LocationTech

Rapid Analytic Development on Near Real-Time Data

Austin Heyne, CCRi
aheyne@ccri.com

CCRi
DATA TO KNOWLEDGE

FOSS4G NORTH AMERICA

The problem

Over 4,400 flights were canceled within, into or out of the United States on Wednesday, according to aviation data services company FlightAware. More than 700 flights have been canceled today, and airlines have canceled over 10,000 flights this month because of the weather.

ABC; March 22, 2018

British Airways has cancelled all flights from Gatwick and Heathrow as computer problems cause disruption worldwide. [...] A spokeswoman for the airline said: "We have experienced a major IT system failure that is causing very severe disruption to our flight operations worldwide."

The Independent; May 27, 2017

Ultra-large container ship CSCL JUPITER ran aground on Scheldt river bank at around 0700 LT Aug 14 at Bath, Zeeland, Netherlands, while proceeding downstream en route from Antwerp to Hamburg. [...] If hull of the giant ship is breached, dramatic situation may well turn into a nightmare.

Maritime Bulletin; August 15, 2017

Then there was the NotPetya cyberattack that hit Danish shipping giant Maersk in June. The crippling attack seized the industry's attention after it cost the company $200 million to $300 million and led to a temporary shutdown of the largest cargo terminal in the Port of Los Angeles.

Risk Management; March 1, 2018

Around half of the flights in Europe on Tuesday face delays after a computer failure at the Eurocontrol center in Brussels, Belgium.

