FAKULTA ELEKTROTECHNICKÁ

České vysoké učení technické v Praze

B4M36DS2 – Database Systems 2

Practical Class 2

Data Formats

Yuliia Prokop

prokoyul@fel.cvut.cz, Telegram **@Yulia_Prokop**

https://cw.fel.cvut.cz/b241/courses/b4m36ds2/start

Data Formats in the DS2 course

CSV (Comma-Separated Values): Plain-text tabular format. Each line is a row; columns are separated by commas. Simple, human-readable, but flat (no nesting) and schema is implicit or external.

XML (eXtensible Markup Language): Textual hierarchical format using tags. Designed to represent nested structures. Verbose; supports attributes, elements, and mixed content.

JSON (JavaScript Object Notation): Lightweight text format representing hierarchical (nested) data via objects and arrays. Widely used in modern systems. Easier to read/write than XML.

BSON (Binary JSON): Binary-encoded serialization of JSON-like documents. Adds type metadata, length prefixes, and is designed for fast parsing and traversal (used e.g., internally by MongoDB).

Parquet: Columnar, binary format designed for analytics. Stores data by column, enabling efficient compression and selective reading. Often used in big data pipelines.

RDF (Resource Description Framework) is a standard model for representing knowledge as **triples** — **subject → predicate → object** — using URIs to identify resources, enabling linked data and semantic interoperability.

Key Differences

Format	Hierarchical?	Human-readable	Schema rigidness	Typical use case	Binary / Text
CSV	No	Yes	Implicit / external	Simple tables, exports/imports	Text
XML	Yes	Yes	Flexible	Document interchange, nested data	Text
JSON	Yes	Yes	Flexible	APIs, document stores, nested documents	Text
BSON	Yes	No	Flexible + typed	Internal storage/transport in NoSQL systems	Binary
Parquet	Yes (logical)	No	Schema-defined	Analytics, large-scale reads	Binary, columnar
RDF	No (graph-based)	Yes	Flexible	Ontologies available via RDFS/OWL Linked Data, semantic web, knowledge graphs, metadata integration Text	Some binary serializations (e.g., HDT)

Syntax Highlights

CSV:

movies.csv

```
title,year
Inception,2010
```

actors.csv

```
title,actor_name,role
Inception,Leonardo DiCaprio,Cobb
Inception,Joseph Gordon-Levitt,Arthur
```

Syntax Highlights

XML:

```
<Movie>
  <Title>Inception</Title>
  <Year>2010</Year>
  <Actors>
    <Actor>
      <Name>Leonardo DiCaprio</Name>
      <Role>Cobb</Role>
    </Actor>
    <Actor>
      <Name>Joseph Gordon-Levitt</Name>
      <Role>Arthur</Role>
    </Actor>
  </Actors>
</Movie>
```

Syntax Highlights

JSON:

```
{
  "title": "Inception",
  "year": 2010,
  "actors": [
    {
      "name": "Leonardo DiCaprio",
      "role": "Cobb"
    },
    {
      "name": "Joseph Gordon-Levitt",
      "role": "Arthur"
    }
  ]
}
```

BSON vs JSON

Illustrative hex breakdown of minimal BSON document for

`{"id":19995,"title":"Avatar"}`:

```
16 00 00 00                                  // total size
10 69 64 00 4b 83 00 00                      // int32 field "id" = 19995
02 74 69 74 6c 65 00                         // string field name "title"
07 00 00 00 41 76 61 74 61 72 00             // length + "Avatar" + null
00                                           // terminator
```

Exercise 1

Convert the given **CSV** data to **JSON** and **XML**

```
id,title,original_language,release_date,vote_average,vote_
count,status,tagline
19995,Avatar,en,2009-12-10,7.2,11800,Released,"Enter the
World of Pandora."
```

8

Exercise 1 - Solution

Convert the given **CSV** data to **JSON** and **XML**

```
id,title,original_language,release_date,vote_average,vote_count,s
tatus,tagline
19995,Avatar, en,2009-12-10,7.2,11800,Released,"Enter the World
of Pandora."
```

```
<movies>
  <movie>
    <id>19995</id>
    <title>Avatar</title>
    <original_language>en</original_language>
    <release_date>2009-12-10</release_date>
    <vote_average>7.2</vote_average>
    <vote_count>11800</vote_count>
    <status>Released</status>
    <tagline>Enter the World of Pandora.</tagline>
  </movie>
</movies>
```

