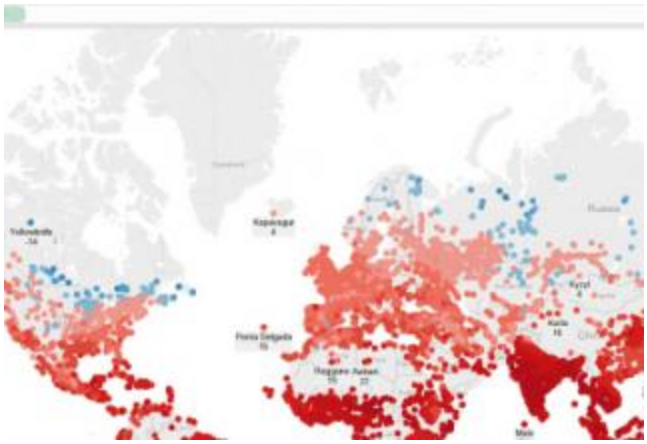
Apache Drill: It's drilliant to query JSON files from Tableau Desktop



June 19, 2015

Uli Bethke (/blog/author/uli-bethke)

Did you know you can run Apache Drill (http://drill.apache.org/) on your laptop? This is great news for business analysts who need to explore complex and semi-structured data. Let's look at a particular example.

A company has implemented a new SaaS based system. This system makes data extracts available over a RESTful API in JSON format. Before the data is loaded and standardised in the corporate data warehouse a business analyst gets tasked with exploring this new data set and a data extract is made available for analysis. A great tool for exploratory data analysis (EDA) is Tableau. Our business analyst immediately gets to work only to realise that querying JSON from

Tableau (http://community.tableau.com/thread/147566 (http://community.tableau.com/thread/147566)) is not straight forward.

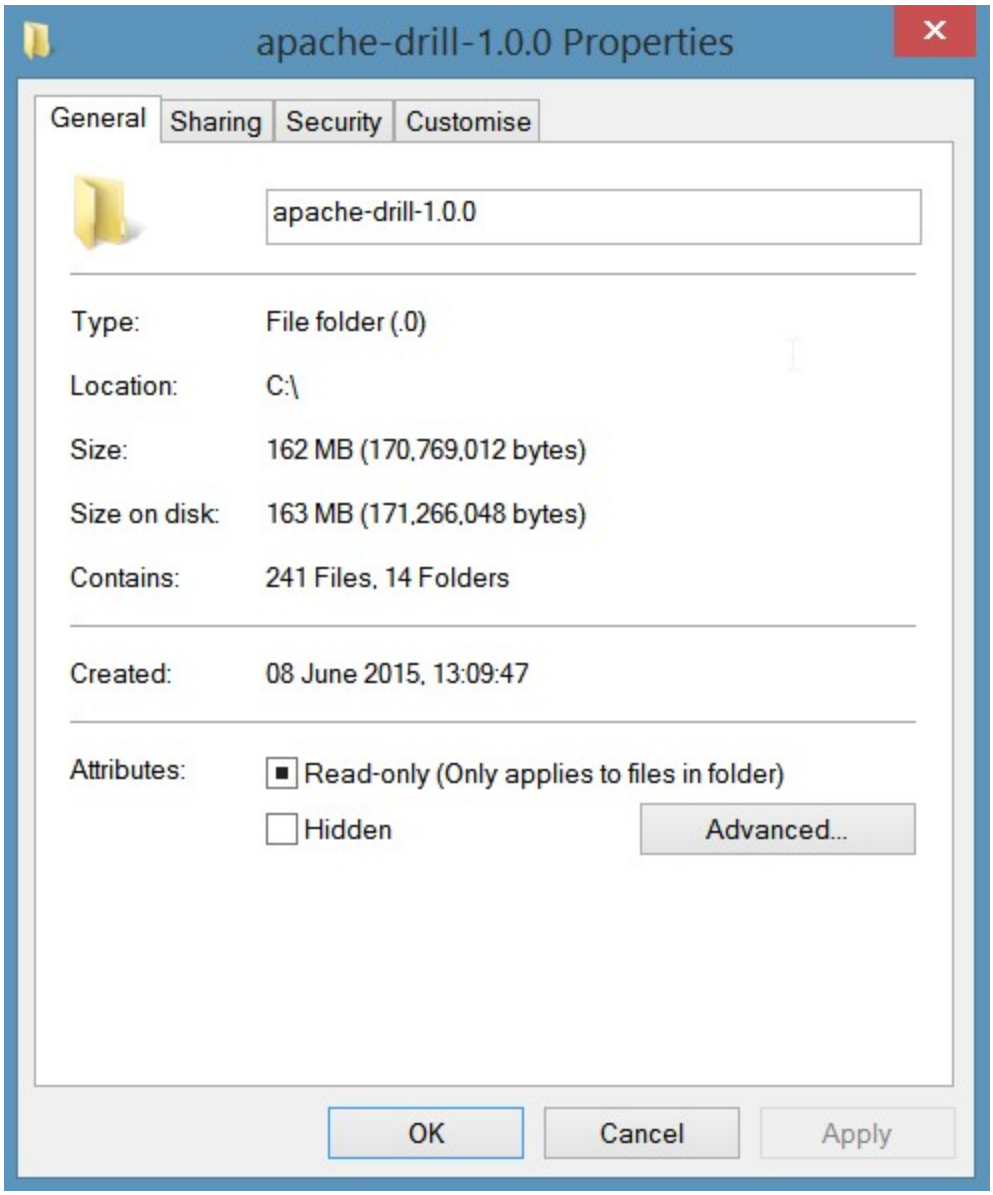
Drill to the Rescue

Our business analyst wonders what to do next. Should she involve the guys from IT to transform the data to something more easily digestible? This could take a week or more. Time is of the essence. There must be another way. After some more googling for a solution she comes across Drill. With Drill she can query JSON data using SQL (https://www.mapr.com/why-hadoop/sql-hadoop/sql-hadoop-details), a skill she is deeply familiar with. Drill also ships an ODBC driver, which allows her to connect with Tableau.

She downloads Drill to her Windows laptop http://drill.apache.org/docs/installing-drill-on-windows/ (http://drill.apache.org/docs/installing-drill-on-windows/).

She checks that she has the Oracle 7 JDK installed.

She then proceeds to install Drill in embedded mode http://drill.apache.org/docs/installing-drill-on-windows/ (http://drill.apache.org/docs/installing-drill-on-windows/). She picks the root of her C:\ drive as the install destination.



Next she double checks that she has set the JAVA_HOME environment variable correctly.

