

Методи та засоби реалізації стадії збору та попереднього опрацювання даних

Виконали:
Косарева А.С.
Яснєв А.С.
Кожевніков Я.І.



Формати даних та метаданих

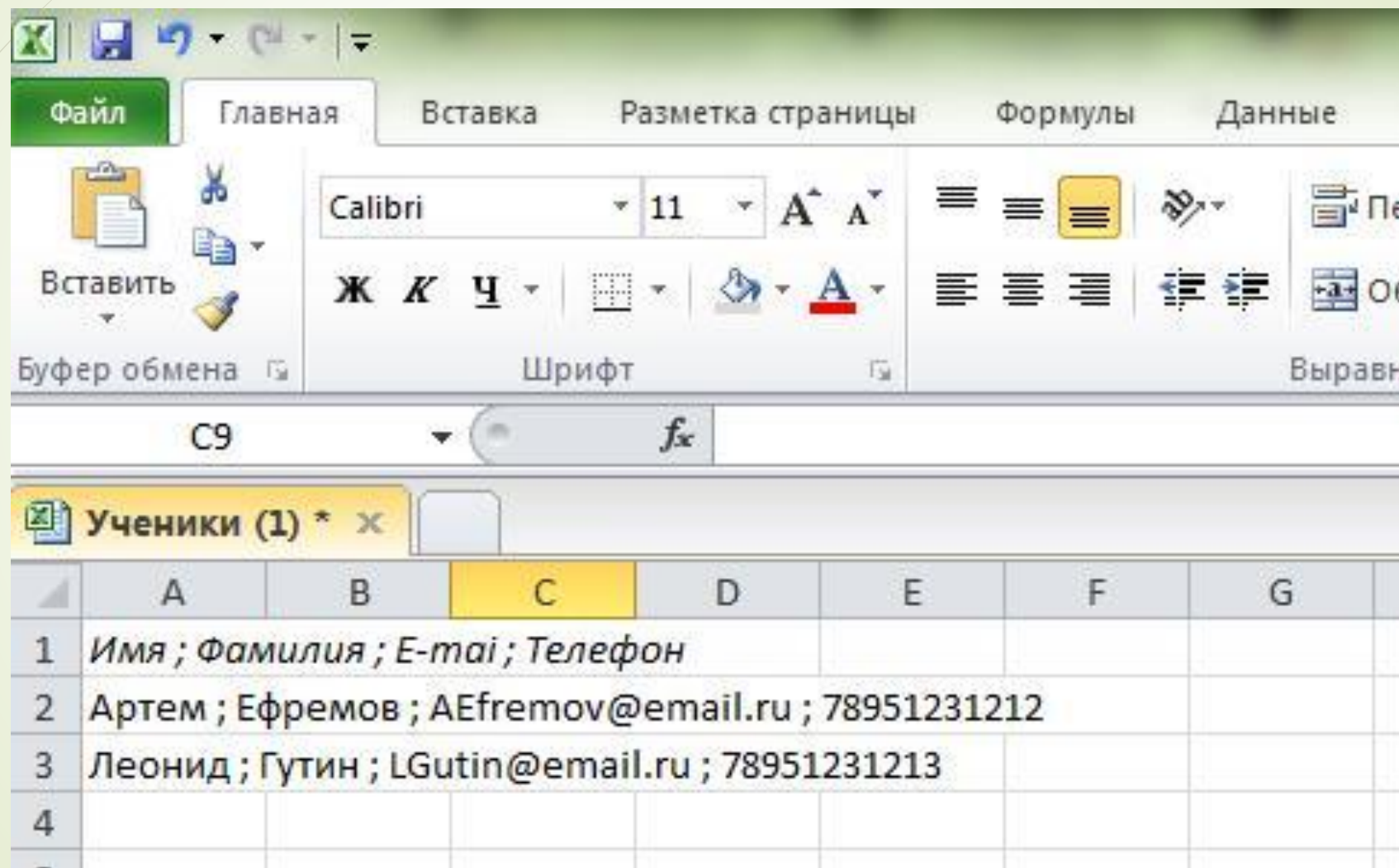
Такі як: csv, xml, xlsx, txt, json; ddf, json-schema

