项目名称：为 openEuler 创建用户轨迹运营看板

一 申请理由：

黄成，成都信息工程大学，信息与计算科学专业，中国共产党党员。跨专业学科，所以在数学和计算机两方面都有所接触。除了我们学过的 Java，我还自学了计算机专业的主干科目，包括计算机网络，数据库等，课余时间出于对开发的热爱，自己摸索用 Java 写过推箱子、贪吃蛇小游戏；自学 python，做过爬虫项目，数据处理项目，目前正在学习《代码整洁之道》和复习《python 编程快速双手——让繁琐工作自动化》，在进行 ACM 比赛中，形成了扎实的 coding 能力。

做过一些传统算法的项目，如眼电信号项目。算法是将前人抽象的，理想的数学概念实体化，解决现实生活中遇到的问题；相比较带我入门的算法项目，我更愿意参与开发项目，开发就是创造，创造让我体会到极大满足。

国内互联网产业已经到达世界巅峰，但是在技术方面，开放创新还和欧美有所差距，希望自己能够以此为契机，走上开源的道路，做时代浪潮的一滴水，为社区的进步添砖加瓦，从社区成就自己到自己成就社区。

二 技术方案：

1.主要功能：

（1）实现用户数据采集，存储，对用户行为进行展示，包含数据采集，OBS 选择,CSS RAW、CSS Enrich 搜索，面板展示，管理员提醒。

（2）实现社区管理自动化，减少运营人员工作量，提高项目推送的准确性，增强用户的参与感和满意度。

2.拟解决的关键问题：本项目针对现阶段主要问题是数据的选取，数据字典，OBS 选取，面板选取，连接代码。

3.解决途径：

通过现有查阅的资料学习、导师指导和了解业界、友商的常用模型。资料来源于博客、手册或使用文档。

4.项目研究的主要内容（模块拆解）：

（1）用户识别：

通过对用户的沉浸度进行区分，对不同用户采取不同强度、不同重要程度的项目推送，有利于提高推送的精确度，用户的满意度与参与度。参考【1】中关于度量框架的讨论文档【2】的可以将用户分为普通用户，contributor,mentor 等。

领域划分：

社区涉及各种各样的开发，用户对不同方向的开发兴趣程度不尽相同，既然

对用户进行了划分，自然可对项目类型做出划分，达到辅助推送的目的。本次Summer2020活动中，OpenEuler社区参加项目多，覆盖范围广，可以参照工作人员对本次项目的分类对整个OpenEuler项目分支进行分类，如Linux桌面相关项目、树莓派相关项目等。【3】

（3）数据选取：

通用的数据包括用户git数据，用户issue，用户fork等数据，对于尝试访问OpenEuler社区的用户可记录访问博客类型（OpenEuler社区暂时无注册功能），参考Stack Overflow中Emilien Macchi activity report，还可以增加诸如Patch sets、Draft Blueprints等数据【4】

（4）OBS选取：

暂定采用ElasticSearch。考虑到数据量以及展示所需的实时性要求，要求数据库必须满足分布式的文档存储、分布式的搜索和分析引擎，参考GitHub【5】，Adobe【6】等有同样需求的网站，同时根据华为云云搜索服务本身也提供了elasticsearch搜索引擎【7】。

（5）面板展示：

暂定采用Grafana。面板的选取通常与OBS的选取相关，在传统ELK体系中，常常采用Kibana作为elasticsearch的web前端展示，由于kibana图形化不完善，没有权限、用户管理等缺陷，不能满足运维的要求，目前Grafana已经有elasticsearch的插件，Grafana更加灵活，提供的功能更加丰富。同时针对Grafana的二次开发软件如Huawei OceanStor metrics in Grafana【9】也可以作为面板的参考样式。

（6）连接代码:

暂定采用python。Python 是一种解释型、面向对象、动态数据类型的高级程序设计语言。其胶水语言的特性，能够快速生成程序的原型或者最终界面。作为这几十年新兴的语言，在系统编程、图像处理、数据库编程等场景下都有广泛的应用，其强大的第三方库，更是为其扩展性提供了无限的可能，所以python作为该项目的第一考虑。在获取数据时暂定使用requests模块、os模块、beautifulSoup模块，与elasticsearch连接时使用elasticsearch模块，数据建模时使用numpy模块和pandas模块，向管理人员发送邮件使用smtplib模块，发送短信需要根据实际使用的平台确定。

（7）运行环境：

暂拟Linux om-mindspore 4.15.0-65-generic #74-Ubuntu SMP。所有软件和代码都需要能在服务器上运行。

【4】https://www.stackalytics.com/report/users/emilienm

【5】https://github.com/collections/projects-that-power-github

【6】
https://www.elastic.co/cn/elasticon/tour/2018/santa-clara/elastic-at-adobe-making-search-smarter-with-machine-learning-at-scale

【8】https://support.huaweicloud.com/css/index.html

【9】https://www.kruyt.org/oceanstor_grafana/

三 时间规划：

第一阶段：方案确定。

7 月 1 日-7 月 7 日：根据项目需求，与导师讨论相关技术栈，了解大概实施路线。

7 月 8 日-7 月 14 日：结合与导师讨论结果，自行学习相关知识，同时确认方案的可行性。

7 月 15 日-7 月 21 日：再次与导师讨论，沟通学习成果，进行技术交流。

7 月 22 日-7 月 28 日：输出设计方案文档。

第二阶段：前期开发。

7 月 29 日-8 月 12 日：开始数据收集，完成后同时开始 OBS 和数据建模。和导师交流遇到的问题。

8 月 13 日-8 月 15 日：交付数据字典报告，建模流程报告，遗留项处理计划，中期报告。

第三阶段：后期开发

8 月 17 日-8 月 21 日：横向比较其他项目面板样式，评估自身样板。根据实际采集的数据同导师交流细化面板样式。

8 月 22 日-9 月 2 日：产出面板代码，管理员提醒代码。

9 月 3 日-9 月 9 日：将代码同导师交流，进一步完善包括备份、异常、崩溃等情况处理。

9 月 10 日-9 月 16 日：完善本地代码，更新 Gitlab 平台，完成收尾。

第四阶段：改进和总结

9 月 17 日-9 月 18 日：验收成果，代码 review，总结经验。

9 月 24 日-9 月 30 日：上交代码，完成面板使用报告，项目文档，结项报告。