大数据的信息基础设施

云计算

陈一帅

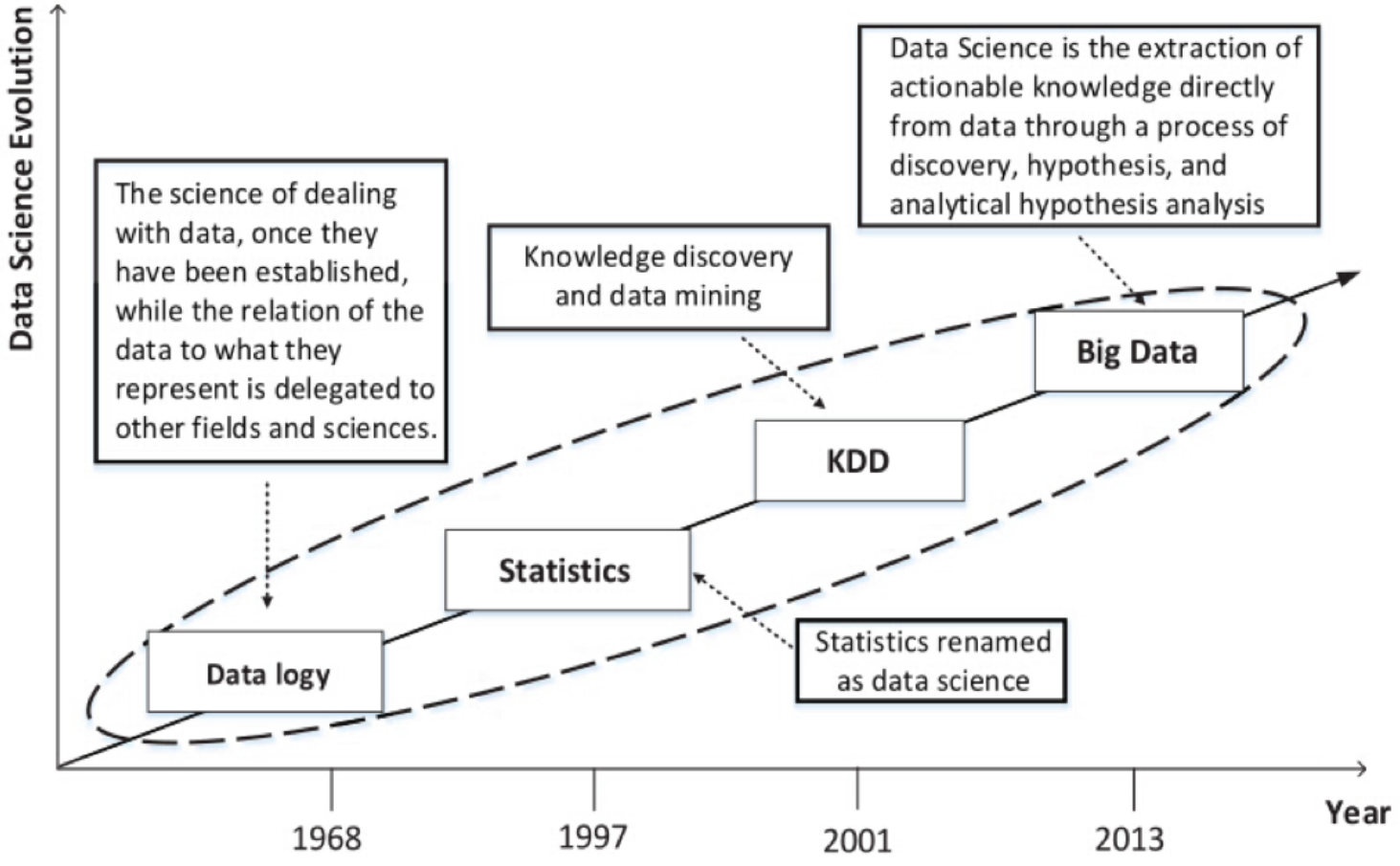
yschen@bjtu.edu.cn

北京交通大学电子信息工程学院

内容

- 云的起源和定义
- 云计算的服务模式
- 云分布式计算的特性

数据科学的演进


Data Science Evolution
The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences.
Knowledge discovery and data mining
Data Science is the extraction of actionable knowledge directly from data through a process of discovery, hypothesis, and analytical hypothesis analysis
Big Data
KDD
Statistics
Data logy
Statistics renamed as data science
1968
1997
2001
2013
Year

