発表概要

AlphaSQL：SQLファイル集合の型・スキーマ解析と自動並列化

松井 誠泰[1,a)]　杉崎 琢人[1,b)]　越塚 登[2,c)]

2020年10月29日発表

機械学習やデータ解析などの技術は強力で柔軟なシステムを実現するが，データと処理の依存関係やデータ・処理それぞれの変更への追随など管理の難しさがある．そこで本発表では，型・スキーマの安全性を保ちながらデータ基盤の構築を行うための AlphaSQL というフレームワークを提案する．AlphaSQL は，SQL ファイル間の依存関係を解決することにより，SQL ファイル集合全体の型・スキーマ解析と並列化を行う．テーブルを作成する SQL 文に対する，作成されたテーブルを参照するクエリの依存関係が，SQL ファイルの静的解析によって自動的に抽出される．結果は有向非循環グラフとして出力され，ユーザは依存関係を視覚的に確認することができる．既存のワークフローツールとは異なり，ユーザは SQL ファイル間の複雑な依存関係に注意したり，並列化のための追加の作業を行う必要はない．また，解析の過程で構文の誤り・型の不整合を含む一般的なエラーが排除されることを，実際の CI での分析結果から確認することができた．一部の SQL は機械学習モデルのトレーニングとデプロイをサポートしはじめており，AlphaSQL は機械学習にも有効だと考えられる．AlphaSQL は Github で公開されており，主に BigQuery で現在使用されているが，AlphaSQL が依存する分析フレームワーク ZetaSQL は Standard SQL2011 とほぼ互換性があり，多くの SQL に対応しやすいと考えられる．https://github.com/Matts966/alphasql

Presentation Abstract

AlphaSQL: Integrated Type/Schema Check and Parallelization for SQL File Set

Masahiro Matsui[1,a)]　Takuto Sugisaki[1,b)]　Noboru Koshizuka[2,c)]

Presented: October 29, 2020

Emerging technologies such as machine learning and data mining realize powerful and flexible systems, however, they cause some problems. For example, management of dependency relationships between data and processing and their changes are typical pains. We present a framework named AlphaSQL to build type/schema safe and efficient data lake, data warehouse, and data mart. AlphaSQL provides integrated type/schema check and parallelization for SQL file set by resolving dependencies between SQL files. The dependencies of table references on the statement creating the tables are automatically resolved by static analysis of SQL files. The framework outputs the result as a directed acyclic graph and users can check the visualization of the dependencies. Unlike other existing workflow tools, users do not have to care about and code the complex dependencies between SQL files. Based on the resolution results, the SQL files are checked to eliminate typical errors including syntax errors, schema errors such as unknown columns and incompatible types, and executed parallelly. In addition, it was confirmed from the actual CI analysis results that typical errors including syntactical errors and type/schema inconsistencies were eliminated in the analysis process. These features are also useful in actual machine learning environment because some SQLs support training and deployement of machine learning models. AlphaSQL is open on Github and currently used mainly for BigQuery, however, we can extend the framework easily because the analysis framework ZetaSQL that AlphaSQL depends on is almost compatible with the standard SQL 2011. https://github.com/Matts966/alphasql

This is the abstract of an unrefereed presentation, and it should not preclude subsequent publication.

[1] 東京大学大学院学際情報学府
Graduate School of Interdisciplinary Information Studies, The University of Tokyo, Bunkyo, Tokyo 113–8654, Japan
[2] 東京大学大学院情報学環
Interfaculty Initiative in Information Studies, The University of Tokyo, Bunkyo, Tokyo 113–8654, Japan

[a)] masahiro.matsui@koshizuka-lab.org
[b)] takuto.sugisaki@koshizuka-lab.org
[c)] noboru@koshizuka-lab.org