SQL (Structured Query Language)

- Based on the Relational Calculus:
  - conjunctive queries
    - aka. **SELECT** blocks
  - set operations
  - update language
  - non first-order features

- **BAG** (multiset) Semantics

- **NULL** values
  - avoid if at all possible

- A **committee** design
  - often more **pragmatic** than logical
  - evolving **standard**:
    - SQL-89, **SQL-92**, SQL-1999,
Three major parts of the language:

1. **DML (Data Manipulation Language)**
   - query language
   - update language

   **Also:** Embedded SQL (SQL/J) and ODBC (JDBC)
   - necessary for application development

2. **DDL (Data Definition Language)**
   - defines *schema* for relations
   - creates (modifies/destroys) database objects.

3. **DCL (Data Control Language)**
   - access control
### SQL Data Types

Values of attributes in SQL:

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>integer</td>
<td>integer (32 bit)</td>
</tr>
<tr>
<td>smallint</td>
<td>integer (16 bit)</td>
</tr>
<tr>
<td>decimal(m,n)</td>
<td>fixed decimal</td>
</tr>
<tr>
<td>float</td>
<td>IEEE float (32 bit)</td>
</tr>
<tr>
<td>char(n)</td>
<td>character string (length n)</td>
</tr>
<tr>
<td>varchar(n)</td>
<td>variable length string (at most n)</td>
</tr>
<tr>
<td>date</td>
<td>year/month/day</td>
</tr>
<tr>
<td>time</td>
<td>hh:mm:ss.ss</td>
</tr>
</tbody>
</table>
Sample Database Revisited

AUTHOR(aid integer, name char(20))

WROTE(author integer, publication char(8))

PUBLICATON(pubid char(8), title char(70))

BOOK(pubid char(8),
    publisher char(50), year integer)

JOURNAL(pubid char(8),
    volume integer, no integer, year integer)

PROCEEDINGS(pubid char(8),
    year integer)

ARTICLE(pubid char(8), crossref char(8),
    startpage integer, endpage integer)

... SQL is NOT case sensitive.
The Basic “SELECT Block”

Basic syntax:

3. SELECT DISTINCT <results>
1. FROM <tables>
2. WHERE <condition>

- Allows formulation of conjunctive (\(\exists, \land\)) queries of the form
  \[\{<\text{results}> | \exists<\text{unused}>. \left(\land <\text{tables}>\right) \land <\text{condition}>\}\]
  \(\Rightarrow\) a conjunction of <tables> with <condition>
  \(\Rightarrow\) <results> specifies values in the resulting tuples, and
  \(\Rightarrow\) <unused> are variables not used in <results>
Example

List all authors in the database:

```
SQL> select distinct *  
2   from author;

AID  NAME          
---  --------------
 1  Toman, David   
 2  Chomicki, Jan  
 3  Saake, Gunter
```

The `FROM` clause cannot be used on its own

⇒ the "SELECT *" notation
⇒ also reveals all attribute names
Variables vs. Attributes

- Relational Calculus uses positional notation, i.e.,
  \[ \text{EMP}(x, y, z) \] is true whenever the \( x \), \( y \), and \( z \) components of an answer can be found as a tuple in the instance of \( \text{EMP} \)
  \[ \Rightarrow \] no need for attribute names
  \[ \Rightarrow \] inconvenient for relations with high arity

- SQL uses coreslations (tuple variables) and attribute names to assign default variable names to components of tuples:
  \[ R \ [\text{AS}] p \] in SQL stands for \( R(p.a_1, \ldots p.a_n) \) in RC
  where \( a_1, \ldots, a_k \) are the attribute names declared for \( R \).
Example

List all publications with at least two authors,

$$\{p \mid \exists a_1, a_2. \text{WROTE}(a_1, p) \land \text{WROTE}(a_2, p) \land a_1 \neq a_2\} :$$

```
SQL> select distinct r1.publication
2   from wrote r1, wrote r2
3   where r1.publication = r2.publication
4   and r1.author != r2.author;
```

