PUC -RJ

MBA em Ciência de dados | MVP Engenharia de Dados

Projeto ETL – Análise de dados sobre o Comércio Exterior Brasileiro

Período: 1997 a 2020

Autor: Wallace de Lima da Silva Conceição

Orientador:

Data: 04/2025

Documentação:

Antes da ingestão dos dados no Databricks, foi criado um diretório no Google Drive para centralizar temporariamente o upload das bases do projeto.
Embora seja sabido que o ideal, em ambientes produtivos, seja o uso de buckets ou data lakes como GCS, S3 ou Azure, essa abordagem foi adotada visando otimizar custos e tempo de desenvolvimento no contexto acadêmico.

O link de acesso à pasta é:Link

Tabela: EXP_COMPLETA

1. ***Catálogo de Dados***

- Nome da Tabela: EXP_COMPLETA
- Descrição da tabela: Registro completo das exportações brasileiras entre 1997 e 2020, com base em dados oficiais da SECEX.
- Fonte: Kaggle (SECEX) - Exportações Brasileiras
- Periodicidade: Mensal (1997 a 2020)
- Volume de Registros: 22.398.370
- Número de Colunas: 11
- Tamanho Aproximado: 1.5 GB
- Primary Key Composta: (CO_ANO, CO_MES, CO_NCM, CO_PAIS, SG_UF_NCM, CO_VIA, CO_URF)

2. ***Metadados – Dicionário de dados***

Campo	Tipo	Descrição
CO_ANO	Integer	Ano da Exportação
CO_MES	Integer	Mês da Exportação
CO_NCM	String	Código da mercadoria exportada (NCM)
CO_UNID	String	Código da unidade de medida
CO_PAIS	String	Código do país de destino

SG_UF_NCM	String	Sigla da unidade federativa de origem
CO_VIA	String	Código do modal logístico utilizado
CO_URF	String	Código da unidade de despacho (URF)
QT_ESTAT	Integer	Quantidade estatística da mercadoria
KG_LIQUIDO	Float	Peso líquido em KG exportado
VL_FOB	Float	Valor FOB (Free on Board da Mercadoria Exportada em dólares)

3. ***Qualidade dos Dados***

- Completude: Todos os campos obrigatórios estão preenchidos; não há valores nulos nas colunas-chave.
- Consistência: Os dados são coerentes entre os campos e compatíveis com tabelas auxiliares.
- Conformidade: Os dados aderem aos padrões estabelecidos para códigos e nomenclaturas.
- Integridade: As relações entre as entidades são mantidas corretamente, garantindo integridade referencial.
- Acurácia: Os valores refletem com precisão as exportações registradas.
- Atualidade: Os dados abrangem o período de 1997 a 2020, estando atualizados até o último ano disponível.

Tabela: IMP_COMPLETA

1. ***Catálogo de Dados***

- Nome da Tabela: IMP_COMPLETA
- Descrição: Registro completo das importações brasileiras entre 1997 e 2020, com base em dados oficiais da SECEX.
- Fonte: Kaggle (SECEX) - Importações Brasileiras
- Periodicidade: Mensal (1997 a 2020)
- Volume de Registros: 33.491.095
- Número de Colunas: 11
- Tamanho Aproximado: 2.21 GB
- Primary Key Composta: (CO_ANO, CO_MES, CO_NCM, CO_PAIS, SG_UF_NCM, CO_VIA, CO_URF)

2. ***Metadados – Dicionário de dados***

Campo	Tipo	Descrição
CO_ANO	Integer	Ano da Importação
CO_MES	Integer	Mês da Importação
CO_NCM	String	Código da mercadoria importada (NCM)
CO_UNID	String	Código da unidade de medida
CO_PAIS	String	Código do país de destino
SG_UF_NCM	String	Silga da unidade federativa de origem
CO_VIA	String	Código do modal logístico utilizado
CO_URF	String	Código da unidade de despacho (URF)
QT_ESTAT	Integer	Quantidade estatística da mercadoria
KG_LIQUIDO	Float	Peso líquido em KG exportado
VL_FOB	Float	Valor FOB (Free on Board da Mercadoria Exportada em dólares)

3. ***Qualidade dos Dados***

- Completude: Todos os campos obrigatórios estão preenchidos; não há valores nulos nas colunas-chave.
- Consistência: Os dados são coerentes entre os campos e compatíveis com tabelas auxiliares.
- Conformidade: Os dados aderem aos padrões estabelecidos para códigos e nomenclaturas.
- Integridade: As relações entre as entidades são mantidas corretamente, garantindo integridade referencial.
- Acurácia: Os valores refletem com precisão as importações registradas.
- Atualidade: Os dados abrangem o período de 1997 a 2020, estando atualizados até o último ano disponível.

