

Youtube Video Analytics

Solution steps

By - Amit Shekhar Bongir

1. Kafka Connect and ADLS Gen2 Sink Connector

1. Create ADLS Gen2 storage *youtubevideoanalytics*

The screenshot displays the Azure portal interface for a newly created Data Lake Storage Gen2 account. The left sidebar shows the navigation menu with categories like Overview, Activity log, Tags, Diagnose and solve problems, Access Control (IAM), Data migration, Events, Storage Explorer (preview), Data storage, Security + networking, and Data management. The main content area is titled 'youtubevideoanalytics' and shows the account's essential information and properties.

Essentials

- Resource group (change): [NetworkWatcherRG](#)
- Location: Central India
- Subscription (change): [Pay-As-You-Go](#)
- Subscription ID: [REDACTED]
- Disk state: Available
- Performance/Access tier: Standard/Hot
- Replication: Locally-redundant storage (LRS)
- Account kind: StorageV2 (general purpose v2)
- Provisioning state: Succeeded
- Created: 10/7/2021, 5:33:53 pm

Tags (change): [project : youtube-video-analytics](#) [year : 2021](#)

Properties | Monitoring | Capabilities (5) | Recommendations | Tutorials | Developer Tools

Data Lake Storage

Property	Value
Hierarchical namespace	Enabled
Default access tier	Hot
Blob public access	Enabled
Blob soft delete	Disabled
Container soft delete	Enabled (7 days)
Versioning	Disabled
Change feed	Disabled
NFS v3	Disabled

File service

Property	Value
Large file share	Disabled
Active Directory	Not configured
Soft delete	Enabled (7 days)
Share capacity	5 TiB

Queue service

Property	Value
CMK support	Disabled

Table service

Security

Property	Value
Require secure transfer for REST API operations	Enabled
Storage account key access	Enabled
Minimum TLS version	Version 1.2
Infrastructure encryption	Disabled

Networking

Property	Value
Allow access from	All networks
Number of private endpoint connections	0
Network routing	Microsoft network routing
Access for trusted Microsoft services	Yes

2. Start Confluent Platform

```
(base) abcoep@amitsb: /home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test
→ confluent local services start
The local commands are intended for a single-node development environment only,
NOT for production usage. https://docs.confluent.io/current/cli/index.html

Using CONFLUENT_CURRENT: /tmp/confluent.414267
ZooKeeper is [UP]
Kafka is [UP]
Schema Registry is [UP]
Kafka REST is [UP]
Connect is [UP]
ksqlDB Server is [UP]
Control Center is [UP]
(base) abcoep@amitsb: /home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test
```

3. Read *category_title.csv* and publish into topic *category_title*

```
→ confluent local services connect connector load csv_spooldir --config csv_spooldir.properties
The local commands are intended for a single-node development environment only,
NOT for production usage. https://docs.confluent.io/current/cli/index.html

{
  "name": "csv_spooldir",
  "config": {
    "connector.class": "com.github.jcustenborder.kafka.connect.spooldir.SpooldirCsvSourceConnector",
    "csv.first.row.as.header": "true",
    "error.path": "/home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test/error",
    "finished.path": "/home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test/YoutubeVideoAnalytics",
    "halt.on.error": "false",
    "input.file.pattern": "category_title.csv",
    "input.path": "/home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test/data",
    "schema.generation.enabled": "true",
    "tasks.max": "1",
    "topic": "category_title",
    "name": "csv_spooldir"
  },
  "tasks": [],
  "type": "source"
}
(base) abcoep@amitsb: /home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test
```

4. Verify data is pushed into the topic

```
(base) abcoep@amitsb: /home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test
→ confluent local services connect connector status csv_spooldir
The local commands are intended for a single-node development environment only,
NOT for production usage. https://docs.confluent.io/current/cli/index.html

{
  "name": "csv_spooldir",
  "connector": {
    "state": "RUNNING",
    "worker_id": "127.0.0.1:8083"
  },
  "tasks": [
    {
      "id": 0,
      "state": "RUNNING",
      "worker_id": "127.0.0.1:8083"
    }
  ],
  "type": "source"
}
(base) abcoep@amitsb: /home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test

(base) abcoep@amitsb: /home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test
→ kafka-avro-console-consumer --topic category_title --from-beginning --bootstrap-server localhost:9092
{"category_id":{"string":"1"},"category":{"string":"Film & Animation"}}
{"category_id":{"string":"2"},"category":{"string":"Autos & Vehicles"}}
{"category_id":{"string":"10"},"category":{"string":"Music"}}
{"category_id":{"string":"15"},"category":{"string":"Pets & Animals"}}
{"category_id":{"string":"17"},"category":{"string":"Sports"}}
{"category_id":{"string":"18"},"category":{"string":"Short Movies"}}
{"category_id":{"string":"19"},"category":{"string":"Travel & Events"}}
{"category_id":{"string":"20"},"category":{"string":"Gaming"}}
{"category_id":{"string":"21"},"category":{"string":"Videoblogging"}}
{"category_id":{"string":"22"},"category":{"string":"People & Blogs"}}
{"category_id":{"string":"23"},"category":{"string":"Comedy"}}
{"category_id":{"string":"24"},"category":{"string":"Entertainment"}}
{"category_id":{"string":"25"},"category":{"string":"News & Politics"}}
{"category_id":{"string":"26"},"category":{"string":"Howto & Style"}}
{"category_id":{"string":"27"},"category":{"string":"Education"}}
{"category_id":{"string":"28"},"category":{"string":"Science & Technology"}}
{"category_id":{"string":"29"},"category":{"string":"Nonprofits & Activism"}}
{"category_id":{"string":"30"},"category":{"string":"Movies"}}
{"category_id":{"string":"31"},"category":{"string":"Anime/Animation"}}
{"category_id":{"string":"32"},"category":{"string":"Action/Adventure"}}
{"category_id":{"string":"33"},"category":{"string":"Classics"}}
{"category_id":{"string":"34"},"category":{"string":"Comedy"}}
{"category_id":{"string":"35"},"category":{"string":"Documentary"}}
{"category_id":{"string":"36"},"category":{"string":"Drama"}}
{"category_id":{"string":"37"},"category":{"string":"Family"}}
{"category_id":{"string":"38"},"category":{"string":"Foreign"}}
{"category_id":{"string":"39"},"category":{"string":"Horror"}}
{"category_id":{"string":"40"},"category":{"string":"Sci-Fi/Fantasy"}}
{"category_id":{"string":"41"},"category":{"string":"Thriller"}}
{"category_id":{"string":"42"},"category":{"string":"Shorts"}}
{"category_id":{"string":"43"},"category":{"string":"Shows"}}
{"category_id":{"string":"44"},"category":{"string":"Trailers"}}
```

