



# Twitch Data Analysis

Polosa Sebastiano e Ungaro Riccardo



# Contesto

**Twitch** è una piattaforma di **video live streaming** incentrata principalmente nello streaming di videogiochi.

- 8'000'000 canali attivi;
- 2'500'000 spettatori attivi;
- 2'000'000 ore di video guardate ad agosto 2021;
- 100'000 live al giorno in media.

Possibili **benefici** di analizzare i dati di questa piattaforma:

- Creare sistemi di raccomandazione per gli spettatori;
- Merchandising;
- Investire su prodotti in tendenza;
- Fornire sistemi evoluti di analisi per supportare gli streamer nella creazione dei contenuti;

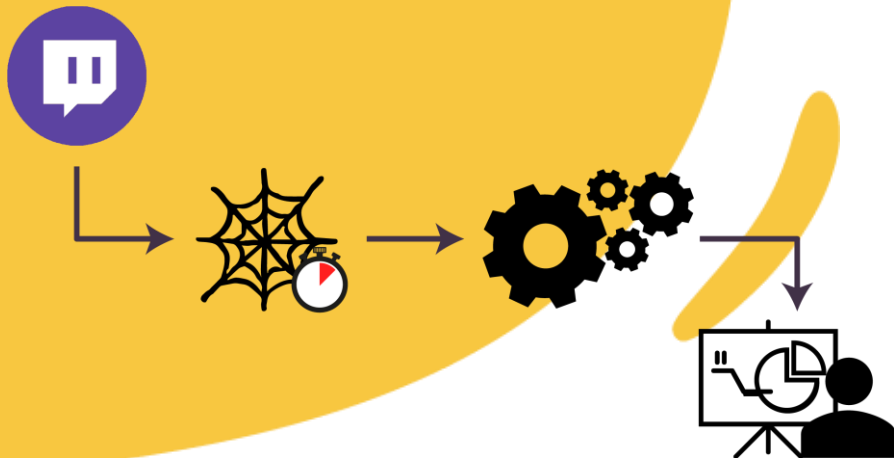


# Dataset

Il **dataset** utilizzato per il testing durante l'implementazione del progetto è formato da due archivi:

- **Broadcaster List**
- **Twitch Dataset**

Questi archivi sono stati creati tramite il recupero delle informazioni attraverso un **crawler** eseguito sulla piattaforma Twitch.



## Broadcaster List

broadcaster ID
----------------

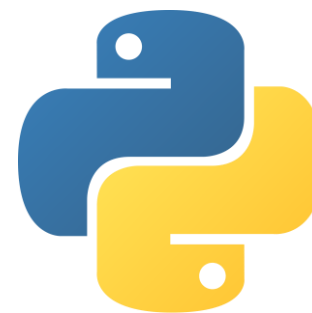
## Twitch Dataset

stream ID
current views
stream created time
game name
broadcaster ID
broadcaster name
delay settings
follower number
partner status
broadcaster language
total views broadcaster
language
bradcaster's created time
playback bitrate
source resolution

# Strumenti Utilizzati

Per la realizzazione del progetto sono stati utilizzati i seguenti strumenti:

- Python
- Kafka
- MongoDB
- Spark
  - Spark Streaming
  - Spark SQL



**Spark**  
*Streaming*

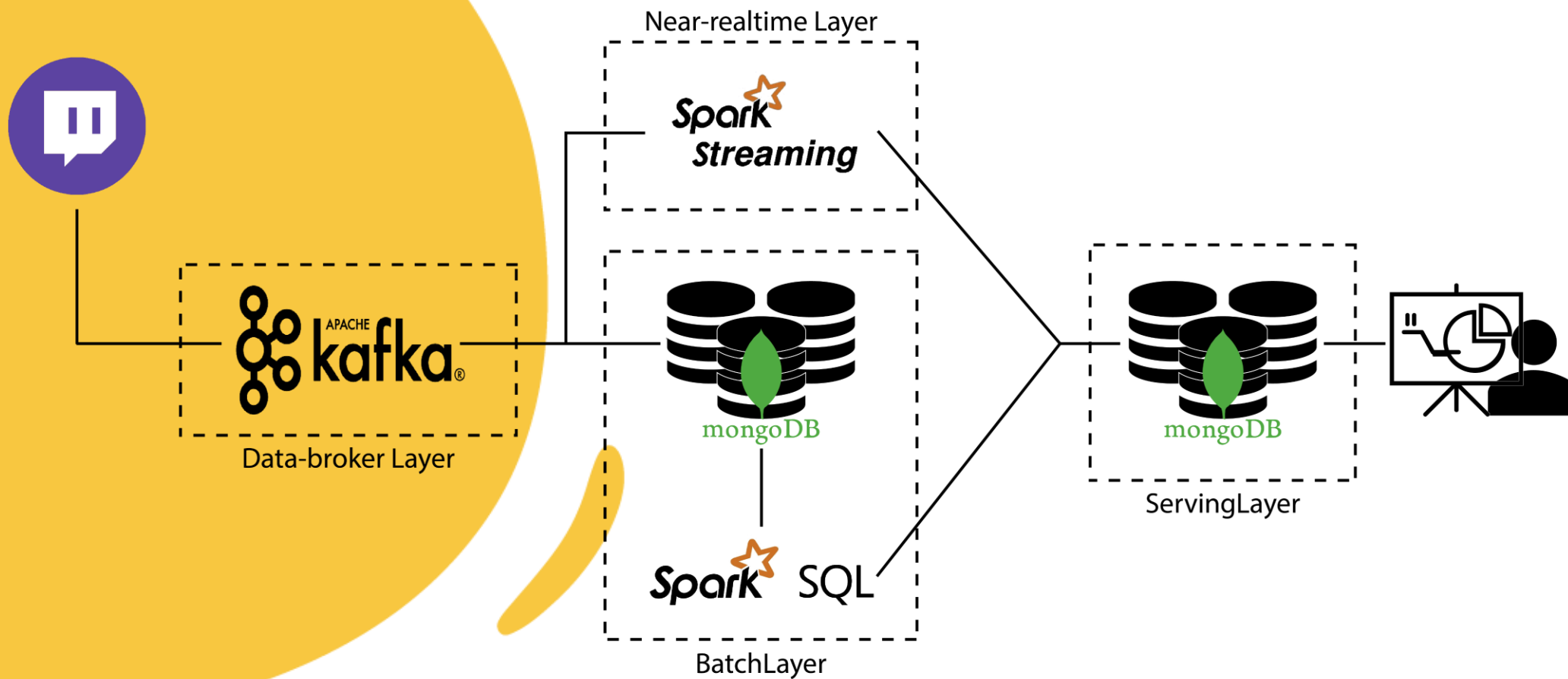


**Spark** SQL



# Architettura $\lambda$

GAME OVER

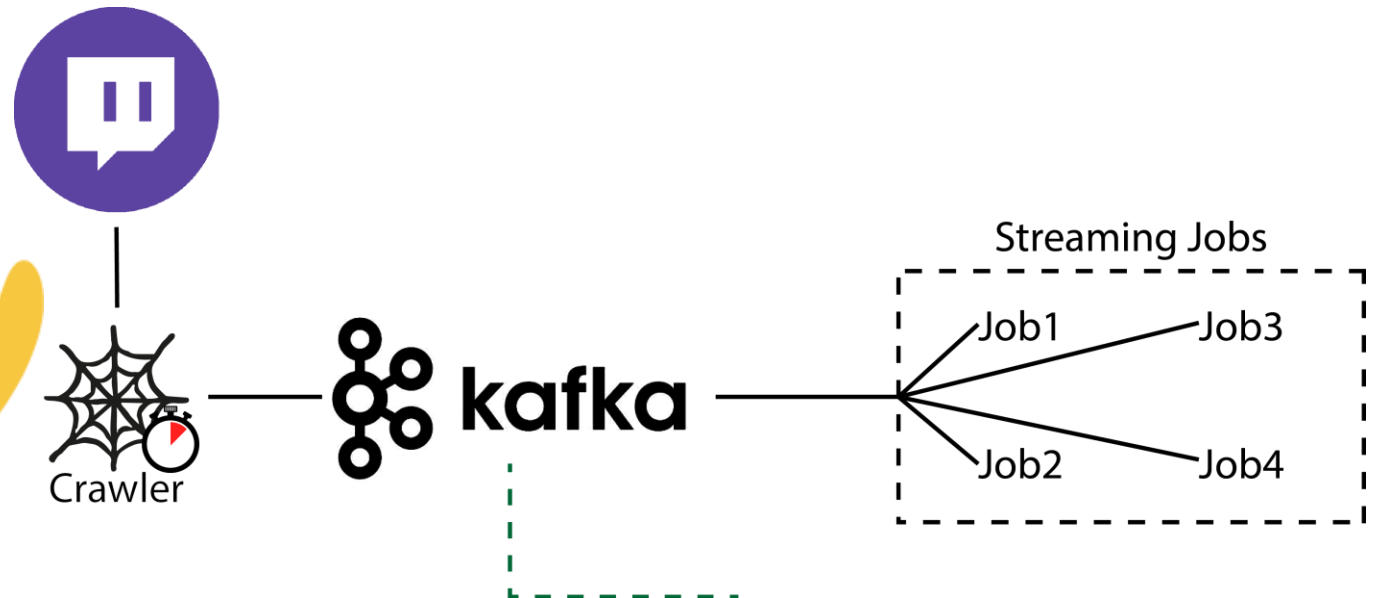


# Streaming Layer

Elaborazione di flussi di dati acquisiti in tempo reale al fine di generare un report in **near-real-time**.  
4 tipologie di Job per analizzare 4 aspetti differenti degli stessi dati.

Sono state utilizzate due tipologie di script per job:

- **Spark streaming** semplificazione ed estrapolazione dei dati d'interesse;
- **Spark SQL** per le interrogazioni.



# Streaming Layer – Job 1 e Job 2



## Job 1 – Ranking by views

Creazione di una classifica ordinata sulla base del **numero di visualizzazioni correnti** degli streaming attualmente in live.

Osservare quali dirette streaming stanno avendo successo **nel momento del crawling**.

Ora in tendenza:



## Job 2 – Ranking by mean of views

Creazione di una classifica ordinata sulla base del **numero medio di visualizzazioni** degli streaming attualmente in live.

Osservare quali dirette streaming hanno avuto maggior successo **nelle ultime 24 ore**.

Contenuti che ti sei perso che potrebbero interessarti:



# Streaming Layer – Job 1 e Job 2



## Job 1 – Ranking by views

stream_id	game_name	current_view	crawl_time
12932973168	Dota 2	29816	2015-02-01 01:15:00
12932549648	StarCraft II: Hea...	27819	2015-02-01 01:15:00
12932994272	Dota 2	24315	2015-02-01 01:15:00
12935159760	Hearthstone: Hero...	19256	2015-02-01 01:15:00
12933578608	Dota 2	16604	2015-02-01 01:15:00
12934530544	League of Legends	16579	2015-02-01 01:15:00
