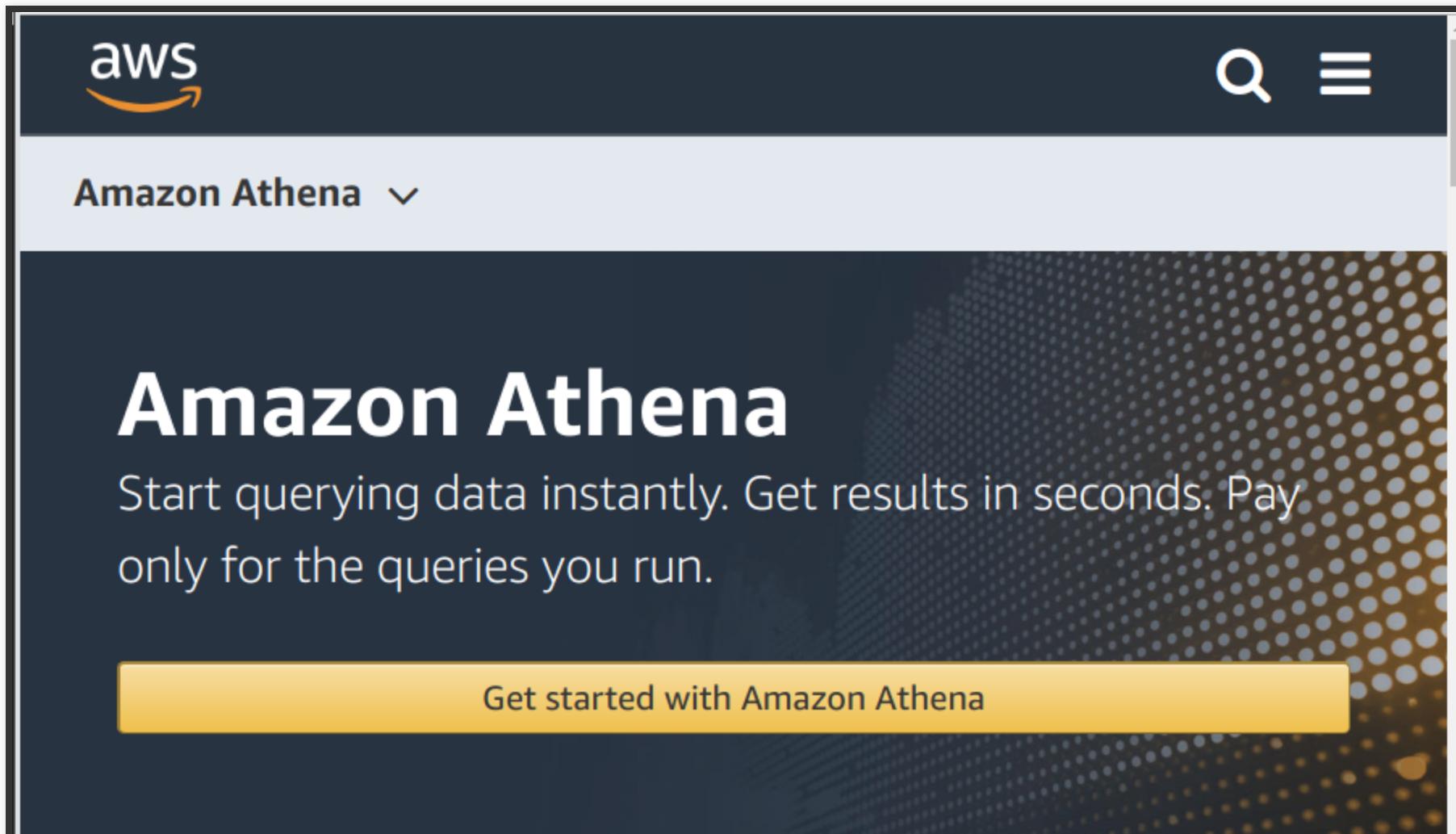


FANTASTIC DATUMS
A IS FOR ATHENA
G IS FOR GLUE
& WHERE TO FIND THEM

AMAZON ATHENA

Service to analyze data in S3 using SQL (prestoDB)



The screenshot shows the Amazon Athena landing page. At the top is the AWS navigation bar with the AWS logo and search bar. Below it, the page title is "Amazon Athena". The main heading is "Amazon Athena" in large white font. The subtext reads: "Start querying data instantly. Get results in seconds. Pay only for the queries you run." A prominent yellow button at the bottom has the text "Get started with Amazon Athena".

