Data Quality Dashboard: Leveraging APEX with Oracle Locator to improve Business Decision Making

Simon Greener,
The Spatial DB Advisor
&
Brendan Soustal,
GIS Manager,
Md Coast Water

Business Decision Making

- Quality decision making requires quality data.
- New data being captured into today's databases should be of good quality due to automated capture techniques (in-field Personal Digital Assistants - PDAs - etc).
- But older data can often be of poor or dubious quality due to lack of data integrity constraints in previous systems.

• If a business is making capital expenditure decisions for say, renewing older assets, its needs the data on which these decisions are made to be accurate in terms of spatial consistency and attribute consistency.

Importance of Data Quality

Reporting

- A data quality report is a very, very good way of presenting to engineers and managers the importance of funding low level data quality updates.

- It provides metrics for financial planning that can be used for:

  - Budgeting geo-spatial data infrastructure expenditure

  AND

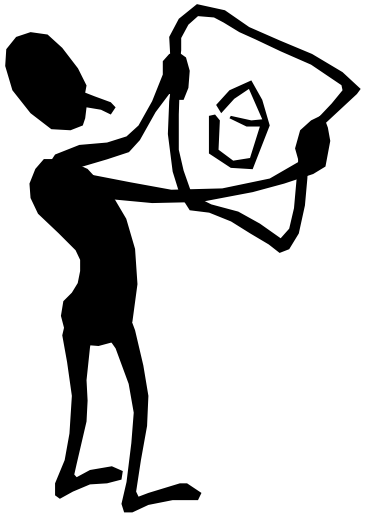
  - Software purchase.

- Because you can demonstrate direct connection to the business profitability of that expenditure.

Simplicity of Output

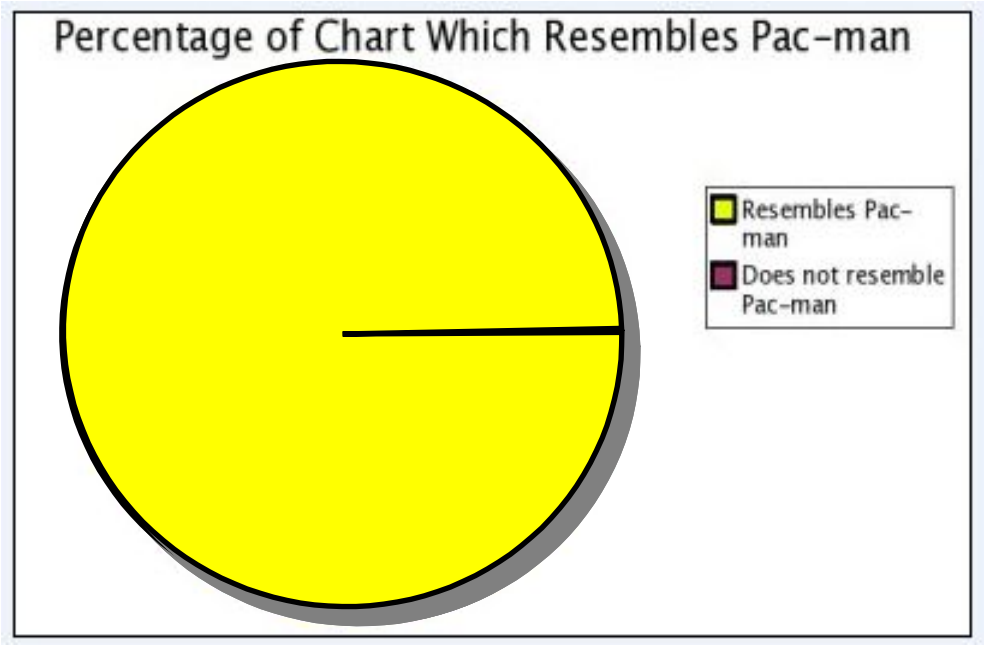
- Presenting reports to upper management requires one to encapsulate the message as simply as possible.

- This is why so many business intelligence "dashboards" make heavy use of charts and graphics.



Or to put it more simply...

- "The higher up the management chain the dumber the reports should be..."



Benchmarking and Output

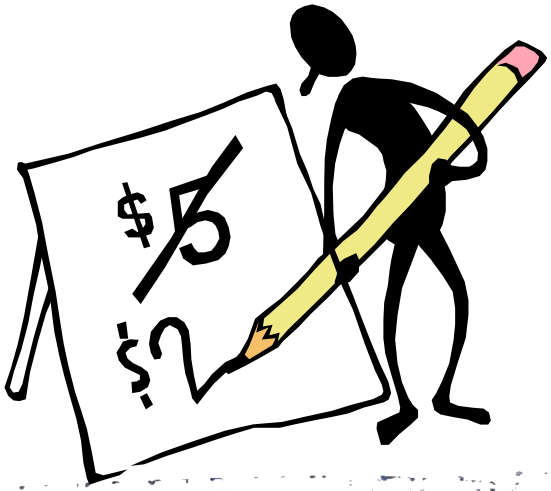
"These dashboard components should be coupled to, or based on, "best practice" metrics that:

-Enable cross-business benchmarking.

