MVP – Engenharia de Dados

Curso: Pós Graduação em Ciência de Dados e Analytics – PUC RIO  
Aluno: Raphael Mageste Duque  
Data: Setembro/2023

# Descrição

Neste trabalho, foi solicitada a construção de um pipeline de dados utilizando tecnologias na nuvem. Foi solicitado também que o pipeline envolva a busca, coleta, modelagem, carga e análise de dados.

O escopo deste projeto inclui fazer uma análise sobre os resultados dos principais campeonatos de futebol da última década e identificar a ocorrência de resultados inesperados (“zebras”) em cada um destes campeonatos.

Para isto, a primeira etapa do projeto consistiu em buscar e coletar fontes de dados com os resultados e as probabilidades de cada partida nos últimos anos.

As bases de dados foram então carregadas em um *Bucket* do Google Cloud Storage, serviço de armazenamento da nuvem Google Cloud Platform – GCP.

Na sequência, foi feita a etapa de modelagem e ETL utilizando a ferramenta Google Data Fusion. Os dados resultantes dos fluxos de ETL foram carregados no BigQuery, constituindo assim o Datawarehouse – DW – do projeto.

Por fim, no BigQuery foram feitas as consultas necessárias para analisar os dados coletados.

Todos os dados presentes no modelo de dados do DW foram catalogados utilizando a ferramenta Google Dataplex.

# Objetivo

Nos últimos anos tivemos um aumento de popularidade da análise de dados em contextos esportivos. Diversas entidades passaram a coletar os mais variados dados de diferentes esportes.

Olhando especificamente para o futebol, estatísticas como a posse de bola e o xG *– expected goals* – possibilitam que profissionais de dados criem modelos preditivos para os jogos. Estes modelos são capazes de definir as probabilidades de cada resultado para uma partida futura.

Porém, um dos motivos do futebol ser tão popular é justamente o fato de gerar resultados que contrariam expectativas – as conhecidas “zebras”.

Desta forma, o intuito deste trabalho é descobrir qual é o campeonato mais emocionante – ou imprevisível – do mundo. Seria a badalada *Premier League*, onde hoje atuam os principais craques do planeta? Ou talvez seria o Campeonato Brasileiro, considerado por alguns como o campeonato mais equilibrado do mundo?

Para responder isto, é feita uma análise em cima da base de dados disponibilizada pelo site FiveThirtyEight.com, um site especializado em coleta e exploração de dados.

O FiveThirtyEight tem uma seção dedicada a esportes, e dentro desta área foi construído um modelo preditivo de futebol, chamado de SPI – *Soccer Power Index*. Através do SPI (um valor de 0 a 100), são definidas as probabilidades de resultado quando duas equipes se enfrentam.

O site disponibiliza uma base de dados com o SPI e a probabilidade calculada para jogos de diferentes campeonatos desde 2016. Os detalhes do modelo preditivo podem ser consultados [aqui](https://fivethirtyeight.com/methodology/how-our-club-soccer-predictions-work/), e os detalhes da base de dados disponibilizada, [aqui](https://github.com/fivethirtyeight/data/tree/master/soccer-spi).

A partir desta base, eis as perguntas que vão nortear este trabalho:

1. Qual é o campeonato mais disputado do mundo (isto é, aquele onde os favoritos mais têm revezes)?
2. O quão imprevisível é o Campeonato Brasileiro? É de fato um campeonato mais disputado que os demais?
3. Existe alguma tendência temporal para o acontecimento das zebras? Elas estão aumentando ou diminuindo?

Na próxima seção, será visto em detalhe a construção do pipeline para responder estas perguntas.

# Detalhamento

## Busca de Dados

Os dados utilizados são os disponibilizados pelo site FiveThirtyEight, através de repositório no GitHub. Segue abaixo o link do repositório:

[data/soccer-spi at master · fivethirtyeight/data (github.com)](https://github.com/fivethirtyeight/data/tree/master/soccer-spi)

São utilizados os seguintes arquivos:

* spi\_matches.csv : contém dados de probabilidades e resultados de diferentes campeonatos de clubes desde 2016.
* spi\_matches\_latest.csv : contém dados de probabilidades e resultados de diferentes campeonatos de clubes considerando apenas a última temporada disponível.
* spi\_matches\_intl.csv : contém dados de probabilidades e resultados de diferentes campeonatos entre seleções.
* spi\_global\_rankings\_intl.csv : contém dados de países (seleções) e a quais confederações eles pertencem.

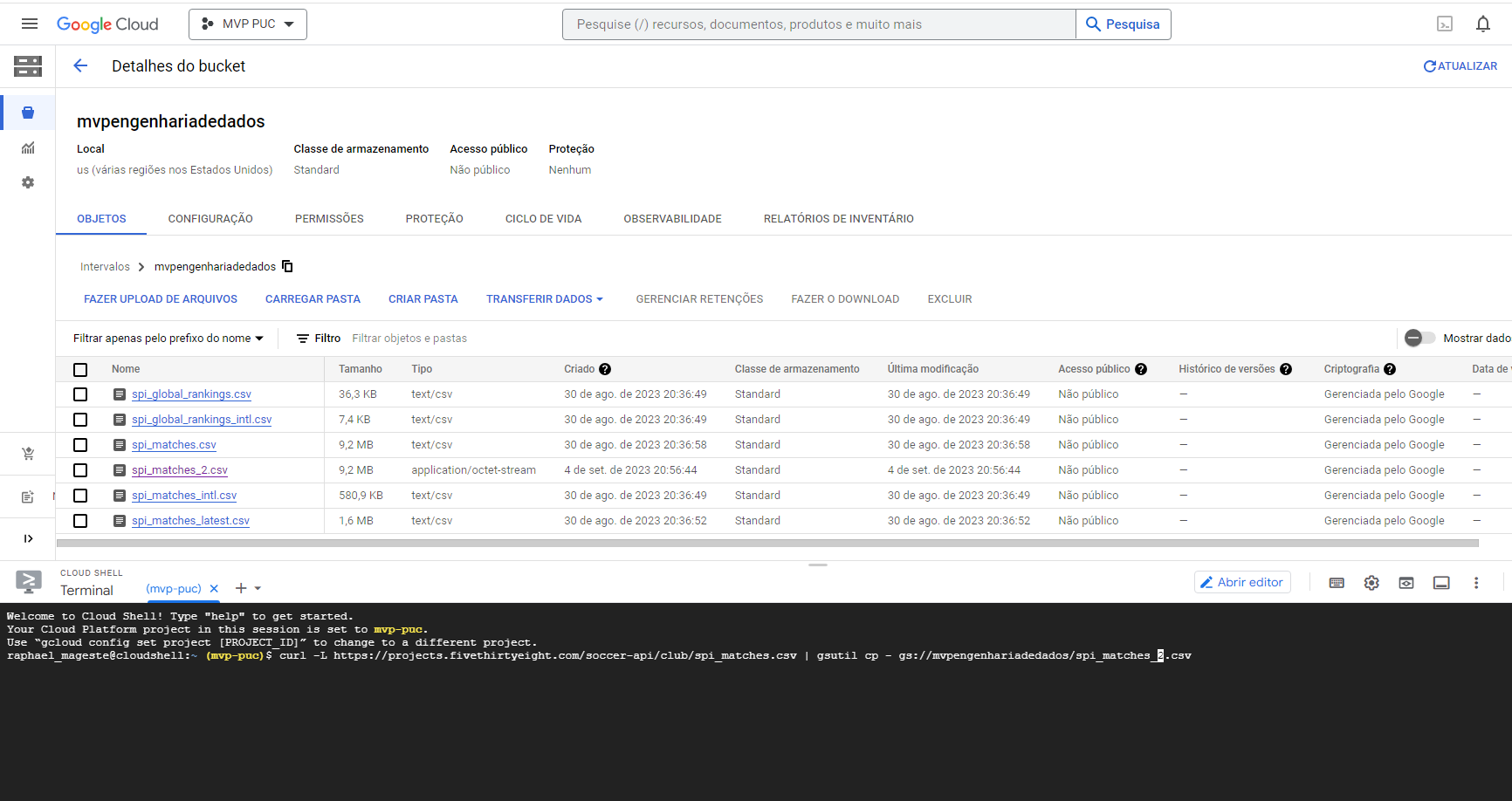
*Documentação da base de dados disponível no GitHub*

## Ingestão de Dados

É feito o upload das bases em .csv para o bucket ‘mvpengenhariadedados’ do Google Cloud Storage. O upload pode ser feito de duas formas:

* Download de arquivos para a máquina local e inserção manual no bucket do Google Cloud Storage;
* Inserção automática através do de linha de comando no console gsutil da GCP.   
  Esta forma foi utilizada para demonstrar a possibilidade de consumir dados direto da fonte, sem a necessidade de baixar arquivos para a máquina local.   
  O comando utilizado está descrito a seguir:

*curl -L https://projects.fivethirtyeight.com/soccer-api/club/spi\_matches.csv | gsutil cp - gs://mvpengenhariadedados/spi\_matches.csv*

*Evidência de upload de arquivo através de linha de comando via shell da GCP*

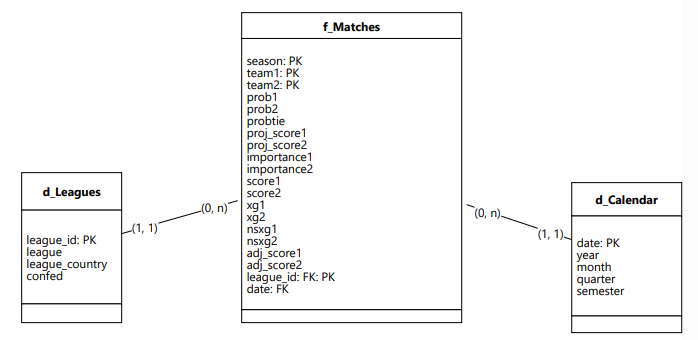
Os arquivos coletados estão disponíveis no bucket *mvpengenhariadedados* (somente leitura):

<https://console.cloud.google.com/storage/browser/mvpengenhariadedados>

## Modelagem

### **Modelo Esquema Estrela**

Após análise preliminar dos dados, fica definido que o projeto será construído no modelo Esquema Estrela, de acordo com o esboço abaixo. Neste modelo temos a tabela fato **Matches**, na qual a PK é definida pela composição dos atributos *season, team1, team2* e *league\_id.*

A tabela fato se relaciona com as dimensões **Leagues** e **Calendar** através das chaves *league\_id* e *date,* respectivamente.

*Modelo Estrela do Projeto*

A tabela fato **f\_matches** é formada pela união de dados de 03 fontes: *spi\_matches.csv*, *spi\_matches\_latest.csv* e *spi\_matches\_intl.csv* .

Todas estas 03 tabelas apresentam a mesma estrutura, detalhada a seguir.

Para a tabela **f\_matches** sãodescartados os campos proj\_score1, proj\_score2, importance1, importance2, nsxg1 e nsxg2, por não serem relevantes para as análises deste trabalho.

|  |  |
| --- | --- |
| Header | Definition |
| season | The season during which the match was played |
| date | The date of the match (YYYY-MM-DD) |
| league\_id | A unique identifier for the league this match was played in |
| league | The name of the league this match was played in |
| team1 | The home team's name |
| team2 | The away team's name |
| spi1 | The home team's overall SPI rating before the match |
| spi2 | The away team's overall SPI rating before the match |
| prob1 | The probability of the home team winning the match |
| prob2 | The probability of the away team winning the match |
| probtie | The probability of match ending in a draw (if applicable) |
| proj\_score1 | The number of goals we expected the home team to score |
| proj\_score2 | The number of goals we expected the away team to score |
| importance1 | The importance of the match for the home team (0-100) |
| importance2 | The importance of the match for the away team (0-100) |
| score1 | The number of goals scored by the home team |
| score2 | The number of goals scored by the away team |
| xg1 | The number of expected goals created by the home team |
| xg2 | The number of expected goals created by the away team |
| nsxg1 | The number of non-shot expected goals created by the home team |
| nsxg2 | The number of non-shot expected goals created by the away team |
| adj\_score1 | The number of goals scored by the home team, adjusted for game state |
| adj\_score2 | The number of goals scored by the home team, adjusted for game state |

*Descrição original dos campos dos arquivos do tipo spi\_matches.csv*

A tabela dimensão d\_leagues é construída a partir de dados de *spi\_matches.csv*, *spi\_matches\_latest.csv* e *spi\_matches\_intl.csv* e também de *spi\_global\_rankings\_intl.csv*.

Das tabelas de matches se pega os campos *league\_id* e *league*. Então é feito uma transformação em cima do campo *league* para a criação do campo *league\_country*. Com este campo *league\_country*, que traz o nome dos países de cada campeonato, é possível fazer um join com o campo *name* da tabela *spi\_global\_rankings\_intl.csv* e assim buscar o campo *confed* para a dimensão.

Abaixo está a estrutura original de *spi\_global\_rankings\_intl.csv:*

|  |  |
| --- | --- |
| Header | Definition |
| rank | The team's current global ranking |
| name | The team's name |
| confed | The confederation the team plays in |
| off | The team's offensive SPI rating |
| def | The team's defensive SPI rating |
| spi | The team's overall SPI rating |

*Descrição original dos campos do arquivo spi\_global\_rankings\_intl.csv*

Por fim, a tabela d\_calendar tem origem nas datas presentes nas tabelas *spi\_matches.csv*, *spi\_matches\_latest.csv* e *spi\_matches\_intl.csv.*

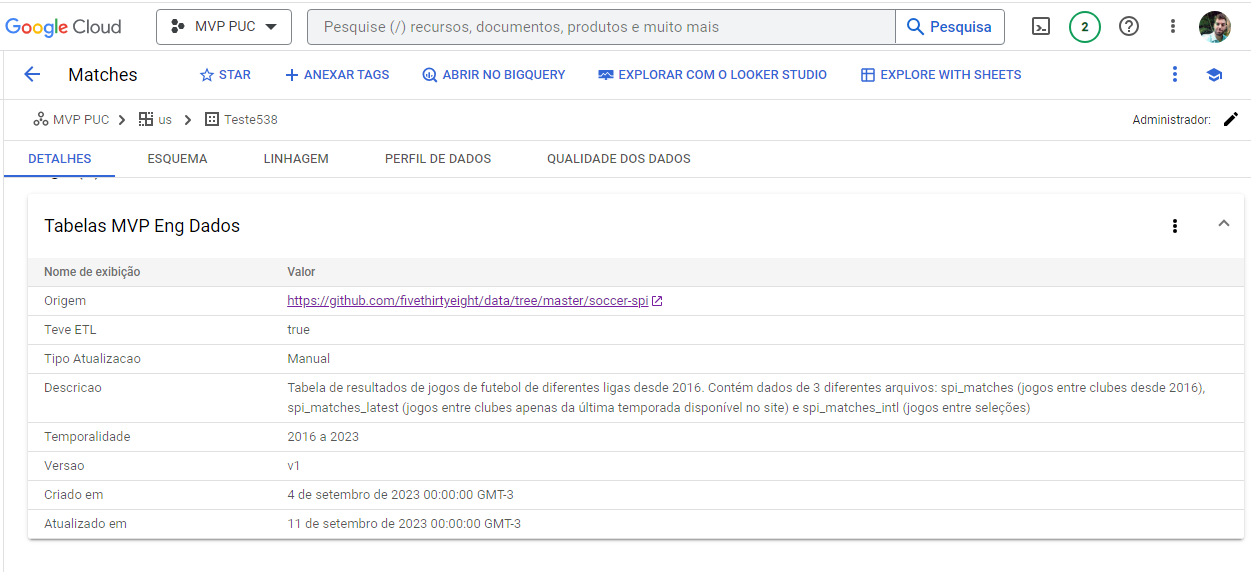
Após a união destas tabelas, é feita uma transformação em cima do campo *date* para obter as datas distintas da tabela fato. Após isso, são feitas transformações em cima da coluna de datas resultante para obter os campos de ano, mês, dia, semestre e trimestre.

### **Catálogo de Dados**

Para melhor descrição dos dados do modelo, foi criado um catálogo de dados através da ferramenta Google Dataplex (que inclui, dentre outras ferramantas, o Google Data Catalog).

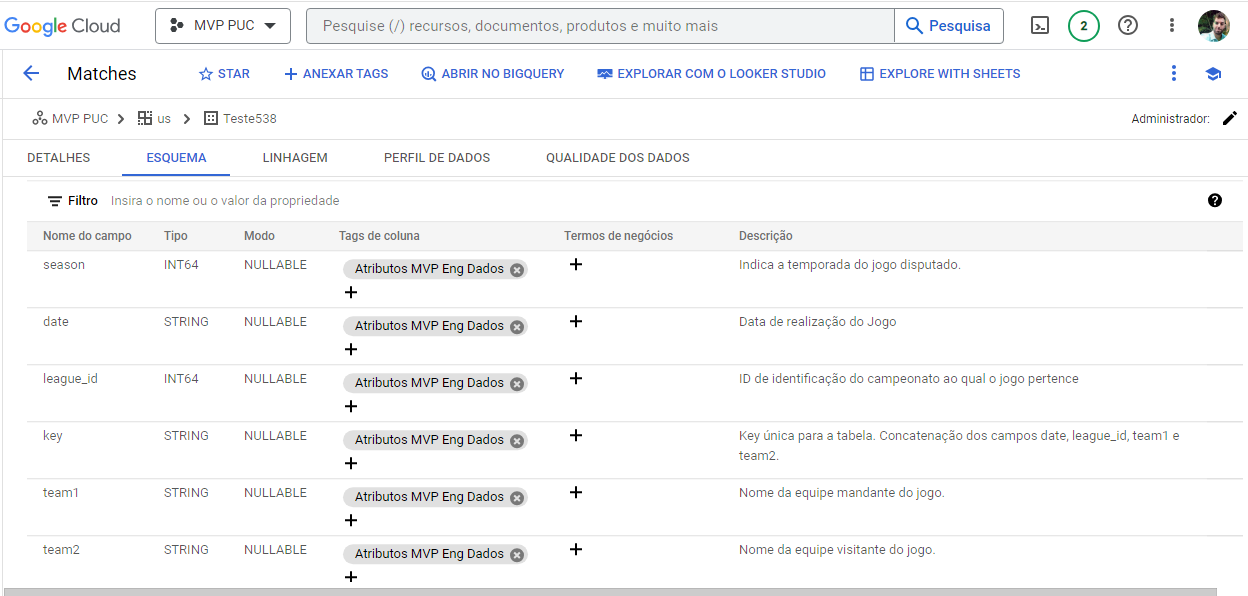
#### Tabela Matches

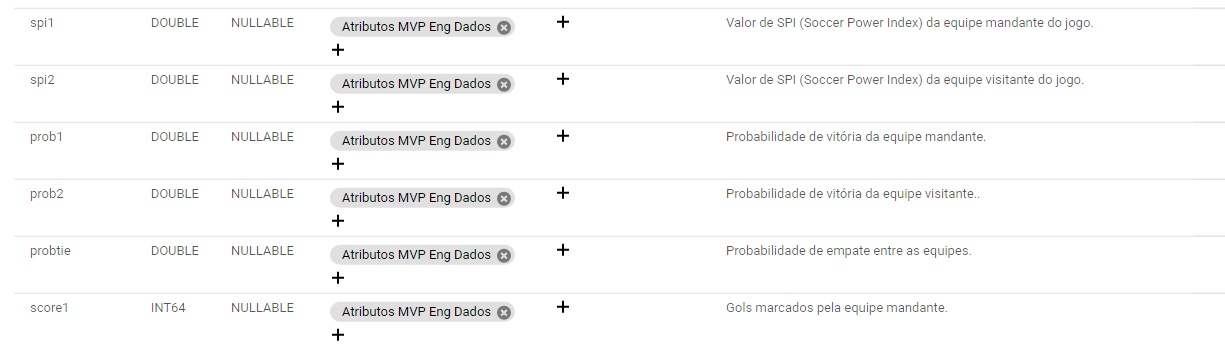
Abaixo estão as evidências de metadados criados para a tabela *matches* no Google Dataplex:

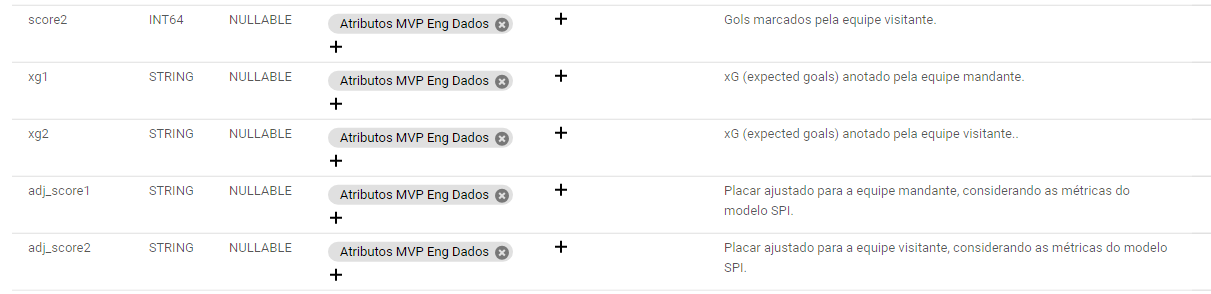


*Metadados da Tabela Matches*

Esquema da tabela *matches* com descrição de campos:



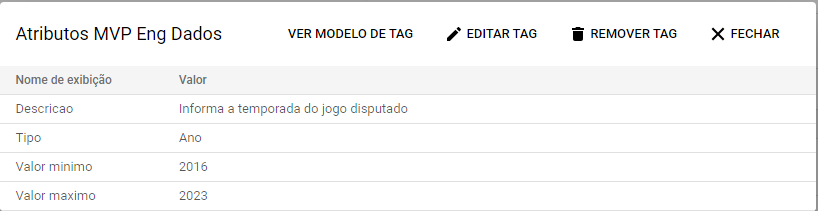




*Schema e descrição dos campos da Tabela Matches*

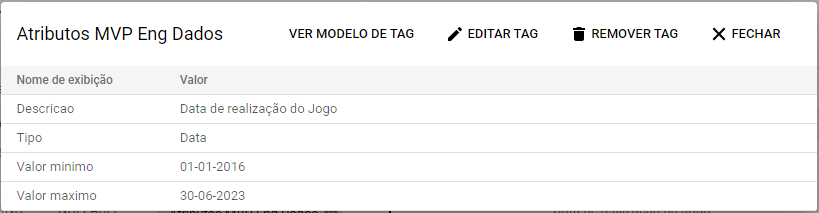
Metadados dos campos da tabela *matches*:

**Season**

****

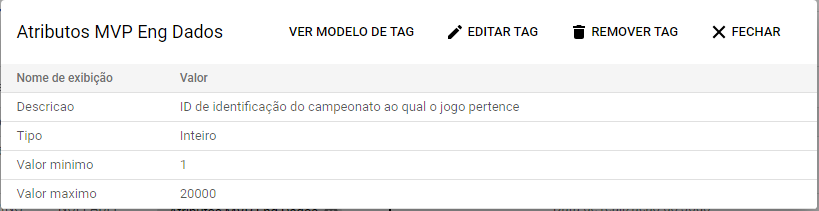
*Metadados do campo Season – Tabela Matches*

**Date**

****

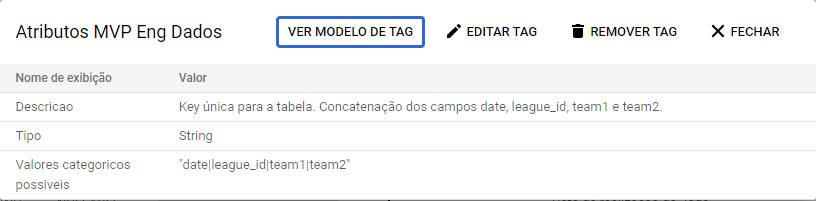
*Metadados do campo Date – Tabela Matches*

**league\_id**

****

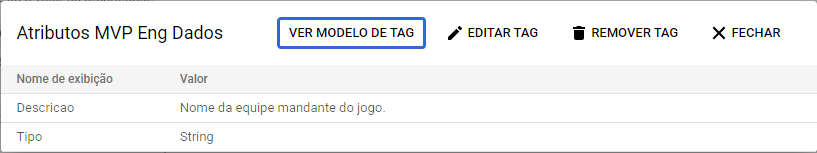
*Metadados do campo league\_id – Tabela Matches*

**Key**

****

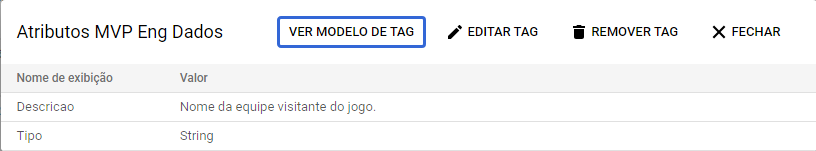
*Metadados do campo Key – Tabela Matches*

**team1**

****

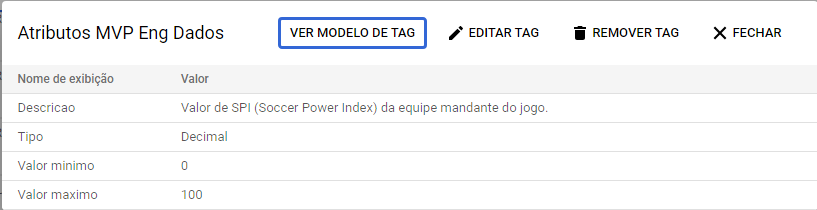
*Metadados do campo team1 – Tabela Matches*

**team2**

****

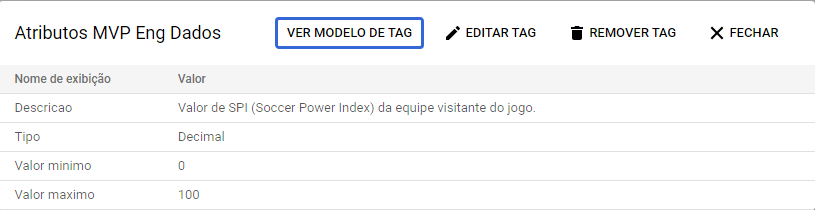
*Metadados do campo team2 – Tabela Matches*

**spi1**

****

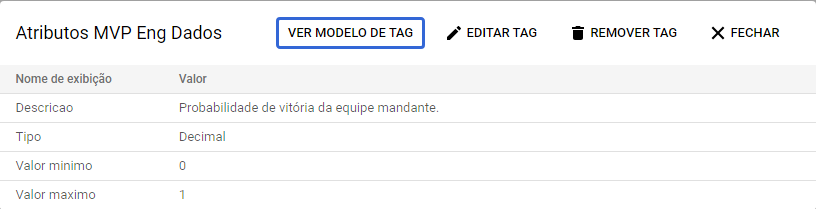
*Metadados do campo spi1 – Tabela Matches*

**spi2**

****

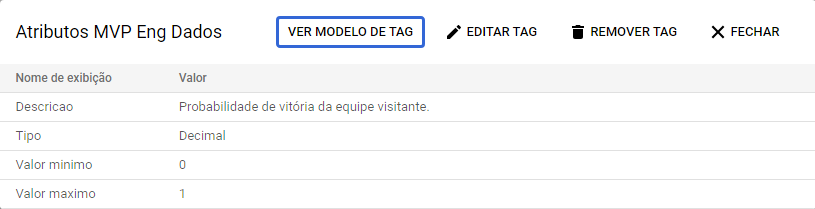
*Metadados do campo spi2 – Tabela Matches*

**prob1**

****

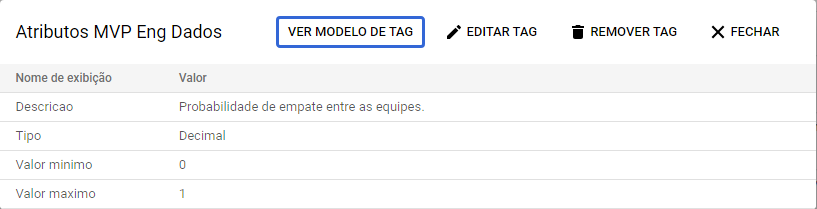
*Metadados do campo prob1 – Tabela Matches*

**prob2**

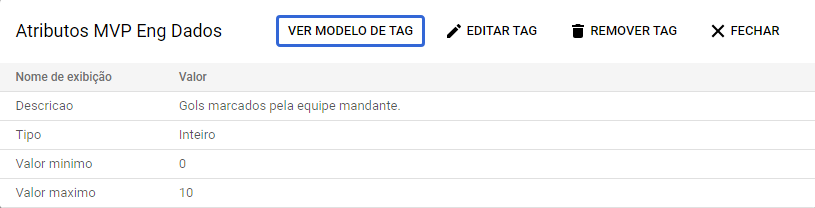
****

*Metadados do campo prob2 – Tabela Matches*

**Probtie**

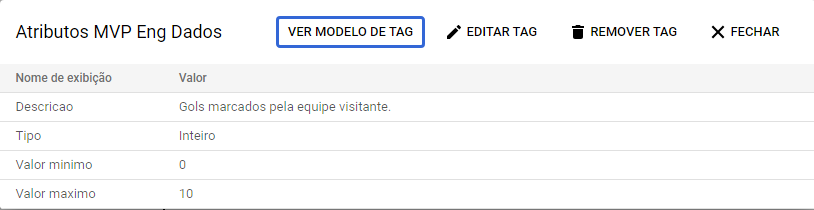
****

*Metadados do campo probtie – Tabela Matches*

**score1**

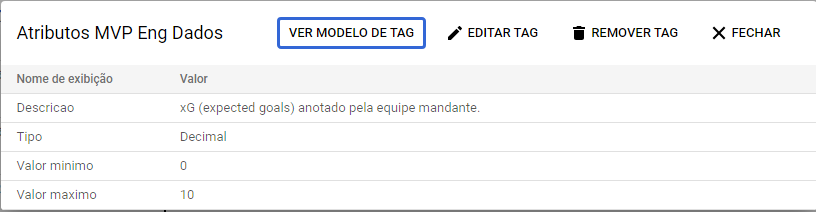
*Metadados do campo score1 – Tabela Matches*

**score2**

****

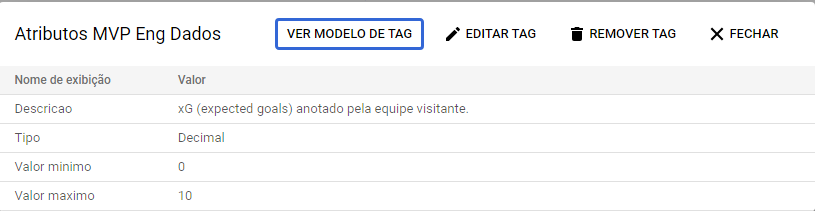
*Metadados do campo score2 – Tabela Matches*

**xg1**

****

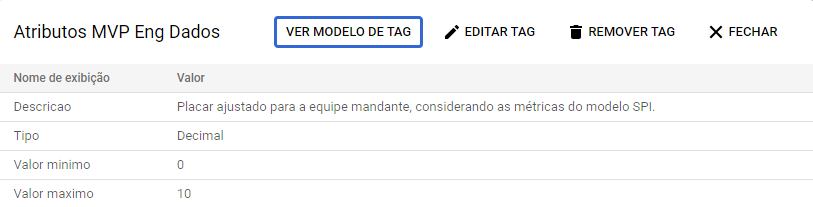
*Metadados do campo xg1 – Tabela Matches*

**xg2**

****

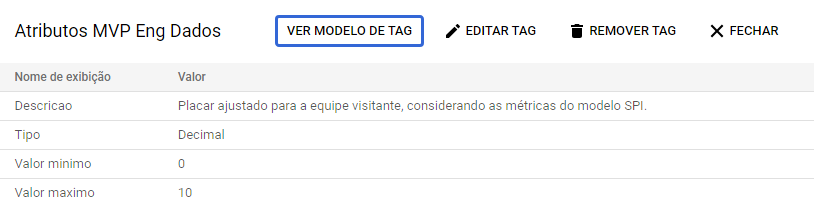
*Metadados do campo xg2 – Tabela Matches*

**adj\_score1**

****

*Metadados do campo adj\_score1 – Tabela Matches*

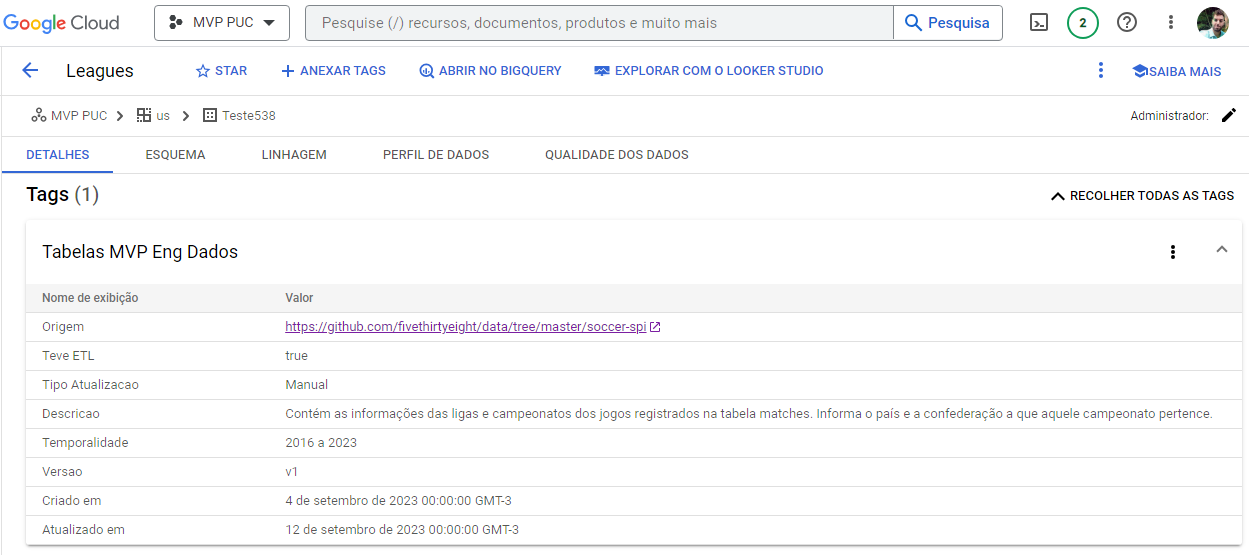
**adj\_score2**

****

*Metadados do campo adj\_score2 – Tabela Matches*

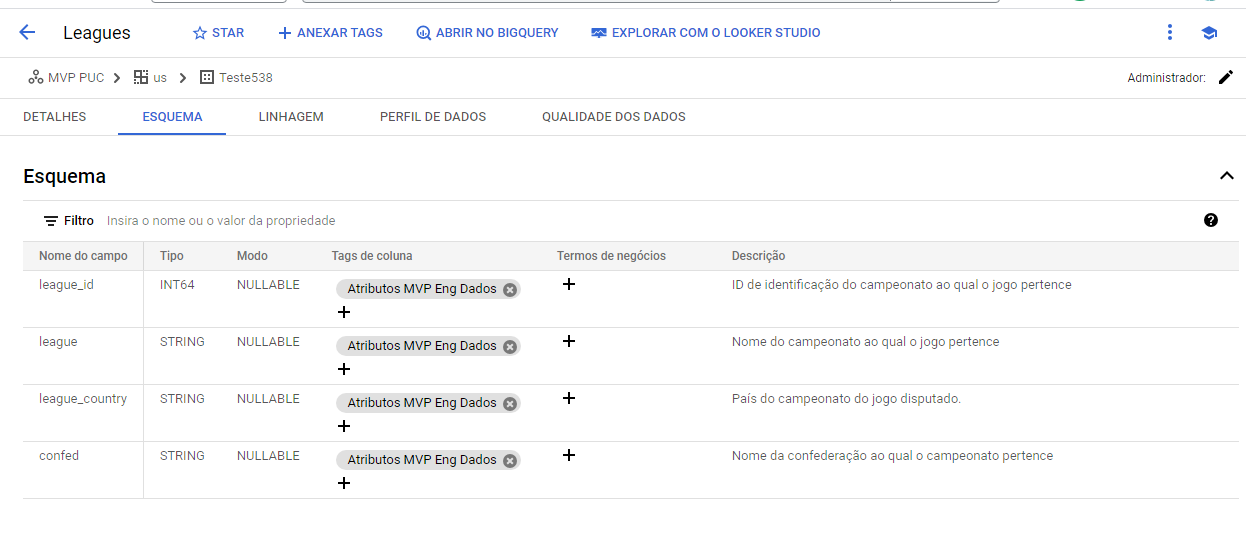
#### Tabela Leagues

Abaixo estão as evidências de metadados criados para a tabela *leagues* no Google Dataplex:



*Metadados da Tabela Leagues*

Esquema da tabela *leagues* com descrição de campos:



*Schema e descrição dos campos da Tabela Leagues*

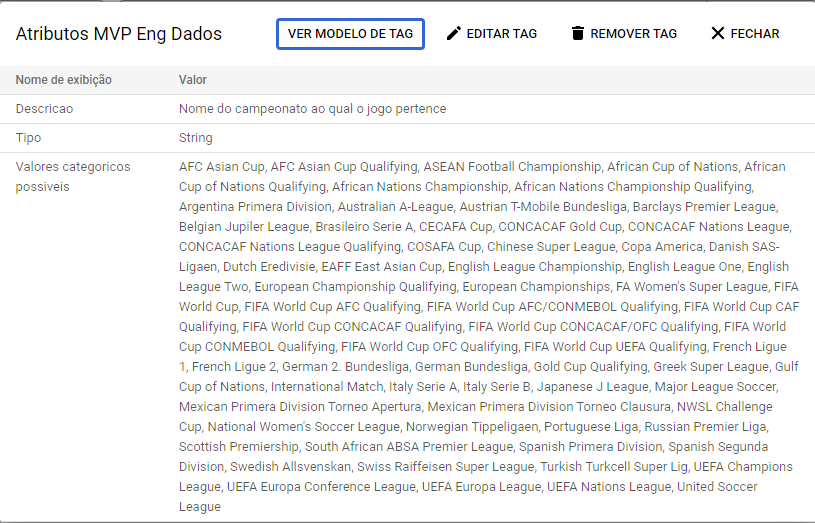
Metadados dos campos da tabela *leagues*:

**league\_id**

****

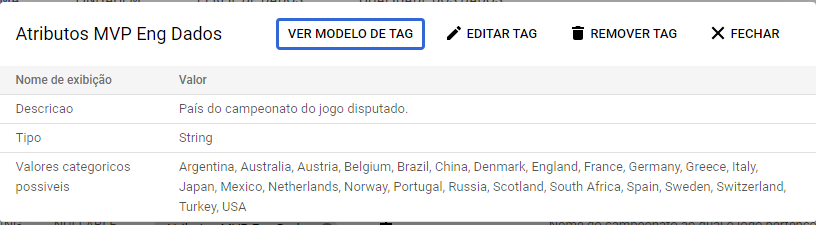
*Metadados do campo league\_id - Tabela Leagues*

**league**

****

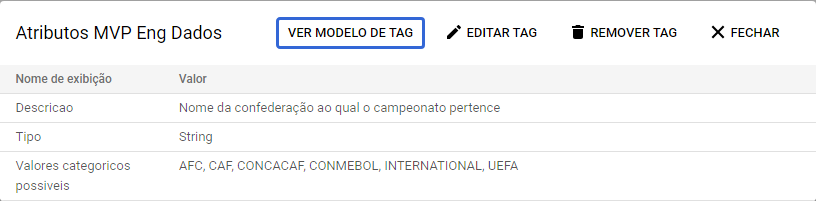
*Metadados do campo league - Tabela Leagues*

**league\_country**

****

*Metadados do campo league\_country - Tabela Leagues*

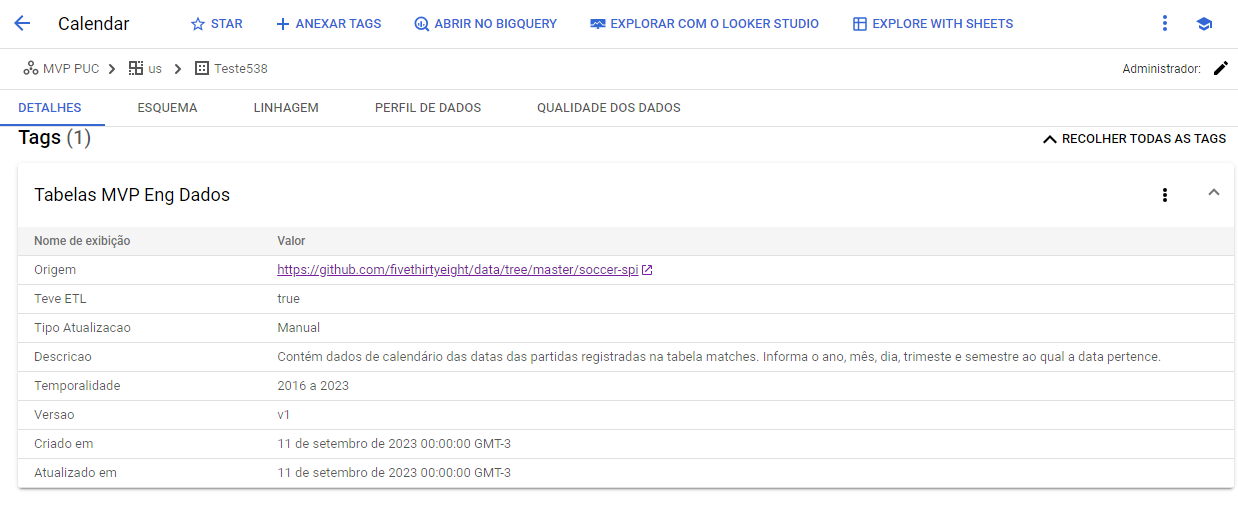
**confed**

****

*Metadados do campo confed - Tabela Leagues*

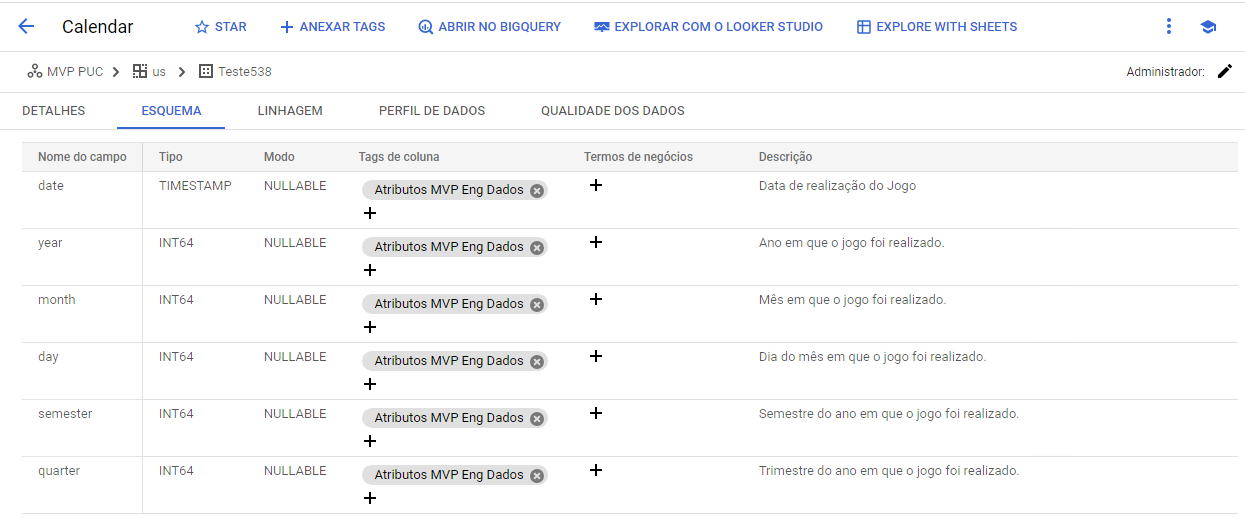
#### Tabela Calendar

Abaixo estão as evidências de metadados criados para a tabela *calendar* no Google Dataplex:



*Metadados da Tabela Calendar*

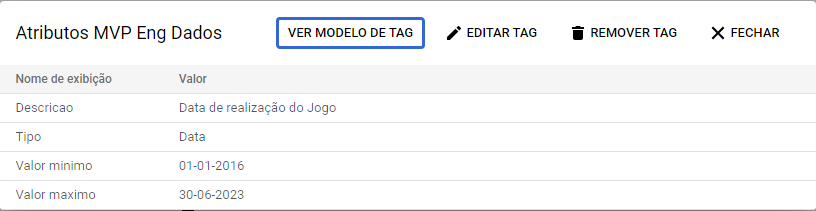
Esquema da tabela *calendar* com descrição de campos:



*Schema e descrição dos campos da tabela Calendar*

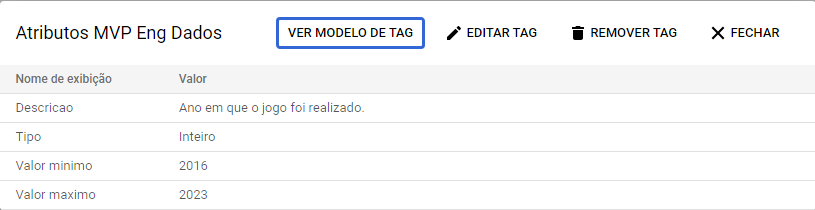
Metadados dos campos da tabela *calendar*:

**date**



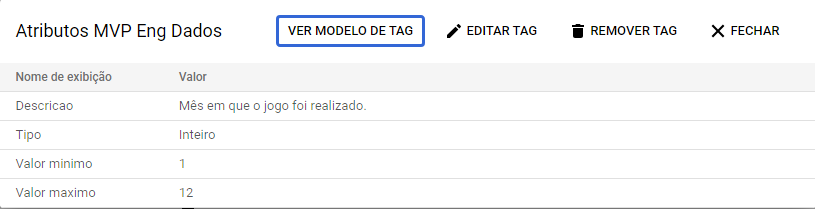
*Metadados do campo date - Tabela Calendar*

**year**

****

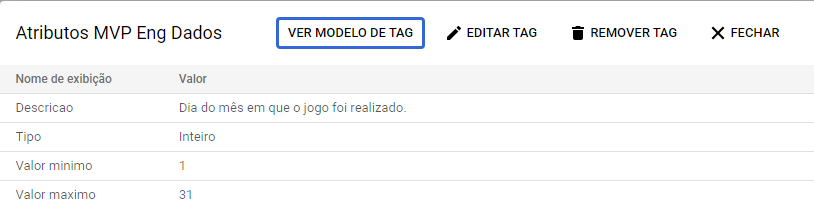
*Metadados do campo year - Tabela Calendar*

**month**

****

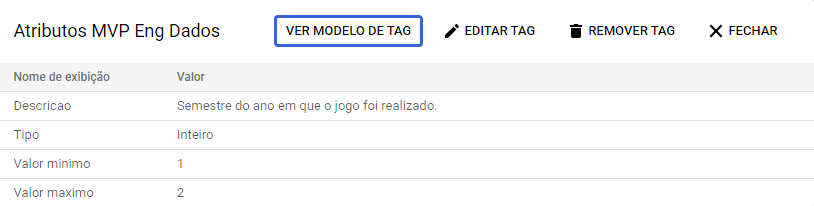
*Metadados do campo month - Tabela Calendar*

**day**

****

*Metadados do campo day - Tabela Calendar*

**semester**

****

*Metadados do campo semester - Tabela Calendar*

**quarter**

****

*Metadados do campo quarter - Tabela Calendar*

### **Linhagem de Dados**

Conforme visto na etapa anterior, as tabelas do modelo tem suas origens em planilhas csv hospedadas no github, são carregadas no Google Bucket Storage e antes de serem criadas no BigQuery são transformadas usando Google Data Fusion.

