FANTASTIC DATUMS

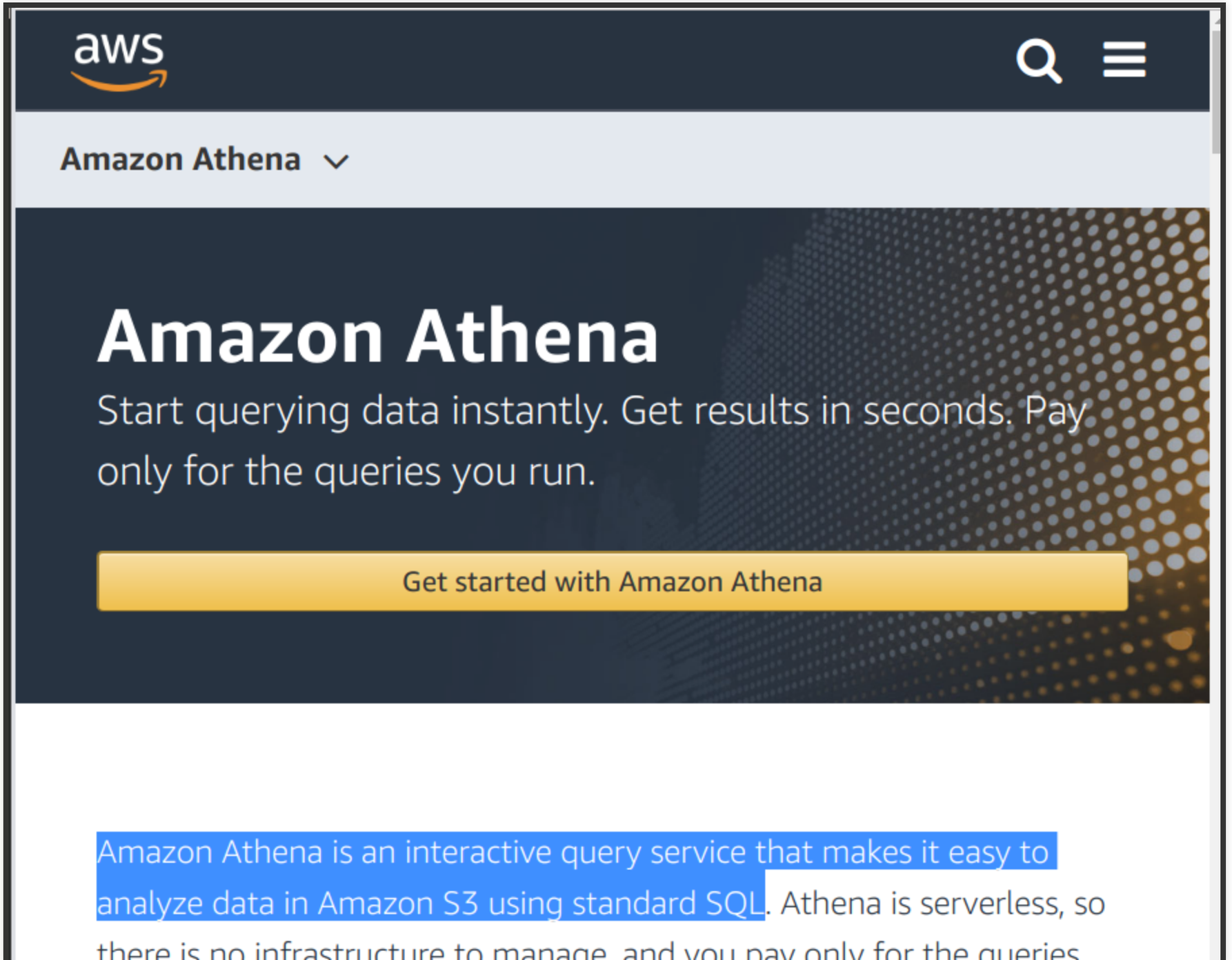
A IS FOR ATHENA

G_IS_FOR_GLUE

& WHERE TO FIND THEM

AMAZON ATHENA

Service to analyze data in S3 using SQL (prestoDB)



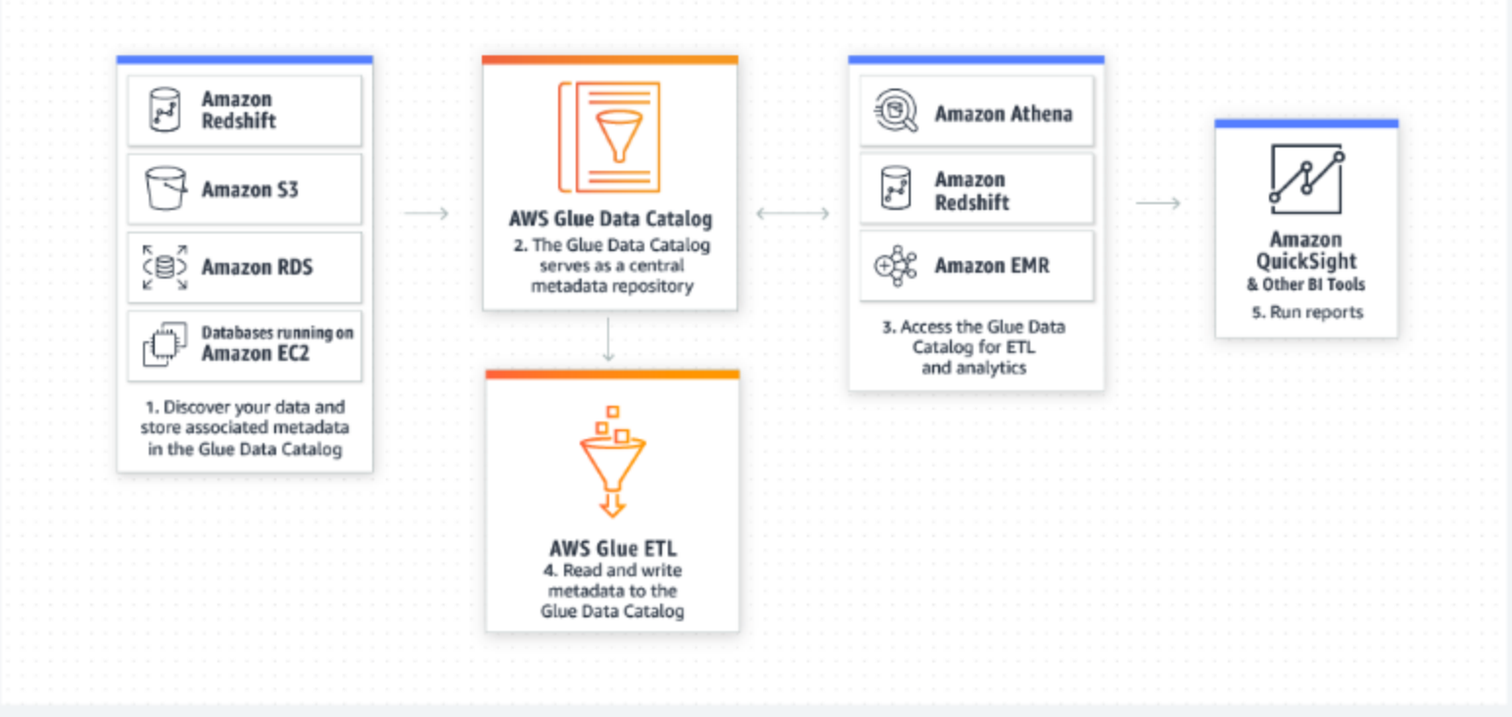
there is no infrastructure to manage, and you pay only for the queries that you run.

