Homework 2: Hive

Studenti: Antimo Barbato
Matricola: M63/1079

1: Introduzione Hive

Hive è un data warehouse costruito sopra Hadoop, progettato per l'archiviazione e l'elaborazione di grandi volumi di dati. Permette di gestire e analizzare i dati in Hadoop usando un linguaggio simile a SQL. Hive è utilizzato per eseguire query sui dati archiviati in **HDFS** (Hadoop Distributed File System).
Hive è molto utile per analizzare enormi dataset attraverso un'architettura distribuita, sfruttando la potenza di calcolo di Hadoop, ma rende l'interazione con i dati molto più semplice grazie alla sua sintassi simile a SQL.

2: Dataset – Videogames Sales

Il dataset "Video Games Sales" (formato csv) offre una panoramica dettagliata delle vendite globali di videogiochi, includendo informazioni come il titolo del gioco, la piattaforma, il genere, l'editore, le vendite regionali (Nord America, Europa, Giappone e Resto del Mondo), le vendite globali e le recensioni. Questo dataset è utile per analizzare le tendenze del mercato dei videogiochi, confrontare le performance degli editori e studiare le preferenze dei giocatori in diverse regioni.
Vediamo nel dettaglio gli attributi relativi al dataset

- **Rank**: La classifica del videogioco basata sul volume di vendite globali.
- **Game Title**: Il nome del videogioco.
- **Platform**: La piattaforma su cui è disponibile il gioco, come PC, PS4, Xbox One, ecc.
- **Year**: L'anno in cui il gioco è stato rilasciato.
- **Genre**: Il genere del gioco, come azione, avventura, corse, ecc.
- **Publisher**: La società che ha pubblicato il gioco.
- **North America**: Il numero di unità vendute in Nord America, in milioni.
- **Europe**: Il numero di unità vendute in Europa, in milioni.
- **Japan**: Il numero di unità vendute in Giappone, in milioni.
- **Rest of World**: Il numero di unità vendute nel resto del mondo, esclusi Nord America, Europa e Giappone, in milioni.
- **Global**: Il numero totale di unità vendute in tutto il mondo, in milioni.
- **Review**: Il punteggio delle recensioni del gioco, su una scala da 1 a 10.

2.1: Piattaforma di analisi dei dati: Databricks

La piattaforma utilizzata per analizzare il dataset è Databricks.
Databricks è una piattaforma di analisi dei dati basata su cloud che facilita l'elaborazione e l'analisi di grandi volumi di dati. Fondata dai creatori di Apache Spark, Databricks offre un ambiente integrato per la gestione dei dati, l'analisi e il machine learning.

Caratteristiche principali di Databricks:

- **Integrazione con Apache Spark**: Databricks sfrutta la potenza di Apache Spark per l'elaborazione distribuita dei dati, consentendo analisi rapide e scalabili.
- **Notebook Collaborativi**: Gli utenti possono creare e condividere notebook interattivi per scrivere codice, eseguire query e visualizzare i risultati in tempo reale.
- **Gestione dei Cluster**: Databricks semplifica la creazione e la gestione dei cluster, permettendo di scalare facilmente le risorse computazionali in base alle esigenze.
- **Supporto Multi-Linguaggio**: La piattaforma supporta diversi linguaggi di programmazione, tra cui Python, SQL, Scala e R, rendendola versatile per vari tipi di analisi.

Nello specifico per analizzare il dataset, si è anzitutto creato un cluster.

Compute

antimo barbato's Cluster

Configurazione | Notebook (1) | Librerie | Log eventi | UI Spark | Log driver | Metriche | App | UI compute Spark - Master

Versione Runtime Databricks

12.2 LTS (includes Apache Spark 3.3.2, Scala 2.12)

Tipo di driver

Community Optimized — Memoria di 15,3 GB e 2 Cores

Istanza

Memoria da 15 GB gratuita: in qualità di utente della Community Edition, il compute terminerà automaticamente dopo un periodo di inattività di una o due ore.
Per maggiori opzioni di configurazione, esegui l'upgrade dell'abbonamento a Databricks.

