# 第7回 東京大学グローバル消費者インテリジェンス寄付講座

# SQLとデータベース

M L E · 岩澤研究室
MATSUO-IWASAWA LAB UTOKYO



### 本日のコンテンツとゴール

- 01
- SQL&データベースを なぜ学ぶのか?

02 手を動かしてSQLを 実行してみよう

03

Next Steps(中級編)

- ・本講義の位置付けと身近な例、データサイエンスとの関係性
- ・データはどう生成されるのか、データをどう管理するか?
- ・データベースとは何か、データのデザイン、SQLとは何か?

- ・SQLを使ったデータベースの構築
- ・テーブル作成、データ挿入
- ・データの抽出、削除、集計

- データベースの中級編に向けて学ぶべきこと
- ・データベース関係の資格試験、参考図書
- ・LLM(ベクトルDB等)との関係等



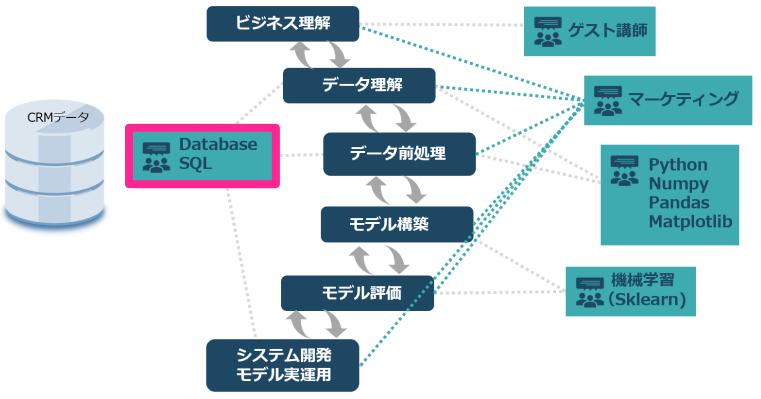


# 01 なぜSQLとデータベースを 学ぶ必要があるのか?

### 再掲

### 実データサイエンスプロジェクトと本講義の関係性

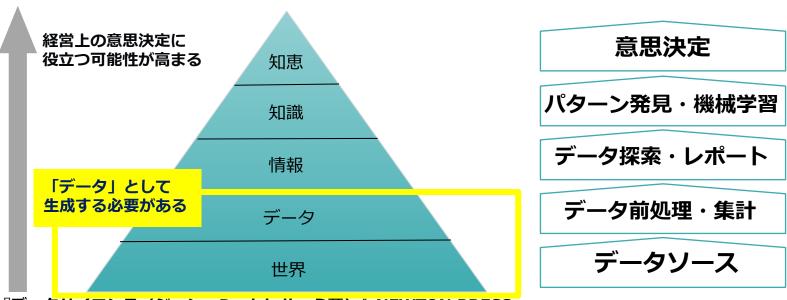
### ビジネス理解からデータ確認と前処理、モデル構築と評価、実運用まで





# 再掲データ分析を意思決定に

★単にデータを蓄積するだけではなく、知恵にまで昇華し意思決定に役立てることが重要



\*参考『データサイエンス(ジョン・D・ケレハーら著)』NEWTON PRESS

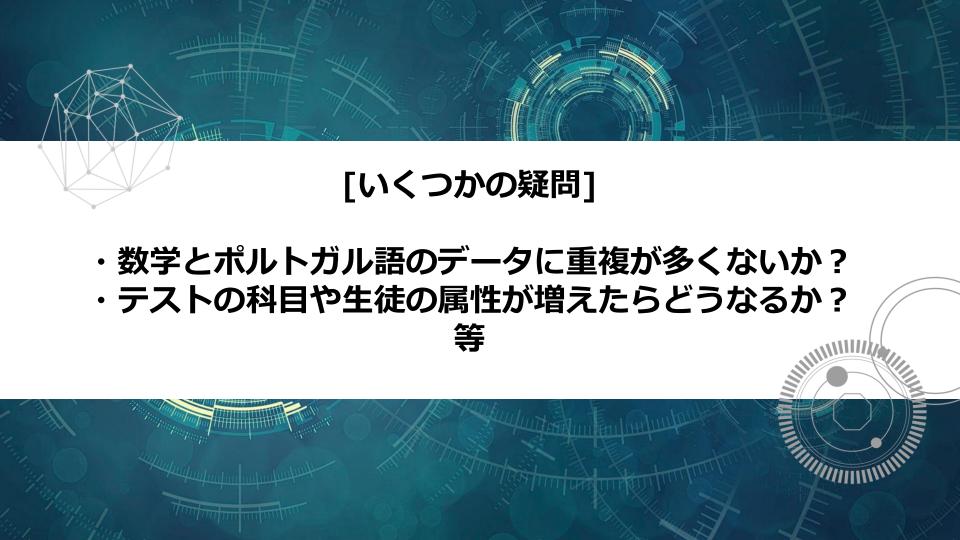
\*出典元: キチン2014年a、ハン、カンバー、ペイ2011年

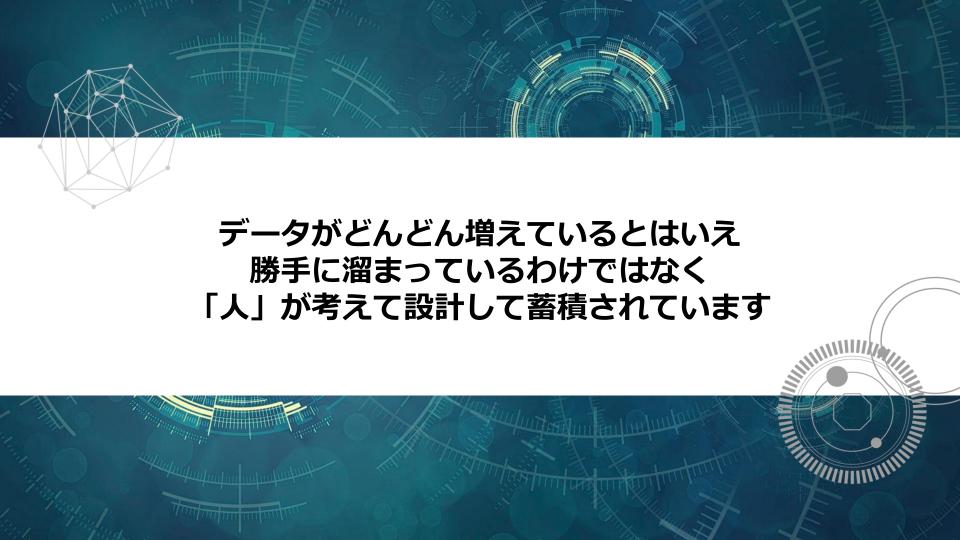
© 2024 東京大学松尾・岩澤研究室 All rights reserved

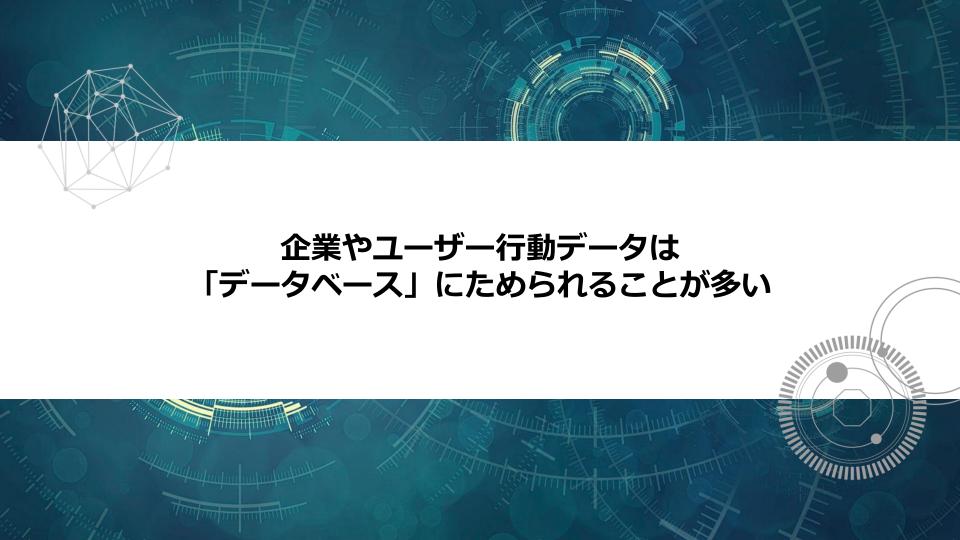
# 再掲 この講義で扱うデータの例

# このデータの形で蓄積することが最適なのか?

school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	 famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
GP	F	18	U	GT3	Α	4	4	at_home	teacher	 4	3	4	1	1	3	6	5	6	6
