**Student Name:**

**Presentation 1 [Spark Streaming] [20 Points]**

**(a)** In typical systems, there are two main approaches for failure recovery: (1) *Replication*, and (2) *Upstream Backup.*

- [5 Points] Describe in 3-4 sentences, what each approach does

- [2 Points] What is a common limitation in these two approaches?

**(b)** [4 Points]What is the difference between a Straggler Node and a Failed Node?

[4 Points] How does Spark Streaming detect and handle each case?

**(c)** [5 Points]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 [3-4 sentences each].

*Left Empty for answers*

**Presentation 2 [Titian] [10 Points]**

**(a)** [3 Points] What is the difference between typical *Spark RDD* and *Titian’s LineageRDD*?

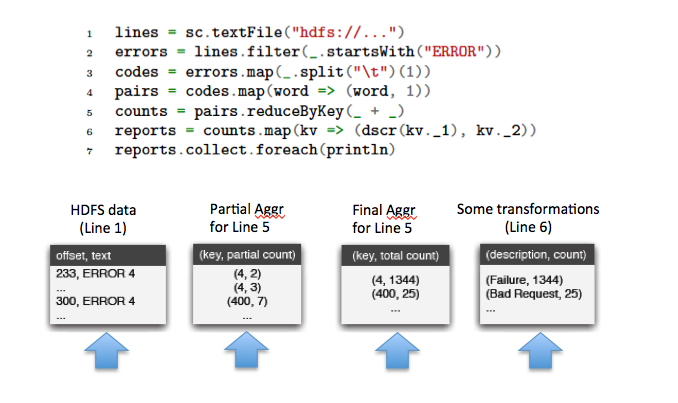
[2 Points] Give two examples of operations you can do on LineageRDD but not the typical RDD?

(b) Titian introduces several “agents” (types of LineageRDDs) to captures the lineage at different stages of the workflow. Below is a workflow from the paper.

[3 Points] Write down under each box (at the arrow) the type of the agent that Titian uses to capture

the lineage.

[2 Points] Indicate which of the agents corresponds to a blocking operation.



*Left Empty for answers*