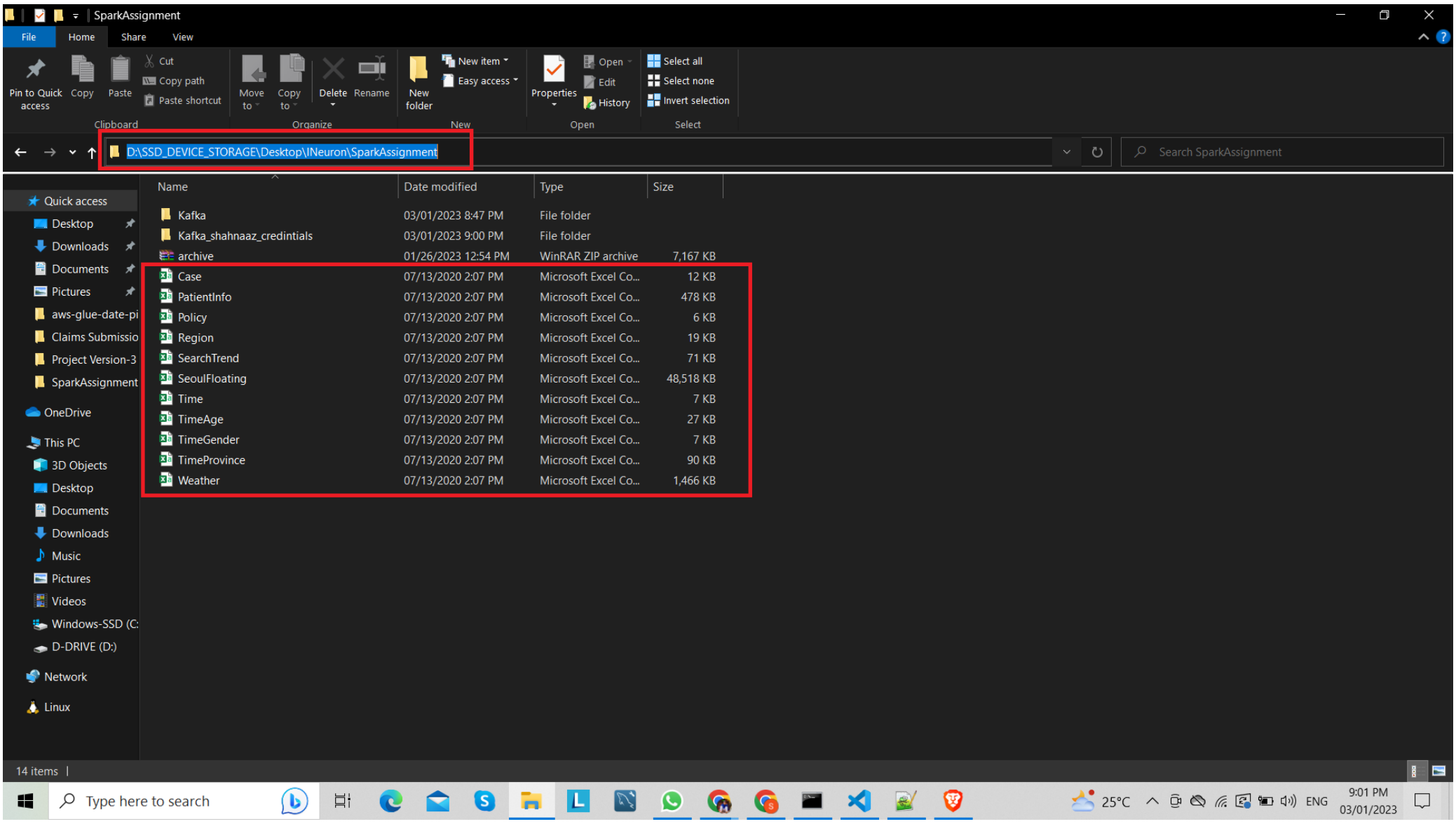
1.Download the data from the given URL :

https://www.kaggle.com/datasets/kimjihoo/coronavirusdataset



2. Create a producer with a python connector in confluent Kafka and stream your data.

Topics - Confluent Cloud

confluent.cloud/environments/env-63ovw6/clusters/lkc-knoy36/topics/case/message-viewer

mystartsearch | Powered by Browse... | New Tab | YouTube | Maps

CONFLUENT

Stream Catalog | LEARN

HOME > ENVIRONMENTS > DEFAULT > CLUSTER_1 > TOPICS >

- Cluster Overview
  - Dashboard
  - Networking
  - API Keys
  - Cluster Settings
- Stream Lineage
- Stream Designer
- Topics
- ksqlDB
- Connectors
- Clients
- Schema Registry

CLI and Tools

Support

case

Overview | Messages | Schema | Configuration

Producers

Bytes in/sec --

Consumers

Bytes out/sec 0

Message fields

- topic
- partition
- offset
- timestamp
- timestampType
- headers
- key
- value
  - case_id
  - province
  - city
  - group

Filter by keyword | Jump to offset | 0 / Partition: 0

+ Produce a new message to this topic

Newest

{"case_id":6100007,"province":"Gyeongsangnam-do","city":"Yangsan-si","group":true,"infection_case":"Soso Seowon","confirmed":3,"latiti…

Partition: 0 Offset: 22 Timestamp: 1677677253769

{"case_id":6000007,"province":"Gyeongsangbuk-do","city":"Chilgok-gun","group":true,"infection_case":"Milal Shelter","confirmed":36,"latitude…

Partition: 0 Offset: 21 Timestamp: 1677677253713

{"case_id":5100005,"province":"Jeollanam-do","city":"-","group":false,"infection_case":"etc","confirmed":4,"latitude":"-","longitude":"-"}

Partition: 0 Offset: 20 Timestamp: 1677677253689

{"case_id":5100004,"province":"Jeollanam-do","city":"-","group":false,"infection_case":"contact with patient","confirmed":4,"latitude":"-","…

Partition: 0 Offset: 19 Timestamp: 1677677253689

Explore Stream Lineage

Description

Add description

Tags

Add tags to this topic

Add business metadata

Date created

Mar 1 2023 12:21 PM

Date modified

--

Retention time

1 week

Retention size

Infinite

Number of partitions

6

Cleanup policy

delete

JSON CSV Download

Type here to search

25°C ENG 9:06 PM 03/01/2023

Topics - Confluent Cloud

confluent.cloud/environments/env-63ovw6/clusters/lkc-knoy36/topics/patientinfo/message-viewer

mystartsearch | Powered by Browse... | New Tab | YouTube | Maps

CONFLUENT

Stream Catalog | LEARN

HOME > ENVIRONMENTS > DEFAULT > CLUSTER_1 > TOPICS >

- Cluster Overview
  - Dashboard
  - Networking
  - API Keys
  - Cluster Settings
- Stream Lineage
- Stream Designer
- Topics
- ksqlDB
- Connectors
- Clients
- Schema Registry

CLI and Tools

Support

patientinfo

Overview | Messages | Schema | Configuration

Producers

Bytes in/sec --

Consumers

Bytes out/sec 5.93K

Message fields

- topic
- partition
- offset
- timestamp
- timestampType
- headers
- key
- value
  - patient_id
  - sex
  - age
  - country

Filter by keyword | Jump to offset | 0 / Partition: 0

+ Produce a new message to this topic

{"patient_id":1000001297,"sex":"NAN","age":"NAN","country":"Korea","province":"Seoul","city":"Yeongdeungpo-gu","infection_case":"NAN","infec…

