实验一：Batch job in cloud

目录

实验概述

在实验一中，我们初识云计算平台，学习华为云提供的弹性云服务器（ECS）。ECS 向客户提供虚拟机，是使用华为云提供的各种服务的基础。

在云平台上，用户通常执行两大类任务：批处理任务与服务型任务。前者执行数据分析、大数据处理等应用，重视吞吐量，可以忍受短暂的资源不足。后者执行后端应用，重视延迟，需要尽量为其提供足够的资源，以避免违反 SLO。这两类任务的混合部署是云计算领域经久不衰的研究课题。实验一主要介绍批处理任务。实验二和三将主要介绍服务型任务。

在实验一中，我们在华为云上购买 ECS，并部署 Hadoop 与 Spark 处理框架，使用 Spark 中 GraphX API 执行简单的数据处理任务。

相关背景资料

建议同学们在开始实验前，先大致浏览以下材料，做到对本实验涉及的基本概念有所了解。

Spark：

官方文档：https://spark.apache.org/

视频解释：https://www.youtube.com/watch?v=ymtq8yjmD9I

Hadoop：

官方文档：https://hadoop.apache.org/

视频解释：https://www.youtube.com/watch?v=aReuLtY0YMI

华为云 ECS：

图解弹性云服务器：https://support.huaweicloud.com/intl/zh-cn/productdesc-ecs/ecs_01_0073.html

什么是弹性云服务器：https://support.huaweicloud.com/intl/zh-cn/productdesc-ecs/zh-cn_topic_0013771112.html

安装与准备

1. 购买华为云 ECS

使用华为云账号登录华为云平台

- 选择`产品 -> 计算 -> 弹性云服务器 ECS -> 立即购买`
- 基础配置

计费模式	区域	CPU 架构	规格	镜像	系统盘
按需计费	华东-上海一	x86	c6.large.2	Ubuntu 18.04	至少 40GB

- 网络配置

推荐使用按流量计费。

网络	安全组	弹性公网 IP
默认的 VPC	Sys-FullAccess	现在购买，公网带宽选择按流量计费，带宽大小选择 10Mbit/s

- 高级配置

1）设置云服务器名称，密码

2）云备份选择暂不购买

- 确认配置

本次实验中，我们需要购入3 台虚拟机，请务必确认收费模式是否为 按需计费 ，注意每小时金额。若误操作可能会导致账户余额不足、无法完成实验。

2. 登录 ECS

创建 ECS 后，可以在 控制台->弹性云服务器（区域务必选择上海一） 中看到弹性公网IP。

推荐使用 Vscode+ssh 插件登录 ECS

https://code.visualstudio.com/docs/remote/ssh

使用SSH工具，输入公网IP、用户名和密码，或 ssh usr@IP 即可登录。

3. 实验环境配置

- 搭建 Java 开发环境，配置好环境变量。
  参考：http://www.oracle.com/technetwork/java/javase/downloads/index.html

- 分布式框架的安装与配置
  Hadoop 配置：http://dblab.xmu.edu.cn/blog/1177-2/
  Spark 配置：http://dblab.xmu.edu.cn/blog/1714-2/
  Sbt(Simple Build Tool)配置：http://dblab.xmu.edu.cn/blog/1307-2/

实验一 Part A：认识 Spark 与 Hadoop

1. 运行 Hadoop wordcount 样例

通过运行 Hadoop 提供的 wordcount 样例，我们可以直观地感受到批处理应用的一种典型模式：先分散执行，再统一收集结果。

- 启动 Hadoop 集群后，执行 `hadoop fs -mkdir /input`，在 hdfs 根目录下新建文件夹。

- 执行 `hadoop fs -put xxx.txt /input` 将需要执行 wordcount 的文本放入新建的文件夹中。

- 执行 `hadoop jar /usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-x.x.x.jar wordcount /input /output`，运行 Hadoop 自带的 wordcount 样例（注意自己部署的 hadoop 版本）。

- 执行 `hadoop fs -cat /output/part-r-00000` 打印结果。

2. 运行 Spark Connected Component 样例

在这一部分中，我们将初步认识 Spark 处理框架。Spark 提供了 GraphX API，可以用于处理图数据。我们将参考 wordcount 应用的打包方法，打包运行一个使用 GraphX API 编写的图处理应用。

- 下载样例数据：

https://github.com/apache/spark/blob/master/data/graphx/

- 参考 wordcount 应用的编写方法：
  http://dblab.xmu.edu.cn/blog/1311-2/
  对样例编译打包成 jar 文件：
  https://github.com/apache/spark/blob/master/examples/src/main/scala/org/apache/spark/examples/graphx/ConnectedComponentsExample.scala

- 将生成的 jar 包通过 spark-submit 提交到 Spark 中运行。

3. 实验二 Part A 实验报告

- 请你给出执行 Hadoop wordcount、Spark connected component 的执行结果截图
- 你认为华为云 ECS 是一种 IaaS 还是一种 PaaS？请谈谈你的看法。
- 请你分析一下 Hadoop 与 Spark 两种处理框架的区别。

注意：实验一 Part A 报告的总篇幅不得超过一页（五号字，不含截图），若超过一页则按不及格处理。

实验一 Part B：使用 Spark 执行 PageRank 算法

1. 编写 PageRank 算法

经过实验一 Part A 部分的学习，相信同学已经对 Spark 框架有了基本的了解。在这一部分中，同学们需要自主地编写一个 PageRank 算法。

1. 下载数据集：
   http://snap.stanford.edu/data/wiki-Vote.html

2. 请你自主地编写一个 PageRank 算法，选出声望最高的前 20 名候选人名单。

2. 实验二 Part B 实验报告

- 请你给出 20 名候选人的名单
- 请你大致描述一下你实现的 PageRank 算法，算法的部署流程以及运行步骤
- 注意：实验二 Part B 报告的总篇幅不得超过两页（五号字），若超过两页则按不及格处理。