Unidade curricular – Sistemas de Informação Analíticos

2.º ano de Tecnologias Digitais e Inteligência Artificial

Ano letivo 2024/2025

Docente: Carlos Miguel Fernandes Francisco

Relatório Final

Marin Cepeleaga – N.º 123550 | Guilherme Real – N.º 124456 |
Ouhao Wu-N.º123542

Maio 2025

Índice

Índice de figuras 3
Etapa 1 – Definição de Objetivos e Requisitos 4
Etapa 2 - Análise e Planeamento 5
Cronograma do Projeto 6
Recursos Necessários 6
Recursos Humanos: 6
Recursos Tecnológicos: 7
Diagrama Entidade-Relação do Data Warehouse: 7
Etapa 3: Modelação de Dados 8
Etapa 4: Seleção e Extração de Dados 10
Etapa 5: Limpeza, Transformação e Carregamento (ETL) 11
Etapa 6: Verificação da qualidade dos dados 12
Etapa 7: Validação da Qualidade dos Dados 13
Verificação de Duplicados 13
Etapa 8 – Análise Visual e Validação no Power BI 15
Etapa 9: Manutenção e Evolução 18
Conclusão e Resultados 19

Índice de figuras

Figura 1 - Modelo ER 8
Figura 2 - Modelo constelação em PowerBI 10
Figura 3 - Carregamento das tabelas do DW no Spoon 12
Figura 4 - Integridade Referencial 13
Figura 5 - Verificação de Duplicados 14
Figura 6 - Consistência dos Dados 14
Figura 7 - Integridade dos Dados 15
Figura 8 - Desvio Padrão 16
Figura 10 - Soma total dos preços de fecho por empresa 17
Figura 11- Média dos preços de fecho por empresa 17

Relatório

Título: ***Sistema Integrado de Data Warehouse para Análise Avançada do Mercado Acionista Brasileiro***

Objetivo: O objetivo deste projeto é desenvolver e implementar um Data Warehouse robusto e eficiente, capaz de consolidar dados históricos das ações negociadas na bolsa de valores brasileira (B3). Com este sistema, pretende-se permitir análises detalhadas que possibilitem identificar tendências setoriais, comparar o desempenho de diferentes empresas e avaliar os riscos associados, contribuindo assim para decisões estratégicas informadas.

Descrição da empresa e do negócio: A Market Analytics é uma empresa especializada na análise avançada de dados financeiros, oferecendo soluções analíticas e consultoria estratégica para investidores e empresas interessadas no mercado de ações brasileiro. Fundada com o compromisso de impulsionar o desempenho dos seus clientes por meio de insights derivados de dados históricos e ferramentas analíticas avançadas, a empresa utiliza tecnologia de ponta para fornecer relatórios detalhados e personalizados, dashboards interativos e previsões estratégicas. O seu objetivo é garantir decisões informadas, melhorar o desempenho financeiro e mitigar riscos no mercado acionista brasileiro.

Serviços principais:

- Consultoria financeira estratégica;
- Desenvolvimento de dashboards analíticos personalizados;
- Análise preditiva e gestão de risco;
- Formação e apoio em ferramentas analíticas.

A missão da Market Analytics é apoiar os seus clientes através da transformação dos dados históricos em insights valiosos, promovendo decisões fundamentadas e a melhoria contínua do desempenho financeiro.

Etapa 1 – Definição de Objetivos e Requisitos

Nesta fase inicial, definiram-se os objetivos e os requisitos essenciais para o desenvolvimento do Data Warehouse.

Questões a responder:

- Quais os setores que, historicamente, apresentam melhor desempenho no mercado acionista brasileiro?

- Quais as empresas que registaram maior rendibilidade, assim como níveis de volatilidade

significativos?

- De que forma reagiram as principais ações em períodos de crise e recuperação económica?

Além disso, foram estabelecidos **requisitos técnicos e funcionais**: **Técnicos:**

- ☐ Implementação de um modelo dimensional (esquema estrela) otimizado para consultas analíticas;
- ☐ Garantia da integridade e qualidade dos dados através de processos ETL (Extração, Transformação e Carregamento) automatizados;
- ☐ ETL eficiente e automatizado;
- ☐ Otimização do desempenho nas consultas históricas e agregações

Funcionais:

- Possibilitar a consulta e análise detalhada dos dados históricos;
- Facilitar a geração de relatórios e dashboards interativos;
- Oferecer uma interface intuitiva para a exploração dos dados consolidados.

Estas questões e requisitos vão orientar toda a execução do projeto, garantindo relevância e eficiência analítica do sistema.

Etapa 2 - Análise e Planeamento

Nesta etapa, procedeu-se à análise detalhada dos dados existentes, definição do plano técnico do projeto e da arquitetura do Data Warehouse:

Avaliação detalhada dos dados disponíveis:

- Análise dos dados históricos das ações, incluindo preços (abertura, máxima, mínima, fechamento) e volume negociado, bem como informações sobre as empresas (setor, segmento, capitalização);
- Informações sobre empresas (setores económicos, segmentos, capitalização).

Planeamento do projeto:

- Definição do escopo do projeto, tendo em conta as necessidades analíticas da Market Analytics e a viabilidade técnica do modelo proposto;
- Análise do Data Warehouse escolhido: optou-se por um modelo dimensional baseado no esquema estrela, que se revela adequado para análises rápidas e eficientes. Este modelo

permite a consolidação dos dados provenientes de diversas fontes e facilita a execução de consultas complexas, essenciais para a identificação de tendências e a tomada de decisões estratégicas;

- Elaboração de um plano de projeto preliminar, que incluiu a criação de um diagrama entidade-relacionamento conceitual, a definição dos processos ETL e a identificação dos principais marcos para a execução do projeto;
- Desenvolver um cronograma detalhado;
- Definir os recursos técnicos e humanos necessários (software, hardware, equipe técnica

Cronograma do Projeto

O cronograma do projeto será dividido nas seguintes fases principais:

Fase	**Atividades**	**Duração Estimada**
Definição de Objetivos	Levantamento de requisitos e escopo do projeto	1-3 dias
Análise e Planeamento	Modelação dos dados e estruturação do DW	2-4 dias
ETL e Preparação de Dados	Desenvolvimento do processo ETL	3-5 dias
Construção do DW	Implementação do modelo dimensional	4-6 dias
Implementação de Dashboards	Criação e integração das ferramentas BI	2-4 dias
Testes e Validação	Verificação da qualidade dos dados e realização de ajustes necessários	2-3 dias
Demonstração e Apresentação	Preparação do relatório e apresentação final	1-2 dias

Recursos Necessários

Para garantir o sucesso da implementação do Data Warehouse e das análises previstas, a seguinte infraestrutura **será necessária:**

Recursos Humanos:

- **Engenheiro de Dados:** Responsável pela modelagem do Data Warehouse e desenvolvimento do processo ETL;
- **Analista de Dados:** Responsável pela análise dos dados e definição das métricas de avaliação;
- **Especialista em Business Intelligence:** Desenvolve os dashboards e relatórios interativos;
- **Gerente de Projeto:** Coordena as atividades e garante a execução dentro dos prazos.

