1. What should be the job responsibility of an IT Architect?

架构师，是一个既需要掌控整体又需要洞悉局部瓶颈并依据具体的业务场景给出解决方案的团队领导型人物。主要职责有：

1. 确认需求，要懂得用户需求，理解用户真正想要什么；
2. 系统分解，一般分为纵向分解和横向分解，纵向分解是将整个系统分层，从而将整体系统分解成下一级的子系统与组件。横向分解是在系统分解成不同的逻辑层或服务后，对逻辑层进行分块，确定层与层之间的关系；
3. 技术选型；
4. 制定技术规格说明，协调所有开发人员。
5. How to implement HDFS in a large Hadoop cluster? Why?

HDFS集群拥有一个NameNode和一些DataNode。NameNode管理文件系统的元数据，DataNode存储实际的数据。客户端通过同NameNode和DataNodes的交互访问文件系统。客户端联系NameNode以获取文件的元数据，而真正的文件I/O操作是直接和DataNode进行交互的。因为HDFS具有高容错性、适合批处理、适合大数据处理支持流式数据访问的优点。

1. What could be the Mahout Strengths for software developers?
2. 技术社区比较活跃，给开发者提供了很好的交流平台；
3. 扩展性和容错性较好，充分利用了MapReduce和HDFS的扩展性和容错性；
4. 100%源代码开源；
5. 易于使用，实现了大部分常用的数据挖掘算法，如聚类算法、分类算法和推荐算法等。
6. What’re the key advantages of Spark?
7. 快：Spark实现了高效的DAG执行引擎，可以通过基于内存来高效地处理数据流。
8. 容易使用：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应用。而且Spark支持交互式的Python和Scala的Shell。
9. 通用性：Spark提供了统一的解决方案。Spark可以用于批处理、交互式查询、实时流处理、机器学习和图计算这些不同类型的处理都可以在同一应用中无缝使用。
10. 可融合性：Spark非常方便的与其他开源产品进行融合。如Spark可以使用Hadoop的YARN和Apache Mesos作为它的资源管理和调度器。Spark也可以不依赖第三方的资源管理器和调度器，使得所有人可以非常容易地部署和使用Spark。
11. Have you tried to play Mahout (see MahoutPractice.doc)? What’s your experience?

Mahout中主要核心三大算法为推荐，聚类及分类算法，推荐中常用的两个算法是”user\_based”和”item\_based”，前者主要通过和你兴趣相似的人来发现新的你感兴趣的东西，而后者则是发现一些和你所喜欢的事物相似的事物。还有一种基于“contend\_based”的推荐算法，它是根据事物所拥有的元数据出现进行事物的推荐，例如，如果你喜欢的电影是陈道明导演的，那么你可以向你的朋友推荐他拍摄的其他影片，推荐的理由就是基于电影的一个属性——导演，这个就是基于内容的推荐。

我尝试的主要流程为：1.构建数据集；2.建立简易推荐器；3.相关解释说明；4.运行结果展示；5.推荐器评估；6.分布式推荐计算。