

大数据技术综述

范颖捷 | 2018年9月

目录 >

CONTENTS

- 1 什么是大数据
- 2 大数据技术概览
- 3 大数据产业生态



1 chapter

什么是大数据

- ✓ 基本特征
- ✓ 应用场景

	传统数据	大数据
数据量	GB → TB	TB → PB以上
速 度	数据量稳定、增长不快	实时产生处理、年增长率超60%
多样性	结构化数据	结构化、半结构化、非结构化数据
价 值	统计报表	机器学习、深度学习

大数据是指超出传统数据库工具收集、存储、管理和分析能力的数据集。与此同时，及时采集、存储、聚合、管理数据，以及对数据深度分析的新技术和新能力，正在快速增长，就像预测计算芯片增长速度的摩尔定律一样。

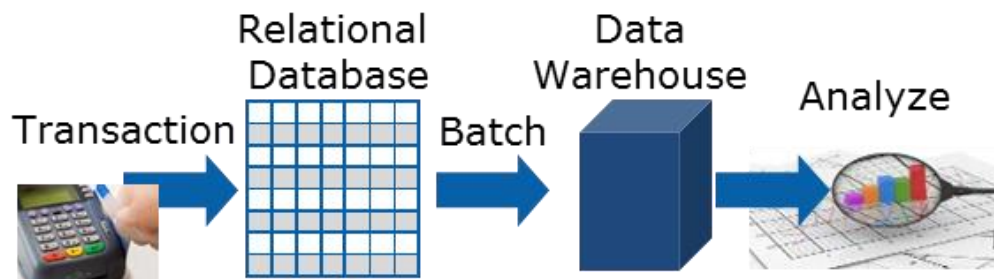
— McKinsey Global Institute

- ✓ 数据规模巨大（Volume）
- ✓ 生成和处理速度极快（Velocity）

- ✓ 数据类型多样（Variety）
- ✓ 价值巨大但密度较低（Value）

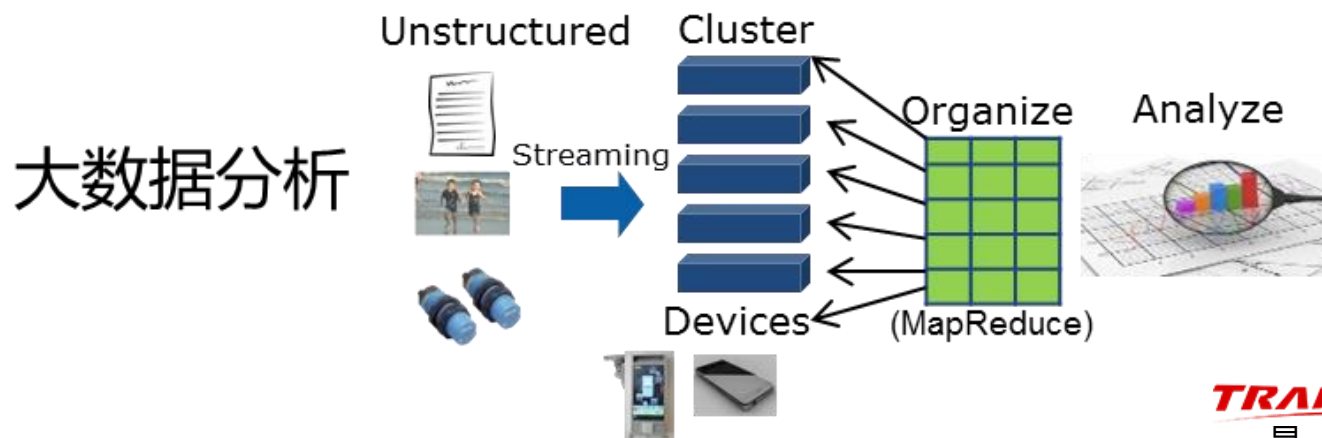
➤ 基于大数据的数据仓库

Traditional Data Analysis



