

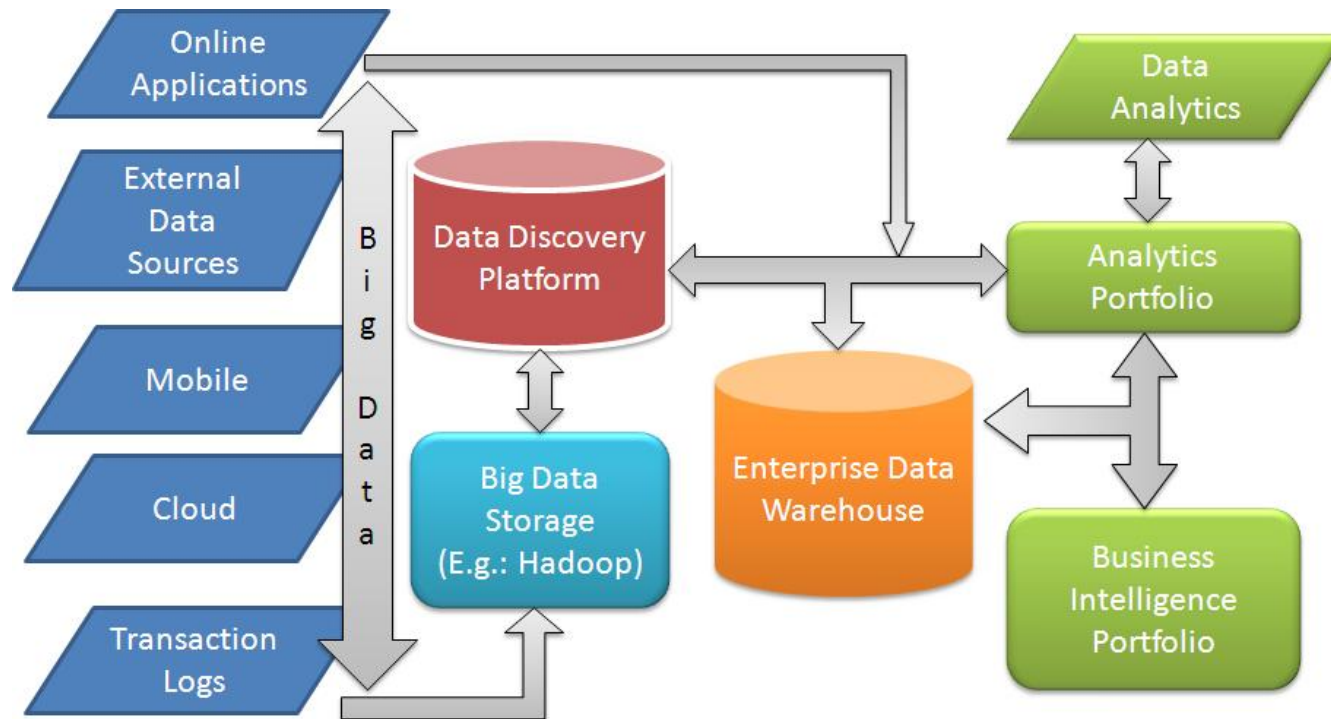
Good Data  $\hat{=}$  Accident

Harry Droogendyk – Stratia Consulting Inc.

