Data Warehousing e Inteligência de Negócio – ICP357
DW no Suporte à Tomada de Decisão – ICP602
Trabalho Final – Parte 1
2025-2

Objetivo Geral
Implementar um ambiente analítico a partir de dados reais, executando todas as fases que usualmente são realizadas em um projeto de solução analítica, desde a coleta dos dados, passando pelas tarefas de modelagem, tratamento e transformação, até a análise. O ambiente analítico deve ser pensado para um tomador de decisão, permitindo que seja possível explorar os dados usando alguma ferramenta de mais alto nível.

Temas
Serão disponibilizados quatro temas distintos, e cada tema será trabalhado por dois grupos. Os temas abrangem domínios distintos como Saúde, Educação, Meio Ambiente e Assistência Social.

- **Tema 1: Notificações de Casos de Dengue no Estado do Rio de Janeiro**
  Os dados de notificações de casos de dengue são provenientes do Sistema de Informação de Agravos de Notificação (Sinan) e disponibilizados no Portal Brasileiro de Dados Abertos como microdado desidentificado (https://dados.gov.br/dados/conjuntos-dados/arboviroses-dengue). O subconjunto a ser utilizados contém dados dos últimos 5 anos (2021–2025) relativos a notificações no estado do Rio de Janeiro.

- **Tema 2: Matrículas no Ensino Técnico na Cidade do Rio de Janeiro**
  Os dados de matrículas no ensino técnico são provenientes da Plataforma Nilo Peçanha (PNP), ambiente virtual de coleta, validação e disseminação das estatísticas oficiais da Rede Federal de Educação Profissional, Científica e Tecnológica, e disponibilizados no Portal Brasileiro de Dados Abertos como microdado desidentificado (https://dados.gov.br/dados/conjuntos-dados/mec-plataforma-nilo-pecanha-pnp). O subconjunto a ser utilizados contém dados de 5 anos (2019-2023) relativos a matrículas nas instituições de Educação Profissional e Tecnológica da cidade do Rio de Janeiro.

- **Tema 3: Multas Ambientais no Brasil**
  Os dados de multas ambientais aplicadas a pessoas físicas ou jurídicas são disponibilizados pelo Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis (Ibama) no Portal Brasileiro de Dados Abertos (https://dados.gov.br/dados/conjuntos-dados/multas-ambientais-distribuidas-por-bens-tutelados). O conjunto completo consiste de dados de autos de infração, com ou sem julgamento, sobre bens tutelados (fauna, flora, unidades de conservação, etc.) que resultaram na aplicação de multa, e abrange dados históricos de todos os estados do Brasil.

- **Tema 4: Famílias no Cadastro Único na Cidade do Rio de Janeiro**
  Os dados sobre famílias no Cadastro Único são disponibilizados como microdado desidentificado pelo Ministério do Desenvolvimento e Assistência Social, Família e Combate à Fome (MDS) no Portal Brasileiro de Dados Abertos (https://dados.gov.br/dados/conjuntos-dados/microdados-amostrais-do-cadastro-unico). O subconjunto a ser utilizado contém dados de 5 anos (2014-2018) relativos a famílias cadastradas na cidade do Rio de Janeiro.

O objetivo do trabalho é pensar um ambiente analítico que permita gerar as evidências necessárias para a tomada de decisão, respondendo a questão do tipo:

- Para o *Tema 1*, como o número de notificações está distribuído pela cidade? Qual o perfil demográfico das pessoas afetadas? Qual a proporção de indivíduos afetados que vieram a óbito? Como tudo isso varia ao longo dos anos? A tendência é de queda ou de aumento no número de notificações?

- Para o *Tema 2*, quais são as áreas de conhecimento com maior procura por matrículas? Está havendo aumento no número de matrículas em cursos a distância? Em que áreas? A tendência é de queda ou de aumento na taxa de evasão? Há diferença no perfil demográfico entre alunos evadidos e aqueles que concluem o curso?

- Para o *Tema 3*, qual o valor total de multas aplicadas em cada ano? Quem são os maiores infratores? Qual é a diferença entre o valor médio das multas aplicadas a pessoas físicas e jurídicas? Quais são os bens mais afetados? Como tudo isso varia ao longo dos anos? E entre as diferentes regiões do país?
- Para o *Tema 4*, como varia o número de famílias cadastradas ao longo dos anos? Qual a renda média dessas famílias? Qual é a condição predominante dos domicílios destas famílias? Qual a proporção das famílias que têm a acesso a saneamento básico? Como tudo isso varia de acordo com a região da cidade onde a família se cadastrou?

As análises pretendidas, bem como o modelo dimensional que irá suportá-las, não devem se limitar aos campos presentes nos datasets, devendo considerar também as informações que podem ser derivadas dos dados já existentes. É fortemente recomendado o uso de conjuntos de dados complementares para enriquecer os dados nos datasets, como, por exemplo, nome do município e UF onde há somente o código IBGE, região onde se encontra o Estado (Norte, Nordeste, Sul, Sudeste, Centro-Oeste), zona onde se localiza o bairro (Norte, Sul, Oeste, Centro), endereço de instituições de ensino ou de Centros de Referência de Assistência Social (CRAS), tamanho da população para que comparações quantitativas possam ser colocadas em escala, etc. Alguns links úteis (não se limitar apenas a estes):

- Códigos dos municípios IBGE
- Centros de Referência de Assistência Social no Rio de Janeiro
- Bairros do Rio de Janeiro por Zona
- Rede Federal de Ensino Profissional e Tecnológico do Rio de Janeiro

Os conjuntos primários de dados a serem utilizados estão disponíveis no Google Drive. Conjuntos adicionais são de livre escolha.

Parte 1 - Entendimento do Domínio e Modelagem

A primeira parte do trabalho inclui o estudo e compreensão do domínio e dos dados, e a construção de um modelo dimensional para representar os dados trabalhados. A entrega deve incluir:

1. Percepção do grupo sobre o tema do trabalho, incluindo uma breve descrição do que é abordado no conjunto de dados, o que entendem como motivação para criar um ambiente analítico para este domínio e uma lista de análises (em alto nível) que acreditam que sejam relevantes e que serão cobertas pelo modelo e ambiente analítico a serem construídos;
2. Caso não estejam utilizando todos os dados, descrição do recorte selecionado para os dados e objetivos associados à seleção (Este subconjunto foi escolhido por quê? O que o grupo acredita que pode fazer com ele?);
3. A modelagem dimensional dos dados;
4. O dicionário de dados do modelo gerado.

Requisitos

- O trabalho deve ser feito em **grupos de 3 ou 4 alunos**. Não serão aceitos trabalhos individuais ou em dupla;
- Na construção do modelo dimensional, considerar as seguinte etapas, descrevendo e documentando as decisões em relação a cada uma delas:
    1. Identificar o fato;
    2. Definir a granularidade do fato;
    3. Identificar as dimensões.
- O modelo dimensional deve obrigatoriamente incluir as dimensões de tempo e de localização geográfica (o que define a localização geográfica do fato varia de acordo com o tema), as demais dimensões vão depender do domínio e das decisões de modelagem;
- Em datasets com um número grande de campos, não é necessário utilizar todos eles, mas o recorte escolhido deve ser justificado através das análises pretendidas sobre o modelo resultante.