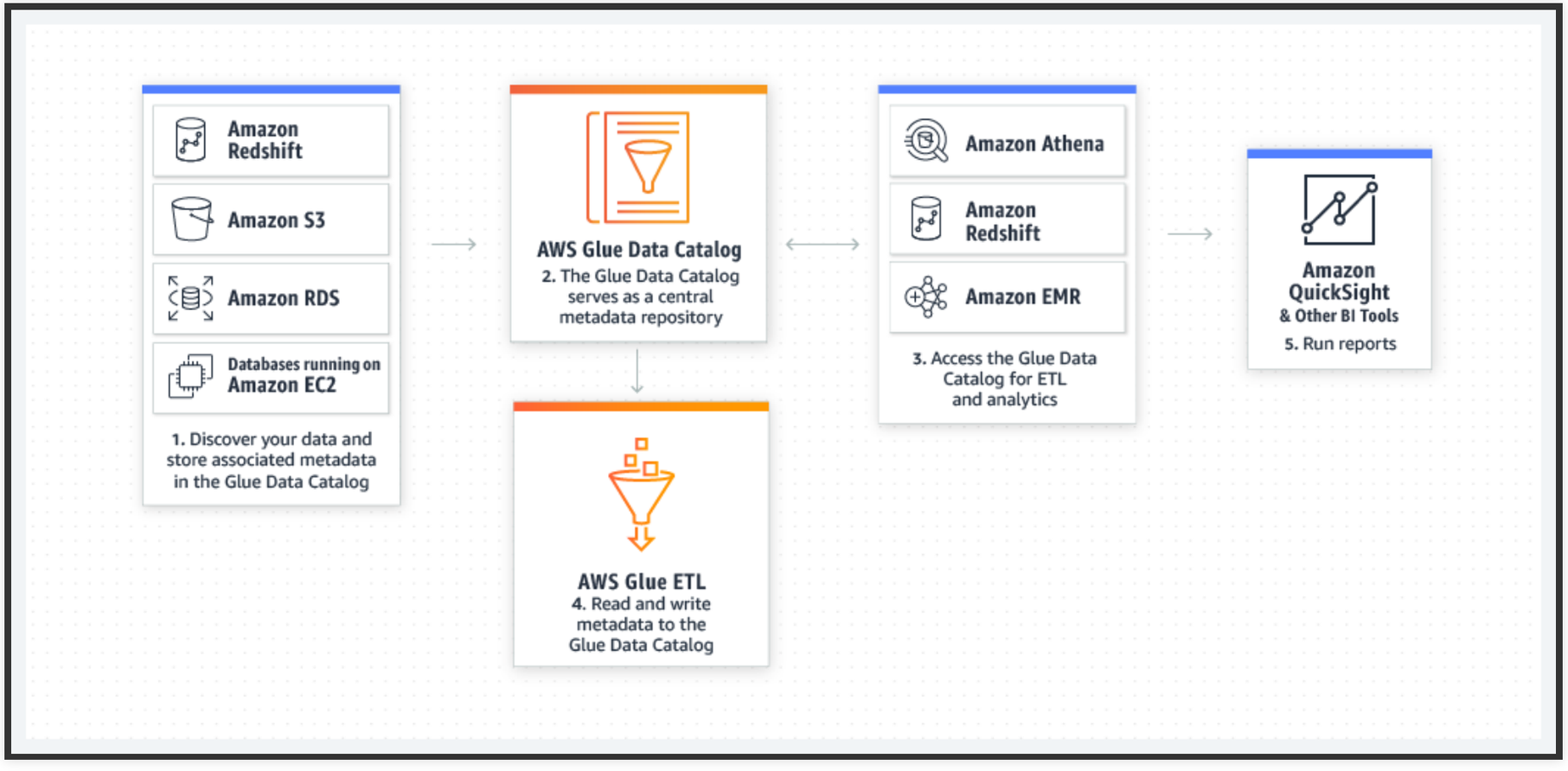
GLUE DATA CATALOG

Hive Compatible Metadata store from AWS Glue

Unified View of Your Data Across Multiple Data Stores

You can use the AWS Glue Data Catalog to quickly discover and search across multiple AWS data sets without moving the data. Once the data is cataloged, it is immediately available for search and query using Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum.




Amazon Redshift
Amazon S3
Amazon RDS
Databases running on Amazon EC2
1. Discover your data and store associated metadata in the Glue Data Catalog
AWS Glue Data Catalog
2. The Glue Data Catalog serves as a central metadata repository
AWS Glue ETL
4. Read and write metadata to the Glue Data Catalog
Amazon Athena
Amazon Redshift
Amazon EMR
3. Access the Glue Data Catalog for ETL and analytics
Amazon QuickSight
& Other BI Tools
5. Run reports

DATA

`core`

- cloudflare, cloudtrail, and elb logs
- prod logs by tag, date, hour

`prod_alerts` / `prod_starterview`

- mirrors of alerts and starterview for data enrichment

`prod_inventory`, `stg_inventory`, `dev_inventory`

- table for each s3 bucket showing s3 meta data (s3 inventory).

DATA: UNIFIED LOGS

`core.prod_unified_logs`

3 days worth of all events, partitioned by `tag`, `log_date`, `hour`

`raw_unified`

database per tag, raw json logs partitioned by `log_date`

`struct_unified`

manually created tables, map json to fields

`prod_event`

- converted from json to read-optimized parquet
- s3://prod-ziprecruiter-datalake-data

