2014年8月28日星期四

HUAWEI ENTERPRISE ICT SOLUTIONS A BETTER WAY

区域医疗卫生大数据分析利用

黄 晓 琴 博士

enterprise.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.

HUAWEI

目录

HUAWEI

区域卫生信息平台的特性

HUAWEI

区域卫生信息平台核心数据

HUAWEI

区域卫生大数据分析洞察难题

1

区卫数据中心建设

分级存储

统一管理

互为备份

2

数据共享和访问

异构数据

数据清洗

数据标准化

3

医疗数据挖掘

数据质量

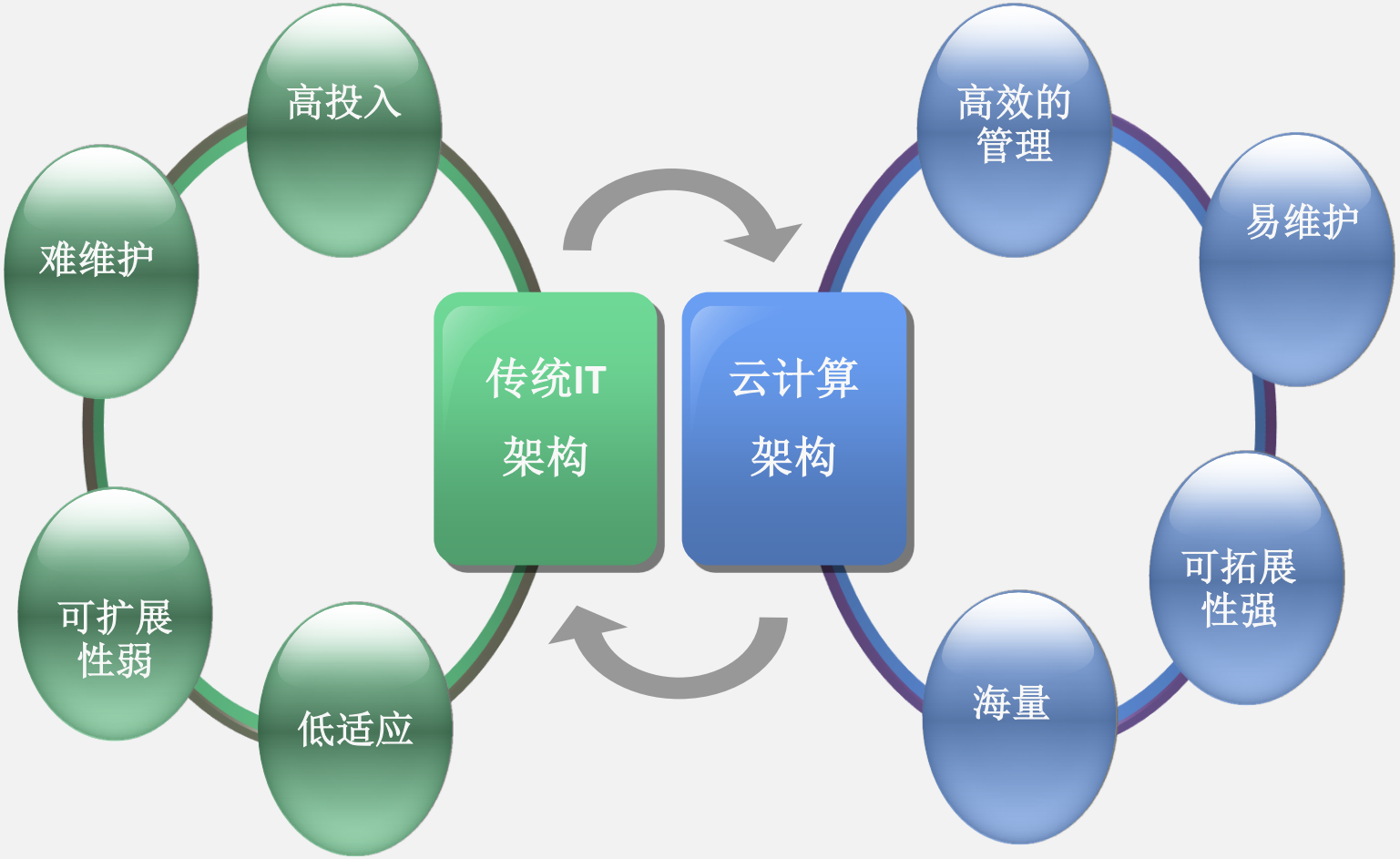
数据建模优化

模型评估

数据可信度到底有多高？

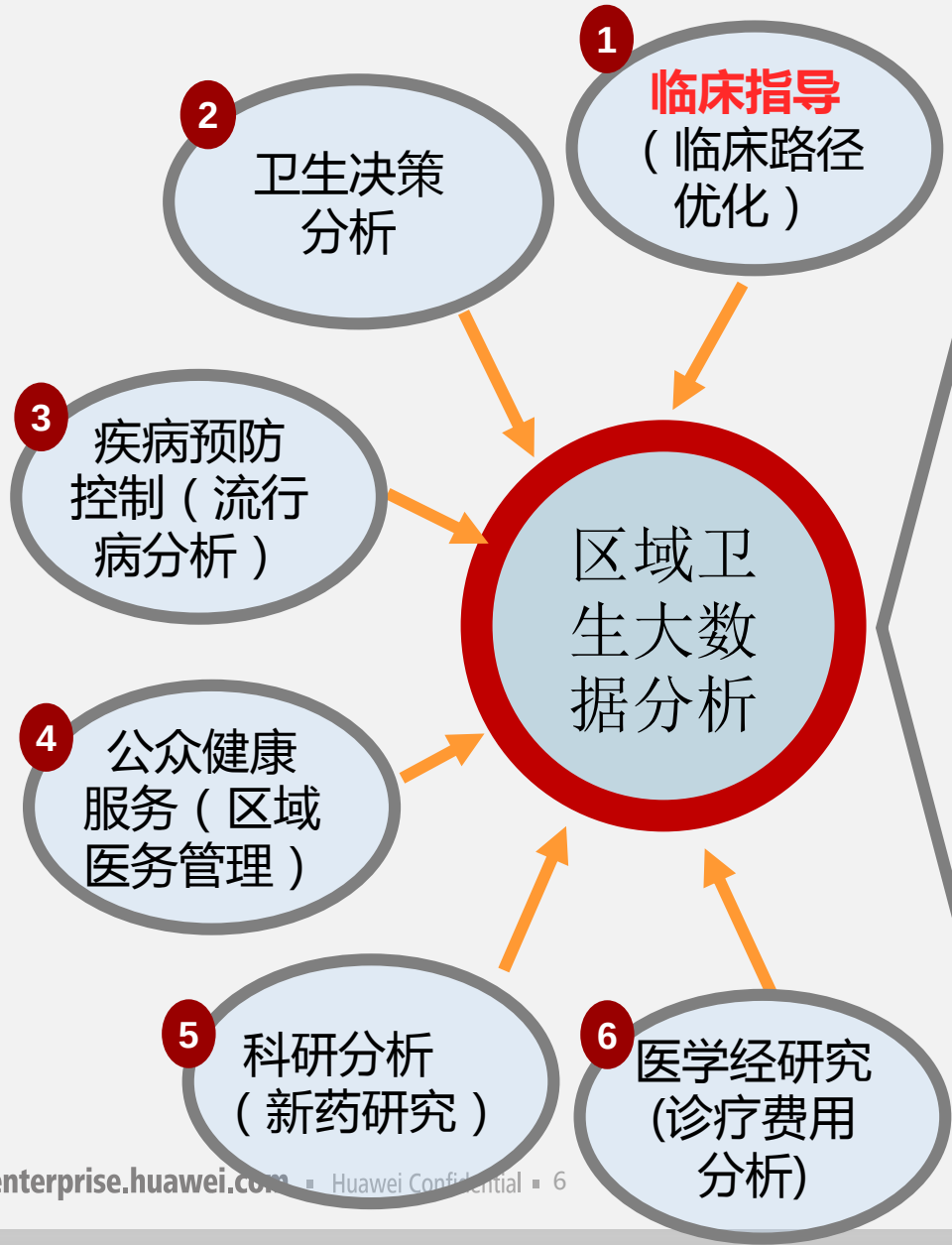
是否能为卫生管理者制定政策提供决策依据？为医生/公卫工作者提供服务？为居民健康提供支撑？

区域卫生大数据分析IT挑战—云计算架构



HUAWEI

区域卫生大数据分析与价值



医疗大数据分析带来的价值

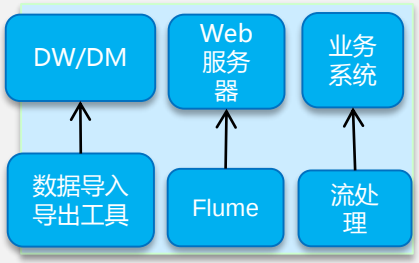
- **提高管理效率**：综合临床和运营相关的有价值的数据
- **提高医疗服务质量**：使得临床策支持系统更为智能的为诊疗提供支持。如药品不良反应、过度使用抗生素等的提醒
- **提高临床科研效率**：如采用大数据进行比较效益研究，评价不同治疗方案对患者的疗效差异
- **降低医疗成本**：利用患者疾病、诊断、用药、治疗、疗效和费用数据，基于成本-效益分析模型

