

# Project Execution Documentation

## 1. 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>
- 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>

Configure Processor

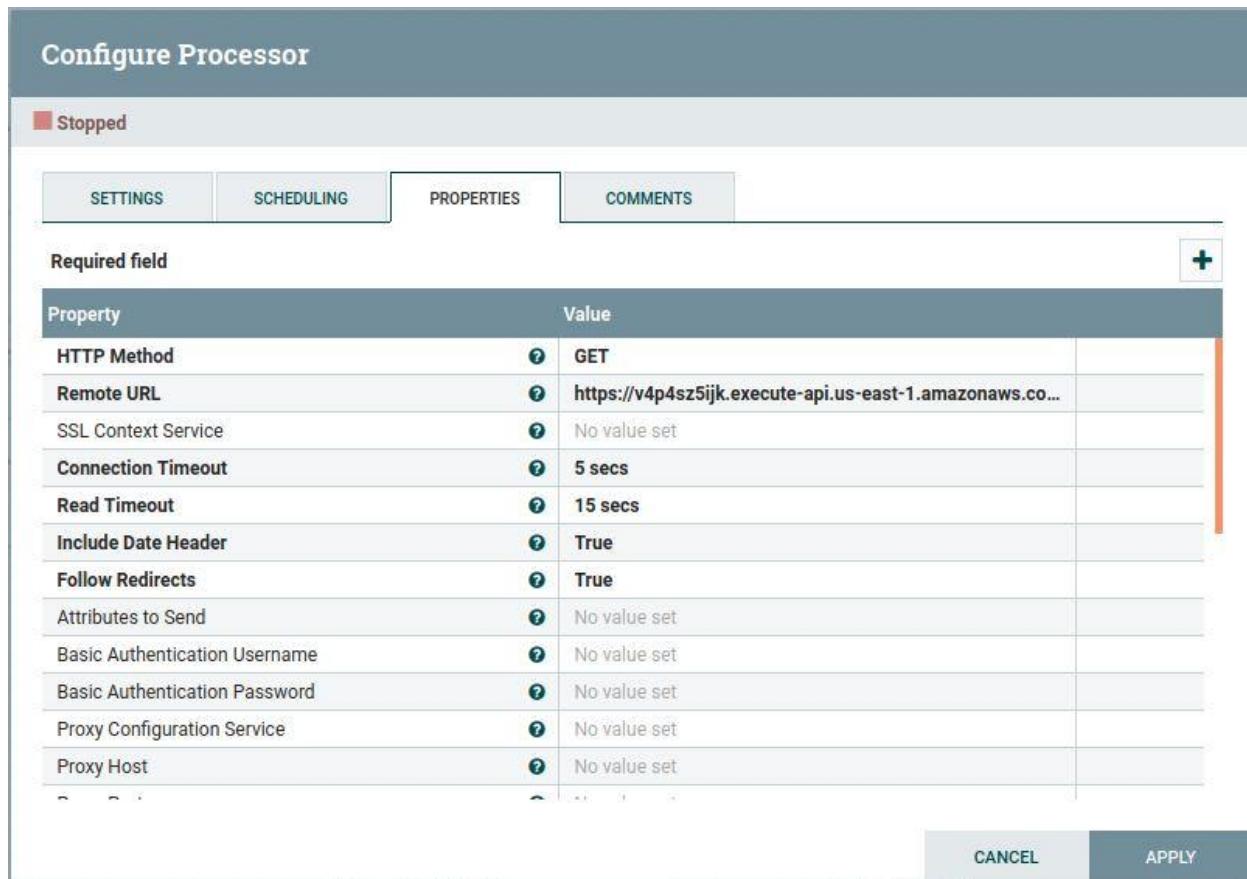
Stopped

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field

Property	Value
HTTP Method	GET
Remote URL	<a href="https://v4p4sz5ijk.execute-api.us-east-1.amazonaws.com/">https://v4p4sz5ijk.execute-api.us-east-1.amazonaws.co...</a>
SSL Context Service	No value set
Connection Timeout	5 secs
Read Timeout	15 secs
Include Date Header	True
Follow Redirects	True
Attributes to Send	No value set
Basic Authentication Username	No value set
Basic Authentication Password	No value set
Proxy Configuration Service	No value set
Proxy Host	No value set

CANCEL APPLY



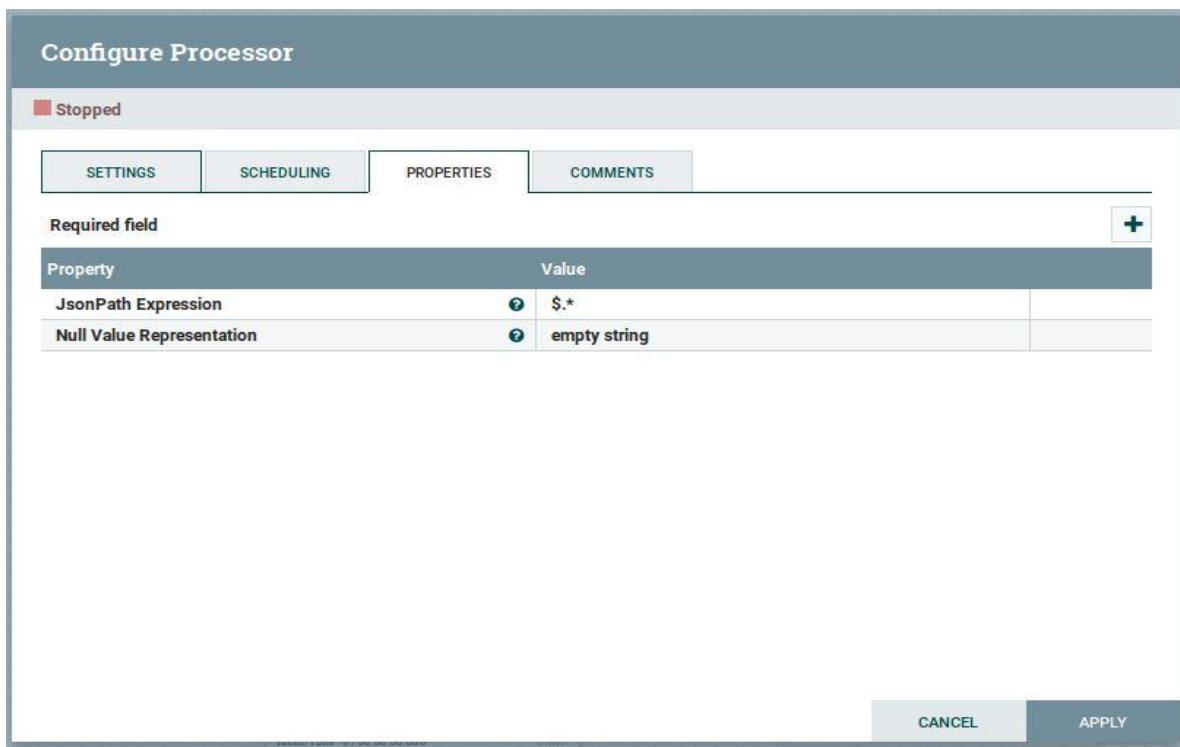
We will get response data as below .



A screenshot of a JSON viewer interface. At the top, there is a "View as:" dropdown set to "original". Below the dropdown, the JSON data is displayed in a code editor-like format with line numbers from 1 to 29 on the left. The JSON structure is an array of objects, with the first object containing various flight and incident details.

```
1 [  
2 {  
3   "Date": "2008-01-02T00:00:00.000Z",  
4   "StateOfOccurrence": "USA",  
5   "Location": "Orlando Sanford, Florida",  
6   "Model": "BOEING 757 200",  
7   "Registration": "G-CEJM",  
8   "Operator": "",  
9   "StateOfRegistry": "GBR",  
10  "FlightPhase": "En route",  
11  "Class": "Incident",  
12  "Fatalities": 0,  
13  "Over2250": true,  
14  "Over5700": true,  
15  "ScheduledCommercial": true,  
16  "InjuryLevel": "None",  
17  "TypeDesignator": "ST75",  
18  "Helicopter": false,  
19  "Airplane": true,  
20  "Engines": 1,  
21  "EngineType": "Piston",  
22  "StateOfOperator": "",  
23  "Official": "",  
24  "Risk": "SCF",  
25  "OccCats": [  
26    "SCF-PP",  
27  ],  
28  "Year": 2008  
},  
29 ]
```

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



A screenshot of the "Configure Processor" dialog for a SplitJson Processor. The title bar says "Configure Processor". Below it, a status bar shows "Stopped". The dialog has four tabs: SETTINGS (selected), SCHEDULING, PROPERTIES, and COMMENTS. The "SETTINGS" tab contains a "Required field" section with a table:

Property	Value
JsonPath Expression	\$.*
Null Value Representation	empty string

At the bottom right are "CANCEL" and "APPLY" buttons.

