

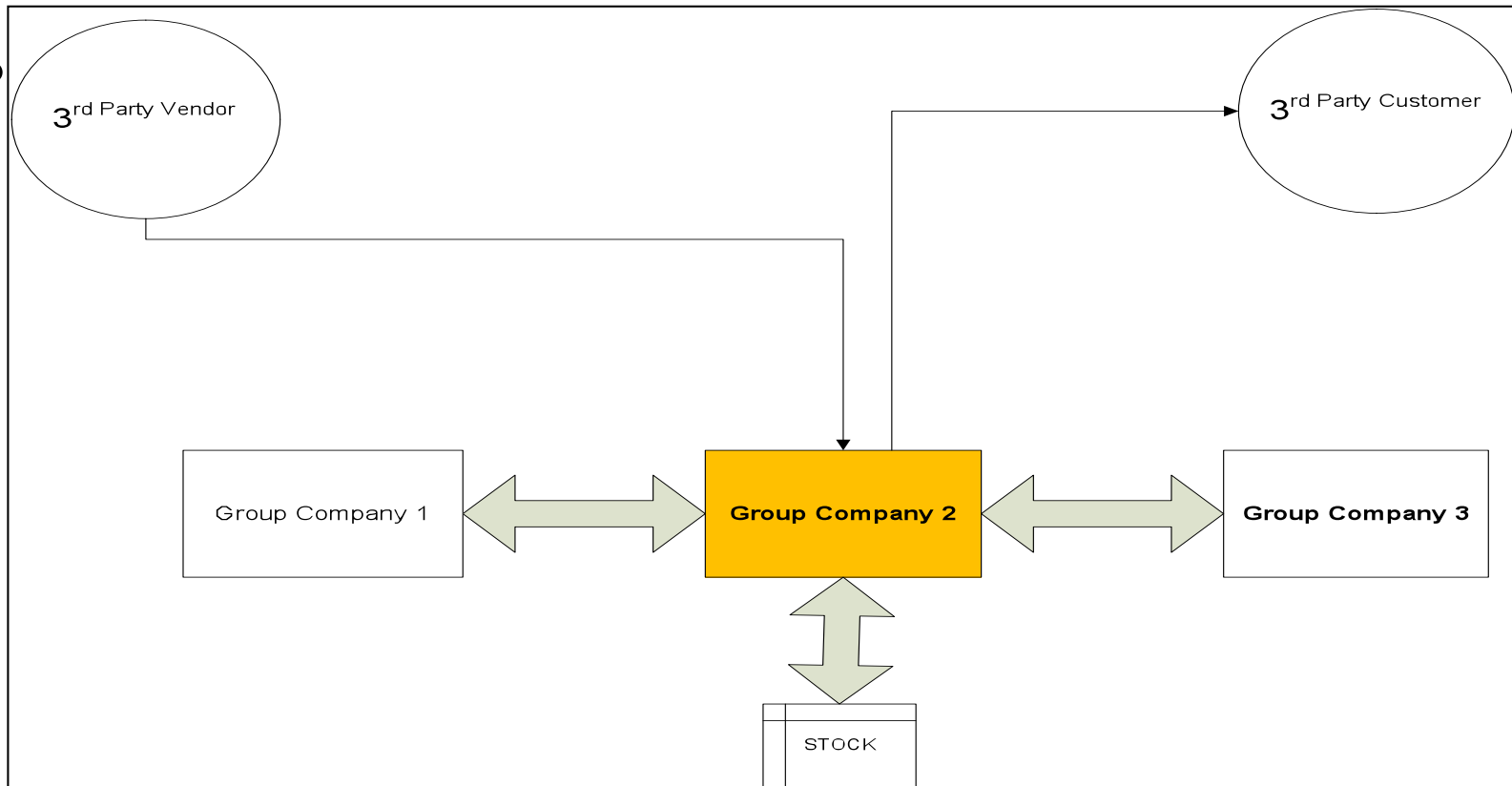
The Data

- Large Number of Workbooks
- Each Workbook has multiple worksheets
- Transaction worksheets have large (LARGE) number of lines (millions of records across the set)
- Conclusion - Only IDEA is going to be any use

The Scope

- Source system is a multi-entity SAP ERP where the entities interact to form Supply Chains
- Each Workbook contains all FI postings in a period for an entity in separate Inputs and Outputs worksheets
- Aim is to combine all the entities and all periods and to profile Supply Chains

Supply Chains



Initial IDEA Challenges

- The Workbooks have passwords and imports failed
- The Workbooks have column totals in value fields at their foot which add non-standard records
- We decide to produce a cloned set of workbooks, check their totals but remove the total rows