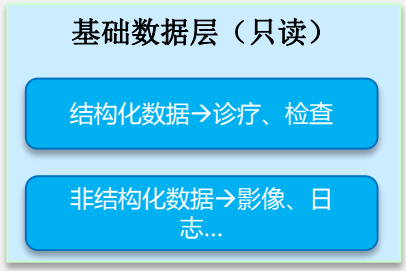
HUAWEI

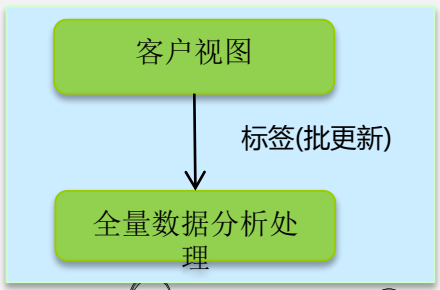
目录

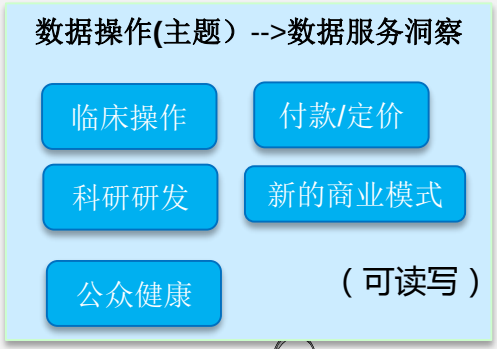
HUAWEI

区域卫生大数据分析的关键技术









数据获取 DataCollect

医院、社区、公卫

- 多数据源
- 分布式收集
- 实时清洗

数据存查 UserProfile

健康档案、电子病历

- 高可靠性
- 高扩展性
- 数据安全
- 实时查询
- 复杂关联
- 存储利用率

客户洞察 UserInsight

行为分析

- 特征管理、模型管理
- 并行计算平台