云与大数据的关系

- 大数据
  - 数据量大
  - 数据产生速度快
  - 数据维度高、类型多
- 要求底层信息基础设施具有可伸缩性（scalability）
- 云提供这种可伸缩性

云的起源

- 大型网络公司业务规模庞大
  - 腾讯、百度、阿里、谷歌、亚马逊、微软
- 为自己的业务建设了大规模信息基础设施
  - 存储、计算、网络
- 这些信息基础设施利用率不高
  - 亚马逊首席执行官对其计算机数量和利用率低感到惊讶
  - 因为这是为高峰负载（圣诞节）设计的
- 开放它们
  - 2006 年 8 月 25 日，亚马逊宣布 EC2
  - 云的诞生

云的定义

- 一台弹性计算机
  - 按需存储和计算
  - 对象存储以及虚拟机和容器的执行
- 一种模型
  - 对可配置资源（网络，服务器，存储，应用程序和服务）共享池的便捷、按需网络访问
  - 这些资源可以以最少的管理工作量或服务提供商交互，进行快速配置和释放

云计算技术的组成

- 存储
  - 低成本，大容量
- 计算
  - 多核处理器
  - 数据中心，服务器集群
- 网络
  - 100G 以太网
  - 大规模高速 Clos 网络连接
  - 最小化任何规模的延迟

内容

- 背景
- 云的起源和定义
- 云计算的服务模式
- 云分布式计算的特性

云的三种服务模式

- Infrastructure as a service (IaaS)
  - 提供基础设施，如虚拟机，存储和网络
  - 用户可以自己管理，在上面部署自己的操作系统和软件
  - Amazon AWS
- Platform as a service (PaaS)
  - 提供软硬件平台，包括硬件和操作系统，开发和管理工具
  - 用户可以利用这些工具，开发和部署他们的软件
  - Microsoft Azure
- Software as a service (SaaS)
  - 具有用户界面的完整应用程序，用户直接用
  - 如 Salesforce.com

云提供的服务 IaaS

- 以 Amazon AWS 为例
- 计算、存储、数据库、网络服务

Table 2.15: Compute, Storage, Database and Networking Services in AWS Cloud

Category	Offering	Service Modules or Short Description
Compute	EC2	Virtual servers in the AWS cloud
	Lambda	Run code in response to events
	EC2 Container Service	Run and manage Docker containers
Storage & Content Delivery	S3	Scalable storage in the AWS cloud
	Elastic File System	Fully management file system for EC2
	Storage Gateway	Integrate on-premises IT facilities with cloud storage
	Glacier	Archive storage in the AWS cloud
	CloudFront	Global content delivery network
Database	RDS	MySQL, Postgres, Oracle, SQL server
	DynamicDB	Predictable and scalable NoSQL data store
	ElastiCache	In-memory Cache
	Redshift	Managed petabyte-scale warehouse service
Networking	VPC	Virtual pivate cloud as isolated cloud resources
	Direct Connect	Dedicated Network Connection to AWS
	Route S3	Scalable DNS and domain name registration

云提供的服务 PaaS

- 以 Amazon AWS 为例
- 应用、移动、分析服务

Table 2.16: Application, Mobile and Analytics Services in the AWS Cloud

Category	Offering	Service Modules or Short Description
Application Services	SQS	Message queue Services
	SWF	Workflow service for coordinating app components
	AppStream	Low latency application streaming
	Elastic Transcoder	Easy-to-use scalable media transcoding
	SES	Email sending and receiving service
	CloudSearch	Managed search service
	API Gateway	Build, deploy and manage APIs
Mobile Services	Cognito	User identity and app data synchronization
	Device Farm	Test Android, Fire OS, and iOS apps on devices in the cloud
	Mobile Analytics	Collect, view and export app analytics
	SNS	Simple push notification Service
Analytics Services	EMR	Managed elastic Hadoop (MapReduce) framework
	Kinesis	Real-time processing of streaming data
	Data Pipeline	Orchestration for data-driven workflows
	Machine Learning	Build machine learning prediction solutions

