Data Storage

Data can be stored in many different formats, but the three main categories are

- Structured data
- semi-structured data
- unstructured data

In short, structured data adheres to a model arrangement of fields and values when it is written to a file. Unstructured data is data that has no particular arrangement, and may need to be structured in some way when read. Semi-structured is a bit of both.

Vocabulary

- categorical data - data that has been divided into groups
- CSV - comma separated value. a row of structured data, with each data point separated by a comma
- data lakes - a central repository for structured or unstructured data.
- data mining - analyzing large sets of data in order to generate new information
- data points - a single fact that can be represented numerically and/or graphically
- delimiter - punctuation used to separate data points in a structured line of data
- EDI - electronic data interchange, used to transfer data between business systems
- IOT - Internet of Things, a network of physical objects embedded with sensors, software, and other technologies for collection and exchanging data.
- JSON - JavaScript Object Notation, a format for storing and transferring data between systems, usually web-based
- metadata - data providing information about one or more aspects of data
- parse - the act of converting a string of data into another type.
- qualitative analysis - using subjective, or inexact information to extract meaning from unstructured data
- quantitative analysis - using math, statistical analysis, or measurements to understand data
- semi-structured data - data which does not conform to a predefined model, but which has some structure identified only when reading the data
- Structured data - data that conforms to a specific, pre-defined model
- traverse - to move across or through
- unstructured data - collected information that is not arranged in any format, and doesn't conform to any model.
- XML - eXtensible Markup language, a loosely structured markup language letting users define their documents on the web

Structured data

Structured data is the most easily recognizable type of data. This is because all the records conform to a well-defined model, with specifically identified data points. A mailing address is a perfect example of structured data.

MAILING ADDRESS MODEL	MAILING ADDRESS DATA	
First Name	Jojo	Frank
Last Name	Jellybeans	Eweenie
P.O. Box or House Number and street.	P.O. Box 123456	1111Anywhere Lane
Apartment number if applicable	Apartment 3	
City, State, Zip	Candyland OH 41111	Coney Island NW 14444

All mailing addresses conform to the identified model. The data is already processed, and conforms to the rules of the model.

Structured data can be used in relational databases and is easily accessible and traversable by both users and machine algorithms. The ability to analyze structured data has been around a while, so there are many tools available to assist with data analysis and visualization. The rigidity of its model makes the data easy to understand, search, sort, and transfer to other systems.

Quantitative analysis is most often performed with structured data, as the data is clean, already processed, and no longer considered “raw” data. Change the structure, however, and all records must also be updated. Since the rules are pre-set in the structure, there is limited scope for creative analysis.

Structured data lends itself well to text data files, Comma Separated Values (CSV) or other delimited data files, and SQL processing.

Comma separated values (CSV)	Tab separated	space separated values
"1997","Ford","E350"	"1997" "Ford" "E350"	"1997" "Ford" "E350"

Structured data is estimated to account for approximately 20% of all data used world-wide, and is the foundation of big-data (*What is structured data?* TIBCO Software. (n.d.). Retrieved March 14, 2022, from https://www.tibco.com/reference-center/what-is-structured-data)

Semi-structured data

Semi-structured data is data that is somewhat, but not rigidly, structured. It is not pure, raw, unstructured data either. It cannot be organized in relational databases, and doesn't have a strict model to follow. It does, however, have a loosely organized framework, yet can also be considered open ended. Email is a great example of semi-structured data.

Any email could have a sender, receiver, subject, data. The user (or inbox rules) can choose to categorize it within a folder, or not, based upon any of these characteristics. However the text of the email can also allow rules to guide it into a specific folder, without really needing to know every word of the email content itself.

All markup languages are considered semi-structured. This includes XML, EDI and even JSON format. They all have metatags to identify the loose structure of the data, but not every record uses every meta tag. JSON data can be collected from any mobile, computing, or IOT device connected from the web. JSON, XML, and even EDI are self-describing, and text-based.

JSON Example	XML	EDI

"name":"Jojo", "age":30,
"car":null
	<note>
	  <to>TFrank</to>
	  <from>Jojo</from>
	  <heading>Lunchr</heading>
	  <body>Let's get pizza!</body>
	</note>
		ISA*00* *00*
		*ZZ*SENDERISA
		*14*0073268795005
		*020226*1534*U*00401*000000001*0*
		T*>~
		GS*PO*SENDERGS*007326879*20020226
		*1534*1*X*004010~
		ST*850*000000001~
		BEG*00*SA*A99999-01**19970214~
		REF*VR*54321~
		ITD*01*3*1**15**16~
		DTM*002*19971219~
		DTM*002*19971219~
		N1*BT*BUYSNACKS
		INC.*9*1223334444~
		N3*P.O. BOX 0000~
		N4*TEMPLE*TX*76503~
		N1*ST*BUYSNACKS
		PORT*9*1223334445~
		N3*1000 N. SAMPLE HIGHWAY~
		N4*ATHENS*GA*30603~
		PO1**16*CA*12.34**CB*000111111*UA
		*002840022222~
		PID*F****CRUNCHY CHIPS LSS~
		PO4*48*7.89*LB~

Semi-structured data is often associated with metadata. Metadata generally describes what the user is seeing in the data. For example, JSON uses name:value pairs, and XML uses descriptive markup tags such as <note> and <to>.

