DWS

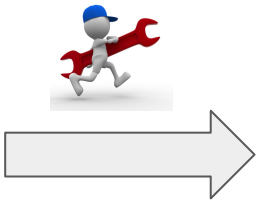
(Data Warehouse Solution for GDELT dataset)

Artsiom Sinitski
Data Engineering Fellow
Insight , New York City

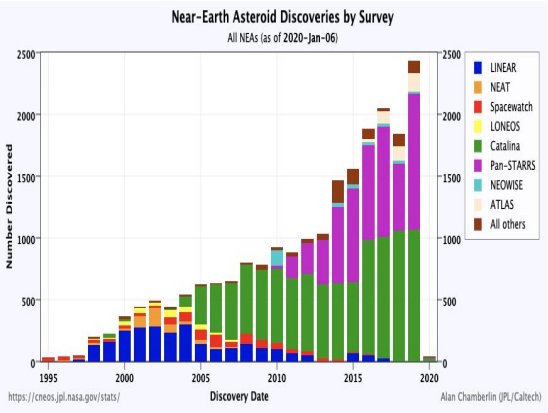
My Motivation

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

FROM THAT



DWS



TO THIS

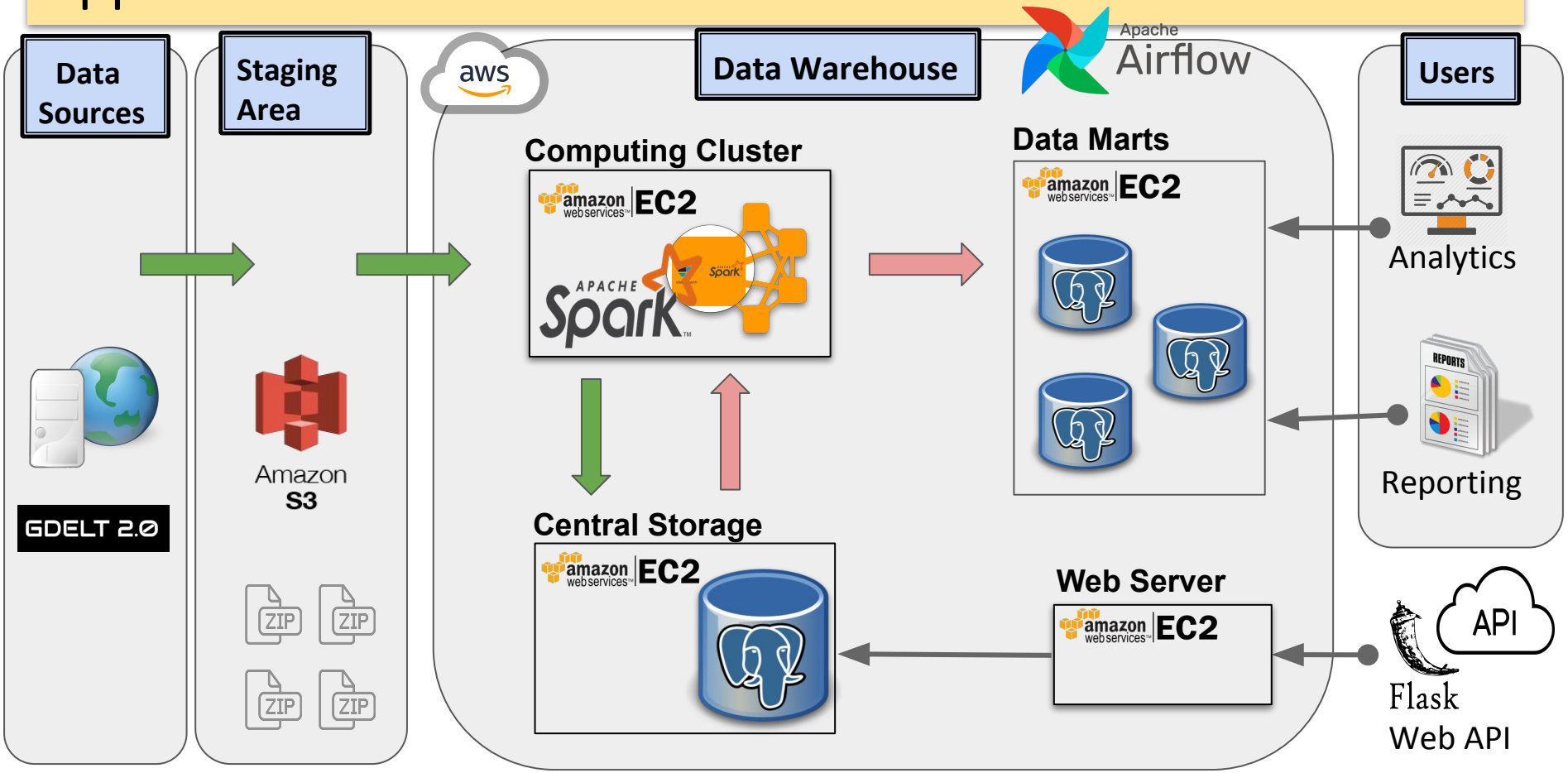
Why do we care?

