DATA ENGINEER – CODING CHALLENGE

Spark Assignment – (20 or 25 mins MAX. ~~This should be cake walk for 4 years of experience~~)

Big Plus:(Spark-Scala is used by 80% of Big Data Developers, and 10 times faster than PySpark)

Do you need Help? Access this. https://spark.apache.org/docs/latest/index.html. (~~Both Python and Spark-scala is here~~)

Follow these Instructions for the best performance in Big Data analysis using Spark-Shell: (~~For Performance, one Should do all of these in Spark DATAFRAME.~~)

- Your Import or Export should be only in S3. Import the Datasets (Products and Categories) to Spark-Shell from <s3 location>. ~~(2 files -AVRO/JSON/PARQUET should not be specified - S3 location)~~.

1. Save the first 20 rows of each dataset in uncompressed parquet file as Result_0.parquet and Result_1.parquet.(~~Snappy compression is default so need change to uncompressed~~)
2. Save only the product ids, names and prices having less than 100 USD as Result_2.csv. (~~Confirms that he/she is good in Filtering the data~~)
3. On the above filtered data, save the top 10 product prices in each category with the following fields - Category_name and Product_price in Results_3.txt as tab delimited. (~~Expert in Join Transformation but professional should give extra care for the question as it is filtered data and tab delimited~~)
4. On the given datasets, find the highest and lowest product price in each category. There should be 2 fields Category_name and Product_details (Product_details should have Highest_product_name, Highest_product_price, Lowest_product_name and Lowest_product_price with pipe delimited). Save the result in default snappy Result_4.parquet format. (~~Proves that you are expert in Aggregation/Calculation/GroupBy/Join and ETL job, note the snappy parquet should be set again since we have changed in step-1~~)