
BICON
BUSINESS INTELLIGENCE
A PART OF
SYSCO

Having Fun with Open Data and Oracle XE

Lets see what you can do with 1 CPU, 11GB data, and a free database

Some preaching towards DBAs included

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Øyvind Isene
@OyvindIsene

ORACLE ACE

BICON BUSINESS INTELLIGENCE

~~http://oisene.blogspot.com~~

https://enesi.no/

SYSCO http://sysco.no

BICON BUSINESS INTELLIGENCE http://www.bicon.no

500+ Technical Experts Helping Peers Globally

ORACLE ACE PROGRAM

ORACLE ACE Director | ORACLE ACE | ORACLE ACE Associate

3 Membership Tiers

- Oracle ACE Director
- Oracle ACE
- Oracle ACE Associate

bit.ly/OracleACEProgram

Connect:

oracle-ace_ww@oracle.com

Facebook.com/oracleaces

@oracleace

ORACLE Developer Community

Nominate yourself or someone you know: acenomination.oracle.com

What the heck?

- Why XE when you can play around with EE?
- Why open data?
- Why should DBAs care about this?
- What has this got to do with autonomous databases?

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Why Open Data?

- Lots of exiting free data set on the net
- Realistic data to practice SQL on
- Motivate you to learn more SQL & PL/SQL
- Test your skills and verify assumptions
- Learn more about a field that interests you

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

SQL for the curious mind

- Answer to a query leads to more questions
- SQL makes it easy to iteratively investigate a data set
- Learn more statistics - useful for everybody
  - Including DBAs involved with optimisation
  - Skew and weird statistical patterns

How can a DBA help others
if he don't know SQL?

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

SQL and PL/SQL for everything

- Almost
- Feel the pressure to learn R, Python, other stuff?
- …but not enough time?
- Try to do it first in the database
- Learn algorithms and methods first

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Improve applications

- SQL, PL/SQL and packages improve for every release
- Almost guaranteed to work better than anything else
- Check out supplied packages and SQL Reference
- Be a good friend with the developers

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Why?

- It’s about latency
- Reduce network trips
- Closeness to data
- Find the fastest algorithm
- Use code tested by many

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

In the database you can

- Load
- Transform
- Massage and wrangle data
- Analyse
- Display with APEX

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

OK, but why XE?

- Self-imposed limit to see how much I can fit into XE
- XE is free
- If you make something great you can release it*
- Smaller footprint and runs almost everywhere
- Faster to install

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Not a fair comparison, but

	Time to create Docker container	Virtual Size of Container
XE (11.2.0.1)	< 4 minutes	3 GB
EE (12.1.0.2)	~ 16 minutes	11 GB

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

XE Limits

- Runs on 1 CPU
- Uses max 1GB RAM
- Max 11GB user data
- Some heavy lifting from EE not included
  - Related to recovery, online operations,
  - Partitioning, management packs
  - Data Mining :-(
  - Oracle Spatial (but Locator is)

In other words:
Still lot of fun stuff in there

XE 18c

- Expected between March and August 2018
- Will have 12GB and more features including advanced compression (giving 40GB real capacity)
- Use 2GB RAM
- 2 CPUs and 4 pluggable databases
- Still no patches
- Yearly releases (meaning less vulnerable)

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Get Started

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Your own lab

- Docker and Vagrant
- Quick installation of XE
- Downloadable VMs from ODC (aka OTN)
- Free cloud trial?
- Needs to be easy and quick - focus on learning

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Express installation

- Linux: rpm -ivh downloads/oracle-xe-11.2.0-1.0.x86_64.rpm
- Windows: Double click something
- Docker: See post by SQL Maria: http://bit.ly/2yeKIQx

BICON BUSINESS INTELLIGENCE A PART OF SYSCO

Increase redo logs

- Archiving turned off by default - OK for fun
- Redo log files are default 50M for 11g
- Increase to say 500M

```
alter database add logfile group 1 '/u01/app/oracle/oradata/XE/redo01.log' size 500M;
alter database add logfile group 2 '/u01/app/oracle/oradata/XE/redo02.log' size 500M;
alter database add logfile group 3 '/u01/app/oracle/oradata/XE/redo03.log' size 500M;
alter database drop logfile group 4;
alter database drop logfile group 5;
alter database drop logfile group 6;
alter database add logfile group 4 '/u01/app/oracle/oradata/XE/redo04.log' size 500M reuse;
```

