CEU CENTRAL EUROPEAN UNIVERSITY

DIFFERENT SHAPES OF DATA

SEMINAR 4

LASZLO SALLO & SZILVESZTER MOLNAR

NOVEMBER, 2017

TOOLS WE TRY TODAY

INFLUXDB

SOLR

NEO4J

SESSION RULES

WE ARE GOING TO USE SOCRATIVE …

SERVER LOCATION AND ACCESS

ALL 3 TOOLS ARE INSTALLED ON AMAZON CLOUD

AND CAN BE ACCESSED AT:

http://ceudsd.net/

GENERAL RULE ABOUT THE ACCESS

DATA SOLUTION CAN BE GENERALLY ACCESSED:

PROGRAMING LANGUAGES (LIKE R, PYTHON)

NOTEBOOKS (ZEPPELIN, JUPYTER)

WEB INTERFACE ✔

HTTP SERVICE ✔


InfluxDB

INFLUXDB - INSTRUCTIONS

BROWSER: http://ceudsd.net/ > Influx DB

INFLUXDB - INSTRUCTIONS

THE DATA IS ALREADY PRE-LOADED FROM

https://s3-us-west-1.amazonaws.com/noaa.water.database.0.9/NOAA_data.txt

LINKS TO HELP YOU:

https://docs.influxdata.com/influxdb/v1.0/query_language/data_exploration/