GP	F	17	U	GT3	Т	1	1	at_home	other	 5	3	3	1	1	3	4	5	5	6
GP	F	15	U	LE3	Т	1	1	at_home	other	 4	3	2	2	3	3	10	7	8	10
GP	F	15	U	GT3	Т	4	2	health	services	 3	2	2	1	1	5	2	15	14	15
GP	F	16	U	GT3	Т	3	3	other	other	 4	3	2	1	2	5	4	6	10	10
MS	М	20	U	LE3	Α	2	2	services	services	 5	5	4	4	5	4	11	9	9	9
MS	М	17	U	LE3	Т	3	1	services	services	 2	4	5	3	4	2	3	14	16	16
MS	М	21	R	GT3	Т	1	1	other	other	 5	5	3	3	3	3	3	10	8	7
MS	М	18	R	LE3	Т	3	2	services	other	 4	4	1	3	4	5	0	11	12	10
MS	М	19	U	LE3	Т	1	1	other	at_home	 3	2	3	3	3	5	5	8	9	9



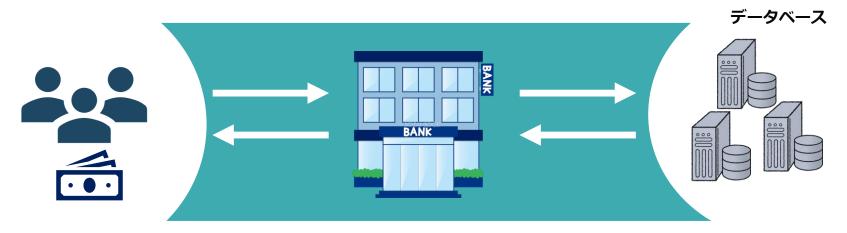




### 日常生活に関わるデータベース

### 様々な業界、業種でデータベースが活用されています

- ✓ATMでのお金の出し入れ、金融取引データ(株や為替の売買など)
- ✓コンビニやスーパーでの買い物 (POSデータ: Point of Sales)
- ✓ECサイト(楽天、Amazon、Yahooショップなど)での購買
- ✓予約システム(飛行機、電車、チケットなど)



# システムの中におけるデータベースのデータ例

# トランザクション(取引、入金など)データ例

Date	CustomerID	LocationNo	Update	Amt	Total
2019/5/27 10:00:00.00	20001	20	2019/5/27/ 10:00:00.00	20000	
2019/5/27 10:10:00.00	20010	25	2019/5/27/ 10:00:00.00	-20000	2019/5/27の10F
2019/5/27 10:11:00.00	20001	20	2019/5/27/ 10:00:00.00	20000	型の19/3/2/0710 顧客20001番の人 東京(20)で