- 迭代计算平台
- 文本挖掘
- 深度学习机器算法
- 自然语言

客户Profile UserProfile

用户刻画

- 基础特征
- 网络特征
- 文本特征
- 多源数据可视
- 知识图谱展示

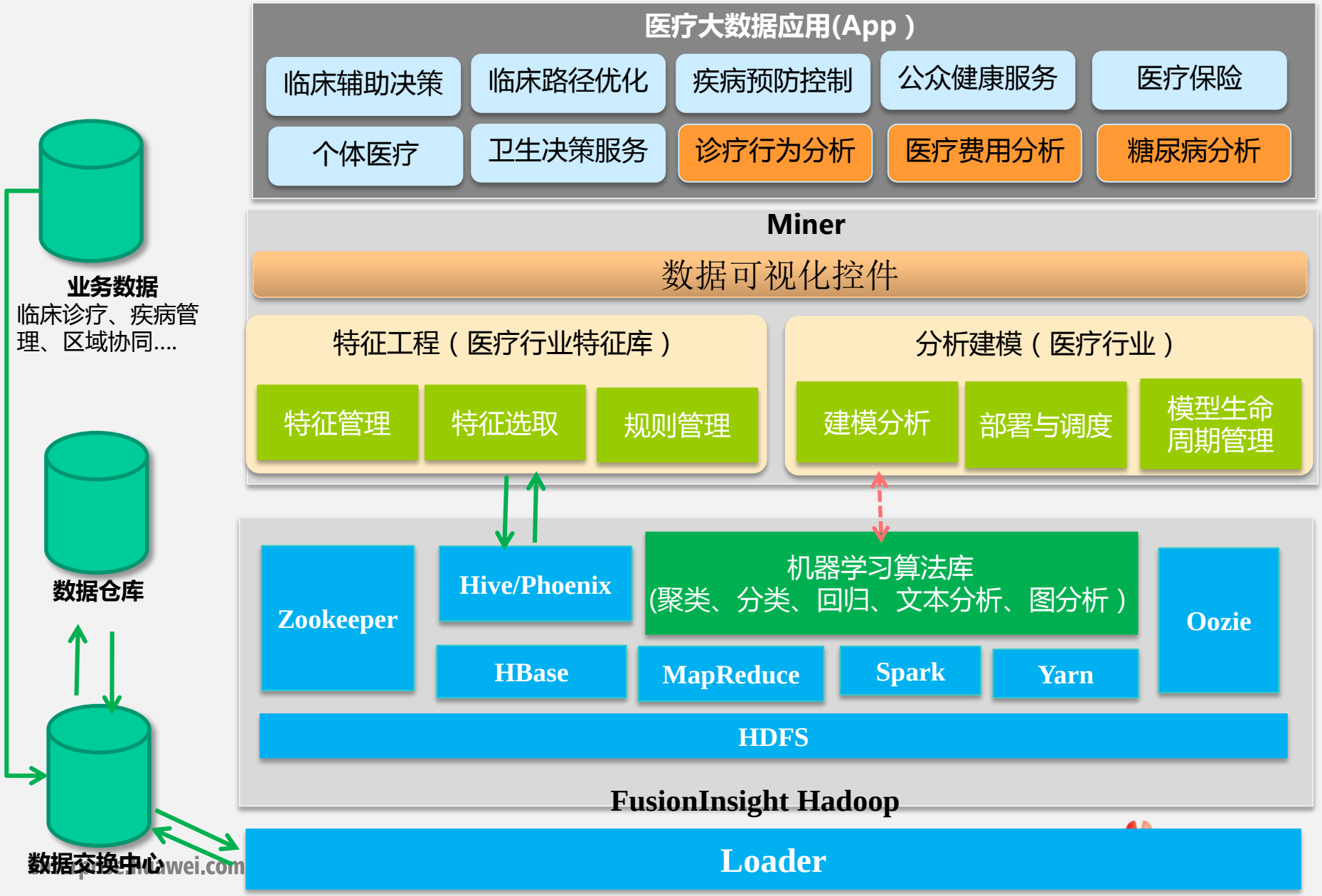
业务分析应用 DataFarm

业务主题分析

- 数据驱动
- 实时响应
- 不同主题
- 插件框架
- 应用解耦

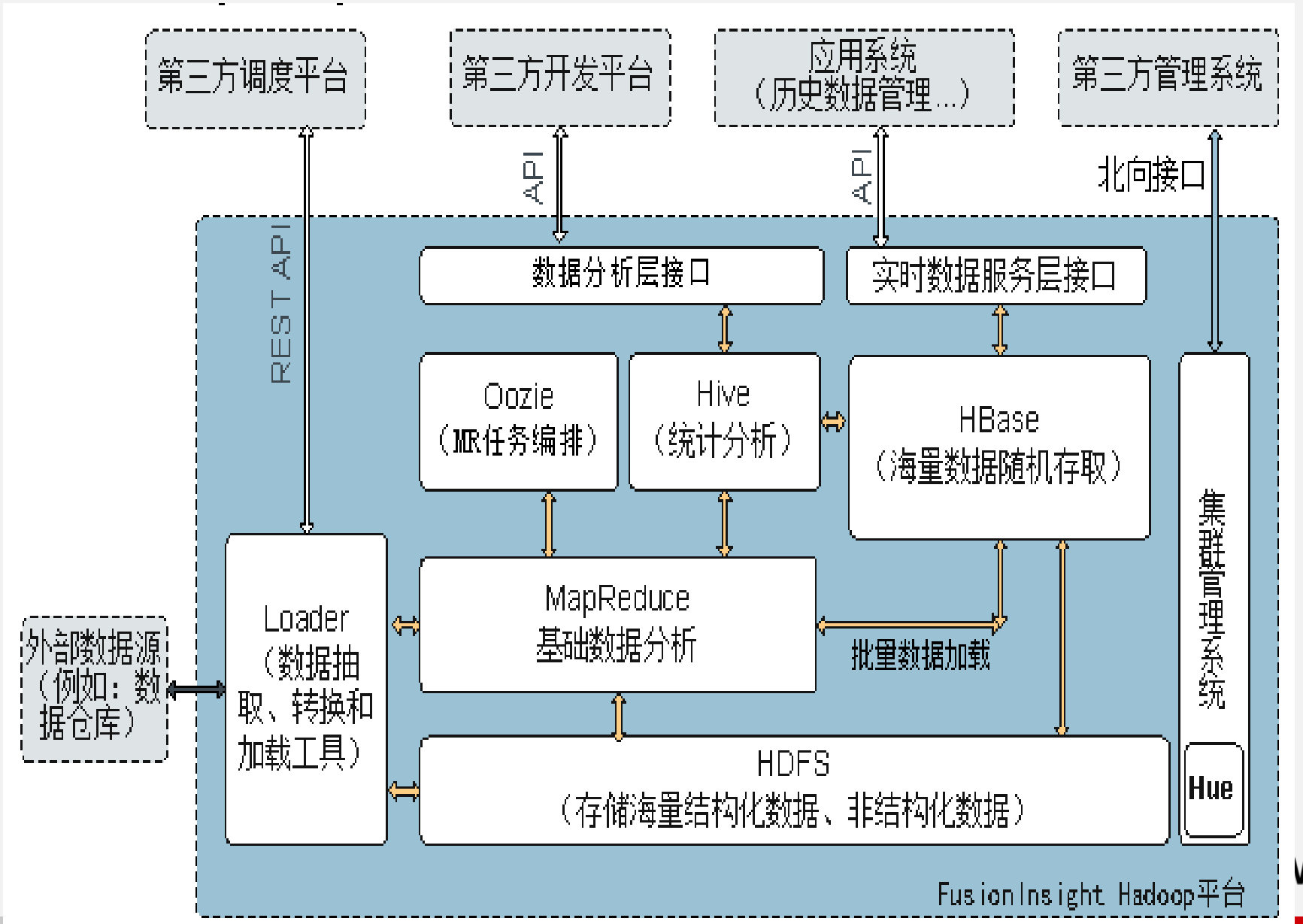
HUAWEI

区域卫生大数据分析挖掘平台系统架构



support.huawei.com

华为FusionInsight hadoop软件架构