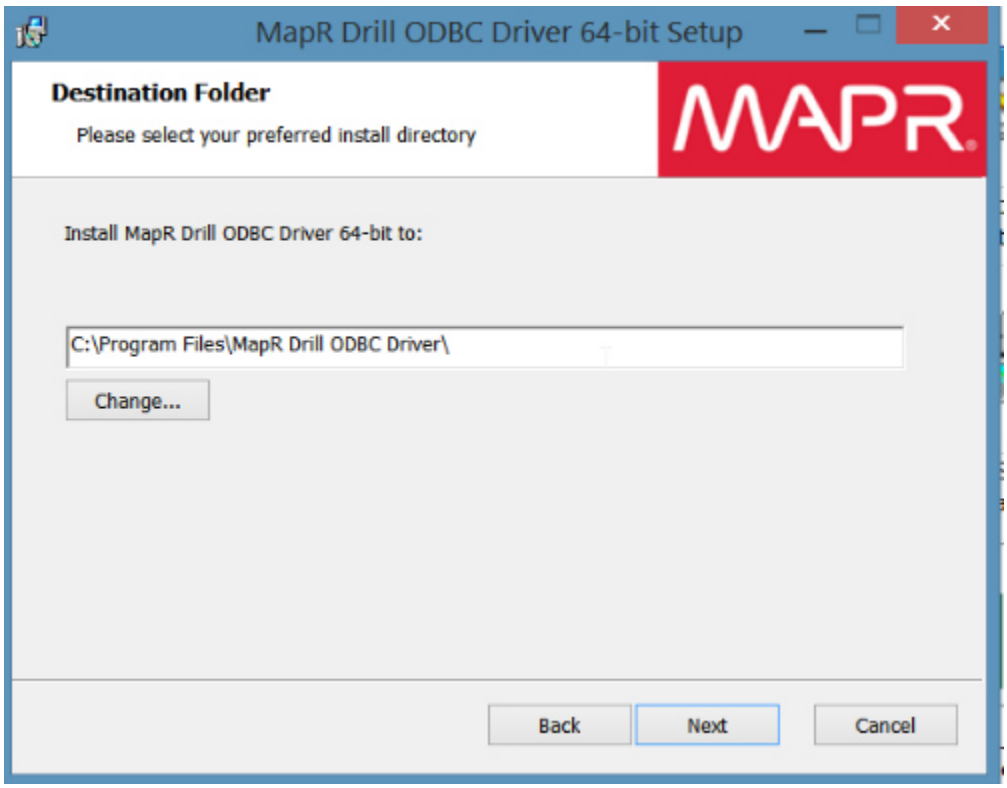
She starts Drill http://drill.apache.org/docs/starting-drill-on-windows/.
(http://drill.apache.org/docs/starting-drill-on-windows/)

This concludes the installation of Drill.

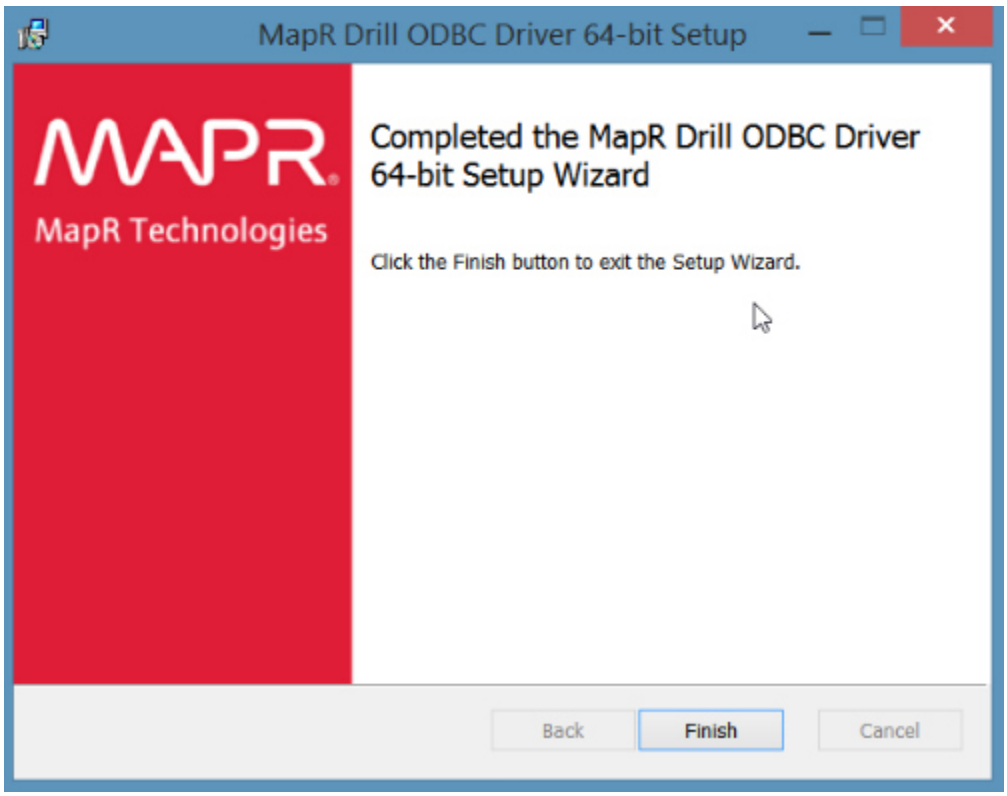
Next she downloads the Drill ODBC driver. (http://package.mapr.com/tools/MapR-ODBC/MapR_Drill/MapRDrill_odbc/ (http://package.mapr.com/tools/MapR-ODBC/MapR_Drill/MapRDrill_odbc/)).

Note: Always make sure that the version of Drill corresponds to the version of the ODBC driver.

She selects the 64 Bit driver as she also runs the 64 Bit version of Tableau. Once the driver has finished downloading she launches the installer.



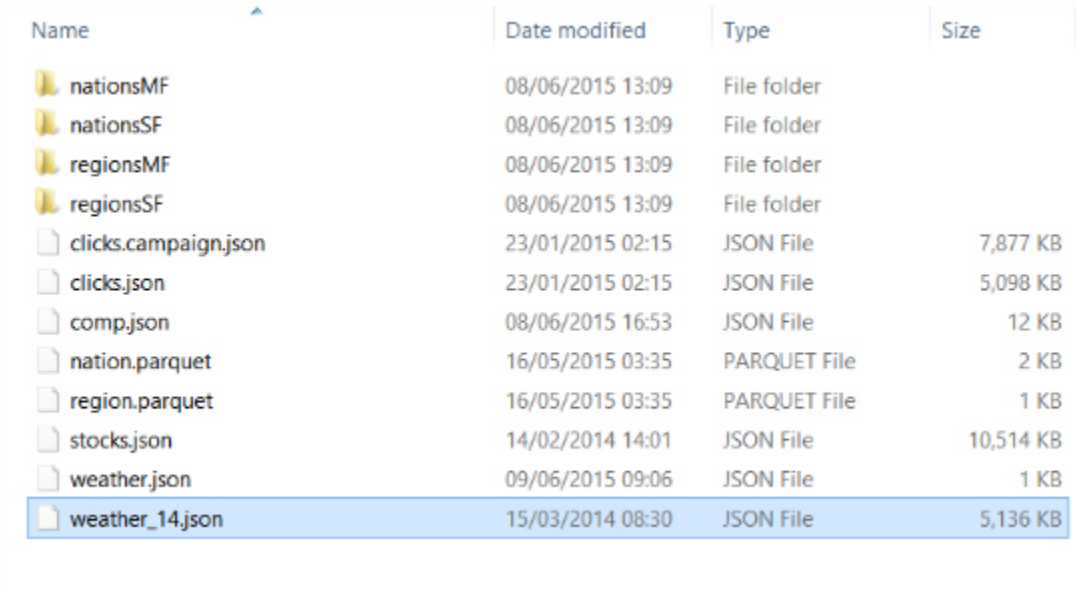
… and waits for the install to complete.



