## 模块A：大数据平台部署管理

本模块用到的所有安装包均在Master节点的/chinaskills目录下。

### Hadoop完全分布式部署:

1. 将JDK包解压到/usr/local/src路径，将完整命令复制粘贴到对应报告中;
2. 修改/root/.bash\_profile文件，设置JDK环境变量，并使环境变量只对当前root用户生效将环境变量配置内容复制粘贴至对应报告中;
3. 从master复制上面步骤配置的JDK环境变量文件到slave1、slave2节点，命令和结果复制粘贴在对应报告中;
4. 配置SSH密钥登录，实现从master登录到slave1，将登录命令和结果复制粘贴在对应报告中;
5. 将配置文件hadoop-env.sh变更内容复制粘贴在对应报告中；
6. 将配置文件core-site.xml变更内容复制粘贴在对应报告中；
7. 初始化Hadoop环境namenode，将命令及结果复制粘贴在对应报告中；
8. 查看master及slave1节点jps进程，将其命令及结果复制粘贴至对应报告中。

### ZooKeeper部署管理:

1. 解压Zookeeper安装包到“/usr/local/src”路径，并修改解压后文件夹名为zookeeper，将修改命令及结果复制粘贴至对应报告中；
2. 设置ZOOKEEPER\_HOME环境变量，并使环境变量只对当前用户生效，命令及环境变量内容复制粘贴至对应报告中；
3. 配置“zoo.cfg”配置文件，将文件变更内容复制粘贴至对应报告中；
4. 修改myid配置文件，将文件变更内容复制粘贴至对应报告中；
5. 启动每个虚拟机上的Zookeeper节点，启动完成之后查看每个节点的zkServer服务状态，命令及结果复制粘贴至对应报告中。

### Flink on yarn部署:

1. 将JDK包解压到/usr/路径，配置环境变量使其生效，将完整命令复制粘贴到对应报告中;
2. 修改/root/.bash\_profile文件，设置Flink环境变量，并使环境变量生效将环境变量配置内容复制粘贴至对应报告中;
3. 在yarn上以per job模式运行 $FLINK\_HOME/examples/batch/WordCount.jar，将运行结果粘贴在报告中。

### Spark on yarn部署：

1、将JDK包解压到/usr/路径，配置环境变量使其生效，将完整命令复制粘贴到对应报告中（若已安装，则可跳过）;

2、将scala包解压到/usr/路径，配置环境变量使其生效，将完整命令复制粘贴到对应报告中（若已安装，则可跳过）;

3、配置/root/.bash\_profile文件，设置Spark环境变量，并使环境变量生效将环境变量配置内容复制粘贴至对应报告中;

4、复制spark-env.sh.template文件并命名为spark-env.sh

5、在spark-env.sh添加JAVA\_HOME，SCALA\_HOME，HADOOP\_CONF\_DIR，SPARK\_LOCAL\_DIRS变量；

6、使用spark on yarn 的模式提交$SPARK\_HOME/examples/jars/spark-examples\_2.11-2.1.1.jar 运行的主类为org.apache.spark.examples.SparkPi，将运行结果粘贴在报告中。

## 模块B：离线数据抽取

### 任务一：抽取数据

1、抽取shtd\_store库中的所有表中新增的数据到Hive的ods层的表中（若ods中没有数据则抽取全部，若有则抽取新增的部分，可以自行设置有无进行训练），分区字段为当前日期，其格式为yyyyMMdd。使用Hadoop指令查看Hive中customer的文件结构;

（例如使用hadoop fs -ls /user/hive/warehouse/ods.db/customer），将结果粘贴在报告中。

## 模块C：离线数据统计

### 任务一：数据清洗(Spark)

1. 将ods层中所有表经过数据清洗（并去重）后抽取到dwd层中，其中将和时间相关的字段全部转为 timestamp类型，时分秒缺失的数字，补充为00:00:00，dwd层表均为最新完整数据，没有分区字段，表需要新建。

2、任意对ods中的某个表运行一个job，删除ods中对应表三天以前分区中的数据，仅保留近三天的（加上当天数据一共三天）分区中的数据，使用Hadoop指令查看Hive中该表的文件结构。

