这可能是学习 Spark Delta Lake 最全的资料

Delta Lake 是一个存储层，为 Apache Spark 和大数据 workloads 提供 ACID 事务能力，其通过写和快照隔离之间的乐观并发控制（optimistic concurrency control），在写入数据期间提供一致性的读取，从而为构建在 HDFS 和云存储上的数据湖（data lakes）带来可靠性。Delta Lake 还提供内置数据版本控制，以便轻松回滚。

为了更好的学习 Delta Lake ，本文收集了互联网上各种关于 Apache Spark Delta Lake 的资料，供大家学习交流，包括
PPT、视频、源码解析、技术文章等；另外，本文也会持续不断更新。

PPT & 视频

目前关于 Delta Lake 的视频和 PPT 基本上都来自 Databricks 的大神们在各种场合分享的，本文对这些分享的资料进行了收集，具体如下：

The Delta Architecture: Delta Lake + Apache Spark Structured Streaming

本分享来自QCon 2019 上海，分享者李潇
数据工程师的纠结与运维的凌乱
- Delta Lake基本原理
- Delta 架构
- Delta 架构的特性
- Delta 架构的经典案例 & Demo
- Delta Lake 社区

PPT下载：https://download.csdn.net/download/w397090770/11930387

New Developments in Open Source Ecosystem: Apache Spark 3.0, Koalas, Delta Lake

本分享来自9月26日的云栖大会，分享者李潇。Apache Spark 3.0 and Koalas的最新进展，本议题相关文章 云栖大会｜Apache Spark 3.0 和 Koalas 最新进展。
完整视频和 PPT 请关注 过往记忆大数据 公众号并回复 spark_yq 获取。x

Delta Lake - Open Source Reliability for Data Lakes

本分享来自 Michael Armbrust，负责 Delta Lake 的首席工程师，也是 Spark SQL 和 Structured Streaming 的核心开发者。这篇 PPT 介绍的比较详细，涉及到 Delta Lake 项目诞生背景、核心功能以及实现原理等。
配套视频 & PPT 下载：关注 开发爱好者社区（bigdata_ai）微信公众号，并回复 2596_1 获取。

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

Making Apache Spark™ Better with Delta Lake

也是 Michael Armbrust 分享的。主题主要包括以下内容：

- Apache Spark 在大数据处理中的作用；
- 使用数据湖作为数据架构的重要组成部分；
- 数据湖可靠性挑战；
- Delta Lake 如何为 Spark 提供可靠的数据
- Delta Lake 具体改进
- 采用 Delta Lake 为您的数据湖提供动力

配套视频 & PPT 下载：关注 开发爱好者社区（bigdata_ai）微信公众号，并回复 2596_2 获取。

Getting Data Ready for Data Science

分享者 Prakash Chockalingam，他是 Databricks 的产品经理。本 PPT 主要内容为：

- 数据科学生命周期
- 数据工程对数据科学的重要性
- 现代数据工程的关键原则

- Delta Lake 如何帮助为分析提供可靠的数据
- 采用 Delta Lake 为您的数据湖提供动力的便利性
- 如何在您的数据基础架构中加入 Delta Lake 以启用数据科学

配套视频下载：关注 开发爱好者社区（bigdata_ai）微信公众号，并回复 2596_3 获取。

Simplify and Scale Data Engineering Pipelines with Delta Lake

分享者 Joe Widen（Databricks 的高级解决方案架构师）以及 Denny Lee（Databricks 的开发人员、倡导者）
配套视频 & PPT 下载：关注 开发爱好者社区（bigdata_ai）微信公众号，并回复 2596_4 获取。

Next-generation scalable data lakes

分享者 Prakash Chockalingam，他是 Databricks 的产品经理。
配套 PPT 下载：关注 开发爱好者社区（bigdata_ai）微信公众号，并回复 2596_5 获取。

Delta Architecture, A Step Beyond Lambda Architecture

分享者 Prakash Chockalingam，他是 Databricks 的产品经理。
Lambda architecture is a popular technique where records are processed by a batch system and streaming system in parallel. The results are then combined during query time to provide a complete answer. Strict latency requirements to process old and recently generated events made this architecture popular. The key downside to this architecture is the development and operational overhead of managing two different systems.

There have been attempts to unify batch and streaming into a single system in the past. Organizations have not been that successful though in those attempts. But, with the advent of Delta Lake, we are seeing lot of our customers adopting a simple continuous data flow model to process data as it arrives. We call this architecture, The Delta Architecture.

In this webinar, we cover the major bottlenecks for adopting a continuous data flow model and how the Delta architecture solves those problems.
配套 PPT 下载：关注 开发爱好者社区（bigdata_ai）微信公众号，并回复 2596_6 获取。

Building Streaming Data Pipelines Using Structured Streaming and Delta Lake

分享者：Tathagata (TD) Das ，Databricks 高级软件工程师。

配套 PPT 下载：关注 开发爱好者社区（bigdata_ai）微信公众号，并回复 2596_7 获取。

技术文章 & 源码解析

- Apache Spark Delta Lake 删除使用及实现原理代码解析
- Apache Spark Delta Lake 更新使用及实现原理代码解析

- Apache Spark Delta Lake 写数据使用及实现原理代码解析
- Apache Spark Delta Lake 事务日志实现源码分析
- 深入理解 Apache Spark Delta Lake 的事务日志（中文）、Diving Into Delta Lake: Unpacking The Transaction Log（英文）
- Apache Spark 社区期待的 Delta Lake 开源了
- Announcing the Delta Lake 0.3.0 Release
- Productionizing Machine Learning with Delta Lake
- Migrating Transactional Data to a Delta Lake using AWS DMS
- Accurately Building Genomic Cohorts at Scale with Delta Lake and Spark SQL
- Efficient Upserts into Data Lakes with Databricks Delta
- Introducing Delta Time Travel for Large Scale Data Lakes
- Processing Petabytes of Data in Seconds with Databricks Delta

帮助文档

- 数砖 Delta Lake 产品文档：Delta Lake
- Delta Lake 开源项目文档：https://delta.io/

项目地址

Delta Lake： https://github.com/delta-io/delta

未来规划

总体来说，未来版本将支持 Python & SQL API。

- 0.4.0 规划
- 长期规划