Exploratory Data Analysis with Tableau

The business analyst is now ready to analyse the data dump. She downloads the weather data from the Open Weather http://openweathermap.org/current (http://openweathermap.org/current) website. The JSON dataset we are interested in is the file http://78.46.48.103/sample/weather_14.json.gz (http://78.46.48.103/sample/weather_14.json.gz) that contains the current weather of 20,000 cities (updated hourly).

She extracts JSON file and copies it into her sample data folder in the Drill install folder C:\apache-drill-1.0.0\sample-data.



Name	Date modified	Type	Size
nationsMF	08/06/2015 13:09	File folder	
nationsSF	08/06/2015 13:09	File folder	
regionsMF	08/06/2015 13:09	File folder	
regionsSF	08/06/2015 13:09	File folder	
clicks.campaign.json	23/01/2015 02:15	JSON File	7,877 KB
clicks.json	23/01/2015 02:15	JSON File	5,098 KB
comp.json	08/06/2015 16:53	JSON File	12 KB
nation.parquet	16/05/2015 03:35	PARQUET File	2 KB
region.parquet	16/05/2015 03:35	PARQUET File	1 KB
stocks.json	14/02/2014 14:01	JSON File	10,514 KB
weather.json	09/06/2015 09:06	JSON File	1 KB
weather_14.json	15/03/2014 08:30	JSON File	5,136 KB

She opens the file in a text editor and copies one record into a JSON formatter.

http://jsonformatter.curiousconcept.com/ (http://jsonformatter.curiousconcept.com/) to get a better understanding of how the data is structured hierarchically.


```
{
  "city":{
    "id":2267057,
    "name":"Lisbon",
    "findname":"LISBON",
    "country":"PT",
```

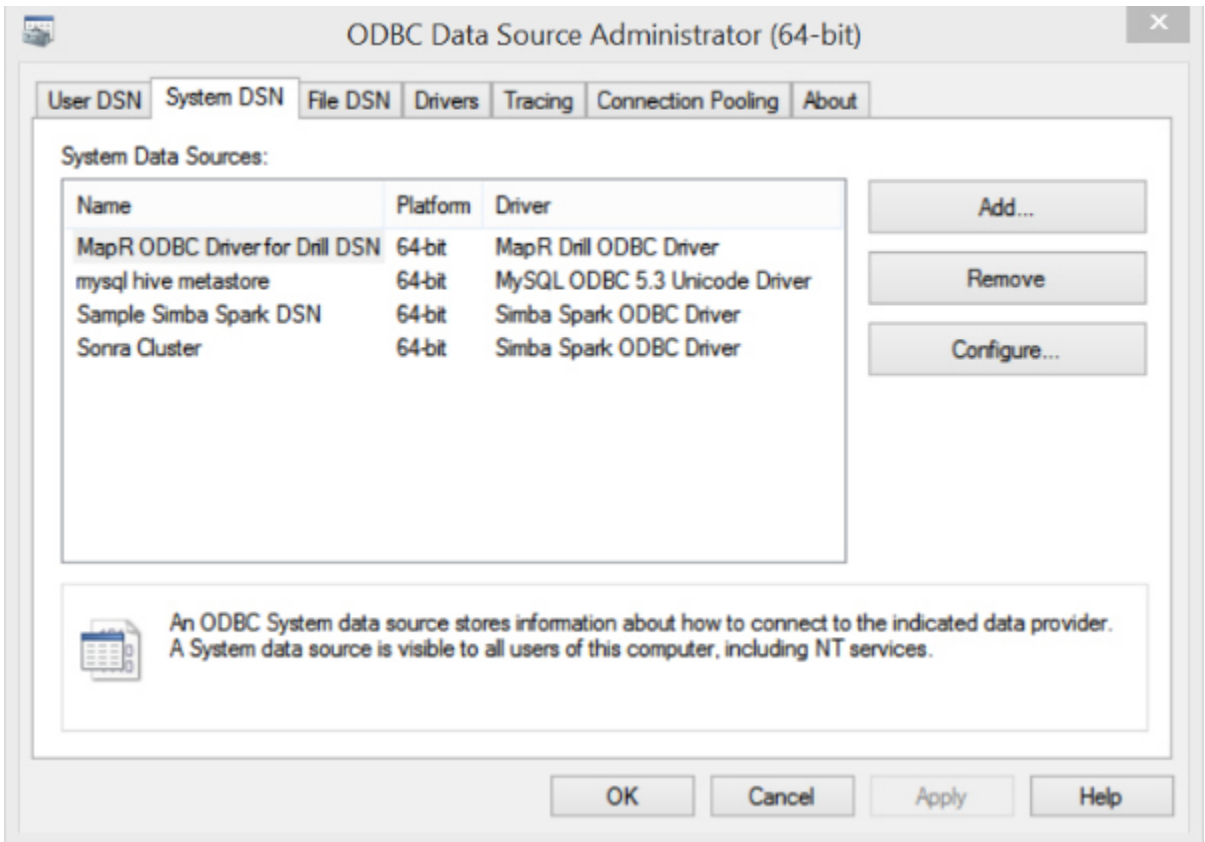
```
  "coord":{
    "lon":-9.13333,
    "lat":38.716671
  },
  "zoom":7
},
"time":1394871602,
"main":{
  "temp":281.29,
  "humidity":82,
  "pressure":1021,
  "temp_min":280.15,
  "temp_max":282.59
},
"wind":{
  "speed":4.85,
  "deg":6.50397
},
"clouds":{
  "all":0
},
"weather":[
  {
    "id":741,
    "main":"Fog",
    "description":"fog",
    "icon":"50d"
  },
  {
    "id":701,
    "main":"Mist",
    "description":"mist",
```

```
        "icon":"50d"
      }
    ]
  }
```


We can see that the JSON document is split into various sections: city, time, main, wind, clouds, weather. We can also see that weather is modelled as an array in this JSON dataset. Weather is multivalued. Each city weather record may contain one or more descriptions. This discovery will become important later on, when we write queries against the data.

