

Curso AWS Data Analytics for IoT

Índice

- [Visão geral do curso](#)
- [Quais as soluções AWS estudadas e o que cada uma atende?](#)
- [Explique três exemplos de atividades que você realizou no laboratório prático](#)
- [Quais as principais lições apreendidas do curso?](#)

Visão geral do curso

O curso **Data Analytics** fornecido pela *AWS Academy* é focado no desenvolvimento das habilidades necessárias para análise de dados e big data utilizando as ferramentas e serviços fornecidos pela AWS. O curso é subdividido em nove laboratórios. O foco dos laboratórios está no desenvolvimento do conhecimento relacionado a utilização do conjunto de ferramentas da AWS.

Quais as soluções AWS estudadas e o que cada uma atende?

O desenvolvimento dos oito laboratórios envolveu a utilização de sete soluções oferecidas pela AWS com foco no processamento de dados para IoT. Entre as muitas soluções disponibilizadas pela AWS, foram exploradas: 1) Amazon Athena, 2) Amazon Redshift, 3) Amazon S3, 4) AWS Glue, 5) AWS IoT Analytics e 6) Amazon Kinesis Data Firehose e 7) Amazon Elasticsearch Service (Amazon ES).

No contexto de soluções, as ferramentas oferecidas pela AWS podem se confundir e por isso explicarei de forma individual o que cada uma delas atende.

1. Amazon Athena

Amazon Athena é um serviço de análise interativo que permite consultar e analisar dados armazenados no Amazon S3 usando SQL padrão. Ele foi projetado para tornar a análise de dados mais rápida, simples e escalável, sem a necessidade de configurar e gerenciar um banco de dados ou infraestrutura complexa. Ele pode ser utilizado para analisar tanto dados não estruturados como dados semiestruturados armazenados no S3.

2. Amazon Redshift

O **Amazon Redshift** é um serviço de *data warehouse* totalmente gerenciável, o que significa que foi projetado para assumir todas as tarefas e responsabilidades operacionais sem a necessidade de intervenção direta do usuário. Isso permite que os usuários se concentrem exclusivamente na análise de dados, sem se preocupar com a complexidade da infraestrutura subjacente. Sua principal aplicação é realizar análises e consultas em grandes volumes de dados. Ele é especialmente útil para empresas e organizações que precisam analisar dados estruturados em busca de insights valiosos. Através do Redshift, os usuários podem executar consultas complexas em tempo hábil, permitindo a extração de informações importantes para tomada de decisões estratégicas.

Uma das características distintivas do Amazon Redshift é sua arquitetura de banco de dados colunar, essa abordagem permite que o Redshift processe consultas analíticas e agregações rapidamente, tornando-o uma escolha ideal para analisar grandes volumes de dados de maneira eficiente. Além disso ele é totalmente escalável, o que significa que pode crescer de acordo com as necessidades das empresas, acomodando volumes de dados em constante expansão. Sua integração com outras ferramentas da AWS e compatibilidade com diversas aplicações de Business Intelligence (BI) facilitam a visualização e comunicação dos resultados da análise.

3. Amazon S3

O **Amazon S3** (Simple Storage Service) é um serviço de armazenamento de objetos altamente escalável e durável. Ele foi criado para armazenar e recuperar grandes quantidades de dados de forma segura e eficiente. Entre as ferramentas disponibilizadas para armazenamento de dados em nuvem, o S3 é uma das opções mais populares devido à sua confiabilidade, acessibilidade e facilidade de uso.

No S3, os dados são armazenados em "objetos", que podem ser qualquer tipo de arquivo digital, como imagens, vídeos, áudios, documentos de texto e muito mais. Cada objeto é identificado por uma chave única e pode ser acessado através de URLs (Uniform Resource Locators) específicos fornecidos pelo serviço.

4. AWS Glue

O **AWS Glue** é um serviço de ETL (Extract, Transform, Load). O ETL é um processo essencial no contexto de análise de dados, que envolve a extração de dados de várias fontes, sua transformação para um formato adequado e a carga dos dados em um destino, como um *data warehouse* ou *data lake*, para fins de análise e tratamento.

