Table of Contents

Introduction to the Project

A. Objective and Scope

The primary objective of this project is to leverage the advanced tools and techniques offered by the Hadoop Distributed Platform (HDP) for the processing and analysis of a comprehensive dataset. Students will utilize these tools to create analytical reports and visually compelling presentations that can serve as valuable resources for decision-makers.

Specifically, the project aims to conduct an in-depth analysis of data related to truck fleets to gain profound insights into the risk factors associated with trucking operations. Large truck accidents remain a leading cause of injuries and deaths in the United States. The overarching goal is to enhance safety within the trucking industry, addressing a critical issue that has led to numerous injuries and fatalities nationwide. The primary objective is to identify dangerous commercial truck drivers across the country.

B. Significance of Cloudera Big Data Ecosystems

Cloudera's Big Data ecosystem plays a pivotal role in this project by providing a robust and scalable platform for data management and analysis. This ecosystem demonstrates its practical relevance by serving as the technological foundation for addressing real-world challenges within the trucking industry.

The project showcases how Cloudera's Big Data solutions can be effectively applied to improve driver safety, ensure regulatory compliance, and proficiently manage risks in the dynamic trucking sector.

Project Phases

Environment Setup and Data Integration

1) Installation of Tableau Drivers

This phase involves the installation of essential Tableau drivers, including JDBC and ODBC connectors. These drivers are crucial as they enable seamless connectivity between the project environment and the Hadoop File System (HDFS), ensuring efficient data access and processing.

2) Compilation of a Geolocation Dataset

In this stage, students will create the Data Definition Language (DDL) for the geolocation table in Hive (see step#3 below) and then proceed to load the geolocation dataset into this table. This dataset includes crucial information such as latitude, longitude, city, state, and essential truck fleet data, such as total mileage, average mileage, and fuel consumption. Additionally, students will create the necessary geolocation data set "geolocation.csv."

3) Create the geolocation Table in Hive

To prepare the geolocation data for analysis, students will create the geolocation table in Hive using the following Data Definition Language (DDL) command:

```
CREATE TABLE geolocation (
 truckid STRING,
 driverid STRING,
 event STRING,
 latitude DOUBLE,
 longitude DOUBLE,
 city STRING,
 state STRING,
 velocity BIGINT,
 event_ind BIGINT,
 idling_ind BIGINT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
TBLPROPERTIES
("skip.header.line.count"="1");
```

This table will serve as the repository for geolocation data within the Hive environment, ensuring it is structured and organized for further use and analysis.

4) Load the Geolocation Dataset

With the ‘**geolocation’** table in Hive, students will proceed to load the geolocation dataset, which includes the "geolocation.csv" file, into **‘geolocation’** Hive table. Loading the dataset into the Hive table is a critical step in preparing the data for subsequent analysis.

By completing this task, students will have the necessary geolocation data ready within the **geolocation** table in Hive, setting the stage for in-depth exploration and visualization.

5) Loading Data into Hadoop File System (HDFS)

In this task, students will transfer the input "geolocation.csv" file into Cloudera VM by utilizing the system copy command (scp). The students will use the concepts they have learned to upload the data into the HIVE instance by utilizing the “LOAD IN PATH” command.

6) Integration of Hive Table with Tableau

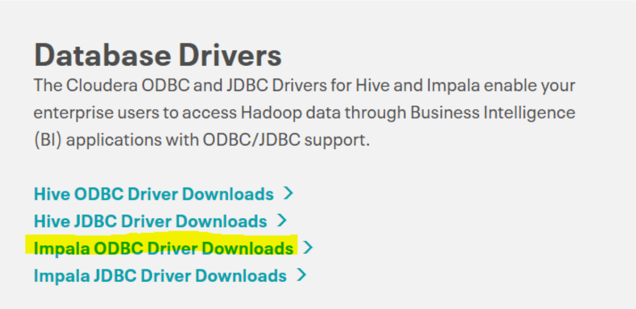
The integration of the Hive table with Tableau is a pivotal step in the project's workflow. This integration ensures that the data stored in HDFS becomes visible and accessible within the Tableau environment, facilitating comprehensive data analysis, exploration, and visualization.