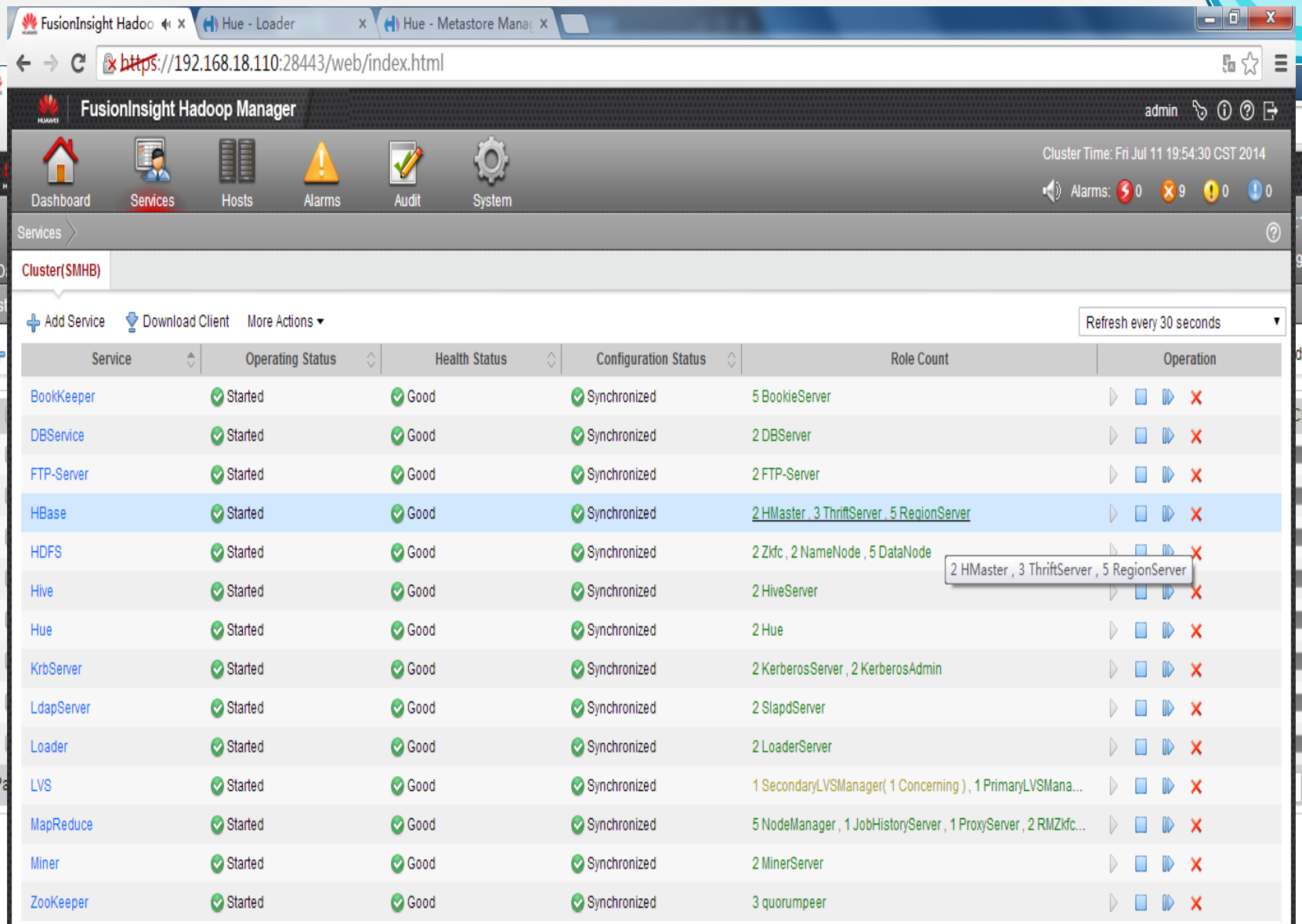


区域卫生大数据分析实施步骤

HUAWEI

区域卫生大数据分析实施路径(0)—搭建分析环境

0
搭建分析环境
软件硬件
及分析工具



区域卫生大数据分析实施路径(1)—业务理解

临床医学方面

- 糖尿病人群身体状况分析(社区医生)
- 糖尿病用药等诊疗手段与疗效的分析（医院医生）

卫生管理循证决策方面

- 糖尿病就诊费用分析（卫生局用）
- 糖尿病就诊行为（医院选择）分析（卫生局用）

总结

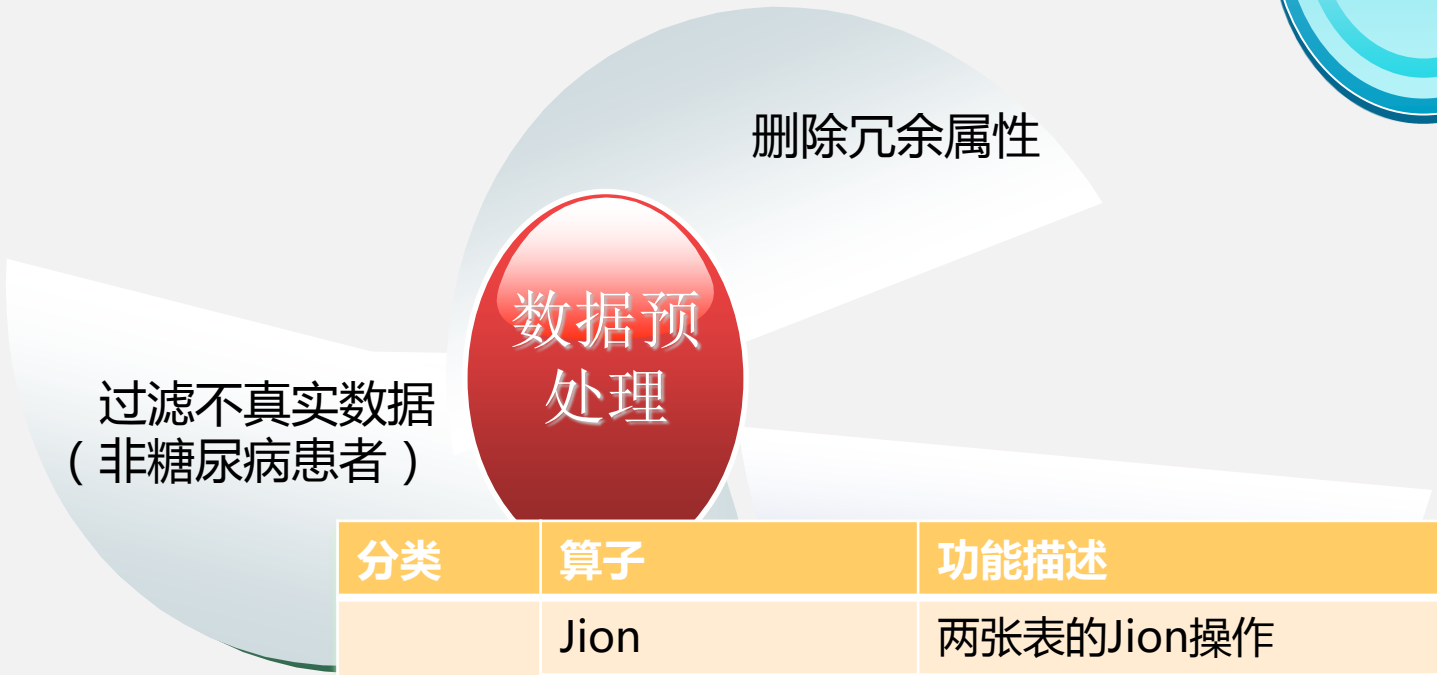
从糖尿病开始分析，后续模型与方法可拓展到其他疾病（如常见疾病—上呼吸道感染，消化系统疾病等，或重大疾病—肿瘤等。

