HIVision

Plataforma de Integração e Inteligência Epidemiológica

Leonardo dos Santos Silva (leonardo.ssilva2@ufrpe.br)
Gabriel Mesquita Gomes (gabriel.mesquitagomes@ufrpe.br)

Fonte dos Dados: https://datasus.saude.gov.br/transferencia-de-arquivos/
Github: https://github.com/gabMesq1812/Projeto_SAD

ETAPA 1 - PLANEJAMENTO

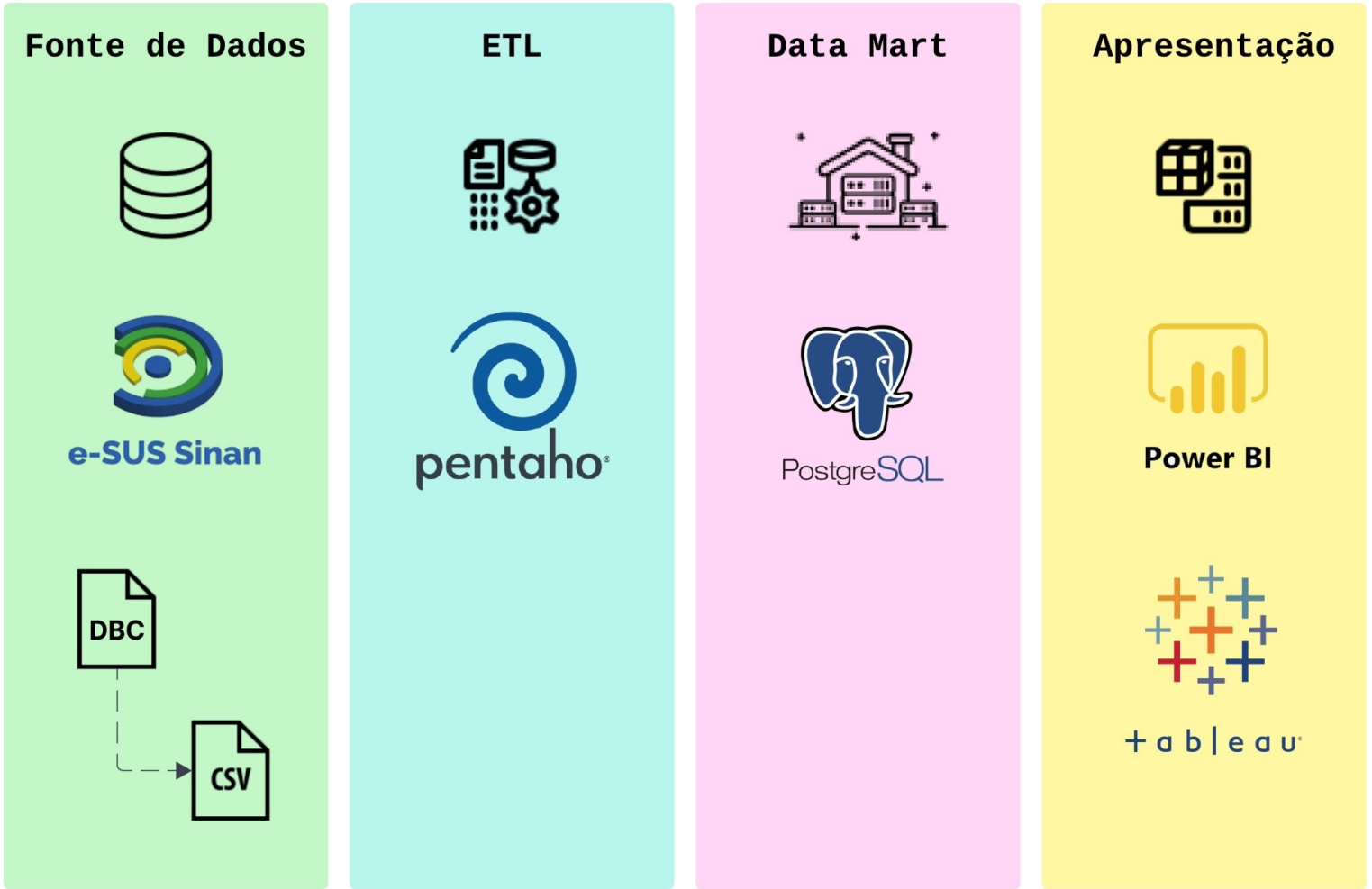
1. Contextualização

O Data Mart HIVision está inserido na área de saúde pública, com foco na vigilância epidemiológica, que tem como finalidade o monitoramento contínuo de doenças e agravos de interesse coletivo, visando identificar tendências, fatores de risco e impactos sobre a população. A análise integrada desses dados é fundamental para subsidiar políticas de prevenção, controle e tratamento, permitindo que gestores e profissionais de saúde tomem decisões baseadas em evidências e direcionem recursos de forma mais eficiente.

2. Escopo/objetivo do Data Mart

O projeto HIVision tem como objetivo desenvolver um Data Mart analítico voltado à integração, análise e visualização de dados epidemiológicos relacionados ao HIV em adultos, no período de 2007 a 2025. A iniciativa busca apoiar a tomada de decisão em políticas públicas de saúde, permitindo análises históricas, exploratórias e preditivas baseadas em informações provenientes do SINAN/DATASUS.

Arquitetura Tecnológica do Data Mart - HIVision



Disponível em:link

Fluxo do Processo de Construção do Data Mart - HIVision

4. Abordagem

Botom-up:

- mais ágil;
- facilita incrementos futuros;
- integração com outros dados do DATASUS.

Star Schema

- simplicidade;
