QCon
全球软件开发大会

货拉拉大数据场景下的稳定性保障实践与思考

王海华 货拉拉

QCon+ 案例研习社

扫码学习大厂案例

学习前沿案例，向行业领先迈进

40 个
热门专题
—
行业专家把关内容筹备，
助你快速掌握最新技术发展趋势

200 个
实战案例
—
了解大厂前沿实战案例，
为 200 个真问题找到最优解

40 场
直播答疑
—
40 位技术大咖，每周分享最新
技术认知，互动答疑

365 天
持续学习
—
视频结合配套 PPT
畅学 365 天

自我介绍

王海华

货拉拉大数据基础架构负责人/架构师

- 6年以上大数据架构经验
- 涉及大数据平台产品/系统架构/安全等方向
- 负责过几千到几万台规模大数据集群和架构

Apache Hive/Spark/Alluxio contributor

QCon
全球软件开发大会

目录

QCon
全球软件开发大会

1 背景和挑战

货拉拉介绍

363
国内城市

58万
月活司机

760万
月活用户

8+
业务线

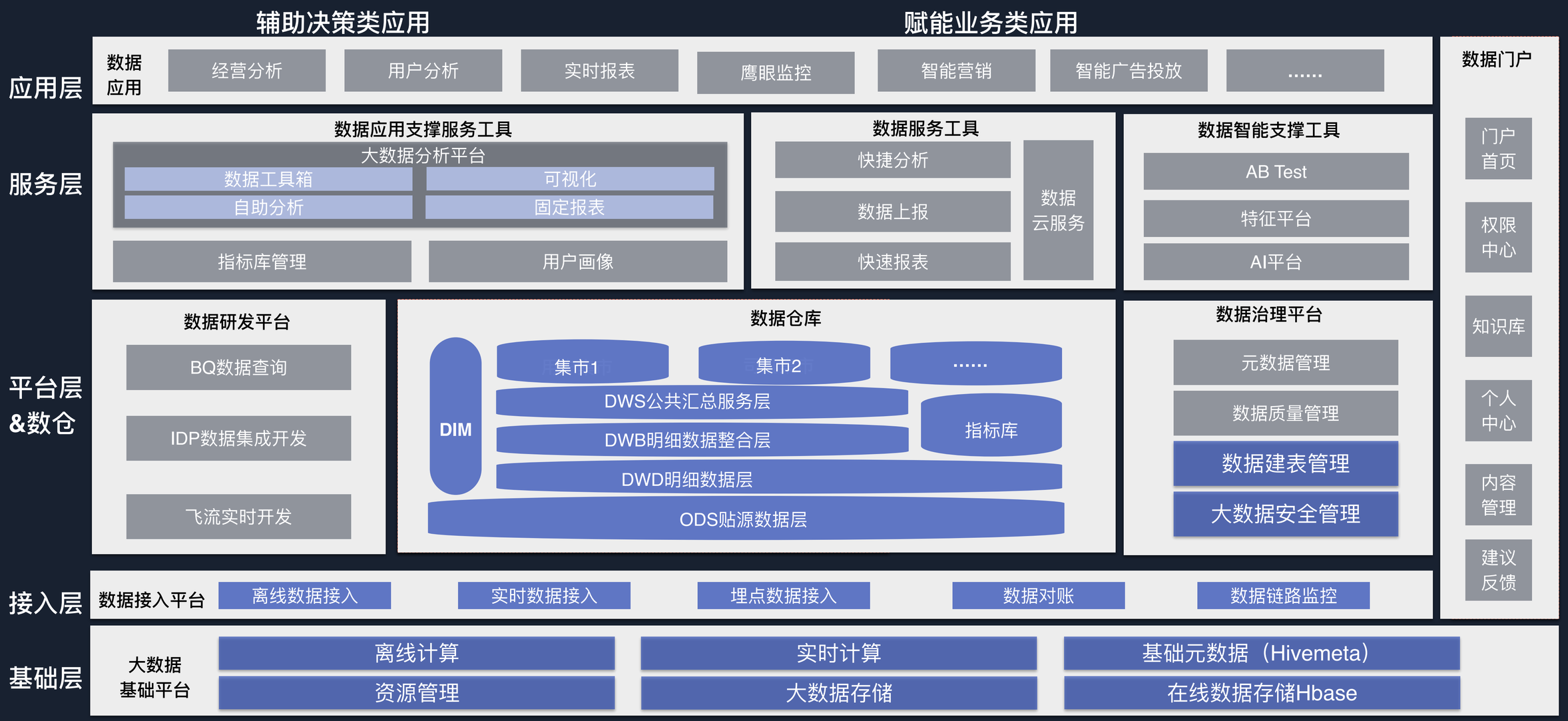
7+
IDC

1000+
机器数

20PB+
存储量

20K+
日均任务数

货拉拉-大数据


辅助决策类应用
赋能业务类应用
应用层
数据应用
经营分析
用户分析
实时报表
鹰眼监控
智能营销
智能广告投放
.....
数据门户
服务层
数据应用支撑服务工具
大数据分析平台
数据工具箱
可视化
自助分析
固定报表
指标库管理
用户画像
数据服务工具
快捷分析
数据上报
快速报表
数据云服务
数据智能支撑工具
AB Test
特征平台
AI平台
门户首页
权限中心
知识库
个人中心
内容管理
建议反馈
平台层&数仓
数据研发平台
BQ数据查询
IDP数据集成开发
飞流实时开发
数据仓库
DIM
集市1
集市2
.....
DWS公共汇总服务层
DWB明细数据整合层
DWD明细数据层
ODS贴源数据层
指标库
数据治理平台
元数据管理
数据质量管理
数据建表管理
大数据安全管理
接入层
数据接入平台
离线数据接入
实时数据接入
埋点数据接入
数据对账
数据链路监控
基础层
大数据基础平台
离线计算
资源管理
实时计算
大数据存储
基础元数据（Hivemeta）
在线数据存储Hbase

