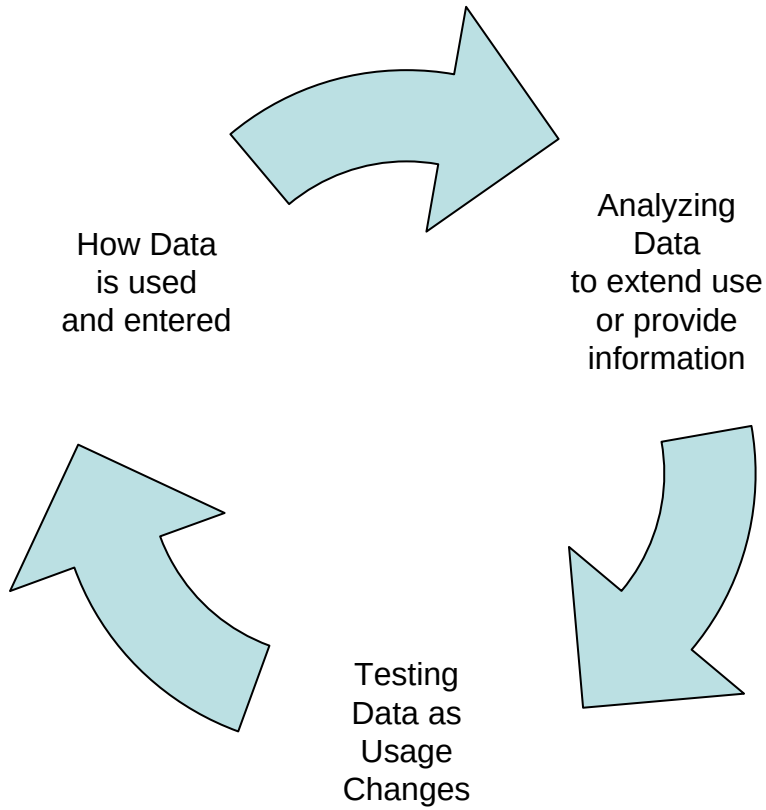
How Long is Long Enough?

Using Statistics to Determine Optimum Field Length

Suzanne Michelle, June 2009

Covering …

- A little background on my team's project
- Considerations for summarizing data
- Data analysis
- Analysis and testing opportunities

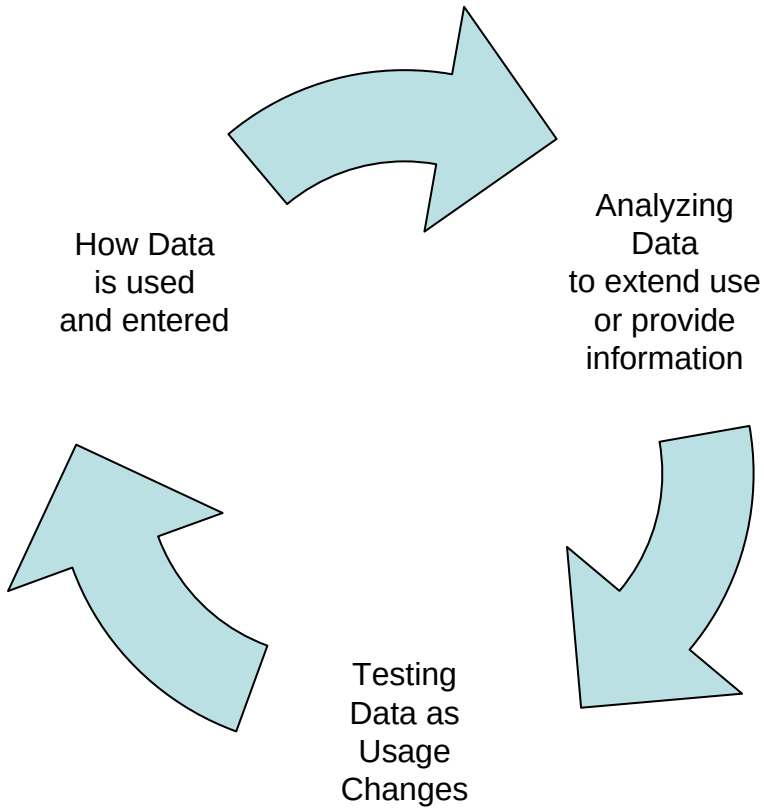


Me?

- Worked with VisiCalc about 1981,
  told by my Graduate Supervisor to learn it
  and teach my fellow classmates how to use it
- Created floppy-disk sized set of relational
  Lotus123 spreadsheets for a city budget
  charge-back system in 1986
- Developed and managed various systems
  for Morgan Bank, W.R.Grace, Perdue Foods,
  and Hershey Foods, and also steel and roofing
  materials plants, among others.
- Began working for NYC Transit in 1994
  on the Unified General Order System (UGOS)

Covering …

- **A little background on my team's project**
- Considerations for summarizing data
- Data analysis
- Analysis and testing opportunities



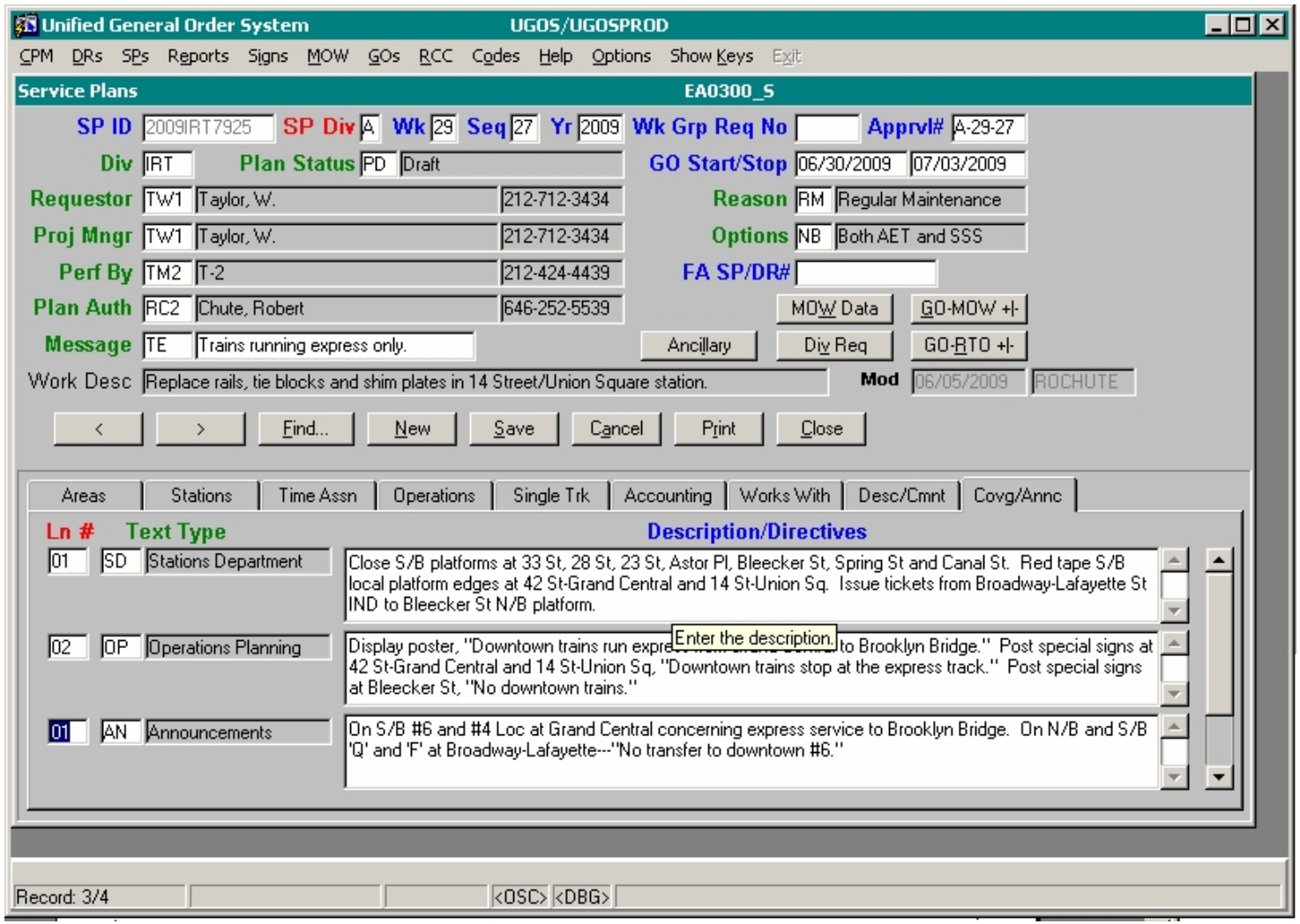
General Orders?

