# 大数据技术综述与发展展望

赵　鹏，朱祎兰

（中国运载火箭技术研究院，北京 100076）

**摘　要：**以大数据、云计算等为代表的新一代信息技术正拉开中国航天数字化能力建设的崭新篇章，数据治理工作正处在勇开新局的关键时期。如何合理应用大数据存储、处理、分析、管理、流通与安全技术，构建航天大数据基础平台传输架构、存储架构及技术架构，保障数据治理规划与标准框架落地，沉淀航天数据资产，构建与航天型号研制体系特点相匹配的数据应用，是一个重要的命题。总结提炼了大数据的主要特征，全面梳理大数据技术体系，并面向不同数据处理场景，归纳适用的大数据技术手段，为航天数据治理的技术架构设计、技术选型等工作推进奠定坚实基础。

**关键词：**大数据；大数据技术体系；数据处理

**中图分类号：**TP391.9　　**文献标识码：**A　　**文章编号：**2096-4080（2022）01-0055-06

# Overview and Development Prospects of Big Data Technology

ZHAO Peng, ZHU Yilan

(China Academy of Launch Vehicle Technology, Beijing 100076, China)

**Abstract:** The new generation of information technology represented by big data and cloud computing is opening a new chapter in the construction of China aerospace's digital capabilities, and the data governance work is also in a critical period. It is a vital thesis that how to apply big data technologies in building the technical architecture of the aerospaces big data basic platform, in ensuring the implementation of the aerospaces data governance framework, in accumulating the data assets in aerospace. This article summarizes and refines the main characteristics of big data, comprehensively sorts out the big data technology system, and summarizes the applicable big data technology methods for different data processing scenarios, which lays solid foundation for the technical architecture design and technology selection of our institute for data governance.

**Key words:** Big data; Big data technology system; Data processing

## 0　引言

纵观整个数字技术的发展历史，自 1980 年前后，随个人计算机开始普及，人类社会经历了 3 次信息化浪潮，数字技术从军事领域走向经济社会各个方面。存储设备容量、CPU 处理能力、网络带宽等基础设施水平快速迭代升级，引发数据的产生、传输、存储、处理方式不断跃迁，在数据、算力和算法的共同繁荣之下，以大数据技术为典型代表的新兴数字技术体系推动第 3 次信息化浪潮席卷全球。大数据技术已然成为人类社会发展的底层驱动力量，推动着生产力、生产关系的深刻变革[1]。

作为技术产品高度复杂、生产组织高度复杂、

**收稿日期：**2021-11-04；**修订日期：**2021-12-24

**作者简介：**赵鹏（1982-），男，博士，高级工程师，主要研究方向为企业数字化和大数据。E-mail：18301502283@163.com

**通信作者简介：**朱祎兰（1991-）女，硕士，工程师，主要研究方向为大数据等。E-mail：zyl_buaa@163.com

经营管理高度复杂的研发、生产一体化科研单位，中国运载火箭技术研究院拥抱大数据，加快迈向以数据赋能生产、以数据驱动经营的新阶段，已成为顺应历史潮流，提升生产经营能力的必然选择。只有充分掌握大数据的基本特征，理清大数据生态体系各类技术及其适用场景，才能在保障数据安全前提下，打通型号产品研制及经营管控各环节数据壁垒，充分激发数据资产价值。

## 1 大数据的主要特征

数据量大、速度快、类型多、复杂性高是大数据的主要自然特征。随着大数据逐步成为驱动数字经济发展的核心要素，使其与劳动、资本、技术、土地一起构成经济新范式，重视和利用数据要素价值已成为社会各界的广泛共识[2]。

### 1.1 体量巨大

对于当前各领域的数据集合，TB、PB 的数据量级单位已不能满足需求，目前已开始使用 EB 和 ZB 进行衡量[3]。

### 1.2 速度快

一般指处理速度与产生速度。大数据往往和人工智能、物联网等技术结合应用，对数据的实时响应要求高。大数据的处理效率又称为“1 秒定律”，即可以在秒级时间内获取分析结果。

### 1.3 维度多

大数据具有多个维度。以人为例，具有性别、年龄、身高、体重、身份证号码、学历、家庭住址等多个属性。数据的多维度、多层次属性应用到社会生产的各个领域，可以加速流程再造，提高生产效率，加速供需信息匹配，提高协同效率，从而创造更大的价值。

### 1.4 复杂性高

大数据复杂性高。由于记录工具不同和应用场景不同，一方面，数据结构不尽相同，呈现出文字、图像、音频、视频等不同的形式；另一方面，在内容逻辑层面也出现看似杂乱无章，实际有章可循的现象。

