Docker

Below are the sql scripts that will be used to initialize the mysql instance:


a_schema.sql


b_data.sql

1. **Script 1** - Create a docker script to start an existing mysql docker instance.

- Below is Script 1


Script1.sh

- Please open the script with an editor of your choice the details for these scripts are inside as comments.

2. **Script 2** - Create a script to import the orders details information into the mysql database.

- Below is Script 2


Script2.sh

- Please open the script with an editor of your choice the details for these scripts are inside as comments.

3. **Script 3** - Write a Scala or Python Spark script that would do the following:

a. Predict the next order date for a customer, based on his purchase patterns.

b. Only customers predicted to buying products in the next week(7 days) must be placed in the mongo database.

I am unclear on this question in terms of making predictions.

4. Explain under which conditions you would use MySQL and Mongo?

In MySQL, the data value is stored in the tables by the MySQL database structure where SQL is used to access them. Schema is used to define the database structure. The prime requirement of the schema is that the rows have the same structure inside the table. It also requires its values to be represented by specific data types. Whereas, in the MongoDB database, the data is stored in JSON-like documents that come in varied structures. To boost up the query speed, it stores the related data sets together. These sets are then accessed by using the MongoDB query language.

The database is schema-free, which means that mobile app developers do not have to define any document structures for creating the documents.

Both the databases MySQL and MongoDB uses indexes for the task of searching data. However, the difference comes in the approach when an index is not defined or found.
When the index is not defined in the case of MySQL index optimization, the database engines scan the entire table to find relevant rows.
Whereas in MongoDB, when the index is not found then each single document in the collection is scanned so that the document offering a match to the query statement could be selected.

MongoDB is known for better controlling large volumes of unrestricted data as compared to that of MySQL. MySQL is considerably slower as compared to MongoDB when you use large volumes of data. MySQL struggles to deal with the high volumes of unstructured data. It is because the data is spread across multiple tables that need to be accessed for writing and reading the data.
When we compare MongoDB with MySQL on the grounds of performance, MongoDB seems to have an advantage.
MySQL uses a rigid table structure model which makes it slower for users to create an application in it.

Databases for MySQL support master-slave replication and master-master replication. With this multi-source replication, you can replicate multiple masters in parallel.
On the other hand, MongoDB supports auto-elections, built-in replication, and sharding. Developers can use auto-elections to set a secondary database that will automatically overtake on the failure of the primary database. Whereas sharding allows horizontal scaling which is considered difficult to implement with MySQL.

MongoDB, it provides enterprise-grade support. This support extends beyond the fix/break model. It provides you with an extended lifecycle support add-on along with round the clock support. This gives the user flexibility to upgrade to a newer version at their own pace rather than with MySQL.

MongoDb can be used when high data availability is your priority along with automatic, fast, and instant data recovery.
If you are working with an unstable schema and need to lower the cost of schema migration.

5. Explain your choice in context of the CAP Theorem?

In context of the CAP Theorm, my choice in-case of Partition taking place I would rather have (AP) Consistency & make trade-off of Availability rather than have Availability & not have Consistency.

I choose this because it takes more effort to fix data that is not consistent across multiple nodes in a distributed system and even though my choice may mean maybe financial revenue declines due to not having availability when a partition takes place it is the safest in terms of not duplicating data and dev maintenance costs & wasting time whilst creating an endless spiral of complications.

Black Friday
On the day of Black-Friday you realize that you have a large number of orders, explain how you would make changes to your data engineering architecture to be more robust, scalable, reliable and real time.

1. What technologies would you use?

I would make use of the following technologies:

- Hadoop File System (HDFS)
- SPARK SQL
- HIVE
- Hadoop cluster
- MapReduce
- Basically, I would go the Bigdata route.

2. Why would you choose these technologies?

I would choose these technologies because they have the ability to handle structured & structured data including processing any type of data extremely fast with less effort and fault tolerance in terms of a pre-defined environment where Data Storage, Data Mining, Data Analytics & Data Visualization, is already catered for with potential to use more tools that this platform has to offer for data i.e. In-memory Databases, Data Lakes & Blockchain etc.

3. What data patterns would you use?

I am unclear on this question. N/A

Order Analysis

Please provide scripts and results for the following:

1. Which day of the week has the most orders?

- Thursday

2. Which time of the day to people order the most?

- Between 17:00 & 20:00

3. Which order does the user buy first most of the time?

- Macbook Pro Laptop & Thinkpad T365 Laptop are the products that the user buys first most of the time.

4. What is the time interval that a user tends to purchase again?

- 2-3 months

5. Write a mysql script on how to delete the duplicate orders, of the latest date, please explain the script in detail?

- Below is the sql script (DeleteDuplicates.sql)

DeleteDuplicates.sql

- Please open the script with an editor of your choice the details for this script are inside as comments.