Data Collection and Modeling

Course 2: Files & File Sharing

Files

Definition

“A computer file is a resource for storing information, which is available to a computer program and is usually based on some kind of durable storage. A file is durable in the sense that it remains available for programs to use after the current program has finished.” (definitions.net)

“A file (...) is a self-contained piece of information available to the operating system and any number of individual programs.” (lifewire.com)

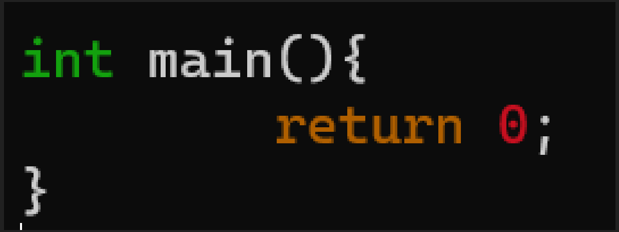
Files

A computer system can distinguish between 3 different file types:

- regular: are used to store data
  - text files: lines of text ended by EOL and a final EOF
  - binary files: bits that can represent custom data
- directory: information containers used to access other files
- special: usually physical devices or communication channels (FIFOs)

Files

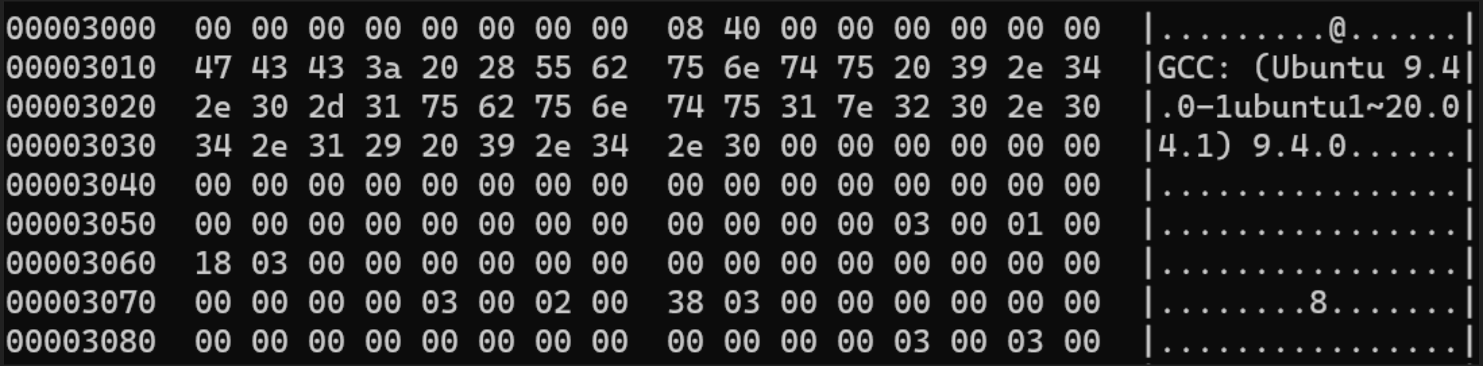
Text



```
int main(){
        return 0;
}
```

compile…

Binary



```
00003000  00 00 00 00 00 00 00 00  08 40 00 00 00 00 00 00  |.........@......|
00003010  47 43 43 3a 20 28 55 62  75 6e 74 75 20 39 2e 34  |GCC: (Ubuntu 9.4|
00003020  2e 30 2d 31 75 62 75 6e  74 75 31 7e 32 30 2e 30  |.0-1ubuntu1~20.0|
00003030  34 2e 31 29 20 39 2e 34  2e 30 00 00 00 00 00 00  |4.1) 9.4.0......|
00003040  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00  |................|
00003050  00 00 00 00 00 00 00 00  00 00 00 00 03 00 01 00  |................|
00003060  18 03 00 00 00 00 00 00  00 00 00 00 00 00 00 00  |................|
00003070  00 00 00 00 03 00 02 00  38 03 00 00 00 00 00 00  |........8.......|
00003080  00 00 00 00 00 00 00 00  00 00 00 00 03 00 03 00  |................|
```