※LocationNoの20は東京とする

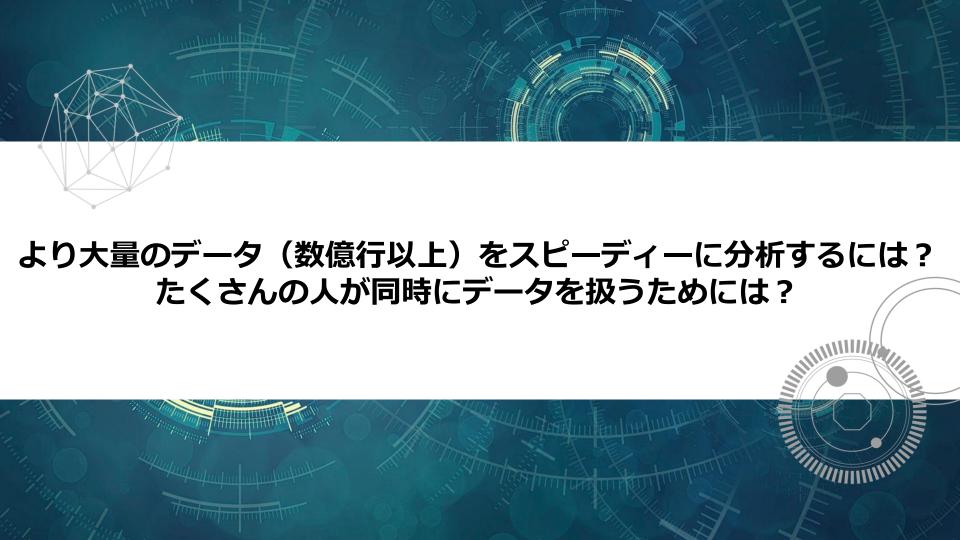
20,000円を入金

# システムの中におけるデータベースのデータ例

# 顧客情報の属性データ例

CustomerID	FM	Job	Update	Place	
20001	F	100	2019/4/27/ 10:00:00.00	60	
20010	М	13	2019/5/27/ 10:00:00.00	-20000	
20011	М	20	2019/5/27/ 10:00:00.00		顧客ID20001の 属性データ





## データベースを使う目的やメリット

### データの一貫性

- ・整合性や重複
- ・参照整合性





# データの独立性

システム処理とは 別で1箇所にまと める、など

※ほかは、セキュリティ観点なども考慮する



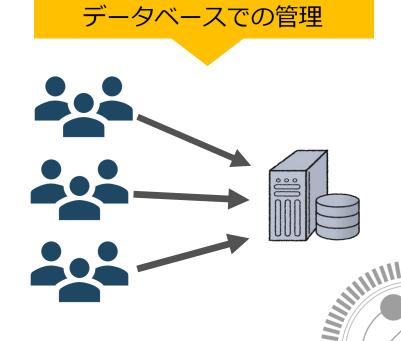
# データへのアクセスやデータの一貫性

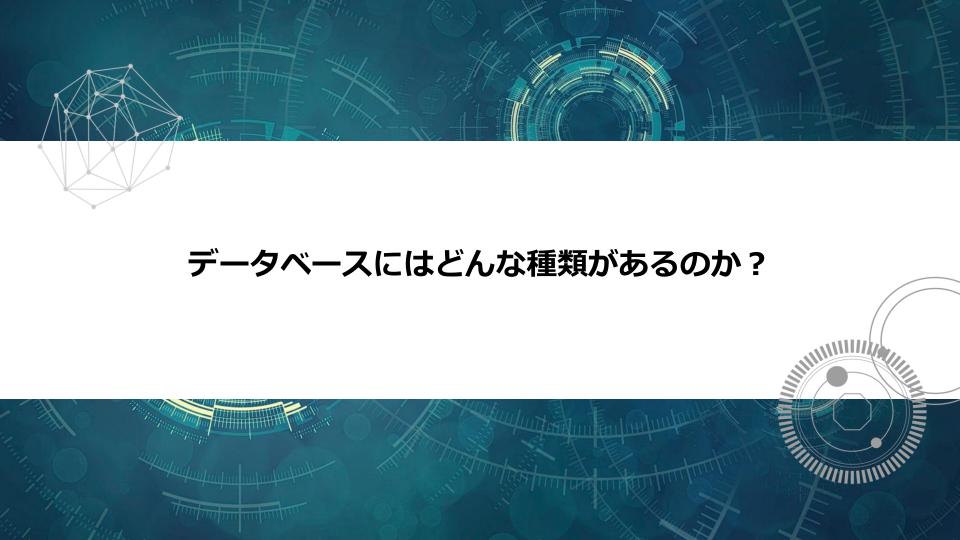
ユーザーはデータの格納場所を意識することなく利用することができ、

一元的に管理されているため、整合性も保てる

# CSV CSV

ファイルでの管理





# データベースの種類

よく使われている

- ▶ リレーショナルデータベース
- ▶ NoSQLデータベース
- ▶ XMLデータベース
- ▶オブジェクト指向型データベース
- ▶階層型データベース



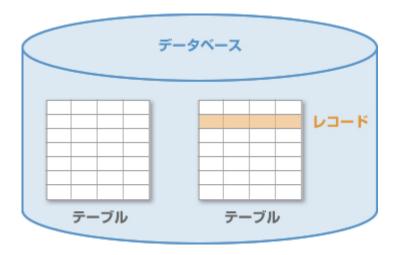


※近年はグラフデータベースも注目されている(後記載)

