



JLTECHWISE
健路网络科技

B I G D A T A

互联网大数据时代

WE HAVE MADE GREAT ACHIEVEMENTS

健路科技



目录

CONTENTS

01

5G通讯数据源介绍

Large data, or huge amounts of data, refers to the need for new processing

02

大数据系统架构及组件

Large data, or huge amounts of data, refers to the need for new processing

03

大数据算法可视化开发基础

Large data, or huge amounts of data, refers to the need for new processing



5G通讯数据源介绍

PART
01



电信全国接入原始数据量概况

MR数据

数据量	文件个数	文件类型
30TB/天	583774	tar.gz/zip

CDR数据

数据量	文件个数	文件类型
40TB/天	1582674	tar.gz/zip/gz

感知DPI数据

数据量	文件个数	文件类型
260TB/天	3685671	tar.gz/gz

移动DPI数据

数据量	文件个数	文件类型
200TB/天	5157618	tar.gz/gz

固网DPI数据

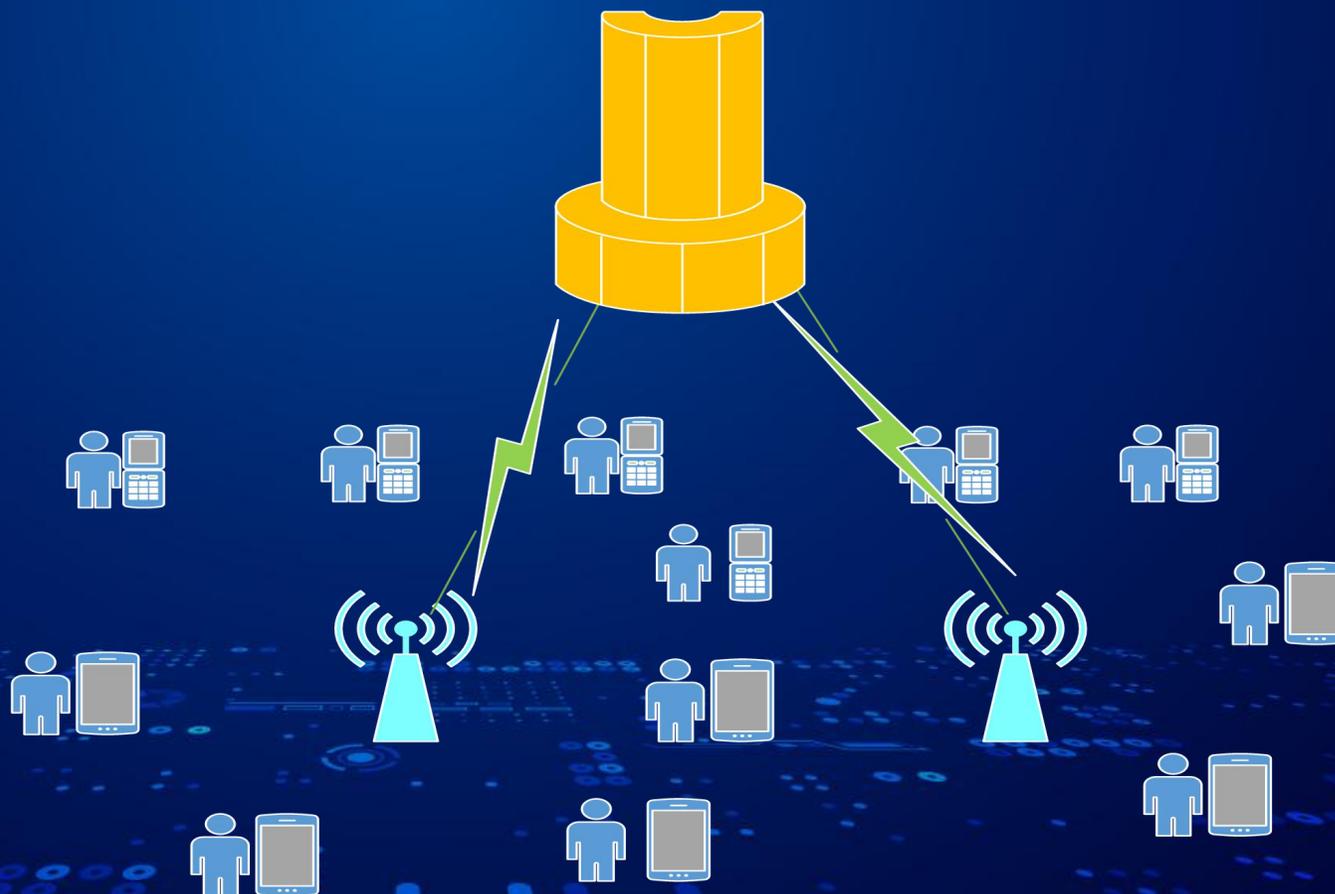
数据量	文件个数	文件类型
50TB/天	1407123	tar.gz/gz/txt

5G数据源——MR数据



JLTECHWISE
健路网络科技

MR (Measurement Report, 测量报告) 是指信息在业务信道上每480ms (信令信道上470ms) 发送一次数据, 这些数据可用于网络评估和优化。



