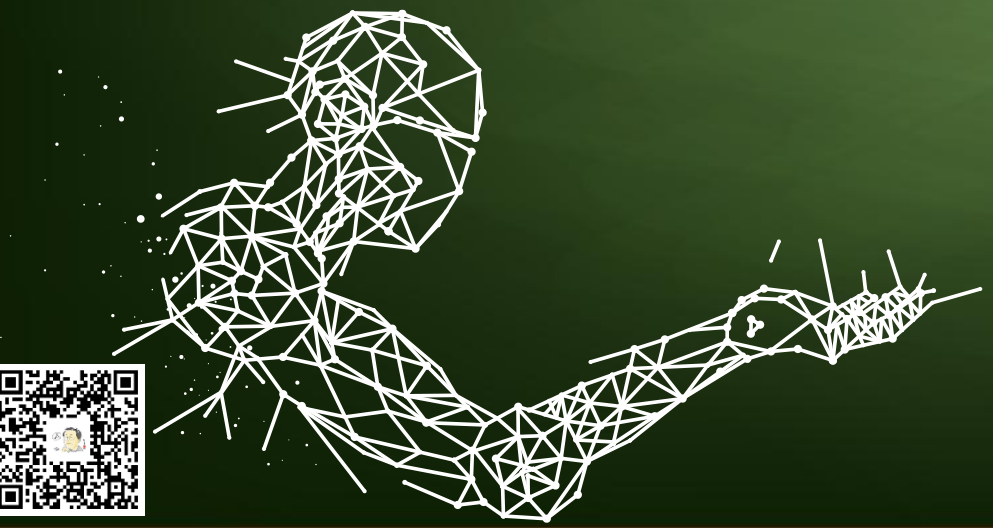
01

第1章 大数据技术概述

授课教师：[李阳源]

15138617705@163.com

主页：https://sites.google.com/view/dalguri



第1章 大数据技术概述


BIG
DATA
ANALYTICS
COMPLEX
SOCIAL
INTERNET
TECHNOLOGIES
STORAGE
SYSTEMS
SOFTWARE
INFORMATION
DATABASES
NETWORKS
RECORDS
SEARCH
PARALLEL
SAN
SIZE
RESEARCH
PETABYTES
INCLUDE
MANAGEMENT
SETS
NEEDED
TIME
LARGE
DISK
SHARED
EXAMPLES
EVERY
ALSO
TOOLS
TARGET
SENSOR
DEFINITION
MOVING
MANAGE
ARCHIVES
PRACTITIONERS
CAPTURE
BUSINESS
DISTRIBUTED
GENOMICS
TERABYTES
ABILITY
PERFORMANCE
STORE
LOGS
COST
MASSIVELY
GROW
DIFFICULTY
HUNDREDS
CAPACITY
BIOLOGICAL
PROCESSING
RELATIONAL
LARGER
CONTINUES
CITATION
TENS
SET
USE
GARTNER
WORKING
ELAPSED
TOLERABLE
CASE
MAY
CURRENT
WITHIN
USING
ONE
PROCESS