# リレーショナルデータベースとは?

### 二次元の表の形式で整理されたデータの集まりで、テーブル間で関係を持つ

# ※イメージとテーブル例



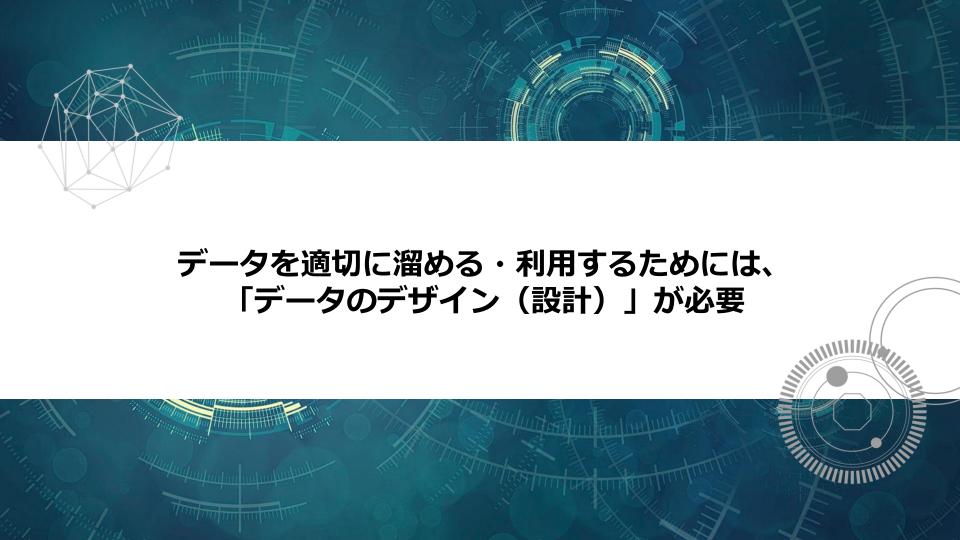
### 1. 顧客マスターテーブル

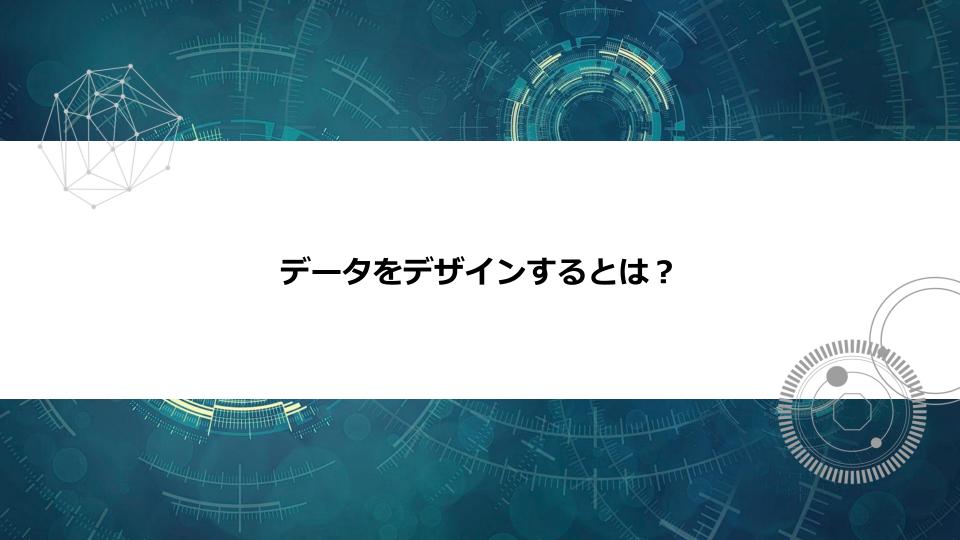
ID	birth_year	city	name
1	1989	Tokyo	Hishoshi
2	1990	Osaka	Akiko
3	1987	Kyoto	Yuki

### 2. 購買履歴テーブル

Date	ID	Commodity	Price
2018-04-01	1	ノート	108
2018-05-01	1	パソコン	60000
2018-05-04	2	バッグ	4000

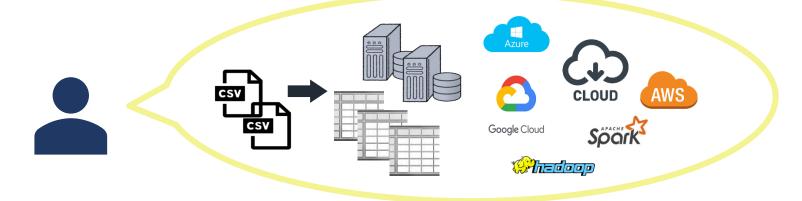
参照URL: http://www.webolve.com/image/basic/aboutdb/fig\_02.jpg





# データのデザインとは?

目的、業務要件に沿って、データインフラ、データベースの構築なども含め、 後の工程でデータをどう処理するか、分析するか等を設計すること

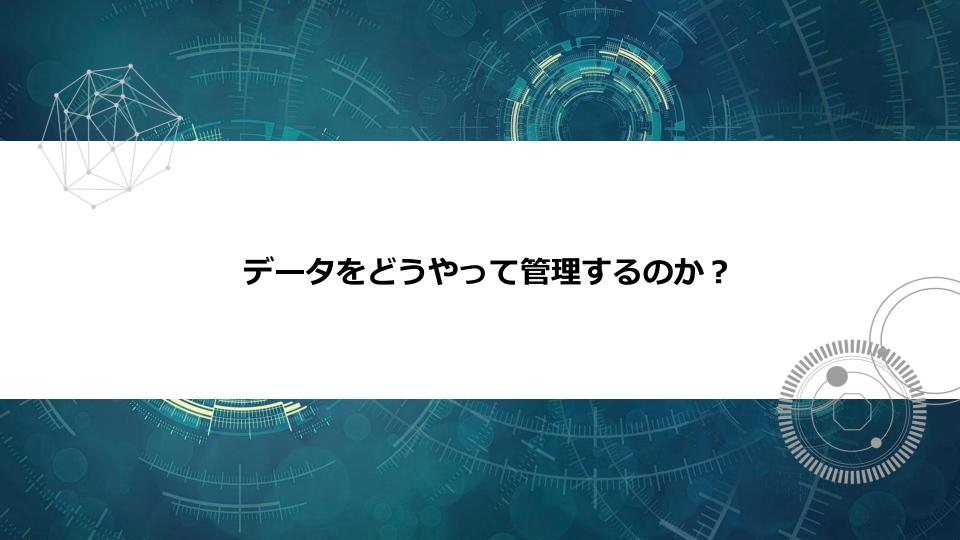


- 注:効率的にデータを処理できるアーキテクト、重複なくデータを抽出できるテーブル設計、 業務知識、クラウド・インフラ知識なども必要
- ▶ キーワード:データベースのデザイン、データモデリング、データアーキテクト、データベースアドミニストレータ、データ整備人、データ構造、ER図、正規化、ETL(後ほど説明)等

