

# 游戏数据处理



# 游戏数据处理

## 游戏 数据 处理

↓      ↓      ↓  
范围      对象      行为



Game Data  
Processing



# 目 录

- 1 ▶ 何为游戏数据处理
- 2 ▶ 游戏数据处理意义
- 3 ▶ 数据处理框架
- 4 ▶ 游戏数据处理案例



# 何为游戏数据处理？

---



WHAT



# 「游戏数据处理」所处位置

「根话题」

学科

理工科

科学与技术(科技)

计算机科学与技术

数据(计算机)

数据处理

游戏数据处理

注：参考知乎话题分类：<https://www.zhihu.com/topic/19554449/organize/entire#anchor-children-topic>



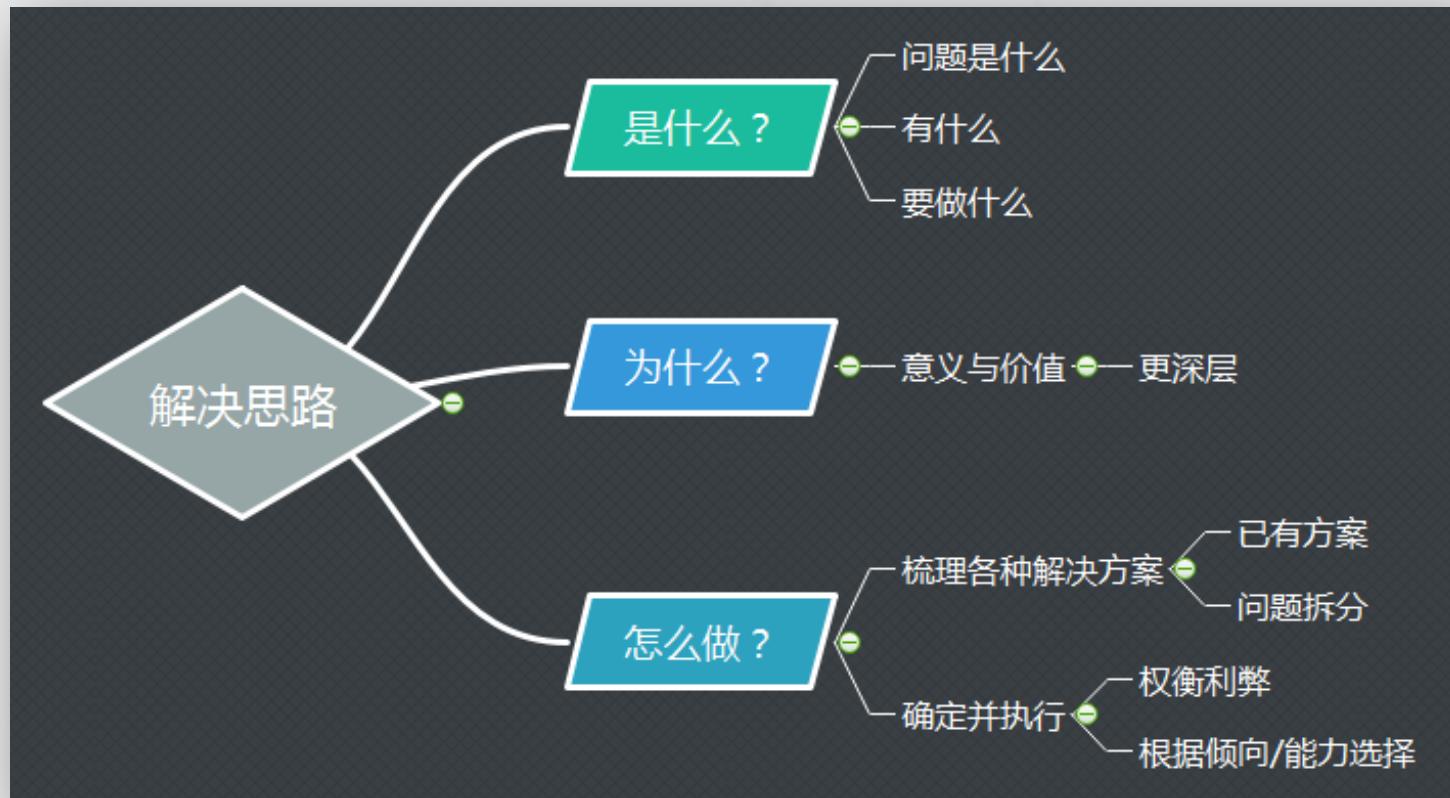
# 生活中的数据处理

**场景一：**老师有个研究本校应届毕业生就业趋势的课题，已有一些原始数据，希望你整理统计完成报告，并按时提交。



# 生活中的数据处理

场景一：本校应届毕业生就业趋势课题报告。



# 生活中的数据处理

场景一：本校应届毕业生就业趋势课题报告。

是什么？	有原始数据，分析历年、各行业就业率情况
为什么？	报告的价值
怎么做？	Excel、代码等方式进行整理统计



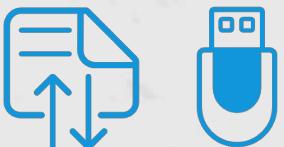
# 生活中的数据处理

场景一：本校应届毕业生就业趋势课题报告。

Step 1

**原始数据**

获取原始数据，  
例如：软件、U  
盘、邮件。



Step 2

**数据存储**

存储原始数据，  
例如：PC。



Step 3

**处理并完成报告**

进行预处理、统  
计与分析，得出  
结论完成报告。



Step 4

**提交报告**

将报告进行提  
交，完成整个过  
程。



# 生活中的数据处理

**场景二：为分析电影数据（例如：各类型评分分布、趋势和提取评论关键词），以豆瓣上数据为例，并做成H5页面。**

## 西虹市首富 (2018)



导演: 闫非 / 彭大魔  
编剧: 闫非 / 彭大魔 / 林炳宝  
主演: 沈腾 / 宋芸桦 / 张一鸣 / 张晨光 / 常远 / 更多...  
类型: 喜剧  
制片国家/地区: 中国大陆  
语言: 汉语普通话  
上映日期: 2018-07-27(中国大陆)  
片长: 118分钟  
又名: Hello Mr. Billionaire



## 一出好戏 (2018)



导演: 黄渤  
编剧: 黄渤 / 张冀 / 郭俊立 / 查慕春 / 崔斯韦 / 邢爱娜 / 黄湛中  
主演: 黄渤 / 舒淇 / 王宝强 / 张艺兴 / 于和伟 / 更多...  
类型: 剧情 / 喜剧  
制片国家/地区: 中国大陆  
语言: 汉语普通话  
上映日期: 2018-08-10(中国大陆)  
片长: 134分钟



# 生活中的数据处理

**场景二：电影数据处理、分析与展示。**

**是什么？** 分析电影各类型评分分布、趋势等

**为什么？** 兴趣、练手

**怎么做？** 数据库管理系统、自然语言处理（NLP）



# 生活中的数据处理

## 场景二：电影数据处理、分析与展示。

Step 1

### 数据获取

通过合作或合规等方式，获取数据。



Step 2

### 数据存储

大容量设备进行存储，搭建数据库系统便于统计分析。

Step 3

### 处理并保存结果

基于系统，进行统计分析，还要NLP技术提取关键词。



Step 4

### 应用开发

开发H5页面，访问结果数据呈现至前端



# 数据&数据处理

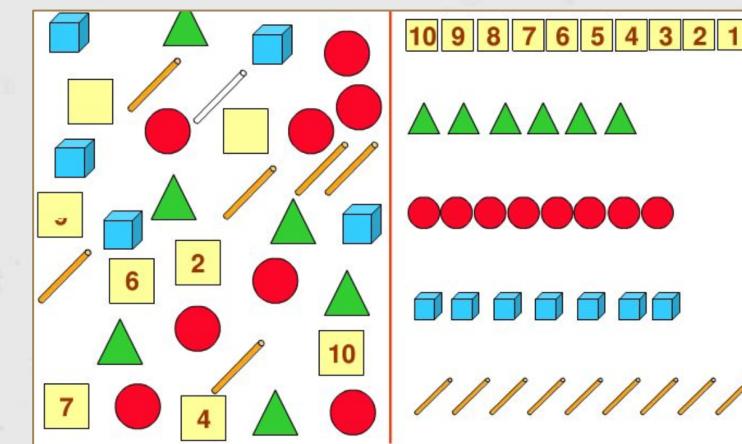
## 数据 (Data) :

