Tableau integrated with Hortonworks Data Platform (HDP) running on IBM Power Systems

Steps for discovering and visualizing data in HDP on IBM Power Systems using Tableau

Beth Hoffman
Narayana Pattipati

December 09, 2017

Tableau is a business intelligence tool that allows data to be discovered and visualized. Tableau supports Hadoop environments as a data source. Read this article for details about how Tableau Desktop was tested to integrate with and visualize data in Hortonworks Data Platform (HDP) on IBM POWER8.

Introduction

Tableau provides a business intelligence (BI) solution called Tableau Desktop. Tableau Desktop provides many features including analysis, dashboards and interactive maps. Tableau supports accessing data in Hadoop environments. Validation testing was performed to verify Tableau's ability to integrate with and visualize data specifically to Hortonworks Data Platform (HDP) on IBM® POWER8® processor based servers. This article provides an overview of the validation tests that were completed.

Objectives

The key objectives for the validation testing of Tableau were to:

1. Configure Tableau to connect to HDP 2.6 running on an IBM POWER8 processor-based server.
2. Extract and visualize sample data from the Hadoop Distributed File System (HDFS) of HDP running on a POWER8 processor-based server.

Test environment

This section lists the high-level components used in the test environment.

Tableau

- Tableau Desktop 10.1 for Microsoft Windows 7

- Hortonworks ODBC Driver for Apache Hive v2.1.5
- A notebook running Windows 7

Hortonworks Data Platform

- HDP version 2.6
- Red Hat Enterprise Linux version 7.2
- Minimum resources: Eight virtual processors, 24 GB memory, 50 GB disk space
- IBM PowerKVM™
- IBM POWER8 processor-based server

Deployment architecture

The deployment architecture is quite simple. Tableau and the Hortonworks ODBC driver were installed and run on a Windows 7 system. HDP was installed and run on a POWER8 server. Tableau and the ODBC driver were configured to connect to HDP. Data in HDP was accessed and visualized by Tableau Desktop. Tests were run in a single-node HDP environment.

Installation and configuration

The section covers installation and configuration of a HDP cluster and vStorm software.

Installing and configuring the HDP cluster

Here are the high-level steps to install and configure the HDP cluster:

1. Follow the installation guide for HDP on Power Systems (see Resources) to install and configure the HDP cluster.
2. Log in to the Ambari server and ensure that all the services are running.
3. Monitor and manage the HDP cluster, Hadoop, and related services through Ambari.

Setting up test data and Hive tables

Download the MovieLens and driver test data, copy the data to HDFS, and create Hive tables.

1. Download the MovieLens data set from here (see the citation in Resources)
2. Follow the instructions here to copy the MovieLens dataset data to HDFS and set up Hive external tables. Use *hive* user ID for the same.
3. Download the driver data file from the Driver Behavior data file from here.
4. Copy the driver data to HDFS.

```
# su – hive
# hadoop fs -mkdir -p /user/hive/dataset/drivers
# hadoop fs -copyFromLocal /home/np/u0014213/Data/truck_event_text_partition.csv /user/hive/dataset/
drivers
# hadoop fs -copyFromLocal /home/np/u0014213/Data/drivers.csv /user/hive/dataset/drivers
# hadoop fs -ls /user/hive/dataset/drivers
Found 2 items
-rw-r--r--  3 hive hdfs     2043 2017-05-21 06:30 /user/hive/dataset/drivers/drivers.csv
-rw-r--r--  3 hive hdfs  2272077 2017-05-21 06:30 /user/hive/dataset/drivers/
truck_event_text_partition.csv
```

5. Create Hive tables for driver data.

```
# su – hive
# hive
hive>create database trucks;
hive> use trucks;

hive> create table drivers
(driverId int,
name string,
ssn bigint,
location string,
certified string,
wageplan string)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
TBLPROPERTIES("skip.header.line.count"="1");

hive> create table truck_events
(driverId int,
truckId int,
eventTime string,
eventType string,
longitude double,
latitude double,
eventKey string,
correlationId bigint,
driverName string,
routeId int,
routeName string)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE
TBLPROPERTIES("skip.header.line.count"="1");

hive> show tables;
OK
drivers
truck_events
```

6. Load the data into the tables from the files in HDFS.

```
hive> LOAD DATA INPATH '/user/hive/dataset/drivers/truck_event_text_partition.csv' overwrite
    into table truck_events;
hive> LOAD DATA INPATH '/user/hive/dataset/drivers/drivers.csv' overwrite into table drivers;
```

7. Cross check the tables to ensure that the data is present by running queries on the tables.

Installing and configuring the Hortonworks ODBC driver

Here are the steps to install and configure the ODBC driver:

1. Download the Hortonworks ODBC driver on Windows 7 (see Resources for the download website).
2. Install and configure the ODBC driver. Follow the instructions in the guide listed in the Resources section.

