

# Project Statement for Milestone 4

Kongqueror

Ross Kugler, Huy Ky, Ben Bordon

## Overview:

At the end of Milestone 4 of the project, teams should have developed a prototype of an end-to-end application with a graphical user interface. Teams should have integrated the graphical user interface with a NoSQL database and Hadoop/Spark framework. In cases where teams are focused on a research topic, they should have validated their findings and compared them against the existing methods.

Teams should have updated their database schema and algorithms for robustness and scalability, and should have validated the scalability of their solution using a large dataset on distributed instances of a NoSQL database and Hadoop/Spark.

All team members are expected to make a significant contribution to the project milestone tasks.

## Project Report Topics:

The report should cover the following subtopics and answer the questions listed:

1. User Interface and Data Visualization:
  - a. Describe the user interface and data visualization components of the software. The user interface should be interactive, taking inputs from a user and providing outputs, i.e., results, to the user. The user interface should include data visualization.

Home Page: does a connection check with MongoDB and Spark. Has an overview of the dataset, including number of videos, number of edges. Has simple figures, including a pie-chart of video categories, and bar chart of top creators.

Network Statistics: allow users to view degree distributions and categorized statistics. User selects a sample size and runs on-demand queries. The result is then presented to user in tables, bar graphs, and pie charts

Top-K Queries: user chooses between Top Categories, Most Viewed Videos, and Highest Rated Videos. User selects parameters (K, minimum ratings) and requests on-demand queries. The results are shown with tables and bar graphs.

Range Queries: Filter videos using multiple criteria. Users send on-demand queries filter by duration, views, and category. Results are visualized using graphs and detailed in tables.

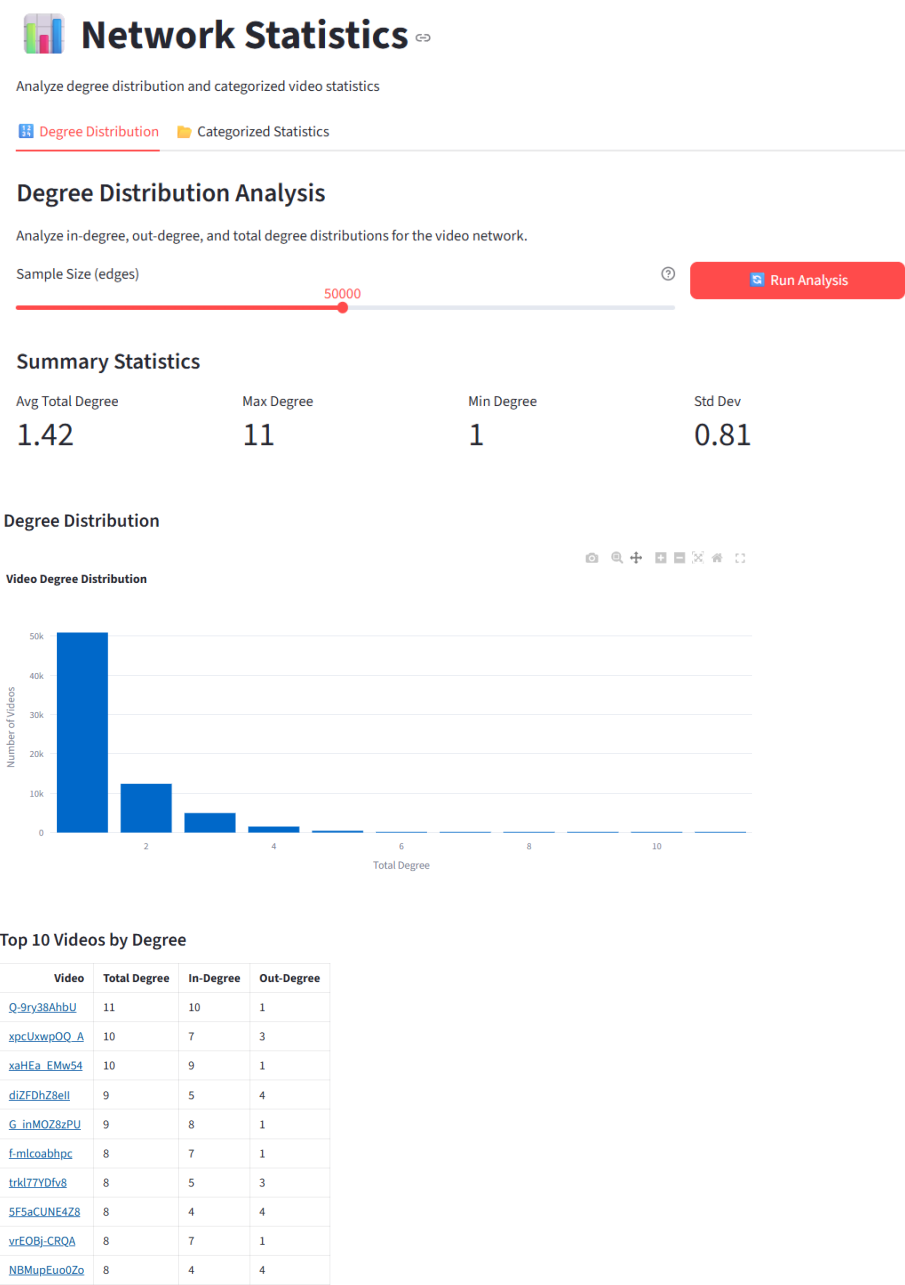
Graph Analytics: Compute PageRank scores and analyze influence.

