**实验11 Hive会计数据仓库高级数据分析**

一、实验目的

1.掌握高级数据分析流程；

2.掌握高级数据分析方法；

3.掌握数据仓库Hive的命令操作；

4.熟练掌握销售业务高级数据分析模型。

二、实验环境

1.实验平台：《会计大数据基础》课程平台

2.操作系统：Linux CentOS 7.2

3.大数据平台：Hadoop 2.9

4.应用工具：Hive

三、实验要求

1.运用销售基本业务数据仓库，实现销售基本业务各个维度的分析；

2.根据查询结果，做出分析结论。

四、实验内容和步骤

1. 启动Hadoop。

**输入命令：**【start-all.sh】

1. 启动 Hive。

**输入命令：**【hive】

1. 进入销售基本业务数据库。

**输入命令：**【use xsjbyw\_dw;】。

1. **业务需求**：因2020年618大促活动即将到来，需要了解2019年6月18日的销量情况。

**分析目标：**为2020年618活动的销量情况做出更好的预测，以便提前准备库存。

**实现过程：**在Hive数据仓库里对各地区各商品2019年6月18日的销量总量情况进行查询。

|  |  |
| --- | --- |
| 事实表、维度表 | sale\_fact(销售事实表）、product\_dim（商品维度表）、customer\_dim（客户维度表）、area\_dim(地域维度表) |
| 数据模型 | 星座模型 |
| 连接字段 | spid（商品编码）、khid (客户编码）、dyid（地域编码） |
| 连接结果 | 将事实表和维度表连接，维度表和维度表连接，获得维度表中更多的数据 |

**HQL语句：【SELECT product\_dim.spmc AS f\_spmc,area\_dim.ssqy AS f\_area,sum(sale\_fact.xssl) AS xszl FROM sale\_fact INNER JOIN product\_dim ON sale\_fact.spid=product\_dim.spid INNER JOIN customer\_dim ON customer\_dim.khid=sale\_fact.khid INNER JOIN area\_dim ON customer\_dim.dyid=area\_dim.dyid where rqid = '2019/6/18' GROUP BY product\_dim.spmc, area\_dim.ssqy;】**

**分析结论：**根据查询结果可知各地区的销量情况，本公司的面包片油炸机在东北片区618活动当天销售最好，而漂洗机在华东片区和华南片区的销售量最低，公司可以考虑到商品销售量的地区分布情况对2020年各类商品库存进行预估和调整。

1. **业务需求：**销售部想确定本公司2019年最受欢迎的商品，需要了解各商品的交易次数情况。

**分析目标：**找出交易次数最高的商品。

**实现过程：**在Hive数据仓库里查询2019年各商品的交易次数，并按降序排序。

|  |  |
| --- | --- |
| 事实表、维度表 | sale\_fact(销售事实表）、product\_dim（商品维度表） |
| 数据模型 | 星座模型 |
| 连接字段 | spid（商品编码) |
| 连接结果 | 将维度表和事实表连接，获得维度表中更多的数据 |

**HQL语句：【SELECT product\_dim.spmc,count(spmc) AS jycs FROM sale\_fact INNER JOIN product\_dim ON product\_dim.spid=sale\_fact.spid GROUP BY product\_dim.spmc ORDER BY jycs DESC;】**

**分析结论：**根据查询结果可知，本公司交易次数最多的商品是圆筒式撒粉机，后续可以深入分析其核心竞争力和销售策略，为新产品开发和推广提供经验。

1. **业务需求：**公司销售情况受淡、旺季的影响较大，如果存在不合理销售安排，公司可能出现商品成本浪费、资金周转不灵等问题，因此需了解各商品的淡旺季及销售情况。

