Faculty of Engineering and Technology			
Ramaiah University of Applied Sciences			
Department	Computer Science and Engineering	Programme	B.Tech
Semester/Batch	7th/2021		
Course Code	20CSE421A	Course Title	Data Science Foundation
Course Leader	Dr. Yatish Bathla/Dr. Mohan Kumar K N		

Assignment - 01			
Register No		Name of Student	

Sections		Marking Scheme	Marks		
			Max Marks	First Examiner Marks	Moderator Marks
Part-A					
	A 1.1	Python language and its usage in data science	04		
	A 1.2	Python syntax and core constructs	06		
	A 1.3	Functions: Namespaces and scopes	05		
	A 1.4	Exception handling	04		
	A1.5	Libraries for Data Science Applications	05		
	A1.6	Data modelling, processing	06		
		Part-A Max Marks	**30**		
Part B 1					
	B 1.1	Design a system for reading data in text format using pandas library	7		
	B 1.2	Design a system for storing data in text format using pandas library	8		
	B 1.3	Design about the task in a parallel system using Celery to obtain square root of a value	5		
	B 1.4	Design about the client and server role in a parallel system using Celery to obtain square root of a value	10		
		Part-B1 Max Marks	**30**		
Part B 2					
	B2.1	Implement Python code for reading data in text format using pandas library	10		
	B2.2	Implement Python code for storing data in text format using pandas library	10		

	B2.3	Implement Python code for the task in a parallel system using Celery to obtain square root of a value	5		
	B2.4	Implement Python code for the client and server role in a parallel system using Celery to obtain square root of a value	15		
		Part-B2 Max Marks	**40**		
		Total Assignment Marks	**100**		

Course Marks Tabulation				
Component- CET B Assignment	**First Examiner**	**Remarks**	**Second Examiner**	**Remarks**
A				
B.1				
B.2				
Marks (Max 50)				
Marks (out of 25)				

Signature of First Examiner Signature of Second Examiner

Please note:

1. Documental evidence for all the components/parts of the assessment such as the reports, photographs, laboratory exam / tool tests are required to be attached to the assignment report in a proper order.
2. The First Examiner is required to mark the comments in RED ink and the Second Examiner's comments should be in GREEN ink.
3. The marks for all the questions of the assignment have to be written only in the **Component – CET B: Assignment** table.
4. If the variation between the marks awarded by the first examiner and the second examiner lies within +/- 3 marks, then the marks allotted by the first examiner is considered to be final. If the variation is more than +/- 3 marks then both the examiners should resolve the issue in consultation with the Chairman BoE.

RAMAIAH UNIVERSITY OF APPLIED SCIENCES

Assignment 1

Term - 1

Instructions to students:

1. The assignment consists of **10** questions: Part A – **6** Question, Part B- **4** Questions.
2. A maximum mark is **100**.
3. The assignment has to be neatly word processed as per the prescribed format.
4. The maximum number of pages should be restricted to **25**.
5. Restrict your report for Part-A to 5 pages only.
6. Restrict your report for Part-B to a maximum of 20 pages.
7. The printed assignment must be submitted to the course leader.
8. **Submission Date: 27/09/2021**
9. **Submission after the due date is not permitted.**
10. **IMPORTANT**: It is essential that all the sources used in preparation of the assignment must be suitably referenced in the text.
11. Marks will be awarded only to the sections and subsections clearly indicated as per the problem statement/exercise/question

Preamble:

The course is intended to teach the design, development, analysis, and evaluation of Python language in Data Science Applications. It starts with the Core concepts of python where objects, functions, syntax, core constructs, standard libraries, external libraries, and basic deployment strategies are explained. Then, Python language for data science is emphasized where building data science tasks and workflows, Core and advanced collections, Operations, Comprehensions, Namespaces and scopes, Generators, Exception handling, Filesystem and OS interface, Data modelling, processing (pandas), data analysis (NumPy), debugging, and testing code (IPython), Interactive visualisations (matplotlib), SciPy, scikit-learn and statsmodels are elaborated. Then, Parallel programming in python is discussed where CPU bound threads, GIL bottleneck and workarounds, Thread pooling, Inter Process Communication (IPC), Distributed computation, Asynchronous operations are explained. Then, testing in python programming is elaborated where Python Unit Testing module, Acceptance Testing of python software, Test Driven Development (TDD) and Behaviour Driven Development (BDD), Debugging and Python debugger are explained. Finally, Advanced Python Programming is analysed where Functional Programming, zip and map, OO Programming, Class Factories and run time attributes, Abstract Base Classes (ABCs) and protocol declaration are examined. Students are trained to modelling, processing and visualization of data for Data Science Applications.

RAMAIAH UNIVERSITY OF APPLIED SCIENCES

PART – A **30 Marks**

A.1 What is Python programming language? Why it is used for data science?

A.2 Identifiers:
A2.1 Why are variable type declarations not used in Python?
A2.2 Why are variable name declarations not used in Python?
A2.3 Why should we avoid the use of the underscore to begin variable names with?

A.3 How are namespaces and variable scopes different from each other?

A.4 Keywords:
A4.1 Name the keyword(s) which is(are) used to raise exceptions.
A4.2 What is the difference between try-except and try-finally?

A.5 What are the essential libraries in Python for data analysis? Please elaborate (at least four of them)

A.6 Data modelling, processing, and visualization:
A6.1 Which library is extremely fast for binary data loading and storage, including support for memory-mapped array?
A6.2 Which two libraries combines to provide interactive features like zooming and panning?
A6.3 Which library is used for flexible and high-performance group by facility, enabling slice and dice, and summarize data sets in a natural way?

PART – B **70 Marks**
B.1 **30 Marks**

Pandas contains high-level data structures and manipulation tools designed to make data analysis fast and easy in Python. Python has become a beloved language for text and file munging due to its simple syntax for interacting with files, intuitive data structures, and convenient features like tuple packing and unpacking. Perform the following:

B1.1 Discuss the Python object, Parsing functions, and most important feature of these functions in pandas for reading data in text format

B1.2 Discuss the in-memory SQLite database used for storing data in python. How pandas improve the issues of not repeating the query to database

Celery is a framework that offers mechanisms to lessen difficulties while creating distributed systems. It works with the concept of distribution of work units (tasks) by exchanging messages among the machines that are interconnected as a network, or local workers. A task is the key concept in Celery; any sort of job we must distribute has to be encapsulated in a task beforehand. It has an architecture based on pluggable components and a mechanism of message exchange that uses a protocol according to a selected message

transport (broker). The client components have the function of creating and dispatching tasks to the brokers.

B1.3 Define a task by using the decorator, which is accessible through an instance of Celery application? What methods are used to dispatch a task? What are the functionalities of broker, please elaborate?

B1.4 Discuss the client role, where app Celery can dispatch the tasks to be executed? Firstly, please explain about the virtual environment. Then, explain about the necessary packages (Basic, Client transmits messages through the broker) of the client. After that, Discuss the server role, which could be broker and result backend. Finally, discuss about the task for square root of a value.

B.2 **40 Marks**

This section is the coding implementation of the Part B.1. Each subsection corresponds to the subsection of previous part (for example B2.1 to B1.1, B2.2 to B1.2 etc.):

B2.1 Implement the code to read text file by using Python functions (with and without comma-delimited). How to assign column names (default and specify name) if header is not available.

B2.2 Implement the code for in-memory SQLite database used for Loading data from SQL into a python object and then insert a few rows of data. Also, implement code using Pandas improve the issues of not repeating the query to database

B2.3 Demonstrate a code to define a task by using the decorator, which is accessible through an instance of Celery application? Then, implement code for a task that will calculate the square root of a value and return a result.

B2.4 Implement the code for the client role, where app Celery can dispatch the tasks to be executed? Firstly, please setup the virtual environment. Then, install the necessary packages (Basic, Client transmits messages through the broker) of the client. After that, please install the server role, which could be broker and result backend. Finally, demonstrate a code about the task for square root of a value.