Hadoop Developer Test

Problem

There's a large retail store, which keeps their products in a relational database. They have a point of sales solution (cash register) that tracks all the sales. The transaction logs of the point of sales (pos) systems are stored on Hadoop. Due to the size of the transaction logs it's not possible to put them in the relational database.

The store wants to run a report on this data to feed their loyalty system. That requires a report on user id containing the most popular product category of the user and their revenue per quarter.

Relational Database

The product meta data resides in a relational database (e.g. MySQL, Postgres, etc.). It can be reached using a JDBC driver.

```
CREATE TABLE products (
  id INT(32) PRIMARY KEY,
  name VARCHAR(255),
  category VARCHAR(255),
  price DOUBLE
);

INSERT INTO products VALUES
(1, 'Banana', 'Fruit', 1.00),
(2, 'Apple', 'Fruit', 0.85),
(3, 'Raspberry', 'Fruit', 1.65),
(4, 'Chocolate sprinkles', 'Decoration', 1.00),
.
.
(9998, 'Windscreen wipers', 'Car accessories', 33.00),
(9999, 'Rear windscreen wipers', 'Car accessories', 35.00);
```

The database contains a maximum of 99999 products. This is a limitation of the pos system.

Log files

The logs files are available in `þ-separated` files per day per pos unit. The files are all in 1 directory on HDFS. And look like: U0001-2013-01-01.log.gz
The files are gzip compressed.

```
DateþOrderIDþProductIDþUserIDþAccMngrIDþQuantity
```

```
2013-01-01þ10000þ1þ1þ1þ10
2013-01-01þ10000þ2þ1þ1þ5
2013-01-01þ10001þ1þ2þ1þ1
.
.
.
2013-12-31þ99992þ9999þ666þ1þ10
```

The files contain a header line and the format is consistent over the files.

Output

The report outputted should contain the user id, most popular product category and the revenue for Q1..Q4.

```
User ID    Most Pop. Prod. Cat.  Revenue Q1 Q2    Q3    Q4
1          Fruits                12.00      20.00 -     -
2          Fruits                8.50       -     -     -
.
.
666        Car accessories       -          -     -     35.00
```

Assignment

Write a process that can process the input data and meta data and generates the required output report. Due to the volume the processing of the log files should be done on hadoop.