- Run NYC Transit train operations
- Coordinate diversions *from* normal service
- Are planned 8 weeks to 5 years ahead of time depending on work / coordination involved
- Are visible to the riding public as service notice signs (for 1 or more actual operations)
- Affect / are affected by any service disruptions
- Reflect operation dependencies
- Interface with other systems (e.g., Accounting)

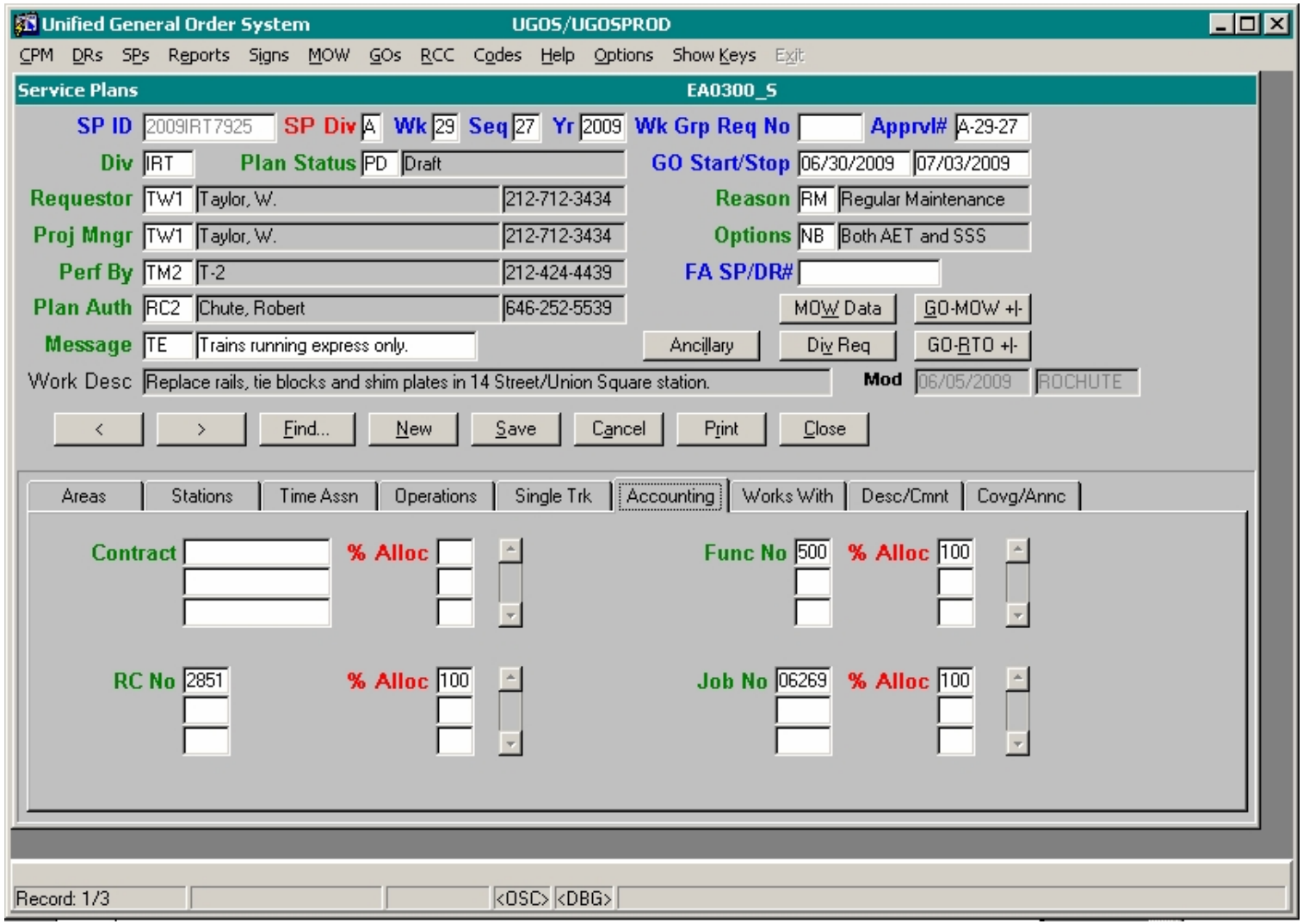
UGOS (You-Goes)

- A Calendar System of Work
- A Decision Support System, with History
- A Coordination Tool, between workgroups and departments
- A Reporting Tool
- Forms6i / 10g DB via MS TS / Citrix

