

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

Create a New Cluster

Find a cluster...

SANDBOX

Cluster0

Version 4.4.3

CONNECT METRICS COLLECTIONS ...

CLUSTER TIER

M0 Sandbox (General)

REGION

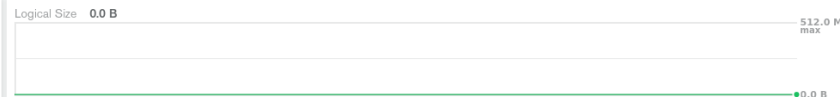
AWS / N. Virginia (us-east-1)

TYPE

Replica Set - 3 nodes

LINKED REALM APP

None Linked



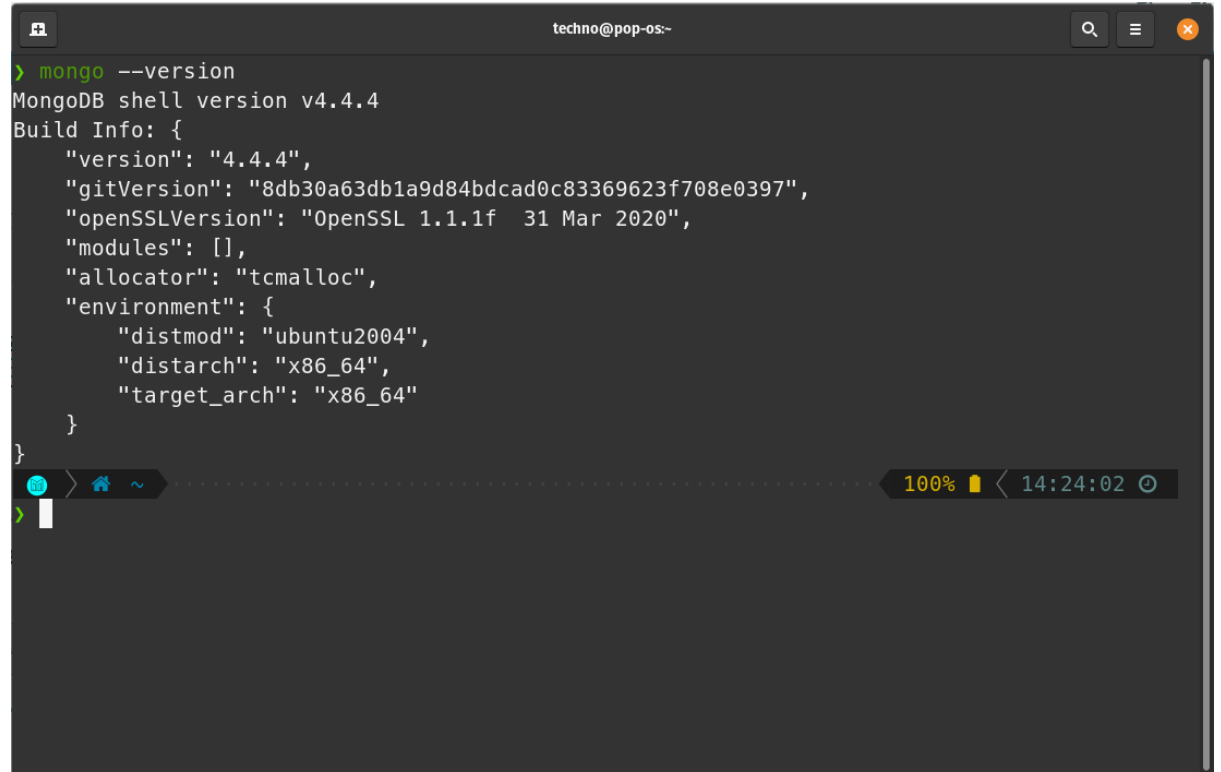
Enhance Your Experience

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

Upgrade

MongoDB

Installed mongodb shell tool



```
techno@pop-os-~  
> 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"  
  }  
}
```

The terminal window shows the output of the `mongo --version` command. The output displays the MongoDB shell version as v4.4.4 and provides detailed build information, including the version, git commit hash, OpenSSL version, modules, allocator, and environment details. The environment details specify the distribution as ubuntu2004, the architecture as x86_64, and the target architecture as x86_64. The terminal window also shows a status bar at the bottom with a 100% battery indicator and the time 14:24:02.

MongoDB

Loaded sample data_set from mongodb for testing connection.

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

ENNO'S ORG - 2021-02-18 > PROJECT 0 > CLUSTERS

Cluster0

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

DATABASES: 8 COLLECTIONS: 21

+ Create Database

NAMESPACES

- 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: 640KB

Find Indexes Schema Anti-Patterns Aggregation Search Indexes

FILTERED ("filter": "example")

QUERY RESULTS 1-20 OF MANY

```
{
  "_id": "10006540",
  "listing_url": "https://www.airbnb.com/rooms/10006540",
  "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 train; • 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 admin

{"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>

Quickstartv2

EditCloneRestartTerminateDelete

ConfigurationNotebooks (1)LibrariesEvent LogSpark UIDriver LogsMetricsAppsSpark Cluster UI - Master

UninstallInstall New

<input type="checkbox"/>	Name	Type	Status	Source
<input type="checkbox"/>	org.mongodb.spark:mongo-spark-connector_2....	Maven	● Installed	

Quickstartv2

EditCloneRestartTerminateDelete

ConfigurationNotebooks (1)LibrariesEvent LogSpark UIDriver LogsMetricsAp

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 Optimized15.3 GB Memory, 2 Cores, 1 DBU

Instance

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

InstancesSparkJDBC/ODBCPermissions

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