Exercise 1 - Solution

JSON

```
[
  {
    "id": 19995,
    "title": "Avatar",
    "original_language": "en",
    "release_date": "2009-12-10",
    "vote_average": 7.2,
    "vote_count": 11800,
    "status": "Released",
    "tagline": "Enter the World of Pandora."
  }
]
```

10

Embedding vs Referencing

Embedded (movie contains full cast objects):

```
{
  "id": 19995,
  "title": "Avatar",
  "cast": [
    { "person_id": 992,
      "name": "Zoe Saldana",
      "character": "Neytiri",
      "cast_id": 15 }
  ]
}
```

***Pros**:* simple movie-detail read.
***Cons**:* duplication when an actor appears in many movies, harder centralized updates.

Embedding vs Referencing

Referenced (separate person documents):

Movie:

```
{ "id": 19995, "title": "Avatar", "cast_person_ids": [992] }
```

Person:

```
{ "person_id": 992,
  "name": "Zoe Saldana",
  "movie_ids": [19995, 12345]
}
```

Pros: normalized, efficient actor-centric queries.

Cons: requires lookups to assemble full movie view.

Embedding vs Referencing

Rules:

- Embed when movie-detail access is dominant and embedded data is stable.
- Reference when related entities are reused or updated independently.
- Hybrid: cache embedded snapshots with invalidation from authoritative referenced documents.

13

Exercise 2

Convert the given verbal description to **CSV**, **JSON** and **XML**

There are two feature films to be archived.

The first, identified as **MovieID 8101** and titled “**Nebula’s Edge**,” was released on 2021-11-12, runs 142 minutes, and belongs to the genres Sci-Fi, Adventure, and Drama.

Its original language is English, while the spoken languages include English and Spanish; it was produced in the USA and Canada.

The production budget amounted to USD 120,000,000.

At the box office, it earned a total of USD 280,000,000, of which the United States contributed USD 98,500,000, Canada USD 12,300,000, the United Kingdom USD 24,700,000, the rest of the European Union excluding the UK USD 83,200,000, Japan USD 14,900,000, Australia USD 9,800,000, and the rest of the world USD 36,600,000.

The credited cast, in billing order, is as follows: Mira Kaur as Cmdr. Ayesha Rao, Daniel Brooks as Elias Ren, Lucía Márquez as Dr. Sofia Vidal, Kenji Mori as Akira Tanaka, and Eva Petrov as Lt. Nadia Volkova.

The principal crew lists Alina Park as director, Alina Park and Jonah Tse as writers, and Marco Alvarez as composer.

The film carries a PG-13 rating and is tagged with the keywords “deep-space mission,” “rogue AI,” and “family.”

Exercise 2-2

Convert the given verbal description to **CSV**, **JSON** and **XML**

The second film, **MovieID 8102**, is titled “**Iron Vale**” and was released on 2023-05-26. It has a runtime of 156 minutes and falls under the genres Action, Thriller, and Mystery.

Its original language is English, with spoken languages including English, French, and German.

The production countries are the UK and Germany.

The budget was USD 95,000,000.

The worldwide box office totaled USD 237,500,000, with regional grosses as follows: United States USD 76,200,000; United Kingdom USD 31,400,000; Germany USD 22,800,000; the remainder of the European Union excluding the UK and Germany USD 46,100,000; Canada USD 9,300,000; Australia USD 7,600,000; Japan USD 15,900,000; and the rest of the world USD 28,200,000.

The cast, in billing order, comprises Tomiwa Adeyemi as Inspector Kade, Sofia Richter as Lena Weiss, Karim Haddad as Yusuf Amari, Aiko Tanaka as Michiko Sato, and Ethan Cole as Marcus Hargrove. The crew credits list Rowan Whitaker as director, Rowan Whitaker and Elise Durant as writers, and Yuki Nakamura as composer.

The rating is PG-13, and the assigned keywords are “cold case,” “industrial town,” and “double-cross.”

Exercise 2 - Solution

See Exercise2 Solution
https://cw.fel.cvut.cz/b251/_media/courses/b4m36ds2/ex2_csv_json_xml.txt

CSV Validator https://toolkitbay.com/tkb/tool/csv-validator

JSON Validator https://codebeautify.org/jsonvalidator

XML Validator https://jsonformatter.org/xml-validator