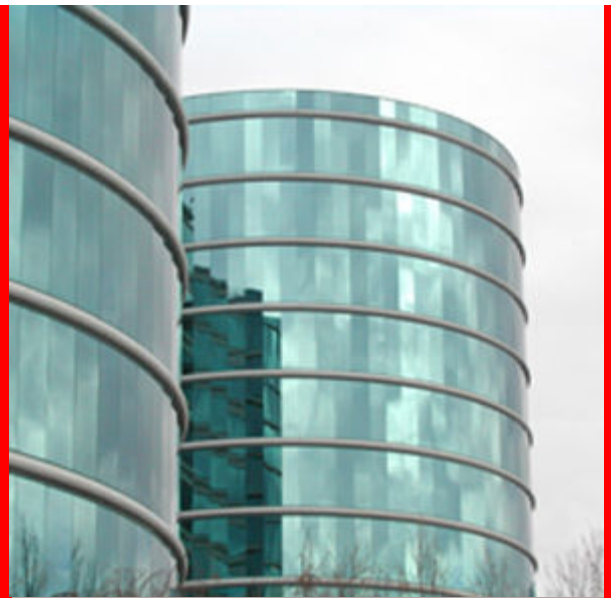
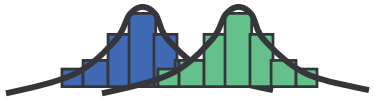

ORACLE®



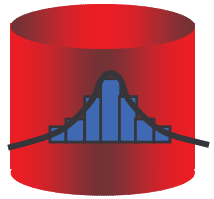
ORACLE®

Oracle Statistical Functions



Know More, Do More, Spend Less!

10g Statistics & SQL Analytics



- Ranking functions
  - rank, dense_rank, cume_dist, percent_rank, ntile
- Window Aggregate functions (moving and cumulative)
  - Avg, sum, min, max, count, variance, stddev, first_value, last_value
- LAG/LEAD functions
  - Direct inter-row reference using offsets
- Reporting Aggregate functions
  - Sum, avg, min, max, variance, stddev, count, ratio_to_report
- Statistical Aggregates
  - Correlation, linear regression family, covariance
- Linear regression
  - Fitting of an ordinary-least-squares regression line to a set of number pairs.
  - Frequently combined with the COVAR_POP, COVAR_SAMP, and CORR functions.
- Descriptive Statistics
  - average, standard deviation, variance, min, max, median (via percentile_count), mode, group-by & roll-up
  - DBMS_STAT_FUNCS: summarizes numerical columns of a table and returns count, min, max, range, mean, stats_mode, variance, standard deviation, median, quantile values, +/- 3 sigma values, top/bottom 5 values
- Correlations
  - Pearson's correlation coefficients, Spearman's and Kendall's (both nonparametric).
- Cross Tabs
  - Enhanced with % statistics: chi squared, phi coefficient, Cramer's V, contingency coefficient, Cohen's kappa
- Hypothesis Testing
  - Student t-test , F-test, Binomial test, Wilcoxon Signed Ranks test, Chi-square, Mann Whitney test, Kolmogorov-Smirnov test, One-way ANOVA
- Distribution Fitting
  - Kolmogorov-Smirnov Test, Anderson-Darling Test, Chi-Squared Test, Normal, Uniform, Weibull, Exponential
- Pareto Analysis (documented)
  - 80:20 rule, cumulative results table

Note: Statistics and SQL Analytics are included in Oracle Database Standard Edition

Descriptive Statistics

- MEDIAN:
  - Takes numeric or datetype values and returns the middle value

```
SELECT EDUCATION, MEDIAN(ANNUAL_INCOME) from
CD_BUYERS GROUP BY EDUCATION;
```

- STATS_MODE:
  - Returns the value that occurs most frequently

```
SELECT EDUCATION, STATS_MODE(ANNUAL_INCOME)
from CD_BUYERS GROUP BY EDUCATION;
```

SUMMARY procedure

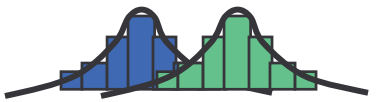
- The SUMMARY procedure is used to summarize a numerical column
- The summary is returned as record of type summaryType
- The record can return the numerical values in the following modes:
  - count
  - min
  - max
  - range
  - mean
  - Cmode
  - Variance
  - stddev
  - median
  - extreme_values
  - top_5_values
  - bottom_5_values
  - quantile values
  - +/- x sigma

SUMMARY procedure

- A subset of data that is returned after execution of the PL/SQL package "summarizes" the use of the different SUMMARY procedures.

