第十四章 大数据管理

目录

* 大数据概述
* 大数据应用
* 大数据管理系统介绍

数据模型的发展

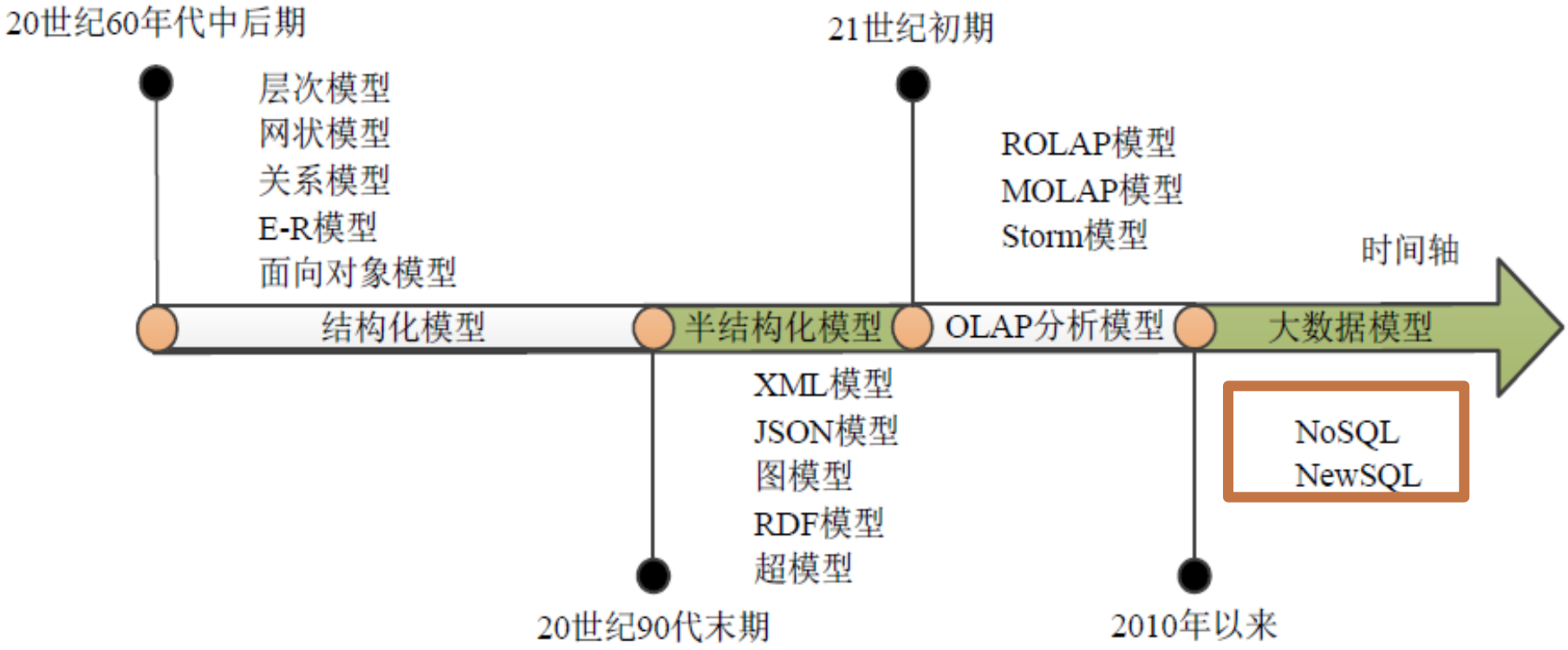


Fig.1 Timeline for the development of data models

图 1 数据模型的发展时间轴

大数据的来源

概念的发展：

- 超大规模数据（20世纪70年代，数百万条）
- 海量数据（21世纪，更大，更丰富的数据集）
- 大数据（2008年9月science发表的big data: science in the Petabyte era）

大数据定义

- 指无法在可容忍的时间里用现有的IT技术和软硬件工具对其进行感知，获取，管理，处理和服务的数据集合。
- 通常被认为是PB（1000TB）或EB或更高数据量级的数据，包括结构化的，半结构化的和非结构化的数据。

I byte=8 bit
1KB=1024 byte
1MB=1024 KB
1G=1024MB
1TB=1024GB
1PB=1024TB
1EB=1024PB

