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.

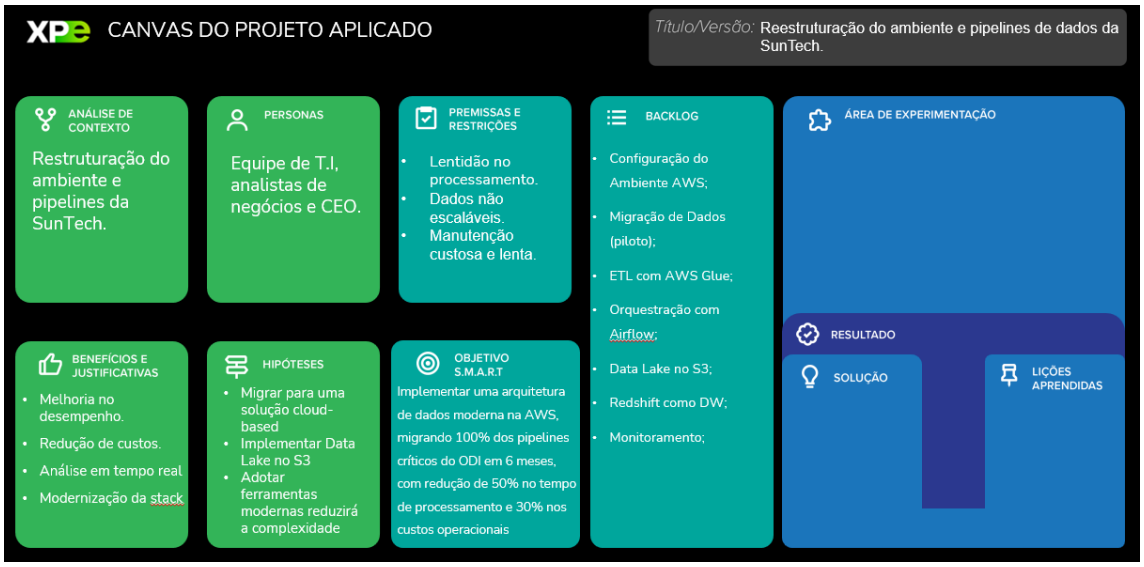
Orientador (a): Moisés Luna

Cataguases, Minas Gerais

18/03/2025

Sumário

1. CANVAS do Projeto Aplicado



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:

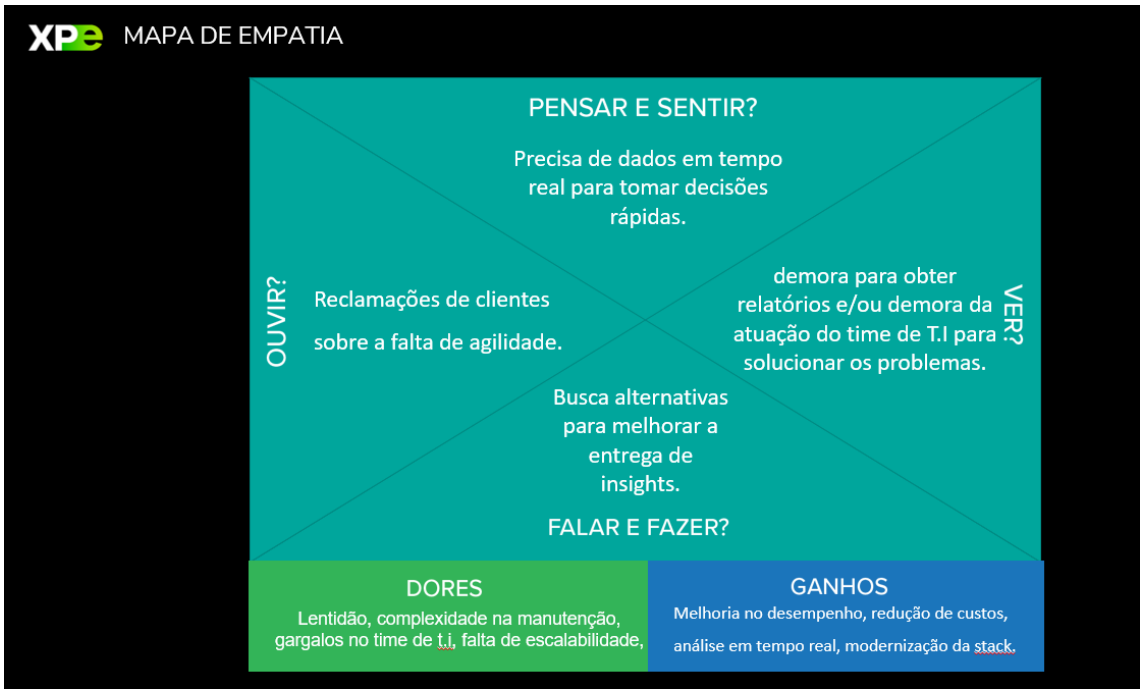
- **Desempenho insuficiente: O ODI não consegue processar grandes volumes de dados de forma eficiente, resultando em pipelines lentos.**
- **Falta de escalabilidade: A infraestrutura atual não suporta o crescimento exponencial dos dados, limitando a capacidade de análise.**
- **Dificuldade de manutenção: 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	**Médio**
Adotar Apache Airflow reduzirá a complexidade.	Baixo	Baixo	Baixo	**Baixo**
Implementar um Data Lake no S3 permitirá análises em tempo real.	Alto	Médio	Alto	**Alto**
Migrar para a nuvem reduzirá os custos em 30%.	Médio	Baixo	Médio	**Médio**

