**Constance: An Intelligent Data Lake System**

**SIGMOD 2016 CCF-A**

Rihan Hai, Sandra Geisler, Christoph Quix, RWTH Aachen University

1. 摘要：

我们的时代挑战，大数据还有很多的困难，特别是在数据的多样性上面。数据来源的多样性经常会造成信息孤岛(information silos)，一个具有异构的模式、不同的查询语言和API的非集成数据管理系统的集合。数据湖系统被认为是解决此类问题的方法，它为原始数据提供了一个具有公共访问接口的无模式存储库。但是，仅仅只是将数据倾倒至数据库而没有任何的元数据管理的话，只会导致“数据沼泽”。为了避免这个，我们提出Constance，一个数据湖系统，对从异构数据源解压的，带有对原始数据复杂的元数据管理，Constance从数据源发现，解压和概括结构化的元数据，并且对数据和元数据标注语义信息以避免模糊性。通过嵌入支持结构化和半结构化的查询重写引擎，Constance为查询处理和数据探索提供了一个统一的接口。在演示中，我们将遍历Constance的所有功能部件。Constance将应用于两个实际的用例，以向与会者展示我们通用和可扩展的Data Lake系统的重要性和有用性。

2. 前言：

数据湖解决了大数据令人畏惧的挑战：“如何轻松使用高度多样化的数据并提供知识？”大量的数据是可用的，但是这些数据往往被信息孤岛所隔离，彼此之间没有联系或者只有很松的内在联系。但是，有价值的洞见往往对数据孤岛中的信息进行综合和集成性的分析后才可以看到。为了填补这个鸿沟，