Pattern Search: The tool allows the user to find patterns in the network. The three patterns we have implemented are related pairs, where video A recommends video B; video chains, where video (a)  $\rightarrow$  (b)  $\rightarrow$  (c); and common recommendations,

where video (a)  $\rightarrow$  (c)  $\leftarrow$  (b). In addition, the user can filter by category, set a maximum number of results, and configure the sample size (edges). After running the algorithm, the tool outputs the results in a table and in a graph visualizer. The user can also save their results into MongoDB.

2. User Queries and Results:
- a. Provide all user queries and their results to describe the overall functionality of the software.

Network Statistics:



Analyze degree distribution and categorized video statistics

[Degree Distribution](#) [Categorized Statistics](#)

## Categorized Statistics

View video statistics grouped by category, length, and view count.

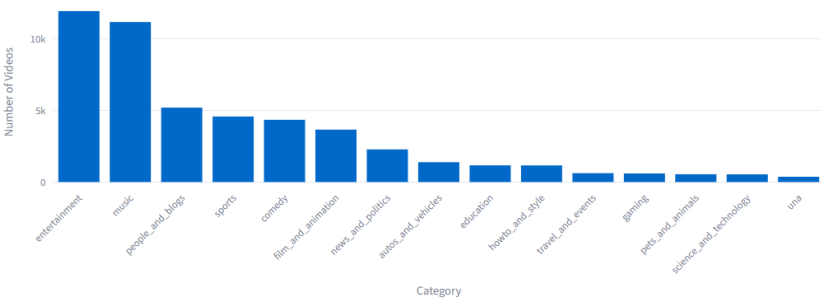
Sample Size (videos)



[Run Analysis](#)

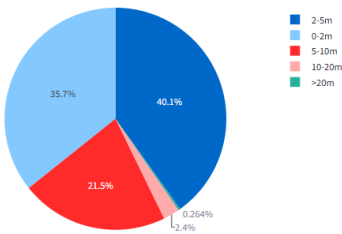
### Videos by Category

Top 15 Video Categories



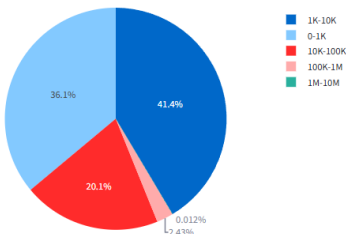
### Length Distribution

Videos by Duration



### View Count Distribution

Videos by View Count



## Top-K queries:



## Top-K Queries

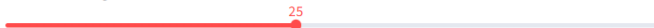
Find the most popular videos, categories, and highest rated content

[Top Categories](#) [Most Viewed Videos](#) [Highest Rated Videos](#)

### Top K Categories by Video Count

Find the categories with the most uploaded videos.

Number of Categories (K)



[Find Top Categories](#)

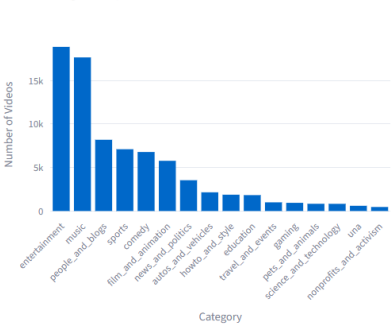
✓ Found top 25 categories!

