

# HISTOGRAMS PRE-12c AND NOW

ANJU GARG

# About me

- More than 11 years of experience in IT Industry
- Sr. Corporate Trainer (Oracle DBA) with Koenig Solutions Pvt. Ltd.
- Oracle blog : <http://oracleinaction.com/>
- Email : anjugarg66@gmail.com
- Oracle Certified

**ORACLE®**

**Certified Associate**

**ORACLE®**

**Certified Professional**

**ORACLE®**

**Certified Professional**

Oracle Database 11g  
Administrator

**ORACLE®**

**Certified Professional**

Oracle Database 12c  
Administrator

**ORACLE®**

**Certified Expert**

Oracle Real Application  
Clusters 11g and  
Grid Infrastructure  
Administrator

**ORACLE®**

**Certified Expert**

Oracle Database 11g  
Performance Tuning

**ORACLE®**

**Certified Expert**

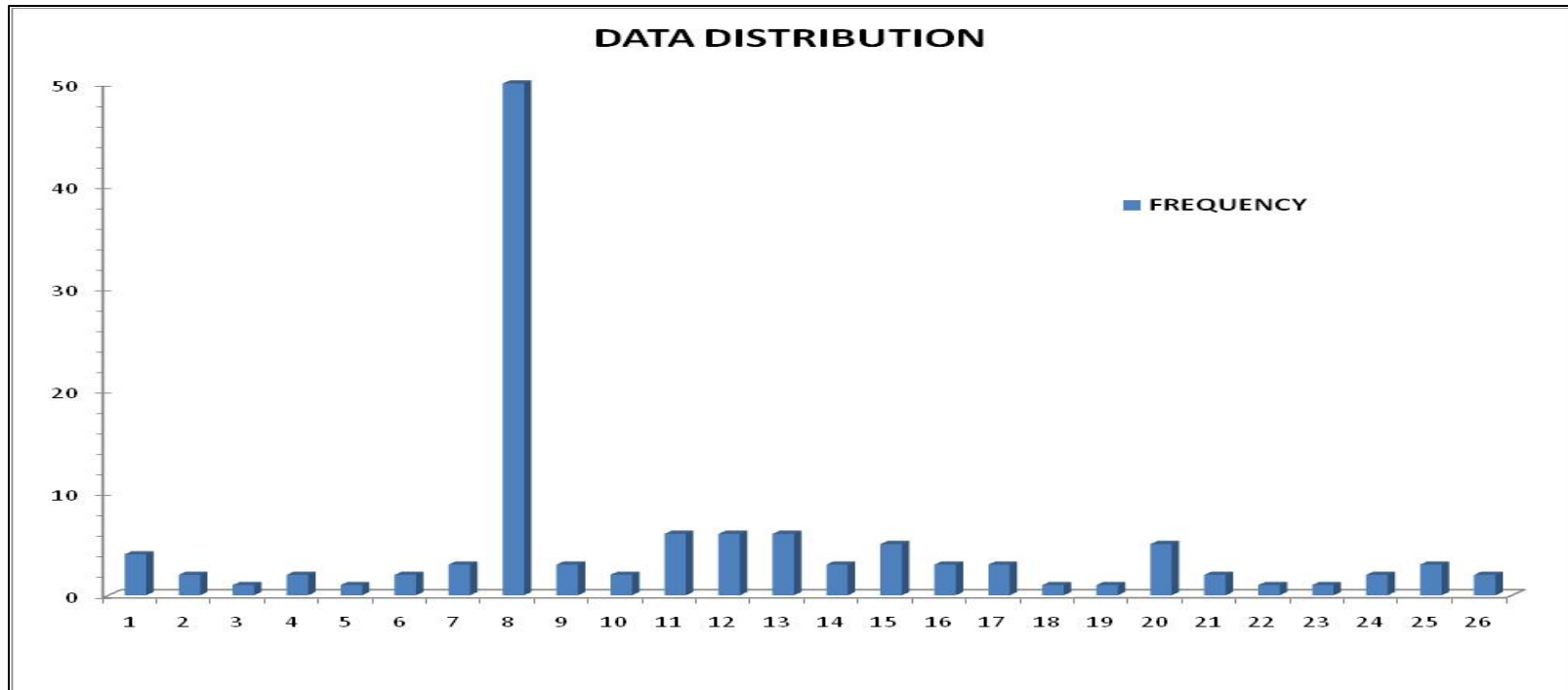
Oracle Database 11g  
Release 2 SQL Tuning

# Agenda

- Need For Histograms
- Pre-12c Histograms
  - Frequency Histograms
  - Height Balanced Histograms
- Issues With Histograms In 11g
- Histograms in 12c
  - Top Frequency Histograms
  - Hybrid Histograms
  - Hybrid Histograms - Corollary
- Conclusion
- References
- Q & A

