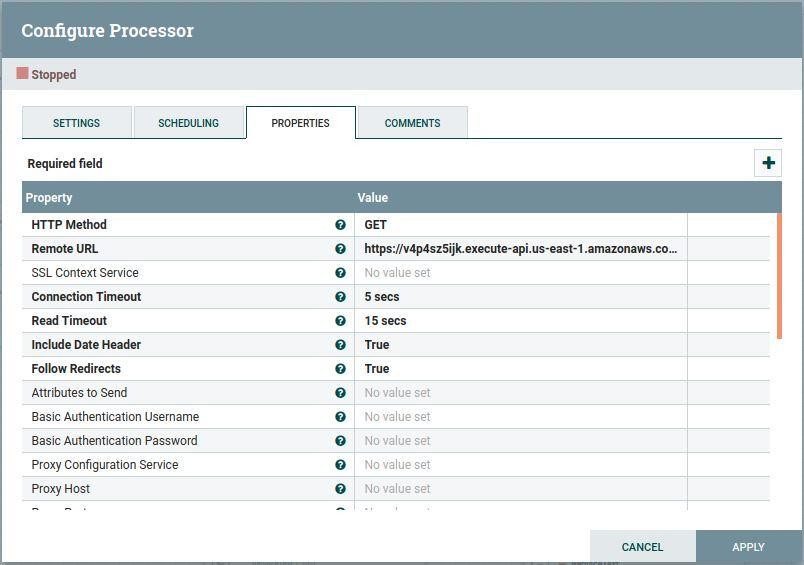
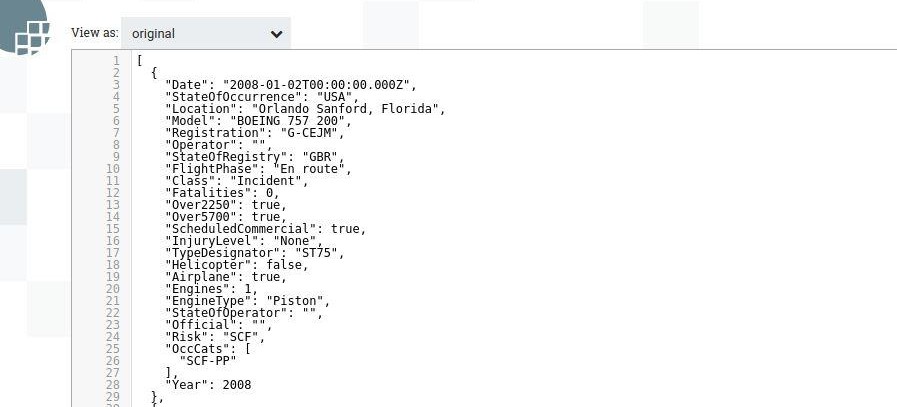
Project Execution Documentation

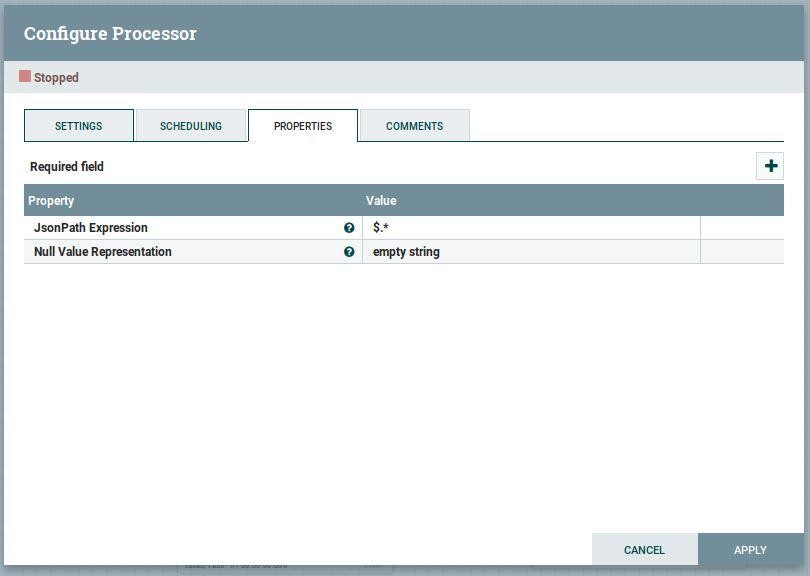
# Data extraction is done using NiFi and store in HDFS and MySQL

* Using **InvokeHttp processor we** will get the airline data from the API by specifying the API key generated from the invokeHTTP processor . which is followed by the below configurations in the processor.
* Data URL : https://[www.icao.int/safety/istars/pages/api-data-service.aspx](http://www.icao.int/safety/istars/pages/api-data-service.aspx)
* How to get API Key : Signup using the above link and get the api key and click on send api key https://v4p4sz5ijk.execute-api.us-east- 1.amazonaws.com/anbdata/occurrences/class/incident



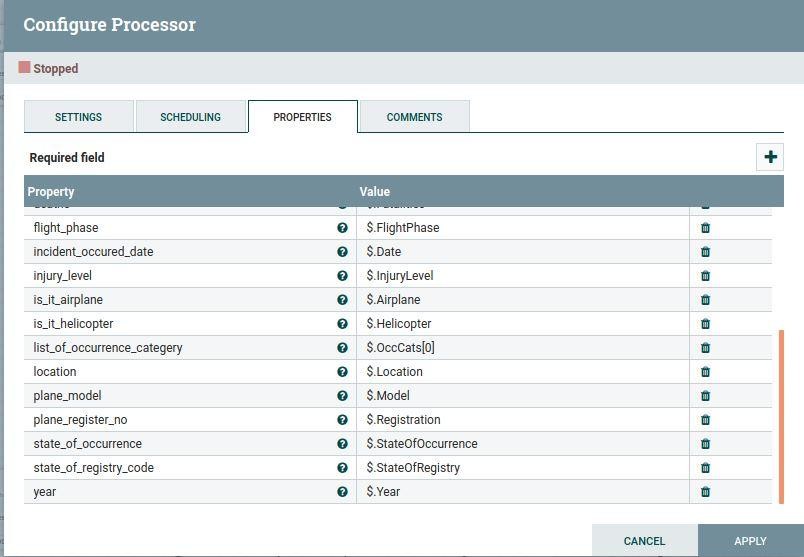
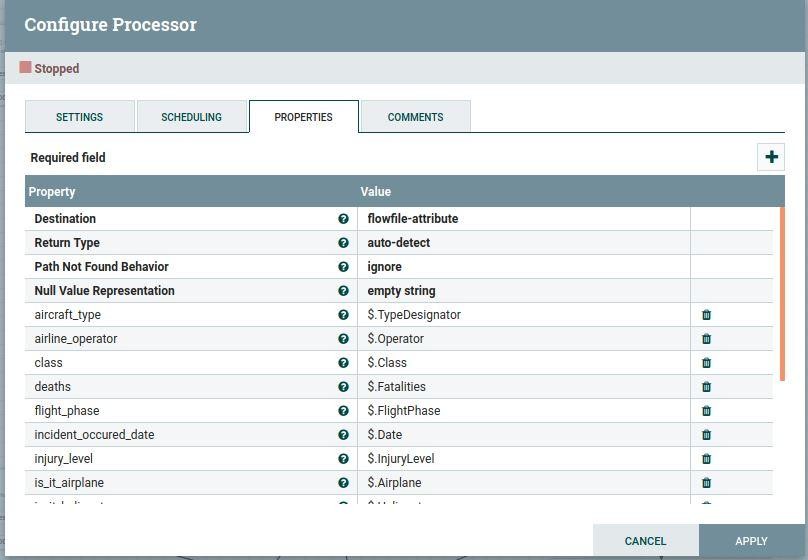
We will get response data as below .

* Split the JSON array from the above data by configuring **SplitJson Processor** as shown below.

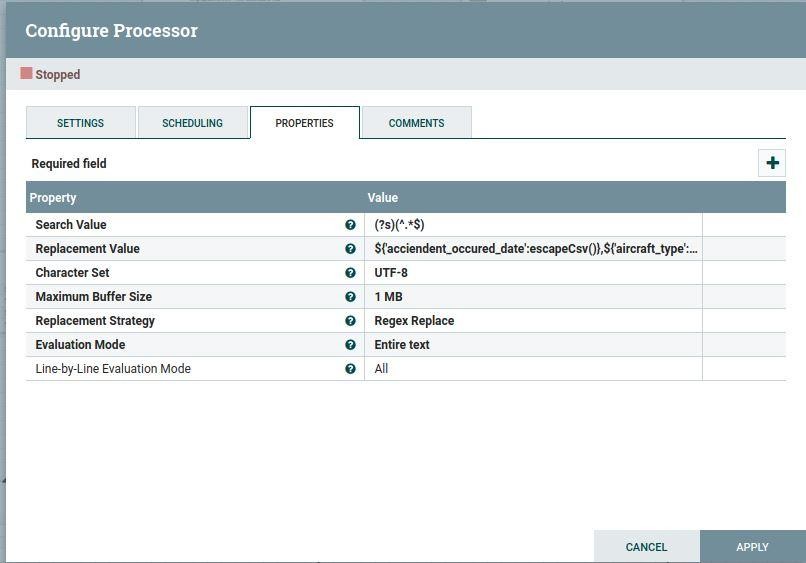


* Evaluate required fields from the data as below images using EvaluateJsonPath to flatten the complex data format to flat csv format to store in HDFS

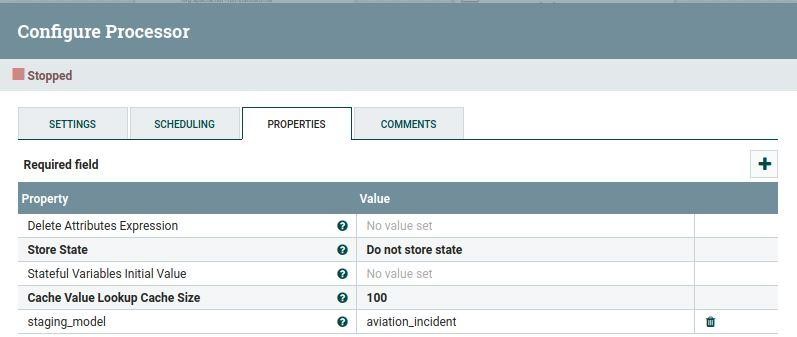
.



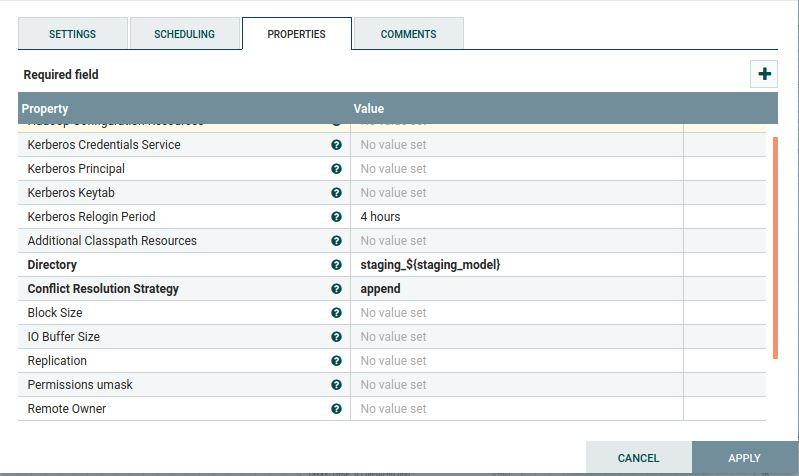
* After evaluating the required fields, convert into CSV format data by placing fields into an order using **ReplaceText** processor as below.



