Фінальний проект

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

thlete_id	name	sex	born	height	weight	country	country_noc	description	special_note
 65649	IvankaBonova	 Female	4April1949	166.0	 55	Bulgaria	BUL	PersonalBest40053	na
112510	NataliyaUryadova	Female	15March1977	184.0	70	RussianFederation	RUS	nan	ListedinOlympians
114973	EssaIsmailRashed	Male	14December1986	165.0	55	Qatar	QAT	PersonalBest10000	ListedinOlympians
30359	PterBoros	Male	12January1908		nan	Hungary	HUN	Between1927and193	na na
50557	RudolfPiowat	Male	28April1900		nan	Czechoslovakia	TCH	RudolfPiowatyjoin	l n
146111	SvetlanaKholomina	Female	9November1997		nan	ROC	ROC	nan	l n
133041	VincentRiendeau	Male	13December1996	178.0	68	Canada	CAN	nan	ListedinOlympians.
110425	TanjaMorel	Female	40ctober1975	164.0	58	Switzerland	SUI	nan	ListedinOlympians.
110705	MaksimShabalin	Male	25January1982	183.0	76	RussianFederation	RUS	nan	ListedinOlympians.
54541	GRegter	Male	6March1916		nan	Netherlands	NED	nan	n n
22721	AristidePontenani	Male	nan		nan	Italy	ITA	nan	n
56266	GoYeongChang	Male	21March1926	167.0	75	RepublicofKorea	KOR	nan	n
82227	MarliesRostock	Female	20April1960		nan	EastGermany	GDR	MarliesRostockwon	WifedivorcedofUwe.
93334	CraigHutchison	Male	26May1975	198.0	97	Canada	CAN	nan	ListedinOlympians.
146013	RaquelOueirs	Female	4March2000	167.0	56	Portugal	POR	nan	i n
109912	VyacheslavKurginyan	Male	22December1986	170.0	65	RussianFederation	RUS	nan	ListedinOlympians.
37019	PhilippeLot	Male	13June1967	189.0	85	France	FRA	nan	ListedinOlympians.
22885	RudyKugeler	Male	11August1928	187.0	86	Luxembourg	LUX		i í n
95497	YoshihiroMiyazaki	Male	10May1930			Japan	JPN	nan	n.
76	RoperBarrett	Male	24November1873		nan	GreatBritain	GBR	RoperBarrettlearn	n.
47023	LornaFrampton	Female	2April1920		nan	GreatBritain		LornaFramptonwasj	ListedinOlvmpians.
42069	IsabelleHberl	Female	11May1959	167.0	67	France	FRA		SisterofPhilippeH.
39033	SjoerdWartena	Male	1May1939			Netherlands	NED		n.
115157	MariyaYakovenko		6January1982			RussianFederation		PersonalBestJT62	n.
207	VirgCsurg		10November1972			Hungary		VirgCsurgwasaprof	
99106	RonnyYeager	Male	16August1952			UnitedStates		RonnyYeagerskiedf	
126257	ShaneSmeltz	Male	29September1981		79	NewZealand			ListedinOlympians.
102554	HwangSeonHong	Male	14July1968			RepublicofKorea	KOR		ListedinOlympians.
24775	FathiKhorshid	Male	11December1937			EgyptUnitedArabRe		Growingupintheagr	
69614	HarryWood	Male	28November1902		nan	GreatBritain		Finishing11thatth	l n
52387	UlfTtken	Male	31May1953	196.0		WestGermany		UlfTtkenwasaWestG	
100169	GalitChait		29January1975			Israel	ISR		
142869	JulienMertine	Male	25June1988		nan	France	l FRA	nan	n
63052	JimmyMooney	Male	25February1920	182.0		Ireland	IRL	Deceased.	
	AnneMetteChristensen		23December1963		nan	Denmark			
136267	ShadrackKipchirchir	Male	22February1989	173.0		UnitedStates	USA	nan	
1005805	JosHansen	Male	8June1932		nan	Luxembourg	LUX		
95761	AlainNavillod	Male	16August1955	169.0		France	l FRA	l nan	
62527	AndreasJosenhans		12September1950			Canada	CAN		
112346	PaulEtiaNdoumb	Male	20April1984			Cameroon	l CMR	l nan	

Зчитуємо дані атлетів з бази.