### 1.5 依附属性强

与传统有形资源不同，大数据具有虚拟性、无形性，无法单独存在，往往需要依赖硬件设备存储，依赖软件平台读取、操作。只有将数据存储在相应介质并通过设备显示，数据才能以更直观的方式被感知、度量、传输、分析与应用，数据质量的好坏、价值的高低才可能被评估。数据的虚拟性、无形性决定了其管理与数据平台管理不可分割，数据的价值与平台算力、算法模型密切相关，倒逼现行资产管理办法升级完善[4]。

### 1.6 关键生产要素

在农业时代，土地是关键生产要素；工业时代以劳动、资本、技术作为关键生产要素；数字时代，随着国家将数据列为第 5 大生产要素[5]，大数据将参与到市场的投入、管理、产出和分配的各个阶段。

## 2 大数据技术体系全景

随着大数据技术体系的不断成熟，内部技术构成不断分化，从面向海量数据的存储、处理、分析等需求的核心技术，延展到数据管理、流通、安全等配套技术，逐渐形成了层次清晰、分工完备的大数据技术体系，如图 1[6]所示。

1）数据基础技术应对多种数据特征产生。针对大数据数据量大、数据源异构多样、数据时效性高等特征催生了高效完成海量异构数据存储与计算的技术需求。在这种需求下，传统集中式计算架构出现难以逾越的瓶颈，传统关系型数据库单机的存储及计算性能有限，出现了分布式存储及分布式计算框架。面向海量结构化及非结构化数据批处理，出现了基于 Hadoop、Hive 和 Spark 生态体系的分布式批处理计算框架；面向时效性数据进行实时计算反馈的需求，出现了 Storm、Flink 及 Spark Streaming 等分布式流处理计算框架[7]。

2）数据管理技术提升数据质量与可用性。随相对基本与急迫的数据存储、计算需求已经在一定程度上得到满足后，如何进行数据管理与沉淀成为了一个主要的需求。由于企业内部大量数据产生链条长、复杂度高，但普遍缺乏有效管理，常常存在数据获取难、准确性低、实时性差、标准混乱等问题[8]，导致数据后续的使用存在众多障碍。在这种情况下，用于数据整合的数据集成技术以及用于实现一系列数据资产管理功能的数据管理技术随之出现。