To achieve this integration, students will identify the correct driver, install the ODBC driver for Impala, and load the tables into Tableau via ETL transactions. This will enable the creation of the necessary charts and visualizations needed for the project.

7) Steps for installing Impala ODBC Driver

a) Go to the https://www.cloudera.com/downloads.html
b) Download Impala ODBC Drive only.

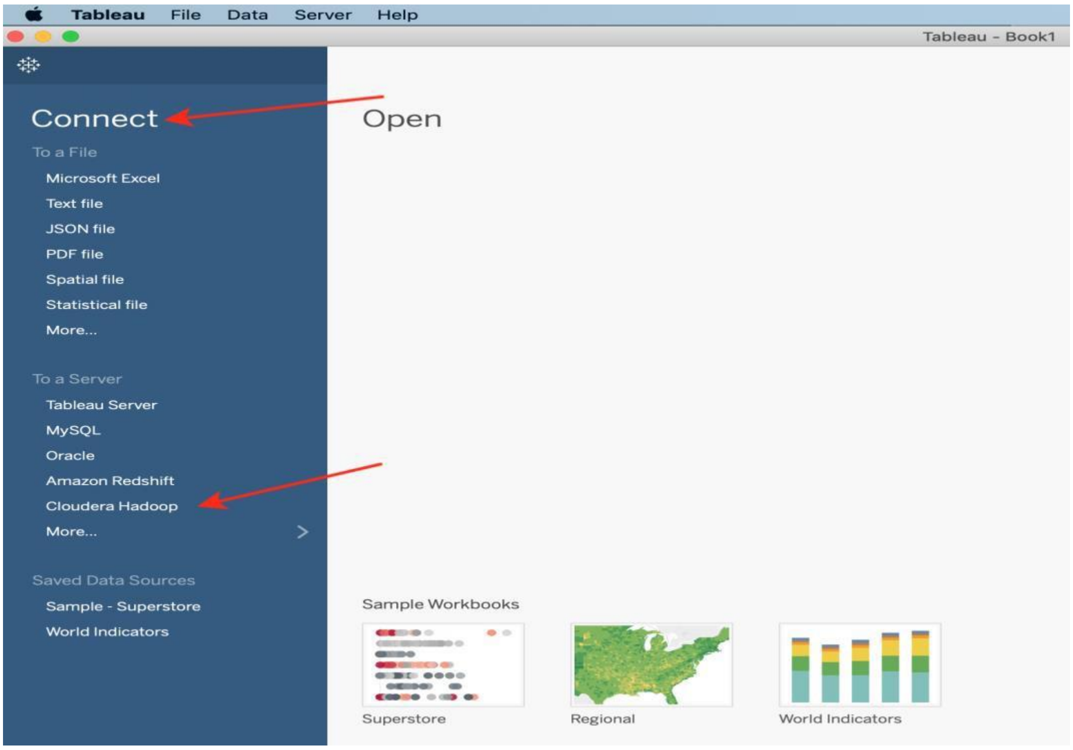
If you are using Tableau as an analytical tool, follow the below steps to connect to HDFS.



