RYSTAD ENERGY

BIG-DATA TECHNOLOGIES FOR PROCESSING AND ANALYZING SPATIAL DATA OF GLOBAL MARINE TRAFFIC

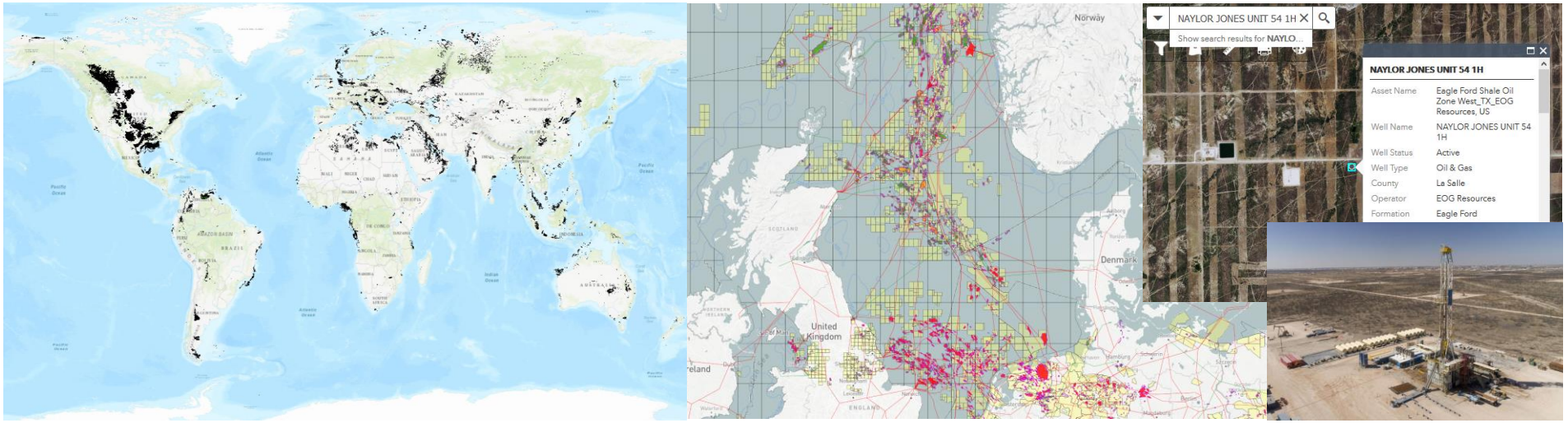
GEOKARTO 2020, KOŠICE

PETER PAVLIČKO

Introduction

- **Rystad Energy** – independent energy research and business intelligence company providing data, tools, analytics and consultancy services to clients exposed to the energy industry across the globe.

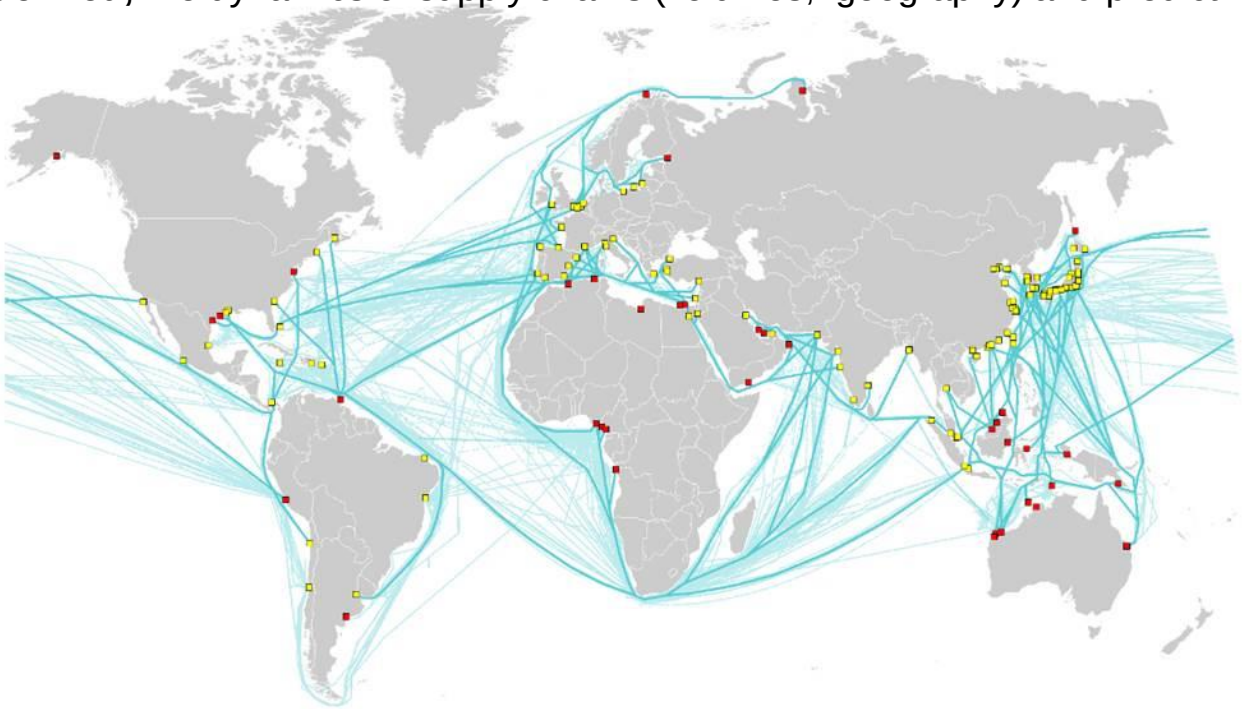
- Importance of spatial data processing, analyses and visualization within the company's business
- Spatial data on global level