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

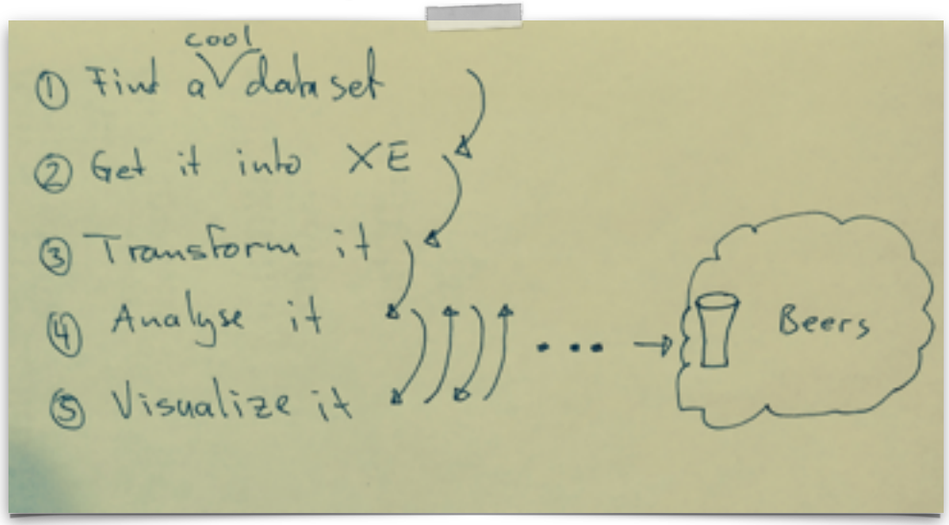
APEX?

- You should
- Version 4.0.2 included in XE 11.2.0.2
- Upgrade to latest - Easy!

```
cd /tmp
unzip apex_5.1.3.zip
sqlplus / as sysdba
@apexins.sql SYSAUX SYSAUX TEMP /i/
sqlplus / as sysdba
@apex_epg_config.sql /tmp
@apxchpwd.sql
```

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

My Workflow


cool
① Find a ✓ data set
② Get it into XE
③ Transform it
④ Analyse it
⑤ Visualize it
Beers

Get motivated

- Find an area that interests you
  - Beer data
  - Government data
  - Your own data from social network
- Look for something in a reasonable format
- Fire up SQL Developer

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Free datasets

- Data.world
- *Data is Plural* by Jeremy Singer-Vine
- kdnuggets.com
- kaggle.com
- Reddit: reddit.com/r/datasets
- Collect you own from gadgets and social networks

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

What can you squeeze into 11GB?

- Many datasets are quite small
  - 30 years of movies: < 1MB
- Different sets to discover weird connections
- Some will not fit:
  - A month of Reddit comments: 47GB JSON
- Advanced Compression not supported
  - But you can use UTL_COMPRESS

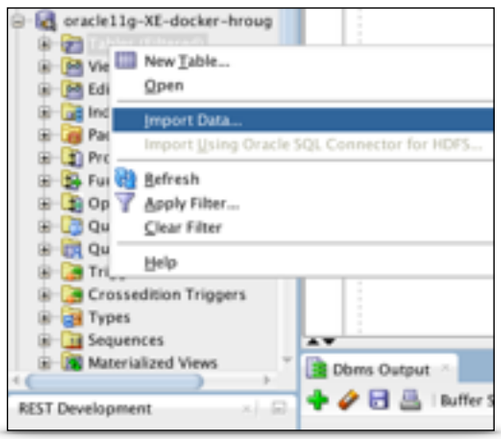
BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Loading data

- Get it in

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Import in SQL Developer



- Easy import for standard formats
  - CSV, XL, text
- Save time by lazy import:
  - large VARCHAR2 columns
  - transform datatypes later

