
Business Intelligence

on IBM i

QUSER Meeting

April, 2015

Session 1

Welcome!

Today's Speaker:



Alan Jordan

*Director of Data Warehouse Technologies,
HelpSystems*

Before we start...



If you are new to Business Intelligence (BI), or Data Warehousing (DW), we need to take a few minutes to get an understanding of what we are talking about and why it's important...

- What is business intelligence?
- Why do I need more than just a query tool?
- Why can't I buy an 'out of the box' solution?

Hopefully we can answer these questions (and others like them) for you....

Business Intelligence

Business intelligence (BI) is a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions.

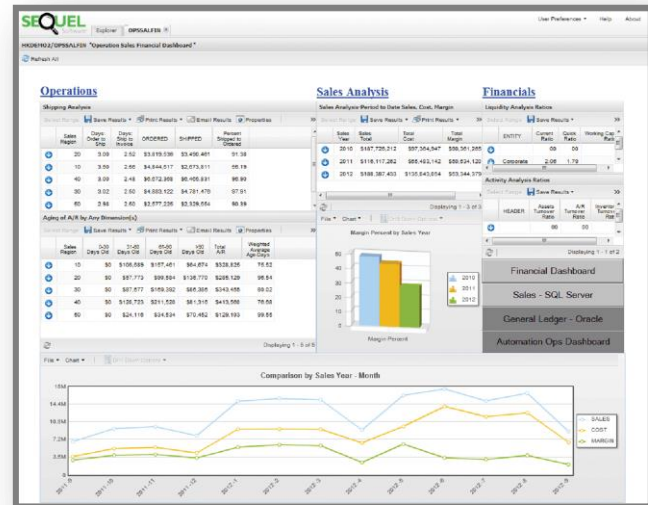
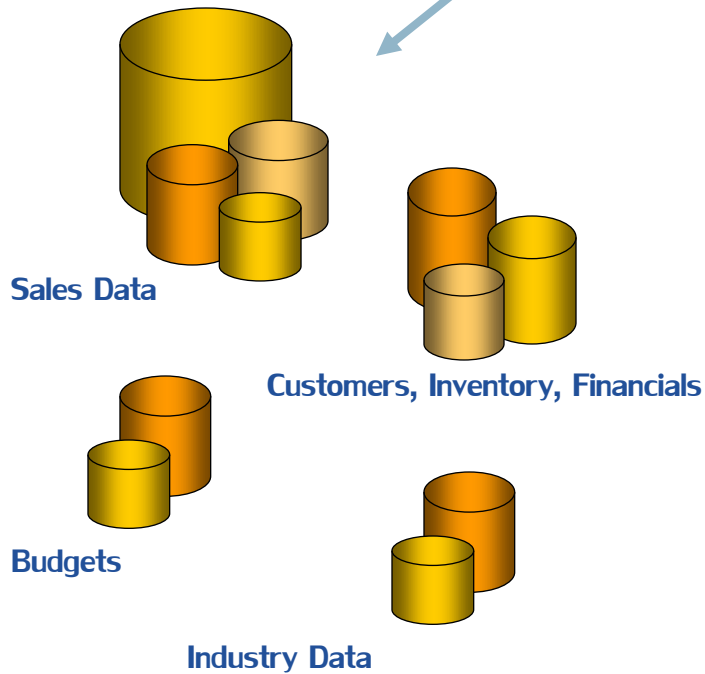
BI applications include data warehouses, data marts, query and reporting, online analytical processing (OLAP), statistical analysis, forecasting, data mining and predictive analytics.

Business intelligence applications can be:

- ✓ Mission-critical and integral to an enterprise's operations or occasional to meet a special requirement
- ✓ Enterprise-wide or local to one division, department, or project

Business Intelligence

In other words, a set of tools and technologies to get from here to here



BI Reporting & Analytics Tools

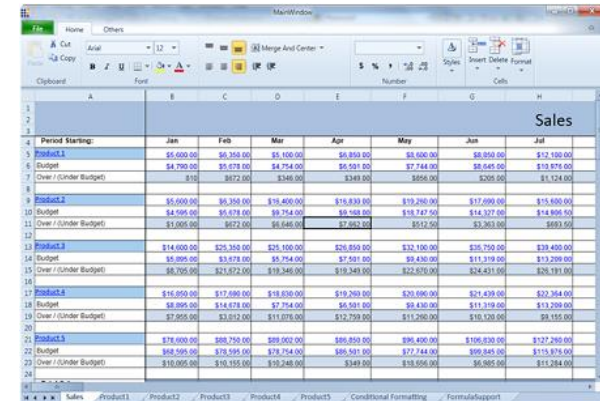
There are many different tools available; each with a different feature set and sometimes a different purpose.

Let's take a look at some of the *types* of these tools.

BI Reporting & Analytics Tools

1. Spreadsheets

- Every organization has dozens, or more likely hundreds or even thousands of spreadsheets
- Some may be ‘sanctioned’ and shared within the company or department
- Many will be private, jealously guarded, secret stashes of data
- Almost always a disjointed, unreliable approach to BI
- Often leads to ‘spreadsheet hell’
- In Australia, referred to as a ‘Claytons’ implementation