O principal objetivo do AWS Glue é simplificar e automatizar o processo de ETL. Ele oferece uma plataforma para a criação, execução e agendamento de fluxos de trabalho ETL sem a necessidade de configurar ou gerenciar a infraestrutura subjacente. Isso permite que os usuários se concentrem na lógica de transformação dos dados, em vez de se preocuparem com a complexidade do ambiente de ETL.

5. AWS IoT Analytics

O AWS IoT Analytics é um serviço que permite a análise de dados coletados de dispositivos como sensores, medidores e máquinas, que além de estar conectados à internet também podem coletar e transmitir dados.

O AWS IoT Analytics facilita a ingestão, processamento, armazenamento e análise de grandes volumes de dados gerados por esses dispositivos IoT. Ele fornece uma plataforma escalável para processar e obter *insights* significativos desses dados, permitindo que as decisões sejam realizadas com base nas informações coletadas.

6. Amazon Kinesis Data Firehose

O **Amazon Kinesis Data Firehose** é um serviço que permite a captura, transformação e carregamento de dados de *streaming* em tempo real para armazenamento e análise. Ele faz parte da família de serviços do Amazon Kinesis, que é projetada para lidar com dados em *streaming* de maneira escalável e eficiente.

O Kinesis Data Firehose é principalmente utilizado para coletar e processar grandes volumes de dados em tempo real, vindos de várias fontes, como dispositivos IoT, logs de aplicativos, eventos de sites, feeds de redes sociais, entre outros. O serviço facilita a entrega desses dados de *streaming* para destinos como o Amazon S3,

o Amazon Redshift, o Amazon Elasticsearch ou até mesmo para a análise em tempo real com o Amazon Kinesis Data Analytics.

7. Amazon Elasticsearch Service

O **Amazon Elasticsearch Service** é um serviço que permite criar, executar e escalar *clusters* Elasticsearch de maneira fácil e eficiente. O Elasticsearch é uma poderosa ferramenta de busca e análise de dados em tempo real, amplamente utilizada para indexar, pesquisar e visualizar grandes volumes de dados não estruturados.

Utilizando o Elasticsearch Service, os usuários podem implantar e configurar *clusters* Elasticsearch sem a necessidade de gerenciar a infraestrutura subjacente como provisionamento de servidores, ajuste de desempenho, aplicação de *patches* e *backups*. Sua implantação permite que os usuários se concentrem na análise de dados em vez de tarefas de gerenciamento de infraestrutura.

8. Identity and Access Management - IAM

O IAM (Identity and Access Management) é um serviço presente na Amazon Web Services (AWS) que permite gerenciar o acesso aos recursos e serviços da AWS de forma segura. Com esse serviço é possível criar e gerenciar identidades (como usuários, grupos e funções) e definir permissões e políticas para controlar o que essas identidades podem fazer nos demais serviços da AWS.

A utilização do IAM é altamente recomendável e até mesmo considerada fundamental para garantir a segurança dos recursos da AWS e para cumprir práticas recomendadas de segurança, como a separação de funções e o controle granular de permissões.

Explique três exemplos de atividades que você realizou no laboratório prático

Lab 1: Store data in Amazon S3

Serviços Utilizados: Access Amazon S3, IAM

Objetivo: Acessar o Amazon Redshift via Console de Gerenciamento da AWS, criar um cluster, carregar e consultar dados do S3 para o Redshift.

Atividades realizadas:

1. Criando um usuário e adicionando ele no grupo

Grupos de usuários (1) [Informações](#)

Um grupo de usuários é um conjunto de usuários do IAM. Use grupos para especificar as permissões para um conjunto de usuários.

Q

Filtre Grupos de usuários por propriedade ou nome de grupo e pressione Enter

☐

Nome do grupo

▼

Usuário

☐

awsusers

Usuários (1) [Informações](#)

Um usuário do IAM é uma identidade com credenciais de longo prazo que é usada para interagir com a AWS em uma conta.