Partition: 0 Offset: 229 Timestamp: 1677679999088

{"patient_id":1000001293,"sex":"NAN","age":"NAN","country":"Korea","province":"Seoul","city":"Yangcheon-gu","infection_case":"overseas inflo…

Partition: 0 Offset: 228 Timestamp: 1677679999076

{"patient_id":1000001288,"sex":"NAN","age":"NAN","country":"Korea","province":"Seoul","city":"Gangnam-gu","infection_case":"Gangnam Yeoksam-…

Partition: 0 Offset: 227 Timestamp: 1677679999060

{"patient_id":1000001285,"sex":"NAN","age":"NAN","country":"Korea","province":"Seoul","city":"Dongdaemun-gu","infection_case":"Gangnam Yeoks…

Partition: 0 Offset: 226 Timestamp: 1677679999052

Explore Stream Lineage

Description

Add description

Tags

Add tags to this topic

Add business metadata

Date created

Mar 1 2023 7:02 PM

Date modified

--

Retention time

1 week

Retention size

Infinite

Number of partitions

6

Cleanup policy

delete

JSON CSV Download

Type here to search

25°C ENG 9:06 PM 03/01/2023

Topics - Confluent Cloud

confluent.cloud/environments/env-63ovw6/clusters/lkc-knoy36/topics/policy/message-viewer

mystartsearch | Powered by Browse... | New Tab | YouTube | Maps

CONFLUENT

Stream Catalog | LEARN

HOME > ENVIRONMENTS > DEFAULT > CLUSTER_1 > TOPICS >

- Cluster Overview
  - Dashboard
  - Networking
  - API Keys
  - Cluster Settings
- Stream Lineage
- Stream Designer
- Topics
- ksqlDB
- Connectors
- Clients
- Schema Registry

CLI and Tools

Support

policy

Overview | Messages | Schema | Configuration

Producers

Bytes in/sec --

Consumers

Bytes out/sec 0

Message fields

- topic
- partition
- offset
- timestamp
- timestampType
- headers
- key
- value
  - policy_id
  - country
  - type
  - gov_policy

Filter by keyword | Jump to offset | 0 / Partition: 0

+ Produce a new message to this topic

{"policy_id":58,"country":"Korea","type":"Transformation","gov_policy":"Wearing of masks","detail":"Mandatory wearing of passenger mask dome…
Partition: 0 Offset: 11 Timestamp: 1677680717995

{"policy_id":54,"country":"Korea","type":"Administrative","gov_policy":"Close bars and clubs","detail":"administrative order to close bars a…
Partition: 0 Offset: 10 Timestamp: 1677680717979

{"policy_id":53,"country":"Korea","type":"Technology","gov_policy":"Electronic Wristbands","detail":"on people who violate self-quarantine m…
Partition: 0 Offset: 9 Timestamp: 1677680717979

{"policy_id":51,"country":"Korea","type":"Technology","gov_policy":"Self-Diagnosis App","detail":"NAN","start_date":"2020-02-12","end_date":…
Partition: 0 Offset: 8 Timestamp: 1677680717971

Explore Stream Lineage

Description
Add description

Tags
Add tags to this topic

Add business metadata

Date created
Mar. 1 2023 7:52 PM

Date modified
--

Retention time
1 week

Retention size
Infinite

Number of partitions
6

Cleanup policy
delete

JSON | CSV | Download

Type here to search | 25°C | ENG | 9:07 PM 03/01/2023

Topics - Confluent Cloud

confluent.cloud/environments/env-63ovw6/clusters/lkc-knoy36/topics/region/message-viewer

mystartsearch | Powered by Browse... | New Tab | YouTube | Maps

CONFLUENT

Stream Catalog | LEARN

HOME > ENVIRONMENTS > DEFAULT > CLUSTER_1 > TOPICS >

- Cluster Overview
  - Dashboard
  - Networking
  - API Keys
  - Cluster Settings
- Stream Lineage
- Stream Designer
- Topics
- ksqlDB
- Connectors
- Clients
- Schema Registry

CLI and Tools

Support

region

Overview | Messages | Schema | Configuration

Producers

Bytes in/sec --

Consumers

Bytes out/sec --

Message fields

- topic
- partition
- offset
- timestamp
- timestampType
- headers
- key
- value
  - code
  - province
  - city
  - latitude

Filter by keyword | Jump to offset | 0 / Partition: 1

+ Produce a new message to this topic

{"code":61120,"province":"Gyeongsangnam-do","city":"Changnyeong-gun","latitude":35.544603,"longitude":128.49233,"elementary_school_count":17…
Partition: 1 Offset: 46 Timestamp: 1677681166745

{"code":61080,"province":"Gyeongsangnam-do","city":"Sancheong-gun","latitude":35.415544,"longitude":127.87349,"elementary_school_count":13,"…
Partition: 1 Offset: 45 Timestamp: 1677681166729

{"code":51190,"province":"Jeollanam-do","city":"Jindo-gun","latitude":34.486834,"longitude":126.263554,"elementary_school_count":10,"kinderg…
Partition: 1 Offset: 44 Timestamp: 1677681166633

{"code":51120,"province":"Jeollanam-do","city":"Sinan-gun","latitude":34.833564,"longitude":126.351636,"elementary_school_count":19,"kinderg…
Partition: 1 Offset: 43 Timestamp: 1677681166609

Explore Stream Lineage

Description
Add description

Tags
Add tags to this topic

Add business metadata

Date created
Mar. 1 2023 7:58 PM

Date modified
--

Retention time
1 week

Retention size
Infinite

Number of partitions
6

Cleanup policy
delete

JSON | CSV | Download

Type here to search | 25°C | ENG | 9:07 PM 03/01/2023

Topics - Confluent Cloud

confluent.cloud/environments/env-63ovw6/clusters/lkc-knoy36/topics/search_trend/message-viewer

mystartsearch | Powered by Browse... | New Tab | YouTube | Maps

CONFLUENT

Stream Catalog | LEARN

HOME > ENVIRONMENTS > DEFAULT > CLUSTER_1 > TOPICS >

- Cluster Overview
  - Dashboard
  - Networking
  - API Keys
  - Cluster Settings
- Stream Lineage
- Stream Designer
- Topics
- ksqlDB
- Connectors
- Clients
- Schema Registry

CLI and Tools

Support

search_trend

Overview | Messages | Schema | Configuration

Producers

Bytes in/sec 0

Consumers

Bytes out/sec 0

Message fields

- topic
- partition
- offset
- timestamp
- timestampType
- headers
- key
- value
  - date
  - cold
  - flu
  - pneumonia

Filter by keyword | Jump to offset | 0 / Partition: 0

+ Produce a new message to this topic

{"date":"2016-08-27","cold":0.12481,"flu":0.01999,"pneumonia":0.12954,"coronavirus":0.00418}
Partition: 0 Offset: 39 Timestamp: 1677681512923

{"date":"2016-08-18","cold":0.12445,"flu":0.01736,"pneumonia":0.14826,"coronavirus":0.00563}
Partition: 0 Offset: 38 Timestamp: 1677681512859

{"date":"2016-08-11","cold":0.09945,"flu":0.01663,"pneumonia":0.14917,"coronavirus":0.00636}
Partition: 0 Offset: 37 Timestamp: 1677681512843

{"date":"2016-08-10","cold":0.09636,"flu":0.01463,"pneumonia":0.15163,"coronavirus":0.00645}
Partition: 0 Offset: 36 Timestamp: 1677681512843