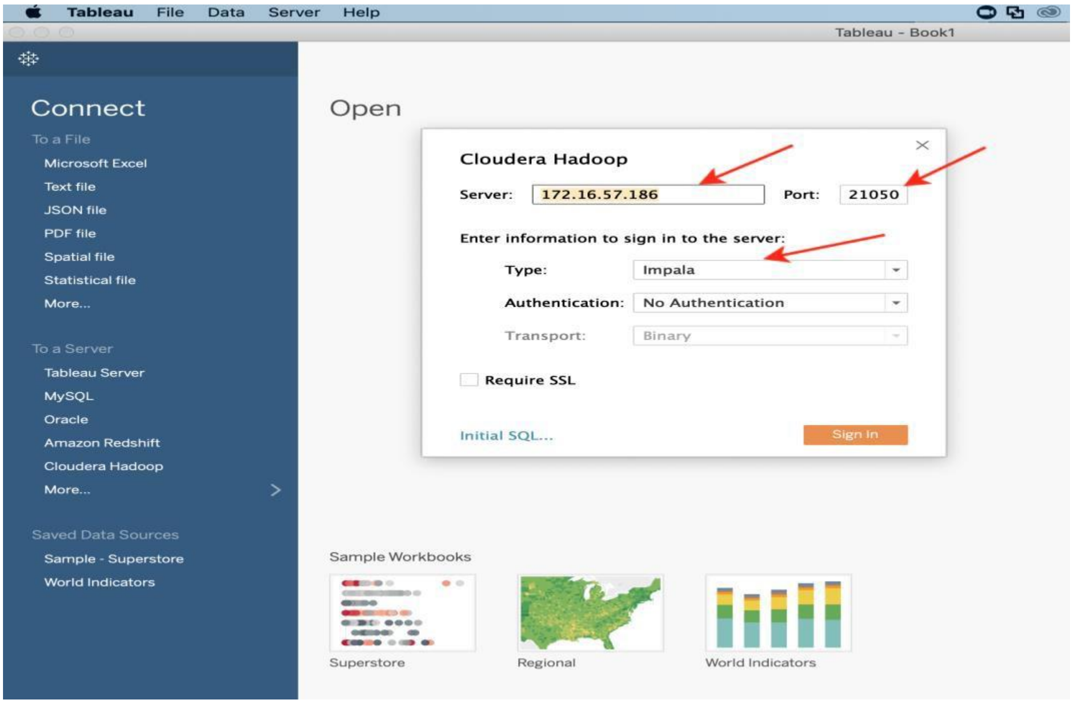
Database Drivers

The Cloudera ODBC and JDBC Drivers for Hive and Impala enable your enterprise users to access Hadoop data through Business Intelligence (BI) applications with ODBC/JDBC support.

Hive ODBC Driver Downloads >
Hive JDBC Driver Downloads >
Impala ODBC Driver Downloads >
Impala JDBC Driver Downloads >



From your Cloudera Image, run the command line ‘hostname -I’ to locate the IP address associated with your Instance. Copy the IP address and inserted into server bar as it shows below. Make sure you are using the port# 21050 and the Type = Impala.



Cloudera Hadoop

Server: 172.16.57.186 Port: 21050

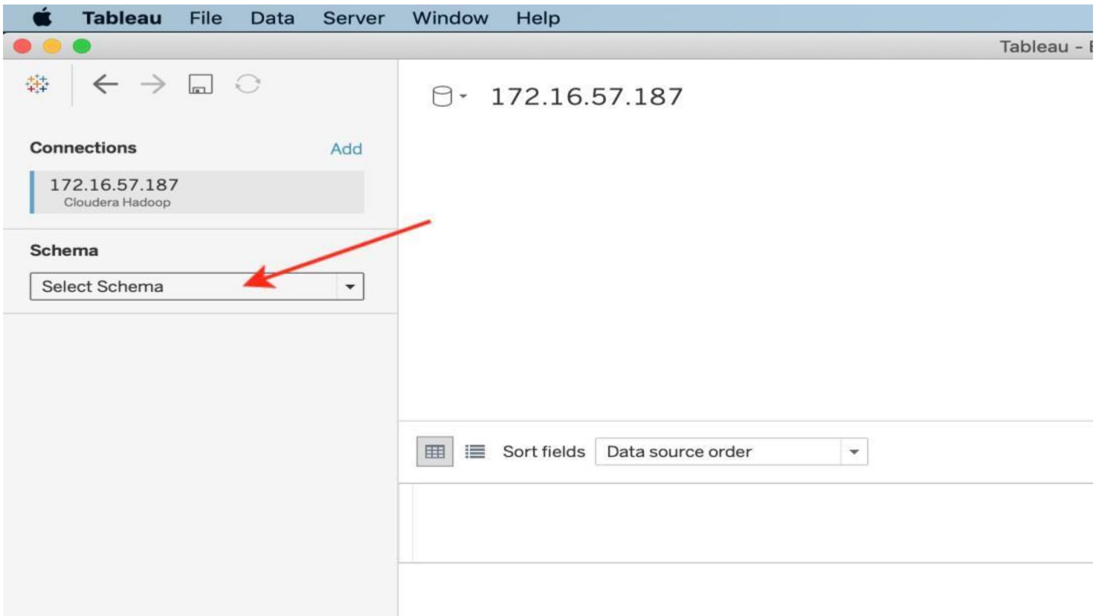
Enter information to sign in to the server:

Type: Impala

Authentication: No Authentication

Transport: Binary

□ Require SSL



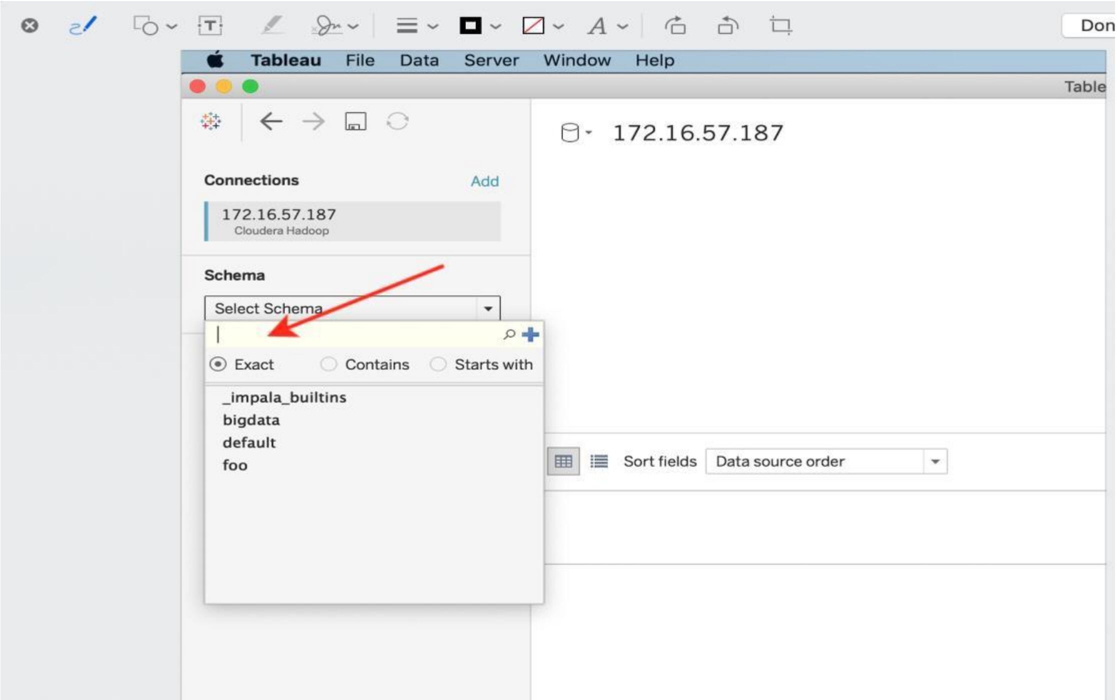
172.16.57.187

Connections: 172.16.57.187 Cloudera Hadoop

Schema: Select Schema

Sort fields: Data source order

Place the cursor at the Schema Bar and hit Enter.



