

You Tube

SPARK STREAMING PROJECT

ANALYZING YOUTUBE TRENDING DATA

Presented by :
Tuğrul Gökçe
Serkan Birol



About Project

	video_id	title	published_at	channel_id	channel_title	category_id	trending_date	view_count	likes	dislikes	comment_count	comments_disabled	ratings_disabled	category_title
1	43SVfVzniAo	Theres SO Much Depth	2022-02-25 19:00:03	UCMhD97ZasArNEhc7o...	Surfnboy Shorts	24	2022-02-26	965914	87246	0	416	FALSE	FALSE	Entertainment
2	mVDFJRM6F9k	Aperture Desk Job Trailer	2022-02-25 17:57:42	UCg0FSqPeiGD_IliPaaAe...	Valve	20	2022-02-26	529996	46042	0	3443	FALSE	FALSE	Gaming
3	SSQrPsQq_q4	Perks of getting your per...	2022-02-25 17:00:27	UCsKVP_4zQ877TEiH_lh...	illymation	1	2022-02-26	415195	40210	0	4221	FALSE	FALSE	Film & Animation
4	SZylCbao4wA	Trope Talk: Love Triangles	2022-02-25 16:00:12	UCodbH5mUeF-m_BsN...	Overly Sarcastic Prod...	27	2022-02-26	352777	30008	0	4013	FALSE	FALSE	Education
5	c6MW-qdNoYA	Elden Ring from the NPC...	2022-02-25 16:00:11	UCVIFCOJw3emIVmBb...	Joel Haver	1	2022-02-26	446349	62904	0	1633	FALSE	FALSE	Film & Animation
6	qDIQf10kHzY	Ukrainian soldiers on fro...	2022-02-25 15:56:52	UCPgLNge0xqQHWM5...	The Telegraph	25	2022-02-26	2045388	21657	0	0	TRUE	FALSE	News & Politics
7	QKOezO8yNdo	Tremors 4: The Legend B...	2022-02-25 14:00:18	UC2lYWWifolht9T47z1Z...	Dead Meat	1	2022-02-26	375999	24850	0	1736	FALSE	FALSE	Film & Animation
8	KUau7vVa5pQ	[1421] US Gov't Says You...	2022-02-25 14:00:12	UCm9K6rby98W8JigLoZ...	LockPickingLawyer	27	2022-02-26	1070979	40269	0	3961	FALSE	FALSE	Education
9	Yl_f0sYMaGk	JUJUTSU KAISEN 0 RELE...	2022-02-25 13:50:43	UC6pGDc4bFGD1_36IKv...	Crunchyroll Collection	1	2022-02-26	938136	41047	0	1161	FALSE	FALSE	Film & Animation
10	bzSAjQZXDOE	Ukraine war: Kyiv braces ...	2022-02-25 13:48:15	UCPgLNge0xqQHWM5...	The Telegraph	25	2022-02-26	5283245	38451	0	0	TRUE	FALSE	News & Politics

Analyzing Youtube stream trend data and publishing it to real-time dashboard.

Our analysis subjects ;

- 1.What are the most viewed channels ?
- 2.What are the most liked categories ?
- 3.What are the most viewed categories ?
- 4.What are the most liked channels ?

Methodology

Flow Management



Reading csv file and
produce record to
Kafka Topic

ETL Jobs
(CSV to JSON, JSON to
SQL etc.)

After Spark Analyzing
session sink data to
NoSQL and RDBMS

Stream Processing



Listen to Kafka Topic

Analyzing stream data
from Kafka Topic using
Scala-Spark.

Writing to a different
Kafka Topic after
analysis

Databases



Importing analyzed data
to NoSQL and RDBMS
systems.

Querying by running KQL
over Elasticsearch.

Dashboard



Creating a dashboard
every 1 second over
real-time data.