Q

Encontrar usuários por nome de usuário ou chave de acesso

☐

Nome do usuário

▼

Grupos

▼

Última atividade

☐

awsuser

awsusers

Nunca

2. Criando um bucket e carregando um arquivo não comprimido e comprimido

4 / 11

criar bucket

Informações

Os buckets são contêineres para dados armazenados no S3. Saiba mais

Configuração geral

Nome do bucket

bucket42alf

O nome do bucket deve ser exclusivo no namespace global e seguir as regras de nomenclatura do bucket. Veja as regras para nomenclatura de buckets

Região da AWS

Leste dos EUA (Norte da Virgínia) us-east-1

Copiar configurações do bucket existente - opcional

Somente as configurações de bucket na configuração a seguir são copiadas.

Escolher bucket

Carregar

Informações

Adicione os arquivos e pastas que você deseja carregar no S3. Para fazer upload de um arquivo maior que 160 GB, use a AWS CLI, o SDK da AWS ou a API REST do Amazon S3. Saiba mais

Arraste e solte aqui os arquivos e pastas para upload ou selecione Adicionar arquivos ou Adicionar pastas.

Arquivos e pastas (1 Total, 337.0 B)

Remover

Adicionar arquivos

Adicionar pasta

Todos os arquivos e pastas desta tabela serão carregados.

Encontrar por nome

< 1 >

	Nome	Pasta	Tipo	Tamanho
<input type="checkbox"/>	lab1.csv	-	text/csv	337.0 B

3. Realizando uma consulta para confirmar que o arquivo foi carregado

5 / 11

Consulta SQL

O Amazon S3 Select oferece suporte apenas ao comando SELECT do SQL. Usando o console do S3, você pode extrair até REST do Amazon S3. Para consultas SQL mais complexas, use o [Amazon Athena](#)

Adicionar SQL a partir de modelos

Executar consulta SQL

```
1 /* Para criar um ponto de referência para gravar consultas SQL, você pode exibir os 5 primeiros  
2 SELECT * FROM s3object s LIMIT 5
```

Resultados da consulta

Os resultados da consulta não estarão disponíveis depois que você escolher **Fechar** ou navegar para outra página. Escolha

Status

✓ 5 Registros retornados com êxito em 978 ms

Bytes retornados: 337 B

Bruto

Formatados

```
CustomerID,First Name,Last Name,Join Date,Street Address,City,State,Phone  
001,Alejandro,Rosalez,12/12/2013,123 Main St.,Baltimore,MD,765-234-2349  
002,Jane,Doe,10/5/2014,456 State St.,Seattle,WA,415-889-4932  
003,John,Stiles,9/20/20016,1980 8th St.,Brooklyn,NY,917-123-9308  
004,Li,Juan,6/29/2011,1323 22nd Ave.,Albany,NY,917-332-3432
```

4. Mudando as propriedades de encriptação
5. Carregando um arquivo não comprimido e comprimido e confirmando seu carregamento

Lab 2: Query Data in Amazon Athena

Serviços Utilizados: Amazon Athena, S3

Objetivo: Acessar o Amazon Athena via Console de Gerenciamento da AWS, criar um banco de dados no Athena, criar uma tabela no Athena e otimizar um banco de dados Athena.

Atividades realizadas:

1. Acessar o S3 (Amazon Simple Storage Service) e copiar o ARN

Amazon S3 > Buckets > c87854a1876996l4372382t1w656541538352-s3bucket-1oo0468vrlyai

c87854a1876996l4372382t1w656541538352-s3bucket-1oo0468vrlyai

Informações

Objetos

Propriedades

Permissões

Métricas

Gerenciamento

Pontos de acesso

Visão geral do bucket

Região da AWS

Leste dos EUA (Norte da Virgínia) us-east-1

Nome de recurso da Amazon (ARN)

 am:aws:s3::c87854a1876996l4372382t1w656541538352-s3bucket-1oo0468vrlyai

Data de criação

26 Jul 2023 07:23:49 PM -03

2. Acessar o Athena e configurar o "Result Location"

Manage settings

Query result location and encryption


Location of query result - optional

Enter an S3 prefix in the current region where the query result will be saved as an object.


Q s3://c87854a1876996l4372382t1w656541538352-s3bucket-1oo04t

X

View




Browse S3




You can create and manage lifecycle rules for this bucket

Use Amazon S3 lifecycle rules to store your query results and metadata cost effectively or to delete them after a period of time.