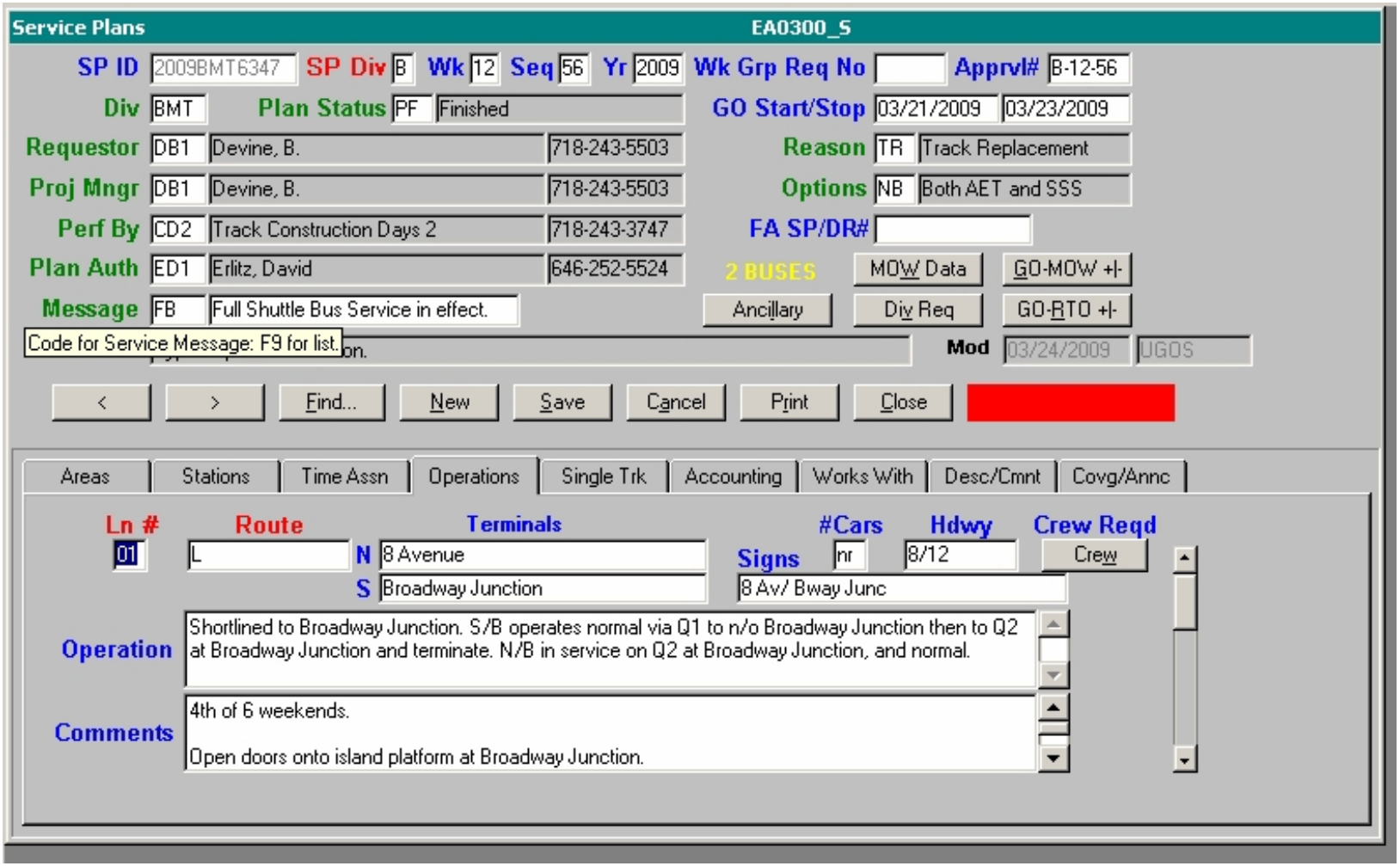
Descriptive Text ...



