P1

大家好，欢迎大家来参加大数据技术讨论。

今天我们继续大数据技术的讨论。

**今天主要有三部分内容**

首先是由我，介绍一下我们这个讨论小组近期讨论和学习的内容，未来的计划等

希望徐老师给出指导意见。

第二部分是杜一凡，会带来一个Spark Spark Streaming的学习分享

第三部分是我们的一个互动主题，本周由陈帅，胡文龙，鹿强，分别带来一个简单的开源项目调研分享。

P2 首先由我开始

第一部分，由我来讲一下

我们这个讨论小组，讨论的目的，讨论的内容，以及未来时间的一个规划。

P3

我们目前可以认为是一个数据工程和数据科学结合的数据团队。但我们大部分人都还没有相对全面的掌握大数据技术的基础知识。这样就只能浮于表面，做一些技术开发。

大数据技术跟普通的编程技术不太一样，有一定的门槛，大家要想一起组建一个大数据工程团队。

一起学习一些大数据基础知识是必要的，所以讨论组的第一个目的就是我们一起学习基于一门大数据课程，掌握大数据的基础知识

**第二个目的**，就是项目需求攻关。 之前一直围绕海洋AIS数据开展我们的数据平台开发。 现在一些新的项目，包括信工所，33基地项目等都和我们现有的架构有很大区别。

**例如信工所，为满足大批量点迹数据的实时上显，需要用到性能更好的Redis数据库，如何满足高效的数据提取，将是我们讨论的主题。**

**又例如33的项目，对数据要进行非常复杂的计算，而且算法不确定，需要适时添加。安师兄在写方案的时候提出了基于算法服务器的松耦合解决方案。**

**那么基于这个思路，如何进行算法流程的描述和配置等，或者说是一个通用的算法引擎，也将是我们讨论的主题。**

因此需要我们就新的项目需求，开展一些专题攻关的项目讨论

第三个目的，就是研究方向的凝练，从陈昭，杜一凡开始，包括孙涛，都会基于数据平台展开硕士，或者博士的论文工作。

对于你们在讨论中的要求，就会更高，需要从实际问题中凝练出研究方向，并提出更好的解决方法。

这就是我们讨论的三个目的。

P4

基于这三个目的，围绕大数据技术，要讨论的内容会非常多。

我想了几天，我按照由易到难，由浅到深，可以将讨论分为三个阶段。

第一个阶段，我们主要关注的是，当下主流大数据平台技术，这个阶段以学习为主，注重基础知识和技术视野

第二个阶段，我们将面向项目需求，进行专题技术讨论

第三个阶段，基于论文的深入讨论 和 项目开发实践

下面我们来详细的看一下这三个阶段更具体的讨论内容

P4

第一阶段，我们已经做了相对详细的规划

主线，就是Big Data infrastructure的课程，主要通过Hadoop，

我们一起学习大数据系统的基础知识，设计思想和开发模式等。

但大数据不等于hadoop，为了让大家拜托这种误区。

主线之外，还有两条辅线，一个是由杜一凡担纲的，Spark主题，和陈昭担纲的SQL on hadoop主题

这两部分是他们两个一直关注的领域，他们将分享他们的一些经验，跟大家一起讨论学习。

此外，还有一个互动环节。上周已经布置给大家。

P5 6 一主两副之外。我们还有一个互动环节

就是在庞大的hadoop生态系统里，找若干个自己感兴趣的开源项目，进行一个分享。

**Hadoop 生态系统**

**Hadoop 生态系统对于我们来说就是一个代码宝库**

**经过十年的发展，他在数据工程和数据科学的各个环节，都有了很高效的开源项目**

**我们要组建一个数据工程，数据科学团队。**

**一个重要的工作，就是整理我们自己有关数据的需求，然后去hadoop生态系统中找相应的项目原型。**

**进行学习和改进。**

**当然，这个过程，是智慧密集型的操作，希望大家都能一起贡献智慧，也能提升自我。**

第一阶段，兴趣导向。第二阶段，需求导向。

本周有陈帅，胡文龙，鹿强，分别就他们各自感兴趣的领域，对Pig，Hbase和xx 准备了一个ppt

分享他们初步学习的成果。

在第一阶段，大家都请准备一个，进行分享。

P7 下面我们用几页PPT，对过去两三周讨论的内容进行一个总结和回顾

这两三周，我们主要基于马里兰大学的大数据基础设施课程展开学习和讨论

我们一起看了大数据的起源。

欧洲大型强子对撞机 LHC

大型综合巡天望远镜

SKA超大天文望远镜网络

Google 2014年已经是 2EB级别的数据库服务级别

大数据平台技术的起源就是数据的爆发，是被数据逼出来的一个学术方向。

P8

我们还讨论了数据库的负载 联机事务处理和联机分析处理

大数据是今年提出的概念，但数据工程的概念由来已久

大数据其实 并没有改变数据工程的负载

只是，优化和丰富了这个流程

P9

然后我们引出了hadoop的内容

Hadoop 内核

P10

由hadoop的内核，经过十年发展，已经衍生出100多个开源项目

如刚才所说，这是我们的宝库。

P11