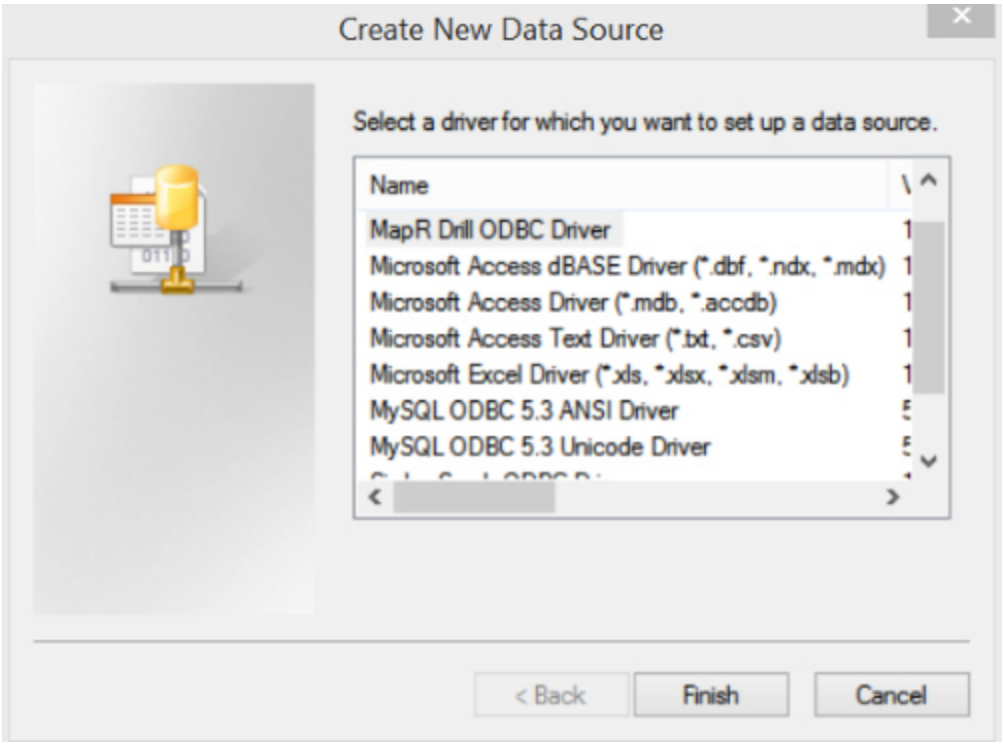
Next she wants to familiarise herself further with the data. Drill Explorer, a tool to visually explore Drill data ships with the Drill ODBC driver.

We can access Drill Explorer via the 64 bit ODBC Administrator in Windows. You can access the 64 bit ODBC driver via C:\WINDOWS\SysWOW64\odbcad32.exe.

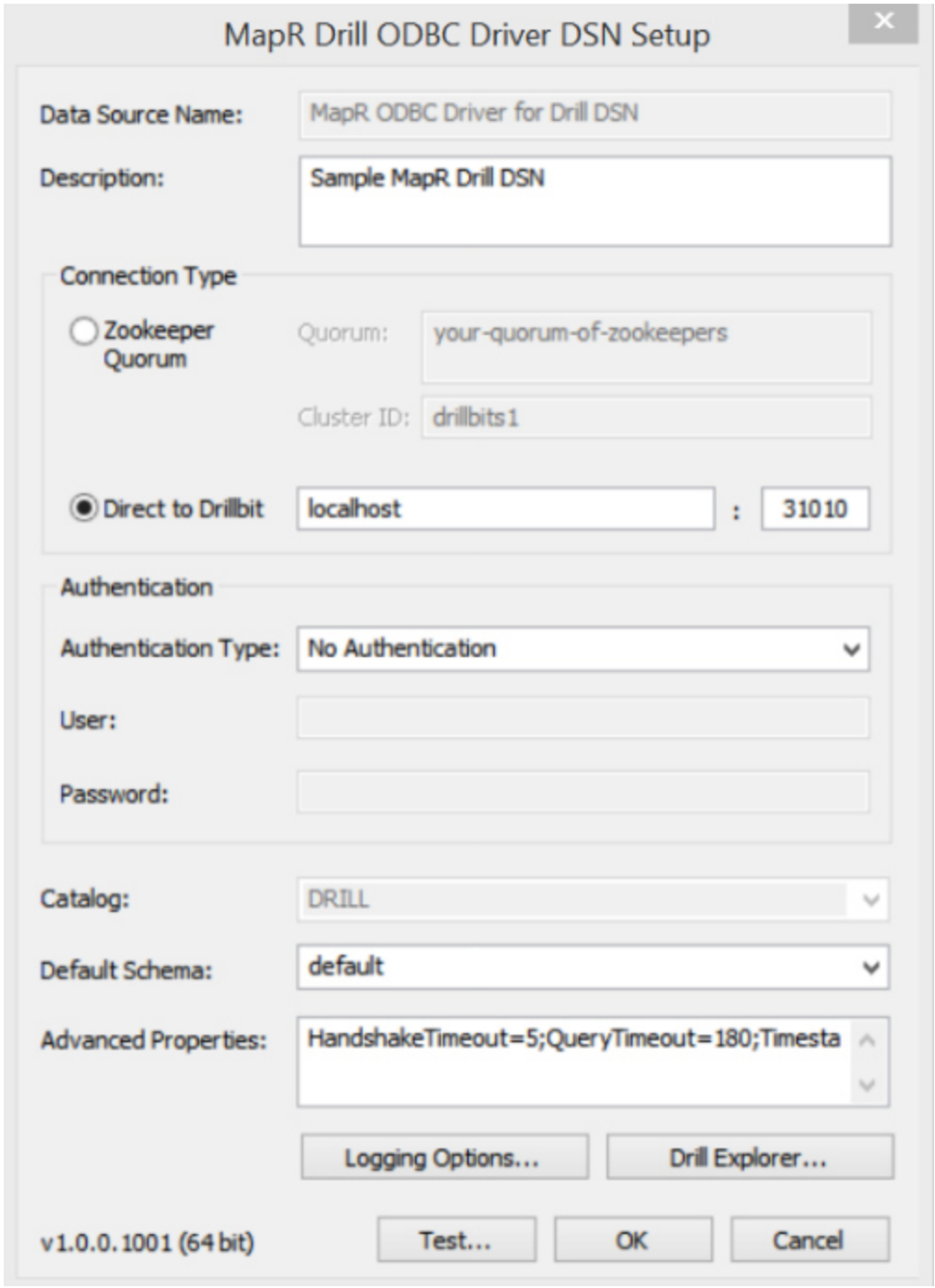


Our business analyst moves to the System DSN tab and there clicks the Add… button.

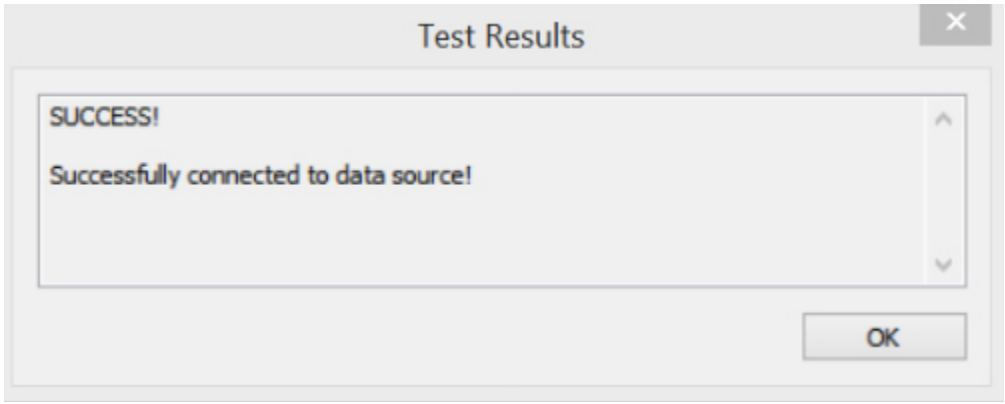
Next she selects the MapR Drill ODBC driver:



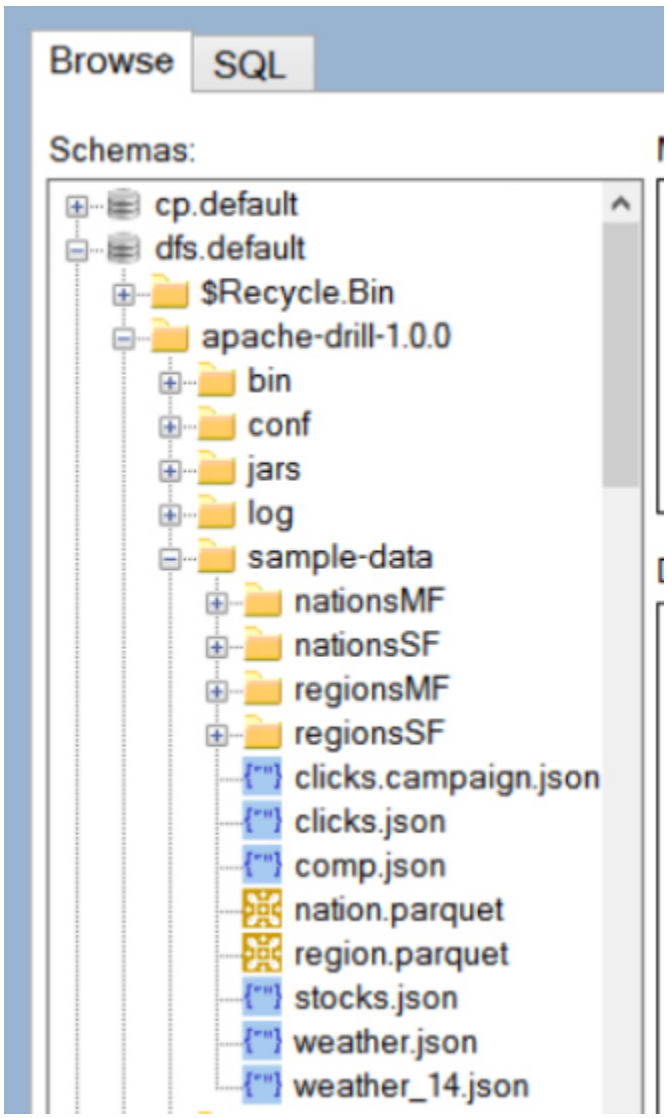
She gives the new data source a name and…



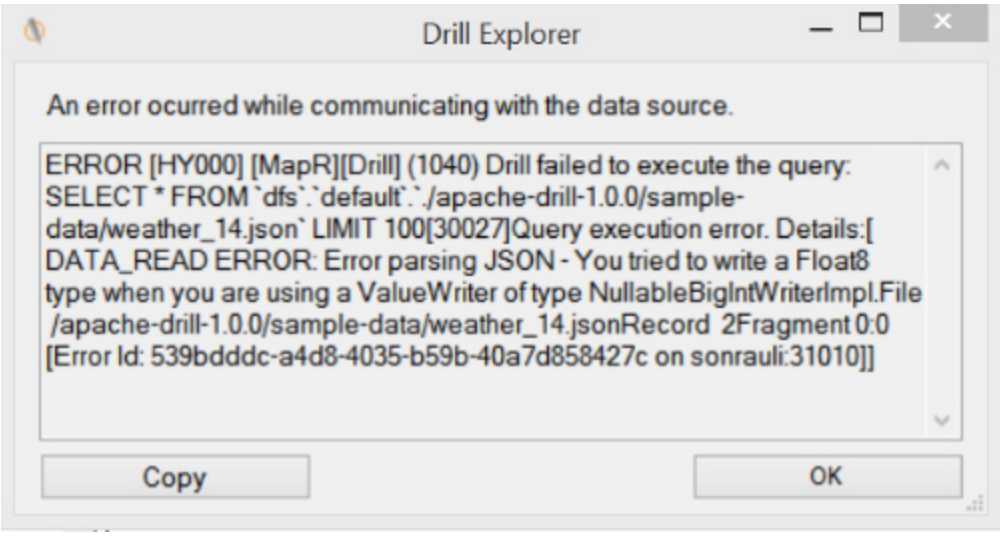
…tests the connection by clicking the Test button:



She is now ready to launch Drill Explorer by clicking the Drill Explorer… button. She navigates to the JSON weather file in the sample data folder and then double clicks the weather_14.json file.



This throws an error:



At this stage our business analyst consults the documentation https://drill.apache.org/docs/json-data-model/ (https://drill.apache.org/docs/json-data-model/) and finds the solution to fix this problem.

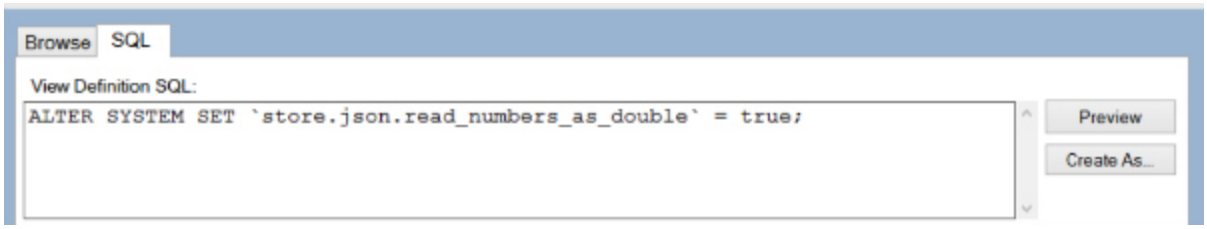
“By default, Drill does not support JSON lists of different types. For example, JSON does not enforce types or distinguish between integers and floating point values. When reading numerical values from a JSON file, Drill distinguishes integers from floating point numbers by the presence or lack of a decimal point. If some numbers in a JSON map or array appear with and without a decimal point, such as 0 and 0.0, Drill throws a schema change error. “

In the weather data set we have exactly this scenario. The field Pressure can be with or without decimal point.

