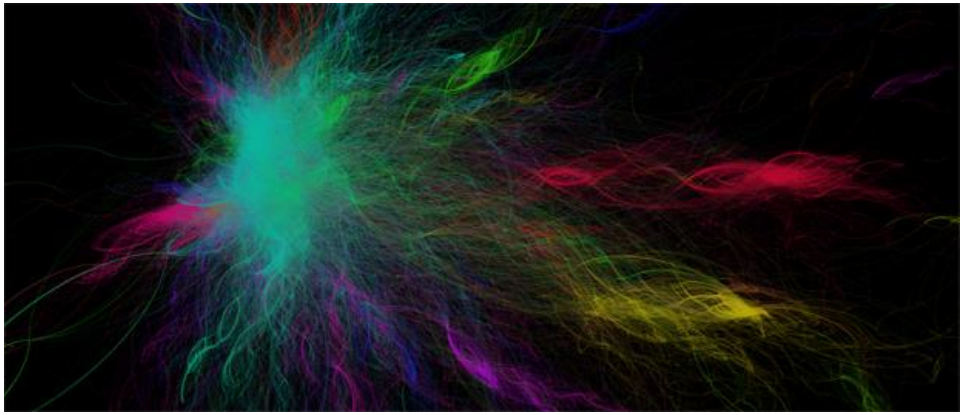
# MapReduce over GDELT Global Knowledge Graph dataset

*Special Topics (CIS 602-01)*



Javier Arechalde

# *Data*

- GDELT Global Knowledge Graph
- Compiles a list of every person, organization, company and location from news reports
- It also provides other features like tone
- Stored every 15 minutes since 2013
- In this project we will focus on the Tone and location

# *Scalability Challenges*

## Size of the data

- Total size of the dataset 2.2 TB
- Problem! Disk size limited to 2 TB
- Drop the columns that we wont use

## Different versions of the dataset

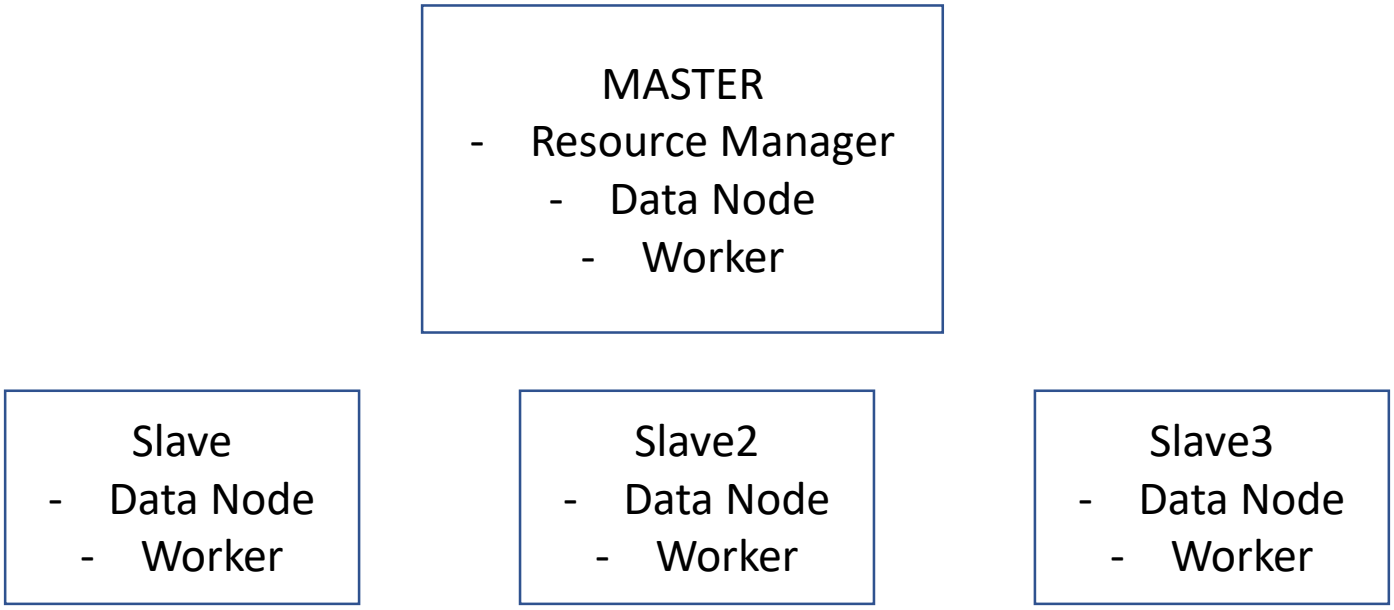
- Data was stored in different format
- In the end it wasn’t a problem

## Erroneous/missing data

- Some of the rows had missing locations or tone
- Used try-except method in Python