`prod_unified_raw_v1`

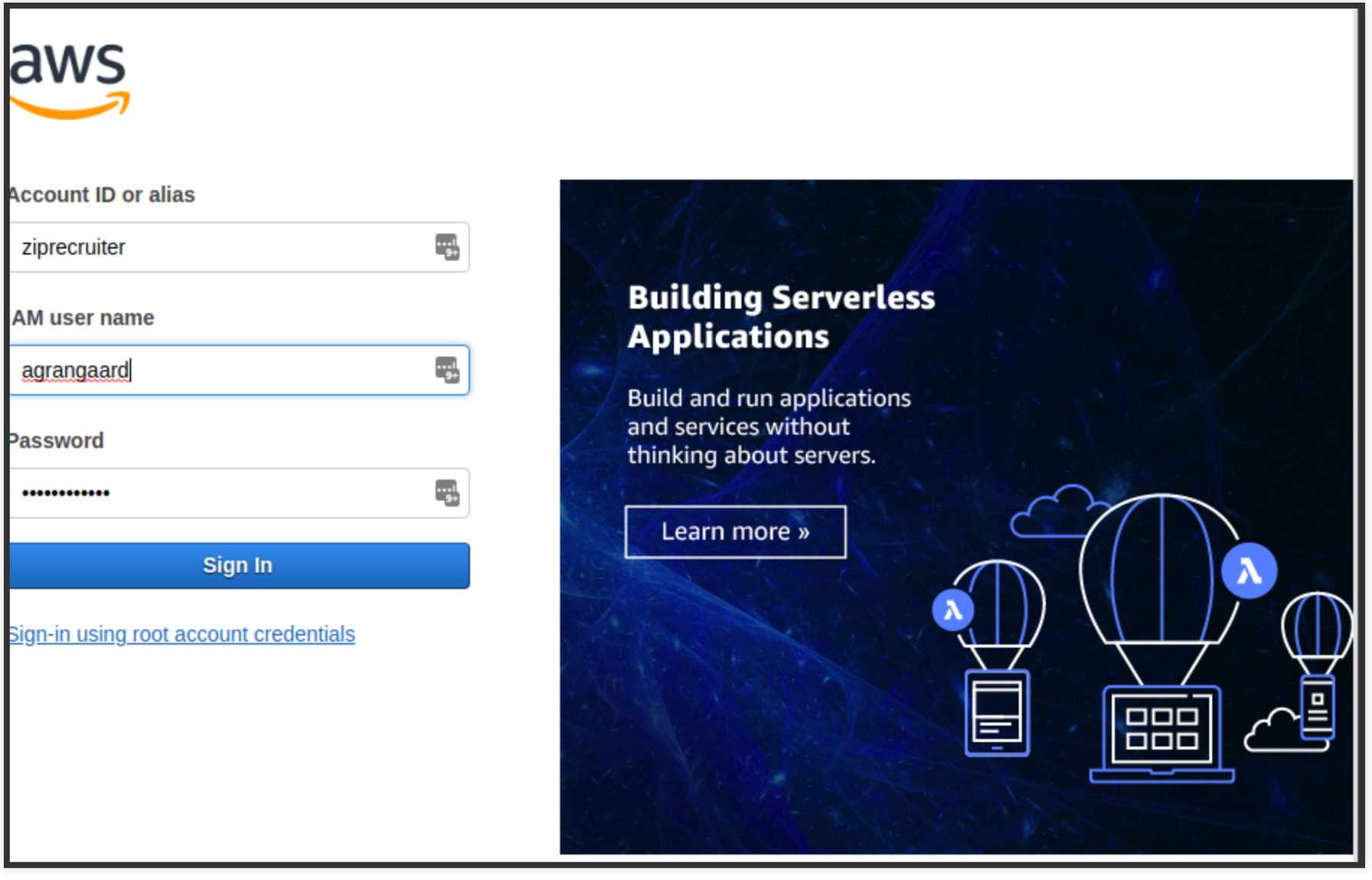
- raw event logs, requires interpreting

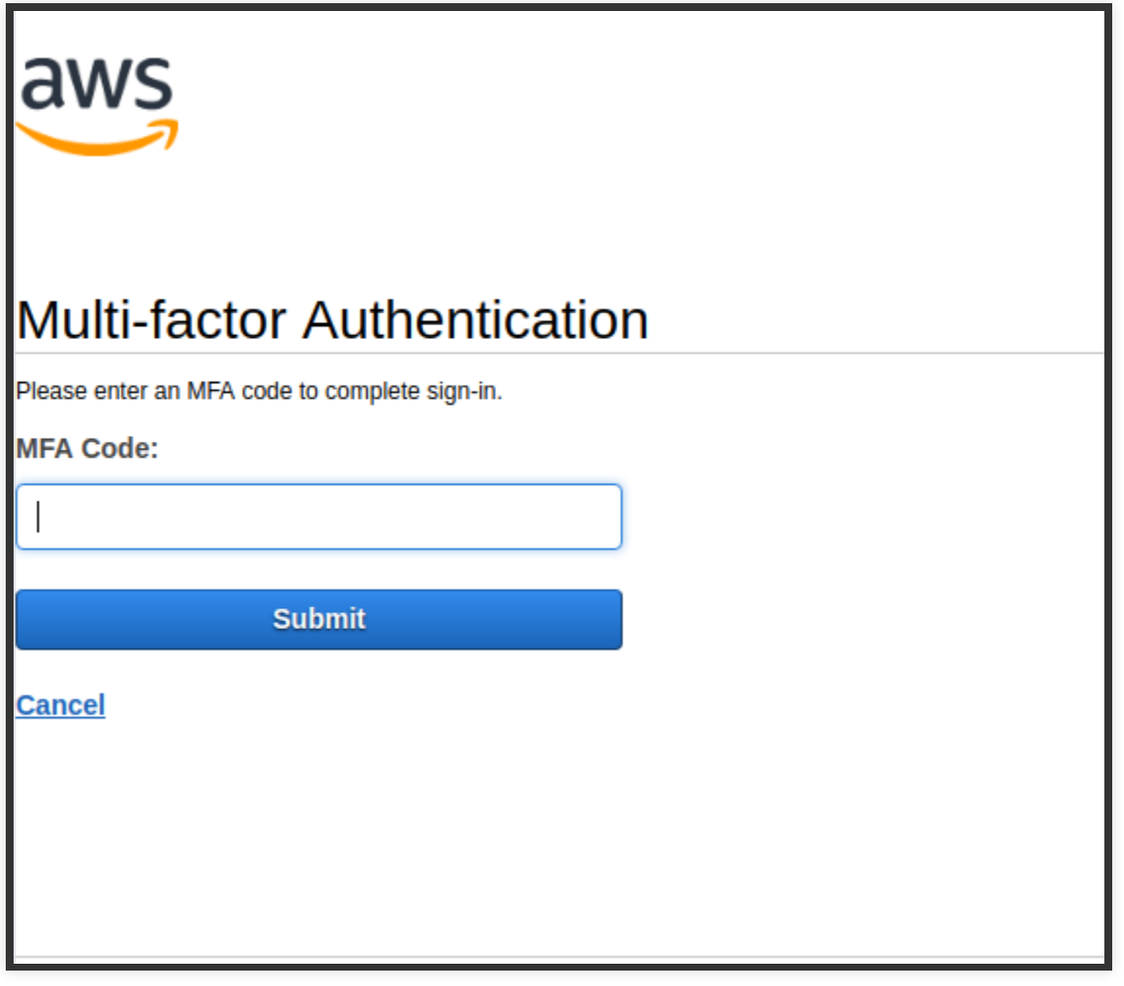
json
- 492 tables! who maintains?