First Steps 1

- Controls !!! If we don't know where we started we will never realise if we mess up the import and subsequent work
- *Thankfully these are present as value column totals and check out*
- Source References - if we don't reference the source entity and period we can't refer back to original data (eg date corruption on import) - *so we added these*

Next IDEA Challenges

- We go for broke - load everything into IDEA *but*
- Append ---The Workbooks are discovered to differ between entities - some have additional columns, some missing columns
- Import Error - Some Workbooks have dodgy date formats and bad data (vlookup failures) so corrupted data is imported

First Steps -1

- Approach now was to take a step back and fix data problems (eg loss of vlookup references and dates) in the workbooks first
- The “go for broke” exercise had also identified that in addition to additional/missing columns we had different field types and field lengths being imported in what appeared to be identical columns ... *Oh well*

Back to Square One

- Then we had a think
- We at least had discovered most if not all of the data problems
- In the absence of some automatic normalisation tools we were going to have to use cunning and guile to get where we need to be ... *To start !*

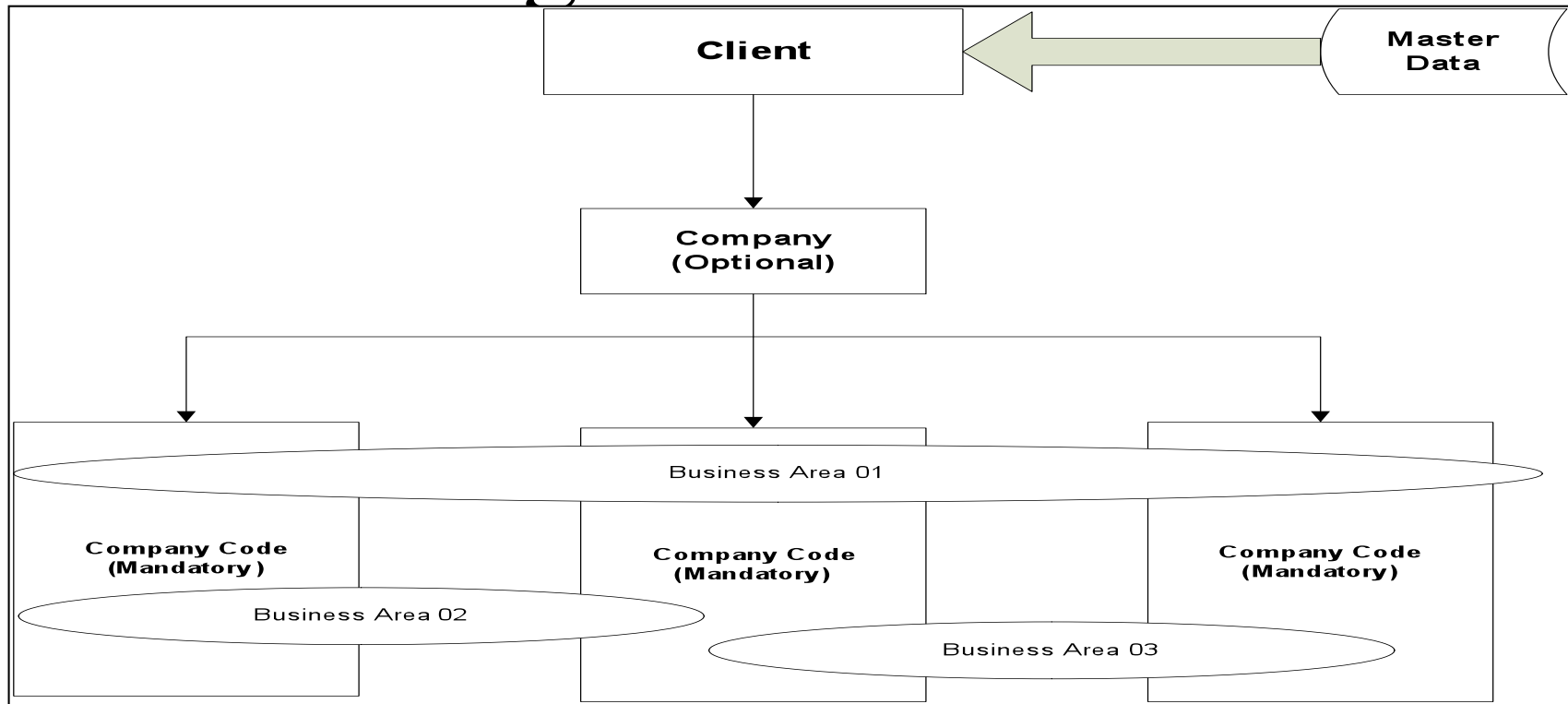
Pilot the Solution

- Since aggregating all the different worksheet formats was too difficult but seemed (*fingers crossed*) mainly to be a problem between entity workbook sets, if we could identify the subset of records we needed on the basis of a pilot exercise on one entity, we could load the sets independently and create extract subsets towards a final aggregation