云提供的服务 PaaS

Table 2.17 Public Clouds Offering Platform-as-a-Service (PaaS) Services (Aug. 2015)

Cloud Name	Languages and Developer Tools	Programming Models supported by provider	Target Applications and Storage Option
Google AppEngine	Python, Java and Eclipse-based IDE	MapReduce, Web programming on demand	Web applications and BigTable Storage
Salesforce.com Force.com	Apex, Eclipse-based IDE, Web-based Wizard	Workflow, Excel-like, Web programming on demand	CRM and add-on App development for Business
Microsoft Azure	.NET, Azure tools for MS Visual Studio	Dryad, Twister, .NET Framework	Enterprise and Web Applications
Amazon Elastic MapReduce	Hive, Pig, Cascading, Java, Ruby, Perl, Python, PHP, R, and C++	MapReduce, Hadoop, Spark,	Data Processing, eMail, and e-Commerce, S3 and WorkDocs

云提供的服务 EMR

- 大数据 EMR（E-MapReduce）PaaS 服务
  - 计算集群（Cluster）
- 节点类型
  - 管理节点：协调数据和任务的分布
  - 核心节点：运行任务，存储数据
  - 任务节点：运行任务
- 安装和配置了各种应用
  - Hadoop MapReduce，YARN（资源管理和分配），HDFS

云提供的服务 SaaS

- Salesforce CRM（客户关系管理）SaaS 服务
  - 销售云：管理客户资料，跟踪商机，优化活动
  - 服务云：创建，跟踪和路由服务案例，包括社交媒体网络服务
  - 市场云：社交营销，从社交媒体中识别销售线索，发现粉丝
  - 数据云：获取和管理 CRM 记录
  - 协作云：业务协作
  - 分析云：基于机器学习的销售绩效分析
  - 定制云：在标准 CRM 应用程序之上创建附加应用程序

部署私有云

- VMware vSphere 数据中心虚拟化套件
- OpenStack 开源云计算系统

内容

- 背景
- 云的起源和定义
- 云计算的服务模式
- 云分布式计算的特性

分布式系统 CAP 特性

- 在包含多台计算机的分布式系统中，有三个可追求的特性
- 一致性（Consistency）
  - 所有计算机同时看到相同的数据
- 可用性（Availability）
  - 每个请求都收到有关是否请求的响应成功或失败
- 分区容限（Partition tolerance）
  - 即使网络故障阻止计算机进行通信，系统仍可继续运行

CAP 定理

- 定理
  - 无法创建具有所有 CAP 三个属性的分布式系统
- 含义
  - 无法在可用性和分区容忍度的同时实现严格的一致性
- 挑战
  - 随着计算机数量增长，系统单元出现故障可能性随之增加
  - 设计人员必须为特定系统选择高一致性还是高可用性

一致性 vs. 可用性

- 对可用性和一致性的选择取决于业务需求
- 例：在电子商务环境中
- 添加购物车
  - 可选择高可用性，保证向购物车添加商品的请求一定成功
  - 可向客户隐藏错误，并在以后进行处理
- 订单提交
  - 可选择高一致性
  - 因为几种服务（信用卡处理，运输和处理，报告）需要同时访问数据

小结

- 背景
- 云的起源和定义
- 云计算的服务模式
- 云分布式计算的特性

练习

- 调研国内外云平台，申请免费学生/试用账号
  - Azure: https://signup.azure.com
  - AWS: https://aws.amazon.com/cn/free/
  - Aliyun: https://free.aliyun.com/
  - 华为云: https://activity.huaweicloud.com/free_test/index.html
  - 腾讯云: https://cloud.tencent.com/act/free
- 调研它们提供的各种服务，指出其服务模式
  - 特别调研大数据服务，指出其服务模式

练习

- 完成报告
  - 【腾讯文档】Lab1、大数据平台调研
  - https://docs.qq.com/doc/DT3BGSFZDQWRQWnRr