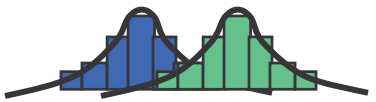
```
Summary of AMOUNT_SOLD:
Count              : 960
Min                : 4
Max                : 8184
Range              : 8180
Mean               : 721.879166666666666666666666666667
Mode(s)            : 24
Variance           : 812566.266051442474800139015660514421
Stddev             : 901.424575908291106316431472900222979
Quantile_5         : 30
Quantile_25        : 140
Median             : 419.9
Quantile_75        : 1008
…
PL/SQL procedure successfully completed
```

Hypothesis Testing



- Parametric Tests
  - Parametric tests make some assumptions about the data.
- Parametric hypothesis tests in Oracle Database 10g include:
  - T-test
  - F-test
  - One-Way ANOVA

T-Test



- T-tests are used to measure the significance of a difference of means.
- T-tests include the following:
  - One-sample T-test
  - Paired-samples T-test
  - Independent-samples T-test (pooled variances)
  - Independent-samples T-test (unpooled variances)

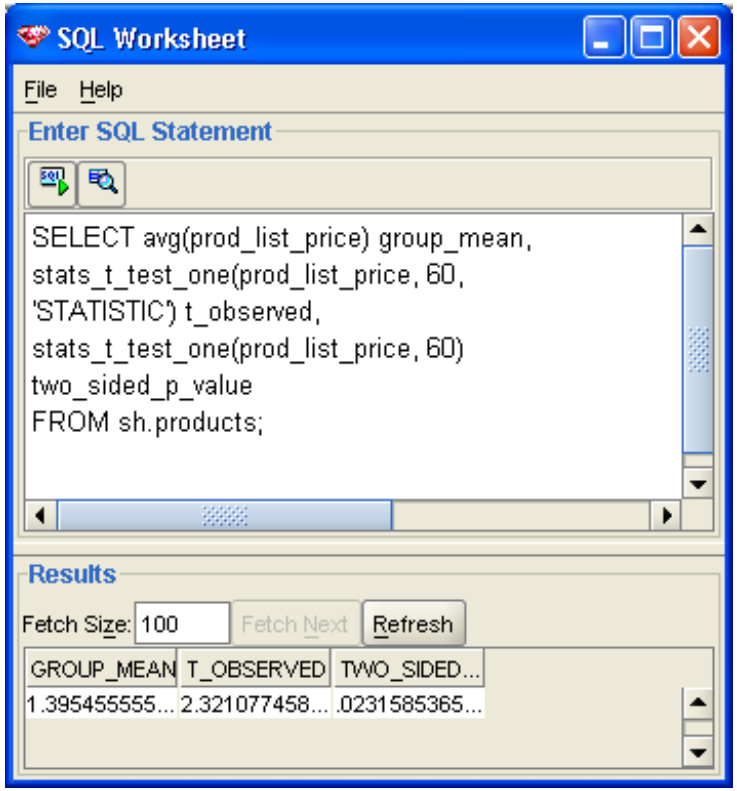
One-Sample T-Test: Example

- This query compares the mean of the product list prices to the assumed value of 60:

```
SELECT avg(prod_list_price) group_mean,
stats_t_test_one(prod_list_price, 60,
'STATISTIC') t_observed,
stats_t_test_one(prod_list_price, 60)
two_sided_p_value
FROM sh.products;
```

- Returns the observed t value and its related two-sided significance:

```
GROUP_MEAN T_OBSERVED
TWO_SIDED_P_VALUE
58.8388154 -1.9191106 .054998562
```



