# Dare Olufunmilayo

# Hadoop & Spark Skills

Q1. Which programming languages give the best performance when using RDDs?

1. **SCALA, JAVA**

Q2. Why is it a problem that the result of a DataFrame operation is “Row”?

1. **This is because it loses compile time “type” safety. Unlike RDDs and Dataset, Dataframe returns a row of different things. The problem is when an engineer intends to do something more meaningful, he has to get back to the compiler as an instance of the row which is prone to error.**

Q3. Why are DataFrame operations in Python and Scala equally fast?

1. **This is because DataFrames support internal Optimization. The engineer need not to worry about how the operation is being carried out. DataFrames support Execution Plan like traditional RDBMs, hence makes non-JVM language like Python equally fast as Scala**

Q4. Your user has written some code in Scala using RDDs. He is joining a couple of large

data sets and then doing some filtering. He’s unhappy with the performance. What would

you advise him?

1. **Unlike in DataFrame and DataSet that support internal optimization, RDDs cannot create any optimal execution plan. Therefore, I will recommend for the user to consider filtering the data first before performing any "REDUCE" based operations such as JOINS**

Q5. Your user plans to ingest a large dataset with many columns covering several years of

data. Most of the queries will be aggregates of a few columns by calendar month. What

advice would you have for the ingestion engineers, with respect to partitioning and file

formats?

1. **Since the dataset is a large one, to benefit from partitions I will recommend for the engineer not to use a gzip file format as RDD supports only one partition. He/She will not be able to split up a gzip file and un-pack it.**

**On the other hand, assuming the data will be ingested into S3, I will suggest to the engineer to ensure that the files are ingested with proper key naming conventions as S3 automatically partition based upon key prefix across the nodes on the sub-system.**

**This is basically done to ensure consistent performance for reads of the objects. A typical example will be each file having a prefix representing each day in format "YYYYMMDD". This is not optimal for read performance in S3 because the partitioning will be based on the first index of the key value, in this case the first figure in "YYYY" which means 1 partition per entire year. Improvement here will be to make code changes to reverse the prefix order to suggest partitioning by "MM". This way the keys will be partitioned by months.**

Q6. Troubleshooting Issues You will be required to troubleshoot issues on the platform, for example latency issues observed in end-to- end data pipelines (ingest via Sqoop, transform, conform, persist). What would you expect to see in the platform (or around the platform) to make your life a bit easier? Assuming these things are there, how would you approach tackling the latency issue?

1. **Proactive Monitoring : I will expect to see proactive monitoring and alerting system for the data platform (monitoring databases, networks, error logs etc…). This will help react to issues quickly, effectively and long before they become problems. Such issues as API credentials expiry can be resolved before a timeout occurs.**

**With adequate monitoring of the data platform, I can understand the workload better and make better recommendations for performance tuning.**

1. **Capacity Planning (Storage and Processing): This should be an on-going task. Better planning for storage and computation based resources can reduce possibilities for issues occurring. I will keep track of the data platform growth through documentation and proactive monitoring.**
2. **Enthusiastic support engineers to provide support: Being enthusiastic to offer support,*( If required, even in the middle of the night )* and getting involved in the end-to-end deployment of data pipelines**
3. **Proper API documentation: Access to proper API documentations on platforms such as Confluence is a very good practice as this improves access to information required to provide support. I will endeavour to keep well written documents on personal experiences , deployed solutions and issues resolved for others to benefit from.**
4. **Adequate testing of proof of concepts and baselining pipe-lines before rolling into production**