

# 世界初のオープンソースETL「Talend Open Studio」

*Talend, Global Leader in Open Source Data Management*

**オープンソースカンファレンス  
2011 Tokyo/Spring**

2011年03月04日(金) 14:00-14:45  
@ 61号館 #202

Talend株式会社  
コンサルティングマネージャー  
正金 秀規

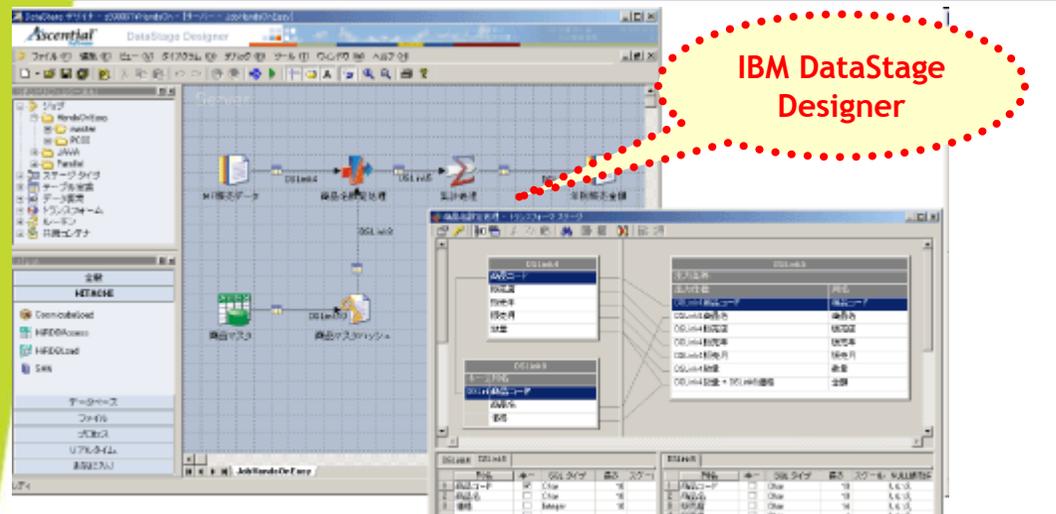
## アジェンダ

- 企業ITにおけるデータ処理基盤
- ETLに期待される役割
- ~~ETLベンダとオープンソースETLの実際~~
- Talend社概要
- Talend製品マップ
- Talend Open Studioで何が出来るか？
- 付加価値と源泉：Talend Forge

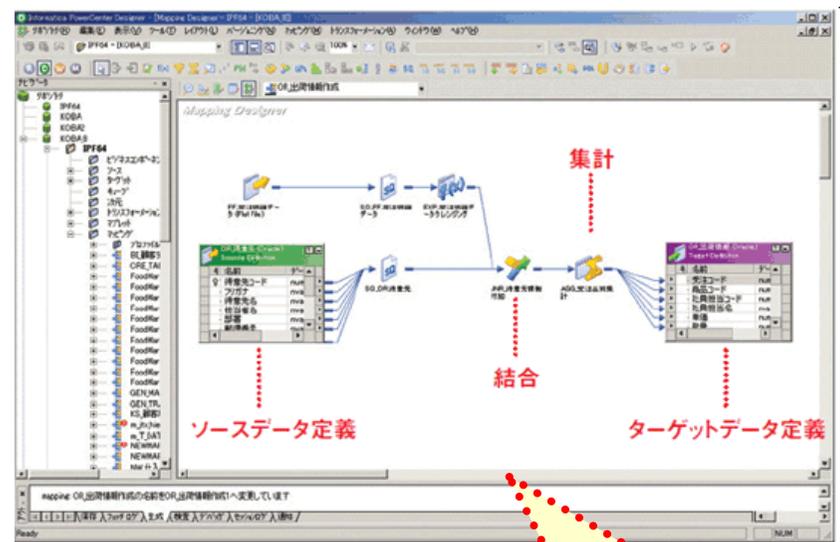




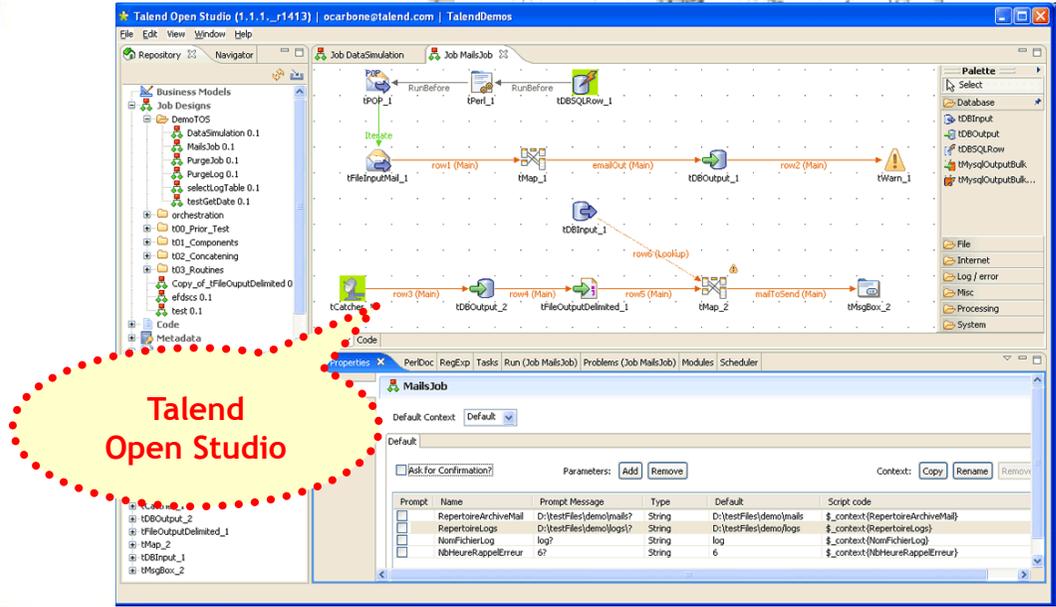
# 企業ITにおけるデータ処理基盤(続き)