未经过处理的原始记录，是对事实、概念或指令的一种表达形式。<sup>[1]</sup>



## 数据处理 (Data Processing) :

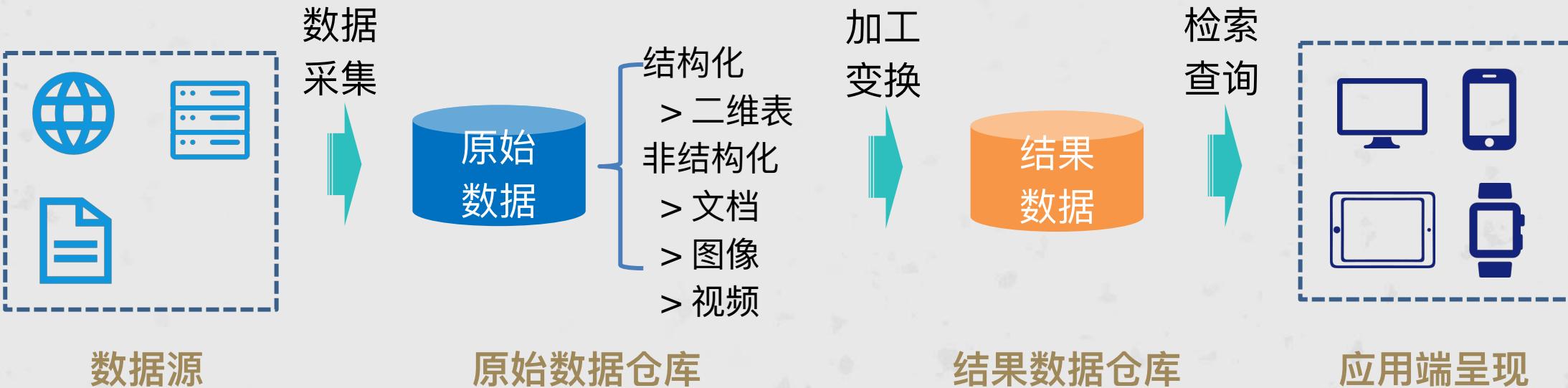
对数据的采集、存储、加工变换、传输和检索，将数据转换为信息的过程。<sup>[2]</sup>



注[1]: <https://zh.wikipedia.org/wiki/%E6%95%B0%E6%8D%AE>

注[2]: [https://en.wikipedia.org/wiki/Data\\_processing](https://en.wikipedia.org/wiki/Data_processing)

# 数据处理流程



# 游戏数据？

《梦幻西游》手游



《第五人格》手游

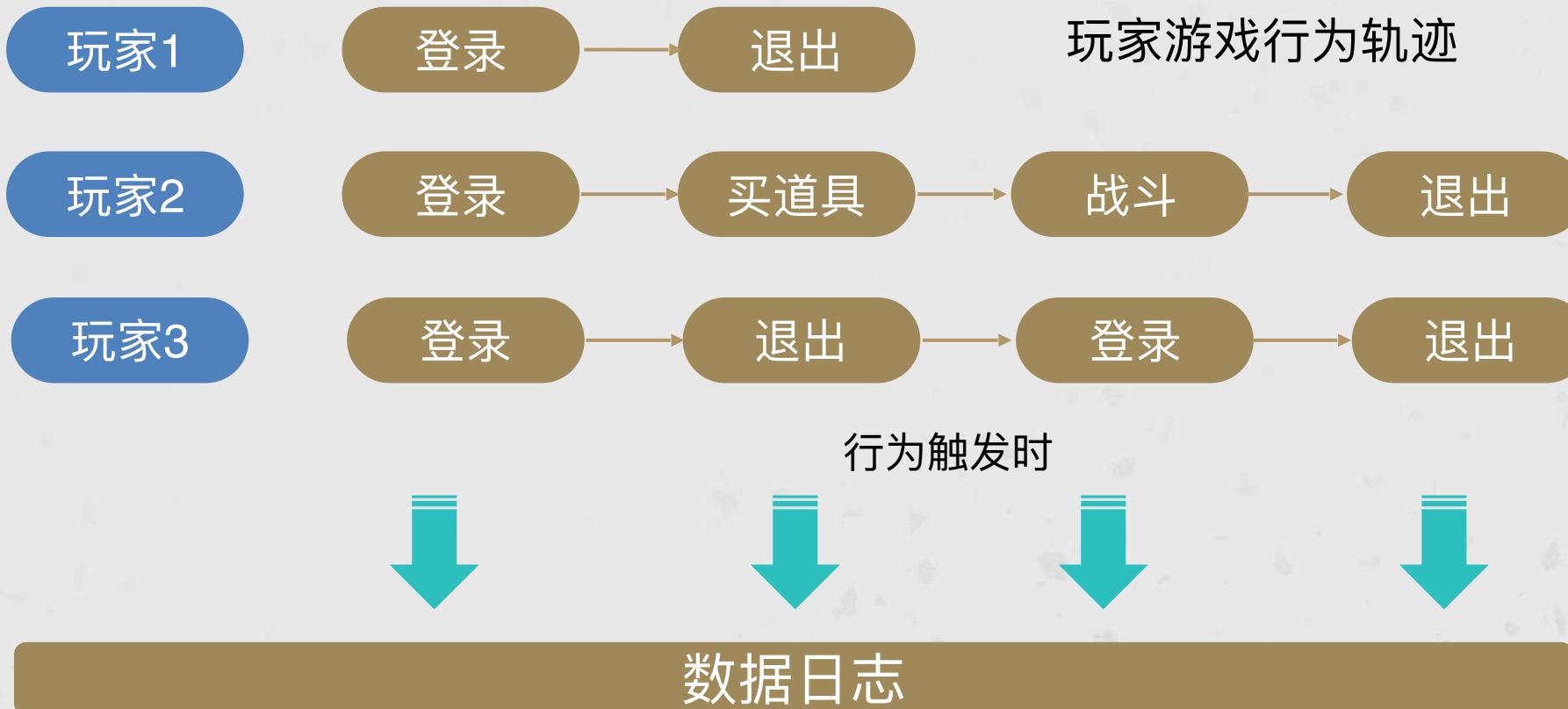


《逆水寒》端游

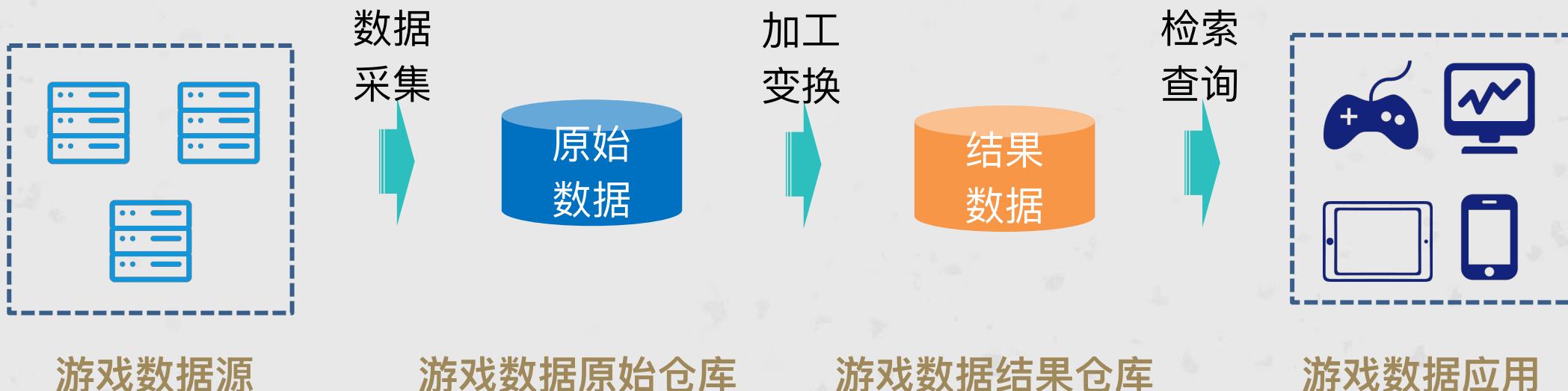


游戏是面向用户的一类产品形态，玩家行为数据是怎样的？

# 游戏数据记录 – 日志



# 游戏数据处理



# 游戏数据处理的意义？

---



WHY



# 游戏数据处理定位

将“躺着”的游戏数据变“活”

的一类工具、手段，  
为游戏上层应用服务。



# 常见数据产品

数据看板

对比走势



# 游戏数据应用

## 应用一：评估一款游戏的运营状态



# 游戏数据应用

## 应用二：评估游戏内的战斗、资源情况



# 游戏数据应用

## 应用三：个性化推荐



# 游戏数据应用

## 应用四：游戏数据分析

月度报告

群体分析

竞品报告

内测评估报告

....



