# PythonとSparkで学ぶPySpark 速習講座

# データエンジニアのための
# 最強のビッグデータ処理エンジンPyspark

~ABC人材のBig Dataを処理しよう バッチ処理~

# はじめに

# 本コースの概要

- Pysparkのバッチ講座になります
  - Pysparkはストリーミング処理も可能ですがそれは別講座にて
- 学べること
  - ケーススタディで実務を例に取ったデータエンジニアリングの流れで紹介
  - PySparkを使う上でハマりやすいチューニングポイントを知ることができます
  - 分散処理の基本を学ぶことができます
- ソースコードはすべてgithubに公開しています
  - https://github.com/yk-st/pyspark_batch

# 本コースの特徴

- 日本で最初のPysparkコースです(おそらく、もしくは少なめ)
- 実務経験から特に重要なポイントに絞り解説を行います
  - よくある関数の羅列ではなく、ストーリじたてで紹介します
  - そのためあまり遠回りはしません
- Pysparkはバッチ処理もストリーミング処理もできますが
  - 本コースは**バッチ処理のコース**です

# 本コースを学ぶ意義

- SparkはABC人材(AI,BigData,Cloud)な人材になるための必須スキルと言っても過言ではありません
  - ABCはもう止められない流れ

- Sparkがスキルセットに存在しているだけで、企業のデータ活用の人材として重宝されます
  - 年収も高めです

# 本コースに適する人

- これからビッグデータの世界で大規模なデータと闘うABC人材になりたい人
  - AI BigData Cloudの頭文字をとった人材のこと
- Pythonを使ったプログラミングを強化したい人
  - Pythonに分散処理というスパイスを加えたい人

# 本コースに適さない人

- Pysparkの熟達者
- Pysparkでストリーミング処理をやってみたい方
  - 別のコースで作成予定
- 機械学習のアルゴリズムを勉強したい方
  - 難しいアルゴリズムは出てきません

