ALIBABA CLOUD

阿里云

云原生数据仓库AnalyticDB MySQL版

客户案例

文档版本：20220711

阿里云

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。

2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。

3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。

4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。

5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。

6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
危险	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	危险 重置操作将丢失用户配置数据。
警告	该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。	警告 重启操作将导致业务中断，恢复业务时间约十分钟。
注意	用于警示信息、补充说明等，是用户必须了解的内容。	注意 权重设置为0，该服务器不会再接受新请求。
说明	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	说明 您也可以通过按Ctrl+A选中全部文件。
>	多级菜单递进。	单击**设置**> **网络**> **设置网络类型**。
粗体	表示按键、菜单、页面名称等UI元素。	在**结果确认**页面，单击**确定**。
Courier字体	命令或代码。	执行 `cd /d C:/window` 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	`bae log list --instanceid` *Instance_ID*
[] 或者 [a|b]	表示可选项，至多选择一个。	`ipconfig [-all|-t]`
{} 或者 {a|b}	表示必选项，至多选择一个。	`switch {active|stand}`

目录

1.交通和物流

1.1. 基于AnalyticDB实现城市公交系统智能化

启迪公交基于AnalyticDB for MySQL和DRDS建设的完整解决方案，将人、车、线、站的大数据资源及相关配套资源进行商业化转换，引领行业提升公交系统的创新能力和服务水平，助力“互联网+城市公交”的提升发展。

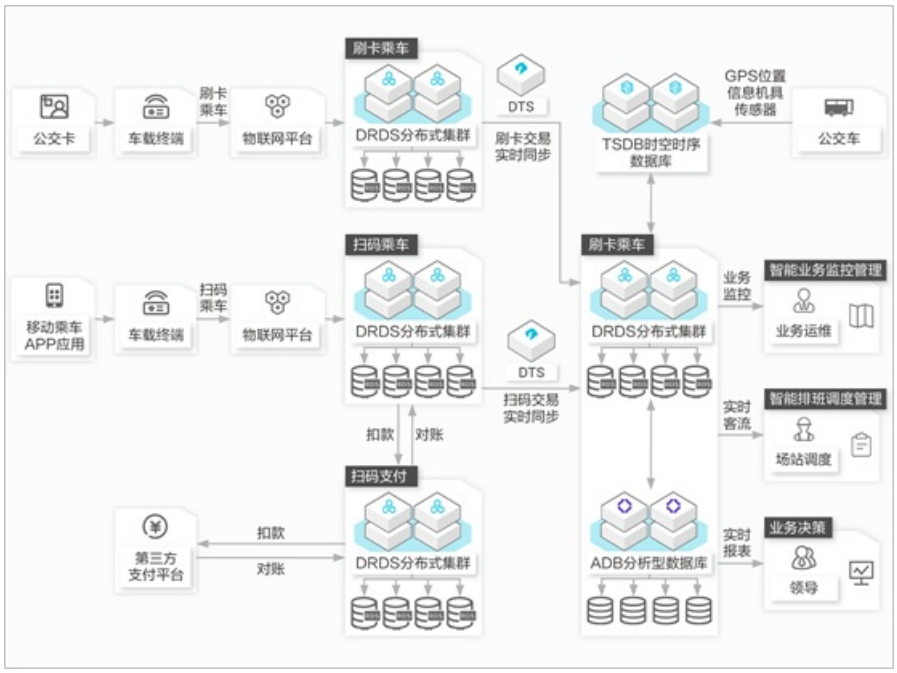
业务挑战

启迪公交（北京）科技股份有限公司（以下简称启迪公交）成立于2018年6月26日，是国内领先的智慧公交系统方案提供商和服务运营商。随着业务不断发展，启迪公交遇到了以下问题：

- 交易量大，并发大。启迪公交是北京市规模最大、车辆最多、车型和计费最复杂的城市公交系统，需要支撑北京市每天几千万笔市民的乘车相关数据，其中早晚高峰可达到每秒上千的并发量级。
- 数据源丰富。支撑北京市2万多辆公交车、车辆GPS机具以及传感器的IoT物联网数据。
- 数据量PB级。预计未来每年的平均数据存储、计算、分析和查询需求将达到PB级以上。

解决方案

针对接近PB级乘车数据和物联网平台数据的存储、计算、分析与查询需求，启迪公交通过多款阿里云数据库产品实现北京市公交系统的智能化。



方案解读：

- 启迪公交采用分布式关系型数据库DRDS（Distributed Relational Database Service）构建全部业务系统，具备海量大数据的处理能力，同时支持高并发、高可用和高度可扩展的弹性伸缩能力。
- 票务管理工作台将用户检录的数据实时同步到分析型数据库MySQL（AnalyticDB for MySQL）中，业务人员即可使用SQL进行自由灵活的计算分析，得出实时客流，然后结合公交车机具上传的监控信息进行实时分析，指挥车辆调度。

客户价值

基于AnalyticDB for MySQL和DRDS建设的公交系统智能解决方案为启迪公交带来以下价值：

- 通过北京公交App，乘客能享受路线规划、站点查询、公交预报、定制公交、旅游等创新类服务。
- 通过智能业务分析系统，公交票务人员能够及时了解运营状态、结算报表和监控报警等信息。
- 通过智能实时客流系统自动分析客流情况，实时监控北京市2万多辆公交车的运行状态，结合车辆和机具运行状态生成调度方案，助力场站调度人员更加合理地制定调度排班方案。
- 通过实时大数据汇总分析系统，协助管理层人员制定更加快速有效的业务决策。

2.互联网
2.1. 客如云精准营销方案

通过AnalyticDB for MySQL和PolarDB以及其他阿里云产品为客如云制定精准营销方案，解决了客如云资源调度缓慢、客户群体模糊以及业务量飙升造成的系统异常等问题，提升了客如云商业价值。