thlete_id	name	sex	born	height	weight	country	country_noc	description	special_note
65649	IvankaBonova	Female	4April1949	166.0	55.0	Bulgaria	BUL	PersonalBest40053	na
112510	NataliyaUryadova	Female	15March1977	184.0	70.0	RussianFederation	RUS	nan	ListedinOlympians
114973	EssaIsmailRashed	Male	14December1986	165.0	55.0	Qatar	QAT	PersonalBest10000	ListedinOlympians
133041	VincentRiendeau	Male	13December1996	178.0	68.0	Canada	CAN		ListedinOlympians
110425	TanjaMorel		40ctober1975		58.0	Switzerland			ListedinOlympians
110705	MaksimShabalin		25January1982		76.0	RussianFederation	RUS	nan	ListedinOlympians
56266			21March1926		75.0	RepublicofKorea	KOR	nan	na
93334	CraigHutchison		26May1975		97.0	Canada	CAN		ListedinOlympians
146013	RaquelQueirs		4March2000		56.0	Portugal	POR	nan	na
109912	VyacheslavKurginyan	Male	22December1986	170.0	65.0	RussianFederation	RUS		ListedinOlympians
37019	PhilippeLot	Male	13June1967	189.0	85.0	France	FRA	nan	ListedinOlympians
22885	RudyKugeler		11August1928	187.0	86.0	Luxembourg	LUX	nan	na
95497	YoshihiroMiyazaki	Male	10May1930	173.0	70.0	Japan	JPN	nan	na
42069	IsabelleHberl	Female	11May1959	167.0	67.0	France	FRA	nan	SisterofPhilippeH
39033	SjoerdWartena	Male	1May1939	189.0	76.0	Netherlands	NED	nan	na
115157	MariyaYakovenko	Female	6January1982	174.0	81.0	RussianFederation	RUS	PersonalBestJT62	na
207	VirgCsurg	Female	10November1972	172.0	63.0	Hungary	HUN	VirgCsurgwasaprof	ListedinOlympians.
99106	RonnyYeager	Male	16August1952	181.0	70.0	UnitedStates	USA	RonnyYeagerskiedf	ListedinOlympians.
126257	ShaneSmeltz	Male	29September1981	184.0	79.0	NewZealand	NZL	nan	ListedinOlympians.
102554	HwangSeonHong	Male	14July1968	183.0	79.0	RepublicofKorea	KOR	nan	ListedinOlympians
52387	UlfTtken	Male	31May1953	196.0	86.0	WestGermany	FRG	UlfTtkenwasaWestG	na
100169	GalitChait	Female	29January1975	153.0	45.0	Israel	ISR	Dance1994WorldCha	ListedinOlympians
63052	JimmyMooney	Male	25February1920	182.0	71.0	Ireland	IRL	Deceased.	na
136267	ShadrackKipchirchir	Male	22February1989	173.0	54.0	UnitedStates	USA	nan	na
95761	AlainNavillod	Male	16August1955	169.0	68.0	France	FRA	nan	n
62527	AndreasJosenhans	Male	12September1950	183.0	81.0	Canada	CAN	nan	na
112346	PaulEtiaNdoumb	Male	20April1984	182.0	78.0	Cameroon	CMR	nan	na
96506	PhilippePoirot	Male	30ctober1958	173.0	66.0	France	FRA	nan	na
8923	OscarChongo	Male	25January1965	155.0	51.0	Zambia	ZAM	nan	na
64560	EricaNixon	Female	15December1953	175.0	67.0	Australia	AUS	PersonalBestLJ6.5	MotherofSteveHook
115463	RyanHall	Male	140ctober1982	180.0	61.0	UnitedStates	USA	Hiswife,SaraHalln	na
125355	AugustoSoares	Male	22August1986	165.0	58.0	TimorLeste	TLS	nan	n
137459	KeijiTanaka	Male	22November1994	172.0	70.0	Japan	JPN	nan	ListedinOlympians.
58561	JzsefFarkas	Male	17March1952	192.0	100.0	Hungary	HUN	JzsefFarkashadthe	na
2821	IlijaLupulesku	Male	300ctober1967	177.0	79.0	IndependentOlympi	IOA	nan	Husbanddivorcedof.
77525	BorisKuznetsov	Male	12June1947	175.0	63.0	SovietUnion	URS	PersonalBest50001	ListedinOlympians
127330	BenjaminMaier	Male	19April1994	182.0	93.0	Austria	AUT	nan	BrotherofRaphaelM.
37731	BirteSiech	Female	19March1967	180.0	75.0	EastGermanyGermany	GDR	RowerBirteSiechwo	na
90206	lhamKrimov	Male	2July1976	180.0	81.0	Azerbaijan	AZE	nan	na
27348	FrantiekKunzo	Male	17September1954	175.0	70.0	Czechoslovakia	TCH	nan	na

