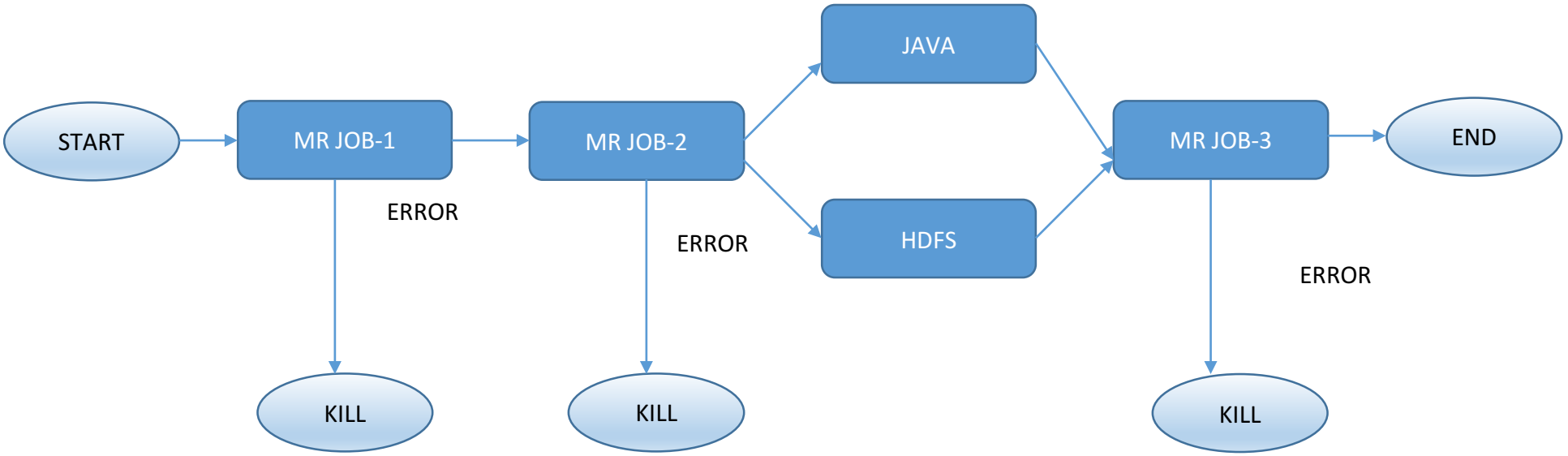
CS 698 Big Data
Flight Data Analysis
Project Report

Submitted by
Sathyaa Erode Kandasamy (se244)
Krithivas Swaminathan (ks583)

a. Oozie workflow diagram:



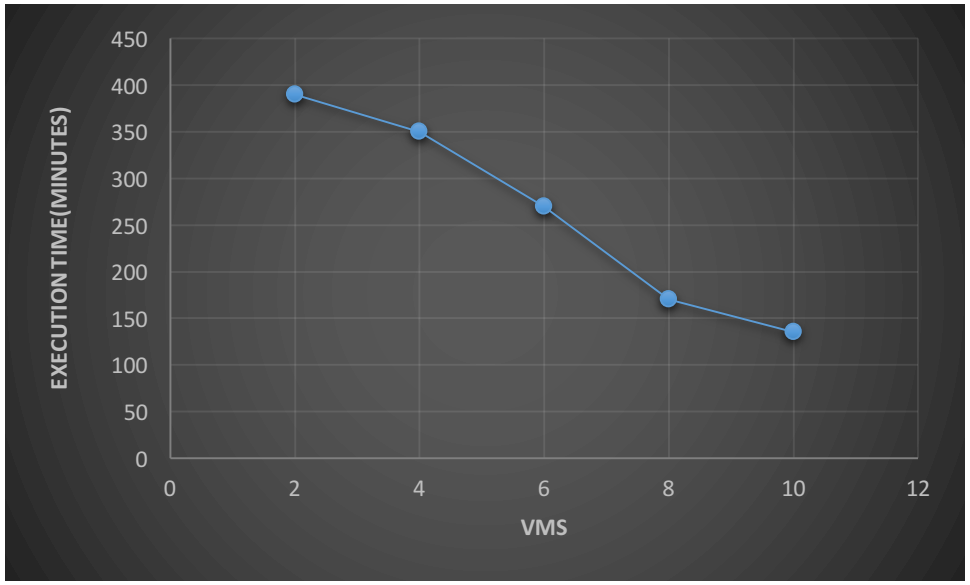
b. A detailed description of the algorithm you designed to solve each of the problems

- The computation takes a set of input key/value pairs, and produces a set of output key/value pairs.
- The user of the MapReduce library expresses the computation as two functions: map and reduce. λ
- Map, written by the user, takes an input pair and produces a set of intermediate key/value pairs.
- The MapReduce library groups together all intermediate values associated with the same intermediate key and passes them to the reduce function. λ
- The reduce function, also written by the user, accepts an intermediate key and a set of values for that key. It merges together these values to form a possibly smaller set of values.

    - Mapper: Identity function for value (k, v) (v, _)
    - Reducer: Identity function (k', _) -> (k', "")

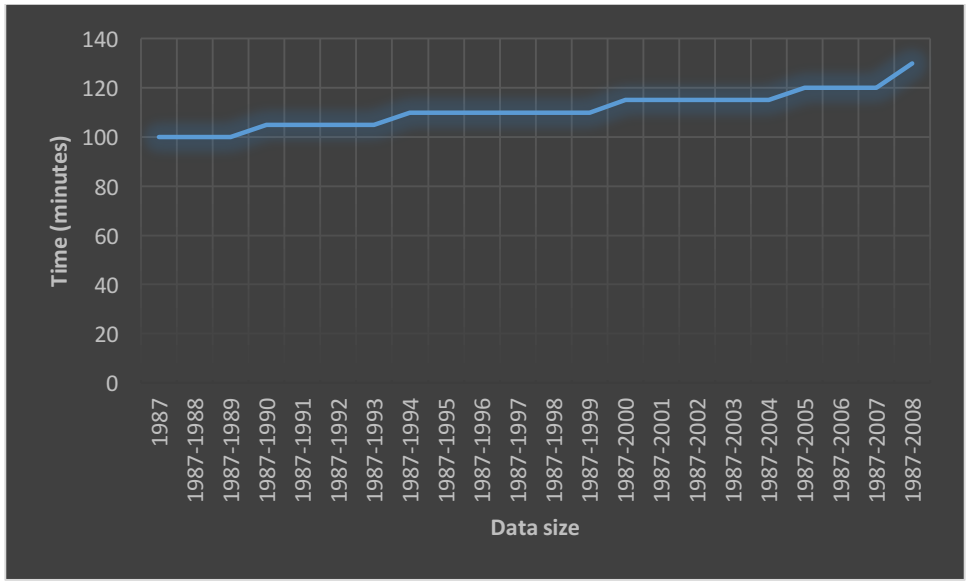
- Typically just zero or one output value is produced per reduce invocation.
- The intermediate values are supplied to the user's reduce function via an iterator.
- This allows us to handle lists of values that are too large to fit in memory.
- We have used parallel algorithm which can also be described a SPMD algorithm.
- The raw data is split into blocks, which is less than 64MB, and the map function is executed on these blocks, producing a pair
- The reduce function handles the iterator to produce the final answer.

c. **A performance measurement plot that compares the workflow execution with the increasing number of VMs**



The execution time seems to decrease as we increase the number of VMs, as there is considerable number of computing power added for the execution

d. A performance measurement plot that compares the workflow execution with increasing data size



As the data size increases, the execution time increases gradually.