Results Table

	Category	Video Count
1	entertainment	18871
2	music	17659
3	people_and_blogs	8192
4	sports	7103
5	comedy	6785
6	film_and_animation	5772
7	news_and_politics	3539
8	autos_and_vehicles	2149
9	howto_and_style	1875
10	education	1827

Distribution Chart

Top 25 Categories by Video Count



📁 Top Categories 📺 Most Viewed Videos ⭐ Highest Rated Videos

Top K Most Viewed Videos

Find the videos with the highest view counts.

Number of Videos (K)

15

5100

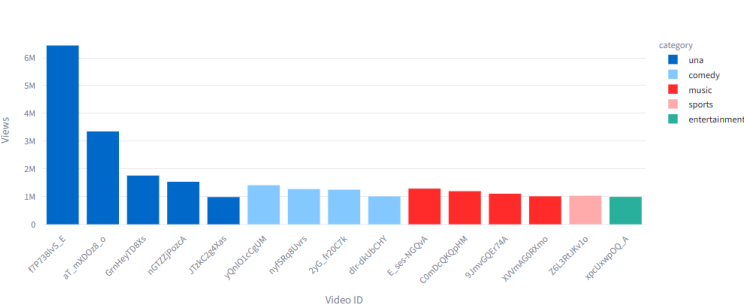
Find Most Viewed

✔ Found top 15 most viewed videos!

Results Table

Video	views	category	rating	num_ratings
f7P738lvS_E	6446580	una	4.34	3904
aT_mXDOu8_o	3349815	una	4.67	4370
GmHwyTDSks	1757440	una	4.69	2432
nGTZJPoscA	1534401	una	4.34	1463
yQnID1cCgUM	1410488	comedy	4.38	895
E_ses-NGQvA	1288157	music	4.86	1202
nyf58o8Uhrs	1270697	comedy	4.84	402
2yG_f20C7k	1249627	comedy	4.10	990
C0mDcQKQpHM	1195238	music	4.80	600
9JmVGQEr74A	1104160	music	4.76	834
Z6L3RtJKvIo	1031723	sports	4.53	730
XVImAG0Rkmo	1010363	music	4.90	757
dIc-dkUbCHY	1005276	comedy	3.05	185
xpcUowpDQ_A	988775	entertainment	4.86	7654
J7dC2e5Kas	982604	una	4.63	180

Top 15 Videos by View Count



📁 Top Categories 📺 Most Viewed Videos ⭐ Highest Rated Videos

Top K Highest Rated Videos

Find the videos with the highest ratings.

Number of Videos (K)

10

Minimum Ratings Required

500

Find Top Rated

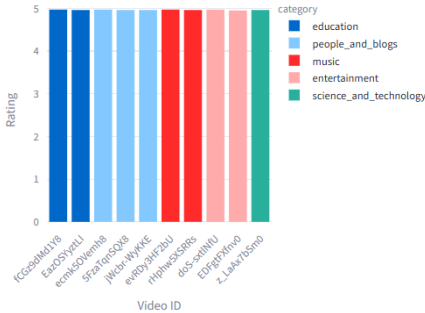
✔ Found top 10 highest rated videos (min 500 ratings)!

Results Table

Video	rating	num_ratings	views	category
<a href="#">fCGz9dMd1Y8</a>	4.97	533	3414	education
<a href="#">ecmk5QVemh8</a>	4.97	960	7804	people_and_blogs
<a href="#">evRDy3HF2bU</a>	4.97	698	5130	music
<a href="#">d0S-sxtlNfU</a>	4.97	502	16914	entertainment
<a href="#">5FzaTqnSQX8</a>	4.96	943	29991	people_and_blogs
<a href="#">rHphw5XSRRs</a>	4.96	698	7832	music
<a href="#">EazOSYztLI</a>	4.96	2023	19899	education
<a href="#">z_LaAx7bSm0</a>	4.96	660	5249	science_and_technology
<a href="#">jWcbr-WyKKE</a>	4.96	802	5068	people_and_blogs
<a href="#">EDFgtFXfny0</a>	4.95	1662	204052	entertainment