5G数据源概览



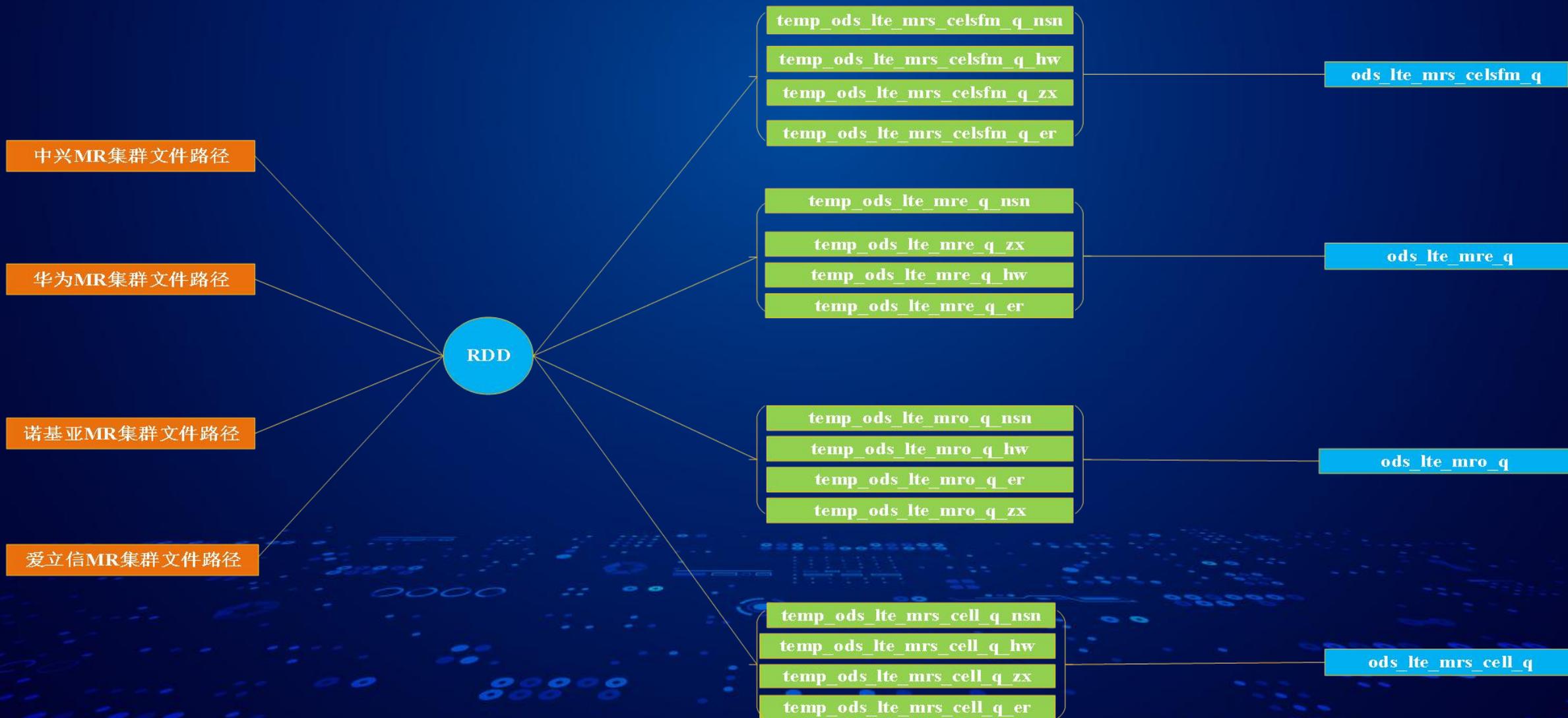
JLTECHWISE
健路网络科技



5G数据源——MR数据



JLTECHWISE
健路网络科技



5G数据源——MR数据



序号	参数	参数类型	说明及书写规范
1	datetime	string	测量时间, 示例: 2019-06-02T04:16:00.000
2	enodebid	int	基站ID
3	cellid	int	小区ID
4	imsi	string	国际移动用户识别码
5	ulsinr	smallint	上行SINR
6	phr	smallint	Power Headroom Report, 发射功率余量
7	aoa	smallint	仅支持: TDD 8天线, 其他模式暂不支持
8	recvbsr	bigint	后台显示名为RecvBsr,上报周期内BSR平均值, 多个承载时选择最大值上报, CMAC上报该值
9	puschprbnum	int	PUSCH PRB采样实际使用个数
10	pdschprbnum	int	PUSCH PRB采样实际使用个数
11	rsrp	smallint	服务小区rsrp
12	dlearfcn	int	服务小区中心频点
13	pci	smallint	服务小区pci
14	rsrq	tinyint	服务小区rsrq
15	ncellcount	tinyint	邻区个数
16	neighcellarfcn	array<int>	邻区频点序列
17	neighpci	array<int>	邻区PCI序列
18	neighcellrsrp	array<int>	邻区RSRP序列
19	neighcellrsrq	array<int>	邻区RSRQ序列
20	ta	int	TUeCmacLocation.swTaDvType1 V3.4之后开启TAResult开关, 输出TUeCmacLocation.TAResult/16

5G数据源——MR数据



序号	参数	参数类型	说明及书写规范
21	cqi	tinyint	中兴：单流CQI或双流中流1的CQI上报值；其他厂商取CQI。取值范围为[0,15]
22	mmecode	tinyint改为smallint	字符串，内容为8位比特，用于MME组内的MME标识
23	mmes1apid	bigint	MME UE S1应用程序标识，用于UE的标识，32位比特
24	mmegroupid	int	MME组的标识，16位比特
25	plmn	string	UE接入E-utran网络使用的PLMN号
26	mdn	string	号码
27	mrlon	double	MR经度
28	mrlat	double	MR纬度
29	neighcellenbid	array<int>	邻区基站ID，数组
30	neighcellid	array<int>	邻区小区ID，数组
31	x_offset_original	Int	X偏移
32	y_offset_original	Int	Y偏移
33	regionid	Smallint	RegionID
34	locationtype	Smallint	定位类型
35	buildingid	String	建筑物名称ID
36	buildingfloor	Int	建筑物楼层ID
37	positionmark	Tinyint	室内外位置区分结果，0：室内，1：室外，2：未知
38	imei	string	串号

呼叫细节记录(CDR)，在PSTN和VOIP系统中所使用的术语，是写给后处理活动数据库的电话呼叫记录。典型的呼叫细节记录包括：呼叫的始发和目的地址、呼叫开始和结束的时间、呼叫的时间长度、打电话的日期、通过网络支付的电话费或服务费以及其他的呼叫细节。CDR主要用于记帐和网络分析。



5G数据源——CDR数据



序号	参数	参数类型	说明及书写规范
1	cdr_enodebid	int	eNB标识
2	cdr_stmsi	bigint	临时UE识别号
3	cdr_mmeues1apid	bigint	UE在MME内的临时标识MmeUeS1APID
4	cdr_enbues1apid	int	UE在eNB内的临时标识eNBUEs1APID
5	cdr_callid	string	呼叫参考号, 在eNB内唯一标识一次用户RRC连接。
6	cdr_traceid	string	跟踪参考号, 在eNB内唯一标识一次用户RRC连接。
7	cdr_starttime	string	生成CDR记录的开始时刻时间戳。
8	cdr_endtime	string	生成CDR记录的结束时刻时间戳。
9	cdr_type	tinyint	生成CDR的类型
10	ac_rrcstarttime	string	UE接入时,RRC连接建立的起始时间戳。
11	ac_rrcestablishmentcause	tinyint	UE随机接入的原因值。
12	cdr_mmecode	smallint	网络的MMEC, 用于标识UE接入时使用的MME池内具体的MME设备。
13	cdr_mmegroupid	int	网络的MME Group ID, 用于标识UE接入时使用的MME池。
14	ac_cellid	int	UE接入时占用的小区标识。
15	ac_pci	int	UE接入时,LTE服务小区的的PCI (物理小区标识)。
16	ac_tai	string	UE接入小区所属的跟踪区标识。
17	ac_calltime	int	UE接入时呼叫建立时长。
18	ac_rrcfailurecause	bigint	UE接入时,RRC连接建立失败的原因值。
19	ac_rrcendtime	string	RRC连接建立的结束时间戳。
20	rls_callreleasetime	string	UE呼叫释放时的时间戳。
21	rls_cellid	int	UE呼叫释放时的小区标识。
22	rls_pci	int	UE呼叫释放时,LTE服务小区的PCI。
23	rls_tai	string	UE接入小区所属的跟踪区标识。
24	rls_callreleasecause	tinyint	UE呼叫释放的原因值。

5G数据源——CDR数据



序号	参数	参数类型	说明及书写规范
25	ho_first_type	tinyint	UE第一次切换的类型。
26	ho_first_srcenbid	int	UE第一次切换时源小区的EnodebID。
27	ho_first_srccellid	int	UE第一次切换时源小区的CellID。
28	ho_first_srculfreq	int	UE第一次切换时源小区上行中心频率信道号。
29	ho_first_srcdlfreq	int	UE第一次切换时源小区下行中心频率信道号。
30	ho_first_srcrsrp	smallint	UE第一次切换时源小区的参考信号接收功率。
31	ho_first_srcrsrq	tinyint	UE第一次切换时源小区的参考信号接收质量。
32	ho_first_tgttenbid	int	UE第一次切换时目标小区的EnodebID。
33	ho_first_tgtcellid	int	UE第一次切换时目标小区的CellID。
34	ho_first_tgtpci	smallint	UE第一次切换时目标E-UTRAN小区的的PCI（物理小区标识）。
44	ho_last_type	tinyint	UE最后一次切换的类型。
45	ho_last_srcenbid	int	UE最后一次切换时源小区的EnodebID。
46	ho_last_srccellid	int	UE最后一次切换时源小区的CellID。
47	ho_last_srculfreq	int	UE最后一次切换时源小区上行中心频率信道号。
48	ho_last_srcdlfreq	int	UE最后一次切换时源小区下行中心频率信道号。
49	ho_last_tgttenbid	int	UE最后一次切换时目标小区的EnodebID。
50	ho_last_tgtcellid	int	UE最后一次切换时目标小区的CellID。
51	ho_last_tgtpci	int	UE最后一次切换时目标E-UTRAN小区的的PCI（物理小区标识）。

