**Data Lake Credit Risk**

**Documentação**

**1. Introdução**

O Data Lake para Análise de Risco de Crédito foi concebido para centralizar e organizar grandes volumes de dados provenientes de diferentes fontes, oferecendo uma base robusta e estruturada para que as equipes de análise e ciência de dados acessem informações confiáveis e bem processadas. Essa infraestrutura é projetada de forma modular, com camadas que representam diferentes estágios de processamento e maturidade dos dados.

Inicialmente, o foco principal do Data Lake é receber e processar arquivos CSV contendo informações sobre clientes de produtos de crédito. Embora seu objetivo inicial seja atender a essa necessidade específica, o sistema foi desenvolvido com flexibilidade para futuras expansões e novos casos de uso.

**Objetivos Principais:**

* Centralizar dados diferentes fontes.
* Validar e transformar dados para garantir a qualidade.
* Suporte para análise e decisão estratégica.
* Preparação para futuros modelos preditivos e análise avançada.

**2. Arquitetura e Tecnologias Utilizadas  
A arquitetura do Data Lake é construída sobre uma série de ferramentas da AWS para garantir escalabilidade, flexibilidade, segurança e desempenho.**

**2.1. Ferramentas e Funcionalidades**

* **AWS S3 (Simple Storage Service):**
  + **Armazena dados nas camadas Raw, Bronze, Silver e Gold, organizados hierarquicamente.**
  + **Garante alta durabilidade e disponibilidade dos dados.**
* **AWS Lambda:**
  + **Orquestra o fluxo de dados entre as camadas, acionando transformações automáticas com base em eventos do S3.**
  + **Gera relatórios de validação e executa funções para iniciar processos no Glue.**
* **AWS Glue:**
  + **Serviço ETL que realiza transformações e validações nas camadas Silver e Gold.**
  + **Automatiza processos de tratamento e consolidação de dados.**
* **IAM (Identity and Access Management):**
  + **Gerencia acessos e permissões, garantindo que apenas usuários e serviços autorizados possam acessar os dados e funcionalidades do sistema.**
  + **Permite a aplicação de políticas de segurança detalhadas para proteger os recursos do Data Lake.**
* **CloudWatch:**
  + **Coleta logs e métricas para monitorar o desempenho das ferramentas e processos no Data Lake.**
  + **Ajuda na identificação e resolução de problemas operacionais por meio de alarmes e visualização em painéis de controle (dashboards).**
* **Amazon SES (Simple Email Service):**
  + **Envia relatórios automatizados sobre dados validados e erros na camada Bronze.**
* **Flask e AWS API Gateway:**
  + **Recebem arquivos enviados via HTTP e armazenam-nos na camada Raw.**
  + **Atuam como interface para interações externas.**
* **PySpark:**
  + **Processa grandes volumes de dados para aplicar transformações complexas e enriquecer os dados.**
* **Amazon Athena:**
  + **Realiza consultas SQL rápidas e eficientes na camada Gold.**

**2.2. Estrutura de Armazenamento**

* **Raw:** Dados brutos enviados pela aplicação Flask.
* **Bronze:** Dados validados e categorizados em "Validated" e "Rejected".
* **Silver:** Dados transformados e enriquecidos, prontos para análises aprofundadas.
* **Gold:** Dados otimizados e consolidados para análises e modelos preditivos.

**3. Book de Variáveis**

**3.1. Introdução**

O **Book de Variáveis** é um guia que descreve as variáveis utilizadas nas camadas Silver e Gold do Data Lake, detalhando nome, tipo de dado e finalidade. Ele facilita a compreensão, padronização e análise dos dados, servindo como referência para as equipes de engenharia e ciência de dados.

