**SJTU公司**

**立项建议书**

项目名称：医疗场景下的数据流动态分析平台

项目组组号：第11组

项目组负责人：胡昊源

联系电话：18856965019

电子邮箱：1124730278@qq.com

**2022 年 10 月**

**1 项目的必要性**

大数据作为新一轮工业革命中最为活跃的技术创新要素，正在全面重构全球生产、流通、分配、消费等领域，对全球竞争、国家治理、经济发展、产业转型、社会生活等方面产生全面深刻影响。大数据是推动数字经济发展的关键生产要素，是重塑国家竞争优势的重大发展机遇，是实现治理能力现代化的重要创新工具，是建设数字中国的关键创新动力。建立“用数据说话、用数据决策、用数据管理、用数据创新”的管理机制，实现基于数据的科学决策，将推动政府管理理念和社会治理模式进步，加快建设与社会主义市场经济体制和中国特色社会主义事业发展相适应的法治政府、创新政府、廉洁政府和服务型政府，逐步实现政府治理能力现代化。

然而，数据的价值并不是恒定不变的。有些数据具有较强的时效性，在业务发生后不久具有很高的价值，但这种价值会随着时间的推移而迅速减少。因此，数据的处理速度变得尤为重要，实时处理数据能够更快地提供数据洞察。

本项目以数据流动态分析平台为基础，对患者就医事件数据流进行注册、共享和应用管理，实现智能应用的新建、编排、调试、封装和部署，并对数据、应用和计算资源的权限进行管理，同时通过用户流程信息聚合，将用户进行类别划分，并通过流程挖掘形成该类别的标准流程模型，在数据流中对用户分类和流程模型进行持续更新，持续输出用户的分类可靠性、类别特征、标准流程模型和流程模型泛化性有效性等指标，帮助医生更加清晰、准确、全面地了解患者的病情，及时制定合适的治疗方案。

本项目平台还可以向制造业、商业、服务业等其他领域拓展应用。

# 2 项目外部条件落实情况

1. 技术基础

近年来，不断发展的流处理和批处理技术为项目的实现提供了可靠的基础。作为大数据较为成熟的解决方案，下面我们将对流处理和批处理进行简要的介绍。

流处理是指基于分布式技术，对流数据进行动态处理，具有较好的伸缩性、实时性和开放性。分布式流处理通常部署于大规模集群中，通常将流数据处理过程抽象为一个有向无环图。调度算法则负责将有向无环图中的组件合理的分配至集群中的可用服务器上。流处理系统作为流处理作业的平台，负责所有集群资源的管理和分配。对于用户提交的流处理作业，流处理系统需考虑该作业所处理的数据量及集群中不同节点的负载，并将其合理的分配到集群的不同作业节点上，能够加快数据处理的速率，提升系统整体吞吐量，均衡集群中的负载，减少集群中的资源占用率。

批处理非常适合需要访问全套记录才能完成的计算工作。例如在计算总数和平均数时，必须将数据集作为一个整体加以处理，而不能将其视作多条记录的集合。这些操作要求在计算进行过程中数据维持自己的状态。需要处理大量数据的任务通常最适合用批处理操作进行处理。无论直接从持久存储设备处理数据集，或首先将数据集载入内存，批处理系统在设计过程中就充分考虑了数据的量，可提供充足的处理资源。由于批处理在应对大量持久数据方面的表现极为出色，因此经常被用于对历史数据进行分析。

2）研发团队

本项目团队共有四名成员，均具有国内一流大学软件工程专业的学习经历，经过系统的学习，成员们已经掌握了软件开发与管理所需要的各种知识与技术，能够运用科学的方法合理解决软件工程中的问题，并在软件开发与管理方面积累了一定的项目经验。

同时，团队还拥有在数据流分析与应用领域专家老师的指导，共同形成了具备扎实的专业知识、较高的专业素质和综合能力的项目团队，高效协作开发与管理。

3）项目组织机制设计

本项目由上海交通大学软件学院硕士生胡昊源担任项目负责人，按照项目管理规范、CMM 质量管理体系和 SCRUM 的开发模式，全权负责对项目的设计与管理，协调本项目各项工作的顺利开展；由公司董事会对本项目的执行情况进行定期跟踪管理；由公司总裁沈备军博士亲自担任本项目的项目总监，接受项目经理的直接汇报，保证项目在预定时间内获得圆满的成功。

4）市场前景和市场基础

随着工业互联网、大数据、机器学习等技术的发展，数据处理的实时性、智能性成为重要需求。特别是在制造业中，对人机物状态数据的实时监控、特征分析、预测决策，有助于提高产能、降低成本，是我国智能制造、高端制造的关键技术。形成基于数据流的大数据执行平台，可以形成数据处理的范式，为智能分析提供应用高效快速构建的工具。因此，本项目市场前景广阔。