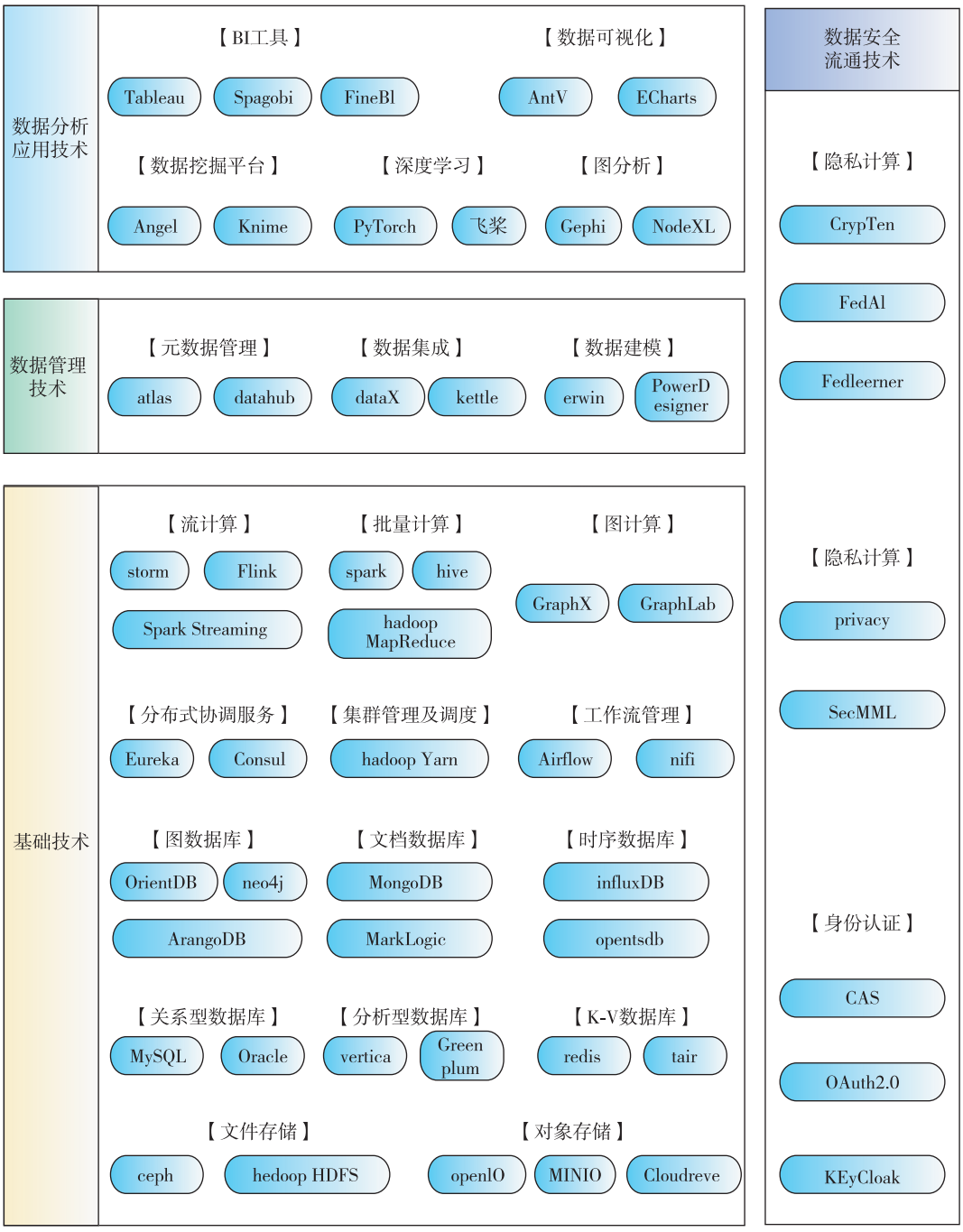


**图 1　大数据技术体系及典型开源软件[6]**

**Fig. 1　Big data technology system and typical open source software[6]**

3）数据分析应用技术挖掘数据价值。为开展数据分析、挖掘数据价值，包括以 BI 工具为代表的统计分析与可视化展现技术，以及以传统机器学习、基于深度神经网络的深度学习为基础的挖掘分析建模技术纷纷涌现，支撑数据价值的挖掘并进一步将分析结果与模型应用于实际业务场景中。

4）数据安全流通技术助力安全合规的数据使用及共享。随着数据价值得到挖掘，数据安全问题也愈发凸显，数据泄露、数据丢失、数据滥用等安全事件层出不穷，如何应对大数据时代下的数据安全威胁，在安全合规的前提下使用及共享数据成为了备受瞩目的问题。访问控制、身份识别、数据加密、数据脱敏、隐私计算等数据保护技术正积极向更加适应大数据场景的方向不断发展。

## 3　面向两类典型数据处理场景的技术架构

大数据处理技术可以分为批处理和流处理两大类。

数据批处理通常处理 $T+1$ 数据，用来支撑以“看”为主的数据应用。批处理非常适合对分布式数据仓库中的历史数据进行分析和计算[9]，例如在计算总数和平均数时，必须将数据集作为一个整体加以处理，而不能将其视作多条记录的集合。这些操作要求在计算进行过程中数据维持自己的状态。数据处理耗时与数据量呈正相关，因此批处理不适合对处理时间要求较高的场合。数据批处理平台通常和 Hadoop、Hive、数据仓库、ETL、维度建模、数据公共层等联系在一起，其典型技术架构如图 2[10]所示。

数据流式处理平台的数据即时处理能力可以达到秒级甚至毫秒级延迟，可以支撑实时化、在线化的数据分析与展现类应用。流处理系统可以处理几乎无限量的数据，但同一时间只能处理一条（真正的流处理）或很少量（微批处理）数据，不同记录间只维持最少量的状态。流式处理非常适合某些类型的工作负载，有近实时处理需求的任务。如分析服务器或应用程序错误日志，以及其他基于时间的衡量指标等。数据流式处理平台的支撑技术主要包含 4 个方面：实时数据采集（如 Flume）、消息中间件（如 Kafka）、流计算框架（如 Storm、Spark、Flink 和 Beam 等）以及实时数据存储（如列族存储的 HBase）。目前主流的实时数据平台也都基于这 4 个方面相关的技术搭建，其典型技术架构如图 3[10]所示。

## 4 大数据技术在中国航天的应用展望

通过在中国航天应用大数据技术与大数据治理理念，形成“全局数据互联，全程业务感知，全域决策智能”的大数据汇聚与分析能力。基于统一数据管理纲领及数据治理工作体系，制定航天型号研制及经营管控各环节的数据标准，保证全局数据模型清晰可控；建成航天特色数据资产全景图，形成数据资产交换、共享、流通模式，构建大数据协同创新体系；打造航天特色全域数据湖，实现全域数据入湖且入湖数据清洁、透明、安全，有效突破航天型号研制及经营管控各环节数据壁垒；依托航天特色全域数据湖，实时感知、全局分析航天型号研制过程及企业经营状况，实现数据支撑决策、数据优化流程；实现全域数据均能按标准实现标准化、规范化采存管理，完成全域数据治理，数据能为科研生产、经营管控工作提供支撑。

