# HDFS

Kristo Raun

Big Data Management 2024

UNIVERSITY OF TARTU

# What is HDFS?

- The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware.

- Source: https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html

# What is HDFS?

- The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware.
- HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware.

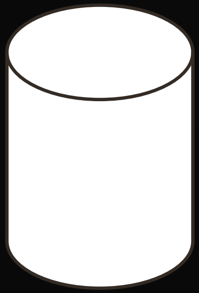
- Source: https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html

# What is HDFS?

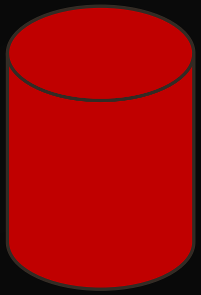
- The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware.
- HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware.
- HDFS provides high throughput access to application data and is suitable for applications that have large data sets.

- Source: https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html

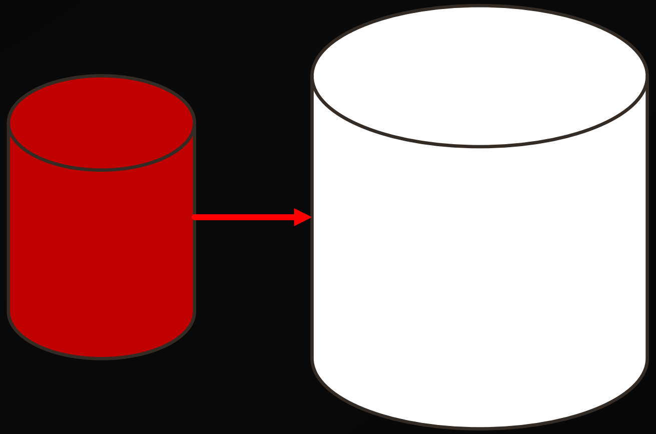
# History of Hadoop



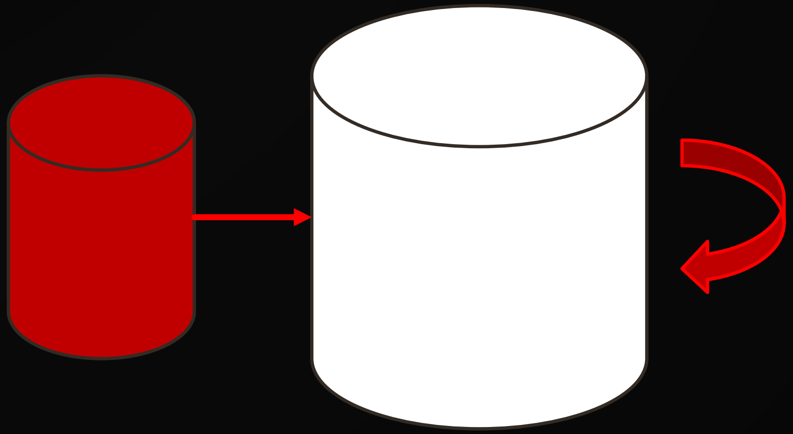
# History of Hadoop



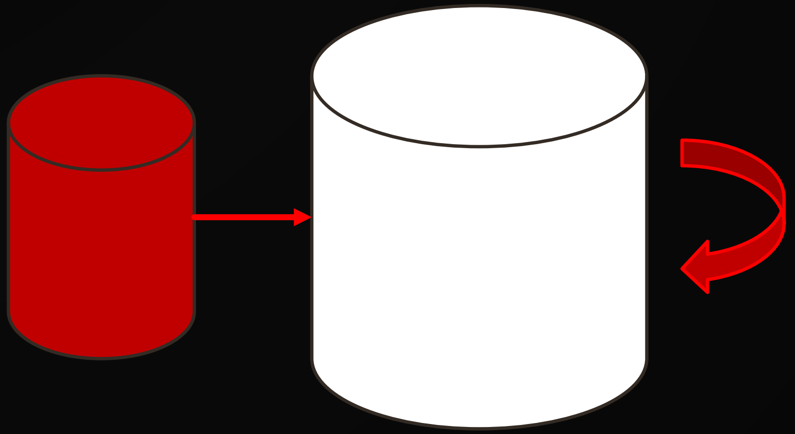
# History of Hadoop



# History of Hadoop

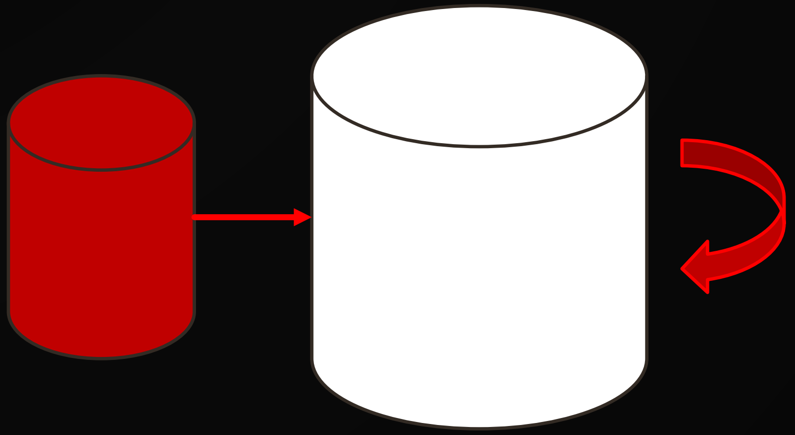


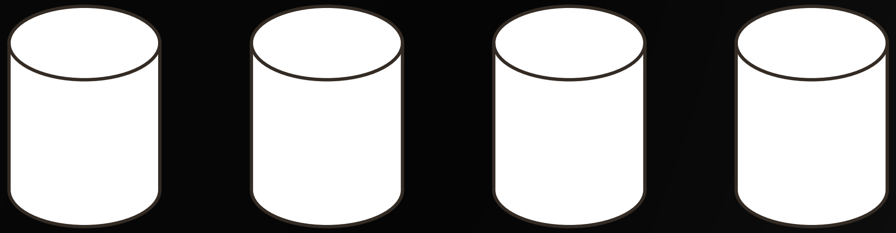
# History of Hadoop



- Difficult to scale (storage and compute)
- Pre-cloud (...-~2006)

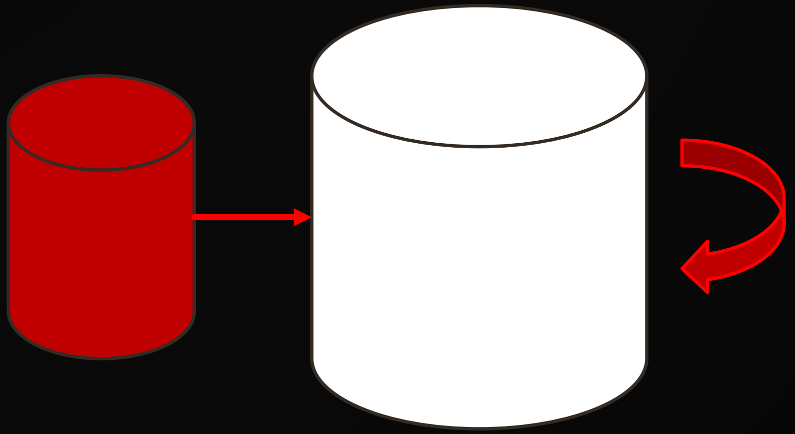
# History of Hadoop

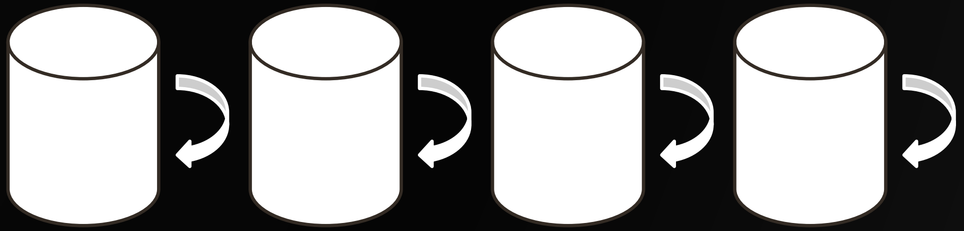




- Difficult to scale (storage and compute)
- Pre-cloud (…-~2006)