CSV



- **comma-separated values** 'значення, розділені комою',
character-separated values 'значення, розділені символом'
- Файловий формат для представлення табличних даних
- Використовується для перенесення даних між базами даних та програмами-редакторами електронних таблиць

Приклад файла CSV-формату:



The screenshot shows the Microsoft Excel interface with the 'Ученики (1)' file open. The ribbon is set to 'Главная' (Home). The font is 'Calibri' and the size is '11'. The text is left-aligned. The data is entered in the following format:

	A	B	C	D	E	F	G
1	Имя ;	Фамилия ;	E-mai ;	Телефон			
2	Артем ;	Ефремов ;	AEfremov@email.ru ;	78951231212			
3	Леонид ;	Гутин ;	LGutin@email.ru ;	78951231213			
4							

XML



- **Extensible Markup Language** – розширювана мова розмітки
- Спрощена підмножина мови розмітки SGML
- Набір базових лексичних та синтаксичних правил для побудови мови описання інформації шляхом застосування простих тегів

Приклад файлу XML-формату:

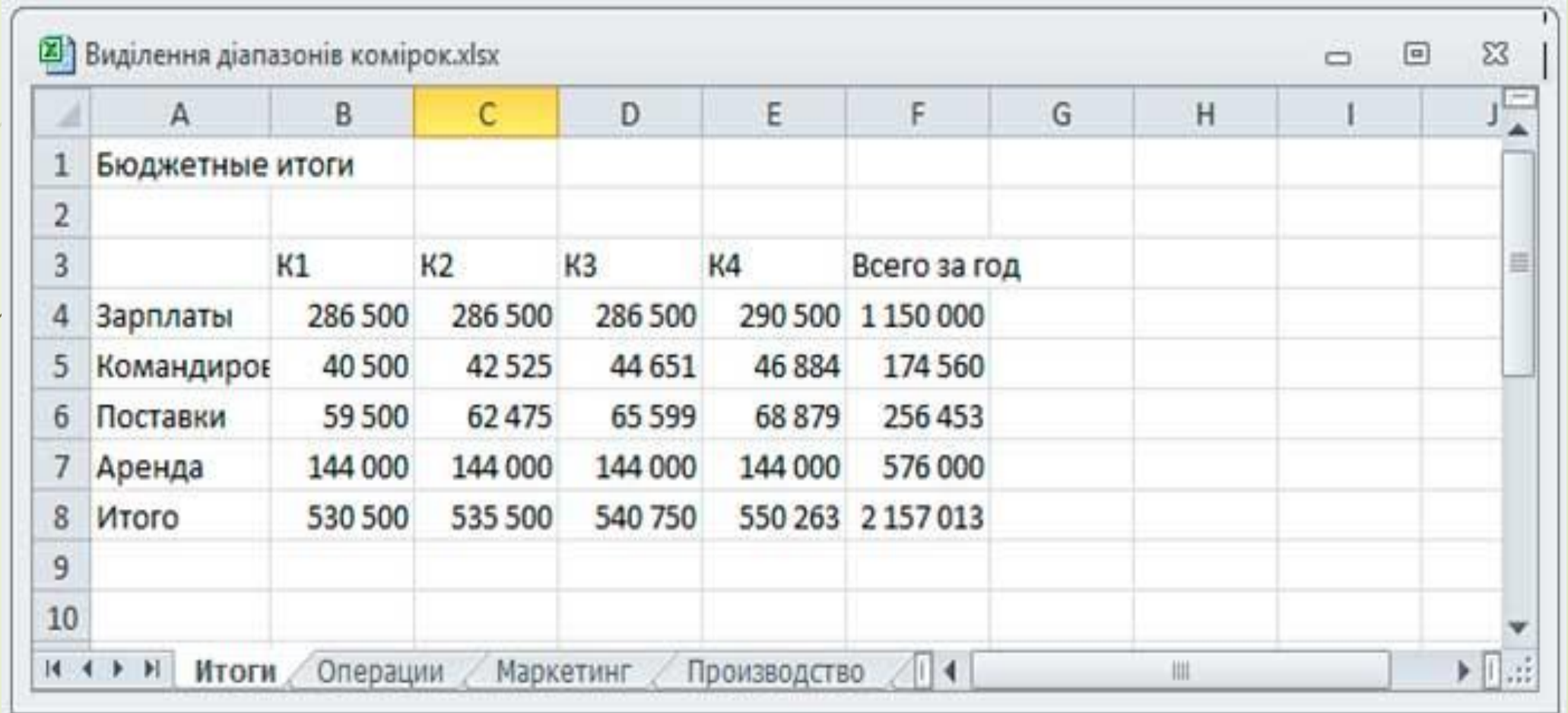
```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema attributeFormDefault="unqualified" elementFormDefault="qualified"
  xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="points">
    <xs:complexType>
      <xs:sequence>
        <xs:element maxOccurs="unbounded" name="point">
          <xs:complexType>
