

Ensuring Data Integrity using Suprtool



A Robelle Tutorial

by Mike Shumko

August 1997

Copyright 1997, Robelle Solutions Technology Inc.



1

Each production database needs programs to ensure that company rules and conventions are being followed. Unfortunately, these programs seldom get written. This tutorial is for those who want to learn how to perform a wide range of integrity checks with Suprtool, Robelle's high-performance data tool on the HP 3000. It will include the following integrity checks:

- Do debits equal the credits?
- Are there customers with two addresses?
- Are all dates within the last 18 months?
- Are there order headers with no line items?
- Are there missing auxiliary records in KSAM or flat files?
- Does price times quantity equal extended-amount in all records?
- Are there at least five status reports per work-order?

Robelle Solutions Technology Inc.
Suite 201, 15399 - 102A Avenue
Surrey, B.C. Canada V3R 7K1

Toll-free: 1.888.ROBELLE
1.888.762.3553)
Telephone: 604.582.1700
Fax: 604.582.1799
E-mail: support@robelle.com
Web: www.robelle.com

Suprtool is a trademark of Robelle Solutions Technology Inc. Other product names and companies may be the trademarks of their respective owners.

For Techies

References

For further information on topics covered in this tutorial, please consult the *Suprtool User Manual*.

What's Inside

	<u>Page</u>
■ Types of integrity checks	3
■ Simple rule verification	4
■ Duplicate detail records	5
■ Correct data values	6
■ Missing pseudo masters	8
■ Masters without details	9
■ Relationships between fields	10
■ Check completeness	11
■ Relationships between files	13
■ Summary	20

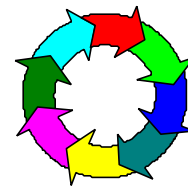
This tutorial will show you how Suprtool can perform a wide range of integrity checks to ensure that data is both correct and complete. We will end the tutorial with a set of job streams that show you complete examples of each integrity check.

For Techies

References

Types of integrity checks

1. Simple rule verification
2. Duplicate detail records
3. Correct data values
4. Missing pseudo master records
5. Master records without attached details
6. Relationships between fields
7. Check completeness
8. Relationships between files



3

You will learn how to use Suprtool to perform the following integrity checks:

1. Do debits equal credits?
2. Customers with two address records
3. Check data values by:
 - pattern (E.g., postal codes)
 - date range (E.g., only the last 18 months)
 - defined subfield (E.g., second character always an "S")
 - list of valid values (E.g., transaction types)
4. Master records kept in KSAM or flat files
5. Order headers without line items
6. Does price x quantity = amount?
7. At least five status records per work order
 - A control record for all 50 states
 - No group of transactions totals a negative amount
8. Do batch totals = total of all batches?

As we go through these examples, feel free to contribute your ideas and we'll see if Suprtool can handle them.

For Techies

All sample output in this tutorial was generated using version 3.7 of Suprtool/iX running on an HP 3000 Series 927 (MPE/iX version 5.0).

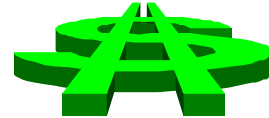
References

Simple rule verification

- Do the debits equal the credits?

```
:run suprtool.pub.robelle
>base      fms.gl,5,reader
>get       d-transactions
>output    $null
>total     trans-amt
>xeq
```

```
Totals (SUN, JAN 29, 1995, 2:56 PM):
TRANS-AMT                81451+
IN=6, OUT=6. CPU-Sec=1. Wall-Sec=1.
```



4

Financial transactions are often processed in batches. If there are programming bugs or if jobs abort, a batch may become unbalanced by losing a credit or debit transaction. You can use the Total command to quickly check whether the sum of all the transactions equals zero.

By default, the sum calculated by the Total command is displayed on the screen (\$STDLIST). You can also append the sum to the output file by using the following commands:

```
set squeeze off
total $file filename,append
```

For Techies

Each pass can total up to 15 fields.

References

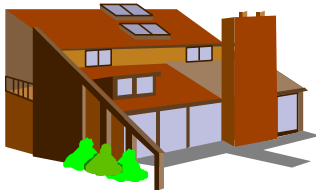
For a complete sample script, see page 14.

To locate an unbalanced batch with the Total option of the Duplicate command, see the example on page 19.

Duplicate detail records

- Are there customers with two address records?

```
>get          d-addresses
>sort         cust-no
>duplicate   only keys
>list standard title "Customers with duplicate addresses"
>xex
```



5

Unlike master datasets, detail datasets do not ensure unique keys. Suprtool's Duplicate command provides an easy way to detect duplicate records.

