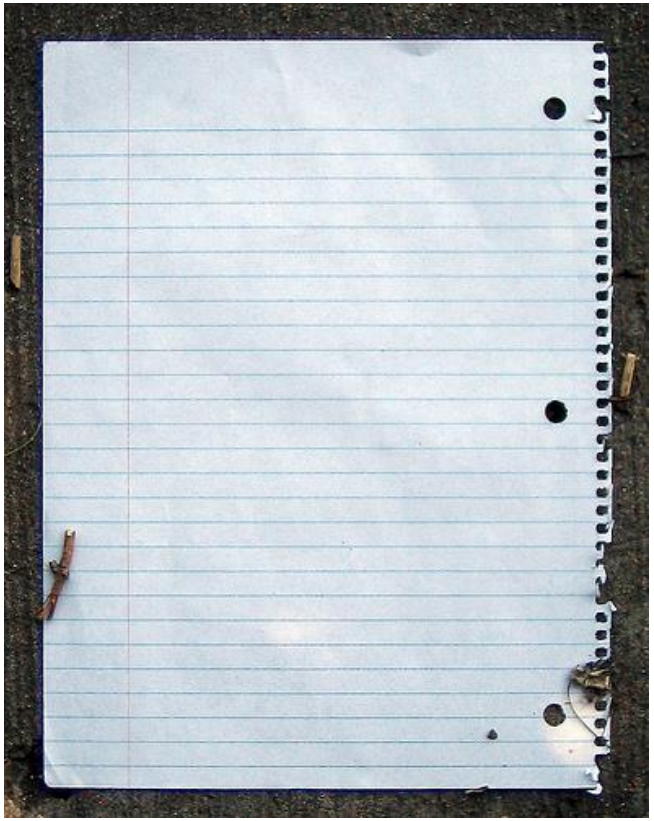
Zettabytes of Data

Department of Computer Science,

National University of Computer & Emerging Sciences,
Islamabad Campus

Data Accumulation

- **Everyone** of you has a **piece of paper**



Data Accumulation

- If you **folded** this **piece of paper** in **half**, it would **now be twice as thick** as it **was before**

- So my question is this: **How many times would you have to fold this paper onto itself to reach *the Moon*?**

How it can be done?

- **500 page** reams are about **2 inches(5 cm) high**.
- That means **one page** is about **0.01 cm high**.

- **And what of the Moon?**

Moon mean distance?

- **Mean distance** from the **Earth** is about **384,000 km**, or about **3.84 x 10^{12} pages away**.

- So ***you'd expect that you'll need an awful lot of foldings to get there***, **right**?

- **41 foldings** will get me **slightly more than halfway** to the Moon,

- So ***how many foldings would be needed, then***?

Data!

