

Faculdade

**XPe**



# RELATÓRIO

---

PROJETO  
APLICADO

XP Educação  
Relatório do Projeto Aplicado

# Do Legacy ao Cloud: Uma Solução para Pipelines de Dados Eficientes

Ana Carolina Anastácio

Orientador(a): Moisés Luna

09/04/2025



ANA CAROLINA ANASTÁCIO

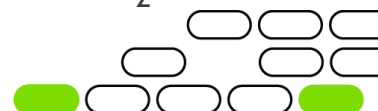
XP EDUCAÇÃO

RELATÓRIO DO PROJETO APLICADO

# Do Legacy ao Cloud: Uma Solução para Pipelines de Dados Eficientes

Relatório de Projeto Aplicado  
desenvolvido para fins de conclusão do  
curso Arquitetura e Engenharia de Dados  
com IA.

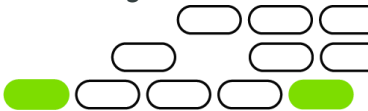
Orientador (a): Moisés Luna



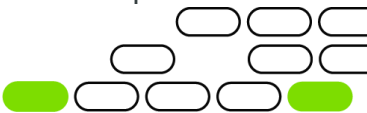
Cataguases, Minas Gerais  
18/03/2025

Sumário

1. CANVAS do Projeto Aplicado	4
Desafio	5
1.1.1 Análise de Contexto	5
1.1.2 Personas	6
1.1.3 Benefícios e Justificativas	7
1.1.4 Hipóteses	8
1.2 Solução	9
1.2.1 Objetivo SMART	9
1.2.2 Premissas e Restrições	11
1.2.3 Backlog de Produto	13
2. Área de Experimentação	162.1 Sprint 1
	16
2.1.1 Solução	16
Evidência do planejamento:	16
Evidência da execução de cada requisito:	16
Evidência dos resultados:	16
2.1.2 Lições Aprendidas	16
2.2 Sprint 2	17
2.2.1 Solução	17
Evidência do planejamento:	17
Evidência da execução de cada requisito:	17
Evidência dos resultados:	17



2.2.2 Lições Aprendidas	17
2.3 Sprint 3	18
2.3.1 Solução	18
Evidência do planejamento:	18
Evidência da execução de cada requisito:	18
Evidência dos resultados:	18
2.3.2 Lições Aprendidas	18
3. Considerações Finais	243.1 Resultados
	19
3.2 Contribuições	19
3.3 Próximos passos	24



## 1. CANVAS do Projeto Aplicado



### 1.1 Desafio

#### 1.1.1 Análise de Contexto

A empresa **SunTech** (fictícia) é uma empresa de médio porte que atua no setor elétrico, oferecendo soluções de análise de dados para clientes. Atualmente, eles utilizam o **Oracle Data Integrator (ODI)** para criar pipelines de dados, extraindo informações de sistemas transacionais, transformando e carregando em um **Data Warehouse (DW)** no Oracle Developer. No entanto, enfrentam os seguintes desafios:

- **Desempenho insuficiente:** O ODI não consegue processar grandes volumes de dados de forma eficiente, resultando em pipelines lentos.
- **Falta de escalabilidade:** A infraestrutura atual não suporta o crescimento exponencial dos dados, limitando a capacidade de análise.
- **Dificuldade de manutenção:** Os processos no ODI são complexos e dependem de uma equipe reduzida, o que gera gargalos operacionais.
- **Limitações na análise em tempo real:** O DW atual não permite análises em tempo real, o que é crítico para decisões estratégicas no setor financeiro.
- **Custos elevados de manutenção:** A infraestrutura on-premises exige investimentos contínuos em hardware e licenças de software.

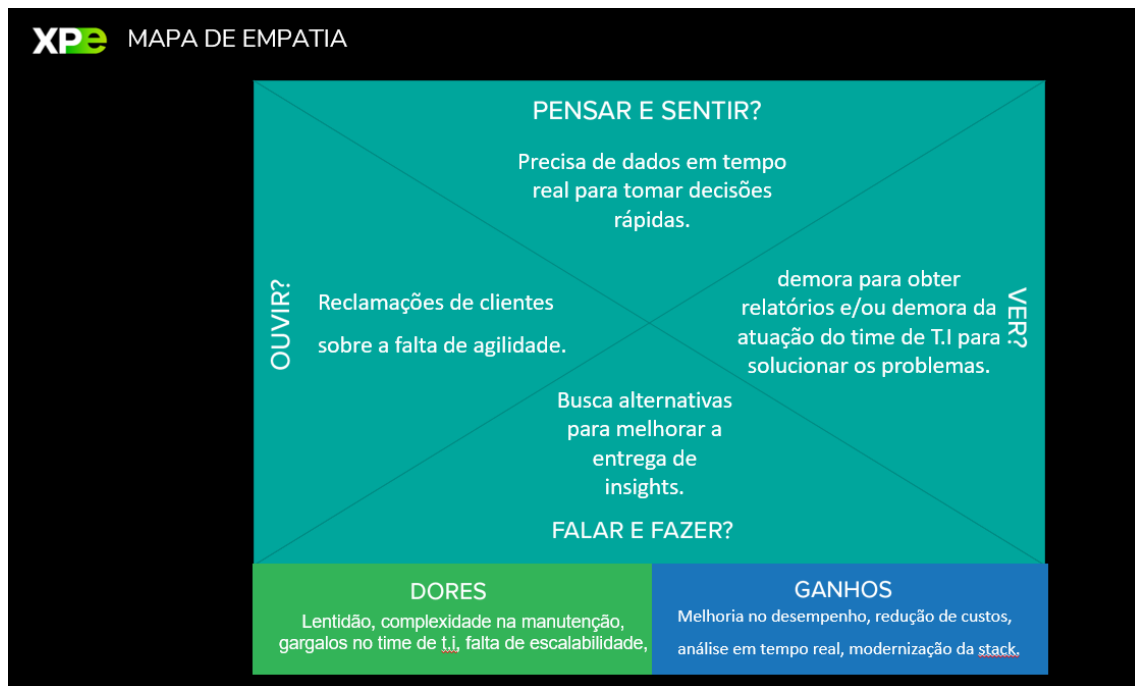
A decisão de modernizar a infraestrutura de dados baseou-se em uma série de evidências e pesquisas conduzidas de forma sistemática. Em primeiro lugar, o feedback da equipe técnica foi fundamental para identificar os gargalos operacionais. Engenheiros de dados e analistas relataram dificuldades recorrentes com a complexidade do ODI e a lentidão dos processos. Além disso, reclamações dos clientes destacaram a insatisfação com a demora para obter relatórios e insights, o que impacta diretamente a experiência do usuário final.

