Data Engineer – Technical Challenge

This is a technical challenge for SIT Data Engineering team. Please send your results from the first 3 questions as soon as possible to this mail account: holger.koch@mail.schwarz. For the fourth question, you can use more time and send us the answer later if you preferer. You can write your answers on paper and send us back photos or type them directly here. You can also develop with your favorite IDE or Notebook and send us the code files by mail or a repository link.

1. Coding Challenge

You can write code in your preferred language (Python, Scala, Java or R) Write a function reverse that:

1. Given an array of integers, find the two positions/indices that sum a specific value
    a. To consider:
        - There is always a solution. There is only one
        - Negative values posible
        - Not previously sorted
        - It fits in memory
2. Print all possible solutions if there more than one.
3. Describe a solution for the previous case when the data does not fit in memory.

Solution:

1&2.

```
public class CheckArrayPosition {

    public static String hasArray(
            int[] arr, int sum) {
        String positions = "";

        for (int i = 0; i < arr.length; i++) {
            for (int j = i + 1; j < arr.length; j++) {
                if (arr[i] + arr[j] == sum) {
                    if (positions.equals("")) {
                        positions = i + " and " + j;
                    } else {
                        positions = positions + ", " + i + " and " + j;
                    }
                }
            }
        }

        return positions;
    }

    public static void main(String[] args) {
        int[] arr = {1, 4, 45, 6, 10, -8, 8, 8};
        int n = 16;

        if (hasArray(arr, n).equals("")) {
            System.out.println("No positions");
        } else {
            System.out.println("The positions are: " + hasArray(arr, n));
        }
    }
}
```

3. Describe a solution for the previous case when the data does not fit in memory.

- Increase the Java memory

2. Spark Challenge

Exercice overview Next exercise is about coding a simple ETL process using Spark. This exercise help us to check out your Spark level at the same time we analyze your coding style. Feel free to use any tool for develop (Notebook, IDE, paper…). You can use the Spark SDK of your choice (preferably Spark 2+) or any other distributed framework like Hadoop/MapReduce, Hive, Pig…

Exercice goal Attach to this document you'll find a "events.csv" file containing users actions. Each action has a timestamp and a possible value, either "open" or "close". We would like you to reduce data temporal granularity to 10 minutes, so that there is only one single row for each 10 minutes. Over this temporal aggregation count how many actions of each type there is per minute. After previous calculation, please compute the average number of actions each 10 minutes. Finally we would like you to compute the top 10 minutes with a bigger amount of "open" action.

Can you do a proposal about how to test this job with unit test, how to test a full pipeline with a integration test and how to release this job on production with data quality check?

3. SQL Challenge

You can write the SQL query or the code necessary to produce the required results.

IMPRESSIONS

Product_id	click	date
1002313003	true	2018-07-10
1002313002	false	2018-07-10
…	….	…

PRODUCTS

Product_id	category_id	price
1002313003	1	10
1002313002	2	15
…	….	…

PURCHASES

Product_id	user_id	date
1002313003	1003431	2018-07-10
1002313002	1003432	2018-07-11
…	….	

1. Given an IMPRESSIONS table with product_id, click (an indicator that the product was clicked), and date, write a query that will tell you the click-through-rate of each product by month
2. Given the above tables write a query that depict the top 3 performing categories in terms of click through rate.
3. Click-through-rate by price tier (0-5, 5-10, 10-15, >15)

Solution:

```
create table PRODUCTS (
  product_id bigint,
  category_id bigint,
  price bigint,
  primary key (product_id)
 );

create table PURCHASES (
  product_id bigint,
  user_id bigint,
  date date
 );

 create table IMPRESSIONS (
  product_id bigint,
  click bigint,
  date date
 );
```

1.

```
SELECT product_id, month(date) as month, COUNT(click) as clickCount
FROM IMPRESSIONS where click = true
GROUP BY product_id, month(date)
```

2.

```
SELECT b.category_id, COUNT(click) as clickCount
FROM IMPRESSIONS a, PRODUCTS b where a.product_id = b.product_id and a.click = true
GROUP BY b.category_id order by clickCount desc limit 3
```

3.

```
SELECT COUNT(click) as clickCount
FROM IMPRESSIONS a, PRODUCTS b where a.product_id = b.product_id and a.click = true
and b.price > 15
```

4. Data Architecture Challenge

We are managing parking lots that a client can check with a mobile app. An app can tell the drive if a parking is full or not. At the moment on entering/leaving the parking a client can scan QR/NFC code on entrance machines and the cost has to be automatically charged when leaving parking. We are interested in monitoring what time the parking is full and what time is not in order to modify prices accordingly. We also would like to create a predictive model that learns what time a client is going to the parking and in order to send him a push message informing how many places are left or if parking is full.

1. What tracking events would you propose? What data model for event analysis? What technologies?
2. How would you design the Backend system? What data model for the Operational system? What technologies?
3. Explain how to combine the operational architecture with the analytical one?
4. Could you propose a process to manage the development lifecycle? And the test and deployment automation?

Solution:

1. **What tracking events would you propose? What data model for event analysis? What technologies?**

The propose traking events:

- A regular client who has purchased a biweekly, monthly, or yearly pass.
- A prepaid client who booked a slot remotely (on the phone or online).
- A walk-in client who neither has a pass nor booked a slot remotely. A slot will be assigned to such a customer based on availability
- Vehicle entering the parking lot
- Vehicle leaving the parking lot

Data model for event analysis:

- Statistical or Business data models for event analysis.