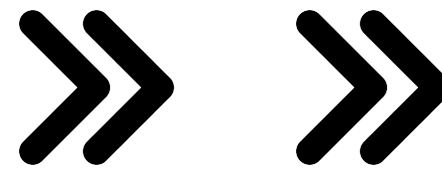
Project Overview



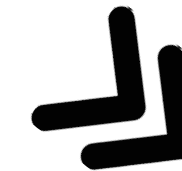
Youtube Trending Data



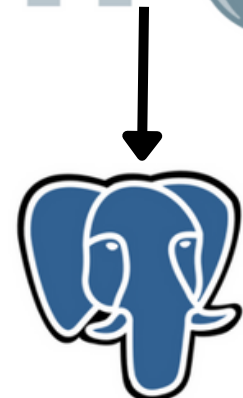
Creating Stream Data for Kafka Topic



Reading data for Analysis



Result of Analysis

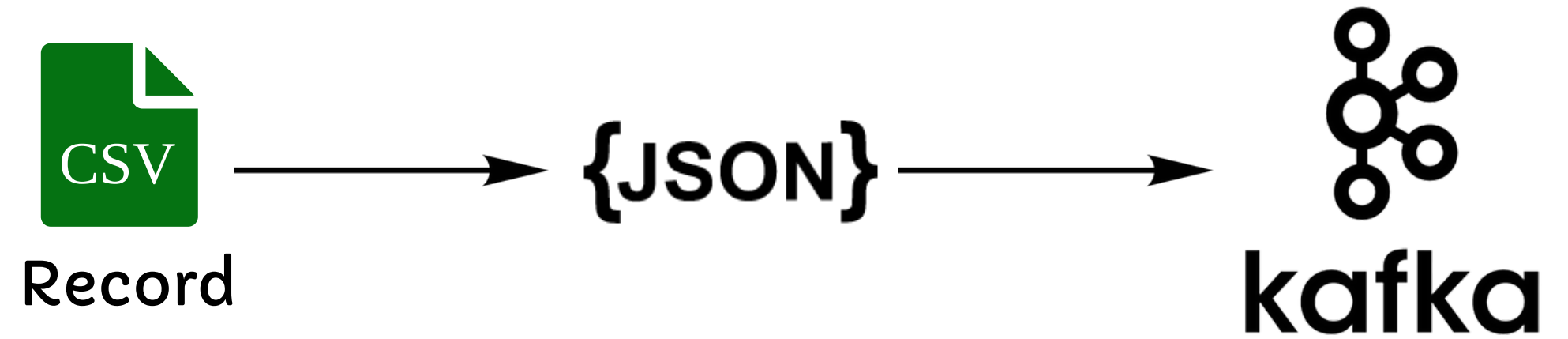
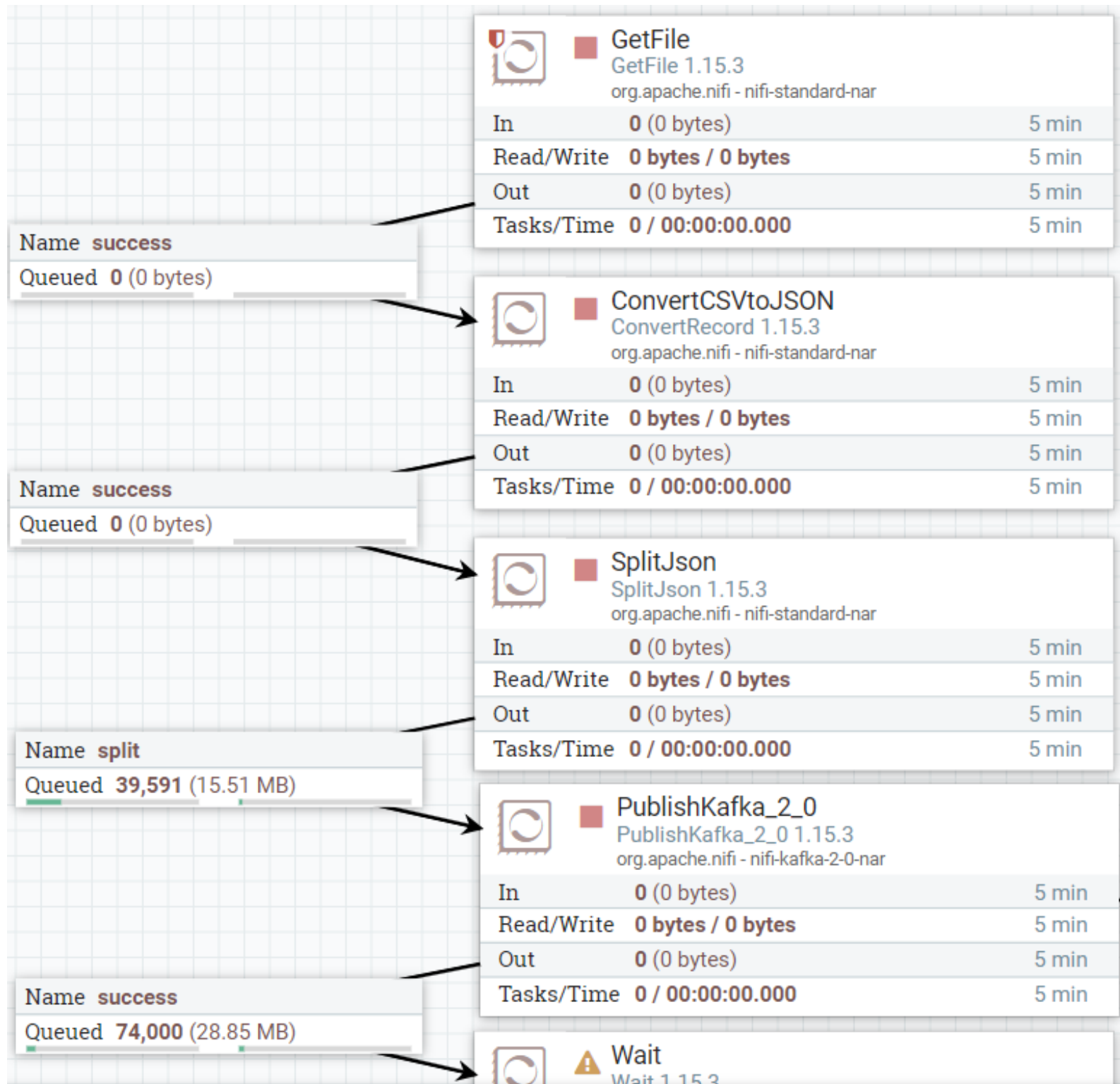


