MODULE 03

DATA SCIENCE BOOTCAMP

Numpy and Pandas

Vertical Institute

Reminder

1. Please turn on Zoom camera the whole duration of classes.

2. At the start of all classes, please rename yourselves to: Name + Last 3 digits and letter of your NRIC. Example: John Tan (123A)

Agenda

- Python Modules
- Numpy
- Pandas
- Types of Joins

© 2024 Vertical Institute

Attendance Photo Taking

MODULE 3: NUMPY AND PANDAS

NumPy

© 2024 Vertical Institute

Python Modules

A python module allows you to logically organize your Python Code.

Module contents available to the caller with the import statement.

Modules include:

- Numpy
- SciPy
- Matplotlib
- Pandas

```python
import <module_name>
```

© 2024 Vertical Institute

NumPy

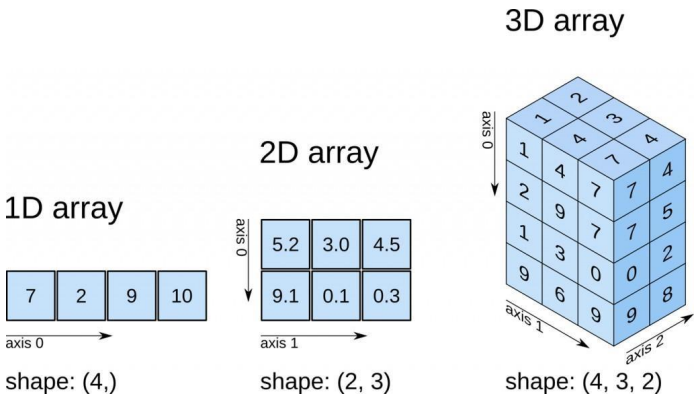
NumPy or Numerical Python is a general-purpose array processing python package for scientific computing. It consists of numerous powerful features inclusive of:

- A robust multi-dimension array object with many useful functions
- An enormous number of routines, including shape manipulation, logical, mathematical &
- many more to operate on NumPy Array objects
- NumPy is also utilized as a generic multi-dimension data container
- A wide set of databases can also be integrated with NumPy

© 2024 Vertical Institute

Numpy Array

- A numpy array is a grid of values, all of the same type, and is indexed by a tuple of nonnegative integers.
- Basically, it is a multidimensional or n-dimensional array of fixed size with homogeneous elements(i.e., the data type of all the elements in the array is the same)



© 2024 Vertical Institute

Difference Between List And Numpy Array

- A user can treat lists as arrays.
- However, user cannot constraint the type of elements stored in a list.
- If you create arrays using the array module, all elements of the array must be of the same type

© 2024 Vertical Institute

Benefits of Numpy Array

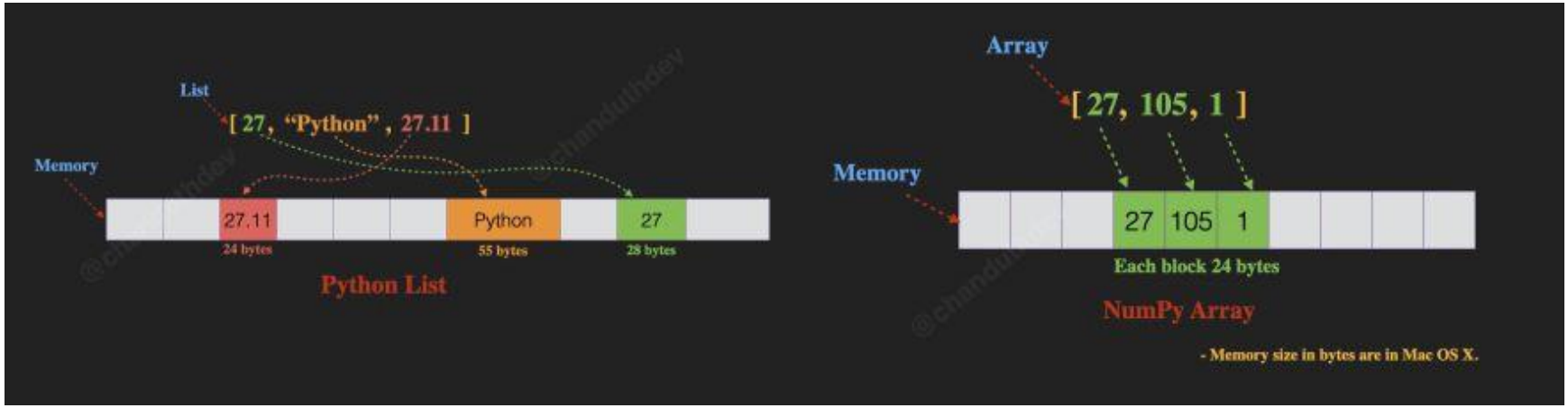
Numpy data structures perform better in:

- **Size** - Numpy data structures take up less space
- **Performance** - they have a need for speed and are faster than lists
- **Functionality** - SciPy and NumPy have optimized functions such as linear algebra operations built in.