# Need For Histograms

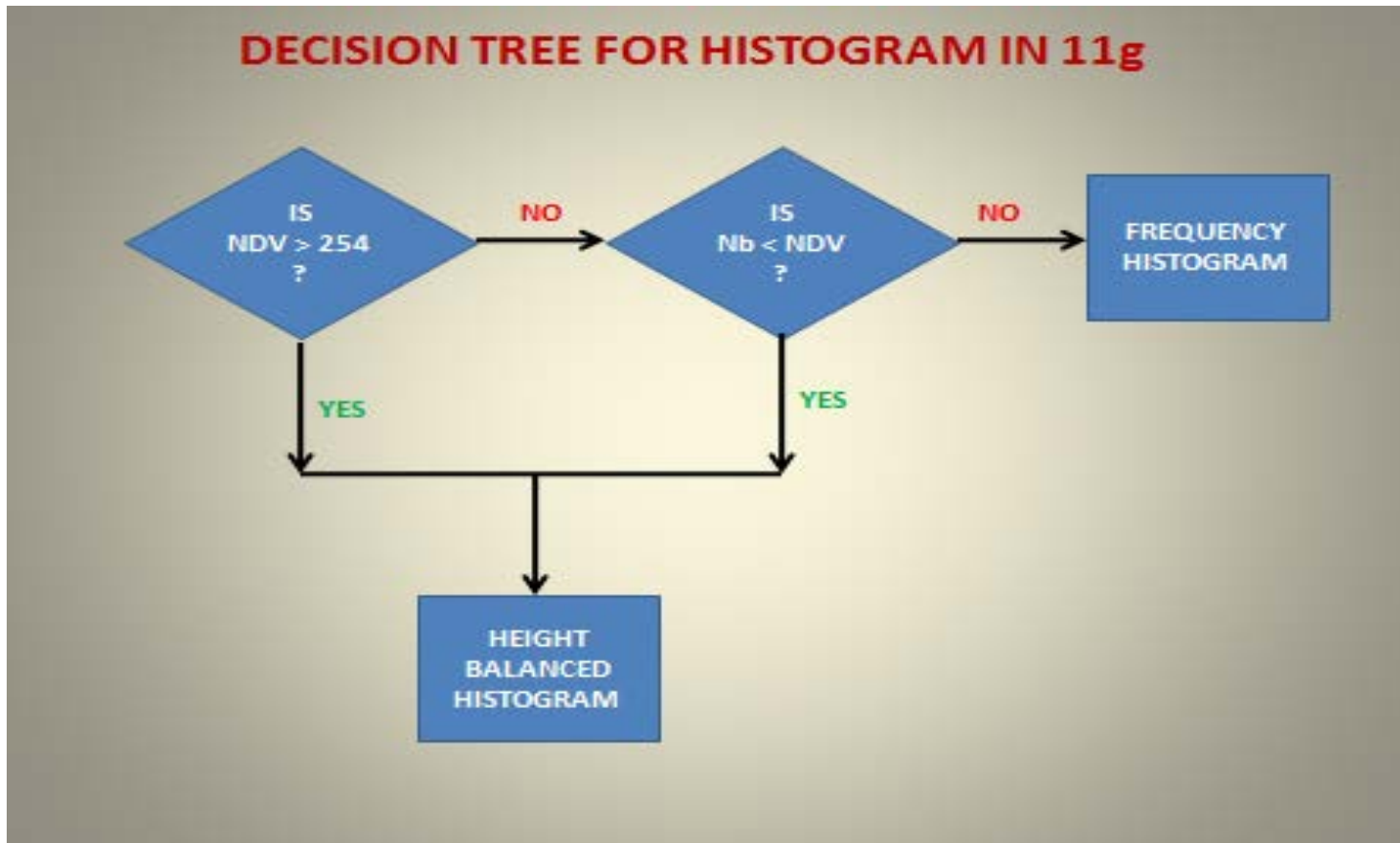
## Skewed Data Distribution



- 26 distinct values in the column ID
- More than 40% rows have ID = 8
- Optimizer assumes uniform distribution and misestimates cardinality
- Results in a bad execution path

# Pre-12c Histograms

- Frequency histograms
- Height balanced histograms



*Nb* – Number of buckets  
*NDV* – No. of Distinct Values

# Frequency Histogram

- $NDV \leq 254$  and  $Nb \geq NDV$
- Records each different value and its exact cardinality.
- **One bucket contains exactly one value.**
- One value resides entirely in one bucket.
- Bucket size = cardinality of the value
- **Precise**

```
SQL>exec dbms_stats.gather_table_stats -  
      (ownname => 'HR', -  
        tabname => 'HIST', -  
        method_opt => 'FOR COLUMNS ID', -  
        cascade => true);
```

```
SQL>select table_name, column_name, histogram,num_distinct,num_buckets  
       where table_name = 'HIST'  
       and    column_name = 'ID';
```