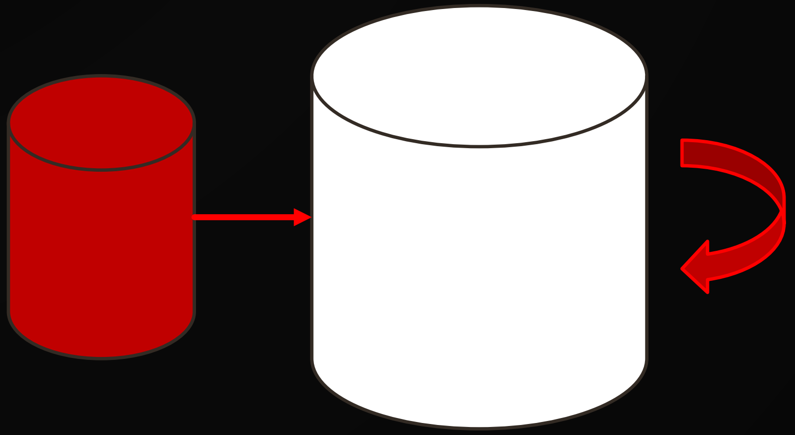
# History of Hadoop

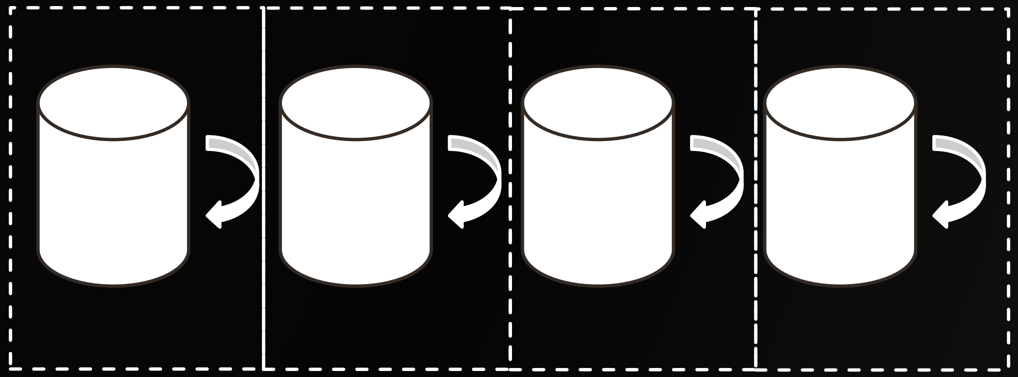




- Difficult to scale (storage and compute)
- Pre-cloud (…-~2006)

