Clickhouse in Telecom (From 0 to 1)

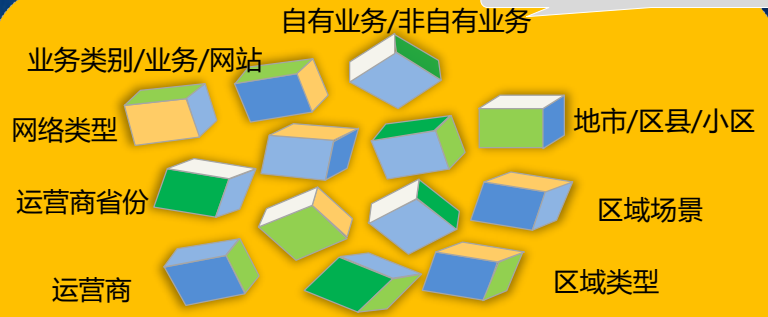
— Dataliance

中国电信G网数据分析典型应用场景

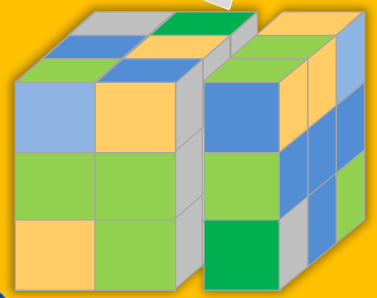
■多维度的用户行为特征分析

对数据业务流量按区域、区域类型、区域场景、业务、终端、SP等多维度进行组合分析，以便掌握用户行为特征

丰富的用户行为特征信息



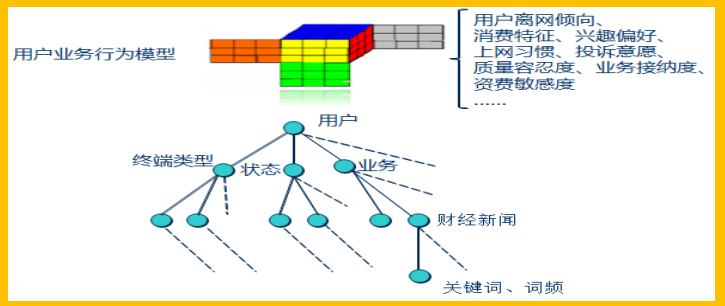
多种维度组合分析



- 2G/3G+区域+业务类型+终端
- 运营商+运营商省份+业务类型
- 区域+区域类型/区域场景+业务
- 区域+区域类型/区域场景+终端
- 区域+自有业务/非自有业务

■基于用户个体的业务消费模型分析

通过业务模型分析、终端业务分析、以及用户区域分析，建立起从业务服务提供端至用户终端的分析手段，再结合经分、BOSS等系统中的业务信息以及用户信息后，就能够实现基于用户个体的业务消费模型分析，进而达到为市场实现精细化营销的目的



中国电信G网数据分析总体技术架构

应用层

用户行为分析 | 质量分析 | 网络安全分析 | 单用户记录

分析引擎

PPP 激活记录	网页浏览记录	邮件收发记录	区域位置信息
寻呼记录	文件下载记录	IP 电话记录	域名信息
重定位记录	重定位记录	……	……
网络信令记录	业务拨测记录	业务信令记录	号码段信息

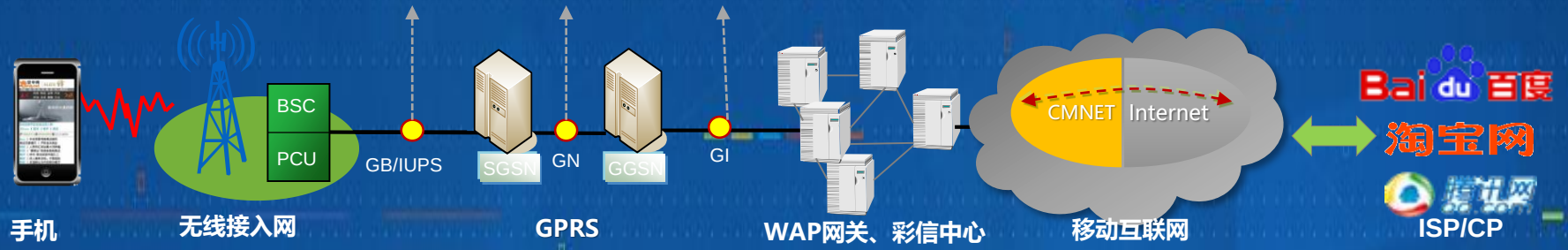
ETL层

按照业务规则清洗数据，完成各类业务数据的抽取、转化、加载

采集层

数据捕获 | 会话管理 | 信令解析 | 业务识别 | CDR合成

网络层



电信级数据处理规模

数据处理规模：

- Ingesting from 网络基站设备、监控设备、骨干网等数据
- 50 billions Entries, ~700G/Day
- 分析后的数据结果可实时呈现在用户分析中心