- uma tabela fato central com dados quantitativos;
- tabelas dimensão conectadas.

5. Usuários

- Gestores de Saúde Pública;
- Analistas e Técnicos de Vigilância Epidemiológica;
- Secretárias de Saúde;
- Pesquisadores e estudantes;

ETAPA 2 - LEVANTAMENTO DAS NECESSIDADES

6. Consultas de Apoio à Decisão

- A epidemia de HIV está aumentando, diminuindo ou estabilizada em nossa região?
- Quem é o perfil mais vulnerável hoje?
- Em quais municípios devemos concentrar esforços de testagem e envio de equipes de saúde?
- O número de diagnósticos em gestantes está caindo?
- A epidemia está afetando de forma desproporcional grupos racializados ou com menor escolaridade?

7. Indicadores do HIVision

Indicador	Descrição
Casos notificados de HIV	Número total de casos de HIV confirmados e registrados no sistema de vigilância em um período específico.
Novos casos de HIV	Número de pessoas diagnosticadas com HIV pela primeira vez durante um período definido
HIV em Gestantes	Contagem de gestantes diagnosticadas com HIV.
Taxa de mortalidade por HIV/AIDS	Número de óbitos por HIV/AIDS por 100.000 habitantes em um período específico.
Pessoas em TARV	Contagem de pessoas em tratamento antirretroviral (TARV) no sistema de saúde.
Casos por sexos	Distribuição de casos de HIV por sexo: masculino e feminino.
Distribuição etária dos casos	Percentual de casos por faixa etária
Testes realizados para HIV	Número total de testes realizados, útil para medir cobertura e rastreamento da população
Tempo médio entre diagnóstico e início do tratamento	Média de dias entre o diagnóstico de HIV e o início da terapia antirretroviral.