RYSTAD ENERGY

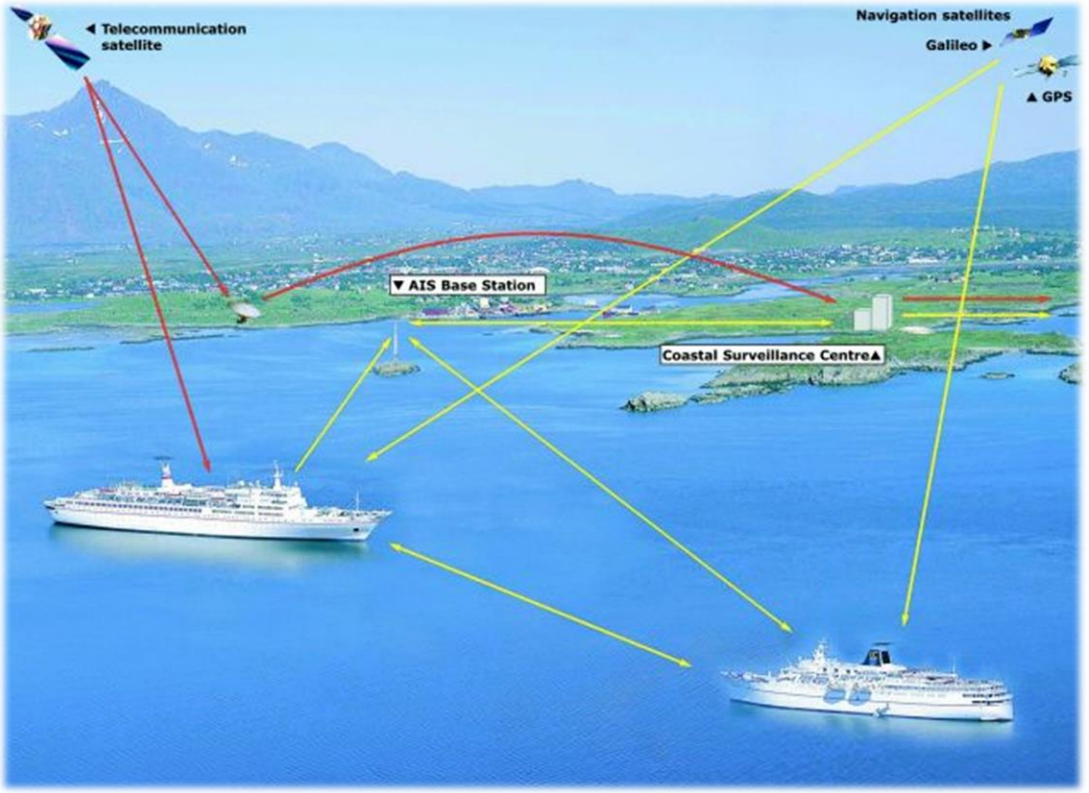
Market challenges

- **Energy Transition** – transformation of global energy sector from fossil-based to zero-carbon energy
- Modelling of global energy system dynamics
  - Increasing demand for information derived from **real-time data**, spatial data play significant role in this process
  - Technological innovations and advances allow **analyzing data** in more complex way and in real-time timeframes and **predict** future scenarios
- Global marine data – important source of data in modelling global energy system
  - **valuable information can be derived,** like dynamics of supply chains (volumes, geography) and predict market trends
  - AIS data – vessels tracking



AIS (Automatic Identification System)

- **AIS** is an **automatic tracking system** that uses transponders on ships and is used by vessel traffic services (VTS)
- **AIS** is intended, primarily, to allow ships to view marine traffic in their area and to be seen by that traffic
- The **International Maritime Organization's** International Convention for the Safety of Life at Sea requires **AIS** to be fitted aboard international voyaging ships with 300 or more gross tonnage (GT), and all passenger ships regardless of size
- several global data providers for AIS data (e.g. Spire, OrbComm, MarineTraffic)
- **OrbComm** – provides **API** for real-time access to AIS data (https://www.orbcomm.com/en/industries/maritime/satellite-ais)
- Paid access to OrbComm services
- Live and historic data (back to 2013)