Pesquisas mostraram que empresas líderes estão migrando para soluções cloud-based, como AWS Glue e Amazon Redshift, que oferecem maior escalabilidade, desempenho e redução de custos. Essas ferramentas modernas permitem não apenas processar grandes volumes de dados de forma eficiente, mas também realizar análises em tempo real, algo que a infraestrutura atual da SunTech não suporta. Por fim, a análise de dados internos revelou que os custos operacionais da infraestrutura on-premises estão aumentando, enquanto a taxa de falhas nos processos ETL cresceu significativamente nos últimos anos.

A migração para uma solução cloud-based traria benefícios tangíveis e intangíveis para a SunTech. Em primeiro lugar, a melhoria no desempenho dos pipelines de dados permitiria processar informações de forma mais rápida e eficiente, reduzindo o tempo necessário para gerar relatórios e insights. Em segundo lugar, a escalabilidade oferecida por ferramentas como o AWS Glue e o Amazon Redshift garantiria que a empresa possa lidar com o crescimento contínuo dos dados sem precisar investir em infraestrutura física adicional.



### 1.1.2 Personas



### 1.1.3 Justificativas



#### Benefícios da Solução:

- Melhoria no desempenho: Pipelines mais rápidos e escaláveis.
- Redução de custos: Menos tempo gasto com manutenção e correção de erros.
- Análise em tempo real: Capacidade de oferecer insights imediatos para clientes.



- Modernização da stack: Adoção de ferramentas alinhadas com as tendências do mercado.

#### 1.1.4 Hipóteses

- Matriz de observações para hipóteses.

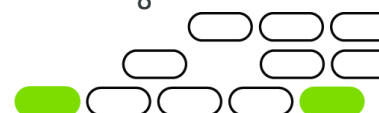


- Identificação do grau de riscos.

Riscos Associados às Suposições				
Suposição	Risco Técnico	Risco Financeiro	Risco Operacional	Grau de Risco
Migrar para o AWS Glue melhorará o desempenho.	Médio	Baixo	Médio	<b>Médio</b>
Adotar Apache Airflow reduzirá a complexidade.	Baixo	Baixo	Baixo	<b>Baixo</b>
Implementar um Data Lake no S3 permitirá análises em tempo real.	Alto	Médio	Alto	<b>Alto</b>
Migrar para a nuvem reduzirá os custos em 30%.	Médio	Baixo	Médio	<b>Médio</b>

Justificativas:

- AWS Glue: Risco técnico médio devido à curva de aprendizado, mas baixo risco financeiro e operacional.



- Apache Airflow: Riscos baixos, pois é uma ferramenta amplamente adotada e de fácil integração.
- Data Lake no S3: Risco alto devido à complexidade de implementação e migração de dados.
- Migração para a nuvem: Risco médio, pois envolve mudanças na infraestrutura, mas com benefícios comprovados.

Dúvida	Risco Técnico	Risco Financeiro	Risco Operacional	Grau de Risco
Qual será o custo total da migração para a nuvem?	Baixo	Alto	Médio	<b>Alto</b>
Quanto tempo levará para migrar todos os dados?	Médio	Médio	Alto	<b>Alto</b>
A equipe terá capacidade para aprender as novas ferramentas?	Médio	Baixo	Alto	<b>Médio</b>
Como garantir a segurança dos dados na nuvem?	Alto	Médio	Alto	<b>Alto</b>
Qual será o impacto nos clientes durante a migração?	Médio	Médio	Alto	<b>Alto</b>

#### Justificativas:

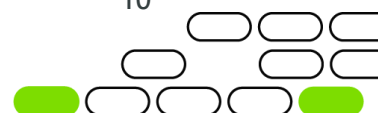
- Custo total da migração: Risco financeiro alto, pois pode haver custos inesperados.
- Tempo de migração: Risco operacional alto, pois a migração pode impactar os prazos.
- Capacidade da equipe: Risco operacional alto, pois a equipe pode enfrentar dificuldades no aprendizado.
- Segurança dos dados: Risco técnico alto, pois a segurança é crítica e qualquer falha pode ser catastrófica.
- Impacto nos clientes: Risco operacional alto, pois a migração pode causar interrupções.

- Priorização de Ideias.

Ideia	Impacto	Esforço	Prioridade
Migrar para o AWS Glue.	Alto	Médio	<b>Alta</b>
Adotar Apache Airflow.	Médio	Baixo	<b>Média</b>
Implementar um Data Lake no S3.	Alto	Alto	<b>Baixa</b>
Migrar para a nuvem para reduzir custos.	Alto	Médio	<b>Alta</b>

#### Critérios de Priorização:

- 1. Migrar para o AWS Glue:**
  - **Impacto Alto:** Melhora significativa no desempenho e escalabilidade.
  - **Esforço Médio:** Requer configuração e integração, mas é viável.
  - **Prioridade Alta:** Solução central para o problema de desempenho.
- 2. Adotar Apache Airflow:**
  - **Impacto Médio:** Reduz a complexidade de manutenção, mas não resolve todos os problemas.
  - **Esforço Baixo:** Fácil de implementar e integrar.
  - **Prioridade Média:** Complementar à migração para o AWS Glue.
- 3. Implementar um Data Lake no S3:**
  - **Impacto Alto:** Permite análises em tempo real.
  - **Esforço Alto:** Complexidade de implementação e migração.
  - **Prioridade Baixa:** Pode ser considerada em uma segunda fase.
- 4. Migrar para a nuvem para reduzir custos:**
  - **Impacto Alto:** Redução significativa de custos operacionais.
  - **Esforço Médio:** Requer planejamento e migração de dados.
  - **Prioridade Alta:** Alinhado com a migração para o AWS Glue.



