SOLUTION SPRINT

MBA ENGENHARIA DE DADOS

Lucas Brandi e Regina Cantele

Versão 2

LISTA DE FIGURAS

SUMÁRIO

1 O CENÁRIO

Modern Data Stack geralmente se refere a uma coleção de tecnologias que compõem uma plataforma de dados nativa da nuvem, que tem como objetivo reduzir a complexidade na execução de uma plataforma de dados tradicional.

Os data warehouses em nuvem com seus recursos de processamento paralelo massivo (MPP) e suporte SQL tornou o processamento de grandes volumes de dados mais rápido e barato. Isso levou ao desenvolvimento de muitas ferramentas de dados nativas da nuvem que são de baixo código, fáceis de integrar, escaláveis e econômicas. Essas ferramentas e tecnologias são chamadas coletivamente de Modern Data Stack (MDS).

Os componentes básicos de uma plataforma de dados (na direção do fluxo de dados) são:

a) Coleta e Rastreamento de Dados: as ferramentas se concentram na redução de problemas de qualidade que surgem devido ao rastreamento de dados mal projetados, implementados incorretamente, perdidos ou atrasados.

b) Ingestão de dados: pipelines que trazem dados brutos de centenas de fontes próprias e de terceiros para o data warehouse.

c) Transformação de Dados: as ferramentas fornecem estruturas que permitem um design de modelo de dados consistente, promovendo a reutilização e testabilidade do código.

d) Armazenamento de dados: fornecer dimensionamento automático sem servidor, desempenho ultrarrápido, economias de escala, melhor governança de dados e alta produtividade do desenvolvedor.

e) Camada de métricas (Headless BI): fica entre os modelos de dados e as ferramentas de BI, permitindo que as equipes de dados definam métricas de forma declarativa em diferentes dimensões. Fornece uma API que converte solicitações de cálculo de métrica em consultas SQL e as executa no data warehouse.

f) Ferramentas de BI: se concentram em permitir a democratização dos dados, tornando mais fácil para qualquer pessoa na organização analisar dados rapidamente e criar relatórios ricos em recursos.

g) ETL reverso: processo de mover dados transformados do data warehouse para sistemas downstream, como operações, finanças, marketing, CRM, vendas e até mesmo de volta ao produto, para facilitar a tomada de decisões operacionais.

h) Orquestração: fornecer gerenciamento de ponta a ponta de cronogramas de fluxo de trabalho, amplo suporte para dependências complexas destes fluxos e integração perfeita com componentes de infraestrutura modernos, como o Kubernetes.

i) Gerenciamento de dados, qualidade e governança: permitir um alto nível de transparência, colaboração e democratização de dados.