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	**Alto**
Quanto tempo levará para migrar todos os dados?	Médio	Médio	Alto	**Alto**
A equipe terá capacidade para aprender as novas ferramentas?	Médio	Baixo	Alto	**Médio**
Como garantir a segurança dos dados na nuvem?	Alto	Médio	Alto	**Alto**
Qual será o impacto nos clientes durante a migração?	Médio	Médio	Alto	**Alto**

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	**Alta**
Adotar Apache Airflow.	Médio	Baixo	**Média**
Implementar um Data Lake no S3.	Alto	Alto	**Baixa**
Migrar para a nuvem para reduzir custos.	Alto	Médio	**Alta**

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.	Paralização 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
Orçamento limitado para contratação de serviços AWS.	Pode exigir otimização de recursos ou priorização de pipelines críticos.
Prazo máximo de 6 meses para conclusão.	Exige cronograma rigoroso e possível redução de escopo se houver atrasos.
Conformidade com LGPD e regulamentações.	Necessidade de revisão extra de segurança e governança de dados na AWS.
Equipe enxuta (sem recursos para contratações).	Sobrecarga da equipe interna e risco de burnout.
Dependência de sistemas legados 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).

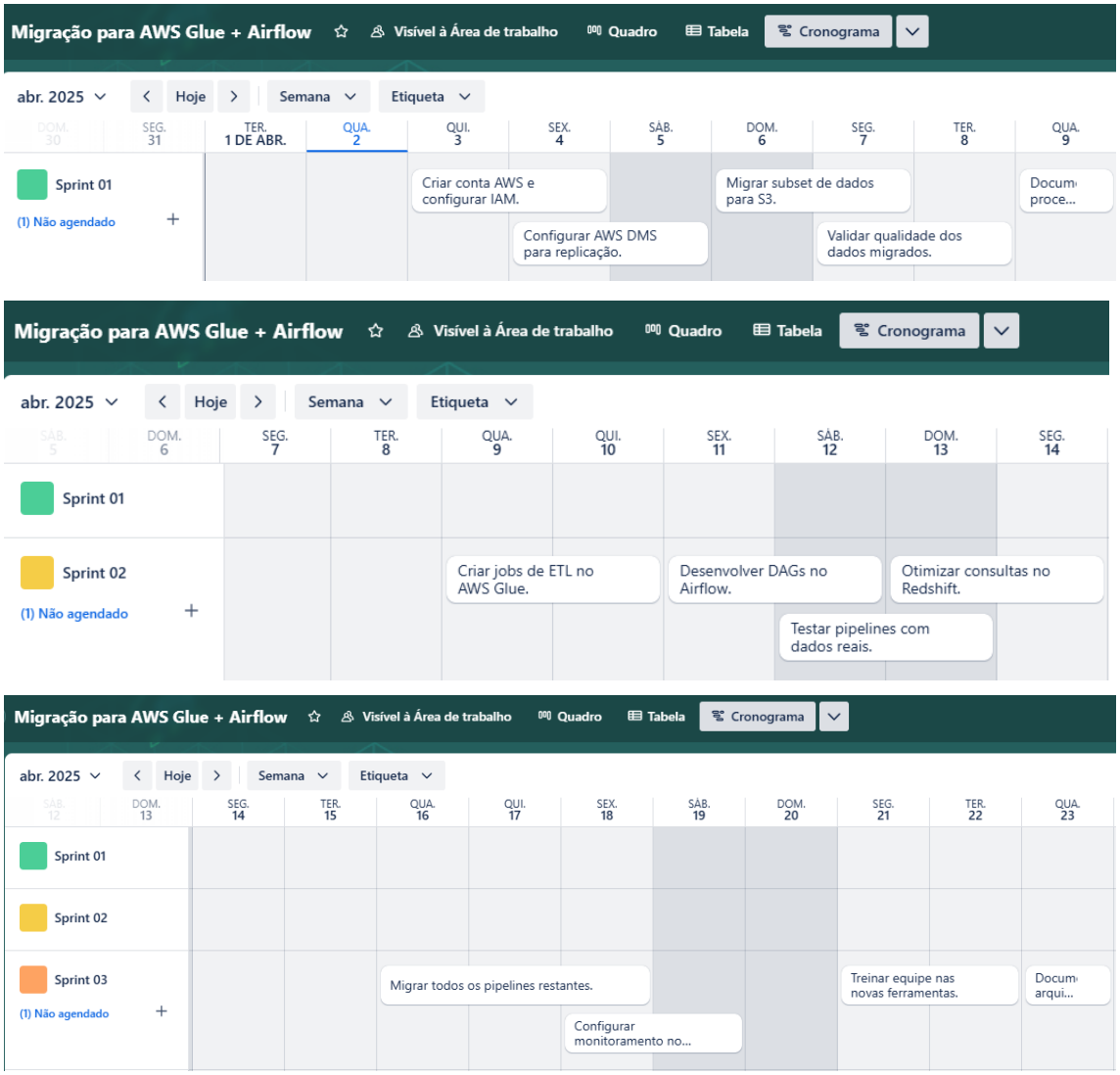
Backlog de Requisitos

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.