Technology and Data processing

- Historic data from 2013-2019
  - approx. **40bil.** records
  - **10+ TB**
- New data generated daily approx. **2GB**
- What technology to use?
- Traditional technologies fail in operational and analytical procedures => Big-data technologies
- Infrastructure clustering, distributed computing

- **ElasticSearch Stack** consists of:
  - **Elasticsearch** - distributed, RESTful search and analytics engine
  - **Logstash** - server-side data processing pipeline that ingests data from a multitude of sources simultaneously, transforms it, and then sends it to your favorite "stash."
  - **Kibana** – dashboards and data visualizations, including maps
- Elasticsearch engine is built over Lucene's library
- ElasticSearch is free, we use "Basic License", more at: https://www.elastic.co/subscriptions

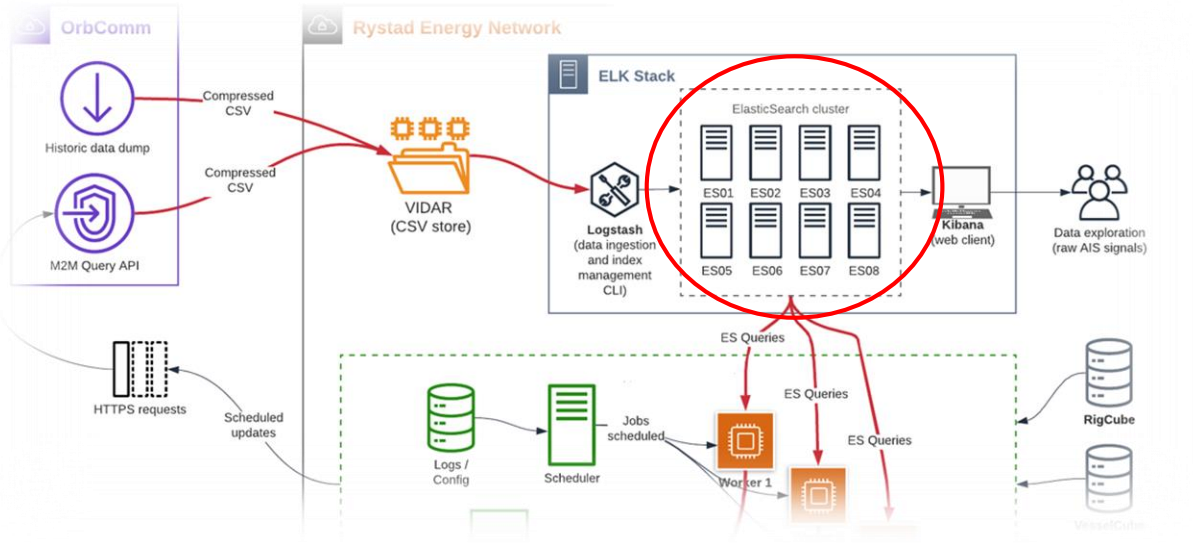
Elasticsearch and geospatial capabilities

- Elasticsearch/Google Maps analogy – search, visualize and analyze (geo-aggregations)
- Raw data stored in CSV files (source of data and backup archive)
- Logstash – data transformation, data in/out, output to Elasticsearch index (JSON)
- Indexing data – process of transforming source data into Elasticsearch index structure
- Elasticsearch spatial data indexing – using block k-d tree (BKD) geo-spatial data structure
  - Geospatial types support: ***geo_point*** (long/lat pairs) and ***geo_shape*** (points, lines, circles, polygons, multi-polygons, etc.)
  - Query capabilities:

RYSTAD ENERGY

Data Indexing and Cluster Setup

- Distributed computing, Elasticsearch cluster setup
- Cluster architecture is horizontally scalable
- HW and performance considerations
- Testing - most important part



Scaling and distribution of load…

Testing and Production system setup

- Several configuration parameters tested
- approx. 100mil documents (db records) indexed in Elastic in about 1 hour (posits1 table)
- some stats for different index and cluster setups:

10mil docs => import dur 11:50	(1 node, 4 shards, no replica, default Logstash configuration /125 batch size) 1.49GB
10mil docs => import dur 7:50 (cca)	(1 node, 4 shards, no replica, default Logstash configuration, batch size 250)
10mil docs => import dur 7:17	(2 nodes, 4 shards, no replica, default Logstash configuration) 1.49GB
10mil docs => import dur 16:29	(1 node, 8 shards, no replica, default Logstash configuration), 1.55GB size
10mil docs => import dur 41:11	(1 node, 24 shards, no replica, default Logstash configuration), 1.68GB size
10mil docs => import dur 5:17	(2 nodes, 4 shards, no replica, default Logstash configuration, batch size 250)
10mil docs => import dur 4:47	(2 nodes, 4 shards, no replica, default Logstash configuration, batch size 500)
100mil docs => import dur 1:58:34	(1 node, 4 shards, no replica, batch size 250), 13.3.GB
100mil docs => import dur 1:13:56	**(2 nodes, 8 shards, no replica, batch size 500), 13.3.GB //final configuration**

- Horizontal scalability – more resources (nodes) increase indexing performance
- Performance was tested on approx**. 1bilion docs** (ais index)

```
1 green open country                 Mx1vuOrbSlajdWwqOIwIOw 1 1        481 0  42.2mb    21mb
2 green open .kibana_task_manager_1  _ZUaY-uFQRGIjgloO-vHeQ 1 1          2 0  43.8kb  21.8kb
3 green open .apm-agent-configuration nbNCYttXTkqOSWF2fJanjw 1 1          0 0    566b    283b
4 green open ais                     kcQ_EcbJTWio93YGvxngcw 8 0 1008137069 0 142.6gb 142.6gb
5 green open .kibana_1               -rggPCCWT3-KcADnaeBXhg 1 1         15 4 112.1kb  59.2kb
6
```