ETAPA 3 - MODELAGEM

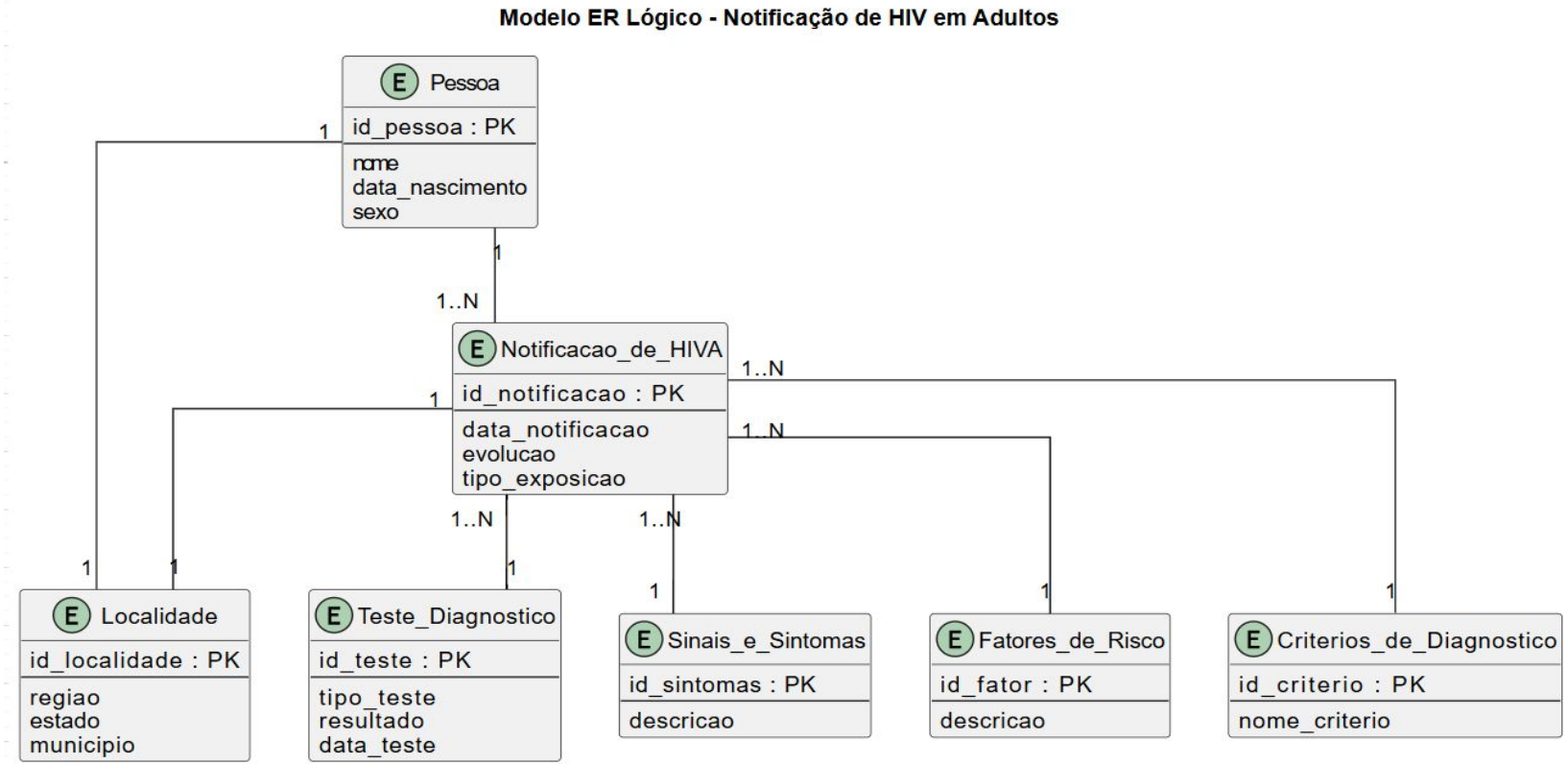
9. Modelo Relacional

9.1) Descrição das entidades e relacionamentos

Relacionamento	Entidade A	Cardinalidade A	Ação realizada por A	Entidade B	Cardinalidade B	Coluna1 no diagrama
1	Pessoa	1	Notifica	Notificação de HIV/AIDS	1..N	FK_PACIENTE (em Notificação)
2	Notificação de HIV/AIDS	1	Local Notificação	Localidade	1	FK_LOCAL_NOTIFICACAO (em Notificação)
3	Pessoa	1	Local Residência	Localidade	1	FK_LOCAL_RESIDENCIA (em Pessoa)
4	Notificação de HIV/AIDS	1..N	Ação Determinada Por	Teste Diagnóstico	1	FK_TESTES (em Notificação)
5	Notificação de HIV/AIDS	1..N	Apresenta	Sinais e Sintomas	1	FK_ANTECEDENTES (em Notificação)
6	Notificação de HIV/AIDS	1..N	Possui	Fatores de Risco	1	FK_FATOR_RISCO (em Notificação)
7	Notificação de HIV/AIDS	1..N	Classificada Por	Critérios de Diagnóstico	1	FK_CRITERIO (em Notificação)

