

Lab 4

[Click here to Register Attendance](https://forms.gle/pesShFDe1ibVfS3L7)

|  |  |
| --- | --- |
| Name |  |
| Date |  |
| Student No |  |
| Student Email |  |

### **Building your own MapReduce Project**

1. Brainstorm ideas for your own MapReduce/HDFS project to solve a real-world problem - this could gather data from a work situation, a small home project, your research project, a small sensor based project using a hobby board e.g. Raspberry Pi or Arduino Yun (we can provide these boards if you need a loan)  
   Enter a brief description of your Project idea here:

|  |
| --- |
|  |

1. Determine what the data schema for this project would look like and design the data fields, making sure to choose a unique Primary Key. Keep your data file flat for this first project - ideally a single plain text, CSV file or similar. Sketch a rough data schema here showing field names, datatypes and identify your primary key field:

|  |
| --- |
|  |

1. You can create a real dataset, generate a synthetic dataset, stream data from an API e.g. Twitter, or use an existing third-party dataset, for the purpose of building out the project. Describe the kind of dataset you have in mind here:

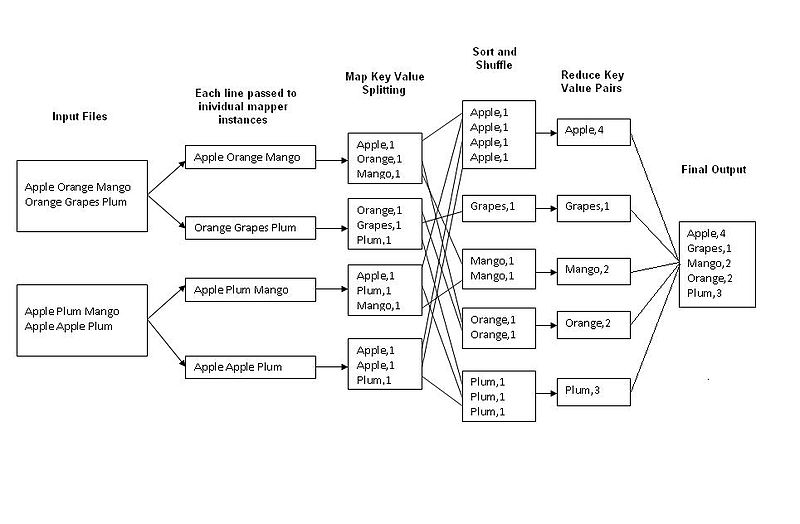
|  |
| --- |
|  |

1. Calculate how much data this system would generate per minute/hour/day/week/month/year

|  |
| --- |
|  |

1. What do it look like in terms of the 3 Vs -   
   Volume (size)  
   Velocity (how quickly it’s growing)   
   & Variety (what different types of data does it include)?

|  |  |
| --- | --- |
| Volume |  |
| Velocity |  |
| Variety |  |

1. In terms of a simple MapReduce paradigm what does the data look like at each of the following stages of the MapReduce operation (map, sort & group, reduce) - sketch an outline below the lecture example:  
   
2. Adapt the Lab 3 code to deal with your dataset and perform perform at least one calculation