# Where does your data come from?

---

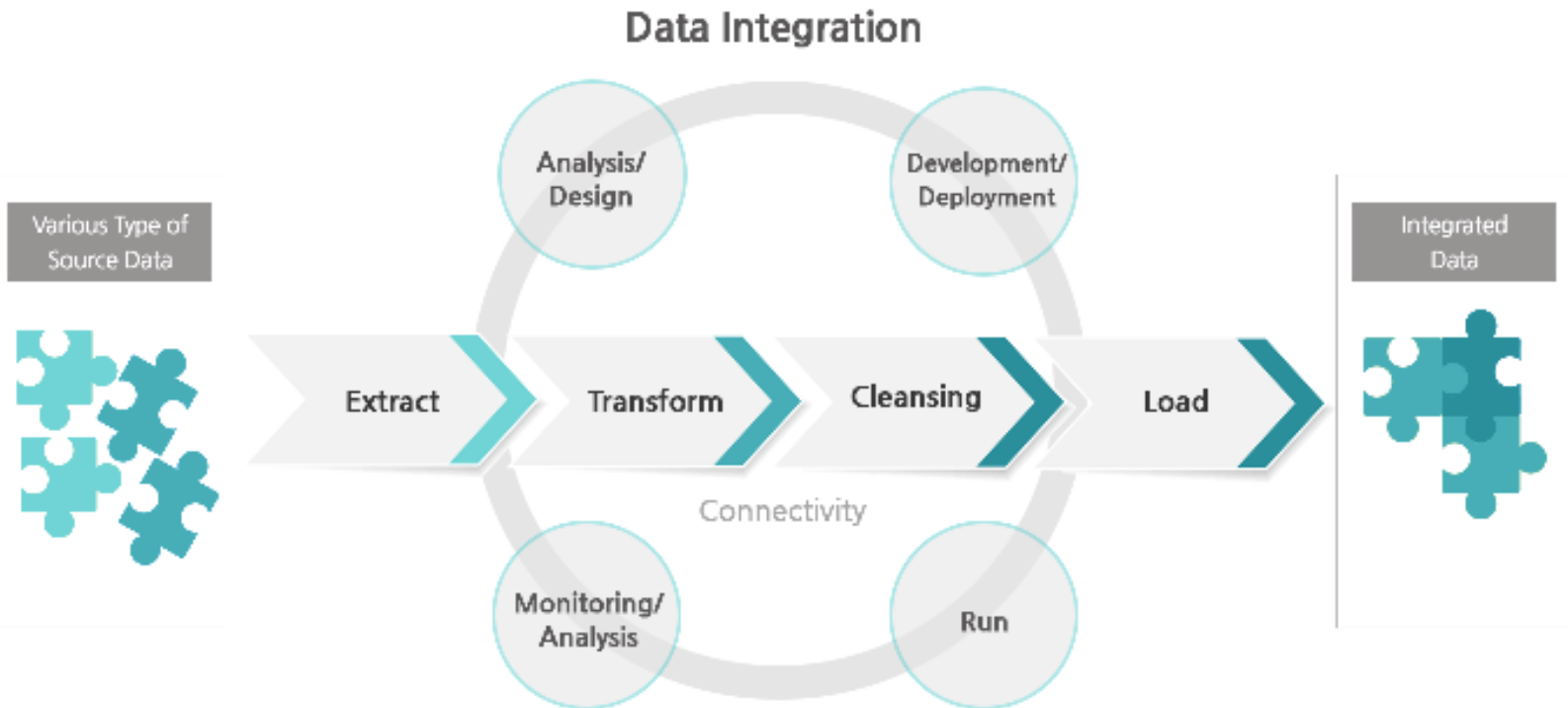
## ► Complex architecture



# How does it get there ?

---

- ▶ Multiple components / steps



# Characteristics of "good data" ?

---



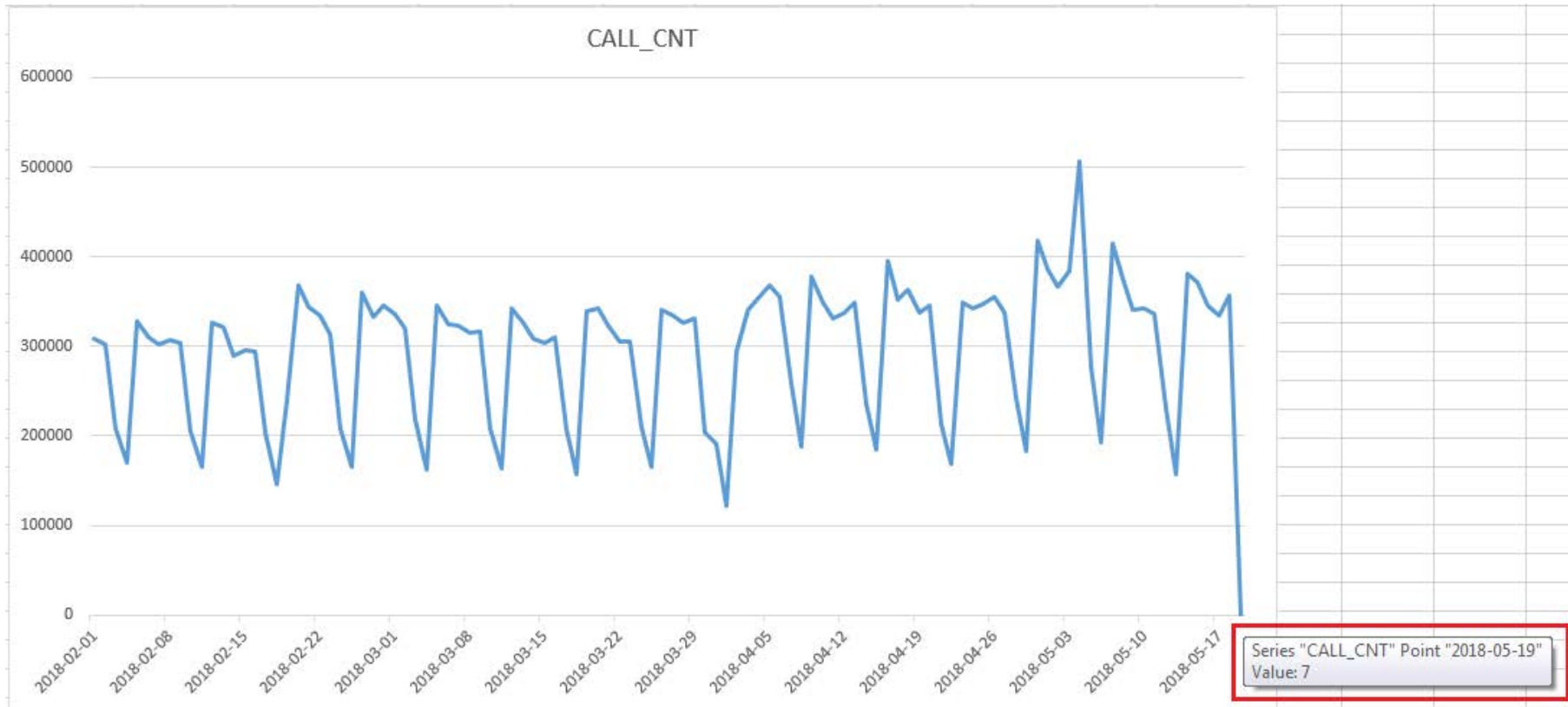
# Characteristics of "good data" ?

---

- ▶ accurate
- ▶ complete
- ▶ timely
- ▶ real
- ▶ representative
- ▶ useful



# What is "good data" ?



- ▶ On morning of May 20<sup>th</sup>, do we have complete IVR data ?
  - ▶ should IVR based *daily* reporting go out ?



# How is "bad data" identified ?

---

## ▶ QA

- ▶ data analysis
- ▶ nulls
- ▶ counts
- ▶ trends
- ▶ outliers

## ▶ logs

## ▶ by end users

- ▶ i.e. *too late* 😊



# How do we effect "good data" ?

---

- ▶ reliable data sources
- ▶ timely extracts
- ▶ robust code
- ▶ simplicity
- ▶ canned processes
- ▶ controls
- ▶ audit
  
- ▶ summarized as "best practices"





# Best Practices

---

- ▶ Ignoring...
  - ▶ architecture
  - ▶ philosophy / design
  - ▶ most of Data Integration 😊
  
- ▶ Focusing on nuts 'n bolts... stuff we *all* can do ...
  - ▶ audit
  - ▶ automation
  - ▶ standards
  
- ▶ Does it *really* pay off ?



# Not Sexy

---

- ▶ growth of analytics
  - ▶ SEXY Data Scientist !
- ▶ presentation layers that pop
- ▶ so, so allllluuuring

REDACTED



- ▶ data work .... ho hum
- ▶ GIGO
- ▶ good data *is* foundational
- ▶ data = steak
  - ▶ keep the sizzle 😊



# Best Practices – really ?!?

---

- ▶ is Data Integration a "thing" anymore?

- ▶ Data Federation / Virtualization

- ▶ aren't we all hackers now ?

- ▶ throwaway solutions

- ▶ technology comes and goes

- ▶ pendulum swings

- ▶ process

- ▶ accountability



# Agenda

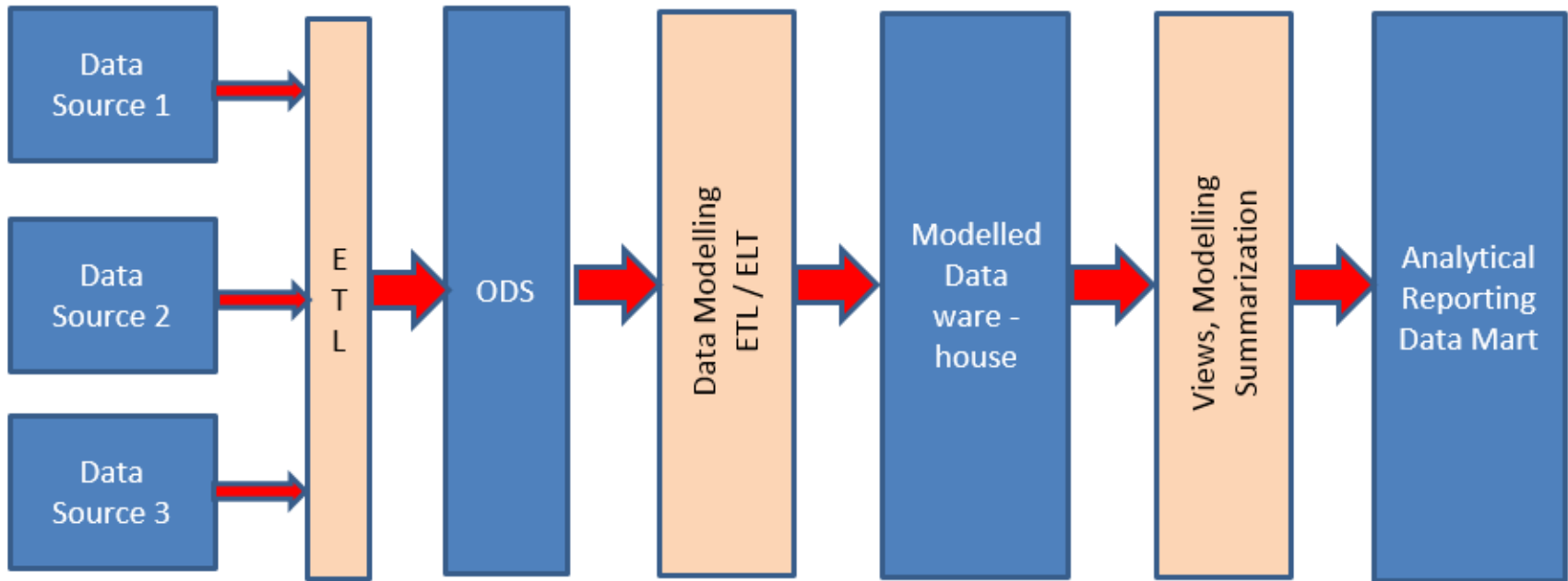
---

- ▶ quick overview of typical architecture
- ▶ tools and techniques
- ▶ control tables
- ▶ constructing jobs and flows
  - ▶ focus functionality
  - ▶ templates
- ▶ benefits



# Data Integration

---



- ▶ "Data Integration is the process of combining data from a heterogeneous set of data stores to create one unified view of all that data", B-Eye Network
- 



# Data Sources

---

- ▶ **projects:**
  - ▶ cheap
  - ▶ on time
  - ▶ correct
- ▶ **pick two of the three!**
  
- ▶ **data sources can be like that:**
  - ▶ accurate
  - ▶ timely
  - ▶ reliable
- ▶ **pick two of the three ?**

Data  
Source 1

Data  
Source 2

Data  
Source 3



# Data Sources

---

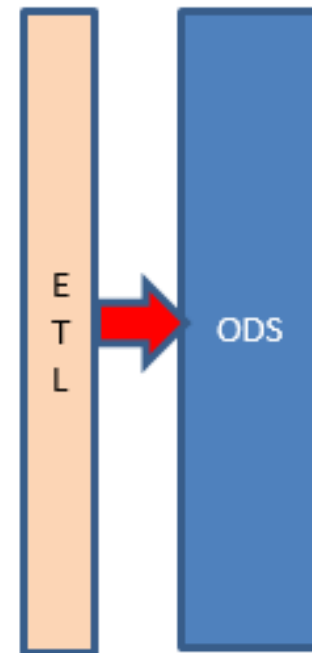
- ▶ know the source data
  - ▶ data dictionary
  - ▶ primary keys
  - ▶ notification when source structure changes
  
- ▶ *reliable* delta criteria ?
  