editi	on edition_io	country_noc	sport	event	result_id	athlete	athlete_id	pos	medal	isTeamSpor
.908SummerOlympi	.cs 5	i ANZ	Athletics	100metres,Men	56265	ErnestHutcheon	647 1 0	DNS	nan	Fals
.908SummerOlympi		ANZ	Athletics	400metres,Men	56313	HenryMurray	64756	DNS	nan	Fals
.908SummerOlymp			Athletics		56338	HarveySutton	64808	3h8r12	nan	Fals
.908SummerOlympi			Athletics					DNS	nan	Fal:
.908SummerOlympi			Athletics	800metres,Men				DNS	nan	Fal
.908SummerOlympi			Athletics	800metres,Men		, , , , , , , , , , , , , , , , , , , ,		DNS	nan	Fal
.908SummerOlympi			Athletics	1,500metres,Men				5h2r12	nan	Fal
.908SummerOlympi			Athletics	1,500metres,Men			· · · · · · · · · · · · · · · · · · ·	ACh3r12	nan	Fal
908SummerOlympi			Athletics				· · · · · · · · · · · · · · · · · · ·	DNS	nan	Fal
.908SummerOlymp:			Athletics	1,500metres,Men				DNS	nan	Fal
908SummerOlympi			Athletics						nan	Fal
908SummerOlymp:			Athletics	5miles,Men				ACh5r12	nan	Fal
908SummerOlympi			Athletics	5miles,Men				DNS	nan	Fal
908SummerOlympi			Athletics	5miles,Men				DNS	nan	Fal
908SummerOlympi			Athletics			VictorAitken		DNF	nan	Fal Fal
908SummerOlympi 908SummerOlympi			Athletics Athletics	Marathon,Men Marathon.Men				DNF DNF	nan	Fal Fal
908SummerOlympi 908SummerOlympi				marathon,men 110metresHurdles.Men				2h1r13	nan nan	Fal Fal
908SummerOlympi 908SummerOlympi				110metreshurdles,Men 400metresHurdles,Men				2h1r13	nan	Fal
908SummerOlympi				3,200metresSteepl				DNS	nan	Fal
908SummerOlympi				3,500metresRaceWa					Bronze	Fal
908SummerOlympi				3,500metresRaceWa				5	nan	Fal
908SummerOlympi				10milesRaceWalk,Men				DNSr22	nan	Fal
908SummerOlympi			Athletics					5h1r12	nan	Fal
908SummerOlympi				10milesRaceWalk,Men				DNS	nan	Fal
908SummerOlympi				StandingHighJump,Men		ErnestHutcheon		AC	nan	Fal
908SummerOlympi			Athletics			ErnestHutcheon		DNS	nan	Fal
908SummerOlympi			Athletics	StandingLongJump,Men		ErnestHutcheon	64710	DNS	nan	Fal
908SummerOlympi	.cs	ANZ	Athletics	TripleJump,Men	56482	ErnestHutcheon	64710	DNS	nan	Fal
908SummerOlympi	.cs 5	ANZ	Boxing	Lightweight, Men	21247	GeorgeBlake	64619	DNS	nan	Fal
908SummerOlympi	.cs 5	ANZ	Boxing	Middleweight,Men	21263	SnowyBaker	45153	2	Silver	Fal
908SummerOlympi	.cs	ANZ	Diving	Springboard, Men	54013	SnowyBaker	45153	6p4r13	nan	Fal
908SummerOlympi	.cs	ANZ	Rugby	Rugby, Men	31505	JohnBarnett	11237	1	Gold	Tr
908SummerOlympi			Rugby	Rugby, Men	31505	PhilCarmichael	11239	1	Gold	Tr
908SummerOlympi			Rugby	Rugby,Men	31505	DanCarroll	11240	1	Gold	Tr
908SummerOlympi			Rugby	Rugby,Men	31505		· Committee of the comm	1	Gold	Tr
908SummerOlymp			Rugby	Rugby,Men	31505	TomGriffin	11242	1	Gold	Tr
908SummerOlymp			Rugby	Rugby,Men	31505	JackHickey	11243	1	Gold	Tr
908SummerOlymp				Rugby,Men		MannieMcArthur		1	Gold	Tr
908SummerOlympi	.cs	ANZ	Rugby	Rugby,Men	31505	ArthurMcCabe	11245	1	Gold	Tr

