



スクエアFreeセミナー



Data Health

オープンソースETL開発ツール Talend Studioのご紹介

Talend株式会社
リードソリューション エンジニア 三浦大洋
2022年2月24日



本日のアジェンダ

セミナーご参加ありがとうございます。TalendのETL開発について詳しくご紹介します

1. Talendとは
2. オープンソースETL開発ツール
Talend Studio
3. Talend Open Studioデモ
4. おわりに





Talendとは



Talend会社概要

企業におけるデータ利活用効率化を支援、データのビジネス成果への変換に貢献



2005年 フランスで創業

本社所在地 米国カリフォルニア州レッドウッドシティ

従業員数 1,500人、26拠点

Talend利用ユーザー 全世界7,250社以上

クラウドエコシステムとの強いパートナーシップ



2006

Data integration
Data Quality

2013

Big Data
Hadoop 2.0

2015

Talend on AWS
Spark
Spark Streaming
API Services
Multi-cloud

2019

Talend on Azure
Data Catalog
Cloud Apps
Self Service Apps

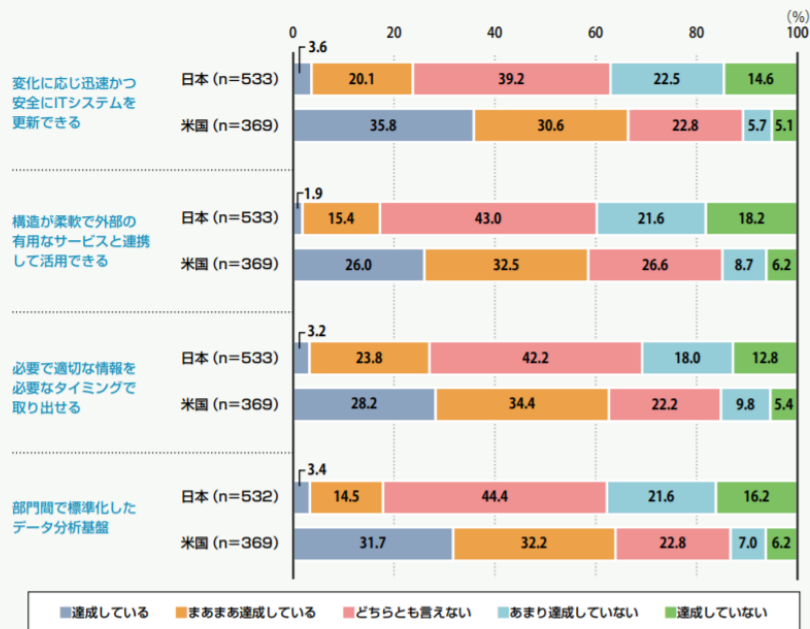
2021

Data Quality Service
Trust Score
Data Inventory

日本企業におけるデータ利活用の現状①

外部サービスとの連携、データ収集の仕組み、データ分析基盤の整備が求められる

図表 14-1 ビジネスニーズに対応するためにITシステムに求められる機能(達成度)

