

作业二实验报告

团队成员分工：

- 祝溢泽218352001：可视化展示
- 宣伟康218352002：环境搭建、代码编写
- 陈绘新218352003：数据接收
- 顾城218352004：环境搭建、代码编写

实验流程

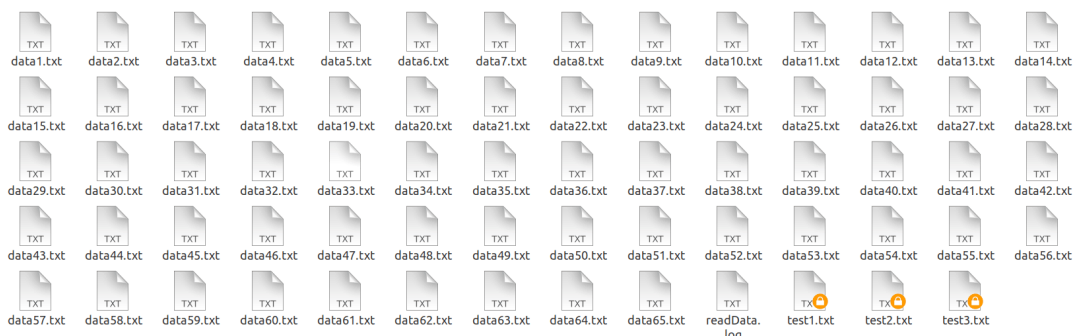
数据获取

- 静态数据获取

```
val reader = session.read.format("jdbc")
    .option("url", "jdbc:hive2://172.29.4.17:10000/default")
    .option("user", "student")
    .option("password", "nju2022")
    .option("driver", "org.apache.hive.jdbc.HiveDriver")
val registerHiveDqldialect = new RegisterHiveSqlDialect()
registerHiveDqldialect.register()
```

- 动态数据获取

- 搭建消费者获取流数据，暂存到txt中



- 在分析数据时，再创建生产者，将这些数据通过生产者发送

实验环境的搭建

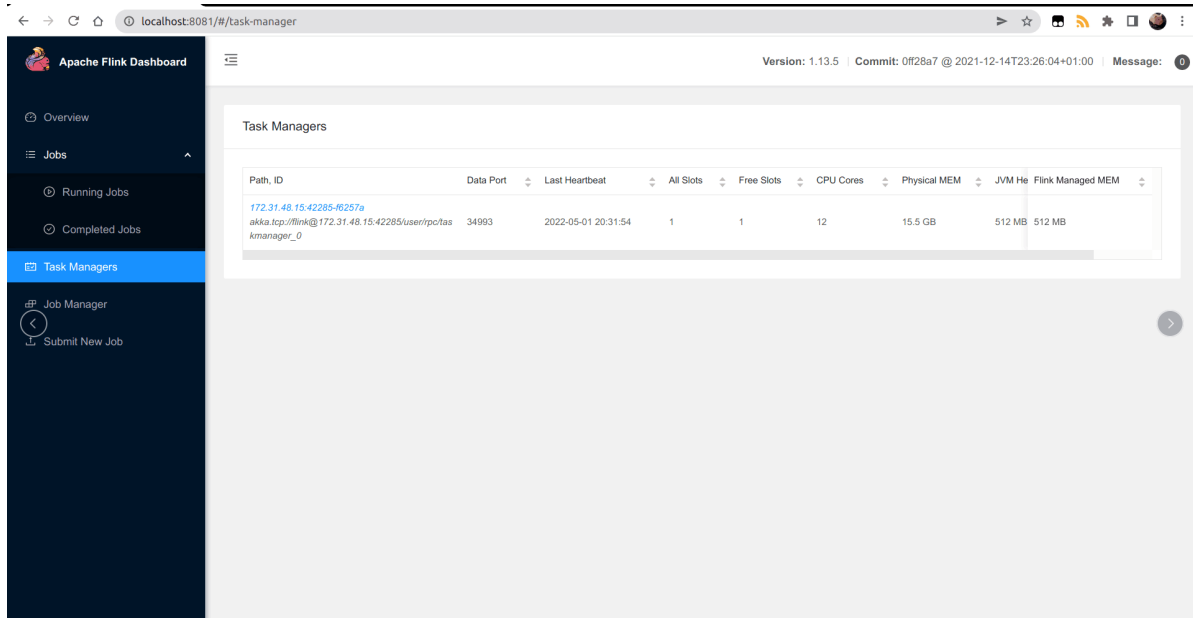
- 搭建Hadoop、Spark

```
jackson@jackson-Lenovo-XiaoXinPro-13API-2019:/usr/local/spark/sbin$ jps
12209 ResourceManager
12052 SecondaryNameNode
13077 Jps
11653 NameNode
10007 NailgunRunner
11831 DataNode
12824 Master
10041 RemoteMavenServer36
13002 Worker
12527 NodeManager
9455 Main
```