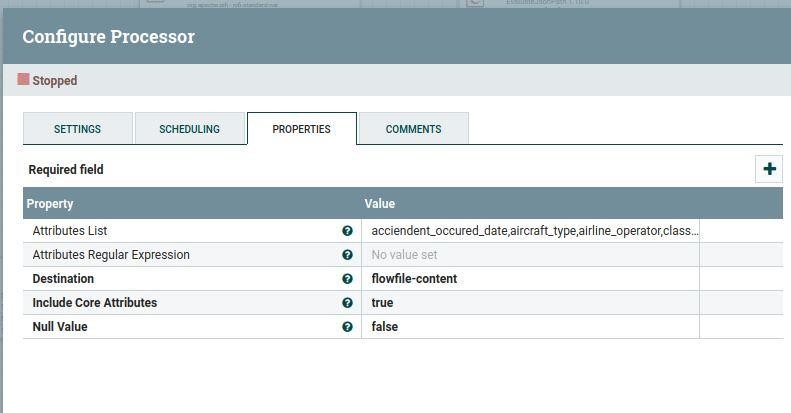
* Specify the name of staging\_model by the table name using **UpdateAttribute** Processor



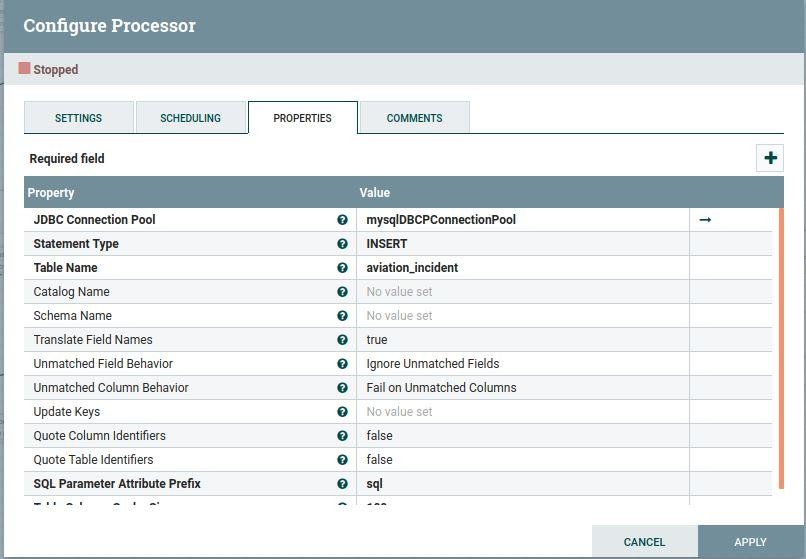
* Then store data into directory in the HDFS using PUTHDFS processor by using following configurations



* At the same time after giving the name to the staging\_model we write attributes into flowfile content as json format by passing the AttributeList using **AttributestoJson** Processor.

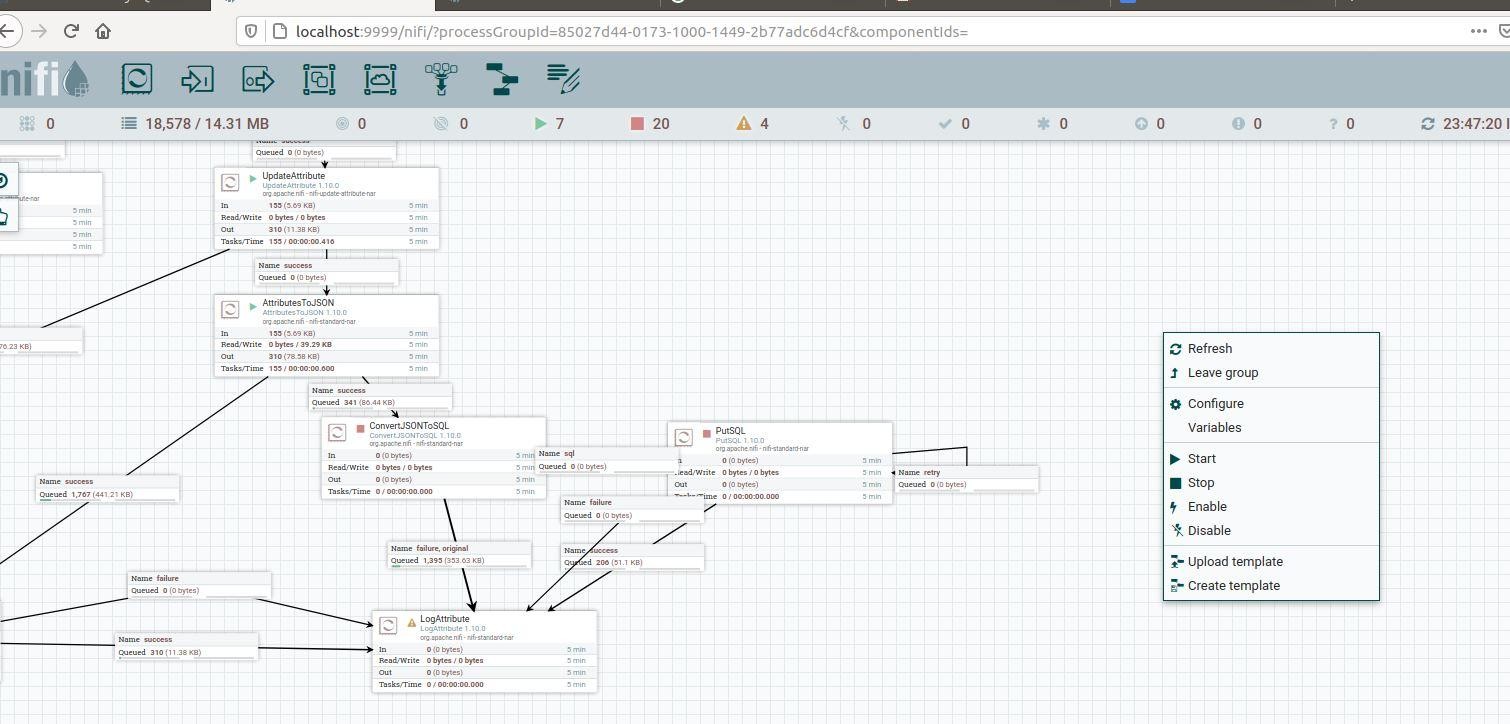


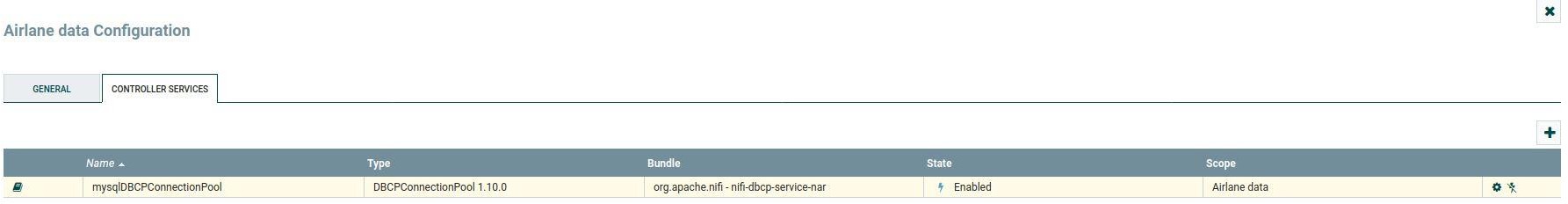
* Then we **convert** a **JSON**-formatted FlowFile into an INSERT **SQL** statement using ConvertJsonToSQL



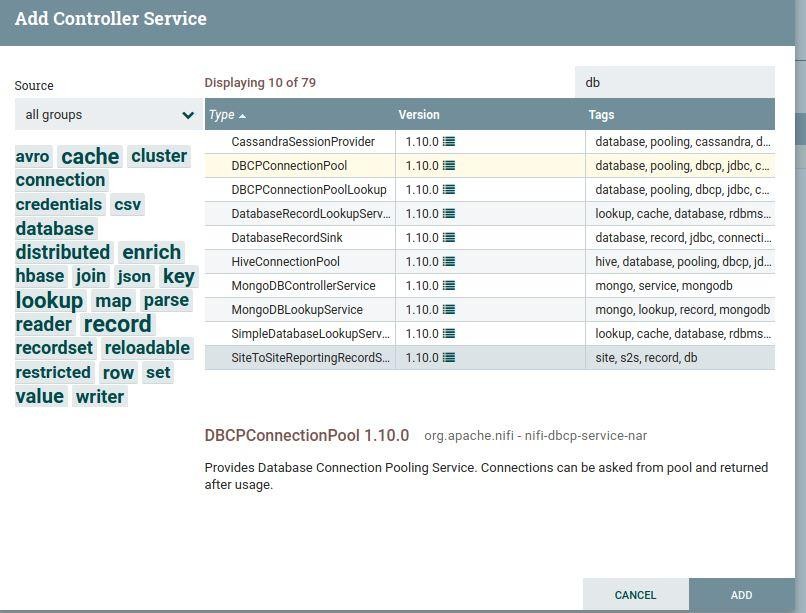
Note : before loading data into the mysql we need to create table in HIVE

# Create JDBConnectionPoolService in nifi ( for mysql)

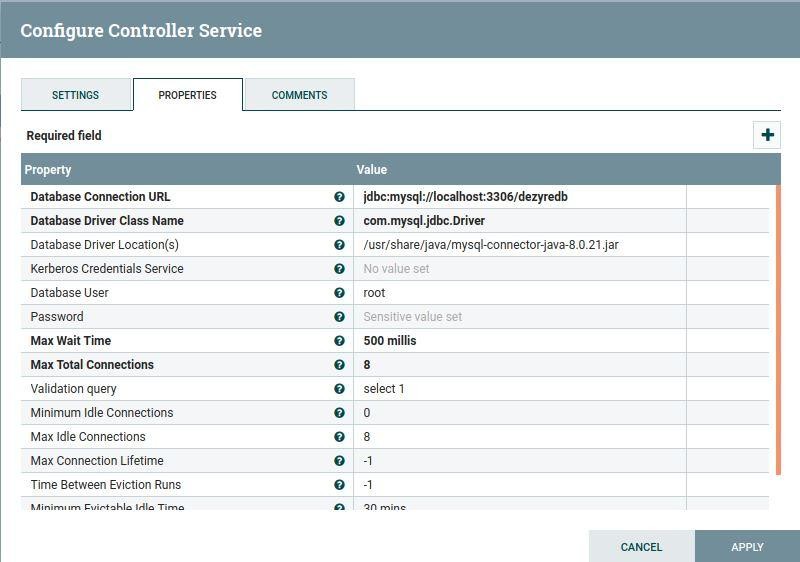
* Right click on the nifi Canvas select the “configure”
* Click on the “ + “ symbol at the right corner.