[Learn more](#) 

Lifecycle configuration



Expected bucket owner - optional

Specify the AWS account ID that you expect to be the owner of your query results output location bucket.

Enter AWS account ID

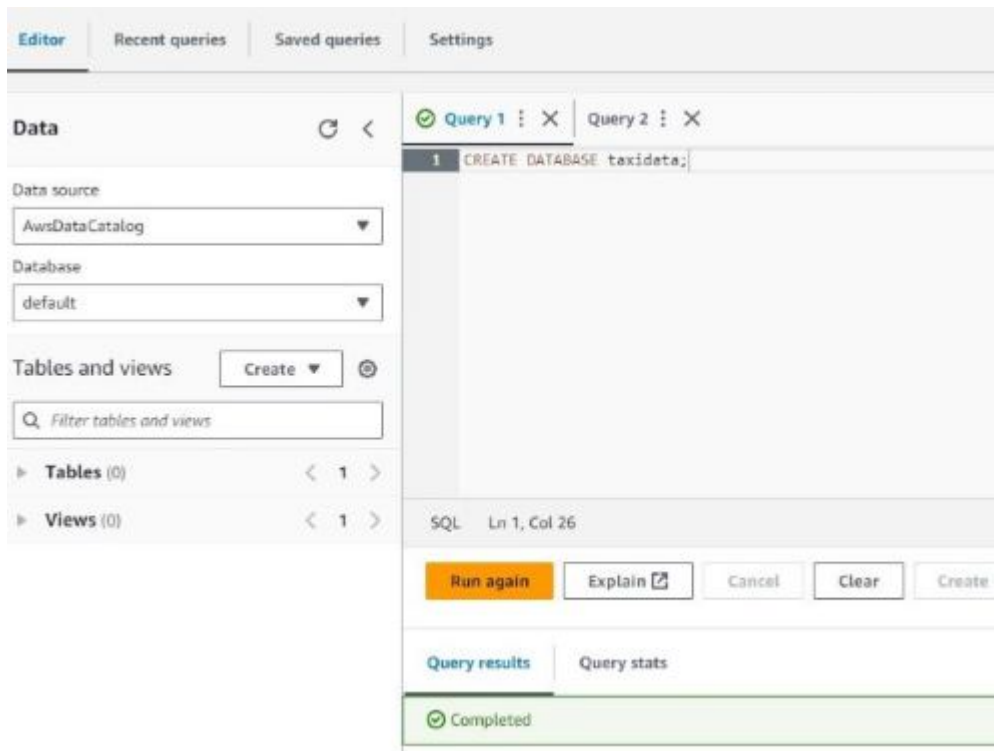
☐ Assign bucket owner full control over query results

Enabling this option grants the owner of the S3 query results bucket full control over the query results. This means that if your query result location is owned by another account, you grant full control over your query results to the other account.

