



第九届中国系统架构师大会  
SYSTEM ARCHITECT CONFERENCE CHINA 2017

# 云+时代大数据平台应用方案

腾讯-陈龙

# 大数据特征



## 数据规模大

企业数据数据规模大部在TB级别以上，像银行电信等行业数据量都在PBI以上，而且每年都是以40%以上的速度增长



## 数据流转快

要在秒级时间范围内给出分析结果，超出这个时间，数据就失去价值了



## 数据类型多

除了以文本为主的结构化数据、以网页数据为代表的半结构数据，也存在大量网络日志、音频、视频、图片、地理位置信息等非结构化数据



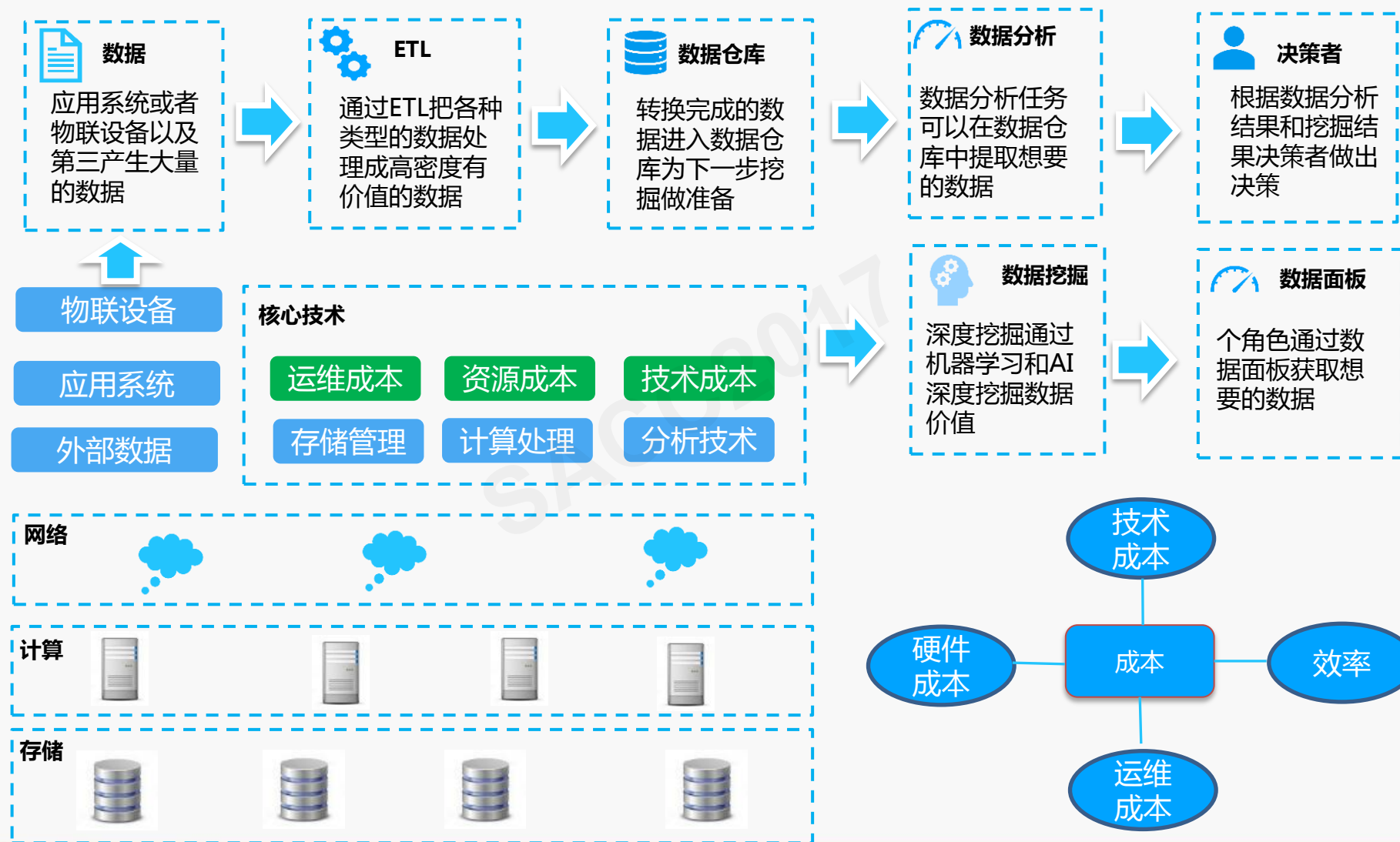
## 价值密度低

海量数据中，如何通过强大的机器算法，更迅速有效地完成数据的价值“提纯”