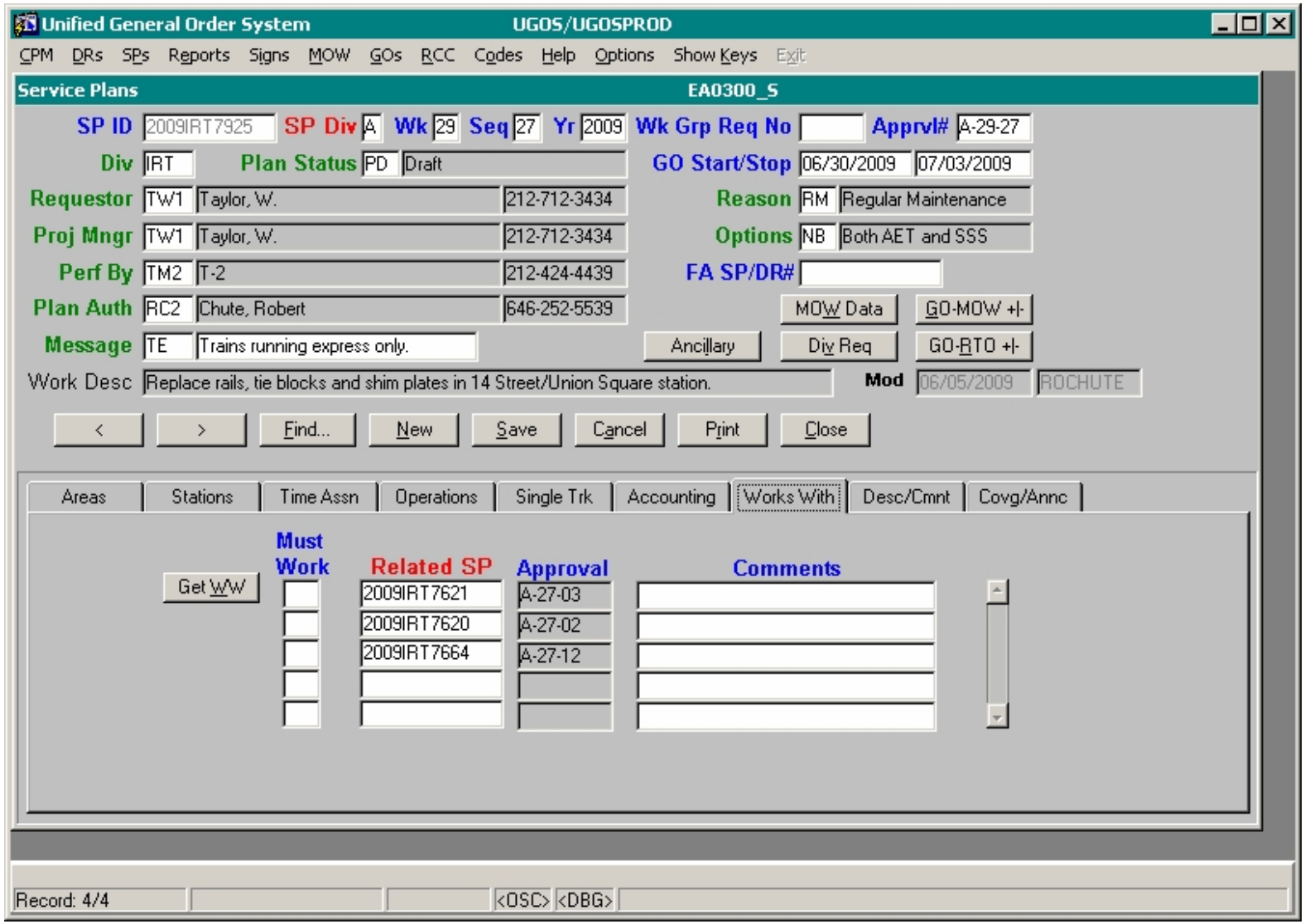
Accounting …


Unified General Order System
UGOS/UGOSPROD
CPM DRs SPs Reports Signs MOW GOs RCC Codes Help Options Show Keys Exit
Service Plans
EA0300_S
SP ID 2009IRT7925 SP Div A Wk 29 Seq 27 Yr 2009 Wk Grp Req No Apprvl# A-29-27
Div IRT Plan Status PD Draft GO Start/Stop 06/30/2009 07/03/2009
Requestor TW1 Taylor, W. 212-712-3434 Reason RM Regular Maintenance
Proj Mngr TW1 Taylor, W. 212-712-3434 Options NB Both AET and SSS
Perf By TM2 T-2 212-424-4439 FA SP/DR#
Plan Auth RC2 Chute, Robert 646-252-5539 MOW Data GO-MOW +|-
Message TE Trains running express only. Ancillary Div Req GO-RTO +|-
Work Desc Replace rails, tie blocks and shim plates in 14 Street/Union Square station. Mod 06/05/2009 ROCHUTE
< > Find... New Save Cancel Print Close
Areas Stations Time Assn Operations Single Trk Accounting Works With Desc/Cmnt Covg/Annc
Contract % Alloc
Func No 500 % Alloc 100
RC No 2851 % Alloc 100
Job No 06269 % Alloc 100
Record: 1/3
<OSC> <DBG>

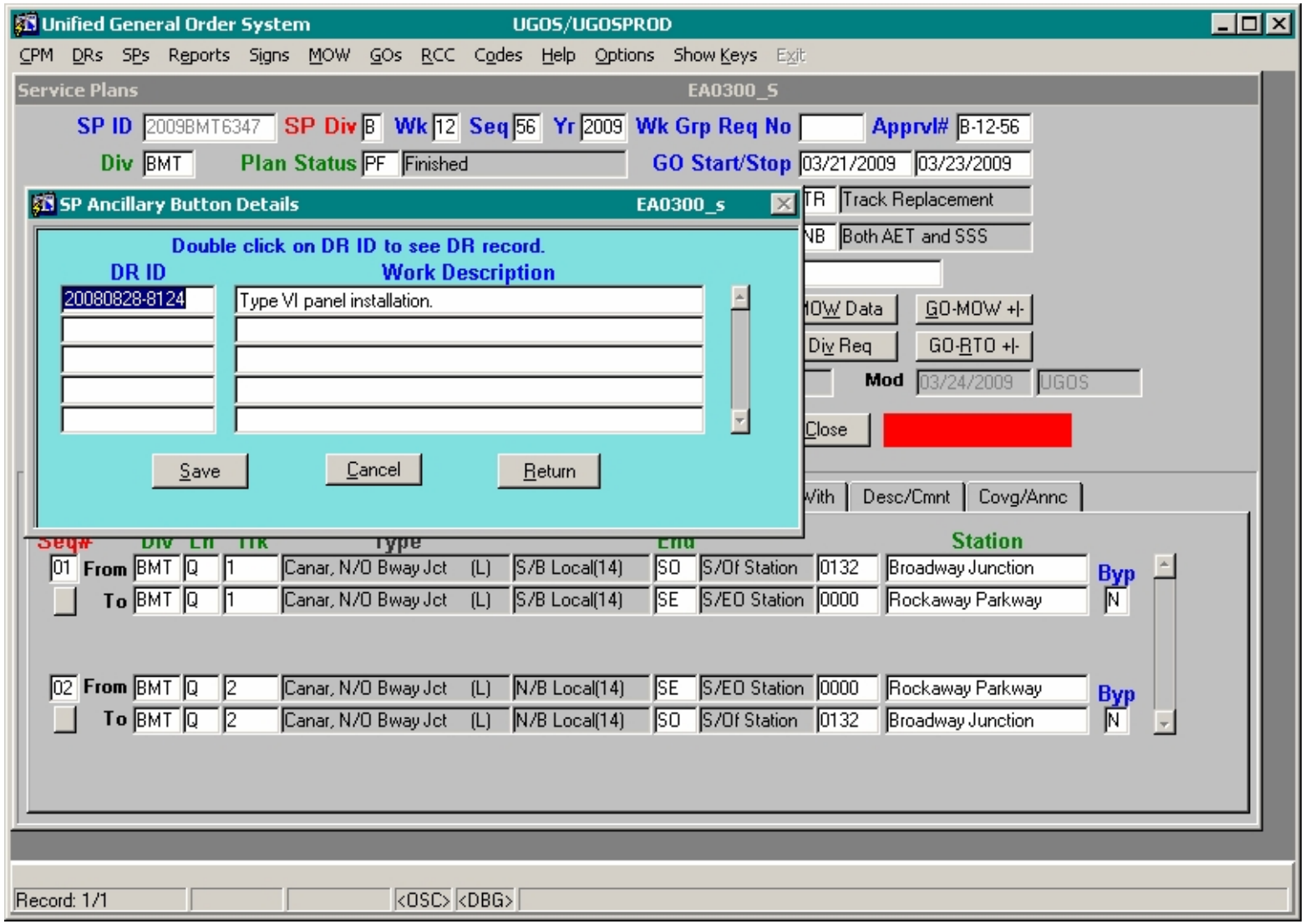
Operations …


Service Plans
EA0300_S
SP ID 2009BMT6347 SP Div B Wk 12 Seq 56 Yr 2009 Wk Grp Req No Apprvl# B-12-56
Div BMT Plan Status PF Finished GO Start/Stop 03/21/2009 03/23/2009
Requestor DB1 Devine, B. 718-243-5503 Reason TR Track Replacement
Proj Mngr DB1 Devine, B. 718-243-5503 Options NB Both AET and SSS
Perf By CD2 Track Construction Days 2 718-243-3747 FA SP/DR#
Plan Auth ED1 Erlitz, David 646-252-5524 2 BUSES MOW Data GO-MOW +|-
Message FB Full Shuttle Bus Service in effect. Ancillary Div Req GO-RTO +|-
Code for Service Message: F9 for list.
Mod 03/24/2009 UGOS
< > Find... New Save Cancel Print Close
Areas Stations Time Assn Operations Single Trk Accounting Works With Desc/Cmnt Covg/Annc
Ln # Route Terminals #Cars Hdwy Crew Reqd
01 L N 8 Avenue Signs nr 8/12 Crew
S Broadway Junction 8 Av/ Bway Junc
Operation Shortlined to Broadway Junction. S/B operates normal via Q1 to n/o Broadway Junction then to Q2 at Broadway Junction and terminate. N/B in service on Q2 at Broadway Junction, and normal.
Comments 4th of 6 weekends.
Open doors onto island platform at Broadway Junction.