            <xs:attribute name="x" type="xs:unsignedShort" use="required" />
            <xs:attribute name="y" type="xs:unsignedShort" use="required" />
          </xs:complexType>
        </xs:element>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```


XLSX



- Стандартний формат робочих книг Excel 2007. У дійсності це стислий ZIP-архів каталогу XML-документів. Є заміною колишнього бінарного формату .xls, хоча і не підтримує макроси з міркувань безпеки.

Приклад файлу XSLX-формату:



Виділення діапазонів комірок.xlsx

	A	B	C	D	E	F	G	H	I	J
1	Бюджетные итоги									
2										
3		K1	K2	K3	K4	Всего за год				
4	Зарплаты	286 500	286 500	286 500	290 500	1 150 000				
5	Командирове	40 500	42 525	44 651	46 884	174 560				
6	Поставки	59 500	62 475	65 599	68 879	256 453				
7	Аренда	144 000	144 000	144 000	144 000	576 000				
8	Итого	530 500	535 500	540 750	550 263	2 157 013				
9										
10										

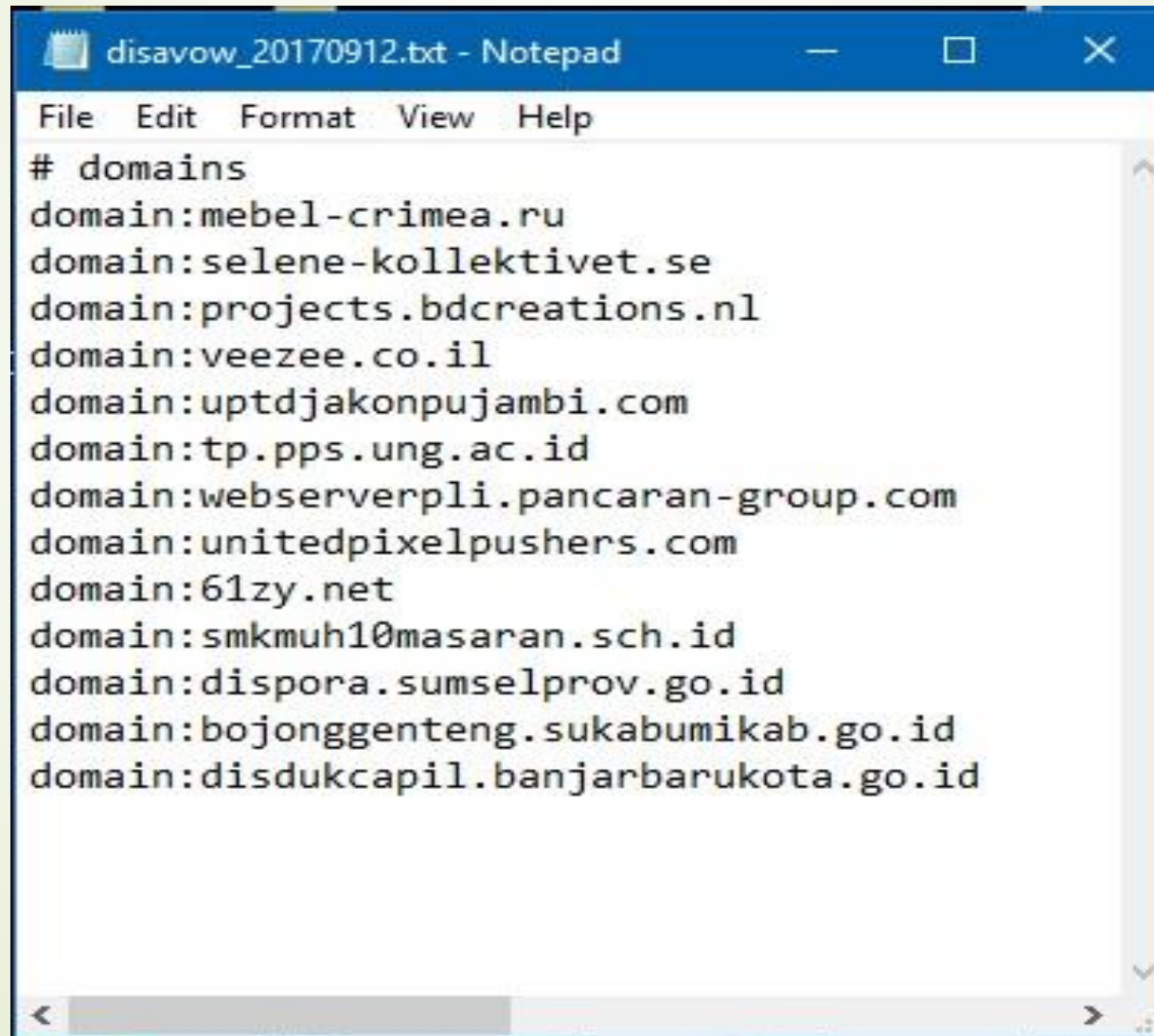
Итоги | Операции | Маркетинг | Производство

TXT



- Комп'ютерний файл, що зберігає текстові дані.
- Послідовність символів, що згрупована у рядки.
- Змінюється завдяки текстовому редактору, що є на всіх ОС.

Приклад файлу TXT-формату:

A screenshot of a Notepad window titled "disavow_20170912.txt - Notepad". The window has a standard menu bar with "File", "Edit", "Format", "View", and "Help". The text content is as follows:

```
# domains
domain:mebel-crimea.ru
domain:selene-kollektivet.se
domain:projects.bdc creations.nl
domain:veezee.co.il
domain:uptdjakonpujambi.com
domain:tp.pps.ung.ac.id
domain:webserverpli.pancaran-group.com
domain:unitedpixelpushers.com
domain:61zy.net
domain:smkmuh10masaran.sch.id
domain:dispورا.sumselprov.go.id
domain:bojonggenteng.sukabumikab.go.id
domain:disdukcapil.banjarbarukota.go.id
```

The text is in a monospaced font. The window includes a scrollbar on the right and a status bar at the bottom.



JSON

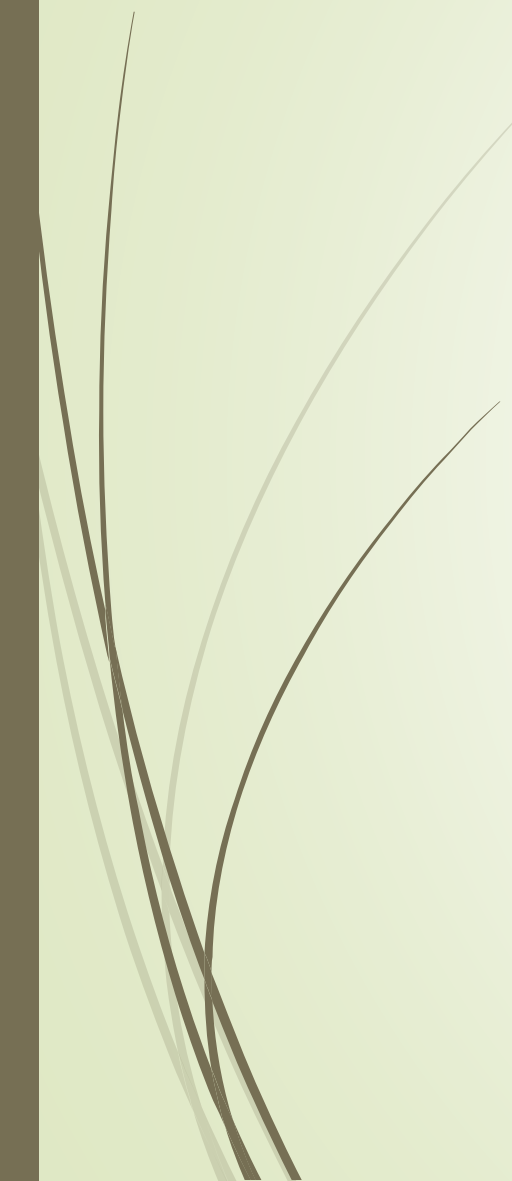
- *JavaScript Object Notation*
- Текстовий формат обміну даними між комп'ютерами
- JSON базується на тексті, може бути прочитаним людиною
- Формат дозволяє описувати об'єкти та інші структури даних
- Розробив і популяризував формат Дуглас Крокфорд.

Приклад файлу JSON-формату:

```
{
  "Time"      : "2014-02-12 14:20:05",
  "Latitude"  : 37.33233141,
  "Longitude" : -122.0312186,
  "Count"     : 101,
  "Comments"  : "Bad Data. SNOW DAY!!",
  "Luma"      : 0,
  "Habitat"   : "Back yard, grass",
  "Types"     : [
    "5",
    "6"
  ],
  "Address"   : {
    "Street"   : "2522 West Georgia Road",
    "City"     : "Piedmont",
    "State"    : "South Carolina",
    "Country"  : "United States",
  }
}
```



JSON-Schema

- Мова описання структури JSON-документа
 - Використовує синтаксис JSON
 - Базується на концепціях XML Schema, та ін.
- 

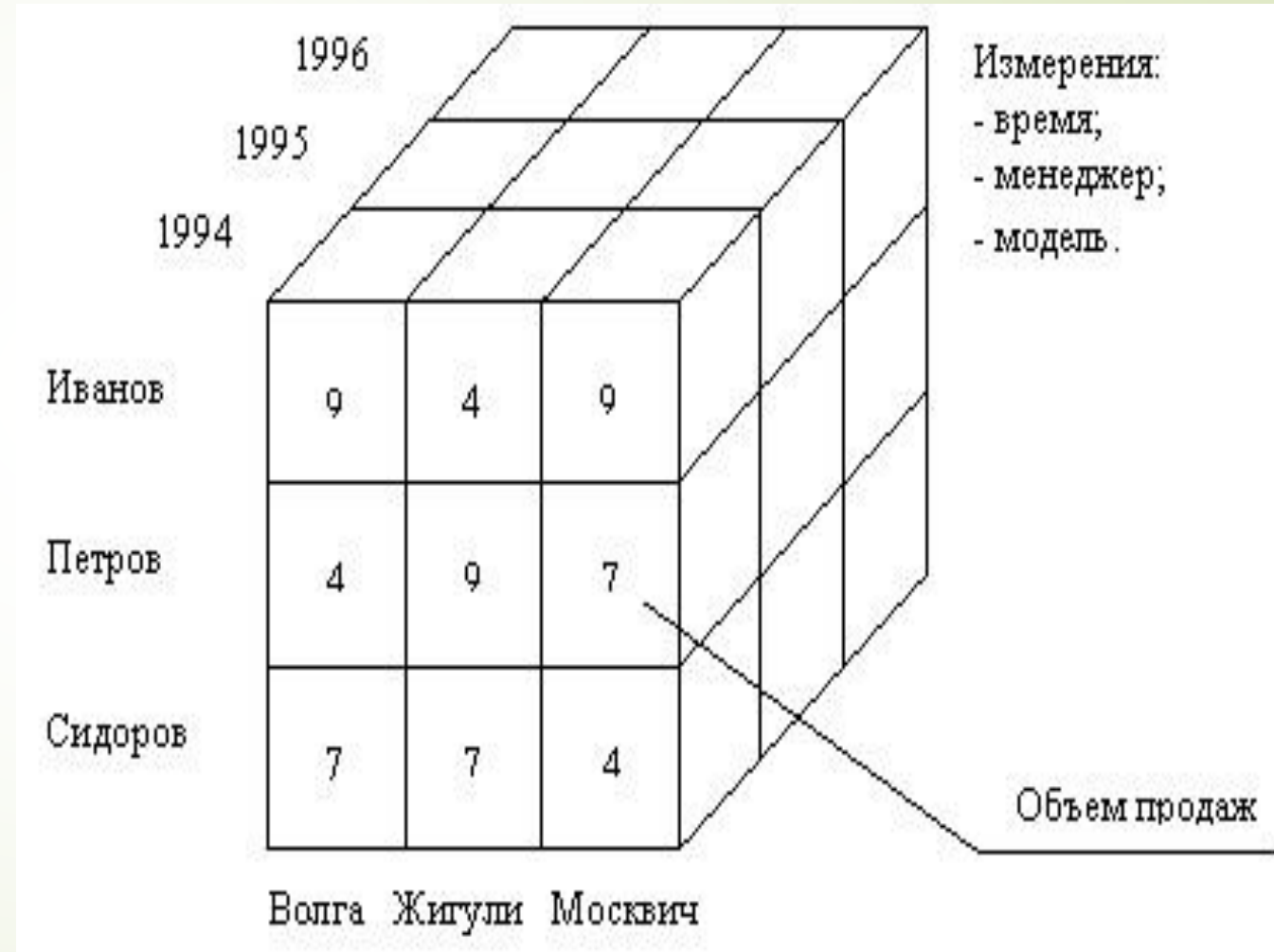