12932518304	Dying Light	16508	2015-02-01 01:15:00
12931574736	League of Legends	14044	2015-02-01 01:15:00
12933966224	Dying Light	11317	2015-02-01 01:15:00
12935517168	Gaming Talk Shows	10138	2015-02-01 01:15:00
12935229856	Dota 2	8863	2015-02-01 01:15:00
12936016272	Dota 2	8635	2015-02-01 01:15:00
12933757376	Dota 2	8447	2015-02-01 01:15:00
12932065776	League of Legends	8187	2015-02-01 01:15:00
12936030864	League of Legends	6528	2015-02-01 01:15:00
12933993312	Counter-Strike: G...	4920	2015-02-01 01:15:00
12931778816	Dying Light	4810	2015-02-01 01:15:00
12932512416	Counter-Strike: G...	4671	2015-02-01 01:15:00
12935833360	World of Warcraft...	4220	2015-02-01 01:15:00
12933140416	The Binding of Is...	3823	2015-02-01 01:15:00

## Job 2 – Ranking by mean of views

stream_id	game_name	sum(current_view)	count	average
12933966224	Dying Light	44721	1	44721.0
12932973168	Dota 2	159885	5	31977.0
12932549648	StarCraft II: Hea...	140787	5	28157.4
12932994272	Dota 2	101718	5	20343.6
12935159760	Hearthstone: Hero...	96551	5	19310.2
12932518304	Dying Light	79886	5	15977.2
12931574736	League of Legends	76945	5	15389.0
12934530544	League of Legends	75330	5	15066.0
12933578608	Dota 2	68002	5	13600.4
12935229856	Dota 2	61658	5	12331.6
12933966224	Left 4 Dead 2	12037	1	12037.0
12933966224	Dying Light	44721	4	11180.25
12933757376	Dota 2	52196	5	10439.2
12936016272	Dota 2	52185	5	10437.0
12935517168	Gaming Talk Shows	51254	5	10250.8
12932065776	League of Legends	37136	5	7427.2
12936030864	League of Legends	30005	5	6001.0
12933993312	Counter-Strike: G...	25813	5	5162.6
12931778816	Dying Light	24516	5	4903.2
12932512416	Counter-Strike: G...	22290	5	4458.0