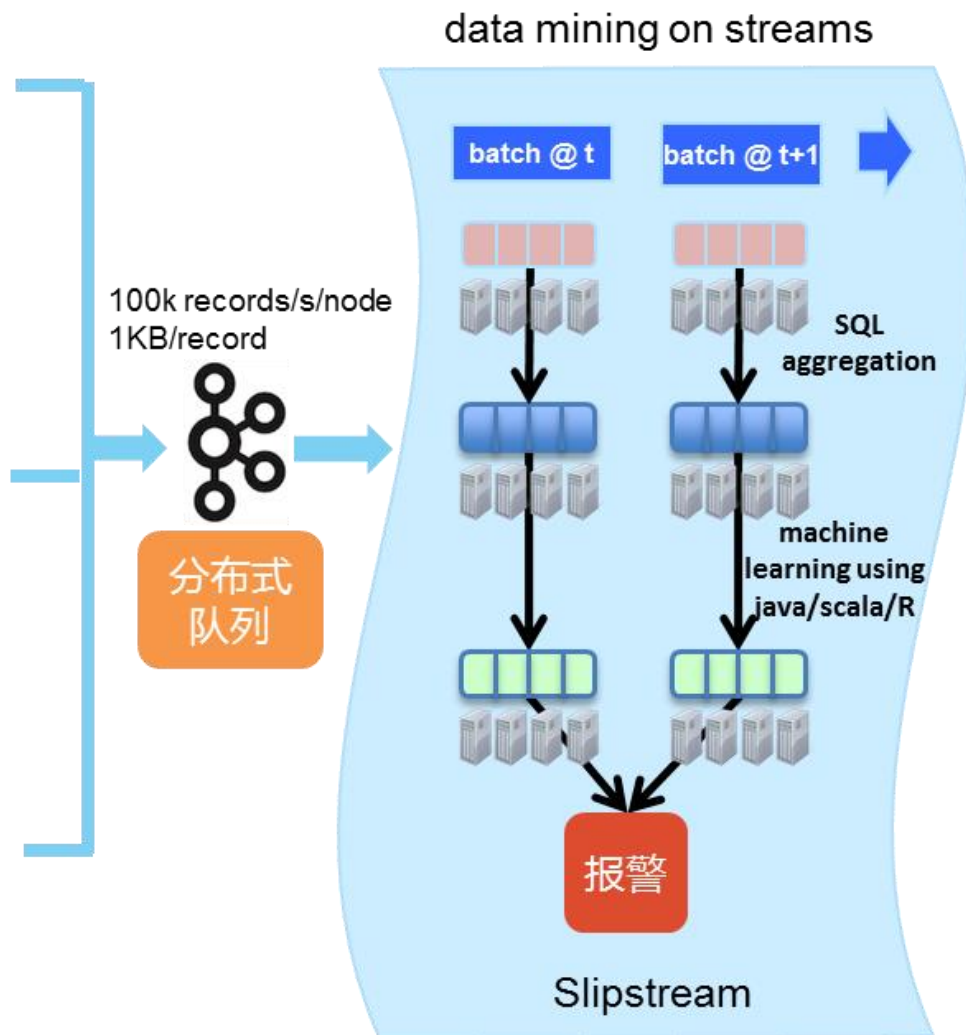
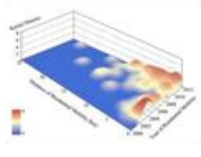
传统数据分析

Big Data Analysis




大数据分析

➤ 基于大数据的实时流处理



1. Streaming processing and batch processing are unified in one programming model
2. SQL and its extension is the unified declarative language for device monitoring and diagnostics.
3. **ANSI SQL 2003 and PL/SQL** are supported on streaming events.
4. Linear Algebra
5. Machine learning

Usage cases in IoT & FS:
Real-time event monitoring
Real-time dashboard & statistics
Real-time outlier detection
Real-time fraud detection



