SÃO
PAULO
TECH
SCHOOL

Arquitetura de soluções em nuvem

ETL e Data Handling

Eduardo Verri

eduardo.verri@sptech.school

ETL

O que é ETL?

- ETL (que significa extrair, transformar, carregar) é um processo de integração de dados usado para combinar dados de várias fontes em um conjunto de dados único e consistente para carregar em um **data warehouse**, **data lake** ou outro sistema de destino.
- Conforme os bancos de dados se tornavam mais populares na década de 1970, o ETL foi introduzido como um processo para integração e carregamento de dados para computação e análise, tornando-se eventualmente o método principal para processar dados em projetos de data warehousing.
- Fornece a base para análise de dados e fluxos de trabalho de **aprendizado de máquina**. Por meio de uma série de business rules. o ETL limpa e organiza dados de forma a atender necessidades específicas do negócio, como relatórios mensais, mas também pode lidar com análises mais avançadas, que podem melhorar processos de back-end ou experiências de usuário final.

Como funciona o ETL - extract

Extrair: Durante a extração de dados, os dados brutos são copiados ou exportados das localizações de origem para uma área de preparação. Os dados podem ser estruturados ou não estruturados. Fontes essas que podem ser: servidores SQL ou NoSQL, sistemas de CRM e ERP, arquivos simples, e-mail, págians da web, etc.

Como funciona o ETL - transform

Transformar: Na área de preparação, os dados brutos passam por processamento de dados. Aqui, os dados são transformados e consolidados para o caso de uso analítico pretendido.

- ✓ Filtrando, limpando, eliminando a duplicação, validando e autenticando os dados.
- ✓ Realizar cálculos, traduções ou resumos com base nos dados brutos. Isso pode incluir a alteração de cabeçalhos de linhas e colunas para consistência, a conversão de moedas ou outras unidades de medida, a edição de strings de texto e muito mais.
- ✓ Realização de auditorias para garantir a qualidade e a conformidade dos dados.
- ✓ Remover, criptografar ou proteger dados regidos por reguladores setoriais ou governamentais.
- ✓ Formatar os dados em tabelas ou tabelas unidas para corresponder ao esquema do armazém de dados de destino.

Como funciona o ETL - load

Carregar: Nesta última etapa, os dados transformados são movidos da área de preparação para um armazém de dados de destino.

Normalmente, isso envolve uma carga inicial de todos os dados, seguida por carregamentos periódicos de mudanças incrementais nos dados e, menos frequentemente, atualizações completas para apagar e substituir os dados no armazém.

Para a maioria das organizações que utilizam ETL, o processo é automatizado, bem definido, contínuo e orientado por lotes. Normalmente, o ETL ocorre fora do horário comercial, quando o tráfego nos sistemas de origem e no data warehouse é mais baixo.

Benefícios e desafios do ETL

As soluções ETL melhoram a qualidade ao realizar a limpeza dos dados antes de carregá-los em um repositório diferente.

Uma operação em lote que consome tempo, o ETL é recomendado com mais frequência para criar repositórios de dados de destino menores que requerem atualizações menos frequentes, enquanto outros métodos de integração de dados, incluindo ELT, captura de mudanças de dados (CDC) e virtualização de dados, são usados para integrar volumes cada vez maiores de dados que mudam ou fluxos de dados em tempo real.

ETL vs ELT

- O **ELT** copia ou exporta os dados das localizações de origem, mas em vez de carregá-los para uma área de preparação para transformação, **ele carrega os dados brutos diretamente no armazenamento de dados de destino para serem transformados conforme necessário**.
- O **ELT** é particularmente útil para conjuntos de **dados não estruturados e de alto volume**, pois o carregamento pode ocorrer diretamente da fonte. O ELT pode ser mais ideal para gerenciamento de **big data**, pois não requer muito planejamento inicial para extração e armazenamento de dados.
- O processo **ETL**, por outro lado, **requer mais definição no início**. Pontos de dados específicos precisam ser identificados para extração, juntamente com quaisquer "chaves" em potencial para integração em sistemas de origem díspares.

Outros métodos de integração de dados

- **Captura de dados de mudança (CDC):** Identifica e captura apenas os dados de origem que foram alterados e move esses dados para o sistema de destino. Pode ser usada para reduzir os recursos necessários durante a etapa de "extração" de ETL; ele também pode ser usado de forma independente para mover dados que foram transformados em um data lake ou outro repositório em tempo real.

- **Data Replication:** Copia as alterações nas fontes de dados em tempo real ou em lotes para um banco de dados central. É frequentemente listada como um método de integração de dados. Na verdade, ele é usado com mais frequência para criar backups para recuperação de desastres.