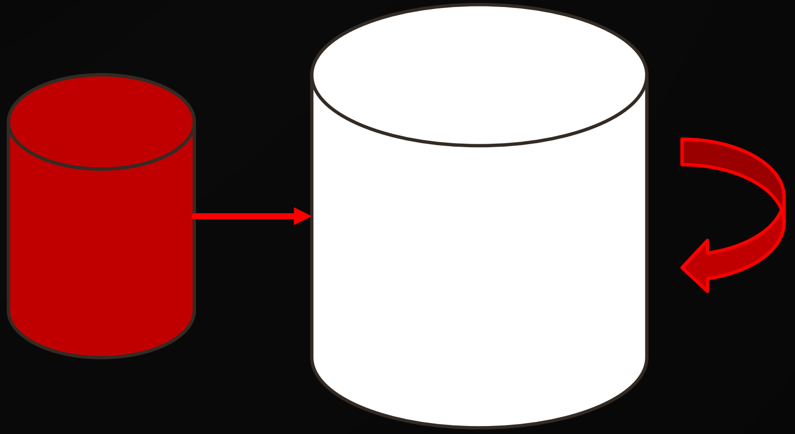
# History of Hadoop

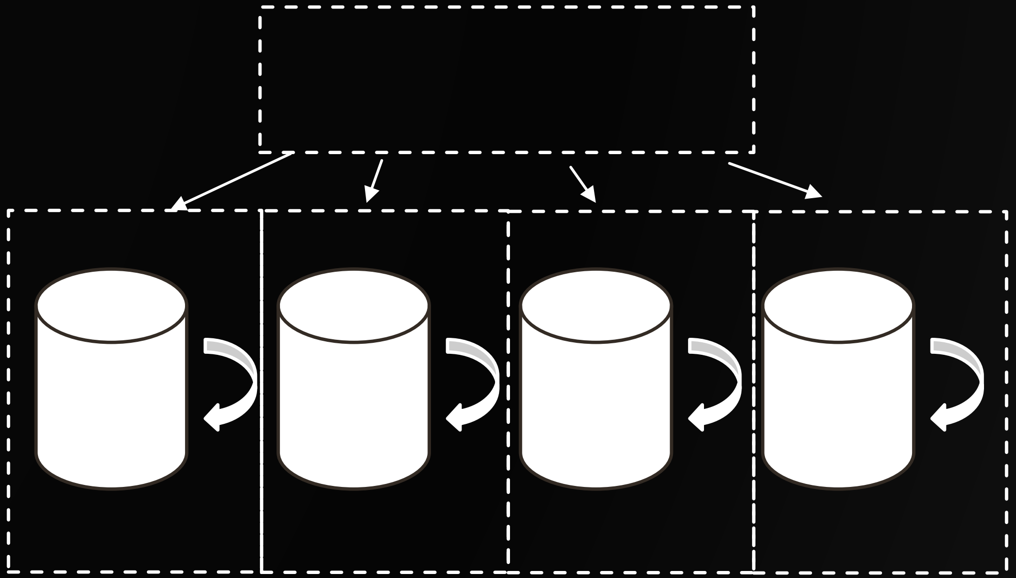




- Difficult to scale (storage and compute)
- Pre-cloud (...-~2006)

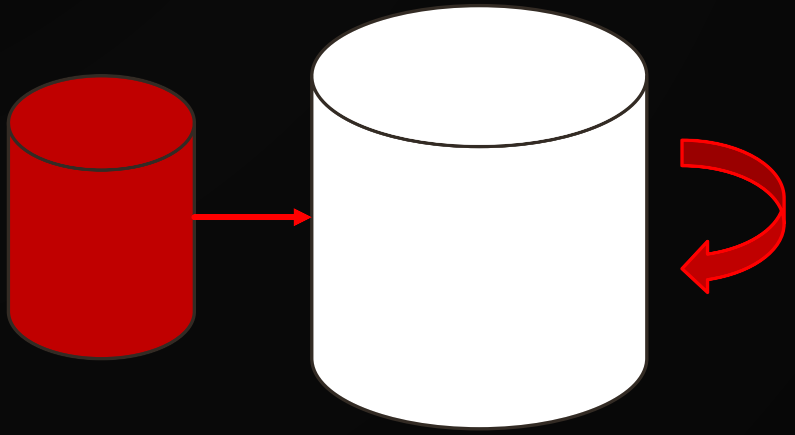
# History of Hadoop

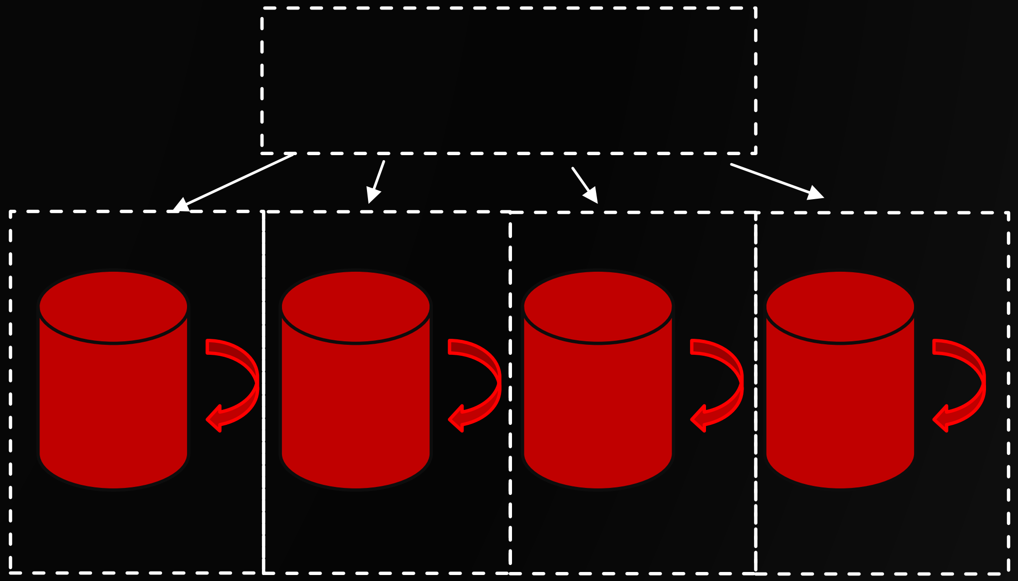




- Difficult to scale (storage and compute)
- Pre-cloud (…-~2006)