9. Modelo Relacional

9.2) Diagrama lógico ER



10. Modelo Dimensional

A. Área de Negócios
    a. Saúde Pública (Vigilância Epidemiológica)
B. Processo
    a. Notificação de Agravos de HIV em Adultos (HIVA)
C. Granularidade
    a. Uma notificação individual de um caso de HIV em adulto.

10. Modelo Dimensional

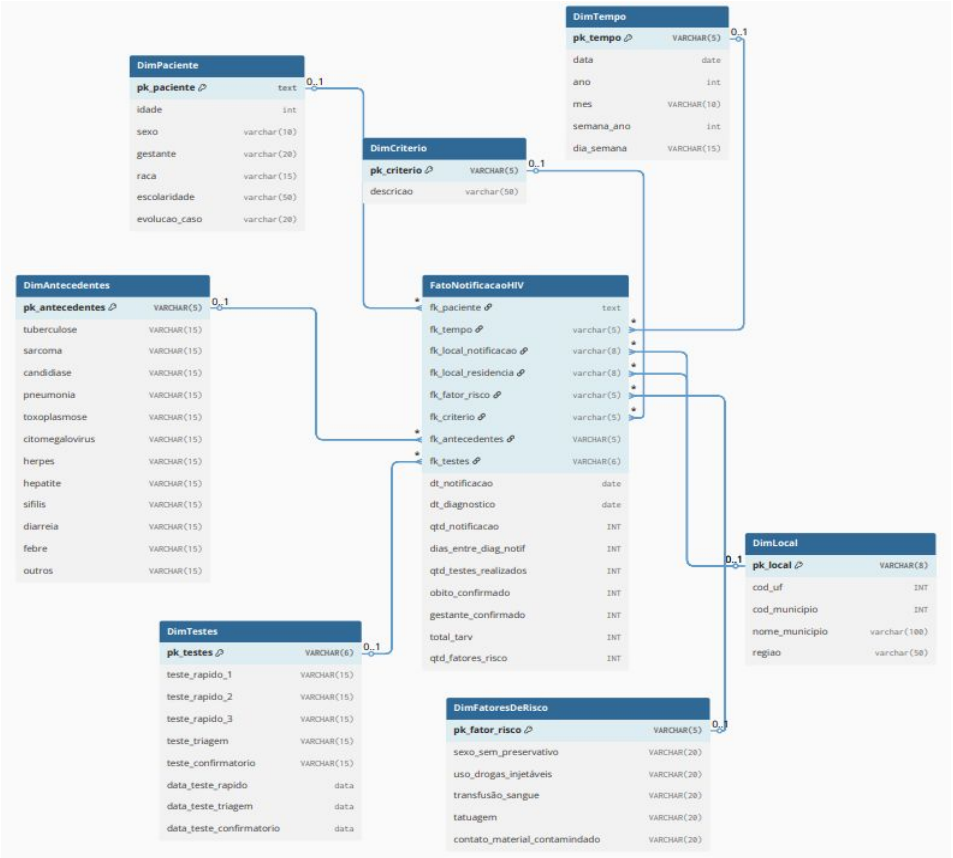
Dimensões	Atributos
DimTempo	Ano → mês → semana → dia
DimLocal	Região → UF → Municipio
DimPaciente	Idade, sexo, raca, escolaridade
DimAntecedentes	Descrição: Presença ou ausencia de outro fator
DimTestes	Descrição: Tipos de testes realizados
Dim FatorDeRisco	Descrição: Fatores de riscos para HIV
DimCriterio	Descrição: Forma de diagnóstico

