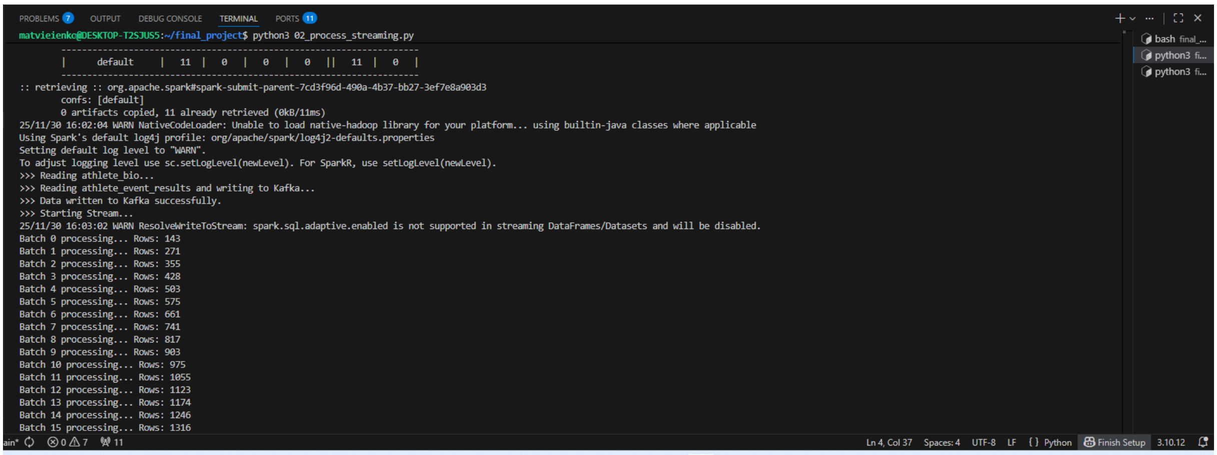
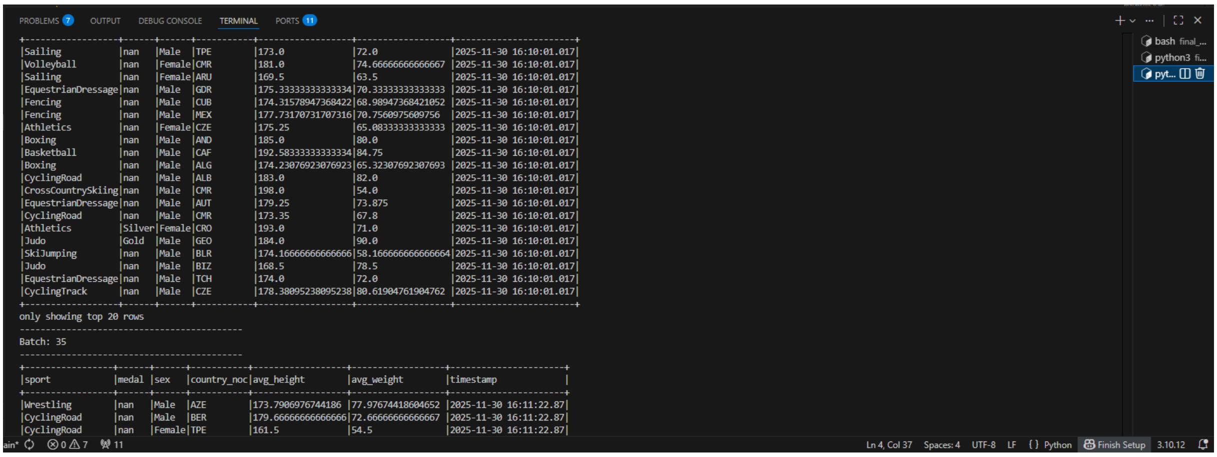
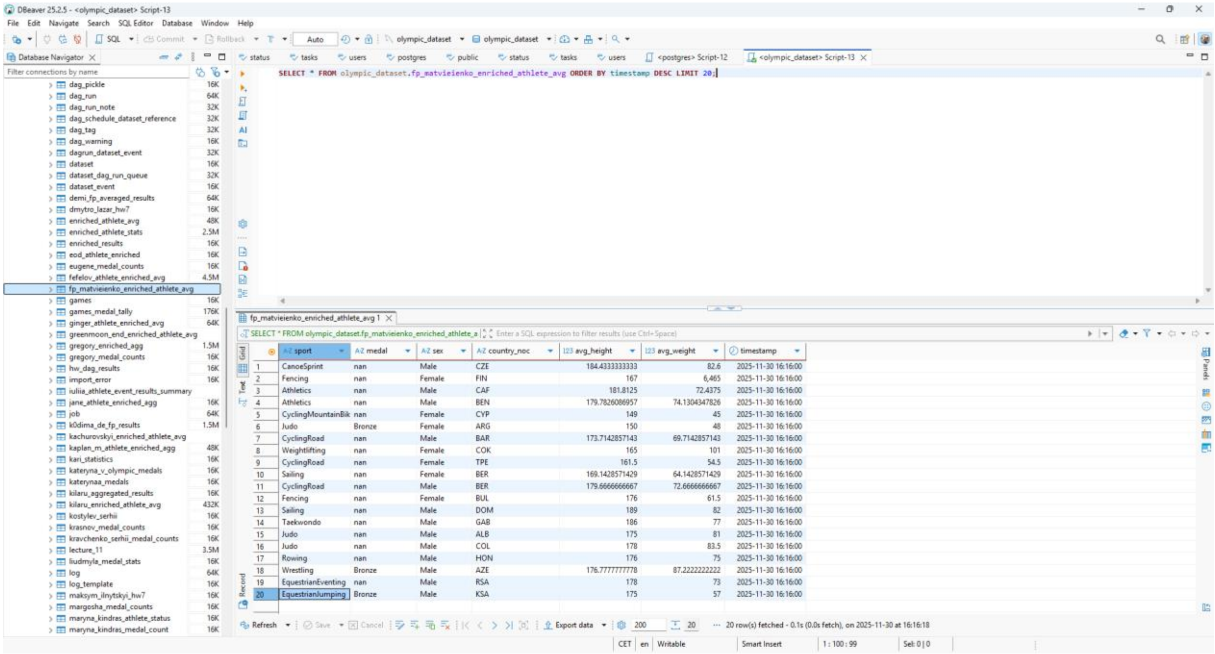
Частина 1. Building an End-to-End Streaming Pipeline



```
matvieienko@DESKTOP-T2SJUS5:~/final_project$ python3 02_process_streaming.py
        ---------------------------------------------------------------------
        |      default     |  11 |  0  |  0  |  0  ||  11 |  0  |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-7cd3f96d-490a-4b37-bb27-3ef7e8a903d3
        confs: [default]
        0 artifacts copied, 11 already retrieved (0kB/11ms)
25/11/30 16:02:04 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j2-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
>>> Reading athlete_bio...
>>> Reading athlete_event_results and writing to Kafka...
>>> Data written to Kafka successfully.
>>> Starting Stream...
25/11/30 16:03:02 WARN ResolveWriteToStream: spark.sql.adaptive.enabled is not supported in streaming DataFrames/Datasets and will be disabled.
Batch 0 processing... Rows: 143
Batch 1 processing... Rows: 271
Batch 2 processing... Rows: 355
Batch 3 processing... Rows: 428
Batch 4 processing... Rows: 503
Batch 5 processing... Rows: 575
Batch 6 processing... Rows: 661
Batch 7 processing... Rows: 741
Batch 8 processing... Rows: 817
Batch 9 processing... Rows: 903
Batch 10 processing... Rows: 975
Batch 11 processing... Rows: 1055
Batch 12 processing... Rows: 1123
Batch 13 processing... Rows: 1174
Batch 14 processing... Rows: 1246
Batch 15 processing... Rows: 1316
```



```
+-----------------+------+------+-----------+------------------+------------------+-----------------------+
|Sailing          |nan   |Male  |TPE        |173.0             |72.0              |2025-11-30 16:10:01.017|
|Volleyball       |nan   |Female|CMR        |181.0             |74.66666666666667 |2025-11-30 16:10:01.017|
|Sailing          |nan   |Female|ARU        |169.5             |63.5              |2025-11-30 16:10:01.017|
|EquestrianDressage|nan  |Male  |GDR        |175.33333333333334|70.33333333333333 |2025-11-30 16:10:01.017|
|Fencing          |nan   |Male  |CUB        |174.31578947368422|68.98947368421052 |2025-11-30 16:10:01.017|
