PUC-RIO – Especialização em Ciência de Dados e Analytics.

MVP da Sprint Engenharia de Dados.

Autor: Paulo Roberto Márcio Batista

**Problema:** Empresa XYZ, gostaria de saber a possibilidade da automatização da coleta de dados de vários formatos até a disponibilização destes dados em um dashboard com informações para o Business com pouca ou nenhuma interação humana. Atualmente a empresa mantém vários profissionais para conseguir gerar tal resultado, que mesmo dada a falta de comunicação, muitas vezes os relatórios aparecem com erros em seus resultados.

**Objetivo:** Este trabalho tem por finalidade desenvolver um protótipo de um pipeline automação da coleta até a entrega dos dados para serem trabalhados por analistas de dados. Para esta engenharia de dados, será utilizado serviços de cloud público, conhecido como Oracle Cloud Infrastructure (OCI).

**Ferramentas da OCI utilizadas:**

IAM – Identity and Access Management: Controle de acessos de usuários e criação de políticas de segurança.

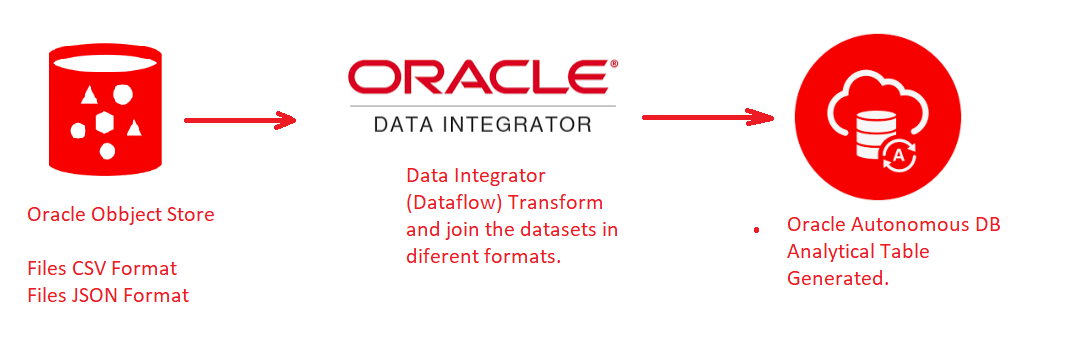
VCN – Virtual Cloud Network: É a rede propriamente, que vai permitir conectar os bancos de dados as áreas de armazenamento, etc...

Oracle Autonomous Database: Banco de dados Oracle OLAP.

Data Integration: Serviço de conexão dos objetos e criação do pipeline de dados (DataFlow).

Object Storage – Buckets: Área para armazenamentos dos dados não estruturados, ou semi-estruturados.

Solução: Para o MVP, a solução apresenta de forma simples e funcional o potencial de obter dados de diferentes formatos, processar (filtrar, remoção de nulos, etc..) e consolidar para futuras análises dos Analistas ou para ferramentas de criação de Dashboard.



Oracle facilita muito o processo de criação de ETL. Vamos aos detalhes da implementação:

**1 – IAM – Políticas de segurança necessária**

Foram utilizadas duas políticas para todo o projeto do MVP.

**Policy:** puc-mvp-private-access-data-integration

**Role:** allow service dataintegration to manage virtual-network-family in compartment PUC-RIO

**Descrição:** Permite conexão entre componentes pela VNC (Rede).

**Policy:** puc-mvp-data-integration-services

**Roles:**

allow group Administrators to manage dis-workspaces in compartment PUC-RIO

allow group Administrators to manage dis-work-requests in compartment PUC-RIO

allow group Administrators to use virtual-network-family in compartment PUC-RIO

allow group Administrators to manage tag-namespaces in compartment PUC-RIO

allow service dataintegration to manage virtual-network-family in compartment PUC-RIO

allow any-user to manage buckets in compartment PUC-RIO where ALL {request.principal.type = 'disworkspace', request.principal.id = 'ocid1.disworkspace.oc1.iad.anuwcljs2ow634yaltgkete5rlfihtlslqufo64ieuao2ym5pt3txibppdpq', request.permission = 'PAR\_MANAGE'}

allow group administratos to use object-family in compartment PUC-RIO

allow any-user to read buckets in compartment PUC-RIO where ALL {request.principal.type = 'disworkspace', request.principal.id = 'ocid1.disworkspace.oc1.iad.anuwcljs2ow634yaltgkete5rlfihtlslqufo64ieuao2ym5pt3txibppdpq', request.operation = 'GetBucket'}

allow any-user to manage objects in compartment PUC-RIO where ALL {request.principal.type = 'disworkspace', request.principal.id = 'ocid1.disworkspace.oc1.iad.anuwcljs2ow634yaltgkete5rlfihtlslqufo64ieuao2ym5pt3txibppdpq'}

