Data storage: simple solutions

File Formats

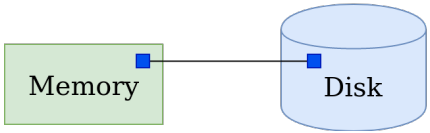
Hiba ALQASIR

2021-2022

télécom
saint-étienne
école d'ingénieurs / nouvelles technologies

Disks vs. memory

- Disk is a persistent storage
- Disk capacity is usually much larger
- Data representation on disks is generally not the same as in memory
- Disks are significantly slower than memory



Disks I/O

- Disk reads/writes in terms of blocks, not bytes.
- How to modify a single byte?
  - Read the entire block containing the byte from the disk;
  - Place the block in the primary memory;
  - Modify that byte;
  - Write the entire block back to disk.
- Disk accesses are a major performance bottleneck

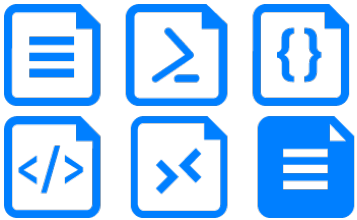
Organizing the data

1. Drive
   - Storage device.
2. Folder (directory)
   - Storage location that takes place on a storage device to hold data.
3. File
   - Document that is generated with programs such as text or image editor.

Files and folders

File

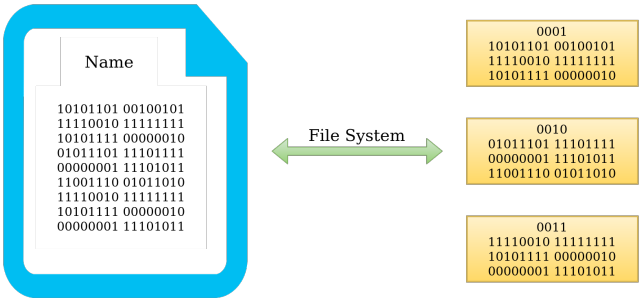
- An array of bytes.
- Associate bytes with a name.



Folder (directory)

- A list of files and directories.
- Associate names with each other.

Files



named sequence of bytes

disk blocks

Goal: operations should have the minimum possible disk access

Operations on files

- Create a new file/ Delete an existing file
- Open/Close a file
- Move a file
- Rename a file
- Read data from file
- Write data to file
- Change the access permissions and attributes of a file

File Attributes

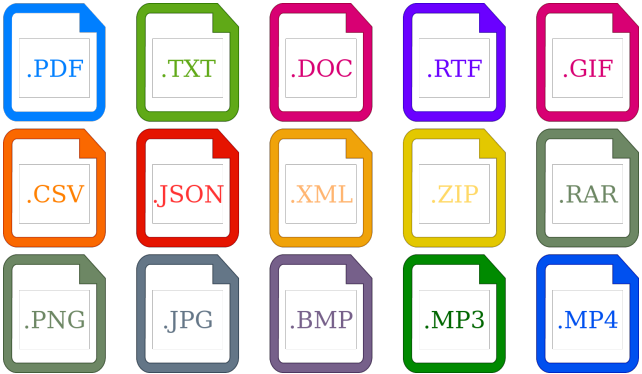
- Name: symbolic file-name, human-readable form.
- Identifier: unique tag; non-human readable (file system).
- Location: pointer to a device and to file location on that device.
- Size: current file size, maximaum possible size.
- Protection: to control who can read, write, execute the file.
- Time, Date and user identification.

File name and contents

- The name and the path of a file must uniquely identify it.
  - No two files can have the same name and path.
- The format of a file is defined by its content, and indicated either by:
  - filename extension,
  - metadata stored inside or outside file.

File formats

- Standard method of encoding data for storage.
- Different file formats are used for different applications.



Serialization

- Translating a data structure into a series of bytes or characters.
- Allows the data be accessed and modified efficiently.
- Allows the recovery of the original data structure.

Serialization formats

- Binary
  - Not human-readable
  - Fast to parse
  - Platform dependent
  - Memory efficient
  - Example: Avro, Thrift, Protocol Buffers
- Text
  - Human-readable
  - Slow to parse
  - Platform independent
  - Lower memory efficiency

Textual formats

Title	Year	Genre	Director	Actors
Cruella	2021	Comedy Crime Drama	Craig Gillespie	Emma Stone Emma Thompson Joel Fry
Cast away	2000	Adventure Drama	Robert Zemeckis	Tom Hanks Helen Hunt Lari White
Joker	2019	Crime Drama	Todd Phillips	Joaquin Phoenix Robert De Niro Zazie Beetz

CSV

- **C**omma **S**eparated **V**alues (CSV)
- Simplest format
- Suitable for storing data organized in a single table
- No hierarchical structure
- Whenever your data does not have a nested structure: USE CSV!

```
Title, Year, Genre, Director, Actors
Cruella, 2021, Comedy/Crime/Drama, "Craig Gillespie", "Emma Stone"/"Emma Thompson"/"Joel Fry"
Cast away, 2000, Adventure/Drama, "Robert Zemeckis", "Tom Hanks"/"Helen Hunt"/"Lari White"
Joker, 2019, Crime/Drame, "Todd Phillips", "Joaquin Phoenix"/"Robert De Niro"/"Zazie Beetz"
```

- Column names are given by the first row (non verbose)
- Column separator: comma, semicolon or *tap*
- Row separator: *newline*
- Easy to edit manually and human-readable.

