arXiv:2305.04258v1 [cs.DB] 7 May 2023

从非结构化到结构化：将聊天机器人对话转换为用于可视化的数据集市模式

Mark Edward M. Gonzales 德拉萨大学，菲律宾马尼拉
mark_gonzales@dlsu.edu.ph

Elyssia Barrie H. Ong 德拉萨大学 马尼拉，菲律宾
elyssia_ong@dlsu.edu.ph

Charibeth K. Cheng 德拉萨大学 马尼拉，菲律宾
charibeth.cheng@dlsu.edu.ph

Ethel Chua Joy Ong 德拉萨大学 菲律宾马尼拉
ethel.ong@dlsu.edu.ph

Judith J. Azcarraga 德拉萨大学 菲律宾马尼拉
judith.azcarraga@dlsu.edu.ph

摘要

学校是开展公共医疗保健干预的主要途径之一。由于资源有限，菲律宾公立学校难以常规开展健康与 wellness 检查，因此部署聊天机器人辅助健康监测系统或可提供一种替代方案。然而，由于自然语言表达丰富，输入内容存在差异，从原始对话中获取洞察并非易事。本文提出了一种将非结构化对话转换为结构化模式的流程。该流程包含四个阶段：*(i)* 通过实体抽取和数据聚合处理对话，*(ii)* 将其作为 NoSQL 文档存储在云端，*(iii)* 将其转换为用于联机分析处理的星型模式，并构建抽取-转换-加载工作流，以及*(iv)* 创建基于 Web 的仪表板，以可视化汇总数据和报告。该仪表板的性能评估显示，存储的对话数量增加 10_5 倍后，上卷、下钻和筛选结果的显示加载时间仅增加约一秒。

CCS 概念

• **信息系统** → **数据仓库**；提取、转换和加载； • **以人为中心的计算** → 信息可视化；

关键词

医疗保健聊天机器人、对话处理、数据库设计、联机分析处理（OLAP）、数据可视化

ACM 参考格式：
Mark Gonzales、Elyssia Barrie H. Ong、Charibeth K. Cheng、Ethel Chua Joy Ong 和 Judith J. Azcarraga。2023。《从非结构化到结构化：将聊天机器人对话转换为用于可视化的数据集市模式》。载于《菲律宾计算机科学大会论文集》（*PCSC2023*）。菲律宾马尼拉，8页。

允许为个人或课堂使用而免费制作本作品的部分或全部内容的数字副本或纸质副本，但不得为牟利或商业利益而制作或分发副本，且副本必须附有本声明及首页的完整引文。本作品中第三方组件的版权必须得到尊重。其他用途请联系所有者/作者。*PCSC2023*，*2023*年*3*月，菲律宾宿务市 © 2023 版权所有者/作者所有。

1 引言

促进小学生的健康与福祉是联合国可持续发展目标 [30] 的重点之一。在菲律宾，学校是开展公共卫生干预的首要场所 [36]，全国和地区项目（例如基本医疗保健计划 [28],教育部健康行动计划 [33], ）以及免费疫苗接种举措 [21], ，均旨在控制学龄儿童可预防疾病的发生率。

尽管做出了这些努力，公立小学开展常规健康评估仍面临一项重大障碍，这可归因于人手不足，因为通常的健康监测方式需要护士或医生协助检查身体系统。此外，学生健康记录通常以纸质文档形式保存；随着时间推移，数据不断累积，这些记录可能变得零散，也难以快速检索、更新并在医疗保健服务提供者之间共享 [4, 41]。

部署自动化健康监测系统，有助于在没有医务人员在场的情况下开展常规健康检查，并将收集到的信息存储在云端。由于儿童可能不习惯使用传统 WIMP（窗口、图标、菜单和指针）界面中高度结构化的操作路径，将聊天机器人集成到该系统中，或可利用聊天机器人拟人化的特点，为用户提供一种熟悉的方式，使他们能够更自由地进行交互和表达自己 [42]。

除了采用具备语音和文本处理能力的聊天机器人来解读音频输入并生成智能响应外，健康监测系统还应从儿童与聊天机器人的对话中提取洞察，在云端更新他们的健康信息，并为医疗专业人员提供分析结果和数据可视化。要实现这些功能，需要一个兼顾对话非结构化特性、自然语言输入表达力以及数据库设计的流程。该流程包括将非结构化对话转换为用于数据可视化的结构化模式。具体而言，该流程旨在提供一种方法来：

- 通过执行实体抽取和数据聚合来处理对话；
- 将处理后的对话存储在云端；

- 将存储的对话转换为便于联机分析处理（OLAP）的模式，并构建自动化的抽取-转换-加载（ETL）工作流；并且
- 创建一个基于 Web 的可视化仪表板，用于汇总儿童与聊天机器人对话的数据。

2 相关研究

聊天机器人是能够使用自然语言与人类用户互动，从而模拟连贯对话的计算机程序 [1, 20]。针对医疗保健领域设计的聊天机器人，能够帮助临床医生轻松整理数字健康记录、检索患者数据，并提供有关已观察到的副作用和药物相互作用的信息 [42]。除了协助医疗专业人员外，有些聊天机器人还用于个人健康监测和管理 [7]。

2.1 对话处理与管理

由于聊天机器人参与基于语音的通信，它们会在对话管理中采用动作选择方法来处理输入语句并返回适当的响应 [9, 24],，而医疗聊天机器人通常会使用医学实体识别 [14] 和语义相似度度量。Quro [14],是一款个性化医疗助理，它利用医学实体识别从用户输入中提取症状；医学实体指任何医学概念（例如体征、症状、疾病或药物），并检测这些实体之间的语义关系。