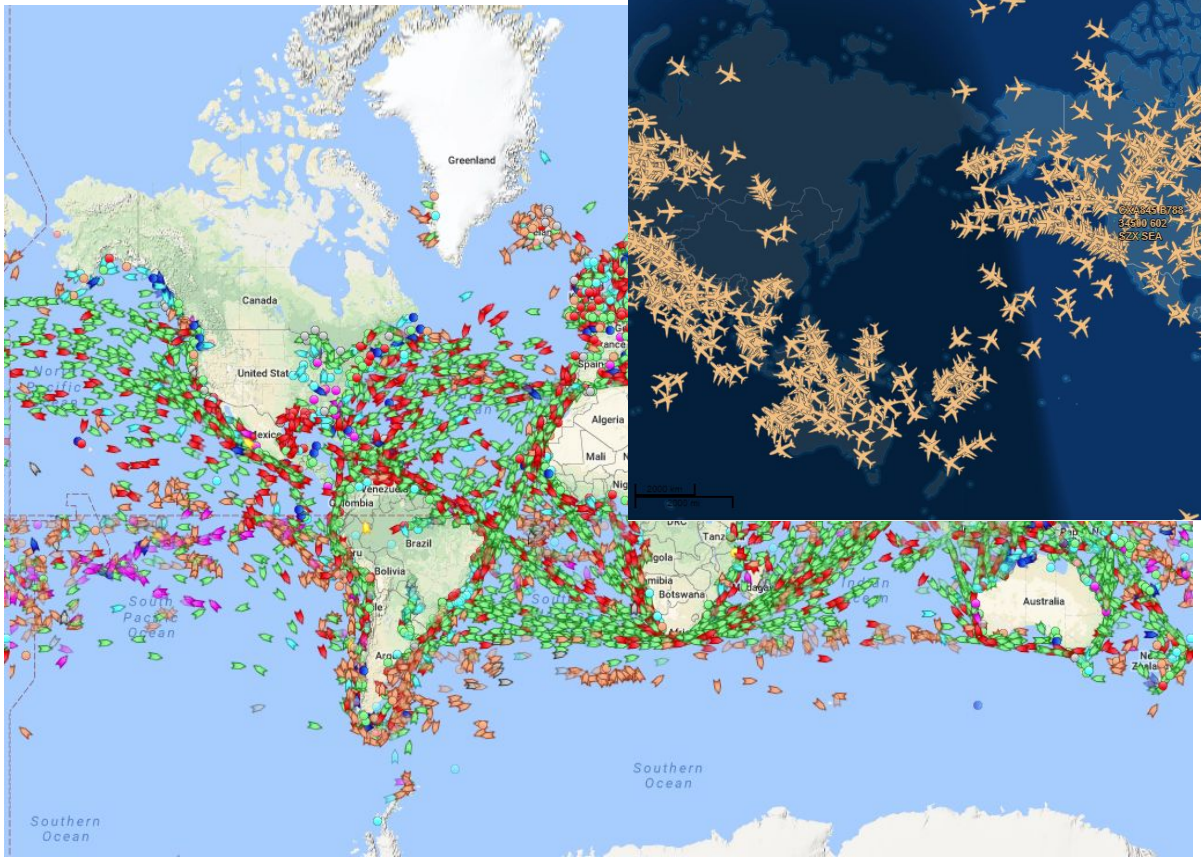
CNBC; April 3, 2018

[...] a massive power outage Sunday afternoon [at Hartsfield-Jackson Atlanta International Airport] left planes and passengers stranded for hours, forced airlines to cancel more than 1,100 flights and created a logistical nightmare during the already-busy holiday travel season.

The Atlanta Journal Constitution; December 18, 2017

CCRi

Are there any data?



Yes. Lots.

Satellite AIS

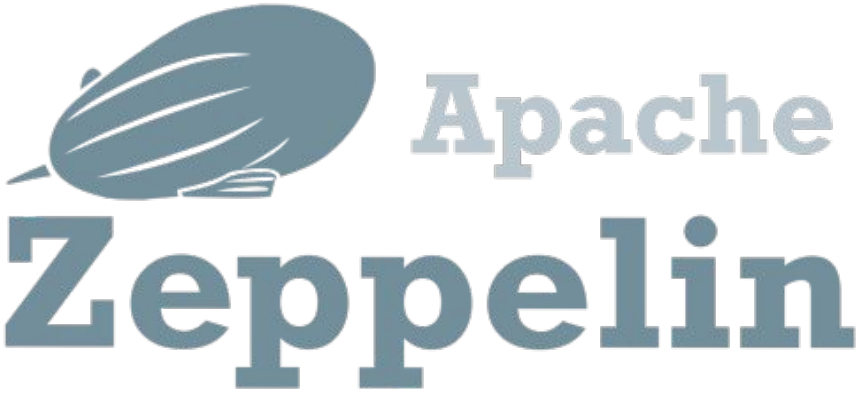
exactEarth

CCRi

How to handle big geo-temporal data?

A suite of tools for persisting, querying, analyzing, and streaming spatio-temporal data at scale... and it can SQL!

Analyst notebooks support innovation

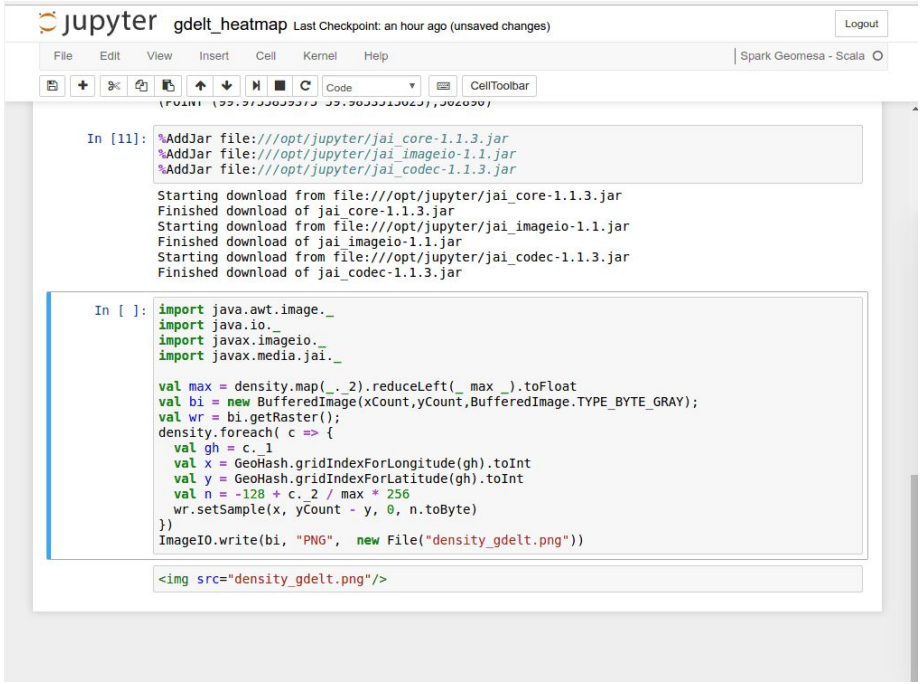


CCRi

Interactive Analysis in Notebooks

Writing (and debugging!) MapReduce / Spark jobs is slow and requires expertise.