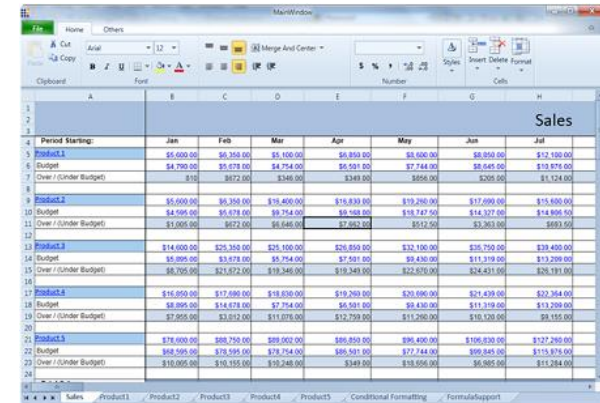


	Jan	Feb	Mar	Apr	May	Jun	Jul
Product 1	\$5,000.00	\$6,350.00	\$1,100.00	\$8,850.00	\$5,000.00	\$8,800.00	\$12,100.00
Budget	\$4,700.00	\$5,070.00	\$4,754.00	\$5,591.00	\$7,744.00	\$8,045.00	\$10,875.00
Over / Under Budget	\$310.00	\$1,280.00	\$346.00	\$3,259.00	\$275.00	\$755.00	\$1,225.00
Product 2	\$5,000.00	\$6,350.00	\$18,400.00	\$15,833.00	\$19,200.00	\$17,000.00	\$15,600.00
Budget	\$4,000.00	\$5,070.00	\$9,754.00	\$8,108.00	\$18,747.00	\$14,327.00	\$14,800.00
Over / Under Budget	\$1,000.00	\$1,280.00	\$8,646.00	\$7,725.00	\$512.00	\$2,673.00	\$860.00
Product 3	\$14,000.00	\$25,350.00	\$25,100.00	\$20,600.00	\$32,100.00	\$18,750.00	\$39,400.00
Budget	\$5,000.00	\$3,070.00	\$5,754.00	\$7,591.00	\$9,430.00	\$11,319.00	\$13,200.00
Over / Under Budget	\$9,000.00	\$22,280.00	\$19,346.00	\$13,009.00	\$22,670.00	\$7,431.00	\$26,200.00
Product 4	\$16,000.00	\$17,000.00	\$18,000.00	\$19,200.00	\$20,000.00	\$21,400.00	\$22,300.00
Budget	\$8,000.00	\$14,070.00	\$7,754.00	\$5,591.00	\$9,430.00	\$11,319.00	\$13,200.00
Over / Under Budget	\$8,000.00	\$3,930.00	\$10,246.00	\$13,609.00	\$10,570.00	\$10,081.00	\$9,100.00
Product 5	\$70,000.00	\$68,750.00	\$89,000.00	\$66,800.00	\$86,400.00	\$106,800.00	\$127,200.00
Budget	\$68,000.00	\$70,070.00	\$78,754.00	\$68,591.00	\$77,744.00	\$89,845.00	\$115,875.00
Over / Under Budget	\$2,000.00	\$18,680.00	\$10,246.00	\$-1,791.00	\$8,656.00	\$16,955.00	\$11,325.00

BI Reporting & Analytics Tools

• Spreadsheets

- Every organization has dozens, or more likely hundreds or even thousands of spreadsheets
- Some may be ‘sanctioned’ and shared within the company or department
- Many will be private, jealously guarded, secret stashes of data
- Almost always a disjointed, unreliable approach to BI
- Often leads to ‘spreadsheet hell’
- In Australia, referred to as a ‘Claytons’ implementation



	Jan	Feb	Mar	Apr	May	Jun	Jul
Product 1	\$5,000.00	\$6,350.00	\$1,100.00	\$8,850.00	\$5,600.00	\$8,800.00	\$12,100.00
Budget	\$4,700.00	\$5,075.00	\$4,754.00	\$5,591.00	\$7,744.00	\$6,645.00	\$10,875.00
Over / Under Budget	\$310.00	\$1,275.00	\$346.00	\$3,259.00	\$-2,144.00	\$2,155.00	\$1,225.00
Product 2	\$5,000.00	\$6,350.00	\$18,450.00	\$15,833.00	\$19,280.00	\$17,090.00	\$15,600.00
Budget	\$4,095.00	\$5,075.00	\$9,754.00	\$8,168.00	\$18,747.00	\$14,327.00	\$14,800.00
Over / Under Budget	\$905.00	\$1,275.00	\$8,696.00	\$7,665.00	\$513.00	\$2,763.00	\$800.00
Product 3	\$14,000.00	\$25,350.00	\$25,100.00	\$26,650.00	\$32,100.00	\$16,750.00	\$39,450.00
Budget	\$5,095.00	\$3,075.00	\$5,754.00	\$7,591.00	\$9,430.00	\$11,319.00	\$13,200.00
Over / Under Budget	\$8,905.00	\$22,275.00	\$19,346.00	\$19,059.00	\$22,670.00	\$5,431.00	\$26,250.00
Product 4	\$16,000.00	\$17,090.00	\$18,650.00	\$19,260.00	\$20,090.00	\$21,430.00	\$22,364.00
Budget	\$8,095.00	\$14,075.00	\$7,754.00	\$5,591.00	\$9,430.00	\$11,319.00	\$13,200.00
Over / Under Budget	\$7,905.00	\$3,015.00	\$10,896.00	\$13,669.00	\$10,660.00	\$10,111.00	\$9,164.00
Product 5	\$78,000.00	\$68,750.00	\$89,002.00	\$66,850.00	\$86,400.00	\$106,830.00	\$127,280.00
Budget	\$68,095.00	\$70,075.00	\$78,754.00	\$66,591.00	\$77,744.00	\$69,845.00	\$115,875.00
Over / Under Budget	\$10,005.00	\$-1,325.00	\$10,248.00	\$3,259.00	\$8,656.00	\$36,985.00	\$11,405.00



Claytons - the drink you have when you're not having a drink!

BI Reporting & Analytics Tools

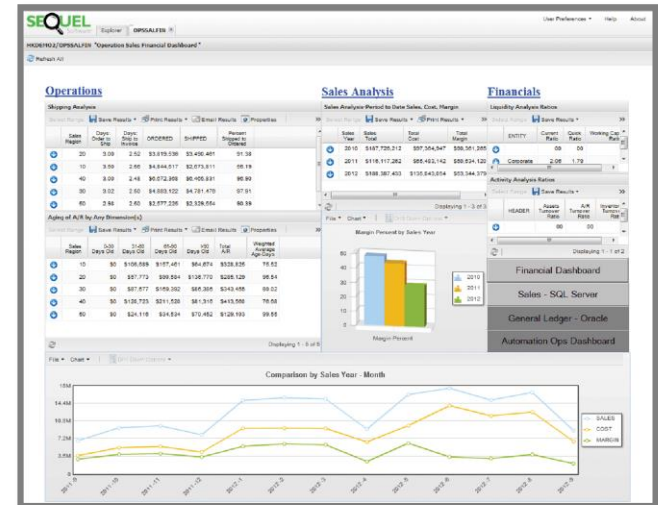
So, you're telling me that the sales forecast you submitted to me last week was based on your intern's fancy formula in this spreadsheet, and you don't know how he came up with it?