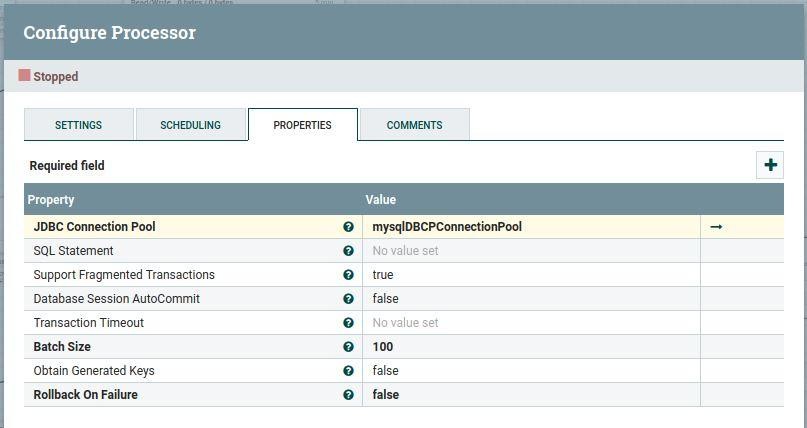
* Select the DBconnectionPool service then click on add



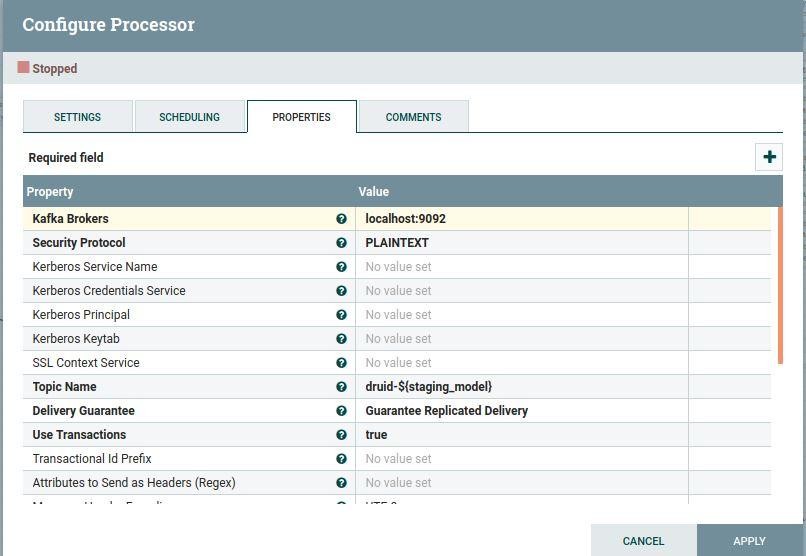
* Configured as in the below image for the mysql Localhost



* + Executes a SQL INSERT command. The content of an incoming FlowFile is expected to be the SQL command to execute using PutSQL processor



* At the same time after writing attributes into flowfile content as json format by passing the AttributeList using **AttributestoJson** Processor.
* Publishing Data to Kafka topic through the Publish Kafka processor by configuring as below in the **PublishKafka** processor



* Once the data is extracted and stored in HDFS. We will next index the data from HDFS into Druid as follows .
* Open the druid UI using VNC server .
* Navigate to Load data tab

# 1. Data ingesting into DRUID from HDFS by indexing from hadoop

To configure HDFS to talk to druid , following configurations need to be made in

/home/ubuntu/apache-druid-0.18.1/conf/druid/single-server/micro- quickstart/\_common/common.runtime.properties file .

Configuration for HDFS

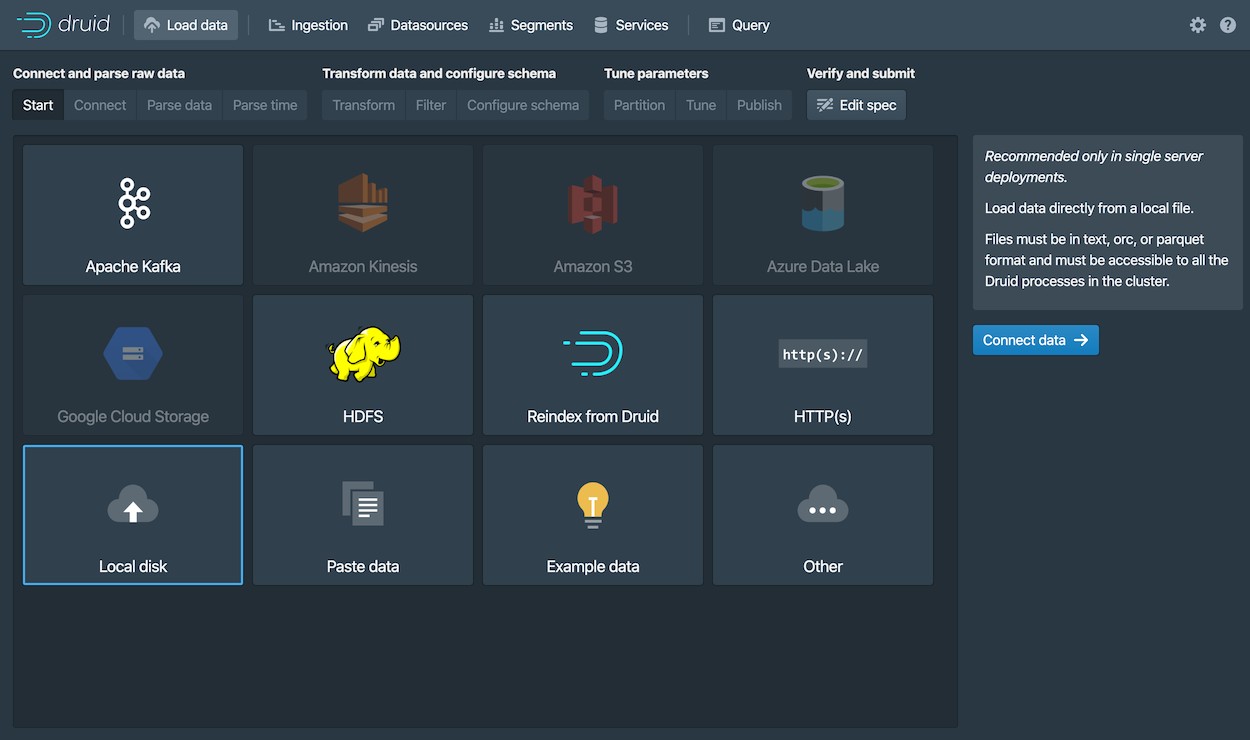
|  |  |  |  |
| --- | --- | --- | --- |
| Property | Possible Values | Descri ption | Default |
| druid.storage.type | hdfs |  | Must be set. |
| druid.storage.storageDirectory |  | Direct ory for storing segme nts. | Must be set. |
| druid.hadoop.security.kerberos. principal | [druid@EXAMPLE.COM](mailto:druid@EXAMPLE.COM) | Princip al user name | empty |
| druid.hadoop.security.kerberos. keytab | /etc/security/keytabs/druid.headlessU ser.keytab | Path to keytab file | empty |

Besides the above settings, you also need to include all Hadoop configuration files (such as core- site.xml, hdfs-site.xml) in the Druid classpath. One way to do this is copying all those files under

${DRUID\_HOME}/conf/\_common.

Now we can load the data from HDFS to Druid .

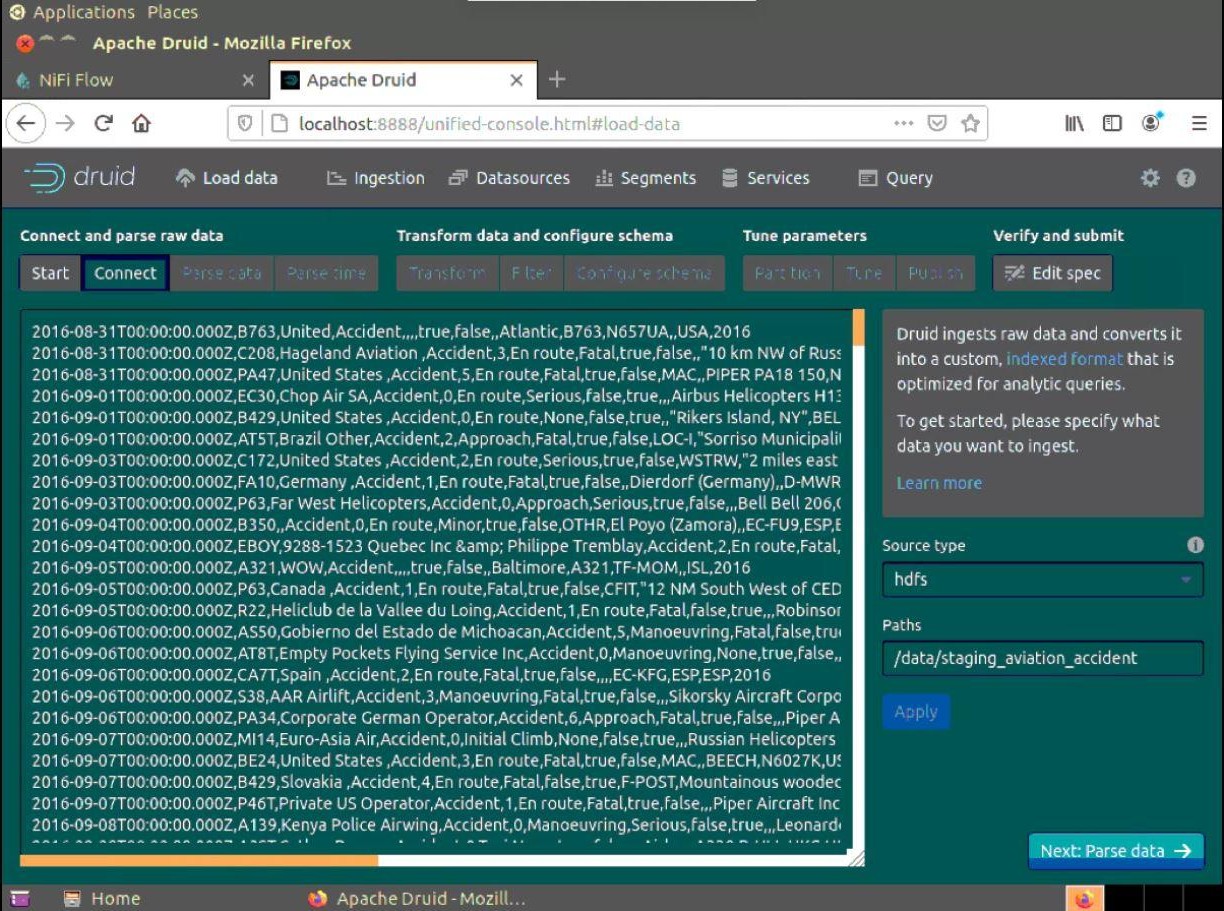
* Click Load data from the Druid console header ().
* Select the HDFS tile and then click Connect data.



Enter the following values:

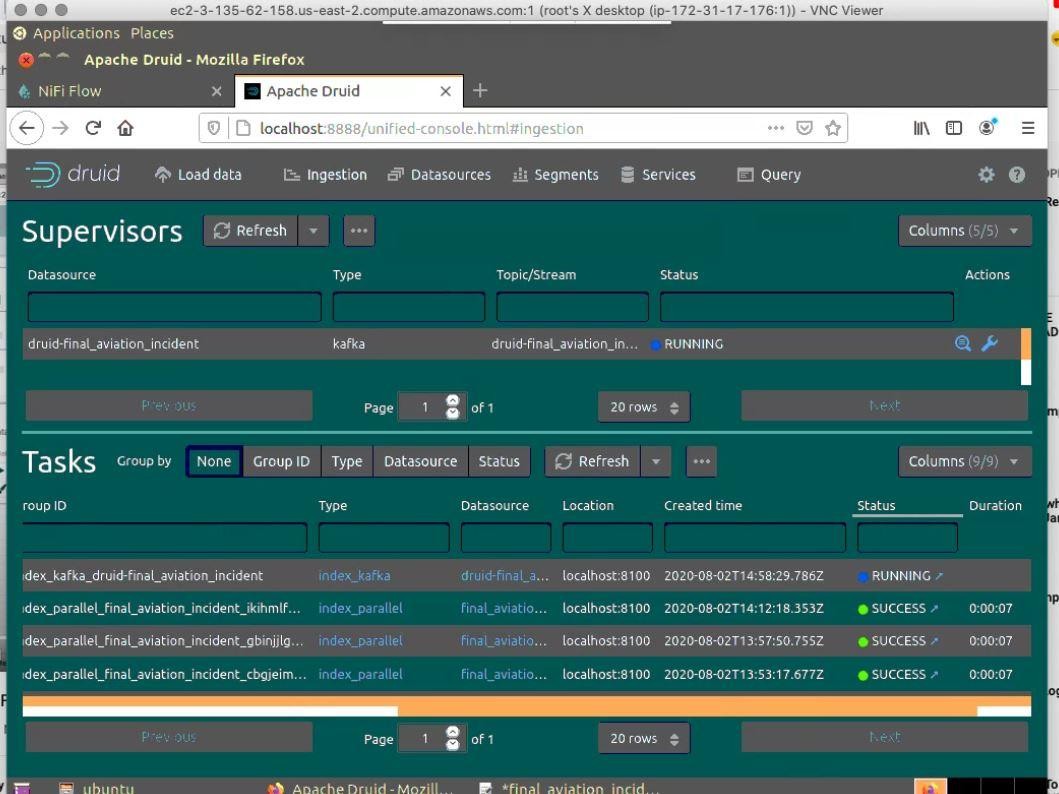
Select the source type **hdfs**

* Base directory: data/staging\_aviation\_incident and click on Apply.
* Data will be displayed in the console as shown in the figure.