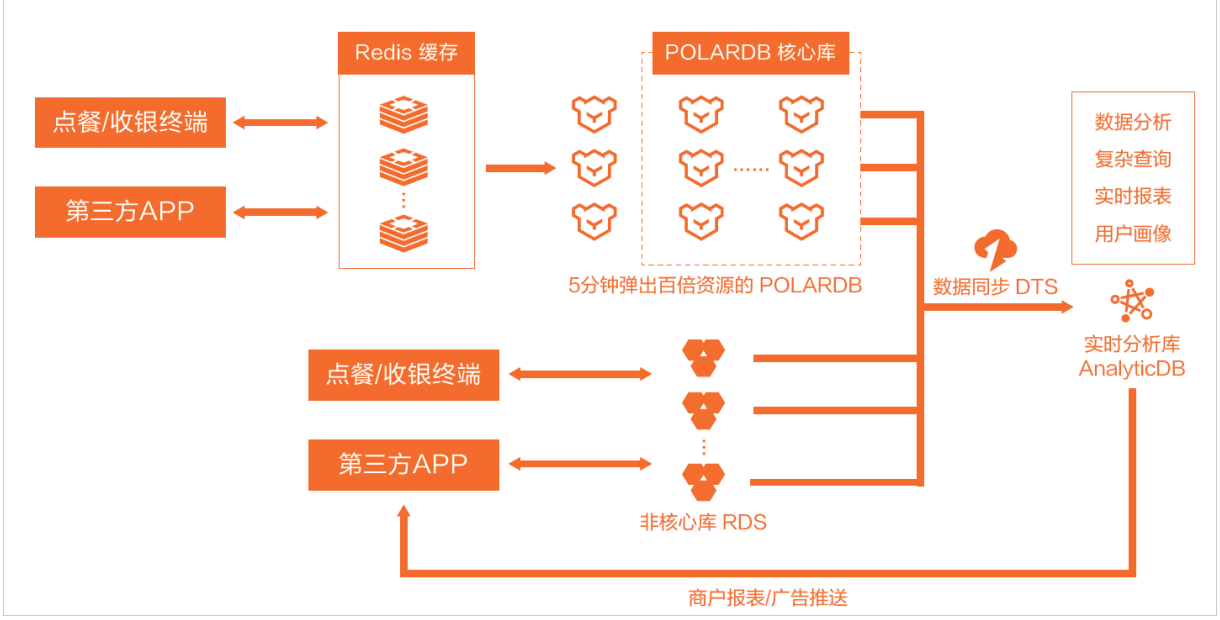
业务痛点

客如云隶属于时时同云科技（北京）有限责任公司，主要为餐饮店提供点餐和收银服务、为餐饮商户提供运营服务，在餐饮和点餐管理领域处于全国领先水平。在客如云业务发展中，以下问题急需解决：

- 商户隔天才能查看运营情况，导致连锁商户资源调配滞后，资源调配效率有待提升。
- 商户希望客如云可以提供更精准的客户画像，例如区分客户的年龄段、 消费习惯和消费层次等，以便区分目标客户群体，为不同类型的客户提供更贴心的餐饮服务，例如情侣套餐、经济套餐、满减打折券等。
- 在元旦、情人节、七 夕、圣诞等节假日以及点餐高峰时段，业务量飙升至平时的4倍，客户点餐和结账非常缓慢甚至经常失败。

解决方案

针对以上问题，阿里云制定以下解决方案：



方案解读：

- 将原系统中分散使用Hive、HBase、kylin、TiDB处理的各类分析业务集中更换为分析型数据库MySQL（AnalyticDB for MySQL）中，确保实时输出分析报表。
- 将业务库数据实时同步到AnalyticDB for MySQL中，商户运营报表每5分钟更新一次。
- 通过AnalyticDB for MySQL构建客户标签系统，开发客户画像分析业务。
- 将23个核心库更换为PolarDB，支持5分钟内最高弹性扩展出百倍资源，应对业务峰值。

客户价值

基于阿里云构建的精准营销方案为客如云带来以下价值：

- 新增报表服务盈利

推出商户报表VIP套餐，每一小时更新一次报表，该套餐当月为客如云带来超过千万元的收入。

- 新增客户画像业务

  新增客户画像精准营销服务，预计上线后每月销售额达到3000万元以上。

- 支持业务峰值
- 以2019年七夕为例，在点餐数量同比增加50%的情况下，平均下单时间仅需2秒，也没有遇到客户点餐、结账缓慢后者失败的问题。

2.2. 赛盒广告精细化管理方案

赛盒科技基于阿里云AnalyticDB for MySQL和Quick BI制定的广告精细化管理方案，提升了广告创建效率和广告业绩，根据实时BI报表数据合理分配广告活动预算，实现企业效益最大化。