De forma resumida, temos:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Fontes | Salvas em | Transformações | Tabela Final | Local |
| *spi\_matches.csv*, *spi\_matches\_latest.csv,* *spi\_matches\_intl.csv* | Google Bucket Storage | União de tabelas, tratamento de nulos e outros via Google Data Fusion | matches | Google BigQuery |
| *spi\_matches.csv*, *spi\_matches\_latest.csv,* *spi\_matches\_intl.csv* | Google Bucket Storage | União de tabelas, remoção de duplicatas e outros via Google Data Fusion | calendar | Google BigQuery |
| *spi\_matches.csv*, *spi\_matches\_latest.csv,* *spi\_matches\_intl.csv,*  *spi\_global\_rankings\_intl.csv* | Google Bucket Storage | União e join de tabelas, tratamento de nulos e outros via Google Data Fusion | leagues | Google BigQuery |

*Linhagem de dados do Projeto*

## ETL

A etapa de ETL foi feita utilizando a ferramenta Data Fusion da GCP. Foram criados 03 fluxos, um para cada tabela do modelo visto acima.

### **Tabela Matches**

Para a construção da tabela matches, é feita a união das 3 tabelas csv: *matches\_latest, matches* e *matches\_international*.

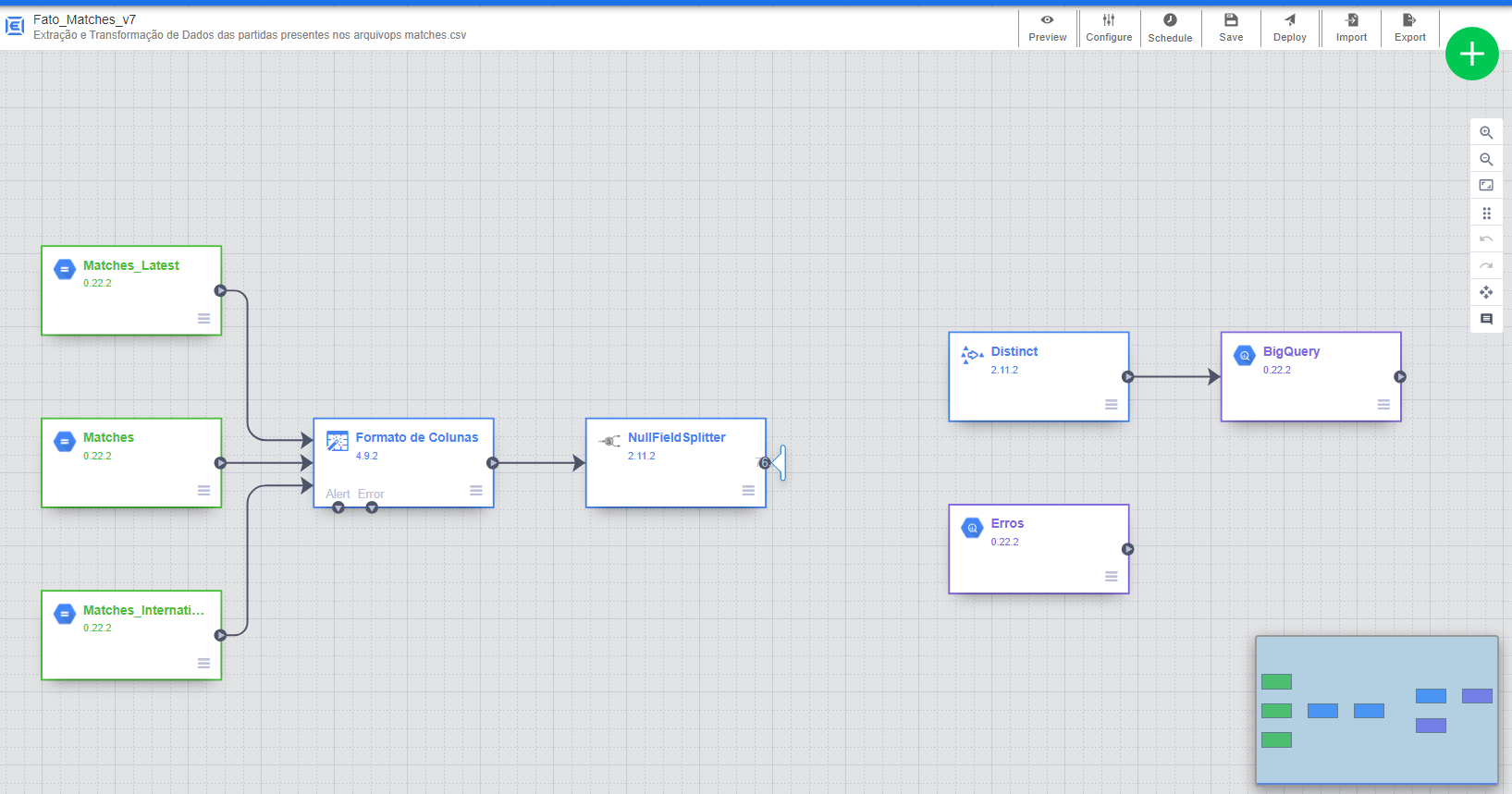
No segundo passo, no **Wrangler**, é feita uma formatação geral das colunas.

Em sequência, o bloco **NullFieldSplitter** faz uma filtragem de valores nulos no conjunto de dados: registros que estejam com nulos no campo de score são classificados como erros e serão salvos à parte em outra tabela do BigQuery (tabela **Erros**).

O fluxo principal segue por sua vez para o bloco **Distinct**, que remove qualquer registro que tenha ficado em duplicata.

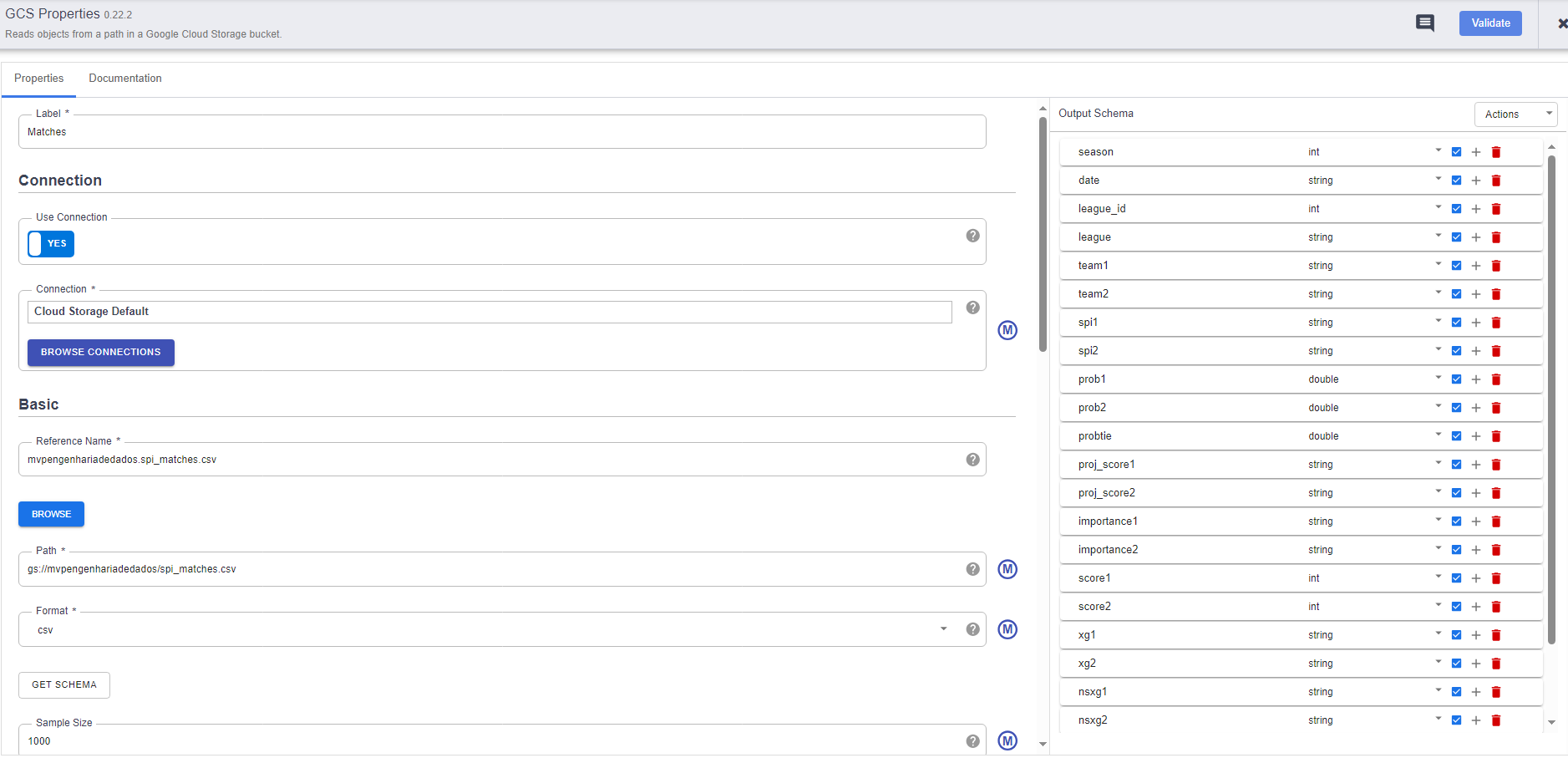
Por fim, a tabela é salva no BigQuery como **Matches**.

Visão geral do fluxo:

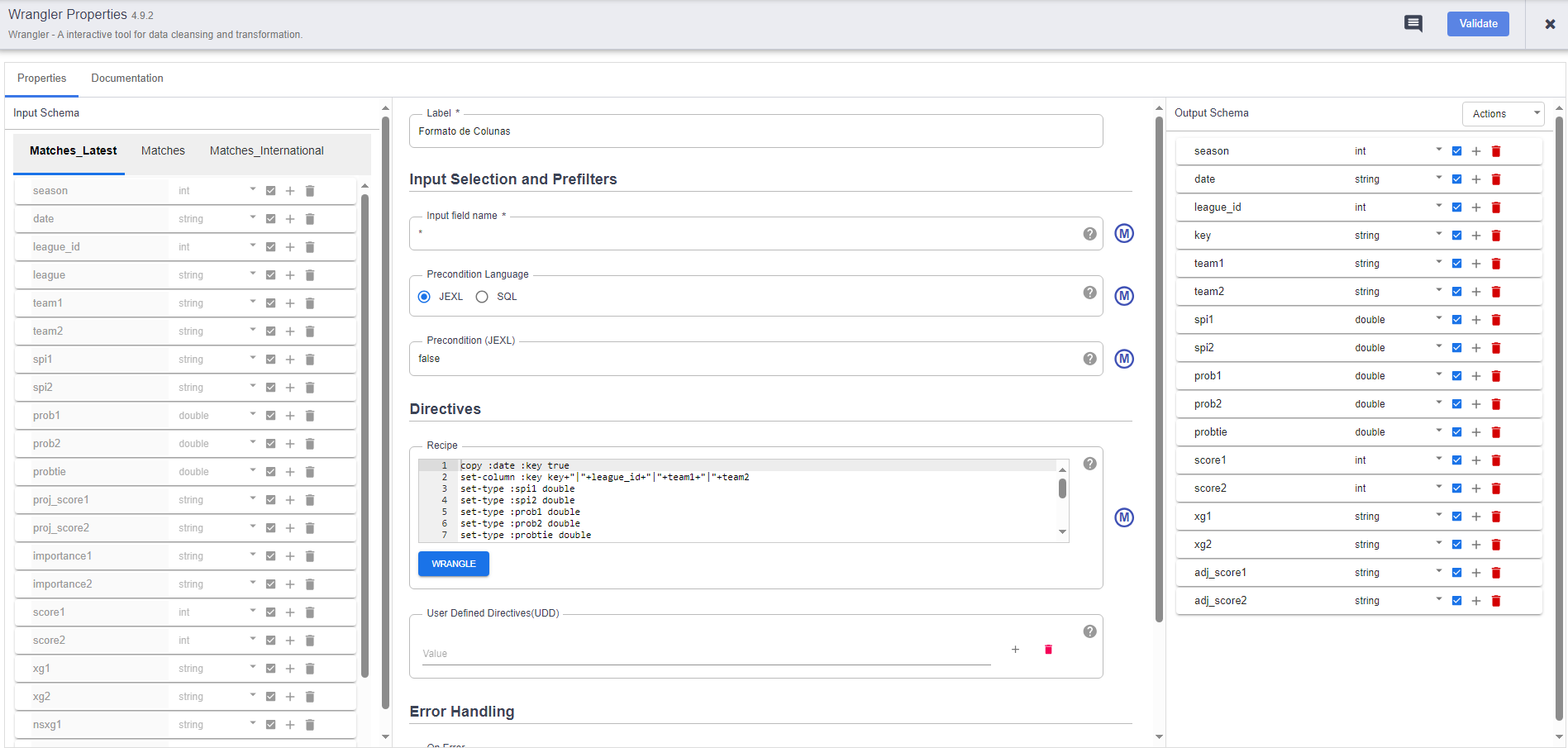


*Fluxo ETL da Tabela Matches*

Configuração geral do bloco de carregamento dos arquivos csv:



Visão geral do bloco WRANGLER e transformações aplicadas:



copy :date :key true

set-column :key key+"|"+league\_id+"|"+team1+"|"+team2

set-type :spi1 double

set-type :spi2 double

set-type :prob1 double

set-type :prob2 double

set-type :probtie double

set-type :league\_id string

drop :league

drop :proj\_score1

drop :proj\_score2

drop :importance1

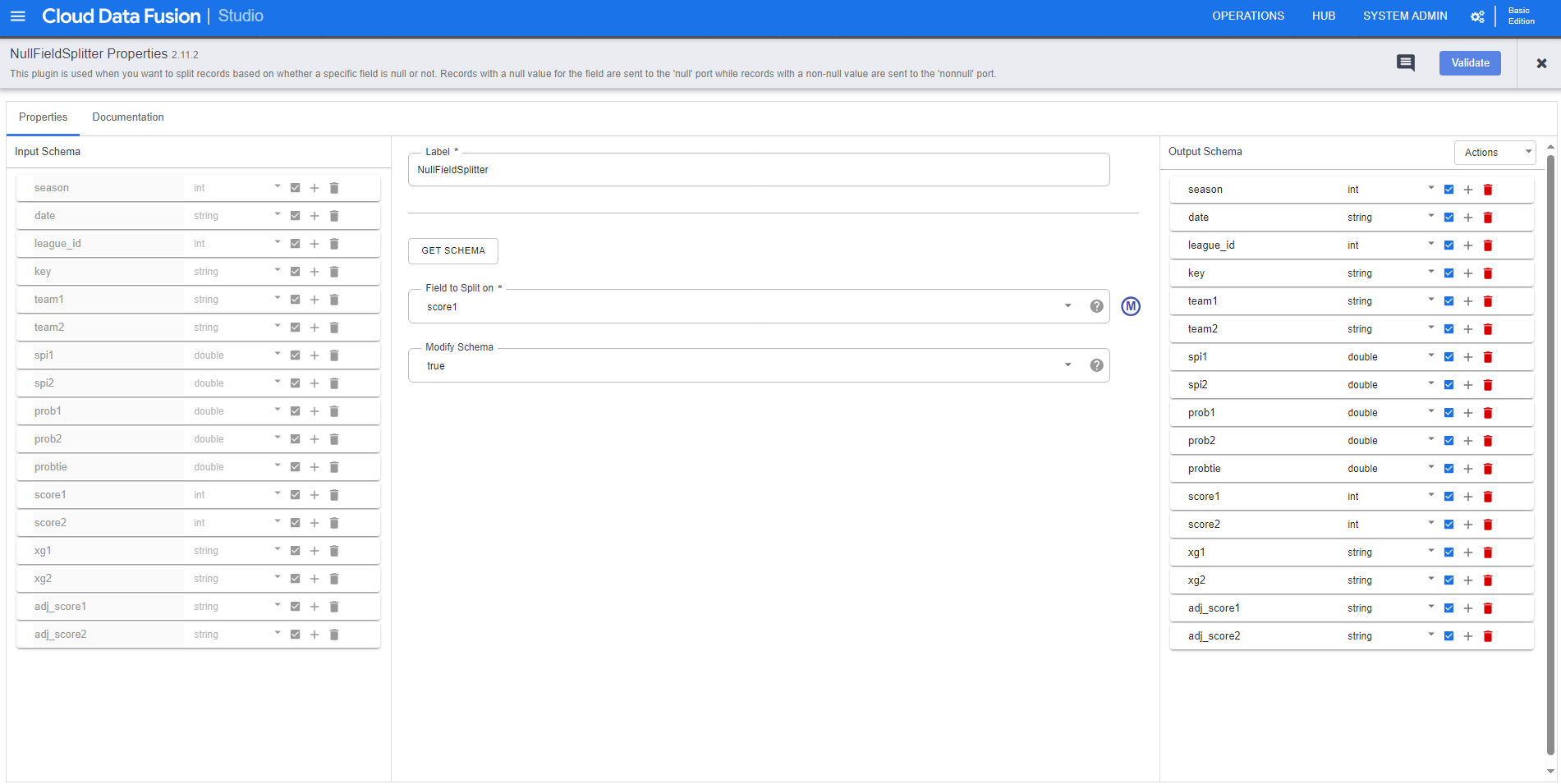
drop :importance2

drop :nsxg1

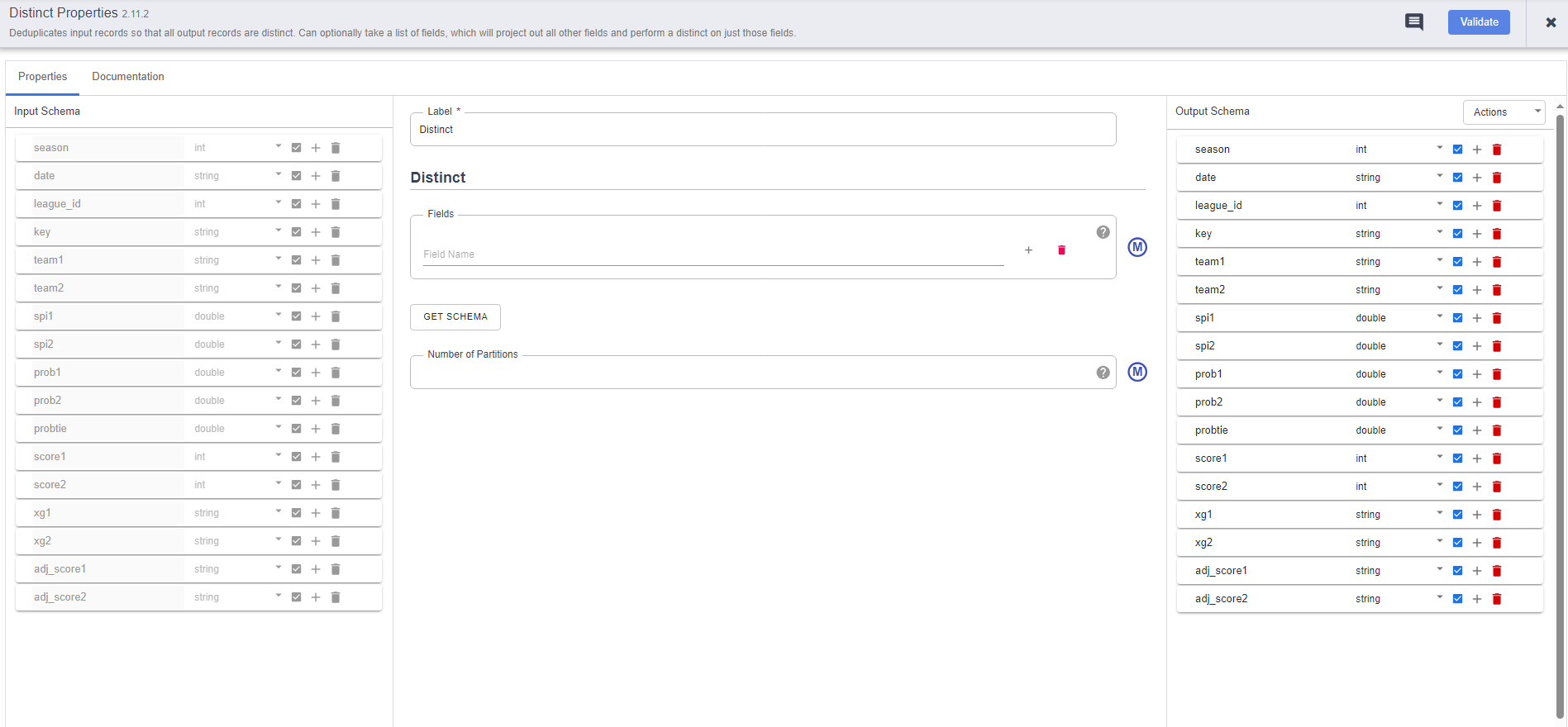
drop :nsxg2

parse-as-simple-date :date yyyy-MM-dd

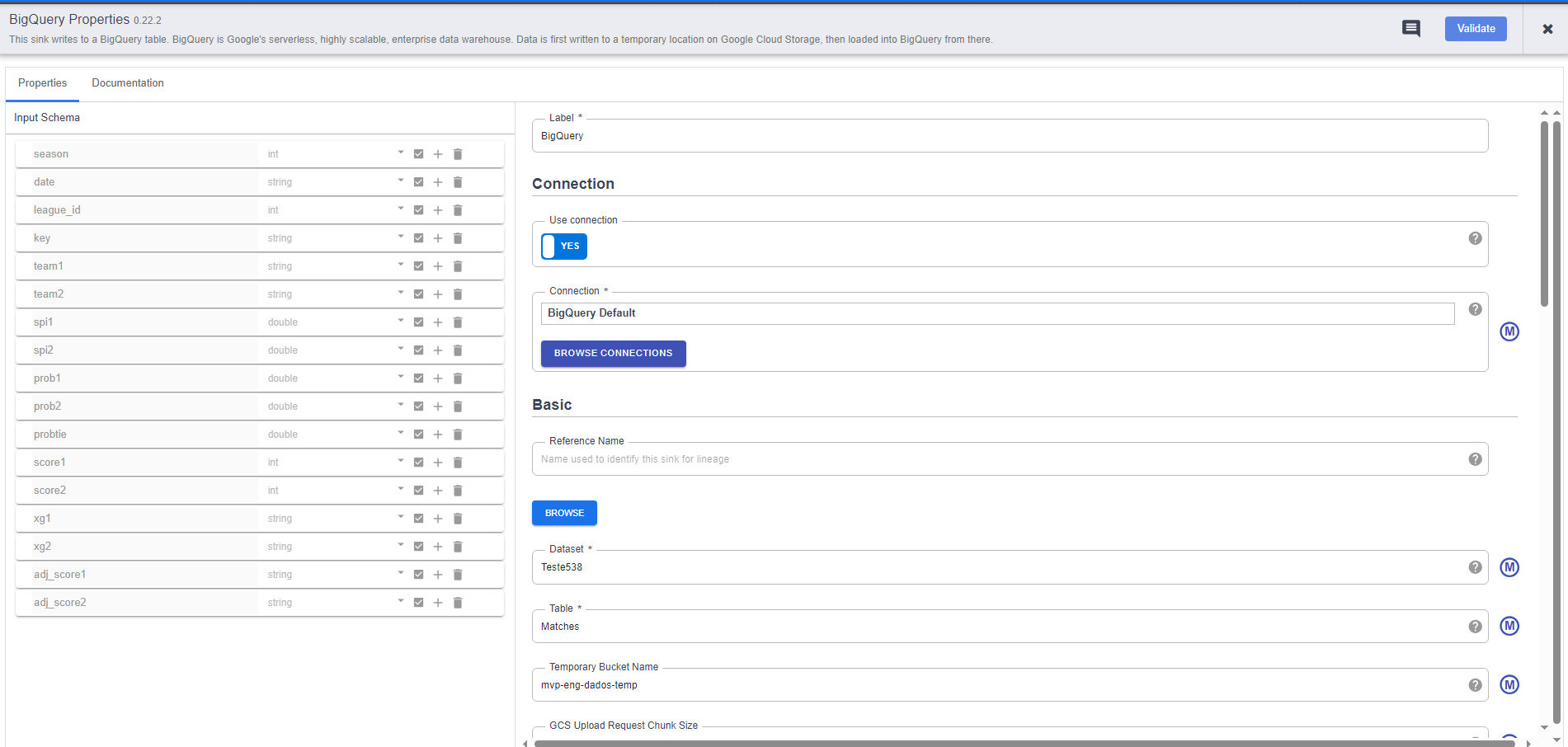
Visão geral do bloco NullFieldSplitter, filtrando valores nulos de score1:



Visão geral do bloco Distinct, que remove duplicatas do conjunto:



Visão geral do bloco do BigQuery, que salva a tabela no BQ:



### **Tabela Leagues**

A tabela **Leagues** também tem origem a partir dos 3 arquivos *csv matches, matches\_latest* e *matches\_international.*

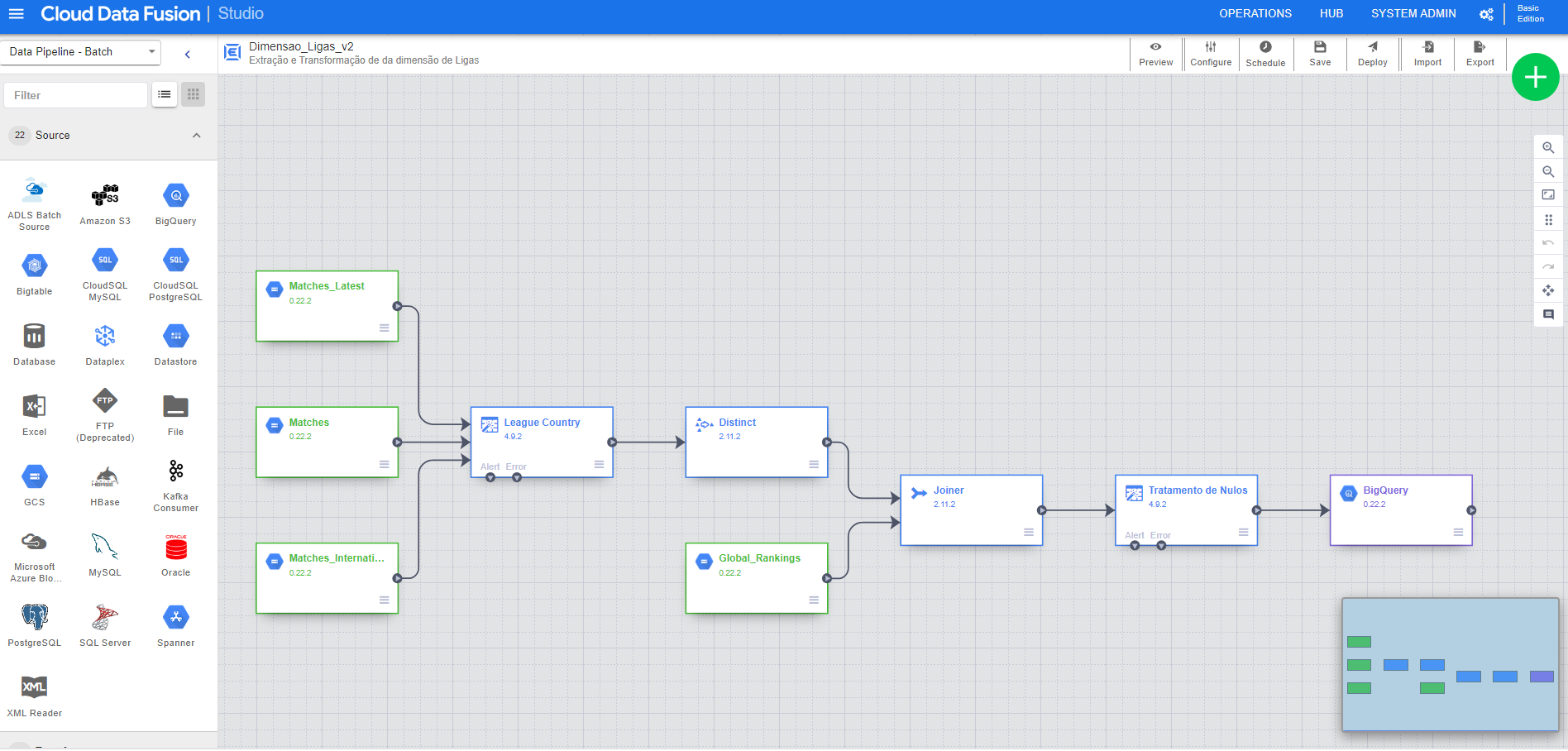
No segundo passo, no **Wrangler**, são feitas transformações a partir do nome das ligas para a criação da coluna *league\_country*, que traz o país de cada liga.

Após remover duplicatas com o bloco **Distinct**, carregamos o csv *global\_rankings* que contém as informações de confederação de cada país.

Assim, juntamos país e confederação em uma única tabela, usando o bloco **Join**.

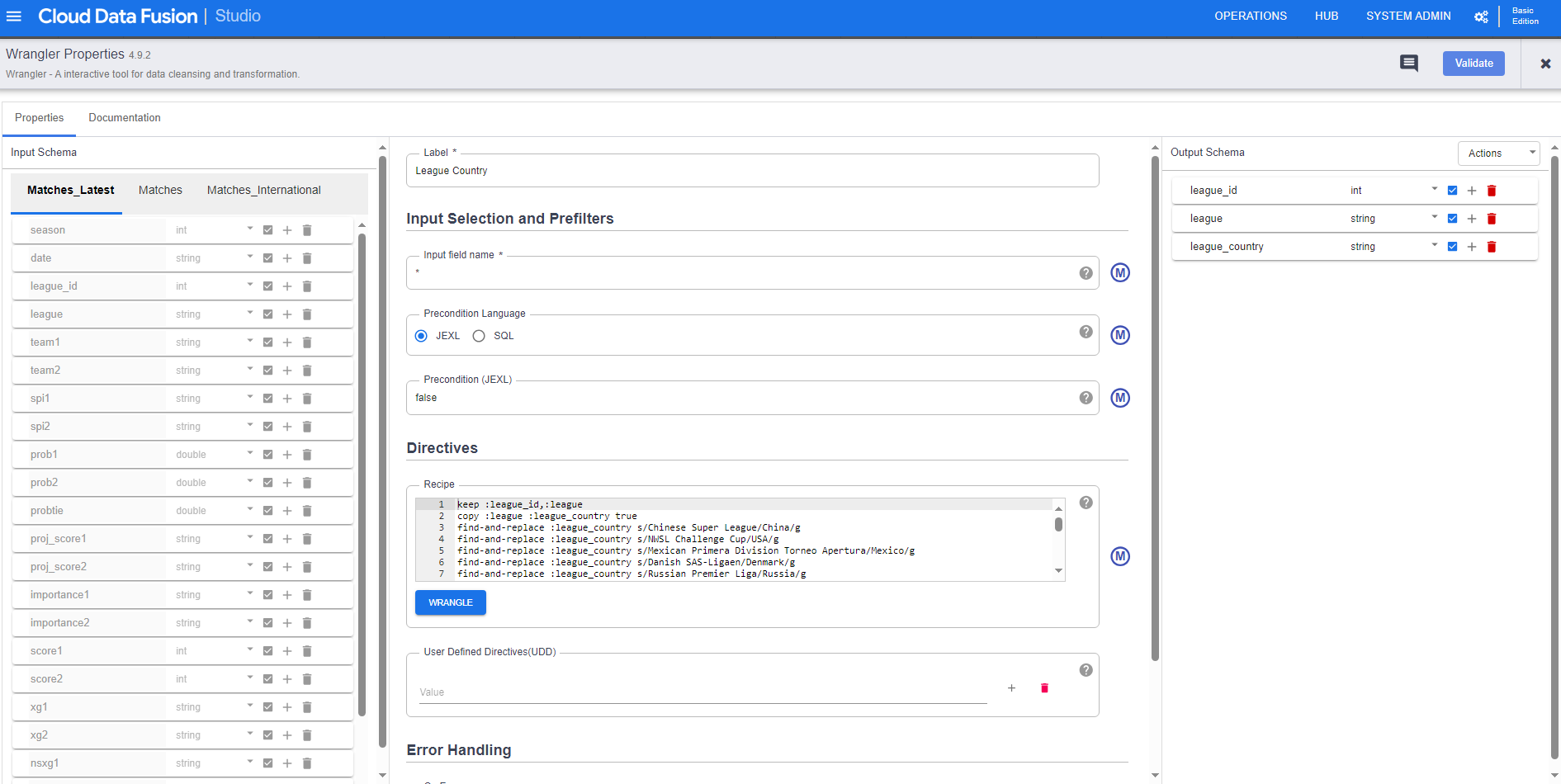
Após o join das tabelas é feito um tratamento de nulos e o resultado é salvo no BigQuery como a tabela **Leagues**.

Visão geral do fluxo:



*Fluxo ETL da Tabela Leagues*

Visão geral do bloco WRANGLER e transformações aplicadas. É criada uma nova coluna na tabela e é feito um de-para manual para atribuir o país da liga a partir de seu nome:



keep :league\_id,:league

copy :league :league\_country true

find-and-replace :league\_country s/Chinese Super League/China/g

find-and-replace :league\_country s/NWSL Challenge Cup/USA/g

find-and-replace :league\_country s/Mexican Primera Division Torneo Apertura/Mexico/g

find-and-replace :league\_country s/Danish SAS-Ligaen/Denmark/g

find-and-replace :league\_country s/Russian Premier Liga/Russia/g

find-and-replace :league\_country s/German 2. Bundesliga/Germany/g

find-and-replace :league\_country s/Swiss Raiffeisen Super League/Switzerland/g

find-and-replace :league\_country s/Austrian T-Mobile Bundesliga/Austria/g

find-and-replace :league\_country s/Belgian Jupiler League/Belgium/g

find-and-replace :league\_country s/English League Championship/England/g

find-and-replace :league\_country s/Scottish Premiership/Scotland/g

find-and-replace :league\_country s/French Ligue 2/France/g

find-and-replace :league\_country s/English League Two/England/g

find-and-replace :league\_country s/English League One/England/g

find-and-replace :league\_country s/South African ABSA Premier League/South Africa/g

find-and-replace :league\_country s/Turkish Turkcell Super Lig/Turkey/g

find-and-replace :league\_country s/Dutch Eredivisie/Netherlands/g

find-and-replace :league\_country s/German Bundesliga/Germany/g

find-and-replace :league\_country s/French Ligue 1/France/g

find-and-replace :league\_country s/Barclays Premier League/England/g

find-and-replace :league\_country s/Portuguese Liga/Portugal/g

find-and-replace :league\_country s/Italy Serie B/Italy/g

find-and-replace :league\_country s/Spanish Segunda Division/Spain/g

find-and-replace :league\_country s/Spanish Primera Division/Spain/g

find-and-replace :league\_country s/Italy Serie A/Italy/g

find-and-replace :league\_country s/Greek Super League/Greece/g

find-and-replace :league\_country s/UEFA Champions League/Europe/g

find-and-replace :league\_country s/UEFA Europa Conference League/Europe/g

find-and-replace :league\_country s/UEFA Europa League/Europe/g

find-and-replace :league\_country s/FA Women's Super League/England/g

find-and-replace :league\_country s/Australian A-League/Australia/g

find-and-replace :league\_country s/Mexican Primera Division Torneo Clausura/Mexico/g

find-and-replace :league\_country s/Argentina Primera Division/Argentina/g

find-and-replace :league\_country s/Japanese J League/Japan/g

find-and-replace :league\_country s/Major League Soccer/USA/g

find-and-replace :league\_country s/United Soccer League/USA/g

find-and-replace :league\_country s/National Women's Soccer League/USA/g

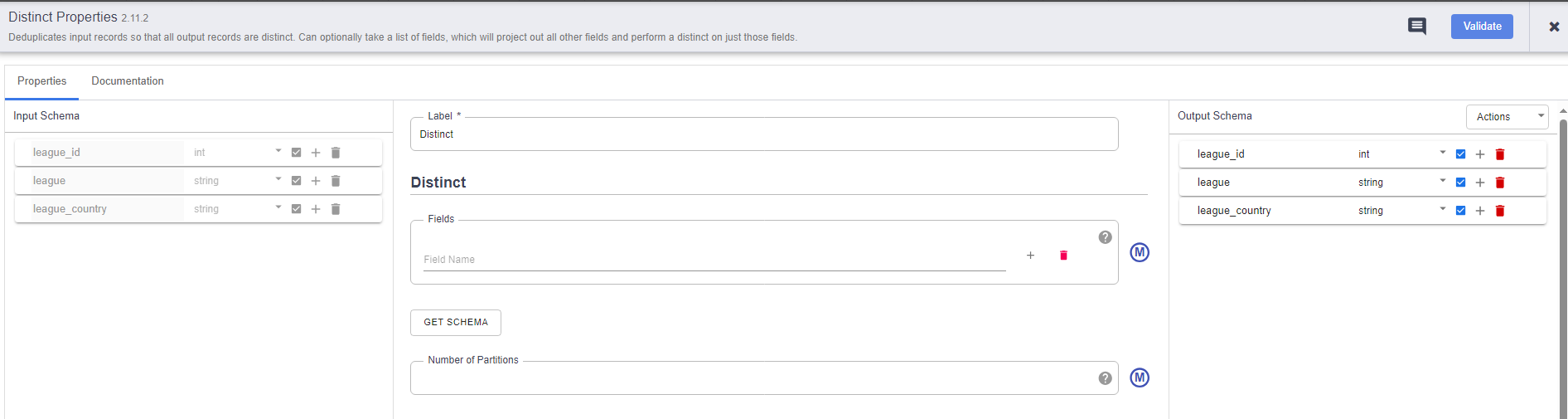
find-and-replace :league\_country s/Swedish Allsvenskan/Sweden/g

find-and-replace :league\_country s/Norwegian Tippeligaen/Norway/g

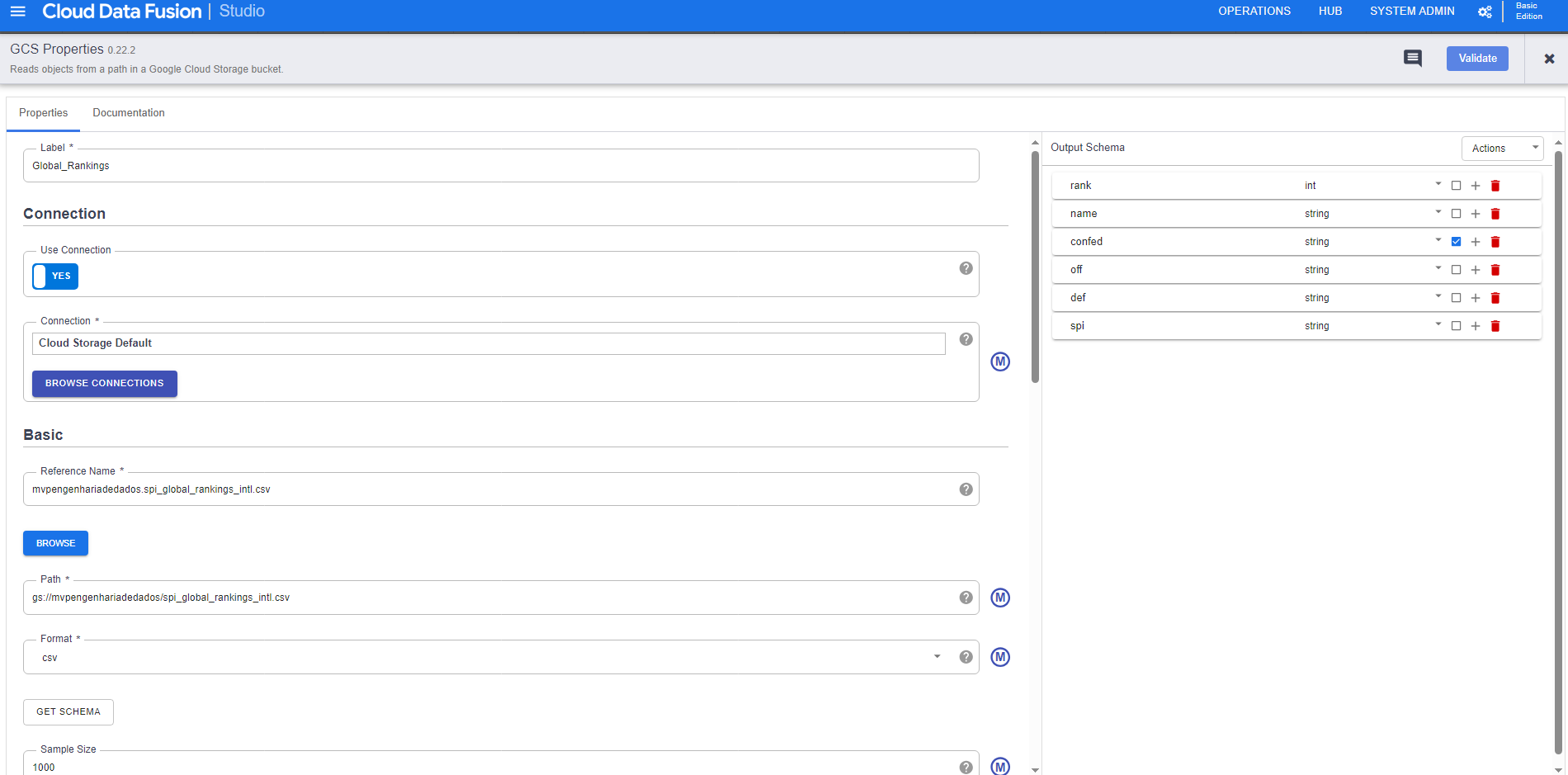
find-and-replace :league\_country s/Brasileiro Série A/Brazil/g