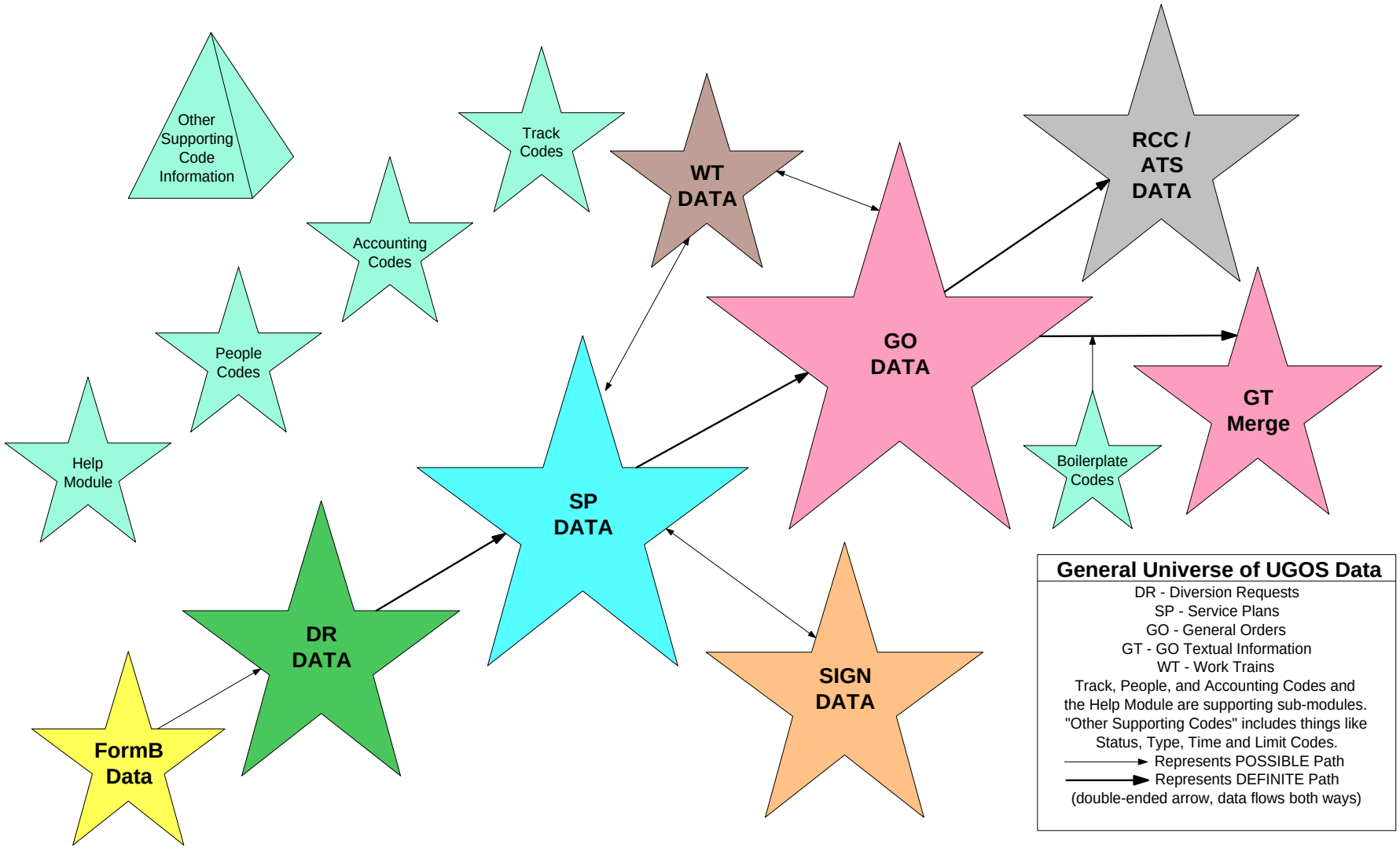
Works With (Siblings) …



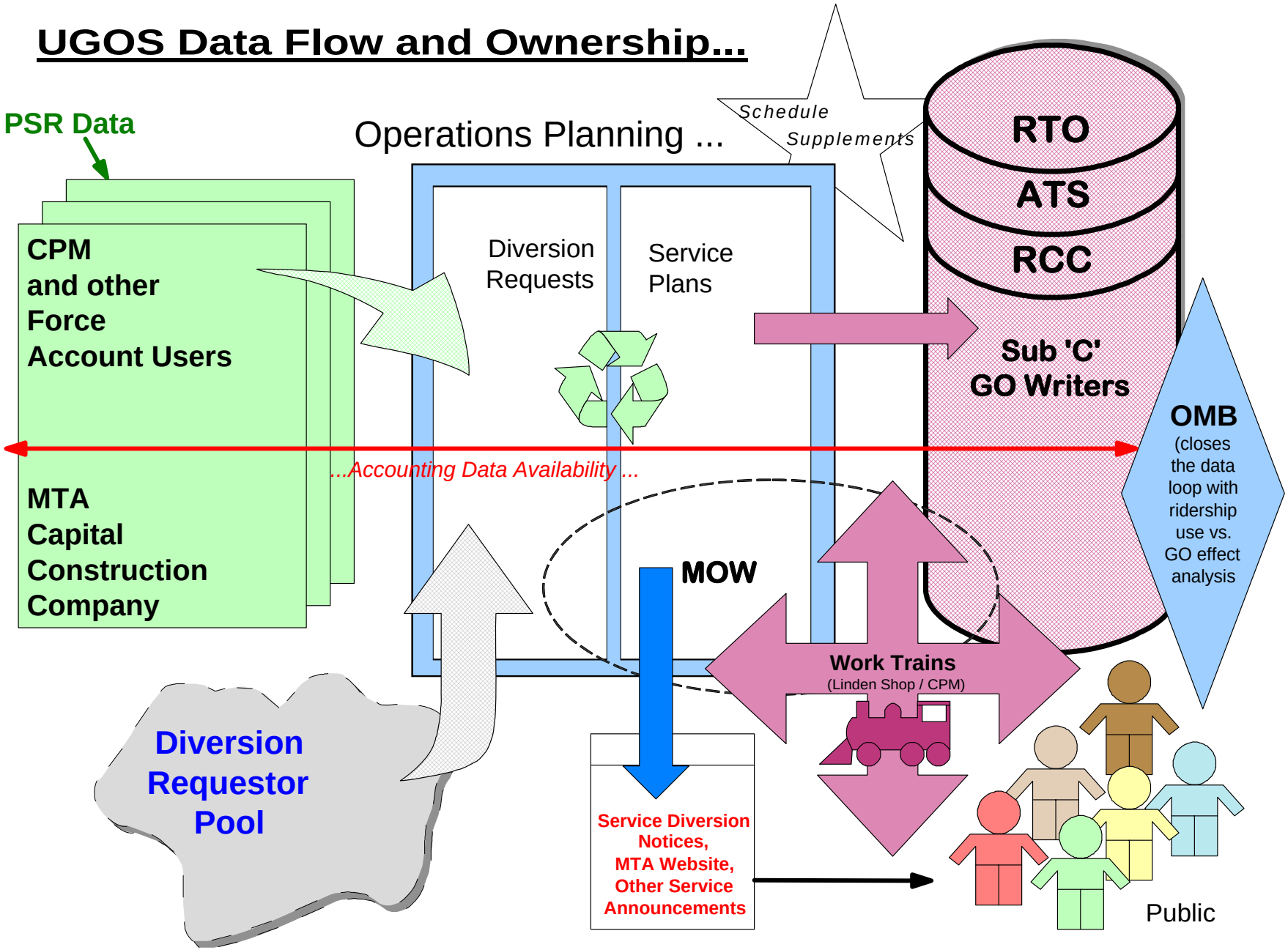
Cross-Referenced data …



UGOS Star Schemas …

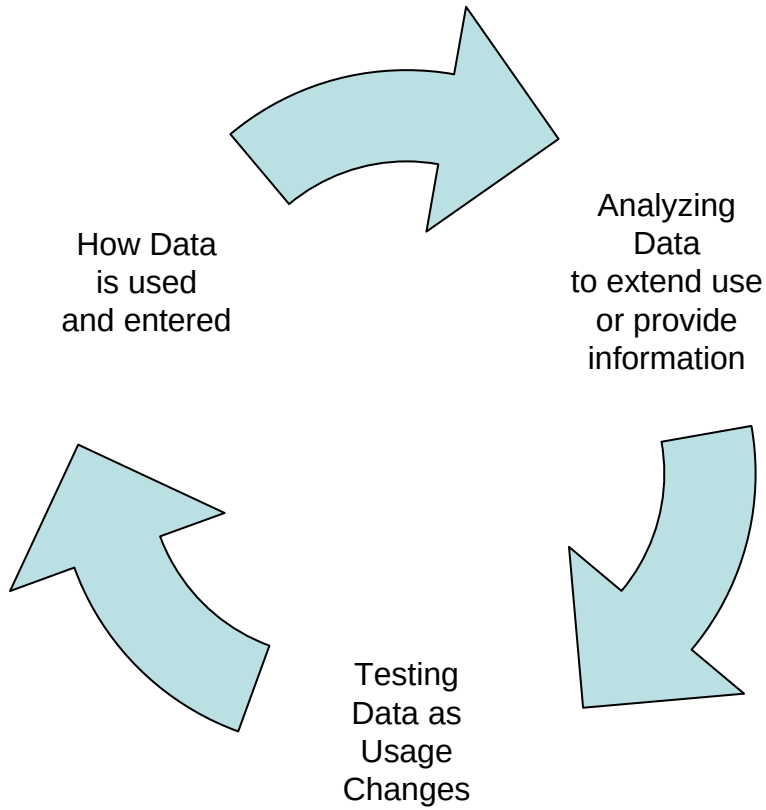


UGOS Data Flow and Ownership...



Covering …

- A little background on my team's project
- **Considerations for summarizing data**
- Data analysis
- Analysis and testing opportunities



Three Questions …

- How to better summarize disparate data types and sources?
- How to allow on-the-fly data extraction (varying columns, criteria, purposes / audiences)?
- How to find / generate appropriate sample data for testing?

Our Solution: 2-fold … VIEWS

- **“Count” views across child tables**
- One record per PK