<table>
<thead>
<tr>
<th>PUBLICATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChSa98</td>
</tr>
<tr>
<td>ChTo98</td>
</tr>
<tr>
<td>ChTo98a</td>
</tr>
</tbody>
</table>

⇒ cannot share a variable (p) in the two WROTE relations
need for explicit equality “r1.publication = r2.publication”
List titles of all books,

\[ \{ t \mid \exists p, b, y. \text{PUBLICATION}(p, t) \land \text{BOOK}(p, b, y) \} \]:

```
SQL> select distinct title
    2    from publication, book
    3    where publication.pubid = book.pubid;
```

<table>
<thead>
<tr>
<th>TITLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logics for Databases and Information Systems</td>
</tr>
</tbody>
</table>

⇒ relations can serve as their own corelations when unambiguous

publication stands for “publication publication”, i.e.,

publication \( \text{publication}(\text{publication}.\text{pubid}, \text{publication}.\text{title}) \)
The "FROM" Clause (summary)

Syntax:

```
FROM  R_1 [ [AS] n_1 ], ..., R_k [ [AS] n_k ]
```

- \( R_i \) are relation (table) names
- \( n_i \) are distinct identifiers
- The clause represents a **conjunction** \( R_1 \land ... \land R_k \)
  - \( \Rightarrow \) all variables of \( R_i \)'s are **distinct**
  - \( \Rightarrow \) we use (co)relation names to resolve ambiguities
- Cannot appear alone
  - \( \Rightarrow \) only as a part of the **select block**
The "SELECT" Clause

Syntax:

\[
\text{SELECT DISTINCT } e_1 \ [\text{AS} \ n_1 ], \ldots, e_k \ [\text{AS} \ n_k ]
\]

1. Eliminate superfluous attributes from answers (∃)
2. Form expressions:
   \[\Rightarrow \] built-in functions applied to values of attributes
3. Give names to attributes in the answer
Standard Expressions

We can create values in the answer tuples using **built-in** functions:

- **On numeric types:**
  
  \[+,-,\ast,/,\ldots\] (usual arithmetic)

- **On strings:**
  
  \[\|\] (concatenation), `substr`, …

- **Constants** (of appropriate types)
  
  "SELECT 1" is a valid query in SQL-92

- **UDF** (user defined functions)

**Note:** all attribute names **MUST** be “present” in the **FROM** clause.
Example

For every article list the number of pages:

SQL> select distinct pubid, endpage-startpage+1
    2   from article;

<table>
<thead>
<tr>
<th>PUBID</th>
<th>ENDPAGE-STARTPAGE+1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChTo98</td>
<td>40</td>
</tr>
<tr>
<td>ChTo98a</td>
<td>28</td>
</tr>
<tr>
<td>Tom97</td>
<td>19</td>
</tr>
</tbody>
</table>
Naming Attributes in the Results

Results of queries \(\iff\) Tables

What are the names of attributes in the result of a `SELECT` clause?

- A single attribute: inherits the name
- An expression: implementation dependent

We can—and should—`explicitly` name the resulting attributes:

\[\Rightarrow "<\text{expr}> \text{ AS } <\text{id}>" \text{ where } <\text{id}> \text{ is the new name}\]
Example

For every article list the number of pages, and name the resulting attributes id, numberofpages:

```
SQL> select distinct pubid as id,
          2       endpage-startpage+1 as numberofpages
          3  from article;

   ID      NUMBEROFPAGES
---------- ------------
ChTo98      40
ChTo98a    28
Tom97       19
```
The "WHERE" Clause

Syntax:

```
WHERE <condition>
```

Additional conditions on tuples that qualify for the answer.