- ▶ use the strengths of the source DB
  - ▶ Oracle hints, partitions, Teradata PPIs
  
- ▶ text files?
  - ▶ SAP Data Surveyor, bless their hearts



# Operational Data Store ( ODS ) Layer

---

- ▶ **often same format as source**
  - ▶ some transformation occur, e.g. Oracle dates
  - ▶ drop null columns
- ▶ ***reliable* delta extracts loading ODS ?**
  - ▶ periodic true-up recasts
  - ▶ CDC
  - ▶ trunc & load
- ▶ **populate ODS audit columns**
  - ▶ load & update timestamps
    - ▶ used by our downstream processes
  - ▶ latest load event



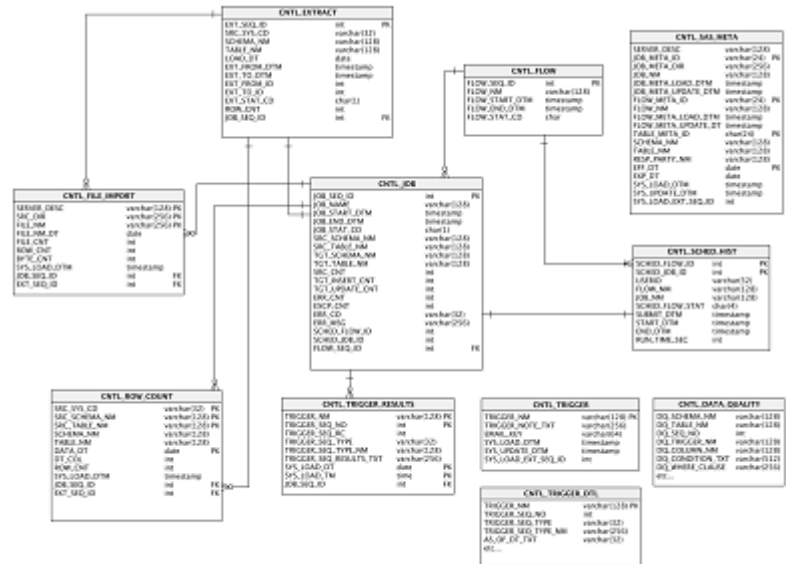


# Modeled Data Loading

- ▶ source may be modeled environment
  - ▶ but we're creating an integrated model

## ▶ DATA MODELS ARE IMPORTANT !

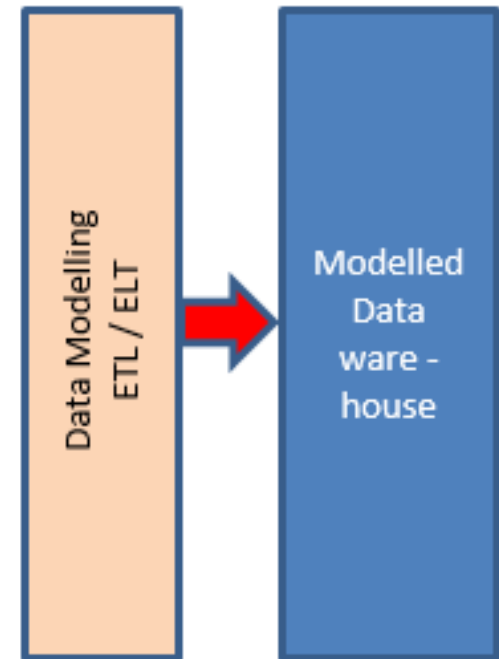
- ▶ normalization
- ▶ naming
  - ▶ same name
  - ▶ standards → contentious !
- ▶ data dictionary



# Modeled Data Loading

---

- ▶ **delta process extracts from ODS**
  - ▶ where ODS update/load timestamp > last extracted timestamp
  - ▶ using source transaction dates is dangerous
    - ▶ data loaded out of order ?
- ▶ **populate Modeled Data audit columns**
  - ▶ load & update timestamps
  - ▶ latest load event
- ▶ **Analytical layer – views, tables**
  - ▶ denormalized, transposed, summarized



# Tools & Techniques

---

- ▶ SAS Data Integration Studio
- ▶ Base SAS



# Tools & Techniques

- ▶ use features of the tool when *practical*

The image shows two overlapping dialog boxes from a SAS tool. The background dialog is 'Loop Properties' and the foreground dialog is 'Teradata Table Loader Properties'.

**Loop Properties Dialog:**

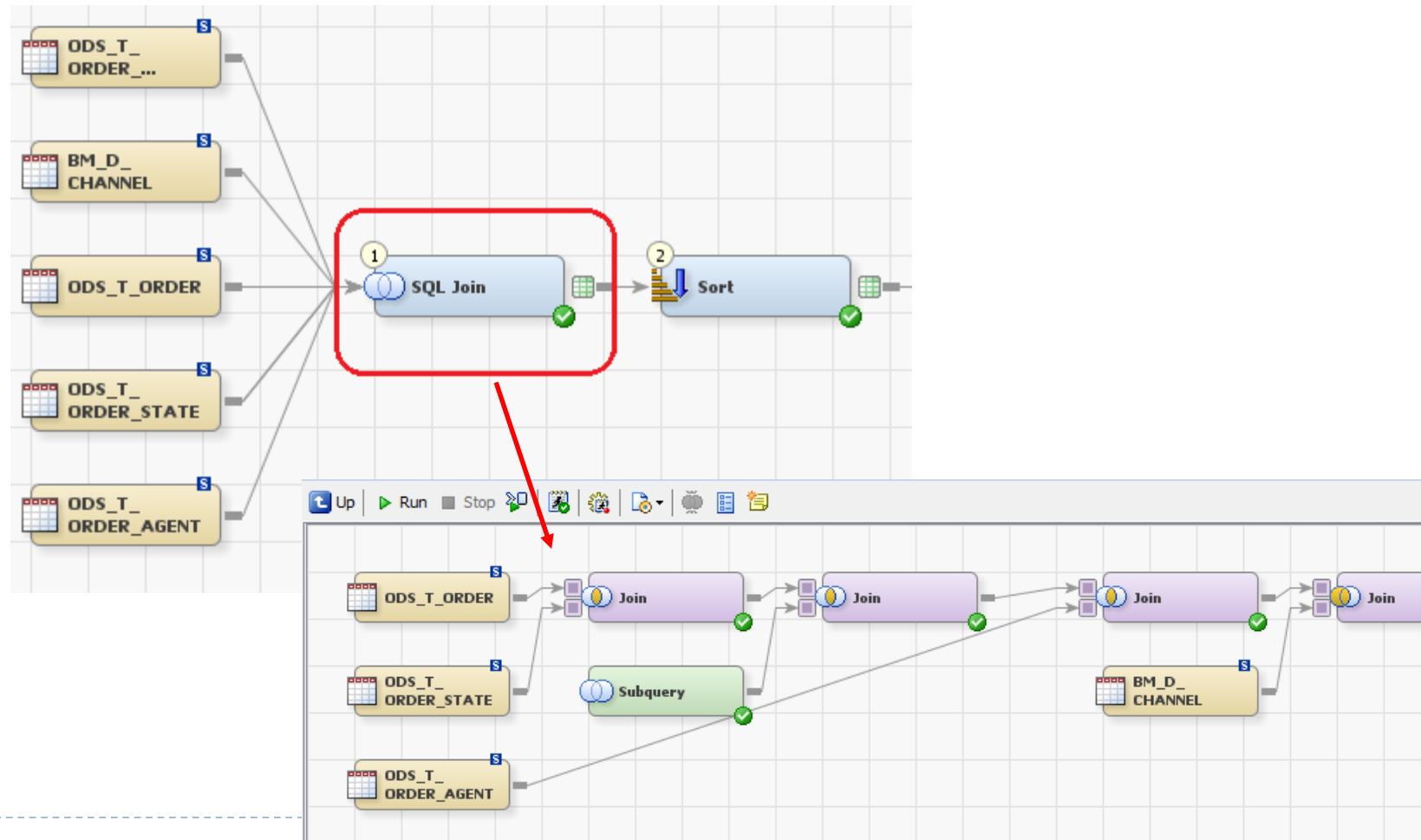
- Tab: Parameter Mapping
- Table with columns: Parameter Name, Macro Variable, Default Value, Mapped Source Column
- Row 1: Parameter Name: ban\_lo, Macro Variable: ban\_lo, Mapped Source Column: nm1\_column\_lo
- Row 2: Parameter Name: ban\_hi, Macro Variable: ban\_hi, Mapped Source Column: (empty)

