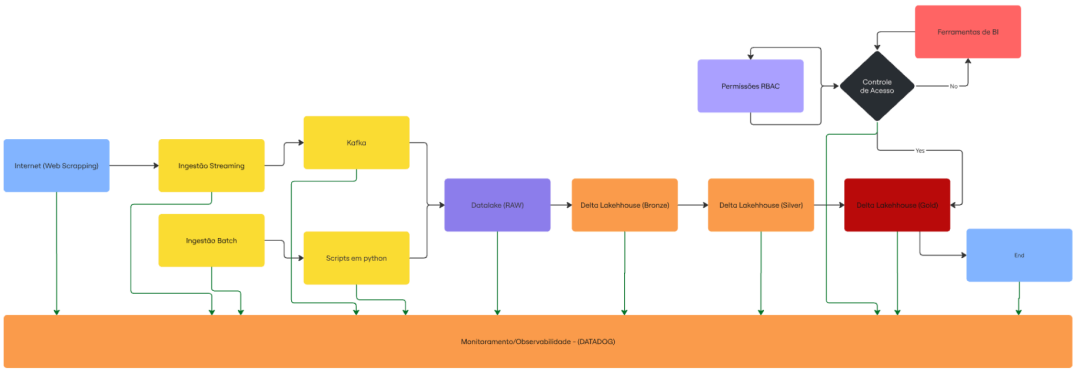
1. Visão Geral

Este documento apresenta a arquitetura proposta para o processamento de dados gerados através da coleta de dados realizada através de web scrapping em páginas da internet, contemplando fluxos de ingestão em tempo real (streaming) e em lotes (batch) visando crescimento da plataforma em um processo de expansão, ou processamento massivo. A solução utiliza uma abordagem moderna de lakehouse, com camadas bem definidas para garantir a qualidade, governança e disponibilidade dos dados para análises.



2. Arquitetura de Ingestão de Dados

2.1 Fontes de Dados

- Páginas de internet (através de web scraping)

2.2 Métodos de Ingestão

- **Streaming**: Captura de dados em tempo real de páginas de internet.
- **Batch**: Processamento em lote de conjuntos de dados, ou processamento massivo para refletir na aplicação.

3. Arquitetura de Armazenamento e Processamento

3.1 Data Lake (Camada Raw)

- Repositório para dados brutos e não estruturados, repositório mantido em arquivos conforme a coleta.
- Armazenamento separado para:
    - Dados de streaming
    - Dados em batch
- Não há unificação de dados nesta camada

- Preservação do formato original dos dados mantendo exatamente da mesma forma conforme foi coletado na internet para fins de auditoria.

3.2 Delta Lakehouse (Camada Bronze)

- Unificação dos dados de streaming e batch em uma única tabela delta.
- Aplicação de schema inferido
- Manutenção da rastreabilidade da origem dos dados
- Primeira camada com formato Delta Lake

3.3 Camada Silver

- Dados refinados e limpos
- Aplicado deduplicação
- Aplicação de transformações básicas
- Validação de qualidade de dados
- Normalização e padronização

3.4 Camada Gold

- Dados modelados no formato star-schema
- Otimização para consultas analíticas
- Criação de agregações e métricas de negócio
- Preparação para consumo por ferramentas de BI

4. Governança de Dados

4.1 Unity Catalog

- Implementação do Unity Catalog para governança centralizada
- Catálogo unificado de metadados
- Controle de acesso granular
- Auditoria de acesso aos dados

4.2 Rastreabilidade

- Registro de linhagem de dados entre camadas
- Monitoramento de transformações
- Histórico de alterações
- Versionamento de dados

4.3 Observabilidade

- Todo o processo conforme exemplificado na figura acima será monitorado pelo software da datadog, mostrando todos os dados referentes ao ambiente

e aplicação, mostrando passo a passo de todo o processo de pipeline de dados

4.4 Orquestração

- Orquestração dos jobs se darão pelo software airflow, onde será o responsável por controlar os agendamentos, execuções e dependências.
- Cada agendamento será representado como todo o processo da imagem representada acima.

5. Fluxo de Processamento

1. **Ingestão**: Dados do aplicativo móvel são capturados via streaming e batch
2. **Armazenamento Raw**: Dados são armazenados em seu formato original no data lake, separados por tipo de ingestão
3. **Processamento Bronze**: Dados são unificados e recebem schema inferido no Delta Lakehouse
4. **Processamento Silver**: Dados são refinados, limpos e transformados
5. **Processamento Gold**: Dados são modelados em star-schema para análises
6. **Consumo**: Dados são disponibilizados para consumo via ferramentas de BI
7. **Observabilidade:** A cada parte do processo será compartilhado dados com o software de observalidade para que possa ser mensurado desde a saúde do ambiente, até a eficiência do processo na realização de uma nova execução dos jobs.

6. Boas Práticas

- O processo será todo versionado utilizando os melhores recursos do GitFlow sendo mantido o seu repositório no GitHub e utilizando com esteira de CI/CD a própria esteira do GitHub Actions, onde o desenvolvimento será realizado para as branchs de feature e a promoção para as branches bloqueadas (develop, homolog a master) deverá ser concedidas somente através de pull requests mediante aprovação de code review e aprovação na execução de testes contínuos validando sempre se não vai fazer algo que já existia parar de funcionar.