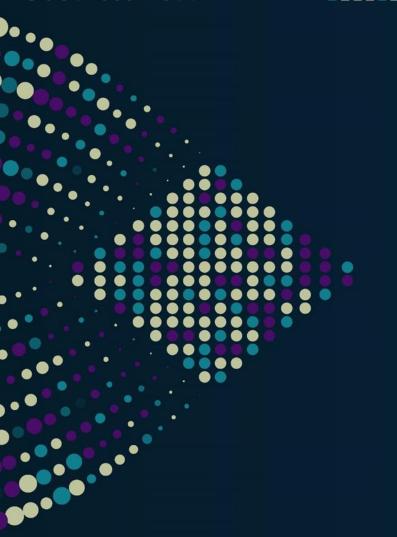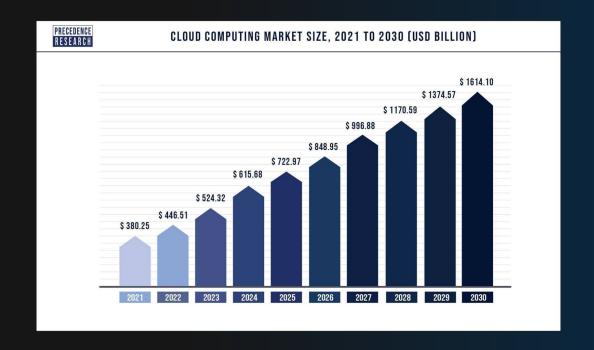# Inside the cloud: A Dive Deep into Cold Data Archiving

Vikranth Etikyala
Senior Staff Software Engineer, SoFi
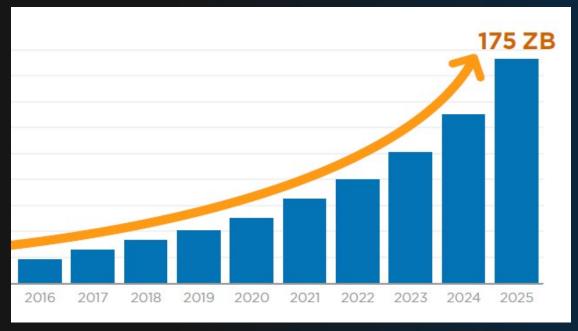
Compute Growth

CLOUD COMPUTING MARKET SIZE, 2021 TO 2030 (USD BILLION)

PRECEDENCE RESEARCH

Year	Value
2021	$ 380.25
2022	$ 446.51
2023	$ 524.32
2024	$ 615.68
2025	$ 722.97
2026	$ 848.95
2027	$ 996.88
2028	$ 1170.59
2029	$ 1374.57
2030	$ 1614.10

SDC 23

Storage Growth

Source: Precedence Reserach Dec 2022

ZetaByte

| Megabytes | Gigabytes | Terabytes | Petabytes | Exabytes | Zettabytes | Yottabytes (1×10^{24}) | Xenotabytes (1×10^{27}) | Shilentobytes (1×10^{30})..... |

1.0 ZB

~1,000 EB

~1,000,000 PB

~1,000,000,000 TB

~1,000,000,000,000 GB

~1,000,000,000,000,000 MB

~ **55,000,000 LTO-9** tapes or 18.0 TB HDDs

Cold Data

- Infrequently accessed
- Not required real time
- Lower priority
- Older or historical data

Why so Cold?

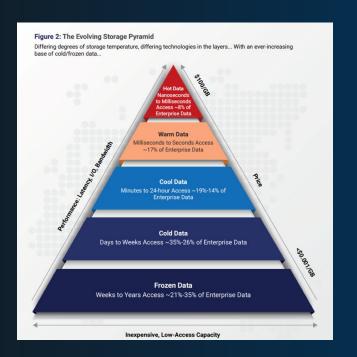- Regulatory Compliance - Banks

- Audit and Legal - Governments

- Historical Records - Genomic Data

- Disaster Recovery - Backup and Restore

- AI/ML Models - ChatGPT

Missing or Deleted Data will most likely be the data you need

Cold Data - Archival and Benefits

- Cost

- Scalability

- Performance

Figure 2: The Evolving Storage Pyramid
Differing degrees of storage temperature, differing technologies in the layers... With an ever-increasing base of cold/frozen data...

