



## Especialização em Ciência de Dados e Analytics

**MVP**

Sprint: Engenharia de Dados

**Erika Kacelnik**

Professores: Vitor Almeida e Silvio Alonso

1 de outubro de 2023

---

### Objetivo

De 22 de agosto a 10 de setembro de 2023, foi disputada em Nova York a 143<sup>a</sup> edição do US Open. Em conjunto com o Australian Open, Roland Garros e Wimbledon, forma o grupo de torneios chamados de Grand Slams, os quatro principais eventos do calendário de tênis mundial. Esses eventos são os que mais somam pontos à classificação do ranking de jogadores e mais atraem espectadores.

Este ano uma disparidade curiosa chamou a atenção dos fãs de tênis. Novak Djokovic, tenista sérvio de 36 anos, venceu o torneio masculino, ganhando seu 24º Grand Slam da carreira e ampliando seu próprio recorde de maior vencedor de Grand Slams da história

entre os homens. Já o título feminino foi conquistado por Coco Gauff, de apenas 19 anos, inaugurando o primeiro Grand Slam de sua carreira.

Além disso, a jornada dos vencedores da edição de 2022 também foi muito diferente. O tenista Carlos Alcaraz alcançou a semifinal do torneio, enquanto Iga Świątek foi eliminada ainda na rodada dos 16 melhores.

A partir dessas duas constatações e observações empíricas dos resultados de tênis ao longo dos últimos anos, este trabalho pretende investigar se esse resultado foi um acaso ou se representa uma tendência do esporte. Portanto, busca-se entender qual é a disparidade da idade entre vencedores masculinos e femininos de Grand Slams. Além disso, entender se, a partir de uma primeira vitória, tenistas homens têm mais chances de acumularem mais títulos do que suas colegas mulheres.

Perguntas elencadas:

- A média de idade das vencedoras de Grand Slams é significativamente mais baixa que a dos vencedores homens?
- É mais provável que um vencedor homem de Grand Slam vença mais torneios do que uma vencedora mulher?

## Busca pelos dados

Foi feita uma busca por bases de dados de tênis em sites como o Kaggle. Apesar de existirem muitas entradas, nenhuma incluía idade dos jogadores a cada vitória, o que era um dado primordial para o projeto.

Em uma busca em blogs e site específicos de tênis, foram encontradas as bases de dados de Jeff Sackmann, referências em estudos do esporte. Elas estão separadas por organização (ATP - circuito masculino e WTA - circuito feminino) e disponíveis em formato de CSV no Github:

[https://github.com/JeffSackmann/tennis\\_wta](https://github.com/JeffSackmann/tennis_wta)  
[https://github.com/JeffSackmann/tennis\\_atp](https://github.com/JeffSackmann/tennis_atp)

github.com/JeffSackmann/tennis\_wta/blob/master/wta\_matches\_1984.csv

**Code** Issues 46 Pull requests Actions Projects Wiki Security Insights

**Files**

master wta\_matches\_1974.csv wta\_matches\_1975.csv wta\_matches\_1976.csv wta\_matches\_1977.csv wta\_matches\_1978.csv wta\_matches\_1979.csv wta\_matches\_1980.csv wta\_matches\_1981.csv wta\_matches\_1982.csv wta\_matches\_1983.csv wta\_matches\_1984.csv wta\_matches\_1985.csv wta\_matches\_1986.csv wta\_matches\_1987.csv wta\_matches\_1988.csv wta\_matches\_1989.csv wta\_matches\_1990.csv

**Preview** Code Blame 2356 lines (2356 loc) · 400 KB

JeffSackmann add match time for various epic matches 9256b55 · last year History

Search this file

1	tourney_id	tourney_name	surface	draw_size	tourney_level	tourney_date	match_num	winner_id	winner_seed	winner_en
2	1984-D001	Fed Cup WG R1: ISR vs PER	Clay	4	D	19840715	1	212336		
3	1984-D001	Fed Cup WG R1: ISR vs PER	Clay	4	D	19840715	2	212337		
4	1984-D002	Fed Cup WG QF: USA vs ITA	Clay	4	D	19840715	1	200585		
5	1984-D002	Fed Cup WG QF: USA vs ITA	Clay	4	D	19840715	2	200422		
6	1984-D003	Fed Cup WG R2: AUT vs ITA	Clay	4	D	19840715	1	200725		
7	1984-D003	Fed Cup WG R2: AUT vs ITA	Clay	4	D	19840715	2	200422		
8	1984-D004	Fed Cup WG R2: YUG vs ISR	Clay	4	D	19840715	1	200628		
9	1984-D004	Fed Cup WG R2: YUG vs ISR	Clay	4	D	19840715	2	200371		
10	1984-D005	Fed Cup WG ConR: CHI vs CAN	Clay	4	D	19840715	1	201174		
11	1984-D005	Fed Cup WG ConR: CHI vs CAN	Clay	4	D	19840715	2	201172		
12	1984-D006	Fed Cup WG ConR: BRA vs NED	Clay	4	D	19840715	1	201975		
13	1984-D006	Fed Cup WG ConR: BRA vs NED	Clay	4	D	19840715	2	200962		

github.com/JeffSackmann/tennis\_atp/blob/master/atp\_matches\_2011.csv

**Code** Issues 30 Pull requests 1 Actions Projects Wiki Security Insights

**Files**

master atp\_matches\_2002.csv atp\_matches\_2003.csv atp\_matches\_2004.csv atp\_matches\_2005.csv atp\_matches\_2006.csv atp\_matches\_2007.csv atp\_matches\_2008.csv atp\_matches\_2009.csv atp\_matches\_2010.csv atp\_matches\_2011.csv atp\_matches\_2012.csv atp\_matches\_2013.csv atp\_matches\_2014.csv atp\_matches\_2015.csv atp\_matches\_2016.csv atp\_matches\_2017.csv atp\_matches\_2018.csv

We can't make this file beautiful and searchable because it's too large.

