BTS – MBDS
Big Data Infrastructure

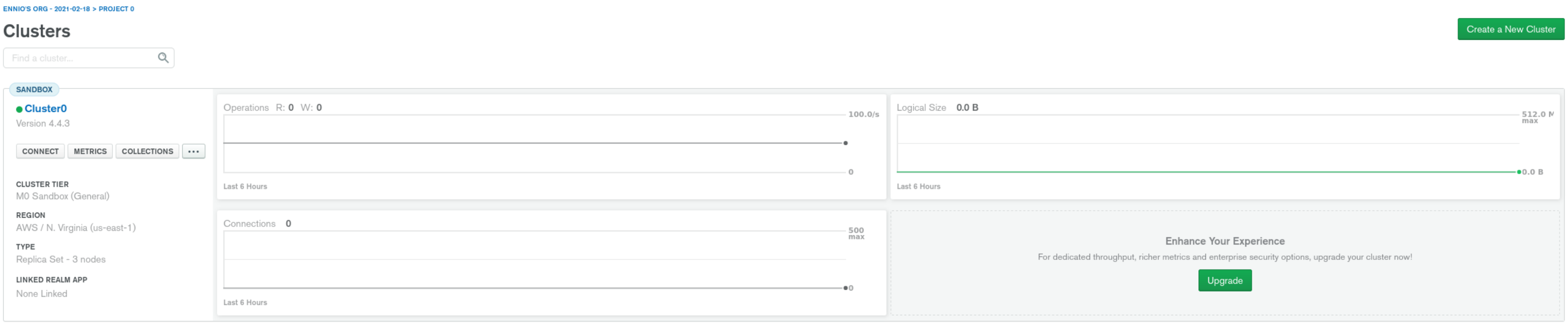
A2: MongoDB and DataBricks

Ennio Maldonado

18 February, 2021

MongoDB

MongoDB cluster creation



ENNIO'S ORG - 2021-02-18 > PROJECT 0

Clusters

Cluster0

Version 4.4.3

CLUSTER TIER
M0 Sandbox (General)

REGION
AWS / N. Virginia (us-east-1)