The Keys option in the Duplicate command refers to the fields specified in the Sort command. The sorted fields do not have to be IMAGE keys. For example,

Input	Output
1111 123 Main St.	2222 33359 East Ave.
2222 15458 West Ave.	
2222 33359 East Ave.	
3333 13 Lake Shore Dr.	

WARNING: If you use the Delete and the Duplicate commands in the same task, Suprtool will delete *much more* than just the duplicate records because the Delete operation occurs in the input phase of Suprtool processing while the Duplicate operation occurs in the output phase. If there are only a few duplicates, then you can manually delete them with Suprtool's Dbedit. For more information, see the *Suprtool User Manual* or call Robelle Technical Support.

For Techies

References

For more information on all the options to Suprtool commands, consult the *Suprtool User Manual*.

For a complete sample script, see page 14.

Correct data values

■ By pattern

```
>if phone-no >< "(###) ###-####" {phone number with area code}
>if product >< "@WIDGET@" {we only sell WIDGETS}
```

■ By date range

```
>item order-date,date,yyyymmdd
>if order-date < $date(*/*-6/last) {orders older than 6 months}
```

(*/*-6/last) is the last day of the month, six months ago

6

Suprtool pattern-matching uses the following operators:

= = selects records that match a pattern
> < selects records that do not match a pattern

You can use these special characters for pattern-matching:

@ zero or more characters of any type
a single numeric character
? a single alphanumeric character
~ zero or more blank characters
& escape character, next character is literal match
E.g., &@ matches the @ character

Date checking requires an Item command to define the date format. YYYYMMDD, YYMMDD, YYMM, PHDATE, ASK and CALENDAR are valid formats that support "<" and ">" comparisons in Suprtool \$date and \$today functions. MMDDYY, MMDDYYYY, DDMMYY, and DDMMYYYY formats can only support the "=" operator.

How would you find dates greater than tomorrow?

For Techies

References

To test for a relative date with the \$today function, see a complete sample script on page 15.

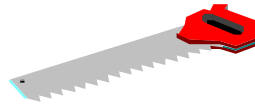
Correct data values continued

- By subfield

```
>define code2,product-code[2],1  
>if code2 <> "S","T" {2nd character must be "S" or "T"}
```

- By data class

```
>if part-code <> alpha {part codes not numbers or spaces}
```



7

The Define command identifies temporary fields that can be used with the If, Extract, Item, Sort, Table and Duplicate commands. Temporary fields are especially useful in systems that reserve space for future fields (E.g., FILLER-80).

The maximum length of a command line is 256 characters. To extend an If command, you can use an ampersand (&) at the end of the line to indicate continuation. For expressions longer than 256 characters, use the \$read function or the Table command.

You can use the following reserved words in an If statement to check whether all the characters in a field are one class:

ALPHA	A-Z, a-z (52 characters)
NUMERIC	0-9 (10 characters)
ALPHANUMERIC	A-Z, a-z, 0-9 (62 characters)
SPECIAL	anything else (194 characters including space, punctuation, Roman-8 letters, binary junk)

How can you find fields with null values?

For Techies

References

For a complete sample script, see page 15.

For an example of a Table command with an If expression, see the slide on the next page.

Missing pseudo master records

- Master records that are kept in KSAM or flat files

```
>input      prodmast      {KSAM master file}
>define     product-code,3,8  {3 byte offset, 8 byte
length}
>extract    product-code
>output     prodkeys,temp,link
>xeg
>get        d-inventory
>table      master-table,product-code,file,prodkeys
>if         not $lookup(master-table,product-code)
>list      standard title "Products without Masters"
8
```

Sometimes systems include master files that are not IMAGE databases (E.g., to facilitate sorted key access to the masters). To check if detail records exist without masters, you can create a table file of the master keys. Then you can read the details using the Table command and the If command with the \$lookup option. Any detail records without keys in the table are "orphans".

For Techies

The Table command ignores any duplicates before it starts sorting and processing the table file.

References

For a complete sample script, see page 16.

Masters without details

- Order headers without line items

```
>get      d-order-items
>extract  order-no
>sort     order-no
>out      orditems,temp,link
>xeg
>get      m-order-header
>table    detail-table,order-no,sorted,orditems
>if       not $lookup(detail-table,order-no)
>list     standard title "Orders without line items"
```

9

To find master records without details, you can use a technique that creates a table file of the detail keys. Then you can read the master records with the Table command followed by the If command with the \$lookup option. Any master records without keys in the table are "childless".