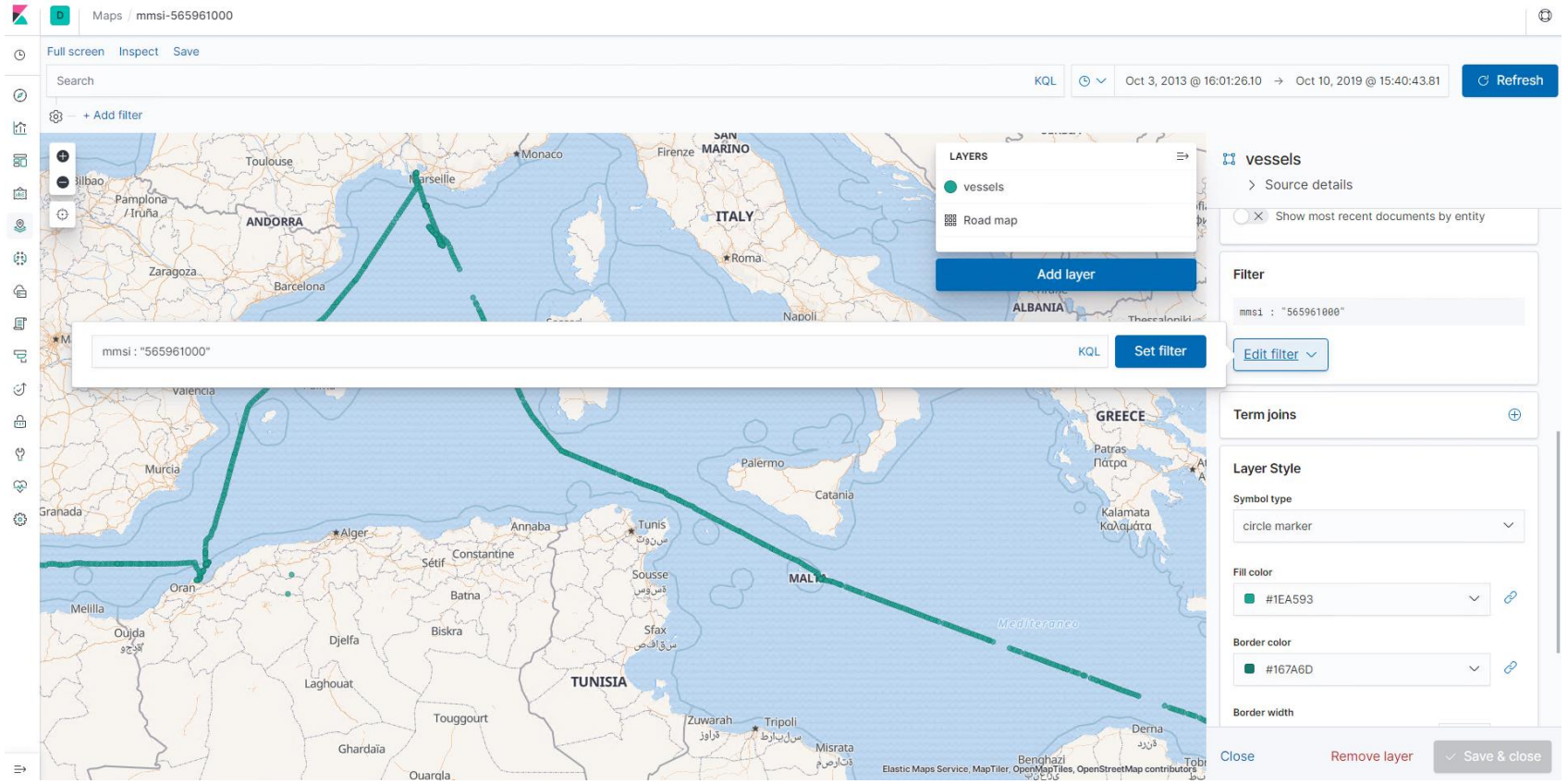
Queries and data analyses

- In general, very fast responses - regular attribute or spatial queries in **ms**
- Elastic uses Query DSL language – JSON interface (native, but supports SQL as well)
- Example of spatial query: “*return all vessels positions defined by given bbox”*

RYSTAD ENERGY

Data Visualization in Kibana (1)

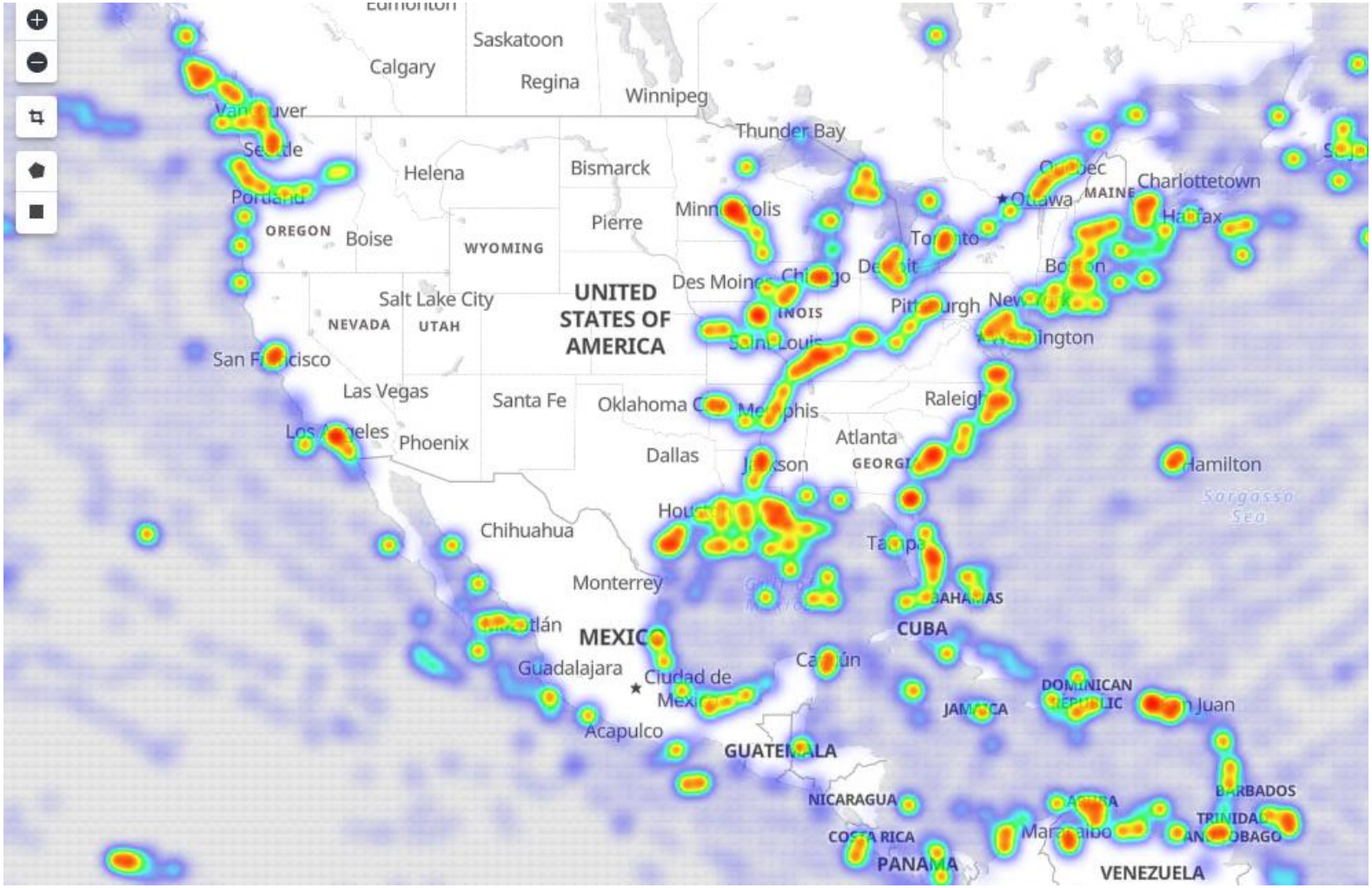
- **Kibana** is an integrated part of Elastic Stack
- **Elastic Map** – module for visualizing spatial data and geo queries results as well as real–time data
- Example: Tracking individual vessel (**mmsi : "565961000“**)



