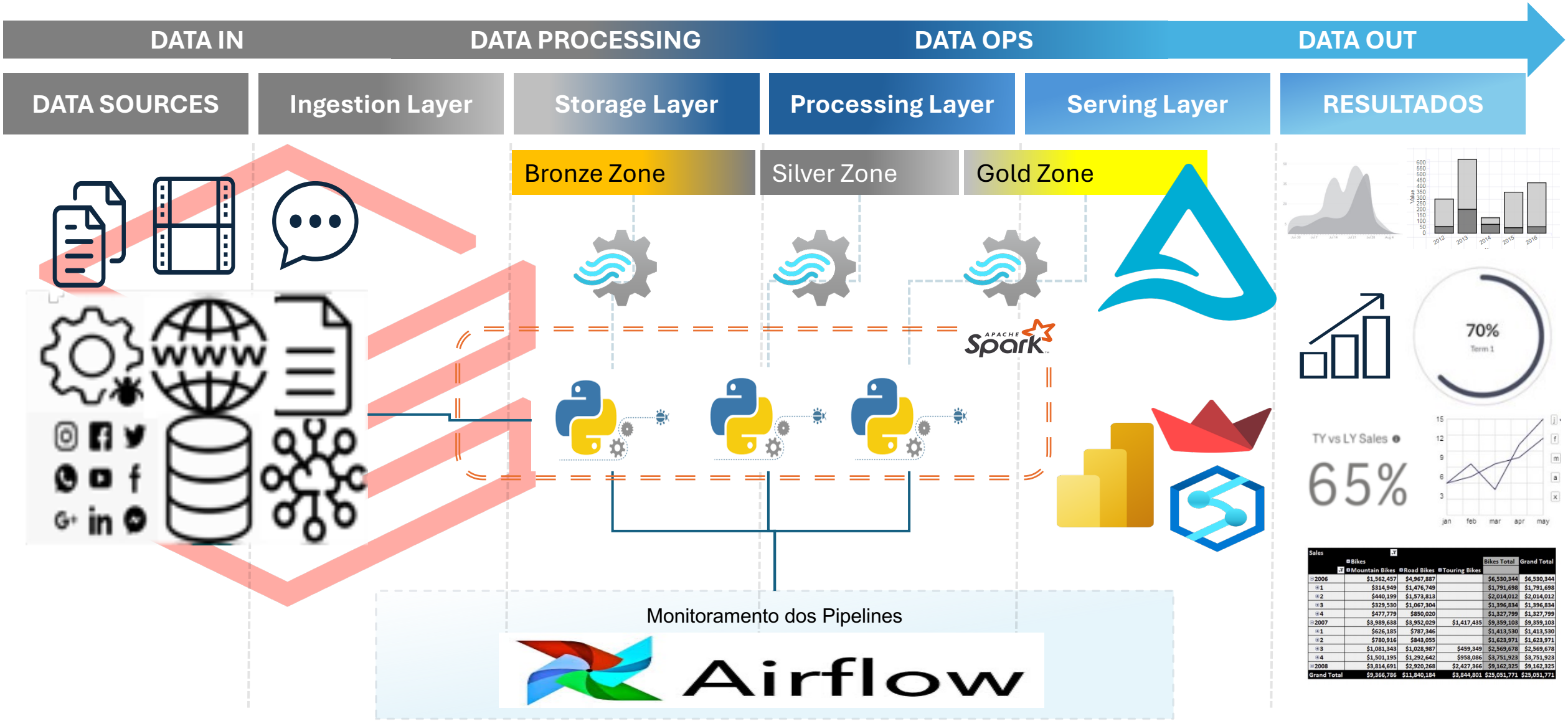
Arquitetura de Lake House para a CondoManage

Anderson Lima Rocha
Especialista **- Engenharia de Dados**

A CondoManage, uma empresa que fornece serviços de administração de condomínios e imobiliárias, está migrando sua infraestrutura de dados para uma arquitetura de Lake House. Essa mudança visa gerenciar e analisar grandes volumes de dados de propriedades, moradores e transações de forma mais eficiente. Para alcançar esse objetivo, a CondoManage contratou um engenheiro de dados para construir uma arquitetura moderna, desenvolver processos de ingestão de dados e implementar transformações de dados usando Spark e Python.