find-and-replace :league s/Brasileiro Série A/Brasileiro Serie A/g

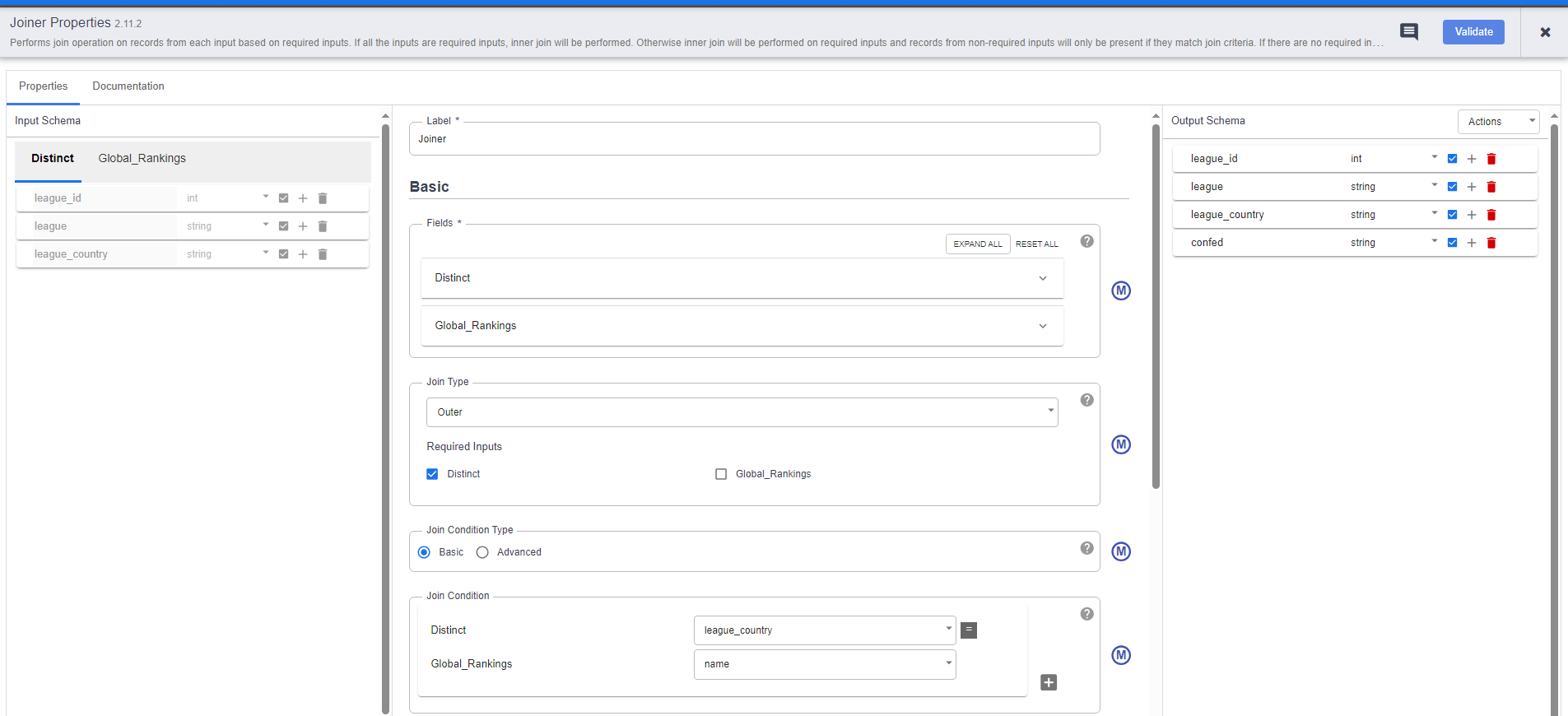
Visão geral do bloco DISTINCT, que remove duplicatas do conjunto:



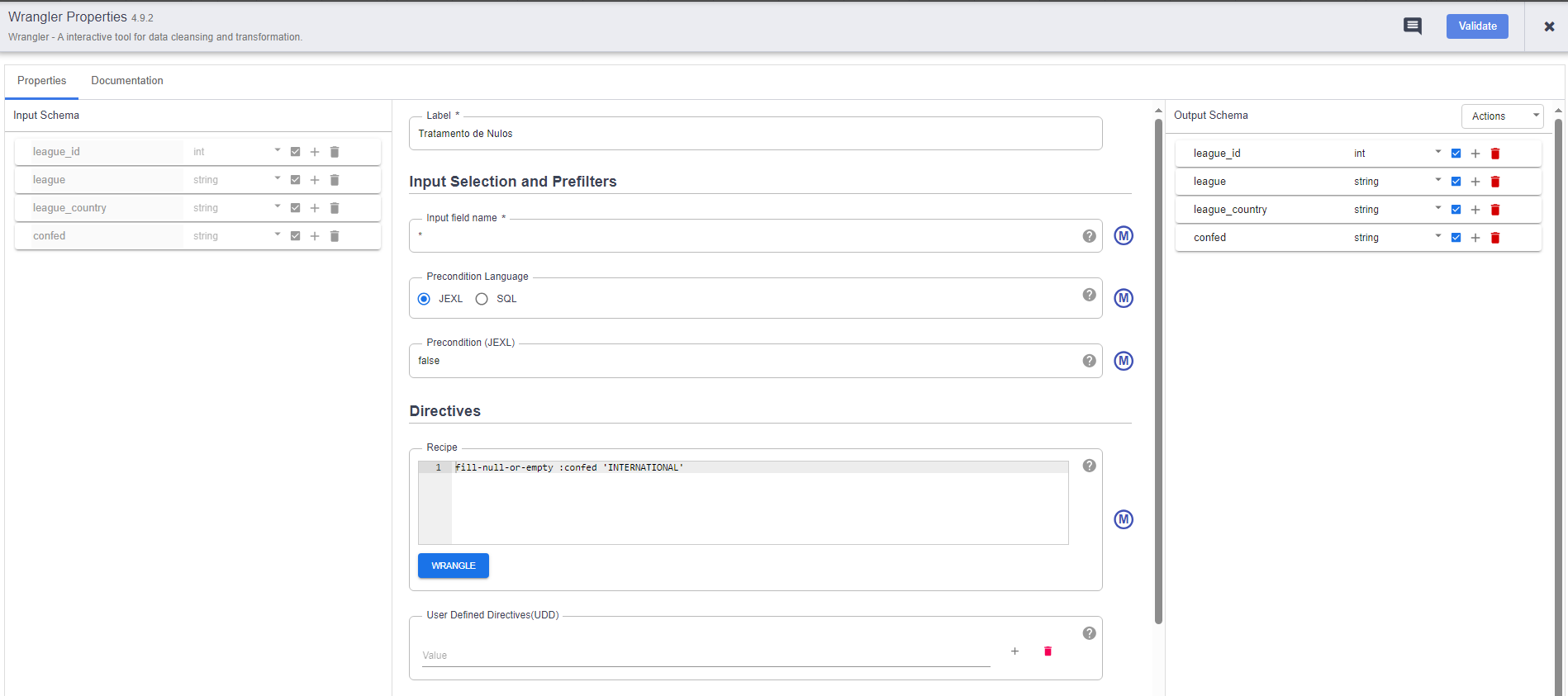
Carregamento do csv *global\_rankings\_intl*:



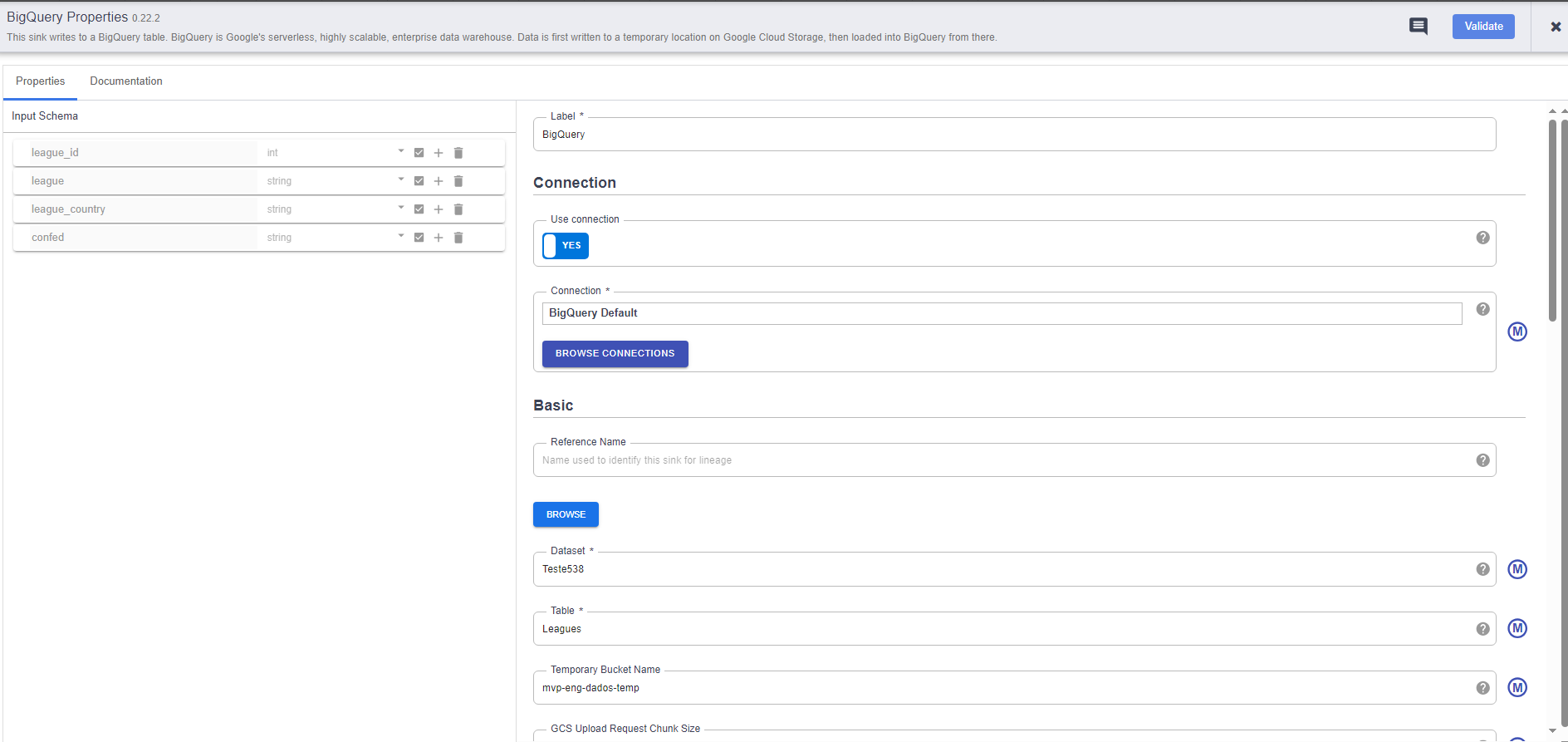
Visão geral do bloco de JOIN. A união é feita entre os campos league\_country (que vem da transformação de *matches*) e name (que vem de *global\_ranking\_intl*):



No bloco de tratamento de nulos, é aplicada uma transformação para atribuir a categoria “INTERNATIONAL” para qualquer registro que tenha ficado com o campo *confed* = nulo :



Por fim, a tabela é salva no BigQuery como Leagues:



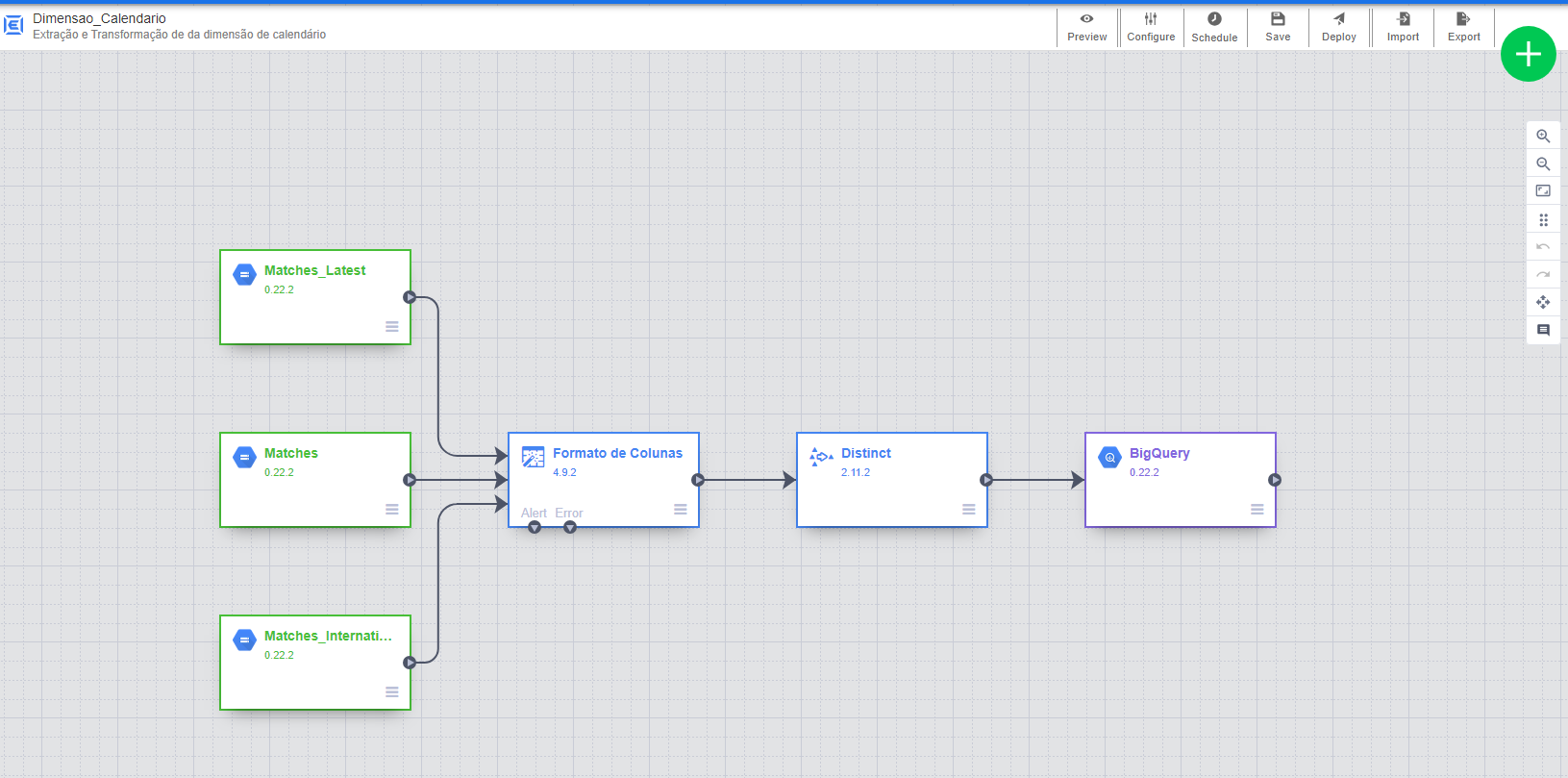
### **Tabela Calendar**

A tabela Calendar também tem origem nos arquivos *csv matches, matches\_latest* e *matches\_intl*.

Após a união destas 3 fontes, são feitas transformações em cima da coluna date para extrair informações das datas presentes no conjunto de dados e a partir daí criar as demais colunas da dimensão de calendário.

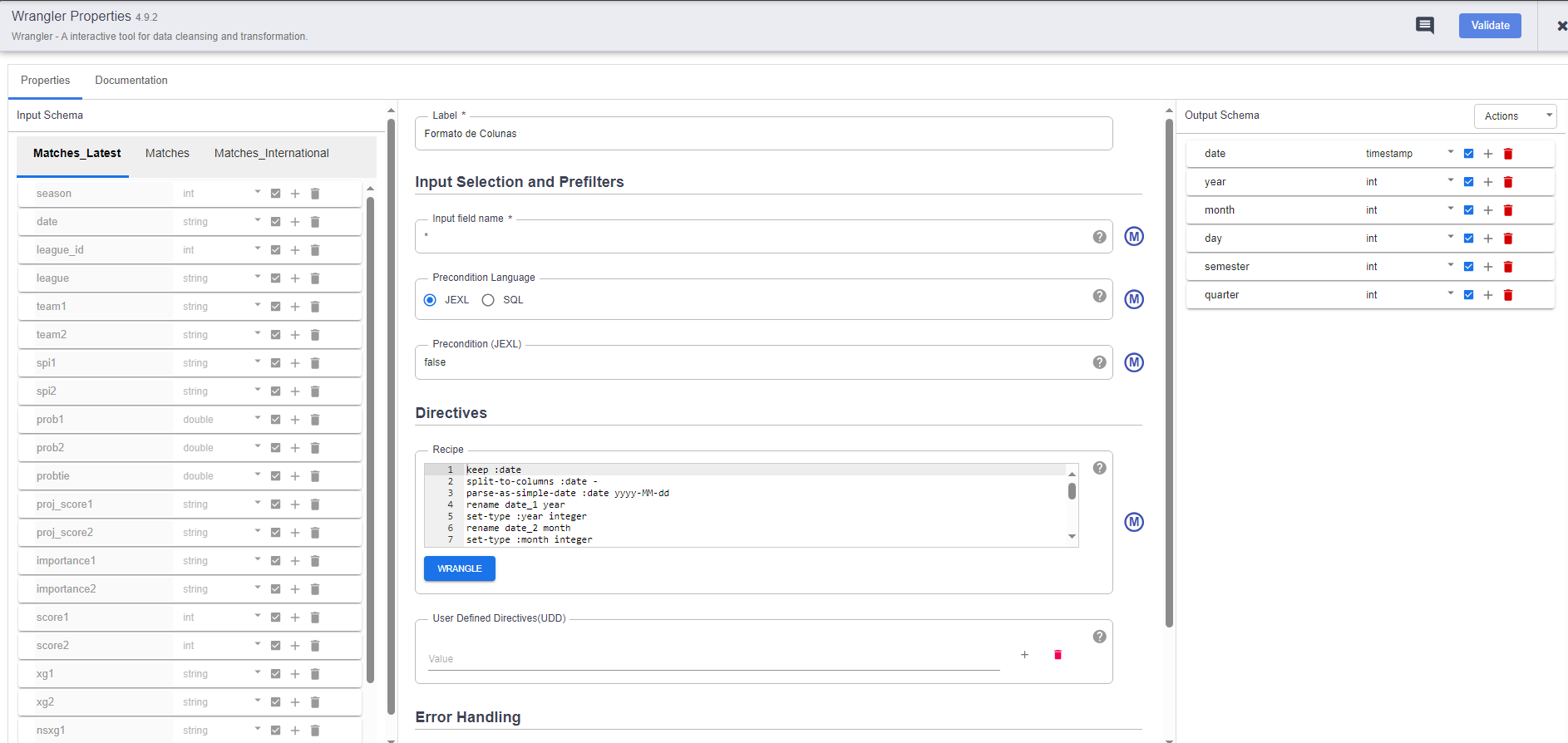
O bloco de Distinct busca filtrar quaisquer duplicatas e na sequência a tabela é salva no BigQuery.

Visão geral do fluxo:



*Fluxo ETL da Tabela Calendar*

Visão geral do bloco WRANGLER e transformações aplicadas. A partir da coluna de data extrai-se os dados de ano, mês e dia. Na sequência, usa-se o campo de mês para calcular os campos de semestre e trimestre:



keep :date

split-to-columns :date -

parse-as-simple-date :date yyyy-MM-dd

rename date\_1 year

set-type :year integer

rename date\_2 month

set-type :month integer

rename date\_3 day

set-type :day integer

copy :month :semester true

set-type :semester double

set-column :semester semester / 6

set-column :semester math:ceil(semester)

set-type :semester integer

copy :month :quarter true

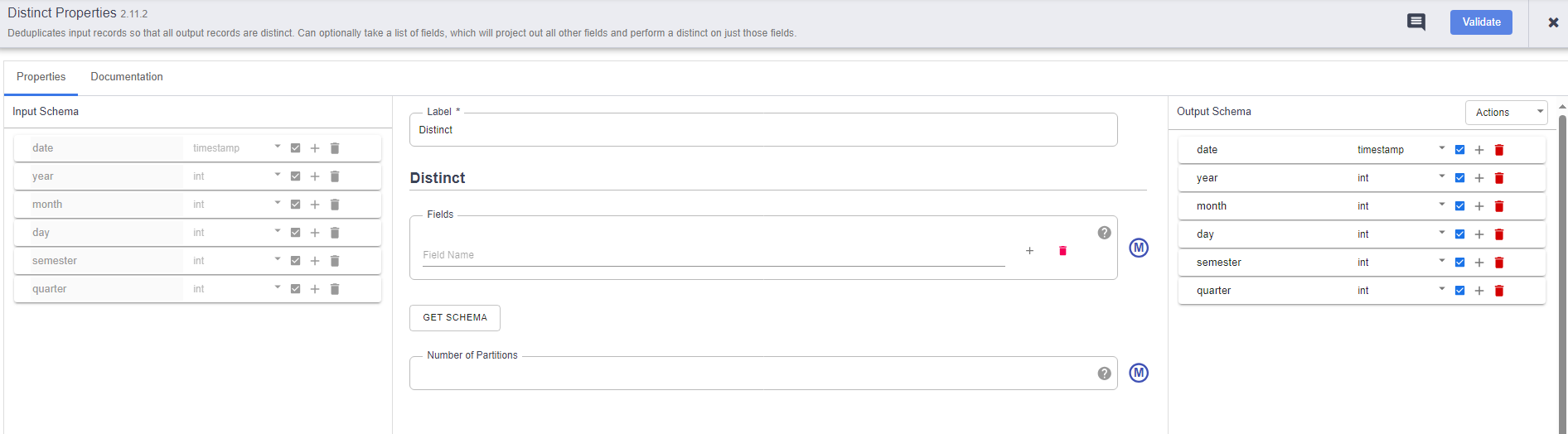
set-type :quarter double

set-column :quarter quarter / 3

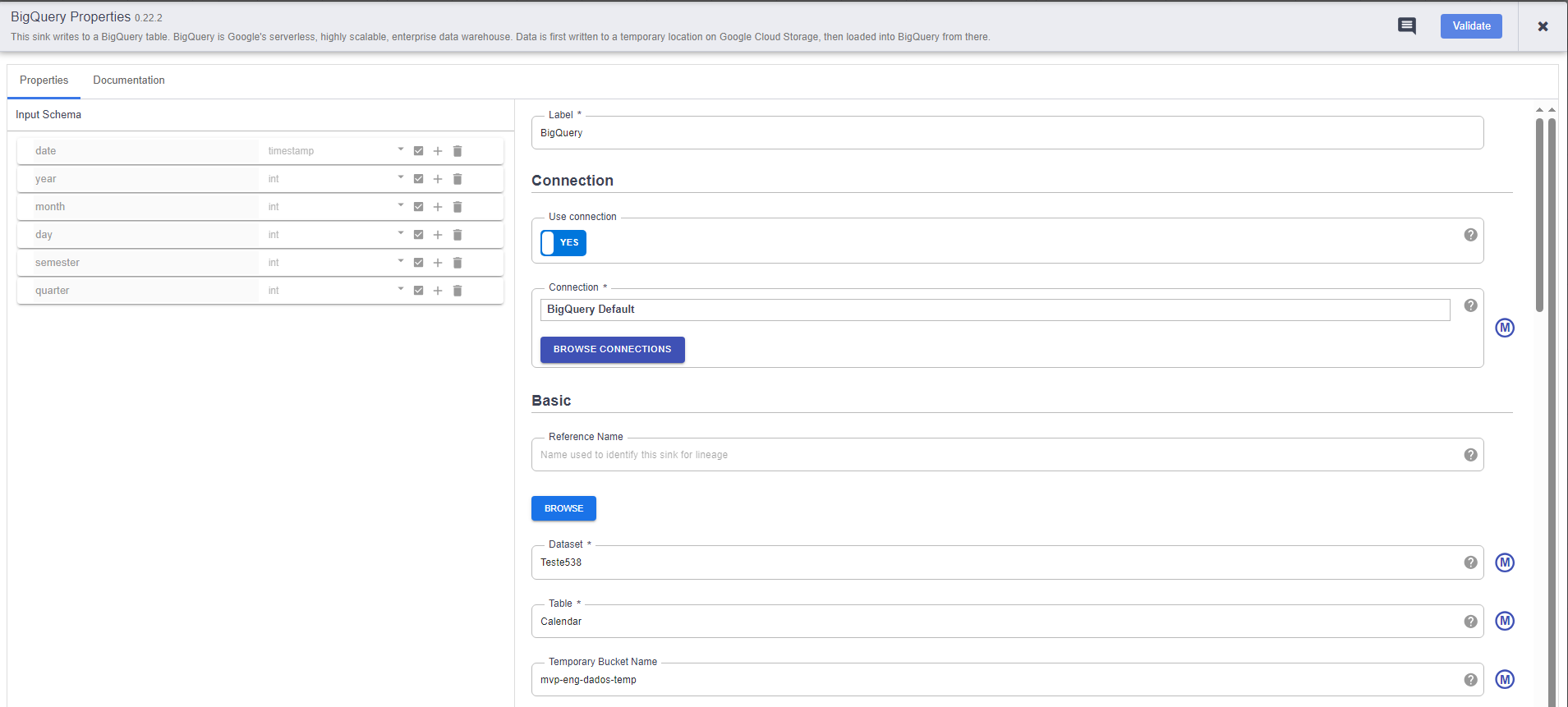
set-column :quarter math:ceil(quarter)

set-type :quarter integer

Visão geral do bloco DISTINCT:



Após remover duplicatas, a tabela é salva no BigQuery como Calendar:



## Análise de Dados

### **Qualidade de Dados**

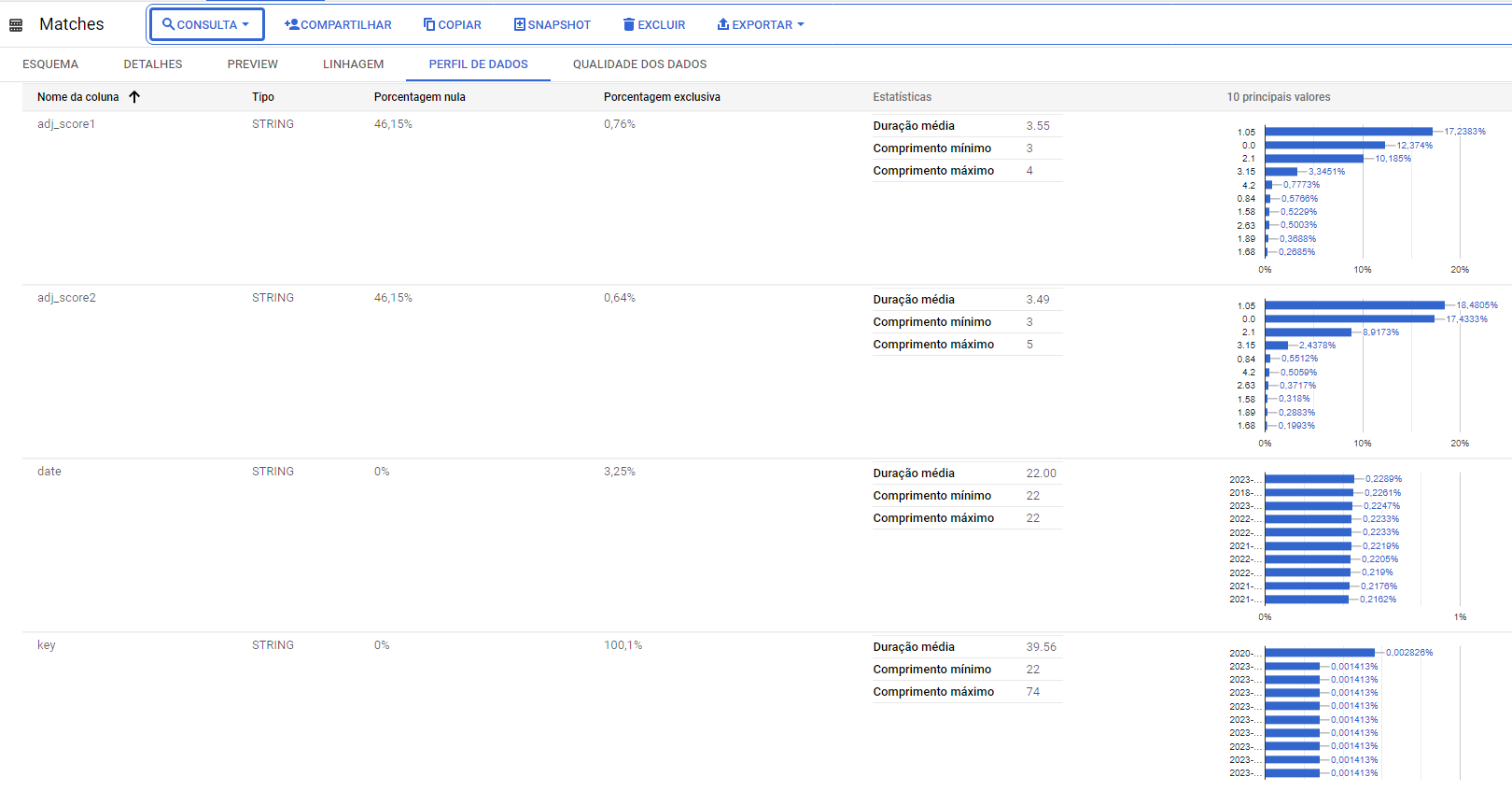
Analisaremos nesta seção a qualidade de dados do conjunto como um todo, analisando os valores dos atributos de cada uma das tabelas. Para auxiliar a análise de qualidade de dados, usaremos nesta seção a funcionalidade de ‘Perfil de Dados’ disponível no BigQuery.

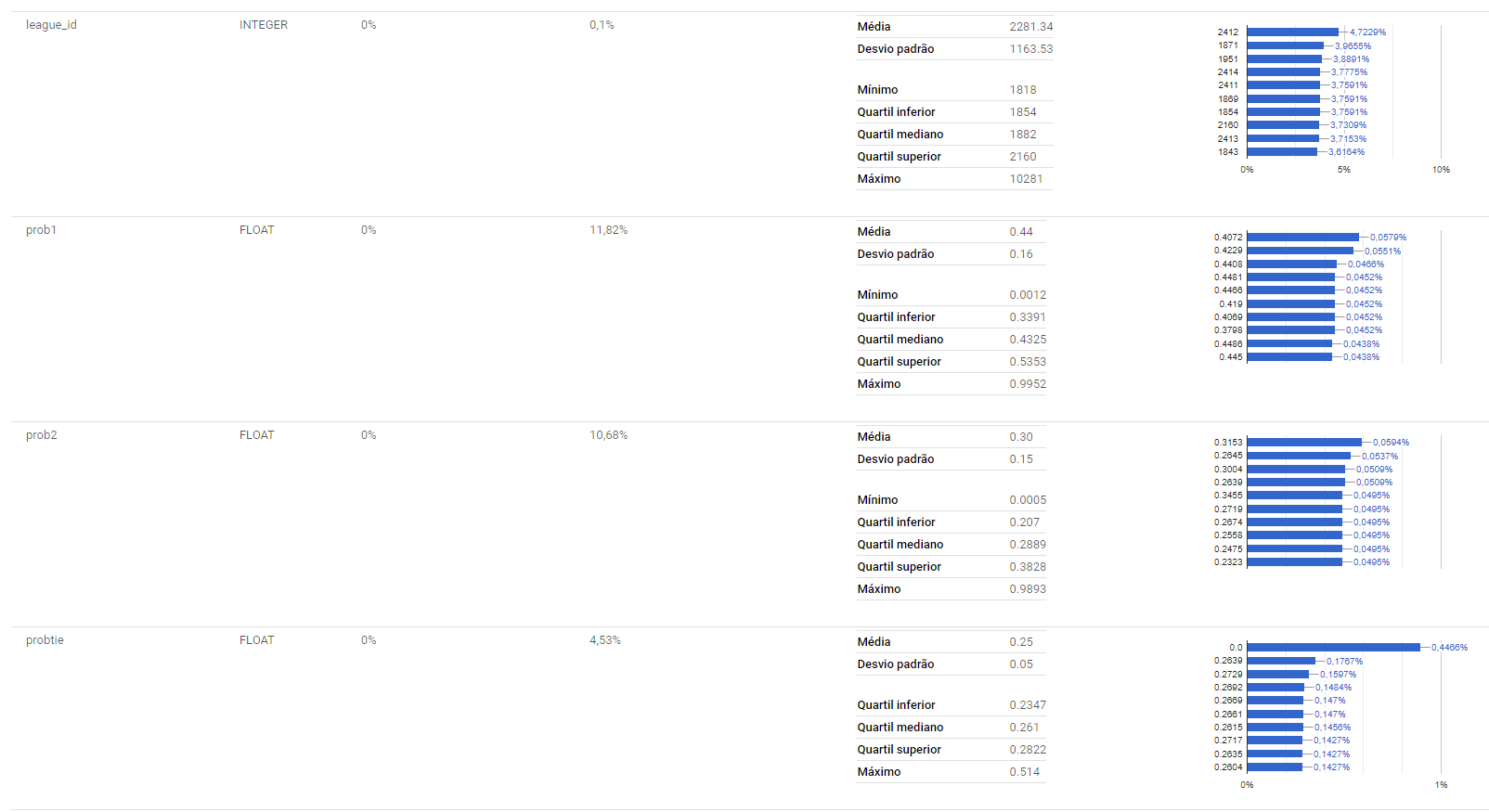
Vale ressaltar que foi feito um tratamento de valores nulos para alguns campos do modelo durante a etapa de ETL. Em particular, foram feitas transformações para filtrar todos os registros da tabela **Matches** nos quais o placar da partida era inválido, isto é, com o campo de *score1* ou *score2* nulos (se não há informação de placar de uma das equipes, não há como saber o resultado da partida, logo não teríamos como classificar como zebra ou não).

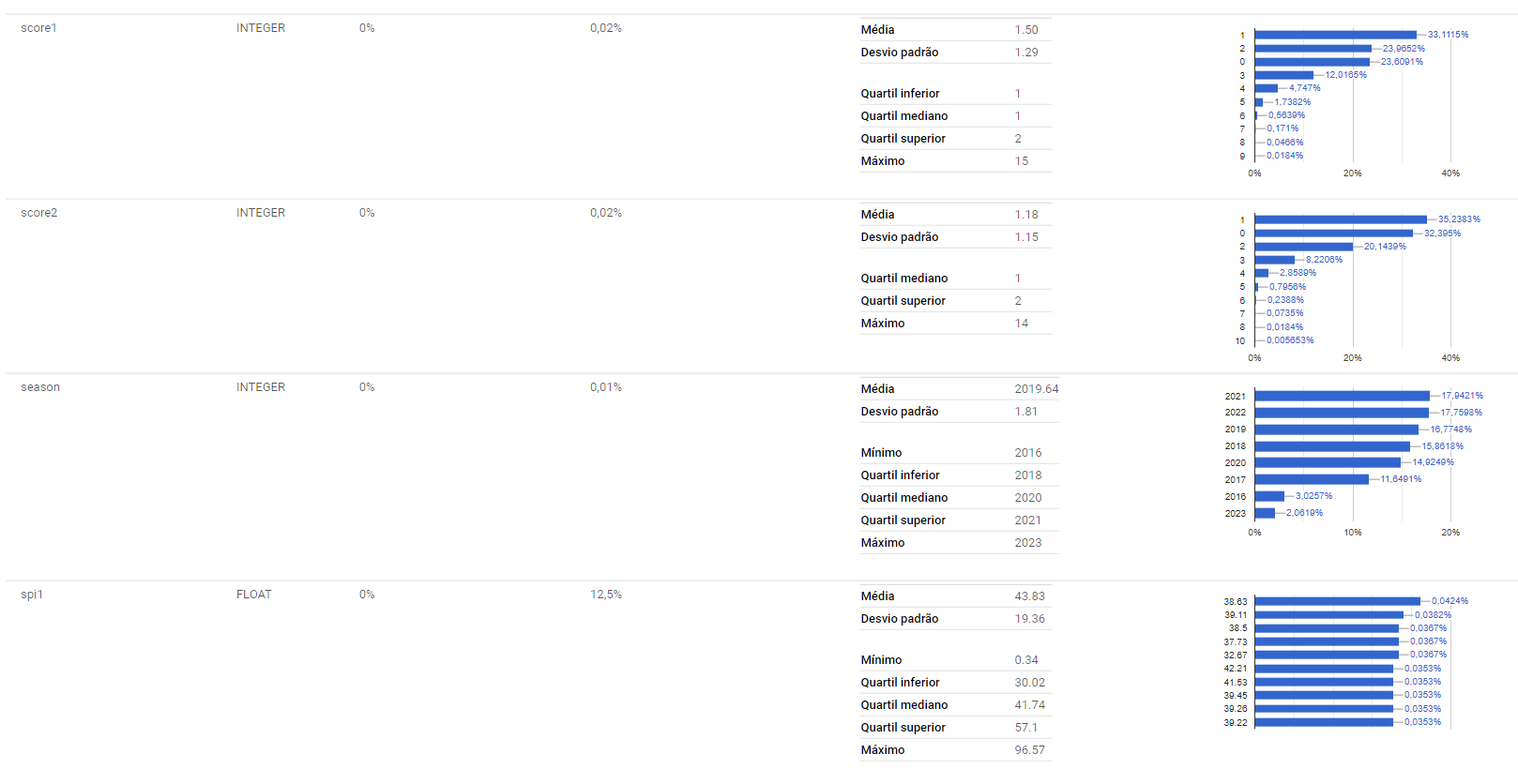
Os registros filtrados nesta etapa de ETL foram salvos na tabela **Erros** do BigQuery. Para fins de comparação, a tabela **Erros** foi salva com 4199 registros, diante de 70671 da tabela **Matches**.  
Isto significa que 5,6% dos dados originais foram filtrados e descartados da análise final. Apesar de 5,6% de registros nulos ser um valor significativo, os outros 70671 registros válidos serão suficientes para a análise proposta nesse projeto.

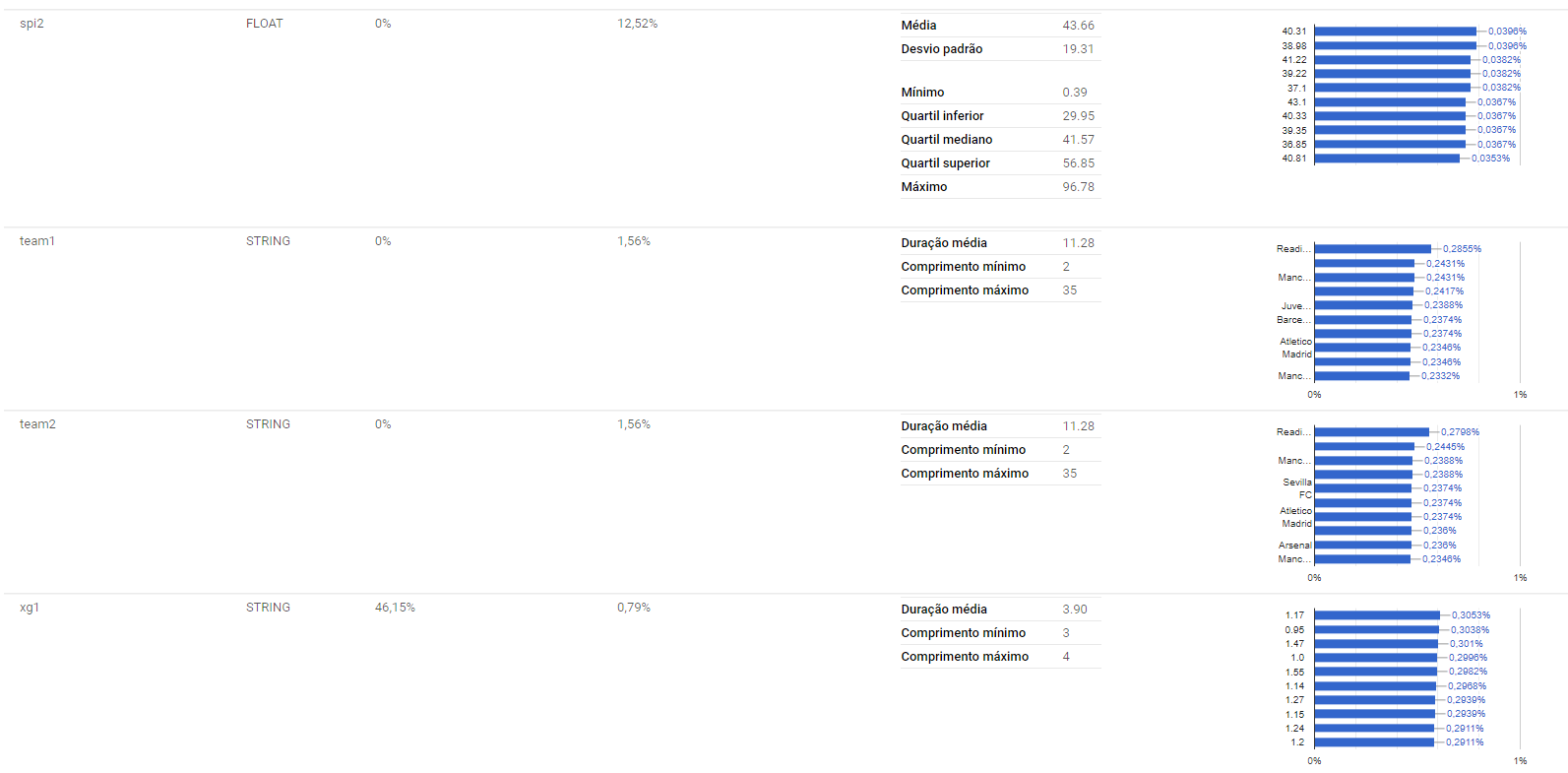
### **Tabela Matches**

Para a tabela Matches, aba Perfil de Dados retorna os seguintes valores:





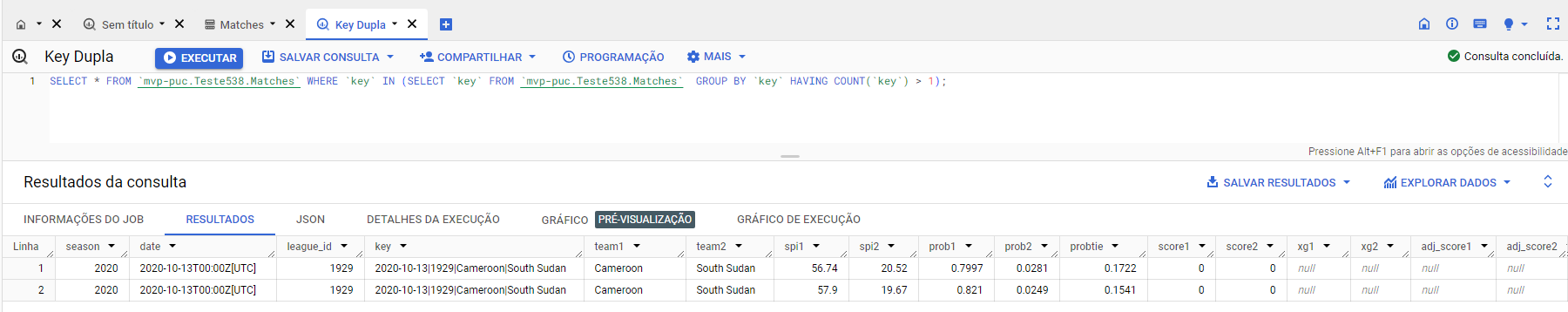






*Visão geral da qualidade de dados dos campos da tabela Matches*

* **adj\_score1**: o campo não está ideal pois contém muitos valores nulos (46,15% dos registros são nulos). De toda forma, este não é um campo determinante para as análises propostas no escopo deste projeto, portanto estes registros nulos não terão impacto no resultado final;
* **adj\_score2**: o campo não está ideal pois contém muitos valores nulos (46,15% dos registros são nulos). De toda forma, este não é um campo determinante para as análises propostas no escopo deste projeto, portanto estes registros nulos não terão impacto no resultado final;
* **date**: o campo está OK, pois não há dados nulos e ele apresenta valores coerentes com os esperados para o campo, sendo todos os registros no formato de data, com o comprimento máximo = comprimento mínimo = 22 (total de caracteres do formato de data: YYYY-MM-DDTHH:MMZ[UTC] );
* **key**: o campo está parcialmente OK: embora esteja correto por não apresentar nenhum valor nulo e estar dentro das regras de formação da key, vemos que há um registro com duas ocorrências.   
  Como estamos falando de um campo que tem o intuito de ser chave da tabela, o esperado seria que não houvesse duplicidade de registros.   
  A query abaixo identifica onde houve o problema:



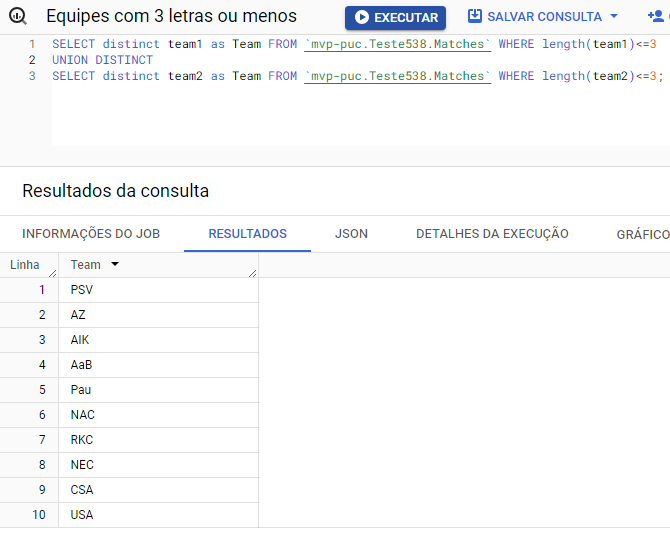
Podemos ver que a partida entre “Cameroon” e “South Sudan” está aparecendo duas vezes na tabela, porém com SPIs e probabilidades diferentes em cada um dos registros.  
 Este é um erro que veio desde a origem dos dados, pois não foi considerada na etapa de transformação a hipótese de se ter registros múltiplos de um mesmo jogo, porém com valores de probabilidade diferentes em cada caso.  
Como se trata de um único registro na amostra de dados, não é um erro que vai comprometer a análise proposta. Mas, de toda forma, fica como ponto de melhoria para uma nova versão do projeto, na qual o ETL deve passar a considerar os campos de probabilidade para a composição da chave da tabela.