BI Reporting & Analytics Tools

• Query & Reporting Tools

- The workhorse of most BI implementations
- Provide both 'canned' reports and drill down capabilities to 'slice & dice'
- Modern tools are very feature rich:
 - Many have web-based (browser) interfaces
 - Should provide dashboards
 - Should be mobile enabled
 - Should be able to email/distribute reports



BI Reporting & Analytics Tools



- Query/400 is **STILL** one of the most commonly used tools in the IBM i community!

```

Work with Queries
Type choices, press Enter.
Option . . . . . -          1=Create, 2=Change, 3=Copy, 4=Delete
                               5=Display, 6=Print definition
                               8=Run in batch, 9=Run
Query . . . . . : _____ Name, F4 for list
Library . . . . . : QGPL      Name, *LIBL, F4 for list

F3=Exit      F4=Prompt      F5=Refresh      F12=Cancel
(C) COPYRIGHT IBM CORP. 1988
  
```

BI Reporting & Analytics Tools

- Query/400 is **STILL** one of the most commonly used tools in the IBM i community!

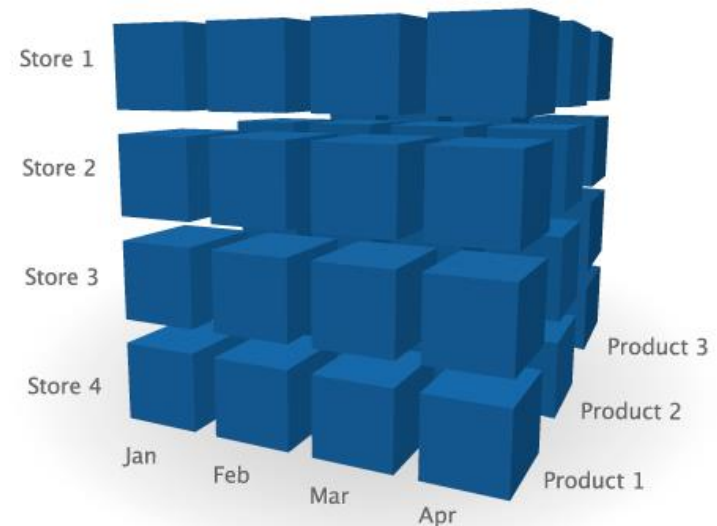


Where it really belongs

BI Reporting & Analytics Tools

- **OLAP Tools**

- Multidimensional databases
- Provide very fast response times when slicing into the data
 - Everything is pre-calculated
 - Access speed comes at the cost of load time and storage requirements
- Proprietary technology/data storage
 - Cannot be accessed via SQL
- Value has diminished over the past decade
 - Modern systems are much faster
 - Query & Reporting tools can drill up or down from one summary level table to another



BI Reporting & Analytics Tools

- **OLAP Tools**
 - The structures built by OLAP tools are often referred to as 'cubes', suggesting 3 axes.

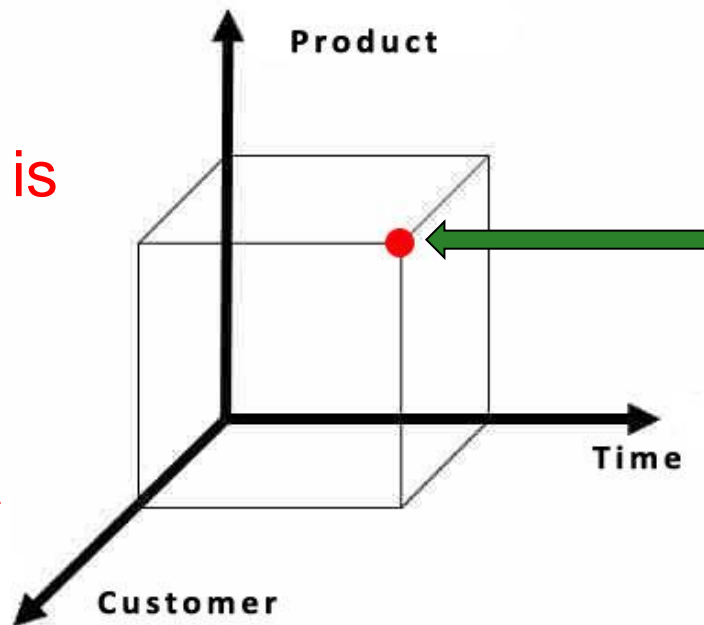
BI Reporting & Analytics Tools

- **OLAP Tools**

- The structures built by OLAP tools are often referred to as 'cubes', suggesting 3 axes.

A 3-dimensional structure (cube) is easy for us to visualize.

Try visualizing a 12-dimensional structure!



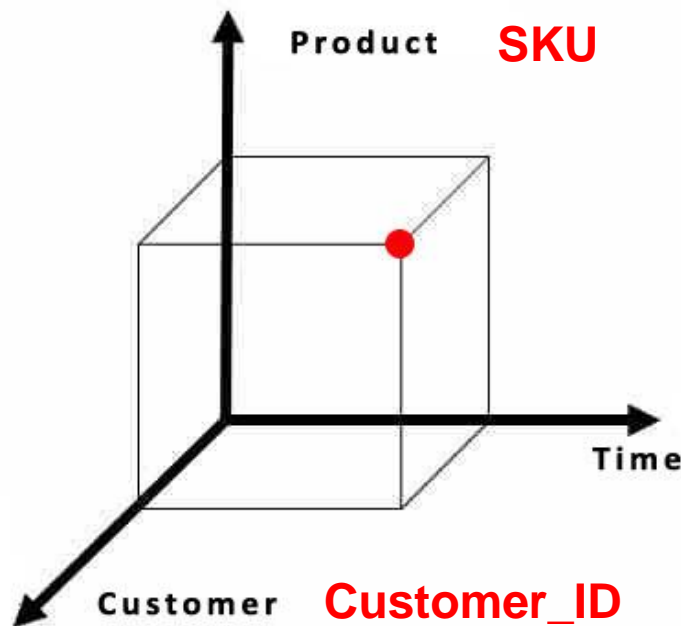
The intersection of the axes (dimensions) contains a data point (fact)

BI Reporting & Analytics Tools

- **OLAP Tools**

- The structures built by OLAP tools are often referred to as 'cubes', suggesting 3 axes.

