Chapter 1

What is Big Data

Text A

Big data is being generated by everything around us at all times. Every digital process and social media exchange produces it. Systems, sensors and mobile devices transmit it. Big data is arriving from multiple sources at an alarming velocity, volume and variety. To extract meaningful value from big data, you need optimal processing power, analytics capabilities and skills.

Big data is a relative term describing a situation where the volume, velocity and variety of data exceed an organization’s storage or compute capacity for accurate and timely decision making.

Some of this data is held in transactional data stores — the byproduct of fast-growing online activity. Machine-to-machine interactions, such as metering, call detail records, environmental sensing and RFID systems, generate their own tidal waves of data. All these forms of data are expanding, and that is coupled with fast-growing streams of unstructured and semi structured data from social media.

However, big data is defined less by volume — which is a constantly moving target — than by its ever-increasing variety, velocity, variability and complexity.

Variety. Up to 85 percent of an organization’s data is unstructured — not numeric — but it still must be folded into quantitative analysis and decision making. Text, video, audio and

New Words and Expressions

sensor/ˈsensə(r)/ *n.*
传感器

transmit/trænsˈmɪt/ *v.*
播送，发射，传送（信号）

velocity/vəˈlɒsəti/ *n.*
速度；速率

extract/ɪkˈstrækt/ *v.*
提取

optimal/ˈɒptɪməl/ *adj.*
最优的，最佳的；优化的

analytics/ˌænəˈlɪtɪks/ *n.*
分析，逻辑分析的方法

exceed/ɪkˈsi:d/ *v.*
超过，胜过

transactional/trænˈzækʃənəl/ *adj.*
交易的，业务的

metering/ˈmi:tərɪŋ/ *n.*
测量（法），测定

tidal /ˈtaɪdl/ *adj.*
潮汐的，潮水的

numeric /nju:ˈmerɪk/ *adj.*
数字的，数值的

quantitative/ˈkwɒntɪtətɪv/ *adj.*
定量的，数量（上）的

other unstructured data require different architecture and technologies for analysis.

Velocity. Thornton May says, "Initiatives such as the use of RFID tags and smart metering are driving an ever greater need to deal with the torrent of data in near-real time. This, coupled with the need and drive to be more agile and deliver insight quicker, is putting tremendous pressure on organizations to build the necessary infrastructure and skill base to react quickly enough."

Variability. In addition to the speed at which data comes your way, the data flows can be highly variable — with daily, seasonal and event-triggered peak loads that can be challenging to manage.

Complexity. Difficulties dealing with data increase with the expanding universe of data sources and are compounded by the need to link, match and transform data across business entities and systems. Organizations need to understand relationships, such as complex hierarchies and data linkages, among all data.

A data environment can become extreme along any of the above dimensions or with a combination of two or all of them at once. However, it is important to understand that not all of your data will be relevant or useful. Organizations must be able to separate the wheat from the chaff and focus on the information that counts — not on the information overload.

What is changing in the realm of big data?

Big data is changing the way people within organizations work together. It is creating a culture in which business and IT leaders must join forces to realize value from all data. Insights from big data can enable all employees to make better decisions — deepening customer engagement, optimizing operations, preventing threats and fraud, and capitalizing on new sources of revenue. But escalating demand for insights requires a fundamentally new approach to architecture, tools and practices.

Competitive advantage: Data is emerging as the world's newest resource for competitive advantage.

Decision making: Decision making is moving from the elite few to the empowered many.

Value of data: As the value of data continues to grow, current systems won't keep pace.

New Words and Expressions

torrent/ˈtɒrənt/ *n.*
奔流

agile/ˈædʒaɪl/ *adj.*
灵活的，机敏的

peak loads
峰值负荷

entity/ˈentəti/ *n.*
实体

hierarchy/ˈhaɪərɑːki/ *n.*
[计]分层，层次，等级制度

linkage/ˈlɪŋkɪdʒ/ *n.*
联系，连接

separate the wheat from the chaff
分清良莠

realm/relm/ *n.*
领域，范围

optimize/ˈɒptɪmaɪz/ *v.*
优化，完善

fraud/frɔːd/ *n.*
诈骗（罪）

revenue/ˈrevənjuː/ *n.*
（公司的）收益，（政府的）税收

escalate/ˈeskəleɪt/ *v.*
（使）增强，（使）扩大

elite/iˈliːt/ *n.*
掌权人物，精英

empower/ɪmˈpaʊər/ *v.*
给（某人）做…的权力，授权

How can you realize the greatest value from big data?

