第10章 大数据分析

【学习目标】

- 了解大数据与云计算
- 掌握大数据存储特点
- 掌握Hadoop开源框架的基本原理
- 掌握Spark开源框架

简介

- 互联网技术的快速发展导致数据的快速膨胀，谷歌公司每天处理超过25PB的数据。包括卫星图片、航拍照片和360度街景图片。人类身体细胞数据各不相同，但是数量最多的接近100万亿个，如果用1个位来表示1个细胞，那么1PB足够记录90个人的身体细胞。Facebook每天处理亿计的图片，信息。在2012年就存储了超过100PB的数据。
- 麦肯锡表示：“大数据值得是所涉及的数据集规模已经超过了传统数据库软件获取、存储、管理和分析的能力”
- IBM则指出：“如果数据集具有数量上，种类多样性，速度极快的这三种特性”，称之为3V（Volume,Variety,Velocity,即容量、多样性、速度），那么，这些数据就称之为大数据。

10.1云计算与大数据

- 由于海量数据的产生，IT基础设施也随着高速发展，特别是高速存储，高速网络的发展壮大。为了对海量数据进行有效的计算，必须最大限度地利用计算和网络资源。计算虚拟化和网络虚拟化对分布式计算、存储和网络资源都提出了很高的要求。
- 从科学研究到零售业，从保险、医疗到互联网，每个行业都在爆发式地产生数据，这种增长甚至超过了人类创造存储器的速度。各行各业，利用大数据的故事层出不穷。

- 江苏省宿迁市公安局以民意为导向，用大数据思维牵引整体工作布局，强力推进“智慧警务”建设落地见效，探寻出一条“强度整合、高度共享、深度运用”的“智慧警务”新路径。2018年10月13日17时21分，宿迁市洋河新区酒家路北一商铺附近，涉嫌故意伤害致人轻伤的犯罪嫌疑人张某还没来得及反应，就被身边的一群巡防民警制伏。原来“智慧警务”从发现嫌疑人的第一个身影开始，他的影像就接连不断地在PGIS平台上标注。短短几十分钟就锁定了嫌疑人的活动区域，并最终确定了准确位置。2018年 3月13日，泗洪县局指挥中心根据人像对比系统的预警指令，迅速锁定在逃嫌疑人位置。街面巡防警力立即开展围捕工作，仅用4分钟就在汽车站附近将潜逃13年的命案在逃嫌疑人张某抓获。