- Outputs can be:
  - Delivered online (dynamically) or
  - Hardcopy (eg Adobe PDF)
- And used in monthly "information management" meetings.

Cost to Implement

- There are many products available for constructing Business Intelligence (BI) "dashboards" eg Oracle Business Intelligence Suite.
- New products that address spatial data quality such as Radius Studio are starting to appear.
- But Oracle's free Application Express coupled with Locator is an excellent, low-cost initial solution to show what is possible.



Apex: Reports vs. Graphics

- APEX can output tabular reports (exportable to Spreadsheets) or as graphics.

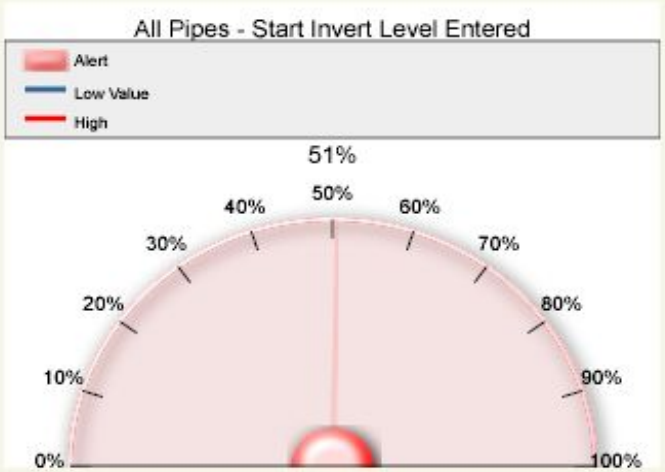
- The table is useful, but the graphic much better when you want to get the point across, by making it red, if the level is below tolerance it really stands out

