デブサミ2008 Developers Summit

みんなまとめて面倒みよう

～真のDBエンジニアになるために必要なこと～

13-E-7

講演者：ミック

自己紹介

名前:ミック

経歴:主にOracleを使ったデータウェアハウス業務に従事するDBエンジニア。

著書:『達人に学ぶ SQL徹底指南書』(翔泳社 2008)
訳書:J.セルコ『SQLパズル 第2版』(翔泳社 2007)

CodeZine(https://codezine.jp/)誌上にてSQLとデータベースについての記事を連載中。

HPのコンテンツ『リレーショナル・データベースの世界』（http://www.geocities.jp/mickindex/database/idx_database.html）。

本セッションの概略

・自著(『指南書』)のテーマと狙いについて
　⇒二つの意味での架橋:初級から中級、原理と実践

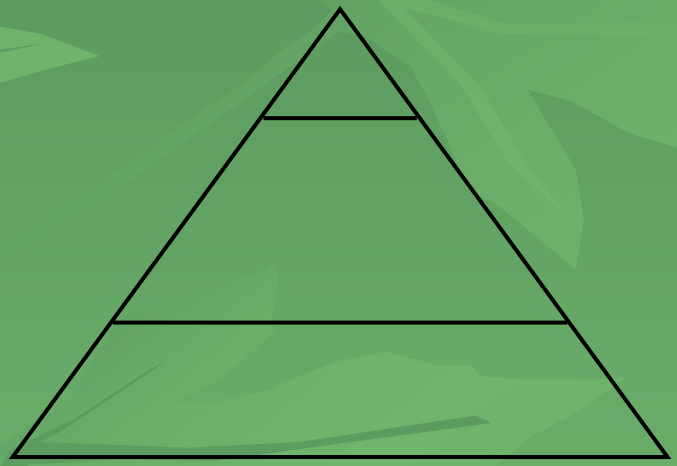
・リレーショナル・データベースという世界の原理を理解する
　⇒二つの基礎:集合論と述語論理
　⇒手続き型言語とSQLの発想の違いを肌で知ろう。

・DBエンジニアとは、どんな仕事か、またどんな仕事であるべきか
　⇒リレーショナル・データベースとアメリカ文化の関係
　　自らの立ち位置を正しくマッピングすること
　⇒面白さの発見の仕方:冒険家型と探偵型

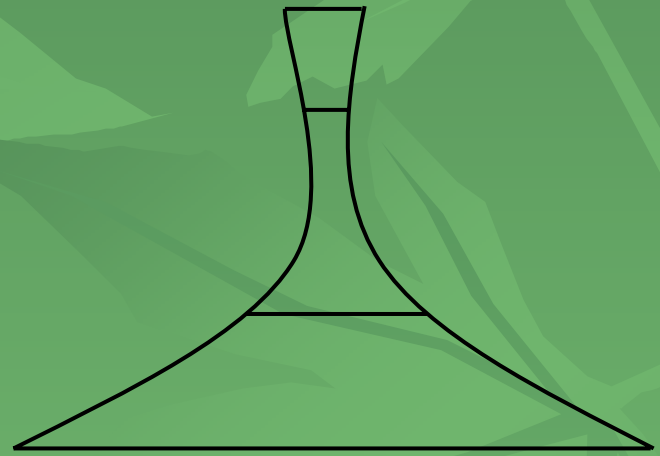
現在の日本のDB界の問題

中堅層が少ない、いびつな形

通常の技術分野：ピラミッド型



日本のDB界：中間のえぐれた花瓶型



技術書の偏りもこれに準じている。そのため、日本のDBエンジニアは、「バイエルの次にいきなりショパンを弾け」と強要されているような状況。

そしてもっと深刻な問題：もしこの花瓶型のまま十年経過すると、分布の形はどうなるか？

SQLの原理を理解しよう

SQLはなぜ理解しにくいのか。
⇒手続き型の考え方に慣れた私たちは、
集合指向（set-oriented）という発想が奇異に感じられる。

・SQLでは手続き、すなわち文を基本単位にしない。
⇒代入、分岐、ループ、ソートといった手続きも一切現れない。

・代わりにSQLは、世界を集合で表現する
SQLの原理は、集合と述語（特性関数）
でも私たちは普通、この二つの概念に馴染みがない
（標準的な学校教育では習わないから）