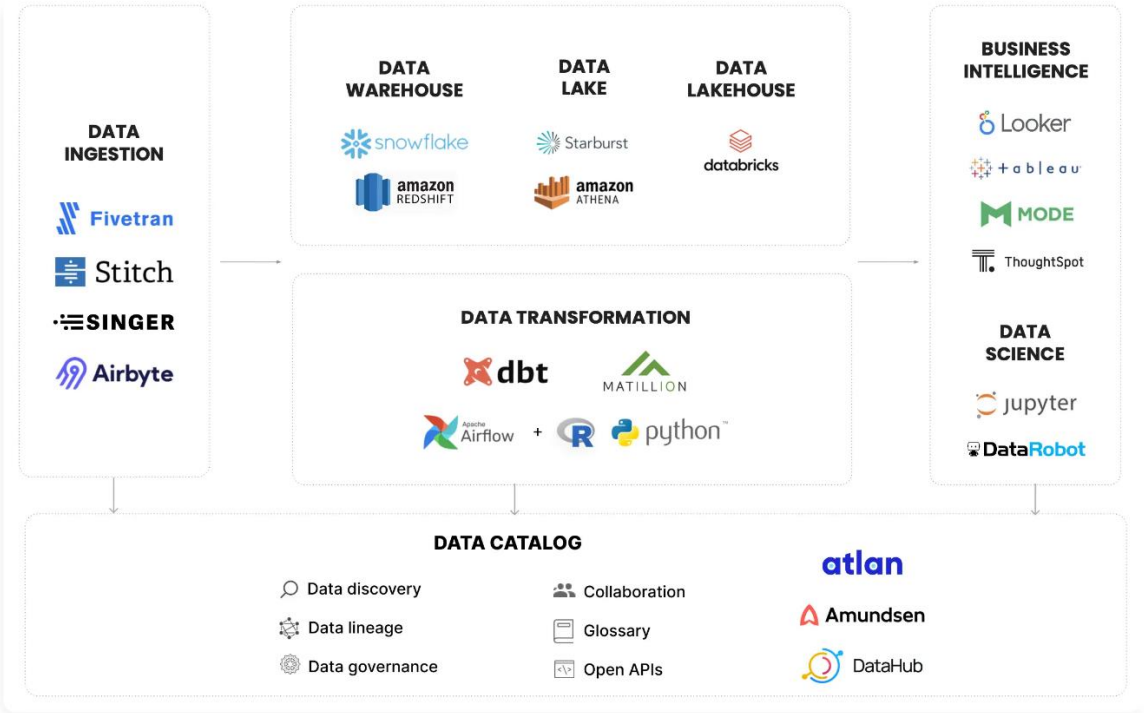

Figura 1 – Visão *Modern Data Stack*
Fonte: Atlan (2022)

Seu histórico contempla desde BigQuery até Dataform, passando pelo Snowflake e Metabase.

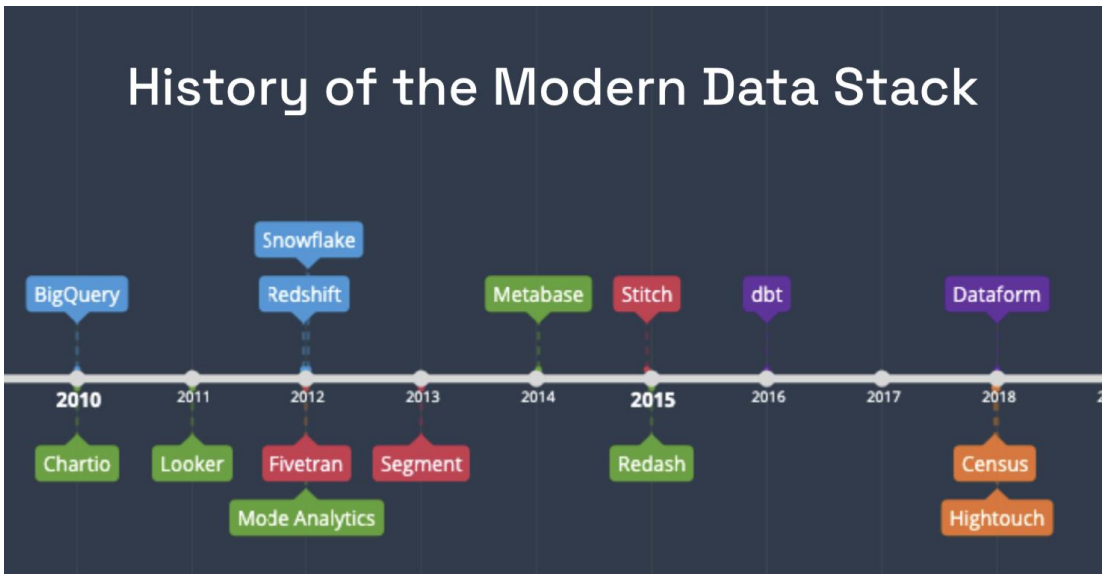

Figura 2 - História da Modern Data Stack
Fonte: Volz (2021)

Os principais recursos da *Modern Data Stack*:

a) Oferecido como um serviço gerenciado: requer configuração mínima ou nenhuma dos usuários e absolutamente nenhuma engenharia necessária.

b) Centrado em torno de um *Cloud Data Warehouse* (CDW): Tudo “simplesmente funciona” de prateleira se as empresas usarem um CDW popular. Ao ser opinativo sobre onde estão os dados, são eliminadas integrações confusas e as ferramentas funcionam bem juntas.

c) Democratiza os dados por meio de um ecossistema centralizado em SQL: as ferramentas são criadas para engenheiros de dados, analistas de dados e usuários de negócios. Esses usuários geralmente sabem mais sobre os dados de uma empresa, por isso faz sentido tentar aprimorá-los, fornecendo a eles ferramentas que falam seu idioma.

d) Cargas de trabalho elásticas: pagar pelo uso. Escalar instantaneamente para lidar com grandes cargas de trabalho. O dinheiro é a única limitação de escala na nuvem moderna.

e) Concentrar nos fluxos de trabalho operacionais: as ferramentas de apontar e clicar são boas para usuários com baixo conhecimento em tecnologia, mas não fazem sentido se não houver um caminho viável para a produção. As ferramentas que compõe a *Modern Data Stack* geralmente são criadas com a automação como uma das características principal.