Tabela: NCM

1. ***Catálogo de Dados***

- Nome da Tabela: NCM
- Descrição: Tabela de códigos NCM, que classifica mercadorias conforme o Mercosul.
- Fonte: Kaggle (SECEX) – Dados Auxiliares
- Periodicidade: Fixa
- Volume de Registros: 13.111
- Tamanho Aproximado: 3 MB
- Número de Colunas: 14
- Primary Key: (CO_NCM)

2. ***Metadados – Dicionário de dados***

Campo	Tipo	Descrição
CO_NCM	String	Código da mercadoria
CO_UNID	String	Código da unidade de medida associada ao NCM
CO_SH6	String	Código SH6 - Sistema Harmonizado, 6 dígitos
CO_PPE	String	Código da Posição Principal de Exportação
CO_PPI	String	Código da Posição Principal de Importação
CO_FAT_AGREG	String	Código do Fator Agregado
CO_CUCI_ITEM	String	Código CUCI (Classificação Uniforme para Comércio Internacional)
CO_CGCE_N3	String	Código da Classificação por Grande Categoria Econômica (nível 3)
CO_SIIT	String	Código da Classificação da Indústria segundo o SIIT
CO_ISIC_CLASSE	String	Código ISIC (Classificação Internacional Industrial Padrão) – classe
CO_EXP_SUBSET	String	Código de subconjunto de exportação (alguns registros estão nulos)
NO_NCM_POR	String	Descrição da mercadoria em português
NO_NCM_ESP	String	Descrição da mercadoria em espanhol
NO_NCM_ING	String	Descrição da mercadoria em inglês

3. ***Qualidade dos Dados***

- Completude: Todos os campos obrigatórios estão preenchidos, exceto CO_EXP_SUBSET que possui valores nulos.
- Consistência: Os dados seguem padrões padronizados do Mercosul e Sistema Harmonizado.
- Conformidade: Formatos e nomenclaturas estão de acordo com classificações internacionais.
- Integridade: As relações entre códigos e descrições são mantidas sem duplicidade.
- Acurácia: As informações representam corretamente as categorias comerciais internacionais.
- Atualidade: Embora fixa, a tabela é válida até nova atualização oficial do NCM.

Tabela: PAIS

1. ***Catálogo de Dados***

- Nome da Tabela: PAIS
- Descrição: Tabela referencial com códigos e nomes de países utilizados nas operações de comércio exterior. Inclui representações padronizadas ISO e nomes em três idiomas.
- Fonte: Kaggle (SECEX) – Dados Auxiliares
- Periodicidade: Fixa
- Volume de Registros: 281
- Número de Colunas: 6
- Tamanho Aproximado: 17kb
- Primary Key: (CO_PAIS)

2. ***Metadados – Dicionário de dados***

Campo	Tipo	Descrição
CO_PAIS	String	Código interno do país (chave primária usada pela SECEX)
CO_PAIS_ISON3	String	Código ISO numérico (ISO 3166-1 numeric)
CO_PAIS_ISOA3	String	Código ISO Alpha-3 do país (ex: BRA, USA, CHN)
NO_PAIS	String	Nome do país em português
NO_PAIS_ING	String	Nome do país em inglês
NO_PAIS_ESP	String	Nome do país em espanhol

3. ***Qualidade dos Dados***

- Completude: Todos os campos estão 100% preenchidos.
- Consistência: Códigos e nomes em diferentes idiomas estão corretamente alinhados.
- Conformidade: Utiliza padrões ISO para nomenclatura e codificação.
- Integridade: Não há duplicidade de registros ou inconsistências entre colunas.
- Acurácia: Os dados correspondem aos padrões internacionais reconhecidos.
- Atualidade: Considerando que os países não mudam frequentemente, os dados são considerados atualizados.

