Seeking Cloud Efficiencies for Social Media Use Cases

Christine Kirkpatrick

Advisor: Dr. Inês Dutra (FCUP)

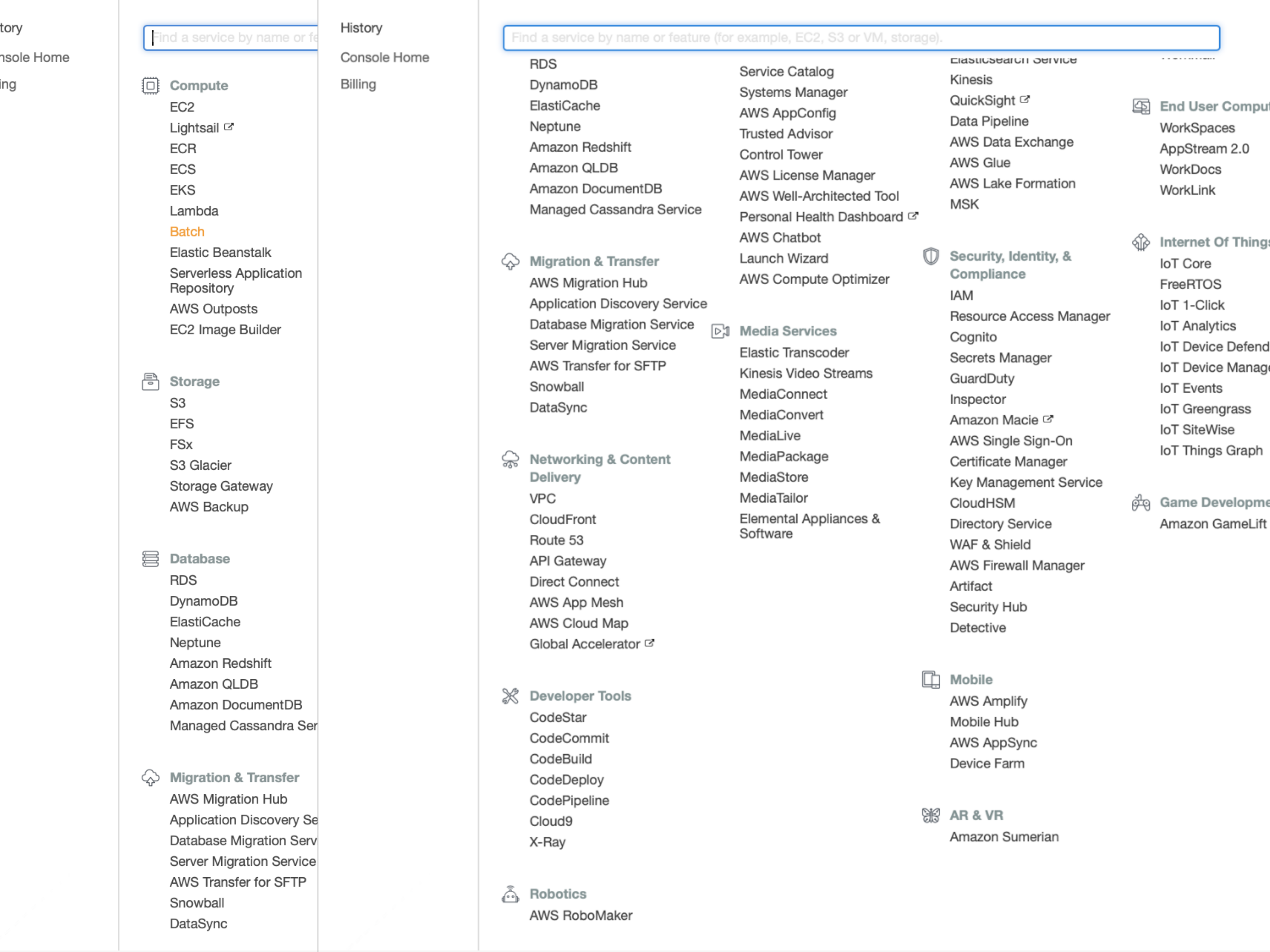
May 2020

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UNIVERSITY OF CALIFORNIA

CONTEXT & OBJECTIVE

- Gap in recommendations for experimental (research) data science loads in the cloud
- Potential efficiencies and optimization points
  - Hardware architecture
  - Software
  - Within applications
- Make better use of finite resources (time, money, CO_2 emissions)



