Data Bhau

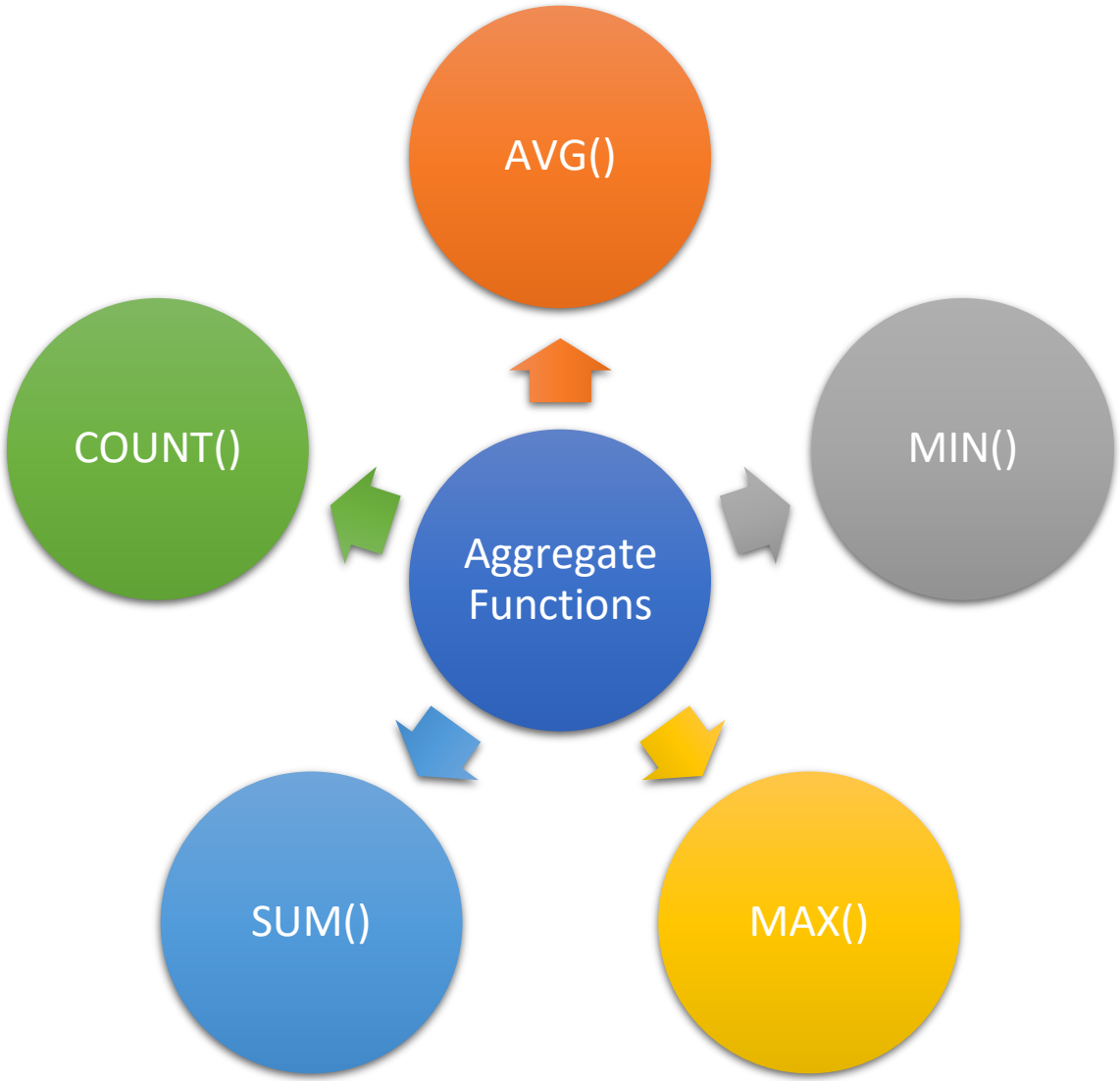
Class 2
Aggregate and Sorting

What is Aggregate Function

An aggregate function in SQL returns one value after performing a specific calculation on multiple values of a particular column.