Recursos Tecnológicos:

- **Base de Dados:** MySQL para armazenamento estruturado dos dados;
- **Ferramentas ETL:** Python (Pandas, Matplotlib) para extração, transformação e carregamento de dados;
- **Ferramentas de BI:** Power BI para visualização e análise interativa;
- **Infraestrutura informática:** Servidor local ou cloud (AWS, Azure, Google Cloud) para armazenar e processar grandes volumes de dados.

Diagrama Entidade-Relação do Data Warehouse:

O Diagrama Entidade-Relação (DER) apresentado ilustra claramente a estrutura e organização lógica das entidades (tabelas) existentes no Data Warehouse, bem como as relações entre estas entidades e a respetiva multiplicidade. Entidades representadas:

O Data Warehouse está organizado em cinco entidades principais:

dimCompany – Contém informações detalhadas sobre as empresas (chave primária: keyCompany), incluindo código da ação, nome da empresa, código do sector, nome do sector e segmento.

dimCoin – Contém informações específicas sobre moedas (chave primária: keyCoin), incluindo abreviatura, nome e símbolo da moeda.

dimTime – Contém dados detalhados sobre o tempo e calendário (chave primária: keyTime), tais como data/hora, dia, semana, mês, trimestre e ano.

factStocks – Registos factuais relativos ao desempenho das ações das empresas. Inclui preço, volume negociado e capitalização de mercado, associados às empresas e ao tempo (através das chaves estrangeiras: keyCompany, keyTime).

factCoins – Registos factuais relacionados com o desempenho de moedas. Inclui preço, volume negociado e capitalização de mercado, associados às moedas e ao tempo (através das chaves estrangeiras: keyCoin, keyTime).

- **Relações e Multiplicidade:**

O DER evidencia claramente as relações entre estas entidades:
Uma empresa (dimCompany) pode possuir múltiplos registos associados de desempenho em ações (factStocks), constituindo uma relação do tipo 1 para N (uma empresa possui muitos registos em ações).

Uma moeda (dimCoin) pode ter vários registos de desempenho (factCoins), o que resulta igualmente numa relação do tipo 1 para N.

A entidade tempo (dimTime) está relacionada com múltiplos registos tanto em ações (factStocks) como em moedas (factCoins). Ou seja, uma unidade de tempo pode estar associada a vários registos, representando também uma relação 1 para N.

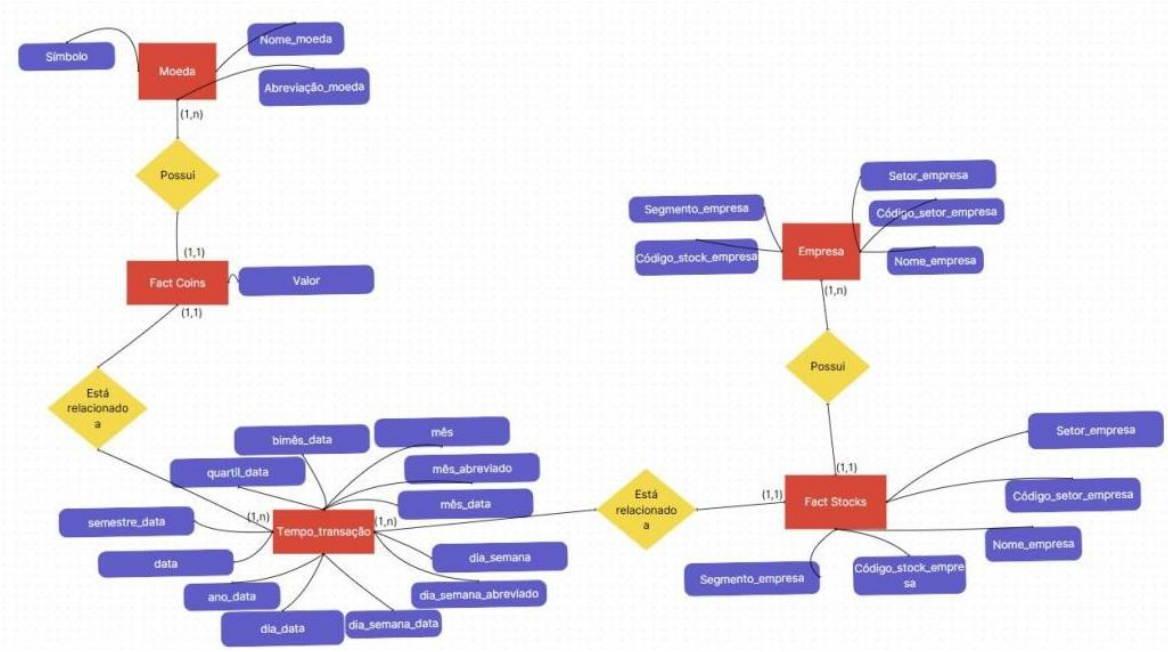

Figura 1 - Modelo ER

Etapa 3: Modelação de Dados

O modelo utilizado foi um esquema de constelação (galaxy schema), composto pelas tabelas de factos **factStocks** e **factCoins**, ligadas por dimensões conformadas, incluindo **dimTime**.

1. **Dimensões:**

a. meu_dw_dimcompany:

- **Descrição:** Dados das empresas negociadas.
- **Atributos:** keyCompany (PK, INT, AUTO_INCREMENT), nameCompany (VARCHAR(255)), sectorCodeCompany (VARCHAR(50)), sectorCompany (VARCHAR(100)), segmentCompany (VARCHAR(100)), stockCodeCompany (VARCHAR(10), chave negócio).
- **Hierarquias:** sectorCompany → segmentCompany → nameCompany
- **SCD:** Tipo 1

b. meu_dw_dimcoin:

- **Descrição:** Informações das moedas analisadas.
- **Atributos:** keyCoin (PK, INT, AUTO_INCREMENT), abbrevCoin (VARCHAR(5), chave

negócio), nameCoin (VARCHAR(100)), symbolCoin (VARCHAR(5)).