近年来，对话管理开始采用语义相似度度量，以替代传统的模式匹配方法 [2, 8], ，从而无需创建脚本模式，避免耗时且容易出错。Adeh 等人 [2]实现了一种模糊句子相似度算法，可接收两个短语或短文本，并根据给定的句法和语义要素，计算二者在含义上的相似度。与基于语义网络和语料库统计的成熟相似度度量相比 [26],，研究发现，采用模糊句子相似度有助于改进规则匹配；不过，该方法在处理含有“不”字的短语时效果不佳，因为这会导致某些规则误触发。这表明，对于具有否定含义的词语，需要特别处理。

模糊匹配常用于医疗聊天机器人，以实现多种目的，既可以简单地改善词语识别，使其不受拼写或排版错误的影响 [27], ，也可以像 IBM Watson Assistant 一样，结合网络钩子和支持向量机来识别医学实体 [35]。

2.2 医疗保健系统数据库设计

医疗保健系统管理着大量异构数据。尤其是随着数字医疗保健数据不断增加，数据的体量和类型成为显著的考量因素。必须从患者处收集分散的信息，因此需要将数据提取并汇总到一个存储库中 [10]。

在此背景下，数据集市和数据仓库可能会发挥重要作用，因为它们是由不同数据源构成的存储库，所有数据源均按照统一的模式进行组织，以生成有助于战略决策的洞察 [38, 39]。由于星型模式具有其 ....., 因此在医疗保健数据管理研究中经常采用 [6, 29, 40] 。

简洁性，从而实现有效的查询处理、快速聚合，并由于连接操作更少而缩短查询执行时间。由于这些模式由一个中心事实表与多个维度表相连接构成，因此维度表的反规范化会引入冗余，进而增加存储需求 [34]。

然而，这些缺点与相对较低的存储成本 [29] 以及用户偏好在大量数据中更快执行查询操作 [6, 22]相比，影响较小。这些特点使星型模式成为更高效执行联机分析处理操作的合理设计选择 [40]。

2.3 医疗保健数据可视化

数据可视化是一种通过对定量数据进行视觉编码来呈现信息的方法 [43], ，使决策者能够更轻松地解读结果并识别模式。因此，可视化被广泛用于向专业人士和大众传递科学信息，包括健康数据。

鉴于可用的数据可视化形式种类繁多，一些研究探讨了目标受众在接触健康相关数据时对呈现形式的偏好。Fortin 等人 [13]研究了女性对如何呈现心脏病、髋部骨折和乳腺癌风险的可视化偏好。他们发现，大多数受访者比起其他传统数据可视化形式，更偏好简单的条形图。

同样，Dolan 和 Iadarola [12] 研究了六种用于呈现癌症及其预防相关信息的可视化形式：增强型条形图、图标阵列、流程图，以及上述可视化形式的另外三种不同组合。研究发现，虽然条形图和流程图是首选形式，但与单独呈现一种可视化形式相比，组合呈现多种数据可视化类型更为可取。

更具创意和复杂的数据呈现格式不断发展，也促使研究者探讨交互性对可视化易懂程度的影响，不过专门关注医疗保健相关数据呈现方式的研究仍然有限 [43]。Ancker、Chan 和 Kukafka [3] 评估了使用图标阵列展示风险信息的交互式图表的可用性。他们的研究结果表明，交互式图表不仅能够呈现风险大小，也能传达与风险相关的感受。这与一些研究的观点一致：通过计算机媒介进行交流时，较高的交互性可能会对用户处理信息的方式、对健康的态度以及对健康相关消息的情感投入产生积极影响 [43]。

3 设计与实现

该聊天机器人使用 DialogFlowCX [17], 开发，这是 Google 于 2021 年推出的一项基于云端的服务，用于设计对话代理。图 1 展示了如何将对话流程模块化为多个较小的对话流程，分别对应 Heartland Community Health Center 改编版《儿科系统回顾》的各个部分 [11]。每个较小的对话流程又由多个相互连接的页面组成，其中包含根据《儿科系统回顾》设计的问题；图 2 展示了一个用于询问过敏症状的示例。

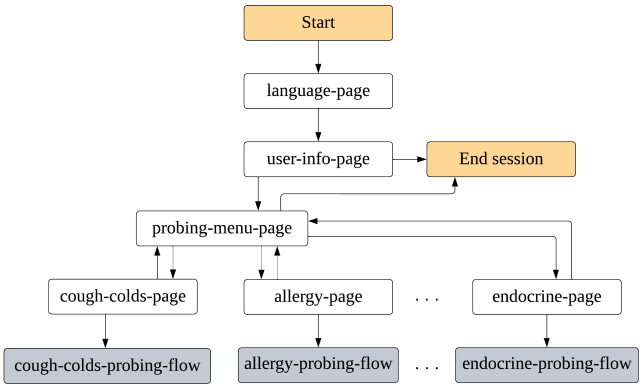

图 1：对话流程的模块化。与聊天机器人的会话始于询问孩子的首选语言和相关信息（用户名及数据隐私同意）。如果孩子不同意，会话便会终止。否则，聊天机器人会继续进行身体系统检查，这一过程在后台由多个模块（灰色）处理。

