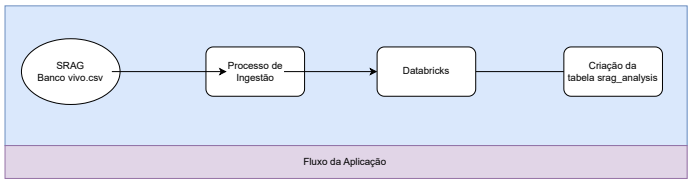

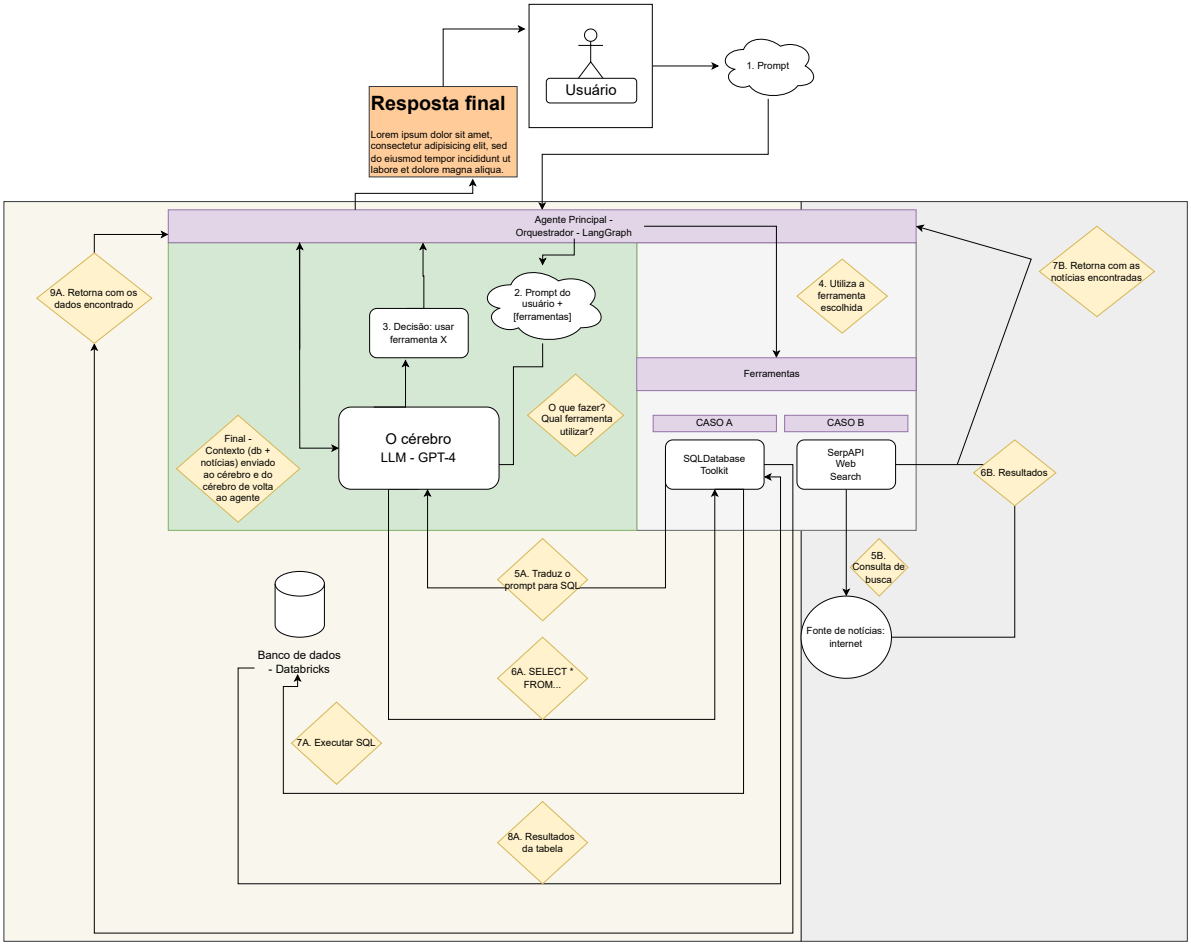

Arquitetura da Solução: Agente de IA com Ferramentas

Este documento descreve a arquitetura conceitual de um agente de IA projetado para responder a perguntas complexas, utilizando fontes de dados estruturadas (banco de dados) e não estruturadas (web).

