Marketing Questions

1. **Convert Click Stream data to sessions**
    a. Below table is the click stream table structure capturing all events from user interaction from columns A to G
        i. Device_id => unique identifier for user device
        ii. Visit date => Date when the user visited the platform
        iii. Visit time => Time when the user visited the site
        iv. Activity kind => Type of activity. Eg. External Ad click, install event, page_view
        v. OS => OS name
        vi. Venture => country code of the platform
    b. We need to create column H “Session_id”
        i. Each session is live for a time window of 60 minutes, if the user comes back after 60 minutes new session is created
        ii. The logic of session id naming convention is
            1. concatenate 's'+device_id + '_' + visit_date + '_' + first visit time of that session
    c. Please feel free to use any technology of your expertise to create the logic of session id creation and share us the code
    d. Caution: The data size is million rows per day and goes to billions on campaign days

	A	B	C	D	E	F	G
1	device_id	visit_date	visit_time	activity_kind	os	venture	session_id
2	1b	20181019	10:00 am	click	ios	SG	s1b_20181019_1000
3	1b	20181019	10:01 am	install	ios	SG	s1b_20181019_1000
4	1c	20181019	10:02 am	page_view	android	SG	s1c_20181019_1002
5	1c	20181019	10:03 am	order	android	SG	s1c_20181019_1002
6	1b	20181019	11:00 am	page_view	ios	SG	s1b_20181019_1100
7	1b	20181019	11:10 am	order	ios	SG	s1b_20181019_1100
8	1c	20181019	12:00 pm	page_view	andorid	SG	s1c_20181019_1200
9							

1. **Data Modelling - From relational to BigData**
    a. Below sales table gets an entry once an order is placed by a user and stored in a mysql database
    For analytics we want this data to be available for data engineering and data science team in hive.
    b. **MySQl table columns**
        i. created_date - datetime when the order was created
        ii. verfied_date - datetime when the order was verified, it might take upto X days for an order to be verified due to cash on delivery, over the counter payment methods, etc.,
        iii. order_number - identifier for the order
        iv. country - country code of the platform from which order was placed
        v. user - user_id
        vi. product - product_id which was purchased
        vii. status - order status which will be updated periodically

c. Can you create the DDL of the table in hive and share the data insertion script as well.
   Assume whole table data dump from mysql is available as a csv file in HDFS every day in location '/data/sales/yyyymmdd/data.csv'

d. Things to keep in mind
   i. The data size is millions of record per day.
   ii. Schema definition and insertion mechanism should be less resource consuming. Also compensating time delay in order verification date and status.
   iii. The table structure should be ideal for data engineers to query as they will be querying for many days
   iv. Order statuses
      1. When user places an order =>
         a. created date column => created date is filled with current time.
         b. verified date column => empty
            (Order is not yet verified (pending payment, anti-fraud checks))
         c. status is set as “created”
      2. When order is verified
         a. created date column => no change to this column
         b. verified date column => filled with current time
            (Order has passed payment & anti-fraud checks)
         c. status is set as “verified”
      3. When order is delivered
         a. created date column => no change to this column
         b. verified date column => no change to this column
         c. status is set as “delivered”
      4. When order is returned
         a. created date column => no change to this column
         b. verified date column => no change to this column
         c. status is set as “returned”

	A	B	C	D	E	F	G
1	created date	verified date	order number	country	user	product	status
2	20181001 10:00:00	2018-10-19 11:00:00	1	SG	a	p1	delivered
3	2018-10-19 12:00:00		2	SG	b	p2	returned
4							
5							
6							