智谱·AI marsoft | 安硕信息 北京交通大学 BEIJING JIAOTONG UNIVERSITY ModelScope 魔搭社区 阿里云

SPM 2023

ChatGLM 金融大模型挑战赛

演讲人姓名：程爽

队员：刘俊、周姿能

nsddd（中国科学院计算技术研究所）

Outline

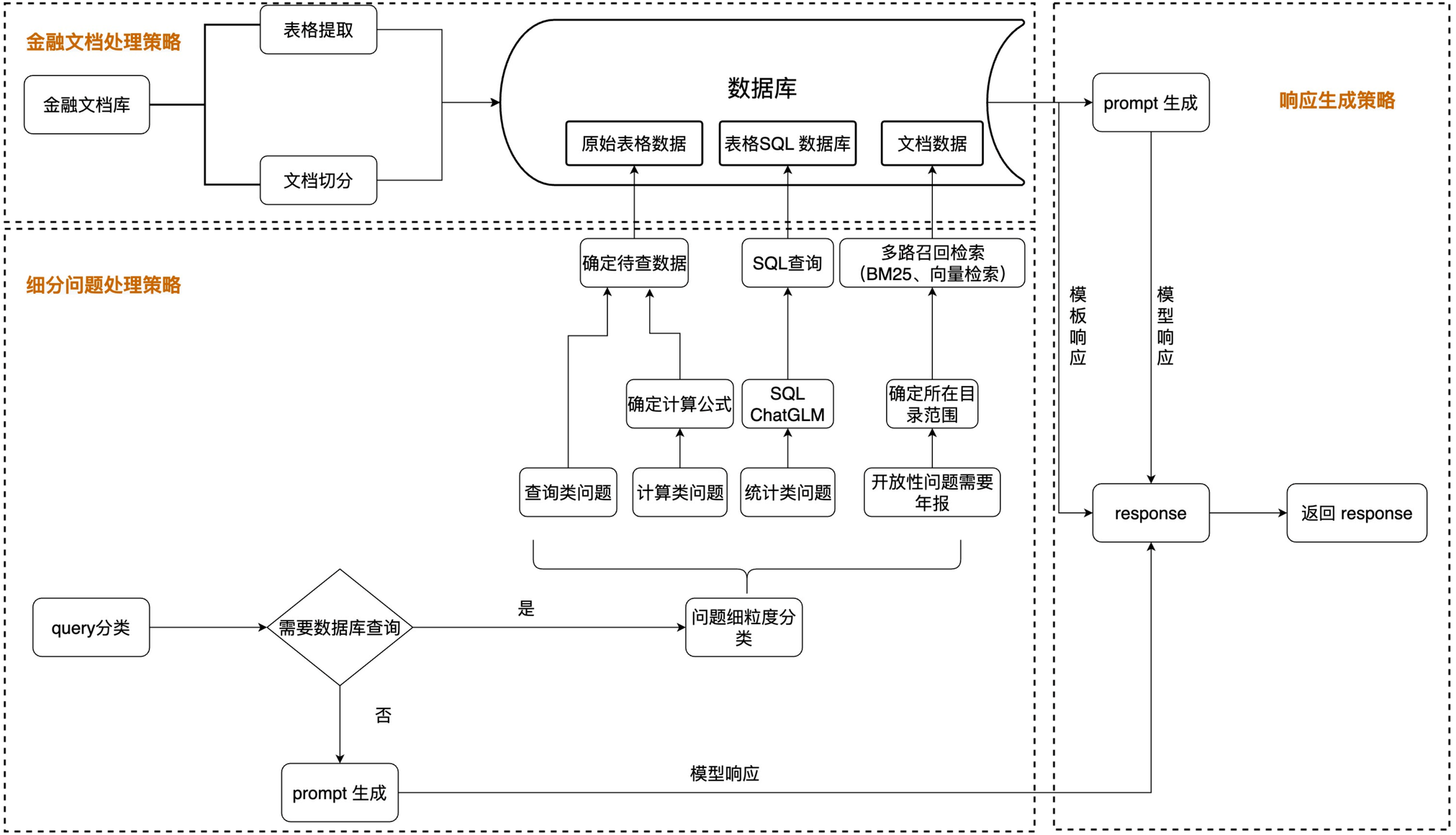
01 方案整体流程

02 金融文档预处理

03 细分问题处理策略

04 回复生成

方案整体流程 总体流程图



方案整体流程

流程设计理念及优势

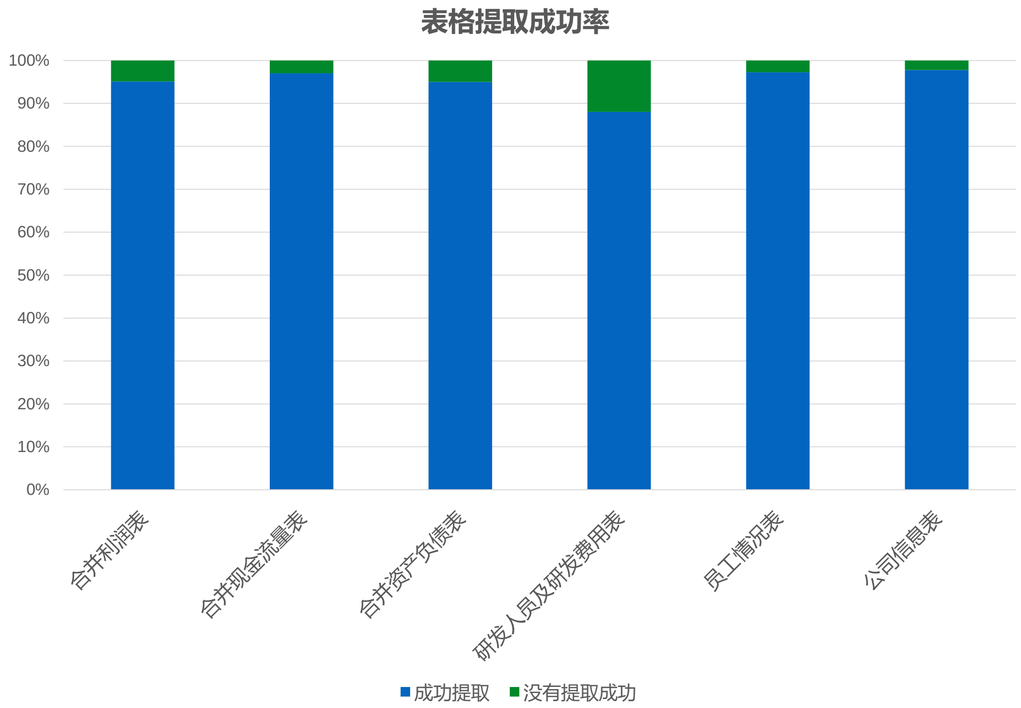
- **流程划分的理论基础**
  - **金融文档预处理**：针对金融文档的特性（如数据密集、专业术语繁多等），设计预处理步骤，将复杂文档转化为易于处理和分析的结构化数据。
  - **问题分类与处理策略**：根据问题的类别和复杂度，采用最适合的处理方法，以提升问题解答的准确性和效率。

- **流程优越性解析**