HUAWEI

区域卫生大数据分析实施路径(2)—数据采集与理解

fph	mxfylb	stfsj	mxxmbm	mxxmmc	mxxmdw	mxxmdj	mxxmsl	mxxmje	ylyl1
45490054	07	2013-10-17 07:50:15	51880312	★葡萄糖测定	次	5	1	5	null
45490054	07	2013-10-17 16:38:34	51880326	★血清总胆固醇测定(化学法或酶法)	次	5	-1	-5	null
45490054	07	2013-10-17 07:50:15	51880326	★血清总胆固醇测定(化学法或酶法)	次	5	1	5	null
45490054	07	2013-10-17 07:50:15	51880328	★血清甘油三酯测定(化学法或酶法)	次	10	1	10	null
45490054	07	2013-10-17 16:38:34	51880328	★血清甘油三酯测定(化学法或酶法)	次	10	-1	-10	null
17284	12	2013-10-07 00:00:00	X00027750470010	[甲]头孢呋辛钠针(西力欣)	支	33.5	6	201	null
17284	12	2013-10-07 00:00:00	X00110650010010	[甲]甲硝唑氯化钠注射液（长富）	支	1.8	2	3.6	null
17284	04	2013-10-07 00:00:00	S12010000400010	护理二级	天	12	1	12	null
17284	03	2013-10-07 00:00:00	S12010001300010	动静脉置管护理(留置针)	天	5	1	5	null
95979	12	2013-10-22 14:09:59	00401	西力欣针(注射用头孢呋辛钠)	瓶	33.5	-4	-134	
2124	12	2013-10-22 13:19:02	23661	松梅乐(鹿瓜多肽注射液)	支	43.9	6	263.4	

区域卫生大数据分析实施路径(3)—数据预处理



分类	算子	功能描述
特征管理（数据预处理）	Jion	两张表的Jion操作
	Replace Missing	表中缺省值和控制的替换处理
	Replace SpecialVale	表中特殊值的替换处理
	Sort	表的排序
	SortParallel	表的排序，支持并行全排序，只支持单列
	Filter	根据条件对数据集进行样本过滤

区域卫生大数据分析实施路径(4)—特征提取

◆特征管理

特征管理是指将原始数据预处理后，在特征库中生成新特征及对特征进行维护的过程

Miner的特征工程包括：

- **特征管理**
- **特征选取**
- **规则管理**

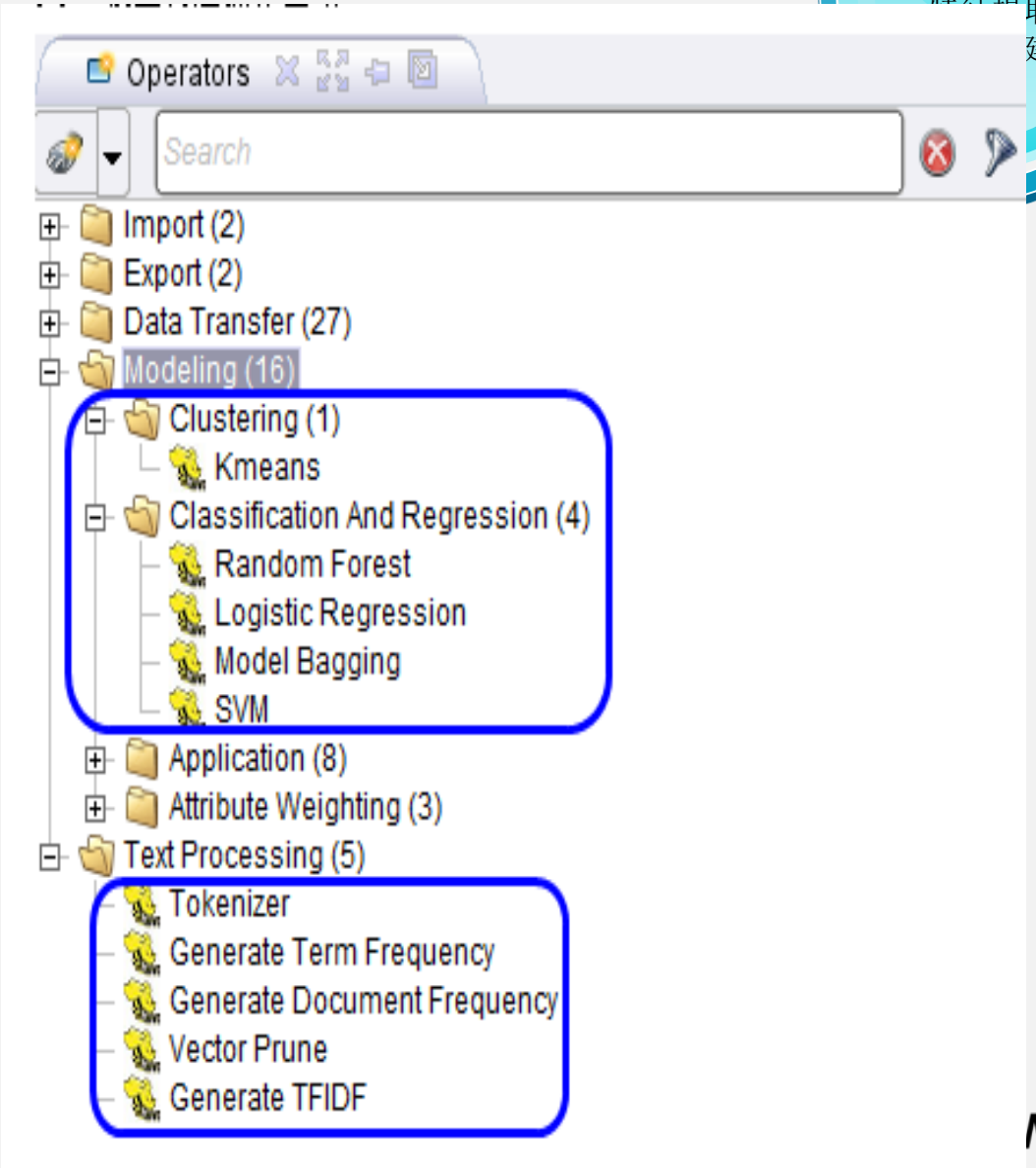
分类	算子	功能描述
特征管理（新增特征）	Set Rlole	设置特征角色
	Transform Attributes	将选取的属性按照所给枚举值生成属性
	Generate ID	生成一个ID特征。
	Normalize	对一个特征或者多个特征进行标准化。
	Select Attributes	选取一个特征或者多个特征。
	Sample by percent	按照比例抽取样本。
	Sample by absoluteSize	按照给定的行数进行取一份样本数据集。
特征选取	Information Gain Ratio	信息增益率算子，根据数据每个特征与目标特征的相关度来计算特征权重。
	Information Gain	
特征排序	Select Weighted Attributes	基于特征权重，将需要的属性选择出来。
规则管理	Generate Attributes	通过对特征进行数据计算、逻辑运算、字符串转换、日期转换生