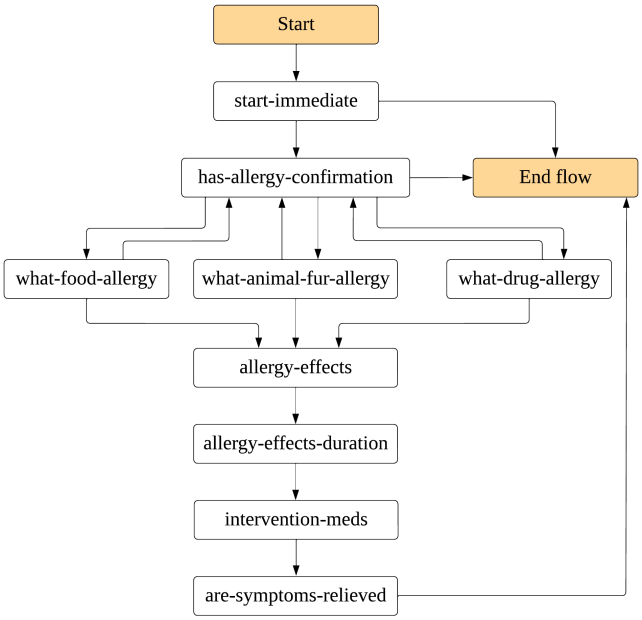

图 2：过敏探查模块的对话流程。DialogFlowCX [17] 将每个流程建模为有限状态机，其中页面是状态，状态处理程序是转移。流程启动后的中间页面用于设置聊天机器人与知识库之间通信的参数。

本节其余部分将介绍将该聊天机器人的非结构化对话转化为用于数据可视化的结构化模式的各个阶段。图 3 展示了框图，涵盖从处理和存储对话，到将其转换为数据集市模式，最后创建用于数据可视化的仪表板的过程。

列表 1：包含实体抽取的对话示例。*C* 表示聊天机器人，而*U* 表示最终用户（孩子）。以最终用户的意图为上下文抽取的实体以粗体显示。

C：你对食物过敏吗？U：**有**C：你对哪些食物过敏？U：**虾** 还有**鸡蛋**C：你对动物毛发过敏吗？U：**不**好 C：你过敏时有什么感觉？U：**呼吸困难**还有**起疹子**C：你用什么药来治疗过敏？U：妈妈会给我涂**软膏**

表 1：接受数据聚合的词语关系。每组词语的参考值以粗体显示。

关系	词语
正字法变体	**肚子**，tyan
拼写错误	**青霉素**，青霉素，青霉素
词形变化	**鸡蛋**, 鸡蛋
同义词	**是的**，是啊，是的，当然。
上位词/下位词	**海鲜**，螃蟹、贻贝、贝类
命名实体	**维生素** , Tiki Tiki , Poten-Cee
译文	**擦伤**，擦伤
省略短语	**3**，3岁，3岁
（语境：年龄）	
语境同义词	**澄清**、无色、白色
（上下文：尿液颜色）	

3.1 对话处理

该流程的第一阶段是处理对话。由于机器人与儿童对话包含开放式问题（例如，你对什么食物过敏？），因此必须考虑自然语言输入的表达性。

3.1.1 自然语言的表达性。 对话处理组件应能够支持两项基本的自然语言理解（NLU）功能：实体抽取和数据聚合。实体抽取涉及根据儿童话语的意图，在其话语中识别关键词或短语。列表 1 给出了聊天机器人与子项之间的一段示例对话；实体以粗体显示。

数据聚合涉及将对话上下文中被视为等价的不同词语或短语进行分组，并为其分配一个统一的参考值。表 1 列举了必须进行聚合的词语关系类别示例；每组词语的参考值以粗体显示。

3.1.2 答案整合器。 为执行实体抽取和数据聚合，我们使用 DialogFlowCX 内置的 Enti- ties 功能 [18] 开发了一个答案整合器。实体条目及其等价词（在 DialogFlowCX 中称为实体同义词 ）列表

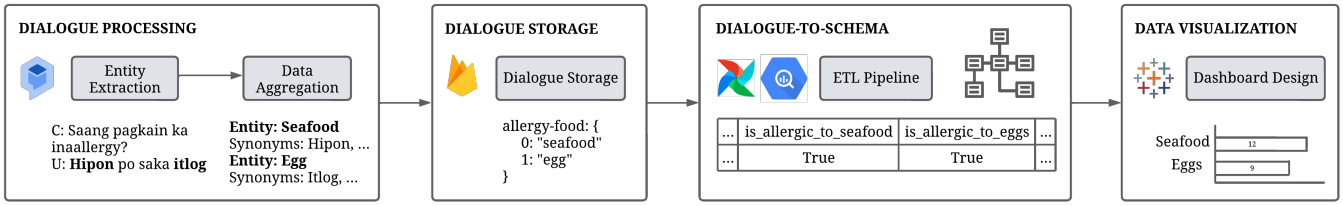

图 3：流程框图。该流程由四个阶段组成：*(i)* 使用 DialogFlowCX 处理对话，*(ii)* 将其存储在 Cloud Firestore 中，*(iii)* 将其转换为星型模式，数据库存储在 BigQuery 中，并在 Apache Airflow 中构建和运行抽取-转换-加载工作流，以及*(iv)* 创建 Tableau 仪表板以进行数据可视化。

表 2：食物过敏实体类型的实体条目

参考值	选定实体同义词
nuts	坚果、花生、木薯
乳制品	乳制品、奶酪、酸奶、冰淇淋、牛奶、奶酪、奶酪、牛奶
egg	鸡蛋、鸡蛋
海鲜	海鲜，螃蟹， 贻贝，贝类，虾，海鲜，螃蟹，贻贝，虾