SQLの多彩な武器

SQLを支える強力な機能の数々：

・CASE式：

SQLで分岐を表現するためになくてはならない。これを使いこなせれば、SQLの関数型言語としての貌が見えてくる。実際、CASE式はLispのcondと同じと考えていい。

・GROUP BY句、PARTITION BY句：

群論の類別（partition cut）を実装したもの。集合を切り分けて新たな集合を作るときに必ずといってよいほど登場する。

・HAVING句：

集合に対する条件、つまり二階の条件を設定するいぶし銀。しかも、これの使い方を学ぶことで、階(order)の概念とSQLの集合指向という理念まで理解できるという教育効果もある。私はSQLにとって最重要の機能の一つだと考えています。

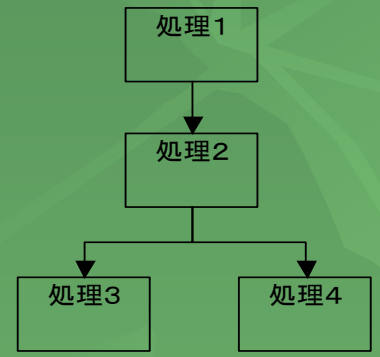
・EXISTS述語：

HAVINGが集合論的な意味での二階の機能であるのに対し、こちらは述語論理における二階の機能。SQLが持つ唯一の高階関数。

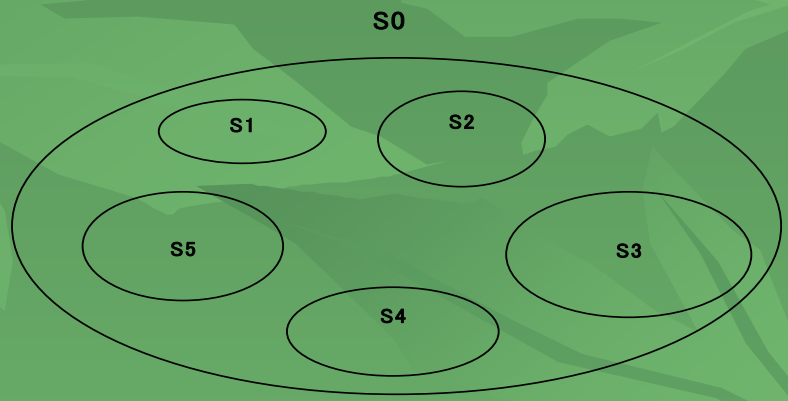
手続き型と集合指向の対比

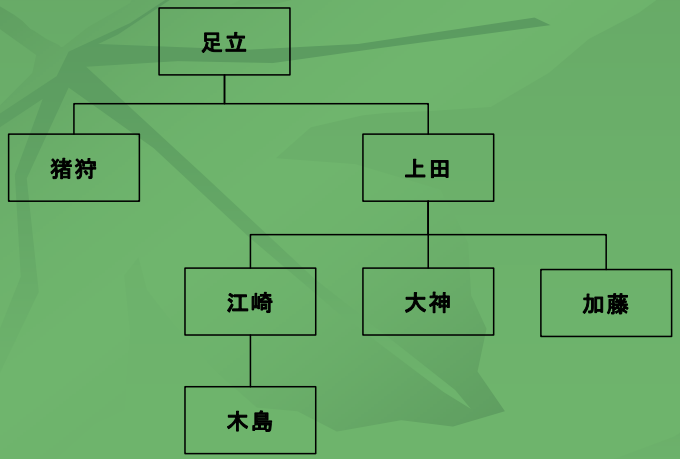
SQLでは集合を次々に組替えて求める集合へ辿り着く

手続き型の思考パターン

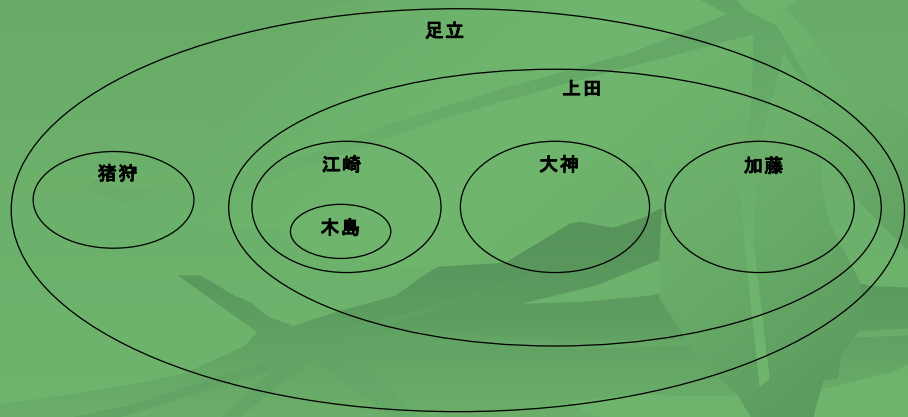


集合指向の思考パターン





（隣接リストモデル）



（入れ子集合モデル）

集合指向と手続き型の考え方の比較

サンプル：メジアン（中央値）の求め方