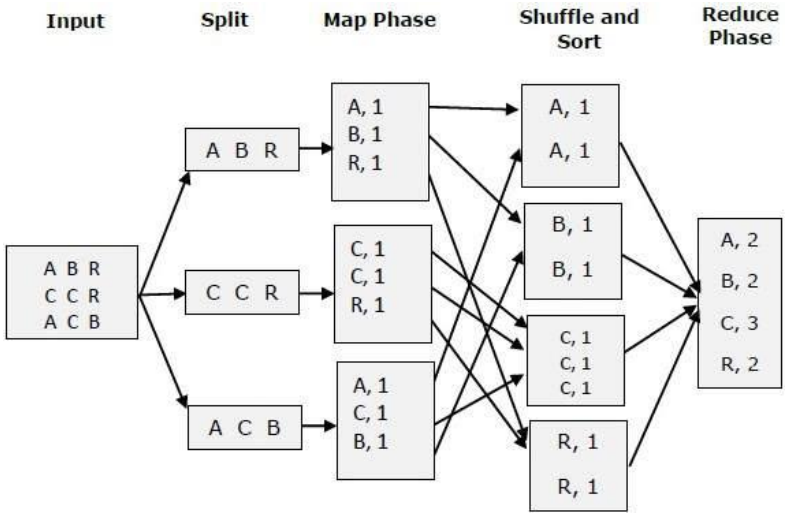
# *Setup*

- Hadoop Multi-node Cluster on Google Cloud Platform
  - 1 vCPU
  - 4 GB RAM
  - 500 GB Disk
  - Ubuntu

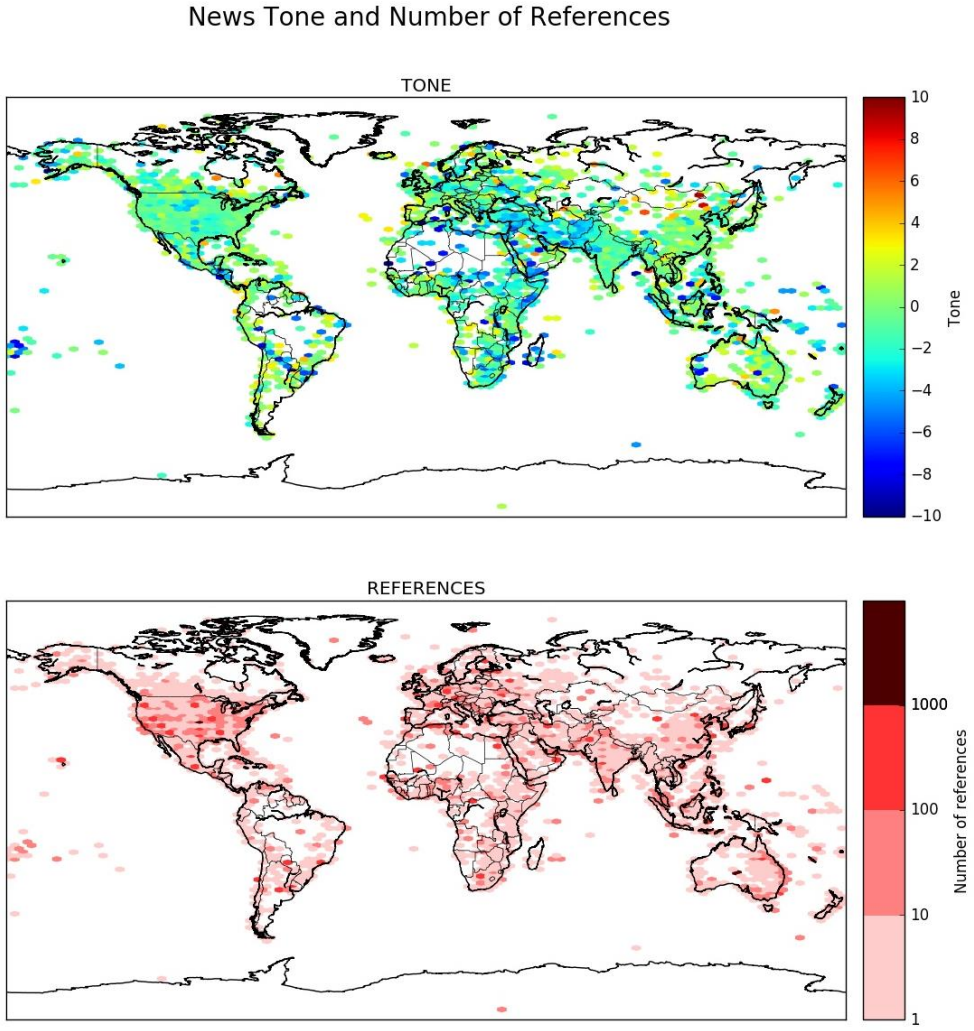


# *Setup*

- Main Function
  - Download files
  - Reduce files
  - Transfer files to HDFS → Data replicated in all nodes
  - MapReduce (YARN)
    - Implemented Mapper: Extracts all locations associated to each news
    - Implemented Reducer: Gets the count and avg. tone for each location



# *Results*


News Tone and Number of References
TONE
10
8
6
4
2
0
−2
−4
−6
−8
−10
Tone
REFERENCES
1000
100
10
1
Number of references

# QUESTIONS?