Ferramentas adequadas para Arquitetura do Sistema de Análise de Dados no Ambiente Acadêmico

Descrevendo as camadas da Arquitetura do Sistema de Análise de Dados no Ambiente Acadêmico

A modelagem da Arquitetura do Sistema de Análise de Dados no Ambiente Acadêmico é dividida em 4 macro áreas, sendo elas: extração de dados, transformação dos dados, data warehouse e business intelligence. Cada parte dessa arquitetura necessita de ferramentas que apoiem o desenvolvimento do processo. Foram feitas buscas e análises sobre ferramentas em cada uma das áreas e foram escolhidas algumas para cada camada da arquitetura.

- **Extração de Dados:**

  Essa camada é aquela que irá conter as fontes de dados relacionadas ao projeto, além das possíveis ferramentas de apoio. Nesta primeira parte pode-se encontrar Bancos de Dados Relacionais e Não-Relacionais, arquivos CSV, XLSX, além de fontes de dados Externas e API's.

- **Transformação dos Dados:**

  Essa camada será responsável por pegar os dados extraídos através da primeira camada e aplicar processos de limpeza de dados, padronização, enriquecimento e transformação necessários para garantir a qualidade e consistência dos dados para a próxima camada. Essa camada também irá remover duplicações, preencher campos ausentes, converter os formatos, dentre vários outros pontos necessários para a arquitetura.

- **Data Warehouse:**

  Essa camada será responsável por centralizar os dados extraídos de várias fontes em um único local, organizando os dados para que se possa fazer consultas complexas e para que a última camada da arquitetura seja aplicada. Os dados serão organizados em modelos dimensionais, podendo utilizar esquemas do modelo estrela ou floco de neve.

- **Business Intelligence:**

  Essa camada será a responsável por gerenciar a visualização das informações e para isso buscou-se analisar várias ferramentas relacionadas a visualização dos dados, geração dos relatórios, entre outros. Dessa forma, dividimos as ferramentas em 2 principais grupos, definidos como ferramentas convencionais de BI e ferramentas orientadas à contexto.

Ferramentas Convencionais: as ferramentas convencionais são aquelas já amplamente utilizadas e validadas no mercado, que até utilizam inteligência artificial em alguns pontos, mas que não permitem interação através de linguagem natural, como Microsoft Power BI, Google Data Studio, Tableau, entre outros.

Por conta das características particulares do projeto, as ferramentas devem atender contextos gerais e contextos específicos de análise - que em grande parte das ferramentas atuais de Business Intelligence conseguem atender - e também identificou-se que uma das formas de se atingir os objetivos seria através de ferramentas orientadas a contexto e comandadas a partir de chats.

Ferramentas orientadas ao contexto: algumas ferramentas se definem como orientadas ao contexto, mas não são eficientes nesse aspecto. Uma das soluções encontradas foi a utilização do Data Dashboard da LenioLabs. Em vista dos fatos mencionados, das limitações das ferramentas atuais, pondera-se em desenvolver uma própria ferramenta que faça essa interpretação de linguagem natural (atendendo aos tipos de escopo abertos e fechados), gere relatórios e possibilite a interação gráfica com elementos.

Assim, a sugestão proposta é que se crie uma ferramenta que atenda a todas essas necessidades, levando-se em conta o cenário ideal e a necessidade do uso de ferramentas orientadas ao contexto.

Em contrapartida, se a prioridade for a interação através de linguagem natural, a sugestão é que se utilize a Data Dashboard LenioLabs.

Por fim, se o foco for em análises convencionais de BI com o uso de consultas em SQL puro ou através da interface gráfica, as ferramentas Microsoft Power BI, Google Data Studio e MetaBase conseguem atender aos requisitos encontrados.

Ferramentas e fluxo de dados para validação da Arquitetura do Sistema de Análise de Dados no Ambiente Acadêmico

Uma arquitetura de BI com data warehouse pode ser dividida em diferentes camadas, cada uma com seu propósito específico. Vamos detalhar as principais ferramentas utilizadas em cada uma dessas camadas:

Camada de Extração, Transformação e Carregamento (ETL): para essa camada podemos utilizar ferramentas que englobam todo o fluxo de dados e os pormenores de mover e processar dados, assim como definí-las em partes e o criarmos por conta própria a estrutura de extração, transformação e carregamento dos dados.