Data in a cube is never at detail level!



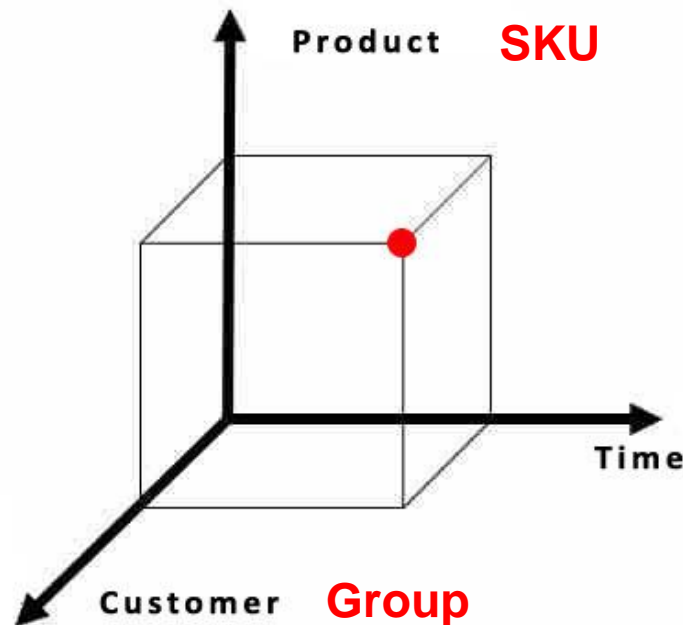
This combination is not sensible!

Date

BI Reporting & Analytics Tools

- **OLAP Tools**

- The structures built by OLAP tools are often referred to as 'cubes', suggesting 3 axes.



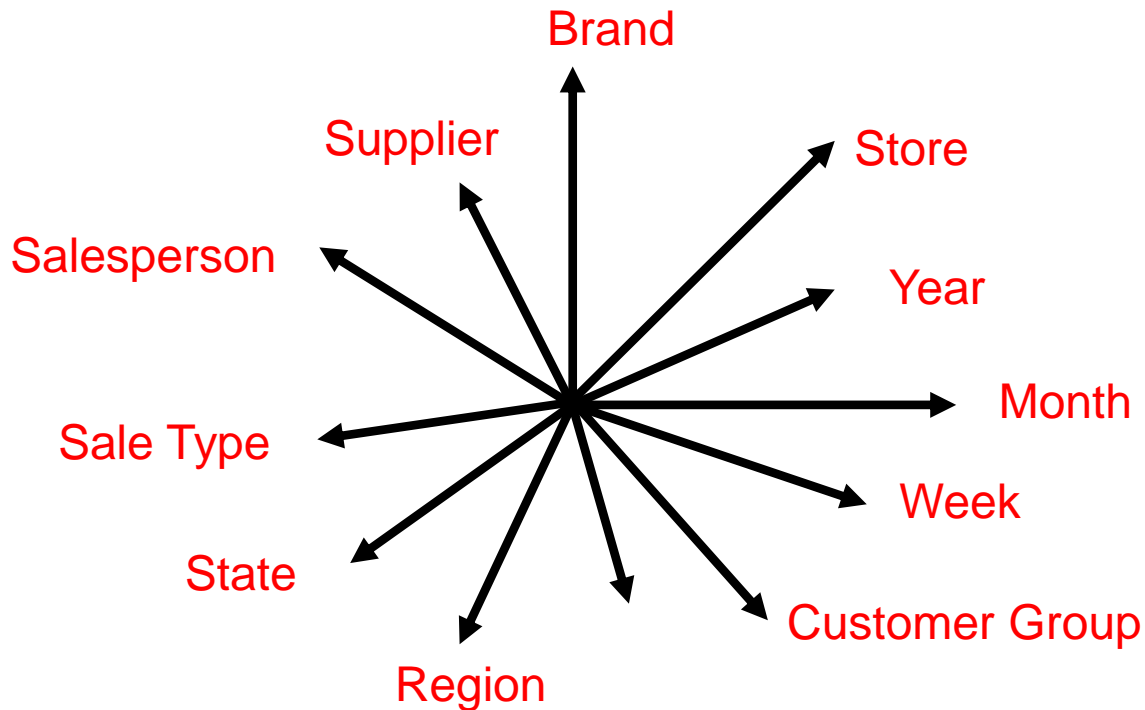
This combination is good!