https://docs.influxdata.com/influxdb/v1.0/query_language/math_operators/

https://docs.influxdata.com/influxdb/v1.0/query_language/functions/

NOAA_WATER_DATABASE SCHEMA

h2o_feet

---

time
level description
water_level
location

h2o_temperature

---

time
degrees
location

~ 80.000 ROWS

average_temperature

---

time
degrees
location

h2o_pH

---

time
pH
location

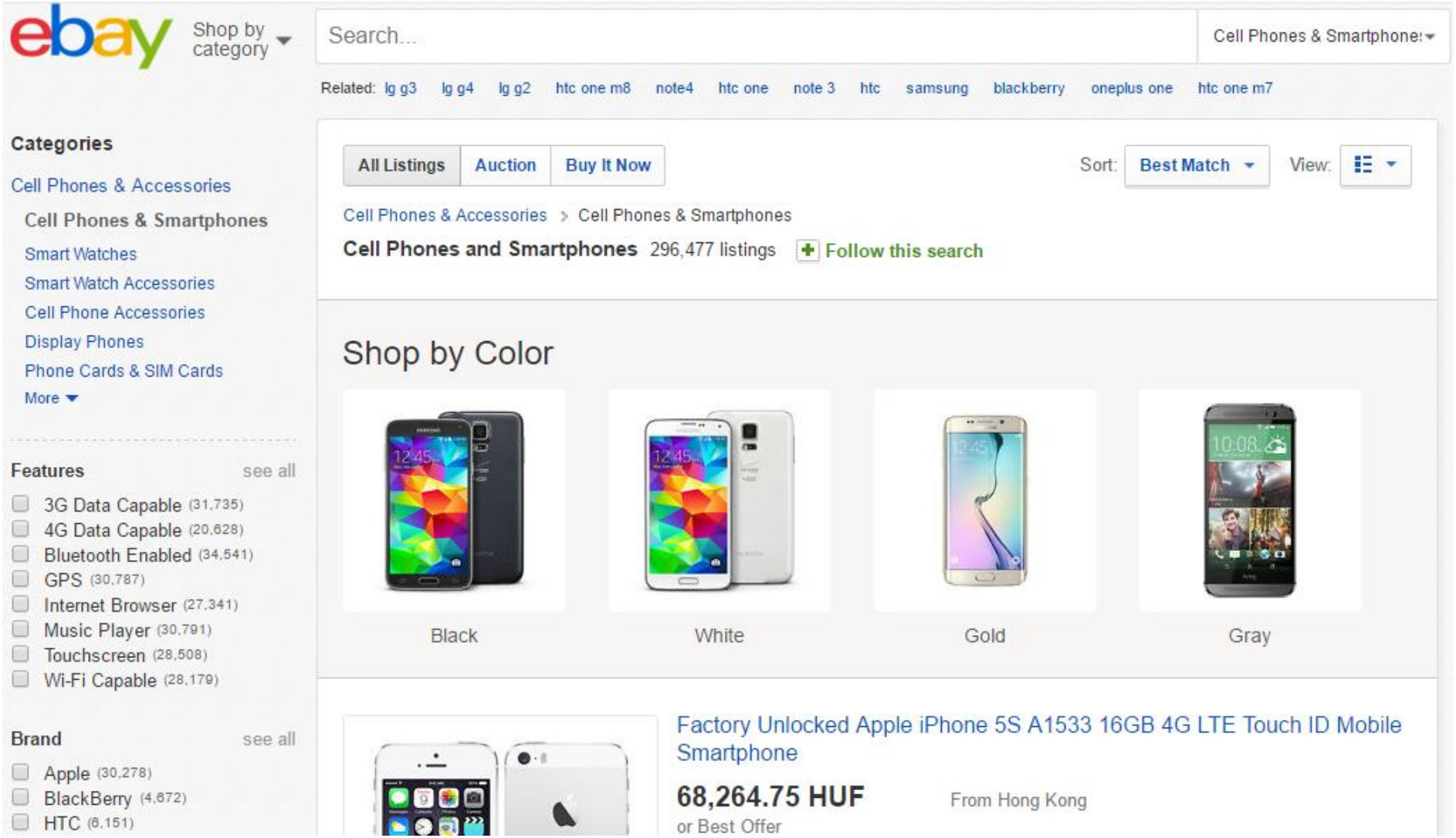
h2o_quality

---

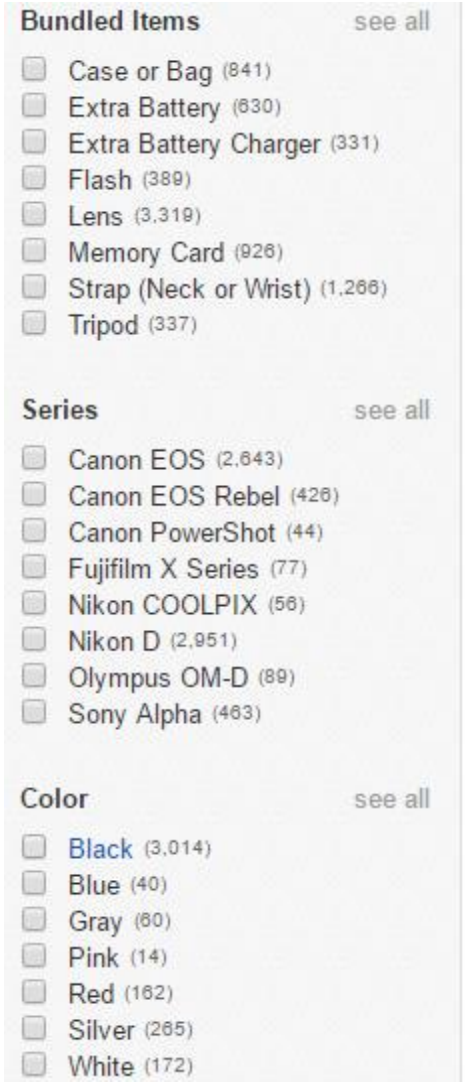
index
index
location
randtag


Solr

FACETED SEARCH



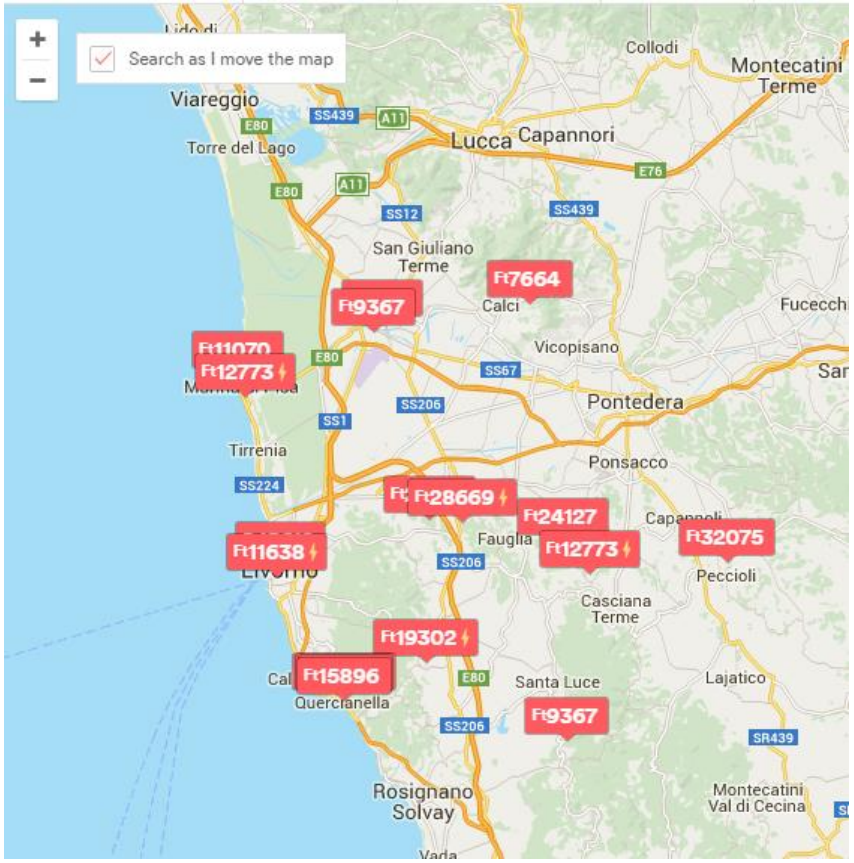
FACETED SEARCH



GUIDED QUERIES

E-COMMERCE: GUIDE YOUR CUSTOMERS
ANALYTICS: SLICING AND DICING

GEOSPATIAL SEARCH



COORDINATE TO ITEMS

AREA BY DIFFERENT SHAPES
PROXIMITY

SOLR - INSTRUCTIONS

BROWSER: http://ceudsd.net/ > Influx DB

THE DATA IS ALREADY PRE-LOADED FROM

https://cran.r-project.org/web/packages/nycflights13/nycflights13.pdf

LINKS WHICH MIGHT HELP YOU:

https://cwiki.apache.org/confluence/display/solr/The+Standard+Query+Parser

https://cran.r-project.org/web/packages/solr/solr.pdf

NYCFLIGHTS SCHEMA

year,month,day Date of departure

dep_time,arr_time Actual departure and arrival times, local tz.

sched_dep_time,sched_arr_time Scheduled departure and arrival times, local tz.

dep_delay,arr_delay Departure and arrival delays, in minutes. Negative times represent early departures/arrivals.

hour,minute Time of scheduled departure broken into hour and minutes.

carrier Two letter carrier abbreviation. See airlines to get name

tailnum Plane tail number

flight Flight number

origin,dest Origin and destination. See airports for additional metadata.

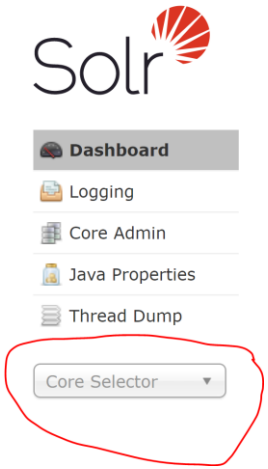
air_time Amount of time spent in the air