DPI (Deep packet inspection, 深度数据包检测) 所谓“深度”是相对“普通的报文解析”而言的。普通报文检测仅会分析数据报文的 IP 五元组 (源地址、目的地址、源端口、目的端口以及协议类型)。

而 DPI 除了支持上述 L2-L4 的报文首部解析之外, 还增加了对 L7 应用层有效载荷 (Payload) 的解析, 可以识别各种应用类型及其内容。

DPI集群文件路径

RDD

fact_pscp_slmme_front

fact_psup_dns_e2e_front

fact_psup_video_e2e_front

fact_psup_http_e2e_front

应用分析: 网络流量构成分析、性能分析、流向分析等;

用户分析: 用户群区分、行为分析、终端分析、趋势分析等;

网元分析: 根据区域属性 (市、区、街道等)、基站负载情况进行分析等;

流量管控: P2P 限速、保证 QoS、带宽保障、网络资源优化等;

安全保障: DDoS 攻击、数据广播风暴、防范恶意病毒攻击等。

针对特定 App 的流量统计: 以微信举例, 主要是做了针对性的协议监控

5G数据源——DPI数据



字段名	类型	字段含义
interface	tinyint	必填字段。接口类型:
xdr_id	string	必填字段。采集解析服务器内唯一的xDR编号。
imsi	string	用户IMSI。
imei	string	终端IMEI/IMEISV。保持信令抽取的原样, 无需进行IMEI和IMEISV间的转化。在eHRPD场景下, 使用MEID填充。
msisdn	string	用户号码。
home_province	string	用户的归属地省份, 枚举定义参见附录G。对国际漫入用户时, 该字段填写用户归属运营商PLMN。
procedure_starttime	bigint	业务流程开始时间, UTC时间, 从1970/1/1 00:00:00开始到当前的毫秒数。
procedure_endtime	bigint	业务流程结束时间, UTC时间, 从1970/1/1 00:00:00开始到当前的毫秒数。
procedure_status	smallint	流程状态 0: 成功; 1: 失败; 255: 超时
request_cause	bigint	流程的请求消息中所包含的Cause值
failure_cause	bigint	流程失败响应消息中的值
mme_ue_s1apid	bigint	MME UE S1AP ID
user_ipv4	string	终端用户的IPv4地址。
user_ipv6	string	终端用户的IPv6地址。
ecgi	string	输出为" MCC-MNC-ECI"格式, 每段均要求按照Integer类型输出。如: 460-11-123456
apn	string	APN
other_tac	bigint	对端小区的TAC, 只在切换过程填写
other_eci	bigint	对端小区的ECI, 只在切换过程填写
mmegroupid	bigint	当前MME组标识
mmecode	bigint	当前MME号
enbid	bigint	eNodeB ID
cellid	string	Cell ID

5G数据源——DPI数据



序号	参数	参数类型	说明及书写规范
1	imsi	string	用户的国际移动用户识别码
2	msisdn	string	用户手机号码, 即MDN
3	imei	string	移动台设备标识, UTF8String 格式, 为一个15 位的十进制数字
4	apn	string	网络接入标识, 用户的帐号, 标识用户名和归属网络, 格式为: user@realm。
5	destinationip	string	用户访问的目标 IPv4或 IPv6 地址
6	destinationport	int	用户访问的目标端口号
7	sourceip	string	用户访问外部网络时使用的 IPv4 或 IPv6 地址
8	sourceport	int	用户访问外部网络时使用的端口号
9	sgwip	string	SGW 的 IPv4 或 IPv6 地址
12	ecgi	string	用户所在小区
14	host	string	主机名根据请求包 HOST 字段提取
15	rattyp	tinyint	1: UTRAN 接入; 6: EUTRAN 接入
16	protocolid	int	协议类型
17	servicetype	string	业务类型
18	starttime	string	业务流开始时间, 格式为 yyyyymmddhhmmss (24 小时制)
19	endtime	string	业务流结束时间, 格式为 yyyyymmddhhmmss (24 小时制)
20	duration	int	持续时间, 单位毫秒根据业务流开始和结束统计得出的时间
21	inputoctets	int	发给用户的业务字节数
22	outputoctets	int	用户发出的业务字节数
23	inputpacket	int	发给用户的业务数据包数量
24	outputpacket	int	用户发出的业务数据包数量

5G数据源——DPI数据



名称	编号
即时通信	1
视频业务	2
P2P业务	3
VoIP业务	4
游戏业务	5
邮件服务	6
电信自营业务	7
阅读学习	8
微博社区	9
地图出行	10
音乐	11
应用商店	12
网上商城	13
影像处理	14
直播业务	15
财经支付	16
云盘服务	17
安全杀毒	18
浏览下载	19
VPN类应用	20
WAP类应用	21
其他网络应用	22

业务大类编号	servicetype	业务大类名称	业务小类编号	业务小类名称	业务细分名称
01	1000101	即时通信	0001	QQ	QQ聊天pc端
01	1000102	即时通信	0001	QQ	网页QQ
01	1000103	即时通信	0001	QQ	移动QQ
01	1000104	即时通信	0001	QQ	qq中转站
01	1000105	即时通信	0001	QQ	qq传文件
01	1000107	即时通信	0001	QQ	QQ秀
01	1000108	即时通信	0001	QQ	QQ语音
01	1000109	即时通信	0001	QQ	QQ视频通话
01	1000201	即时通信	0002	微信	微信摇一摇
01	1000202	即时通信	0002	微信	微信
01	1000203	即时通信	0002	微信	微信视频聊天
01	1000204	即时通信	0002	微信	网页微信
01	1000205	即时通信	0002	微信	微信接收文本聊天
01	1000206	即时通信	0002	微信	微信语音对讲
01	1000207	即时通信	0002	微信	微信传送图片
01	1000208	即时通信	0002	微信	微信传送视频
01	1000209	即时通信	0002	微信	微信实时对讲
01	1000210	即时通信	0002	微信	微信接收图片
01	1000211	即时通信	0002	微信	微信接收视频
01	1000212	即时通信	0002	微信	微信朋友圈
01	1000213	即时通信	0002	微信	微信发送文本聊天
01	1000214	即时通信	0002	微信	微信电话本语音聊天