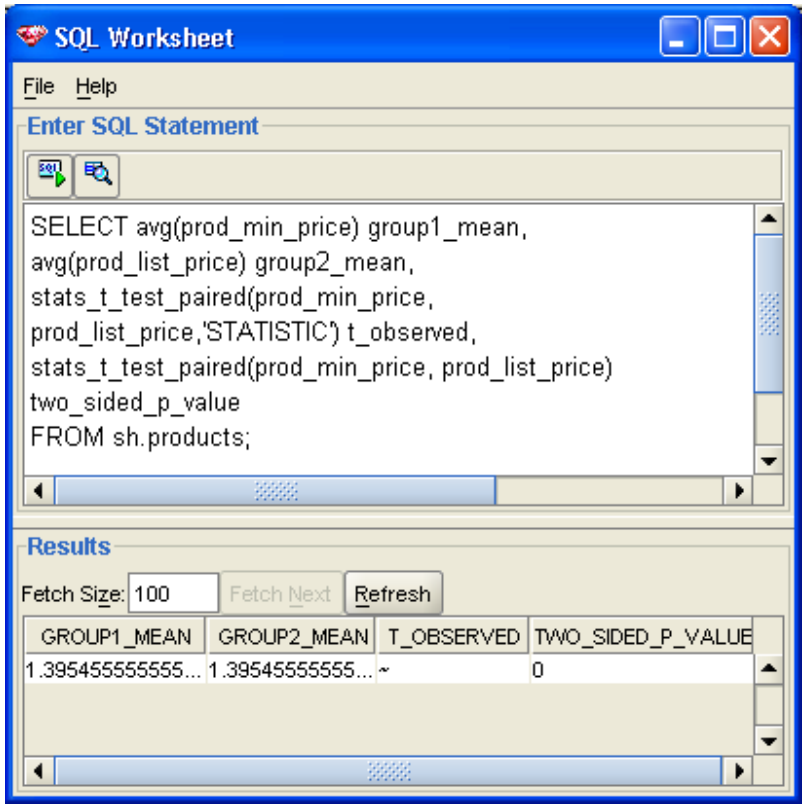
Paired Samples T-Test: Example

- This query compares the mean of the product minimum prices to the mean of the product list prices:

```
SELECT avg(prod_min_price) group1_mean,
avg(prod_list_price) group2_mean,
stats_t_test_paired(prod_min_price,
prod_list_price,'STATISTIC') t_observed,
stats_t_test_paired(prod_min_price,
prod_list_price)
two_sided_p_value
FROM sh.products;
```

- Returns the observed t value and its related two-sided significance:

```
GROUP1_MEAN GROUP2_MEAN
T_OBSERVED TWO_SIDED_P_VALUE
40.5285721 58.8388154 -85.84078 0
```



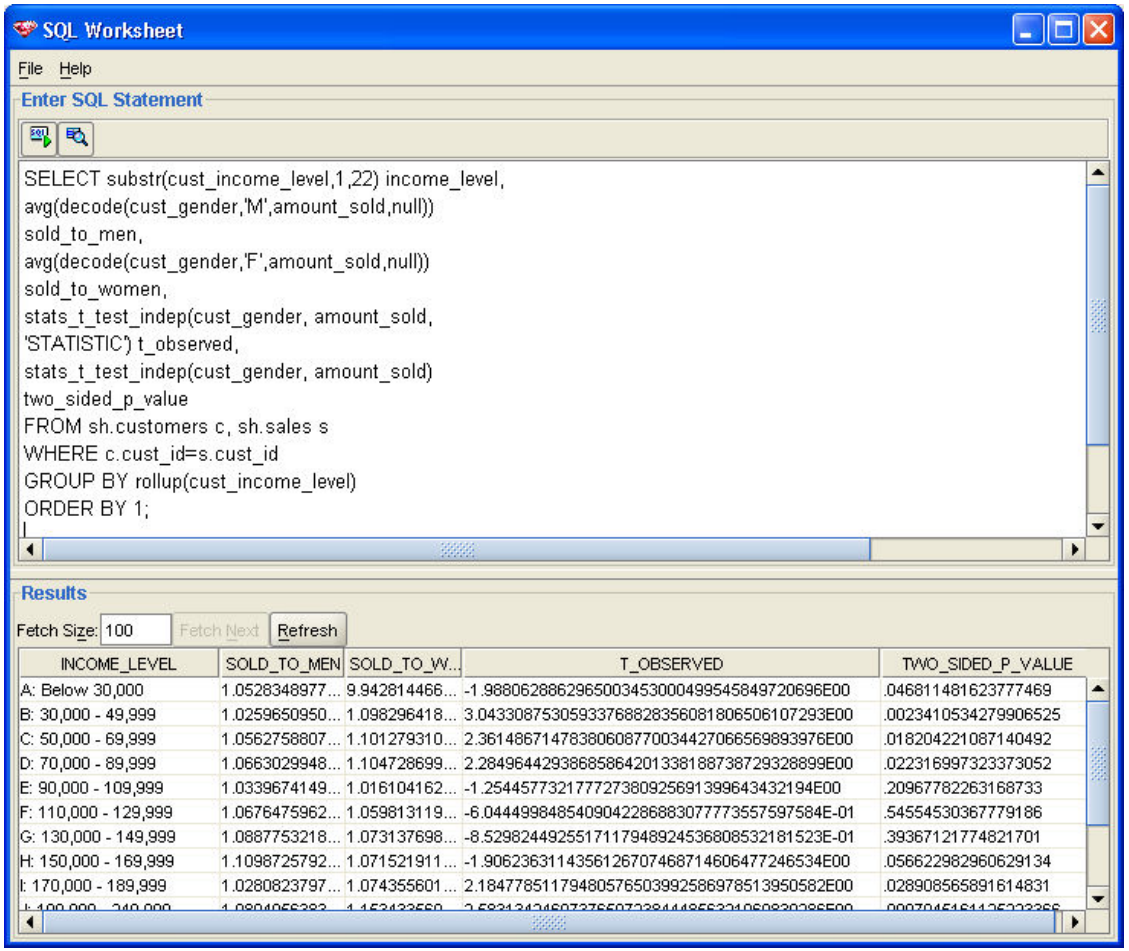
Independent Samples T-Test (Pooled Variances): Example

