基于ClickHouse的星表快速合并方法

汇报人：黄智鹏

时间：2022年7月20日

目录

CONTENTS

1、背景介绍

背景介绍

科学家们借助越来越先进和精确的天文望远镜、卫星等观测设备，正逐步探索宇宙奥秘，揭开其神秘的面纱。

天文观测逐渐进入到了全波段巡天观测的大数据时代，相关的观测数据呈指数型增长，对数据处理和利用提出了全新的挑战。

天文数据分布分散，研究人员需借助基于交叉证认的数据融合技术，融合多波段信息，从而挖掘出隐含在数据中的潜在价值。

2、交叉证认

交叉证认

证认意义

天文数据的开放性和共享性导致虚拟天文台中存储了大量不同来源、不同内容和不同格式的星表数据。
科研人员需要调取不同星表数据做横向比对或融合时，可利用交叉证认技术获取分布在不同星表文件中属于同一天体的属性信息。

实现原理

不同星表中只有位置信息是共有且一致的，因此只能通过比较两个天体是否处于同一位置才能确定它们是否为同一天体

实现过程

1、从星表A中取出一条待匹配记录，取另一星表作为匹配数据源，分别获取其赤经赤纬位置信息
2、A中待匹配天体记录与B中星表记录逐一计算球面角距离
3、若计算所得球面角距离在星表误差允许范围内，则判定两记录为同一天体的不同观测记录

HTM球面索引+
K-D树快速交叉证认算法

MPI并行编程

ThreadPool多线程技术

基于规则化的
MatchEx交叉证认方法

基于MapReduce框架
交叉证认算法

前人工作

近些年，各国的天文、计算机领域的研究人员，针对大规模天文星表数据下进行交叉证认这一问题进行了诸多研究，提出了一些新颖的算法和思路，并取得了不错的性能表现。

目前研究进展

通过引入层次化索引思想，应用HEALPix伪球面划分技术建立数据索引，借助ClickHouse数据库技术实现大规模天文数据分块连续存放，快速检索，从而提升交叉证认速度

