TECH VAULT

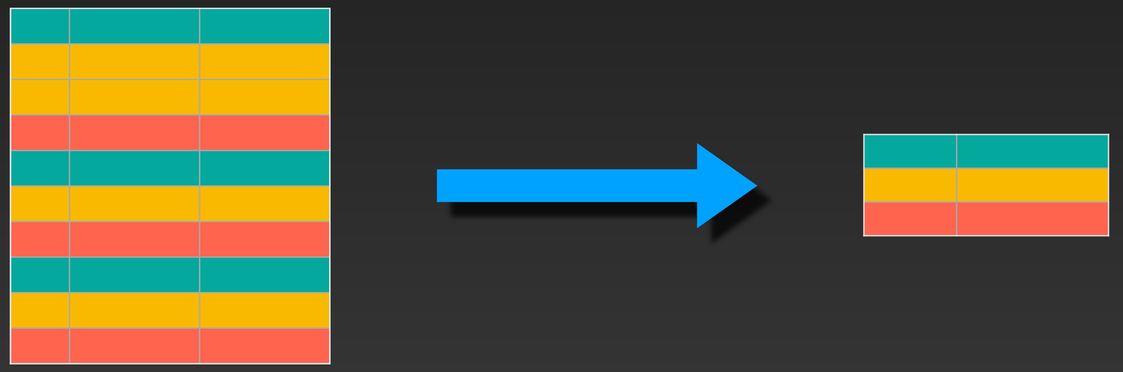
AGGREGATION ALGORITHMS

- SORT-BASED
- HASH-BASED

AMR ELHELW

Aggregation

Combining and summarizing multiple data records into a single result set based on some criteria.



Example

Total Sales Amount by Product Category

Sales

id	date	item	category	amount
101	2022-04-10	TV	Electronics	$800
102	2023-05-07	Chair	Furniture	$120
103	2021-10-04	Blender	Kitchen	$350
104	2021-10-04	Table	Furniture	$300
105	2023-06-20	Coffee Maker	Kitchen	$120
106	2022-07-25	Printer	Electronics	$80

```
SELECT category, SUM(amount) AS total
FROM Sales
GROUP BY category
```

category	total
Electronics	$880
Furniture	$420
Kitchen	$470

```
SELECT SUM(amount) AS total
FROM Sales
```

Total Sales Amount

total
$1770

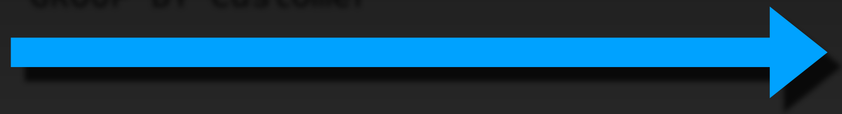
Example

Number of Orders per customer

OrderInfo

id	date	customer
221	2021-05-10	Bob
222	2023-05-07	Alice
223	2021-09-03	Bob
224	2018-11-24	John
225	2023-06-20	Alice
226	2022-07-25	Alice

```
SELECT customer, COUNT(*) AS num_orders
FROM OrderInfo
GROUP BY customer
```



customer	num_orders
Bob	2
Alice	3
John	1

Example

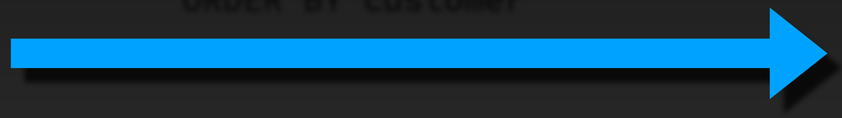
Sorted list of customers

OrderInfo

id	date	customer
221	2021-05-10	Bob
222	2023-05-07	Alice
223	2021-09-03	Bob
224	2018-11-24	John
225	2023-06-20	Alice
226	2022-07-25	Alice

```
SELECT customer
FROM OrderInfo
GROUP BY customer
ORDER BY customer
```

```
SELECT DISTINCT customer
FROM OrderInfo
ORDER BY customer
```



