https://itstep.by/

Урок 4
(Аналитик Данных)
Data Analyst

Dr. Sergey Postnikov

Сергей Постников

Iteration 2, 17-Jan-2022, Sergey P.

Содержание

- …
- Экосистема данных аналитика
- Типы данных
- Типы форматов файлов
- Источники данных
- Языки обработки данных
- Обзор хранилищ данных
- RDBMS, NoSQL
- Data Marts, Data Lakes, ETL, and Data Pipelines
- Foundations of Big Data
- Big Data Processing Tools
- …

* извлекать (mine)

** обрабатывать (wrangling)

Итоги

Вы узнаете о **роли, обязанностях и наборах навыков**, необходимых для работы аналитиком данных, и о том, как выглядит **обычный день из жизни аналитика** данных.

Вы также узнаете о различных **типах** структур данных, **форматах** файлов, **источниках** данных и **языках**, которые специалисты по данным используют в своих повседневных задачах.

Вы получите представление о различных типах хранилищ данных, таких как базы данных, хранилища данных, витрины данных, озера данных и конвейеры данных.

Кроме того, вы узнаете о процессе (ETL), который используется для извлечения, преобразования и загрузки данных в хранилища данных.

Вы получите базовое представление о больших данных и инструментах обработки больших данных, таких как Hadoop, распределенная файловая система Hadoop (HDFS), Hive и Spark.

Цели

Понять, как может выглядеть день из жизни аналитика данных

Описывать и различать реляционные и нереляционные системы управления базами данных.

Объяснять различные типы структур данных, форматы файлов и источники данных.

Объяснить особенности и использование различных языков, используемых специалистами по данным.

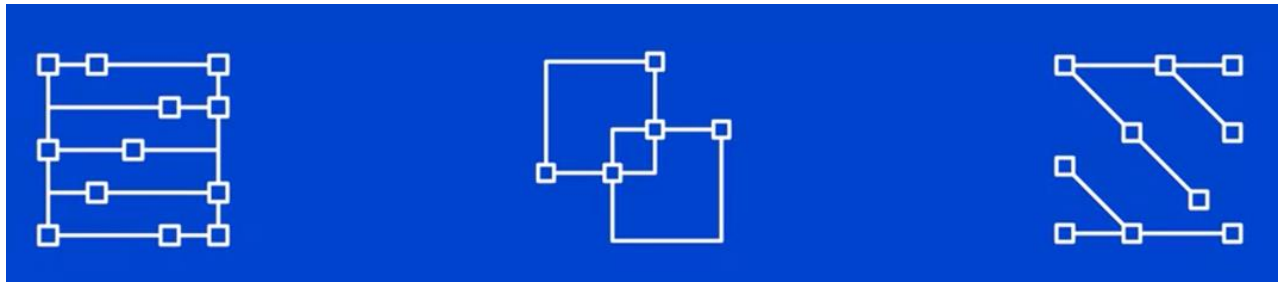
Описывать, как работают хранилища данных, витрины данных, озера данных и конвейеры данных.

Объяснять, как работает процесс извлечения, преобразования и загрузки, чтобы подготовить необработанные данные для анализа.

Объяснять, что такое большие данные, и обобщить особенности и использование некоторых инструментов обработки больших данных.

Какой тип данных обычно используется в базах данных и электронных таблицах??

1. Unstructured data
2. Structured data
3. Social media content
4. Semi-structured data



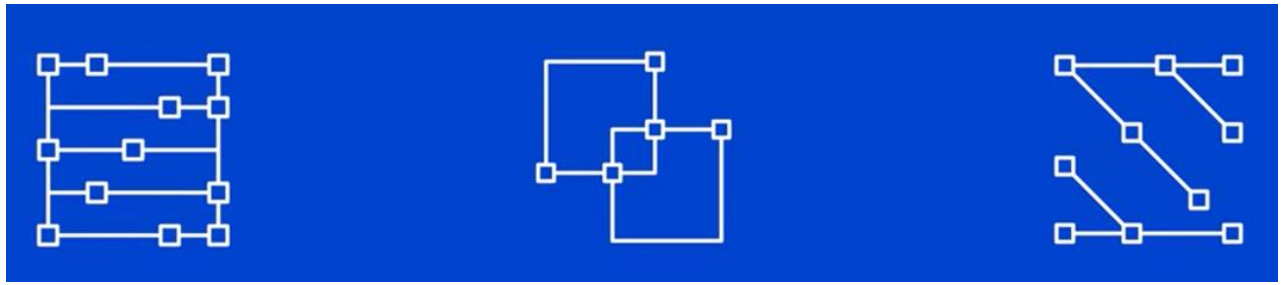
Какой тип данных обычно используется в базах данных и электронных таблицах?

1. Unstructured data
2. Structured data
3. Social media content
4. Semi-structured data

Структурированные данные, то есть данные, которые можно аккуратно организовать в строки и столбцы, обычно хорошо подходят для баз данных и электронных таблиц.

Какой из этих источников данных является примером полуструктурированных данных?

1. Social media feeds
2. Documents
3. Emails
4. Network and web logs



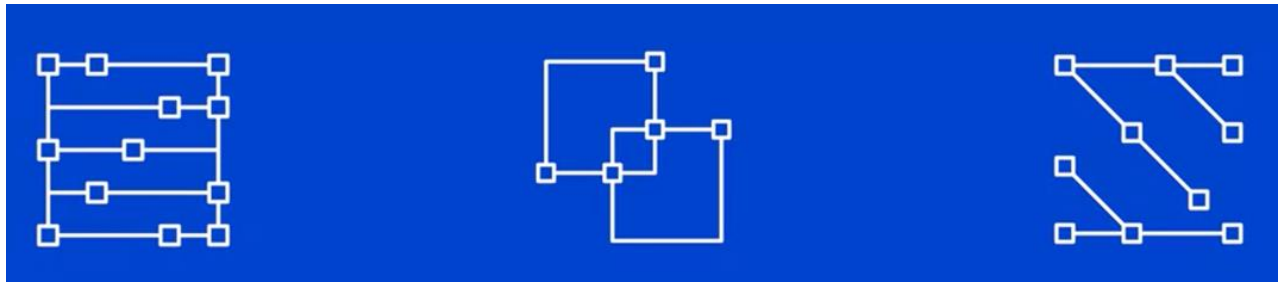
Какой из этих источников данных является примером полуструктурированных данных?

1. Social media feeds
2. Documents
3. Emails
4. Network and web logs

Электронные письма являются источником полуструктурированных данных, поскольку они содержат данные с некоторыми организационными свойствами, но не следуют жесткой схеме.

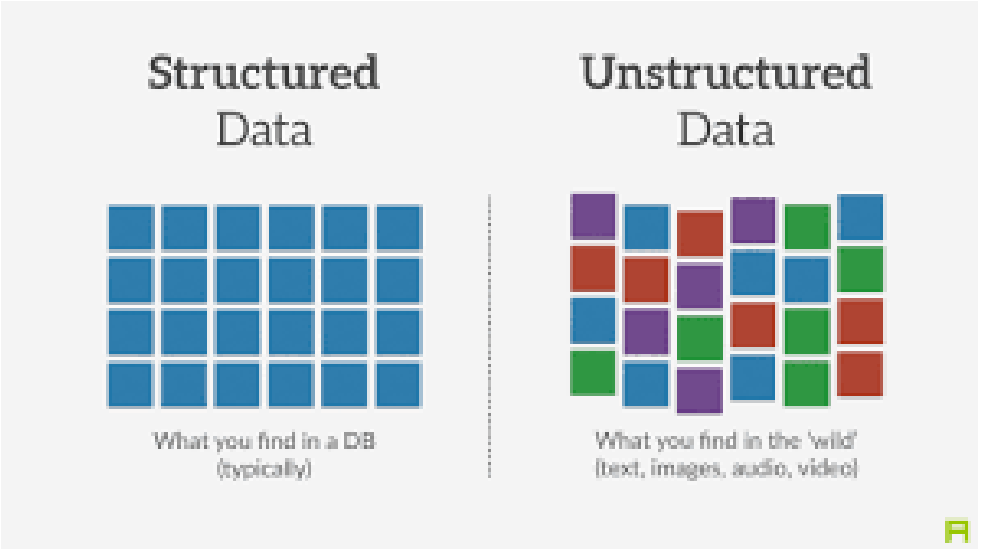
Что из следующего является примером неструктурированных данных?

1. Zipped files
2. Spreadsheets
3. XML
4. Video and audio files



Что из следующего является примером неструктурированных данных?