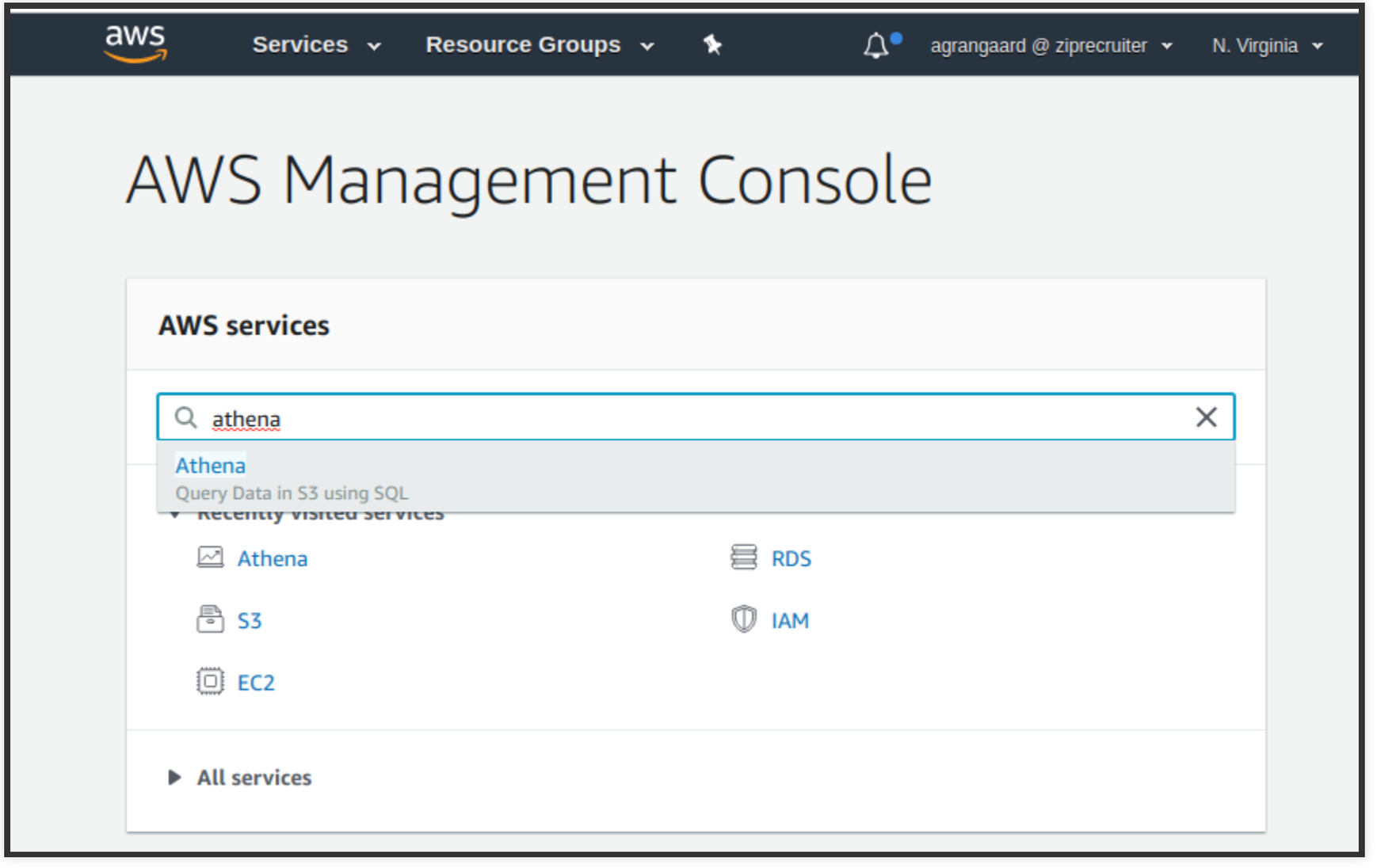
QUERY ATHENA

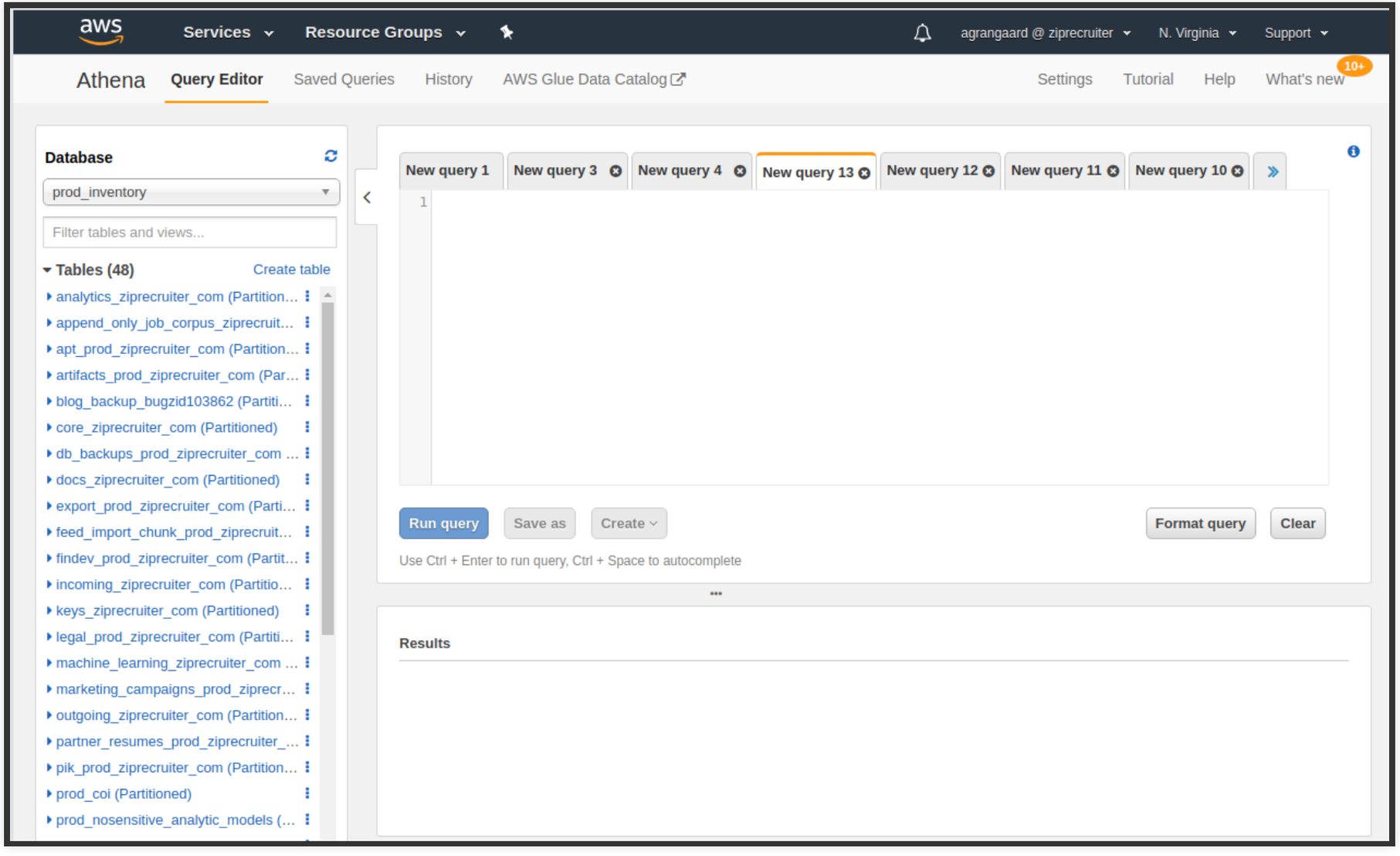
1. `athena` tool in AWS Console.
2. API, using `boto` or similar tool.