Explore Stream Lineage

Description
Add description

Tags
Add tags to this topic

Add business metadata

Date created
Mar. 1 2023 8:06 PM

Date modified
--

Retention time
1 week

Retention size
Infinite

Number of partitions
6

Cleanup policy
delete

JSON CSV Download

Type here to search

25°C ENG 9:08 PM 03/01/2023

Topics - Confluent Cloud

confluent.cloud/environments/env-63ovw6/clusters/lkc-knoy36/topics/SeoulFloating/message-viewer

mystartsearch | Powered by Browse... | New Tab | YouTube | Maps

CONFLUENT

Stream Catalog | LEARN

HOME > ENVIRONMENTS > DEFAULT > CLUSTER_1 > TOPICS >

- Cluster Overview
  - Dashboard
  - Networking
  - API Keys
  - Cluster Settings
- Stream Lineage
- Stream Designer
- Topics
- ksqlDB
- Connectors
- Clients
- Schema Registry

CLI and Tools

Support

SeoulFloating

Overview | Messages | Schema | Configuration

Producers

Bytes in/sec 0

Consumers

Bytes out/sec 0

Message fields

- topic
- partition
- offset
- timestamp
- timestampType
- headers
- key
- value
  - date
  - hour
  - birth_year
  - sex

Filter by keyword | Jump to offset | 0 / Partition: 0

+ Produce a new message to this topic

{"date":"2020-01-01","hour":3,"birth_year":20,"sex":"male","province":"Seoul","city":"Jung-gu","fp_num":11460}
Partition: 0 Offset: 29 Timestamp: 1677681936282

{"date":"2020-01-01","hour":3,"birth_year":20,"sex":"male","province":"Seoul","city":"Gwanak-gu","fp_num":45910}
Partition: 0 Offset: 28 Timestamp: 1677681936266

{"date":"2020-01-01","hour":3,"birth_year":20,"sex":"male","province":"Seoul","city":"Guro-gu","fp_num":27040}
Partition: 0 Offset: 27 Timestamp: 1677681936258

{"date":"2020-01-01","hour":3,"birth_year":20,"sex":"female","province":"Seoul","city":"Geumcheon-gu","fp_num":15010}
Partition: 0 Offset: 26 Timestamp: 1677681936250

Explore Stream Lineage

Description
Add description

Tags
Add tags to this topic

Add business metadata

Date created
Mar. 1 2023 8:12 PM

Date modified
--

Retention time
1 week

Retention size
Infinite

Number of partitions
6

Cleanup policy
delete

JSON CSV Download

Type here to search

25°C ENG 9:08 PM 03/01/2023

Topics - Confluent Cloud

confluent.cloud/environments/env-63ovw6/clusters/lkc-knoy36/topics/Time/message-viewer

mystartsearch | Powered by Browse... | New Tab | YouTube | Maps

CONFLUENT

Stream Catalog | LEARN

HOME > ENVIRONMENTS > DEFAULT > CLUSTER_1 > TOPICS >

- Cluster Overview
  - Dashboard
  - Networking
  - API Keys
  - Cluster Settings
- Stream Lineage
- Stream Designer
- Topics
- ksqlDB
- Connectors
- Clients
- Schema Registry

CLI and Tools

Support

Time

Overview | Messages | Schema | Configuration

Producers

Bytes in/sec 0

Consumers

Bytes out/sec 0

Message fields

- topic
- partition
- offset
- timestamp
- timestampType
- headers
- key
- value
  - date
  - time
  - test
  - negative

Filter by keyword | Jump to offset | 0 / Partition: 0

+ Produce a new message to this topic

{"date":"2020-06-23","time":0,"test":1196012,"negative":1161250,"confirmed":12484,"released":10908,"deceased":281}
Partition: 0 Offset: 21 Timestamp: 1677682500098

{"date":"2020-06-21","time":0,"test":1176463,"negative":1143971,"confirmed":12421,"released":10868,"deceased":280}
Partition: 0 Offset: 20 Timestamp: 1677682500098

{"date":"2020-06-11","time":0,"test":1066888,"negative":1029447,"confirmed":11947,"released":10654,"deceased":276}
Partition: 0 Offset: 19 Timestamp: 1677682500034

{"date":"2020-06-09","time":0,"test":1035997,"negative":996686,"confirmed":11852,"released":10589,"deceased":274}
Partition: 0 Offset: 18 Timestamp: 1677682500034

Explore Stream Lineage

Description
Add description

Tags
Add tags to this topic

Add business metadata

Date created
Mar. 1 2023 8:22 PM

Date modified
--

Retention time
1 week

Retention size
Infinite

Number of partitions
6

Cleanup policy
delete

JSON CSV Download

Type here to search | 25°C | ENG | 9:09 PM 03/01/2023

Topics - Confluent Cloud

confluent.cloud/environments/env-63ovw6/clusters/lkc-knoy36/topics/time_province/message-viewer

mystartsearch | Powered by Browse... | New Tab | YouTube | Maps

CONFLUENT

Stream Catalog | LEARN

HOME > ENVIRONMENTS > DEFAULT > CLUSTER_1 > TOPICS >

- Cluster Overview
  - Dashboard
  - Networking
  - API Keys
  - Cluster Settings
- Stream Lineage
- Stream Designer
- Topics
- ksqlDB
- Connectors
- Clients
- Schema Registry

CLI and Tools

Support

time_province

Overview | Messages | Schema | Configuration

Producers

Bytes in/sec 0

Consumers

Bytes out/sec 0

Message fields

- topic
- partition
- offset
- timestamp
- timestampType
- headers
- key
- value
  - date
  - time
  - province
  - confirmed

Filter by keyword | Jump to offset | 0 / Partition: 0

+ Produce a new message to this topic

{"date":"2020-02-07","time":"16","province":"Daejeon","confirmed":0,"released":0,"deceased":0}
Partition: 0 Offset: 49 Timestamp: 1677683770240

{"date":"2020-02-07","time":"16","province":"Gwangju","confirmed":1,"released":0,"deceased":0}
Partition: 0 Offset: 48 Timestamp: 1677683770239

{"date":"2020-02-07","time":"16","province":"Busan","confirmed":0,"released":0,"deceased":0}
Partition: 0 Offset: 47 Timestamp: 1677683770238

{"date":"2020-02-06","time":"16","province":"Jeju-do","confirmed":0,"released":0,"deceased":0}
Partition: 0 Offset: 46 Timestamp: 1677683770237

Explore Stream Lineage

Description
Add description

Tags
Add tags to this topic

Add business metadata

Date created
Mar. 1 2023 8:43 PM

Date modified
--

Retention time
1 week

Retention size
Infinite

Number of partitions
6

Cleanup policy
delete

JSON CSV Download

Type here to search | 25°C | ENG | 9:09 PM 03/01/2023

Topics - Confluent Cloud

confluent.cloud/environments/env-63ovw6/clusters/lkc-knoy36/topics/timeage/message-viewer

mystartsearch | Powered by Browse... | New Tab | YouTube | Maps

CONFLUENT

Stream Catalog | LEARN

HOME > ENVIRONMENTS > DEFAULT > CLUSTER_1 > TOPICS >

- Cluster Overview
  - Dashboard
  - Networking
  - API Keys
  - Cluster Settings
- Stream Lineage
- Stream Designer
- Topics
- ksqlDB
- Connectors
- Clients
- Schema Registry