目录

CONTENT

- 大数据时代
- 大数据概念
- 大数据的影响
- 大数据关键技术
- 大数据计算模式
- 代表性大数据技术



1.4 大数据关键技术

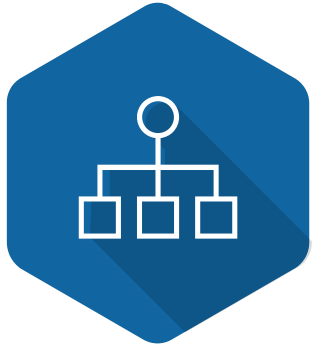
1.4 大数据关键技术



数据采集



数据存储与管理

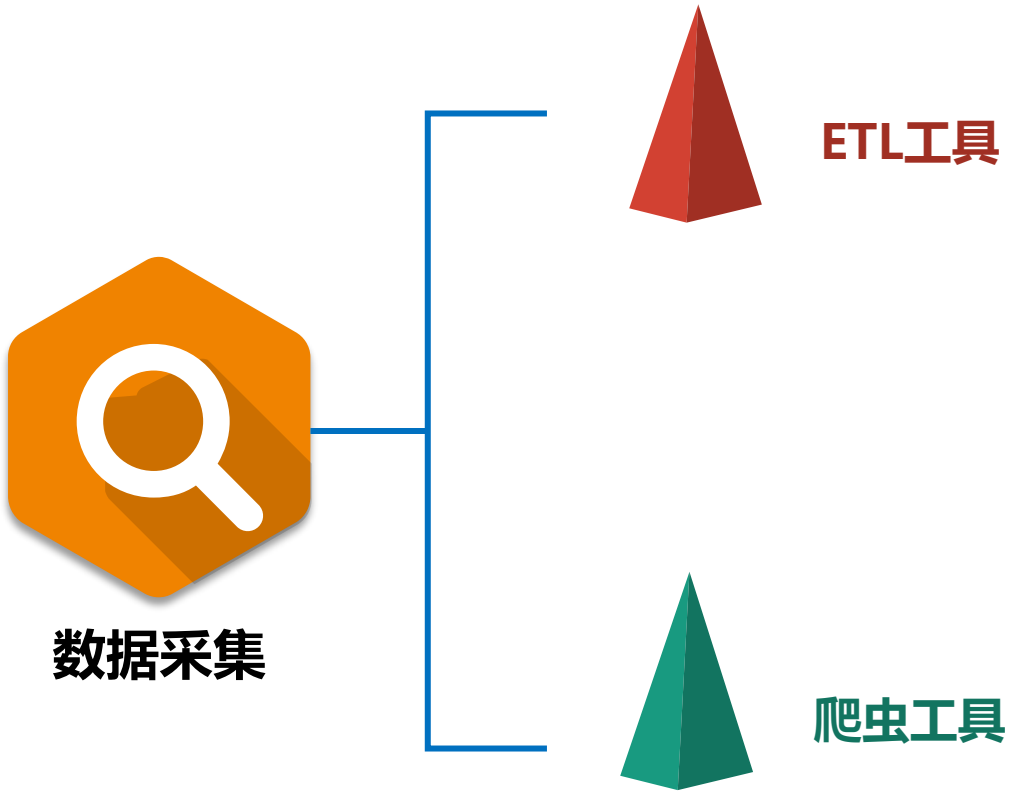


数据处理与分析



数据隐私与安全

1.4 大数据关键技术

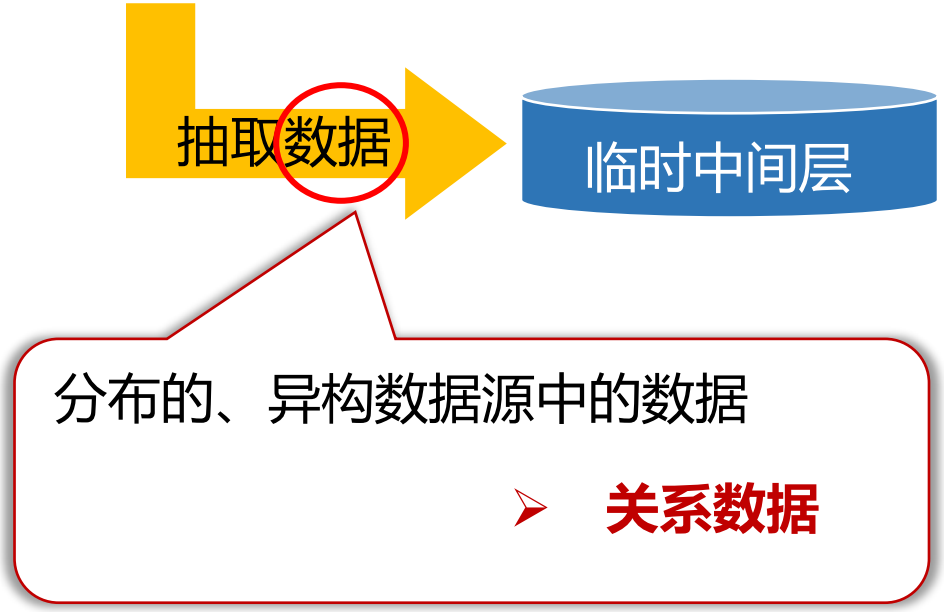


1.4 大数据关键技术



数据采集

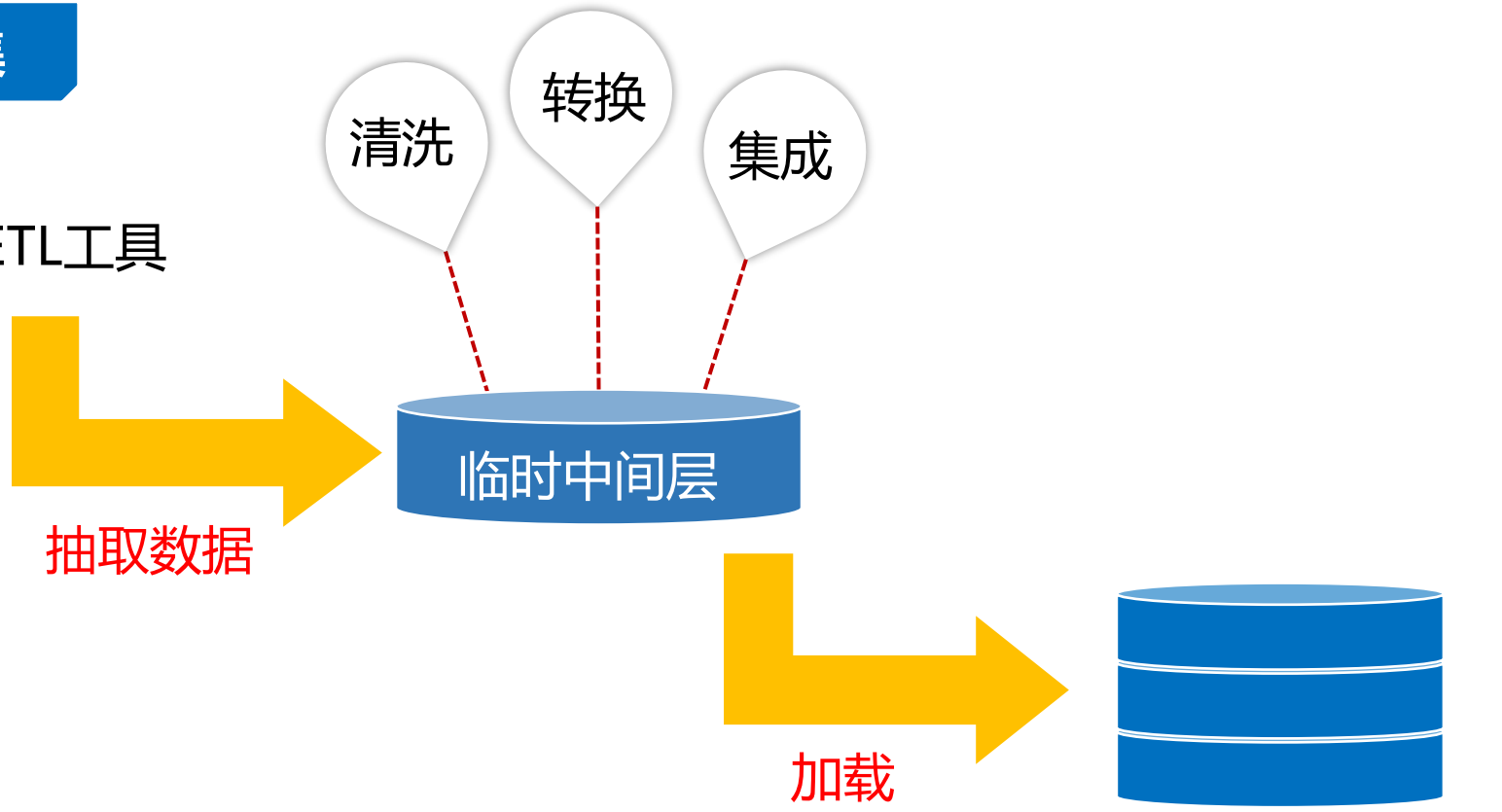
- ETL工具



1.4 大数据关键技术

数据采集

- ETL工具





1.4 大数据关键技术