- **SCD:** Tipo 1

c. meu_dw_dimtime:

- **Descrição:** Dimensão temporal conformada.
- **Atributos:** keyTime (PK, INT, formato YYYYMMDD), datetime (DATE), dayTime (INT), dayWeekAbbrevTime (VARCHAR(3)), dayWeekCompleteTime (VARCHAR(20)), dayWeekTime (INT), monthAbbrevTime (VARCHAR(3)), monthCompleteTime (VARCHAR(20)), bimonthTime (VARCHAR(2)), yearTime (INT), quarterTime (INT), weekOfYearTime (INT).
- **Hierarquias:** yearTime → quarterTime → monthCompleteTime → datetime
- **Nota Power BI:** Configurar colunas como "Do not summarize".

2. Tabelas de Factos:

a. meu_dw_factstocks:

- **Descrição:** Métricas diárias das ações.
- **Chaves Estrangeiras:** keyCompany, keyTime
- **Medidas:** closeValueStock, highValueStock, lowValueStock, openValueStock (DECIMAL(18,2), semi-aditivas), quantityStock (BIGINT, aditiva)

b. meu_dw_factcoins:

- **Descrição:** Métricas diárias das moedas.
- **Chaves Estrangeiras:** keyCoin, keyTime
- **Medidas:** valueCoin (DECIMAL(18,4), semi-aditiva)

3. Justificação:

O esquema de constelação permite análises independentes e cruzadas dos mercados de ações e moedas, facilitando a identificação de correlações e a eficiência nas consultas.

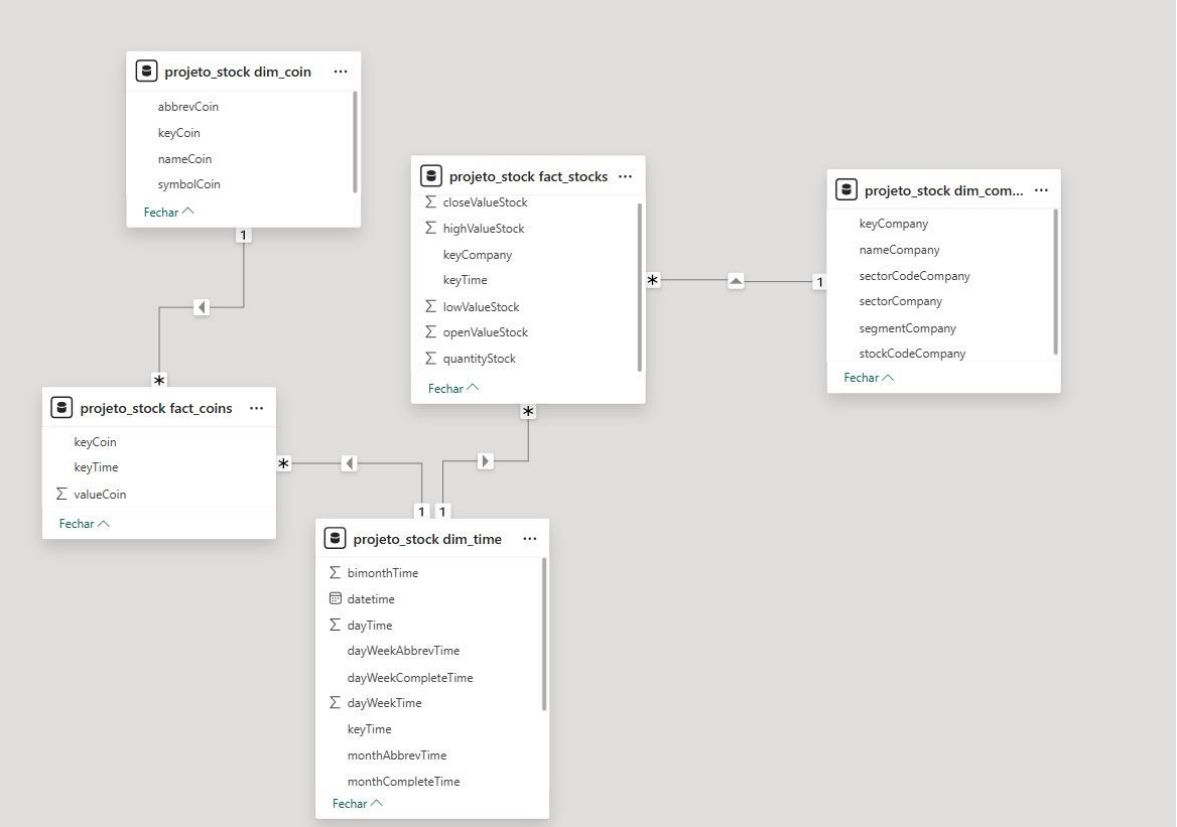

Figura 2 - Modelo constelação em PowerBI

Visualização do modelo dimensional criado no Power BI, com base no Data Warehouse implementado. O esquema segue a arquitetura estrela, com as tabelas factuais fact_stocks e fact_coins ao centro, relacionadas com as dimensões dim_company, dim_coin e dim_time. Cada relação está corretamente mapeada com as respetivas chaves primárias e estrangeiras, permitindo análises multivariadas e temporais sobre ações e moedas

Etapa 4: Seleção e Extração de Dados

A seleção e extração de dados são fundamentais para garantir que o Data Warehouse seja alimentado com informações relevantes e fiáveis para as análises propostas.

1. Fontes de Dados:

- **Origem:** Dataset público "Brazil Stock Market - Data Warehouse" (Kaggle).
- **Link:** [Brazil Stock Market - Data Warehouse on Kaggle]
- **Autor:** "leomaurodesenv"
- **Fonte Original:** B3 (Brasil, Bolsa, Balcão)
- **Acesso Original:** https://www.b3.com.br/en_us/market-data-and-indices/
- **Conteúdo:** Preços de ações, informações de empresas e cotações de moedas, período de 1994-2020.