Month

BI Reporting & Analytics Tools

- **OLAP Tools**

- In most cases, many axes (dimensions) will be implemented



BI Reporting & Analytics Tools

• Data Mining

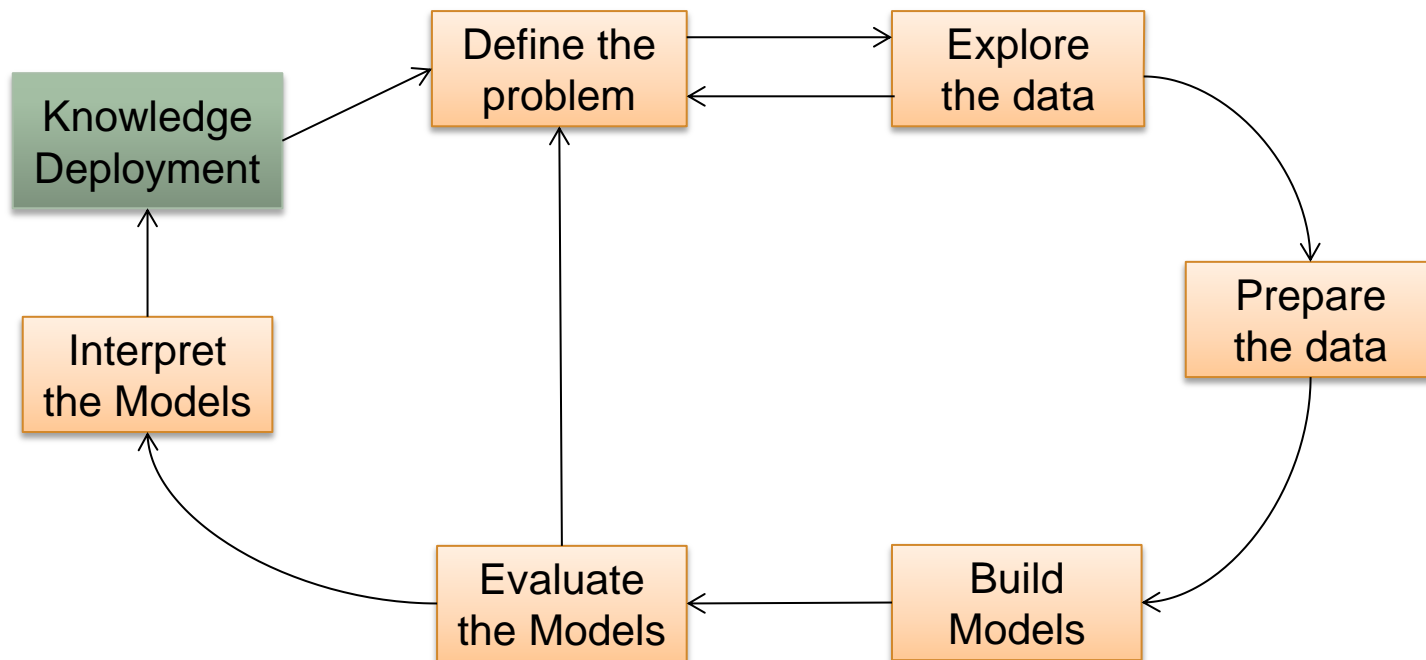
- The discovery of patterns in large sets of data, using statistical analysis.
- Commonly misused as a buzzword – the majority of organizations who say they are doing data mining are not!
- Requires very careful preparation of the data (to be mined). This can take weeks or even months
- NOT something you usually have the skills to do in-house.
 - A consulting engagement
 - Can be very expensive to undertake



BI Reporting & Analytics Tools

- **Data Mining**

- The process is quite involved



BI Reporting & Analytics Tools

- **Data Mining**

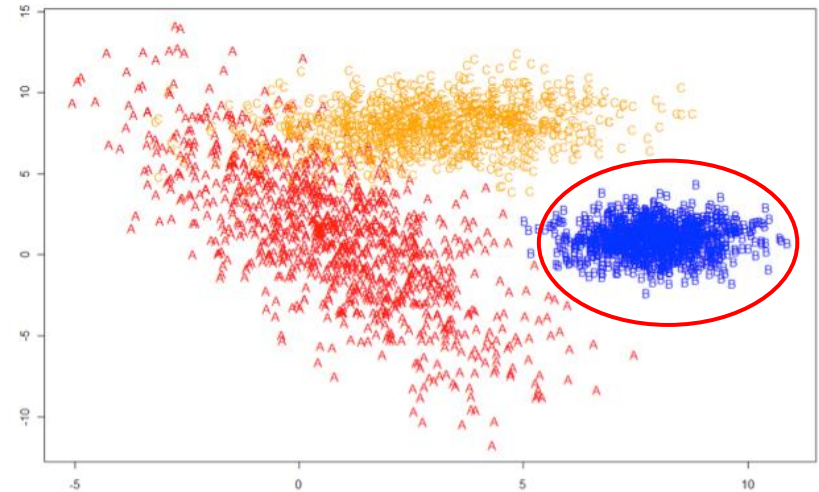
- Beware of fools gold!

An insurance company mined it's data to understand who its customers were.

Results showed a concentration of customers who were 24-30 years old and drove 40 minutes or more to work each day.

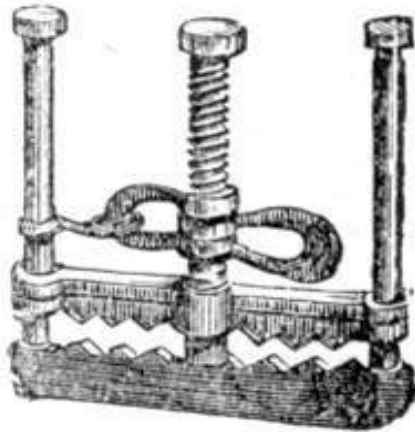
Marketing used this data to plan drive-time radio advertising to go after this demographic

Luckily, before it went to air, someone thought to analyze this group of customers using conventional BI - and found they were the LEAST profitable of its customers!



BI Reporting & Analytics Tools

- **Data Mining has its critics**



Torture numbers enough
and they'll confess to
anything!