All Pipe End Invert Exists

End Invert Exists	Percentage	Total Pipes
10872	50.7	21441

1 - 1

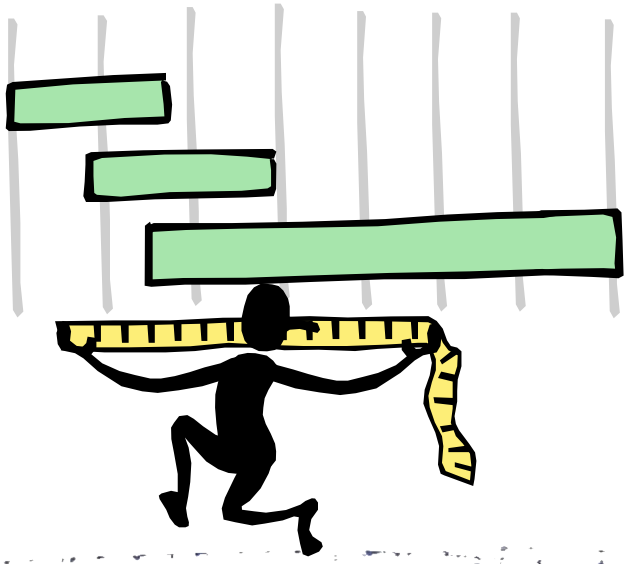
All Pipe End Invert Level Entered Chart



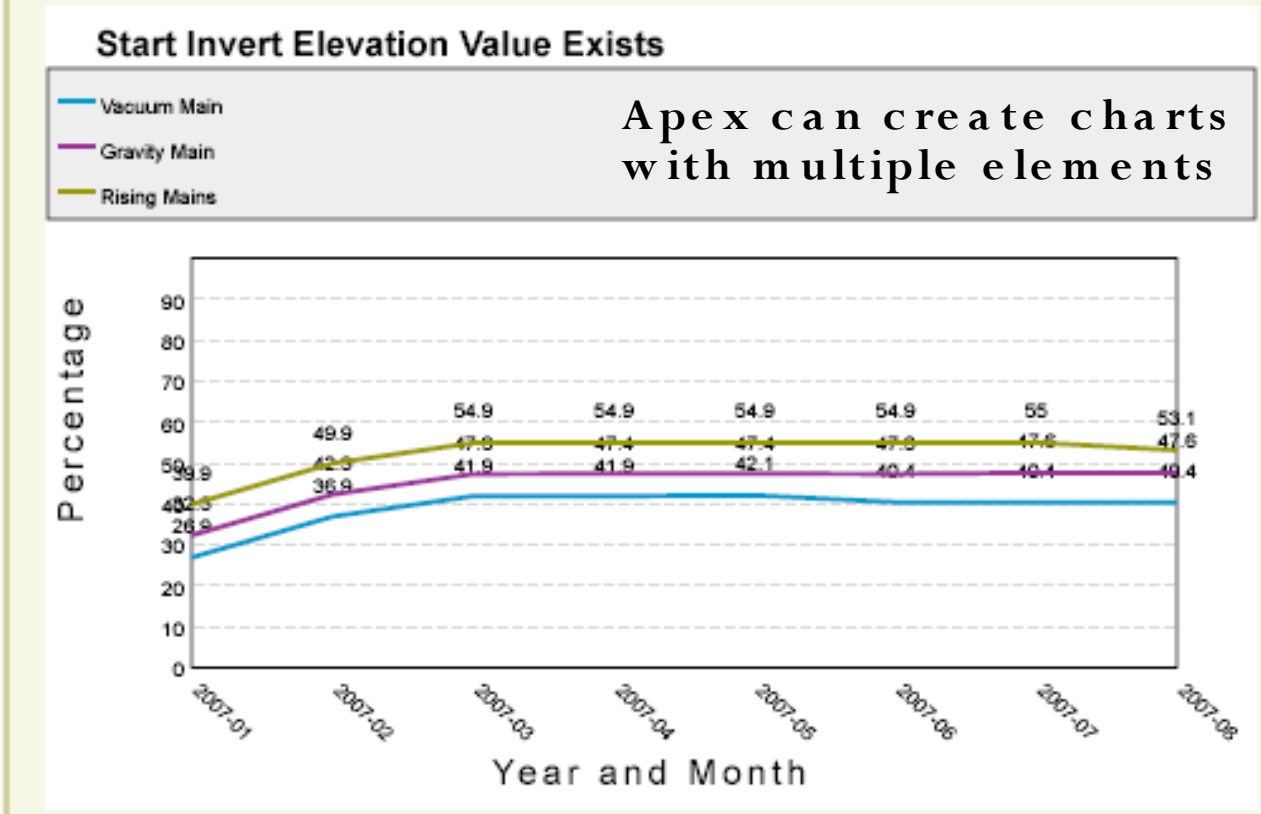
Scenario 1 - Tracking data quality over time

Upper management would like to know how data quality (DQ) improvement initiatives for older data are progressing in a financial year.

- Allows accurate measure of what work has been done, what should be targeted.
- Immediate "Key Result Area" for GIS Manager for a financial year is to improve DQ in the sewer network.
- Need to examine Sewer network "wholistically" by its 3 main elements:
  - Gravity Mains
  - Rising Mains
  - Vacuum Mains



APEX Chart


Start Invert Elevation Entry over Time
Start Invert Elevation Value Exists
Vacuum Main
Gravity Main
Rising Mains
Apex can create charts with multiple elements
Percentage
90
80
70
60
50
40
30
20
10
0
49.9
54.9
54.9
54.9
54.9
55
53.1
47.6
42.3
41.9
41.9
42.1
36.9
26.9
2007-01
2007-02
2007-03
2007-04
2007-05
2007-06
2007-07
2007-08
Year and Month

APEX – Multi-line Chart SQL

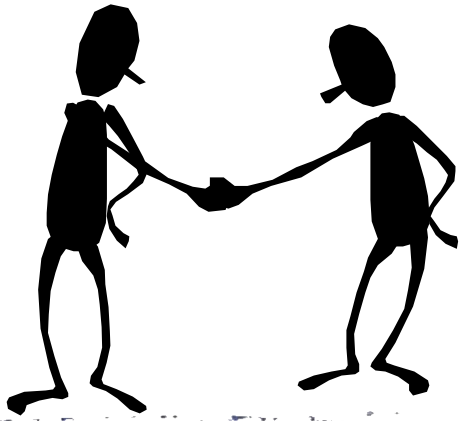
```
select NULL Link,
       year_month label,
       start_percent value
from (select year_month,
             start_percent
        from apex_pipe_stats
       where pipe_network = 'Sewer'
         and Pipe_element = 'Vacuum Main'
         and Pipe_Measure = 'InvElev'
     order by 1 )

select NULL Link,
       year_month label,
       start_percent value
from (select year_month,
             start_percent
        from apex_pipe_stats
       where pipe_network = 'Sewer'
         and Pipe_element = 'Gravity Main'
         and Pipe_Measure = 'InvElev'
     order by 1 )

select NULL Link,
       year_month label,
       start_percent value
from (select year_month,
             start_percent
        from apex_pipe_stats
       where pipe_network = 'Sewer'
         and Pipe_element = 'Rising Main'
         and Pipe_Measure = 'InvElev'
     order by 1 )
```

