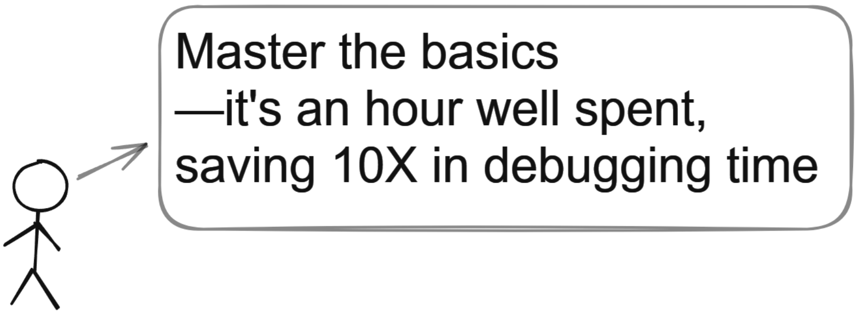
Advanced Python - Project1

Hello and welcome to your first project, In our first week, we tried to cover the basics you need to start with Python programming. ***Sometimes when we start learning anything and we are excited about it, we skip the basics just to see quick results***. It might look fun at the beginning because we are achieving quick results but you will have no clue how to debug any issue you might face when things don’t go with our happy flow (*Which is a frequent thing to happen with coding*)



> `Now, start thinking and analyzing. If any concept is unclear, go back and write some code.`

Last week, we covered the following topics:

- Python Data Types
- Conditional Branching and Loops
- Python Functions

☑ ~~Go back and write some code if any concept is unclear~~

As a Python developer working in data manipulation, your role is crucial in *extracting, transforming, and loading (ETL)* data. You will be responsible for processing and manipulating data from various sources to meet the company's needs. To successfully perform your tasks, you will need to track and manage different data sources. This involves understanding the structure and format of each data source, as well as identifying any potential inconsistencies or errors.

One important aspect of tracking data sources is ensuring *data integrity*. You will need to verify the accuracy and completeness of the data, as well as handle any missing or corrupted data. This may involve implementing data validation techniques and performing quality checks.

Furthermore, you will need to establish efficient workflows and processes to handle the data. This includes designing and implementing data pipelines, which involve extracting data from different sources, transforming it into the desired format, and loading it into the target systems or databases.

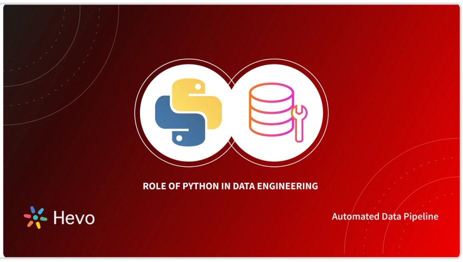
In addition, as a data manipulator, you may also need to perform data cleansing and data wrangling tasks. This involves cleaning and organizing the data, removing duplicates, standardizing formats, and handling any inconsistencies. Overall, as a Python developer working with data, your role as a data manipulator and ETL specialist requires a strong understanding of data structures, programming concepts, and the ability to track and manage various data sources effectively.