tory

nsole Home

ing

Find a service by name or fe

Compute
- EC2
- Lightsail
- ECR
- ECS
- EKS
- Lambda
- Batch
- Elastic Beanstalk
- Serverless Application Repository
- AWS Outposts
- EC2 Image Builder

Storage
- S3
- EFS
- FSx
- S3 Glacier
- Storage Gateway
- AWS Backup

Database
- RDS
- DynamoDB
- ElastiCache
- Neptune
- Amazon Redshift
- Amazon QLDB
- Amazon DocumentDB
- Managed Cassandra Ser

Migration & Transfer
- AWS Migration Hub
- Application Discovery Se
- Database Migration Serv
- Server Migration Service
- AWS Transfer for SFTP
- Snowball
- DataSync

History

Console Home

Billing

Find a service by name or feature (for example, EC2, S3 or VM, storage).

- RDS
- DynamoDB
- ElastiCache
- Neptune
- Amazon Redshift
- Amazon QLDB
- Amazon DocumentDB
- Managed Cassandra Service

Migration & Transfer
- AWS Migration Hub
- Application Discovery Service
- Database Migration Service
- Server Migration Service
- AWS Transfer for SFTP
- Snowball
- DataSync

Networking & Content Delivery
- VPC
- CloudFront
- Route 53
- API Gateway
- Direct Connect
- AWS App Mesh
- AWS Cloud Map
- Global Accelerator

Developer Tools
- CodeStar
- CodeCommit
- CodeBuild
- CodeDeploy
- CodePipeline
- Cloud9
- X-Ray

Robotics
- AWS RoboMaker

- Service Catalog
- Systems Manager
- AWS AppConfig
- Trusted Advisor
- Control Tower
- AWS License Manager
- AWS Well-Architected Tool
- Personal Health Dashboard
- AWS Chatbot
- Launch Wizard
- AWS Compute Optimizer

Media Services
- Elastic Transcoder
- Kinesis Video Streams
- MediaConnect
- MediaConvert
- MediaLive
- MediaPackage
- MediaStore
- MediaTailor
- Elemental Appliances & Software

- Elasticsearch Service
- Kinesis
- QuickSight
- Data Pipeline
- AWS Data Exchange
- AWS Glue
- AWS Lake Formation
- MSK

Security, Identity, & Compliance
- IAM
- Resource Access Manager
- Cognito
- Secrets Manager
- GuardDuty
- Inspector
- Amazon Macie
- AWS Single Sign-On
- Certificate Manager
- Key Management Service
- CloudHSM
- Directory Service
- WAF & Shield
- AWS Firewall Manager
- Artifact
- Security Hub
- Detective