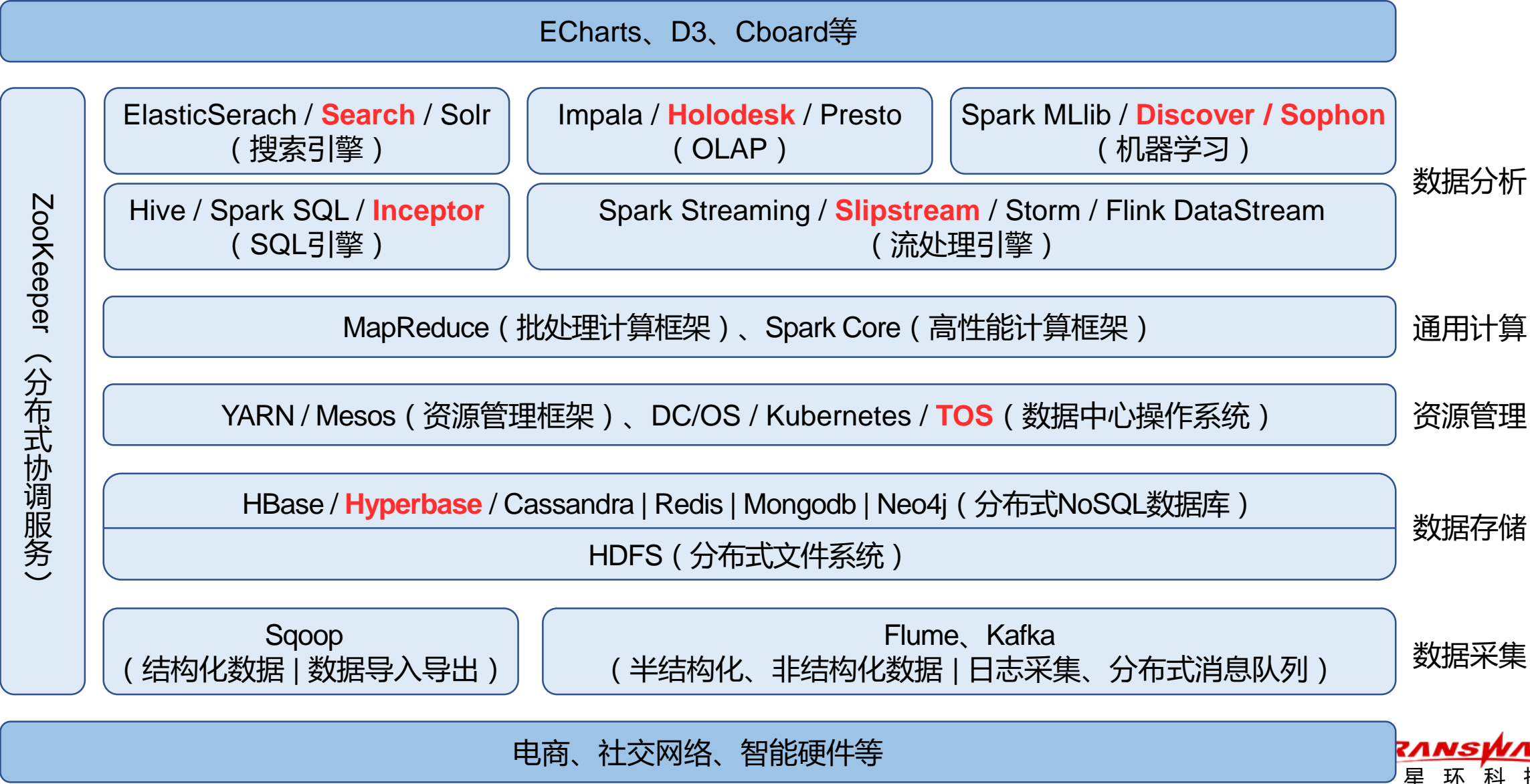
2 chapter

大数据技术概览

- ✓ Hadoop编年史
- ✓ 大数据技术体系

阶段	时间	事件
前Hadoop时代	2002.10	Doug Cutting、Mike Cafarella创建了开源网页爬虫项目Nutch
	2003.10	Google发表了Google File System论文
	2004.07	Doug Cutting、Mike Cafarella在Nutch中实现了GFS的功能
	2004.10	Google发表了MapReduce论文
	2005.02	Mike Cafarella在Nutch中实现了MapReduce的功能
	2006.01	Doug Cutting加入Yahoo，将Hadoop发展成一个可在网络上运行的系统
	2006.02	Apache Hadoop项目正式启动，并支持MapReduce和HDFS独立发展
	2006.02	Yahoo的网格计算团队采用Hadoop技术
	2006.03	Yahoo建立了第一个用于开发的Hadoop集群
	2006.04	第一个Apache Hadoop版本发布
	2006.11	Google发表了Bigtable论文
	2007.04	Yahoo Hadoop集群发展成两个1000个节点的集群
	2008.01	Hadoop成为Apache顶级项目
	2008.02	Yahoo运行了世界最大的Hadoop应用，宣布其搜索引擎产品部署在一个拥有一万个内核的Hadoop集群上

阶段	时间	事件
Hadoop时代	2008.06	Hadoop的第一个SQL框架Hive成为Hadoop子项目
	2008.08	第一个Hadoop商业化公司Cloudera成立