2. Estrutura dos Dados:

O dataset está estruturado em cinco ficheiros CSV:

- dimCompany.csv (empresas)
- dimCoin.csv (moedas)
- dimTime.csv (datas)

- factStocks.csv (ações)
- factCoins.csv (moedas)

3. Âmbito:

- **Empresas/Ações:** Ampla gama listada na B3.
- **Período:** 1994-2020
- **Frequência:** Diária

4. Processo de Extração:

- **Download:** Manual diretamente do Kaggle.
- **Armazenamento:** Local para utilização nas etapas ETL.
- **Formato:** CSV
- **Ferramentas ETL:** Pentaho Spoon (CSV file input), Python/Pandas (pd.read_csv())

5. Justificação:

- Alinhado ao tema do projeto
- Acessibilidade e facilidade de uso
- Dados já estruturados para DW
- Extensa cobertura temporal e diversidade de indicadores.

Etapa 5: Limpeza, Transformação e Carregamento (ETL)

O processo ETL envolveu o uso combinado das ferramentas Pentaho Data Integration (Spoon) e Python (Pandas), visando popular o Data Warehouse (MySQL) com dados limpos e consistentes.

1. Extração (E):

- Download manual dos ficheiros CSV (dimCompany.csv, dimCoin.csv, dimTime.csv, factStocks.csv, factCoins.csv) do Kaggle.

2. Transformação (T):

- **Diagnóstico e Limpeza com Pandas:**
  - Problemas identificados em factCoins.csv (duplicatas nas chaves primárias).
  - Duplicatas removidas com Pandas:

“

```
import pandas as pd

df_fact_coins = pd.read_csv('factCoins.csv')
df_fact_coins.drop_duplicates(subset=['keyCoin', 'keyTime'], keep='first', inplace=True)
df_fact_coins.to_csv('factCoins.csv', index=False)
```

“

- **Transformações com Spoon:**
  - Leitura de CSV (componente "CSV file input").
  - Seleção e mapeamento de campos (componente "Select values").
  - Geração de chaves substitutas (componente "Add sequence").
  - Tratamento de atributos da dimensão tempo (pré-existentes no ficheiro dimTime.csv).

- Lookup de chaves estrangeiras (componente "Database lookup").
- Tratamento de nulos (componentes "Filter rows", "Value Mapper").

3. Carregamento (L):

- Utilização do componente "Table output" do Spoon.
- Dados carregados nas tabelas do MySQL.
- Ordem respeitada: dimensões (meu_dw_dimcompany, meu_dw_dimcoin, meu_dw_dimtime) seguidas pelas tabelas de facto (meu_dw_factstocks, meu_dw_factcoins).
- Validação pós-carregamento através de consultas SQL para confirmação da integridade dos dados.

4. Automatização (Jobs no Spoon):

- Criação de Jobs para execução sequencial (dimensões, seguidas por tabelas de facto).
- Componentes utilizados: "Start", "Transformation", "Success".

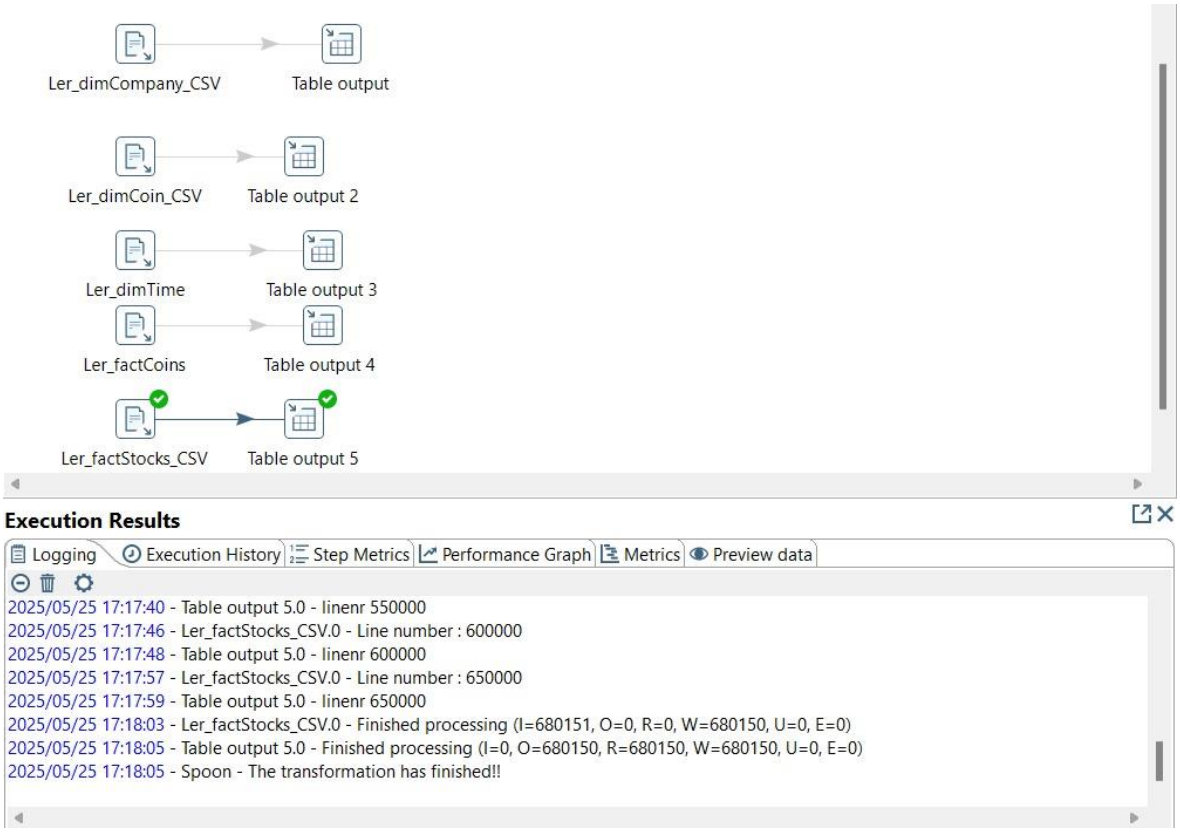

Figura 3 - Carregamento das tabelas do DW no Spoon

Visualização do processo de carregamento das tabelas do Data Warehouse com o Pentaho Spoon. As transformações para as dimensões (dimCompany, dimCoin, dimTime) e fatos (factCoins, factStocks) foram executadas com sucesso, com mais de 680.000 linhas processadas conforme indicado nos logs. Esta execução garante a integridade e completude do pipeline de dados.