# 游戏数据应用

## 应用五：游戏数据接口



# 游戏数据应用

其他应用？



# 游戏数据应用小结



# 数据处理框架



SOLUTION



# 游戏数据特点



## 数据量超大

记录玩家在游戏中的各种细节信息。



## 数据不统一

数据源不一，各游戏数据内容不一。



## 计算要求高

数据规模大，计算要足够稳定。

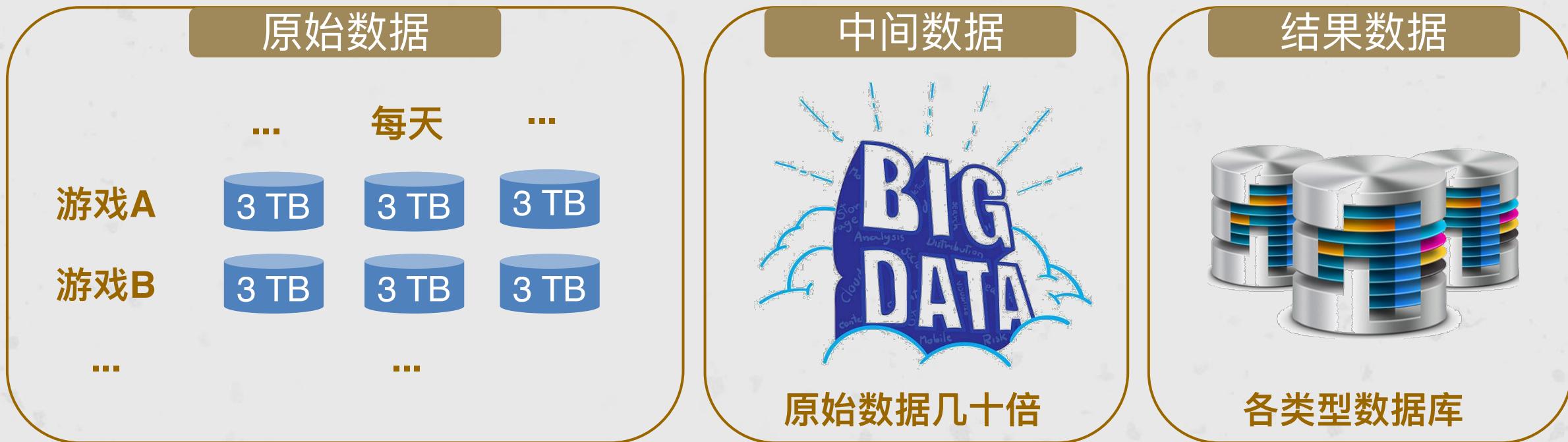


## 应用场景多

能应对离线、实时计算、交互查询等各类应用。



# 游戏数据量



PB级



# 单机访问速度有多快

多年来磁盘存储容量快速增加，其访问速度却未能  
与时俱进（基于成本考虑，大部分为机械硬盘）。

年份	容量	速度	读取全盘时间
90年代	1370MB	4.4MB/s	5分钟
现在	1TB	50MB/s~100MB/s	将近3小时



# 大数据解决方案

众人拾柴火焰高

单机=>多机



# 大数据框架佼佼者 – Hadoop

Doug Cutting



Hadoop



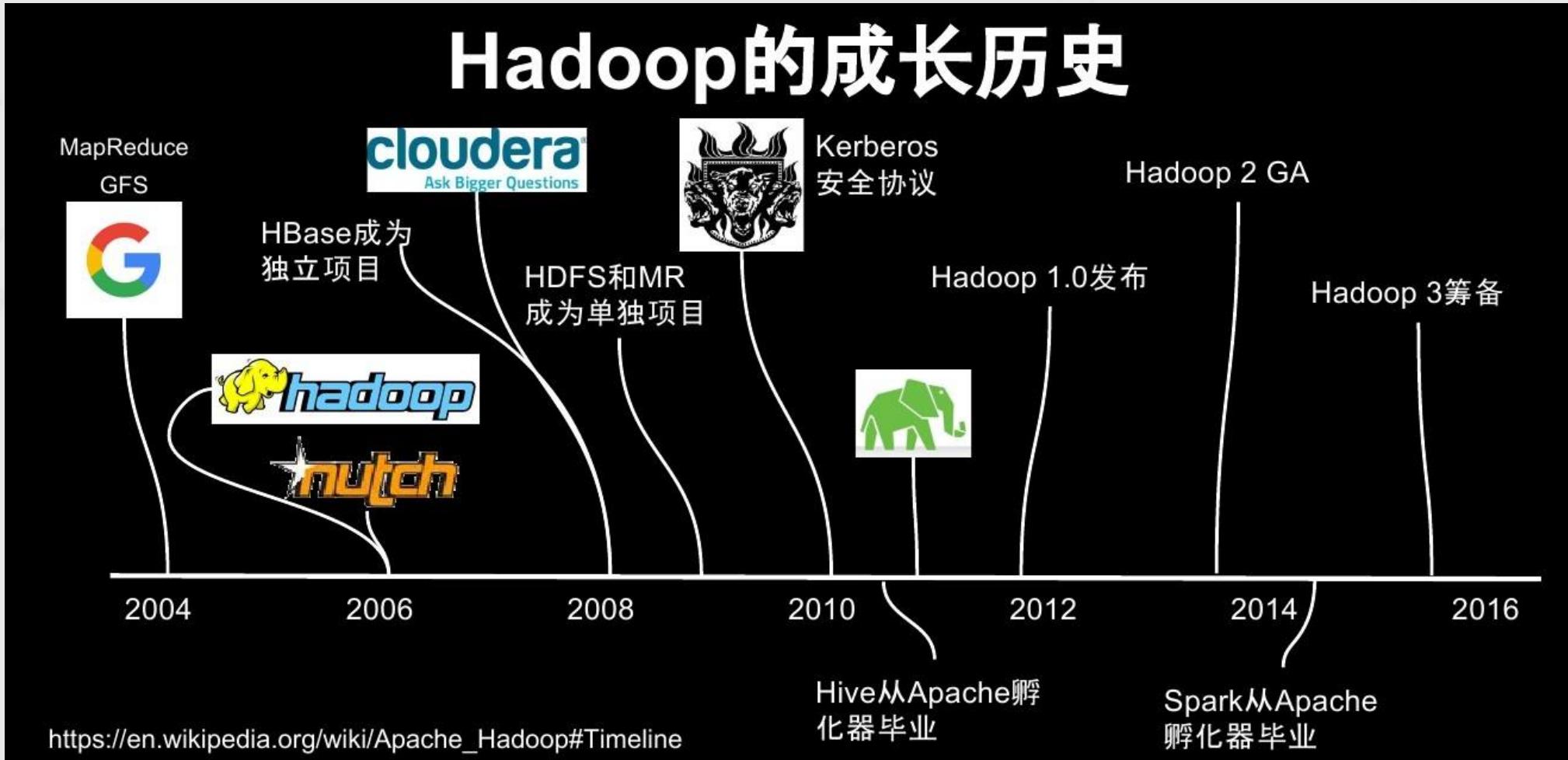
他的儿子有个叫Hadoop的大象玩具

Lucene、Nutch 、  
Hadoop等项目发起人。



# Hadoop成长历史

Hadoop是一个开源的框架，可编写和运行分布式应用来处理大规模数据。



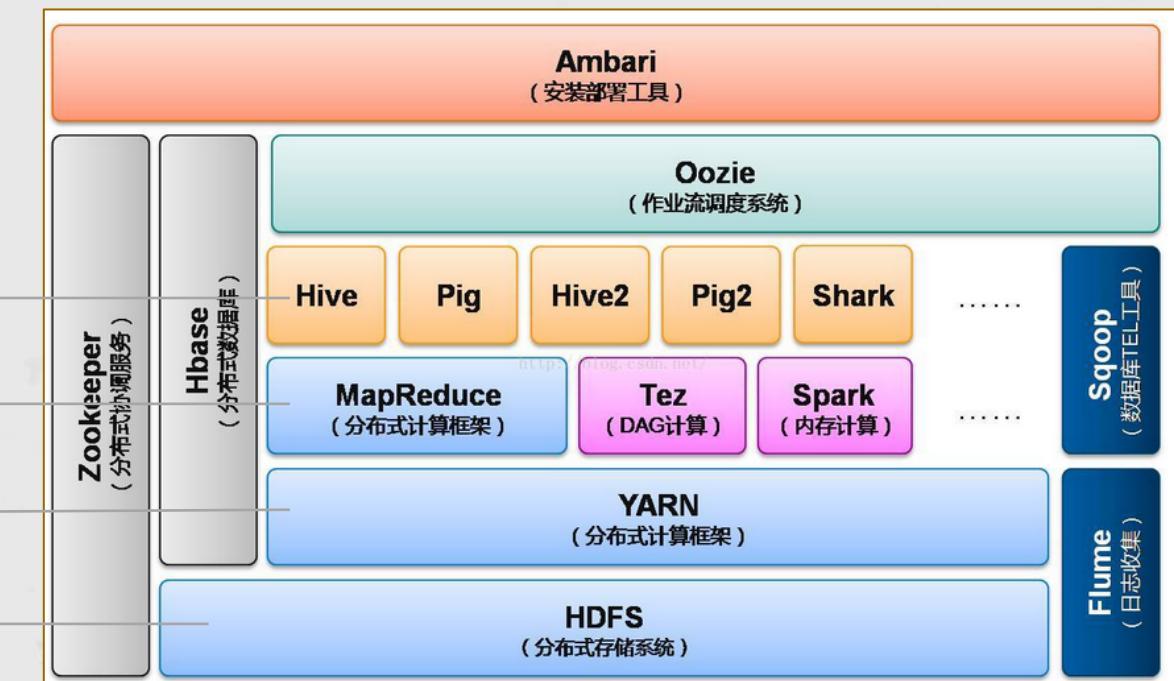
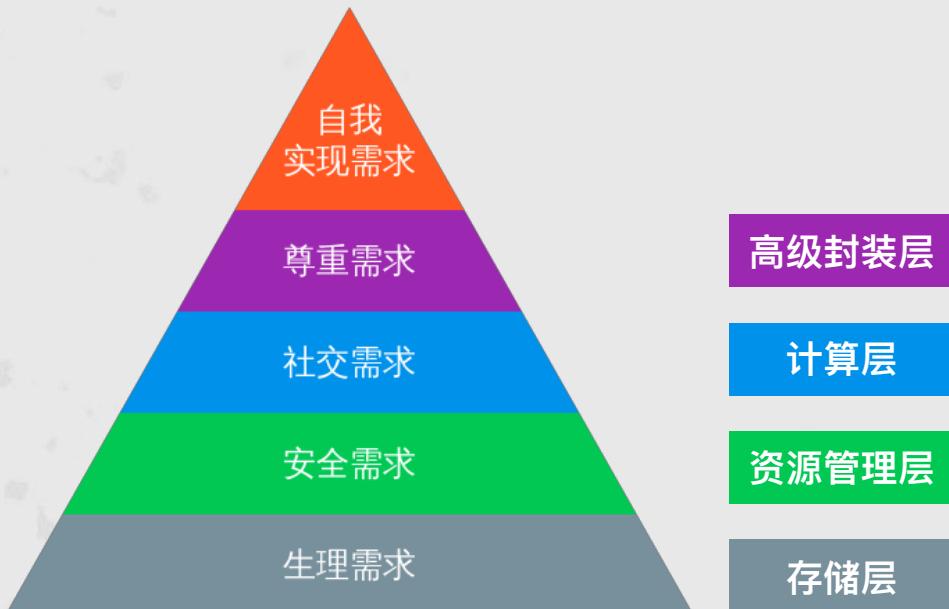
