DWS

(Data Warehouse Solution for GDELT dataset)

Artsiom Sinitski

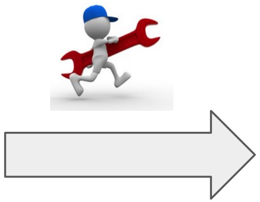
Data Engineering Fellow

Insight , New York City

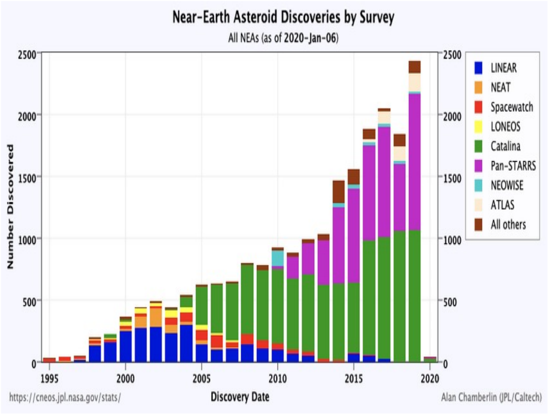
My Motivation

- Is to create a solution for collecting, managing and analysing GDELT data set effortlessly
- Target users are business analysts

FROM THAT



DWS



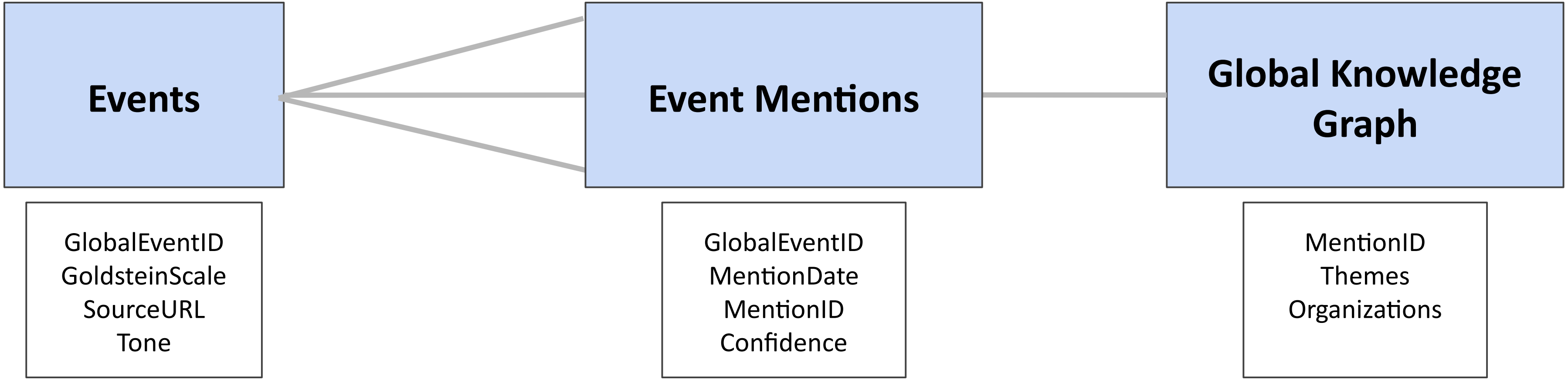
TO THIS

Why do we care?

- Time between a question and the answer is much shorter and the business decisions are made faster
- This helps to lower operating expenses and to develop a more effective business strategy faster
- Which contributes to increased profits

Dataset

- Global Database of Events, Language & Tone (GDELT) - collection of world’s broadcast, print and web news
- Volume: ~2.5 TB per year (CSV format)