业务痛点

深圳赛盒科技有限公司（赛盒科技）是一家以高科技创意为核心的技术服务公司，随着用户量的增长，赛盒科技面临以下业务挑战：

- 高并发实时需求

  用户在第三方电商平台通过关键词搜索时产生高并发实时入库需求。

- 海量历史数据的存储需求

  每位用户的广告数据存储在独立的物理库中，部分用户单张表每年将存储超过1亿行的记录，如果历史数据保留一年，300位用户一年预计产生18TB的数据。

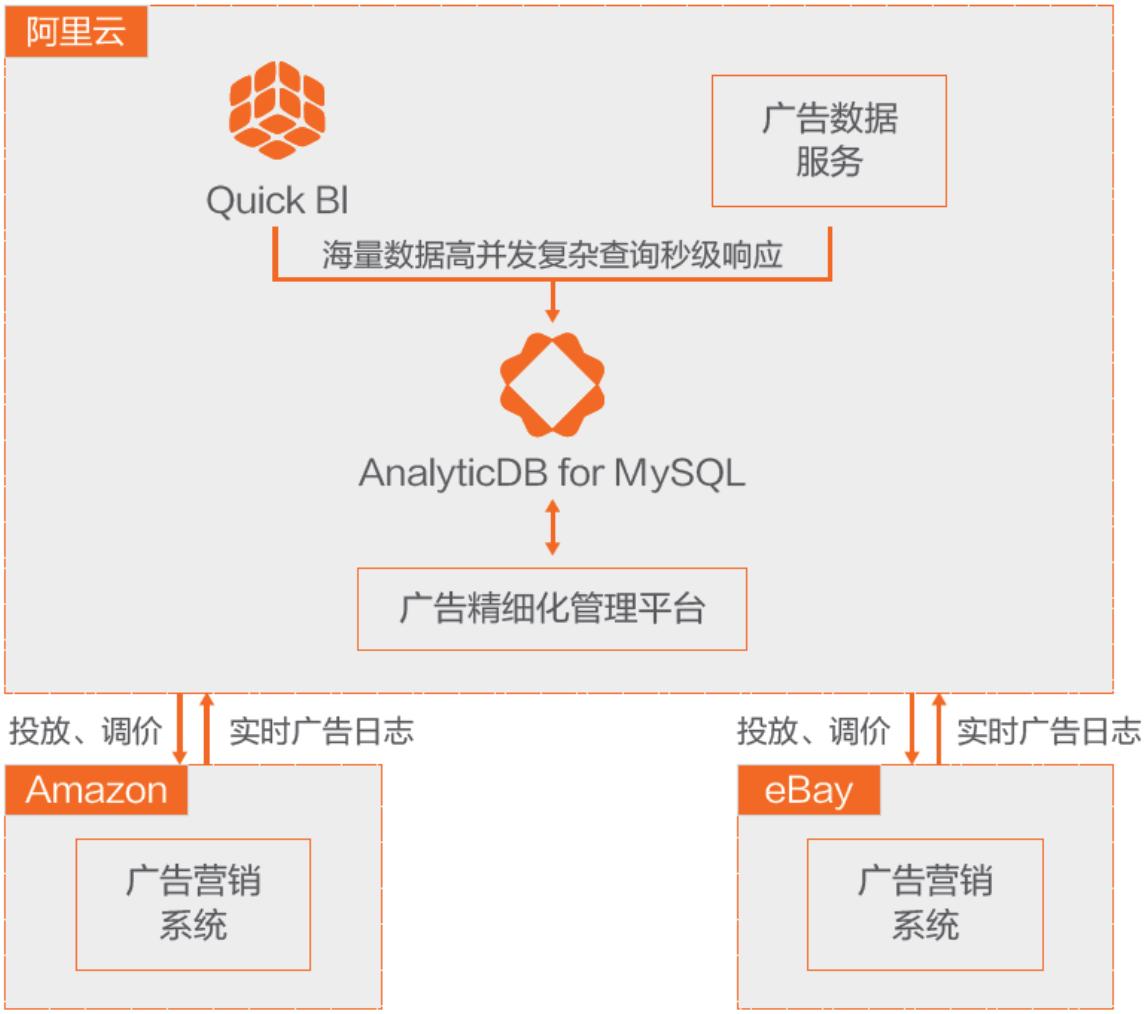
- 对数据库的实时、高速查询能力需求

  用户通过仪表板实时了解广告的转化率和投放收益比等指标，业务高峰时可达100+QPS（Queries-per-second）的复杂查询。

- 对数据库查询响应时长的需求
- 实际业务中要求数据库的查询相应时间小于20秒，经过测试，其他第三方OLTP（On-Line Transaction Processing）数据库中部分GROUP BY语句的响应时间超过40秒，难以满足业务需求。

解决方案

针对以上问题，阿里云通过分析型数据库MySQL（AnalyticDB for MySQL）和Quick BI制定了以下解决方案：



上述方案中采用AnalyticDB for MySQL存放业务数据，AnalyticDB for MySQL的配置为集群版16个C8节点，可支撑上万TPS（Transactions-per-second），TB级数据复杂查询秒级响应，同时支持弹性扩展节点数和存储空间。

客户价值

基于AnalyticDB for MySQL和Quick BI的广告精细化管理方案为赛盒科技带来以下价值：

- 通过统一管理多个电商站点、多个店铺的关键词广告，解决用户在多个电商站点间来回切换操作的问题。
- 通过对海量数据的实时统计分析，识别出转化率高的搜索词，帮助用户快速创建广告活动，提升广告创建效率和广告业绩。
- 通过对海量数据的统计分析，用户可以按时段或业务变化智能调价，避免错过调价黄金时段。
- 用户可以实时了解广告投放产生的商品毛利润，让每一笔收益都清晰可见。
- 快速查看分析高效、低效关键词，合理分配广告活动预算，效益最大化。