Semi-structured allows businesses to collate data from multiple sources, as well as transfer between systems as needed. Data mining tools help to index, model, and search unstructured data to extract information.

Unstructured Data

Unstructured data is data in the most raw format. Its objective, and unorganized. It has no formal structure, so cannot be placed in a relational database, and searching on it using traditional techniques is ineffective.

How could you place the contents of an audio, video, or image file to a structure? Social media posts, shopping habits, browsing data are all forms of unstructured data. It is only when the data is read, that a schema or model can loosely be developed. There is no native format, making it adaptable, and usable by larger pools of data. It is not restricted to specific data, and can be collected very quickly from its native format as it doesn't need to be pre-defined to a format. Pools of unstructured data are often referred to as "data lakes".

Extracting and analyzing information from unstructured data is what data scientists do. They can use the data to look for trends, patterns, and connections to other, unrelated data schemas. Unstructured data is not quantitative, but more categorical, or qualitative.

Activity

Resources

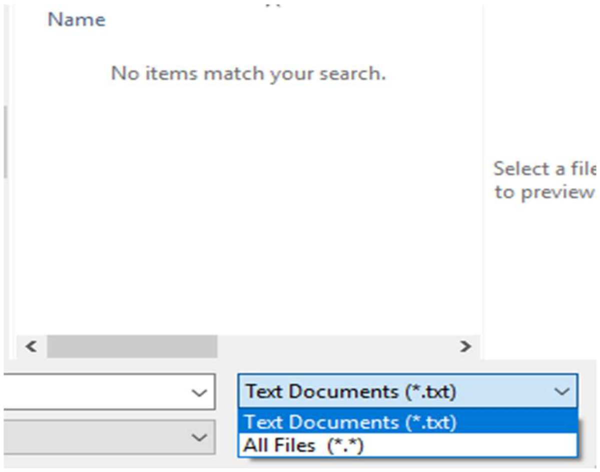
https://www.kaggle.com/

Warm-up

Download the 2015 flight delays and cancellations data from 2015 at https://www.kaggle.com/usdot/flight-delays . Decompress (unzip) the file.

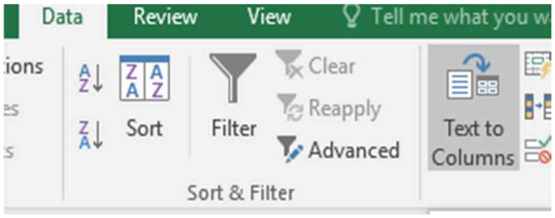
Steps

1. Start notepad.exe.
2. Click on File -> Open and navigate to the extracted airlines.csv data file.
3. Change the file type to "All File (*.*)
4. Select airlines.csv. Notice the format and the locations of the commas?
5. Close the file.
6. Re-open the file in MS Excel. Notice how Excel automatically converts the CSV data points to separate columns?
7. Close the file.
8. Re-open the file in Notepad, all files. Change



the commas to another delimiter (a space, tab, or nay punctuation symbol).

9. Save and close.
10. Reopen the file in MS Excel. What happened? CSV is the most commonly accepted structured format for transferring text data between systems.
11. To separate the data into columns, click the column, then using the menus click Data -> Text-to-columns



12. Click "NEXT" and select your delimiter in the supplied list, or click OTHER and enter the punctuation used.
13. click "FINISH".
14. MS Excel can parse your data to columns, if it knows what the delimiter is.
15. Practice again on a dataset of your choosing, until you feel comfortable parsing data into columns using MS Excel.

Extensions of the Activity

1. **SQL:** Have students review the structured data at https://www.w3schools.com/whatis/whatis_sql.asp and create a model describing the structure, and types of data. Have students include potential field sizes and any constraints to help prepare them for data cleaning needs.
2. **Kaggle or Jupyter Notebook:** Have students read the structure about the airlines.csv file using python. First, students will create a new notebook. Then, students can enter the following commands. Instructors could also include a worksheet pre-activity where students predict what each line is doing (if there is no experience with Python), and identify what various commands are for. The code below assumes the airlines.csv file is extracted to the "archive" directory, inside the user's Downloads folder. Brainstorm how the information provided is useful. Have students create a model of the structure.

```
import pandas
df = pandas.read_csv("Downloads\archive\airlines.csv")
print df.head() # prints the first 6 rows of the table.
print df.head(40) # prints the first 40 rows of the table.
print df.info(verbose = True) # gives a summary of the dataframe
```

3. **Semi-structured data:** Download and explore a semi-structured XML data set at https://data.cityofnewyork.us/Recreation/Museums-and-galleries/kcrm-j9hh and have students

identify key differences between the airlines.csv and this data. See if students can find both the structured and unstructured elements. Have students suggest ways to structure the unstructured pieces of data.

4. **Unstructured data:** Browse example set of unstructured data at https://docs.atlas.mongodb.com/sample-data/sample-airbnb/#std-label-sample-airbnb. While students may think the name:value pairs are a form of structured data, be sure to point out what makes it unstructured.

This material is based upon work supported by the National Science Foundation under Grant DUE 2055411.Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.