5. Consume topic `category_title` with ADLS Gen2 Sink Connector

```
(base) abcoep@amitsb: /home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test
→ confluent local services connect connector load datalake-sink --config datalake.properties
The local commands are intended for a single-node development environment only,
NOT for production usage. https://docs.confluent.io/current/cli/index.html

{
  "name": "datalake-sink",
  "config": {
    "avro.codec": "snappy",
    "azure.datalake.gen2.account.name": "youtubevideoanalytics",
    "azure.datalake.gen2.client.id": " ",
    "azure.datalake.gen2.client.key": " ",
    "azure.datalake.gen2.token.endpoint": "https://login.microsoftonline.com/ /oauth2/token",
    "confluent.topic.bootstrap.servers": "localhost:9092",
    "confluent.topic.replication.factor": "1",
    "connector.class": "io.confluent.connect.azure.datalake.gen2.AzureDataLakeGen2SinkConnector",
    "flush.size": "33",
    "format.class": "io.confluent.connect.azure.storage.format.avro.AvroFormat",
    "tasks.max": "1",
    "topics": "category_title",
    "name": "datalake-sink"
  },
  "tasks": [],
  "type": "sink"
}
```

6. Verify `category_title` Avro data stored in ADLS Gen2 container `topics`

Home > Storage accounts > youtubevideoanalytics >

topics ...

Container

Search (Ctrl+/) « Upload + Add Directory Refresh Rename Delete Change tier Acquire lease Break lease

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

Authentication method: Access key (Switch to Azure AD User Account)

Location: topics / category_title / partition=0

Search blobs by prefix (case-sensitive)

Name	Modified	Access tier	Blob type	Size	Lease state	
<input type="checkbox"/> [.]						...
<input type="checkbox"/> category_title+0+0000000000.avro	7/17/2021, 1:06:16 AM	Hot (Inferred)	Block blob	892 B	Available	...
<input type="checkbox"/> category_title+0+0000000033.avro	7/17/2021, 1:06:16 AM	Hot (Inferred)	Block blob	0 B	Available	...

7. Similarly load data from USvideos.csv. Unload Kafka CSV Connector and load again as schema is different now.

```
(base) abcoep@amitsb: /home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test
→ confluent local services connect connector unload csv_spooldir
The local commands are intended for a single-node development environment only,
NOT for production usage. https://docs.confluent.io/current/cli/index.html

Success.
(base) abcoep@amitsb: /home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test
```

```
(base) abcoep@amitsb: /home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test
→ confluent local services connect connector load csv_spooldir --config csv_spooldir.properties
The local commands are intended for a single-node development environment only,
NOT for production usage. https://docs.confluent.io/current/cli/index.html

{
  "name": "csv_spooldir",
  "config": {
    "connector.class": "com.github.jcstenborder.kafka.connect.spooldir.SpooldirCsvSourceConnector",
    "csv.first.row.as.header": "true",
    "error.path": "/home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test/error",
    "finished.path": "/home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test/YoutubeVideoAnalytics",
    "halt.on.error": "false",
    "input.file.pattern": "USvideos.csv",
    "input.path": "/home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test/data",
    "schema.generation.enabled": "true",
    "tasks.max": "1",
    "topic": "USvideos",
    "name": "csv_spooldir"
  },
  "tasks": [],
  "type": "source"
}
(base) abcoep@amitsb: /home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test
→ confluent local services connect connector status csv_spooldir
The local commands are intended for a single-node development environment only,
NOT for production usage. https://docs.confluent.io/current/cli/index.html
```

```
{
  "name": "csv_spooldir",
  "connector": {
    "state": "RUNNING",
    "worker_id": "127.0.0.1:8083"
  },
  "tasks": [
    {
      "id": 0,
      "state": "RUNNING",
      "worker_id": "127.0.0.1:8083"
    }
  ]
}
```

```
(base) abcoep@amitsb: /home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test
→ confluent local services connect connector load datalake-sink --config datalake.properties
The local commands are intended for a single-node development environment only,
NOT for production usage. https://docs.confluent.io/current/cli/index.html

{
  "name": "datalake-sink",
  "config": {
    "avro.codec": "snappy",
    "azure.datalake.gen2.account.name": "youtubevideoanalytics",
    "azure.datalake.gen2.client.id": " ",
    "azure.datalake.gen2.client.key": " ",
    "azure.datalake.gen2.token.endpoint": "https://login.microsoftonline.com/ /oauth2/token",
    "confluent.topic.bootstrap.servers": "localhost:9092",
    "confluent.topic.replication.factor": "1",
    "connector.class": "io.confluent.connect.azure.datalake.gen2.AzureDataLakeGen2SinkConnector",
    "flush.size": "40948",
    "format.class": "io.confluent.connect.azure.storage.format.avro.AvroFormat",
    "tasks.max": "1",
    "topics": "USvideos",
    "name": "datalake-sink"
  },
  "tasks": [],
  "type": "sink"
}
```

Home

>

Storage accounts

>

youtubevideanalytics

topics

Container

Search (Ctrl+I)

Upload

Add Directory

Refresh

Rename

Delete

Change tier

Acquire lease

Break lease

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

Authentication method: Access key (Switch to Azure AD User Account)

