Proposta de trabalho final - Big Data

CAIO ALMEIDA SANTOS
HIGOR BATISTA FERNANDES
JOÃO PEDRO BRUNO MACHADO
ROBSON FERREIRA DOS SANTOS JUNIOR
VINICIUS PASSOS OLIVEIRA

Fonte de Dados e Motivação

Fonte de Dados:

- Série Histórica de Preços de Combustíveis e de GLP – Agência Nacional do Petróleo (ANP)
- Lei do Petróleo (Lei nº 9478/1997, artigo 8º)
- Preços praticados por revendedores de combustíveis automotivos e de gás liquefeito de petróleo (GLP P13)
- Pesquisa semanal de preços
- Disponível em: https://dados.gov.br/dados/conjuntos-dados/serie-historica-de-precos-de-combustiveis-e-de-glp

Contexto e Motivação:

- Monitorar a evolução e comportamento do preço de combustíveis no Brasil entre 2004 e 2021
- Identificar padrões de variação por ano, tipos de combustível
- Possibilitar comparações históricas (ex.: impactos de crises, políticas de subsídio, flutuações cambiais)

Estrutura dos Arquivos CSV Originais

Os arquivos CSV contém dados detalhados sobre os preços de venda de combustíveis (gasolina, etanol, diesel, GNV, GLP) em diversos municípios brasileiros.

Formato dos arquivos CSV (Raw):

- Regiao - Sigla
- Estado - Sigla
- Municipio
- CNPJ da Revenda
- Nome da Rua
- Numero Rua
- Complemento
- Bairro
- Cep
- Produto
- Data da Coleta
- Valor de Venda
- Valor de Compra
- Unidade de Medida
- Bandeira

Exemplos de dados do CSV

Regiao - Sigla	Estado - Sigla	Municipio	Revenda	CNPJ da Revenda	Nome da Rua	Numero Rua	Complemento	Bairro	Cep	Produto	Data da Coleta	Valor de Venda	Valor de Compra	Unidade de Medida	Bandeira
SE	SP	GUARULHOS	AUTO POSTO	49.051.667/	RODOVIA PRE	S/N	KM 210,5-SENT SP/R	BONSUCESSO	07178-580	GASOLINA	11/05/2004	1,967	1,6623	R$ / litro	PETROBRAS DISTRIBUIDORA S.A.
SE	SP	GUARULHOS	AUTO POSTO	49.051.667/	RODOVIA PRE	S/N	KM 210,5-SENT SP/R	BONSUCESSO	07178-580	ETANOL	11/05/2004	0,899	0,6282	R$ / litro	PETROBRAS DISTRIBUIDORA S.A.
SE	SP	GUARULHOS	AUTO POSTO	49.051.667/	RODOVIA PRE	S/N	KM 210,5-SENT SP/R	BONSUCESSO	07178-580	DIESEL	11/05/2004	1,299	1,1704	R$ / litro	PETROBRAS DISTRIBUIDORA S.A.
SE	SP	SOROCABA	COMPETRO C	00.003.188/	RUA HUMBER	306		JARDIM ZULM	18061-000	GASOLINA	10/05/2004	1,85	1,67	R$ / litro	BRANCA
SE	SP	SOROCABA	COMPETRO C	00.003.188/	RUA HUMBER	306		JARDIM ZULM	18061-000	ETANOL	10/05/2004	0,78	0,48	R$ / litro	BRANCA
SE	SP	SOROCABA	COMPETRO C	00.003.188/	RUA HUMBER	306		JARDIM ZULM	18061-000	DIESEL	10/05/2004	1,29	1,216	R$ / litro	BRANCA
CO	DF	BRASILIA	GASOL COMB	00.603.738/	QI-QUADRA I	S/N		TAGUATINGA	72315-000	GASOLINA	10/05/2004	2,03	1,7021	R$ / litro	PETROBRAS DISTRIBUIDORA S.A.
CO	DF	BRASILIA	GASOL COMB	00.603.738/	QI-QUADRA I	S/N		TAGUATINGA	72315-000	ETANOL	10/05/2004	1,29	0,8437	R$ / litro	PETROBRAS DISTRIBUIDORA S.A.
CO	DF	BRASILIA	GASOL COMB	00.603.738/	QI-QUADRA I	S/N		TAGUATINGA	72315-000	DIESEL	10/05/2004	1,46	1,2487	R$ / litro	PETROBRAS DISTRIBUIDORA S.A.
NE	BA	SALVADOR	PETROBRAS D	34.274.233/	RUA EDISTIO	474		STIEP	41770-395	GASOLINA	11/05/2004	1,91	1,7059	R$ / litro	PETROBRAS DISTRIBUIDORA S.A.