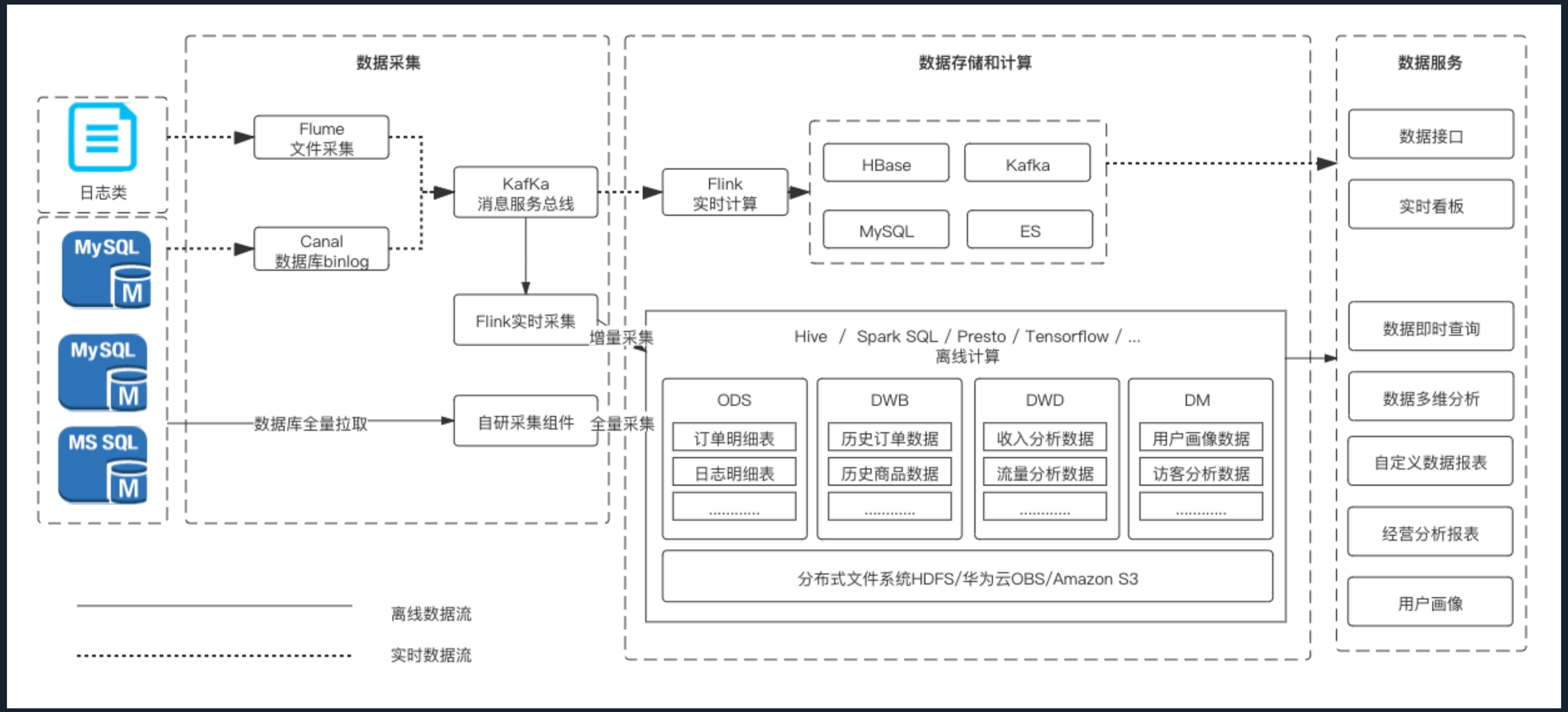
背景和挑战

01 大数据领域下稳定性保障的特殊性

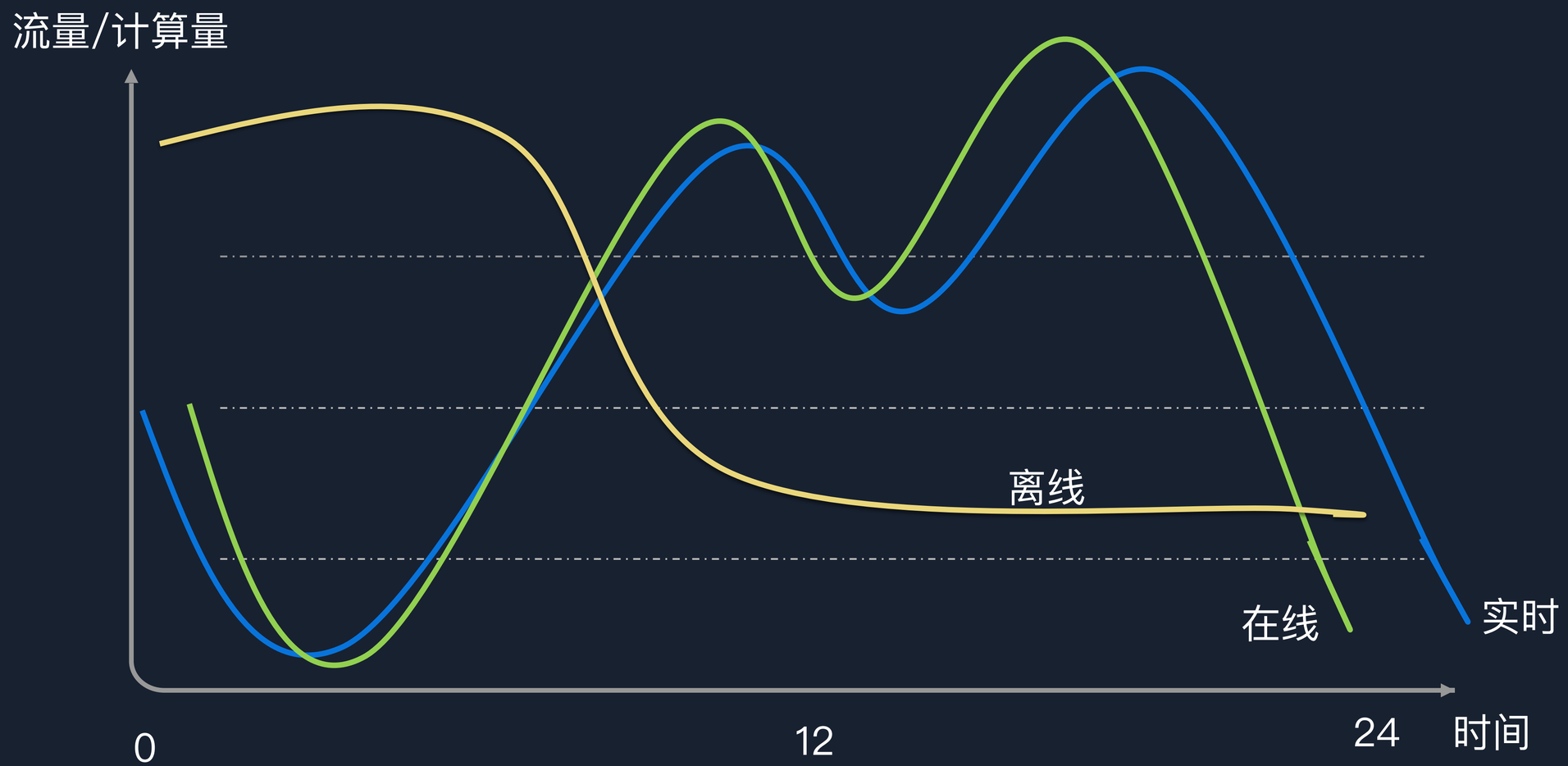
02 大数据领域下场景多样性(在线/实时/离线)

03 开源软件基本能力和生产需求之间的巨大差距

背景和挑战-特殊性


数据采集
数据存储和计算
数据服务
日志类
MySQL
MySQL
MS SQL
Flume
文件采集
Canal
数据库binlog
数据库全量拉取
KafKa
消息服务总线
Flink实时采集
自研采集组件
增量采集
全量采集
Flink
实时计算
HBase
Kafka
MySQL
ES
Hive / Spark SQL / Presto / Tensorflow / ...
离线计算
ODS
订单明细表
日志明细表
.....
DWB
历史订单数据
历史商品数据
.....
DWD
收入分析数据
流量分析数据
.....
DM
用户画像数据
访客分析数据
.....
分布式文件系统HDFS/华为云OBS/Amazon S3
数据接口
实时看板
数据即时查询
数据多维分析
自定义数据报表
经营分析报表
用户画像
离线数据流
实时数据流

背景和挑战-场景多样性



背景和挑战-开源软件生产差距

1. Hadoop生态软件Bug多
2. 只提供数据存储、数据计算等基本能力

项目	(截止2021.05.20) Issue数	Bug数	Bug占比
Hadoop-common	15444	7762	50.25%
HDFS	1264	3173	39.8%
Spark	35058	14482	41.3%