# 自己紹介

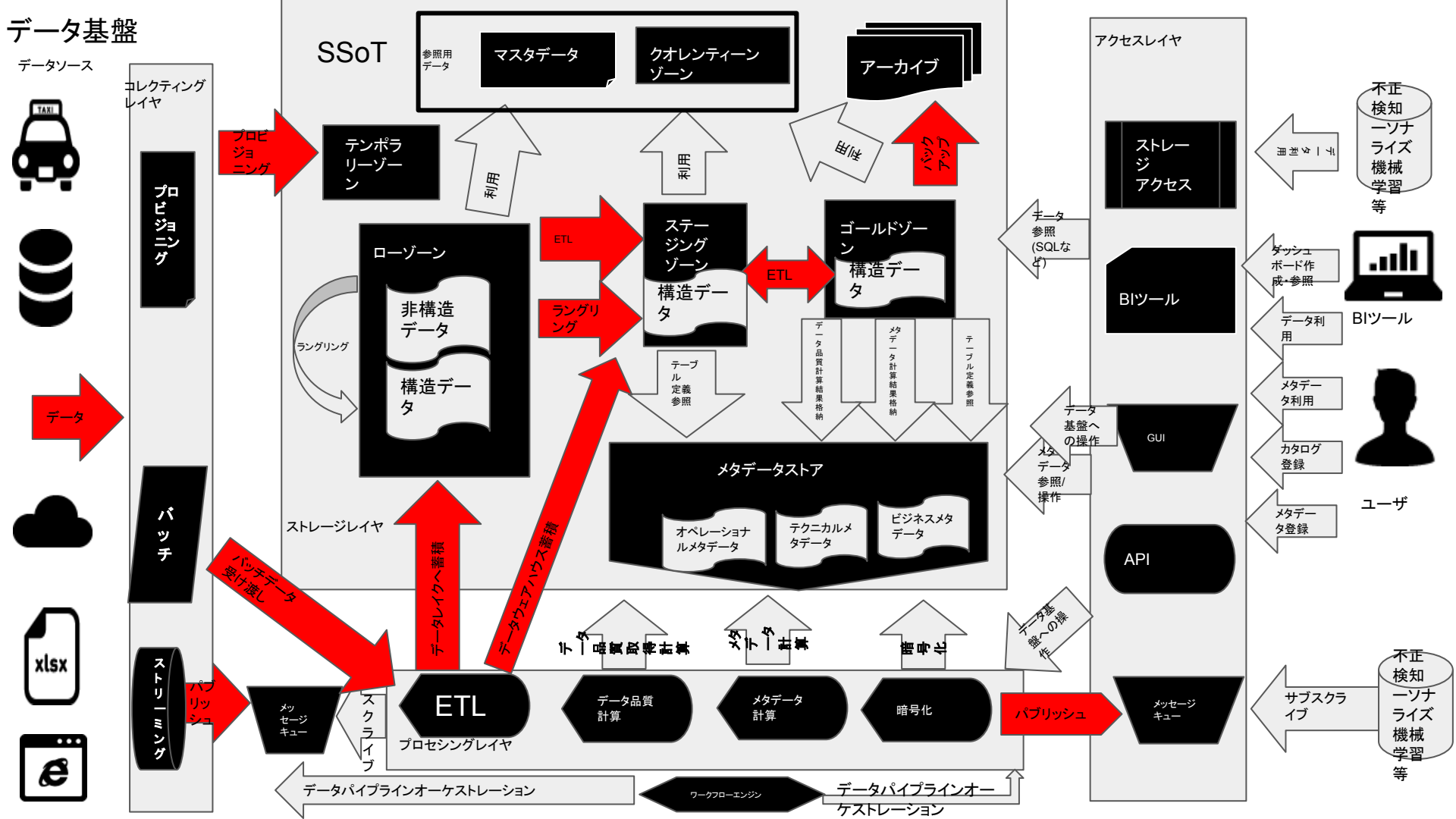


- データエンジニア
- 数PBクラスのデータレイクやデータウェアハウスアーキテクトを担当
- データ処理(データラングリング、データ品質の可視化処理などなど)
- 2021/12末くらいに拙書のデータエンジニアリング書籍が出ます