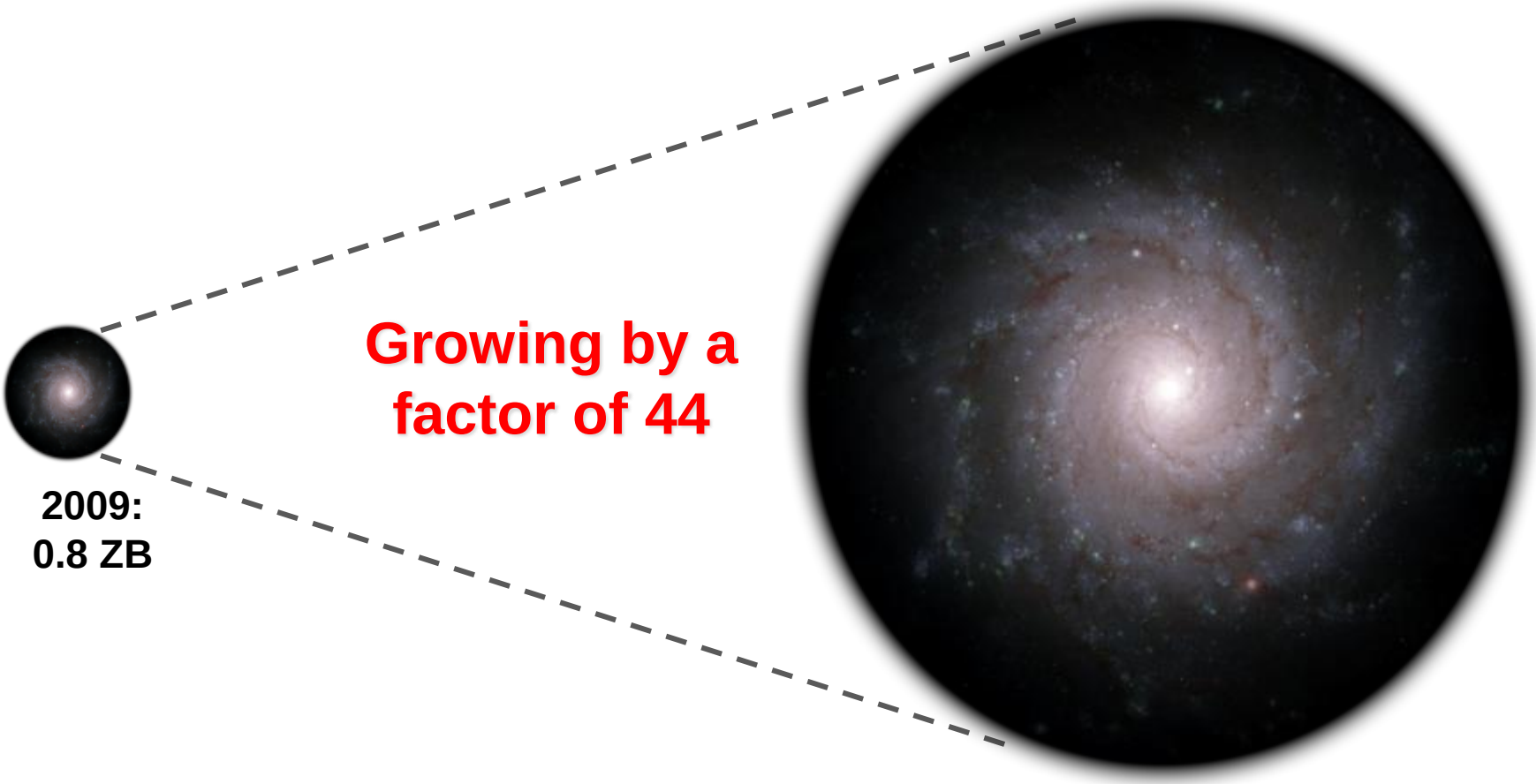
- We **live** in the **data age** and it is **difficult for a single system** to ***store***, ***process*** and ***analyze*** all of it !!
  - **Facebook** hosts more than **240 billion photos**, **growing** at **7 petabytes per month**.
    - **In 2020, it is about 4 petabytes per day**
  - **By 2025, 463 exabytes of data will be generated by humans each day**
  - **By 2024, number of emails will be 361 billion every day**
  - **Cloud storage by 2025, 200+ zetabytes**

The zettabytes of data

- **It's not easy to measure the total volume of data stored electronically,**
  - An **estimate** of the "**digital universe**" was **4.4 zettabytes** in **2013**
  - forecasting a tenfold growth **by 2020** to **44 zettabytes**.

- **But what is this zettabyte??**

The Digital Universe 2009-2020

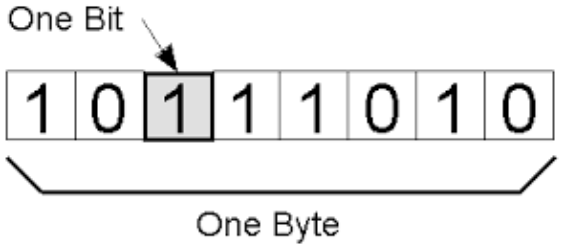


One Zettabyte (ZB) = 1 trillion gigabytes

How Big Is A Petabyte, Exabyte, Zettabyte, or A Yottabyte?

Bytes(8 Bits)

- **1 bit**: A **binary decision**
- **1 byte**: A **single character**
- **8 bytes**: A **single word**



How Big Is A Petabyte, Exabyte, Zettabyte, Or A Yottabyte?

Kilobyte (1000 Bytes, 1024 for memory)

- **1 Kilobyte**: A very **short story**
- **10 Kilobytes**: An **encyclopaedic page**
- **50 Kilobytes**: A **compressed document image page**
- **100 Kilobytes**: A **low-resolution photograph**

How Big Is A Petabyte, Exabyte, Zettabyte, Or A Yottabyte?