2.3. 互联网行业实时BI分析

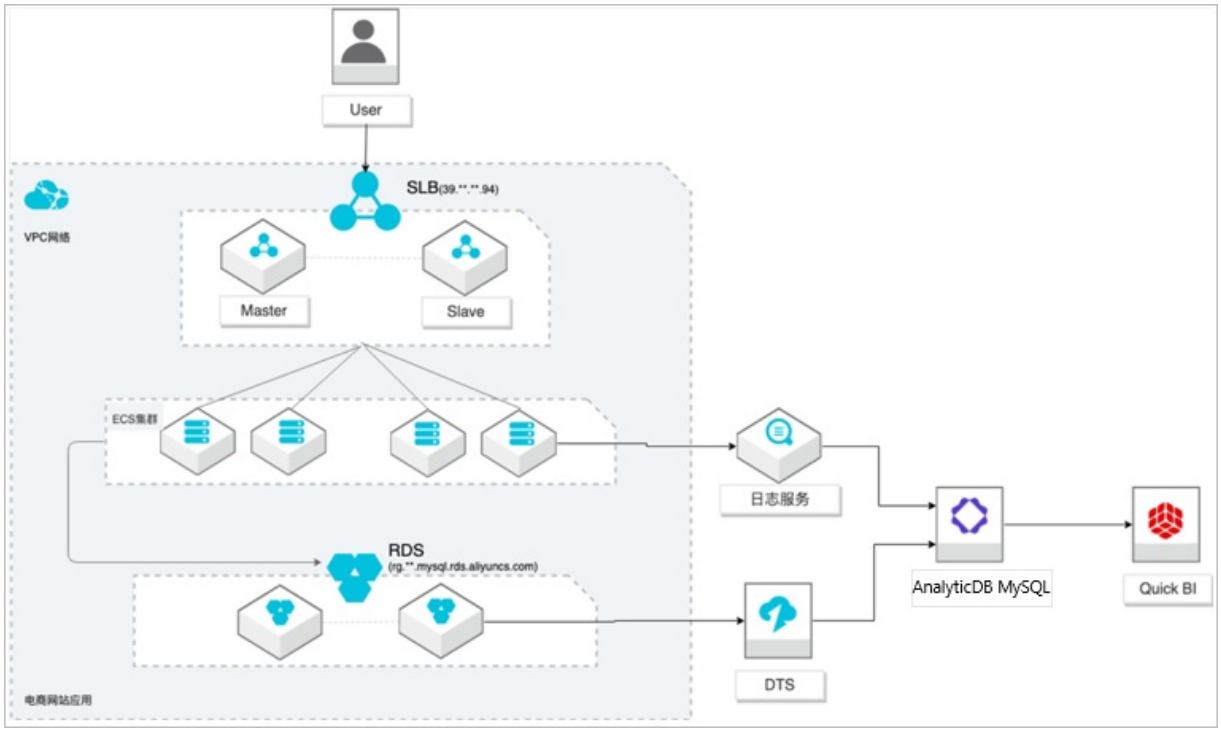
本文以某互联网公司为例，介绍如何将DB业务数据和日志数据实时同步到云原生数据仓库AnalyticDB MySQL版中，然后通过Quick BI进行实时可视化数据分析。

相对于传统的关系型数据库，云原生数据仓库AnalyticDB MySQL版只需要毫秒级时间，即可实现万亿数据高并发多维实时分析。

应用范围

- 互联网公司的网站、App以及小程序应用内BI分析场景。
- 在线运营以及运营指标实时化分析等场景。
- 可扩展到各类网站BI分析场景使用。

架构图



详细实现方式请参见互联网、电商及游戏行业实时BI分析。

客户价值

- 1小时短平快即可实现实时数据分析平台建设，无需掌握Hadoop\Spark\Flink\Presto\Impala等复杂的大数据技术。
- 操作简单快捷，全程拖拽式配置，无需编码。
- 业务实时指标数据延时在1分钟以内。

3.新零售

3.1. AnalyticDB助力Flowerplus业务高速发展

基于MySQL生态打造的HTAP（Hybrid Transaction and Analytical Process）数据库解决方案（RDS MySQL+AnalyticDB for MySQL）有效支撑了Flowerplus的业务鲜花售卖。基于AnalyticDB for MySQL快速分析海量数据的结果，优化用户的采购环节、订单分析、营销活动、业务报警等重要业务，助力Flowerplus业务快速发展的同时，为未来业务发展提供足够的扩展性。

业务痛点

Flowerplus（花加）的业务主要涉及鲜花采购、售卖、物流，需要通过BI报表分析鲜花库存情况、采购链路、物流进展、业务转化率、商品售罄报警等，同时也要对海量用户订单进行业务分析。由此可见，Flowerplus对大数据分析的实时性要求较高，而传统的MySQL数据库无法满足这一需求。

- 复杂数据查询性能

  使用传统MySQL数据库对订单、商品流量、采购、业务转化率以及商品售罄报警等分析时速度较慢，数据达到千万级或者亿级时，复杂查询报表返回很慢或者根本无法返回，无法正常支撑报表和BI业务。

- 数据实时性

  部分报表对返回速度要求较高，要求秒级返回。

- 数据兼容性

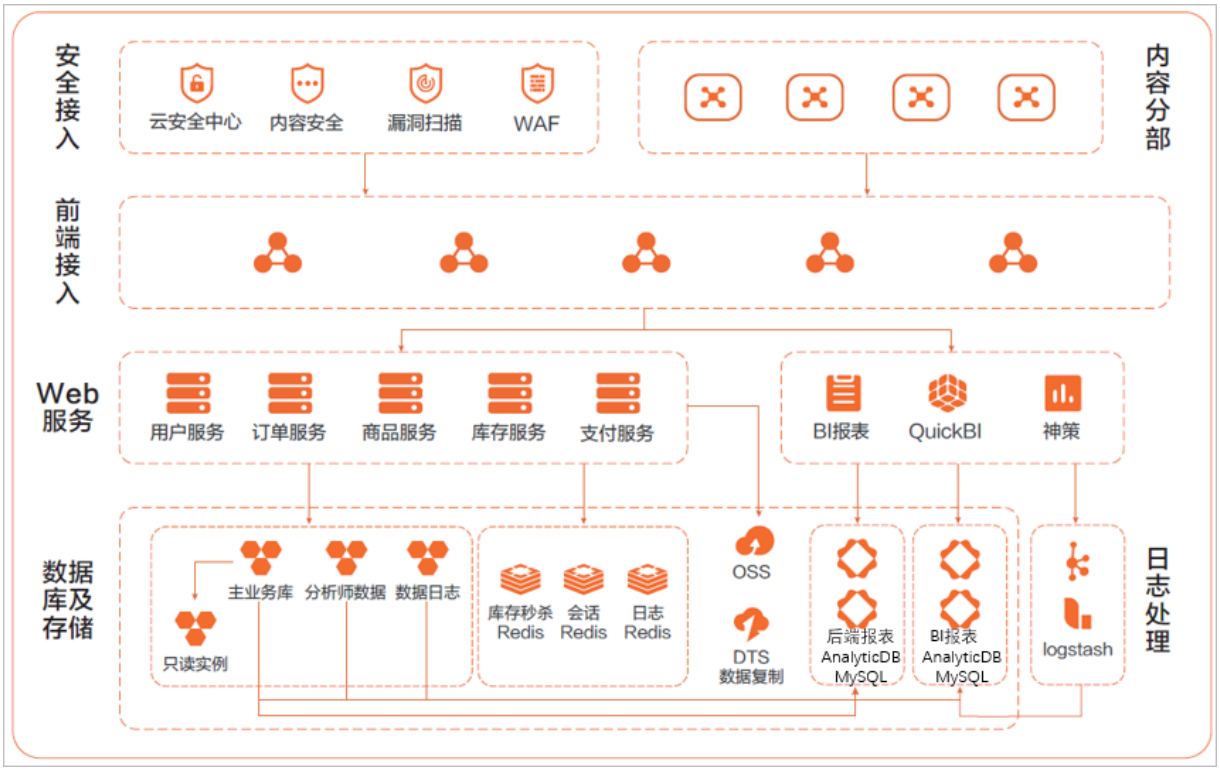
  Flowerplus原有系统中主要使用RDS MySQL、PolarDB MySQL、DRDS等MySQL生态数据库，Flowerplus的研发团队希望分析存储产品能兼容MySQL数据库语法，降低研发团队的使用成本。