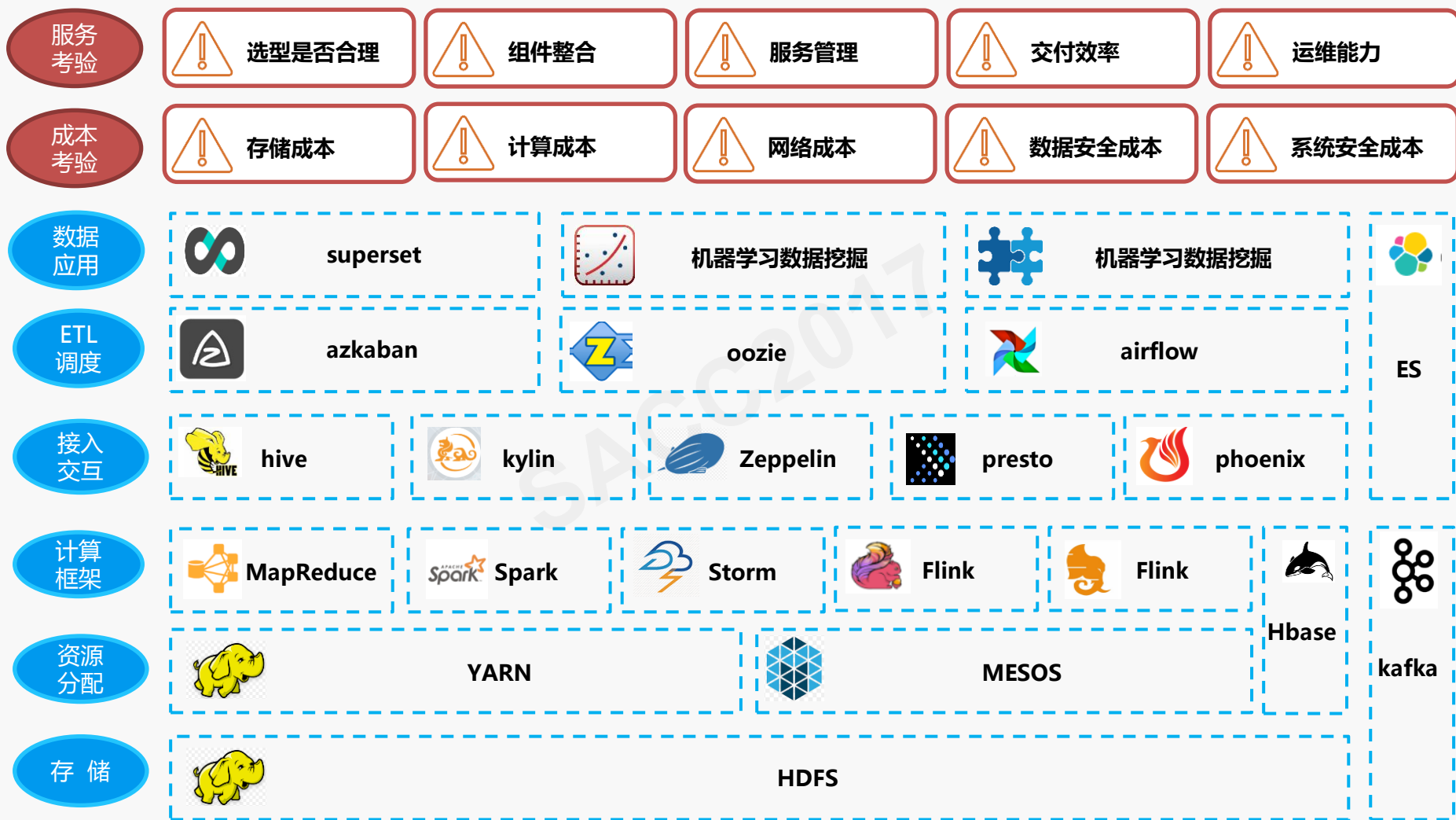
4V



# 企业大数据应用现状



# 大数据解决方案现状



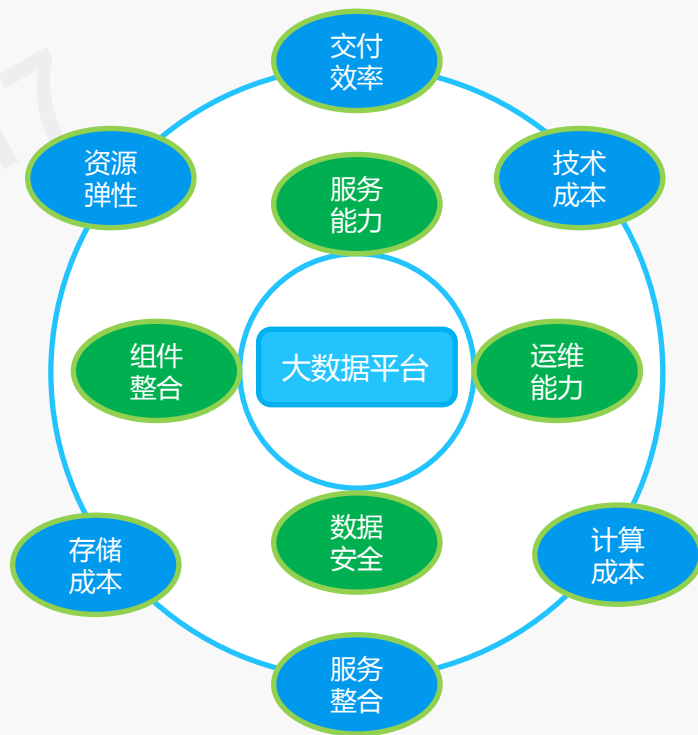
# 大数据解决方案现状

	cloudera®	Hortonworks	社区自建
服务选型	×	×	×
组件整合	部分	部分	×
服务管理	部分	部分	×
交付效率	中	中	低
运维能力	中	中	差
数据安全	×	×	×
技术支持	×	×	×
服务整合	×	×	×
计算成本	高	高	高
存储成本	高	高	高
网络成本	高	高	高



需要什么样的平台


价值最大化、聚焦业务  
成本最小化



# 云环境下的大数据基础平台

 平台服务化


 云消息服务

 kafka

 云服务

....