Записуємо в кафка-топік та зчитуємо дані по результатам змагань з перетворенням даних у датафрейм.

atch: 2							
spont.		cov	country_noc	boight	- ra voi abt	+ 	+imactam
sport	medal	sex	country_noc	avg_neight	avg_weight		timestam
Athletics	nan	Male	CAF	181.8	72.4	 2024-12-21	04.55.
Athletics		Male	NULL	183.2		2024-12-21	
IceHockey		Female	NULL	167.0		2024-12-21	
Judo		Male	ALB			2024-12-21	
EquestrianEventing		Female	NULL			2024-12-21	
Shooting		Male	NULL			2024-12-21	
Taekwondo		Male	ARM			2024-12-21	
ArtisticGymnastics		Female	BOL	150.0		2024-12-21	
FreestyleSkiing		Female	NULL	165.4		2024-12-21	
CrossCountrySkiing		Male	ARM			2024-12-21	
CyclingTrack			NULL			2024-12-21	
CyclingRoad			BER			2024-12-21	
Swimming			CMR	179.5		2024-12-21	
Athletics		Female	ALB			2024-12-21	
Shooting		Male	BLR		97.0	2024-12-21	04:55:
Swimming		Male	DJI	180.0	70.0	2024-12-21	04:55:
Shooting	Bronze	Male	BLR	175.3	75.7	2024-12-21	04:55:
CanoeSlalom			NULL	177.8	73.1	2024-12-21	04:55:
Rowing	nan	Female	NULL	176.4	83.6	2024-12-21	04:55:
						2024 42 24	04 55
Swimming nly showing top 20 atch: 3	nan	Female	BEN 	167.0	57.0	2024-12-21 +	
Swimming	nan	Female	BEN	167.0	57.0		
Swimming	nan rows		BEN			+	timestam
Swimming	nan rows	sex	country_noc	avg_height	avg_weight		timestam
Swimming nly showing top 20 atch: 3 sport Athletics	rows rows medal	sex Male	country_noc	avg_height	avg_weight	+	 timestam
Swimming	rows medal silver nan	sex Male	country_noc NULL	avg_height 183.2 167.0	avg_weight 78.2 63.8	+	timestam 04:55:
Swimming	rows rows medal Silver nan nan	sex Male Female	country_noc NULL NULL GDR	avg_height 183.2 167.0 164.0	avg_weight 78.2 63.8 60.5	+	timestam 04:55: 04:55:
Swimming	rows rows medal Silver nan nan	sex Male Female Female	country_noc NULL NULL GDR NULL	avg_height 183.2 167.0 164.0	avg_weight 78.2 63.8 60.5 95.4	+	timestam 04:55: 04:55: 04:55:
Swimming	rows rows medal Silver nan nan silver	sex Male Female Female Male	country_noc NULL NULL GDR NULL	avg_height 183.2 167.0 164.0 166.9	avg_weight 78.2 63.8 60.5 95.4	+	timestam 04:55: 04:55: 04:55: 04:55:
Swimming nly showing top 20 atch: 3 sport Athletics IceHockey CanoeSlalom EquestrianEventing Badminton Taekwondo	nan rows medal silver nan nan Silver	sex Male Female Female	country_noc NULL NULL GDR NULL INA	avg_height 183.2 167.0 164.0 166.9 174.2	78.2 63.8 60.5 95.4 70.7	+	timestam
Swimming	rows rows medal silver nan nan silver nan nan nan	sex Male Female Female Male Male	country_noc NULL NULL GDR NULL	avg_height 183.2 167.0 164.0 166.9 174.2	78.2 63.8 60.5 95.4 70.7 80.0	+	timestam 04:55: 04:55: 04:55: 04:55: 04:55: 04:55:
Swimming nly showing top 20 atch: 3 sport Athletics IceHockey CanoeSlalom EquestrianEventing Badminton Taekwondo	rows rows medal silver nan silver nan nan nan nan	sex Male Female Female Male Male	country_noc NULL NULL GDR NULL INA ARM BOL	avg_height 183.2 167.0 164.0 166.9 174.2 185.0	avg_weight 78.2 63.8 60.5 95.4 70.7 80.0 40.0	+	timestam
Swimming	rows rows medal Silver nan silver nan nan nan nan	sex Male Female Female Male Male Male Male	country_noc NULL NULL GDR NULL INA ARM BOL	avg_height 183.2 167.0 164.0 166.9 174.2 185.0 150.0	avg_weight 78.2 63.8 60.5 95.4 70.7 80.0 40.0 65.9	+	timestam