**Teradata Table Loader Properties Dialog:**

- Table: LND\_BM\_T\_NM1\_CHRG\_HIST (Output)
- Tab: Teradata Options
- Section: Advanced > FastLoad
- Section: \* FastLoad utility (FASTLOAD):
  - Text: Specifies using Teradata's bulk-load capability by loading rows of data as one unit. FastLoad can load only empty tables; it cannot append to a table that already contains data.
  - Value: YES
- Section: Error tables name (BL\_LOG):
  - Text: Specifies the names of the error tables that are created when you are using the FastLoad facility. By default, the errors are logged in Teradata tables named SAS\_FASTLOAD\_ERRS1\_<random> and SAS\_FASTLOAD\_ERRS2\_<random>.
  - Value: &\_errTable
- Section: Control file (BL\_CONTROL): (Value: Reset)

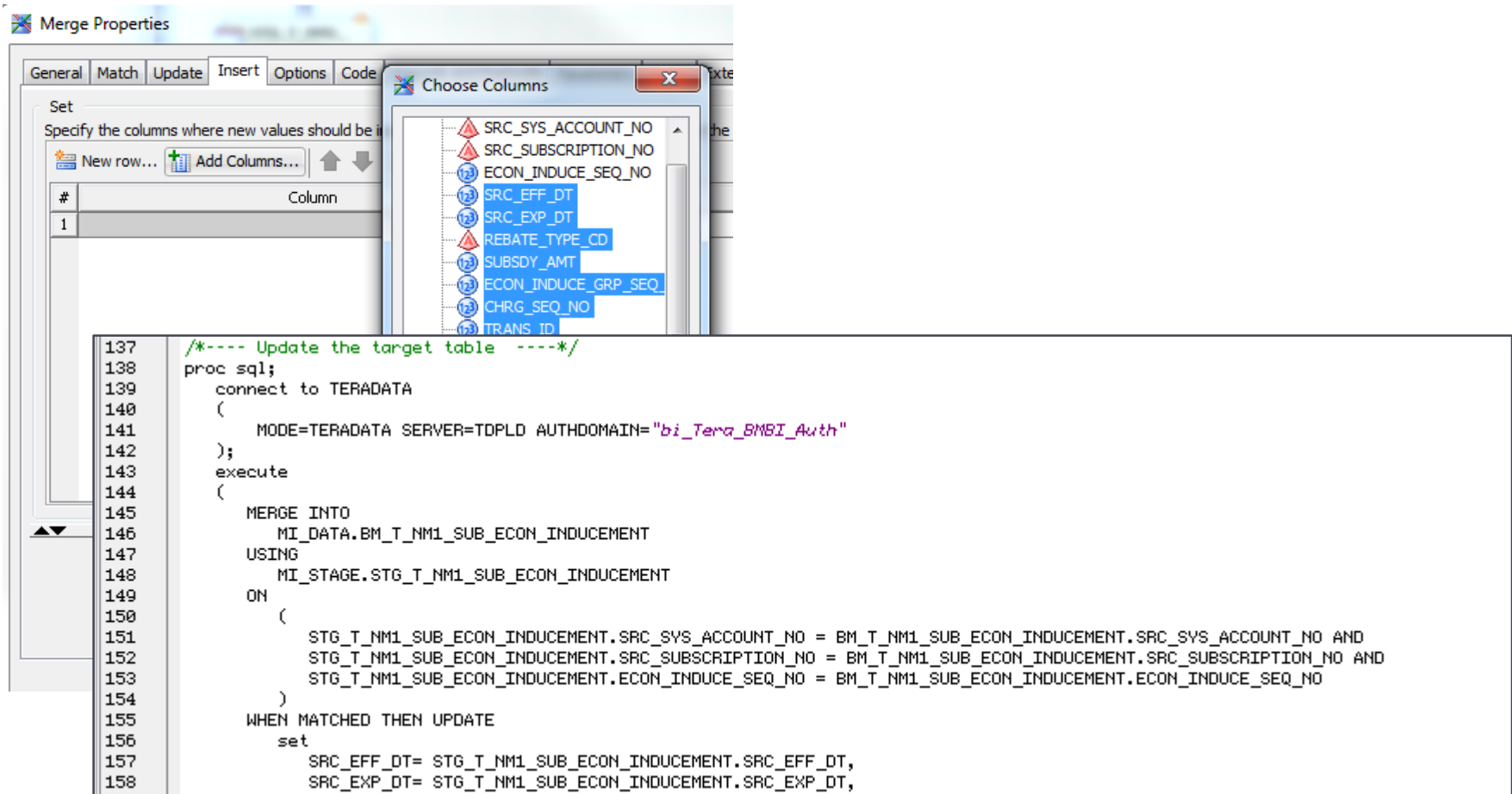
# Tools & Techniques

- ▶ do *not* use all features of the tool



# Tools & Techniques

- ▶ do *not* use all features of the tool



The screenshot displays a database tool interface. The 'Merge Properties' dialog is open, showing the 'Update' tab. A 'Choose Columns' dialog is also open, listing columns with checkboxes. Below the dialog is a code editor window showing a SQL script for a merge operation.

```
137 /*---- Update the target table ----*/
138 proc sql;
139   connect to TERADATA
140   (
141     MODE=TERADATA SERVER=TDPLD AUTHDOMAIN="bi_Tera_BMBI_Auth"
142   );
143   execute
144   (
145     MERGE INTO
146     MI_DATA.BM_T_NM1_SUB_ECON_INDUCEMENT
147     USING
148     MI_STAGE.STG_T_NM1_SUB_ECON_INDUCEMENT
149     ON
150     (
151       STG_T_NM1_SUB_ECON_INDUCEMENT.SRC_SYS_ACCOUNT_NO = BM_T_NM1_SUB_ECON_INDUCEMENT.SRC_SYS_ACCOUNT_NO AND
152       STG_T_NM1_SUB_ECON_INDUCEMENT.SRC_SUBSCRIPTION_NO = BM_T_NM1_SUB_ECON_INDUCEMENT.SRC_SUBSCRIPTION_NO AND
153       STG_T_NM1_SUB_ECON_INDUCEMENT.ECON_INDUCE_SEQ_NO = BM_T_NM1_SUB_ECON_INDUCEMENT.ECON_INDUCE_SEQ_NO
154     )
155     WHEN MATCHED THEN UPDATE
156     set
157     SRC_EFF_DT= STG_T_NM1_SUB_ECON_INDUCEMENT.SRC_EFF_DT,
158     SRC_EXP_DT= STG_T_NM1_SUB_ECON_INDUCEMENT.SRC_EXP_DT,
```

