官宣，Apache Hudi 正式成为 Apache 顶级项目

2020年6月4日，马萨诸塞州韦克菲尔德（Wakefield, MA）—— Apache 软件基金会（ASF），超过350个开源项目和计划的全志愿者开发人员、管理人员和孵化器，正式宣布 Apache Hudi 成为顶级项目（Top-Level Project 、TLP）。

如果想及时了
解Spark、Hadoop或者Hbase相关的文章，欢迎关注微信公共帐号：iteblog_hadoop

Apache Hudi (Hadoop Upserts delete and Incrementals) 数据湖技术支持在Apache Hadoop 兼容的云存储和分布式文件系统之上进行流处理。该项目最初于 2016 年由 Uber 开发(代号和发音为“Hoodie”)，2017年开源，并于2019年1月提交给 Apache 孵化器。

“在孵化器中学习和发展 Apache 之路是一次有益的经历” Apache HUdi 的副总裁维诺斯·钱达尔(Vinoth Chandar)说。“作为一个社区，我们为我们共同推进这个项目所取得的进步感到谦虚，同时，也为未来的挑战感到兴奋。”

Apache Hudi 用于在 Apache Hadoop 分布式文件系统（HDFS）或云存储上提供诸如 upserts 和增量变更流（incremental change streams）等流处理原语来管理 PB 级的数据湖。Hudi 数据湖提供了新的数据，同时比传统的批处理效率高一个数量级。功能包括:

- 支持快速、可插拔索引的更新/删除
- 以事务的形式提交/回滚数据
- 以流处理的方式更改从 Hudi 表捕获的数据
- 支持Apache Hive，Apache Spark，Apache Impala 和 Presto 查询引擎
- 内置数据提取工具，支持 Apache Kafka，Apache Sqoop 和其他常见数据源
- 通过管理文件大小，存储布局来优化查询性能
- 基于行的快速摄取格式，并支持异步压缩为列格式

- 时间线元数据以进行审计跟踪

Apache Hudi 目前在阿里巴巴集团、EMIS Health、Linknovate、Tathastu.AI，腾讯和 Uber 等组织中使用，并且 Amazon Web Services 中的 Amazon EMR 也对其提供支持。
详细的使用列表可以参见 https://hudi.apache.org/docs/powered_by.html

本文翻译自 The Apache Software Foundation Announces Apache® Hudi™ as a Top-Level Project