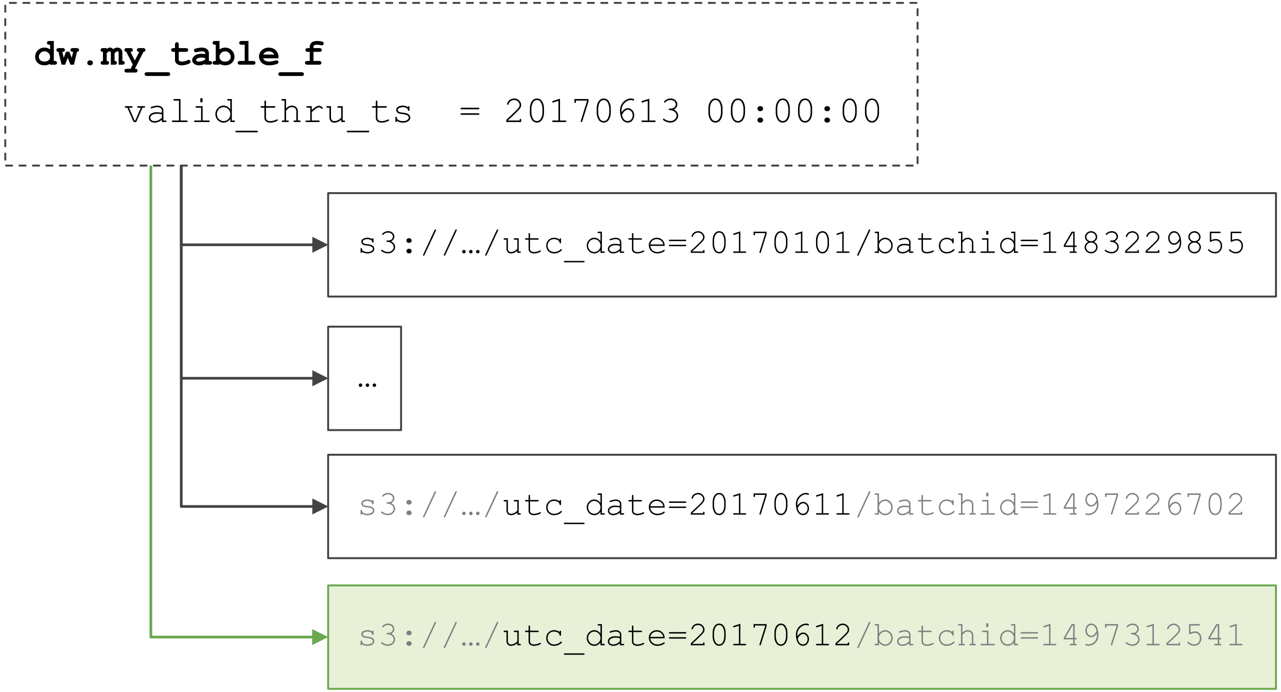
…

s3://…/utc_date=20170611/batchid=1497226702

s3://…/utc_date=20170612/batchid=1497312541

WAP *Stage-3: Publish*

ETL Pattern



dw.my_table_f

valid_thru_ts = 20170613 00:00:00

s3://…/utc_date=20170101/batchid=1483229855

…

s3://…/utc_date=20170611/batchid=1497226702

s3://…/utc_date=20170612/batchid=1497312541

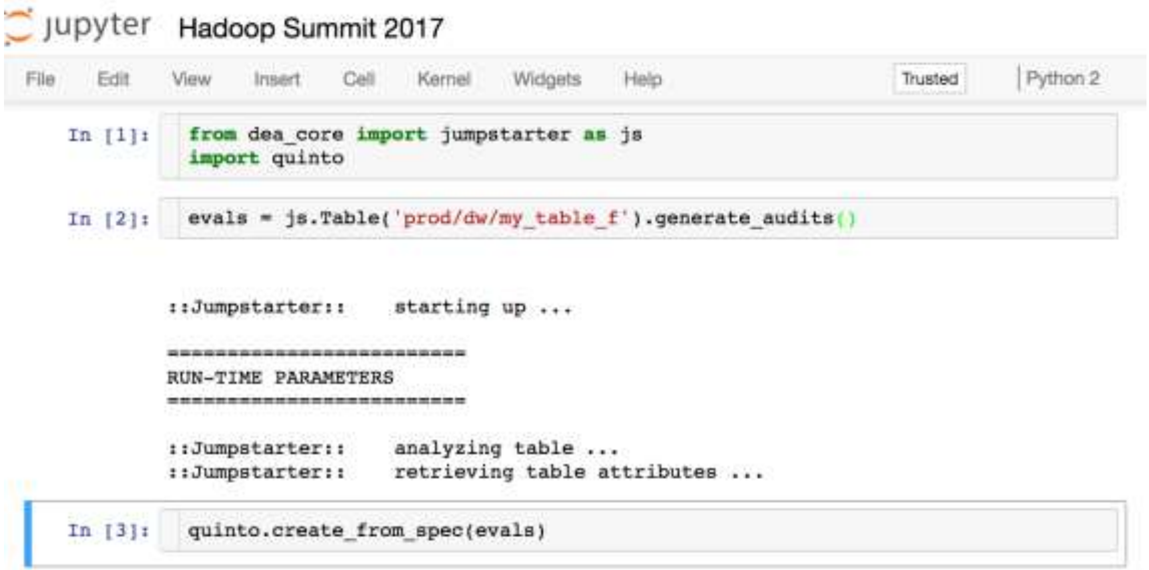
Python Libraries.

Jumpstarter.

Python Library

Quinto evaluations

- intelligent recommendations
