Search

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 <u>Apache Drill (http://drill.apache.org/)</u> 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 <u>SQL (https://www.mapr.com/why-hadoop/sql-hadoop-details)</u>, 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 <u>http://drill.apache.org/docs/installing-drill-on-windows/</u> (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 <u>http://drill.apache.org/docs/installing-drill-on-</u> <u>windows/ (http://drill.apache.org/docs/installing-drill-on-windows/)</u>. She picks the root of her C:\ drive as the install destination.

1	apache-drill-1.0.0 Prop	perties	x
General Sharin	g Security Customise		
	apache-drill-1.0.0		
Туре:	File folder (.0)		
Location:	C:\		
Size:	162 MB (170,769,012 bytes)		
Size on disk:	163 MB (171,266,048 bytes)		
Contains:	241 Files, 14 Folders		
Created:	08 June 2015, 13:09:47		
Attributes:	Read-only (Only applies to	files in folder)	
	Hidden	Advanced	
	OK Ca	Apply	·

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

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

This concludes the installation of Drill.

Next she downloads the Drill ODBC driver. (<u>http://package.mapr.com/tools/MapR-ODBC/MapR_Drill_odbc/ (http://package.mapr.com/tools/MapR-ODBC/MapR_Drill_odbc/ (http://package.mapr.com/tools/MapR-ODBC/MapR_Drill_odbc/)</u>).

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.

MapR Drill ODBC	Driver 64-bi	t Setup	_ 🗆 🗙
Destination Folder Please select your preferred install directory		W	NR.
Install MapR Drill ODBC Driver 64-bit to:			
C:\Program Files\MapR Drill ODBC Driver\ Change	T		
	Back	Next	Cancel

... and waits for the install to complete.

MapR I	Drill ODBC Driver 64-bit Setup 🛛 🗕 🗖 🗙
MAPR.	Completed the MapR Drill ODBC Driver 64-bit Setup Wizard
MapR Technologies	Click the Finish button to exit the Setup Wizard.
	L a
	Back Finish Cancel

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 (<a href="http://openweat

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	Туре	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.

		OD	BC Data	Source	Administra	tor (64	4-bit)		
er DSN	System DSN	File DSN	Drivers	Tracing	Connection P	ooling	About		
ystem Da	ata Sources:								
Name			Platform	Driver				Add	
MapR O mysql hiv	DBC Driver for ve metastore	Drill DSN	64-bit 64-bit	Map R Dri MySQL O	I ODBC Driver DBC 5.3 Unico	de Drive	er [Remo	ove
Sonra Cl	uster		64-bit	Simba Spark ODBC Driver		Configu	ure		
	An ODBC Sys A System data	tem data s a source is	ource sto visible to	res informa all users of	tion about how this computer,	to conn includin	ect to th g NT set	e indicated data rvices.	a provider.
					ок	Canc	el	Anniv	Help

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

Next she selects the MapR Drill ODBC driver:

	Name	11
-	MapR Drill ODBC Driver	1
OTT D	Microsoft Access dBASE Driver (*.dbf, *.ndx, *.mdx)	1
	Microsoft Access Driver (*.mdb, *.accdb)	1
	Microsoft Access Text Driver (*.txt, *.csv)	1
	Microsoft Excel Driver (*xls, *xlsx, *xlsm, *xlsb)	1
	MySQL ODBC 5.3 ANSI Driver	Ę
	MySQL ODBC 5.3 Unicode Driver	٤.,
		>

She gives the new data source a name and...

Mapf	R Drill OD	BC Driver DSN Setup		×
Data Source Name:	MapR ODB	C Driver for Drill DSN		
Description:	Sample Ma	pR Drill DSN		
Connection Type				
O Zookeeper Quorum	Quorum:	your-quorum-of-zookeepe	rs	
	Cluster ID:	drillbits1		
Direct to Drillbit	localhost		: 31010	
Authentication				
Authentication Type:	No Authen	tication		~
User:				
Password:				
Catalog:	DRILL			~
Default Schema:	default			¥
Advanced Properties:	Handshake	Timeout=5;QueryTimeout=	180;Timesta	^ ~
	Logging	p Options Drill	Explorer	
v1.0.0.1001 (64 bit)		Test OK	Cancel	

...tests the connection by clicking the Test button:

Test Results	>
SUCCESS!	^
Successfully connected to data source!	

