流计算技术与GreenPlum结合进行数据处理

— 杨旭钧　Bloom.io CTO

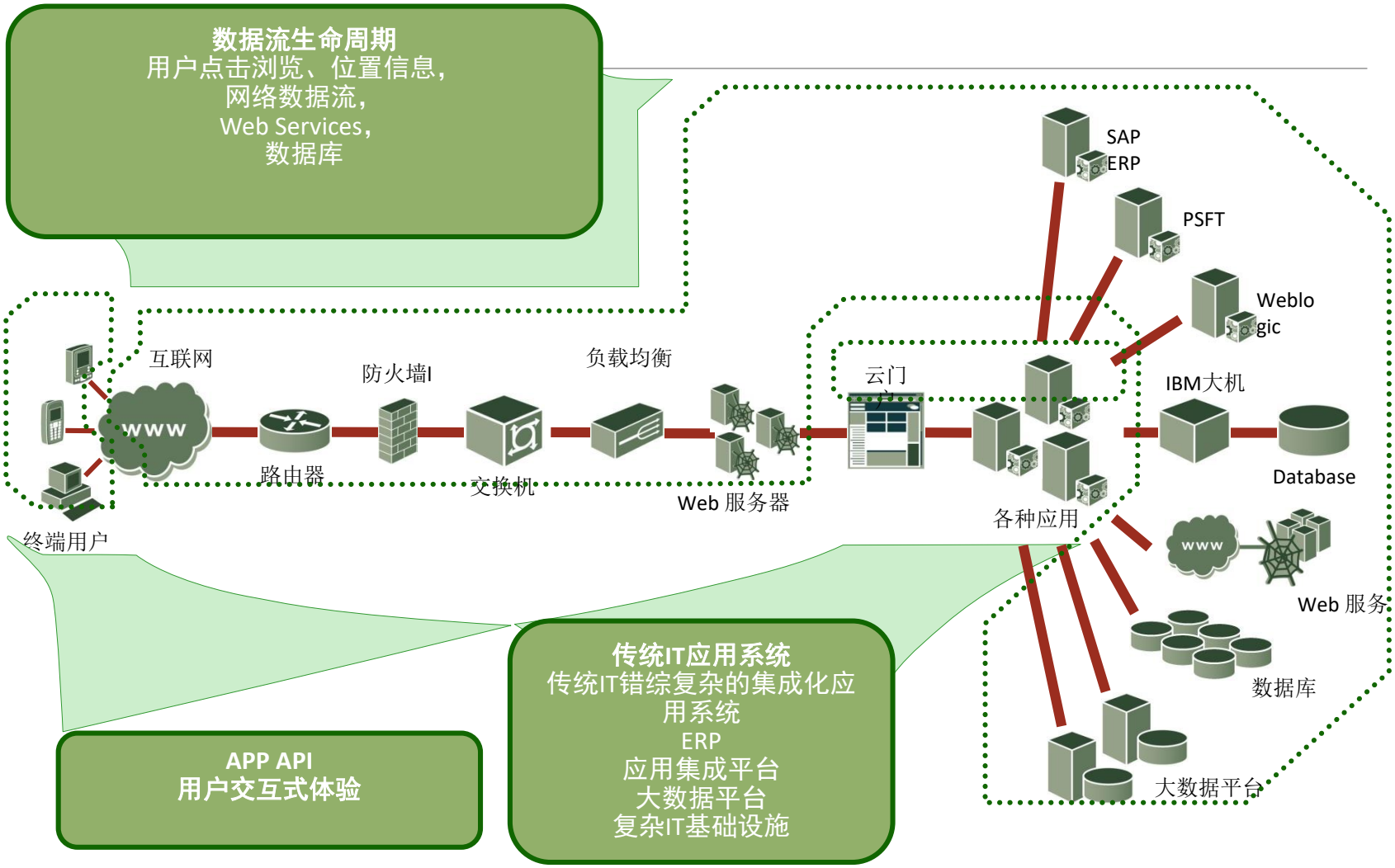
Bloom Software Ltd 公司介绍

Bloom Software 是一家专注于山寨 Google 技术的公司。

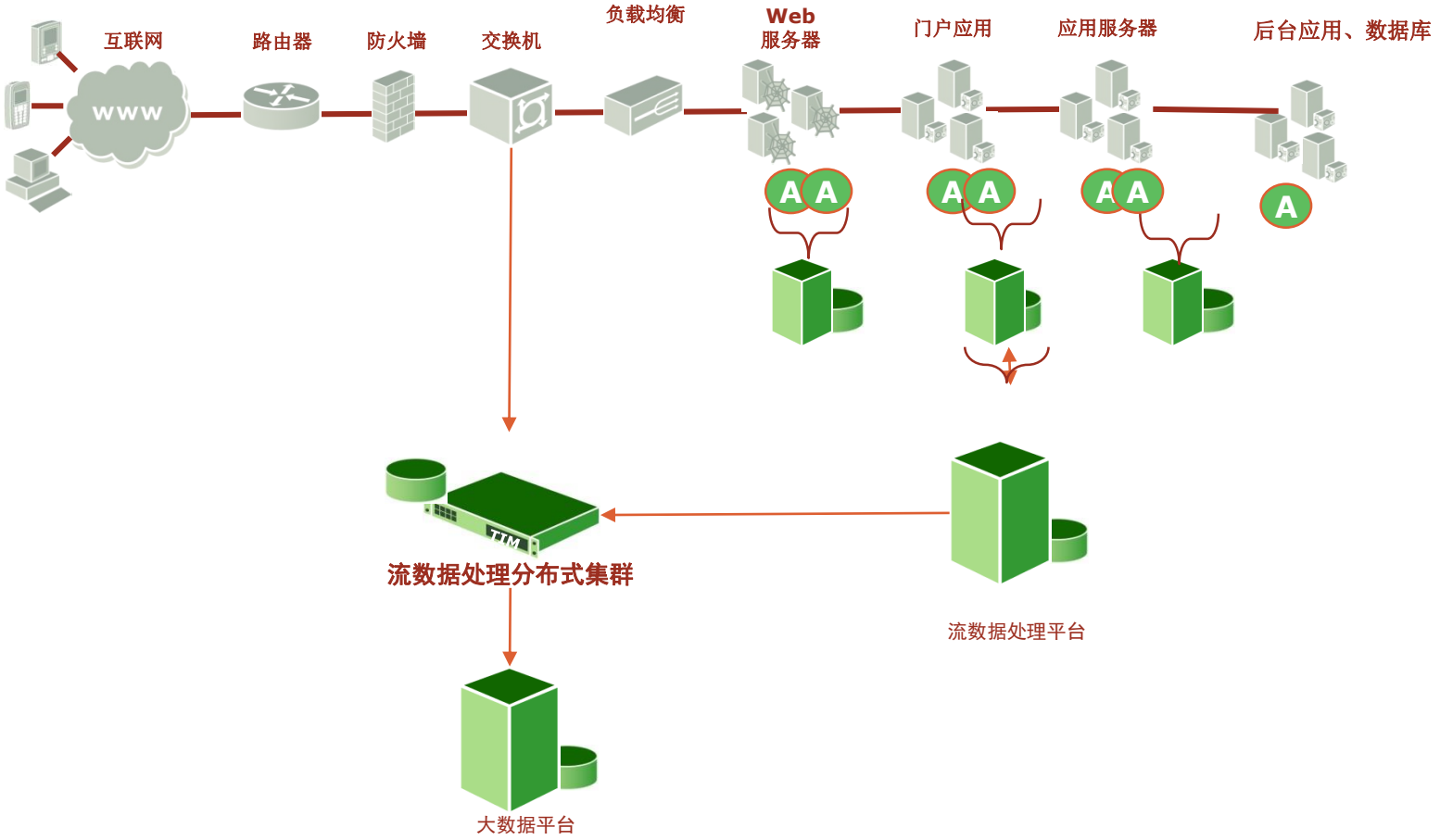
Bloom Stream Process Plateform就是Google Data Flow的山寨版实现，