 企业应用服务

 企业服务

 人工智能深度应用

 机器学习

 数据可视化

 智能BI

 可视交互

 托管Hadoop计算服务

 离线处理

 流式计算

 实时数据库

 ETL

 云服务

弹性

效率

海量

 虚拟网络

安全

高效

 云存储

 云数据库

 对象存储


 KV存储

 文档数据库

 专业技术支持

 海量计算资源保证

 低运维和开发成本

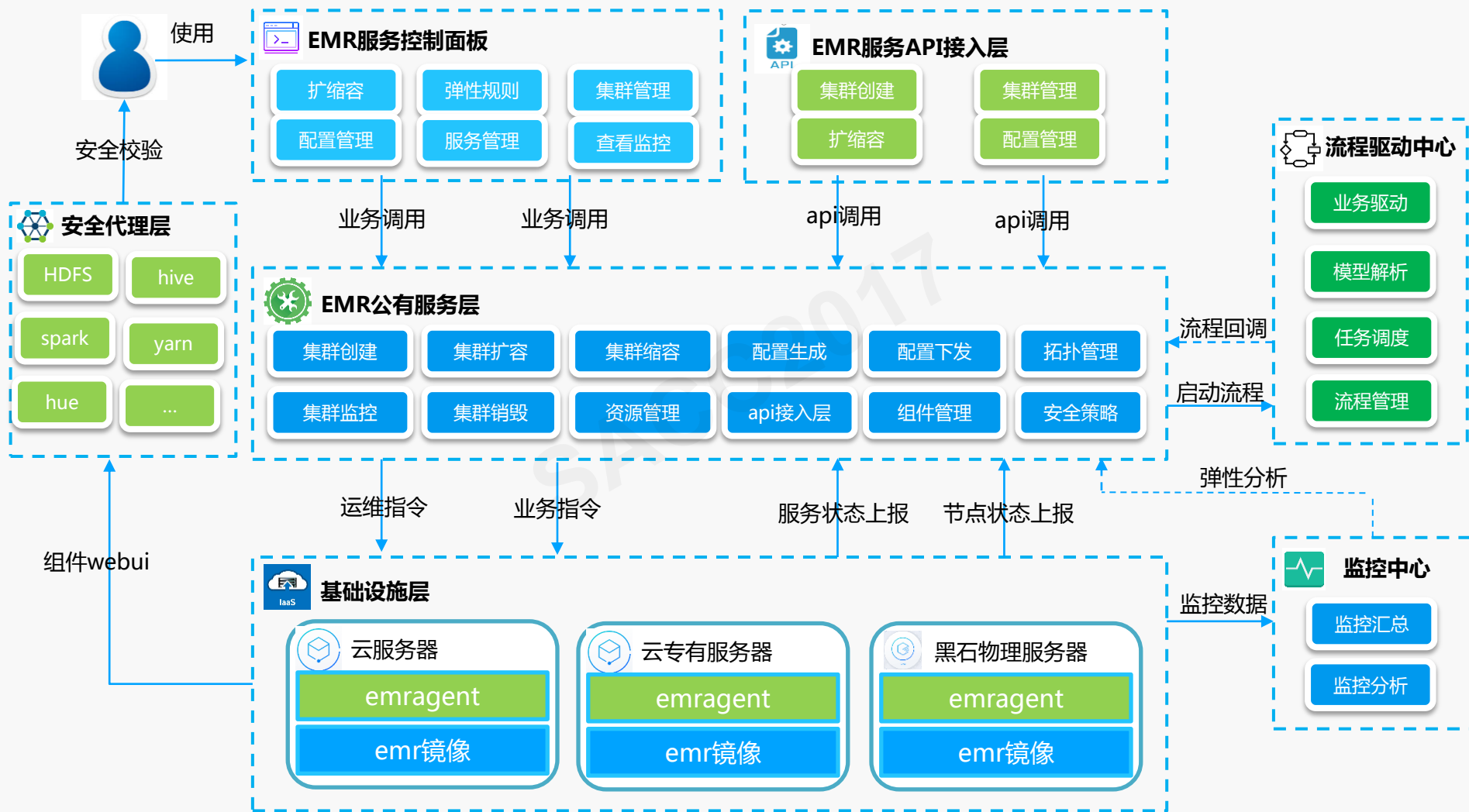
 计算存储分离

 服务深度整合

 快速交付

 资源弹性

# 腾讯云公有云大数据平台实践(EMR)



# 服务流程化



- 流程设计器设计业务流程
- 流程管理系统管理流程
- 流程监控
- 流程告警
- 流程mock
- 通过流程重用业务功能

- 流程热加载
- 自动重试
- 步骤跳过
- 自定义配置
- 简化业务开发
- 业务过程可视

- ❑ 业务实现原子功能
- ❑ 代码高度解耦
- ❑ 代码高度复用
- ❑ 代码维护简单
- ❑ 代码结构高度可扩展
- ❑ 控制逻辑和业务分离



# 服务模型

## 套件集合

服务A

服务B

服务C

服务...



## 套件集合

套件是软件配置的集合，套件内的软件之间的版本兼容性在集成前都做过处理

## 组件集合

hadoop

hive

hbase

....



## 组件集合

组件集合里是一个一个的单个软件，由软件和软件版本组成，比如hadoop-2.7.3

## 服务组

hdfs

yarn

spark

....