Benefits

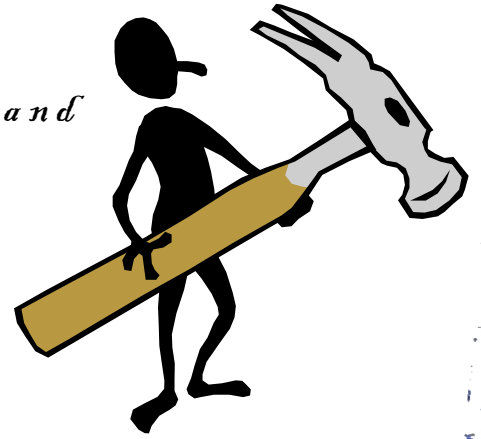
- Multiple constituent elements in network presented as a single group (as it is managed)
- Graph over time enables periodic reporting eg quarterly
- Graph can show effects of multiple initiatives



Scenario 2 - Refurbishment

- Upper management wishes to understand the capital expenditure that would be required to refurbish current infrastructure.
- The calculation requires one to measure the current material/pipe material in the network.
- This can be viewed in two ways:
  - As a % of the total number of pipes and
  - As a % by total length.

The difference can be important!



Scenario 2 – Refurbishment Chart

Vacuum Main Material Percentage Length

Material	Length	Percentage
PE	15270.8	77.5
PVC-U	4434.8	22.5
Total	19705.6	100

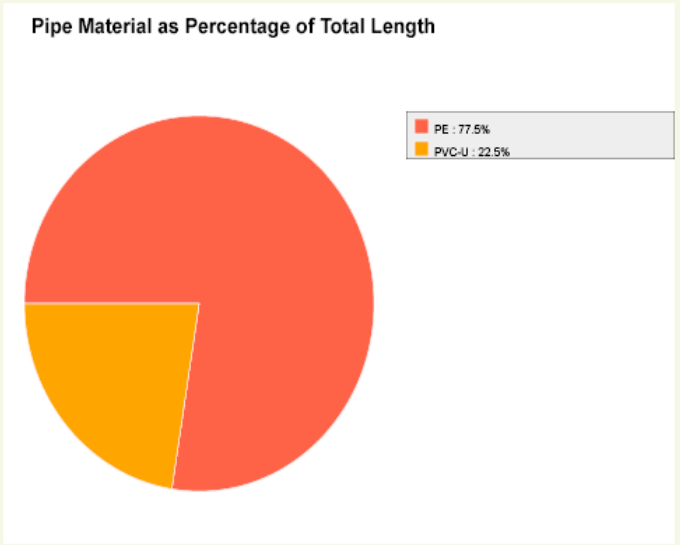
Export to CSV

Vacuum Main Material Percentage Number

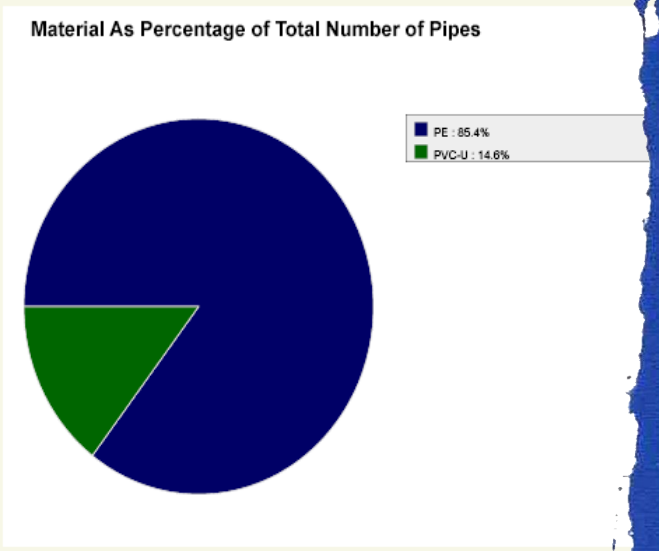
Material	Count	Percentage
PE	1219	85.4
PVC-U	209	14.6
Total	1428	100

