

# Greenplum在医疗大数据领域的应用

大数据驱动的临床研究

胡锡峰

国家健康医疗大数据研究院 山东大学健康医疗大数据研究院 2021.07.03(济南)

## Ⅰ报告提纲



- 1 "医学数据学"
  - 2 我们关注的问题
  - 3 GP带来了什么
  - 4 致谢

## ■ "医学数据学"学科平台

GREENPLUM DATABASE

■ 医学数据学 (Medical Dataology )

是以数据为中心,以计算为工具,从计算思维/数据思维角度,



**Medical Science** 

解析生物 医学世界 的奥秘,



解决医学 实践中的 问题



**Healthcare** 

的一门新型交叉医学分支学科。

### "医学数据学"学科平台

■ 健康大数据流--特点:多源异构、高噪稀疏

生命历程 危险暴露



疾病进程

健康→低危状态→高危状态→早期病变→疾病状态→预后康复-----→死亡

疾病/死亡 R

危险因

和县

R 防治措施

零级预防 →一级预防 → 二级预防(早诊、早治) →临床诊疗 →三级预防→临终关怀

健康促进 危险因子检测/ 健康教育

早期病变筛查及高 风险预测/干预 危个体识别和早诊

早治

预后评 临床规范诊治/个 性化诊疗/管理

康复诊 疗

四级预 防临床 诊疗

住院前

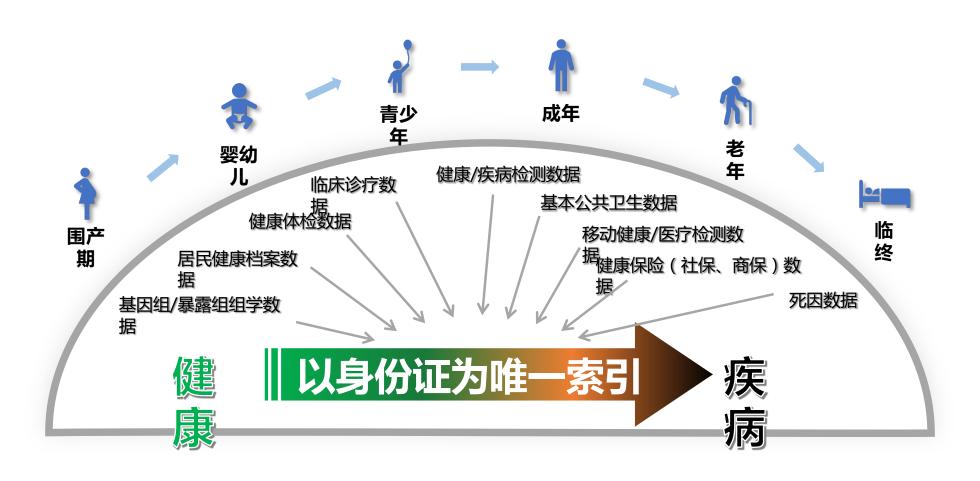
住院

住院后

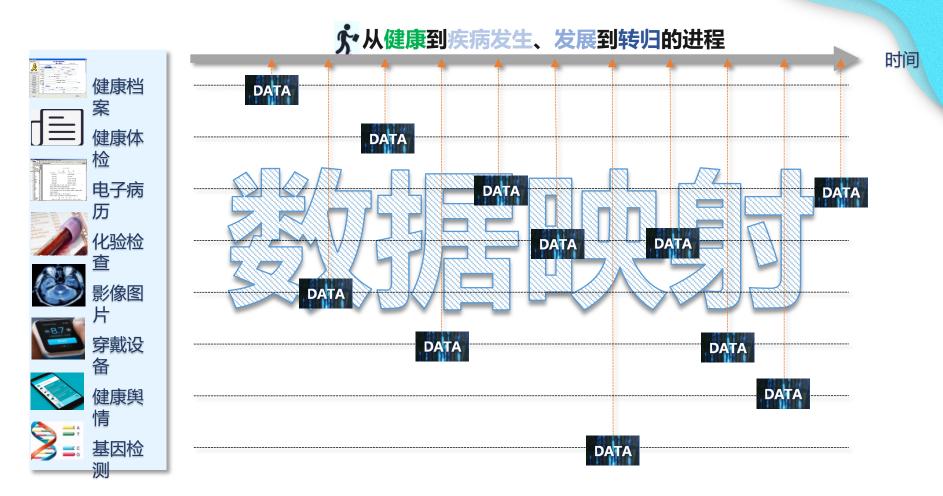
源源不断的健康医疗大数据流。



### ■数据密集型医学新时代:数据大≠大数据



## 数据密集型医学新时代:多源异构高噪稀疏健康医疗大数据



### "医学数据学"学科平台

"理论方法创新和人才培养平台"与"社会服务和产业孵化平台"有机整合的新型医学数据学/健康大数据共建共享平台。



以双一流学科建设为己任,构筑"医学数据学" 学科和"健康大数据共享"两大平台(4000m²

> 医学数据学迈入国际一流 为生物医学研究团队提供数据支撑

以技术转化和社会服务为己任,打造健康 大数据产业孵化和创新创业两大平台(2万

助力国家健康医疗大数据北方中心 助推济南国际医学中心

## 国家健康医疗大数据研究院

## ■数据驱动的医学研究云平台:山东省卫生信息云平台



#### 全员人口信息数据集(2018版)

包括居民个人基本信息、社保信息、住房信息、居民 诚信失信信息等4个数据集



#### 基础资源数据集(2018版)

包括区域总体情况、机构、从业人员、固定资产和运营监管等5个业务域49个基本数据集,以及数据元所需使用的 值域码表



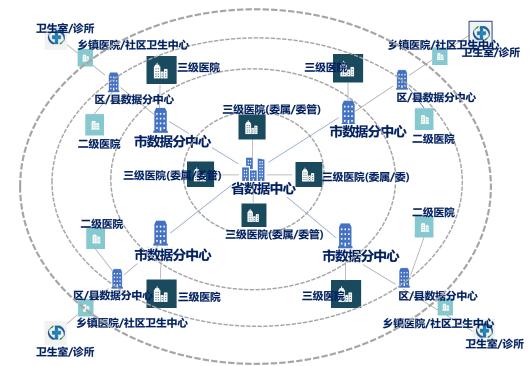
#### 公共卫生数据集(2018版)