Events

GlobalEventID
GoldsteinScale
SourceURL
Tone

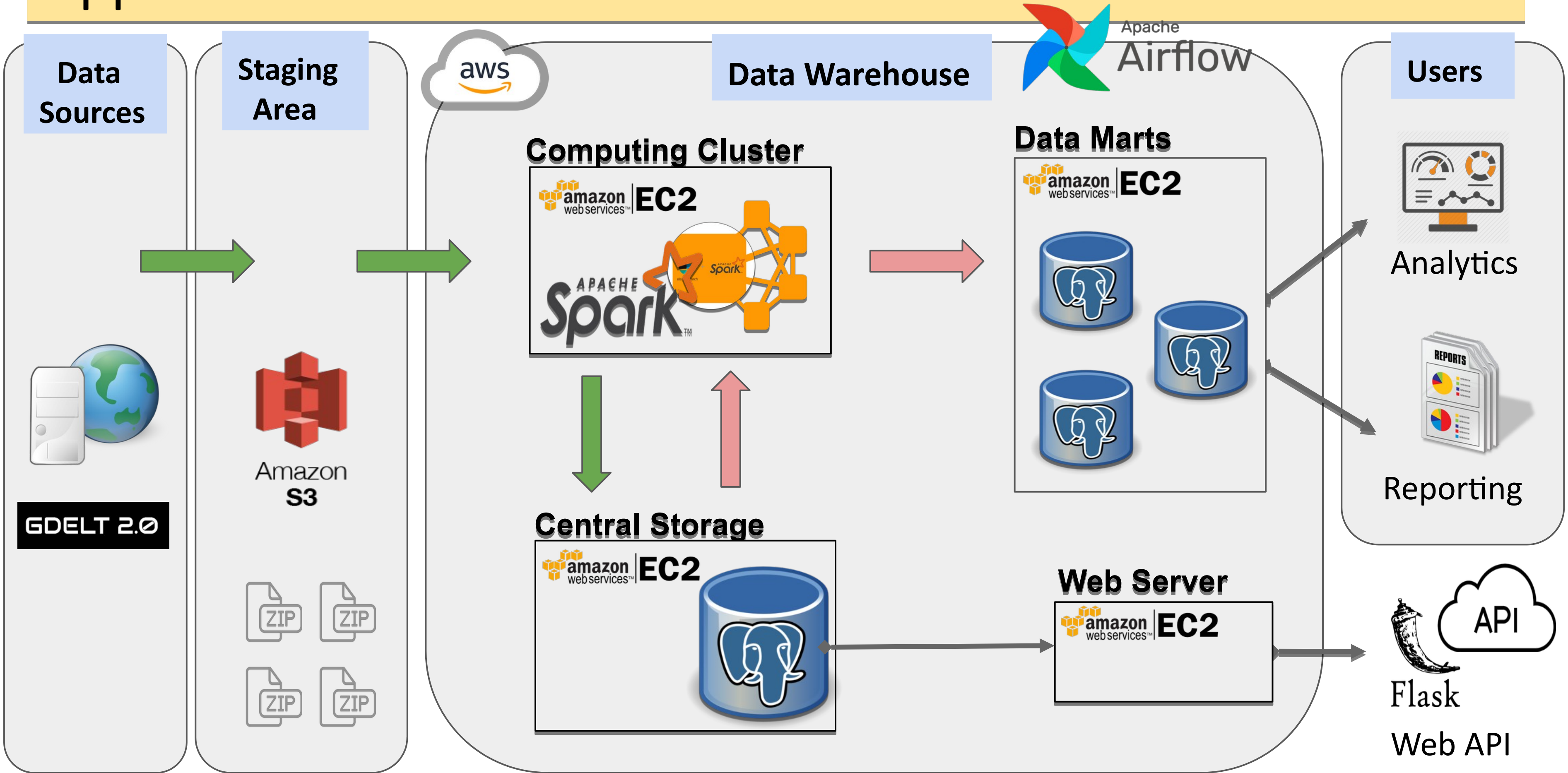
Event Mentions

GlobalEventID
MentionDate
MentionID
Confidence

Global Knowledge Graph

MentionID
Themes
Organizations

Approach



Data Sources

Staging Area

Data Warehouse

Users

Computing Cluster

Data Marts

Central Storage

Web Server

Analytics

Reporting

Web API

Approach Memo (1 of 2)