# 再掲 講義で扱うデータの例

school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	 famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
GP	F	18	U	GT3	А	4	4	at_home	teacher	 4	3	4	1	1	3	6	5	6	6
GP	F	17	U	GT3	Т	1	1	at_home	other	 5	3	3	1	1	3	4	5	5	6
GP	F	15	U	LE3	Т	1	1	at_home	other	 4	3	2	2	3	3	10	7	8	10
GP	F	15	U	GT3	Т	4	2	health	services	 3	2	2	1	1	5	2	15	14	15
GP	F	16	U	GT3	Т	3	3	other	other	 4	3	2	1	2	5	4	6	10	10
MS	М	20	U	LE3	Α	2	2	services	services	 5	5	4	4	5	4	11	9	9	9
MS	М	17	U	LE3	Т	3	1	services	services	 2	4	5	3	4	2	3	14	16	16
MS	М	21	R	GT3	Т	1	1	other	other	 5	5	3	3	3	3	3	10	8	7
MS	М	18	R	LE3	Т	3	2	services	other	 4	4	1	3	4	5	0	11	12	10
MS	М	19	U	LE3	Т	1	1	other	at_home	 3	2	3	3	3	5	5	8	9	9

# ▶ ビジネス要件、システム要件などによって、データの持たせ方は変わる



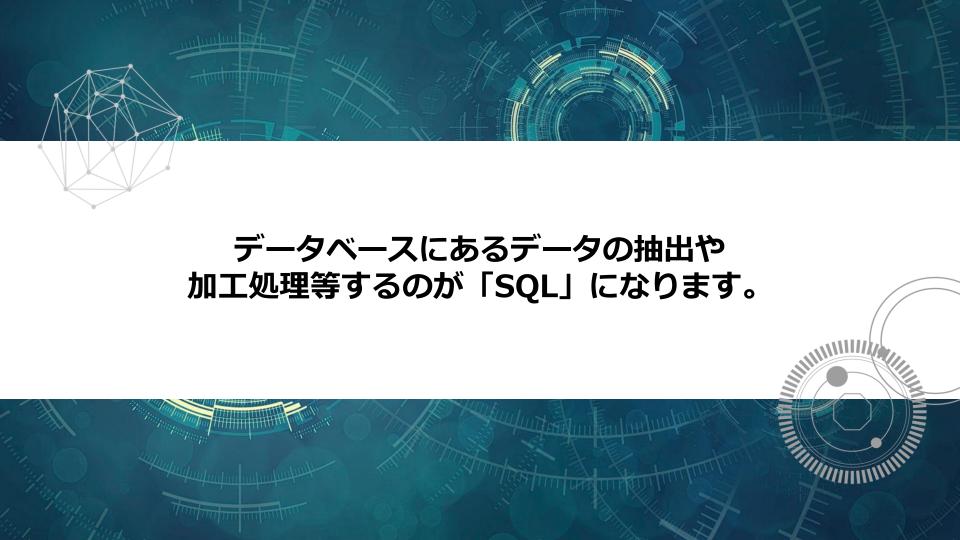
# **DBMS(Database management system)の種類**

DBMSとはデータベースの機能を提供するソフトウェア

DBMSのうちOracle、MySQLやPostgreSQL、SQLServerなどがある



参照URL: https://www.indiamart.com/global-standards-for/oracle-sql-mysql.html



### 再掲

# システムの中におけるデータベースのデータ例

★SQLを使えば、データ抽出の条件を指定するなどして、データ分析できる

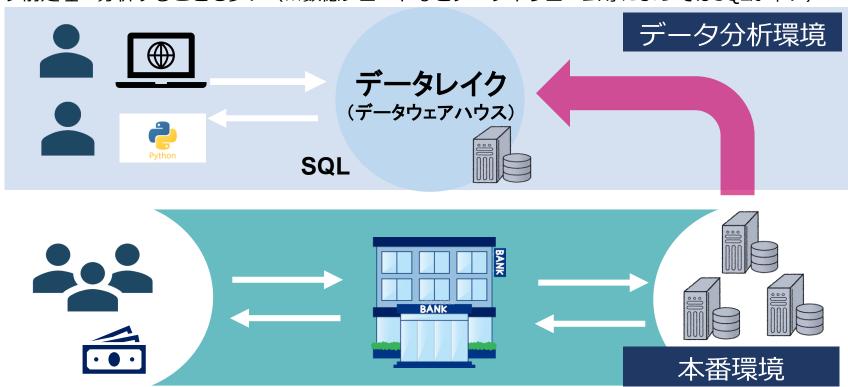
## トランザクション(取引、入金など)データ例

Date	CustomerID LocationNo Upo		Update	Amt		Total	
2019/5/27 10:00:00.00	20001	20	2019/5/27/ 10:00:00.00	20000			
2019/5/27 10:10:00.00	20010	25	2019/5/27/ 10:00:00.00	-20000	2019	9/5/27の10	·哇/:-
2019/5/27 10:11:00.00	20001	20	2019/5/27/ 10:00:00.00	20000	顧客	ジャックログ 20001番の。 〔(20)で	
	·		·			100円を3全	

※LocationNoの20は東京とする

# データサイエンスとデータベースの関係性

データを蓄積する環境を構築(データレイク等)し、そのデータをSQLやpython等を使ってデータ前処理・分析することも多い(※数億レコードなどデータボリューム等によってはSQLが早い)



# 再掲 データ分析の8~9割はデータ加工処理など

# これを関係データベースの設計して、データ加工処理してテーブルにいれる(※ETL)

データとして厄介な例(※以下は金融の半構造化データ、テラバイト級/monthで約20億行以上)

```
2016-01-01 10:10:10:000 8=FIX.4.4/x019=122/x019=122/x0135=D/x0134=215/x0149=CLIENT12/x0152=2010022519:41:57.316/x0156=B/x019=1
/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=2010022519:39:52.020/x0110=072/x01
\sqrt{x}019=122\sqrt{x}011=Marcel\sqrt{x}0111=13346\sqrt{x}0121=1\sqrt{x}0140=2\sqrt{x}0144=5\sqrt{x}0154=1\sqrt{x}0159=0\sqrt{x}0160=20100225-19:39:52.020 \sqrt{x}0110=072\sqrt{x}012016-01-
0110:10:10:0008=FIX.4.4/x019=122/x0135=D/x0134=215/x0149=CLIENT12/x0152=2010022519:41:57.316/x0156=B
/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=2010022519:39:52.020/x0110=072
\sqrt{x}01/x019=122/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-19:39:52.020/x0110=072/x01 2016-01
10:10:10 000 8=FIX 4 4/x019=122/x0135=D/x0134=215/x0149=CI IFNT12/x0152=20100225-
19:41:57.316/x0156=B/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-
19:39:52.020/x0110=072/x01/x019=122/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-
19:39:52 020/x0110=072/x01 2016-01-01 10:10:10 000 8=FIX 4 4/x019=122/x0135=D/x0134=215/x0149=CLIFNT12/x0152=20100225-
19:41:57.316/x0156=B/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-
19:39:52.020/x0110=072/x018=FIX.4.4/x019=122/x0135=D/x0134=215/x0149=CLIENT12/x0152=20100225-
19:41:57.316/x0156=B/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-
19:39:52.020/x0110=072/x018=FIX.4.4/x019=122/x0135=D/x0134=215/x0149=CLIENT12/x0152=20100225-
19:39:52.020/x0110=072/x018=FIX.4.4/x019=122/x0135=D/x0134=215/x0149=CLIENT12/x0152=20100225-
```