Mobile
- AWS Amplify
- Mobile Hub
- AWS AppSync
- Device Farm

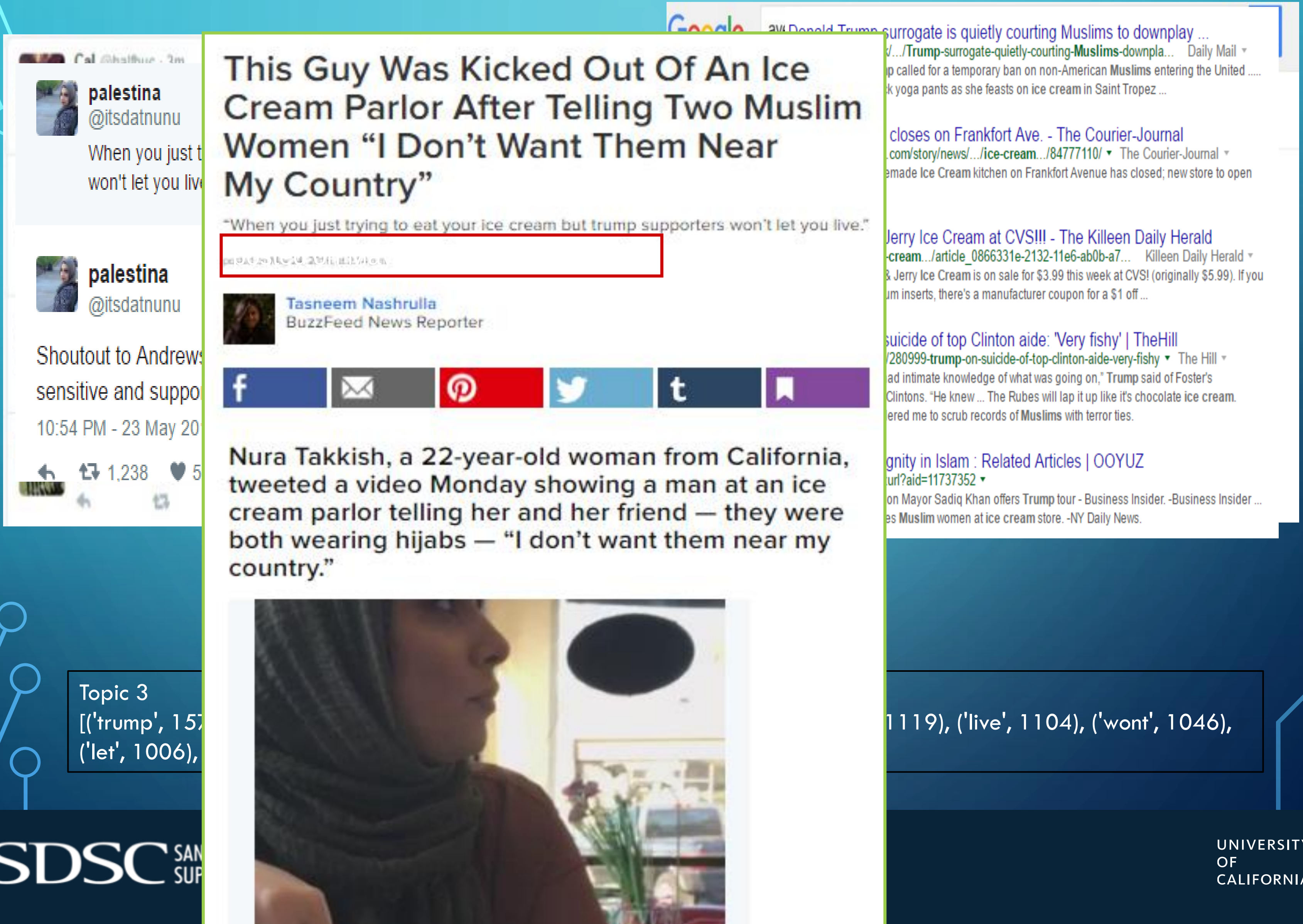
AR & VR
- Amazon Sumerian

End User Comput
- WorkSpaces
- AppStream 2.0
- WorkDocs
- WorkLink

Internet Of Thing
- IoT Core
- FreeRTOS
- IoT 1-Click
- IoT Analytics
- IoT Device Defend
- IoT Device Manag
- IoT Events
- IoT Greengrass
- IoT SiteWise
- IoT Things Graph

Game Developme
- Amazon GameLift

EVENT DETECTION WITH TWITTER DATA



Cal @halfhue · 3m

palestina
@itsdatnunu

When you just t
won't let you liv

palestina
@itsdatnunu

Shoutout to Andrews
sensitive and suppo
10:54 PM - 23 May 20

1,238 5

This Guy Was Kicked Out Of An Ice Cream Parlor After Telling Two Muslim Women "I Don't Want Them Near My Country"

"When you just trying to eat your ice cream but trump supporters won't let you live."