Title	Year	Genre	Director	Actors
Cruella	2021	Comedy/Crime/Drama	Craig Gillespie	Emma Stone"/"Emma Thompson"/"Joel Fry
Cast away	2000	Adventure/Drama	Robert Zemeckis	Tom Hanks"/"Helen Hunt"/"Lari White
Joker	2019	Crime/Drame	Todd Phillips	Joaquin Phoenix"/"Robert De Niro"/"Zazie Beetz

- Can be read without difficulty in all programming languages
- Supported by a wide range of applications (Hadoop, Spark, kafka).
- Can be accessed in text editors and in Microsoft Excel, Apple Numbers, Google Sheets, OpenOffice and LibreOffice

- No support for null values, same as empty values
- No guarantee of being splittable
- Non-standardized format, each with its own interpretations
- No support for schema evolution.

XML

- eXtensible Markup Language (XML)
- Most common format
- Support for hierarchical structures

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<root>
  <row>
    <Title>Cruella</Title>
    <Year>2021</Year>
    <Genre>Comedy</Genre>
    <Genre>Crime</Genre>
    <Genre>Drama</Genre>
    <Director firstname="Craig" lastname="Gillespie"/>
    <Actors>Emma Stone</Actors>
    <Actors>Emma Thompson</Actors>
    <Actors>Joel Fry</Actors>
  </row>
</root>
```

- XML is a tree node types: root, elements, text, attribute, comments, ...
- Verbose, especially for human readers

```
−<root>
  −<row>
      <Title>Cruella</Title>
      <Year>2021</Year>
      <Genre>Comedy</Genre>
      <Genre>Crime</Genre>
      <Genre>Drama</Genre>
      <Director firstname="Craig" lastname="Gillespie"/>
      <Actors>Emma Stone</Actors>
      <Actors>Emma Thompson</Actors>
      <Actors>Joel Fry</Actors>
    </row>
  </root>
```

- Platform-independent
- Editing software helps users avoid errors.

- Not splittable.
- Very verbose (large footprint)
- Not very readable
- Redundant
- Higher storage and transport costs when the volume of data is large.

JSON

- **J**ava**S**cript **O**bjet **N**otation (JSON)
- Popular in the web
- Lightweight text format compared to XML
- Support for hierarchical structures

```
[
  {
    "Title": "Cruella",
    "Year": 2021,
    "Genre": [
      "Comedy",
      "Crime",
      "Drama"
    ],
    "Director": {
      "First name": "Craig",
      "Last name": "Gillespie"
    },
    "Actors": [
      "Emma Stone",
      "Emma Thompson",
      "Joel Fry"
    ]
  }
]
```

- Booleans, numbers, strings, arrays, objects (dictionaries)
- Very verbose, but human readable

- Widely used for NoSQL databases (MongoDB, Couchbase, and Azure Cosmos DB).
- Widely supported by many software.
- Relatively easy to implement in several languages.

JSON

Disadvantages

- Not splittable
- Lacks indexing as many text formats
- Verbose (large footprint)
- Not easy to parse

YAML

- **Y**et **A**nother **M**arkup **L**anguage (YAML)
- Commonly used for configuration files
- Support for hierarchical model & simple relational scheme

```yaml
---
- Title: Cruella
  Year: 2021
  Genre:
  - Comedy
  - Crime
  - Drama
  Director:
   First name: Craig
   Last name: Gillespie
  Actors:
  - Emma Stone
  - Emma Thompson
  - Joel Fry
```

- Same as JSON, but uses indentation instead of brackets
- Human-readable (minimal syntax)

- Many programming languages has support for reading and writing YAML
- Some source-code editors make editing YAML easier (Emacs)

- Editing is difficult in case of large files
- The absence of terminators.
- Inconsistent implementations because of the complexity of the standard.

GeoJSON

- **Geo**graphic **JSON** (GeoJSON)
- Encode a variety of geographic data structures.

GeoJSON

- Geometry object: basically the location information.
  1. Point
  2. LineString
  3. Polygon
  4. MultiPoint
  5. MultiLineString
  6. MultiPolygon
  7. GeometryCollection
- Feature object: a geometry object with additional properties.
- FeatureCollection object: a list of feature objects.

```
{
  "type": "Feature",
  "geometry": {
    "type": "Point",
    "coordinates": [125.6, 10.1]
  },
  "properties": {
    "name": "Dinagat Islands"
  }
}
```

- Every object contains a member named “type”
- Almost all of the geometry objects also contain a member named “coordinates” listed in order as:
  - longitude
  - latitude
  - elevation (optional)

Exercise

From the following table create four files in CSV, XML, JSON and YAML format.

Rank	City	Member State	Official population	Date of census
1	Berlin	Germany	3,669,495	31 December 2019
2	Madrid	Spain	3,348,536	1 February 2020
3	Rome	Italy	2,856,133	31 December 2018
4	Paris	France	2,175,601	1 January 2019

Table from wikipedia.org

1. Choose your favorite text editor (Notepad, Textedit, Sublime Text, Notepad++).
2. Enter the text data.
3. Save this file with the appropriate extension.