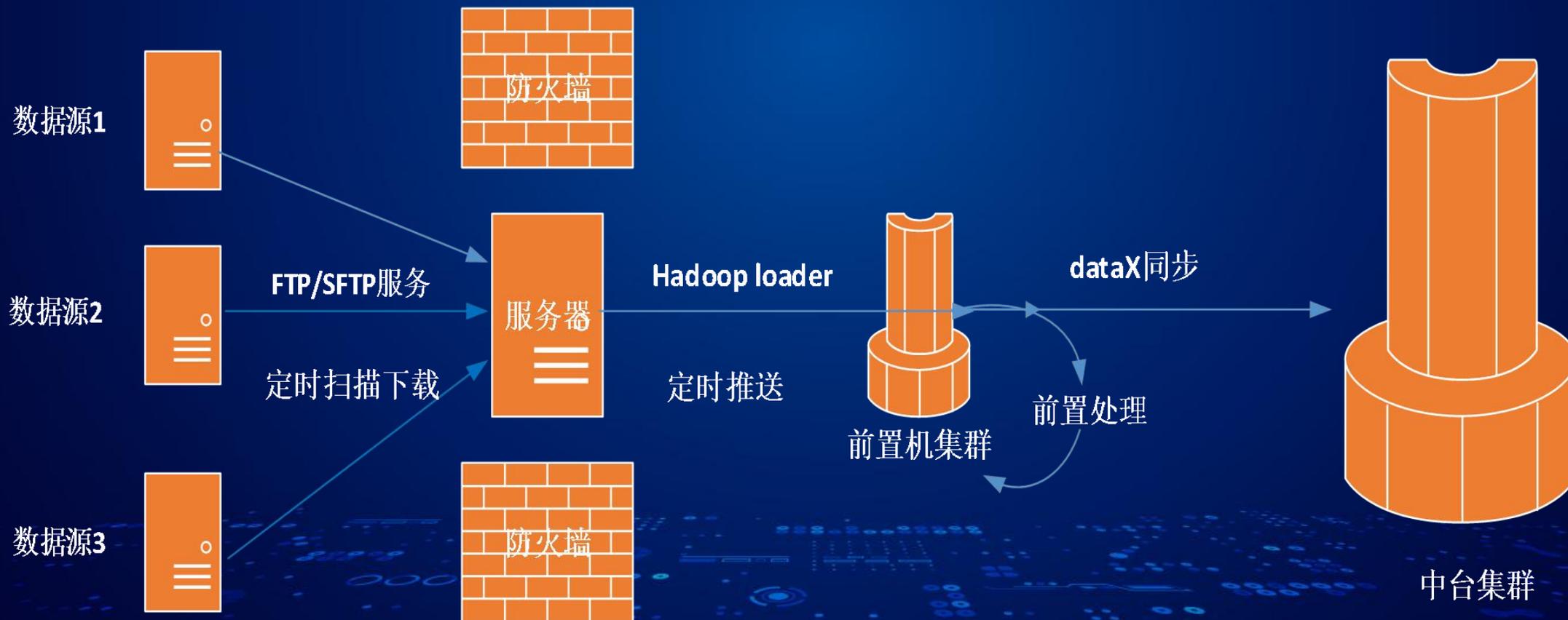
大数据系统架构及组件

PART
02

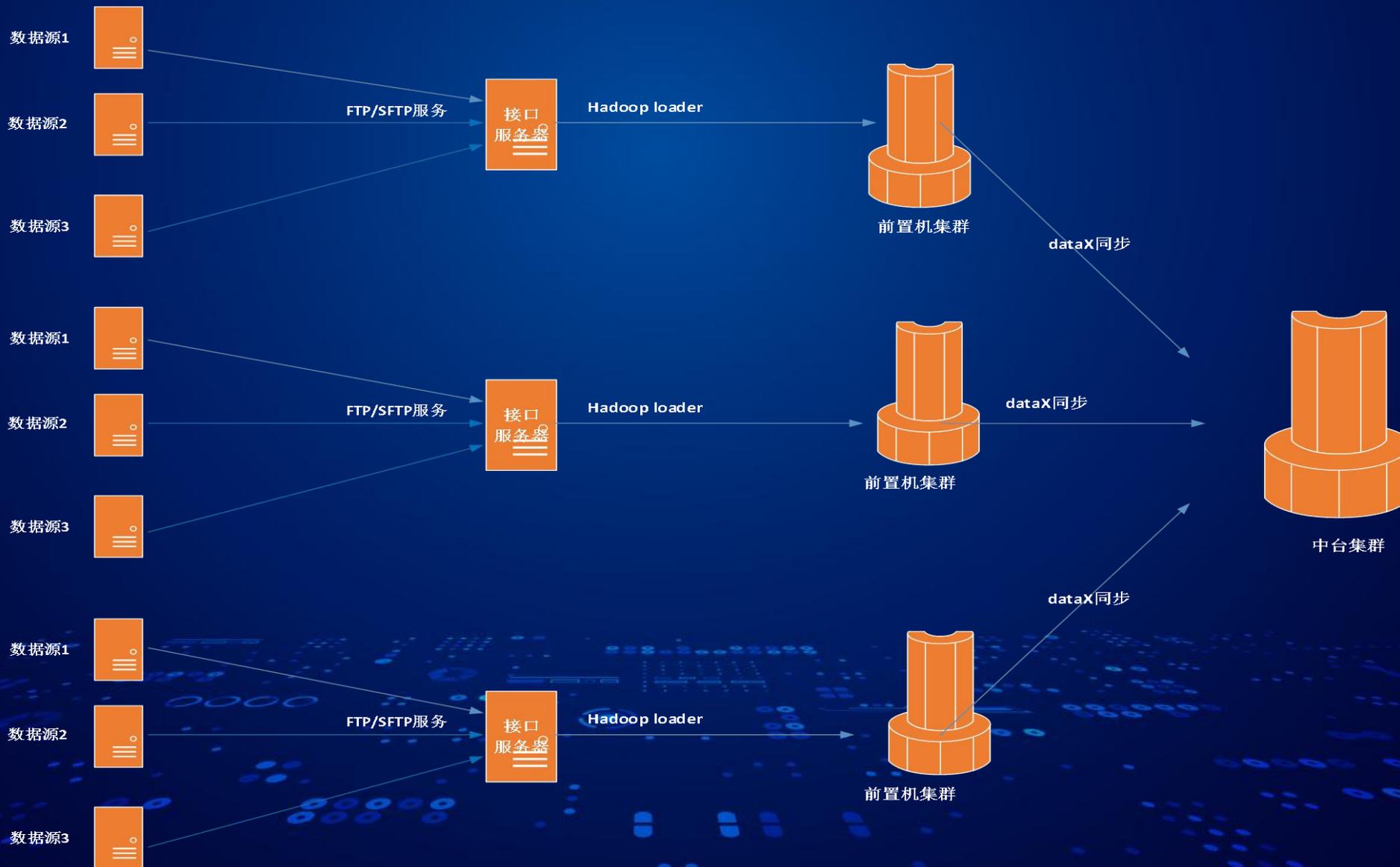
大数据架构及组件



JLTECHWISE
健路网络科技



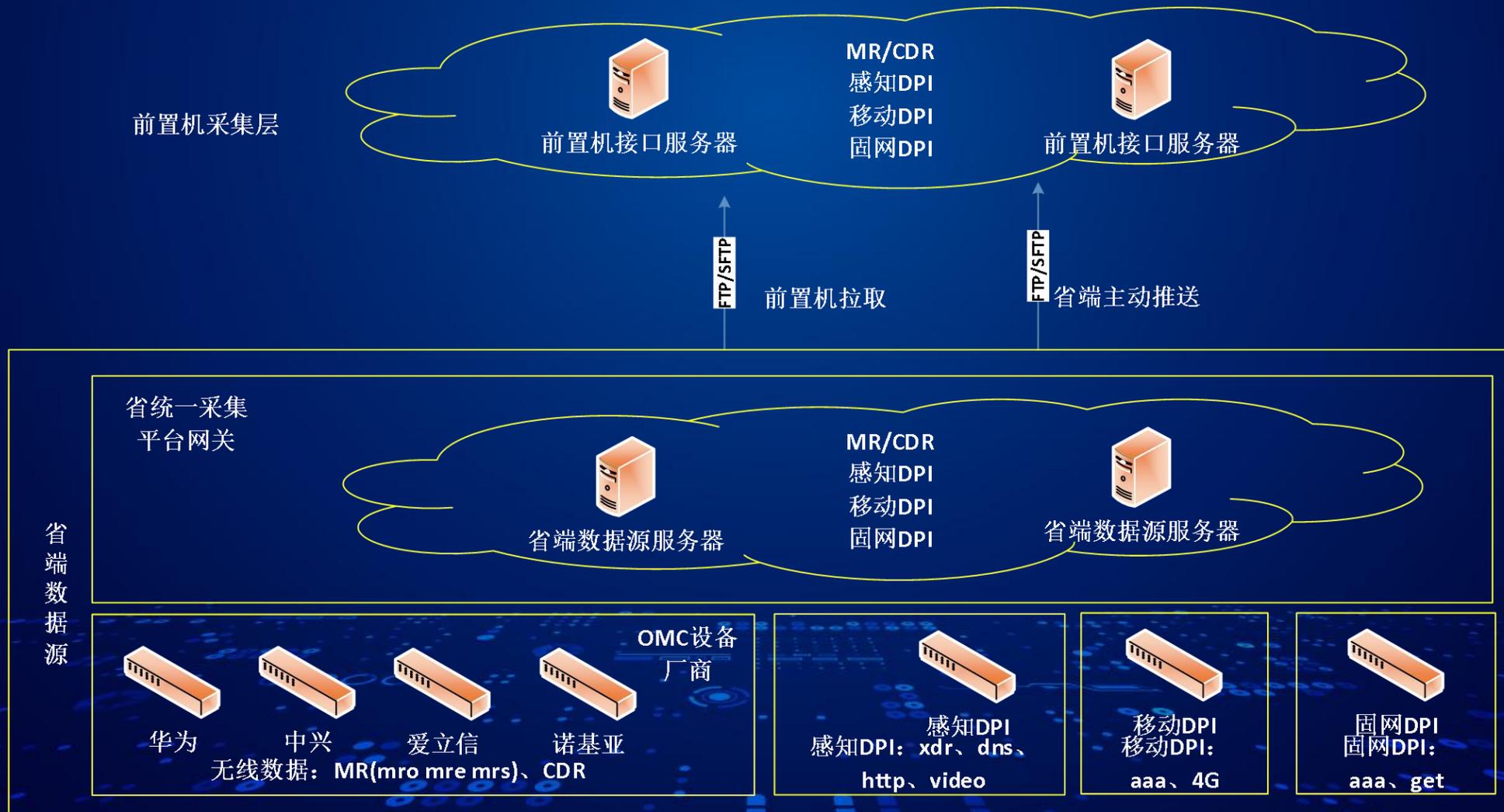
大数据架构及组件



大数据架构及组件



JLTECHWISE
健路网络科技



大数据架构及组件



JLTECHWISE
健路网络科技