A long development cycle for an analytic saps energy and creativity. The answer to both is interactive 'notebook' servers like Apache Zeppelin and Jupyter (formerly iPython Notebook).



What's missing?

We have...

- a problem
- big geo-time data
- big geo-time indexing
- interactive analyst notebook

We still need...

- a place to bring these all together
- to do it cheaply
- to do it flexibly and in a scalable way

Storage Options

- GPU RAM
- RAM
- Attached Disk (SSD, Platter)
- Cloud Blob Storage (S3, WASB)
- Cloud archive storage (Glacier)

Faster / **$$$**

Slower / $

Storage Options

- GPU RAM
- RAM
- Attached Disk (SSD, Platter)
- Cloud Blob Storage (S3, WASB)
- Cloud archive storage (Glacier)

Faster / **$$$**

Slower / $

Distributed Databases

Originally:

- Databases like Accumulo and HBase used HDFS to store data.
- HDFS needs 3-5x storage for replication and extra space.
- Compute and disk scaled together!

Distributed Databases

Originally:

- Accumulo and HBase used HDFS to store data.
- HDFS needs 3-5x storage for replication and extra space.
- Compute and disk scaled together!

Today:

- Accumulo and HBase can use Azure Blob storage and S3*
- Cloud storage scales
- Compute and disk scale separately!

* Accumulo works on Azure blob storage. Some modifications would required to support S3. Consult your cloud professional for details, interactions, and additional suggestions.

Distributed Databases

Originally:

- Accumulo and HBase used HDFS to store data.
- HDFS needs 3-5x storage for replication and extra space
- Compute and disk scaled together!

Today:

- Accumulo and HBase can use Azure Blob storage and S3*
- Cloud storage scales
- Compute and disk scale separately!

Take away:

GeoMesa HBase on S3 works great!

Others have used GeoMesa Accumulo on Azure...

Try it today!

* Accumulo works on Azure blob storage. Some modifications would required to support S3. Consult your cloud professional for details, interactions, and additional suggestions.

“Ditching the Database?”

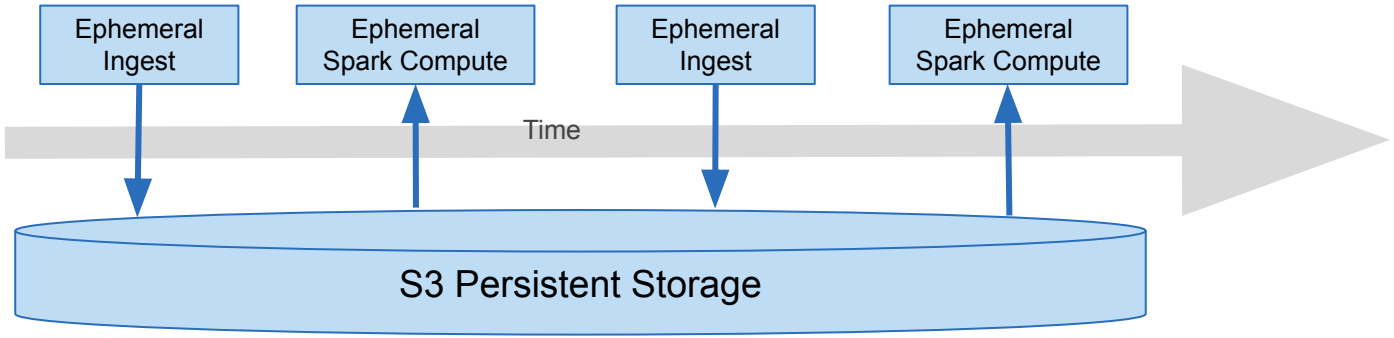
- Blobstores/Filesystems are key-value stores.
- Building a key-value store on top of a key-value store is kinda' redundant.

- GeoMesa can load and parse GDELT S3 files into Spark.
- GDELT on S3 is organized by date-keys.

- What would happen if we ‘ditched’ HBase and wrote files by space-time keys in cloud native storage?
- What parts of a ‘database’ do we really need to store/index non-updating, ascending temporal data streams?

GeoMesa FileSystem Datastore

- Serverless architecture
  - Standalone files in S3
  - Ephemeral compute for ingest and query
- Configurable partition schemes (geo + time)
- Parquet file format
  - Column-based storage (great for SQL!)
- Works great with Spark (intended for batch analysis)



Hurricane Harvey's Effect on Fuel Prices

Can oil tanker positions predict prices?