O que foi feito no Projeto

- **Ingestão dos dados**: Download automático de centenas de arquivos CSV do portal da ANP

- **Pré-processamento: Limpeza, padronização de colunas, conversão de tipos e criação de colunas derivadas (lucro, semestre, mandato).**

- **Modelagem dimensional: Criação de um modelo estrela com tabelas fato e dimensões (produto, tempo, local).**

O que foi feito no Projeto

- **Consultas analíticas: Cálculo da média de preços por mandato, semestre e tipo de combustível.**

- **Visualizações: Geração de gráficos de séries temporais com Matplotlib.**

- **Métricas de desempenho: Tempo de execução das etapas e tamanho das zonas de dados (raw, trusted, refined).**

Relevância do Processamento

Benefícios:

Identificar padrões de variação (ex: inflação sazonal).

Correlacionar eventos históricos (ex: crises econômicas) com picos de preço.

Impacto:

Base para estudos econômicos e logísticos.

Transparência para o cidadão.

Tipo de Processamento

Batch (Lote)

Características:

Os dados são processados em grandes conjuntos, não em tempo real. Ideal para análises históricas e ETL de grandes volumes.

PySpark: API Python para Apache Spark, poderosa para processamento distribuído de grandes volumes de dados.

Spark SQL: Módulo do Spark para trabalhar com dados estruturados usando consultas SQL, facilitando a manipulação e análise.

Por que é Big Data?

Volume dos Dados:

- A série histórica da ANP compreende um grande volume de dados, acumulados ao longo de muitos anos (2004-2024).
- São disponibilizados diversos arquivos, separados por ano e/ou semestre e tipo de combustível, resultando em uma quantidade significativa de dados a serem processados e gerenciados

Variedade de Arquivos:

- Dados em formatos diferentes (CSV, Excel)
- Diferentes níveis de agregação (regional, estadual, municipal)

Velocidade:

- Possibilidade de lidar com diversos novos arquivos a cada ciclo de atualização dos dados

Por que é Big Data?

Veracidade dos Dados:

- Possibilidade de dados faltantes ou inconsistentes.
  a. Erros humanos
  b. Dados faltantes
  c. Inconsistências/Falta de padronização

Valor das informações e análises:

- A extração de informações úteis para diversos contextos, tais como:
  a. Entender padrões de mercado
  b. Avaliar políticas públicas
  c. Ajudar consumidores e empresas no planejamento financeiro
  d. Investigar tendência e razões para variação nos preços

Por que é desafiador processar esses dados?

Tamanho dos arquivos

- Volume massivo de dados coletados entre 2004 e 2021

Desafio

- Não basta usar pandas/Excel; é necessário uma ferramenta de processamento distribuído como o spark

Qualidade dos dados

Dados reais, coletados ao longo de muitos anos, geralmente têm problemas como:

- Dados faltantes, Formatos diferentes ou Erros de digitação

Desafio

- Tratamento de dados para padronização antes da análise ser realizada

Por que é desafiador processar esses dados?

Conversão para Parquet

- Compactação melhor (menor tamanho)
- Leituras mais rápidas

Desafio

- Converter de CSV para Parquet demanda processamento e espaço
- Schema bem definido
- Particionamento eficiente com foco em evitar arquivos desnecessários

Arquitetura Proposta : Camadas do Datalake

Raw (Bruta):

- CSVs originais da ANP (sem alterações).
- Função: Preservar dados crus para auditoria.