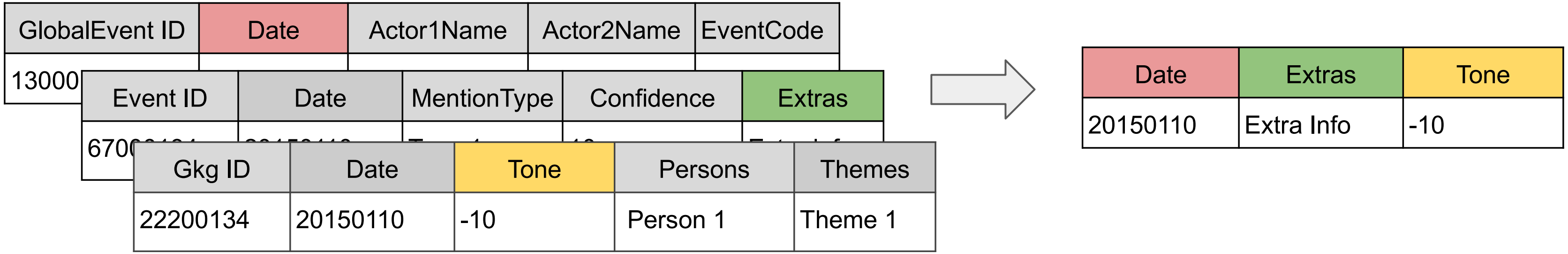
- Data Warehouse Solution is hosted in the AWS Cloud
- AWS S3 object store is used for “Staging Area”
- DWS servers are run on AWS EC2 instances
    - Apache Spark cluster nodes
    - Central Storage and Data Mart databases
    - Web server (Flask)
- The current data is downloaded and combined with the historical data by Apache Airflow which schedules ETLs processes once per day

Approach Memo (2 of 2)

- Central Storage ETL (➡) - extracts the raw data from GDELT web site, enforces GDELT schema onto it using Apache Spark engine and lastly, saves the data into Central Storage
- Data Mart ETL (➡) - extracts a data subset from Central Storage, breaks it up into appropriate tables (according to the "star" data schema) and loads the tables into Data Marts

Trade Offs

- Retrieving data from multiple large tables is very slow!
  - Create a separate table on subset



GlobalEvent ID	Date	Actor1Name	Actor2Name	EventCode
13000				

Event ID	Date	MentionType	Confidence	Extras
670				

Gkg ID	Date	Tone	Persons	Themes
22200134	20150110	-10	Person 1	Theme 1

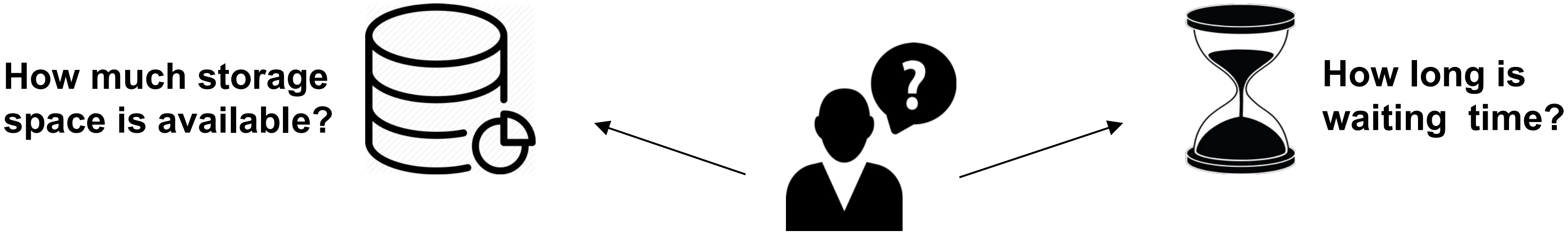
Date	Extras	Tone
20150110	Extra Info	-10

  - Also, consider denormalizing data and indexing tables
- Data redundancy vs. time complexity of O(m*n*k)

Trade Offs (contd.)

