ASSIGNMENT 2

Joel Jacob John

1) What is the difference between a list and a tuple in Python?

List:

(i)Mutable: Lists elements can be modified.

(ii)Defined using square brackets [ ].

Tuple:

(i)Immutable: Tuple elements cannot be changed.

(ii)Defined using parentheses ( ).

2.How can you iterate through a list in Python?

You can iterate through a list using various methods, including for loops and list comprehensions.

3) How do you handle exceptions in Python?

Exceptions are handled using try and except.

```
try:
    result = 10 / 0
except ZeroDivisionError:
    print("Division by zero is not allowed.")
else:
    print("No exceptions occurred.")
```

4)What are list comprehensions in Python?

List comprehension provides a concise way to create lists based on existing lists.

```
l = [2, 4, 6, 8, 10]
new_l = [x+x for x in l]
```

5)What is the purpose of the if __name__ == "__main__" statement?

This statement is used to ensure that the code block following it is only executed when the Python script is run directly as the main program just like how C++ has int main() function.

6)What is the purpose of the with statement in Python?

The with statement ensures that certain operations, like opening and closing a file, are properly handled. For example, it's often used with file handling to ensure the file is properly closed when done.

7)What are the key features of Spark?

Apache Spark is a powerful, distributed data processing framework with several key features, including:

(i)In-memory data processing for speed.

(ii)Support for machine learning and graph processing.

(iii)Ease of use through high-level APIs in languages like Python and Scala.

8)What are Resilient Distributed Datasets (RDDs) in Spark?

A resilient distributed dataset (RDD) is a collection of elements partitioned across the nodes of the cluster that can be operated on in parallel. It is the fundamental data structure of Apache Spark and the core of Spark. RDDs are immutable and can contain any type of Python, Java, or Scala objects, including user-defined classes. RDDs are resilient, meaning they can be recomputed in case of failures.

9)What is the difference between a DataFrame and an RDD in Spark?

A Resilient Distributed Dataset (RDD) is a fundamental data structure in Spark that represents an immutable, distributed collection of objects. RDDs are fault-tolerant and can be cached in memory for faster processing. They are low-level APIs that provide more control over the data but with lower-level optimizations. On the other hand, a DataFrame is a distributed collection of data organized into named columns similar to an SQL table. It has a schema, which defines the types and names of its columns, and each row represents a single record or observation.

10)What is Spark's ecosystem?

Spark ecosystem refers to the set of components and tools that work with Apache Spark, a general purpose cluster computing system.

Spark ecosystem includes Spark SQL, Spark Streaming, Spark Machine learning (MLlib), Spark GraphX, and Spark R.