为深入剖析大数据技术在中国航天数据治理中的潜在应用场景和价值，本文选取建模仿真这一高度依赖模型，同时又产生大量数据的领域[11]作为典型代表进行分析。某航天研究所积累了同

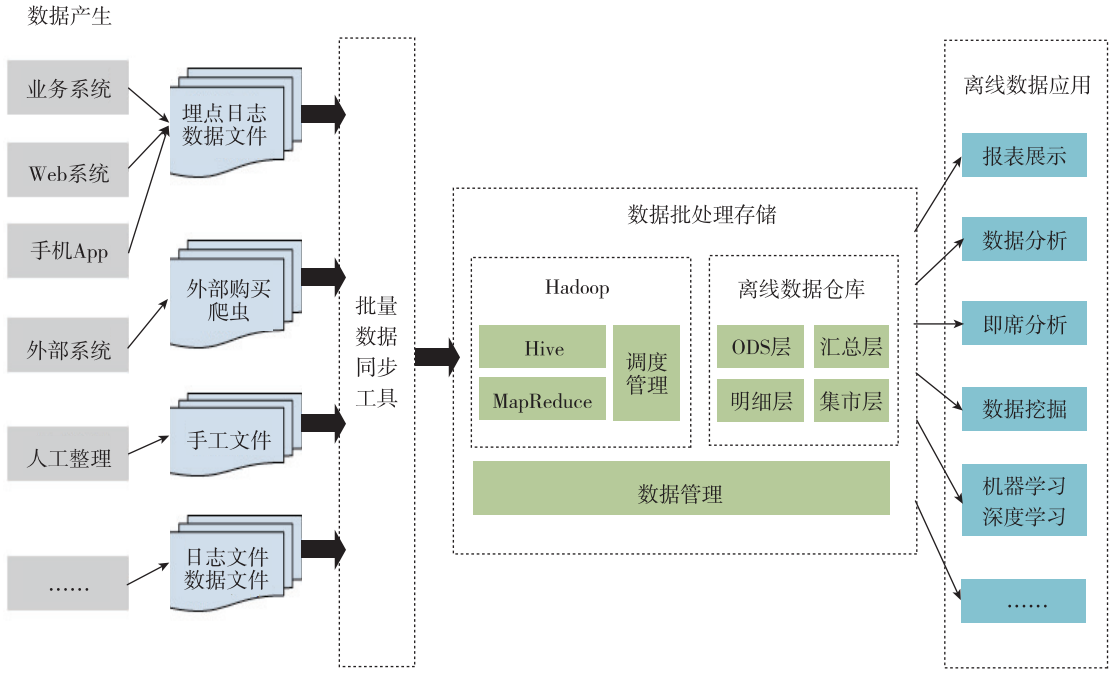


**图 2 面向数据批处理的技术架构设计[10]**

**Fig. 2 Technical architecture design for data batch processing[10]**

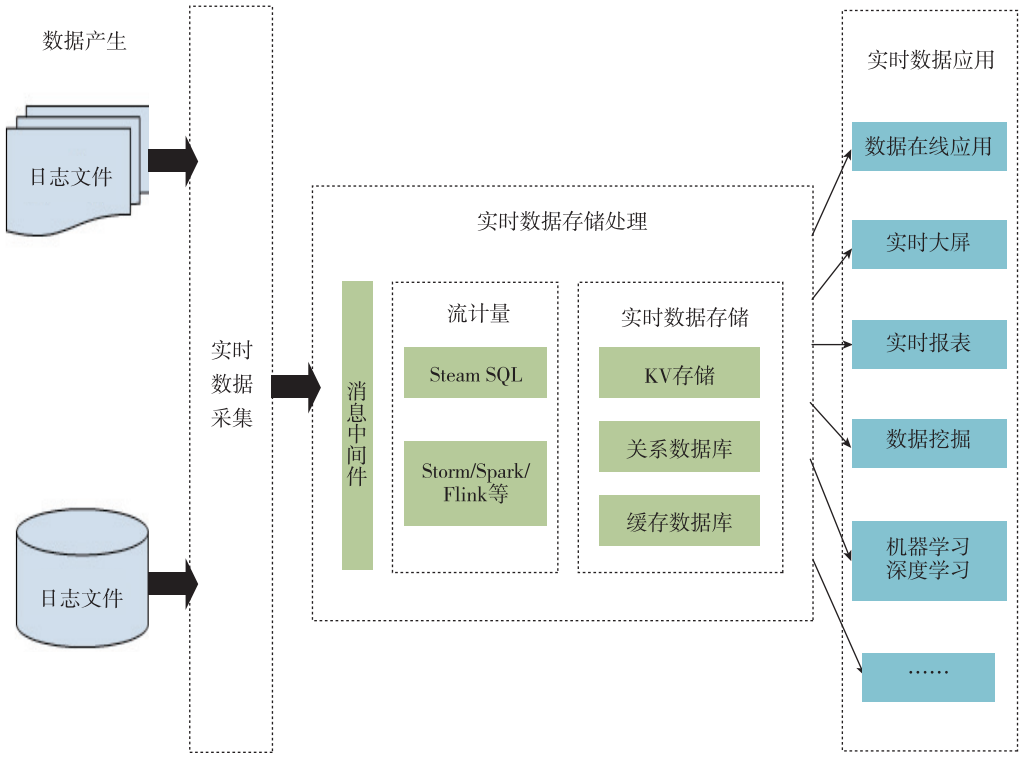


**图 3　面向数据流式处理的技术架构设计**

**Fig. 3　Technical architecture design for data stream processing**

类型相似型号的大量物理真实实测试验条件与实测性能数据，通过应用大数据分析技术，一方面对运载器飞行中的遥测参数值、变化趋势以及关联参数间表征的状态是否一致等展开分析，另一方面建立基于真实实测数据的产品测试性能预测模型，在理论仿真与物理试验测试之间，扩展一条新的性能预测方法，既提高性能测试试验效率，又能提高仿真模拟计算的预测精度。基于型号产品试验时序大数据，计算各项试验参数和飞行器状态参数之间的关系，完成对不同机器学习算法模型预测效果的分析，针对每类试验参数优选出预测精确最高的拟合训练模型，以支撑在不同试验场景中对飞行器多状态参数综合预测评估。

在这一案例中，通过应用大数据分析技术，基于试验时序大数据，实现部件技术状态与遥测参数之间的联系以及遥测参数之间相关性构建，实现基于虚拟试验的产品状态预测评估，完善了试验评估的技术手段，提升仿真准确率，协助缩减重复性的高耗资物理试验，节约成本。