- Standard atomic conditions:
  1. equality: =, != (on all types)
  2. order: <, <=, >, >=, <> (on numeric and string types)

- Conditions may involve *expressions*
  ⇒ similar conditions as in the **SELECT** clause
Example(s)

Find all journals printed since 1997:

```
SQL> select * from journal where year >= 1997;

PUBID    VOLUME    NO    YEAR
-------- ---------- ---------- ----------
JLP-3-98  35        3      1998
```

Find all articles with more than 20 pages:

```
SQL> select * from article
2       where endpage-startpage > 20;

PUBID    CROSSREF    STARTPAGE    ENDPAGE
-------- ----------- ---------- ----------
ChTo98   ChSa98      31          70
ChTo98a  JLP-3-98    263         290
```
Boolean Connectives

Atomic conditions can be combined using **boolean connectives**:

- **AND** (conjunction)
- **OR** (disjunction)
- **NOT** (negation)
Example

List all publications with at least two authors:

```sql
SQL> select distinct r1.publication
2  from wrote r1, wrote r2
3  where r1.publication = r2.publication
4  and not r1.author = r2.author;

PUBLICATION
------------
ChSa98
ChTo98
ChTo98a
```
Summary

- Simple SELECT block accounts for many queries
  ⇒ all in ∃, ∧ fragment of relational calculus

- Additional features
  - alternative names for relations
  - expressions and naming in the output
  - built-in atomic predicates and boolean connectives

- Well defined semantics (declarative and operational)
Complex Queries in SQL

- So far we can write only $\exists, \land$ queries
  - ⇒ the SELECT BLOCK queries
  - ⇒ not sufficient to cover all safe RC queries

- Remaining connectives:
  1. $\lor, \neg$: are expressed using set operations
     - ⇒ easy to enforce range-restriction requirements
  2. $\forall$: rewrite using negation and $\exists$
     - ⇒ the same for $\rightarrow$, $\leftrightarrow$, etc.
Set Operations at Glance

Answers to Select Blocks are **relations** (sets of tuples)

⇒ we can apply **set operations** on them

- **Set union:** \( Q_1 \text{ UNION } Q_2 \)
  ⇒ the set of tuples in \( Q_1 \) or in \( Q_2 \).
  ⇒ used to express “or”.

- **Set difference:** \( Q_1 \text{ EXCEPT } Q_2 \)
  ⇒ the set of tuples in \( Q_1 \) but not in \( Q_2 \).
  ⇒ used to express “and not”.

- **Set intersection:** \( Q_1 \text{ INTERSECT } Q_2 \)
  ⇒ the set of tuples in both \( Q_1 \) and \( Q_2 \).
  ⇒ used to express “and” (redundant, rarely used).

\( Q_1 \) and \( Q_2 \) must have **union-compatible** signatures:

⇒ same number and types of attributes
Example: Union

List all publication ids for books or journals:

```sql
SQL> (select distinct pubid from book)
2   union
3   (select distinct pubid from journal);

<table>
<thead>
<tr>
<th>PUBID</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChSa98</td>
</tr>
<tr>
<td>JLP-3-98</td>
</tr>
</tbody>
</table>
```
Example: Set Difference

List all publication ids except those for articles:

```
SQL> (select distinct pubid from publication)
  2    except
  3  (select distinct pubid from article);
```

<table>
<thead>
<tr>
<th>PUBID</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChSa98</td>
</tr>
<tr>
<td>DOOD97</td>
</tr>
<tr>
<td>JLP-3-98</td>
</tr>
</tbody>
</table>

(University of Waterloo)
What About Nesting of Queries?

We can use SELECT Blocks (and other set operations) as arguments of set operations.

What if we need to use a set operation inside of a SELECT Block?

- Can use **distributive laws**
  \[(A \lor B) \land C \equiv (A \land C) \lor (B \land C)\]
  \[(A \land C) \lor (B \land C)\]
  ⇒ often **very** cumbersome

- Nest set operation inside a select block.
  ⇒ **common table expressions**
Naming (Sub-)queries

Idea:
Queries denote **relations**. We provide a **naming** mechanism that allows us to assign names to (results of) queries.