AWS CONSOLE ATHENA

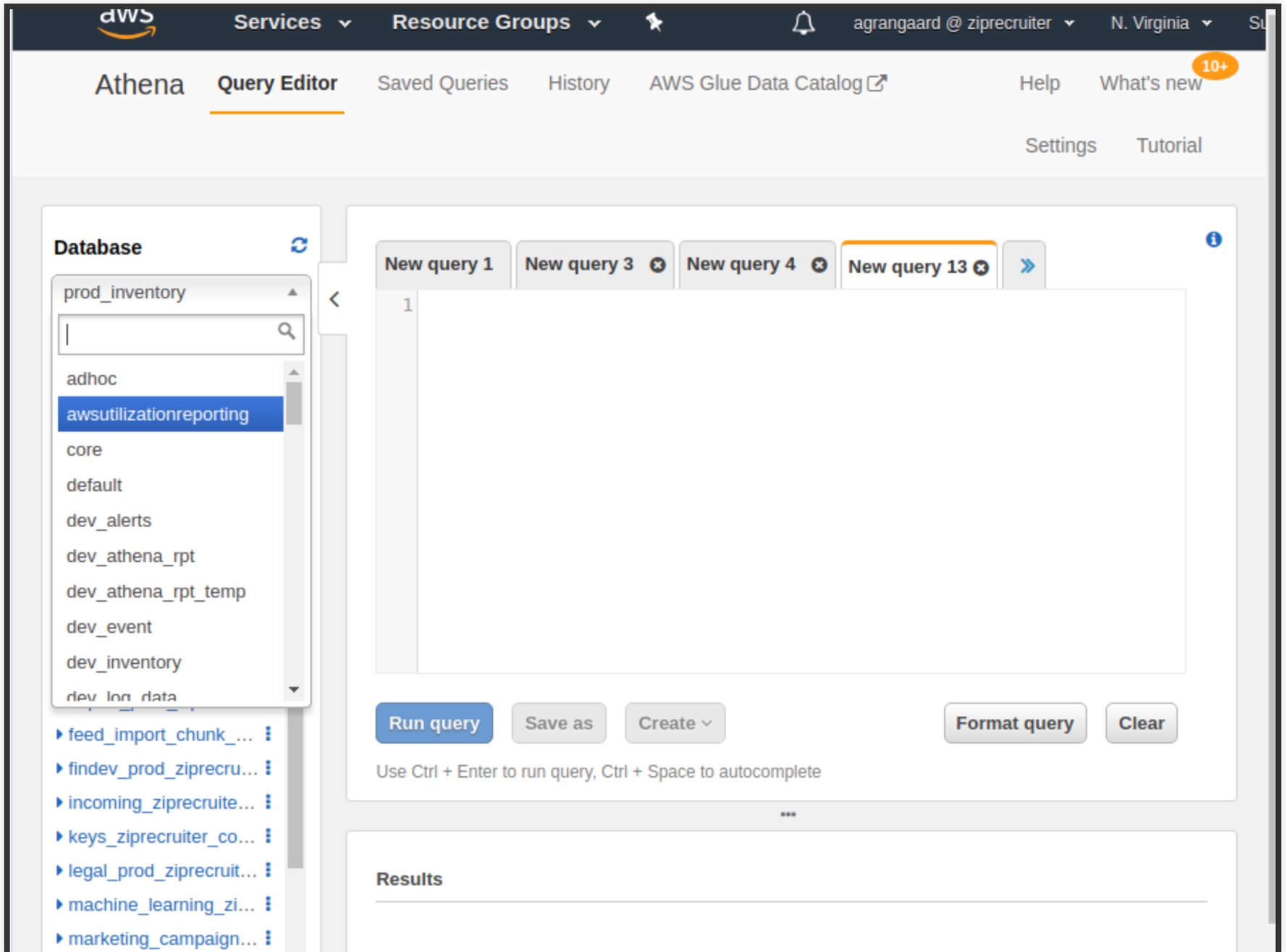


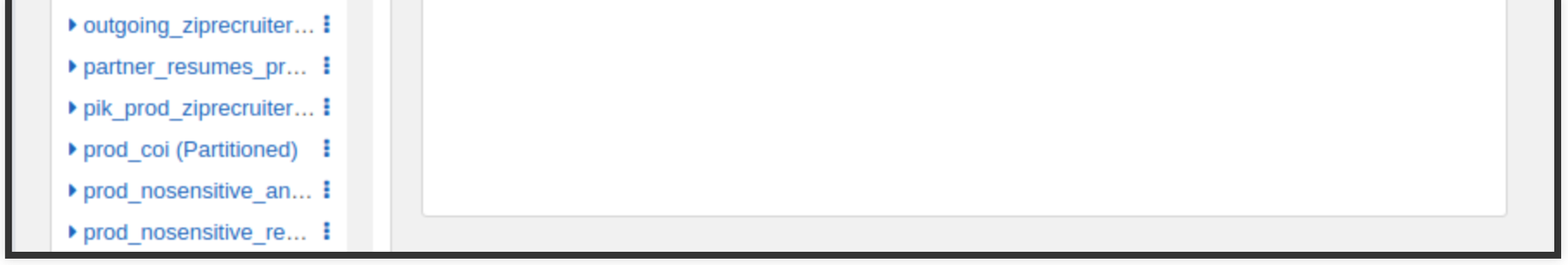

aws
Multi-factor Authentication
Please enter an MFA code to complete sign-in.
MFA Code:
Submit
Cancel


aws
Services
Resource Groups
agrangaard @ ziprecruiter
N. Virginia
AWS Management Console
AWS services
athena
Athena
Query Data in S3 using SQL
Athena
RDS
S3
IAM
EC2
All services

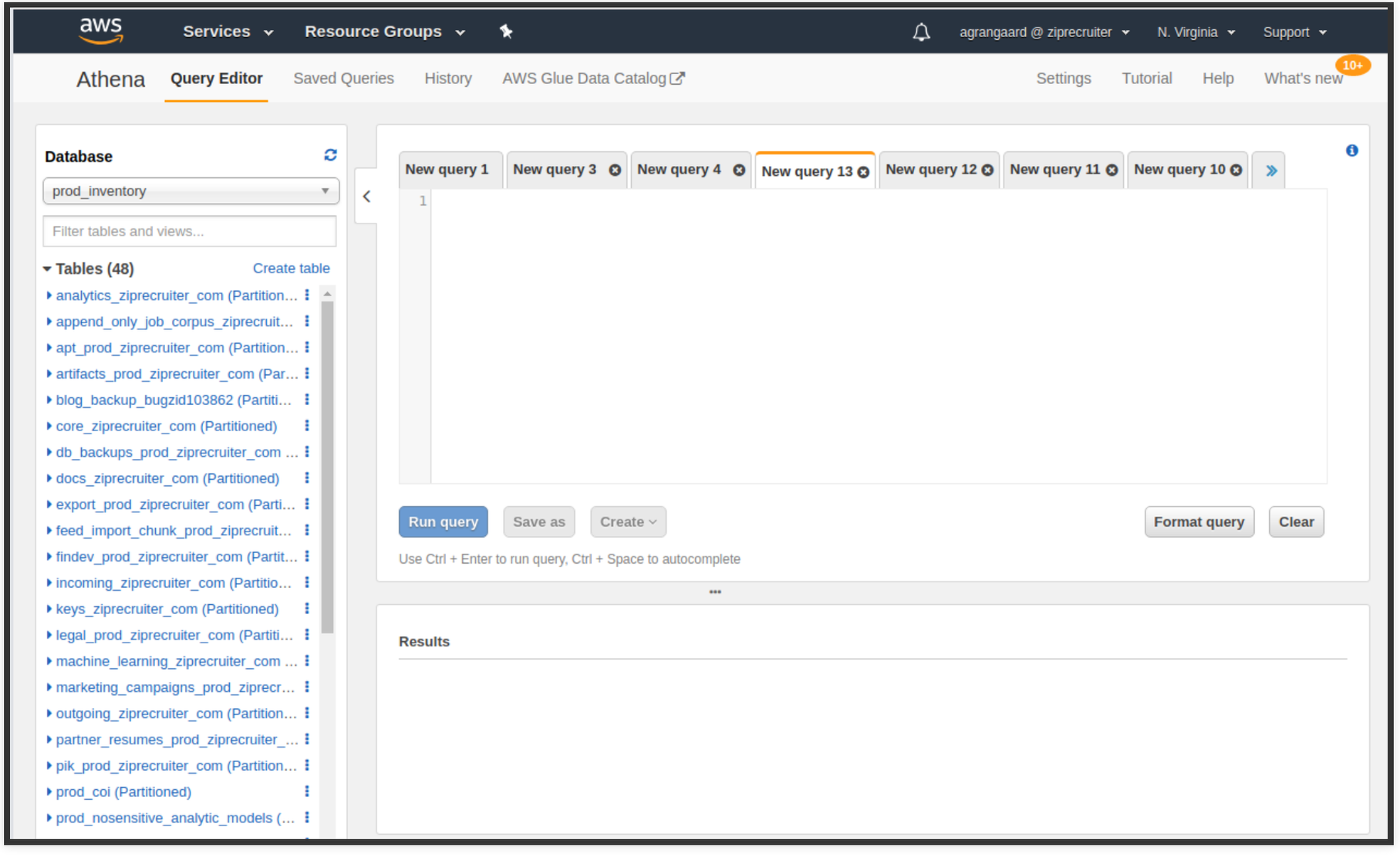

aws
Services
Resource Groups
agrangaard @ ziprecruiter
N. Virginia
Support
Athena
Query Editor
Saved Queries
History
AWS Glue Data Catalog
Settings
Tutorial
Help
What's new
10+
Database
prod_inventory
Filter tables and views...
Tables (48)
Create table
analytics_ziprecruiter_com (Partition...
append_only_job_corpus_ziprecruit...
apt_prod_ziprecruiter_com (Partition...
artifacts_prod_ziprecruiter_com (Par...
blog_backup_bugzid103862 (Partiti...
core_ziprecruiter_com (Partitioned)
db_backups_prod_ziprecruiter_com ...
docs_ziprecruiter_com (Partitioned)
export_prod_ziprecruiter_com (Parti...
feed_import_chunk_prod_ziprecruit...
findev_prod_ziprecruiter_com (Partit...
incoming_ziprecruiter_com (Partitio...
keys_ziprecruiter_com (Partitioned)
legal_prod_ziprecruiter_com (Partiti...
machine_learning_ziprecruiter_com ...
marketing_campaigns_prod_ziprecr...
outgoing_ziprecruiter_com (Partition...
partner_resumes_prod_ziprecruiter_...
pik_prod_ziprecruiter_com (Partition...
prod_coi (Partitioned)
prod_nosensitive_analytic_models (...
New query 1
New query 3
New query 4
New query 13
New query 12
New query 11
New query 10
1
Run query
Save as
Create
Format query
Clear
Use Ctrl + Enter to run query, Ctrl + Space to autocomplete
Results