还会陆续推出 Google BigQuery，Google Analytics 等山寨版。

传统IT环境下数据流生命周期



大数据环境下数据流生命周期



Bloom BSP实时流数据处理平台

统一实时消息流数据集成平台有如下的功能特点：

— 内存级数据流处理

— 混合云迁移

— 摄入和回放数据流

— 数据库云容灾备份

— IoT物联网消息流传输

Bloom BSP 产品特点和优势

优势一：

BSP实现了自己的分析引擎，可对流入数据进行毫秒级实时分析。

优势二：

BSP集成了你可以想到的几乎所有MQ平台，可以灵活适配各种MQ，如JMS,ActiveMQ,ZeroMQ,Kafka等。

优势三：

BSP兼容多种消息格式或数据格式，如Msg,XML,JSON,TXT,Log日志,图片,PCAP,TCP/UDP包等

优势四：

BSP拥有内存级流数据处理平台，并在其上面加入了SQL语言，可进行多种复杂流式SQL查询。

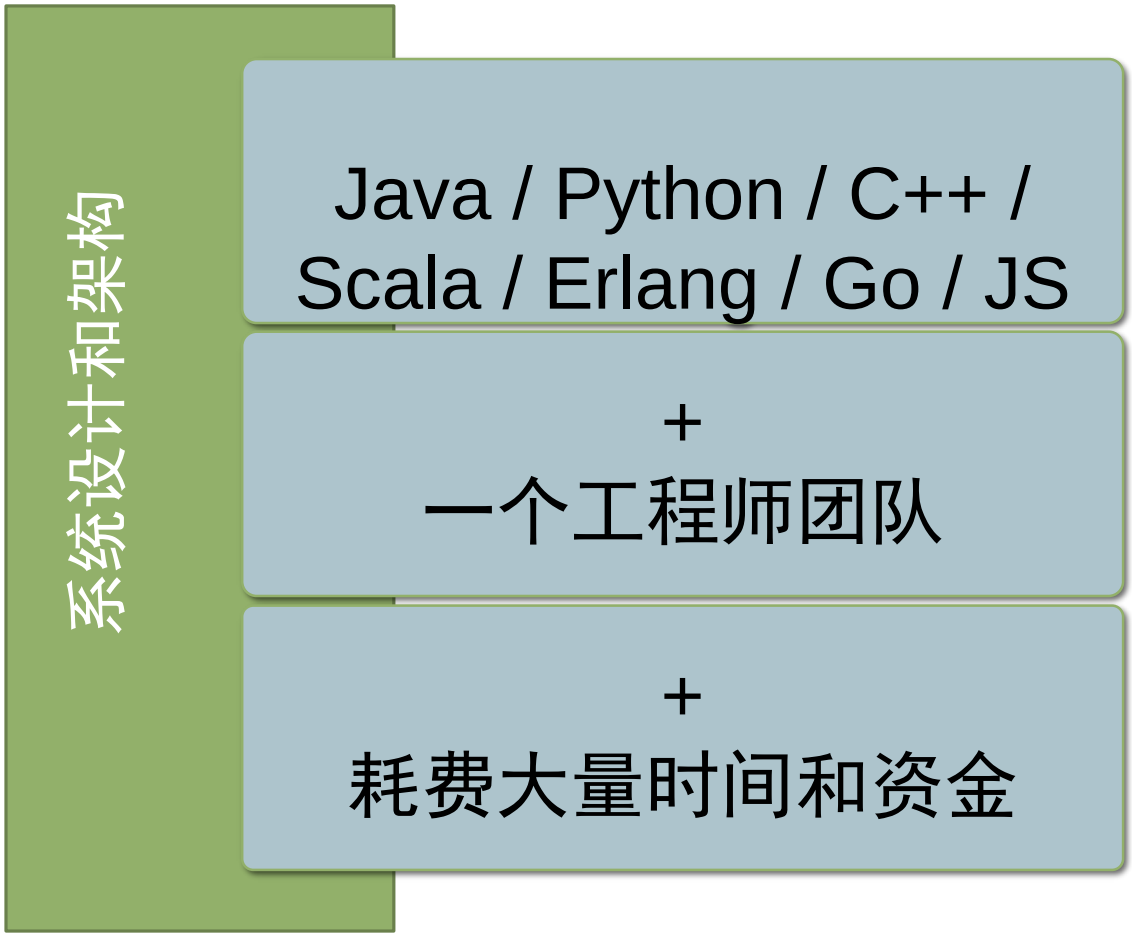
你需要构建一个实时分析引擎应用

持续流处理和实时查询

选择1: 开源(需要掌握各种开源组件)

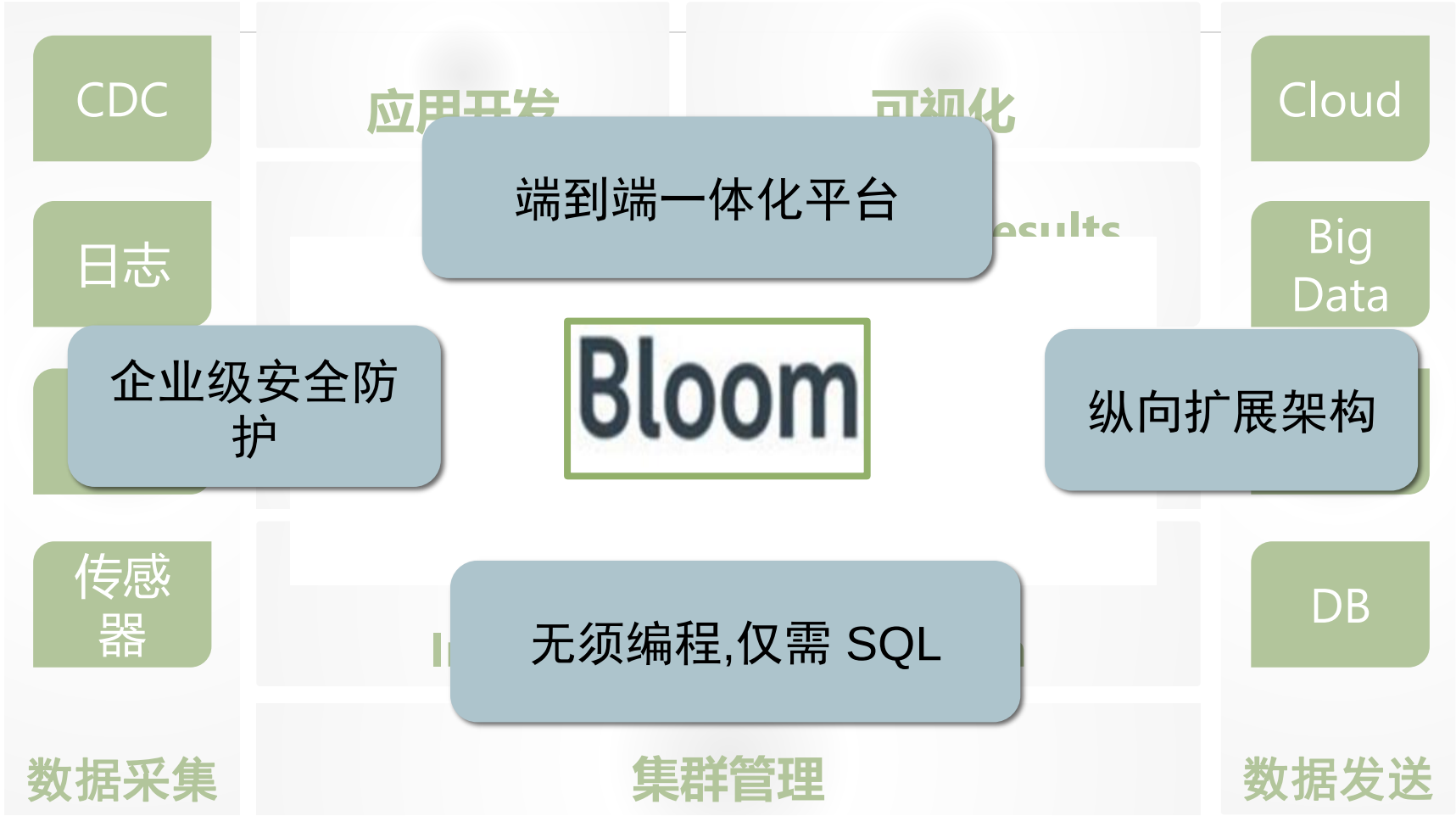
持续流数据处理和实时查询

选择2: 开源 (耗费资源较大…)



选择 2: Bloom Stream Platform

持续流处理和实时查询



选择 2: BSP (No brainer..)