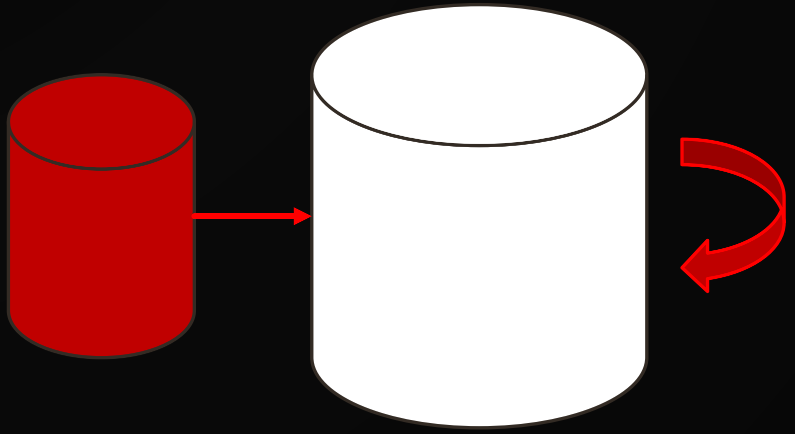
# History of Hadoop

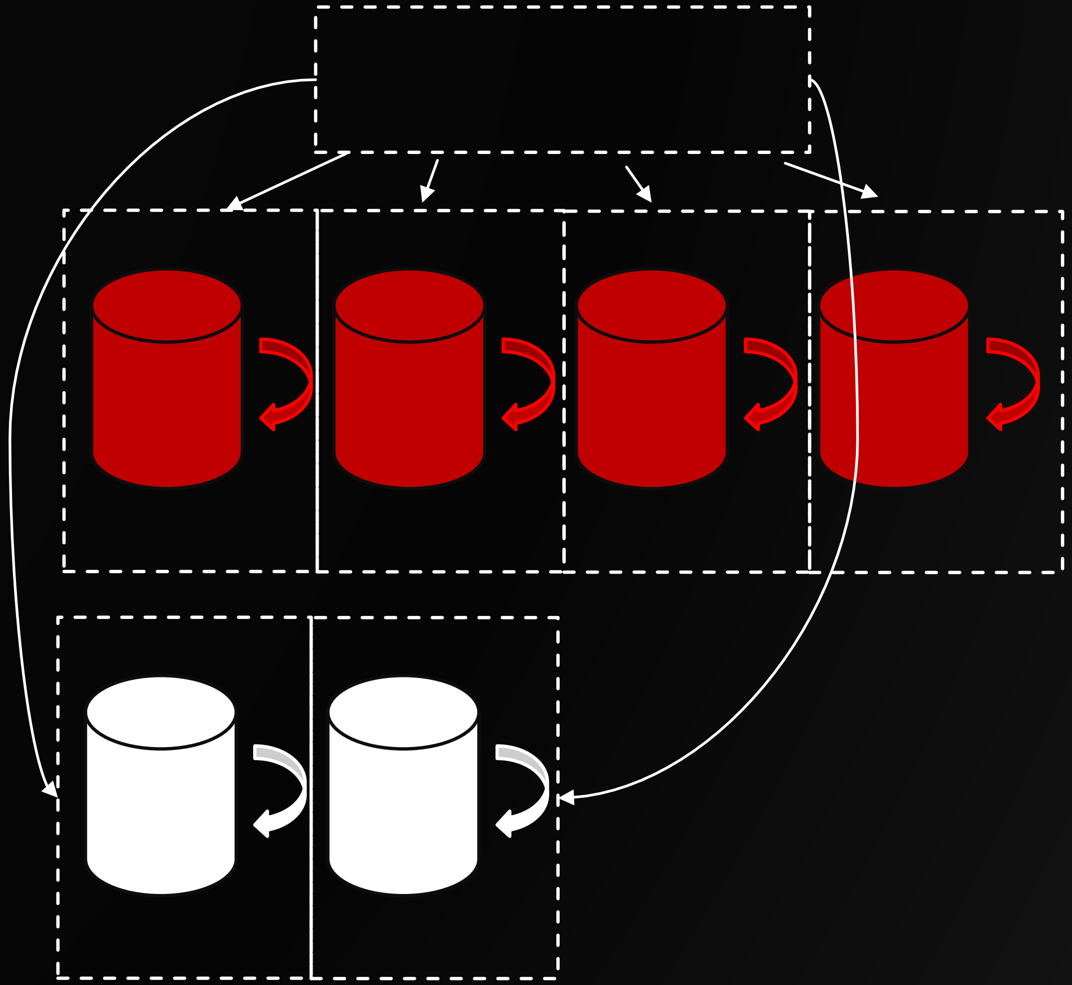




- Difficult to scale (storage and compute)
- Pre-cloud (…-~2006)