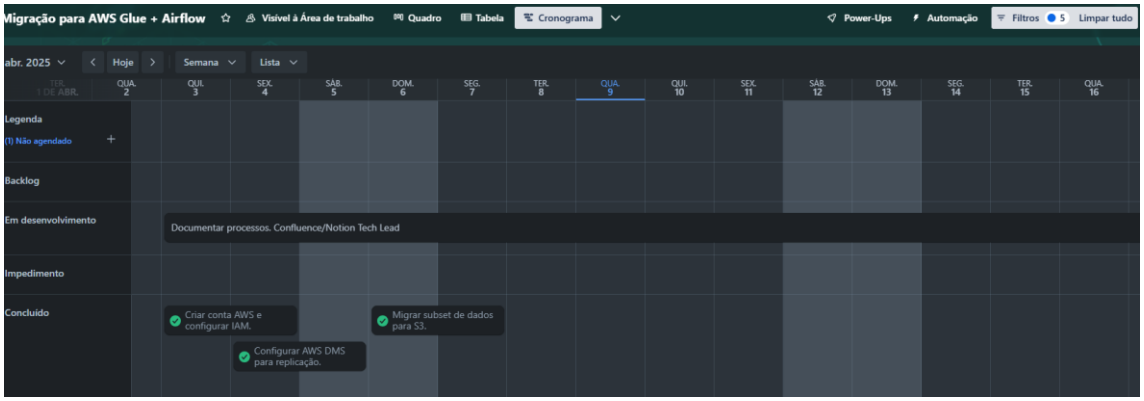


2. Área de Experimentação

2.1 Sprint 1

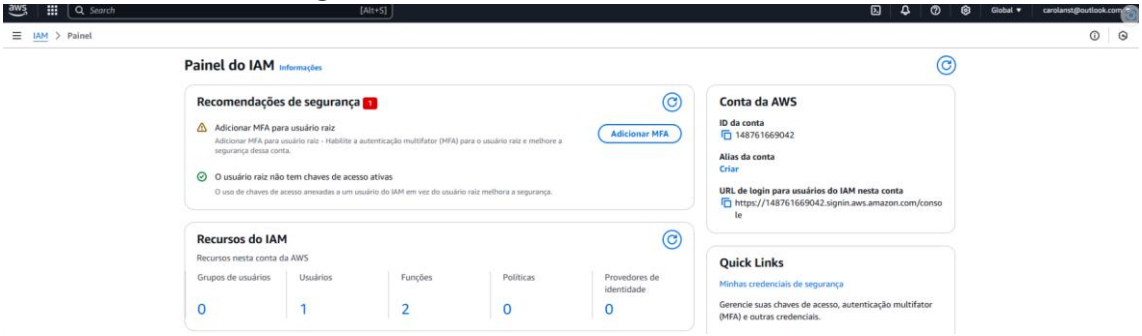
2.1.1 Solução

Planejamento:

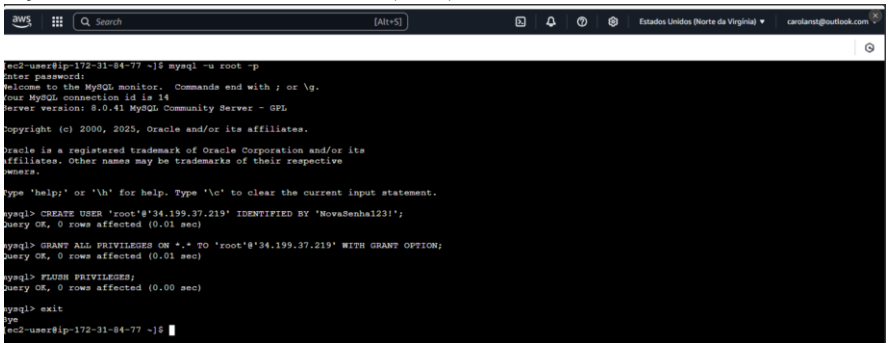


- Evidência da execução de cada requisito:

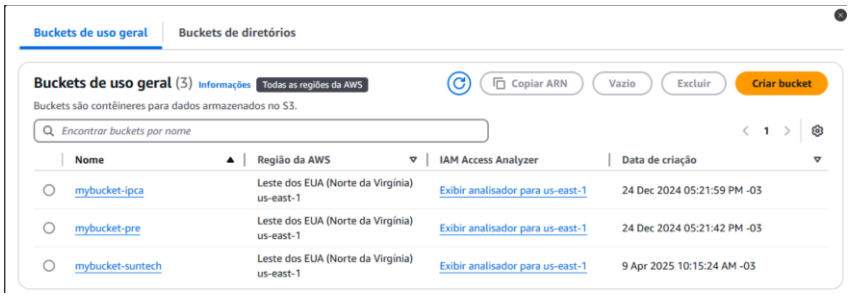
1. Criar conta AWS e configurar IAM



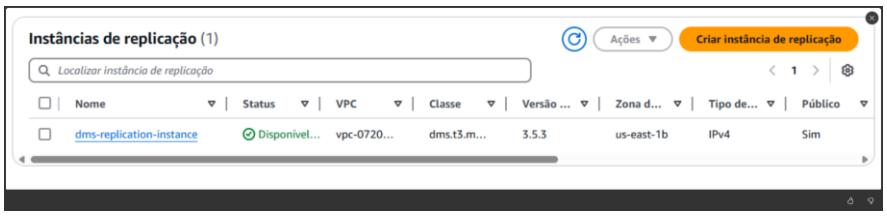
2. Configurar AWS DMS para replicação.
   - MySQL instalado em uma VM (EC2)