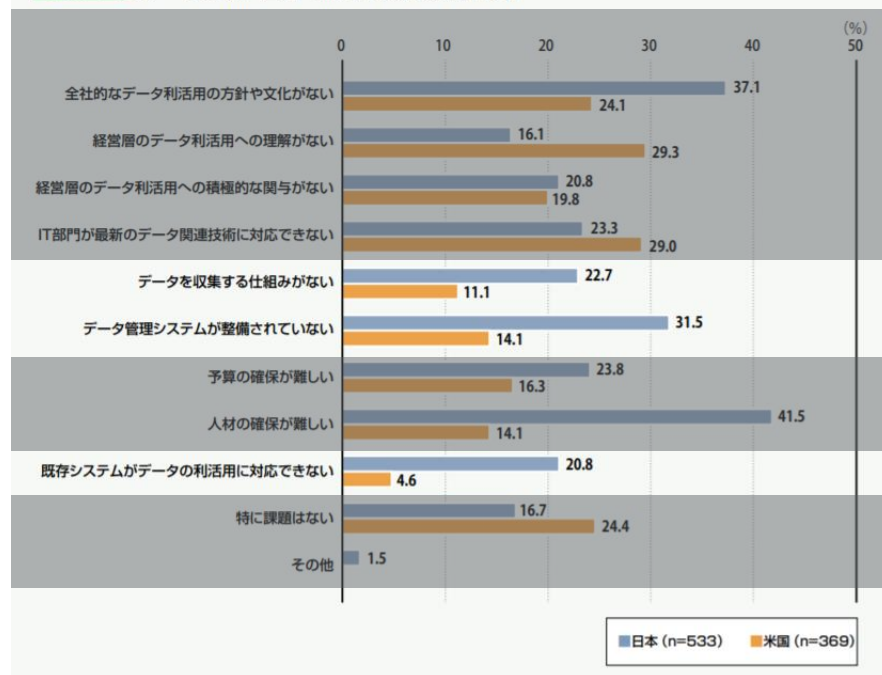


- データを分析し、顧客の真のニーズを捉えて早期にサービスを立上げ、改善を繰り返すことで顧客価値を高めていくことも重要である
- 米国企業では「まあまあ達成している」も含めると**5割以上**の企業が「変化に応じた迅速かつ安全なITシステムの更新」や「外部サービスとの連携」といったビジネスニーズを満たしている
- 独立行政法人 情報処理推進機構「DX白書2021」より

日本企業におけるデータ利活用の現状②

データ収集の仕組みやデータ管理システムの整備が課題

図表14-4 データ整備・管理・流通の課題(複数回答)



- 予測困難な外部環境変化に俊敏に対応するために、データに基づき経営や現場の意思決定を行う**データドリブン経営**の重要性が高まっている。しかし、日本企業は「適切な情報を必要なタイミングで取り出せる」ことや「部門間で標準化したデータ分析基盤」の実現といったビジネスニーズへの**対応が十分できていない**。
- 独立行政法人 情報処理推進機構「DX白書2021」より

3つのデータ活用市場で選択されるTalend

多くの企業がTalendを使用してデータ活用を効率化しビジネスゴールを達成してます



データ民主化
データ分析



データマネジメント
データガバナンス



AI活用
オートメーション

データ連携基盤 (ETL)