There are Lies, Damned Lies
... and Statistics

Mark Twain

BI Reporting & Analytics Tools

- **Predictive Analytics**

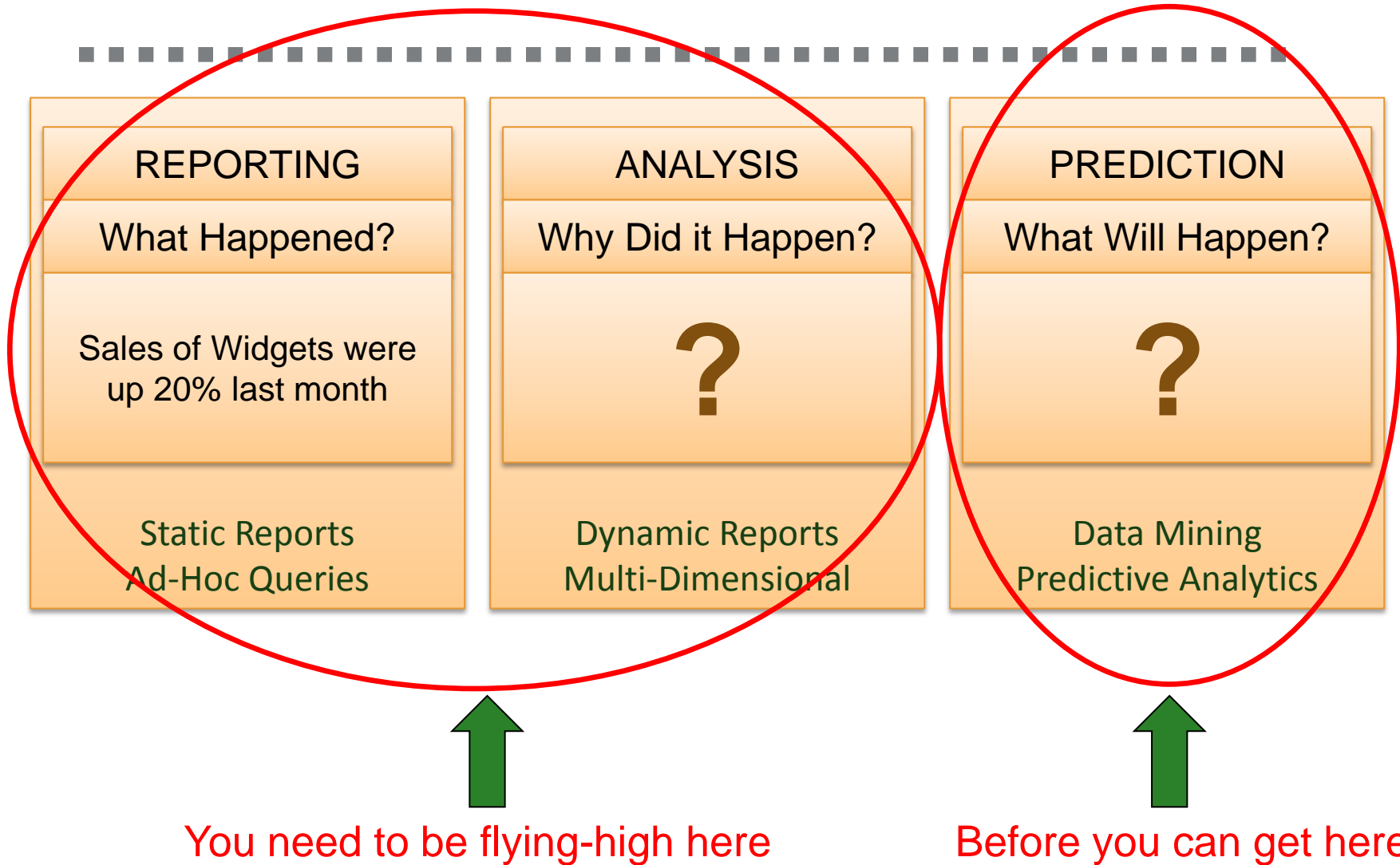
- The next step beyond data mining.
- Applying data mining in conjunction with machine learning and even artificial intelligence to make predictions about what will happen in the future, based on patterns in available historical data.
- Credit scoring
- Earthquake prediction
- Weather forecasting
- Fraud detection



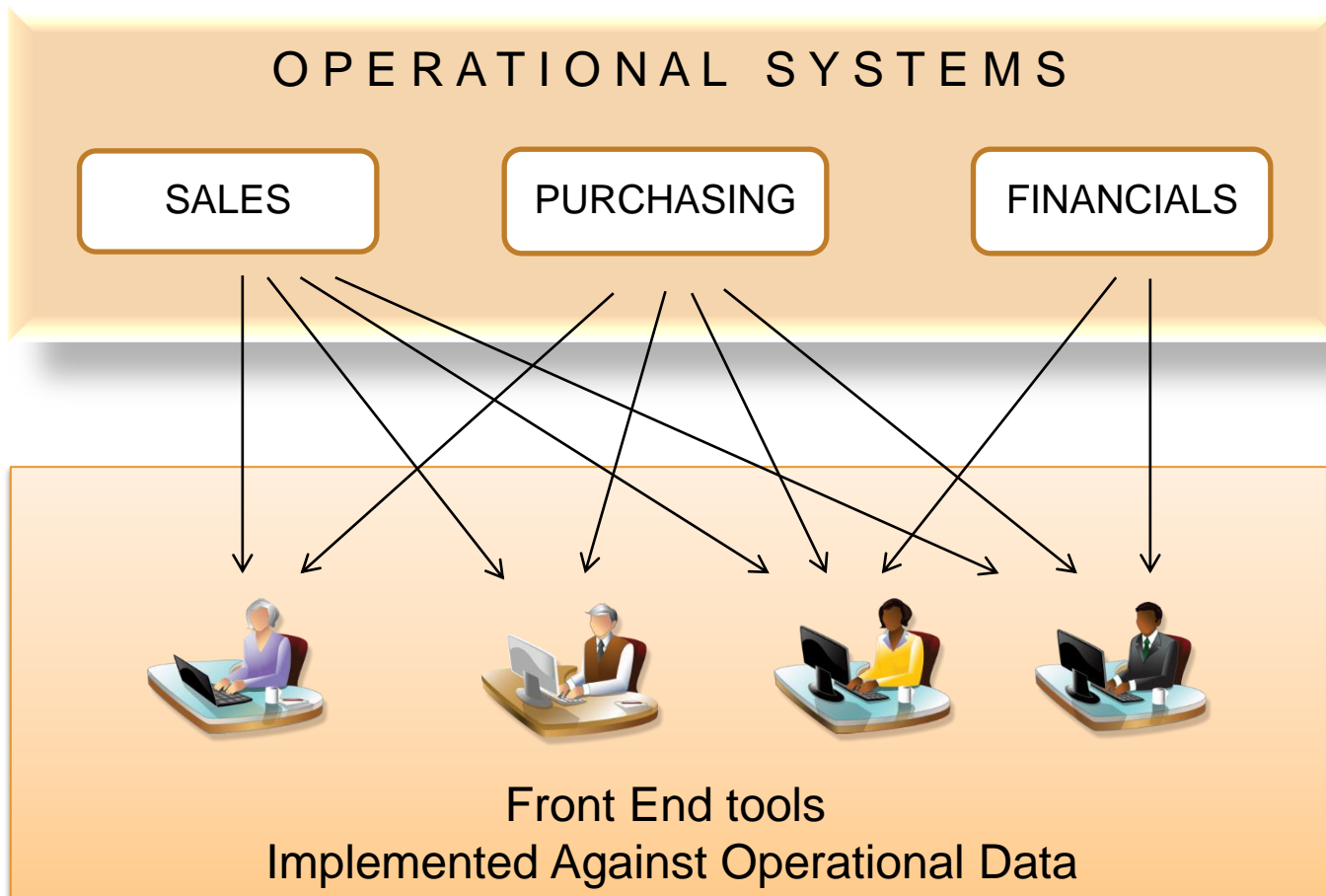
Business Intelligence



Business Intelligence



Simple Implementation



Simple Implementation

-
- **Many Small to Medium size organizations begin with this approach**
 - Single toolset needed
 - Low cost
 - **It works well with in many cases**
 - **But it can have its challenges**

