DS203 – Big Data Analytics (Spark)

LIVE

BIG DATA ANALYSIS OF YOUTUBE TRENDING VIDEOS ACROSS COUNTRIES

Tools: PySpark, SQL, Python, Matplotlib

Presented By: Abisola Ojewole

December, 2025

```
# Filtering 1: Videos with over 1 million views

spark.sql("""
SELECT country, title, views
FROM youtube
WHERE views > 1000000
ORDER BY views DESC
LIMIT 10
""").show(truncate=False)
```

```
# Aggregate 2: Total views per country

spark.sql("""
SELECT country,
       SUM(views) AS total_views
FROM youtube
GROUP BY country
ORDER BY total_views DESC
""").show()
```

Objectives

- Analyze large-scale YouTube trending data using Spark
- Compare video performance across countries
- Examine both popularity and engagement.
- Apply Big Data concepts: filtering, aggregation, joins, and window functions

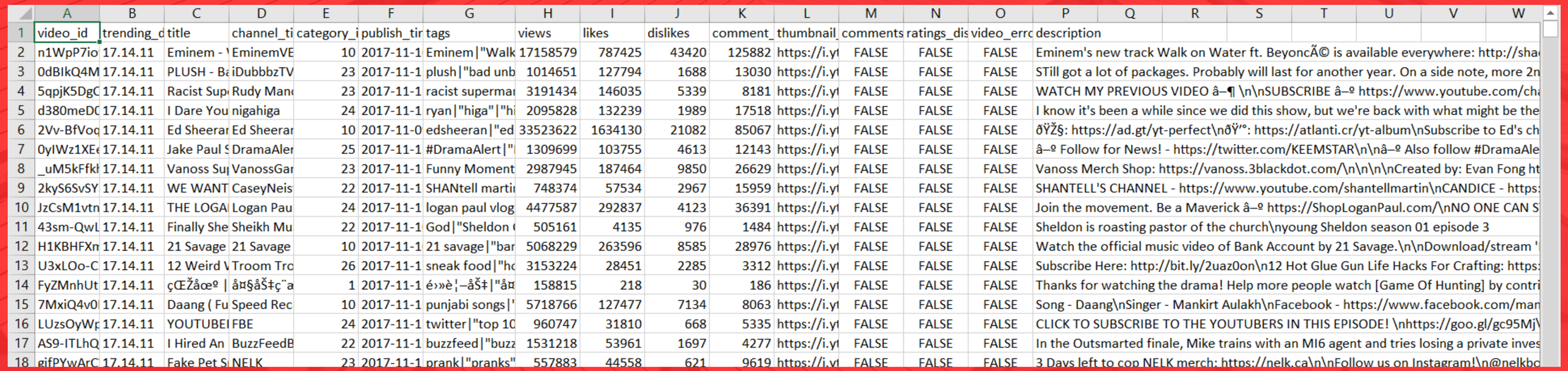
Why Apache Spark?

- Handles large datasets efficiently
- Supports distributed processing
- Faster than traditional tools like Excel or Pandas
- Allows SQL-style queries and advanced analytics

