Turing AI 万维图灵 | 大数据系列课程

大数据技术-第一章：Hadoop大数据概述

Hadoop与云计算、Spark

CONTENTS

01. 云计算概念

02 云计算特点

03. Hadoop与云计算关系

04. Spark概念

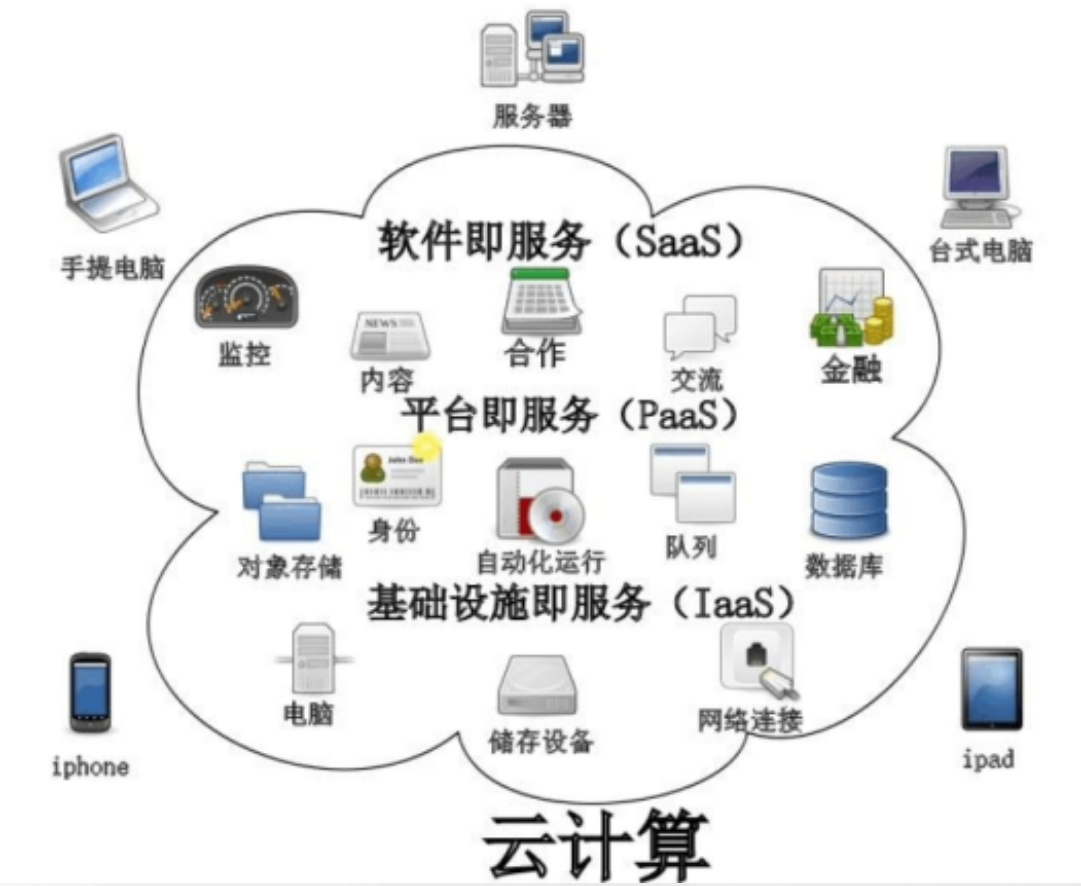
05. Spark特点

06. Hadoop与Spark关系

01

云计算概念

云计算概念



云计算是一种可以通过网络方便地接入共享资源池，按需获取计算资源（包括网络、服务器、存储、应用、服务等）的服务模型。共享资源池中的资源可以通过较少的管理代价和简单业务交互过程而快速部署和发布。

02

云计算特点

云计算特点

- 按需提供服务：以服务的形式为用户提供应用程序、数据存储、基础设施等资源，并可以根据用户需求自动分配资源，而不需要管理员的干预。比如亚马逊弹性计算云（Amazon EC2），用户可以通过Web表单提交自己需要的配置给亚马逊，从而动态获得计算能力，这些配置包括CPU核数、内存大小、磁盘大小等等。