大数据的特点

* 巨量：人均5.2TB
* 多样：文本，图像，图形，音频，视频，博客等
* 快变：实时性，动态，快速产生
* 价值：潜在，巨大。
* Volume,Variety,Velocity,Variability,Veracity,
* Complexity,Value

大数据应用

* 互联网文本大数据管理与挖掘

--互联网文本大数据管理与挖掘

* 基于大数据分析的用户建模

--基于大数据分析的用户建模

* 。 。 。

大数据模型

Nosql 数据管理系统

* Non-relational
* Not only SQL

Newsql数据管理系统

* 介于关系模型与NOSQL模型之间

NOSQL 数据模型

- Key-value
- Big Table：按列存放，支持时间戳及版本控制等元数据的存贮
- Document：支持复杂的结构定义，支持数据库索引的定义。
- Graph，G(v,e) v为节点集合，每个节点具有若干属性，e 是边的集合，也可以有若干属性。支持图结果的各种基本算法。

NOSQL 数据模型
Key-Value 模型

- 键$k1$ 对应的值 $value=\{11,22,33\}$
- 键$k2$ 对应的值是一个字符串数组 {Name:Jim,Tel:1234}
- Key-Value 模型支持任意格式的值存储.

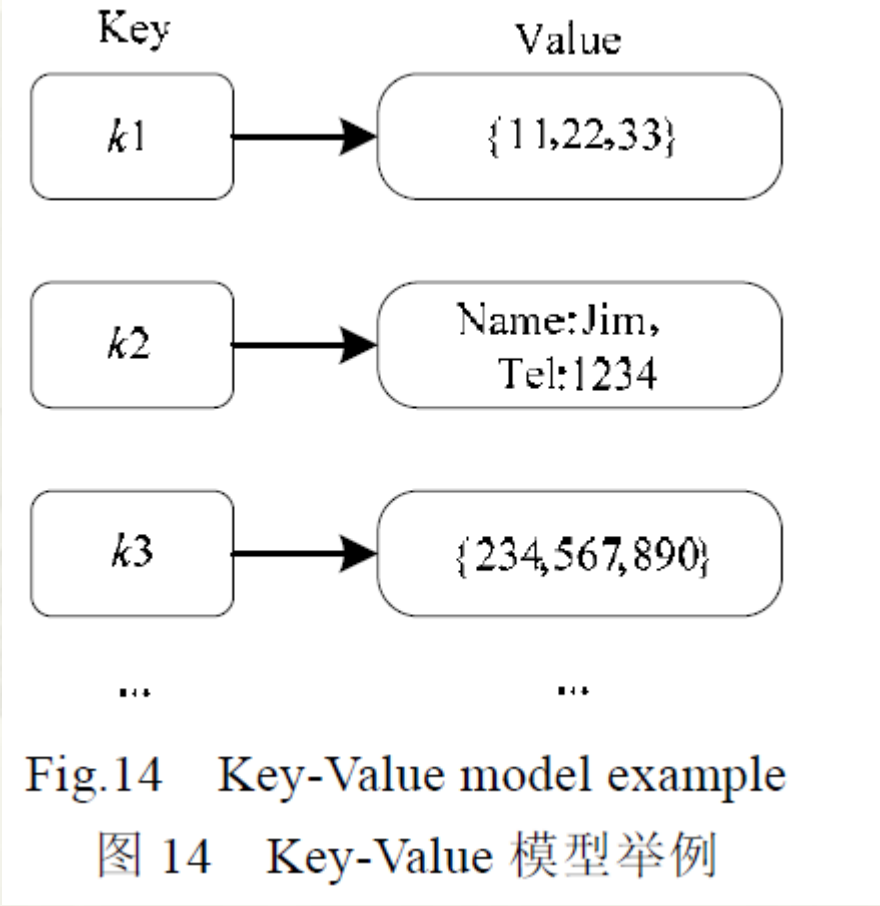


Fig.14 Key-Value model example

图 14 Key-Value 模型举例

NOSQL 数据模型
Key-Document 模型

- 数据用文档来表示。
- 面向集合：每个集合有一个唯一标识，存储在集合中的文档没有数量限制
- 无需定义模式。

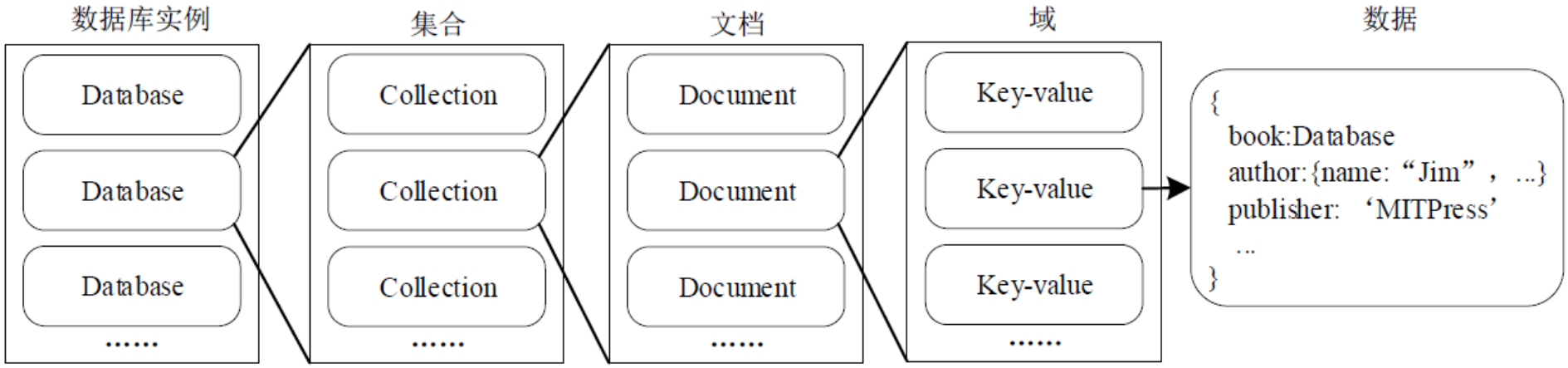


Fig.15 Key-Document model example

图 15 Key-Document 模型举例

NOSQL 数据模型
Key-Column 模型（bigTable）

* 稀疏的、分布式的、持久化的多维排序图, 并通过字典顺序来组织数据, 支持动态扩展, 以达到负载均衡.