TABLE_NAME	COLUMN_NAME	HISTOGRAM	NUM_DISTINCT	NUM_BUCKETS
HIST	ID	<b>FREQUENCY</b>	<b>26</b>	<b>26</b>

*Nb* – Number of buckets

*NDV* – No. of Distinct Values

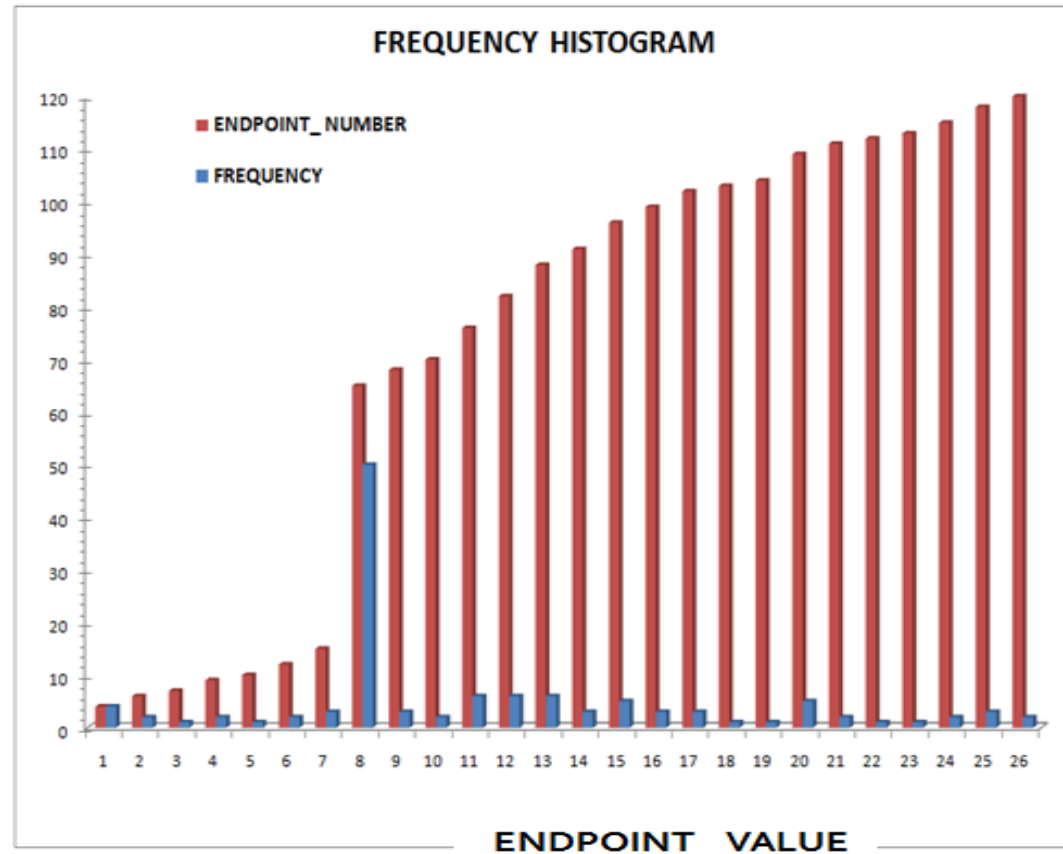
# Frequency Histogram

```
SQL>select ENDPOINT_VALUE, ENDPOINT_NUMBER
       from dba_histograms
       where table_name = 'HIST'
       and    column_name = 'ID';
```

ENDPOINT_VALUE	ENDPOINT_NUMBER
1	4
2	6
3	7
4	9
5	10
6	12
7	15
8	65
9	68
10	70
11	76
12	82
13	88
14	91
15	96
16	99
17	102
18	103
19	104
20	109
21	111
22	112
23	113
24	115
25	118
26	120

**Table 2.2**

ENDPOINT\_VALUE - The value in a bucket.  
ENDPOINT\_NUMBER - Cumulative frequency



**Fig 2.2**

# Frequency Histogram

- Optimizer uses frequency histogram to estimate cardinality accurately
- Uses FTS access path as desired even though column ID is indexed

```
SQL>explain plan for select * from hr.hist where id = 8;  
      select * from table(dbms_xplan.display);
```

PLAN\_TABLE\_OUTPUT

-----  
Plan hash value: 538080257  
-----

Id	Operation	Name	Rows	Bytes	Cost (%CPU)	Time	
0	SELECT STATEMENT		50	50200	7 (0)	00:00:01	
* 1	<b>TABLE ACCESS FULL</b>	HIST	<b>50</b>	50200	7 (0)	00:00:01	

-----



# Height-Balanced Histogram

- $NDV > 254$  or  $Nb < NDV$
- Distributes the count of all rows evenly across all histogram buckets
- **All buckets have almost the same number of rows**
- **Less precise**
- No. of buckets = 20 ( $< NDV (=26)$  )

```
DB11g>exec dbms_stats.gather_table_stats -  
          (ownname => 'HR', -  
            tabname => 'HIST', -  
            method_opt => 'FOR COLUMNS ID size 20', -  
            cascade => true);
```

```
DB11g>select table_name,column_name,histogram,num_distinct,num_buckets  
       from dba_tab_col_statistics  
       where table_name = 'HIST' and column_name = 'ID';
```

