Name: Shakib Ibna Shameem Polock. ID: 2307673

Q1. Lesson Learned.

Ans: Form this exercise I got some basic ideas about spark structured streaming and its related terminologies. It is a scalable stream micro batch processing engine built in Spark SQL that process data in a small batch jobs. It also provides high level API’s for continuous stream updates in tables. There are a few keywords that are related to spark structured streaming. Input Source allows to get streaming data into structured streaming with various built-in options like File Source, Kafka Source, Rate Source and Socket Source. For the writing the streamed data various sink is available such as File Sink, Kafka Sink, Foreach Sink, Memory Sink and Console Sink. Output mode defines how the result will appear in the output. Append mode add only the new rows in the table after the last executed trigger. Complete Mode shows the whole result table and the Update mode shows the rows that are updated since the last trigger. The trigger settings in Spark Structured Streaming selects how data is processed. Unspecified runs in micro-batch mode, generating batches upon completion. Fixed Interval Micro-batches triggers batches at set intervals and One-time Micro-batch processes all data in a single batch and Continuous with Fixed Checkpoint Interval offers experimental continuous processing with fixed checkpoint intervals for fault tolerance.

Q2. Difference between methods and operations provided for structured streaming using sparksession and spark streamingcontext?

Ans: In Spark, there are two main approaches for processing streaming data: Structured Streaming with SparkSession and Legacy Streaming with SparkStreamingContext. These two approaches offer different levels of operatins and API’s for working with streaming data. Structured Streaming and SparkSession provides high-level, DataFrame-based API. This API enables to create and manipulate streaming DataFrames using familiar DataFrame operations. For instance, streaming data can be read using the readStream method, transformations can be applied using DataFrame methods like filter, groupBy, and agg, and data can be written using the writeStream method.

On the other hand, Streaming with SparkStreamingContext provides a lower-level API that enables to create and manipulate Discretized Streams, which are represented as sequences of RDDs (Resilient Distributed Datasets). Operations like map, reduceByKey, and saveAsTextFiles are applied directly to DStreams. It provides more control and flexibility over the stream processing. The choice depends on the context of the use case.