- 弹性扩展存储空间

  Flowerplus当前部分报表业务量数据已达到数亿级，未来可能超10亿，Flowerplus希望分析存储产品能够弹性扩展。

解决方案

针对上述Flowerplus遇到的业务痛点，阿里云制定以下方案。



方案解读：

- 使用分析型数据库MySQL（AnalyticDB for MySQL）替换MySQL进行OLAP（Online Analytical Processing）数据分析，提升业务报表和BI报表分析速度，确保业务报表和BI报表快速返回。
- AnalyticDB for MySQL2.0和3.0均兼容MySQL语法，其中AnalyticDB for MySQL 3.0的技术架构和MySQL数据库更贴合。
- 通过增加节点提升AnalyticDB for MySQL的并发能力和并行计算能力，达到横向扩展的目的，保证在业务报表数据量和查询复杂程度提高的情况下，依然能够以相对稳定的返回速度执行报表查询。

客户价值

RDS MySQL+AnalyticDB for MySQL的HTAP（Hybrid Transaction and Analytical Process）数据库解决方案为Flowerplus带来以下价值：

- 通过AnalyticDB for MySQL的快速分析能力，提升Flowerplus的数据分析效率，帮助企业更加快速的进行业务优化。
- 通过AnalyticDB for MySQL进行报表分析时，数据分析性能提升了2~10倍，极大的提升了业务体验。
- AnalyticDB for MySQL基于多节点集群架构，相比传统关系型数据库，具有更高的扩展性和灵活性，降低了Flowerplus由于数据量和访问量变大带来的维护成本。

4.金融

4.1. 聚合支付方案

阿里云从业务扩展性、数据流动性、服务高可用三方面入手，通过多款云数据库产品为利楚扫呗制定聚合支付方案，解决利楚扫呗在业务扩展期遇到的数据存储空间不足、读写性能下降以及大数据分析空缺等问题。

业务痛点

武汉利楚商务服务有限公司（利楚扫呗）已覆盖全国400多个地级市，旗下商户数量70万家，每天处理交易数大约1200万笔，年受理交易金额2000亿元，成功为上百个行业提供聚合支付综合解决方案。随着业务持续扩张，利楚扫呗面临以下问题：