为每种实体类型手动构建了相应内容，以涵盖表 1 中的词语关系。表 2 给出了食物过敏实体类型的示例，其中包括坚果、乳制品、鸡蛋和海鲜等参考值。

实体条目和同义词列表主要基于对227名6至7岁儿童进行系统儿科评估时记录的对话，这些评估由合作护士协助完成。随后，担任领域专家的护士对列表进行了扩充。

答案整合器通过与实体条目进行模糊匹配，从儿童话语中提取实体。为此，会调用 DialogFlowCX 的模糊匹配功能 [19]。随后，将提取出的实体映射到其参考值，以进行数据聚合。如果模糊匹配返回空集，则会触发 sys.no-match-default 事件，聊天机器人会再次提示儿童作答。

正如 Liu 和 Sundar 的研究 [27],以及 HealthConsultantBot [35], 所展示的那样，与精确匹配或基于正则表达式的匹配相比，采用模糊匹配的优势在于，它能够在一定程度上处理拼写错误、正字法变体和词形变化，而无需手动将这些形式添加到实体同义词集合中，也无需创建硬编码的正则表达式模式；随着聊天机器人的复杂性增加，这些模式可能会变得难以管理或不再适用。

这种方法还能够处理多词实体，这一点尤其重要，因为典型的他加禄语和比萨亚语句子结构会将代词主语置于否定词和动词之间。例如，*hindi ko po alam* 和*wala kokabalo* 会与*hindi alam* 和*wala kabalo*进行模糊匹配，而这两者都映射到参考值*don't know*。

然而，对于编辑距离较小的词，它可能会产生误匹配；例如，*hipon*（虾）可能会与 *sipon* 匹配

表 3：列表 1 中示例对话的文档表示

字段	值
会话_标识符	123456789
用户名	juan-dela-cruz
sex	M
数据_隐私_同意	True
过敏_食物	海鲜，鸡蛋
过敏_动物_皮毛	None
过敏_毛毡	呼吸困难、皮疹
过敏_干预	软膏

（感冒），前提是后者已是实体中最接近的近似匹配。为减轻这一问题，我们手动将拼写相似但对应不同参考值的词添加到实体条目和同义词列表中。对于应与精确值匹配的响应（例如用户名），也关闭了模糊匹配。

3.2 对话存储

该流程的第二阶段是将每段处理后的对话保存为 JSON文档，并存储在托管于 Cloud Firestore 的集合中。Cloud Firestore [16]，是一种可扩展的 NoSQL 数据库，可与其他 Google Cloud Platform 服务（包括 DialogFlowCX）无缝集成。该文档中的字段-值对对应一个问题，以及答案整合器确定的实体参考值，用于记录子项的响应。

由于对话本质上是一系列问题和响应，因此，与关系数据库 [39]基于表格的方式相比，将其建模为由字段-值对构成的 NoSQL 文档，可以实现直接且灵活的表示。此外，某些字段（例如食物过敏，以及过敏反应期间出现的症状）应能够容纳多个值。对于 JSON文档而言，这并非问题；但关系数据库对多值数据类型的支持因数据库管理系统（DBMS） [20]而异。

表 3 展示了列表 1 中示例对话的文档表示中相关的字段-值对；其中还包括用于存储用户和会话信息的字段。由于每个文档都对应于与聊天机器人的一个会话，因此 Cloud Firestore 生成的文档 ID 同时也充当会话ID。

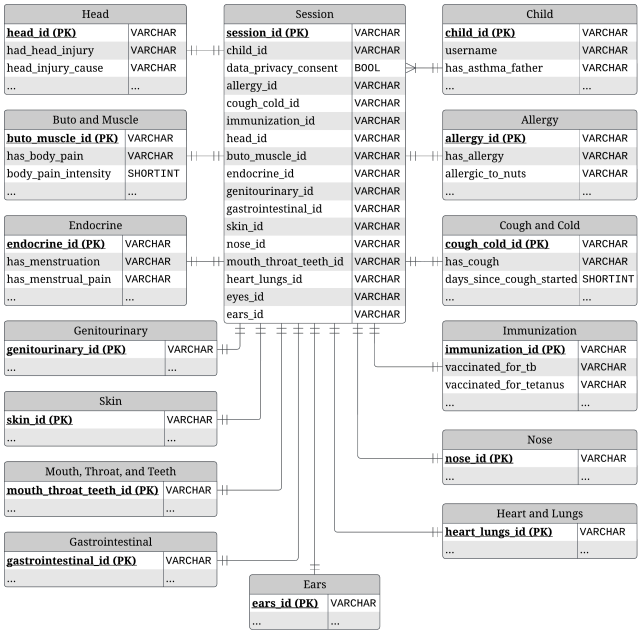

图4：用于将处理后的对话转换为联机分析处理的星型模式。将 Cloud Firestore 中的集合拆分为这些表，旨在与 DialogFlowCX 中对话流程的模块化保持一致。大多数列采用可变长度字符串（可变长度字符型）作为数据类型，以容纳是、否和不知道等回答。

3.3 对话到模式转换

该流程的第三个阶段是将存储的对话转换为用于联机分析处理的结构化数据集市模式。为了充分发挥关系数据库的适用性以及 SQL 在执行 OLAP 操作方面的表达能力，我们设计了星型模式和 ETL流水线。

3.3.1 星型模式设计。数据库设计采用星型模式。如图4的实体关系图所示，该模式由（*i*）一个事实表构成，用于存储会话ID及其与维度表相连的键；以及（*ii*）14个维度表构成，分别对应改编版《儿科系统回顾》中的各个部分。事实表的主键是会话ID，而每个维度表的主键则是在插入行时生成的代理键。