PostgreSQL



kibana

Project Details



```
[{"video_id": "xK5qVtcVmk", "title": "Roosil - Love Me Easy (Official Audio)", "published_at": "2022-02-14T17:50:13Z", "channel_id": "UCUtmzHuW43BF0UP-Xl0T7KA", "channel_title": "Roosil 2x", "category_id": 10, "trending_date": "2022-02-19 00:00:00+00:00", "view_count": 458758, "likes": 29118, "dislikes": 0, "comment_count": 1201, "comments_disabled": false, "ratings_disabled": false, "category_title": "Music"}, {"video_id": "sUEG2uie6Qo", "title": "I Busted 72 Chapter 3 Myths That 100% Work...", "published_at": "2022-02-14T20:45:43Z", "channel_id": "UCGNs7Q5TLpOzx-oIocgkQA", "channel_title": "Txns", "category_id": 20, "trending_date": "2022-02-19 00:00:00+00:00", "view_count": 982704, "likes": 12611, "dislikes": 0, "comment_count": 349, "comments_disabled": false, "ratings_disabled": false, "category_title": "Gaming"}, {"video_id": "i#NAME?", "title": "Newegg Responded (Sort Of)", "published_at": "2022-02-15T15:14:13Z", "channel_id": "UCHis72whgZi9w6d6FhwGGHA", "channel_title": "Gamers Nexus", "category_id": 20, "trending_date": "2022-02-19 00:00:00+00:00", "view_count": 811654, "likes": 57956, "dislikes": 0, "comment_count": 5579, "comments_disabled": false, "ratings_disabled": false, "category_title": "Gaming"}, {"video_id": "G5W1dWxSS-4", "title": "New Ringu Killer The Onryo- New Survivor, Mori, Perks, Add-Ons & more | Dead by Daylight PTB", "published_at": "2022-02-15T17:32:00Z", "channel_id": "UCUDO-z-2Zd4bfG9m4AkK_uA", "channel_title": "not Otzdarva", "category_id": 20, "trending_date": "2022-02-19 00:00:00+00:00", "view_count": 841017, "likes": 34823, "dislikes": 0, "comment_count": 2446, "comments_disabled": false, "ratings_disabled": false, "category_title": "Gaming"}, {"video_id": "rB73q8tp3tM", "title": "Minecraft, But Every Minute It Gets More Scary...", "published_at": "2022-02-15T22:07:19Z", "channel_id": "UCKYb5XB8e-SOSEgLiJLSOdTw", "channel_title": "UnspeakableGaming", "category_id": 20, "trending_date": "2022-02-19 00:00:00+00:00", "view_count": 679458, "likes": 16219, "dislikes": 0, "comment_count": 996, "comments_disabled": false, "ratings_disabled": false, "category_title": "Gaming"}, {"video_id": "CG1Pkmeg7o", "title": "Cyberpunk 2077 Just Got a 60GB New Update", "published_at": "2022-02-15T19:36:45Z", "channel_id": "UCLdPU0IBv-O0mTWtUBAOnTg", "channel_title": "JuiceHead", "category_id": 20, "trending_date": "2022-02-19 00:00:00+00:00", "view_count": 409098, "likes": 13058, "dislikes": 0, "comment_count": 1687, "comments_disabled": false, "ratings_disabled": false, "category_title": "Gaming"}, {"video_id": "bgICoQj_SyS", "title": "Dead by Daylight | Sadako Rising | Reveal Trailer", "published_at": "2022-02-15T16:00:09Z", "channel_id": "UCASgsFdGbwjfdawl3rOXiwQ", "channel_title": "Dead by Daylight", "category_id": 20, "trending_date": "2022-02-19 00:00:00+00:00", "view_count": 675342, "likes": 41170, "dislikes": 0, "comment_count": 