

## Spark Delta Lake 现在由Linux基金会托管，将成为数据湖的开放标准



如果想及时了解Spark、Hadoop或者HBase相关的文章，欢迎关注微信公众号：iteblog\_hadoop

一年一度的 Spark + AI Summit Europe

峰会于2019年10月15-17日在欧洲的阿姆斯特丹举行。在10月16日 数砖和 Linux 基金会共同宣布 Delta Lake 和 将成为一个 Linux 基金会项目（参考：[The Delta Lake Project Turns to Linux Foundation to Become the Open Standard for Data Lakes](#)

）。该项目旨在建立一个开放的标准来管理数据湖中的大量数据，同时 Delta Lake 同样基于 Apache 2.0软件许可证开源。



**Databricks**

88,424 followers

5h • Edited •  Anyone



Announced at [#SparkAISummit](#): **The Linux Foundation** will host the open source project Delta Lake. This will allow for an open governance model that encourages participation and technical contribution, and will provide a framework for long-term stewardship by an ecosystem invested in Delta Lake's success. <https://lnkd.in/eafEpnA>



## Delta Lake Now Hosted by the Linux Foundation to Become the Open Standard for Data Lakes

[databricks.com](https://databricks.com)

如果想及时了解Spark、Hadoop或者HBase相关的文章，欢迎关注微信公众号：iteblog\_hadoop

Delta Lake 致力于提高数据湖的可靠性和可扩展性。它的高级抽象和保证，包括ACID事务和时间旅行，极大地简化了真实世界数据工程体系结构的复杂性。自从今年四月底的 Spark + Ai 峰会数砖宣布开源 Delta Lake 项目（[重磅 | Apache Spark 社区期待的 Delta Lake 开源了](#)）。该项目每个月被部署在数千个组织中，成为数据和人工智能架构中不可或缺的基础组件。

为了进一步推动社区的采用和发展，数砖决定与Linux 基金会合作，利用他们的平台和丰富的经验来培育有影响力的开源项目，包括Linux本身、Jenkins和Kubernetes。数砖与阿里巴巴、Booz Allen Hamilton、英特尔和Starburst 公司一起宣布，不仅要为Apache Spark的Delta Lake 开发支持，还要对Apache Hive、Apache Nifi 和Presto 的开发进行支持。

## 为更健壮的数据湖提供丰富的特性集

正如前面所讨论的，Delta Lake 使数据湖中的数据更容易处理，也更健壮。它旨在解决数据湖中常见的许多问题。例如，不完整的数据摄入可能导致数据损坏；为了解决这个问题Delta Lake 引入了 ACID事务，支持多个数据管道并发地向数据湖读写数据。提供数据湖的数据源可能无法提供完整的列数据或正确的数据类型，因此 Schema Enforcement 可以防止坏数据导入到数据湖中从而导致数据损坏。更改数据捕获和 update/delete/upsert的支持允许非Append 的作业在数据湖上正常工作，这对于GDPR/CCPA来说是必须的。更多关于Delta Lake 的介绍可以参见过往记忆大数据之前的文章: [重磅 | Apache Spark 社区期待的 Delta Lake 开源了](#)。

Delta Lake 的功能还在不断增强，比如 Delta Lake 0.4.0 支持 Python Api，支持一些简单功能的 SQL 处理，参见过往记忆大数据之前的文章: [Spark Delta Lake 0.4.0 发布，支持 Python API 和部分 SQL](#)。相信在未来 Delta Lake 还会支持更多的功能。

## 数据湖的开放性和可扩展性

Delta Lake 设计的主要原则是开放性和可扩展性。Delta Lake将所有数据和元数据存储于云对象存储中，采用开放协议设计，利用了JSON 和 Apache Parquet 等现有的开放格式。这种开放性不仅消除了厂商锁定的风险，而且对于数据科学、机器学习和SQL的无数不同用例来说，也是至关重要的。

为了确保项目的长期发展和社区的发展，所以数砖与Linux基金会合作，进一步发扬这种开放精神。

## 开放 Delta Lake 管理和社区参与

很高兴Linux基金会现在将Delta Lake作为中立的项目，并采用开放治理模型来鼓励参与和技术贡献。这将有助于为 Delta Lake 项目的长期管理提供一个框架，并为 Delta Lake 社区建立良好的生态系统，开放 Delta Lake 存储的标准。我们相信，这种方法将有助于确保存储在 Delta Lake 的数据保持开放和可访问性，同时推动更多的创新和发展，以解决这一领域的挑战性问题。

Databricks 团队为data & AI生态系统创建并贡献了各种开源项目，包括Apache Spark、MLflow、Koalas和Delta Lake。数砖团队将继续参与开源社区，因为这是将新功能引入市场的最快、最全面的方式；能够建立一个可持续的、健康的业务，同时与社区保持联系，以确保项目不会将客户锁定在专有系统或数据格式中。

## 关于 Linux 基金会

Linux基金会是一个非盈利性的联盟，其目的在于协调和推动Linux系统的发展，以及宣传、保护和规范Linux，该组织是2007年，由开源码发展实验室（Open Source Development Labs，OSDL）与自由标准组织（Free Standards Group，FSG）联合起来成立的，其中MeeGo是Linux基金会管理下的Linux操作系统。

本文翻译自 <https://databricks.com/blog/2019/10/16/delta-lake-now-hosted-by-the-linux-foundation-to-become-the-open-standard-for-data-lakes.html>

本博客文章除特别声明，全部都是原创！  
原创文章版权归过往记忆大数据（[过往记忆](#)）所有，未经许可不得转载。  
本文链接: **【】**（ ）