# Streaming Data Consume from the kafka topic into Druid data source

* Here we are consuming the data from the kafka topic by submitting the kafka-supervisor.json file
* Copy the code from supervisor.json file code into load data by configuring Kafka broker URL and topic name as localhost:9092 , topic name



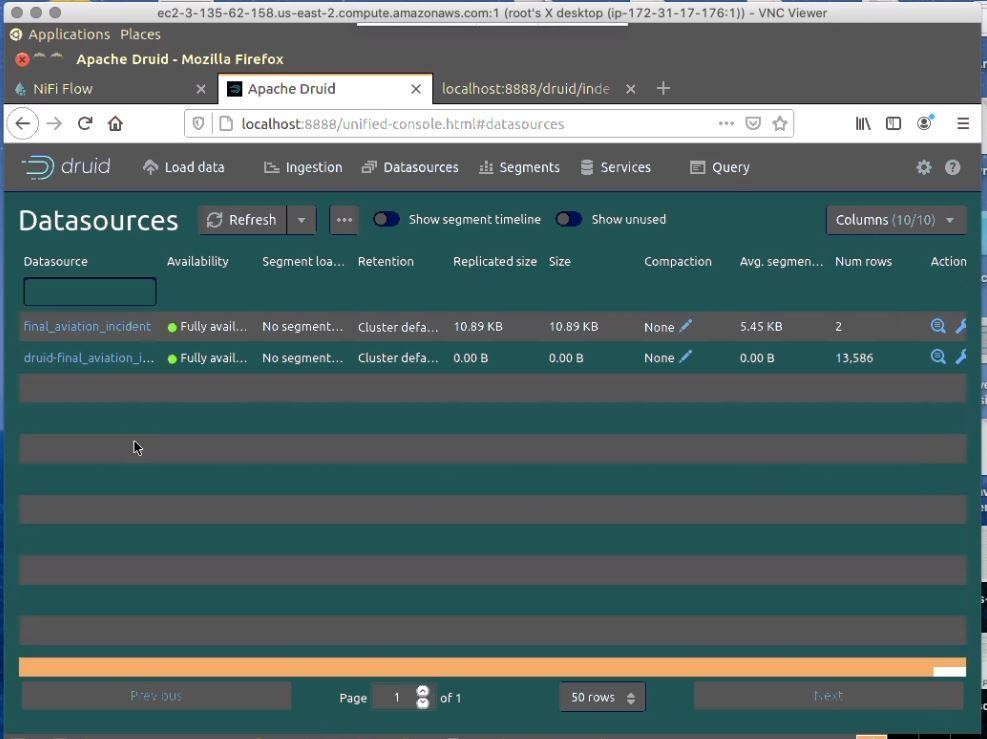
# Data Querying from Druid

You can now see the data as a datasource in the console and try out a query, as follows:

* Click Data Sources from the console header.

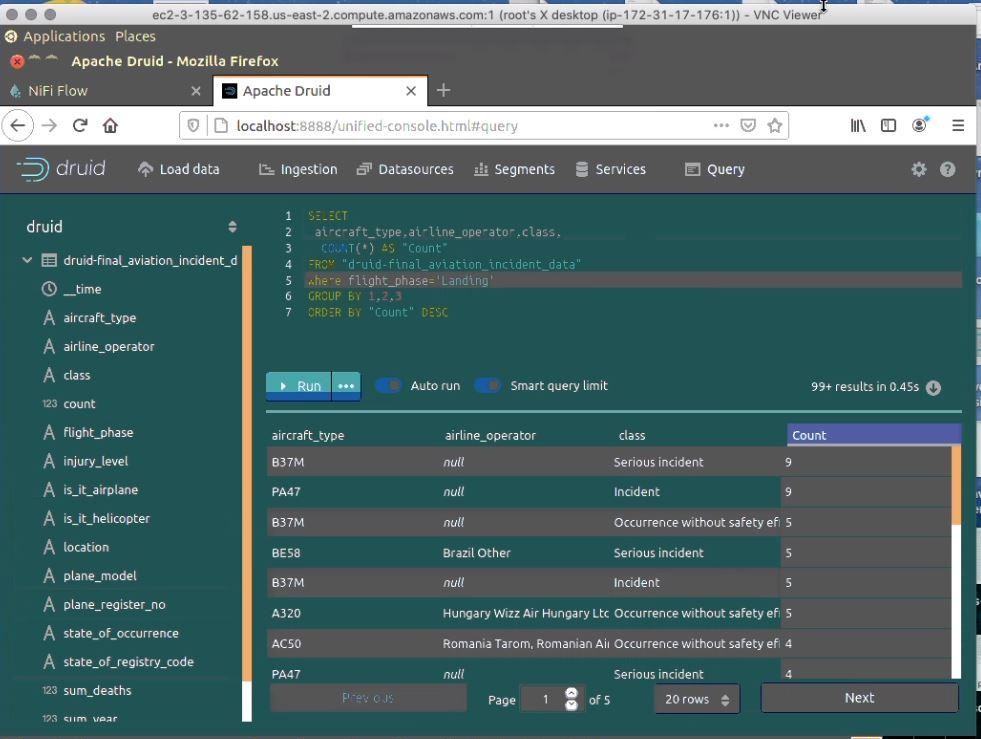
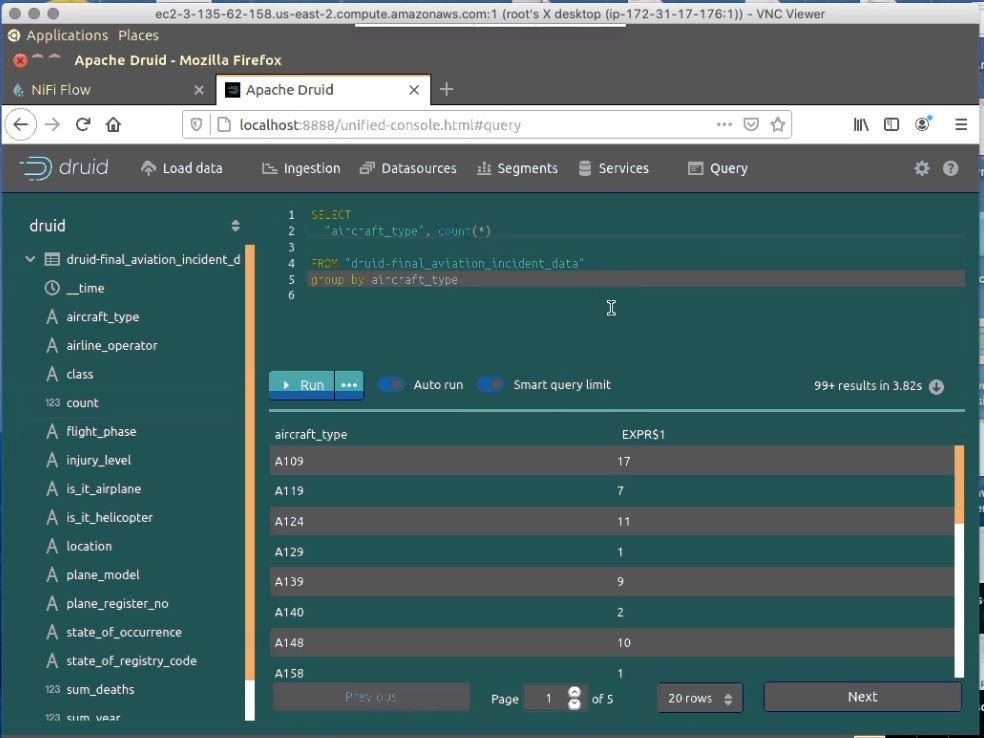
If the wikipedia datasource doesn't appear, wait a few moments for the segment to finish loading. A datasource is queryable once it is shown to be "Fully available" in the Availability column.

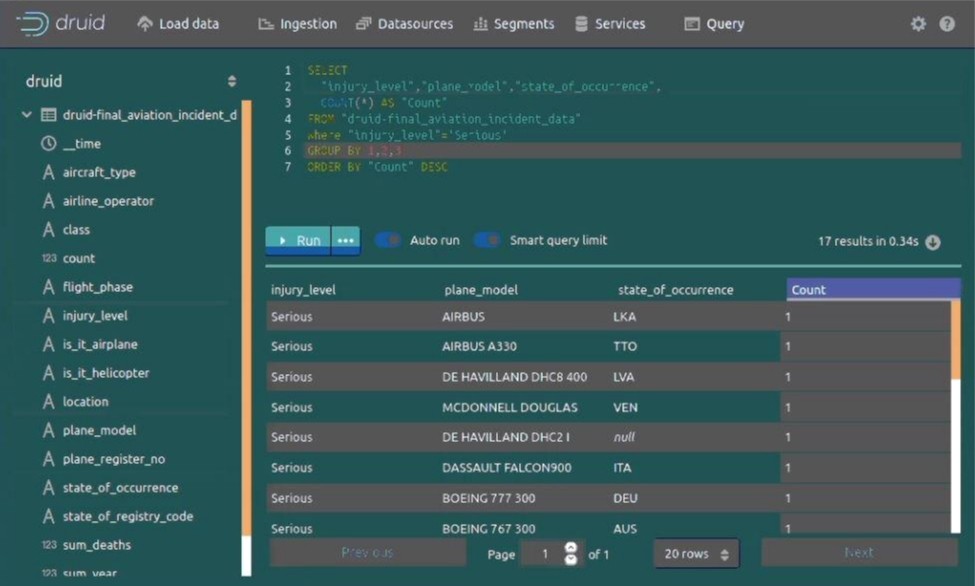
* When the datasource is available, open the Actions menu () for that datasource and choose Query with SQL.

