1. **Explain the core changes made in Hadoop 2.x.**

**ANS:**

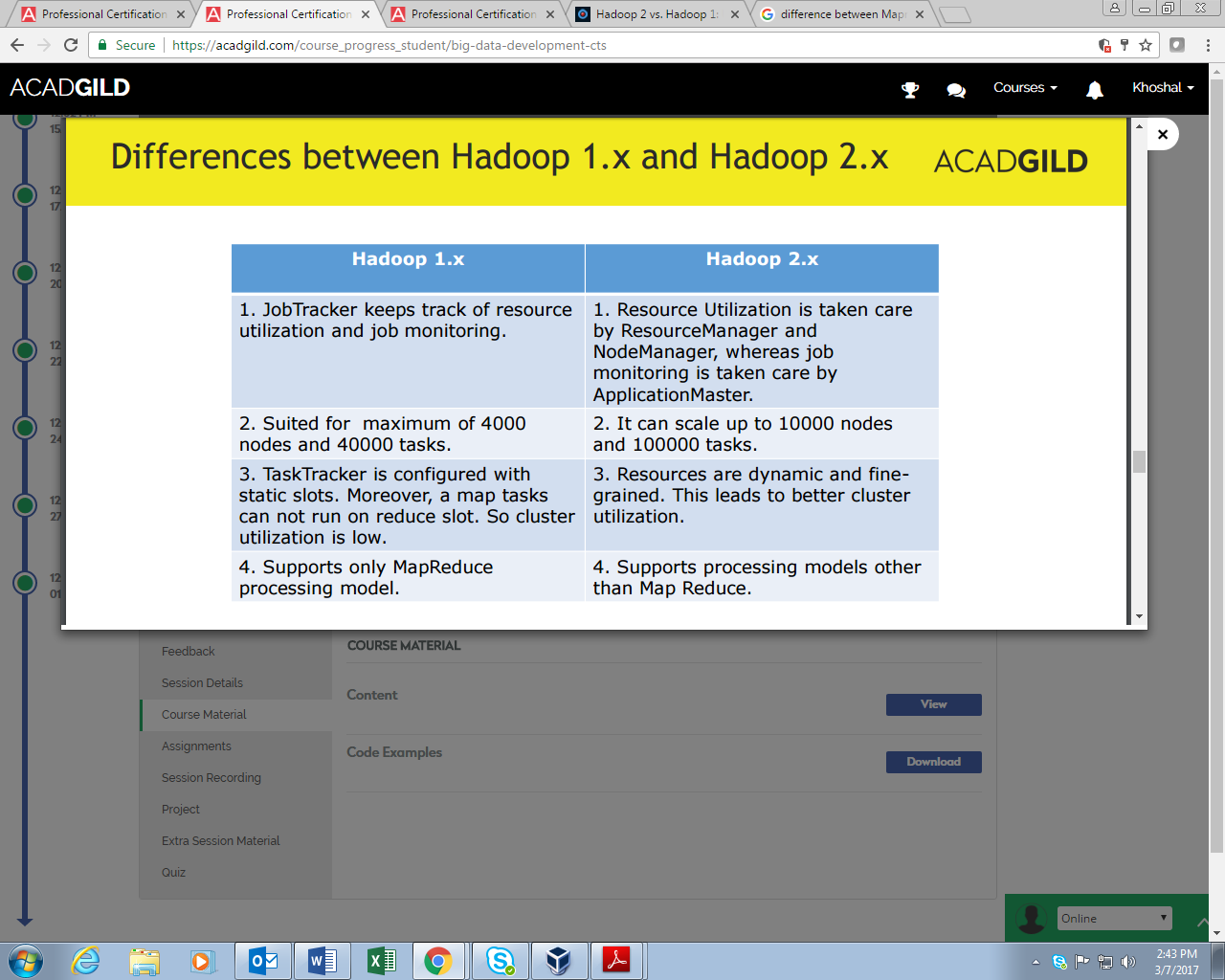
The Hadoop developers rewrote the major components of the file system to form the Hadoop 2.x platform. The two most important advances in the Hadoop 2.x platform are the introduction of HDFS federation and resource manager YARN (Yet Another Resource Manager).

HDFS is the Hadoop file system and comprises of two major components such as

1) Namespaces and 2) Blocks Storage Services.

The Namespace service manages operations on files and directories, such as creating and modifying files and directories.

The block storage service implements data node cluster management, block operations and replication.



**Difference between Hadoop 1.x and Hadoop 2.x**

The major differences between Hadoop 1.x and 2.x is listed below.

In Hadoop 1.x, a single Namenode manages the entire namespace for a Hadoop cluster. With HDFS federation, multiple Namenode servers manage namespaces and this allows for horizontal scaling, performance improvements and many more name spaces.

The HDFS federation allows existing namenode configurations to run without changes. For

Hadoop administrators, moving to HDFS federation requires formatting Namenodes.

**YARN:**

HDFS federation brings measures of scalability and reliability to Hadoop. YARN, the other major advancement in Hadoop 2.x. It brings significant performance improvements for some applications, supports additional processing models, and implements a more flexible execution engine.

YARN is a resource manager that was created by separating the processing engine and resource management capabilities of MapReduce as it was implemented in Hadoop 1.x.

YARN is often called the OS (Operating System) of the Hadoop because it manages and monitors workloads and maintaining a multi-storey environment, implementing security controls, and managing high availability features of Hadoop. YARN supports multiple processing models in addition to MapReduce.

The most significant benefits of this is that we are no longer limited to working the often I/O intensive high latency MapReduce framework. This advance means Hadoop users should be familiar with the Plus and the minus of the new processing models and understand what and when to apply them to case specific problems.

**HIVE:**

Hive is the most popular SQL-in-Hadoop option and the Hadoop community has invested heavily in making

Hive faster, and more scalable, and more supportive of SQL operations. This work has been done under the Stinger initiative.

**TEZ:**

Further speed improvements will come with the production release of TEZ, which is a workflow for executing programs using a directed acyclic graph model.

**SPARK:**

Spark is a top-level Apache project that leverages memory to improve the speed of large-scale data analytical problems.