Hot Data
Nanoseconds to Milliseconds Access ~8% of Enterprise Data

Warm Data
Milliseconds to Seconds Access ~17% of Enterprise Data

Cool Data
Minutes to 24-hour Access ~19%-14% of Enterprise Data

Cold Data
Days to Weeks Access ~35%-26% of Enterprise Data

Frozen Data
Weeks to Years Access ~21%-35% of Enterprise Data

$100/GB

Price

$0.001/GB

Performance: Latency, I/O, Bandwidth

Inexpensive, Low-Access Capacity

Archival Solutions on Cloud

- AWS S3 - Glacier
- Azure Archive Storage
- Google Cloud Storage
- Oracle Archive Storage
- IBM Cloud Object Storage

Cloud Archival Benefits - First Order

- Serverless

- Scalable and Elastic

- Native Cloud Integrations

- Lifecycle - Auto Transition to Archives - AI/ML, Date Based

- Security

- Service Level Agreements

Cloud Archival Benefits - Second Order

- Data Center Management - Host, Racks, Temperature
- Energy - Carbon Footprint
- Capacity Management
- Operations Team
- Distributed System Problems
- Storage Media Advancements

Example: S3 Scale

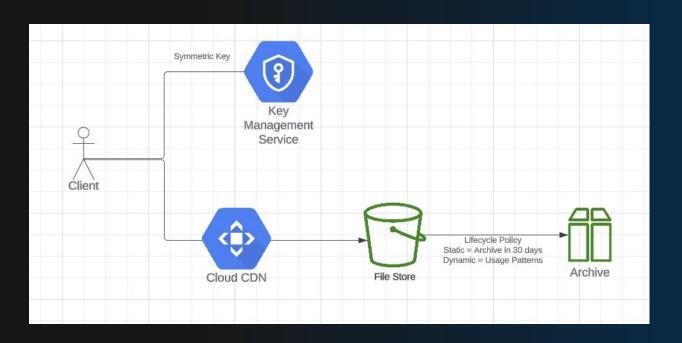|	
Capacity and throughput	Amazon S3 holds more than **280 trillion objects** and averages over **100 million requests per second**
Events	Every day, Amazon S3 sends over **125 billion event notifications** to serverless applications
Replication	Customers use Amazon S3 Replication to **move more than 100 PB** of data per week
Cold Storage Retrieval	Every day, customers **restore more than 1PB** from the S3 Glacier Flexible Retrieval and S3 Glacier Deep Archive storage classes
Data Integrity Checks	Amazon S3 performs over **4 billion checksum computations per second**
Cost Optimization	On average, customers using Amazon S3 Storage Lens advanced metrics and recommendations have obtained **cost savings 6x greater** than the Storage Lens cost in the first six months of using it.
Flexibility	**Hundreds of thousands of data lakes** are built on Amazon S3

Access Patterns

- Lower cost storage is slow

- Reads and Writes are async

- Staging - Writes

- Restores - Reads

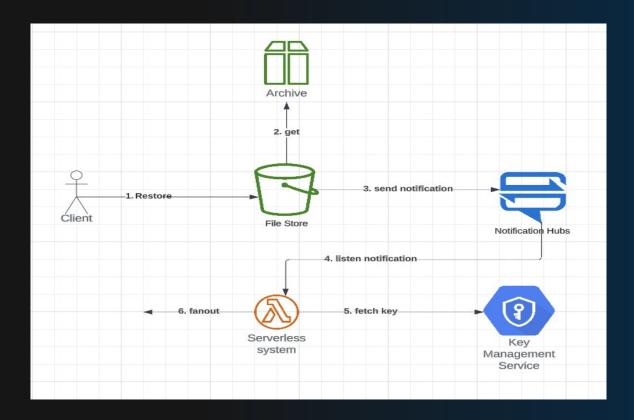- Prefer large objects for restore performance - Part Upload