+

声明式 UI, SQL

+

充分利用现有资源

+

分钟级构建部署一整套大数据环境

BSP 核心技术

High Performance MQ

+

In Memory Computing

+

Stream SQL Query

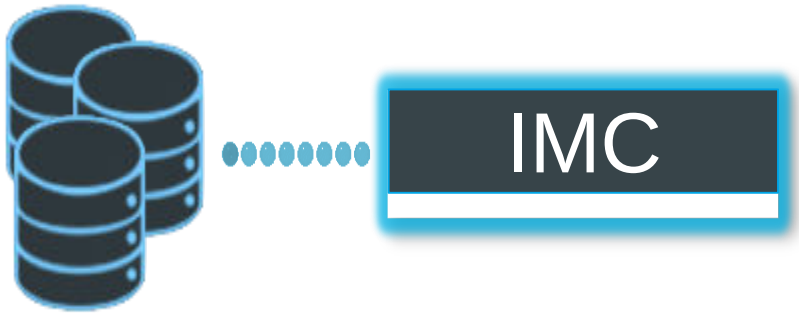
满足企业所有大数据分析需求

- 金融
- 电商
- 供应链
- HR
- 营销
- CRM
- 制造业
- 支持

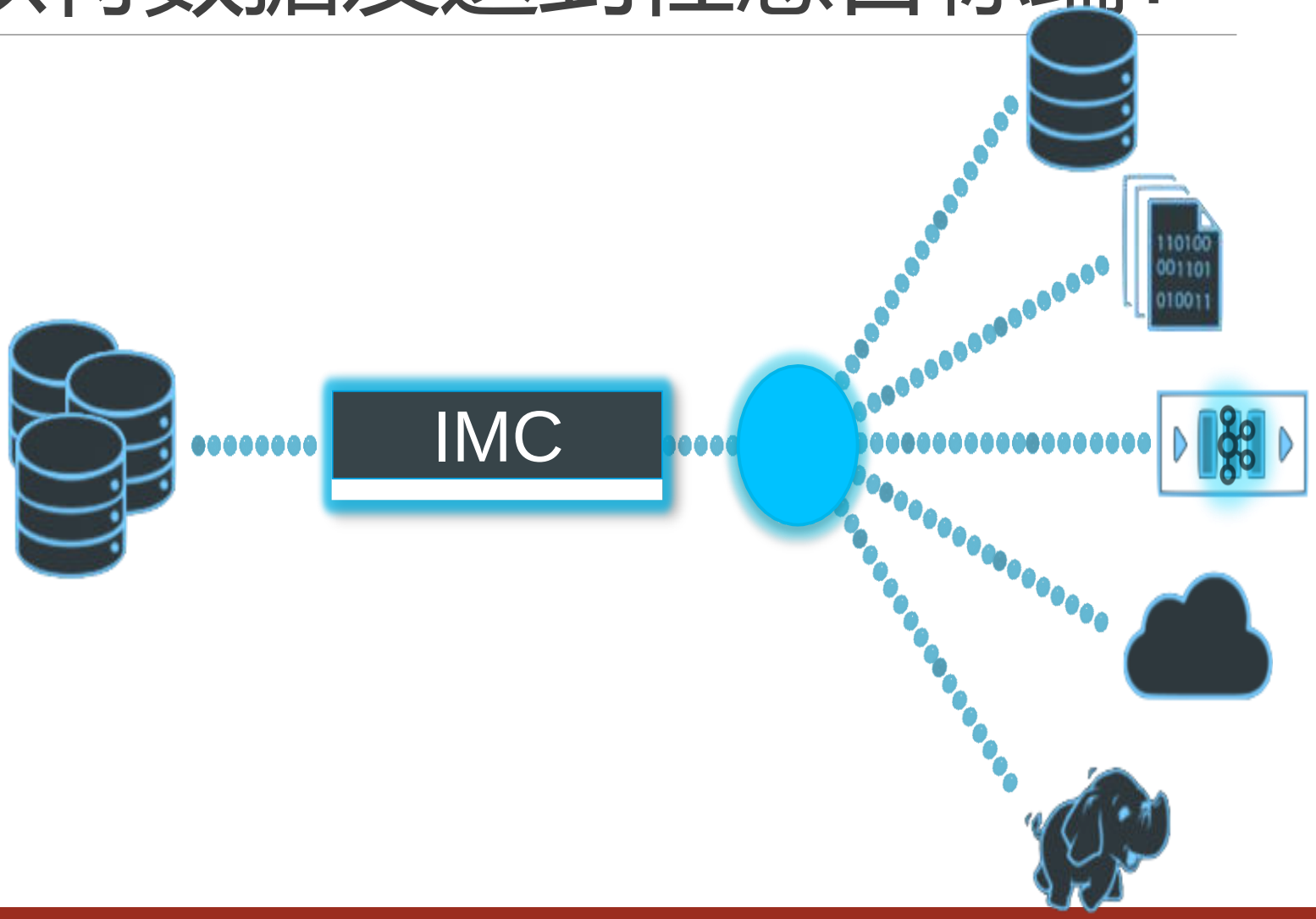
以及所有主流数据库

- Oracle
- MySQL
- SQL Server
- DB2
- Postgres
- HP NonStop
- Mainframe

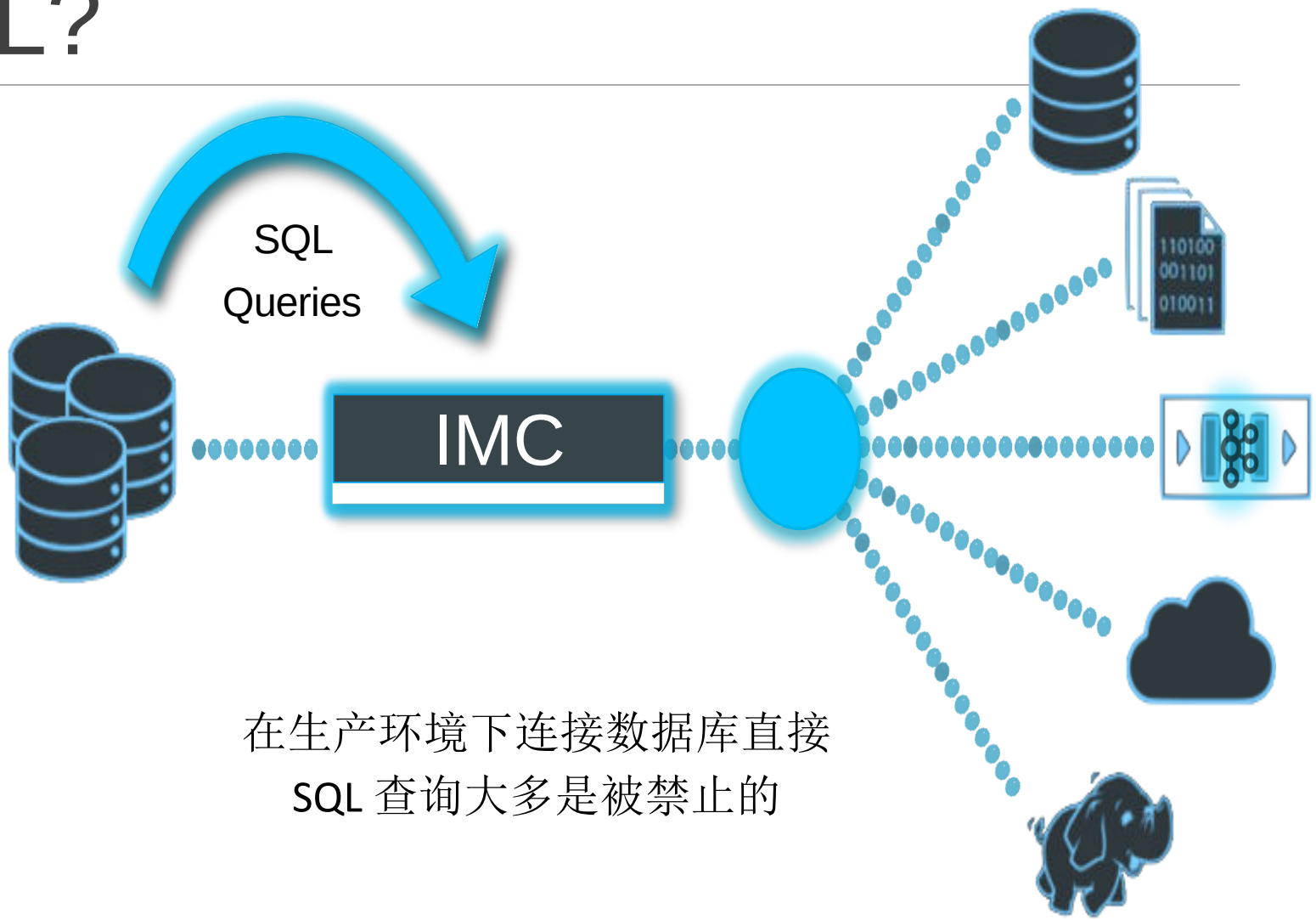
怎样将数据送入IMC?



可以将数据发送到任意目标端?

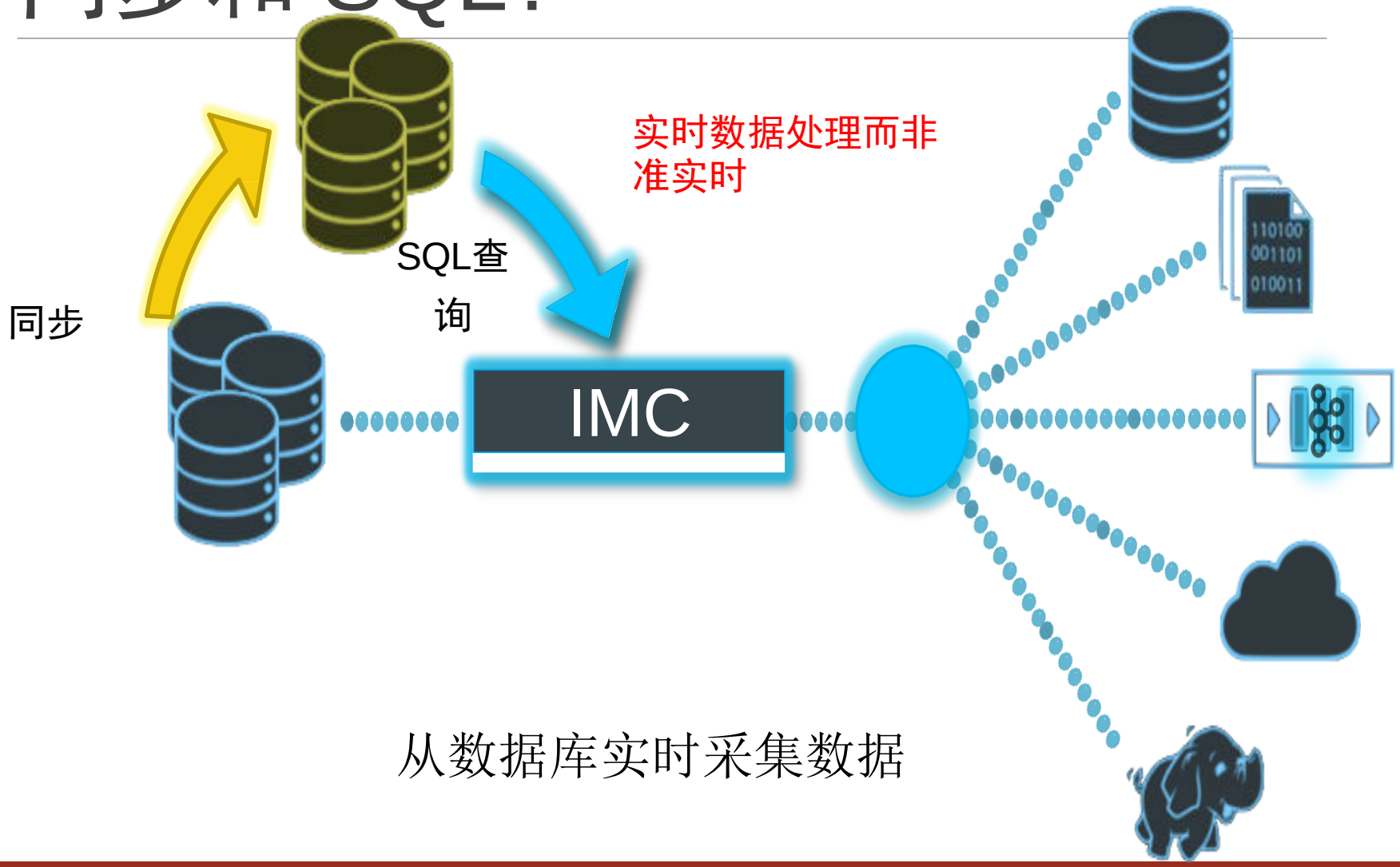


SQL?



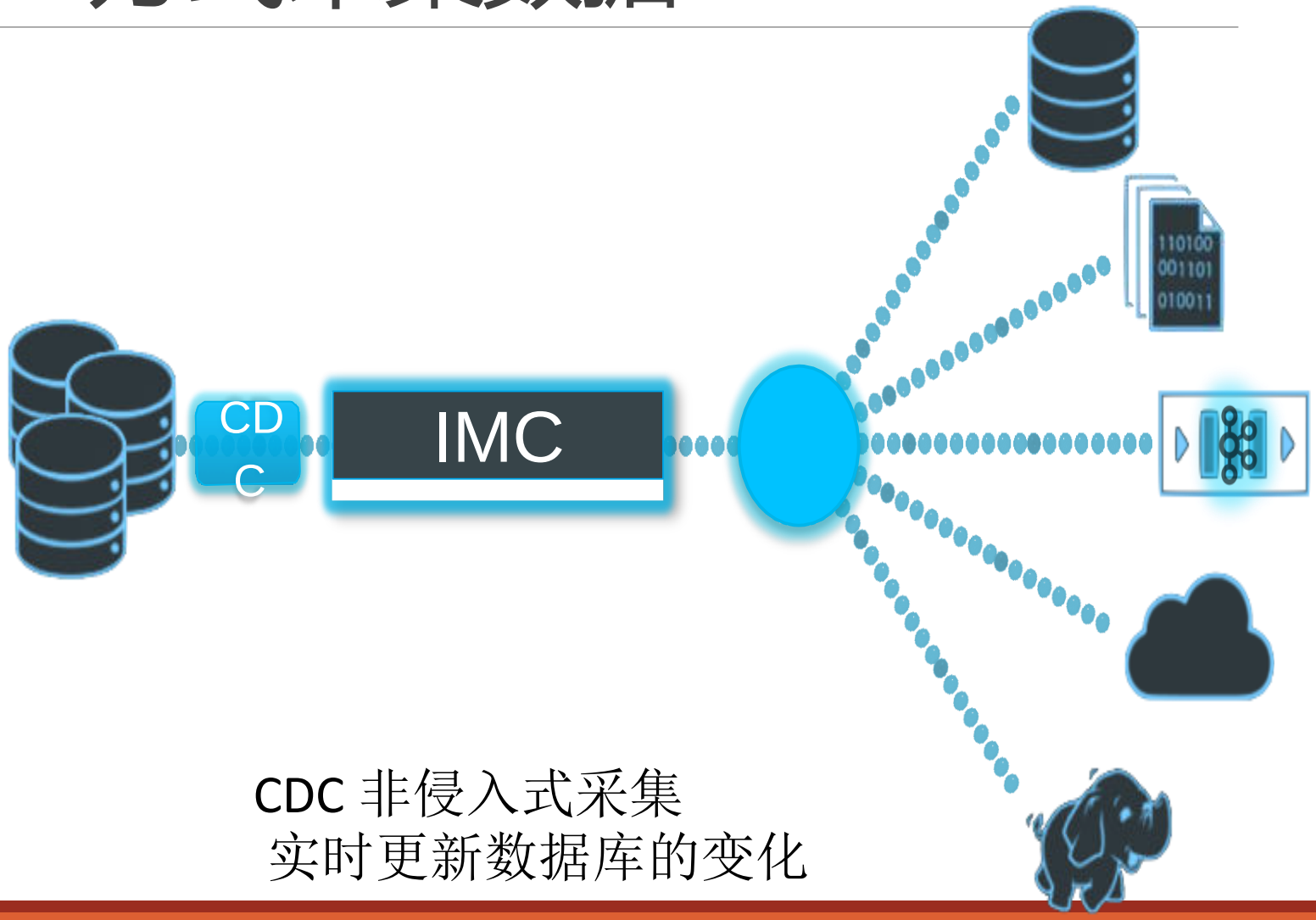
在生产环境下连接数据库直接
SQL 查询大多是被禁止的

同步和 SQL?