数据采集后进行存储，存储到哪里？

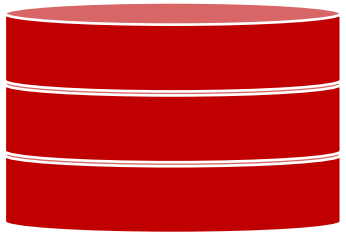


1.4 大数据关键技术

数据存储和管理



文件系统

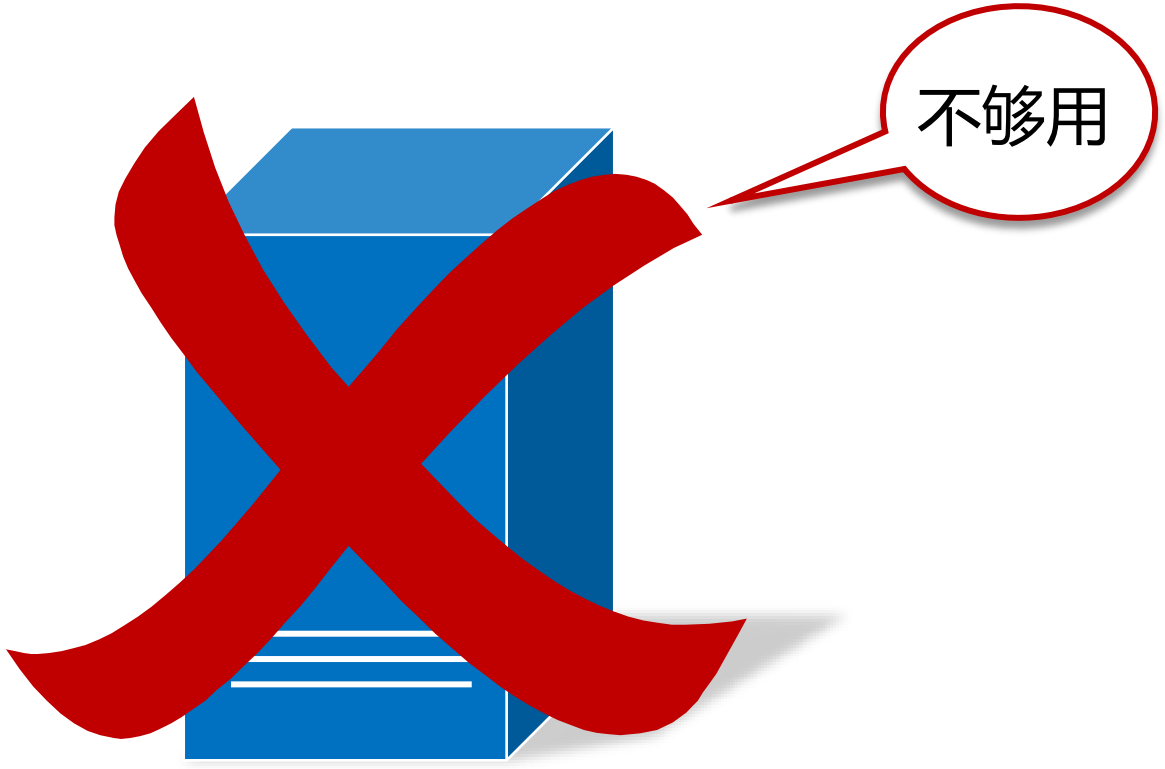


数据库



1.4 大数据关键技术

数据存储和管理



单机文件系统

1.4 大数据关键技术



数据存储和管理

- 分布式环境
  - 分布式文件系统
  - 分布式数据库

1.4 大数据关键技术

数据存储和管理

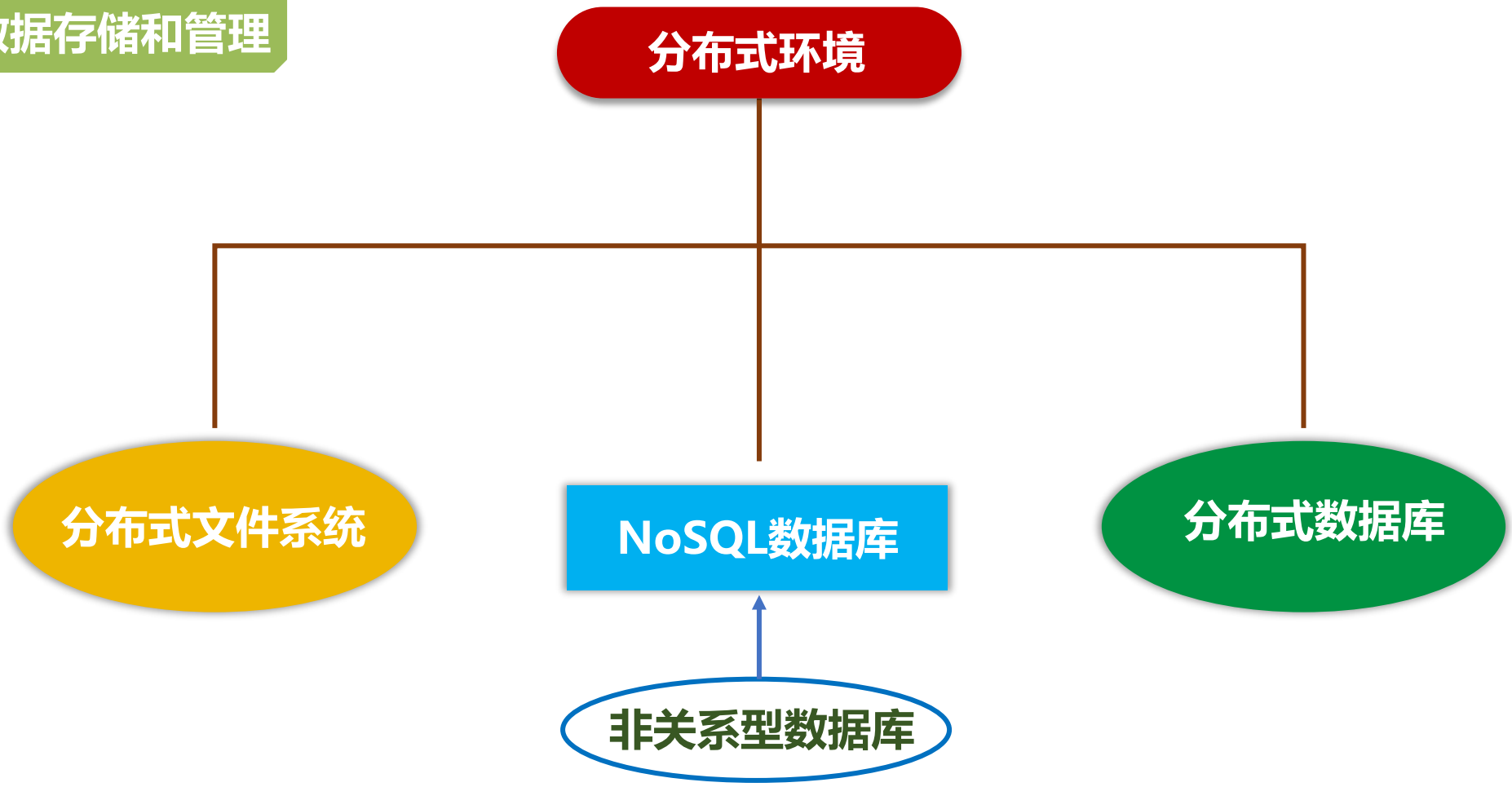


原有关系型数据库

1.4 大数据关键技术



数据存储和管理





1.4 大数据关键技术



数据处理与分析



高效计算



1.4 大数据关键技术

数据处理与分析

分布式架构