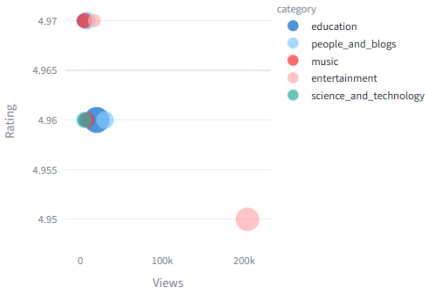
Rating Distribution

Top 10 Videos by Rating



Rating vs Views

Rating vs Views Correlation



Range queries:

Range Queries

Filter videos by category, duration, views, and other criteria

☒ Duration Range ☐ Views Range

Filter Videos by Duration

Find all videos in a specific category with duration within a specified range.

Select Category

comedy

Duration Range (seconds)

200

930

03600

Maximum Results

20

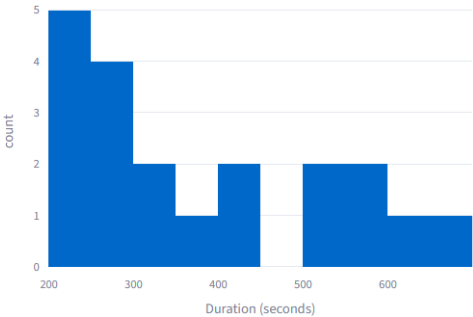
Search

Found 20 videos!

Results: comedy videos (200-930 sec)

Video	Duration	Views	Rating
<a href="#">MVbv6r_tKnE</a>	4:57	775457	4.75
<a href="#">jk2iR8mIZyM</a>	4:08	571747	4.70
<a href="#">P11J07k51ss</a>	4:16	571279	4.87
<a href="#">Wtposdpspn8</a>	3:21	482292	4.80
<a href="#">zn6c53V1LrE</a>	4:15	444657	4.70
<a href="#">Daq2GhJWmmI</a>	9:30	415215	4.70
<a href="#">EI_VtKutD30</a>	4:22	405018	4.19
<a href="#">6ihpfM0KXMU</a>	10:24	398407	4.85
<a href="#">tXhak_gPQgQ</a>	8:59	393957	4.57
<a href="#">Kv9SGYPO8gw</a>	5:58	391924	4.42
<a href="#">wJPVvk9Uju9s</a>	3:33	371730	4.83

Duration Distribution (comedy)

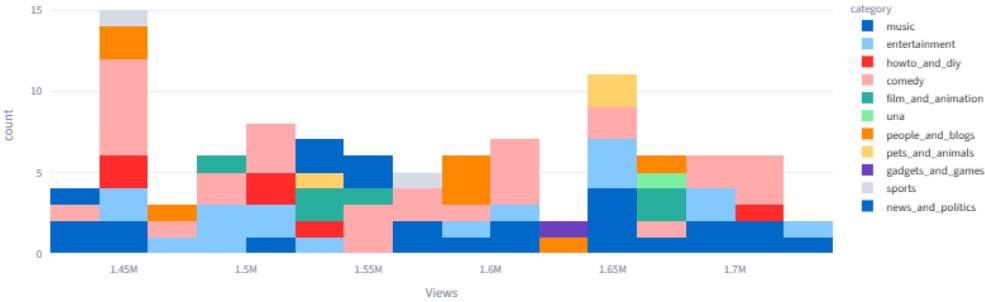


Range Query (View Count All categories 450,000 to 1740000 view)

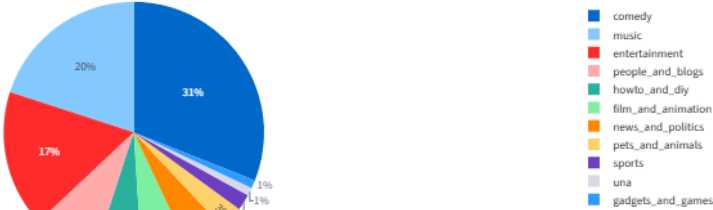
Results: Videos in all categories with 450,000-1,740,000 views