# 3 项目目标和创新点

## 3.1 项目目标

项目采用动态流程分析作为验证，将患者就医事件作为数据流输入，经过预处理后将其进行压缩存储，同时通过用户流程信息聚合，将用户进行类别划分，并通过流程挖掘形成该类别的标准流程模型，在数据流中对用户分类和流程模型进行持续更新，持续输出用户的分类可靠性、类别特征、标准流程模型和流程模型泛化性有效性等指标。

## 3.2 创新点

1）本项目平台是基于数据流的，为企业数据处理的实时性和智能性提供保障，在最大程度上帮助企业优化服务。云计算、物联网等新兴信息技术和应用模式的快速发展，推动人类社会迈入大数据新时代。在大数据时代，数据的时效性日益突出，数据的流式特征更加明显。流式数据蕴含更多信息，能够提炼更丰富的知识，这些知识将在更高的层面、以更广的视角、在更大的范围内帮助用户提高洞察力、提升决策力，为人类社会创造前所未有的大价值。

本项目将数据流作为输入，通过预处理、信息聚合、类别划分、流程挖掘等处理，将数据流充分转化为有价值的、可利用的信息，并在数据流中对相关模型等进行持续更新，用于监控管理、决策分析等服务的优化。

2）本项目为数据处理分析操作提供可视化界面，直观地呈现结果，便于统计与理解。本项目采用流程编排模式将应用涉及的数据特征监控或数据处理等操作进行编排，形成从输入数据流到输出数据的操作管道，对每个操作都提供统计图等形式的可视化展示界面，在调试和应用时都可以查看每个操作输出的可视化结果，使整个数据流分析过程透明化，便于用户监督、理解与后续分析。

3）本项目提供编排式低代码应用构建，方便非技术人员的操作与理解本项目采用的流程编排模式。允许用户根据业务需求，通过简单的拖拽或添加等操作，加入相应的数据处理，同时，也提供算法的管理功能，能够通过上传文件、在线编辑的方式导入或者修改算法，可作为智能分析应用高效快速构建的工具，操作简单，同时满足开发人员和业务人员快速开发的需求，具有多场景多领域的适应能力，便于后续将其推广到不同领域的应用中。