## 1.2 Solução

### 1.2.1 Objetivo SMART

Implementar uma arquitetura de dados moderna na AWS, utilizando AWS Glue para ETL, Apache Airflow para orquestração e Data Lake no S3 para armazenamento, migrando 100% dos **pipelines críticos** do ODI em 6 meses, com redução de 50% no tempo de processamento e 30% nos custos operacionais, além de viabilizar análises em tempo real.

#### Critérios SMART Aplicados

##### 1. Específico (S)

- O quê? Migrar pipelines do ODI para AWS Glue + Airflow + S3.
- Como? Usando AWS Glue para ETL, Airflow para orquestração e S3 como Data Lake.
- Para quê? Reduzir tempo/custos e habilitar análises em tempo real.

##### 2. Mensurável (M)

- Métricas:
  - % de pipelines migrados (100% críticos).
  - Tempo de ETL (redução de 50%).
  - Custos operacionais (redução de 30%).

##### 3. Atingível (A)

- Recursos:
  - AWS Glue e Airflow já validados em POCs.
  - Equipe treinada (ou em treinamento) nas ferramentas.
- Escalonável: Migração em fases (piloto → produção).

##### 4. Relevante (R)



- Alinha-se às metas de:
  - TI: Modernização e redução de custos.
  - Negócio: Agilidade na geração de insights.

## 5. Temporal (T)

- Prazo total: 6 meses.
- Marcos:
  - Mês 2: Piloto com 20% dos pipelines.
  - Mês 4: 70% migrados.
  - Mês 6: 100% concluído + otimizações.

### 1.2.2 Escopo do Projeto

#### Premissas

Premissa	Impacto se Não For Verdadeira
A equipe conseguirá se capacitar em AWS Glue e Airflow dentro do prazo.	Atrasos na migração e aumento de custos com treinamentos adicionais ou contratação de especialistas.
A infraestrutura da AWS terá disponibilidade contínua durante o projeto.	Paralisação dos trabalhos e risco de não cumprimento dos prazos.
Os dados atuais no Oracle Developer estão íntegros e consistentes.	Necessidade de limpeza de dados durante a migração, aumentando o tempo e custo do projeto.
A diretoria aprovará o orçamento para migração e operação na AWS.	Limitação de recursos, podendo inviabilizar a adoção total da solução cloud.
Não haverá mudanças significativas nos requisitos durante o projeto.	Retrabalho e realinhamento de escopo, afetando prazos e custos.

## Restrições

Restrição	Como Impacta o Projeto
<b>Orçamento limitado</b> para contratação de serviços AWS.	Pode exigir otimização de recursos ou priorização de pipelines críticos.
<b>Prazo máximo de 6 meses</b> para conclusão.	Exige cronograma rigoroso e possível redução de escopo se houver atrasos.
<b>Conformidade com LGPD</b> e regulamentações.	Necessidade de revisão extra de segurança e governança de dados na AWS.
<b>Equipe enxuta</b> (sem recursos para contratações).	Sobrecarga da equipe interna e risco de burnout.
<b>Dependência de sistemas legados</b> que não podem ser desativados imediatamente.	Arquitetura híbrida temporária (ODI + AWS), aumentando complexidade.

## Como Mitigar Riscos Derivados de Premissas e Restrições

### 1. Para Premissas:

- **Testar pequenos cenários antes da migração total** (ex: validar capacitação da equipe com um piloto).
- **Ter um plano B** (ex: contrato com consultorias especializadas em AWS).

### 2. Para Restrições:

- **Otimizar recursos:** Usar instâncias AWS com custo-benefício (ex: Redshift Serverless).
- **Priorizar pipelines críticos:** Migrar primeiro os ETLs com maior impacto no negócio.
- **Monitorar compliance:** Usar AWS Artifact para garantir conformidade com LGPD.

### 1.2.3 Cronograma de Ações Planejada

#### Ferramentas Utilizadas

Ferramenta	Finalidade
Trello	Gestão das tarefas (Kanban com sprints).
AWS Glue	ETL (extração, transformação e carga).
Apache Airflow	Orquestração dos pipelines.
Amazon S3	Data Lake (armazenamento de dados brutos/processados).
Amazon Redshift	Data Warehouse (análises e dashboards).
AWS DMS	Migração dos dados do Oracle Developer para a AWS.
Git/GitHub	Versionamento de scripts (Python, SQL, DAGs do Airflow).
Power BI/Tableau	Visualização de dados (conectado ao Redshift).

---

#### Backlog de Requisitos

##### Requisitos Técnicos:

1. Configuração do Ambiente AWS
  - Criar contas e permissões (IAM).
  - Configurar VPC, segurança e acesso.
2. Migração de Dados
  - Configurar AWS DMS para replicar dados do Oracle Developer para o S3.
  - Validar integridade dos dados migrados.
3. ETL com AWS Glue
  - Criar jobs no Glue para transformar dados no S3.
  - Definir crawlers para catalogação automática.
4. Orquestração com Airflow
  - Desenvolver DAGs para agendamento e monitoramento dos pipelines.



- Configurar alertas para falhas.