Megabyte (1 000 000 Bytes)

- **2 Megabytes**: A **high resolution photograph**
- **20 Megabytes**: A **box of floppy disks**
- **700 Megabytes**: A **CD-ROM**

How Big Is A Petabyte, Exabyte, Zettabyte, Or A Yottabyte?

Gigabyte (1 000 000 000 Bytes)

- **1 Gigabyte**: **A movie at TV quality**

How Big Is A Petabyte, Exabyte, Zettabyte, Or A Yottabyte?

Terabyte (1 000 000 000 000 Bytes)

- 1 Terabyte: 50000 trees made into paper and printed
- 2 Terabytes: My external (pathetic) HD
- 10 Terabytes: The printed collection of the US Library of Congress

How Big Is A Petabyte, Exabyte, Zettabyte, Or A Yottabyte?

Petabyte (1 000 000 000 000 000 Bytes)

- **2 Petabytes**: **All US academic research libraries**
- **200 Petabytes**: **All printed material**

How Big Is A Petabyte, Exabyte, Zettabyte, Or A Yottabyte?

Exabyte (1 000 000 000 000 000 000 Bytes)

- **5 Exabytes**: **All words ever spoken by human beings.**

How Big Is A Petabyte, Exabyte, Zettabyte, Or A Yottabyte?

Zettabyte (1 000 000 000 000 000 000 000 Bytes)

- **1.9 zettabytes** is the informational equivalent to **every person** on **earth receiving 174 newspapers per day.**
- **all human speech** **ever spoken** at **42 zettabytes** if digitized as **16 kHz 16-bit audio**.

And it goes on and on …

- **Yottabyte, Xenottabyte, Shilentnobyte, Domegemegrottebyte ….**

Multiples of bytes v · d · e				
SI decimal prefixes		Binary usage	IEC binary prefixes	
Name (Symbol)	Value		Name (Symbol)	Value
kilobyte (kB)	10^3	2^{10}	kibibyte (KiB)	2^{10}
megabyte (MB)	10^6	2^{20}	mebibyte (MiB)	2^{20}
gigabyte (GB)	10^9	2^{30}	gibibyte (GiB)	2^{30}
terabyte (TB)	10^{12}	2^{40}	tebibyte (TiB)	2^{40}
petabyte (PB)	10^{15}	2^{50}	pebibyte (PiB)	2^{50}
exabyte (EB)	10^{18}	2^{60}	exbibyte (EiB)	2^{60}
zettabyte (ZB)	10^{21}	2^{70}	zebibyte (ZiB)	2^{70}
yottabyte (YB)	10^{24}	2^{80}	yobibyte (YiB)	2^{80}

Google in 2010

- **Eric Schmidt** (CEO Google 2001-2011): **Every 2 Days** we create **as much information as we did up to 2003**