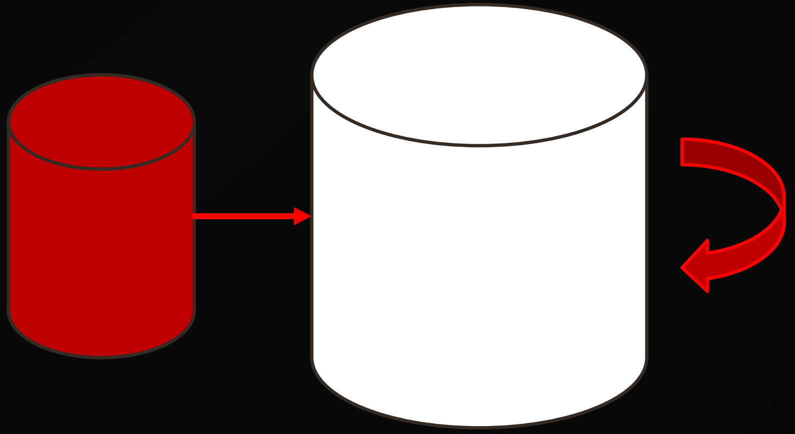
# History of Hadoop



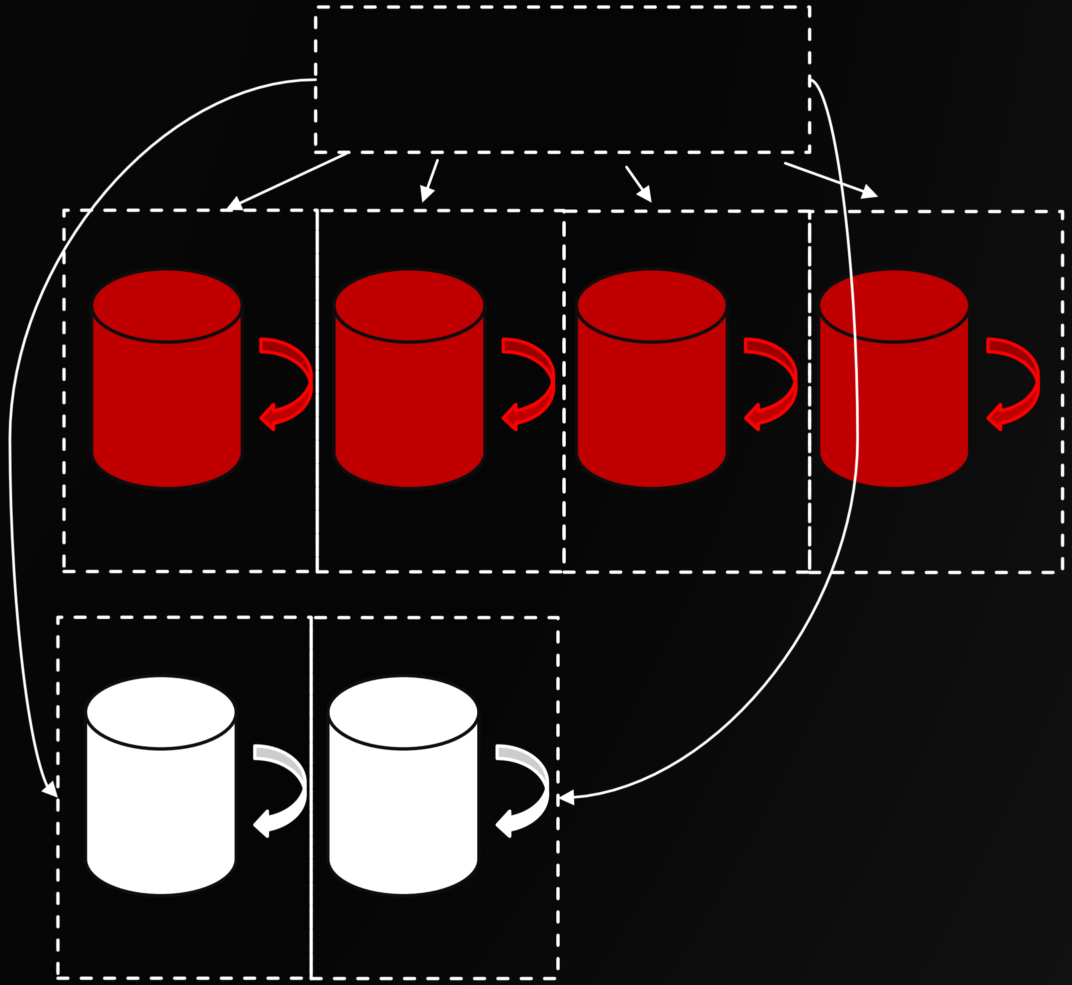


- Difficult to scale (storage and compute)
- Pre-cloud (…-~2006)

# History of Hadoop



- Difficult to scale (storage and compute)
- Pre-cloud (...-~2006)



The Google File System (2003)

MapReduce: Simplified Data Processing on Large Clusters (2004)

# HDFS vs S3 vs …

# HDFS vs S3 vs …

- In 2006 AWS launched S3 storage in cloud.

