Part1- By Avinash Kumar

Duration – 180 minutes
Exam Pattern:

- **20 MCQ Questions(MCQ had only one correct choice)**
- **Coding Challenge- 19 Questions**

You can switch between MCs and coding challenges any time.

MCQs were medium and difficulty level, basic knowledge of spark architecture, execution and advanced knowledge of caching and memory management is required. Wide and narrow transformations.

MCQ will consist questions such as:

1. **Architecture questions**-High level designs of cluster given and you will be asked to choose the best cluster in respect to following scenarios. **Cluster experience most traffic, Cluster most prone to out-of-memory exception, Cluster most likely to complete task in case of job Failure** etc.
2. **Configuration questions**- Default **spark.sql.shuffle.partitions**(200), Default **spark.executor.memory, Spark node limitation**(128) etc.
3. Choose the narrow or wide transformation from the given List.
4. Choose the action or transformation from the given List.
5. **Accumulators**- True and False questions Accumulators will not apply catalyst optimization, Accumulators default type etc.
6. Difference between Partition and Coalesce.
7. Minimum 3 questions on **Stages, Job, Task, partition** such as which is the lowest level among the 4 ,State the hierarchy, At what level physical plan is made etc.
8. Correct syntax to register a UDF.
9. True and False w.r.t Physical plan, and transformation

Coding Challenges:

1. **String Manipulation**- expect at least 2 questions e.g.:- From the given Dataframe of String we need to remove a particular String from each row then we need to break the sentence into words in separate rows then count the distinct words.

2. Remove duplicates then OrderBy a particular dataframe on 1st column, nulls first then 2nd column .Note this question actually uses asc_nulls_first which is rarely used).

3. Given 2 dataframe need to create a third dataframe which contains common of the given 2 dataframe.

4. Need to have proper understanding of limit(),first(), take() functions as they are present in most transformation.

5. **Date manipulation**- Question on Date Manipulation will surely be there need to understand to_date(), from_unixtime(),

from_utc_timestamp(), unix_timestamp() .I had a question where timestamp was given in long format and need to create 3 columns Date in given format yyyy-MM-dd, Year, Month(Jan,Feb etc).
2nd question timestamp was given in long format need to get the given format.

6. Questions included persist a dataframe to DISK_ONLY (Ans: df.persist(StorageLevel.DISK_ONLY)).

7. Print the default partition(Ans:- spark.sparkContext.defaultParallelism)

8. Question on write a Dataframe into given number of partition.

9. Question on BucketBy on a particular column with given number also the path.

10. Question on read a given csv by defining your schema in DDL format and then giving various read options.

11. Question on creating a json nested schema and validating.

12. Sample a given dataframe with replacement and given fraction. Split it into train and test.

13. Question on Groupby and pivot is important.

14. One Join question with instruction as only one column on which the join condition is applied should come.

15. Create Dataframe using the giving range. This can have 2 approach **spark.createDataFrame(range(20),IntegerType())**this will create a value column.
spark.range(your_range) this will create an id column of your range.

16. Question on explode and create a particular dataframe.

17. Window function usage should be clear as there are questions on rank, lead, lag were there however they were not explicitly mentioned.

18. One Udf question where you need to register the given function as UDF and perform some transformation.

Take Aways:

1. Time Management is the Key, though 180 minutes sounds much time .I am sure you will miss questions if you are not careful.

2. Don't spend too much time on MCQs, either you know them or not, try to complete them, I targeted 20 minutes. Save time for the coding challenges as you might get stuck in one of them.

3. MCQ's comprise just 10% of your total score Coding challenges will Comprise 90% of your score so strictly do not waste time searching for correct answer in docs.

4. Start with the easiest or the most comfortable CC first, as it gives you confidence to move on to difficult ones. They are not marked so keep navigating and use pen and paper to mark which one you want to try again.

5. If you are stuck on a CC, please move on and come back to it later if time permits.

6. Each coding challenge is in it's own notebook. The first few cells outline the problem statement. Then there is a cell that includes a setup script for the individual challenge. Depending on the challenge, it will initialize spark. Basically you will be given some code and stubbed method to fill in. There are comments prompting you along.

7. There is also a cell to test your method(s) on a provided sample set but the sample set won't be visible to you and you need to debug on your own why a particular test case failed.

8. Memorization of dataframe/sql API is not needed. But you need to know where to find the available functions. To make sure this please go through different functions and when they can be applied here:
   https://spark.apache.org/docs/latest/api/python/pyspark.sql.html
   https://docs.databricks.com/

9. If you are not comfortable with dataframe you can switch to spark sql, but most of the questions required to have dataframe API knowledge.

10. Having a small experience in databricks community edition will help you feel comfortable.

11. Make sure after completing each coding questions you go to clear tab and click **clear State & Run All**. It will be given in instruction page too if you don’t do this your code might fail to compile despite having the correct answer.

12. The exam will be conducted by ProctorU they are databricks partner. There are no center option available for this certification and you need to take it personally.