10. Modelo Dimensional

Métrica	Tipo	Descrição
dt_notificacao	Não-aditiva	Data em que o caso foi notificado
dt_diagnostico	Não-aditiva	Data em que o diagnóstico foi confirmado
qtd_notificacao	Aditiva	Quantidade de notificações registradas
dias_entre_diag_notif	Semi-aditiva	Diferença em dias entre diagnóstico e notificação
qtd_testes_realizados	Aditiva	Total de testes realizados em um período ou local
obito_confirmado	Aditiva	Número de óbitos confirmados
gestante_confirmado	Aditiva	Número de gestantes confirmadas com a doença
total_tarv	Semi-aditiva	Total de pacientes em tratamento antirretroviral
qtd_fatores_risco	Semi-aditiva	Número de fatores de risco associados

10. Modelo Dimensional

Disponível em: [Star Schema]



10. Modelo Dimensional

k_pacient	fk_tempo	cal_notifi	cal_reside	_fator_ris	fk_criterio	anteceden	fk_testes	_notificac	_diagnosti	l_notifica	entre_diag	estes_reali	o_confirm	nte_confir	total_tarv	_fatores_r
PAC001	2025W01	LOC001	RES001	FR01	CR01	ANT01	TST01	2025-01-0	2025-01-0	1	2	2	0	0	3	1
PAC002	2025W02	LOC002	RES002	FR02	CR02	ANT02	TST02	2025-01-1	2025-01-0	1	2	3	0	1	4	2
PAC003	2025W02	LOC001	RES003	FR03	CR01	ANT03	TST03	2025-01-1	2025-01-1	1	2	2	0	0	5	1
PAC004	2025W03	LOC003	RES004	FR02	CR03	ANT02	TST01	2025-01-2	2025-01-1	1	2	1	1	0	4	2
PAC005	2025W03	LOC002	RES005	FR01	CR02	ANT01	TST02	2025-01-2	2025-01-2	1	2	3	0	1	5	1
PAC006	2025W04	LOC001	RES006	FR03	CR03	ANT03	TST03	2025-01-2	2025-01-2	1	2	2	0	0	3	3
PAC007	2025W05	LOC003	RES007	FR02	CR01	ANT02	TST01	2025-02-0	2025-02-0	1	2	1	1	0	6	2
PAC008	2025W05	LOC002	RES008	FR01	CR02	ANT01	TST02	2025-02-0	2025-02-0	1	2	2	0	1	4	1
PAC009	2025W06	LOC001	RES009	FR03	CR03	ANT03	TST03	2025-02-1	2025-02-0	1	2	3	0	0	5	3
PAC010	2025W06	LOC003	RES010	FR02	CR02	ANT02	TST01	2025-02-1	2025-02-1	1	2	1	1	0	6	2