RYSTAD ENERGY

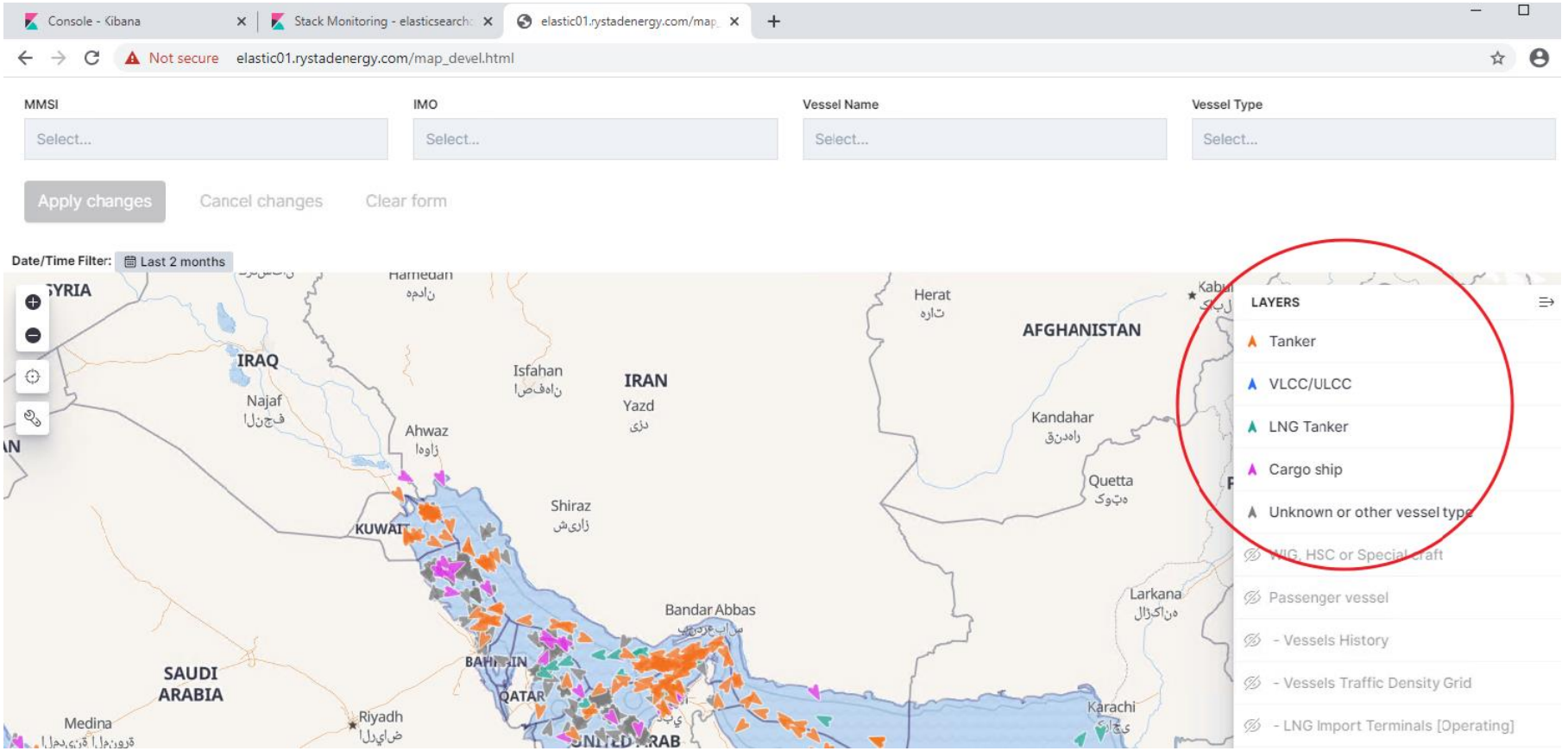
Data Visualization in Kibana (2)

- **Geo-aggregations** - vessels traffic density maps



Dashboards and web map applications in Kibana

- Real-time data monitoring
- Custom applications builder – no coding for creating powerful map applications with search and filtering options