- Benchmarking results:

	Execution Time	Data set size
Multi-table Join Select	1.5 mins	~ 16.4 mil records (Events table) ~ 55 mil records (Mentions table)
Subset Table Select	2 sec	~ 38.6 thousand records (Subset table)



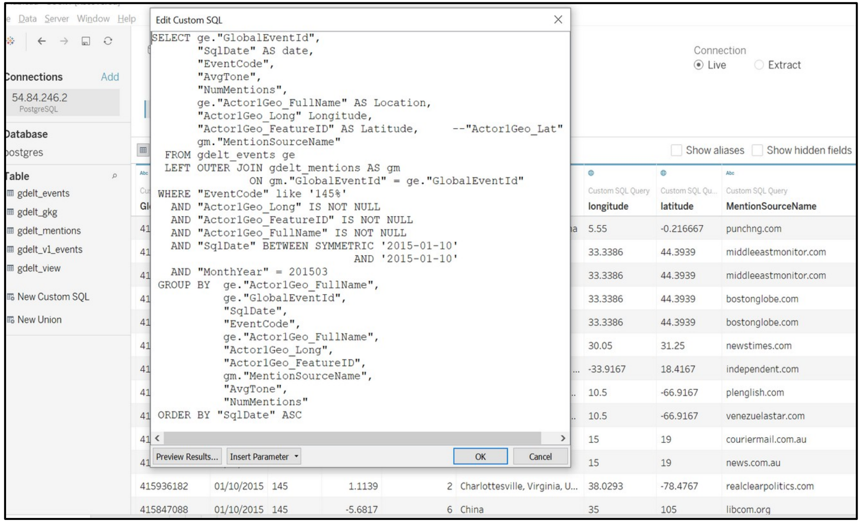
How much storage space is available?

How long is waiting time?

DWS Demo

- DWS Demo

Countries where violent protests occurred on January, 10 of 2015



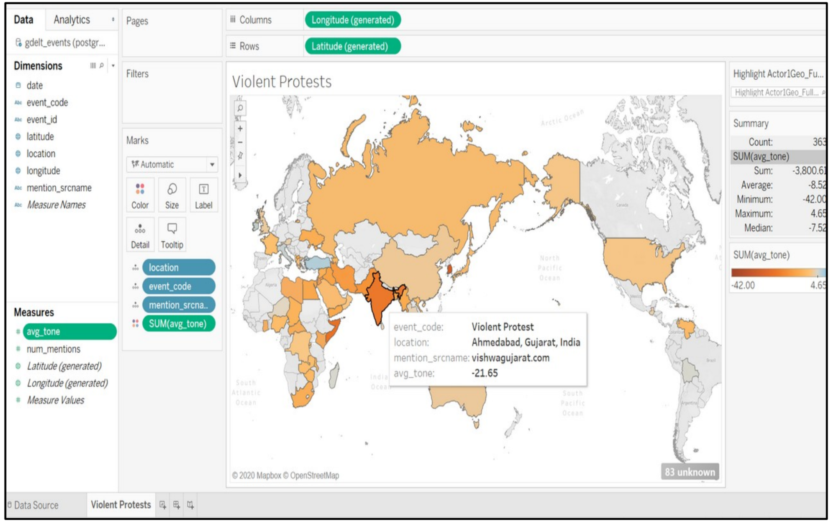
```
SELECT ge."GlobalEventId",
       "SqlDate" AS date,
       "EventCode",
       "AvgTone",
       "NumMentions",
       ge."Actor1Geo_FullName" AS Location,
       "Actor1Geo_Long" Longitude,
       "Actor1Geo_FeatureID" AS Latitude,     --"Actor1Geo_Lat"
       gm."MentionSourceName"
  FROM gdelt_events ge
  LEFT OUTER JOIN gdelt_mentions AS gm
               ON gm."GlobalEventId" = ge."GlobalEventId"
 WHERE "EventCode" like '145%'
   AND "Actor1Geo_Long" IS NOT NULL
   AND "Actor1Geo_FeatureID" IS NOT NULL
   AND "Actor1Geo_FullName" IS NOT NULL
   AND "SqlDate" BETWEEN SYMMETRIC '2015-01-10'
                               AND '2015-01-10'
   AND "MonthYear" = 201503
 GROUP BY  ge."Actor1Geo_FullName",
           ge."GlobalEventId",
           "SqlDate",
           "EventCode",
           ge."Actor1Geo_FullName",
           "Actor1Geo_Long",
           "Actor1Geo_FeatureID",
           gm."MentionSourceName",
           "AvgTone",
           "NumMentions"
 ORDER BY "SqlDate" ASC
```