Installing and configuring Tableau

Here are the steps to install and configure Tableau:

1. Go to the Tableau download page (see Resources) to download Tableau Desktop on Windows 7.

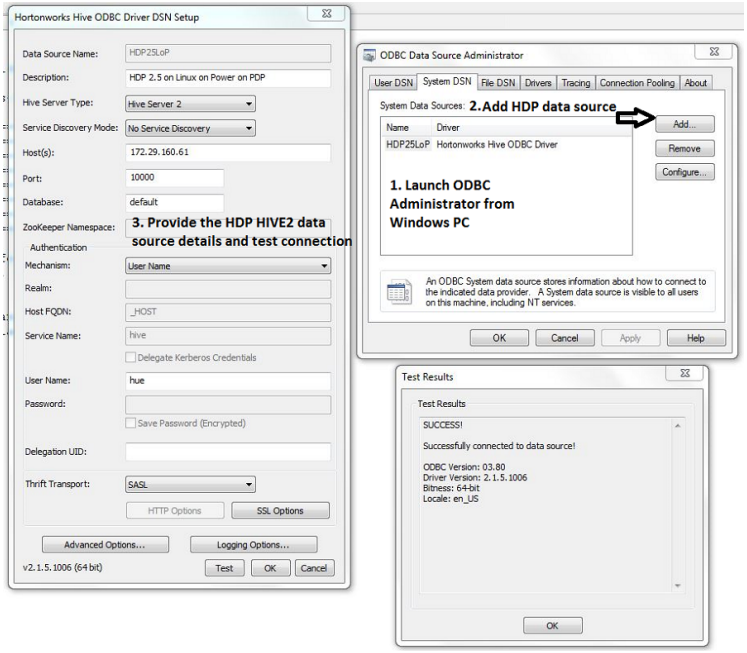
2. Follow the prompts to install it in the Windows 7 system.

Connecting HDP to Tableau

Here are the steps to configure the connection between HDP and Tableau.

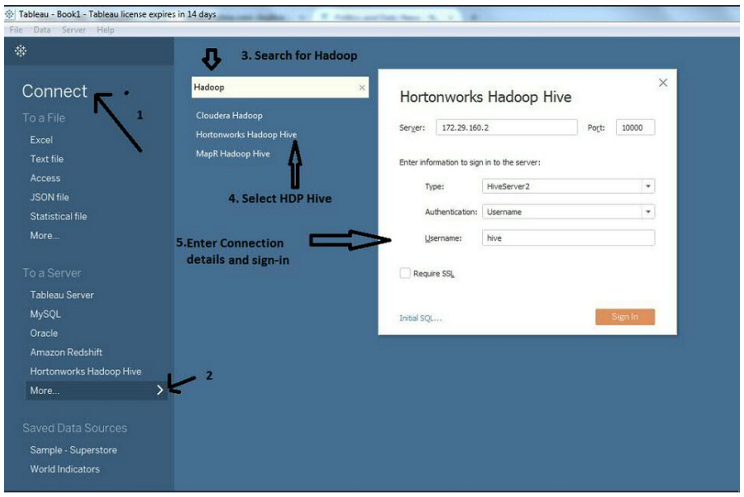
1. Launch the ODBC Administrator from the Windows system and add a data source for Hortonworks Hive as shown in Figure 1.