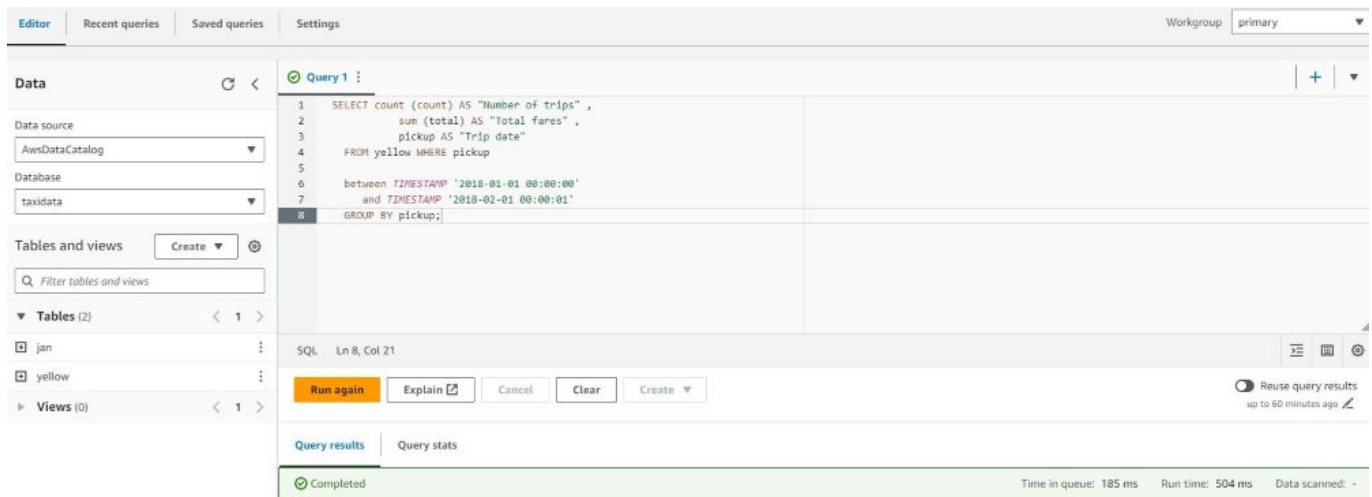
☐ Encrypt query results

4. Criar uma query em SQL

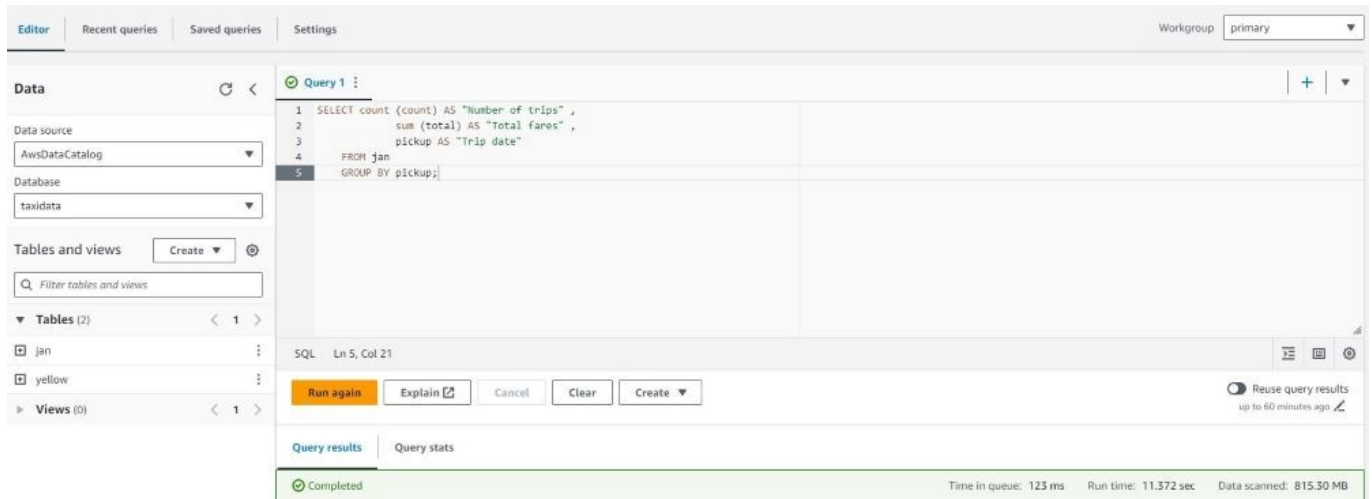
7 / 11



5. Criar uma tabela usando o bucket criado no S3 para armazenar os resultados das consultas executadas no Athena
 6. Otimização do Banco de dados por meio do particionamento da tabela
- Utilizando consulta para dados não divididos em buckets