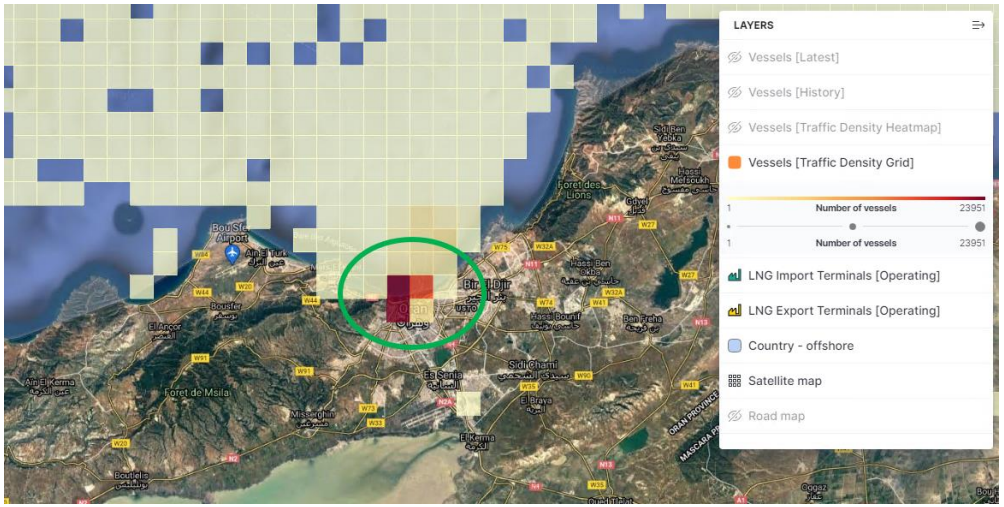
RYSTAD ENERGY

Automated ways of deriving spatial data from Elasticsearch

RYSTAD ENERGY

Geo_hash aggregation method for data extraction

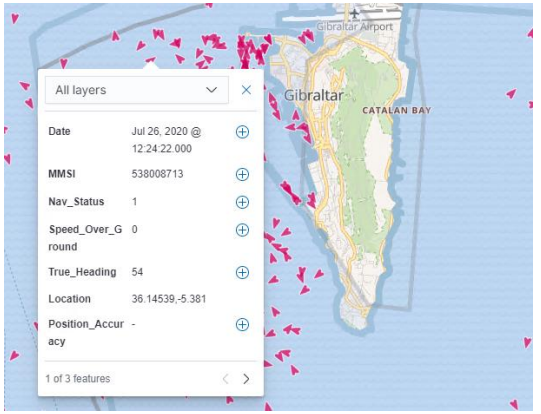
- Geo_hash aggregation – a method for grouping points into buckets that represent cells in a grid (multi-bucket aggregation)
- Automatic localization of LNG terminals and mooring points with high-accuracy
- Important for future vessels tracking and location tagging

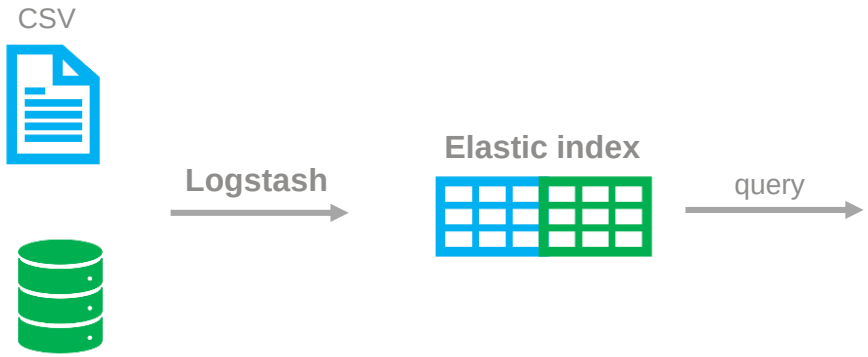


RYSTAD ENERGY

Enriching data during indexing

- Elasticsearch index
  - “Flat” data structure (compared to traditional relational database)
  - Designed for fast queries and aggregations
- *delete_by_query* API operations allowed, but not recommended on big indexes
- Enriching data methods – Logstash lookup plugins (jdbc, dns, elasticsearch…)