- Columns for standard selections (type, status, date-related) and at least 1 column per child type counted
  (sometimes more – denormalized, to suite analysis needs)

- **“Aggregate” views, one per select star**
- Use count views to analyze lengths, data combinations, to find test samples
- Often done when design begins
  (but we did not do it – benefit now is we know the data very well)

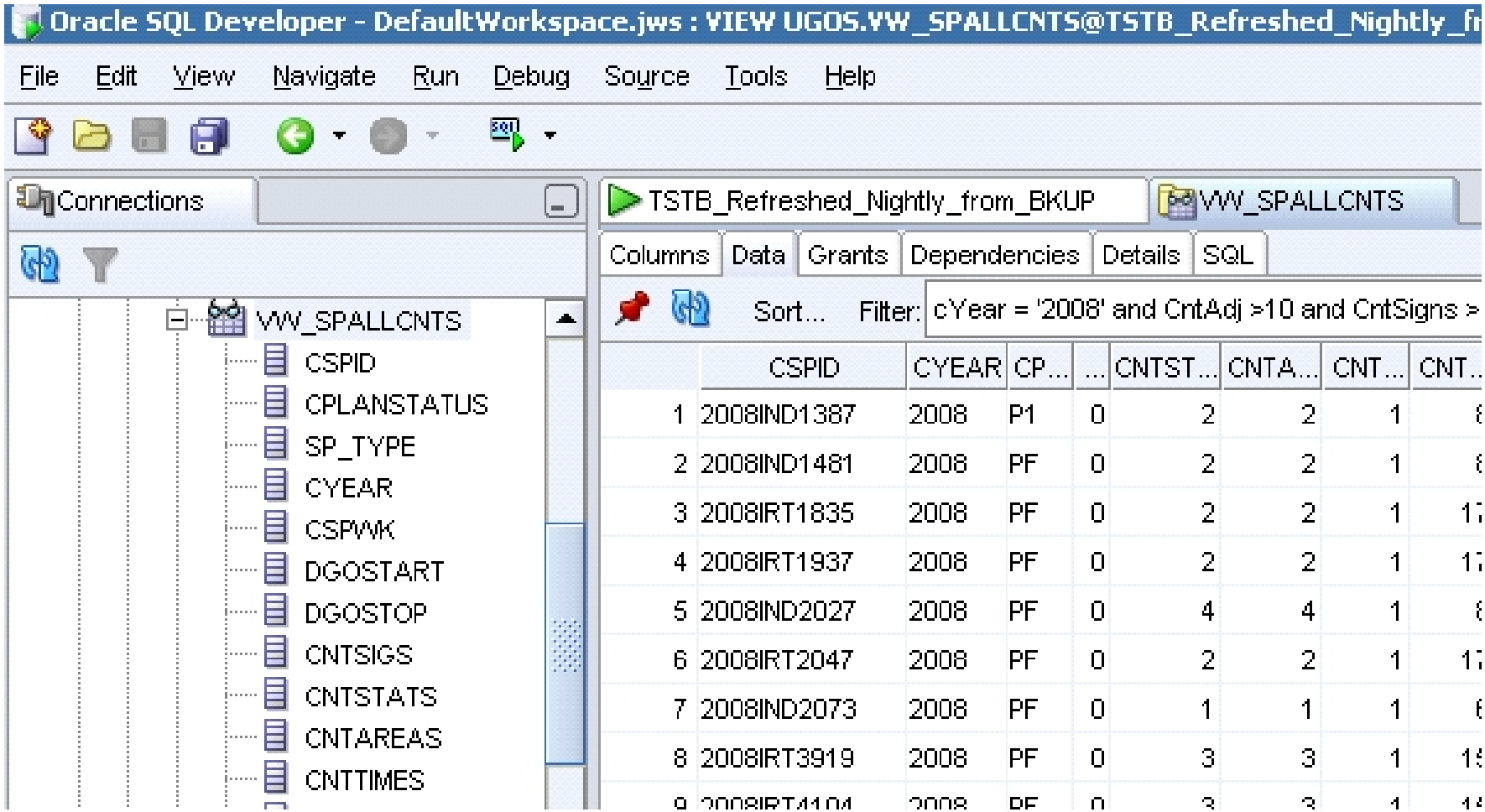
Our Solution: 2-fold … WITH

- Using Count view summary data to select actual data …for function, procedure, and error handling purposes
- Using Oracle analytic functions, aggregate the data (one record per type of child record summarized) for a reporting / extract view