## 服务组

服务组是一个软件提供的功能集合，比如hadoop提供了HDFS,YARN,那么HDFS是一个服务组

## 服务节点

datanode

rm

nm

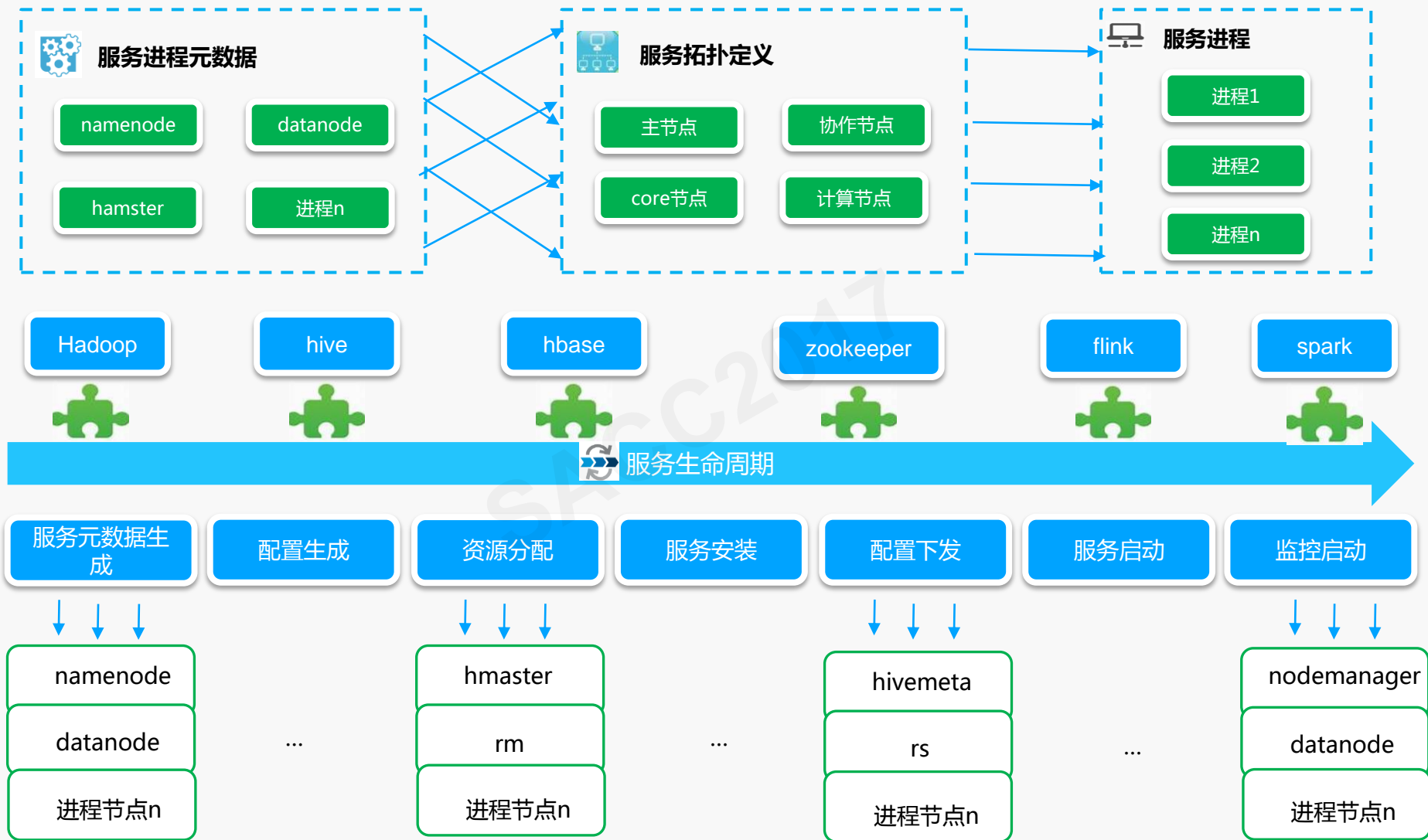
....