- This query compares the mean of amount sold between men and women:SELECT avg(prod_min_price) group1_mean,

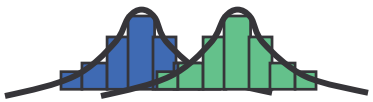
```
SELECT substr(cust_income_level,1,22) income_level,
avg(decode(cust_gender,'M',amount_sold,null))
sold_to_men,
avg(decode(cust_gender,'F',amount_sold,null))
sold_to_women,
stats_t_test_indep(cust_gender, amount_sold,
'STATISTIC') t_observed,
stats_t_test_indep(cust_gender, amount_sold)
two_sided_p_value
FROM sh.customers c, sh.sales s
WHERE c.cust_id=s.cust_id
GROUP BY rollup(cust_income_level)
ORDER BY 1;
```

Independent Samples T-Test (Pooled Variances): Example

- Independent Samples T-Test (Pooled Variances): Example



F-Test

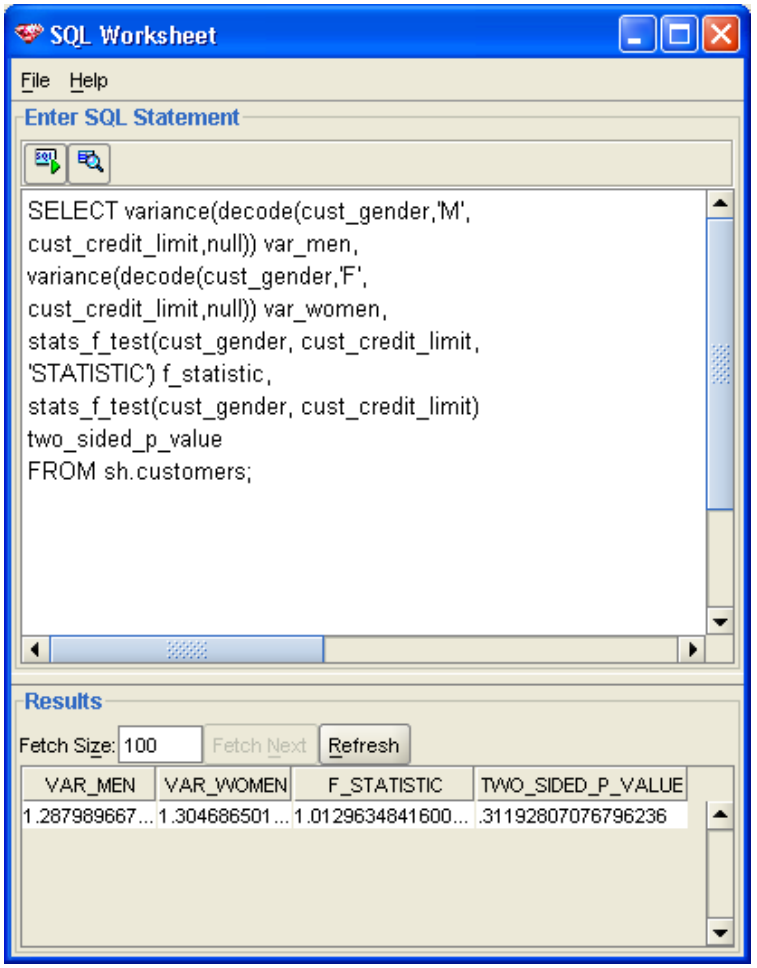


- **This query compares the variance in the credit limit between male and female customers:**

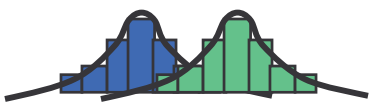
```
SELECT variance(decode(cust_gender,'M',
cust_credit_limit,null)) var_men,
variance(decode(cust_gender,'F',