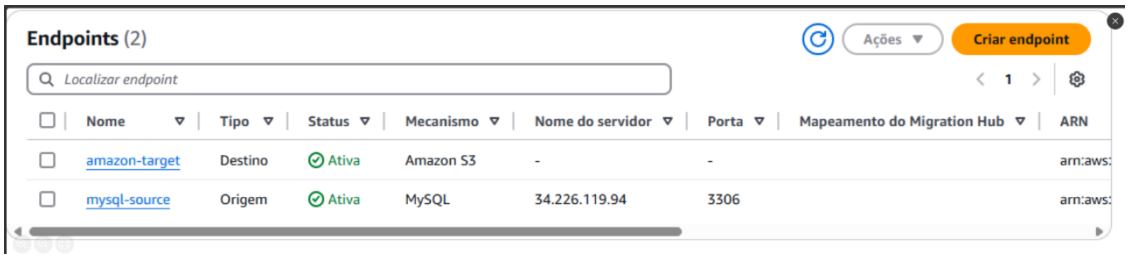
   - **Bucket S3** na AWS para receber os dados.



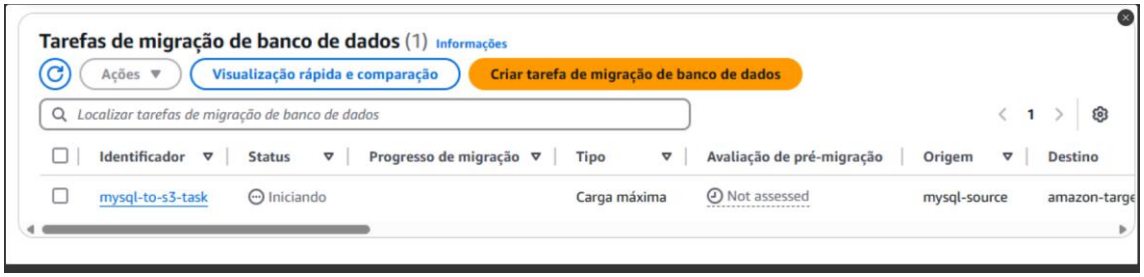
- Criar uma "Replication Instance" no DMS



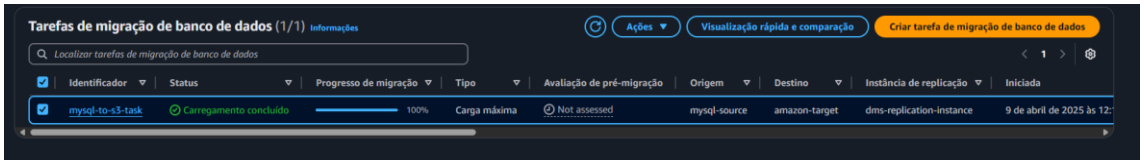
3. Configurar os Endpoints



4. Criar a "Replication Task"



- Evidência dos resultados:



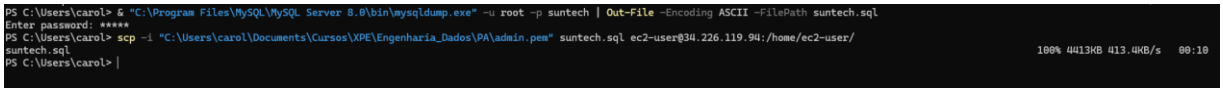
2.1.2 Retrospectiva da Sprint

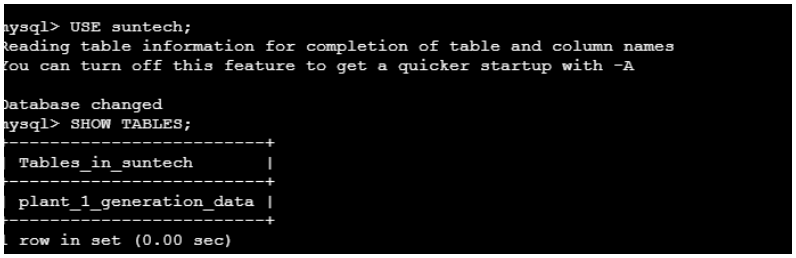
Durante o desenvolvimento, nos deparamos com erros de firewall ao tentar acessar o banco de dados localmente. A situação exigia configurações específicas que eu não conseguiria resolver sozinha. Diante disso, tínhamos duas alternativas para dar continuidade ao projeto:

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.