Export to CSV

Vacuum Main Material Percentage Length



Vacuum Main Material Percentage Number



Scenario 2 – Refurbishment SQL

```
SELECT a.Pipe_Link,   /* SQL for % by Total Network Length */
       a.Pipe_Material,
       ROUND( (a.matrl_length/b.total_length) * 100,1 ) AS Percentage_By_Length
  FROM (SELECT ilv.pipe_matrl       AS Pipe_Link,
               ilv.pipe_matrl       AS Pipe_Material,
               SUM(SDO_GEOM.SDO_LENGTH(ilv.geometry,0.5)) AS matrl_length
          FROM SP_SewVPipe ilv
         GROUP BY ilv.pipe_matrl) a,
       (SELECT SUM(SDO_GEOM.SDO_LENGTH(tl.geometry,0.5)) AS total_length
          FROM SP_SewVPipe tl ) b

SELECT a.Pipe_Link,   /* SQL for % by Total Number of Pipes */
       a.Pipe_Material,
       ROUND( (a.matrl_count/b.total_count) * 100,1 ) AS Percentage_By_Length
  FROM (SELECT ilv.pipe_matrl AS Pipe_Link,
               ilv.pipe_matrl AS Pipe_Material,
               COUNT(*)   AS matrl_count
          FROM SP_SewVPipe ilv
         GROUP BY ilv.pipe_matrl) a,
       (SELECT COUNT(*) AS total_count
          FROM sp_SewVPipe) b
```

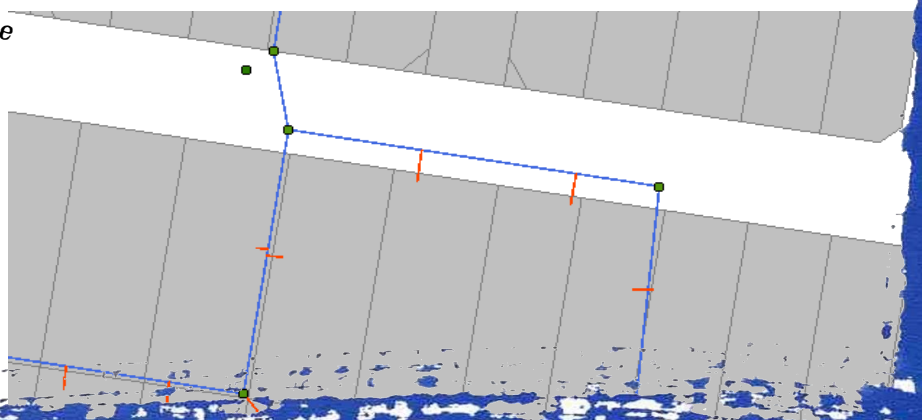
Scenario 3 – Service Delivery

- **Business Problem**

- Efficient service delivery, notification of maintenance activities, billing etc requires water and sewerage services to be associated with land parcels/buildings.

- For example, if work is being considered on a particular sewer gravity main, then one would like to produce a "mail out" to all households that are serviced by that main.

- But how can we be sure that connectors are servicing the property?



Spatial Referential Integrity

- In the following example, the start of the (sewer) service should be connected a gravity pipe, and end of service should be within a parcel, and that the parcel's primary key that is recorded as a foreign key in the connector is in fact correct.

- This is an example of "spatial referential integrity".

- We only want to report those for which this relationship is not correct.

(The following SQL does not test all aspects of this requirement.)

Spatial Connections Chart

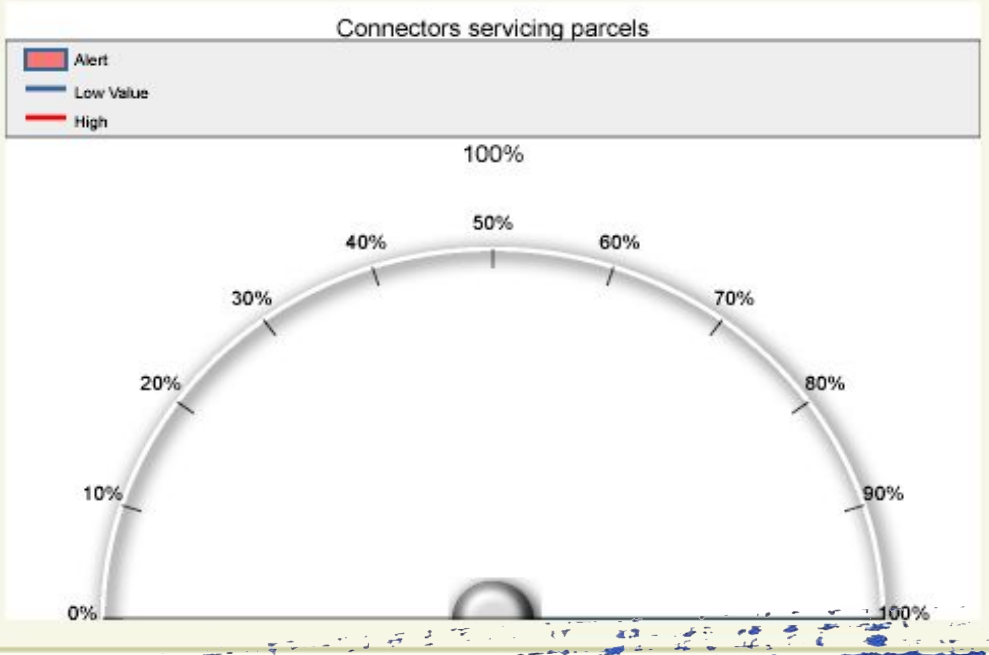
Operational Service Parcel Connections

No Parcel	No Parcel Percentage	Has Parcel	Has Parcel Percentage	Total Connections
89	.3	30563	99.7	30652