基于位置的服务，网络优化

- 网优
  - 例如. 重新路由来电到另外一个基站， 如果检测到有网络拥塞存在
- 基于位置的营销
  - 匹配点击事件到订阅者资料; 如果匹配 说明是位置敏感性广告
- 挑战: 交互式实时控制台
  - 简单的规则 - *(CallDroppedCount > threshold) 然后告警*
  - 或者, 复杂 (OLAP 查询)
  - *TopK, 趋势分析, Join查询, 与历史数据关联*

目前的查询场景

需要强大的查询分析引擎

Comparison with Clickhouse and Hadoop

Why Choose Clickhouse? Drop Hadoop

- Hadoop Cluster has a poor performance , that is too slow to be valid.
- Hadoop Cluster is fat.
- Cant execute to query data (PB) in real time.

Clickhouse

- Rich functions
- Perfect performance
- Structure flat not fat
- Flexible way in execution

Clickhouse功能特点与优势

数据库内压缩

采用了业内领先的压缩技术，提高性能的同时，显著地减少存储数据所需的空间。客户可以将所用空间减少3-10倍，并提高有效的I/O性能。

千万亿字节规模的数据加载操作

高性能的并行数据装载器可以在所有节点上同步执行操作，装载速度超过50W条/秒。

随地访问数据

不管数据的位置、格式或存储介质如何，都可以从数据库向外部数据源执行查询操作，并行向数据库返回数据。

动态扩展

对数据仓库进行便捷的小规模或大规模扩展，同时避免高成本的设备或SMP服务器升级。

集中管理

提供集群级管理工具和资源，帮助管理人员像管理一台服务器一样管理整个多维实时分析平台。

G 网大数据平台架构演化

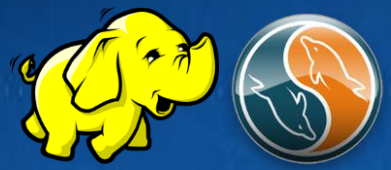
2000~2009

Oracle

2009~2013

Greenplum

2013~2017



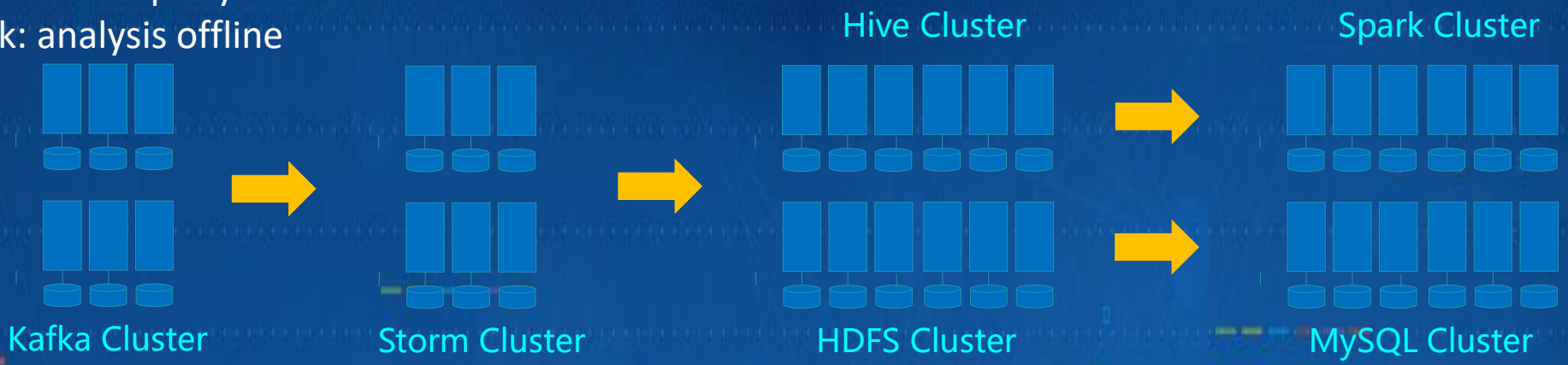
Hadoop+MySQL

2017+

Clickhouse+Mongo

Technology Architecture Before Clickhouse

Legacy Architecture： Kafka+Storm+Hive+Spark+MySQL
Kafka: collect and aggregate data
Storm: wrangle data
Hive: ad-hot query data
Spark: analysis offline



Make a Migration to Clickhouse

Speed up ~560X!