- 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)

Approach


Data Sources
GDELT 2.0
Staging Area
Amazon S3
ZIP
ZIP
ZIP
ZIP
aws
Data Warehouse
Apache Airflow
Computing Cluster
amazon web services EC2
APACHE Spark
Data Marts
amazon web services EC2
Central Storage
amazon web services EC2
Web Server
amazon web services EC2
Users
Analytics
REPORTS
Reporting
API
Flask
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)

- <u>Central Storage ETL</u> ( ) - 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
- <u>Data Mart ETL</u> ( ) - 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!
  - Work with a subset of data (data marts)

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

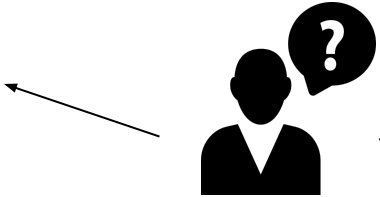
  - Also, consider denormalizing data and indexing tables
- Data redundancy vs. time complexity of O(m*n*k)

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?

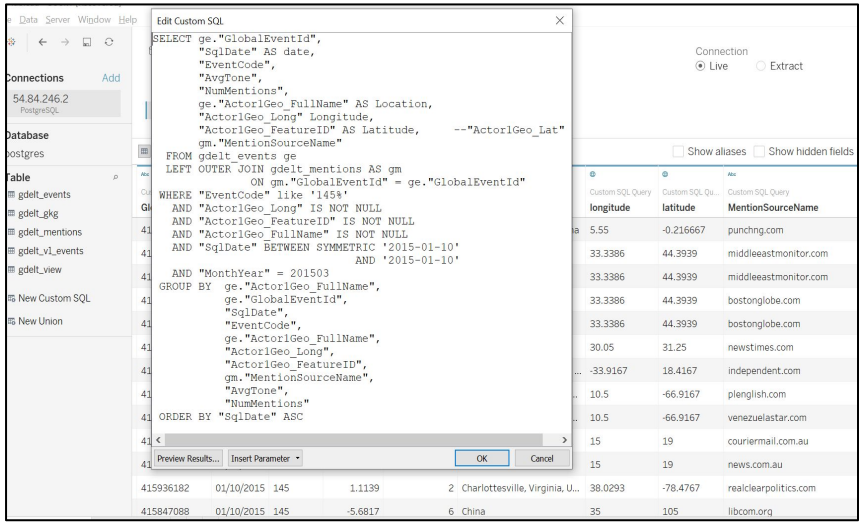


How long is waiting time?

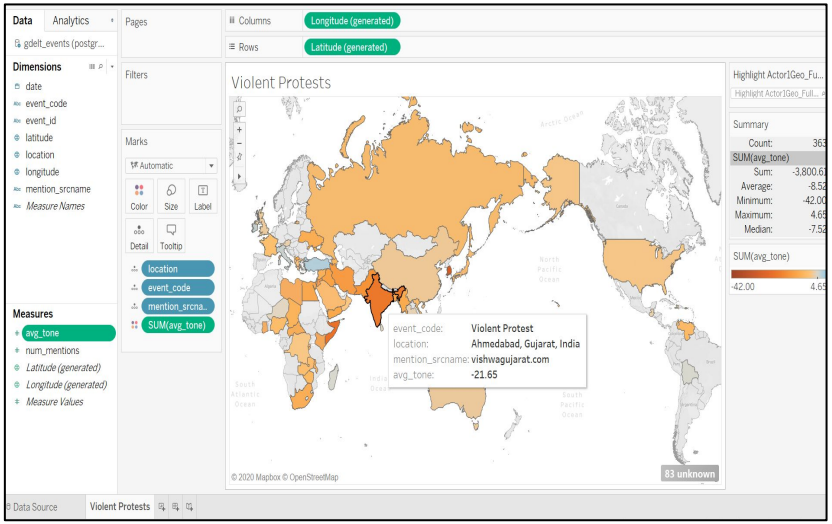
DWS Demo

- [DWS Demo]