Count View Code Snippet …

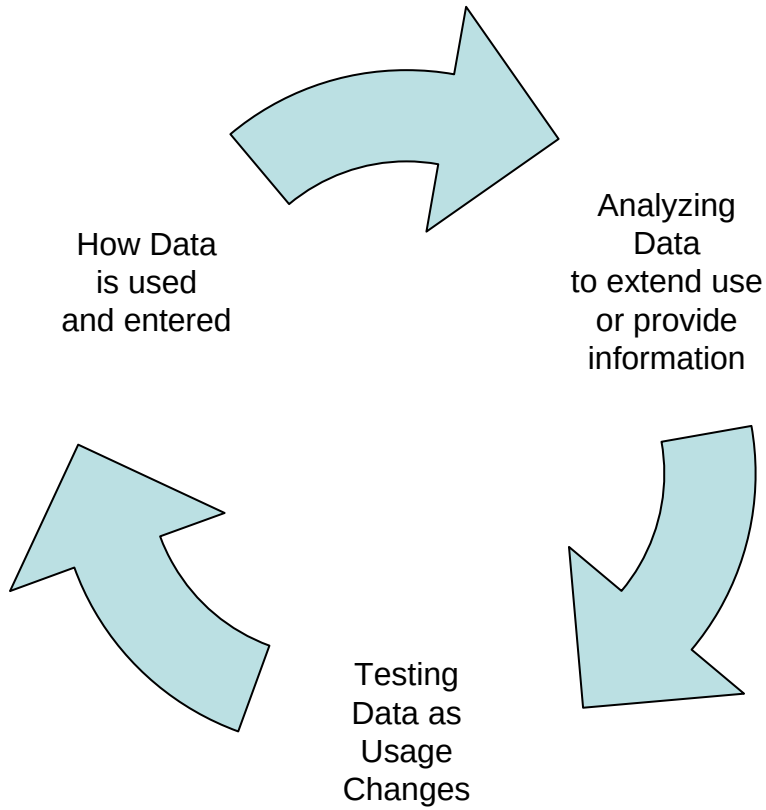
```
Create View VW_SPALLCNTS (CSPID, cPlanStatus, cYear,
       CNTTIMES, CNTEXCPS, CNTTEXT, CNTADJ, CNTWW, CNTDRs)
AS select CSPID, cPlanStatus, cYear,
          (select count(*) from SP3TA Three
                           where Three.CSPID = Base.CSPID
                             and Three.CDATETYPE <> 'AE'),
          (select count(*) from SP3TA Threa
                           where Threa.CSPID = Base.CSPID
                             and Threa.CDATETYPE = 'AE'),
          (select count(*) from SP4DescFour
                           where Four.CSPID  = Base.CSPID
                             and Four.CTEXTTYPE <> 'CN'),
          (select count(*) from SP5Adj Five
                           where Five.CSPID  = Base.CSPID),
          (select count(*) from SP6WWs Six
                           where Six.CSPID   = Base.CSPID),
          (select count(*) from DRXRef Sxtn
                           where Sxtn.CSPID  = Base.CSPID)
     from SP0SERVICEPLANS Base;
```

In SQL Developer …



Covering …

- A little background on my team's project
- Considerations for summarizing data
- **Data analysis**
- Analysis and testing opportunities



MAIN Question:

How long should the Aggregate column be?

We looked at the summary data using Excel, across all columns, in each "All Count" view

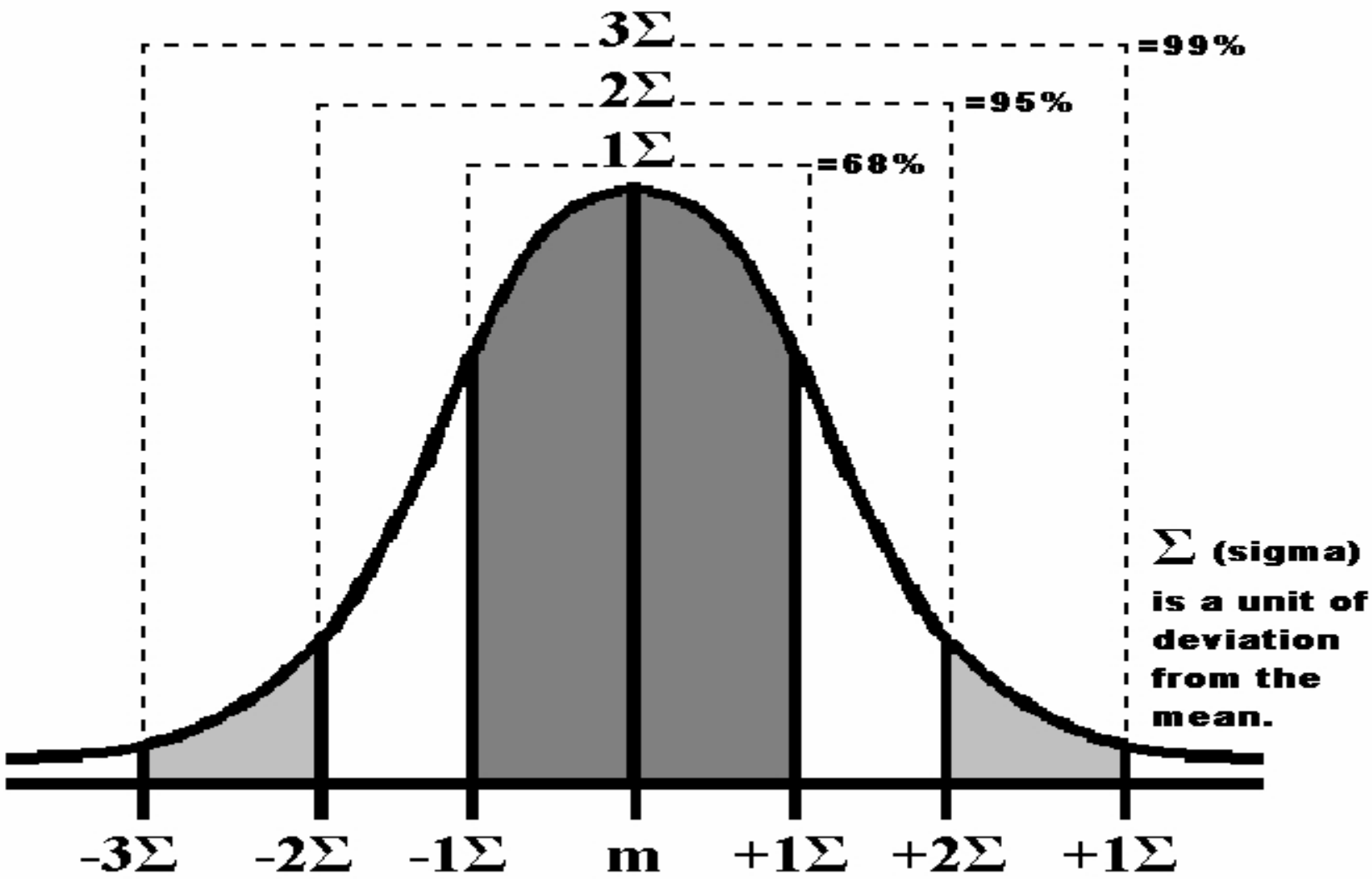
ROUGH Answer: 1000 characters was a nice round number …

- Too large for Accounting Data (but so what?)
- Too small for Text Data (but we knew we could not aggregate all text and a summary field existed in one child table)
- Just right for almost all other data, with error traps in place
  (but how to find the records that could cause error?)

