Project Report

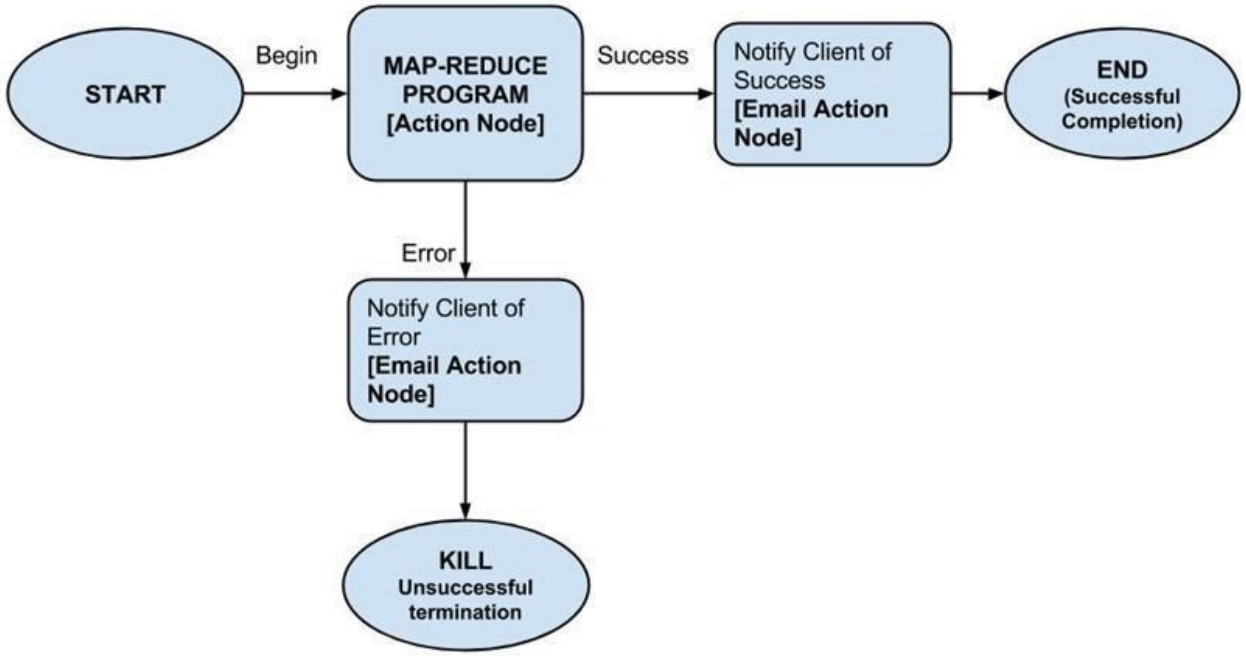
Flight Data Analysis

Submitted by

Harshil Shah (has37@njit.edu)

Praveen Kumar (pm424@njit.edu)

- **Basic Oozie Workflow Diagram :**


START
Begin
MAP-REDUCE PROGRAM [Action Node]
Success
Notify Client of Success [Email Action Node]
END (Successful Completion)
Error
Notify Client of Error [Email Action Node]
KILL Unsuccessful termination

- ## **Algorithm :**

Step 1:

The input data can be divided into n number of chunks depending upon the amount of data and processing capacity of individual unit.

partition (k’, number of partitions) ? partition for k’

Often a simple hash of the key, e.g., hash(k’) mod n

Divides up key space for parallel reduce operations

Step 2 :

Next, it is passed to the mapper functions. Please note that all the chunks are processed simultaneously at the same time, which embraces the parallel processing of data.

map (k, v) ? <k’, v’>

Step 3:

After that, shuffling happens which leads to aggregation of similar patterns.

Step 4:

Finally, reducers combine them all to get a consolidated output as per the logic.

reduce (k’, v’) ? <k’, v’>*. All values with the same key are sent to the same reducer.

Mini-reducers that run in memory after the map phase.