If you are reading large datasets (over 100,000 records), it is usually faster to sort a table file when you are creating it than to use the File option of the Table command. Suprtool's Sort command uses our proprietary sort routines while the Table command uses an insertion sort. Remember to use the Sorted option with the Table command instead of the File option.

For Techies

References

For a complete sample script, see page 16.

Relationships between fields

- Does price multiplied by quantity always equal amount?

```
>get      d-sales  
>if      price * quantity <> amount  
>list    standard title "Price * Qty not = Amount"
```

- Student Exercises

- Is the sales tax always 7% of the sales amount?
- Is the pay amount correct? It should be the pay rate times the regular hours plus time-and-a-half for overtime hours.



10

In most systems, there are many built-in formulas that need to be checked for compliance. The If command supports arithmetic operations in the following order of precedence:

- (...) Parentheses have the highest precedence
- NOT Take the opposite (logical negation)
- AND Both conditions must be true
- OR One or the other condition must be true
- Unary minus
- * / Multiplication and division, higher than + and -
- + - Addition and subtraction, use parentheses where needed

You cannot perform arithmetic operations on byte-type fields, and arithmetic expressions cannot start with a numeric constant (E.g., $2 + a = 10$ is invalid). If you do have a byte-type field consisting entirely of numeric digits, you can redefine the field to a display-type and use the new field name in an If command.

For Techies

References

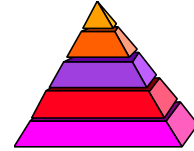
For a complete sample script, see page 17.

Check completeness

- At least five status records per work order

```

>get          d-process
>extract     work-order
>sort        work-order
>duplicate   none keys count total status-code
>output      worksum,link,temp
>xeq
>input       worksum
>if          st-count <> 5 or st-total-1 <> 150
>list        standard title "Non-Standard work-orders"
    
```



11

Lets say that each work order goes through 5 stages and that each stage generates a detail record with a status-code of 10, 20, 30, 40 and 50 respectively.

You can use the Duplicate command to summarize, count and subtotal the status-codes for each work order. In this scenario, a standard work order will count up to five and subtotal the status-codes to 150 (10 + 20 + 30 + 40 + 50).

The Total option of the Duplicate command generates a new variable for each field selected, and names the variables ST-TOTAL-1, ST-TOTAL-2 etc. One field counts the five stages and its variable name is ST-COUNT. The Define command can later rename these fields so they are more meaningful.

```

>define      work-orders, st-count
>define      status-checksum, st-total-1
>if work-orders <> 5 or status-checksum <>
150
    
```

For Techies

References

For a complete sample script, see page 18.

Check completeness using HowMessy

- Only 5 records per work order

```
>input loadfile
>if dataset = "D-PROCESS" and &
searchfield = "WORK-ORDER" and &
(maxchain <> 5 or avechain <> 5.00)
```

- Only one address per customer

```
>input loadfile
>if dataset = "D-ADDRESSES" and &
maxchain > 1
```

12

Your Suprtool tape also includes the bonus program HowMessy which generates statistics for a database. When it produces the report, it also creates a temporary self-describing file called Loadfile. You can write detailed job stream that examine the HowMessy output. Here are some of the fields from that file:

```
DATABASE
DATASET
DATASETTYPE
CAPACITY
ENTRIES
LOADFACTOR
SECONDARIES
HIGHWATER
SEARCHFIELD
MAXCHAIN
AVECHAIN
STDDEVIATION
```

For Techies

References

Relationships between files

- Batch control total must equal the sum of batch records
 1. Create a link file of the batch control records with batch-amt
 2. Create a link file of the summarized transaction records
 3. Use Suprlink to merge the two files

```
>input mergfile
>define actual-total,st-total-1
>if batch-amt <> actual-total
>list standard title "Batches with incorrect totals"
```

13

Some systems batch their transaction entries and keep a control record in a master file. To check whether the sum of the detail records equals the control amount, you can create two self-describing files. One file contains the control totals from the control records and the other file reads the detail records and uses the Total option of the Duplicate command to create a summary record for each batch. You can use Suprlink to merge the two files and then compare the two fields with the If command.

For Techies

References

For a complete sample script, see page 19.