Location: topics / USVideos / partition=0

Search blobs by prefix (case-sensitive)

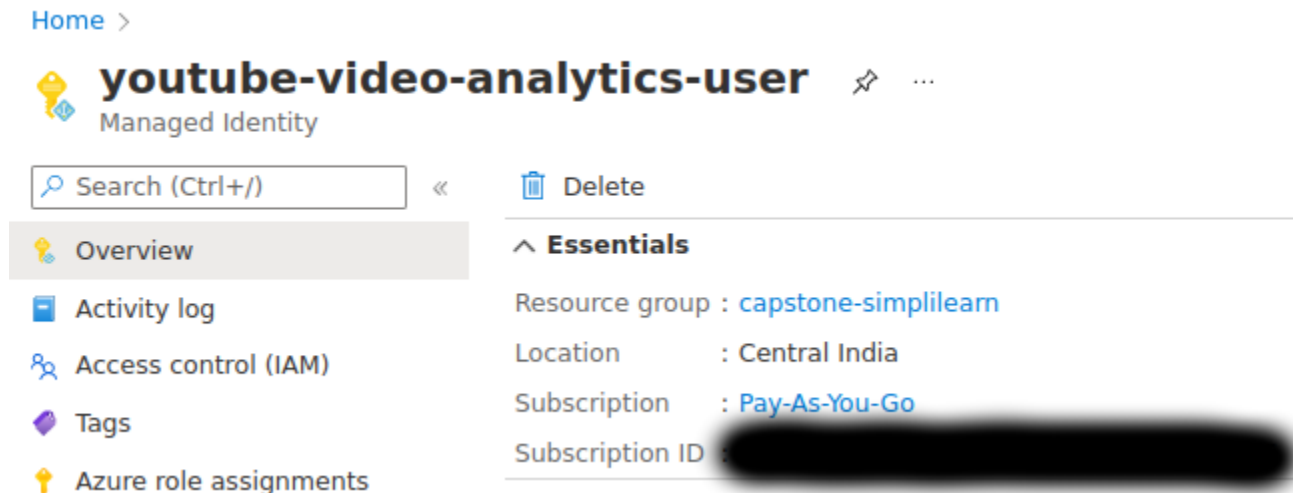
Name	Modified	Access tier	Blob type	Size	Lease state	
<input type="checkbox"/> [..]						...
<input type="checkbox"/> USVideos+0+0000000000.avro	7/17/2021, 1:55:12 AM	Hot (Inferred)	Block blob	10.57 MiB	Available	...
<input type="checkbox"/> USVideos+0+0000040948.avro	7/17/2021, 1:55:12 AM	Hot (Inferred)	Block blob	0 B	Available	...

```
(base) abcoep@amitsh: /home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test
→ confluent local services stop
The local commands are intended for a single-node development environment only,
NOT for production usage. https://docs.confluent.io/current/cli/index.html

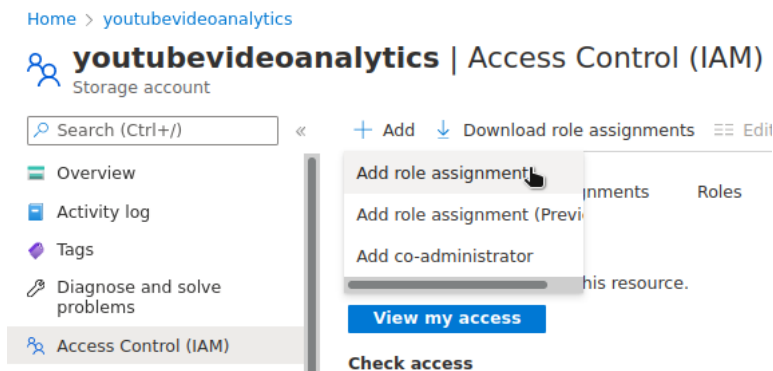
Using CONFLUENT_CURRENT: /tmp/confluent.599572
Stopping Control Center
Control Center is [DOWN]
Stopping ksqlDB Server
ksqlDB Server is [DOWN]
Stopping Connect
Connect is [DOWN]
Stopping Kafka REST
Kafka REST is [DOWN]
Stopping Schema Registry
Schema Registry is [DOWN]
Stopping Kafka
Kafka is [DOWN]
Stopping ZooKeeper
ZooKeeper is [DOWN]
(base) abcoep@amitsh: /home/abcoep/Amit_HD/IT/Big_Data/kafka-spooldir-test
```


2. Creating Azure HDInsight Spark cluster

1. Create user *youtube-video-analytics-user*



2. Grant 'Storage Blob Data Owner' role of storage account *youtubevideoanalytics* to *youtube-video-analytics-user*



Add role assignment ×

Role ⓘ

Storage Blob Data Owner ⓘ ▼

Assign access to ⓘ

User, group, or service principal ▼

Select ⓘ

youtube-video-analytics-user

No users, groups, or service principals f...

Selected members:



youtube-video-analytic:

[Remove](#)

3. Create HDInsight Spark cluster

Microsoft Azure

Search resources, services, and docs (G+)

[Home](#) >

Create HDInsight cluster

New to HDInsight? Get started with our [training resources](#).
Create a managed HDInsight cluster. Select from Spark, Kafka, Hadoop, Storm, and more. [Learn more](#)

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription *

Pay-As-You-Go

Resource group *

capstone-simplilearn

[Create new](#)

Cluster details

Name your cluster, pick a region, and choose a cluster type and version. [Learn more](#)

Cluster name *

youtube-video-analytics

Region *

Central India

Cluster type *

Spark
[Change](#)

Version *

Spark 2.4 (HDI 4.0)

Set *topics* as the container for the cluster's storage

Assign *youtube-video-analytics-user* to represent the cluster

Home

 >

Create HDInsight cluster

Basics

Storage

Security + networking

Configuration + pricing

Tags

Review + create

Select or create storage accounts that will be used for the cluster's logs, job input, and job output. Configure the cluster's access to these accounts, if needed.