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.

Browse	SQL	
Schemas:		
	default	^
e- 🗐 dfs	default	
	\$Recyc	le.Bin
🖨 🗎	apache	-drill-1.0.0
	📄 bin	
	📄 conf	
	🚞 jars	
·	log	
	sam sam	ple-data
	• n	ationsMF
		ationsSF
		egionsMF
		licks compaign ison
		licks.campaign.json
	{""} c	omp ison
	- Sid n	ation.parquet
		egion.parquet
	{""} s	tocks.json
	{***} w	veather.json
	{*"} w	eather_14.json

This throws an error:

An error ocur	red while	communicating w	vith the data sou	Irce.		
ERROR [HY0 SELECT * FF data/weather	000] [Ma ROM `dfs _14.json	pR][Drill] (1040) Dr `.`default`.`./apach ` LIMIT 100[30027]	rill failed to exe e-drill-1.0.0/sar Query executio	cute the que nple- in error. Det	ery: ails:[^
DATA_REAL type when yo /apache-drill [Error Id: 539t	D ERROF u are usi I-1.0.0/sa odddc-a	R: Error parsing JS ng a ValueWriter of mple-data/weathe Id8-4035-b59b-40a	SON - You tried of type Nullable er_14.jsonReco a7d858427c on	to write a Fl BigIntWrite rd 2Fragm sonrauli:31	oat8 rlmpl.F ent 0:0 010]]	ile v

At this stage our business analyst consults the documentation <u>https://drill.apache.org/docs/json-data-model/</u> 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;

Browse	SQL				
View De	finition S	AL:			
ALTER	SYSTE	M SET	<pre>`store.json.read_numbers_as_double` = true;</pre>	^	Preview
					Create As
				~	

She confirms by clicking Preview:

Ø	Drill Explorer	_ 🗆 🗡
Browse	SQL	
View Def	inition SQL:	
ALTER	SYSTEM SET `store.json.read_numbers_as_double` = true;	Create As
Total Nu	mber of Records: 1	
► 1	store.json.read_numbers_as_double updated.	

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

Browse SQL Schemas: M Browse SQL Schemas: M Schemas: M Schemas: M Schemas: M Schemas: M Schemas: M Schemas: M Schemas: Schemas: M Schemas: Schemas: M Schemas: Schemas: Sch	Metadata:		
Schemas: M Bell cp.default Bell dfs.default SRecycle.Bin Bell apache-dnll-1.0.0 Bell bin Bell conf Bell jars Bell log Bell sample-data Bell nationsMF	Metadata:		
Configure Second Data	Data Previe		
nationsSF regionsMF regionsSF clicks.campaign.json clicks.ison comp.json mation.parquet stocks.json weather.json weather.json git.properties KEYS LICENSE	▶ 1 2 3 4 5 6 6 7 8 9 10 11 11 <	city <td <td<="" td=""></td>	

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:

```
      Browse
      SQL

      View Definition SQL:
      CREATE OR REPLACE VIEW dfs.tmp.city_weather

      As
      SELECT

      t.city['id'] AS city_id,
      t.city['courty'] AS country_cd,

      t.city['court']['lon'] AS city_lon,
      t.city['court']['lat'] AS city_lat,

      t.city['courd']['lat'] AS city_lat,
      to_timestamp(t.'time') AS datetime,

      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 swather_desc_2
      PROM 'dfs'.'default'.'./apache-drill-1.0.0/sample-data/weather_14.json' AS 't';

      Total Number of Records:1
      ok summary
      1
```

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/ 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 <u>http://drill.apache.org/docs/selecting-nested-data-for-a-column/</u>

(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.

