八斗学院 2020 年最新大厂•大数据面试题（内部整理）

2020-04-15

某 1 公司面试

笔试

1.写出你最熟悉的一个大数据组件，并说明其原理

2.写出你了解的机器学习算法，并选择其中一种简述其原理

3.新浪网站和头条的新闻推荐算法，试分析其业务逻辑和技术实现

4.写出 hive 和 hbase 的区别

5.有数据类型

0001，2018-01-02
0003，2019-11-02
0002，2017-11-03

实现 mr 代码，要求对时间排序，输出结果为序号+用户 id
如 1. 0002
2. 0001
3. 0003

语言不限 java,python,scala 均可

6.已知日志数据存储在 hdfs 上，路径为'/log/20181003',其中有许多日志文件，如 a1.log,a2.log,a3.log...

其中数据格式为：
192.168.22.1,POST/app1/x1

192.168.127.1,get/app2/x2

192.168.101.2,POST/app3/x3

语言不限，可以用 java,python,scala

1)要求用 spark 开发一个应用，统计 pv,uv(对 ip 地址去重)

2）把上边的 rdd 转换成 DataFrame,用 sparksql 进行统计

7.什么是数据倾斜？如何解决？

面试

1.hive 如何从 mysql 中获取数据？

2.表之间有关联，如何放到 hive 中，在 mysql 中做关联，还是在 hive 中，为什么？

3.hbase 中的表如何设计？假设我有个字段用户名和时间，有时候单个查询，有时候一起查询，要如何设计？

4.为什么要使用 flume 做数据采集？agent 进程会占用一些资源，假设资源不够用了，你怎么办？

5.之前做过的项目架构介绍一下，集群规模，数据量多少，你参与了哪些部分的设计？为什么要选用这些组件？

6.流式处理，flink 有了解过吗？与 spark 有什么不同？

7.数据清洗时，是否需要对尽可能多的数据字段做持久化？如何实现？

8.有了解过一些企业级的大数据平台吗？是否有动手搭建过？

9.redis 的 hash 原理，分布式情况下，如果一个节点挂了，数据会怎么变化？

10.索引是什么？怎么实现？

11.mysql 的存储原理

12.分布式情况下如何保证数据一致性？

13.mr 原理，spark 和传统 mr 相比，为什么速度快？

14.为什么选用 redis 做数据库？有什么优点？

某 2 公司直接面试

面试

1、项目规模，一天/月数据量，各组件版本

2、spark 2 和 spark 1 版本有什么区别

3、项目上经常遇到那些问题，如何解决的

4、Hive 元数据存储了哪些信息

5、数据去重怎么做

6、udf udaf udtf 有什么区别

7、项目上的数据仓库分层怎么做的

8、spark streaming 和 spark 的联系和区别

9、推荐系统了解吗？你们推荐有什么算法做的

10、kafka 是如何保证数据的安全和可靠性的

11、kafka 的数据是有序的吗

12、spark 的优化你们一般从哪些方面做的

13、spark streaming 计算速度远远小于 kafka 缓存的数据，怎么解决

14、spark streaming 对接 kafka 的两种方式有什么区别

15、数据质量如何如何监控的

某 3 公司算法面试

题目

1：写出推荐系统的框架思路（线上比试）

2：编程实现二叉树的中序遍历（线上比试）

3：深度学习不收敛的原因（线上比试）

4：LR 模型的推导

5：随机森林的原理

6：则怎样处理连续数据

7：随机森林的损失函数是什么

8：分别利用 MR 和 Spark 实现求解：每个月男人收入的平均值即：1 月所有男人的收入平均值，2 月所有男人的平均值.....

数据的格式为 月份 性别 收入
1 男 1 3000
3 女 1 2000
2 男 2 20000
.....
求解思路(要尽量高效率)
MR 思路： 在 mapreduce 的基础上结合 combiner 和 cleanup 函数
Spark 思路：（如果用 groupByKey 效率可能回低）可以考虑用 combineByKey 这个算子