Issues with Operational Data

Challenges

Operational data can be complex and difficult to understand

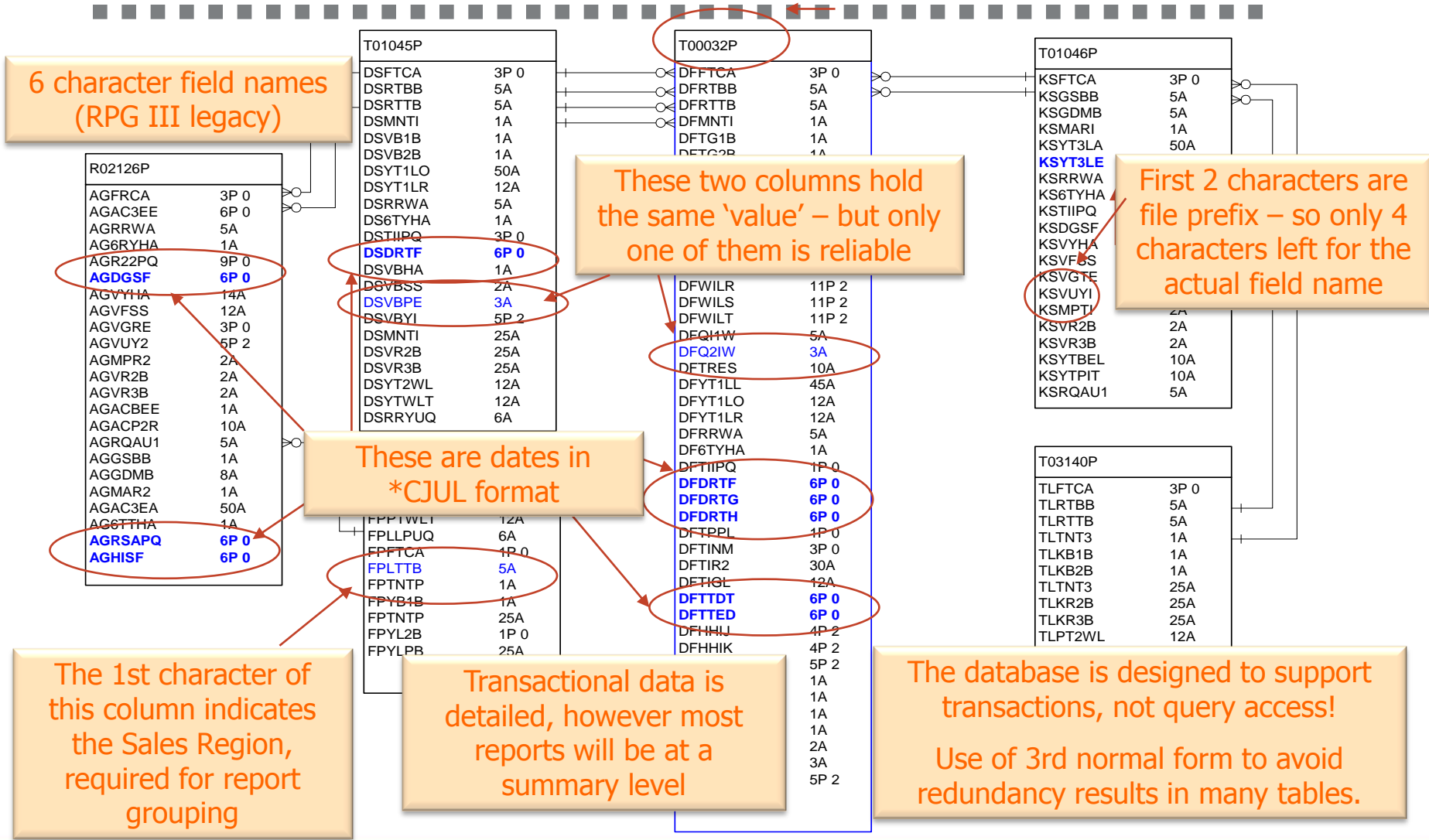
- Many tables in a transaction 'schema'
- Cryptic table and column names
- Often need to calculate or derive values
- Dates are numeric values

Causes

The database was designed for the application to access – not you

- Principle of 3rd normal form
- Legacy of earlier restrictions (6 character field names, no true date support)
- Inconsistencies: different developers, merger of applications

Operational Data Example



Issues with Operational Data

Challenge

- Questionable or unknown data quality
- The data may be correct – but you don't understand it correctly

Causes

- Bugs in the application
- Inconsistent data entry
- Incorrect data conversions

Issues with Operational Data

Data Quality Example

2005: Valparaiso, Indiana

Somehow a property assessment value for this home was incorrectly changed to **\$400M** in the property tax database.



The expected **\$8M** property tax revenue was included in the county budget, but the homeowner (of course) did not pay the bill.

The county had a huge revenue shortfall.

The school district was forced to return **\$2.7M**.

All extracurricular activities and sports were cancelled that year.

Just because of 1 bad data value!

Issues with Operational Data

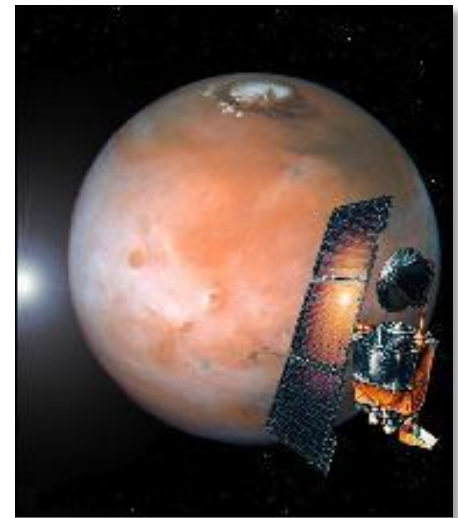
Sometimes it is not the data itself, but our understanding of it

The 1999 NASA Mars Climate Observer mission failed because of a data interpretation problem.