データカタログ基盤

データレイク構築

マスターデータ管理

ビッグデータ処理

セルフサービスアプリ

クラウド移行

データストリーミング

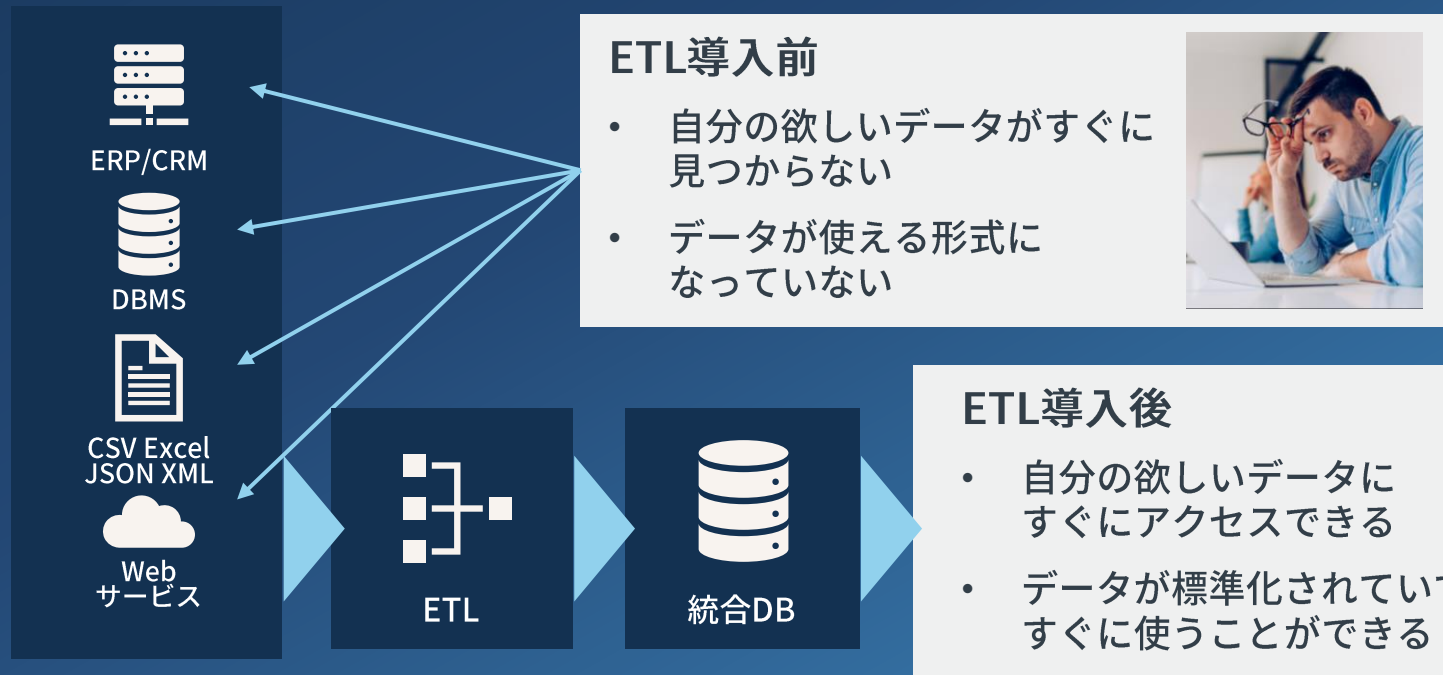


Talend Data Fabric / Talend Professional Service

企業におけるデータ利活用を効率化する
統合データマネジメントプラットフォームの構築を支援

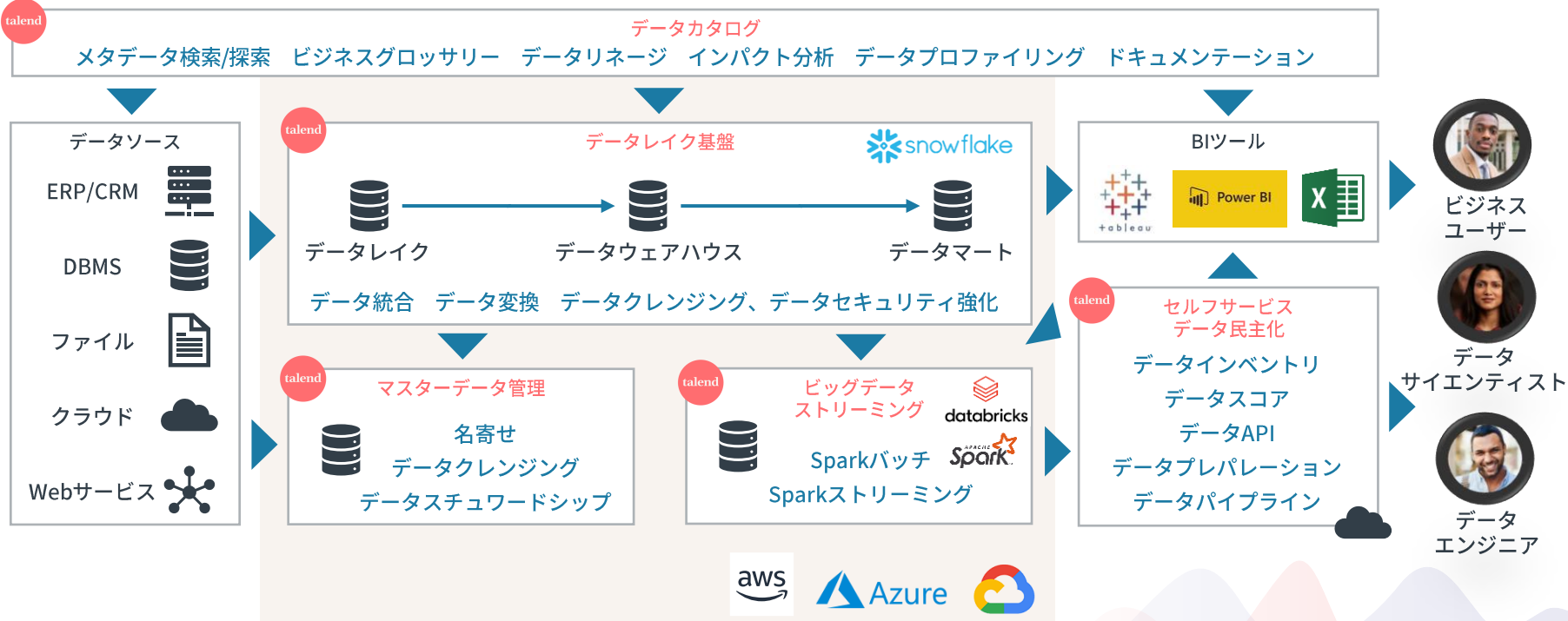
Why ETL

ETL：業務システムからデータを抽出（Extract）、使える形式に変換（Transfer）
DWHなどの統合DBにロード（Load）する仕組み



Talendで実現するデータ活用基盤

データ活用の様々なニーズに対応する統合型データマネジメントプラットフォーム





オープンソースETL開発ツール Talend Studio



Talend Studio

データ統合/データ連携の統合開発環境



2,000以上のコネクタ
とコンポーネント



汎用コードの
自動生成



ドラッグ&ドロップツ
ールによる直感的なUI

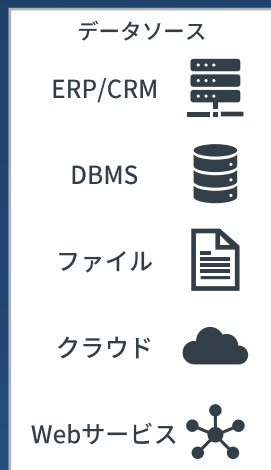


共有リポジトリ

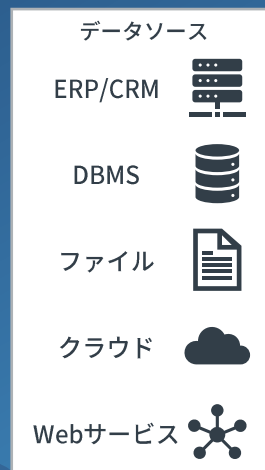
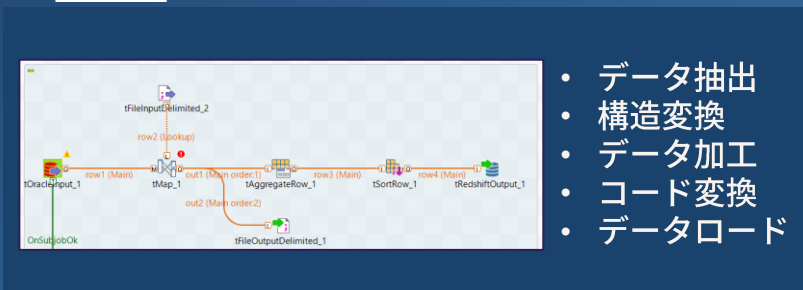


データ統合（ETL）アプリケーションの開発

- GUIベースのデータ統合/データ連携開発環境
- 機能別部品を接続するシンプルな操作性によりデータ統合のフローを容易に開発
- データ統合の様々な局面をサポートする充実した機能
- 開発とデバッグの同時進行、仕様書やテストの自動生成といった支援機能が豊富



GUI統合開発環境
Talend Studio