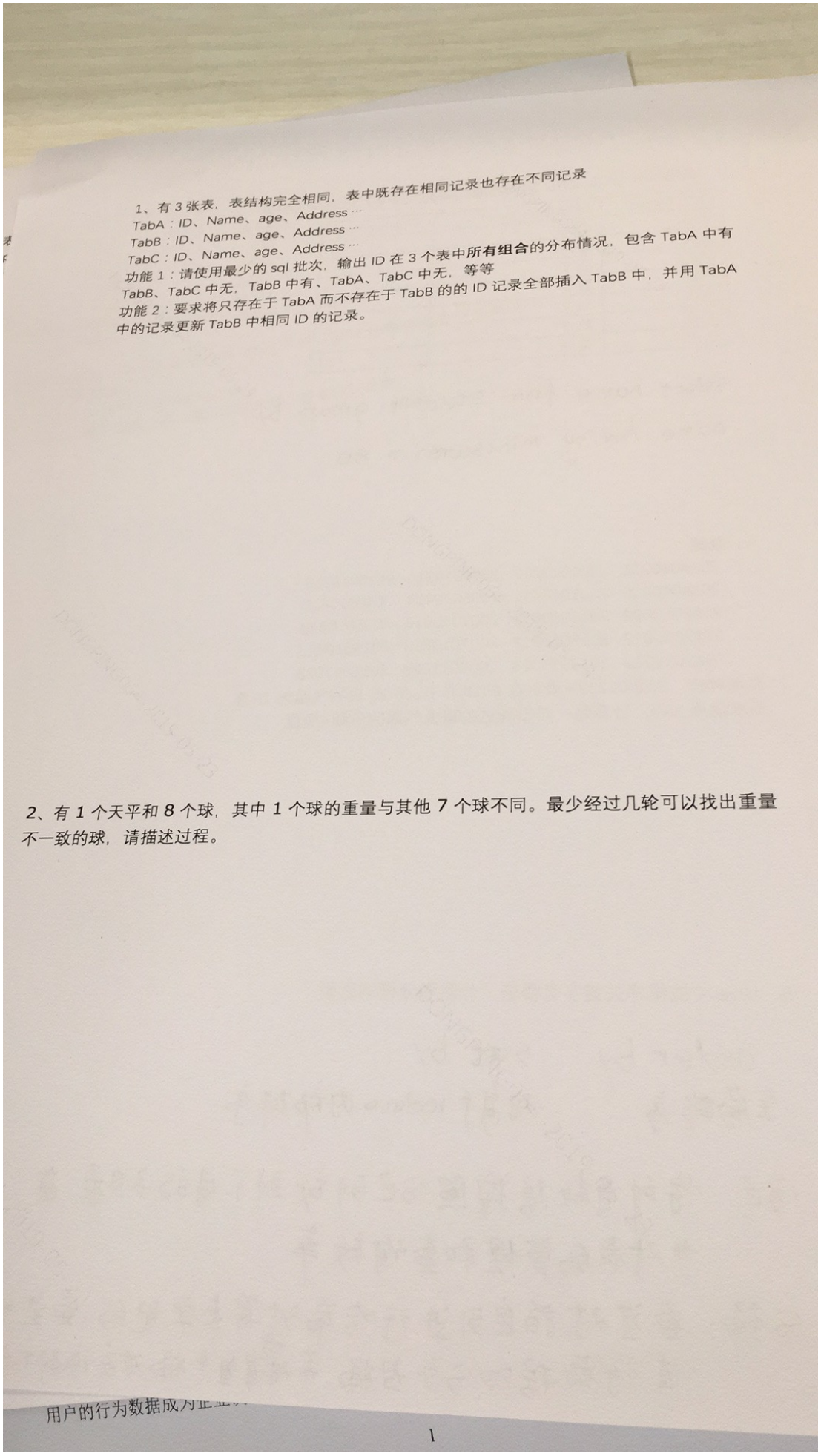
9：编程实现有序矩阵的定位查找，

10：编程实现二分查找

某 4 公司直面

题目

1. hive执行哪些操作时会触发MR，哪些操作不能触发MR？
2. hive触发MR转换过程？
3. 如何处理数据倾斜？
4. RDD的五大特性是哪几特性？
5. spark作业执行流程？
6. spark sql与RDD之间如何转换？
7. flume生产中如何设计？
8. flume的三个组件？
9. kafka架构？
10. 生产中数据量？
11. 介绍工作中项目选型、数据量？
12. 介绍广播变更？
13. yarn的工作流程？
14. 介绍spark中的隐式转换与使用？