- 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

**Configure Processor**

■ Stopped

SETTINGS	SCHEDULING	PROPERTIES	COMMENTS																										
<b>Required field</b> <table border="1"> <thead> <tr> <th>Property</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Destination</td> <td>flowfile-attribute</td> </tr> <tr> <td>Return Type</td> <td>auto-detect</td> </tr> <tr> <td>Path Not Found Behavior</td> <td>ignore</td> </tr> <tr> <td>Null Value Representation</td> <td>empty string</td> </tr> <tr> <td>aircraft_type</td> <td>\$.TypeDesignator</td> </tr> <tr> <td>airline_operator</td> <td>\$.Operator</td> </tr> <tr> <td>class</td> <td>\$.Class</td> </tr> <tr> <td>deaths</td> <td>\$.Fatalities</td> </tr> <tr> <td>flight_phase</td> <td>\$.FlightPhase</td> </tr> <tr> <td>incident_occurred_date</td> <td>\$.Date</td> </tr> <tr> <td>injury_level</td> <td>\$.InjuryLevel</td> </tr> <tr> <td>is_it_airplane</td> <td>\$.Airplane</td> </tr> </tbody> </table>				Property	Value	Destination	flowfile-attribute	Return Type	auto-detect	Path Not Found Behavior	ignore	Null Value Representation	empty string	aircraft_type	\$.TypeDesignator	airline_operator	\$.Operator	class	\$.Class	deaths	\$.Fatalities	flight_phase	\$.FlightPhase	incident_occurred_date	\$.Date	injury_level	\$.InjuryLevel	is_it_airplane	\$.Airplane
Property	Value																												
Destination	flowfile-attribute																												
Return Type	auto-detect																												
Path Not Found Behavior	ignore																												
Null Value Representation	empty string																												
aircraft_type	\$.TypeDesignator																												
airline_operator	\$.Operator																												
class	\$.Class																												
deaths	\$.Fatalities																												
flight_phase	\$.FlightPhase																												
incident_occurred_date	\$.Date																												
injury_level	\$.InjuryLevel																												
is_it_airplane	\$.Airplane																												
<span style="border: 1px solid #ccc; padding: 2px;">+</span> <span>CANCEL</span> <span>APPLY</span>																													

**Configure Processor**

■ Stopped

SETTINGS	SCHEDULING	PROPERTIES	COMMENTS																										
<b>Required field</b> <table border="1"> <thead> <tr> <th>Property</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>flight_phase</td> <td>\$.FlightPhase</td> </tr> <tr> <td>incident_occurred_date</td> <td>\$.Date</td> </tr> <tr> <td>injury_level</td> <td>\$.InjuryLevel</td> </tr> <tr> <td>is_it_airplane</td> <td>\$.Airplane</td> </tr> <tr> <td>is_it_helicopter</td> <td>\$.Helicopter</td> </tr> <tr> <td>list_of_occurrence_category</td> <td>\$.OccCats[0]</td> </tr> <tr> <td>location</td> <td>\$.Location</td> </tr> <tr> <td>plane_model</td> <td>\$.Model</td> </tr> <tr> <td>plane_register_no</td> <td>\$.Registration</td> </tr> <tr> <td>state_of_occurrence</td> <td>\$.StateOfOccurrence</td> </tr> <tr> <td>state_of_registry_code</td> <td>\$.StateOfRegistry</td> </tr> <tr> <td>year</td> <td>\$.Year</td> </tr> </tbody> </table>				Property	Value	flight_phase	\$.FlightPhase	incident_occurred_date	\$.Date	injury_level	\$.InjuryLevel	is_it_airplane	\$.Airplane	is_it_helicopter	\$.Helicopter	list_of_occurrence_category	\$.OccCats[0]	location	\$.Location	plane_model	\$.Model	plane_register_no	\$.Registration	state_of_occurrence	\$.StateOfOccurrence	state_of_registry_code	\$.StateOfRegistry	year	\$.Year
Property	Value																												
flight_phase	\$.FlightPhase																												
incident_occurred_date	\$.Date																												
injury_level	\$.InjuryLevel																												
is_it_airplane	\$.Airplane																												
is_it_helicopter	\$.Helicopter																												
list_of_occurrence_category	\$.OccCats[0]																												
location	\$.Location																												
plane_model	\$.Model																												
plane_register_no	\$.Registration																												
state_of_occurrence	\$.StateOfOccurrence																												
state_of_registry_code	\$.StateOfRegistry																												
year	\$.Year																												
<span style="border: 1px solid #ccc; padding: 2px;">+</span> <span>CANCEL</span> <span>APPLY</span>																													

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

**Configure Processor**

Stopped

SETTINGS	SCHEDULING	PROPERTIES	COMMENTS																
Required field																			
<table border="1"> <thead> <tr> <th>Property</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Search Value</td> <td>(?s)(^.*\$)</td> </tr> <tr> <td>Replacement Value</td> <td>\$(accident_occurred_date:escapeCsv()),\$/aircraft_type:...)</td> </tr> <tr> <td>Character Set</td> <td>UTF-8</td> </tr> <tr> <td>Maximum Buffer Size</td> <td>1 MB</td> </tr> <tr> <td>Replacement Strategy</td> <td>Regex Replace</td> </tr> <tr> <td>Evaluation Mode</td> <td>Entire text</td> </tr> <tr> <td>Line-by-Line Evaluation Mode</td> <td>All</td> </tr> </tbody> </table>		Property	Value	Search Value	(?s)(^.*\$)	Replacement Value	\$(accident_occurred_date:escapeCsv()),\$/aircraft_type:...)	Character Set	UTF-8	Maximum Buffer Size	1 MB	Replacement Strategy	Regex Replace	Evaluation Mode	Entire text	Line-by-Line Evaluation Mode	All	<a href="#">+</a>	
Property	Value																		
Search Value	(?s)(^.*\$)																		
Replacement Value	\$(accident_occurred_date:escapeCsv()),\$/aircraft_type:...)																		
Character Set	UTF-8																		
Maximum Buffer Size	1 MB																		
Replacement Strategy	Regex Replace																		
Evaluation Mode	Entire text																		
Line-by-Line Evaluation Mode	All																		
			CANCEL	APPLY															

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

## Configure Processor

Stopped

SETTINGS	SCHEDULING	PROPERTIES	COMMENTS																						
Required field																									
<table border="1"> <thead> <tr> <th>Property</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Delete Attributes Expression</td> <td>?</td> <td>No value set</td> <td></td> </tr> <tr> <td><b>Store State</b></td> <td>?</td> <td><b>Do not store state</b></td> <td></td> </tr> <tr> <td>Stateful Variables Initial Value</td> <td>?</td> <td>No value set</td> <td></td> </tr> <tr> <td><b>Cache Value Lookup Cache Size</b></td> <td>?</td> <td><b>100</b></td> <td></td> </tr> <tr> <td>staging_model</td> <td>?</td> <td>aviation_incident</td> <td></td> </tr> </tbody> </table>				Property	Value	Delete Attributes Expression	?	No value set		<b>Store State</b>	?	<b>Do not store state</b>		Stateful Variables Initial Value	?	No value set		<b>Cache Value Lookup Cache Size</b>	?	<b>100</b>		staging_model	?	aviation_incident	
Property	Value																								
Delete Attributes Expression	?	No value set																							
<b>Store State</b>	?	<b>Do not store state</b>																							
Stateful Variables Initial Value	?	No value set																							
<b>Cache Value Lookup Cache Size</b>	?	<b>100</b>																							
staging_model	?	aviation_incident																							

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

SETTINGS	SCHEDULING	PROPERTIES	COMMENTS																																																		
Required field																																																					
<table border="1"> <thead> <tr> <th>Property</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Kerberos Credentials Service</td> <td>?</td> <td>No value set</td> <td></td> </tr> <tr> <td>Kerberos Principal</td> <td>?</td> <td>No value set</td> <td></td> </tr> <tr> <td>Kerberos Keytab</td> <td>?</td> <td>No value set</td> <td></td> </tr> <tr> <td>Kerberos Relogin Period</td> <td>?</td> <td>4 hours</td> <td></td> </tr> <tr> <td>Additional Classpath Resources</td> <td>?</td> <td>No value set</td> <td></td> </tr> <tr> <td><b>Directory</b></td> <td>?</td> <td><b>staging_\${staging_model}</b></td> <td></td> </tr> <tr> <td><b>Conflict Resolution Strategy</b></td> <td>?</td> <td><b>append</b></td> <td></td> </tr> <tr> <td>Block Size</td> <td>?</td> <td>No value set</td> <td></td> </tr> <tr> <td>IO Buffer Size</td> <td>?</td> <td>No value set</td> <td></td> </tr> <tr> <td>Replication</td> <td>?</td> <td>No value set</td> <td></td> </tr> <tr> <td>Permissions umask</td> <td>?</td> <td>No value set</td> <td></td> </tr> <tr> <td>Remote Owner</td> <td>?</td> <td>No value set</td> <td></td> </tr> </tbody> </table>				Property	Value	Kerberos Credentials Service	?	No value set		Kerberos Principal	?	No value set		Kerberos Keytab	?	No value set		Kerberos Relogin Period	?	4 hours		Additional Classpath Resources	?	No value set		<b>Directory</b>	?	<b>staging_\${staging_model}</b>		<b>Conflict Resolution Strategy</b>	?	<b>append</b>		Block Size	?	No value set		IO Buffer Size	?	No value set		Replication	?	No value set		Permissions umask	?	No value set		Remote Owner	?	No value set	
Property	Value																																																				
Kerberos Credentials Service	?	No value set																																																			
Kerberos Principal	?	No value set																																																			
Kerberos Keytab	?	No value set																																																			
Kerberos Relogin Period	?	4 hours																																																			
Additional Classpath Resources	?	No value set																																																			
<b>Directory</b>	?	<b>staging_\${staging_model}</b>																																																			
<b>Conflict Resolution Strategy</b>	?	<b>append</b>																																																			
Block Size	?	No value set																																																			
IO Buffer Size	?	No value set																																																			
Replication	?	No value set																																																			
Permissions umask	?	No value set																																																			
Remote Owner	?	No value set																																																			

- 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.

**Configure Processor**

Stopped

SETTINGS	SCHEDULING	PROPERTIES	COMMENTS												
<b>Required field</b>															
<table border="1"> <thead> <tr> <th>Property</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Attributes List</td> <td>accident_occurred_date,aircraft_type,airline_operator,class...</td> </tr> <tr> <td>Attributes Regular Expression</td> <td>No value set</td> </tr> <tr> <td>Destination</td> <td>flowfile-content</td> </tr> <tr> <td>Include Core Attributes</td> <td>true</td> </tr> <tr> <td>Null Value</td> <td>false</td> </tr> </tbody> </table>				Property	Value	Attributes List	accident_occurred_date,aircraft_type,airline_operator,class...	Attributes Regular Expression	No value set	Destination	flowfile-content	Include Core Attributes	true	Null Value	false
Property	Value														
Attributes List	accident_occurred_date,aircraft_type,airline_operator,class...														
Attributes Regular Expression	No value set														
Destination	flowfile-content														
Include Core Attributes	true														
Null Value	false														