- 存储空间

  利楚扫呗近3个月的数据存储量已超过2TB，业务还在持续扩张，预计单体RDS的存储空间将不够用。

- 高并发

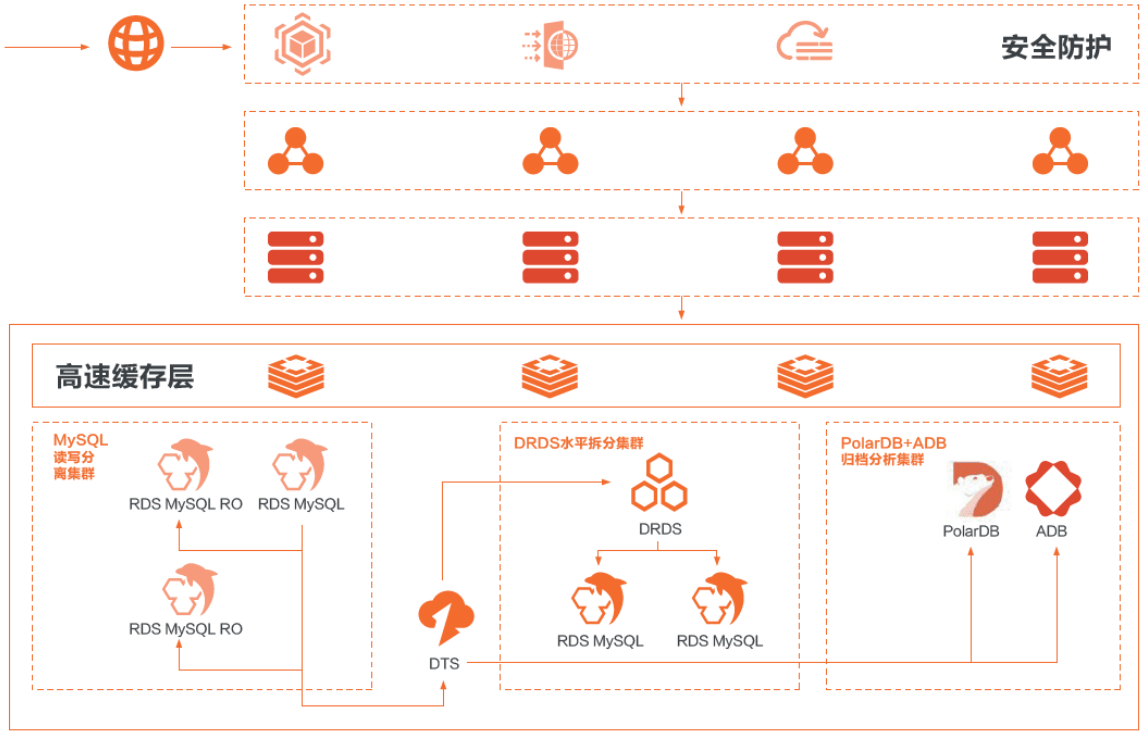
  随着接入商户数量的增加，每日的流水持续增长，数据库读写性能明显下降。

- 分析型需求

  随着数据量的增大，数据查询涉及的量级呈指数级上升，针对商户等大数据量场景的分析查询，单体MySQL已无法满足需求。

解决方案

阿里云通过多款云数据库产品为利楚扫呗制定以下解决方案：



方案解读：

- 使用DRDS分库分表将数据库进行水平拆分，有效解决订单的存储上限及业务高峰时的高并发压力，借助

了DRDS对MySQL语法的高度兼容能力，轻松从原来单体数据库升级到现在的分布式架构。

- 使用PolarDB进行数据归档，满足超大数据量的存储需求，借助存储、计算分离以及与MySQL 100%的兼容性，确保原有业务不做修改。
- 使用Redis来做缓存，通过缓存提高读取效率；使用RDS只读实例中的读库进行分流，自动均衡读流量。
- 使用AnalyticDB for MySQL对海量流水数据进行即时查询，满足运营的实时性要求，快速精准进行客户营销。

客户价值

- 业务可线性拓展

  依托Redis缓存和DRDS水平或者垂直拆分数据库技术，解决了高并发、存储容量扩展以及在线计算扩展的难题，使系统具备了存储100TB数据的能力，同时也具备10万TPS（Transactions-per-second）、百万QPS（Queries-per-second）的支撑能力，足以支撑利楚扫呗业务扩展至当前业务量的5~10倍，解决了在业务快速增长过程中对数据库存储量和并发量带来的挑战，确保在业务快速增长的情况下，每一位终端用户都能享受平滑的支付体验。

- 让数据流动起来

  借助DTS的数据实时同步功能，让客户的数据实时流动起来。借助PolarDB和AnalyticDB for MySQL的大数据处理能力，决策报表的生成时间从分钟级别降低至秒级，大部分报表可以在10秒内生成，部分特别复杂的报表生成时间为1分钟。业务行为和业务决策平滑对接，决策者可以通过报表数据快速进行业务决策，实现通过计算数据创造业务价值。

- 提供7×24小时的高可用服务

  以业务7×24小时运行标准设计，提供弹性升降配、在线扩容、SQL审计分析、只读实例等功能，阿里云可提供金融级别的安全性和稳定性。

5.生物

5.1. 高效基因序列检索助力快速分析肺炎病毒

云原生数据仓库AnalyticDB MySQL版是云端托管的PB级高并发低延时数据仓库，通过AnalyticDB MySQL版向量检索功能构建的基因检索系统，支持毫秒级针对10亿级别的向量数据进行查询分析，更加快速、高效的为肺炎病毒防控、研发治疗药物以及相关疫苗提供帮助。

基因序列检索技术应用范围和现状

基因序列检索技术主要应用于以下场景：

- 用于肺炎病毒的溯源和分析，找到病毒宿主，做好有效防范。
- 用于分析肺炎病毒的复制和传播过程，为研发治疗药物和疫苗提供帮助。
- 用于检索与肺炎病毒相似的病毒基因序列。

当前的基因匹配算法太慢，迫切需要高效匹配算法进行基因序列检索。阿里云AnalyticDB MySQL版团队将基因序列片段转化成对应的1024维特征向量，将两个基因片段的匹配问题，转换成了两个向量间的距离计算问题，从而大大降低了计算开销，实现毫秒级返回相关基因片段，完成基因片段的首次筛选。然后，使用基因相似计算BLAST算法，完成基因相似度的精确排查，从而高效率完成基因序列的匹配计算。匹配算法从原来 `O(M+N)` 的复杂度降低到 `O(1)` 。同时，阿里云AnalyticDB MySQL版提供强大的机器学习分析工具，通过基因转向量技术，将局部的和疾病相关的关键靶点基因片段转成特征向量，用于基因药物的研发，大大加速了基因分析过程。

AnalyticDB MySQL版基因检索系统

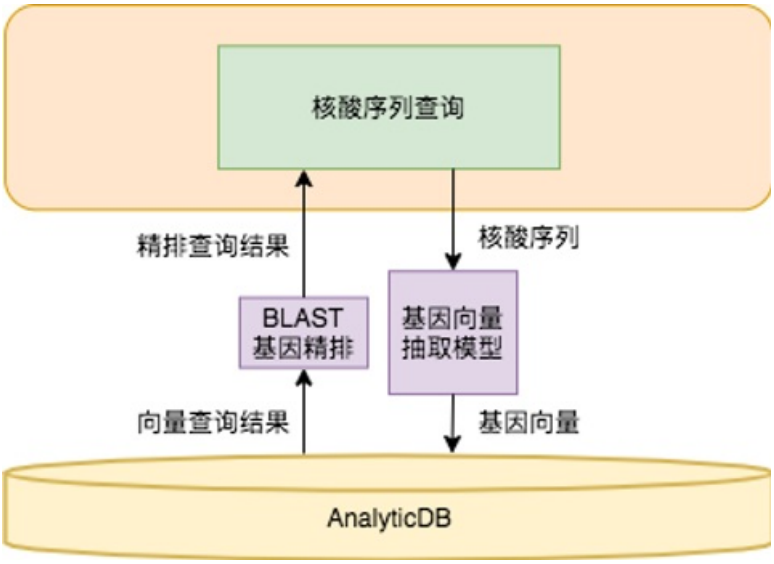
肺炎病毒的RNA序列可以用一串核酸序列（又称碱基序列）表示，RNA序列含有有四种核苷酸，分别用A、C、G和T表示，分别代表腺嘌呤、胞嘧啶、鸟嘌呤、胸腺嘧啶。每个字母代表一种碱基，无间隔排列在一起。每一个物种的RNA序列均不相同但又有规律，基因检索系统可以通过输入一串病毒的基因片段，检索相似的基因，用来对病毒的RNA序列进行分析。