Data Types

- Observational
- Experimental
- Derived (compiled)
- Simulation
- Canonical (reference)

Data Formats

Many formats: text, numeric, multimedia, domain specific, device specific (medical imagistics)

Accessible format characteristics:

- Uncompressed
- Standard character encoding
- Non proprietary
- Open standards
- Popular among community

File Formats

Represents the structure of data stored inside a file

Formats can be grouped in various categories

Usually possible to convert from one format to another within same category

docs.fileformat.com

fileinfo.com/filetypes/

File Formats

Format	High Confidence	Medium Confidence	Low Confidence
Text	Plain text UTF-8 BOM (.txt) XML with schema (.xml)	Plain text ISO 8859 (.txt) HTML (.html, .htm), CSS	Microsoft word (.doc) Wordperfect (.wpd)
Raster	Uncompressed TIFF (.tiff) True color 24bit PNG (.png)	Compressed TIFF (.tiff) GIF (.gif), BMP, 8 bit PNG	Rawfile Photoshop (.psd)
Vector	SVG (.svg)	Computer Graphics Metafile (.cgm)	Macromedia Flash (.swf), Postscript (.eps)
Containers	TAR (.tar) No compressed ZIP (.zip)	Compressed ZIP (.zip)	

File Formats

Format	High Confidence	Medium Confidence	Low Confidence
Spreadsheet / Database	Separated values (.tsv, .csv), Delimited text	dBASE (.dbf), MS Excel (xlsx), SAS (.sas)	Excel (.xls)
Multimedia	Uncompressed AVI, FFV1/Matroska (.mkv)	MPEG-1, MPEG-2, Motion JPEG 2000 (.jp2)	Windows MEdia Video (wmv), RealAudio (.ra)
Digitized documents	Uncompressed TIFF PDF/A-1		
Programs		Source code	Executable

File Sharing

Definition

“the practice of making computer files available to other users of a network, in particular the illicit sharing of music and video via the internet.” (Oxford Languages)

“File sharing is a productivity tool that allows select users to share files with one another remotely. Files can be shared with just one or two individuals or entire organizations. File sharing also works with nearly any file type, so users can exchange text documents, images, audio and video files, PowerPoint slides and more.” (Mitel)

Types of sharing

- Operating system file sharing
  - Sharing files between users using network layer (FTP)
- Internet file sharing
  - P2P
  - File sync and sharing services
  - Portals
  - Rsync
  - Data sync