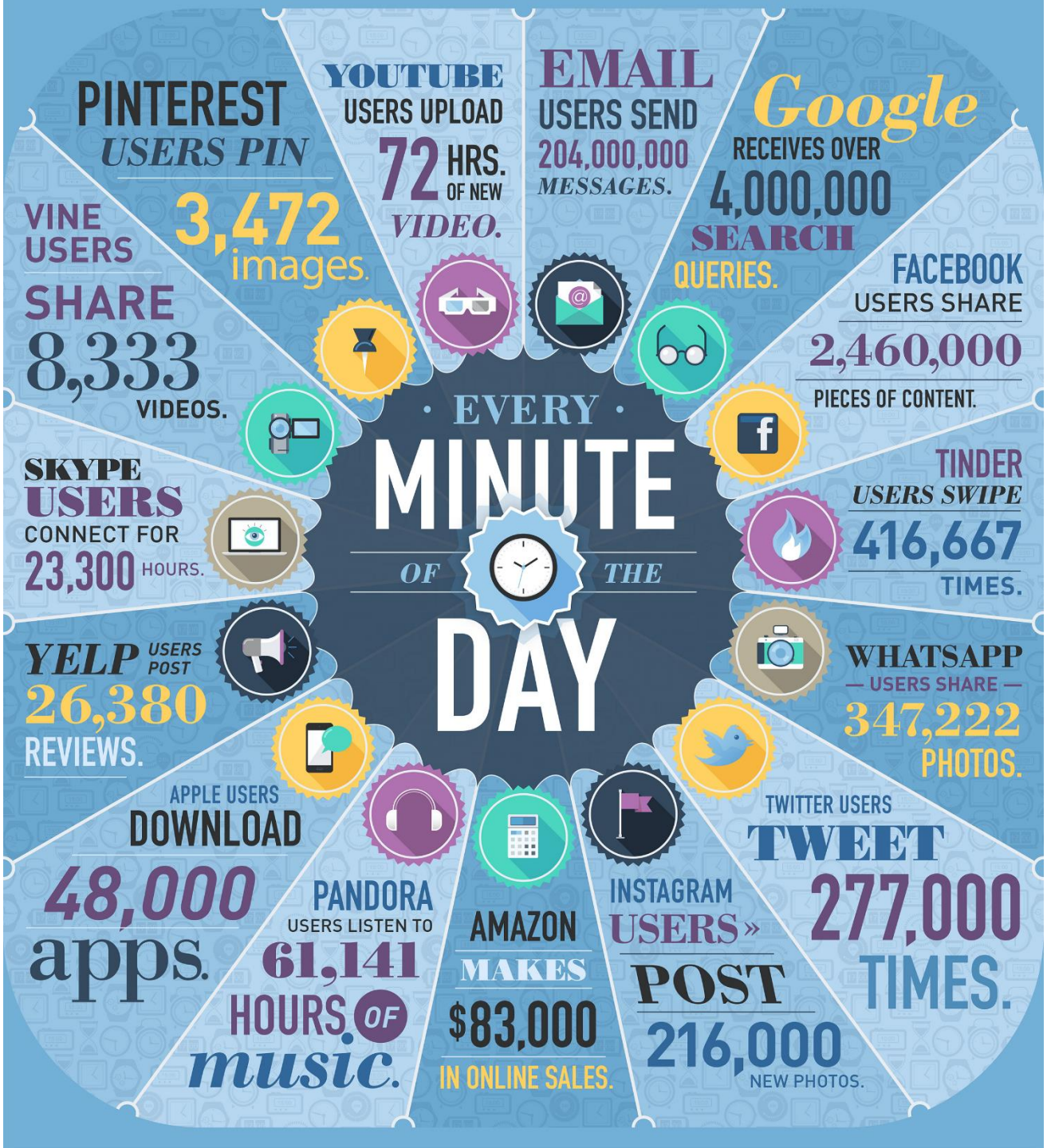
- "***The real issue is user-generated content***," He noted that ***pictures***, ***instant messages***, and ***tweets*** all add to this

Internet Traffic

- **Internet Traffic reaches around 1.1 Zettabytes in 2016**

Year	Global Internet Traffic
1992	100 GB per day
1997	100 GB per hour
2002	100 GBps
2007	2,000 GBps
2015	20,235 GBps
2020	**61,386 GBps**

Source: Cisco VNI, 2016


PINTEREST
USERS PIN
3,472
images.
YOUTUBE
USERS UPLOAD
72 HRS.
OF NEW
VIDEO.
EMAIL
USERS SEND
204,000,000
MESSAGES.
Google
RECEIVES OVER
4,000,000
SEARCH
QUERIES.
VINE
USERS
SHARE
8,333
VIDEOS.
FACEBOOK
USERS SHARE
2,460,000
PIECES OF CONTENT.
EVERY
MINUTE
OF THE
DAY
SKYPE
USERS
CONNECT FOR
23,300 HOURS.
TINDER
USERS SWIPE
416,667
TIMES.
YELP USERS POST
26,380
REVIEWS.
WHATSAPP
— USERS SHARE —
347,222
PHOTOS.
APPLE USERS
DOWNLOAD
48,000
apps.
PANDORA
USERS LISTEN TO
61,141
HOURS OF
music.
AMAZON
MAKES
$83,000
IN ONLINE SALES.
INSTAGRAM
USERS »
POST
216,000
NEW PHOTOS.
TWITTER USERS
TWEET
277,000
TIMES.

Data, data, data !!

- **Too much data being produced** (which should not be considered a problem)

- All this **creates several challenges** in the **storage environment**

Data, data, data !!

- **Organizations no longer have to merely manage their own data**;

- **Success** in the **future** will be **dictated** to a **large extent by** their **ability to extract value** from **other organizations' data**

So the data is big !!

- **What it should be called**, obviously the **Big Data** ☺
- **Big data** is a **term** that **describes** the **large volume** of **data**
- It's not the amount of data that's important. It's ***what organizations do with the data that matters***.