172.16.57.187

Connections: 172.16.57.187 Cloudera Hadoop

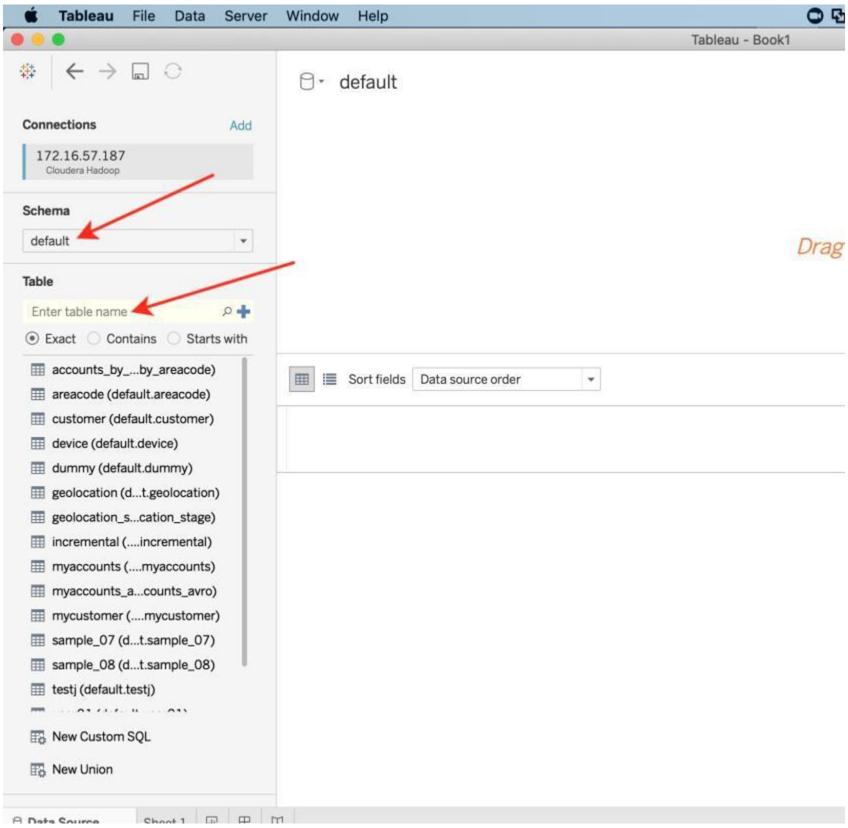
Schema: Select Schema

● Exact ○ Contains ○ Starts with

_impala_builtins
bigdata
default
foo

Sort fields: Data source order

Place the cursor at the Table bar and hit Enter. Then you will start getting the Hive/Impala tables.



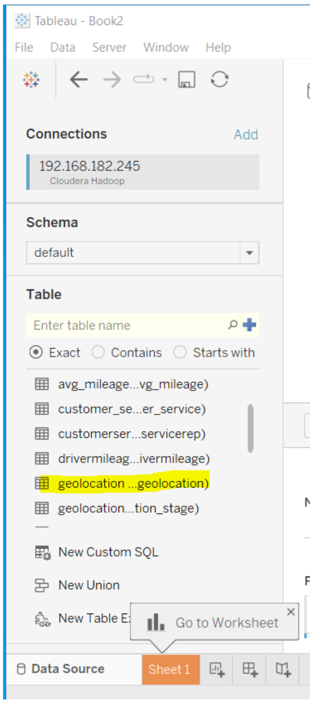
Connections: 172.16.57.187 (Cloudera Hadoop)

Schema: default

Table: Enter table name — ● Exact ○ Contains ○ Starts with

- accounts_by_...by_areacode)
- areacode (default.areacode)
- customer (default.customer)
- device (default.device)
- dummy (default.dummy)
- geolocation (d...t.geolocation)
- geolocation_s...cation_stage)
- incremental (....incremental)
- myaccounts (....myaccounts)
- myaccounts_a...counts_avro)
- mycustomer (....mycustomer)
- sample_07 (d...t.sample_07)
- sample_08 (d...t.sample_08)
- testj (default.testj)

Drag geolocation table to the middle of the sheet and click on “Update now “as it shows below.