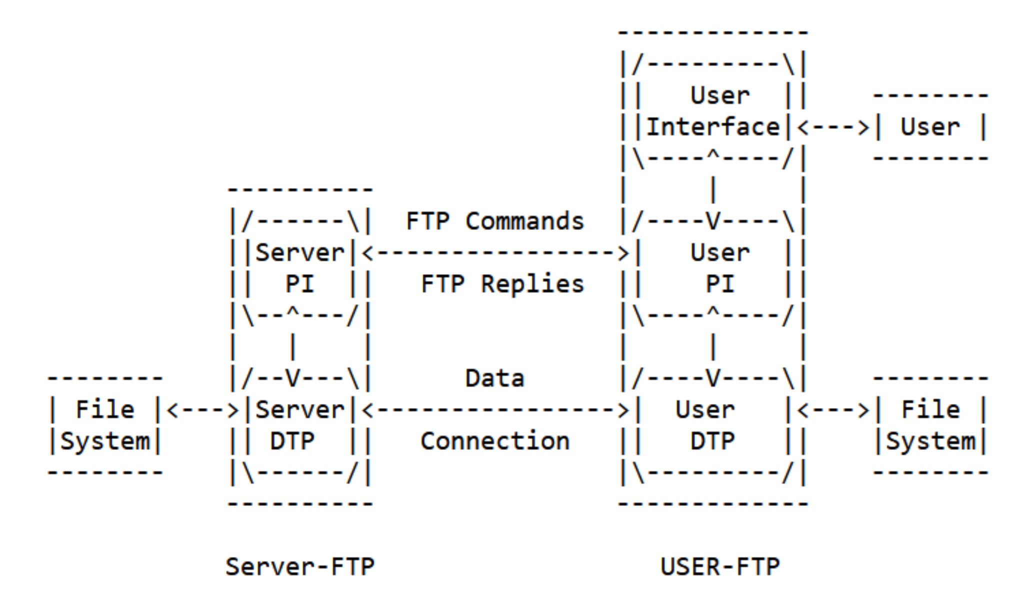
FTP

File Transfer Protocol (RFC 959) is used to communicate and transfer files between computers on a TCP/IP network

Objectives:

- Promote sharing of files
- Encourage use of remote computers
- Prevent against variations in file storage systems between hosts
- Transfer data in a reliable and efficient manner

FTP



```
                                            -------------
                                            |/---------\|
                                            ||   User  ||    --------
                                            ||Interface|<--->| User |
                                            |\----^----/|    --------
                  ----------                |     |     |
                  |/------\|  FTP Commands  |/----V----\|
                  ||Server|<---------------->|   User  ||
                  ||  PI  ||   FTP Replies  ||    PI   ||
                  |\--^---/|                |\----^----/|
                  |   |    |                |     |     |
      --------    |/--V---\|      Data      |/----V----\|    --------
      | File |<--->|Server|<---------------->|  User   |<--->| File |
      |System|    || DTP  ||   Connection   ||   DTP   ||    |System|
      --------    |\------/|                |\---------/|    --------
                  ----------                -------------

                  Server-FTP                   USER-FTP
```

The FTP Model

FTP Apps

- FileZilla: open source, multithreaded transfers, SFTP, FTPS; all OSs
- Cyberduck: open source, cloud storage browser, SFTP, WebDav, AWS S3; Windows and MacOS
- FireFTP: open source, Firefox add-on (up to 57, Waterfox after that) , FTP, SFTP, FTPS; all OSs
- WinSCP: popular, many features, FTP, SFTP, SCP, FTPS, WebDAV, AWS S3; Windows

P2P

File-sharing technology that allows the users to access files over a network. Users/computer accounts in this network are called to as peers; they request files from other peers using TCP or UDP connections.

Such a P2P network allows communication without a server; there is no central server for handling requests, the peers interacting without the requirement of a central server.

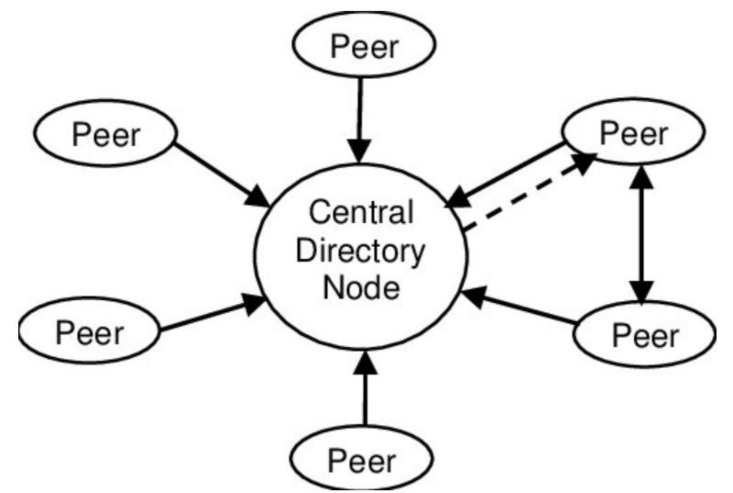
For a peer request there may be multiple peers that have a copy of the file. There are 3 architectures that allow communication with these peers:

- Centralized (Central Directory)
- Pure (Query Flooding)
- Hybrid

P2P - Centralized

- similar to client-server architecture: it maintains a central server for directory services
- the peers inform central server about their IP address and the files available for sharing
- the server queries the peers at regular intervals to check if still connected

