

**BlueShift Academy**

PROJETO INTEGRADO

Avaliação da correlação entre o desempenho dos atletas nas olimpíadas frente ao investimento realizado por estado e o IDH (Índice de Desenvolvimento Humano)

2021

São Paulo

Projeto Integrado.

Responsável líder: Henrique Dantas Marques

Responsáveis técnicos: Marcelo Pereira, Marcia Tamazoto Longhi, Rafael Thoaldo Souza, Vinicius Alexandre de Andrade, Wilton Simões de Oliveira.

Aprovadores:

|  |  |  |
| --- | --- | --- |
| Nome | Cargo/Função | Assinatura |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Histórico de revisões

|  |  |  |  |
| --- | --- | --- | --- |
| Versão | Data | Autor | Descrição de mudança |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Sumário**

[1. Introdução 4](#_Toc82422828)

[2. Solicitação 4](#_Toc82422829)

[3. Premissas da Solução 5](#_Toc82422830)

[3.1 Origem e especificação dos dados 5](#_Toc82422831)

[3.2 Ambiente de desenvolvimento 5](#_Toc82422832)

[4. Modelo da arquitetura sugerida 6](#_Toc82422833)

[5. Extração dos arquivos 6](#_Toc82422834)

[5.1 Dados da Bolsa Atleta 7](#_Toc82422835)

[5.2 Dados do IDH 8](#_Toc82422836)

[5.3 Dados de medalhas conquistadas nas Olimpíadas de Tóquio 9](#_Toc82422837)

[6. Dashboards 10](#_Toc82422838)

[7. Desafios 10](#_Toc82422839)

[8. Conclusão 11](#_Toc82422840)

1. Introdução

Este documento tem por finalidade descrever o desenvolvimento e a evolução dos procedimentos adotados para o Projeto Integrado “Olimpíadas - Investimentos nos esportes e a correlação com o índice de desenvolvimento humano “.

1. Solicitação

Com o evento das Olimpíadas de Tóquio, a decisão por um tema atual foi o ponto crucial para decisão e direção dos estudos e análises, focando nos benefícios distribuídos no país para a área de esportes, e também analisar como e se o índice de desenvolvimento humano exerce alguma influência nessa área.

As bases de dados consideradas para os estudos do projeto foram pesquisadas nos arquivos disponibilizados para:

* Beneficiados pelo programa ‘Bolsa Atleta’;
* PNUD - Programa das Nações Unidas para o Desenvolvimento;
* Medalhas conquistadas nas Olimpíadas de Tokio.

Para o desenvolvimento do projeto, os trabalhos foram divididos em etapas:

* Análise dos sites que com dados necessários para o projeto bem como a definição da coleta dos mesmos (API, *Webscraping*);
* Consumo dos dados com script em Python para tratamento e seleção das informações relevantes para o processo;
* Complementação de dados com informações referente a Unidade Federal, visto que as análises foram concentradas a nível estadual;
* Esses dados foram disponibilizados no Blob Storage do Azure para ingestão no Power Bi.

Para viabilizações e estudos do projeto, foram consideradas as informações referentes a:

* Valores dos benefícios do programa Bolsa Atleta distribuídos no estado por tipo de benefício e modalidade de esporte;
* Índices de desenvolvimento humano (IDH) adquirido por estado, geral, renda, educação e longevidade;
* Resultado da Olimpíada de Tóquio conquistada pelos atletas nacionais.

1. Premissas da Solução

## **3.1 Origem e especificação dos dados**

Os datasets foram extraídos dos repositórios públicos:

**Links Bolsa Atleta:**

“https://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?data=13/05/2021&jornal=515&pagina=43&totalArquivos=412”

"https://www.in.gov.br/en/web/dou/-/portaria-mc-n-629-de-12-de-maio-de-2021-319600420"

"https://www.in.gov.br/en/web/dou/-/portaria-mc-n-629-de-12-de-maio-de-2021-anexo-unico-continuacao-319559187"

**Índice IDH:** https://www.br.undp.org/content/brazil/pt/home/idh0rankings/idhm-uf-2010.html

**Lista UF’s:** "https://mundoeducacao.uol.com.br/geografia/estados-brasil.htm"

**Quadro de medalhas:** “https://ge.globo.com/olimpiadas/noticia/quantas-medalhas-o-brasil-ja-ganhou-nas-olimpiadas-2020-veja-quadro.ghtml”

