数据库性能优化：

1. 索引优化---主键列加索引，常用关联列、where搜索列加索引
2. 分区表优化---范围（range）、列表（list）、哈希、组合分区（数据量大通过先分表，再分表分区）
3. 执行计划优化（优化器hints优化）---CBO代价优化器（cost、time、cpu）

Sql句子性能数据跑的慢优化：

1. 查看执行计划（f5或者PLSQL执行计划窗口、命令窗口查看），索引是否失效
2. Sql语句优化增加索引，过滤、分组、关联、排序时都可以加快
3. 使用分区，减少查询的数量
4. 使用hints，增加并行parallel,
5. 查看存储过程，减少游标使用（用merge into使用）
6. 减少关联的数据量，先过滤再关联
7. 涉及并集语句，多使用union all
8. group by 代替 distinct
9. 表过多，可以使用中间表优化

开窗函数

排序分析----区别

Row\_number 1234 依次递增

Rank 1134 占同排名位置

Dense\_rank 11234 不占同排名位置

位移分析

聚合分析

Kellte优化原因

1.创建mysql资源库报错，因为boolean（0或1，y或n）类型问题实际保存为

Tinyint类型，需要手动修改

1. 乱码问题，数据库连接在选项中添加charaterEncoding=utf-8，资源库连接需要加上，否则在mac下出现乱码
2. 死锁问题：用sort、lookup等组件造成死锁，多并发数据插入更新失败

正则表达式

正则表达式通常被用来检索、替换那些符合某个模式(规则)的文本，如检索字符串中所有的数字，日志中所有带error的信息，比like匹配更加灵活。

部分规则如下

\* 表示匹配所有

(0-9a-zA-Z] 或者 \d表示匹配任意的字母和数字

\s+表示匹配多个空格

^和$表示开始和结束

oracle 中有4个函数

1. regexp\_like 匹配：一般作为过滤条件

2. regexp\_substr截取

3. regexp\_instr查找

4. regexp\_replace 替换

例如

select str from tmp where regexp\_like(str,'^A\d+’,’i’);--’i’ 忽略大小写

：不考虑大小写，包含A和多个数字，如str为12A123B，匹配到A123，所以结果成立

每个单词解析如下：

^A表示实际字母 A，且以A开头

\d表示数字

+表示前面的数字有多个

i是ingore的简写，表示忽略大小写

存储过程优化：

1. 尽量少使用游标（用loop循环忘记关闭游标，for 循环）--建临时表，用merge into；
2. 存储过程有多个更新操作最后提交，进行异常处理，回滚操作，并在存储中执行开始、异常日志记录及结束执行；
3. 建立临时表，避免sql执行多次；
4. sql优化办法

索引失效：(模型数空运最快）

1. 索引列进行运算（各种+、-、\*、\、！）
2. 使用函数
3. 隐式转换
4. 空值、模糊匹配、不等式
5. 最左原则

数据倾斜原因：

1. Key分部不均匀
2. 业务数据本身特性
3. 建表考虑不周，关联字段不符合
4. HQL语句本身存在数据倾斜

数据倾斜优化：

1. 减少数据量，利用分区或分桶，分桶用的少，列多的表做裁剪
2. 语句优化：临时表，group by 代替distinct
3. 数据配置优化：开启map端的聚合，set hive.map.aggr = true
4. 存储格式优化
5. 根据业务逻辑实现整体优化
6. 实际业务场景：列中空值赋值或不参与关联，关联列数据类型不一致

Hive优化：

1. 避免数据倾斜
2. 合理使用分区和分桶，把数据分开存储
3. 使用语句优化
4. 参数配置优化---平台工程师或数据组长

HIVE存储格式：

1. textfile 默认格式，行存储
2. Sequencefile 二进制文件，行存储可分开压缩
3. Rcfile 数据按行分块，列存储
4. Orc 数据按行分块，列存储，效率比rcfile高，是其的改良版本

Hive是什么？

1. 基于hadoop的一个数据仓库工具
2. 可以用于数据抽取、转换、下载，这是一种可以存储、查询和分析存储在hadoop中的大数据数据的机制
3. 将结构化的数据文件映射为一张数据库表，提供sql查询功能

数据建模

1.三范式建模-业务数据库

第一范式（1NF）：字段不可分；

第二范式（2NF）：有主键，非主键字段依赖主键；

第三范式（3NF）：每列与主键有直接关系，非主键字段不能相互依赖；

2.维度建模法

Hive sql 中4个by区别：

order by：全局排序，只有一个reducer，单机处理，效率低，一般不用；

Sort by：分区内排序；

Distrbute by：类型partition，进行分区，结合sort by使用，这个语句之前；

Cluster by：当srot by 和Distrbute by字段一样，就可以用此，只能升序排序，不能指定排序desc和asc

Oracle中数据相乘用exp() 和 ln()：exp(sum( ln(字符名)))