TABLE_NAME	COLUMN_NAME	HISTOGRAM	NUM_DISTINCT	NUM_BUCKETS
HIST	ID	HEIGHT BALANCED	26	20

*Nb* – Number of buckets

*NDV* – No. of Distinct Values

# Height-Balanced Histogram

```
DB11g>select ENDPOINT_VALUE,ENDPOINT_NUMBER
        from dba_histograms
        where table_name = 'HIST'
        and column_name = 'ID';
```

ENDPOINT_VALUE	ENDPOINT_NUMBER
1	0
2	1
6	2
8	10
9	11
11	12
12	13
13	14
14	15
15	16
17	17
20	18
24	19
26	20

14 rows selected.

Table 2.5

ENDPOINT _ NUMBER	VALUES						ENDPOINT _ VALUE
1	1	1	1	1	2	2	2
2	3	4	4	5	6	6	6
3	7	7	7	8	8	8	8
4	8	8	8	8	8	8	8
5	8	8	8	8	8	8	8
6	8	8	8	8	8	8	8
7	8	8	8	8	8	8	8
8	8	8	8	8	8	8	8
9	8	8	8	8	8	8	8
10	8	8	8	8	8	8	8
11	8	8	8	8	8	9	9
12	9	9	10	10	11	11	11
13	11	11	11	11	12	12	12
14	12	12	12	12	13	13	13
15	13	13	13	13	14	14	14
16	14	15	15	15	15	15	15
17	16	16	16	17	17	17	17
18	18	19	20	20	20	20	20
19	20	21	21	22	23	24	24
20	24	25	25	25	26	26	26
HEIGHT- BALANCED HISTOGRAM							

Fig 2.3

Bucket size = Total no. of rows / Nb = 120 / 20 = 6  
 ENDPOINT\_NUMBER - unique identifier of a bucket  
 ENDPOINT\_VALUE - largest value stored in a bucket

# Height-Balanced Histogram

- One bucket can have multiple values.
- One value can span multiple buckets.
- Multiple buckets (3 – 10) with same ENDPOINT\_VALUE (8) compressed into a single bucket with the highest endpoint number (bucket 10).
- A **popular value** (8) occurs as an endpoint value of multiple buckets.
- Any value that is not popular is a **non-popular value**.

# Height-Balanced Histogram

## Popular values

- Cardinality of popular value = (Bucket size)\*(no. of buckets with value as endpoint)
- ID = 8 is an end point of 8 buckets
- **Estimated cardinality = (6 \* 8) = 48 (Actual = 50)**

```
DB11g> explain plan for select * from hr.hist where id =8;
        select * from table(dbms_xplan.display);
```

```
PLAN_TABLE_OUTPUT
```

```
-----
Plan hash value: 538080257
```

```
-----
| Id  | Operation                | Name | Rows  | Bytes | Cost (%CPU)| Time     |
-----|-----|-----|-----|-----|-----|-----|
|    0 | SELECT STATEMENT         |      |     48 | 48192 |       7   (0)| 00:00:01 |
|*   1 |  TABLE ACCESS FULL      | HIST |     48 | 48192 |       7   (0)| 00:00:01 |
-----
```

```
Predicate Information (identified by operation id):
```

```
-----
  2 - access("ID">=8)
```

# Height-Balanced Histogram

## Non-Popular values (Endpoint)

- Cardinality of non-popular value = (number of rows in table) \* density
- ID = 15 is an end point of one bucket
- **Estimated cardinality = 3 (Actual = 5)**

```
DB11g>explain plan for select * from hr.hist where id = 15;  
      select * from table(dbms_xplan.display);
```

```
PLAN_TABLE_OUTPUT
```

```
-----  
Plan hash value: 4058847011  
-----
```

Id	Operation	Name	Rows	Bytes	Cost (%CPU)	Time
0	SELECT STATEMENT		3	3012	2 (0)	00:00:01
1	TABLE ACCESS BY INDEX ROWID	HIST	<b>3</b>	3012	2 (0)	00:00:01
* 2	INDEX RANGE SCAN	HIST_IDX	3		1 (0)	00:00:01

```
-----  
Predicate Information (identified by operation id):  
-----
```

```
2 - access("ID"=15)
```

# Height-Balanced Histogram

## Non-Popular values (Non-endpoint)

- Cardinality of non-popular value = (num of rows in table) \* density
- ID = 3 is not an endpoint
- **Estimated cardinality = 3 (Actual = 1)**

```
DB11g>explain plan for select * from hr.hist where id = 3;  
      select * from table(dbms_xplan.display);
```

```
PLAN_TABLE_OUTPUT
```

```
-----  
Plan hash value: 4058847011  
-----
```