TYPE
Replica Set - 3 nodes

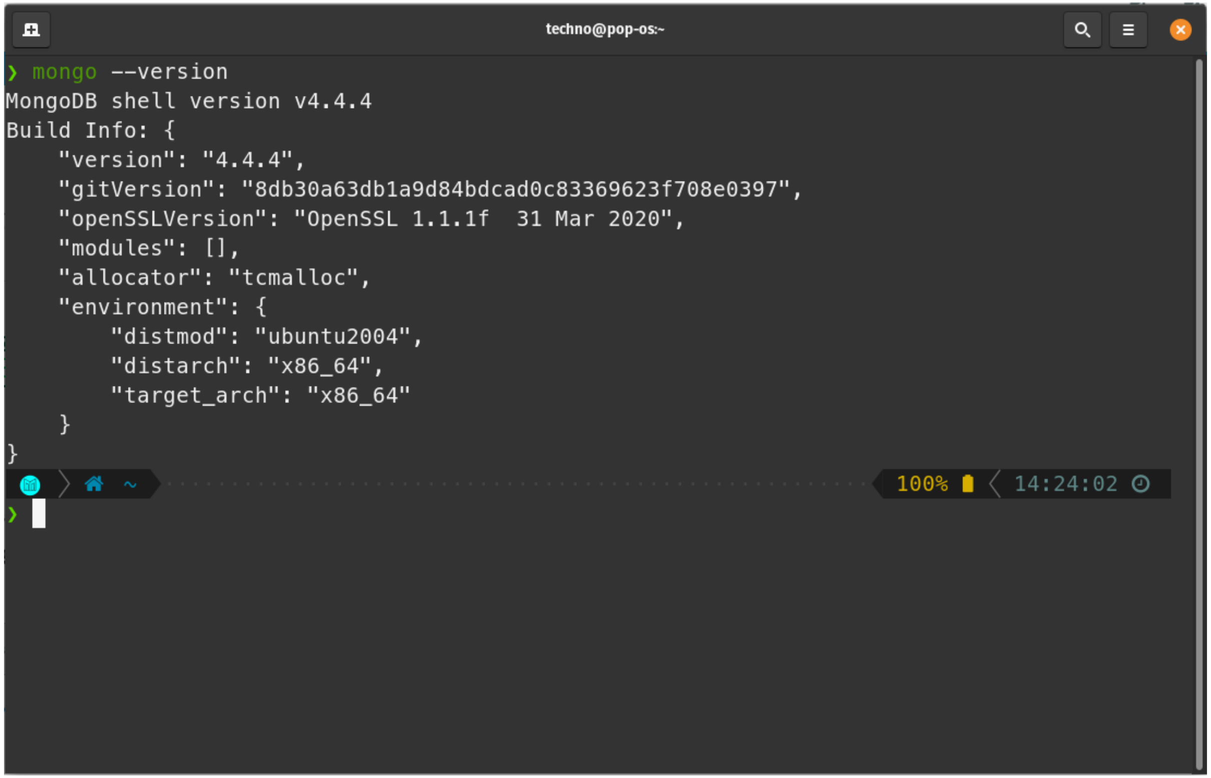
LINKED REALM APP
None Linked

Enhance Your Experience

For dedicated throughput, richer metrics and enterprise security options, upgrade your cluster now!

MongoDB

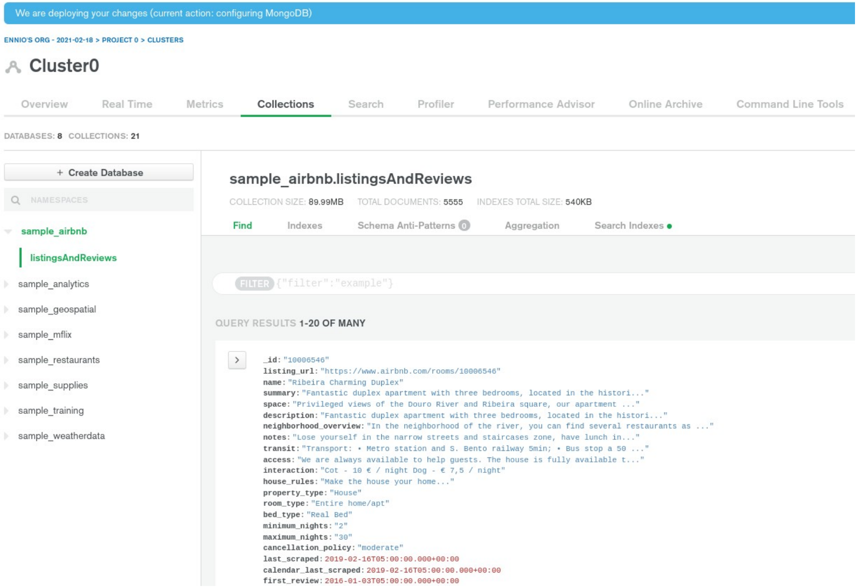
Installed mongodb shell tool



```
❯ mongo --version
MongoDB shell version v4.4.4
Build Info: {
    "version": "4.4.4",
    "gitVersion": "8db30a63db1a9d84bdcad0c83369623f708e0397",
    "openSSLVersion": "OpenSSL 1.1.1f  31 Mar 2020",
    "modules": [],
    "allocator": "tcmalloc",
    "environment": {
        "distmod": "ubuntu2004",
        "distarch": "x86_64",
        "target_arch": "x86_64"
    }
}
 ~                                                    100%  14:24:02
❯
```

MongoDB

Loaded sample data_set from mongodb for testing connection.



We are deploying your changes (current action: configuring MongoDB)

Cluster0

DATABASES: 8 COLLECTIONS: 21

- sample_airbnb
  - listingsAndReviews
- sample_analytics
- sample_geospatial
- sample_mflix
- sample_restaurants
- sample_supplies
- sample_training
- sample_weatherdata