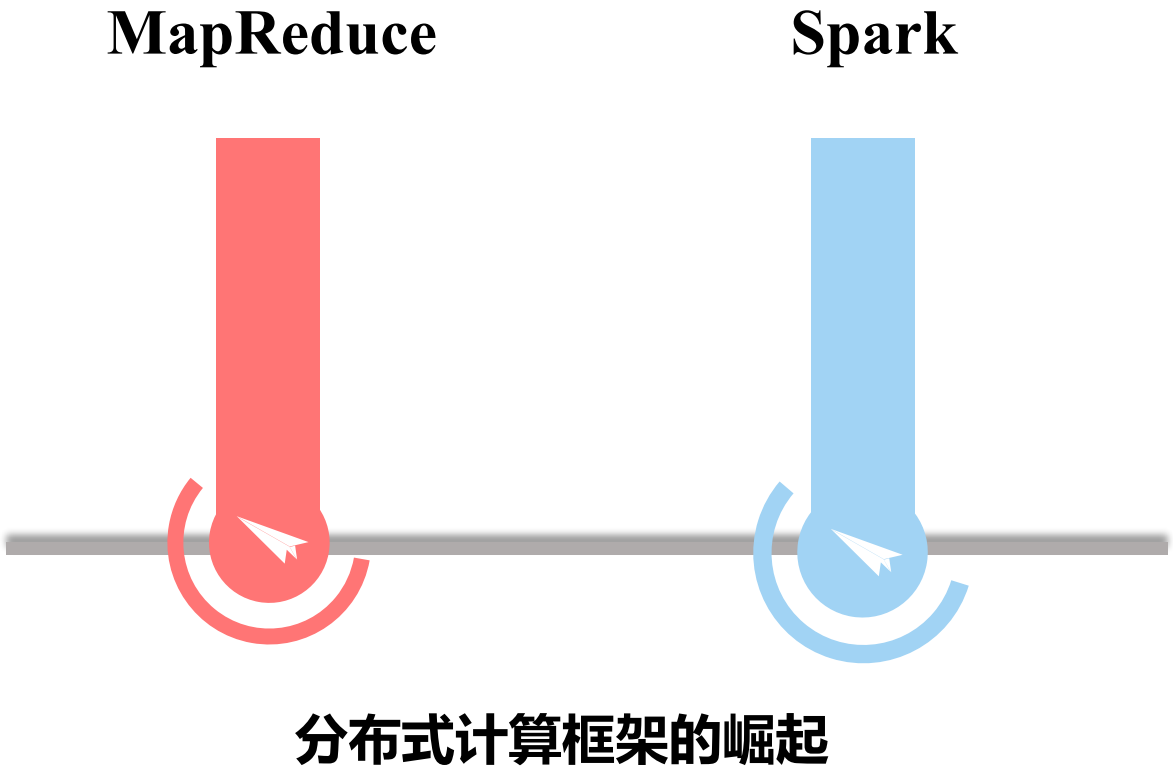
MapReduce

Spark



数据处理与分析

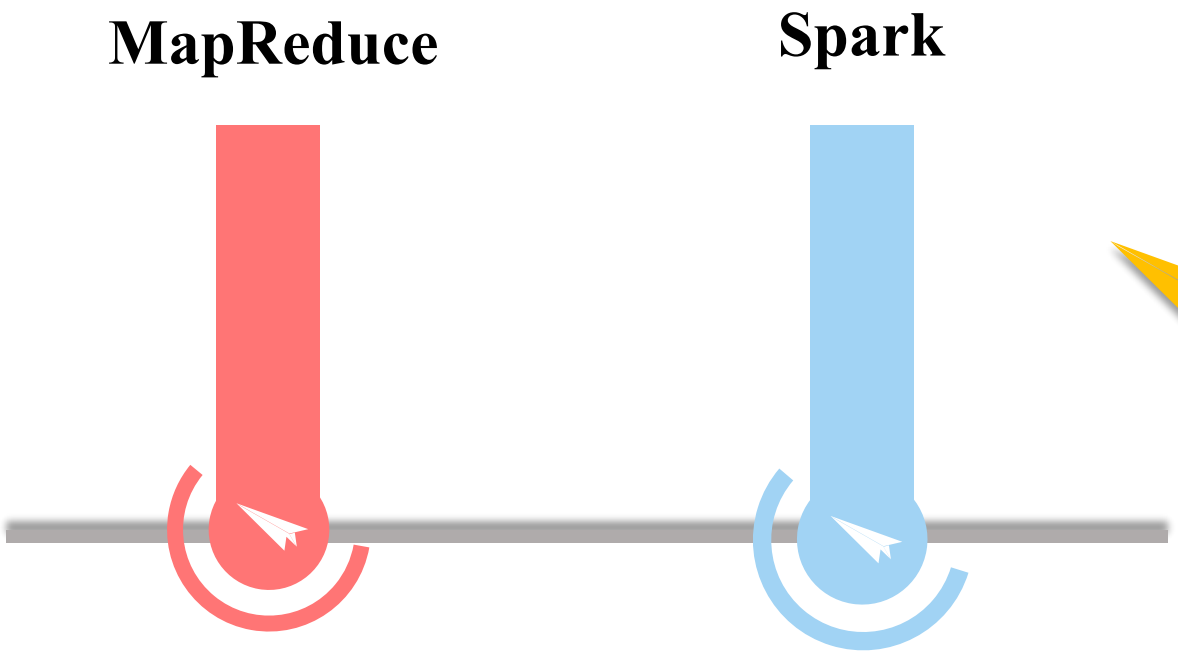
分布式架构



分布式计算框架的崛起

1.4 大数据关键技术

数据处理与分析



解决数据分布式计算

分布式计算框架的崛起

1.4 大数据关键技术

数据隐私和安全



- 隐私数据保护体系
- 数据安全保护体系

1.4 大数据关键技术



- 数据存储与管理
- 数据处理与分析

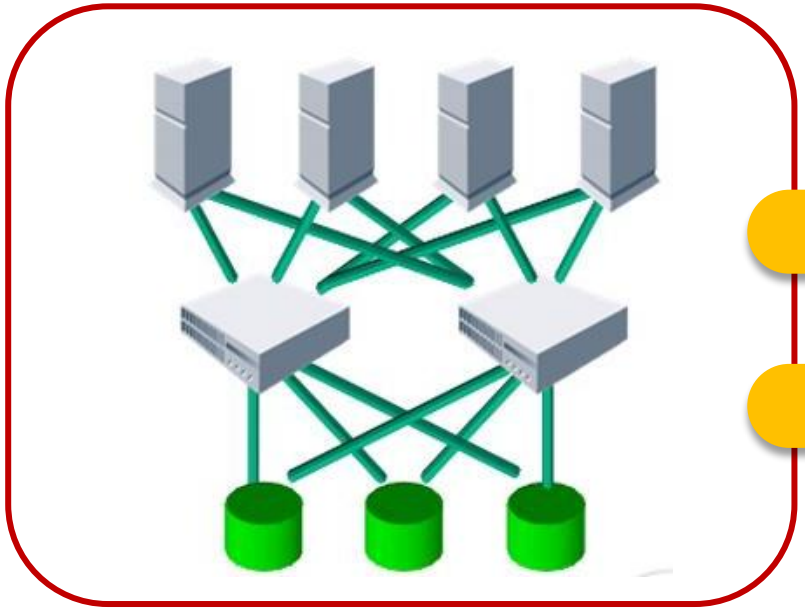
1.4 大数据关键技术



- 分布式存储
  - 解决数据存储问题
- 分布式处理
  - 解决数据高效计算问题

1.4 大数据关键技术

分布式存储



GFS

HDFS

Google

将数据分布式的存储