Connection
● Live ○ Extract

□ Show aliases □ Show hidden fields

Gl						Custom SQL Query longitude	Custom SQL Qu... latitude	Custom SQL Query MentionSourceName
41					a	5.55	-0.216667	punchng.com
41						33.3386	44.3939	middleeastmonitor.com
41						33.3386	44.3939	middleeastmonitor.com
41						33.3386	44.3939	bostonglobe.com
41						33.3386	44.3939	bostonglobe.com
41						30.05	31.25	newstimes.com
41					...	-33.9167	18.4167	independent.com
41					..	10.5	-66.9167	plenglish.com
41					..	10.5	-66.9167	venezuelastar.com
41						15	19	couriermail.com.au
41						15	19	news.com.au
415936182	01/10/2015	145	1.1139	2	Charlottesville, Virginia, U...	38.0293	-78.4767	realclearpolitics.com
415847088	01/10/2015	145	-5.6817	6	China	35	105	libcom.org

Step 1. Get the data from DWS w/ Tableau



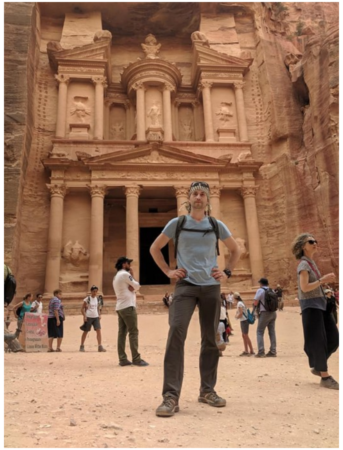
event_code: Violent Protest
location: Ahmedabad, Gujarat, India
mention_srcname: vishwagujarat.com
avg_tone: -21.65

Summary
Count: 363
SUM(avg_tone)
Sum: -3,800.61
Average: -8.52
Minimum: -42.00
Maximum: 4.65
Median: -7.52