4）本项目允许不同应用之间分享数据处理算法，方便技术人员同样能够高效快速的进行开发，减少开发负担。

# 4 项目方案和可行性分析

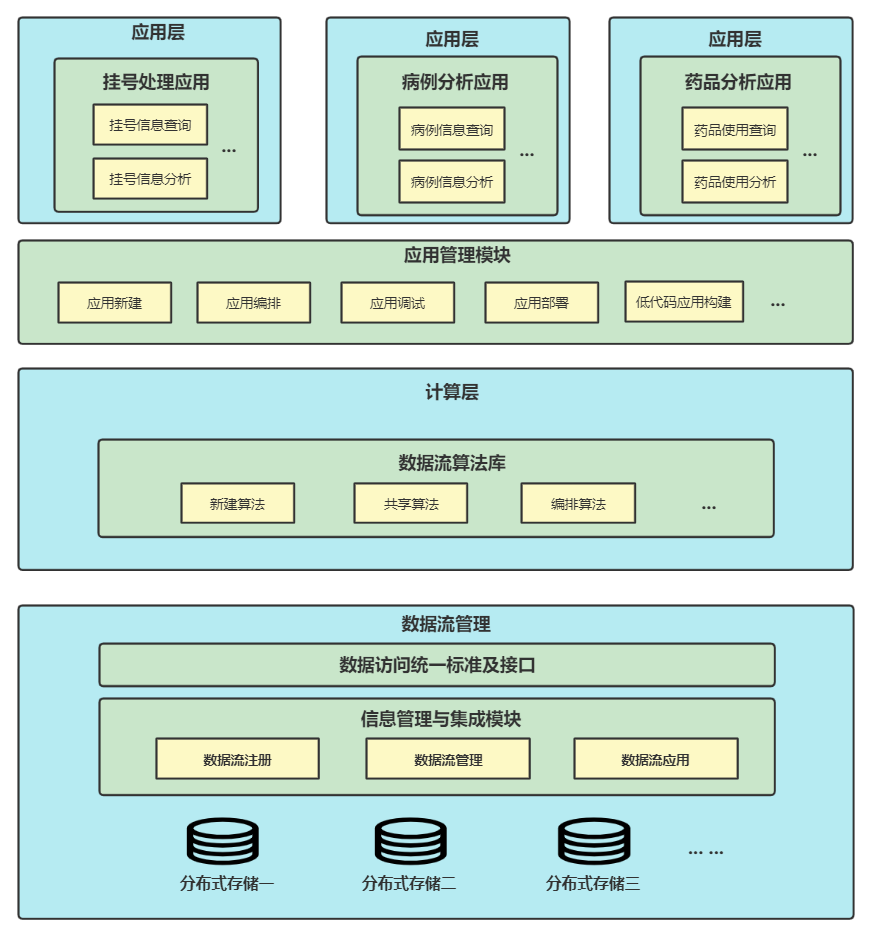
## 4.1 项目前景

随着工业互联网、大数据、机器学习等技术的发展，数据处理的实时性、智能性成为重要需求。特别是在制造业中，对人机物状态数据的实时监控、特征分析、预测决策，有助于提高产能、降低成本，是我国智能制造、高端制造的关键技术。本项目市场前景广阔。

项目由数据流管理、应用管理、权限管理、基础设施管理四部分组成。

对于数据流管理，我们需要实现数据流的注册、共享、应用。对于应用管理，我们需要实现智能应用的新建、编排、调试、封装、部署。对于权限管理，我们需要实现数据权限、应用权限、计算资源权限等。对于基础设施管理，我们需要实现基础算法管理、分布式计算资源管理等。

## 4.2 技术方案



1）开发方法：基于UML的面向对象方法。

2）数据流管理：数据流的注册、共享、应用。

a.数据流注册：可采用 kafka，storm，flink 等流引擎。

b.数据流共享：通过数据权限限定数据流的共享范围。

c.数据流应用：可在数据流上创建 Operator 用于数据特征的监控或数据的处理 Operator 可以通过统计图等方式形成可视化展示界面。

3）应用管理：智能应用的新建、编排、调试、封装、部署。

a.应用新建：包含应用的输入数据流的选择、输出数据或数据流的创建、应用的名称权限的设置。

b.应用编排：采用流程编排模式将应用涉及的 Operator 进行编排，形成从输入数据流到输出数据的 Operator pipeline。

c.应用调试：编排后的应用可以开启调试模式，在输入数据流上观测每个 Operator 的输出，但输出结果不形成输出流。

d.应用封装部署：可将编排后的应用持久化，可采用 docker 格式进行封装，部署后的应用可以开始和停止。

e.应用查看：可以查看应用的 Operator pipeline，集成显示 Operator 的可视化界面。

4）权限管理：数据权限、应用权限、计算资源权限

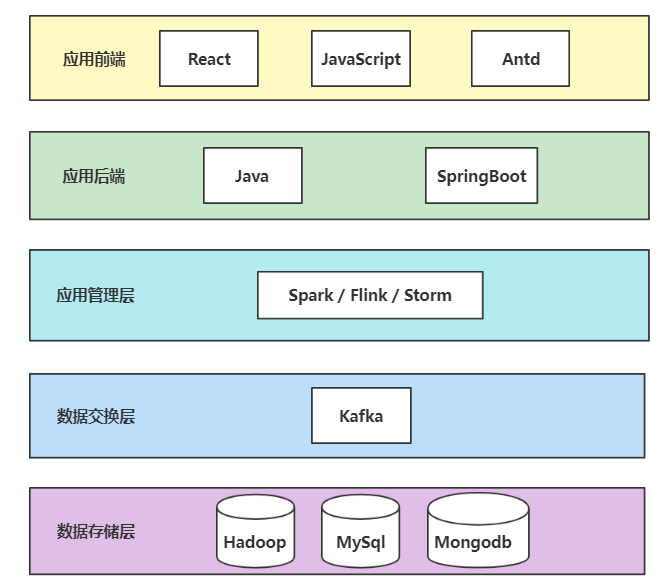
可通过业务组、工作组、应用组等组织结构的划分为各类资源设置权限，也可添加特定人员的权限。

5）基础设施管理：基础算法管理、分布式计算资源管理

a.基础算法管理：可通过上传文件、在线编辑的方式导入算法，可以对算法进行增删改查。

b.分布式计算资源管理：可以查看现有资源使用情况、应用部署情况，可以将应用部署到指定的计算资源上。

## 4.3 技术栈



1）应用前端：React、JavaScript、Antd

2）应用后端：Java、SpringBoot

3）计算与管理：Spark / Flink / Storm

4）数据交换：Kafka

5）数据存储：Hadoop、MySql、Mongodb

## 4.4 可行性分析

1）项目市场前景广阔

当前，随着工业互联网、机器学习、云计算等技术的快速发展，信息技术与

传统产业加速融合，数字经济蓬勃发展，数据已成为经济社会运行不可或缺的关键基础，与此同时，各个行业对数据处理的实时性和智能性的需求也越来越多，对数据进行实时监控、特征分析、预测决策等关键技术的要求也越来越高。