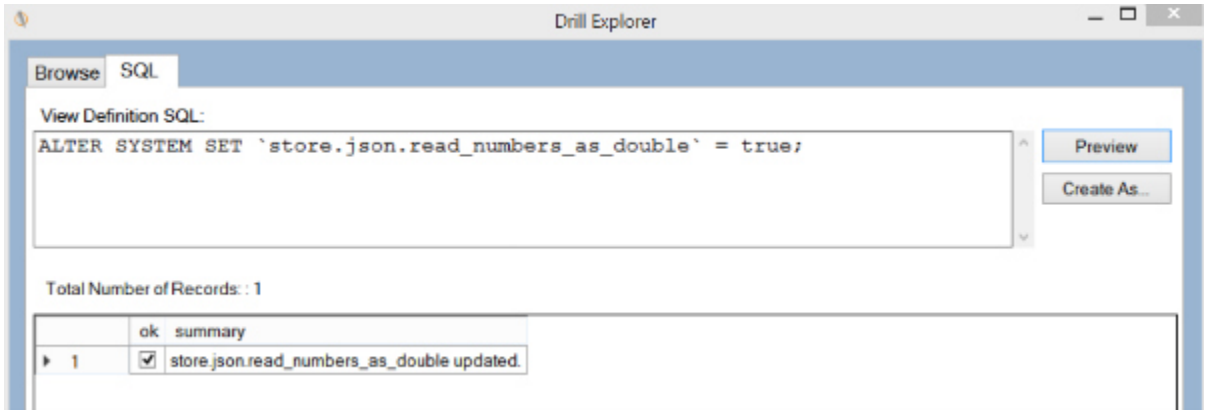
```
{"temp":297.15,"pressure":1020,"humidity":33,"temp_min":297.15,
":288.64,"temp_min":288.64,"temp_max":288.64,"pressure":835.59,
```

The solution is to set the `store.json.read_numbers_as_double` property to true. In the Drill Explorer the BA switches to the SQL tab and issues the following command:

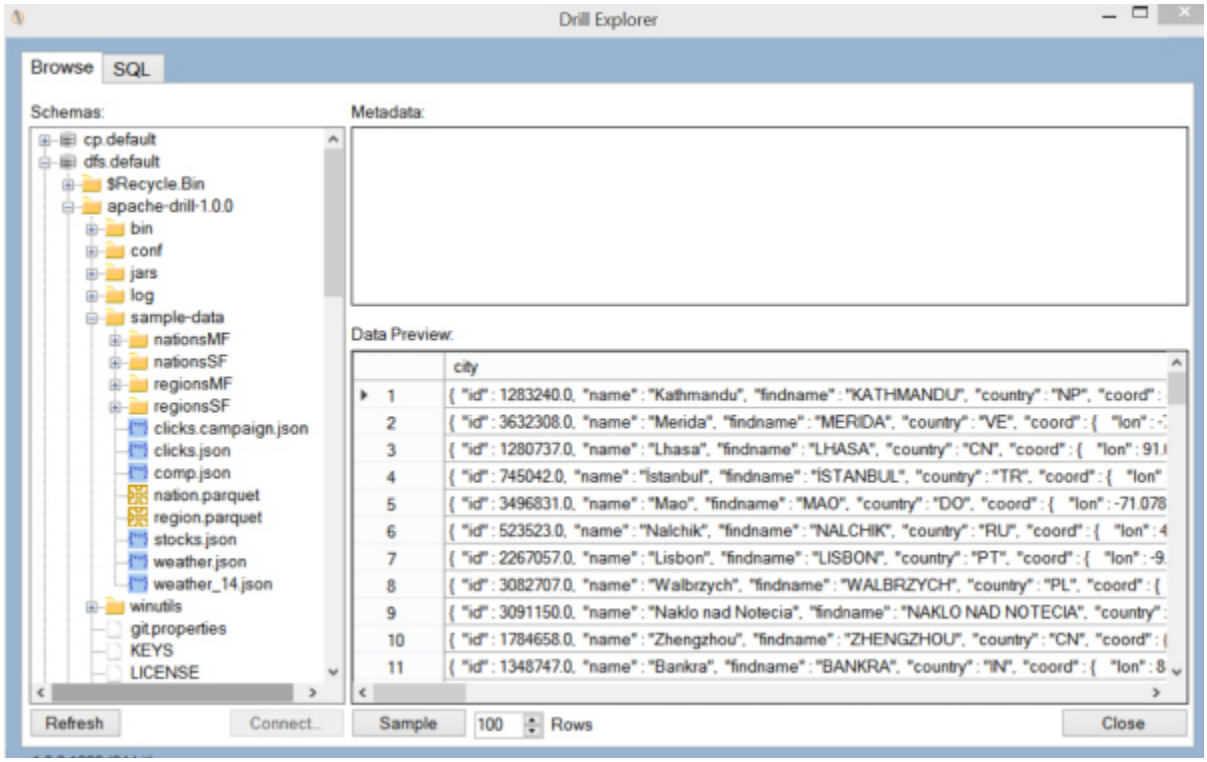
```
ALTER SYSTEM SET `store.json.read_numbers_as_double` = true;
```



She confirms by clicking Preview:



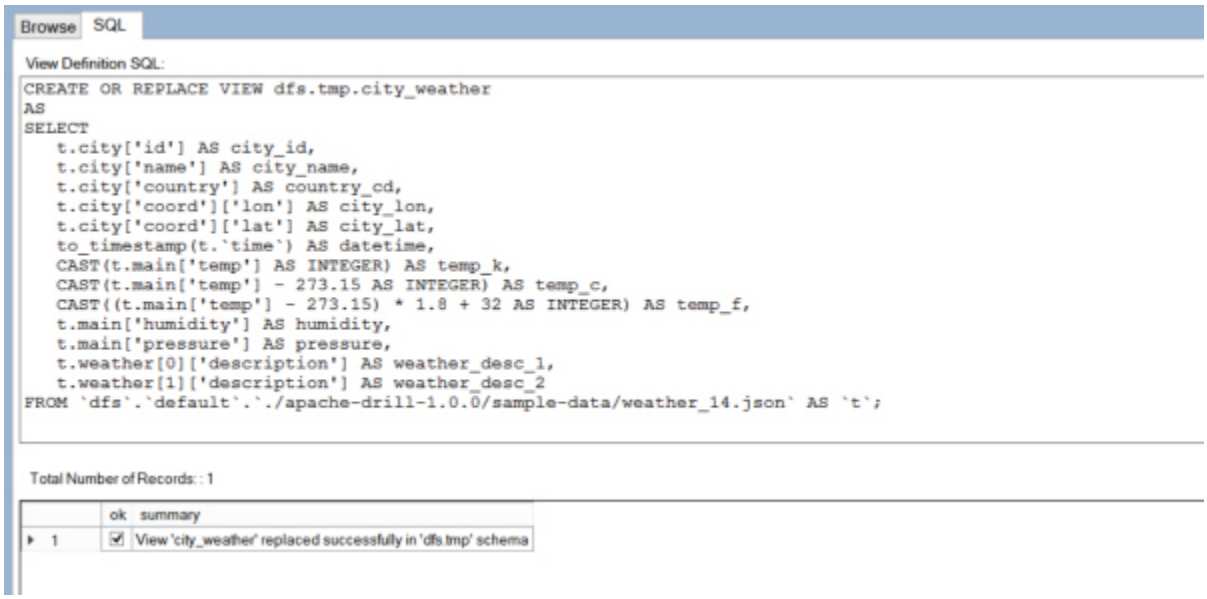
Next she switches back to the Browse tab and double clicks the weather14.json file et voilá



The business analyst is now ready to further explore this data in Tableau. For her analysis she needs the following data points:

- Country Code
- City Name
- Geo Coordinates
- Time
- Temperature in Kelvin, Celcius, and Fahrenheit
- Humidity
- Pressure
- Weather Description

She creates a View over her dataset that will then be exposed to Tableau:



CREATE OR REPLACE VIEW dfs.tmp.city_weather

AS

SELECT

t.city['id'] AS city_id,

t.city['name'] AS city_name,

t.city['country'] AS country_cd,

```
t.city['coord']['lon'] AS city_lon,

t.city['coord']['lat'] AS city_lat,

to_timestamp(t.`time`) AS datetime,

CAST(t.main['temp'] AS INTEGER) AS temp_k,

CAST(t.main['temp'] - 273.15 AS INTEGER) AS temp_c,

CAST((t.main['temp'] - 273.15) * 1.8 + 32 AS INTEGER) AS temp_f,

t.main['humidity'] AS humidity,

t.main['pressure'] AS pressure,

t.weather[0]['description'] AS weather_desc_1,

t.weather[1]['description'] AS weather_desc_2

FROM `dfs`.`default`.`./apache-drill-1.0.0/sample-data/weather_14.json` AS
`t`;
```

There are a couple of things worthwhile noting:

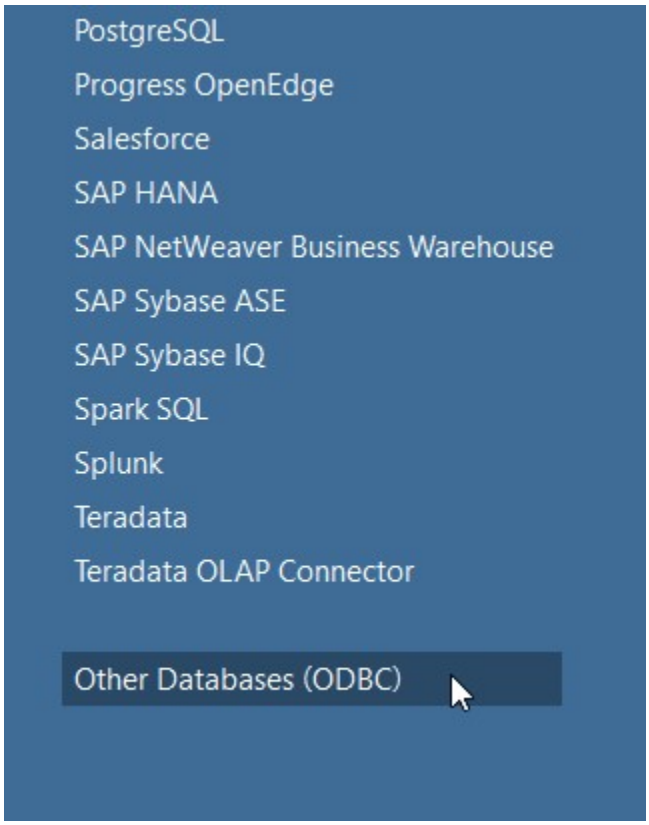
- When we create a view in Drill we need to create it in a workspace that is writable. By default that is the tmp workspace, which is already created. You can set up workspaces via the Web UI under the Storage plugin registration tab. You access the Web UI via http://<IP address>:8047/storage. As we run Drill in Embedded mode this would be http://localhost:8047/storage (http://localhost:8047/storage). For more details refer to the storage plugin configuration section http://drill.apache.org/docs/plugin-configuration-basics/ (http://drill.apache.org/docs/plugin-configuration-basics/) in the documentation.
- When drilling into the hierarchy of the JSON dataset we need to create an alias for the datastore. In our case this is the letter t. When referencing attributes in the view we need to prefix the column hierarchy with this alias.

- The Drill function to_timestamp converts the UNIX timestamp to a DATE
- In Drill the keyword time is a reserved word. We need to enclose the time attribute in our JSON file with backticks t.`time` so that it can be interpreted correctly.
- We apply a couple of transformation to the temperature data to convert it from Kelvin to Celcius and Fahrenheit.
- As weather is an array in our JSON file it may hold multiple values. We can access each value by referencing the individual members in the array. In our case we want to retrieve the first two members in the array. For more detailed information refer to the documentation http://drill.apache.org/docs/selecting-nested-data-for-a-column/ (http://drill.apache.org/docs/selecting-nested-data-for-a-column/).

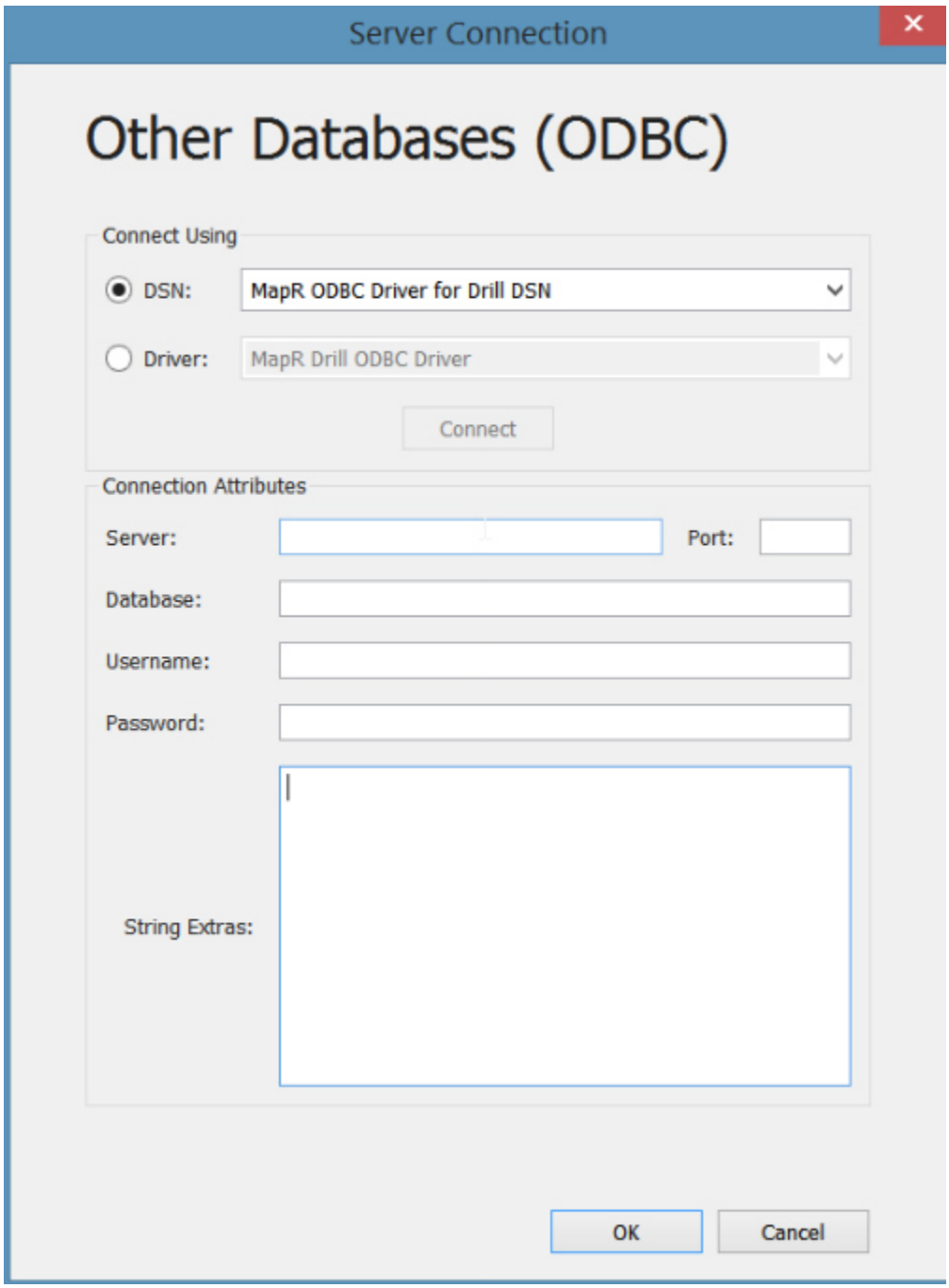
Drill and Tableau a match made in heaven

Our business analyst finally is ready to run queries against the JSON weather dataset in Tableau.