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to manage, and you pay only for the queries.

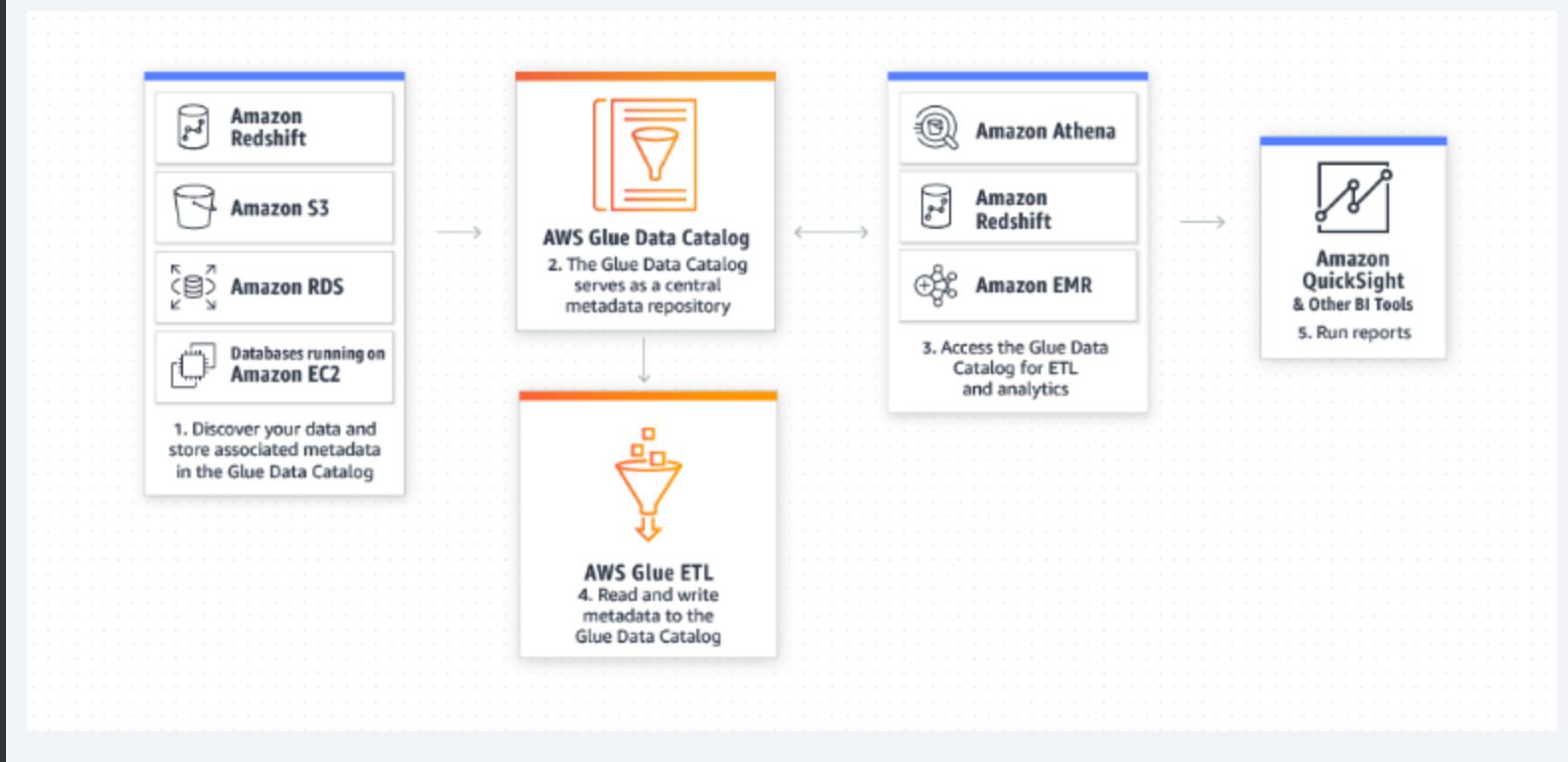
there is no infrastructure to manage, and you pay only for the queries that you run.