保障需求和保障目标

场景	价值	稳定性矛盾	保障目标
数据采集和存储	数据存储可靠性是大数据的生命线	数据丢失	数据可靠性100%
离线核心数据链路和报表	高管日常决策首要依据	数据延迟	核心链路数据延迟 >= 1次/每月
数据准确性	业务支撑和数据赋能的基础	离线、实时报表数据错误	核心数据准确性100%
大数据核心服务	抢单、风控、实时营销等核心链路数据服务	稳定性无保障，冒烟事故多	可用性 >= 99.95% 单次不可用时间 >= 10min
大数据核心产品	数据研发、数据应用等大数据能力输出	稳定性，出现过大面积长时间不可用故障	可用性 >= 99.9% 单次不可用时间 <= 30min

QCon
全球软件开发大会

2 能力保障

能力保障

分场景保障
To distinguish the scene

链路高可用
link high availability

故障隔离
Fault isolation

容量规划
Capacity planning

区分场景

在线场景

- 延迟敏感：毫秒级
- 可用性要求高

离线场景

- 延迟不敏感：分钟到小时级别
- 吞吐高，资源利用率高
- 可用性要求中等

实时场景

- 延时敏感：亚秒到秒级
- 可用性要求较高

分场景保障

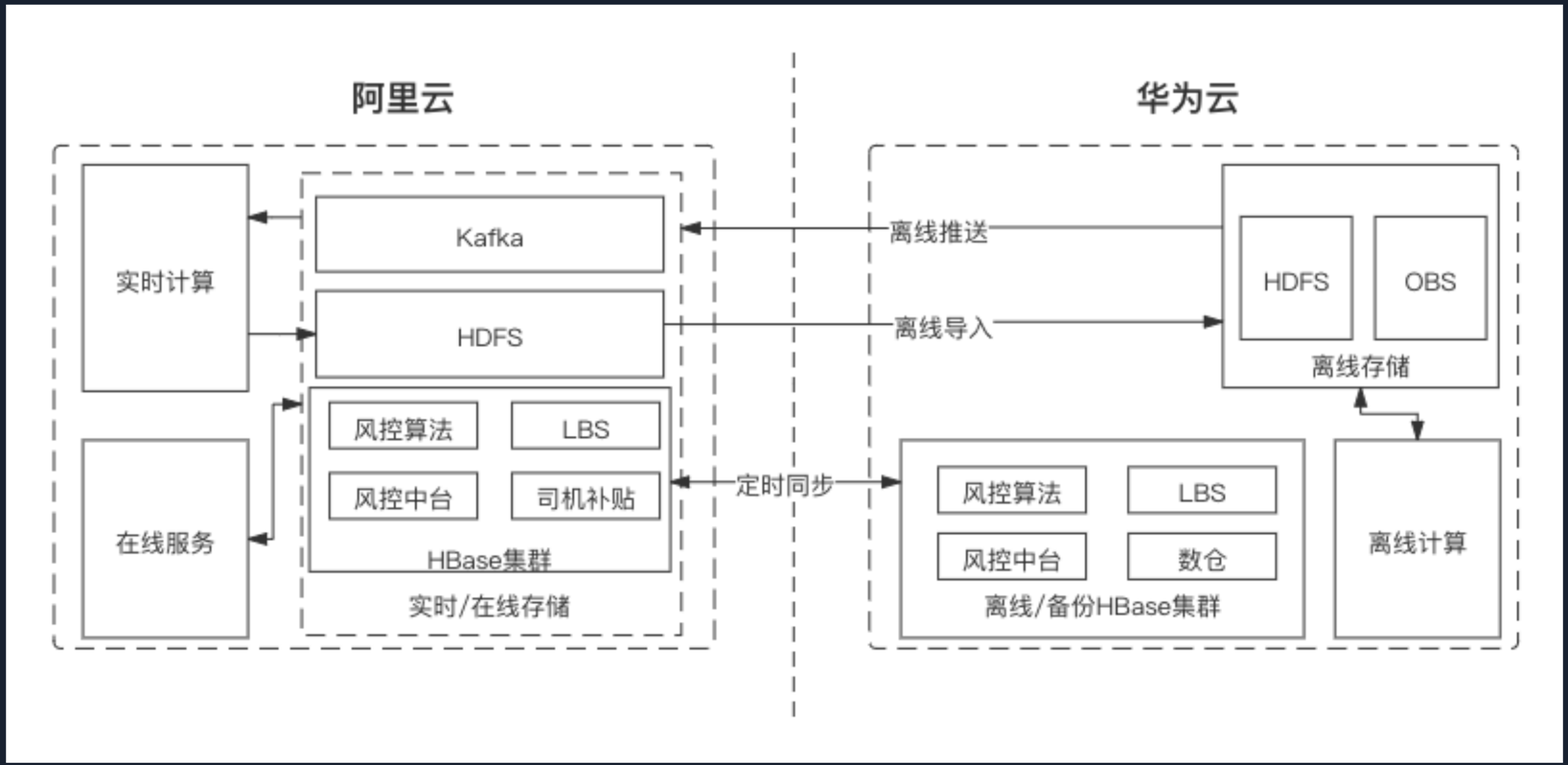
场景	保障需求	保障策略
在线	高： 99.95%可靠性意味着一个月停机时间<= 21.6分钟 单次停机时间要求10分钟内	事前保障，故障预案，熔断降级
离线	一般： 99%可靠性意味着一个月停机时间<= 432分钟 单次停机要求60分钟内	优先事后保障，着重是发现、响应和恢复能力
实时	高： 可用性要求比在线略低	与在线相近，降级能力，恢复能力