Video ID	Views	Rating	# Ratings	Category	Duration (s)
1 T0wQqy9tU		1727996	4.46	424 music	204
2 w2xZv6ZDko		1726429	0.00	0 entertainment	243
3 Mvz8H6iLQ		1718281	4.61	7815 music	139
4 STQbnhuuEM		1713974	4.61	2741 howto_and_diy	39
5 uetdP_fj9Mc		1713532	4.25	5806 comedy	44
6 0hkrudQ9Dw		1710477	3.68	3471 comedy	25
7 cuHvudQ9Dh		1707028	4.67	8973 comedy	187
8 FfHCBa9gr		1701447	4.63	4089 music	252
9 QHQ37dhu3Y		1693436	2.67	154 entertainment	302
10 Gd_Au8mpdQ		1689810	4.69	1121 music	208

Views Distribution



Category Distribution in Results



## Graph Analytics:

### Graph Analytics

Advanced network analysis using Spark GraphFrames for large-scale graph mining

 PageRank Analysis  Community Detection  Centrality Metrics  Network Visualization

#### Interactive Network Visualization

Visualize graph structure with insights from GraphFrames analysis.

Sample Size



Layout Algorithm

Color By



Show Labels

100



Spring

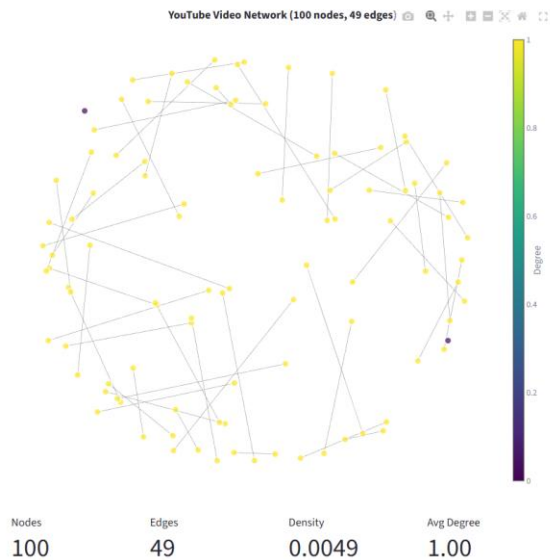


Degree



 Generate Visualization

Visualizing 100 nodes and 49 edges



## Centrality Metrics

Measure video importance using various centrality algorithms.

Sample Size (edges)

30000

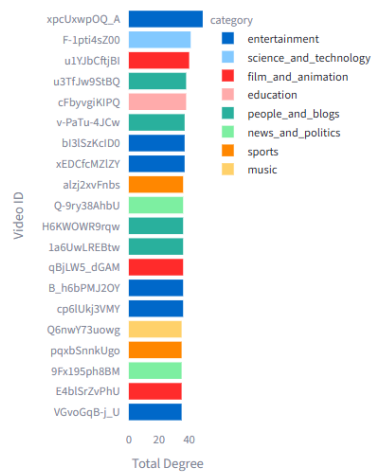
Calculate Centrality Metrics

Analyzing network: 20941 nodes, 30000 edges

Centrality metrics calculated!

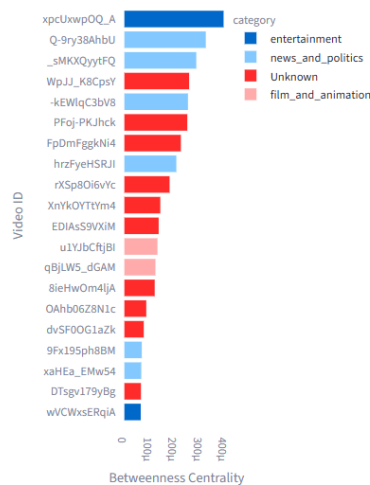
### Top Videos by Degree Centrality

Highest Degree Centrality



### Top Videos by Betweenness Centrality

Highest Betweenness (Bridge Videos)



Video	Total Degree	Category
<a href="#">xpcUxwpOQ_A</a>	49	entertainment
<a href="#">F-1pti4sZ00</a>	41	science_and_technology
<a href="#">u1YJbCftjBI</a>	40	film_and_animation
<a href="#">cFbyvgiKlPQ</a>	38	education
<a href="#">u3TfJw9StBQ</a>	38	people_and_blogs
<a href="#">xEDCfcMZlZY</a>	37	entertainment
<a href="#">bl3lSzKclD0</a>	37	entertainment
<a href="#">v-PaTu-4JCw</a>	37	people_and_blogs
<a href="#">Q-9ry38AhbU</a>	36	news_and_politics



## Community Detection

Identify clusters of highly connected videos using Label Propagation algorithm.