- Aggregate functions in SQL —- avg, count, sum, min, max, etc.
- All aggregate functions ignore NULL values when it performs the calculation, except for the count function.
- We have to use aggregate functions with the GROUP BY and HAVING clauses of the SELECT statement

Aggregate Functions in SQL



MIN()

Query Execution

```
Select min(distance) from tutorial.flights
```

BEFORE EXECUTION

flight_number	distance	destination_city
2454	1919	Atlanta
2456	1069	Fort Lauderdale
2459	1035	West Palm Beach/Palm Bea...
2462	2079	Detroit
2464	1005	New York
2466	1250	Salt Lake City
2471	1182	New Orleans
2474	1598	San Juan
2476	1481	Detroit
2479	1182	New Orleans
2480	1235	Dallas/Fort Worth

AFTER EXECUTION

min
31

AVG()

Query Execution

```
Select avg(distance) from tutorial.flights
```

BEFORE EXECUTION

flight_number	distance	destination_city
2454	1919	Atlanta
2456	1069	Fort Lauderdale
2459	1035	West Palm Beach/Palm Bea...
2462	2079	Detroit
2464	1005	New York
2466	1250	Salt Lake City
2471	1182	New Orleans
2474	1598	San Juan
2476	1481	Detroit
2479	1182	New Orleans
2480	1235	Dallas/Fort Worth

AFTER EXECUTION

avg
808.3815

COUNT()

Query Execution

```
Select count(flight_number) from tutorial.flights
```

BEFORE EXECUTION

flight_number	distance	destination_city
2454	1919	Atlanta
2456	1069	Fort Lauderdale
2459	1035	West Palm Beach/Palm Bea...
2462	2079	Detroit
2464	1005	New York
2466	1250	Salt Lake City
2471	1182	New Orleans
2474	1598	San Juan
2476	1481	Detroit
2479	1182	New Orleans
2480	1235	Dallas/Fort Worth

AFTER EXECUTION

count
593842

GROUP BY CLAUSE

The GROUP BY clause is a SQL command that is used to group rows that have the same values.

- It is used to summarize data from the database.
- The queries that contain the GROUP BY clause are called grouped queries and only return a single row for every grouped item.
- GROUP BY clause applied on column(s) can be used to get unique records for those columns
- Syntax: select column_1, avg(column_2) from table_name Group by column_1

GROUP BY WITH A SINGLE COLUMN

Query Execution

```
select airline_code,airline_name from tutorial.flights
group by 1,2;
```

BEFORE EXECUTION

Source

air_traffic_delay	airline_code	airline_name	arrival_delay	cancellation_reason
0	DL	Delta Air Lines In...	-17	
0	DL	Delta Air Lines In...	-17	
0	DL	Delta Air Lines In...	-20	
0	DL	Delta Air Lines In...	3	
0	DL	Delta Air Lines In...	-25	
0	DL	Delta Air Lines In...	-17	
0	DL	Delta Air Lines In...	-21	
0	DL	Delta Air Lines In...	-9	
0	DL	Delta Air Lines In...	-9	
0	DL	Delta Air Lines In...	-30	
0	DL	Delta Air Lines In...	-9	

Query Executed

AFTER EXECUTION

airline_code	airline_name
G4	Allegiant Air
AS	Alaska Airlines Inc.
UA	United Air Lines Inc.
YX	Republic Airline
YV	Mesa Airlines Inc.
DL	Delta Air Lines Inc.
AA	American Airlines Inc.
EV	ExpressJet Airlines LLC
NK	Spirit Air Lines

GROUP BY CLAUSE WITH AGGREGATE FUNCTIONS

SQL Aggregate functions aggregate data across the entire column/dataset. However, if we want to aggregate the data at a certain part of table or at certain unique values of a column, it is achieved by using aggregate function with Group By clause.

Typical Structure of an Aggregate Function with Group By:

```
SELECT column_1, group/aggregate function(column_2)
FROM table_name
WHERE condition
GROUP BY column_1
```

GROUP BY WITH AGGREGATE FUNCTION - EXAMPLE

Query Execution

```
select state, count(city) from tutorial.city_populations group by state
```

BEFORE EXECUTION

city	state	population_estimate_2012	id
New York	NY	8336697	1
Los Angeles	CA	3857799	2
Chicago	IL	2714856	3
Houston	TX	2160821	4
Philadelphia	PA	1547607	5
Phoenix	AZ	1488750	6
San Antonio	TX	1382951	7
San Diego	CA	1338348	8
Dallas	TX	1241162	9
San Jose	CA	982765	10
Austin	TX	842592	11
Jacksonville	FL	836507	12
Indianapolis	IA	834852	13
San Francisco	CA	825863	14

AFTER EXECUTION

state	count
NV	1
OH	2
NY	1
NE	1
FL	2
MI	2
CA	8
OR	1
TX	7
KY	1
NM	1
DC	1

FILTERING DATA WITH AGGREGATE FUNCTION

As we learnt about grouping data within rows, we might be tempted to filter certain groups from our data.

```
SELECT column_1, COUNT(columns_2) AS sales_per_day
FROM sales
WHERE count(column_2) > 1
GROUP BY column_1
```

THIS PIECE OF CODE IS NOT GOING TO WORK , BUT WHY ?

WHY IS THERE AN ERROR?

Aggregate functions are not allowed in the WHERE clause because the WHERE clause is evaluated before the GROUP BY clause—there aren't any groups yet to perform calculations on.