Technologies:

- MySQL, Hibernate, Java, Spring Boot
- Java Event Handlers will be used

2. **How would you design the Backend system? What data model for the Operational system? What technologies?**

Design the backend system that are given below:

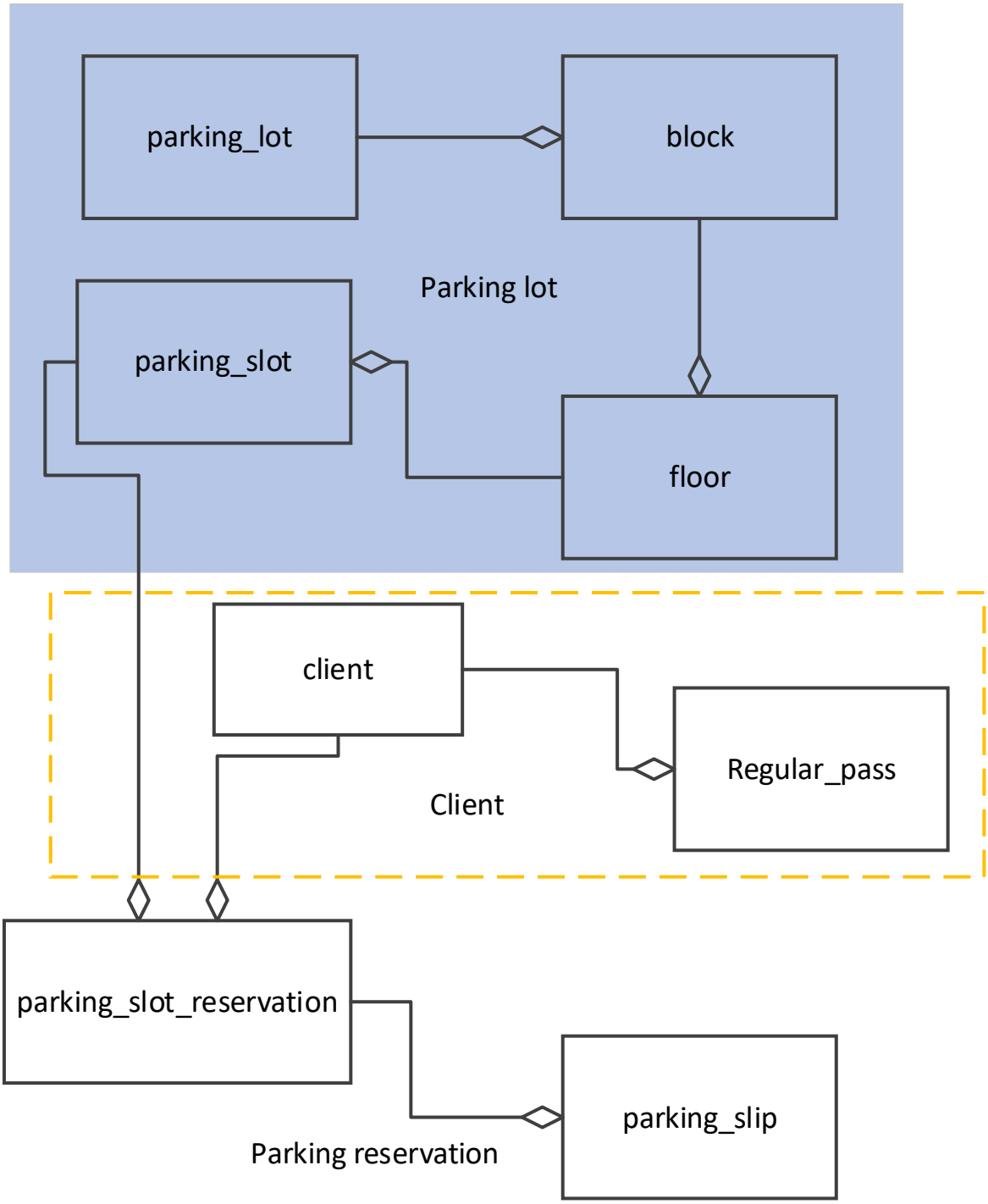

Figure 4.1: Backend System (Data model)

Data model for the Operational system:

- Parking lot
- Customer
- Parking reservation

Technologies:

- MySQL, Hibernate, Java, Spring Boot
- Backend Services With Java and Spring Boot
- Database with MySql

3. Explain how to combine the operational architecture with the analytical one?

- Waterfall development or iterative development.
- Operational architecture is the mobile app and the events which will update the status of the parking, full or empty to occuply the vehicle.
- Analytical one, is to analyse how and what are the timings the parking will be full or fee for parking.

4.Could you propose a process to manage the development lifecycle? And the test and deployment automation?

- Yes, I could propose to manage the software development lifecycle (SDLC). I would like to waterfall methodology for software development.
- Development Methodology
  Iterative Development
  Features are prioritized and developed in iterations

- Continuous Integration
  - Build
  - Test
  - Merge
  - Continuous Delivery
  - Continuous Deployment
- Automated Deployment
- Using PipeLine Scripts from the Project Repository which has the source code