大数据架构及组件



JLTECHWISE
健路网络科技

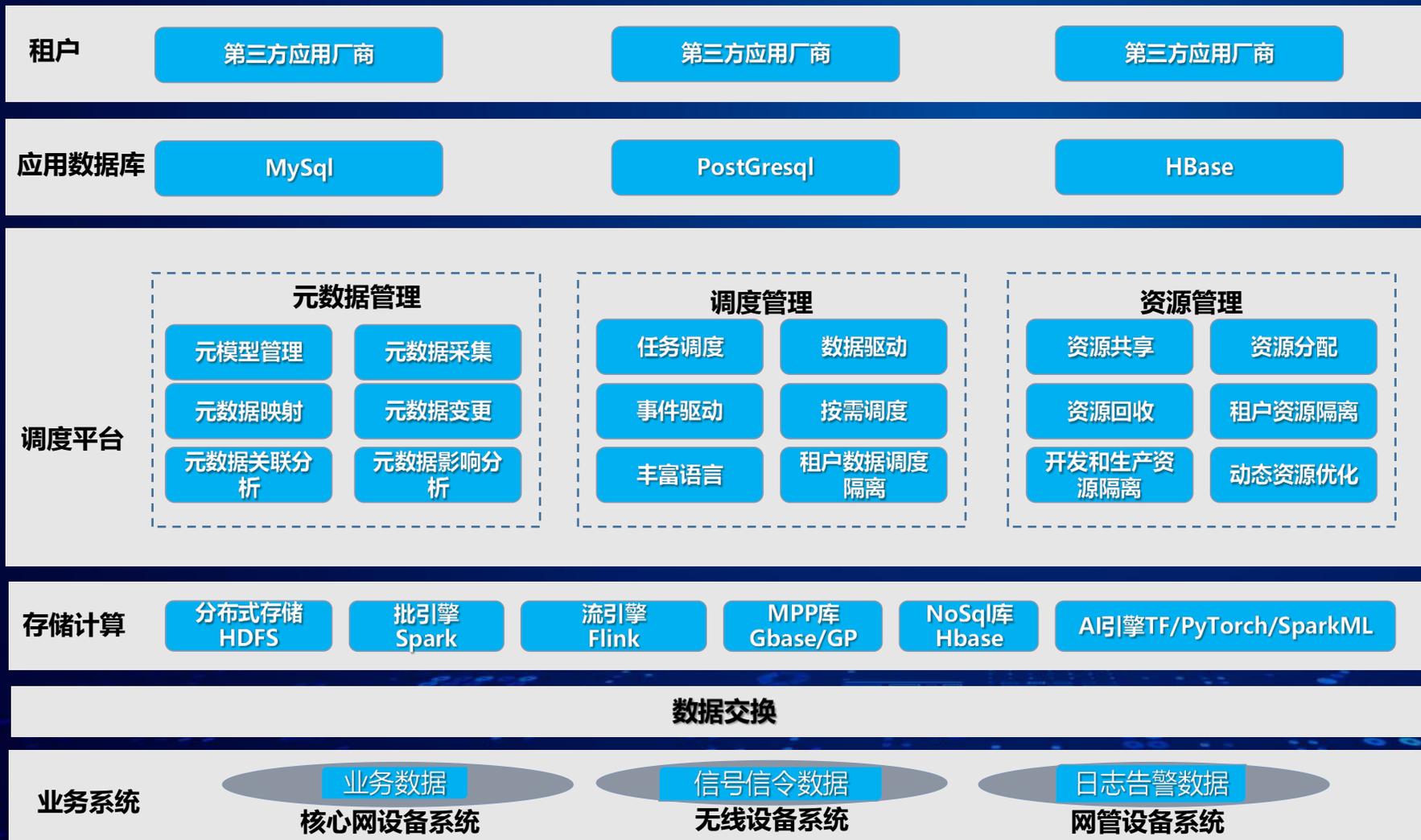
■ 云、数、智合一的开放式的共建、共享大数据平台的科研创新平台



大数据架构及组件



JLTECHWISE
健路网络科技



OLTP与OLAP融合

数据汇聚入数据湖前业务系统进行OLTP分析，入数据湖后基于调度引擎的数据清洗、融合、汇聚能力完成OLAP分析，OLAP分析后的专题数据推入应用数据库再次进行OLTP分析，从而提高数据质量