链路高可用

1. 定义关键路径和关键依赖
2. 关键路径系统高可用
3. 尽量弱依赖

故障隔离

1. IDC故障隔离
2. 业务故障隔离
3. 分场景故障隔离



容量规划

货拉拉

容量梳理

1. 确定容量指标
2. 压测确认容量最大水位
3. 包含自身容量和外部资源依赖容量

容量指标

1. 业务层/服务层/VM/OS
2. 外部依赖容量
3. 指标分层

容量监控和预警

1. 告警包含ERROR、WARN两个级别
2. 根据历史趋势设定预警值，能做到容量预警

容量监控

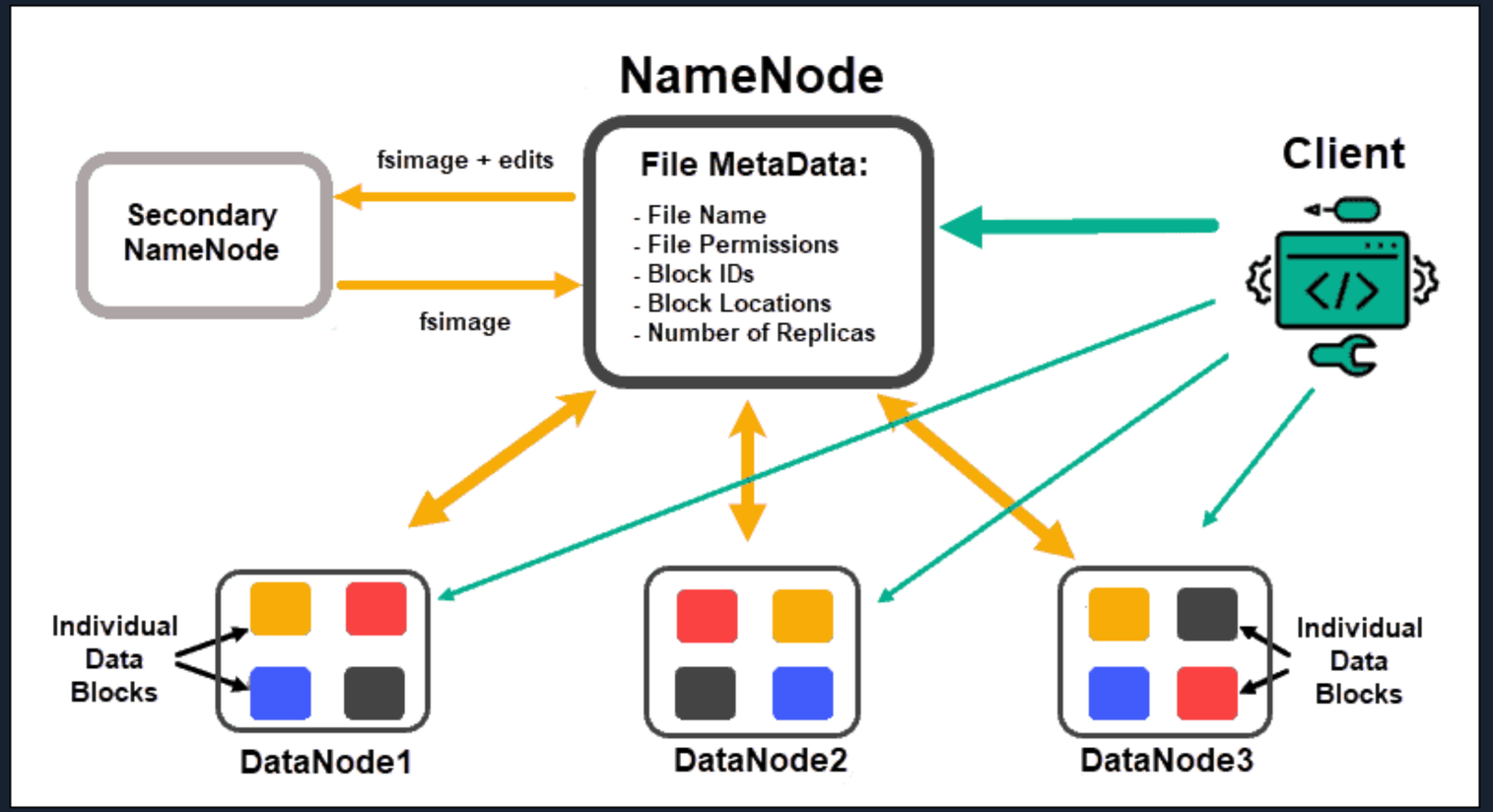
1. 从以系统指标梳理为起点，以容量预警作为目标
2. 指标分级，一级指标覆盖ERROR级别告警