使用大约六千万的星表数据作为测试数据，当前能够将单一天体多次记录的检索时间降低至数十毫秒，相较于大多数方法数分钟的性能表现有了显著提升

3、球面索引技术

伪球面索引

HTM

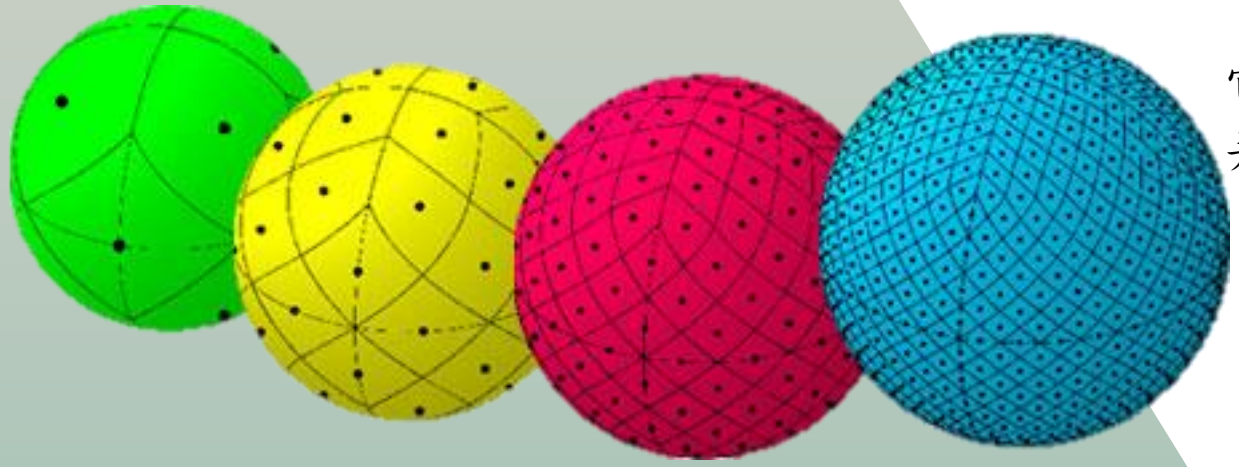
Hierarchical Triangular Mesh

Q3C

Quad Tree Cube

HEAPix

Hierarchical Equal Area
isoLatitude Pixelisation



HEAPix

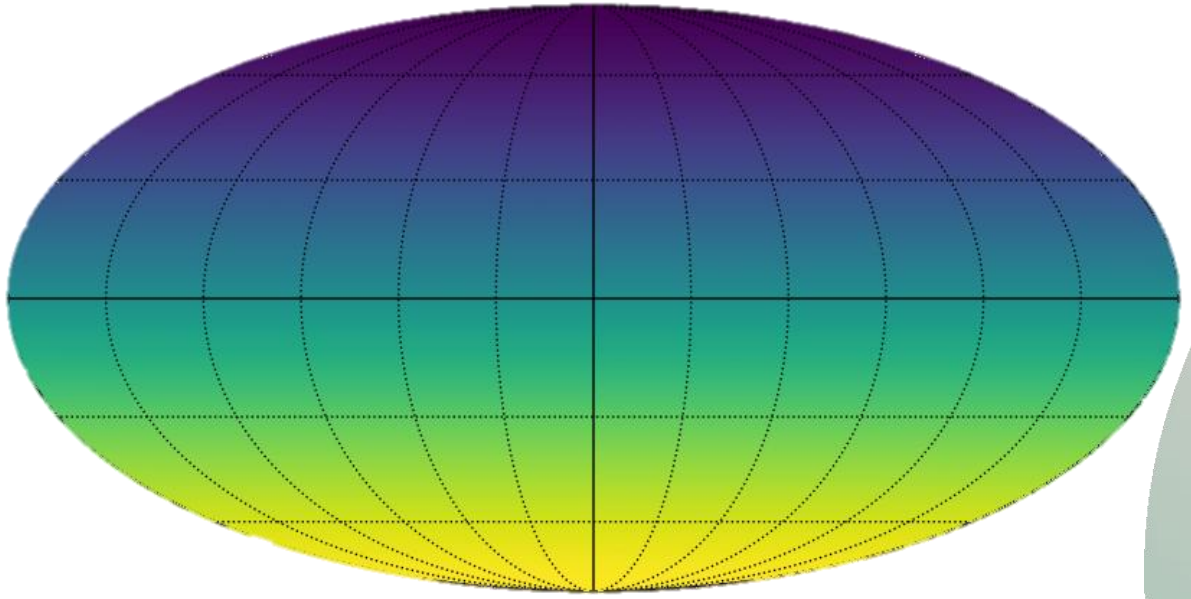
HEALPix is an acronym for Hierarchical Equal Area isoLatitude Pixelation of a sphere.

HEALPix球面索引方法最初是为了解决宇宙微波背景辐射实验的数据处理和分析需求而开发的。

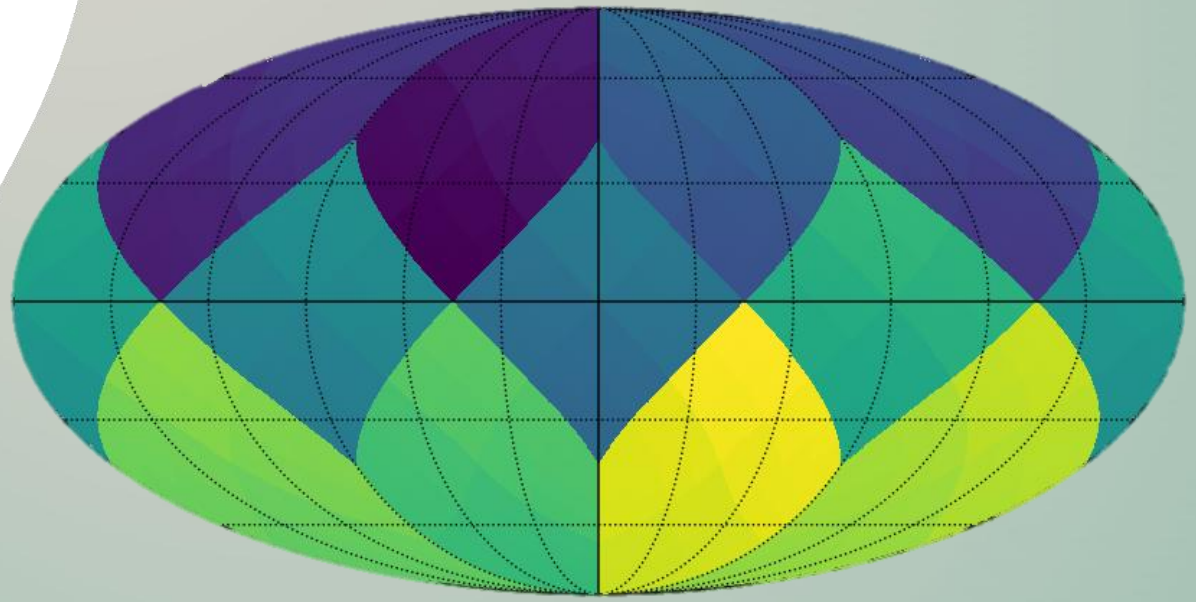
它的划分原理与HTM类似，都是使用基础的几何图形，并采用四叉树结构对天球球面进行层次划分。