元数据系统管理

基于元数据驱动系统业务流转，基于一站式元数据管理实现系统数据360度的管理和维护

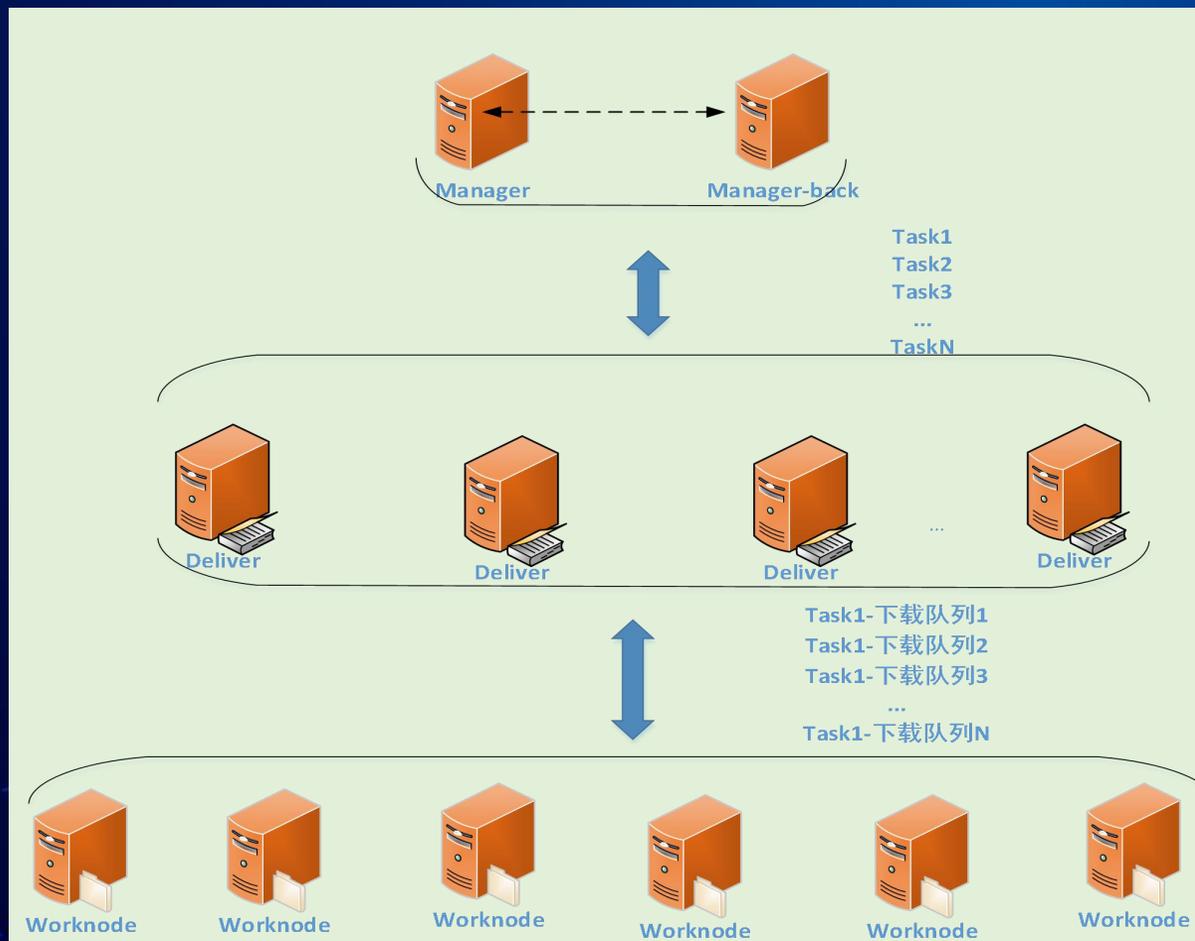
统一调度管理和资源管理

丰富的批处理、实时流、AI调度引擎；丰富算法类型和业务逻辑编排，满足多种业务场景需求

统一租户管理

租户间数据、调度队列、资源全隔离，同时提供租户间数据共享通道

□ 主备，分布式架构，高并发、高可靠、高扩展



■ Manager

解析数据任务配置信息，获取调度周期、数据存放目录、文件名规则、是否解压、压缩、合并等信息，并将任务信息均衡分发给deliver节点

■ Deliver

接收任务信息，获取数据扫描路径，获取并过滤扫描到的数据信息，并将数据信息均衡分发给worker

■ Worker

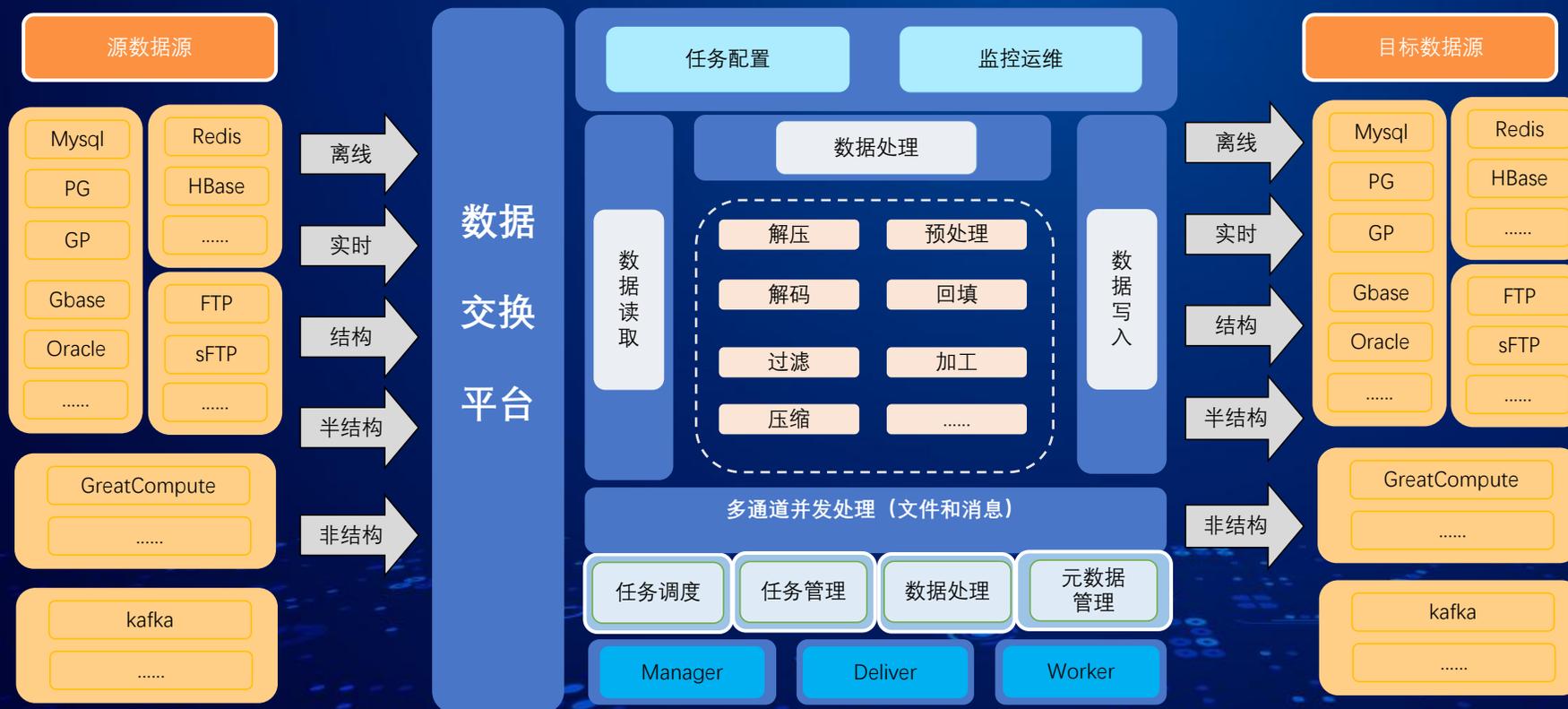
接收数据信息，实现数据下载上传

大数据架构及组件——数据交换



JLTECHWISE
健路网络科技

□ 目标：实现跨地域跨集群海量多种数据存储系统、多种异构数据源、多文件格式、多粒度的标准化高质量离线批量、实时数据采集；



- **N+种数据存储系统**
大数据集群、关系型数据库、非关系型数据库之间的结构化数据和非结构化数据交换
- **实时和离线同步**
支持实时流、离线批量数据的全量和增量数据交换
- **高性能高可靠**
分布式部署、负载均衡多节点并行同步，数据不落盘
- **任务配置统一建模，一套配置多数据源复用**
一套配置支持多空间、多时间、多文件的同步，配置简单高效
- **任务数据联动监控**

