1116面试总结

1. 人品方面：一定要记住自己的项目工作地址，详细到某区某大厦，虽然问这么详细大概率就是已经穿帮了，但是也要装得像。
2. 技术方面：（被问到的一些答不上来也不明白的）

首先是hadoop大数据平台问得比较多，今天基本都问了，要了解大数据平台的原理和相关的组件和工具

Hadoop集群具体来说包含两个集群：HDFS集群和YARN集群，两者逻辑上分离，但物理上常在一起。

（1）HDFS集群：负责海量数据的存储，集群中的角色主要有 NameNode / DataNode/SecondaryNameNode。

（2）YARN集群：负责海量数据运算时的资源调度，集群中的角色主要有 ResourceManager /NodeManager

（3）MapReduce：它其实是一个应用程序开发包。

Hive是面向Hadoop生态系统的数据仓库。它让用户可以使用HiveQL查询和管理大数据，这是一种类似SQL的语言。

**Sqoop**

企业经常需要在关系数据库与Hadoop之间传输数据，而Sqoop就是能完成这项任务的一款工具。它可以将数据导入到Hive或HBase，并从Hadoop导出到关系数据库管理系统(RDBMS)。

**Spark**

作为MapReduce之外的一种选择，Spark是一种数据处理引擎。它声称，用在内存中时，其速度比MapReduce最多快100倍;用在磁盘上时，其速度比MapReduce最多快10倍。它可以与Hadoop和Apache Mesos一起使用，也可以独立使用。

报表的开发：带参数的报表开发—>在帆软中通过SQL语句将参数设置，有特殊符号WHERE 查询字段='${查询字段}'

填报报表的开发：1.定义数据连接2.获取数据3.填报报表设计（重点）4.报表填报属性设置（还需要明天去教室实验下，只知道有这个可以手动在帆软中进行增删改的功能）

3、还有linux xshell shell Hadoop 被问混乱了，有没有人能帮我捋捋？

4、SQL相关的：碰到连续的字段查询，首先想到分组和伪劣，例如连续三天登录，用日期减去伪劣计算出相同的值大于等于3即可。另外还有问到not exists 和 not in 的性能，请尽量不要使用not in(它会调用子查询)，而尽量使用not exists（它会调用关联子查询）。如果子查询中返回的任意一条记录含有空值，则查询将不返回任何记录。如果子查询字段有非空限制，这时可以使用not in。如果查询语句使用了not in，那么对内外表都进行全表扫描，没有用到索引；而not exists的子查询依然能用到表上的索引。所以无论哪个表大，用not exists都比not in 要快。（面试官给的答案貌似是 not exists 只需要对表进行一次扫描，而not in 每一次的比对都要对全表进行扫描）