## 5. Data Lake no S3

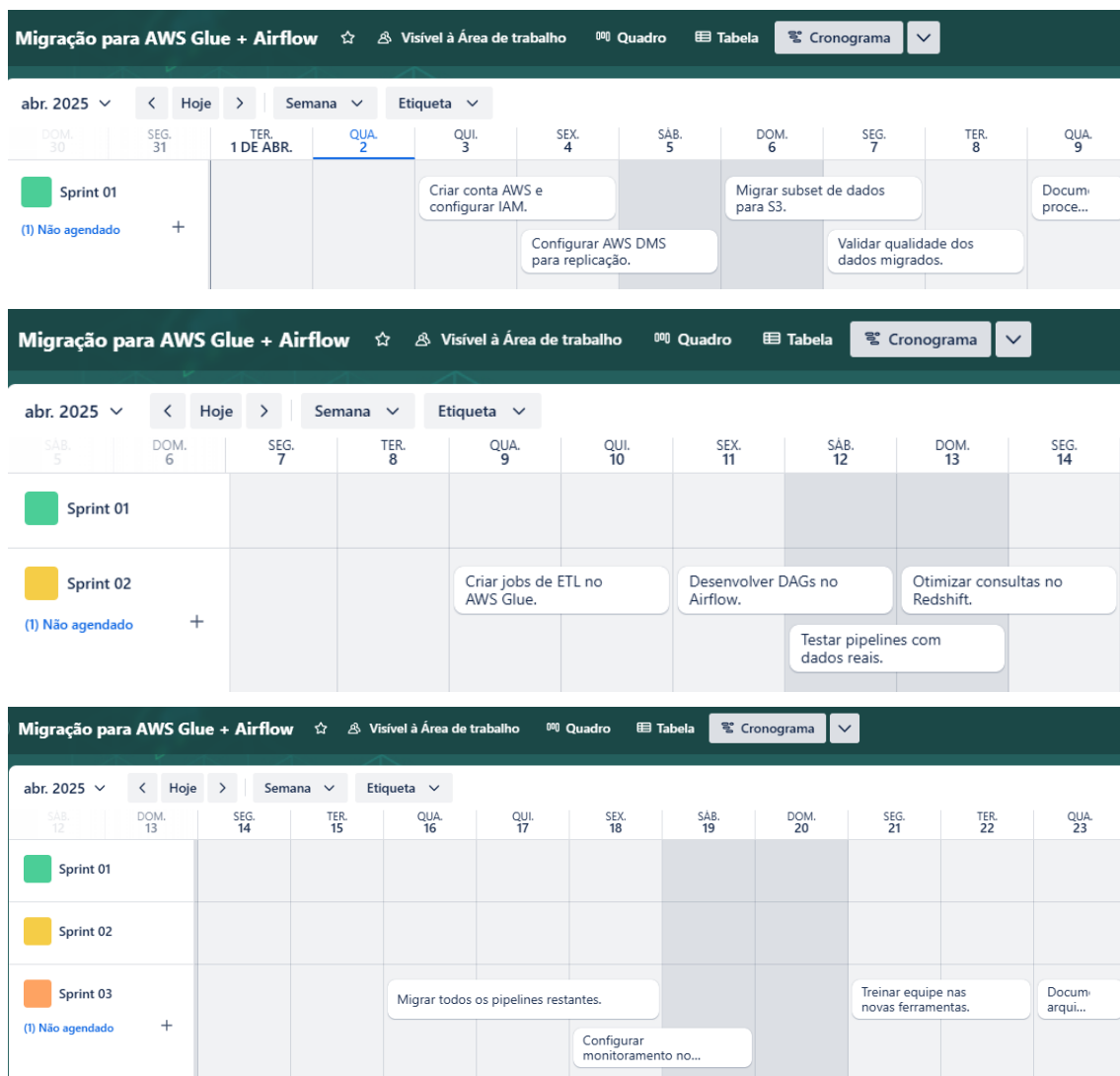
- Estruturar buckets por camada (raw, processed, curated).
- Implementar políticas de retenção e ciclo de vida.

## 6. Redshift como DW

- Modelar tabelas otimizadas para consultas.
- Configurar conexão com ferramentas de BI.

## 7. Monitoramento

- Configurar CloudWatch para métricas de desempenho.
- Alertas para custos e falhas.



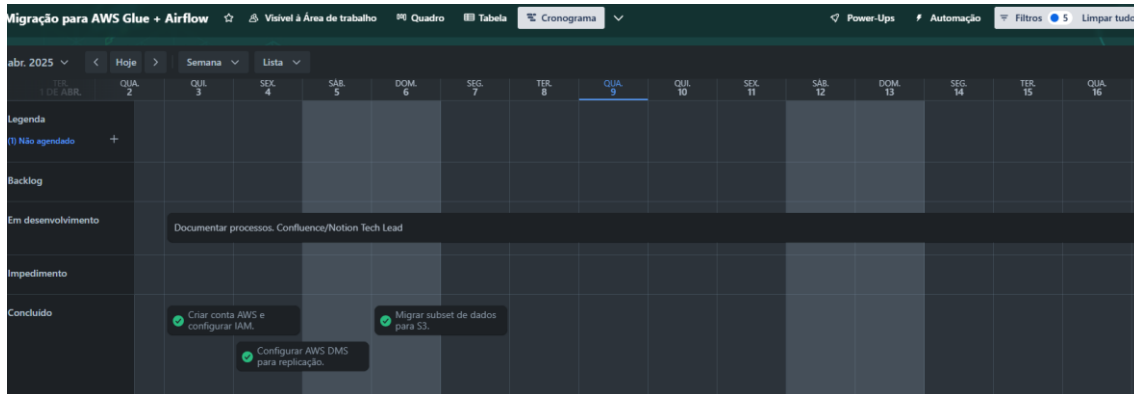


## 2. Área de Experimentação

### 2.1 Sprint 1

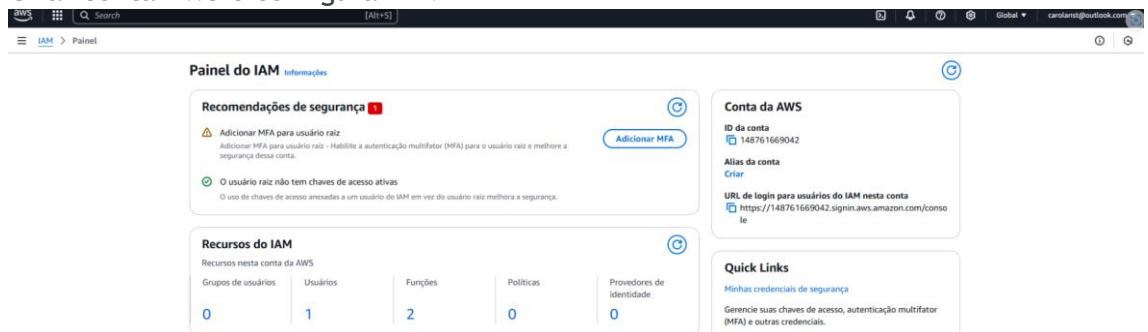
#### 2.1.1 Solução

Planejamento:



#### • Evidência da execução de cada requisito:

##### 1. Criar conta AWS e configurar IAM



##### 2. Configurar AWS DMS para replicação.

- MySQL instalado em uma VM (EC2)

```

ec2-user@ip-172-31-84-77 ~$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor. Commands end with ; or \g.
Your MySQL connection id is 14
Server version: 8.0.41 MySQL Community Server - GPL

Copyright (c) 2000, 2025, Oracle and/or its affiliates.
Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

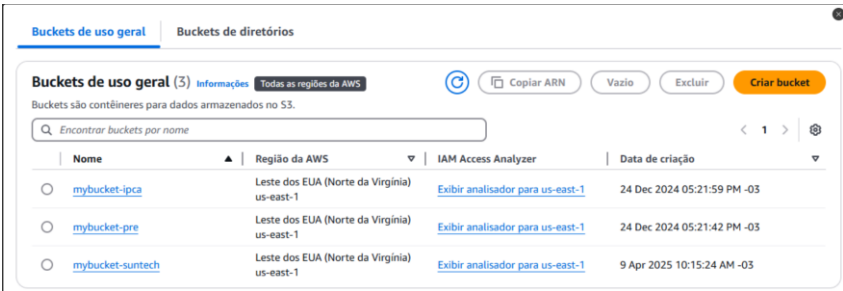
mysql> CREATE USER 'root'@'34.199.37.219' IDENTIFIED BY 'NovaSenha123!';
Query OK, 0 rows affected (0.01 sec)

mysql> GRANT ALL PRIVILEGES ON *.* TO 'root'@'34.199.37.219' WITH GRANT OPTION;
Query OK, 0 rows affected (0.01 sec)

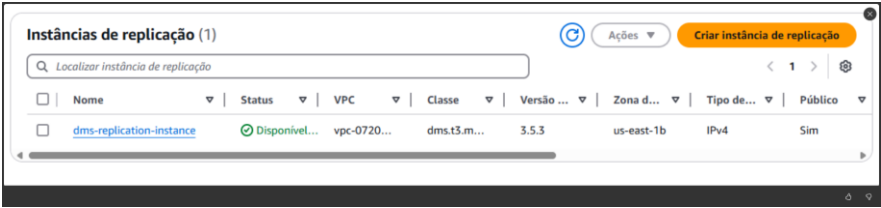
mysql> FLUSH PRIVILEGES;
Query OK, 0 rows affected (0.00 sec)

mysql> exit
Bye
ec2-user@ip-172-31-84-77 ~$
    
```

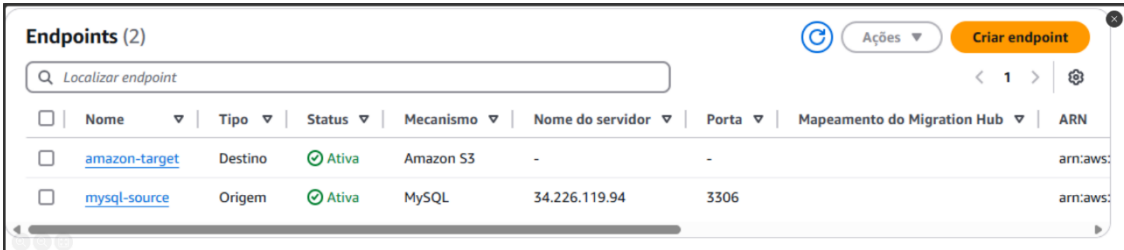
- Bucket S3 na AWS para receber os dados.



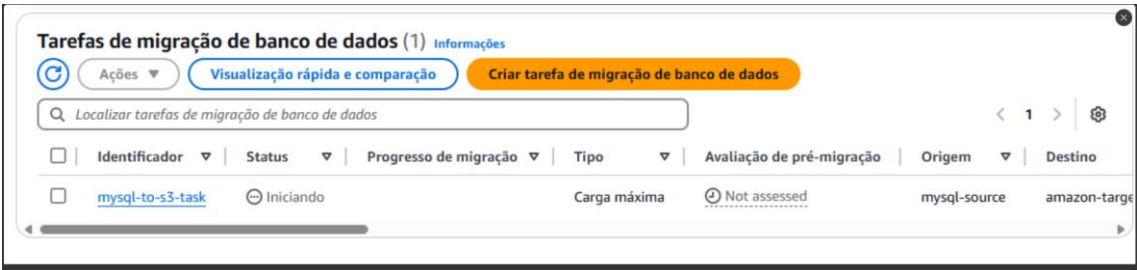
- Criar uma "Replication Instance" no DMS



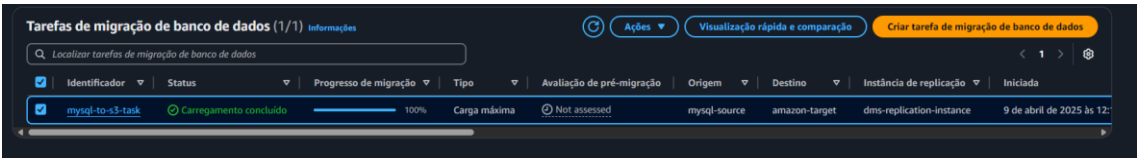
3. Configurar os Endpoints

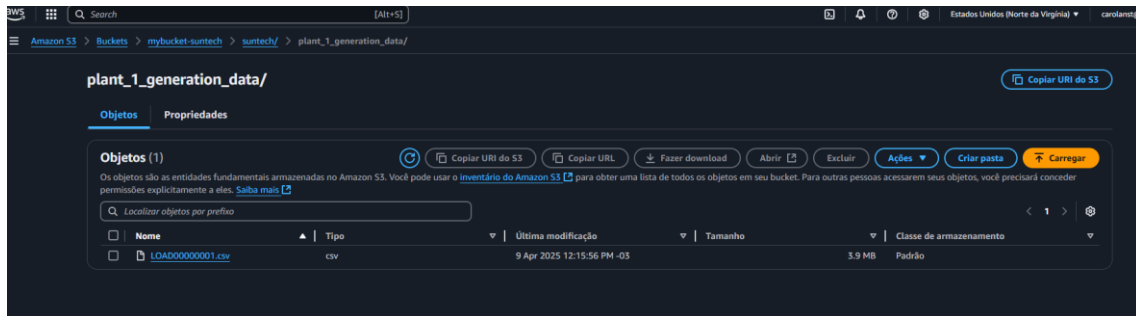


4. Criar a "Replication Task"



- Evidência dos resultados:





### 2.1.2 Retrospectiva da Sprint

Durante o desenvolvimento, nos deparamos com erros de firewall ao tentar acessar o banco de dados localmente. A situação exigia configurações específicas que eu não conseguiria resolver sozinha. Diante disso, tínhamos duas alternativas para dar continuidade ao projeto:

1. Exportar os dados manualmente em formato CSV e carregá-los diretamente no bucket S3;
2. Instalar uma instância virtual do MySQL na EC2 e realizar a migração por meio do AWS DMS, conforme previsto inicialmente.