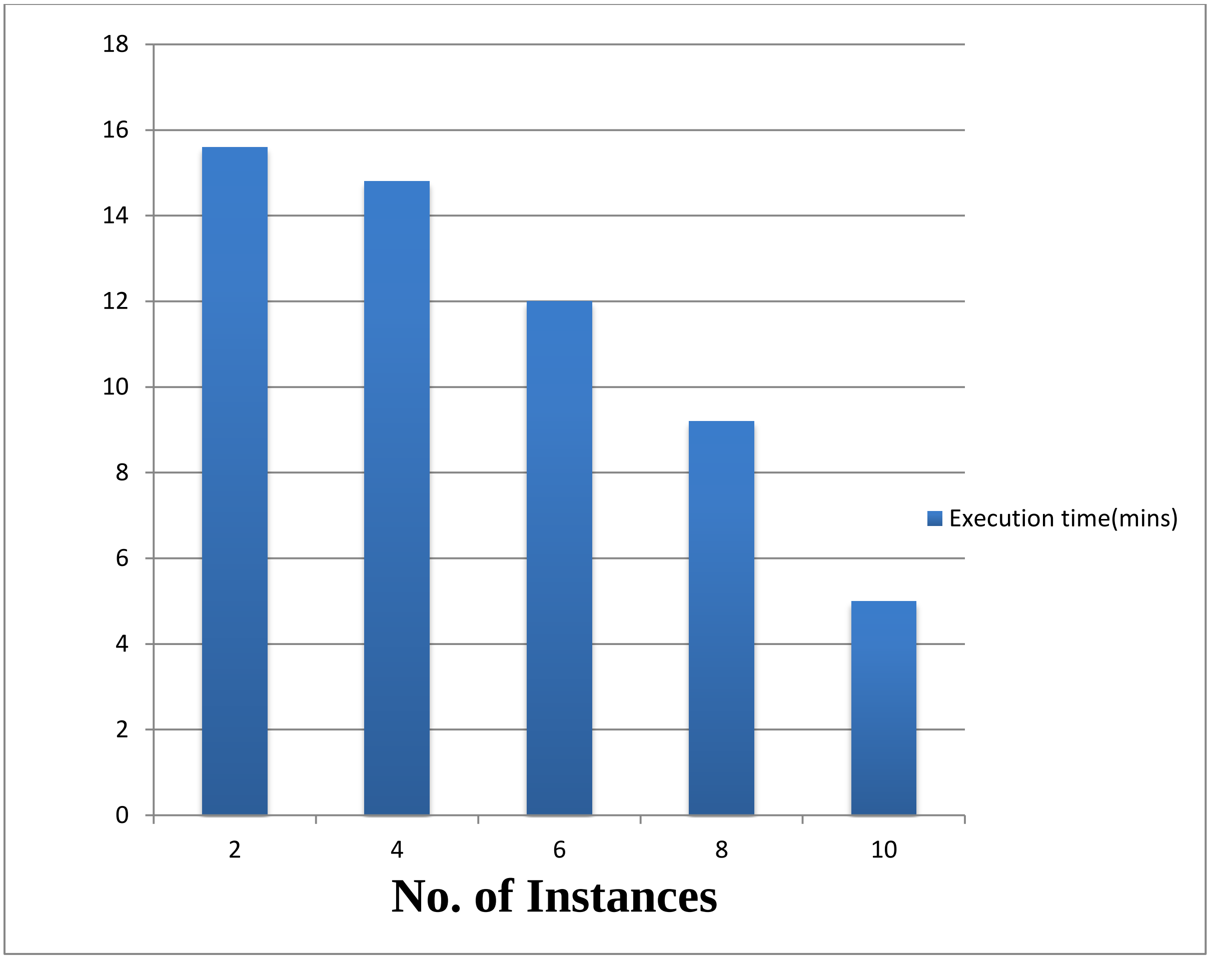
Used as an optimization to reduce network traffic.

Step 5:

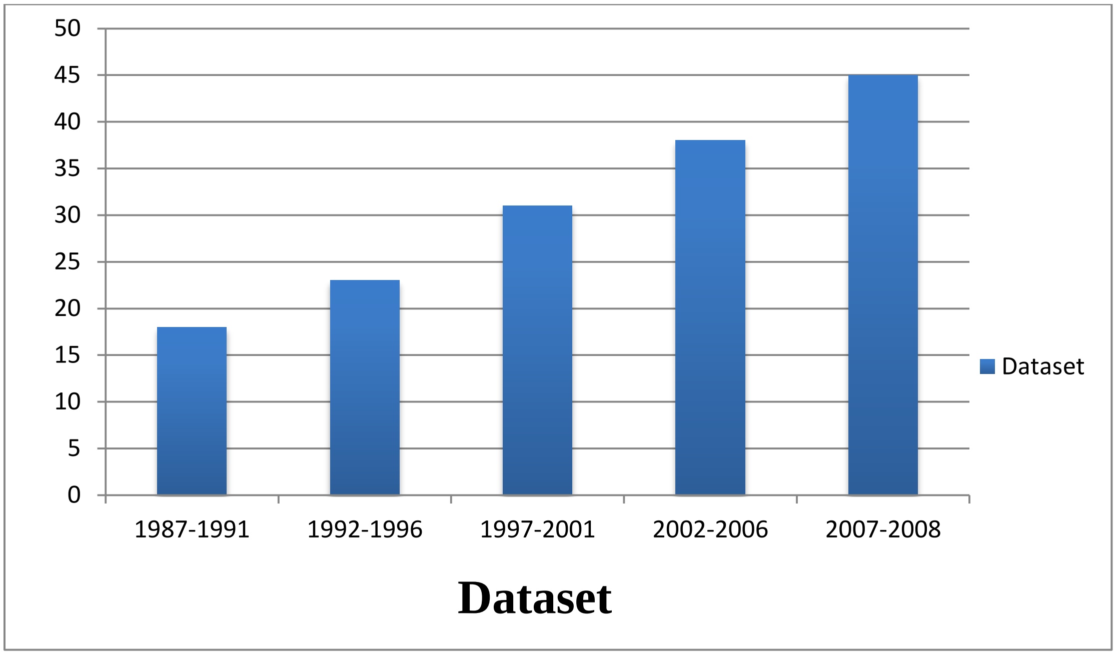
This algorithm embraces scalability as depending on the size of the input data, we can keep increasing the number of the parallel processing units.

- ## Performance Analysis:

A performance measurement plot that compares the workflow execution time in response to an increasing number of VMs used for processing the entire data set (22 years) and an in-depth discussion on the observed performance comparison results.



A performance measurement plot that compares the workflow execution time in response to an increasing data size (from 1 year to 22 years) and an in-depth discussion on the observed performance comparison results.


Dataset
Dataset