Countries where violent protests occurred on January, 10 of 2015



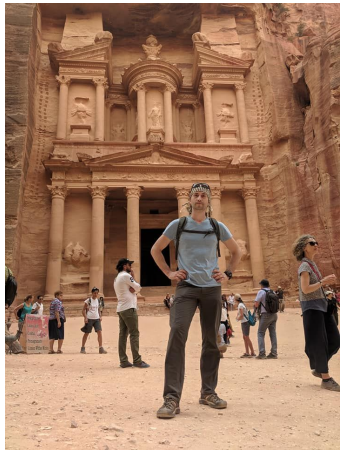
Step 1. Get the data from DWS w/ Tableau



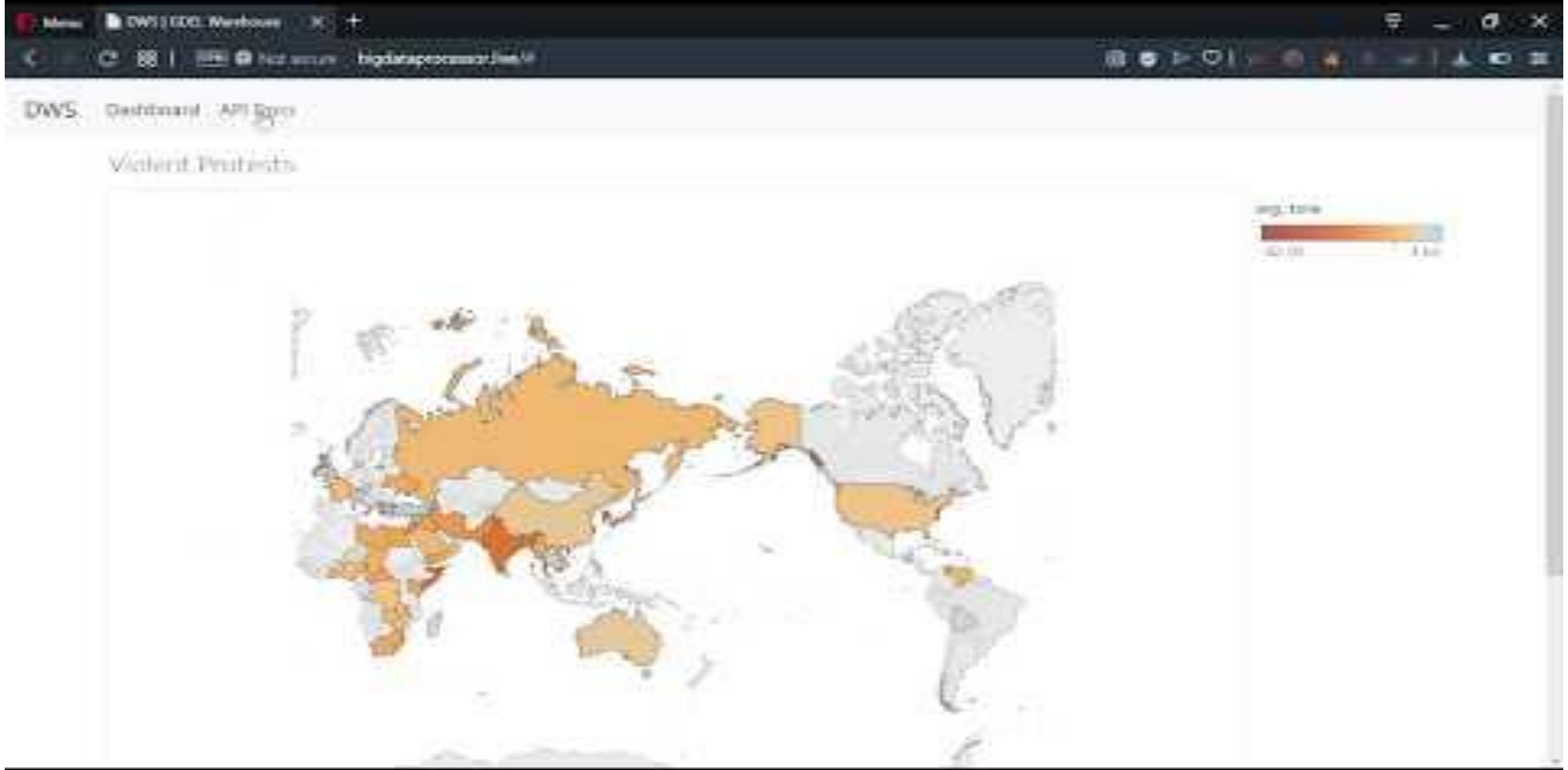
Step 2. Analyze and visualize the data

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


DWS
Dashboard
API Docs
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