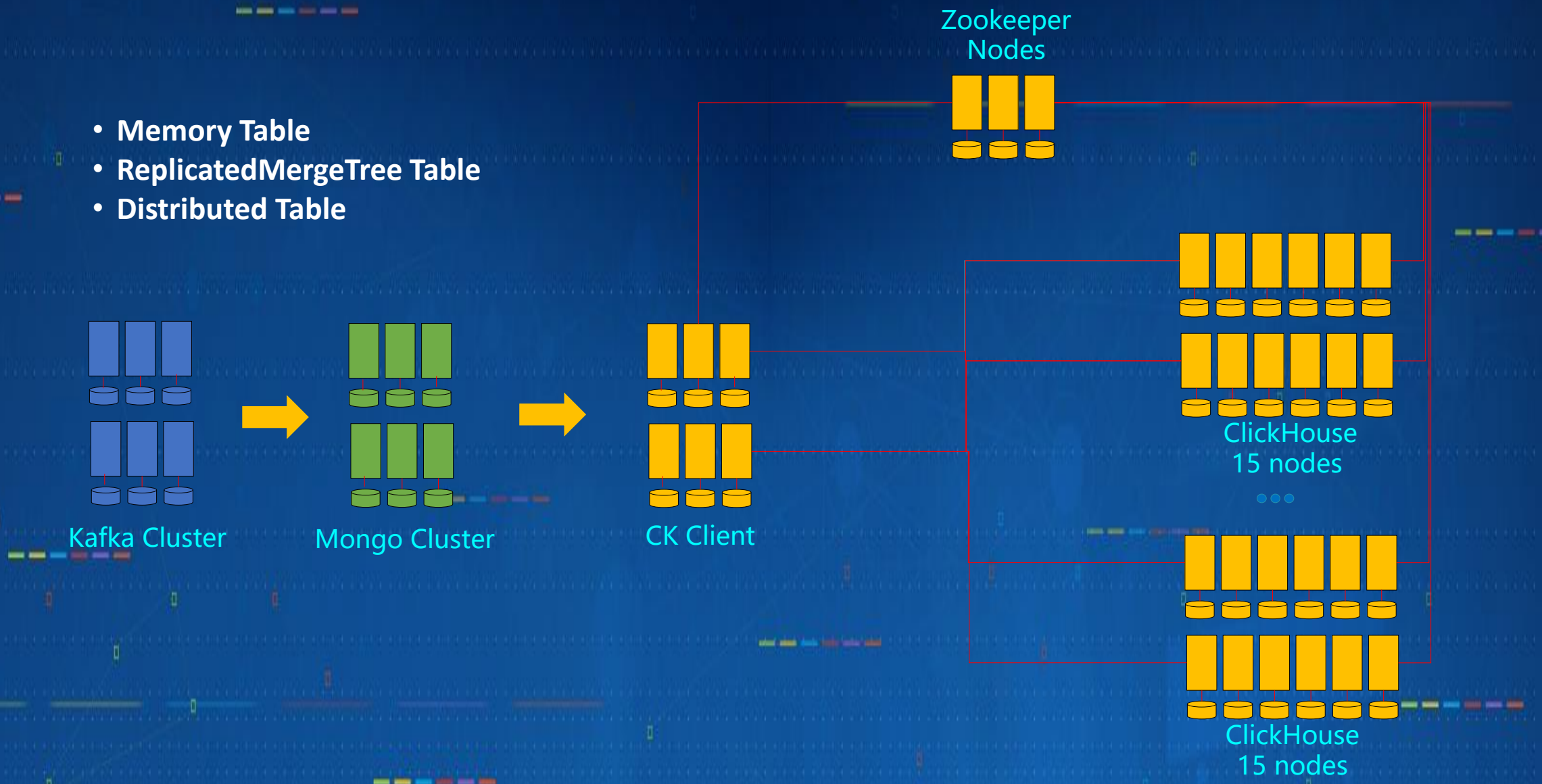
Elapsed Time 80s -0.3s

放弃 HDFS、Hive、Spark联合解决方案

- 全表扫描慢、数据过滤消耗时间
- 离线分析难以处理大数据量