1. Zipped files
2. Spreadsheets
3. XML
4. Video and audio files



Обзор экосистемы данных **аналитика**

Экосистема аналитика данных включает

- инфраструктуру,
- программное обеспечение,
- инструменты,
- фреймворки
- и процессы , используемые для
  - сбора,
  - очистки,
  - анализа (& mine - извлечения),
  - обработки (wrangling)
  - и визуализации данных.

Overview

A Data Analyst's ecosystem includes the infrastructure, software, tools, frameworks, and processes used to

Gather Data

Clean Data

Mine Data

Visualize Data

В зависимости от того, насколько четко определена структура данных, данные можно разделить на следующие категории:

- **Структурированные** данные, то есть данные, которые хорошо организованы в форматах, которые можно хранить в базах данных.

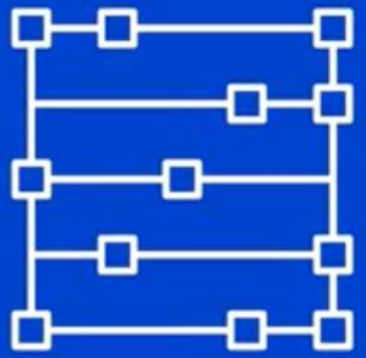
- **Полуструктурированные** данные, то есть данные в частично организованной и частично в свободной форме.

- **Неструктурированные** данные, то есть данные, которые не могут быть организованы обычным образом в строки и столбцы.

Data

имя отправителя и получателя, но также содержит содержимое электронной почты

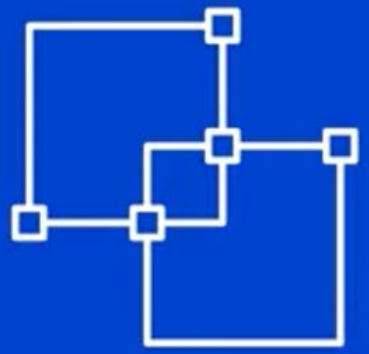
фотографии, видео, текстовые файлы, PDF-файлы и контент в социальных сетях



Structured

Data that follows a rigid format and can be organized into rows and columns.

обычно отображаются в базах данных и электронных таблицах



Semi-structured

Mix of data that has consistent characteristics and data that does not conform to a rigid structure.

Unstructured

Data that is complex and mostly qualitative information that cannot be structured into rows and columns.

Тип данных определяет

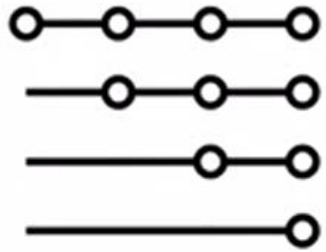
- **тип хранилищ** данных, в которых данные могут быть собраны и сохранены,

- а также **инструменты**, которые могут использоваться для запроса или обработки данных.