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

External table

- Easier to automate
- Works best on data with a good structure

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

1.5 mill. beer tastings from Beer Advocate, 1999-2011

DBMS_LOB

- Read entire text file into a CLOB
- Easier for ridiculous formats

BICON BUSINESS INTELLIGENCE A PART OF SYSCO

Example with DBMS_LOB

```
create table file_clob(id number, text clob);

declare
  l_file BFILE ;
  l_clob clob ;
  l_dest_offset integer := 1;
  l_src_offset integer := 1;
  l_lang_context integer := 0;
  l_warning integer ;
begin
  insert into file_clob values(1,empty_clob()) returning text into l_clob;
  l_file := bfilename('TMP_DIR','Beeradvocate.txt');
  dbms_lob.fileopen( l_file, dbms_lob.FILE_READONLY );
  DBMS_LOB.LOADCLOBFROMFILE (
   dest_lob      => l_clob,
   src_bfile   => l_file,
   amount      => DBMS_LOB.LOBMAXSIZE,
   dest_offset  => l_dest_offset,
   src_offset   => l_src_offset,
   bfile_csid   => 873,
   lang_context => l_lang_context,
   warning      => l_warning);
  dbms_lob.fileclose( l_file );
  dbms_output.put_line('warning: ' || l_warning);
  COMMIT;
end;
/
```

BICON BUSINESS INTELLIGENCE A PART OF SYSCO

REST

- Widely used when integrating
- Many sites offers a REST API
- Easier to automate - use scheduler in db
- Not difficult to write PL/SQL to consume a web service with UTL_HTTP.

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

REST in peace

- REST is even easier with APEX
  - Create reports directly on REST data source
- Some sites offers real time data only
  - Need to fetch regularly to get historical data

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

JSON

- Used a lot in datasets
- Good support in 12c
- Transform to and from JSON with standard functions
- Open source libraries on Github for 11g

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Other Tools

- SQL Loader
- Open Source tools
- Write your own
- *But I prefer SQL Developer for simple ad-hoc*

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Transform

- Clean up other people's mess

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

SQL

- Verify structure and look for bad data
- Create Table as Select (CTAS)
- Probably the fastest way to convert data
- Add new features (columns) and aggregations

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

PL/SQL

- When you can’t do it in SQL
- Read line by line and transform with PL/SQL
- Learn regular expressions!
  - REGEXP_SUBSTR
  - REGEXP_COUNT
  - REGEXP_INSTR
  - REGEXP_REPLACE

BICON BUSINESS INTELLIGENCE A PART OF SYSCO

Field and Record Separators

- Structured data has a field and a record separator
- Useful to have routines to search for these and split accordingly
- Wild variations in various ugly formats

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Parsing file from Beeradvocate

A BEER Table

```
declare
 l_pos integer := 1;
 l_end_of_record integer ;
 l_line varchar2(32767);
 l_length integer;
 l_clob clob;
 l_beer beer%rowtype;
begin
  select text into l_clob
  from file_clob
  where id=1;
  l_length := dbms_lob.getlength(l_clob);
  while l_pos < l_length - 5 loop
    l_end_of_record := dbms_lob.instr(l_clob,chr(9) || chr(9) || chr(10) || chr(10),l_pos,1);
    l_line := dbms_lob.substr(l_clob,l_end_of_record - l_pos,l_pos);
    l_beer.name := regexp_substr(l_line,'beer\/name:\s+(.+)\s',1,1,'i',1);
    l_beer.id := regexp_substr(l_line,'beer\/beerId:\s+(\d+)\s',1,1,'i',1);
    l_beer.brewerid := regexp_substr(l_line,'beer\/brewerId:\s+(\d+)\s',1,1,'i',1);
    l_beer.abv := regexp_substr(l_line,'beer\/ABV:\s+(\d+\.\d+)\s',1,1,'i',1);
    l_beer.style := regexp_substr(l_line,'beer\/style:\s+(.+)\s',1,1,'i',1);
    insert into beer values l_beer;
    l_pos := l_end_of_record + 4;
  end loop;
end;
/
```

```
create table beer(
   name varchar2(500),
   id integer,
   brewerid integer,
   abv number,
   style varchar2(100));
```