**3.2. Variáveis Consolidadas**

|  |  |  |
| --- | --- | --- |
| **Variável** | **Tipo** | **Descrição** |
| **Camada Silver** | | |
| idade\_da\_pessoa | int | Idade da pessoa que está solicitando o empréstimo. |
| renda\_da\_pessoa | float | Renda anual da pessoa. |
| propriedade\_da\_casa\_da\_pessoa | str | Status da propriedade da casa da pessoa (ex: Alugada, Própria, Hipoteca, etc). |
| tempo\_de\_emprego | int | Tempo, em anos, de emprego da. |
| intenção\_do\_empréstimo | str | Intenção do empréstimo (ex: Consolidação de Dívidas, Educação, etc.). |
| classificação\_do\_empréstimo | str | Classificação do empréstimo (ex: A, B, C). |
| quantidade\_do\_empréstimo | float | Valor total do empréstimo solicitado. |
| taxa\_de\_juros\_do\_empréstimo | float | Taxa de juros do empréstimo. |
| status\_do\_empréstimo | str | Status atual do empréstimo (ex: Aprovado, Negado, Pendente). |
| percentual\_de\_renda\_do\_empréstimo | float | Percentual da renda comprometido com o empréstimo. |
| historico\_de\_inadimplencia | str | Histórico de inadimplência registrado no banco de dados de crédito. |
| tempo\_de\_crédito\_da\_pessoa | int | Tempo de histórico de crédito da pessoa (em anos). |
| data\_de\_extracao | date | Data de extração dos dados. |

|  |  |  |
| --- | --- | --- |
| **Variável** | **Tipo** | **Descrição** |
| **Camada Gold** | | |
| quantidade\_total\_empréstimos | int | Quantidade total de empréstimos na classificação do empréstimo. |
| valor\_total\_empréstimos | float | Valor total dos empréstimos na classificação do empréstimo. |
| valor\_médio\_empréstimos | float | Valor médio dos empréstimos na classificação do empréstimo. |
| porcentual\_inadimplentes | float | Percentual de inadimplentes na classificação do empréstimo. |
| faixa\_de\_tempo\_de\_emprego | str | Faixa de tempo de emprego (ex: 0-1, 1-3, 3-5, 5 ou maior em anos). |
| quantidade\_inadimplente | int | Quantidade de empréstimos inadimplentes em cada faixa de tempo de emprego. |
| percentual\_inadimplência | float | Percentual de inadimplência em cada faixa de tempo de emprego. |
| idade\_média | float | Idade média das pessoas que solicitaram o empréstimo com base na intenção. |
| renda\_média | float | Renda média das pessoas que solicitaram o empréstimo com base na intenção. |
| tempo\_médio\_de\_emprego | float | Tempo médio de emprego, em anos, das pessoas que solicitaram o empréstimo com base na intenção. |
| faixa\_de\_renda | str | Faixa de renda anual da pessoa (ex: R$0 a R$17.999,99, R$18.000,00 a R$35.999,99, etc.). |
| faixa\_tempo\_credito | str | Faixa de tempo de crédito, em anos (ex: 0-1, 1-3, 3-5, 5 ou maior). |
| quantidade\_total\_empréstimos | int | Quantidade total de empréstimos em cada faixa (ex: faixa de renda ou tempo de crédito). |
| quantidade\_inadimplente | int | Quantidade de empréstimos inadimplentes em cada faixa (ex: faixa de renda ou tempo de crédito). |
| propriedade\_da\_casa\_da\_pessoa | str | Tipo de propriedade da casa da pessoa (ex: Alugar, Possuir, Hipoteca, etc.). |
| percentual\_inadimplência | float | Percentual de inadimplência de acordo com o tipo de propriedade ou faixa categorizada. |

**4. Fluxo de Dados e Processamento**

**4.1. Ingestão de Dados**

**Aplicação Flask e AWS API Gateway:**

* A aplicação Flask recebe arquivos via HTTP POST e os envia para o S3 (credit-risk-datalake).
* **Pasta Raw:** Armazenamento de dados brutos sem validação adicional.

**4.2. Camada Bronze**

A camada Bronze realiza a validação dos dados para garantir que apenas dados válidos avancem para a camada seguinte.

* **Objetivos:**
  + Validar os dados recebidos.
  + Sinalizar para o responsável a entrada de dados no Data Lake.
* **Fluxo:**
  + A AWS Lambda é acionada sempre que um arquivo é enviado para Raw.
  + Arquivos aprovados são movidos para **Bronze/Validated**, enquanto arquivos rejeitados vão para **Bronze/Rejected**.
  + Relatórios de validação são gerados e enviados para o e-mail do responsável cadastrado via Amazon SES e para **Bronze/Relatorio.**
* **Estrutura:**
  + **Bronze/Validated:** Dados validados prontos para transformação.
  + **Bronze/Rejected:** Arquivos com erros ou inconsistências.
  + **Bronze/Relatorio:** Relatórios de inconsistências gerado.

