

# CS 338

## Lecture 22

### Storage and Indexing - Part 3

### Additional Slides

Collin Roberts

August 4, 2026

# Contents

## 1 Slide 04 - Detailed Computations

- 0. Setup for both questions
- 1. How many users do you expect daily?
- 2. Maximum number of concurrent users

## 2 Slide 05 - Data Backup

- 1. Online backup (hot backup)
- 2. Offline backup (archive)

## 3 Slide 06 - Incremental Backup

- 1. Incremental Backup
- 2. Data repository models

- 1 RAID5, with 10K HDDs
- 2 Each HDD's data transfer rate: 50 MBps
- 3 Total data transfer rate:  $10K * 50 \text{ MBps} = 500 \text{ GBps}$

1M users per day  $\approx$  12 users per second.

Check:

$$\begin{aligned}\# \text{ of seconds per day} &= 60 \cdot 60 \cdot 24 \\ &= 84,600, \text{ so that we get} \\ \frac{1,000,000}{84,600} &\approx 12\end{aligned}$$

**Assumption:** To render a page required 5MB of data.

$$5MB/3s \approx 1.7MBps$$

$$500GBps/1.7MBps \approx 294,118 \text{ visitors every 3 seconds}$$

Check that this yields 8.5 billion visitors per day:

$$\begin{aligned} \#3s \text{ intervals per day} &= 86,400/3 \\ &= 28,800, \text{ so that we get} \\ 294,118 \cdot 28,800 &\approx 8.5B. \end{aligned}$$

You can unplug the old drive and plug in the new one without interruption.

- 1 RAID 1 and 5 can protect against one HDD failure.
  - 1 No interruption or data loss occurs if one HDD fails.
  - 2 This is called online backup.
  - 3 It provides instant backup while the system is up and running.  
Guarding against just one failure may not sound like much.  
But for 100 HDDs, this increases reliability from 50% to 85%!
- 2 However, more than 1 simultaneous HDD failure could still happen.
- 3 What if there are two HDD failures?
- 4 Or if a fire destroys everything, how do we protect our data against that?

## What should we do?

- 1 In a business setting, besides RAID, you should also implement offline backup, e.g. back up all your data at the end of each day.
- 2 So even if the whole system fails, you still can recover data from the day before.
- 3 So you have RAID providing instant backup for one HDD failure, you also have offline backup to prevent complete failure.
- 4 You also have SMART technology that can predict and warn you about possible HDD failure.
- 5 Your data is reasonably safe.
- 6 Timing, type of failure, freshness of the data are up for discussion.

It sounds simple, right?

- 1 Facebook generates 4 petabytes of data per day.
- 2 A petabytes is a million gigabytes.
- 3 All that data is stored in what is known as the Hive, which contains about 300 petabytes of data in total.
- 4 Making a copy of the Hive regularly is not practical (too time-consuming) nor cost-effective (too expensive).
- 5 Alternative: incremental back-up - only maintain copy of changes regularly and make full copy occasionally.

- 1 An incremental backup is a backup type that only copies data that has been changed or created since the previous backup activity was created.
- 2 An incremental backup approach is used when the amount of data that has to be protected is too voluminous to do a full backup of that data every day.
- 3 How many students access their bank account today? In W20, only 2%. It means only 2% of bank data needs to be backed up on a daily basis.
- 4 For example, suppose that you created a full backup on Monday and used incremental backups for the rest of the week.
- 5 Tuesday's scheduled backup would only contain the data that has changed since Monday.
- 6 Wednesday's backup files would only contain the data that has changed since Tuesday, and so on.

- 7 The primary disadvantage of incremental backups is that they can be time-consuming to restore.
- 8 Going back to the previous example, suppose that you wanted to restore the backup from Wednesday.
- 9 To do so, you would have to first restore Monday's full backup.
- 10 After that, you would have to restore Tuesday's backup, followed by Wednesday's.
- 11 In addition, if any of the backup media happens to be missing or damaged, then you will experience incomplete data recovery.
- 12 Food for thought (FYI):  
<https://searchdatabackup.techtarget.com/tip/Data-backup-types-explained-Full-incremental-different>

- 1 Any backup strategy starts with a concept of a **data repository**.
- 2 The backup data needs to be stored somehow and probably should be organized to a degree.
- 3 It can be as simple as a sheet of paper with a list of all backup tapes and the dates they were written or a more sophisticated setup with a computerized index, catalog, or even a relational database.
- 4 Different repository models have different advantages.
- 5 This is closely related to choosing a **backup rotation scheme**.

# Data Repository: Unstructured

- 1 An unstructured repository may simply be a stack of floppy disks or CD-R/DVD-R media with minimal information about what was backed up and when.
- 2 This is the easiest to implement, but probably the least likely to achieve a high level of recoverability.

## Data Repository: Full only / System imaging

- 1 A repository of this type contains complete system images from one or more specific points in time.
- 2 This technology is frequently used by computer technicians to record **known good** configurations.
- 3 Imaging is generally more useful for deploying a standard configuration to many systems rather than as a tool for making ongoing backups of diverse systems.

# Data Repository: Incremental / Differential

- 1 An incremental style repository aims to make it more feasible to store backups from more points in time by organizing the data into increments of change between points in time.
- 2 This eliminates the need to store duplicate copies of unchanged data.
- 3 Typically to start out, a full backup (of all files) is made.
- 4 After that, any number of incremental or differential backups can be made.

# Data Repository: Incremental / Differential

- 5 Restoring a whole system to a certain point in time would require locating the last full backup taken previous to that time and all the incremental / differential backups that cover the period of time between the full backup and the particular point in time to which the system is supposed to be restored.
- 6 Additionally, some backup systems can reorganize the repository to synthesize full backups from a series of incrementals.

# Data Repository: Incremental / Differential

- 7 **Note:** Different implementations of backup systems frequently use specialized or conflicting definitions of these terms.
- 8 The most relevant characteristic of a type of incremental backup is which reference point is it using to check for changes.
- 9 By a common definition, a differential backup copies files that have been created or changed since the last full backup, regardless of whether any other backups have been since then and an incremental backup refers only to a backup that includes just the changes made since the most recent backup of any type.
- 10 Other variations include multi-level incrementals and incremental backups that compare parts of files instead of just the whole file.

# Data Repository: Reverse Delta

- 1 A reverse delta type repository stores a recent “mirror” of the source data and a series of differences between the mirror in its current state and its previous states.
- 2 A reverse delta backup will start with a normal full backup.
- 3 After the full backup is performed, the system will periodically synchronize the full backup with the live copy, while storing the data necessary to reconstruct older versions.
- 4 This can either be done using hard links, or using binary diffs.
- 5 This system works particularly well for large, slowly changing, data sets.

# Data Repository: Reverse Delta

- 6 Examples of programs that use this method are rdiff-backup and Time Machine.
- 7 Instead of scheduling periodic backups, the system immediately logs every change on the host system.
- 8 This is generally done by saving byte or block-level differences rather than file-level differences.
- 9 It differs from simple disk mirroring in that it enables a roll-back of the log and thus restoration of old image of data.