Primary storage

Select or create a storage account that will be the default location for cluster logs and other output.

Primary storage type *

Azure Data Lake Storage Gen2

Primary storage account *

youtubevideoanalytics

Filesystem * ⓘ

topics

Identity

Select a user-assigned managed identity to represent the cluster for Azure Data Lake Gen2 Storage account access. Only identities with access to the selected storage account are listed. Assign the managed identity to the 'Storage Blob Data Owner' role on the storage account. [Learn more](#)

User-assigned managed identity * ⓘ

youtube-video-analytics-user

Enable Load-based Autoscaling

[Home](#) > [HDInsight clusters](#) >

Create HDInsight cluster ...

Configure your cluster's size and performance, and view estimated cost information.

The cost estimate represented in the table does not include subscription discounts or costs related to storage, networking, or data transfer.

i This configuration will use 9 to 11 of 12 available cores in the Central India 1.
[View cores usage](#)

+ Add application

Node type	Node size	Number ...	Estimated c...
Head node	E2 V3 (2 Cores, 16 GB RAM), 12.61 l... ▼	2	25.22 INR
Zookeeper n...	D1 v2 (1 Core, 3.5 GB RAM), 7.31 IN... ▼	3	21.94 INR
Worker node	E2 V3 (2 Cores, 16 GB RAM), 12.61 l... ▼	1 ✓	
<input checked="" type="checkbox"/> Enable autoscale Learn More	<div>Autoscale type <input checked="" type="radio"/> Load-based <input type="radio"/> Schedule-based</div> <div>Min: 1 ✓ Max: 2 ✓ 12.61 to 25.22 INR</div> <div>Load-based autoscale will scale the number of worker nodes used based on the cluster's activity.</div>		

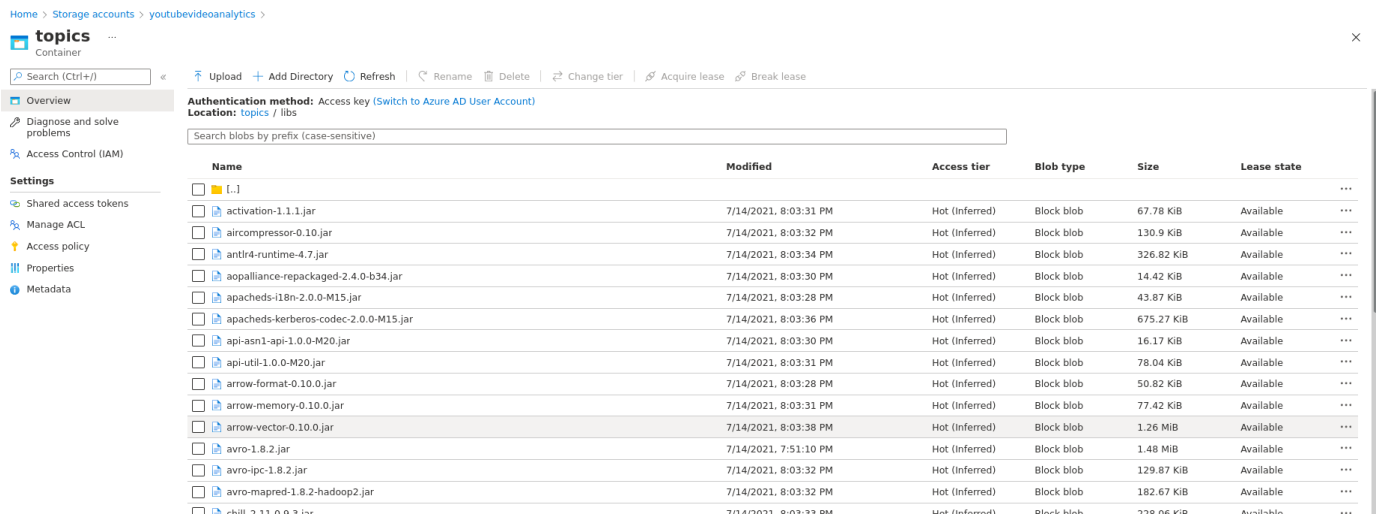
Total estimated cost/hour 59.76 to 72.37 INR

Script actions

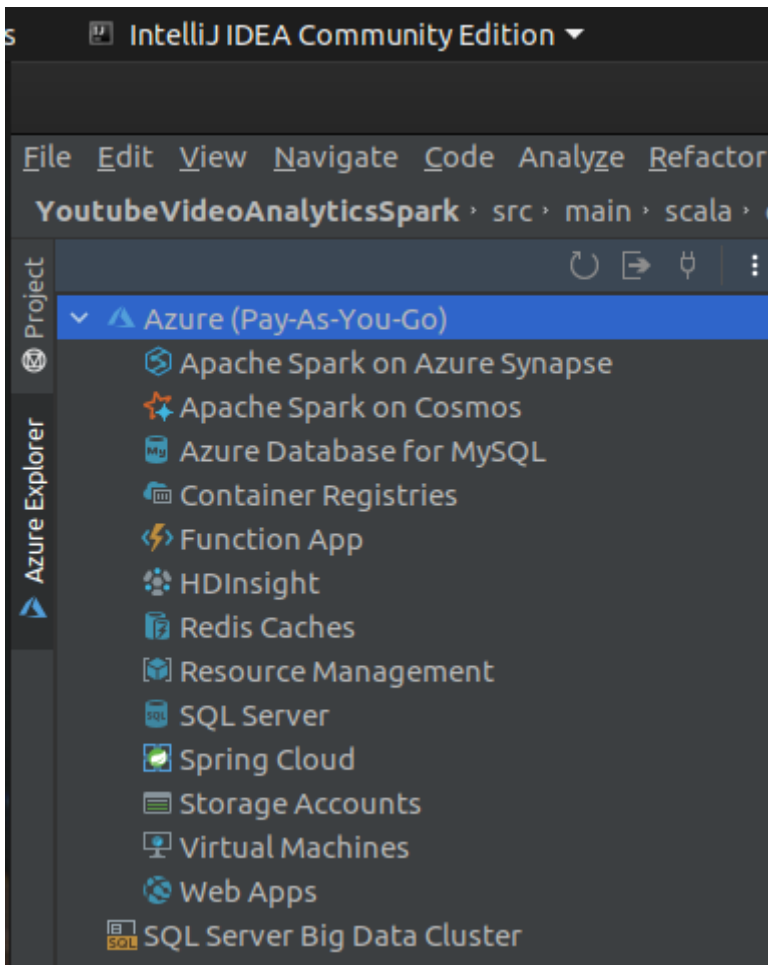
Use script actions to run custom PowerShell or Bash scripts on cluster nodes during cluster provisioning. [Learn about script actions](#)