Id	Operation	Name	Rows	Bytes	Cost (%CPU)	Time
0	SELECT STATEMENT		3	3012	2 (0)	00:00:01
1	TABLE ACCESS BY INDEX ROWID	HIST	3	3012	2 (0)	00:00:01
* 2	INDEX RANGE SCAN	HIST_IDX	3		1 (0)	00:00:01

```
-----  
Predicate Information (identified by operation id):  
-----
```

```
2 - access("ID"=3)
```

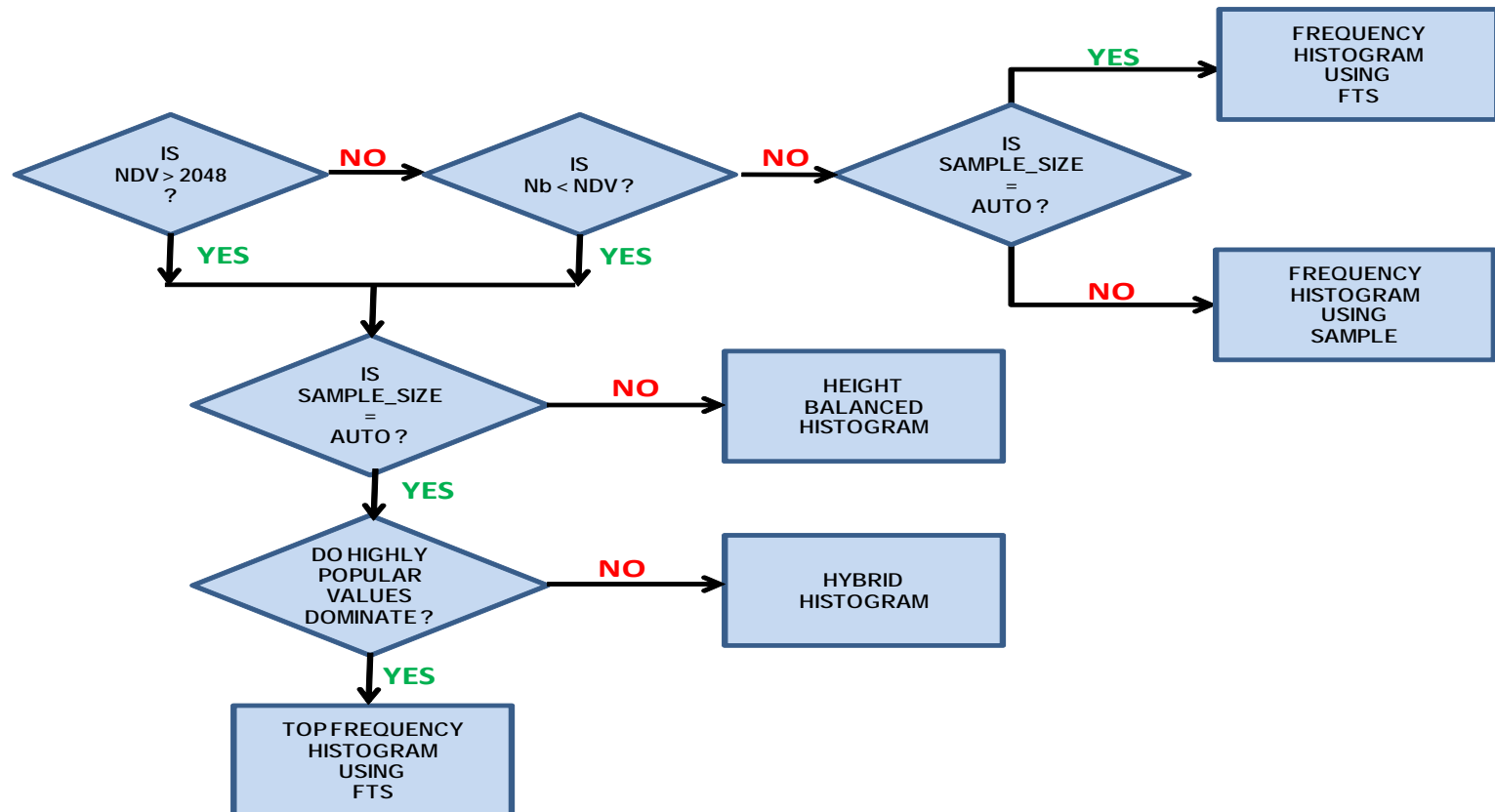
# Drawbacks Of Pre-12c Histograms

- **Frequency histograms**
  - Accurate but can be created only for  $NDV \leq 254$ .
- **Height balanced histograms**
  - May misestimate cardinality for both popular and non-popular values.
  - Produces inaccurate estimates for values that are almost popular.
  - A value which is endpoint of one bucket and almost fills up another bucket value might be considered unpopular.
  - Could result in a sub-optimal execution plan being chosen.