IBM DataStage Designer



PowerCenter Mapping Designer



Talend Open Studio

2000年代～：  
サーバ機(Linux / 商用UNIX / Windows)  
上でのETLソフトウェアによる実装  
データの抽出・処理・ローディングという一連  
のデータフローをGUI上で組立て、そのまま  
処理として走る

## ETLに期待される役割

ETLの語源 : **E**xtract **T**ransform **L**oading の頭文字を抜粋した造語

そもそもETLは、全てのデータ処理を「抽出」「変換」「登録」の大きく三つの処理に分類したアプリケーション処理方式。DWHの父：米国ビル・インモン（William H. Inmon）氏により、統合履歴管理型DB構築に不可欠なソリューションとして定義された言葉

- Extract（抽出） : 処理対象のデータをシステムより抽出  
Transform（変換） : 抽出したデータを業務ロジックに従い変換  
Loading（登録） : 変換したデータを目的のデータベースに登録

DWH構築用途から、現在では以下のように広範囲で活用が進む！

データ移行

データ連携  
システム連携

基幹業務  
バッチ処理

名寄せ処理

リアルタイム  
連携

マスターデータ  
管理

クラウド  
オンプレミス  
連携

マッシュアップ  
基盤、、、等

# Talend社概要

- OSSを基本としたデータマネジメント製品のリーダー
- 未上場、VC支援による経営
- グローバル展開と導入ユーザ



**San Francisco (Los Altos)**

コーポレート

**Orange County (Irvine)**

営業マーケティング

R&D

技術支援

**Boston (Burlington)**

営業マーケティング

R&D

**New York (Tarrytown)**

営業マーケティング

技術支援

**Paris (Suresnes)**

コーポレート

営業マーケティング

R&D

技術支援

**Utrecht**

営業マーケティング

**Milan (Curno)**

営業マーケティング

技術支援

**Beijing**

R&D

技術支援

**Tokyo**

営業マーケティング