（例如使用hadoop fs -ls /user/hive/warehouse/ods.db/customer），将结果粘贴在报告中。

### 任务二：指标计算(Spark)

1、根据dwd层中的订单表和客户表，找出消费额最高的五位用户，存入mysql中；

2、根据订单表，客户表，国家表，区域表，找出下单数量最多的五个国家，存入mysql中；

3、统计连续两个月下单的用户数量，存入mysql中；

4、统计未退回的订单数量，存入mysql中。

## 模块D：数据采集与实时计算

### 任务一：实时数据采集

1、使用Flume采集26001端口的socket数据，将数据存入到Kafka的Topic中（topic名称为order，分区数为2），将Flume配置粘贴在报告中。

### 任务二：使用Flink处理Kafka中的数据

1、使用Flink消费kafka中的数据，统计实时营业额存入redis中；

2、使用Flink消费kafka中的数据，统计已下单用户数存入redis中。

## 模块E：数据可视化

结合模块C结果mysql中的数据和模块D结果redis中的数据，使用任意语言编写RESTful风格API，前端使用vue.js（2.0及以上），结合Echarts（5.0及以上）完成下列题目。

### 任务一：柱状图呈现5位消费额度最高的用户

### 任务二：条形图呈现5位消费额度最高的国家

### 任务三：实时展示营业额和下单用户数

## 附录：补充说明

### 电商数据表名

|  |  |
| --- | --- |
| **表名称** | **说明** |
| PART | 表示零件的信息 |
| SUPPLIER | 表示供货商的信息 |
| PARTSUPP | 表示供货商的零件信息 |
| CUSTOMER | 表示消费者的信息 |
| ORDERS | 表示订单的信息 |
| LINEITEM | 表示订单明细的信息 |
| NATION | 表示国家的信息 |
| REGION | 表示地区信息 |

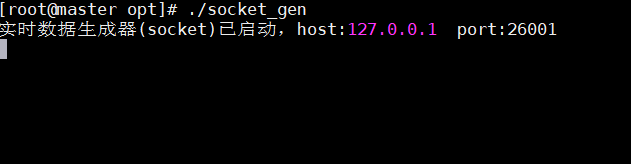
### 电商数据表关系



|  |  |
| --- | --- |
| 表字段注释 （1）supplier表 |  |
| suppkey | 供应商序号 |
| name | 供应商名称 |
| address | 地址 |
| nationkey | 国家代码 |
| phone | 电话 |
| acctbal | 余货 |
| comment | 备注 |
| （2）region表 |  |
| regionkey | 地区代码 |
| name | 地区名称 |
| comment | 备注 |
| （3）customer表 |  |
| custkey | 主键 |
| name | 姓名 |
| address | 地址 |
| nationkey | 外键 |
| phone | 电话 |
| acctbal | 余额 |
| mktsegment |  |
| comment | 备注 |
| （4）part配件表 |  |
| PARTKEY | 配件序号 |
| name | 配件名 |
| mfgr | 厂商 |
| brand | 品牌 |
| type | 类型 |
| size | 大小 |
| container | 包装 |
| retailprice | 零售价 |
| comment | 备注 |
| （5）partsupp配件供应表 |  |
| partkey | 配件序号 |
| suppkey | 供应商序号 |
| availqty | 供应量 |
| supplycost | 批发价 |
| comment | 备注 |
| （6）orders零售订单表 |  |
| orderkey | 订单编号 |
| custkey | 顾客序号 |
| orderstatus | 订单状态 |
| totalprice | 总价 |
| orderdate | 下单日期 |
| orderpriority | 订单优先级 |
| clerk | 收银员 |
| shippriority | 发货优先级 |
| comment | 备注 |
| （7）lineitem订单明细表 |  |
| orderkey | 订单编号 |
| partkey | 配件序号 |
| suppkey | 供应商序号 |
| linenumber | 流水号 |
| quantity | 数量 |
| extendedprice | 价格 |
| discount | 折扣 |
| Tax | 税 |
| returnflag | 退回标记（R拒绝退回 A接收退回 N 默认状态，无退回申请 ） |
| linestatus | 明细状态（O已发货，F没发货） |
| shipdate | 发货日期 |
| commitdate | 预计到达日期 |
| receiptdate | 实际到达日期 |
| shipinstruct | 运单处理策略 |
| shipmode | 运输途径 |
| comment | 备注 |

### 实时数据生成器脚本socket\_gen使用说明

1、利用terminal执行脚本，启动之后提示对应的socket的地址信息（注：由于Flume监听端口是作为服务端，所以应先启动Flume监听再启动该实时数据脚本）



2、将Flume接收到脚本的数据存入Kafka，在Master节点使用以下命令可测试接收到的数据

kafka-console-consumer.sh --bootstrap-server ‘master节点ip’:9092 --group test11 --topic order