### ETLとは?

Extract(抽出)、Transform(変換)、Load(読み込み)の頭文字で、 データをスムーズに分析したり、機械学習のモデルに読み込むことが目的

### O1 抽出 [Extract]

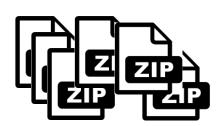
2016-01-01 10:10:10.000

8=FIX.4.4/x019=122/x019=122/x0135=D/x0134=215/x 0149=CLIENT12/x0152=2010022519:41:57.316/x0156 =B/x019=122

/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144= 5/x0154=1/x0159=0/x0160=2010022519:39:52.020/x0 110=072/x01

/x019=122/x011=Marcel/x0111=13346/x0121=1/x0140 =2/x0144=5/x0154=1/x0159=0/x0160=20100225-19:39:52.020 /x0110=072/x012016-01-

0110:10:10.0008=FIX.4.4/x019=122/x0135=D/x0134= 215/x0149=CLIENT12/x0152=2010022519:41:57.316/x 0156=B=



### 02 変換 [ Transform ]

Date, 2016-01-01 10:10:10.000

8,FIX.4.4

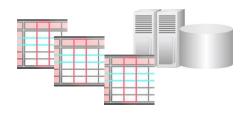
19,12

135,D

134,215

149,CLIENT ....

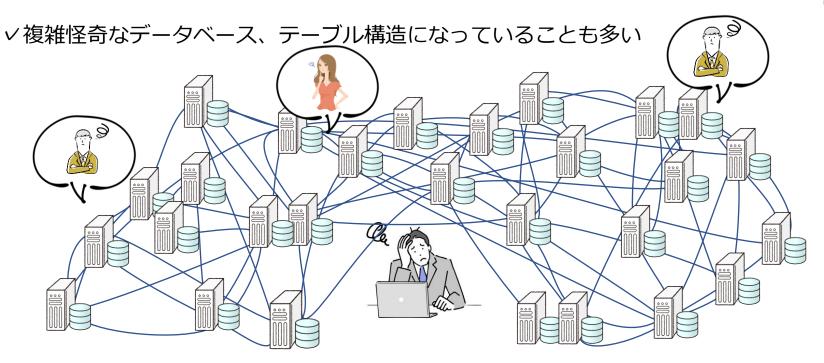
# **03** 読み込み [ Load ]



Date	8	19	135	134
2016/1/1 10:00:00.00	FIX.4.4	12	D	215
2016/1/1 10:00:01.00	FIX.4.4	12	D	215
2016/1/1 10:00:10.00	FIX.4.4	12	D	215

### 89%もの企業がレガシーシステムに悩んでいる

∨途中で改修されたシステムとの連携がたくさん



参照: Experian 社の「2019 Global Data Management Research」レポート ※今後は生成AI等の活用で整理される可能性もあり

# (再掲) データ戦略が重要

データを活用していくためには、様々な顧客データ(属性データ、行動データなど)を 統合したり、連携処理したり、管理していく必要があるため「データ戦略」が重要となる



## 参考文献 データ活用システム



# 初級~中級



『データ活用システム開発ガイド』 (東京化学同人)

[キーワード] MVP, 構造化データと非構造化データ,MySQL,WebAPI, ABテスト,Cloud(GCP,BigQuery),データパイプライン, データレイク,データウェアハウス, etc



# 初級~中級

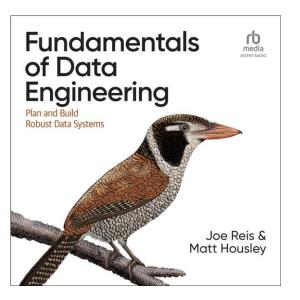


『データ分析基盤入門』 (技術評論社)

[キーワード] MVP, 構造化データと非構造化データ, MySQL, WebAPI, ABテスト,Cloud(GCP,BigQuery),データパイプライン, データレイク,データウェアハウス, etc



# 初級



[キーワード] データエンジニア、データエンジニアリングライフサイクル(生成、 保存、取り込み、変換、提供)、DataOps、データアーキ

テクチャ、データメッシュ、API、クエリ、データモデリング、ストリ

ーミングなど

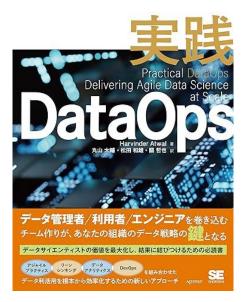
Fundamentals of Data Engineering: Plan and Build Robust Data Systems J (オライリー)



# データ戦略とDataOps



# 初級~中級



『実践DataOps』 (翔泳社) [キーワード]

データサイエンスの問題点。データ戦略、リーンシンキング、アジャイル、システム思考、DataOps、DevOps、組織づくり、





# SQL書き方の基本(Pandasでいうと?)



```
select
   column_name1,
                            カラムと集約関数等を指定
   column_name2,
   sum(column_name3)
from
                            データを指定
   TABLENAME
where
   column_name1 == "aaaa"
                            条件を指定
group by
   column_name1,
   column name2
                            軸を指定
```