customer
Alice
Bob
John

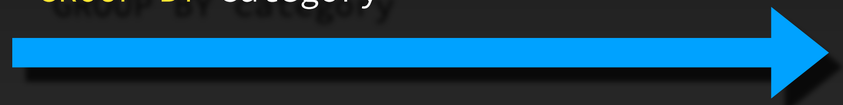
Example

Total Sales Amount by Product Category
for the year 2022

Sales

id	date	item	category	amount
101	2022-04-10	TV	Electronics	$800
102	2022-05-07	Chair	Furniture	$120
~~103~~	~~2021-10-04~~	~~Blender~~	~~Kitchen~~	~~$350~~
104	2022-10-04	Table	Furniture	$300
~~105~~	~~2023-06-20~~	~~Coffee Maker~~	~~Kitchen~~	~~$120~~
106	2022-07-25	Printer	Electronics	$80

```
SELECT category, SUM(amount) AS total
FROM Sales
WHERE EXTRACT(YEAR FROM date) = 2022
GROUP BY category
```



category	total
Electronics	$880
Furniture	$420

Aggregation Steps

- Split input rows into groups
  - Only if there is a GROUP BY or DISTINCT
  - Otherwise, treat everything as 1 big group
- Combine values within each group
  - May use functions: COUNT, MIN, MAX, SUM, AVG, etc.
- Output one row per group

Example

Total Sales Amount by Product Category

Sales

id	date	item	category	amount
101	2022-04-10	TV	Electronics	$800
102	2023-05-07	Chair	Furniture	$120
103	2021-10-04	Blender	Kitchen	$350
104	2021-10-04	Table	Furniture	$300
105	2023-06-20	Coffee Maker	Kitchen	$120
106	2022-07-25	Printer	Electronics	$80

id	date	item	category	amount
102	2023-05-07	Chair	Furniture	$120
104	2021-10-04	Table	Furniture	$300
101	2022-04-10	TV	Electronics	$800
106	2022-07-25	Printer	Electronics	$80
103	2021-10-04	Blender	Kitchen	$350
105	2023-06-20	Coffee Maker	Kitchen	$120

category	total
Furniture	$420
Electronics	$880
Kitchen	$470

```
SELECT category, SUM(amount) AS total
FROM Sales
GROUP BY category
```

Aggregate function

Grouping Key

Amr Elhelw's TECH VAULT

Example

List all categories (duplicate elimination)

Sales

id	date	item	category	amount
101	2022-04-10	TV	Electronics	$800
102	2023-05-07	Chair	Furniture	$120
103	2021-10-04	Blender	Kitchen	$350
104	2021-10-04	Table	Furniture	$300
105	2023-06-20	Coffee Maker	Kitchen	$120
106	2022-07-25	Printer	Electronics	$80

id	date	item	category	amount
102	2023-05-07	Chair	Furniture	$120
104	2021-10-04	Table	Furniture	$300
101	2022-04-10	TV	Electronics	$800
106	2022-07-25	Printer	Electronics	$80
103	2021-10-04	Blender	Kitchen	$350
105	2023-06-20	Coffee Maker	Kitchen	$120

category
Furniture
Electronics
Kitchen

```
SELECT category
FROM Sales
GROUP BY category
```

⟷

```
SELECT DISTINCT category
FROM Sales
```

Grouping Key

Sort-based Aggregation

optimizer uses it if the grouping key is already sorted based on this key through:
index on that key or
data comes from merge join on the same key(because is sorts the data based on this key)

* the output will be sorted

Sort-based Aggregation

List all categories (duplicate elimination)