1、有 3 张表，表结构完全相同，表中既存在相同记录也存在不同记录
TabA：ID、Name、age、Address …
TabB：ID、Name、age、Address …
TabC：ID、Name、age、Address …
功能 1：请使用最少的 sql 批次，输出 ID 在 3 个表中**所有组合**的分布情况，包含 TabA 中有 TabB、TabC 中无，TabB 中有、TabA、TabC 中无，等等
功能 2：要求将只存在于 TabA 而不存在于 TabB 的的 ID 记录全部插入 TabB 中，并用 TabA 中的记录更新 TabB 中相同 ID 的记录。

2、有 1 个天平和 8 个球，其中 1 个球的重量与其他 7 个球不同。最少经过几轮可以找出重量不一致的球，请描述过程。

用户的行为数据成为企业

1

3、请用 hive SQL 语句 查询出每门课都大于 80 分的学生姓名

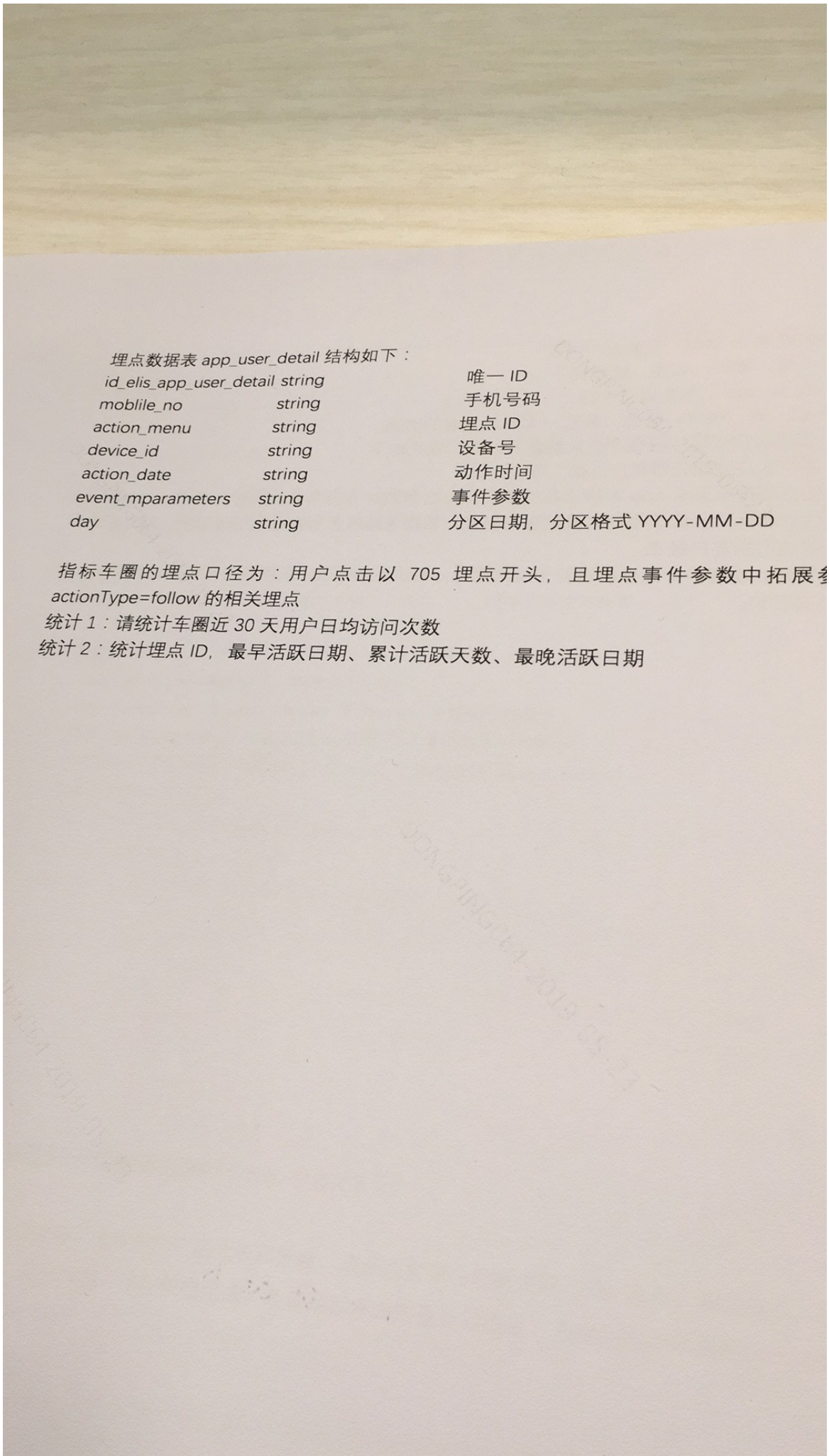
姓名(name)	课程(subject)	得分(score)
张三	语文	89
张三	数学	76
李四	语文	66
李四	数学	95
王五	语文	81
王五	数学	90
王五	英语	100