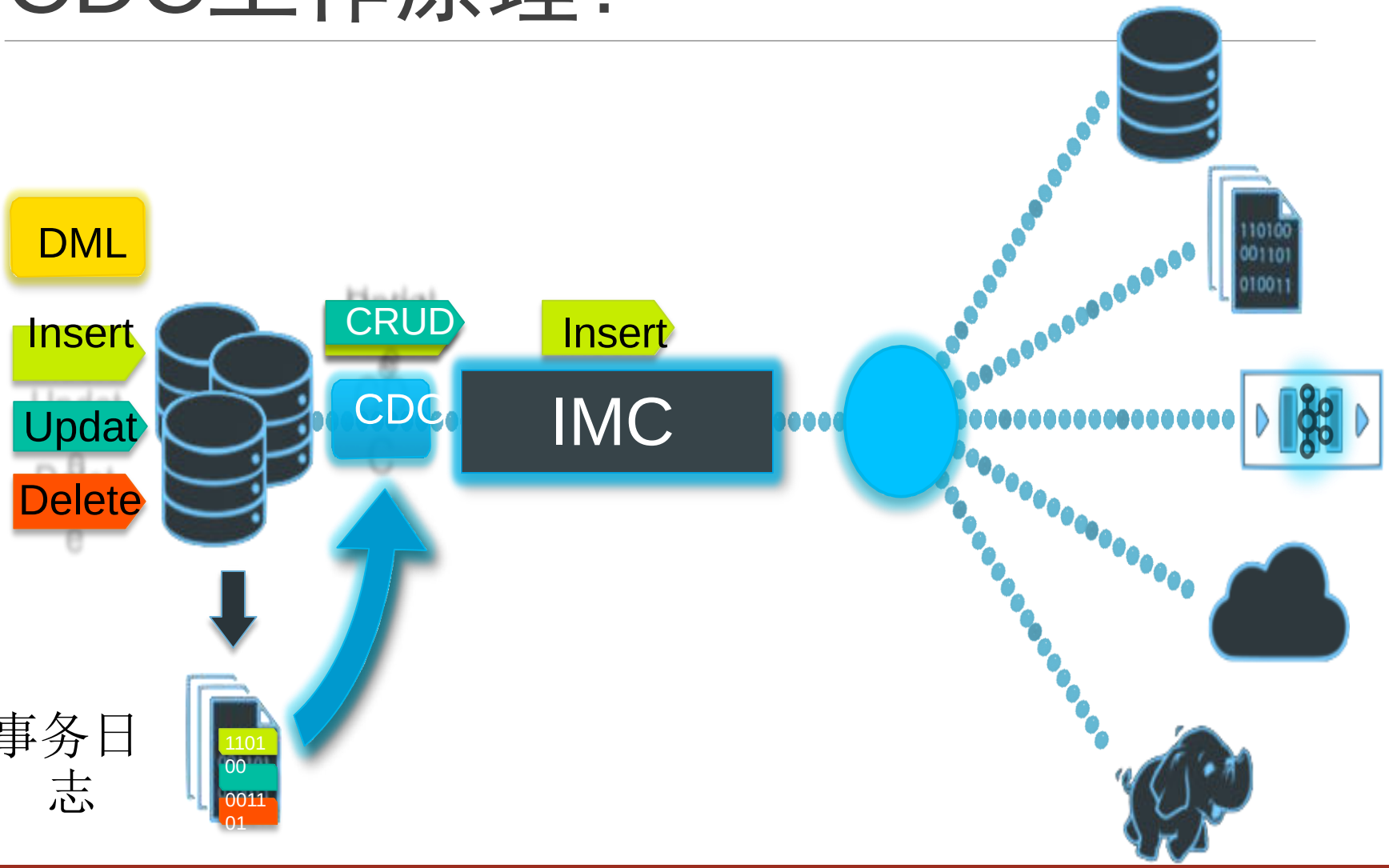
从数据库实时采集数据

CDC方式采集数据

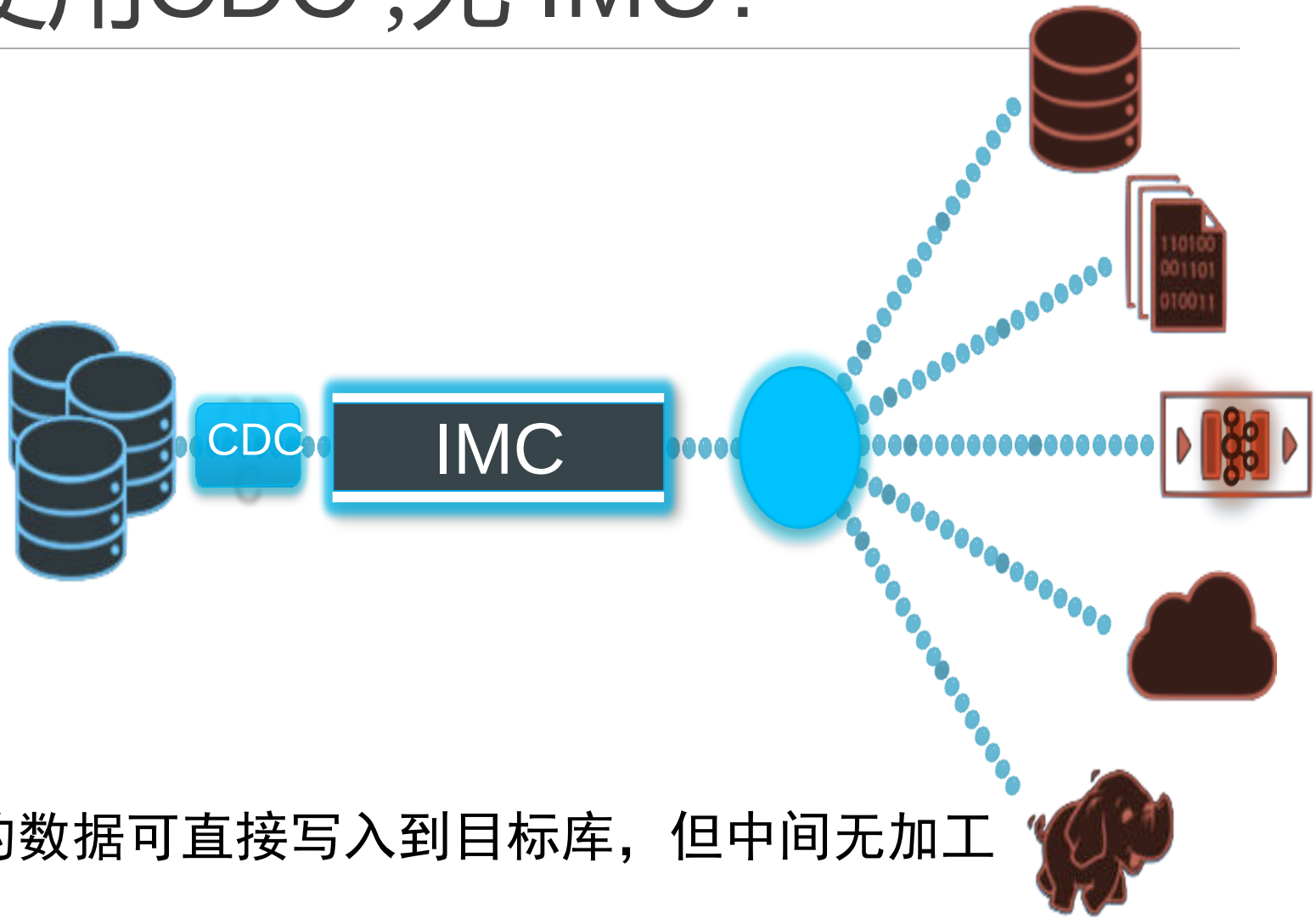


CDC 非侵入式采集
实时更新数据库的变化

CDC工作原理?