Trusted (Confiança):

- Dados convertidos para Parquet (apenas colunas relevantes).
- Função: Dados limpos, eficientes e prontos para análise.

Arquitetura Proposta : Camadas do Datalake

Refined (Refinada):

- **Modelo dimensional (ex: tabelas de fato/dimensão).**

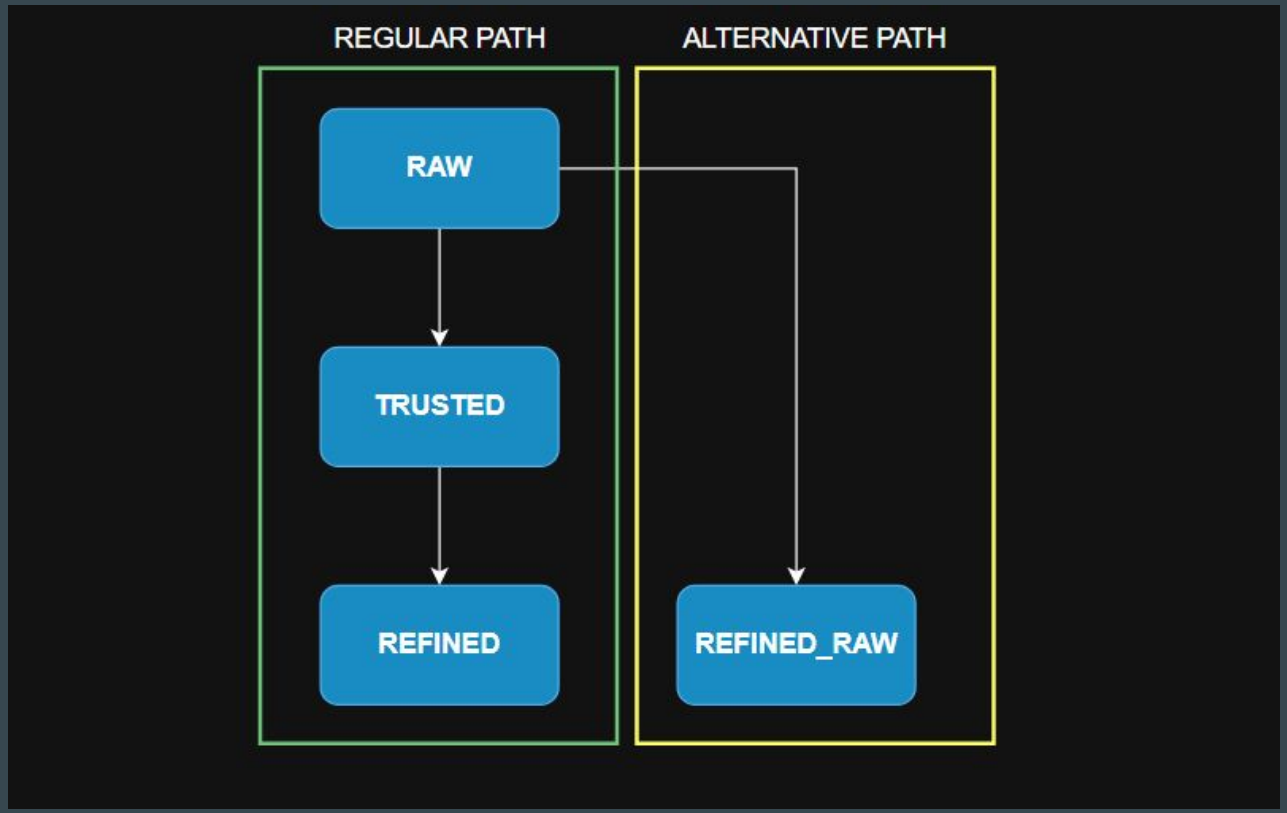
  Exemplo: Fato: Preço, Data_ID, Local_ID, Produto_ID. Dimensão: Data (ano, mês), Local (região, UF), Produto (tipo, bandeira).