4、数据：

2014010216 2001010212 2008010216 2010010216
2014010410 2001010411 2008010414 2010010410
2012010609 2013010619 2007010619 2015010649
2012010812 2013010812 2007010812 2015010812
2012011023 2013011023 2007011023 2015011023

数据解释：2010012325 表示在 2010 年 01 月 23 日的气温为 25 度
要求使用 hive，计算每一年出现过的最大气温的日期+温度

5、Hive 中的排序关键字有哪些？分区和分桶的区别？



埋点数据表 app_user_detail 结构如下：

id_elis_app_user_detail	string	唯一 ID
moblile_no	string	手机号码
action_menu	string	埋点 ID
device_id	string	设备号
action_date	string	动作时间
event_mparameters	string	事件参数
day	string	分区日期，分区格式 YYYY-MM-DD

指标车圈的埋点口径为：用户点击以 705 埋点开头，且埋点事件参数中拓展参
actionType=follow 的相关埋点
统计 1：请统计车圈近 30 天用户日均访问次数
统计 2：统计埋点 ID，最早活跃日期、累计活跃天数、最晚活跃日期

京东面试

面试

1、项目规模，一天/月数据量，各组件版本？

数据规模一般 100M 数据由 300 万条数据 数据量上百 G 条数达到几十亿条数据
美团数据规模：负责每天数百 GB 的数据存储和分析

2、Spark 2.x 和 Spark 1.x 版本的区别？

1）Spark2.x 实现了 Spark sql 和 Hive Sql 操作 API 的统一。
2）Spark2.0 中引入了 SparkSession 的概念，它为用户提供了一个统一的切入点来使用 Spark 的各项功能，统一了旧的 SQLContext 与 HiveContext
3）统一 DataFrames 和 Datasets 的 API
4）Spark Streaming 基于 Spark SQL(DataFrame / Dataset)构建了 high-level API，使得 Spark Streaming 充分受益 Spark SQL 的易用性和性能提升

3、项目中的遇见的问题，如何解决？【讲述数据倾斜】

4、Hive 元数据存储了哪些信息？

存储了 hive 中所有表格的信息，包括表格的名字，表格的字段，字段的类型就是表的定义

5、数据去重怎么做？【UDF 使用】

在 hive 数据清洗这里总结三种常用的去重方式
1.distinct
2.group by
3.row_number()
实例：

```
SELECT tel, link_name, certificate_no, certificate_type, modify_time
  FROM order_info
 WHERE deleted = 'F'
   AND pay_status = 'payed'
   AND create_time >= to_date('2017-04-23', 'yyyy-MM-dd')
   AND create_time < to_date('2017-04-24', 'yyyy-MM-dd')
   AND row_number() over(PARTITION BY tel ORDER BY tel DESC) = 1
```

上面 SQL 对某一字段（tel）排序后分区去重，这样避免了其对不相干字段的
数据干扰，影响数据处理的效率推荐方法三

6、udf, udaf, udtf 有什么区别？

UDF：用户自定义普通函数，1 对 1 关系，常用于 select 语句
UDAF：用户自定义聚合函数，多对 1 关系,常用于 group by 语句
UDTF：用户自定义表生成函数,1 对多关系 分词输入一句话输出多个单词