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

**Configure Processor**

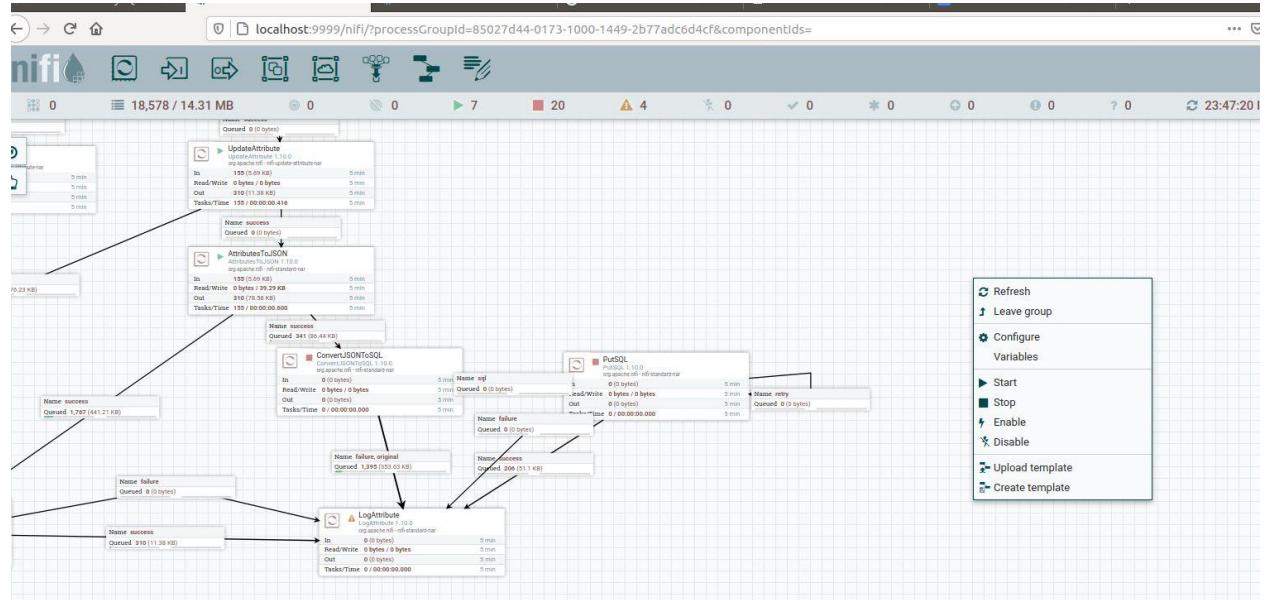
Stopped

SETTINGS	SCHEDULING	PROPERTIES	COMMENTS																										
<b>Required field</b>																													
<table border="1"> <thead> <tr> <th>Property</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>JDBC Connection Pool</td> <td>mysqlDBCPConnectionPool</td> </tr> <tr> <td>Statement Type</td> <td>INSERT</td> </tr> <tr> <td>Table Name</td> <td>aviation_incident</td> </tr> <tr> <td>Catalog Name</td> <td>No value set</td> </tr> <tr> <td>Schema Name</td> <td>No value set</td> </tr> <tr> <td>Translate Field Names</td> <td>true</td> </tr> <tr> <td>Unmatched Field Behavior</td> <td>Ignore Unmatched Fields</td> </tr> <tr> <td>Unmatched Column Behavior</td> <td>Fail on Unmatched Columns</td> </tr> <tr> <td>Update Keys</td> <td>No value set</td> </tr> <tr> <td>Quote Column Identifiers</td> <td>false</td> </tr> <tr> <td>Quote Table Identifiers</td> <td>false</td> </tr> <tr> <td>SQL Parameter Attribute Prefix</td> <td>sql</td> </tr> </tbody> </table>				Property	Value	JDBC Connection Pool	mysqlDBCPConnectionPool	Statement Type	INSERT	Table Name	aviation_incident	Catalog Name	No value set	Schema Name	No value set	Translate Field Names	true	Unmatched Field Behavior	Ignore Unmatched Fields	Unmatched Column Behavior	Fail on Unmatched Columns	Update Keys	No value set	Quote Column Identifiers	false	Quote Table Identifiers	false	SQL Parameter Attribute Prefix	sql
Property	Value																												
JDBC Connection Pool	mysqlDBCPConnectionPool																												
Statement Type	INSERT																												
Table Name	aviation_incident																												
Catalog Name	No value set																												
Schema Name	No value set																												
Translate Field Names	true																												
Unmatched Field Behavior	Ignore Unmatched Fields																												
Unmatched Column Behavior	Fail on Unmatched Columns																												
Update Keys	No value set																												
Quote Column Identifiers	false																												
Quote Table Identifiers	false																												
SQL Parameter Attribute Prefix	sql																												
		<input type="button" value="CANCEL"/> <input type="button" value="APPLY"/>																											

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

## 2. Create JDBCConnectionPoolService in nifi ( for mysql)

- Right click on the nifi Canvas select the “configure”



- Click on the “+“ symbol at the right corner.

Name	Type	Bundle	State	Scope
mysqlDBCPConnectionPool	DBCPConnectionPool 1.10.0	org.apache.nifi - nifi-dbc-service-nar	Enabled	Airline data

- Select the DBconnectionPool service then click on add

Add Controller Service

Source	Type	Version	Tags
all groups			db
<b>avro</b>	CassandraSessionProvider	1.10.0	database, pooling, cassandra, d...
<b>cache</b>	DBCPConnectionPool	1.10.0	database, pooling, dbcp, jdbc, c...
<b>cluster</b>	DBCPConnectionPoolLookup	1.10.0	database, pooling, dbcp, jdbc, c...
<b>connection</b>	DatabaseRecordLookupServ...	1.10.0	lookup, cache, database, rdbms...
<b>credentials</b>	DatabaseRecordSink	1.10.0	database, record, jdbc, connecti...
<b>csv</b>	HiveConnectionPool	1.10.0	hive, database, pooling, dbcp, jd...
<b>database</b>	MongoDBControllerService	1.10.0	mongo, service, mongodb
<b>distributed</b>	MongoDBLookupService	1.10.0	mongo, lookup, record, mongodb
<b>enrich</b>	SimpleDatabaseLookupServ...	1.10.0	lookup, cache, database, rdbms...
<b>hbbase</b>	SiteToSiteReportingRecordS...	1.10.0	site, s2s, record, db
<b>join</b>			
<b>json</b>			
<b>key</b>			
<b>lookup</b>			
<b>map</b>			
<b>parse</b>			
<b>reader</b>			
<b>record</b>			
<b>recordset</b>			
<b>reloadable</b>			
<b>restricted</b>			
<b>row</b>			
<b>set</b>			
<b>value</b>			
<b>writer</b>			

DBCPConnectionPool 1.10.0 org.apache.nifi - nifi-dbc-pool-service-nar  
Provides Database Connection Pooling Service. Connections can be asked from pool and returned after usage.

CANCEL ADD

- Configured as in the below image for the mysql Localhost

Configure Controller Service

SETTINGS	PROPERTIES	COMMENTS																														
Required field																																
<table border="1"> <thead> <tr> <th>Property</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Database Connection URL</td> <td>jdbc:mysql://localhost:3306/dezyredb</td> </tr> <tr> <td>Database Driver Class Name</td> <td>com.mysql.jdbc.Driver</td> </tr> <tr> <td>Database Driver Location(s)</td> <td>/usr/share/java/mysql-connector-java-8.0.21.jar</td> </tr> <tr> <td>Kerberos Credentials Service</td> <td>No value set</td> </tr> <tr> <td>Database User</td> <td>root</td> </tr> <tr> <td>Password</td> <td>Sensitive value set</td> </tr> <tr> <td>Max Wait Time</td> <td>500 millis</td> </tr> <tr> <td>Max Total Connections</td> <td>8</td> </tr> <tr> <td>Validation query</td> <td>select 1</td> </tr> <tr> <td>Minimum Idle Connections</td> <td>0</td> </tr> <tr> <td>Max Idle Connections</td> <td>8</td> </tr> <tr> <td>Max Connection Lifetime</td> <td>-1</td> </tr> <tr> <td>Time Between Eviction Runs</td> <td>-1</td> </tr> <tr> <td>Minimum Evictable Idle Time</td> <td>30 mins</td> </tr> </tbody> </table>			Property	Value	Database Connection URL	jdbc:mysql://localhost:3306/dezyredb	Database Driver Class Name	com.mysql.jdbc.Driver	Database Driver Location(s)	/usr/share/java/mysql-connector-java-8.0.21.jar	Kerberos Credentials Service	No value set	Database User	root	Password	Sensitive value set	Max Wait Time	500 millis	Max Total Connections	8	Validation query	select 1	Minimum Idle Connections	0	Max Idle Connections	8	Max Connection Lifetime	-1	Time Between Eviction Runs	-1	Minimum Evictable Idle Time	30 mins
Property	Value																															
Database Connection URL	jdbc:mysql://localhost:3306/dezyredb																															
Database Driver Class Name	com.mysql.jdbc.Driver																															
Database Driver Location(s)	/usr/share/java/mysql-connector-java-8.0.21.jar																															
Kerberos Credentials Service	No value set																															
Database User	root																															
Password	Sensitive value set																															
Max Wait Time	500 millis																															
Max Total Connections	8																															
Validation query	select 1																															
Minimum Idle Connections	0																															
Max Idle Connections	8																															
Max Connection Lifetime	-1																															
Time Between Eviction Runs	-1																															
Minimum Evictable Idle Time	30 mins																															
<p>CANCEL APPLY</p>																																

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

## Configure Processor

Stopped

SETTINGS	SCHEDULING	PROPERTIES	COMMENTS																		
Required field																					
<table border="1"> <thead> <tr> <th>Property</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>JDBC Connection Pool</td> <td>mysqlDBCPConnectionPool</td> </tr> <tr> <td>SQL Statement</td> <td>No value set</td> </tr> <tr> <td>Support Fragmented Transactions</td> <td>true</td> </tr> <tr> <td>Database Session AutoCommit</td> <td>false</td> </tr> <tr> <td>Transaction Timeout</td> <td>No value set</td> </tr> <tr> <td>Batch Size</td> <td>100</td> </tr> <tr> <td>Obtain Generated Keys</td> <td>false</td> </tr> <tr> <td>Rollback On Failure</td> <td>false</td> </tr> </tbody> </table>				Property	Value	JDBC Connection Pool	mysqlDBCPConnectionPool	SQL Statement	No value set	Support Fragmented Transactions	true	Database Session AutoCommit	false	Transaction Timeout	No value set	Batch Size	100	Obtain Generated Keys	false	Rollback On Failure	false
Property	Value																				
JDBC Connection Pool	mysqlDBCPConnectionPool																				
SQL Statement	No value set																				
Support Fragmented Transactions	true																				
Database Session AutoCommit	false																				
Transaction Timeout	No value set																				
Batch Size	100																				
Obtain Generated Keys	false																				
Rollback On Failure	false																				