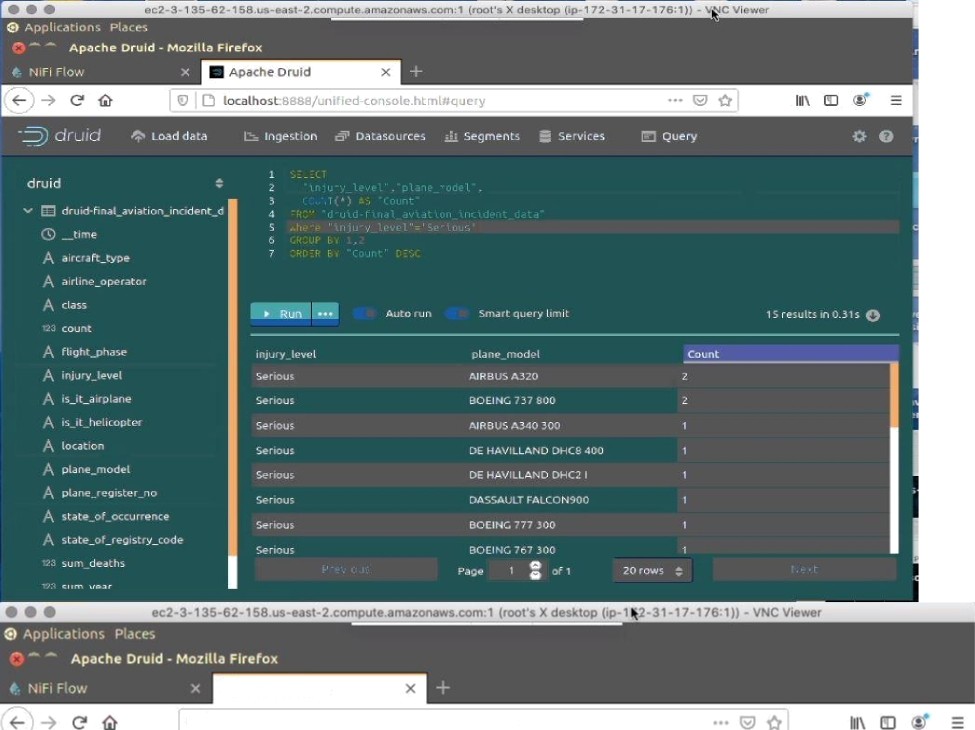


1. Run the pre populated query, SELECT \* FROM "druid-final\_aviation\_incident" to see the results.
2. Then, run the queries given in the Druid SQL code

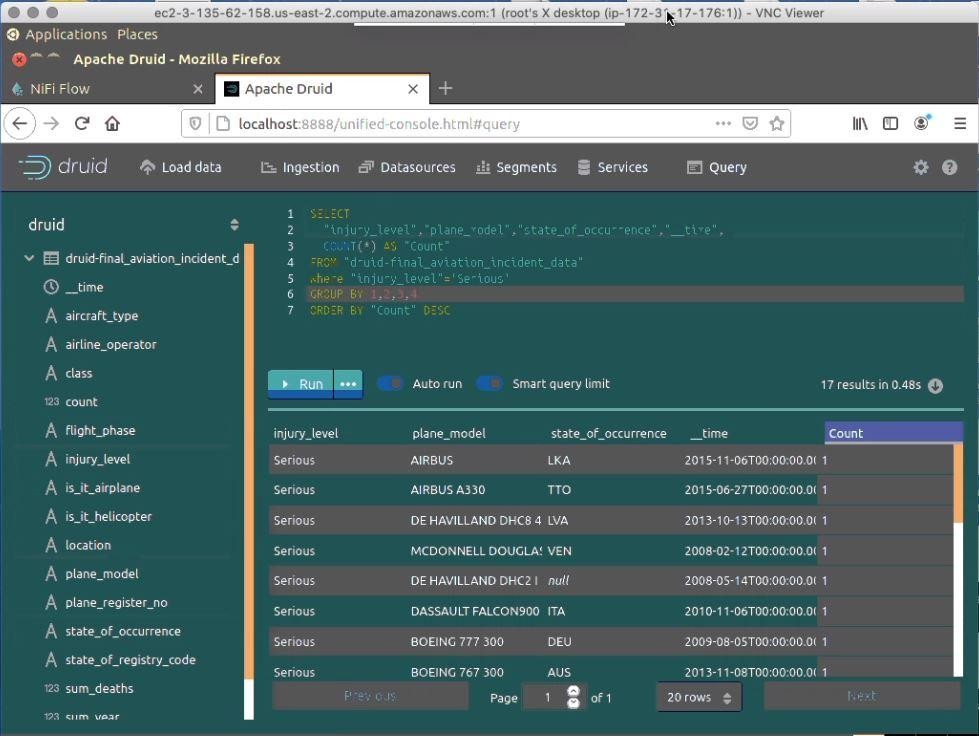
3.

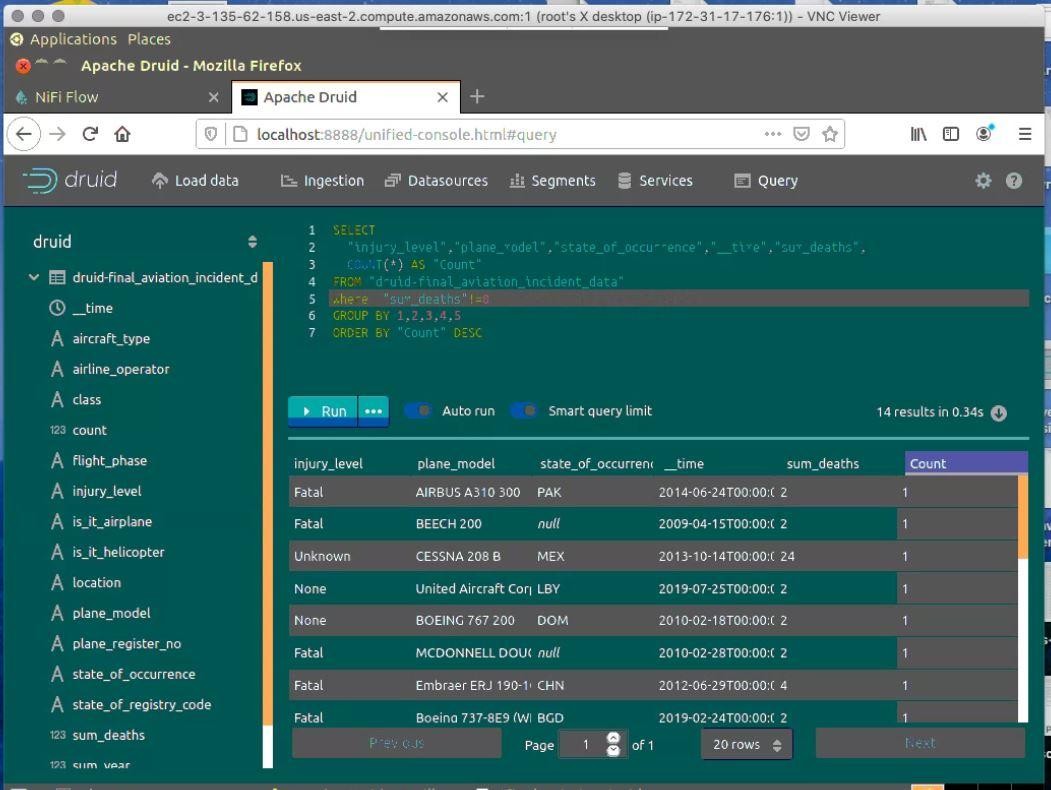






pache Dru! d





# Data Querying from HIVE

Login into Hive console (beeline) by giving the following commands

root@ip-172-31-17-176:/home/ubuntu/apache-hive-2.1.0-bin# bin/beeline

ls: cannot access '/home/ubuntu/apache-hive-2.1.0-bin/lib/hive-jdbc-\*-standalone.jar': No such file or directory

Beeline version 2.1.0 by Apache Hive beeline> !connect jdbc:hive2:// Connecting to jdbc:hive2://

Enter username for jdbc:hive2://: admin Enter password for jdbc:hive2://: admin

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/home/ubuntu/apache-hive-2.1.0-bin/lib/log4j-slf4j-impl- 2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/home/ubuntu/hadoop-2.7.3/share/hadoop/common/lib/slf4j- log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: See <http://www.slf4j.org/codes.html#multiple_bindings>for an explanation. SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

20/08/02 18:50:57 [main]: WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

20/08/02 18:51:01 [main]: WARN session.SessionState: METASTORE\_FILTER\_HOOK will be ignored, since hive.security.authorization.manager is set to instance of HiveAuthorizerFactory.

Connected to: Apache Hive (version 2.1.0) Driver: Hive JDBC (version 2.1.0)

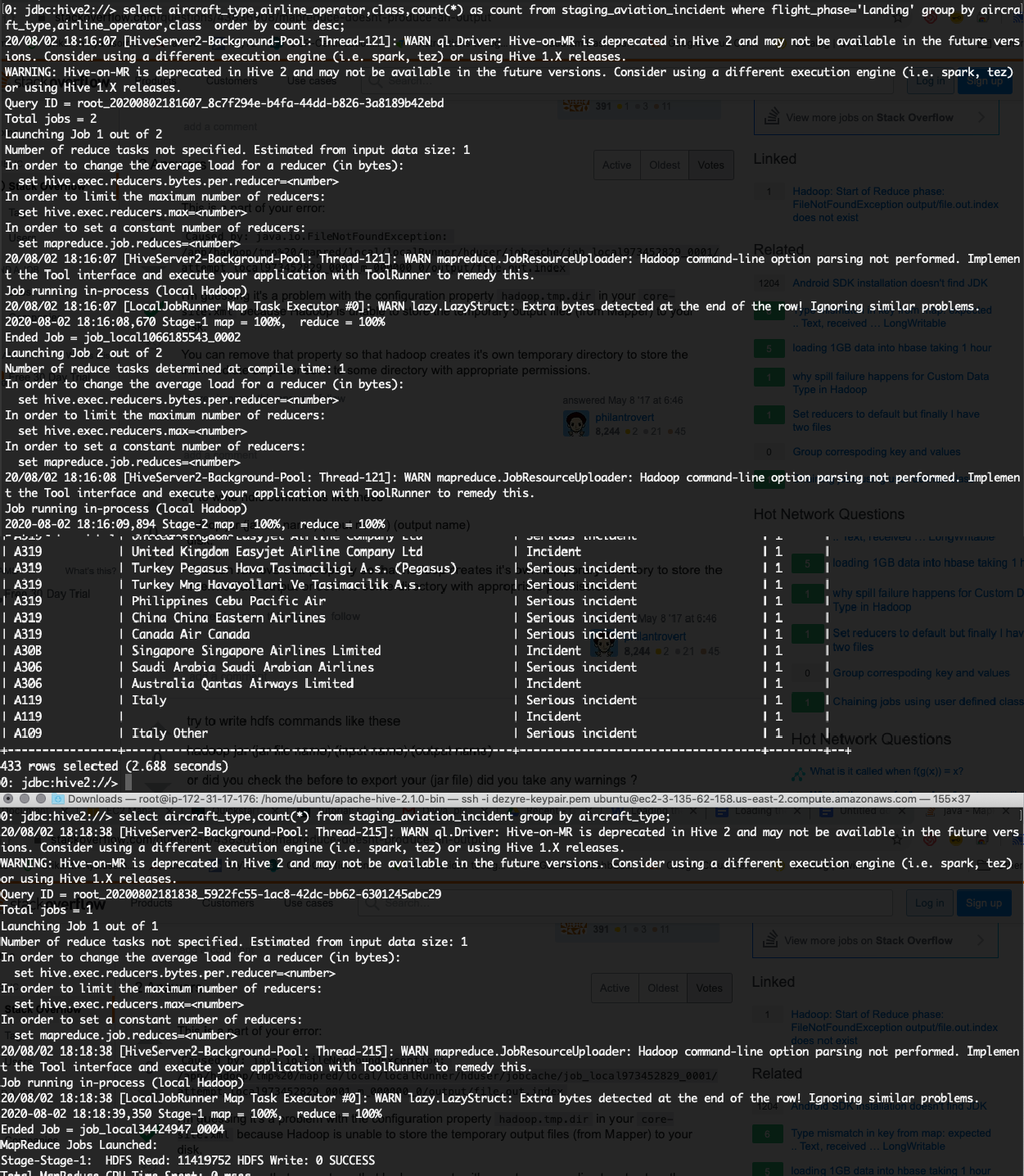
20/08/02 18:51:01 [main]: WARN jdbc.HiveConnection: Request to set autoCommit to false; Hive does not support autoCommit=false.