2572, "comments_disabled": false, "ratings_disabled": false, "category_title": "Gaming"}, {"video_id": "8eH2GloVwhg", "title": "Go To Bed - Cyanide & Happiness Shorts", "published_at": "2022-02-17T17:00:05Z", "channel_id": "UCWKCrrtCF6sgKrdx0US-Idw", "channel_title": "ExplosmEntertainment", "category_id": 23, "trending_date": "2022-02-19 00:00:00+00:00", "view_count": 572194, "likes": 39213, "dislikes": 0, "comment_count": 882, "comments_disabled": false, "ratings_disabled": false, "category_title": "Comedy"}, {"video_id": "mXGnPoNF054", "title": "The Lawsuit of Shane Dawson", "published_at": "2022-02-15T22:04:03Z", "channel_id": "UCV9v_KinVpV-snHe3C3nlhvA", "channel_title": "shane", "category_id": 23, "trending_date": "2022-02-19 00:00:00+00:00", "view_count": 2386146, "likes": 132687, "dislikes": 0, "comment_count": 18207, "comments_disabled": false, "ratings_disabled": false, "category_title": "Comedy"}, {"video_id": "pKAcITUm4LM", "title": "what is this EXACTLY? 🤔#shorts", "published_at": "2022-02-18T17:30:06Z", "channel_id": "UCHWgDpLoyJgtFXoAaaStu40", "channel_title": "lionfield", "category_id": 23, "trending_date": "2022-02-19 00:00:00+00:00", "view_count": 1474696, "likes": 129090, "dislikes": 0, "comment_count": 670, "comments_disabled": false, "ratings_disabled": false, "category_title": "Comedy"}, {"video_id": "6nlyBZPMR7k", "title": "So I tried Cyberpunk's NEXT GEN Update", "published_at": "2022-02-15T21:47:01Z", "channel_id": "UCsnEJAjrwfuU7T2yYv0WsA", "channel_title": "Kevduit", "category_id": 20, "trending_date": "2022-02-19 00:00:00+00:00", "view_count": 383898, "likes": 15090, "dislikes": 0, "comment_count": 1068, "comments_disabled": false, "ratings_disabled": false, "category_title": "Gaming"}, {"video_id": "Fm4E8v3_2s", "title": "Fortnite *NEW* Best GYRO Controller Settings & New Aim Update & Flick Stick Settings (XBOX/PS5)", "published_at": "2022-02-15T22:57:06Z", "channel_id": "UCM SXHBV7033NzyGccDug", "channel_title": "Disc", "category_id": 20, "trending_date": "2022-02-19 00:00:00+00:00", "view_count": 149097, "likes": 3391, "dislikes": 0, "comment_count": 464, "comments_disabled": false, "ratings_disabled": false, "category_title": "Gaming"}, {"video_id": "Y4Wwqz3ng_U", "title": "This Kinda SUCKS! - Cyberpunk 2077 Next Gen Update is HERE (PS5/Xbox)", "published_at": "2022-02-15T18:03:04Z", "channel_id": "UCiHFS5txB007jckRk-k8XQ", "channel_title": "DreamcastGuy", "category_id": 20, "trending_date": "2022-02-19 00:00:00+00:00", "view_count": 227405, "likes": 4959, "dislikes": 0, "comment_count": 2621, "comments_disabled": false, "ratings_disabled": false, "category_title": "Gaming"}, {"video_id": "e6M1juzA-wU", "title": "LAZAR vs SIDEMEN", "published_at": "2022-02-16T21:06:37Z", "channel_id": "UCZ3Amkn5JtbxXCe05a4peoQ", "channel_title": "LazarLazar", "category_id": 20, "trending_date": "2022-02-19 00:00:00+00:00", "view_count": 860576, "likes": 65177, "dislikes": 0, "comment_count": 1662, "comments_disabled": false, "ratings_disabled": false, "category_title": "Gaming"}]
```