CLI and Tools

Support

timeage

Overview | Messages | Schema | Configuration

Producers

Bytes in/sec 0

Consumers

Bytes out/sec 0

Message fields

- topic
- partition
- offset
- timestamp
- timestampType
- headers
- key
- value
  - date
  - time
  - age
  - confirmed

Filter by keyword | Jump to offset | 0 / Partition: 0

+ Produce a new message to this topic

{"date":"2020-03-31","time":0,"age":"70s","confirmed":651,"deceased":46}
Partition: 0 Offset: 49 Timestamp: 1677682931700

{"date":"2020-03-31","time":0,"age":"30s","confirmed":1012,"deceased":1}
Partition: 0 Offset: 48 Timestamp: 1677682931692

{"date":"2020-03-30","time":0,"age":"20s","confirmed":2630,"deceased":0}
Partition: 0 Offset: 47 Timestamp: 1677682931676

{"date":"2020-03-30","time":0,"age":"0s","confirmed":112,"deceased":0}
Partition: 0 Offset: 46 Timestamp: 1677682931676

Explore Stream Lineage

Description
Add description

Tags
Add tags to this topic

Add business metadata

Date created
Mar. 1 2023 8:27 PM

Date modified
--

Retention time
1 week

Retention size
Infinite

Number of partitions
6

Cleanup policy
delete

JSON CSV Download

Type here to search | 25°C | ENG | 9:10 PM 03/01/2023

Topics - Confluent Cloud

confluent.cloud/environments/env-63ovw6/clusters/lkc-knoy36/topics/timegender/message-viewer

mystartsearch | Powered by Browse... | New Tab | YouTube | Maps

CONFLUENT

Stream Catalog | LEARN

HOME > ENVIRONMENTS > DEFAULT > CLUSTER_1 > TOPICS >

- Cluster Overview
  - Dashboard
  - Networking
  - API Keys
  - Cluster Settings
- Stream Lineage
- Stream Designer
- Topics
- ksqlDB
- Connectors
- Clients
- Schema Registry

CLI and Tools

Support

timegender

Overview | Messages | Schema | Configuration

Producers

Bytes in/sec 0

Consumers

Bytes out/sec 0

Message fields

- topic
- partition
- offset
- timestamp
- timestampType
- headers
- key
- value
  - date
  - time
  - sex
  - confirmed

Filter by keyword | Jump to offset | 0 / Partition: 0

+ Produce a new message to this topic

{"date":"2020-06-30","time":"0","sex":"female","confirmed":7305,"deceased":131}
Partition: 0 Offset: 37 Timestamp: 1677683414742

{"date":"2020-06-28","time":"0","sex":"male","confirmed":5450,"deceased":151}
Partition: 0 Offset: 36 Timestamp: 1677683414738

{"date":"2020-06-24","time":"0","sex":"female","confirmed":7190,"deceased":131}
Partition: 0 Offset: 35 Timestamp: 1677683414733

{"date":"2020-06-19","time":"0","sex":"female","confirmed":7100,"deceased":130}
Partition: 0 Offset: 34 Timestamp: 1677683414727

Explore Stream Lineage

Description
Add description

Tags
Add tags to this topic

Add business metadata

Date created
Mar. 1 2023 8:37 PM

Date modified
--

Retention time
1 week

Retention size
Infinite

Number of partitions
6

Cleanup policy
delete

JSON CSV Download

Type here to search | 25°C | ENG | 9:10 PM 03/01/2023

Topics - Confluent Cloud

confluent.cloud/environments/env-63ovw6/clusters/lkc-knoy36/topics/weather/message-viewer

mystartsearch | Powered by Browse... | New Tab | YouTube | Maps

CONFLUENT

Stream Catalog | LEARN

HOME > ENVIRONMENTS > DEFAULT > CLUSTER_1 > TOPICS >

- Cluster Overview
  - Dashboard
  - Networking
  - API Keys
  - Cluster Settings
- Stream Lineage
- Stream Designer
- Topics
- ksqlDB
- Connectors
- Clients
- Schema Registry

CLI and Tools

Support

weather

Overview | Messages | Schema | Configuration

Producers

Bytes in/sec 0

Consumers

Bytes out/sec 0

Message fields

- topic
- partition
- offset
- timestamp
- timestampType
- headers
- key
- value
  - code
  - province
  - date
  - avg_temp

Filter by keyword | Jump to offset | 0 / Partition: 0

+ Produce a new message to this topic

{"code":30000,"province":"Gangwon-do","date":"2016-01-18","avg_temp":-1.1,"min_temp":-9.7,"max_temp":4.5,"precipitation":0,"max_wind_speed":…
Partition: 0 Offset: 49 Timestamp: 1677684023808

{"code":13000,"province":"Gwangju","date":"2016-01-18","avg_temp":-0.8,"min_temp":-4.4,"max_temp":3.2,"precipitation":4.1,"max_wind_speed":4…
Partition: 0 Offset: 48 Timestamp: 1677684023803

{"code":14000,"province":"Incheon","date":"2016-01-17","avg_temp":2.5,"min_temp":1,"max_temp":5.2,"precipitation":0,"max_wind_speed":4.7,"mo…
Partition: 0 Offset: 47 Timestamp: 1677684023789

{"code":70000,"province":"Jeju-do","date":"2016-01-16","avg_temp":7.5,"min_temp":4.2,"max_temp":11.1,"precipitation":0,"max_wind_speed":4.3,…
Partition: 0 Offset: 46 Timestamp: 1677684023785