Swimming	rows rows medal silver nan nan silver nan nan nan nan nan nan	sex Male Female Female Male Male Male Female	country_noc NULL NULL GDR NULL INA ARM BOL EGY NULL BEL	avg_height 183.2 167.0 164.0 166.9 174.2 185.0 150.0 165.8 165.4	avg_weight 78.2 63.8 60.5 95.4 70.7 80.0 40.0 65.9 59.0	+	timestam
Swimming	rows rows medal silver nan nan silver nan nan nan nan nan nan	sex Male Female Female Male Male Male Female Male Male Female	country_noc NULL NULL GDR NULL INA ARM BOL EGY NULL BEL	avg_height 183.2 167.0 164.0 166.9 174.2 185.0 150.0 165.8 165.4 169.6	avg_weight 78.2 63.8 60.5 95.4 70.7 80.0 40.0 65.9 59.0 62.0	+	timestam
Swimming	rows rows medal Silver nan nan silver nan nan nan nan nan nan nan n	sex Female Female Male Male Female Male Female Male Female	country_noc NULL NULL GDR NULL INA ARM BOL EGY NULL BEL ARM NULL	avg_height 183.2 167.0 166.9 174.2 185.0 150.0 165.8 165.4 169.6	avg_weight 78.2 63.8 60.5 95.4 70.7 80.0 40.0 65.9 59.0 62.0 67.1	+	timestam
Swimming	rows rows medal Silver nan nan nan nan nan nan nan n	sex Female Female Male Male Male Female Male Female Male Male Male	country_noc NULL NULL GDR NULL INA ARM BOL EGY NULL BEL ARM NULL CAF	avg_height 183.2 167.0 164.0 166.9 174.2 185.0 150.0 165.8 165.4 169.6 170.5	avg_weight 78.2 63.8 60.5 95.4 70.7 80.0 40.0 65.9 59.0 62.0 67.1 135.9	+	timestam 04:55: 04:55: 04:55: 04:55: 04:55: 04:55: 04:55: 04:55: 04:55:
Swimming	rows rows medal Silver nan nan nan nan nan nan nan n	sex Female Female Female Male Male Female Female Male Male Male Male	country_noc NULL NULL GDR NULL INA ARM BOL EGY NULL BEL ARM NULL CAF	avg_height 183.2 167.0 164.0 166.9 174.2 185.0 150.0 165.8 165.4 169.6 170.5	avg_weight 78.2 63.8 60.5 95.4 70.7 80.0 40.0 65.9 59.0 62.0 67.1 135.9 72.4 81.0	+	timestam
Swimming	rows rows medal silver nan nan nan nan nan nan nan nan nan	sex Female Female Female Male Male Female Female Male Male Male Male	country_noc NULL NULL GDR NULL INA ARM BOL EGY NULL ARM NULL CAF ALB	avg_height 183.2 167.0 164.0 166.9 174.2 185.0 165.8 165.4 169.6 170.5 178.6 181.8	avg_weight 78.2 63.8 60.5 95.4 70.7 80.0 65.9 59.0 62.0 67.1 135.9 72.4 81.0 80.3	+	timestam
Swimming	rows rows medal medal nan nan nan nan nan nan nan	sex Female Female Male Male Male Male Male Male Male M	country_noc NULL NULL GDR NULL INA ARM BOL BEL ARM NULL CAF ALB NULL CZE GDR	avg_height	avg_weight 78.2 63.8 60.5 95.4 70.7 80.0 65.9 62.0 67.1 135.9 72.4 81.0 80.3 63.2	+	timestam
Swimming	rows rows medal medal nan nan silver nan nan nan nan nan nan silver nan nan nan silver nan silver silver silver silver	sex Female Female Male Male Male Male Male Male Male M	country_noc NULL NULL GDR NULL INA ARM BOL BEL ARM NULL CAF ALB NULL CZE GDR	avg_height 183.2 167.0 164.0 166.9 174.2 185.0 150.0 165.8 169.6 170.5 178.6 181.8 175.0 176.3	avg_weight 78.2 63.8 60.5 95.4 70.7 80.0 40.0 65.9 59.0 67.1 135.9 72.4 81.0 80.3 63.2 80.6 72.4	+	timestam
Swimming	rows rows medal medal nan nan nan nan nan nan nan	sex Male Female Female Male Male Female Male Male Male Male Male Male Male M	country_noc NULL NULL GDR NULL INA ARM BOL BEL ARM NULL CAF ALB NULL CZE GDR	avg_height 183.2 167.0 164.0 166.9 174.2 185.0 150.0 165.8 169.6 170.5 178.6 181.8 175.0 176.3 172.0	avg_weight 78.2 63.8 60.5 95.4 70.7 80.0 40.0 65.9 59.0 67.1 135.9 72.4 81.0 80.3 63.2 80.6 72.4 72.7	+	timestam