A *Modern Data Stack* deve evoluir para contemplar *data mesh*, *metrics laye*r, *Active Metadata & terceira geração de Data Catalog* e *data observability*.

Data Mesh – malha de dados - não é uma plataforma ou um serviço para poder comprar na prateleira. É um conceito de design com algumas características como propriedade distribuída, design baseado em domínio, descoberta de dados e padrões de envio de produtos de dados - todos os quais valem a pena tentar operacionalizar em sua organização.

Metrics layer tem o objetivo de resolver o problema das métricas em Business Intelligence. As métricas são críticas para avaliar e impulsionar o crescimento de uma empresa, mas elas vêm enfrentando dificuldades há anos. Geralmente estão divididas em diferentes ferramentas de dados, com diferentes definições para a mesma métrica em diferentes equipes ou painéis. Grandes empresas como o Airbnb anunciaram a construção de uma plataforma de métricas própria para resolver esse problema. Airbnb criou a camada denominada Minerva.

Os catálogos de dados de terceira geração e metadados ativos são construídos em torno de diversos ativos de dados, “grandes metadados”, visibilidade de dados de ponta a ponta e colaboração incorporada. As plataformas de metadados ativas atuam como plataformas bidirecionais — elas não apenas reúnem metadados em um único armazenamento, mas também aproveitam “metadados reversos” para disponibilizar metadados em fluxos de trabalho diários.

Data observability surgiu do “tempo de inatividade de dados” - se referindo a períodos de tempo em que os dados são parciais, errôneos, ausentes ou imprecisos. O tempo de inatividade de dados faz parte da vida normal de uma equipe de dados há anos. Foi aí que surgiu a observabilidade dos dados – a ideia de monitoramento, rastreamento e triagem de incidentes para evitar o tempo de inatividade.

Somado a esta visão, Inteligência Artificial entra fortemente numa camada. Qualquer um que tenha trabalhado em ciência de dados na última década provavelmente está familiarizado com a “Hierarquia de Necessidades da Ciência de Dados”.

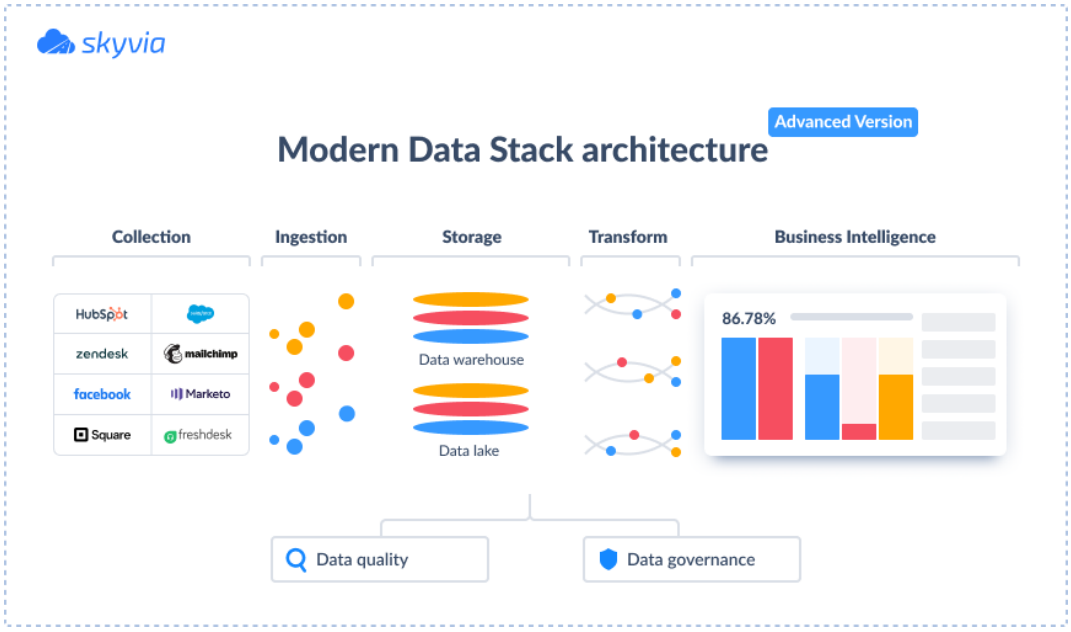

Figura 3 - Hierarquia de Necessidades da Ciência de Dados
Fonte: MAKSYMIUK (2022)

