HASH Join's
Methods of Joining

• Data Merge
• Format (Informat)
• SQL Join
• HASH Table
Data Merge Pseudo Code

• Data step to Load Data 1
• Data step Load Data 2
• If not indexed then sort 1 and/or 2
• Data ; Merge Data1 Data2
Format (Informat)

• Data step to Load Data 1 and create Informat data set.

• Proc Format CNTLIN= to Load Data 1 into format library.

• Data step Load Data 2
  – Use informat as a lookup to add extra value(s).

Note that 1 format would be required for each data element desired in the Join
SQL Join Pseudo Code

• Data step to Load Data 1
• Data step Load Data 2
• Proc SQL join
Joining Methods crude read/write examples

- **Data Assumption**
  - 1,000,000 Records in the main table
  - 1,000 Records in the secondary (lookup) table.

- **Read Write Examples Full Join**

<table>
<thead>
<tr>
<th>Read Write Examples</th>
<th>Read</th>
<th>Write</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Merge</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Load 1</td>
<td>1,000</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Load 2</td>
<td>1,000,000</td>
<td>1,000,000</td>
<td>1,000,000</td>
</tr>
<tr>
<td>Merge</td>
<td>1,001,000</td>
<td>1,000,000</td>
<td>2,001,000</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1,001,000</td>
<td>2,001,000</td>
<td>3,002,000</td>
</tr>
<tr>
<td><strong>Format/Informat</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Load 1</td>
<td>1,000</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Proc Format</td>
<td>1,000</td>
<td>1,000</td>
<td>2,000</td>
</tr>
<tr>
<td>Load 2</td>
<td>1,000,000</td>
<td>1,000,000</td>
<td>1,000,000</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1,000</td>
<td>1,002,000</td>
<td>1,003,000</td>
</tr>
<tr>
<td><strong>Proc SQL</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Load 1</td>
<td>1,000</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Load 2</td>
<td>1,000,000</td>
<td>1,000,000</td>
<td>1,000,000</td>
</tr>
<tr>
<td>Join</td>
<td>1,001,000</td>
<td>1,000,000</td>
<td>2,001,000</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1,001,000</td>
<td>2,001,000</td>
<td>3,002,000</td>
</tr>
</tbody>
</table>
HASH in SAS

• DOW/Array based Hash Tables (Dorfmann)
  – Hand coded techniques

• Hash Object
  – Introduction
  – Unique Properties
Code example for Hash Table

- DATA PREVFILE (KEEP=ClientID  Balance SalesPersonCode) ;
- DECLARE HASH BUID () ;
- RC  =BUID.DEFINEKEY ('SalesPersonCode') ;
- RC  =BUID.DEFINEDATA ('BusinessUnit') ;
- RC  =BUID.DEFINEDONE() ;

DO UNTIL (EOF1) ;
SET REPTONE (KEEP=SalesPersonCode BusinessUnit WHERE=(BusinessUnit EQ 3)) END=EOF1 ;
RC  =BUID.ADD() ;
END ;

DO UNTIL (EOF2) ;
INFILE PREVFILE END=EOF2 ;
INPUT
@014 SalesPersonCode  $7.
@;  
RC  =BUID.FIND() ;
IF REPTIND EQ 3 THEN DO ;
INPUT
@007 ClientID  $6.
@033 Balance  PD8.2
;
IF RC EQ 0 THEN OUTPUT ;
END ;
END ;
Credits

• Paul M. Dorfman

• SUGI/SGF Paper (Found on Lex Jansen’s Site)
  – Snell, Gregg P. “Think FAST! User Memory Tables (Hasing) for Faster Merging.”

• SAS Hash Objects: A Programmer’s Guide
  – Michele M. Burlew, SAS Institute Press
Questions