CCRi
DATA TO KNOWLEDGE

Setup

- Import GeoMesa dependency
- Create dataframe backed by GeoMesa relation
- Create SQL temporary view so we can query it

```
import org.locationtech.geomesa.spark._

val dataFrame = spark.read
  .format("geomesa")
  .option("fs.encoding","parquet")
  .option("fs.path","s3a://fsds-data/")
  .option("geomesa.feature", "AIS")
  .load()
dataFrame.createOrReplaceTempView("fsdsAIS")
```

FINISHED

```
import org.locationtech.geomesa.spark._
dataFrame: org.apache.spark.sql.DataFrame = [__fid__: string, mmsi: bigint ... 36 more fields]
```

Took 1 sec. Last updated by anonymous at January 30 2018, 12:32:20 PM. (outdated)

CCRi

Subselect Data

- Create rough subselection of data
  - Bound by time
  - Bound by bounding box roughy around the Gulf of Mexico
- Create a new temporary view from this subselection
- Cache the data (pull into memory)

```
%sql

CREATE OR REPLACE TEMP VIEW harvey as (
    select
        *
    from
        fsdsAIS
    where
        /* Harvey formed on Aug 17 */
        /* We pad more than a month so we can later crop off some */
        /* null data and still have a clean month to examine. */
        dtg > cast('2017-07-10' as timestamp)
        and
        /* Harvey dissipated on Sept 2 */
        dtg < cast('2017-10-3' as timestamp)
        and
        /* Approximate bounding box for the Gulf of Mexico */
        st_intersects(st_makeBBOX(-100.3051, 15.5020, -73.6962, 33.8272), position)
)
```

FINISHED

```
%sql

cache table harvey
```

FINISHED

Data Exploration

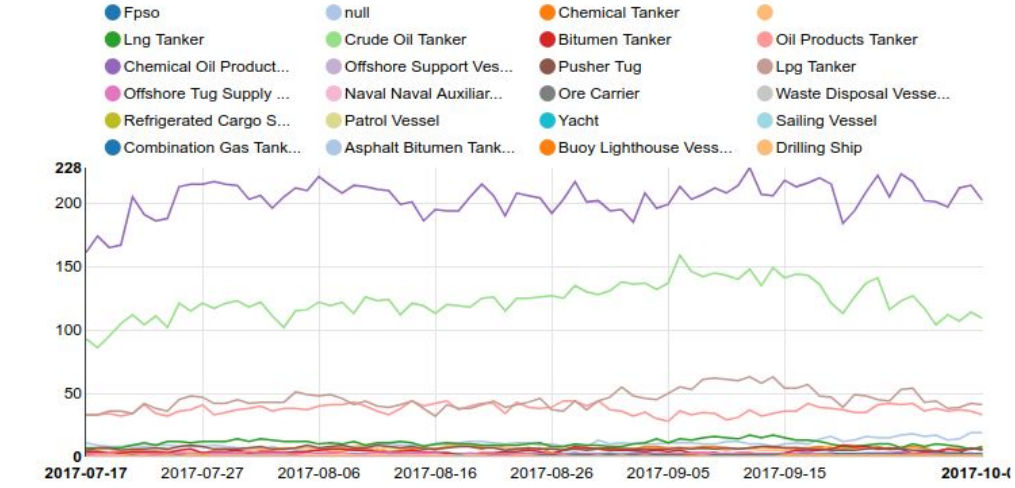
- Query for Tankers in the Gulf
- Get counts for each type of Tanker
- Group the counts by day
- Grach counts to see trends

```
%sql

select
    count(distinct mmsi),
    vessel_type_sub,
    dtg
from (
    select
        mmsi,
        date_sub(dtg, 0) as dtg,
        vessel_type_sub
    from
        harvey
    where
        vessel_type like "Tanker"
        and
        dtg > cast('2017-07-17' as timestamp)
        and
        dtg < cast('2017-10-3' as timestamp)
)
group by
    dtg,
    vessel_type_sub
order by
    dtg
```

FINISHED

settings



Took 5 sec. Last updated by anonymous at January 30 2018, 11:35:08 AM.