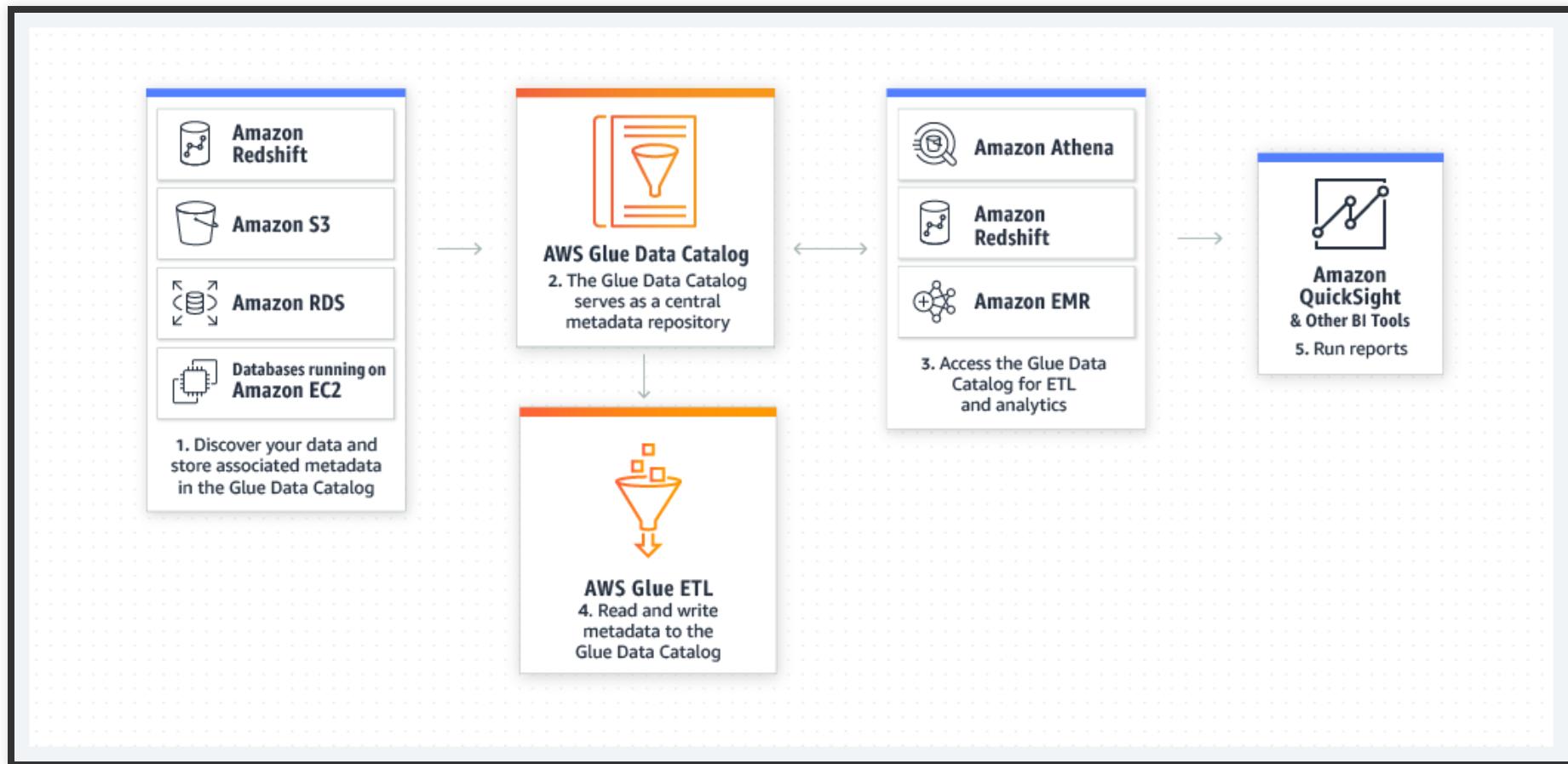
GLUE DATA CATALOG

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.





DATA

core

- cloudfare, 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

The AWS sign-in page shows the following fields:

- Account ID or alias: ziprecruiter
- AM user name: agrangaard
- Password: (redacted)

Sign In

[Sign-in using root account credentials](#)

Building Serverless Applications

Build and run applications and services without thinking about servers.

[Learn more »](#)