# Tools & Techniques

---

- ▶ autocall macro to generate MERGE
  - ▶ DB dictionary columns
  - ▶ macro parms to customize update / insert

```
1
2  %macro do_initial;
3
4      %if &_etl_run_type = INITIAL %then %do;
5          %td_trunc( schema = &_tgtSchema, table = &_tgtTable )
6          %errorhandling_etl(xtr_seq_id=&_seqid, job_seq_id=&_jobseqid, schema=&_schema, table=&_table, srccsys=
7      %end;
8
9  %mend;
10
11 %do_initial
12
13 %td_update_insert (  trans_db      = &_stgSchema,
14                      trans_table   = &_stgTable,
15                      master_db     = &_tgtSchema,
16                      master_table  = &_tgtTable,
17                      join_cols     = %str(SRC_SYS_ACCOUNT_NO SRC_SUBSCRIPTION_NO ECON_INDUCE_SEQ_NO),
18                      create_history | N
19                      );
20
21 %errorhandling_etl(xtr_seq_id=&_seqid, job_seq_id=&_jobseqid, schema=&_schema, table=&_table, srccsys=&_srccsys
```

# Control Tables

---

- ▶ Data Integration activity must be recorded
  - ▶ structured and accessible
- ▶ three reasons
  - ▶ control
  - ▶ audit
  - ▶ automation





# Control Tables

## Control

- ▶ regulate delta processes
  - ▶ extract completion
    - ▶ store max(mod\_dtm) extracted from source in control table
  - ▶ next extract
    - ▶ retrieve stored max(mod\_dtm) from extract control table

```
1
2  %let _rowCount=0;
3  %let _maxXtrDTM=;
4
5  %let _dtmTypeNumFormat = %eval(20 + &_dtmType).&_dtmType;
6
7  proc sql noprint;
8      connect to teradata ( &bi360_connect_string );
9
10     select coalesce(max_dtm, "&_xtrFromDTM"DT) format &_dtmTypeNumFormat. , row_count
11     into :_maxXtrDTM, :_rowCount
12     from connection to teradata (
13         select max(SYS_SRC_DELTA_DTM) as max_dtm
14             , count(*) as row_count
15         from &_stgSchema..&_stgTable
16     );
17 ;
18 quit;
19
```

# Control Tables




---

## Audit

- ▶ stuff happens, e.g. data anomalies
- ▶ helpful historical data captured in control tables
  - ▶ job start / end times
  - ▶ job order, job failures
  - ▶ reruns, recasts
  - ▶ record counts
    - ▶ extracted, inserted, updated, deleted, exceptions



- ▶ **All** DB tables contain audit data as well

17	 SYS_LOAD_EVENT	SYS_LOAD_EVENT	Numeric	8	11.
18	 SYS_LOAD_DTM	SYS_LOAD_DTM	Numeric	8	DATETIME19.
19	 SYS_UPDATE_DTM	SYS_UPDATE_DTM	Numeric	8	DATETIME19.



# Control Tables

---

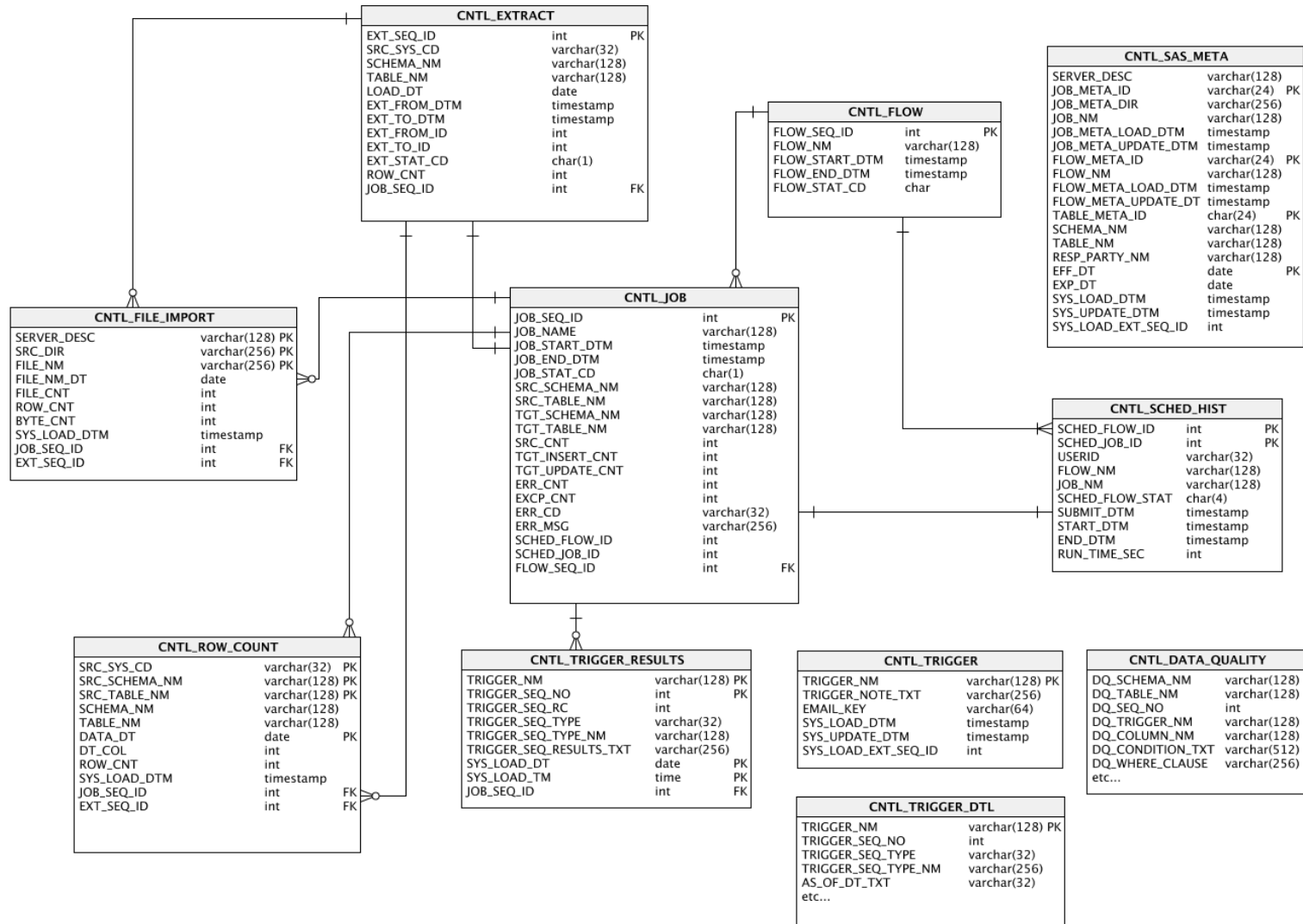
## Automation

- ▶ control data + audit data → automation
- ▶ data quality checks
- ▶ triggers and scheduling

trigger_nm	trigger_seq_no	trigger_type	trigger_type_nm	as_of_dt	trigger_warning_fail_ind
JF_TD_DLY_3002_EXEC_SOURCE_PORT_IN_REPORT	10	FLOW_COMPLETE	TD_KILLB_DIM_MIDATA_DLY	CURRENT_DATE	FAIL
JF_TD_DLY_3002_EXEC_SOURCE_PORT_IN_REPORT	20	FLOW_COMPLETE	TD_FOUND_DIM_DAILY	CURRENT_DATE	FAIL
JF_TD_DLY_3002_EXEC_SOURCE_PORT_IN_REPORT	30	JOB_COMPLETE	J_MI_3024_Load_BM_T_PORT	CURRENT_DATE	FAIL
JF_TD_DLY_3002_EXEC_SOURCE_PORT_IN_REPORT	40	JOB_COMPLETE	J_TD_DLY_1018_SRC_STORE_DEF_Load_S	CURRENT_DATE	FAIL
JF_TD_DLY_3002_EXEC_SOURCE_PORT_IN_REPORT	50	JOB_COMPLETE	J_MI_2133_Load_TD_Dim_Subscriber_I	CURRENT_DATE	FAIL



# Control Tables



# CNTL\_EXTRACT

---

- ▶ each ETL process creates at least one entry
- ▶ describes source data
  - ▶ system
  - ▶ schema . table
- ▶ records each extracts' max(mod\_dtm), row count
- ▶ FK to CNTL\_JOB

CNTL_EXTRACT			
EXT_SEQ_ID	int		PK
SRC_SYS_CD	varchar(32)		
SCHEMA_NM	varchar(128)		
TABLE_NM	varchar(128)		
LOAD_DT	date		
EXT_FROM_DTM	timestamp		
EXT_TO_DTM	timestamp		
EXT_FROM_ID	int		
EXT_TO_ID	int		
EXT_STAT_CD	char(1)		
ROW_CNT	int		
JOB_SEQ_ID	int		FK