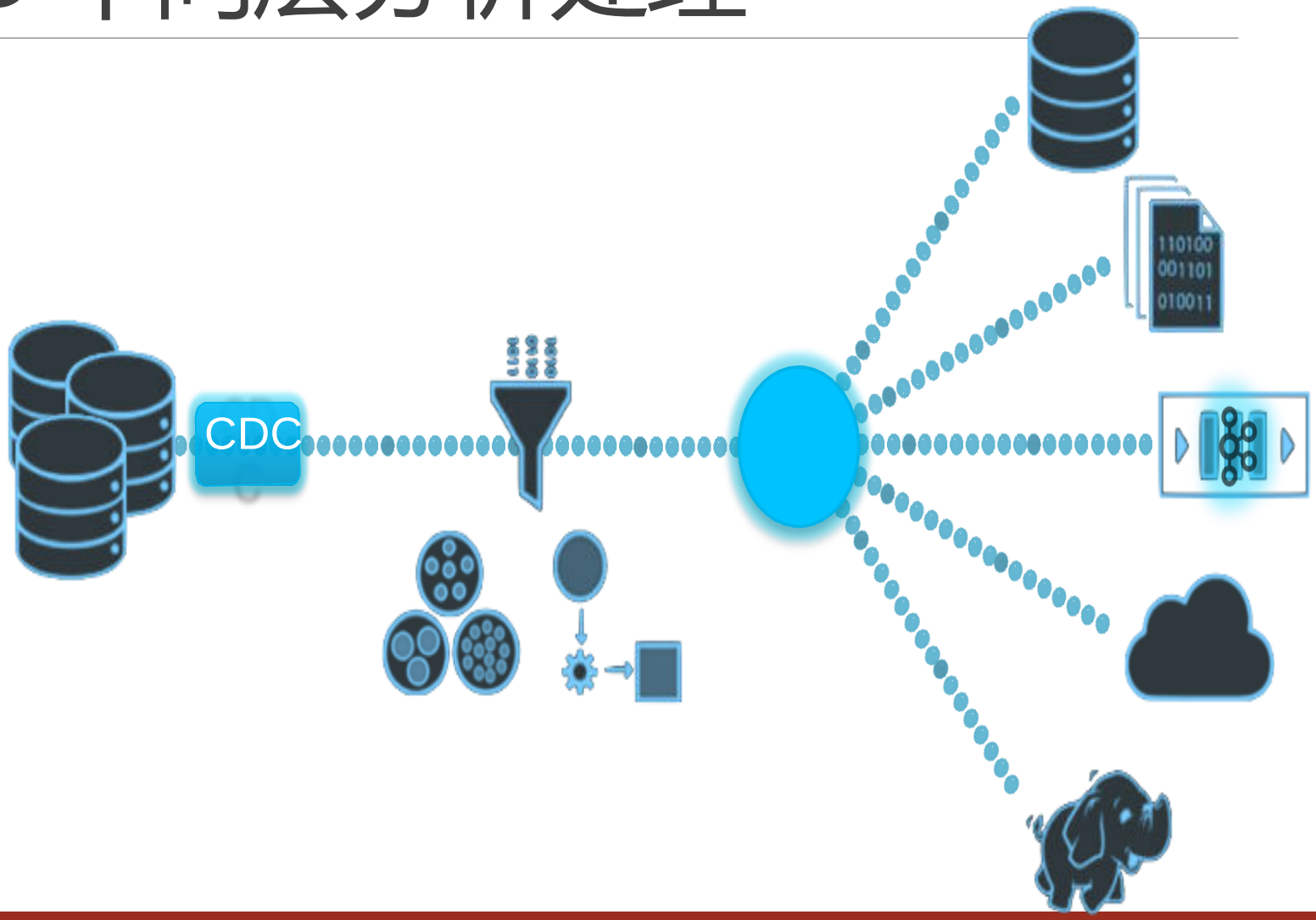

DML
Insert
Updat
Delete
CRUD
CDC
Insert
IMC
事务日志
1101
00
0011
01
110100
001101
010011

仅使用CDC ,无 IMC?

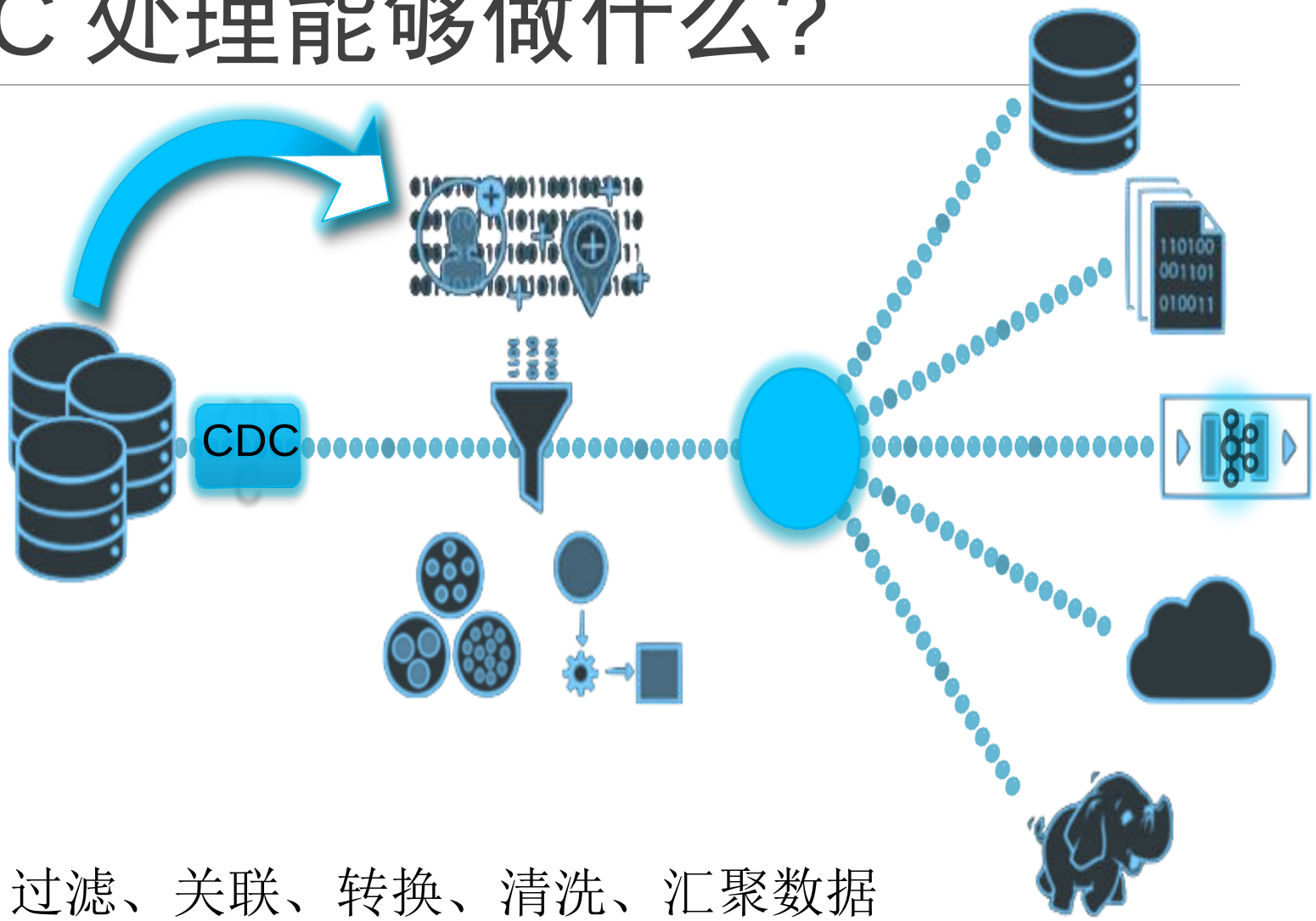


CDC的数据可直接写入到目标库，但中间无加工

IMC 中间层分析处理

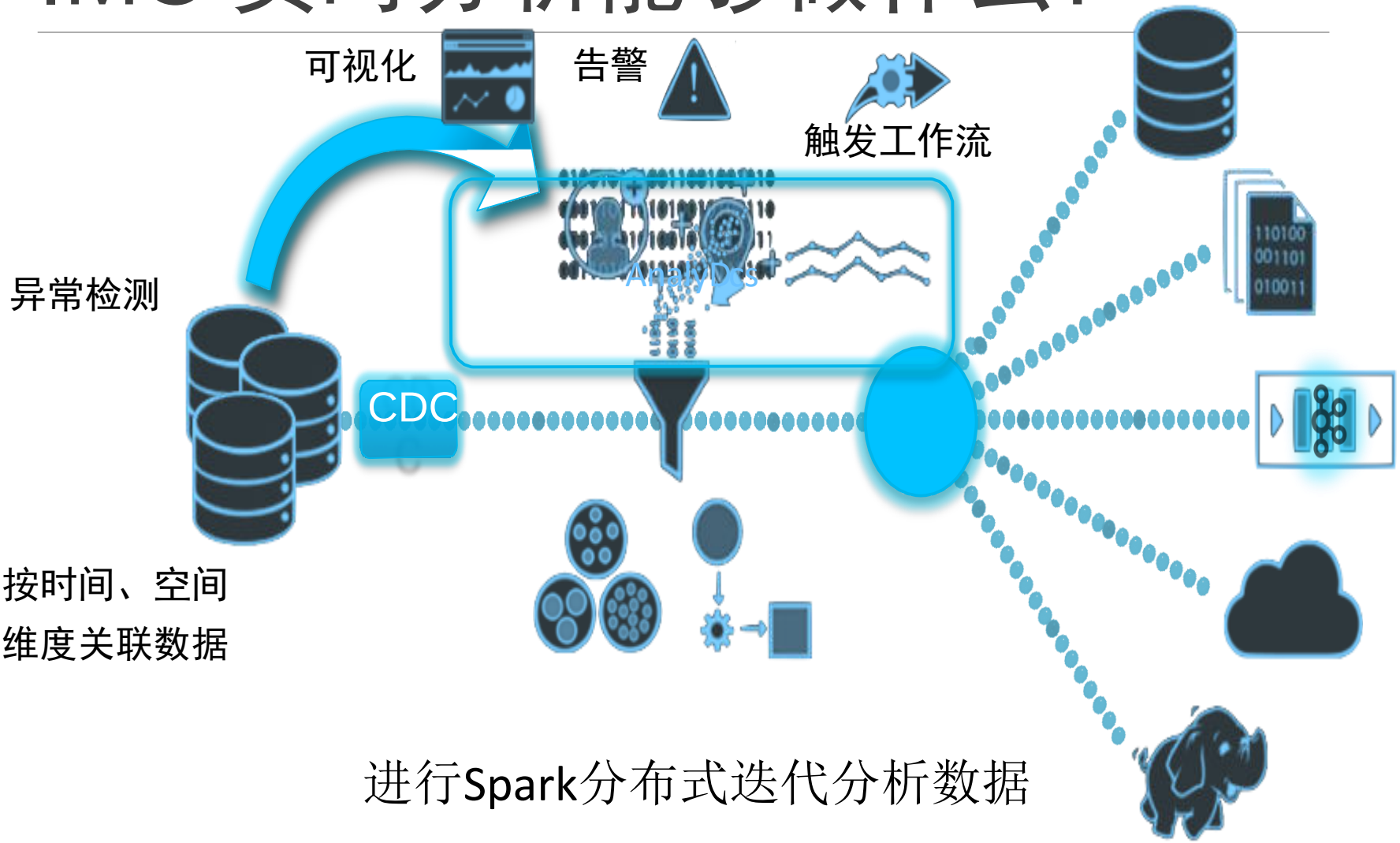


IMC 处理能够做什么?



