Family name:…………………………………………………………………………………… Given name:………………………………………………

1) (30%) How does HDFS decide where to place the different chunks of a file?

- HDFS has a master/slave architecture. An HDFS cluster consists of a single NameNode, a master server that manages the file system namespace and regulates access to files by clients.
- There are a number of DataNodes, usually one per node in the cluster, which manage storage attached to the nodes that they run on.
- Afile is split into one or more blocks and these blocks are stored in a set of DataNodes.
- The NameNode executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes.
- The DataNodes are responsible for serving read and write requests from the file system's clients.
- The DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode.

2) (40%) What indicates to Spark query optimizer the end of a stage and the beginning of the next one?

- The execution of a Spark job is based on its DAG representation and RDDs dependencies, which are automatically optimized.
- The scheduler examines the DAG and builds a new data structure with stages. This data structure can be seen as the execution plan of a Spark job.
- Stages are actually subgraphs of the DAG defined with the objective of maximizing the number of narrow dependencies inside, while the boundaries of the stages are the wide dependencies.

3) (30%) Which are the four axis of the Devil's Quadrangle of performance measures?

1-…Time……………………………………………………………………………………

2-…Cost……………………………………………………………………………………

3-…Quality…………………………………………………………………………………

4- Flexibility…………………………………………………………………………………