## 5　结论

在中国航天多年的复杂型号产品研制过程中，沉淀了大量数据资产，同时，随着产品数字化水平、数字化生产水平的大幅攀升，大量鲜活研制数据源源不断产生，应用大数据技术激活数据资产、发掘数据价值的条件已然成熟。本文全面梳理了大数据的基本特征、大数据技术体系，并面向批、流两类数据处理场景归纳了典型技术架构，结合建模仿真场景，展望了大数据技术在计算及建模仿真领域的应用前景，为数据治理工作打下坚实基础。

## 参考文献

[1] 林子雨. 大数据技术原理与应用[M].北京:人民邮电出版社,2017.

[2] 王璟璇,窦悦,黄倩倩,等.全国一体化大数据中心引领下超大规模数据要素市场的体系架构与推进路径[J]. 电子政务, 2021(6): 20-28.

[3] 董西成.大数据技术体系详解:原理、架构与实践[M].北京:机械工业出版社,2018.

[4] 梅宏. 数据治理之论[M].北京:中国人民大学出版社,2020.

[5] 关于构建更加完善的要素市场化配置体制机制的意见[J].中国产经,2020(8):1-4.

[6] 中国信息通信研究院. 大数据白皮书(2020 年)[R]. 2020.

[7] 谢朝阳.大数据:规划、实施、运维[M].北京:电子工业出版社,2018.
[8] 华为公司数据管理部.华为数据之道[M].北京:机械工业出版社, 2020.
[9] 张旭,戴丽,闾赛华,等.数据中台架构:企业数据化最佳实践[M].北京: 电子工业出版社,2020.
[10] 朱松岭. 离线和实时大数据开发实战[M].北京:机械工业出版社,2018.
[11] 毕长剑.大数据时代建模与仿真面临的挑战[J].计算机仿真,2014,31(1):1-3+17.

**引用格式**: 赵鹏,朱祎兰. 大数据技术综述与发展展望[J].宇航总体技术,2022,6(1):55-60.

**Citation**: Zhao P, Zhu Y L. Overview and development prospects of big data technology [J]. Astronautical Systems Engineering Technology, 2022, 6(1): 55-60.