数据集

|  |  |
| --- | --- |
| 数据集 | 参数 |
| Basic | 数据集规模scale\_factor  数据倾斜程度degree,输入范围[0..1]浮点数，0代表均匀分布，1代表高倾斜程度  数据集生成路径HDFS |
| TPC-H | 数据集规模scale\_factor,输入范围为1,30,100,300,1000,3000,10000,30000,100000,单位为G  生成数据文件数num\_files,输入范围>=2  Host节点地址文件host\_list  TPC-H工具本地目录  数据输出的HDFS地址hdfs\_path |
| TPC-DS | 同上 |

执行query

Basic

表结构



覆盖query

|  |  |
| --- | --- |
| 语句 | 特征 |
| select | 单表查询 |
| where | where条件查询 |
|  |  |

TPC-H

**表结构**

TPC-H包括8张表（表上有些约束等需要满足）

PART：表示零件的信息

SUPPLIER：表示供货商的信息

PARTSUPP：表示供货商的零件的信息。

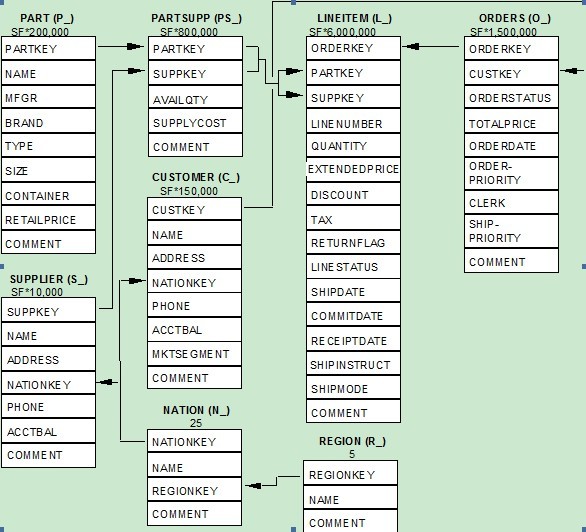
CUSTOMER：表示消费者的信息。

ORDERS：表示订单的信息。

LINEITEM：表示在线商品的信息。

NATION：表示国家的信息。

REGION：表示地区的信息。



覆盖query

|  |  |  |
| --- | --- | --- |
| 语句 | 特征 | 相关表数目 |
| Q01 | 统计查询 | 1 |
| Q02 | WHERE条件中，使用子查询(=) | 5 |
| Q03 | 多表关联统计查询，并统计(SUM) | 3 |
| Q04 | WHERE条件中，使用子查询(EXISTS)，并统计(COUNT) | 2 |
| Q05 | 多表关联查询(=)，并统计(SUM) | 6 |
| Q06 | 条件(BETWEEN AND)查询，并统计(SUM) | 1 |
| Q07 | 带有FROM子查询，从结果集中统计(SUM) | 5 |
| Q08 | 带有FROM多表子查询，从结果集中的查询列上带有逻辑判断(WHEN THEN ELSE)的统计(SUM) | 7 |
| Q09 | 带有FROM多表子查询，查询表中使用函数(EXTRACT)，从结果集中统计(SUM) | 6 |
| Q10 | 多表条件查询(>=, <)，并统计(SUM) | 4 |
| Q11 | 在GROUP BY中使用比较条件(HAVING >)，比较值从子查询中查出 | 4 |
| Q12 | 带有逻辑判断(WHEN AND/ WHEN OR)的查询，并统计(SUM) | 2 |
| Q13 | 带有FROM子查询，子查询中使用外联结 | 2 |
| Q14 | 使用逻辑判断(WHEN ELSE)的查询 | 2 |
| Q15 | 使用视图和表关联查询 | 2 |
| Q16 | 在WHERE子句中使用子查询，使用IN/ NOT IN判断条件，并统计(COUNT) | 4 |
| Q17 | 在WHERE子句中使用子查询，使用<比较，使用了AVG函数 | 2 |
| Q18 | 在WHERE子句中使用IN条件从子查询结果中比较 | 3 |
| Q19 | 多条件比较查询 | 2 |
| Q20 | WHERE条件子查询(三层) | 5 |
| Q21 | 在WHERE条件中使用子查询，使用EXISTS和NOT EXISTS判断 | 4 |
| Q22 | 在WHERE条件中使用判断子查询、IN、NOT EXISTS，并统计(SUM、COUNT)查询结果 | 2 |