1. Objetivo da Arquitetura

A arquitetura foi desenhada para criar um assistente autônomo capaz de orquestrar múltiplas ferramentas.

O agente recebe uma pergunta em linguagem natural, entende a intenção do usuário, seleciona a ferramenta apropriada (consulta a um banco de dados interno ou busca na web)

e sintetiza as informações coletadas para fornecer uma resposta final coesa.

2. Visão Geral dos Componentes

O sistema é dividido em duas áreas principais: um processo de preparação de dados e o fluxo da aplicação em tempo real.

A. Processo de Ingestão
Este é um pipeline de dados executado separadamente (no topo do diagrama).

- **Banco Original:** A fonte de dados bruta.
- **Processo no Databricks:** Os dados são extraídos, limpos, transformados (ETL) e carregados em uma tabela otimizada para análise.
- **Tabela p/ Análise:** O banco de dados final que o agente consultará.

B. Componentes da Aplicação (Fluxo Principal)

- **Usuário:** O iniciador do fluxo, que envia um **Prompt** (pergunta) e recebe a **Resposta Final**.
- **Agente Principal (Orquestrador):** Construído com LangGraph/ReAct. Ele gerencia o estado da conversa e decide qual passo tomar.
- **O Cérebro (LLM - GPT-4):** O modelo de linguagem grande que serve como o núcleo de inteligência. Ele é usado para três tarefas críticas:
    1. **Decisão:** Escolher qual ferramenta o agente deve usar.
    2. **Tradução:** Converter a pergunta do usuário em uma consulta SQL (para a Ferramenta A).
    3. **Geração:** Formular a resposta final com base nos dados coletados.
- **Ferramentas:**
    - **Ferramenta A (SQLDatabase Ruler):** Uma ferramenta especializada para interagir com o banco de dados Databricks.
    - **Ferramenta B (SerpAPI Search):** Uma ferramenta para realizar buscas em tempo real na internet (Fontes de Notícias).

3. Fluxo de Execução (Passo a Passo)

O fluxo de interação do usuário com o agente segue a sequência numerada no diagrama:

1. O **Usuário** envia um **Prompt** (pergunta) para o sistema.
2. O **Agente Principal** recebe o prompt e o encaminha, junto com a lista de ferramentas disponíveis, para **O Cérebro (LLM)**.
3. O **LLM** analisa a intenção da pergunta e decide qual ferramenta é a mais adequada para respondê-la, informando sua decisão ao Agente.
4. O **Agente** invoca a ferramenta selecionada pelo LLM.

Cenário A: Consulta ao Banco de Dados (Ferramenta A)

5A. A **Ferramenta A** é ativada. Ela envia a pergunta do usuário para o **LLM (Cérebro)** com a instrução de traduzi-la.
6A. O **LLM** gera uma consulta **SQL** precisa e a retorna para a Tool A.
7A. A **Ferramenta A** executa a consulta SQL diretamente no **Banco de Dados (Databricks)**.
8A. O Banco de Dados retorna os **Resultados da Tabela** (dados).
9A. A **Ferramenta A** envia os dados encontrados de volta ao Agente como uma "Observação".

Cenário B: Busca na Web (TOOL B)

B. A **Ferramenta B** é ativada.
5B. A ferramenta executa uma consulta na **SerpAPI** para buscar informações nas **Fontes de Notícias (Internet)**.
6B. Os **Resultados** da busca são retornados para a Ferramenta B.
7B. A **Ferramenta B** envia os links e snippets encontrados de volta ao Agente como uma "Observação".

Conclusão do Fluxo

9. **Geração da Resposta:** O **Agente Principal** coleta todas as "Observações" (sejam os dados do banco ou os resultados da web) e as envia como **Contexto Final** para **O Cérebro (LLM)**.
10. O **LLM** sintetiza todas as informações de contexto e gera uma **Resposta Final** coesa e em linguagem natural para o usuário.