Tabela: PAIS_BLOCO

1. ***Catálogo de Dados***

- Nome da Tabela: PAIS_BLOCO
- Descrição: Relação entre países e os blocos econômicos aos quais pertencem.
- Fonte: Kaggle (SECEX) – Dados Auxiliares
- Periodicidade: Fixa
- Volume de Registros: 323
- Número de Colunas: 5
- Tamanho Aproximado: 24kb
- Primary Key: (CO_PAIS + CO_BLOCO)

2. ***Metadados – Dicionário de dados***

Campo	Tipo	Descrição
CO_PAIS	String	Código do país
CO_BLOCO	String	Código do bloco econômico
NO_BLOCO	String	Nome do bloco econômico em português
NO_BLOCO_ING	String	Nome do bloco econômico em inglês
NO_BLOCO_ESP	String	Nome do bloco econômico em espanhol

3. ***Qualidade dos Dados***

- Completude: Todos os campos estão 100% preenchidos.
- Consistência: Relações entre países e blocos estão bem estruturadas e sem conflito.
- Conformidade: Os nomes e códigos seguem nomenclaturas padronizadas e multilíngue.
- Integridade: Relacionamentos entre país e bloco mantêm coerência referencial.
- Acurácia: Os vínculos entre países e blocos são condizentes com acordos internacionais.
- Atualidade: Estrutura válida até mudança oficial em tratados econômicos.

Tabela: UF

1. ***Catálogo de Dados***

- Nome da Tabela: UF
- Descrição: Lista das Unidades Federativas do Brasil, contendo códigos, siglas, nomes e respectivas regiões geográficas..
- Fonte: Kaggle (SECEX) – Dados Auxiliares
- Periodicidade: Fixa
- Volume de Registros: 34
- Número de Colunas: 4
- Tamanho Aproximado: 2kb
- Primary Key: (CO_UF)

2. ***Metadados – Dicionário de dados***

Campo	Tipo	Descrição
CO_UF	String	Código da unidade federativa
SG_UF	String	Sigla da UF
NO_UF	String	Nome completo da UF
NO_REGIAO	String	Nome da região geográfica

3. ***Qualidade dos Dados***

- Completude: Todos os campos estão 100% preenchidos.
- Consistência: Siglas, nomes e regiões condizem com a estrutura federativa brasileira.
- Conformidade: Os nomes seguem convenções oficiais do IBGE.
- Integridade: Não há duplicidade ou conflito entre códigos e siglas.
- Acurácia: Os dados representam corretamente todas as 27 UFs e subdivisões necessárias.
- Atualidade: Dados fixos, válidos até alteração oficial da divisão político-administrativa do Brasil.

Tabela: UF_MUN

1. ***Catálogo de Dados***

- Nome da Tabela: UF_MUN
- Descrição: Lista de municípios do Brasil com vinculação À sigla da UF.
- Fonte: Kaggle (SECEX) – Dados Auxiliares
- Periodicidade: Fixa
- Volume de Registros: 5570
- Número de Colunas: 4
- Tamanho Aproximado: 247kb
- Primary Key: (CO_MUN_GEO)

2. ***Metadados – Dicionário de dados***

Campo	Tipo	Descrição
CO_MUN_GEO	String	Código geográfico do município (IBGE)
NO_MUN	String	Nome do município
NO_MUN_MIN	String	Nome abreviado do município
SG_UF	String	Sigla da unidade federativa

3. ***Qualidade dos Dados***

- Completude: Todos os campos estão 100% preenchidos.
- Consistência: Os códigos de município estão corretamente relacionados às UFs.
- Conformidade: As nomenclaturas seguem o padrão oficial do IBGE.
- Integridade: Relações entre municípios e UFs são consistentes e não ambíguas.
- Acurácia: Os nomes e códigos geográficos refletem corretamente a estrutura territorial brasileira.
- Atualidade: Estrutura válida até eventuais mudanças administrativas nos municípios.

Tabela: URF

1. ***Catálogo de Dados***

- Nome da Tabela: URF
- Descrição: Lista das Unidades de Receita Federal (alfândegas).
- Fonte: Kaggle (SECEX) – Dados Auxiliares
- Periodicidade: Fixa
- Volume de Registros: 275

- Número de Colunas: 2
- Tamanho Aproximado: 11kb
- Primary Key: (CO_URF)

2. ***Metadados – Dicionário de dados***

Campo	Tipo	Descrição
CO_URF	String	Código da unidade de despacho aduaneira
NO_URF	String	Nome da unidade

3. ***Qualidade dos Dados***

- Completude: Todos os campos estão completamente preenchidos.
- Consistência: Códigos e nomes estão em conformidade com a estrutura oficial da Receita Federal.
- Conformidade: Padrões de nomenclatura e codificação seguem órgãos governamentais.
- Integridade: Relação direta e única entre código e nome da URF.
- Acurácia: Representa com precisão todas as unidades operacionais de despacho no país.
- Atualidade: Estrutura válida até alteração ou reestruturação administrativa da Receita Federal.