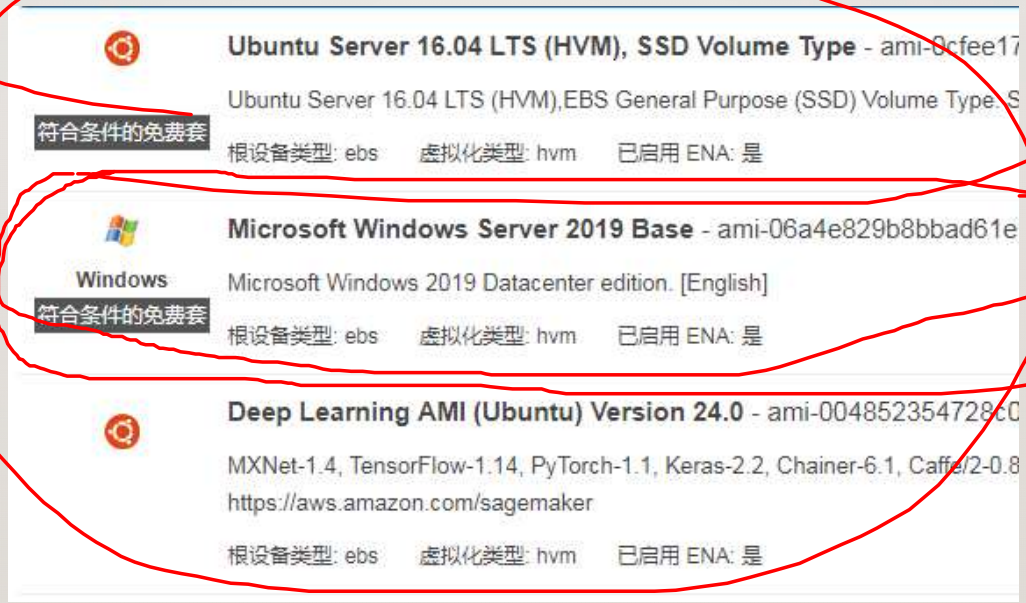

"十一五"国家科技支撑计划项目——全国警用地理信息基础平台应用技术研究与规模应用示范
在线人数：12 访问次数：90106 2014/06/19 欢迎访问!业务超管
全国警用地理信息基础平台
许昌市公安局警用地理信息系统
信息查询
业务应用
数据维护
平台服务
运行管理
数据采集
人口综合分析
实有人口分类统计
实有人口状况统计
碰撞比对
漫游
放大
缩小
全图
测距
测面
鹰眼
导航
打印
清除
人员轨迹图类型
日期段
- 2014-06-19
居住轨迹
黑色
上网轨迹
绿色
住宿轨迹
红色
展示
停止
清除
时间轴
（1）张燕伟:2013-03-25
（4）张燕伟:许昌市开发区日月潭酒店
（2）张燕伟:2014-07-26
矢量地图
影像地图
矢量叠加地图
姓名:张燕伟
人口类别:常住人口
身份证号:411002198604090024
地址:瑞丰苑小区5号楼
【详细信息】
【房屋信息】
【周边查询】
许昌市公安局
许昌县
© 2010 PGIS
系统时间: 2014年06月19日 11:22:10
公安部信息中心 版权所有

- 在贵阳市修文县的猕猴桃如今也插上了科技的翅膀，形成了新的产销业态。猕猴桃已实施二维码追溯体系，手机一扫，就能了解猕猴桃的身份信息。借助二维码防伪标，每个果园有一个二维码身份证。消费者只要用手机扫描一下猕猴桃上的二维码，就可以显示出产地在哪儿、园主是谁、田间管理、采摘时间、出库时间等信息，全程跟踪，全程溯源。与此同时，每次扫描之后，后台都可以进行流向监测。通过分析挖掘扫描数据，可以获得中国南北方市场口感偏好、各大城市的销售情况等信息。

“云计算”（CLOUD COMPUTING）

- 是一种基于互联网的计算方式，通过这种方式，共享的软硬件资源和信息可以按照需求提供给计算机和其他设备。云计算为我们提供了跨地域、高可靠性、按需付费、快速部署的能力。随着云计算的发展，大数据正成为云计算面临的一个重大挑战。
- 用户不需要了解“云”中基础设施的细节，不必具有相应的专业知识。云计算是一种虚拟化的资源，并且意味着计算能力也可作为一种商品通过互联网进行流通。

公有云


Ubuntu Server 16.04 LTS (HVM), SSD Volume Type - ami-0cfee17
Ubuntu Server 16.04 LTS (HVM),EBS General Purpose (SSD) Volume Type. S
符合条件的免费套
根设备类型: ebs 虚拟化类型: hvm 已启用 ENA: 是
Microsoft Windows Server 2019 Base - ami-06a4e829b8bbad61e
Windows
Microsoft Windows 2019 Datacenter edition. [English]
符合条件的免费套
根设备类型: ebs 虚拟化类型: hvm 已启用 ENA: 是
Deep Learning AMI (Ubuntu) Version 24.0 - ami-004852354728c0
MXNet-1.4, TensorFlow-1.14, PyTorch-1.1, Keras-2.2, Chainer-6.1, Caffe/2-0.8
https://aws.amazon.com/sagemaker
根设备类型: ebs 虚拟化类型: hvm 已启用 ENA: 是

私有云

- 往往只针对特定客户群提供服务，比如一个企业内部可以在自己的数据中心搭建私有云。或者是个人以家庭为单位提供家庭成员的云服务。

云计算的服务

- 云计算一般包括三个层次的服务：基础设施即服务（IaaS），平台即服务（PaaS）和软件即服务（SaaS）。这里的分层，指的是在基础设施，系统平台、软件应用上来实现。
- 信息技术的发展主要解决的是云计算中结构化数据的存储、处理和应用。结构化数据的特征是“逻辑性强”。然而，显示社会中大量数据事实上没有“显著”的逻辑关系。如某个时态的交通状况、人流状况等。它的特征是随时、海量与弹性的，如一个图片天气分析会有几百PB的数据。一个社会事件的爆发也是突然的，并且可以产生大量的相关数据（微博、视频、文章）等

