BIG DATA

Hands On Session - 1

MapReduce

K V Subramaniam
Computer Science and Engineering

Overview

BIG DATA

MapReduce Handson Setup

In order to execute the wordcount application, we need

Ubuntu –virtualbox (example Ubuntu 20.04)

Java version 8

Hadoop v3.2.1

Wordcount application

Overview

MapReduce is a programming model.

Implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster.

Mapper method performs filtering and sorting, and a Reduce method, performs an aggregate operation.

We aim to solve a real world problem using MapReduce.

Problem Statement

Find the number of cars in every city which use gas as a mode of fuel using MapReduce.

BIG DATA

MapReduce Hands-on Exercise

SPECIFICATIONS

1. Ubuntu 16.04+
2. Hadoop: 3.2
3. Python: 2.x/3.x
4. Java: 1.8
5. Dataset: You will be using the modifie
   Dataset" for this session. Please dow
   the below Google Drive link:
   https://drive.google.com/open?id=1GxEaY_aAlkMHfJN2Z1C
   vt1O1yNtCp1gN

Columns of the Dataset : The columns are indexed from [0-25] (Ex. Transmission is the 11th index)

Sample output :

City	Number of Cars that use Gas
Bangalore	10
Chennai	12

Actual output to be in a text file with each line of the answer having the pair <cityname> <number> .

Pointers & Clues

Python Coders:

start the code with the python shebang:

#!/usr/bin/python

use sys module to read from stdin

Java Coders - A sample word count program can be found here:

https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html