New skills are needed to fully harness the power of big data. Though courses are being offered to prepare a new generation of big data experts, it will take some time to get them into the workforce. Meanwhile, leading organizations are developing new roles, focusing on key challenges and creating new business models to gain the most from big data.

- *Discover the new role of data scientist*

Gartner finds that by 2015, the demand for data and analytics resources will reach 4.4 million jobs globally, but only one-third of those jobs will be filled. The emerging role of data scientist is meant to fill that skills gap.

- *Be proactive about privacy, security and governance*

While big data can provide significant value, it also presents significant risk. Organizations must be proactive about privacy, security and governance to ensure all data and insights are protected and secure.

- *Create new business models with big data*

From data-driven marketing and ad targeting to the connected car, big data is fueling product innovation and new revenue opportunities for many organizations.

Employ the most effective big data technology

To gain the competitive advantage that big data holds, you need to infuse analytics everywhere, make speed a differentiator, and exploit value in all types of data. This requires an infrastructure that can manage and process exploding volumes of structured and unstructured data — in motion as well as at rest — and protect data privacy and security.

Big data technology

Big data technology must support search, development, governance and analytics services for all data types — from transaction and application data to machine and sensor data to social, image and geospatial data, and more.

- *Systems*

Your infrastructure must capitalize on real-time information flowing through your organization. It must be optimized for analytics to respond dynamically — with automated business processes, better agility and improved economics — to the increasing demands of big data.

New Words and Expressions

harness/ˈhɑ:nɪs/ *vt.*	利用
proactive/ˌprəʊˈæktɪv/ *adj.*	积极主动的，前摄的
governance/ˈgʌvənəns/ *n.*	管理，统治
innovation/ˌɪnəˈveɪʃn/ *n.*	改革，创新
infuse/ɪnˈfju:z/ *vt.*	注入，灌输
differentiator/dɪfəˈrenʃɪeɪtə/ *n.*	区分者，微分器
exploit/ɪkˈsplɔɪt/ *vt.*	开拓，开采
agility /əˈdʒɪlətɪ/ *n.*	敏捷，灵活

- *Privacy*

To protect your reputation and brand, your platform must comprise stringent policies and practices around privacy and data protection, safeguarding all of the data and insights on which your business relies.

- *Governance*

The right platform instills trust, so you can act with confidence. It controls how information is created, shared, cleansed, consolidated, protected, maintained, retired and integrated within your enterprise.

- *Storage*

To achieve economies and efficiencies, you must run certain analytics close to the data, while it is in motion. But for data you elect to store, your infrastructure must embody a defensible disposal strategy that reduces the run rate of storage, legal expense and risk.

- *Security*

As you infuse analytics into your organization, data security becomes more central to your competitive advantage profile. Your infrastructure must have strong security measures built in to guard your organization against internal and external threats.

- *Cloud*

To relieve the pressure that big data is placing on your IT infrastructure, you can host big data and analytics solutions on the cloud. Achieve the scalability, flexibility, expandability and economics that will provide competitive advantage into the future.

New Words and Expressions	
reputation/ˌrepjuˈteɪʃn/ *n.*	名声
comprise/kəmˈpraɪz/ *vt.*	包含，包括
safeguard/ˈseɪfgɑ:d/ *vt.*	防护，保卫
instill /ɪnˈstɪl/ *vt.*	逐渐使某人获得（某种可取的品质）
cleanse/klenz/ *vt.*	净化，清洗
consolidate/kənˈsɒlɪdeɪt/ *vt.*	统一，合并
integrate/ˈɪntɪgreɪt/ *vt.*	使一体化
efficiency/ɪˈfɪʃnsi/ *n.*	功效，效率
defensible/dɪˈfensəbl/ *adj.*	能防御的
scalability/skeɪləˈbɪlɪtɪ/ *n.*	可量测性
flexibility/ˌfleksəˈbɪlətɪ/ *n.*	灵活性
expandability/ɪksˈpændəbɪlɪtɪ/ *n.*	扩展性

Note:

The text is adapted from the website:
http://www.ibm.com/big-data/us/en/.