JSON | CSV | Download

Explore Stream Lineage

Description
Add description

Tags
Add tags to this topic

Add business metadata

Date created
Mar. 1 2023 8:49 PM

Date modified
--

Retention time
1 week

Retention size
Infinite

Number of partitions
6

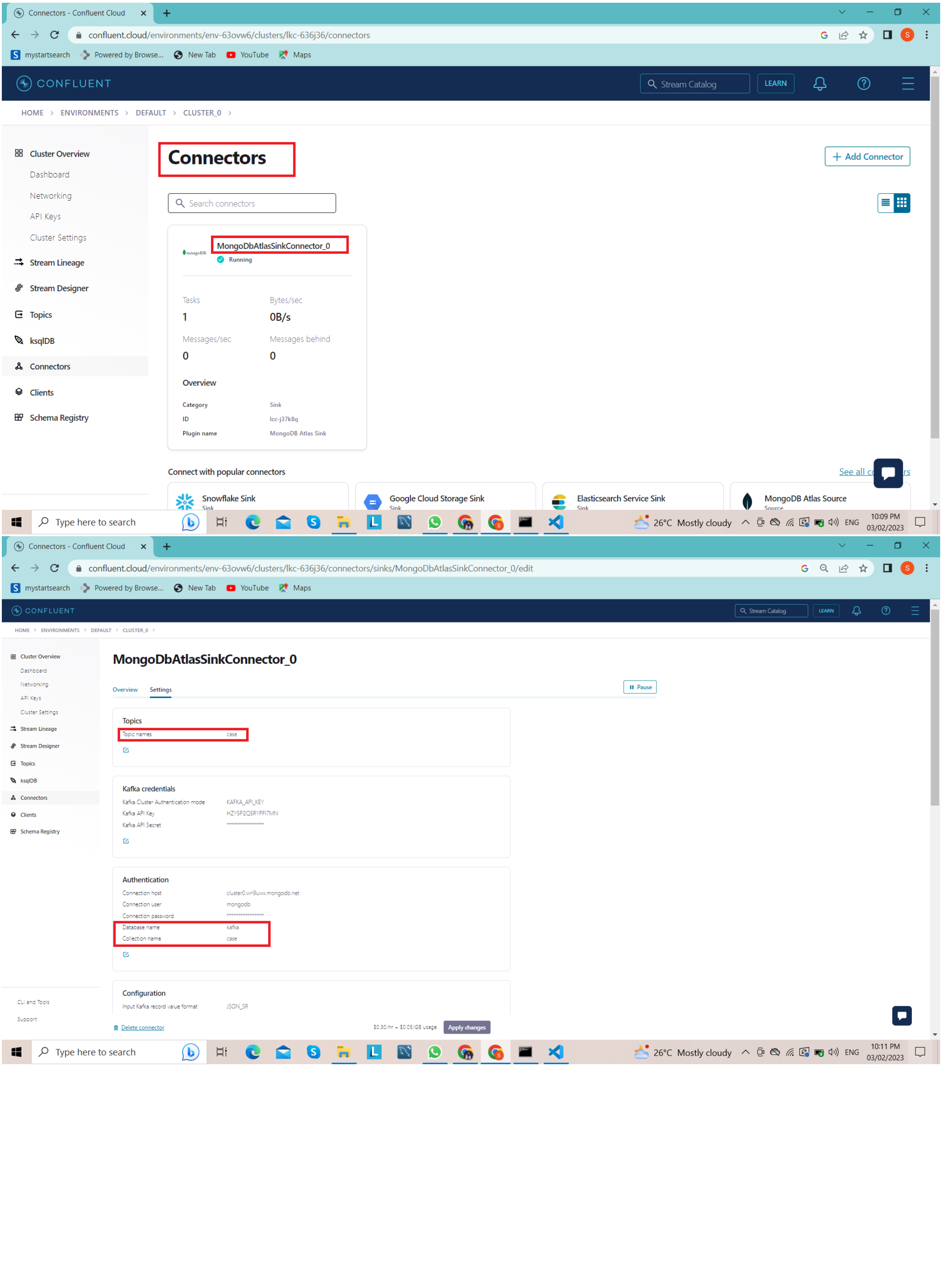
Cleanup policy
delete

Type here to search

25°C ENG 9:11 PM 03/01/2023

3. Consume your data through the python connector and dump it in mongo dB atlas.

Note: Here in the dataset you will be finding a multiple files you need to use all file for the Kafka and mongo dB



Invensis Technologies Pvt. Ltd. - / | Mail - abdulqadir.syed - Network | hcare - INV_TEAM - Slack | Data | Cloud: MongoDB Cloud

cloud.mongodb.com/v2/63a3f850eb7a144c3202b002#/metrics/replicaSet/6400bb19de519e522e409747/explorer/kafka/case/find

Atlas | Syedabdul's ... | Access Manager | Billing | All Clusters | Get Help | Syedabdul

Project 0 | Data Services | App Services | Charts

DEPLOYMENT
- Database
- Data Lake PREVIEW

SERVICES
- Triggers
- Data API
- Data Federation
- Search

SECURITY
- Database Access
- Network Access
- Advanced

Goto

SYEDABDUL'S ORG - 2022-12-22 > PROJECT 0 > DATABASES

Cluster0

VERSION 5.0.15 | REGION GCP Iowa (us-central1)

Overview | Real Time | Metrics | Collections | Search | Profiler | Performance Advisor | Online Archive | Cmd Line Tools

DATABASES: 1 COLLECTIONS: 1

VISUALIZE YOUR DATA | REFRESH

+ Create Database

Search Namespaces

- kafka
  - case

kafka.case

STORAGE SIZE: 28KB | LOGICAL DATA SIZE: 32.11KB | TOTAL DOCUMENTS: 174 | INDEXES TOTAL SIZE: 24KB

Find | Indexes | Schema Anti-Patterns 0 | Aggregation | Search Indexes

INSERT DOCUMENT

FILTER { field: 'value' } | OPTIONS | Apply | Reset

QUERY RESULTS: 161-174 OF 174

```
_id: ObjectId('6400d035e4b28f74a00fdaa7')
infection_case: "etc"
province: "Sejong"
city: "-"
latitude: "-"
case_id: 1700006
confirmed: 1
group: false
longitude: "-"
```

```
_id: ObjectId('6400d035e4b28f74a00fdaa8')
infection_case: "Anyang Gunpo Pastors Group"
province: "Gyeonggi-do"
city: "Anyang-si"
latitude: "37.381784"
case_id: 2000011
```

Type here to search | 26°C Mostly cloudy | ENG | 10:15 PM 03/02/2023

Invensis Technologies Pvt. Ltd. - / | Mail - abdulqadir.syed - Network | hcare - INV_TEAM - Slack | Data | Cloud: MongoDB Cloud

cloud.mongodb.com/v2/63a3f850eb7a144c3202b002#/metrics/replicaSet/6400bb19de519e522e409747/explorer/kafka/patientinfo/find

Atlas | Syedabdul's ... | Access Manager | Billing | All Clusters | Get Help | Syedabdul

Project 0 | Data Services | App Services | Charts

DEPLOYMENT
- Database
- Data Lake PREVIEW

SERVICES
- Triggers
- Data API
- Data Federation
- Search

SECURITY
- Database Access
- Network Access
- Advanced

Goto

SYEDABDUL'S ORG - 2022-12-22 > PROJECT 0 > DATABASES

Cluster0

VERSION 5.0.15 | REGION GCP Iowa (us-central1)

Overview | Real Time | Metrics | Collections | Search | Profiler | Performance Advisor | Online Archive | Cmd Line Tools

DATABASES: 1 COLLECTIONS: 2

VISUALIZE YOUR DATA | REFRESH

+ Create Database

Search Namespaces

- kafka
  - case
  - patientinfo

kafka.patientinfo

STORAGE SIZE: 712KB | LOGICAL DATA SIZE: 1.69MB | TOTAL DOCUMENTS: 5165 | INDEXES TOTAL SIZE: 332KB

Find | Indexes | Schema Anti-Patterns 0 | Aggregation | Search Indexes

INSERT DOCUMENT

FILTER { field: 'value' } | OPTIONS | Apply | Reset

QUERY RESULTS: 1-20 OF MANY

```
_id: ObjectId('6400d327e4b28f74a00fdaba')
country: "Korea"
infection_case: "contact with patient"
city: "Jongno-gu"
sex: "male"
confirmed_date: "2020-01-30"
released_date: "2020-02-19"
contact_number: "17"
deceased_date: "NAN"
infected_by: "2002000001"
province: "Seoul"
symptom_onset_date: "NAN"
patient_id: 1000000003
state: "released"
age: "50s"
```

Type here to search | 26°C Mostly cloudy | ENG | 10:22 PM 03/02/2023

Invensis Technologies Pvt. Ltd. - / | Mail - abdulqadir.syed - Network | hcare - INV_TEAM - Slack | Data | Cloud: MongoDB Cloud

cloud.mongodb.com/v2/63a3f850eb7a144c3202b002#/metrics/replicaSet/6400bb19de519e522e409747/explorer/kafka/policy/find

Atlas | Syedabdul's ... | Access Manager | Billing | All Clusters | Get Help | Syedabdul

Project 0 | Data Services | App Services | Charts

DEPLOYMENT
- Database
- Data Lake PREVIEW

SERVICES
- Triggers
- Data API
- Data Federation
- Search

SECURITY
- Database Access
- Network Access
- Advanced

Goto

SYEDABDUL'S ORG - 2022-12-22 > PROJECT 0 > DATABASES

Cluster0

VERSION 5.0.15 | REGION GCP Iowa (us-central1)