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

## Configure Processor

Stopped

SETTINGS	SCHEDULING	PROPERTIES	COMMENTS																										
Required field																													
<table border="1"> <thead> <tr> <th>Property</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Kafka Brokers</td> <td>localhost:9092</td> </tr> <tr> <td>Security Protocol</td> <td>PLAINTEXT</td> </tr> <tr> <td>Kerberos Service Name</td> <td>No value set</td> </tr> <tr> <td>Kerberos Credentials Service</td> <td>No value set</td> </tr> <tr> <td>Kerberos Principal</td> <td>No value set</td> </tr> <tr> <td>Kerberos Keytab</td> <td>No value set</td> </tr> <tr> <td>SSL Context Service</td> <td>No value set</td> </tr> <tr> <td>Topic Name</td> <td>druid-\${staging_model}</td> </tr> <tr> <td>Delivery Guarantee</td> <td>Guarantee Replicated Delivery</td> </tr> <tr> <td>Use Transactions</td> <td>true</td> </tr> <tr> <td>Transactional Id Prefix</td> <td>No value set</td> </tr> <tr> <td>Attributes to Send as Headers (Regex)</td> <td>No value set</td> </tr> </tbody> </table>				Property	Value	Kafka Brokers	localhost:9092	Security Protocol	PLAINTEXT	Kerberos Service Name	No value set	Kerberos Credentials Service	No value set	Kerberos Principal	No value set	Kerberos Keytab	No value set	SSL Context Service	No value set	Topic Name	druid-\${staging_model}	Delivery Guarantee	Guarantee Replicated Delivery	Use Transactions	true	Transactional Id Prefix	No value set	Attributes to Send as Headers (Regex)	No value set
Property	Value																												
Kafka Brokers	localhost:9092																												
Security Protocol	PLAINTEXT																												
Kerberos Service Name	No value set																												
Kerberos Credentials Service	No value set																												
Kerberos Principal	No value set																												
Kerberos Keytab	No value set																												
SSL Context Service	No value set																												
Topic Name	druid-\${staging_model}																												
Delivery Guarantee	Guarantee Replicated Delivery																												
Use Transactions	true																												
Transactional Id Prefix	No value set																												
Attributes to Send as Headers (Regex)	No value set																												
		CANCEL	APPLY																										

- 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 .

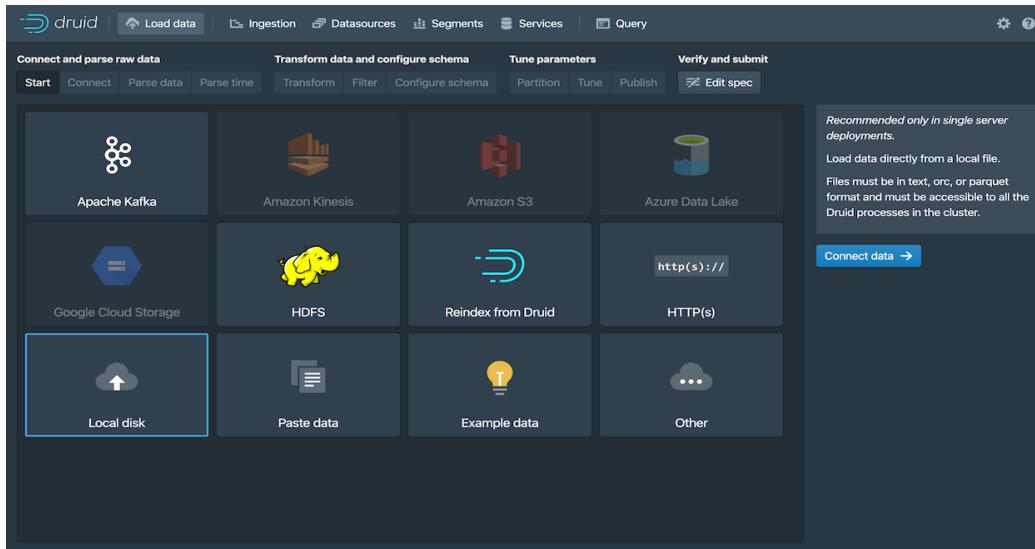
Configuration for HDFS

Property	Possible Values	Description	Default
druid.storage.type	hdfs		Must be set.
druid.storage.storageDirectory		Directory for storing segments.	Must be set.
druid.hadoop.security.kerberos.principal	druid@EXAMPLE.COM	Principal user name	empty
druid.hadoop.security.kerberos.keytab	/etc/security/keytabs/druid.headlessUser.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.

The screenshot shows the Apache Druid unified console interface. In the 'Load data' tab, under 'Connect and parse raw data', there is a large text area containing a list of aviation accident logs from 2016. These logs include details such as date, location, and accident severity. To the right, in the 'Verify and submit' section, there is a configuration panel for a new data source. The 'Source type' dropdown is set to 'hdbs', and the 'Paths' field contains the path '/data/staging\_aviation\_accident'. A blue 'Apply' button is visible at the bottom of this panel. A message box on the right provides instructions for getting started with Druid.

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

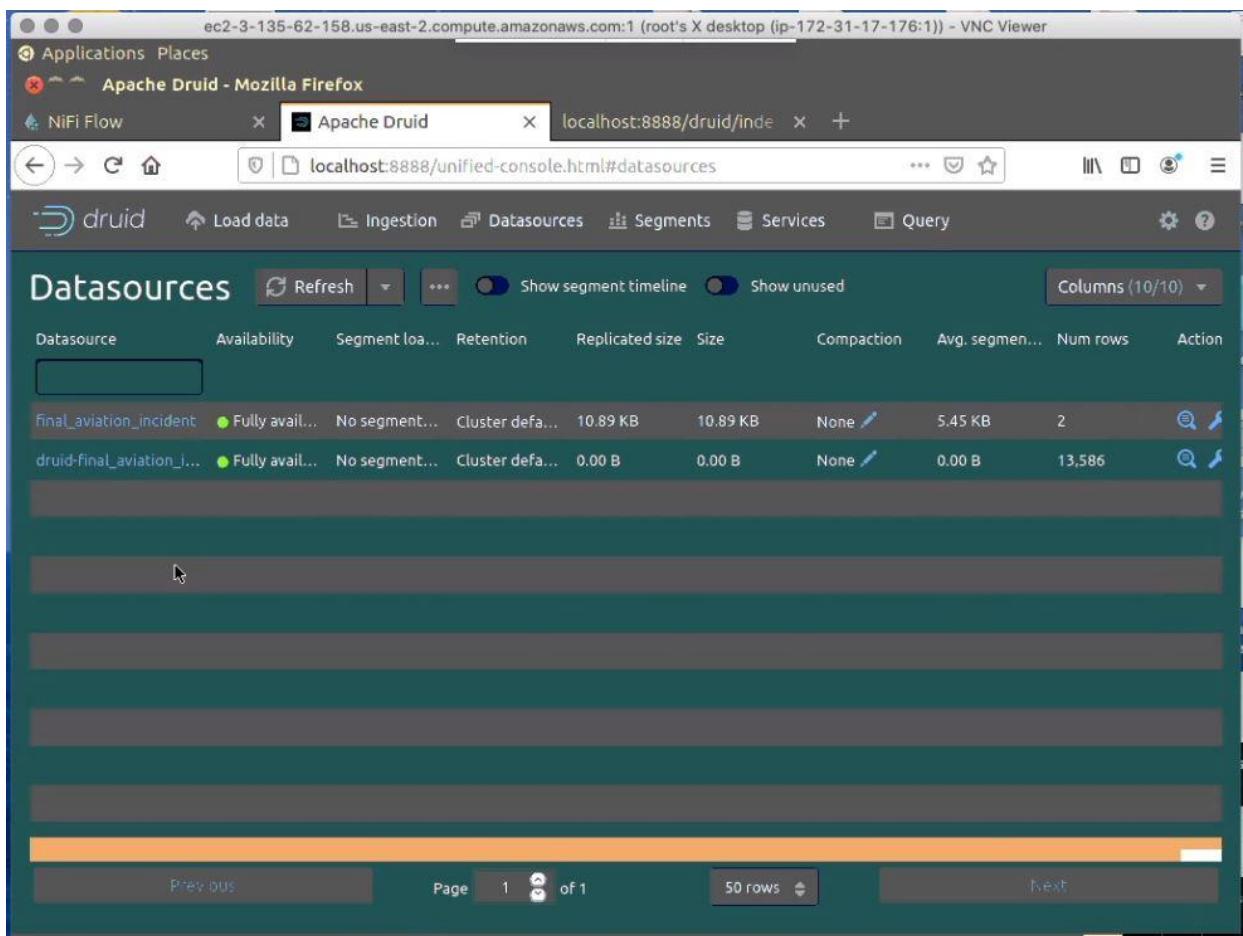
- Here we are consuming the data from the kafka topic by submitting the kafka-supervisor.json file
- Copy the code from supervisor.json file code into load data by configuring Kafka broker URL and topic name as localhost:9092 , topic name

The screenshot shows the Apache Druid unified console interface. In the 'Ingestion' tab, the 'Supervisors' section displays a table with one row for 'druid-final\_aviation\_incident'. The 'Type' column is 'kafka', the 'Topic/Stream' column is 'druid-final\_aviation\_in...', and the 'Status' column is 'RUNNING'. Below this, the 'Tasks' section shows a table with four rows. The 'Group ID' column lists task names like 'index\_kafka\_druid-final\_aviation\_incident', 'index\_parallel\_final\_aviation\_incident\_ikihmlf...', etc. The 'Type' column shows 'index\_kafka' or 'index\_parallel'. The 'Datasource' column is 'druid-final\_aviation\_incident'. The 'Location' column shows 'localhost:8100'. The 'Created time' and 'Status' columns show the tasks were created on 2020-08-02T14:58:29.786Z and are marked as 'SUCCESS'. The 'Duration' column shows times around 0:00:07. Navigation buttons for 'Previous', 'Page 1', 'of 1', '20 rows', and 'Next' are visible at the bottom of both tables.