Weights（データ数が奇数）

student_id	weight
A100	50
A101	55
A124	55
B343	60
B346	72
C563	72
C345	72

メジアン = 60

Weights（データ数が偶数）

student_id	weight
A100	50
A101	55
A124	55
B343	60
B346	72
C563	72
C345	72
C478	90

メジアン = 66

集合指向でメジアンを求める:コード

特性関数を利用して、集合を上位と下位に分割

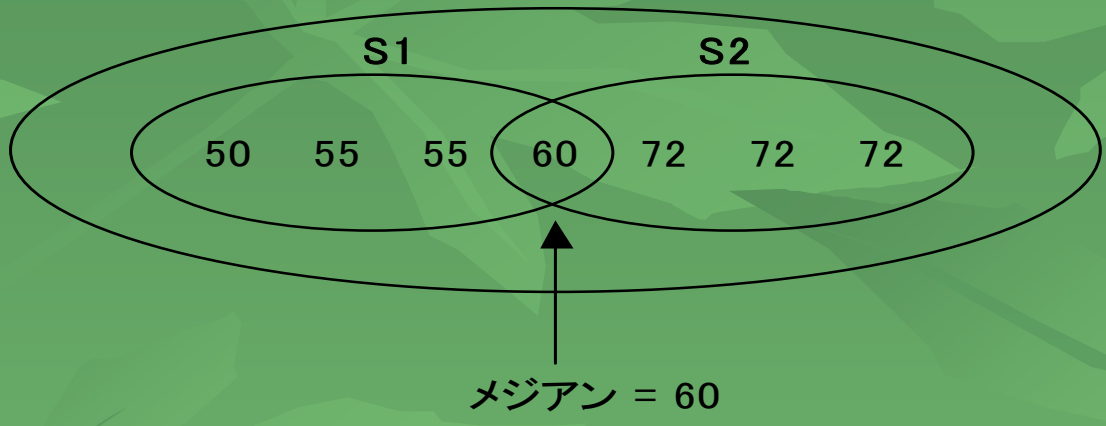
```
SELECT AVG(weight)
  FROM (SELECT W1.weight
          FROM Weights W1, Weights W2
         GROUP BY W1.weight
           --S1(下位集合)の条件
        HAVING SUM(CASE WHEN W2.weight >= W1.weight THEN 1 ELSE 0 END)
                   >= COUNT(*) / 2
          --S2(上位集合)の条件
           AND SUM(CASE WHEN W2.weight <= W1.weight THEN 1 ELSE 0 END)
                   >= COUNT(*) / 2 ) TMP;
```

(『指南書』1-4「HAVING句の力」参照)