注：<http://www.useit.com.cn/thread-13075-1-1.html>



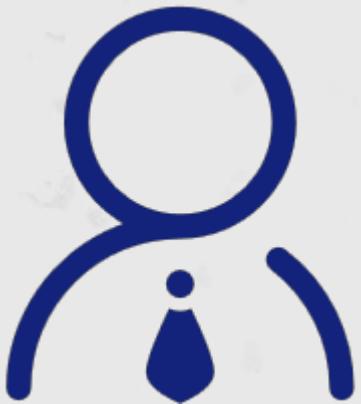
# Hadoop集群简介

**Hadoop**三大核心概念：**HDFS**、**YARN** 和 **MapReduce**（以下简称**MR**）。

逐渐演变为现在的四层架构：



# Hadoop在手



产品经理大大

游戏即将上线，  
对于运营指标计  
算有什么方案？



方案我有：

- 先解决存储问题-**HDFS**；
- 再解决资源管理-**YARN**；
- 接着分布式计算-**MR**；
- 最后更高效开发-**Hive**。



# 解决最基本的”生理需求“

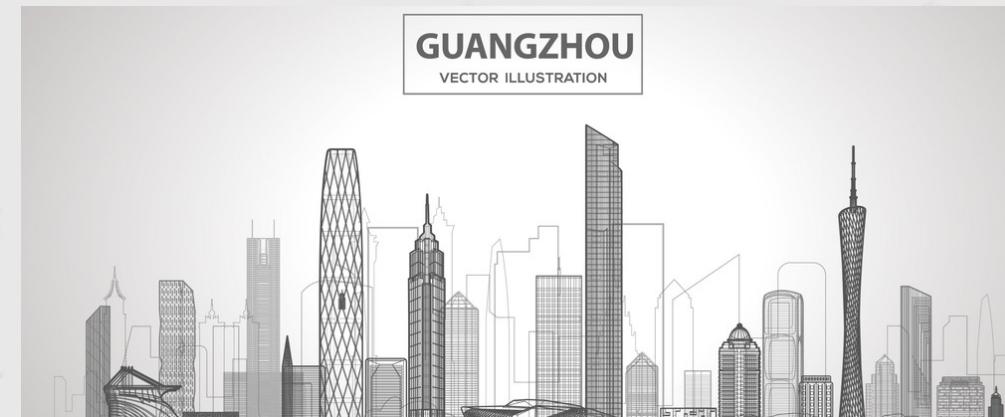
## ◆ HDFS

**Hadoop Distributed File System,**  
是Hadoop项目的核心子项目，是分布式计算的存储  
管理基础。



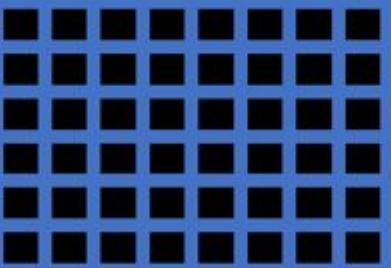
# HDFS作用?

存储海量数据



# HDFS处理文件的思路

文件转换成数据块



数据块分布式处理



# HDFS优缺点

优势

- 高容错性，可自动恢复；
- 适合大数据处理，TB、甚至PB级数据；
- 可构件在廉价机器上。



优缺点



不足

- 不适用于实时查询，延迟较高；
- 不适合很多小文件存取；
- 并发写入、文件随机修改问题。



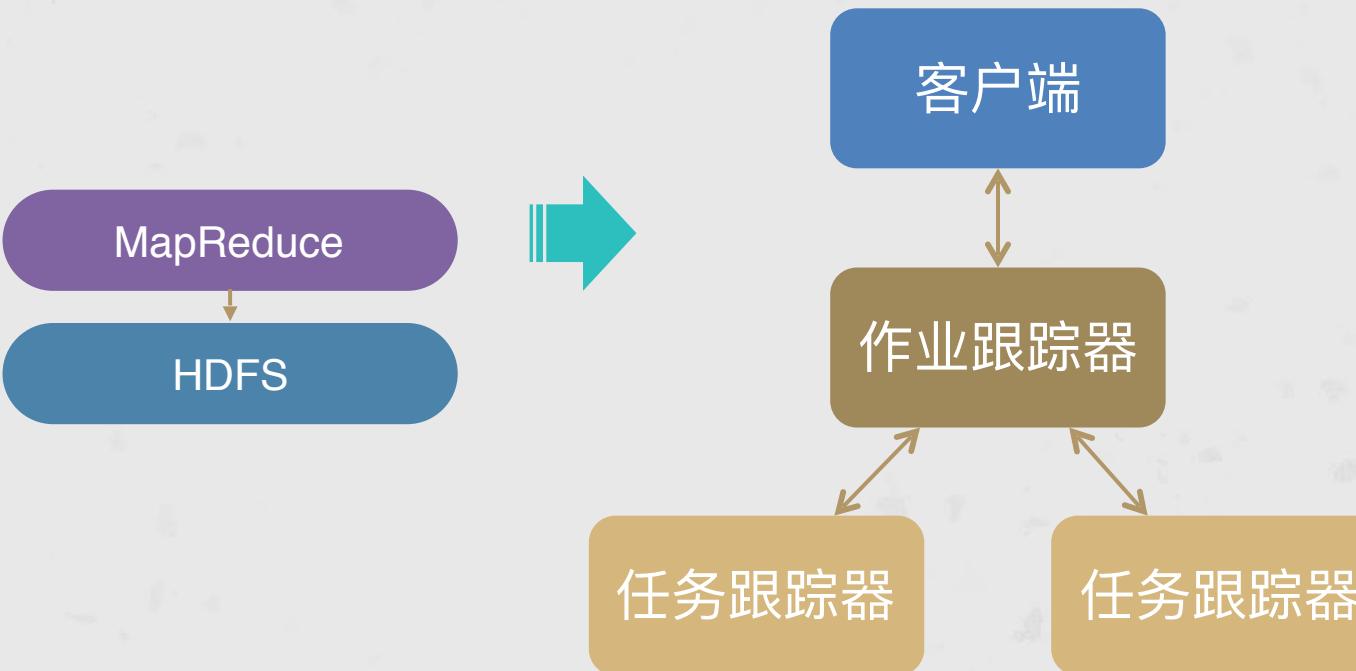
# 放心地解决“安全需求”