## 4. 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.



The screenshot shows the Apache Druid Datasources page in Mozilla Firefox. The URL in the address bar is `localhost:8888/unified-console.html#datasources`. The page displays a table of data sources with the following columns: Datasource, Availability, Segment lo..., Retention, Replicated size, Size, Compaction, Avg. segmen..., Num rows, and Action. Two rows are visible:

Datasource	Availability	Segment lo...	Retention	Replicated size	Size	Compaction	Avg. segmen...	Num rows	Action
final_aviation_incident	Fully avail...	No segment...	Cluster defa...	10.89 KB	10.89 KB	None	5.45 KB	2	 
druid-final_aviation_i...	Fully avail...	No segment...	Cluster defa...	0.00 B	0.00 B	None	0.00 B	13,586	 

At the bottom of the page, there are navigation buttons for Previous, Page 1 of 1, 50 rows, and Next.

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

ec2-3-135-62-158.us-east-2.compute.amazonaws.com:1 (root's X desktop (ip-172-31-17-176:1)) - VNC Viewer

Applications Places

Apache Druid - Mozilla Firefox

NiFi Flow Apache Druid

localhost:8888/unified-console.html#query

druid

Load data Ingestion Datasources Segments Services Query

```

druid
  druid-final_aviation_incident_d
    _time
    aircraft_type
    airline_operator
    class
    count
    flight_phase
    injury_level
    is_it_airplane
    is_it_helicopter
    location
    plane_model
    plane_register_no
    state_of_occurrence
    state_of_registry_code
    sum_deaths
    sum_year

```

1 SELECT  
2 "aircraft\_type", count(\*)  
3  
4 FROM "druid-final\_aviation\_incident\_data"  
5 group by aircraft\_type  
6

Run Auto run Smart query limit 99+ results in 3.82s

aircraft_type	EXPR\$1
A109	17
A119	7
A124	11
A129	1
A139	9
A140	2
A148	10
A158	1

Previous Page 1 of 5 Next 20 rows

3.

ec2-3-135-62-158.us-east-2.compute.amazonaws.com:1 (root's X desktop (ip-172-31-17-176:1)) - VNC Viewer

Applications Places

Apache Druid - Mozilla Firefox

NiFi Flow Apache Druid

localhost:8888/unified-console.html#query

druid

Load data Ingestion Datasources Segments Services Query

```

druid
  druid-final_aviation_incident_d
    _time
    aircraft_type
    airline_operator
    class
    count
    flight_phase
    injury_level
    is_it_airplane
    is_it_helicopter
    location
    plane_model
    plane_register_no
    state_of_occurrence
    state_of_registry_code
    sum_deaths
    sum_year

```

1 SELECT  
2 aircraft\_type,airline\_operator,class,  
3 COUNT(\*) AS "Count"  
4 FROM "druid-final\_aviation\_incident\_data"  
5 where flight\_phase='Landing'  
6 GROUP BY 1,2,3  
7 ORDER BY "Count" DESC

Run Auto run Smart query limit 99+ results in 0.45s

aircraft_type	airline_operator	class	Count
B37M	null	Serious Incident	9
PA47	null	Incident	9
B37M	null	Occurrence without safety eff	5
BES8	Brazil Other	Serious Incident	5
B37M	null	Incident	5
A320	Hungary Wizz Air Hungary Ltc	Occurrence without safety eff	5
AC50	Romania Tarom, Romanian Air	Occurrence without safety eff	4
PA47	null	Serious Incident	4

Previous Page 1 of 5 Next 20 rows

ec2-3-135-62-158.us-east-2.compute.amazonaws.com:1 (root's X desktop (ip-172-31-17-176:1)) - VNC Viewer

Applications Places

Apache Druid - Mozilla Firefox

NiFi Flow Apache Druid

localhost:8888/unified-console.html#query

druid

druid-final\_aviation\_incident\_d

- ⌚ \_\_time
- ⚠ aircraft\_type
- ⚠ airline\_operator
- ⚠ class
- 123 count
- ⚠ flight\_phase
- ⚠ injury\_level
- ⚠ is\_it\_airplane
- ⚠ is\_it\_helicopter
- ⚠ location
- ⚠ plane\_model
- ⚠ plane\_register\_no
- ⚠ state\_of\_occurrence
- ⚠ state\_of\_registry\_code
- 123 sum\_deaths
- 123 sum\_year

```
1 SELECT
2   "injury_level", "plane_model",
3   COUNT(*) AS "Count"
4 FROM "druid-final_aviation_incident_data"
5 WHERE "injury_level"='Serious'
6 GROUP BY 1,2
7 ORDER BY "Count" DESC
```

Run Auto run Smart query limit 15 results in 0.31s

injury_level	plane_model	Count
Serious	AIRBUS A320	2
Serious	BOEING 737 800	2
Serious	AIRBUS A340 300	1
Serious	DE HAVILLAND DHC8 400	1
Serious	DE HAVILLAND DHC2 I	1
Serious	DASSAULT FALCON900	1
Serious	BOEING 777 300	1
Serious	BOEING 767 300	1

Previous Page 1 of 1 20 rows Next

ec2-3-135-62-158.us-east-2.compute.amazonaws.com:1 (root's X desktop (ip-172-31-17-176:1)) - VNC Viewer

Applications Places

Apache Druid - Mozilla Firefox

NiFi Flow Apache Druid

localhost:8888/unified-console.html#query

druid

druid-final\_aviation\_incident\_d

- ⌚ \_\_time
- ⚠ aircraft\_type
- ⚠ airline\_operator
- ⚠ class
- 123 count
- ⚠ flight\_phase
- ⚠ injury\_level
- ⚠ is\_it\_airplane
- ⚠ is\_it\_helicopter
- ⚠ location
- ⚠ plane\_model
- ⚠ plane\_register\_no
- ⚠ state\_of\_occurrence
- ⚠ state\_of\_registry\_code
- 123 sum\_deaths
- 123 sum\_year

```
1 SELECT
2   "injury_level", "plane_model", "state_of_occurrence",
3   COUNT(*) AS "Count"
4 FROM "druid-final_aviation_incident_data"
5 WHERE "injury_level"='Serious'
6 GROUP BY 1,2,3
7 ORDER BY "Count" DESC
```

Run Auto run Smart query limit 17 results in 0.34s

injury_level	plane_model	state_of_occurrence	Count
Serious	AIRBUS	LKA	1
Serious	AIRBUS A330	TTO	1
Serious	DE HAVILLAND DHC8 400	LVA	1
Serious	MCDONNELL DOUGLAS	VEN	1
Serious	DE HAVILLAND DHC2 I	null	1
Serious	DASSAULT FALCON900	ITA	1
Serious	BOEING 777 300	DEU	1
Serious	BOEING 767 300	AUS	1

Previous Page 1 of 1 20 rows Next

ec2-3-136-62-158.us-east-2.compute.amazonaws.com:1 (root's X desktop (ip-172-31-17-176:1)) - VNC Viewer

Applications Places

Apache Druid - Mozilla Firefox

NiFi Flow Apache Druid

localhost:8888/unified-console.html#query

druid Load data Ingestion Datasources Segments Services Query

druid

druid-final\_aviation\_incident\_d

⌚ \_time  
A aircraft\_type  
A airline\_operator  
A class  
123 count  
A flight\_phase  
A injury\_level  
A is\_it\_airplane  
A is\_it\_helicopter  
A location  
A plane\_model  
A plane\_register\_no  
A state\_of\_occurrence  
A state\_of\_registry\_code  
123 sum\_deaths  
123 sum\_year

```
1 SELECT
2   "injury_level", "plane_model", "state_of_occurrence", "_time",
3   COUNT(*) AS "Count"
4   FROM "druid-final_aviation_incident_data"
5   WHERE "injury_level"='Serious'
6   GROUP BY 1,2,3,4
7   ORDER BY "Count" DESC
```

Run Auto run Smart query limit 17 results in 0.48s

injury_level	plane_model	state_of_occurrence	_time	Count
Serious	AIRBUS	LKA	2015-11-06T00:00:00.000Z	1
Serious	AIRBUS A330	TTO	2015-06-27T00:00:00.000Z	1
Serious	DE HAVILLAND DHC8 4	LVA	2013-10-13T00:00:00.000Z	1
Serious	MCDONNELL DOUGLAS	VEN	2008-02-12T00:00:00.000Z	1
Serious	DE HAVILLAND DHC2 I	null	2008-05-14T00:00:00.000Z	1
Serious	DASSAULT FALCON900	ITA	2010-11-06T00:00:00.000Z	1
Serious	BOEING 777 300	DEU	2009-08-05T00:00:00.000Z	1
Serious	BOEING 767 300	AUS	2013-11-08T00:00:00.000Z	1

Previous Page 1 of 1 Next 20 rows

The screenshot shows the Apache Druid unified console interface. On the left, there is a sidebar with a tree view of datasets and their columns. The main area displays a query results table.

**Query:**

```

1  SELECT
2      "injury_level", "plane_model", "state_of_occurrence", "__time", "sum_deaths",
3      COUNT(*) AS "Count"
4  FROM "druid-final_aviation_incident_data"
5  WHERE "sum_deaths"!=0
6  GROUP BY 1,2,3,4,5
7  ORDER BY "Count" DESC

```

**Run Options:** Run, Auto run, Smart query limit

**Results:** 14 results in 0.34s

injury_level	plane_model	state_of_occurrence	__time	sum_deaths	Count
Fatal	AIRBUS A310 300	PAK	2014-06-24T00:00:00	2	1
Fatal	BEECH 200	null	2009-04-15T00:00:00	2	1
Unknown	CESSNA 208 B	MEX	2013-10-14T00:00:00	24	1
None	United Aircraft Corp LBY		2019-07-25T00:00:00	2	1
None	BOEING 767 200	DOM	2010-02-18T00:00:00	2	1
Fatal	MCDONNELL DOUG null		2010-02-28T00:00:00	2	1
Fatal	Embraer ERJ 190-100 CHN		2012-06-29T00:00:00	4	1
Fatal	Boeing 737-8E9 (WL BGD)		2019-02-24T00:00:00	2	1