A ideia é que cada camada seja baseada na que está abaixo dela, e a Figura 4 “Comparativo Necessidade para IA e MDS”, ilustra as dependências necessárias para chegar a um ponto em que se possa começar a resolver problemas de IA. Se uma empresa não tem uma história sólida de como está coletando, armazenando e modificando dados, qualquer projeto de ciência de

dados está condenado antes de começar porque as bases sobre as quais se baseia mudam rapidamente.

Figura 4 - Comparativo Necessidade para IA e MDS
Fonte: Volz (2021)

A IA representa uma enorme oportunidade de crescimento para muitas empresas e nada como agilizar este crescimento com MDS.

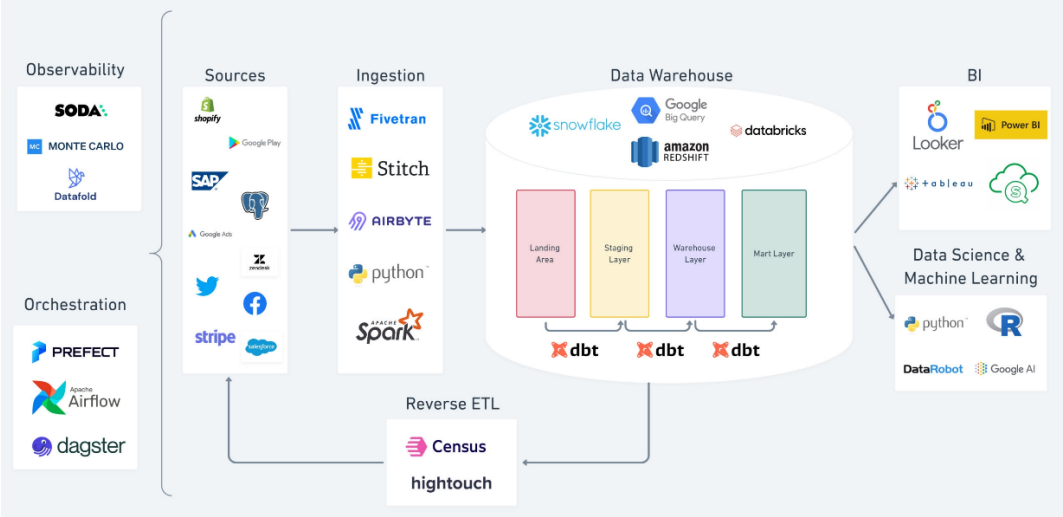

Figura 5 - Exemplo MDS na era IA
Fonte: Filtenborg (2023)

2 O DESAFIO

Atualmente a arquitetura de dados das empresas atende parcialmente as demandas de negócios e faz uso de alguns componentes como apresentado na Figura 7 - Solução AS IS.

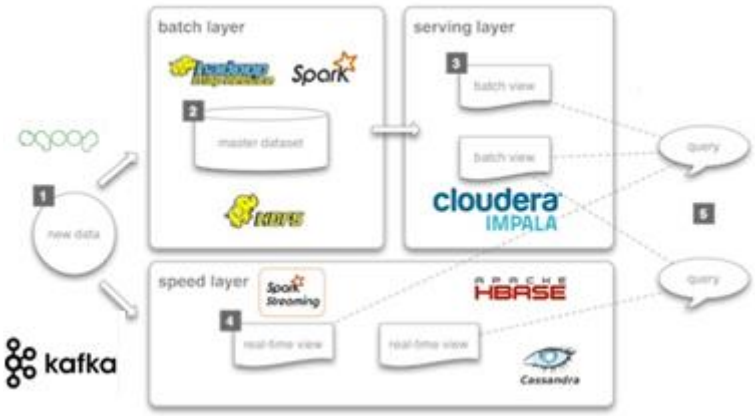

Figura 6 – Solução AS IS
Fonte: Elaborada pelos autores (2023)

As principais dores das áreas de negócios das empresas podem ser resumidas em:

- ✓ As demandas de dados não estão sendo atendidas dentro do prazo;
- ✓ As empresas ainda não conseguem colocar os modelos de *Advanced Analytics* em produção;
- ✓ Os dados não estão confiáveis, mesmo com vários processos de engenharia de dados implementados;
- ✓ Os dados não estão seguros, as informações sensíveis estão sendo acessadas pelos engenheiros de dados sem restrições;
- ✓ O departamento de TI ainda não impulsionou o *Data Driven* para as áreas de negócios;
- ✓ Modelos básicos de recomendação, sortimento de vendas, elasticidade de preços ainda não foram implementados.