© 2024 Vertical Institute

Difference Between Numpy Array & Python Lists

Numpy array tend to be more efficient in memory and storage.



© 2024 Vertical Institute

MODULE 3: NUMPY AND PANDAS

Pandas

© 2024 Vertical Institute

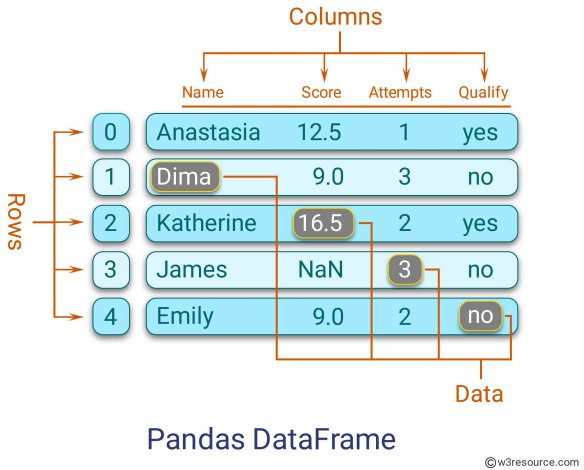
What is Pandas?

- Pandas is an open source Python package that is most widely used for data science/ data analysis and machine learning tasks.
- It offers powerful, expressive and flexible data structures that make data manipulation and analysis easy, among many other things. The DataFrame is one of these structures.

© 2024 Vertical Institute

DataFrame

DataFrame is **a 2-dimensional labeled data structure with columns of potentially different types.**



© 2024 Vertical Institute

What You Can Do With DataFrames Using Pandas

Pandas makes it simple to do many of the time consuming, repetitive tasks associated with working with data, including:

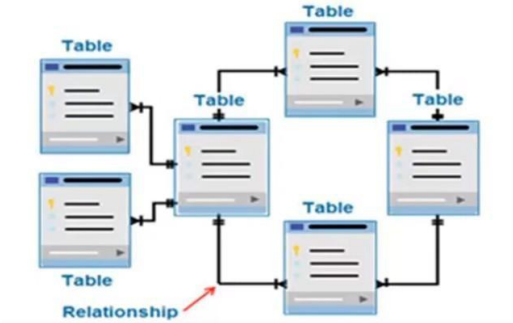
- Data cleansing
- Data fill
- Data normalization
- Merges and joins
- Data visualization
- Statistical analysis
- Data inspection
- Loading and saving data
- And much more

© 2024 Vertical Institute

Real World Use Case With Pandas

In general, the data in financial institutions are stored in relational data format. Most of the raw data are not in their best quality and are stored in different databases.

Hence, we utilized the pandas functions to efficiently clean and join data among the different data tables so as to perform data analytics of quality. Python programs are often utilized pandas to automate the data cleaning process for their data analytics use.

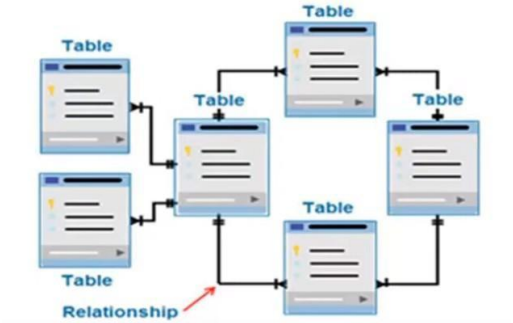


© 2024 Vertical Institute

Real World Use Case With Pandas

In general, the data in financial institutions are stored in relational data format. Most of the raw data are not in their best quality and are stored in different databases.

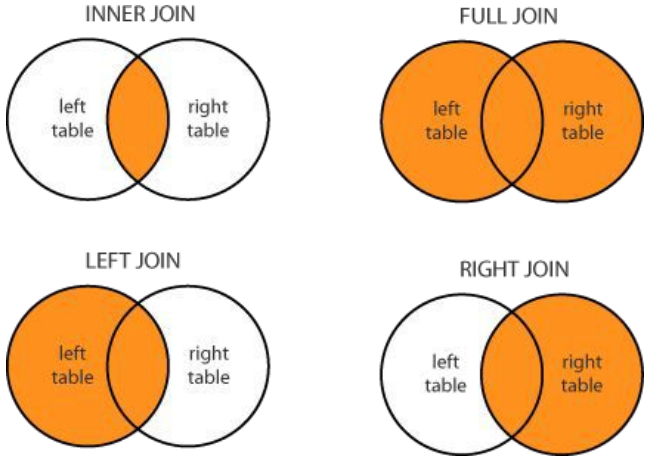
Hence, we utilized the pandas functions to efficiently clean and join data among the different data tables so as to perform data analytics of quality. Python programs are often utilized pandas to automate the data cleaning process for their data analytics use.