采用星型模式的优势在于，维度表经过非规范化处理，因此表的数量更少 [22]。因此，执行 OLAP 操作所需的连接语句数量也随之减少，从而简化查询并加快执行速度 [39]。得益于这些优势，多个医疗保健系统采用了星型模式，因为这种数据库设计使大规模异构数据的管理更加容易 [6, 29, 34]。

数据库设计中一个重要的考量是如何处理包含多个值的字段。此模式采用的方法是将这些值拆分到多个列中；这是可行的，因为答案整合模块（第 3.1.2 节）会枚举聊天机器人识别出的所有实体引用值。举例来说，

表4：表3中文档的部分表格表示

字段（列）	值
会话表	
会话 _id	123456789
子项 _id	424242424
过敏_标识符	787878787
...	...
子项表	
子项 _id	424242424
用户名	juan-dela-cruz
...	...
过敏表	
过敏_标识符	787878787
对__坚果过敏	no
对__鸡蛋过敏	yes
对__海鲜过敏	yes
...	...
感到__呼吸困难	yes
出现_皮疹	yes
...	...
干预措施_涂抹了_软膏	yes
干预措施_远离__过敏原	no
...	...

表4展示了依据此模式转换表 3 中文档时涉及的部分相关表和字段（列）；由于子项和过敏信息存储在独立表中，因此分配了代理键（即 child_id 和 allergy_id）。

虽然这种方法需要在模式中为实体引用值列表的每次扩充新增一列，但考虑到本项目中的实体引用值（例如，医学文献或检查清单中的常见食物过敏和过敏反应症状 [31, 32]）预计不会频繁变化，因此这是可以接受的。

考虑过的其他方案包括使用多值数据类型存储多个值，或将它们存储为逗号分隔字符串。然而，前一种方案会在迁移时引入额外的复杂性，因为对多值数据类型的支持取决于数据库管理系统 [20]。与此同时，后一种方案被视为一种反模式，因为在提取汇总统计数据之前，必须先解析逗号分隔字符串，这会使联机分析处理聚合查询更加复杂且低效 [23]。

设计雪花模式是另一种曾考虑过的方法。尽管这种方法具有可扩展性，但每个多值字段都需要创建两张表（即一张用于存储可能值的表，以及一张映射表），因此会带来开销。强制执行这种程度的规范化会增加连接语句的数量，进而延长联机分析处理查询的执行时间 [39]。

最后，尽管将每个值分别存储在一行中可作为一种可能的反规范化策略，但需要注意的是，与医疗聊天机器人的对话中包含若干

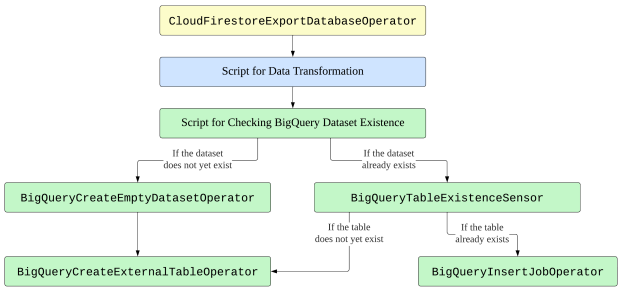

图 5：ETL流水线的有向无环图表示。在提取阶段（黄色），从 Cloud Firestore 中获取存储处理后的对话的文档。在转换阶段（蓝色），Python脚本会修改这些文档，使其符合模式。最后，在加载阶段（绿色），将它们作为行存储在相应的BigQuery表中。

多选题，因此会导致大量冗余，并显著增加存储需求。这可能会给将聊天机器人整合到全国性健康监测系统中带来挑战。

3.3.2 抽取-转换-加载流水线。 根据第 3.3.1 节所述的星型模式，我们创建了一个关系数据库，并将其存储在 BigQuery 中[15], ; BigQuery 是 Google Cloud 提供的数据仓储平台即服务。我们构建了一条抽取-转换-加载流水线，用于自动化并安排将 Cloud Firestore 中处理后的对话传输到 BigQuery 数据库的任务。图 5 展示了该流水线的高层有向无环图，该流水线使用 Apache Airflow [5], 实现；Apache Airflow 是面向数据工程的工作流管理平台。

转换阶段的脚本包含两个主要功能。第一个功能是将JSON文档中的字段映射到星型模式中各自对应的表。由于与系统回顾相关的字段名称前缀能够表明其所属的表，因此该功能的实现相当直接。第二个功能负责处理多值字段，并按照第 3.3.1 节中讨论的逻辑将其拆分为多个字段。

3.4 数据摘要与可视化

该流程的最后阶段是创建一个基于 Web 的仪表板，用于汇总和可视化儿童对话中来自聊天机器人的健康数据。该仪表板使用可视分析平台 Tableau 创建。为获取数据，我们使用Tableau的BigQuery驱动程序在 BigQuery 数据库与仪表板之间建立了实时连接。

3.4.1 仪表板设计。 该仪表板由多个图表组成，分别对应改编版《儿科系统回顾》中的相关问题。其设计旨在为多种联机分析处理 操作提供交互式且直观的支持。因此，该设计遵循底层数据层级，使用户能够更改粒度，即根据人口统计信息（即 年龄 和 性别）执行上卷和下钻，如图 6 所示。