Export to CSV

1 - 1

Operational Service Parcel Connections

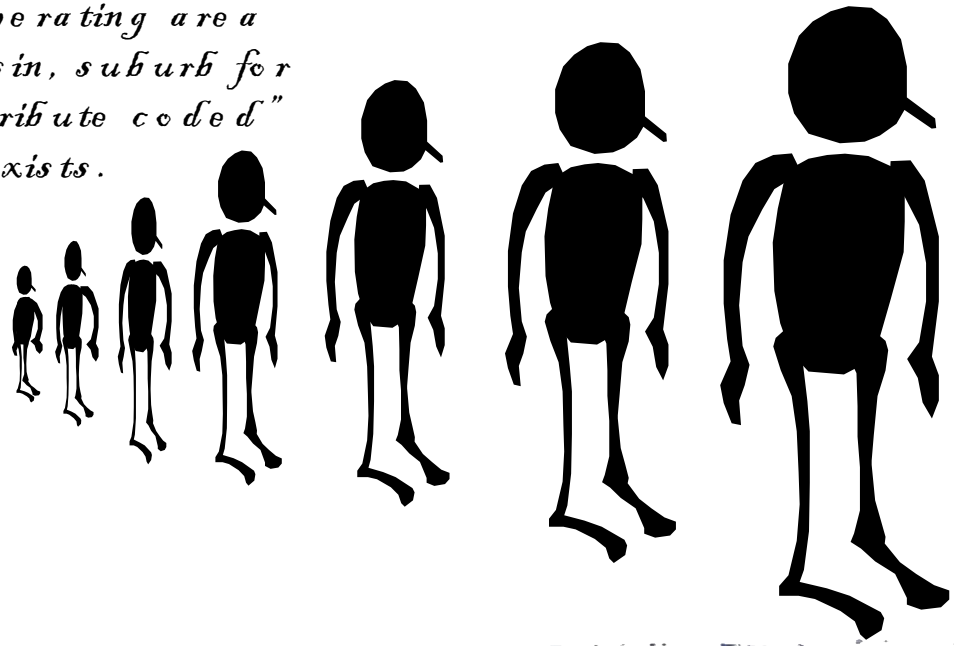


Spatial Connections - SQL

```
SELECT ss.GID,
       ss.prcl_gid
 FROM all_sdo_geom_metadata asgm,
       sp_sewserv ss
       INNER JOIN
       sp_parcel sp
       ON ( sp.gid = ss.prcl_gid )
 WHERE (     asgm.owner       = 'MCW_GIS'
         AND asgm.table_name  = 'SP_SEWSERV'
         AND asgm.column_name = 'GEOMETRY' )
   AND ss.sstatus = 'OPERATING'
   AND ss.prcl_gid is not null
   AND SDO_GEOM.RELATE(sp.geometry,
                       asgm.diminfo,
                       'mask=DETERMINE',
                       ss.geometry,
                       asgm.diminfo)
        = 'DISJOINT';
```

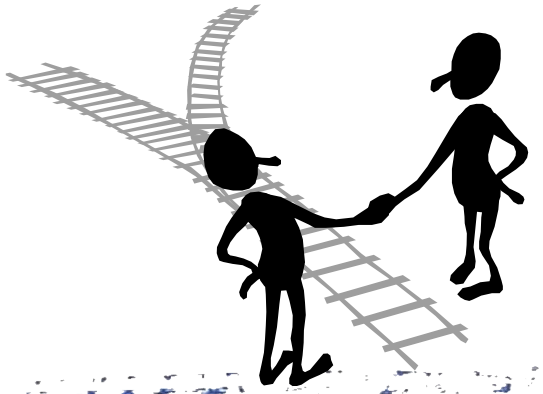
Scenario 3 : Conclusion

- Note that the output of a spatial query is not necessarily a map!
  - Apex can handle spatial because Oracle can!
  - Reporting of spatial data quality could occur by "rollup" by reporting / operating area
    eg sewer basin, suburb for which no "attribute coded" foreign key exists.