分级	作用	告警级别	常见系统指标
一级指标	发现问题和定位问题	**ERROR**、WARN	系统服务能力和存储能力 时延/吞吐/错误率 CPU/Memory/Network/IO
二级指标	定位问题	WARN	系统内部状态，例如外部资源读写效率，GCTime，线程池繁忙数量等

容量规划-NameNode

主要功能：

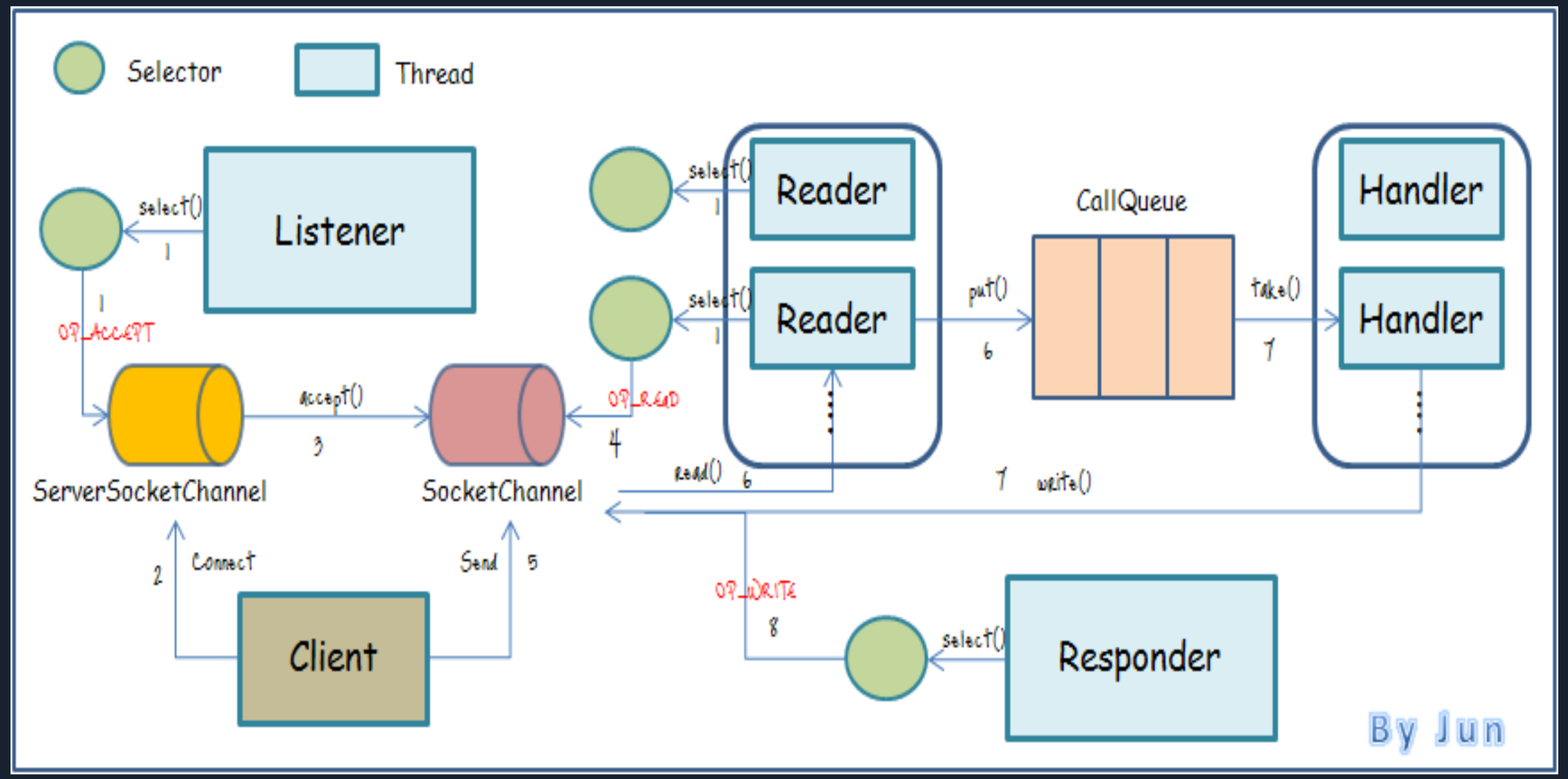
1. 存储元数据管理
2. 元数据读写服务
3. Datanode节点管理



容量规划-NameNode

内部实现：

1. Listener
2. Reader
3. CallQueue
4. Handler
5. Responder



容量规划-NameNode

指标分级	指标名	压测容量值	告警阈值
一级指标	RpcProcessingTime RpcProcessingNumops RpcQueueTime CPU user avg/load/IOUtil/Ioawait/Network	RpcProcessingNumops 100K Metadata storage 500M	CPU user avg 20min > 30% RpcProcessingNumops avg > 60K/s …
二级指标	ReaderRunnableCount RpcQueueLength Get/Delete/RenameRpcNumops	N/A	ReaderRunnableCount avg 30min > 30% …