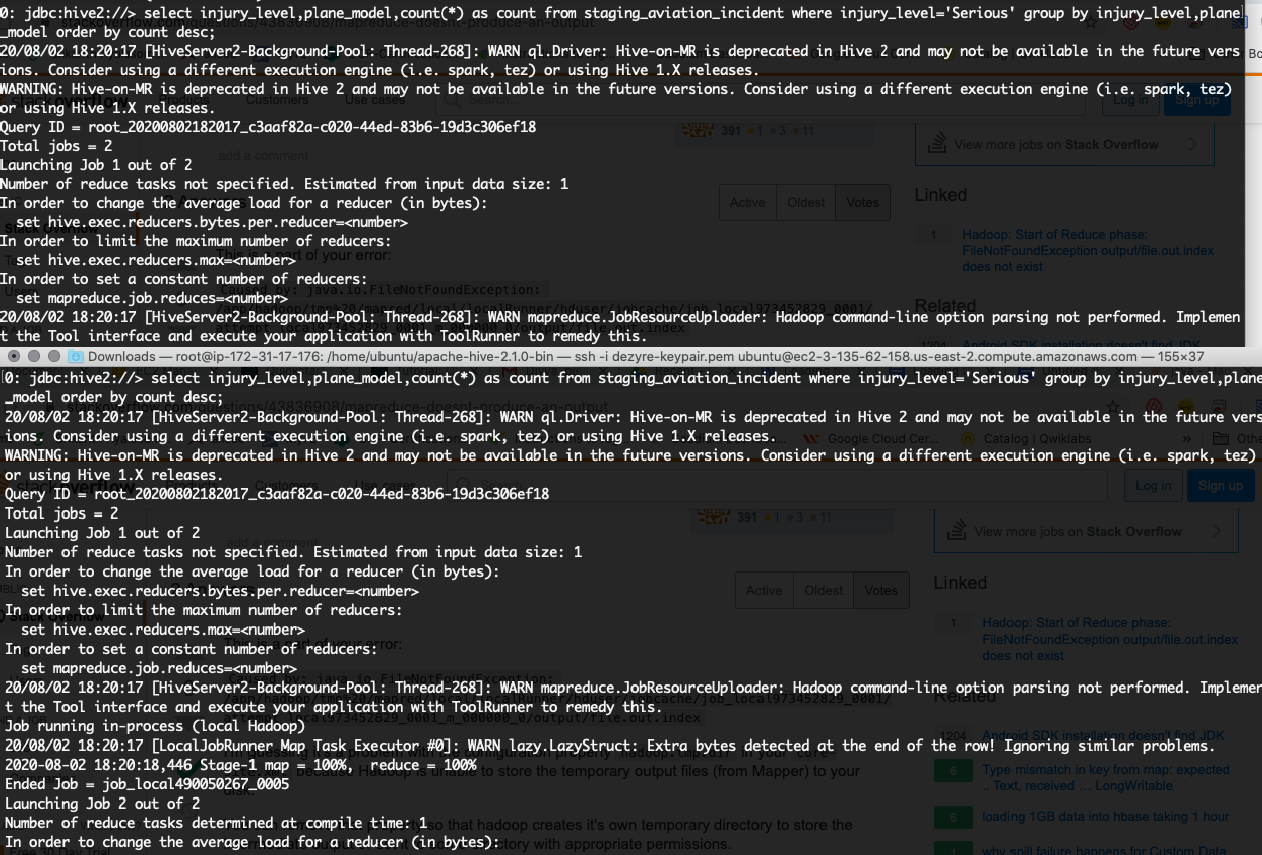
Transaction isolation: TRANSACTION\_REPEATABLE\_READ 0: jdbc:hive2://>

Then give the following commands

0: jdbc:hive2://> SHOW DATABASES; 0: jdbc:hive2://>USE dezyredb ;

Then run the below queries from the HIVE queries code document :





pTcric roleI

I cour1t I

13 rows selected {Z. 559 seconds)

B: j dbc: h fvc2 : //›

Jbc . NJ L vcz . X/ sc Lcct to Jute r r1c I bent occu rcJ Ju tc us dut c of cu reJ rt r I s LurJe re L sts r no is £r1 re rlus I rom sto r r1 uv r ut I art I rJ c r c cr1 - w-1ere - r -I

2B/B8/02 18:37:03 Ld0c0835^-0a\*f-401J-h019-bc9JlG8u4B29 mu r j: WSRN {cry.LaayStruct: Pxtrc bytcs Jctcctcd a hc cnJ of tñc row! Ignoring s s Our problczs.

3 zoos solcctcJ (B.1B4 scconJs)

B: dbc: ft I ver: ZZ> scler t to\_du t c( I r1r r der1t\_oc r u r cd\_du t e9 cs du te\_oc r u r ed , rt r r o(piuuc\_re9 r sts r\_rlow us pi ri\_r e9\_r1u i •roi st cp r riq\_uv I ut r or1\_ r rlc r den t wfie re II I 9fi t\_p Hose—' L end I r19 ' Grid I rid u ry\_1cvc1— ' Se r r ous ' 1 r i r t 18 :

DK

28ZB8 Z82 18 : 37: 83 bd Be88354-8uff -481d -b8l9 -be9d l68«4829 mo r rid : h'ARN 10 zy. L ozyS t met : fizi r 0 0 yt c s Act cc t ed oi th c end o• th c row! I$ r1or I u9 s mI I lo r o rob1c ms .

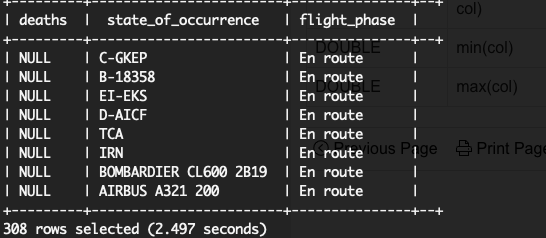
l datc\_occurcd t oln\_<cg\_num l l 2B08-l2-0E t EH-IQR l

l 2Bl5-11-0E t TC-JOA l l 2Bl1B8-12 I RP-C8E07 l

3 rows scI cc t c d {0. 1B4 sc cor ds3 B: dbc: I rvc2. //>

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| I | Serrou5 | BDEZN€ 737 80B BAS |  | 1 | 2 | I |
| l | ScrLou5 | AIRBUS A320 |  | l | 2 | l |
| l | Serious | BDEZN€ 737 88B |  | 1 | 2 | I |
| l | ScrLou5 | DE HAVILLAND DHC2 | I | l | 2 | l |
| l | Serious | DE HAVILLAND DHC8 | 4BB | l | 1 | l |
| l | ScrtouS | DASSAULT FXLCON9BB |  | l | 1 | l |
| l | Scrtou5 | BDEZN€ 7G7 38B |  | 1 | 1 | I |
| l | ScrtouS | BDEZN€ 737 98B |  | 1 | 1 | I |
| l | Scrtou5 | AZRBUS A348 3BB |  | 1 | 1 | I |
| l | ScrtouS | AIRBUS A330 |  | l | 1 | l |
| l | Scrtou5 | AIRBUS |  | l | 1 | l |
| l | Scr t Out | Papua |  | l | 1 | l |
| l | Scrtou5 | Hcsscrl” |  | 1 | 1 | l |

|  |  |  |
| --- | --- | --- |
| l ZB08—IN—0E | t EH—YQH | l |
| l 2B15—11—0E | I TC—JOE | l |
| l 2B1l-B8-12 | t R°-C8E07 | l |



# Data Querying from MySQL

Login to MYSQL using the following commands

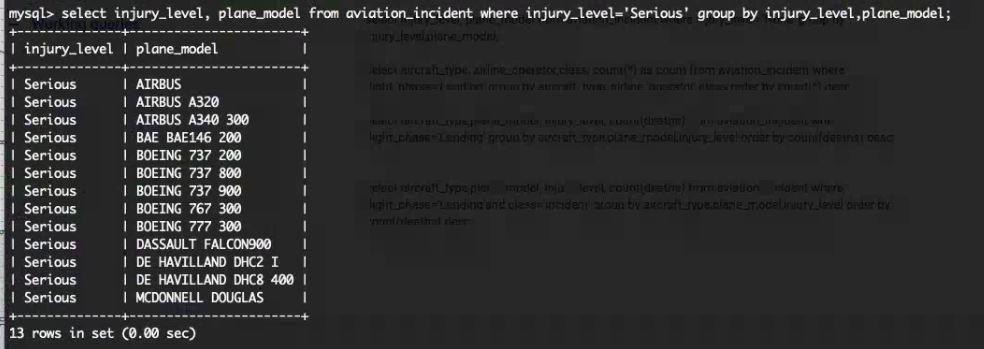
root@ip-172-31-17-176:/home/ubuntu# sudo mysql -u root -p Enter Password : root

mysql>

mysql> SHOW DATABASES;

mysql> USE dezyredb;

Then run the below queries according to MYSQL run book code :



GL 51



Z 5



M 06



AlRBdS

156 rcms n set .01 se



SERI ES

EXI STS



C484



fi319

EMBR/\ER



BE33



EON

%onc



1A10

BE9L



rysgl: seier• or •c•o\*i\_iype, nlone\_model, i<,u•y\_level, count(deo ns) from ovio ion\_ir.‹ dens where rli9 •.pncse- ' L«ndin9'°"d c:oss ’ i nriden•’ g-otp by oi<rrof=\_\*ype



BD70B



SERIES

EXTSTS

1 booe

|  |  |  |
| --- | --- | --- |
| BAB.1RDIER | | |
| B E fi C1 i | 9B |  |
| CSNADAIR | |

CL60



I AI



B3W

BDEIñG

737



PIPER



done



unKnoY•”ñ



FR



N

one

nknmn

EMBRAE R

|  |
| --- |
| C.337 |
|  |
|  |
| CRJX |

|  |  |  |  |
| --- | --- | --- | --- |
|  | |  | |
|  |  | | |
| JUL FSTRE@ | | |  |
| BDMBARDIEE | | |



FV8RAJ- R



Hi n or



DE

CVI L LAID



rDKKEft

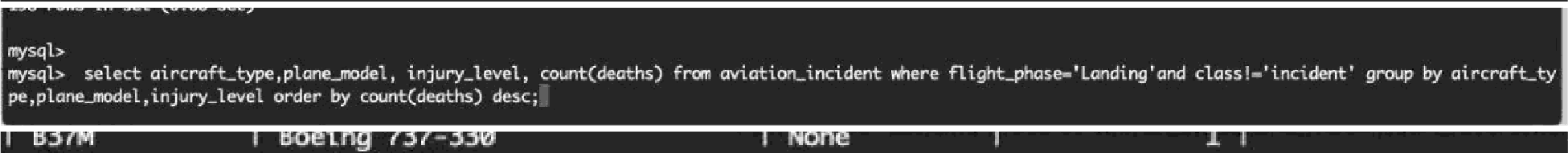
nz



|  |
| --- |
|  |
|  |
|  |
|  |

BAE

AVRDL46RJ



90 res

AIRBUS

B £EC!