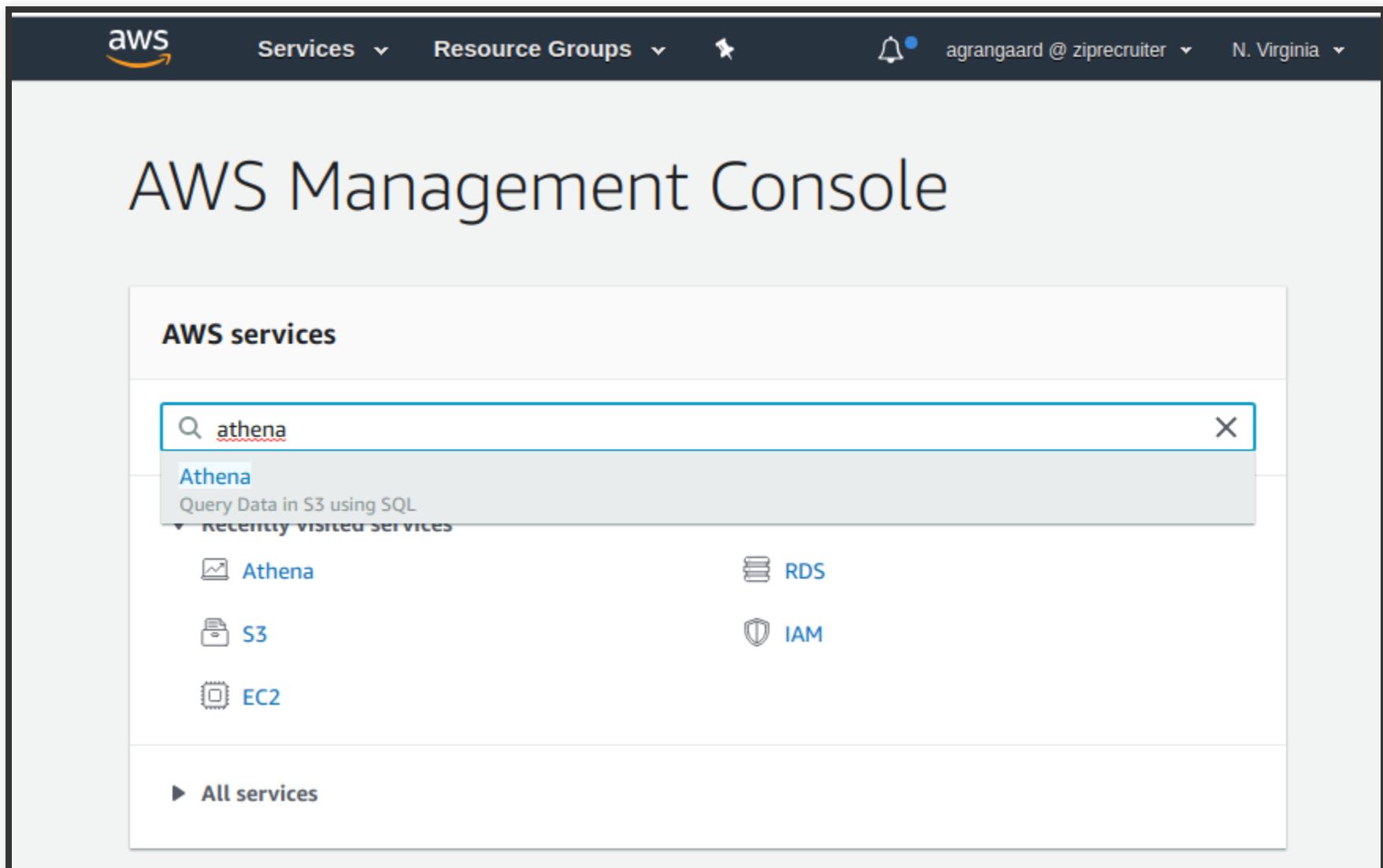
Multi-factor Authentication

Please enter an MFA code to complete sign-in.

MFA Code:

Submit

[Cancel](#)



AWS Athena Query Editor

Database: prod_inventory

Tables (48):

- analytics_ziprecruiter_com (Partitioned)
- append_only_job_corpus_ziprecruit...
- apt_prod_ziprecruiter_com (Partitioned)
- 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 (...)

Queries:

- New query 1
- New query 3
- New query 4
- New query 13 (selected)
- New query 12
- New query 11
- New query 10

Buttons:

- Run query
- Save as
- Create
- Format query
- Clear

Results:

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

SHOW DATABASES

The screenshot shows the AWS Athena Query Editor interface. At the top, there is a navigation bar with the AWS logo, 'Services' dropdown, 'Resource Groups' dropdown, a bell icon, user information 'agrangaard @ ziprecruiter', location 'N. Virginia', and a search bar. Below the navigation bar, the 'Athena' tab is selected, followed by 'Query Editor' (which is underlined), 'Saved Queries', 'History', 'AWS Glue Data Catalog' (with a blue arrow icon), 'Help', 'What's new' (with a '10+' badge), 'Settings', and 'Tutorial'.

On the left side, there is a sidebar titled 'Database' with a search bar and a list of databases. The database 'awsutilizationreporting' is currently selected and highlighted in blue. Other listed databases include 'prod_inventory', 'adhoc', 'core', 'default', 'dev_alerts', 'dev_athena_rpt', 'dev_athena_rpt_temp', 'dev_event', 'dev_inventory', 'dev_log_data', 'feed_import_chunk...', 'findev_prod_ziprecru...', 'incoming_ziprecru...', 'keys_ziprecruiter_co...', 'legal_prod_ziprecruit...', 'machine_learning_zi...', and 'marketing_campaign...'.

The main workspace shows a horizontal tab bar with four tabs: 'New query 1', 'New query 3', 'New query 4', and 'New query 13'. The fourth tab, 'New query 13', is currently active and highlighted in orange. Below the tabs, there is a large empty area where the results of the query would be displayed.

At the bottom of the workspace, there are several buttons: 'Run query' (blue), 'Save as', 'Create', 'Format query', and 'Clear'. A tooltip message 'Use Ctrl + Enter to run query, Ctrl + Space to autocomplete' is displayed above the 'Format query' button. Below these buttons, there is a section labeled 'Results' which is currently empty.