Sales

id	date	item	category	amount
101	2022-04-10	TV	Electronics	$800
102	2023-05-07	Chair	Furniture	$120
103	2021-10-04	Blender	Kitchen	$350
104	2021-10-04	Table	Furniture	$300
105	2023-06-20	Coffee Maker	Kitchen	$120
106	2022-07-25	Printer	Electronics	$80

Sort by Grouping Key

id	date	item	category	amount
106	2022-07-25	Printer	Electronics	$80
101	2022-04-10	TV	Electronics	$800
104	2021-10-04	Table	Furniture	$300
102	2023-05-07	Chair	Furniture	$120
103	2021-10-04	Blender	Kitchen	$350
105	2023-06-20	Coffee Maker	Kitchen	$120

Result

category
Electronics
Furniture
Kitchen

```
SELECT category
FROM Sales
GROUP BY category
```

```
SELECT DISTINCT category
FROM Sales
```

Sort-based Aggregation

Total Sales Amount by Product Category

Sales

id	date	item	category	amount
101	2022-04-10	TV	Electronics	$800
102	2023-05-07	Chair	Furniture	$120
103	2021-10-04	Blender	Kitchen	$350
104	2021-10-04	Table	Furniture	$300
105	2023-06-20	Coffee Maker	Kitchen	$120
106	2022-07-25	Printer	Electronics	$80

Sort by Grouping Key

id	date	item	category	amount
106	2022-07-25	Printer	Electronics	$80
101	2022-04-10	TV	Electronics	$800
104	2021-10-04	Table	Furniture	$300
102	2023-05-07	Chair	Furniture	$120
103	2021-10-04	Blender	Kitchen	$350
105	2023-06-20	Coffee Maker	Kitchen	$120

Result

category	total
Electronics	$880
Furniture	$420
Kitchen	$470

```
SELECT category, SUM(amount) AS total
FROM Sales
GROUP BY category
```

Key	Electronics
Sum	80

Aggregate functions

Function	How to update?
SUM	Add new value for each row
COUNT	Add 1 for each row
MAX	Use new value if greater than previous max
MIN	Use new value if less than previous min
AVERAGE	Keep both sum, count Compute sum/count at the end of each group

Hash-based Aggregation

Hash-based aggregation

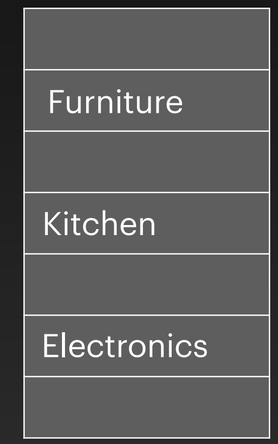
List all categories (duplicate elimination)

Sales

id	date	item	category	amount
101	2022-04-10	TV	Electronics	$800
102	2023-05-07	Chair	Furniture	$120
103	2021-10-04	Blender	Kitchen	$350
104	2021-10-04	Table	Furniture	$300
105	2023-06-20	Coffee Maker	Kitchen	$120
106	2022-07-25	Printer	Electronics	$80

Hash Function

Hash Table



Result

category
Furniture
Kitchen
Electronics

```
SELECT category
FROM Sales
GROUP BY category
```

⟺

```
SELECT DISTINCT category
FROM Sales
```

Hash-based aggregation

Average Sales Amount by Product Category

Sales

id	date	item	category	amount
101	2022-04-10	TV	Electronics	$800
102	2023-05-07	Chair	Furniture	$120
103	2021-10-04	Blender	Kitchen	$350
104	2021-10-04	Table	Furniture	$300
105	2023-06-20	Coffee Maker	Kitchen	$120
106	2022-07-25	Printer	Electronics	$80

Hash Function

Hash Table

Furniture	Sum: 420, Count: 2
Kitchen	Sum: 470, Count: 2
Electronics	Sum: 880, Count: 2

Result

category	Average
Furniture	$210
Kitchen	$235
Electronics	$440

```
SELECT category, AVG(amount) AS average
FROM Sales
GROUP BY category
```