## **3.2 Ambiente de desenvolvimento**

Para o desenvolvimento do projeto, foram necessários liberação de acessos, ferramentas e os ambientes de desenvolvimento da empresa Blueshift:

* Azure Databricks Microsoft
* Azure Blob Storage (Containers)
* Power BI

1. Modelo da arquitetura sugerida

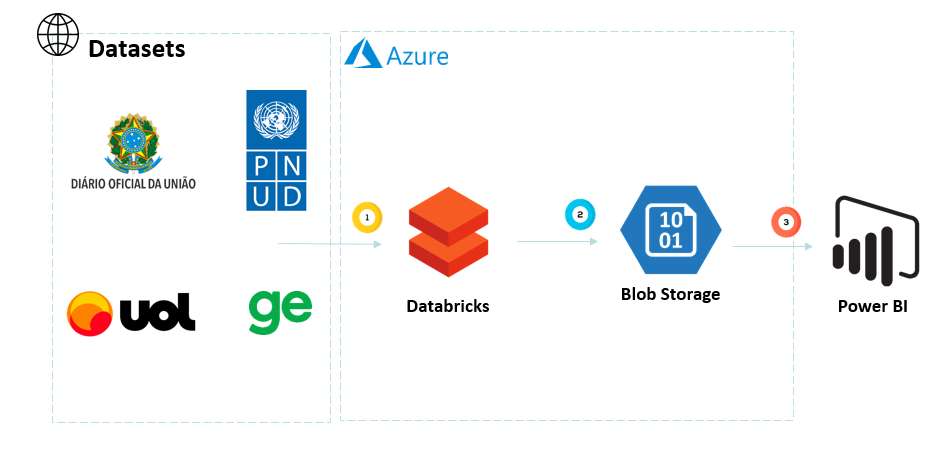
Com a base de dados selecionada para o processo, a solução para o desenvolvimento do projeto foi elaborada conforme demonstrado nas sequências da figura abaixo.

Figura 1 - Arquitetura do projeto integrado

**Fonte:** Própria autoria.

Descrição da arquitetura:

* Requisição dos dados dos sites públicos por meio da biblioteca Koalas e tratamento dos dados com scripts em Pyspark;
* Transferência dos datasets para um contêiner do Blob Storage em formato .csv;
* Visualização dos dados por meio de *dashboards* no Power BI.

1. Extração dos arquivos

A extração dos arquivos utilizados no projeto foi extraída e submetida a scripts Spark para tratamento das informações conforme descrito no projeto incluso no Github: “https://github.com/HenriqueMarques27/ProjetoIntegrado”.

## **5.1** **Dados da Bolsa Atleta**

Os dados foram extraídos das bibliotecas abaixo e tratados com script em Spark como descrito na Figura 2.

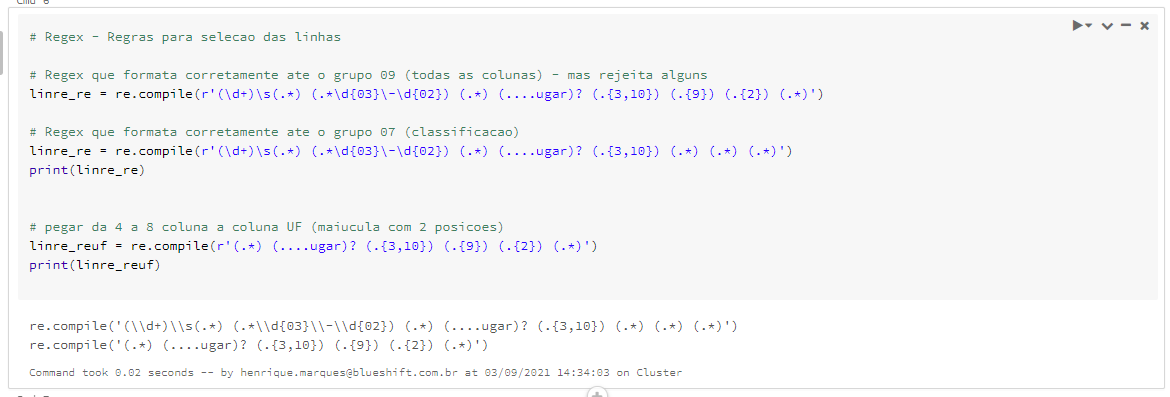
Figura 2 – Requisição do Diário Oficial da União