3. Running YoutubeVideoAnalytics Spark Scala application on HDInsight cluster

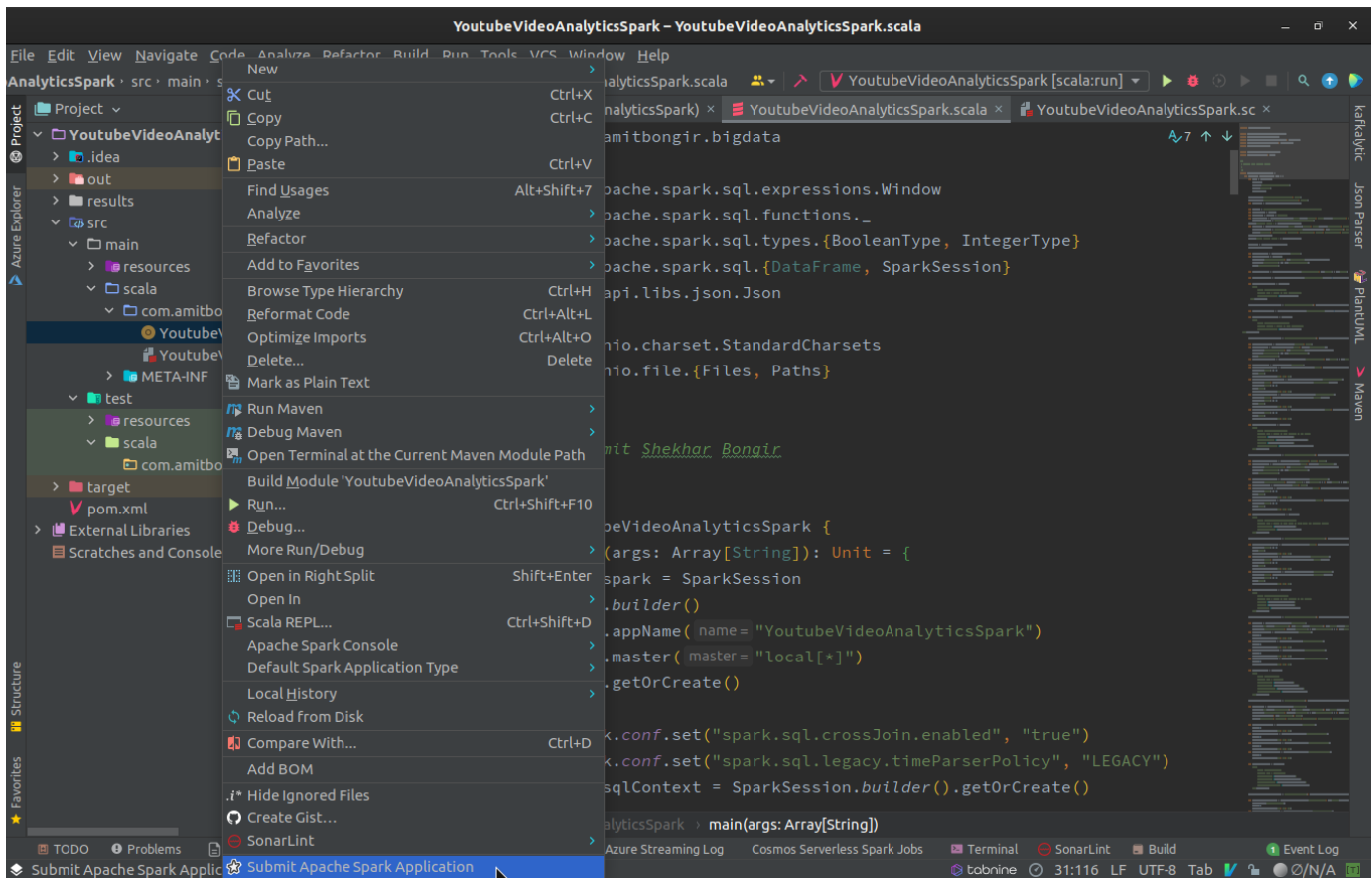
1. Copy all dependencies of the application to libs folder in *topics* container



