

# 目录

## Contents

---

一	项目任务需求
二	工作进展
三	分工与时间计划

仅做项目展示、  
请勿外传

# 项目背景：多项目复用、技术栈亟需更新

2023年预计交付**3项**生产业务系统，**2项**科研项目系统，其它小程序调查软件若干

现状

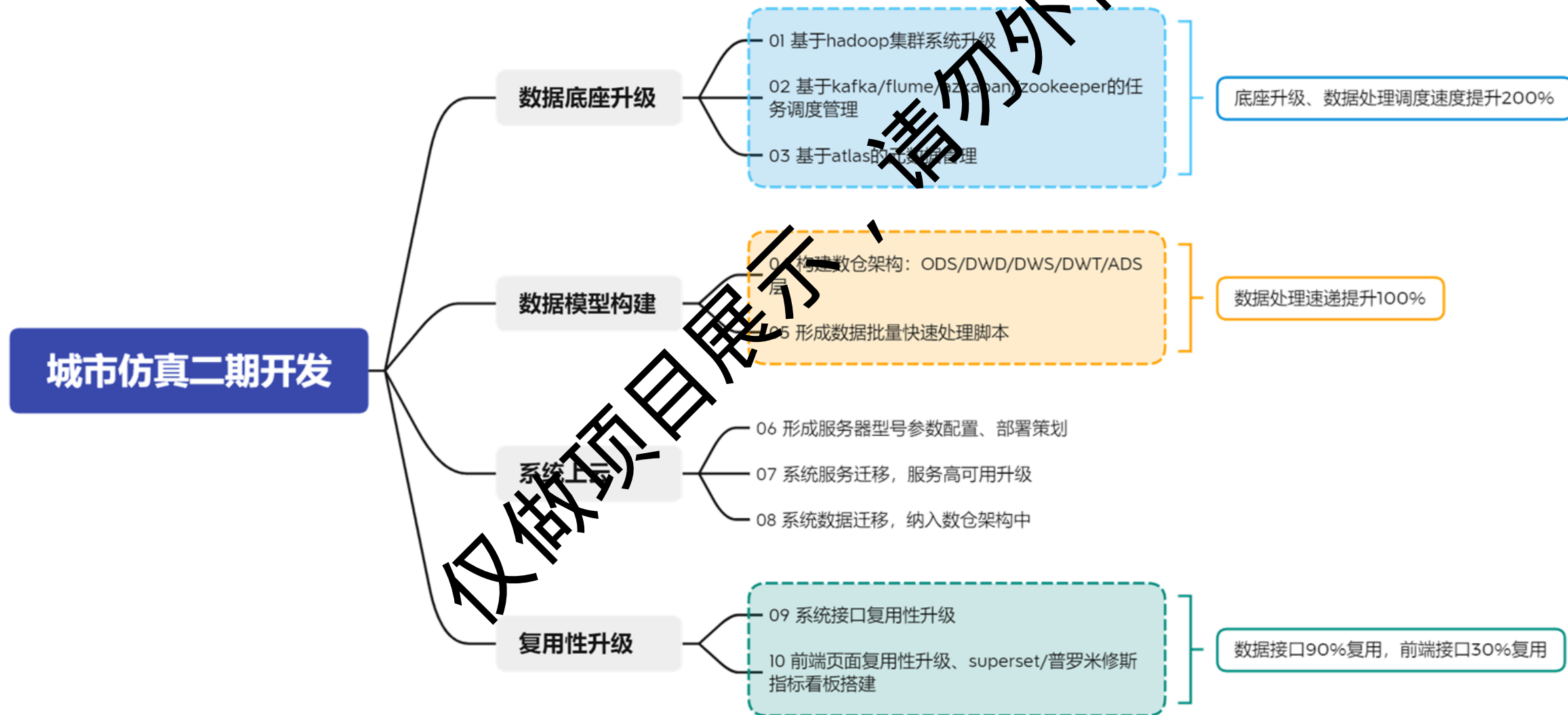
- ❑ 单节点服务器，无专人运维；
- ❑ 系统开发、测试环境混用；
- ❑ 数据存储无标准、数据运维非自动；
- ❑ 交付产品
- ❑ 数据质量、数据安全、数据更新未进行长远考虑。

预期

- ❑ 云上服务器；
- ❑ 测试与开发环境分开，权限控制；
- ❑ 数据字段定义统一，数据自动运维；
- ❑ 交付产品稳定运行，可进行权限控制与用户登录访问监控、版本管理；
- ❑ 进行开发工作接口复用性升级，前端组件复用；
- ❑ 进行数据质量、数据安全

