# 引言

## 选题背景

## 文献综述

## 研究目的及意义

## 论文主要工作

## 论文章节安排

# 相关技术背景

# 需求分析

## 功能需求分析

### 存储层需求

存储层主要解决数据采集、数据存储、数据流转换等问题。为上层计算层的相应模块提供数据支撑。

#### 研报数据采集模块

**研报数据基本情况**

卖方研究报告由券商研究所发布为主，可以从两个维度进行分类。一个维度是从获取角度，主要包括资讯平台上的公开研报、金融终端订阅的付费研报、分析师团队微信公众号内容。另一个维度是从研究报告的类别，研究报告类别角度又分为两个层次。一个是按照研究领域进行分类，新财富2020榜单将研究领域分为了30个领域；另一个是按照研报的内容，大致可分为行业研究报告、公司研究报告、宏观策略、券商晨报。

从采集器的设计角度，需要对不同数据源和采集频率差别较大的情况分别设计研报数据采集器。

**数据源**

采集模块需要涵盖的数据源包括：1. 东方财富研报中心 2. 东方财富公告大全。

东方财富研报中心，涵盖部分券商的公开研究报告，包括行业研究报告、公司研究报告、宏观研报、新股研报、策略报告、券商晨报。其中宏观研究、策略研报、券商晨报可以归为行业研究报告类型，新股研报可以归为公司研究报告。两类报告的内容在结构上会有所不同，数据处理的流程会有差异，从灵活性出发需要针对性的需要设计两套研报采集器来定期获取两类研究报告数据。

上市公司公告分为定期公告和临时公告，公告发出后，卖方研究会及时发出相应点评。因此对于需要及时获取的事件点评类研究报告，可以根据将公告作为线索来获取点评报告。因此需要设计一个独特的采集器，首先获取公告大全中相应的公告信息，然后根据公告对应的上市公司去及时获取对应公司的事件点评报告。

**数据采集器**

行业研究报告的发布往往没有特定时间点，一般聚焦在某一段时间的行业热点问题，但是为了实现行业热点挖掘功能，必须至少满足日频的采集频率。

公司研究报告的发布，一类是和行业研究报告发布具备同样特性的不定期覆盖报告，另一类是事件驱动的点评类报告，其中以上市公司公告事件为主要。因此针对公司研究报告需要针对性的设计两个采集器，一个用来进行日频的批量采集，一个是根据收盘后公告线索进行针对性的采集。

根据上述对数据源的分析，需要实现三类数据采集器。

1. 行业研究报告采集器（日频采集，覆盖所有研究领域）
2. 公司研究报告采集器（日频采集，覆盖所有上市公司）
3. 公告为线索的实时采集器（以公告为线索，针对性覆盖公告对应公司研报）

#### 研报数据存储模块

**数据结构多样性**

一篇研报的数据由4部分组成：研报摘要、标题及预测等相关信息、研报内容。研报摘要一般是文字形式的非结构化数据；研报内容一般为PDF格式的文件，少数为PPT格式的文件；标题及预测等相关信息为结构化数据。因此需要存储模块能够对文字内容、文件这类非结构化数据以及结构化数据都具有较好的支撑。

**数据存储量**

研报数据的原始形态一般为PDF格式的文件，其中包含大量的图片。平均每篇研报的大小在10MB左右，根据每日500篇研报测算，每日新增存储在5GB。系统应能够存储近前后10年内的所有研报，存储需求大约为40TB。

**一致性需求**

一个研究领域和一家上市公司相关的研究报告应能够尽量存储在同一个表结构中。由于数据源的多样性，和采集器并发的不确定性，需要存储模块能够在性能要求范围内解决一致性问题。

### 计算层需求

计算层基于存储层的数据进行转换，最终为应用层的操作型应用提供数据。

#### 研报内容结构化模块

研报内容一般以PDF文件的形式进行发布，对数据分析应用并不友好。因此需要将研报内容结构化。

研报内容以PDF形式进入数据存储层后，结构化模块对存储进来的研究报告进行处理，将一篇研究报告变成半结构化的数据。

研究报告一般逻辑性较强，描述内容的目的性明确，篇章结构化的转换能够提高研报的精简性同时方便热点挖掘模块的计算。

用途主要有两个：

1. 篇章结构化，为结构化的研报阅读功能提供数据基础
2. 为热点挖掘模块的计算提供数据基础

#### 研报热点挖掘模块

研报热点挖掘模块通过分析研报内容结构化模块处理后的数据，为热点挖掘模块提供数据支撑。

### 应用层需求

**结构化的研报阅读**

**市场热点发现**

## 性能需求分析

### 研报数据采集模块

不同研究报告类型的发布周期与频次有着很大的不同。从买方投研需求角度，一般报告以日频更新即可满足需求，但是对于突发事件的点评类报告却需要足够的实时性，采集的及时性应在分钟级别。

### 研报数据存储模块

### 研报内容结构化模块

### 研报热点挖掘模块

## 可行性分析

# 系统设计

### 整体架构设计

### 存储层设计

#### 研报采集模块设计

#### 研报数据存储模块设计

### 计算层设计

#### 研报内容结构化模块设计

#### 研报热点挖掘模块设计

### 应用层设计

#### 市场热点发现

#### 结构化研报阅读

# 系统实现

### 研报采集模块实现

### 研报数据存储模块实现

### 研报内容结构化模块实现

### 研报热点挖掘模块实现

# 系统测试与验证

## 系统测试概述

## 系统测试环境

## 系统功能测试

# 总结与展望

## 论文工作总结

## 工作展望

# 参考文献

# 致谢