IPCA INSTITUTO POLITÉCNICO DO CÁVADO E DO AVE

ANÁLISE DE DADOS DE F1 (2016-2024)

INTEGRAÇÃO DE SISTEMAS DE INFORMAÇÃO

Aluno:

Vítor Leite (a25446)

Docente:

Luís Ferreira

ENGENHARIA DE SISTEMAS INFORMÁTICOS

Outubro, 2025

IPCA INSTITUTO POLITÉCNICO DO CÁVADO E DO AVE

ANÁLISE DE DADOS DE F1 (2016-2024)

INTEGRAÇÃO DE SISTEMAS DE INFORMAÇÃO

Aluno:

VÍtor Leite (a25446)

Docente:

Luís Ferreira

ENGENHARIA DE SISTEMAS INFORMÁTICOS

Outubro, 2025

Resumo

Este projeto apresenta o desenvolvimento de um processo completo de integração e transformação de dados (ETL) aplicado ao contexto da Fórmula 1, recorrendo à ferramenta Pentaho Data Integration (PDI). O principal objetivo foi consolidar informação proveniente de diferentes fontes de dados, de forma a gerar estatísticas fiáveis sobre o desempenho dos pilotos entre as temporadas de 2016 e 2024.

O trabalho envolveu a extração de dados a partir de ficheiros CSV obtidos no Kaggle, o seu tratamento e cruzamento através de operações de junção e filtragem, e a posterior agregação de indicadores como total de pontos, número de corridas e médias de posições. O resultado final foi a criação de ficheiros de saída em múltiplos formatos — Excel, Texto e JSON.

Este projeto demonstra a aplicabilidade prática das ferramentas ETL na área da análise de dados, evidenciando a importância da automatização e da integração de fontes distintas para a produção de informação consolidada e de valor analítico.

Índice

Introdução

O projeto recorreu a diversas etapas de transformação no Pentaho, incluindo uniões (joins) entre tabelas de resultados, corridas, pilotos e construtores, bem como operações de limpeza e normalização de dados.

Após o pré-processamento, os dados foram agrupados por piloto, somando os pontos e o número de corridas, e calculando as médias das posições de largada e chegada.

O resultado final é exportado em vários formatos (Excel, TXT e JSON), permitindo o seu uso em relatórios, dashboards ou outras ferramentas analíticas.

Este fluxo de transformação possibilita uma análise consistente, automática e facilmente atualizável, bastando substituir os ficheiros de entrada por dados de novas temporadas.

Objetivos do Trabalho

• Criar um processo ETL completo usando o Pentaho PDI.

• Extrair, transformar e carregar dados da Fórmula 1.

• Gerar ficheiros de saída nos formatos Excel, TXT e JSON.

Processo ETL

A implementação do processo ETL foi efetuada em várias etapas distintas, desde a extração dos dados até à geração dos resultados finais.

Extração

O processo inicia-se com a extração de dados brutos provenientes de quatro ficheiros CSV distintos:

Results.csv – contém os resultados de cada corrida, indicando o piloto, posição, pontos, voltas, tempos, entre outros campos.

Races.csv – descreve as corridas, com informações como o ano, o circuito e o identificador único de cada corrida (raceId).

Drivers.csv – armazena os dados dos pilotos, incluindo o seu nome, número, nacionalidade e identificador (driverId).

Constructors.csv – contém as informações das equipas (construtores), com o seu nome e identificador (constructorId).

Cada um destes ficheiros é lido através de um step de entrada no Pentaho (“CSV Input”), que faz a importação dos dados originais para o processo.

Estes datasets representam a base do projeto e são a matéria-prima para as transformações posteriores.

Transformação

Preparação e Ordenação dos Dados

Antes de proceder às junções entre tabelas, é necessário garantir que os dados estão ordenados de forma consistente pelas suas chaves de ligação.

No Pentaho, o step Merge Join exige que ambos os fluxos estejam ordenados pelo campo que será utilizado na junção.

Para isso, são usados vários steps “Sort Rows”:

Sort_Results_by_raceId e Sort_Races_by_raceId, para preparar os dados de resultados e corridas.

Sort_Results_Races_by_driverId e Sort_Drivers_by_driverId, para organizar os pilotos e resultados pelo identificador do piloto.

Sort_Constructors_by_constructorId, que prepara as equipas para a junção final.

Esta preparação assegura que todas as ligações entre datasets ocorrem de forma correta e sem perda de registos.

Junção dos Dados (Merge Join)

Depois da ordenação, são realizadas três junções sequenciais, utilizando sempre o tipo INNER JOIN para garantir que apenas os dados com correspondência em todas as tabelas são incluídos.

- Join_Results_Races – Junta os resultados das corridas (Results) com as informações das próprias corridas (Races), utilizando o campo raceId como chave comum.

O resultado é um dataset que relaciona cada corrida ao seu contexto (ano, circuito, etc.).

- Join_Results_Races_Drivers – Adiciona a informação dos pilotos (Drivers), através do campo driverId.

A partir deste ponto, cada linha representa o desempenho de um piloto específico numa corrida específica.

- Join_Results_Races_Drivers_Constructors – Faz a última junção, agora com a tabela das equipas (Constructors), usando o campo constructorId.

O resultado final desta etapa contém todos os dados integrados: piloto, corrida, equipa e resultados de desempenho.