## ◆ YARN

**Yet Another Resource Negotiator,**  
是Hadoop资源管理器，为集群在利用率、资源统一  
管理和数据共享等方面带来了巨大好处。



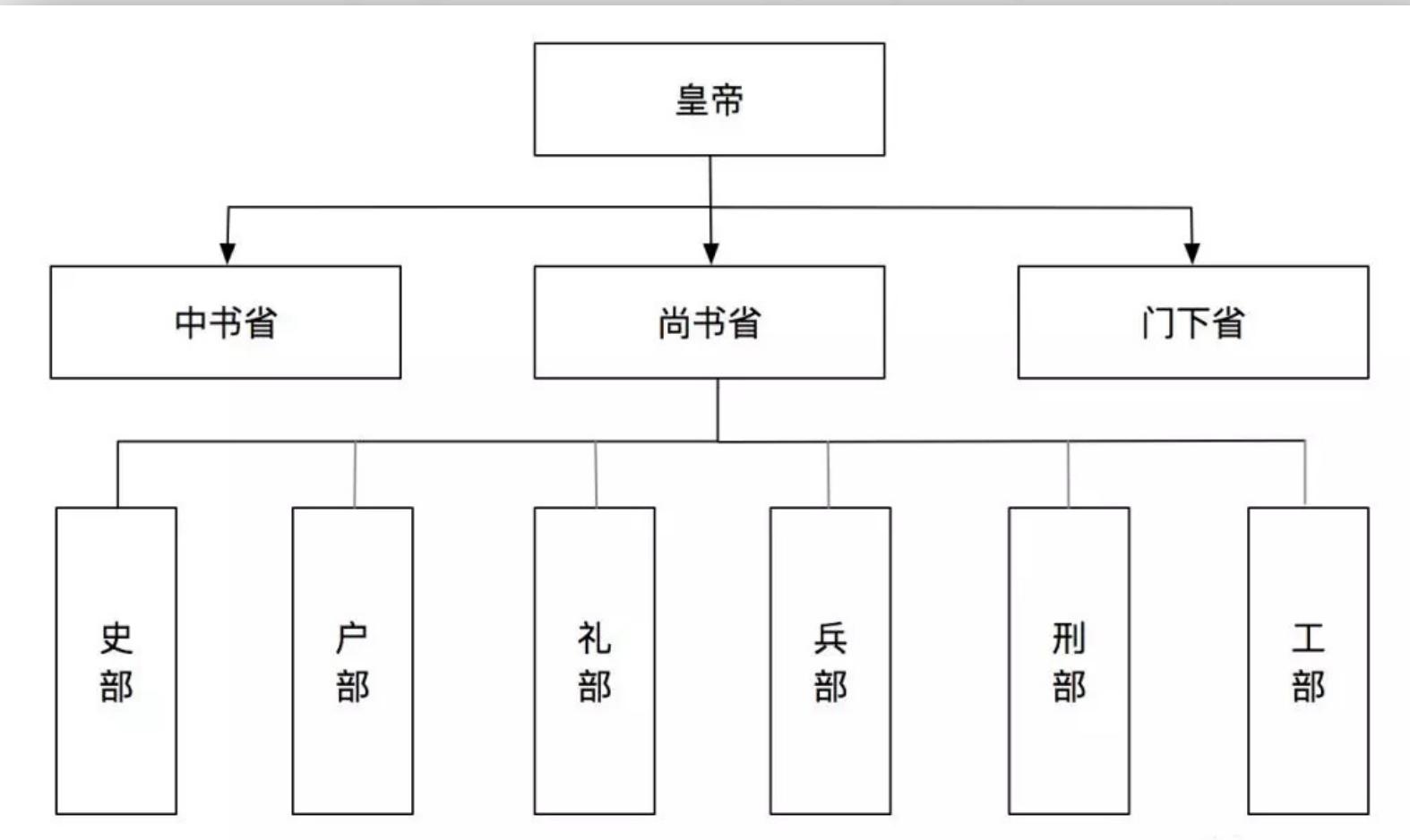
# 没有YARN的日子



事无巨细  
活活累死



# 如何管理



# Hadoop管理页面

← → C ⓘ 不安全 | gdc-nn01-testing.i.nease.net:8088/cluster/cluster



## About the Cluster

Cluster Metrics				应用数量				内存资源				节点信息	
Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	
13374	0	3	13371	7	15 GB	156 GB	0 B	7	78	0	<u>13</u>	0	

Cluster ID: 1522241135238 → RM状态  
 ResourceManager state: STARTED  
 ResourceManager HA state: active  
 ResourceManager RMStateStore: org.apache.hadoop.yarn.server.resourcemanager.recovery.ZKRMStateStore  
 ResourceManager started on: 星期三 三月 28 20:45:35 +0800 2018  
 版本 → ResourceManager version: 2.6.0-cdh5.6.0 from Unknown by whke source checksum a1cb02ff3db5d3b3deb416e22703cf on 2016-03-28  
 Hadoop version: 2.6.0-cdh5.6.0 from Unknown by whke source checksum c2393bf01dc7b2d15446ff11d272441 on 2016-03-28

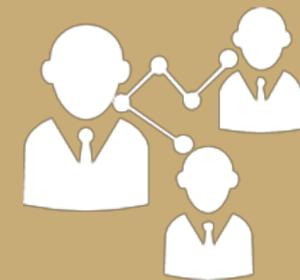
- 总览（上图以测试集群为例）：[http://\\${host}:8088/cluster/cluster](http://${host}:8088/cluster/cluster)。
- 各节点信息：[http://\\${host}:8088/cluster/nodes](http://${host}:8088/cluster/nodes)。



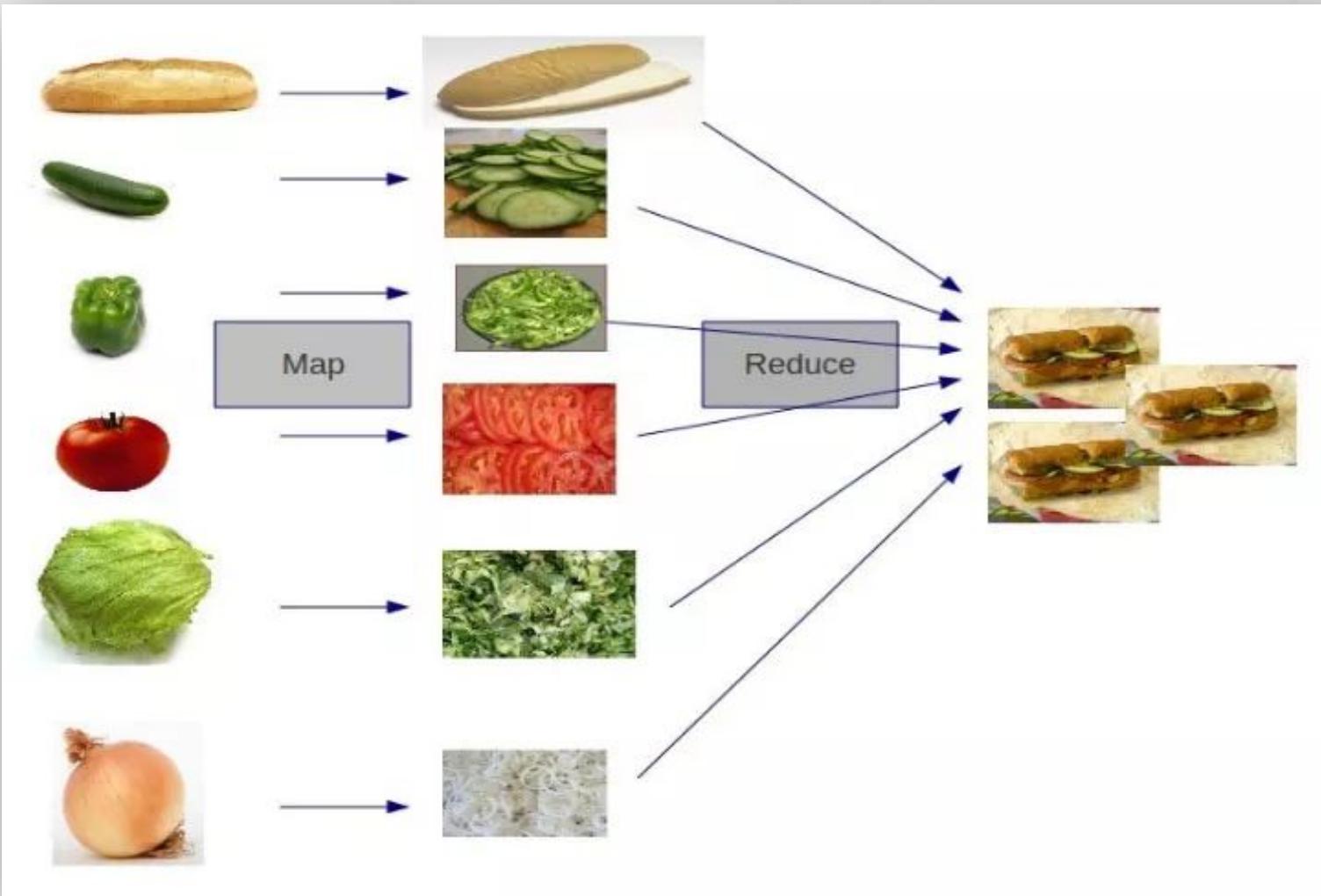
# 数据结合解决“社交需求”