JeffSackmann add heights for ~300 players a707e4f · last year History

Preview Code Blame 3016 lines (3016 loc) · 598 KB

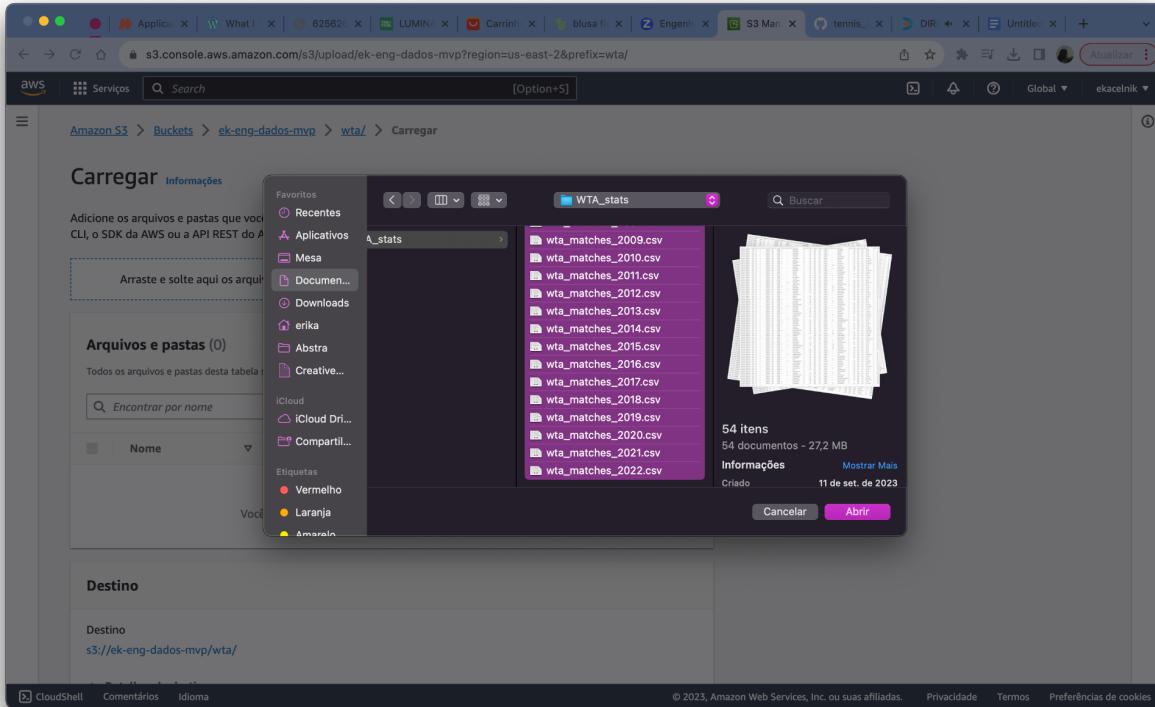
Raw Download Edit

```
1 tourney_id,tourney_name,surface,draw_size,tourney_level,tourney_date,match_num,winner_id,winner_seed,winner_entry,winner_name,winner_hand,
2 2011-339,Brisbane,Hard,32,A,20110102,1,104417,1,,Robin Soderling,R,193,SWE,26.3,-105992,0,Ryan Harrison,R,183,USA,18.6,6-2 6-4,3,R32,66,8,
3 2011-339,Brisbane,Hard,32,A,20110102,2,103582,,Michael Berrer,L,193,GER,30.5,104534,,Dudi Sela,R,175,ISR,25.7,1-6 7-6(3) 6-2,3,R32,152,6
4 2011-339,Brisbane,Hard,32,A,20110102,3,105051,,Matthew Ebden,R,188,AUS,23.1,105357,,WC,John Millman,R,183,AUS,21.5,4-6 6-2 6-4,3,R32,104
5 2011-339,Brisbane,Hard,32,A,20110102,4,104797,8,,Denis Istomin,R,188,UZB,24.3,105217,,,Thiemo De Bakker,R,193,NED,22.2,7-6(5) 6-4,3,R32,9
6 2011-339,Brisbane,Hard,32,A,20110102,5,103888,4,,Mardy Fish,R,188,USA,29.0,105173,,O,Adrian Mannarino,L,183,FRA,22.5,6-1 6-4,3,R32,79,11,(4)
7 2011-339,Brisbane,Hard,32,A,20110102,6,103285,,Radek Stepanek,R,185,CZE,32.1,104735,,,Tobias Kamke,R,180,GER,24.6,5-7 6-1 6-4,3,R32,124,(4)
8 2011-339,Brisbane,Hard,32,A,20110102,7,105575,,Ricardas Berankis,R,175,LTU,20.5,103096,,Arnaud Clement,R,173,FRA,33.0,6-4 6-3,3,R32,86
9 2011-339,Brisbane,Hard,32,A,20110102,8,104252,7,,Florian Mayer,R,190,GER,27.2,106071,,WC,Bernard Tomic,R,193,AUS,18.2,6-2 6-2,3,R32,55,12
10 2011-339,Brisbane,Hard,32,A,20110102,9,103852,6,,Feliciano Lopez,L,188,ESP,29.2,104332,,Philipp Petzschner,R,185,GER,26.7,6-4 7-6(11),3,f
11 2011-339,Brisbane,Hard,32,A,20110102,10,104731,,Kevin Anderson,R,203,RSA,24.6,103429,,LL,Peter Luczak,R,183,AUS,31.3,6-4 6-4,3,R32,72,18
12 2011-339,Brisbane,Hard,32,A,20110102,11,105053,,Santiago Giraldo,R,188,COL,23.1,104468,,Gilles Simon,R,183,FRA,26.0,6-2 6-3,3,R32,86,2,(4)
13 2011-339,Brisbane,Hard,32,A,20110102,12,103794,,Benjamin Becker,R,178,GER,29.5,104269,3,,Fernando Verdasco,L,188,ESP,27.1,6-1 6-7(2) 6-3
14 2011-339,Brisbane,Hard,32,A,20110102,13,104571,5,,Marcos Baghdatis,R,183,CYP,25.5,103722,,Florent Serra,R,180,FRA,29.8,6-3 7-5 6-4,3,R32
15 2011-339,Brisbane,Hard,32,A,20110102,14,103997,,Lukasz Kubot,R,190,POL,28.6,104978,,Daniel Brands,R,196,GER,23.4,6-2 6-2,3,R32,78,7,1,5
16 2011-339,Brisbane,Hard,32,A,20110102,15,105238,,Alexandr Dolgopolov,R,180,UKR,22.1,104214,,Igor Andreev,R,185,RUS,27.4,6-4 6-2,3,R32,67
17 2011-339,Brisbane,Hard,32,A,20110102,16,104053,2,,Andy Roddick,R,188,USA,28.3,104594,,WC,Marinko Matosevic,R,194,AUS,25.4,6-3 6-2,3,R32,72
18 2011-339,Brisbane,Hard,32,A,20110102,17,104417,1,,Robin Soderling,R,193,SWE,26.3,103582,,Michael Berrer,L,193,GER,30.5,6-3 7-6(7),3,f
19 2011-339,Brisbane,Hard,32,A,20110102,18,105051,,Matthew Ebden,R,188,AUS,23.1,104797,8,,Denis Istomin,R,188,UZB,24.3,6-4 6-4,3,R16,73,5,(4)
20 2011-339,Brisbane,Hard,32,A,20110102,19,103285,,Radek Stepanek,R,185,CZE,32.1,103888,4,,Mardy Fish,R,188,USA,29.0,6-3 6-1,3,R16,78,3,2,(4)
21 2011-339,Brisbane,Hard,32,A,20110102,20,104252,7,,Florian Mayer,R,190,GER,27.2,105575,,Ricardas Berankis,R,175,LTU,20.5,6-4 4-6 6-4,3,R16,56,
```