Hive-for循环（确定需要次数），while（无法确定次数）----for或while do done

存储过程怎么防止自己调用的时候避免其他人在调用，写个文件、表控制，判断是否有人在调用；

一张宽表（100多个字段）跟20多张表做关联，怎么写存储过程，分部关联，先几张表做关联。

分区表前一天能跑，隔天跑不了，无数据变化：1.查看日记记录，看看有没有异常进行分析；有张分区表某一天数据插入变多，其他张分区表数据还是很少：分区表数据变多，插入数据存在null，关联字段不对；

关联机制强制切换：hints功能；

连续天数问题：留存问题，建伪列（row\_number分析函数，对id分片，日期排序，命名列-rn，根据id、rn-日期这列分组，having count（1）>=连续登入次数）；

拉链问题

1. 建临时表-需要更新的数据 union all 拉链表中没有的数据
2. 将拉链表没有的数据放入拉链表，结束日期为最大值
3. 利用外关联的方式找出拉链表中变化的数据，将结束日期改为数据抽取时间或当前时间减1
4. 将拉链表中不需要改变的数据加进来，union all

帆软和帆bi区别

前者：直接连接数据库，性能需要数据库支撑，灵活定制复杂报表，面向技术人员，完成业务部门需求，完成的报表由底层员工和领导使用

后者：提供实时数据和抽取数据两种数据计算引擎，多维数据分析，面向业务人员，有业务人员进行仪表板设计分析

前者中的所有数据源可以通过服务器数据集在后者读取使用，前者报表页面挂载在后者查看和使用

Mapping定义

mapping是目标表和源表中的字段的映射关系，包括取数的逻辑和表结构一起写在数据字典里。

Hive数据质量检测-写成脚本定时检测

1. 检查hive元数据---0行检测（表不应该存在0行数据），写成脚本（查看表结构，避免影响下游，为下游任务执行节省资源）
2. 阈值检测：提供一些类sql语句

Mysql数据库和oracle数据库区别：

1. 前者免费，使用许多安全功能（用户名、密码、配置文件、本地身份验证等），后者收费，后者没有表空间、角色管理、快照、同义词和包及自动存储管理，使用三个参数验证用户信息（用户名、密码、位置）；
2. 主键、单引号处理，字符串模糊搜索（后者用不能用索引效率慢）；
3. 日期函数转换区别；

主键、外键和索引的区别：



锁和死锁的定义

在用户在进行修改和更新数据，锁定当前的表或记录，不允许其他用户修改；

死锁是两个会话在争抢对方的资源造成等待的现象，死锁找管理员kill它（plsq ldeveloper工具-会话窗口，找到对应会话kill它）。

视图和物化（物理）视图的区别

视图是不存储任何数据，只有定义，转化为对应的sql查询；物化视图将数据转化为表，存储数据，查询数据不用关联一大堆表，如果表很大会在临时表空间内做大量操作。

Hive运行语句方式：

-e：不进入hive的交互窗口执行sql语句，hive语句是 “”双引号包围SQL语句，可以多行使用；分割。

-f：执行脚本中sql语句，把所有SQL语句写在一个.sql结尾的文件中执行；

-hiveconf:hive 用户配置属性，如hive 运行的参数。

Shell脚本赋权：chmod + x 命令

拉链表的场景（银行存款项目中的账户余额，信用卡中的额度使用余额，客户信息变化，订单信息变化）

Sqoop命令及参数

命令：sqoop import 单表导入

Sqoop import-all-tables 批量导入

Sqoop export 导出

Sqoop job 多任务创建

参数:执行参数、模式参数、保存文件参数、格式化参数、限定参数、控制参数、增量导入参数、验证参数、hive相关参数

Sqoop export 增量导出参数（update-mode模式 updateonly或allowinsert,前者是更新变化的数据，后者除了更新变化的数据还有插入新的数据）

Hive 静态分区与动态分区区别：

1. 前者手动指定分区名，后者根据查询语句自动生成分区名
2. 前者存在分区中一条数据都没有，后者每个分区至少一条数据
3. 后者比前者更消耗性能；前者可以用load或者insert插入数据，后者仅能用insert插入

Sqoop数据采集和load data （load data local inpath ‘/XX/XXX/’ overwrite into table 表名---若去掉local，默认在分布式文件系统上）导入数据至hive区别：

1. 后者适合提供了外部数据文件，该方式使用不对，一般大厂数据源一般都是数据库定时任务抽取数据即可，除非外包数据以文件形式提供
2. 前者Sqoop是可以双向互导

存储过程与自定义函数区别：

1. 前者可以返回多个值或者参数，后者是会返回一个值
2. 前者一般独立执行，后者作为sql语句的组成部分使用
3. 前者实现的功能比较复杂，后者针对性比较强
4. 前者比接口调用快

Merge into 替代游标增量（事实表，存储过程）

ETL工具区别：

Informatica与kettle的区别

1. 前者是商业软件，后者是开源软件、
2. 两者均是gui图形界面（用户图形界面），难度差不多
3. 后者是基于java的etl开发工具，使用过程需要借用JVM(java虚拟机)，在数据抽取速度和大数据处理能力远不如前者
4. 前者是将客户端和服务端进行安装的，而后者则没有但需要基于JVM,故后者部署相对简单
5. 前者较稳定
6. 前者是由data quality组件来保证数据的质量，两者均支持sql语句
7. 均有监控日志功能，前者的监控日志相对详细一些
8. 前者支持各类结构化、非结构化的数据源，后者支持数据库、数据文件