But, there is a type of clause that allows us to filter, perform aggregations, and it is evaluated after the GROUP BY clause: the HAVING clause.

The HAVING clause is like a WHERE clause for your groups.

HAVING CLAUSE

The HAVING clause was added to SQL because the WHERE keyword cannot be used with aggregate functions.

```
SELECT column_1,sum(column_2)
FROM table_name
WHERE condition
GROUP BY column 1
HAVING sum(column_2) > 30
```

HAVING CLAUSE - EXAMPLE

Employee

EmployeeID	Ename	DeptID	Salary
1001	John	2	4000
1002	Anna	1	3500
1003	James	1	2500
1004	David	2	5000
1005	Mark	2	3000
1006	Steve	3	4500
1007	Alice	3	3500

```
SELECT DeptID, AVG(Salary)
FROM Employee
GROUP BY DeptID;
```

GROUP BY Employee Table using DeptID

DeptID	AVG(Salary)
1	3000.00
2	4000.00
3	4250.00

```
SELECT DeptID, AVG(Salary)
FROM Employee
GROUP BY DeptID
HAVING AVG(Salary) > 3000;
```

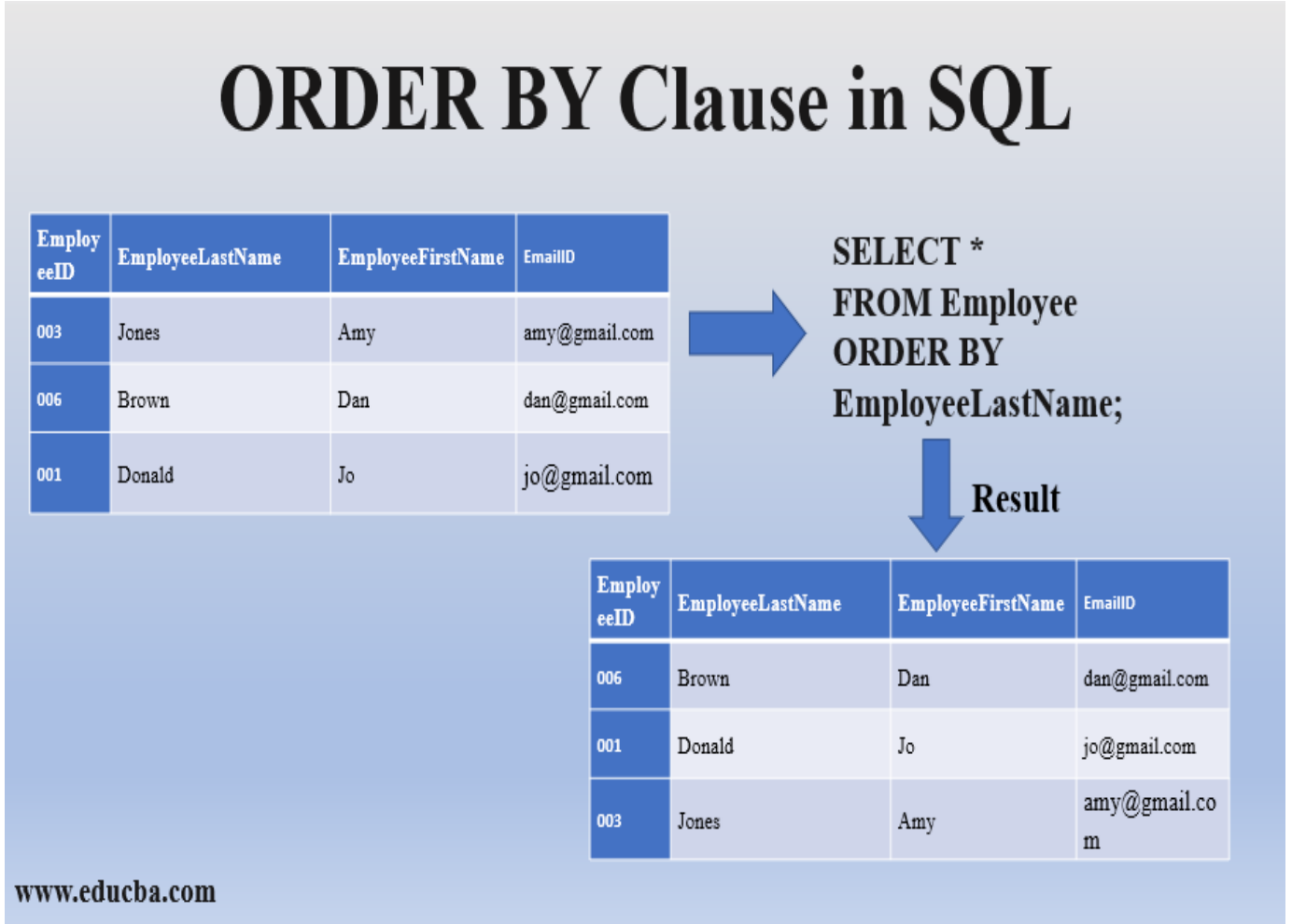
HAVING

DeptID	AVG(Salary)
2	4000.00
3	4250.00

ORDER BY

The ORDER BY clause in SQL is used to sort the extracted data in either ascending or descending according to one or more columns.