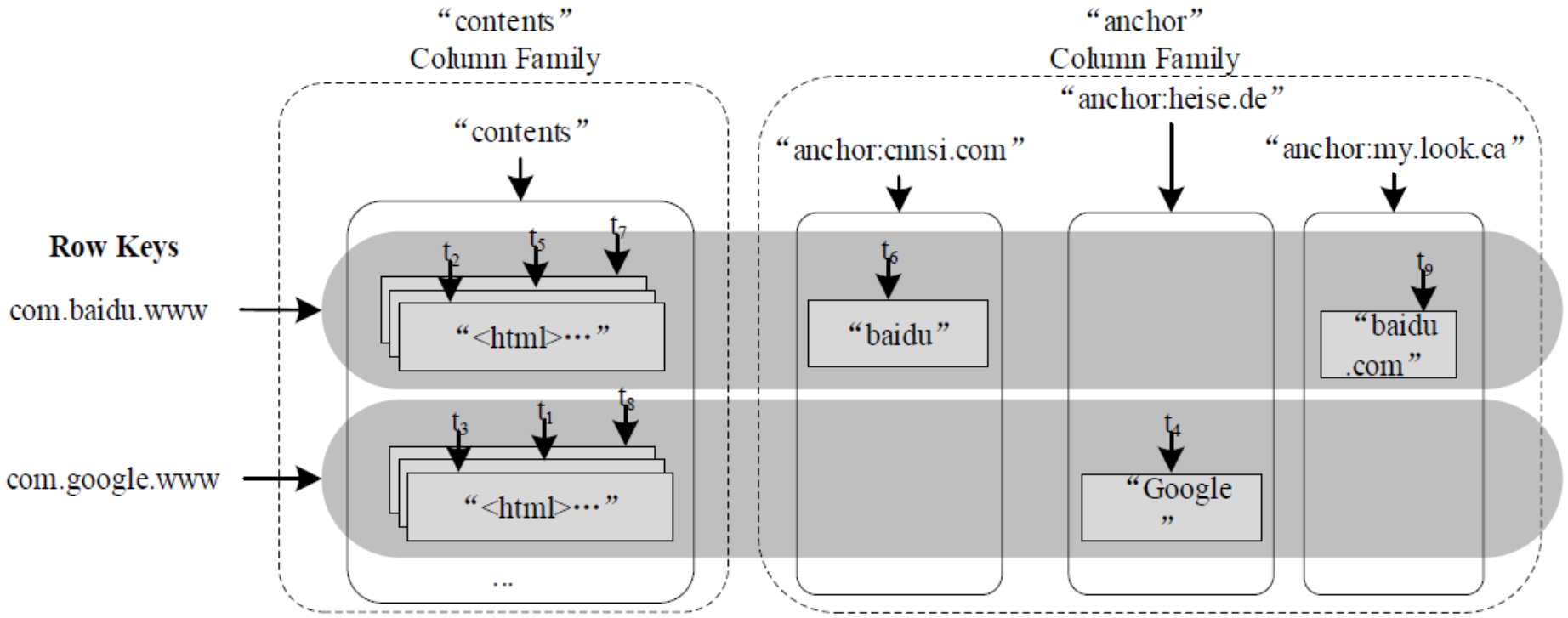


Fig.16 Key-Column model example

图 16 Key-Column 模型举例

三种NOSQL模型对比

Table 6 NoSQL database instance comparison

表 6 NoSQL 数据库实例对比

数据模型	Key-Value 模型			Key-Document 模型			Key-Column 模型	
数据库实例	Redis	Memcached	LevelDB	MongoDB	DynamoDB	CouchDB	HBase	Cassandra
实现语言	C	C	C++	C++	–	Erlang	Java	Java
二级索引	不支持	不支持	不支持	支持	支持	支持	不支持	受限制的
SQL	不支持	不支持	不支持	不支持	不支持	不支持	不支持	CQL
触发器	不支持	不支持	不支持	不支持	支持	支持	支持	支持
数据划分	分区	不支持	不支持	分区	分区	分区	分区	分区
数据副本	主从复制; 多主节点	不支持	不支持	主从复制	–	主\从复制; 主\主复制	可选择地 复制因子	可选择地 复制因子
外键	不支持	不支持	不支持	不支持	不支持	不支持	不支持	不支持
事务特性	支持	无	无	无	无	无	无	无
操作并发	支持	支持	支持	支持	支持	支持	支持	支持
数据持久	支持	不支持	支持	支持	支持	支持	支持	支持
内存计算	支持	支持	支持	支持	–	不支持	不支持	不支持

NEWSQL模型

* 使用SQL语言作为应用之间交互的主要机制。
* 遵循ACID
* 使用“无锁”的并发机制。
* 结合NOSQL和传统SQL系统的优点

模型的对比

Table 1 Comparison of RDBMS, NoSQL, and NewSQL features

表 1 RDBMS、NoSQL 和 NewSQL 特点比较

	RDBMS	NoSQL	NewSQL
SQL	支持	不支持	支持
宿主机	单机	多机/分布式	多机/分布式
类型	关系型	非关系型	关系型
模式	表	key-(value,column,document)	二者都支持
物理存储	磁盘+缓存	磁盘+缓存	磁盘+缓存
特性	ACID	CAP,BASE	ACID
查询复杂度	低	高	高
一致性	高	最终一致性	高
可用性	故障转移	高	高
可扩展性	垂直扩展	水平扩展	水平扩展
复制	可配置	可配置	自动
安全性	高	低	低
大数据处理	支持,但效率低	支持	充分支持
OLTP	不完全支持	支持	充分支持

大数据管理的新格局

- 面向**操作性**应用：基于行存贮的关系数据库系统，并行数据库系统，实时计算的内存数据库系统，NoSQL 系统，以及结合 nosql 和关系的新型系统（VoltDB）
- 面向**分析型**应用：列存贮数据库（MonetDB）和基于列存贮技术的内存数据库（MonetDB,VectorWise,Hana），以及采用MapReduced 技术，面向分析应用的Nosql系统。