- **Função: Acelerar consultas analíticas.**

Refined_raw (Refinada a partir da camada raw):

- **Tem os mesmos dados da camada refined pardrão,**
- **Função: verificar o tempo de processamento e criação da camada Refined a partir da camada Raw.**

Arquitetura Proposta : Camadas do Datalake



Fluxo de Processamento

Pipeline com PySpark

1-Leitura: Do arquivo CSV

2-Camada Trusted: Filtro de colunas (select) + conversão para Parquet.

3-Camada Refined: Construção do modelo estrela via Spark SQL.

4-Análise: Consultas agregadas (ex: variação média anual por região). Visualização (ex: gráficos de linha com Matplotlib).

Resultados das execuções

- **Verificarmos os diferentes resultados das execuções para poder identificar possíveis mudanças**
- **Depois de fazer essas execuções observamos que todas tiveram razoavelmente os mesmos restultados**

Resultados das execuções

- **Primeira execução**

```
Chegada de arquivos (Raw - Refined): 2 minutos e 22 segundos
Chegada de arquivos (Trusted - Refined): 0 minutos e 44 segundos
Tempo de consulta (Raw): 1 minutos e 42 segundos
Tempo de consulta (Trusted): 0 minutos e 9 segundos
Tempo de consulta (Refined): 0 minutos e 7 segundos
Tempo de consulta (Raw - Refined): 0 minutos e 6 segundos
Tamanho do diretório (Raw): 3315.22 MB
Tamanho do diretório (Trusted): 286.74 MB
Tamanho do diretório (Refined): 158.55 MB
Tamanho do diretório (Raw - Refined): 163.09 MB
```

Resultados das execuções

- **Segunda execução**

```
Chegada de arquivos (Raw - Refined): 2 minutos e 23 segundos
Chegada de arquivos (Trusted - Refined): 0 minutos e 38 segundos
Tempo de consulta (Raw): 2 minutos e 12 segundos
Tempo de consulta (Trusted): 0 minutos e 8 segundos
Tempo de consulta (Refined): 0 minutos e 6 segundos
Tempo de consulta (Raw - Refined): 0 minutos e 6 segundos
Tamanho do diretório (Raw): 3315.22 MB
Tamanho do diretório (Trusted): 286.74 MB
Tamanho do diretório (Refined): 158.55 MB
Tamanho do diretório (Raw - Refined): 157.67 MB
```

Resultados das execuções

- **Terceira execução**

```
Chegada de arquivos (Raw - Refined): 2 minutos e 46 segundos
Chegada de arquivos (Trusted - Refined): 0 minutos e 42 segundos
Tempo de consulta (Raw): 2 minutos e 27 segundos
Tempo de consulta (Trusted): 0 minutos e 9 segundos
Tempo de consulta (Refined): 0 minutos e 6 segundos
Tempo de consulta (Raw - Refined): 0 minutos e 6 segundos
Tamanho do diretório (Raw): 3315.22 MB
Tamanho do diretório (Trusted): 286.74 MB
Tamanho do diretório (Refined): 158.55 MB
Tamanho do diretório (Raw - Refined): 157.67 MB
```

Resultados das execuções

- **Quarta execução**

```
Chegada de arquivos (Raw - Refined): 2 minutos e 15 segundos
Chegada de arquivos (Trusted - Refined): 0 minutos e 37 segundos
Tempo de consulta (Raw): 2 minutos e 33 segundos
Tempo de consulta (Trusted): 0 minutos e 8 segundos
Tempo de consulta (Refined): 0 minutos e 6 segundos
Tempo de consulta (Raw - Refined): 0 minutos e 6 segundos
Tamanho do diretório (Raw): 3315.22 MB
Tamanho do diretório (Trusted): 286.74 MB
Tamanho do diretório (Refined): 158.55 MB
Tamanho do diretório (Raw - Refined): 163.09 MB
```

Referências e Links Úteis

Dados ANP:
https://dados.gov.br/dados/conjuntos-dados/serie-historica-de-precos-de-combustiveis-e-de-glp

Documentação PySpark:
https://spark.apache.org/docs/latest/api/python/index.html