**分析目标：**分析各商品销量的淡旺季情况，做出合理战略计划。

**实现过程：**在Hive数据仓库里查询各季度的各商品销售量和销售额以分析每个季度的商品销售变化趋势。

|  |  |
| --- | --- |
| 事实表、维度表 | sale\_fact(销售事实表）、product\_dim（商品维度表）、time\_dim（时间维度表） |
| 数据模型 | 星座模型 |
| 连接字段 | spid（商品编码)、rqid（日期编码) |
| 连接结果 | 将维度表和事实表连接，获得维度表中更多的数据 |

**HQL语句：【SELECT product\_dim.spmc,time\_dim.quarters,sum(sale\_fact.xsje) AS total\_xsje,sum(sale\_fact.xssl) AS total\_xssl FROM sale\_fact INNER JOIN time\_dim ON sale\_fact.rqid=time\_dim.rqid INNER JOIN product\_dim ON product\_dim.spid=sale\_fact.spid GROUP BY product\_dim.spmc,time\_dim.quarters ORDER BY product\_dim.spmc,time\_dim.quarters;】**

**分析结论:** 以中级油油炸机（20#电）为例，由查询结果可知此商品没有明显的形式旺季，但销售淡季位于秋季，可能因为秋季对食品的需求变化，而影响此商品销售的时间，故公司在做出发展战略的时候，应把这个问题考虑进去。

1. **业务需求：**销售部为了确定关键客户，对客户购买情况和购买偏好进行分析以调整对关键客户的商品折扣和销售策略。

**分析目标：**分析不同客户的购买力和购买偏好以定位关键客户。

**实现过程：**在Hive数据仓库里查询各公司购买商品数量的情况，并列出购买的商品名称和商品总数量，按销售总量降序排序。

|  |  |
| --- | --- |
| 事实表、维度表 | sale\_fact(销售事实表）、product\_dim（商品维度表）、customer\_dim（客户维度表） |
| 数据模型 | 星座模型 |
| 连接字段 | spid（商品编码）、khid (客户编码） |
| 连接结果 | 将维度表和事实表连接，获得维度表中更多的数据 |

**HQL语句：【****SELECT customer\_dim.khmc,product\_dim.spmc,sum(sale\_fact.xssl) AS xszl FROM sale\_fact INNER JOIN customer\_dim ON sale\_fact.khid=customer\_dim.khid INNER JOIN product\_dim ON sale\_fact.spid=product\_dim.spid GROUP BY customer\_dim.khmc,product\_dim.spmc ORDER BY xszl DESC;】**

**分析结论：**以圆筒式撒粉机销售量分析为例，根据查询结果可知株洲湘味馆连锁店客户和绿悠源食品有限公司购买圆筒式撒粉机数量最多，说明这些客户购买力较强，可能为圆筒式撒粉机的关键客户，需要公司尽力保证这些客户的满意度。

1. **业务需求：**进一步查询购买力最强的客户在2019年的其他各类商品购买情况，以保证今年为其提供充足的库存。

**分析目标：**查询购买商品金额最多的客户公司，并列出购买的商品名称、购买数量、商品总金额。

**实现过程：**在Hive数据仓库里查询购买商品金额最多的公司，并列出购买的商品名称和商品总数量，按销售总量降序排序。

|  |  |
| --- | --- |
| 事实表、维度表 | sale\_fact(销售事实表）、product\_dim（商品维度表）、customer\_dim（客户维度表） |
| 数据模型 | 星座模型 |
| 连接字段 | spid（商品编码）、khid (客户编码） |
| 连接结果 | 将维度表和事实表连接，获得维度表中更多的数据 |

**HQL语句：**【**SELECT customer\_dim.khmc,product\_dim.spmc,sum(sale\_fact.xssl), sum(sale\_fact.xsje) FROM (SELECT customer\_dim.khid,sum(xsje) b FROM sale\_fact INNER JOIN customer\_dim ON sale\_fact.khid=customer\_dim.khid GROUP BY customer\_dim.khid ORDER BY b DESC limit 1) a INNER JOIN sale\_fact ON sale\_fact.khid=a.khid INNER JOIN customer\_dim ON a.khid=customer\_dim.khid INNER JOIN product\_dim ON product\_dim.spid=sale\_fact.spid GROUP BY product\_dim.spmc, customer\_dim.khmc;**】