区域卫生大数据分析实施路径(4)—模型构建

分析建模

指选取合适的模型算法，通过特征化的训练集作为输入进行训练生成评估模型，并对准确率和识别率进行评价。Miner的分析建模包括：

- **模型构建**
- **部署与调度**
- **模型生命周期管理**



区域卫生大数据分析实施路径(4)—模型构建案例

糖尿病用药等诊疗手段与疗效的分析（医院医生）

输入	输出	验证方式
个体基本特征、患病病情状况、不同的治疗手段及其用药、疗效等特征变量	基于测试集数据由预测模型给出当前病人的治疗手段与用药建议	基于测试集数据由预测模型给出当前病人的治疗手段与用药建议，和有实际疗效的糖尿病病人的治疗方式与用药情况进行比对

HUAWEI

区域卫生大数据分析实施路径(5)—模型评估

模型评估

根据作业运行结果，分析应用模型是否满足实际业务需求的过程。如果评估结果不理想，需要重新进行数据理解再构建模型

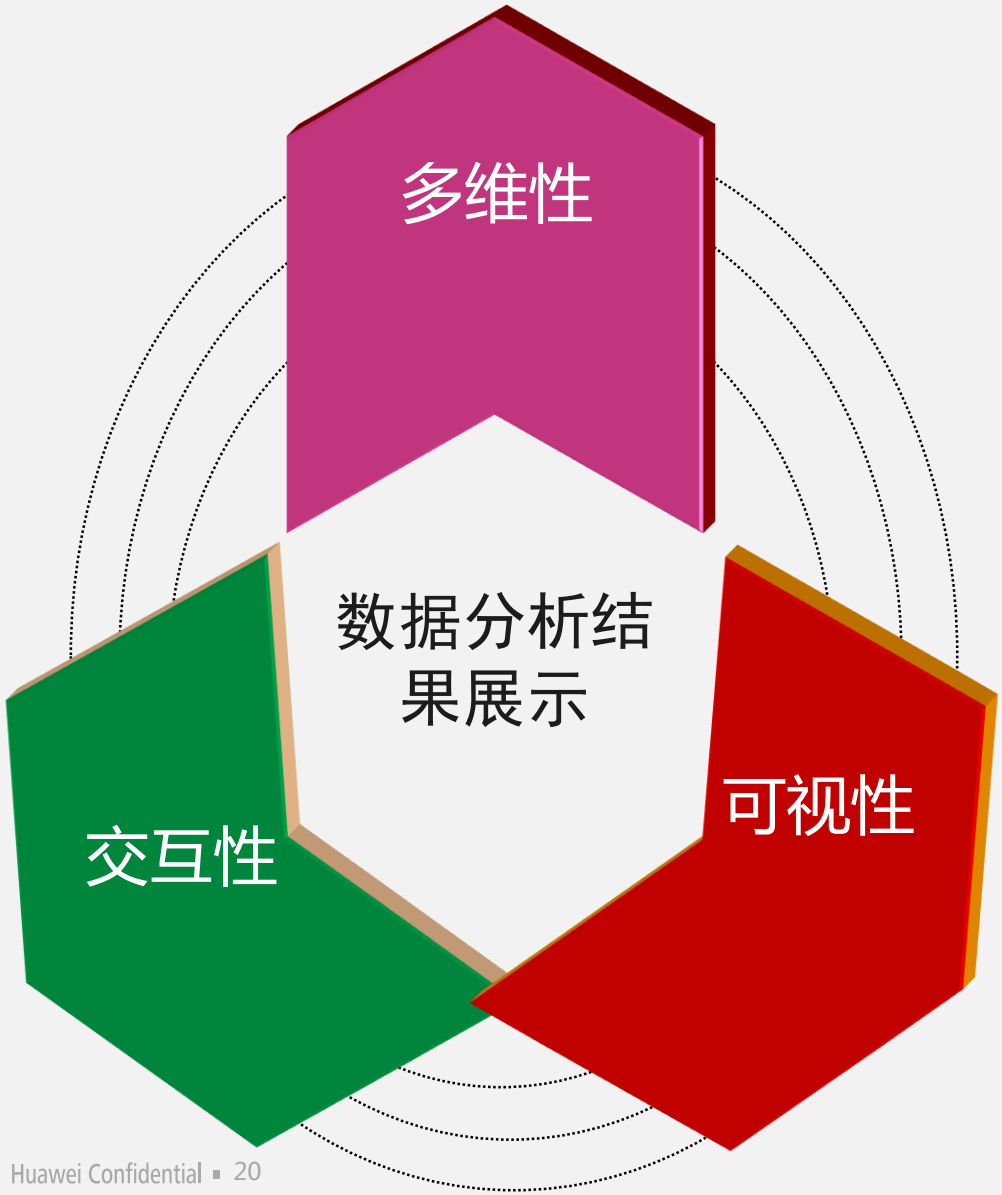
- ◆模型准确率评估
- ◆模型转化率评估
- ◆业务需求评估
- ◆反馈

模型优化

根据作业运行效果，发现结果有偏差，可通过对模型进行优化，重新构建模型进行优化，重新构建模型后再应用的过程。