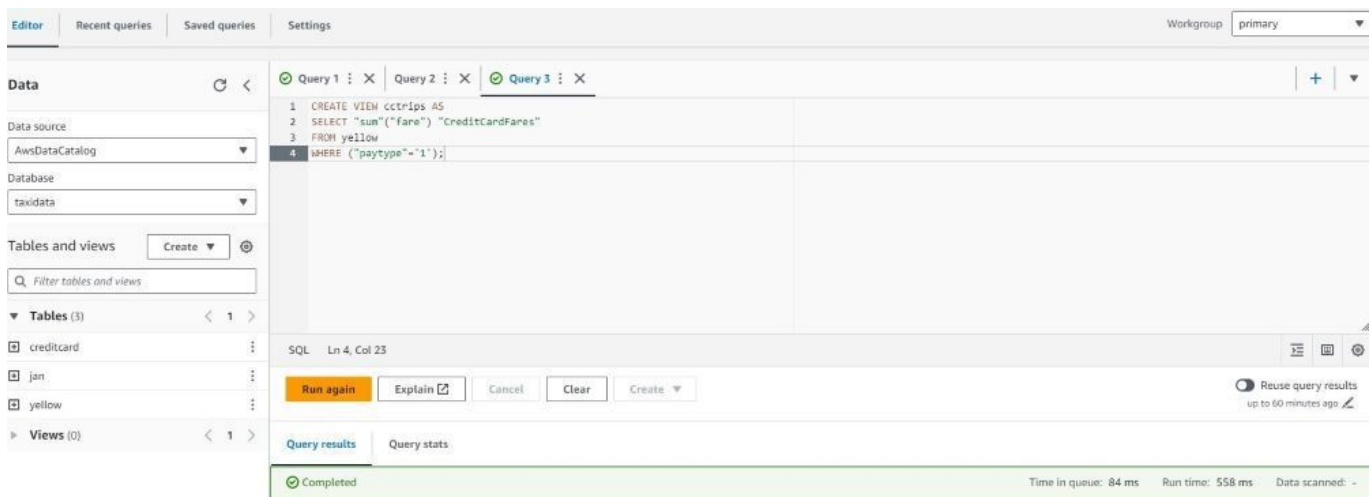


- Utilizando consulta para dados divididos em buckets



7. Testando o particionando dos dados

8. Criar views no Athena com o objetivo de ocultar a complexidade das consultas e otimizar o desempenho



Lab 3: Query data in Amazon S3 with Amazon Athena and AWS Glue

Serviços Utilizados: Amazon Glue, S3

Objetivo: Acessar o AWS Glue via Console de Gerenciamento da AWS, criar um rastreador (crawler) e um banco de dados no AWS Glue, consultar dados do S3.

Atividades realizadas:

1. Acessando o AWS Glue e realizando as configurações iniciais do rastreador

AWS Glue > Crawlers > Add crawler

Step 1
Set crawler properties

Step 2
Choose data sources and classifiers

Step 3
Configure security settings

Step 4
Set output and scheduling

Step 5
Review and create

Review and create

Step 1: Set crawler properties

Set crawler properties

Name

wather

Description

-

Tags

-

Step 2: Choose data sources and classifiers

Choose data sources and classifiers

Data sources (1) info

The list of data sources to be scanned by the crawler.

Type

Data source

Parameters

S3

s3://c87854a1876998143779781w794395796128-querybuc...

Recrawl all

Step 3: Configure security settings

Configure security settings

IAM role

gluelab

Security configuration

-

Lake Formation configuration

-

Step 4: Set output and scheduling

Set output and scheduling

Database

weatherdata

Table prefix - optional

-

Maximum table threshold - optional

-

Schedule

On demand

Cancel

Previous

Create crawler

2. Iniciando o rastreador
3. Revisando os metadados criados pelo AWS Glue

Tables

table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

Tables (1)

View and manage all available tables.

Last updated (UTC)
July 27, 2023 at 24:24:55

Refresh

Delete

Data quality New

Add tables using crawler

Add table

Filter tables

<input type="checkbox"/>	Name	Database	Location	Classification	Deprecated	View data
<input type="checkbox"/>	csv	weatherdata	s3://noaa-ghcn-pds/csv/	CSV	-	Table data

4. Editando o schema criado
5. Consultando a tabela usando o AWS Glue Data Catalog
6. Criando uma tabela

Data

Data source
AwsDataCatalog

Database
weatherdata

Tables and views
Filter tables and views

Tables (2)

csv
Partitioned

late20th

Views (0)

Query 1 : X

Query 2 : X

Query 3 : X

1 CREATE table weatherdata.late20th

2 WITH (

3 format='PARQUET', external_location='s3://c87854a1876998143779781w794395796128-querybucket-4hh2rytsecv0/1ab3/'

