

# 代表的な拡張機能の紹介

—In-Database Processing—

株式会社KSKアナリティクス



# 1. In-Database Processing Extension

## In-Database Processingとは

今回はIn-Database Processing Extensionのご紹介をしたいと思います。In-Database Processing（データベース内処理）は、その名の通りデータベースシステム内でデータの準備や前処理を行うことを指します。

生のデータセットをローカル環境にダウンロードして前処理を行う為には、大量のデータに耐えうるNW環境や端末の性能が必要な上、ダウンロードにも前処理にも長い時間が掛かります。更にクラウドなどからデータを取得する際に、容量に応じた課金が発生することもあります。

In-Database Processingでは処理はデータベース内で行われ、実際に取得するのはプロセスの結果のみなので、時間や費用などの各種のコストを節約することが出来ます。

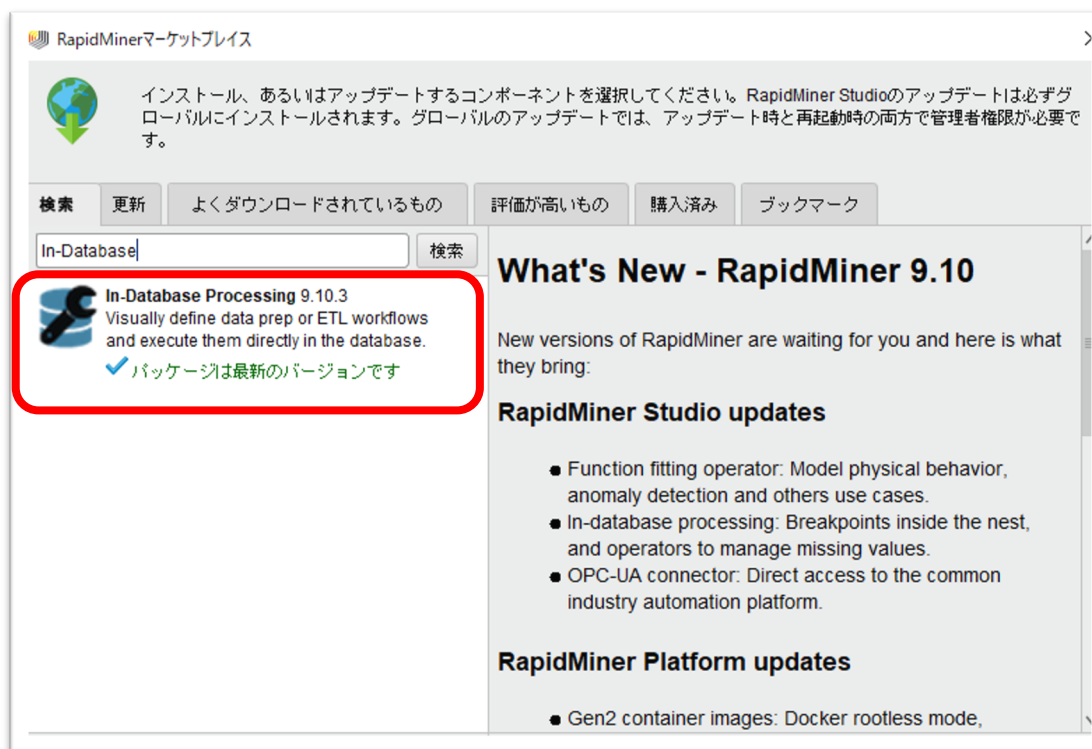
RapidMinerのプロセス上でIn-Database Processingを実装することで、上記メリットに加えて、SQLで書く必要がなくなりますので、より効率的にデータ処理を行うことが出来るExtensionです。

# 1. In-Database Processing Extension

## Extensionの準備

以前ご紹介した通り、マーケットプレイスからインストールか、下記WEBページからダウンロードし手動で適用します。

[https://marketplace.rapidminer.com/UpdateServer/faces/product\\_details.xhtml?productId=rmx\\_in\\_database\\_processing](https://marketplace.rapidminer.com/UpdateServer/faces/product_details.xhtml?productId=rmx_in_database_processing)