distance Distance flown

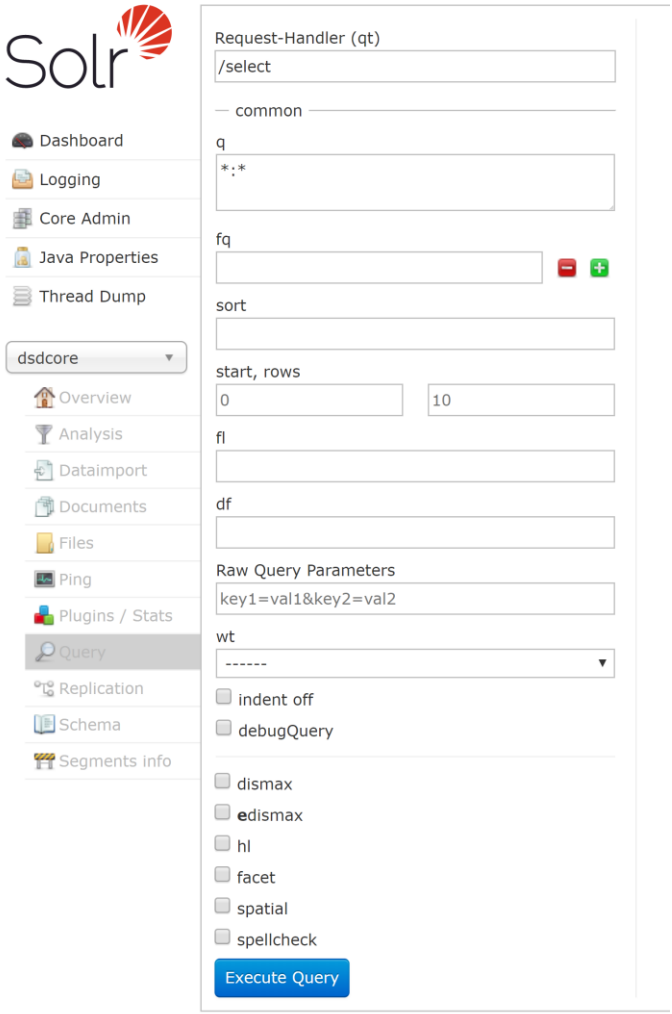
time_hour Scheduled date and hour of the flight as a POSIXct date. Along with origin, can be used to join flights data to weather data.

SOLR ADMIN

select dsd core



explore menu items, especially “query”