Não necessariamente, embora Python seja uma escolha viável para codificação de tarefas ETL, existe a obrigatoriedade de os desenvolvedores usarem Python e podem recorrer a outras linguagens de programação para extração e carregamento de dados.

- *Python para ETL:* pode assumir várias formas, dependendo dos requisitos técnicos, dos objetivos de negócios, das bibliotecas com as quais as ferramentas existentes são compatíveis e do quanto os desenvolvedores sentem que precisam trabalhar do zero. Dessa maneira, no contexto do ETL, devemos estruturar o gerenciamento de fluxo de trabalho organiza atividades de engenharia e manutenção, e os aplicativos de fluxo de trabalho também podem automatizar as próprias tarefas de ETL (i.e. As principais ferramentas de gerenciamento de fluxo de trabalho mais populares são Airflow e Luigi) e a questão de mover e processar dados (ie.. Que pode acessar bibliotecas que extraem, processam e transportam dados, como o pandas, Beautiful Soup e Odo).
- *Informatica PowerCenter*: uma ferramenta líder no mercado de integração de dados, projetada para extrair dados brutos de várias fontes, transformá-los de acordo com as regras de negócio e carregá-los no data warehouse.
- *Pentaho Data Integration*: uma poderosa ferramenta de ETL Open Source que faz parte da suíte de ferramentas Pentaho, uma plataforma de análise de código aberto que oferece recursos abrangentes de BI (Business Intelligence), ETL e mineração de dados. Também conta com uma ampla variedade de conectores, transformação de dados avançada e agendamento e orquestração de tarefas.

Camada de Armazenamento de Dados: nesta camada, um data warehouse é projetado para dar suporte a uma ferramenta de Business Intelligence (BI) e consiste em um conjunto de tecnologias, processos e estruturas de armazenamento de dados, com o objetivo de coletar, organizar, integrar e disponibilizar informações para análise e tomada de decisão.

É importante lembrar que a configuração específica de um data warehouse e a integração com uma ferramenta de BI podem variar de acordo com as necessidades e requisitos da organização, bem como as tecnologias e ferramentas utilizadas.

- *Oracle Database*: É um SGBD robusto e escalável, frequentemente utilizado como a solução de armazenamento de dados em um data warehouse. Ele oferece recursos avançados para armazenar, organizar e consultar grandes volumes de dados.
- *Microsoft SQL Server*: Assim como o Oracle Database, o Microsoft SQL Server também é amplamente utilizado como plataforma de armazenamento de dados em um data warehouse.

Camada de Transformação e Modelagem de Dados:

- *Apache Hadoop*: O Apache Hadoop é um framework open-source que permite o armazenamento e processamento distribuído de grandes volumes de dados. É comumente utilizado para processar dados não estruturados ou semiestruturados e prepará-los para a carga no data warehouse.
- *Apache Spark*: O Apache Spark é uma ferramenta de processamento de dados em memória, que oferece recursos avançados de transformação e modelagem de dados. Ele pode ser utilizado para pré-processar e transformar os dados antes de serem carregados no data warehouse.

Camada de Análise e Visualização de Dados:

- *Tableau*: O Tableau é uma ferramenta de visualização de dados avançada, que permite criar dashboards interativos, relatórios e análises exploratórias. Ele se

integra facilmente com data warehouses e oferece recursos avançados de visualização e análise.

- *Power BI*: O Power BI, da Microsoft, é outra ferramenta popular para a criação de dashboards e relatórios interativos. Ele se conecta a várias fontes de dados, incluindo data warehouses, e oferece recursos avançados de visualização e análise.
- *Data Dashboard LenioLabs*: Além de ser orientada a contexto, e acessada através de API's, essa ferramenta também pode ser utilizada diretamente em código, já que também é uma biblioteca TypeScript. Porém, essa ferramenta possui análises gráficas e relatórios bem limitados e sem muita interação com os gráficos e informações. Além disso, ela não possui uma IA por trás de seu código e na verdade ela faz uso do ChatGPT para o processamento de linguagem natural.

É importante ressaltar que a escolha das ferramentas deve ser feita considerando as necessidades, objetivos e restrições da organização, bem como as habilidades e conhecimentos da equipe responsável pela implantação e manutenção da arquitetura de BI com data warehouse e a disponibilidade de ferramentas no mercado.