Optamos pela segunda opção, como demonstrado nas evidências abaixo. Instalamos o MySQL na EC2 e realizamos um *dump* dos dados da máquina local diretamente em um diretório da instância, possibilitando a migração para o bucket. Nesse primeiro momento, não foi possível fazer a validação dos dados. Começaremos a partir da segunda sprint.

Sabíamos dos desafios de segurança envolvidos nesse processo, e acredito que, em um ambiente corporativo com uma equipe de segurança atuante, esse tipo de situação seria rapidamente resolvido — já que as liberações de firewall estariam contempladas na arquitetura desde o início.

```
PS C:\Users\carol> & "C:\Program Files\MySQL\MySQL Server 8.0\bin\mysqldump.exe" -u root -p suntech | Out-File -Encoding ASCII -FilePath suntech.sql
Enter password: ****
PS C:\Users\carol> scp -i "C:\Users\carol\Documents\Cursos\XPE\Engenharia_Dados\PA\admin.pem" suntech.sql ec2-user@34.226.119.94:/home/ec2-user/
suntech.sql
PS C:\Users\carol> |
100% 4413KB 413.4KB/s 00:10
```

```
mysql> USE suntech;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> SHOW TABLES;
+-----+
| Tables_in_suntech |
+-----+
| plant_1_generation_data |
+-----+
1 row in set (0.00 sec)
```

Exportar arquivo local:

```
& "C:\Program Files\MySQL\MySQL Server 8.0\bin\mysqldump.exe" -u root -p suntech
| Out-File -Encoding ASCII -FilePath suntech.sql
```

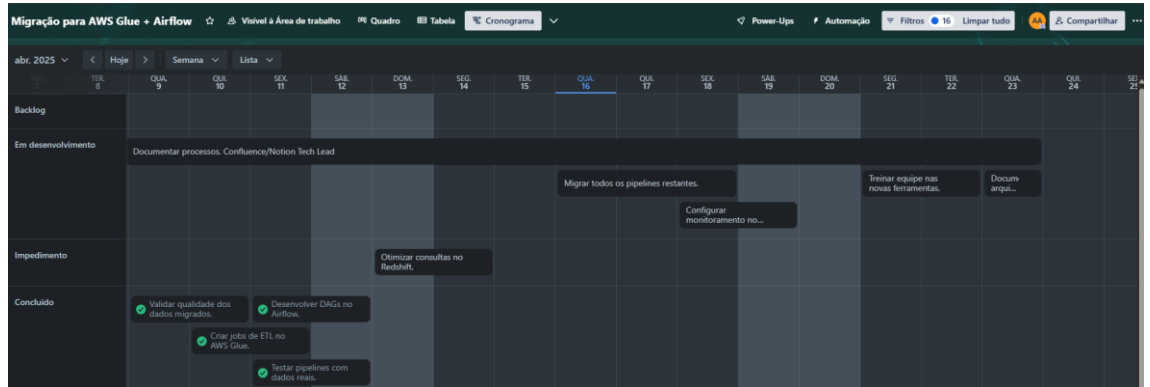
Migrar para Ec2:

```
scp -i "C:\Users\carol\Documents\Cursos\XPE\Engenharia_Dados\PA\admin.pem"
suntech.sql ec2-user@34.226.119.94:/home/ec2-user/
```

## 2.2 Sprint 2

### 2.2.1 Solução

Planejamento:



- Evidência da execução de cada requisito:

1. Criar um Crawler para catalogar os dados no S3.

**Crawlers**  
A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

**Crawlers (1)** info  
View and manage all available crawlers.

Filter crawlers

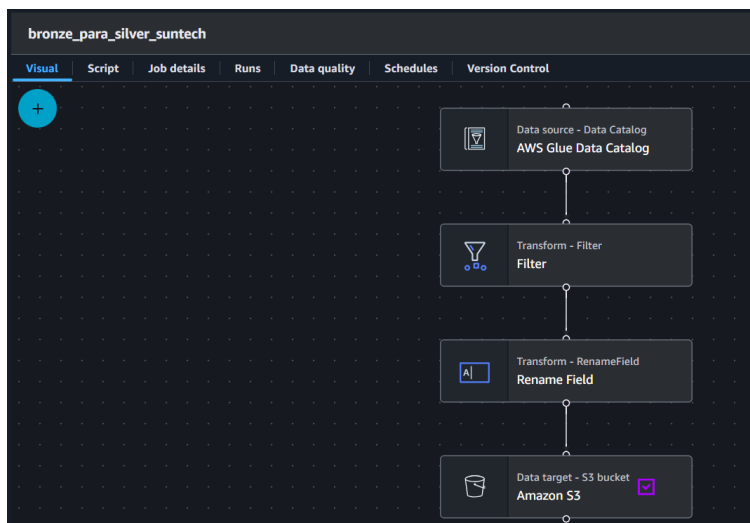
Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from last run
crawler-suntech	Ready		Succeeded	April 15, 2025 at 23:54:25	View log	1 updated

2. Criar o job ETL no Glue Studio ou via script (preferência: Python).

**Your jobs (2)** info  
Filter jobs by property

Job name	Type	Created by	Last modified	AWS Glue version
silver_para_gold_suntech	Glue ETL	Visual	16/04/2025, 21:07:46	5.0
bronze_para_silver_suntech	Glue ETL	Visual	15/04/2025, 20:26:45	5.0

3. Definir transformações iniciais (limpeza, normalização, etc.).



#### 4. Validar se os dados foram corretamente inseridos no Glue Data Catalog.

**Tables**  
A 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.

View and manage all available tables.