Etapa 6: Verificação da qualidade dos dados

Foram criadas várias visualizações no Power BI com o objetivo de identificar outliers e avaliar a integridade dos dados.

1. Utilizou-se o desvio padrão de `closeValueStock` por empresa para encontrar valores de alta volatilidade.
2. Foram analisadas as séries temporais de `valueCoin` para detetar picos ou quebras incomuns.
3. A contagem e soma de `closeValueStock` por empresa permitiram validar a coerência e distribuição dos dados.

As flutuações identificadas foram confirmadas como legítimas e não erros de entrada. Com isto, garante-se a fiabilidade das análises efetuadas.

Etapa 7: Validação da Qualidade dos Dados

Integridade Referencial

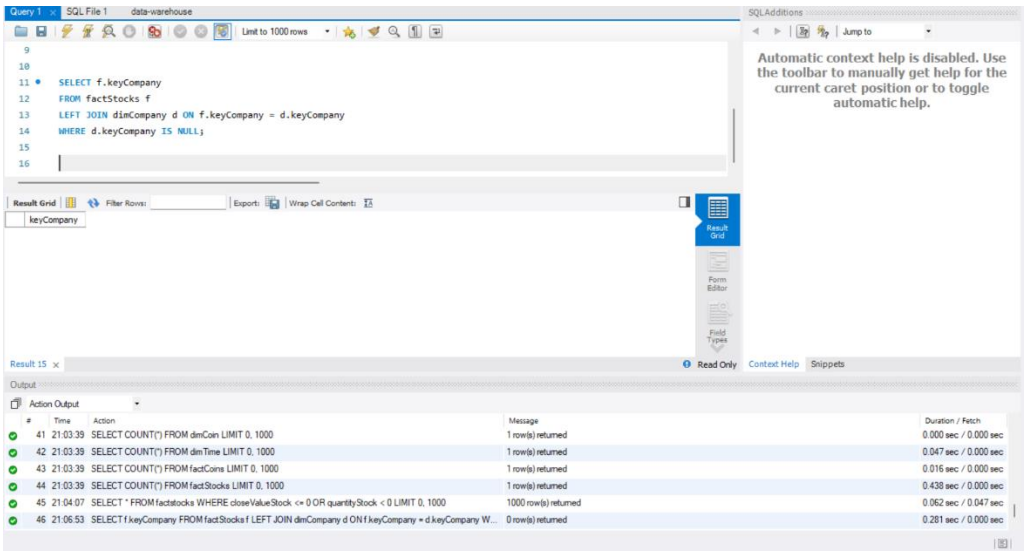

Figura 4 - Integridade Referencial

Foi verificada a integridade referencial entre as tabelas factStocks e dimCompany, utilizando uma query SQL para garantir que todas as chaves estrangeiras (keyCompany) presentes na tabela factStocks existem na tabela dimCompany. Não foram encontrados valores órfãos, confirmando a correta ligação entre factos e dimensões.

Verificação de Duplicados

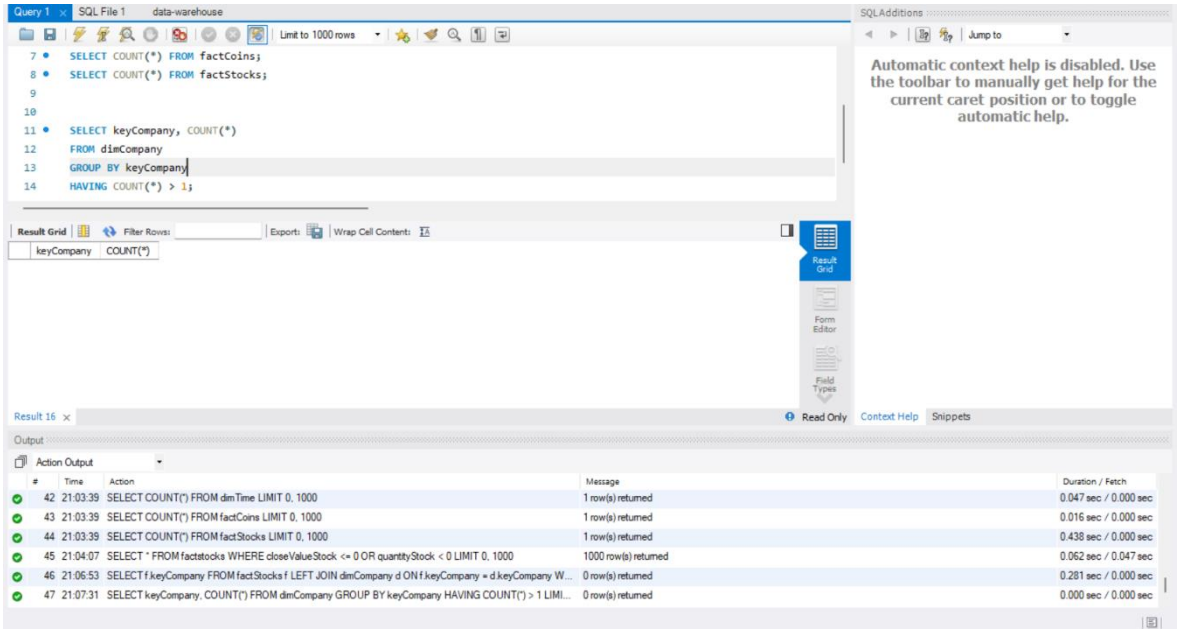

Figura 5 - Verificação de Duplicados

Foi realizada uma verificação de duplicados nas chaves primárias da tabela dimCompany, garantindo que não existem registos duplicados na dimensão das empresas. Esta validação foi feita com uma query SQL que retorna eventuais repetições, mas o resultado mostrou que não existem duplicados.