	S	erver Conne	ction	
Other	Data	bases	(ODBC)
Connect Usin	9			
O DSN:	MapR ODBC	Driver for Drill DS	N	~
O Driver:	MapR Drill O	DBC Driver		~
		Connect		
Connection A	ttributes			
Server:		1	Po	ort:
Database:				
Username:				
Password:				
String Extra	as:			

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

 city_weather (dfs. Converted to Online Database 100800 	tmp.cit	y_weather)									ě	Die O Bried	0	Add.,
Server MapR ODBC Driver for Drill DSN L.	(ity.we	te												
Schema														
dhimp -														
Table														
Enter table name P +														
8 Exact Contain: O Starts wi														
E city,weather_ity,weather)		Copy Show allows Show allows												
Ro New Custom S22	City Id	Oby Name	Country Cd	CityLon	CRY LM	Datetime	Tomp K.	temp C	Temp F	Humidity	Pressure	Weather Desc 1	Weather Desc 2	
	1283240	Kathmandu	MP	85.317	27.717	15/03/2014 08:20	297	2	4 75	33.000	1,020.00	scattered clouds	null	1
	3632308	Merida	WE .	-71,145	8.998	15/85/2014 08:20	289	1	5 66	99,000	835.59	Sky is Clear	rolf	
	1280737	Uhasa	CN	91,180	29,650	15/83/2014 08:20	201		4 4	12,000	611.65	Sky is Clear	null	
	345042.00	Istanbul	TR	28.983	41.085	15/85/2014 08:20	283	1	• <u>,</u> ×	87,000	1,021.00	Sky is Clear	roll	
	3496821	Map	DO	-71.070	19.552	15/83/2014 08:20	295	2	2 9	100,000	1,020:65	light rain	null	
	\$23523.00	Natchik	8.2	43.679	43.498	15/03/2014 08:20	286	1	1 10	30,000	1,018.00	overcast clouds	null	
	2267857	Lisbon	PT	-8.133	38,717	15/05/2014 08:20	281		8 40	82.000	1,021.00	fog	mist	
	3062707	Walbrzych	я.	16.284	\$0.771	15/83/2014 08:20	278		4 4	99,000	967.80	light rain	ruit	
	3091150	Naklo ned Notocla	R.	17.682	53,142	15/85/2014 08:29	250		7 48	77,000	994.00	light rain	mail	
	1784658	Zhengzhou	CN	113,649	34,758	15/83/2014 08:20	299	2	5 71	25.000	1,012.00	moderate tain	null	

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)





el iv er in gfa st e

st ti

(/blog/author/uli-bethke)

<u>Uli Bethke (/blog/author/uli-</u>	m
<u>bethke)</u>	e- v
CO-FOUNDER, SONRA	al
I lli Bethke is the co-founder	u A-
of Course	S
or <u>Sonra</u>	ql
(http://www.sonra.io). Sonra	- h
is a Big Data consulting	а
company in Ireland and a	d
partner company of MapR,	0
the only enterprise-ready	р 2
Hadoop distribution. Sonra	S
provide services and	0
accelerators for data	C
warehouse offload and data	e
lake implementations on	S
MapR	0
	cı al
Uli is a data visionary and	&
provides thought leadership	c a
in the architecture and	m
implementation of data driven	p ai
applications. He has led	g
some of the largest and most	n =
complex data warehouse	2
implementations in Europe.	0 1
	5
Uli is the chair of the <u>Hadoop</u>	-
User Group Ireland	0
(http://www.meetup.com/ha	ci
doop-user-group-ireland/).	aı _
You can get in touch with Uli	BI
by connecting with him on	g)

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

(/ st re а m in g-ar c hi te ct ur eu si n gа р а С h ek af k am а pr st re а m s ? S 0 ur С е = S

o ci

al		
&		
С		
а		
m		
р		
ai		

I