# データベースとテーブルの作成:練習問題(5分)

#### 練習問題①

新しいテーブル「meibo2」を作成してください。作成後はテーブルが完成しているか、確認しましょう。データのカラム名とその型はなんでも構いません。

#### 練習問題②

新しく作ったテーブルにデータ を追加してみましょう。追加後 は、確かにデータが入っている か確認してください。



# データの検索と更新、削除、テーブルカラムの変更:練習問題(5分)

#### 練習問題①

テーブルmeiboにて、id=4の人レコードを抽出してください。

#### 練習問題②

上のmeiboテーブルでid=8の人のclassを7にアップデートしてください。select 文でアップデートを確認できたら、id=8の人のclassを1に戻してください。

#### 練習問題③

※必須問題:以降の問題で使います。

上記と同じテーブルmeiboに新しい列heightという身長を追加してください。さらに、id=1から4までの人は150、id=5から6までの人は155、id=7から8までの人は160でアップデートしてください。



# データの集計、演算、並び替え:練習問題(5分)

#### 練習問題①

上記と同じmeiboテーブルで、異なる年齢は何種類でしょうか。

#### 練習問題②

身長が一番小さい人、大きい人の身長をそれぞれ求めてください。

#### 練習問題③

身長が155以上で、classが3のレコードを抽出してください。



#### 練習問題①

meiboテーブルにてクラスごとの平均身長を求めてみましょう

#### 練習問題②

上に加えて、クラスごとの人数、一番小さい人と、一番大きな人の身長も それぞれのクラスで算出しましょう。

#### 練習問題②

年齢が13歳より上の人に絞って、上記と同じ項目を求めてください

# 複数テーブルの利用:練習問題(5分)

#### 練習問題①

新しくcardtbというテーブルを作成してください。ただし、カラムは id,point,money (すべてint)にしてください。さらに、以下のデータを挿入してください。

(id,point,money) = (1,100,1000),(3,50,500),(4,30,600),(5,10,10),(7,100,1000),(8,2000,100)

#### 練習問題②

meiboテーブルにidをキーとして、上記のテーブルを内部結合させてください。

#### 練習問題③

meiboテーブルにidをキーとして、上記のテーブルを外部結合させてください。



#### 練習問題①

meiboテーブルで、heightが155未満の場合は"below\_155"、155ぴったりならば"equal\_155"、155より大きいならば"over\_155"と名付けて、テーブルを表示させてください。

#### 練習問題②

上記のテーブルとサブクエリの考え方を利用して、それぞれのheightLevel(練習問題1で付与した列)の人数を求めてください。

#### 練習問題③

上記のテーブルとサブクエリの考え方を利用して、それぞれのheightLevel(練習問題1で付与した列)の平均年齢を求めてください。

### 補足

# なぜストアドプロシージャが必要なのか?

- ★ストアドプロシージャとは、SQLを使った一連の処理をデータベースで 実行するプログラムのこと
- ∨ SQLを手続き型プログラムのように書ける(OracleのPL/SQLなど)
- ∨ Pythonなど他言語で処理することもあるが、ネットワークレーテンシーの問題や、 システム構成やデータアーキテクトの目的による
- ✓ 金融システムなどの処理はパフォーマンスが求められ同じサーバー内にあるSQLプロシージャ (PL/SQLなど)で実行することが多い



### 補足

# グラフデータベース(近年注目されているデータベース)

# ★グラフデータベースを使うと、データ間の新しい関係性やつながりを発見できる

- ✓ ソーシャルネットワーク、マスタデータ管理、地理空間、リコメンデーション、不正検知などの領域応用される
- ✓ リレーショナルデータベースでは、つながりのある関係を扱うのが難しい(SQLのクエリが複雑になりがち)
- ∨ 機械学習モデル構築時の特徴量エンジニアリングにも役立つ

SWANTINE PRODUCT LOW LONG TO THE PROPERTY OF T

参照:https://thefintechtimes.com/neo4j-why-graph-technology-is-the-key-to-fraud-detection/

#### リレーショナルデータベースとグラフデータベース(Neo4j)との比較

データの深さ	RDBMS 実行時間(秒)	グラフデータベース 実行時間(秒)	返されたレコード数		
2	0.016	0.01	~2500	1000倍以上の	
3	30.267	0.168	~11万	パフ	オーマンス
4	1543.505	1.359	~60万		
5	未完了	2.132	~80万		4



参照: 『グラフデータベース』 (オライリー) の p17

© 2024 東京大学松尾・岩澤研究室 All rights reserved



# 今後データベース関連で学習すべきこと

01 SQLの中級者への道 (後で紹介する参考図書を参照)

正規化、パフォーマンスチューニング・インデクシング、より複雑なSQL処理、データベース設計やアーキテクト、SQLを使ったシステム開発、テクニカルドキュメント

※自動で管理やチューニング等やってくれるサービスもある ※Oracleなどテクニカルドキュメントがかなり充実している https://docs.oracle.com/cd/F39414 01/index.html

03

NoSQL、Hadoop、Spark (分散処理システム) やグラフデータベース ベクトルデータベース等





02 Pythonや他言語などとの連携



04

クラウドサービス

AmazonのAWS、GoogleのGCP、Microsoftの Azureなど。RedshiftやBigQuery、最近は Snowflakeなどの各社独自のデータベースやデー タウェアハウス等もある。

※たとえば、GoogleのBigQueryなどを使えば、SQLでMLモデルを構築することも可能

# データベースに関連する資格

### 資格試験

- 1. 基本情報処理
- 2. 応用情報処理
- 3. データベーススペシャリスト
- (4. ネットワークスペシャリスト)

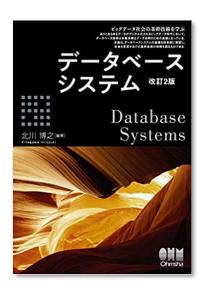


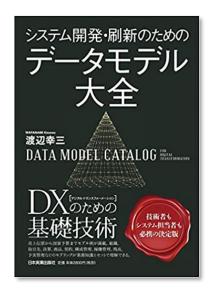
# データベース全般



# データベース の概要、設計などの基礎~中級







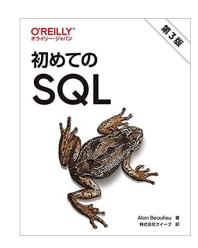


# 基礎

『初めてのSQL』 (オライリー・ジャパン)