7、项目上数仓分层如何做的

8、Spark Streaming 和 Spark 联系和区别？

9、推荐系统了解吗？里面涉及到的算法？

叙述推荐系统流程，CB，CF，NB

10、Kafka 如何保证数据的安全性和可靠性？

11、Kafka 的数据是有序的吗？

12、Spark 优化？

13、Spark Streaming 计算速度远远小于 Kafka 缓存的数据，怎么解决？

14、Spark Streaming 对接 Kafka 的两种方式的区别？

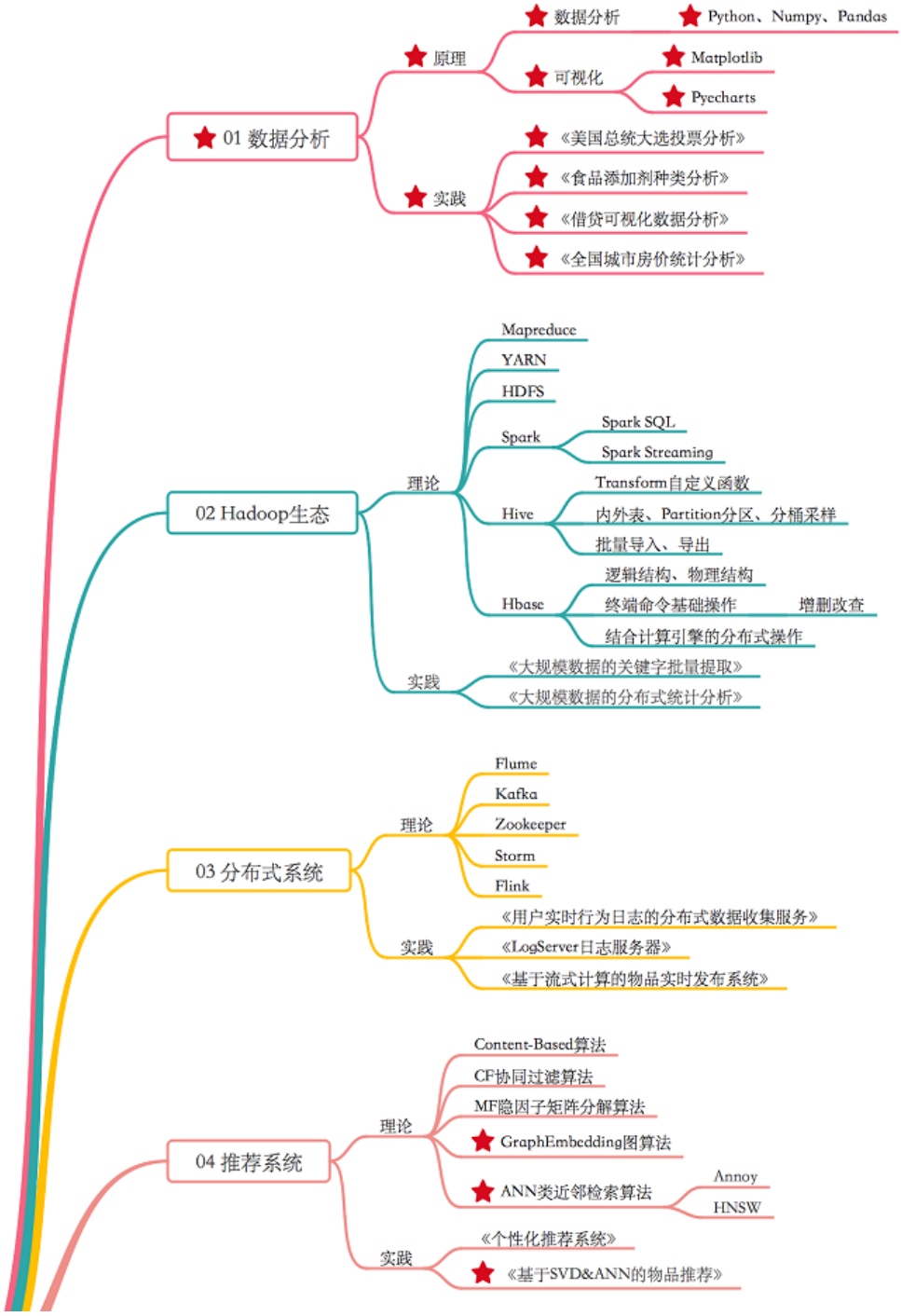
15、数据质量如何监控？

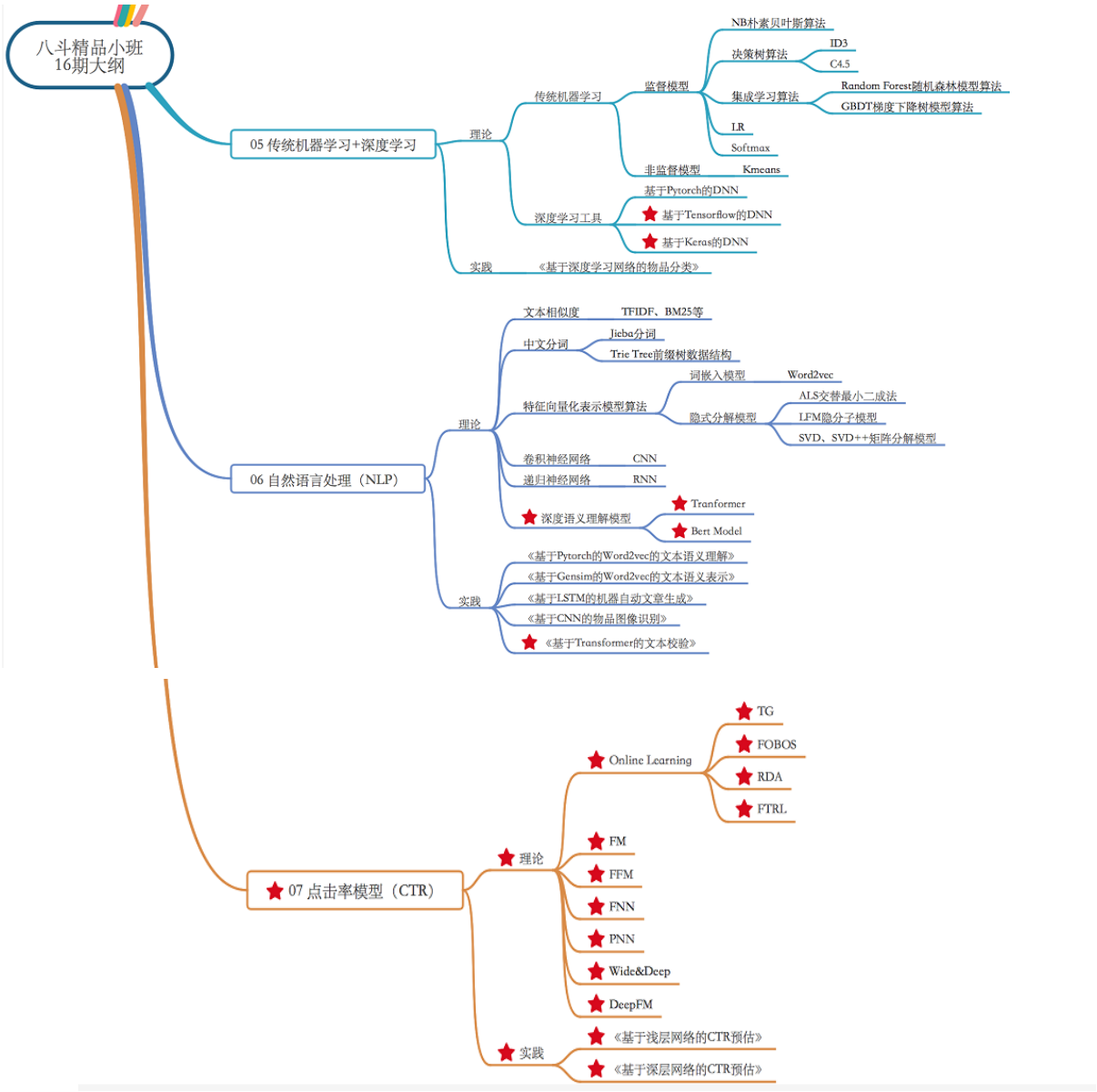
16、Flume 断点续连？

17、Spark 时间窗口怎么设置？是基于业务还是基于架构

最后分享

八斗 2020 年最新课程大纲





学习目标 1

解决大数据业务问题，重点在于策略工程相结合，从离线到在线、从工程到策略、从局部到全局掌控，培养架构师思维和问题解决能力。

学习目标 2

生态主要组件逐点击破，数据上下游关联方法，从微观组件到宏观架构建设，具备全局设计和搭建能力

2020-04-15 整理发布
八斗学院

学习目标 3

灵活运用生态组件，模拟真实用户行为，打通整条数据通路，从生产到消费，培养数据 pipeline 全局开发思维和动手能力

学习目标 4

大数据挖掘算法，配合常用 hadoop大数据处理工具，熟悉策略开发通路，达到大数据处理的目的