

	传统数据	大数据
数据量	$GB \to TB$	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)

- ✓ 数据类型多样 (Variety)
- ✓ 生成和处理速度极快 (Velocity)
- ✓ 价值巨大但密度较低 (Value)



ZooKeeper(分布式协调服务)

ECharts、D3、Cboard等

数据可视化

ElasticSerach / Search / Solr (搜索引擎) Impala / Holodesk / Presto (OLAP)

Spark MLlib / Discover / Sophon (机器学习)

数据分析

Hive / Spark SQL / Inceptor (SQL引擎)

Spark Streaming / Slipstream / Storm / Flink DataStream (流处理引擎)

MapReduce (批处理框架)、Spark Core (高性能计算框架)

通用计算

YARN (资源管理框架)、Mesos / Kubernetes / TOS (数据中心操作系统)

资源管理

HDFS(分布式文件系统)、HBase / Hyperbase / Cassandra / Redis / Mongodb(分布式NoSQL数据库)

数据存储

Sqoop (结构化数据 | 数据导入导出) Flume、Kafka (半结构化、非结构化数据 | 日志采集、分布式消息队列)

数据采集

电商、社交网络、智能硬件等

数据源



TDH技术架构