  - **金融文档预处理**：此步骤有效提升了处理效率，减少了错误和冗余，为后续步骤提供了清晰、准确的输入。
  - **问题分类与处理策略**：此策略实现了针对性的解决方案，显著提升了问题解答的准确性和效率。

金融文档预处理

表格提取及文档拆分

- **表格提取**
  - **基于PDF转html提取：**首先将PDF文档转化为HTML格式，然后采用有限状态机的方法来抽取六种主要的表格（包括合并利润表、合并现金流表、合并资产负债表、研发人员及研发费用表、员工情况表和公司信息表）。
  - **基于PDF转txt提取：**此方法利用TXT格式的上下文内容约束来抽取表格。
  - **整合：**由于以上两种抽取方法可能导致表格信息的部分缺失，我们同时采用这两种策略来确保表格信息的完整性。



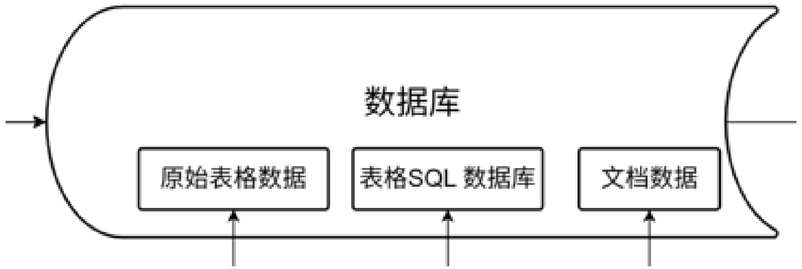
- **文档拆分**
  - **基于目录拆分文档：**此步骤有效提升了处理效率，减少了错误和冗余，为后续步骤提供了清晰、准确的输入。