Data Exploration

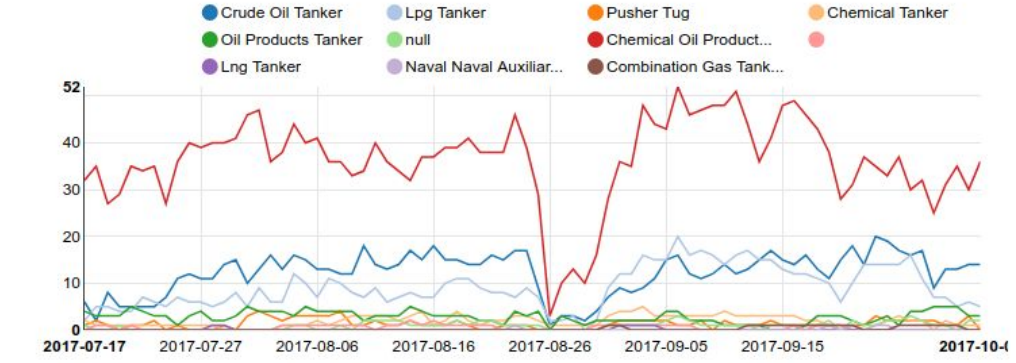
- Restrict our search to just Trinity Bay

```
%sql

select
    count(distinct mmsi),
    vessel_type_sub,
    dtg
from (
    select
        mmsi,
        date_sub(dtg, 0) as dtg,
        vessel_type_sub
    from
        harvey
    where
        /* Buffer a point in Trinity Bay by 50 Km */
        st_contains(st_bufferPoint(st_makePoint(-94.867089, 29.580314), 50000), position)
        and
        vessel_type like "Tanker"
        and
        dtg > cast('2017-07-17' as timestamp)
        and
        dtg < cast('2017-10-3' as timestamp)
)
group by
    dtg,
    vessel_type_sub
order by
    dtg
```

FINISHED

settings



Took 10 sec. Last updated by anonymous at January 30 2018, 11:48:20 AM.

Data Exploration

- Create a new temporary view of the number of ships in Trinity Bay

```
%sql

CREATE OR REPLACE TEMP VIEW ships as (
    /* Select the total number of tankers around huston during Harvey grouped by day */
    select
        count(distinct mmsi) as num_ships,
        /* Convert to 00:00:00 */
        date_sub(dtg, 0) as dtg_sub
    from (
        /* select ships around Houston during Harvey */
        select
            mmsi,
            dtg,
            vessel_type_sub
        from
            harvey
        where
            /* This is the Trinity Bay buffer for Houston */
            st_contains(st_bufferPoint(st_makePoint(-94.867089, 29.580314), 50000), position)
    )
    where
        /* Only get tankers */
        vessel_type like "Tanker"
    group by
        dtg_sub
)
```

FINISHED

Extra Data

- Pull in Gas price data
  - Acquired from EIA.gov
  - Two Gas Price Indexes
    - NYH: New York Harbor
    - GC: Gulf Coast
- Create temporary view so we can analyze with SQL

```
val gasPrices = sqlContext.read
    .format("csv")
    .option("header", "true")
    .option("treatEmptyValuesAsNulls", "true")
    .option("inferSchema", "true")
    .option("mode","DROPMALFORMED")
    .option("timestampFormat", "yyyy-MM-dd HH:mm:ss")
    .load("s3://fsds-data/csv/gasoline_spotprice_daily.csv")
gasPrices.createOrReplaceTempView("gasPrices")
```