- 宽带网络访问：用户可以通过各种终端设备，比如智能手机、笔记本电脑、PC机等，随时随地通过互联网访问云计算服务。

- 资源池化：资源以共享池的方式统一管理。通过虚拟化技术，将资源分享给不同的用户，而资源的存放、管理以及分配策略对用户是透明的。

- 高可伸缩性：服务的规模可以快速伸缩，来自动适应业务负载的变化。这样就保证了用户使用的资源与业务所需要的资源一致性，从而避免了因为服务器过载或者冗余造成服务质量下降或者资源的浪费。

云计算特点

- 可量化服务：云计算服务中心可以通过监控软件监控用户的使用情况，从而根据资源的使用情况对提供的服务进行计费。
- 大规模：承载云计算的集群规模非常巨大，一般达到数万台服务器以上。从集群规模来看，云计算赋予了用户前所未有的计算能力。
- 服务非常廉价:云服务可以采用非常廉价的PC Server来构建，而不需要非常昂贵的小型机。另外云服务的公用性和通用性，极大的提升了资源利用率，从而大幅降低使用成本。

03

Hadoop与云计算关系

Hadoop与云计算关系

- IaaS(Infrastructure as a Service)：

它的含义是基础设施即服务。比如，阿里云主机提供的就是基础设施服务，我们可以直接购买阿里云主机服务。

- PaaS(Platform as a Service)：

它的含义是平台即服务。比如，阿里云主机上已经部署好Hadoop集群，可以为我们提供大数据平台服务，我们直接购买平台的计算能力跑自己的应用即可。

- SaaS(Software as a Service)：

它的含义是软件即服务，比如阿里云平台已经部署好具体项目应用，我们直接购买账号使用它们提供的软件服务即可。

总的来说，云计算是一种运营模式，而Hadoop是一种技术手段，对云计算提供支撑。

04

Spark概念

Spark是基于内存计算的大数据并行计算框架。Spark基于内存计算，提高了在大数据环境下数据处理的实时性，同时保证了高容错性和高可伸缩性，允许用户将Spark部署在大量的廉价硬件之上，形成集群提高并行计算能力。

05

Spark特点

Spark特点

- 运行速度快

Spark 1.0核心代码只有4万行，这是由于Scala语言的简洁和丰富的表达力，以及Spark充分利用和集成Hadoop等其他第三方组件。同时着眼于大数据处理，数据处理速度是至关重要的，Spark通过将中间结果缓存在内存减少磁盘I/O来达到性能的提升。

- 易用性

Spark支持Java、Python和Scala的API，还支持超过80种高级算法，使用户可以快速构建不同的应用。而且Spark支持交互式的Python和Scala的shell，可以非常方便地在这些shell中使用Spark集群来验证解决问题的方法。

Spark特点

- 支持复杂查询

Spark支持复杂查询。在简单的“map”及“reduce”操作之外，Spark还支持SQL查询、流式计算、机器学习和图算法。同时，用户可以在同一个工作流中无缝搭配这些计算范式。

- 实时的流处理

对比MapReduce只能处理离线数据，Spark还能支持实时流计算。Spark Streaming主要用来对数据进行实时处理，而Hadoop在拥有了YARN之后，也可以借助其他的工具进行流式计算。

- 容错性

Spark引进了弹性分布式数据集RDD(Resilient Distributed Dataset) 的抽象，它是分布在一组结点中的只读对象集合，这些集合是弹性的，如果数据集一部分丢失，则可以根据"血统"对它们进行重建。另外在RDD计算时可以通过CheckPoint来实现容错。

06

Hadoop与Spark关系

Hadoop与Spark关系

	Spark	Hadoop
流式计算	Streaming	无
批计算	Core	MapReduce
图计算	GraphX	无
集群学习	MLib	Mahout
Sql	DataFrame	Hive

Turing AI 万维图灵 | 大数据系列课程

大数据

BIG DATA

智 / 能 / 科 / 技　　放 / 眼 / 未 / 来