BE9L



CSZS

CESSNA

CESSMA

5Z5



An -?4R'/

Nor-e



MD

HE L ICOPTE



EL60

C/t D/tlR



tLf4

C•UL f”STRCV

Nor:c



NCDOONE I

DOJCLAS



HCDOMNE LL

DDJ6LAS



30

None



BE10

BEE€i



ROCWELL





A386

A330



KA\‘ANAfiH



RL7R



1. **Visualisation of data in Quicksight AWS**

To provide access to Quicksight user from MYSQl follow the steps below Login into MYSQL and create Quciksight user and grant required permissions

mysql> use mysql;

Reading table information for completion of table and column names You can turn off this feature to get a quicker startup with -A

Database changed mysql> flush privileges;

Query OK, 0 rows affected (0.00 sec)

mysql> GRANT ALL PRIVILEGES ON \*.\* TO 'root'@'3.129.44.2' IDENTIFIED BY 'root';

Query OK, 0 rows affected, 1 warning (0.00 sec)

mysql> FLUSH PRIVILEGES;

Query OK, 0 rows affected (0.00 sec)

mysql>

mysql> CREATE USER 'x'@'ec2-52-15-247-168.us-east-2.compute.amazonaws.com' IDENTIFIED BY 'root'; Query OK, 0 rows affected (0.00 sec)

mysql>

mysql> GRANT ALL PRIVILEGES ON \*.\* TO 'x'@'ec2-52-15-247-168.us-east-2.compute.amazonaws.com' WITH GRANT OPTION;

Query OK, 0 rows affected (0.00 sec)

mysql> CREATE USER 'x'@'%' IDENTIFIED BY 'root';

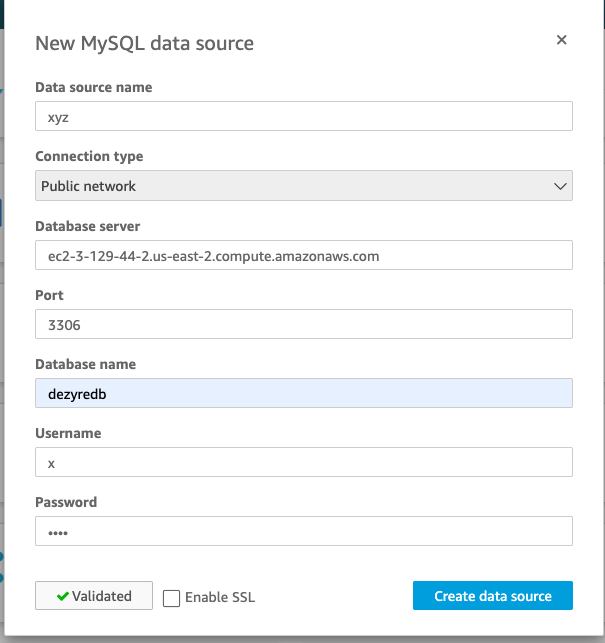
Query OK, 0 rows affected (0.00 sec)

mysql> GRANT ALL PRIVILEGES ON \*.\* TO 'x'@'%' WITH GRANT OPTION;

Query OK, 0 rows affected (0.00 sec)

mysql>

Then go to quicksight console and Create new dataset and select MYSQL and configure required parameters as follows :



Once the connection is validated, click on Create datasource Then it will show the list of databases available in MYSQL Choose the aviation\_incident database and click on continue .

Then it shows the data as follows from which we can create required visualisations