# 项目任务书要求：数据底座产品化重构

根据项目任务书，研究内容分为四部分：**升级底座、数据建模、系统上云、复用性升级**



# 目录

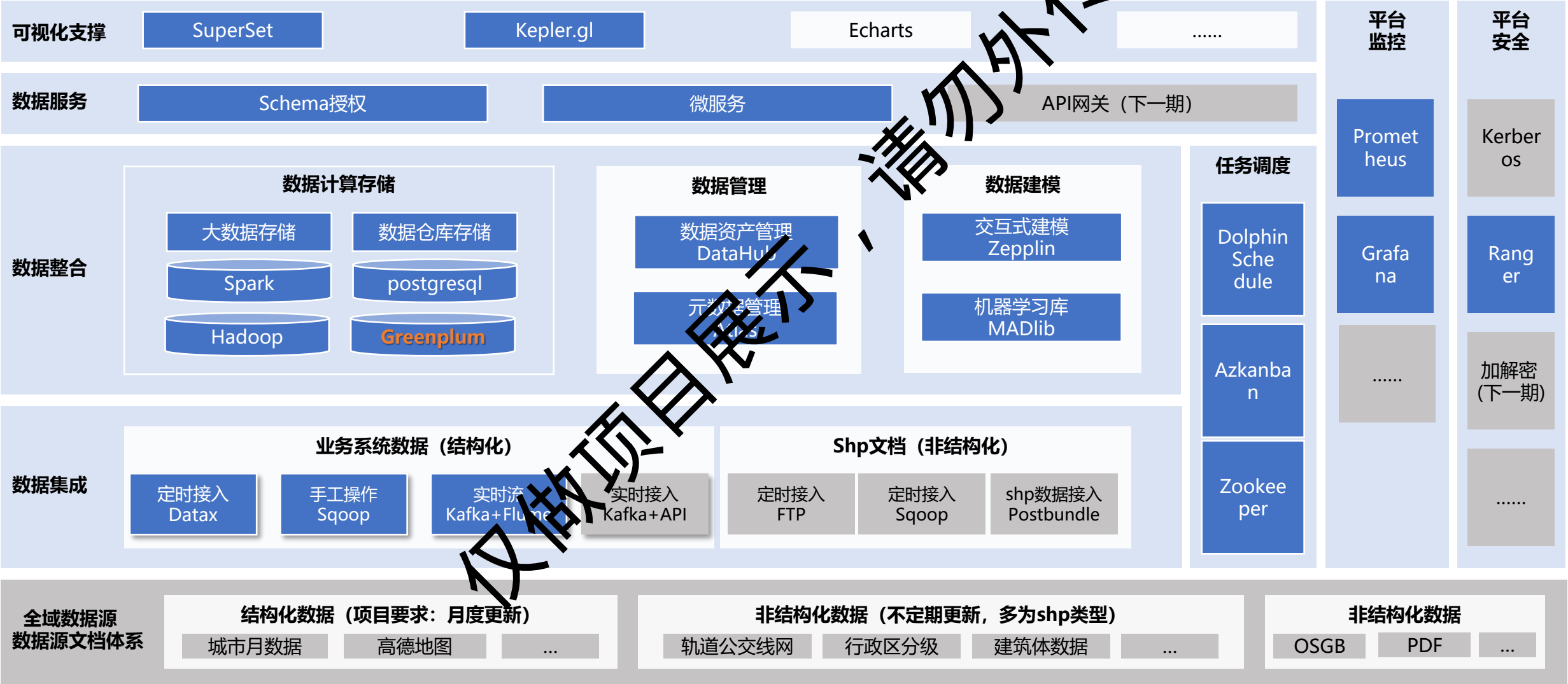
## Contents

---

一	项目任务需求
二	工作进展
三	分工与时间计划

# 系统整体架构：重存储、重调度、轻计算

设计完成了系统整体架构，通过了行业专家评审



# 技术选型：以稳定版本为主

## 本次主要技术选型

- ❑ 数据采集传输: **Flume-1.9**, **Kafka-2.4.1**, **Sqoop**, Logstash, **Datax**, FTP
- ❑ 数据存储: **Postgresql-15**, MySQL, **HDFS-3.1.3**, HBase, Redis, MongoDB
- ❑ 数据计算: **Hive-3.0.0**, Tez, **Spark**, Flink, Storm
- ❑ 数据查询: Presto, Kylin, Impala, Druid, ClickHouse, Doris
- ❑ 数据可视化: Echarts, **Superset**, QuickBI, DataV
- ❑ 任务调度: **Azkaban-3.8.4**, Oozie, **DolphinScheduler**, Airflow
- ❑ 集群监控: Zabbix, **Prometheus**, Grafana
- ❑ 元数据管理: **Atlas-2.0**
- ❑ 权限管理: **Ranger-2.0**, Sentry
- ❑ 数据资产管理: **DataHub**