|Fencing          |nan   |Male  |MEX        |177.73170731707316|70.7560975609756  |2025-11-30 16:10:01.017|
|Athletics        |nan   |Female|CZE        |175.25            |65.08333333333333 |2025-11-30 16:10:01.017|
|Boxing           |nan   |Male  |AND        |185.0             |80.0              |2025-11-30 16:10:01.017|
|Basketball       |nan   |Male  |CAF        |192.58333333333334|84.75             |2025-11-30 16:10:01.017|
|Boxing           |nan   |Male  |ALG        |174.23076923076923|65.32307692307693 |2025-11-30 16:10:01.017|
|CyclingRoad      |nan   |Male  |ALB        |183.0             |82.0              |2025-11-30 16:10:01.017|
|CrossCountrySkiing|nan  |Male  |CMR        |198.0             |54.0              |2025-11-30 16:10:01.017|
|EquestrianDressage|nan  |Male  |AUT        |179.25            |73.875            |2025-11-30 16:10:01.017|
|CyclingRoad      |nan   |Male  |CMR        |173.35            |67.8              |2025-11-30 16:10:01.017|
|Athletics        |Silver|Female|CRO        |193.0             |71.0              |2025-11-30 16:10:01.017|
|Judo             |Gold  |Male  |GEO        |184.0             |90.0              |2025-11-30 16:10:01.017|
|SkiJumping       |nan   |Male  |BLR        |174.16666666666666|58.166666666666664|2025-11-30 16:10:01.017|
|Judo             |nan   |Male  |BIZ        |168.5             |78.5              |2025-11-30 16:10:01.017|
|EquestrianDressage|nan  |Male  |TCH        |174.0             |72.0              |2025-11-30 16:10:01.017|
|CyclingTrack     |nan   |Male  |CZE        |178.38095238095238|80.61904761904762 |2025-11-30 16:10:01.017|
+-----------------+------+------+-----------+------------------+------------------+-----------------------+
only showing top 20 rows
-------------------------------------------
Batch: 35
-------------------------------------------
+-----------------+------+------+-----------+------------------+------------------+----------------------+
|sport            |medal |sex   |country_noc|avg_height        |avg_weight        |timestamp             |
+-----------------+------+------+-----------+------------------+------------------+----------------------+
|Wrestling        |nan   |Male  |AZE        |173.7906976744186 |77.97674418604652 |2025-11-30 16:11:22.87|
