**University of California, Santa Cruz - Silicon Valley Extension**

**Class: DBDA.X407 Hadoop: Distributed Processing of Big Data**

**Full Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**===== Midterm ======**

**8 Questions - 40 points**

**a1.** Select the one that **IS NOT** one the 5 V’s of Big Data? (5 points)

1. Velocity.
2. Variety.
3. Volume
4. Veracity.
5. Validity.

**a2.** How the Spark context is used in a Spark program? (5 points)

1. It allows storing cached variables.
2. Is the handle to connect to a Spark cluster.
3. It’s used as a transformation.
4. It’s used as a type of RDD.
5. None of the above.

**a3.** What does the command **distcp** do? (5 points)

1. Allows to run distributed jobs in the cluster.
2. Sets the limits on storage in HDFS.
3. It’s a utility program used to perform parallel copy between clusters.
4. This program is deprecated and it’s not used Hadoop anymore.
5. None of the above.

**a4.** What’s the benefit of Hadoop Job Queues? (5 points)

1. They make jobs more secure.
2. They define resource allocation rules and allow several groups to concurrently use the cluster.
3. By queuing the jobs, Hadoop can make sure only one job will be running at any point in time.
4. Job Queues are used to store logs about the Hadoop jobs.
5. None of the above.

**a5.** How the Attempts, Jobs and Tasks are related? (5 points)

1. A Job contain an attempt and each attempt is made of tasks.
2. A job is made of tasks and each task can contain one or more attempts.
3. An attempt represents a full try of Job execution including its tasks.
4. Jobs contains tasks and attempts are not related to either jobs or tasks.
5. None of the above.

**a6.** a Spark function named ‘my\_operation’ receives an RDD as input and returns an RDD.

What can you say about this function**?** (5 points)

1. This function is an action.
2. This function is a transformation.
3. It can be an action or transformation.
4. It’s not an action nor a transformation just a function.
5. None of the above.

**a7.** Explain in your own words what is a Spark RDD. (1-2 paragraphs). (10 points)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_