**Fonte:** Própria autoria.

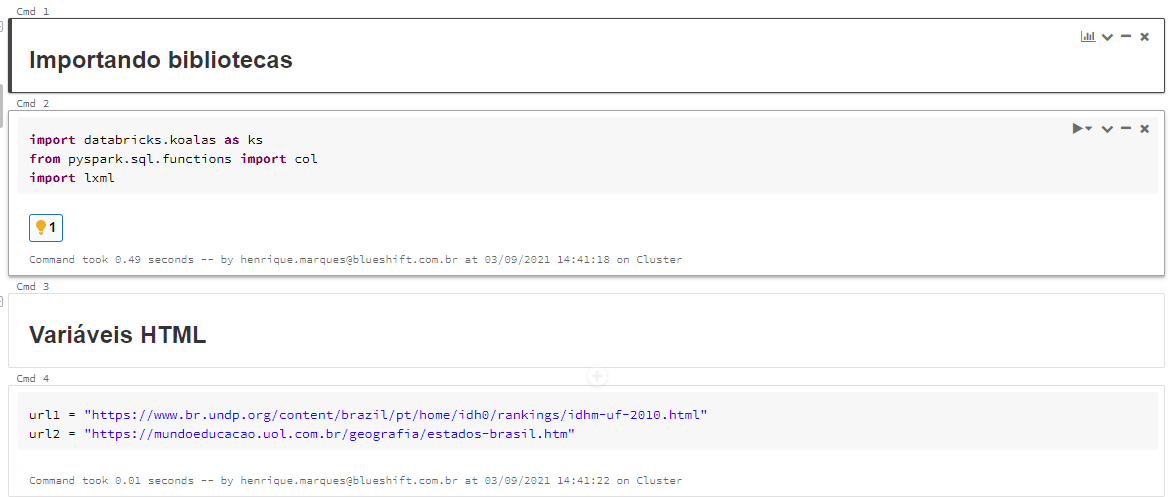
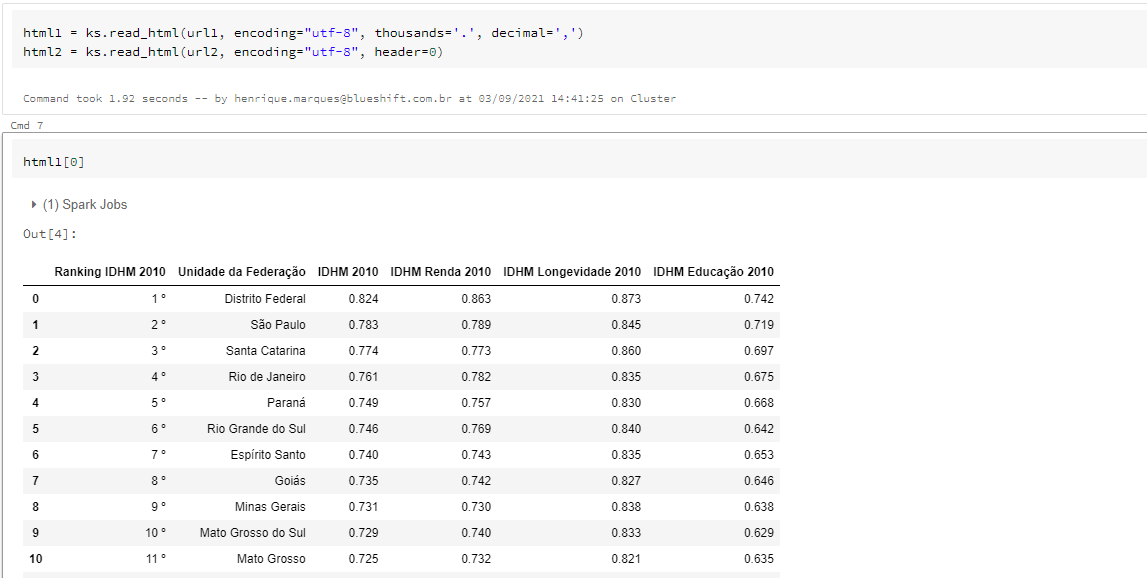
Identificamos que haviam dados faltantes no processo, esses dados foram extraídos e gravados em arquivo no formato .pdf, foram tratados com a linguagem *Python* por meio da função Regex (Figura 3) para posterior complementação com os dados processados no passo anterior.