- By default ORDER BY sorts the data in ascending order.
- We can use the keyword DESC to sort the data in descending order and the keyword ASC to sort in ascending order.



TOP & LAST

It specifies the number of records to return from top and usually used to get a quick view of the schema (rows/columns) of the database when you start looking at a database for the first time.

```
Select top 8 * from population_table
```

city	state	population_estimate_2012	id
Arlington	TX	375600	50
Wichita	KS	385577	49
Cleveland	OH	390928	48
Minneapolis	MN	392880	47
Tulsa	OK	393987	46
Oakland	CA	400740	45
Miami	FL	413892	44
Omaha	NE	421570	43

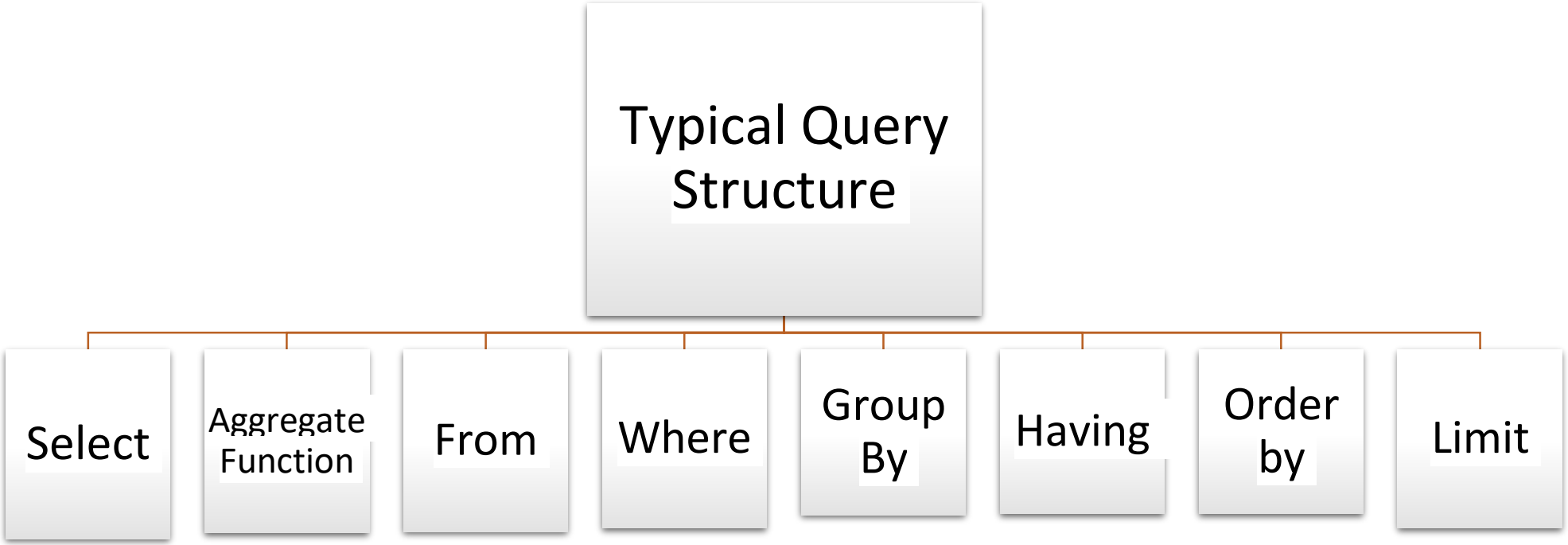
LIMIT

The LIMIT clause is used to set an upper limit on the number of tuples/records returned by SQL.

```
Select * from flights limit 5
```

air_traffic_delay	airline_code	airline_name	arrival_delay	cancellation_reason
0	DL	Delta Air Lines In...	-17	
0	DL	Delta Air Lines In...	-17	
0	DL	Delta Air Lines In...	-20	
0	DL	Delta Air Lines In...	3	
0	DL	Delta Air Lines In...	-25	

Standard Aggregate Query Structure



Thanks for attending

DataBhau communication platforms:

- **Whatsapp :** https://chat.whatsapp.com/BaP2CAajm9597J8LwzYE3j
- **Linkedin Handle:** https://www.linkedin.com/company/databhau/
- **Instructors Linkedin Handles:**
  - Shrey Jain : https://www.linkedin.com/in/shrey-jain-74a90b13b/
  - Harsh Katyayan : https://www.linkedin.com/in/harsh-katyayan-a2248316b/