为方便演示AnalyticDB MySQL版基因片段检索方法，我们从GenBank下载了大量病毒的RNA片段，并将GenBank内部关于病毒的论文以及Google Scholar中相关病毒的论文导入AnalyticDB MySQL版基因检索数据库中。

AnalyticDB MySQL版会将肺炎病毒的序列上传到AnalyticDB MySQL版基因检索系统中，AnalyticDB MySQL版基因检索系统只需几毫秒即可检索到相似的基因片段（当前示例系统只返回匹配度超过0.8的基因片段）。从返回的基因片段得出穿山甲携带的肺炎病毒（GD/P1L）、蝙蝠携带的肺炎病毒（RaTG13）以及SARS和MARS病毒，其中GD/P1L的序列匹配度最高为0.974，由此推断出肺炎病毒很可能是通过穿山甲传染到人的。

如果RNA片段非常相似，说明这两个RNA可能有相似的蛋白质表达和结构。通过基因检索工具，可以看到SARS和MARS与肺炎病毒的匹配度为0.8以上，说明可以将一些SARS或者MARS的研究成果应用到肺炎病毒上。系统提取了每种病毒的论文，通过文本分类算法，将论文划分为检测类、疫苗类和药物类。其中，对SARS有效的荧光定量PCR检测，目前正应用于肺炎病毒的检测；基因疫苗的方法以及诱导体内免疫疫苗的方法，也正在展开研究；治疗药物中瑞德西韦以及相关的干扰素也都用于肺炎病毒的治疗上。

实现架构



AnalyticDB MySQL版基因检索系统中，AnalyticDB MySQL版负责存储和查询所有结构化数据（例如基因序列的长度，长度包含基因的论文名称、基因种类、DNA或者RNA等）和基因序列产生的特征向量。查询数据时，通过基因向量抽取模型将基因转化成向量，在AnalyticDB MySQL版向量数据库中进行粗排检索，然后在返回的向量匹配结果集中使用经典的BLAST算法进行精确检索，返回最相似的基因序列。

AnalyticDB MySQL版基因检索系统的核心是基因向量抽取模型，该模块可以将核苷酸序列转化成向量。目前AnalyticDB MySQL版抽取了各种病毒的RNA全部序列样本进行训练，可以非常方便的对病毒的RNA进行相似度计算。同时，基因向量抽取模型也可以扩展应用于其他物种基因检索。

基因向量抽取算法

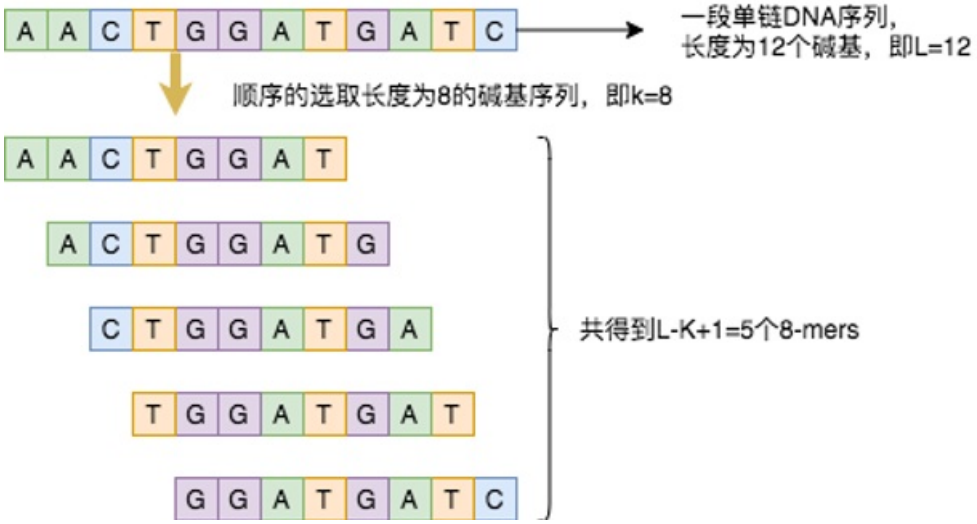
目前词向量技术已经非常成熟，被广泛应用于机器翻译、阅读理解、语义分析等相关领域，并取得了巨大成功。词向量化采用了分布式语义的方法表示一个词的含义，一个词的含义就是这个词所处的上下文语境。例如高中英语中的完形填空题，一篇短文空出10个地方，学生根据空缺词的上下文语境选择合适的词。也就是说上下文语境能够准确的表达这个词，如果某位同学选择了正确的选词，表示该同学理解了空缺词的含义。因此，通过上下文词的关系，采用词向量算法，可以为每个词生成一个向量，通过计算两个词向量之间的相似度，得到两个词的相似度。

同样的道理，基因序列的排列具有一定的规律，并且每一部分基因序列所表达的功能和含义不同。可以将很长的基因序列划分成小的单元片段（也就是词）进行分析，并且这些词也有上下文语境，这些词相互连接、相互作用共同完成相对应的功能，形成合理的表达。因此，生物科学家们采用词向量算法对基因序列单元进行向量化，两个基因单元相似度很高，说明需要这两个基因单元共同来表达和完成相应的功能。

总体而言，AnalyticDB MySQL版基因向量抽取算法分为三步：

1. 在氨基酸序列中定义词

   生物信息学中用K-mers来分析氨基酸序列，k-mer是指将核酸序列分成包含k个碱基的字符串，即从一段连续的核酸序列中迭代选取长度为K个碱基的序列，若核酸序列长度为L，k-mer长度为K，那么可以得到L-K+1个k-mers。例如下图所示，假设某序列长度为12，设定选取的k-mer长度为8，则得到（12-8+1=5）5个5-mers。这些k-mer，就是氨基酸序列中词。



