《大数据管理》第2章作业 2023.9.20

大数据2101班 李嘉鹏 U202115652

1.（1）共同点：都使用了xml查询（SQL for XML），将查询结果转换为xml格式输出；

不同之处：两条语句所用的模式不同，分别为raw和auto，其中raw返回的行作为元素，列值作为元素的属性，可以将每行的元素命名为自定义的名称，而auto返回表名对应节点名称的元素，每列的属性作为元素的属性输出，可形成简单嵌套结构。

（2）语句（a）的查询结果：

<teacher.tid=“1” teacher.tname=“张一” student.sid=“101” student.sname=“李思” />

<teacher.tid=“1” teacher.tname=“张一” student.sid=“102” student.sname=“王学” />

<teacher.tid=“2” teacher.tname=“赵二” student.sid=“103” student.sname=“孙行” />

语句（b）的查询结果：

<teacher tid=“1” tname=“张一”>

<student sid=“101” sname=“李思” />

</teacher>

<teacher tid=“1” tname=“张一”>

<student sid=“102” sname=“王学” />

</teacher>

<teacher tid=“2” tname=“赵二”>

<student sid=“103” sname=“孙行” />

</teacher>

2.在HiveQL上SQL与hadoop的并行计算框架相结合的两种方式：①通过类似SQL的语法提供Hadoop上的数据管理与查询处理能力，例如Distribute by/Sort by/Order by/Cluster by等，还能支持元组的连接。HiveQL还支持常见的SQL数据结构以及数组、结构体、映射数据类型；②在查询功能上，HiveQL支持嵌入MapReduce程序。

3.在SQL中应用R脚本决策树数据挖掘代码的主要执行流程：

（1）创建决策树模型：基于客户会员卡类型member\_card和客户表其它相关属性以及store\_sales属性，创建一个决策树模型，这需要首先定义客户会员卡类型member\_card与其它属性之间的关系，然后将客户表customer和sales\_fact表连接，并将其作为训练模型的数据。

（2）创建决策树模型存储表：为决策树数据挖掘模型创建表，将生成的二进制对象R包输出结果存储在表中，存储模型的表使用varbinary类型的列属性。

（3）保存决策树模型存储表：运行第一步创建的存储过程，将其产生的模型存在表中。

（4）使用存储在表中的模型进行预测，预测数据通过SQL命令输入，通过存储过程调用生成的决策树模型，预测结果显示输入数据原始值与预测值。

4.（1）三个阶段的评语分别是“ok”、“good”、“excellent”。

MySQL中插入该记录的面向json拓展的SQL语句为：

INSERT INTO R1 (sno, cno, eval) VALUES (1, 3, {evaluation: [{“stage”: 1, “result”: “ok”}, {“stage”: 2, “result”: “good”}, {“stage”: 3, “result”: “excellent”}]});

（2）面向json拓展的SQL查询语句为：

SELECT sno, cno, evaluation->’$[0]’ FROM R1 WHERE sno=1 AND cno=3

UNION ALL

SELECT sno, cno, evaluation->’$[1]’ FROM R1 WHERE sno=1 AND cno=3;