|CyclingRoad      |nan   |Male  |BER        |179.66666666666666|72.66666666666667 |2025-11-30 16:11:22.87|
|CyclingRoad      |nan   |Female|TPE        |161.5             |54.5              |2025-11-30 16:11:22.87|
```

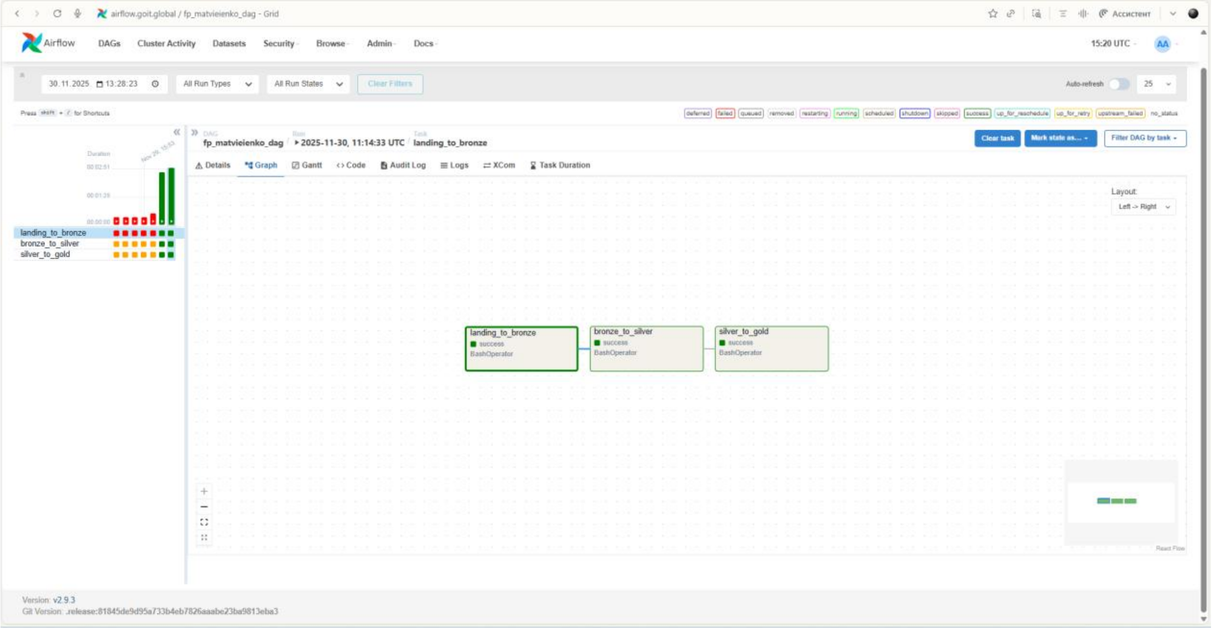


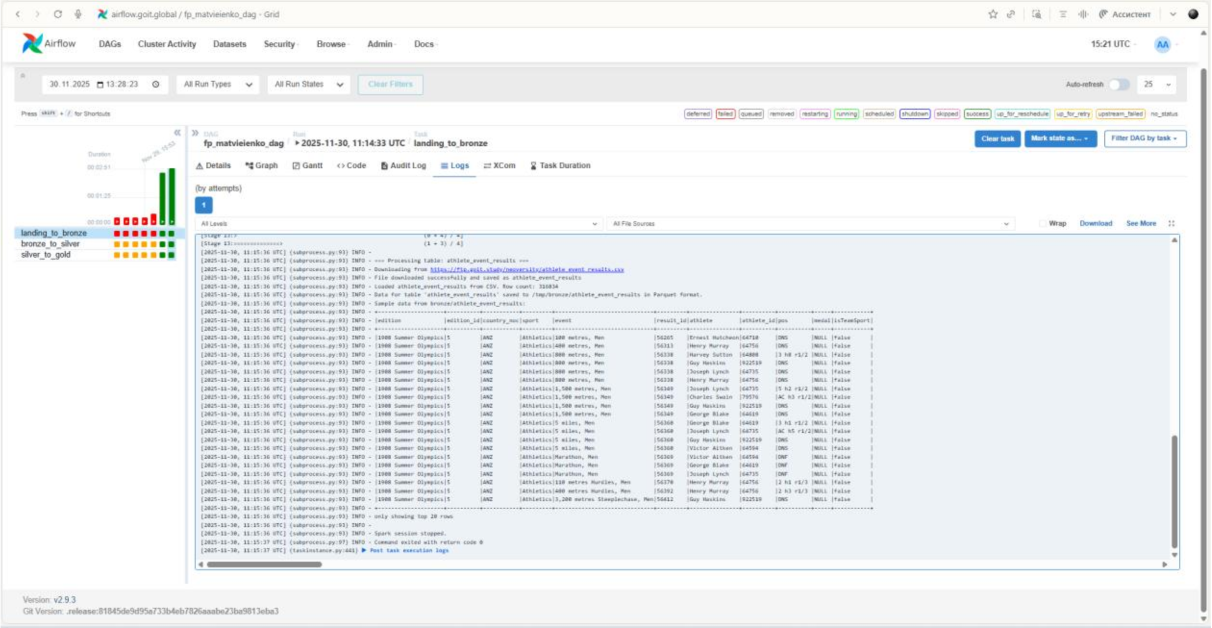
```sql
SELECT * FROM olympic_dataset.fp_matvieienko_enriched_athlete_avg ORDER BY timestamp DESC LIMIT 20;
```

	sport	medal	sex	country_noc	avg_height	avg_weight	timestamp
1	CanoeSprint	nan	Male	CZE	184.4333333333	82.6	2025-11-30 16:16:00
2	Fencing	nan	Female	FIN	167	6,465	2025-11-30 16:16:00
3	Athletics	nan	Male	CAF	181.8125	72.4375	2025-11-30 16:16:00
4	Athletics	nan	Male	BEN	179.7826086957	74.1304347826	2025-11-30 16:16:00
5	CyclingMountainBik	nan	Female	CYP	149	45	2025-11-30 16:16:00
6	Judo	Bronze	Female	ARG	150	48	2025-11-30 16:16:00
7	CyclingRoad	nan	Male	BAR	173.7142857143	69.7142857143	2025-11-30 16:16:00
8	Weightlifting	nan	Female	COK	165	101	2025-11-30 16:16:00
9	CyclingRoad	nan	Female	TPE	161.5	54.5	2025-11-30 16:16:00
10	Sailing	nan	Female	BER	169.1428571429	64.1428571429	2025-11-30 16:16:00
11	CyclingRoad	nan	Male	BER	179.6666666667	72.6666666667	2025-11-30 16:16:00
12	Fencing	nan	Female	BUL	176	61.5	2025-11-30 16:16:00
13	Sailing	nan	Male	DOM	189	82	2025-11-30 16:16:00
14	Taekwondo	nan	Male	GAB	186	77	2025-11-30 16:16:00
15	Judo	nan	Male	ALB	175	81	2025-11-30 16:16:00
16	Judo	nan	Male	COL	178	83.5	2025-11-30 16:16:00
17	Rowing	nan	Male	HON	176	75	2025-11-30 16:16:00
18	Wrestling	Bronze	Male	AZE	176.7777777778	87.2222222222	2025-11-30 16:16:00
19	EquestrianEventing	nan	Male	RSA	178	73	2025-11-30 16:16:00
20	EquestrianJumping	Bronze	Male	KSA	175	57	2025-11-30 16:16:00

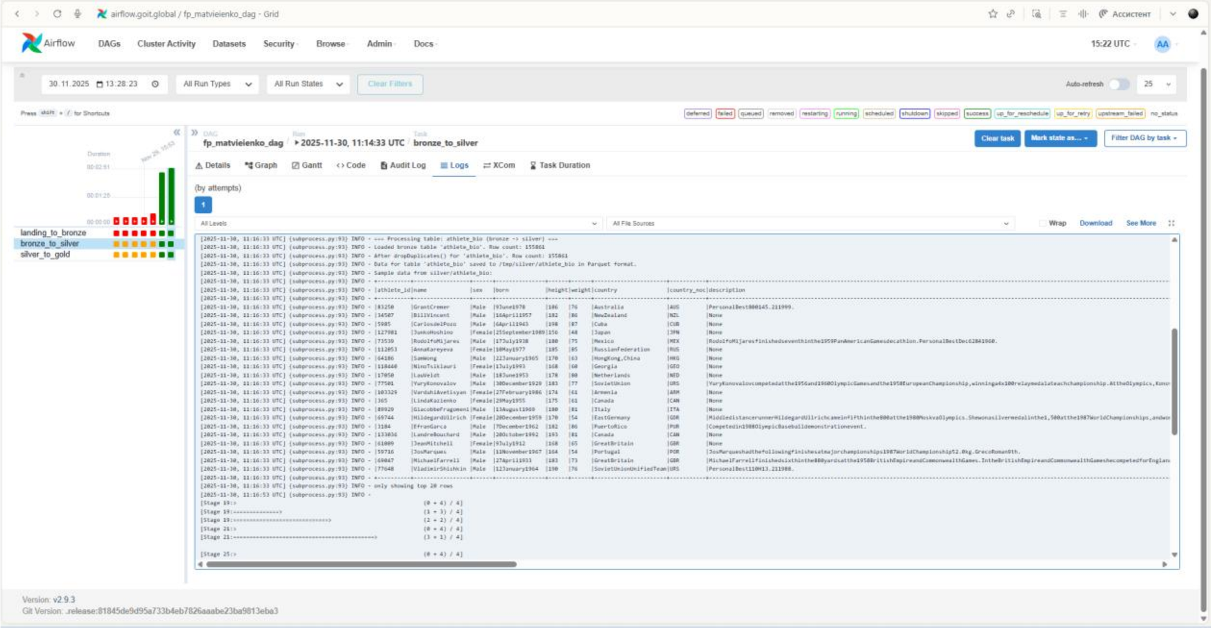
20 row(s) fetched - 0.1s (0.0s fetch), on 2025-11-30 at 16:16:18

Частина 2. Building an End-to-End Batch Data Lake



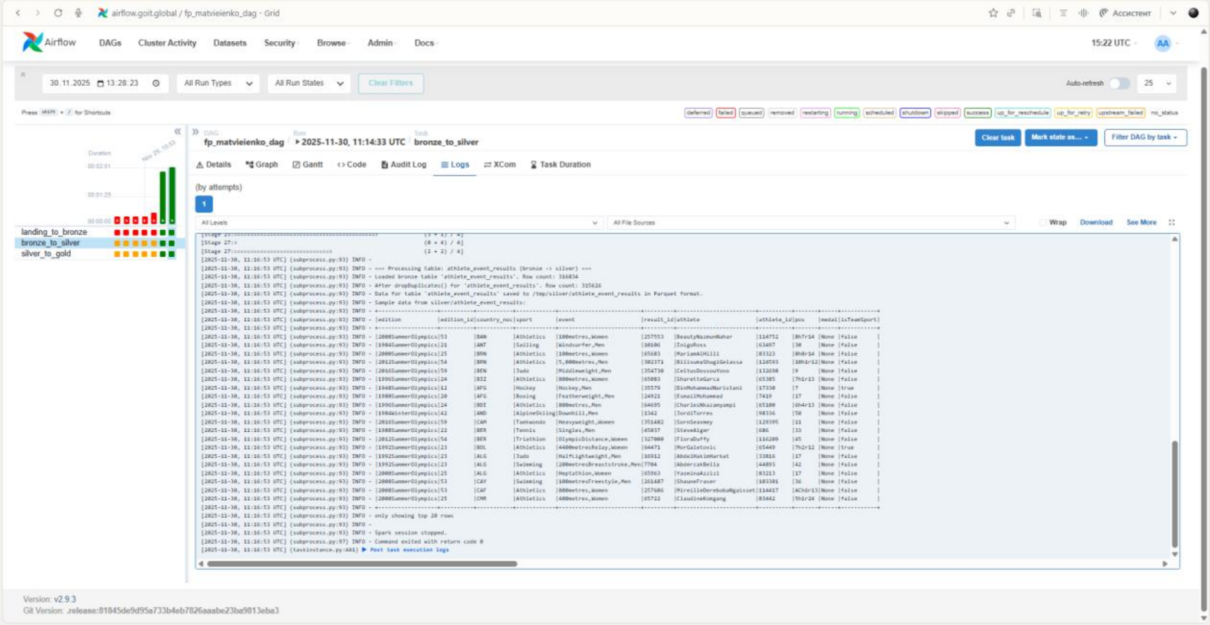


```
[Stage 13:>                                                         (0 + 4) / 4]
[Stage 13:===============>                                          (1 + 3) / 4]
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO -
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - === Processing table: athlete_event_results ===