Data

Data can come in a variety of file formats, such as

Relational Database



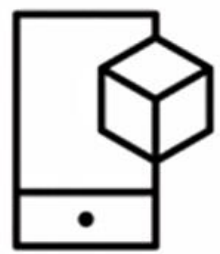
Non-Relational Database

APIs

Web Services

Data Streams

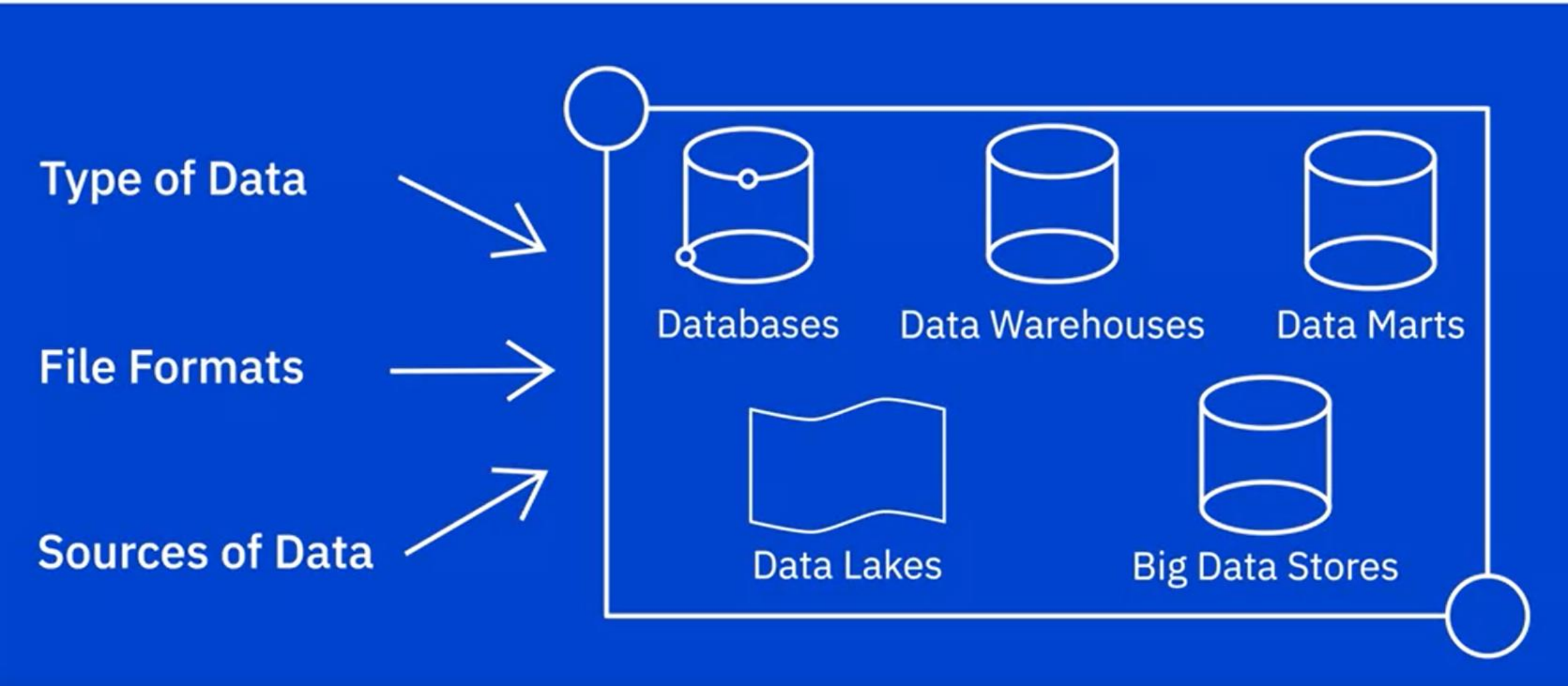
Social Platforms



Sensor Devices

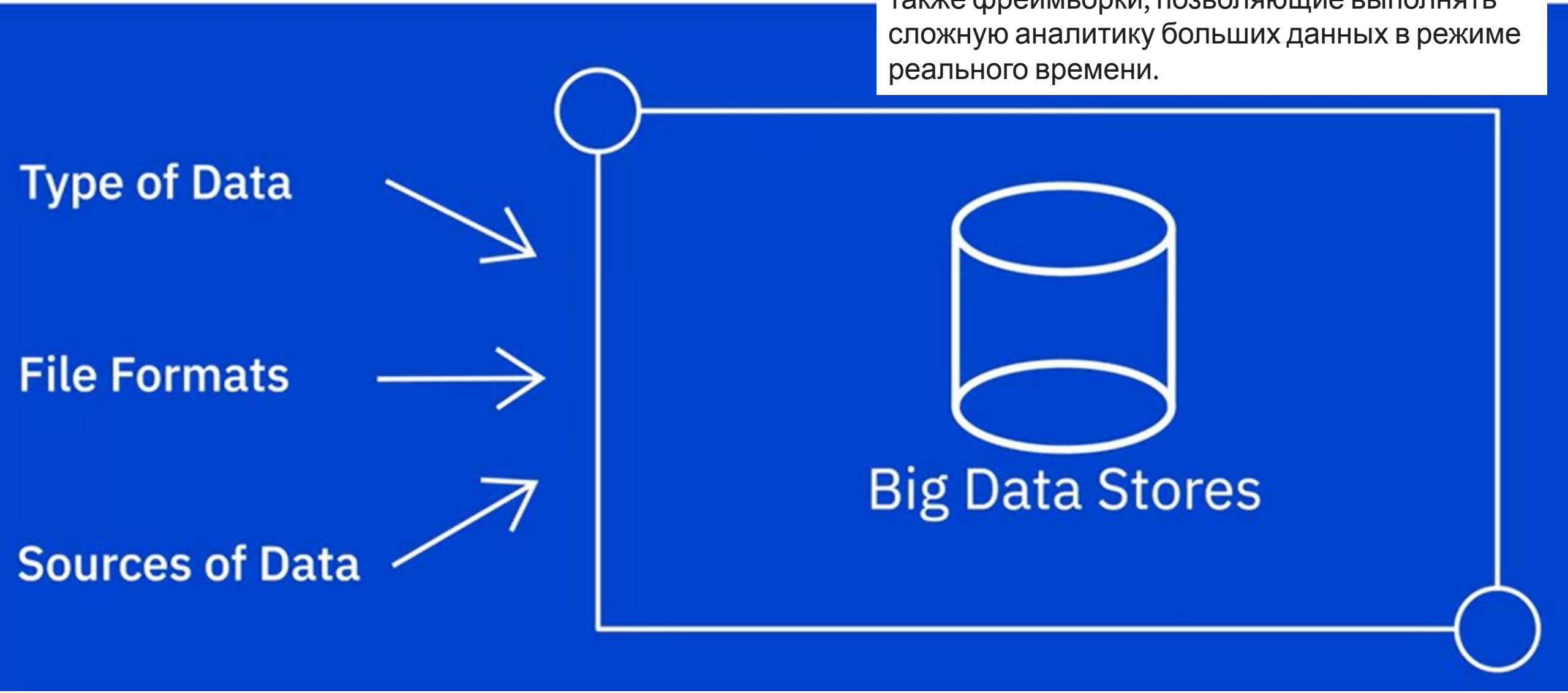
Data Repositories

Тип, формат и источники данных влияют на тип хранилищ данных, которые можно использовать для сбора, хранения, очистки, анализа и извлечения данных для анализа.



Data Repositories

если вы работаете с большими данными, вам понадобятся хранилища больших данных, которые позволяют хранить и обрабатывать большие объемы высокоскоростных данных, а также фреймворки, позволяющие выполнять сложную аналитику больших данных в режиме реального времени.



Languages

Languages available in the Data Analyst Ecosystem:

Query languages
For example, SQL for querying and manipulating data

Programming languages
For example, Python for developing data applications

Shell and Scripting languages
For repetitive operational tasks

Data Analysts Ecosystem

Automated tools, frameworks, and processes for all stages of the analytics process are part of the Data Analysts ecosystem

Электронные таблицы, ноутбуки Jupyter и IBM Cognos Analytics …

Gathering, Extracting, Transforming, and Loading Data

Data Wrangling and Cleaning

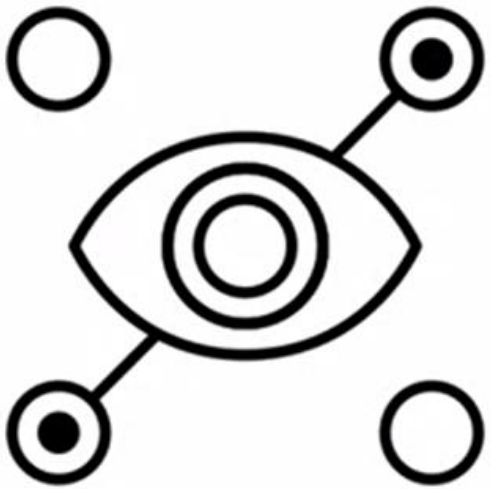
Data Analysis and Mining

Data Visualization

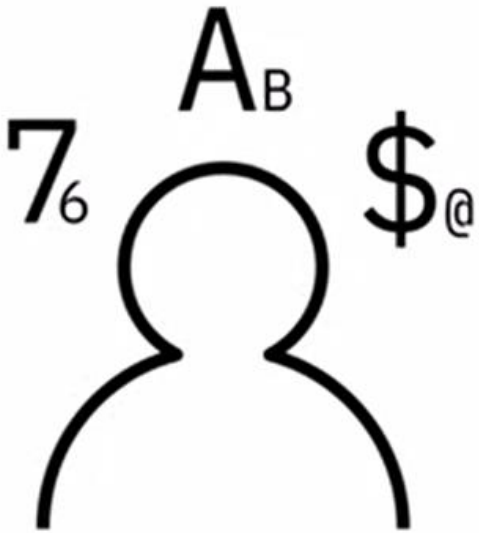
Типы данных

Данные — это **неорганизованная** информация, которая **обрабатывается**, чтобы сделать ее **ценной** и значимой.

What is Data?



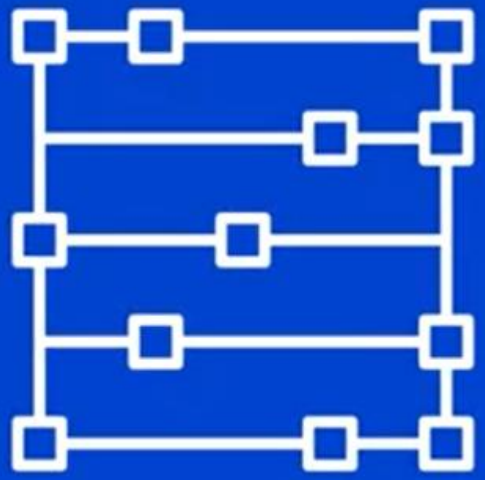
Facts
Observations
Perceptions



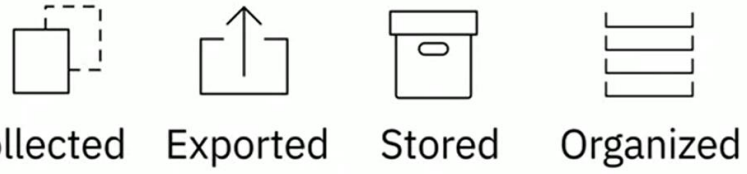
Numbers
Characters
Symbols

Images

Structured Data



Facts Numbers

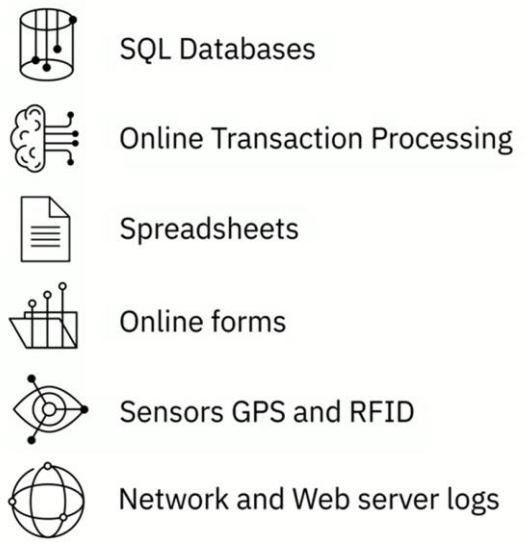


можете легко изучить **структурированные** данные с помощью стандартных методов и **инструментов** анализа данных

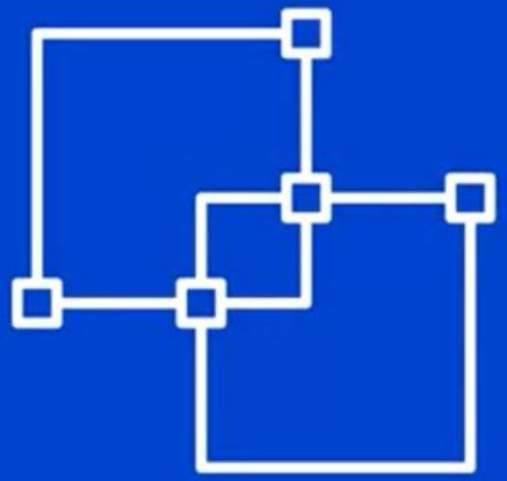
- Has a well-defined structure
- Can be stored in well-defined schemas
- Can be represented in a tabular manner with rows and columns

OLTP

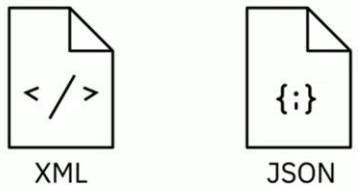
Excel и Google



Semi-Structured Data



- Has some organizational properties but lacks a fixed or rigid schema
- Cannot be stored in the form of rows and columns as in databases
- Contains tags and elements, or metadata, which is used to group data and organize it in a hierarchy



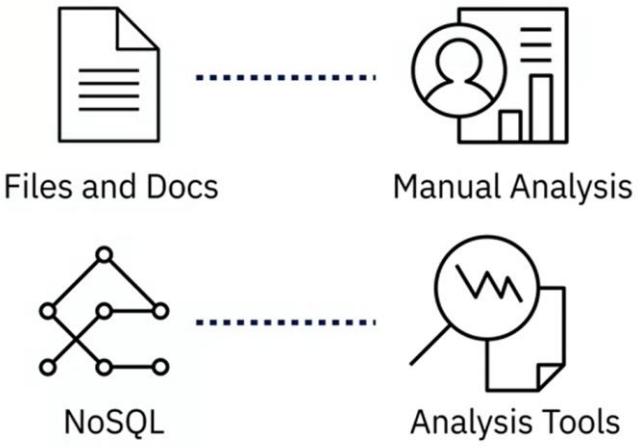
Allow users to

широко используются для хранения и **обмена**

Из разных источников

Unstructured Data

- Does not have an easily identifiable structure
- Cannot be organized in a mainstream relational database in the form of rows and columns
- Does not follow any particular format, sequence, semantics, or rules



могут справляться с **неоднородностью** источников и имеют множество приложений для бизнес-аналитики и аналитики

- Web pages
- Social media feeds
- Images in varied file formats
- Video and Audio files
- Documents and PDF files
- PowerPoint presentations
- Media logs
- Surveys

To summarize

- Structured data is data that is well organized in formats that can be stored in databases and lends itself to standard data analysis methods and tools
- Semi-structured data is data that is somewhat organized and relies on meta tags for grouping and hierarchy
- Unstructured data is data that is not conventionally organized in the form of rows and columns in a particular format

Типы форматов файлов

Важно понимать

- базовую структуру форматов файлов,

а также их

- преимущества и ограничения.