⇒ can be used later in place of (base) relations.

Syntax:

```
WITH fool [opt-schema-1] AS ( <query-1-goes-here> ),
    ...
foon [opt-schema-n] AS ( <query-n-goes-here> )
<query-that-uses-fool-...-foon-as-table-names>
```
Example

List all publication titles for books or journals:

```sql
SQL> with bookorjournal(pubid) as
2   ( (select distinct pubid from book)
3     union
4     (select distinct pubid from journal)
5   )
6  select distinct title
7  from publication, bookorjournal
8  where publication.pubid = bookorjournal.pubid;

TITLE
-----------------------------
Logics for Databases and Information Systems
Journal of Logic Programming
```
The **FROM** clause revisited

- Using the **WITH** mechanism is sometimes cumbersome:
  - ⇒ we don’t want to name every subexpression

- SQL-92 allows us to **inline** queries in the **FROM** clause:
  - FROM ..., ( <query-here> ) <id>,...
  - ⇒ <id> stands for the result of <query-here>.
  - ⇒ unlike for base relations, <id> is **mandatory**.

- In “old” SQL (SQL-89) this does NOT work; *views* were the only option.
Example

List all publication titles for journals or books:

```sql
SQL> select distinct title
2   from publication,
3   ( (select distinct pubid from journal)
4      union
5     (select distinct pubid from book) ) jb
6   where publication.pubid = jb.pubid;

TITLE
------------------------
Logics for Databases and Information Systems
Journal of Logic Programming
```
Can’t we just use OR instead of UNION?

- A common mistake:
  ⇒ use of OR in the WHERE clause instead of the UNION operator

- An incorrect solution:
  
  ```sql
  select distinct title
  from publication, book, journal
  where publication.pubid = book.pubid
  or publication.pubid = journal.pubid
  ```

- Often works, but consider where there are no books.
Summary on First-Order SQL

- SQL introduced so far captures all of RC (relational calculus)
  - optionally with duplicate semantics
  - powerful (many queries can be expressed)
  - efficient (PTIME, LOGSPACE)

- Shortcomings:
  - some queries are hard to write (syntactic sugar)
  - no “counting” (aggregation)
  - no “path in graph” (recursion)
WHERE Subqueries

- Additional (complex) search conditions
  ⇒ query-based search predicates

- Advantages
  - simplifies writing queries with negation

- Drawbacks
  - complicated semantics
    (especially when duplicates are involved)
  - very easy to make mistakes

- VERY COMMONLY used to formulate queries
Overview of **WHERE** Subqueries

- Presence/absence of a *single* value in a query
  
  \[ \text{<attr> IN ( <query> )} \]
  
  \[ \text{<attr> NOT IN ( <query> )} \]

- Relationship of a value to some/all values in a query
  
  \[ \text{<attr> op SOME ( <query> )} \]
  
  \[ \text{<attr> op ALL ( <query> )} \]

- Emptiness/non-emptiness of a query
  
  \[ \text{EXISTS ( <query> )} \]
  
  \[ \text{NOT EXISTS ( <query> )} \]

In the first two cases \(<query>\) must be unary.
Example: “<attr> in (<query>)”

```
SQL> select distinct title
    2  from publication
    3  where pubid in (select pubid from article);
```

<table>
<thead>
<tr>
<th>TITLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal Logic in Information Systems</td>
</tr>
<tr>
<td>Datalog with Integer Periodicity Constraints</td>
</tr>
<tr>
<td>Point-Based Temporal Extension of Temporal SQL</td>
</tr>
</tbody>
</table>
“Pure” SQL Equivalence

Nesting in the WHERE clause is mere syntactic sugar:

```sql
select r.b
from r
where r.a in ( select distinct b
    select b
    from s
) as s
where r.a = s.b
```

All of the remaining constructs can be rewritten in similar fashion.
Example: “<attr> not in (<query>)”