它还允许用户根据人口统计信息动态筛选数据，从而灵活地切分和分析数据和响应，如图 7 所示。动态筛选器仪表板不仅用于将所有图表整合到同一视图中，还用于展示图表组合并增强交互性。这与为用户提供灵活演示方式的初衷相符，使他们能够更轻松地理解所提供的健康数据，并积极促进其处理这些信息的过程 [43]。

按照 Lechner 和 Fruhling 的研究中提到的公共卫生仪表板设计通用指南， [25], 动态筛选器还允许查看者使用下钻功能，从更概括的摘要中展示更详细的可视化内容，并根据年龄和性别呈现数据比较。

数据可视化主要由三种图表构成：（*i*）适用于仅允许单一响应的查询的饼图（例如，比较对坚果过敏与不过敏的儿童人数），（*ii*）适用于允许多个响应的查询的条形图（例如，比较患有各种常见过敏症的儿童人数），以及（*iii*）适用于下钻查询的堆叠条形图（例如，比较患有各种常见过敏症的儿童人数，并进一步按年龄细分）。

3.4.2 性能评估。通过测量不同视图的加载时间来评估仪表板的性能。测试在一台配备 AMD RyzenTM 7 5800H 八核处理器（处理器基准频率为 3.20 GHz）、16.0 GB 内存、475 GB 硬盘和 Microsoft Windows 10（64 位）操作系统的计算机上进行。为排除互联网连接和带宽等混杂变量的影响，所有测试均在 Tableau 桌面版上进行，数据则从本地托管的 MySQL 实时服务器获取。

除了初始加载时间外，我们还测量了所选 OLAP 操作的加载时间，以了解聊天机器人存储会话（对话）数量不同时的性能。所选操作如下：*(i)* 上卷、*(ii)* 单层下钻（按年龄）、*(iii)* 双层下钻（先按年龄，再按性别）、*(iv)* 单条件筛选（按响应）以及*(v)* 双条件筛选（按年龄和性别）。表 5 展示了五次试验的平均结果。

初始加载时间包括从服务器获取数据、联接表、缓存数据，以及生成首批面向用户的图表。由于这段时间大部分用于与服务器通信并为交互操作准备数据，因此逻辑模式的设计（第 3.3.1 节）至关重要。此外，如图 8 所示，存储条目数量增加 10^5 倍后，上卷、下钻和筛选结果的显示加载时间仅增加约一秒。

降低这些 OLAP 操作结果的展示开销，需要在数据缓存与数据新鲜度之间进行权衡。此仪表板采用 Tableau 内置的下钻、上卷和下钻功能，这些功能对缓存的数据进行操作，而非采用计算开销更大的方式，为每个 OLAP 查询都与服务器通信。缓存的数据会定期更新，默认间隔为 12 小时 [37]。在本项目背景下，这种权衡是可以接受的，因为目标是汇总一段时间内的数据，而不是生成实时分析报告。



(a) 汇总 (b) 按年龄下钻 (c) 按性别细分 (d) 按年龄再按性别下钻

图 6：对过敏数据应用上卷和下钻。可通过仪表板左侧面板中的下拉菜单执行上卷和下钻操作。与图 6a 中的汇总统计数据相比，图 6b 至图 6d 中的堆叠条形图展示了粒度更细的数据。左侧的图表比较各类常见过敏的统计数据，右侧的图表则聚焦于某种特定过敏（通过右侧面板中的下拉菜单选择）。

表 5：显示所选联机分析处理操作结果的加载时间（秒）

存储数量 会话	初始加载	上卷	下钻 按年龄	下钻 先按年龄，再按性别	筛选 按响应	筛选 按年龄和性别
10	16.089	5.063	5.068	5.076	1.080	3.056
100	17.031	5.070	5.071	6.012	1.088	4.007
1000	17.065	5.080	5.076	6.016	1.092	4.015
10000	21.082	5.094	5.093	6.017	2.000	4.029
100000	23.032	5.096	6.016	6.020	2.002	4.032
1000000	331.044	6.013	6.020	6.026	2.032	4.036



(a) 按人口统计信息筛选 (b) 按响应筛选

图 7：筛选过敏数据。为了提供更有针对性的洞察，可以将图 6 中的下钻操作与切片与切块操作结合使用；这些操作可通过左侧面板中的年龄和性别筛选器执行（图 7a）；右侧面板中的响应筛选下拉菜单也会根据所选的汇总可视化动态变化（图 7b ）。仪表板左侧显示所有响应的汇总统计，右侧显示筛选后的图表；二者并列展示，使用户无需切换视图即可进行比较并了解其背景。

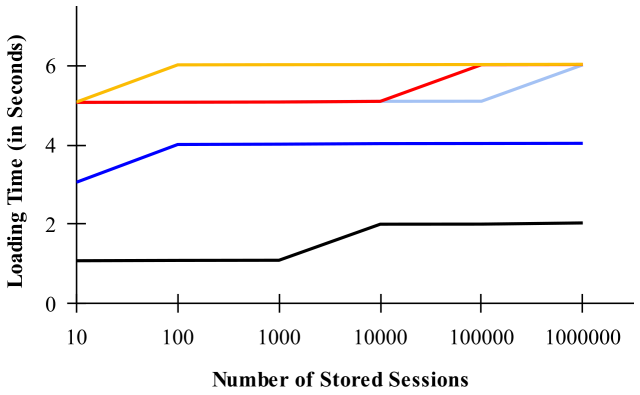

图 8：所选 OLAP 操作加载时间的线性-对数图。该图显示了汇总（天蓝色）、按年龄下钻（红色）、按年龄和性别下钻（黄色）、按响应筛选（黑色）以及按年龄和性别筛选（蓝色）的加载时间。