本项目基于数据流，通过信息聚合、类别划分、流程挖掘，形成数据处理的标准模型，为智能分析提供应用高效、快速构建的工具，可以向医疗、制造业、金融业等不同领域的行业推广，具有广阔的应用前景。

2）符合我国相关政策和规划

近年来，我国高度重视大数据在推进社会经济发展中的地位和作用，出台了

一系列政策来推动大数据产业的发展。而本项目形成了基于数据流的大数据动态分析平台，有助于企业对数据进行实时处理、智能分析，在此基础上进行创新和发展，进一步响应国家“数字经济”的发展战略，推动大数据与实体经济的深度融合。

3）项目团队综合能力强，具有专业指导，高效协作。

本项目团队共有四名成员，均具有国内一流大学软件工程专业的学习经历，经过系统的学习，成员们已经掌握了软件开发与管理所需要的各种知识与技术，能够运用科学的方法合理解决软件工程中的问题，并在软件开发与管理方面积累了一定的项目经验。

同时，团队还拥有在数据流分析与应用领域专家老师的指导，共同形成了具备扎实的专业知识、较高的专业素质和综合能力的项目团队，高效协作开发与管理。

4）经济社会效益良好

本项目目前主要针对医疗领域，对患者就医事件进行流程建模和统一分析，

帮助医院更好地利用数据和技术，改善服务，解决问题，提高效率。而随着业务的成熟，本项目产品能够形成数据处理的范式，快速地为企业构建智能分析工具，且高效地投入应用，扩展到医疗、制造业、金融业等不同的领域，具有广阔的市场前景。随着该项目产品的不断推广，帮助企业推动服务转型、激发业务创新，降低生产成本，可创造巨大的经济效益。同时，有助于企业顺应大数据发展潮流，对中国经济的发展，中国企业国际竞争力的提升起巨大的推动作用，具有良好的社会效益。

# 5 计划进度

本课题计划从 2022 年 10 月起至 2022 年 1 月止，用 3 个月时间完成。考虑到课题周期较短且团队规模较小，项目预计采用 SCRUM 的软件开发过程进行开发。课题的风险管理使用主动风险策略。在进行风险识别与优先级排序后，按照优先级设计迭代。

## 5.1 风险分析

通过对有经验的项目组成员与专家进行访谈、对项目的技术框架进行调查以及定期的会议讨论识别出如下表所示的风险。



### 5.1.1 项目风险

1. 需求风险

本项目以医疗业的数据流业务场景为主要的参考依据，其需求有一定概率将无法完全覆盖实际使用的所有需求。因此，本课题的项目需求在后期的开发中可能会不够明确。届时需要在识别到该风险的第一时间重新分析需求，并且尽可能将其产生的影响降低到最小。