Spark | JDBC/ODBC

Configurazione Spark

spark.databricks.rocksDB.fileManager.useCommitService false

Variabili di ambiente

PYSPARK_PYTHON=/databricks/python3/bin/python3

Successivamente si è caricato il dataset

default.videogames_2_csv — Aggiorna

antimo barbato's Cluster

Schema:

	col_name	data_type	comment
1	index	int	null
2	Rank	int	null
3	Game Title	string	null
4	Platform	string	null
5	Year	double	null
6	Genre	string	null
7	Publisher	string	null
8	North America	double	null
9	Europe	double	null
10	Japan	double	null
11	Rest of World	double	null
12	Global	double	null
13	Review	double	null

Dati campione:

	index	Rank	Game Title	Platform	Year	Genre	Publisher	North America	Europe	Japan
1	0	1	Wii Sports	Wii	2006	Sports	Nintendo	40.43	28.39	3.77
2	1	2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81
3	2	3	Mario Kart Wii	Wii	2008	Racing	Nintendo	14.5	12.22	3.63
4	3	4	Wii Sports Resort	Wii	2009	Sports	Nintendo	14.82	10.51	3.18
5	4	5	Tetris	GB	1989	Puzzle	Nintendo	23.2	2.26	4.22
6	5	6	New Super Mario Bros.	DS	2006	Platform	Nintendo	10.85	8.87	6.48
7	6	7	Wii Play	Wii	2006	Misc	Nintendo	13.83	9.11	2.93
8	7	8	Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28
9	8	9	New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	13.35	6.48	4.66
10	9	10	Nintendogs	DS	2005	Simulation	Nintendo	9.02	10.81	1.93
11	10	11	Pokémon Gold / Silver Version	GB	1999	Role-Playing	Nintendo	9	6.18	7.2

2.3: Query con Hive

Non è stata effettuata alcuna operazione di pre-processing dei dati, si è preferito filtrare i dati escludendo i valori nulli con ''IS NOT NULL''.

	col_name	data_type
1	index	int
2	Rank	int
3	Game Title	string
4	Platform	string
5	Year	double
6	Genre	string
7	Publisher	string
8	North America	double
9	Europe	double
10	Japan	double
11	Rest of World	double
12	Global	double
13	Review	double

FIGURA 1: TABELLA OTTENUTA CON DESCRIBE

Query 1 : Conteggio dei giochi con lo stesso Game Title e ordinati in ordine decrescente

```
SELECT `Game Title` AS game_name, COUNT(`Game Title`) AS game_count
FROM videogames_3_csv
WHERE `Game Title` IS NOT NULL
GROUP BY `Game Title`
ORDER BY game_count DESC;
```

	game_name	game_count
1	FIFA Soccer 08	6
2	LEGO Indiana Jones: The Original Adventures	6
3	LEGO Batman: The Videogame	6
4	The Simpsons Game	5
5	Star Wars: The Force Unleashed	5
6	WWE SmackDown vs Raw 2008	5
7	Pro Evolution Soccer 2008	5
8	FIFA Soccer 10	5
9	LEGO Star Wars: The Complete Saga	4
10	Madden NFL 08	4
11	The Sims 3	4
12	LEGO Star Wars: The Video Game	4
13	Madden NFL 07	4
14	Need for Speed Carbon	4
15	Spider-Man 2	4

Query 2: Media vendite globali ordinate in ordine decrescente

```
SELECT Genre, AVG(GLOBAL) AS AVG_Globale
FROM videogames_3_csv
WHERE GLOBAL IS NOT NULL AND Genre IS NOT NULL
GROUP BY Genre
```

```
ORDER BY AVG_Globale DESC;
```

	Genre	AVG_Globale
1	Platform	3.166170212765958
2	Role-Playing	2.868208092485551
3	Shooter	2.7048543689320415
4	Misc	2.68
5	Puzzle	2.469318181818182
6	Racing	2.429032258064516
7	Action	2.317345454545455
8	Sports	2.282824675324676
9	Simulation	2.2297826086956514
10	Adventure	2.1768181818181813
11	Fighting	1.9761904761904763
12	Strategy	1.9449999999999998

Concentriamoci sui primi 3 per la prossima query per vedere quale Publisher ne ha prodotti maggiormente

Query 3: Filtriamo in base ai 3 generi, raggruppiamo in base al publisher e contiamo il numero di giochi prodotti per genere con relativo ordinamento decrescente

```
SELECT Genre, Publisher, COUNT(*) as Genre_game_count
FROM videogames_3_csv
WHERE Genre IN ('Platform' , 'Role-Playing', 'Shooter')
      AND
      Genre IS NOT NULL
      AND
      Publisher IS NOT NULL
GROUP BY Publisher, Genre
ORDER BY Genre_game_count DESC
```

