

# 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~

# PythonとSparkで学ぶ データエンジニアリング (PySpark) 速習講座

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](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、ストーリーじたてで紹介します
  - そのためあまり遠回りはありません
- 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

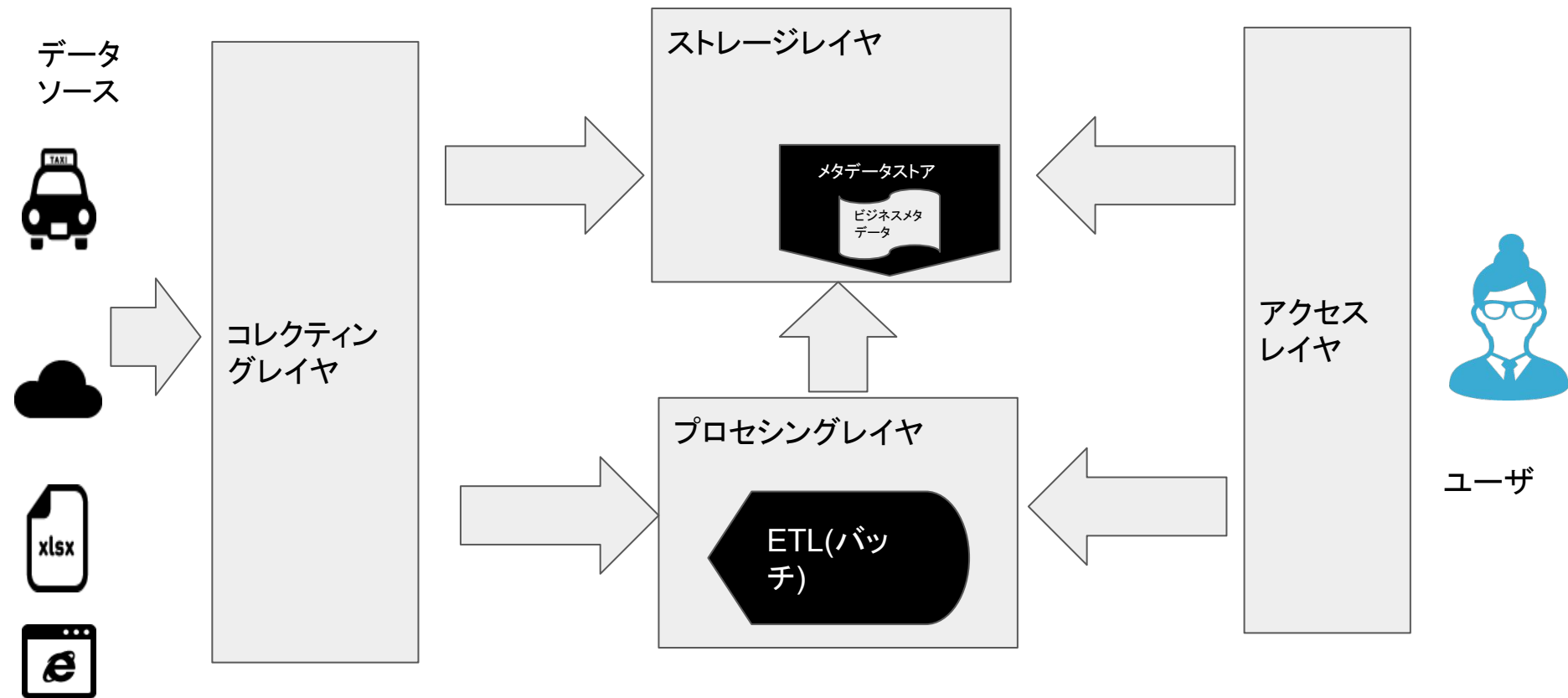
## ビッグデータ基盤

## においてどこに対する

## データエンジニアリングなのか？



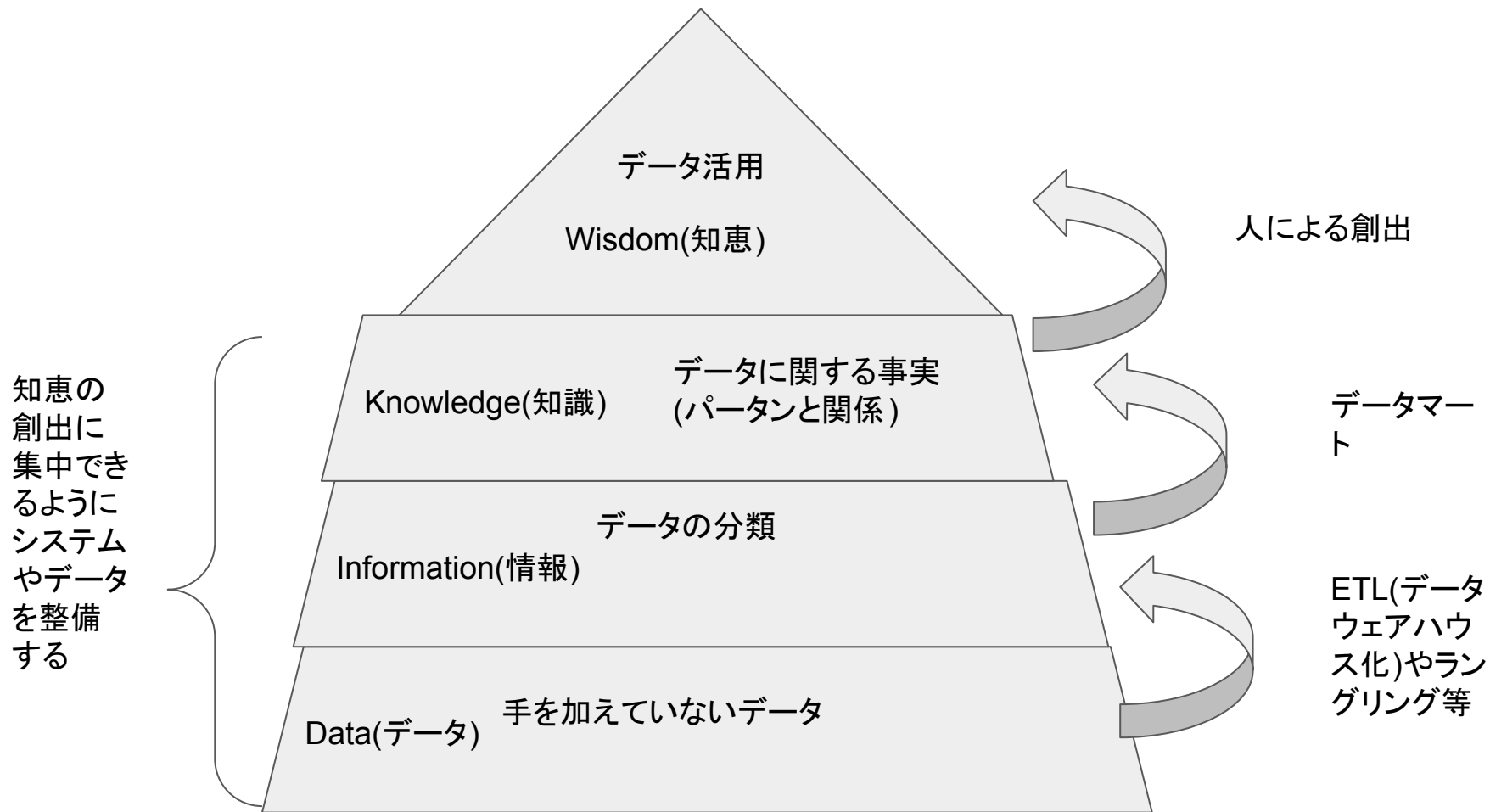
# 今回のコースはどこに当たる？



# Sparkの紹介とインストール

# 目次

# Pyspark Basics





# 目次

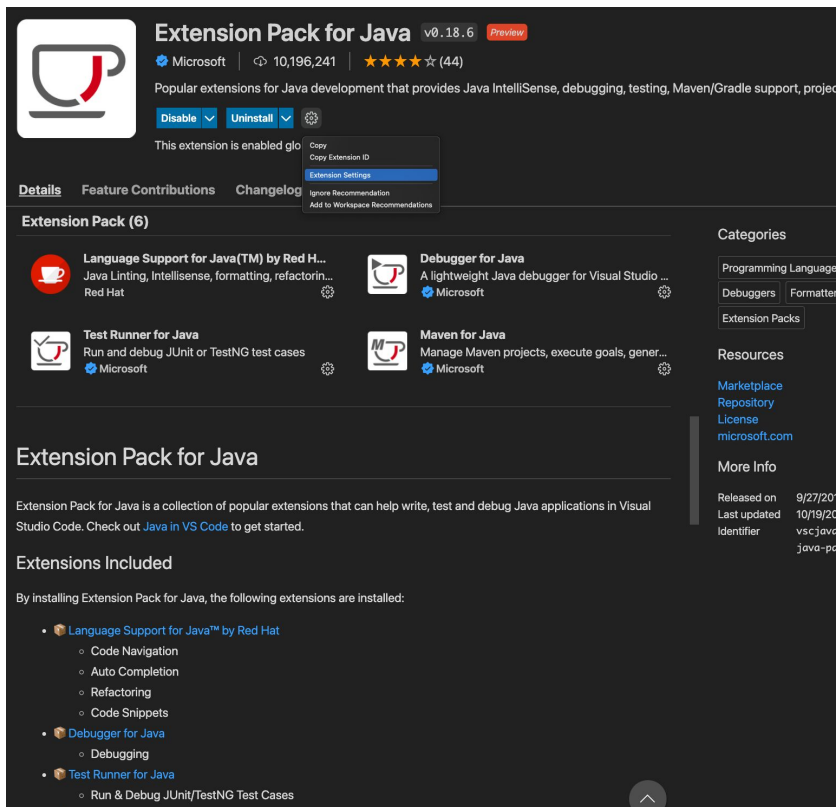
# Pyspark for SQL

# 目次

# Pyspark in Production

# 目次

# Javaのインストール



**Extension Pack for Java** v0.18.6 Preview

Microsoft | 10,196,241 | ★★★★★ (44)

Popular extensions for Java development that provides Java IntelliSense, debugging, testing, Maven/Gradle support, project management, and more.

[Disable](#) [Uninstall](#) [Extension Settings](#)

This extension is enabled globally. [Copy Extension ID](#) [Copy Extension ID](#) [Extension Settings](#) [Ignore Recommendation](#) [Add to Workspace Recommendations](#)

**Details** Feature Contributions Changelog

**Extension Pack (6)**

- Language Support for Java(TM) by Red Hat**  
Java Linting, IntelliSense, formatting, refactorin...  
Red Hat
- Debugger for Java**  
A lightweight Java debugger for Visual Studio...  
Microsoft
- Test Runner for Java**  
Run and debug JUnit or TestNG test cases  
Microsoft
- Maven for Java**  
Manage Maven projects, execute goals, gener...  
Microsoft

**Categories**

- Programming Language
- Debuggers
- Formatters
- Extension Packs

**Resources**

- [Marketplace](#)
- [Repository](#)
- [License](#)
- [microsoft.com](#)