- 体验不好、速度不快
- 难于支持即席查询分析

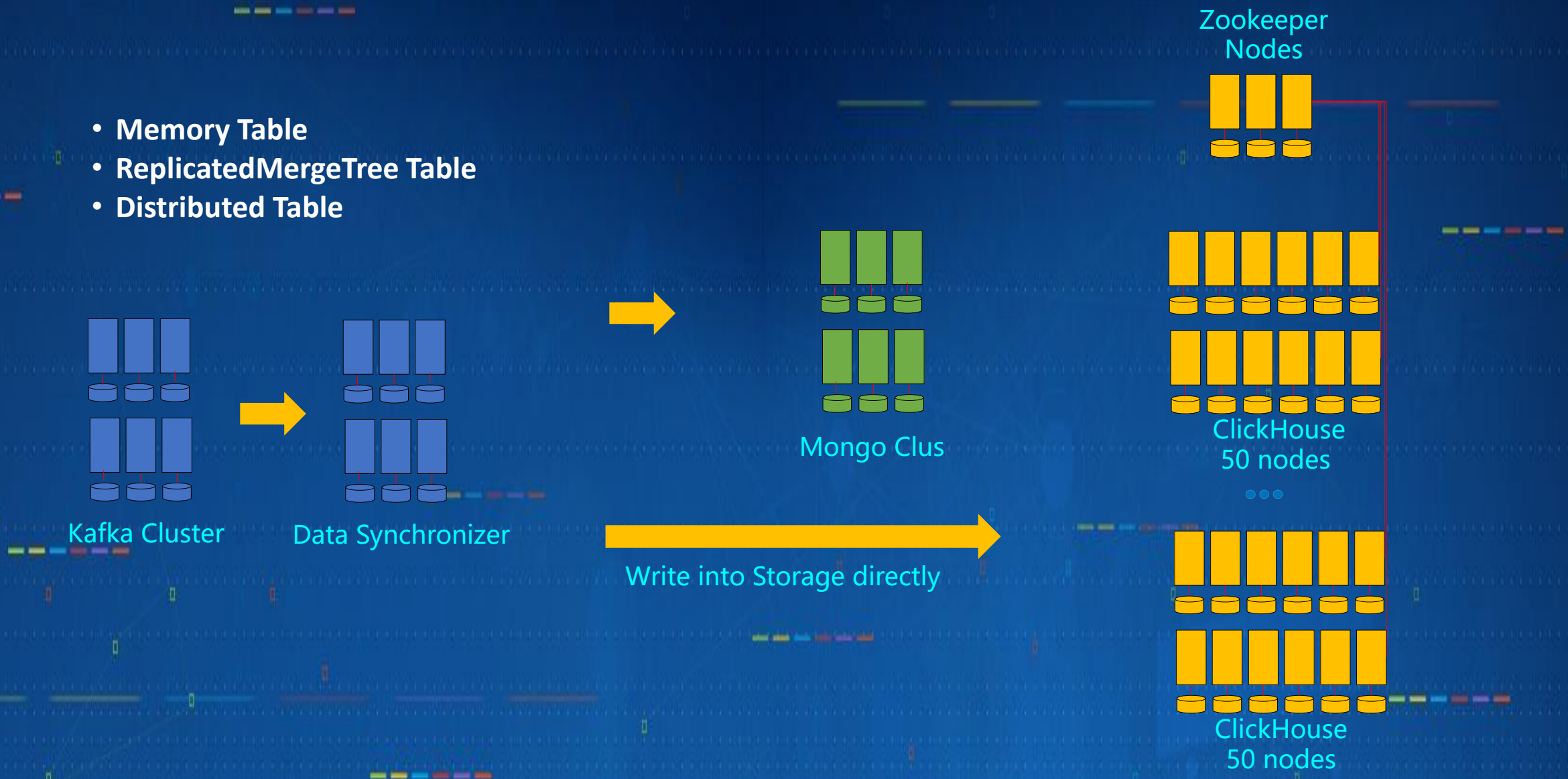
Technology Architecture After Clickhouse