- ◆通过调整参数
- ◆更换算子

区域卫生大数据分析实施路径(6)—模型应用



HUAWEI

区域卫生大数据分析实施路径(7)—应用效果评估

华为Fusioninsight hadoop 大数据产品介绍

HDFS：分布式文件系统

MapReduce：并行计算处理

Hbase：NoSQL数据库

Hive：SQL转MR处理工具

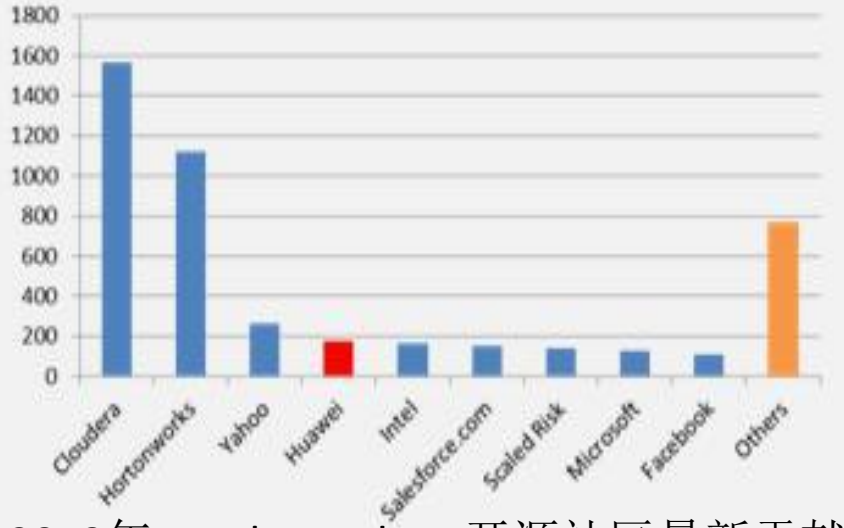
Spark：迭代并行处理

Impala：基于HBase SQL查询引擎

Oozie：工作流处理

Zookeeper：分布式系统协同

OM Server：操作维护与管理

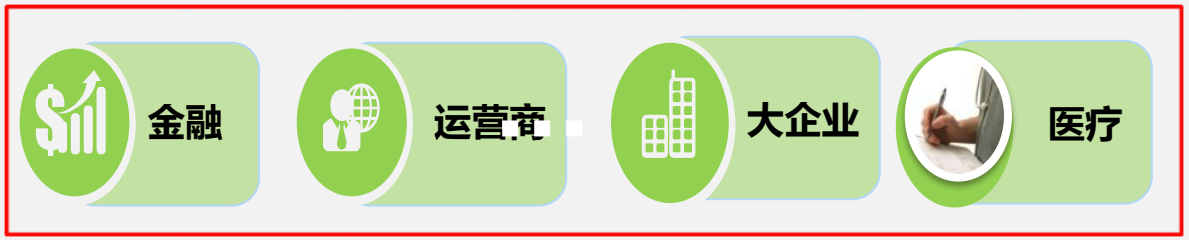


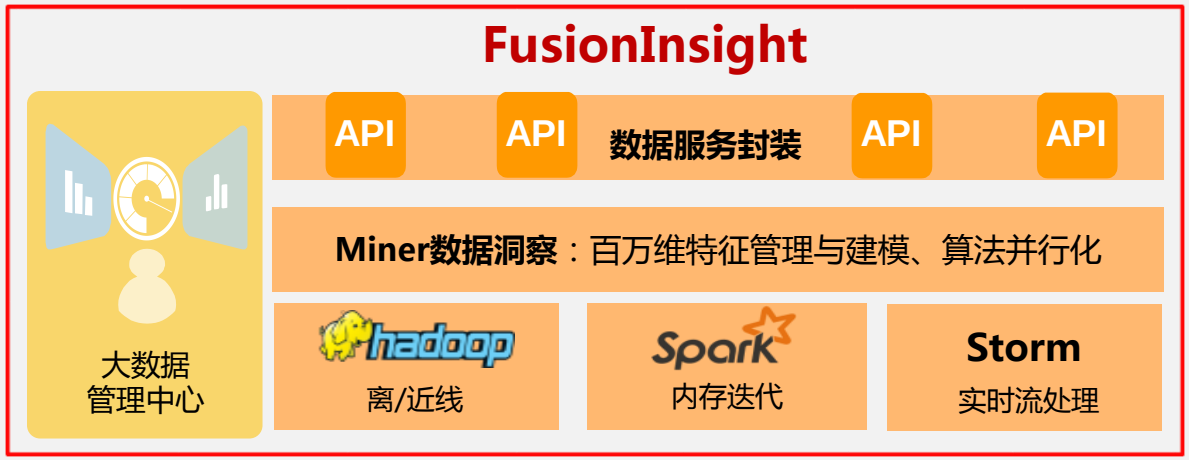
2013年Apache Hadoop开源社区最新贡献量

在社区贡献的基础上，华为公司于2011年推出了企业级大数据解决方案FusionInsight。**华为FusionInsight是企业级大数据存储、查询、分析的统一平台**