All author-publication ids for all publications except books and journals:

```sql
SQL> select * 
  2  from wrote 
  3  where publication not in ( 
  4     ( select pubid from book ) 
  5     union 
  6     ( select pubid from journal ) );
```

<table>
<thead>
<tr>
<th>AUTHOR</th>
<th>PUBLICAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ChTo98</td>
</tr>
<tr>
<td>1</td>
<td>ChTo98a</td>
</tr>
<tr>
<td>1</td>
<td>Tom97</td>
</tr>
<tr>
<td>2</td>
<td>ChTo98</td>
</tr>
<tr>
<td>2</td>
<td>ChTo98a</td>
</tr>
</tbody>
</table>

...search conditions may contain complex queries.
"<attr> not in (<query>)" (cont.)

...another formulation:

```sql
SQL> select *
   2  from wrote
   3  where publication not in ( 
   4   select pubid from book
   5  ) and publication not in ( 
   6   select pubid from journal
   7  )
```

...and may be combined using boolean connectives.
Example: “<attr> op SOME/ALL (<query>)”

Find the longest articles (a way expressing max):

```
SQL> select distinct pubid
  2  from article
  3  where endpage-startpage >= all (  
  4    select endpage-startpage 
  5    from  article 
  6  );

PUBID
------
ChTo98
```

“<attr> = some (<query>)” ≡ “<attr> in (<query>)”
“<attr> <> all (<query>)” ≡ “<attr> not in (<query>)”
Parametric Subqueries

- So far, subqueries were independent of the main query
  ⇒ not correlated
  ⇒ not much fun (good only for simple queries)

- SQL allows parametric (correlated) subqueries.

Parametric subqueries have the form “<query>” mentioning

<attr>_1,<attr>_2,...

where <attr>_i is an attribute in the main query.

The truth of a predicate defined by a subquery is determined for each substitution (tuple) in the main query:

1. instantiate all the parameters and
2. check for the truth value as before . . .
Example: “EXISTS (<query>)”

Parametric subqueries are most common for “existential” subqueries:

```
SQL> select *  
  2  from wrote r  
  3  where exists ( select *  
  4       from wrote s  
  5  where r.publication = s.publication  
  6   and r.author <> s.author );

    AUTHOR   PUBLICAT
--------- --------
     1 ChTo98
     1 ChTo98a
     2 ChTo98
     2 ChTo98a
     2 ChSa98
     3 ChSa98
```
Example: “NOT EXISTS (<query>)”

It is easy to now complement conditions:

```sql
SQL> select *
2   from wrote r
3   where not exists ( 
4       select *
5       from wrote s
6       where r.publication = s.publication
7       and r.author <> s.author
8   );

    AUTHOR    PUBLICAT
----------- --------
     1  Tom97
```
Example: “<attr> IN (<query>)”

```
SQL> select *
2     from wrote r
3     where publication in ( 
4         select publication 
5         from wrote s 
6         where r.author <> s.author 
7     );

  AUTHOR  PUBLICAT
----------  -------
         1  ChTo98
         1  ChTo98a
         2  ChTo98
         2  ChTo98a
         2  ChSa98
         3  ChSa98
```
More levels of Nesting

- WHERE subqueries are **just queries**
  - we can nest again and again and ... 
  - every nested subquery can use attributes from the enclosing queries as parameters.
  - correct naming is imperative

- Used to formulate very complex **search conditions**
  - attributes present **in the subquery only**
    - CANNOT be used to construct the result(s).
Example

List all authors who always publish with someone else:

```sql
SQL> select distinct a1.name
2   from author a1, author a2
3   where not exists (
4       select *
5           from publication p, wrote w1
6           where p.pubid = w1.publication
7           and a1.aid = w1.author
8       and a2.aid not in (  
9           select author  
10              from wrote  
11              where publication = p.pubid  
12              and author <> a1.aid  
13        )
14 );
```
Summary

- WHERE subqueries enable easy formulation of queries of the form “All x in R such that (a part of) x doesn’t appear in S”.
- Subqueries only stand for **WHERE conditions** ⇒ **CANNOT** be used to produce results.
- You can use **input parameters**, but these must be **bound** in the main query.
- All of these are just a syntactic sugar and can be expressed using queries nested in the **FROM** clause.
  - but it might be quite hard . . .
  - and it is easy to make mistakes (be **very** careful)
How do we Modify a Database?