	2008.11	Apache Pig的第一个版本发布
	2009.03	Cloudera推出世界上首个Hadoop发行版——CDH，并完全开放源码
	2009.07	MapReduce和HDFS成为Hadoop子项目
	2010.05	HBase脱离Hadoop项目，成为Apache顶级项目
	2010.09	Hive脱离Hadoop项目，成为Apache顶级项目
	2010.09	Pig脱离Hadoop项目，成为Apache顶级项目
	2011.01	ZooKeeper脱离Hadoop项目，成为Apache顶级项目
	2012.03	HDFS NameNode HA加入Hadoop主版本
	2012.08	YARN成为Hadoop子项目
后Hadoop时代	2013.11	星环科技发布了国内首个全面支持Spark和Hadoop2.0的大数据基础平台软件——TDH
	2014.02	Spark代替MapReduce成为Hadoop的缺省计算引擎，并成为Apache顶级项目
	2015.10	Cloudera公布继HBase以后的第一个Hadoop原生存储替代方案——Kudu



➤ HDFS

- 概念

- Hadoop分布式文件系统（Hadoop Distributed File System）
- 在开源大数据技术体系中，地位无可替代

- 特点

- 高容错：数据多副本，副本丢失后自动恢复
- 高可用：NameNode HA，安全模式
- 高扩展：10K节点规模
- 简单一致性模型：一次写入多次读取，支持追加，不允许修改
- 流式数据访问：批量读而非随机读，关注吞吐量而非时间
- 大规模数据集：典型文件大小GB~TB级，百万以上文件数量，PB以上数据规模
- 构建成本低且安全可靠：运行在大量的廉价商用机器上，硬件错误是常态，提供容错机制