1.58 million records, 8 minutes with XE on i7

BICON BUSINESS INTELLIGENCE A PART OF SYSCO

Final version with all attributes

REVIEWS

Columns | Constraints | Grants | Statistics | Triggers

Refresh: 0

	COLUMN_NAME	DATA_TYPE	NULLABLE	DA
1	BEER_NAME	VARCHAR2(200 BYTE)	Yes	(n
2	BEER_ID	NUMBER(38,0)	Yes	(n
3	BREWERID	NUMBER(38,0)	Yes	(n
4	ABV	NUMBER	Yes	(n
5	STYLE	VARCHAR2(200 BYTE)	Yes	(n
6	APPEARANCE	NUMBER	Yes	(n
7	AROMA	NUMBER	Yes	(n
8	PALATE	NUMBER	Yes	(n
9	TASTE	NUMBER	Yes	(n
10	OVERALL	NUMBER	Yes	(n
11	TIME_EPOCH	NUMBER	Yes	(n
12	PROFILE_NAME	VARCHAR2(100 BYTE)	Yes	(n
13	TEXT	VARCHAR2(4000 BYTE)	Yes	(n

NUM_ROWS	1586197
BLOCKS	20297
AVG_ROW_LEN	86
SAMPLE_SIZE	1586197

COLUMN_NAME	NUM_DISTINCT
TEXT	433
PROFILE_NAME	33383
TIME_EPOCH	1577547
OVERALL	10
TASTE	9
PALATE	9
AROMA	9
APPEARANCE	10
STYLE	104
ABV	530
BREWERID	5840
BEER_ID	66055
BEER_NAME	56857

1.59 mill. rows
Size on disk: 1.5 GB
Segment in db: 156 MB

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Learning Opportunities

- Oracle Text on review text
- Analytical / statistical functions
- Find other beers that matches your taste
- Are there faults in the data?
- How do you calculate a DATE from Unix epoch?

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Add Virtual Column

- This dataset uses UNIX epoch time (seconds since 1970-01-01)
- Easily derive a DATE column:

```
alter table reviews
add review_date date generated always as
(date '1970-01-01' + time_epoch /60/60/24);
```

BICON BUSINESS INTELLIGENCE A PART OF SYSCO

Analyse it

- This is supposed to be fun

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

What is the average score?

What is the longest period a person has been drinking beer every day?

Is Friday really beer day?

What is the most popular style?

Do people get better over time to find good beer?

Do beers with higher ABV get better scores?

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

What is the most reviewed style?

```
select style, count(*)
from beer
group by style
order by 2 desc;
```

STYLE	COUNT(*)
American IPA	117563
American Double / Imperial IPA	85958
American Pale Ale (APA)	63451
Russian Imperial Stout	54109
American Double / Imperial Stout	50698
American Porter	50461
American Amber / Red Ale	45741
Belgian Strong Dark Ale	37724
Fruit / Vegetable Beer	33853
American Strong Ale	31939

BICON BUSINESS INTELLIGENCE A PART OF SYSCO

But if you check the score…

```
select style,round(avg(overall),2) avg_score,
  round(stddev(overall),2) stddev_score
from reviews
group by style
order by 2 desc;
```

Style	Avg Score	Std Dev
Gueuze	4.09	0.64
American Wild Ale	4.09	0.65
Quadrupel (Quad)	4.07	0.63
Lambic - Unblended	4.05	0.66
American Double / Imperial Stout	4.03	0.67
Russian Imperial Stout	4.02	0.64
Weizenbock	4.01	0.6
American Double / Imperial IPA	4	0.64
Flanders Red Ale	3.99	0.68
Eisbock	3.98	0.63

BICON BUSINESS INTELLIGENCE
A PART OF SYSCO

Do Beers with Higher ABV Get Better Scores?

```
select style,round(corr(abv,overall),2)  Correlation,
  count(*) cnt
from reviews
where review_date between date '2011-01-01'
   and date '2011-12-31'
group by style
order by 2 desc;
```

STYLE	CORRELATION	CNT
Chile Beer	0.39	504
Dortmunder / Export Lager	0.35	670
Faro	0.35	184
Vienna Lager	0.32	1189
Bière de Champagne / Bière Brut	0.29	347
Happoshu	0.29	25

BICON BUSINESS INTELLIGENCE
A PART OF SYSCO

Longest Period

```
select profile_name,min(review_day), max(review_day), count(*) days_drinking
from (
  select profile_name, review_day,
    review_day - row_number() over (partition by profile_name order by review_day) lb
  from (
            select profile_name,trunc(review_date) review_day
            from reviews
            group by profile_name, trunc(review_date)
       )
)
group by profile_name,lb
order by days_drinking desc;
```