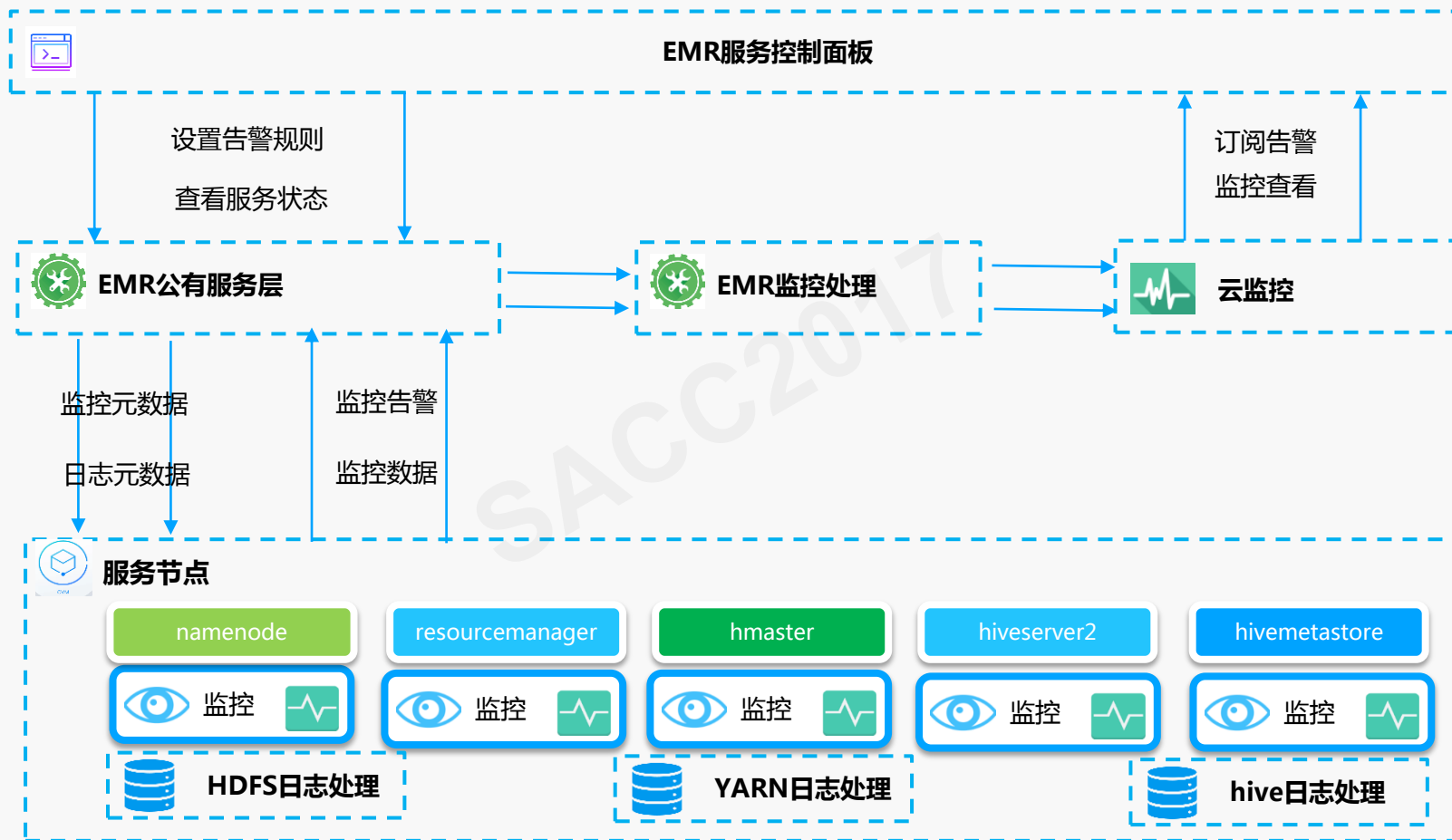
## 服务节点

一个服务组要想提供服务，必须由多种进程提供服务，服务节点可以理解作为一种进程

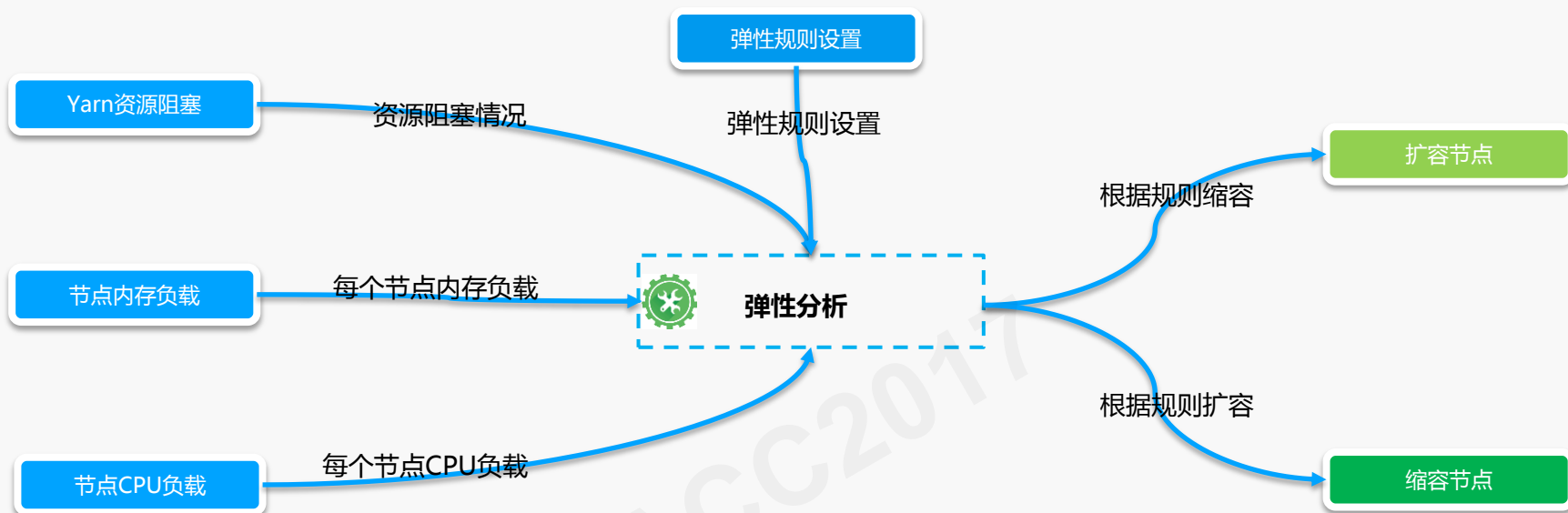
# 服务模型



# 服务管理



# 服务弹性



参数1：任务阻塞因子