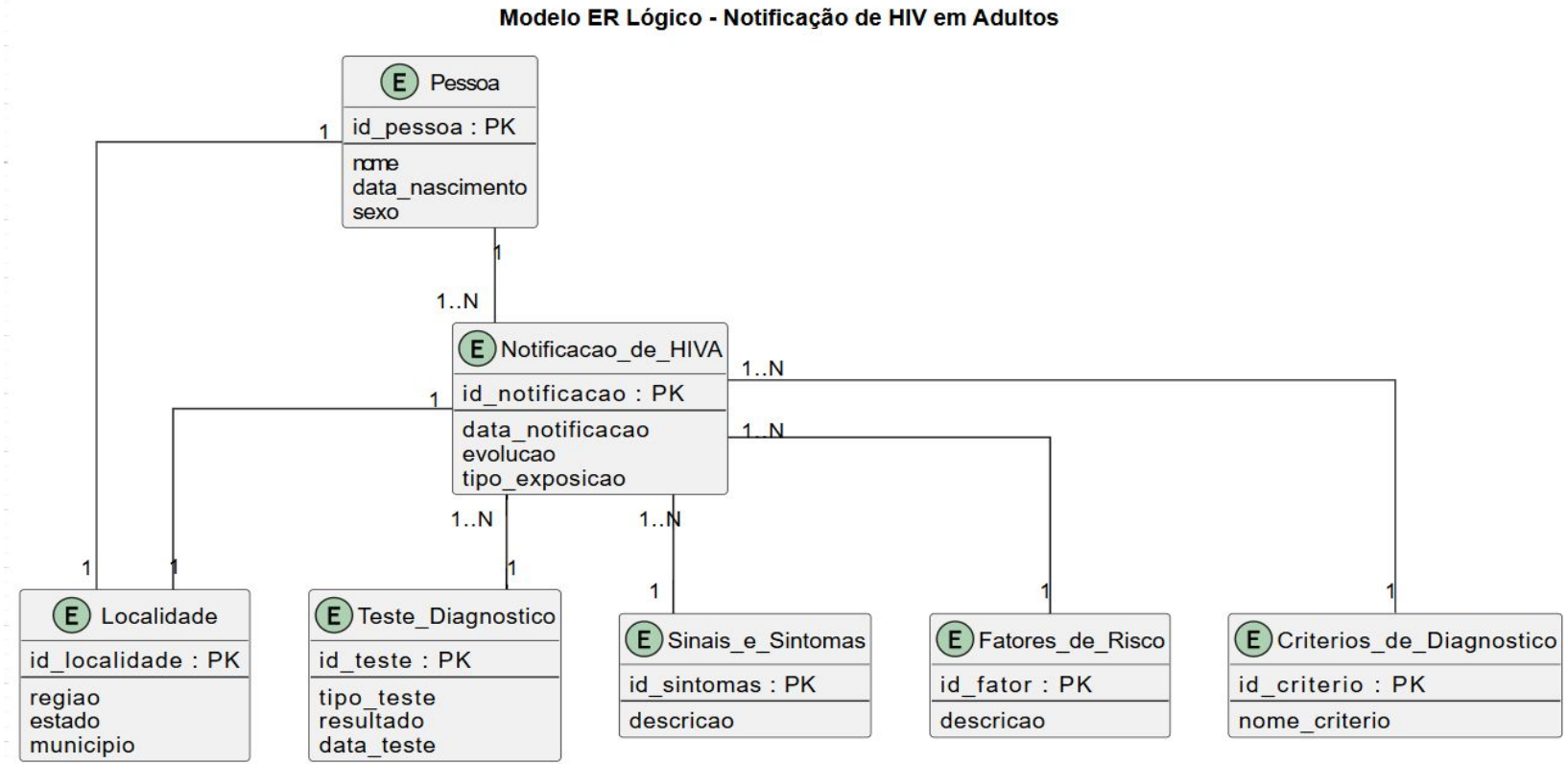
10. Modelo Dimensional do Data Mart (lógico)

H. Estimativa de espaço

(539.535 linhas)x(91 bytes)=49.097.685+20% = 58.917.222 bytes = 58,9 Mb

ETAPA 4 - PROJETO FÍSICO DO BD

11. Modelo Relacional do Data Mart (físico)



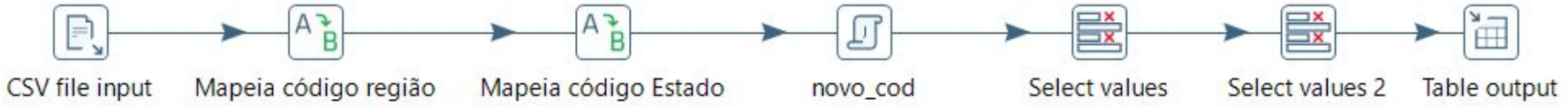
ETAPA 5 - EXTRAÇÃO, TRANSFORMAÇÃO E CARGA

12. Plano de Carga da Dimensão Tempo

Na dimensão tempo, realizei uma carga simples porém essencial. Primeiro importo o conjunto de datas e gero uma chave surrogate via Add Sequence. Em seguida, calculei atributos temporais como ano, mês e dia. Depois mapeei os nomes dos dias e meses usando Value Mapper, para facilitar análises no BI. Por fim, selecionei somente os campos necessários e carrego tudo na tabela dim_tempo no Data Mart.