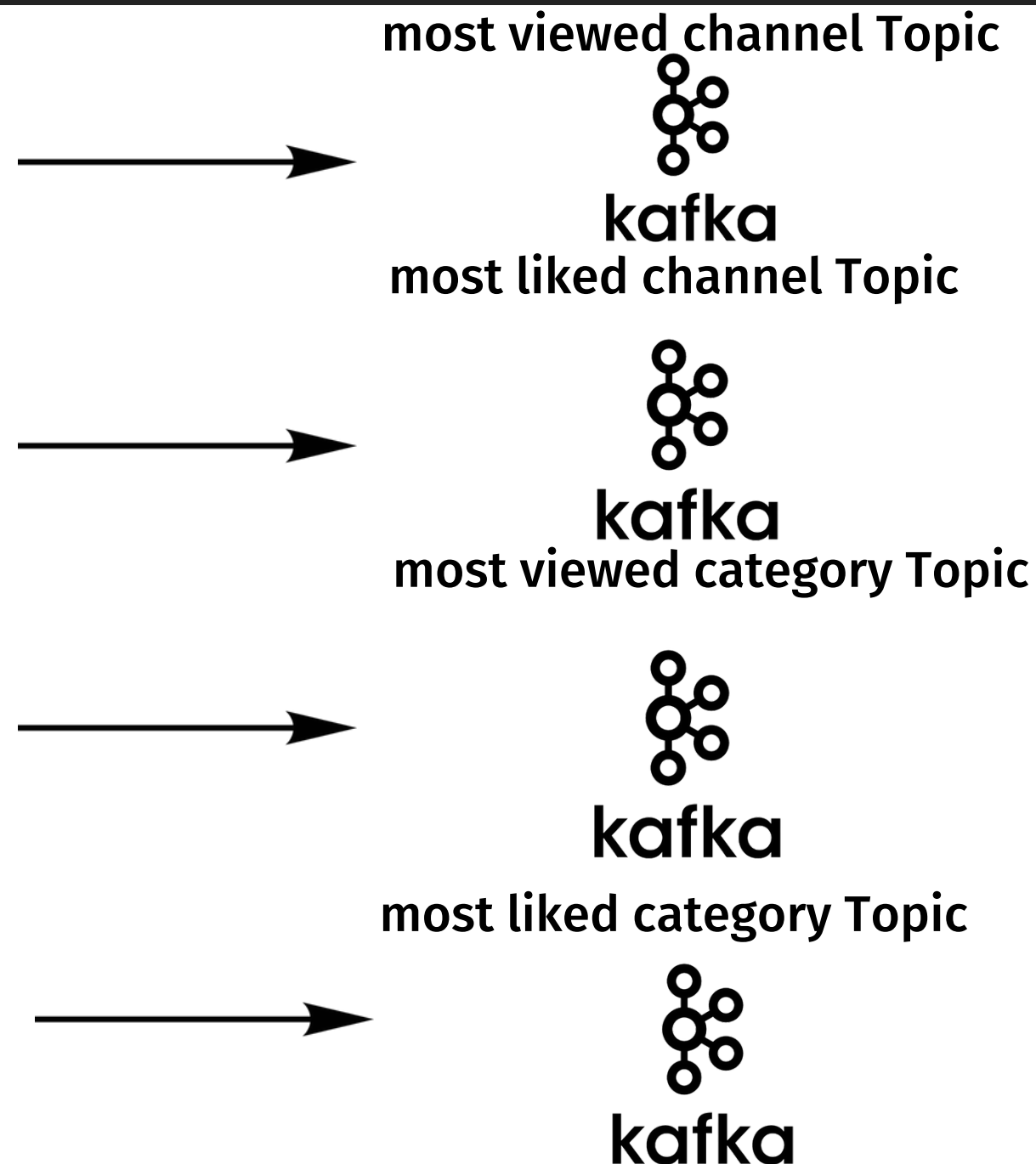
Analyzing Youtube Data with Spark Streaming

```
scala> val mostviewedchannelDf = youtubeSchemaDf.groupBy("trending_date","channel_title").agg(sum("view_count").as("total_view_count_by_channel"))
mostviewedchannelDf: org.apache.spark.sql.DataFrame = [trending_date: string, channel_title: string ... 1 more field]