	Genre	Publisher	Genre_game_count
1	Platform	Nintendo	68
2	Shooter	Electronic Arts	44
3	Role-Playing	Nintendo	38
4	Shooter	Activision	33
5	Platform	Sega	30
6	Role-Playing	Square Enix	24
7	Shooter	Ubisoft	22
8	Platform	Sony Computer Entertainme...	21
9	Role-Playing	Square	16
10	Platform	THQ	16

Continuiamo analizzando a questo punto le vendite globali di Nintendo per il genere Platform

Query 4: Filtro per il genere “Platform” e Publisher “Nintendo.
Calcolo la media delle vendite globali per il publisher Nintendo

```
SELECT Publisher, Genre, AVG(Global) AS AVG_Vendite_Globali
FROM videogames_3_csv
WHERE Publisher = 'Nintendo'
    AND
    Genre = 'Platform'
    AND
    Global IS NOT NULL
    AND
    Publisher IS NOT NULL
    AND
    Genre IS NOT NULL
GROUP BY Publisher, Genre;
```

	Publisher	Genre	AVG_Vendite_Globali
1	Nintendo	Platform	5.671323529411764

Successivamente calcolo la media delle vendite annue di nintendo

Query 5: Filtro per il Publisher “Nintendo” e calcolo la media di vendite globali annue di Nintendo

```
SELECT Publisher, Year, AVG(Global) AS AVG_Vendite_Globali_Annue
FROM videogames_3_csv
WHERE Publisher = 'Nintendo'
      AND
      Global IS NOT NULL
GROUP BY Year, Publisher
ORDER BY AVG_Vendite_Globali_Annue DESC
```

	Publisher	Year	AVG_Vendite_Globali_Annue
1	Nintendo	2006	11.232352941176472
2	Nintendo	1989	9.833333333333334
3	Nintendo	1985	9.704
4	Nintendo	2009	8.651428571428571
5	Nintendo	1988	6.860000000000001
6	Nintendo	2005	6.820588235294117
7	Nintendo	1984	6.184285714285715
8	Nintendo	2008	6.159230769230769
9	Nintendo	1999	5.602727272727273
10	Nintendo	2007	5.6

Query 6: Filtro in base a Review e Global per ottenere i giochi con valutazione maggiore di 8 e con vendite superiori a 20 milioni

```
SELECT `Game Title`, Review, Global
FROM videogames_3_csv
WHERE Review > 8
      AND
      Global > 20
      AND
      Review IS NOT NULL
      AND
      Global IS NOT NULL
ORDER BY Global DESC
```

	Game Title	1.2 Review	1.2 Global
1	Wii Sports	76.28	81.12
2	Super Mario Bros.	91	40.24
3	Mario Kart Wii	82.07	33.55
4	Wii Sports Resort	82.65	31.52
5	Tetris	88	30.26
6	New Super Mario Bros.	90	29.08
7	Wii Play	61.64	28.71
8	Duck Hunt	84	28.31
9	New Super Mario Bros. Wii	88.18	26.75
10	Nintendogs	85	24.5
11	Pokémon Gold / Silver Version	89	23.1
12	Wii Fit	81.2	22.74
13	Mario Kart DS	91.34	22.47
14	Wii Fit Plus	80.83	21.15
15	Grand Theft Auto: San Andreas	95.08	20.81

A seguito di tale risultato si è deciso di inserire nella select anche l'attributo relativo all'anno, per verificare il primo gioco a che anno fa riferimento, considerando che nintendo ha le migliori vendite medie annue nel 2006

```
SELECT `Game Title`, Year, Review, Global
FROM videogames_3_csv
WHERE Review > 8
      AND
      Global > 20
      AND
      Review IS NOT NULL
      AND
      Global IS NOT NULL
ORDER BY Global DESC
LIMIT 10
```

Come si può notare nella top 10 rientrano 3 giochi Nintendo con un numero di vendite globali pari a 81 milioni, 29 milioni e 28 milioni per quanto riguarda l'anno 2006.
A questo punto può essere interessante vedete tutti i giochi venduti da nintendo nel 2006.