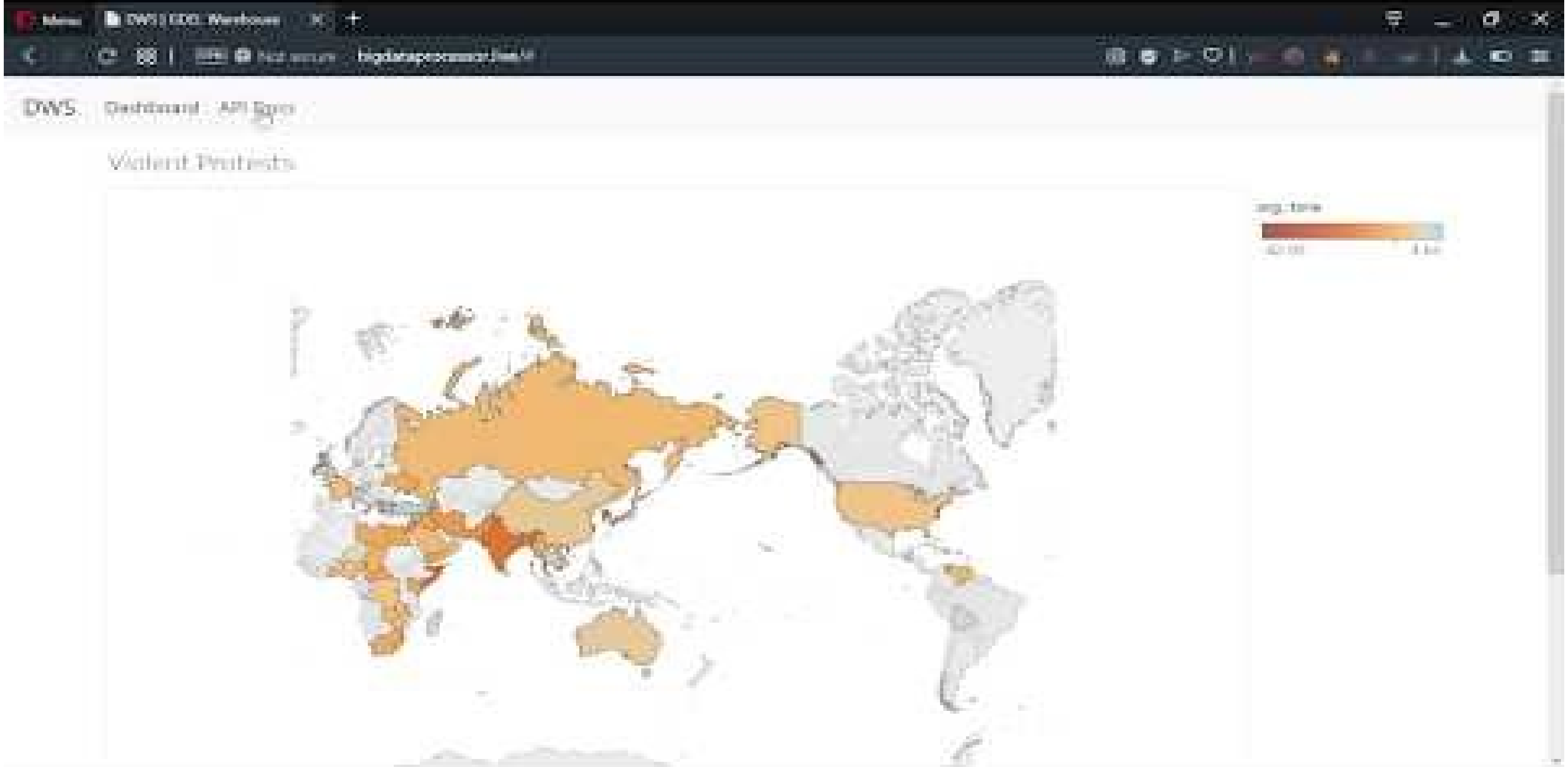
Step 2. Analyze and visualize the data

Artsiom Sinitski

- Has a professional background as software engineer and consultant in the supply chain management field
- Is a passionate traveller and photographer



https://www.instagram.com/artsiom_sinitski/



Violent Protests

Technology Stack Choices (1 of 2)

- Apache Spark vs. Flink
  - Fast batch processing
  - Diverse ecosystem of libraries (ML, GraphX, etc.)
- PostgreSQL vs. MySQL
  - Better SQL standard compliance
  - Large and active development community
  - Supports multiple CPUs and concurrent writes

Technology Stack Choices (2 of 2)

- Flask vs Django
  - Simple and lightweight web framework
  - Automatic API documentation (via Swagger)
- Airflow vs. Luigi
  - Strong and active development community
  - Scheduler support (“set it and forget it”)
  - Support for distributed execution
  - Intuitive web UI

Future Project Work

- Implement Service layer (data lineage) to give more visibility into the errors root cause in a data analytics process.
- Scale data warehouse storage automatically, as it grows
- Replace PostgreSQL with a columnar database (?)
  - Citus (PostgreSQL extension)
  - Presto

References

- Gdelt: Global data on events, location, and tone, 1979–2012

By Kalev Leetaru, Philip A. Schrodt

- GDELT Data Format Codebook