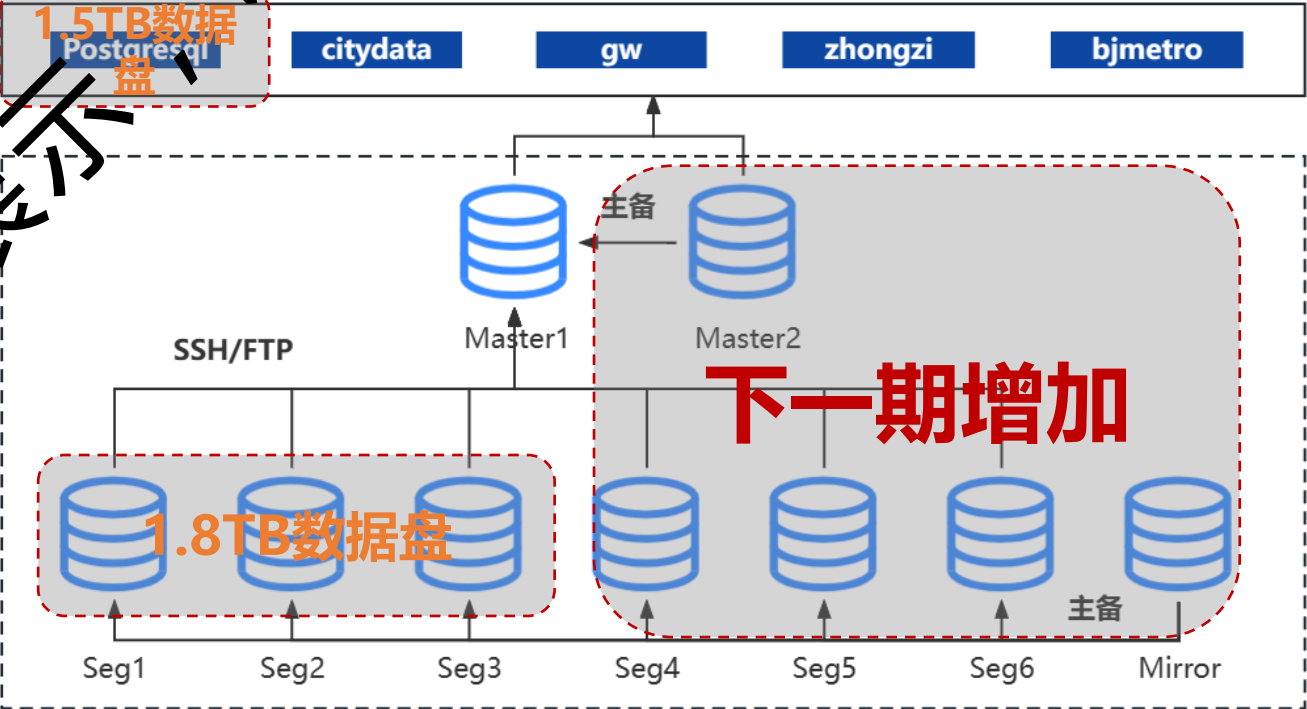
# 服务器选型：服务器轻量化瘦身

本次主要服务器配置，考虑数据时效性与可靠性要求，数据节点由6个减少到3个，删除镜像部署

产品名称	规格	描述	数量
云服务器 ECS	ecs.g6.large	华北3 通用型型 8核 32G 系统盘 ESSD 100G 无带宽	4
云服务器 ECS	ecs.g6.large	华北3 通用型4核 16G 系统盘 ESSD 100G 无带宽	6
云服务器 ECS	ecs.r6.2xlarge	华北3 内存型8核 64G 系统盘 ESSD 200G 无带宽	1
云服务器 ECS	ecs.c6.2xlarge	华北3 计算型16核 32G 系统盘 ESSD 100G 系统盘ESSD: 500g无带宽	1
云服务器 ECS	ecs.c6.2xlarge	华北3 计算型4核 16G 系统盘 ESSD 100G数据盘ESSD500G 无带宽	3
数据库	PostgreSQL 15.0	华北3 12核48G独享型 ESSD PL1 云盘 1500GB	1
负载均衡 SLB	标准	华北3	1
弹性公网IP	20M	华北3 按量	1
共享流量包	10TB	华北3 亚太全时	1
云安全中心	企业版	日志分析: 600GB防勒索病毒: 750GB	15

阿里云配置

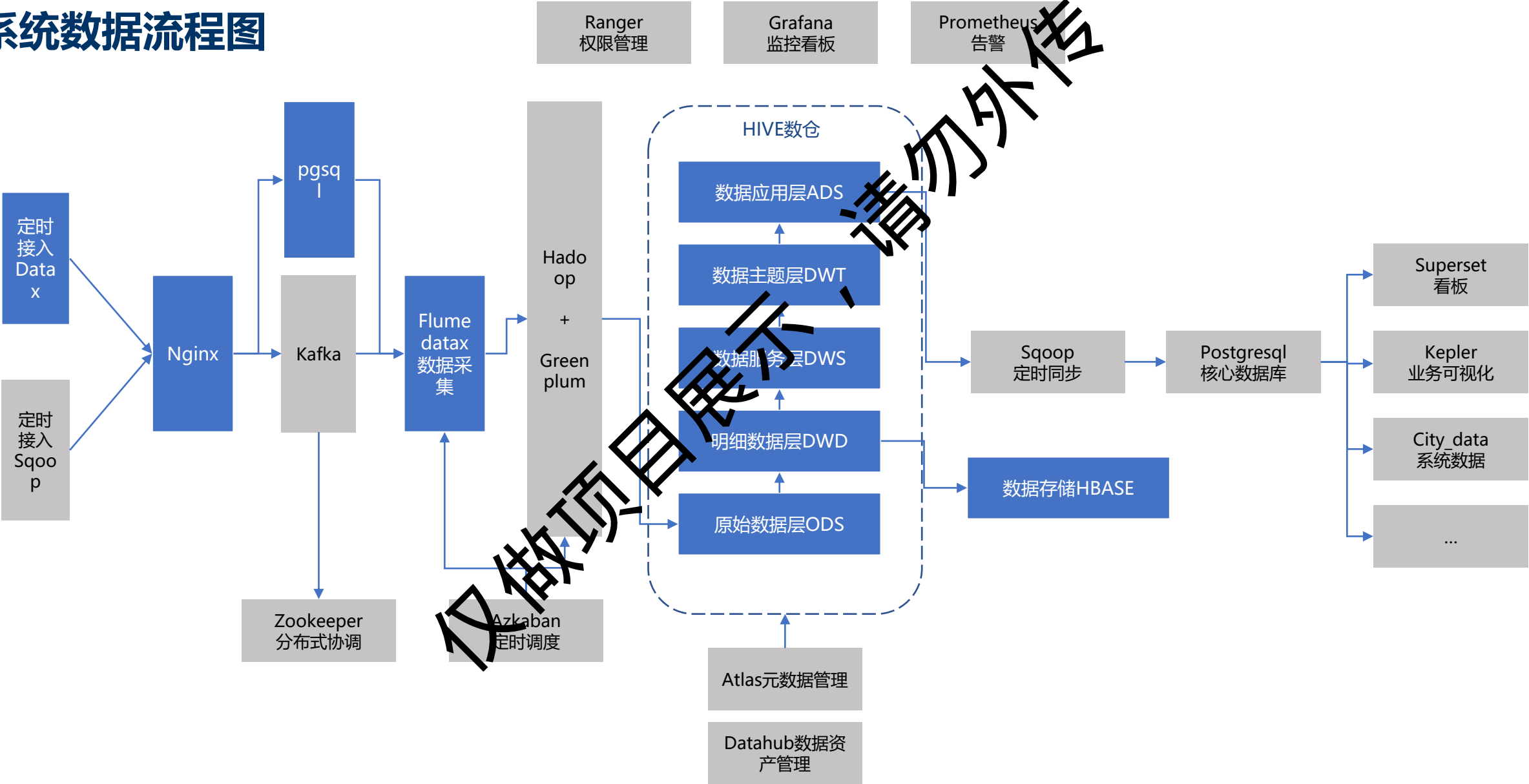
目前DELL生产环境 400GB，测试环境0.8TB（包含部分生产环境数据），未来数据更新300GB以内，20%冗余。2.0TB满足需求，master节点16核，32GB，与目前DELL物理机性能接近，满足需求