You can read more here or any other search you do

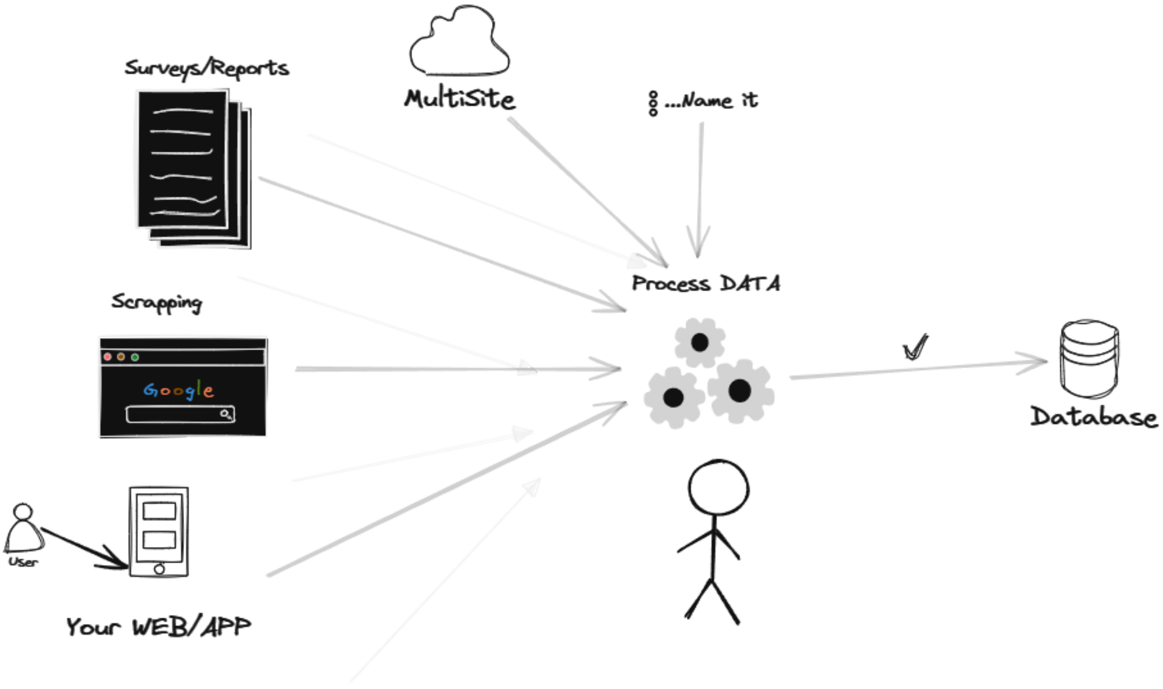
Role of Python for Data Engineering: 4 Critical Aspects | Hevo

Unlock the power of Python for data engineering. Leverage Python libraries and tools for effective data engineering and analysis.

https://hevodata.com/learn/python-for-data-engineering/



💡 Now, theoretically, we can say that there is an infinite number of data sources you might have, meaning a lot of different sources, and it all depends on the business needs. Can you think of a scenario where you are providing a platform that reads sensor data? What would be the data source and data size?



Let’s Start !

You are hired as a python developer, and you are excited to your first task meeting,

Hello! We are starting with a new client who has his own data source
Manager
Sure, can I know what type of source we are talking about?
You
Well, supposed to be csv format
OK, I need to read about CSV and also what is the final shape we want!
Sure, will write a dedicated task for you and keep me updated!

Quick Tip: When working on any task, don't hesitate to seek clarification if something is unclear. Your understanding matters, and asking questions early on is a sign of diligence. Remember, there's no shame in seeking clarity, and it's far better than proceeding with uncertainties

Now, till your manager come back to you, you might need to quick search from your side such as:

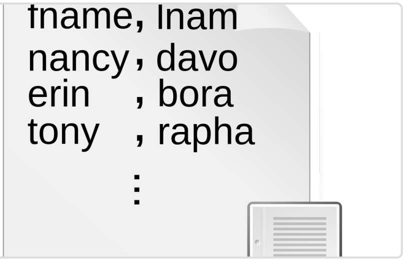
- ☐ Read about csv file/ other type such as tsv
- ☐ Check how to read csv file in python (Regardless the processing, sure you must read it)
- ☐ Check databases format that you have in your company
- ☐ What is data cleaning?

Resources:

Comma-separated values

Comma-separated values (CSV) is a text file format that uses commas to separate values. A CSV file stores tabular data in plain text, where each line of the file typically represents one data record. Each record

W https://en.wikipedia.org/wiki/Comma-separated_values



How to read csv files in Python (without Pandas)

Reading CSV files are a basic and important first step in getting data. In this article I will go over the basic Python read functions…

https://medium.com/@AIWatson/how-to-read-csv-files-in-python-without-pandas-b693fc7ea3b7

Cleaning data in a CSV file using Python:

Data scientists spend a large amount of their time cleaning datasets and getting them down to a form with which they can work. In fact, a lot of data scientists argue that the initial steps of obtaining and cleaning data

https://www.linkedin.com/pulse/cleaning-data-csv-file-using-python-sheetal-dhande-dandge-phd-/

After one day, your manager came back and shared this email for you:

“Hello

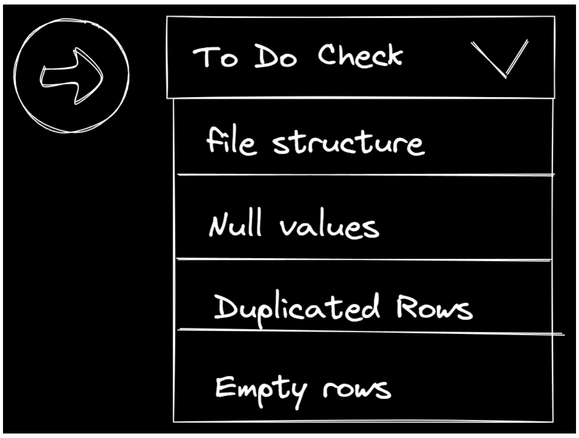
As discussed yesterday,

We will process with an example file till we have the full requirements from the client.