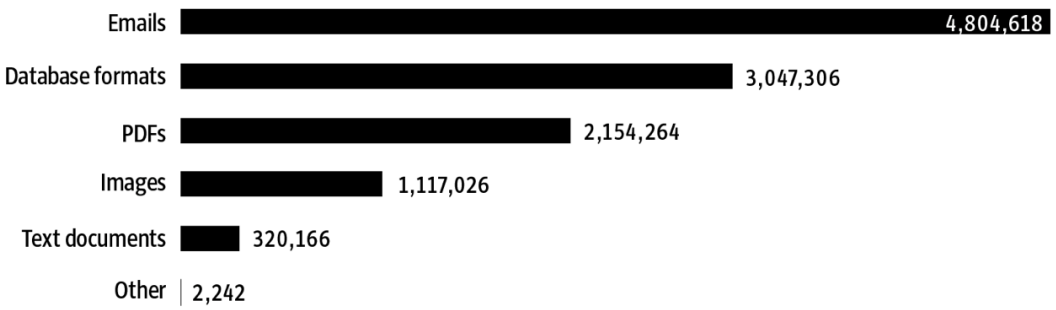



neo4j

USE CASE – PANAMA PAPERS

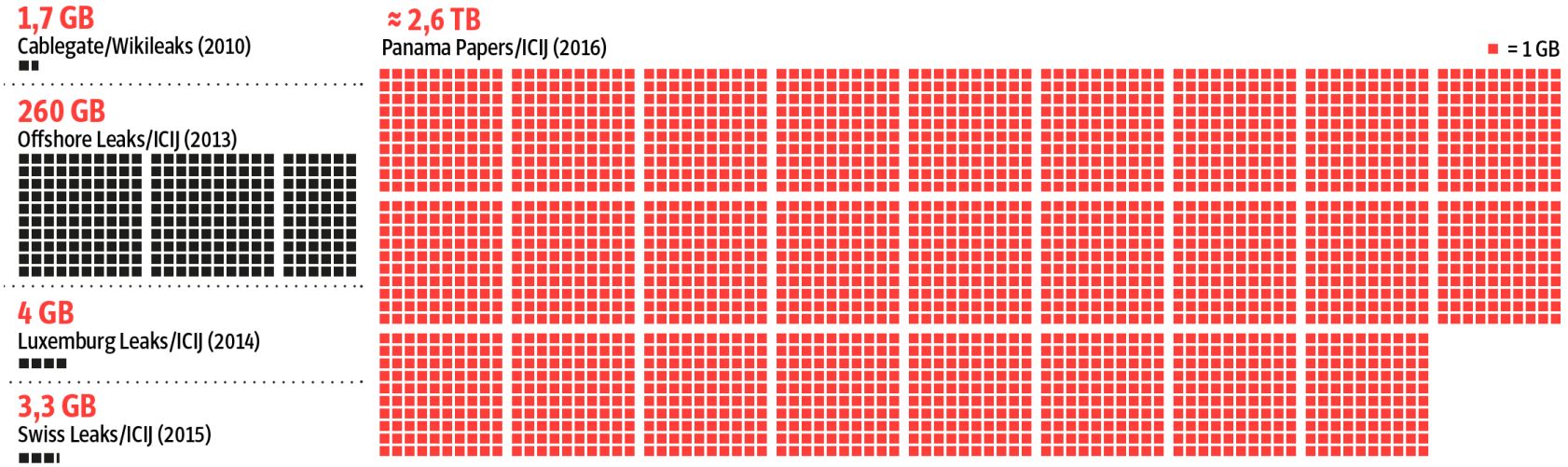
The structure of the leak

The 11,5 millionen contain the following file types