cust_credit_limit,null)) var_women,
stats_f_test(cust_gender, cust_credit_limit,
'STATISTIC') f_statistic,
stats_f_test(cust_gender, cust_credit_limit)
two_sided_p_value
FROM sh.customers;
```

- **Returns the observed f value and its related two-sided significance:**

```
VAR_MEN VAR_WOMEN F_STATISTIC
 TWO_SIDED_P_VALUE
12275828.1 12562439.8 1.02334765 .082516768
```



One-Way ANOVA

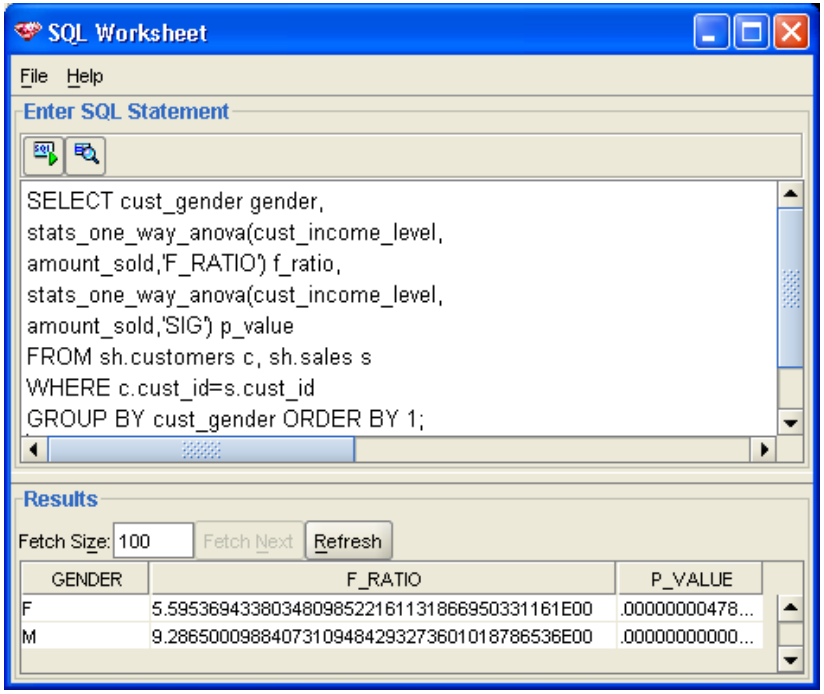


- **This query compares the average amount of product sold to customers with different income levels:**

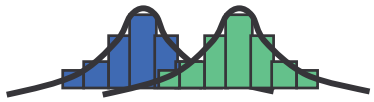
```
SELECT cust_gender gender,
stats_one_way_anova(cust_income_level,
amount_sold,'F_RATIO') f_ratio,
stats_one_way_anova(cust_income_level,
amount_sold,'SIG') p_value
FROM sh.customers c, sh.sales s
WHERE c.cust_id=s.cust_id
GROUP BY cust_gender ORDER BY 1;
```

- **Returns the one-way ANOVA significance and splits this on a per-gender basis:**

```
G F_RATIO P_VALUE
F 60.7217737 0
M 33.0952715 2.6910E-71
```