物理部署图

# 数据流图

系统数据流程图





# 集群资源策划

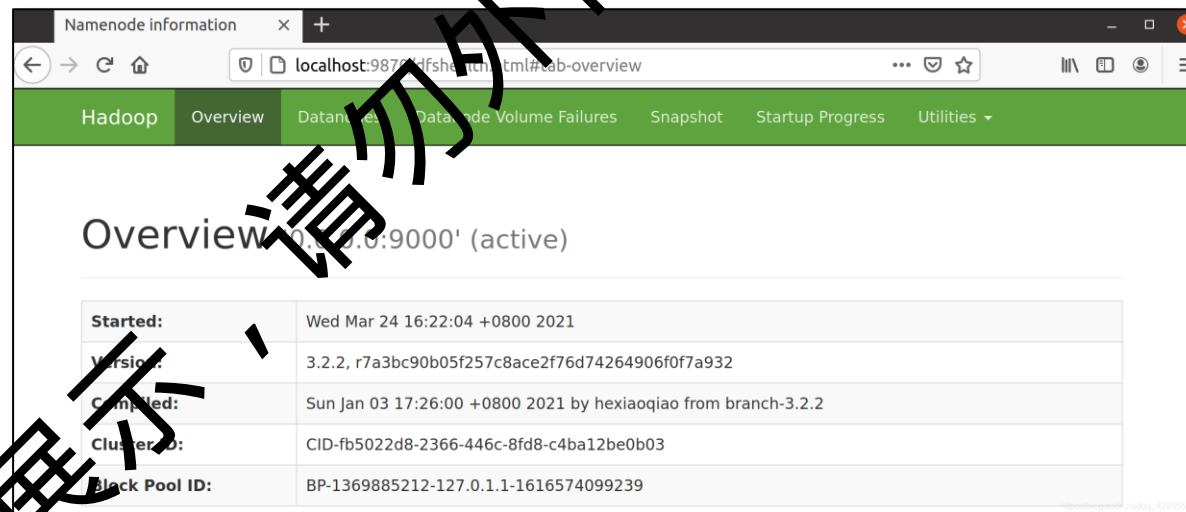
- (1) 消耗内存的分开，分散master节点压力；
- (2) 数据传输数据比较紧密的放在一起（Kafka 、Zookeeper）；
- (3) 客户端在master节点，方便外部访问；
- (4) 数据调度与集群在一台机器（Hive和Azkaban Executor）

1-master	2-datanode	3-datanode	4-datanode
nn	dn	dn	dn
	rm	nm	nm
	nm		
			zk
			kafka
			Flume/sqoop
	Hbase	Hbase	
hive			
mysql			
spark			
Azkaban			

服务名称	服务	1-master	2-datanode	3-datanode	4-datanode
HDFS	NameNode	√			
	DataNode	√	√	√	√
	SecondaryNameNode				√
	NodeManager	√	√	√	√
	Resourcemanager		√	√	
Zookeeper	Zookeeper Server	√	√	√	√
Flume（采集日志）	Flume	√	√	√	
Kafka	Kafka	√	√	√	√
Flume（消费Kafka）	Flume				√
Hive	Hive	√			
MySQL	MySQL	√			
Sqoop	Sqoop	√			
Presto	Coordinator	√			
	Worker		√	√	√
Azkaban	AzkabanWebServer	√			
	AzkabanExecutorServer	√			
Spark		√			
Kylin		√			
HBase	HMaster	√			
	HRegionServer	√	√	√	√
Superset		√			
Atlas		√			
Solr	Jar	√			
服务数总计		19	8	8	8