包括基本信息、健康体检、儿童保健、妇女保健、疾病控制、疾病管理、计划生育、医疗服务等8个业务域43个基本数据集



#### 电子病历数据集(2018版)

包括门急诊病历、检验结果明细、检查记录、入院记录、住院病案首页、中医住院病案首页、住院其他疾病诊断、出院记录等12个基本数据集,以及数据元所需使用的值域码表。



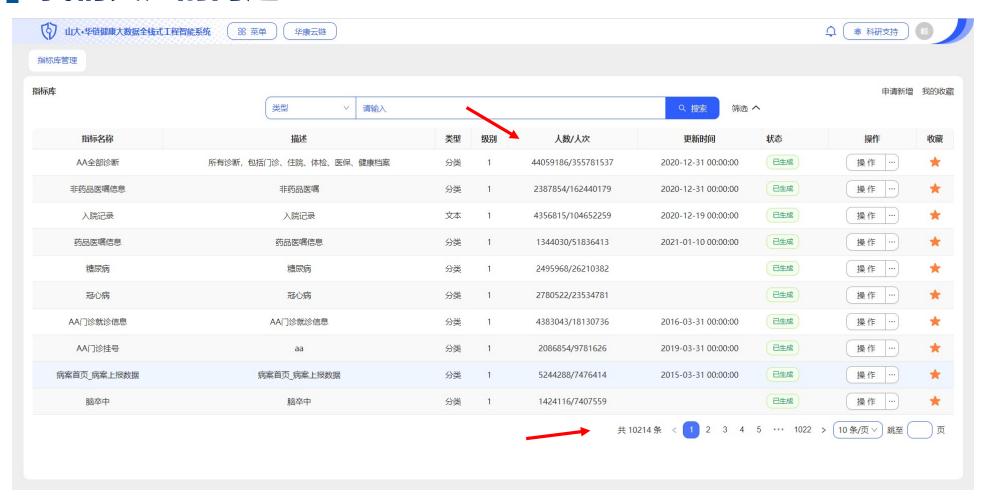


## Ⅰ报告提纲



- 1 "医学数据学"
- ② 2 我们关注的问题
  - 3 GP带来了什么
  - 4 致谢

## ■ 我们关注的问题



### ■ 我们关注的问题

- 入门门槛不能太高:使用对象以非计算机专业的教授和学生为主
- 传统关系型数据库(要分库分表来解决查询效率问题)
- 现有的数据分析工具多数要求数据离线,不安全
- 基因数据放在HBASE中,维护不便
- 实时性需求较少
- 我们没有热数据、温数据、冷数据的分区

## Ⅰ报告提纲



- 1 "医学数据学"
- 2 我们关注的问题
- 3 GP带来了什么
  - 4 致谢

## (1)入门门槛较低

- 标准SQL语法,上手容易
- 与PostgreSQL兼容,便于迁移
- 开放授权,文档规范,社区活跃
- GP自带的运维工具(备份、迁移、测试、优化等)



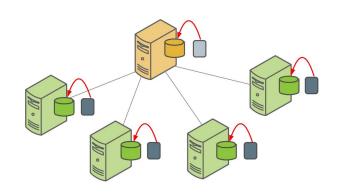






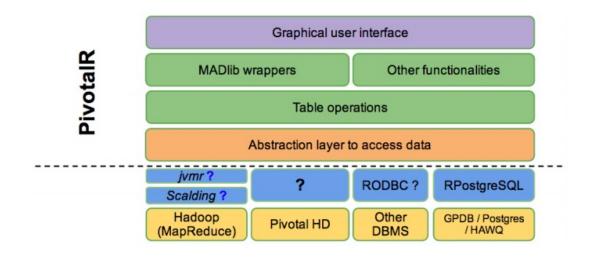
## (2)MPP架构的优势

- 不需要分库分表,方便运维
- 并行计算,提升查询、分析效率
- 支持线性扩容,满足医疗行业多变需求
- 高可用支持,不间断恢复,保证电子病历、健康档案、基因测序等敏感数据安全



## (3)对R语言的深度支持

- R语言是医学、生信领域最常用的统计工具
- 利用PivotalR包和MADlib机器学习库,充分利用并行计算优势,突破R语言性能限制



## (4)面向分析的关系型数据库

- 适合基因数据存储和计算:数据量大、计算密集
- 多表关联查询和复杂查询效率高
- 所有数据在线,原始数据低频使用
- 对非结构化文本数据(医嘱、诊断、检验)的支持

**GPText** 







微信技术讨论群 微信搜索添加"gp\_assistant" 加入技术讨论



微信公众号 搜索添加"Greenplum中文社区" 技术干货、行业热点、活动预告



# Greenplum中文社区 https://cn.greenplum.org

文档 · 博文 · 问答 · 项目 · 下载