Foi feita uma conferência manual de cerca de 20 entradas antes da escolha pelo uso dos dados, e tomado como ano inicial da análise o ano de 1969, o início da chamada Open Era do tênis, cujos padrões de organização se mantém até os dias de hoje.

## Coleta

Os CSVs estão organizados por ano no repositório. Os dados foram baixados para uma máquina local e inseridos manualmente em um bucket do S3, na AWS.



Carregar Informações

Adicione os arquivos e pastas que você deseja carregar no S3. Para fazer upload de um arquivo maior que 160 GB, use a AWS CLI, o SDK da AWS ou a API REST do Amazon S3. [Saiba mais](#)

Arraste e solte aqui os arquivos e pastas para upload ou selecione Adicionar arquivos ou Adicionar pastas.

Arquivos e pastas (54 Total, 26.0 MB)					
	Nome	Pasta	Tipo	Tamanho	
<input type="checkbox"/>	wta_matches_1969....	-	text/csv	449.5 KB	
<input type="checkbox"/>	wta_matches_1970....	-	text/csv	455.1 KB	
<input type="checkbox"/>	wta_matches_1971....	-	text/csv	386.6 KB	
<input type="checkbox"/>	wta_matches_1972....	-	text/csv	450.2 KB	
<input type="checkbox"/>	wta_matches_1973....	-	text/csv	449.5 KB	
<input type="checkbox"/>	wta_matches_1974....	-	text/csv	425.3 KB	
<input type="checkbox"/>	wta_matches_1975....	-	text/csv	409.4 KB	
<input type="checkbox"/>	wta_matches_1976....	-	text/csv	344.7 KB	
<input type="checkbox"/>	wta_matches_1977....	-	text/csv	369.9 KB	
<input type="checkbox"/>	wta_matches_1978....	-	text/csv	433.4 KB	

Todos os arquivos e pastas desta tabela serão carregados.

Encontrar por nome < 1 2 3 4 5 6 >

CloudShell Comentários Idioma © 2023, Amazon Web Services, Inc. ou suas afiliadas. Privacidade Termos Preferências de cookies

Upload bem-sucedido  
Visualize os detalhes abaixo.

Arquivos e pastas | Configuração

Arquivos e pastas (54 Total, 26.0 MB)

Nome	Pasta	Tipo	Tamanho	Status	Erro
wta_matches_1969.csv	-	text/csv	449.5 KB	<input checked="" type="checkbox"/> Bem-sucedida	-
wta_matches_1970.csv	-	text/csv	455.1 KB	<input checked="" type="checkbox"/> Bem-sucedida	-
wta_matches_1971.csv	-	text/csv	386.6 KB	<input checked="" type="checkbox"/> Bem-sucedida	-
wta_matches_1972.csv	-	text/csv	450.2 KB	<input checked="" type="checkbox"/> Bem-sucedida	-
wta_matches_1973.csv	-	text/csv	449.5 KB	<input checked="" type="checkbox"/> Bem-sucedida	-
wta_matches_1974.csv	-	text/csv	425.3 KB	<input checked="" type="checkbox"/> Bem-sucedida	-
wta_matches_1975.csv	-	text/csv	409.4 KB	<input checked="" type="checkbox"/> Bem-sucedida	-
wta_matches_1976.csv	-	text/csv	344.7 KB	<input checked="" type="checkbox"/> Bem-sucedida	-
wta_matches_1977.csv	-	text/csv	369.9 KB	<input checked="" type="checkbox"/> Bem-sucedida	-
wta_matches_1978.csv	-	text/csv	433.4 KB	<input checked="" type="checkbox"/> Bem-sucedida	-

CloudShell Comentários Idioma © 2023, Amazon Web Services, Inc. ou suas afiliadas. Privacidade Termos Preferências de cookies

## **Modelagem**

Como mencionado anteriormente, os dados foram baixados do repositório aberto de dados de Jeff Sackmann sobre tênis, que estão licenciados sob a licença Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Os dados foram modelados em duas tabelas flat: uma para os dados do masculino (circuito ATP) e feminino (circuito WTA).