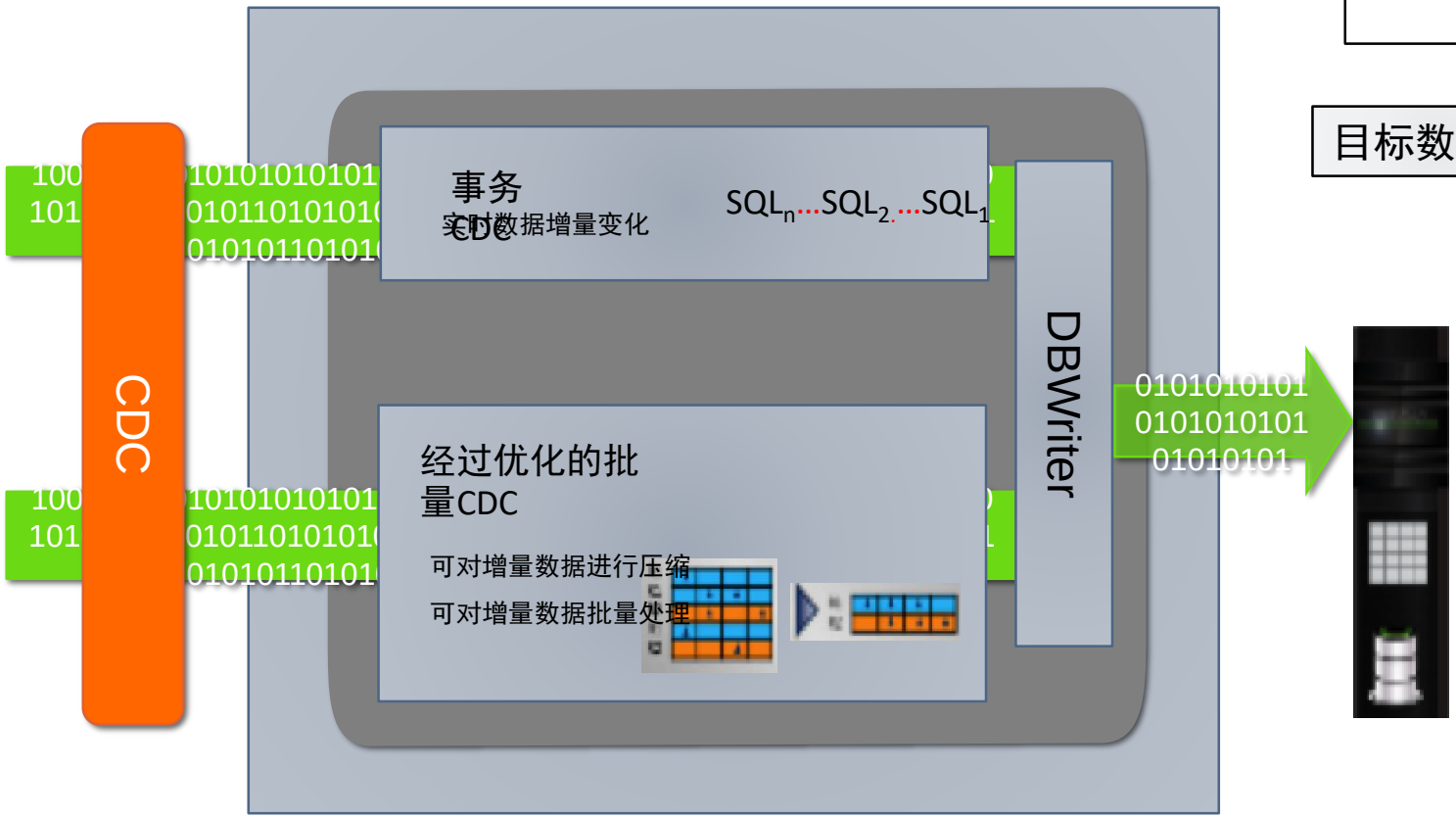
过滤、关联、转换、清洗、汇聚数据

IMC 实时分析能够做什么？

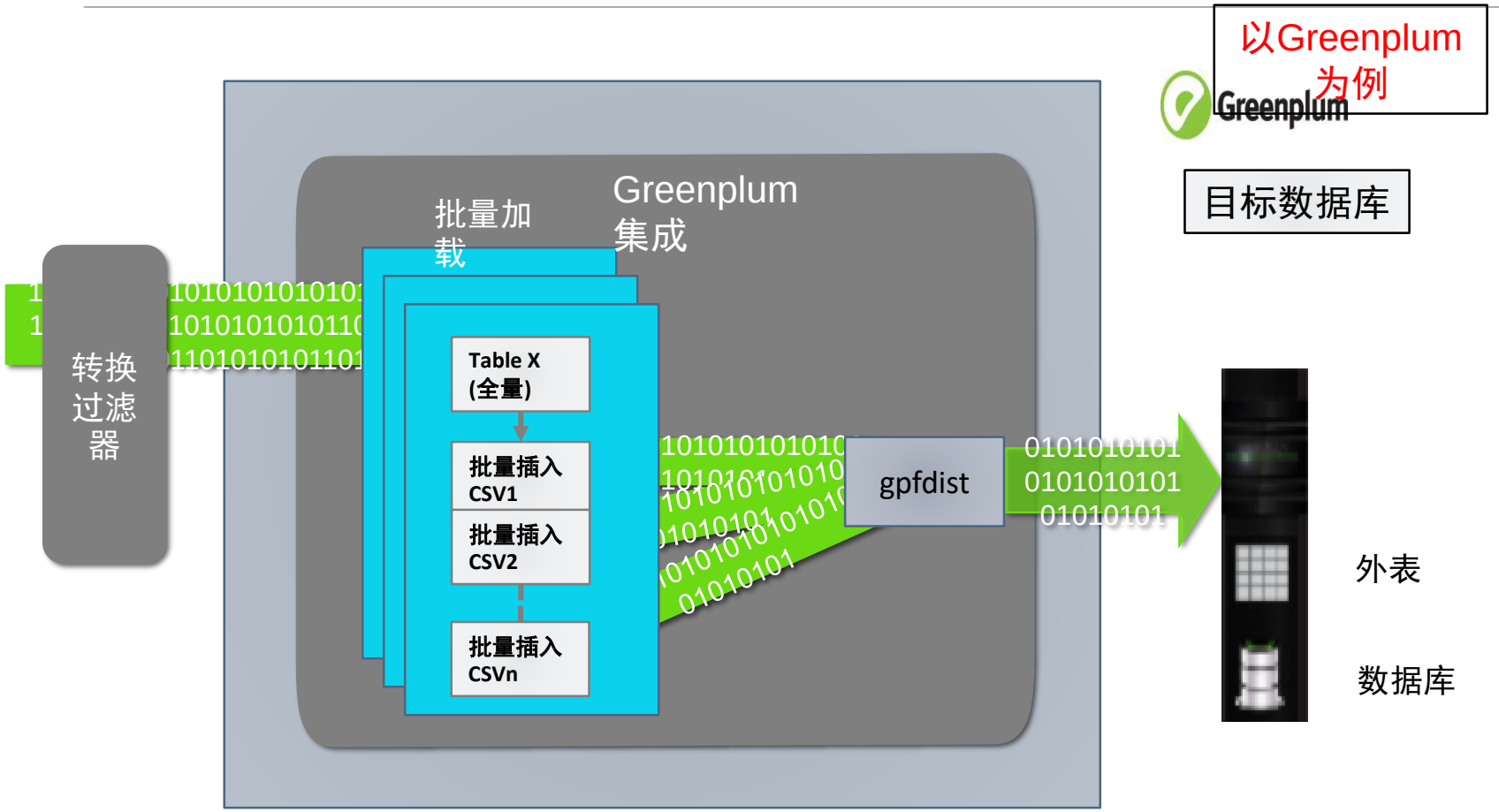


BSP 流数据处理架构原理

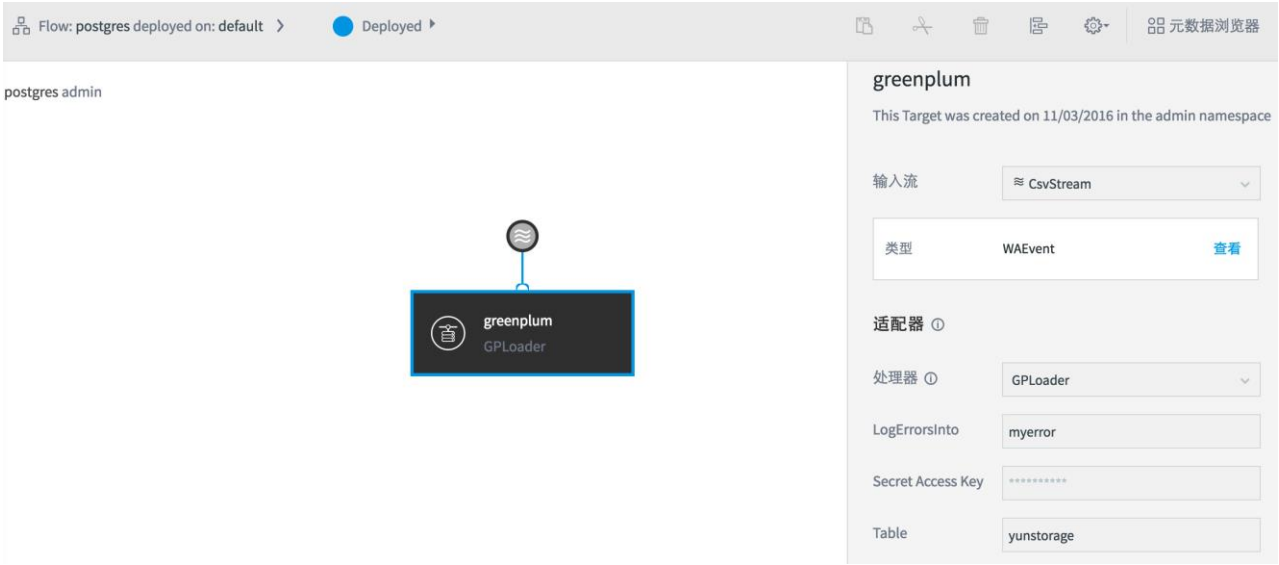
以Greenplum为例



BSP 流数据处理架构原理

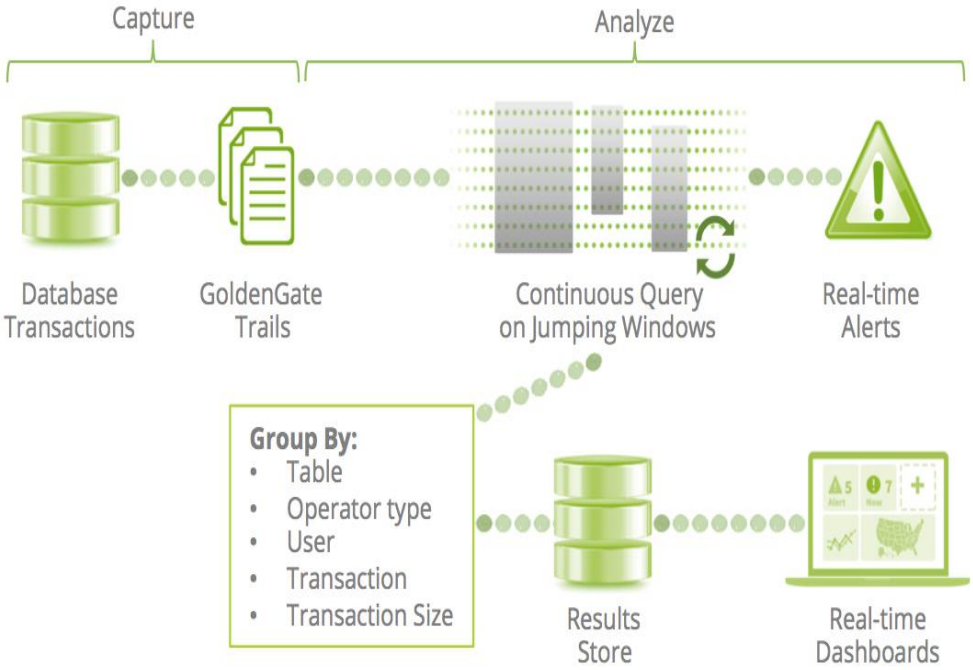


并行加载OSS的TXT到Greenplum



OSS与Greenplum参数配置

BSP CDC处理流程



OGG 适配器：

从OGG或数据库中采集增量数据。

Window 窗口：

定义基于时间窗口的数据分析。

CQ 查询：

通过消息订阅的方式支持消息流查询。

流数据保存：

流数据可以保存下来，供应用查询检索使用。

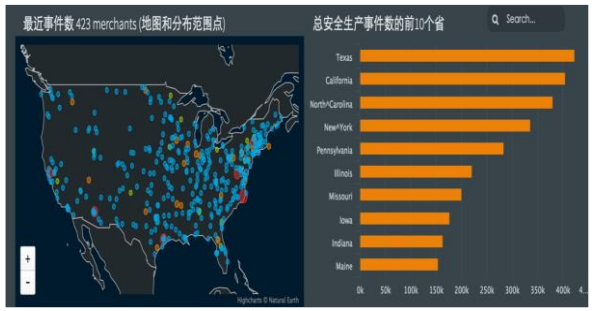