# 工作项：01基于hadoop集群系统升级

- ❑ 服务器IP设置✓;
- ❑ 免密登录设置✓;
- ❑ 集群配置安装✓;
- ❑ 节点均衡✓;
- ❑ 数据压缩设置✓;
- ❑ Hadoop性能调优✓;
- ❑ Yarn配置✓
- ❑ 空间引擎升级; ✓



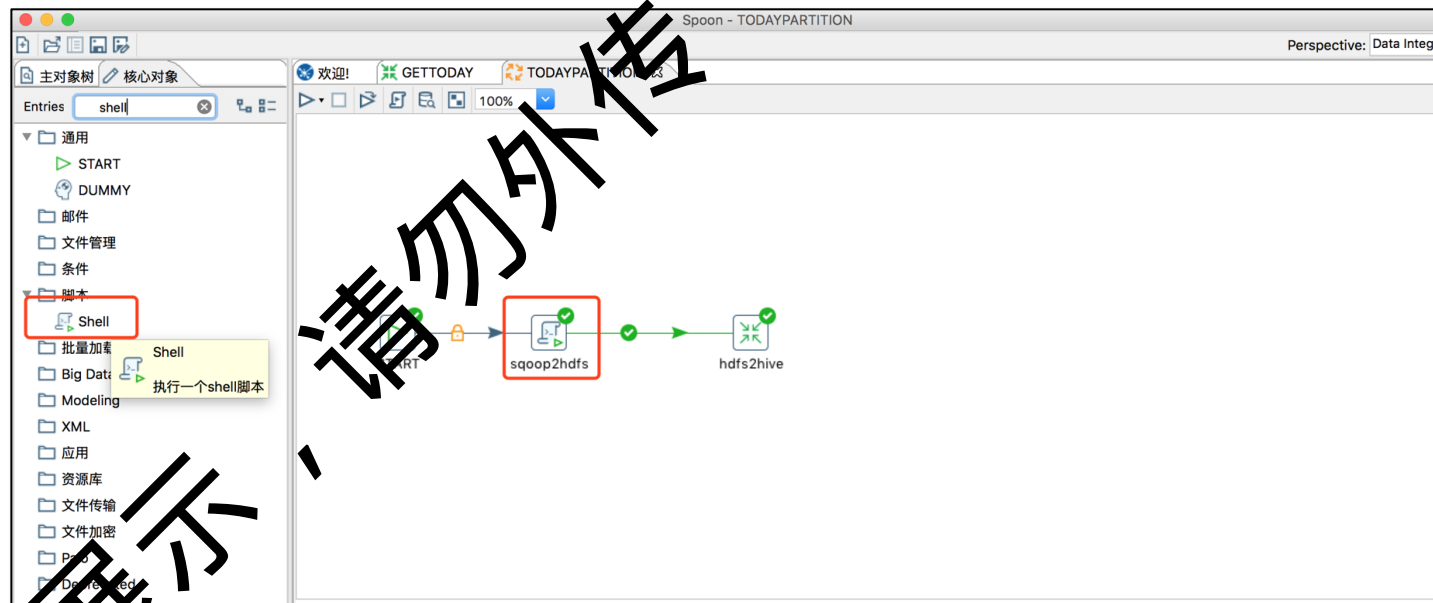
Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
✓hadoop102:9866 (172.20.31.10:9866)	http://hadoop102:9864	0s	8m	98.3 GB	0	24 KB (0%)	3.1.3
✓hadoop103:9866 (172.18.110.10:9866)	http://hadoop103:9864	2s	8m	98.3 GB	0	24 KB (0%)	3.1.3
✓hadoop104:9866 (172.20.31.40:9866)	http://hadoop104:9864	0s	8m	98.3 GB	0	24 KB (0%)	3.1.3
✓hadoop105:9866 (172.18.110.221:9866)	http://hadoop105:9864	2s	8m	98.3 GB	0	24 KB (0%)	3.1.3

Showing 1 to 4 of 4 entries

Previous 1 Next

# 工作项：02基于dolphin的数据调度

- ❑ 确定数据源接入方式与数据类型；
- ❑ Zookeeper安装✓；
- ❑ Kafka安装✓；
- ❑ Flume安装； ✕
- ❑ dataX安装； ✓
- ❑ Sqoop安装； ✕
- ❑ Kafka消费配置✓；
- ❑ Azkaban自动调度配置； ✕
- ❑ 同步策略配置



# 工作项：03基于atlas的元数据管理

- ❑ Atlas安装; ×
- ❑ Atlas配置; ×
- ❑ Atlas元数据导入; ×
- ❑ Atlas血缘分析; ×
- ❑ Datahub安装部署 (轻量化) ✓