Welcome to SAP

- SAP is a highly integrated ERP with many sub-modules
- Once transactions in the sub-modules cause an accounting event, a posting ends up in the accounting module FI
- There are lots of postings and we had buckets of them labelled “inputs” and “outputs”

SAP Organisational Model



What we knew 1

- The entities for which we had FI postings were “company codes” which are legal entities with their own P&L and Balance Sheet for external accounting
- In SAP “company” is something different and corresponds to the optional grouping of company codes for consolidation reporting - think accounting groups

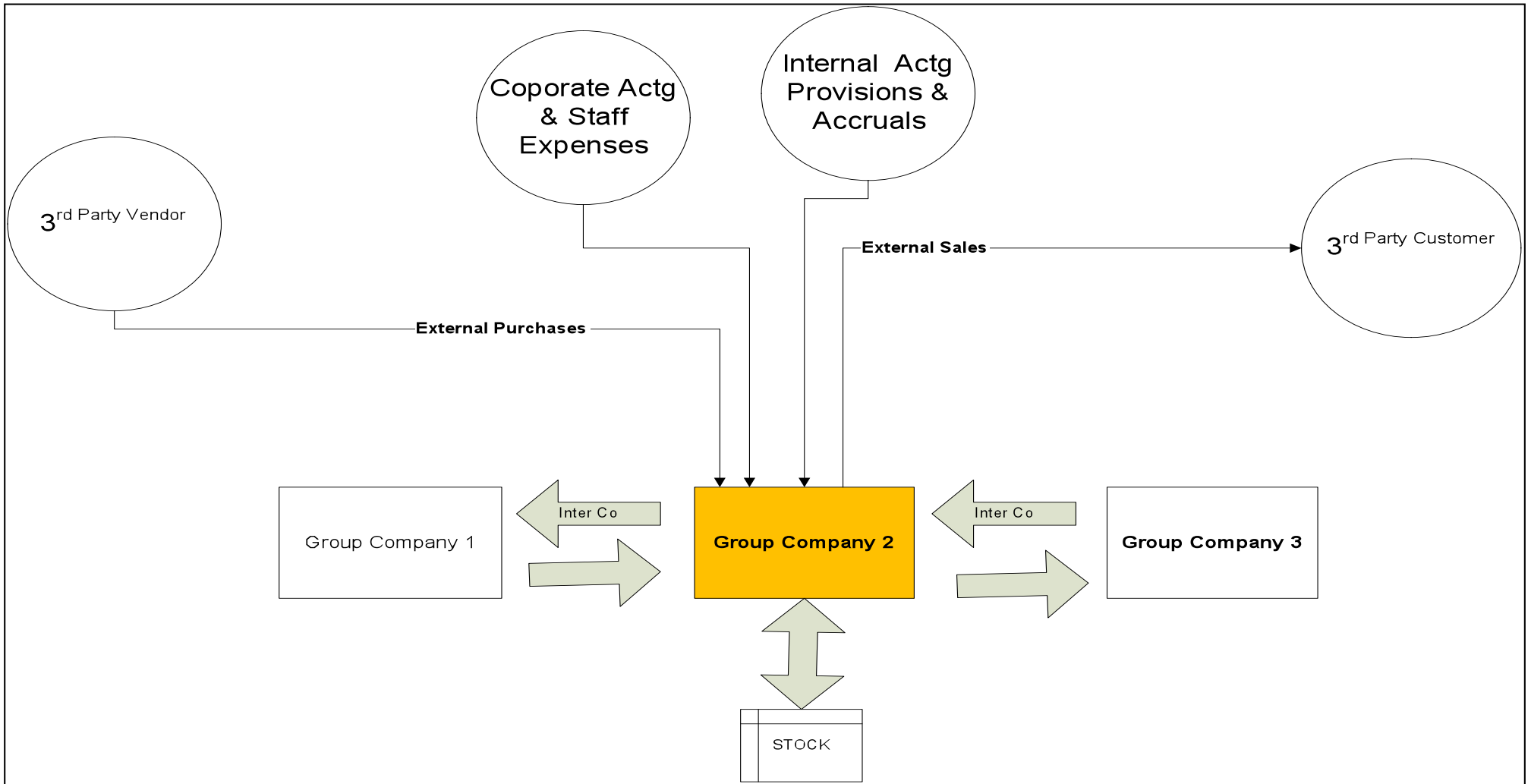
What we knew 2

- This organisation also implemented the optional “Business Areas” which can be used across “Company Codes” , so a coding of business activities shared by multiple legal entities
- Thankfully many of the other optional SAP organisational types eg “Sales Areas” did not make an appearance (...*so far...*)