Desenho da Arquitetura do Lake House



Sales	Bikes			Bikes Total	Grand Total
	Mountain Bikes	Road Bikes	Touring Bikes		
2006	$1,562,457	$4,967,887		$6,530,344	$6,530,344
1	$314,949	$1,476,749		$1,791,698	$1,791,698
2	$440,199	$1,573,813		$2,014,012	$2,014,012
3	$329,530	$1,067,304		$1,396,834	$1,396,834
4	$477,779	$850,020		$1,327,799	$1,327,799
2007	$3,989,638	$3,952,029	$1,417,435	$9,359,103	$9,359,103
1	$626,185	$787,346		$1,413,530	$1,413,530
2	$780,916	$843,055		$1,623,971	$1,623,971
3	$1,081,343	$1,028,987	$459,349	$2,569,678	$2,569,678
4	$1,501,195	$1,292,642	$958,086	$3,751,923	$3,751,923
2008	$3,814,691	$2,920,268	$2,427,366	$9,162,325	$9,162,325
Grand Total	$9,366,786	$11,840,184	$3,844,801	$25,051,771	$25,051,771

Camada de Ingestão de Dados

1

Fontes de Dados

A CondoManage coleta dados de diversas fontes, incluindo APIs de provedores de serviços, arquivos CSV com informações de moradores e sistemas internos da plataforma.

2

Extração de Dados

O script em Python extrai os dados das tabelas de condomínios, moradores, imóveis e transações no banco de dados PostgreSQL.

3

Armazenamento no Data Lake (Bonze Zone)

Os dados processados são armazenados na camada Bronze do Data Lake, preservando a estrutura original para análise posterior, num formato otimizado em Parquert.

Camada de Armazenamento de Dados

Data Lake

O Data Lake da CondoManage armazenará dados brutos e processados de diversas fontes, incluindo dados estruturados, semi e não estruturados, em tecnologia **DELTA LAKE**, sendo ele dividido em 3 camadas: Bronze, Silver e Gold.

Camada Bronze

Área que armazenará dados brutos, preservando a estrutura original para análise e recuperação de informações históricas.

Camada Silver

Área que armazenará os dados processados e organizados, para facilitar as consultas e analises, atendendo as necessidades operacionais e táticas e ações de Ciencias de Dados.

Camada Gold

Servir como base de dados refinados para antender tomadas de decisões Táticas e Estratégicas, como um DW, para consumo direto de Dashboards e modelos de Machine Learning.

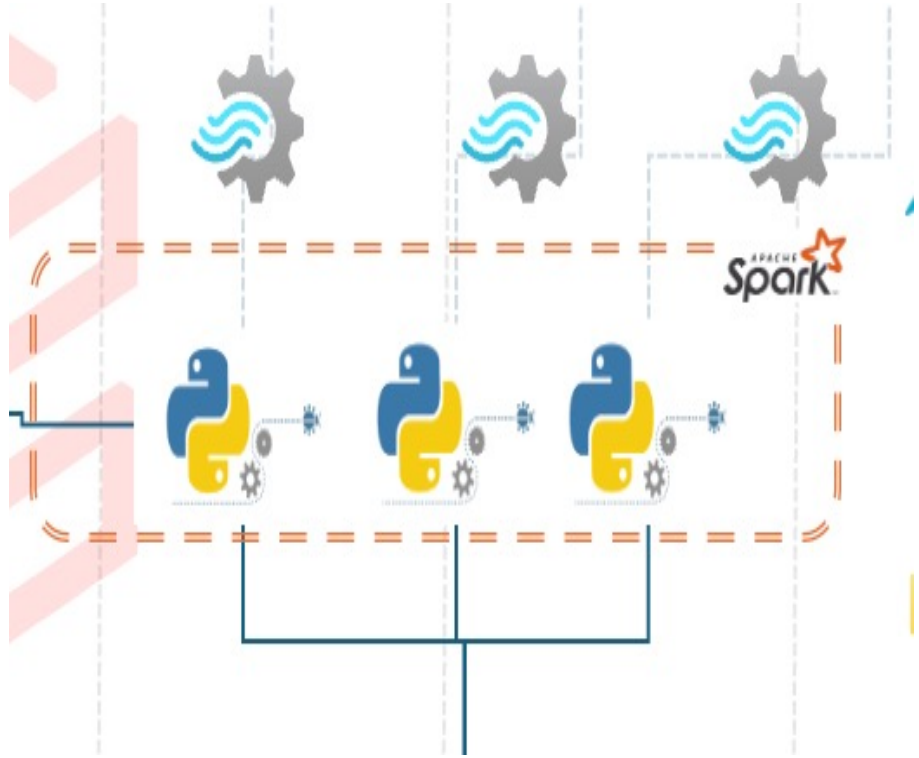
Camada de Processamento de Dados

1. **Processamento Batch**
   O processamento batch permite a análise de grandes volumes de dados em intervalos regulares, ideal para tarefas como relatórios e análises agregadas.

2. **Processamento Stream**
   O processamento stream permite o processamento de dados em tempo real, ideal para análises de dados de sensores, eventos de uso e monitoramento de sistemas.