A abstração de camadas de infraestrutura operacionais e integrações complexas são chaves para entrega de mais valor para as áreas de negócios. Conceitos de desacoplamento inclusive da camada central de dados a partir de soluções baseadas em *Modern Data Stack* garantem maior velocidade, eficiência e agilidade para a engenharia de dados.

Nesta jornada, a adoção de uma plataforma baseada na *Modern Data Stack* é a estratégia central.

Nosso desafio será implementar uma solução MDS considerando o contexto do Airbnb e dados abertos.

Inside Airbnb (http://insideairbnb.com/) é um site de investigação lançado por Murray Cox em 2016 para relatar e visualizar dados raspados - obtidos via crawler - na empresa de mercado de aluguel de imóveis Airbnb. Ele traz um conjunto independente de ferramentas e dados não comerciais que permite explorar como Airbnb está realmente sendo usado nas cidades pelo mundo.

Inside Airbnb fornece filtros e métricas importantes para analisar informações públicas disponíveis sobre as acomodações do Airbnb de uma cidade.

A plataforma disponibiliza um dashboard para análise dos dados por cidade.

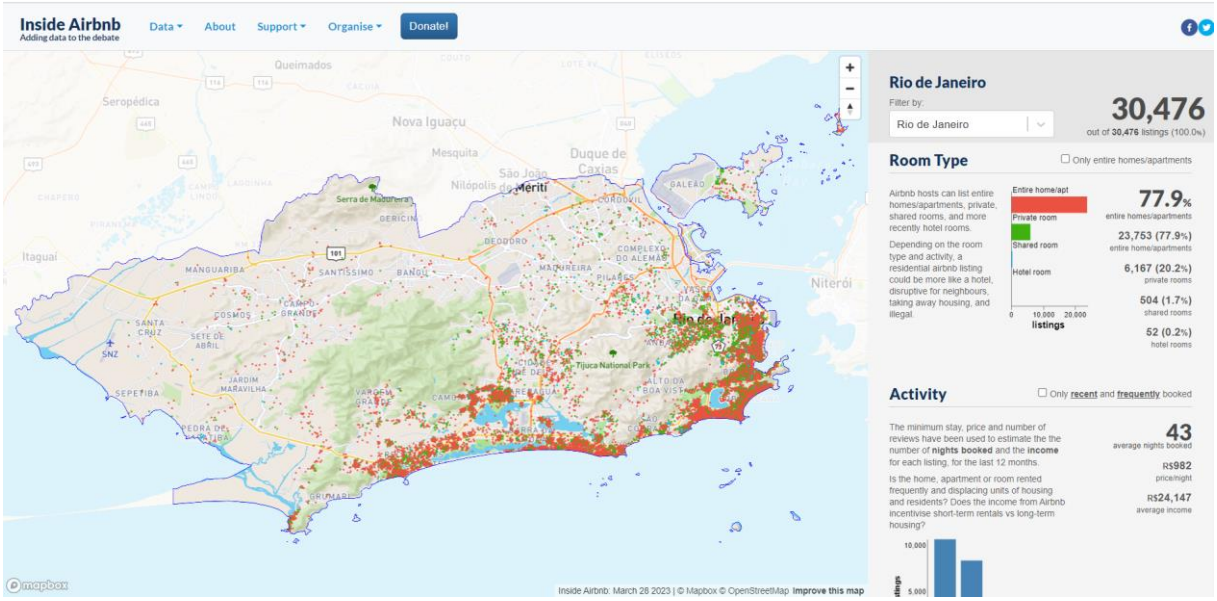

Figura 7 - Inside Airbnb - Rio Janeiro
Fonte: Inside (2023)

Para cada dataset disponibilizado existe um dicionário de dados (https://docs.google.com/spreadsheets/d/1iWCNJcSutYqpULSQHlNyGInUvHg2BoUGoNRIGa6Szc4/edit?usp=sharing) para facilitar a exploração e análise de dados. Os datasets são listings.csv, reviews.csv, calendar.csv e neighbourhoods.csv. Os dados sobre Rio de Janeiro de março de 2023 podem ser obtidos diretamente na plataforma (http://insideairbnb.com/get-the-data).