The Big Data Explosion

- There **are three types of Big Data**:
  1. **Big Structured Data,**
  2. **Big Semi-structured Data**
  3. **Big Unstructured Data**

The Big Data Explosion

- **All three require one** or **more** of "**three V's**",
  - **Big Data** refers to **any set of data** that **comes in**:
    - **great Volumes**
    - has a **large Variety** of information
    - and/or is consumed at **high Velocity**

Big Structured Data

- It **concerns all data** which **can be stored in database**, in **table** with **rows** and **columns**

- They have **relationnal key** and can be **easily mapped into pre-designed fields.**

- **Accounts for only 5 to 10% of all data**.

Big Semi-Structured Data

- It **doesn't reside** in a **relational database**
- **But does have some organizational properties** that **make it easier to analyze**

- With **some process** you **can store them** in **relational database**, **but not always**.

- **Examples of semi-structured** : **CSV**, **XML**, **JSON**, **NoSQL** databases are considered as semi structured. **Accounts for only 5 to 10% of all data.**

Big Unstructured Data

- **Refers** to **all** **data** that **users best understand** as **files**

- It **often include** **text** and **multimedia content**
  - **Examples**: **e-mail messages**, **word processing** documents, **videos**, **photos**, **audio** files, **presentations**, **webpages**, etc.

- **Unstructured** **data represent** around **80% of data**.

Data, data, data !!

- The **good news** is that **Big data is here**.

- The **bad news** is that we are **struggling** to **store** and **analyze it**.

why can't we just Load and Analyze?

- **The core of problem:**
  - **storage capacities** of hard drives **have increased**
  - **access speeds — have not kept up**.

- **1990** - **1,370 MB**, transfer **speed of 4.4 MB/s,** all the **data** from a **full drive can be read** in **around five minutes**.

- **Over 20 years later**, **1-terabyte drives** are the **norm**, but the **transfer speed** is **around 100 MB/s**, so it takes **more than two and a half hours to read** all the data off the disk.

- This is a **long time** to **read all data** on a **single drive** — and **writing is even slower**

- **What can be done?**

Concurrent Access

- The **obvious way** to **reduce the time** is to **read from multiple disks at once**.

- **Imagine** if we had **100 drives**, **each holding one hundredth** of the **data**.
  - Working in parallel, **we could read the data in under two minutes.**

Concurrent Access

- Using **only one hundredth** of a **disk** → **it seem wasteful**
- But **we can store 100 datasets**, each of which is **1 terabyte**, and **provide shared access to them**
- **However, it is not that simple !!**

Concurrent Access

- The **first problem** to **solve is hardware failure**:
  - as soon as **you start using many pieces** of **hardware**, the **chance that one will fail** is **fairly high**
  - A **common solution** (to **avoid data loss**): **Replication**
  - **Replication**: **redundant copies of the data** are **kept** by the system so that in the **event of failure**, there is **another copy available**

Concurrent Access

- The **second problem** is that **most analysis tasks need to be able to combined the data** in **some way**:
  - **data read** from **one disk** **may need** to **be combined** with **data** from **any** of the **other 99 disks**.
  - **Various distributed systems allow data** to be **combined** from **multiple sources**
    - **Doing this correctly is notoriously challenging**

Hadoop - Why ?

- **Need** to **process huge datasets** on **large clusters** of **computers**
- **Very expensive** to **build reliability** into **each application**
- **Nodes fail every day**
  - **Failure** is **expected**, **rather than exceptional**
  - **Mean time between failures for 1 node = 3 years**
- **Need a common infrastructure**
  - **Efficient**, **reliable**, **easy to use**
  - **Open Source**

So what is this Hadoop?

- A **framework** for **distributed processing** of **large data sets** across **clusters of computers** using **simple programming models**

- Can **scale up** from **single servers** to **thousands of machines**, each offering **local computation** and **storage**

- **Designed** to **detect** and **handle failures** at the application layer (so delivering a **highly-available services**)

Typical Hadoop Cluster

Image from http://wiki.apache.org/hadoop-data/attachments/HadoopPresentations/attachments/aw-apachecon-eu-2009.pdf

Who uses Hadoop?

- Amazon, Facebook, Google, Twitter, New York Times, Veoh, Yahoo! …. many more