**More Info**

- Released on 9/27/2019
- Last updated 10/19/2020
- Identifier vscjv
- java-pd

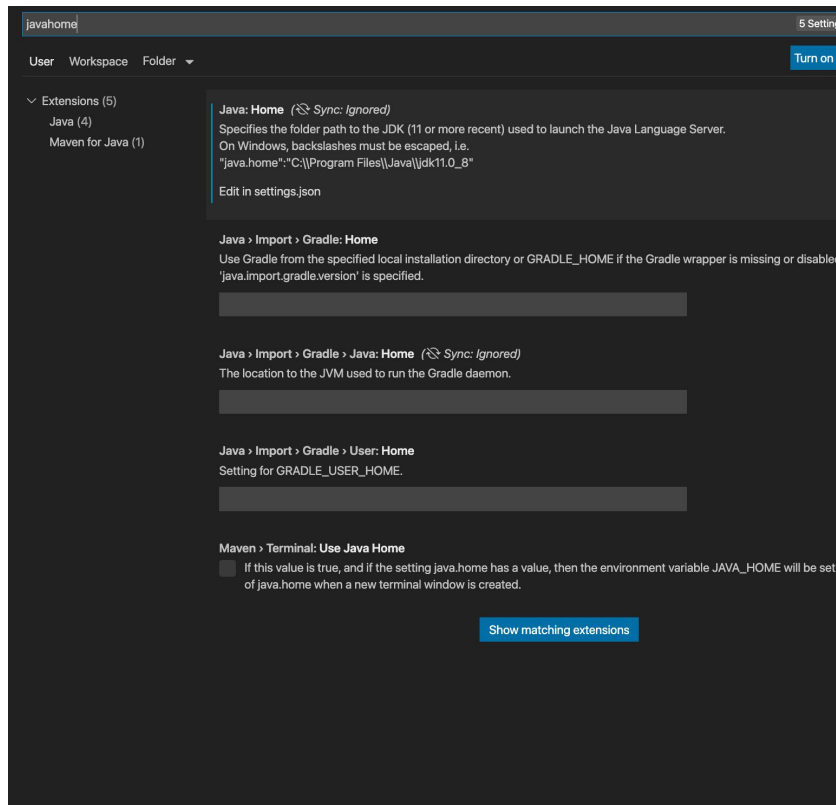
**Extension Pack for Java**

Extension Pack for Java is a collection of popular extensions that can help write, test and debug Java applications in Visual Studio Code. Check out [Java in VS Code](#) to get started.

**Extensions Included**

By installing Extension Pack for Java, the following extensions are installed:

- **Language Support for Java™ by Red Hat**
  - Code Navigation
  - Auto Completion
  - Refactoring
  - Code Snippets
- **Debugger for Java**
  - Debugging
- **Test Runner for Java**
  - Run & Debug JUnit/TestNG Test Cases



javahome 5 Settings

User Workspace Folder [Turn on](#)

Extensions (5)  
Java (4)  
Maven for Java (1)

**Java: Home** (Sync: Ignored)  
Specifies the folder path to the JDK (11 or more recent) used to launch the Java Language Server.  
On Windows, backslashes must be escaped, i.e.  
"java.home": "C:\\Program Files\\Java\\jdk11.0\_8"  
[Edit in settings.json](#)

**Java > Import > Gradle: Home**  
Use Gradle from the specified local installation directory or GRADLE\_HOME if the Gradle wrapper is missing or disabled.  
'java.import.gradle.version' is specified.

**Java > Import > Gradle > Java: Home** (Sync: Ignored)  
The location to the JVM used to run the Gradle daemon.

**Java > Import > Gradle > User: Home**  
Setting for GRADLE\_USER\_HOME.

**Maven > Terminal: Use Java Home**  
☐ If this value is true, and if the setting java.home has a value, then the environment variable JAVA\_HOME will be set of java.home when a new terminal window is created.

[Show matching extensions](#)

# Java Homeを追加

Users > saitouyuuki > Library > Application Support > Code > User > {} settings.json > [abc] java.home

```
1 {  
2   "workbench.colorTheme": "Default Dark+",  
3   "window.zoomLevel": 3,  
4   "editor.suggestSelection": "first",  
5   "vsintellicode.modify.editor.suggestSelection": "automaticallyOverrodeDefaultValue",  
6   "java.home": "/opt/homebrew/opt/openjdk@11/bin"  
7 }
```

No suggestions.