# CNTL\_JOB

---

- ▶ job start / end time, status
- ▶ source schema . table
- ▶ target schema . table
- ▶ various row counts
  - ▶ error and exceptions !!
  - ▶ error codes and descriptions
- ▶ FK to CNTL\_FLOW

CNTL_JOB			
JOB_SEQ_ID	int		PK
JOB_NAME	varchar(128)		
JOB_START_DTM	timestamp		
JOB_END_DTM	timestamp		
JOB_STAT_CD	char(1)		
SRC_SCHEMA_NM	varchar(128)		
SRC_TABLE_NM	varchar(128)		
TGT_SCHEMA_NM	varchar(128)		
TGT_TABLE_NM	varchar(128)		
SRC_CNT	int		
TGT_INSERT_CNT	int		
TGT_UPDATE_CNT	int		
ERR_CNT	int		
EXCP_CNT	int		
ERR_CD	varchar(32)		
ERR_MSG	varchar(256)		
SCHED_FLOW_ID	int		
SCHED_JOB_ID	int		
FLOW_SEQ_ID	int		FK



# CNTL\_FLOW

---

- ▶ flows are related jobs
- ▶ flow start / end time, status
- ▶ relatively boring

CNTL_FLOW			
FLOW_SEQ_ID	int		PK
FLOW_NM	varchar(128)		
FLOW_START_DTM	timestamp		
FLOW_END_DTM	timestamp		
FLOW_STAT_CD	char		

## Autocall Macros

- |             |            |                        |
|-------------|------------|------------------------|
| ▶ %flowOpen | %flowClose | Flow open/close jobs   |
| ▶ %jobOpen  | %jobClose  | Job pre / post-code    |
| ▶ %extOpen  | %extClose  | Job pre-code/last step |



# Data Quality & Automation

---

- ▶ ETL control tables capture job info
  - ▶ rows extracted today
  - ▶ delta timestamp values
  
- ▶ But do they tell us we have "good data" ??

Data Activity – more important	Activity Date
IVR – 150k calls / business day	CALL_DT
Orders – 18k / business day	ORDER_DT
SCD2 Customer Master - 5% new rows / day	EFF_DT





# Data Quality & Automation

---

- ▶ **IVR Data**

- ▶ track agents & customers

- ▶ normally 150K calls / business day

- ▶ CALL\_XFER\_DT = yesterday

- ▶ extracted 155K rows of call data today

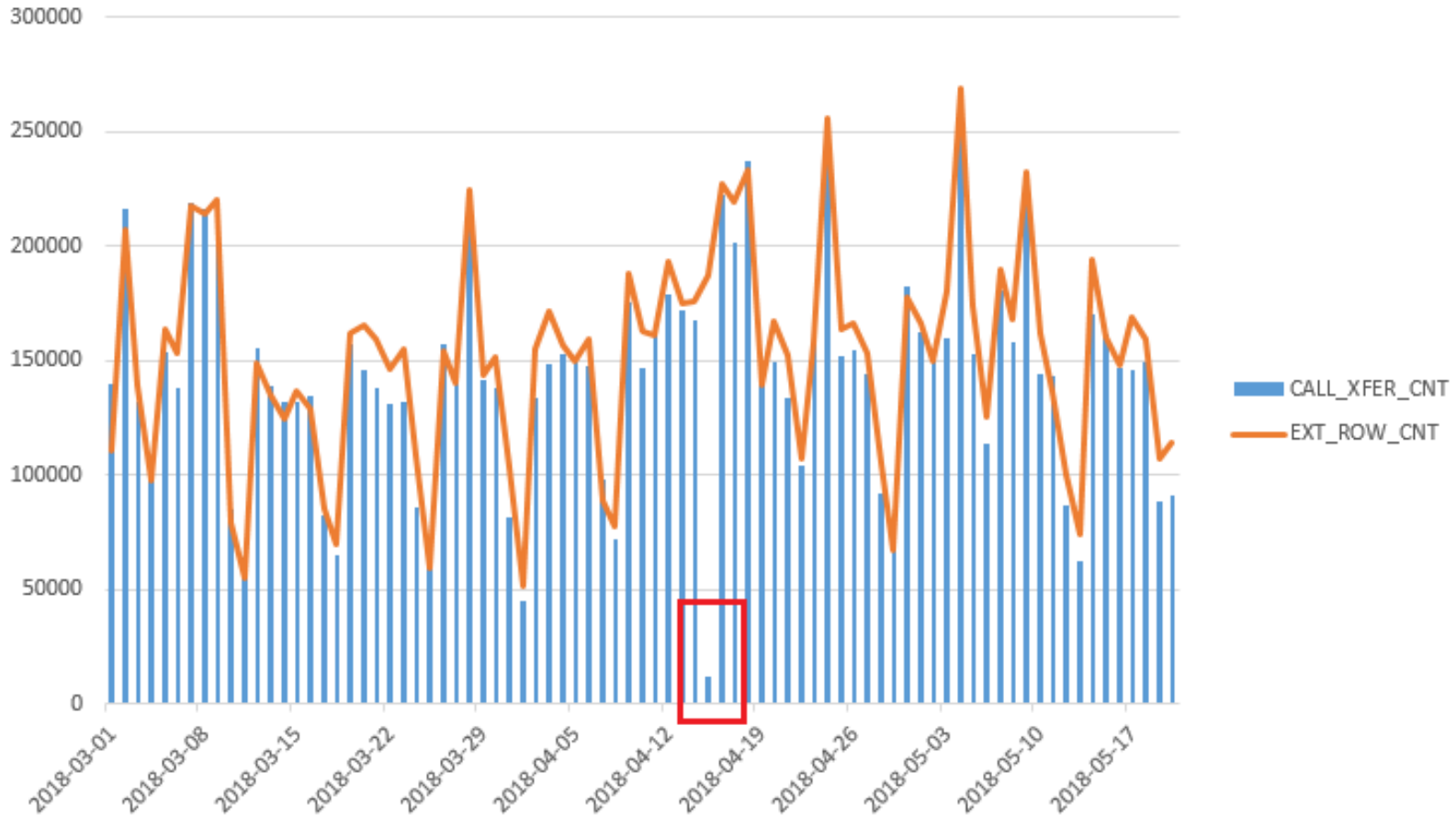
- ▶ source system burped
  - ▶ 152K have CALL\_XFER\_DT of two days ago

- ▶ is IVR data "complete" for today ?

---



# Data Quality & Automation



- ▶ consistent extract row count, call xfer count not so ...



# CNTL\_ROW\_COUNT

---

- ▶ source system identifiers
- ▶ date column and data date value, row count
- ▶ FK to ETL Control tables
- ▶ rolling averages
  - ▶ +/- counts or percentages
- ▶ loaded by macros in jobs

CNTL_ROW_COUNT		
SRC_SYS_CD	varchar(32)	PK
SRC_SCHEMA_NM	varchar(128)	PK
SRC_TABLE_NM	varchar(128)	PK
SCHEMA_NM	varchar(128)	
TABLE_NM	varchar(128)	
DATA_DT	date	PK
DT_COL	int	
ROW_CNT	int	
SYS_LOAD_DTM	timestamp	
JOB_SEQ_ID	int	FK
EXT_SEQ_ID	int	FK



# CNTL\_TRIGGER

- ▶ regulate downstream processes
- ▶ Billing ODS done and Customer Master updated ?
  - ▶ Billing modeled flow can start
- ▶ row count and data quality metrics
- ▶ trigger results