Comprehension

Blank Filling

1. Big data is being generated by everything around us at all times. Every________ and ___________ produces it.

2. Big data is arriving from multiple sources at an alarming ______, ________ and _______. To extract meaningful value from big data, you need ___________ power, ________capabilities and skills.

3. Big data is a relative term describing a situation where the volume, velocity and variety of data exceed an organization's _______ or _______ capacity for accurate and timely _______________.
4. Insights from big data can enable all employees to make better decisions — deepening ___________, optimizing _______, preventing__________, and capitalizing on new sources of________.
5. Meanwhile, leading organizations are developing _______, focusing on key _______ and creating new _________ to gain the most from big data.
6. To gain the competitive advantage that big data holds, you need to infuse _______ everywhere, make speed a differentiator, and exploit ________ in all types of data.
7. Big data technology must support _______, _________, _________ and _________ services for all data types — from ___________ data to __________machine and sensor data to ________________ data, and more.
8. To relieve the pressure that big data is placing on your IT infrastructure, you can host big data and analytics solutions on the _______.
9. IBM data scientists break big data into four dimensions:__________________________.

Content Questions

1. What is the definition of big data?
2. What are the characteristics of big data?
3. What is the background of big data?
4. What does the big data technology do?
5. What is the value of digging big data?

Answers

Blank Filling

1. digital process; social media exchange
2. velocity; volume; variety; optimal processing; analytics
3. storage; compute; decision making
4. customer engagement; operations; threats and fraud; revenue
5. new roles; challenges; business models
6. analytics; value
7. search; development; governance; analytics; transaction and application; social; image and geospatial
8. cloud
9. volume, variety, velocity and veracity

Content Questions

1. "Big data" is a massive, high-growth and diversified information asset that requires a new processing model which have greater decision-making power, insight into discovery

and process optimization capabilities.

2. Volume, Variety, Velocity, Value.
3. Big data is being generated by everything around us at all times. Every digital process and social media exchange produces it. Systems, sensors and mobile devices transmit it. Big data is arriving from multiple sources at an alarming velocity, volume and variety.
4. Big data technology must support search, development, governance and analytics services for all data types — from transaction and application data to machine and sensor data to social, image and geospatial data, and more.
5. The value of digging big data is similar to sandy gold rush, digging sparse but valuable information from massive amounts of data.

参考译文

我们身边的一切每时每刻都在产生大量的数据。每个数字流程和社交媒体的互动都会产生数据。这些数据通过系统、传感器和移动设备进行传输。大数据源于众多数据源，其产生速度、数据量和多样性都迅速增长。要从大数据中提取有意义的信息价值，需要最优的处理能力、分析能力和技术。

大数据是描述数据量、数据获得速度和数据多样性的名词术语，大数据受限于硬件设施，从而使一些公司在存储空间、计算资源方面不能提供准确、实时的分析结果。

其中一些数据存储在基于事务模型的数据库中——这是快速增长的在线活动的副产品。机器对机器的交互，如计量、通话细节记录、环境感测和 RFID 系统，产生自己的大量数据。所有这些形式的数据正在快速增长，同时，来自社交媒体的非结构化和半结构化数据也在飞速增长。

然而，大数据并不定义为“量”，而指的是不断增加的数据种类、数据产生速度、数据多样性和复杂性。

多样性。组织中的数据有高达 85%的部分是非结构化的（非数字形式），但其必须转化为数字形式，以用于定量分析和决策。文本、视频、音频和其他非结构化数据需要不同的架构和技术进行分析。

高速。Thornton May 指出：“射频识别（RFID）和智能计量等新技术的使用，正推动人们越来越迫切地需要实时地处理海量数据，并更加敏捷快速地给出见解，这给公司带来巨大的压力，必须建立必要的基础设施和技能库，以迅速作出反应。”

变化性。除了数据传输的速度之外，数据流可能是高度可变的，日常的、季节性的和事件触发的峰值负载都可能对管理带来挑战。

复杂性。随着数据源不断增多，处理数据的难度越来越大，需要在业务实体和系统之间链接、匹配和转换数据。组织需要了解所有数据之间的关系，例如复杂的层次结构和数据链接。

数据环境可以在上述任何方面变得极端，更不用说上述几方面还可能组合出现。但是，重要的是要了解并不是所有的数据都是相关的或有用的。为了应对信息过载，组织需要具备筛选重点信息并关注有效信息的能力。