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

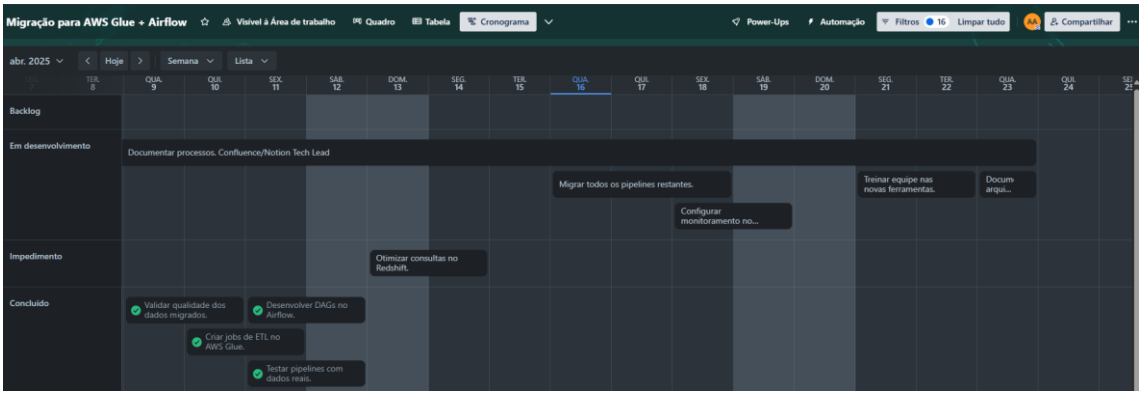
Migrar para Ec2:

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

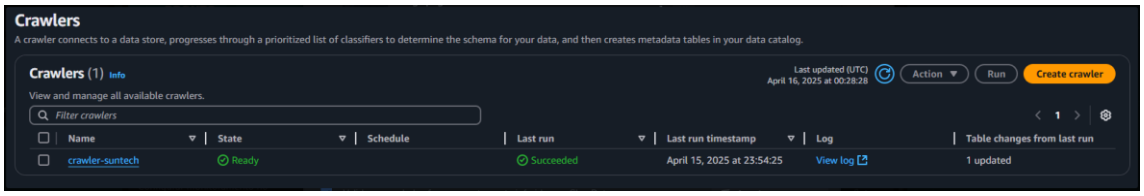
2.2.1 Solução

Planejamento:

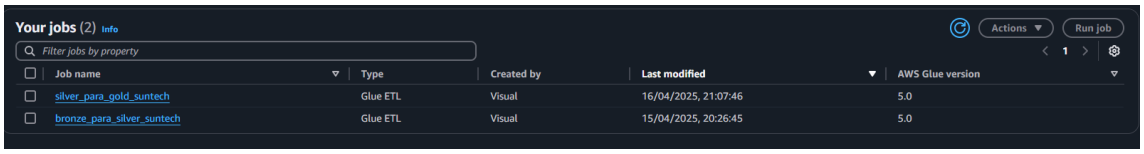


- Evidência da execução de cada requisito:

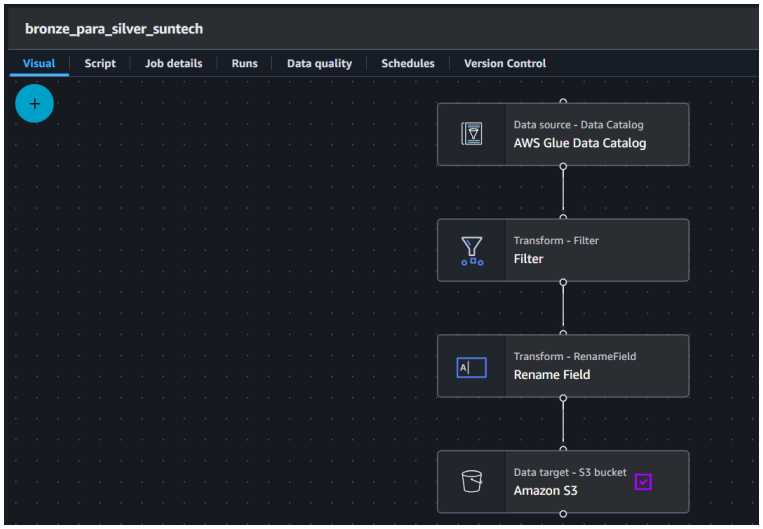
1. Criar um Crawler para catalogar os dados no S3.



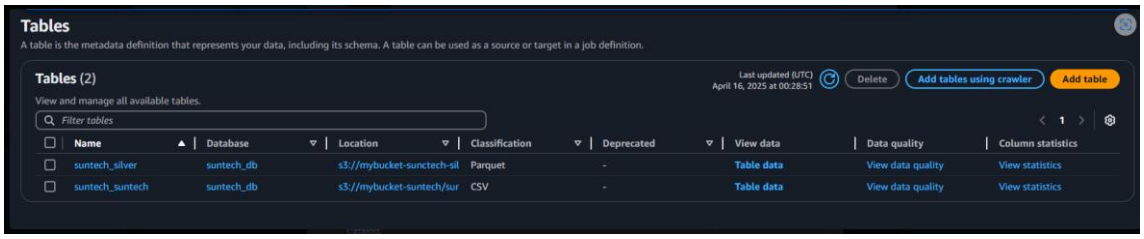
2. Criar o job ETL no Glue Studio ou via script (preferência: Python).