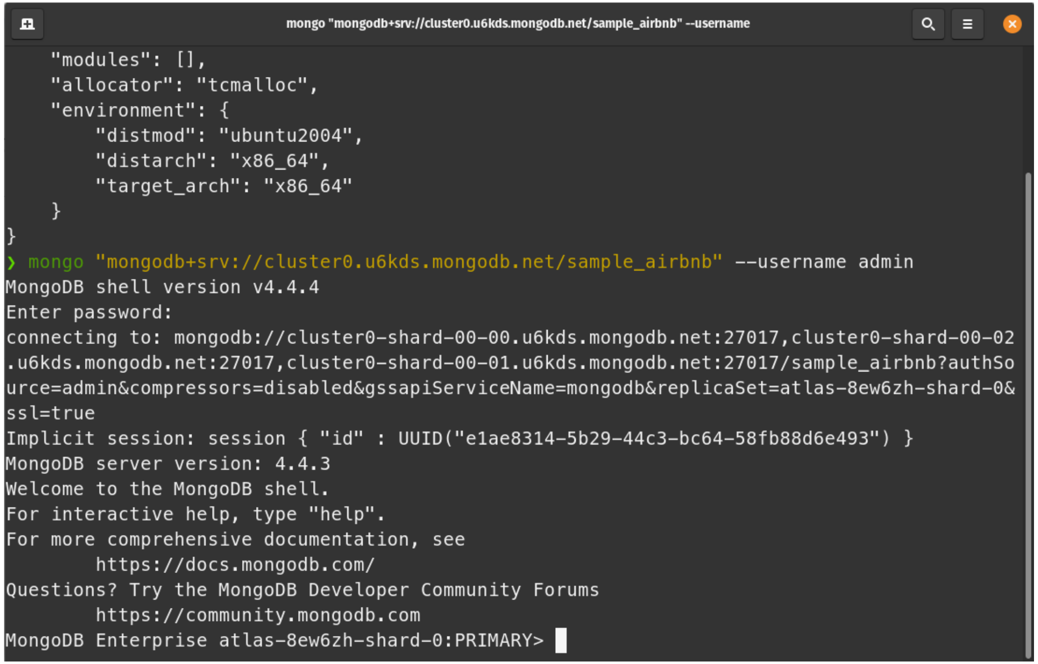
sample_airbnb.listingsAndReviews

COLLECTION SIZE: 89.99MB TOTAL DOCUMENTS: 5555 INDEXES TOTAL SIZE: 540KB

FILTER {"filter":"example"}

QUERY RESULTS 1-20 OF MANY

```
_id: "10006546"
listing_url: "https://www.airbnb.com/rooms/10006546"
name: "Ribeira Charming Duplex"
summary: "Fantastic duplex apartment with three bedrooms, located in the histori..."
space: "Privileged views of the Douro River and Ribeira square, our apartment ..."
description: "Fantastic duplex apartment with three bedrooms, located in the histori..."
neighborhood_overview: "In the neighborhood of the river, you can find several restaurants as ..."
notes: "Lose yourself in the narrow streets and staircases zone, have lunch in..."
transit: "Transport: • Metro station and S. Bento railway 5min; • Bus stop a 50 ..."
access: "We are always available to help guests. The house is fully available t..."
interaction: "Cot - 10 € / night Dog - € 7,5 / night"
house_rules: "Make the house your home..."
property_type: "House"
room_type: "Entire home/apt"
bed_type: "Real Bed"
minimum_nights: "2"
maximum_nights: "30"
cancellation_policy: "moderate"
last_scraped: 2019-02-16T05:00:00.000+00:00
calendar_last_scraped: 2019-02-16T05:00:00.000+00:00
first_review: 2016-01-03T05:00:00.000+00:00
```