Figure 1. Hortonworks Hive ODBC driver setup



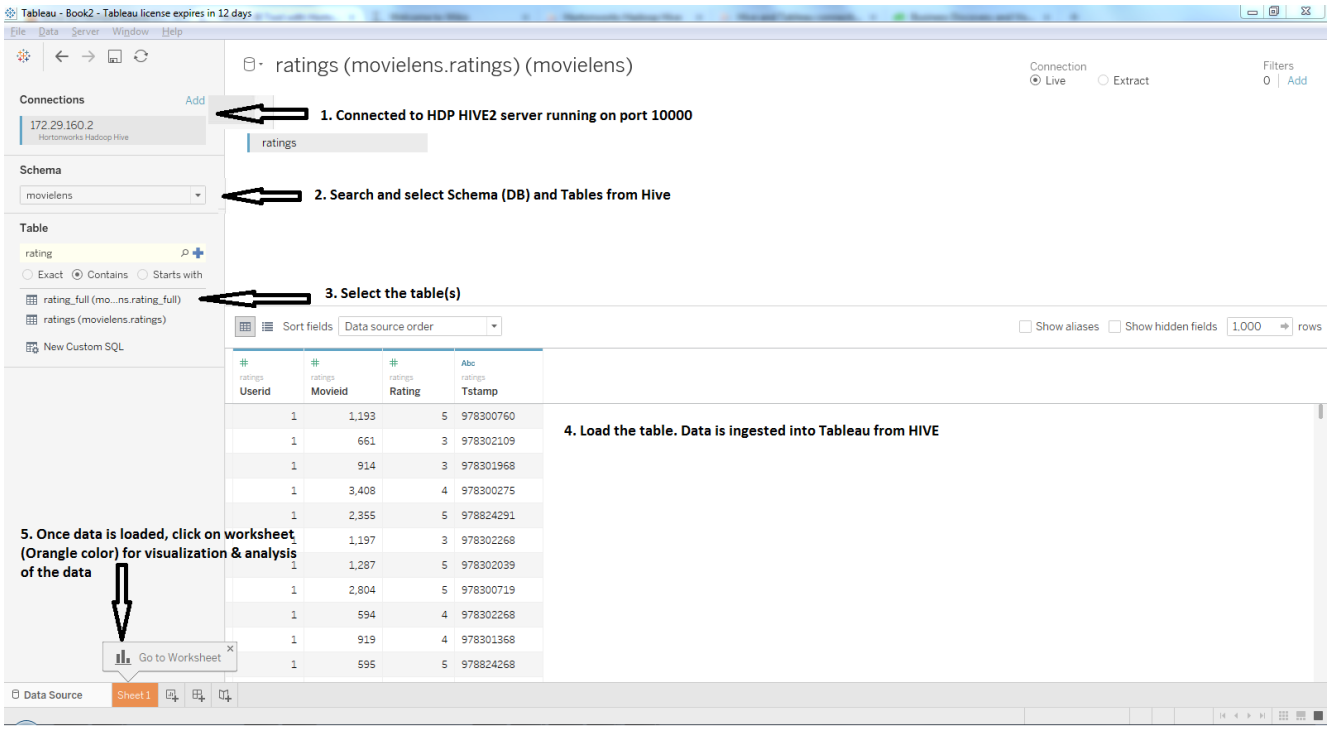
2. In Windows 7, launch Tableau Desktop and configure the connection to HDP as shown in Figure 2.

Figure 2. Main Tableau screen



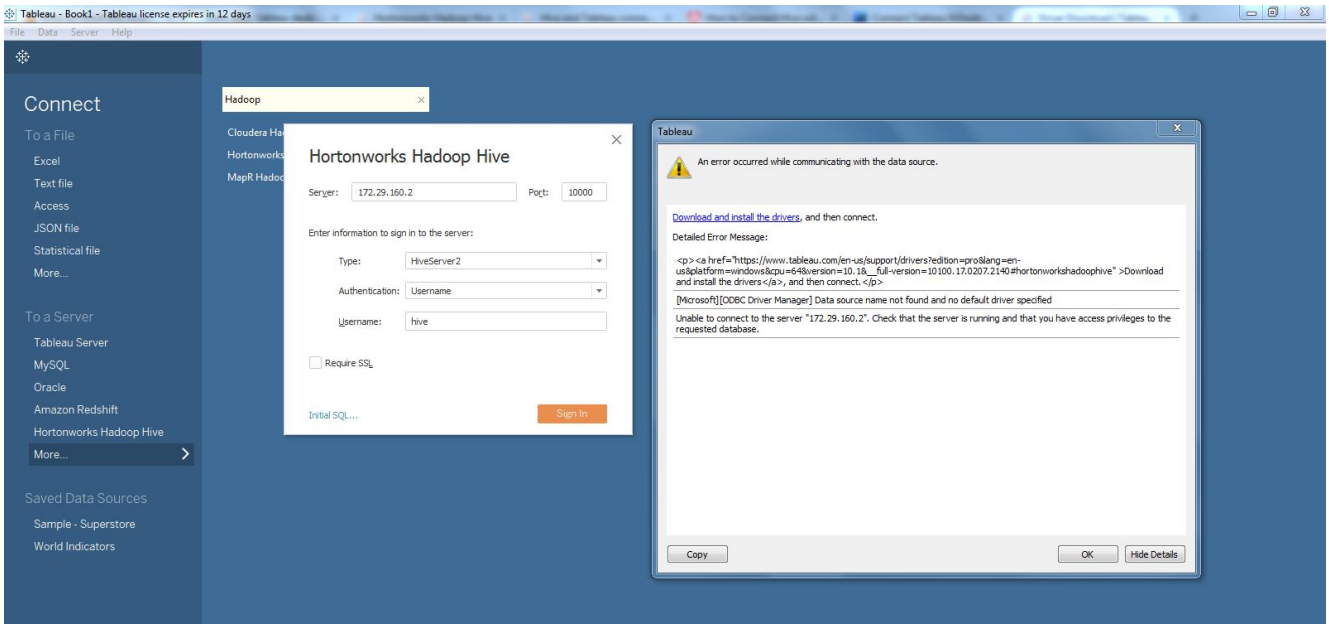
3. Connect to the HIVE2 server running on HDP 2.6 instance running on the IBM Power8-based server as shown in Figure 3. Select the schema (DB) and tables from Hive. Load the data so the data is ingested into Tableau from Hive. Now you are ready to start analyzing.