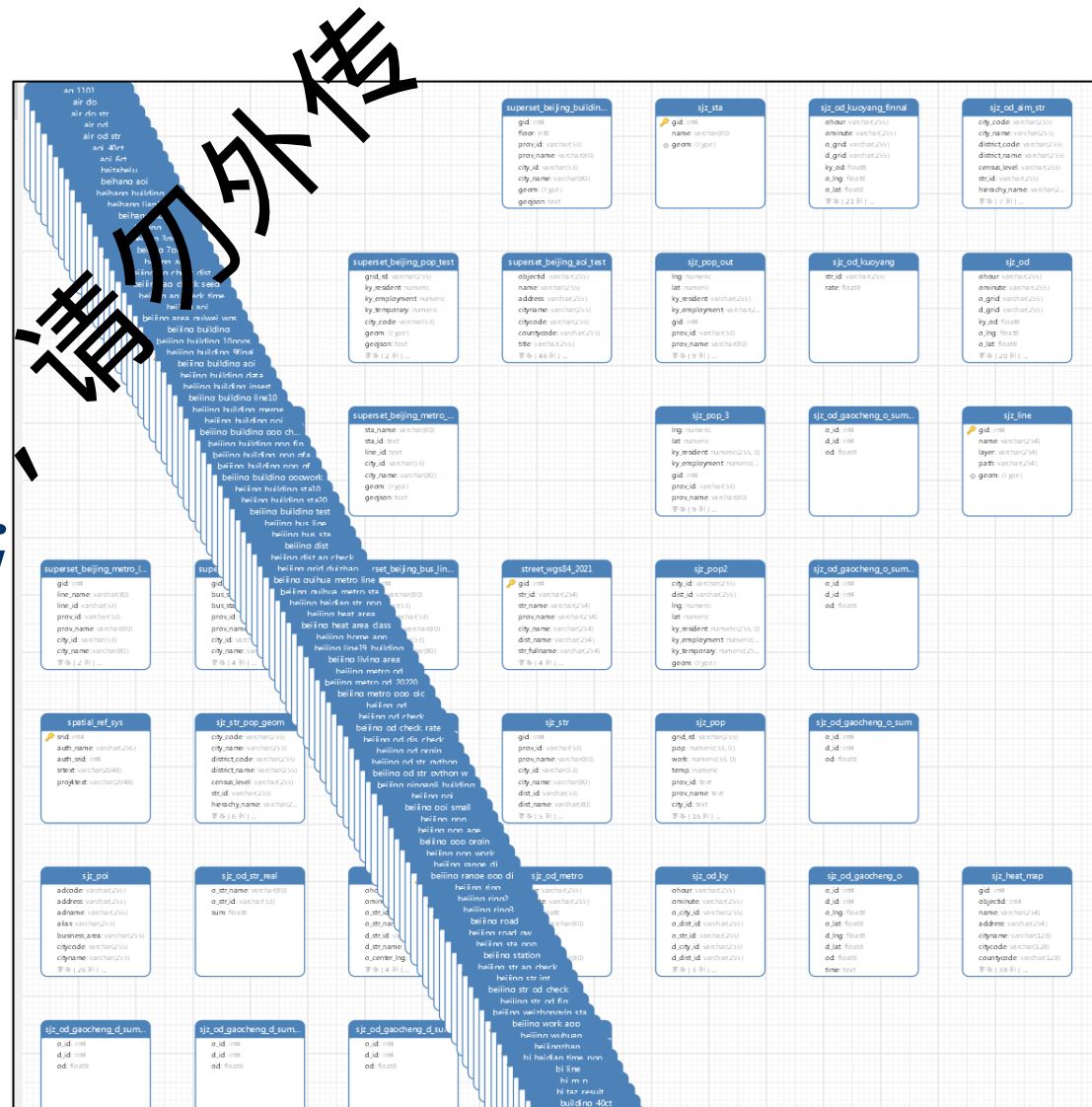
The screenshot shows the Apache Atlas web interface displaying the schema of the 'hera\_job' table. The 'Schema' tab is selected, showing a table with columns and their data types. The table is titled 'hera\_job (hive\_table)' and has a 'Term' field. The table is showing 1-25 rows.

Name	Owner	Type	Tags
op_time	hadoop	string	+
script	hadoop	string	+
timezone	hadoop	string	+
repeat_run	hadoop	tinyint	+
owner	hadoop	string	+
run_type	hadoop	string	+
host_group_id	hadoop	tinyint	+
must_end_minute	hadoop	int	+
gmt_modified	hadoop	bigint	+
group_id	hadoop	int	+
history_id	hadoop	bigint	+
last_end_time	hadoop	string	+
post_processors	hadoop	string	+
start_time	hadoop	string	+
is_valid	hadoop	tinyint	+
auto	hadoop	tinyint	+

仅做项目展示

# 工作项：04数仓模型构建

- ❑ postgresql安装✓；
- ❑ 确定数据表命名规范与字段类型✓；
- ❑ Hive数仓建模✓；
- ❑ 梳理整合层数据规范✓；
- ❑ Hive表构建（核心100个，总计300个）✓；
- ❑ 数据标准构建✓；
- ❑ 业务数据导入✓



# 工作项：05批量数据处理脚本

- 批量结构化数据导入✓；
- 批量非结构数据导入✓；
- 批量SHP数据导入✓；
- ODS-DWS ✓；
- DWS-DWT ✓；
- DWT-ADS ✓；
- 批量坐标系转换、批量建筑体人口生成、批量数据清洗等等功能性脚本✓；
- 步行可达性计算、TransCAD基础数据生成✓

