## 文件数据需求

Hdfs位置：172.16.14.31

文件格式：

文件名：

时间戳（10位）+源ip+目的ip+源port+目的port+文件类型(html,doc,ppt, etc.)+”.data”

其中，这6个字段，以下划线”\_”进行分割。

正文：

第一行写入该文件的原文件名。

从第二行开始，写入原文件中的文本内容。

存储路径：

存在hdfs的路径。

/spark/data/files\_data/raw\_data/年月日/时/分

例如：/spark/data/files\_data/raw\_data/20180709/08/02

## 分词（重构中）

### 数据源

Html页面的文本，以及office文件的文本内容。

### 模型算法

1. 分时间段，读取时间段内的所有文本
2. 调用ansj分词库，对文本进行分词
3. 在调用ansj时，加入自定义分词。
4. 将同一时间段内的数据整合成一份，存到hdfs

### 需要重构的内容

1. 转成从hdfs读数据
2. 增加对自定义短语的识别能力（通过加入字典）←2.0a

## 数据分类（重构中，进度预估在80%左右）

### 数据源

已分词的文件

### 模型算法

1. 从hdfs里读取多份数据，整合成一份
2. 使用改进型LDA模型，对数据进行训练。
3. 从LDA模型中，导出每个类别下权重最高的若干个词，并以此定义该类别
4. 从LDA模型中，导出每个词的主题向量。根据主题向量，推算词的主题。
5. 将LDA模型保存到hdfs上。

### 需要重构的内容

1. 性能优化
2. 改进算法以支持自定义类别的训练
3. 提供加入少量新样本后的快速再训练功能
4. 新的模型保存方案

## 文件的静态属性（重构中）

### 数据源

已分词的文件

### 模型算法

1、得出每份文件的五元组和文件类型（一般写在文件名里）

2、得到文件的大小。

3、对文本内容的字符串，计算md5

4、结果存表

### 重构的内容

1、从hdfs上读数据

## 数据分布（重构中）

### 数据源

已分词的文件，LDA模型，url与对应应用表

### 模型算法

1. 推断已分词文件的主题
2. 根据文件的url，推断所属应用
3. 根据文件的md5值，记录文件的流转。

### 重构的内容

1. 从hdsf读数据
2. 加载新的模型格式

## 敏感数据定义（不是我们需要开发的）

### 数据源

敏感词数据库

### 数据格式

1. 用户自定义添加敏感词。
2. 敏感词有3个额外属性：是否敏感，敏感等级，包含范围（标题，或正文，或两者皆有）。
3. 敏感文件检测规则包括2点：仅适用特定主题下的敏感词，仅适用一定敏感等级以上的敏感词

## 发现数据泄露之敏感文件访问下载（重构）

### 数据源

已分词的文件，敏感词数据库，敏感文件检测规则

### 模型算法

1. 加载敏感词数据库
2. 根据敏感文件检测规则，检测敏感文件
3. 把敏感文件报警。

### 重构内容

1、从hdfs上读数据

## 发现数据泄露之违反数据基线（重构中）

### 设计原理

根据数据的类别，统计出用户使用数据的习惯，包括：曾用该类别的次数，在一天的同一时间段中使用该类别的数据量大小。以此构造数据基线。

### 数据源

已分词的文件，LDA模型

### 特征处理

根据用户的历史数据中包含的词，通过LDA计算出词的类别向量，进而计算出数据的类别向量，最终推断出数据的类别。

### 模型算法

构建数据基线：

1. 得出数据的类别
2. 用户对应类别的使用计数上+1
3. 计算当天，该时段内，使用某个类别的数据的总量
4. 把使用某类别的次数，以及历史上，在特定时间段，使用某类别数据的数量，一起作为用户的数据基线。
5. 把用户的数据基线存入HDFS中，待查

检测违反数据基线：

1. 得出数据的类别

2、根据待检测数据的类别，检索该用户历史上使用该类别的次数。如果用户很少或从不使用该类别，则警报

3、统计用户在每个时间段中使用某类别数据的量。如果使用的量超过了历史上该时间段中的量的最大值，则警报。

### 重构的内容

1、从hdsf读数据

2、加载新的模型格式

## 热词、主题、文件、用户的云图（重构中）

### 数据源

已分词的文件、LDA模型

### 模型算法

1. 对文件推断出主题，并对词计算tf-idf，作为词的权重
2. 统计热词、主题、文件、用户的频率
3. 根据频率，分别构造云图。

### 重构的内容

1、从hdsf读数据

## 文件流转（重构中）

### 数据源

已分词的文件

### 模型算法

1. 根据文件的文本内容，计算md5
2. 记录一定时间周期内的所有md5值，以及对应的源ip，目的ip
3. 将这段时间内，每个md5按照出现次数，以及对应的源ip目的ip，构建传播网络。

### 重构的内容

1、从hdsf读数据