随着项目的推进，项目的需求极有可能随着开发人员对技术的深入了解，业务场景的扩大等因素而变更。届时也需要在第一时间对其做出处理，以确保项目的进度能够按照计划推进。

2）进度风险

本项目的工期较短，人力资源也并不充裕，进度风险发生的概率较高。

3）人力风险

小组成员中有部分成员缺少相关的项目经验，并且对项目中使用的技术框架不够了解。需要今早在开发初期解决这一风险，以避免后期编码工作较多的构建阶段的进度无法按时推进。

小组成员在地理上并不处于同一地区，因此会议等交流活动将需要在线上进行。需要小组成员更加积极的在开发过程中进行交流，以避免交流的不足导致的反工。

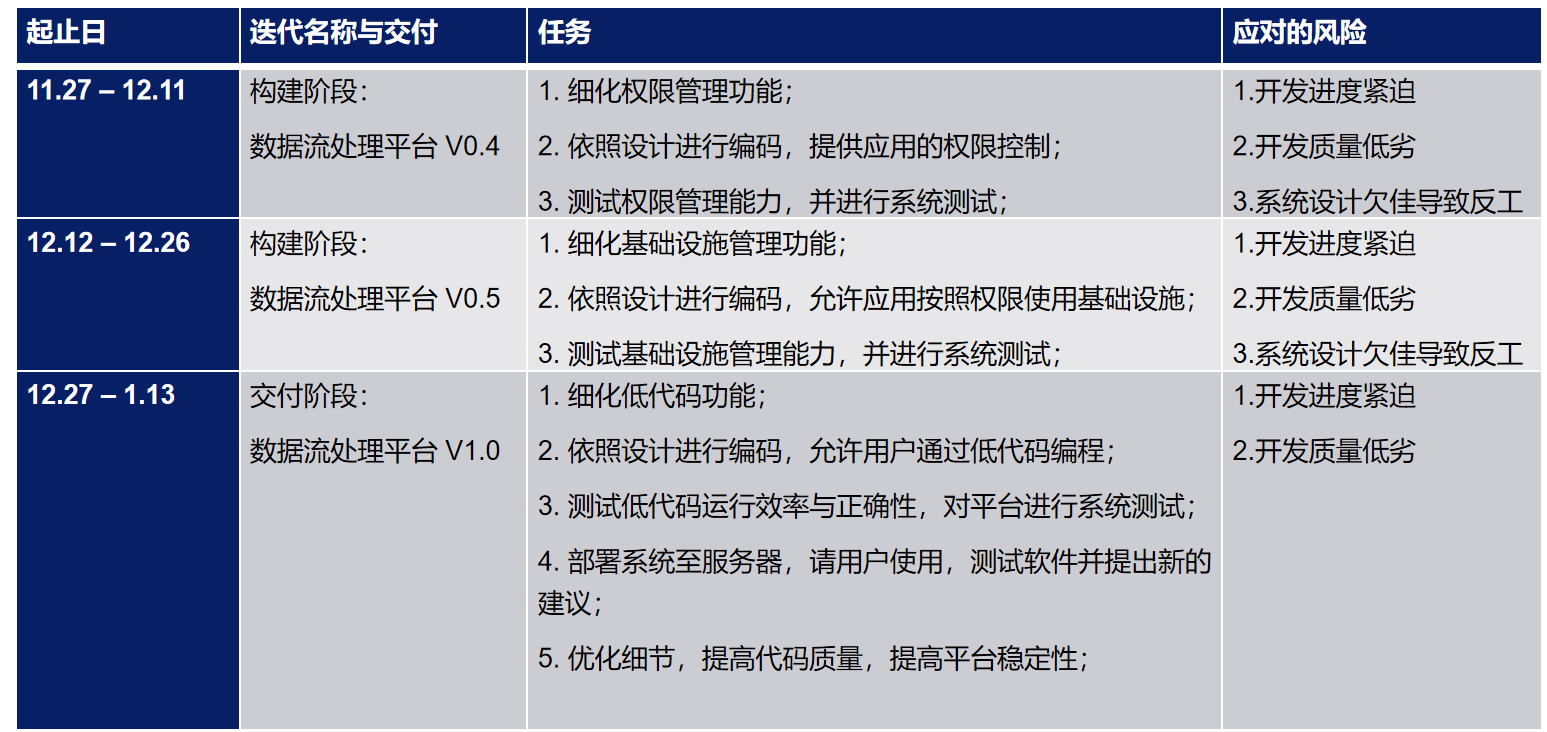
4）技术风险

开发人员对技术框架不熟悉，没有相关开发经验。同时，项目在早期架构上设计欠佳，从而可能导致技术重构。

随着项目的推进，我们可以通过技术预研，快速获取技术经验。同时，通过培训让项目组成员获取相关技术。培训是项目组补充技术缺乏的主要手段，请教有经验的同学或老师，将项目组缺乏的技术及时补齐。

## 5.2 迭代

项目采用SCRUM进行过程开发，结合上述风险与项目周期，以14天为一次迭代设计了如下的迭代计划。

项目分为先启阶段、精化阶段、构建阶段和交付阶段。其中先启阶段不占用迭代时间，主要用于明确项目规模，确定项目边界，制订验收标准，了解环境、主要需求及约束条件，识别关键用例等。精化阶段、构建阶段和交付阶段具体迭代计划如下图所示：

# 6 项目预期成果

- 立项申请书

- 软件开发计划

- 迭代计划，每个迭代一份计划

- 迭代评估报告，每个迭代一份报告

- 软件需求规约文档

- 软件架构文档

- 源代码

- 软件测试计划

- 软件安装包

- 项目总结报告

# 7 项目社会经济效益

随着工业互联网、大数据、机器学习等技术的发展，数据处理的实时性、智能性成为重要需求。特别是在制造业中，对人机物状态数据的实时监控、特征分析、预测决策，有助于提高产能、降低成本，是我国智能制造、高端制造的关键技术。形成基于数据流的大数据执行平台，可以形成数据处理的范式，为智能分析提供应用高效快速构建的工具。