1.4 大数据关键技术

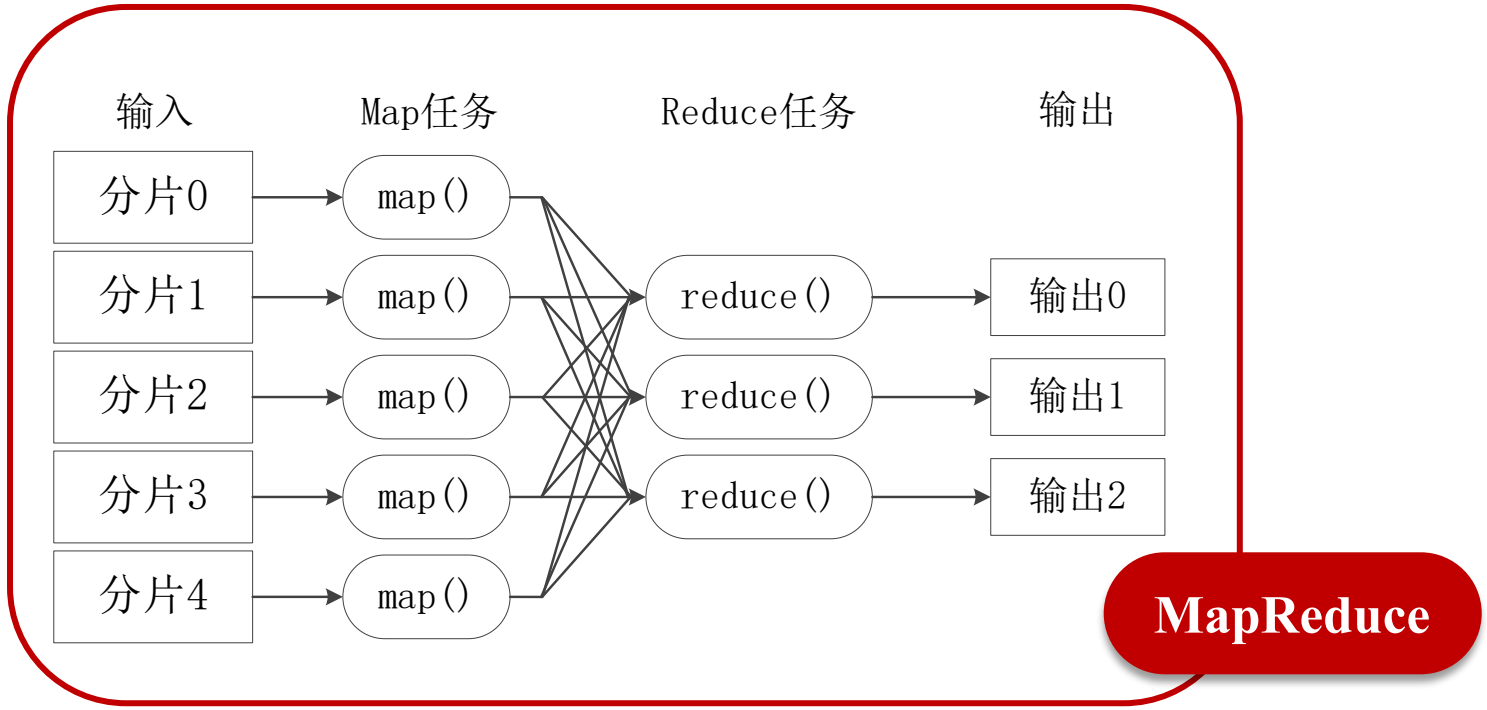
分布式处理



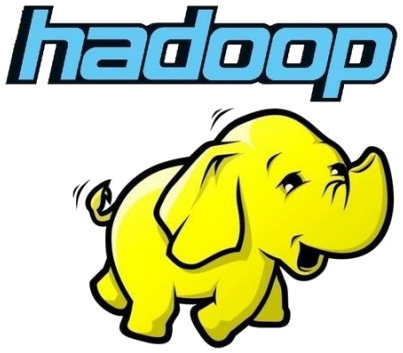
单台机器计算　　分布式存储

1.4 大数据关键技术

分布式处理



1.4 大数据关键技术



分布式存储框架

HDFS

分布式处理框架

MapReduce



1.4 大数据关键技术

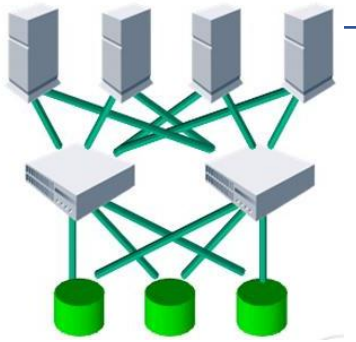


大数据技术以谷歌公司技术为代表

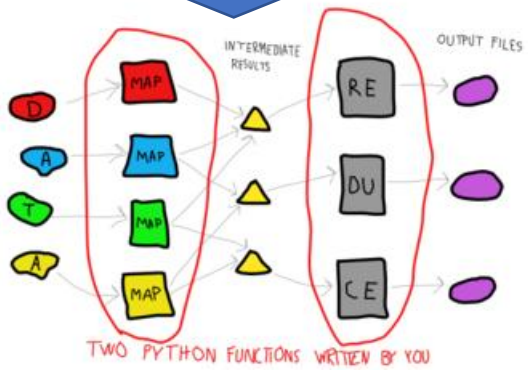
Google

1.4 大数据关键技术





分布式存储



分布式处理

GFS\HDFS
BigTable\HBase
NoSQL（键值、列族、图形、文档数据库）