**分析结论：**由查询结果可知购买商品金额最多的公司是株洲湘味馆连锁店，该客户在圆筒式撒粉机、中级油油炸机（20#电）商品上的购买力是最强的，但在真空油炸机商品上的购买力较弱，公司应该先保障该客的购买力较强的商品供应，以保证客户满意度。

1. **业务需求**：一般情况下，销售单价可能会影响客户的购买力，那么分析根据各地区经济水平制定的销售单价对销售数量产生影响情况是很有必要的。

**分析目标：**分析公司各地区的各商品制定单价是否合理。

**实现过程：**在Hive数据仓库里对查询各个地区各种商品每个月的平均销售单价和平均销售数量，并对平均销售数量进行降序排序。

|  |  |
| --- | --- |
| 事实表、维度表 | sale\_fact(销售事实表）、product\_dim（商品维度表）、customer\_dim（客户维度表）、area\_dim(地域维度表)、time\_dim(时间维度表) |
| 数据模型 | 星座模型 |
| 连接字段 | spid（商品编码）、khid (客户编码）、dyid（地域编码）、rqid（日期编码） |
| 连接结果 | 将维度表和事实表连接，维度表和维度表连接，获得维度表中更多的数据 |

**HQL语句：【****SELECT area\_dim.ssqy AS f\_area,product\_dim.spmc AS f\_spmc,time\_dim.months AS f\_months, sum(xsje)/sum(xssl) AS pjxsdj, avg(xssl) AS pjxssl FROM sale\_fact INNER JOIN product\_dim ON sale\_fact.spid=product\_dim.spid INNER JOIN customer\_dim ON customer\_dim.khid=sale\_fact.khid INNER JOIN time\_dim ON sale\_fact.rqid=time\_dim.rqid INNER JOIN area\_dim ON customer\_dim.dyid=area\_dim.dyid GROUP BY area\_dim.ssqy, product\_dim.spmc, time\_dim.months ORDER BY pjxsdj DESC,pjxssl DESC;】**

**分析结论：**以圆筒式撒粉机为例，根据查询结果可知圆筒式撒粉机在华南片区的6、8、11、5月销售数量最高，单价在21500左右。由此得出结论可知，经济类似的片区可以把圆筒式撒粉机的单价制定为21500元左右。其他片区也可以从多方面进行相应的分析和调整。

1. **业务需求：**因公司年底进行业绩考核，需要了解各区域各商品的销售量情况。

**分析目标：**查询不同地区销售量较高和较少和商品，思考需要是否调整销售策略。

**实现过程：**在Hive数据仓库里对各个地区各种商品按销售量进行降序排序。

|  |  |
| --- | --- |
| 事实表、维度表 | sale\_fact(销售事实表）、product\_dim（商品维度表）、customer\_dim（客户维度表）、area\_dim(地域维度表) |
| 数据模型 | 星座模型 |
| 连接字段 | spid（商品编码）、khid (客户编码）、dyid（地域编码） |
| 连接结果 | 将维度表和事实表连接，维度表和维度表连接，获得维度表中更多的数据 |

**HQL语句：【SELECT area\_dim.ssqy,product\_dim.spmc,sum(sale\_fact.xssl) AS xsl FROM sale\_fact INNER JOIN customer\_dim ON sale\_fact.khid=customer\_dim.khid INNER JOIN product\_dim ON product\_dim.spid=sale\_fact.spid INNER JOIN area\_dim ON customer\_dim.dyid=area\_dim.dyid GROUP BY area\_dim.ssqy,product\_dim.spmc ORDER BY xsl DESC;】**