**4.3. Camada Silver**

Na camada Silver, os dados passam por transformações e enriquecimentos para se tornarem mais úteis para análises.

* **Objetivos:**
  + Garantir a qualidade dos dados recebidos e possibilitar uma estrutura de dados sólida e confiável
* **Fluxo:**
  + A AWS Lambda é acionada sempre que um arquivo é enviado para Validated.
  + L**impeza de dados**, como remoção de duplicatas e preenchimento de valores nulos e adição da coluna data\_da\_extração com a data de chegada do arquivo
  + **Transformações** para unificação de formatos (ex.: tradução de colunas para português, transformação de datas).
  + Organização dos dados em **tabelas estruturadas** na pasta **Silver** (usando formatos otimizados como Parquet).
* **Estrutura:**
  + **Silver/(**nome do arquivo**)**: Dados tratados.

**4.4. Camada Gold**

A camada Gold é onde os dados são consolidados para análises e modelos preditivos.

* **Objetivos:**
  + Criar **tabelas analíticas** consolidadas.
  + Realizar **agregações e cálculos estatísticos** para ajudar na análise do risco de crédito.
  + Possibilitar **uso** **de dados analíticos** para Analistas e Cientistas.
* **Processo:**
  + **Entrada de dados:** Dados extraídos da camada Silver em formato Parquet.
  + **Transformações:** Cálculos e agregações para criar métricas e categorizações, como faixas de renda e tempo de crédito.
  + **Tabelas geradas:** Exemplos de tabelas incluem "Resumo de Risco de Crédito", "Inadimplência por Faixa de Renda", entre outras.
  + **Armazenamento:** Tabelas armazenadas no S3 em diretórios temáticos, otimizadas para consultas analíticas.
* **Estrutura:**
  + **Gold/(**nome do arquivo**)/(**nome da tabela**)**: Dados gerados.

**5. Qualidade dos Dados e Governança**

**5.1. Garantia de Qualidade**

A qualidade dos dados é garantida por meio de:

* Validações automatizadas na camada Bronze.
* Limpeza e enriquecimento de dados nas camadas Silver.
* Agregação e consolidação de dados na camada Gold para promover insights consistentes.

**5.2. Governança de Dados**

A governança é assegurada por:

* Logs detalhados de erro em todas as etapas, armazenados no S3 para auditoria.
* Regras de acesso controladas, garantindo a segurança dos dados.

**6. Processos Automatizados e Orquestração**

A orquestração do fluxo de dados é feita principalmente pela AWS Lambda, que automatiza a movimentação de dados entre as camadas. O AWS Glue executa as transformações complexas, e o Amazon SES garante a entrega automatizada de relatórios.

**7. Consultas e Análises**

A camada Gold foi projetada para ser otimizada para consultas rápidas e eficientes. O Amazon Athena é utilizado para consultas SQL diretas, permitindo que analistas e cientistas de dados acessem dados prontos para análise.

**8. Resultados Esperados e Impacto**

**8.1. Resultados Esperados**

* **Dados estruturados e otimizados** para consultas rápidas e análises detalhadas.
* **Facilidade de análise** devido à segmentação e agregação dos dados.
* **Suporte à tomada de decisão** estratégica no processo de concessão de crédito.

**8.2. Impacto**

* Melhoria da qualidade dos dados disponibilizados para os analistas.
* Otimização do tempo de resposta para consultas e análises exploratórias.
* Fornecimento de uma base sólida para a construção de modelos preditivos de risco de crédito.

**9. Considerações Finais**

Este Data Lake foi projetado para ser uma solução escalável, robusta e eficiente, que organiza e transforma grandes volumes de dados em insights acionáveis para a análise de risco de crédito. Com uma arquitetura baseada em camadas e ferramentas da AWS, ele oferece alta performance, confiabilidade e flexibilidade, adaptando-se às mudanças e novas demandas de dados. A implementação dessa arquitetura garante um sistema de dados mais seguro, acessível e pronto para suportar decisões estratégicas e preditivas no domínio de risco de crédito.