技術支援

**London (Maidenhead)**

営業マーケティング

技術支援

**Nuremberg**

営業マーケティング

技術支援

**Bonn**

営業マーケティング

R&D

技術支援

**Munich**

営業マーケティング

## Talend社概要：誰がTalendを産んだのか



**Bertrand Diard**  
Co-founder and CEO  
ベルトランド・ディアド  
創業者兼最高経営責任者

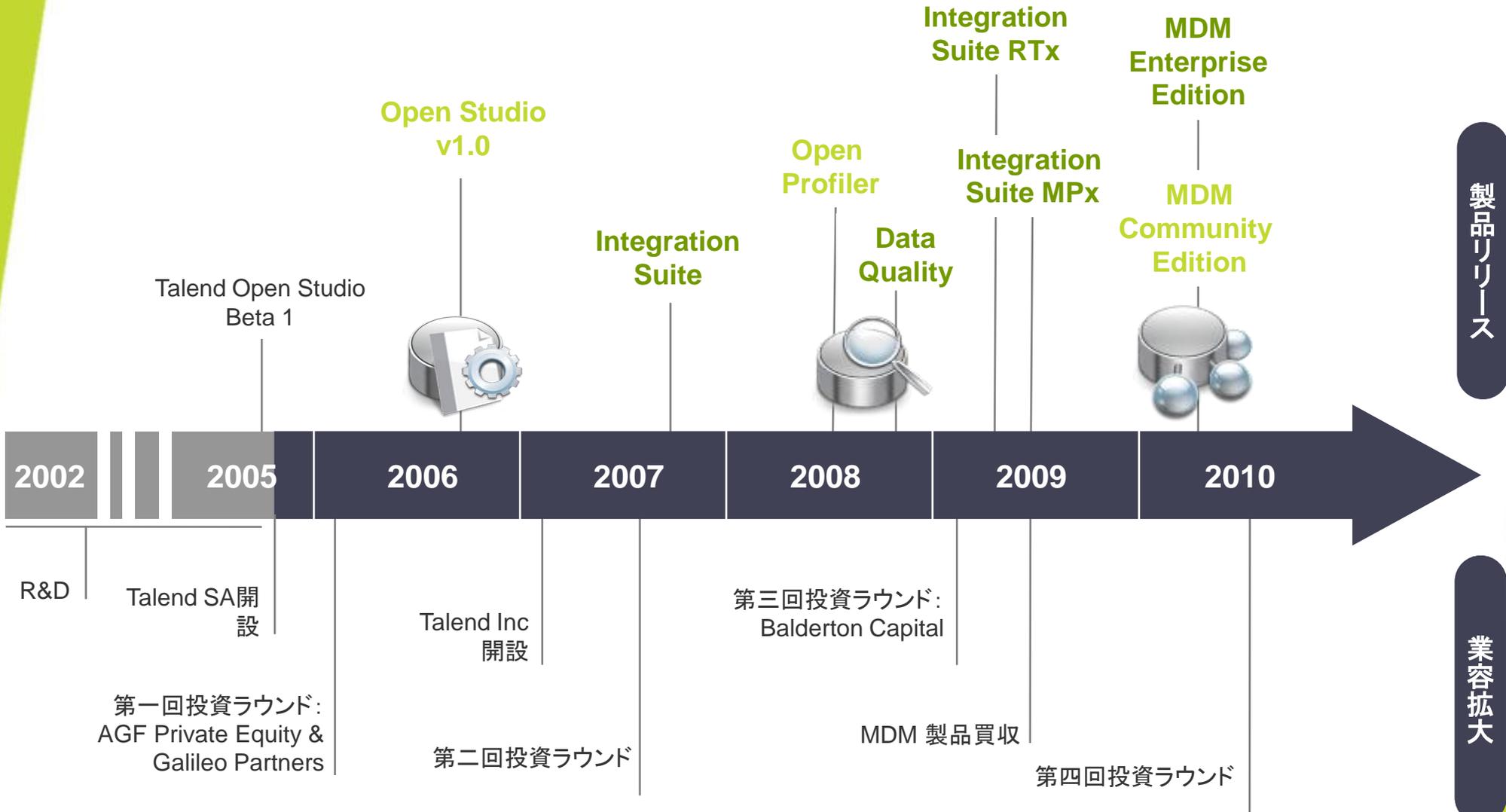


**Fabrice Bonan**  
Co-founder and COO  
ファブリス・ボナン  
創業者兼再考執行責任者



**Cédric Carbone**  
CTO  
セドリック・カルボン  
最高技術責任者

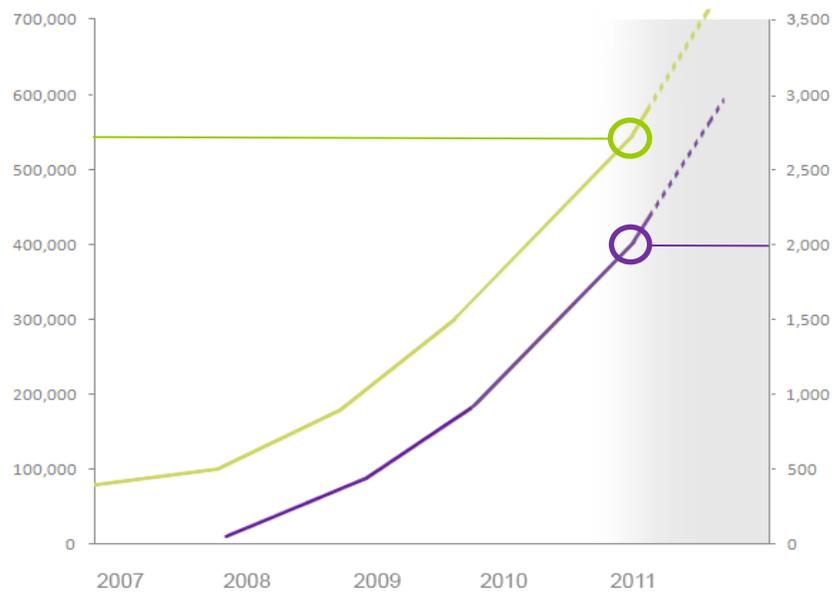
# Talend社概要:沿革



## Talend社概要:ハイライト

### 高い市場認知度！

- 1,200万ダウンロード
- 55万以上のユーザ
- 2,000社の有償版顧客



1 ダウンロード / 分  
Talend Open Studio



100 新規ユーザ / 月

## Talend社概要: 導入顧客例

### 金融・保険業



### サービス業



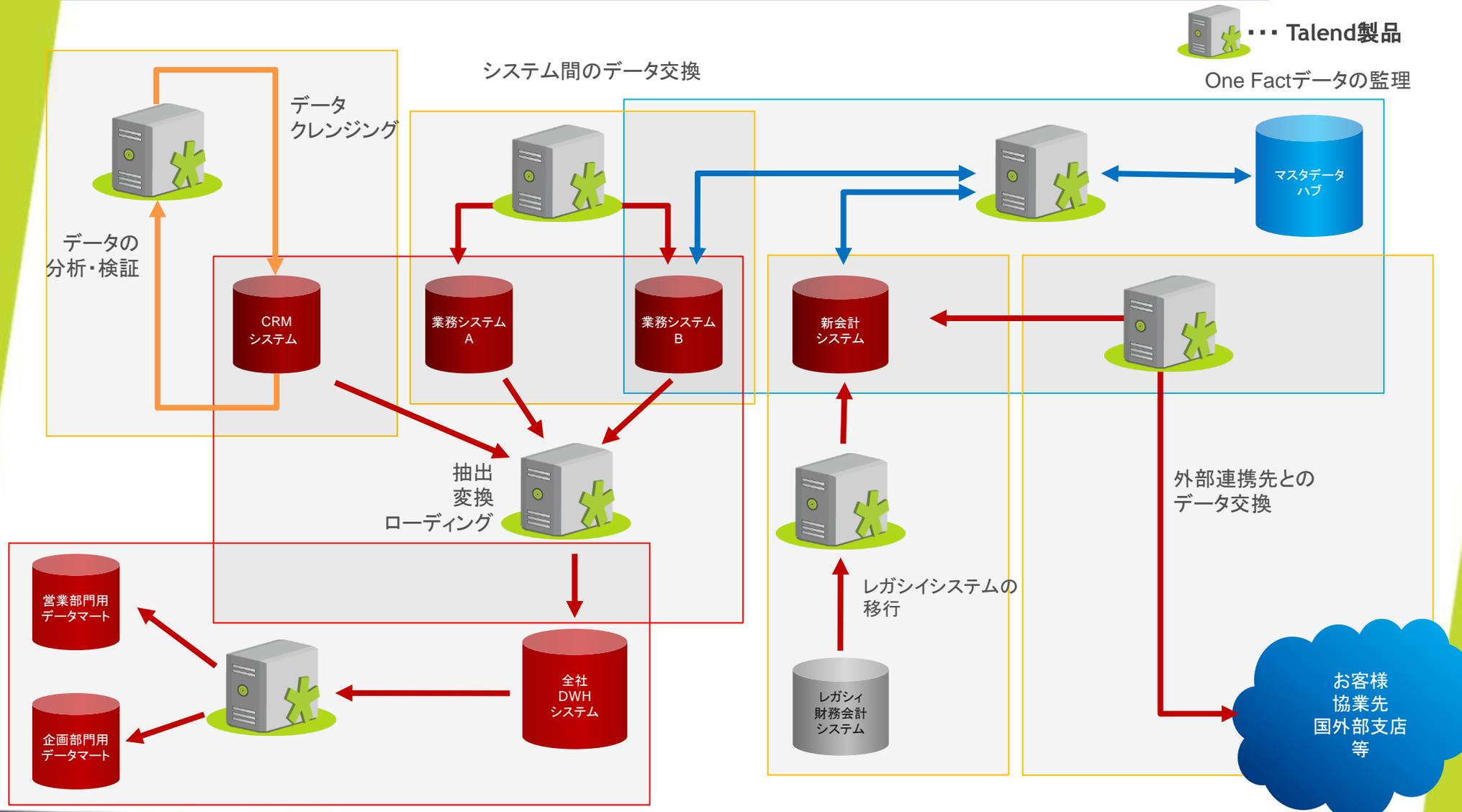
### 製造・小売・流通サービス



### 公共団体 教育機関



# Talend製品マップ: 企業ITにおける位置づけ



... Talend製品  
One Factデータの監理

お客様  
協業先  
国外部支店  
等

# Talend製品マップ: GPL製品 & 商用ライセンス製品

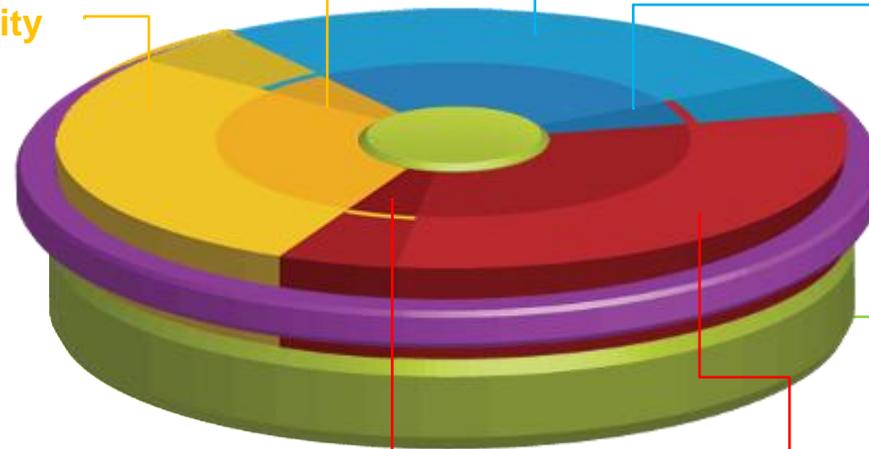
## Talend Open Profiler

- データの内容・品質検証
- GPLv2製品、無制限に使用可能
- 品質指標の作成が可能

## Data Quality

### Talend Data Quality

- クレンジングと検知
- クレンジング用  
コンポーネント
- 検証レポート機能
- データ品質に関する  
ポータル機能