QCon
全球软件开发大会

3 流程规范保障

流程规范保障

研发和发布规范

01

故障管理规范

02

系统和数据研发规范


系统研发
系统设计规范
代码规范检测
Code review
Sonar检测
单元/集成测试
灰度上线
变更窗口
发布审核
设 计
研 发
测 试
上 线
数据研发
数据模型
设计规范
SQLScan
静态检测
SQL Review
数据质量
检测
试运行验证
准确性、资源

系统发布规范

01

发布窗口

1. 业务低峰期，非节假日前一天
2. 离线12-18点，在线/实时20 – 24点

02

发布内容和用户通知

1、非标准附加详细命令

2、通知对应业务方和值班人员

03

验收

1、稳定性验收

2、功能和性能验收

3、可回滚、发布后oncall

04

审核

1、日常审核，变更数量限制

2、节假日封版，紧急变更流程

3、审核规范执行情况

故障定级规范

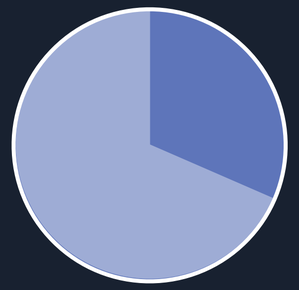
No measurement, no improvement

作用：稳定性度量

故障定级规范

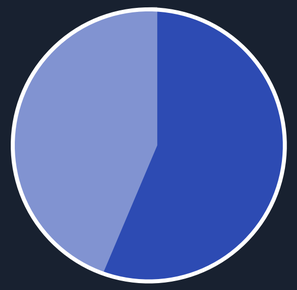
稳定性保障目标	核心指标	故障等级严重程度
数据可靠性	表数据丢失率	表数据丢失比例 是否可找回
数据准确性	数据重要等级 是否业务先发现	是否业务先发现 数据重要程度
离线核心数据链路和报表	报表延迟时间	报表产出与预期差距
核心服务可用性	资损/服务停机时间	资损大小 服务停机时间长短 停机时间是否高峰期 是否核心链路
核心产品可用性	服务停机时间	功能损失比例 服务停机时间长短 停机时间是否高峰期

系统复盘规范



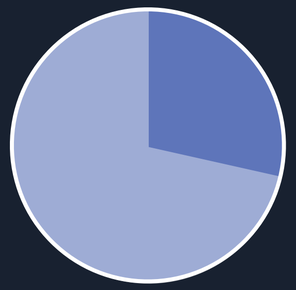
时间线与事实对齐

1. 保证大家对于事实有相同的感受
2. 区分为发生、发现、定位、恢复阶段



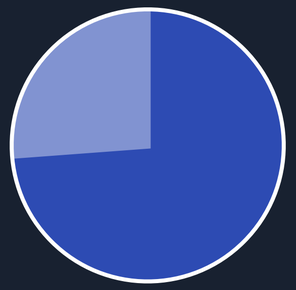
存在问题讨论

1. 发现和引入
2. 定位
3. 恢复
4. 根本原因



整改事项确定

1. 与问题一一对应
2. 事项明确、有负责人和完成时间



整改跟进

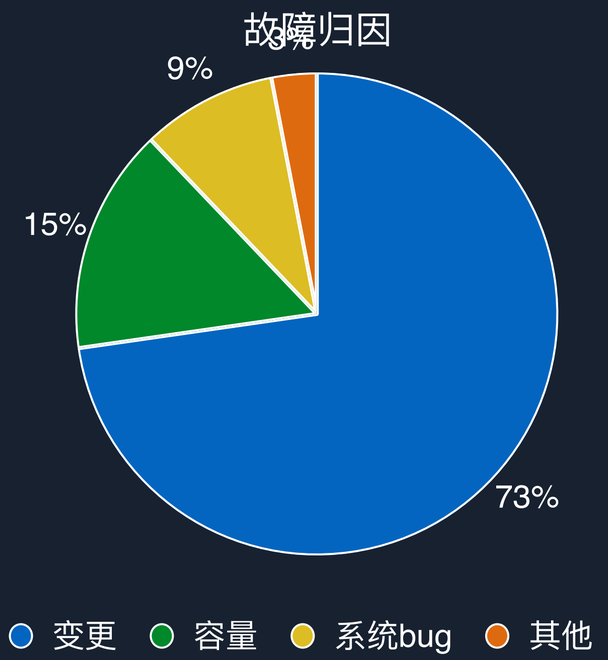
1. 由责任人/故障跟进小组跟进
2. 复盘结果用户信息同步

QCon
全球软件开发大会

4 组织保障

SRE

- **72%**故障来源于发布和变更：没有变更就没有“伤害”
- 稳定性 **or** 产品研发需求**PK**谁胜利?
- 研发为新功能负责，**SRE**为稳定性负责
- 专职 SRE team?