- Memory Table
- ReplicatedMergeTree Table
- Distributed Table

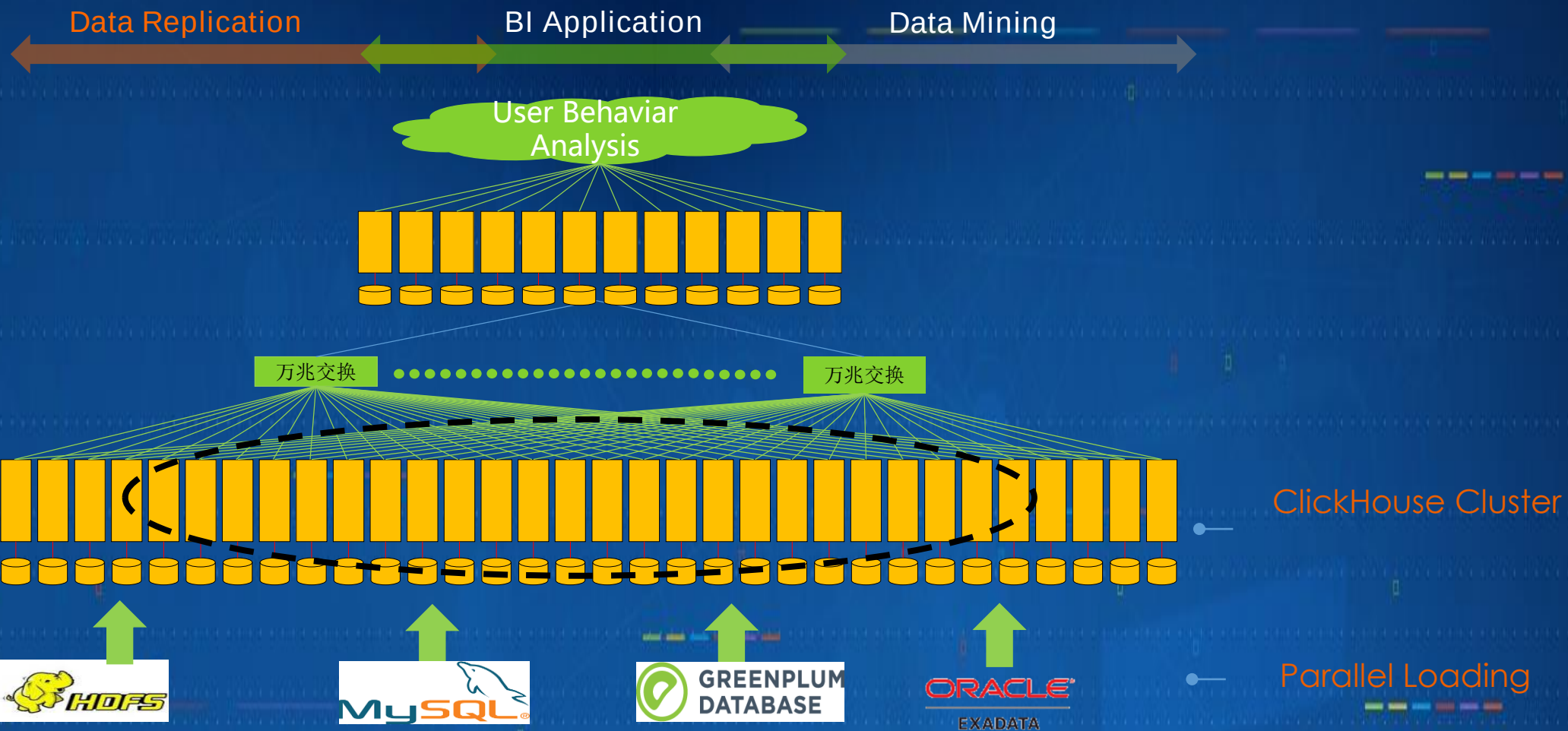


Technology Architecture In the furture

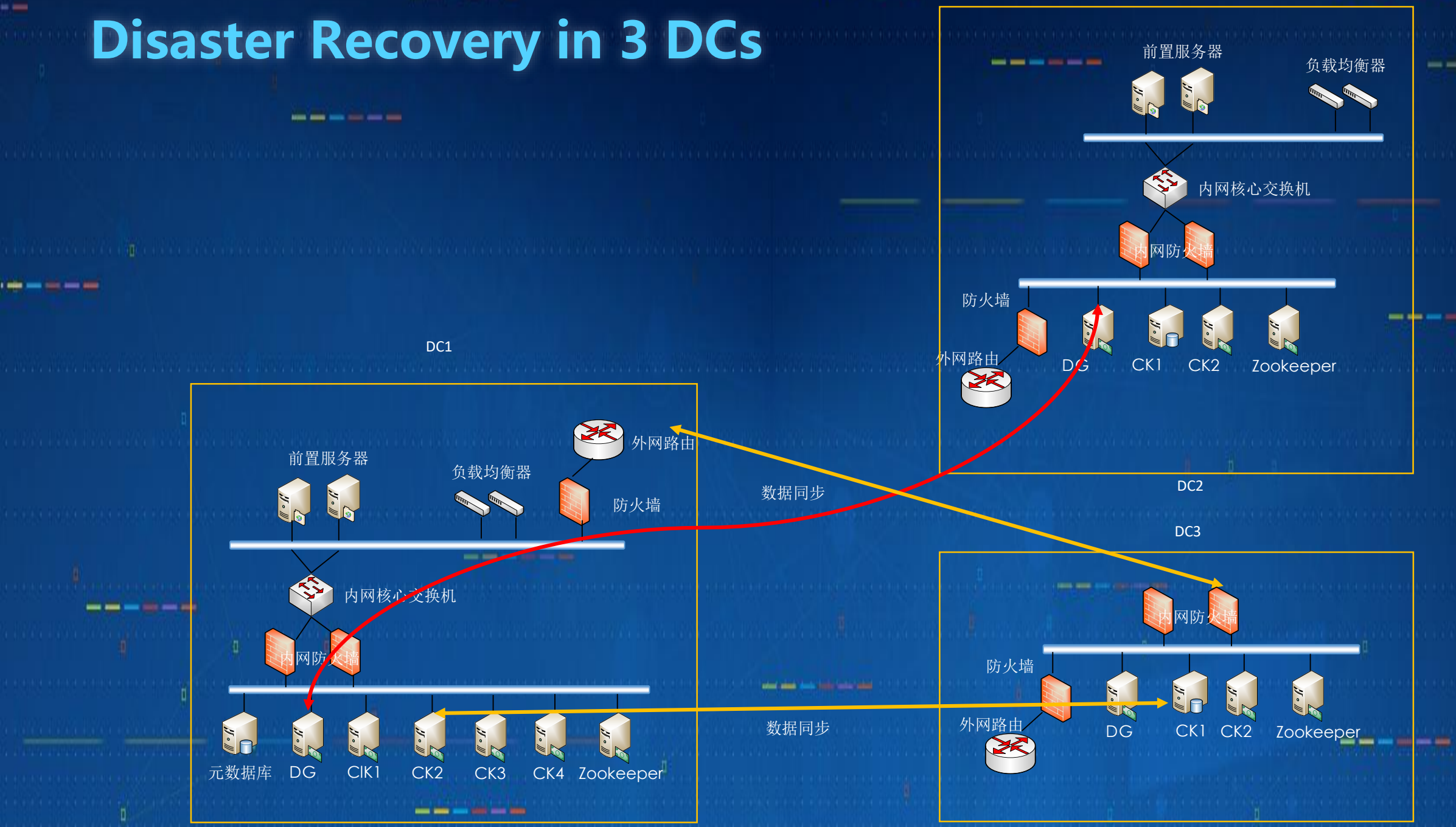
- **Memory Table**
- **ReplicatedMergeTree Table**