4 结论

从原始对话中提取洞察，需要兼顾自然语言的表达能力和数据库方面的考量。本文提出了一个四阶段流程，用于将非结构化对话转换为适用于可视化的结构化数据集市模式。首先，对对话进行处理，提取实体使用 DialogFlowCX 进行抽取和数据聚合。随后，处理后的对话以文档形式存储在 Cloud Firestore 中。然后，将其转换为用于联机分析处理的星型模式，关系数据库存储在 BigQuery 中，并使用 Apache Airflow 构建 ETL 工作流。最后，基于 Web 的交互式可视化用于汇总与聊天机器人对话中的数据，为健康分析提供洞察。

对该仪表板性能的评估显示，将与聊天机器人的存储会话（对话）数量增加 10^5 倍后，显示上卷、下钻和筛选结果的加载时间仅增加了约一秒。该可视化提供的分析结果可作为了解菲律宾公立学校学生常见疾病、症状和健康问题的信息来源。这些结果为医疗专业人员，以及负责医疗预算编制和干预规划的人员提供了有价值的洞察。

未来的工作方向包括：通过结合本体驱动词典扩充实体同义词列表，并探索多语言词嵌入，以改进答案整合器模块；同时与领域专家验证当前的关键词和实体同义词。还可以开展可用性测试，作为改进后续版本可视化仪表板的设计和交互性的依据。