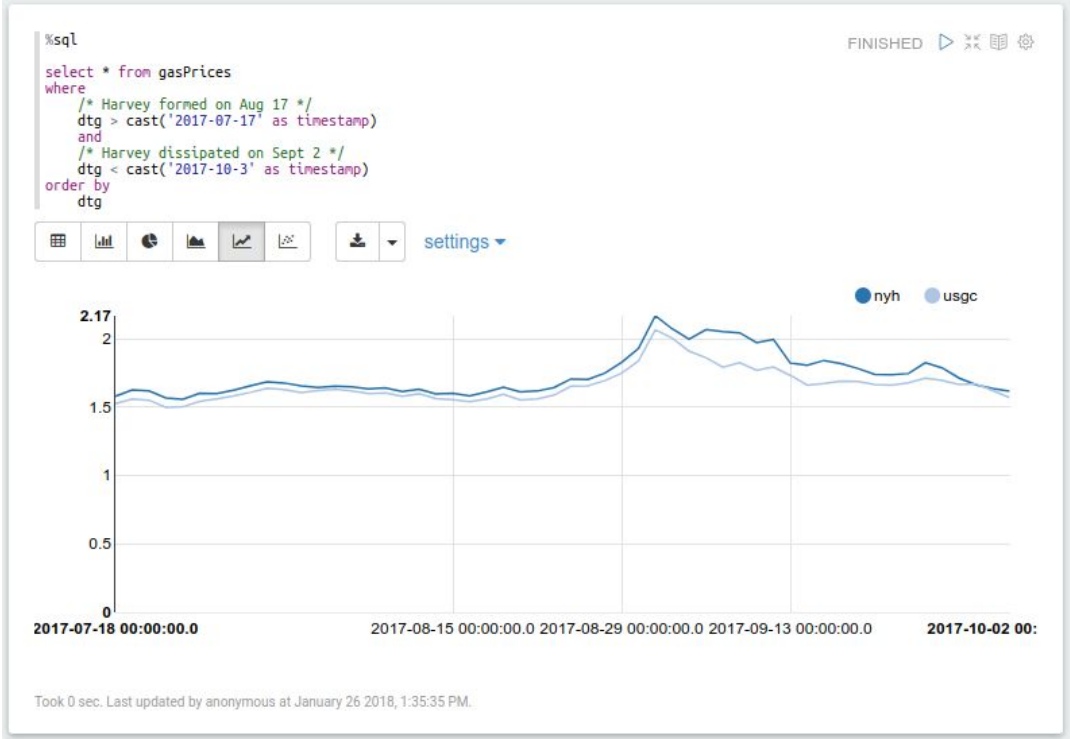
FINISHED

```
gasPrices: org.apache.spark.sql.DataFrame = [dtg: timestamp, nyh: double ... 1 more field]
```

Took 8 sec. Last updated by anonymous at January 30 2018, 11:45:21 AM. (outdated)

Data Exploration

- Graph data over time period of Harvey
- Notice we don't have daily values



Data Exploration

- Create temporary view of gas price data around our time of interest

```
%sql

CREATE OR REPLACE TEMP VIEW gas as (
    /* Select the Gas prices by day */
    select
        date_sub(dtg, 0) as dtg_sub,
        nyh as nyh_value,
        usgc as usgc_value
    from
        gasPrices
    where
        dtg > cast('2017-07-10' as timestamp)
        and
        dtg < cast('2017-10-02' as timestamp)
)
```

FINISHED

Data Exploration

- Backfill the price data with the last value to give us day-continuous data
- Min/Max Normalize gas and ship counts
- Graph gas prices and ship counts together

```
%sql
select
    /* Calculate min/max normalized values */
    dtg,
    ((nyh - min_nyh) / (max_nyh - min_nyh)) as nyh_norm,
    ((usgc - min_usgc) / (max_usgc - min_usgc)) as usgc_norm,
    ((num_ships - min_num_ships) / (max_num_ships - min_num_ships)) as ships
from (
    select
        /* Prepare for min/max normalization */
        dtg,
        nyh,
        min(nyh) over () as min_nyh,
        max(nyh) over () as max_nyh,
        usgc,
        min(usgc) over () as min_usgc,
        max(usgc) over () as max_usgc,
        num_ships,
        min(num_ships) over () as min_num_ships,
        max(num_ships) over () as max_num_ships
    from (
        /* Get from the JOIN */
        select
            /* Fill null values with last value to make graph continuous */
            ifnull(gas.nyh_value, last(gas.nyh_value,true) over (order by ships.dtg_sub)) as nyh,
            ifnull(gas.usgc_value, last(gas.usgc_value,true) over (order by ships.dtg_sub)) as usgc,
            num_ships,
            ships.dtg_sub as dtg
        from
            ships LEFT JOIN gas
        on
            ships.dtg_sub = gas.dtg_sub
    )
)
where
    /* Cut off lead dates that have null fuel values */
    dtg > cast('2017-07-17' as timestamp)
    and
    dtg < cast('2017-10-02' as timestamp)
order by
    dtg
```

FINISHED

settings

Took 11 sec. Last updated by anonymous at January 30 2018, 11:56:05 AM.

CCRi

Questions?

Find out more at http://geomesa.org
Connect with us on Gitter: https://gitter.im/locationtech/geomesa
See applications at CCRi's blog: http://www.ccri.com/blog/