# 姓名:刘昕宸

# 学号:2015214450

# 教学班号:0002

# 序号:60

# Big Data

Big data is a general term used to describe the voluminous amount of unstructred and semistructed data a company creates—data that would take too much time and cost too much money to load into a lational database for analysis. Although big data doesn’t refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data.

A primary goal for looking at big data is to discover repeatable business patterns. It’s generally accepted that unstructred data, most of which is located in text files, accounts for at least 80% of an organization’s data. If left unmanaged, the sheer volume of unstructured data that’s generatedeach year within an enterprise can be costly in terms of storage. Unmanaged data can also pose a liability if information cannot be located in the event of a computing audit or lawsuit.

Big data analytics is often associated with cloud computing because the analysis of large data sets in real-time requires a framework like MapReduce to distribute the work among tens, hundreds or even thousands of computers.

Big data analytics is the process of examining large amounts of data of a variety of types to uncover hidden patterns, unknown correlations and other useful information. Such information can provide competitive advantages over rival organizations and result in business benefits, such as more effective marketing and increased revenue.

The primary goal of big data analytics is to help companies make better business decisions by enabling data scientists and other users to analyse huge volumes of translation data as well as other data sources that may be left untapped by conventional business intelligence(BI) programs. These other data sources may include Web server logs and Internet clickstream data, social media activity reports, mobile-phone call detail records and information captures by sensors. Some people exclusively associate big data and big data analytics with unstructred data of that sort, but consulting firms like Garther Inc. and Forrester Research Inc. also consider translations and other structured data to be valid forms of big data.

Big data analytics can be done with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics and data warehouses.Furthermore, traditional data warehouses may not be able to handle the processing demands posed by big data.As a result, a new class of big data techology has emerged and is being used in many big data analytics environments. The techologies form the core of an open source software framework that supports the processing of large data sets across clustered systems.

Potential pitfalls that can trip up organizations on big data analytics initiatives include a lack of internal analytics skills and the high cost of hiring experienced analytics professionals, plus challenges in integrating Hadoop systems and data warehouses, although vendors are starting to offer software connectors between those technologies.

Big data management is the organization, administration and governance of large volumes of both structured and unstructured data.

The goal of big data management is to ensure a high level of data quality and accessibility for business intelligence and big data analytics applications.Corporations, government agencies and other organizatoins employ big data management strategies to help them contend with fast-growing pools of data,typically involving many terabytes or even petabytes of information saved in a variety of file formats. Effective big data management helps companies locate valuable information in large sets of unstructured data and semi-structured data from a variety of sources, including call detail records, system logs and social media sites.

Most big data environments go beyond relational databases and tradtional data warehouses platforms to incorporate techologies that are suited to processing and storing nontransactional forms of data. The increasing focus on collecting and analyzing big data is shaping new platforms that combine the traditional data warehouses with big data systems in a logical data warehousing architecture. As part of the process, they must decide what data must be kept for compliance reasons, what data can be disposed of and what data should be kept and analyzed in order to improve current business process or provide a business with a competitive advantage. The process requires careful data classification so that ultimately, smaller sets of data can be analyzed quickly and productively.

Big data as a service (BdaaS) is the delivery of statistical analysis tools or information by an outside provider that helps organizations understand and use insights gained from large information sets in order to gain a competitive advantage.

Given the immense amount of unstructured data generated on a regular basis, BdaaS is intended to free up organization resources by taking advantage of the predictive analytics skills of an outside provider to manage and assess large data sets, rather than hiring in-house staff for those functions. It can take the form of software that assists with data processing or a contract for the services of a team of data sctentists.

BdaaS is a form of manages services, similar to Software as a Service or Infratructure as a Service.It often relies upon cloud storage to preserve continual data access for the organization that owns the information as well as the provider working with it.

Unstructured data is a generic label for describing any corporate information that is not in a database. Unstructured data can be textual or non-textual. Textual unstructured data is generated in media like E-mail messages, PowerPoint presentation, Word documents, collaboration software and instant messages. Non-textual unstructured data is generated in media like JPEG images, MP3 audio files and Flash video files.