- ▶ outgoing_ziprecruiter... □
- ▶ partner_resumes_pr... □
- ▶ pik_prod_ziprecruiter... □
- ▶ prod_coi (Partitioned) □
- ▶ prod_nosensitive_an... □
- ▶ prod_nosensitive_re... □

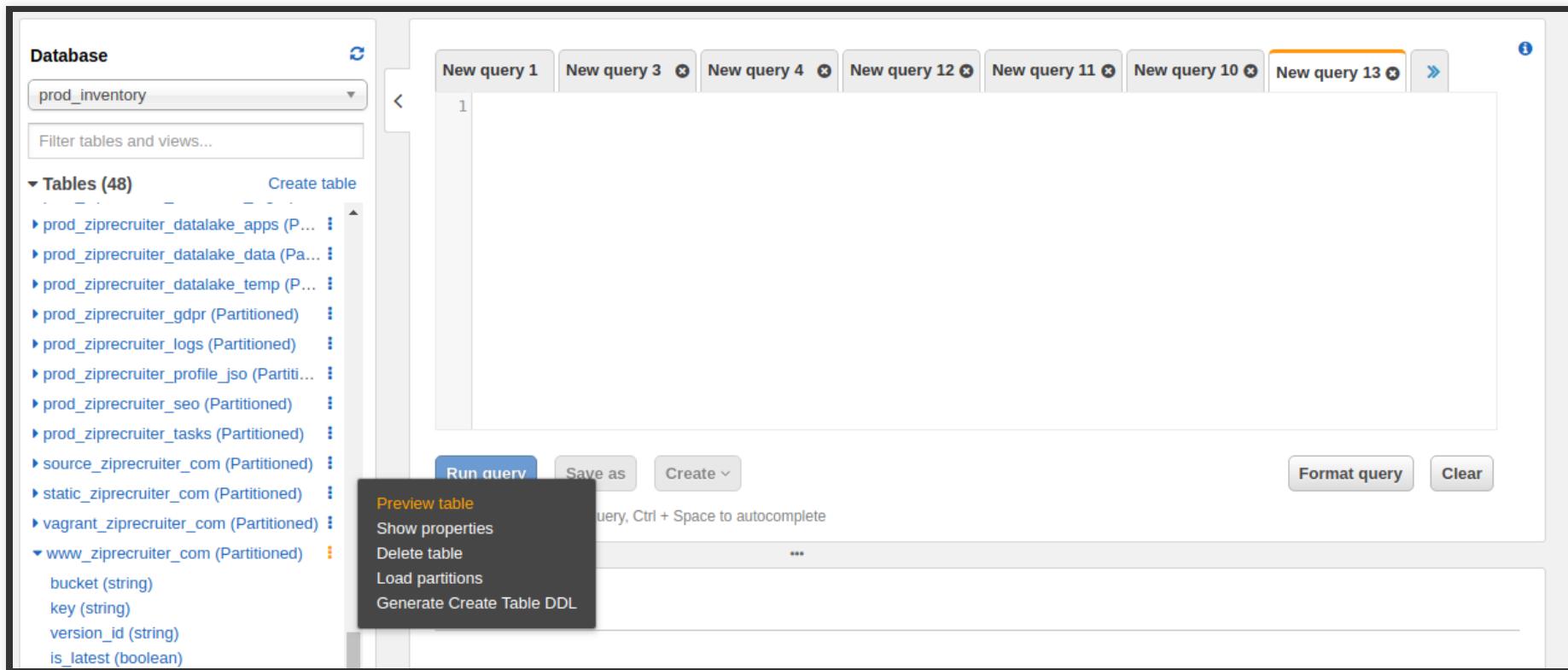
SHOW TABLES IN DATABASE

The screenshot shows the AWS Athena Query Editor interface. At the top, there's a navigation bar with the AWS logo, 'Services' dropdown, 'Resource Groups' dropdown, a bell icon, user information ('agrangaard @ ziprecruiter'), location ('N. Virginia'), and 'Support' dropdown. Below the navigation bar, the 'Athena' tab is selected, followed by 'Query Editor' (which is underlined), 'Saved Queries', 'History', and 'AWS Glue Data Catalog'. On the far right, there are 'Settings', 'Tutorial', 'Help', and 'What's new' links, with a '10+' badge above 'What's new'.