Это понимание поможет вам принимать правильные **решения** по форматам, наилучшим образом **подходящим** для ваших

- **данных**
- и **производительности**.

Standard file formats:

1. Delimited text file formats, or .CSV
2. Microsoft Excel Open .XML Spreadsheet, or .XLSX
3. Extensible Markup Language, or .XML
4. Portable Document Format, or .PDF
5. JavaScript Object Notation, or .JSON

Delimited text files

```
Manufacturer, Model, Sales_in_thousands, __year_resale_value, Vehicle_type, Price_in_thousands
Acura, Integra, 16.919, 16.36, Passenger, 21.5
Acura, TL, 39.384, 19.875, Passenger, 28.4
Acura, CL, 14.114, 18.225, Passenger, 14
Acura, RL, 8.588, 29.725, Passenger, 42
Audi, A4, 20.397, 22.255, Passenger, 23.99
Audi, A6, 18.78, 23.555, Passenger, 33.95
Audi, A8, 1.38, 39, Passenger, 62
BMW, 323i, 19.747, Passenger, 26.99
BMW, 328i, 9.231, 28.675, Passenger, 33.4
BMW, 528i, 17.527, 36.125, Passenger, 38.9
Buick, Century, 91.561, 12.475, Passenger, 21.975
```

.CSV

.TSV

Manufacturer	Model	Sales_in_thousands	__year_resale_value
Acura	Integra	16.919	16.36 Passenger 21.5
Acura	TL	39.384	19.875 Passenger 28.4
Acura	CL	14.114	18.225 Passenger
Acura	RL	8.588	29.725 Passenger 42
Audi	A4	20.397	22.255 Passenger 23.99
Audi	A6	18.78	23.555 Passenger 33.95
Audi	A8	1.38	39 Passenger 62
BMW	323i	19.747	Passenger 26.99
BMW	328i	9.231	28.675 Passenger 33.4
BMW	528i	17.527	36.125 Passenger 38.9
Buick	Century	91.561	12.475 Passenger 21.975

Delimiters also represent one of various means to specify boundaries i

Files used to store data as text

Each value is separated by a delimiter

Delimiter - A sequence of one or more characters for specifying the boundary between independent entities or values.

Comma, Tab, Colon, Vertical Bar, Space

Они могут обрабатываться практически всеми существующими приложениями.

Comma-separated values

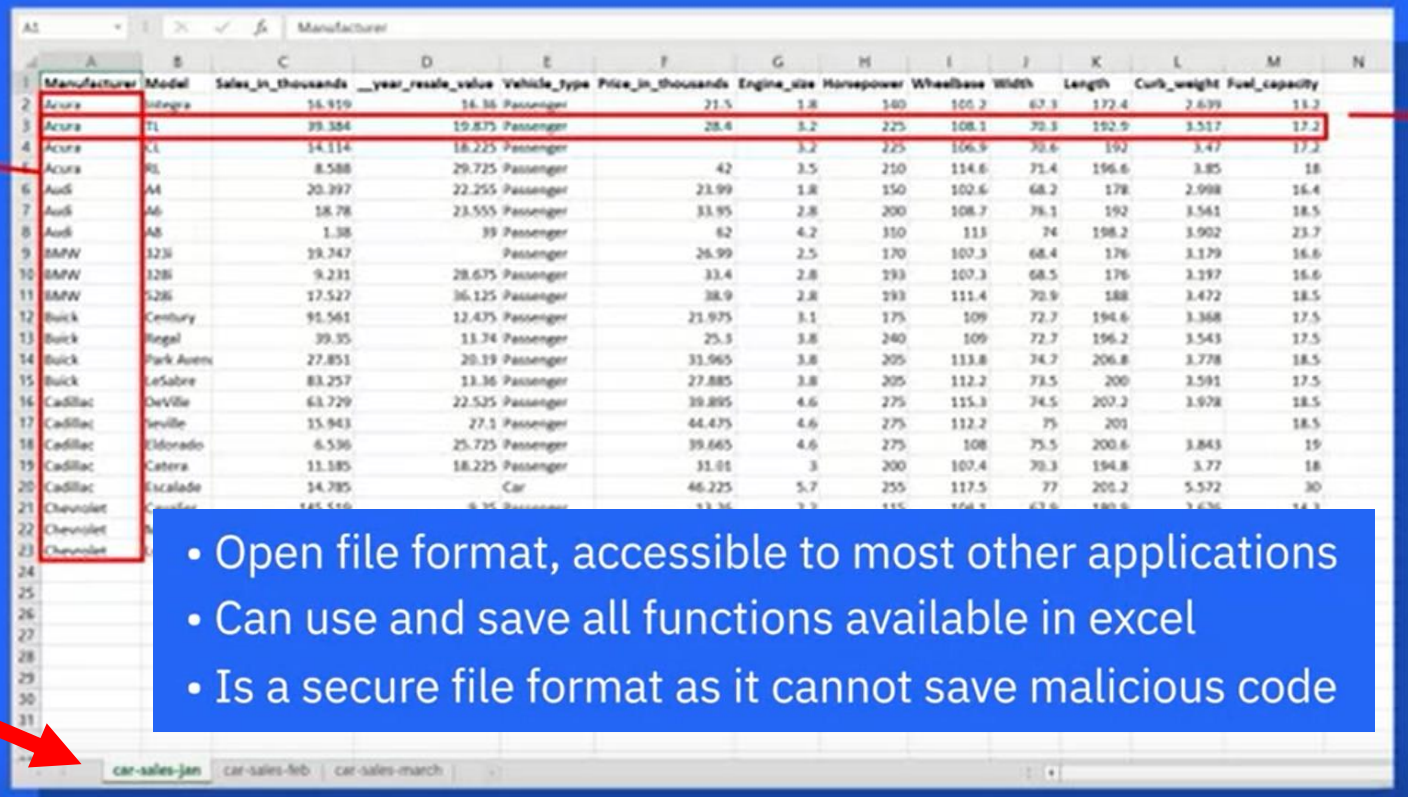
Tab-separated values

Microsoft Excel Open XML Spreadsheet, or .XLSX

Is a Microsoft Excel Open XML file format that falls under the spreadsheet file format. It is an XML-based file format created by Microsoft.