SHOW DATABASES

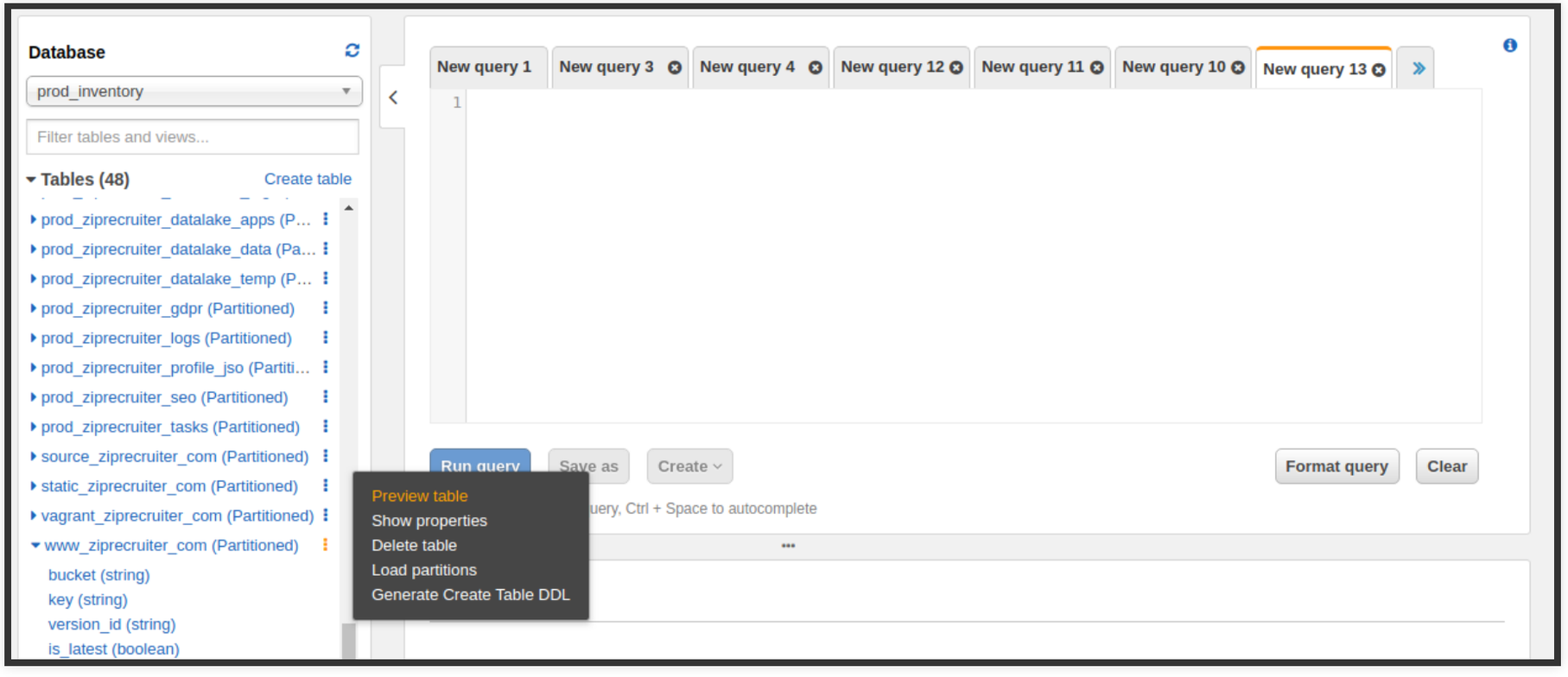



outgoing_ziprecruiter…
partner_resumes_pr…
pik_prod_ziprecruiter…
prod_coi (Partitioned)
prod_nosensitive_an…
prod_nosensitive_re…

SHOW TABLES IN DATABASE



PREVIEW TABLE



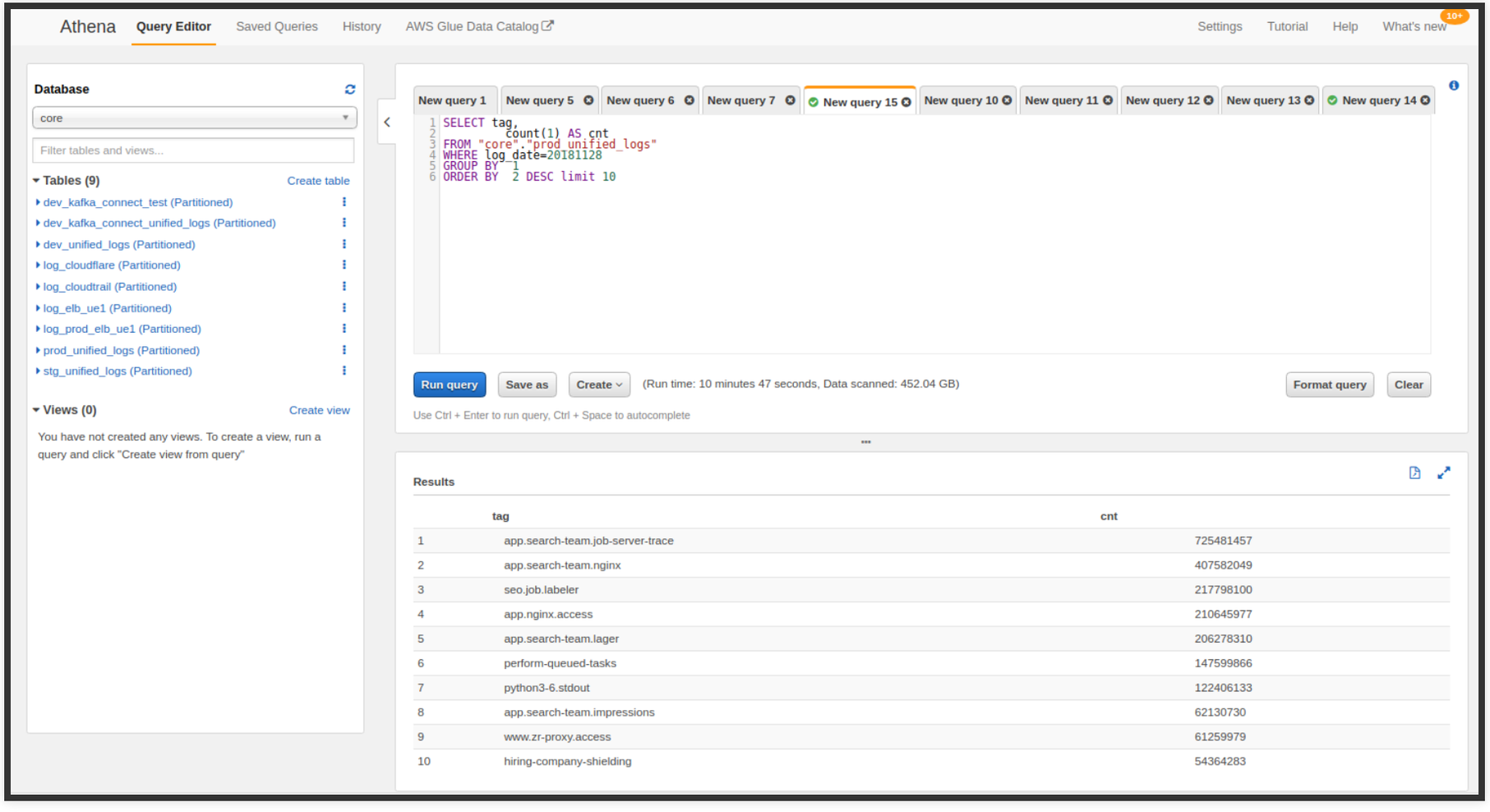
EXAMPLES

PROD LOGS

```
SELECT tag,
         count(1) AS cnt
FROM "core"."prod_unified_logs"
WHERE log_date=20181128
GROUP BY  1
ORDER BY  2 DESC limit 10
```

10 minute query scans 450 gb of data. ($2.50 query)

	tag	cnt
1	app.search-team.job-server-trace	725481457
2	app.search-team.nginx	407582049
3	seo.job.labeler	217798100
4	app.nginx.access	210645977
5	app.search-team.lager	206278310
6	perform-queued-tasks	147599866
7	python3-6.stdout	122406133
8	app.search-team.impressions	62130730
9	www.zr-proxy.access	61259979
10	hiring-company-shielding	54364283