* **league\_id**: o campo está OK, pois não há dados nulos e ele apresenta valores coerentes com os esperados para o campo, sendo o valor mínimo=1818 e o valor máximo=10281;
* **prob1**: o campo está OK, pois não há dados nulos e ele apresenta valores coerentes com os esperados para o campo, estando todos eles dentro do range de 0 (0%) a 1 (100%);
* **prob2**: o campo está OK, pois não há dados nulos e ele apresenta valores coerentes com os esperados para o campo, estando todos eles dentro do range de 0 (0%) a 1 (100%);
* **probtie**: o campo está OK, pois não há dados nulos e ele apresenta valores coerentes com os esperados para o campo, estando todos eles dentro do range de 0 (0%) a 1 (100%);
* **score1**: o campo está OK, pois não há dados nulos e ele apresenta valores coerentes com os esperados para o campo, sem números negativos ou decimais.   
  Porém, vemos que tivemos valores maiores que 10, que era o range esperado e definido no Catálogo de Dados.  
  Isso significa que tivemos partidas onde a equipe mandante marcou mais de 10 gols. A query abaixo mostra quais foram esses jogos:



Tivemos em toda a base de dados 7 jogos onde a equipe 1 ou a equipe 2 anotou mais de 10 gols no jogo. Embora este seja um valor que a princípio não era esperado no range de valores, podemos interpretar que não são valores errados.   
De fato, ao olhar para a probabilidade das equipes envolvidas em cada um destes jogos, podemos ver que se tratava de jogos de grande disparidade de força, com uma das equipes muito mais forte – com uma probabilidade de vitória – muito maior que a outra.  
Desta forma, entende-se que são valores também coerentes, apesar de raros.

* **score2**: o campo está OK, pois não há dados nulos e ele apresenta valores coerentes com os esperados para o campo, sem números negativos ou decimais. A análise do range de dados feita acima vale também para este campo;
* **season**: o campo está OK, pois não há dados nulos e ele apresenta valores coerentes com os esperados para o campo, sendo o valor mínimo=2016 e o valor máximo=2023;
* **spi1**: o campo está OK, pois não há dados nulos e ele apresenta valores coerentes com os esperados para o campo, dentro do range 0-100;
* **spi2**: o campo está OK, pois não há dados nulos e ele apresenta valores coerentes com os esperados para o campo, dentro do range 0-100;
* **team1**: o campo está OK, pois não há dados nulos e ele apresenta valores coerentes com os esperados para o campo. Podemos desconfiar da informação de ‘comprimento mínimo’ = 2, pois isso significa que alguma equipe tem apenas 2 letras no nome.   
  Para verificar a qualidade destes registros, rodamos a seguinte query:

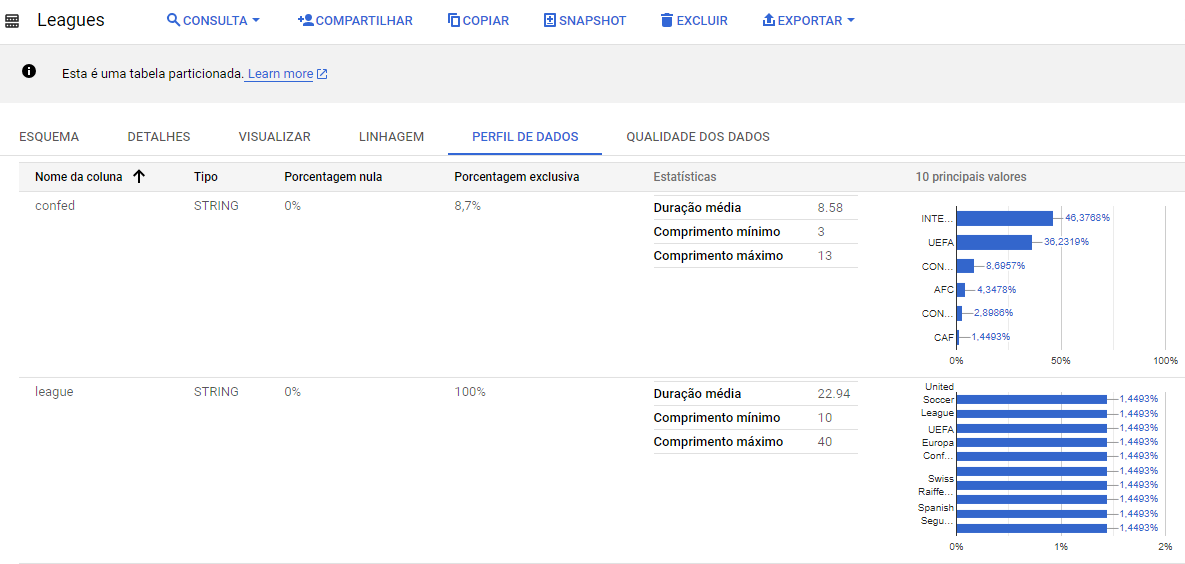


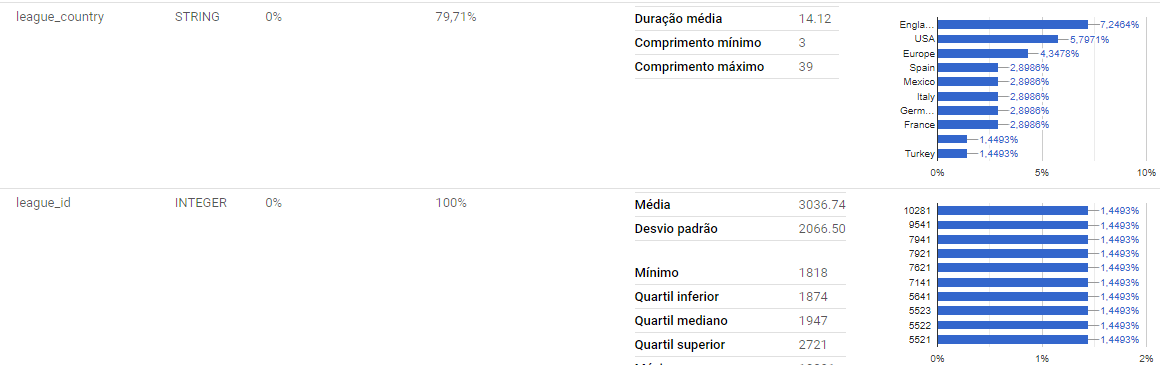
Esta query mostra que de fato temos equipes com nomes de apenas 2 ou 3 letras, que no caso representam siglas/abreviações dos nomes completos das mesmas.   
Desta forma, confirmamos que o campo está correto;

* **team2**: o campo está OK, pois não há dados nulos e ele apresenta valores coerentes com os esperados para o campo. Os valores de comprimento mínimo são válidos, conforme visto na query acima;
* **xg1**: o campo não está ideal pois contém muitos valores nulos (46,15% dos registros são nulos). De toda forma, este não é um campo determinante para as análises propostas no escopo deste projeto, portanto estes registros nulos não terão impacto no resultado final;
* **xg2** : o campo não está ideal pois contém muitos valores nulos (46,15% dos registros são nulos). De toda forma, este não é um campo determinante para as análises propostas no escopo deste projeto, portanto estes registros nulos não terão impacto no resultado final;

### **Tabela Leagues**

Para a tabela Leagues, aba Perfil de Dados retorna os seguintes valores:





*Visão geral da qualidade de dados dos campos da tabela Leagues*

* **confed**: o campo está OK por não ter nulos, porém chama a atenção a quantidade de ligas pertencentes à confederação “INTERNACIONAL”. Na query abaixo podemos ver quais são essas ligas:

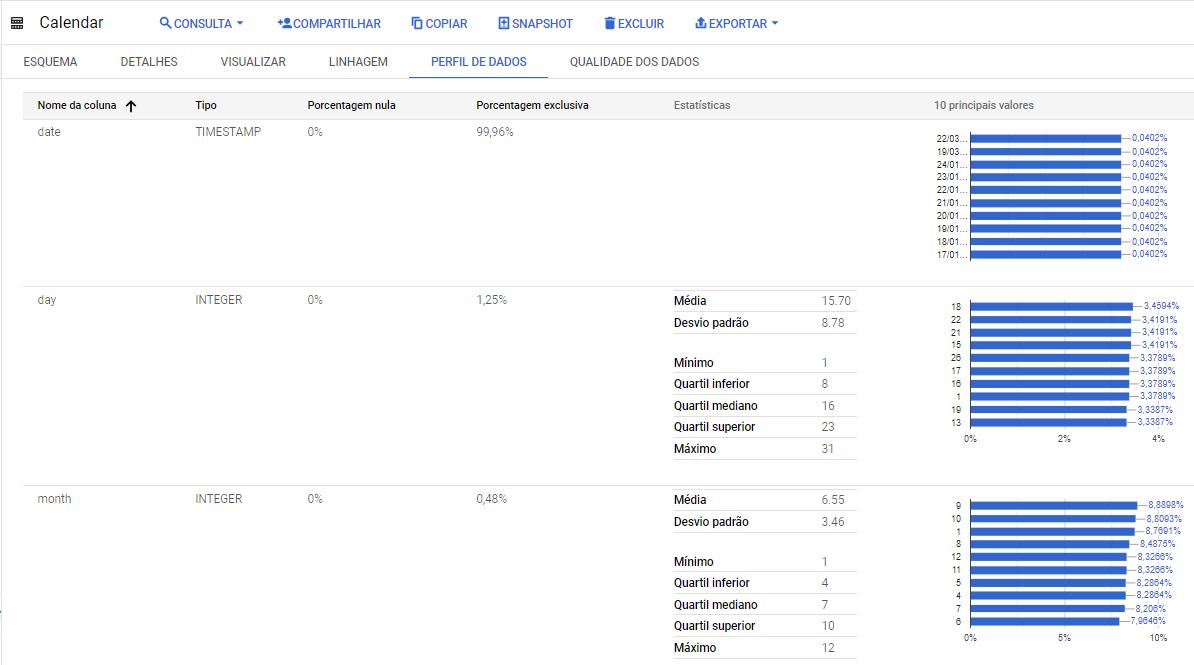
****

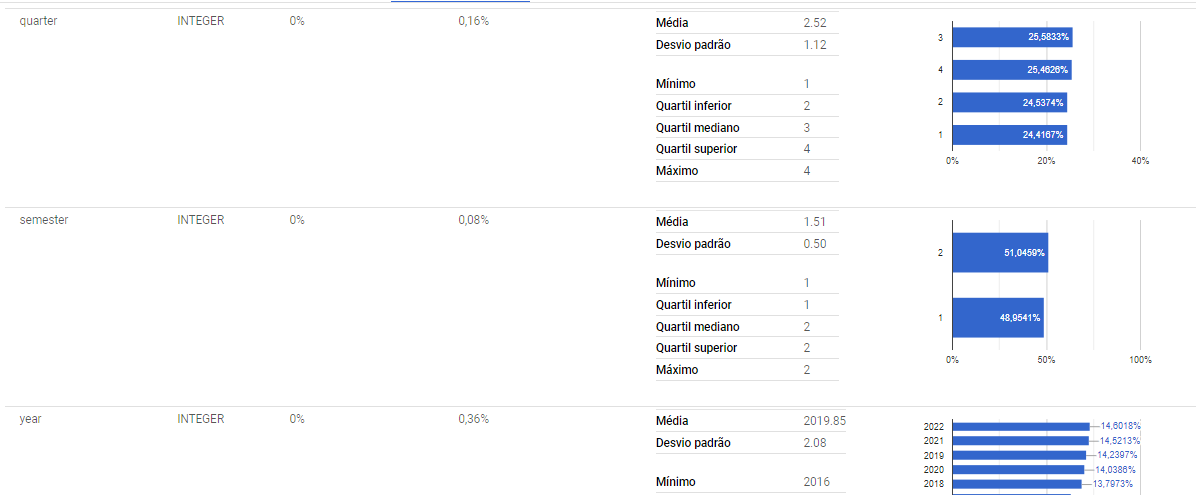
Pode-se ver que as ligas de fato são entre seleções, caindo na categoria INTERNATIONAL, porém, algumas competições poderiam também ser consideradas dentro das confederações continentais. Por exemplo, AFC Asian Cup pertence à AFC, Copa America pertence à CONMEBOL e assim por diante.  
Embora não seja exatamente um erro, é um fator que poderia dar mais profundidade às análises.

* **league**: o campo está OK, pois não há dados nulos, os valores são coerentes com os esperados para o campo e cada liga está registrando apenas uma vez, conforme esperado dentro da dimensão;
* **league\_country**: o campo está OK, pois não há dados nulos, os valores são coerentes com os esperados para o campo e a variação dentre a quantidade de cada registro está dentro do esperado: no caso, pode-se perceber que temos mais ligas da Inglaterra e dos EUA na base de dados;
* **league\_ID**: o campo está OK, pois não há dados nulos, os valores são coerentes com os esperados para o campo e cada ID está registrando apenas uma vez, conforme esperado dentro da dimensão;

### **Tabela Calendar**

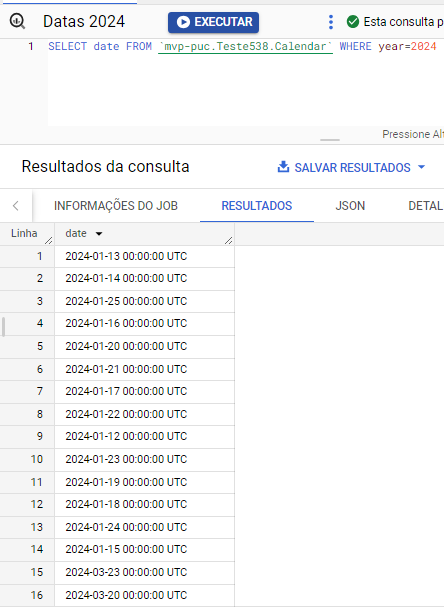
Para a tabela Calendar, aba Perfil de Dados retorna os seguintes valores:





*Visão geral da qualidade de dados dos campos da tabela Calendar*

* **date**: o campo está OK, pois não há dados nulos e vemos que não há datas repetidas. Isso pode ser visto através do gráfico ’10 principais valores’, onde cada data tem apenas 1 registro;
* **day**: o campo está OK, pois não há dados nulos e ele apresenta valores coerentes com os esperados para o campo, sendo o valor mínimo=1 e o valor máximo=31;
* **month**: o campo está OK, pois não há dados nulos e ele apresenta valores coerentes com os esperados para o campo, sendo o valor mínimo=1 e o valor máximo=12;
* **quarter**: o campo está OK, pois não há dados nulos e ele apresenta valores coerentes com os esperados para o campo, sendo o valor mínimo=1 e o valor máximo=4;
* **semester**: o campo está OK, pois não há dados nulos e ele apresenta valores coerentes com os esperados para o campo, sendo o valor mínimo=1 e o valor máximo=2;
* **year**: não há dados nulos no campo, mas ele apresenta valores fora dos esperados para o campo. O valor mínimo=2016 está OK, porém o valor máximo=2024 não faz sentido, pois não há ainda resultados para jogos de 2024. Com a query abaixo podemos ver quais são esses registros:



Ou seja, temos 16 registros que fogem do range adequado do campo. Considerando que a tabela tem 2486 datas no total, podemos dizer que há um erro de 0,64% neste campo, o que não compromete a análise do projeto.

### **Solução das Perguntas**

Nesta seção, vamos demonstrar e analisar queries em cima do conjunto de dados para responder as perguntas propostas no escopo. Recapitulando, o objetivo era responder os 03 pontos abaixo:

1. Qual é o campeonato mais disputado do mundo (isto é, aquele onde os favoritos mais têm revezes)?
2. O quão imprevisível é o Campeonato Brasileiro? É de fato um campeonato mais disputado que os demais?
3. Existe alguma tendência temporal para o acontecimento das zebras? Elas estão aumentando ou diminuindo?

Vamos analisar cada pergunta individualmente:

#### Campeonato com mais Zebras

Para responder esta pergunta, vamos partir da seguinte abordagem: uma zebra é quando uma equipe favorita antes de um jogo – isto é, uma equipe com mais probabilidade de vitória – não consegue um resultado positivo.

Ou ainda, podemos olhar uma zebra como sendo um jogo onde uma equipe menos favorita (com menos probabilidade de vencer que a outra equipe) conseguiu um resultado positivo.

Sob um ponto de vista matemático, podemos pensar que um jogo de futebol pode ter 03 resultados: vitória da equipe mandante, vitória da equipe visitante ou empate. Em um jogo idealmente equilibrado, as probabilidades de resultado seriam de 1/3 para cada resultado (33,3%).   
  
Para simplificar nossa análise, vamos considerar como zebra apenas situações onde uma equipe era mais favorita que a adversária e perdeu. Desta forma, estaremos desconsiderando situações onde uma equipe mais favorita apenas empatou.

Para completar a análise, precisamos definir um valor de probabilidade que signifique um grande favoritismo, a ponto de uma derrota ser considerada zebra.

Este ponto é importante pois, em um duelo mais equilibrado, onde por exemplo a equipe 1 tem 40% de probabilidade de vitória, a equipe 2 tem 35% de probabilidade e o empate tem 25% de acontecer, uma eventual vitória da equipe 2 não deve ser considerada como zebra.

Nesse contexto, definimos aqui de forma arbitrária que a zebra acontece em jogos onde a equipe mais favorita antes do jogo tinha 50% ou mais de probabilidade de vitória e o adversário tinha no máximo 25% de probabilidade de vencer, de forma que a equipe favorita era cotada com uma probabilidade no mínimo 2x maior de obter a vitória.

Isto é traduzido na seguinte query:

SELECT A.league\_id,B.league, COUNT(\*) as zebras, C.total\_jogos, ROUND(COUNT(\*)/C.total\_jogos,3) as pct\_zebras

FROM `mvp-puc.Teste538.Matches` A

  LEFT JOIN (SELECT league\_id, league

            FROM `mvp-puc.Teste538.Leagues`) B

  ON A.league\_id = B.league\_id

    LEFT JOIN (SELECT league\_id, COUNT(\*) as total\_jogos

              FROM `mvp-puc.Teste538.Matches`

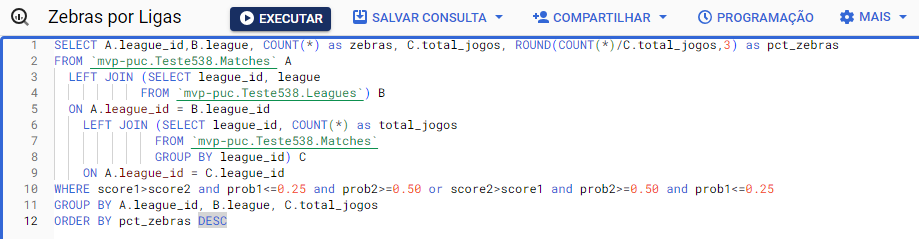
              GROUP BY league\_id) C

    ON A.league\_id = C.league\_id

WHERE score1>score2 and prob1<=0.25 and prob2>=0.50 or score2>score1 and prob2>=0.50 and prob1<=0.25

GROUP BY A.league\_id, B.league, C.total\_jogos

ORDER BY pct\_zebras DESC



Nesta query são feitos dois joins. Partindo de dentro pra fora, pegamos primeiro o total de jogos de cada liga dentro da base (alias C). O segundo join é para trazer o nome das ligas presente na dimensão Leagues, e isso é feito no join com a tabela de alias B.

Por fim, realizamos SELECT do id da liga, o nome da mesma, o total de jogos de cada liga e o total de jogos “zebra” em cada liga, isto é, o total de jogos onde o time A tinha uma probabilidade de vitória de 50% ou mais e o time B tinha uma probabilidade de 25% ou menos, e mesmo assim o time B venceu.

Para determinar a vitória de um time, no caso, comparamos os campos de score de cada equipe, isto é, a quantidade de gols que cada uma marcou. Se a equipe A marcou mais gols que B, então A venceu o jogo.

A condição presente no WHERE contempla zebras tanto de mandantes quanto de visitantes.

Adicionamos ainda uma coluna de total de zebras sobre total de jogos, para poder avaliar o % de zebras em um determinado campeonato.

