BIG DATA ANALYTICS
KAFKA AND SPARK STREAMING
HOMEWORK – 9

1. **What is Apache Spark Streaming?**
   Apache Spark Streaming is a scalable fault-tolerant streaming processing system that natively supports both batch and streaming workloads. Spark Streaming is the most-used and important spark component. Spark Streaming architecture focusses on programming perks for spark developers owing to its ever-growing user base- CloudPhysics, Uber, eBay, Amazon, ClearStory, Yahoo, Pinterest, Netflix, etc.

2. **What are DStreams?**
   Spark Streaming receives live input data streams and divides the data into batches, which are then processed by the Spark engine to generate the final stream of results in batches.

3. **What are DStreams?**
   Spark Streaming provides a high-level abstraction called discretized stream or DStream, which represents a continuous stream of data. It is received from source or from a processed data stream generated by transforming the i/p stream. Internally, a DStream is represented by a continuous series of RDD contains data from a certain level.

4. **What is a Streaming Context object?**
   A Streaming Context object can be created from a SparkContext object. A SparkContext represents the connection to a spark cluster and can be used to create RDDs, accumulators, and broadcast variables on that cluster.

5. **What are some of the common transformations on DStreams supported by Spark Streaming?**
   Common transformations on Dstreams supported by spark streaming are :
   - map()
   - flatmap()
   - filter()
   - reduce()
   - groupBy()

6. **What are the output operations that can be performed on DStreams?**
   The output operations allow the DStream's data to be pushed to external systems like db or file systems. The output operations in DStream are :
   - Print()
   - saveAsTextFiles(prefix,[suffix])
   - saveAsObjectFiles(prefix,[suffix])
   - saveAsHadoopFiles(prefix,[suffoix])
   - foreachRDD()