致谢

本研究属于一个由科学技术部——菲律宾工业、能源与新兴技术研究与发展委员会（DOST-PCIEERD）资助的项目。

参考文献

[1] Alaa A. Abd-alrazaq、Mohannad Alajlani、Ali Abdallah Alalwan、Bridgette M. Bewick、Peter Gardner 和 Mowafa Househ。2019。心理健康聊天机器人的功能概述：范围综述。《国际医学信息学杂志》132 (2019)，103978。[2] Naeemeh Adel、Keeley Crockett、David Chandran 和 Joao P. Carvalho。2020。使用模糊语义相似度量解读对话系统中的人类响应。载于 2020 年 IEEE 国际模糊系统会议（FUZZ-IEEE）。1 8 . 10.1109 48607.2020.9177605 –。https://doi org/ /FUZZ Jessica S. [3] 2009 Ancker、Connie Chan 和 Rita Kukafka。。用于表达健康风险的交互式图形：开发与定性评估。《健康传播杂志》14, 5 (2009), 461–475。PMID: 19657926。[4] Carl Abelardo T Antonio、Ivy D Patdu 和 Alvin B Marcelo。2016。菲律宾健康信息隐私：政策与实践中的趋势和挑战。《Acta Medica Philippina》50, 4 (2016)。[5] Apache。 [n. d.]。Apache Airflow。https://airflow.apache.org/ [6] Bremang Appah 和 David Amos。2018。用于卫生服务决策的数据多维数据模型。《国际计算机科学工程技术期刊》3, 3 (2018), 1–6。[7] Mary Bates。2019。医疗保健聊天机器人来帮忙了。《IEEE Pulse》10, 3 (2019), 12 14 . 10.1109 .2019.2911816 –。https://doi org/ /MPULS Nivedita [8] Bhirud、Subhash Tataale、Sayali Randive 和 Shubham Nahar。2019。医疗保健领域聊天机器人的文献综述。《国际科学与技术研究期刊》8, 7 (2019), 225–231。[9] Deeno Burgan。2016。对话系统与对话管理。技术报告。DST Group Edinburgh，澳大利亚南澳大利亚州爱丁堡。[10] Vincenza Carchiolo、Alessandro Longheu、Michele Malgeri 和 Giuseppe Mangioni。2015。多源基于代理的医疗保健数据采集。载于 2015 年计算机科学与信息系统联合会议（FedCSIS）。1723–1729。[11] Heartland Community Health Center。 [n. d.]。儿科系统回顾（10 岁及以下）。https://www.heartlandhealth.org/wp-content/uploads/2012/06/Review-of-Systems-10-and-Under-2.14.17.pdf [12] James G Dolan 和 Stephen Iadarola。2008。低概率事件的风险沟通格式：患者偏好的探索性研究。《BMC 医学信息学与决策制定》8 (2008 年 4 月)，14。[13] JM Fortin、LK Hirota、BE Bond、AM O'Connor 和 NF Col。2001。确定患者对风险估计沟通方式的偏好：描述性试点研究。《BMC 医学信息学与决策制定》1 (2001)，2。[14] Shameek Ghosh、Sammi Bhatia 和 Abhi Bhatia。2018。Quro：借助个性化面向聊天机器人的对话系统检查用户症状。《健康技术与信息学研究》252 (2018), 51—56。[15] Google。2022。BigQuery。https://cloud.google.com/bigquery/docs/introduction [16] Google。2022。Cloud Firestore。https://firebase.google.com/docs/firestore [17] Google。2022。DialogFlowCX。https://cloud.google.com/dialogflow/cx/docs [18] Google。2022。实体。https://cloud.google.com/dialogflow/cx/docs/concept/entity [19] Google。2022。模糊匹配。https://cloud.google.com/dialogflow/cx/docs/concept/entity-fuzzy [20] PostgreSQL 全球开发组。 [n. d.]。8.15。数组。https://www.postgresql.org/docs/current/arrays.html [21] Nanette Guadalquiver。2018。第 6 1040126 区 160 万名 [22] 2008 公立学校学生将可享免费卫生服务。https://www.pna.gov.ph/articles/ Eric Johnson 和 Joshua Jones。。SQL Server 数据建模开发者指南：涵盖 SQL Server 2005 和 2008。Addison-Wesley Professional。[23] Bill Karwin 和 SQL 反模式。2010。避免数据库编程中的陷阱。The Pragmatic Bookshelf (2010), 15–155。[24] Liliana Laranjo、Adam G Dunn、Huong Ly Tong、Ahmet Baki Kocaballi、Jessica Chen、Rabia Bashir、Didi Surian、Blanca Gallego、Farah Magrabi、Annie Y S Lau 和 Enrico Coiera。2018。医疗保健领域的对话代理：系统综述。《美国医学信息学会会刊》25, 9 (2018 年 7 月)，1248–1258。[25] Bettina Lechner 和 Ann Fruhling。2014。公共卫生仪表板设计指南探索。载于《商业中的人机交互》，Fiona Fui-Hoon Nah（编）。Springer International Publishing，Cham，49–59。[26] Y. Li、D. McLean、Z.A. Bandar、J.D. O'Shea 和 K. Crockett。2006。基于语义网络和语料库统计的句子相似度。《IEEE 知识与数据工程汇刊》18, 8 (2006), 1138–1150。[27] Bingjie Liu 和 S. Shyam Sundar。2018。机器是否应表达同情与共情？与健康咨询聊天机器人的实验。《网络心理学、行为与社交网络》21, 10 (2018), 625–636。PMID: 30334655。[28] Bella Monse、Habib Benzian、Ella Naliponguit、Vincente Belizario、Alexander Schratz 和 Wim van Palenstein Helderman。2013。Fit for School 健康结果研究——一项纵向调查，用于评估菲律宾综合学校卫生计划的健康影响。《BMC 公共卫生》13 (2013 年 3 月), 256 3610233 。[29] https://europepmc.org/articles/PMC Lekha Narra、Tony Sahama 和 Peta Stapleton。2015。临床数据仓储：管理健康数据的商业分析方法。载于第八届澳大拉西亚卫生信息学与知识管理研讨会论文集*[《信息技术研究与实践会议》（CRPIT），第 164]*, 卷，A Maeder 和 J Warren（编）。澳大利亚计算机协会，澳大利亚，101–104。[30] 联合国。 [n. d.]。3. 确保各年龄段所有人的健康生活并促进福祉。 3 https://sdgs.un.org/goals/goal[31] 澳大利亚临床免疫学与过敏学会。 [n. d.]。过敏性鼻炎（花粉症）检查清单。http://www.allergy.org.au/images/stories/aer/infobulletins/pdf/ascia_allergic_rhinitis_checklist.pdf [32] 卫生部科迪勒拉卫生发展中心。 [n. d.]。食物过敏。https://caro.doh.gov.ph/food-allergies/ [33] 菲律宾共和国。2019。教育部健康行动计划：健康周。https://www.deped.gov.ph/wp-content/uploads/2019/07/DM_s2019_083.pdf [34] Samuel Z. Oliva 和 Joaquim C. Felipe。2018。通过数据仓储分析框架优化公共医疗管理。《IFAC 论文集》51, 27 (2018), 407–412。第十届 IFAC 生物与医疗系统研讨会 BMS 2018。[35] Asad Rahman、Madiha Liaqat、Ali Javeed 和 Farman Hassan。2022。健康顾问机器人：用于疾病预测的初级卫生保健监测聊天机器人。(2022)。[36] Ryan Richard Ruff、Bella Monse、Denise Duijster、Gina Santos Itchon、Ella Naliponguit 和 Habib Benzian。2020。基于学校的策略预防菲律宾儿童龋齿的有效性：一项整群随机试验。《社区牙科与口腔流行病学》n/a, n/a (2020 年 2 月)。[37] Salesforce。 [n. d.]。设置数据新鲜度策略。https://help.tableau.com/current/online/en-us/data_freshness_policy.htm [38] Nor Fazlida Mohd Sani、Lilly Suriani Affendey、Rahmita Wirza O.K. Rahmat、Mohd. Hasan Selamat、Rabiah Abdul Kadir 和 Mohd. Zamrin Dimon。2012。智能医疗临床系统的概念数据库设计。载于 2012 年先进计算机科学应用与技术国际会议（ACSAT）。 472 475 . 10.1109 .2012.41 –。https://doi org/ / [39] ACSAT Abraham Silberschatz、Henry F Korth 和 S Sudarshan。2020。《数据库系统概念》第 7 版。[40] Talib M. J. Al Taleb、Sami Hasan 和 Yaqoob Yousif Mahd。2021。门诊医疗保健的联机分析处理（OLAP）操作。《伊拉克科学杂志》(2021 年 1 月)，225–231。[41] Joseph Thomas。2009。病历与医疗过失问题。《印度泌尿学杂志》25, 3 2009 384 388 2779965 (7 月)，—。[42] https://europepmc.org/articles/PMC Ziang Xiao、Michelle X. Zhou、Q. Vera Liao、Gloria Mark、Changyan Chi、Wenxi Chen 和 Huahai Yang。2020。告诉我关于你自己的事：使用人工智能驱动的聊天机器人开展包含开放式问题的对话式调查。《ACM 人机交互汇刊》27, 3，文章 15 (2020 年 6 月)，37 页。[43] Fan Yang。2020。健康与风险传播的数据可视化。John Wiley & Sons, Ltd，第 13 章，213–232。