## ◆ MapReduce(MR)

MapReduce用于大规模数据集的并行运算，分为两个阶段：Map和Reduce。  
极大方便将程序运行在分布式系统上。



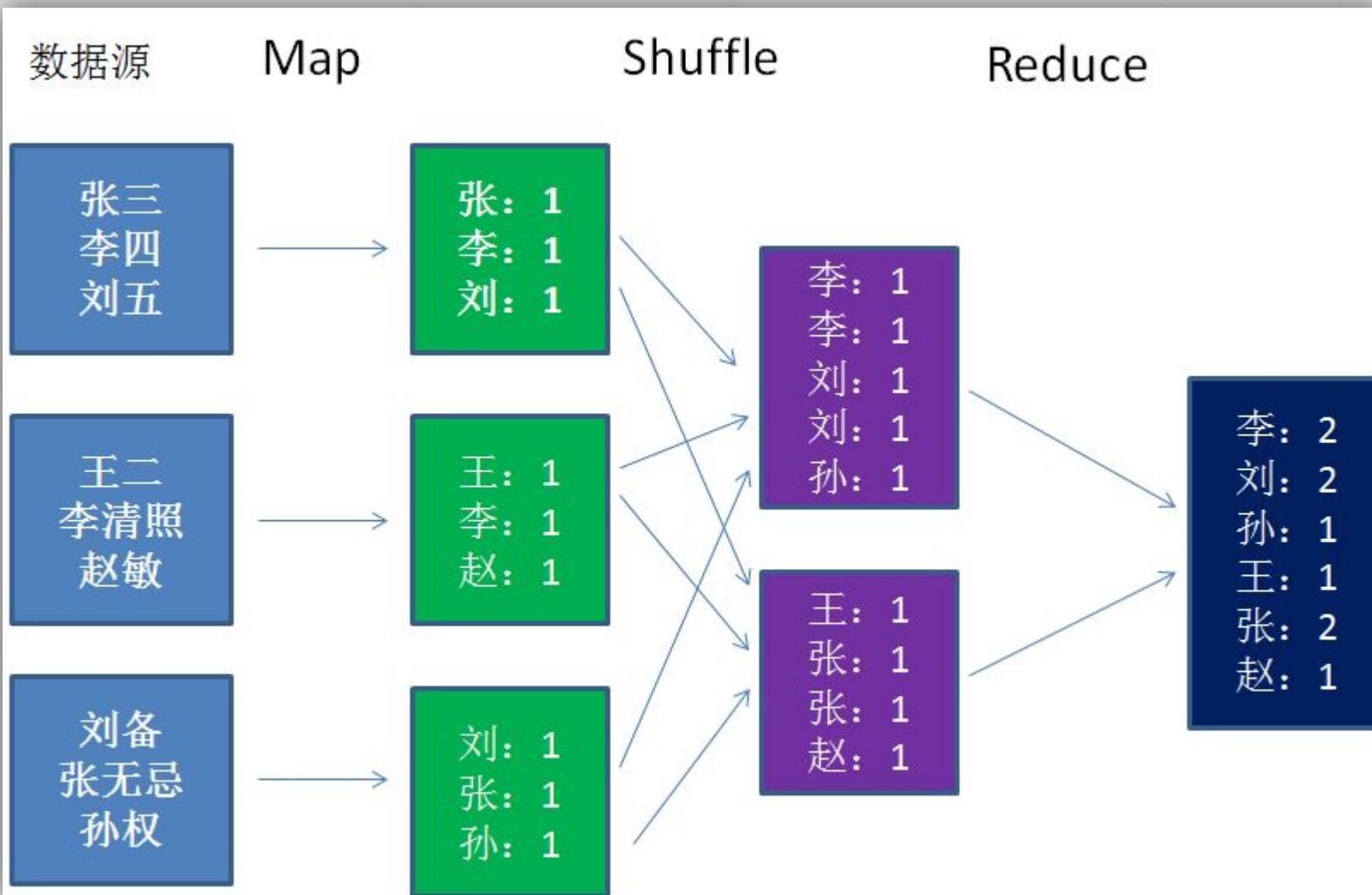
# MapReduce是什么



注: <https://blog.csdn.net/bjweimengshu/article/details/79295013>



# MapReduce举例 – 全国所有姓氏的人数



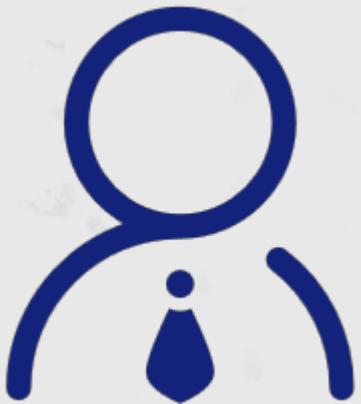
# 超高效率解决“尊重需求”

## ◆ Hive

Hive将结构化数据映射为数据库表，  
可方便使用SQL来完成海量数据的统计和分析。



# MR实现计算需求



产品经理大大

给我算下，A游戏七夕那天男性玩家的数量和平均在线时长。



- Map怎么实现？
- Reduce怎么实现？
- 数据怎么读取写入？
- 代码编译、作业提交
- ... ...

# Hive实现计算需求



产品经理大大

给我算下，A游戏七夕那天男性玩家的数量和平均在线时长。



- A游戏: game='A'
- 七夕: dt='七夕'
- 男性: gender='M'
- 数量: count(1)
- 平均在线: avg(online)
- 一句Hive SQL搞定！



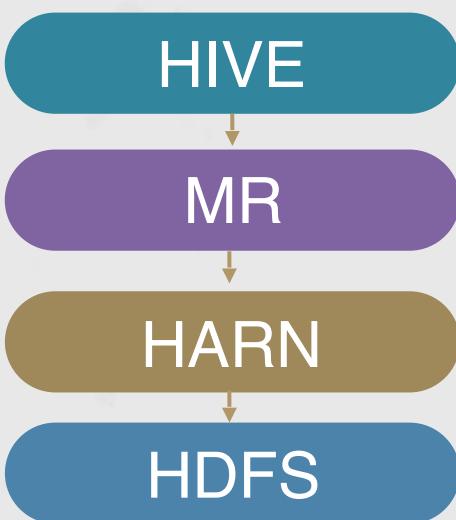
# 数据库语言 – 沟通的桥梁

- **SQL** (Structured Query Language) , 结构化查询语言。
- 可简单理解为：对数据的增删查改。

注[1]: <http://sql.yehyeh.net/content/Syntax.php>



# Hive的定位及原理



Hive实质就是将 **Hive SQL** 语句转换为  
**MapReduce** 任务运行。

底层为：元数据+数据文件（HDFS）



# 都直接用Hive就好？

MR



Hive



实现难易

相对复杂

简单

可处理数据

结构化、非结构化文本

结构化

表达能力

相对比较强大

部分场景无法表达

效率

设计得当可调优

自动生成优化不够

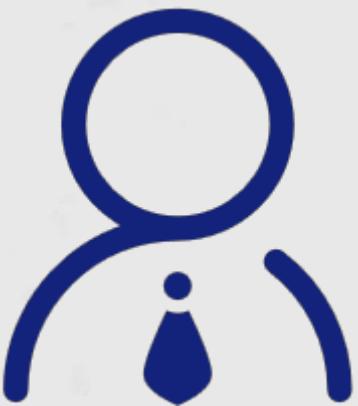


# 游戏数据处理案例

▶ HOW



# 指标统计



BOSS

游戏日志已经记录在各台游戏服务器了，每天的总日志量大概**1TB**。  
统计今年7月各服务器上角色的登录数、消费情况以及各玩法参与情况。



# 分析问题

是什么?

有游戏日志，进行游戏指标统计

为什么?

洞察玩家状态

怎么做?

