Quiz 7 - More on Spark

1. Which part of SPARK is in charge of creating RDDs?

- **Driver Program**
- Local CPU
- Storage
- Spark Executor
- Worker Node

2. How does lazy evaluation work in Spark?

- Transformations are queued and executed at a certain threshold.
- **Transformations are not executed until the action stage.**
- Actions are queued and executed at a certain threshold.
- Actions are not executed until the transformation stage.

3. What are the consequences of lazy evaluation as mentioned in lecture?

- **Errors sometimes do not show up until the action stage.**
- Hiccups within the system during queue execution.
- There are no consequences.

4. What is a wide transformation?

- **A transformation that requires data shuffling across node partitions.**
- Transformations that take a lot of nodes to complete.
- A longer time-taking transformation compared to narrow transformations.
- The name for the most used transformations.

5. Where does the data for each worker node get sent to after a collect function is called?

- Other Worker Nodes
- Spark Streaming
- **Spark Context**
- None; Stays in the Same Node
- Spark SQL

6. What are DataFrames?

- A special type of data node that contains framework to manipulate SQL.
- **A column like data format that can be read by Spark SQL.**
- A type of narrow transformation.

7. Can RDD's be converted into DataFrames directly without manipulation?

- Yes
- **No: lines have to be converted into row.**
- No: RDD’s needed to be made relational first.
- No: RDD’s cannot be converted into DataFrames.

8. What is the function of Spark SQL as mentioned in lecture? (Choose 3)

- Efficient data manipulation using SQL like structure.
- **Enables relational queries on Spark.**
- **Deploy business intelligence tools over Spark.**
- **Connect to variety of databases.**
- Better ability to manipulate big data.
- Better worker node interpolation.

9. What is a triplet in GraphX?

- A type of data to contain vertex info.
- **A type of data to contain the information on connections between vertices and edges.**
- A type of data to contain both edge and vertex info.
- A type of data to contain edge info.