# HDFS vs S3 vs …

- In 2006 AWS launched S3 storage in cloud.
- ~2010: Azure Blob Storage, Google Cloud Storage

# HDFS vs S3 vs …

- In 2006 AWS launched S3 storage in cloud.
- ~2010: Azure Blob Storage, Google Cloud Storage
- ~2015: MinIO (S3 compatible, self-hosted)

# HDFS vs S3 vs …

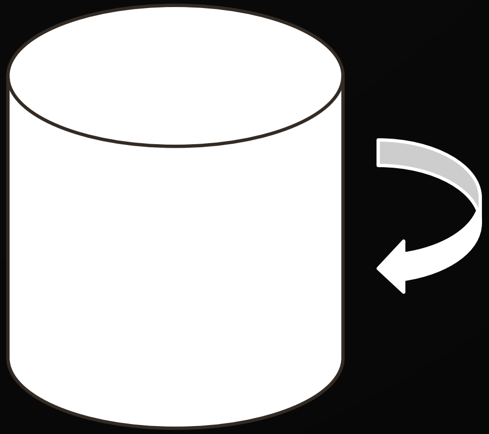
- In 2006 AWS launched S3 storage in cloud.
- ~2010: Azure Blob Storage, Google Cloud Storage
- ~2015: MinIO (S3 compatible, self-hosted)

Object storage

# HDFS vs S3 vs ...

- In 2006 AWS launched S3 storage in cloud.
- ~2010: Azure Blob Storage, Google Cloud Storage
- ~2015: MinIO (S3 compatible, self-hosted)

Object storage

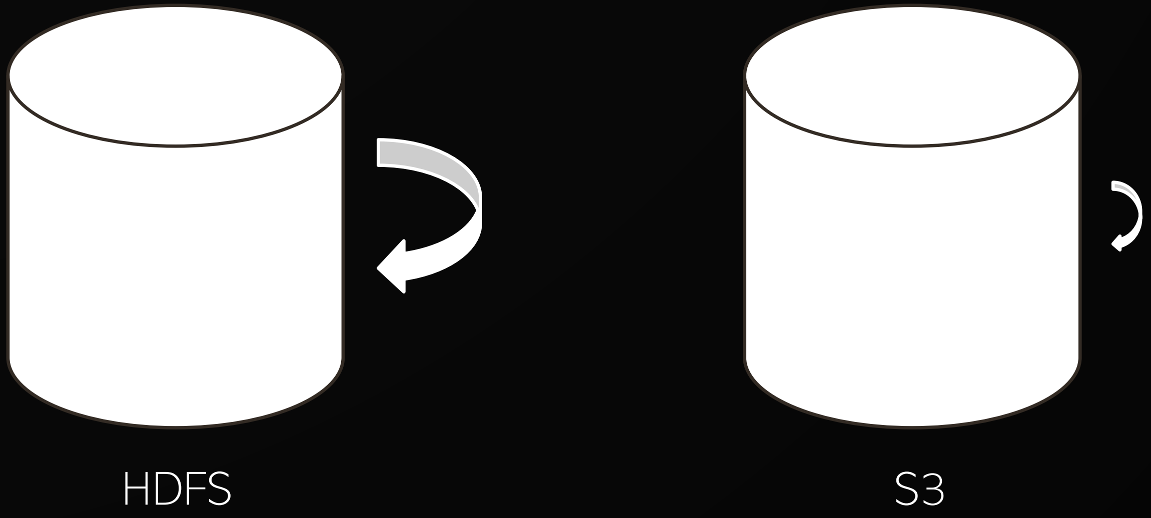


HDFS

# HDFS vs S3 vs ...

- In 2006 AWS launched S3 storage in cloud.
- ~2010: Azure Blob Storage, Google Cloud Storage
- ~2015: MinIO (S3 compatible, self-hosted)