Columns

Rows



- Open file format, accessible to most other applications
- Can use and save all functions available in excel
- Is a secure file format as it cannot save malicious code

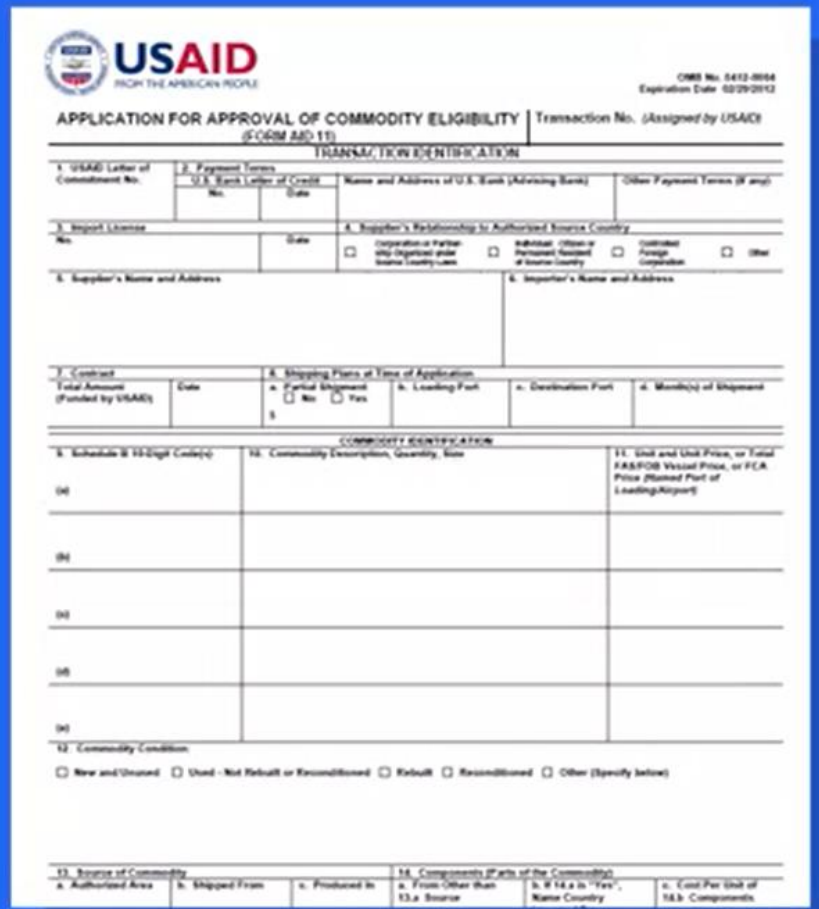
Extensible Markup Language or .XML

```
<?xml version="1.0"?>
<car-specs>

<manufacturer>Acura<manufacturer>

<model>Integra<model>

<sales_in-thousands>16.919<sales_in-thousands>

<year_resale_value>16.36<year_resale_value>

<vehicle_type>Passenger<vehicle_type>

<car-specs>
```

Extensible Markup Language, or XML, is a markup language with set rules for encoding data.

- Readable by both humans and machines
- Self-descriptive language
- Similar to .HTML in some respects
- Does not use predefined tags like .HTML does
- Platform independent
- Programming language independent
- Makes it simpler to share data between systems

Portable Document Format or PDF



USAID
FROM THE AMERICAN PEOPLE

OMB No. 0412-0016
Expiration Date: 02/29/2012

APPLICATION FOR APPROVAL OF COMMODITY ELIGIBILITY (FORM AID 11) | Transaction No. (Assigned by USAID)

TRANSACTION IDENTIFICATION

1. USAID Letter of Commitment No.
2. Payment Terms
U.S. Bank Letter of Credit: No. / Date
Name and Address of U.S. Bank (Advising Bank)
Other Payment Terms (If any)
3. Import License: No. / Date
4. Supplier's Relationship to Authorized Source Country
☐ Corporation or Partner Wholly Organized under Source Country Laws ☐ Individual Citizen or Permanent Resident of Source Country ☐ Controlled Foreign Corporation ☐ Other
5. Supplier's Name and Address
6. Importer's Name and Address
7. Contract: Total Amount (Funded by USAID) / Date
8. Shipping Plans at Time of Application
a. Partial Shipment ☐ No ☐ Yes
b. Loading Port
c. Destination Port
d. Month(s) of Shipment

COMMODITY IDENTIFICATION

9. Schedule B 10-Digit Code(s)
10. Commodity Description, Quantity, Size
11. Unit and Unit Price, or Total FAS/FOB Vessel Price, or FCA Price (Named Port of Loading/Airport)
(a)
(b)
(c)
(d)
(e)
12. Commodity Condition
☐ New and Unused ☐ Used - Not Rebuilt or Reconditioned ☐ Rebuilt ☐ Reconditioned ☐ Other (Specify below)
13. Source of Commodity
a. Authorized Area
b. Shipped From
c. Produced In
14. Components (Parts of the Commodity)
a. From Other than 13.a Source
b. If 14.a is "Yes", Name Country
c. Cost Per Unit of 14.b Components

Portable Document Format, or PDF,Is a file format developed by Adobe to present documents independent of application software, hardware, and operating systems.

- Can be viewed the same way on any device
- Is frequently used in legal and financial documents
- Can also be used to fill in data for forms

JavaScript Object Notation or JSON

```
{
    "Employee": [
        {
          "id":"1",
          "Manufacturer": "Audi",
          "Model": "Integra",
        },
        {
          "id":"2",
          "Manufacturer": "Buick",
          "Model": "LeSabre",
        },
        {
          "id":"3",
          "Manufacturer": "Cadillac",
          "Model": "Escalade",
        }
    ]
}
```

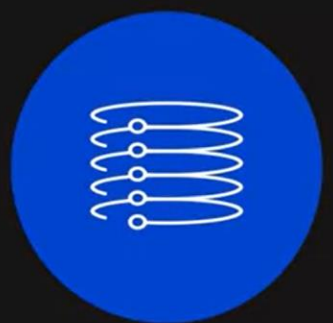
JavaScript Object Notation, or JSON, is a text-based open standard designed for transmitting structured data over the web.

- Language-independent data format
- Can be read in any programming language
- Easy to use
- Compatible with a wide range of browsers
- Considered as one of the best tools for sharing data

многие API и веб-службы возвращают данные как JSON

Common sources of data:

Data sources have never been as dynamic and diverse as they are today.



Relational Databases

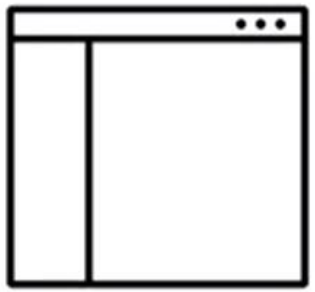
Flat files and XML Datasets

APIs and Web Services

Web Scraping

Data Streams and Feeds

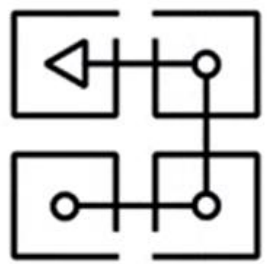
Relational Databases



Business activities

Customer transactions

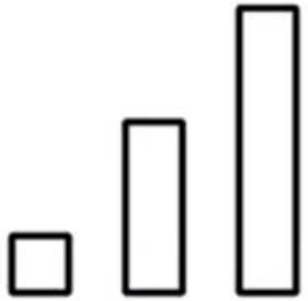
Human resource activities

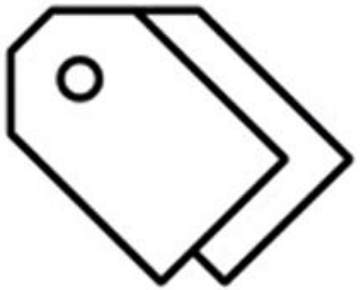


Workflows

Relational Databases

Store structured data that can be leveraged for analysis



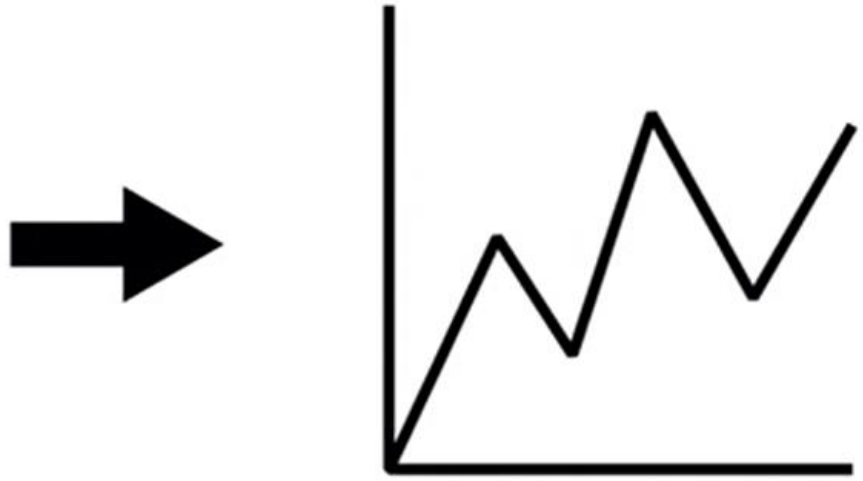


Например, данные из системы розничных транзакций могут использоваться для анализа продаж в разных регионах