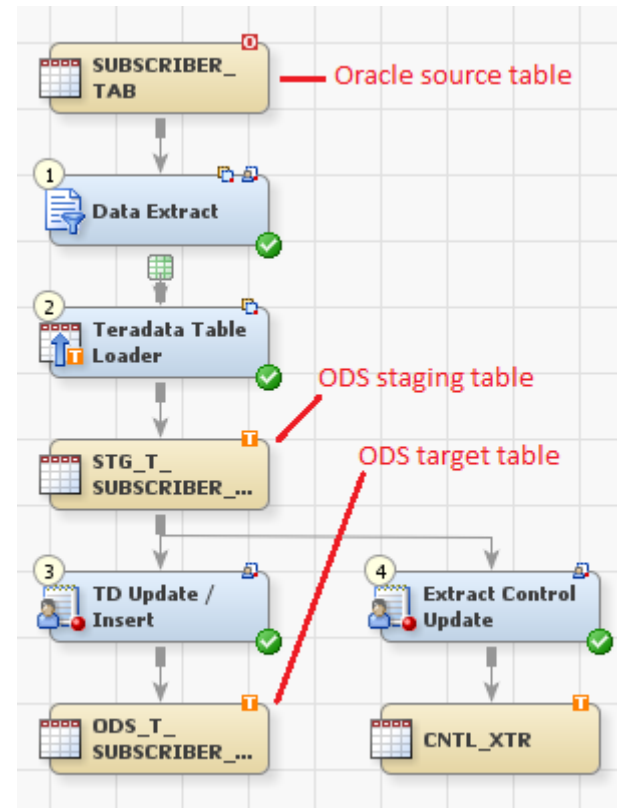
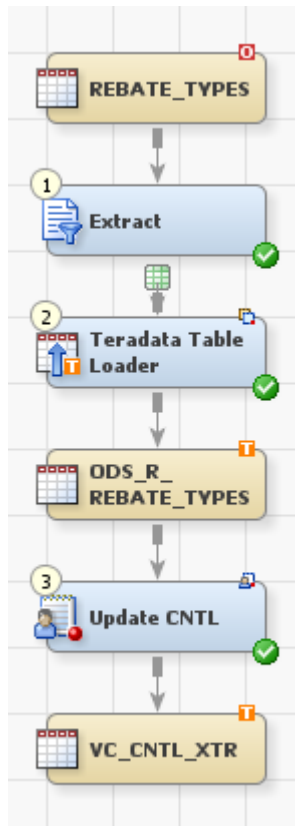
CNTL_TRIGGER	
TRIGGER_NM	varchar(128) PK
TRIGGER_NOTE_TXT	varchar(256)
EMAIL_KEY	varchar(64)
SYS_LOAD_DTM	timestamp
SYS_UPDATE_DTM	timestamp
SYS_LOAD_EXT_SEQ_ID	int

CNTL_TRIGGER_DTL	
TRIGGER_NM	varchar(128) PK
TRIGGER_SEQ_NO	int
TRIGGER_SEQ_TYPE	varchar(32)
TRIGGER_SEQ_TYPE_NM	varchar(256)

CNTL_TRIGGER_RESULTS		
TRIGGER_NM	varchar(128) PK	
TRIGGER_SEQ_NO	int	PK
TRIGGER_SEQ_RC	int	
TRIGGER_SEQ_TYPE	varchar(32)	
TRIGGER_SEQ_TYPE_NM	varchar(128)	
TRIGGER_SEQ_RESULTS_TXT	varchar(256)	
SYS_LOAD_DT	date	PK
SYS_LOAD_TM	time	PK
JOB_SEQ_ID	int	FK