Page: 1 of 1    20 rows

## 5. 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](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 :

```

0: jdbc:hive2://> select aircraft_type,airline_operator,class,count(*) as count from staging_aviation_incident where flight_phase='Landing' group by aircraft_type,airline_operator,class order by count desc;
20/08/02 18:16:07 [HiveServer2-Background-Pool: Thread-121]: WARN ql.Driver: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20200802181607_8c7f294e-b4fa-44dd-b826-3a8189b4ebd
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
20/08/02 18:16:07 [HiveServer2-Background-Pool: Thread-121]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
It's a problem with the configuration property hadoop.tmp.dir in your core-site.xml. It's temporary directory to store the temporary output files (from Mapper) to your local temporary directory.
20/08/02 18:16:08 [LocalJobRunner Map Task Executor #0]: WARN lazy.LazyStruct: Extra bytes detected at the end of the row! Ignoring similar problems.
2020-08-02 18:16:08 Stage-1 map = 100%, reduce = 100%
Ended Job = job_local1066185543_0002
Launching Job 2 out of 2
You can remove that property so that hadoop creates its own temporary directory to store the Number of reduce tasks determined at compile time: 1 some directory with appropriate permissions.
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
20/08/02 18:16:08 [HiveServer2-Background-Pool: Thread-121]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
2020-08-02 18:16:09 Stage-2 map = 100%, a reduce = 100% (output name)
+-----+-----+-----+-----+
| A319 | United Kingdom Easyjet Airline Company Ltd | Serious incident | Incident | |
| A319 | Turkey Pegasus Hava Tasimacılık A.s. (Pegasus) sales it's | Serious incident | Serious incident |
| A319 | What's this? | Turkey Mng Havayolları Ve Tasimacılık A.s. (Pegasus) sales it's | Serious incident | Serious incident |
| A319 | Day Trial | Philippines Cebu Pacific Air | Serious incident | Serious incident |
| A319 | | China China Eastern Airlines follow | Serious incident | Serious incident |
| A319 | | Canada Air Canada | Serious incident | Serious incident |
| A308 | | Singapore Singapore Airlines Limited | Serious incident | Serious incident |
| A306 | | Saudi Arabia Saudi Arabian Airlines | Serious incident | Serious incident |
| A306 | | Australia Qantas Airways Limited | Serious incident | Incident |
| A119 | | Italy | Serious incident | Serious incident |
| A119 | | | Incident | Incident |
| A109 | | try to write hdfs commands like these | Serious incident | Serious incident |
| A109 | | Italy Other | Serious incident | Serious incident |
+-----+-----+-----+-----+
433 rows selected (2.688 seconds)
0: jdbc:hive2://> or did you check the before to export your (jar file) did you take any warnings ?
0: jdbc:hive2://> select aircraft_type,count(*) from staging_aviation_incident group by aircraft_type;
20/08/02 18:18:38 [HiveServer2-Background-Pool: Thread-215]: WARN ql.Driver: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20200802181838_5922fc55-1ac8-42dc-bb62-6301245abc29
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
20/08/02 18:18:38 [HiveServer2-Background-Pool: Thread-215]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
20/08/02 18:18:38 [LocalJobRunner Map Task Executor #0]: WARN lazy.LazyStruct: Extra bytes detected at the end of the row! Ignoring similar problems.
2020-08-02 18:18:39,350 Stage-1 map = 100%, reduce = 100%
Ended Job = job_local34424947_0004 because Hadoop is unable to store the temporary output files (from Mapper) to your MapReduce Jobs Launched: disk
Stage-Stage-1: HDFS Read: 11419752 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec

```

```

| X15 | 1 |
| Y141 | 1 |
| YK40 | 3 |
| YK42 | 9 |
| YK55 | 1 |
| Z42 | 1 |
| ZA6 | 2 |
+-----+----+
try to write hdfs commands like

489 rows selected (1.352 seconds)

0: jdbc:hive2://> select injury_level,plane_model,count(*) as count from staging_aviation_incident where injury_level='Serious' group by injury_level,plane_model order by count desc;
20/08/02 18:20:17 [HiveServer2-Background-Pool: Thread-268]: WARN ql.Driver: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20200802182017_c3aaaf82a-c020-44ed-83b6-19d3c306ef18
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
20/08/02 18:20:17 [HiveServer2-Background-Pool: Thread-268]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
0: jdbc:hive2://> select injury_level,plane_model,count(*) as count from staging_aviation_incident where injury_level='Serious' group by injury_level,plane_model order by count desc;
20/08/02 18:20:17 [HiveServer2-Background-Pool: Thread-268]: WARN ql.Driver: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20200802182017_c3aaaf82a-c020-44ed-83b6-19d3c306ef18
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
20/08/02 18:20:17 [HiveServer2-Background-Pool: Thread-268]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (Local Hadoop)
20/08/02 18:20:17 [LocalJobRunner Map Task Executor #0]: WARN lazy.LazyStruct: Extra bytes detected at the end of the row! Ignoring similar problems.
2020-08-02 18:20:18,446 Stage-1 map = 100%, reduce = 100%
Ended Job = job_local490050267_0005
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1 so that hadoop creates it's own temporary directory to store the
In order to change the average load for a reducer (in bytes): story with appropriate permissions.

add a comment

```

```

+-----+-----+-----+
| injury_level | plane_model | count |
+-----+-----+-----+
| Serious | BOEING 737 800 8AS | 2 |
| Serious | AIRBUS A320 | 2 |
| Serious | BOEING 737 800 | 2 |
| Serious | DE HAVILLAND DHC2 | 1 |
| Serious | DE HAVILLAND DHC8 400 | 1 |
| Serious | DASSAULT FALCON900 | 1 |
| Serious | BOEING 767 300 | 1 |
| Serious | BOEING 737 900 | 1 |
| Serious | AIRBUS A340 300 | 1 |
| Serious | AIRBUS A330 | 1 |
| Serious | AIRBUS | 1 |
| Serious | Papua | 1 |
| Serious | Hessen" | 1 |
+-----+-----+-----+
13 rows selected (2.559 seconds)
0: jdbc:hive2://> or did you check the before to export your (jar file
0: jdbc:hive2://> select to_date(incident_occurred_date) as date_occurred, rtrim(plane_register_no) as pln_reg_num from staging_aviation_incident where flight_phase='Landing' and injury_level='Serious' limit 10 ;
OK
20/08/02 18:37:03 [d0c08354-0aff-401d-b019-ba9d168a4029 main]: WARN lazy.LazyStruct: Extra bytes detected at the end of the row! Ignoring similar problems.
+-----+-----+-----+
| date_occurred | pln_reg_num | Type
+-----+-----+-----+
| 2008-12-06 | VH-VQR | It will give the reversed string of X
| 2015-11-06 | TC-JOA | It will fetch and gives str, which is right-padded with pad to a length of length(integer value)
| 2011-08-12 | RP-C8607 | g pad|
+-----+-----+-----+
3 rows selected (0.104 seconds)
0: jdbc:hive2://> list_or_occurrence_category, location, plane_model, plane_register_no, state_or_occurrence, state_or_registry_code, year) (state=42000,code=10004)
0: jdbc:hive2://> select to_date(incident_occurred_date) as date_occurred,rtrim(plane_register_no) as pln_reg_num from staging_aviation_incident where flight_phase='Landing' and injury_level='Serious' limit 10 ;
OK
20/08/02 18:37:03 [d0c08354-0aff-401d-b019-ba9d168a4029 main]: WARN lazy.LazyStruct: Extra bytes detected at the end of the row! Ignoring similar problems.
+-----+-----+-----+
| date_occurred | pln_reg_num | Type
+-----+-----+-----+
| 2008-12-06 | VH-VQR | It will give the reversed string of X
| 2015-11-06 | TC-JOA | It will fetch and gives str, which is right-padded with pad to a length of length(integer value)
| 2011-08-12 | RP-C8607 | g pad|
+-----+-----+-----+
3 rows selected (0.104 seconds)
0: jdbc:hive2://>

```

deaths	state_of_occurrence	flight_phase		col)
NULL	C-GKEP	DOUBLE	min(col)	
NULL	B-18358	En route		
NULL	EI-EKS	En route		max(col)
NULL	D-AICF	En route		
NULL	TCA	En route		
NULL	IRN	En route		
NULL	BOMBARDIER CL600 2B19	En route		
NULL	AIRBUS A321 200	En route		

308 rows selected (2.497 seconds)

## 6. 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 :

```

mysql> select injury_level, plane_model from aviation_incident where injury_level='Serious' group by injury_level,plane_model;
+-----+-----+
| injury_level | plane_model |
+-----+-----+
| Serious | AIRBUS |
| Serious | AIRBUS A320 |
| Serious | AIRBUS A340 300 |
| Serious | BAE BAE146 200 |
| Serious | BOEING 737 200 |
| Serious | BOEING 737 800 |
| Serious | BOEING 737 900 |
| Serious | BOEING 767 300 |
| Serious | BOEING 777 300 |
| Serious | DASSAULT FALCON900 |
| Serious | DE HAVILLAND DHC2 I |
| Serious | DE HAVILLAND DHC8 400 |
| Serious | MCDONNELL DOUGLAS |
+-----+-----+
13 rows in set (0.00 sec)

```

```

mysql> select aircraft_type,plane_model, injury_level, count(deaths) from aviation_incident where flight_phase='Landing' group by aircraft_type,plane_model,injury_level order by count(deaths) desc;