[illegible]

Tasneem Nashrulla
BuzzFeed News Reporter

Nura Takkish, a 22-year-old woman from California, tweeted a video Monday showing a man at an ice cream parlor telling her and her friend — they were both wearing hijabs — "I don't want them near my country."

Google

Donald Trump surrogate is quietly courting Muslims to downplay ...
/.../Trump-surrogate-quietly-courting-Muslims-downpla... Daily Mail
p called for a temporary ban on non-American Muslims entering the United .....
k yoga pants as she feasts on ice cream in Saint Tropez ...

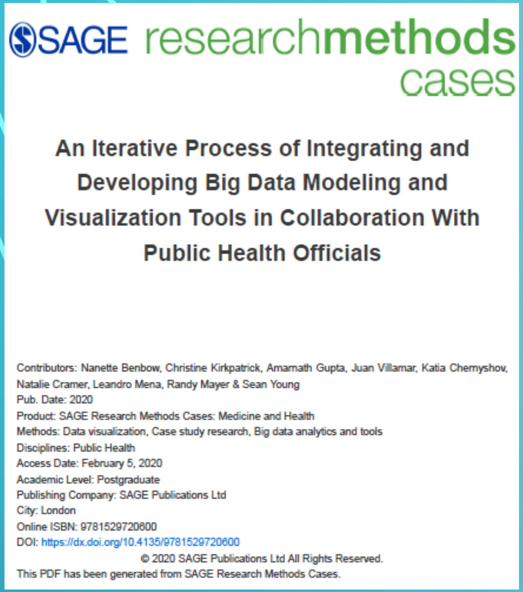
closes on Frankfort Ave. - The Courier-Journal
.com/story/news/.../ice-cream.../84777110/ The Courier-Journal
emade Ice Cream kitchen on Frankfort Avenue has closed; new store to open

Jerry Ice Cream at CVS!!! - The Killeen Daily Herald
-cream.../article_0866331e-2132-11e6-ab0b-a7... Killeen Daily Herald
& Jerry Ice Cream is on sale for $3.99 this week at CVS! (originally $5.99). If you
um inserts, there's a manufacturer coupon for a $1 off ...

suicide of top Clinton aide: 'Very fishy' | TheHill
/280999-trump-on-suicide-of-top-clinton-aide-very-fishy The Hill
ad intimate knowledge of what was going on," Trump said of Foster's
Clintons. "He knew ... The Rubes will lap it up like it's chocolate ice cream.
ered me to scrub records of Muslims with terror ties.

gnity in Islam : Related Articles | OOYUZ
url?aid=11737352
on Mayor Sadiq Khan offers Trump tour - Business Insider. -Business Insider ...
es Muslim women at ice cream store. -NY Daily News.

Topic 3
[('trump', 157 1119), ('live', 1104), ('wont', 1046),
('let', 1006),



SAGE researchmethods cases

An Iterative Process of Integrating and Developing Big Data Modeling and Visualization Tools in Collaboration With Public Health Officials

Contributors: Nanette Benbow, Christine Kirkpatrick, Amarnath Gupta, Juan Villamar, Katia Chernyshov, Natalie Cramer, Leandro Mena, Randy Mayer & Sean Young
Pub. Date: 2020
Product: SAGE Research Methods Cases: Medicine and Health
Methods: Data visualization, Case study research, Big data analytics and tools
Disciplines: Public Health
Access Date: February 5, 2020
Academic Level: Postgraduate
Publishing Company: SAGE Publications Ltd
City: London
Online ISBN: 9781529720600
DOI: https://dx.doi.org/10.4135/9781529720600
© 2020 SAGE Publications Ltd All Rights Reserved.
This PDF has been generated from SAGE Research Methods Cases.