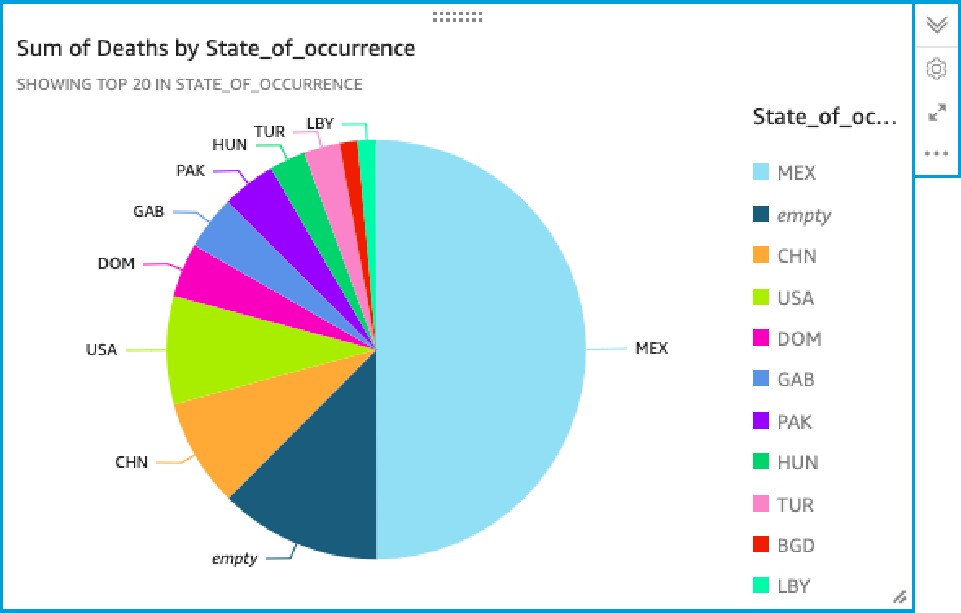


Add undo

+

Data set

dual ice aviation incident

@ Fields list



@ d eaths

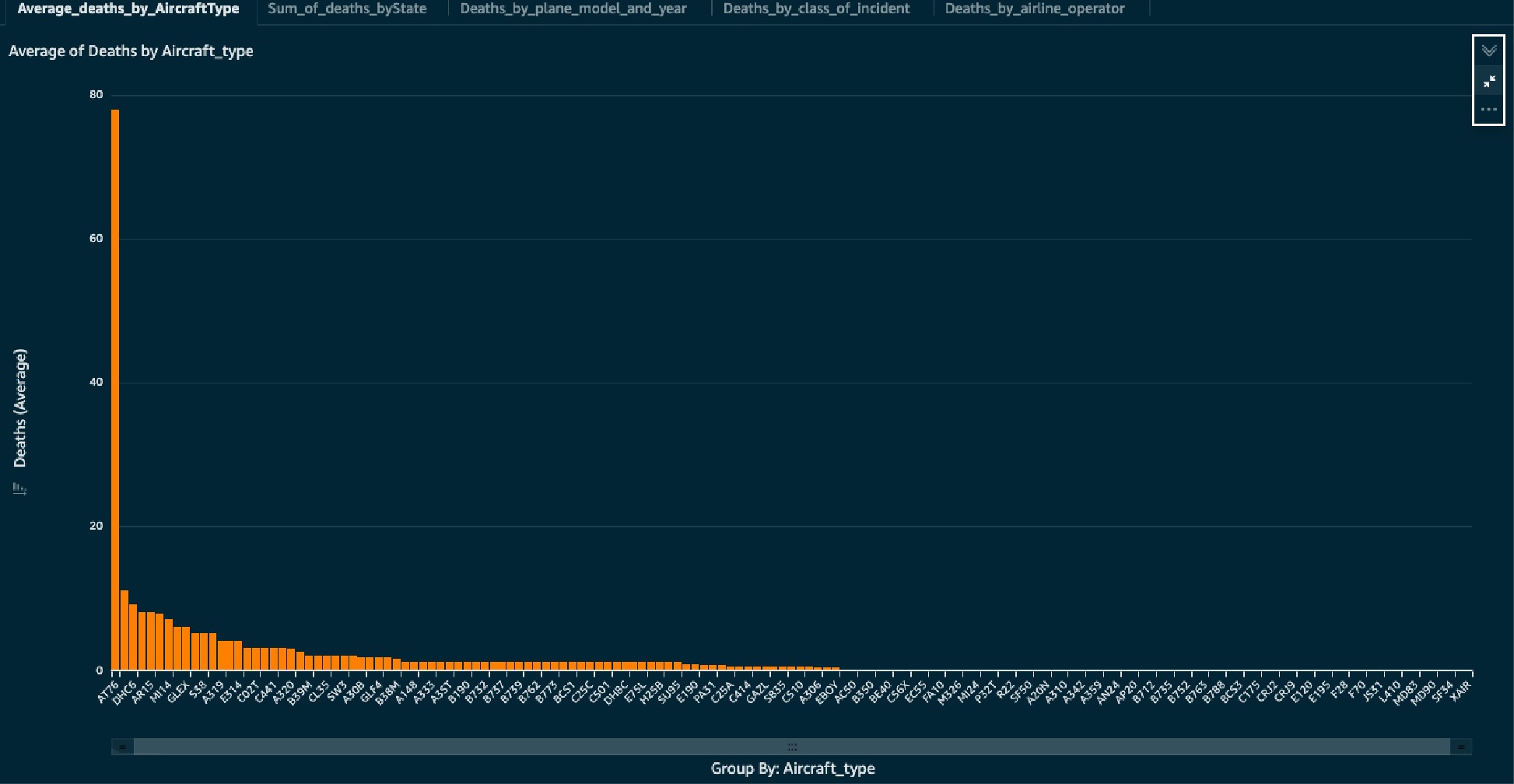
Visual types

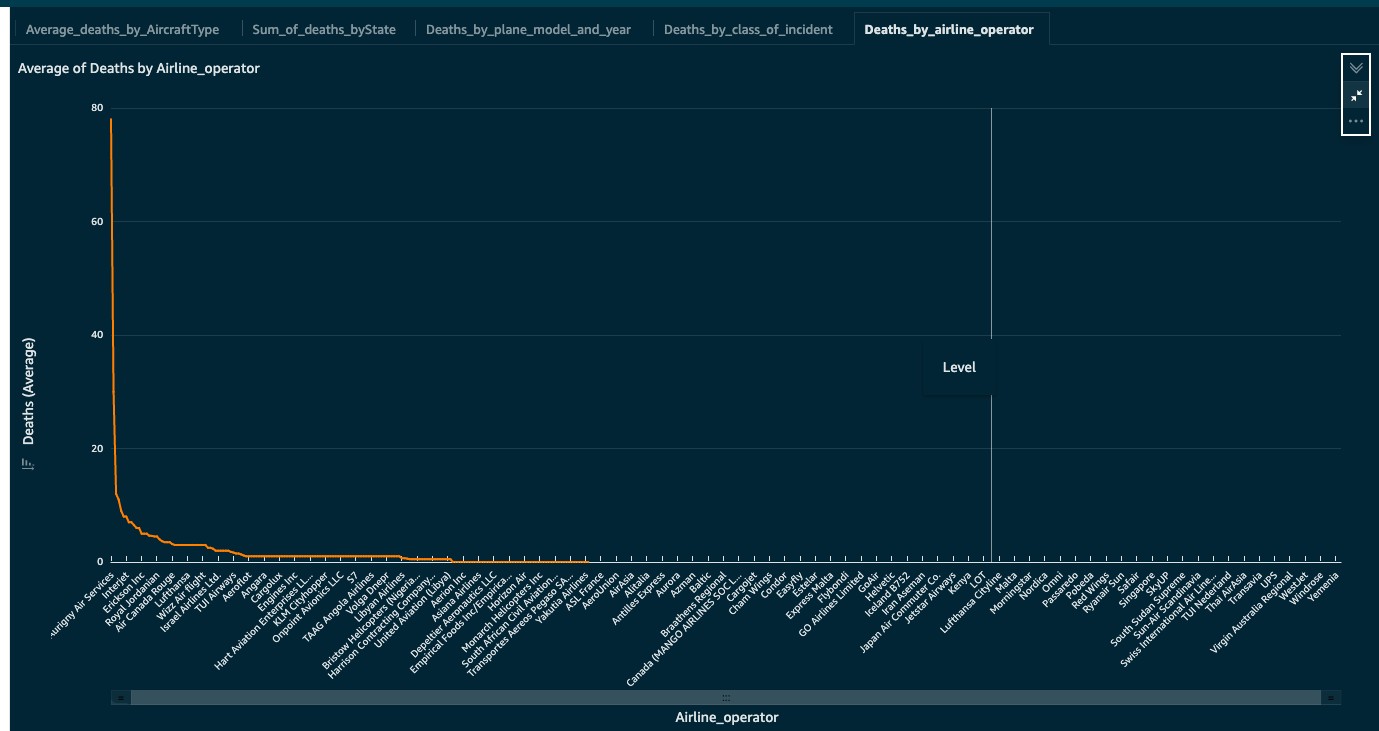
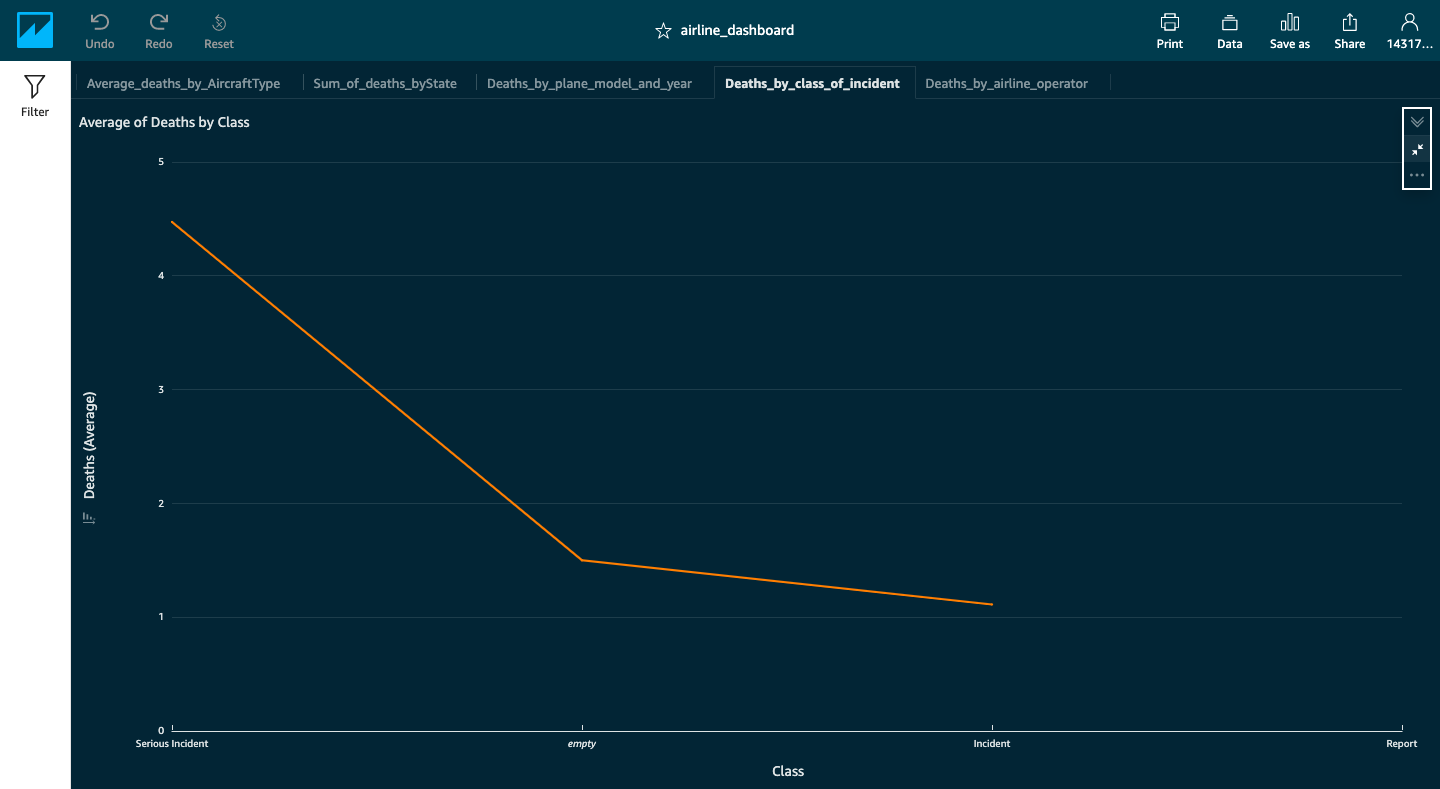
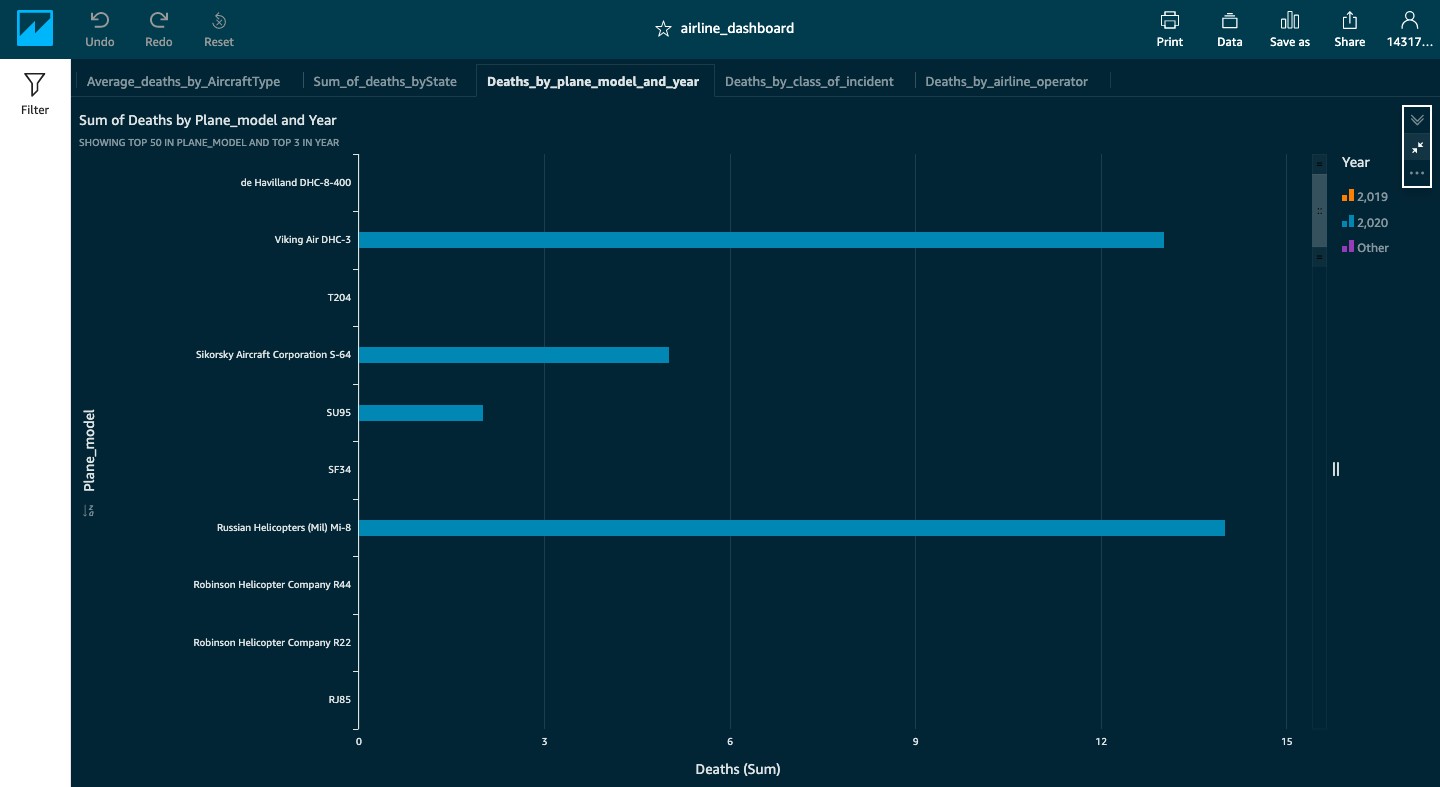
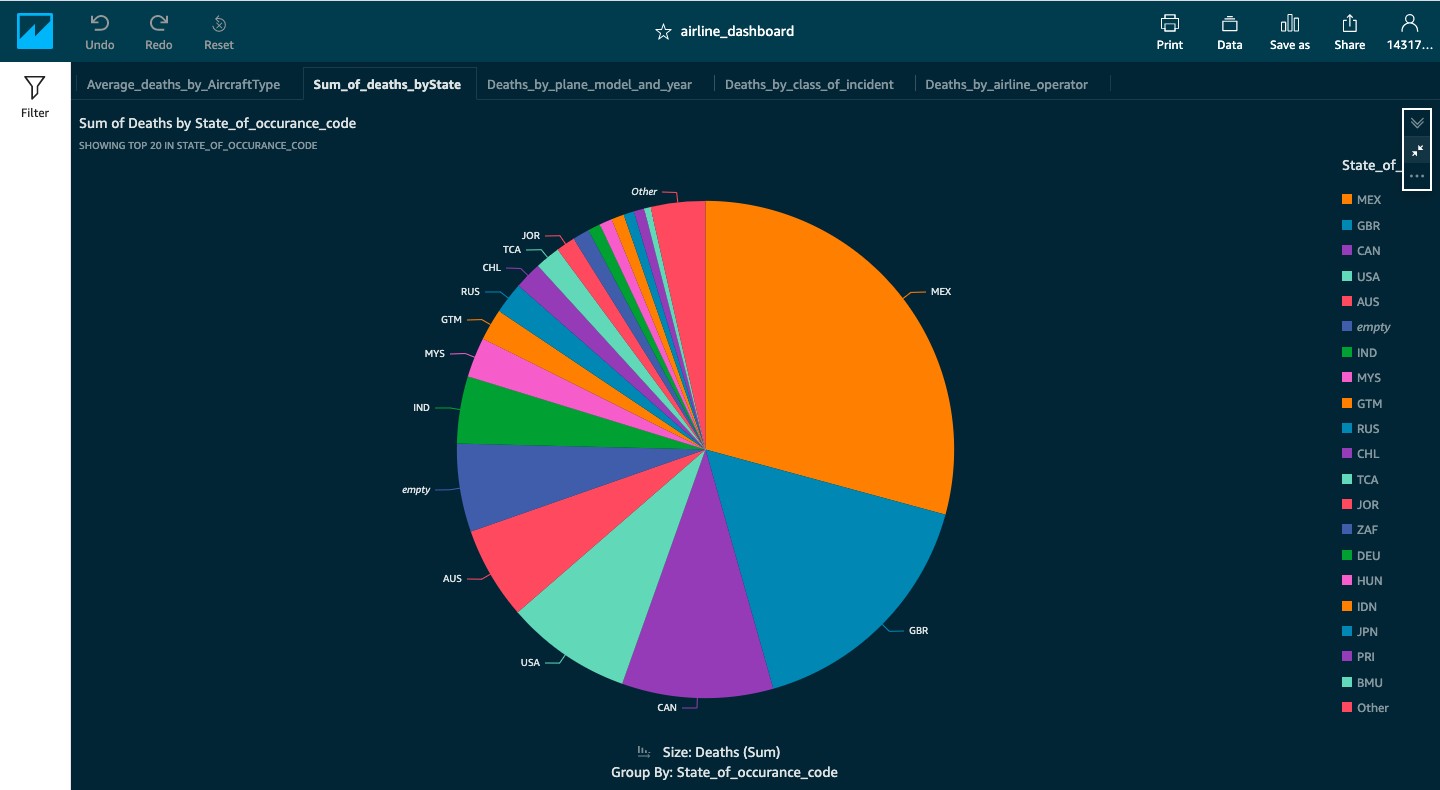
@

›ettl ngs

••• rind mets Grev#/coter 0 ut \_ i\_o«ur «e \*aiue P a‹aihs lsuml







**Extended project :** We compare the performance of Hive/Druid/MySQL with variable workloads in project recordings.

Discuss on project extendibility using RDS and Sqoop