*NDV – No. of Distinct Values*

# Histograms in 12c

- Frequency histograms can be created for up to 2048 distinct values .
- Introduces two new types of histograms
  - Top-n-frequency histograms
  - Hybrid histograms





# Top Frequency Histograms

- Useful when a small number of distinct values dominate the data set.
- Capture highly popular values by ignoring statistically insignificant unpopular values.
- **Prerequisites**
  - $NDV > Nb$
  - The percentage of rows occupied by the top  $Nb$  frequent values is equal to or greater than threshold  $p$ , where  $p$  is  $(1-(1/Nb))*100$ .
  - The **ESTIMATE\_PERCENT** is set to **AUTO\_SAMPLE\_SIZE** in the DBMS\_STATS statistics gathering procedure.
- **Features**
  - One value per bucket
  - One value contained entirely in one bucket
  - Cardinality of the value = Bucket size
  - Variable bucket size
  - Precise for all the endpoints captured

# Top Frequency Histogram

- NDV (26) > Nb (20)
- Threshold p for 20 buckets =  $(1 - (1/Nb)) * 100 = (1 - (1/20)) * 100 = 95.0$
- Top 20 most popular values occupy more than 95% of rows
- **ESTIMATE\_PERCENT = AUTO\_SAMPLE\_SIZE** (default)
- Precisely captures cardinality of values occurring top Nb (20) times.
- 6 (= NDV - Nb) Values occurring least no. of times are not captured.

```
DB12c>exec dbms_stats.gather_table_stats -
      (ownname => 'HR', tabname => 'HIST', method_opt => 'FOR COLUMNS ID size
20', cascade => true);
      select table_name, column_name, histogram, num_distinct, num_buckets
      from    dba_tab_col_statistics
      where   table_name = 'HIST' and column_name = 'ID';
```

TABLE_NAME	COLUMN_NAME	HISTOGRAM	NUM_DISTINCT	NUM_BUCKETS
HIST	ID	TOP-FREQUENCY	26	20

*Nb – Number of buckets*

*NDV – No. of Distinct Values*

# Top Frequency Histogram

```
DB12c>set pagesize 100
       select ENDPOINT_VALUE, ENDPOINT_NUMBER
       from dba_histograms
       where table_name = 'HIST'
       and column_name = 'ID';
```

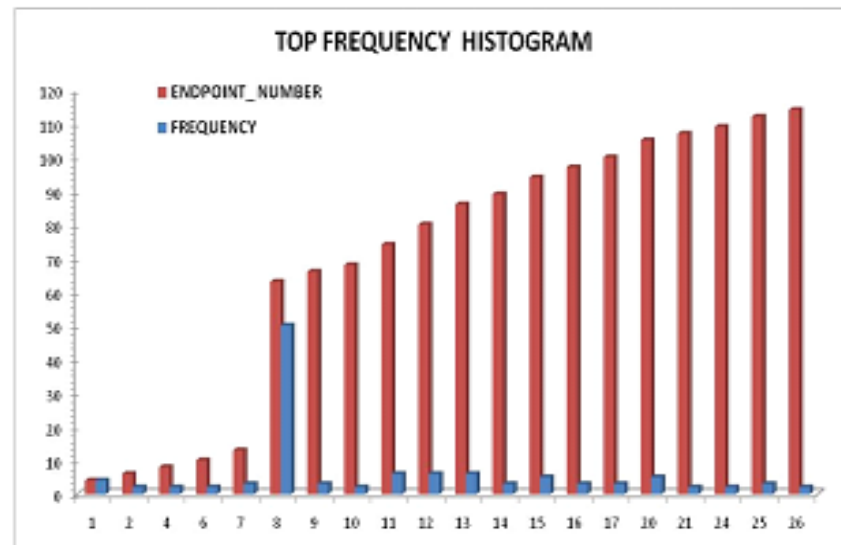
ENDPOINT_VALUE	ENDPOINT_NUMBER
1	4
2	6
4	8
6	10
7	13
8	63
9	66
10	68
11	74
12	80
13	86
14	89
15	94
16	97
17	100
20	105
21	107
24	109
25	112
26	114

20 rows selected

**Table 3.3**

ENDPOINT\_VALUE - Key value (ID)

ENDPOINT\_NUMBER - Cumulative Frequency



**Fig 3.2**

# Top Frequency Histogram

## Non-Popular values (Endpoint)

- Top Frequency histogram makes **accurate cardinality estimate** for unpopular value ID = 15
  - is one of the top 20 most frequently occurring values
  - Is an endpoint
  - Occurs 5 times

```
DB12c>explain plan for select * from hr.hist where id = 15;  
      select * from table(dbms_xplan.display);
```

```
PLAN_TABLE_OUTPUT
```

```
-----  
Plan hash value: 3950962134  
-----
```