Figure 3. Connecting to HDP



Note: If the ODBC driver for Hortonworks Hive is not installed, you will get the error shown in Figure 4 while connecting to HDP.

Figure 4. Example of Tableau communication error



Visualization and analysis in Tableau

Using the Tableau Desktop, select the columns of data for visualization and analysis. Figures 5-9 show examples of analysis and visualization that were tested.

Figure 5. Tableau visualization example 1

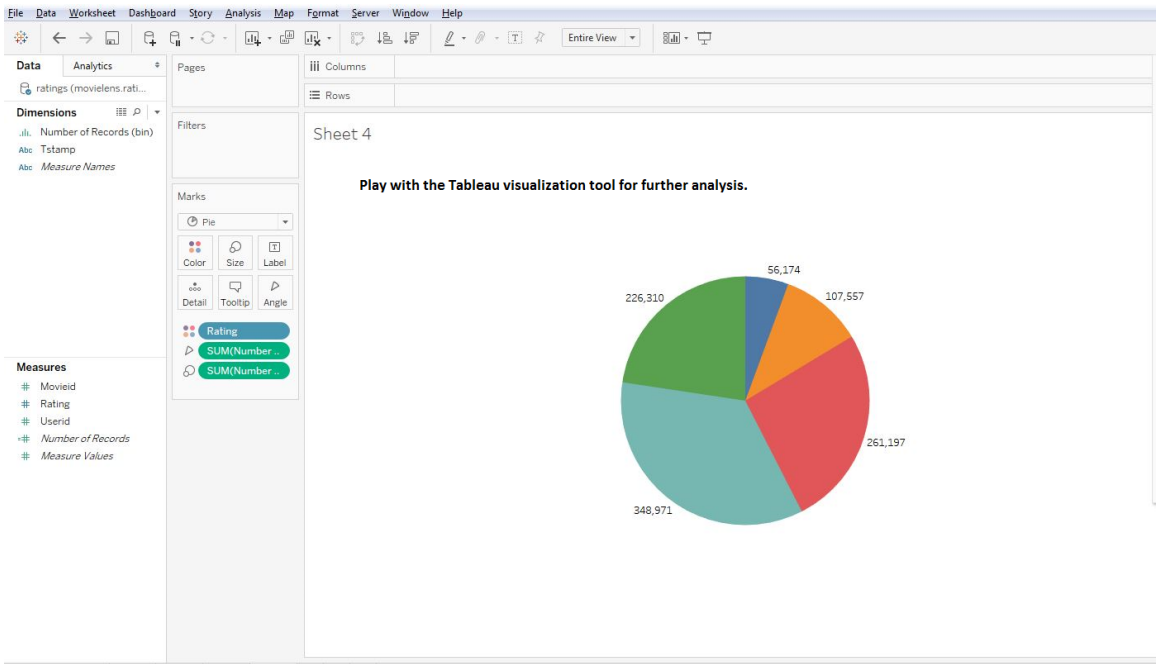


Figure 6. Tableau visualization example 2

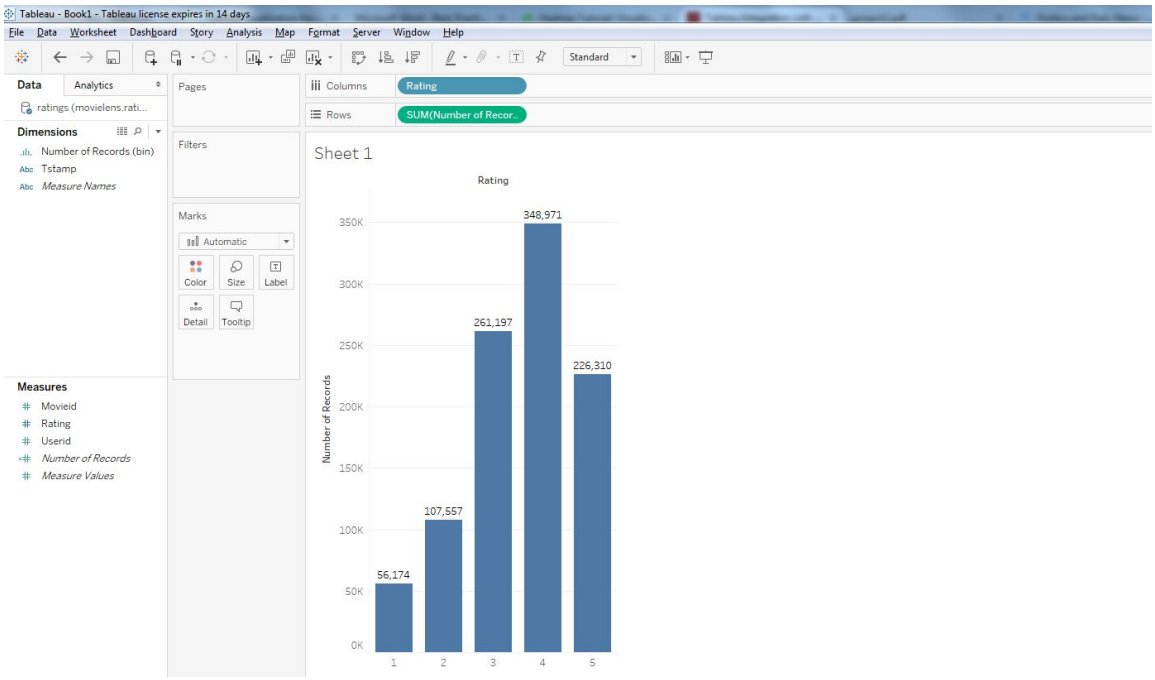


Figure 7. Tableau visualization example 3

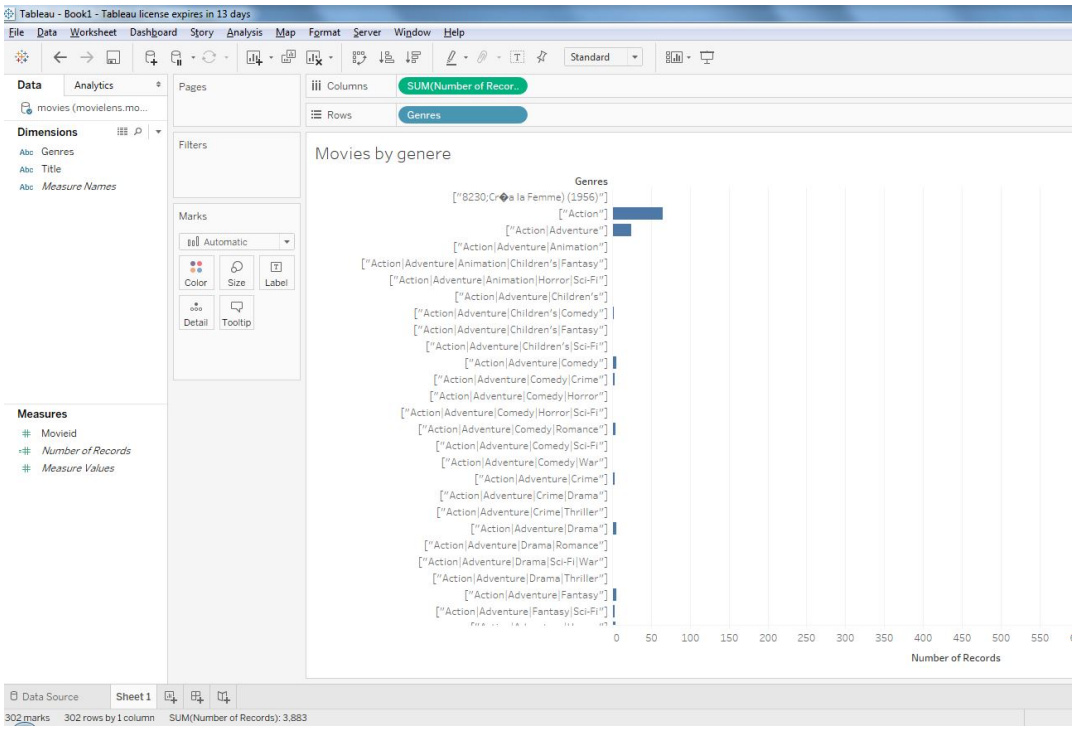
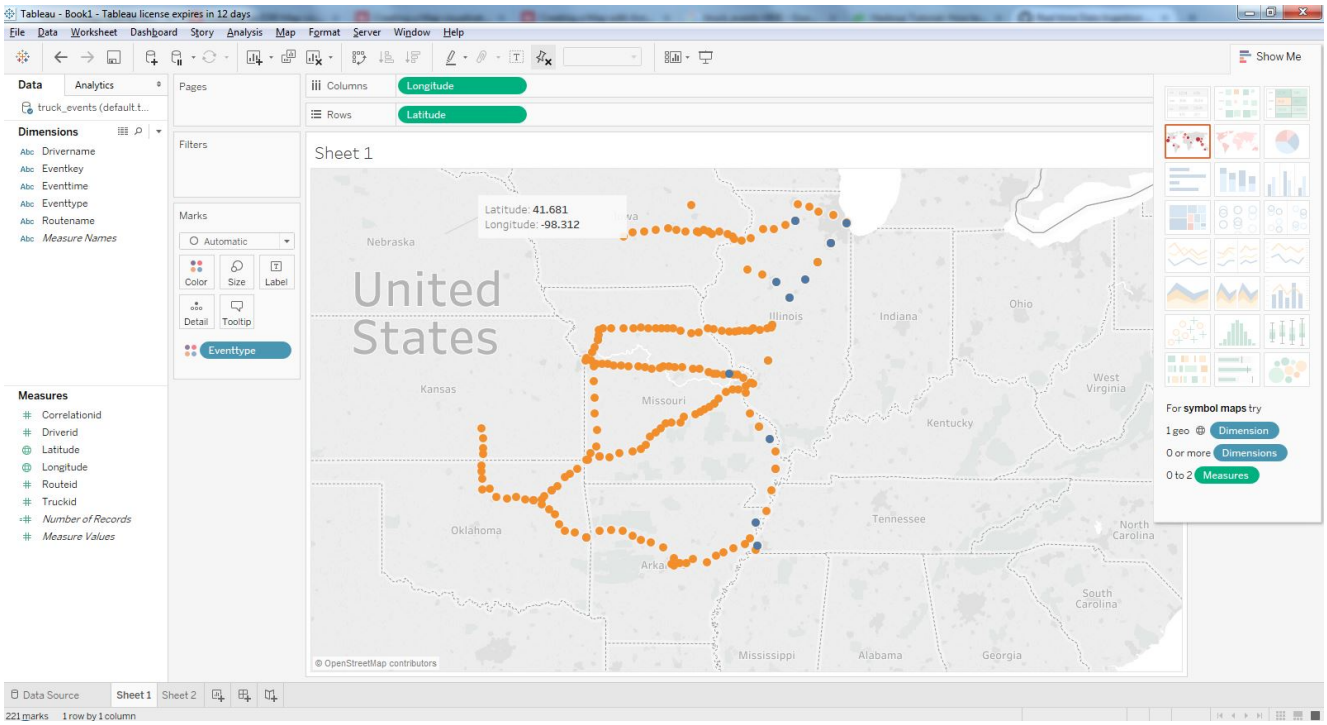


Figure 8. Tableau visualization example 4



Resources

- Hortonworks Data Platform: Apache Ambari Installation for IBM Power Systems
- Hortonworks ODBC Driver for Apache Hive v2.1.5 download web page
- Hortonworks ODBC installation and configuration guide
- Tableau website
- Tableau Desktop download web page
- Driver Behavior database
- ISV solution ecosystem for Hortonworks on IBM Power Systems
- MovieLens dataset
- MovieLens data set citation:

  ```
  F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages. DOI=http://dx.doi.org/10.1145/2827872
  ```