13. Plano de Carga da Dimensão Localização

A dimensão Localização começa com a leitura de um CSV, disponível no site do IBGE, contendo códigos municipais e estaduais. Em seguida realizei dois mapeamentos: o primeiro converte o código da região para o nome da região, e o segundo converte os códigos de estado para suas siglas. Depois gerei um novo código que será a chave da dimensão. Finalizo ajustando e padronizando os campos com dois Select Values, e por fim insiro os dados tratados na tabela dim_localizacao do Data Warehouse.



14. Plano de Carga da Dimensão Paciente

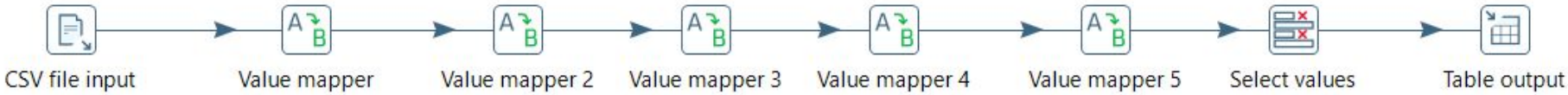
A Dimensão Paciente tem como finalidade consolidar e padronizar as informações demográficas dos indivíduos registrados no sistema de vigilância epidemiológica do SUS. Nela, começo lendo um CSV com os dados brutos das notificações registrados no sistema. Em seguida, utilizo um step para converter códigos de sexo, raça/cor e outras características em valores padronizados e legíveis. Depois filtro e organizo os campos com o Select Values, garantindo que apenas atributos importantes sejam levados para o DM. Por fim, esses dados são carregados na tabela dim_paciente.

15. Plano de Carga da Dimensão Critério

A Dimensão Critério tem como objetivo representar e padronizar os critérios de classificação utilizados no processo de notificação e confirmação dos casos registrados no sistema de vigilância epidemiológica. Essa dimensão fornece suporte analítico ao permitir segmentações baseadas no tipo de critério adotado, tal como laboratorial, clínico ou epidemiológico. A carga inicia-se com a leitura dos dados brutos por meio do CSV File Input, seguida pelo Value Mapper, que realiza a padronização dos códigos de critério, convertendo valores numéricos ou abreviados em suas respectivas descrições textuais. Em seguida, o Select Values organiza os atributos que farão parte da dimensão, selecionando apenas os campos relevantes, ajustando tipos de dados e eliminando informações desnecessárias. Por fim, o Table Output grava os registros tratados na tabela **dim_criterio**, concluindo o processo de carga de forma estruturada e consistente.

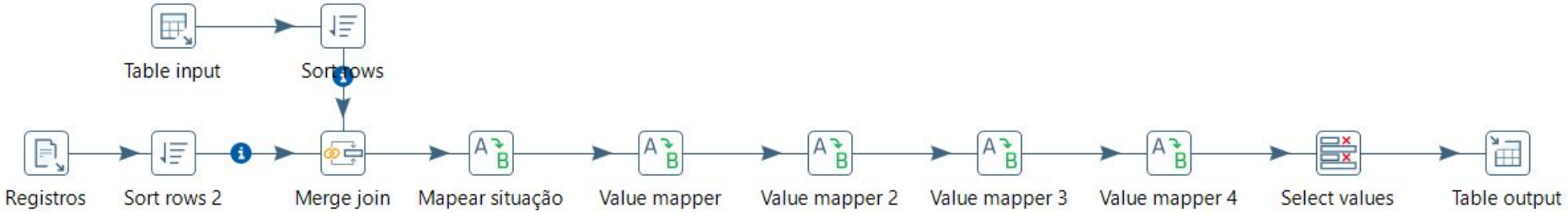
16. Plano de Carga da Dimensão Testes

A transformação da dimensão testes inicia-se com a leitura estruturada do arquivo CSV fornecido no sistema de vigilância epidemiológica do SUS. contendo os registros brutos, que são então submetidos a uma sequência de cinco etapas de Value Mapper, responsáveis por padronizar e recodificar valores específicos das colunas a fim de garantir consistência semântica e integridade referencial. Após essa normalização progressiva dos atributos, aplica-se o step Select Values, no qual são selecionados apenas os campos relevantes para compor a dimensão e descartados aqueles desnecessários ao modelo dimensional. Por fim, os dados tratados e devidamente padronizados são inseridos na tabela de destino por meio do step Table Output, concluindo o processo de construção da dimensão.