Dataset Overview

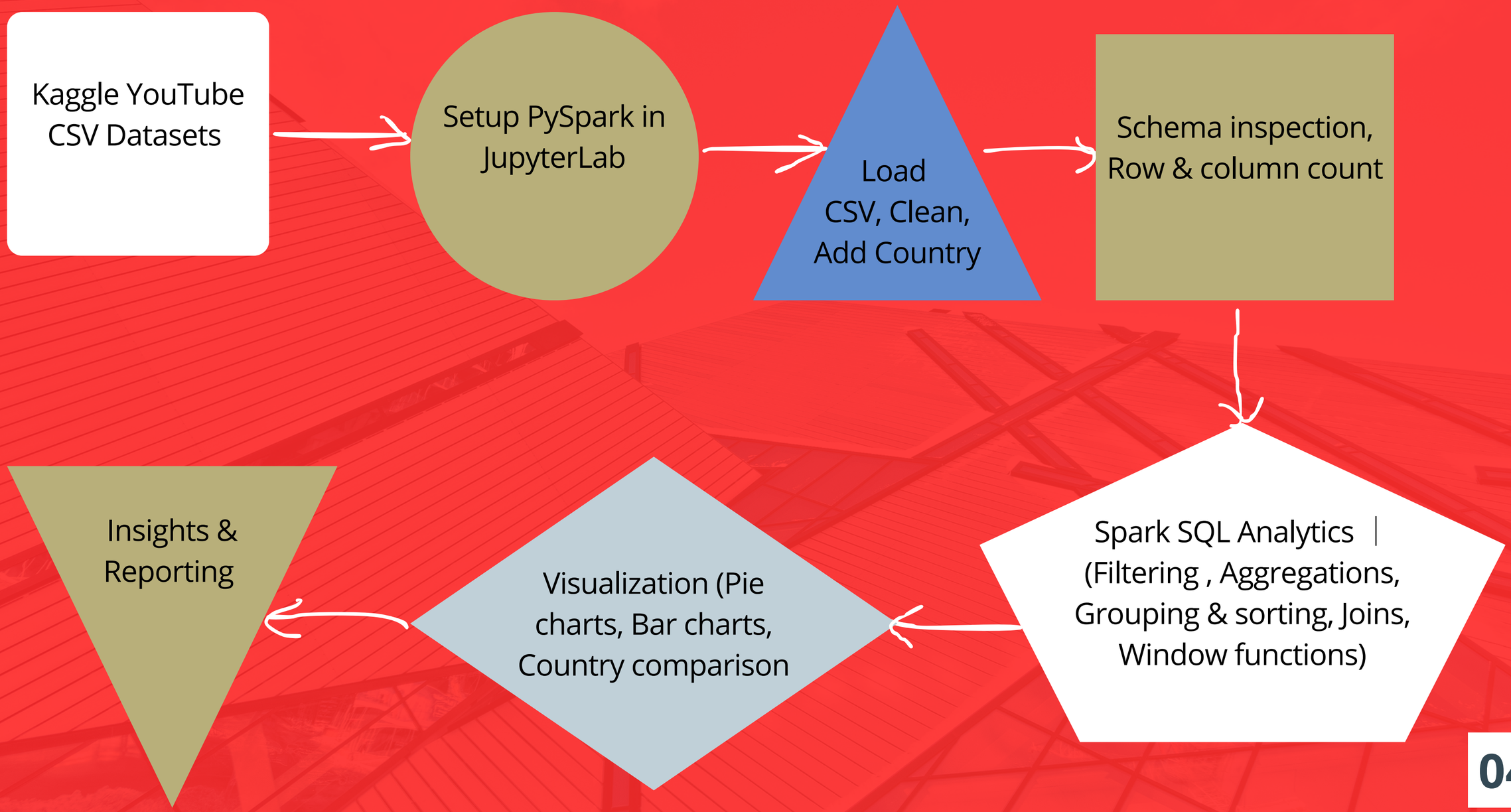
- YouTube Trending Videos dataset (539.22 MB)
- 10 countries: USA, Great Britain, Germany, Canada, Russia, Mexico, South Korea, Japan, India, France
- 361,424 records and 17 columns after cleaning
- Multiple CSV files (one per country)