- 创建Kafka

```
etl
mytopic
testAll
transaction
====创建kafka主题命令:====
/usr/local/kafka/bin/kafka-topics.sh --create --zookeeper 172.31.48.15:2181 --replication-factor 1 --partitions 1 --topic 主题名
====等待3秒...====
====Jps====
18817 Main
323142 Worker
322958 Master
322542 ResourceManager
31148 QuorumPeerMain
31149 Kafka
57614 Launcher
45331 StandaloneSessionClusterEndpoint
306608 org.eclipse.equinox.launcher_1.6.300.v20210813-1054.jar
45622 TaskManagerRunner
322879 NodeManager
57661 MyFlinkSinkClickhouse
```

- Flink的搭建



数据转换

```
SingleOutputStreamOperator<Object> dataStream = source.map((value) -> {
    if (value == null) {
        return null;
    } else {
        JSONObject jsonObject = JSONObject.parseObject(value);
        String eventType = jsonObject.getString("eventType");
        if (eventType == null) {
            return null;
        }
    }
})
```

根据不同的eventType转化为不同的java对象(部分举例)

```
if ("sa".equals(eventType)) {
    return jsonObject.getJSONObject("eventBody", dm_v_tr_sa_mx.class);
} else if ("shop".equals(eventType)){
    return jsonObject.getJSONObject("eventBody", dm_hlw_shop_info.class);
} else if ("djk_info".equals(eventType)){
    return jsonObject.getJSONObject("eventBody", dm_v_as_djk_info.class);
} else if ("djkfq".equals(eventType)){
    return jsonObject.getJSONObject("eventBody", dm_v_as_djkfq_info.class);
}
```

数据存储

- 数据写入clickhouse

```
StreamExecutionEnvironment env =
StreamExecutionEnvironment.getExecutionEnvironment();
    String topic = "etl";
    Properties props = new Properties();
    props.setProperty("bootstrap.servers", "172.31.48.15:9092");
    props.setProperty("group.id", "consumer-group");
    props.setProperty("key.deserializer",
"org.apache.kafka.common.serialization.StringDeserializer");
    props.setProperty("value.deserializer",
"org.apache.kafka.common.serialization.StringDeserializer");
    FlinkKafkaConsumer010<String> consumer = new FlinkKafkaConsumer010(topic,
new SimpleStringSchema(), props);
    consumer.setStartFromGroupOffsets();
    consumer.setStartFromEarliest();
    DataStreamSource<String> source = env.addSource(consumer);
```

实现对pri_cust_contact_info表的ETL

```
    rdf = session.sql("select * from UseLessData where contact not in('无',
'null', '', '-')")

    rdf = rdf.withColumn("contact_phone",

        when(rdf.col("con_type") === "TEL" || rdf.col("con_type") === "OTH"
            || rdf.col("con_type") === "MOB",
            col("contact")))

    rdf = rdf.withColumn("contact_address",
        when(rdf.col("con_type").notEqual("TEL") &&
rdf.col("con_type").notEqual("OTH")
            && rdf.col("con_type").notEqual("MOB"),
            col("contact")))
    //    删去其他字段
    rdf = rdf.drop("con_type", "contact", "sys_source", "create_date",
"update_date")

    rdf = rdf.dropDuplicates()
    rdf.createTempView("tempView")

    val tempView = session.sql("select a.uid, " +
        "concat_ws(',',collect_list(a.contact_phone)) as contact_phone,\n " +
        "concat_ws(',',collect_list(a.contact_address)) as contact_address\n " +
        "from tempView a\n" +
        "group by a.uid"
    )
    rdf = tempView.drop("con_rn")
    res = rdf
    println(res.count())
}
```