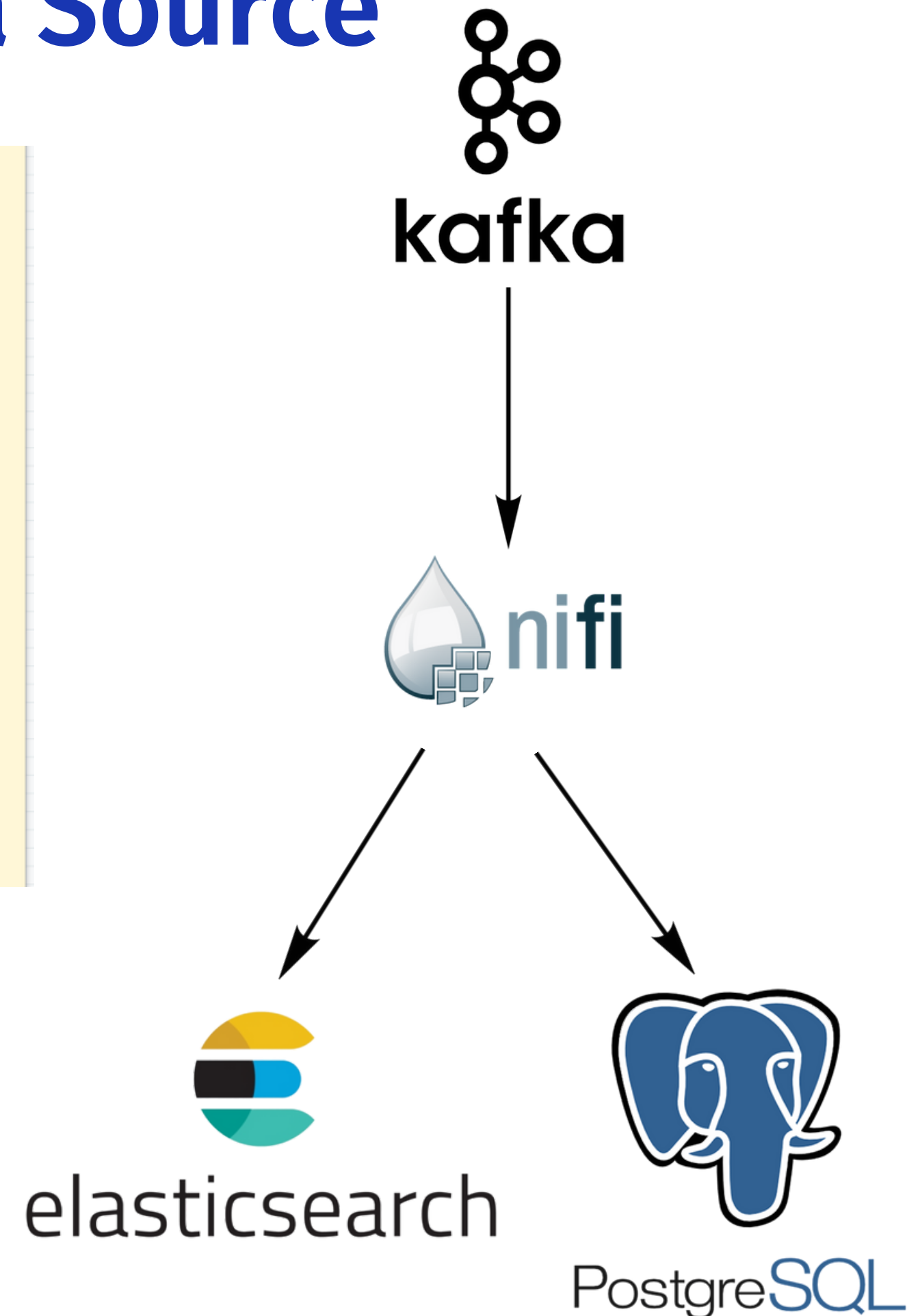
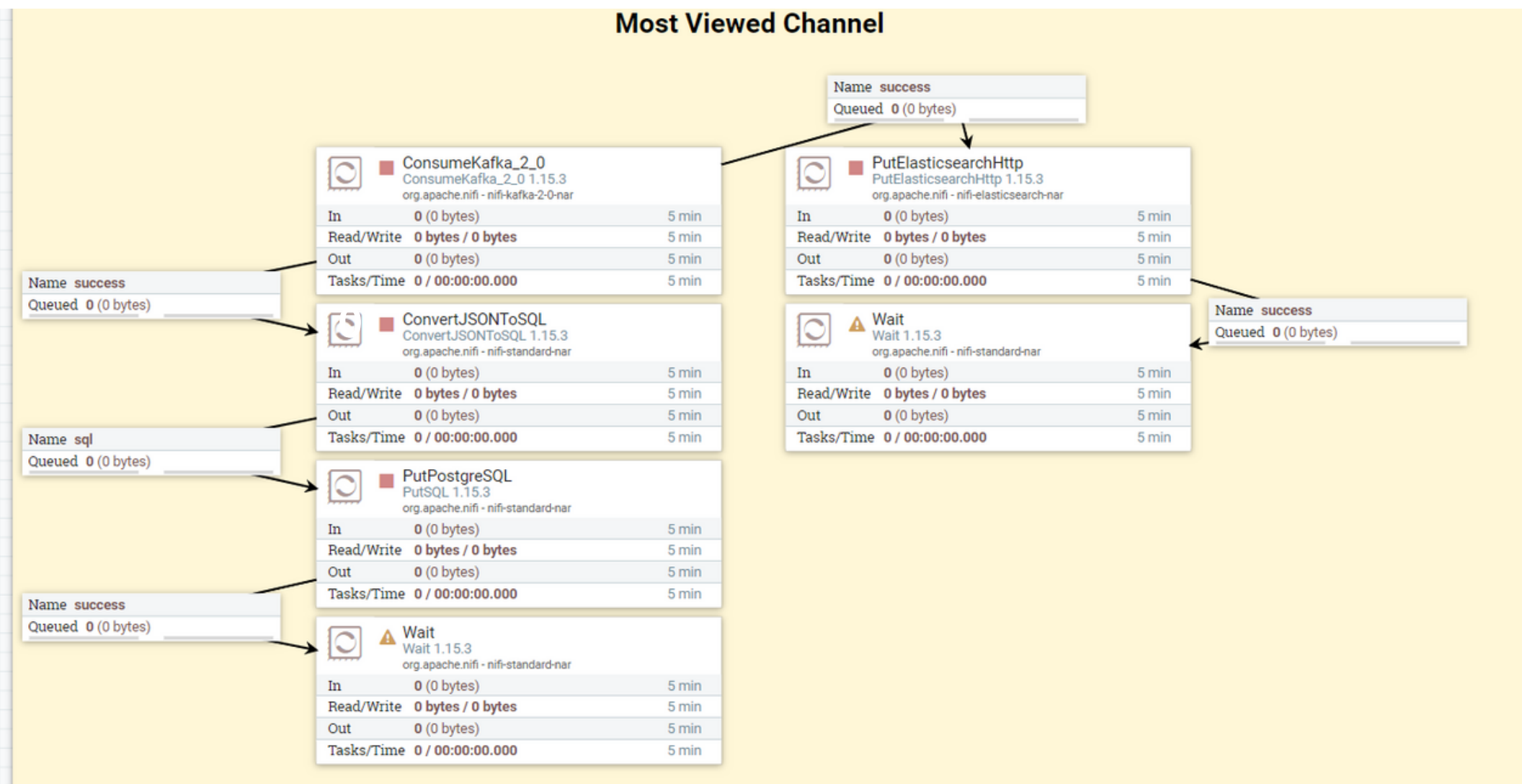
scala> mostviewedchannelDf.printSchema
root
|-- trending_date: string (nullable = true)
|-- channel_title: string (nullable = true)
|-- total_view_count_by_channel: long (nullable = true)

scala> import org.apache.spark.sql.streaming.Trigger.ProcessingTime