Algumas perguntas possíveis sobre os dados disponibilizados do Airbnb, em qualquer bairro ou em toda a cidade, são:

a) Quantas acomodações existem num bairro e onde ficam?

b) Quantas casas e apartamentos estão sendo alugados com frequência para turistas e não para residentes de longa duração?

c) Quanto os anfitriões ganham com aluguel para turistas?

d) Quais hosts estão administrando uma empresa com várias listagens e onde estão?

e) Que tipo de acomodação é mais comum no Airbnb numa localidade?

f) Qual é a diferença de preço entre os diferentes tipos de acomodações?

g) Quais são as regiões mais caras para ficar?

E pode-se criar um modelo analítico para escolher o melhor bairro para compra de um imóvel pensando num retorno com sua locação?

Pode ser necessário obter dados externos como do Zap Imóveis para confrontar os dados imobiliários de mercado. Uma biblioteca já foi construída com este objetivo (https://pypi.org/project/zapimoveis-scraper/).

O desafio será construir uma plataforma de dados moderna (MDS) na nuvem e realizar integrações em real time para compor a solução. Para isso, vamos considerar as bases de dados do Inside Airbnb para o Rio de Janeiro e dados imobiliários sobre imóveis na cidade.

A expectativa é uma arquitetura completa para a ingestão, transformação e análise de datasets com governança de dados e análise de dados; e uma solução que seja robusta a ponto de ser implementada. Sendo assim, levar em consideração fatores como atualização dos dados de forma constante, manutenção, escala e funcionalidade do projeto.

Requisitos Técnicos:

a) No mínimo uma ferramenta utilizada para ingestão de dados;

b) Uma camada de Storage, preferencialmente um *Cloud Data Warehouse*;

c) Uma camada de transformação diretamente no *Storage*;

d) No mínimo um dashboard ou report analisando os dados transformados;

e) No mínimo um modelo analítico envolvendo técnicas de aprendizado de máquina;

f) Uma solução para governança e catálogo de dados.

Caso seja necessária uma ferramenta de orquestração (Airflow, Prefect, entre outras), não é necessário subir clusters na nuvem para a conclusão do projeto, o desenvolvimento das DAG's ou qualquer tipo de agendamento pode ser feito local para gerar as evidências mesmo que apenas para uma única execução.

Exemplo: Um script Python foi utilizado para extrair e carregar os dados do Mongo para o DW. A evidência foi feita utilizando a execução em uma máquina local, mas para garantir a execução recorrente, a implementação final necessitará de um agendamento (ex.: Airflow). Os scripts ou comandos para o agendamento/orquestração deverão estar presentes e mencionados na arquitetura projetada.

Assim como para orquestração, dependendo da solução de governança selecionada, não é necessário provisionar um cluster, apenas exemplos desenvolvidos localmente são suficientes.

3 ENTREGÁVEIS

1. Arquitetura projetada

   Definir as camadas e tecnologias escolhidas considerando os conceitos da *Modern Data Stack.* A camada fast, para resolver os problemas de dados real time, **não será cobrada em uma fase de implementação**, mas será avaliada na arquitetura projetada.

Entregar um comparativo baseado em critérios técnicos, negócio e custos para cada tecnologia escolhida.

Garantir que o agendamento/orquestração dos processos, a governança dos dados e a observabilidade do pipeline estejam presentes na arquitetura caso necessários para a solução.

Diagramas são muito bem-vindos para a visualização da arquitetura de forma eficiente, é importante identificar as possíveis tecnologias do mercado para cada camada da arquitetura e selecionar a que mais se adequa aos requisitos do caso de uso trazendo o melhor custo x benefício para a solução.

Entregável 1

a) Detalhes da arquitetura projetada.

b) Comparativo dos cenários baseado em critérios técnicos, negócios e custos.

Formato: (ppt ou doc)

2. Escolha dos datasets

a) Baixar datasets adicionais para enriquecer as análises e implementar o modelo preditivo para identificar as melhores regiões para compra de imóveis com objetivo de retorno com aluguel;

b) Analisar os datasets e construir reports/dashboards para responder as perguntas sobre o airbnb e/ou visualizar os resultados do modelo.

