Quiz 5 - Pipeline and Tools

1. What is data-parallelism as defined in lecture?

- Having multiple multiple data pipelines at the same time.
- Simultaneously processing input data from multiple cores.
- **Running the same function simultaneously for the partitions of a data set on multiple cores.**
- At each step of the data pipeline, process values simultaneously by using multiple cores.

2. Of the following, which procedure best generalizes big data procedures such as (but not limited to) the map reduce process?

- split->sort->merge
- **split->do->merge**
- split->map->shuffle and sort->reduce
- split ->shuffle and sort->map->reduce

3. What are the three layers for the Hadoop Ecosystem? (Choose 3)

- Data Manipulation and Integration
- **Data Management and Storage**
- **Data Integration and Processing**
- **Coordination and Workflow Management**
- Data Creation and Storage

4. What are the 5 key points in order to categorize big data systems?

- **Execution model, Latency, Scalability, Programming Language, Fault Tolerance**
- Coordination, Latency, Productivity, Speed, Fault Tolerance
- Execution model, Speed, Scalability, Flexibility, Fault Tolerance
- Coordination, Latency, Productivity, Flexibility, Fault Tolerance

5. What is the lambda architecture as shown in lecture?

- **A type of hybrid data processing architecture.**
- A type of architecture that only contains part of the data processing method.
- A type of swappable data processing layer.
- An architecture that natively supports lambda calculus.

6. Which of the following scenarios is NOT an aggregation operation?

- Counting the total number of data per type.
- Averaging the total number of data per type.
- **Removing undefined values.**
- Counting the total number of data.

7. What usually happens to data when aggregated as mentioned in lecture?

- Data become organized.
- **Data becomes smaller.**
- Data becomes personalized.
- Data becomes faster to process.

8. What is K-means clustering?

- Divide samples using k lines.
- Classify data by k decisions.
- **Group samples into k clusters.**
- Classify data by k actions.

9. Why is Hadoop not a good platform for machine learning as mentioned in lecture? (Choose 4)

- Too massive.
- Requires nodes and multiple machines.
- **Bottleneck using HDFS.**
- **Map and Reduce Based Computation.**
- Unable to support machine learning.
- **No interactive shell and streaming.**
- **Java support only.**

10. What are the layers (parts) of Spark? (Choose 5)

- **SparkSQL**
- **Graphx**
- **MLlib**
- Spark Graph
- **Spark Core**
- Spark RDD
- **Spark Streaming**
- Worker Node

11. What is in-memory processing?

- Having the pipeline completely in disk.
- Writing data to disk between pipeline steps.
- **Writing data to memory between pipeline steps.**
- Having the pipeline completely in memory.
- Having the input completely in disk.
- Having the input completely in memory.