The main area is divided into several sections:

- Database:** A dropdown menu set to 'prod_inventory'. Below it is a 'Filter tables and views...' input field.
- Tables (48):** A list of 48 tables, each preceded by a blue triangle icon indicating they are expandable. The list includes:
 - analytics_ziprecruiter_com (Partitioned)
 - append_only_job_corpus_ziprecruit...
 - apt_prod_ziprecruiter_com (Partition...
 - artifacts_prod_ziprecruiter_com (Par...
 - blog_backup_buzzid103862 (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 (...)
- Queries:** A row of tabs for managing queries: 'New query 1', 'New query 3', 'New query 4', 'New query 13' (which is currently active), 'New query 12', 'New query 11', 'New query 10', and a 'Next' button.
- Action Buttons:** Buttons for 'Run query', 'Save as', 'Create', 'Format query', and 'Clear'.
- Text Input:** A placeholder text 'Use Ctrl + Enter to run query, Ctrl + Space to autocomplete'.
- Results:** A large, empty table where results will be displayed.

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

Database: core

Tables (9): dev_kafka_connect_test, dev_kafka_connect_unified_logs, dev_unified_logs, log.cloudflare, log_cloudfail, log_elb_ue1, log_prod_elb_ue1, prod_unified_logs, stg_unified_logs

Views (0)

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_size_tb
  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-latest.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-latest.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_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

AWS Services Resource Groups agrangaard @ ziprecruiter N. Virginia Sub

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
- dev_log_data
- feed_import_chunk...
- findev_prod_ziprecru...
- incoming_ziprecruite...
- keys_ziprecruiter_co...
- legal_prod_ziprecruit...
- machine_learning_zi...
- marketing_campaign...
- outgoing_ziprecruiter...
- partner_resumes_pr...
- pik_prod_ziprecruiter...

New query 1 New query 3 × New query 4 × **New query 13 ×** > i

Run query Save as Create Format query Clear

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

Results

This screenshot shows the AWS Athena Query Editor interface. The top navigation bar includes links for Services, Resource Groups, a user profile (agrangaard @ ziprecruiter), and regions (N. Virginia). Below the navigation is a secondary menu with links for Athena, Query Editor (which is selected and highlighted in blue), Saved Queries, History, AWS Glue Data Catalog, Help, and What's new (with a notification count of 10+). Further down are links for Settings and Tutorial.

The main workspace is divided into several sections. On the left, there is a sidebar titled "Database" containing a list of available databases. One database, "awsutilizationreporting", is currently selected and highlighted in blue. The list also includes "prod_inventory", "adhoc", "core", "default", "dev_alerts", "dev_athena_rpt", "dev_athena_rpt_temp", "dev_event", "dev_inventory", "dev_log_data", and several other entries starting with "feed_import_chunk...", "findev_prod_ziprecru...", etc. A search bar is located within the database sidebar.

In the center, a horizontal row of buttons allows users to "New query", "Run query", "Save as", "Create", "Format query", and "Clear". Below this row is a placeholder text: "Use Ctrl + Enter to run query, Ctrl + Space to autocomplete".

The right side of the interface features a large, empty "Results" pane where query results would be displayed.

- ▶ prod_coi (Partitioned) ▶
- ▶ prod_nosensitive_an... ▶
- ▶ prod_nosensitive_re... ▶

Database 

prod_inventory 

Filter tables and views...

Tables (48) 

- prod_ziprecruiter_datalake_apps (Partitioned)
- prod_ziprecruiter_datalake_data (Partitioned)
- prod_ziprecruiter_datalake_temp (Partitioned)
- prod_ziprecruiter_gdpr (Partitioned)
- prod_ziprecruiter_logs (Partitioned)
- prod_ziprecruiter_profile_jso (Partitioned)
- 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)

New query 1 New query 3  New query 4  New query 12  New query 11  New query 10  New query 13  New query 13 

1

Run query Save as Create  Format query Clear

Preview table Ctrl + Space to autocomplete
Show properties
Delete table
Load partitions
Generate Create Table DDL

...