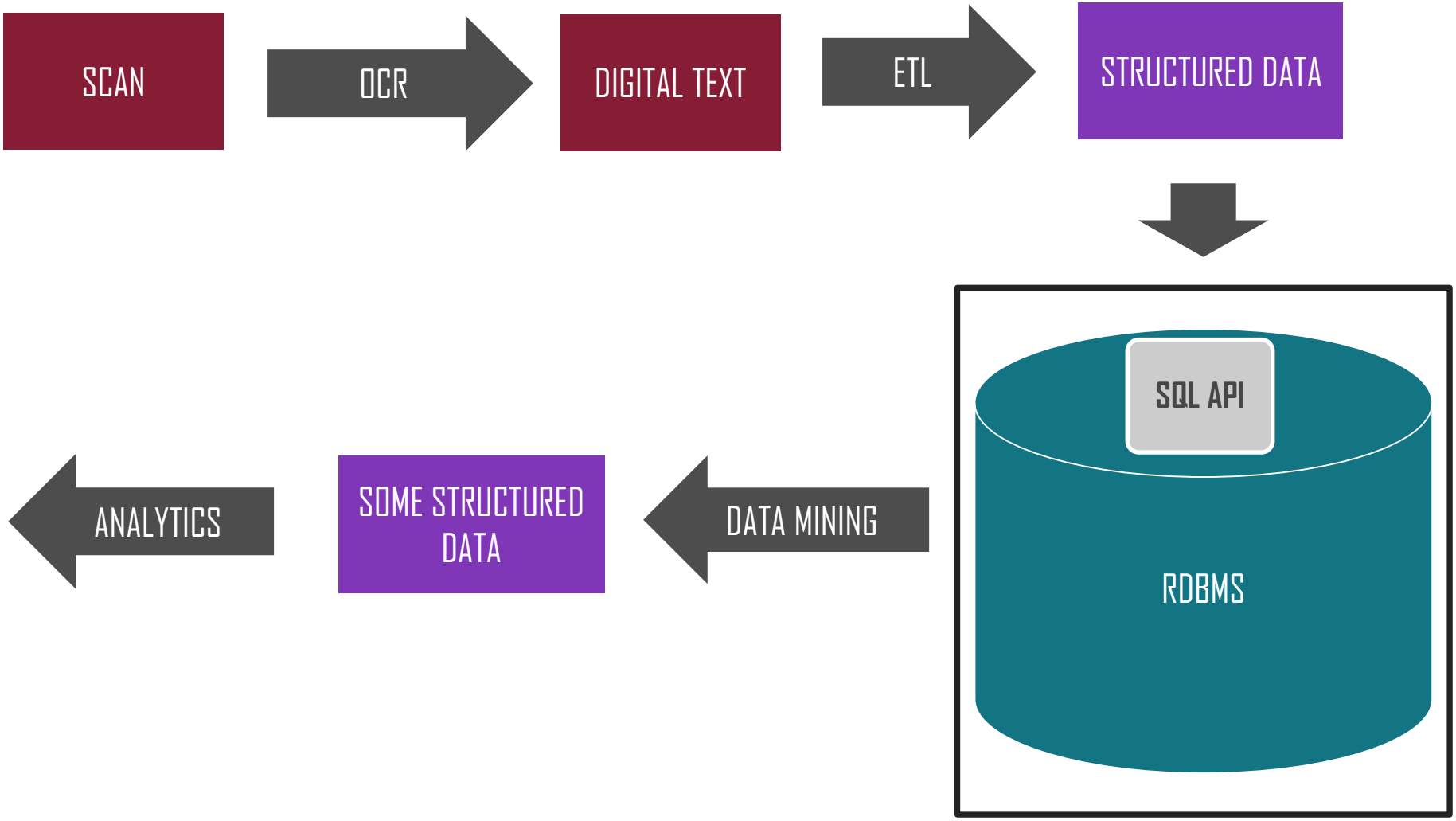


The scale of the leak

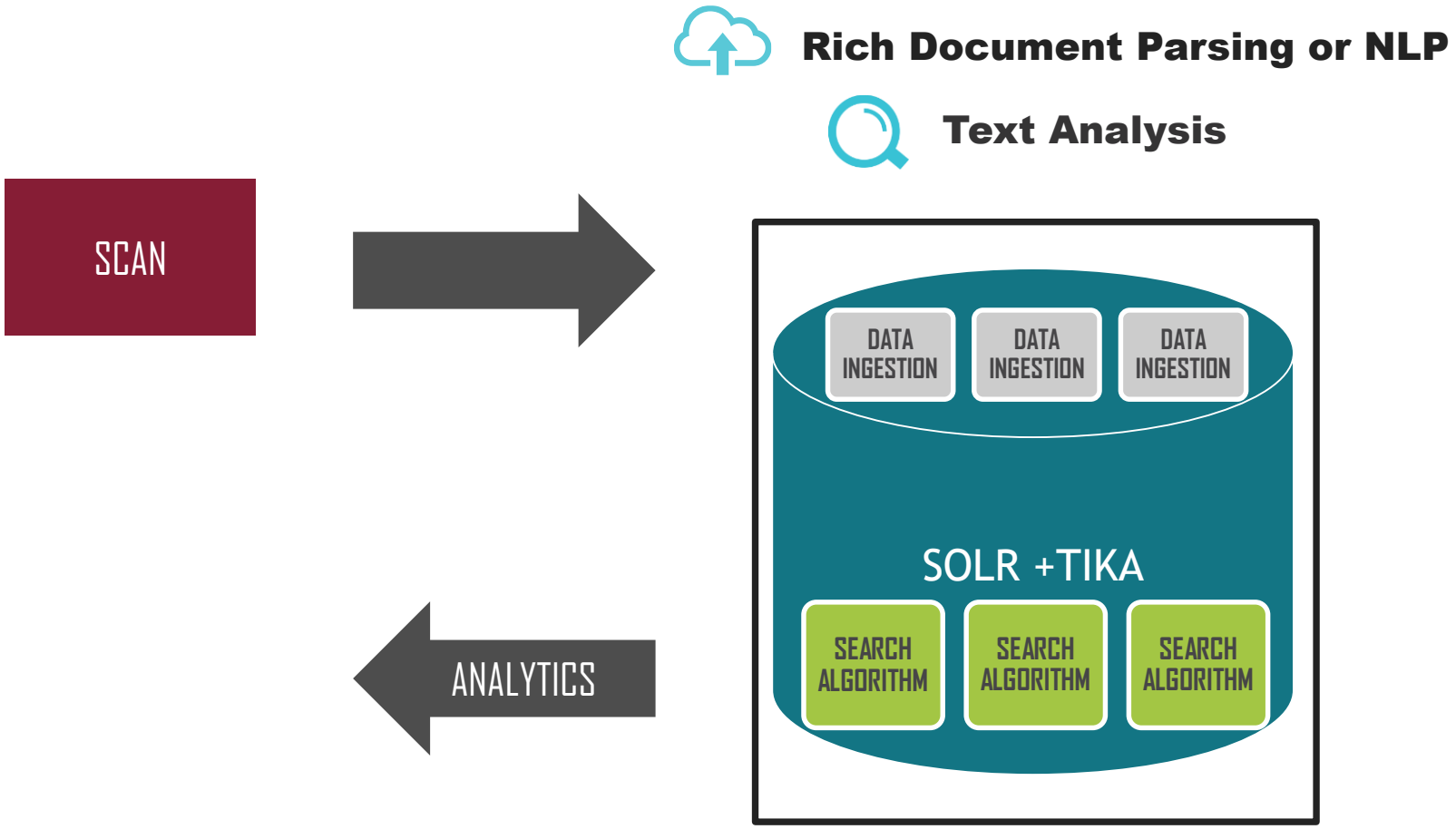
Volume of data compared to previous leaks



