**《Storm实时大数据处理》课程教学大纲**

一、课程基本信息

课程代码：18100112

课程名称：Storm实时大数据处理

英文名称：Storm

课程类别：选修

学 时：32

学 分：2

适用对象:大数据专业

考核方式：考试

先修课程：java, Hadoop, unix, lunix,离散数学

二、课程简介

Storm是一个免费开源的分布式实时计算系统。Storm能秦松可靠地处理无界的数据流，就像Hadoop对数据进行批处理；但是Storm能持续地运作下去，并且要掌握Storm相对简单，开发人员可以使用任何编程语言对它进行操作，而得到满意的结果。

Apache Storm is a free and open source distributed realtime computation system. Apache Storm makes it easy to reliably process unbounded streams of data, doing for realtime processing what Hadoop did for batch processing. Apache Storm is simple, can be used with any programming language, and is a lot of fun to use!

三、课程性质与教学目的

本课程是大数据专业的必修课。Storm是目前完善且主流的开源环境，用于进行实时的分布式计算，处理时效性较强的问题效果较好。为了使学生接触到主流的大数据处理技术，开设本门课程。教学目的主要是要求学生了解Storm的基本原理，可以进行开发环境与生产环境的配置。能对某些问题提出解决思路，在有参考代码的前提下进行二次开发。

四、教学内容及要求

**第一章 基础概念模块**

1. 目的与要求
2. 了解Storm的发展过程，目前有哪些企业使用Storm
3. 了解Storm的基本概念，配置，序列化，容错机制等
4. 理解Storm中的拓扑

1. 教学内容

第一节

1. 主要内容：Storm的基本概念，如元组，流，龙卷，闪电，拓扑，主控节点与工作节点等
2. 主要知识点：龙卷，闪电和拓扑。龙卷是拓扑的流的来源，是一个拓扑中产生源数据流的组件。闪电是流的处理节点。拓扑是Storm中运行的一个实时应用程序，因为各个组件的消息流动而形成逻辑上的拓扑。

第二节

1. 主要内容：容错机制，可靠性机制，消息传递机制
2. 主要知识点：容错机制，当一个工作进程死亡，主控会尝试重启，如果连续多次重启失败，则重新分配进程。可靠性机制，创建新链接通知主控，完成单一元组通知主控。

第三节

1. 主要内容：开发环境，生产环境，命令客户端
2. 主要知识点：开发环境，需要jdk，maven，eclipsa。生产环境，集群环境。

1. 思考与实践

会使用本地模式。尝试自主组建集群。

1. 教学方法与手段

课堂教授，多媒体教学

**第二章 基本组件模块**

1. 目的与要求
2. 理解龙卷，闪电，拓扑

1. 教学内容

第一节

1. 主要内容：拓扑
2. 主要知识点：一个简单的拓扑，拓扑的常见模式
3. 问题与应用：本地模式下运行拓扑，生产集群上运行拓扑

第二节

1. 主要内容：龙卷
2. 基本概念和知识点：龙卷获取数据的方式，常用的龙卷
3. 问题与应用：学习编写龙卷类

第三节

1. 主要内容：闪电
2. 基本概念和知识点：可靠与不可靠的闪电，复合流
3. 问题与应用：学习编写闪电类

1. 思考与实践

能结合具体问题进行相应部分代码的编写，尝试分析网购平台进行购物节促销活动涉及到的实时数据处理问题，写出结构图，流程图。

1. 教学方法与手段

课堂教授，多媒体教学

**第三章 分布式远程调用详解**

1. 目的与要求
2. 理解分布式远程调用

1. 教学内容

第一节

1. 主要内容：线性分布式远程调用拓扑构建器
2. 主要知识点：输入参数为string类型的函数，6中多态方法，线性远程分布式调用的工作原理

第二节

1. 主要内容：本地模式的远程分布式调用
2. 基本概念和知识点：局部远程分布式调用对象，局部集群

第三节

1. 主要内容：远程模式的分布式调用
2. 基本概念和知识点：启动服务器，配置服务器的地址，提交拓扑到集群

1. 思考与实践

微博粉丝集去重

1. 教学方法与手段

课堂教授，多媒体教学

**第四章 事务拓扑详解**

1. 目的与要求
2. 理解事务拓扑

1. 教学内容

第一节

1. 主要内容：事务拓扑
2. 主要知识点：事务拓扑的核心是保护数据处理的严格有序

第二节

1. 主要内容：事务实拓扑的细节
2. 基本概念和知识点：管理状态，协调事务，故障检测，封装

第三节

1. 主要内容：事务拓扑的实现细节
2. 基本概念和知识点：事务龙卷的工作原理，异常处理

1. 思考与实践

计算指定输入流的元组的全局计数

1. 教学方法与手段

课堂教授，多媒体教学

**第五章 Storm的集群搭建**

1. 目的与要求
2. 了解集群搭建步骤

1. 教学内容

第一节

1. 主要内容：准备工作
2. 主要知识点：搭建ZooKeeper集群，安装Storm本地以来，下载发行版本，配置yaml文件，启动守护进程
3. 问题与应用：本地模式的配置

第二节

1. 主要内容：集群安装
2. 基本概念和知识点：配置host文件，配置静态IP，集群SSH无密码，修改主机名，同步时间，安装JDK
3. 问题与应用：配置集群若干节点

第三节

1. 主要内容：准备开发环境
2. 基本概念和知识点：如何管理Storm，使用Maven管理项目，部署伪分布的ZookKeeper，部署伪分布的Storm集群
3. 问题与应用：项目提交至集群运行

1. 思考与实践

能运行微博粉丝集去重任务，能运行模拟购物节促销活动实时数据处理

1. 教学方法与手段

课堂教授，多媒体教学

五、各教学环节学时分配

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **教学环节**  **教学时数**  **课程内容** | **讲**  **课** | **习**  **题**  **课** | **讨**  **论**  **课** | **实验** | **实习** | **其他教学环节** | **小**  **计** |
| 第一章 | 2 |  |  | 2 |  |  | 4 |
| 第二章 | 8 |  |  | 6 |  |  | 14 |
| 第三章 | 2 |  |  | 2 |  |  | 4 |
| 第四章 | 2 |  |  | 2 |  |  | 4 |
| 第五章 | 6 |  |  |  |  |  | 6 |
| 合计 |  |  |  |  |  |  | 32 |

六、课程考核

（一）考试

（二）成绩构成

平时成绩占比： 40% 期末考试占比：60%

（三）成绩考核标准

七、推荐教材和教学参考资源

1.http://storm-project.net/

2.http;//github.com/nathanmarz/storm

3.http;//mathanmarz.com

4.http://github.com/nathanmarz/storm-starter

5.http://github.com/nathanmarz/storm-deploy

6.Jonathan Leibiusky, Gabriel Eisbruch, Dario Siomonassi. Getting Started with Storm[M].O’Reilly Media, 2012

7.张广骏.实时大数据分析，基于Storm，Spark技术的实时应用。北京;清华大学出版社。2018年1月

大纲修订人： 邹杨 修订日期：2020.12

大纲审定人： 审定日期：