Cada tabela possui dados de todas as partidas realizadas nos circuitos desde 1969. Cada linha da tabela representa um jogo, com 49 colunas que nos informam dados distintos sobre a partida, como informações sobre a duração, o torneio, o jogador vencedor e o jogador perdedor.

As colunas mais utilizadas para a análise proposta serão:

- tourney\_id - string - identificador único da edição do torneio, composto pelo ano de realização e o código do torneio.
- tourney\_name - string - nome do torneio, restrito aos torneios reconhecidos no calendário oficial de cada circuito.
- tourney\_date - bigint - data de cada partida, sempre em concordância com o ano na coluna tourney\_id.
- winner\_id - bigint - código identificador único de cada jogador vencedor.
- winner\_name - string - nome único de cada jogador vencedor.
- winner\_age - double - idade de cada jogador vencedor, que entende-se ter um valor mínimo de 15 anos e máximo de 45 anos.

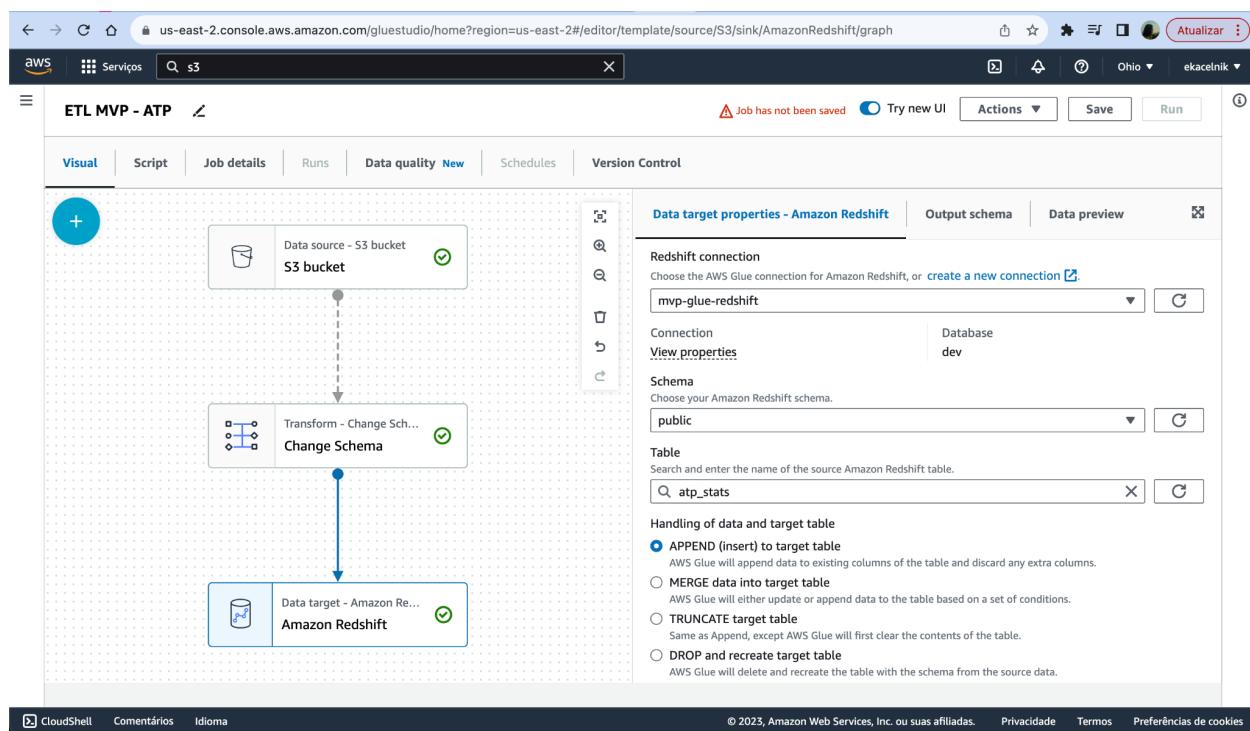
Segue abaixo o schema criado com o AWS Glue Data Catalog com os domínios de cada coluna das tabelas:

Schema (49)		
View and manage the table schema.		
#	Column name	Data type
1	tourney_id	string
2	tourney_name	string
3	surface	string
4	draw_size	bigint
5	tourney_level	string
6	tourney_date	bigint
7	match_num	bigint
8	winner_id	bigint
9	winner_seed	string
10	winner_entry	string
11	winner_name	string
12	winner_hand	string
13	winner_ht	bigint
14	winner_loc	string
15	winner_age	double
16	loser_id	bigint
17	loser_seed	string
18	loser_entry	string
19	loser_name	string
20	loser_hand	string
21	loser_ht	bigint
22	loser_loc	string
23	loser_age	double
24	score	string
25	best_of	bigint
26	round	string
27	minutes	bigint
28	w_ace	bigint
29	w_df	bigint
30	w_svpt	bigint
31	w_1stin	bigint
32	w_1stwon	bigint
33	w_2ndwon	bigint
34	w_svgtms	bigint
35	w_bpssaved	bigint
36	w_bpffaced	bigint
37	l_ace	bigint
38	l_df	bigint
39	l_svpt	bigint
40	l_1stin	bigint
41	l_1stwon	bigint
42	l_2ndwon	bigint
43	l_svgtms	bigint
44	l_bpssaved	bigint
45	l_bpffaced	bigint
46	winner_rank	bigint
47	winner_rank_points	bigint
48	loser_rank	bigint
49	loser_rank_points	bigint

## Carga

Seguindo a inserção dos dados no S3, a AWS continuou sendo usada para as próximas etapas do trabalho. O AWS Glue e seu editor visual AWS Glue Studio foram selecionados para o processo de ETL e a ferramenta de Data Warehouse AWS Redshift foi utilizada para carga dos dados tratados.