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - Downloading from https://ftp.goit.study/neoversity/athlete_event_results.csv
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - File downloaded successfully and saved as athlete_event_results
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - Loaded athlete_event_results from CSV. Row count: 316834
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - Data for table 'athlete_event_results' saved to /tmp/bronze/athlete_event_results in Parquet format.
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - Sample data from bronze/athlete_event_results:
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - +--------------------+----------+-----------+---------+------------------------------+---------+---------------+----------+----------+-----+-----------+
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |edition             |edition_id|country_noc|sport    |event                         |result_id|athlete        |athlete_id|pos       |medal|isTeamSport|
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - +--------------------+----------+-----------+---------+------------------------------+---------+---------------+----------+----------+-----+-----------+
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|100 metres, Men               |56265    |Ernest Hutchcon|64710     |DNS       |NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|400 metres, Men               |56313    |Henry Murray   |64756     |DNS       |NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|800 metres, Men               |56338    |Harvey Sutton  |64808     |3 h8 r1/2 |NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|800 metres, Men               |56338    |Guy Haskins    |922519    |DNS       |NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|800 metres, Men               |56338    |Joseph Lynch   |64735     |DNS       |NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|800 metres, Men               |56338    |Henry Murray   |64756     |DNS       |NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|1,500 metres, Men             |56349    |Joseph Lynch   |64735     |5 h2 r1/2 |NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|1,500 metres, Men             |56349    |Charles Swain  |79576     |AC h3 r1/2|NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|1,500 metres, Men             |56349    |Guy Haskins    |922519    |DNS       |NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|1,500 metres, Men             |56349    |George Blake   |64619     |DNS       |NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|5 miles, Men                  |56360    |George Blake   |64619     |3 h1 r1/2 |NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|5 miles, Men                  |56360    |Joseph Lynch   |64735     |AC h5 r1/2|NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|5 miles, Men                  |56360    |Guy Haskins    |922519    |DNS       |NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|5 miles, Men                  |56360    |Victor Aitken  |64594     |DNS       |NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|Marathon, Men                 |56369    |Victor Aitken  |64594     |DNF       |NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|Marathon, Men                 |56369    |George Blake   |64619     |DNF       |NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|Marathon, Men                 |56369    |Joseph Lynch   |64735     |DNF       |NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|110 metres Hurdles, Men       |56370    |Henry Murray   |64756     |2 h1 r1/3 |NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|400 metres Hurdles, Men       |56392    |Henry Murray   |64756     |2 h3 r1/3 |NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - |1908 Summer Olympics|5         |ANZ        |Athletics|3,200 metres Steeplechase, Men|56412    |Guy Haskins    |922519    |DNS       |NULL |false      |
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - +--------------------+----------+-----------+---------+------------------------------+---------+---------------+----------+----------+-----+-----------+
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - only showing top 20 rows
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO -
[2025-11-30, 11:15:36 UTC] {subprocess.py:93} INFO - Spark session stopped.