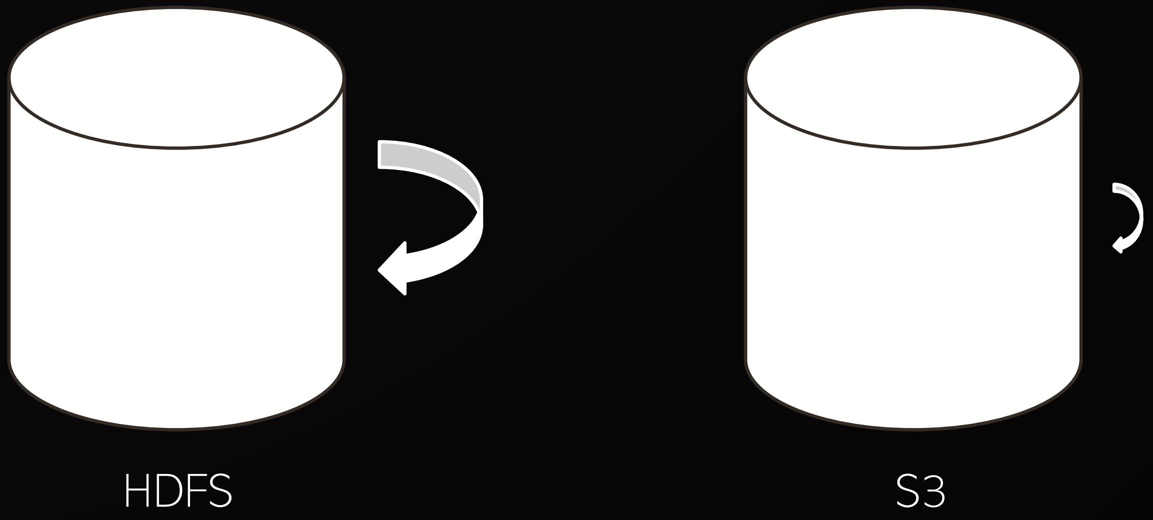
} Object storage



# HDFS vs S3 vs ...

- In 2006 AWS launched S3 storage in cloud.
- ~2010: Azure Blob Storage, Google Cloud Storage
- ~2015: MinIO (S3 compatible, self-hosted)

Object storage

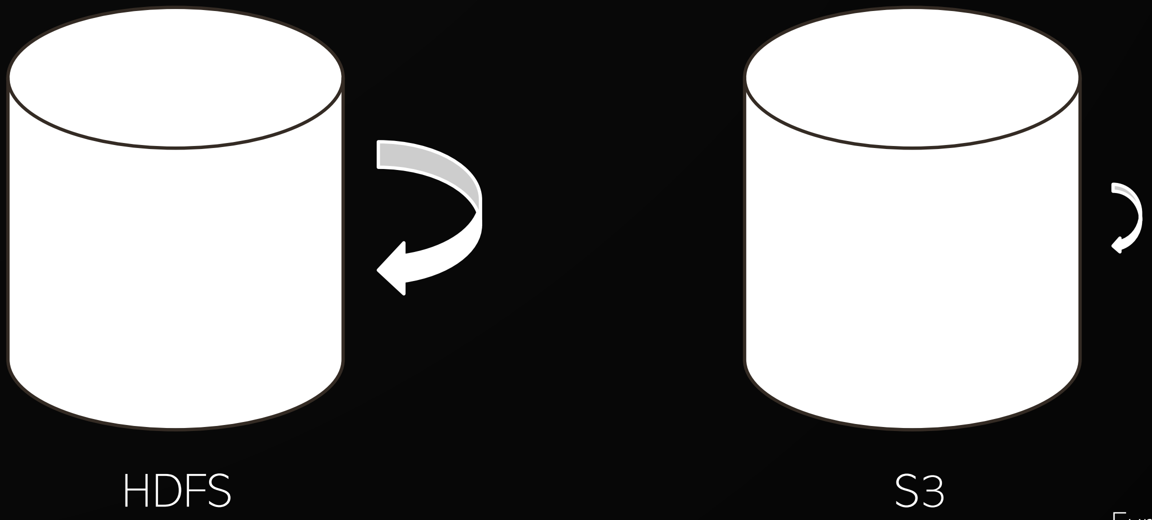


S3: more focused on storage

# HDFS vs S3 vs ...

- In 2006 AWS launched S3 storage in cloud.
- ~2010: Azure Blob Storage, Google Cloud Storage
- ~2015: MinIO (S3 compatible, self-hosted)

Object storage



S3: more focused on storage

Further reading:
https://jonascleveland.com/hdfs-vs-s3/
https://www.databricks.com/blog/2017/05/31/top-5-reasons-for-choosing-s3-over-hdfs.html