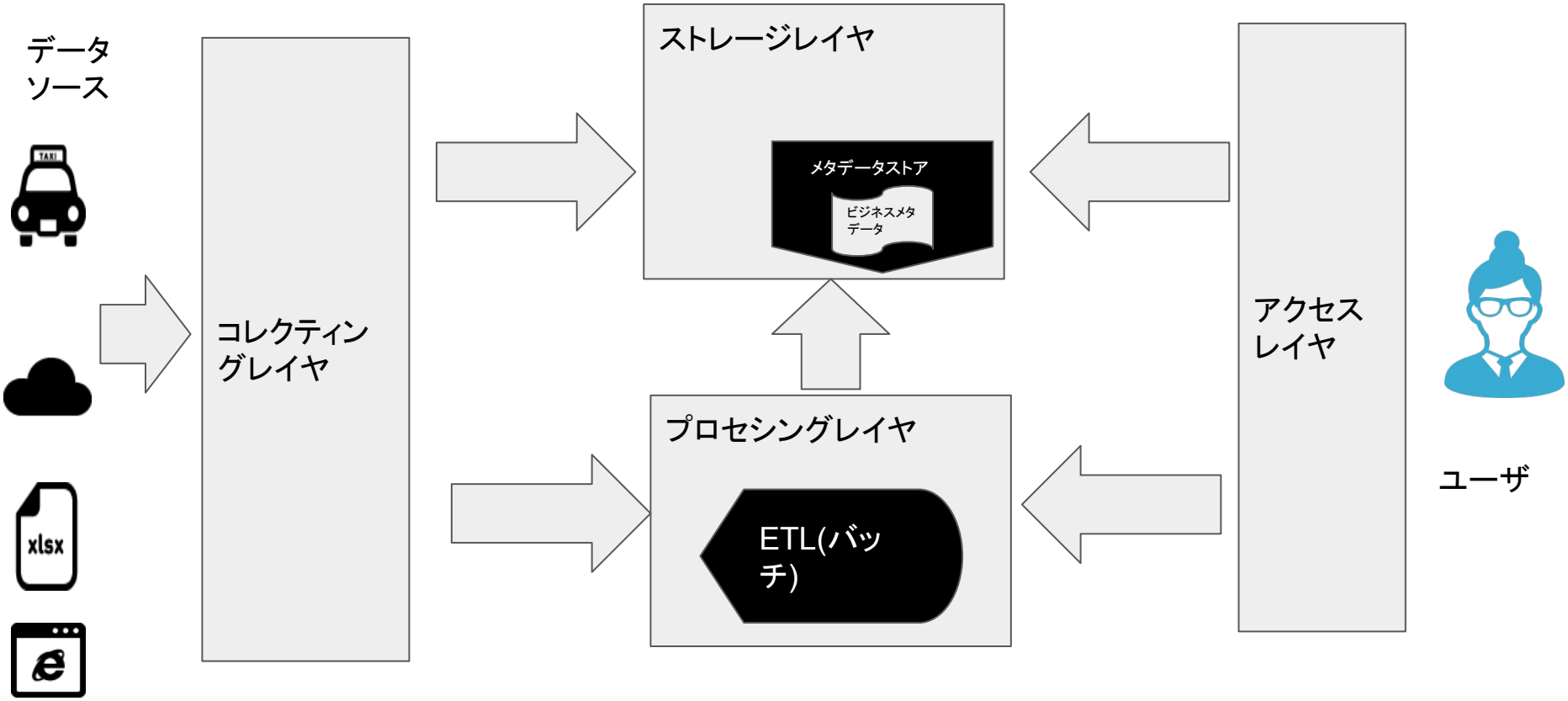
# 本コースの役割

ビッグデータ基盤
においてどこに対する
データエンジニアリングなのか？

# データ基盤


データソース
コレクティングレイヤ
プロビジョニング
データ
バッチ
xlsx
ストリーミング
SSoT
参照用データ
マスタデータ
クオレンティーンゾーン
アーカイブ
テンポラリーゾーン
ローゾーン
非構造データ
構造データ
ラングリング
ETL
ステージングゾーン
構造データ
ゴールドゾーン
構造データ
テーブル定義参照
データ品質計算結果格納
メタデータ計算結果格納
テーブル定義参照
メタデータストア
オペレーショナルメタデータ
テクニカルメタデータ
ビジネスメタデータ
ストレージレイヤ
データレイクへ蓄積
データウェアハウス蓄積
バッチデータ受け渡し
パブリッシュ
メッセージキュー
スクライブ
ETL
プロセシングレイヤ
データ品質計算
メタデータ計算
暗号化
パブリッシュ
メッセージキュー
データパイプラインオーケストレーション
ワークフローエンジン
データパイプラインオーケストレーション
アクセスレイヤ
ストレージアクセス
BIツール
API
GUI
データ参照(SQLなど)
メタデータ参照/操作
データ基盤へのデータ登録
データ基盤への操作
不正検知
ーソナライズ
機械学習等
ダッシュボード作成・参照
BIツール
データ利用
メタデータ利用
カタログ登録
メタデータ登録
ユーザ
サブスクライブ
不正検知
ーソナライズ
機械学習等

