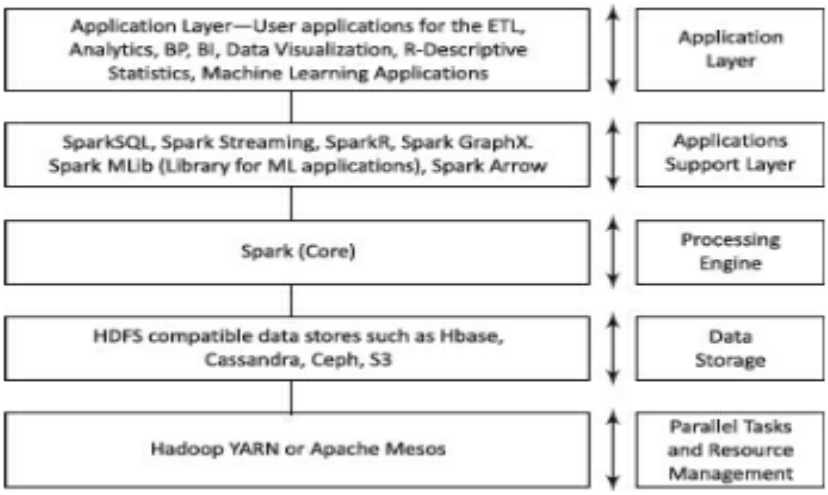
UE18CS322-Big Data- Unit 3

Question Bank and answers

1. **What do you mean by Acyclic Data? Does that mean repeatedly reading and writing?**
2. **Compare Spark and Hadoop in terms of speed, processing etc.**
3. **How does Spark process as a fast and general compute engine? What does expressive programming model mean?**

Execution engine uses both in-memory and on-disk computing. Intermediate results save in-memory and spill over to disk. Contains API to define Resilient Distributed Datasets (RDDs). RDD is a programming abstraction. RDD is the core concept in Spark framework. RDD represents a collection of Object Stores distributed across many compute nodes for parallel processing. Spark stores data in RDD on different partitions. A table has partitions into columns or rows. Similarly, an RDD can also be considered as a table in a database that can hold any type of data. RDDs are also fault tolerant.

4. **List the main components of Spark stack and the functions of each.**



5. **Consider a text file named jigsaw_puzzle_info.txt in /home/user directory. The three lines in the text file are:**

```
puzzle_code_Garden 10725 pieces 100 puzzle_cost_USD 1.35
puzzle_code_Jungle 31047 pieces 300 puzzle_cost_USD 2.85
puzzle_code_School 81049 pieces 800 puzzle_cost_Cents 90
```

(i) How will you create RDD puzzle_Costs from jigsaw_puzzle_info.txt? Use Spark Context
(sc).
(ii) A new RDD must have the lines having the string "puzzle_cost_USD". How will you use transform command to get new RDD textFile, puzzle_cost_USD? How many lines will puzzle_cost_USD possess?
(iii) How will an action command display first line from the filtered text?
Solution
(i) RDD creates from text file at Spark Core using the following command:
>>> puzzle_Costs = sc.textFile("jigsaw_puzzle_info.txt") [sc stands for SparkContext.] Alternatively, without using sc then create RDD using the following command: puzzle_Costs = spark.read.textFile("jigsaw_puzzle_info.txt").rdd (ii) A transformation command is filter(). The following statement does the transformation using filter(): >>>puzzle_cost_USD = puzzle_Costs.filter (pyspark line: "puzzle_cost_USD" in line)
puzzle_cost_USD RDD will have first two lines only. (Third line has the cost in cents.) (iii) An action command to get the first line is first(). The following statement does the action using first(): >>> puzzle_Costs_USD.first() The result puzzle_cost_USD first line will display as follows: ## puzzle_code_Garden 10725 pieces 100 puzzle_cost_USD 1.35

6. How iteration is done in Spark?
7. What type of Challenges in memory processing and is that got rectified in Spark?
8. How do you say the datastructure is distributed in Spark?
9. What is the meaning of RDD and how to create it?
10. Is split in map reduce and paritition in Spark is same?
11. Why does the spark create partitions?
12. What is lineage?
13. Does transformations are executed on demand?
14. Does actions are encountered in Spark following lineage graph?
15. Can you use Spark to access and analyse data stored in Cassandra databases?
16. What are the languages supported by Apache Spark for developing big data applications?
17. Explain about the different cluster managers in Apache Spark.

18. What is Executor Memory in a Spark application? Please explain.
19. What are the various data sources available in SparkSQL?
20. What do you understand by Pair RDD?
21. What is Dataframe and Dataset in Spark? When and why do we need to create Dataframe and Dataset in Spark? Please explain with some examples.

Note : Apart from these questions, do concentrate on given a dataset how can you create RDD and perform the various transformations and the same applies to dataframe as well. Do look into the extra spark questions, that would be fair enough for you to understand the various actions and transformations.

Go through the slides for scala and complexity lectures.