HUAWEI

FusionInsight: 企业级大数据处理、分析挖掘平台优势





•智慧

– 全量建模，深刻洞察

– 存储自动分级

•实时

– 在线实时处理

– 领先的存储性能

•可信

– 全组件HA，异地容灾

– 全分布式架构，N+M数据保护

•易用

– 数据全生命周期管理

– 自定义Dashborad 、二次开发助手

HUAWEI

华为FusionInsight端到端竞争力构筑

UserProfile

- 采纳社区精华做企业增强
- SQL on HBase
- 实时交互查询
- 搜索引擎

UserInsight

- 分析挖掘的使能框架
- 特征管理，模型管理
- Spark、Shark、MTLib
- 文本挖掘
- 交互式探索

OceanStor9000

存储

- 存储和计算全融合架构
- 本地化IO提升性能
- 快照、异步远程复制
- 全文检索

DataFarm

云

- 联合客户孵化数据服务平台
- RTMQ
- 实时流Storm
- 生命周期管理
- 服务治理
- 二次开发

DataCollect

- 分布式、实时、多协议
- 文本搜索

HUAWEI

华为大数据：数据分析和挖掘领域的顶尖人才，多项创新成果

美国、香港、深圳、西安

Dr. Hang Li 李航

- 中央研究院Noah Ark Lab首席科学家
- 原微软亚洲研究院主任研究员
- 多个国际会议领域主席
- 个人拥有二十多项美国发明专利

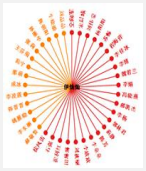
Prof. Qiang Yang 杨强

- 中央研究院Noah Ark Lab主任
- 世界级数据挖掘和人工智能专家
- 香港科技大学教授
- IEEE Fellow，IAPR Fellow

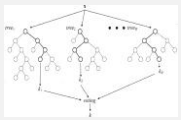
Wenyuan Dai 戴文渊

- 中央研究院Noah Ark Lab主任研究员
- 负责大数据相关的计算金融、、推荐引擎、计算视觉的研究

人物画像

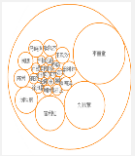


倾向预测



主题提取

关系估计



特征管理分析

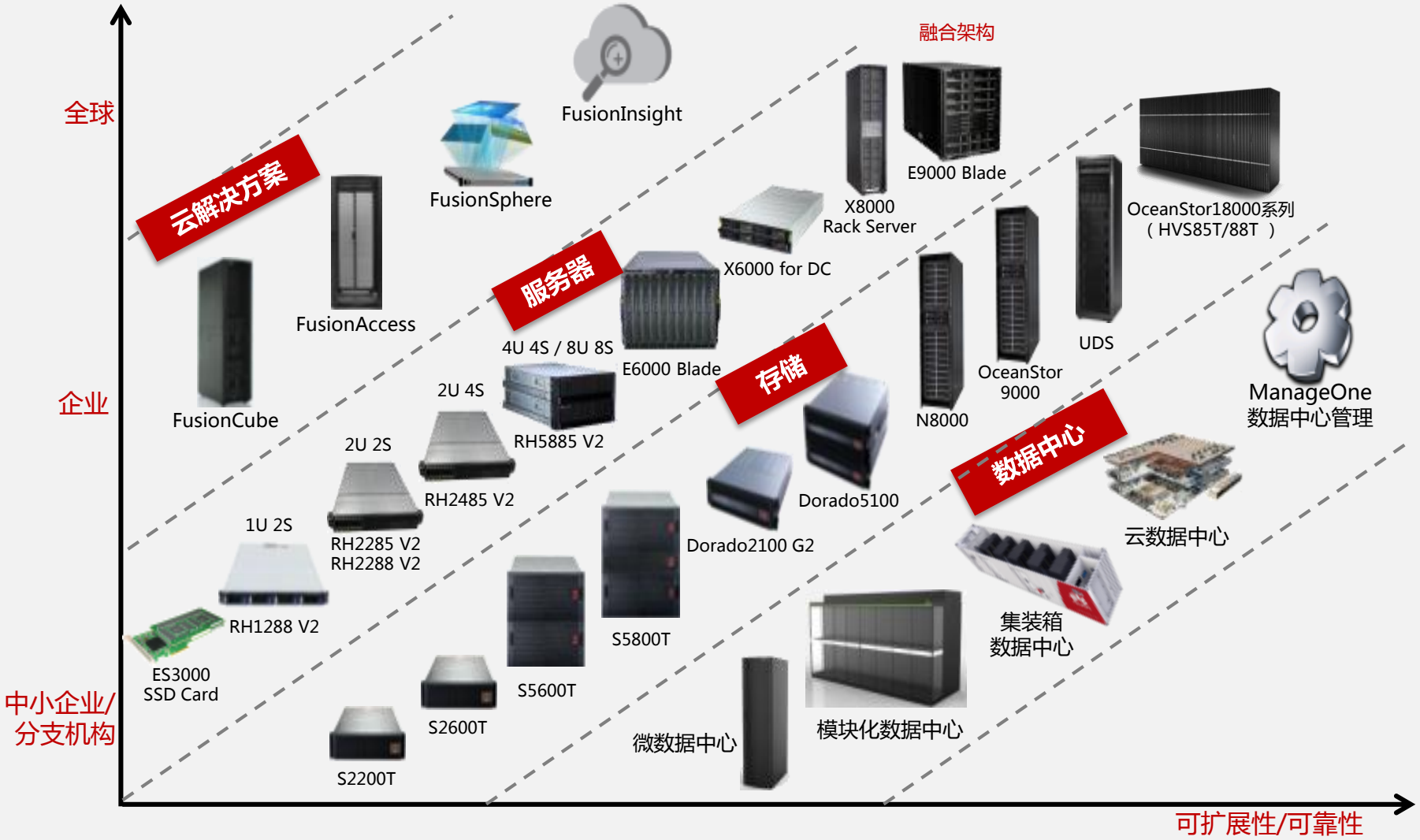
OCEAN

自动特征构建

WHALE

HUAWEI

云及大数据解决方案—华为全系列可扩展和高可靠的IT产品



目录

HUAWEI

华为云计算和大数据已覆盖国内全行业客户

金融	能源&大企业	媒资	公共部门	教育
深交所	大港油田	凤凰卫视	西安铁路局	华中科技大学
香港 Infocast	国家电网	中央电视台	吉林社保	上海海事大学
中国银行	沈飞集团	新华社	广东海事局	清远职业学院
中信信托	榆林神华	广东广电	福建工商云	上海中学

HUAWEI ENTERPRISE ICT SOLUTIONS A BETTER WAY

Copyright©2012 Huawei Technologies Co., Ltd. All Rights Reserved.
The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.