Ultra Tendency GmbH
August-Bebel-Str. 46
39326 Colbitz
DE271390765 (USt-IdNr.)

ULTRA TENDENCY
Pioneering the Future

Ultra Tendency Data Engineering Challenge
Practical Assignment

This assignment sheet consists of three tasks. We recommend downloading Cloudera Quickstart VM to complete the assignment as it contains all components (HDFS, Spark Streaming, Yarn, HBase, Impala) required. Please do your coding in Scala or Java.

Task #1: Data Ingestion

Write a simulator for IoT devices. The device will generate data according to the following template

```
{
        “data”: {
                “deviceId”: “11c1310e-c0c2-461b-a4eb-f6bf8da2d23c“,
                “temperature”: 12,
                “location”: {
                        “latitude”: “52.14691120000001”,
                        “longitude”: “11.658838699999933”
                },
                “time”: “1509793231”
        }
}
```

The properties are defined in the following way

Property name	Data type	Comment
deviceId	UUID	The unique ID of the device sending the data.
temperature	Integer	The temperature measured by the device.
latitude	Long	The latitude of the position of the device.
longitude	Long	The longitude of the position of the device.
time	Timestamp	The time of the signal as a Unix timestamp.

The simulator should simulate three different devices and needs to send a signal to Apache Kafka every second. Implement the simulator as a long running service and create Kafka topic(s) as needed.

ULTRA TENDENCY
Pioneering the Future

Ultra Tendency GmbH
August-Bebel-Str. 46
39326 Colbitz
DE271390765 (USt-IdNr.)

Task #2: Data transformation

The data ingested in task #1 needs to be transformed into an entity and stored for long-term analysis. Create a Spark Streaming job, which reads the data from Kafka and stores it into an HBase table. The HBase table needs to store a data point into a single row. A row should contain the data in its raw format as well as the entity representation of the data point. Note that the timestamp needs to be converted into human-readable format using the pattern “yyyy-MM-dd'T'HH:mm:ssXXX”. Design and create the HBase table and implement the Spark Streaming job.

Task #3: Data analysis

After the data is stored in HBase it needs to be analyzed. The data analyst responsible for the analysis prefers to use SQL as the query language. Create an Impala table on top of the HBase table created in task #2. Also, document queries for the following use cases:

1. The maximum temperatures measured for every device.
2. The amount of data points aggregated for every device.
3. The highest temperature measured on a given day for every device.

Submit you project (without binaries) and a few screenshots showing you solution in practice.