Uber 向 Apache 软件基金会提交开源大数据存储库 Hudi

快速管理和访问 PB 级数据的能力对于整个数据生态系统的可伸缩增长是至关重要的。尽管如此，这种对规模和速度的综合需求并不总是自然地适合现有的批处理和流系统架构。

Hudi 于 2016 年以“Hoodie”为代号开发，旨在解决 Uber
大数据生态系统中需要插入更新和增量消费原语的摄取管道和 ETL
管道的低效问题。为了与更广泛的大数据社区分享这些好处，Uber 在 2017 年开源了 Hudi。

2019 年 1 月，我们向 Apache 孵化器提交了 Hudi ，从而进一步推进了我们的开源承诺，保证 Apache Hudi 可以在 Apache 软件基金会的开放治理和指导下长期可持续性地增长。

Hudi 联合创始人 Vinoth Chandar 说：“考虑到 Uber 使用了这么多优秀的 Apache
项目，我们相信 Apache 社区驱动的开源开发方式将使我们能够与不同的贡献者合作，发展 Apache Hudi。我们期待与 Apache 软件基金会合作，实现最佳实践，并为项目带来新的想法。”

随着时间的推移，在大数据开源社区的帮助下，Hudi
已经发展成为一个通用的大数据存储系统，使得以下特性成为可能：

- 摄取和查询引擎之间的快照隔离，包括 Apache Hive、Presto 和 Apache Spark；
- 支持回滚和存储点，可以恢复数据集；
- 自动管理文件大小和布局，以优化查询性能和目录清单；
- 准实时摄取，为查询提供最新数据；
- 实时数据和列数据的异步压缩。

为了证明 Hudi 的可扩展性，目前 Uber 使用它管理着 4000 多个表，这些表存储了几 PB 的数据，同时将 Apache Hadoop 仓库访问延迟从几个小时降低到 30 分钟以下。Hudi 还为数百个增量数据管道提供了支撑，与该公司以前使用的解决方案相比，它的成本更低，效率更高。

未来，Hudi 将与 Apache 软件基金会进行合作，有关技术文档和社区参与指南，请参阅 Apache Hudi 项目页面。
本文原文：Uber Submits Hudi, an Open Source Big Data Library, to The Apache Software Foundation