- **Distributed Table**



Technology Architecture In the furture



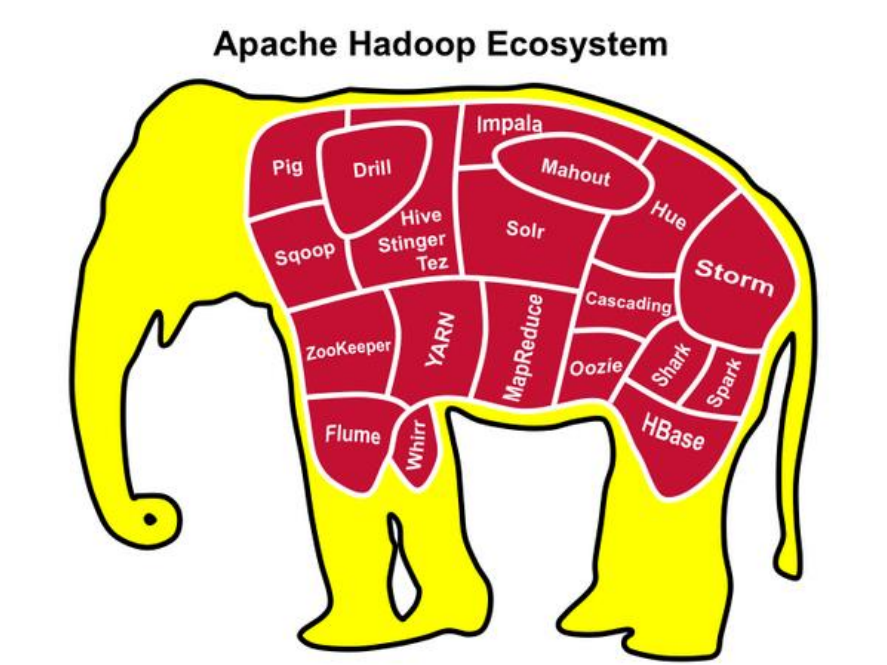
Disaster Recovery in 3 DCs



Need to work

- DML SQL(Update , Delete)
- SQL99/2003
- Automated Operation Tools
- Clickhouse on HDFS(like Hawq)