故障跟进小组

目标：故障复盘率100%，整改事项完成率100%

全局回顾故障，系统性解决方案

输出故障定期报告、优化流程制度

技术委员会

- 打破组织壁垒，提升技术和稳定性底线
- 负责技术评审、研发规范、架构统筹规划
- 注重为结果负责，避免形式化

QCon
全球软件开发大会

5 总结与思考

总结

能力保障

分场景保障

链路高可用

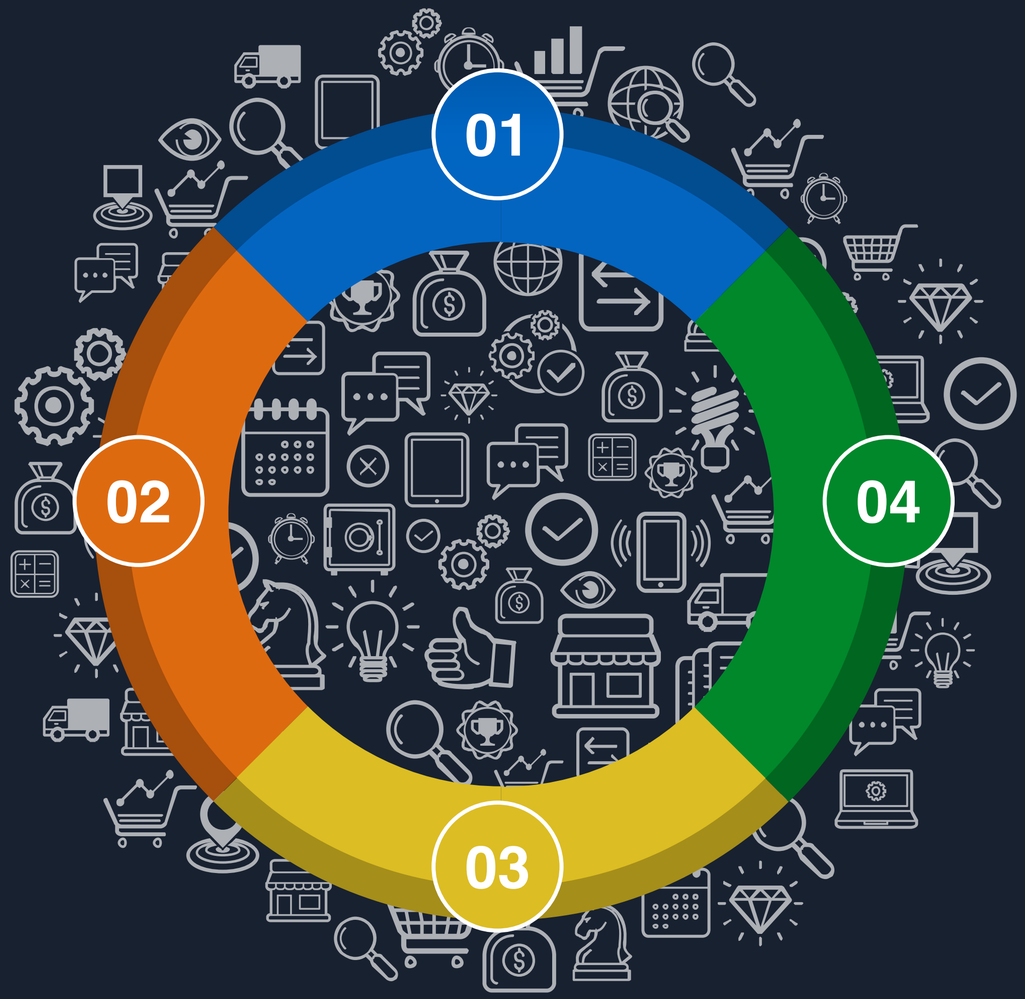
故障隔离

容量规划

流程规范

研发和发布规范

故障管理规范



保障目标

数据准确性/可靠性

核心数据链路稳定性

核心服务/产品稳定性

组织保障

SRE团队

故障跟进小组

技术委员会

一些思考

- 稳定性的建设是风险控制能力建设，而非靠运气
- 稳定性的提升依靠事实和数据，而非靠感觉
- 稳定性的目标实现靠端到端体系化建设，而非靠单点突破
- “道阻且长，行则将至，行而不辍，未来可期”


THANKS

InfoQ | 写作平台

InfoQ 写作平台是 InfoQ 开放给**开发者的高端技术社区**，创作者可以在这里**自由创作和发布内容**。

写作平台将为创作者提供**签约、培训、资金扶持**等一系列权益，助力作者成长为高精尖技术人才；同时也为企业提供**品牌、活动打造、内容传播**等服务，与伙伴一同成长。

扫码申请创作者

企业/个人均可申请

扫码进入写作平台

打开技术大门