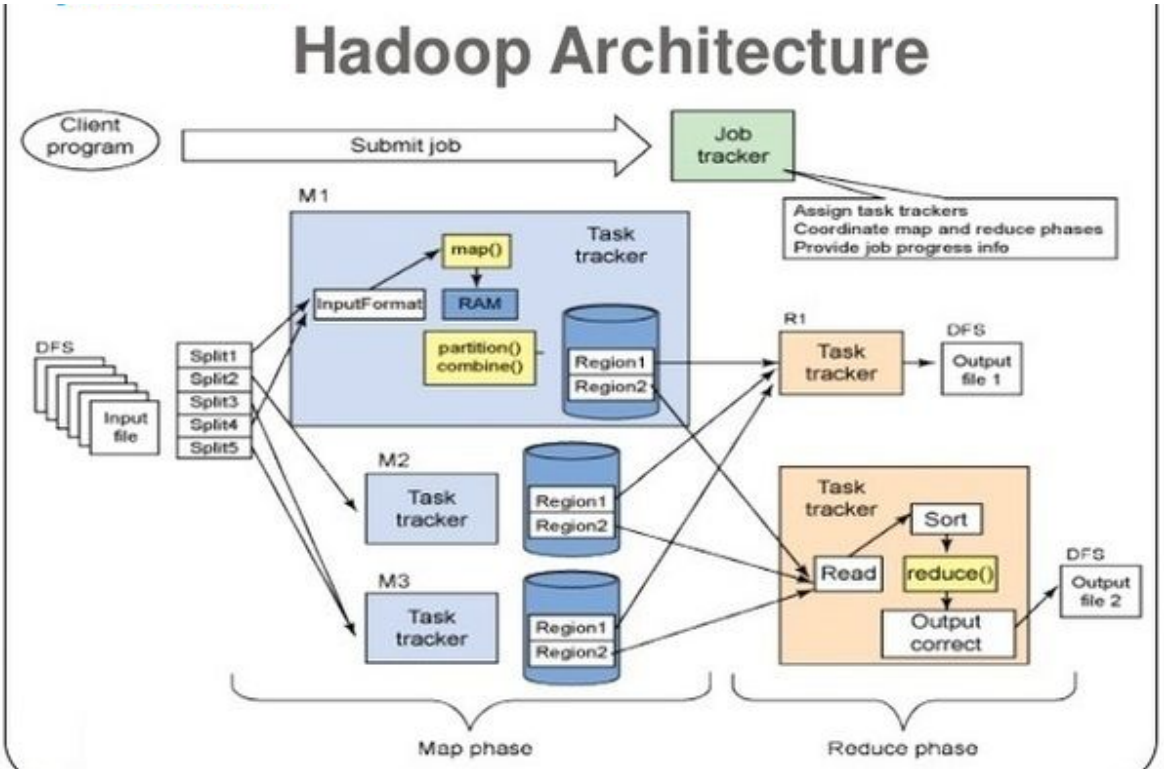
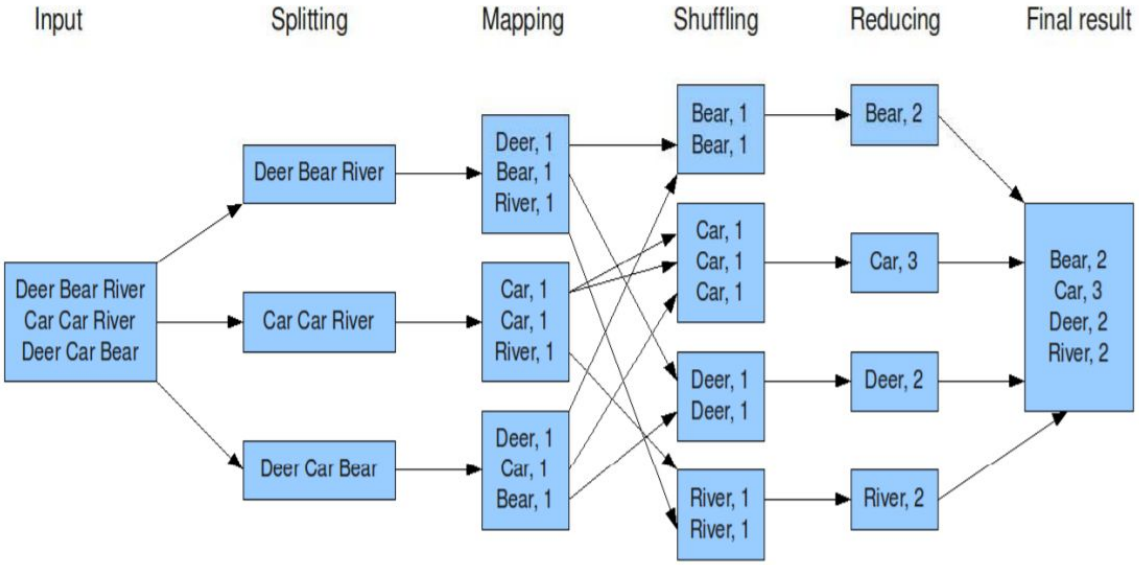
Hadoop map reduce



Reminder



Read the original MapReduce paper by Dean and Ghemawat You can find the paper at http://research.google.com/archive/mapreduce.html

1. Answer the following questions:
    a. How do the input keys/values, the intermediate keys/values and the output keys/values relate?
    b. How does MapReduce deal with node failures?
    c. What is the meaning and the implication of locality? How is it used?
    d. Which problem is addressed by introducing a combiner function to the MapReduce model?
2. Go over the WordCount example and run it on your system https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html#Example%3A+WordCount+v1.0
   Important note: Make sure you delete the output directory before running your application again. Otherwise you will get an error from Hadoop indicating that the directory already exists

3. Implement one executable Hadoop MapReduce program to perform the inner join of two tables based on “Student ID” . You can create sample data in below format. Or generate using https://mockaroo.com/

T1

STUDENT_ID	NAME	YEAR_OF_BIRTH

T2

STUDENT_ID	SEMESTER_NUM	AVG	MAX

4. Implement one executable Hadoop MapReduce program to calculate the highest temperature for **every** given year. (generate data according the format) Attention generate the data with **different** samples for **every** year

Year	Temp

5. Implement one executable Hadoop MapReduce program to calculate AVG temperature for **every** given year. (generate data according the format)

   Attention generate the data with few different samples for every year

Year	Temp