	Game Title	1.2 Year	1.2 Review	1.2 Global
1	Wii Sports	2006	76.28	81.12
2	Super Mario Bros.	1985	91	40.24
3	Mario Kart Wii	2008	82.07	33.55
4	Wii Sports Resort	2009	82.65	31.52
5	Tetris	1989	88	30.26
6	New Super Mario Bros.	2006	90	29.08
7	Wii Play	2006	61.64	28.71
8	Duck Hunt	1984	84	28.31
9	New Super Mario Bros. ...	2009	88.18	26.75
10	Nintendogs	2005	85	24.5

Query 7: Filtro per Publisher e Anno.
Ordino in base alle vendite globali per ottenere una lista dei giochi più venduti di nintendo nel 2006

```
SELECT `Game Title`, Publisher, Year, Global
FROM videogames_3_csv
WHERE Publisher = "Nintendo"
      AND
      Year = "2006"
ORDER BY Global DESC
```

	Game Title	Publisher	1.2 Year	1.2 Global
1	Wii Sports	Nintendo	2006	81.12
2	New Super Mario Bros.	Nintendo	2006	29.08
3	Wii Play	Nintendo	2006	28.71
4	Pokémon Diamond / Pearl Version	Nintendo	2006	18.05
5	The Legend of Zelda: Twilight Princess	Nintendo	2006	6.76
6	Clubhouse Games	Nintendo	2006	3.35
7	English Training: Have Fun Improving Your Skills!	Nintendo	2006	3.33
8	Personal Trainer: Cooking	Nintendo	2006	3.09
9	WarioWare: Smooth Moves	Nintendo	2006	2.84
10	Yoshi's Island DS	Nintendo	2006	2.69
11	Pokémon Ranger	Nintendo	2006	2.11
12	Tetris DS	Nintendo	2006	2.07
13	Kirby Squeak Squad	Nintendo	2006	1.91
14	The Legend of Zelda: Twilight Princess	Nintendo	2006	1.59
15	Mario Hoops 3 on 3	Nintendo	2006	1.57

Query 8: Raggruppo in base a platform e year e effettuo il conteggio per ottenere il numero di giochi in base all'anno.

```
SELECT Platform, Year, COUNT(*) AS Giochi_Annui_Piattaforma
FROM videogames_3_csv
GROUP BY Platform, Year
ORDER BY Giochi_Annui_Piattaforma DESC
```

	Platform	1.2 Year	Giochi_Annui_Piattaforma
1	PS2	2002	66
2	PS2	2004	64
3	PS2	2003	59
4	PS	1998	55
5	PS2	2005	48
6	PS	1999	45
7	PS2	2001	43
8	PS3	2008	43
9	X360	2010	43
10	X360	2008	42

Query 9: Raggruppo per Publisher e effettuo la somma relativa alle vendite sia regionali che globali e ordino in base alle globali

```
SELECT
    Publisher,
    SUM(`North America`) AS Vendite_NA,
    SUM(Europe) AS Vendite_EU,
    SUM(Japan) AS Vendite_JP,
    SUM(`Rest of World`) AS Vendite_RW,
    SUM(Global) AS Vendite_Globali
FROM videogames_3_csv
GROUP BY Publisher
ORDER BY Vendite_Globali DESC;
```

	Publisher	1.2 Vendite_NA	1.2 Vendite_EU	1.2 Vendite_JP	1.2 Vendite_RW	1.2 Vendite_Globali
1	Nintendo	688.47	341.80000000000047	338.0399999999997	80.57999999999998	1448.8400000000013
2	Electronic Arts	352.3499999999997	204.00000000000009	8.729999999999984	68.36000000000008	633.3600000000001
3	Sony Computer Entertainment	167.58999999999992	119.19	50.27000000000001	40.51000000000004	377.60999999999996
4	Activision	219.20999999999995	113.52000000000002	3.6599999999999953	35.070000000000014	371.41999999999996
5	Take-Two Interactive	112.63999999999999	67.43	3.8199999999999963	24.499999999999996	208.4199999999999
6	Ubisoft	103.22999999999996	69.61	1.790000000000001	21.69000000000001	196.31999999999994
7	Microsoft Game Studios	110.68000000000002	43.220000000000006	2.3700000000000006	13.369999999999992	169.73000000000005
8	THQ	80.19000000000004	47.00000000000002	2.5799999999999987	13.269999999999996	142.98
9	Sega	60.65000000000001	39.820000000000014	11.889999999999999	10.309999999999993	122.67
10	Capcom	49.41999999999999	24.310000000000013	32.91000000000001	7.76	114.33000000000001
11	Konami Digital Entertainment	30.41	43.57	21.710000000000008	12.009999999999998	107.66999999999999
12	Namco Bandai Games	30.279999999999994	17.040000000000003	18.750000000000004	5.599999999999997	71.68999999999998
13	Square Enix	21.930000000000003	12.019999999999998	25.559999999999995	5.08	64.59
14	LucasArts	35.17999999999999	19.100000000000005	0.2000000000000000...	6.649999999999999	61.10999999999999
15	Eidos Interactive	28.86	20.46	2.7799999999999994	4.16	56.249999999999986