```

```

+-----+-----+-----+-----+-----+-----+-----+
| aircraft_type | plane_model | flight_phase | group_injury_level | count(deaths) | airline | sort_order |
+-----+-----+-----+-----+-----+-----+-----+
| A359 | AIRBUS A300 600 | None | group_injury_level | 1 | prior_to_landing | 1 |
| E145 | EMBRAER EMB145 ER | None | group_injury_level | 1 | prior_to_landing | 1 |
| AN3 | Antonov An-24RV | None | group_injury_level | 1 | prior_to_landing | 1 |
| SC7 | SHORT SC7 | Unknown | group_injury_level | 1 | prior_to_landing | 1 |
| A504 | BAE AVRO146RJ 300 | None | group_injury_level | 1 | prior_to_landing | 1 |
| AVIN metallurgie | BAE AVRO146RJ 300 | None | group_injury_level | 1 | prior_to_landing | 1 |
| A345 | AIRBUS A340 300 | None | group_injury_level | 1 | prior_to_landing | 1 |
| GLST | BOMBARDIER BD700 1A10 NO SERIES EXISTS | None | group_injury_level | 1 | prior_to_landing | 1 |
| DH94 | DE HAVILLAND DHC8 300 | None | group_injury_level | 1 | prior_to_landing | 1 |
| CL60 | CANADAIR CL600 1A11 | Unknown | group_injury_level | 1 | prior_to_landing | 1 |
| UZ124N AL-PRH BEECH 90 F90 | None | group_injury_level | 1 | prior_to_landing | 1 |
| SJET | None | Minor | group_injury_level | 1 | prior_to_landing | 1 |
| A310 | AIRBUS A310 | None | group_injury_level | 1 | prior_to_landing | 1 |
| RC70 | ROCKWELL | None | group_injury_level | 1 | prior_to_landing | 1 |
| L355 |LEARJET 25 D | None | group_injury_level | 1 | prior_to_landing | 1 |
| B39M | BOEING 737 900 | None | group_injury_level | 1 | prior_to_landing | 1 |
| GLF5 | GULFSTREAM GV SP G550 | None | group_injury_level | 1 | prior_to_landing | 1 |
| F28 | FOKKER F28 100 | None | group_injury_level | 1 | prior_to_landing | 1 |
| F15 | MCDONNELL DOUGLAS | None | group_injury_level | 1 | prior_to_landing | 1 |
| DH80 | DE HAVILLAND | None | group_injury_level | 1 | prior_to_landing | 1 |
| BE10 | BEECH B100 | None | group_injury_level | 1 | prior_to_landing | 1 |
| BE9L | BEECH 90 | None | group_injury_level | 1 | prior_to_landing | 1 |
| MD83 | MCDONNELL DOUGLAS | None | group_injury_level | 1 | prior_to_landing | 1 |
| A306 | Airbus A330-202 | None | group_injury_level | 1 | prior_to_landing | 1 |
| BE9L | BEECH 90 F90 | None | group_injury_level | 1 | prior_to_landing | 1 |
+-----+-----+-----+-----+-----+-----+-----+
304 rows in set (0.01 sec)

```

```

mysql>

```

```

mysql> select aircraft_type,plane_model, injury_level, count(deaths) from aviation_incident where flight_phase='Landing' and class='incident' group by aircraft_type,plane_model,injury_level order by count(deaths) desc;

```

A359	AIRBUS A300 600	select distinct	None	aircraf	1	minlev
E145	EMBRAER EMB145 ER	flight_phase	None	group by aircraf_type,plane_m	1	injury_level
AN3	Antonov An-24RV		None		1	
SC7	SHORT SC7		Unknown		1	
A504	BAE AVR0146RJ 300	flight_phase	None	group by aircraf_type,plane_m, injury_level	1	minlev
AVIN	BAE AVR0146RJ 300	and class	None	incident group by aircraf_type,pla	1	
A345	AIRBUS A340 300	count(incide	None		1	
GL5T	BOMBARDIER BD700 1A10 NO SERIES EXISTS		None		1	
DH94	DE HAVILLAND DHC8 300		None		1	
CL60	CANADAIR CL600 1A11		Unknown		1	
U21	BEECH 90 F90		None		1	
SJET			Minor		1	
A310	AIRBUS A310		None		1	
RC70	ROCKWELL		None		1	
LJ55	LEARJET 25 D		None		1	
B39M	BOEING 737 900		None		1	
GLF5	GULFSTREAM GV SP G550		None		1	
F28	FOKKER F28 100		None		1	
F15	MCDONNELL DOUGLAS	select distinct	None	aircraf_type,plane_m, injury_level, count(incide	1	minlev
DH80	DE HAVILLAND	flight_phase	None	and class)	1	injury_level
BE10	BEECH B100	count(incide	None		1	
BE9L	BEECH 90		None		1	
MD83	MCDONNELL DOUGLAS				1	
A306	Airbus A330-202				1	
BE9L	BEECH 90 F90		None		1	

304 rows in set (0.00 sec)

```
mysql> select aircraft_type, plane_model, injury_level, count(deaths) from aviation_incident where flight_phase='Landing' and class='incident' group by aircraft_type, plane_model, injury_level order by count(deaths) desc;
```

A310	AIRBUS A310	Flight phase	None	group by aircraft_type, plane_model, injury_level	1
E145	EMBRAER EMB145 ER		None		1
BE33	RAYTHEON		None		1
VAMP	DE HAVILLAND		None		1
E75L	EMBRAER EMB145 ER		None		1
GL7T	BOMBARDIER BD700 1A10 NO SERIES EXISTS		None		1
BE9L	BEECH 90		None		1
CL60	CANADAIR CL600 1A11		Unknown		1
MD82	MCDONNELL DOUGLAS	Working query	None		1
WW23	IAI 1123	elect injury_level, plane_model from aviation_incident			1
B39M	BOEING 737 900	injury_level, plane_model	None		1
PA47	PIPER PA23 250		Unknown		1
C337	CESSNA	select injury_level, plane_model from aviation_incident			1
A504	BAE AVR0146RJ 300	injury_level, plane_model	None		1
GLF4	GULFSTREAM		None		1
CRJX	BOMBARDIER	select aircraft_type, plane_model, class, count(*) from aviation_incident			1
B37M	BOEING 747 400	flight_phase	Unknown	group by aircraft_type, plane_model	1
E55P	EMBRAER EMB505 PHENOM 300		Minor		1
DH80	DE HAVILLAND	flight_phase	None	group by aircraft_type, plane_model, injury_level	1
F28	FOKKER F28 100		None		1
AVIN	BAE AVR0146RJ 300		None		1
DH94	DE HAVILLAND DHC8 300	select aircraft_type, plane_model, injury_level, count(*) from			1
E190	EMBRAER ERJ190	flight_phase	None	group by aircraft_type, plane_model	1
PA47	PIPER PA28	injury_level	Unknown		1

156 rows in set (0.01 sec)

256 rows in set (0.00 sec)

```
mysql>
mysql> select aircraft_type, plane_model, injury_level, count(deaths) from aviation_incident where flight_phase='Landing' and class!= 'incident' group by aircraft_type, plane_model, injury_level order by count(deaths) desc;
```

B37M	Boeing 737-300	Flight phase	None	group by aircraft_type, plane_model, injury_level	1
BE9L	BEECH 90 F90		None		1
A345	AIRBUS A340 300		None		1
SJET			Minor		1
CS25	CESSNA 525 A		None		1
C404	CESSNA 404 NO SERIES EXISTS		None		1
AN3	Antonov An-24RV		None		1
C25M	CESSNA 525 A		None		1
ZA6	MD HELICOPTER 902		None		1
CL60	CANADAIR		None		1
GLF4	GULFSTREAM		None		1
FA10		Working query	Minor		1
C04T	CESSNA 404 NO SERIES EXISTS		None		1
MD81	MCDONNELL DOUGLAS		None		1
MD90	MCDONNELL DOUGLAS MD90 30		None		1
A359	AIRBUS A300 600		None		1
GLF5	GULFSTREAM GV SP G550		None		1
BE10	BEECH B100		None		1
V10	ROCKWELL		None		1
A306	Airbus A330-202	Working query	None	group by aircraft_type, plane_model, injury_level	1
COL3	KAVANAGH B350		None		1
RC70	ROCKWELL		None		1

198 rows in set (0.00 sec)

## 7. Visualisation of data in Quicksight AWS

To provide access to Quicksight user from MYSQL follow the steps below

Login into MYSQL and create Quicksight 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 :