Outros métodos de integração de dados

- **A virtualização de dados:** Usa uma camada de abstração de software para criar uma visão unificada, integrada e totalmente utilizável dos dados, sem copiar, transformar ou carregar fisicamente os dados de origem em um sistema de destino. Como por exemplo a criação de **views** no banco de dados.

- **A integração de dados de fluxo (SDI):** Consome continuamente fluxos de dados em tempo real, transforma-os e carrega-os em um sistema de destino para análise. Permitindo um armazenamento de dados para potencializar a análise, o aprendizado de máquina e os aplicativos em tempo real para melhorar a experiência do cliente, a detecção de fraudes e muito mais.

Data cleaning ou data transformation?

- Limpeza de dados é o processo que remove dados que não pertencem ao seu conjunto de dados. Transformação de dados é o processo de conversão de dados de um formato ou estrutura para outro.
- Processos de transformação também podem ser chamados de data wrangling, ou data munging, transformando e mapeando dados de um formato de dados "bruto" (raw) para outro formato para armazenamento e análise.

DATA CLEANING

O que é data cleaning

- Limpeza de dados é o processo de corrigir ou remover dados incorretos, corrompidos, formatados incorretamente, duplicados ou incompletos dentro de um conjunto de dados. Ao combinar várias fontes de dados, há muitas oportunidades para que os dados sejam duplicados ou rotulados incorretamente.
- Se os dados estiverem incorretos, os resultados e algoritmos não serão confiáveis, mesmo que pareçam corretos.

Como limpar os dados

Etapa 1: Remova observações duplicadas ou irrelevantes

Remova observações indesejadas do seu conjunto de dados, incluindo observações duplicadas ou irrelevantes. Observações duplicadas ocorrerão com mais frequência durante a coleta de dados. Quando você combina conjuntos de dados de vários lugares, extrai dados ou recebe dados de clientes ou vários departamentos, há oportunidades de criar dados duplicados. Observações irrelevantes são quando você percebe observações que não se encaixam no problema específico que você está tentando analisar.

Por exemplo, se você deseja analisar dados sobre clientes da geração Y, mas seu conjunto de dados inclui gerações mais velhas, você pode remover essas observações irrelevantes. Isso pode tornar a análise mais eficiente e minimizar a distração do seu alvo principal, além de criar um conjunto de dados mais gerenciável e com melhor desempenho.

Como limpar os dados

Etapa 2: Corrigir erros estruturais

Erros estruturais ocorrem quando você mede ou transfere dados e percebe convenções de nomenclatura estranhas, erros de digitação ou capitalização incorreta. Essas inconsistências podem causar categorias ou classes rotuladas incorretamente. Por exemplo, você pode encontrar “N/A” e “Não aplicável” ambos aparecendo, mas eles devem ser analisados como a mesma categoria.

Como limpar os dados

Etapa 3: Filtrar valores discrepantes indesejados

Frequentemente, haverá observações únicas onde, à primeira vista, elas não parecem se encaixar nos dados que você está analisando. Se você tiver um motivo legítimo para remover um outlier, como entrada de dados imprópria, isso ajudará no desempenho dos dados com os quais você está trabalhando.

No entanto, às vezes é o aparecimento de um outlier que provará uma teoria na qual você está trabalhando. Lembre-se: só porque um outlier existe, não significa que ele esteja incorreto. Esta etapa é necessária para determinar a validade desse número. Se um outlier provar ser irrelevante para a análise ou for um erro, considere removê-lo.

Como limpar os dados

Etapa 4: lidar com dados ausentes

Você não pode ignorar dados ausentes porque muitos algoritmos não aceitarão valores ausentes. Existem algumas maneiras de lidar com dados ausentes. Nenhuma delas é ideal, mas ambas podem ser consideradas.

1. Como primeira opção, você pode descartar observações que tenham valores ausentes, mas fazer isso descartará ou perderá informações, então esteja ciente disso antes de removê-las.
2. Como segunda opção, você pode inserir valores ausentes com base em outras observações; novamente, há uma oportunidade de perder a integridade dos dados porque você pode estar operando a partir de suposições e não de observações reais.
3. Como terceira opção, você pode alterar a maneira como os dados são usados para navegar efetivamente por valores nulos.

Como limpar os dados

Etapa 5: Validar e fazer controle de qualidade

No final do processo de limpeza de dados, você deve ser capaz de responder a estas perguntas como parte da validação básica:

Os dados fazem sentido?

Os dados seguem as regras apropriadas para seu campo?

Eles provam ou refutam sua teoria de trabalho, ou trazem algum insight à tona?