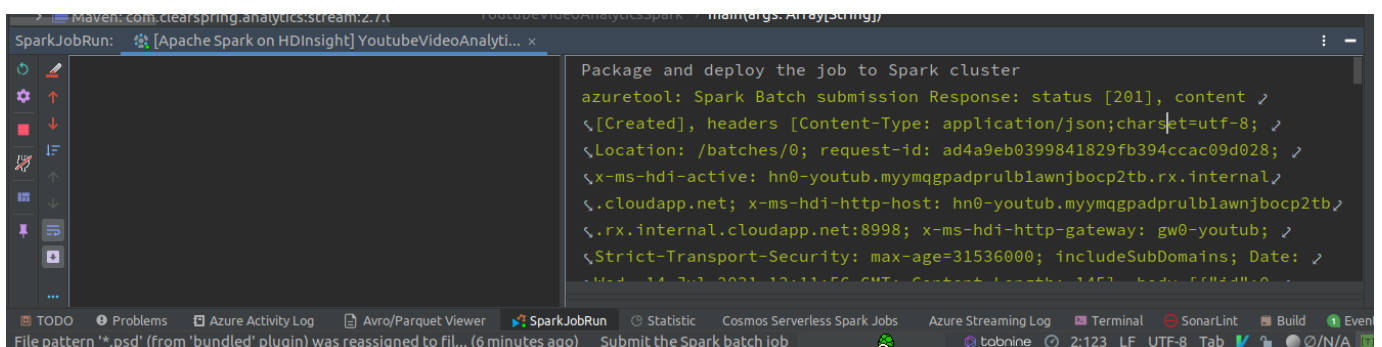
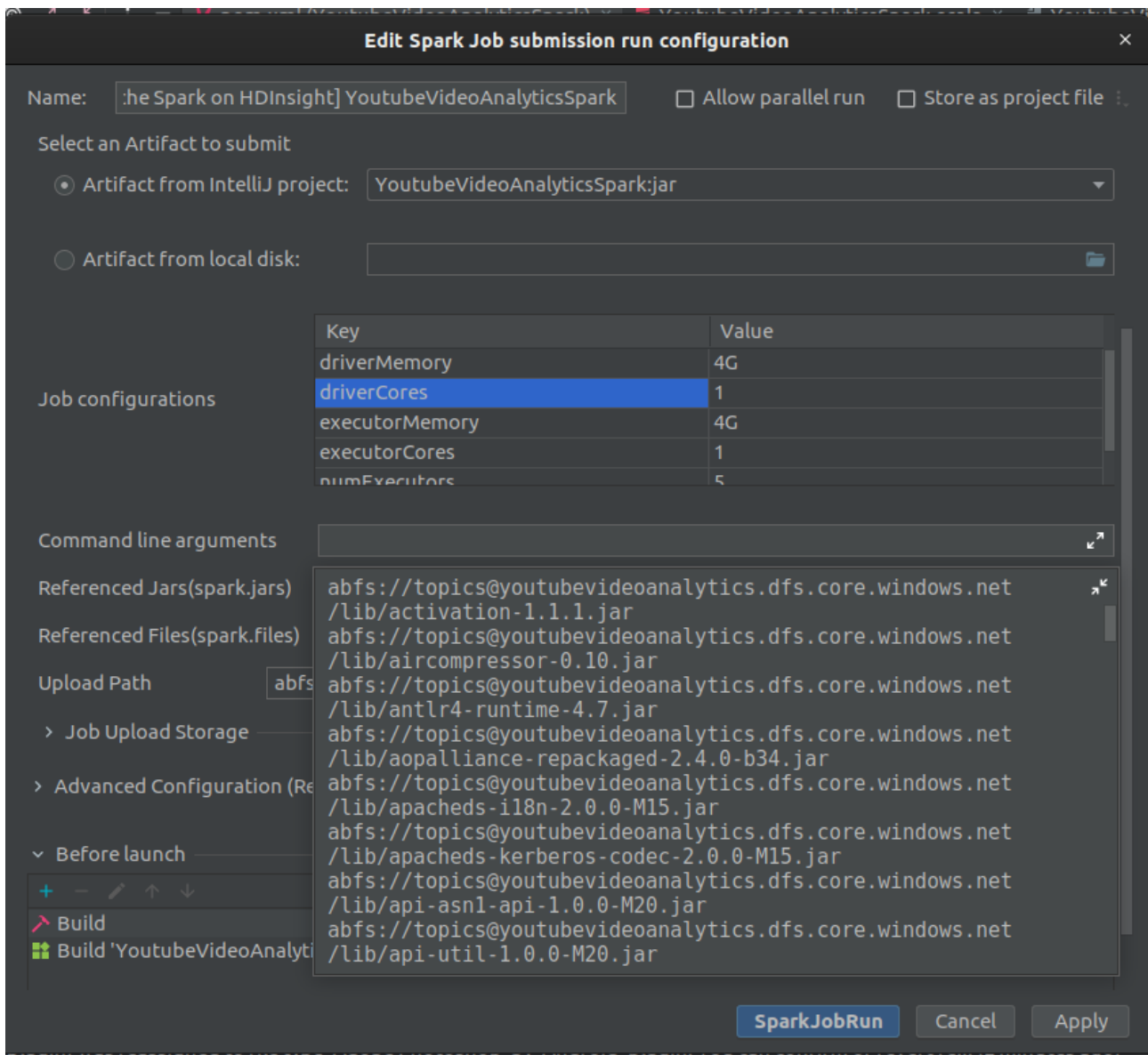
2. After installing Azure Toolkit for IntelliJ plugin in the IDE, sign into the Azure Explorer



3. Submit the application to the HDInsight cluster



4. Add Reference jars previously copied to the libs folder in topics container and run



Category count: 32

root

```
|-- video_id: string (nullable = true)
|-- category_id: string (nullable = true)
|-- category: string (nullable = true)
|-- trending_date: string (nullable = true)
|-- title: string (nullable = true)
|-- channel_title: string (nullable = true)
|-- publish_time: string (nullable = true)
|-- tags: string (nullable = true)
|-- views: string (nullable = true)
|-- likes: string (nullable = true)
|-- dislikes: string (nullable = true)
|-- comment_count: string (nullable = true)
|-- comments_disabled: string (nullable =
true)
|-- ratings_disabled: string (nullable =
true)
```

Video records count: 40948

root

```
|-- video_id: string (nullable = true)
|-- category_id: string (nullable = true)
|-- category: string (nullable = true)
|-- trending_date: date (nullable = true)
|-- title: string (nullable = true)
```

Problems Azure Activity Log Avro/Parquet Viewer SparkJobRun
are up-to-date (3 minutes ago) Spark batch job com.amitbong

5. After completion, verify the results stored in the *results* folder in *topics* container

↑ Upload

✚ Add Directory

↻ Refresh

|

🔄 Rename

🗑 Delete

|

↔ Change tier










Authentication method:

Access key [\(Switch to Azure AD User Account\)](#)