Overview | Real Time | Metrics | Collections | Search | Profiler | Performance Advisor | Online Archive | Cmd Line Tools

DATABASES: 1 COLLECTIONS: 3

VISUALIZE YOUR DATA | REFRESH

+ Create Database

Search Namespaces

- kafka
  - case
  - patientinfo
  - policy

kafka.policy

STORAGE SIZE: 24KB | LOGICAL DATA SIZE: 12.67KB | TOTAL DOCUMENTS: 61 | INDEXES TOTAL SIZE: 20KB

Find | Indexes | Schema Anti-Patterns 0 | Aggregation | Search Indexes

INSERT DOCUMENT

FILTER { field: 'value' } | OPTIONS | Apply | Reset

QUERY RESULTS: 1-20 OF MANY

```
_id: ObjectId('6400d543e4b28f74a00feee8')
end_date: "NAN"
country: "Korea"
gov_policy: "Special Immigration Procedure"
policy_id: 6
detail: "from Hong Kong"
type: "Immigration"
start_date: "2020-02-12"
```

```
_id: ObjectId('6400d543e4b28f74a00feee9')
end_date: "NAN"
country: "Korea"
gov_policy: "Emergency Use Authorization of Diagnostic Kit"
policy_id: 23
detail: "4th EUA"
type: "Health"
```

Type here to search | 24°C Mostly cloudy | ENG | 10:30 PM 03/02/2023

Invensis Technologies Pvt. Ltd. - / | Mail - abdulqadir.syed - Network | hcare - INV_TEAM - Slack | Data | Cloud: MongoDB Cloud

cloud.mongodb.com/v2/63a3f850eb7a144c3202b002#/metrics/replicaSet/6400bb19de519e522e409747/explorer/kafka/region/find

Atlas | Syedabdul's ... | Access Manager | Billing | All Clusters | Get Help | Syedabdul

Project 0 | Data Services | App Services | Charts

DEPLOYMENT
- Database
- Data Lake PREVIEW

SERVICES
- Triggers
- Data API
- Data Federation
- Search

SECURITY
- Database Access
- Network Access
- Advanced

Goto

SYEDABDUL'S ORG - 2022-12-22 > PROJECT 0 > DATABASES

Cluster0

VERSION 5.0.15 | REGION GCP Iowa (us-central1)

Overview | Real Time | Metrics | Collections | Search | Profiler | Performance Advisor | Online Archive | Cmd Line Tools

DATABASES: 1 COLLECTIONS: 4

VISUALIZE YOUR DATA | REFRESH

+ Create Database

Search Namespaces

- kafka
  - case
  - patientinfo
  - policy
  - region

kafka.region

STORAGE SIZE: 40KB | LOGICAL DATA SIZE: 76.42KB | TOTAL DOCUMENTS: 244 | INDEXES TOTAL SIZE: 24KB

Find | Indexes | Schema Anti-Patterns 0 | Aggregation | Search Indexes

INSERT DOCUMENT

FILTER { field: 'value' } | OPTIONS | Apply | Reset

QUERY RESULTS: 1-20 OF MANY

```
_id: ObjectId('6400d66ee4b28f74a00fef26')
elderly_population_ratio: 15.38
code: 10000
province: "Seoul"
elementary_school_count: 607
city: "Seoul"
latitude: 37.566953
kindergarten_count: 830
university_count: 48
academy_ratio: 1.44
nursing_home_count: 22739
elderly_alone_ratio: 5.8
longitude: 126.977977
```

```
_id: ObjectId('6400d66ee4b28f74a00fef27')
elderly_population_ratio: 14.39
```

Type here to search | 24°C Mostly cloudy | ENG | 10:37 PM 03/02/2023

Invensis Technologies Pvt. Ltd. - / | Mail - abdulqadir.syed - Network | hcare - INV_TEAM - Slack | Data | Cloud: MongoDB Cloud

cloud.mongodb.com/v2/63a3f850eb7a144c3202b002#/metrics/replicaSet/6400bb19de519e522e409747/explorer/kafka/search_trend/find

Atlas | Syedabdul's ... | Access Manager | Billing | All Clusters | Get Help | Syedabdul

Project 0 | Data Services | App Services | Charts

DEPLOYMENT
- Database
- Data Lake PREVIEW

SERVICES
- Triggers
- Data API
- Data Federation
- Search

SECURITY
- Database Access
- Network Access
- Advanced

Goto

SYEDABDUL'S ORG - 2022-12-22 > PROJECT 0 > DATABASES

Cluster0

VERSION 5.0.15 | REGION GCP Iowa (us-central1)

Overview | Real Time | Metrics | Collections | Search | Profiler | Performance Advisor | Online Archive | Cmd Line Tools

DATABASES: 1 COLLECTIONS: 5

VISUALIZE YOUR DATA | REFRESH

+ Create Database

Search Namespaces

- kafka
  - case
  - patientinfo
  - policy
  - region
  - search_trend

kafka.search_trend

STORAGE SIZE: 108KB | LOGICAL DATA SIZE: 176.89KB | TOTAL DOCUMENTS: 1642 | INDEXES TOTAL SIZE: 64KB

Find | Indexes | Schema Anti-Patterns 0 | Aggregation | Search Indexes

INSERT DOCUMENT

FILTER { field: 'value' } | OPTIONS | Apply | Reset

QUERY RESULTS: 1-20 OF MANY

```
_id: ObjectId('6400d83be4b28f74a00ff01b')
date: "2016-01-02"
pneumonia: 0.20826
coronavirus: 0.0089
cold: 0.13372
flu: 0.17135
```

```
_id: ObjectId('6400d83be4b28f74a00ff01c')
date: "2016-01-04"
pneumonia: 0.29008
coronavirus: 0.01145
cold: 0.17463
flu: 0.18626
```

Type here to search | 24°C Mostly cloudy | ENG | 10:41 PM 03/02/2023

Invensis Technologies Pvt. Ltd. - / | Mail - abdulqadir.syed - Network | hcare - INV_TEAM - Slack | Data | Cloud: MongoDB Cloud

cloud.mongodb.com/v2/63a3f850eb7a144c3202b002#/metrics/replicaSet/6400bb19de519e522e409747/explorer/kafka/Time/find

Atlas | Syedabdul's ... | Access Manager | Billing | All Clusters | Get Help | Syedabdul

Project 0 | Data Services | App Services | Charts

DEPLOYMENT
- Database
- Data Lake PREVIEW

SERVICES
- Triggers
- Data API
- Data Federation
- Search

SECURITY
- Database Access
- Network Access
- Advanced

Goto

SYEDABDUL'S ORG - 2022-12-22 > PROJECT 0 > DATABASES

Cluster0

VERSION 5.0.15 | REGION GCP Iowa (us-central1)

Overview | Real Time | Metrics | Collections | Search | Profiler | Performance Advisor | Online Archive | Cmd Line Tools

DATABASES: 1 COLLECTIONS: 6

VISUALIZE YOUR DATA | REFRESH

+ Create Database

Search Namespaces

- kafka
  - Time
  - oase
  - patientinfo
  - policy
  - region
  - search_trend

kafka.Time

STORAGE SIZE: 4KB | LOGICAL DATA SIZE: 22.92KB | TOTAL DOCUMENTS: 163 | INDEXES TOTAL SIZE: 4KB

Find | Indexes | Schema Anti-Patterns 0 | Aggregation | Search Indexes

INSERT DOCUMENT

FILTER { field: 'value' } | OPTIONS | Apply | Reset

QUERY RESULTS: 1-20 OF MANY

```
_id: ObjectId('6400d905e4b28f74a00ff686')
date: "2020-01-29"
negative: 155
deceased: 0
test: 187
time: 16
confirmed: 4
released: 0
```

```
_id: ObjectId('6400d905e4b28f74a00ff687')
date: "2020-01-31"
negative: 245
deceased: 0
test: 312
time: 16
confirmed: 11
```