大数据领域的变化是什么？

大数据正在改变组织内部人员的合作方式。它正在创造一种文化，企业和IT领导者必须共同努力，使大数据的价值得以体现。来自大数据的分析结论可以使所有员工做出更好的决策——深化客户参与，优化运营，防止威胁和欺诈，以及探索新的收入来源。但是，由于洞察力需求的不断增长，需要一种全新的方法来构建、使用和实践。

竞争优势：数据正在成为世界上最新的竞争优势资源。

决策过程：决策正在从精英阶层转向被赋予权力的许多人。

数据价值：随着数据价值的不断增长，目前的系统将不能保持同步。

如何从大数据中获得最大价值？

大数据的能量需要新的技术来发掘，虽然一些课程正在培养新一代大数据专家，但需要一段时间他们才能工作。同时，领先企业正在发挥新的作用，重点关注重大挑战，创造新的商业模式，从大数据中获取最大收益。

- 发现数据科学家的新角色

Gartner公司发现，到2015年，对数据和分析资源的需求将在全球创造440万个工作岗位，但只有三分之一的岗位得到落实。数据科学家这一新兴角色意在填补这一技能缺口。

- 积极主动关注隐私、安全和管理

虽然大数据可以提供重要的价值，但也存在重大风险。公司机构必须积极主动地了解隐私、安全和管理，以确保所有数据和分析得到妥善保护。

- 使用大数据创建新业务模式

从数据驱动的营销、广告定向投放到联网汽车，大数据推动了许多公司的产品创新和新的收入机会。

采用最有效的大数据技术

为了获得大数据所具有的竞争优势，人们可以在任何客户端上输入分析数据，使速度成为一个产生区别的主要因素，并深度挖掘不同类型数据的价值。因此必须设计一个完善的基础架构，可以管理和处理以指数级增长的结构化和非结构化的数据量（包括静态数据与动态数据），同时保护数据的隐私和安全。

大数据技术

大数据技术必须支持所有数据类型的搜索、开发、管理和分析服务，从交易数据、应用程序数据到机器和传感器数据，以及社交化信息、图像和地理空间数据等。

- 系统

大数据基础设施必须利用流经公司组织的实时信息，同时它必须对数据分析进行优化以便动态响应，包括自动化业务流程、高便捷性和高性价比，以满足大数据日益增长的需求。

- 隐私

为了保护公司的声誉和品牌，大数据平台必须包含有关隐私和数据保护的严格策略和机制，保护公司业务所依靠的数据和未来规划。

- 管理

拥有可信度的完善大数据平台可以让用户或企业在使用时更加放心。它控制如何在企

业中创建、共享、清理、整合、保护、维护、删除和集成信息。

- 存储

为了实现经济性和高效性，必须在运行过程中执行与数据关系密切的特定分析。但是，对于用户选择存储的数据，平台的基础架构必须体现出可防范的处置策略，从而减少运行存储系统的费用、法律费用和风险。

- 安全

当企业将分析数据上传到大数据平台时，数据安全将成为企业竞争优势的核心。大数据的基础架构必须具有强大的安全措施，以保护企业免受内部和外部威胁。

- 云

为了减轻大数据在IT基础设施上的压力，可以在云端托管大数据和分析解决方案，以实现可伸缩性、灵活性、可扩展性和经济性，为未来提供竞争优势。

Text B

Big data is increasingly becoming a factor in production, market competitiveness. Cutting-edge analysis technologies are making inroads into all areas of life and changing our day-to-day existence. Sensor technology, biometric identification and the general trend towards a convergence of information and communication technologies are driving the big data movement.

Huge challenges must be overcome if the benefits are to be leveraged effectively. Matters of concern alongside increasing volumes of data, varying data structures and real-time processing include data security, data privacy policies that are in urgent need of reform and the rising quality expectations of the stakeholders.

Using sensors, a multitude of data sets and specific algorithms, automatic predictions could soon be made about particular behavioral tendencies (and not just online) on the basis of simple correlations. The way in which people think about data and data analysis will gradually change as well, in addition to the technological possibilities.