HEALPix像素块排序方式

HEALPix RING Order



HEALPix NESTED Order



HEALPix优势

HEALPix作为一种天文学常用的天区索引方法，被广泛应用于天文图像处理、星表交叉证认等场景，相较于另外两种球面划分方法HTM和Q3C，拥有几点突出优势

等面积性

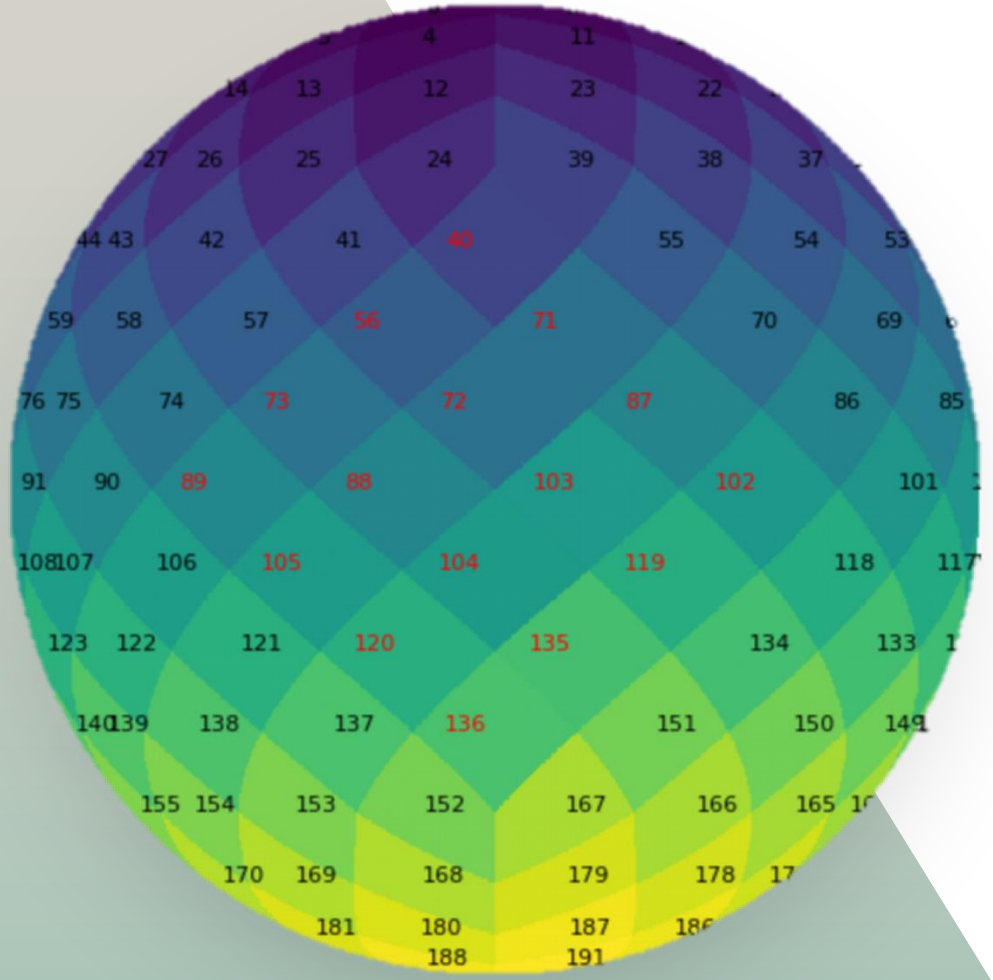
区块之间计算时间均衡，
易实现负载均衡

区块精度高

最高支持29级划分，
精度达0.4毫角秒

通用普遍

被众多天文机构和任务所使用，
支持多种编程语言（支持Python）



边界漏源问题

星表数据会因天文设备所处的环境和设备精度的不同而产生一定程度的偏差，称之为误差半径。

误差半径的存在，导致位于划分块边界位置的记录，可能会被错误地划分到目标块周围的邻接块内

4、ClickHouse

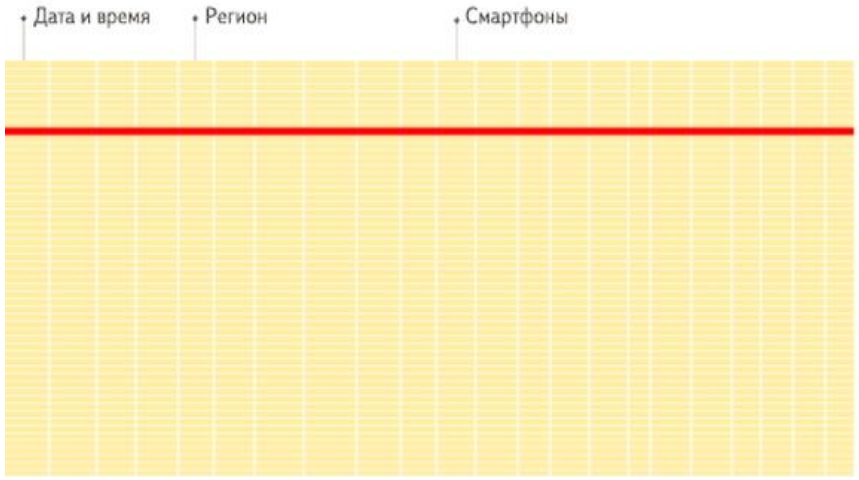