Type here to search | 24°C Mostly cloudy | ENG | 10:43 PM 03/02/2023

Invensis Technologies Pvt. Ltd. - / | Mail - abdulqadir.syed - Network | hcare - INV_TEAM - Slack | Data | Cloud: MongoDB Cloud

cloud.mongodb.com/v2/63a3f850eb7a144c3202b002#/metrics/replicaSet/6400bb19de519e522e409747/explorer/kafka/timeage/find

Atlas | Syedabdul's ... | Access Manager | Billing | All Clusters | Get Help | Syedabdul

Project 0 | Data Services | App Services | Charts

DEPLOYMENT
- Database
- Data Lake PREVIEW

SERVICES
- Triggers
- Data API
- Data Federation
- Search

SECURITY
- Database Access
- Network Access
- Advanced

Goto

SYEDABDUL'S ORG - 2022-12-22 > PROJECT 0 > DATABASES

Cluster0

VERSION 5.0.15 | REGION GCP Iowa (us-central1)

Overview | Real Time | Metrics | Collections | Search | Profiler | Performance Advisor | Online Archive | Cmd Line Tools

DATABASES: 1 COLLECTIONS: 7

VISUALIZE YOUR DATA | REFRESH

+ Create Database

Search Namespaces

- kafka
  - Time
  - oase
  - patientinfo
  - policy
  - region
  - search_trend
  - timeage

kafka.timeage

STORAGE SIZE: 4KB | LOGICAL DATA SIZE: 113.67KB | TOTAL DOCUMENTS: 1089 | INDEXES TOTAL SIZE: 4KB

Find | Indexes | Schema Anti-Patterns 0 | Aggregation | Search Indexes

INSERT DOCUMENT

FILTER { field: 'value' } | OPTIONS | Apply | Reset

QUERY RESULTS: 1-20 OF MANY

```
_id: ObjectId('6400d985e4b28f74a00ff72a')
date: "2020-03-02"
deceased: 0
time: 0
confirmed: 32
age: "0s"
```

```
_id: ObjectId('6400d985e4b28f74a00ff72b')
date: "2020-03-02"
deceased: 0
time: 0
confirmed: 169
age: "10s"
```

Type here to search | 24°C Mostly cloudy | ENG | 10:45 PM 03/02/2023

Invensis Technologies Pvt. Ltd. - / | Mail - abdulqadir.syed - Network | hcare - INV_TEAM - Slack | Data | Cloud: MongoDB Cloud

cloud.mongodb.com/v2/63a3f850eb7a144c3202b002#/metrics/replicaSet/6400bb19de519e522e409747/explorer/kafka/timegender/find

Atlas | Syedabdul's ... | Access Manager | Billing | All Clusters | Get Help | Syedabdul

Project 0 | Data Services | App Services | Charts

DEPLOYMENT
- Database
- Data Lake PREVIEW

SERVICES
- Triggers
- Data API
- Data Federation
- Search

SECURITY
- Database Access
- Network Access
- Advanced

Goto

SYEDABDUL'S ORG - 2022-12-22 > PROJECT 0 > DATABASES

Cluster0

VERSION 5.0.15 | REGION GCP Iowa (us-central1)

Overview | Real Time | Metrics | Collections | Search | Profiler | Performance Advisor | Online Archive | Cmd Line Tools

DATABASES: 1 COLLECTIONS: 8

VISUALIZE YOUR DATA | REFRESH

+ Create Database

Search Namespaces

- kafka
  - Time
  - oase
  - patientinfo
  - policy
  - region
  - search_trend
  - timeage
  - timegender

kafka.timegender

STORAGE SIZE: 4KB | LOGICAL DATA SIZE: 25.29KB | TOTAL DOCUMENTS: 242 | INDEXES TOTAL SIZE: 4KB

Find | Indexes | Schema Anti-Patterns 0 | Aggregation | Search Indexes

INSERT DOCUMENT

FILTER { field: 'value' } | OPTIONS | Apply | Reset

QUERY RESULTS: 1-20 OF MANY

```
_id: ObjectId('6400d9fde4b28f74a00ffb6c')
date: "2020-03-04"
deceased: 20
sex: "male"
time: "0"
confirmed: 1996
```

```
_id: ObjectId('6400d9fde4b28f74a00ffb6d')
date: "2020-03-04"
deceased: 12
sex: "female"
time: "0"
confirmed: 3332
```

Type here to search | 24°C Mostly cloudy | ENG | 10:47 PM 03/02/2023

Invensis Technologies Pvt. Ltd. - / | Mail - abdulqadir.syed - Network | hcare - INV_TEAM - Slack | Data | Cloud: MongoDB Cloud

cloud.mongodb.com/v2/63a3f850eb7a144c3202b002#/metrics/replicaSet/6400bb19de519e522e409747/explorer/kafka/time_province/find

Atlas | Syedabdul's ... | Access Manager | Billing | All Clusters | Get Help | Syedabdul

Project 0 | Data Services | App Services | Charts

DEPLOYMENT
- Database
- Data Lake PREVIEW

SERVICES
- Triggers
- Data API
- Data Federation
- Search

SECURITY
- Database Access
- Network Access
- Advanced

Goto

SYEDABDUL'S ORG - 2022-12-22 > PROJECT 0 > DATABASES

Cluster0

VERSION 5.0.15 | REGION GCP Iowa (us-central1)

Overview | Real Time | Metrics | Collections | Search | Profiler | Performance Advisor | Online Archive | Cmd Line Tools

DATABASES: 1 COLLECTIONS: 10 | VISUALIZE YOUR DATA | REFRESH

+ Create Database

Search Namespaces

- kafka
  - Time
  - oase
  - patientinfo
  - policy
  - region
  - search_trend
  - time_province
  - timeage
  - timegender
  - weather

kafka.time_province

STORAGE SIZE: 112KB | LOGICAL DATA SIZE: 365.22KB | TOTAL DOCUMENTS: 2771 | INDEXES TOTAL SIZE: 100KB

Find | Indexes | Schema Anti-Patterns 0 | Aggregation | Search Indexes

INSERT DOCUMENT

FILTER { field: 'value' } | OPTIONS | Apply | Reset

QUERY RESULTS: 1-20 OF MANY

```
_id: ObjectId('6400da4fe4b28f74a00ffc5f')
date: "2020-01-20"
deceased: 0
province: "Gwangju"
time: "16"
confirmed: 0
released: 0
```

```
_id: ObjectId('6400da4fe4b28f74a00ffc60')
date: "2020-01-20"
deceased: 0
province: "Chungcheongbuk-do"
time: "16"
confirmed: 0
released: 0
```

Type here to search | 24°C Mostly cloudy | ENG | 10:51 PM 03/02/2023

Invensis Technologies Pvt. Ltd. - / | Mail - abdulqadir.syed - Network | hcare - INV_TEAM - Slack | Data | Cloud: MongoDB Cloud

cloud.mongodb.com/v2/63a3f850eb7a144c3202b002#/metrics/replicaSet/6400bb19de519e522e409747/explorer/kafka/weather/find

