Quiz 3

Presentation1: Spark Streaming

(a) Typical systems,there are two approaches for failure discovery (1) Replication and (2) Upstream backup

-Describe what each approach does.

-What is a common limitation in these two approaches?

(b) What is the difference between a Straggler Node and a Failed Node?

How does the Spark Streaming detect and handle each case?

(c) Spark Streaming introduced several optimizations to Spark for efficiency. Two of these optimizations are Timestep Pipelining and Lineage Cutoff. Describe what does each optimization do?

(a) Replication : Node replication use a separate set of nodes process the same stream along
with the main nodes. These nodes act as hot failover nodes. On failure, the system turn to check these replication nodes. These nodes are also called hot failover nodes.

Upstream backup: Each node maintain a backup of all records until they are knowledged the downstream finish their work and reach a checkpoint. On failure, system feed backuped record to a node to recompute results.

The common limitation is that neither approach handles stragglers.

(b) Straggler Node is a node that runs slowlier than other node, while a failure is running time error that a node can not continue and must stop.

(c) Timestep Pipelining: scheduler could allow submitting tasks from the next timestep before the current one has finished.

Lineage Cutoff: The scheduler will forget lineage after an RDD has been checkpointed, so that the state won't grow to arbitrarily large.

Presentation 2: Titian

(a) What is the difference between typical Spark RDD and Titian's Lineage RDD?

Give two examples of operations you can do on Lineage RDD but not the typical RDD?

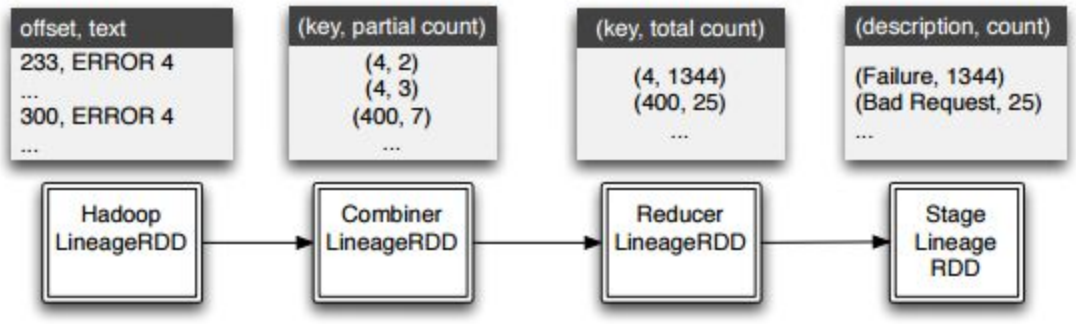
(b) Titian introduces several agents (type of Lineage RDDs) to captures the lineage at different stages of the workflow. Below is a workflow from the paper.

-Write down under each box (at the arrow) the type of the agent that titian uses to capture the lineage

-Indicate which of the agents corresponds to a blocking operation.

(a). Lineage RDD is an extension of typical RDD. Lineage RDD offer data tracing functionalities.
e.g.goBack or goNext

(b).



The Combiner Lineage RDD and Reducer Lineage RDD correspond to blocking operation.