[2025-11-30, 11:15:37 UTC] {subprocess.py:97} INFO - Command exited with return code 0
[2025-11-30, 11:15:37 UTC] {taskinstance.py:441} ▶ Post task execution logs
```

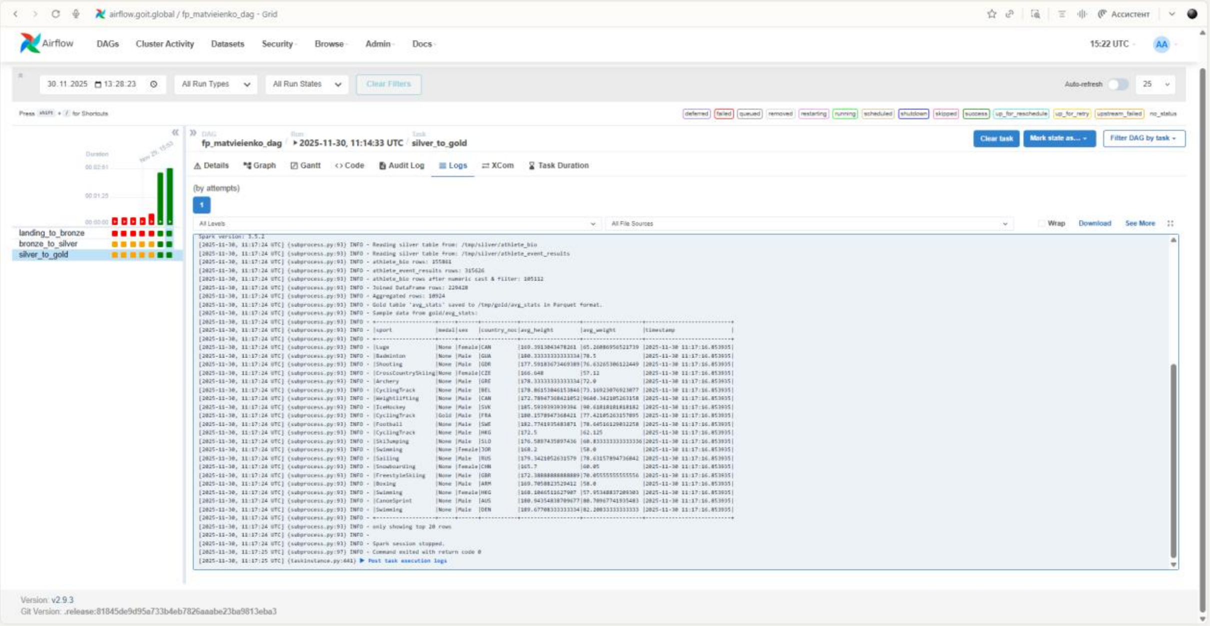


```
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - === Processing table: athlete_bio (bronze -> silver) ===
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - Loaded bronze table 'athlete_bio'. Row count: 155861
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - After dropDuplicates() for 'athlete_bio'. Row count: 155861
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - Data for table 'athlete_bio' saved to /tmp/silver/athlete_bio in Parquet format.
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - Sample data from silver/athlete_bio:
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - +----------+-----------------+------+---------------+------+------+----------------------+-----------+---------------------------------------------------------------------------------------------------------------------------------------------
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |athlete_id|name             |sex   |born           |height|weight|country               |country_noc|description
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - +----------+-----------------+------+---------------+------+------+----------------------+-----------+---------------------------------------------------------------------------------------------------------------------------------------------
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |83250     |GrantCremer      |Male  |9June1978      |186   |76    |Australia             |AUS        |PersonalBest800145.211999.
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |34507     |BillVincent      |Male  |16April1957    |182   |86    |NewZealand            |NZL        |None
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |5985      |CarlosdelPozo    |Male  |6April1943     |198   |87    |Cuba                  |CUB        |None
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |127981    |JunkoHoshino     |Female|25September1989|156   |48    |Japan                 |JPN        |None
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |73530     |RodolfoMijares   |Male  |17July1938     |180   |75    |Mexico                |MEX        |RodolfoMijaresfinishedseventhinthe1959PanAmericanGamesdecathlon.PersonalBestDec62841960.