## Talend Open Studio

- データフローの構築
- GPLv2製品、無制限に使用可能
- 多機能・高速データプロセスエンジン
- 450+ のコンポーネントが利用可能

## Data Integration

## Talend Integration Suite

- ミッションクリティカルなデータ運用を実現
- チーム開発機能
- 自動デプロイ、ロードバランシング、HA
- ジョブフロー制御機能
- 運用監理機能

## Talend MDM Enterprise Edition

- 全社を俯瞰したマスタデータ管理
- 権限管理・制御
- 妥当性ルールの変換
- 高度なワークフローエンジン

## MDM

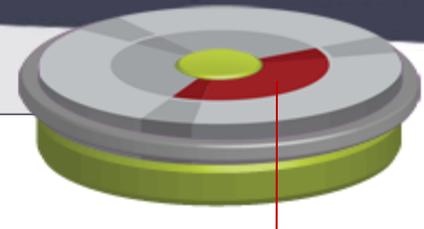
## Talend MDM Community Edition

- コミュニティベースのマスタデータ管理
- GPLv2製品、無制限に使用可能
- XMLベースのアクティブデータモデル
- 業務ユーザー向け軽量GUI

## Talend Unified Platform

- 標準テクノロジーを製品基盤に採用
- GUI : Eclipse, ブラウザ
- リポジトリ: Subversion, RDBMS

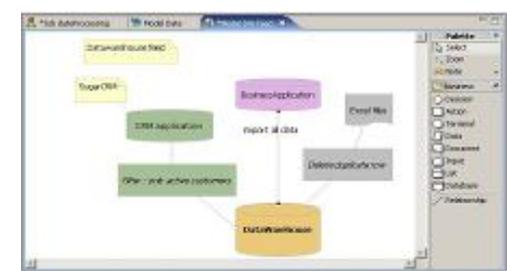
# Talend Open Studioで何が出来るか？



**Data Integration**

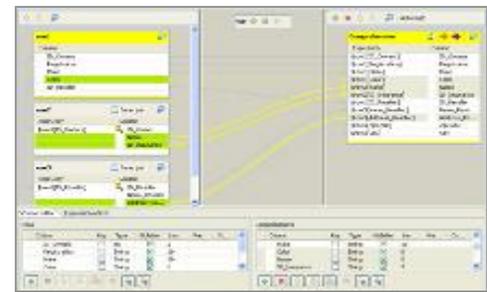
## 世界初のオープンソースによるデータ統合製品：

- ビジネスモデラ  
⇒ *ビジネスフロー作成機能*
- ジョブデザイナー  
⇒ *データ処理とジョブフローをGUIベースで定義*
- メタデータマネージャ  
⇒ *スキーマ定義を自動収集*



## 主要機能：

- ビジネスフローモデリング機能
- 堅牢で拡張性に富んだ処理構築が可能
- 広範にわたるシステム接続をサポート:450+コンポーネント
- 設計⇔設定⇔実行⇔デバッグの開発製造工程をシームレスに支援するリアルタイムデバッグ機能
- 設計・設定内容を自動文書化



**GNU GPL, LGPL**

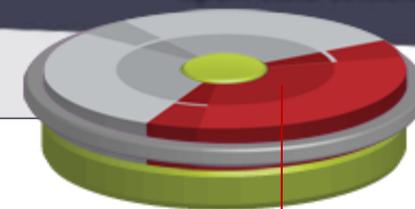
# 提供コンポーネント



## コンポーネント

- 450+