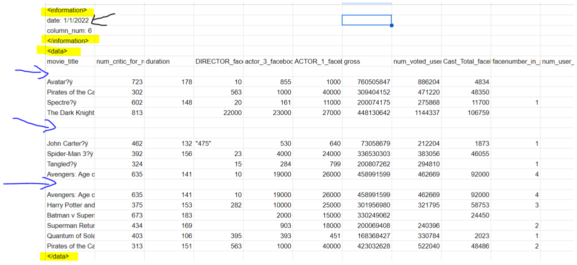
However, regardless of the information that will be hold in the file

Take into the consideration the following points

Plus, find attached the example file, thanks”



General points to consider



<information>										
date: 1/1/2022										
column_num: 6										
</information>										
<data>										
movie_title	num_critic_for_r	duration	DIRECTOR_fac	actor_3_faceboo	ACTOR_1_facel	gross	num_voted_use	Cast_Total_facel	facenumber_in_	num_user_
Avatar?ÿ	723	178	10	855	1000	760505847	886204	4834		
Pirates of the Ca	302		563	1000	40000	309404152	471220	48350		
Spectre?ÿ	602	148	20	161	11000	200074175	275868	11700	1	
The Dark Knight	813		22000	23000	27000	448130642	1144337	106759		
John Carter?ÿ	462	132	"475"	530	640	73058679	212204	1873	1	
Spider-Man 3?ÿ	392	156	23	4000	24000	336530303	383056	46055		
Tangled?ÿ	324		15	284	799	200807262	294810		1	
Avengers: Age o	635	141	10	19000	26000	458991599	462669	92000	4	
Avengers: Age o	635	141	10	19000	26000	458991599	462669	92000	4	
Harry Potter and	375	153	282	10000	25000	301956980	321795	58753	3	
Batman v Super	673	183		2000	15000	330249062		24450		
Superman Retur	434	169		903	18000	200069408	240396		2	
Quantum of Sola	403	106	395	393	451	168368427	330784	2023	1	
Pirates of the Ca	313	151	563	1000	40000	423032628	522040	48486	2	
</data>										

- Yellow: Headers , Blue: Empty rows, Black: Headers

As a first time, you will be chocked that the task might seems very big and you don't know from where to start. The best thing to do after discussing the task with the manager is to divide the task into subtasks:

- **Task Overwhelm:**
  - Feeling overwhelmed? It's common for big tasks.
  - Don't know where to start? Break it down.
- **Subtask Strategy:**
  - Discuss the task with your manager.
  - Break the task into smaller, manageable subtasks.
- **Demonstrate Understanding:**
  - Breaking down tasks shows your capability to understand.
- **Measurable Progress:**
  - Easier to measure progress in stages.
  - If a week passes, showcase what you've achieved in parts.
- **Debugging and Adaptation:**
  - Smaller tasks are easier to debug and adapt.
- **Effective Communication:**
  - Clarity in your approach facilitates effective communication with the team.

💡 **Take a moment to break down your tasks into manageable steps.**

GREAT!

Now, At east we know what shall we do:

□ Read and try how to read csv file in python (without pandas)

□ After reading, think about structure, do we need to check it ? every time or what do you think ?

□ Check Empty and Duplicated rows

□ Based on the number of columns in the header, you are supposed to generate new file with the needed columns. For example if the header has (num_cols:5) → means your output file will have 5 processed and clean columns

Programming Tips😕

- Debug and Debug, don’t write the steps in one shot
- Read and Read, Don’t Copy Code without understanding
- As long as you don’t start, it will be always hard to understand the full steps

Submission:

- Python Notebook, well documented
- Tested Code, you can test other file if you want
- Readable Code