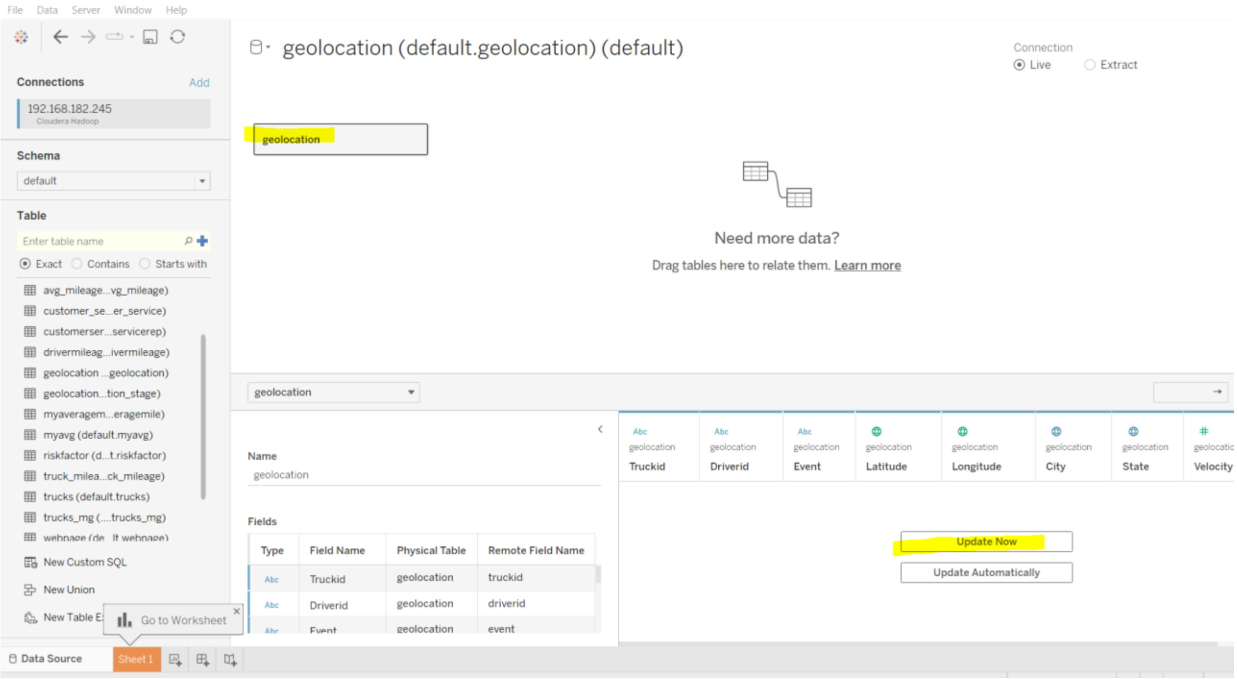


Connections: 192.168.182.245 (Cloudera Hadoop)

Schema: default

Table: Enter table name — ● Exact ○ Contains ○ Starts with

- avg_mileage...vg_mileage)
- customer_se...er_service)
- customerser...servicerep)
- drivermileag...ivermileage)
- geolocation ...geolocation)
- geolocation...tion_stage)

