Technical Test for Prospective Employees

Instructions:

- **You will have 2 hours to complete the following technical test.**
- **There are two sections, an applied section and a general questions section.**
- **For the applied section, please provide in this word document the output as well as the code used to produce the output. Code should be written in PySpark, Scala or Java (preferred) but if python is also acceptable.**
- **If you cannot replicate what is asked of you, please put into words how best you would achieve the given task.**

Applied Section:

1. Generate the following data frame*:

Input 1	Input 2	Input 3	Description
Green	1	1.3	DescriptionONE
Green	0	1.445	DescriptionONE
Blue	4	1.2	DescriptionTWO
Red	5	1.3	Description THREE
Yellow	7	1.325	Description Four
Red	9	1.4	DescriptionONE
Red	6	1.72158	Description THREE
Blue	5	1	DescriptionONE

*Where all data is in string format.

2. Remove the space in the string for 'Description THREE' and 'Description Four'. Also, capitalize the letters in 'Four' for 'Description Four'.
3. Change all of the numbers in 'Input 3' to show four decimal places.
4. Generate another data frame as follows*:

Green	Night
Yellow	Morning
Red	Afternoon
Blue	Evening

*Where all data is in string format.

5. Add column headers to this data frame where column 1 is called 'Input 1' and column 2 is called 'Day Period'.

Technical Test for Prospective Employees

6. Left join the table from 4 to the table in 1.

7. Add a column to 6 that generates random dates in 'YYYYMMDD' format. Call this column 'Date'.

8. Filter from this table all values that are less than '1.31' in 'Input 3' and not equal to 'Red' or 'Green' in 'Input 1'.

9. Create a flag that identifies all records that are greater than the middle date amongst the dates generated and that are also greater than 1 in 'Input 2'.

10. Create a for loop that will sum all values in 'Input 3' by 'Description' and divide by the minimum value in 'Input 2', then put them in separate data frames.

11. Create a function that will carry out 5-10. Add exception handling.

12. Create a basic spark-submit function that will call 11. (Path statements can be local machine)

General questions:

What is an RDD, is it different from a data frame?

Why do we use spark?

What is the difference between spark.sql and dataframe operations? Which should you use when?

Assuming there is a sqldb table named database.table and I wanted to delete particular records how would I do that? How would I change the values in columnA in the same database.table from 'a' to 'b'?

What is sparkmagic, where can I use it and why would I use it?

What is modularization of code? Give an instance of where you would use it? What are the pros and cons?

What is logging? Why is it important? Give an example of a situation where logging would be useful.

What is spark context? How do you start one? Should you put a sc.stop() at the end of all scripts?

What is a JAR file? When and why do I need it?