仅供项目展示，请勿外传

# 工作项：06服务器部署配置 07服务迁移 08 系统迁移

- 环境配置（基础组件安装）√；
- 服务迁移（现有服务）√；
- 数据库迁移√；
- 基础组件搭建、网络配置√（Nginx负载等配置）

仅做项目展示、请勿外传

# 工作项：09系统接口复用性升级

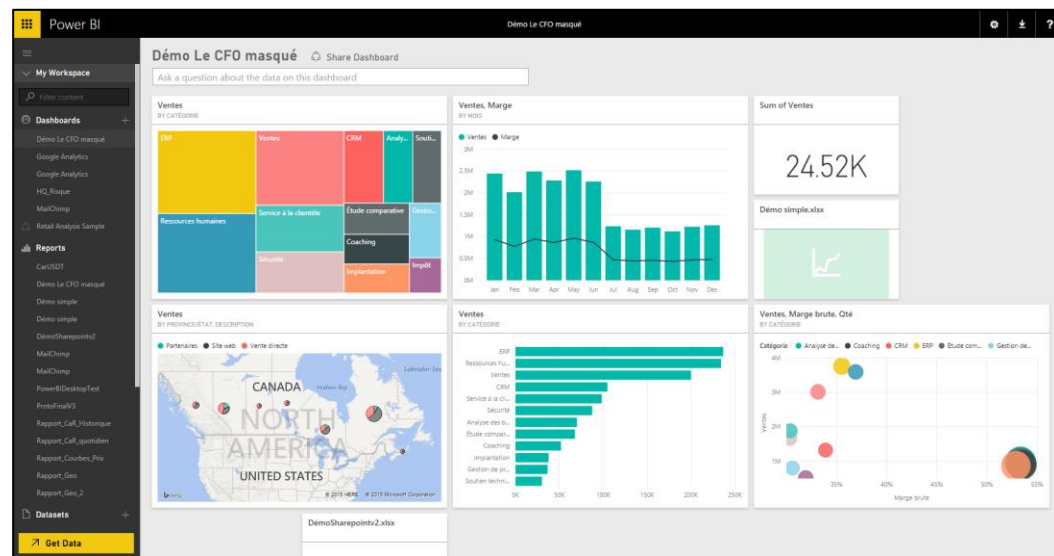
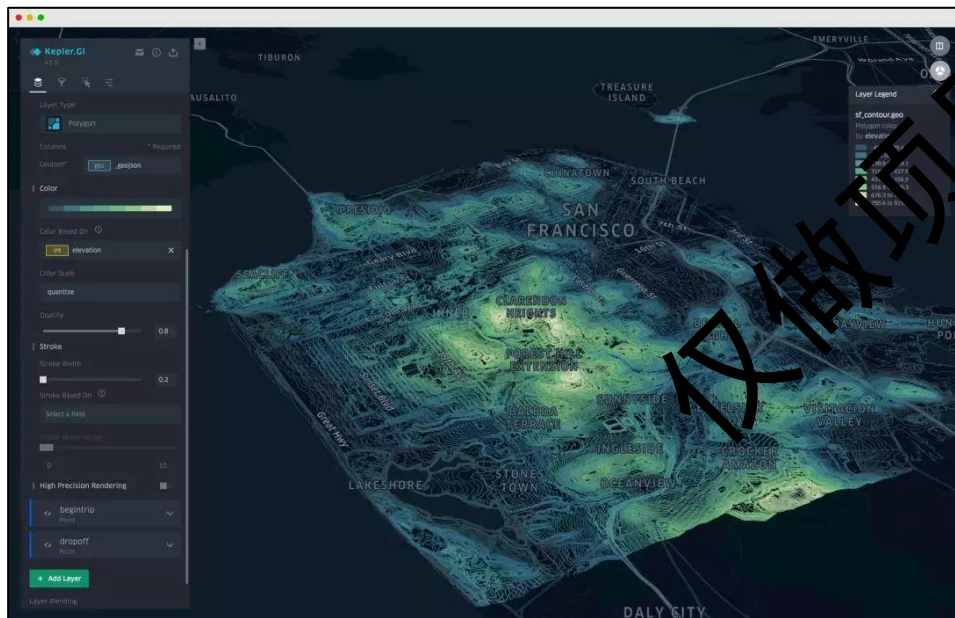
- 轻量级网络协议通信✓；
- 多个服务间实现负载均衡✓；
- 服务的超时重试，限流、熔断、服务发现✓；
- 服务链路跟踪✓；
- 微服务网关✓；
- 引入分布式服务事务组件✓；