Athena
Query Editor
Saved Queries
History
AWS Glue Data Catalog
Settings
Tutorial
Help
What's new
10+
Database
core
Filter tables and views...
Tables (9)
Create table
dev_kafka_connect_test (Partitioned)
dev_kafka_connect_unified_logs (Partitioned)
dev_unified_logs (Partitioned)
log_cloudflare (Partitioned)
log_cloudtrail (Partitioned)
log_elb_ue1 (Partitioned)
log_prod_elb_ue1 (Partitioned)
prod_unified_logs (Partitioned)
stg_unified_logs (Partitioned)
Views (0)
Create view
You have not created any views. To create a view, run a query and click "Create view from query"
New query 1
New query 5
New query 6
New query 7
New query 15
New query 10
New query 11
New query 12
New query 13
New query 14
1 SELECT tag,
2 count(1) AS cnt
3 FROM "core"."prod_unified_logs"
4 WHERE log_date=20181128
5 GROUP BY 1
6 ORDER BY 2 DESC limit 10
Run query
Save as
Create
(Run time: 10 minutes 47 seconds, Data scanned: 452.04 GB)
Format query
Clear
Use Ctrl + Enter to run query, Ctrl + Space to autocomplete
Results
tag
cnt
1 app.search-team.job-server-trace 725481457
2 app.search-team.nginx 407582049
3 seo.job.labeler 217798100
4 app.nginx.access 210645977
5 app.search-team.lager 206278310
6 perform-queued-tasks 147599866
7 python3-6.stdout 122406133
8 app.search-team.impressions 62130730
9 www.zr-proxy.access 61259979
10 hiring-company-shielding 54364283

S3 INVENTORY

WWW_ZIPRECRUITER_COM

metadata over time

```
SELECT dt,
         is_latest,
         count(key) AS key,
         sum(size) / (1024 * 1024) /1024 / 1024 AS total_si
FROM "prod_inventory"."www_ziprecruiter_com"
WHERE (dt='2018-10-26-08-00'
        OR dt='2018-11-26-08-00')

GROUP BY  1,2
ORDER BY  1 , 3 DESC ,4 DESC limit 10;
```

dt	is_latest	key	total_size_tb
2018-10-26-08-00	false	3120810	725
2018-10-26-08-00	true	17664	0
2018-11-26-08-00	false	345177	70
2018-11-26-08-00	true	18638	0

WWW_ZIPRECRUITER_COM

How many copies of feed/zr-lastest.xml

```
SELECT count(key) AS cnt,
         key,
         dt,
         sum(size) / (1024 * 1024) /1024 AS total_size_gb
FROM "prod_inventory"."www_ziprecruiter_com"
WHERE (dt='2018-10-26-08-00'
        OR dt='2018-11-26-08-00')
        AND key='feed/zr-latest.xml'
GROUP BY  2,3
ORDER BY  1 DESC ,4 DESC limit 10;
```

cnt	key	dt	total_size_gb
323306	feed/zr-latest.xml	2018-10-26-08-00	2495
12461	feed/zr-latest.xml	2018-11-26-08-00	89

WWW_ZIPSTG_COM

How many copies of feed/zr-lastest.xml

```
SELECT count(key) AS cnt,
         key,
         dt,
         sum(size) / (1024 * 1024) /1024 AS total_size_gb
FROM "prod_inventory"."www_zipstg_com"
WHERE (dt='2018-10-26-08-00'
        OR dt='2018-11-26-08-00')
        AND key='feed/zr-latest.xml'
GROUP BY  2,3
ORDER BY  1 DESC ,4 DESC limit 10;
#+END_SRC sql

#+BEGIN_SRC
  cnt   key      dt       total_size_gb
281373  feed/zr-latest.xml      2018-10-26-08-00        32
28060   feed/zr-latest.xml      2018-11-26-08-00        3
```

YOPASS BUCKET VIA AWS CLI

```
% aws s3 ls s3://yopass.ziprecruiter.com/
                           PRE static/
2018-08-27 11:25:53         19 _redirects
2018-08-27 11:25:54        196 asset-manifest.json
2018-08-27 11:25:54       5558 favicon.ico
2018-08-27 11:25:54      21813 favicon.png
2018-08-27 11:25:54       1044 index.html
2018-08-27 11:25:54        317 manifest.json
2018-08-27 11:25:54         26 robots.txt
2018-08-27 11:25:54       3288 service-worker.js

% aws s3 ls s3://yopass.ziprecruiter.com/static/
                           PRE css/
                           PRE js/

% aws s3 ls s3://yopass.ziprecruiter.com/static/css/
2018-08-27 11:25:54     137510 main.6f022e07.css
2018-08-27 11:25:54     221969 main.6f022e07.css.map

% aws s3 ls s3://yopass.ziprecruiter.com/static/js/
2018-08-27 11:25:54     632051 main.6371b98d.js
2018-08-27 11:25:54    3534838 main.6371b98d.js.map
```

YOPASS BUCKET VIA S3_INVENTORY

```
SELECT
  key
  , is_latest
  , is_delete_marker
  , size

FROM "prod_inventory"."yopass_ziprecruiter_com"
WHERE
  dt='2018-11-24-08-00'
ORDER BY key DESC
```

Results

	key	is_latest	is_delete_marker	size
1	static/js/main.6371b98d.js.map	true	false	3534838
2	static/js/main.6371b98d.js	true	false	632051
3	static/css/main.6f022e07.css.map	true	false	221969
4	static/css/main.6f022e07.css	true	false	137510
5	service-worker.js	true	false	3288
6	robots.txt	true	false	26
7	manifest.json	true	false	317
8	index.html	true	false	1044
9	favicon.png	true	false	21813
10	favicon.ico	true	false	5558
11	asset-manifest.json	true	false	196
12	_redirects	true	false	19

AWS CMDLINE

ALL DATABASES

```
aws glue get-databases | jq -c '.DatabaseList[]|.Name'
```

```
"adhoc"
"core"
"default"
"dev_alerts"
"dev_log_data"
"dev_reach"
"dev_reporting"
"dev_spamhaus"
"dev_starterview"
"dev_zr_finance"
"dev_zr_shared"
"es"
"hrxml"
"logs_test"
"prod"
"prod_alerts"
"prod_athena_rpt"
"prod_event"
"prod_jobs"
"prod_log_data"
```

"prod_reach"

"prod_reach_test"
"prod_reach_test2"
"prod_reporting"
"prod_reporting_test"
"prod_spamhaus"
"prod_starterview"
"prod_static"
"prod_tracking"
"prod_unified_raw_v1"
"prod_zr_finance"
"prod_zr_shared"
"raw_unified"
"reach2"
"redshift_archive"
"sburke"
"staging"
"stg_athena_rpt_temp"
"stg_raw_unified"
"stg_unified_raw_v1"
"struct_unified"
"tjones"
"ziprank"