The Summarisation Trap

- Using summarisation on key fields to get a handle on the data is an exploratory data technique but with large numbers of diverse transaction types you need to have a model of the data to work against
- Sometimes you keep summarising in the hope that a pattern will emerge - unfortunately with big data sets there are lots of patterns *But what do they mean ?*

The Model



INPUTS	Explorative Analysis	OUTPUTS	Explorative Analysis
DocumentNo	first2char summary	DocumentNo	first2char summary
Business Area	summary	Business Area	summary
Posting Date		Pstng Date	
Reference		Reference	
Doc. Date		Doc. Date	
Type	summary	Type	summary
Vendor	summary	Customer	summary
Vendor Name	summary	Customer Name	summary
VAT Registration No.		VAT Registration No.	
Cty	Vendor Country	Cty	Customer Country
G/L	summary	G/L	summary
GL Account Name	summary	GL Account Name	summary
Values incl FX			
Assignment	as PO with line number	Assignment	
PO Number			
Document Header Text		Document Header Text	
Text	summary	Text	summary
User field A		User field A	summary
Tx	summary	Tx	summary
VAT rate	summary	VAT rate	summary
Plnt		Plnt	
Cty		Cty	
Ship-To		Ship-To	
Cty		Cty	
Material	summary	Material	summary
Material Name	summary	Material Name	summary

More is sometimes less !!

	Name
inputs Action reason.xlsx	outputs BUSA sumy.xlsx
inputs Material VRate.XLS	outputs Check sumy.xlsx
inputs Material.XLS	outputs Checks Performed.xlsx
inputs Plnt Shipto.XLS	outputs customer sumy.xlsx
inputs Ref sumy.xlsx	outputs DocFirst2 sumy.XLS
Inputs Round 1000 No material.XLS	outputs GL sumy.xlsx
Inputs round 1000.XLS	outputs material sumy.xlsx
inputs round1000 sumy.XLS	outputs Plnt Shipto Material Su...
inputs TX Check and Correct.xlsx	outputs reason is incorrect VAT ...
inputs TX sumy.XLS	outputs reason sumy.xlsx
inputs Type sumy.XLS	outputs round1000.xlsx
inputs Vat Rate sumy.XLS	outputs Text Sumy.xlsx
inputs vendor blank and material blank and materia...	outputs TX sumy.XLS
inputs vendor blank and material blank.xlsx	outputs Type sumy.xlsx
inputs vendor blank and material non-blank.xlsx	outputs UserField A sumy.xlsx
inputs vendor blank.xlsx	outputs VAT Rate sumy.xlsx
Inputs Vendor Sumy.XLS	
inputs VRate Correct VC.XLS	

Finding Signposts

- Happily we were now in a position to ask some sensible questions about the codings used for key fields (SAP provides a default set but they are often extensible)
- Business Area
- Document Type
- Document Number (the first 2 digits are significant for which ledger is posted)

Directing Extraction

- Using the model and the codes we were able to identify the elements of the model in the pilot entity inputs and outputs
- We had initially tried the Vendor/Customer Fields but suspected they were only part of the story
- We needed to leave all internal postings behind but keep relevant intercompany ones

Not There Yet But....

- So for the pilot entity we had developed a relationship between input and output sets following the model which can be reconciled to the original controls and through them to the analysis of P&L and Balance Sheet postings for periods
- So positive progress

Codes to use

All Records with Vendor/Customer Names (External & Internal)

Business Area			PLUS		DOCFIRST		Inwards	Outwards
	Inward	Outward	Documents Type	Inward	Outward	2		
<blank>	No	No	CI	No	Yes	16	Yes	Yes
0	No	No	CQ	No	Yes	21	Yes	Yes
10	Yes	Yes	DP	Yes	No	22	Yes	Yes
11	Yes	Yes	GA	Yes	Yes	31	Yes	Yes
14	Yes	Yes	GD	Yes	No	32	Yes	Yes
2	Yes	Yes	GE	Yes	Yes	36	Yes	Yes
20	No	No	RR	Yes	No	37	Yes	Yes
3	Yes	Yes	RV	No	Yes	44	Yes	Yes
4	Yes	Yes	V3	Yes	No	51	Yes	Yes
5	No	No	VL	Yes	No	52	Yes	Yes
6	Yes	Yes	VQ	Yes	No	54	Yes	Yes
8	Yes	Yes				55	Yes	Yes
9	Yes	Yes				59	Yes	Yes

Further IDEA Challenges

- We now have to apply the extraction criteria to the other workbook sets individually (mostly avoiding the inconsistent fields) and then add these extracts together (overcoming the field type and length problems with append).... Maybe a job for a batch process ???