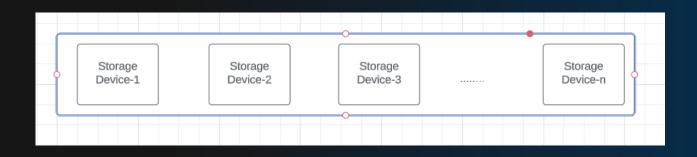- Events after completion

Data Ingress - Native Cloud Integrations

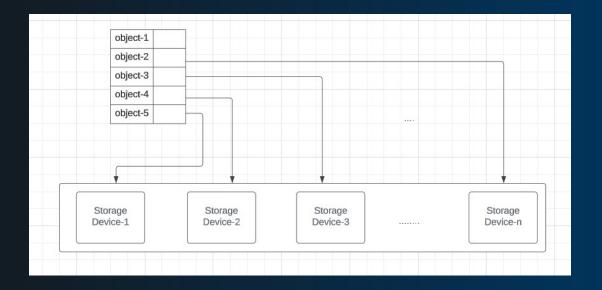# Data Egress - Native Cloud Integrations

Building a Storage System

- Enterprise Grade
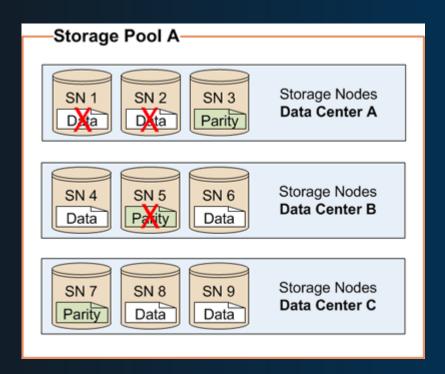- Multiple Storage Devices
- Distributed System

Distributed System

- Replication

- Horizontal Scaling

- Heat Management

- Metadata Store

- Vertical Scaling

Replication

- Erasure Coding

Horizontal Scaling

- Heat Management

- Control Plane

- Cells

Metadata Store

- Database
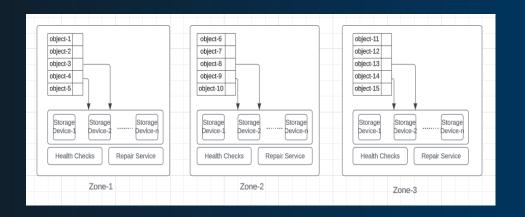- Consistency
- Replication
- Caches
- Ordering writes

Vertical Scaling

- Hot partitions

- Control Plane

- Cells

Problems to Consider

- Data Integrity - Checksums

- Durability - Repairs

- Blast Radius - Zones
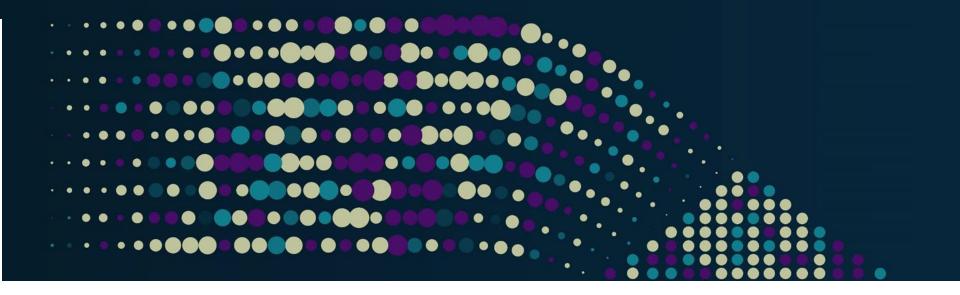
- Storage Device Availability - Health checks

Lifecycle System

- Asynchronous System

- Last Updated Time Transition

- AI/ML based Transition

Thank you

Please take a moment to rate this session.

SDC 23