- Kaggle (Trending YouTube Video Statistics)
- Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	video_id	trending_d	title	channel_ti	category_i	publish_tir	tags	views	likes	dislikes	comment_	thumbnail_	comments	ratings_dis	video_erro	description							
2	n1WpP7io	17.14.11	Eminem - \	EminemVE	10	2017-11-1	Eminem|"Walk	17158579	787425	43420	125882	https://i.yt	FALSE	FALSE	FALSE	Eminem's new track Walk on Water ft. BeyoncÃ© is available everywhere: http://sha							
3	0dBIkQ4M	17.14.11	PLUSH - Ba	iDubbbzTV	23	2017-11-1	plush|"bad unb	1014651	127794	1688	13030	https://i.yt	FALSE	FALSE	FALSE	STill got a lot of packages. Probably will last for another year. On a side note, more 2n							
4	5qpjK5DgC	17.14.11	Racist Sup	Rudy Mane	23	2017-11-1	racist superma	3191434	146035	5339	8181	https://i.yt	FALSE	FALSE	FALSE	WATCH MY PREVIOUS VIDEO â–¶ \n\nSUBSCRIBE â–º https://www.youtube.com/cha							
5	d380meD0	17.14.11	I Dare You	nigahiga	24	2017-11-1	ryan|"higa"|"hi	2095828	132239	1989	17518	https://i.yt	FALSE	FALSE	FALSE	I know it's been a while since we did this show, but we're back with what might be the							
6	2Vv-BfVoq	17.14.11	Ed Sheeran	Ed Sheeran	10	2017-11-0	edsheeran|"ed	33523622	1634130	21082	85067	https://i.yt	FALSE	FALSE	FALSE	ðŸŽ§: https://ad.gt/yt-perfect\nðŸ’°: https://atlanti.cr/yt-album\nSubscribe to Ed's ch							
7	0yIWz1XE	17.14.11	Jake Paul S	DramaAler	25	2017-11-1	#DramaAlert|"	1309699	103755	4613	12143	https://i.yt	FALSE	FALSE	FALSE	â–º Follow for News! - https://twitter.com/KEEMSTAR\n\nâ–º Also follow #DramaAle							
8	_uM5kFfkI	17.14.11	Vanoss Su	VanossGar	23	2017-11-1	Funny Moment	2987945	187464	9850	26629	https://i.yt	FALSE	FALSE	FALSE	Vanoss Merch Shop: https://vanoss.3blackdot.com/\n\n\n\nCreated by: Evan Fong ht							
9	2kyS6SvSY	17.14.11	WE WANT	CaseyNeis	22	2017-11-1	SHANtell martii	748374	57534	2967	15959	https://i.yt	FALSE	FALSE	FALSE	SHANTELL'S CHANNEL - https://www.youtube.com/shantellmartin\nCANDICE - https:							
10	JzCsM1vtn	17.14.11	THE LOGA	Logan Pau	24	2017-11-1	logan paul vlog	4477587	292837	4123	36391	https://i.yt	FALSE	FALSE	FALSE	Join the movement. Be a Maverick â–º https://ShopLoganPaul.com/\nNO ONE CAN S							
11	43sm-QwL	17.14.11	Finally She	Sheikh Mu	22	2017-11-1	God|"Sheldon (	505161	4135	976	1484	https://i.yt	FALSE	FALSE	FALSE	Sheldon is roasting pastor of the church\nyoung Sheldon season 01 episode 3							
12	H1KBHFXn	17.14.11	21 Savage	21 Savage	10	2017-11-1	21 savage|"bar	5068229	263596	8585	28976	https://i.yt	FALSE	FALSE	FALSE	Watch the official music video of Bank Account by 21 Savage.\n\nDownload/stream '							
13	U3xLOo-C	17.14.11	12 Weird \	Troom Tro	26	2017-11-1	sneak food|"ho	3153224	28451	2285	3312	https://i.yt	FALSE	FALSE	FALSE	Subscribe Here: http://bit.ly/2uaz0on\n12 Hot Glue Gun Life Hacks For Crafting: https:							
14	FyZMnhUt	17.14.11	çŒŽåœº |	å¤§å Š‡ç¨a	1	2017-11-1	é›»è¦–â Š‡|"å¤	158815	218	30	186	https://i.yt	FALSE	FALSE	FALSE	Thanks for watching the drama! Help more people watch [Game Of Hunting] by contri							
15	7MxiQ4v0	17.14.11	Daang (Fu	Speed Rec	10	2017-11-1	punjabi songs|'	5718766	127477	7134	8063	https://i.yt	FALSE	FALSE	FALSE	Song - Daang\nSinger - Mankirt Aulakh\nFacebook - https://www.facebook.com/man							
16	LUzsOyWp	17.14.11	YOUTUBEI	FBE	24	2017-11-1	twitter|"top 10	960747	31810	668	5335	https://i.yt	FALSE	FALSE	FALSE	CLICK TO SUBSCRIBE TO THE YOUTUBERS IN THIS EPISODE! \nhttps://goo.gl/gc95MjV							
17	AS9-ITLhQ	17.14.11	I Hired An	BuzzFeedB	22	2017-11-1	buzzfeed|"buzz	1531218	53961	1697	4277	https://i.yt	FALSE	FALSE	FALSE	In the Outsmarted finale, Mike trains with an MI6 agent and tries losing a private inves							
18	gifPYwArC	17.14.11	Fake Pet S	NELK	23	2017-11-1	prank|"pranks"	557883	44558	621	9619	https://i.yt	FALSE	FALSE	FALSE	3 Days left to cop NELK merch: https://nelk.ca\n\nFollow us on Instagram!\n@nelkbo							