Full sample job stream for correct data values

```
!job dbcheck,mgr.sales,db
!
!comment Integrity Check: 3. Correct Data Values
!comment This job runs Suprtool to print these exception reports:
!comment   Phone numbers not in the form (604) 582-1700
!comment   Product names that do not include the word "WIDGET"
!comment   Order dates older than 6 months or beyond tomorrow
!comment   Product codes whose 2nd character is not "S" or "T"
!comment   Part codes that are not ALL letters
!
!file suprlist;dev = lp
!run suprtool.pub.robelle
base    sales.db,5,reader
get     m-customer
if      phone-no >< "(###) ###-####"
list    standard title "Irregular Phone numbers"
xeq

get     m-product
if      description >< "@WIDGET@"
list    standard title "Non-WIDGET Product names"
xeq

get     d-order
item    order-date,date,yyyymmdd
if      order-date < $date(*/*-6/last)
list    standard title "Orders older than 6 months ago"
xeq

get     d-order
if      order-date > $today(+1)
list    standard title "Orders made after tomorrow"
xeq

get     m-product
define  code2,product-code[2],1
if      code2 <> "S","T"
list    standard title "Products whose 2nd char. is not S or T"
xeq

get     m-inventory
if      part-code <> alpha
list    standard title "Parts that are not ALL letters"
xeq
```

Full sample job stream for missing pseudo masters and details

```
!job dbcheck,mgr.sales,db
!
!comment Integrity Check: 4. Missing Pseudo Master Records
!
!comment This job runs Suprtool to list all the products in
!comment the inventory detail set that are missing from the
!comment KSAM prodmast file.
!comment It also lists all order header records that have no
!comment detail line items.
!
!file suprlist;dev = lp
!run suprtool.pub.robelle
  input    prodmast                {KSAM master file}
  define   product-code, 3, 8
  extract  product-code
  output   prodkeys, temp, link
  xeq

  base     sales,5,reader
  get      d-inventory
  table    master-table,product-code,file,prodkeys
  if       not $lookup(master-table,product-code)
  list     standard title "Products without Masters"
  xeq

!
!comment Integrity Check: 5. Masters Without Details
!
  get      d-order-items
  extract  order-no
  sort     order-no
  out      orditems, temp, link
  xeq

  get      m-order-header
  table    detail-table, order-no, sorted, orditems
  if       not $lookup(detail-table, order-no)
  list     standard title "Orders without line items"
  exit

!
```

Full sample job stream for relationships between fields

```
!job dbcheck,mgr.sales,db
!
!comment Integrity Check: 6. Relationships Between Fields
!
!comment This job runs Suprtool to list the following
!comment situations:
!comment     sales price * quantity not equal to amount
!comment     sales tax not equal to 7%
!comment     sales commissions greater than 15%
!comment     payroll payments with computation errors
!comment
!
!file suprlist;dev = lp
!run suprtool.pub.robelle
base     sales,5,reader
get      d-sales
if       price * quantity <> amount
list     standard title "Price * Qty not = Amount"
xeq

get      d-sales
if       (amount * 0.07) <> sales-tax
list     standard title "Sales tax not = 7% of sales amount"
xeq

get      d-commissions
if       (commission-owed / sales-amt) > 0.15
list     standard title "Commissions greater than 15%"
xeq

base     payroll,5,reader
get      d-payments
if       (reg-hrs * pay-rate) + (ot-hrs * pay-rate * 1.5) &
        <> pay-amount
list     standard title "Paycheck computation errors"
exit
!
!eoj
```


Full sample job stream for completeness test

```
!job dbcheck,mgr.sales,db
!
!comment Integrity Check: 7. Completeness
!
!comment This job runs Suprtool to verify that the Control
!comment Dataset has sales tax records for the 50 valid
!comment states and that all 50 state codes are present.
!
!comment M-CONTROL is the dataset of control table records,
!comment the key control-key, 2 bytes = table type and
!comment 2 bytes = table value.
!comment Table 15 is for State Sales Tax records. E.g.,"15CA"
!comment The file STATES is a 50 line file with valid state
!comment codes.
!
!file suprlist;dev = lp
!run suprtool.pub.robelle
  base  gl,5,reader
  get   m-control
  define type, control-key, 2
  define value, control-key[3], 2
  table state-table, value, file, states
  if    type = "15" and not $lookup(state-table, value)
  list  standard title "Invalid State Sales tax records"
  xeq

  get   m-control
  if    type = "15"
  extract value
  output mcontrol, temp
  xeq

  input states
  define value, 1, 2
  table tax-table, value, file, mcontrol
  if    not $lookup(tax-table, value)
  list  standard &
        title "States missing from the Sales & Tax table"
  exit
!
```