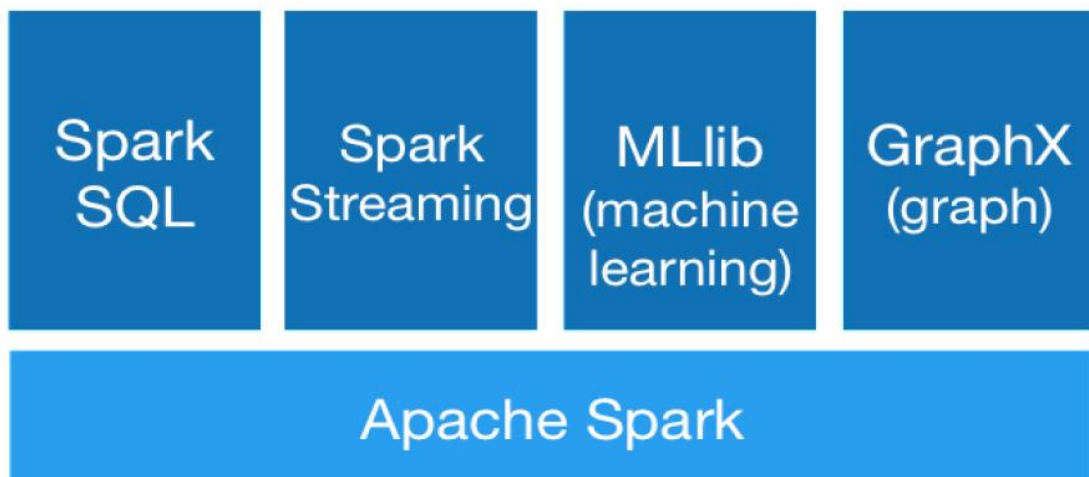
➤ MapReduce

- 概念
 - 面向批处理的分布式计算框架
 - 编程模型：将MapReduce程序分为Map、Reduce两个阶段
- 核心思想
 - 分而治之，分布式计算
 - 移动计算，而非移动数据
- 特点
 - 高容错：任务失败，自动调度到其他节点重新执行
 - 高扩展：计算能力随着节点数增加，近似线性递增
 - 适用于海量数据的离线批处理
 - 降低了分布式编程的门槛



➤ Spark

- 由加州大学伯克利分校的AMP实验室开源
- 高性能分布式通用计算引擎
 - Spark Core: 基础计算框架（批处理、交互式分析）
 - Spark SQL: SQL引擎（海量结构化数据的高性能查询）
 - Spark Streaming: 实时流处理（微批）
 - Spark MLlib: 机器学习
 - Spark GraphX: 图计算
- 采用Scala语言开发
- 特点
 - 计算高效: 内存计算、Cache缓存机制、DAG引擎、多线程池模型
 - 通用易用: 适用于批处理、交互式计算、流处理、机器学习、图计算等多种场景
 - 运行模式多样: Local、Standalone、YARN/Mesos



➤ YARN

- 概念
 - Yet Another Resource Negotiator, 另一种资源管理器
 - 为了解决Hadoop 1.x中MapReduce的先天缺陷
 - 分布式通用资源管理系统
 - 负责集群资源的统一管理
 - 从Hadoop 2.x开始, YARN成为Hadoop的核心组件
- 特点
 - 专注于资源管理和作业调度
 - 通用: 适用各种计算框架, 如: MapReduce、Spark
 - 高可用: ResourceManager高可用、HDFS高可用
 - 高扩展



➤ Hive

- 概念
 - Hadoop数据仓库：企业决策支持
 - SQL引擎：对海量结构化数据进行高性能的SQL查询
 - 采用HDFS或HBase为数据存储
 - 采用MapReduce或Spark为计算框架
- 特点
 - 提供类SQL查询语言
 - 支持命令行或JDBC/ODBC
 - 提供灵活的扩展性
 - 提供复杂数据类型、扩展函数、脚本等



➤ HBase

- 概念

- Hadoop Database
- Google BigTable的开源实现
- 分布式NoSQL数据库
- 列式存储：主要用于半结构化、非结构化数据
- 采用HDFS为文件存储系统

- 特点

- 高性能：支持高并发写入和查询
- 高可用：HDFS高可用、Region高可用
- 高扩展：数据自动切分和分布，可动态扩容，无需停机
- 海量存储：单表可容纳数十亿行，上百万列




➤ Elasticsearch

- 开源的分布式全文检索引擎
- 基于Lucene实现全文数据的快速存储、搜索和分析
- 处理大规模数据：PB级以上
- 具有较强的扩展性，集群规模可达上百台
- 首选的分布式搜索引擎



elastic



3 chapter

大数据产业生态

✓ 产品分布

数据分析与人工智能

SAS

SPSS

AzureML

PAI

Sophon

大数据基础平台

CDH

HDP

MapR

FI

TDH

数据中心操作系统

DC/OS

Kubernetes

TOS

Docker / CoreOS

AWS
阿里云

TDC



Q&A

TRANSWARP
星环科技

温故知新

- 大数据的基本特征是什么？
- Hadoop经历了几个发展阶段，各有什么特点？
- 大数据技术体系大致分为几层？每层包含哪些技术？
- Apache Hadoop项目包含哪些子项目？简述一下它们的功能。
- Spark包含哪些组件？简述一下它们的功能。