少=>单机，大=>大数据处理框架



# 游戏数据处理思路

问题：针对原始游戏日志的数据统计。

Step 1

**游戏日志收集**  
原始数据在服  
务器，需要”  
捞”过来。



Step 2

**数据存储**  
搭建集群，用分  
布式文件系统存  
储。



Step 3

**数据处理与统计**  
利用集群的计算  
引擎进行预处理  
和统计。

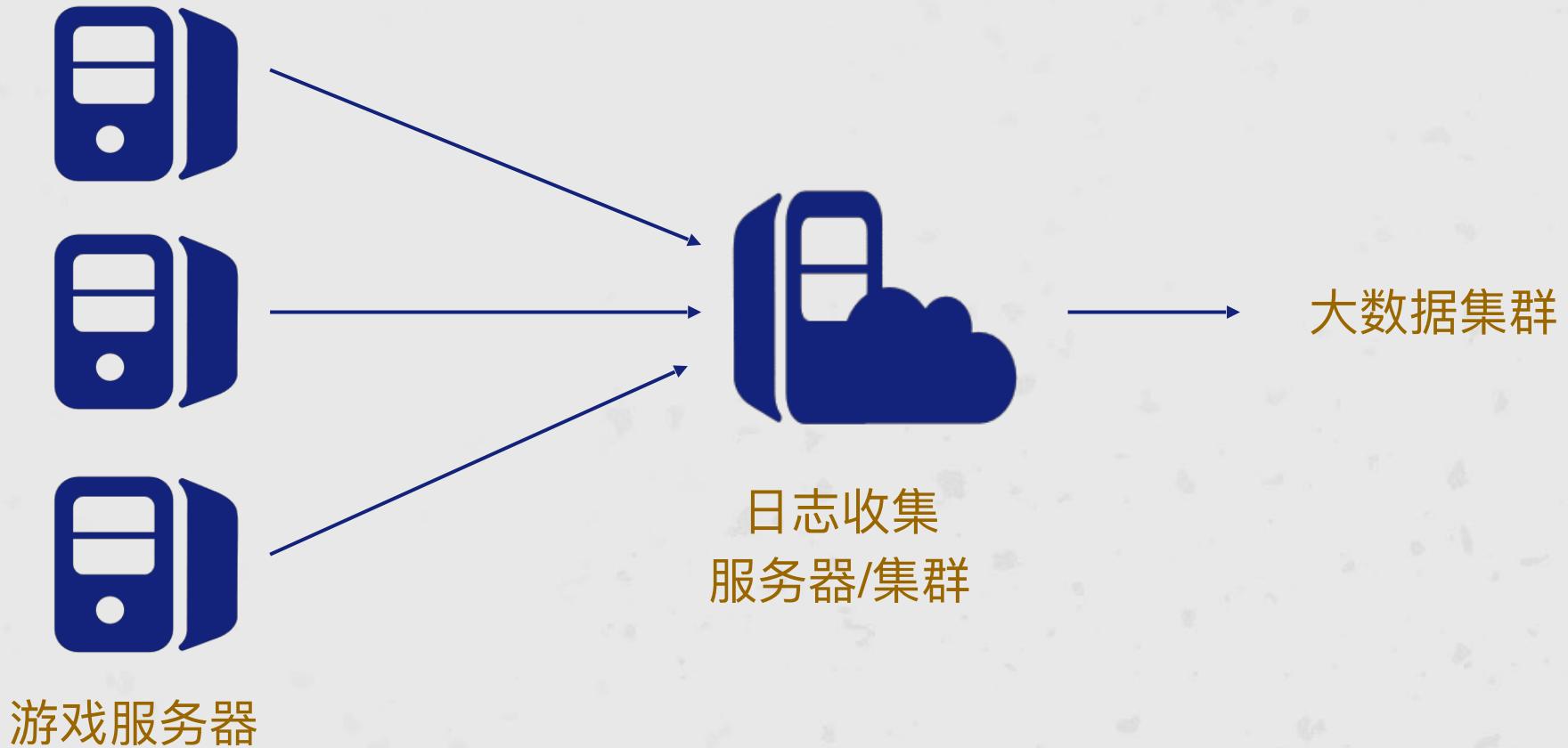


Step 4

**数据展示**  
数据展示并提交



# 游戏日志收集

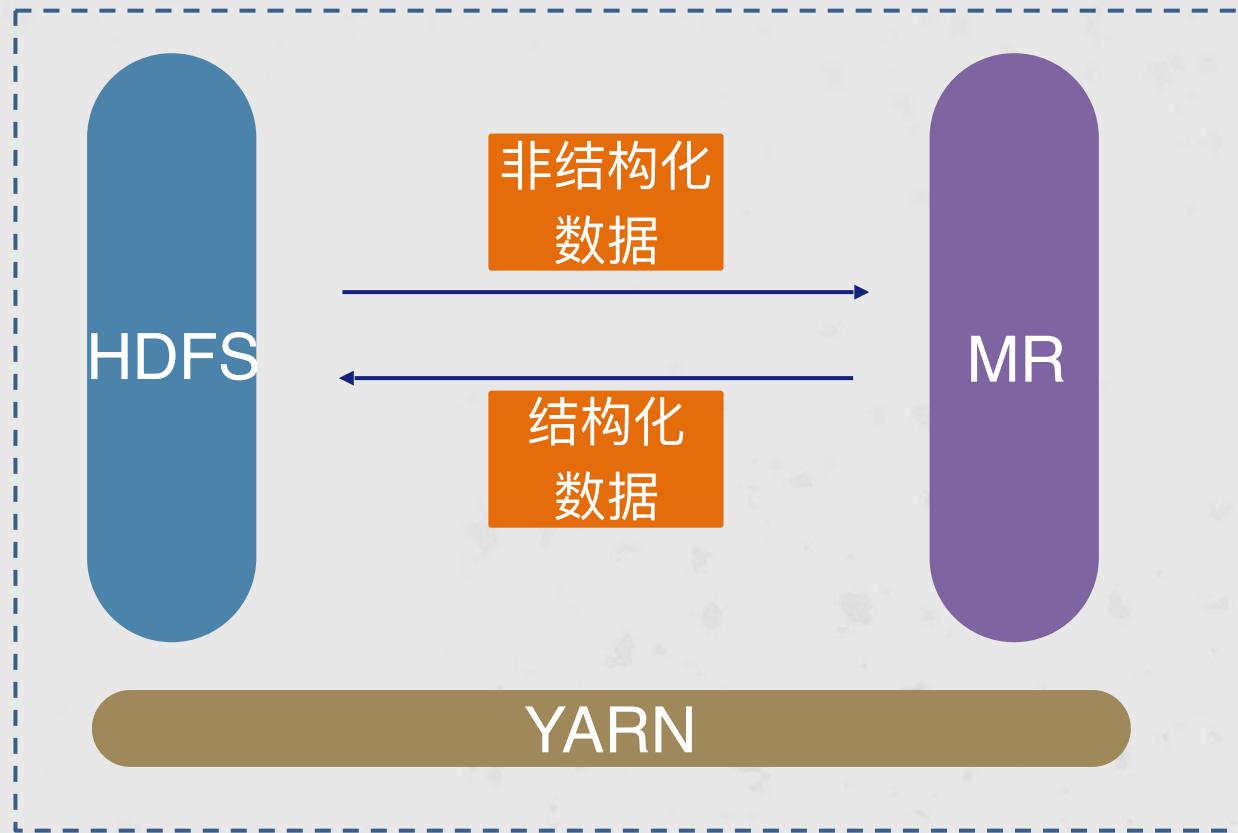


# 数据存储



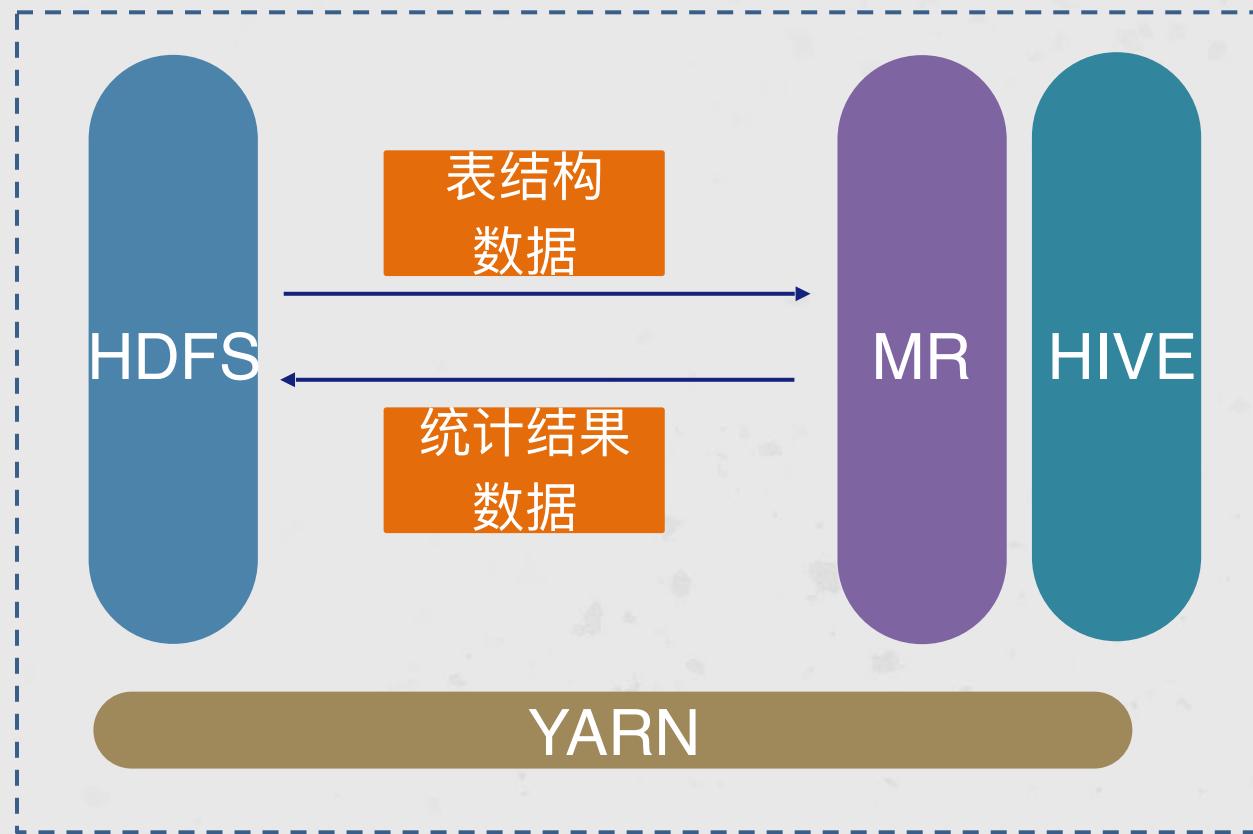
# 数据预处理

## Hadoop集群



# 数据统计

## Hadoop集群



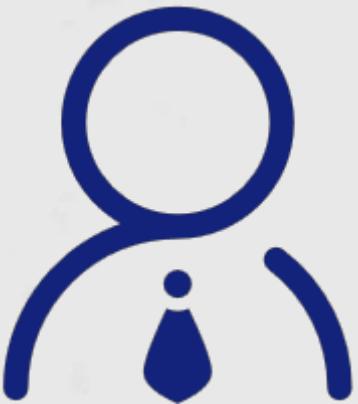
# 数据展示

## 结果示例

服务器	角色登录数	总付费	...
501	XXXXXX	XXXXXXXX	
502	-	-	
503	-	-	