数据库技术

ClickHouse

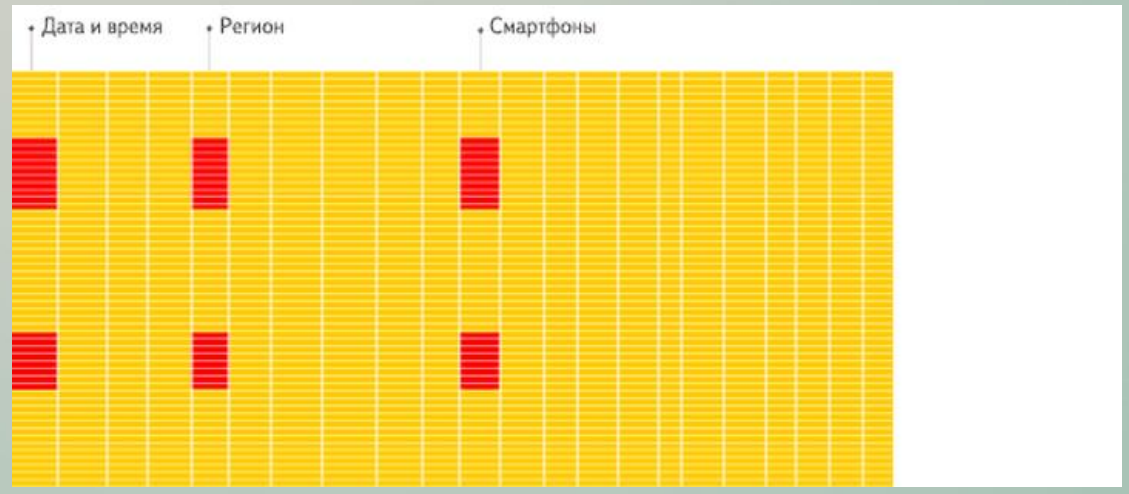
ClickHouse是俄罗斯于2016年开源的一个用于联机分析的列式数据库管理系统，主要用于在线分析处理查询，能够使用SQL查询，并实时生成数据分析报告。

行列数据库对比

行式存储数据库



列式存储数据库



ClickHouse特性

- 列式存储
- 数据压缩
- 多核心并行处理
- 多服务器分布式处理
- 向量引擎
- 数据索引

高效检索

列式存储

数据按列进行文件存储，进行数据检索时减少了磁盘扫描读取的范围，提升检索性能

高压缩率

依照索引粒度，按批次获取数据并进行分块压缩。默认使用LZ4压缩算法，减少数据大小，降低存储空间，加速数据传输过程

稀疏索引

按照主键对数据进行排序，并建立稀疏索引，帮助DBMS在短时间内完成对数据特定值或范围的查找

感谢大家的聆听与指导！

感谢大家的聆听与指导！

感谢大家的聆听与指导！