大数据架构及组件——数仓构建



JLTECHWISE

构建一体化数据资源体系，实现应用和数据解耦，保证数出同源，数据的一致性、完整性



■ ODS层

数据源表

■ DWD层

数据关联、融合

■ DWS层

根据业务主题，清洗生成主题库

■ CDS层

根据业务专题，清洗生成专题库，快速支撑上层应用构建

大数据架构及组件——应用举例

- 全面掌握网络覆盖质量，多维度分析网络覆盖，支撑问题小区专项工作。弱、过、重和MOD3干扰小区由2020年1月**7.5万个**降至11月份**1.95万个**。
- 全面支撑2019/2020年校园优化专项：全国**4461**所校园场景覆盖、用户、业务负荷分析，发现感知质差区域**820**个，高负荷小区**3.3万个**。

全网质量分析

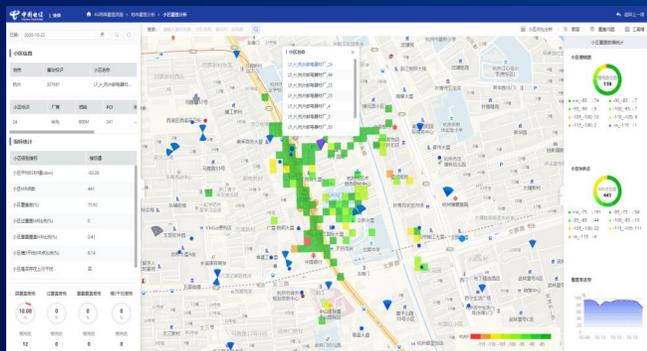
□ 整体覆盖呈现

- 全国、省级、地市、区县网络覆盖分析



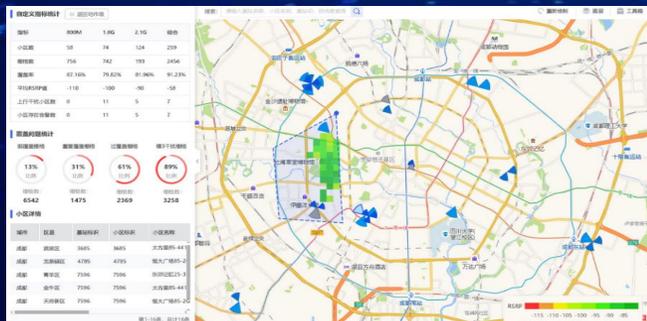
□ 覆盖问题分析

- 小区级覆盖分析;
- 栅格级覆盖分析;
- 弱、过、重叠和Mod3问题小区分析。



□ 灵活覆盖监控

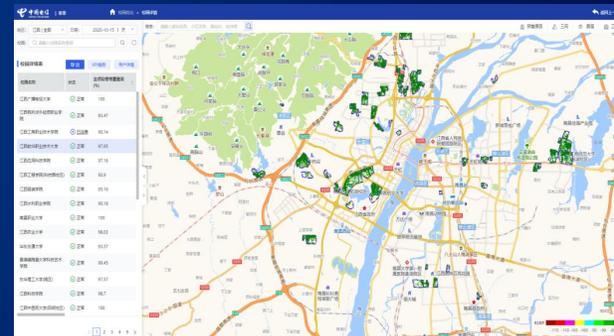
- 自定义区域分析
- 应用场景：例如马拉松、世博会、演唱会等临时重大活动可自定义区域进行覆盖分析保障。