Tabela: VIA

1. ***Catálogo de Dados***

- Nome da Tabela: VIA
- Descrição: Tipos de modais logísticos utilizados na importação e exportação.
- Fonte: Kaggle (SECEX) – Dados Auxiliares
- Periodicidade: Fixa
- Volume de Registros: 17
- Número de Colunas: 2
- Tamanho Aproximado: 1kb
- Primary Key: (CO_VIA)

2. ***Metadados – Dicionário de dados***

Campo	Tipo	Descrição
CO_VIA	String	Código do modal logístico
NO_VIA	String	Descrição do modal (exemplo: marítimo, etc)

4. ***Qualidade dos Dados***

- Completude: Todos os registros estão completos e sem valores nulos.
- Consistência: Códigos e descrições estão corretamente relacionados e não ambíguos.
- Conformidade: Os nomes dos modais seguem padrões logísticos reconhecidos.
- Integridade: Cada código corresponde exclusivamente a uma descrição de modal.
- Acurácia: Os dados representam corretamente as modalidades logísticas utilizadas na prática.
- Atualidade: Tabela válida até inclusão de novos modais ou mudança oficial de nomenclatura.

Resumo Geral da Documentação:

Tabela	Registros Aprocimados	Colunas	Tamanho Estimado	Tipo de Chave
EXP_COMPLETA	22.398.370	11	1.5 GB	Primária Composta
IMP_COMPLETA	33.491.095	11	2.26 GB	Primária Composta
NCM	13.11	14	1.4MB	Primária Simples
PAIS	281	6	17 kb	Primária Simples
PAIS_BLOCO	323	5	24 kb	Primária Simples
UF	34	4	2 kb	Primária Simples
UF_MUN	5.570	4	247 kb	Primária Simples
URF	275	2	11 kb	Primária Simples
VIA	17	2	1 kb	Primária Simples

Padrão de Qualidade Adotado

Cada tabela foi avaliada segundo as **dimensões de qualidade do DAMA-DMBOK**:

- Completude
- Consistência
- Conformidade
- Integridade Referencial
- Acurácia
- Atualidade

Observações para Modelagem e Ingestão

- As tabelas **EXP_COMPLETA** e **IMP_COMPLETA** serão as fatores centrais do modelo, ligadas às demais via chaves como **CO_NCM, CO_PAIS, SG_UF_NCM, CO_URF, e CO_VIA**.
- Durante a ingestão em Databricks, recomenda-se a conversão de todos os campos de identificadores numéricos para string, mantendo zeros à esquerda e padrão de joins.
- Todas as tabelas auxiliares são dimensões estáticas que suportam consistência, validação e enriquecimento analítico.
- Tabelas fato serão criadas com particionamento por ano e mês, otimizando análise temporal.

- Os dados podem ser usados para:
    - Análises unificadas (JOIN EXP + IMP)
    - Drill-down geográfico e de produtos
    - Séries temporais de desempenho comercial

Tabelas Fato

Tabela Fato	Descrição
EXP_COMPLETA	Registra cada operação de exportação mensal por NCM, país, UF, via, etc.
IMP_COMPLETA	Registra cada operação de importação mensal pelos mesmos eixos

Tabelas Dimensão

Dimensão	Chave Primária	Finalidade
NCM	CO_NCM	Classificação da mercadoria exportada/importada
PAIS	CO_PAIS	Informações sobre países de origem/destino
UF	SG_UF	Unidades Federativas (origem/destino interno)
UF_MUN	CO_MUN_GEO	Municípios brasileiros (não utilizados diretamente nas tabelas fato nesse modelo, mas disponível)
URF	CO_URF	Unidades de despacho da Receita Federal (alfândegas)
VIA	CO_VIA	Modal Logístico utilizado (aéreo, marítimo, rodoviário, etc.)
PAIS_BLOCO	(CO_PAIS, CO_BLOCO)	Associação entre países e blocos económicos.

Relacionamentos (Dimensões ->Fato)

FATO	Fato	Chave Estrangeira
EXP_COMPLETA e IMP_COMPLETA (mesma estrutura)	CO_NCM	NCM
	CO_PAIS	PAIS
	SG_UF_NCM	UF
	CO_URF	URF
	CO_VIA	VIA