**Descrição:** Estas roles são necessárias para autorização de acesso do Data Integrator (Data Flow), com object storage e autonomous db.

**2 – Object Storage (Armazenamento)**

Foi criado um bucket chamado Datalake. Onde serão armazenado os dados não estruturado, semi-estruturado.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Arquivos em diferentes formatos

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente com confiança média

**3 – Oracle Autonomous Database (Analítical Database)**

Foi criando um banco de dados para armazenagem dos dados já tratados, resultado de todo processamento dos dados realizados pelo serviço de data integration.

Tela de celular com publicação numa rede social

Descrição gerada automaticamente

Foi Ativado o monitoramento para análise dos picos de uso, durante a ingestão automática dos dados pelo serviço de integração de dados.

Gráfico

Descrição gerada automaticamente

**4 – Data Integration (Construção do pipeline de automação ETL).**

Foi um workspace para desenvolvimento do pipeline de dados.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

4.1 – Foi criado dois data assets. Um que contém o dados de origem (source) no caso **Data\_Lake**, e outro como destino (target) chamado **Data\_Warehouse**.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

4.1.1 – Data\_Lake e Data\_Warehouse (Source/target) Conexões.

Aqui as conexões lógicas entre o serviço de integração de dados, com o bucket físico.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Aqui, como no item anterior, conecta o target com o Oracle Autonomous DB

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

**5 – Dataflow (Construção do pipeline).**

Primeiro passo foi criar um projeto para o Data Flow, chamado DI\_LAB.

Interface gráfica do usuário, Tabela

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Oracle Data flow, permite automatizar todo o fluxo de dados, obtendo dados de qualquer origem, neste MVP, optei por utilizar somente Buckets com arquivos de diferentes formatos. Porém poderia ser outros bancos oracles, ou de outros provedores de cloud publica como Azure ou AWS. Criando assim um pipeline multi-cloud.

Diagrama

Descrição gerada automaticamente

Descrição do pipeline:

1. Ingestão de aquivos CSV. Neste caso ingestão de clientes.
2. Ingestão de arquivos JSON. Neste caso ingestão das receitas geradas pelos clientes.
3. Base the Clientes (Customers), passa por uma filtragem de somente clientes residentes nos Estados Unidos (US).
4. Base de receitas (Revenue), passa para uma expressão de fitragem (FILL IN). Insere receita 0 em campos nulos.
5. Join\_1, realiza o relacionamento de 1 para muitos, ou seja, 1 cliente para muitas receitas.
6. ADB (Autonomous DB), recebe as tabelas consolidadas, já tratadas e validadas para a relização dos trabalhos dos analisas de dados.

Abaixo exemplo da propriedade da filtragem de cliente do US.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

**6 – Autonomous DB Insights**

Uma vez finalizado o pipeline e executado, o próprio Autonomous DB pode gerar Dashboards, customizados para atender os executivos. Apesar de fora do escopo deste MVP, o Autonomous DB integra Machine Learning, podendo desta forma realizar predições sobre o pipeline gerado.

Tela de celular com publicação numa rede social

Descrição gerada automaticamente

**7 - Conclusão**

Sim é possível, automatizar todo os processo conhecido como ETL (Extract, Tranform and Load). Com isso este MVP demonstra a possibilidade do uso da nuvem pública da Oracle para tal finalidade. Como se trata de um MVP, podemos observar o valor do uso da OCI com a integração entre seus componentes de forma gráfica e facilitada.

Como continuidade de um projeto futuro, ficou demonstrando também a possibilidade de geração de Insighs de negócios, e como comentado ainda a possibilidade de utilizar Machine Learning dentro do Autonomous DB para realização de inferências.