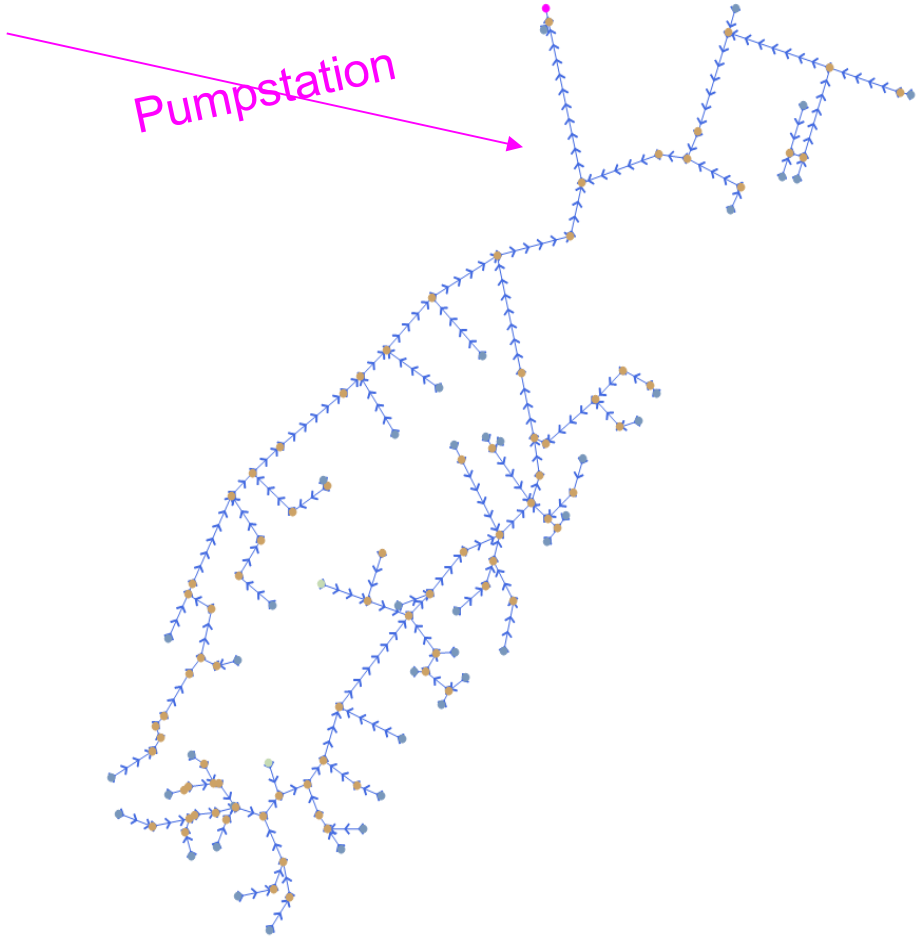


Scenario 4 : Network Tracing

- Problem : Should we upgrade a particular pump in a sub-network?
  - We need to look at spatial data quality across a whole network or sub-network not just a single element.
    - Is coding of a single invert elevation on one node correct in relation to the nodes before and after it?
  - We can configure Apex reports to consume data from a simple hierarchical query!



A sub-network



A simple network trace ...

```
SELECT wp.pipe_id,.
       wp.start_node,.
       wp.end_node,.
       wp.pipe_diameter,.
       wp.pipe_material,.
       SDO_GEOM.SDO_LENGTH(wp.geometry,0.01).
       AS pipe_length,.
       LEVEL.
 FROM sp_sewgpipe wp.
    START WITH OID = .
    (SELECT sgj.oid.
    FROM sp_sewnode snj.
 INNER JOIN.
 sp_sewgpipe sgj.
 ON ( sgj.end_node = snj.node_id).
 WHERE snj.node_type = 'PUMPSTATION'.).

 CONNECT BY NOCYCLE PRIOR start_node = end_node;
```

Result ...

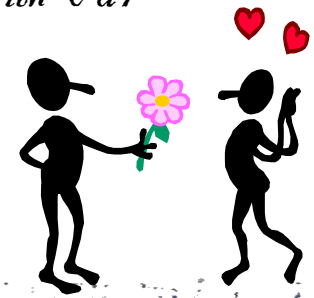
PIPE_ID	START_NODE	END_NODE	PIPE_DIAMETER	PIPE_MATERIAL	PIPE_LENGTH	LEVEL
18918	18024	12210	150	PVC-U	11.6317825	1
13715	18028	18024	150	PVC-U	6.89768229	2
18917	11204	18024	150	PVC-U	133.368761	2
13445	11202	11204	150	PVC-U	71.2474506	3
13437	11201	11202	150	PVC-U	24.7904271	4
13438	11200	11201	150	PVC-U	23.6856216	5
.....						
22642	22844	22843	150	PVC-U	27.0856449	10
22641	21548	22844	150	PVC-U	8.41058057	11

127 rows selected.

We can pop this into Apex as a simple HTML report (with download to Excel)!

Spicing it up...

- The hierarchical query is great but we can "spice up" the report by including Oracle's really cool SQL99 based "analytic" functions.
  - Eg GROUP BY CUBE(), ROLLUP() etc
- But included in the SQL99 standard are set of "windowing" functions that allow us to calculate "moving" averages etc and not just an average at the end of each grouping set!
- The following query shows what can be done with our network...