金融文档预处理

数据库构建

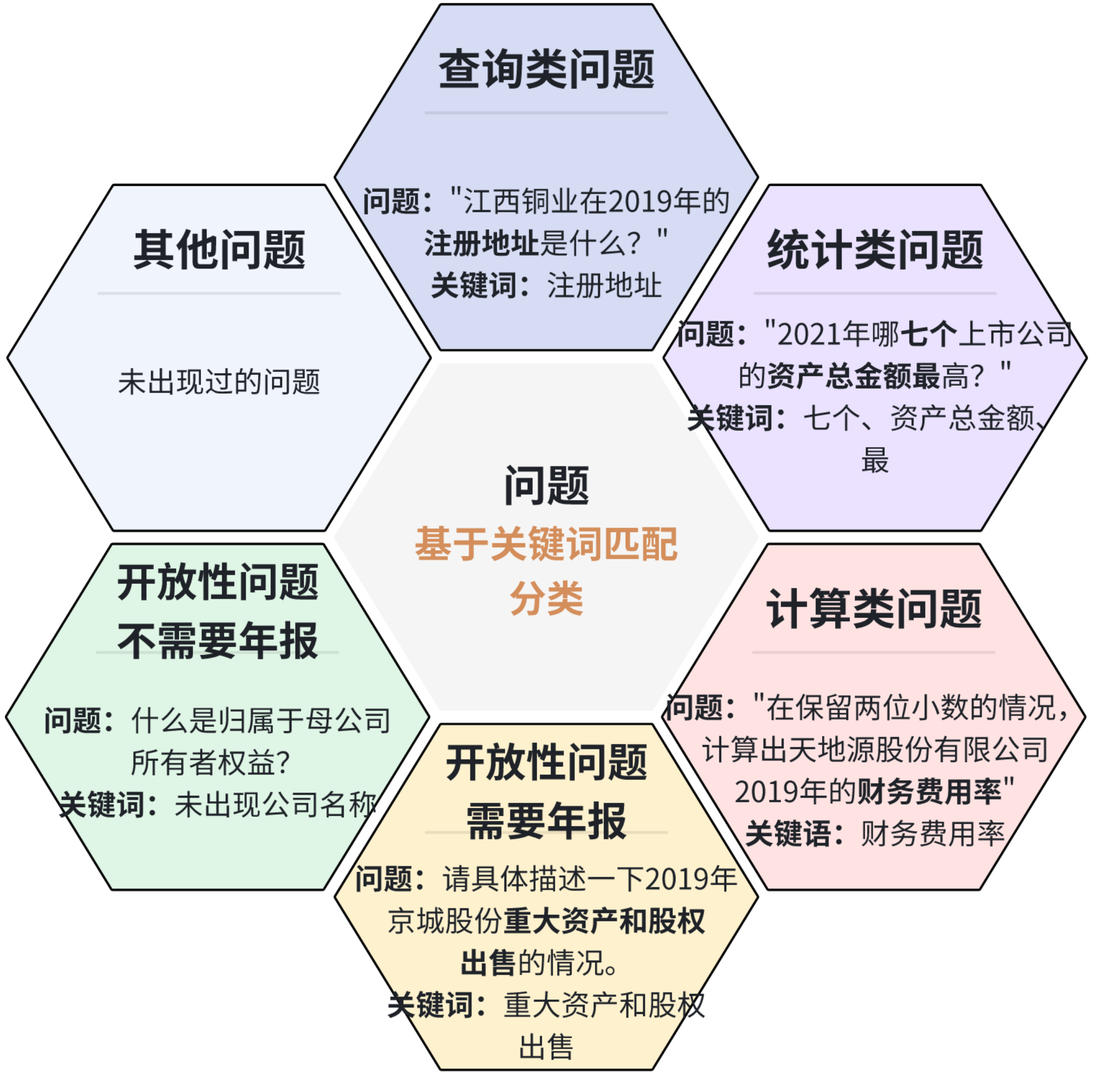
- 数据库组成
  - **原始表格数据：** 包括合并利润表、合并现金流表、合并资产负债表、研发人员及研发费用表、员工情况表和公司信息表 ，回答查询类问题。
  - **表格SQL数据库：**这是一种关系型数据库，它将原始表格数据转化为更加结构化的形式，用于回答统计类问题。
  - **文档向量数据库：**这是一种非关系型数据库，它将数据存储为一系列文档。每个文档都包含多个键值对。这种数据库适用于回答开放性问题。