データベースとSQLの基本を幅広く







SQLの練習ウェブサイト

https://sqlzoo.net/wiki/SQL\_Tutorial



# 基礎

『基礎からのMySQL 第3版 (基礎からシリーズ)』 (SBクリエイティブ)



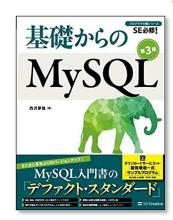
データベースとSQLの基本から学べる + PHPも少し

# 基礎

『おうちで学べるデータベースのきほん』 (翔泳社)



データベースとSQLの復習







# SQLやデータベース中級レベル

# 中級~応用



『ビッグデータ分析・ 活用のためのSQLレシピ』 (マイナビ出版)



『プログラマのためのSQL』 (翔泳社)



『データベース実践入門』 (技術評論社)



『SQLクックブック』 (オライリー社)



# 参考文献 データベース応用レベル

# 中級~応用



『ビッグデータ分析ーシステムと 開発がしっかりとわかる教科書ー』 (技術評論社)



『データ指向アプリケーション デザイン』 (オライリー社)



『データマネジメント 知識体系ガイド』 (日経BP社)

#### **Database System Concepts**

SEVENTH EDITION



Database System
 Concepts
 (McGrawHill)

# データの作成と機械学習



# 中級

**Human-in-the-Loop** 機械学習 人間参加型 AIのための 能動学習とアノテーション

Data-Centric AI

機械学習アルゴリズムを高度化することよりも、正しいデータを正しくアノテーションして作成することのほうが何倍も価値がある(本書引用)

『Human-in-the-Loop 機械学習』 (共立出版)

#### 1番オススメの学習方法は?

★掲示板ウェブサイトや簡単なニュース配信アプリなど開発することで、 データベースだけではなく、関連スキルも身に付けることができる。

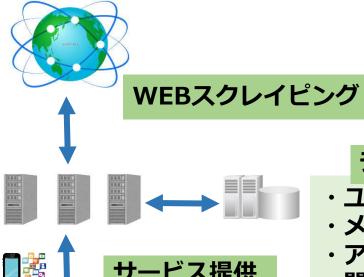
# クラウド環境

- AWS
- · GCP
- · Azure, etc



# ウェブアプリ開発

- PHP · JavaScript
- Django(Python)
- ∙ HTML∙ CSS etc



サービス提供

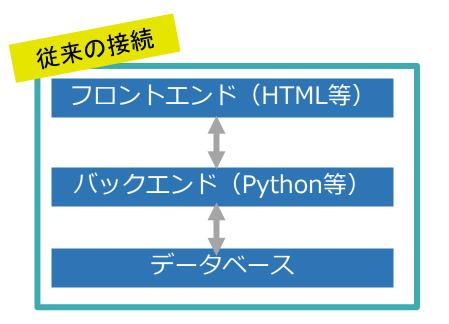
・リコメンドetc

# データベース

- ・ユーザー管理
- ・メッセージ管理
- ・アクセスログ管理
- ・関連データの保存 etc

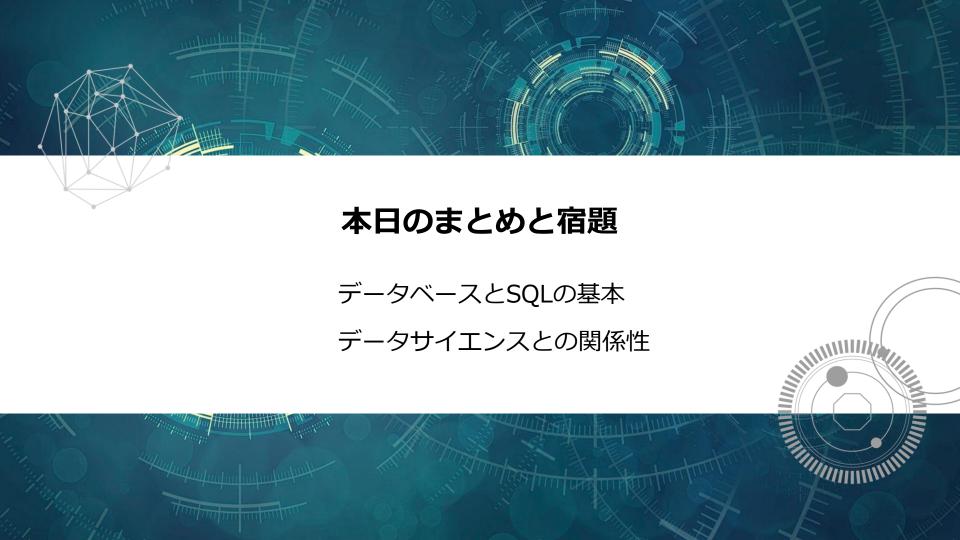
#### おまけ(Web3との関係性):中央集権的なデータベースからブロックチェーンへ??

★Web3では、従来のバックエンドではなく、ブロックチェーンに対してデータを読み書きすることになる。





- ※参考『SolidityとEthereumによる実践スマートコントラクト開発』(オライリー社)
- ※参考『ブロックチェーン実践入門』(オーム社)
- ※Web3:ブロックチェーンを簡単に操作できるAPIを提供する



# 本日のまとめ

- で データベースとは何か、データベースとデータサイエンス データベース、DBMS、データ設計、SQL、データベースと データサイエンスの関係性
- で データベースとテーブルの作成 データベース作成、テーブル、show databases,use,show tables, create, select,主キー
- 03 データの挿入、検索、更新、カラムの作成 where, in, like, update, delete, alter
- 04 データの集計、複数テーブルの利用、case文 group by, having,union,union all, inner join(内部結合),outer join(外部結合),自己結合
- 05 今後の学習コンテンツ データベース中級(インデクシング、複雑なSQL)、NoSQL、クラウドなど





M 尾·岩澤研究室 MATSUO-IWASAWA LAB UTOKYO