Talend Studio : あらゆるシステムと接続

 SAP Business Warehouse	 Box	 Amazon DynamoDB	 Amazon MySQL
 Amazon Aurora	 Amazon EMR	 MQTT	 Amazon Kinesis
 Machine Learning	 Apache Kafka	 MapReduce	 SOAP
 REST API	 Microsoft Azure Storage	 MongoDB	 Microsoft HDInsight
 Google BigQuery	 SAP HANA	 Salesforce	 Amazon Redshift
 PostgreSQL	 Oracle	 Amazon S3	 Exasol
 Dropbox	 Microsoft .net	 ParAccel	 SPSS
 SAS	 Google Cloud Platform	 Amazon Web Services	 SageX3

 ServiceNow	 Mondrian	 Amazon RDS Oracle	 Centric CRM
 SAP S/4HANA	 Netezza	 NetSuite	 MySQL
 Microsoft Dynamics CRM	 Microsoft SQL Server	 Microsoft Dynamics AX	 SAP MaxDB
 H2	 Apache Derby DB	 Marketo	 Java DB
 EDI x 12	 Apache Parquet	 HQM and JMS	 Ingres
 Excel	 Apache Orc	 LDAP	 Google Cloud Pub/Sub
 SCP	 Apache Avro	 Google Cloud DataProc	 Apache Storm
 HyperSQL	 Amazon SQS	 Microsoft Azure CosmosDB	 HP Vertica