# Streaming Layer – Job 3

## Job 3 – Trend games

Determinare il numero di streaming attivi per ogni categoria con lo scopo di calcolare quali siano quelle maggiormente *streammate*.

Comprendere quali sono le categorie che gli streamer preferiscono e studiarne la distribuzione.

game_name	count
League of Legends	1451
Dying Light	991
Destiny	468
Counter-Strike: G...	427
null	415
Minecraft	393
Call of Duty: Adv...	345
Grand Theft Auto V	341
World of Warcraft...	239
Call of Duty®: Ad...	223
H1Z1	214
Dota 2	211
FIFA 15	187
Battlefield 4	164
Madden NFL 15	159
Hearthstone: Hero...	155
NBA 2K15	150
Heroes of the Storm	113
Smite	102
DayZ	77

game_name	count
League of Legends	1429
Dying Light	952
Destiny	463
null	427
Counter-Strike: G...	418
Minecraft	389
Grand Theft Auto V	349
Call of Duty: Adv...	346
World of Warcraft...	234
Call of Duty®: Ad...	219
Dota 2	219
H1Z1	209
FIFA 15	190
Battlefield 4	157
Hearthstone: Hero...	157
NBA 2K15	151
Madden NFL 15	150
Heroes of the Storm	122
Smite	101
Far Cry 4	85


# Streaming Layer – Job 4



## Job 4 – Views percentage

Determinare quale sia la **percentuale degli iscritti** ad un canale **che stanno guardando la live** in corso.


Comprendere quali live hanno avuto **maggior successo** tra gli iscritti e scovare anomalie.



**CHANNEL NAME**  
92618 [ISCRIVITI](#)

Segui la live:


LIVE




29816

Altri video del canale:

TERMINATO



TERMINATO



# Streaming Layer – Job 4



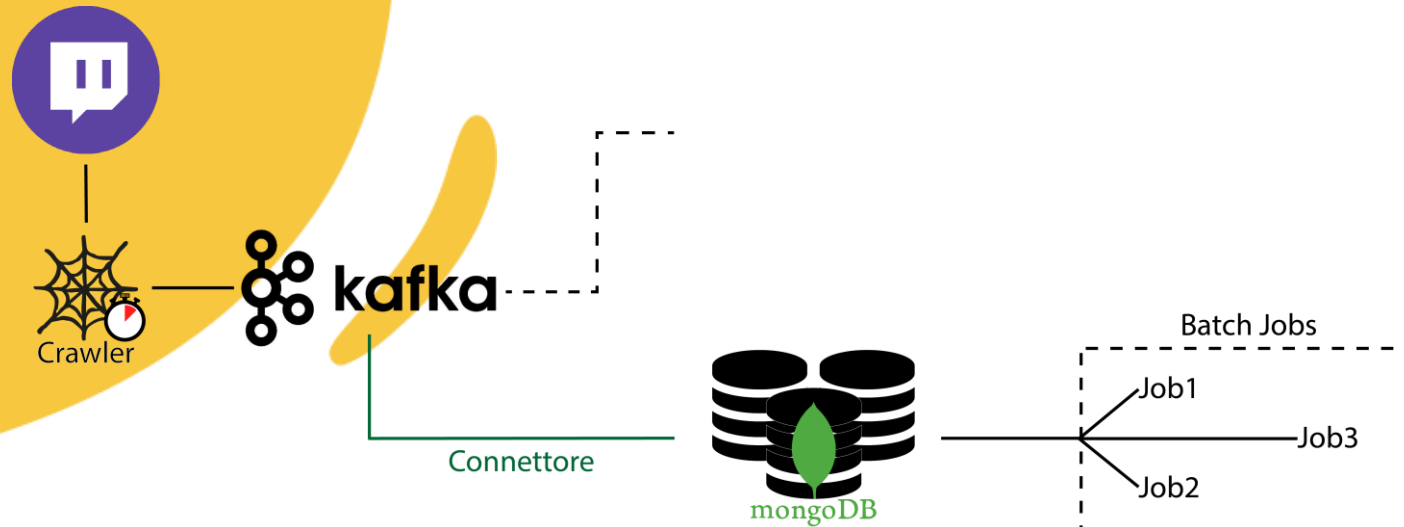
stream_id	game_name	current_view	broadcaster_id	broadcaster_name	follower_number	view_percentage	crawl_time
12935516256	Battlefield™ Hard...	83	80759669	scooby3751	2	41.5	2015-02-01 01:15:00
12931831792	Call of Duty: Adv...	119	37894502	thecodman11	3	39.666668	2015-02-01 01:15:00
12935582464	Dying Light	307	55920369	based_gordeez	10	30.7	2015-02-01 01:15:00
12936224880	APB Reloaded	49	81487127	animebhopper	4	12.25	2015-02-01 01:15:00
12935070496	World of Tanks	21	75422343	xumpbiu_jiuc	2	10.5	2015-02-01 01:15:00
12934335120	League of Legends	89	50688254	loyaltylol	9	9.888889	2015-02-01 01:15:00
12929952208	Minecraft	397	80702628	thearcalypse	49	8.102041	2015-02-01 01:15:00
12935912144	RuneScape	8	81030935	brandirex	1	8.0	2015-02-01 01:15:00
12936076544	FF14	8	52916159	lyrisbrue	1	8.0	2015-02-01 01:15:00
12936452688	Battlefield™ Hard...	15	81497008	youngturk44	2	7.5	2015-02-01 01:15:00
12930780752	null	14	79847857	redstone10th	2	7.0	2015-02-01 01:15:00
12936372256	Lords of the Fallen	7	51470823	loosdevil	1	7.0	2015-02-01 01:15:00
12936173456	Pro Evolution Soc...	6	78260104	davidrebirth77	1	6.0	2015-02-01 01:15:00
12935698288	Fibbage: The Hila...	6	45841918	jpeeper	1	6.0	2015-02-01 01:15:00
12935775888	FIFA 15	6	68116639	the_hawkz	1	6.0	2015-02-01 01:15:00
12936128656	LIMBO	6	80823315	thirsty4chicken	1	6.0	2015-02-01 01:15:00
12935838880	Raven's cry	18	40520823	frolik12	3	6.0	2015-02-01 01:15:00
12936261232	World of Tanks	5	80633741	dokeid_	1	5.0	2015-02-01 01:15:00
12936396352	Grand Theft Auto:...	5	65224056	frank207	1	5.0	2015-02-01 01:15:00
12936230960	Life Is Strange™	5	52468541	vindicatednoxus	1	5.0	2015-02-01 01:15:00

# Batch Layer

Analisi di set di dati di grandi dimensioni raccolti per un lungo periodo di tempo (giorni, settimane o mesi) prima di essere processati.

Nell'architettura proposta sono state utilizzate le seguenti tecnologie:

- **Kafka** che, tramite un connettore, permette il salvataggio persistente dei dati in un database;
- **MongoDB** per la memorizzazione dei dati all'interno del database (data\_lake);
- **SparkSQL** per il processing dei dati;



# Batch Layer – Job 1



## Job 1 – Le piattaforme preferite dai creatori di contenuti

- Diversi possibili approcci alla piattaforma, ad esempio utilizzare un pc o creare un contenuto direttamente da console (ps4 o xbox);
- Tre differenti file di testo denominati `all_broadcaster_dict`, `ps4_broadcaster_dict` e `xbox_broadcaster_dict` contengono gli ID dei broadcaster divisi per piattaforma di utilizzo;
- Utilizzo del programma python `upload_data_to_mongo.py` affinché questi tre file .txt siano caricabili su MongoDB rispettivamente in tre collezioni distinte;
- Utilizzo di PyMongo e SparkSQL per produrre l'output desiderato;

all_broadcaster	ps4_broadcaster	xbox_broadcaster	pc_broadcaster
2388705	702705	426483	1259517

