Transform Operation

The `transform()` operation

- Overview: `transform` operation (and `transformWith` variant) allow arbitrary RDD-to-RDD functions to be applied on a DStream.
- It can be used to apply any RDD operation that is not exposed in the DStream API.
- For example, the functionality of joining every batch in a data stream with another dataset is not directly exposed in the DStream API. However, you can easily use `transform` to do this.

Example Application

- Real-time data cleaning by joining the input data stream with precomputed spam information, and then filtering based on it.

```
spamInfoRDD = sc.pickleFile(...)  # RDD containing spam information


# join data stream with spam information to do data cleaning
cleanedDStream = wordCounts.transform(lambda rdd:
rdd.join(spamInfoRDD).filter(...))
```

Important Note

- functions supplied to `Transform()` get called in every batch interval.
- You can create operations for RDDs that vary by time.
- The RDD operations, the number of partitions in the DStream, the specific broadcast variables, can all be changed between batches.