This method is well explained in KISS series on Analytics, part 3, Mind the Gap

BICON BUSINESS INTELLIGENCE A PART OF SYSCO

The Results Are In

mikesgroove	2008-04-01	2008-07-14	105
atsprings	2010-07-06	2010-10-14	101
cvstrickland	2008-05-12	2008-08-07	88
Daniellobo	2010-01-13	2010-04-02	80
jwinship83	2009-05-22	2009-08-07	78
Zorro	2004-01-27	2004-04-06	71
Gusler	2002-10-06	2002-12-12	68
Jadjunk	2010-12-19	2011-02-23	67
superdedooperboy	2008-08-14	2008-10-18	66

Query above took 8 sec.
A simple inspection to verify:

```
select review_date
from reviews
where profile_name='mikesgroove'
and review_date >= date '2008-04-01'
order by 1;
```

BICON BUSINESS INTELLIGENCE
A PART OF SYSCO

DBMS_FREQUENT_ITEMSET

- A hidden gem from 10g
- Sort of data mining
- Find items that occur together - basket analysis
- Use for taste recommendations

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

From doc

```
DBMS_FREQUENT_ITEMSET.FI_TRANSACTIONAL (
   tranx_cursor          IN    SYSREFCURSOR,
   support_threshold     IN    NUMBER,
   itemset_length_min    IN    NUMBER,
   itemset_length_max    IN    NUMBER,
   including_items       IN    SYS_REFCURSOR DEFAULT NULL,
   excluding_items       IN    SYS_REFCURSOR DEFAULT NULL)
  RETURN TABLE OF ROW (
     itemset [Nested Table of Item Type DERIVED FROM tranx_cursor],
     support        NUMBER,
     length         NUMBER,
     total_tranx    NUMBER);
```

BICON BUSINESS INTELLIGENCE A PART OF SYSCO

Table with three most popular

```
create table popular_style_list
as select profile_name,style,avg_score
from (
  select profile_name,style,round(avg(overall),1) avg_score,
  row_number() over (partition by profile_name order by avg(overall) desc) rn
  from reviews
  group by profile_name,style
)
where rn <=3;
```

The will profile_name serve as the transaction id
The three most popular styles are chosen for each person

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

A new type

```
create or replace type fi_style_nt as table of varchar2(200)
/
```

BICON BUSINESS INTELLIGENCE A PART OF SYSCO

Find most common combinations

```
SELECT CAST(itemset as fi_style_nt) itemset, support, length, total_tranx
   FROM table(DBMS_FREQUENT_ITEMSET.FI_TRANSACTIONAL(
                CURSOR(SELECT profile_name,style
                         FROM popular_style_list),
                0.01,
                2,
                3,
                NULL,
                NULL
                ))
order by support desc;
```

BICON BUSINESS INTELLIGENCE A PART OF SYSCO

Script Output | Query Result

SQL | All Rows Fetched: 3 in 0.074 seconds

	ITEMSET	SUPPORT	LENGTH	TOTAL_TRANX
1	HROUG.FI_STYLE_NT('American Double / Imperial IPA', 'American IPA')	693	2	33383
2	HROUG.FI_STYLE_NT('American IPA', 'American Pale Ale (APA)')	360	2	33383
3	HROUG.FI_STYLE_NT('American Double / Imperial IPA', 'American Double / Imperial Stout')	357	2	33383

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Get recommendations

```
SELECT CAST(itemset as fi_style_nt) itemset, support, length, total_tranx
   FROM table(DBMS_FREQUENT_ITEMSET.FI_TRANSACTIONAL(
                CURSOR(SELECT profile_name,style
                         FROM popular_style_list),
                0.01,
                2,
                3,
                CURSOR(SELECT * FROM table(fi_style_nt
                                           ('American IPA'))),
                NULL
                ))
order by support desc;
```

BICON BUSINESS INTELLIGENCE A PART OF SYSCO

	ITEMSET	SUPPORT	LENGTH	TOTAL_TRANX
1	HROUG.FI_STYLE_NT('American Double / Imperial IPA', 'American IPA')	693	2	33383
2	HROUG.FI_STYLE_NT('American IPA', 'American Pale Ale (APA)')	360	2	33383

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Visualise it

- More fun?

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Do they get better at finding good beer?