仅做项目展示，请勿外传



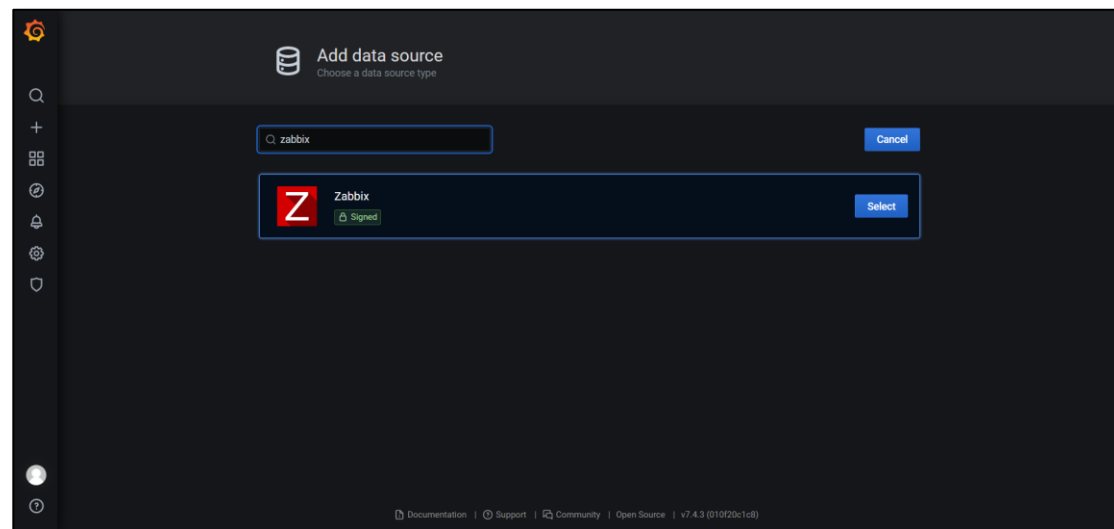
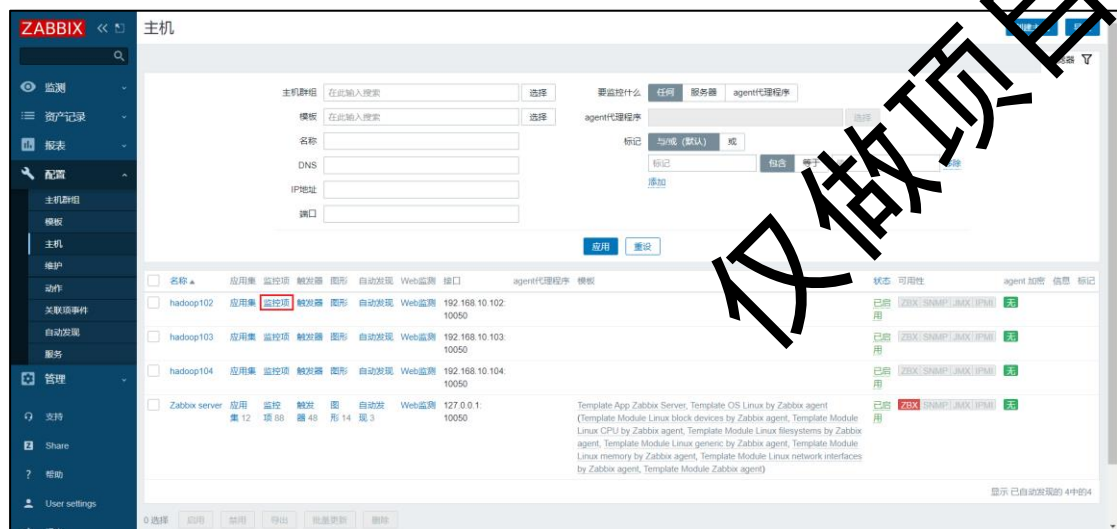
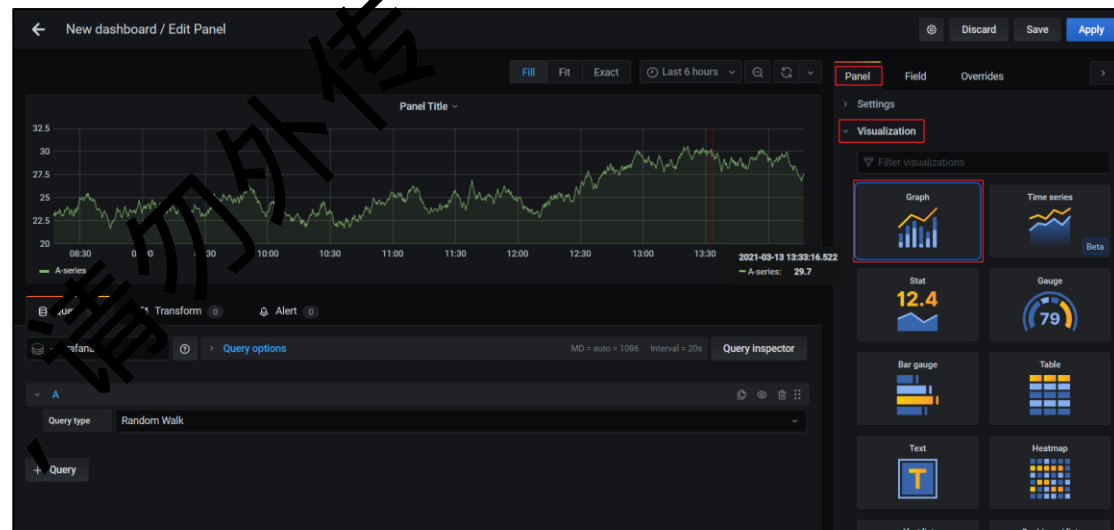
# 工作项：10前端页面复用性升级，看板搭建

- ❑ Python环境安装✓
- ❑ Superset安装;
- ❑ Kepler接口研究;
- ❑ Superset看板搭建;
- ❑ MS powerBI;



# 工作项：11项目运维（专人运维）

- ❑ Zabbix安装;
- ❑ Zabbix触发配置;
- ❑ 部署Grafana;
- ❑ 创建dashboard;
- ❑ 运维shell脚本编写;
- ❑ 服务器运维（监控、安全、权限配置、启停）



# 工作项：12进一步提升数据精细度

- ❑ 从车站点位细化至车站出入口；
- ❑ 建筑体白模细化至高精度三维楼宇；
- ❑ 轨道网分区间双向优化；



仅做项目展示，请勿外传



# 目录

## Contents

---

一	项目任务需求
二	工作进展
三	分工与时间计划

## 时间安排：周例会讨论进度，月专家会评审

- ❑ 项目前期（1-3月）以基础组件安装、云环境安装为主；
- ❑ 项目中期（2-4月）以基础数仓模型构建、后端复用性升级为主；
- ❑ 项目后期（5-6月）以前端、运维、监控设计为主。

