Brainstorm Projeto Big Data

🗓️ **Prazo Final:**

10 de dez. de 2025

👥 **Participantes**

- @Rafaela Biaze RA: 6324518
- @Alison RA: 6324005
- @Daniele Fagundes Ra: 6324661
- @Ronaldo Canavezzi RA: 6324536
- @Giovanna Sabino RA: 6324089

Projeto de Big Data: Correlação entre Transtornos de Humor e Suicídio

1. **Visão Geral do Projeto**

Problema:

Existe uma lacuna no entendimento de como a incidência de casos graves de transtorno de humor (refletidos em internações ou prevalência) se correlaciona geograficamente com as taxas de mortalidade por suicídio. Especificamente, o projeto visa investigar por que essa correlação não é homogênea, já que alguns países com alta taxa de suicídio apresentam baixa taxa de depressão registrada.

Hipótese de Trabalho (Atualizada):

"Nem todo país que tem uma taxa de suicídio alta possui uma taxa de depressão alta em seus registros; porém, a depressão não deixa de ser o transtorno de humor predominante que mais causa mortalidade por suicídio em escala global."

Objetivo Principal:

Desenvolver uma solução completa de dados (end-to-end) capaz de ingerir dados globais da OMS, processá-los para garantir qualidade e consistência, e apresentar visualizações que identifiquem tanto a correlação predominante quanto os países outliers (fora da curva), confirmando a tese proposta.

Justificativa Técnica:

O projeto demonstrará a aplicação prática dos 5 Vs do Big Data (especialmente Variedade e Veracidade – lidando com a subnotificação de dados), utilizando uma arquitetura moderna de Lakehouse para tratamento de dados brutos via API.

---

2. **Escopo e Fontes de Dados**

2.1 O que está incluído (In-Scope)

- Coleta automatizada de dados via API pública.
- Criação de Data Lake com camadas (Raw, Bronze, Silver, Gold).
- Análise exploratória e identificação de discrepâncias geográficas na correlação.

2.2 Fontes de Dados

- Origem: World Health Organization (WHO) - Global Health Observatory (GHO).
- API Endpoint: GHO OData API

Indicadores Chave:

- Suicídio: Suicide mortality rate (Código estimado: SDGSUICIDE).
- Depressão: Depressive disorders prevalence ou Hospitalization rates (Código a confirmar: ex. MH_12).

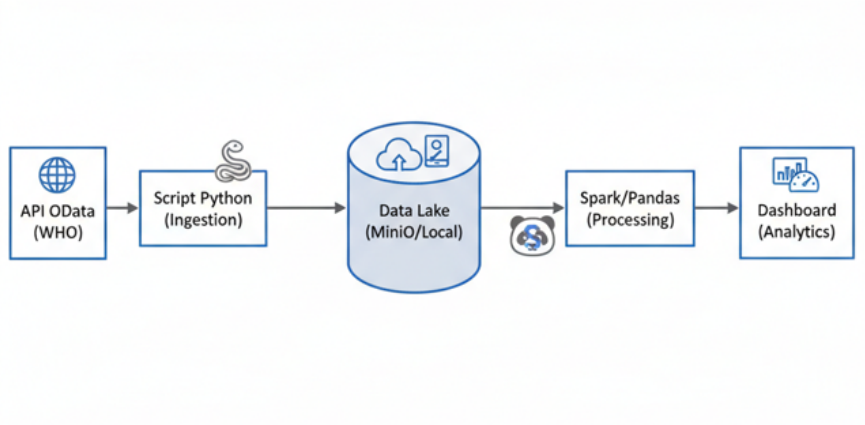
Formato: JSON (OData).

---

3. **Arquitetura da Solução**

O pipeline de dados seguirá o fluxo de ingestão, processamento, armazenamento e análise, utilizando os componentes do repositório (HDFS, Spark, Postgres).

3.1 Diagrama de Componentes



API OData (WHO) ➔ Script Python (Ingestão) ➔ Data Lake (HDFS/Spark) ➔ Spark (Processamento) ➔ PostgreSQL (Gold) ➔ Dashboard (Superset/Power BI)

3.2 Camadas de Dados (Data Lake)

Camada Raw (Bronze): Dados brutos em formato JSON, exatamente como recebidos da API.

- Objetivo: Histórico imutável e reprocessamento se necessário.

Camada Silver (Trusted): Dados limpos, convertidos para formato colunar (Parquet), com tipagem corrigida e nulos tratados.

- Transformações: Deduplicação, limpeza de colunas inúteis, padronização de nomes de países.

Camada Gold (Refined): Dados agregados prontos para análise.

- Modelo: Tabela única contendo Região | Ano | Taxa_Suicidio | Taxa_Depressao.

---

4. **Stack Tecnológico e Ferramentas**

Componente	Tecnologia	Justificativa
Ingestão	Python (requests)	Simplicidade e facilidade de conexão com APIs REST/OData.
Processamento	Apache Spark (PySpark)	Manipulação eficiente de dataframes em ambiente distribuído (HDFS).
Armazenamento	Hadoop HDFS / PostgreSQL	HDFS como Data Lake (Bronze/Silver). Postgres para a camada Gold de consumo.
Orquestração	Jupyter Notebooks	Controle do fluxo de execução ETL e ambiente de desenvolvimento.

Visualização	Power BI	Criação de dashboards interativos, idealmente com Scatter Plots para visualizar a correlação vs. discrepâncias.
Versionamento	Git / GitHub	Controle de versão do código e documentação.

---

5. **Organização da Equipe e Responsabilidades**

Engenheiro de Dados (Ingestão/Storage): [@Rafaela Biaze] - Responsável pelos scripts de coleta e estrutura de pastas do Lake (HDFS).

Engenheiro de Dados (Processamento): [@Ronaldo Canavezzi] - Responsável pela limpeza (Silver) e agregação (Gold) com Spark.

Analista de Dados: [@Giovanna] - Responsável pelas definições de regras de negócio e validação das discrepâncias.

Engenheiro de Visualização: [@Alison] - Responsável pelo Dashboard e storytelling dos dados (ênfase na nova tese).

Líder de Documentação: [@Daniele] - Garantir que o Confluence e o README estejam atualizados.

---

6. **Cronograma e Próximos Passos**

Semana 1:

1. Validar os códigos exatos dos indicadores na API da WHO.
2. Configurar o repositório Git com a estrutura de pastas (/docs, /src, /data).
3. Desenvolver script de ingestão (Camada Bronze).
4. Desenvolver script de limpeza (Camada Silver).

Semana 2:

1. Criar tabela agregada (Camada Gold) com as duas variáveis cruzadas.
2. Gerar primeira versão do Dashboard focando na visualização da tese (correlação vs. outliers).

Semana 3:

1. Revisão final da documentação.
2. Gravação/Ensaio da apresentação.

---

7. **Notas Importantes**

Risco da Tese (Veracidade): A baixa taxa de depressão em países com alto suicídio pode não significar a ausência da doença, mas sim subnotificação devido à falta de acesso a diagnóstico ou estigma cultural. O projeto deve tratar essa limitação na análise.

Disponibilidade de Dados: Caso a API da WHO não forneça dados de "internação" globalmente, utilizaremos dados de "prevalência" como proxy para transtornos de humor, mantendo a validade da análise de correlação.

Qualidade dos Dados: Países com subnotificação de suicídio ou depressão podem enviesar a análise, sendo o foco do projeto a interpretação desses vieses.