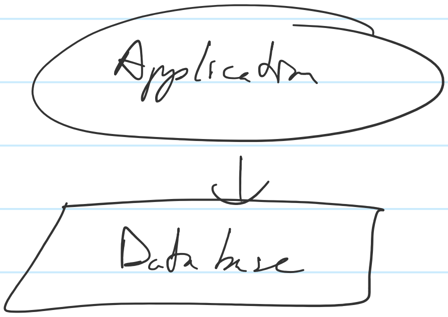
Big Data Systems

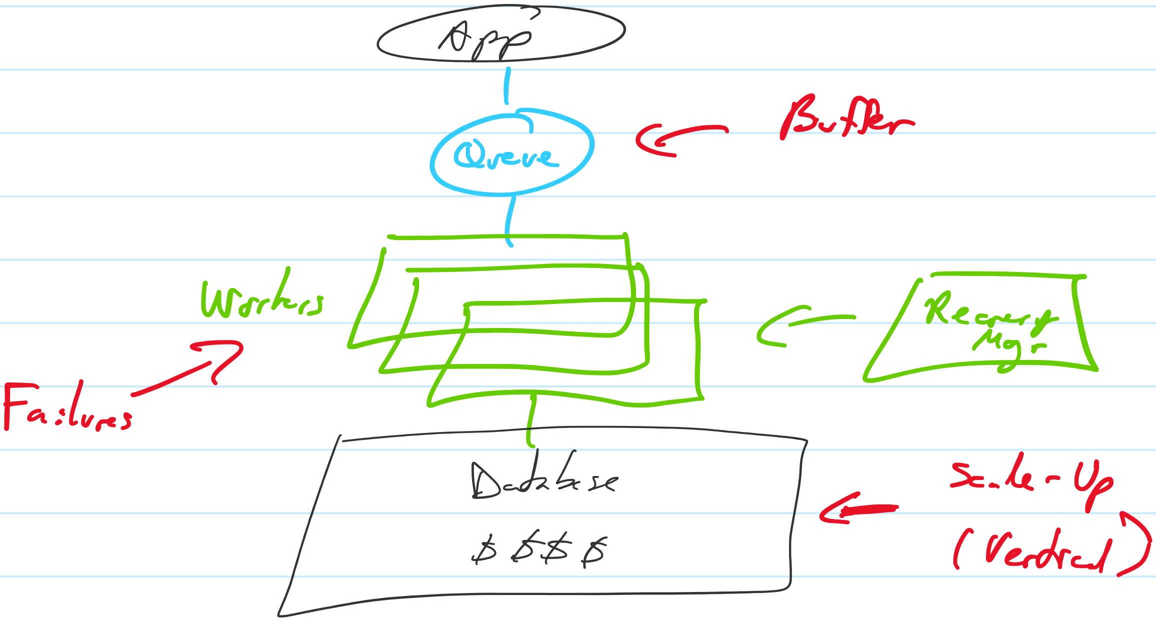
Tuesday, January 10, 2023 7:12 AM

- Big Data Technology
  - Previously:

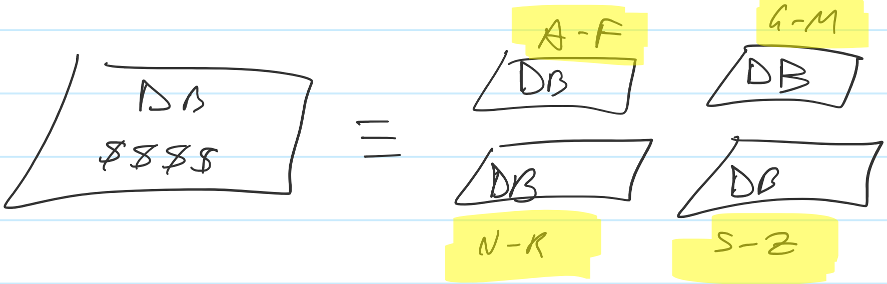


* MVC

- Issues
  - DB can't handle > X writes/sec



* Sharding (DB-Level)



↳ Recovery Mgr for DB

- System Rethinking
  - Fault-Tolerance
  - Low-Latency (IOPS)
  - High-Throughput (IOPS)
  - Scalability (Infinite)
    - Compute } HW / Cloud
    - Storage
  - Flexibility
    - Variations (Unstructured)
    - Interactions (Queries)
    - Change Mgmt (Schema Drift)
  - Extensibility
  - Ease of Use

- Big Data Attributes
  - Volume (Throughput)
    - Size Scale: GB, TB, PB ....
    - "Data at Rest"
    - Storage Tech Limitations
      - Size of disk
      - Capacity per Server
      - Storage I/O
  - Velocity (Latency)
    - "Data in Motion"
    - Stream Scale: 1d, 1h, 1m, 1s, <1s
      [illegible]

- Stream Scale: 1d, 1h, 1m, 1s, <1s
- Compute Tech Limitations
  - Speed of GPU/CPU/TPU...
  - Interconnect speed
  - Memory speed/capacity

- Variety
  - Unstructured Data
  - "Schemaless"
  - Significant storage/compute overhead

  * Structured Data: Schema → Fields/Types

  * Semi-Structured: Common Case

- File Systems (Storage) *NFS
  - Tied to running OS (POSIX)
  - Coupled to specific computer/hardware
  - Filesystems over Network → Internet
    - HTTP: GET/PUT

      read ↑ GET, write ↑ PUT

- Object Store
  - Independent of individual machines/architectures
  - Independent of client software/platform

  * Acessible via Web tech → REST

  * AWS: S3

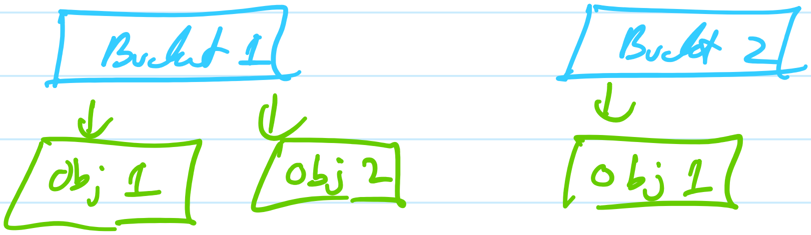
- Simple Storage Service
- Scalable
- Integrates w/ other AWS services

* Ceph
* Minio

- Concept
  - Buckets
  - Objects
  - keys

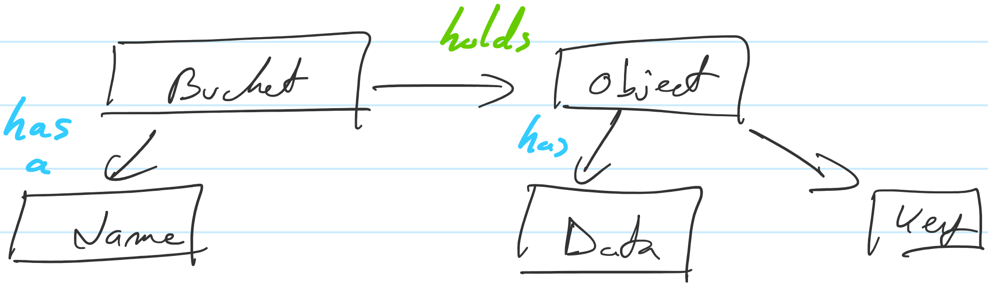
Global Uniqueness

Bucket Name + Object Name (key)



* File <-> object (key)

↳ WORM



mybucket/hello.txt

mybucket/temp/hello.txt