$a = \frac{dy}{dt}$   $y$ 为阻塞任务的变化曲线,  $a$ 为任务阻塞变化率

参数2：集群总体CPU负载  
负载

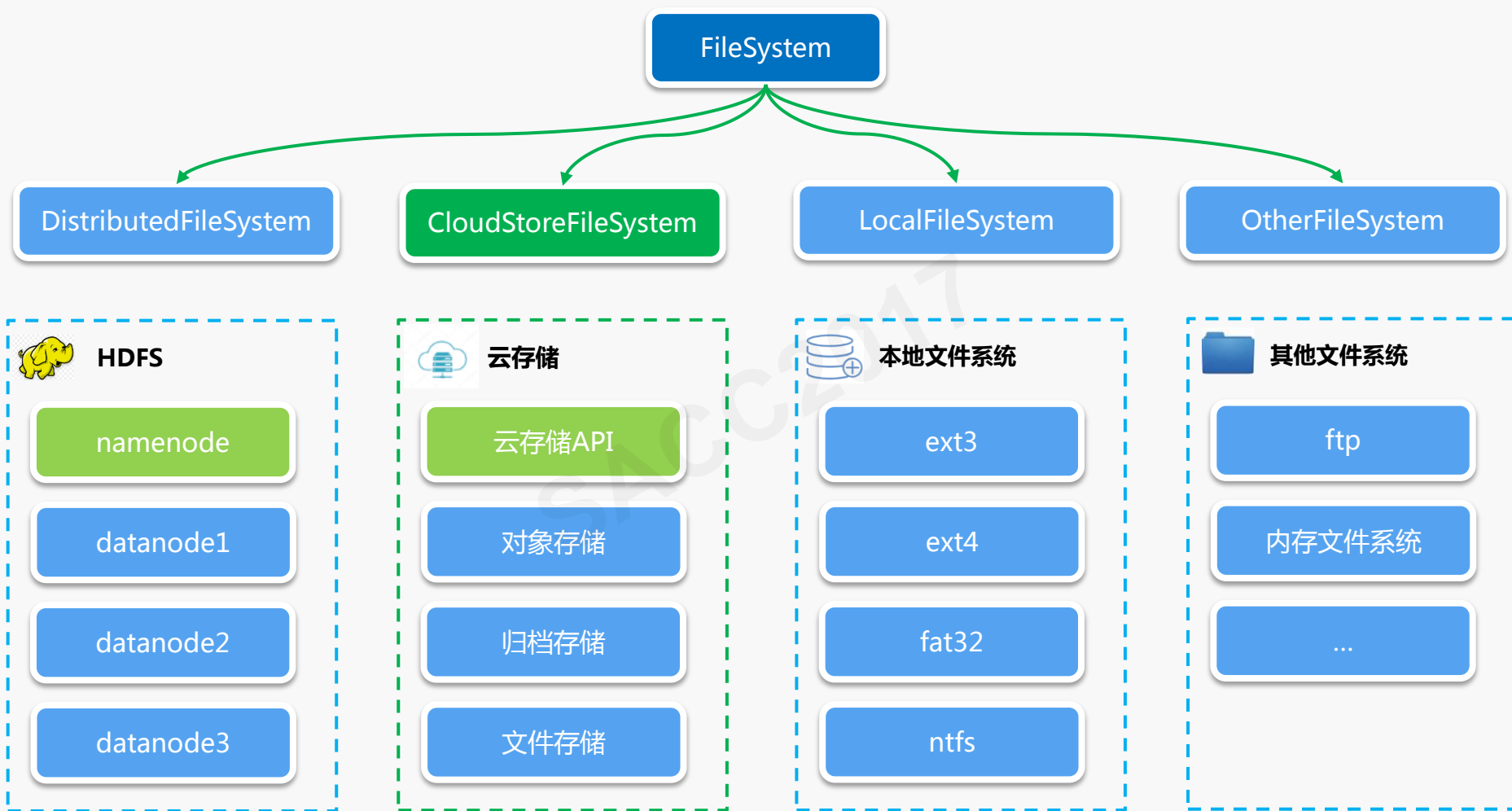
$b = \sum_{n=1}^m (\frac{load1+load2+\dots+loadk}{k})$   $loadn$ 为过去一段时间点某个时间点某个节点的