The information contained in unstructured data is not always easy to locate. It requires that data in both electronic and hard copy documents and other media be scanned so a search application can parse out concepts based on words used in specific contexts.This is called semantic search.It is also referred to as enterprise search.

In custom-facing business, the information contained in unstructured data can be analyzed to improve customer relationship management and relationship marketing. As social media application like Twitter and Facebook go mainstream, the growth of unstructured data is expected to far outplace the growth of structured data.

Generally ,data mining (sometimes called data or knoeledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information—information that can be used to increase revenue, cut costs, or both. Data mining software is one of a number of analytical tools for analytical tools for analyzing data. It allows users to analyze data from many different dimensionsor angels, categorizes it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Data mining parameters include:

Association—looking for patterns where one event is connected to another event.

Sequence or path analysis—looking for patterns where one event leads to another later event.

Classification—looking for new patterns(may result in a charge in the way the data is organized but that’s OK).

Clustering—finding and visually documenting groups of facts not previously known.

Forecasting—discovering patterns in data that can lead to reasonable predictions about the future(This area of mining is known as predictive analytics).

Data mining techniques are used in many reaearch areas, including mathematics, cybernetics, genetics and marketing. Web mining, a type of data mining used in customer relationship management(CRM), takes advaneage of the huge amount of information gathered by a Web site to look for patterns in user behavior.

A data warehouse is a central repository for all or significiant parts of the data that an enterpeise’s various business systems collect. The term was coined by W.H.Inmon. IBM sometimes uses the term “information warehouse”.

Typically, a data warehouse is housed on an enterprise mainframe server. Data from various online transaction processing(OLTP) application and other sources is selectively extracted and organized on the data warehouse database for the use by analytical applications and user queries.Data warehousing emphasizes the capture of data from diverse sources for useful analysis and access, but does not generally start from the point-of-view of the end user or knowledge worker who may need access to speciallized, sometimes local databases. The latter idea is known as the data mart.

Application of data warehouses include data mining, Web Mining and decision support systems(DSS).

# 中文翻译

## 大数据

大数据是用来描述公司产生的浩繁的非结构化和半结构化数据的一个通用术语——要把这些数据加载到关系型数据库来分析会耗费大量时间和金钱。虽然大数据并没有涉及任何具体数量，但通常在谈论拍字节和艾字节时使用该术语。

观察大数据的一个主要目的是发现可重复的业务模式。人们普遍承认，非结构化数据，其中大部分在文本文件中，占一个组织中至少有80%的数据。如果不加以管理，则企业每一年产生的全部非结构化数据就会花费巨额的存储费用。如果审计或诉讼时不能找到信息，那么不加管理的非结构化数据也可能带来法律责任。

大数据分析往往是与云计算相关的，因为实时分析大型数据集需要像MapReduce这样的一个框架来将任务分布到数十台、几百台甚至上千台的计算机上。

大数据分析研究大量的多种类型的数据，以揭示隐藏的模式、未知的相互关系及其他有用的信息。这些信息可以提供有竞争力的优势以超过对手组织，产生商业利益，例如更有效地营销和增加收入。

大数据分析的主要目的是，让数据科学家和其他用户分析数据巨大的业务数据以及可能没有被传统的商业智能（BI）程序利用的其他数据源来帮助企业做出更好的业务决策。这些其他的数据源可能包括Web服务器日志和互联网点击流数据、社交媒体活动报告、移动电话的呼叫详细记录和传感器捕获的信息。有些人专门对这类非结构化数据进行大数据分析，而像Gatner公司和Forrester研究公司这样的咨询公司也把业务数据和其他结构化数据当做有效的大数据形式。

大数据分析可以用工具软件常用的部分高级分析等学科的预测分析和数据仓库。此外，传统数据仓库无法处理大数据所带来的需求。因此，一个新的大数据技术类已经出现并被用于许多大数据分析环境。该技术的一个开源软件框架，支持大数据集在集群系统的处理核心。

潜在的陷阱，阻碍组织对大数据分析的举措，包括缺乏内部的分析技能和聘请有经验的分析人员成本高，再加上整合Hadoop系统与数据仓库的挑战，尽管供应商开始提供软件连接器之间的技术。