O resultado da query está a seguir, ordenado do maior percentual de zebras para o menor:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| rank | league\_id | league | zebras | total\_jogos | pct\_zebras |
| 1 | 5522 | CECAFA Cup | 3 | 10 | 0,3 |
| 2 | 7621 | CONCACAF Nations League Qualifying | 4 | 18 | 0,222 |
| 3 | 5521 | Gulf Cup of Nations | 3 | 23 | 0,13 |
| 4 | 1908 | FIFA World Cup | 8 | 64 | 0,125 |
| 5 | 4802 | African Nations Championship | 7 | 59 | 0,119 |
| 6 | 1912 | FIFA World Cup OFC Qualifying | 1 | 10 | 0,1 |
| 7 | 1933 | AFC Asian Cup | 5 | 51 | 0,098 |
| 8 | 1820 | UEFA Europa League | 101 | 1093 | 0,092 |
| 9 | 7941 | CONCACAF Nations League | 18 | 207 | 0,087 |
| 10 | 1947 | Japanese J League | 135 | 1596 | 0,085 |
| 11 | 1874 | Swedish Allsvenskan | 131 | 1538 | 0,085 |
| 12 | 1832 | Belgian Jupiler League | 126 | 1492 | 0,084 |
| 13 | 1827 | Austrian T-Mobile Bundesliga | 94 | 1153 | 0,082 |
| 14 | 1837 | Danish SAS-Ligaen | 85 | 1064 | 0,08 |
| 15 | 2411 | Barclays Premier League | 205 | 2660 | 0,077 |
| 16 | 1859 | Norwegian Tippeligaen | 116 | 1523 | 0,076 |
| 17 | 2414 | English League Two | 200 | 2673 | 0,075 |
| 18 | 1882 | Turkish Turkcell Super Lig | 155 | 2060 | 0,075 |
| 19 | 2080 | AFC Asian Cup Qualifying | 3 | 40 | 0,075 |
| 20 | 1866 | Russian Premier Liga | 104 | 1435 | 0,072 |
| 21 | 1818 | UEFA Champions League | 63 | 869 | 0,072 |
| 22 | 1951 | Major League Soccer | 195 | 2752 | 0,071 |
| 23 | 10281 | UEFA Europa Conference League | 20 | 282 | 0,071 |
| 24 | 1879 | Swiss Raiffeisen Super League | 77 | 1080 | 0,071 |
| 25 | 1849 | Dutch Eredivisie | 122 | 1762 | 0,069 |
| 26 | 1845 | German Bundesliga | 148 | 2142 | 0,069 |
| 27 | 4801 | African Nations Championship Qualifying | 7 | 102 | 0,069 |
| 28 | 1854 | Italy Serie A | 184 | 2660 | 0,069 |
| 29 | 2160 | United Soccer League | 181 | 2640 | 0,069 |
| 30 | 1869 | Spanish Primera Division | 178 | 2660 | 0,067 |
| 31 | 4582 | National Women's Soccer League | 45 | 670 | 0,067 |
| 32 | 1921 | African Cup of Nations | 7 | 104 | 0,067 |
| 33 | 2105 | Brasileiro Serie A | 157 | 2380 | 0,066 |
| 34 | 1929 | International Match | 73 | 1109 | 0,066 |
| 35 | 1843 | French Ligue 1 | 165 | 2559 | 0,064 |
| 36 | 5523 | COSAFA Cup | 3 | 47 | 0,064 |
| 37 | 2413 | English League One | 165 | 2629 | 0,063 |
| 38 | 1911 | FIFA World Cup AFC Qualifying | 14 | 230 | 0,061 |
| 39 | 1884 | Greek Super League | 59 | 960 | 0,061 |
| 40 | 1844 | French Ligue 2 | 130 | 2179 | 0,06 |
| 41 | 1940 | European Championships | 3 | 51 | 0,059 |
| 42 | 1913 | FIFA World Cup CONMEBOL Qualifying | 5 | 89 | 0,056 |
| 43 | 7141 | UEFA Nations League | 18 | 330 | 0,055 |
| 44 | 1948 | Australian A-League | 42 | 775 | 0,054 |
| 45 | 1975 | Mexican Primera Division Torneo Clausura | 58 | 1101 | 0,053 |
| 46 | 2412 | English League Championship | 176 | 3342 | 0,053 |
| 47 | 1846 | German 2. Bundesliga | 97 | 1836 | 0,053 |
| 48 | 2081 | African Cup of Nations Qualifying | 14 | 273 | 0,051 |
| 49 | 7921 | FA Women's Super League | 34 | 688 | 0,049 |
| 50 | 1979 | Chinese Super League | 18 | 369 | 0,049 |
| 51 | 1856 | Italy Serie B | 117 | 2382 | 0,049 |
| 52 | 1931 | CONCACAF Gold Cup | 3 | 62 | 0,048 |
| 53 | 2417 | Scottish Premiership | 63 | 1313 | 0,048 |
| 54 | 1952 | Mexican Primera Division Torneo Apertura | 50 | 1032 | 0,048 |
| 55 | 1983 | South African ABSA Premier League | 40 | 876 | 0,046 |
| 56 | 5641 | Argentina Primera Division | 86 | 1959 | 0,044 |
| 57 | 1871 | Spanish Segunda Division | 118 | 2806 | 0,042 |
| 58 | 1864 | Portuguese Liga | 73 | 1836 | 0,04 |
| 59 | 1910 | FIFA World Cup CAF Qualifying | 6 | 158 | 0,038 |
| 60 | 2721 | Copa America | 2 | 54 | 0,037 |
| 61 | 1909 | FIFA World Cup UEFA Qualifying | 9 | 258 | 0,035 |
| 62 | 1919 | FIFA World Cup CONCACAF Qualifying | 4 | 118 | 0,034 |
| 63 | 1941 | European Championship Qualifying | 9 | 308 | 0,029 |
| 64 | 5501 | ASEAN Football Championship | 1 | 52 | 0,019 |

*% de zebras por campeonato, do maior para o menor*

Vemos através destes resultados que o torneio que registrou mais zebras foi a *CECAFA CUP* isto é, com 3 zebras em 10 jogos. Porém, por se tratar de um torneio de baixa amostragem, este resultado pode ser visto como pouco conclusivo.

De fato, analisando os demais resultados da query como um todo, vemos que os primeiros resultados são majoritariamente de torneios entre seleções, que tem naturalmente uma quantidade menor de jogos na base do que as ligas nacionais.

Portanto, para termos um resultado mais coerente, vamos rodar a query novamente, desta vez excluindo os torneios de seleções. Para isso, vamos trazer também na query a confederação de cada torneio.

A nova query fica assim:

SELECT A.league\_id,B.league,B.confed, COUNT(\*) as zebras, C.total\_jogos, ROUND(COUNT(\*)/C.total\_jogos,3) as pct\_zebras

FROM `mvp-puc.Teste538.Matches` A

  LEFT JOIN (SELECT league\_id, league, confed

            FROM `mvp-puc.Teste538.Leagues`) B

  ON A.league\_id = B.league\_id

    LEFT JOIN (SELECT league\_id, COUNT(\*) as total\_jogos

              FROM `mvp-puc.Teste538.Matches`

              GROUP BY league\_id) C

    ON A.league\_id = C.league\_id

WHERE (confed NOT LIKE ("INTERNATIONAL") AND score1>score2 AND prob1<=0.25 AND prob2>=0.50) OR (confed NOT LIKE ("INTERNATIONAL") AND score2>score1 AND prob1>=0.50 AND prob2<=0.25)

GROUP BY A.league\_id, B.league, B.confed, C.total\_jogos

ORDER BY pct\_zebras DESC

E seus resultados estão aqui:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| rank | league\_id | league | confed | zebras | total\_jogos | pct\_zebras |
| 1 | 1947 | Japanese J League | AFC | 135 | 1596 | 0,085 |
| 2 | 1874 | Swedish Allsvenskan | UEFA | 131 | 1538 | 0,085 |
| 3 | 1832 | Belgian Jupiler League | UEFA | 126 | 1492 | 0,084 |
| 4 | 1827 | Austrian T-Mobile Bundesliga | UEFA | 94 | 1153 | 0,082 |
| 5 | 1837 | Danish SAS-Ligaen | UEFA | 85 | 1064 | 0,08 |
| 6 | 2411 | Barclays Premier League | UEFA | 205 | 2660 | 0,077 |
| 7 | 1859 | Norwegian Tippeligaen | UEFA | 116 | 1523 | 0,076 |
| 8 | 2414 | English League Two | UEFA | 200 | 2673 | 0,075 |
| 9 | 1882 | Turkish Turkcell Super Lig | UEFA | 155 | 2060 | 0,075 |
| 10 | 1866 | Russian Premier Liga | UEFA | 104 | 1435 | 0,072 |
| 11 | 1879 | Swiss Raiffeisen Super League | UEFA | 77 | 1080 | 0,071 |
| 12 | 1951 | Major League Soccer | CONCACAF | 195 | 2752 | 0,071 |
| 13 | 2160 | United Soccer League | CONCACAF | 181 | 2640 | 0,069 |
| 14 | 1849 | Dutch Eredivisie | UEFA | 122 | 1762 | 0,069 |
| 15 | 1854 | Italy Serie A | UEFA | 184 | 2660 | 0,069 |
| 16 | 1845 | German Bundesliga | UEFA | 148 | 2142 | 0,069 |
| 17 | 4582 | National Women's Soccer League | CONCACAF | 45 | 670 | 0,067 |
| 18 | 1869 | Spanish Primera Division | UEFA | 178 | 2660 | 0,067 |
| 19 | 2105 | Brasileiro Serie A | CONMEBOL | 157 | 2380 | 0,066 |
| 20 | 1843 | French Ligue 1 | UEFA | 165 | 2559 | 0,064 |
| 21 | 2413 | English League One | UEFA | 165 | 2629 | 0,063 |
| 22 | 1884 | Greek Super League | UEFA | 59 | 960 | 0,061 |
| 23 | 1844 | French Ligue 2 | UEFA | 130 | 2179 | 0,06 |
| 24 | 1948 | Australian A-League | AFC | 42 | 775 | 0,054 |
| 25 | 1846 | German 2. Bundesliga | UEFA | 97 | 1836 | 0,053 |
| 26 | 2412 | English League Championship | UEFA | 176 | 3342 | 0,053 |
| 27 | 1975 | Mexican Primera Division Torneo Clausura | CONCACAF | 58 | 1101 | 0,053 |
| 28 | 1856 | Italy Serie B | UEFA | 117 | 2382 | 0,049 |
| 29 | 1979 | Chinese Super League | AFC | 18 | 369 | 0,049 |
| 30 | 7921 | FA Women's Super League | UEFA | 34 | 688 | 0,049 |
| 31 | 2417 | Scottish Premiership | UEFA | 63 | 1313 | 0,048 |
| 32 | 1952 | Mexican Primera Division Torneo Apertura | CONCACAF | 50 | 1032 | 0,048 |
| 33 | 1983 | South African ABSA Premier League | CAF | 40 | 876 | 0,046 |
| 34 | 5641 | Argentina Primera Division | CONMEBOL | 86 | 1959 | 0,044 |
| 35 | 1871 | Spanish Segunda Division | UEFA | 118 | 2806 | 0,042 |
| 36 | 1864 | Portuguese Liga | UEFA | 73 | 1836 | 0,04 |

*% de zebras por campeonato, do maior para o menor (apenas campeonatos nacionais)*

Agora temos resultados um pouco mais interessantes. Dentro destas condições, vemos que a liga com mais zebras é a ***Japanese J League***, a primeira divisão do futebol japonês, com 135 zebras em 1596 jogos disputados no período, um total de 8,5% de zebras.

Praticamente empatado com a ***J League*** está a ***Swedish Allsveskan***, a primeira divisão da Suécia, com uma porcentagem de zebras de também 8,5%.

No outro extremo dos resultados temos *a* ***Portuguese Liga***, isto é, o Campeonato Português, como o torneio com menos zebras, percentualmente falando. Foram apenas 73 zebras em 1836 jogos avaliados, o que significa que em apenas 4% dos jogos do Campeonato Português houve alguma zebra.

Desta forma, respondemos a primeira de nossas perguntas: **O campeonato japonês – a J League – pode ser considerado o campeonato mais disputado do mundo**: ou pelo menos, é onde as zebras mais acontecem, e os azarões mais conseguem surpreender os favoritos.

#### O Campeonato Brasileiro é o mais disputado do mundo?

Para responder esta pergunta, podemos olhar novamente para a tabela retornada na query anterior.

Nela vemos que o campeonato brasileiro – “***Brasileiro Serie A”*** – ocupa a 19ª posição no ranking de ligas com mais zebras, bem no meio da classificação.

No período analisado, contabilizou-se 157 zebras em 2380 jogos no Brasileirão, um total de 6,6% de zebras.

Naturalmente que torneios de futebol envolvem muitas outras características para podermos estimar uma competitividade como um todo, mas sob o ponto de vista deste trabalho apenas, pode-se dizer que o Brasileirão não é o campeonato mais disputado do mundo.

Afinal de contas, os azarões por aqui não vencem os favoritos na mesma proporção de outros campeonatos.

Como uma observação complementar, vale a pena analisarmos a posição da ***Premier League***, o campeonato inglês, considerado por muitos especialistas o mais disputado do mundo.

De fato, a ***Premier League*** figura nas primeiras posições deste ranking, estando em 6º lugar com 7,7% de zebras em seus jogos, validando de certa forma o posicionamento dos que a consideram como uma liga competitiva.

#### As zebras estão aumentando ou diminuindo?

Por fim, vamos fazer uma análise temporal sobre a frequência das zebras considerando este conjunto de dados.

Vamos estruturar uma query que contabilize o total de zebras por ano, o percentual por ano e avaliar se há alguma tendência de queda ou aumento.

SELECT A.season, COUNT(\*) as zebras, C.total\_jogos, ROUND(COUNT(\*)/C.total\_jogos,3) as pct\_zebras

FROM `mvp-puc.Teste538.Matches` A

  LEFT JOIN (SELECT league\_id, league, confed

            FROM `mvp-puc.Teste538.Leagues`) B

  ON A.league\_id = B.league\_id

    LEFT JOIN (SELECT season, COUNT(\*) as total\_jogos

              FROM `mvp-puc.Teste538.Matches`

              GROUP BY season) C

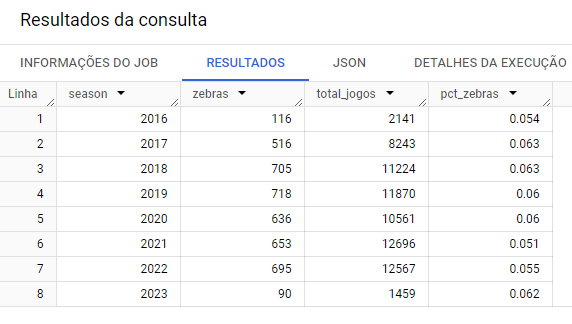
    ON A.season = C.season

WHERE (confed NOT LIKE ("INTERNATIONAL") AND score1>score2 AND prob1<=0.25 AND prob2>=0.50) OR (confed NOT LIKE ("INTERNATIONAL") AND score2>score1 AND prob1>=0.50 AND prob2<=0.25)

GROUP BY A.season, C.total\_jogos

ORDER BY season ASC

Nesta query mantemos a condição de avaliar apenas jogos de ligas, excluindo partidas entre seleções. O agrupamento agora é feito por temporadas. O resultado é visto na sequência:



*Evolução do % de zebras ao longo dos anos*

Ou, de forma gráfica:



*Evolução do % de zebras ao longo dos anos – visão gráfica*

Por estes dados, podemos concluir que não há exatamente uma tendência definida para a frequência de zebras no futebol nos últimos anos. De 2017 a 2020 o total de zebras por temporadas se permaneceu estável, sem grandes variações.

Em 2021 houve uma queda maior no registro de zebras, porém nas temporadas seguintes (2022 e 2023) já se observa um retorno aos valores observados entre 2017 e 2020.

A mediana destes dados é de 6%, enquanto que a média é de 5,85%.

Ao que parece, a frequência de zebras gira em torno de 6% independente do ano e campeonato, com valores discrepantes a este valor sendo mais situações de *outliers*.

# Comentários Finais

### Avaliação Geral

O projeto se mostrou bem sucedido, conseguindo atingir os objetivos propostos.

Vimos que existem alguns campeonatos mais propensos a ocorrências de zebras, como por exemplo o campeonato japonês (J League) e o campeonato sueco (Allsvenskan).

Dentre as principais ligas do mundo, o destaque foi o campeonato inglês, que figurou em 6º no ranking geral de zebras por ligas. Podemos interpretar este resultado como um indicativo de que, de fato, a liga inglesa é uma das mais competitivas do mundo.   
Isto porque uma maior ocorrência de zebras indica o quão competitivas são as equipes consideradas menos favoritas neste campeonato, sendo capazes de derrotar os oponentes mais fortes mais vezes do que em outras ligas.

Vimos também que, ao contrário do que indica o senso comum, o campeonato brasileiro não é considerado o mais competitivo do mundo (pelo menos não por esta ótica de frequência de zebras).   
Dentro da metodologia adotada, o campeonato brasileiro registrou um percentual de zebras próximo da média do conjunto analisado. Isto talvez seja um reflexo da consolidação de forças de algumas das principais equipes do país nos últimos anos, como Flamengo e Palmeiras, reduzindo assim a quantidade de zebras nas edições do campeonato.

No outro extremo dos resultados, vimos que a liga portuguesa é a que apresentou o menor percentual de zebras dentre as ligas avaliadas. Em outras palavras, podemos interpretar que a liga portuguesa é a mais previsível dentre as analisadas, ou ainda, a menos competitiva.  
Isto faz sentido quando olhamos para o histórico da liga portuguesa: 84 das 86 edições já disputadas foram vencidas ou por Benfica, ou por Sporting, ou por Porto.   
Ou seja, é um campeonato no qual os menos favoritos de fato tem muita dificuldade de superar os mais favoritos.

Vale também ressaltar que entre jogos de seleções nacionais se observou um percentual maior de ocorrência de zebras. Isto pode ser explicado pelo fato de jogos entre seleções serem menos frequentes. Como as seleções jogam menos ao longo dos anos, fica mais difícil determinar a real probabilidade de uma seleção vencer a outra.   
Além da menor quantidade de jogos para uma análise amostral, deve-se considerar também a hipótese de seleções sofrerem mais com a falta de entrosamento do que clubes, e assim estarem mais propensas a não jogarem dentro de todo o seu potencial previsto, o que pode gerar resultados fora do esperado.   
Como exemplos, podemos citar a campanha da Alemanha na Copa do Mundo 2018 (derrotas inesperadas) e a campanha de Marrocos na Copa do Mundo 2022 (vitórias inesperadas).

De toda forma, devido ao volume menor de jogos entre seleções, vale a pena uma análise mais aprofundada, com uma amostragem maior de jogos (incluindo mais anos de análise, no caso).

Por fim, foi feita uma análise temporal do percentual de zebras de todo o conjunto de jogos ao longo dos 8 anos analisados (2016-2023). Nesta análise, vimos que o percentual se mantém estável ao longo dos anos, em torno de 6%, sofrendo poucas oscilações no geral.   
Assim, não foi possível encontrar uma tendência de crescimento ou diminuição de zebras ao longo deste período.

### Desafios Técnicos

Foram poucos os casos de dificuldades técnicas para a execução do projeto.

Os maiores desafios foram lidar com registros de jogos com placares nulos e lidar com a falta de informação dos países a que cada campeonato pertence, tendo que construir esta classificação via Transformação.

Quanto aos nulos, vimos que cerca de 5% de todos os registros não possuíam placares válidos, e por isso foram descartados da análise. Embora um valor considerável, os outros 95% foram suficientes para fazer a análise de zebras de cada campeonato.

E quanto aos países e confederações a que cada campeonato pertence, foi possível categorizar as ligas nacionais, porém os campeonatos continentais, assim como os torneios entre seleções ficaram com esta informação incompleta.

Isto não impediu de avaliar a quantidade de zebras por campeonato, porém caso fosse de interesse avaliar o percentual de zebras no nível de confederações, por exemplo, o resultado não seria tão claro. Este, inclusive, é um dos pontos de melhoria para uma próxima versão deste trabalho.

Um último ponto a se considerar é que o ETL da tabela *matches* gerou um *schema* no qual o campo date está como **string**. Naturalmente isto é inadequado, pois *date* deveria ser um campo de **timestamp**, tal qual está na tabela *calendar*.   
Desta forma, para fazer joins entre estas tabelas se faz necessário um passo a mais de transformação de tipos de dados.   
Assim, este então é um outro ponto a ser revisitado neste projeto. O ideal é que o ETL já salve a tabela *matches* no BigQuery com o *schema* correto.

### Próximos Passos

Além de revisitar os pontos técnicos listados acima, listamos aqui passos adicionais para este projeto que poderiam levar a análises mais aprofundadas, capazes de gerar ainda mais insights:

1. **Analisar a quantidade de zebras por confederações**

Uma vez que a dimensão *leagues* estiver com os países e confederações de cada liga 100% adequados, vale analisar o percentual de zebras no nível de confederações.

Assim, poderemos avaliar se há mais zebras em um determinado continente ou se todos eles apresentam valores similares.

1. **Aumentar o grão da análise temporal para meses**

Na análise temporal feita, vimos que o percentual de zebras se mantém estável ao longo do ano. Porém, vale a pena refazer a análise considerando os meses do ano.   
Isto porque, ao longo de uma temporada de futebol, é possível identificar diferentes momentos.

Por exemplo, será que as zebras ocorrem mais nos primeiros meses da temporada, quando as equipes ainda estão se entrosando? Ou será que elas ocorrem mais nos últimos meses, com o maior cansaço das equipes?

1. **Avaliar as zebras a partir de outras métricas**

Neste projeto consideramos zebra como sendo a vitória de um time menos favorito quando este time tinha 25% ou menos de probabilidade de vitória e seu adversário 50% ou mais.

Pode ser interessante aprofundar esta análise para contemplar casos de empates. Vale a pena também comparar novos resultados variando os limites definidos para a zebra (por exemplo, considerar casos onde o time menos favorito tinha só 20% de chances ou menos e ainda assim venceu).  
A criação e análise destes novos cenários ajudariam a trazer mais insights sobre a competitividade de cada liga.

1. **Avaliar as zebras por equipes**

O escopo deste projeto analisou a quantidade de zebras por ligas, mas será que existe alguma equipe dentre as analisadas mais propensa a cometer (ou sofrer) zebras?

Enfim, estas são algumas das sugestões imediatas para uma sequência do trabalho. Seria possível ainda buscar expandir a análise para anos anteriores, porém para isto seria necessário encontrar outra fonte de dados, pois as usadas neste projeto contemplam dados apenas de 2016 a 2023.

Outra linha de análise que pode ser feita a partir destes dados seria avaliar as zebras considerando a posição de cada equipe no campeonato quando elas se enfrentam. Isto é, avaliar a zebra como sendo uma vitória de uma equipe da parte de baixo da tabela contra outra de uma da parte superior da tabela.

É uma linha de raciocínio interessante, mas que para isso seria necessário construir uma nova base de dados que seja capaz de somar os pontos ganhos de cada equipe jogo a jogo e ainda registrar a posição de cada rodada a rodada. Dada a complexidade, fica como ideia para um próximo projeto.