集团大数据平台项目整改方案规划

我最近对集群的了解，提出下面六点整改建议

一、增加一套测试集群环境

上线之前，没有测试环境。

一般有4个环境：开发环境(DEV) 测试环境(UAT) ,仿真环境,生产环境(PROD)

至少三个环境,也可以说是系统开发的三个阶段：开发->测试->上线

测试集群配置:cpu 8核 内存32G 硬盘300G 数量3台

二、集群优化

方案A：物理隔离

拆分成2个集群，共用一个管理控制台(cm)

独立的HDFS底层，同时把hbase、kafka、zk分离出来。

优点：

1.hbase稳定,不会受其他任务影响，特别是大的离线分析任务。

2.hbase的机器，可以把内存省出来，放到spark，或impala的节点上。做到资源合理分配。

缺点：

1.hive回流到hbase是跨集群操作。没有同集群方便，要通过认证和外部网络。之前写的程序需要修改。

2.hive不能直接关联hbase建外部扩展表。

方案B：逻辑隔离

一个集群，共用底层hdfs，在部署hbase的节点上，去掉yarn。跑mapreduce的任务不会调度到hbase的节点上。这也是cloudera公司官网推荐的方案。

优点：改动较小，之前写的程序不需要改动。

缺点：由于共用hdfs，IO还是有影响

对于我们公司，我建议选方案A。支付行业对hbase要求高。

三、任务调度问题

当有任务依赖其他任务时，不管对错，依然执行。

举例：当其中有一个基表有问题，导入数据有问题。后续依赖它的每一个任务都是有问题的。跑批没有终止，仍然跑完所有任务。
解决方法：任务执行完写日志。新的任务先读日志，判断依赖的表是否完整。

四、改善用户体验

1.申请2台跳板机
原因：
每一个用户要想使用大数据，都要自己配置环境，都要走流程开通端口，申请相关权限。我们还要提供技术支持。耽误大量时间。在跳板机上统一配置好环境。他们就直接就可以使用了，又能保证数据的安全。
2.成立工作站(jupyter)。
原因：
做数据挖掘、模型分析的和BI的同事，并不熟悉大数据环境。让他们直接连接集群风险很大.提供一个工作站，web的方式，让他们直接可以使用python，R，spark，scala等

五、安全问题。

1.权限整改
新颜将取消个人账户，根据小组划分账户。
2.限制下载数据
3.hbase数据备份
定时做快照，相对重要的表15天之内的快照，资源充足的情况下，我们将跨机房备份。

六、生产问题

遇到到问题，要记录在wiki上。总结问题，提出改善方案。

确定好方案后，我将规划一下工作，需要哪些资源，大概什么时候能完成。