Sample Size (edges) ? Max Iterations ?

50000 - + 5 - +

[Detect Communities](#)

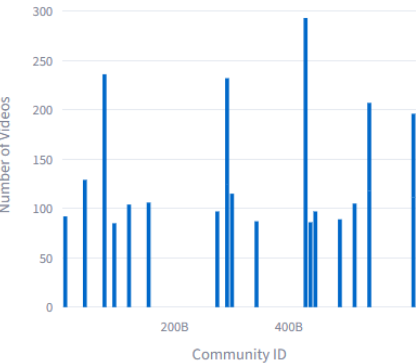
☐ Running community detection algorithm...

✓ Found 2372 communities!

## Community Size Distribution ↔ Community Size Categories



Top 20 Largest Communities



## PageRank Analysis

Identify the most influential videos in the network based on link structure.

Sample Size (edges) ? Reset Probability ? Max Iterations ?

30000 - + 0.15 5 - +

[Run PageRank Analysis](#)

☐ Running PageRank algorithm on video network...

Loaded 30,000 edges for analysis

Identified 20,941 unique videos

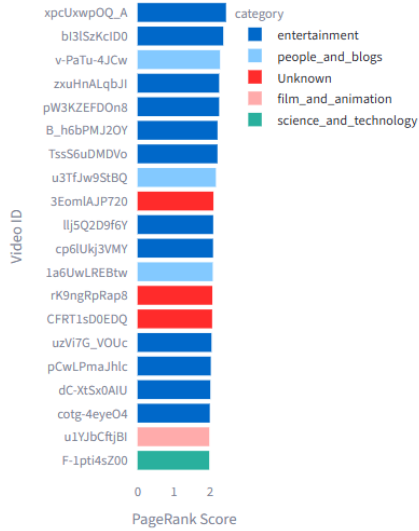
☐ Computing PageRank (max 5 iterations)...

✔ PageRank completed successfully!

