实验二：使用 Flume、Thrift、Kafka、HBase 进行数据收集、传输与存储

实验目的：

（1）掌握 Flume、Thrift、Kafka、HBase 的安装；
（2）掌握 Flume、Thrift、Kafka、HBase 的使用。

实验平台：

操作系统：Ubuntu 18.04 LTS

实验内容：

1、 下载相关软件。

Flume 版本：apache-flume-1.9.0
Thrift 版本：thrift-0.13.0
Kafka 版本：kafka-2.13-2.6.1
HBase 版本：apache-hbase-2.3.4

2、 启动一个 Flume 服务器进行数据收集。本次实验要求 Flume 服务器以 HTTP Source 作为收集端。

本次收集的数据有两种类型，分别是“Order Detail”和“Transaction”。两者都为一个 JSON 格式的字典数据。“Order Detail”数据的包括 consumerId，itemId，itemCategory，amount，money 字段。“Transaction”数据的键包括 createTime，paymentTime，deliveryTime，completeTime 字段。

本次实验提供一个具有数据生成功能的服务端，会自动向 Flume 端口通过 HTTP 发送数据。在发送给 Flume 时，报文头部字段中包含了键“key”，值为“<rowkey>-<Order Detail/Transaction>”，在对接 KafkaSink 时,Flume 会自动将 HTTP 的报文中 headers 包含的 key 键的值作为 Kafka 数据三元组中<key>字段的值。报文内容字段为收集到的数据。以下是服务端生成数据的几个示例：

① 头部字段：

```
‘header’:{
    …
    ‘key’: ‘000001-Order Detail’,
}
```

内容字段：

```
‘data’:{
    ‘consumerId’: ‘41341’,
    ‘itemId’: ‘1057499’,
    ‘itemCategory’: ‘2’,
    ‘amout’: ‘1’,
    ‘money’: ‘462.8’,
}
```

② 头部字段：

```
‘header’:{
    …
    ‘key’: ‘000001-Transaction’,
}
```

内容字段：

```
‘data’:{
    ‘createTime’: ‘2020-4-16 9:21:09’,
    ‘paymentTime’: ‘2020-4-16 10:14:47’,
}
```

③ 头部字段：

```
‘header’:{
    …
    ‘key’: ‘000001-Transaction’,
}
```

内容：

```
‘data’:{
    ‘completeTime’: ‘2020-7-28 12:10:40’,
}
```

注意：对于“Order Detail”类型的数据，一定包含了所有的键值对；而对于“Transaction”类型的数据，有可能不包含所有的键值对。

3、 以伪分布式方式启动的三个 Kafka 服务器进行数据传输，broker.id 分别设置为 0，1，2。创建一个名为“Kafka_Orders”的 Topic 并使用第一个服务器（broker.id=0）接收来自 Flume 的数据。

4、 以伪分布式方式启动 HBase，并创建一个名为“HBase_Orders”的表，“HBase_Orders”表包括两个列族，“Order Detail”与“Transaction”，其中：

- “Order Detail”列族包含 5 个“consumerId”，“itemId”，“itemCategory”，“amount”，“money”列标识符；
- “Transaction”列族包含 4 个“createTime”，“paymentTime”，“deliveryTime”，“completeTime”列标识符。

5、编写 HBaseConsumer，与第三个 Kafka 服务器（broker_id=2）进行连接，读取 Topic 为 “Kafka_Orders”的数据，并写入 HBase 的“HBase_Orders”表中。

6、(*选做内容) 使用 Java API 创建本次实验中需要使用到的“HBase_Orders”表，并定时查看“HBase_Orders”表内的数据量，完成最简单的统计。

7、(*选做内容)使用 Thrift，并自定义 Flume 的 HTTP Source 中 Handler 组件，将 Flume 接收到的数据序列化，使得数据在 Flume 与 Kafka 中以字节流的形式进行传输。在 HBaseConsumer 中，对字节流进行反序列化，读取对象的内容并写入 HBase 中。

8、完成实验报告。

注：本次实验成绩选做内容各占 10%（共 20%），其他部分占比 80%