- 60%は、Talendコミュニティにより設計開発
- 全て無償で使用可能
- Q&Aとサポートは、Talendで担当

## サブスクリプション製品: Talend Integration Suite

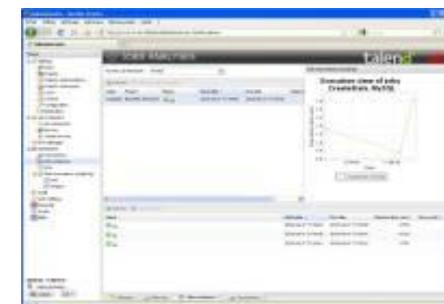


Data Integration

Talend Open Studioをベースにチーム開発機能、運用・監理機能を追加し、ミッションクリティカルにも対応したデータ処理基盤！  
e-mail / Support Portalを通じたSLAを含むサポートサービスを提供！

### 追加される主要機能:

- 各種ウィザード機能、データプレビュー機能
- 共有リポジトリ機能
- Joblet機能による処理の共通化
- 自動配布機能
- CDC: チェンジデータキャプチャ機能
- コマンドライン/I/Fの提供
- Jobコンダクタによる  
タイム & イベントベーススケジューラ機能
- 仮想サーバ化による  
フェイルオーバー、ロードバランシング機能
- 運用監視ダッシュボードによる統合監理機能



### Talend Integration Suiteの価値:

- 開發生産性のさらなる向上
- 開発製造作業の共有・共通化と最適化された  
コンポーネント配布が可能
- ミッションクリティカル基盤を提供
- 統合運用監理基盤を提供

**Talend商用ライセンス**

## Talend Integration Suite: Edition別機能表

Edition	設計 / 文書化	製造 / 実装	検証	ジョブ配布	実行 / 運用管理
MPx					Hadoop   FileScale
Enterprise				高可用性	ロードバランシング   フェイルオーバー
RTx				SOA マネージャ	
Professional		ビジネス ルール   API	ディスタントラン	実行計画   イベントスケジューラ	エラーリカバリ   ダッシュボード
Team	Auto Doc	リファレンスプロジェクト   Jobデザイナ +	CDC   コマンドライン	タイムスケジューラ   Jobコンダクタ	AMC (アクティビティ モニタリングコンソール)
	共有リポジトリ / SVN				
Talend Open Studio	ビジネス モデラ	Job デザイナ	コンポーネント	コンテキスト	バージョン管理