NewSQL（如：SQL Azure）

MapReduce
Spark
Flink

1.4 大数据关键技术理解测验（选择题）



第1题

以下哪一项不属于大数据关键技术的组成部分？

A. 数据采集
B. 数据分析
C. 数据可视化
D. 数据隐私与安全

1.4 大数据关键技术理解测验（选择题）



第2题

ETL 工具中，“T”代表的含义是？
A. 储存（Transfer）
B. 转换（Transform）
C. 传输（Transmit）
D. 跟踪（Track）

1.4 大数据关键技术理解测验（选择题）



第3题

关于分布式文件系统，下列说法正确的是？

A. 所有数据存储在一台机器中

B. 分布式文件系统无法容错

C. GFS 是谷歌提出的分布式文件系统架构

D. HDFS 只能用于结构化数据存储

1.4 大数据关键技术理解测验（选择题）



第4题

下列哪一种不是典型的 NoSQL 数据库类型？

A. 文档型（ Document ）

B. 列族型（ Column family ）

C. 图形型（ Graph ）

D. 行表型（ Row-table ）

1.4 大数据关键技术理解测验（选择题）



第5题

在 MapReduce 中，负责初步数据处理（如提取关键字）的阶段是？

A. Shuffle

B. Join

C. Map

D. Reduce

1.4 大数据关键技术理解测验（选择题）



第6题

以下关于 Spark 的描述，不正确的是？
A. Spark 支持内存计算，适合迭代处理
B. Spark 是一个流式处理框架
C. Spark 相比 MapReduce 在性能上更优
D. Spark 是谷歌最早开发的开源项目



1.4 大数据关键技术理解测验（选择题）



第7题

以下哪项是大数据技术用来保护用户数据隐私的重要措施？

A. 数据冗余复制

B. 数据标准化

C. 数据加密与访问控制

D. 数据分片存储

1.4 大数据关键技术理解测验（选择题）



第8题

为什么关系型数据库难以适应大数据场景？

A. 数据更新太频繁

B. 数据体量太小

C. 缺乏复杂结构支持，扩展性差

D. 无法存储图片



1.4 大数据关键技术理解测验（选择题）



第9题

下列哪种技术主要用于大规模数据的分布式存储？

A. MapReduce

B. Spark

C. HDFS

D. Flink



1.4 大数据关键技术理解测验（选择题）



第10题

以下哪一项最能体现“处理与存储分离”的大数据架构思想？

A. SQL 与 NoSQL 混合使用

B. Spark 与 HDFS 的协作

C. ETL 工具的使用流程

D. Hadoop 集群的容错设计



Thank you