Workflow


Kaggle YouTube CSV Datasets
Setup PySpark in JupyterLab
Load CSV, Clean, Add Country
Schema inspection, Row & column count
Spark SQL Analytics | (Filtering , Aggregations, Grouping & sorting, Joins, Window functions)
Visualization (Pie charts, Bar charts, Country comparison
Insights & Reporting

```python
# load, clean,and add country column

from pyspark.sql.functions import col, lit

dfs = []

for country, filename in files.items():
    df = spark.read.csv(
        base_path + "\\" + filename,
        header=True,
        quote='"',
        escape='"',
        multiLine=True,
        mode="PERMISSIVE"
    )

    df = (
        df.dropna(how="all")
          .withColumn("views", col("views").cast("long"))
          .withColumn("likes", col("likes").cast("long"))
          .withColumn("dislikes", col("dislikes").cast("long"))
          .withColumn("comment_count", col("comment_count").cast("long"))
          .fillna({"views":0, "likes":0, "dislikes":0, "comment_count":0})
          .dropDuplicates(["video_id","trending_date"])
          .withColumn("country", lit(country))
    )

    dfs.append(df)
```

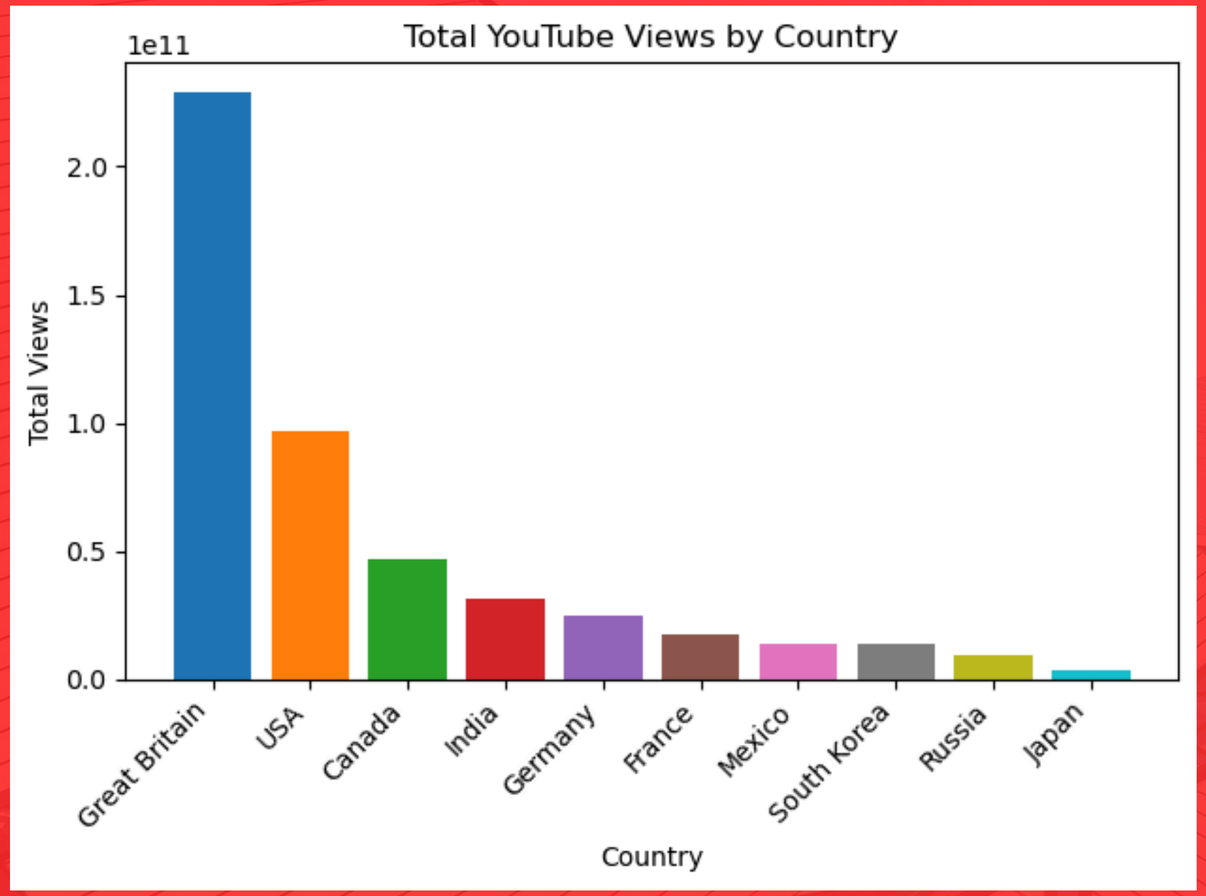
Data preparation

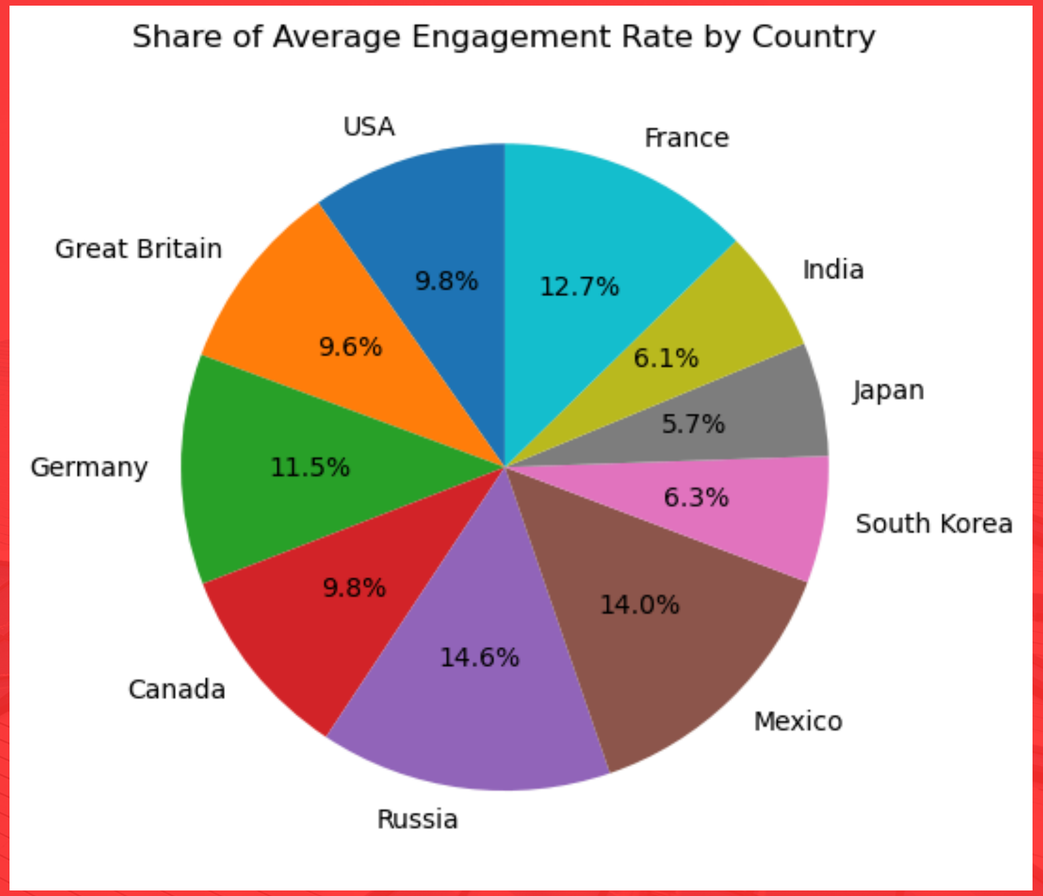
[22]:
```python
# Window 1: Top 3 videos per country

spark.sql("""
SELECT *
FROM (
    SELECT country, title, views,
           DENSE_RANK() OVER (PARTITION BY country ORDER BY views DESC) AS rank
    FROM youtube
)
WHERE rank <= 3
ORDER BY country, rank
""").show(truncate=False)
```