import org.apache.spark.sql.streaming.Trigger.ProcessingTime

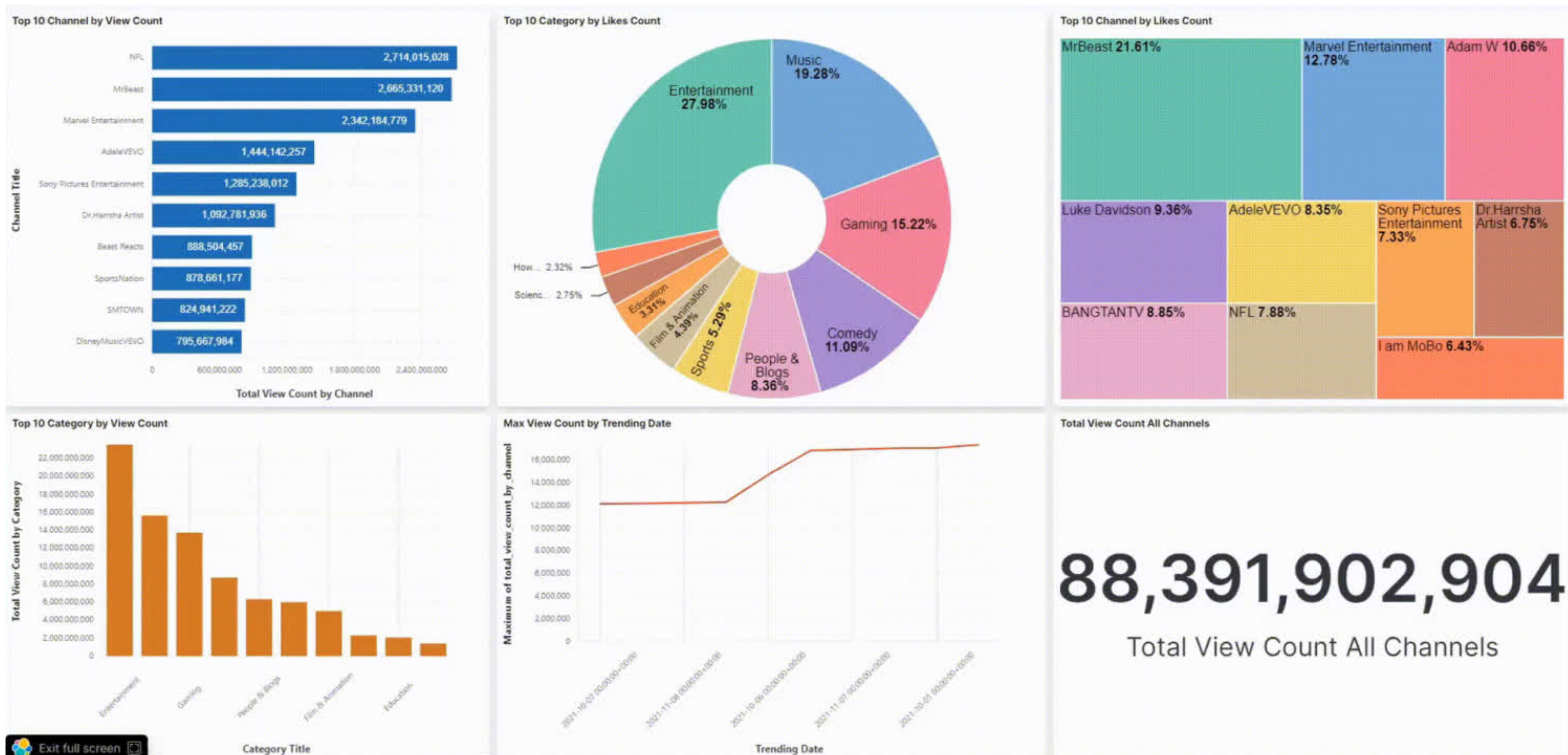
scala> val mostviewedchannelDfQuery = mostviewedchannelDf.selectExpr("CAST(trending_date as STRING)","CAST(channel_title AS STRING)", "CAST(total_view_count_by_channel AS INTEGER)", "to_json(
struct(*) AS value)").writeStream.format("kafka").option("checkpointLocation", "/home/tugrulgkc34/consumermostviewedchannel").option("failOnDataLoss", "false").outputMode("complete").option("
kafka.bootstrap.servers", "localhost:9092").option("topic", "mostviewedchannel").start()
```



Writing Data Source



Real-time Dashboard with Kibana





Thank you!

Feel free to approach us if you have any questions.