校园区域分析

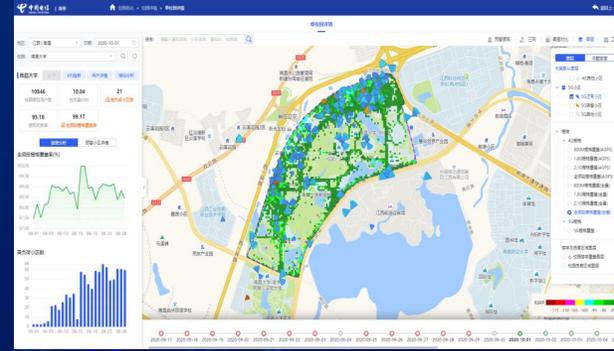
□ 区域校园概览

- 支持多维校园统计分析;
- 关键指标排名;
- 异常校园工单对接评估。



□ 单校园详情

- 负荷等关键指标历史趋势呈现;
- 校园小区、校园用户分析;
- 覆盖、感知、工参等多图层渲染灵活选择。



高速场景分析

□ 问题路段识别

- 利用AGPS MR实现对高速覆盖的监控，自动判断覆盖及感知问题路段，并生成工单。





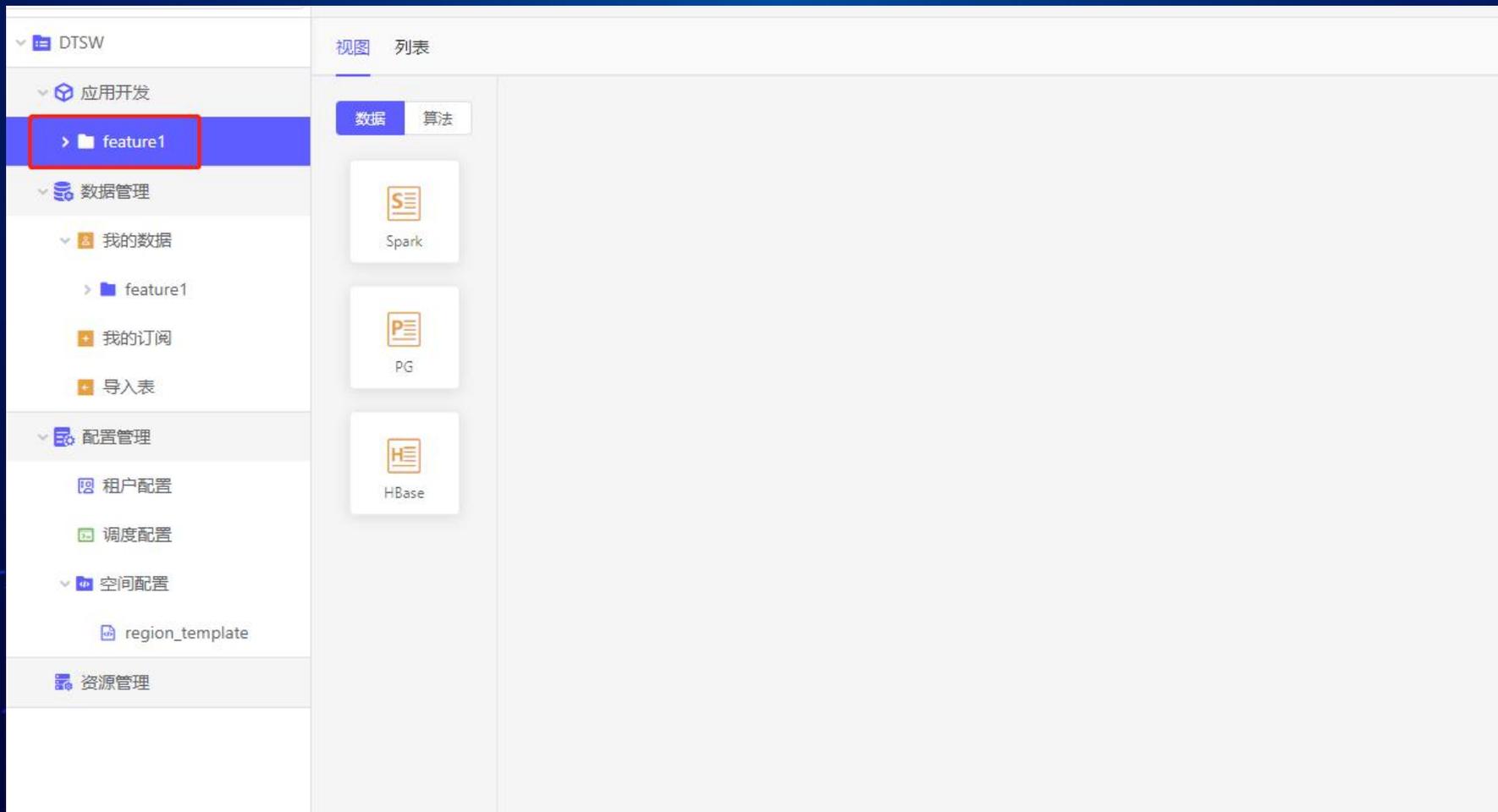
JLTECHWISE
健路网络科技

大数据算法可视化 开发基础



PART
03

支持项目按功能划分，以目录的形式进行项目功能管理，一个目录即一个功能，对应该功能的画布。
新建目录成功，就会在左侧项目树显示该级目录，单击这个目录，就会显示这个目录的画布

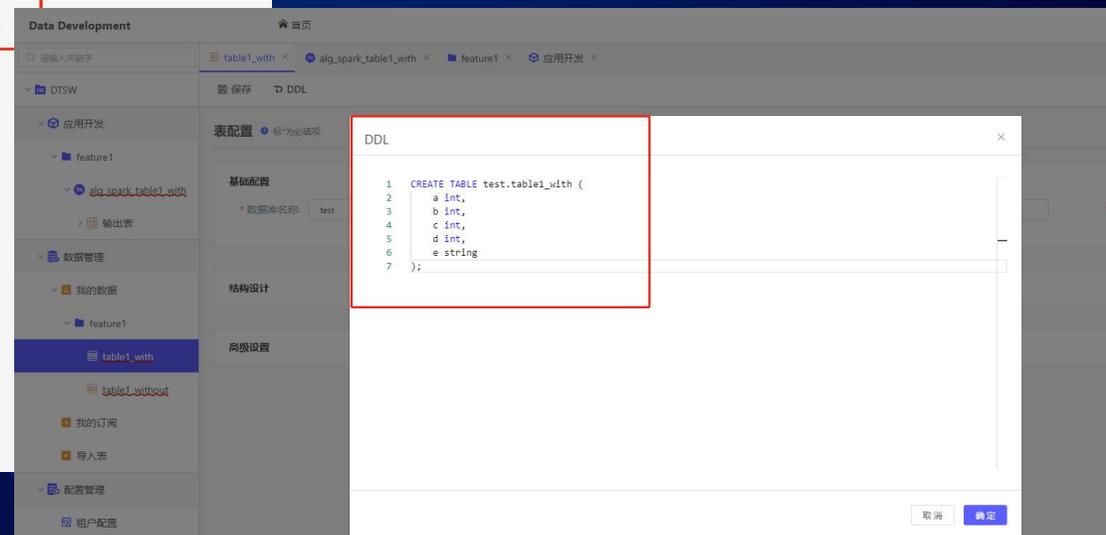
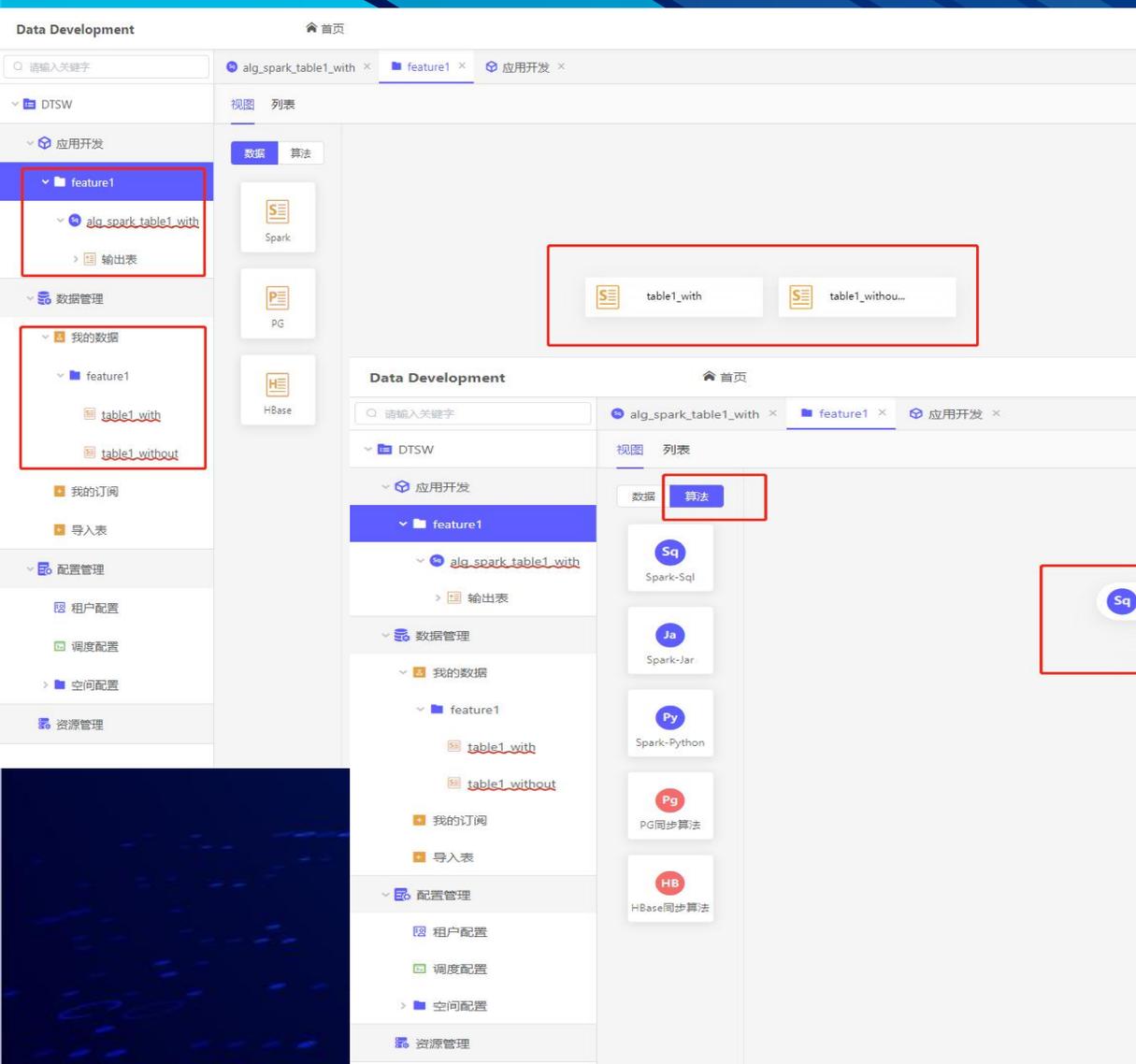


大数据算法可视化开发基础



JLTECHWISE
健路网络科技

双击算法结点或在左侧项目点击算法，会进入到算法开发页面。
对于spark-sql类的算法，算法开发页面有sql脚本实现开发，
算法配置和资源配置两个tab页面。点击算法配置





JLTECHWISE
健路网络科技



THANKS!