Id	Operation	Name	Rows	Bytes	Cost(%CPU)	Time
0	SELECT STATEMENT		5	5020	2 (0)	00:00:01
1	TABLE ACCESS BY INDEX ROWID BATCHED	HIST	5	5020	2 (0)	00:00:01
* 2	INDEX RANGE SCAN	HIST_IDX	5		1 (0)	00:00:01

```
-----  
Predicate Information (identified by operation id):  
-----
```

```
2 - access("ID"=15)
```

# Top Frequency Histogram

## Non-Popular values (Non-endpoint)

- Top Frequency histogram makes **accurate cardinality estimate** for unpopular value ID = 3
  - is not one of the top 20 most frequently occurring values
  - Is not an endpoint
  - Occurs once

```
DB12c>explain plan for select * from hr.hist where id = 3;
          select * from table(dbms_xplan.display);
PLAN_TABLE_OUTPUT
-----
Plan hash value: 3950962134
-----
| Id | Operation                                | Name   | Rows | Bytes | Cost (%CPU) | Time
-----
|  0 | SELECT STATEMENT                        |        |    1 | 1004 | 2   (0)    | 00:00:01
|  1 |   TABLE ACCESS BY INDEX ROWID BATCHED | HIST   |    1 | 1004 | 2   (0)    | 00:00:01
|*  2 |    INDEX RANGE SCAN                     | HIST_IDX |    1 |      | 1   (0)    | 00:00:01
-----
Predicate Information (identified by operation id):
-----
   2 - access("ID"=3)
```

# Top Frequency Histograms

## Conclusion

- The occurrences of popular values are accurately captured at the expense of not capturing the data for least occurring values.
- Accurately estimates cardinality of top  $Nb$  popular values.
- Does not capture  $(NDV - Nb)$  values occurring least no. of times.
- Useful for cases when data set is dominated by a small no. of values.

*$Nb$  – Number of buckets*

*$NDV$  – No. of Distinct Values*

# Hybrid Histograms

- Combines characteristics of both height-based histograms and frequency histograms.
- **Correctly estimates the frequency of endpoints**
  - stores the `ENDPOINT_REPEAT_COUNT` - number of times the endpoint value is repeated.
- **Captures more endpoints**
  - stores all occurrences of every value in the same bucket - makes available buckets to store more endpoints.
- **Pre-requisites**
  - $Nb < NDV$
  - The criteria for top frequency histograms do not apply i.e.  
Percentage of rows occupied by the top  $Nb$  most popular values is less than  $p$   
where  $p = (1 - (1/Nb)) * 100$ .
  - The sampling percentage is **`AUTO_SAMPLE_SIZE`**.

*Nb – Number of buckets*

*NDV – No. of Distinct Values*

# Hybrid Histogram

- On deleting 20 rows with ID = 8 from table HR.HIST , it qualifies for hybrid histogram creation.
- $Nb(20) < NDV(26)$
- Threshold  $p$  for 20 buckets =  $p = (1 - (1/nb)) * 100 = (1 - (1/20)) * 100 = 95.0$
- Top 20 most popular ID's occupy less than 95% of total rows.

```
DB12c>exec dbms_stats.gather_table_stats -  
          (ownname => 'HR',tabname => 'HIST', -  
           method_opt => 'FOR COLUMNS ID size 20', -  
           cascade => true);
```

```
DB12c>select table_name,column_name, histogram,num_distinct,num_buckets  
       from dba_tab_col_statistics  
       where table_name = 'HIST' and column_name = 'ID';
```

TABLE_NAME	COLUMN_NAME	HISTOGRAM	NUM_DISTINCT	NUM_BUCKETS
HIST	ID	HYBRID	26	20



# Hybrid Histogram

```
DB12c>select ENDPOINT_VALUE, ENDPOINT_NUMBER,
        ENDPOINT_REPEAT_COUNT RPT_CNT
        from dba_histograms
        where table_name = 'HIST'
        and    column_name = 'ID';
```

ENDPOINT_VALUE	ENDPOINT_NUMBER	RPT_CNT
1	4	4
3	7	1
5	10	1
7	15	3
8	45	30
10	50	2
11	56	6
12	62	6
13	68	6
14	71	3
15	76	5
16	79	3
17	82	3
19	84	1
20	89	5
21	91	2
22	92	1
23	93	1
24	95	2
26	100	2

```
20 rows selected.
```

**ENDPOINT\_VALUE** : The largest value in a bucket

ENDPOINT\_NUMBER : Cumulative frequency

ENDPOINT\_REPEAT\_COUNT: No. of times the endpoint value occurs

[illegible]

# Hybrid Histograms

## Conclusion

- Hybrid histograms have features of both frequency and height balanced histograms
- **Features similar to frequency histograms**
  - All occurrences of a value are placed in one bucket
  - ENDPOINT\_NUMBER stores cumulative frequency
  - Variable bucket size
- **Features similar to Height Balanced histograms**
  - One bucket can have multiple values.
- Captures more endpoints
- Now we have a better estimate of the data distribution of “non-popular” values.
- Hybrid histograms are an improvement over height balanced histograms when top frequency histograms cannot be used

# Hybrid Histograms : Corollary

- If no more than two values are stored in one bucket, frequency of all the values captured in buckets can be accurately estimated.