504	-	-	
505	-	-	
...	...	...	



# 实时计算



BOSS

很赞，数据很有价值！当天的角色登录数很重要，希望能实时展示，这个不难做到吧？



# 再次分析

是什么?

有游戏日志，游戏指标实时计算

为什么?

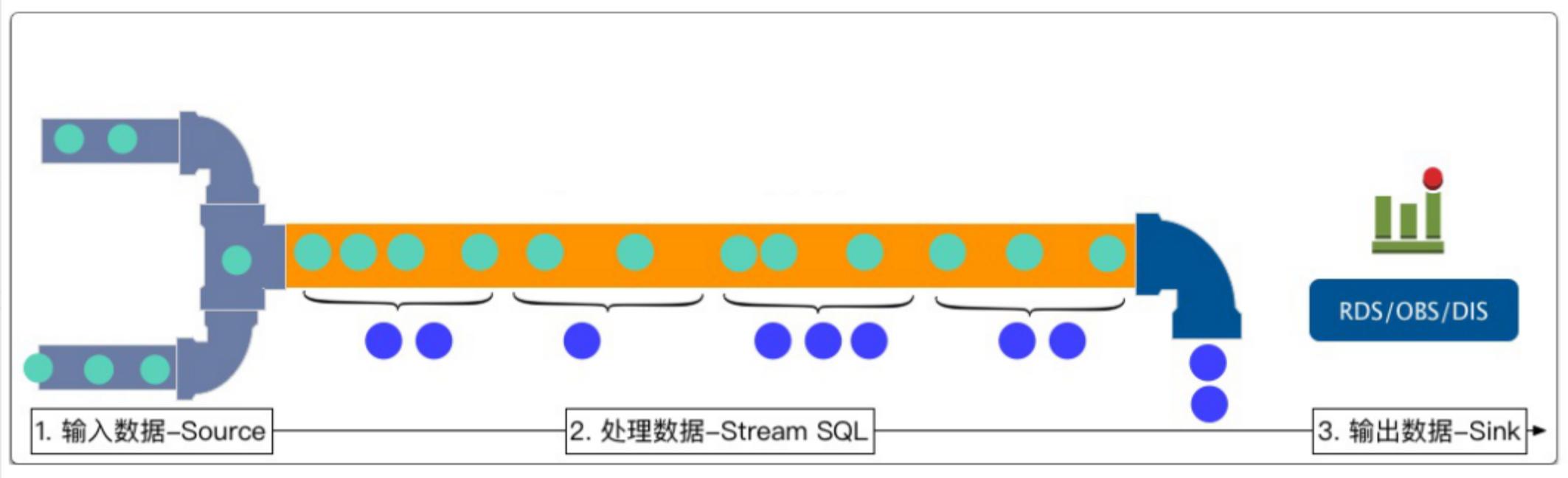
实时监控，便于决策

怎么做?

HDFS本身就是高延迟，需要从日志  
开始实时，采用实时计算框架



# 实时计算思路



- 流式：逐条处理
- 小批量：小数据块处理



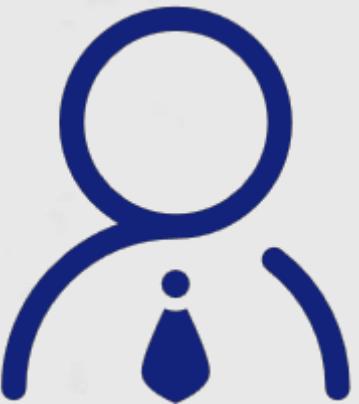
# 游戏日志收集



日志收集  
服务器/集群



# 更多应用



BOSS

利用行为数据进行商品推荐 ...  
实时监控作弊行为 ...  
玩家用户画像 ...

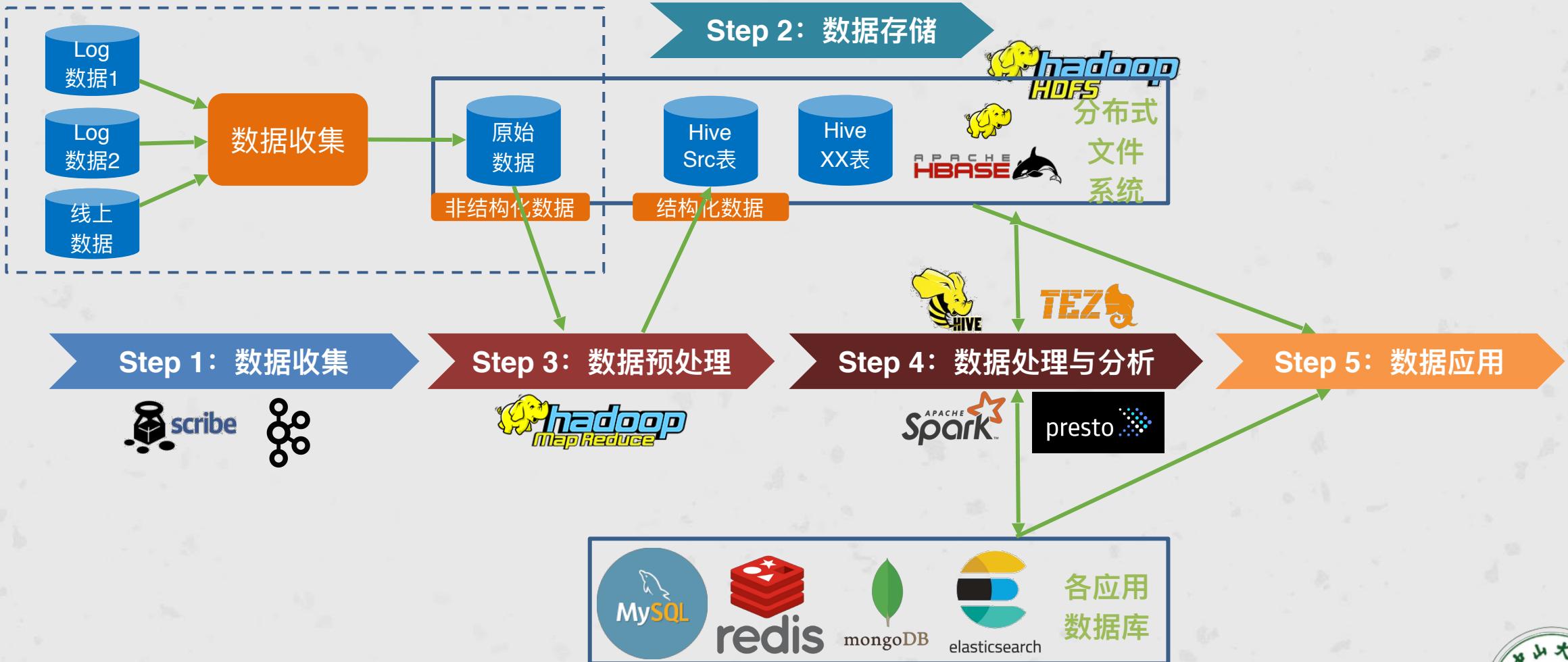


# 数据集群架构

原生Hadoop集群无法满足很多业务需求，因此需要融合更多组件。



# 游戏数据处理主要技术概览



游戏数据处理的总体流程





## Q & A

►感谢大家的观看和参与