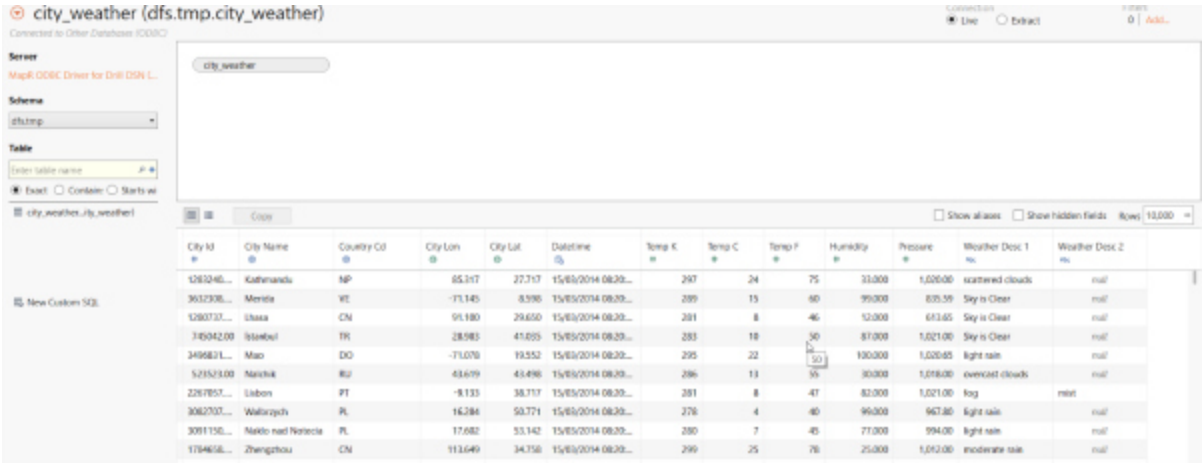
She launches Tableau and selects Other Databases to connect to Drill.



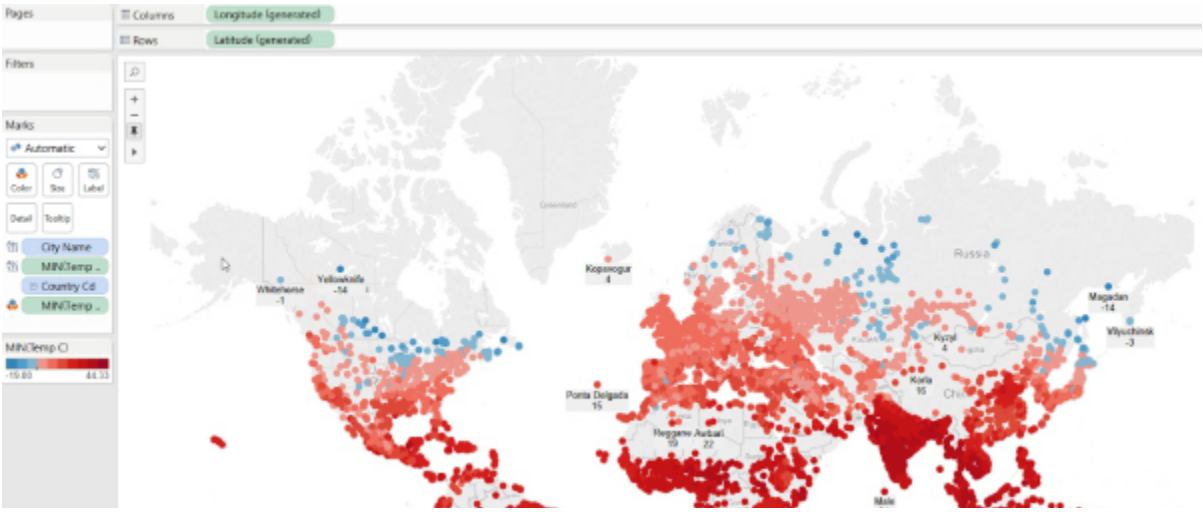
Next she selects the Data Source Name (DSN) that she created earlier on through 64 bit ODBC Administrator.



Next she selects the Schema dfs.tmp and drags the city_weather view across.



In a last step our business analysts creates a map with world temperatures…



…and explores a particular region in more detail:

What's Next

Of course, Drill is not only drilliant to query data on a single laptop. It can be deployed in a clustered environment to query large volumes of data at scale with low latency and high concurrency.

References:

- Drill documentation (https://drill.apache.org/docs/)
- MapR Blog (https://www.mapr.com/search/site/drill)
- More Details and Options to try Drill (https://www.mapr.com/products/apache-drill)

Delivering Fastest Time-to-Value for SQL-on-Hadoop

Read this paper to learn about: How Drill enables self-service data analysis, An example scenario - analyzing Twitter JSON data with Drill, How Drill compares to Hive and Impala.

(https://www.mapr.com/delivering-fastest-ti

148

BLOG SIGN UP

Sign up and get the top posts from each week delivered to your inbox every Friday!

* Email Address

Subscribe

(/blog/author/uli-bethke)

Uli Bethke (/blog/author/uli-bethke)

CO-FOUNDER, SONRA

Uli Bethke is the co-founder of Sonra (http://www.sonra.io). Sonra is a Big Data consulting company in Ireland and a partner company of MapR, the only enterprise-ready Hadoop distribution. Sonra provide services and accelerators for data warehouse offload and data lake implementations on MapR.

Uli is a data visionary and provides thought leadership in the architecture and implementation of data driven applications. He has led some of the largest and most complex data warehouse implementations in Europe.

Uli is the chair of the Hadoop User Group Ireland (http://www.meetup.com/hadoop-user-group-ireland/). You can get in touch with Uli by connecting with him on

me-value-sql-hadoop?source=Social&campaign=2015_Social_Blog)

LinkedIn (https://www.linkedin.com/in/ulibethke) or Twitter (https://twitter.com/ubethke).

FOLLOW MAPR

Follow @mapr 38K followers

Dev Ops Hub RSS (https://www.mapr.com/devops.xml)

Big Data Hub RSS (https://www.mapr.com/bigdata.xml)

STREAMING DATA ARCHITECTURE:
New Designs Using Apache Kafka and MapR Streams

(/streaming-architecture-using-apache-kafka-mapr-streams?source=Soci