参数3：集群总体内存使用率  
节点的内存使用率

$c = \sum_{n=1}^m (\frac{usage1+usage2+\dots+usagek}{k})$   $usagen$ 为过去一段时间某个时间点的某个

参数 $a, b, c$  共同决定集群是需要扩容还是需要缩容

# 计算存储分离



# 组件深度优化整合



## 参数优化

HDFS参数优化

YARN参数优化

HIVE参数优化

Hbase参数优化

.....



## 环境整合

Lzo,sanppy等压缩支持

版本兼容性处理

多版本python支持

Spark集群学习库支持

.....



## 社区patch

Hive-14029

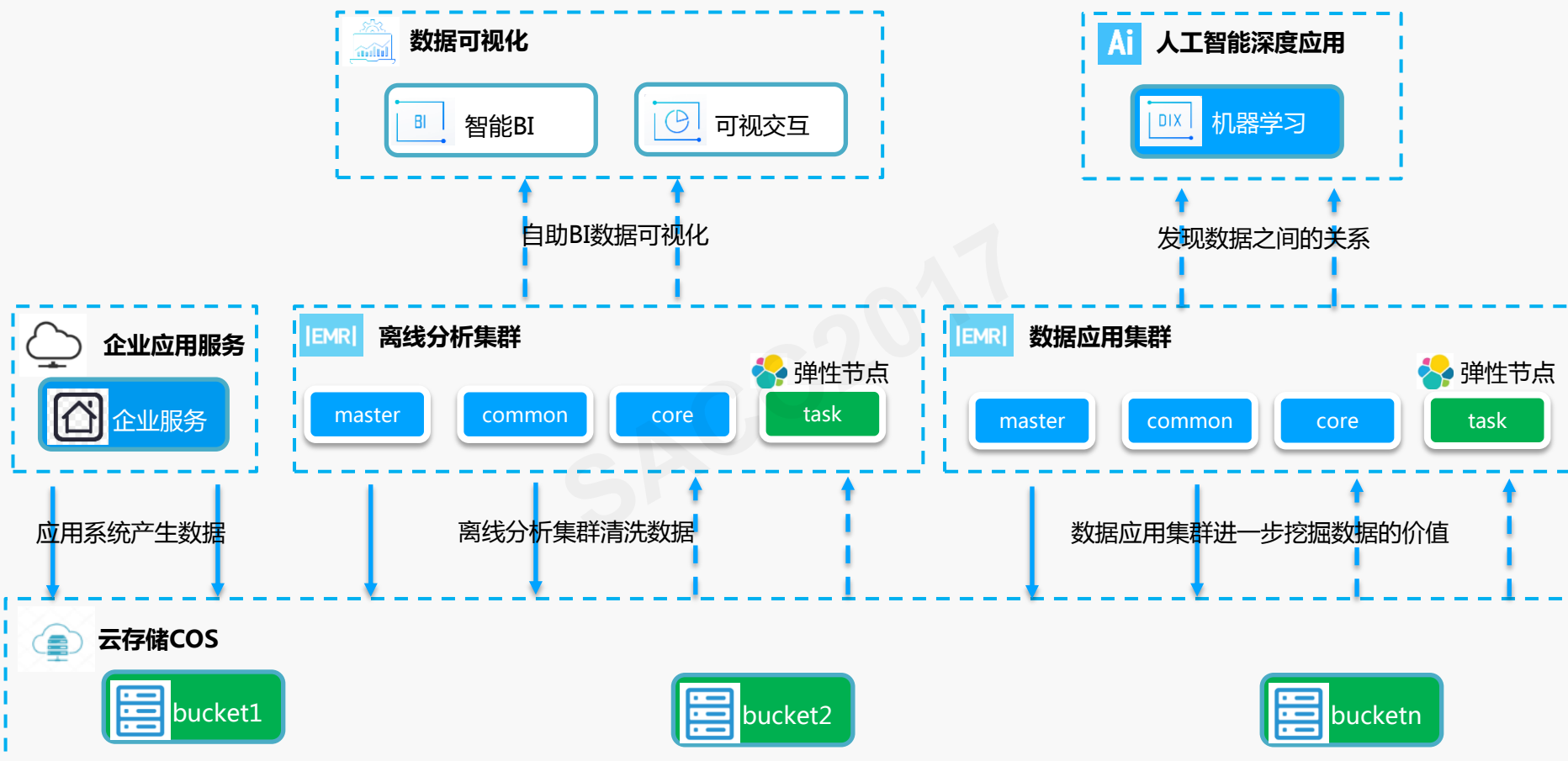
Hive-15355

Hive支持中文注释

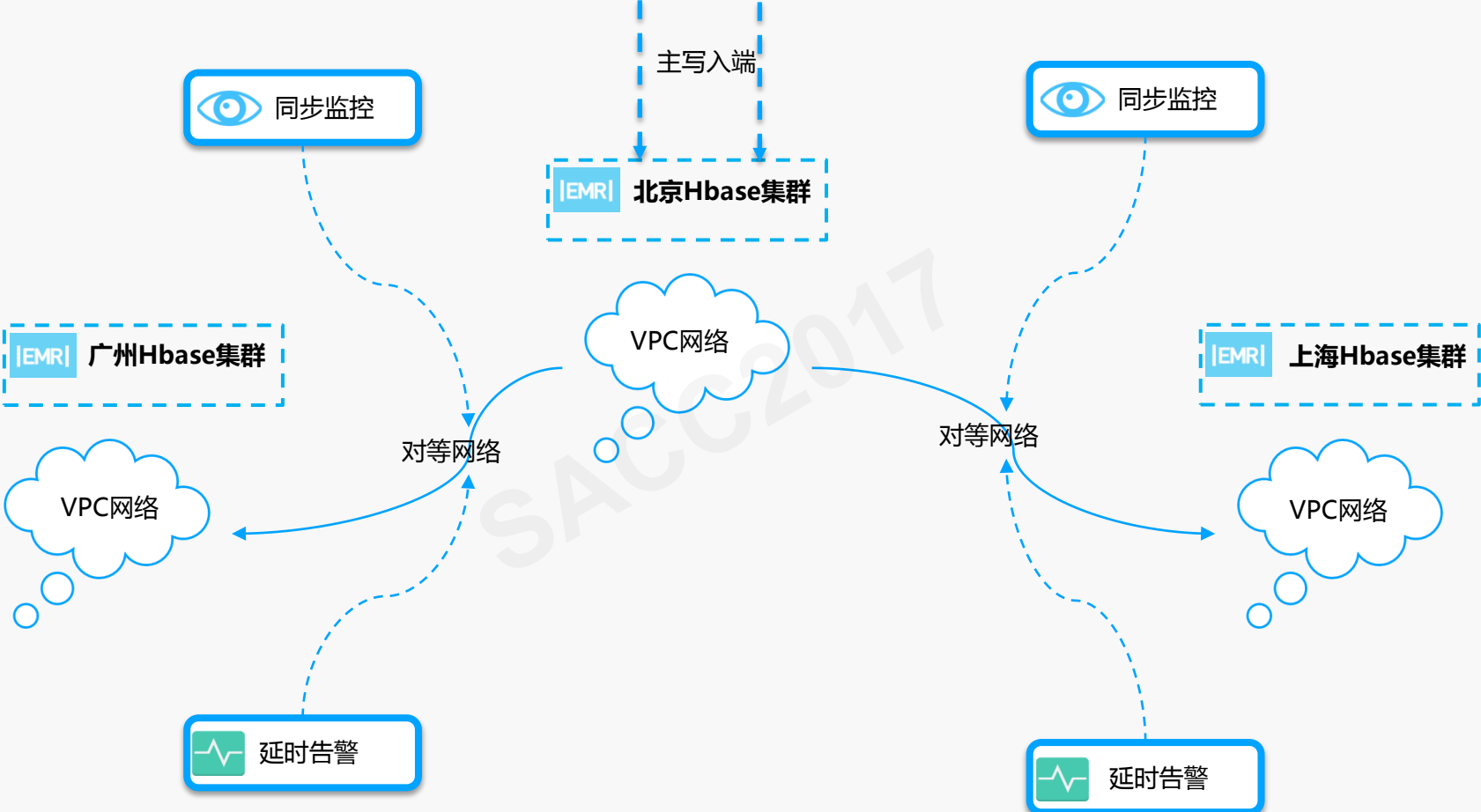
Hbase-16993

.....