```
+-------------+-----------------------------------------------------------------------------+---------+----+
|country      |title                                                                        |views    |rank|
+-------------+-----------------------------------------------------------------------------+---------+----+
|Canada       |YouTube Rewind: The Shape of 2017 | #YouTubeRewind                           |137843120|1   |
|Canada       |YouTube Rewind: The Shape of 2017 | #YouTubeRewind                           |125431369|2   |
|Canada       |YouTube Rewind: The Shape of 2017 | #YouTubeRewind                           |113876217|3   |
|France       |YouTube Rewind: The Shape of 2017 | #YouTubeRewind                           |100911567|1   |
|France       |YouTube Rewind: The Shape of 2017 | #YouTubeRewind                           |75969469 |2   |
|France       |BTS (방탄소년단) 'FAKE LOVE' Official MV                                       |65396157 |3   |
|Germany      |YouTube Rewind: The Shape of 2017 | #YouTubeRewind                           |113876217|1   |
|Germany      |YouTube Rewind: The Shape of 2017 | #YouTubeRewind                           |100911567|2   |
|Germany      |Marvel Studios' Avengers: Infinity War Official Trailer                      |80360459 |3   |
|Great Britain|Nicky Jam x J. Balvin - X (EQUIS) | Video Oficial | Prod. Afro Bros & Jeon|424538912|1   |
|Great Britain|Nicky Jam x J. Balvin - X (EQUIS) | Video Oficial | Prod. Afro Bros & Jeon|413586699|2   |
