Building a Data Anonymization Platform on AWS

Laszlo Török, Engineering Lead, Telefónica NEXT

Özkan Can, Senior Solutions Architect, Amazon Web Services
Cloud-native application storage

Özkan Can,
Senior Solutions Architect,
Amazon Web Services
Agenda

1. Why cloud-native?
2. What is cloud-native?
3. What does this mean for Storage?
Why cloud-native?
Why cloud-native?

1. Reliability
2. Frugality
3. Security
4. Elasticity & scale
5. Performance
6. Better design
SO WHAT DOES THE FUTURE LOOK LIKE?

ALL THE CODE YOU EVER WRITE IS BUSINESS LOGIC
So what is cloud-native or modern?
Paths to cloud-native

Assess and prioritize, app by app

Pick path to modernization

Lift & shift: data center → EC2

Re-platform: VMs → containers

Refactor: monolith → microservices

Re-invent: host fleets → serverless
Traditional app. example, in-cloud

Application Load Balancer

Amazon EC2 Auto Scaling

Amazon Aurora MySQL
Traits of modern applications

Security in every layer.

Built to scale on demand.

Continously integrated and deployed.

Microservice-oriented and API-backed.

Leveraging purpose-built databases and cloud-native storage options.
What this means for storage
File vs Block vs Object

**Block Storage**
*Raw Storage*
Data organized as an array of unrelated blocks
Host File System places data on disk
e.g.: Microsoft NTFS, Unix ZFS

**File Storage**
Unrelated data blocks managed by a file (serving) system
Native file system places data on disk

**Object Storage**
Stores Virtual containers that encapsulate the data, data attributes, and metadata
API Access to data
Metadata Driven, Policy-based, etc
File vs Block vs Object

File
- Amazon Elastic File System
- Amazon FSx for Lustre
- Amazon FSx for Windows File Server

Block
- Amazon Elastic Block Store (EBS)

Object
- Amazon Simple Storage Service (S3)
- Amazon S3 Glacier
- Amazon FSx for Windows File Server
Storage options for cloud-native applications

Amazon EC2
Amazon EFS
AWS Storage Gateway Family
Amazon FSx for Windows File Server
Amazon EBS
Amazon S3
Amazon FSx for Lustre

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Storage Services on AWS

- Amazon Elastic Block Store (EBS)
- Amazon Elastic File System
- Amazon FSx for Lustre
- Amazon FSx for Windows File Server
- Amazon Simple Storage Service (S3)
- Amazon S3 Glacier
- AWS Snowball
- AWS Snowball Edge
- AWS Snowmobile
- AWS Backup
- AWS Storage Gateway
Magic Quadrant for Public Cloud Storage Services, Worldwide – 2018

Positioned furthest for completeness of vision and highest for ability to execute in each report since inception in 2014

Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner’s research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.
Benefits of Amazon S3

- Website hosting
- Media Master files
- Big Data
- File Sharing
- Data Analytics
- Mobile sync & backup
- Disaster Recover
- Archive
- Content Distribution
- Re-creatable data
- Backup & Restore
- Dynamic Websites
- Enterprise applications
- Manageability
- Performance
- Security
- Scalability
- Durability
- Availability
Amazon Simple Storage Service (S3)

**Collect**
- The most ways to move data in/out
- Security that helps the CISO
- Automated cost reduction tools

**Store**
- Designed for 99.999999999% durability
- Unmatched security and compliance capabilities
- Replication options across regions

**Analyze**
- On-demand analytics
- Built-in support for SQL expressions with S3 Select
- Detailed data on usage patterns and access
Your choice of Amazon S3 storage classes

- **S3 Standard**
  - Active, frequently accessed data
  - Milliseconds access
  - > 3 AZ
  - From: $0.0210/GB
  - Monitoring fee per obj.
  - Min storage duration

- **S3 Intelligent-Tiering (NEW!)**
  - Data with changing access pattern
  - Milliseconds access
  - > 3 AZ
  - From: $0.0210 to $0.0125/GB
  - Retrieval fee per GB
  - Min storage duration
  - Min object size

- **S3 Standard-IA**
  - Infrequently accessed data
  - Milliseconds access
  - > 3 AZ
  - From: $0.0125/GB
  - Min storage duration
  - Min object size

- **S3 One Zone-IA**
  - Re-creatable less accessed data
  - Milliseconds access
  - 1 AZ
  - From: $0.0100/GB
  - Retrieval fee per GB
  - Min object size

- **S3 Glacier (NEW!)**
  - Archive data
  - Hours access
  - > 3 AZ
  - From: $0.0040/GB
  - Retrieval fee per GB
  - Min object size