- multiple tiers of coverage
- configurable rules



```
In [1]:  from dea_core import jumpstarter as js
         import quinto
```

```
In [2]:  evals = js.Table('prod/dw/my_table_f').generate_audits()
```

```
::Jumpstarter::    starting up ...

=========================
RUN-TIME PARAMETERS
=========================

::Jumpstarter::    analyzing table ...
::Jumpstarter::    retrieving table attributes ...
```

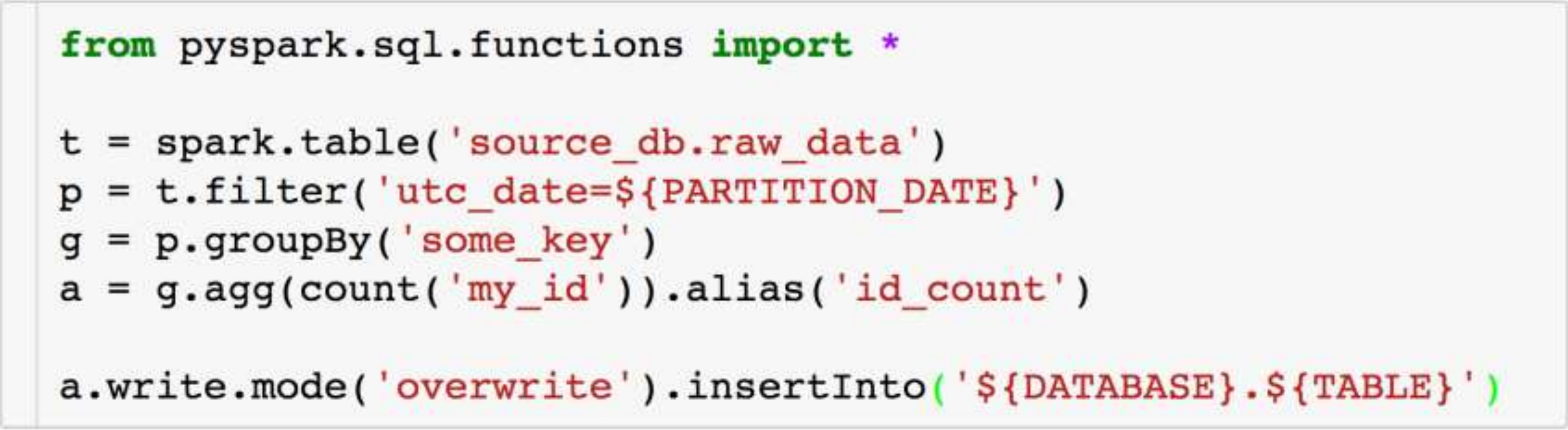
```
In [3]:  quinto.create_from_spec(evals)
```

WAP.