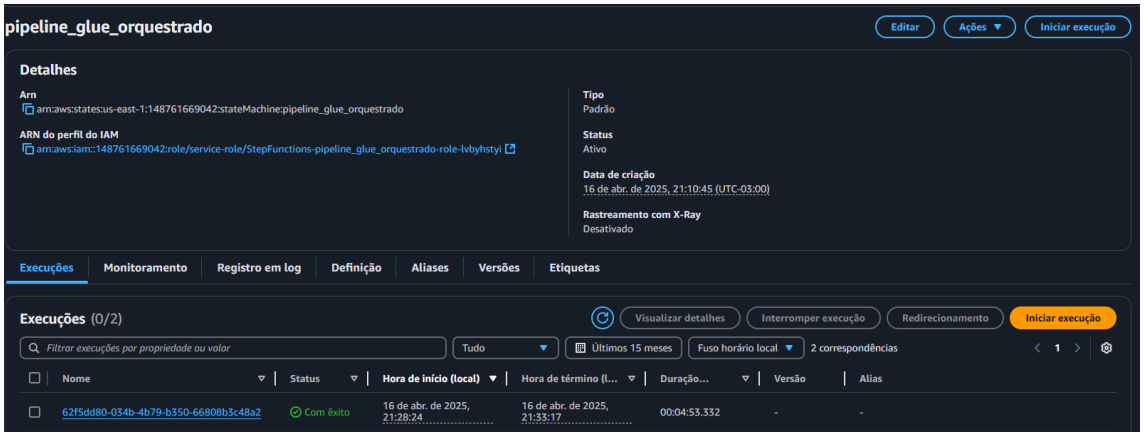
3. Definir transformações iniciais (limpeza, normalização, etc.).



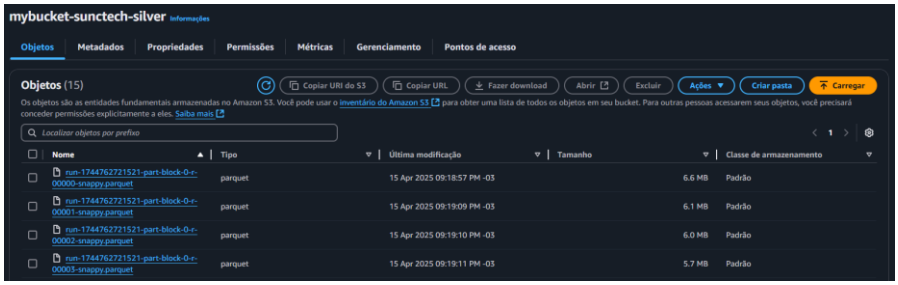
4. Validar se os dados foram corretamente inseridos no Glue Data Catalog.

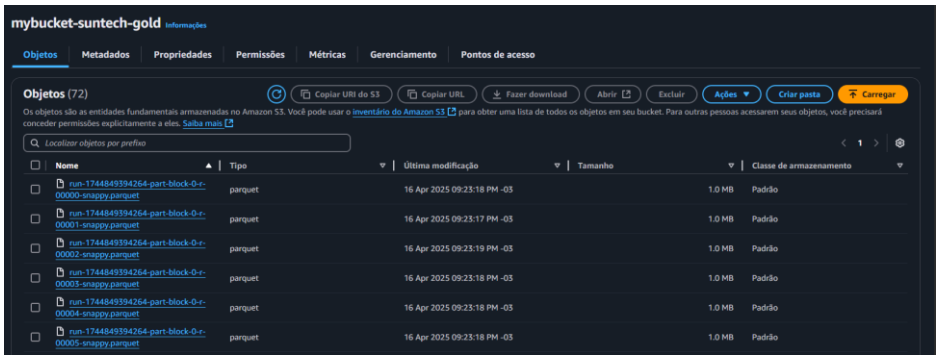


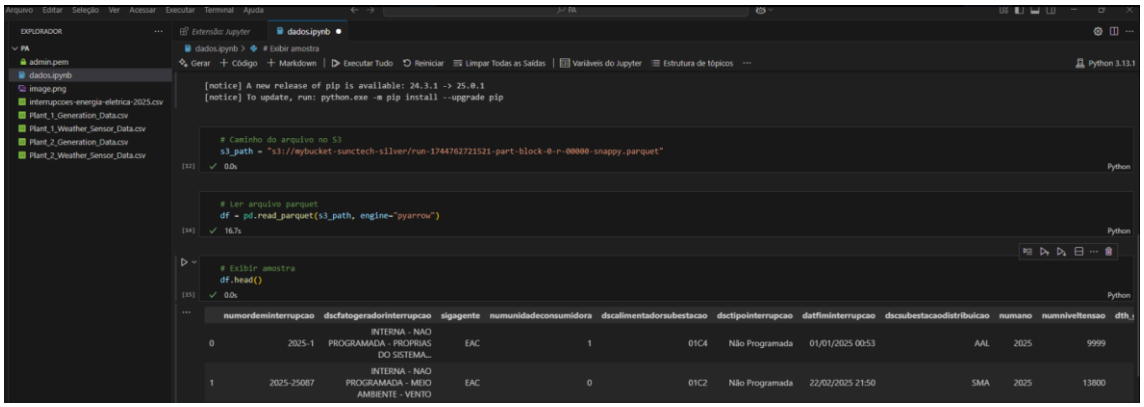
5. Usar Step Functions / Criar uma DAG simples que orquestre um job Glue.