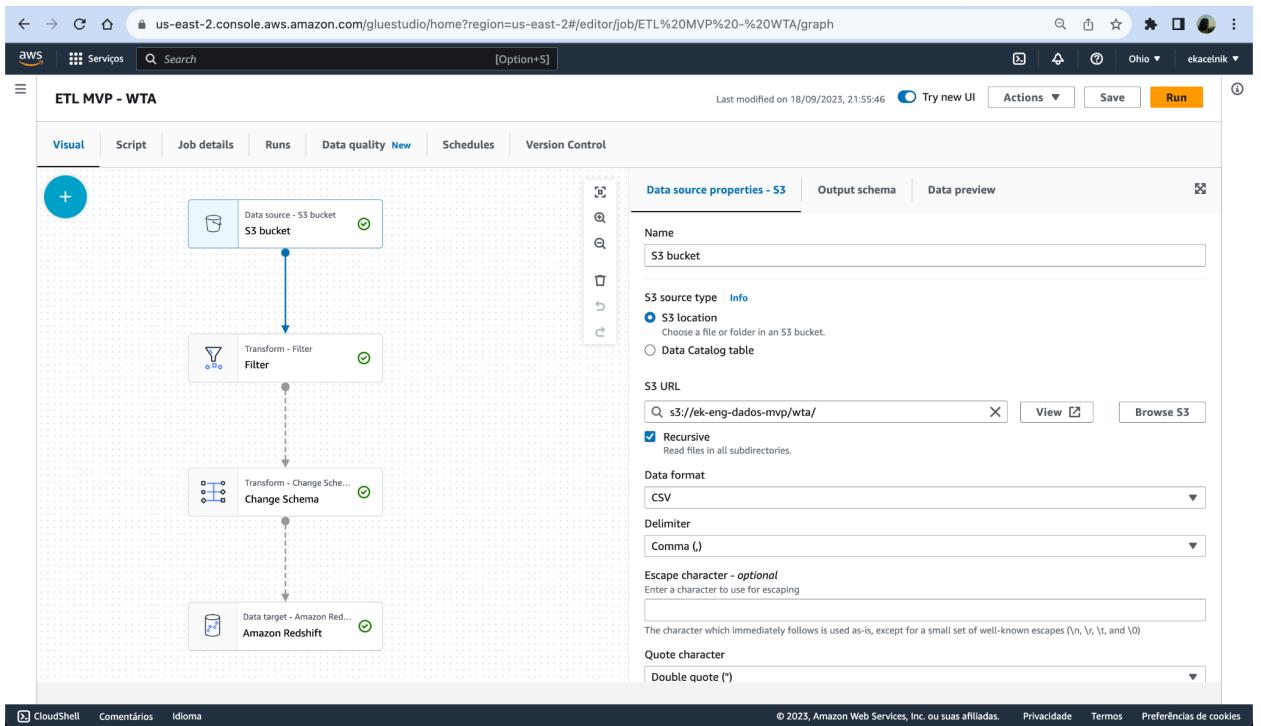
Primeiro foi feito um processo de ETL com apenas 3 etapas, com a intenção de uma filtragem de informações posterior ser feita em SQL dentro do Redshift.



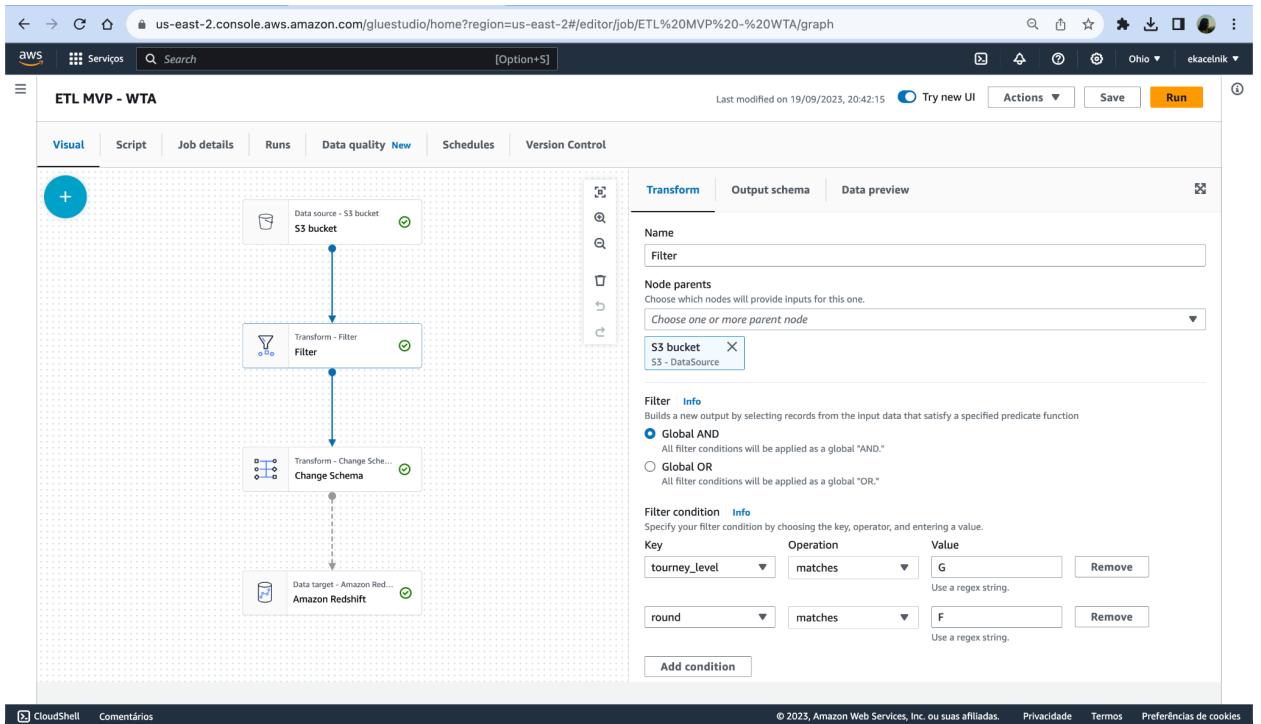
No entanto, como as bases de dados utilizadas eram muito extensas e só era necessário resultados de finais de campeonatos Grand Slam, seria necessário fazer a mesma filtragem em todas as queries. Por isso, foi refeito o ETL incluindo já o processo de filtragem.