Relational Databases

Store structured data that can be leveraged for analysis

Customer relationship management system



Sales projections

Flat File and XML Datasets

Public

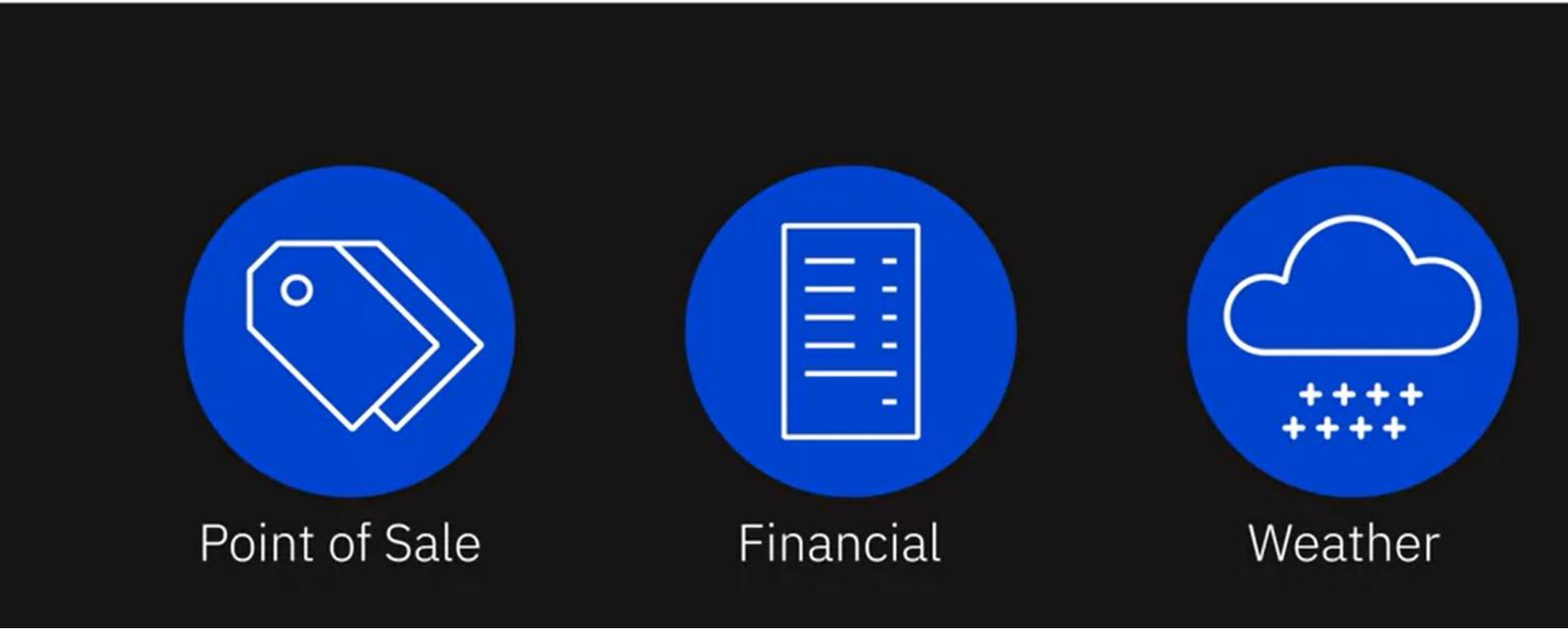
Private

Flat File and XML Datasets

Demographic

Economic

Flat File and XML Datasets



Flat File and XML Datasets

Flat File and XML Datasets

- Flat files
- Spreadsheet files
- XML documents

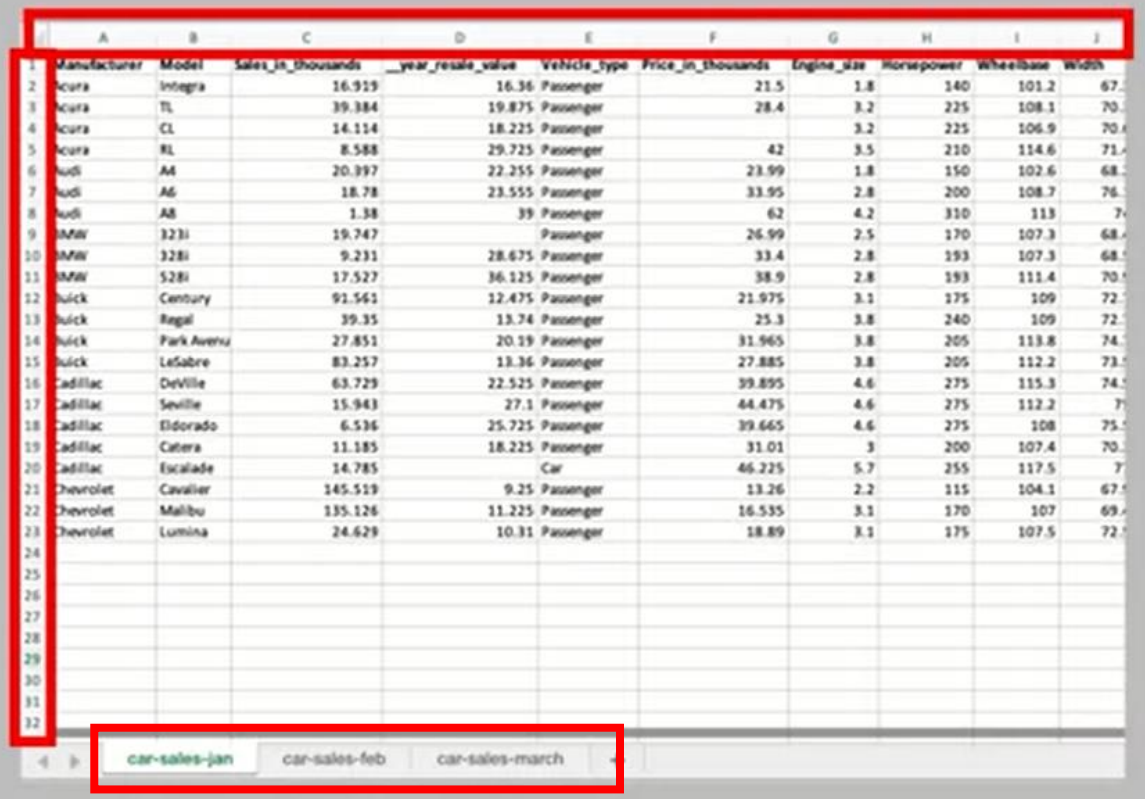
Flat files

- Store data in plain text format
- Each line, or row, is one record
- Each value is separated by a delimiter
- All of the data in a flat file maps to a single table
- Most common flat file format is .CSV

```
"Manufacturer", "Model", "Sales_in_thousands", "__year_resale_value", "Vehicle_type", "Price_in_thousands"
"Acura", "Integra", "16.919", "16.36", "Passenger", "21.5"
"Acura", "TL", "39.384", "19.875", "Passenger", "28.4"
"Acura", "CL", "14.114", "18.225", "Passenger", "14"
"Acura", "RL", "8.588", "29.725", "Passenger", "42"
"Audi", "A4", "20.397", "22.255", "Passenger", "23.99"
"Audi", "A6", "18.78", "23.555", "Passenger", "33.95"
"Audi", "A8", "1.38", "39", "Passenger", "62"
"BMW", "323i", "19.747", "Passenger", "26.99"
"BMW", "328i", "9.231", "28.675", "Passenger", "33.4"
"BMW", "528i", "17.527", "36.125", "Passenger", "38.9"
"Buick", "Century", "91.561", "12.475", "Passenger", "21.975"
```

```
"EMPNO","ENAME","JOB","MGR","HIREDATE","SAL","COMM","DEPTNO"
9999,"ADAMS","CLERK",7788,23-MAY-1987 12.00.00,1100,,20
7369,"SMITH","CLERK",7902,17-DEC-1980 12.00.00,800,,20
7499,"ALLEN","SALESMAN",7698,20-FEB-1981 12.00.00,1600,300,30
7521,"WARD","SALESMAN",7698,22-FEB-1981 12.00.00,1250,500,30
7566,"JONES","MANAGER",7839,02-APR-1981 12.00.00,2975,,20
7654,"MARTIN","SALESMAN",7698,28-SEP-1981 12.00.00,1250,1400,30
7698,"BLAKE","MANAGER",7839,01-MAY-1981 12.00.00,2850,,30
7782,"CLARK","MANAGER",7839,09-JUN-1981 12.00.00,2450,,10
7788,"SCOTT","ANALYST",7566,19-APR-1987 12.00.00,3000,,20
7839,"KING","PRESIDENT",,17-NOV-1981 12.00.00,5000,,10
7844,"TURNER","SALESMAN",7698,08-SEP-1981 12.00.00,1500,0,30
7876,"ADAMS","CLERK",7788,23-MAY-1987 12.00.00,1100,,20
7900,"JAMES","CLERK",7698,03-DEC-1981 12.00.00,950,,30
7902,"FORD","ANALYST",7566,03-DEC-1981 12.00.00,3000,,20
7934,"MILLER","CLERK",7782,23-JAN-1982 12.00.00,1300,,10
```