# Batch Layer – Job 2

GAME  
OVER

## Job 2 – Il gioco preferito dagli iscritti

- Produrre, per ogni Streamer, un resoconto contenente il gioco più seguito sulla base delle visualizzazioni totali mensili;
- Aiuta a tener traccia dei giochi preferiti dai propri iscritti e consiglia contenuti che potrebbero aver maggior successo;
- Stabilire una connessione con MongoDB ed elaborazione dati contenuti nella collezione twitch al fine di produrre la classifica desiderata;

broadcasterID	broadcasterName	gameName	max(currentViews)
29578325	beyondthesummit	Dota 2	34846
30220059	esltv_sc2	StarCraft II: Hea...	27293
24954143	dotacinema	Dota 2	24142
29795919	nl_kripp	Hearthstone: Hero...	18725
28633266	starladder3	Dota 2	17817
28036688	trick2g	League of Legends	17044
1518077	goldglove	Dying Light	15252
36794584	riotgames2	League of Legends	14314
28633177	starladder1	Dota 2	13070
7951350	cryaotic	Left 4 Dead 2	12037
28633298	starladder4	Dota 2	11039
31478096	mym_alkapone	Gaming Talk Shows	10315
29769280	beyondthesummit2	Dota 2	10148
32803072	bestrivenna	League of Legends	5880
54706574	theoriginalweed	Counter-Strike: G...	5113
38881685	flosd	Dying Light	5080
14293484	voyboy	League of Legends	4832
30080840	tsm_theoddone	XCOM: Enemy Within	4282
37701508	phantoml0rd	Counter-Strike: G...	4137
23524577	swifty	World of Warcraft...	3984

# Batch Layer – Job 3

## Job 3 – Top 25 Games and Streamer

- Tre resoconti utili sia ai creatori di contenuti che agli editori di videogiochi;
- Connessione con MongoDB;
- DataFrame per le interrogazioni;
- Top 25 dei giochi con più contenuti sulla piattaforma nel mese d'interesse;
- Top 25 giochi più seguiti del mese;
- Top 25 streamer più seguiti del mese;

gameName	count	broadcasterName	sum(currentViews)
League of Legends	1514	beyondthesummit	34846
Dying Light	1096	esltv_sc2	27293
Destiny	519	dotacinema	24142
-1	472	nl_kripp	18725
Counter-Strike: G...	443	starladder3	17817
Minecraft	418	trick2g	17044
Call of Duty: Adv...	394	goldglove	15252
Grand Theft Auto V	369	riotgames2	14314
World of Warcraft...	253	starladder1	13070
H1Z1	235	cryaotic	12037
Dota 2	225	starladder4	11039
Call of Duty®: Ad...	225	mym_alkapone	10315
Battlefield 4	177	beyondthesummit2	10148
Madden NFL 15	170	bestrivenna	5880
Hearthstone: Hero...	167	theoriginalweed	5113

gameName	sum(currentViews)
Dota 2	118968
League of Legends	68964
StarCraft II: Hea...	32536
Dying Light	28404
Hearthstone: Hero...	28213
Counter-Strike: G...	18311
Gaming Talk Shows	16824
Left 4 Dead 2	12162
World of Warcraft...	11699
Minecraft	10004
H1Z1	8027
The Binding of Is...	5086
Destiny	4753
XCOM: Enemy Within	4294
Call of Duty: Adv...	4290
RuneScape	3487
Grand Theft Auto V	3472
Resident Evil Arc...	3359
Magic: The Gathering	2952
Heroes of the Storm	2731



# Conclusioni e Sviluppi Futuri

- Architettura Lambda è una delle architetture maggiormente utilizzate nel mondo dei Big Data;
- Utilizzo di diverse tecnologie in grado di interagire e collaborare tra loro per uno scopo comune;
- Sviluppi futuri:
  - Un'interfaccia web per la visualizzazione dei risultati dei vari job;
  - Nuovi job:
    - **Streaming**: monitorare il numero degli iscritti durante una live;
    - **Batch**: numero di streaming che vengono avviati utilizzando un pc o direttamente una console.
  - Implementazione di un secondo crawler;
  - Predizione di trend utilizzano opportune librerie di Machine Learning (Mlib).





# Grazie per la vostra attenzione

Polosa Sebastiano e Ungaro Riccardo