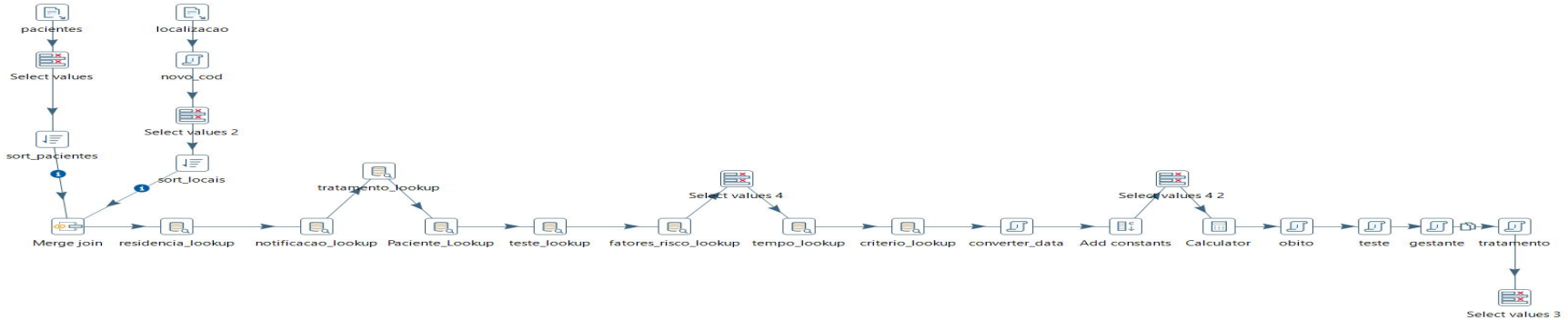
17. Plano de Carga da Dimensão Fatores de Risco

O processo de carga da Dimensão Fatores de Risco inicia com a leitura dos registros brutos do SINAN e sua ordenação, seguida de um merge join com a base auxiliar para complementar informações. Em seguida, aplicamos uma etapa de padronização composta por vários Value Mappers, responsável por traduzir códigos e corrigir nomenclaturas conforme a gramática oficial do SINAN, além do mapeamento da situação do fator de risco. Após essa harmonização, selecionamos apenas os atributos relevantes para a modelagem dimensional e carregamos o resultado final na tabela de dimensão, garantindo consistência, integridade e padronização dos fatores de risco utilizados nas análises epidemiológicas.



18. Plano de Carga da Fato

A carga da tabela fato inicia com a integração dos dados de pacientes e localização, previamente selecionados e ordenados, por meio de um merge join que consolida informações demográficas e territoriais. Em seguida, o processo percorre uma série de lookups das dimensões, garantindo a substituição dos valores brutos pelos respectivos códigos dimensionais. Após isso, são aplicadas transformações complementares, como conversão de datas, criação de atributos derivados através do uso de constantes e cálculos, além da normalização de campos específicos, como óbito, teste, gestante e tratamento. Por fim, os campos finais são selecionados e estruturados para compor a tabela fato, assegurando integridade referencial, padronização e completude dos registros epidemiológicos para análise no data mart.



ETAPA 6 - APLICAÇÃO OLAP e PAINEL DE BORDO

19. Consulta OLAP 1

Na primeira consulta OLAP, analisamos o número de casos por localização e por sexo, além da evolução temporal das notificações.

20. Consulta OLAP 2

Análise do número de notificações por etnia(raça), nível de escolaridade e faixa etária.

21. Consulta OLAP 3

Análise da evolução temporal do número de óbitos por localização.

21. Consulta OLAP 4

Análise do número dos casos em gestantes por escolaridade, além de filtrar esses dados por localização e faixa etária. Além disso, analisamos também os casos onde houve contaminação por acidente.

22. Consulta OLAP 5

As duas últimas consultas refere-se a analisar esses dados para as cidades do Estado de Pernambuco.

Referências

1. Dados fontes (abertos)

Anexos

Apêndices