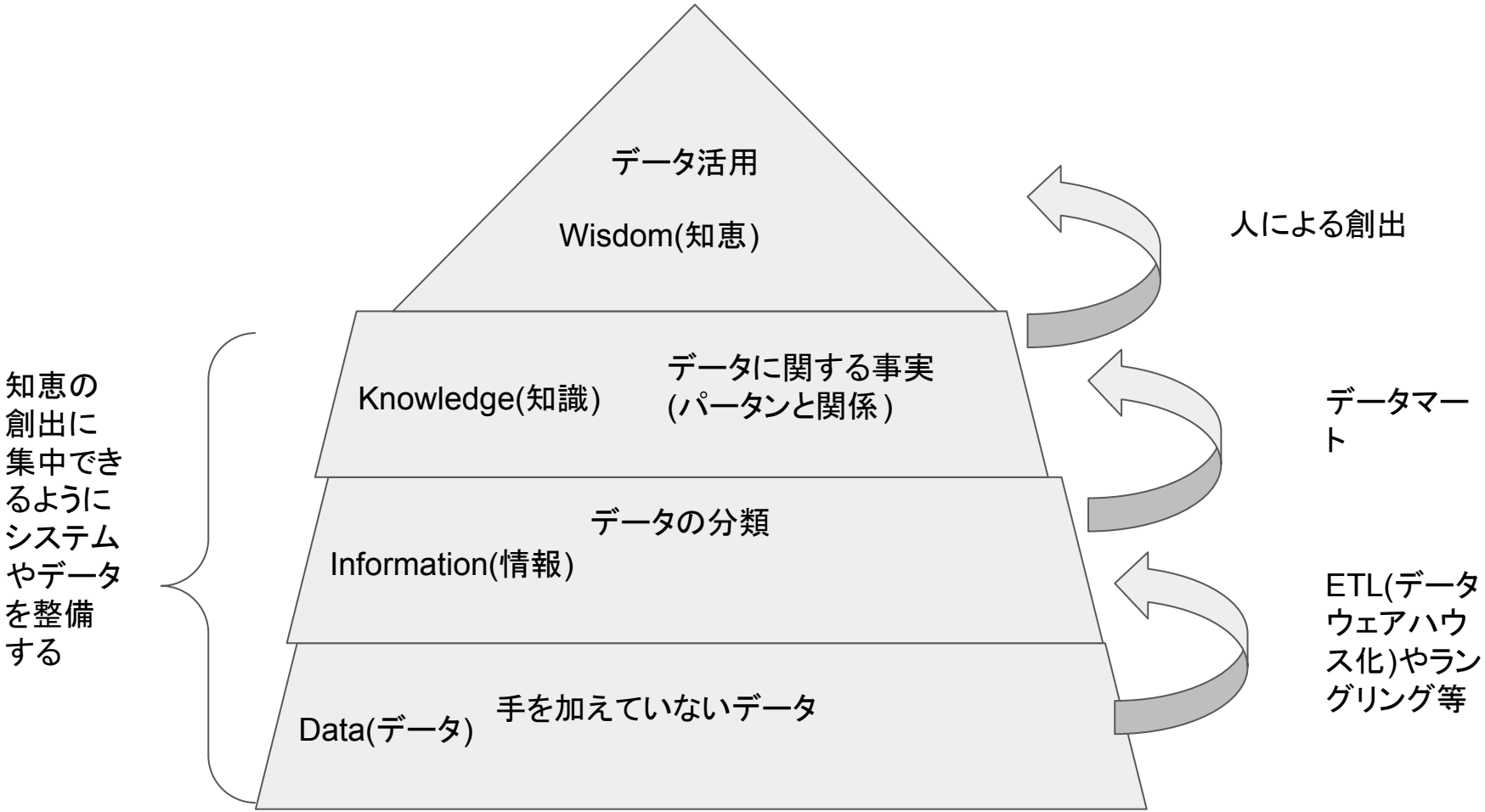
# 今回のコースはどこに当たる？



# Sparkの紹介とインストール

# 目次

# Pyspark Basics


データ活用
Wisdom(知恵)
人による創出
Knowledge(知識)
データに関する事実
(パータンと関係)
データマート
データの分類
Information(情報)
ETL(データウェアハウス化)やラングリング等
手を加えていないデータ
Data(データ)
知恵の創出に集中できるようにシステムやデータを整備する

# 目次

# Pyspark for SQL

# 目次

# Pyspark in Production

# 目次