- Evidência dos resultados:







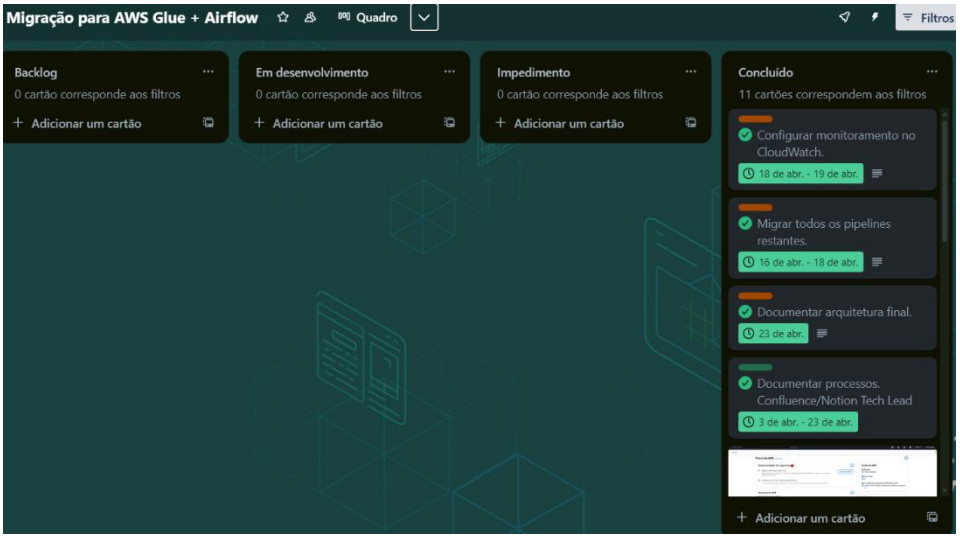
2.2.2 Retrospectiva da Sprint

Durante o desenvolvimento, nos deparamos com uma cobrança inesperada de um dos recursos utilizados. Em um ambiente corporativo isso seria previsto, mas é importante salientar a importância do domínio das ferramentas para não gerar custos extras. A nível de estudos, foi necessário deletar os componentes uma vez que já foi registrado as evidências neste documento.

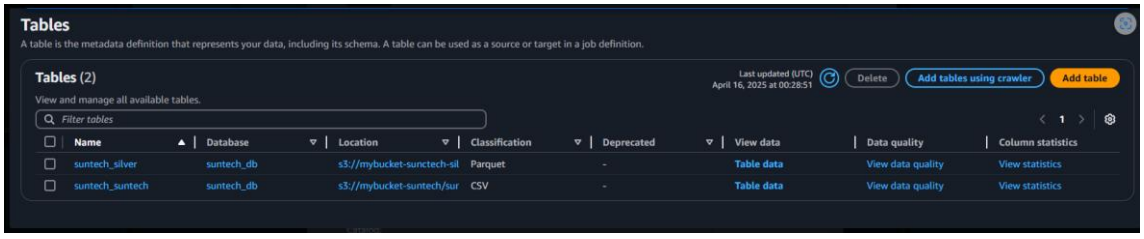
Seguindo com a segurança de utilizar ferramentas gratuitas para este projeto, substituímos o Apache Airflow pelo Step Functions da AWS, para criação das DAGs e ao invés de usarmos Athena para visualização dos dados, utilizamos o Jupyter .