Figura 3 – Extração pela função Regex

**Fonte:** Própria autoria.

## **5.2** **Dados do IDH**

Os dados foram extraídos do site público do Programa das Nações Unidas para o Desenvolvimento (“url1”) como descrito na Figura 4. As informações referentes a Unidade Federativa da biblioteca abaixo foram complementadas com a requisição do site do UOL (“url2”)

Figura 4 – Extração dos Índices de Desenvolvimento Humano

**Fonte:** Própria autoria.

## **5.3** **Dados de medalhas conquistadas nas Olimpíadas de Tóquio**

Os dados do quadro de medalhas foram extraídos do site do Globo Esporte a requisição com a biblioteca do Koalas e script Spark está descrita na Figura 5.

Figura 5 – Requisição do Globo Esporte das medalhas conquistadas

**Fonte:** Própria autoria.

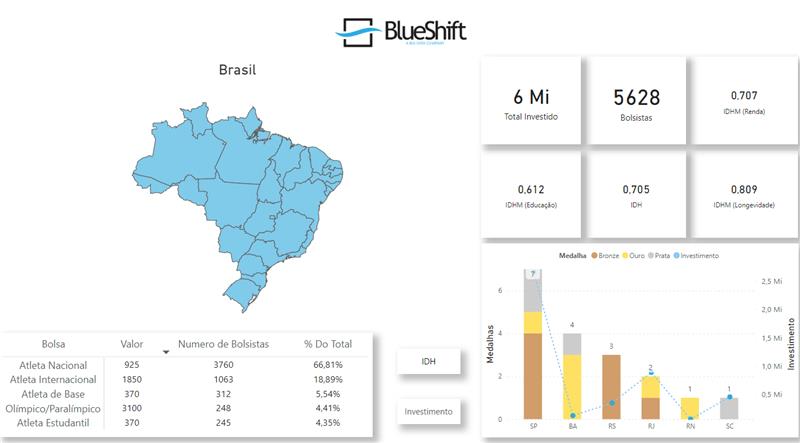
**5.4 Ingestão dos arquivos no Azure Blob Storage**

Os arquivos com os dados tratados e complementados foram enviados para um contêiner do Blob Storage do projeto (Figura 6) para que os dados pudessem ficar disponíveis para serem consumidos pelo Power BI para construção dos dashboards.

Figura 6 – Envio dos dados tratados para o Blob Storage

Fonte: Própria autoria.

1. Dashboards

Figura 7 – Dashboard no Power BI

Fonte: Própria autoria.

1. Desafios

O desenvolvimento deste projeto foi muito rico em estudo e aprendizagem e também contou com muitos desafios, desde o início do mesmo. As principais dificuldade encontradas estão descritas a seguir.

* Definição do tema:

A escolha do tema para o projeto foi uma das etapas mais difíceis, pois quando o tema era definido, não haviam datasets com as informações necessárias para o desenvolvimento, ou quando escolhemos um tema e encontramos os datasets, verificamos que já havia um projeto semelhante disponível.

* Obtenção e disposição dos dados:

Identificamos que os datasets disponíveis com os dados nacionais não estavam disponibilizados no formato necessário para extração e tratamento das informações.

* Extração e tratamentos de dados no formato .pdf:

A etapa de extração e tratamento de dados dos arquivos no formato .pdf foi concluída após tentativas com vários scripts por Regex.

* Automatização dos processos:

A automatização dos processos de extração dos dados foi efetuada com várias adaptações em relação a extração dos dados dos datasets.

1. Conclusão

Analisando as informações demonstradas no gráfico, verificamos:

* A concentração de investimento e número de atletas foi maior no estado de São Paulo em relação ao país; o estado teve uma contribuição significativa no retorno dos investimentos, devido ao maior número de medalhas conquistadas nas Olimpíadas;
* Os estados da Bahia e Rio Grande do Norte tem investimentos e índice de IDH menores, mas tiveram um retorno melhor com a conquista de medalhas;
* Alguns estados com IDH alto também não tiveram retorno de medalhas.

Chegamos à conclusão de que não existe a correlação entre o desempenho dos atletas nas Olimpíadas frente aos investimentos realizados por estado e o IDH (Índice de Desenvolvimento Humano).