Location:

topics / results

Search blobs by prefix (case-sensitive)

Name	Modified	Access
<input type="checkbox"/>  [..]		
<input type="checkbox"/>  bottom_3_user_interaction		
<input type="checkbox"/>  top_3_by_category_monthly		
<input type="checkbox"/>  top_3_by_category_yearly		
<input type="checkbox"/>  top_3_categories		
<input type="checkbox"/>  top_3_channels		
<input type="checkbox"/>  top_3_likes_dislikes_ratio_monthly		
<input type="checkbox"/>  top_3_user_interaction		
<input type="checkbox"/>  top_3_videos		

↑ Upload

✚ Add Directory

↻ Refresh

|

🔄 Rename

🗑 Delete

|

↔ Change tier

🔒 Acquire lease

🔓 Break lease





Authentication method:

Access key [\(Switch to Azure AD User Account\)](#)

Location:

topics / results / top_3_channels

Search blobs by prefix (case-sensitive)

Name	Modified	Access tier	Blob type	Size	Lease state
<input type="checkbox"/>  [..]					...
<input type="checkbox"/>  _SUCCESS	7/15/2021, 2:18:56 AM	Hot (Inferred)	Block blob	0 B	Available ...
<input type="checkbox"/>  part-00000-c63f9384-dbc5-427d-a035-f3d512169e17-...	7/15/2021, 2:18:55 AM	Hot (Inferred)	Block blob	0 B	Available ...
<input type="checkbox"/>  part-00001-c63f9384-dbc5-427d-a035-f3d512169e17-...	7/15/2021, 2:18:56 AM	Hot (Inferred)	Block blob	2.19 KiB	Available ...

«

Authentication method:


Access key [\(Switch to Azure AD User Account\)](#)


Location:


topics / results / top_3_channels


Search blobs by prefix...

Name

☐  **[..]** ...

☐  **_SUCCESS** ...

☐  **part-00000-c63f9384-d...** ...

☒  **part-00001-c63f9384-d...** ...

results/top_3_channels/part-00001-c63f9384-dbc5-427d-a0...

Blob

📄 Save

✕ Discard

↓ Download

↻ Refresh

|

🗑 Delete

Overview

Versions

Edit

Generate SAS

1 [{"dislikes_ratio":{"channel_title":"Rob Andretti","like_dislike_ratio":-1.0,"total_comment_count":

2

4. Creating Cosmos database and container

[Home](#) > [Azure Cosmos DB](#) > [Select API option](#) >

Create Azure Cosmos DB Account - Core (SQL)

Basics Global Distribution Networking Backup Policy Encryption

Azure Cosmos DB is a fully managed NoSQL database service for building scalable database, multiple containers included. [Learn more](#)

Project Details

Select the subscription to manage deployed resources and costs. Use resource group

Subscription *	<input type="text" value="Pay-As-You-Go"/>
Resource Group *	<input type="text" value="capstone-simplilearn"/> Create new

Instance Details

Account Name *	<input type="text" value="youtube-video-analytics"/>
Location *	<input type="text" value="(Asia Pacific) Central India"/>
Capacity mode ⓘ	<input type="radio"/> Provisioned throughput <input checked="" type="radio"/> Serverless Learn more about capacity mode

New Database ×

* Database id ⓘ

New Container ×

* Database id ⓘ

☐ Create new ☒ Use existing

* Container id ⓘ

* Partition key ⓘ

Unique keys ⓘ

+ Add unique key

> Advanced

5. Copying results/insights JSON files from topics container into insights Cosmos DB using Azure Data Factory

[Home](#) > [Data factories](#) >

Create Data Factory ...

Basics Git configuration Networking Advanced Tags Review + create

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * ⓘ

Pay-As-You-Go

Resource group * ⓘ

capstone-simplilearn

[Create new](#)

Instance details

Region * ⓘ

Central India

Name * ⓘ

youtube-video-analytics

Version * ⓘ

V2

Copy Data tool


- 1 Properties
- 2 Source
- 3 Target
- 4 Settings
- 5 Review and finish

Use Copy Data Tool to perform a one-time or scheduled data load from 90+ data sources. Follow the wizard experience to specify your data loading settings, and let the Copy Data Tool generate the linked services. [Learn more](#)


Properties

Select copy data task type and configure task schedule

Task type



Built-in copy task
You will get single pipeline to copy data from 90+ data source easily.



Metadata-driven copy task (Preview)
Metadata is required to be stored in external control tables to load data at large-scale.

You will get single pipeline to quickly copy objects from data source store to destination in a very intuitive manner.

Task cadence or task schedule *

☒ Run once now ☐ Schedule ☐ Tumbling window

Microsoft Azureyoutube-video-analytics

amitbongir@yahoo.in

DEFAULT DIRECTORY

»

Copy Data tool

1 Properties

2 Source

3 Dataset

4 Configuration

5 Target

6 Settings

7 Review and finish

Source data store

Specify the source data store for the copy task. You can use an existing data store c

Source type

Azure Data Lake Storage Gen2

Connection *

Select...

+ New connection

< Previous

Next >

New connection (Azure Data Lake Storage Gen2)

Name *

YoutubeVideoAnalyticsInsights

Description

Connect via integration runtime *

AutoResolveIntegrationRuntime

Authentication method

Account key

Account selection method

From Azure subscription

Enter manually

Azure subscription

Storage account name *

youtubevideanalytics

Test connection

To linked service

To file path

topics / results

Annotations

+ New

Parameters

Create

Connection successful

Test connection

Cancel

Copy Data tool

1 Properties

2 Source

3 Dataset

4 Configuration

5 Target

6 Settings

7 Review and finish

Source data store

Specify the source data store for the copy task. You can use an existing data store c

Source type

Azure Data Lake Storage Gen2

Connection *

YoutubeVideoAnalyticsInsights

Edit

+ New co

File or folder

If the identity you use to access the data store only has permission to subdirectory instead of the entire account, specify the path to browse.