Foram, então, criados dois Jobs idênticos: um para tratar dos dados da WTA e outro da ATP. Foram realizadas as mesmas seguintes etapas em ambos:

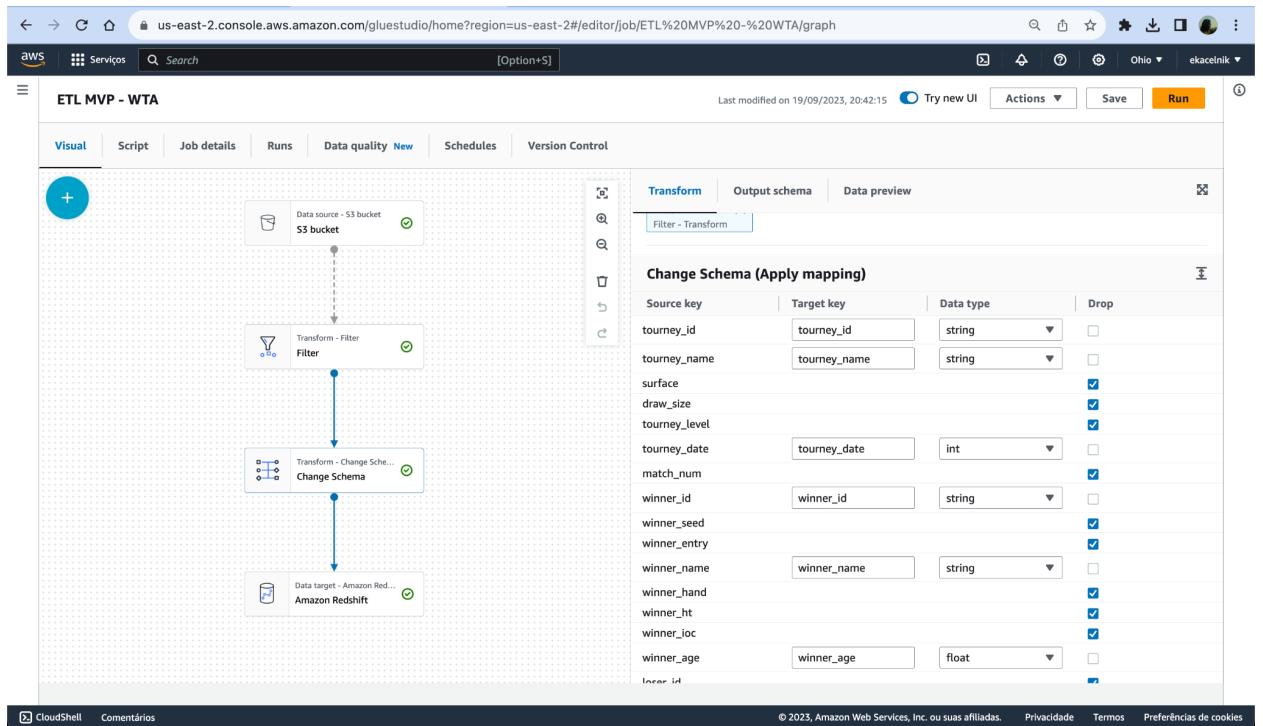
1. Configuração do nódulo "Data source - S3 bucket" para extrair, em formato de CSV, os dados das pastas "wta" e "atp" do bucket "ek-eng-dados-mvp" do S3.



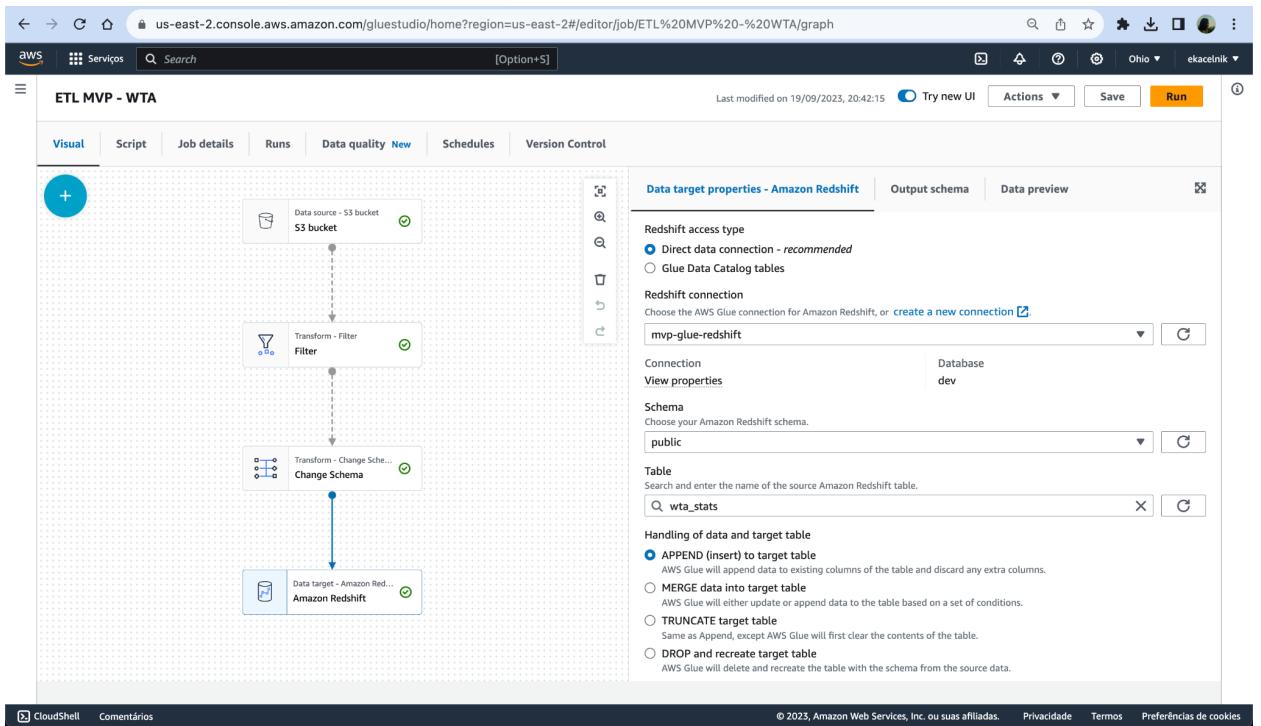
2. No intuito de reunir apenas os dados a respeito de campeões de Grand Slams, no nódulo "Transform - Filter", foi realizada uma filtragem apenas dos dados cuja categoria "tourney\_level" fosse "G" (abreviação de Grand Slam) e (AND) cuja categoria "round" fosse "F" (abreviação de final).



