Prompt para Desenvolvimento de Sistema de Integração de Dados Públicos (IA Generativa)

Objetivo Geral

Desenvolver um sistema robusto e escalável para integração, cruzamento e análise de múltiplas bases de dados públicas brasileiras, com foco em apoiar a formulação e avaliação de políticas públicas e a gestão governamental. O sistema deve permitir que usuários não técnicos realizem pesquisas complexas e visualizem insights através de dashboards interativos.

Contexto e Importância

Atualmente, dados governamentais estão dispersos em diversas fontes, dificultando a análise integrada e a tomada de decisões baseada em evidências. Este projeto visa centralizar essas informações, padronizá-las e torná-las acessíveis para análises aprofundadas, promovendo maior transparência, eficiência e eficácia na gestão pública.

Fontes de Dados Iniciais (MVP)

O projeto deve começar com a integração das seguintes bases de dados públicas brasileiras, com potencial para expansão futura:

- **IBGE (SIDRA):** Dados demográficos, socioeconômicos, geográficos (população, PIB, IDH, malhas territoriais).
- **CGU (Portal da Transparência):** Gastos federais, convênios, servidores, empresas sancionadas.
- **DATASUS (OpenDataSUS):** Dados de saúde (internações, óbitos, nascidos vivos, vacinação).
- **INEP (Censo Escolar):** Dados da educação básica e superior (matrículas, IDEB).

Arquitetura de Dados Proposta (Data Lakehouse)

A arquitetura deve seguir o modelo Data Lakehouse, estruturada em camadas para garantir flexibilidade, qualidade e performance:

1. **Camada Bruta (Bronze/Landing):** Armazenamento dos dados originais, sem transformações, diretamente das fontes. Deve ser imutável e servir como histórico.

2. **Camada Tratada (Silver):** Limpeza, padronização e enriquecimento dos dados. Crucial para esta camada é a **padronização de chaves de cruzamento**, como o código de município de 7 dígitos do IBGE, para permitir a integração entre diferentes bases.
3. **Camada Validada (Gold/Analytics):** Dados agregados e modelados, otimizados para consumo por ferramentas de BI e análise, com métricas e indicadores pré-calculados.

Pilha Tecnológica Recomendada

Para a implementação, sugere-se a seguinte pilha tecnológica, priorizando soluções open-source e escaláveis:

- **Orquestração e Ingestão (ETL/ELT):**
  - **Airbyte:** Para extração e carregamento de dados de diversas fontes (APIs, arquivos, bancos de dados).
  - **Apache Airflow ou Prefect:** Para orquestração e agendamento de pipelines de dados.
- **Armazenamento e Processamento:**
  - **PostgreSQL com PostGIS:** Para armazenamento de dados relacionais e suporte a análises geoespaciais (essencial para dados públicos brasileiros).
  - **ClickHouse (opcional):** Para cenários que exijam consultas analíticas de alta performance em grandes volumes de dados.
- **Transformação e Modelagem:**
  - **dbt (data build tool):** Para transformar e modelar dados na camada Silver e Gold, utilizando SQL, com versionamento de código (GitHub) e testes automatizados.
- **Visualização e Exploração:**
  - **Metabase ou Apache Superset:** Para criação de dashboards interativos e exploração de dados por usuários finais.
  - **Streamlit (opcional):** Para desenvolvimento de aplicações de dados personalizadas e simuladores.

Funcionalidades Desejadas

O sistema deve oferecer as seguintes funcionalidades:

- **Coleta Automatizada:** Pipelines automatizados para extração e atualização periódica dos dados das fontes públicas.
- **Padronização e Qualidade de Dados:** Mecanismos para limpeza, padronização e validação dos dados, garantindo a integridade e consistência.

- **Cruzamento de Dados:** Capacidade de cruzar informações entre diferentes bases utilizando chaves comuns (ex: código de município, CPF/CNPJ anonimizado).
- **Interface de Pesquisa:** Uma interface intuitiva que permita aos usuários realizar pesquisas complexas, aplicando filtros e agrupamentos.
- **Dashboards Interativos:** Geração de dashboards e relatórios visuais que apresentem os principais indicadores e insights para políticas públicas.
- **Exportação de Dados:** Funcionalidade para exportar dados e resultados de análises em formatos comuns (CSV, Excel).

Roadmap de Implementação (MVP)

O desenvolvimento deve seguir uma abordagem iterativa, começando com um MVP:

1. **Fase 1: MVP de Saúde e Demografia:** Integrar DATASUS e IBGE (população e malhas territoriais) para analisar indicadores de saúde por município.
2. **Fase 2: Expansão para Educação:** Adicionar dados do INEP para cruzar com demografia e saúde, avaliando o impacto de políticas educacionais.
3. **Fase 3: Transparência e Finanças:** Incluir dados do Portal da Transparência para análise de gastos e convênios por município/estado.

Requisitos Não Funcionais

- **Escalabilidade:** A arquitetura deve ser capaz de lidar com o crescimento do volume de dados e o aumento do número de fontes.
- **Segurança:** Implementação de boas práticas de segurança para acesso aos dados e ao sistema.
- **Qualidade de Dados:** Mecanismos de monitoramento e alerta para problemas de qualidade de dados.
- **Documentação:** Documentação completa da arquitetura, pipelines de dados, modelos e uso do sistema.

Entrega Esperada

O resultado esperado deste prompt é um plano de projeto detalhado, incluindo:

- **Código-fonte:** Scripts Python para ETL, modelos dbt, configurações de orquestração (Airflow/Prefect).
- **Esquemas de Banco de Dados:** Definições das tabelas e relacionamentos no PostgreSQL/ClickHouse.

- **Configurações de BI:** Modelos de dashboards de exemplo no Metabase/Superset.
- **Documentação Técnica:** Detalhamento da arquitetura, instruções de setup e deploy, guia de uso.
- **Estimativa de Recursos:** Requisitos de infraestrutura e tempo de desenvolvimento.

Este prompt fornece uma base sólida para que uma IA generativa possa auxiliar no desenvolvimento de um sistema complexo e de alto impacto para a gestão pública.