Computing missing values, etc.

```
SELECT ROWNUM as row_num,
       LEVEL as row_level,
       wp.end_invelev,
       wp.start_invelev,
       wp.pipe_gradient,
       SDO_GEOM.SDO_LENGTH(wp.geometry,0.01) as pipe_length,
       CASE WHEN wp.end_invelev   is not null
             AND wp.start_invelev is not null
             AND wp.pipe_length   is not null
            THEN CASE WHEN ( wp.start_invelev - wp.end_invelev )
                           <> 0
                      THEN ROUND(wp.pipe_length /
                                 (wp.start_invelev - wp.end_invelev),3)
                      ELSE 0
                 END
            ELSE
               NULL
            END AS comp_gradient
  FROM sp_sewgpipe wp
 START WITH OID = (SELECT wp.oid
                     FROM sp_sewnode snj
                          INNER JOIN
                          sp_sewgpipe wp
                          ON ( wp.end_node = snj.node_id)
                    WHERE snj.node_type = 'PUMPSTATION')
 CONNECT BY NOCYCLE PRIOR start_node = end_node;
```

Now use "windowing" functions to compute rolling total length (SUM) and average gradient (AVG)...

```
Prompt Use row_level to do some "pretty printing"...

SELECT LPAD('_', row_level - 1, '_')  ||
       ' Inv(End,Srt)->(' ||
         DECODE(h.end_invelev,  NULL,'NULL',h.end_invelev) ||
       ','                          ||
         DECODE(h.start_invelev,NULL,'NULL',h.start_invelev) ||
       ') Grdnt(DB,Qry)->(' ||
         DECODE(h.pipe_gradient,NULL,'NULL',h.pipe_gradient) ||
       ','                          ||
         DECODE(h.comp_gradient,NULL,'NULL',h.pipe_gradient) ||
       ') MvTotLen('        ||
         ROUND( SUM(h.pipe_length)
                OVER (ORDER BY row_num ROWS UNBOUNDED PRECEDING)
               ,3) ||
       ') MvAvgGrdnt('         ||
         ROUND( AVG(h.comp_gradient)
                OVER (ORDER BY row_num ROWS UNBOUNDED PRECEDING),3) ||
       ')'
       As Information
FROM (
       SELECT ... statement from previous slide
     ) h;
```

Result...

```
INFORMATION
--------------------------------------------------------------------------------
Inv(End,Srt)->(0,-.685) Grdnt(DB,Qry)->(0,0) MvTotLen(401.296) MvAvgGrdnt(-16.981)
_ Inv(End,Srt)->(-.685,0) Grdnt(DB,Qry)->(0,0) MvTotLen(639.267) MvAvgGrdnt(-3.456)
_ Inv(End,Srt)->(-.685,-3.1) Grdnt(DB,Qry)->(0,0) MvTotLen(5240.489) MvAvgGrdnt(-20.712)
__ Inv(End,Srt)->(-3.1,-2.07) Grdnt(DB,Qry)->(0,0) MvTotLen(7698.526) MvAvgGrdnt(1.759)
___ Inv(End,Srt)->(-1.88,-2.07) Grdnt(DB,Qry)->(0,0) MvTotLen(8553.796) MvAvgGrdnt(-24.688)
____ Inv(End,Srt)->(-1.88,-1.64) Grdnt(DB,Qry)->(0,0) MvTotLen(9370.945) MvAvgGrdnt(-4.125)
_____ Inv(End,Srt)->(-1.64,-1.04) Grdnt(DB,Qry)->(0,0) MvTotLen(12291.457) MvAvgGrdnt(16.62)
______ Inv(End,Srt)->(-1.04,0) Grdnt(DB,Qry)->(0,0) MvTotLen(13020.131) MvAvgGrdnt(17.081)
______ Inv(End,Srt)->(-1.04,-.26) Grdnt(DB,Qry)->(0,0) MvTotLen(16243.283) MvAvgGrdnt(28.491)
...
_________ Inv(End,Srt)->(0,0) Grdnt(DB,Qry)->(0,0) MvTotLen(141705.275) MvAvgGrdnt(58.46)
__________ Inv(End,Srt)->(0,0) Grdnt(DB,Qry)->(0,0) MvTotLen(141995.437) MvAvgGrdnt(57.999)

127 rows selected.
```

Conclusion

- What we have tried to show is that:
  - Application Express + Simple Spatial/Attribute SQL = An effective, initial, data quality Dashboard
  - Will help GIS Managers deliver quality data into their business's decision making, and
  - Provide measureable input into the budgetry process.
  - Increase business Return On Investment for its spatial data, database and software investment.