Sample Analysis … “Works With” …

```
with DATA as (select cyear, cspid, cntww,
                     (case when cntww >50 then '10 = 51+'
                           when cntww >45 then '09 = 46-50'
                           when cntww >40 then '08 = 41-45'
                           when cntww >35 then '07 = 36-40'
                           when cntww >30 then '06 = 31-35'
                           when cntww >25 then '05 = 26-30'
                           when cntww >20 then '04 = 21-35'
                           when cntww >15 then '03 = 16-20'
                           when cntww >10 then '02 = 11-15'
                           when cntww >5  then '01 = 6-10'
                           when cntww >0  then '00 = 1- 5'
                           else '00 = 0' end) as Catg
                 from vw_SPALLCNTS)
select catg as grp,
              count(cspid) as cnt_sp, sum(cntww) as sum_ww
  from DATA
   group by catg order by catg;
```

A Simple Bell Curve …



Sample Analysis … “Works With” …

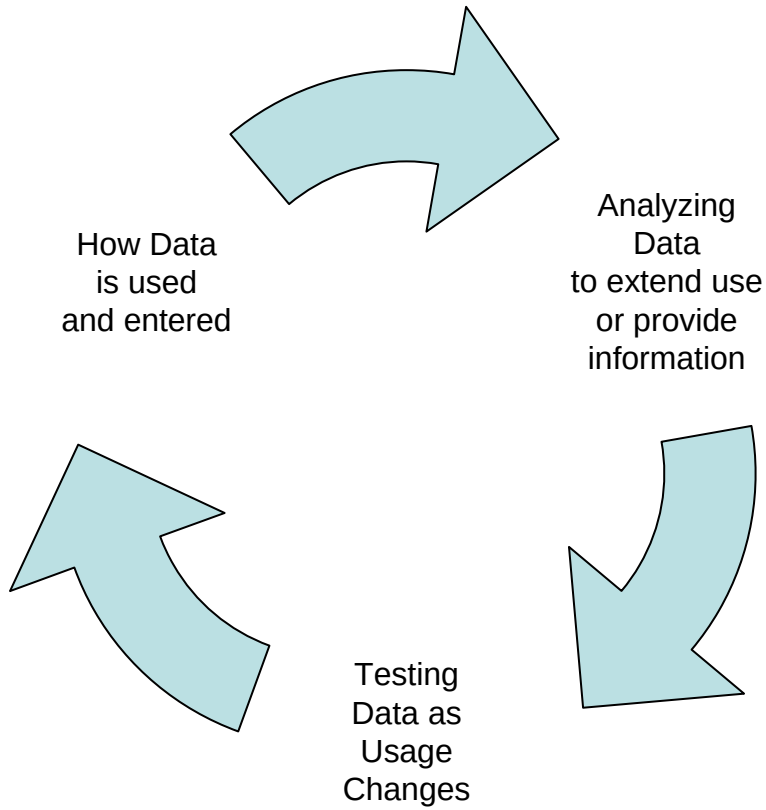
GRP	CNT_SP	SUM_WW	PerCnt	PerAccum	PerTot	PerAccTot
00 = 0	15393	0	34.44	34.44	0.00	0.00
00 = 1- 5	12935	38206	28.94	63.38	13.15	13.15
01 = 6-10	8914	68775	19.94	83.33	23.67	36.82
02 = 11-15	3783	47925	8.46	91.79	16.50	53.32
03 = 16-20	1728	30712	3.87	95.66	10.57	63.89
04 = 21-35	737	16757	1.65	97.31	5.77	69.66
05 = 26-30	334	9284	0.75	98.05	3.20	72.86
06 = 31-35	182	5938	0.41	98.46	2.04	74.90
07 = 36-40	115	4361	0.26	98.72	1.50	76.40
08 = 41-45	83	3554	0.19	98.90	1.22	77.63
09 = 46-50	55	2634	0.12	99.03	0.91	78.53
10 = 51+	435	62367	0.97	100.00	21.47	100.00
	44694	290513				

So the REAL pattern is …

```
with ALIAS as (select [identifier], [SumItem],
                      (case when SumItem >50 then '10 = 51+'
                            when SumItem >45 then '09 = 46-50'
                            when SumItem >40 then '08 = 41-45'
                            when SumItem >35 then '07 = 36-40'
                            when SumItem >30 then '06 = 31-35'
                            when SumItem >25 then '05 = 26-30'
                            when SumItem >20 then '04 = 21-35'
                            when SumItem >15 then '03 = 16-20'
                            when SumItem >10 then '02 = 11-15'
                            when SumItem >5  then '01 = 6-10'
                            when SumItem >0  then '00 = 1- 5'
                            else '00 = 0' end) as Category
                from CountView)
select Category,
       count(identifier) as CatCount, sum(SumItem) as CatSum
  from ALIAS
   group by Category order by Category;
```

Covering …

- A little background on my team's project
- Considerations for summarizing data
- Data analysis
- **Analysis and testing opportunities**



Examples and Uses …

- For UGOS, counts of text and counts of service adjustments are in newsletter article … analysis was done for each type of record where records were flattened into a single data entry.
- SOME types of data flatten in several ways, each important to a different user group or purpose, e.g., station codes versus station text.
- We've worked out a way to add in and aggregate new types of data as these become known, but length and "fit" analysis has to be done for each type of data, to avoid "overstuff" errors (our self-imposed 1,000 char limit).

Uses for Testing …

```
with data as (select cSPID from vw_SPALLCNTS
                where (cntCont+CntFunc+CntJobs+CntRCNs) >3
                  and cYear >'2006')
select cspid||': '||substr(get_Accounting(cSPID,'A'),1,100) as
   useful
  from data where rownum <2;
```

2007IRT8077: CNs:C-33293 JNs:15705 RNs:2832 FNs:500

```
with data as (select cSPID from vw_SPALLCNTS
                where (cntCont+CntFunc+CntJobs+CntRCNs) =0
                  and cYear >'2006')
select cspid||': '||substr(get_Accounting(cSPID,'A'),1,100) as
   useful
  from data where rownum <2;
```

2007IND7996: No Acctg Data

```
select GET_ACCOUNTING('2007IRT2524','P') as useful from dual;
```

Bad Type Entered

```
select GET_ACCOUNTING('2007I24','A') as useful from dual;
```

Bad Key Length

Questions?

Suzanne Michelle

Suzanne.Michelle@nyct.com