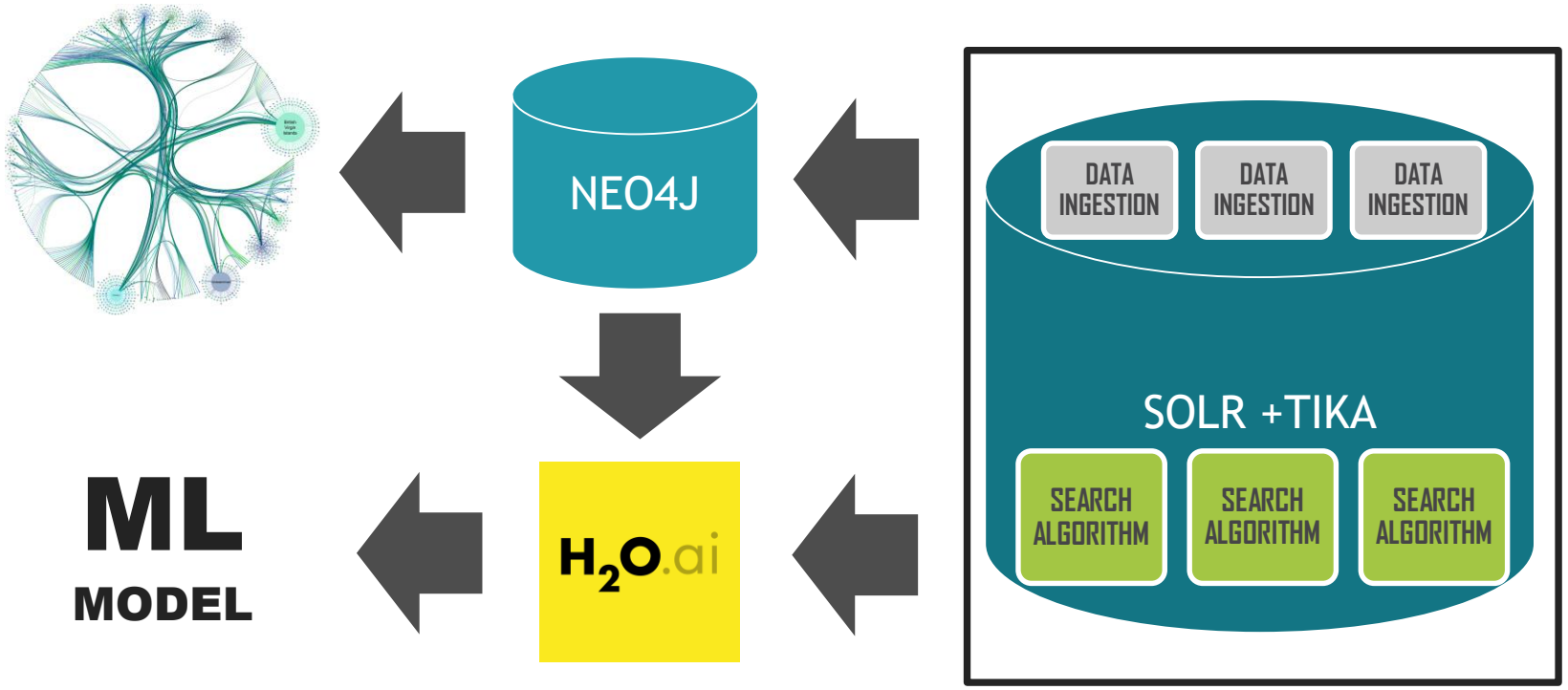
PANAMA PAPERS – CLASSICAL FLOW



WITH SEARCH ENGINE



ADVANCED ANALYTICS



NEO4J - INSTRUCTIONS

BROWSER: http://ceudsd.net/ > Neo4j

PASSWORD: ceudsd357

THE DATA IS ALREADY PRE-LOADED FROM

https://offshoreleaks.icij.org/pages/database

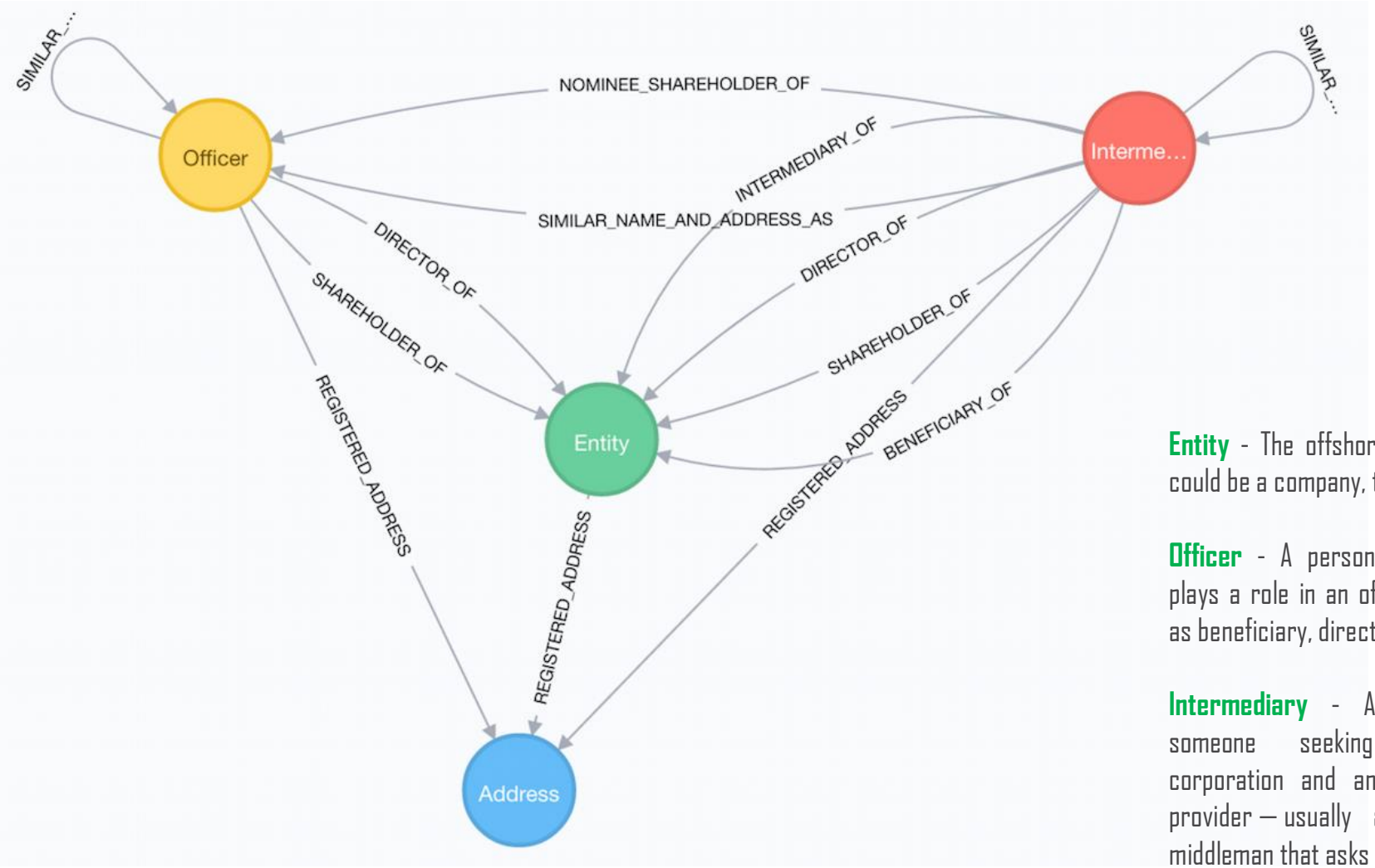
LINKS WHICH MIGHT HELP YOU:

https://neo4j.com/developer/cypher-query-language/

http://neo4j.com/docs/developer-manual/current/cypher/

https://cloudfront-files-1.publicintegrity.org/offshoreleaks/neo4j/guide/index.html

PANAMA PAPERS SCHEMA



Entity - The offshore legal entity. This could be a company, trust, foundation ...

Officer - A person or company who plays a role in an offshore entity, such as beneficiary, director, or shareholder.

Intermediary - A go-between for someone seeking an offshore corporation and an offshore service provider — usually a law-firm or a middleman that asks an offshore service provider to create an offshore firm.

Address - The registered address