3. **Spark**
   O Spark é uma ferramenta de processamento distribuído ideal para trabalhar com grandes volumes de dados no Data Lake, proporcionando alta performance e escalabilidade.

4. **Transformações e Agregações**
   Os dados são transformados, agregados e enriquecidos durante o processamento, preparando-os para análise e visualização.



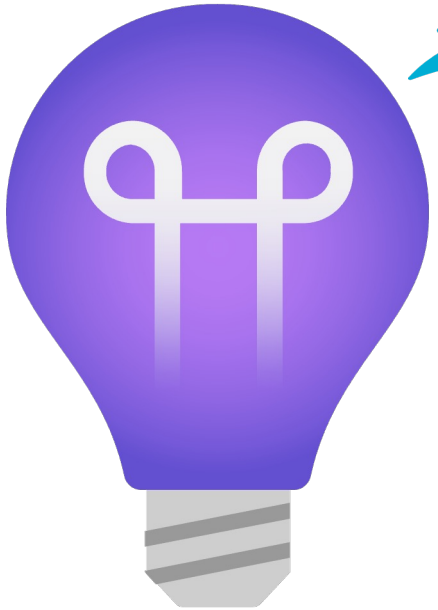
Camadas de Serviços

API de Dados

A camada permitirá que outras aplicações acessem os dados processados do Data Lake (DW), integrando dados com dashboards, relatórios e ferramentas de BI.

Data Warehouse

O Data Warehouse estará na camada Gold, com dados processados e agregados, otimizados para consulta rápida e análise de negócios.

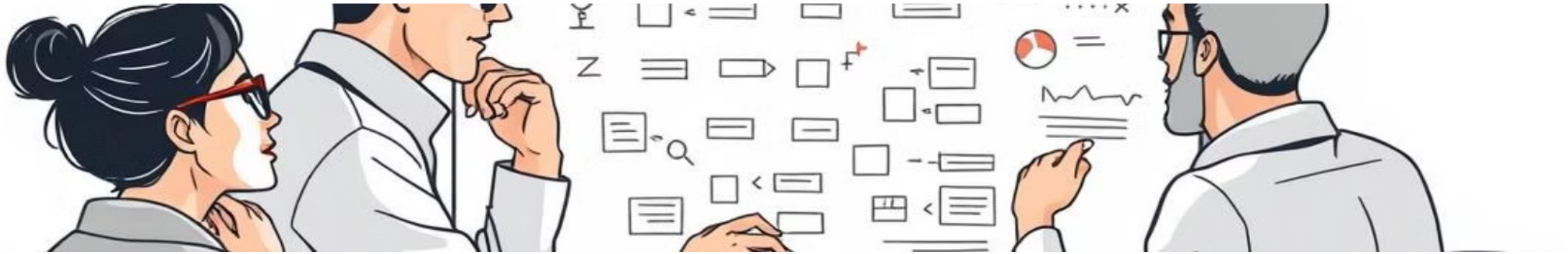


KIP’s Corporativos

Dashboards interativos fornecem visualizações personalizadas de dados, permitindo análises de alto nível e monitoramento de indicadores chave.

Implementação da Arquitetura

Ferramentas	Descrição
Delta Lake	Armazenamento de dados em nuvem, escalável e seguro.
Azure Databricks	Plataforma para processamento de dados com Spark, facilitando o desenvolvimento, armazenamento e segurança.
Power BI	Ferramenta de BI para visualização e análise de dados, com dashboards interativos e relatórios personalizados.
Apache Airflow	Monitoramento e orquestração de pipelines de processamento.
Python	Será utilizado para extração e ingestão de dados e em conjunto com o Spark (pySpark) para processor e tartar as informações



Documentação

A documentação é essencial para garantir a compreensão e a manutenção da arquitetura de Lake House e dos pipelines de processamento. A CondoManage utiliza uma variedade de métodos de documentação, incluindo documentação de código, diagramas de arquitetura, documentação de processos e wiki interna. A documentação é mantida atualizada e acessível a todos os membros da equipe, garantindo a consistência e a clareza das informações.

Documentação de Código

Documentação detalhada do código-fonte, incluindo comentários e descrições de funções, versionamentos de códigos e processos.

Diagramas de Arquitetura

Diagramas visuais que representão a arquitetura do Lake House e dos pipelines de processamento.

Documentação de Processos

Documentação detalhada dos processos de ingestão, processamento e transformação de dados.

Wiki Interna

Um repositório centralizado de informações sobre a arquitetura de Lake House, os pipelines de processamento e outros recursos relevantes e consulta rápida.

Obrigado!

Anderson Lima Rocha

E-mail: andersonlimarocha@gmail.com

Fone/WhatsApp: (11) 96128-9707