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |112053    |AnnaKareyeva     |Female|10May1977      |185   |85    |RussianFederation     |RUS        |None
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |64186     |SamWong          |Male  |22January1965  |170   |63    |HongKong,China        |HKG        |None
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |118440    |NinoTsiklauri    |Female|1July1993      |168   |60    |Georgia               |GEO        |None
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |17050     |LauVeldt         |Male  |18June1953     |178   |80    |Netherlands           |NED        |None
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |77501     |YuryKonovalov    |Male  |30December1929 |183   |77    |SovietUnion           |URS        |YuryKonovalovcompetedatthe1956and1960OlympicGamesandthe1958EuropeanChampionship,winninga4x100relaymedalateachchampionship.AttheOlympics,Kono
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |103329    |VarduhiAvetisyan |Female|27February1986 |174   |61    |Armenia               |ARM        |None
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |365       |LindaKazienko    |Female|29May1955      |175   |61    |Canada                |CAN        |None
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |89929     |GiacobbeFragomeni|Male  |13August1969   |180   |81    |Italy                 |ITA        |None
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |69744     |HildegardUllrich |Female|20December1959 |170   |54    |EastGermany           |GDR        |MiddledistancerunnerHildegardUllrichcameinfifthinthe800atthe1980MoskvaOlympics.Shewonasilvermedalinthe1,500atthe1987WorldChampionships,andwo
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |3184      |EfranGarca       |Male  |7December1962  |182   |86    |PuertoRico            |PUR        |Competedin1988OlympicBaseballdemonstrationevent.
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |133036    |LandreBouchard   |Male  |20October1992  |193   |81    |Canada                |CAN        |None
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |61009     |JeanMitchell     |Female|9July1912      |168   |65    |GreatBritain          |GBR        |None
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |59716     |JosMarques       |Male  |11November1967 |164   |54    |Portugal              |POR        |JosMarqueshadthefollowingfinishesatmajorchampionships1987WorldChampionship52.0kg.GrecoRoman9th.
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |69047     |MichaelFarrell   |Male  |27April1933    |183   |73    |GreatBritain          |GBR        |MichaelFarrellfinishedsixthinthe880yardsatthe1958BritishEmpireandCommonwealthGames.IntheBritishEmpireandCommonwealthGameshecompetedforEnglan
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - |77648     |VladimirShishkin |Male  |12January1964  |190   |76    |SovietUnionUnifiedTeam|URS        |PersonalBest110H13.211988.
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - +----------+-----------------+------+---------------+------+------+----------------------+-----------+---------------------------------------------------------------------------------------------------------------------------------------------
[2025-11-30, 11:16:33 UTC] {subprocess.py:93} INFO - only showing top 20 rows
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO -
[Stage 19:>                                                         (0 + 4) / 4]
[Stage 19:==============>                                           (1 + 3) / 4]
[Stage 19:=============================>                            (2 + 2) / 4]
[Stage 21:>                                                         (0 + 4) / 4]
[Stage 21:===========================================>              (3 + 1) / 4]

[Stage 25:>                                                         (0 + 4) / 4]
```



```
[Stage 25:==============================================>           (3 + 1) / 4]
[Stage 27:>                                                          (0 + 4) / 4]
[Stage 27:=============================>                             (2 + 2) / 4]
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO -
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - === Processing table: athlete_event_results (bronze -> silver) ===
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - Loaded bronze table 'athlete_event_results'. Row count: 316834
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - After dropDuplicates() for 'athlete_event_results'. Row count: 315626
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - Data for table 'athlete_event_results' saved to /tmp/silver/athlete_event_results in Parquet format.
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - Sample data from silver/athlete_event_results:
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - +------------------+----------+-----------+------------+-------------------------+---------+------------------------+----------+-------+-----+-----------+
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |edition           |edition_id|country_noc|sport       |event                    |result_id|athlete                 |athlete_id|pos    |medal|isTeamSport|
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - +------------------+----------+-----------+------------+-------------------------+---------+------------------------+----------+-------+-----+-----------+
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |2008SummerOlympics|53        |BAN        |Athletics   |100metres,Women          |257553   |BeautyNazmunNahar       |114752    |8h7r14 |None |false      |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |1984SummerOlympics|21        |ANT        |Sailing     |Windsurfer,Men           |10106    |InigoRoss               |63497     |30     |None |false      |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |2000SummerOlympics|25        |BRN        |Athletics   |100metres,Women          |65683    |MariamAlHilli           |83323     |8h8r14 |None |false      |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |2012SummerOlympics|54        |BRN        |Athletics   |5,000metres,Men          |302371   |BilisumaShugiGelassa    |124593    |10h1r12|None |false      |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |2016SummerOlympics|59        |BEN        |Judo        |Middleweight,Men         |354730   |CeltusDossouYovo        |132698    |9      |None |false      |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |1996SummerOlympics|24        |BIZ        |Athletics   |800metres,Women          |65083    |SharetteGarca           |65385     |7h1r13 |None |false      |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |1948SummerOlympics|12        |AFG        |Hockey      |Hockey,Men               |35579    |DinMohammadNuristani    |17330     |7      |None |true       |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |1980SummerOlympics|20        |AFG        |Boxing      |Featherweight,Men        |24921    |EsmailMohammad          |7419      |17     |None |false      |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |1996SummerOlympics|24        |BDI        |Athletics   |800metres,Men            |64695    |CharlesNkazamyampi      |65100     |6h4r13 |None |false      |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |1984WinterOlympics|42        |AND        |AlpineSkiing|Downhill,Men             |1342     |JordiTorres             |98336     |50     |None |false      |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |2016SummerOlympics|59        |CAM        |Taekwondo   |Heavyweight,Women        |351482   |SornSeavmey             |129395    |11     |None |false      |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |1988SummerOlympics|22        |BER        |Tennis      |Singles,Men              |45037    |SteveAlger              |686       |33     |None |false      |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |2012SummerOlympics|54        |BER        |Triathlon   |OlympicDistance,Women    |327000   |FloraDuffy              |116209    |45     |None |false      |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |1992SummerOlympics|23        |BOL        |Athletics   |4x400metresRelay,Women   |64471    |MorGaletovic            |65449     |7h2r12 |None |true       |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |1992SummerOlympics|23        |ALG        |Judo        |HalfLightweight,Men      |16912    |AbdelHakimHarkat        |33816     |17     |None |false      |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |1992SummerOlympics|23        |ALG        |Swimming    |200metresBreaststroke,Men|7704     |AbderzakBella           |44893     |42     |None |false      |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |2000SummerOlympics|25        |ALG        |Athletics   |Heptathlon,Women         |65963    |YasminaAzzizi           |83213     |17     |None |false      |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |2008SummerOlympics|53        |CAY        |Swimming    |100metresFreestyle,Men   |261487   |ShauneFraser            |103381    |36     |None |false      |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |2008SummerOlympics|53        |CAF        |Athletics   |800metres,Women          |257606   |MireilleDerebonaNgaisset|114417    |ACh6r13|None |false      |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - |2000SummerOlympics|25        |CMR        |Athletics   |400metres,Women          |65722    |ClaudineKomgang         |83442     |5h1r24 |None |false      |
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - +------------------+----------+-----------+------------+-------------------------+---------+------------------------+----------+-------+-----+-----------+
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - only showing top 20 rows
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO -
[2025-11-30, 11:16:53 UTC] {subprocess.py:93} INFO - Spark session stopped.
[2025-11-30, 11:16:53 UTC] {subprocess.py:97} INFO - Command exited with return code 0
[2025-11-30, 11:16:53 UTC] {taskinstance.py:441} ▶ Post task execution logs
```



```
Spark version: 3.5.2
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - Reading silver table from: /tmp/silver/athlete_bio
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - Reading silver table from: /tmp/silver/athlete_event_results
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - athlete_bio rows: 155861
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - athlete_event_results rows: 315626
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - athlete_bio rows after numeric cast & filter: 105112
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - Joined DataFrame rows: 229428
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - Aggregated rows: 10924
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - Gold table 'avg_stats' saved to /tmp/gold/avg_stats in Parquet format.
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - Sample data from gold/avg_stats:
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - +------------------+-----+------+-----------+------------------+------------------+--------------------------+
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |sport             |medal|sex   |country_noc|avg_height        |avg_weight        |timestamp                 |
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - +------------------+-----+------+-----------+------------------+------------------+--------------------------+
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |Luge              |None |Female|CAN        |169.3913043478261 |65.26086956521739 |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |Badminton         |None |Male  |GUA        |180.33333333333334|78.5              |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |Shooting          |None |Male  |GDR        |177.59183673469389|76.63265306122449 |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |CrossCountrySkiing|None |Female|CZE        |166.048           |57.12             |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |Archery           |None |Male  |GRE        |178.33333333333334|72.0              |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |CyclingTrack      |None |Male  |BEL        |178.86153846153846|73.16923076923077 |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |Weightlifting     |None |Male  |CAN        |172.78947368421052|9640.342105263158 |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |IceHockey         |None |Male  |SVK        |185.5939393939394 |90.61818181818182 |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |CyclingTrack      |Gold |Male  |FRA        |180.1578947368421 |77.42105263157895 |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |Football          |None |Male  |SWE        |182.7741935483871 |78.64516129032258 |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |CyclingTrack      |None |Male  |HKG        |172.5             |62.125            |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |SkiJumping        |None |Male  |SLO        |176.5897435897436 |68.83333333333336 |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |Swimming          |None |Female|JOR        |168.2             |58.0              |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |Sailing           |None |Male  |RUS        |179.3421052631579 |78.63157894736842 |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |Snowboarding      |None |Female|CHN        |165.7             |60.05             |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |FreestyleSkiing   |None |Male  |GBR        |172.38888888888889|70.05555555555556 |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |Boxing            |None |Male  |ARM        |169.7058823529412 |58.0              |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |Swimming          |None |Female|HKG        |168.1046511627907 |57.95348837209303 |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |CanoeSprint       |None |Male  |AUS        |180.94354838709677|80.70967741935483 |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - |Swimming          |None |Male  |DEN        |189.67708333333334|82.20833333333333 |2025-11-30 11:17:16.853935|
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - +------------------+-----+------+-----------+------------------+------------------+--------------------------+
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - only showing top 20 rows
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO -
[2025-11-30, 11:17:24 UTC] {subprocess.py:93} INFO - Spark session stopped.
[2025-11-30, 11:17:25 UTC] {subprocess.py:97} INFO - Command exited with return code 0
[2025-11-30, 11:17:25 UTC] {taskinstance.py:441} ▶ Post task execution logs
```