TWITTER TO SUPPLEMENT PUBLIC HEALTH SURVEILLANCE

- Twitter API - #condom = #condom* not "#condom"
- Parameters: #condom includes #condomínio if language=all

USING EMOJIS TO CLASSIFY TWEETS & PROFILES

Research Plan

1. Compile 6 months of tweets into a database

   ETA: 5.5 months!
2. Develop classificati
3. Create training set
4. Machine learning
5. Refine criteria
6. Integrate to social r ance

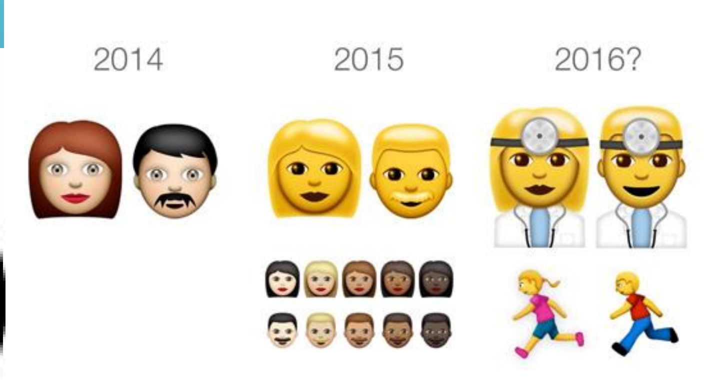


Image source: https://blog.emojipedia.org/unicode-and-the-emoji-gender-gap/

‘BIG DATA’ CHALLENGES

- Streaming data
- Ingestion failure
- Supporting polystores
- Many researchers/teams expect SQL
- Big data platforms out of reach ($$)

DESIGN & METHOD: SPEEDING UP DATA INGESTION

- Classifying tweets → data ingestion refinement
- Start at the beginning
  - PostgreSQL database
  - Twitter’s developer schema
- Keep obvious improvements
  - RAM disk on cloud instance handling ingestion

Heron Tweet API

xe.large

- 2 vCPUs
- 64GB (32 as RAM disk)