Kill Hadoop using clickhouse



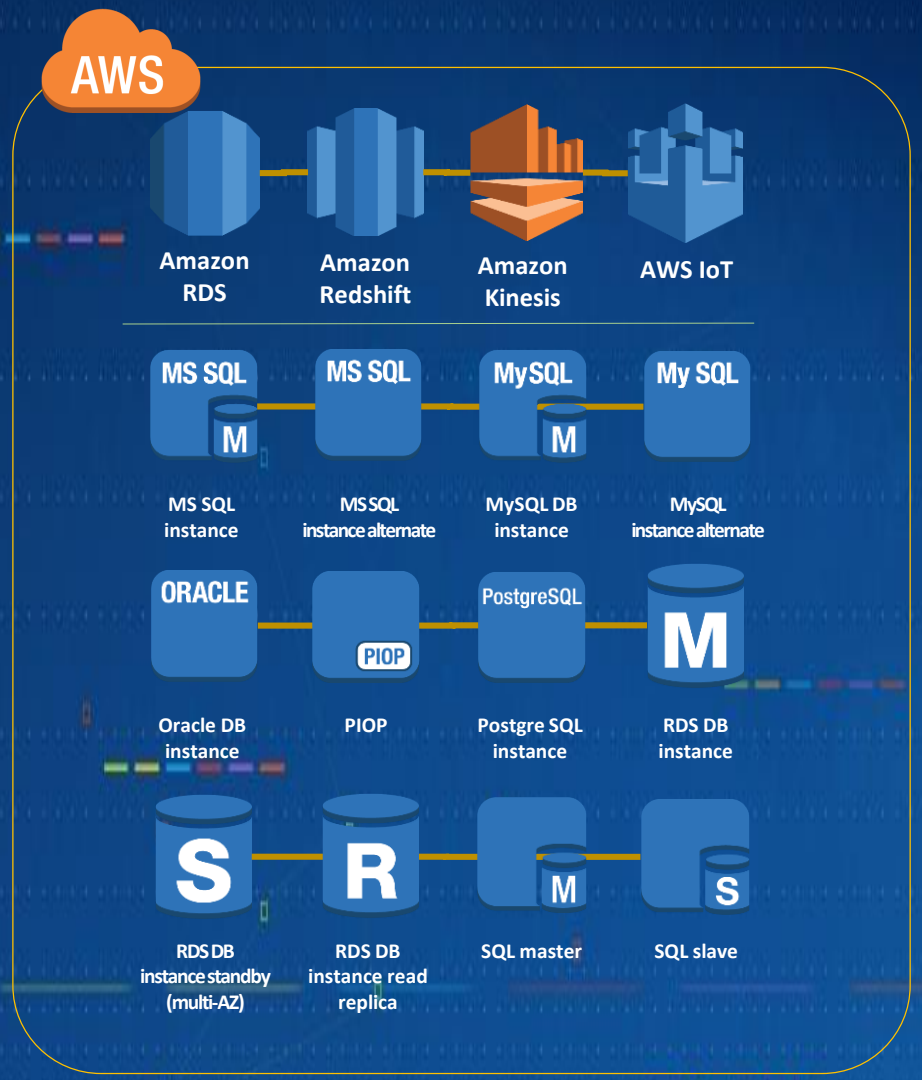
一只大象拆分后，有价值的东西所剩不多

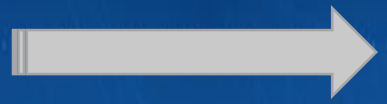
- Kafka
- HDFS
- Spark

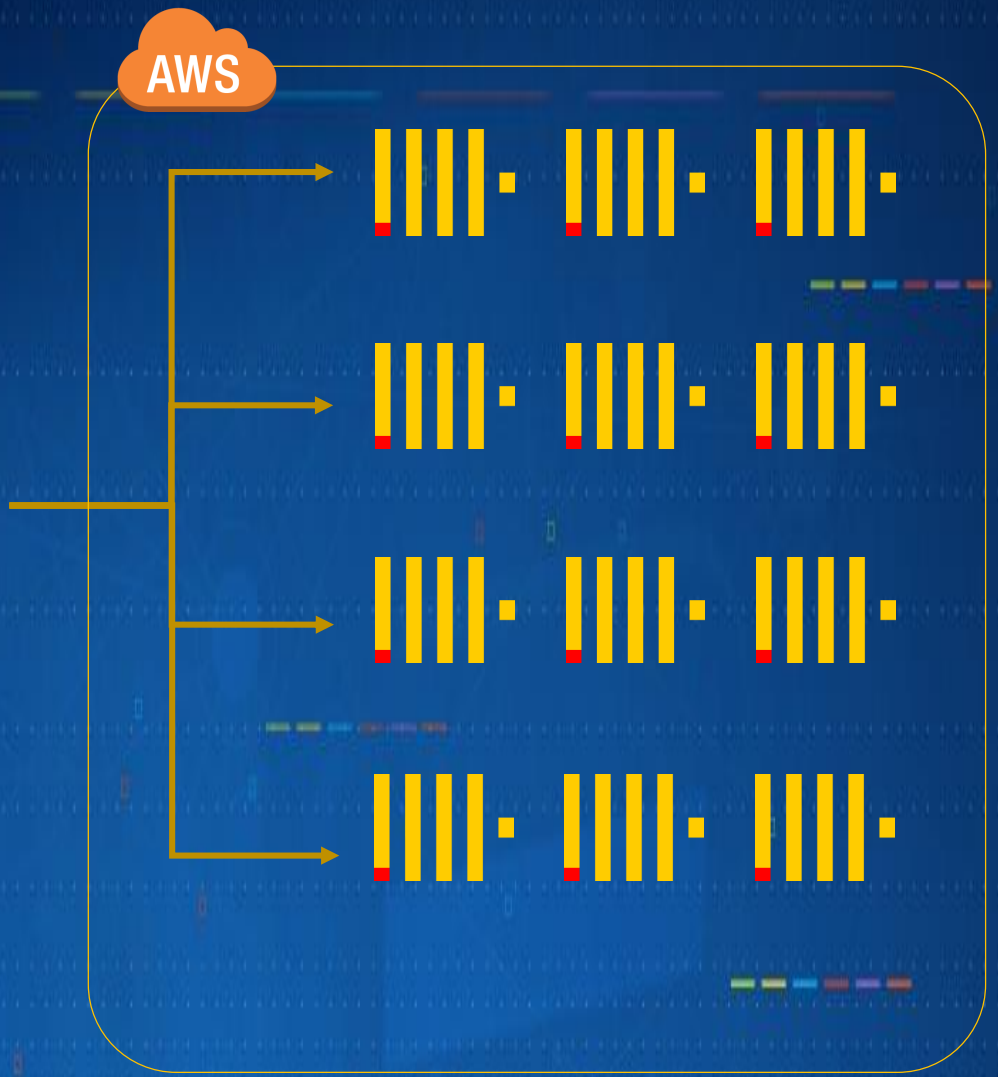
ClickHouse on AWS

提供 ClickHouse Cloud云服务