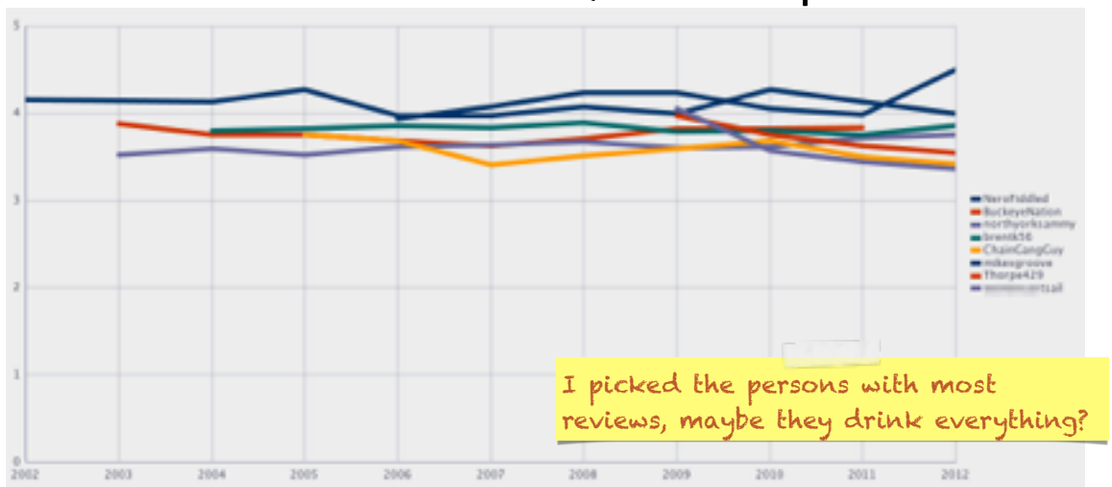
```
select profile_name, to_char(review_date,'yyyy') year
   , round(avg(overall),2) avg_overall_year
from (
       select profile_name, review_date,overall,count(*) over (partition by profile_name) no_of_reviews
       from reviews
       )
where no_of_reviews > 3200
group by profile_name,to_char(review_date,'yyyy')
order by profile_name,year;
```

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Not so easy to see

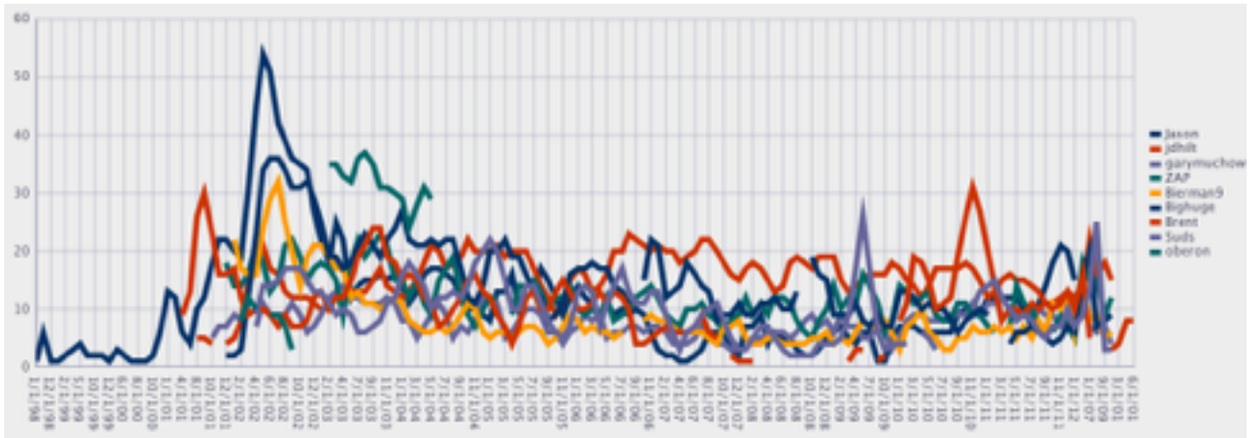
BuckeyeNation	2003	3.88
BuckeyeNation	2004	3.75
BuckeyeNation	2005	3.75
BuckeyeNation	2006	3.69
BuckeyeNation	2007	3.63
BuckeyeNation	2008	3.71
BuckeyeNation	2009	3.82
BuckeyeNation	2010	3.82
BuckeyeNation	2011	3.84
ChainGangGuy	2005	3.76
ChainGangGuy	2006	3.69
ChainGangGuy	2007	3.41
ChainGangGuy	2008	3.51
ChainGangGuy	2009	3.6