JSON-STAT

- Формат JSON-stat - це простий формат JSON для поширення даних
- Базується на кубічній моделі
- Найбільш вживною формою поширення даних є таблична форма.
- У цій моделі кубів набори даних організовані за розмірами.
- Розміри організовані за категоріями.
- Використовує синтаксис JSON

Приклад моделі куба

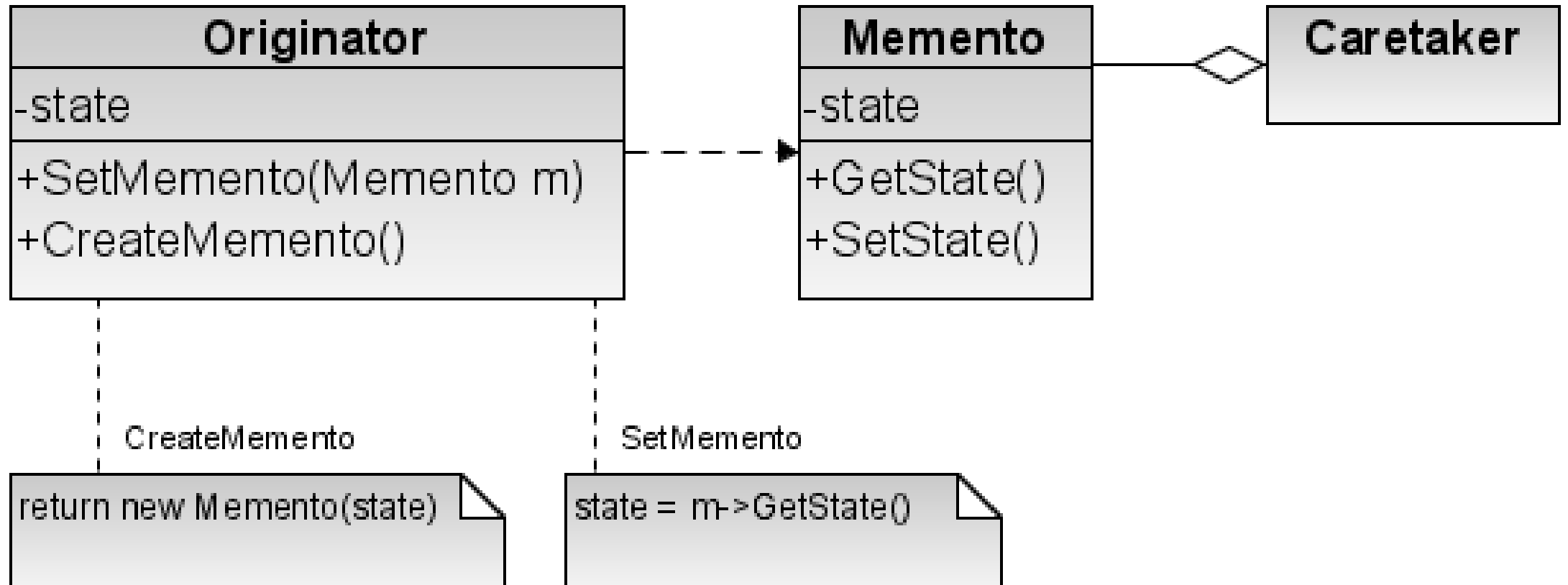
- Для кожного виміру створюється об'єднання, що з'єднує його з фактографічним об'єктом





Використання шаблонів
проектування для реалізації
підсистеми збору,
перетворення та перевірки
даних

MEMENTO



Data Mapper

