# Data Application Lab

## Friday Quiz

1. **What is Big Data?**

Big data is defined as the voluminous amount of structured, unstructured or semi-structured data that has huge potential for mining but is so large that it cannot be processed using traditional database systems.

Big data is characterized by its high velocity, volume and variety that requires cost effective and innovative methods for information processing to draw meaningful business insights.

2. **What is a block in HDFS?**

The minimum amount of data that can be read or written is generally referred to as a "block" in HDFS. The default size of a block in HDFS is 64MB.

3. **Differentiate between Structured and Unstructured data.**

Data which can be stored in traditional database systems in the form of rows and columns can be referred to as Structured Data. Data which can be stored only partially in traditional database systems can be referred to as semi structured data. Unorganized and raw data that cannot be

categorized as semi structured or structured data is referred to as unstructured data.

4. **On what concept the Hadoop framework works?**

HDFS: Hadoop Distributed File System is the java based file system for scalable and reliable storage of large datasets. Data in HDFS is stored in the form of blocks and it operates on the Master Slave Architecture.

Hadoop MapReduce: This is a java based programming paradigm of Hadoop framework that provides scalability across various Hadoop clusters. MapReduce distributes the workload into various tasks that can run in parallel. Hadoop jobs perform 2 separate tasks job. The map job breaks down the data sets into key-value pairs or tuples. The reduce job then takes the output of the map job and combines the data tuples to into smaller set of tuples. The reduce job is always performed after the map job is executed.