3. A próxima etapa também foi de transformação, desta vez do tipo "Transform - Change Schema". Foram mantidos apenas os campos cujas informações eram relevantes ao problema: tourney\_id, tourney\_name, tourney\_date, winner\_id, winner\_name e winner\_age. O campo "tourney\_date" foi convertido para int, que foi a forma mais fácil encontrada de possibilitar filtros de data, e "winner\_age" para float, para possibilitar cálculos.



4. No nódulo final de "Data target - Amazon Redshift", foi configurada a conexão com o Redshift e com o schema e a tabela já criados. Assim, foi feito o carregamento dos dados transformados para o banco.



## 5. Por fim, foi feito o registro de todas as execuções de ambos os Jobs.

The screenshot shows the "Runs" tab in the AWS Glue Studio interface. It displays a table of job runs for the "ETL MVP - WTA" job. The table has columns for "Run status", "Retries", "Start time", "End time", "Duration", "Capacity (DPU)", "Worker type", and "Glue version". There are five entries, all of which are marked as "Succeeded". The most recent run started at 09/13/2023 21:49:08 and ended at 09/13/2023 21:50:50. The "Table View" button is highlighted. Below the table, a specific run for 09/13/2023 21:49:08 is expanded, showing detailed information such as Job name (ETL MVP - WTA), Id (jr\_4566057c33e597ece5fe120a33598254395b45ef16db1f828b186e6333cd5474), Run status (Succeeded), Start-up time (15 seconds), and Security configuration (-). At the bottom, there are copyright and legal links.

## Análise

### Qualidade dos dados

O conjunto de dados utilizado foi extensamente tratado antes de ser disponibilizado e, consequentemente, selecionado para este trabalho. Cabe, no entanto, uma breve análise de valores para cada coluna, a fim de demonstrar a qualidade dos dados e entender se algum problema se apresentará para a solução do problema.

Cada tourney\_id deve aparecer apenas uma vez, dado que temos dados de um vencedor a cada ano. A query testada em cada tabela retornou 219 IDs distintas de torneio, que equivale ao número de linhas, provando-se correta.

A coluna tourney\_name deveria ser populada por apenas 4 resultados distintos, dado que foram filtrados apenas Grand Slams. A query, no entanto, retornou com 6 resultados. Uma breve investigação concluiu que isso se deve a uma extraordinária segunda edição anual do Aberto da Australia que ocorreu em 1977 e uma mudança de sintaxe do nome do Aberto dos Estados Unidos, de "US Open" para "Us Open" a partir de 2019. Dado que nenhuma dessas alterações prejudica a qualidade dos dados, não foram necessárias alterações.

Para o `winner_id` e `winner_name`, que representam o nome e ID únicos de cada vencedor, não se pode existir mais do que 219 resultados distintos, que seria o resultado caso cada vencedor tivesse ganhado uma única vez. Ambas as tabelas tiveram resultados compatíveis.

Para as colunas com domínio `integer`, `winner_age` e `tourney_date`, foi utilizado a análise de valores mínimos e máximos para entender se o conjunto tinha algum outlier. Os resultados, demonstrados na tabela abaixo, também se alinharam com as expectativas: que as datas estejam entre 1969 e 2023 e as idades entre 15 e 45 anos.

Tabela	Coluna	Valor mínimo	Valor máximo
WTA	<code>winner_age</code>	16.2	35.3
WTA	<code>tourney_date</code>	19690120	20230828
ATP	<code>winner_age</code>	17.2	37.1
ATP	<code>tourney_date</code>	19690120	20230828

## Solução do problema

Relembrando as duas perguntas elencadas no início do trabalho:

- A média de idade das vencedoras de Grand Slams é significativamente mais baixa que a dos vencedores homens?
- É mais provável que um vencedor homem de Grand Slam vença mais torneios do que uma vencedora mulher?

Iniciou-se a solução calculando as médias de idade dos vencedores da tabela WTA e ATP.

```
1 select avg(winner_age_double) from public.atp_stats
```

Result 1 (1)

avg
25.904566210045658

```
1 select avg(winner_age_float) from public.wta_stats
```

Result 1 (1)

avg
24.49908681547261

Os resultados foram muito semelhantes: a média de idade de vencedores homens foi 25.9, enquanto a de mulheres foi 24.5.

Entende-se, portanto, que observando a história do tênis, a expectativa de idade de um vencedor de Grand Slam é bem semelhante entre gêneros.