大数据管理是组织、管理和管理大量的结构化和非结构化数据。

大数据管理的目标是确保高层次的数据质量和可访问性的商业智能和大数据分析应用。公司、政府机构和其他organizatoins聘请大数据管理策略，帮助他们应付快速增长的数据池，通常涉及许多百万兆字节甚至PB的保存在不同的文件格式信息。有效的大数据管理有助于企业在大量的非结构化数据和半结构化数据从各种来源，包括呼叫详细记录，系统日志和社交媒体网站定位有价值的信息。

数据环境中最大的超越关系型数据库和传统的数据仓库平台整合技术，适用于处理和存储数据的非事务性的形式。在收集和分析大数据的重点是塑造新的平台，将传统的数据仓库与大数据系统相结合的逻辑数据仓库体系结构。作为过程的一部分，他们必须决定什么样的数据必须遵守的原因，什么样的数据可以被处理和什么数据应保持和分析，以改善目前的业务流程或提供一个具有竞争优势的业务。这个过程需要细致的数据分类，最终，较小的数据集分析可以快速有效。

作为一个服务大数据（BDaaS）是由外部供应商，可以帮助企业了解和使用大量的信息集，获得了获得竞争优势的见解，统计分析工具和信息的传递。

鉴于在定期的基础上产生的非结构化数据量巨大，BDaaS意在提高组织资源利用的管理和评估大数据集的外部供应商的预测分析能力，而不是雇佣内部员工为这些功能。它可以协助数据处理软件，或一个团队的数据科学家服务合同的形式。

BDaaS是一种管理服务，类似于软件作为一种服务或基础设施建设作为服务。它往往依赖于云存储保持持续的数据访问的组织拥有以及供应商信息的工作。

非结构化数据是用于描述任何不在数据库中的任何公司信息的通用标签。非结构化数据可以是文本的或非文本的。文本的非结构化数据是媒体如电子邮件，生成PowerPoint演示，word文档，协作软件和即时消息。非文本的非结构化数据是媒体如JPEG图像生成，MP3音频文件和视频文件。

在非结构化数据中包含的信息并不总是很容易找到。它要求在电子和硬拷贝文档和其他媒体的数据进行扫描，这样一个搜索应用程序可以解析出在特定上下文中使用的词的概念，这被称为语义搜索，它也被称为企业搜索。

在面向客户的业务中，可以分析非结构化数据中所包含的信息，以提高客户关系管理和关系营销。随着社会化媒体的应用如推特和脸谱网的主流，非结构化数据的增长预期远取代结构化数据的增长。

一般来说，数据挖掘（有时称为数据或知识发现）是从不同的角度分析数据的过程，并总结成有用的信息，可以用来增加收入，降低成本，或两者。数据挖掘软件是分析数据分析工具的一种分析工具。它允许用户从不同的角度来分析维度，数据归类，并总结确定关系。从技术上讲，数据挖掘是在大型关系数据库中的几十个领域中发现相关性或模式的过程。

数据挖掘参数包括：

寻找一个事件连接到另一个事件的模式的关联。

寻找一个事件导致另一个事件的模式的序列或路径分析。

分类寻找新的模式（可能会导致在数据的方式是有组织的收费，但这是确定的）。

聚类发现和视觉记录组的事实，而不是以前所知的。

预测发现模式的数据，可以导致合理的预测未来（这一领域的采矿被称为预测分析）。

数据挖掘技术应用于许多研究领域，包括数学、控制论、遗传学和营销。Web挖掘，一种数据挖掘技术在客户关系管理（CRM），以庞大的信息量advaneage聚集在网站上寻找用户行为模式。

数据仓库是一个为所有或一个企业的各种业务系统收集数据的中央存储库的重要部分。这个词是由W.H.Inmon。IBM有时使用术语“信息库”。

通常情况下，数据仓库位于企业主机服务器上。从各种在线事务处理（OLTP）和其他来源的数据，应用选择性地提取并用分析应用和用户查询使用数据仓库数据库组织数据仓库强调从有用的分析和访问各种数据源的数据捕获，但一般不从的角度对最终用户或知识工作者可能需要访问本地数据库专业，有时。后者的想法被称为数据集市。

数据仓库的应用包括数据挖掘、Web挖掘和决策支持系统（决策支持系统）。