Assignment 2

1. (20 pts) What is virtualization? What are the benefits we can have with it?

Virtualization is creating a virtual version of something. You could create virtual computer storage devices, computer network resources, or hardware platforms.

Virtualization reduces the number of physical servers any given company needs on site, cuts energy consumption, creates independent user environments, and overall saves tons of money on IT costs and web hosting for servers.

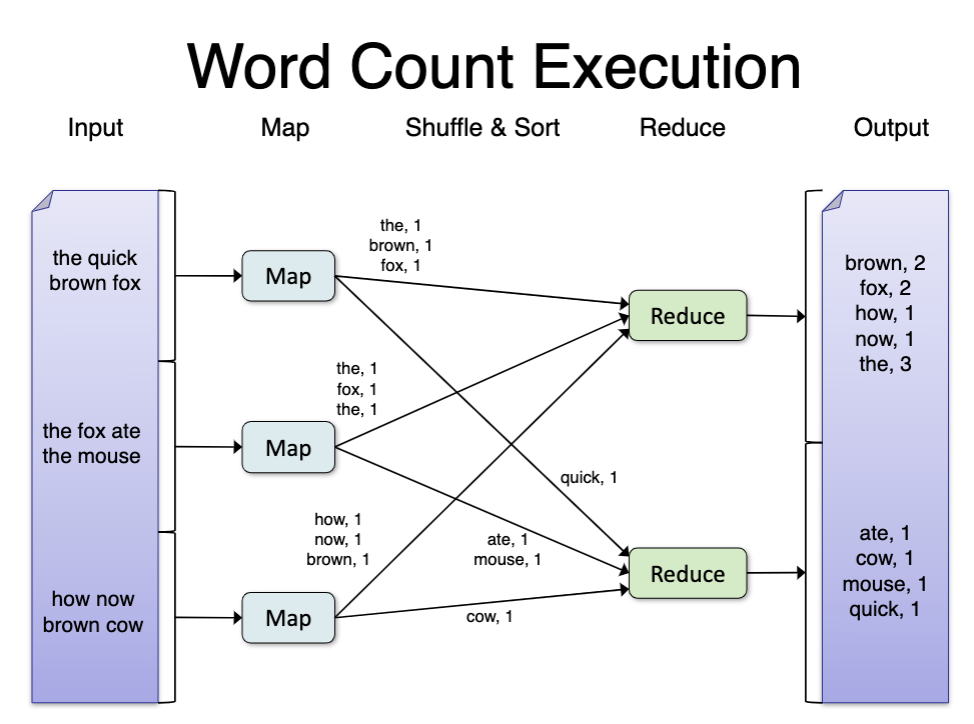
1. (20 pts) Based on virtualization technologies, how many types can we classify them? Please give a brief statement on each of them.

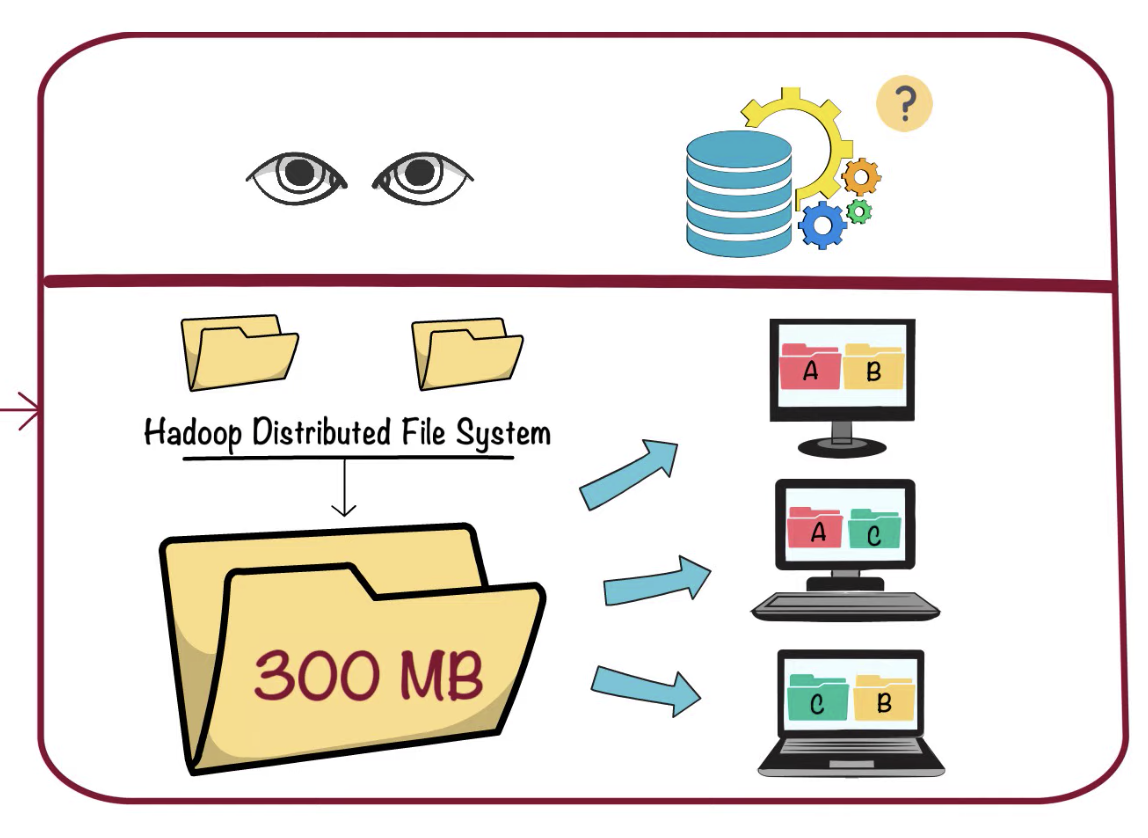
Container Virtualization - Virtual machine where guests share the same filesystem and kernels and provides extra level of security.

Full Virtualization - Runs unmodified guests and simulates bios, communicates through ACPI and BIOS.

Paravirtualization - Works as a guard and passes instructions directly to CPU and devices by working as a guard.

1. (20 pts) Core Hadoop has two main systems, what are they?

Hadoop/MapReduce - big data processing infrastructure that filtering, sorting, and reducing

HDFS (Hadoop Distributed File System) - A way of storing files to keep high bandwidth, high availability, and stay fault tolerant

1. (20 pts) How does MapReduce work? Please give a brief explanation, figures and diagrams could be used if necessary.

MapReduce frameworks consist of three operations:

Map: the map function is applied to the local data and writes output to a temporary storage. A separate master node makes sure that only one copy of input is processed

Shuffle: The data redistributes data based on the output keys received from the map function. Data with the same keys are grouped together

Reduce: Each group of output is processed per key at the same time

1. (20 pts) How many main component in HDFS and what are the responsibilities of each component?

RHadoop - Statistical data processing using R

Mahout - Machine Learning tool

Hive - Tool for NoSQL (supports semi-structured/unstructured data)

Pig - Tool for NoSQL (supports semi-structured/unstructured data)

Sqoop - Tool for getting data into and out of the hadoop file system