 Microsoft Azure SQL Data Warehouse	 Open Bravo	 Microsoft Azure Data Lake Store	 Palo
 Interbase	 XML	 Microsoft Azure Queue Storage	 Apache Flume
 Microsoft Azure Blob Storage	 Cassandra	 Informix	 Snowflake
 Microsoft Azure Table Storage	 Hive	 Delimited File	 EDB
 Greenplum	 Cloudera Alux	 Elastic	 Sqoop
 Neo4j	 TDM	 MemSQL	 Cloudera Impala
 CouchDB	 Couchbase	 Apache Hbase	 Google Storage
 FTP	 HL7	 Riak	 EDIFACT

 Google Drive	 Pig	 Firebird	 MarkLogic
 HCatalog	 HP CloudStart	 SAP Sybase	 SAP ECC
 SugarCRM	 SQLite	 Teradata	 Microsoft Azure
 MapR Streams	 Microsoft OLE		

詳細は以下のURLを参照:

https://help.talend.com/r_eader/nQI6Cat_RfMT4hy0BHZLSA/Azb5~coWQdOJ0Wd6M0YOfg

Talend Studio：データ加工/変換処理

データクレンジングに必要とされる機能を豊富に提供
それぞれの機能は独立しているため処理プロセスに応じて自由に組み合わせ可能

構造変換

スキーマ変換（カラム分割/結合/追加/削除）
複数スキーマのジョイン、単一スキーマの分割
正規化/非正規化、構造化⇄半構造化

データ整形

コード変換
データ型変換
フォーマット変換

フィルタリング

除外データ抽出
（ルールベース、ルックアップ、重複削除）

データ変換

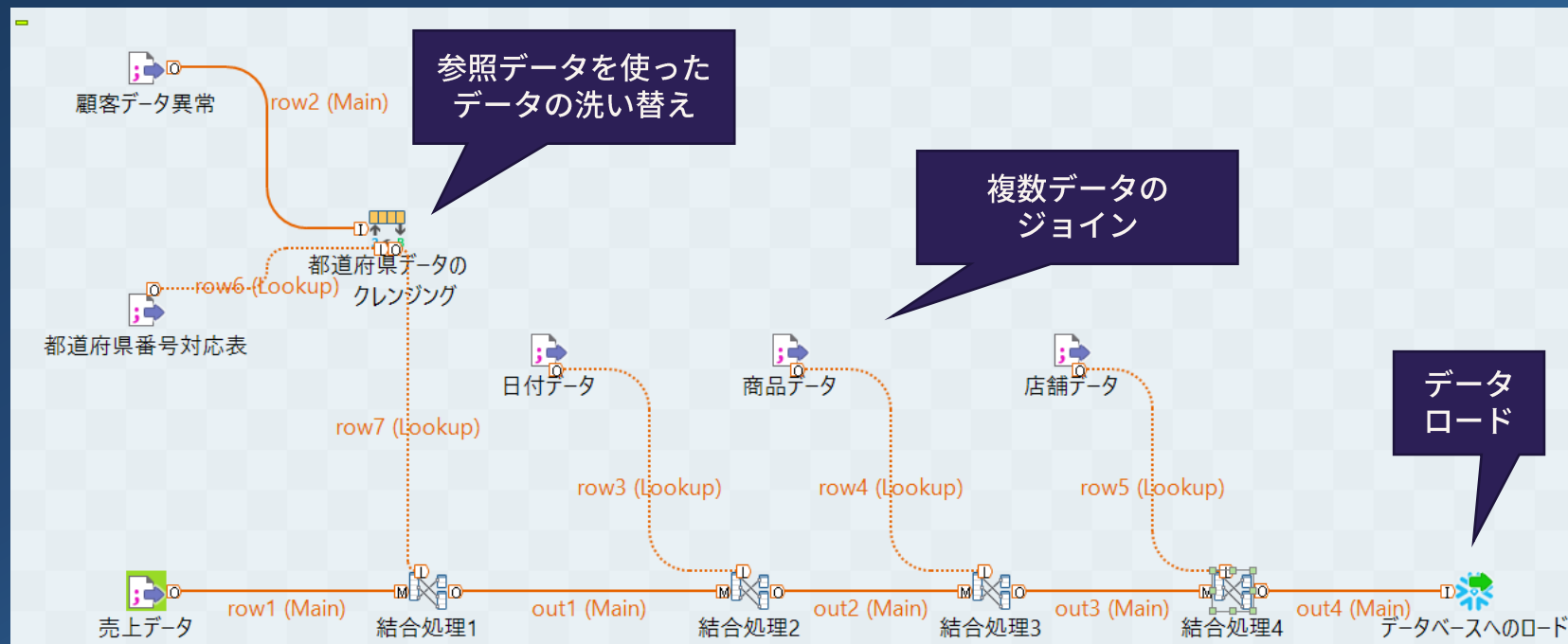
文字列操作、数値演算、日付処理
置換（ルックアップ、ルールベース）
クレンジング