Total no. of entries in the bucket = BUCKET\_SIZE  
= Difference of 2 consecutive ENDPOINT\_NUMBER's

Frequency of endpoint = ENDPOINT\_REPEAT\_COUNT

If BUCKET\_SIZE = ENDPOINT\_REPEAT\_COUNT (e.g. Bucket no. 5)

Bucket filled with ENDPOINT\_VALUE (8)

Frequency of endpoint = ENDPOINT\_REPEAT\_COUNT (30)

else (e.g. Bucket no. 4)

Bucket has two entries : endpoint (7) and non-endpoint (6)

Frequency of endpoint (7) = ENDPOINT\_REPEAT\_COUNT (3)

Frequency of non endpoint (6) = (BUCKET\_SIZE - ENDPOINT\_REPEAT\_COUNT)  
= ( 5 - 3 ) = 2

# Hybrid Histograms : Corollary

- Frequency of values up to twice the number of buckets should be estimated accurately using hybrid histogram.
- Create a histogram with No. of buckets =  $NDV/2 = 26/2 = 13$

```
DB12c>exec dbms_stats.gather_table_stats -  
      (ownname => 'HR', tabname => 'HIST',method_opt => 'FOR COLUMNS ID  
size 13', cascade => true);
```

```
DB12c> select table_name, column_name, histogram, num_distinct,  
              num_buckets  
        from dba_tab_col_statistics  
       where table_name = 'HIST' and column_name = 'ID';
```

TABLE_NAME	COLUMN_NAME	HISTOGRAM	NUM_DISTINCT	NUM_BUCKETS
HIST	ID	HYBRID	26	13

*Nb – Number of buckets*  
*NDV – No. of Distinct Values*

# Hybrid Histograms: Corollary

```
DB12c>select  ENDPOINT_VALUE
              ENDPOINT,
              ENDPOINT_NUMBER
              ENDPOINT_NO,
              endpoint_repeat_count
              rpt_cnt
from    dba_histograms
where   table_name = 'HIST'
and     column_name = 'ID';
```

ENDPOINT	ENDPOINT_NO	RPT_CNT
----------	-------------	---------

1	4	4
5	10	1
8	45	30
11	56	6
12	62	6
13	68	6
15	76	5
17	82	3
20	89	5
22	92	1
23	93	1
24	95	2
26	100	2

13 rows selected.

*ENDPOINT\_VALUE* : Largest value in a bucket

**ENDPOINT NUMBER :** Cumulative frequency

**ENDPOINT\_REPEAT\_COUNT**: Frequency of endpoint

[illegible][illegible]

# Hybrid Histograms : Corollary

- Actual histogram
  - some of the buckets have more than two values
  - cardinality of some of the non endpoint values cannot be accurately determined.
- Hypothetical histogram
  - All the buckets have two values each.
  - cardinality of all ( $2 \cdot Nb$ ) values can be accurately determined.
- Further enhancement?
  - $Nb \geq NDV / 2$ 
    - Hybrid histogram with no more than two values per bucket
    - Accurately estimates frequency of all the distinct values
  - $Nb < NDV / 2$ 
    - Top Frequency Hybrid histogram with no more than two values per bucket
    - Captures ( $2 \cdot Nb$ ) most popular values
    - Accurately estimates frequency of all the captured values

*Nb – Number of buckets*

*NDV – No. of Distinct Values*

# Conclusion

- In 12c, frequency histogram can be created for  $NDV \leq 2048$  .
- Top Frequency and Hybrid histograms are designed to overcome flaws of Height-Balanced histograms.
- Top frequency and Hybrid histograms are created only if `ESTIMATE_PERCENT = AUTO_SAMPLE_SIZE` .
- Top frequency histogram accurately estimates the frequencies for only top occurring values if a small number of values dominate the data set.
- Hybrid histogram has features of both frequency and height balanced histograms
- Hybrid histogram captures more endpoints and estimates their frequency accurately.

*Nb – Number of buckets*

*NDV – No. of Distinct Values*

# References

- [http://docs.oracle.com/database/121/TGSQL/tgsql\\_histo.htm#TGSQL366](http://docs.oracle.com/database/121/TGSQL/tgsql_histo.htm#TGSQL366)
- <http://jimczuprynski.files.wordpress.com/2014/04/czuprynski-select-q2-2014.pdf>
- <http://jonathanlewis.wordpress.com/2013/09/01/histograms/>





# Thank You!

ANJU GARG

Email: [anjugarg66@gmail.com](mailto:anjugarg66@gmail.com)

Blog: <http://oracleinaction.com>