Filter tables

Name	Database	Location	Classification	Deprecated	View data	Data quality	Column statistics
<input type="checkbox"/> suntech_silver	suntech_db	s3://mybucket-suntech-sil	Parquet	-	<a href="#">Table data</a>	<a href="#">View data quality</a>	<a href="#">View statistics</a>
<input type="checkbox"/> suntech_suntech	suntech_db	s3://mybucket-suntech/sur	CSV	-	<a href="#">Table data</a>	<a href="#">View data quality</a>	<a href="#">View statistics</a>

#### 5. Usar Step Functions / Criar uma DAG simples que orquestre um job Glue.

**pipeline\_glue\_orquestrado** [Editar](#) [Ações](#) [Iniciar execução](#)

**Detalhes**

Arn: [arn:aws:statesus-east-1:148761669042:stateMachine:pipeline\\_glue\\_orquestrado](#)

ARN do perfil do IAM: [arn:aws:iam::148761669042:role/service-role/StepFunctions-pipeline\\_glue\\_orquestrado-role-lvbyhstyi](#)

Tipo: Padrão

Status: Ativo

Data de criação: 16 de abr. de 2025, 21:10:45 (UTC-03:00)

Rastreamento com X-Ray: Desativado

**Execuções** | Monitoramento | Registro em log | Definição | Aliases | Versões | Etiquetas

**Execuções (0/2)** [Visualizar detalhes](#) [Interromper execução](#) [Redirecionamento](#) [Iniciar execução](#)

Filterar execuções por propriedade ou valor

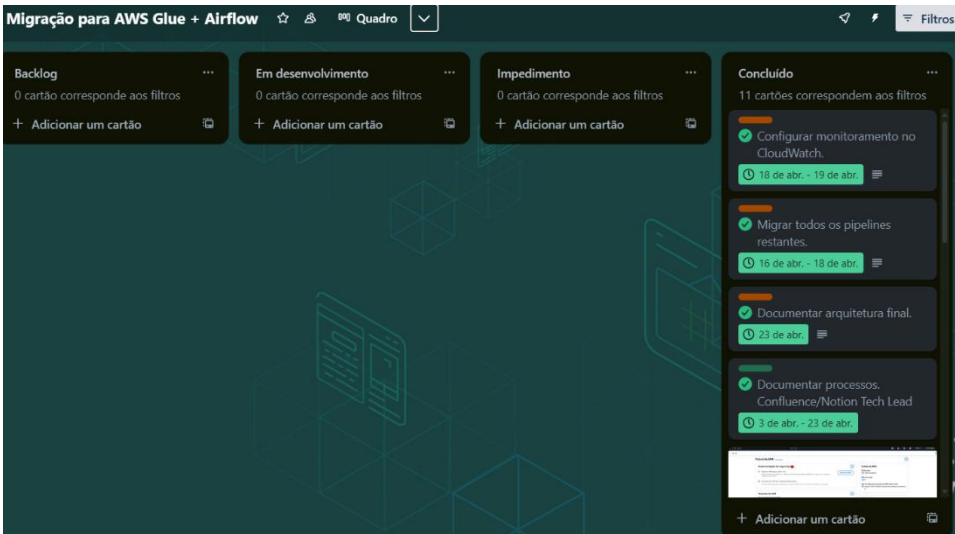
Nome	Status	Hora de início (local)	Hora de término (l...	Duração...	Versão	Alias
<a href="#">62f5dd80-034b-4b79-b350-66808b3c48a2</a>	Com êxito	16 de abr. de 2025, 21:28:24	16 de abr. de 2025, 21:33:17	00:04:53.332	-	-



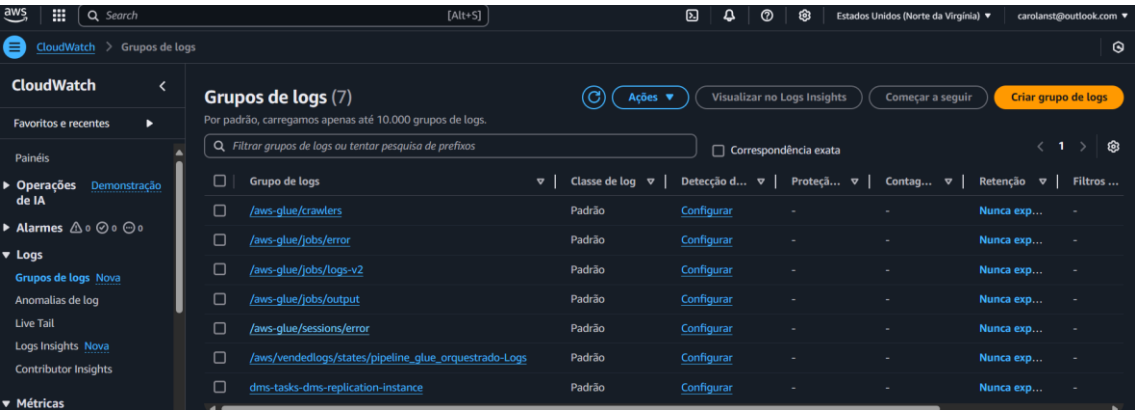
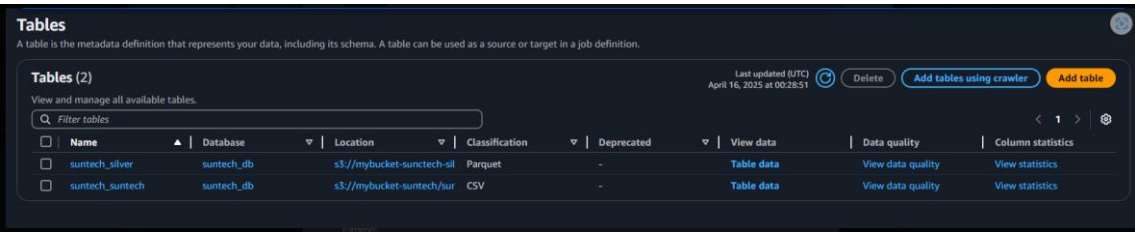


2.3 Sprint 3

2.3.1 Solução



- Evidência da execução de cada requisito:



## Evidência dos resultados:

mybucket-suntech-silver

Objetos (15)