4) AS SELECT date, type, observation FROM csv

5 WHERE date/10000 between 1950 and 2015;

SQL Ln 3, Col 107

Run again

Explain

Cancel

Clear

Create

Query results

Query stats

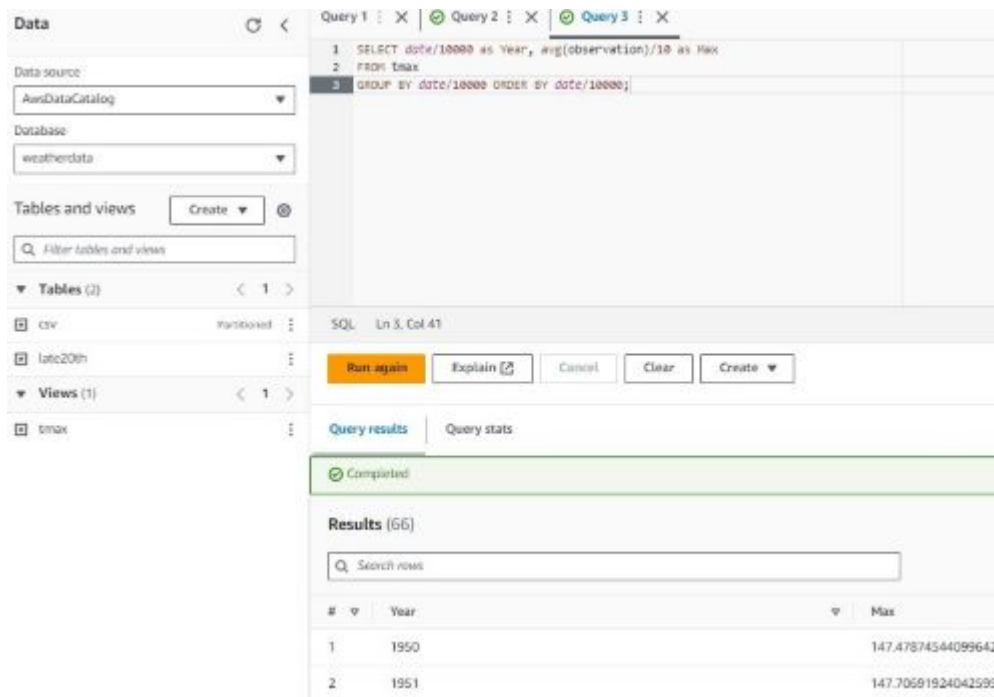
Completed

Time in queue: 2.724 sec

Run time: 1 min 57.963 sec

Data scanned: 200.71 GB

7. Criando uma vis o e executando uma consulta a partir dos dados selecionados
8. Executando uma consulta



The screenshot displays the AWS Data Catalog console interface. On the left, the 'Data' sidebar shows the 'Data source' as 'AwsDataCatalog' and the 'Database' as 'weatherdata'. Below this, a list of tables and views is shown, including 'csv', 'late20th', and 'tmax'. The main panel shows a SQL query editor with the following query:

```
1 SELECT date/10000 as Year, avg(observation)/30 as Max
2 FROM tmax
3 GROUP BY date/10000 ORDER BY date/10000;
```

Below the query editor, there are buttons for 'Run again', 'Explain', 'Cancel', 'Clear', and 'Create'. The 'Query results' tab is active, showing a 'Completed' status. The results are displayed in a table with 66 rows, showing the average maximum temperature for each year from 1950 to 1951.

#	Year	Max
1	1950	147.47874544099642
2	1951	147.70691924042595

Quais as principais lições aprendidas do curso?

A realização do curso foi de extrema importância para minha formação na área de tecnologia, principalmente relacionada a minha área de atuação que se enquadra no contexto da ciência de dados. Essa experiência foi ainda mais enriquecedora para mim, uma vez que pude estar em contato com ferramentas tão importantes como as oferecidas pela AWS.

Algumas lições importante foram retiradas dessa experiência como a autonomia necessária para conclusão dos laboratórios, além dessa lição posso citar inúmeras outras como:

- Conhecimento das tecnologias da AWS
- Experiência com a análise de dados em nuvem
- Conhecimento relacionado ao processamento de dados em tempo real
- Conhecimento prático para integração entre os serviços AWS