Limpeza e Normalização dos Campos

Com os dados já integrados, é necessário fazer uma limpeza e padronização dos campos, de forma a remover informações desnecessárias e uniformizar os nomes das colunas.

Essa tarefa é executada com o step Select Values, chamado Cleaning.

Esta etapa reduz a quantidade de dados processados e melhora a legibilidade do dataset.

Criação do Nome Completo do Piloto

O passo seguinte, chamado CleanName, utiliza o step Modified JavaScript Value para criar um novo campo FullName, juntando o primeiro nome (forename) e o apelido (surname).

O script também elimina caracteres inválidos e espaços repetidos, garantindo uma padronização visual:

```
var FullName = Name + " " + surname;
FullName = FullName.replace(/[^A-Za-zÀ-ÿ\s]/g, '').replace(/\s+/g, ' ').trim();
var Clean_FullName = FullName;
```

Desta forma, o nome de cada piloto passa a estar completo e limpo, pronto para ser exibido nas estatísticas finais.

Filtragem Temporal

Como o objetivo do projeto é analisar os dados entre 2016 e 2024, foi adicionado um step de filtro chamado Filter_2016_2024.

Neste passo, o Pentaho seleciona apenas as linhas cujo campo year se encontra dentro desse intervalo:

```
year >= 2016 AND year <= 2024
```

Assim, as corridas anteriores a 2016 ou posteriores a 2024 são descartadas do processo.

Agrupamento e Cálculo de Estatísticas

Após a filtragem, entra-se na parte mais analítica do processo: o Group by. Este passo consolida as informações e calcula estatísticas para cada piloto e equipa.

O agrupamento é feito pelos campos:

driverId

Ou seja, cada linha do resultado representa um piloto específico.

Os campos agregados são:

Full Name – último valor do nome completo do piloto.

Driver Number – último valor do número do piloto.

Total Races – contagem de corridas disputadas (Number of Values (N) sobre o campo raceId).

Total Points – soma total dos pontos obtidos (Sum sobre o campo points).

AVG Start Position – média das posições de partida (Average (Mean) sobre grid).

AVG Finish Position – média das posições finais (Average (Mean) sobre positionOrder).

Este agrupamento permite perceber, por exemplo, quantas corridas um piloto fez, quantos pontos acumulou e qual a sua média de performance, tudo isto por equipa.

Ordenação dos Resultados

Depois do agrupamento, é aplicado um step de ordenação chamado Order_By_TotalPoints.

Aqui, os dados são ordenados pelo campo “Total Points” em ordem decrescente, para que os pilotos (ou piloto-equipa) com mais pontos apareçam no topo da listagem.

Exportação

Na fase final do processo, o fluxo de dados é direcionado para três diferentes saídas:

- Microsoft Excel Writer – gera o ficheiro Estatisticas_2016_2024.xlsx.
- Text File Output – cria o ficheiro Estatisticas_2016_2024.txt.
- JSON Output – exporta os mesmos dados no formato Estatisticas_2016_2024.json.

Desta forma, o mesmo conjunto de informações pode ser utilizado em ambientes analíticos distintos ou partilhado facilmente com outros sistemas.

Para complementar a análise desenvolvida no Pentaho, foram criados dois gráficos interativos no Power BI com base nos ficheiros de saída gerados pelo processo ETL.

O primeiro gráfico apresenta o Top 15 pilotos com maior número total de pontos entre 2016 e 2024, permitindo uma visualização clara do desempenho global de cada piloto.

O segundo gráfico ilustra a relação entre a posição média de partida e a posição média de chegada, evidenciando a consistência e evolução dos pilotos ao longo das corridas.

Estes elementos visuais reforçam a interpretação dos resultados e demonstram a utilidade prática dos dados transformados no processo de integração.

Resultados

O resultado final é uma tabela consolidada onde cada linha representa um piloto numa equipa específica, incluindo o número de corridas, o total de pontos e as médias de posição de partida e chegada.

Este modelo foi escolhido porque, ao longo dos anos, alguns pilotos mudaram de equipa.

Ao agrupar também pelo identificador da equipa (constructorId), é possível manter a fidelidade histórica dos dados e evitar confusões com pilotos repetidos em equipas diferentes.

O projeto demonstra um pipeline ETL completo e funcional:

- lê dados brutos de diversas fontes,
- aplica transformações e cálculos complexos,
- e entrega saídas organizadas e consistentes em múltiplos formatos.

Em suma, este trabalho representa um fluxo de dados realista e bem estruturado, com todas as fases essenciais de um processo ETL implementadas de forma clara e justificada.

Conclusão

Com este processo de ETL desenvolvido no Pentaho, foi possível centralizar, limpar e consolidar dados complexos de várias fontes de forma estruturada e automatizada.

O resultado permite identificar rapidamente os pilotos mais consistentes e de melhor desempenho entre 2016 e 2024, servindo de base para análises desportivas ou estatísticas.

A metodologia aplicada demonstra o potencial das ferramentas de integração de dados no apoio à tomada de decisão baseada em dados.

Além disso, o fluxo construído é escalável e pode ser facilmente adaptado para incluir novas métricas (como vitórias, pódios ou equipas), garantindo uma evolução contínua da análise estatística no contexto da Fórmula 1.

Bibliografia

https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020