Talend フォミニクストワーシジョンセンター (TAC)

## 特徴と差別化要因: 技術的観点

### Javaのコードジェネレータである

- 環境/プラットフォームの制限が少ない実行ファイル形式 (Javaアプリケーション)
- インタプリタコードによる実行時のオーバヘッドを削減しており、実行時のCPU資源消費が少ない
- H/W遊休資産の流用が可能である
- Gridコントロールにより、要求に応じて必要な環境にジョブを配布し実行！

### 標準的かつオープンなテクノロジーに立脚

(Eclipse, Java, SQL, XML, Apache Tomcat, Subversion, etc)

- 既存の技術スキルが流用可能
- 習得・習熟に要するコストが少ない

### 多機能であり非常に柔軟な拡張性

- 既存のJavaルーチンを埋め込むことが可能
- 「自分で」コンポーネントの作成が可能である
- JMS/MOM連携、Loop処理、リアルタイム連携、LDAP連携等について標準機能で実現可能
- Low CostでSalesforce.comとの連携の仕組みを実装可能
- 勿論、コミュニティで製造されたコンポーネントも使用可能

### 統合化されたコンポーネント管理、運用監視機能

- メタデータ、ジョブ、ドキュメント等、プロジェクトに必要な全ての成果物を集中管理しバージョンコントロール可能
- 運用時の統合化された管理・監視環境を提供

## 処理性能に関する製品比較

ETL benchmarks v1.1 (2009年02月時点)より考察

[http://www.manapps.tm.fr/pdfETL/ETLBenchmarks\\_Manapps%20090203.pdf](http://www.manapps.tm.fr/pdfETL/ETLBenchmarks_Manapps%20090203.pdf)



- 仏ManApps社が、IBM DataStage Server & PX, Informatica PowerCenter, Talend Open Studio, Pentahoの5製品を対象に実施。Creative Commonsのライセンスにて公開済み
- 11のテストシナリオに対して10万件、100万件、500万件、2,000万件とスケールアップして検証
- 検証環境：
  - OS : Microsoft Windows XP Professional Edition SP2
  - CPU : Intel Core2 Duo 2.0GHz
  - Memory : 4GB
  - JVM : JVM1.6.0\_87

### 【結果】

#	製品	スコア
1	PowerCenter 8.1.1	353 points
<b>2</b>	<b>Talend Open Studio 2.4.1</b>	<b>333 points</b>
3	DataStage PX 7.5	239 points
4	DataStage Server 7.5	199 points
5	Pentaho Data Integration 3.0.0	148 points

※1位を5point、2位を4point、...としてスコアを集計

### 前提:

1. Talendは、NonチューニングでありMPxも使用していない
2. Informatica社は自社のコンサルタントがチューニングを実施
3. DataStage, PowerCenterは、並列処理を使用

### 考察:

1. 100万件までは、概ねTalendが最も良いスコアである
  2. 集計処理は、PowerCenterが最も良いスコアである
  3. ELT機能は、TalendとPowerCenterが双璧である
  4. 処理内容が複雑になる(ルックアップしマッチしないデータをリジェクトするなど現実的なロジック)とデータ量に依らずTalendが最も良いスコアである
- ※もう少し長いロウサイズでサーバ機で実行する必要があり、各々、現実的な範囲でチューニングした結果の比較が必要と思われる

# SaaS対応: Salesforce.com

## Talend Open Studio

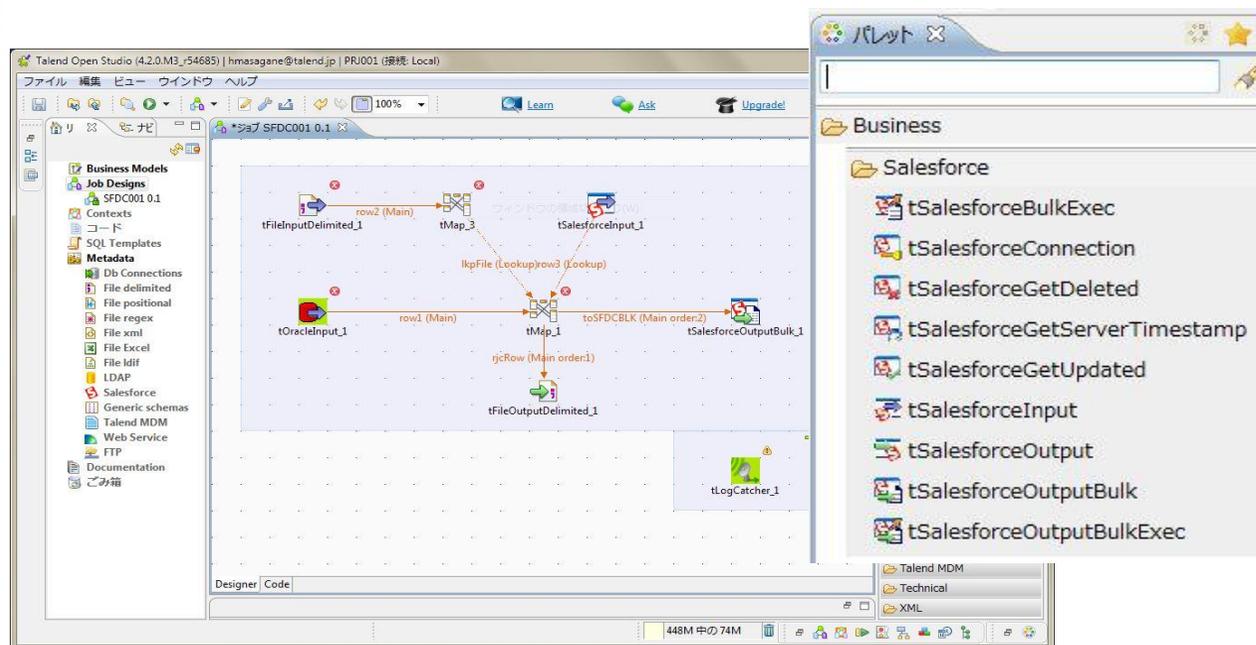
### Talend Integration Suiteの役割例:

クラウド間連携、クラウド⇔オンプレミス間連携例

- 個別部門システムより業務要件等に従いソースデータデータを抽出統合、重複削除など実施して素ファイルを作成
- 素ファイルをSalesforce.comにオブジェクトごとに適宜挿入・更新・削除を実施。またはBulk APIを使用したコンポーネントで一括反映
- 社内システムに必要となるデータをSalesforce.comより条件指定して抽出
- Salesforce.comデータを動的参照して社内システム用データを作成
- Salesforce.comで発生するソースデータを抽出して社内でバックアップを構築する等

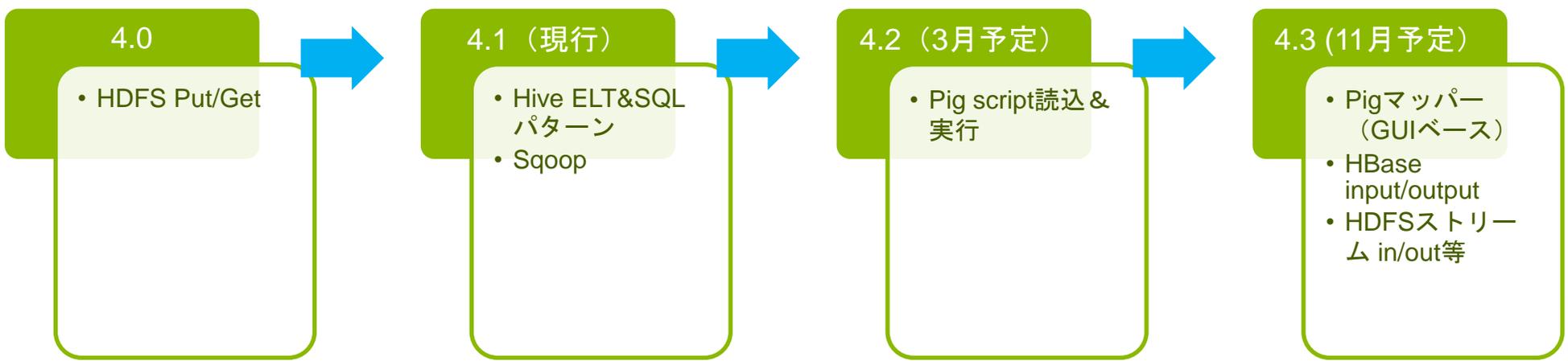
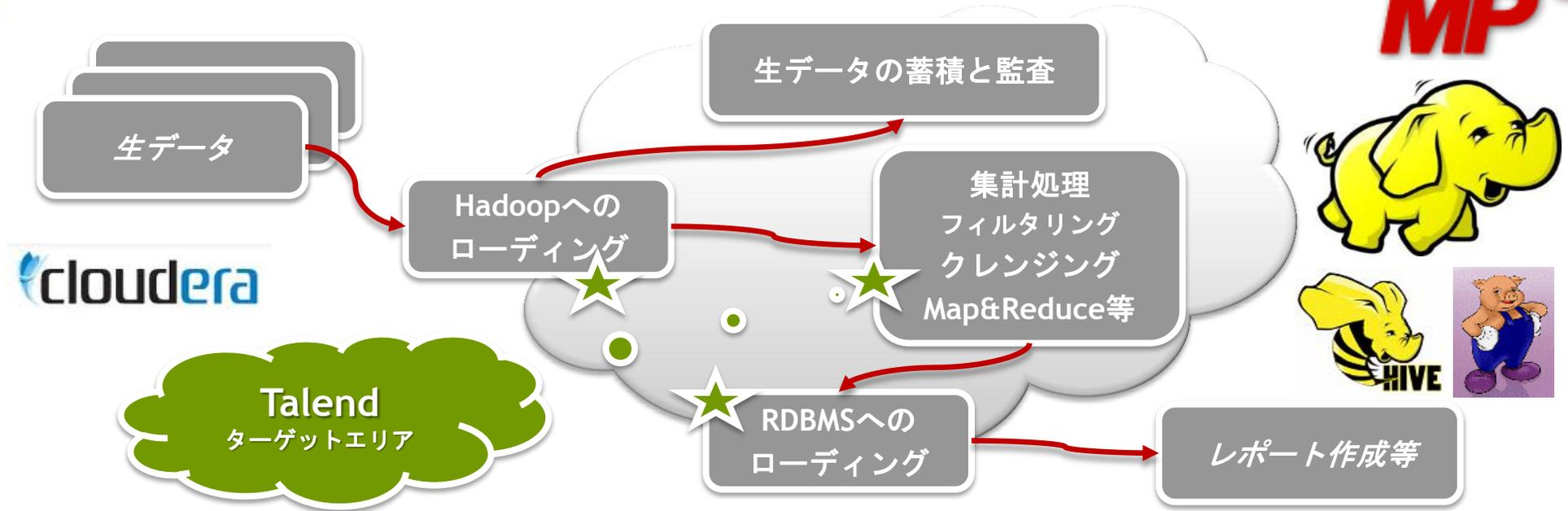
### Salesforce.com用コンポーネント