Tweet pre-processing

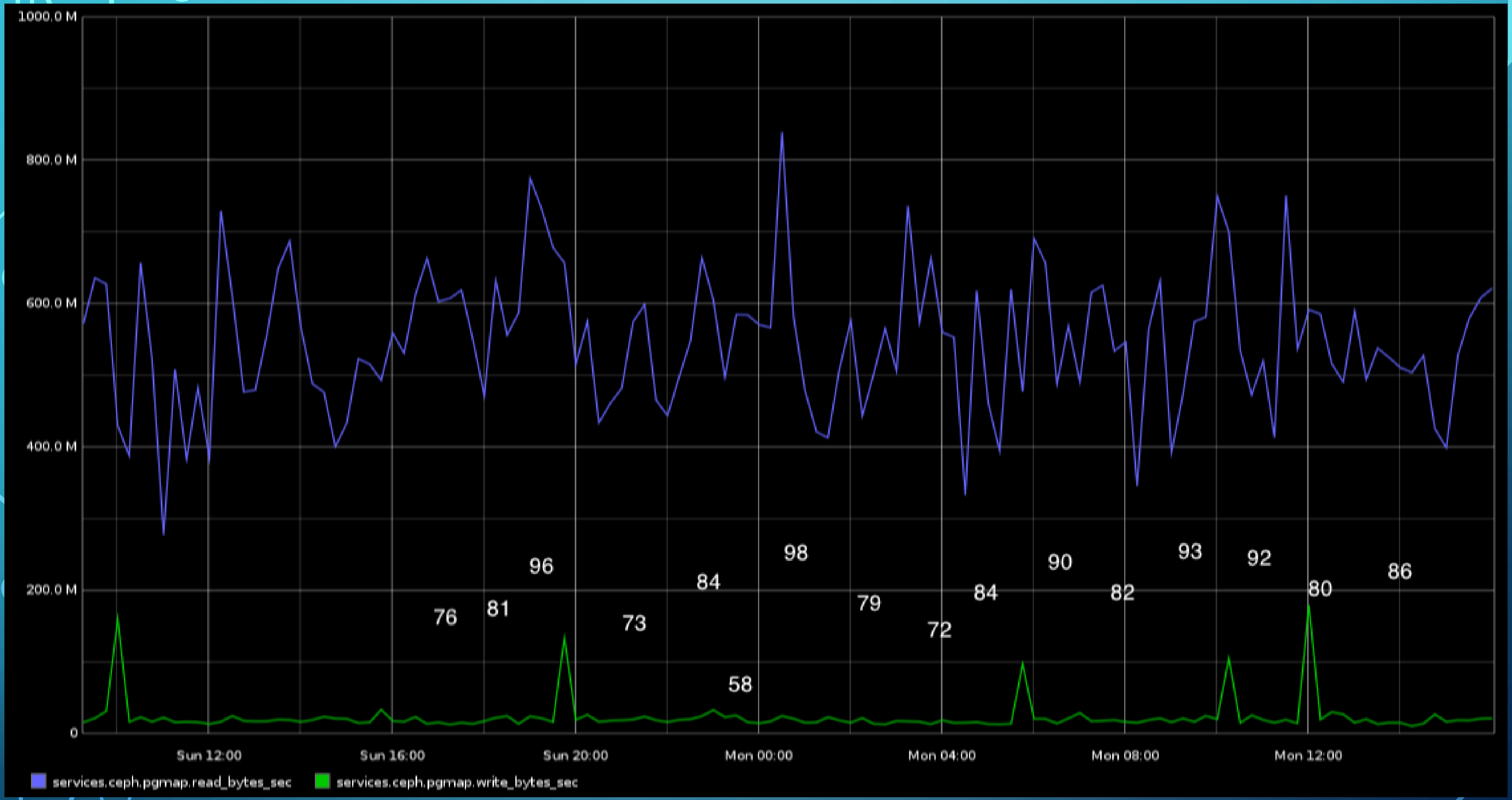
m1.medium

- 1 vCPU

Postgres DB

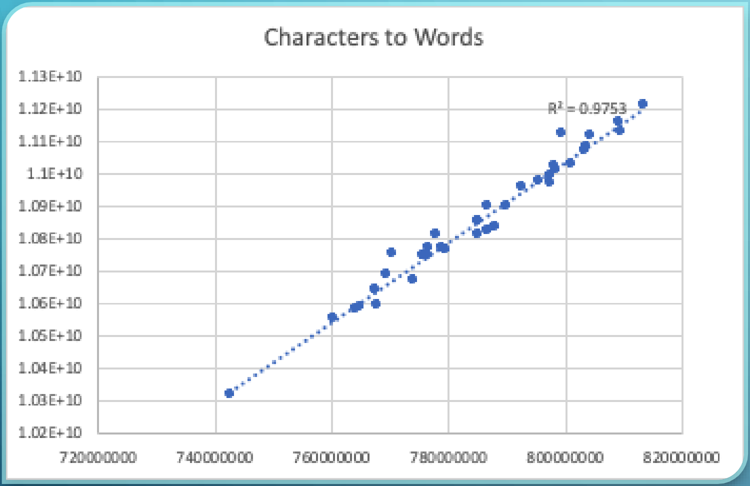
R.1xlarge

- 4 vCPUs
- 32 GB RAM



	Writes (with outliers removed)	Writes (raw)	Reads
Mean	18,799,954	22,726,801.95	544,754,417
STDEV	8,456,525	23,998,263.25	96,574,868
Coefficient of variation	45%	106%	18%

ASSUMPTION CHECK: DATA UNIFORMITY & PRE-PROCESSING



- Tweets written to 2.5M record files
- Duplicate keys – pre-processing opportunity?