Napster

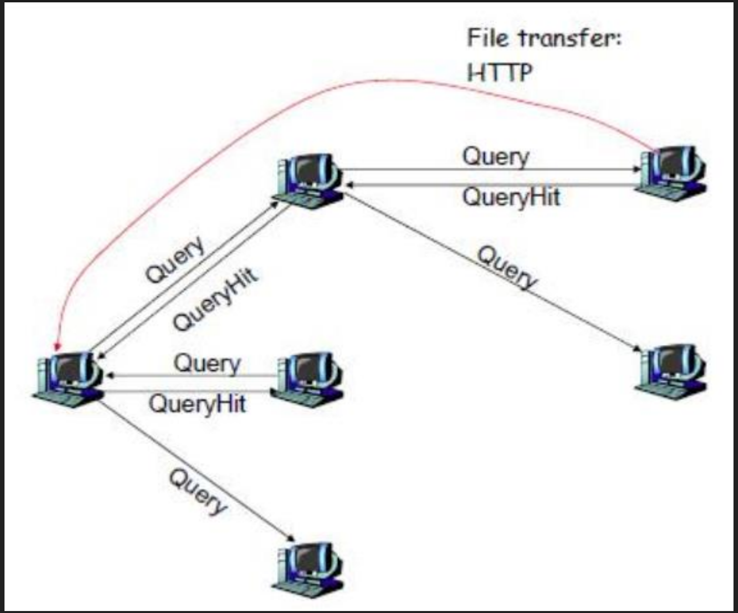


source: researchgate.net

P2P - Pure

- a node searching for a file contacts all neighbors in the system, they contact their own neighbors and so on until a "hit" is obtained (file is located)
- this process assumes no knowledge about the network topology
- if there are multiple hits, the client can select one from the list of peers

Gnutella

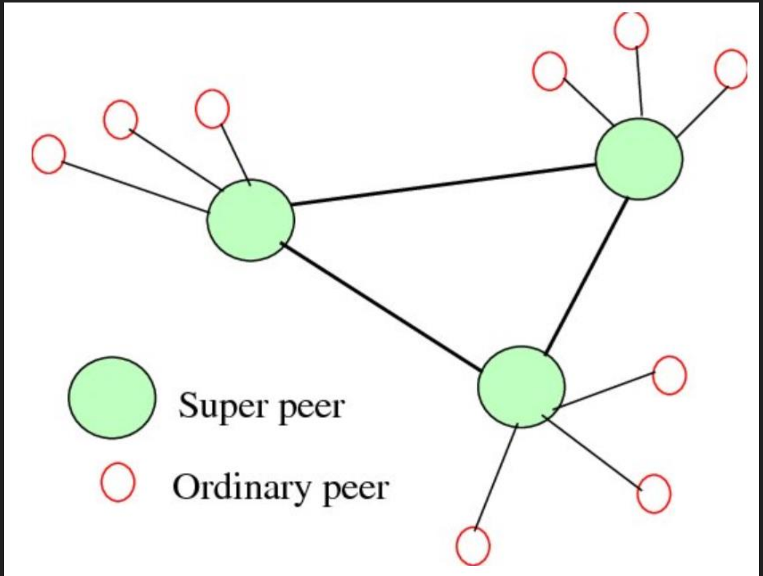


source: Wikipedia

P2P - Hybrid

- makes use of both centralized and pure
- there is no centralized server for query processing
- not all pears are equal; ones with high bandwidth and stable connectivity are called leaders/supernodes
- each supernode has a subset of the rest of th peers and indexes their IPs and shared files

KaZaA



source: researchgate.net

File sync / File Sharing

- file syncing is the process of file updating (in real-time) across multiple devices. It allows concurrent work from more than one person on the same file
- there are two types of syncing:
    - one-way sync: files are updated from one single source to multiple target locations; no data is sent back for update to the source
    - two-way sync: there are several locations that work together using two-way communication between every pair in the system
- usually cloud based
- Google Drive, IDrive, Sync.com, Microsoft Onedrive