云计算和传统计算机

- 传统的计算机设计与软件都是以解决结构化数据为主，对“非结构化”数据的解决有些力不从心。所以，以社交网络、电子商务和移动通信为主的社会将会处理大量的非结构化数据为主。这从很大程度上，需要依赖云计算。

云计算和大数据

- 云计算和大数据是相互依存的。以云计算为基础的信息存储、挖掘手段为知识生产提供了工具，而通过对大数据分析、预测会使得决策更加精准和及时。从另一方面说，云计算是一种理论、技术架构，以处理大数据为主。大数据技术是云计算技术的扩展。
- 大数据为云计算大规模与分布式的计算能力提供了广泛的资源，解决了传统计算机和数据库无法完成的任务

10.2大数据存储的特点

- 大容量及高可扩展性
- 大数据存储主要计算来源包括社交网站、个人信息、科学研究数据、在线事务、系统日志以及传感和遥控数据等。各种应用系统源源不断地产生着大量的数据。尤其是社交网站的兴起，更加快乐数据增长的书读。大数据一般可达到几个PB的信息量，传统的NAS一般很难达到这个级别的存储容量。因此除了巨大的存储容量外，大数据存储还必须拥有一定的可扩容能力。
- 高可用性
- 对于大数据应用和服务来说，数据是其价值所在。因此，存储系统的可用性至关重要。平均无故障时间(MTTF)和平均维修时间(MTTR)是衡量存储系统可用性的两个主要指标。传统存储系统一般采用RAID、数据通道冗余等方式保证数据的高可用性和高可靠性。除了这些传统的技术手段外，大数据存储还会采用其他一些技术。比如，分布式存储系统中多采用简单明了的多副本来实现数据冗余;针对RAID导致的数据冗余率过高或者大容量磁盘的修复时间过长等问题，近年来学术界和工业界研究或采用了其他的编码方式。
- 3.高性能
- 在考量大数据存储性能时，吞吐率、延时和IOPS是其中几个较为重要的指标。对于些实时事务分析系统，存储的响应速度至关重要;而在其他一些大数据应用场景中，每秒处理的事务数则可能是最重要的影响因素。大数据存储系统的设计往往需要在大容高可扩展性、高可用性和高性能等特性间做出一个权衡。
- 4.安全性
- 大数据具有巨大的潜在商业价值，这也是大数据分析和数据挖掘兴起的重要原因之。因此，数据安全对于企业来说至关重要。数据的安全性体现在存储如何保证数据完整性和持久化等方面。在云计算、云存储行业风生水起的大背景下，如何在多租户环境中保护好用户隐私和数据安全成了大数据存储面临的一个亟待解决的新挑战。

10.2大数据存储的特点

- 5.自管理和自修复
- 随着数据量的增加和数据结构的多样化，大数据存储的系统架构也变得更加复杂，管理和维护便成了一大难题。这个问题在分布式存储中尤其突出，因此，能够实现自我管理、监测及自我修复将成为大数据存储系统的重要特性之一。
- 6.成本
- 大数据存储系统的成本包括存储成本、使用成本和维护成本等。如何有效降低单位存储给企业带来的成本问题，在大数据背景下显得极为重要。如果大数据存储的成本降不下来，PB级的数据量将会让很多中小型企业在管理上有非常大的成本支出。
- 7. 访问接口的多样性
- 同一份数据可能会被多个部门、用户或者应用来访问、处理和分析。不同的应用系统由于业务不同可能会采用不同的数据访问方式。因此，大数据存储系统需要提供多种接口来支持不同的应用系统。

10.3大数据应用

- 大数据正在向我国实体经济各领域渗透融合，进入全方位、广渗透的新阶段。融合范围日益宽广。大数据与实体经济的融合正在从部分先导领域如零售、医疗保健、安防等生活服务和公共服务领域向农业、制造业生产、供应领域拓展。目前，金融、汽车、餐饮、物流等各行各业都融入了大数据，应用领域日益丰富全面。融合深度逐步加深。中国信通院调查数据显示 50%的受访企业2017 年对大数据的投入比 2016 年提升 6.7%，其中，25.5%的企业对大数据投入增加超过 50%，32.7%的企业在数据方面的投入增加在 50%以内。