- Naive approach:
  
  $DBSTART$;
  $r_1 := Q_1(DB);$  
  ...  
  $r_k := Q_k(DB);$  
  $DBCOMMIT$;

- Not an acceptable solution in practice
Incremental Updates

Idea

Tables are large but **updates are small** ⇒ Incremental updates

1. Insertion of a tuples (**INSERT**)
   ⇒ constant tuple
   ⇒ results of queries

2. Deletion of tuples (**DELETE**)
   ⇒ based on match of a condition

3. Modification of tuples (**UPDATE**)
   ⇒ allows updating “in place”
   ⇒ based on match of a condition
SQL Insert

- One constant tuple (or a fixed number):

  \[
  \text{INSERT INTO } r[(a_1, \ldots, a_k)] \\
  \text{VALUES } (v_1, \ldots, v_k)
  \]

  \[\Rightarrow\] adds tuples \((v_1, \ldots, v_k)\) to \(r\).
  \[\Rightarrow\] the type of \((v_1, \ldots, v_k)\) must match the schema definition of \(r\).

- Multiple tuples (generated by a query):

  \[
  \text{INSERT INTO } r (Q)
  \]

  \[\Rightarrow\] adds result of \(Q\) to \(r\)
Example: insertion of a tuple

Add a new author:

SQL> insert into author
2    values (4, 'Niwinski, Damian',
3    'zls.mimuw.edu.pl/~niwinski');

1 row created.

SQL> select distinct aid, name, url from author;

<table>
<thead>
<tr>
<th>AID</th>
<th>NAME</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Toman, David</td>
<td>db.uwaterloo.ca/~david</td>
</tr>
<tr>
<td>2</td>
<td>Chomicki, Jan</td>
<td>cs.monmouth.edu/~chomicki</td>
</tr>
<tr>
<td>3</td>
<td>Saake, Gunter</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Niwinski, Damian</td>
<td>zls.mimuw.edu.pl/~niwinski</td>
</tr>
</tbody>
</table>
Example: use of a query

Add a new author (without looking up author id):

```sql
SQL> insert into author (  
2         select max(aid)+1, 'Snodgrass, Richard T.',  
3               'www.cs.arizona.edu/people/rts'  
4         from author );
1 row created.

SQL> select distinct aid, name from author;

AID NAME
--- ---------------
1 Toman, David   
2 Chomicki, Jan  
3 Saake, Gunter  
4 Damian Niwinski
5 Snodgrass, Richard T.
```
SQL Delete

- Deletion using a condition:
  
  ```sql
  DELETE FROM r
  WHERE cond
  
  ⇒ deletes all tuples that match cond.
  ```

- Deletion using cursors (later)
  
  ⇒ available in embedded SQL
  ⇒ only way to delete one out of two duplicate tuples
Delete all publications that are not articles or the collections an article appears in:

```sql
SQL> delete from publication
    2   where pubid not in (  
    3     select pubid  
    4     from article  
    5   ) and pubid not in (  
    6     select crossref  
    7     from article  
    8   );

0 rows deleted.
```
SQL Update

- Two components:
  1. an update statement (*SET*)
     ⇒ an assignment of values to attributes.
  2. a search condition (*WHERE*)

- Syntax:

```
UPDATE r
SET <update statement>
WHERE <condition>
```
Example

```sql
SQL> update author
    2  set url = 'brics.dk/~david'
    3  where aid in ( 
    4      select aid
    5      from author
    6      where name like 'Toman%'
    7    )
;
1 row updated.

SQL> select * from author;

AID  NAME          URL
---  -------------- ---------------------
 1    Toman, David  //brics.dk/~david
...
```
Support for Transactions

The DBMS guarantees noninterference (serializability) of all data access requests to tables in a database instance.