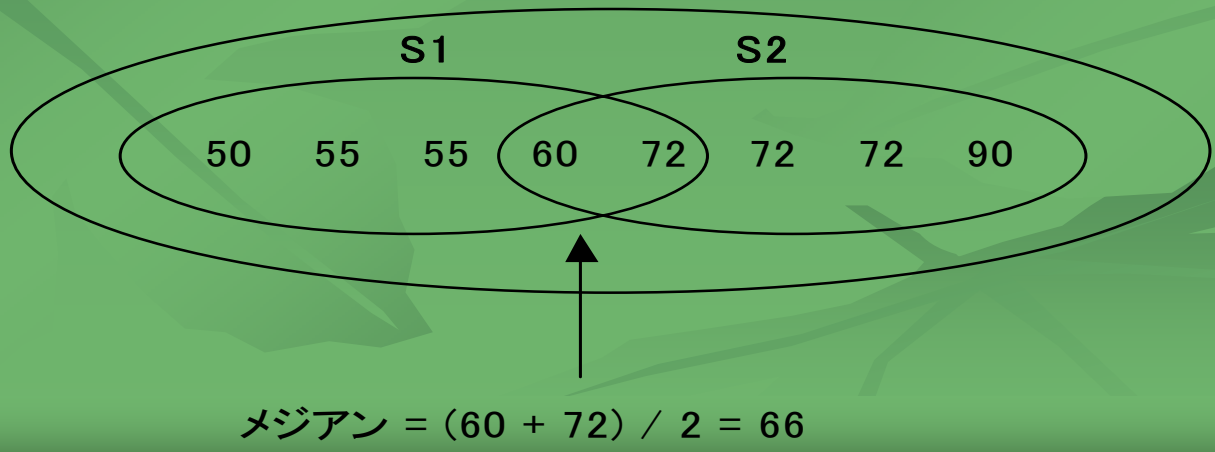
集合指向でメジアンを求める：図解

入れ子の集合をイメージしよう

奇数のケース



偶数のケース



手続き型でメジアンを求める

順序とソートの概念を前面に押し出す

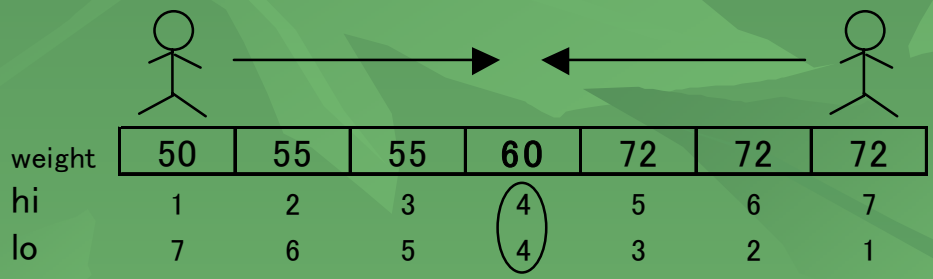
```
SELECT AVG(weight) AS median
  FROM (SELECT
          weight,
          ROW_NUMBER() OVER (ORDER BY weight ASC,  student_id ASC)  AS hi,
          ROW_NUMBER() OVER (ORDER BY weight DESC, student_id DESC) AS lo
        FROM Weights) TMP
 WHERE hi IN (lo, lo +1 , lo -1);
```

(J.Celko『Joe Celko‘s Analytics And OLAP in SQL』「8.1 OLAP Functionality」参照)