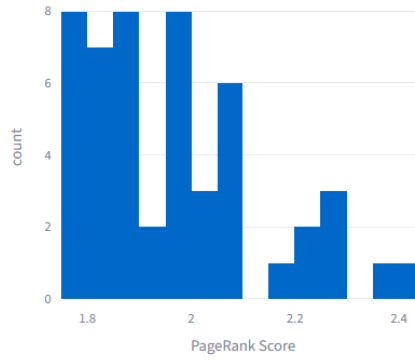
## 🏆 Top 20 Most Influential Videos

## 📊 PageRank Distribution

### Videos with Highest PageRank Scores



### PageRank Score Distribution (Top 50)

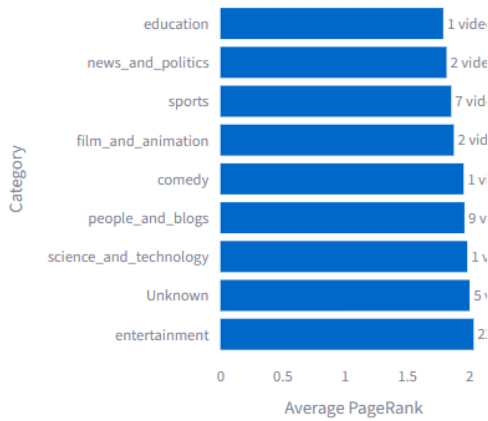


## 📁 PageRank by Category

### Average PageRank by Category

## 📁 PageRank by Category

### Average PageRank by Category



📄 Top 50 Videos by PageRank

Video	PageRank Score	Category	Uploader
<a href="#">xpcUxwpQQ_A</a>	2.444577	entertainment	meepmeepmeepow
<a href="#">bJ3ISzKcIDQ</a>	2.368115	entertainment	MzViSyon
<a href="#">v-PaTu-4JCw</a>	2.278786	people_and_blogs	bootylicious0714
<a href="#">pW3KZEFDO8</a>	2.254238	entertainment	MzViSyon
<a href="#">zxuHnALqbJl</a>	2.254238	entertainment	MzViSyon
<a href="#">B_h6bPMJ2OY</a>	2.210008	entertainment	tacksy
<a href="#">TssS6uDMDVo</a>	2.208477	entertainment	idance4me22
<a href="#">u3TfJw9StBQ</a>	2.163861	people_and_blogs	11Rohit11
<a href="#">3EomIAJP72Q</a>	2.096532	Unknown	Unknown
<a href="#">llj5Q2D9f6Y</a>	2.090472	entertainment	evr716
<a href="#">cp6lUkj3VMY</a>	2.086060	entertainment	fayakoxxx
<a href="#">1a6UwLREBtw</a>	2.077284	people_and_blogs	zocroc1234
<a href="#">rK9ngRpRap8</a>	2.066103	Unknown	Unknown
<a href="#">CFRT1sD0EDQ</a>	2.060772	Unknown	Unknown
<a href="#">uzVi7G_VOUC</a>	2.042572	entertainment	lanefan37

Pattern Search (Related Pairs)

About Pattern Search

This analysis finds patterns of connected videos in the recommendation network:

- **Related Pairs:**  $(a) \rightarrow (b)$  - Video A recommends Video B
- **Video Chains:**  $(a) \rightarrow (b) \rightarrow (c)$  - Recommendation chains
- **Common Recommendations:**  $(a) \rightarrow (c) \leftarrow (b)$  - Videos that share recommendations

These patterns help identify recommendation clusters and content relationships.

Search Configuration

Filter by Category

All Categories

Maximum Results

50

Pattern Type

Related Pairs (a→b)

Edge Sample Size

50000

Find Patterns

Found 50 results!

Related Pairs (a→b) (all categories)

Results Table

	Source Video (a)	Related Video (b)
1	→mklyh90bc	5rkuDqKGgw
2	→mklyh90bc	DVxvSBMTYA
3	→mklyh90bc	L_GUfJ_RLV
4	→mklyh90bc	QBC0nfS8lwg
5	→mklyh90bc	QNowKEZfRE
6	→mklyh90bc	SLPWwUJ_XIE
7	→mklyh90bc	UR2CqPMwUt
8	→mklyh90bc	WGb0m1Ud548
9	→mklyh90bc	_THardKR8BQ
10	→mklyh90bc	duMScvRTLJB

Pattern Visualization

Top Source Videos by Outgoing Links



Pattern Search (Video Chains)

### About Pattern Search

This analysis finds patterns of connected videos in the recommendation network:

- **Related Pairs:**  $(a) \rightarrow (b)$  - Video A recommends Video B
- **Video Chains:**  $(a) \rightarrow (b) \rightarrow (c)$  - Recommendation chains
- **Common Recommendations:**  $(a) \rightarrow (c) \leftarrow (b)$  - Videos that share recommendations

These patterns help identify recommendation clusters and content relationships.

### Search Configuration

Filter by Category: All Categories

Pattern Type: Video Chains (a→b→c)

Maximum Results: 50

Edge Sample Size: 50000

**Find Patterns**

Found 50 results!

#### Video Chains (a→b→c) (all categories)

##### Results Table

Video A	Video B	Video C
1 - GbUuHcSM	-8tfeEC-A	IwGVRQfQxkE
2 - GbUuHcSM	-8tfeEC-A	QwE0FJ7Tg
3 - GbUuHcSM	-8tfeEC-A	MuT4_WQw00
4 - GbUuHcSM	-8tfeEC-A	LYk7_J2Gdli
5 - GbUuHcSM	-8tfeEC-A	SpETa1DuDg
6 - HHVQakJSM	-4THz_yd40	n_n0DSuH26g
7 - HHVQakJSM	-4THz_yd40	mmAb-s-3uWA
8 - HHVQakJSM	-4THz_yd40	Mfpc0WV10A
9 - HHVQakJSM	-4THz_yd40	B1DQ8j4u8M

##### Pattern Visualization

###### Most Common Bridge Videos (B in A→B→C)



## Pattern Search (Common Recommendations)

### Pattern Search

Find subgraph patterns in the YouTube recommendation network

#### About Pattern Search

This analysis finds patterns of connected videos in the recommendation network:

- **Related Pairs:**  $(a) \rightarrow (b)$  - Video A recommends Video B
- **Video Chains:**  $(a) \rightarrow (b) \rightarrow (c)$  - Recommendation chains
- **Common Recommendations:**  $(a) \rightarrow (c) \leftarrow (b)$  - Videos that share recommendations

These patterns help identify recommendation clusters and content relationships.