**分析结论：**以华东片区为例，根据查询结构可知，华东片区的圆筒式撒粉机和中级油油炸机（20#电）的销量情况最好，因此可以知道该地区这2个商品制定的发展战略值得其他地区借鉴。而真空油炸机的销售量较低，理应调整。结合当地的经济水平可知，单从这方面分析还不够，还应结合历年销量情况结合分析。

1. **业务需求：**公司在去年对每个区域都制定了计划销售金额，检验计划的落实情况，并找出计划没有落实的原因，以及在此基础上制定新一年的年度计划销售金额。

**分析目标：**通过与计划销售金额进行对比，获得各地区每月的完成计划情况相对指标。为新一年销售计划的制定提供数据支撑。计算公式：计划完成程度相对指标=实际完成数/计划任务数×100%。

**实现过程：**在Hive数据仓库中，对各地区每月计划完成程度相对指标进行查询。

|  |  |
| --- | --- |
| 事实表、维度表 | sale\_fact(销售事实表）、area\_dim（地域维度表）、customer\_dim（客户维度表）、xsplan\_fact(销售计划事实表) |
| 数据模型 | 星座模型 |
| 连接字段 | khid（客户编码）、dyid（地域编码） |
| 连接结果 | 将事实表和事实表通过维度表相连接，获得维度表中更多的数据 |

**HQL语句：**【**SELECT a.ssqy,a.real\_xsje,b.plan\_xsje,a.real\_xsje/b.plan\_xsje AS jhwcd FROM (SELECT area\_dim.dyid,area\_dim.ssqy, sum(xsje) AS real\_xsje FROM sale\_fact INNER JOIN customer\_dim ON customer\_dim.khid=sale\_fact.khid INNER JOIN area\_dim ON area\_dim.dyid=customer\_dim.dyid GROUP BY area\_dim.dyid, area\_dim.ssqy)a INNER JOIN (SELECT area\_dim.dyid,area\_dim.ssqy AS b\_area, sum(xsplan\_fact.jhxsje) AS plan\_xsje FROM xsplan\_fact INNER JOIN area\_dim ON area\_dim.dyid= xsplan\_fact.dyid GROUP BY area\_dim.dyid,area\_dim.ssqy ORDER BY b\_area)b ON a.dyid=b.dyid order by jhwcd desc;**】

**分析结论：**从统计结果来看，除西南片区、西北片区完成销售计划外，其他地区均为达到计划销售额，说明其他地区的销售计划设定较高而为达到预期。后续可以找到未达到计划销售额地区的具体月份，分析具体原因。

1. **业务需求：**营业收入是企业长期稳定盈利的基础和源泉。分析各类商品营收情况，以便安排后续生产优化，对营收较差的商品进行改进，对营收较好的可以适当的增加库存储备量。

**分析目标：**获得2019年12月份各商品的主营业务收入增长率，找出营业收入增长率波动较大的商品，以便后续生产对商品的生产改进和储备。

**实现过程：**在Hive数据仓库中对2019年12月份各商品的主营业务收入增长率进行查询。

|  |  |
| --- | --- |
| 事实表、维度表 | sale\_fact(销售事实表）、time\_dim（时间维度表）、product\_dim（商品维度表） |
| 数据模型 | 星座模型 |
| 连接字段 | rqid（日期编码)、spid（商品编码) |
| 连接结果 | 将维度表和事实表连接，获得维度表中更多的数据 |

**HQL语句：**【**SELECT a.spid,spmc, (axsje-bxsje)/bxsje as zyywzzl FROM (SELECT sale\_fact.spid,sum(xsje) AS axsje FROM sale\_fact join time\_dim ON sale\_fact.rqid = time\_dim.rqid where months = '12' GROUP BY spid) a join (SELECT spid,sum(xsje) AS bxsje FROM sale\_fact join time\_dim ON sale\_fact.rqid = time\_dim.rqid where months = '11' GROUP BY spid) b ON a.spid = b.spid join product\_dim ON a.spid = product\_dim.spid order by zyywzzl desc;】**