Manufacturer	Model	Sales_in_thousands	__year_resale_value	Vehicle_type	Price_in_thousands
Acura	Integra	16.919	16.36	Passenger	21.5
Acura	TL	39.384	19.875	Passenger	28.4
Acura	CL	14.114	18.225	Passenger	
Acura	RL	8.588	29.725	Passenger	42
Audi	A4	20.397	22.255	Passenger	23.99
Audi	A6	18.78	23.555	Passenger	33.95
Audi	A8	1.38	39	Passenger	62
BMW	323i	19.747		Passenger	26.99
BMW	328i	9.231	28.675	Passenger	33.4
BMW	528i	17.527	36.125	Passenger	38.9
Buick	Century	91.561	12.475	Passenger	21.975



Manufacturer	Model	Sales_in_thousands	__year_resale_value	Vehicle_type	Price_in_thousands	Engine_size	Horsepower	Wheelbase	Width
Acura	Integra	16.919	16.36	Passenger	21.5	1.8	140	101.2	67.
Acura	TL	39.384	19.875	Passenger	28.4	3.2	225	108.1	70.
Acura	CL	14.114	18.225	Passenger		3.2	225	106.9	70.
Acura	RL	8.588	29.725	Passenger	42	3.5	210	114.6	71.
Audi	A4	20.397	22.255	Passenger	23.99	1.8	150	102.6	68.
Audi	A6	18.78	23.555	Passenger	33.95	2.8	200	108.7	76.
Audi	A8	1.38	39	Passenger	62	4.2	310	113	7
BMW	323i	19.747		Passenger	26.99	2.5	170	107.3	68.
BMW	328i	9.231	28.675	Passenger	33.4	2.8	193	107.3	68.
BMW	528i	17.527	36.125	Passenger	38.9	2.8	193	111.4	70.
Buick	Century	91.561	12.475	Passenger	21.975	3.1	175	109	72.
Buick	Regal	39.35	13.74	Passenger	25.3	3.8	240	109	72.
Buick	Park Avenu	27.851	20.19	Passenger	31.965	3.8	205	113.8	74.
Buick	LeSabre	83.257	13.36	Passenger	27.885	3.8	205	112.2	73.
Cadillac	DeVille	63.729	22.525	Passenger	39.895	4.6	275	115.3	74.
Cadillac	Seville	15.943	27.1	Passenger	44.475	4.6	275	112.2	7
Cadillac	Eldorado	6.536	25.725	Passenger	39.665	4.6	275	108	75.
Cadillac	Catera	11.185	18.225	Passenger	31.01	3	200	107.4	70.
Cadillac	Escalade	14.785		Car	46.225	5.7	255	117.5	7
Chevrolet	Cavalier	145.519	9.25	Passenger	13.26	2.2	115	104.1	67.
Chevrolet	Malibu	135.126	11.225	Passenger	16.535	3.1	170	107	69.
Chevrolet	Lumina	24.629	10.31	Passenger	18.89	3.1	175	107.5	72.

car-sales-jan | car-sales-feb | car-sales-march

Spreadsheet files

- Special type of flat files
- Organize data in a tabular format
- Can contain multiple worksheets
- .XLS or .XLSX are common spreadsheet formats
- Other formats include Google Sheets, Apple Numbers, and LibreOffice Calc

XML files

- Contain data values that are identified or marked up using tags
- Can support complex data structures
- Common uses include online surveys, bank statements, and other unstructured data sets

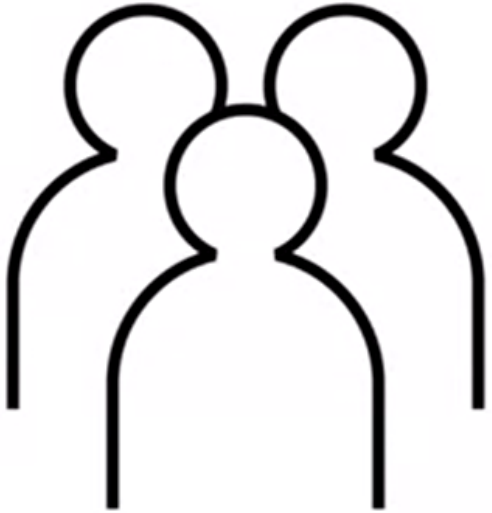
```
<?xml version="1.0"?>
<car-specs>

<manufacturer>Acura<manufacturer>

<model>Integra<model>

<sales_in-thousands>16.919<sales_in-thousands>

<year_resale_value>16.36<year_resale_value>

<vehicle_type>Passenger<vehicle_type>

<car-specs>
```

APIs and Web Services

Application Program Interfaces

Web Services

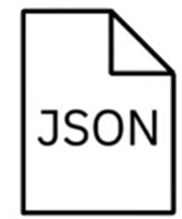
получать данные для обработки или анализа



APIs and Web Services

Web requests

Network requests



Popular examples of APIs

интеллектуальный анализ мнений или анализ настроений, который заключается в том, чтобы суммировать оценку и критика по конкретной теме, такой как политика правительства, продукт, услуга или удовлетворенность клиентов в целом

Twitter and Facebook APIs
for customer sentiment analysis

Stock Market APIs
for trading and analysis

цены на акции и сырьевые товары
, прибыль на акцию и исторические цены

Data Lookup and Validation APIs
for cleaning and co-relating data

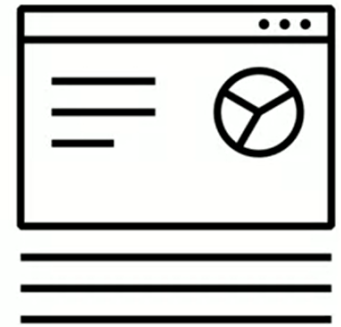
к какому городу или штату принадлежит почтовый индекс. Извлечение данных из БД.

Web scraping

- Extract relevant data from unstructured sources
- Also known as Screen scraping, Web harvesting, and Web data extraction
- Downloads specific data based on defined parameters
- Can extract text, contact information. images, videos, product items, and more...

Popular web scraping tools:

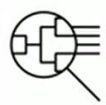
- BeautifulSoup
- Scrapy
- Pandas
- Selenium



Popular uses:

Providing price comparisons by collecting product details from retailer, manufacturers, and eCommerce websites

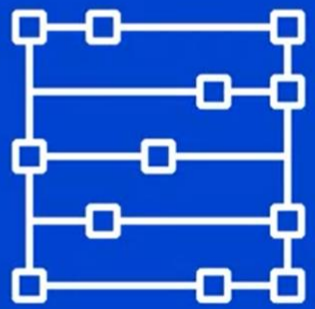
Generating sales leads through public data sources



Extracting data from posts and authors on various forums and communities

Collecting training and testing datasets for machine learning models

Data Streams and feeds



Aggregating streams of data flowing from instruments, IoT devices and applications, GPS data from cars, computer programs, websites, and social media posts

Popular technologies used to process data streams include:

kafka

APACHE Spark™

APACHE STORM™

Stock and market tickers for financial trading

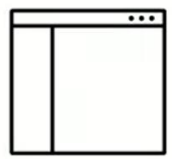
Retail transaction streams for predicting demand and supply chain management

Surveillance and video feeds for threat detection

RSS (or Really Simple Syndication) feeds

Capturing updated data from online forums and news sites where data is refreshed on an ongoing basis.

Online forums



News sites

Social media feeds for sentiment analysis

Sensor data feeds for monitoring industrial or farming machinery

Web click feeds for monitoring web performance and improving design

Real-time flight events for rebooking and rescheduling

- Данные **поступают** в самых разных форматах файлов, таких как текстовые файлы с разделителями, электронные таблицы, XML, PDF и JSON, каждый из которых имеет свой собственный список преимуществ и ограничений использования.

- Данные **извлекаются** из нескольких источников данных, от реляционных и нереляционных баз данных до API, веб-сервисов, потоков данных, социальных платформ и сенсорных устройств.

- После того, как данные идентифицированы и собраны из разных источников, их необходимо поместить в репозиторий данных, чтобы их можно было **подготовить для анализа**. Тип, формат и источники данных влияют на тип используемого репозитория данных.

Языки, имеющие отношение к работе специалистов по обработке данных.

Они могут быть классифицированы как

- языки **запросов**,
- языки **программирования**
- и сценарии **оболочки**.

Знание по крайней мере **одного языка в каждой категории** имеет важное значение для любого специалиста по обработке данных.