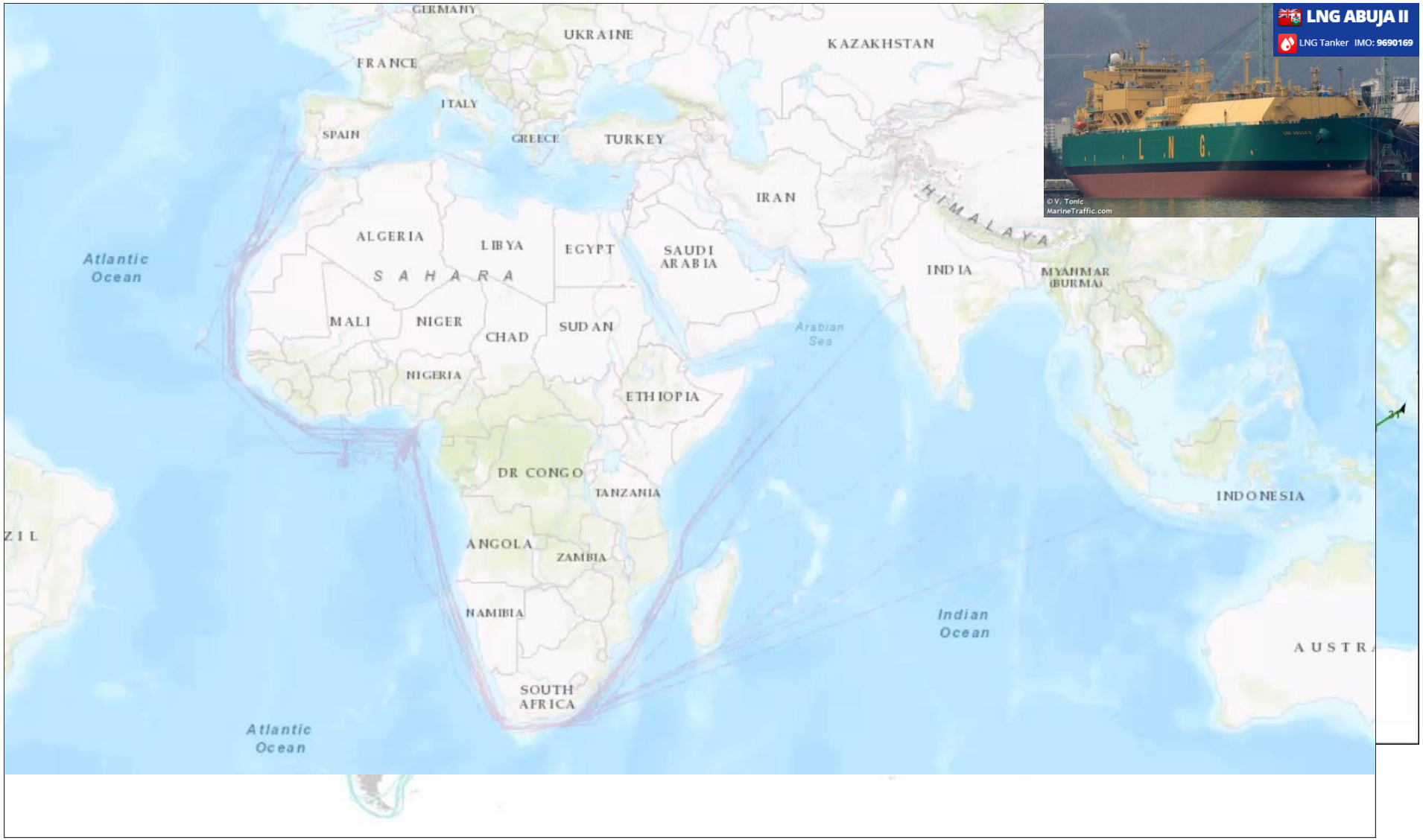





```
"_source" : {
        "path" : "//vidar/OrbComm/Streamed data/2020/07/04/csv/20200704_15h_ITU123_1.csv",
        "True_Heading" : 266,
        "Location" : "-8.59381,132.01533",
        "Turn_Rate" : -123,
        "@timestamp" : "2020-07-04T15:47:58.023Z",
        "Course_over_Ground" : 1031,
        "@version" : "1",
        "Date" : "2020-07-04 15:47:11",
        "fingerprint" : "20200704154711338736684",
        "Speed_Over_Ground" : 114,
        "Message_Type" : 3,
        "MMSI" : "338736684",
        "Nav_Status" : 15,
        "Timestamp" : "2020-07-04 15:47:47“,
        “VesselType”: “VLCC”,
        “Destination”: “Gibraltar”,
        “VesselCategory”: “Tanker”

}
```

LNG tradeflows



RYSTAD ENERGY

Conclusion

- New technologies bring new possibilities to analyze existing data in real-time and predict future trends
- Increasing focus and investments on big-data technologies in private sector (not just a buzzword), including geospatial industry
- Case-study on global vessels data related to energy sector
  - Applicable to similar projects (sensor data, GPS positions, etc.)
- New challenges for developing and implementing big-data algorithms to deal with spatial datasets

RYSTAD ENERGY

Rystad Energy is an independent energy consulting services and business intelligence data firm offering global databases, strategy advisory and research products for E&P and oil service companies, investors, investment banks and governments. Rystad Energy is headquartered in Oslo, Norway.

Headquarters
Rystad Energy
Fjordalléen 16, 0250 Oslo, Norway

Americas +1 (281)-231-2600
EMEA +47 908 87 700
Asia Pacific +65 690 93 715

Email: support@rystadenergy.com