数据管理技术新格局

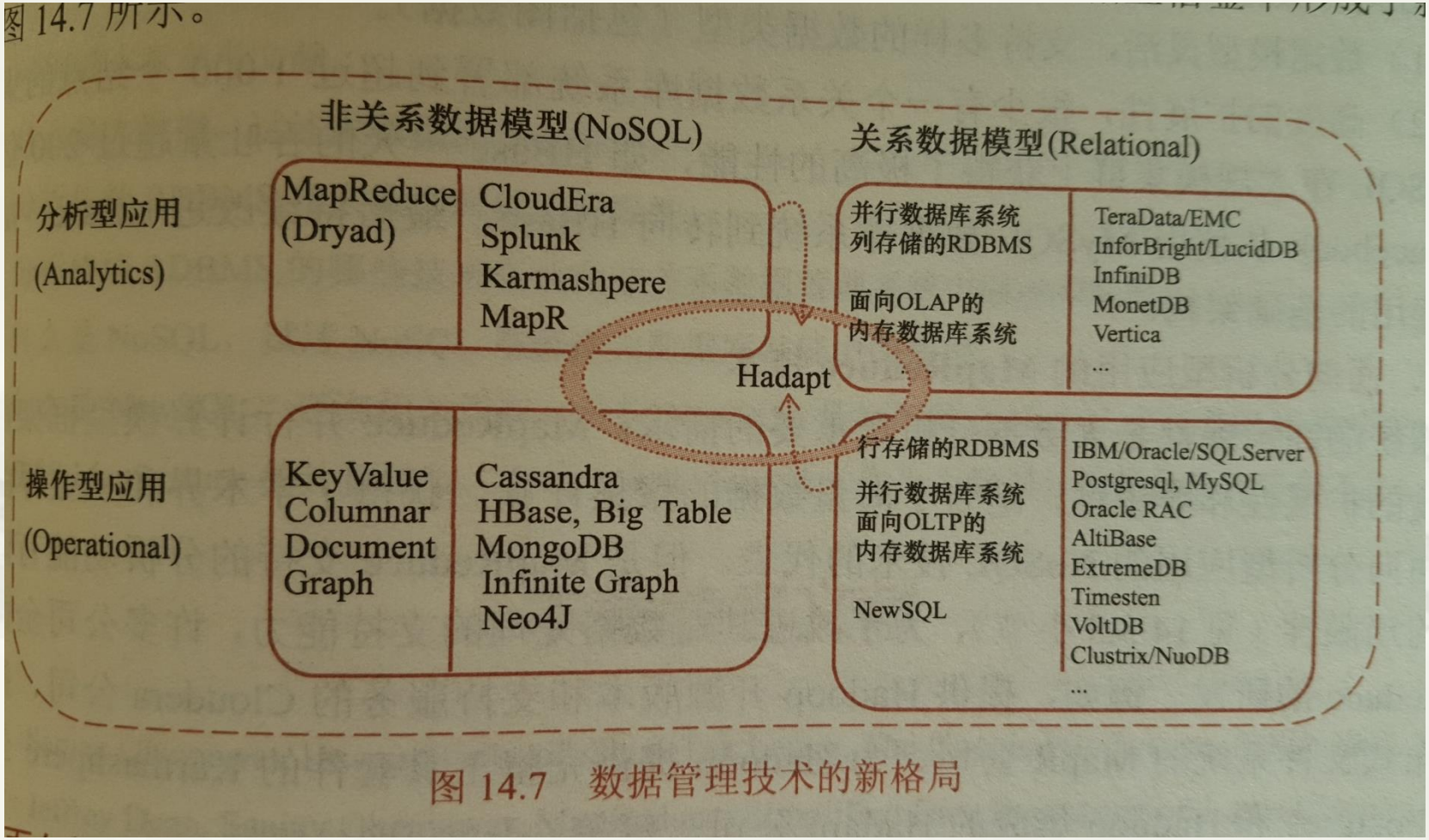


图 14.7　数据管理技术的新格局

大数据管理的新格局

* 面向操作型应用的数据库技术

基于行存贮的关系数据库系统，面向实时计算的内存数据库系统，形成新的NEWSQL系统

* 面向分析型应用的关系数据库技术

面向分析的列存贮数据库和内存数据库，内存数据库利用大内存，多核CPU等硬件系统。

* 面向操作型应用的NOSQL技术