- **S3 Glacier Deep Archive**
  - Archive data
  - Hours access
  - > 3 AZ
  - From: $0.00099/GB
  - Retrieval fee per GB
  - Min storage duration
  - Min object size

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Data Lakes from AWS

- Machine Learning
- Analytics
- Data Lake on AWS
- On-premises Data Movement
- Real-time Data Movement

Open and comprehensive
Secure
Scalable and durable
Lowest cost
Tiered Storage to Optimize Price/Performance

Lowest Cost

• Tiered storage to optimize price/performance
  • S3 Standard
  • S3 Standard—Infrequent Access
  • S3 One Zone—Infrequent Access
  • Amazon Glacier

• Migrate between tiers based on lifecycle policies
• Store data at $0.023/GB/month with S3
• Store data at $0.004/GB/month with Glacier
Building a Data Anonymization Platform on AWS

Laszlo Török, Engineering Lead, Telefónica NEXT
Who we are, what we do
We are part of the global telco group Telefónica

Telefónica Germany’s network has the most customer lines
An average day of the Telefónica Germany Network

45M
Customer Lines

5B+
Network Events
Very valuable data

Help predict capacities in regional transport in Berlin-Brandenburg in the ProTrain project, supported by the German Ministry of Transport

Model city traffic for special events like football matches with partner Intraplan

Retail applications: store location planning
So bewegt sich Deutschland

Die neue Datenbasis für nachhaltige Mobilität: anonymisierte Mobilfunkdaten

https://next.telefonica.de/so-bewegt-sich-deutschland
Public Viewing zur Fußball-WM 2018
Analysiert mit anonymisierten Mobilfunkdaten von Telefónica NEXT

Einzugsgebiete

In welcher Stadt kamen die meisten Besucher?


Wie alt waren die meisten Besucher?

- 25-29 Jahre: 14%
- 30-34 Jahre: 13%
- 35-39 Jahre: 12%

Kamen mehr Frauen oder Männer?

- Frauen: 44%
- Männer: 56%

Besuchsaufkommen über den Tag verteilt - Die Spielzeit ist in Orange hervorgehoben

Woher kamen die meisten Touristen?

- USA
- Vereinigtes Königreich
- Spanien
- Vereinigtes Königreich
- Polen
- Österreich
- USA
- Vereinigtes Königreich
- Polen
Very sensitive data

Protected by German privacy laws (TMG, BDSG)

We are committed to protect privacy and give users control over their data
Valuable insights vs Individual privacy?

Most value comes from aggregate **analysis of groups, not individuals.**
Telefónica’s Data Anonymization Platform
Building a scalable data storage in the cloud
Context

Secure Storage

Stream + Batch

Analyst Group

Batch

Analyst Group

Stream

Analyst Group
Requirements for our Data Hub

- Allow streaming and bulk access to event data
- Cost-effective long-term storage
- Flexible access management & secure storage
- Low maintenance / ops overhead
Do it yourself on-premise?
Fully programmable provisioning
Can we do better?

1. Events pulled, by pulling time intervals of interest

2. Once stored, data is accessed infrequently

3. Retrieval latency (mins, hours) acceptable
Optimizing S3 storage costs

Sort into hour sized buckets {yyyy}/{mm}/{DD}/{HH}/
First month of production

Top 5 items on our end of month AWS bills

1. Lambda
2. Kinesis
3. Cloudwatch (?!?)
4. S3
5. KMS (~ on par with S3)
Use compression – fewer Kinesis shards needed
What went wrong?

~1000s invocations / sec, constantly

PROTIP: Too many Lambda invocations can get costly
Data Hub v2 – Fargate to the rescue

NEXT Data Hub
- Inbound Kinesis Stream
  - triggers
  - writes to
    - Transformation Step
      - writes to
        - Analyst Kinesis Stream
- Intermediary S3 bucket
  - concatenated and compressed by
  - writes to
    - Batch Archiver
      - writes to
        - Long-term archive in S3/Glacier

Anonymized event data stream

Real-time access
On-demand bulk access

Data Analyst
Access Management via IAM

Data Hub Account

Analyst IAM Role

Analysts' Account

Analyst User Account

Local Admin

PROTIP: Delegate user management if possible
Monitoring via CloudWatch
Recap
Lessons learnt

- On-demand, near-instant, programmatic provisioning of IT resources = Superpower
- More fully-managed service = Less DIY ops
- Using fully managed services not always lower TCO
- Optimizing for data access patterns can bring significant €€€ savings
Thank you!

We are hiring!

László Török
Engineering Lead, Big Data Privacy Services
laszlo.toeroek1@telefonica.com
https://next.telefonica.de
@telefonicaNEXT
linkedin.com/company/telefónicaNEXT
Please complete the session survey.