## 2. オペレータ紹介

### In-Database Nest



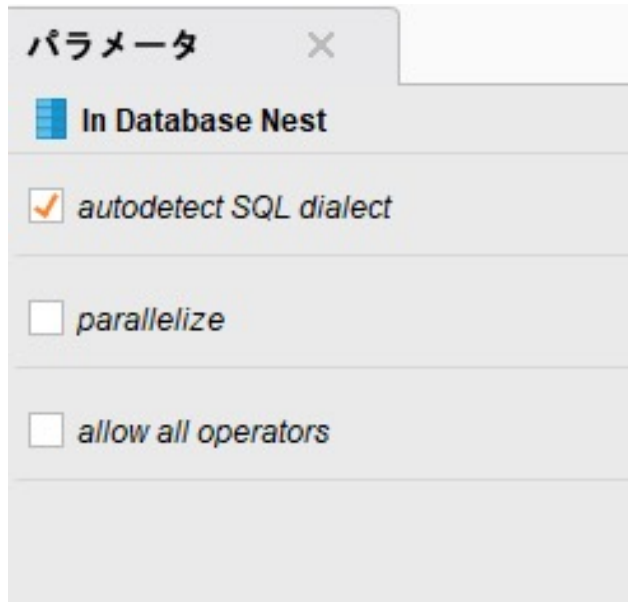
- In Database Nest

このオペレータに任意のデータベースを接続し、サブプロセスで構築されたプロセスをSQLに変換し、そのコードを更に任意のデータベースに送信します。

この内部では、In-Database Processingグループのオペレータのみを使用できます。各出力は、そのポートに接続されているオペレータによって生成されたクエリのインメモリデータセットの結果を配信します。並列化を有効にすると、これらのクエリがデータベース内で同時に実行されることがあります。

## 2. オペレータ紹介

### In-Database Nest



#### パラメータ解説

- autodetect SQL dialect

接続設定から自動的にSQLの形式を設定します。

通常はオンになっています。手動設定したい場合は外します。

- parallelize (並列化)

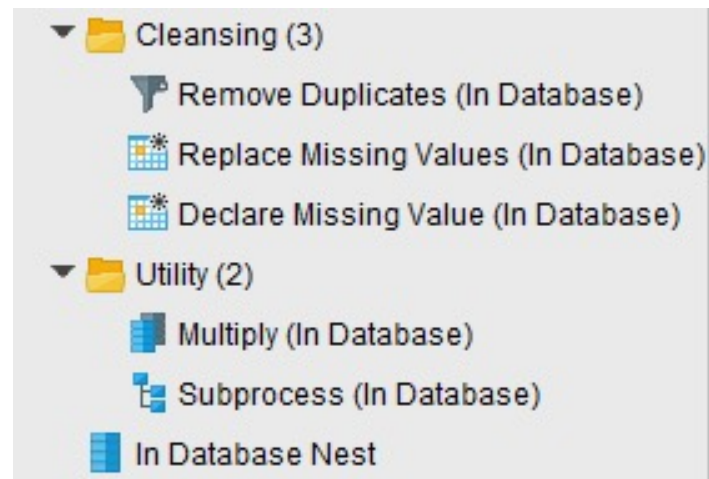
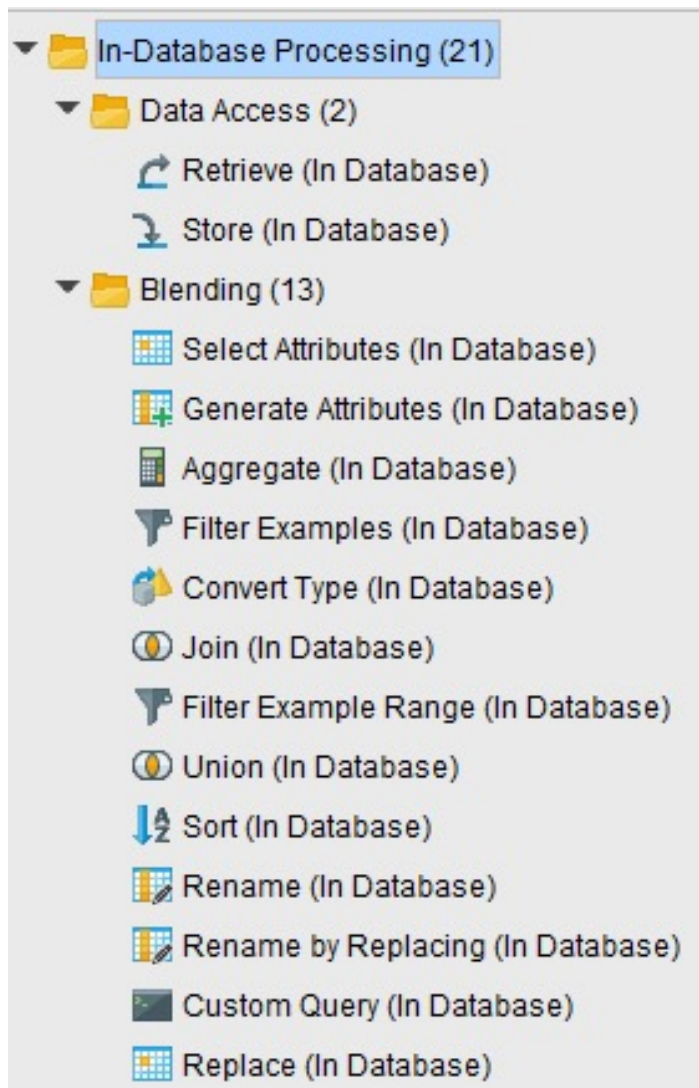
チェックすると、複数のクエリを同時にデータベースに送信します。複数のネスト出力には複数のクエリが必要ですが、それらを並行して実行します。この場合、ソースデータベースの負荷が大幅に増加する可能性があることに注意してください。

- allow all operators

In-Database Processingグループのオペレータ以外の使用を許可します。ただし、意図しない動作をする可能性があり、動作保証もないためチェックは推奨しません。万一の際のみの利用としてください。

