**Similarity Group-by (SGB)**

- SGB extends the standard grouping operator to group similar or approximate values.
- The main goal of SGB is to generate more meaningful and useful similarity-based groupings than those of the regular group-by while maintaining:
  - Low running time
  - Good scalability properties
  - Efficient integration with the query processing engine.

**SGB: Three Instances**

**Unsupervised SGB**

- SELECT select_ops, FROM table, WHERE GROUP BY col_name [MAXIMUM_GROUP_DIAMETER d], ...
- No extra data provided to guide the process. Classes control group size and group compactness.

**Supervised Similarity Group Around**

- SELECT select_ops, FROM table, WHERE GROUP BY col_name NAME central-points [MAXIMUM_GROUP_DIAMETER d], ...
- Groups tuples around a set of guiding points. Each tuple is assigned to the group of its closest central point.
- Classes control group size and group compactness.

**Supervised Similarity Group with Delimiters**

- SELECT Arg(Temperature), Arg(Pressure) FROM SensorsReadings GROUP BY Temperature DELIMITED BY SELECT value and Thresholds, Pressure DELIMITED BY ([SILS]), MAXIMUM_ELEMENT_SEPARATION 3
- Forms groups based on a set of delimiting points.

**Exploiting SGB in Decision Support System Dashboards**

- Studying groups of large-volume customers with similar buying power.
- Studying profit of a line of parts around marketing campaigns.
- Studying groups of orders around revenue levels of interest.

**DSS Architecture**

- Parser: Extended the grammar rules. Extended the parse trees and query-free structures.
- Planner/Optimizer: Made use of the RMS input plan tree of aggregation nodes to process the reference points. Each internal aggregation node processes 1 SGA and 1 or more GAS SGAs can be ordered to reduce number of flowing tuples.
- The executor: Hash-based approach used to maintain the formed groups. Single plane sweeping approach used to form the groups. The tuples to be grouped and the reference points are processed simultaneously. Data tuples and reference points are sorted before being processed by the aggregation node.

**SGB Implementation (PostgreSQL)**

- PostgreSQL
- Decision Support Dashboards
- Central points
- Data points
- Sweeping Plane
- Group 1
- Group 2
- Performance Evaluation (TPC-H)
- Performance of generating similarity groups with GB vs. SGB
- Performance while increasing dataset size
- Performance while increasing SGAs