Oracle与hive区别

1. 数据类型不一致，前者是number表示数字，后者分别用int和double来表示整数和小数；
2. Hive建表，需增加参数、存储格式stored as，location保存的位置，分隔符（默认’\t’）
3. Hive数据存储在hafs，oracle数据存储在本地文件系统或者块设备；
4. Hive不建议对数据进行改写（使用overwrite），oracle支持数据改写（delete、update）
5. Hive支持大规模数据计算，oracle支持数据规模较小
6. Hive支持分区不支持索引，oracle支持索引加快查询
7. Hive不支持子查询作为字段和过滤条件，不支持非等值连接
8. Hive没有rowid、rownum伪列，可用row\_number

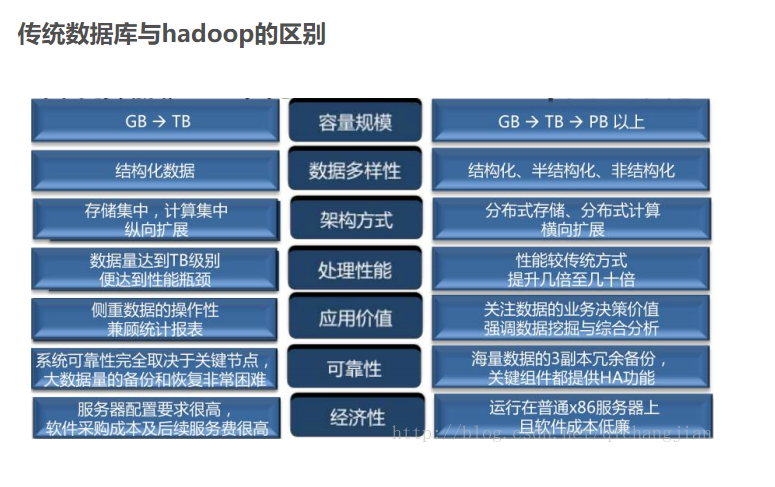
Hivesql 与sparksql区别

1. 前者是基于hadoop下的数仓工具，数据存储在hafs，计算基于MapReduce（任务调度：线程级别），海量数据数据，实时性能较差；
2. 后者是大规模数据处理设计高效计算引擎，内存计算，（任务调度：进程级别），基于

hash的分布式聚合

1. 后者实现内存计算模型，性能高于前者很多
2. 窗口函数中hivesql若没有可以不填写partition by order by，会默认不指定执行，但是sparksql不支持需要填写完整才支持





**工具：**

**Linux命令语句**

查看文件 vi 或cat 文件名

创建新目录 mkdir

切换目录 cd

查看内存 free -h

查看存储空间 df -h

删除文件 rm

移动文件 mv

查找文件 find

查找文件xx.txt 中含‘a’内容 grep或awk a xx.txt

查找etc/hosts包含192的行 grep 或awk 192 etc/hosts

替换文件中内容 sed -i 's/原字符串/新字符串/' /home/1.txt

sed -i 's/原字符串/新字符串/g' /home/1.txt

crontab命令一览：

crontab -e: 编辑当前用户的定时任务列表

crontab -l: 查看当前用户的定时任务列表

crontab -r: 删除当前用户的定时任务列表

定时任务执行命令和脚本

**hdfs操作命令：**

hadoop fs -help

hadoop fs -mkdir hdfs://hadoop-master:9000/mytable ---创建目录下的文件夹

hadoop fs -mkdir hdfs://hadoop-master:9000/mytable/test ---创建目录下的文件夹

hadoop fs -put /mydir/dept.txt hdfs://hadoop-master:9000/mytable/test/ ---把数据文件上传到hadoop

hadoop fs -rmdir hdfs://hadoop-master:9000/mytable/test ---删除目录下的文件夹

hadoop fs -rm -r -f hdfs://hadoop-master:9000/mytable/test ----删除非空目录

hadoop fs -rm -f hdfs://hadoop-master:9000/mytable/test/dept1.txt ---删除目录下的文件

hadoop fs -cat hdfs://hadoop-master:9000/mytable/test/dept1.txt ---查看目录下的文件

**Shell脚本写啥**

平时开发的时候，写shell脚本头部部分，注释，定义变量（引用美元符号$），日记记录（echo），命令异常判断，判断是否为0,运行写到文件内

if [ $? -eq 0] ; then

echo ‘’数据导入成功，时间：’’ $(date)

Else

Echo’’数据导入失败，时间：’’ $(date)

Fi

**存储格式用的什么**

默认是textfile文本格式，我们也会使用ORC格式

Kettle工具 中并发多---日志表被锁、依赖设置问题

索引与组合索引区别，遵从最左边原则