Introduction to Data Management CSE 344

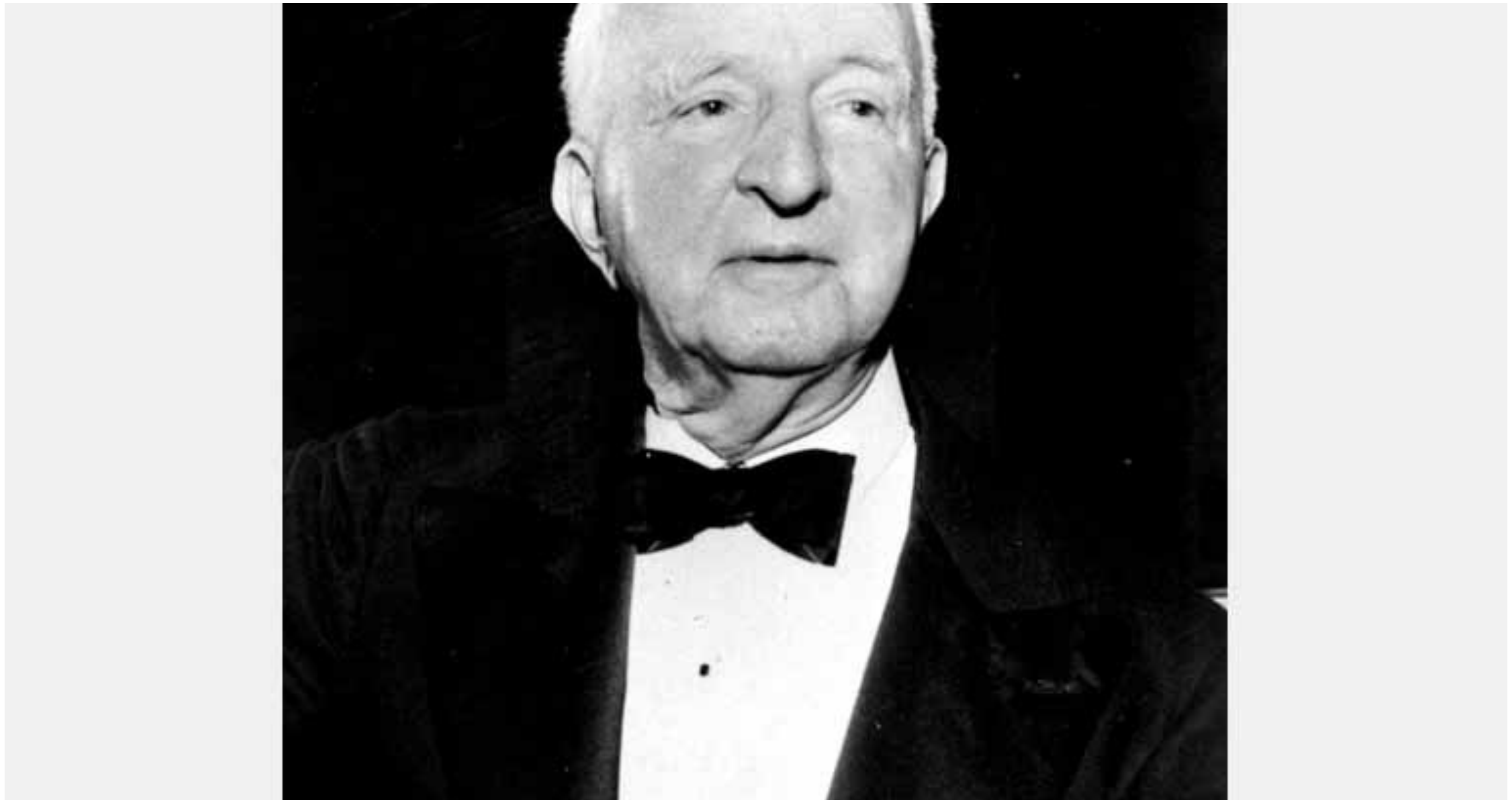
Lectures 5: Aggregates in SQL

Daniel Halperin

Announcements

- Webquiz 2 posted this morning
- Homework 1 is due on Thursday (01/16)

(Random detour:) Who is this?



http://content.lib.washington.edu/cdm4/item_viewer.php?CISOROOT=/portraits&CISOPTR=117&CISOBOX=1&REC=5

Does this help?