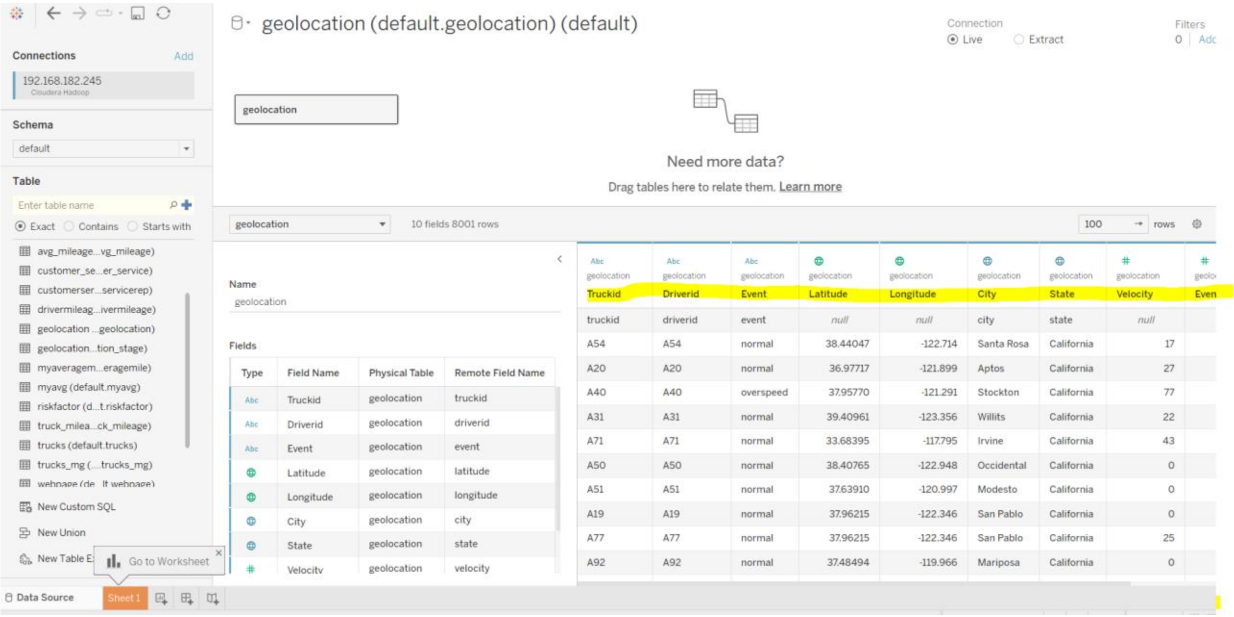

File Data Server Window Help
Connections Add
192.168.182.245
Cloudera Hadoop
Schema
default
Table
Enter table name
Exact Contains Starts with
avg_mileage...vg_mileage)
customer_se...er_service)
customerser...servicerep)
drivermileag...ivermileage)
geolocation ...geolocation)
geolocation...tion_stage)
myaveragem...eragemile)
myavg (default.myavg)
riskfactor (d...t.riskfactor)
truck_milea...ck_mileage)
trucks (default.trucks)
trucks_mg (....trucks_mg)
webpage (de...lt.webpage)
New Custom SQL
New Union
New Table E
Go to Worksheet
Data Source
Sheet 1
geolocation (default.geolocation) (default)
Connection
Live
Extract
geolocation
Need more data?
Drag tables here to relate them. Learn more
geolocation
Truckid
Driverid
Event
Latitude
Longitude
City
State
Velocity
Update Now
Update Automatically

Name
geolocation

Fields

Type	Field Name	Physical Table	Remote Field Name
Abc	Truckid	geolocation	truckid
Abc	Driverid	geolocation	driverid
Abc	Event	geolocation	event

You should have geolocation data migrated into Tableau as it shows below.


Connections Add
192.168.182.245
Cloudera Hadoop
Schema
default
Table
Enter table name
Exact Contains Starts with
avg_mileage...vg_mileage)
customer_se...er_service)
customerser...servicerep)
drivermileag...ivermileage)
geolocation ...geolocation)
geolocation...tion_stage)
myaveragem...eragemile)
myavg (default.myavg)
riskfactor (d...t.riskfactor)
truck_milea...ck_mileage)
trucks (default.trucks)
trucks_mg (....trucks_mg)
webpage (de...lt.webpage)
New Custom SQL
New Union
New Table E
Go to Worksheet
Data Source
Sheet 1
geolocation (default.geolocation) (default)
Connection
Live
Extract
Filters
0 Add
geolocation
Need more data?
Drag tables here to relate them. Learn more
geolocation
10 fields 8001 rows
100 rows

Name
geolocation

Fields

Type	Field Name	Physical Table	Remote Field Name
Abc	Truckid	geolocation	truckid
Abc	Driverid	geolocation	driverid
Abc	Event	geolocation	event
🌐	Latitude	geolocation	latitude
🌐	Longitude	geolocation	longitude
🌐	City	geolocation	city
🌐	State	geolocation	state
#	Velocity	geolocation	velocity

Truckid	Driverid	Event	Latitude	Longitude	City	State	Velocity	Even
truckid	driverid	event	null	null	city	state	null	
A54	A54	normal	38.44047	-122.714	Santa Rosa	California	17	
A20	A20	normal	36.97717	-121.899	Aptos	California	27	
A40	A40	overspeed	37.95770	-121.291	Stockton	California	77	
A31	A31	normal	39.40961	-123.356	Willits	California	22	
A71	A71	normal	33.68395	-117.795	Irvine	California	43	
A50	A50	normal	38.40765	-122.948	Occidental	California	0	
A51	A51	normal	37.63910	-120.997	Modesto	California	0	
A19	A19	normal	37.96215	-122.346	San Pablo	California	0	
A77	A77	normal	37.96215	-122.346	San Pablo	California	25	
A92	A92	normal	37.48494	-119.966	Mariposa	California	0	