Full sample job stream for relationship test

```
!job dbcheck,mgr.sales,db
!
!comment Integrity Check: 8. Relationships Between Files
!
!file suprlist;dev = lp
!run suprtool.pub.robelle
    base      gl,5,reader
    get        m-batch
    sort       batch-no
    extract    batch-no, batch-amt
    output     mbatch, temp, link
    xeq

    get        d-transactions
    sort       batch-no
    duplicate  none keys total trans-amt
    extract    batch-no
    output     dtrans, temp, link
    xeq

    link       input dtrans
    link       link mbatch
    link       output mergfile, temp
    link       exit

    input      mergfile
    define     actual-total, st-total-1
    if         batch-amt <> actual-total
    list       standard title "Batches with incorrect
totals"
    exit

!
!eoj
```

Full sample job stream for invalid dates

```
!job dbcheck,mgr.sales,db
!  
!run suprtool.pub.robelle  
base      sales,5,reader  
  
get        d-sales  
item      order-date,date,ccyymmdd  
if         $invalid(order-date)  
extract   order-date  
sort      order-date  
duplicate none keys count  
list      standard title "Invalid date values, counted"  
xeq  
  
exit  
!eoj
```

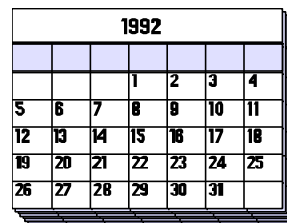
Full sample job stream for comparing records in two files

```
!job dbcheck,mgr.sales,db
!  
!run suprtool.pub.robelle  
base    sales,5,reader  
get     d-sales  
define  address-1,address  
extract order-no, address-1  
sort    order-no  
output  file1 link  
xeq  
  
base    archive,5,reader  
get     old-sales  
define  address-2,address  
extract order-no, address-2  
sort    order-no  
output  file2 link  
xeq  
  
link    input    file1  
link    link     file2  
link    output   file3  
link    xeq  
  
input   file3  
if      address-1 <> address-2  
list    standard title "Addresses don't match between files"  
xeq  
  
exit  
!eoj
```

Invalid dates

- Dates that don't make sense

```
>item      order-date,date,ccyyymmdd  
>if        $invalid(order-date)
```



1992						
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

15

To find date values that are syntactically incorrect, use the \$invalid function of the If command.

\$invalid looks for dates whose values don't fit the rules for the date type defined in the Item command. For example, a month of 0, a day of 99, or a day of 31 for month 11. Of course, \$invalid takes leap years into account.

To find only the valid dates, use IF NOT \$INVALID(field-name).

You may not think you have any invalid dates (your application program does a good job of validation), but do you have any flag values that have special meaning? E.g., 999999, -1, 0, NEVER, NOTADATE, *****.

The sample jobstream shows you all invalid dates and flag values, and how often they appear in the file.

For Techies

\$invalid was introduced in Suprtool version 4.0, in 1997.

References

For a complete sample script, see page 'Sample-21'.

Comparing records in two files

- Records in two files should have the same values in the address field

```
>link      input      file1
>link      link        file2
>link      output     file3
>link      xeq

>input     file3
>if        address-1 <> address-2
>list      standard title "Addresses don't match"
```

16

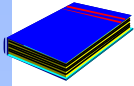
If you have two files that are supposed to mirror each other, you can use Suprtool to look for discrepancies.

You need to be able to match the records based on some common field. Then in the matching records you can look for differences. Create a composite file using Suprlink that has all the fields you are interested in, from both files.

For Techies

References

For a complete sample script, see page 'Sample-22'.



Summary of Suprtool Integrity Checks

- Easy to validate many important facts and relationships
- No need to write programs
- Fast execution
- Sample job streams can be modified

17

One way to make the jobs more intelligent is to initialize a variable to zero. Then after each Suprtool task, increment it by the value of the variable SuprtoolOutCount. If at the end of the job, the value of the new variable is still zero then the job's \$Stdlist can be deleted.

```
!job dbcheck,mgr.sales,db;outclass=,5
!
!setvar errorcount, 0
!
!run suprtool.pub.robelle
...      { task # 1 }
setvar errorcount, errorcount + suprtooloutcount
...      { task # 2 }
setvar errorcount, errorcount + suprtooloutcount
exit
!
!if errorcount = 0 then
!  set stdlist=delete
!endif
!eoj
```

For Techies

Suprtool automatically generates the variable SuprtoolOutCount with the number of records selected.

References