© 2024 Vertical Institute

Type of Joins

A Join clause is used to combine rows from two or more tables, based on a related column between them.



© 2024 Vertical Institute

Type of Joins

Join Clause	Description
(INNER) JOIN	Returns records that have matching values in both tables
LEFT (OUTER) JOIN	Returns all records from the left table, and the matched records from the right table
RIGHT (OUTER) JOIN	Returns all records from the right table, and the matched records from the left table
FULL (OUTER) JOIN	Returns all records when there is a match in either left or right table

© 2024 Vertical Institute

Pd.merge()

When you want to combine data objects based on one or more keys in a similar way to a relational database, merge() is the tool you need.

left

	key1	key2	A	B
0	K0	K0	A0	B0
1	K0	K1	A1	B1
2	K1	K0	A2	B2
3	K2	K1	A3	B3

right

	key1	key2	C	D
0	K0	K0	C0	D0
1	K1	K0	C1	D1
2	K1	K0	C2	D2
3	K2	K0	C3	D3

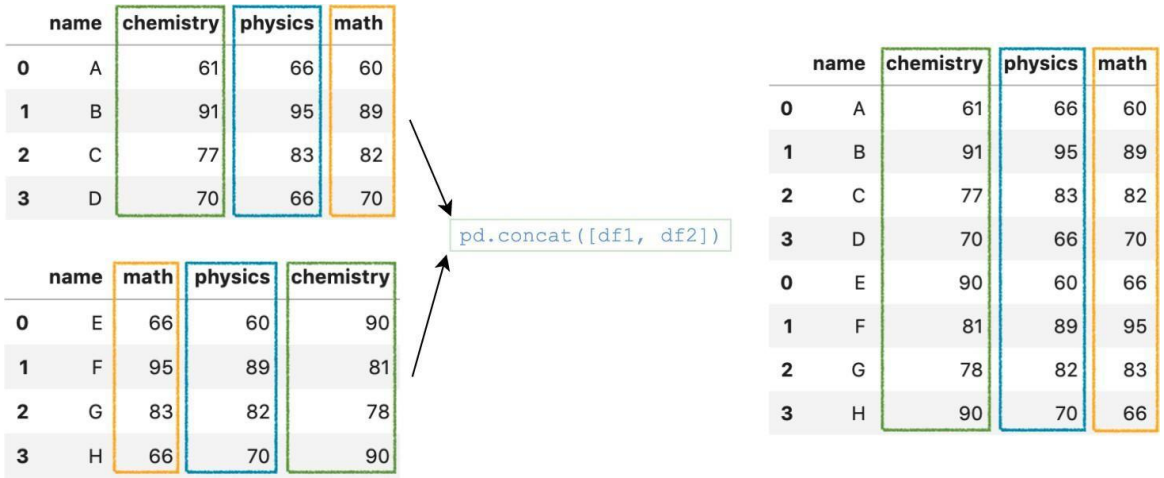
Result

	key1	key2	A	B	C	D
0	K0	K0	A0	B0	C0	D0
1	K0	K1	A1	B1	NaN	NaN
2	K1	K0	A2	B2	C1	D1
3	K1	K0	A2	B2	C2	D2
4	K2	K1	A3	B3	NaN	NaN
5	K2	K0	NaN	NaN	C3	D3

© 2024 Vertical Institute

Pd.concat()

The **concat** function does all of the heavy lifting of performing concatenation operations along an axis. It is similar to appending the data rows of different dataset.



	name	chemistry	physics	math
0	A	61	66	60
1	B	91	95	89
2	C	77	83	82
3	D	70	66	70

	name	math	physics	chemistry
0	E	66	60	90
1	F	95	89	81
2	G	83	82	78
3	H	66	70	90

	name	chemistry	physics	math
0	A	61	66	60
1	B	91	95	89
2	C	77	83	82
3	D	70	66	70
0	E	90	60	66
1	F	81	89	95
2	G	78	82	83
3	H	90	70	66

© 2024 Vertical Institute

Pandas Operations

In the jupyter notebook, we will cover a lot of the pandas operations that are essential to the data analytics process such as:

- Slice and dice of datasets by different selection methods
- Assignment of new values to dataframe
- Filtering with multiple conditions
- Mathematical operations
- Join and appending datasets

© 2024 Vertical Institute

Recap time!

What are your favorite takeaways on NumPy & Pandas today?

Let's share with each other!

Some things to take note...

Link and resource could be accessed in the Learning Portal.

https://elearn.verticalinstitute.com/users/sign_in

© 2024 Vertical Institute

Attendance Photo Taking

MODULE 03

DATA SCIENCE BOOTCAMP

Thank you!

Vertical Institute