Quiz 5 - Intro to MapReduce

1. What does IaaS provide?

- Software On-Demand
- Computing Environment
- **Hardware Only**

2. What does PaaS provide?

- **Computing Environment**
- Software On-Demand
- Hardware Only

3. What does SaaS provide?

- Computing Environment
- Hardware Only
- **Software On-Demand**

4. What are the two key components of HDFS and what are they used for?

- FASTA for genome sequence and Rasters for geospatial data.
- NameNode for block storage and Data Node for metadata.
- **NameNode for metadata and DataNode for block storage.**

5. What is the job of the NameNode?

- **Coordinate operations and assigns tasks to Data Nodes**
- Listens from DataNode for block creation, deletion, and replication.
- For gene sequencing calculations.

6. What is the order of the three steps to Map Reduce?

- Map -> Reduce -> Shuffle and Sort
- Shuffle and Sort -> Reduce -> Map
- **Map -> Shuffle and Sort -> Reduce**
- Shuffle and Sort -> Map -> Reduce

7. What is a benefit of using pre-built Hadoop images?

- Guaranteed hardware support.
- Less software choices to choose from.
- **Quick prototyping, deploying, and validating of projects.**
- Quick prototyping, deploying, and guaranteed bug free.

8. What is an example of open-source tools built for Hadoop and what does it do?

- Giraph, for SQL-like queries.
- Pig, for real-time and in-memory processing of big data.
- Zookeeper, analyze social graphs.
- **Zookeeper, management system for animal named related components**

9. What is the difference between low level interfaces and high level interfaces?

- **Low level deals with storage and scheduling while high level deals with interactivity.**
- Low level deals with interactivity while high level deals with storage and scheduling.

10. Which of the following are problems to look out for when integrating your project with Hadoop?

- **Random Data Access**
- **Infrastructure Replacement**
- Data Level Parallelism
- **Task Level Parallelism**
- **Advanced Alogrithms**

11. As covered in the slides, which of the following are the major goals of Hadoop?

- **Facilitate a Shared Environment**
- **Provide Value for Data**
- Latency Sensitive Tasks
- **Enable Scalability**
- **Handle Fault Tolerance**
- **Optimized for a Variety of Data Types**

12. What is the purpose of YARN?

- **Allows various applications to run on the same Hadoop cluster.**
- Enables large scale data across clusters.
- Implementation of Map Reduce.

13. What are the two main components for a data computation framework that were described in the slides?

- Resource Manager and Container
- Applications Master and Container
- Node Manager and Applications Master
- **Resource Manager and Node Manager**
- Node Manager and Container