Você consegue encontrar tendências nos dados para ajudar a formar sua próxima teoria?

Se não, é por causa de um problema de qualidade de dados?

Conclusões falsas por causa de dados incorretos ou "sujos" podem informar uma estratégia de negócios e tomada de decisão ruins.

Conclusões falsas podem levar a um momento embaraçoso em uma reunião de relatórios quando você percebe que seus dados não resistem a apuração.

Antes de chegar lá, é importante criar uma cultura de dados de qualidade em sua organização. Para fazer isso, você deve documentar as ferramentas que pode usar para criar essa cultura e o que a qualidade de dados significa para você.

DATA TRANSFORMATION

O que é data transformation

- A transformação de dados é uma parte crítica do processo de integração de dados no qual **dados brutos são convertidos em um formato ou estrutura unificada**. A transformação de dados garante a compatibilidade com sistemas de destino e melhora a qualidade e a usabilidade dos dados. É um aspecto essencial das práticas de gerenciamento de dados, incluindo data wrangling, análise de dados e data warehousing.
- Por exemplo, uma transformação de dados pode ser tão direta quanto converter um campo de data (por exemplo: MM/DD/AA) em outro, ou dividir uma única coluna do Excel em duas.
- Outras funções avançadas de engenharia de dados incluem normalização de dados, que define relacionamentos entre pontos de dados; e enriquecimento de dados, que complementa informações existentes com conjuntos de dados de terceiros.

Tipos de data transformation

- **Limpeza de dados:** a limpeza de dados melhora a qualidade dos dados corrigindo erros e inconsistências, como a eliminação de registros duplicados.
- **Agregação de dados:** a agregação de dados resume os dados combinando vários registros em um único valor ou conjunto de dados.
- **Normalização de dados:** A normalização de dados padroniza os dados, trazendo todos os valores para uma escala ou formato comum, como valores numéricos de 1 a 10.
- **Codificação de dados:** A codificação de dados converte dados categóricos em um formato numérico, facilitando a análise. Por exemplo, a codificação de dados pode atribuir um número exclusivo a cada categoria de dados.

Tipos de data transformation

- **Enriquecimento de dados:** o enriquecimento de dados aprimora os dados adicionando informações relevantes de fontes externas, como dados demográficos de terceiros ou metadados relevantes.
- **Imputação de dados:** a imputação de dados substitui dados ausentes por valores plausíveis. Por exemplo, pode substituir valores ausentes pela mediana ou valor médio.
- **Divisão de dados:** a divisão de dados divide os dados em subconjuntos para diferentes propósitos. Por exemplo, engenheiros podem dividir um conjunto de dados para usar um para treinamento e outro para teste em aprendizado de máquina.

Tipos de data transformation

- **Discretização de dados:** Na discretização de dados, os dados são convertidos em buckets ou intervalos discretos em um processo às vezes chamado de binning. Como exemplo, a discretização pode ser usada em um ambiente de saúde para traduzir dados como idade do paciente em categorias como "infantil" ou "adulto".
- **Generalização de dados:** A generalização de dados abstrai grandes conjuntos de dados em um formato de nível superior ou resumo, reduzindo detalhes e tornando os dados mais fáceis de entender.
- **Visualização de dados:** A visualização de dados representa os dados graficamente, revelando padrões ou insights que podem não ser imediatamente óbvios.

ETL data transformation vs. ELT data transformation

Extrair, transformar, carregar: Transformação de dados em uma área de preparação.

- No processo ETL, um subconjunto predeterminado de dados estruturados é extraído de sua fonte, onde é transformado em uma área de preparação ou servidor de processamento secundário antes de ser carregado em seu sistema de destino. ETL é mais adequado para armazenamento local e conjuntos de dados menores.
- No entanto, ETL pode ser preferível em cenários com necessidades específicas de qualidade e consistência de dados, pois etapas mais rigorosas de limpeza e validação de dados podem ser introduzidas.
- ETL também pode ser necessário para proteger dados confidenciais, como informações protegidas, durante a migração.