Introduction

Query languages

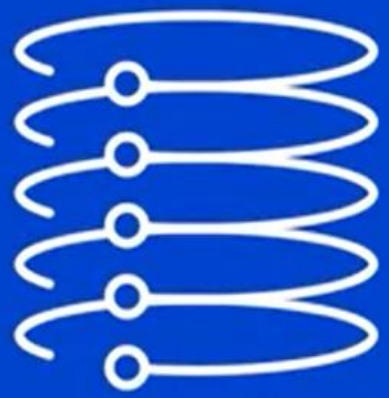
Programming languages

Shell scripting

Introduction

- Query languages are designed for accessing and manipulating data in a database (SQL)
- Programming languages are designed for developing applications and controlling application behavior (Python, R, Java)
- Shell and Scripting languages are ideal for repetitive and time-consuming operational tasks (Unix/Linux Shell, PowerShell)

SQL



SQL, or Structured Query Language, is a querying language designed for accessing and manipulating information from, mostly, though not exclusively, relational databases.

Using SQL, you can:

- Insert, update, and delete records in a database
- Create new databases, tables, and views
- Write stored procedures

Advantages of using SQL:

- SQL is portable and platform independent
- Can be used for querying data in a wide variety of databases and data repositories
- Has a simple syntax that is similar to the English language
- Its syntax allows developers to write programs with fewer lines of code using basic keywords
- Can retrieve large amounts of data quickly and efficiently
- Runs on an interpreter system

Python

Python is a widely-used open-source, general-purpose, high-level programming language.

Its syntax allows programmers to express their concepts in fewer lines of code

Has in-built functions for frequently used concepts

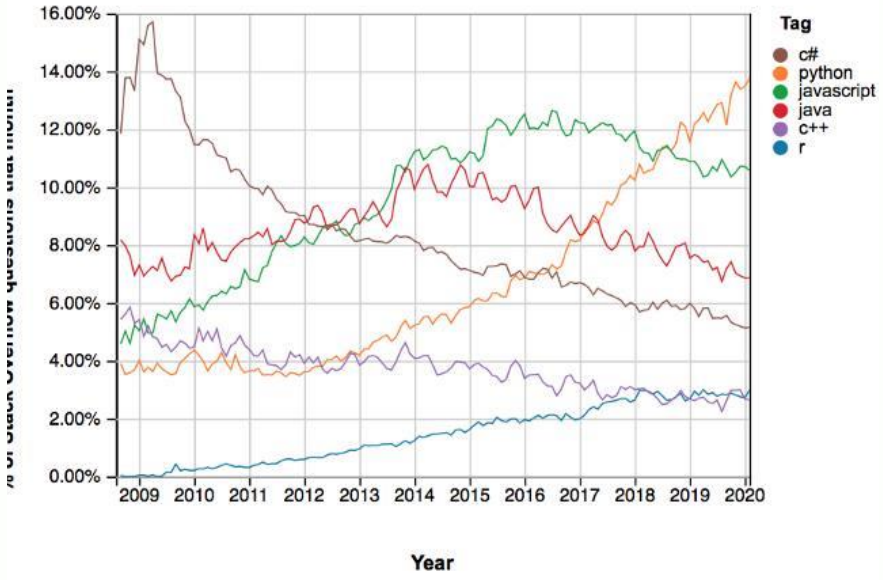
An ideal tool for beginning programmers because of its focus on simplicity and readability

Supports multiple programming paradigms – object-oriented, imperative, functional, and procedural

Great for performing high-computational tasks in large volumes of data

Python is one of the fastest-growing programming languages in the world.

- Easy to learn
- Open-source
- Can be ported to multiple platforms
- Has widespread community support
- Provides open-source libraries for data manipulation, data visualization, statistics, mathematics



Its vast array of libraries and functionalities also include:

- Pandas for data cleaning and analysis
- Numpy and Scipy, for statistical analysis
- Beautifulsoup and Scrapy for web scraping
- Matplotlib and Seaborn to visually represent data in the form of bar graphs, histogram, and pie-charts
- Opency for image processing

R-programming

R is an open-source programming language and environment for data analysis, data visualization, machine learning, and statistics.

Widely used for:

- Developing statistical software
- Performing data analytics
- Creating compelling visualizations

Key benefits:

- Open-source
- Platform-independent
- Can be paired with many programming languages
- Highly extensible
- Facilitates the handling of structured and unstructured data
- Includes libraries such as Ggplot2 and Plotly that offer aesthetic graphical plots to its users
- Allows data and scripts to be embedded in reports
- Allows creation of interactive web apps
- Can be used for developing statistical tools

Java

Java is an object-oriented, class-based, and platform-independent programming language originally developed by Sun Microsystems.

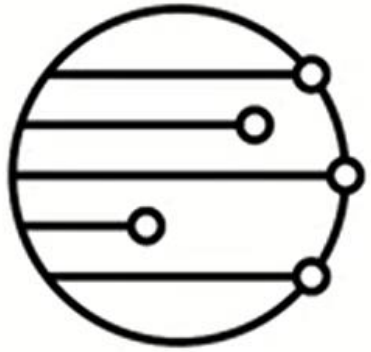
- One of the top-ranked programming languages used today
- Used in a number of data analytics processes – cleaning data, importing and exporting data, statistical analysis, data visualization
- Used in the development of big data frameworks and tools – Hadoop, Hive, Spark
- Well-suited for speed-critical projects

Unix/ Linux Shell

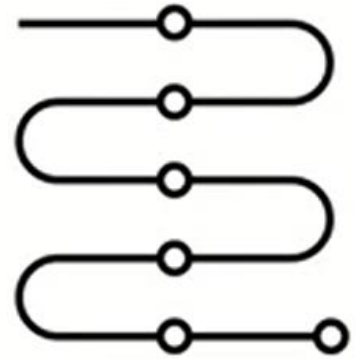
A Unix/Linux Shell is a computer program written for the UNIX shell. It is a series of UNIX commands written in a plain text file to accomplish a specific task.

Typical operations performed by shell scripts include:

- File manipulation
- Program execution
- System administration tasks such as disk backups and evaluating system logs
- Installation scripts for complex programs
- Executing routine backups
- Running batches

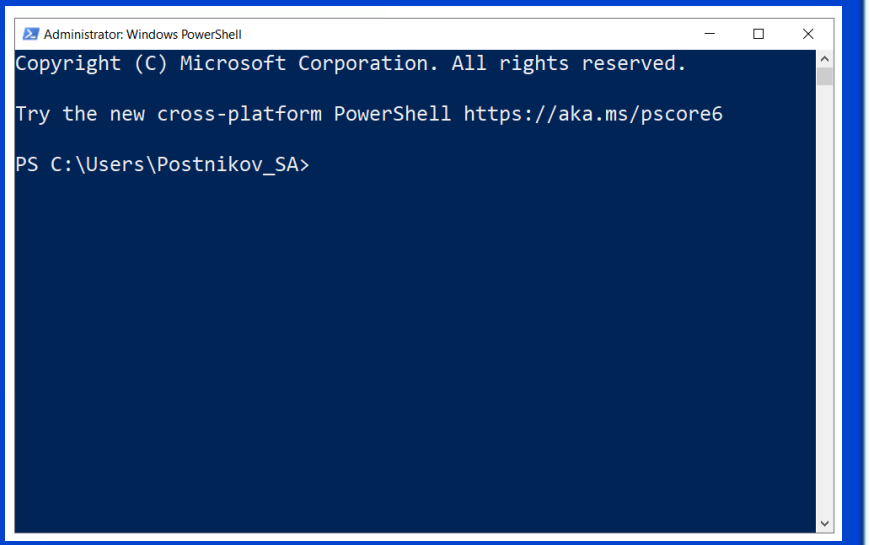


Написание сценария оболочки быстро и легко.



Это наиболее полезно для повторяющихся задач, выполнение которых может занять много времени, введя одну строку за раз.

PowerShell



PowerShell is a cross-platform automation tool and configuration framework by Microsoft that is optimized for working with structured data formats, such as JSON, CSV, XML, and REST APIs, websites, and office applications.

- Consists of command-line shell and scripting language
- Is object-based and can be used to filter, sort, measure, group, and compare objects as they pass through a data pipeline
- Used for data mining, building GUIs, creating charts, dashboards, and interactive reports

Специалисты по данным нуждаются в множестве языков

которые могут помочь им извлекать, подготавливать и анализировать данные. Их можно классифицировать как:

- Языки **запросов**, такие как SQL, используемые для доступа к данным из баз данных и управления ими.

- Языки **программирования**, такие как Python, R и Java, для разработки приложений и управления их поведением.

- Языки **оболочки** и сценариев, такие как Unix/Linux Shell и PowerShell, для автоматизации повторяющихся рабочих задач.


HOW'S THE BIG DATA PROJECT COMING ALONG, HOSKINS?
© D.Fletcher for CloudTweaks.com