Entregável 2

a) Detalhes dos datasets escolhidos – url, metadados, data, tamanho, entre outros.

b) Análises, relatórios ou dashboards para o negócio respondendo as perguntas relacionando dados do Airbnb correlacionando com data set obtido sobre imóveis.

Formato: (ppt, doc, pdf ou markdown (caso entregue diretamente no github))

3. Implementar a arquitetura:

   a) Criar um *Cloud Data Warehouse* como Snowflake, Redshift, BigQuery ou Databricks Delta Lake;

   b) Utilizar um Data Integration Service, como Fivetran, Stitch ou Airbyte para ingestão dos datasets selecionados;

   c) Transformar os dados com uma ferramenta de transformação, como dbt ou dataform;

   Criar os metadados numa ferramenta de data catalog;

   Criar relatórios ou *dashboards* em uma ferramenta de BI que responda as perguntas das áreas de negócio, como Tableau, Power BI ou Metabase;

Entregável 3

a) Detalhes da instalação e configuração realizada para implementação da arquitetura sugerida para solução.

b) Script(s)/comandos utilizados na transformação de dados, modelo(s) de Machine Learning e qualquer outra etapa adicional/opcional à arquitetura.

c) Definição detalhada da modelagem de dados.

Formato: (ppt, doc ou pdf, e repositório e repositório no github com os scripts utilizados)

REFERÊNCIAS

AIRBNB. **Media assets: 2022 Summer Release**. 2022. Disponível em: <https://news.airbnb.com/media-assets/category/2022-summer-release/> Acesso em: 20 jun. 2023.

AIRBNB. **Sobre Nós**. 2023. Disponível em: <https://news.airbnb.com/br/about-us/> Acesso em: 20 jun. 2023.

ATLAN. **What Is Modern Data Stack: History, Components, Platforms, and the Future**. 2022. Disponível em: <https://atlan.com/modern-data-stack-101/> Acesso em: 20 jun. 2023.

CONTINUAL, T. **The Modern Data Stack Ecosystem - Fall 2021 Edition.** 2021. Disponível em: <https://continual.ai/post/the-modern-data-stack-ecosystem-fall-2021-edition> Acesso em: 20 jun. 2023.

FILTENBORG, M. **A trend dissected: The modern data stack**. 2023. Disponível em: <https://bitestreams.com/blog/modern_data_stack/> Acesso em: 20 jun. 2023.

HEIDMANN, L. **Demystifying the Modern Data Stack**. 2021.Disponível em: <https://blog.dataiku.com/demystifying-the-modern-data-stack> Acesso em: 20 jun. 2023.

KAARNE, Jenni. **Supporting specialists in applying self-service analytics tools in sales analysis**. 2020. Disponível em: <https://aaltodoc.aalto.fi/handle/123456789/44992>. Acesso em: 20 jun. 2023.

MAKSYMIUK, V. **How to Build a Modern Data Stack in 2022**. 2022. Disponível em: <https://blog.skyvia.com/modern-data-stack/> Acesso em: 20 jun. 2023.

MILNER, T. **Wardley mapping the Modern Data Stack**. 2022. Disponível em: https://dev.to/aws-builders/wardley-mapping-the-modern-data-stack-1h6g Acesso em: 20 jun. 2023.

PRUKALPA. **The Future of the Modern Data Stack in 2022**. 2022. Disponível em: https://towardsdatascience.com/the-future-of-the-modern-data-stack-in-2022-4f4c91bb778f> Acesso em: 20 jun. 2023.

RAHAYU, S., HATI, H., EZNI, T. , ARGA HANANTO, A., YULIATI, E**. A decade of systematic literature review on Airbnb: the sharing economy from a multiple stakeholder perspective**, Heliyon, VoL. 7, n. 10, 2021, doi: https://doi.org/10.1016/j.heliyon.2021.e08222.

SHARMA, Abhishek. **Why You Need Agile Business Intelligence and Data Analytics.** 2020. Disponível em: <https://dzone.com/articles/what-is-agile-business-intelligence-and-data-analy> Acesso em: 20 jun. 2023.

VOLZ, J. **The Future of the Modern Data Stack**. 2021. Disponível em: https://continual.ai/post/the-future-of-the-modern-data-stack Acesso em: 20 jun. 2023.