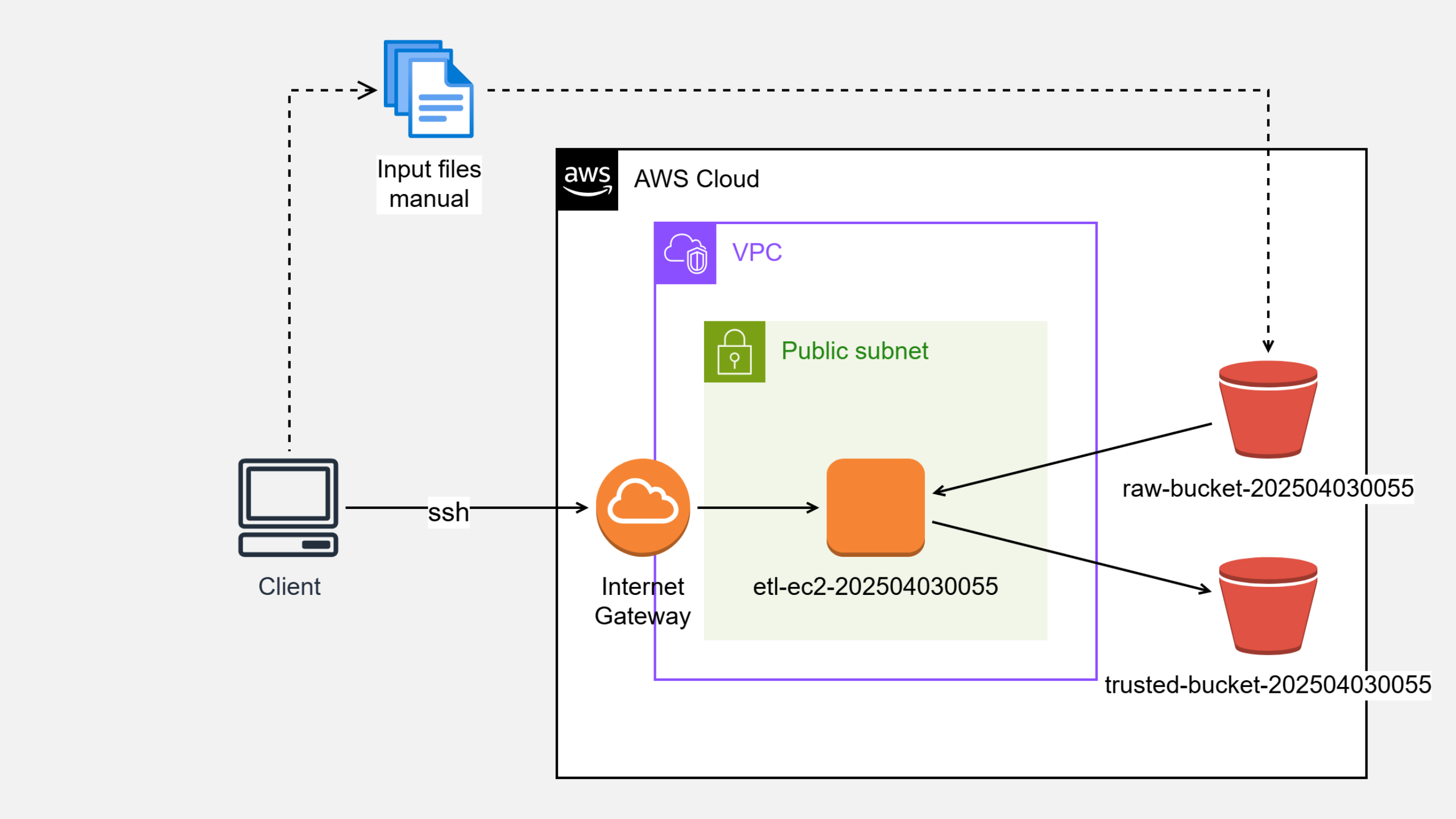
ETL data transformation vs. ELT data transformation

Extrair, carregar, transformar: Transformando dados na nuvem

No processo ELT, as informações são extraídas de fontes de dados e carregadas no sistema de destino baseado em nuvem, onde são transformadas. Essa abordagem, ao aproveitar o poder da computação em nuvem, normalmente permite um processamento mais rápido e um gerenciamento de dados mais ágil.

Ela também pode ser usada com dados não estruturados, como imagens. Com a vantagem da computação baseada em nuvem e do poder de armazenamento, o processo ELT se beneficia de maior escalabilidade.

Lab. Automação ETL


Input files
manual
aws
AWS Cloud
VPC
Public subnet
raw-bucket-202504030055
ssh
Client
Internet
Gateway
etl-ec2-202504030055
trusted-bucket-202504030055

Para esse Lab:

- Crie dois buckets – um para receber o arquivo bruto manualmente e um para receber o arquivo tratado.
- Cria uma instância na sub-rede pública padrão da AWS que vai hospedar nossos scripts python para realizar a ETL.
- Configurar a instância com o ambiente virtual assim como mostrado no Moodle.
- Ajustar os arquivos python para o Bucket pessoal.
- Testar o arquivo base disponibilizado.
- Melhorar o processo.

Agradeço
a sua atenção!

SÃO
PAULO
TECH
SCHOOL