## New MySQL data source

**Data source name**  
xyz

**Connection type**  
Public network

**Database server**  
ec2-3-129-44-2.us-east-2.compute.amazonaws.com

**Port**  
3306

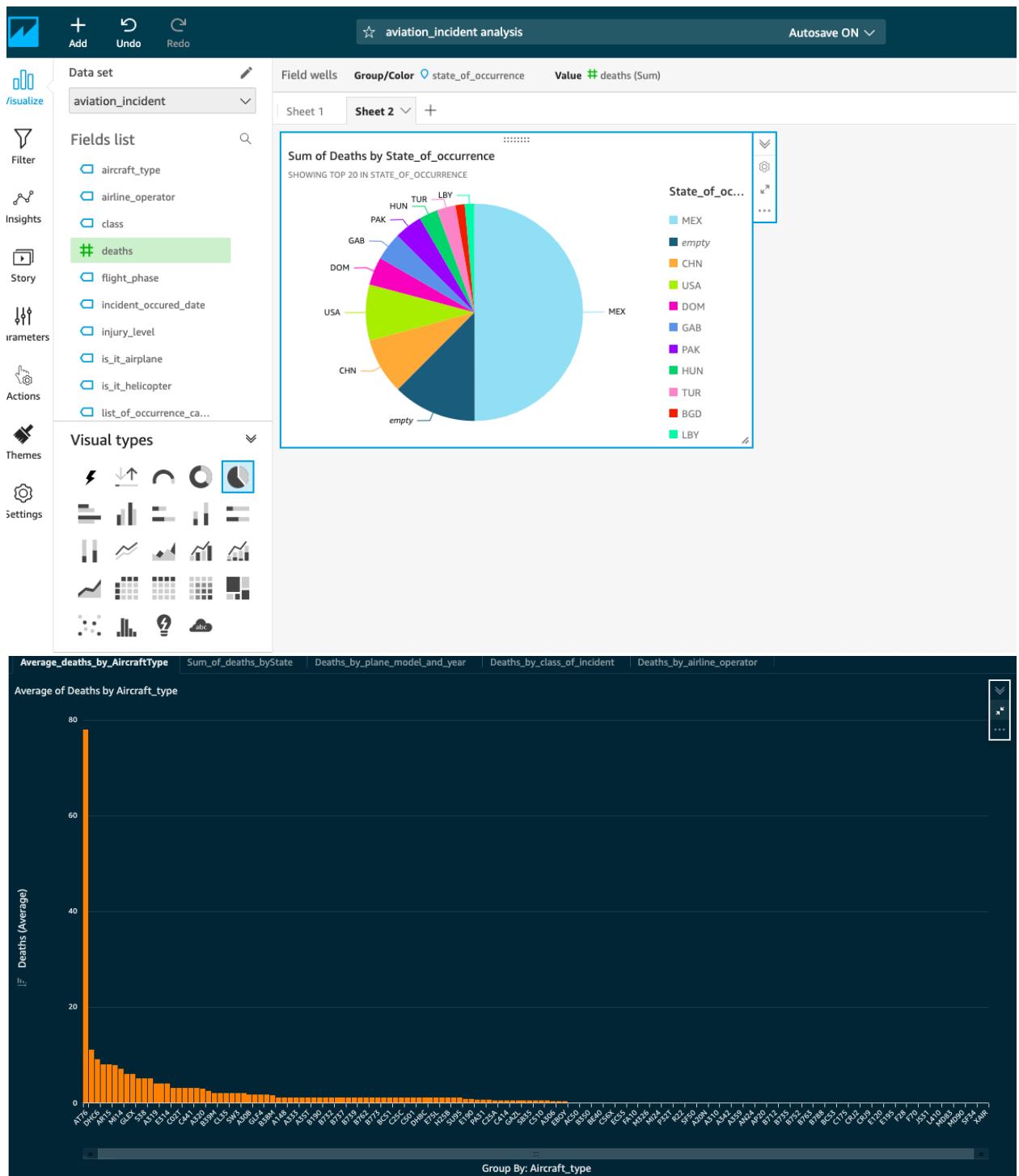
**Database name**  
dezyredb

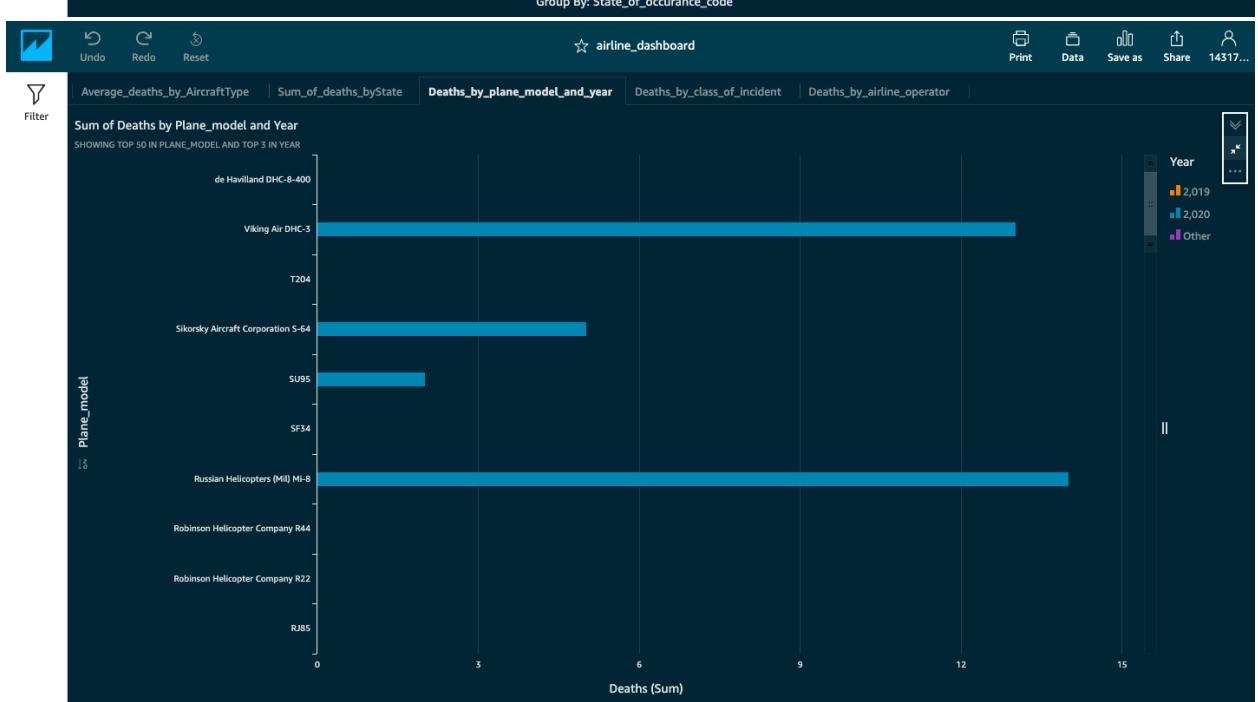
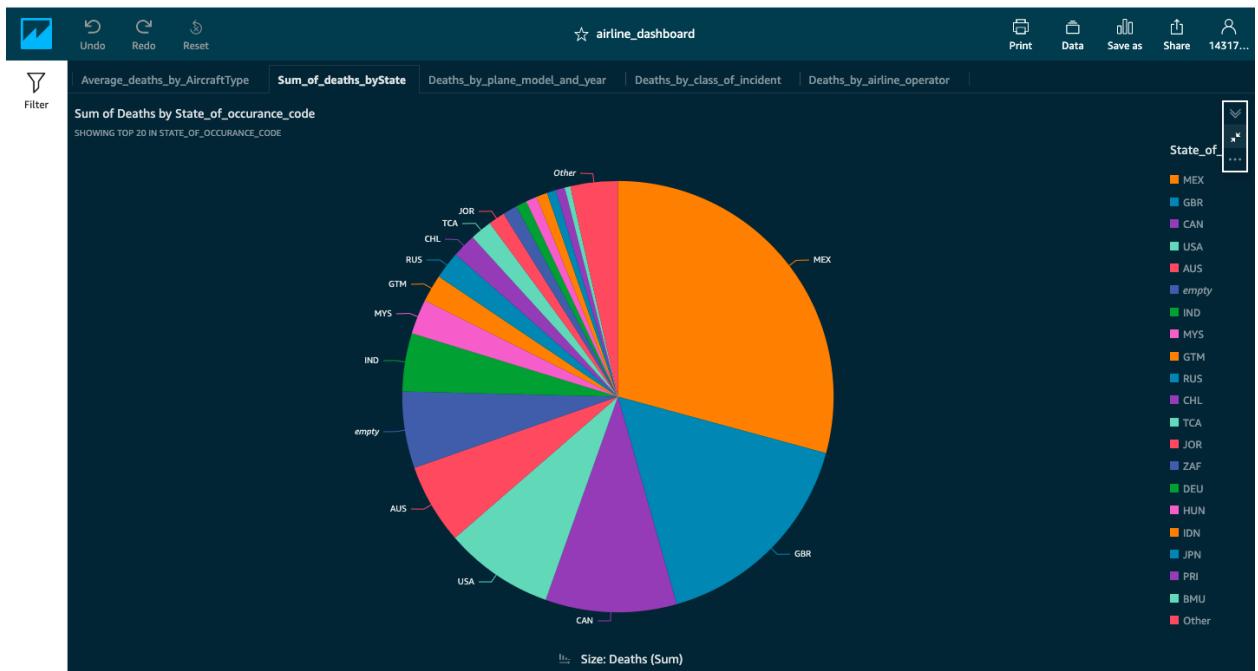
**Username**  
X

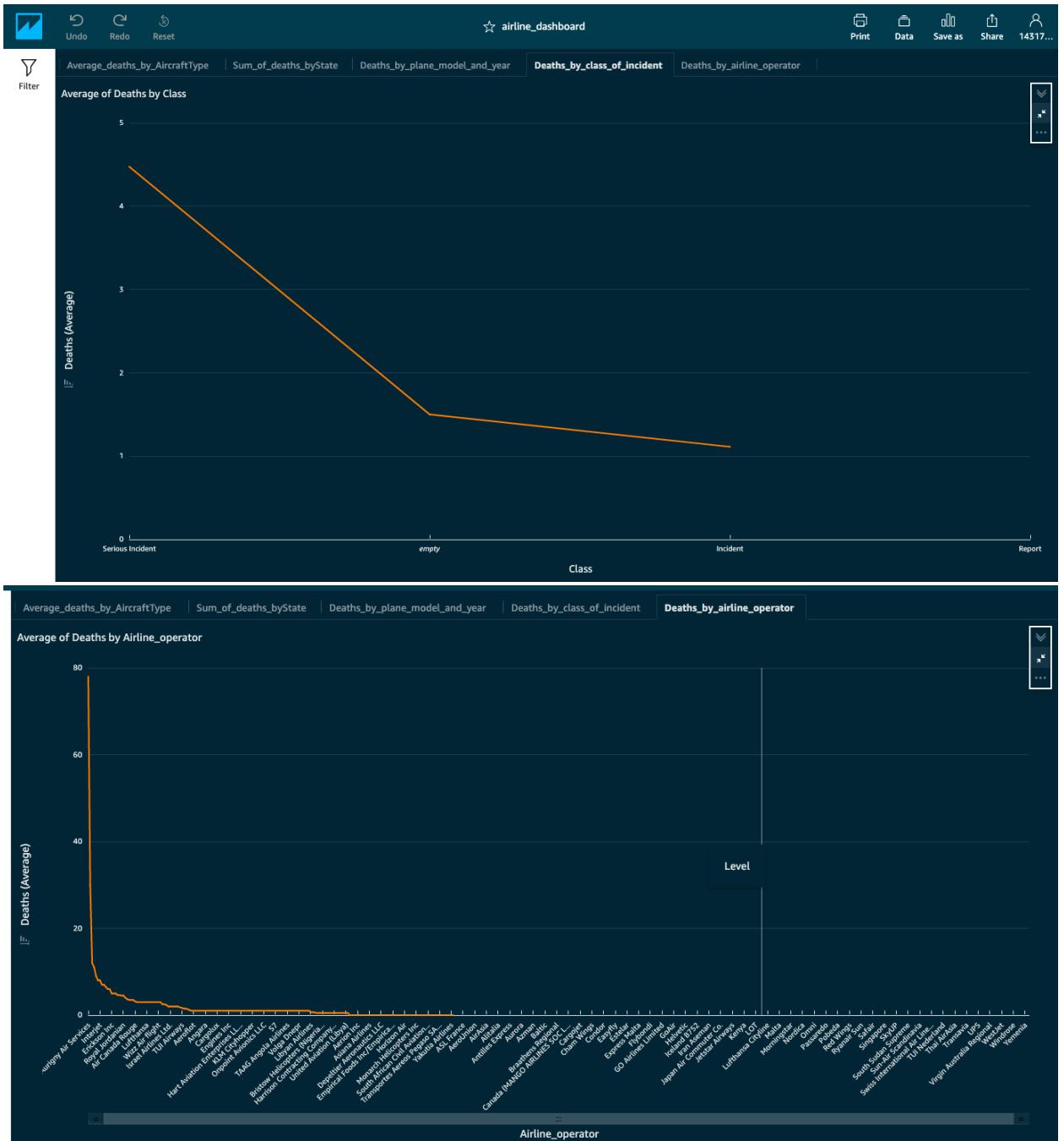
**Password**  
\*\*\*\*

Validated    Enable SSL   **Create data source**

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







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

Discuss on project extensibility using RDS and Sqoop