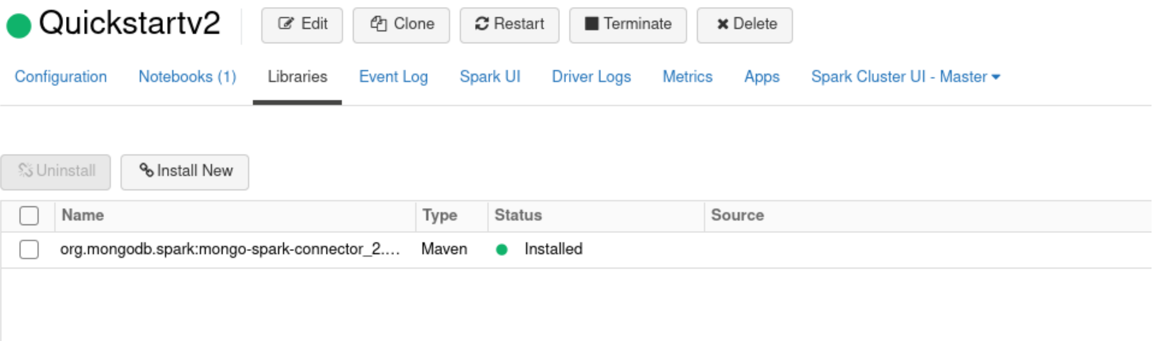
mongo "mongodb+srv://cluster0.u6kds.mongodb.net/sample_airbnb" --username

```
    "modules": [],
    "allocator": "tcmalloc",
    "environment": {
        "distmod": "ubuntu2004",
        "distarch": "x86_64",
        "target_arch": "x86_64"
    }
}
❯ mongo "mongodb+srv://cluster0.u6kds.mongodb.net/sample_airbnb" --username admin
MongoDB shell version v4.4.4
Enter password:
connecting to: mongodb://cluster0-shard-00-00.u6kds.mongodb.net:27017,cluster0-shard-00-02
.u6kds.mongodb.net:27017,cluster0-shard-00-01.u6kds.mongodb.net:27017/sample_airbnb?authSo
urce=admin&compressors=disabled&gssapiServiceName=mongodb&replicaSet=atlas-8ew6zh-shard-0&
ssl=true
Implicit session: session { "id" : UUID("e1ae8314-5b29-44c3-bc64-58fb88d6e493") }
MongoDB server version: 4.4.3
Welcome to the MongoDB shell.
For interactive help, type "help".
For more comprehensive documentation, see
        https://docs.mongodb.com/
Questions? Try the MongoDB Developer Community Forums
        https://community.mongodb.com
MongoDB Enterprise atlas-8ew6zh-shard-0:PRIMARY>
```

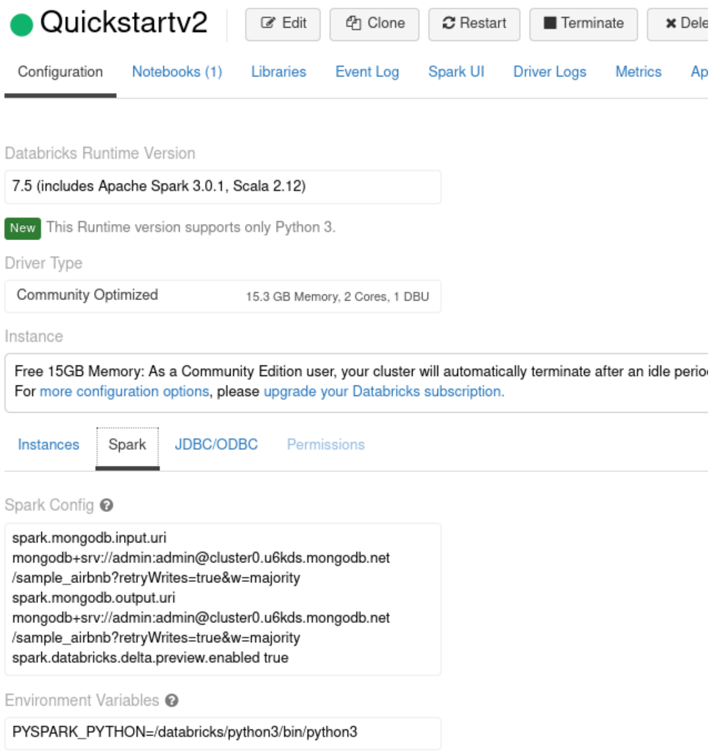
MongoDB on DataBricks

Connect MongoDB to DataBricks followed this guide:

https://docs.databricks.com/data/data-sources/mongodb.html



□	Name	Type	Status	Source
□	org.mongodb.spark:mongo-spark-connector_2....	Maven	● Installed	



Databricks Runtime Version

7.5 (includes Apache Spark 3.0.1, Scala 2.12)

New This Runtime version supports only Python 3.

Driver Type

Community Optimized 15.3 GB Memory, 2 Cores, 1 DBU

Instance

Free 15GB Memory: As a Community Edition user, your cluster will automatically terminate after an idle perio
For more configuration options, please upgrade your Databricks subscription.

Spark Config

```
spark.mongodb.input.uri
mongodb+srv://admin:admin@cluster0.u6kds.mongodb.net
/sample_airbnb?retryWrites=true&w=majority
spark.mongodb.output.uri
mongodb+srv://admin:admin@cluster0.u6kds.mongodb.net
/sample_airbnb?retryWrites=true&w=majority
spark.databricks.delta.preview.enabled true
```

Environment Variables

```
PYSPARK_PYTHON=/databricks/python3/bin/python3
```