Guia para Integração de Bases de Dados Públicas: Foco em Políticas e Gestão Pública

A criação de um sistema centralizado para integração de bases de dados públicas é um passo fundamental para a modernização da gestão pública e a formulação de políticas baseadas em evidências. Este documento apresenta uma estratégia estruturada para iniciar esse projeto, detalhando as fontes de dados essenciais, a arquitetura tecnológica recomendada e um roteiro de implementação.

1. Identificação das Fontes de Dados Estratégicas

Para um sistema voltado à gestão pública, a escolha das bases deve priorizar aquelas que oferecem maior granularidade e potencial de cruzamento. A tabela abaixo destaca as principais fontes brasileiras:

Órgão / Fonte	Principais Dados Disponíveis	Utilidade na Gestão Pública
IBGE (SIDRA/Censo)	População, PIB, IDH, malhas territoriais	Base para denominadores e recortes geográficos.
CGU (Portal Transparência)	Gastos, convênios, servidores, sanções	Fiscalização, auditoria e controle orçamentário.
DATASUS (OpenDataSUS)	Internações, óbitos, vacinação, nascidos	Planejamento de saúde e vigilância epidemiológica.
INEP (Censo Escolar)	Matrículas, infraestrutura, IDEB	Avaliação da qualidade e cobertura educacional.
MTE (RAIS/CAGED)	Emprego formal, renda, ocupações	Monitoramento do mercado de trabalho local.
IPEA (IpeaData)	Séries históricas socioeconômicas	Análise de tendências de longo prazo.

2. Arquitetura de Referência: O Data Lakehouse

A arquitetura mais moderna e eficiente para este tipo de projeto é o **Data Lakehouse**, que combina a flexibilidade de armazenamento de um Data Lake com a performance de consulta de um Data Warehouse. Recomenda-se a estruturação em três camadas principais:

1. **Camada Bruta (Bronze/Landing):** Armazena os arquivos originais (CSV, JSON, XML) exatamente como foram extraídos das fontes oficiais. Serve como backup histórico e permite reprocessar dados se as regras de negócio mudarem.
2. **Camada Tratada (Silver):** Onde ocorre a limpeza e padronização. O passo mais crítico aqui é a **unificação das chaves de cruzamento**, como o código de município de 7 dígitos do IBGE, que permite "colar" dados do DATASUS com dados do INEP, por exemplo.
3. **Camada Validada (Gold/Analytics):** Contém tabelas prontas para consumo, com métricas calculadas e agregações que facilitam a criação de dashboards e relatórios.

3. Pilha Tecnológica Recomendada

Para garantir escalabilidade e baixo custo inicial, sugere-se a seguinte combinação de ferramentas:

- **Ingestão e Orquestração:** Ferramentas como **Airbyte** (open-source) facilitam a extração de APIs e arquivos. O **Apache Airflow** ou **Prefect** são ideais para agendar as atualizações (ex: baixar novos dados do CAGED todo mês).
- **Armazenamento e Processamento:** O **BigQuery (Google Cloud)** é uma excelente escolha por ser "serverless" e cobrar apenas pelo uso. Para soluções locais, o **PostgreSQL** com a extensão **PostGIS** é indispensável para análises espaciais e geográficas.
- **Transformação de Dados:** O **dbt (data build tool)** é a ferramenta padrão para transformar dados usando SQL, permitindo versionamento no GitHub e testes automáticos de qualidade.
- **Visualização:** O **Metabase** oferece uma interface amigável para gestores que não dominam SQL, enquanto o **Apache Superset** é mais robusto para visualizações complexas.

4. Roadmap de Implementação (Passo a Passo)

> "O segredo para o sucesso de projetos de dados complexos é começar pequeno, mas com uma base sólida e escalável."

1. **Definição do MVP (Mínimo Produto Viável):** Escolha um tema prioritário (ex: "Eficiência da Saúde Primária") e integre apenas duas bases (ex: DATASUS + IBGE).
2. **Padronização Geográfica:** Crie uma tabela mestre de municípios e estados para garantir que todos os cruzamentos futuros funcionem sem erros de grafia ou códigos divergentes.

3. **Automação da Primeira Base:** Configure um pipeline automático para uma base simples (ex: Portal da Transparência) para testar o fluxo completo da extração à visualização.
4. **Governança e Documentação:** Registre a origem de cada dado, a data da última atualização e o dicionário de variáveis. Em gestão pública, a transparência do método é tão importante quanto o resultado.

Referências

[1] Base dos Dados. Arquitetura de um data lakehouse Escalável: Case Real com GCP, GitHub e Metabase. Disponível em: https://basedosdados.org/blog/artigo-arquitetura-de-um-data-lakehouse-escalavel. Acesso em: 27 fev. 2026.