- **Transaction starts with first access of the database**
  - ⇒ until it sees:
- **COMMIT:** make changes permanent
  
  SQL> commit;
  Commit complete.
- **or ROLLBACK:** discard changes
  
  SQL> rollback;
  Rollback complete.
Aggregation

Standard and very useful extension of First-Order Queries.

- Aggregate (column) functions are introduced to
  - find number of tuples in a relation
  - add values of an attribute (over the whole relation)
  - find minimal/maximal values of an attribute

- Can apply to groups of tuples that have equal values for (selected) attributes

- Can NOT be expressed in Relational Calculus
Aggregation in SQL

The same in SQL syntax:

```
SELECT  x1, ..., xk, aggl, ..., aggl
FROM    Q
GROUP BY x1, ..., xk
```

Restrictions:

- All attributes in the SELECT clause that are NOT in the scope of an aggregate function MUST appear in the GROUP BY clause.
- aggl are of the form `count(*)`, `count(<expr>)`, `sum(<expr>)`, `min(<expr>)`, `max(<expr>)`, or `avg(<expr>)` where `<expr>` is usually an attribute of Q (and usually not in the GROUP BY clause).
Operational Reading

1. Partition the input relation to groups with equal values of **grouping** attributes
2. On each of these partitions apply the **aggregate function**
3. Collect the results and form the answer
Example (count)

For each publication, count the number of authors:

```
SQL> select publication, count(author)  
     2   from wrote  
     3   group by publication;

PUBLICAT COUNT(AUTHOR)  
-------- --------------
ChSa98    2           
ChTo98    2           
ChTo98a   2           
Tom97     1
```

(University of Waterloo)
Example (sum)

For each author, count the number of article pages:

```
SQL> select author, sum(endpage-startpage+1) as pgs
  2   from wrote, article
  3   where publication=pubid
  4   group by author;

<table>
<thead>
<tr>
<th>AUTHOR</th>
<th>PGS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>87</td>
</tr>
<tr>
<td>2</td>
<td>68</td>
</tr>
</tbody>
</table>

... not quite correct: it doesn’t list 0 pages for author 3.
The **HAVING** clause

- The **WHERE** clause can’t impose conditions on values of aggregates
  \[ \Rightarrow \text{WHERE clause has to be used \textbf{before} \textsc{GROUP BY}} \]

- SQL allows a **HAVING** clause instead
  \[ \Rightarrow \text{like WHERE, but for aggregate values...} \]

- The aggregate functions used in the **HAVING** clause may be different from those in the **SELECT** clause; the grouping, however, is common.

---

The **HAVING** clause is mere **SYNTACTIC SUGAR**

... and can be replaced by a nested query and a **WHERE** clause.
Example

List publications with exactly one author:

```
SQL> select publication, count(author)
    2    from wrote
    3  group by publication
    4  having count(author) = 1;

PUBLICAT  COUNT(AUTHOR)
---------  ---------
Tom97        1
```

... This query *can* be written without aggregation.
Example (revisited.)

For every author, count the number of books and articles:

SQL> select distinct aid, name, count(publication) 
    2     from author, ( 
    3     ( select distinct author, publication 
    4         from wrote, book 
    5         where publication = pubid ) 
    6     union all 
    7     ( select distinct author, publication 
    8         from wrote, article 
    9         where publication = pubid ) ) ba 
10    where aid = author 
11    group by name, aid;
Summary

- SQL covered so far:
  1. Simple SELECT BLOCK
  2. Set operations
  3. Formulation of complex queries, nesting of queries, and views
  4. Updating Data
  5. Aggregation

- This covers pretty much all of the useful SQL DML
  ⇒ the Bad and Ugly coming next . . .