## 大数据管理-概念、技术与挑战

大数据最大的优点是针对传统手段捕捉到 的数据之外的非结构化数据。这意味着不 能保证输入的数据是完整的，清洗过的和 没有任何的错误。这使它更有挑战性，但 同时它提供了在数据中获得更多的洞察力 的范围。

正如Google的首席经济学家Hal Varian所 说¨”，数据是广泛可用的，所缺乏的是从中提取出 知识的能力．数据收集的根本目的是根据需求从数 据中提取有用的知识，并将其应用到具体的领域之 中．不同领域的大数据应用有不同的特点。

整个大数据的处理流程可以定义为在合适工具 的辅助下，对广泛异构的数据源进行抽取和集成，结 果按照一定的标准统一存储．利用合适的数据分析 技术对存储的数据进行分析，从中提取有益的知识 并利用恰当的方式将结果展现给终端用户．具体来说 可以分为数据抽取与集成、数据分析以及数据解释．

．从数据集成模型来看，现有的数据抽取与集 成方式可以大致分为以下4种类型心⋯：基于物化或 ETI。方法的引擎(materialization or ETL engine)、基 于联邦数据库或中间件方法的引擎(federation engine or mediator)、基于数据流方法的引擎(stream engine) 及基于搜索引擎的方法(search engine)．

)数据量大并不一定意味着数据价值的增加， 相反这往往意味着数据噪音的增多．因此在数据分 析之前必须进行数据清洗等预处理工作，但是预处 理如此大量的数据对于机器硬件以及算法都是严峻 的考验．

如果将各种大数据的应用比作一辆辆“汽车”， 支撑起这些“汽车”运行的“高速公路”就是云计算． 正是云计算技术在数据存储、管理与分析等方面的 支撑，才使得大数据有用武之地．

，Google于2006年首先提出了云计算的概念

Bigtable， Dynamo，PNUTS等的成功促使人们开始对关系数 据库进行反思，由此产生了一批未采用关系模型的 数据库，这些方案现在被统一称为NoSQL(not only SQL)．NoSQL并没有一个准确的定义，但一般认为 NoSQL数据库应当具有以下的特征[42I：模式自由 (schema—free)、支持简易备份(easy replication support)、简单的应用程序接口(simple APl)、最终 一致性(或者说支持BASE特性，不支持ACID)、支 持海量数据(huge amount of data)．

综上所述，大数据时代的数据存在着如下几个 特点：多源异构；分布广泛；动态增长；先有数据后有 模式．正是这些与传统数据管理迥然不同的特点，使 得大数据时代的数据管理面临着新的挑战，