CONLCUSION: THE KEY IS THE KEY

- Speeding ingestion → No more duplicate keys
- Auto-incrementing integer = 3x quicker

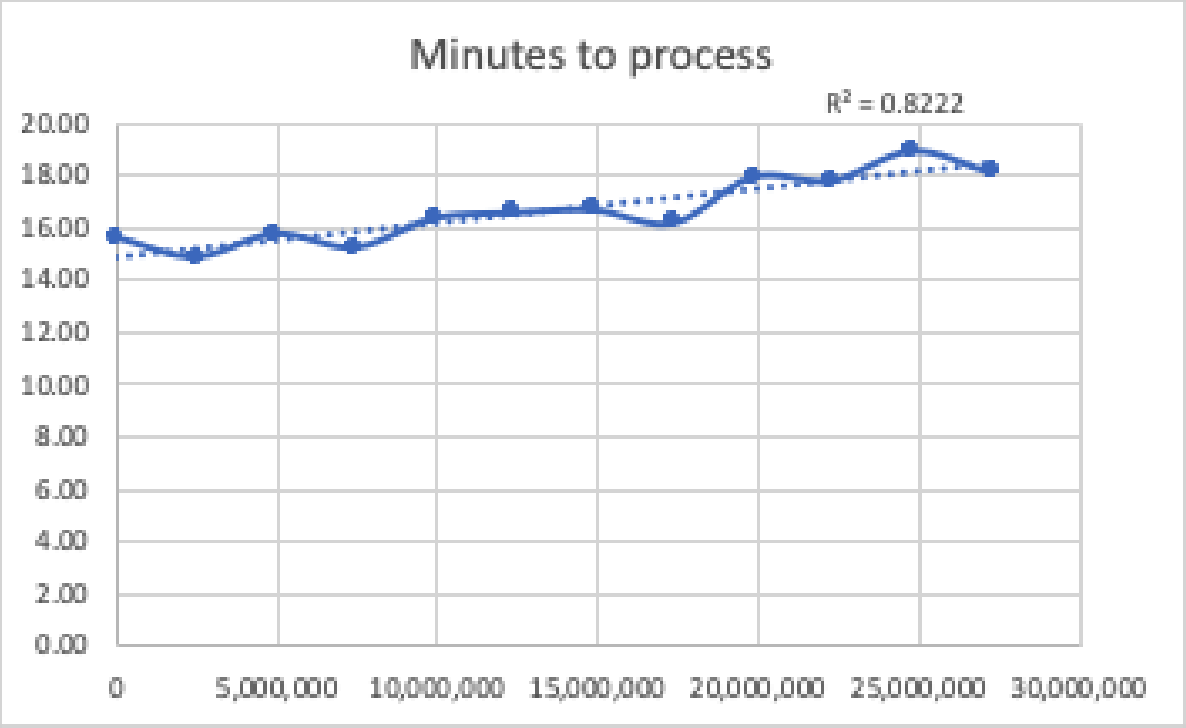
ALTER TABLE ONLY public.tweet ADD CONSTRAINT tweet_id_pk PRIMARY KEY (tweet_id);

CREAT

CREAT

were r

ALTER



‘TIME TO SCIENCE’ PAIN POINTS

Request: Add more GPUs and SSDs
Solution: *Adjust database schema*

Request: Add dedicated SSDs to Apache Solr nodes
Solution: *Fix Java error in web UI*

… still searching for issues to solve with cloud architecture solutions!

POWER, MONEY, CO_2 SAVED FROM SCHEMA TWEAK

Primary Key	Time (hrs)	Compute Cost	Storage	Power (kW)	Power Cost	Total Cost	Savings	kW Saved
Tweet ID	678.7	$433.02 US €399.72	$42.76 US €39.47	27.1	$4.76 US €4.39	$480.54 US €443.59		
Integer	195.5	$124.74 US €115.15	$12.32 US €11.37	7.8	$1.37 US €1.26	$138.44 US €127.79	$342.11 US €315.80	19.3

- Saved equivalent of burning ~7 kg of coal or
- Charging your cellphone 1,740 times

FUTURE WORK

Skip SQL databases (provide SQL-like UIs)

Amazon S3 Select - open source equivalent?

Work through pain points on cloud loads for architecture recommendations

Study filesystem impact on clouds.

THANK YOU!

CHRISTINE@SDSC.EDU
@SUPERCHRISTINEK