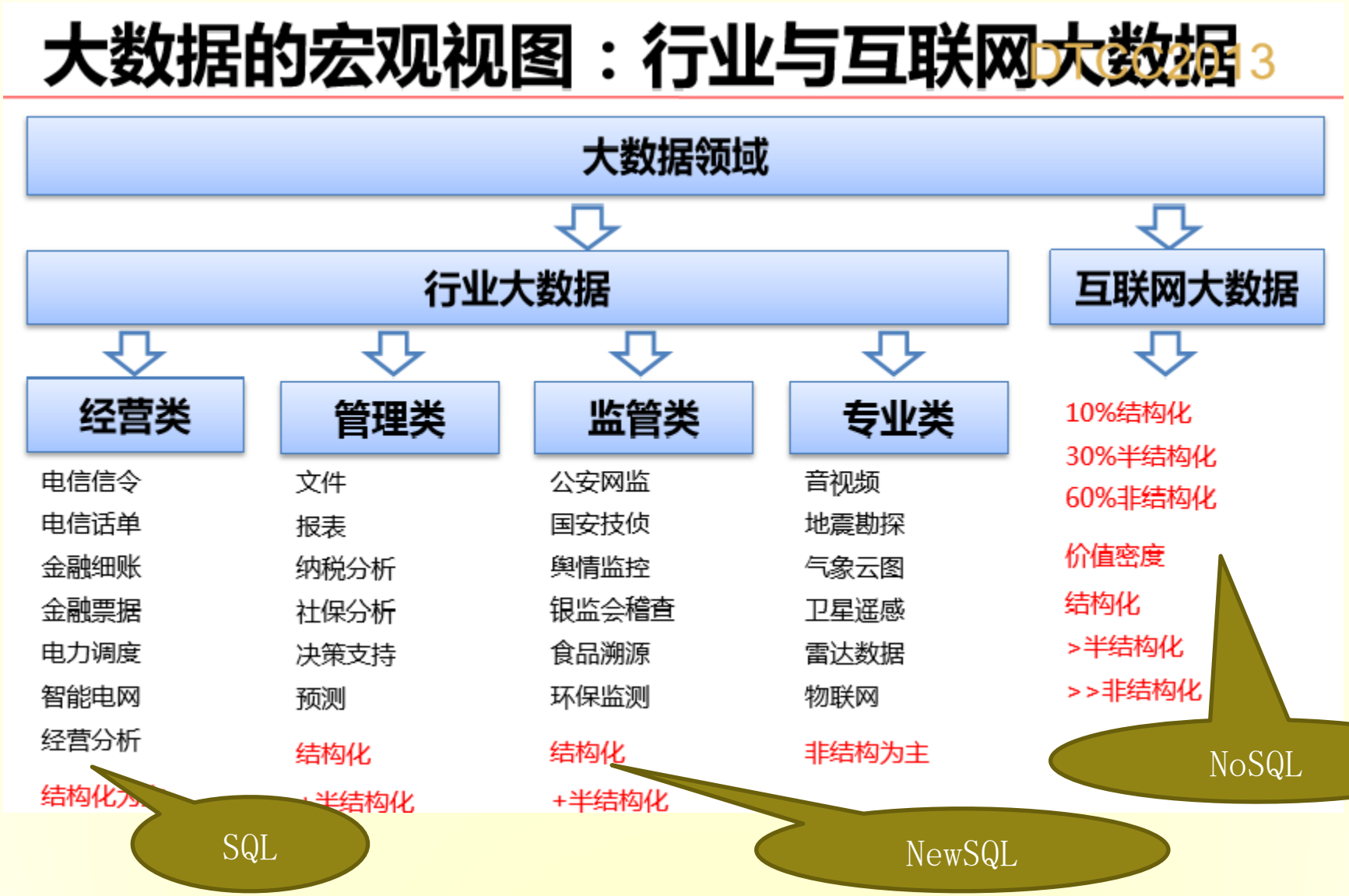
NOSQL系统，数据模型灵活，扩展性好

* 面向分析型的MAPreduce技术

并行计算的框架，简单，高度的扩展性和容错性，适合海量数据的聚集计算。

The state of the art

from DTCC2014



参考文献

* 大数据管理：概念、技术与挑战
* 大数据分析与高速数据更新
* 大数据的一个重要方面：数据可用性
* 大数据隐私管理
* 分布式大数据函数依赖发现
* 大数据图数据匹配技术综述