2. 找到氨基酸序列的上下文语境，将基因序列中的词转换成1024维向量。

   对于词向量算法而言，另一个重要的问题就是上下文的语境。AnalyticDB MySQL版基因向量抽取算法在氨基酸片段中选择一个长度为L的窗口，该窗口内的氨基酸片段可认定为在同一语境内。例如CTGGATGA是一段核酸序列，选取了长度为10的窗口，AnalyticDB MySQL版基因向量抽取算法将CTGGATGA转换成5个5-mers即{AACTG, ACTGG, CTGGA, GGATG, GATGA}。对于其中一个5-mer {CTGGA}而言，另外四个{AACTG, ACTGG, GGATG, GATGA}5-mers就是当前5-mer {CTGGA} 的上下文语境。AnalyticDB MySQL版基因向量抽取算法套用词向量空间训练模型，对已有生物基因的k-mers进行训练，便可将一个k-mer（基因序列中的一个词）转换成1024维向量。

3. 类似于词向量模型，k-mer向量模型也拥有和词向量模型相似的数理计算性质。

   - 向量减法：$||vec(ACGAT) - vec(GAT)|| \approx ||vec(AC)||$
   - 向量加法：$||vec(AC) + vec(ATC)|| \approx ||vec(ACATC)||$

   向量减法公式说明核苷酸序列ACGAT的向量减去GAT序列的向量和AC的序列向量距离非常接近。向量加法公式说明核苷酸序列AC的向量加上ATC序列的向量和ACATC序列的向量距离也很接近。因此，根据这些数理特征，计算一个长氨基酸序列向量时，可以将这个序列中每一段的k-mer序列进行累加，最后进行归一化就能得到整个氨基酸序列的向量。同时，为提升精度，可以将基因片段看作一个文本，使用doc2vec函数将整个序列转换成向量进行计算。为进一步验证算法性能，AnalyticDB MySQL版基因向量抽取算法计算了常用于基因检索库中的BLAST[6]算法序列与基因转向量l2距离序列的相似度，两个序列的斯皮尔曼等级相关系数是0.839。 以上得出结论，将DNA序列转换成向量用于相似基因片段的初次筛选，是有效且可行的。

向量检索功能概述

一般包含向量检索的应用系统中，开发者通常会使用向量检索引擎（例如Faiss）存储向量数据，然后使用关系型数据库存储结构化数据。因此，查询时也需要交替查询两个系统，明显额外增加了开发人员的工作量，数据查询性能也不是最优。

AnalyticDB MySQL版是云端托管的PB级高并发低延时数据仓库，可以毫秒级针对10亿级别的向量数据进行查询，100毫秒级别的响应时间。AnalyticDB MySQL版全面兼容MySQL协议以及SQL:2003 语法标准，其向量检索功能支持对图像、文本推荐、声纹、核苷酸序列等相似性进行查询和分析，目前在多个城市的安防项目中已大规模部署了AnalyticDB MySQL版。

AnalyticDB MySQL版支持结构化和非结构化数据的近似检索和分析，通过SQL接口即可快速搭建基因检索或者基因+结构化数据混合检索等系统。在混合检索场景中AnalyticDB MySQL版的优化器会根据数据的分布和查询条件选择最优执行计划，在保证数据召回率的同时，得到最优的性能。例如，通过以下一条SQL即可检索RNA核酸序列。

```
-- 查找RNA和提交的序列向量相近的基因序列。
select  title, # 文章名
        length, # 基因长度
        type, # mRNA或DNA等
        l2_distance(feature, array[-0.017,-0.032,...]::real[]) as distance # 向量距离
from demo.paper a, demo.dna_feature b
where a.id = b.id
order by distance; # 用向量相似度排序
```

上述SQL中表demo.paper用于存储上传的每篇文章的基本信息，demo.dna_feature存储各个物种的基因序列对应的向量。通过基因转向量模型，将要检索的基因转成向量[-0.017,-0.032,...]，然后在AnalyticDB MySQL版数据库中进行检索。

当前系统也支持结构化信息+非结构化信息（核苷酸序列）的混合检索，例如查找和冠状病毒相关的类似基因片段时，只需要在SQL中增加 `where title like'%COVID-19%'` 即可。

附录

- [1] Mikolov Tomas; et al. (2013). "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781.
- [2] Mikolov Tomas, Sutskever Ilya, Chen Kai, Corrado, Greg S. and Dean Jeff (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems. arXiv:1310.4546. Bibcode:2013arXiv1310.4546M.
- [3] Mapleson Daniel, Garcia Accinelli, Gonzalo, Kettleborough George, Wright Jonathan and Clavijo, Bernardo J. (2016). "KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies". Bioinformatics. 33(4): 574-576. doi:10.1093/bioinformatics/btw663. ISSN 1367-4803. PMC 5408915. PMID 27797770.
- [4] Quoc Le and Tomas Mikolov. (2014). Distributed representations of sentences and documents. In International Conference on Machine Learning, pages 1188-1196.
- [5] 人类基因组hg38, http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/hg38.chromFa.tar.gz.
- [7] Julia Piantadosi, Phil Howlett and John Boland. (2007). "Matching the grade correlation coefficient using a copula with maximum disorder", Journal of Industrial and Management Optimization, 3 (2), 305-312.
- [8] Stephen Woloszynek, Zhengqiao Zhao, Jian Chen and Gail L. Rosen. (2019). "16s rRNA sequence embeddings: Meaningful numeric feature representations of nucleotide sequences that are convenient for downstream analyses", PLoS Computational Biology, 15(2), e1006721.
- [9] James K. Senter, Taylor M. Royalty, Andrew D. Steen and Amir Sadovnik. (2019) "Unaligned Sequence Similarity Search Using Deep Learning.", arXiv e-prints.
- [10] Ng Patrick. (2017) dna2vec: consistent vector representations of variable-length k-mers. arXiv preprint, arXiv:1701.06279.