#### Search Configuration

Filter by Category: All Categories

Pattern Type: Common Recommendations (a→c←b)

Maximum Results: 50

Edge Sample Size: 50000

**Find Patterns**

Found 50 results!

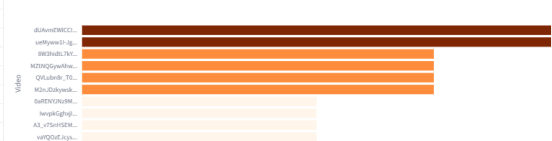
#### Common Recommendations (a→c←b) (all categories)

##### Results Table

Video A	Video B	Common Target (c)
1 - mkyh90Bc	-8tfeEC-A	r_Cu0RAEEge
2 - 0R69A3CVU	-ALmISUKMN04	BW0N4SL7KX
3 - 0R69A3CVU	-A1CvPpH8BA	BW0N4SL7KX
4 - 0R69A3CVU	-51w6RHF8c	BW0N4SL7KX
5 - 0R69A3CVU	-ALmISUKMN04	M2hL0dlyewk
6 - 0R69A3CVU	-A1CvPpH8BA	M2hL0dlyewk
7 - 0R69A3CVU	-49K0tqZG04	M2hL0dlyewk
8 - 0R69A3CVU	-ALmISUKMN04	M2hL0dlyewk

##### Pattern Visualization

###### Most Commonly Recommended Videos



### 3. Scalability:

- Describe how data is stored in a cluster configuration of a NoSQL database. Provide performance benchmarks for data ingestion/query in a cluster deployment. How do the results compare to the performance benchmarks on the reduced dataset or non-cluster deployment?

Our NoSQL database is stored locally with replica sets (through MongoDB config). We have not done sharding or cluster deployment yet.

- b. Describe how the algorithms are run on a cluster configuration of Hadoop/Spark. Provide performance benchmarks for algorithm in a cluster deployment. How do the results compare to the performance benchmarks on the reduced dataset or non-cluster deployment?

Our algorithms can run in Spark's standalone mode with multiple worker nodes based on number of cores of the system.

For testing purposes and simplicity, we are running our application in local mode. Our app will run in Standalone mode for the actual demo.

- c. What hardware (computers with CPU, memory and disk storage) did the team use to test the scalability of the solution?

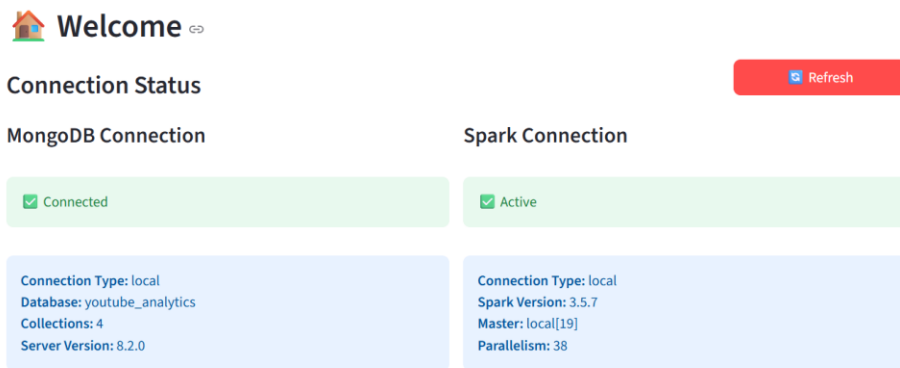
We all tested on our local computers. They have different numbers of logical clusters. These are the number of worker threads:

Ross: master: local[15]

Ben: master: local[7]

Harry: master: local[19]

Our Home Page shows the configuration:



4. Source Code:

- a. Provide the source code of your application prototype, including user interface, data ingestion, data query, and analytics algorithms, in a ZIP file. Make sure to **exclude** the dataset.

Included in submission and repo: [https://github.com/rk3026/Youtube\\_Analyzer](https://github.com/rk3026/Youtube_Analyzer)