# 基于云的计算存储分离应用模式

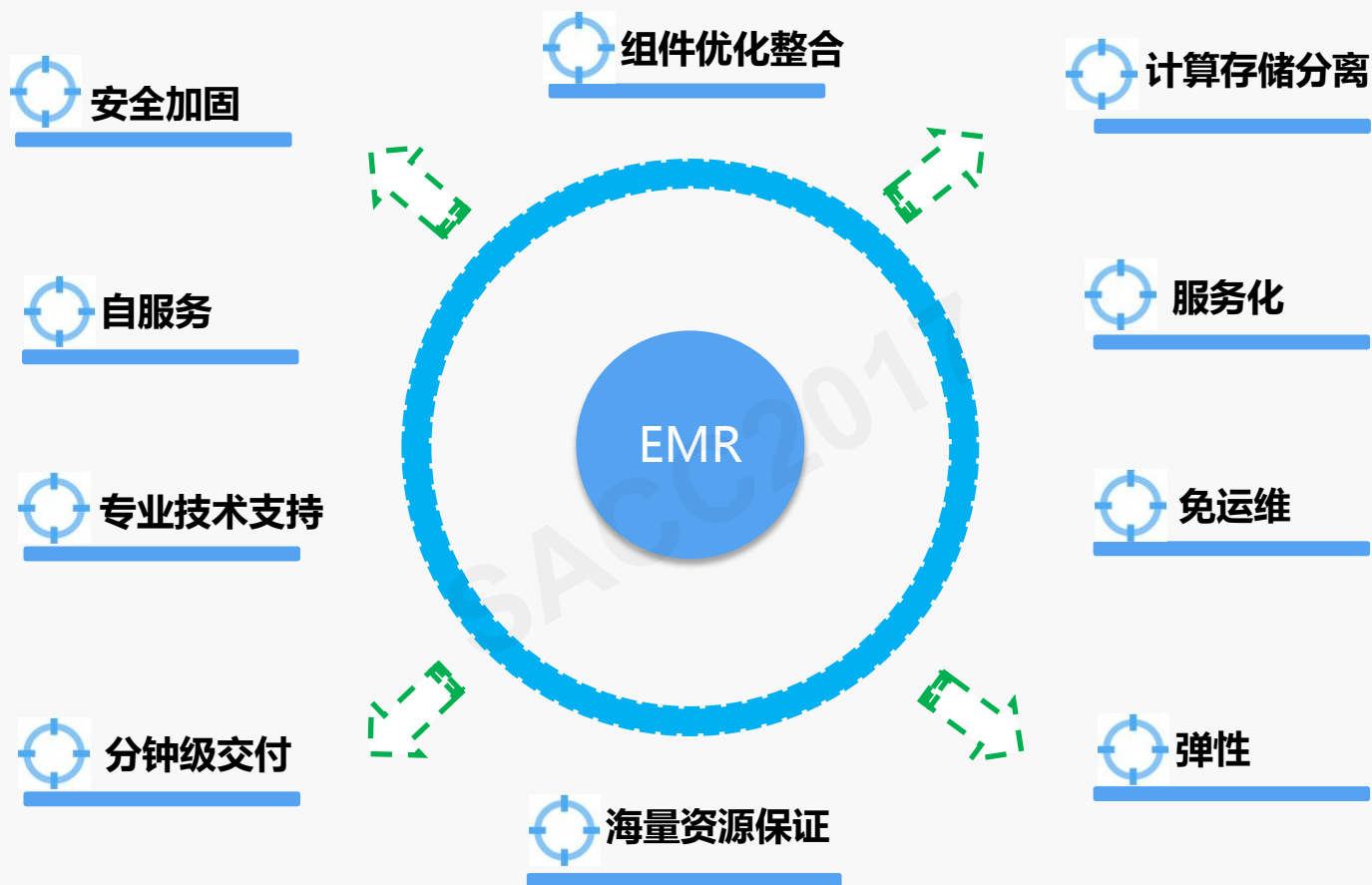


# 基于云虚拟子网的海量数据高可靠应用

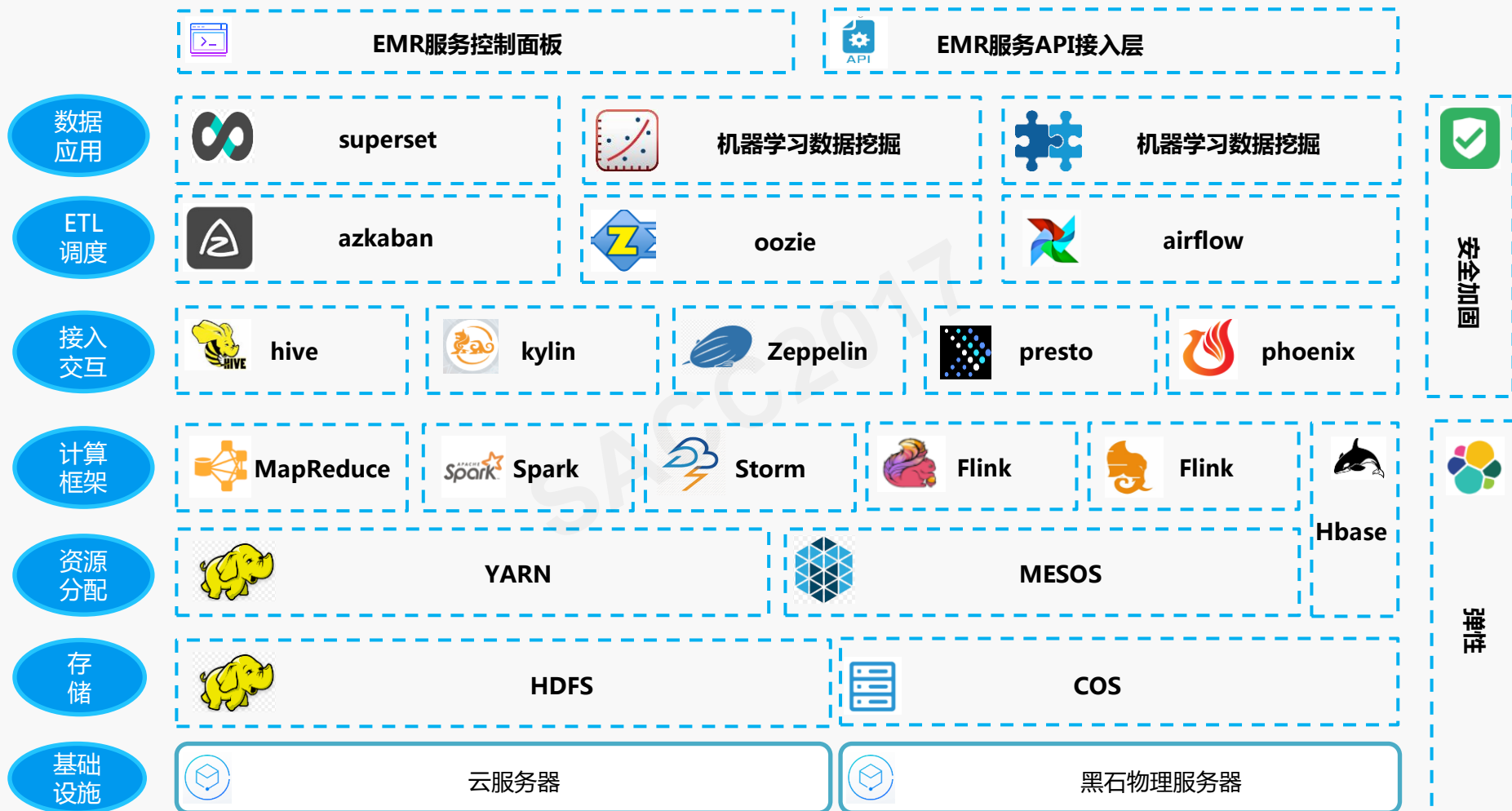




# 腾讯云EMR服务



# 腾讯云EMR服务



THANKS