- 优点及效率
  - **数据整合：**统一管理和查询各种源的数据。
  - **高效查询：**通过SQL快速寻找所需信息。
  - **扩展性：**随着数据增长，数据库容易扩展。
  - **灵活性：**文档向量数据库可以灵活处理各种结构的数据。
  - **降低冗余：**数据库设计避免数据重复，节省空间并提高效率。



细分问题处理策略

问题分类



细分问题处理策略

关键词扩充

- 关键词扩充方案
  - **GPT4扩充关键词：**利用GPT4的强大语言理解能力，对现有关键词进行同义词扩充。这种方法不仅丰富了我们的关键词库，而且提高了模型对不同表述的鲁棒性。
  - **双向扩展关键词：**采用双向扩展策略，即在问题匹配和查询表格两个方向进行关键词扩充。这样做可以更全面地捕捉到问题的信息，同时提高了关键词在分类后查询表格中的匹配度，从而提高分类的准确性。

问题匹配关键词	唯一标识符	查询表格关键词
['博士及以上', '博士', '博士学位', '博士研究生', '博士生', '博士后', '博士人数', '博士数量', '在读博士', '博士学历', '博士员工', '博士及以上的员工人数']	博士	博士及以上,博士以上,博士及以上学历,博士研究生,博士学历

细分问题处理策略

查询类问题

- **关键词匹配**
  - 从“**问题匹配关键词**”库中为问题匹配相应的关键词
  - 采用字典树结构，有效防止关键词的重复匹配
  - 识别多年份问题，对多年份问题采用多年报查询策略
- **表格查询**
  - 查询到的关键词被映射到“**唯一标识符**”
  - 通过“唯一标识符”进一步映射到“**查询表格关键词**”
  - 通过“**查询表格关键词**”查询表格中对应值

问题
2019年四川英杰电气股份有限公司硕士人数是什么?

→ **关键词匹配**
硕士人数

→ **唯一标识符**
硕士

→ **查询表格关键词**
硕士研究生, 硕士研究生, 硕士以上, 硕士及以上, 研究生以上, 硕士学历, 研究生及以上, 研究生, 硕士研究生及以上, 硕士及其以上, 硕士及以上学历, 博硕士研究生学历

细分问题处理策略

统计类问题

- **数据库构造**
  - 将提取的六类表格合成SQL数据库
- **SQL ChatGLM 全参微调**
  - 训练数据集来源：Cspider、DuSQL、NL2SQL：
  - 训练数据QA对构造
- **SQL语句生成**
  - 通过将金融问题按照上述模板构造，训练数据集经过 SQL ChatGLM处理后，可以生成对应的SQL语句。

"query": "你是一个自然语言到SQL转换专家，你的任务是将金融领域问题，转换成对应的SQL查询：生成结果只含SQL语句。\n 问题：哪些城市不属于需要帮扶的贫困城市，并给出它们所在的省。\n 查询需要用到的数据库以及对应的字段如下：表1：城市，可用字段：['城市', '所属省份', '词条id']表2：对口帮扶城市，可用字段：['贫困城市id']SQL查询:",
"answer": "select 城市 , 所属省份 from 城市 where 词条id not in (select 贫困城市id from 对口帮扶城市)"

2019-2021年哪些家上市公司货币总额均位列前十?	[('上海汽车集团股份有限公司',), ('中航工业产融控股股份有限公司',), ('上海建工集团股份有限公司',), ('上海建工集团股份有限公司',), ('新城控股集团股份有限公司',), ('新城控股集团股份有限公司',), ('东方财富信息股份有限公司',), ('新城控股集团股份有限公司',), ('厦门建发股份有限公司',), ('中远海运控股股份有限公司',)]	SELECT 公司名称 FROM fin_report WHERE 年份 IN ('2019', '2020', '2021') AND 货币资金 IS NOT NULL ORDER BY 货币资金 DESC LIMIT 10;

细分问题处理策略

计算类问题

- **建立公式库**
  - 由于金融领域关键词严谨的表述，可以为需要计算的关键词进行标注
  - 为关键词建立公式库
  - 根据召回的关键字匹配公式
- **查询相关指标数值**
  - 和查询类问题查询流程相同