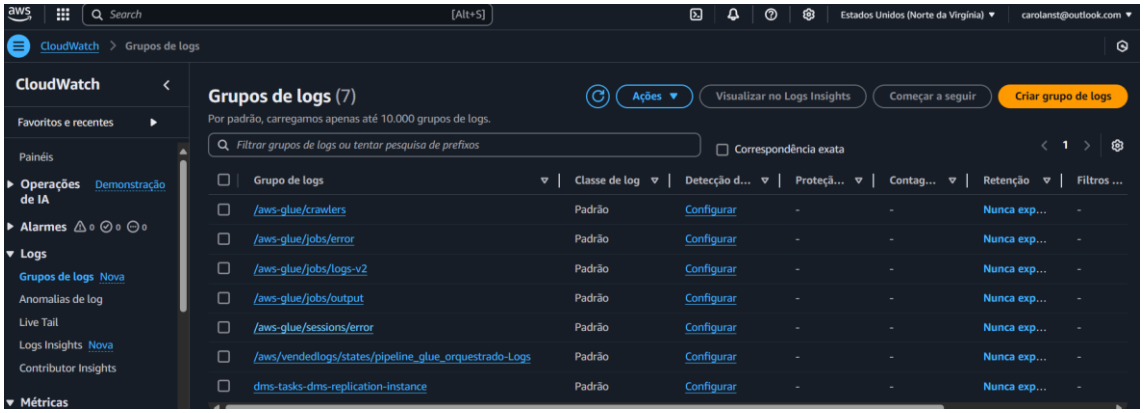
2.3 Sprint 3

2.3.1 Solução

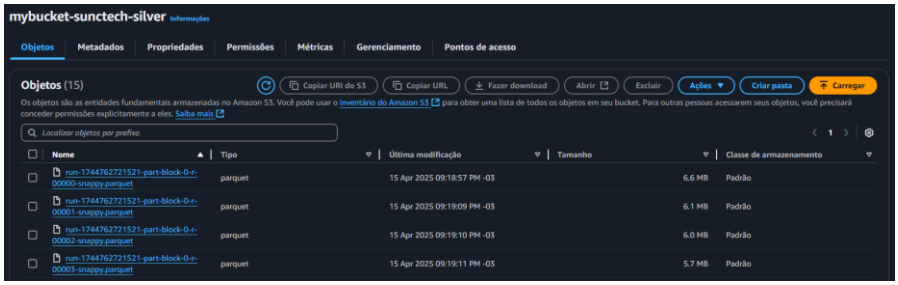


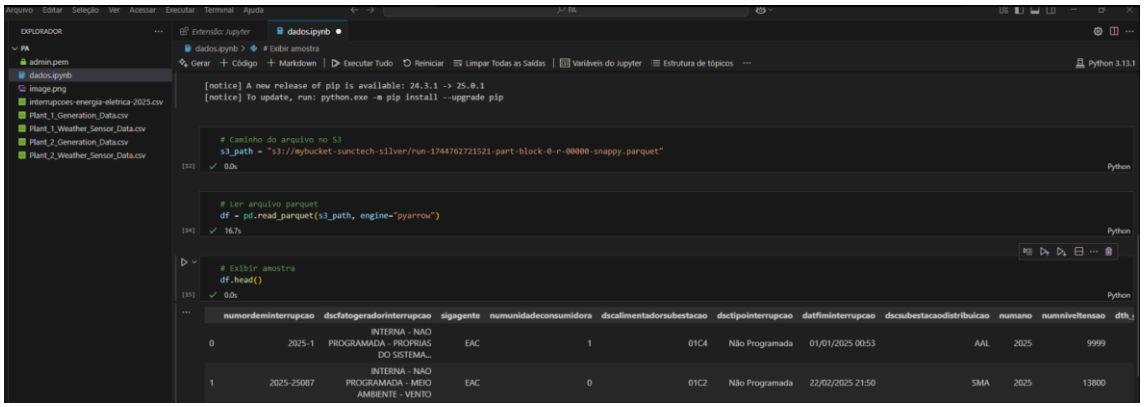
- Evidência da execução de cada requisito:

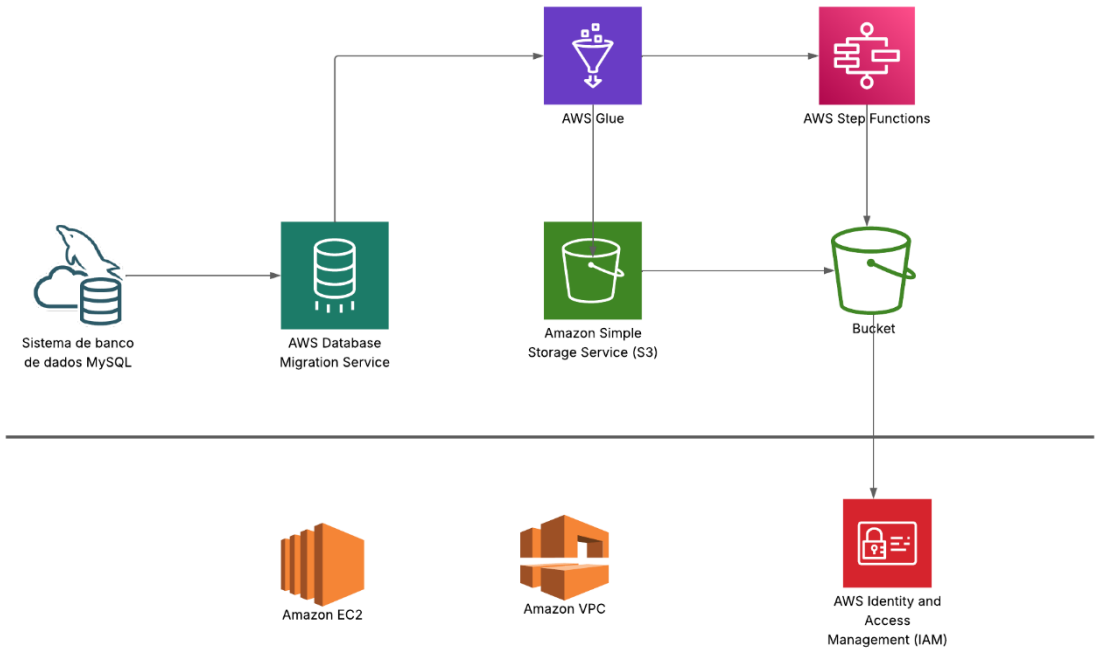




- Evidência dos resultados:







2.3.2 Retrospectiva da Sprint

Nessa última etapa, não tivemos imprevistos nem novos desenvolvimentos, alemda 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.