Результат потокового обчислення середніх ваги та росту атлетів для кожного виду спорту, типу медалі або її відсутності, статі, країни.

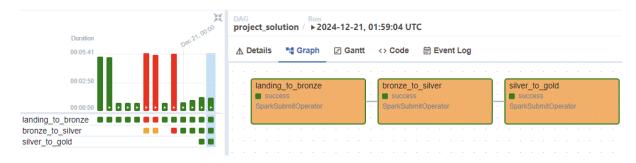
А ось п.6 з першої частини завдання я не зміг побороти.

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

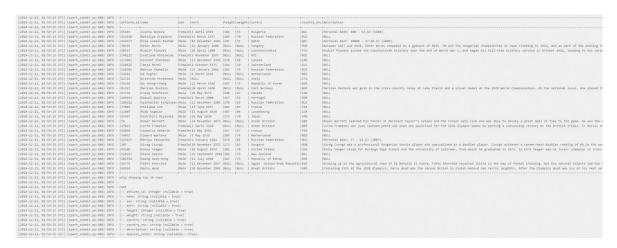
landing_to_bronze.py:

```
[2024-12-21, 01:59:08 UTC] {spark_submit.py:488} INFO - Script landing_to_bronze started...
[2024-12-21, 01:59:08 UTC] {spark_submit.py:488} INFO - Downloading from https://ftp.goit.study/neoversity/athlete_bio.csv
[2024-12-21, 01:59:12 UTC] {spark_submit.py:488} INFO - File downloaded successfully and saved as dags/vvd/athlete_bio.csv
[2024-12-21, 01:59:12 UTC] {spark_submit.py:488} INFO - Downloading from https://ftp.goit.study/neoversity/athlete_event_results.csv
[2024-12-21, 01:59:14 UTC] {spark_submit.py:488} INFO - File downloaded successfully and saved as dags/vvd/athlete_event_results.csv
```