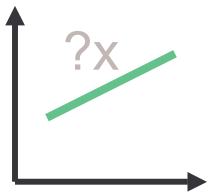


Hypothesis Testing with Nonparametric Tests



- Nonparametric tests are used when certain assumptions about the data are questionable.
- This may include the difference between samples that are not normally distributed.
- All tests involving ordinal scales (in which data is ranked) are nonparametric.
- Nonparametric tests supported in Oracle Database 10g:
  - Binomial test
  - Wilcoxon Signed Ranks test
  - Mann-Whitney test
  - Kolmogorov-Smirnov test

Correlation Functions



- The CORR_S and CORR_K functions support nonparametric or rank correlation (finding correlations between expressions that are ordinal scaled).
- Correlation coefficients take on a value ranging from –1 to 1, where:
  - 1 indicates a perfect relationship
  - –1 indicates a perfect inverse relationship
  - 0 indicates no relationship
- The following query determines whether there is a correlation between the weight class and the list price of a product, using Spearman's correlation:

```
select
CORR_S(weight_class,
list_price)
coefficient,
CORR_S(weight_class,
list_price,
'TWO_SIDED_SIG')
p_value
from
oe.product_information
where product_status =
'orderable';

COEFFICIENT P_VALUE
----------- ----------
.323518292 1.4204E-07
```

Cross Tabulations

- This query analyzes the strength of the association between customer income level and gender using a cross tabulation:

```
SELECT stats_crosstab(cust_gender, cust_income_level,
'CHISQ_OBS') chi_squared,
stats_crosstab(cust_gender, cust_income_level,
'CHISQ_SIG') p_value,
stats_crosstab(cust_gender, cust_income_level,
'PHI_COEFFICIENT') phi_coefficient
FROM sh.customers;
```

- Returns the observed p_value and its related phi coefficient significance:

```
CHI_SQUARED P_VALUE PHI_COEFFICIENT
11.7903518 .379606468 .136802208
```

Cross Tabulations

- The STATS_CROSSTAB function takes as arguments two expressions (the two variables being analyzed) and a value that determines which test to perform. These values include the following:
  - CHISQ_OBS (observed value of chi-squared)
  - CHISQ_SIG (significance of observed chi-squared)
  - CHISQ_DF (degree of freedom for chi-squared)
  - PHI_COEFFICIENT (phi coefficient)
  - CRAMERS_V (Cramer's V statistic)
  - CONT_COEFFICIENT (contingency coefficient)
  - COHENS_K (Cohen's kappa)
- The function returns all values as specified by the third argument. The default is CHISQ_SIG

Distribution-Fitting Functions

- Distribution-fitting functions in Oracle Database 10g include the following:
  - NORMAL_DIST_FIT function
  - UNIFORM_DIST_FIT function
  - POISSON_DIST_FIT function
  - WEIBULL_DIST_FIT function
  - EXPONENTIAL_DIST_FIT function
- These functions test how well a sample of values “fits” a particular distribution
- The IN parameter of each function specifies which of the tests to use to measure the fit

Why Statistics in the Database?

- Fewer moving parts
- Natural integration with database-driven applications
- Security
- Availability for mission critical applications
- Real-world scalability
- Leverage 30+ years of experience with ever *advancing* database technology
  - *e.g. 10gR2 Sort 10X performance improvements*

In-Database Statistics

Advantages

- Straightforward inclusion within interesting and arbitrarily complex queries
- Resource management benefits
- Oracle scalability, fully parallelizable
- Real-time computations
- Scalable and performant batch computations

… a true differentiator

Computing statistics is just the beginning

- How do you integrate the results into the business?
  - Drive reports…export to Discoverer, XML Publisher
  - Populate dashboard KPIs
  - Any tool that talks to Oracle can leverage results, combining statistical results with other data
- Integration in the database and across Oracle stack unleashes many more opportunities to leverage your results in spontaneous and unexpected ways

Oracle Database 10g is “changing the economics of analytics.”
Bill Hostmann, Gartner

ODM Value Proposition for IT

- Statistical functions
  - Provides the key statistical functions
- Secure
  - Data remains in the database – single version of truth
  - Leverage database security options
- Scalable solution
  - Analyze more data
  - Can leverage Real Application Clusters (RAC)
- Low TCO compared to other Statistics vendors
  - Lower cost for purchase, deployment, development, maintenance
- Increased efficiency
  - Seamlessly query, summarize and analyze the same data
  - Speeds data to information and knowledge


QUESTIONS
ANSWERS


ORACLE®