Options

Binary copy

Recursively

Enable partition discovery

Max concurrent connections

Filter by last modified

Start time (UTC)

End time (UTC)

Browse

Select a file or folder.

Root folder > topics > results

bottom_3_user_interaction

top_3_by_category_monthly

top_3_by_category_yearly

top_3_categories

top_3_channels

top_3_likes_dislikes_ratio_monthly

top_3_user_interaction

top_3_videos

Copy Data tool

✓ Properties

✓ Source

3 Target

• Dataset

○ Configuration

4 Settings

5 Review and finish

Destination data store

Specify the destination data store for the copy task. You can use an existing data store connection or specify a new data store.

Target type

Azure Cosmos DB (SQL API)

Connection *

Select...

New connection

< Previous

Next >

New connection (Azure Cosmos DB (SQL API))

Name *

YoutubeVideoAnalyticsInsightsSQL

Description

Connect via integration runtime *

AutoResolveIntegrationRuntime

Authentication method

Connection string

Connection string

Azure Key Vault

Account selection method

From Azure subscription

Enter manually

Azure subscription

Azure Cosmos DB account name *

youtube-video-analytics

Database name *

insights

Additional connection properties

+ New

Annotations

+ New

Create

Connection successful

Test connection

Cancel

Copy Data tool

✓ Properties

✓ Source

3 Target

• Dataset

○ Configuration

Destination data store

Specify the destination data store for the copy task. You can use an existing data store connection or specify a new data store.

Target type

Azure Cosmos DB (SQL API)

Connection *

YoutubeVideoAnalyticsInsigh.✓

Edit

New connection

Source

Target

▼ Azure Data Lake Storage Gen2 file → usvideos

Copy Data tool

✓ Properties

✓ Source

✓ Target

4 Settings

5 Review and finish

youtube-video-analytics

Settings

Enter name and description for the copy data task, more options for data movement

Task name *

YoutubeVideoAnalyticsInsightsCopy

Task description

Data consistency verification

Fault tolerance

Enable logging

Enable staging

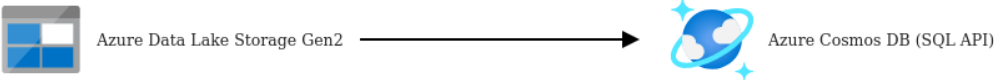
▶ Advanced

Copy Data tool

- ✓ Properties
- ✓ Source
- ✓ Target
- ✓ Settings
- 5 Review and finish
- Review
- Deployment

Summary

You are running pipeline to copy data from Azure Data Lake Storage Gen2 to Azure Cosmos DB (SQL API).



Properties

[Edit](#)

Task name YoutubeVideoAnalyticsInsightsCopy

Task description

[Edit](#)

Source

Connection name YoutubeVideoAnalyticsInsights

Dataset name SourceDataset_dnw

Folder path results

[Edit](#)

Target

Connection name YoutubeVideoAnalyticsInsightsSQL

Dataset name DestinationDataset_dnw

Collection usvideos

Copy settings

[Edit](#)

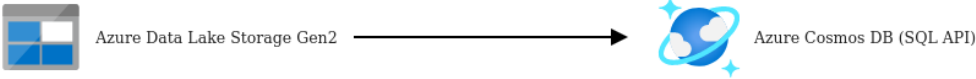
Timeout 7.00:00:00

[< Previous](#)

[Next >](#)

Copy Data tool

- ✓ Properties
- ✓ Source
- ✓ Target
- ✓ Settings
- 5 Review and finish
- Review
- Deployment



Deployment complete

▸ Validate copy runtime environment ✓

Deployment step	Status
> Creating datasets	Succeeded ✓
> Creating pipelines	Succeeded ✓
> Running pipelines	Succeeded ✓

Datasets and pipelines have been created. You can now monitor and edit the copy pipelines or click finish to close Copy Data Tool.

Insights now available for query through Cosmos SQL API

Home > Azure Cosmos DB > youtube-video-analytics

»

youtube-video-analytics | Data Explorer

...

Azure Cosmos DB account

New Item

Update

Discard

Delete

Upload Item

SQL API

Items

DATA

insights

usvideos

Items

Settings

Stored Procedures

User Defined Functions

Triggers

NOTEBOOKS

Gallery

My Notebooks

SELECT * FROM c

Edit Filter

id	/id
top3LikesDislik...	top3LikesDislik...
top3Categories	top3Categories
bottom3UserIn...	bottom3UserIn...
top3Channels	top3Channels
top3ByCategor...	top3ByCategor...
top3Videos	top3Videos
top3UserIntera...	top3UserIntera...
top3ByCategor...	top3ByCategor...

Load more

```
1 {
2   "id": "top3UserInteraction",
3   "name": "Top 3 videos by user interaction",
4   "user_interaction": [
5     {
6       "category": "Music",
7       "category_id": "18",
8       "channel_title": "ChildishGambinoVEVO",
9       "comment_count": 517232,
10      "comments_disabled": false,
11      "dislikes": 343541,
12      "likes": 5023450,
13      "publish_time": "2018-05-06T04:00:07.000Z",
14      "ratings_disabled": false,
15      "tags": "Childish Gambino|\"Rap|\"This Is America|\"mcDJ Recording",
16      "title": "Childish Gambino - This Is America (Official Video)",
17      "trending_date": "2018-06-02",
18      "user_interaction": 231096146,
19      "video_id": "VY0jWnS4cMY",
20      "views": 225211923
21    },
22    {
23      "category": "Entertainment",
24      "category_id": "24",
25      "channel_title": "YouTube Spotlight",
26      "comment_count": 810698,
27      "comments_disabled": false,
28      "dislikes": 1643059,
29      "likes": 3093544,
30      "publish_time": "2017-12-06T17:58:51.000Z",
31      "ratings_disabled": false,
32      "tags": "Rewind|\"Rewind 2017|\"youtube rewind 2017|\"#YouTubeRwi",
33      "title": "YouTube Rewind: The Shape of 2017 | #YouTubeRewind"
```