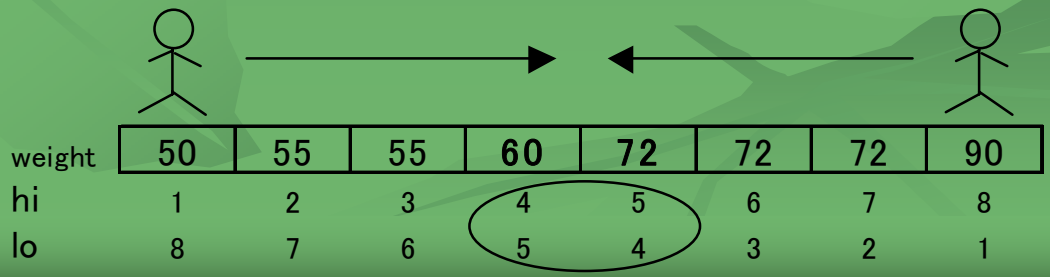
手続き型でメジアンを求める：図解

二人の旅人が世界の中心を探す

1．奇数の場合、ちょうど4番で出会う。
メジアン = 60



2．偶数の場合、4番と5番の中間で出会う。
メジアン = (60 + 72) / 2 = 66



これまでのまとめ

これからのSQLの方向性：

基本的には集合指向の考え方をベースにしつつ
手続き型の考えもミックスしていくハイブリッド型になる。

ex. SQL:2003ではOLAP関数が追加され、「行の順序」の概念がSQLに持ち込まれた。

また、シーケンス・オブジェクトも標準化され、これから「連番」を扱う機能も強化されていく可能性は高い。

～ちょっと脱線：面白いことの見つけ方～

新しいことを見つける二つのロールモデル：

1．冒険家型
- 誰も見たことのない新天地を探しにでかける。
  バイタリティと独創性が必要。
  素材にこだわる派。

2．探偵型
- 警察が捜査し終わった現場に新しい光をあてる。
  わりと偏執的で細かいことにこだわる。深読み深掘りが好き。
  調理にこだわる派。

（※この両者の対比は、歴史学者の論文の書き方や、コックの料理の仕方をイメージすると理解しやすいかも。）

データベースとアメリカ文化

問題：なぜデータベースの本場は、アジアでもヨーロッパでもなく、アメリカなのか？

（ヒント：GoogleやAmazonのような企業が、なぜ日本ではなくアメリカで生まれたのか、という理由とも関わってくる）

正しさを評価する（アメリカの）二つの基準軸

1. 実証主義：「データが多ければ多いほど正しい」

⇒統計分析、データマイニングの基礎となる考え方

アメリカの政府や企業がデータを大量に集めるのが好きなのはこのため。

（Googleの保持するデータ規模は4ペタバイト。）

2. 民主主義：「支持する人が多いほど正しい」

⇒マーケティングの基礎となる考え方。多数決主義。

例：映画の結末を複数用意して、投票で正式な結末を決定する。

大統領を国民投票で選ぶのも、裁判が陪審員制なのもこれによる。

科学的事実を多数決で決める行き過ぎも⇒1925年の「進化論裁判」

両者に共通するのは、「多いことはいいことだ」の精神

多いことはいいこと、か

イアン・エアーズ『その数学が戦略を決める』（文藝春秋 2007）

訳者の山形浩生さんも驚く売れ行き好調なデータマイニング万歳本。第6章の見出しの一つは「あらゆるところにデータベースが」。

「われわれはいま、馬と蒸気機関の競争のような歴史的瞬間にいる。直感や経験に基づく専門技能がデータ分析に次々に負けているのだ。・・・・・・企業も政府も、意志決定をますますデータベースに頼るようになっている。」(p.20)

いまやアメリカでは、データベースによるマーケティング戦略はあらゆる業種で行われている（コンビニの売れ筋分析から出会い系サイトの相手診断まで）。

※一方、こうしたデータ重視の風潮に対する批判も（アメリカ国内にさえ）ある。

ex.「結局、失敗したときのもっともらしい言い訳、逃げ道作りなのではないか」

「かえって不平等や差別を助長し、社会を住みづらくするのではないか」

データベースと実証主義の精神

「データを積み重ねる」という点で、統計分析が重視される。当然、大量データを処理するために、データベースも重要になる。

一方、日本には伝統的に意志決定のプロセスに実証性を組み込む習慣がない（※）。

⇒データベースを本当に活用しようとするなら、道具だけでなく、実証主義の精神も導入する必要がある。

※もっとも、アメリカも昔から実証主義を重視していたわけではない。

それなりに痛い目を見て学習した結果。

「アメリカには、実証的根拠のない空虚な理論を経済政策や社会政策に採用したために、あたら公共財（税）を無駄に費やしてきたという、苦い経験がある。」（谷岡一郎『「社会調査」のウソ』p.102）

データベースの未来

・一つは従来どおりのOLTP

データ量は少なく、長期間の保持もしない代わりにトランザクション制御が重要。

・もう一つは、データウェアハウスの極端化

テラマイニングと称されるような大規模データを用いた統計分析ツール

おそらく、アメリカ主導が続く限り、この用途がこれから多くなる。

しかし本当に日本に根付くか？