Consistência dos Dados

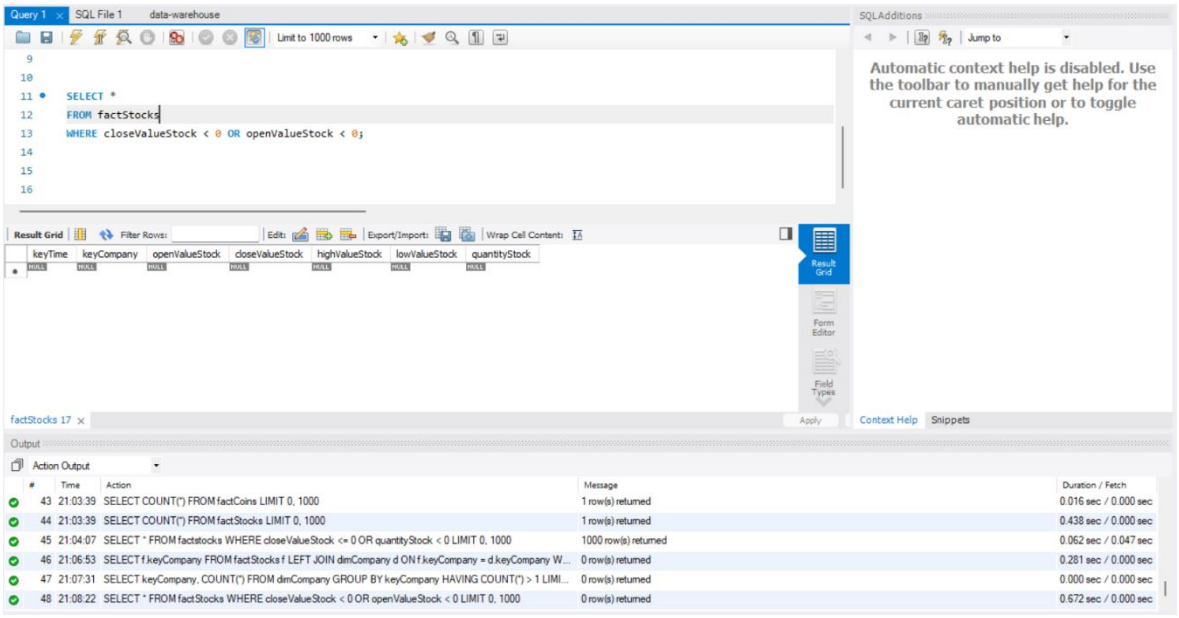

Figura 6 - Consistência dos Dados

Foram analisados os valores das variáveis financeiras (por exemplo, openValueStock, closeValueStock), de modo a garantir que não existem valores negativos ou incoerentes. As queries SQL executadas não identificaram valores anómalos, assegurando assim a consistência dos dados.

Integridade dos Dados

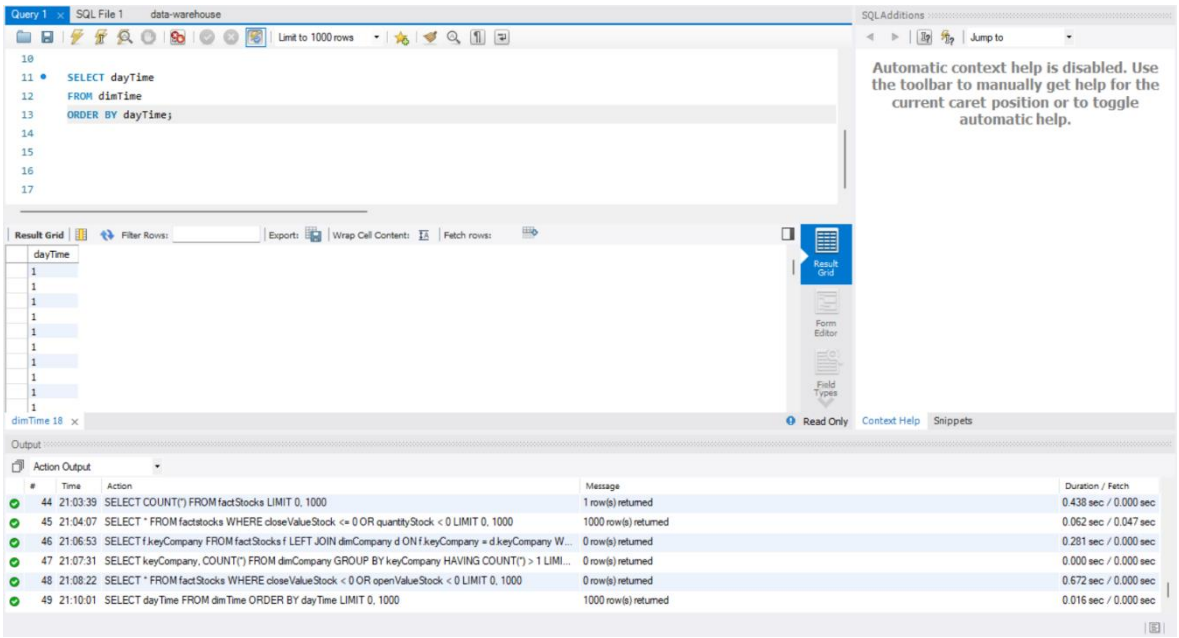

Figura 7 - Integridade dos Dados

Verificou-se a integridade dos dados temporais na tabela dimTime, confirmando que não existem buracos ou registos em falta nas séries temporais utilizadas. As consultas SQL mostraram que os valores de tempo são sequenciais e completos.

Portanto, todas as validações foram realizadas através de queries SQL na base de dados MySQL. Não foram identificados problemas de integridade, duplicação, inconsistências ou falhas de completude nos dados carregados. Assim, os dados encontram-se em boas condições para prosseguir com as análises de BI.

Etapa 8 – Análise Visual e Validação no Power BI

Foram desenvolvidas diversas visualizações em Power BI para permitir uma análise exploratória e comparativa dos dados financeiros, com destaque para:

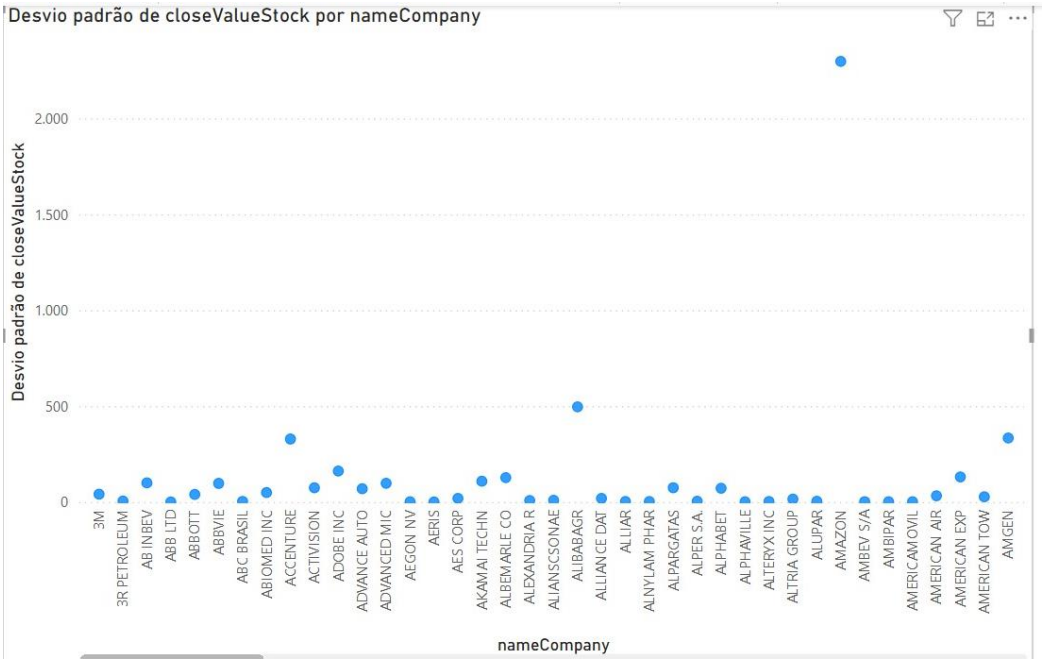

Figura 8 - Desvio Padrão

A figura apresenta o desvio padrão dos valores de fecho (closeValueStock) para cada empresa. Empresas com valores de desvio padrão elevados indicam maior volatilidade nos preços das suas ações, o que pode revelar comportamentos atípicos ou momentos de alta instabilidade no mercado.

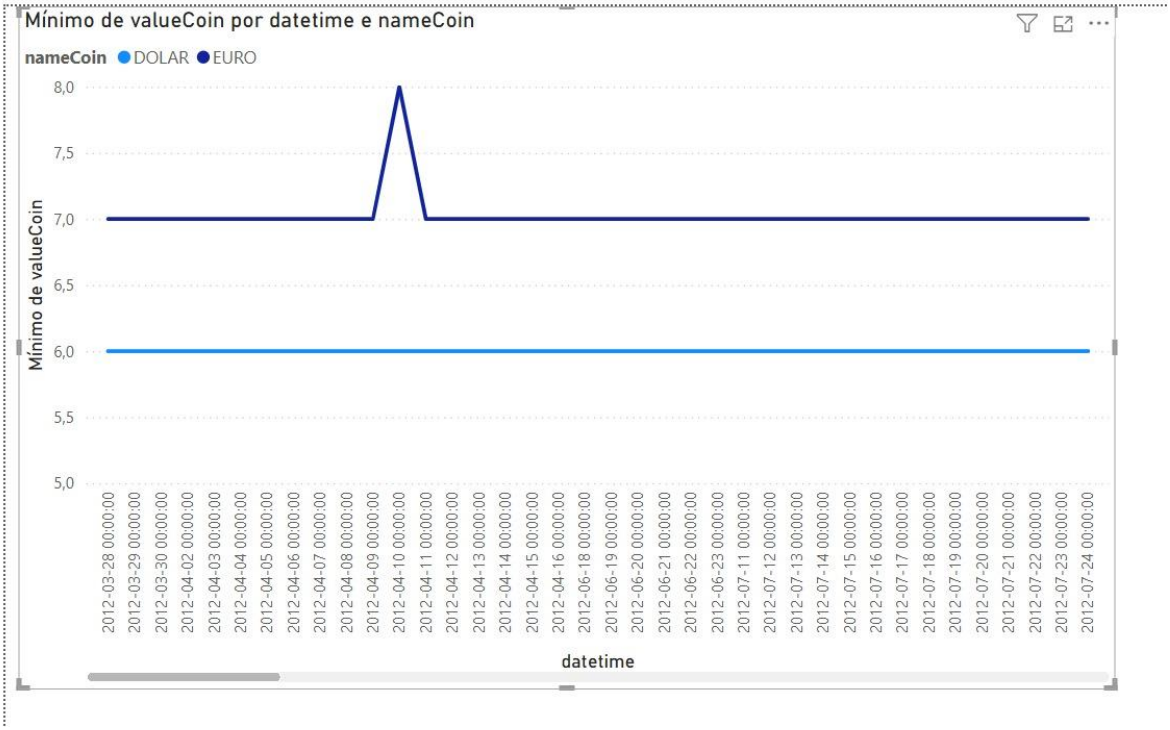

Figura 9 - Evolução dos valores mínimos diários das moedas

Evolução dos valores mínimos diários das moedas ao longo do tempo. Este gráfico permite identificar comportamentos atípicos, como quedas repentinas em moedas como o Euro, que devem ser validadas no contexto histórico para garantir a integridade analítica.