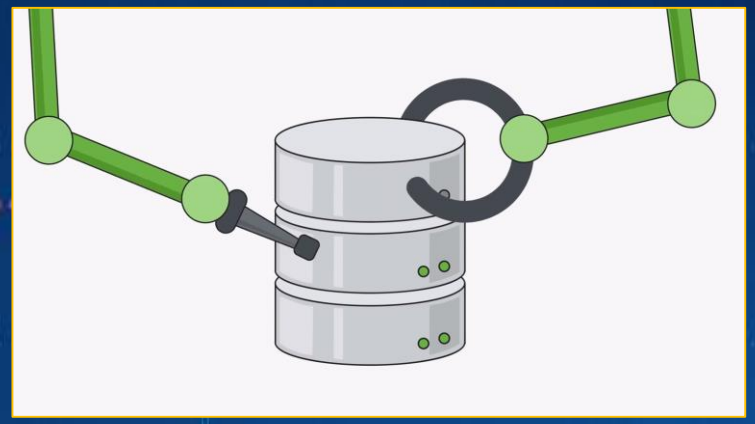
Clickhouse On AWS







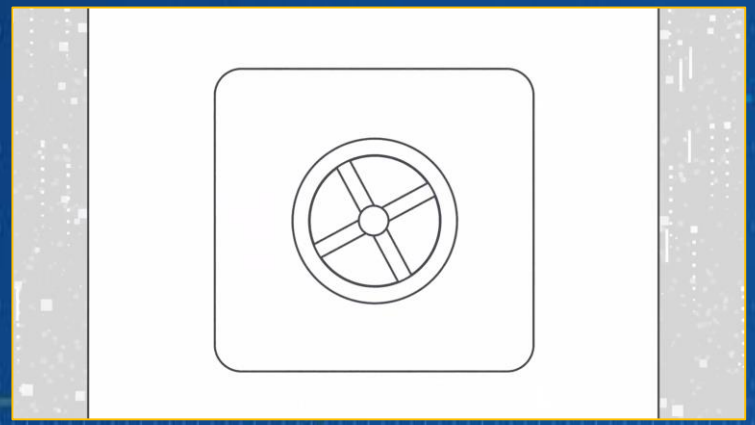
构建Clickhouse as a service平台



自动化

按需使用

水平扩展



安全可靠

高可用

自动备份

公司招聘:

Database kernel developer

Clickhouse integration developer

欢迎加入我们