BSP 大数据算法分析

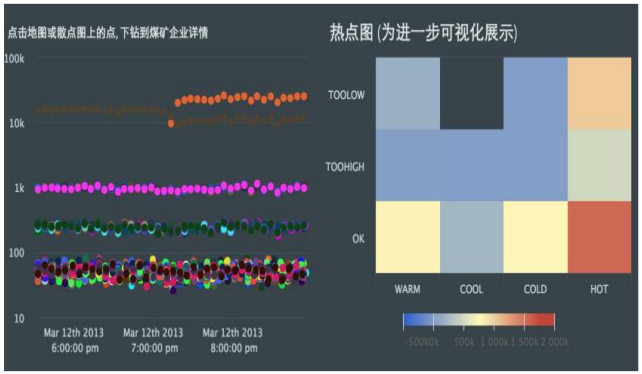
在数据分析方面，本平台支持数据解析、模型分析和模型获取结果等流程。在模型上，包含通用统计、分类、聚类、人工神经网络、朴素贝叶斯、向量分析、回归模型和随机时序分析等数学模型。

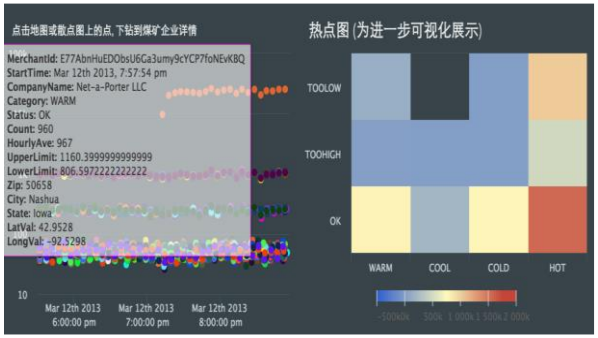
算法名称	算法类型
决策树	分类算法
梯度提升决策树	分类算法
随机森林	分类算法
Logistic 回归	分类算法
支持向量机	分类算法
多项式朴素贝叶斯	分类算法
回归树	回归算法
梯度提升回归树	回归算法
回归森林	回归算法
K 均值	聚类算法

BSP 数据可视化展示

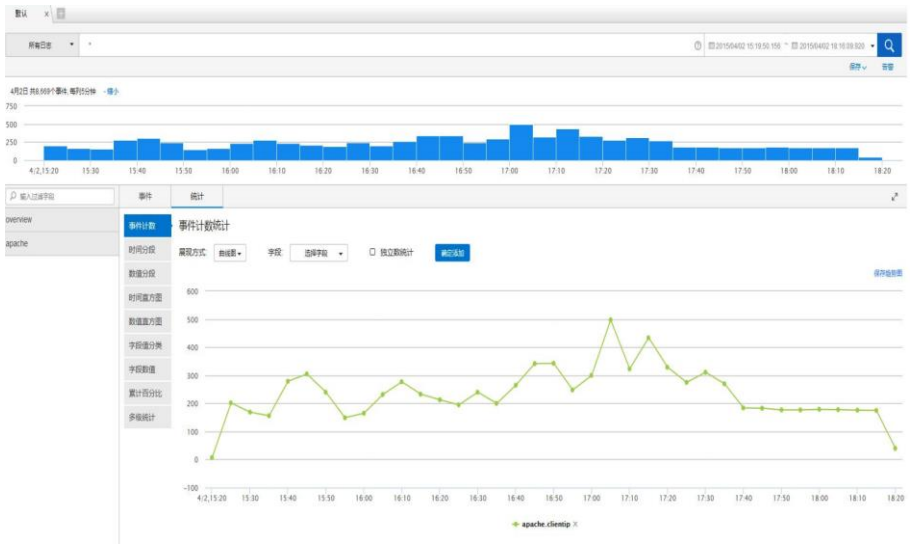
在数据可视化展示方面，本平台支持根据具体视觉效果展示为柱状图、折线图、散点图、K线图、饼图、雷达图、和弦图、力导布局图、地图、仪表盘、漏斗图、孤岛图、多维度堆积图等。