DBeaver 22.0.3 - pri_cust_contact_info			
文件(F) 编辑(E) 导航(N) 搜索(A) SQL 编辑器 数据库(D) 窗口(W) 帮助(H)			
Auto			
localhost 3 default			
属性 数据 ER 图			
localhost 3 etl 表 pri_cust_contact_info			
输入表格名称的一部分			
pri_cust_contact_info			
输入一个 SQL 表达式来过滤结果 (使用 Ctrl+Space)			
uid	contact_phone	contact_address	
140932199205213694	17849159906	太原万柏区双拥路19号	
230125199808294512	15114200523	辽宁省朝阳县根德乡麒麟村二组0181号	
231002197511170393	13675698777	大丰鼎盛农业有限公司	
232623197504251596	18630778485	江苏省盐城市大丰区上海花园	
320125197109012072	13611577799	南京市鼓楼区中山北路215-1号1136室	
320223195803193890	15312209905	江苏省盐城市大丰区星河名园5号506室	
320304197807232899		江苏省盐城市亭湖区解放路2号金色花园2幢1004室	
320325196004010595	15862053837	邳州市夫山明远村26号	
320721196406025497	15501415399	江苏省连云港市赣榆区青口镇赣榆公安局	
320721196508105391	17712248660	江苏省赣榆县青口镇下口村渔业四队口	
320721196912089485	16651058047	江苏省连云港市赣榆县青口镇大朱旭社区八队758号	
320721197103275493	13921494989	江苏省连云港市赣榆区海头镇海后村	
320721198109205394	13961312908	江苏省赣榆县青口镇下口村渔业六队288号	
320721640602549	15501415399	江苏省连云港市赣榆区青口镇赣榆公安局	
320721710327549	13921494989	江苏省连云港市赣榆区海头镇海后村	
32082919870131502X	15851016525	江苏省大丰市新丰镇太兴村一组83号	
320882198606044841	18967031829	浙江省开化县苏庄镇大坂湾村 2 7 2 号	

实现对流式数据的存储

DBeaver 21.3.0 - dm_v_tr_sa_mx													
文件(F) 编辑(E) 导航(N) 搜索(A) SQL 编辑器 数据库(D) 窗口(W) 帮助(H)													
Auto													
localhost <N/A>													
属性 数据 ER 图													
localhost dm 表 dm_v_tr_sa													
输入表格名称的一部分													
dm_v_tr_sa_mx													
输入一个 SQL 表达式来过滤结果 (使用 Ctrl+Space)													
uid	card_no	cust_name	acct_no	det_n	curr_type	tran_teller_no	cr_amt	ba					
654123198611223715	6230661373600740037	常勇	4558560747715467032444	4,260	CNY	320982900801	1,300	9,010					
654123198611223715	6230661373600740037	常勇	4558560747715467032444	4,261	CNY	320099900N04	11,000	30,010					
654123198611223715	6230661373600740037	常勇	4558560747715467032444	4,259	CNY	320982900801	2,630	7,710					
140421199810115391	6230661373679030286	乔悲渺	4558560577715467470169	1,752	CNY	320982046N07	0	1,973					
14243019590214937X	6230661373672323985	沿礼冀	4558566927549467898609	1,322	CNY	320982046K09	3,550	14,953					
230702197512283426	6230661373604549111	牛吉	4558560437715467096507	169	CNY	320982023N84	0	576					
232326198509013447	6230661373604070134	赵晓璐	4558560447715467410489	99	CNY	320982028N22	0	803					
320219196712254686	623066137367594422	龚络魔	4558560467715467097376	105	CNY	320099900F36	0	9,575					
320825197905196974	6230661373600330425	摆欧拍	4558560207715467402196	115	CNY	320982034N38	0	2					
320825197905196974	6230661373600330425	摆欧拍	4558560207715467402196	114	CNY	320982034N08	6	8					
320828196910304444	6224521393671029931	于茅那	4558560717715467627434	379	CNY	320982052N04	0	1,495					
320828196910304444	6224521393671029931	于茅那	4558560717715467627434	380	CNY	320982052N79	0	1,345					
320911196403064229	6230661373600912040	庞框弯	4558560507715467418596	741	CNY	320982900801	539.76	1,017					
320911196403064229	6230661373600912040	庞框弯	4558560507715467418596	739	CNY	320982900801	185	10					
320911196403064229	6230661373600912040	庞框弯	4558560507715467418596	740	CNY	320982900801	10,293	20					
320911196403140973	6230661373604777548	辛烦厅	4558560507715467461348	276	CNY	320982044N24	0	1,221					
320911197507181440	6230661373600749087	常番舅	4558560747715467413783	75	CNY	320902021A15	0	1,758					
320911197507181440	6230661373600749087	常番舅	4558560747715467413783	74	CNY	320902021A15	0	3,758					
320911197507181440	6230661373600749087	常番舅	4558560747715467413783	76	CNY	320902021A15	0	758					
32091119760531354X	6230661373604770584	裴鹿汪	4558560117715467032477	221	CNY	320982002N39	0	24					
32091119760531354X	6230661373604770584	裴鹿汪	4558560117715467032477	223	CNY	320982900801	3,000	3,014					
32091119760531354X	6230661373604770584	裴鹿汪	4558560117715467032477	222	CNY	320982002N49	0	14					
32091119760531354X	6230661373604770584	裴鹿汪	4558560117715467032477	224	CNY	320982002N83	0	14					
320911197802246497	6230661373608510523	乔贵桂	4558560747715467182152	5	CNY	320099900F28	0	45					

可视化展示

我们选取了shop该表，用python做了一个可视化展示。该表展示了日期与销售额之间的关系。