Atlas | Syedabdul's ... | Access Manager | Billing | All Clusters | Get Help | Syedabdul

Project 0 | Data Services | App Services | Charts

DEPLOYMENT
- Database
- Data Lake PREVIEW

SERVICES
- Triggers
- Data API
- Data Federation
- Search

SECURITY
- Database Access
- Network Access
- Advanced

Goto

SYEDABDUL'S ORG - 2022-12-22 > PROJECT 0 > DATABASES

Cluster0

VERSION 5.0.15 | REGION GCP Iowa (us-central1)

Overview | Real Time | Metrics | Collections | Search | Profiler | Performance Advisor | Online Archive | Cmd Line Tools

DATABASES: 1 COLLECTIONS: 10 | VISUALIZE YOUR DATA | REFRESH

+ Create Database

Search Namespaces

- kafka
  - Time
  - oase
  - patientinfo
  - policy
  - region
  - search_trend
  - time_province
  - timeage
  - timegender
  - weather

kafka.weather

STORAGE SIZE: 2.71MB | LOGICAL DATA SIZE: 6.09MB | TOTAL DOCUMENTS: 26271 | INDEXES TOTAL SIZE: 1.48MB

Find | Indexes | Schema Anti-Patterns 0 | Aggregation | Search Indexes

INSERT DOCUMENT

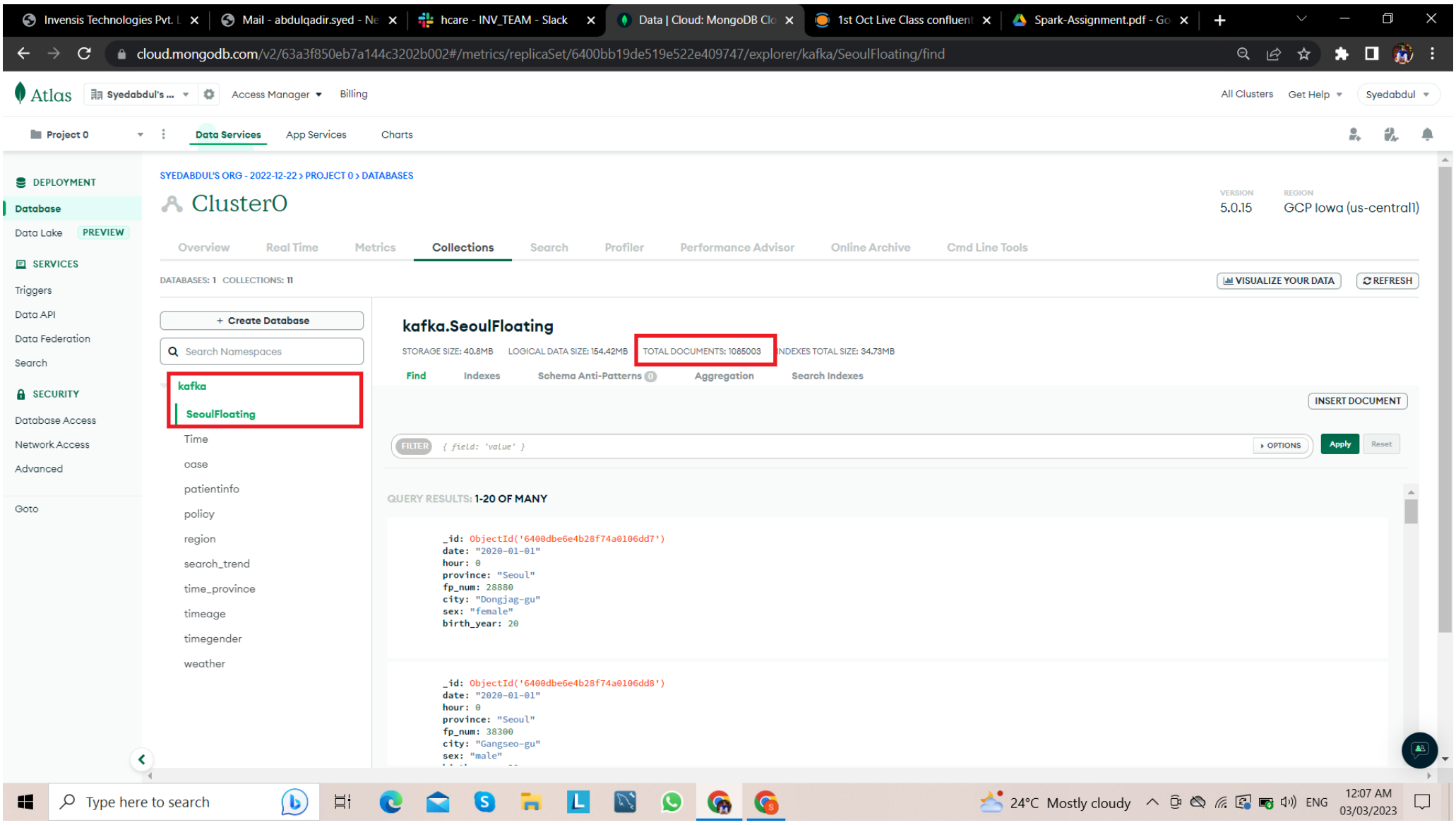
FILTER { field: 'value' } | OPTIONS | Apply | Reset

QUERY RESULTS: 1-20 OF MANY

```
_id: ObjectId('6400dae3e4b28f74a0100733')
date: "2016-01-01"
precipitation: 0
avg_relative_humidity: 52.1
code: 11000
province: "Busan"
avg_temp: 5.3
min_temp: 1.1
most_wind_direction: 340
max_wind_speed: 7.4
max_temp: 10.9
```

```
_id: ObjectId('6400dae3e4b28f74a0100734')
date: "2016-01-01"
precipitation: 0
avg_relative_humidity: 70.5
```

Type here to search | 24°C Mostly cloudy | ENG | 10:54 PM 03/02/2023



4. Collect your data as a pyspark dataframe and perform different operations.

Note: Consider only three files for creating a data frame among all case, region and Time Province

a. Read the data, show it and Count the number of records.

b. Describe the data with a describe function.

c. If there is any duplicate value drop it.

d. Use limit function for showcasing a limited number of records.

e. If you find the column name is not suitable, change the column name.[optional]

f. Select the subset of the columns.

g. If there is any null value, fill it with any random value or drop it.

h. Filter the data based on different columns or variables and do the best analysis.

For example: We can filter a data frame using multiple conditions using AND(&), OR(|) and NOT(~) conditions. For example, we may want to find out all the different infection case in Daegu Province with more than 10 confirmed cases.

i. Sort the number of confirmed cases. Confirmed column is there in the dataset. Check with descending sort also.

j. In case of any wrong data type, cast that data type from integer to string or string to integer.

k. Use group by on top of province and city column and agg it with sum of confirmed cases. For example df.groupBy(["province","city"]).agg(function.sum("confirmed")

l. For joins we will need one more file. you can use region file. User different different join methods, for example cases. join(regions, ['province', 'city'],how='left') You can do your best analysis.

5. If you want, you can also use SQL with data frames. Let us try to run some SQL on the cases table.

For example:

cases.registerTempTable('cases table')

newDF = sqlContext.sql('select * from cases table where confirmed>100') newDF.show()

Here is a example how you can use df for sql now you can perform

various operations with GROUP BY, HAVING, AND ORDER BY

ABOVE MENTION QUERIES ARE IN ,Mongodb_Pyspark.ipynb file