フロー制御

ジョイン/ユニオン
除外データの分離、データによるフロー分岐
前処理/後処理
繰り返し処理/パラレル処理

データ統合アプリケーションの例

売上データと顧客データ、商品データ、店舗データなどを結合してDWHにロードする例
顧客データの都道府県に異常があり、対応表をもとにクレンジング



ハンドコーディングとの比較

統合開発ツール「Talend Studio」を使ったGUI開発による開発生産性向上

- GUIによるバッチジョブ開発フレームワークを提供
- インプット/アウトプットとデータ加工処理が分離
- 複数のアプローチから最適なものを選択可能（GUI/プログラミング）
- 実データやサンプルデータを使用し結果を確認しながら開発
- 既存Javaプログラムの流用



Talend Open Studioデモ



Talend Open Studioデモ

分散しているデータを統合しデータをクレンジングするETLバッチの開発



データ統合
開発 (ETL)



データ
クオリティ



システム連携
開発 (API)



データ
カタログ



データ
インベントリ



データ
パイプライン



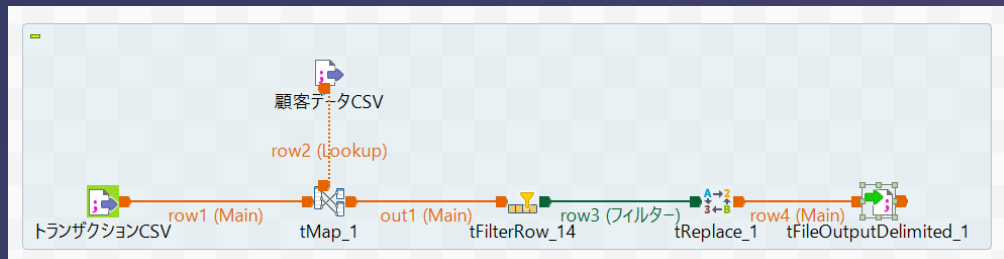
データ
プレパレーション



データ
スチュワード
シップ

デモシナリオ

CSV形式の顧客データと売上データを
マージし一部のデータを検索/置換。
東京のデータだけを抽出しファイルに
出力するジョブを開発します





おわりに



Talend Open Studioを使い始める

Talend Open Studio：オープンソース版（無償版）Talend Studio

Talend Open Studioのダウンロード

- <https://www.talend.com/jp/products/talend-open-studio/>
- <https://sourceforge.net/projects/talend-studio/>

オンラインヘルプ

- <https://help.talend.com/r/ja-JP/8.0/studio-user-guide-open-studio-for-data-integration>
- <https://help.talend.com/r/ja-JP/8.0/studio-getting-started-guide-open-studio-for-data-integration>

Talend Open Studioには何が必要ですか？

Talend Open Studioをダウンロードする前に、お使いのシステムがインストール要件を満たしていることをご確認ください：

		
メモリ	ディスク容量	オペレーティングシステム
8GBを推奨	20GB以上を推奨	Windows 10、Catalina 10.15、Mojave 10.14、 またはHigh Sierra 10.13

 Linux 版Talend Open Studio for Data Integration インストール&アップグレードガイド

3 個のバリエーションから選択 8.0

 Talend Open Studio for Data Integration ユーザーガイド 8.0

 Talend Open Studio for Data Integration 入門ガイド 8.0



健全なデータで
ビジネスを健全に

お問い合わせ先
<https://www.talend.com/jp/contact-sales/>