|Great Britain|Nicky Jam x J. Balvin - X (EQUIS) | Video Oficial | Prod. Afro Bros & Jeon|402650804|3   |
|India        |YouTube Rewind: The Shape of 2017 | #YouTubeRewind                           |125432237|1   |
|India        |YouTube Rewind: The Shape of 2017 | #YouTubeRewind                           |113876217|2   |
|India        |YouTube Rewind: The Shape of 2017 | #YouTubeRewind                           |100911567|3   |
```

Data Analysis





Visualization

Key Insights

Main Findings:

- High views ≠ high engagement
- Great Britain leads in total views, but not always engagement
- Engagement varies significantly by country
- Trending behavior differs across regions
- YouTube Rewind: The Shape of 2017 is one of the top 3 videos trending in more than 5 countries.

country	category_id	video_count
Germany	24	15292
India	24	14511
Great Britain	10	13698
Canada	24	13451
Mexico	24	13361
Russia	22	10244
USA	24	9943
France	24	9819
Great Britain	24	9055
South Korea	24	8306
Mexico	22	8070
South Korea	25	6896
South Korea	22	6601
USA	10	6467
Germany	22	5988
Russia	24	5886
France	22	5719
Russia	25	5353
India	25	4649
Japan	24	4359

Challenges & Solutions

Challenges:

- Large CSV files with text fields
- Slow ingestion on Windows
- Duplicate records due to daily trending

Solutions:

- Incremental file loading
- Explicit data cleaning in Spark

Conclusion

- Spark is effective for Big Data analytics
- Engagement metrics add deeper insight
- Country-level analysis reveals meaningful patterns

THANK YOU