Line Chart in SQL Developer



BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Number of styles over time



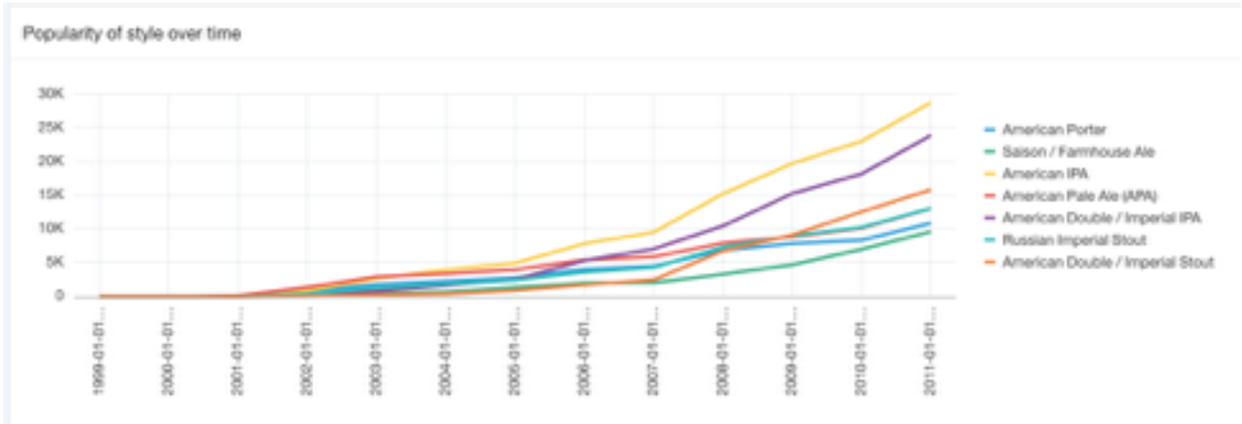
BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

```
select *
from (
   select profile_name, review_month,
     round(avg(no_of_styles) over (partition by profile_name order by review_month
     range between interval '2' month preceding and current row)) moving_avg,
     dense_Rank() over (order by months_drinking desc) rn
   from (
      select profile_name,trunc(review_date,'MM') review_month, count(distinct style) no_of_styles,
        count(*) over (partition by profile_name) months_drinking
      from reviews
      group by profile_name,trunc(review_date,'MM'))
   )
where rn <= 10
order by rn,2 ;
```

Charts in Apex

- Latest version has several option for nice charts
- Check out Sample App for charts in APEX 5.1

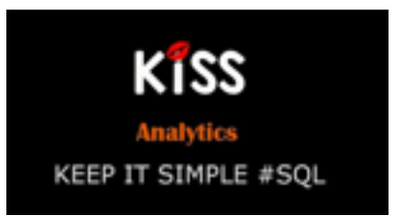
BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO



BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Conclusion

- XE is powerful for many purposes
- Open data lets you learn on realistic data
- Remember to have fun
- Limits sometimes make it easier to prioritise



BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO


#OOW15
KISS
MY
ASS
KEEP IT SIMPLY #SQL
MAKE YOUR
APPLICATIONS SMARTER in SQL
Credits to Connor McDonald
BICON
BUSINESS INTELLIGENCE

A new world for DBAs

- Cloud, autonomous database…?

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

Some trends

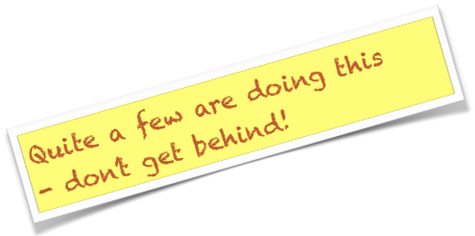
- Cloud reality shows the need for new skills
- Autonomous database might work - advances in AI
- DBA spends less time on tedious tasks
- … and more on what?

I'm done with patching, backup, extending data files

BICON BUSINESS INTELLIGENCE A PART OF SYSCO

We need to improve on

- Security
- Data quality
- Integration
- Analysis



BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO

DBA should learn more

- Data governance
- Data modelling
- Some coding
- Security
- How to utilise the database better

Are DBAs too busy helping others?
Need to reserve time for learning

BICON
BUSINESS INTELLIGENCE
A PART OF SYSCO