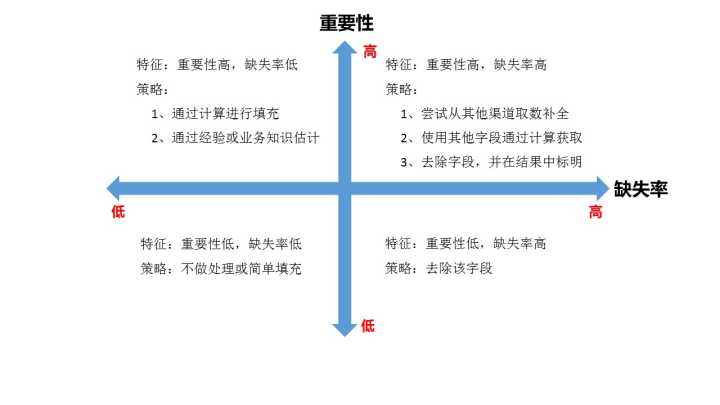
## 第一步数据预处理

1. **选取处理工具：**鉴于数据实时传输，处理工具选择kafka与SparkStreaming。
2. **数据查看：**数据查看可以大致分为两方面。一是查看数据的元数据信息，包括数据来源，字段解释，字段的逻辑关系。二是从数据源抽取一部分数据，人工查看，加深对数据的了解，发现问题，为后续的数据清洗做准备。

## 第二步去除或补全有缺失的数据

1. **确定缺失值范围：**根据数据的重要程度以及数据的缺失程度评估缺失分数，然后按照缺失比例和字段重要性，分别制定策略，可用下图表示：



1. **去除不需要的数据：**这一步主要是为了减少数据的冗余。建议做好数据备份，避免后面的业务逻辑需要用到某些字段。或者先在少量数据上试验成功之后再对全量数据进行处理。
2. **填充缺失内容：**某些没有的数据值需要进行填充，避免后续程序在对数据进行分析时出现空值错误。数据填充方法大概分为三种：
   1. 以业务知识或经验推测填充缺失值
   2. 以同一指标的计算结果（均值、中位数、众数等）填充缺失值
   3. 以不同指标的计算结果填充缺失值

前两种方法比较好理解。关于第三种方法，举个最简单的例子：年龄字段缺失，但是有屏蔽后六位的身份证号，从中提取需要的信息。

**4、重新取数：**如果某些指标非常重要又缺失率高，那就需要和取数人员或业务人员了解，是否有其他渠道可以取到相关数据。

以上，简单的梳理了**缺失值**清洗的步骤，但其中有一些内容远比我说的复杂，比如填充缺失值。很多讲统计方法或统计工具的书籍会提到相关方法，有兴趣的各位可以自行深入了解。

## 第三步数据内容清洗

如果数据是由系统日志而来，那么通常在格式和内容方面，会与元数据的描述一致。而如果数据是由人工收集或用户填写而来，则有很大可能性在格式和内容上存在一些问题，简单来说，格式内容问题有以下几类：

1. **时间、日期、数值、全半角等显示格式不一致**

这种问题通常与输入端有关，在整合多来源数据时也有可能遇到，将其处理成一致的某种格式即可

1. **内容中有不该存在的字符**

如在数值字段中却混入了字符，导致在数据类型转换时发生错误。因此该步骤需要对字段进行匹配去掉不该有的字符。

1. **内容与该字段应有内容不符**

出现该问题的原因可能有多种，例如数据列对照错误，数据解析错误。对于内容与字段不符合的情况需要慎重处理。

流程设计

Kafka->streaming->Datashuffle->json->kafka

-

Property{

分隔符号；

字段名（）

字段类型

字段格式

。。。。。。

}

Data class{

名称

类型

格式

处理策略

}

Datashuffle class （data:String）{

接收数据

清洗数据（数据不完整，缺值填充，等）

数据导出 （json）

}