CORE DATABASE TABLE

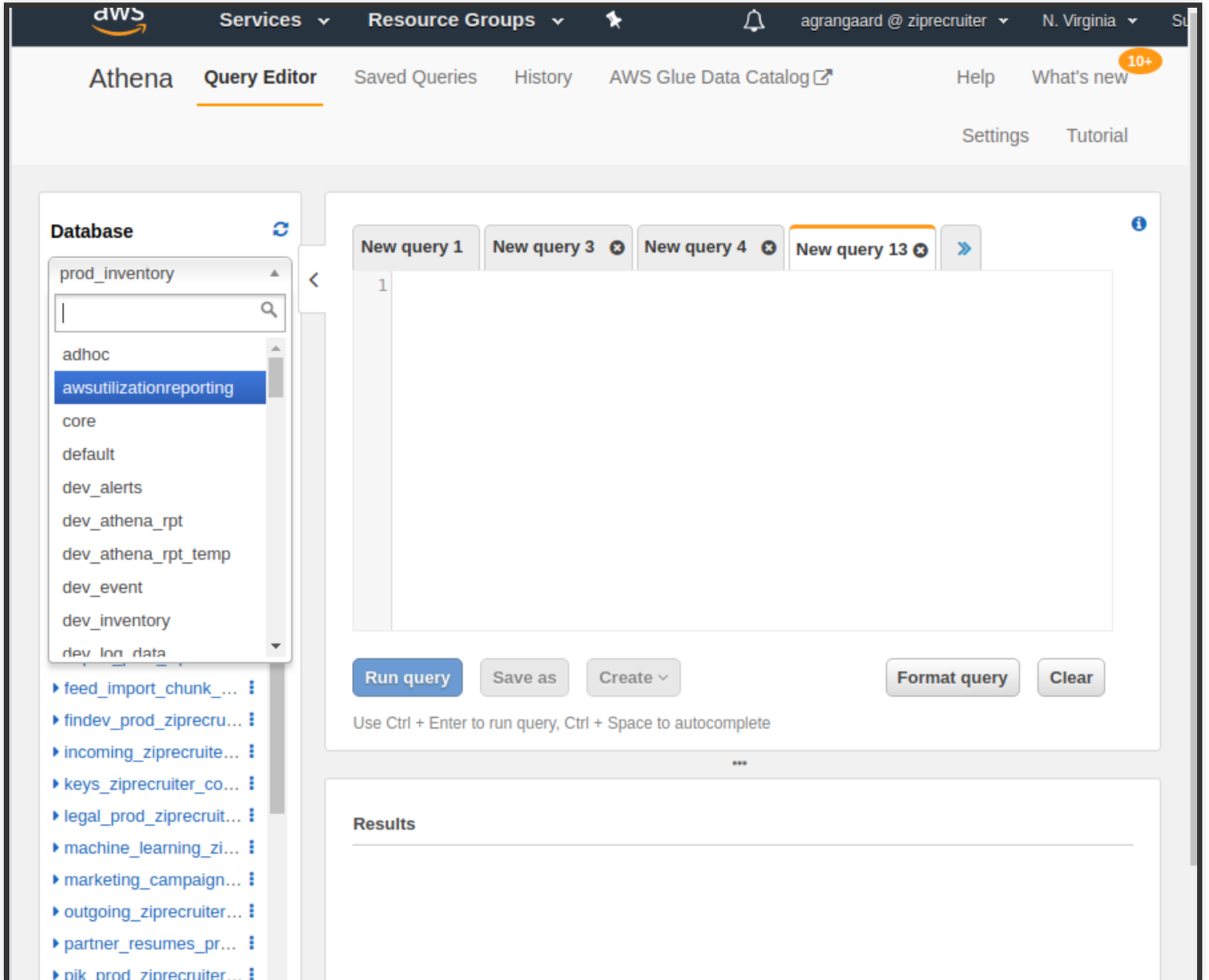
```
aws glue get-tables --database-name=core |jq '.TableList[].
```

```
"dev_kafka_connect_test"
"dev_kafka_connect_unified_logs"
"dev_unified_logs"
"log_cloudflare"
"log_cloudtrail"
"log_elb_ue1"
"log_prod_elb_ue1"
"prod_unified_logs"
"stg_unified_logs"
```

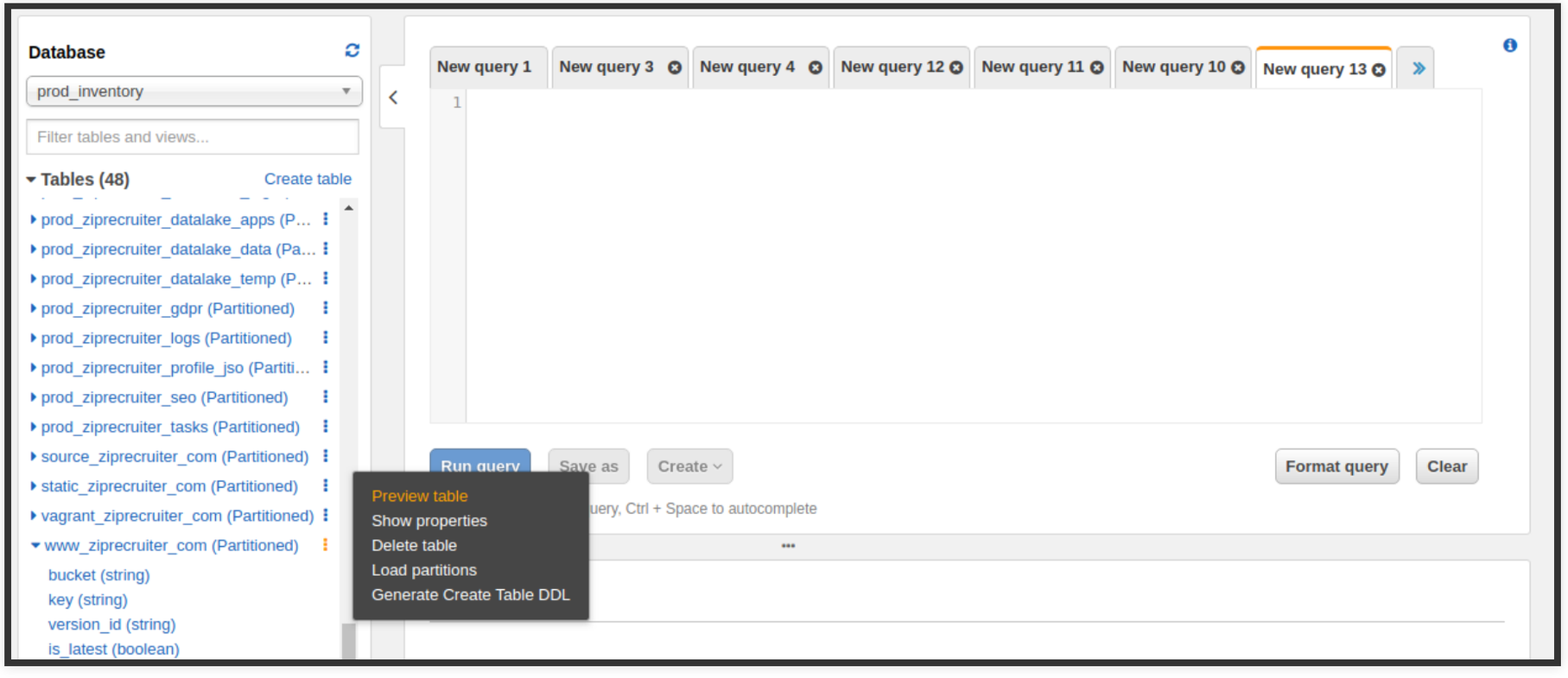
FIN(N)

FIN

EXTRA IMAGES TODO


Services
Resource Groups
agrangaard @ ziprecruiter
N. Virginia
Athena
Query Editor
Saved Queries
History
AWS Glue Data Catalog
Help
What's new
10+
Settings
Tutorial
Database
prod_inventory
adhoc
awsutilizationreporting
core
default
dev_alerts
dev_athena_rpt
dev_athena_rpt_temp
dev_event
dev_inventory
feed_import_chunk_...
findev_prod_ziprecru...
incoming_ziprecruite...
keys_ziprecruiter_co...
legal_prod_ziprecruit...
machine_learning_zi...
marketing_campaign...
outgoing_ziprecruiter...
partner_resumes_pr...
New query 1
New query 3
New query 4
New query 13
Run query
Save as
Create
Format query
Clear
Use Ctrl + Enter to run query, Ctrl + Space to autocomplete
Results


prod_coi (Partitioned)
prod_nosensitive_an...
prod_nosensitive_re...


Database
prod_inventory
Filter tables and views...
Tables (48)
Create table
prod_ziprecruiter_datalake_apps (P...
prod_ziprecruiter_datalake_data (Pa...
prod_ziprecruiter_datalake_temp (P...
prod_ziprecruiter_gdpr (Partitioned)
prod_ziprecruiter_logs (Partitioned)
prod_ziprecruiter_profile_jso (Partiti...
prod_ziprecruiter_seo (Partitioned)
prod_ziprecruiter_tasks (Partitioned)
source_ziprecruiter_com (Partitioned)
static_ziprecruiter_com (Partitioned)
vagrant_ziprecruiter_com (Partitioned)
www_ziprecruiter_com (Partitioned)
bucket (string)
key (string)
version_id (string)
is_latest (boolean)
Preview table
Show properties
Delete table
Load partitions
Generate Create Table DDL
New query 1
New query 3
New query 4
New query 12
New query 11
New query 10
New query 13
1
Run query
Save as
Create
Format query
Clear
uery, Ctrl + Space to autocomplete