Portals

Client portals represent a secure online location that clients can access at any time to view and manage files (upload, download)

Allows:

- easy share files in a secure manner by creating a link to a public or private share
- limit the maximum number of downloads for a shared file
- automatically expire a share link after a certain time
- anonymous file uploads

Files are always in sync and protected by regular backups

Portals

- Huddle: cloud based, audit trails (timestamps for files), enhanced (government grade) security, HIPAA and GDPR compliant, branded client portals
- Citrix Sharefile: enhanced security using password and device lock, flexible storage (on cloud or on-premise storage)
- Dropbox for Business: third party app integrations (Zoom, Slack), top-tier security, fast file transferring, task management systems
- FileCloud: flexible hosting (public, private), syncs across all devices

rsync

stands for remote sync

is a remote and local file synchronization tool

uses an algorithm to minimize the amount of data copied by only moving the portions of files that have changed

in order to use rsync to sync with a remote system, SSH access need to be configured between local and remote machines and rsync installed on both systems

written in C as a single threaded application; the algorithm is a type of delta encoding, and is used for minimizing network usage

Data Sync

The process of synchronizing data between two or more devices and updating changes automatically between them to maintain consistency, ensures accurate, secure, compliant data

Input data gets cleaned, checked for errors, removed duplication and checked for consistency

We consider to be local synchronization when devices and computers use the same local network, and to be remote synchronization when it takes place over a mobile network

Data changes must upgrade every system in real-time to avoid mistakes, prevent privacy breaches, and ensure that the most up-to-date data is the only information available

Data synchronization ensures that all records are consistent, all the time

Data Sync Types

There are some data sync methods that can update more than one copy of a file at a time, and some may update only one

- File Synchronization: most used for home backups, external hard drives, or updating portable data via flash drive
- Version Control: synchronizing solution for files that can be altered by more than one user at the same time
- Distributed File Systems: multiple file versions must be synced at the same time on different devices, those devices must always be connected for the distributed file system to work
- Mirror Computing: provides different sources with an exact copy of a data set. Useful for backup, provides an exact copy to one other location

Data Sync Challenges

- security: data that is updated in different locations hass to meet regulatory standards and privacy laws
- data quality: updates have to maintaining strict integrity of information within a secure environment
- management: data management has to be done in real-time to ensure accuracy and prevent errors
- performance: usually data synchronization is done using ETL (5 steps)
    - Extraction from the source
    - Transfer
    - Transformation
    - Transfer
    - Load to target

When dealing with a large volume of data, synchronization must be a priority to keep performance

- data complexity: formats may change according to needs and technological modifications; data should be available for both old and new systems operations

ETL - E

Data needs to be managed from various sources and saved in a destination system. In this first step of the ETL process, structured and unstructured data is imported and consolidated into a single repository. Volumes of data can be extracted from a wide range of data sources, including:

- Databases and legacy systems
- Cloud, hybrid, and on-premises environments
- Sales and marketing applications
- Mobile devices and apps
- CRM systems
- Data storage platforms
- Data warehouses
- Analytics tools

ETL -T

Data quality can be assured by applying rules and regulations. Data transformation consists of several sub-processes:

- Cleansing — deal with inconsistencies and missing values
- Standardization — apply formatting rules
- Deduplication — remove redundant data
- Verification — unusable data is removed and anomalies are flagged
- Sorting — organize data based on types
- Other data quality tasks — additional/optional applied rules to improve data quality

Transformation is usually the most important/consistent part of the ETL process

ETL - L

- load the transformed data into a new destination (usually data lake or data warehouse). Data can be loaded in its entirety (full load) or incremental load:
- full loading: everything that comes from the transformation step goes into new, unique records in the destination.
- incremental loading: compares incoming data with what’s already in the destination and loads additional records if new information is found