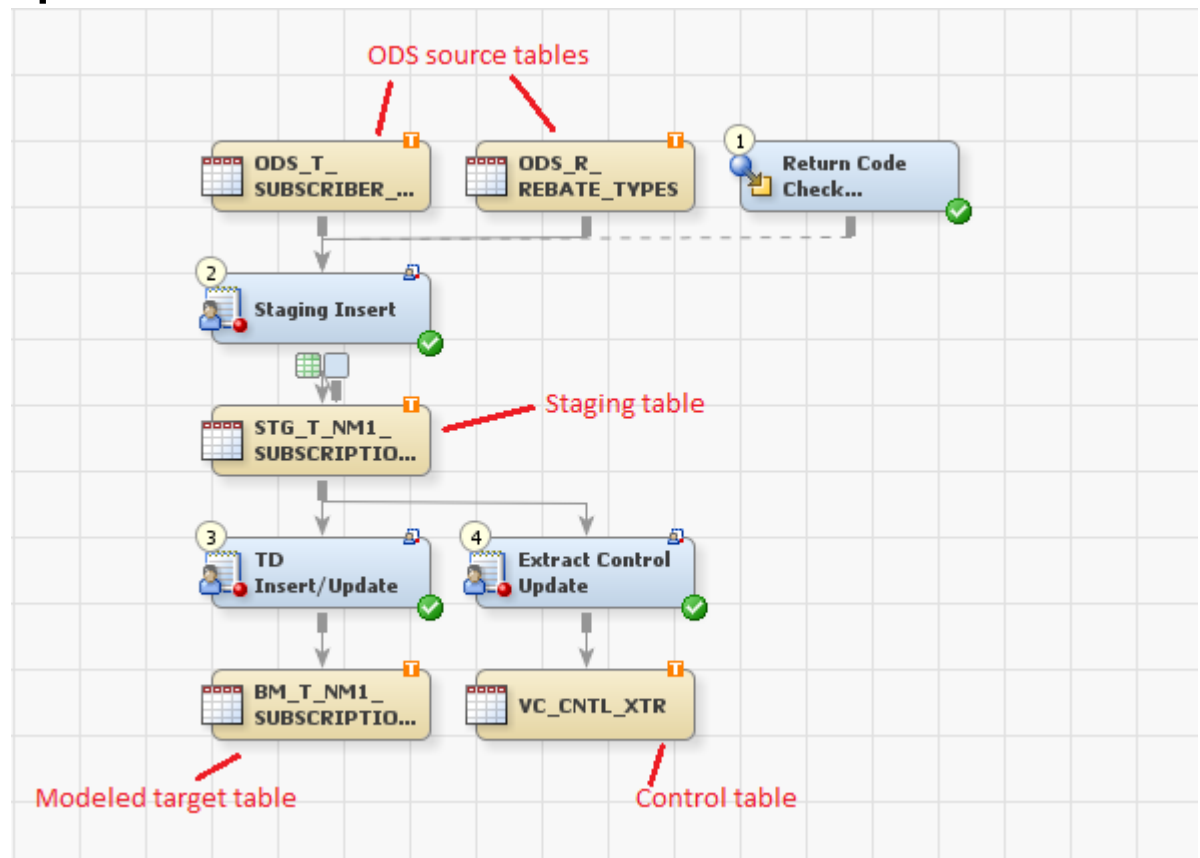
# Focus Functionality

- ▶ One target table per job - KISS
- ▶ ODS layer generally 1:1



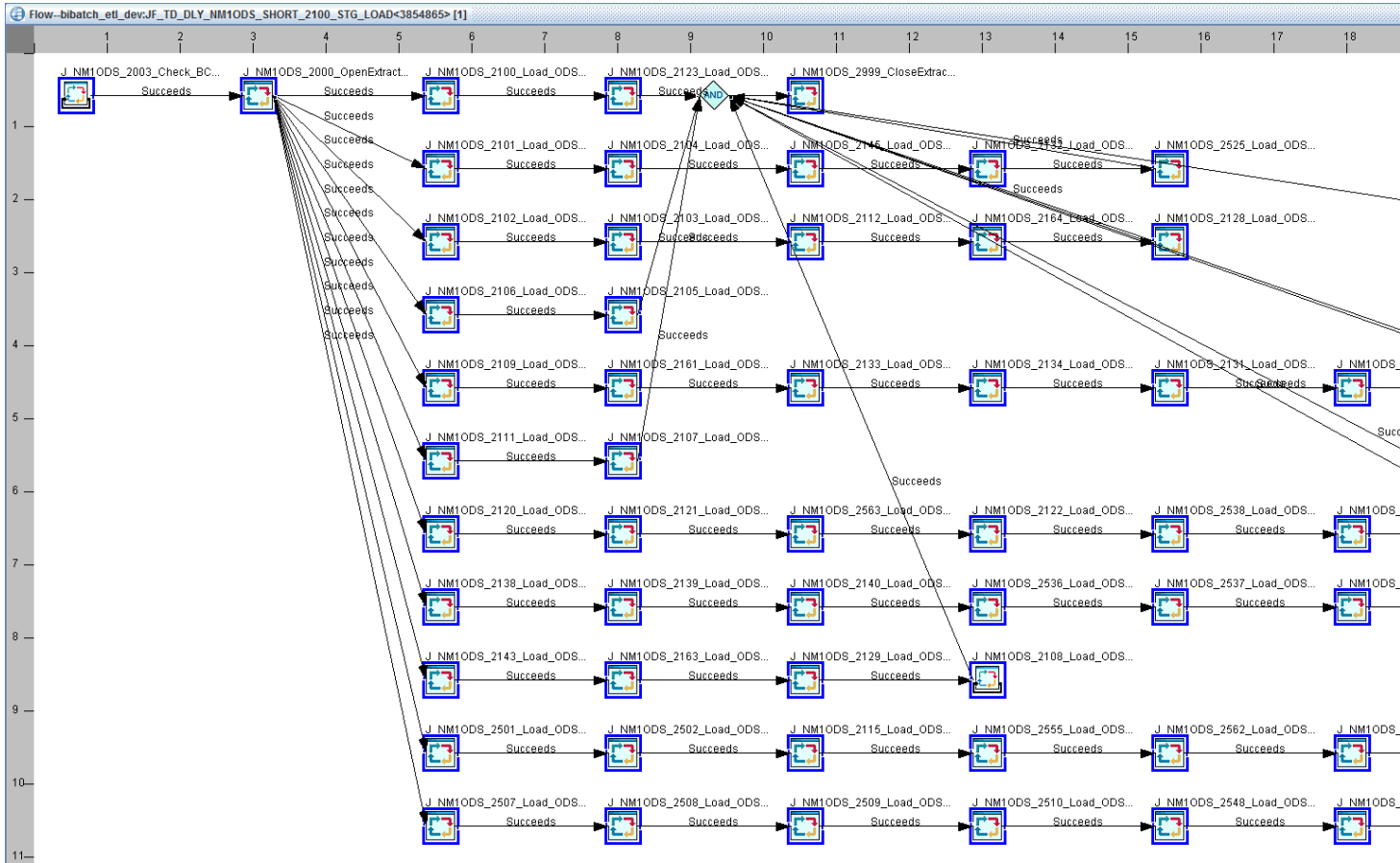
# Focus Functionality

- ▶ ETLs into modeled environment integrate
- ▶ in this case, two input ODS tables
- ▶ one target table



# Focus Functionality

## ▶ Group source to ODS jobs



# Focus Functionality

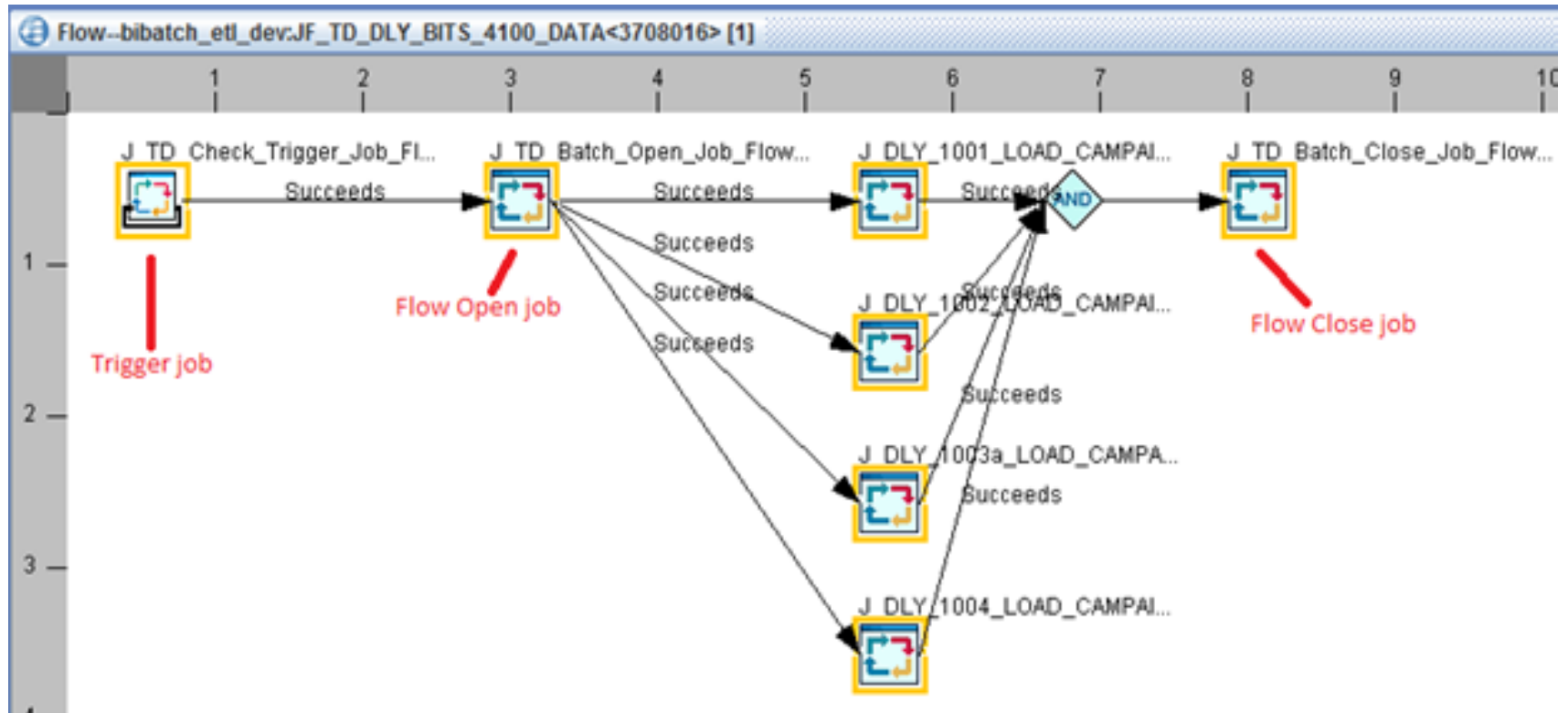
---

- ▶ why group jobs in "stages" like this?
  - ▶ source DB access window restrictions
  - ▶ better sense of data readiness
    - ▶ e.g. Billing ODS data is complete
  - ▶ modeled integration jobs require multiple ODS streams
  - ▶ triggers and data quality
  
- ▶ KISS





# Flow Template



# Job Template

---

- ▶ **pre-code** → establish parameters
  - ▶ database connections
  - ▶ macro variables for source, staging, target schemas / tables
- ▶ **job code uses macro variables assigned in pre-code**
- ▶ **autocall macros**
  - ▶ control table open/close,
  - ▶ for common ETL activities
    - ▶ table truncation, drop table, upsert
  - ▶ capture data quality info



# So What ?!?

---

- ▶ standards are hard
  - ▶ compliance is harder
- ▶ tyranny of the urgent
  - ▶ git 'er done !
- ▶ "one-off"
- ▶ does it pay off ?



# So What ?!?

---

- ▶ **templates & standards reduce:**
  - ▶ onboarding time
  - ▶ development time
  - ▶ maintenance effort
  - ▶ errors
  - ▶ test cycles
  - ▶ concentrate on creative business solutions
  
- ▶ **control & audit**
  - ▶ what happened ?
  - ▶ data driven automation



# So What ?!?

---

- ▶ Control Tables answer questions
  - ▶ data anomalies
    - ▶ counts, values, trends, gaps
    - ▶ Teradata fastload errors – missing data
  - ▶ dashboards & reports
    - ▶ batch schedule progress
    - ▶ data readiness and quality
    - ▶ automatic notification



# So What ?!?

---

- ▶ Control Tables answer questions
  - ▶ identifying process gaps
    - ▶ jobs updating DB tables with no stats
    - ▶ escalating processing time
    - ▶ job failures / reasons
      - only occurring on *one* node ?!?
    - ▶ schedule optimization



# Conclusion

---

- ▶ best practices beget good data
- ▶ standards & reliable controls pay dividends
  - ▶ through entire SDLC
- ▶ KISS
  - ▶ Keep It Simple & Standardized



# Author

---

Harry Droogendyk  
Stratia Consulting Inc.

[www.stratia.ca/papers](http://www.stratia.ca/papers)