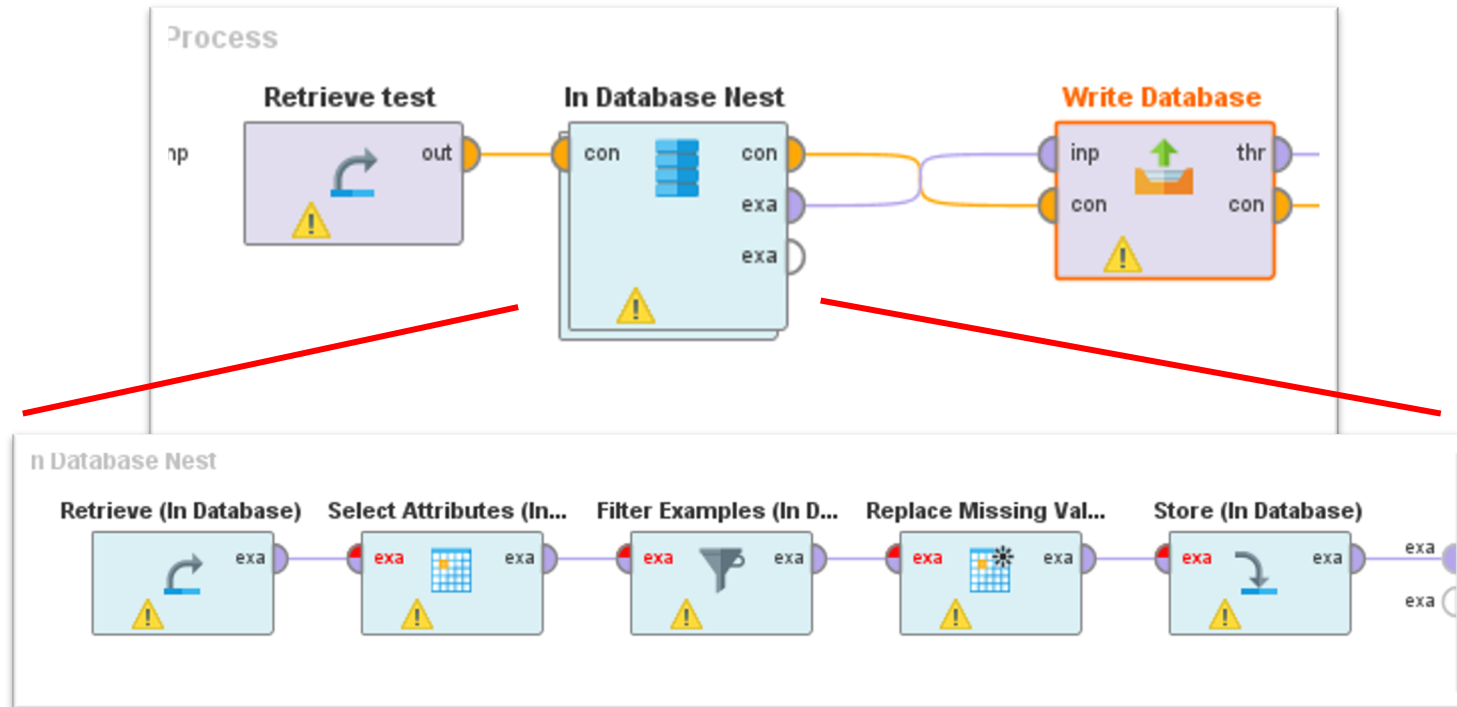
## 2. オペレーター紹介

### 収録オペレーター一覧



# 3. 使用方法

## 構築例



In-Database用のオペレータを使用していますが、各オペレータの使い方は、通常のものとはほとんど変わりません。RetrieveとStoreについてはSchemaとTableを指定する必要があります。

## 3. 使用方法

### メリット

In-Database Processingでは、データベースシステムの処理能力を活用による処理速度の向上や、通信量の節約による高速化と省コストが図れる上、RapidMinerで実装するとSQLを書く必要が無いため、SQL知識が無くてもデータベース上の処理を視覚的に実行できます。

SQLに慣れている方にとっても、RapidMinerでの処理を行うことは組織内での展開力という意味で大きなメリットがあります。

この拡張機能は現在、Google BigQuery（OAuth 2経由）、PostgreSQL、MySQL、MSSQL、およびOracleをサポートしています。ユーザーの需要に応じて、データベースとクラウドエンジンのさらなるサポートが将来的に計画されています。



## 4. おわりに

### 更なる活用へ向けて

In-Database Processing Extensionはいかがでしたでしょうか。

データベース上のデータ処理は組織内の大規模データを利活用する上で非常に重要な手法ですが、SQL知識が必要なため、誰でもできるものではありませんでしたが、RapidMinerと連携することで通常の前処理と同じように行うことが出来ます。

このExtensionはRapidMiner社によるサポートもされており、更新頻度も高い拡張機能になっています。サポートするデータベース形式を増やしていくことも計画されていますので、更なる進化が期待できます！