Para responder a segunda pergunta em uma análise mais simplificada, buscou-se entender quantos vencedores de Grand Slams de cada gênero existem, a fim de calcular uma média de títulos por pessoa por gênero.

```
▶ Run ⚡ Limit 100 Explain Isolated session ⓘ Serverless: de... dev
1 select count(distinct winner_name) from public.wta_stats
```

Result 1 (1)

	count
□	59

```
▶ Run ⚡ Limit 100 Explain Isolated session ⓘ Serverless: de... dev
1 select count(distinct winner_name) from public.atp_stats
```

Result 1 (1)

	count
□	57

As queries informaram que existem 57 vencedores distintos na categoria masculina, e 59 na feminina.

Analisando a média em relação a 219 títulos:

$$219 \div 57 = 3.8$$

$$219 \div 59 = 3.7$$

Assim como para a primeira pergunta, um resultado semelhante entre gêneros foi encontrado: a média de Grand Slams por vencedor masculino é 3.8 e por vencedora feminina é 3.7.

O resultado da análise, portanto, é de que, observando o conjunto histórico de campeões desde 1969, o fator gênero praticamente não influencia na idade do vencedor, nem em sua quantidade de títulos.

É evidente, porém, que esses valores são aproximações que não levam em conta a existência de certos grandes vencedores históricos que acumulam muitos mais títulos do que a média.

Refletindo sobre isso e sobre a motivação para o trabalho - a vitória recente do 24º Grand Slam de um jogador de 36 anos - foi decidido investigar um pouco mais a fundo se a existência de grandes vencedores na última década tiveram algum impacto nas médias calculadas anteriormente - e na percepção pública de uma disparidade.

Portanto, foram calculadas as médias de idade de vencedores dos últimos 10 anos:

	ATP	WTA
2014	27.3	29
2015	28.5	33.5
2016	29.5	28.5
2017	33.3	25.8
2018	32.7	26.3
2019	32.5	22.8
2020	31.3	22.7
2021	31.8	23.1
2022	31.4	22.7
2023	32	22.5

Os valores se aproximam nos primeiros anos, mas as idades de campeãs femininas passam por uma brusca queda a partir de 2016, enquanto as idades dos campeões masculinos se elevam um pouco e se mantêm mais altas.

Foi também calculada a média de Grand Slams por vencedor nos últimos 10 anos. Ao repetir a query de contar vencedores distintos, encontram-se 9 vencedores masculinos e 22 femininos.

A screenshot of a PostgreSQL terminal window. The top bar includes buttons for 'Run' (highlighted in blue), 'Limit 100', 'Explain', 'Isolated session', and 'Serverless'. The main area contains the following SQL code:

```
1 select count(distinct winner_id)
2 from public.atp_stats
3 where tourney_date_int > 20140000
```

The results section shows a single row labeled 'Result 1 (1)' with a table header 'count' and a value of '9'.

	count
	9

A screenshot of a PostgreSQL terminal window. The top bar includes buttons for 'Run' (highlighted in blue), 'Limit 100', 'Explain', 'Isolated session', and 'Serverless: de...'. The main area contains the following SQL code:

```
1 select count(distinct winner_id)
2 from public.wta_stats
3 where tourney_date > 20140000
```

The results section shows a single row labeled 'Result 1 (1)' with a table header 'count' and a value of '22'.

	count
	22

Como foram disputados 39 torneios neste período:

$$39 \div 9 = 4.3$$

$$39 \div 22 = 1.7$$

O resultado é de uma média de 4.3 Grand Slams por vencedor masculino e 1.7 por vencedora feminina.

Analizando este resultado em conjunto com a tabela de médias, é perceptível uma disparidade grande com os valores históricos calculados anteriormente. Nos anos recentes, os vencedores masculinos têm uma média de idade significativamente mais alta e mais chances de acumular múltiplos títulos.

Para chegar a conclusões de causa, caberia aqui uma análise qualitativa completa dos campeões recentes. Porém, é plausível supor que o surgimento de jogadores com carreiras longevas e recordistas, como foi na última década com Novak Djokovic, afetam as estatísticas e momentaneamente a percepção do público a respeito de qual é a idade típica de um campeão de Grand Slam.

## **Autoavaliação**

Este trabalho foi um desafio muito novo para mim, visto que este é meu primeiro contato com o estudo formal de disciplinas de dados. Antes só havia aprendido SQL por conta própria e Python, mas não para o tópico de análise de dados. Trabalho na área de tecnologia há cerca de 1 ano e, com vontade de entender alguns dos assuntos que me cercam no dia-a-dia, ingressei na pós-graduação, com Engenharia de Dados sendo minha primeira sprint.

Selecionei a AWS por ser o ambiente utilizado na minha empresa e foi muito interessante, além de conhecer alguns dos serviços, como Redshift e Glue, entender aspectos de funcionamento da Cloud em si, como máquinas virtuais e permissionamento. Houve desafios com pequenas configurações de roles, mas tudo pôde ser resolvido com tranquilidade lendo guias de ajuda e pesquisando dúvidas online.

Creio que os objetivos do trabalho puderam ser respondidos e, com eles, pôde ser contada uma história interessante e relevante. No entanto, sei que propus perguntas simples dada a minha inexperiência com o assunto e foco em aprender, nesse sprint, o fluxo de ETL propriamente dito. Adoraria, após realizar a sprint de Análise de Dados, voltar a esses dados e poder responder as perguntas de forma mais completa com Python. Inclusive, acho que a evolução anual das idades e títulos poderia ser melhor contada com gráficos.

Dito isso, acredito que explorei bem as possibilidades dentro do que propus e fiquei feliz em conseguir iterar e aumentar um pouco a complexidade do trabalho conforme fui ganhando confiança, melhorando o fluxo de ETL e acrescentando perguntas ao objetivo inicial.

Obrigada aos professores pelas aulas e apoio!