企业研发经费与利润比值=研发费用/净利润
企业研发经费与营业收入比值=研发费用/营业收入
研发人员占职工人数比例=研发人员数/职工总数
流动比率=流动资产/流动负债
速动比率=(流动资产-存货)/流动负债
企业硕士及以上人员占职工人数比例=(硕士人数 + 博士及以上人数)/职工总数
企业研发经费占费用比例=研发费用/(销售费用+财务费用+管理费用+研发费用
营业利润率=营业利润/营业收入
资产负债比率=总负债/资产总额
现金比率=货币资金/流动负债
非流动负债比率=非流动负债/总负债
流动负债比率=流动负债/总负债
净资产收益率=净利润/净资产
净利润率=净利润/营业收入
营业成本率=营业成本/营业收入
管理费用率=管理费用/营业收入
财务费用率=财务费用/营业收入
毛利率=(营业收入-营业成本)/营业收入
净资产增长率=(净资产-上年净资产)/上年净资产
三费比重=(销售费用+管理费用+财务费用)/营业收入
投资收益占营业收入比率=投资收益/营业收入
销售费用增长率=(销售费用-上年销售费用)/上年销售费用
财务费用增长率=(财务费用-上年财务费用)/上年财务费用
管理费用增长率=(管理费用-上年管理费用)/上年管理费用
研发费用增长率=(研发费用-上年研发费用)/上年研发费用
总负债增长率=(总负债-上年总负债)/上年总负债
流动负债增长率=(流动负债-上年流动负债)/上年流动负债
货币资金增长率=(货币资金-上年货币资金)/上年货币资金
固定资产增长率=(固定资产-上年固定资产)/上年固定资产
无形资产增长率=(无形资产-上年无形资产)/上年无形资产
总资产增长率=(资产总额-上年资产总额)/上年资产总额
营业收入增长率=(营业收入-上年营业收入)/上年营业收入
营业利润增长率=(营业利润-上年营业利润)/上年营业利润
净利润增长率=(净利润-上年净利润)/上年净利润
现金及现金等价物增长率=(现金及现金等价物-上年现金及现金等价物)/上年现金及现金等价物

细分问题处理策略

开放性问题

- **开放性问题需要年报**
  - **目录选择**：对年报的目录进行选择。不同的问题可能需要查阅年报中的不同部分。例如，关于公司财务的问题可能需要查阅财务报告部分，准确地定位到年报中的相关部分，提高回复精准度
  - **多路检索召回**：为了提高召回的准确度，采用多路检索召回策略。具体来说，我同时采用向量检索和BM25检索这两种方法进行召回，然后将召回的内容合并。从而提高召回的准确度。

- **开放性问题不需要年报**
  - **Prompt 工程**：尝试不同的提示模板，选择合适的prompt 激活模型在金融领域知识问答的能力。

```
PROMPT_TEMPLATE_4 = """作为金融行业的咨询分析助手，我希望你充当一个经验丰富的企业年报分析专家，熟悉企业企业年报的内容，包括财务报表、经营业绩、风险因素、管理层讨论与分析等方面；擅长财务分析，理解会计原理和财务报表，包括利润表、资产负债表和现金流量表以及股票和债券市场的相关知识。简洁和专业地回答我关于经济和证券的一些问题。

问题是：{question}
答案："""
```

回复生成

输出标准模板构建

模板或 prompt 示例如下：

```
# 数值 单位元 无公式
MATCH_TEMPLATE_1 = """{stock}在{year}的{keyword}是{res}元。"""
# 数值 单位元 带公式
MATCH_TEMPLATE_2 = """根据公式{keyword}={formula}，得出{stock}在{year}的{keyword}是{res}元"""
# 数值 单位% 带公式
MATCH_TEMPLATE_3 = """根据公式{keyword}={formula}，得出{stock}在{year}的{keyword}是{res}%。"""
# 比率
MATCH_TEMPLATE_4 = “” “根据公式{keyword}={formula}，得出{stock}在{year}的{keyword}是
{res}。” “”
……
PROMPT_SQL_TO_TEXT="""
已知一个问题"{question}” 的sql 查询结果为{sql_result}，请重新组织语言回答该问题
"""
```


智谱·AI

marsoft
安硕信息

北京交通大學
BEIJING JIAOTONG UNIVERSITY

ModelScope
魔 搭 社 区

阿里云

Thanks