Os objetos são as entidades fundamentais armazenadas no Amazon S3. Você pode usar o [inventário do Amazon S3](#) para obter uma lista de todos os objetos em seu bucket. Para outras pessoas acessarem seus objetos, você precisará conceder permissões explicitamente a eles. Saiba mais

Localizar objetos por prefixo

Nome	Tipo	Última modificação	Tamanho	Classe de armazenamento
run-1744762721521-part-block-0-r-00000-snappy.parquet	parquet	15 Apr 2025 09:18:57 PM -03	6.6 MB	Padrão
run-1744762721521-part-block-0-r-00001-snappy.parquet	parquet	15 Apr 2025 09:19:09 PM -03	6.1 MB	Padrão
run-1744762721521-part-block-0-r-00002-snappy.parquet	parquet	15 Apr 2025 09:19:10 PM -03	6.0 MB	Padrão
run-1744762721521-part-block-0-r-00003-snappy.parquet	parquet	15 Apr 2025 09:19:11 PM -03	5.7 MB	Padrão

```

[notice] A new release of pip is available: 24.3.1 -> 25.0.1
[notice] To update, run: python.exe -m pip install --upgrade pip

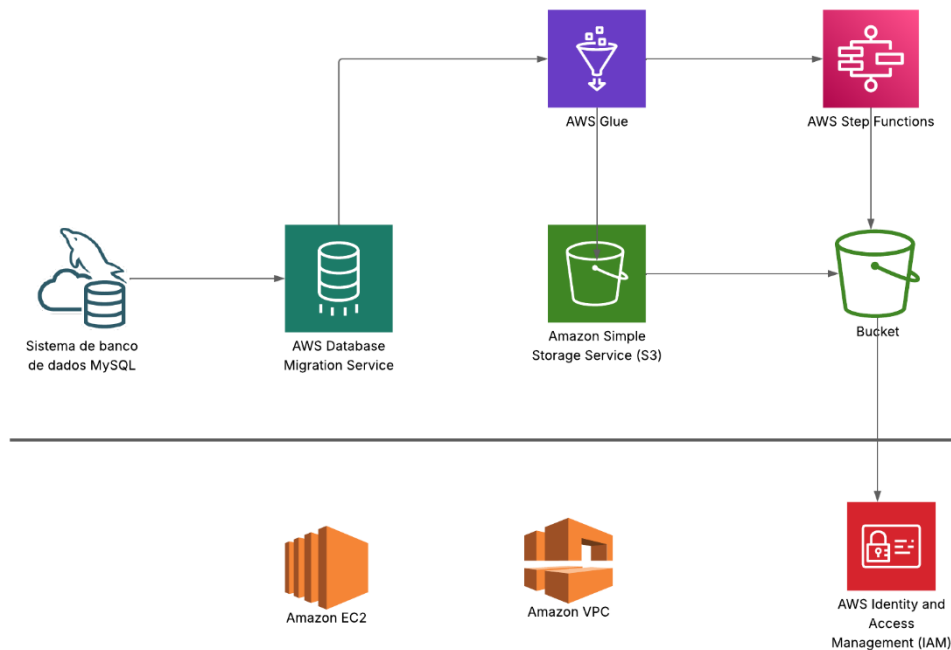
# Caminho do arquivo no S3
s3_path = "s3://mybucket-suntech-silver/run-1744762721521-part-block-0-r-00000-snappy.parquet"

# Ler arquivo parquet
df = pd.read_parquet(s3_path, engine="pyarrow")

# Exibir amostra
df.head()

```

numordeminterrupcao	dsctatogeneradorinterrupcao	sigagente	numunidadeconsumidora	dsctatogeneradorinterrupcao	dsctipointerrupcao	dsctatointerrupcao	dsctatointerrupcao	numano	numinvtensao	dtb...
0	INTERNA - NAO PROGRAMADA - PROPRIAS DO SISTEMA	EAC	1	01C4	Não Programada	01/01/2025 00:53	AAL	2025	9999	
1	INTERNA - NAO PROGRAMADA - MEIO AMBIENTE - VENTO	EAC	0	01C2	Não Programada	22/02/2025 21:50	SMA	2025	13800	



### 2.3.2 Retrospectiva da Sprint

Nessa última etapa, não tivemos imprevistos nem novos desenvolvimentos, além da configuração do CloudWatch. Apenas foi executado o pipeline com mais volumes de dados e validados, e o desenho da arquitetura.



## 3. Considerações Finais

### 3.1 Resultados

O Projeto Aplicado teve como objetivo principal a migração de dados de um banco MySQL on-premise para a nuvem AWS, utilizando ferramentas gratuitas com foco na construção de um pipeline de dados eficiente. Entre os principais resultados, destaca-se a migração bem-sucedida via AWS DMS, a criação de um data lake segmentado por camadas (bronze, silver e gold) no Amazon S3 e a implementação de processos ETL com o AWS Glue.

Entre os aspectos positivos, evidenciam-se a viabilidade técnica da solução, a integração eficiente entre os serviços AWS e o aprendizado prático proporcionado. Como ponto negativo, observou-se a complexidade inicial na configuração de permissões (IAM) e na escrita dos scripts de transformação no AWS Glue.

Dentre as principais dificuldades enfrentadas, destacam-se a integração segura com o banco on-premise, a curva de aprendizado das ferramentas AWS e a falta de domínio das ferramentas que acarretou cobranças inesperadas. Ainda assim, o projeto proporcionou uma experiência significativa, consolidando conhecimentos em arquitetura de dados, engenharia de dados em nuvem e boas práticas de ETL, alinhados com as demandas do mercado atual.

### 3.2 Próximos passos

Para aprimorar a solução proposta, recomenda-se a adoção de ferramentas de orquestração, como AWS Step Functions, e a implementação de monitoramento com Amazon CloudWatch. A validação da qualidade dos dados também é essencial para garantir confiabilidade ao pipeline.

Adicionalmente, a integração com ferramentas de BI, como o Amazon QuickSight, permitirá a criação de dashboards interativos, ampliando o valor analítico da solução. Por fim, a utilização de bases analíticas como o Amazon Redshift pode melhorar a performance em cenários de grandes volumes de dados.