DISTINGUISHED SERVICE AS A REGENT
OF THE UNIVERSITY OF WASHINGTON
THE BUILDING WAS GIVEN THE NAME OF
WINLOCK W. MILLER HALL
AS A TRIBUTE TO HIS UNTIRING EFFORTS
TO ENSURE THE ORDERLY DEVELOPMENT
OF ITS CAMPUS

http://content.lib.washington.edu/cdm4/item_viewer.php?CISOROOT=/uwcampus&CISOPTR=1649

Winlock W Miller (of Miller Hall :)

- UW Regent (managers of univ.) for 35 years between 1913 and 1953
  - Usually full of executives from major instutions
  - Current Board of Regents Chair is former Alaska Airlines CEO, etc.
- Winlock, WA is named after him
- Father was Gen'l William Winlock Miller (confusing, I know), first mayor of Olympia and land speculator.
- (I think) WA has some interesting history!

Refresh your memory

```
> SELECT * FROM Purchase;
pid          product      price        quantity     month
----------   ----------   ----------   ----------   ----------
1            bagel        1.99         20           september
2            bagel        2.5          12           december
3            banana       0.99         9            september
4            banana       1.59         9            february
5            gizmo        99.99        5            february
6            gizmo        99.99        3            march
7            gizmo        49.99        3            april
8            gadget       89.99        3            january
9            gadget       89.99        3            february
10           gadget       49.99        3            march
11           orange       NULL         5            may
12           orange       1.29         34           january
```

Refresh your memory

How do we…

- Compute the total number of sales?
- Compute the total number of products sold?
- Compute the total number of each product sold?
- Compute the gross $ spent on of each product?
  (qty * price)
- Compute the average gross $ of each product?
  (2 ways)

Refresh your memory

How do we…

- Compute the gross monthly sales in $?
  (units * price/unit)
- Sort the months from most sales to least?
- Find all the unique prices? (2 ways)

HAVING Clause

Same query as earlier, except that we consider only products that had at least 30 sales.

```
SELECT      product, sum(price*quantity)
FROM        Purchase
WHERE       price > 1
GROUP BY product
HAVING      Sum(quantity) > 30
```

HAVING clause contains conditions on aggregates.

WHERE vs HAVING

- WHERE condition is applied to individual rows
  - The rows may or may not contribute to the aggregate
  - No aggregates allowed here

- HAVING condition is applied to the entire group
  - Entire group is returned, or not at all
  - May use aggregate functions in the group

Aggregates and Joins

```
create table Product
  (pid int primary key,
   pname varchar(15),
   manufacturer varchar(15));

insert into product values(1,'bagel','Sunshine Co.');
insert into product values(2,'banana','BusyHands');
insert into product values(3,'gizmo','GizmoWorks');
insert into product values(4,'gadget','BusyHands');
insert into product values(5,'powerGizmo','PowerWorks');
```

Aggregate + Join Example

What do these queries mean?

```
SELECT x.manufacturer, count(*)
FROM Product x, Purchase y
WHERE x.pname = y.product
GROUP BY  x.manufacturer
```

```
SELECT x.manufacturer, y.month, count(*)
FROM Product x, Purchase y
WHERE x.pname = y.product
GROUP BY  x.manufacturer, y.month
```

General form of Grouping and Aggregation

```
SELECT    S
FROM      R1,…,Rn
WHERE     C1
GROUP BY  a1,…,ak
HAVING    C2
```

Why ?

S = may contain attributes $a_1,\dots,a_k$ and/or any aggregates but NO OTHER ATTRIBUTES

C1 = is any condition on the attributes in $R_1,\dots,R_n$

C2 = is any condition on aggregate expressions and on attributes $a_1,\dots,a_k$

Semantics of SQL With Group-By

```
SELECT    S
FROM      R_1,…,R_n
WHERE     C1
GROUP BY  a_1,…,a_k
HAVING    C2
```

Evaluation steps:

1. Evaluate FROM-WHERE using Nested Loop Semantics
2. Group by the attributes $a_1,\ldots,a_k$
3. Apply condition C2 to each group (may have aggregates)
4. Compute aggregates in S and return the result

Empty Groups

- In the result of a group by query, there is one row per group in the result
- No group can be empty!
- In particular, count(*) is never 0

```
SELECT x.manufacturer, count(*)
FROM Product x, Purchase y
WHERE x.pname = y.product
GROUP BY x.manufacturer
```

What if there are no purchases for a manufacturer

Empty Groups: Example

```
SELECT product, count(*)
FROM purchase
GROUP BY product
```

5 groups in our example dataset

```
SELECT product, count(*)
FROM purchase
WHERE price > 2.0
GROUP BY product
```

3 groups in our example dataset

Empty Group Problem

```
SELECT x.manufacturer, count(*)
FROM Product x, Purchase y
WHERE x.pname = y.product
GROUP BY x.manufacturer
```

What if there are no purchases for a manufacturer

Empty Group Solution: Outer Join

```
SELECT x.manufacturer, count(y.pid)
FROM Product x LEFT OUTER JOIN Purchase y
ON x.pname = y.product
GROUP BY x.manufacturer
```

- 1) List all manufacturers with more than 10 items sold. Return the manufacturer name and the number of items sold.