Big data is more than just IT

Many decision-makers in all kinds of sectors have recognized that big data is no longer purely the preserve of IT. Big data is instead becoming a movement that brings together cutting-edge internet technologies and analysis techniques in order for large, extendable and above all differently structured data sets to be captured, stored and analyzed. This gives big data a broad, international dimension with different knowledge-based outcomes

New Words and Expressions

cutting-edge
前沿的

inroad /ˈɪnrəʊd/ *n.*
进展

convergence/kənˈvɜ:dʒəns/ *n.*
会聚，聚集，收敛

dimension/daɪˈmenʃn/ *n.*
[数]次元，度，维

and expectations with regard to increasing growth and efficiency. But above all, big data provides scope for experimentation, innovation and creativity, offers a wealth of potential new data combinations and is therefore ideal for discovering unexpected correlations. It could be used to create new business models, products and services and to drive innovation.

The information management big data and analytics capabilities include:

Data Management & Warehouse: Gain industry-leading database performance across multiple workloads while lowering administration, storage, development and server costs; Realize extreme speed with capabilities optimized for analytics workloads such as deep analytics, and benefit from workload-optimized systems that can be up and running in hours.

Hadoop System: Bring the power of Apache Hadoop to the enterprise with application accelerators, analytics, visualization, development tools, performance and security features.

Stream Computing: Efficiently deliver real-time analytic processing on constantly changing data in motion and enable descriptive and predictive analytics to support real-time decisions. Capture and analyze all data, all the time, just in time. With stream computing, store less, analyze more and make better decisions faster.

Content Management: Enable comprehensive content lifecycle and document management with cost-effective control of existing and new types of content with scale, security and stability.

Information Integration & Governance: Build confidence in big data with the ability to integrate, understand, manage and govern data appropriately across its lifecycle.

New Words and Expressions

Stream Computing
流计算

Note:

The text is adapted from the website:
https://www-01.ibm.com/software/data/bigdata/.

参考译文

大数据越来越成为影响生产和市场竞争力的因素。先进的分析技术正在进入生活的各个方面，改变我们的日常生活。传感器技术、生物识别和信息通信技术融合的趋势正在驱动大数据快速发展。

要有效利用这些好处，必须克服巨大的挑战。大数据亟待解决的重要问题包括数据量的不断增长、数据结构的不断变化和需求的实时处理、数据的安全性、迫切需要改革的数据隐私策略以及利益相关者不断提高的期望值。

使用传感器、多种数据集和特定算法可以在简单的相关性基础上就特定的行为倾向（而不只是在线）做出自动预测。除了技术可能性之外，人们对数据和数据分析的思考方式也将逐渐改变。

各行业的利益相关者：大数据不仅仅是 IT

各行业的决策者都认识到，大数据不再仅仅是 IT 部门独占的领域。相反，大数据将汇集先进的互联网技术和分析技术，以便进行对大型的、可扩展的和各种不同结构的数据集的捕捉、存储和分析。这为大数据提供了广泛的国际应用空间、不同的知识成果和对效率日益增长的期望。但最重要的是，大数据为实验、创新和创造力的发展空间，提供了大量潜在的新数据组合，因此大数据分析是发现数据之间意外相关性的最佳方式。它可以用于创建新的商业模式、产品和服务，并能推动创新。

信息管理大数据和分析功能包括以下方面。

数据管理和仓库：在降低管理、存储、开发和服务器成本的同时，大数据可在多个工作负载下获得行业领先的数据库性能；通过对分析工作负载（如深度分析）进行优化的功能，实现极高的速度，并可从数小时内启动和运行的工作负载优化系统中获益。

Hadoop 系统：通过应用加速器、分析、可视化、开发工具、性能和安全功能，将 Apache Hadoop 的强大功能带入企业。

流计算：有效地为不断变化的运动数据提供实时分析处理，并支持描述性和预测性分析，以支持实时决策，无时无刻地捕捉并分析所有的数据。使用流计算，可以减少存储空间，做更多的分析，更快地做出更好的决策。

内容管理：通过规模、安全性和稳定性，对现有和新类型的内容进行成本效益的控制，实现全面的内容生命周期和文档管理。

信息集成和治理：建立对大数据的信心，并在整个周期中适当地集成、推断、管理和支配数据。