Зчитуємо таблиці



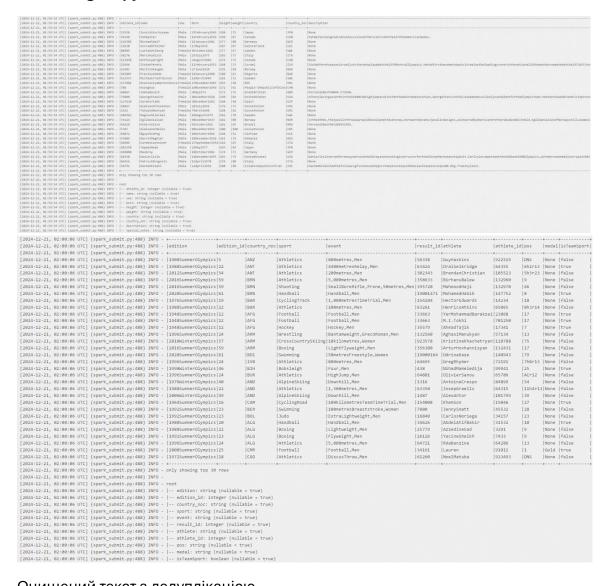
Скрін відпрацьованого DAG-файлу



[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	+					+	+	+	-+	+	+	+
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	edition	le.	edition_id cou	ntry_noc spo	ort	event	resu	lt_id athlete	athlete_	id pos	medal	isTeamSpor
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	+					+	+	+	-+	+	+	+
		{spark_submit.py:488}					Ath	letics	100 metres, Men	5626	5 Ernest Hutched	n 64710	DNS	NULL	false
		{spark_submit.py:488}							400 metres, Men	5631	3 Henry Murray	64756	DNS	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Atl	letics	800 metres, Men	5633	8 Harvey Sutton	64808	3 h8 r1/2	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	800 metres, Men	5633	8 Guy Haskins	922519	DNS	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	800 metres, Men	5633	8 Joseph Lynch	64735	IDNS	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	800 metres, Men	5633	8 Henry Murray	64756	DNS	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Atl	letics	1,500 metres, Men	5634	9 Joseph Lynch	64735	5 h2 r1/2	NULL	[false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	1,500 metres, Men	5634	9 Charles Swain	79576	AC h3 r1/	2 NULL	[false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	1,500 metres, Men	5634	9 Guy Haskins	922519	DNS	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	1,500 metres, Men	5634	9 George Blake	64619	DNS	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	5 miles, Men	5636	0 George Blake	64619	3 h1 r1/2	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	5 miles, Men	5636	8 Joseph Lynch	64735	AC h5 r1/	2 NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	S miles, Men	5636	0 Guy Haskins	922519	DNS	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	5 miles, Men	5636	0 Victor Aitken	64594	DNS	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	Marathon, Men	5636	9 Victor Aitken	64594	DNF	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	Marathon, Men	5636	9 George Blake	64619	DNF	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	Marathon, Men	5636	9 Joseph Lynch	64735	DNF	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	110 metres Hurdles, Men	5637	8 Henry Murray	64756	2 h1 r1/3	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	400 metres Hurdles, Men	5639	2 Henry Murray	64756	2 h3 r1/3	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	3,200 metres Steeplechase	, Men 5641	2 Guy Haskins	922519	DNS	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	3,500 metres Race Walk, M	en 5642	1 Harry Kerr	64719	3	Bronze	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	3,500 metres Race Walk, Me	en 5642	1 Albert Rowland	64788	15	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	10 miles Race Walk, Men	5642	7 Harry Kerr	64719	DNS r2/2	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	10 miles Race Walk, Men	5642	7 Albert Rowland	64788	5 h1 r1/2	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics S	5 ANZ	Ath	letics	10 miles Race Walk, Men	5642	7 Victor Aitken	64594	DNS	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	Standing High Jump, Men	5645	4 Ernest Hutched	n 64710	AC	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	Long Jump, Men	5646	7 Ernest Hutched	n 64710	DNS	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Ath	letics	Standing Long Jump, Men	5647	5 Ernest Hutched	n 64718	DNS	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Atl	letics	Triple Jump, Men	5648	2 Ernest Hutched	n 64710	DNS	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	1908 Summer	Olympics 5	5 ANZ	Box	ing	Lightweight, Men	2124	7 George Blake	64619	DNS	NULL	false
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	+					+	+	+	-+	+	+	+
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	only showing	top 30 row	WS									
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -												
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	root											
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	edition:	string (no	ullable = true)								
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	edition_i	ld: integer	r (nullable = 1	true)								
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	country_n	oc: string	g (nullable = 1	true)								
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	sport: st	ring (nul)	lable = true)									
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	event: st	ring (null	lable = true)									
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	result_id	: integer	(nullable = to	rue)								
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	athlete:	string (nu	ullable = true)								
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	athlete_i	d: integer	r (nullable = 1	true)								
[2024-12-21,	01:59:29 UTC]	{spark_submit.py:488}	INFO -	pos: stri	ing (nullab	ble = true)									
		{spark_submit.py:488}													
	01:50:20 UTC1	{spark submit.py:488}	INFO -	isTeamSpo	ort: booles	an (nullable =	true)								

Після запису у форматі parquet робимо виводи в консоль по 30 рядків таблиць та їхні структури.

silver_to_gold.py:



Очищений текст з дедуплікацією.

silver_to_gold.py:

[2024-12-21, 02:00:21 UTC]	(snark submit nv:488)	TNEO -	+	+	+	4	+	+		
[2024-12-21, 02:00:21 UTC]				medal		bio_country_noc	lave weight	avg_height	timestamp	
[2024-12-21, 02:00:21 UTC]				1=====	, Jan	1020_0000007_000	1	10.97	Lamestonp	
[2024-12-21, 02:00:21 UTC]				Silver	[Male	INOR	87.6	185.8	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]				None	010000			172.68421052631578		
[2024-12-21, 02:00:21 UTC]					Female		57.41353383458647			02:00:15.438191
[2024-12-21, 02:00:21 UTC]								189.677083333333334		
[2024-12-21, 02:00:21 UTC]				None				180.8740157480315		
[2024-12-21, 02:00:21 UTC]				None	Male	ISWE	78.64516129032258	182.7741935483871	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]				Silver	 Female	IJPN	12569.0	1160.6	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]				None						02:00:15.438191
[2024-12-21, 02:00:21 UTC]				None	Male	Isvk	90.68553459119497	185.62264150943398	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]				None	Male	IGDR	78.63934426229508	178.0	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]					[Male			172.78947368421052		
[2024-12-21, 02:00:21 UTC]				None	Female			168.1046511627907		
[2024-12-21, 02:00:21 UTC]				None	Male	IRUS	179.36	179.38	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]				Silver			70.6666666666667	174.1666666666666		
[2024-12-21, 02:00:21 UTC]				None						02:00:15.438191
[2024-12-21, 02:00:21 UTC]				None	Male			178.86153846153846	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]				None	Female	EST	61.5	166.875	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]	{spark submit.py:488}	INFO -	Archery	None	Male	IGRE	72.0	178.33333333333334	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]				None	Male	KEN	96.0	186.0	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]	{spark submit.py:488}	INFO -	Sailing	None	Male	BER	81.47619047619048	179.04761904761904	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]	{spark_submit.py:488}	INFO -	FreestyleSkiing	Bronze	Female	NOR	62.0	169.0	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]	{spark_submit.py:488}	INFO -	CrossCountrySkiing	None	Female	NOR	59.21621621621622	170.18243243243242	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]	{spark submit.py:488}	INFO -	Volleyball	None	Female	PER	65.4375	173.25	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]	{spark_submit.py:488}	INFO -	Luge	Silver	Female	GDR	72.4	172.4	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]	{spark_submit.py:488}	INFO -	Sailing	Silver	Male	GDR	82.83333333333333	178.6666666666666	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]	{spark submit.py:488}	INFO -	Archery	None	Male	NZL	76.25	178.25	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]	{spark_submit.py:488}	INFO -	Athletics	None	Female	MLT	55.36363636363637	163.0	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]	{spark_submit.py:488}	INFO -	FreestyleSkiing	None	Male	NOR	79.83333333333333	181.16666666666666	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]	{spark_submit.py:488}	INFO -	Snowboarding	None	Female	CHN	60.05	165.7	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]	{spark_submit.py:488}	INFO -	CanoeSprint	Gold	Male	NZL	84.3	178.9	2024-12-21	02:00:15.438191
[2024-12-21, 02:00:21 UTC]	{spark_submit.py:488}	INFO -	+	+	+	+	+	+	+	
[2024-12-21, 02:00:21 UTC]	{spark_submit.py:488}	INFO -	only showing top 30	rows						
[2024-12-21, 02:00:21 UTC]	{spark submit.py:488}	INFO -								
[2024-12-21, 02:00:21 UTC]	{spark_submit.py:488}	INFO -	root							
[2024-12-21, 02:00:21 UTC]	{spark_submit.py:488}	INFO -	sport: string (nullabl	e = tru	e)				
[2024-12-21, 02:00:21 UTC]	{spark_submit.py:488}	INFO -	medal: string (nullabl	e = tru	e)				
[2024-12-21, 02:00:21 UTC]	{spark_submit.py:488}	INFO -	sex: string (nu	llable	= true)					
[2024-12-21, 02:00:21 UTC]	{spark_submit.py:488}	INFO -	bio_country_noc	: strin	g (null	able = true)				
[2024-12-21, 02:00:21 UTC]	{spark_submit.py:488}	INFO -	avg_weight: dou	ble (nu	llable	= true)				
[2024-12-21, 02:00:21 UTC]	{spark_submit.py:488}	INFO -	avg_height: dou	ble (nu	llable	= true)				
[2024-12-21, 02:00:21 UTC]	{spark_submit.py:488}	INFO -	timestamp: time	stamp (nullabl	e = false)				

Цільовий файл. У структурі видно типи стовпчиків (переведено y double)