BSP 统一日志分析和监控



非结构化转结构化：

可将非结构化的日志数据转换为结构化数据。

实时查询：

可对日志中的关键信息进行SQL流查询。

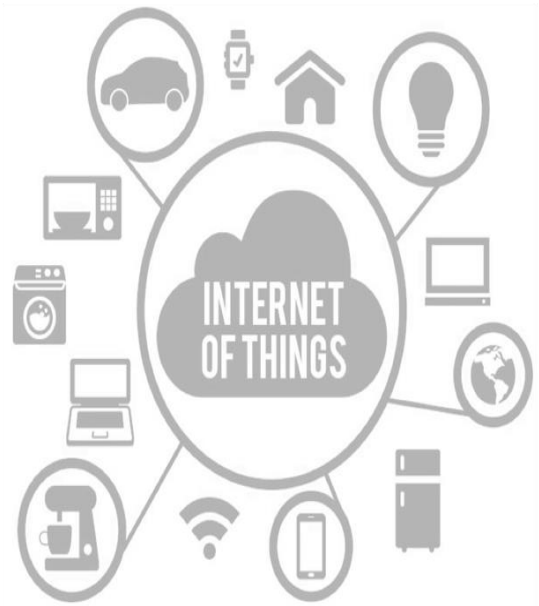
日志关联分析：

可将多种日志进行关联分析。

日志检索：

通过ES可对日志进行快速检索。

BSP IoT数据采集和处理



设备传感器数据采集：

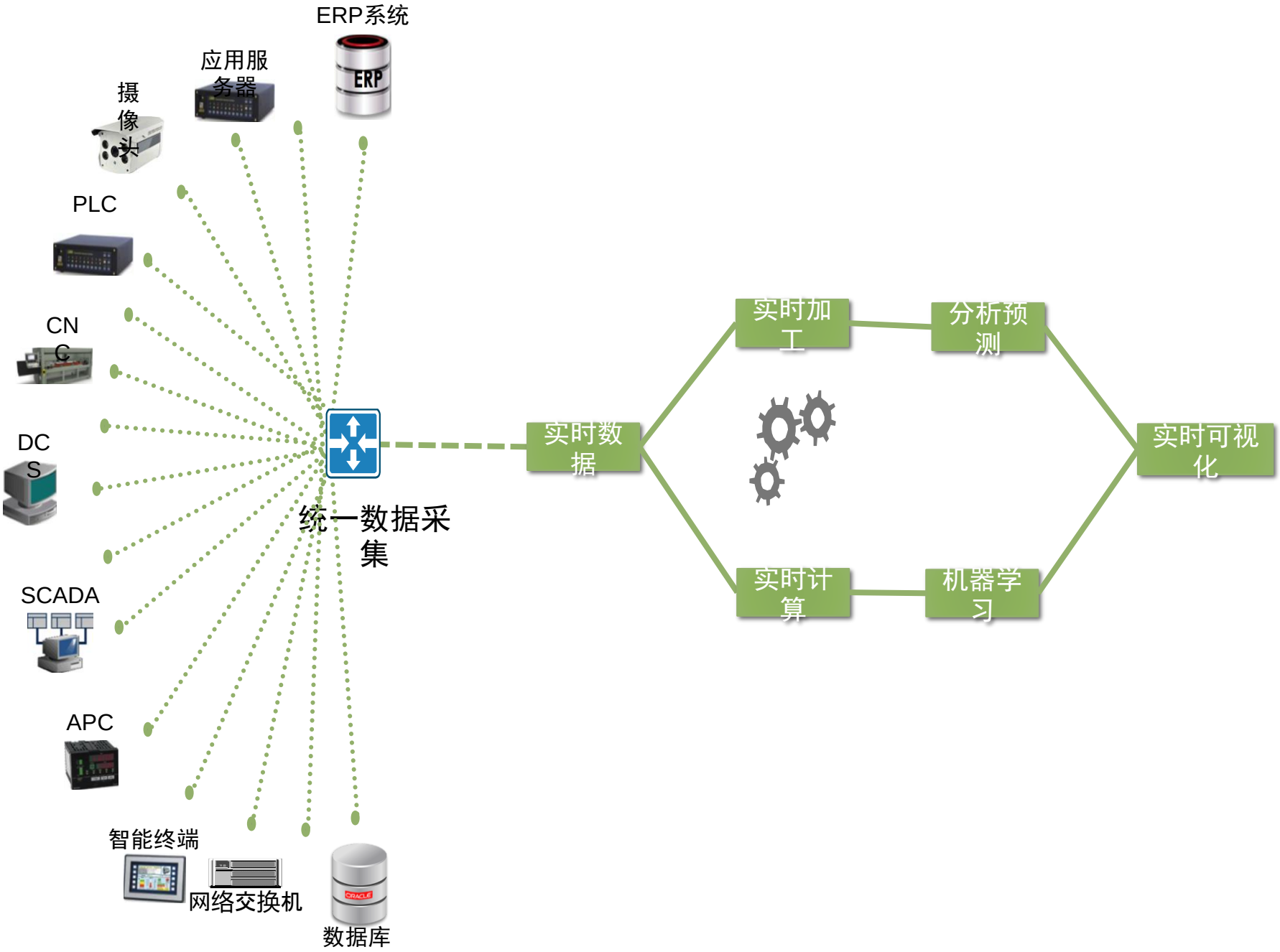
可对机器数据进行采集。

设备传感器数据分析：

通过分析引擎可对数据进行分析。

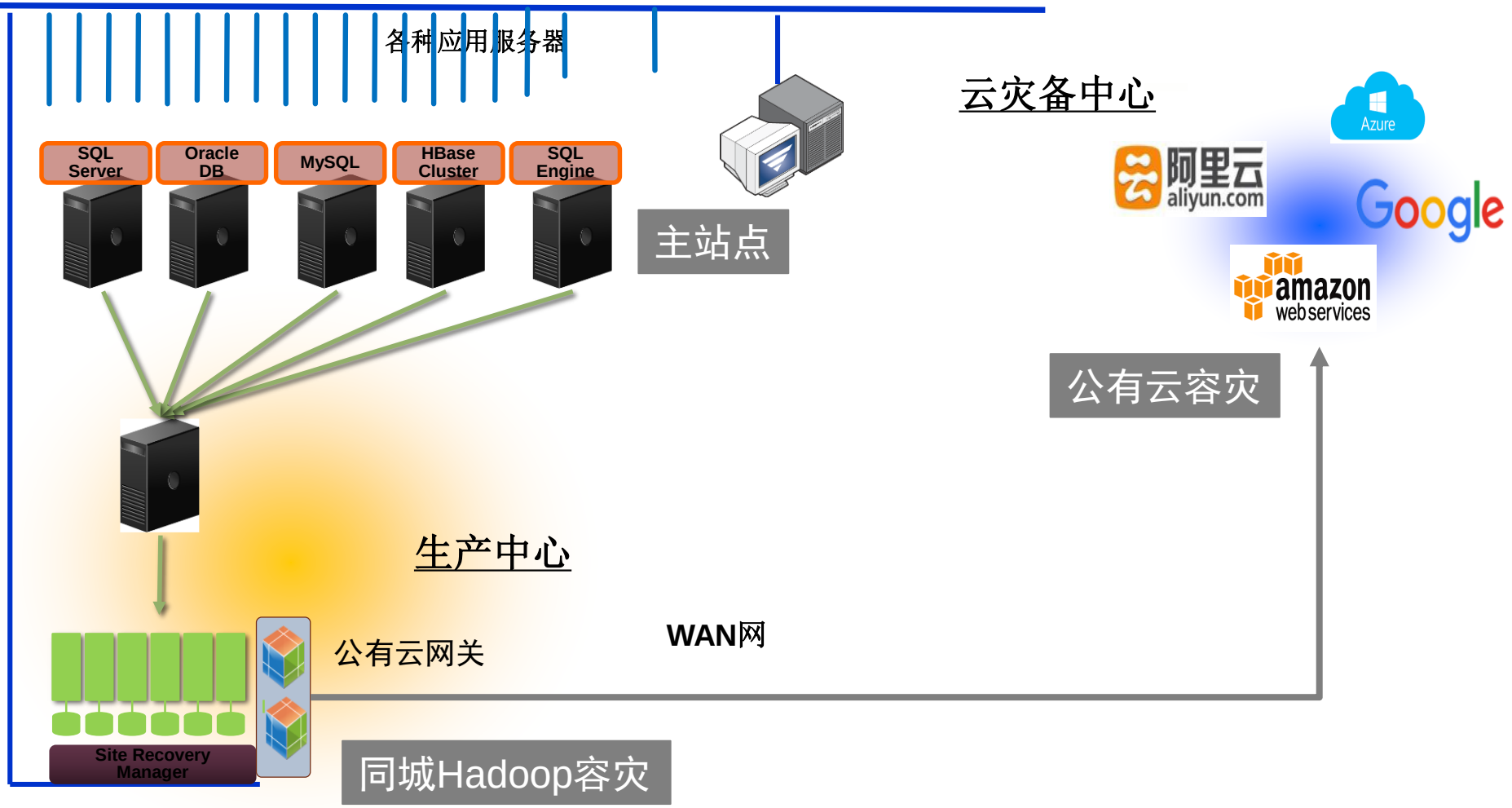
位置数据实时处理：

可对位置数据进行实时处理，反馈给终端设备路径规划或位置结果。



公有云支持

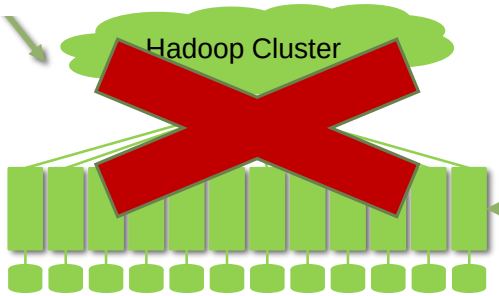
大数据混合云容灾架构介绍



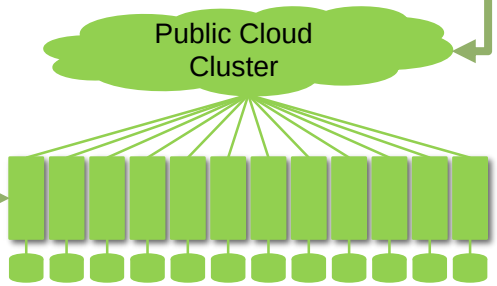
高可用性

故障后连接

应用连接到集群



事务复制;
集群保持同步



关键业务:

- 呼叫中心
- 电商
- 银行
- 电信
- 金融市场
- ….

关键任务复制 – 双活配置- 事务零损失

故障后自动连接

应用连接到
集群

Hadoop Cluster

应用连接到集群

事务复制;
集群保持同步

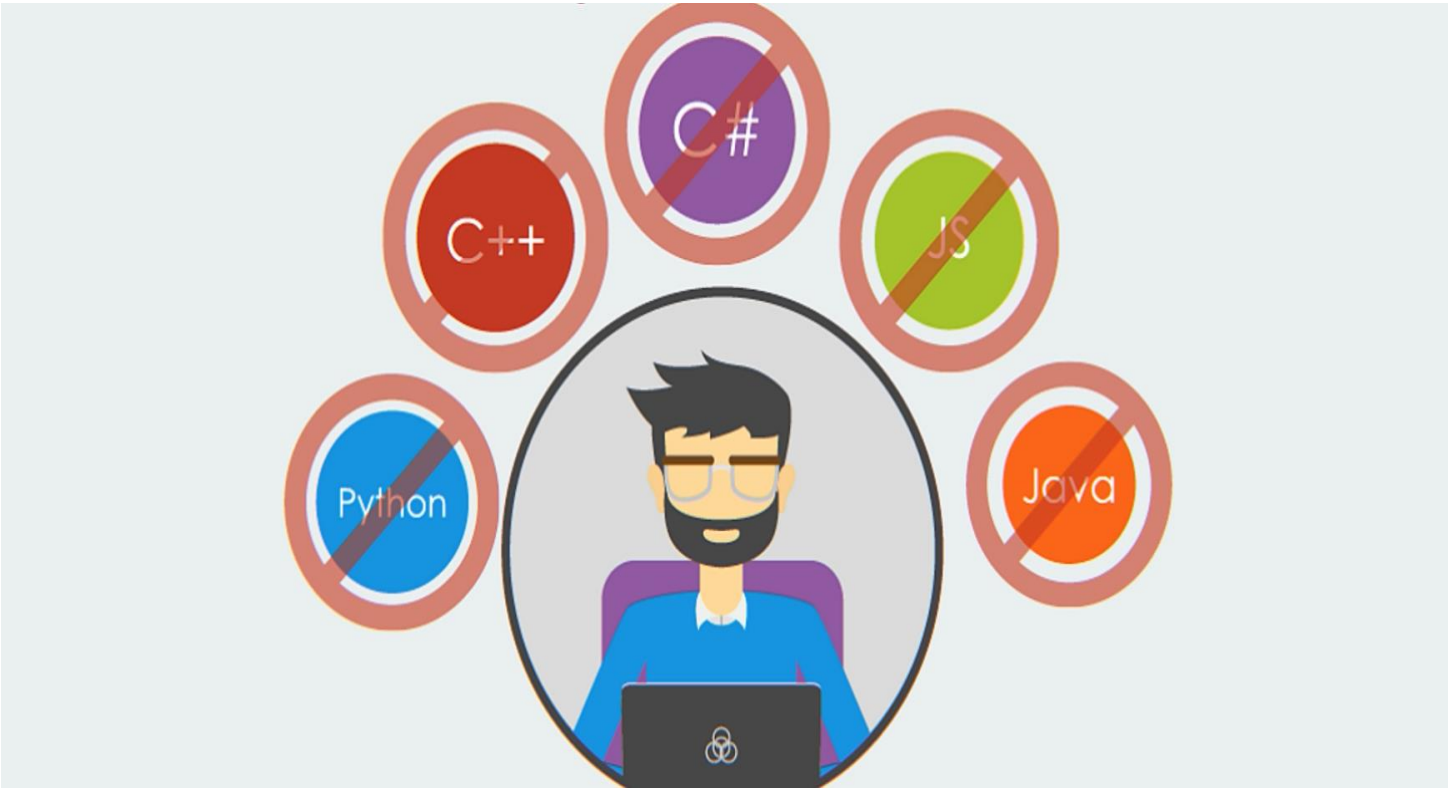
Public Cloud
Cluster

关键业务:

- 呼叫中心
- 电商
- 银行
- 电信
- 金融市场
- ….

彻底解放软件开发的生产力

DBA Friendly



感谢大家