Python Library

Minimal requirements

- parameterized destination table

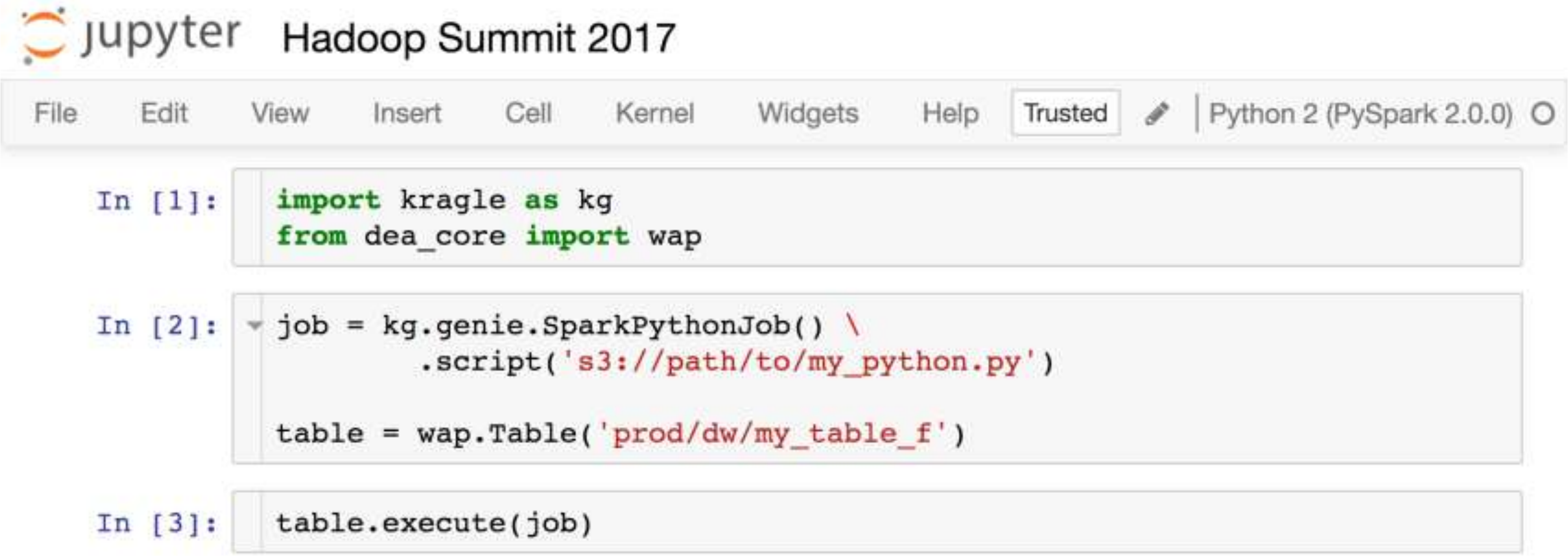


```
from pyspark.sql.functions import *

t = spark.table('source_db.raw_data')
p = t.filter('utc_date=${PARTITION_DATE}')
g = p.groupBy('some_key')
a = g.agg(count('my_id')).alias('id_count')

a.write.mode('overwrite').insertInto('${DATABASE}.${TABLE}')
```

WAP.

Python Library

Running WAP.



In [1]:

```
import kragle as kg
from dea_core import wap
```

In [2]:

```
job = kg.genie.SparkPythonJob() \
        .script('s3://path/to/my_python.py')

table = wap.Table('prod/dw/my_table_f')
```

In [3]:

```
table.execute(job)
```

NETFLIX

Scalable Data

What’s Next.

- additional Metacat statistics
- robust anomaly detection (RAD)
- complete migration for all prod tables

Tips & Lessons Learned.

- Query-based solution may be “good enough” for many.
- Not all tables need quality coverage.
- One size rarely fits all tables.
- Build components, not “all-or-nothing” frameworks.

Thank you!

MICHELLE UFFORD

mufford@netflix.com

twitter.com/MichelleUfford

NETFLIX DATA

techblog.netflix.com

medium.com/netflix-techblog

twitter.com/NetflixData

tinyurl.com/NetflixData

WE’RE HIRING! jobs.netflix.com