Data Lake Olímpico - Análise Completa dos Jogos Olímpicos

Arquitetura RAW → BRONZE → GOLD

Resumo Executivo

Este projeto implementa uma **arquitetura moderna de Data Lake** para análise abrangente dos Jogos Olímpicos, integrando dados históricos da World Olympedia com informações oficiais de Paris 2024. O objetivo é demonstrar boas práticas de engenharia de dados e gerar insights estratégicos sobre a evolução olímpica através de análises estatísticas rigorosas.

Datasets Integrados

Fonte	Período	Registros	Descrição
World Olympedia	1896-2020	155.861 atletas	Dados históricos completos
Paris 2024	2024	11.113 atletas	Dados oficiais dos jogos
Paris 2024	2024	2.315 medalhas	Resultados completos
Total	**1896-2024**	**169.289 registros**	**128 anos de dados**

Arquitetura do Data Lake

Camadas Implementadas

```
raw/          # Dados brutos originais + metadados JSON
bronze/       # Dados processados e otimizados (Parquet)
gold/         # Análises finais e visualizações
```

RAW Layer - Dados Brutos

- **24 arquivos** de dados originais em formato CSV
- **Metadados JSON** descritivos para cada dataset
- **Preservação** do formato original para auditoria
- **Cobertura completa** de 13 datasets de Paris 2024

BRONZE Layer - Dados Processados

- **20 arquivos** convertidos para formato Parquet otimizado
- **Limpeza e padronização** de dados
- **Integração** entre diferentes fontes
- **Metadados técnicos** estruturados

GOLD Layer - Análises e Insights

- **13 arquivos** de análises finais
- **Visualizações profissionais** em alta resolução
- **Relatórios executivos** em formato JSON
- **Dashboard consolidado** com 6 visualizações

Questões Analíticas Respondidas

1. Evolução da Distribuição de Medalhas por País (1986-2024)

Metodologia: Análise de correlação entre dados históricos e Paris 2024, com estatísticas descritivas completas.

Principais Descobertas: - **Estados Unidos** mantêm liderança com 5.249 medalhas totais - **Correlação forte** (r=0.756) entre tradição histórica e performance atual - **Top 5 países** concentram 65% das medalhas analisadas - **Efeito país-sede** beneficiou significativamente a França

2. Crescimento de Modalidades em Participação (1986-2024)

Metodologia: Análise de 55 modalidades com cálculo de quartis e distribuição estatística.

Principais Descobertas: - **Atletismo domina** com 2.018 participantes (18% do total) - **Distribuição desigual**: Top 10 modalidades concentram 70% dos atletas - **Modalidades tradicionais** mantêm alta participação global - **55 modalidades diferentes** garantem diversidade olímpica

3. Evolução da Proporção por Sexo nas Modalidades (1980-2024)

Metodologia: Análise temporal de 5 décadas com cálculo de percentuais e boxplots.

Principais Descobertas: - **Crescimento significativo**: Participação feminina de 25% (1980) para 46.2% (2020) - **Progresso consistente** sem retrocessos ao longo das décadas - **35+ modalidades** alcançaram paridade (40-60% feminino) em Paris 2024 - **Tendência de equilíbrio** crescente entre gêneros

---

Tecnologias e Metodologia

Stack Tecnológico

- **Python**: Linguagem principal para processamento e análise
- **Pandas**: Manipulação e transformação de dados
- **Parquet**: Formato otimizado para analytics de alta performance
- **Matplotlib/Seaborn**: Visualizações profissionais e estatísticas
- **JSON**: Metadados estruturados com schema técnico
- **Jupyter**: Análise interativa e relatórios executivos

Boas Práticas Implementadas

- **Arquitetura em camadas** para separação de responsabilidades
- **Metadados completos** para governança de dados
- **Formato Parquet** para otimização de consultas
- **Versionamento Git** para controle de mudanças
- **Documentação técnica** abrangente

Como Executar

Pré-requisitos

```
pip install -r requirements.txt
```

Pipeline Completo

```
# Pipeline principal (RAW → BRONZE → GOLD)
python enhanced_pipeline.py

# Correção de gráficos (se necessário)
python final_pipeline.py

# Análise interativa
jupyter notebook olympics_final_report.ipynb
```

Estrutura de Execução

1. **RAW**: Criação automática de metadados para dados brutos
2. **BRONZE**: Processamento e conversão para Parquet otimizado
3. **GOLD**: Geração de análises estatísticas e visualizações

---

Arquivos Principais

Scripts de Processamento

- `enhanced_pipeline.py` - Pipeline principal com análises completas
- `final_pipeline.py` - Correções e refinamentos finais
- `download_paris2024.py` - Download automático dos dados Paris 2024

Relatórios e Análises

- `olympics_final_report.ipynb` - **Relatório executivo completo**
- `relatorio_completo.json` - Resumo técnico das análises
- `dashboard_corrected.png` - Dashboard executivo com 6 visualizações

Configuração

- `requirements.txt` - Dependências Python
- `metadata_schema.json` - Schema técnico dos metadados
- `README.md` - Documentação completa do projeto

Resultados e Visualizações

Dashboard Executivo

Visualização consolidada com 6 gráficos integrados: - Distribuição de medalhas por país (pizza) - Top modalidades por participação (barras) - Evolução histórica por gênero (área) - Correlação histórico vs atual (scatter) - Distribuição de participantes (histograma) - Paridade de gênero (boxplot)

Análises Específicas

- **Medalhas por país**: Gráfico de barras + correlação
- **Crescimento de modalidades**: Top 15 + distribuição estatística
- **Evolução por gênero**: 4 visualizações integradas

Estatísticas Consolidadas

- **169.289 registros** processados
- **55 modalidades** analisadas
- **20 países** no ranking principal
- **5 décadas** de evolução histórica

Próximos Passos

Expansões Recomendadas

- Incluir dados de Los Angeles 2028 quando disponíveis
- Adicionar métricas de performance (tempos, recordes)
- Implementar análises preditivas baseadas em machine learning
- Desenvolver dashboard interativo com Streamlit/Dash

Melhorias Técnicas

- Automatização do pipeline com Apache Airflow
- Implementação de testes unitários
- Integração com cloud storage (AWS S3)
- API REST para consulta de dados

Contribuição e Licença

Demonstração de Competências

Este projeto demonstra: - **Arquitetura moderna** de Data Lake - **Boas práticas** de engenharia de dados - **Análises estatísticas** rigorosas - **Visualizações profissionais** de alta qualidade - **Documentação técnica** completa

Estrutura Final

- **57 arquivos** organizados em camadas
- **100% cobertura** de metadados
- **Código versionado** e documentado
- **Análises reproduzíveis** e auditáveis

Projeto desenvolvido como demonstração de competências em Ciência de Dados e Engenharia de Dados

Data Lake Olímpico - Transformando dados em insights estratégicos