- **tSalesforceConnection**  
Salesforce.comへのSOAP接続を確立
- **tSalesforceGetServerTimestamp**  
Salesforce.comのサーバ時間を取得
- **tSalesforceGetUpdated**  
Salesforce.com内の論理更新される以前のデータを日時分秒範囲指定で取得
- **tSalesforceGetDeleted**  
Salesforce.com内の論理削除される以前のデータを日時分秒範囲指定で取得
- **tSalesforceInput**  
Salesforce.comのオブジェクト単位で抽出条件を付けてデータを抽出
- **tSalesforceOutput**  
Salesforce.comのオブジェクトに対してデータを挿入/更新/削除/UPSERTを実施
- **tSalesforceBulkExec**  
Salesforce.comのオブジェクトに対してバルクでファイルデータを挿入/更新/UPSERTを実施
- **tSalesforceOutputBulk**  
Salesforce.comのオブジェクトへ反映するファイルの準備を行う
- **tSalesforceOutputBulkExec**  
tSalesforceOutputBulk, tSalesforceBulkExecの動作要素を一つで実施



# Hadoop対応 / Cloudera社提携

**MP<sup>x</sup>**



## 付加価値と源泉: Talend Forge

- フォーラム数: 11
- 投稿総数: 40,000+ (80+ポスト/日)
- 登録ユーザ: 6,000+
- βテスター: 1,000+
- Talend Exchangeコンポーネント: 330+
- Talend Babili(国際化): 単体62,000+

<http://www.talendforge.org>

**Thank you**

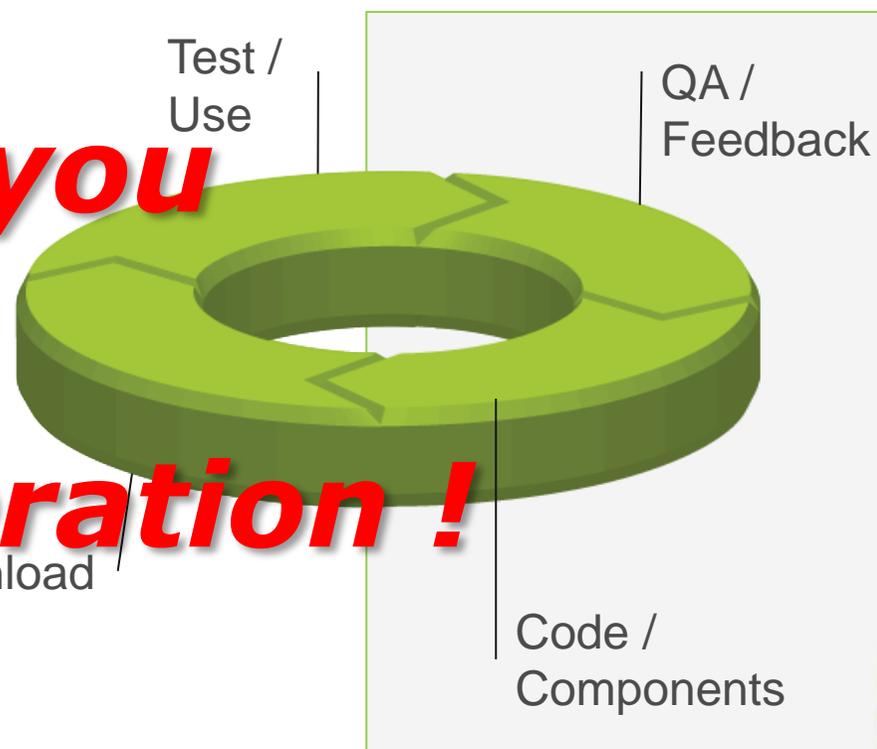
**for**

**your cooperation !**



### コミュニティベースのプロジェクト: 例

- Excel Report add-ins
- コネクタ: BIRT, Google Apps., etc.
- DataStageからの移行ツール: ETL Converter  
⇒ SourceForgeよりダウンロード可能!



Please visit !

*<http://jp.talend.com/index.php>*  
*<http://www.talendforge.org>*