**分析结论：**从营业收入增长率来看，真空油炸机和漂洗机的销售情况最为可观，在接下来的一个月可以适当的增加真空油炸机和漂洗机的储备。值得注意的是，圆筒式撒粉机和面包片油炸机营收处于负增长状态。这应该进行市场调研，了解市场情况，找出导致其12月销路不好的原因。

五、实验拓展

请思考以下问题，写出命令语句，并在《会计大数据基础》课程平台上完成相应操作，将实验过程截图记录到对应题目。

* + - 1. **业务需求：**双十一即将到来，为了编制明年计划库存，查询各地区各商品2019年11月11日的销量总量情况。

**实现过程：**在Hive数据仓库里对各地区各商品2019年11月11日的销量总量情况进行查询。

**HQL语句：【SELECT product\_dim.spmc,area\_dim.ssqy,SUM(xssl) xszl FROM sale\_fact LEFT JOIN product\_dim ON product\_dim.spid=sale\_fact.spid INNER JOIN customer\_dim ON sale\_fact.khid=customer\_dim.khid INNER JOIN area\_dim ON area\_dim.dyid=customer\_dim.dyid where rqid = '2019/11/11' GROUP BY product\_dim.spmc, area\_dim.ssqy;】**

* + - 1. **业务需求：**公司想要查看今年各个地区各种商品销售额情况并按销售额降序排序，以便能够直观看出商品的销售额情况。

**实现过程：**在Hive数据仓库里对各地区各商品销售额情况进行查询，并按销售额降序排序。

**HQL语句：【SELECT product\_dim.spmc,area\_dim.ssqy,sum(xsje) AS xszje FROM sale\_fact INNER JOIN product\_dim ON product\_dim.spid=sale\_fact.spid INNER JOIN customer\_dim ON sale\_fact.khid=customer\_dim.khid INNER JOIN area\_dim ON area\_dim.dyid=customer\_dim.dyid GROUP BY product\_dim.spmc,area\_dim.ssqy ORDER BY xszje DESC;】**

* + - 1. **业务需求：**销售部想要查询销售额最高的商品，以及购买它的客户名称，为商品的折扣调整提供数据。

**实现过程：**在Hive数据仓库里查询销售额最高的商品，并列出购买的客户名称，按销售总量降序排序。

**HQL语句：**【SELECT a.spmc,customer\_dim.khmc ,sum(sale\_fact.xsje)AS spxse FROM (SELECT product\_dim.spid, product\_dim.spmc,sum(xsje) b FROM sale\_fact INNER JOIN product\_dim ON sale\_fact.spid=product\_dim.spid GROUP BY product\_dim.spid, product\_dim.spmc ORDER BY b DESC limit 1) a INNER JOIN sale\_fact ON sale\_fact.spid=a.spid INNER JOIN customer\_dim ON sale\_fact.khid=customer\_dim.khid GROUP BY a.spmc,customer\_dim.khmc ORDER BY spxse DESC;】

* + - 1. **业务需求：**公司想要查询2019年1月较12月的主营业务收入增长率，以分析在2019年年初到年末的营收情况，以进行年终的总结。

**实现过程：**在Hive数据仓库中对2019年1月较12月份各商品的主营业务收入增长率进行查询。

**HQL语句：**【SELECT a.spid,spmc, (axsje-bxsje)/bxsje FROM (SELECT spid,sum(xsje) AS axsje FROM sale\_fact join time\_dim ON sale\_fact.rqid = time\_dim.rqid where months = '12' GROUP BY spid) a join (SELECT spid,sum(xsje) AS bxsje FROM sale\_fact join time\_dim ON sale\_fact.rqid = time\_dim.rqid where months = '1' GROUP BY spid) b ON a.spid = b.spid join product\_dim ON a.spid = product\_dim.spid;】