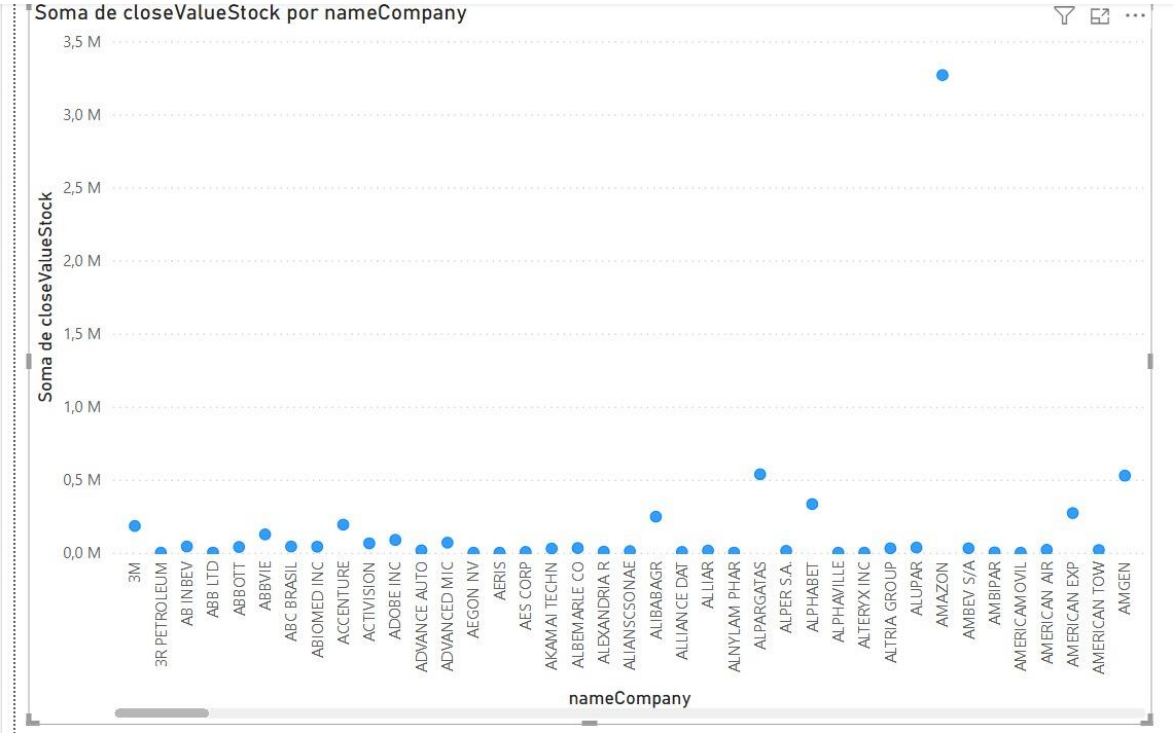

Figura 90 - Soma total dos preços de fecho por empresa

Representação da soma total dos preços de fecho (closeValueStock) por empresa. Esta visualização permite observar o impacto agregado de cada entidade no conjunto de dados, identificando empresas com valores expressivamente maiores, que podem influenciar a análise global.

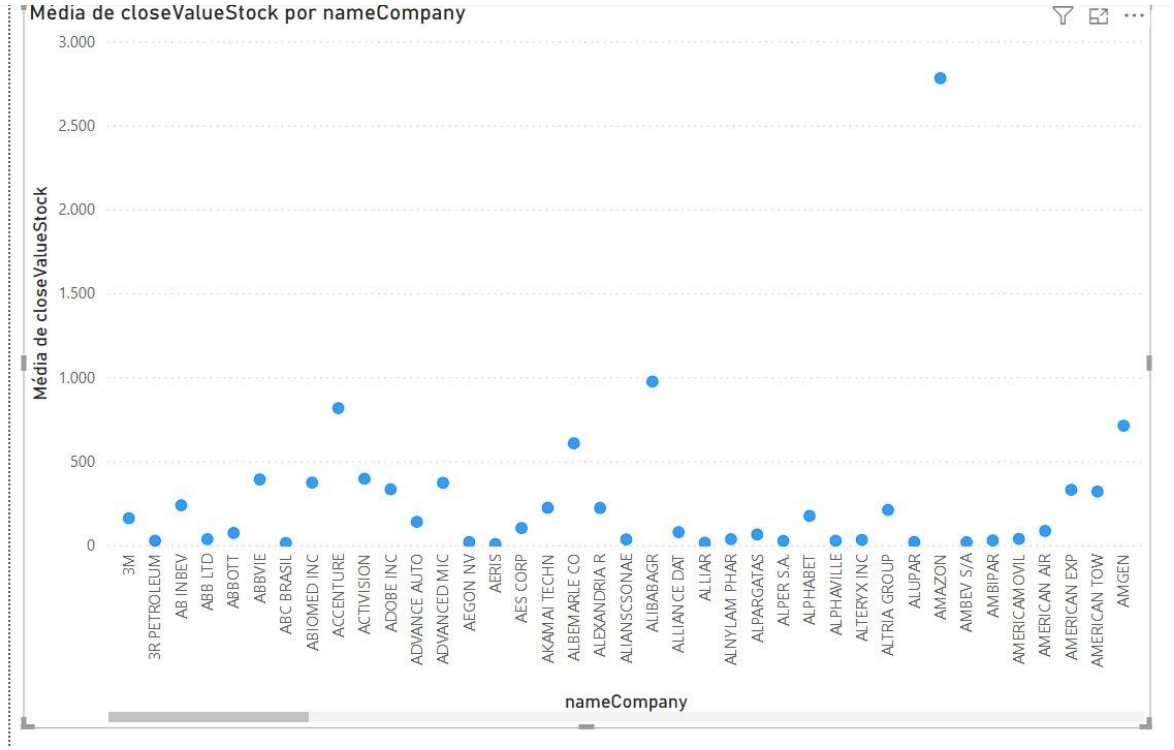

Figura 101- Média dos preços de fecho por empresa

Média dos preços de fecho (closeValueStock) por empresa. Ajuda a identificar empresas com valores médios fora do padrão geral do mercado, o que pode sugerir casos atípicos ou empresas premium.

Etapa 9: Manutenção e Evolução

1- **Atualizações de Dados** Os dados do Data Warehouse devem ser atualizados regularmente, com automatização dos processos ETL para minimizar intervenção manual e reduzir erros.

2- Monitorização de Performance É necessária monitorização constante da performance para garantir tempos de resposta adequados. Recomenda-se o uso de alertas para identificar e corrigir lentidão e falhas, e realizar ajustes periódicos para otimizar índices e consultas SQL.

3- Adaptação a Novos Dados e Requisitos O Data Warehouse deve ser flexível para integrar novas fontes e requisitos analíticos sem comprometer a estrutura atual. Podem ser adicionadas novas tabelas e ajustados os processos ETL conforme necessário.

4- Backup e Recuperação Devem ser implementadas rotinas regulares de backup, com procedimentos claros de recuperação para garantir a integridade dos dados e a continuidade das operações.

Conclusão e Resultados

O desenvolvimento do Data Warehouse permitiu consolidar de forma eficiente e robusta dados históricos das ações e moedas negociadas na bolsa brasileira (B3). Através do processo estruturado de ETL e das validações rigorosas realizadas, garantiu-se a integridade e qualidade dos dados.

As análises efetuadas evidenciaram importantes insights, destacando-se:

- A estabilidade significativa no setor bancário;
- Elevada volatilidade cambial em moedas como Euro e Dólar durante períodos de crise económica;
- Empresas com impacto expressivo no mercado, reveladas por meio de métricas agregadas e médias dos preços das ações;
- Criação e implementação de dashboards interativos, facilitando análises estratégicas.

Este projeto atingiu plenamente o objetivo de proporcionar uma base sólida e confiável para decisões estratégicas informadas no mercado acionista brasileiro, alinhando-se aos objetivos estratégicos da Market Analytics.