Thrust calculation data was provided in the US scale of pounds/square foot, but was interpreted as metric numbers representing newtons/second.

This resulted in the wrong amount of thrust being used to slow it down, resulting in failure to go into orbit. It probably crashed on Mars.

A \$300M mission failed because of a simple mistake!



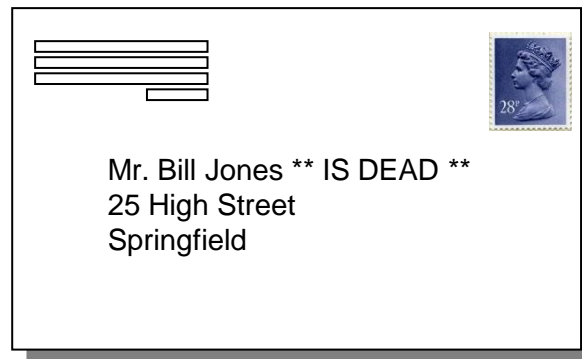
Issues with Operational Data

Sometimes, it is how we use it...

A travel agency used telemarketing to sell vacations to its past customers

On occasion, it happened that a customer had passed away. Their system would not let them delete the customer, since there were transaction records tied to it. Someone came up with the idea of appending the customer name with "*** IS DEAD **", so operators would not call and upset the family of the deceased.

This worked fine until the company switched to direct mail. Imagine the grief caused to Mrs. Jones when she received this letter



*This really DID
happen!
(in the UK)*

Cost of Poor Data Quality

A 2014 Report from Artemis Ventures indicates that poor data quality costs US businesses

\$3.1 Trillion per year!

An estimate from the US Insurance Data Management Association puts the cost of poor data quality at

15% to 20%

of operating revenue

Issues with Operational Data



Challenge

Different applications/databases/platforms

- Totally different structures – but related information



Very difficult, if not impossible to join tables across databases, different security, availability, etc.

Issues with Operational Data

Challenge

Multiple instances of same table, with duplicate key values

Customer File - US	
CUSTNO	CUSTNAME
1001	John Smith
1002	Mary Jones
1003	Chris Anderson
1004	David Perry

Customer File - Canada	
CUSTNO	CUSTNAME
1001	Harry Potter
1002	Jeremy Carr
1003	Penny Hayes
1004	Debbie Thornton

or different versions of same entity

- Incompatible data types
- Duplicates

Customer File - US	
CUSTNO	CUSTNAME
1001	John Smith
1002	Mary Jones
1003	Chris Anderson
1004	David Perry

Customer File - Canada	
CUSTID	CUSTNAM
AA234	Julie Johnson
AA235	Fred Hunter
AB670	John Smith
BD309	Alan Jordan

Issues with Operational Data

Challenge

Changing attributes

2011	100	Acme Flooring	Small Retailer	Jenny Brown
2013	100	Acme Flooring	Major Retailer	Jenny Brown
2014	100	Acme Flooring	Major Retailer	Rob McAdam

2011 Report

2011 Sales by Sales Rep/Customer Group

Acme Flooring	250,000
Regal Rugs	150,000
Total Small Retailer	400,000
Carpet Warehouse	2,500,000
Hardwood Hank	2,100,000
Total Major Retailer	4,600,000
Total Jenny Brown	5,000,000

Same report, re-run in 2014

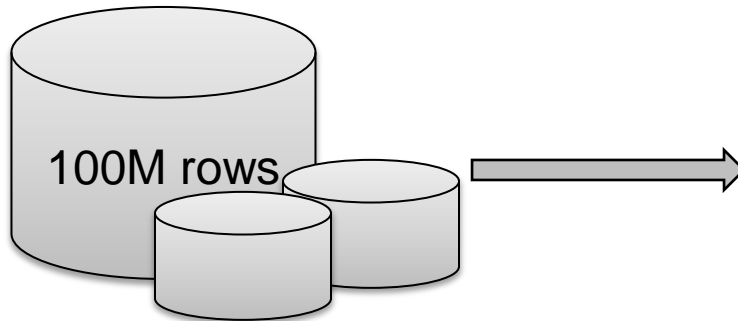
2011 Sales by Sales Rep/Customer Group

Regal Rugs	150,000
Total Small Retailer	150,000
Carpet Warehouse	2,500,000
Hardwood Hank	2,100,000
Total Major Retailer	4,600,000
Total Jenny Brown	4,750,000

Issues with Operational Data

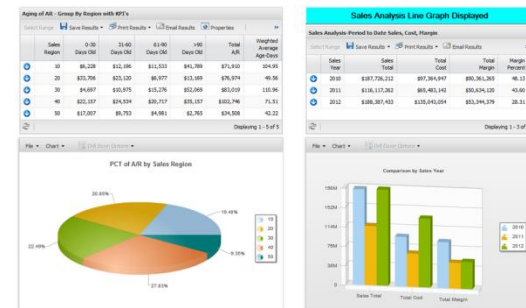
Poor Performance

- Large transaction table
- Many related tables
- Most reports are at a summary level
- Reports and queries are long running and consume significant system resources



- Shipping Analysis
- IT's Hyperlinks Doc
- Sales Client Report
- Customer Detail Rpt
- SEQUEL Website
- Web Calculator

SALES Dashboard KPI - Key Performance Dashboard



Issues with Operational Data

Inconsistent Results

- Maintenance changes during the day can be a problem



You are performing analysis at the Customer Group level, happily slicing and dicing away at the data.

Suddenly, the numbers are all out of whack.

What happened?

Someone performed customer maintenance and changed the Group for one or more customers.

But you don't know that!



The Wrong Solution

These issues are often solved in an ad-hoc way

Create “extract files” and write RPG programs to load them

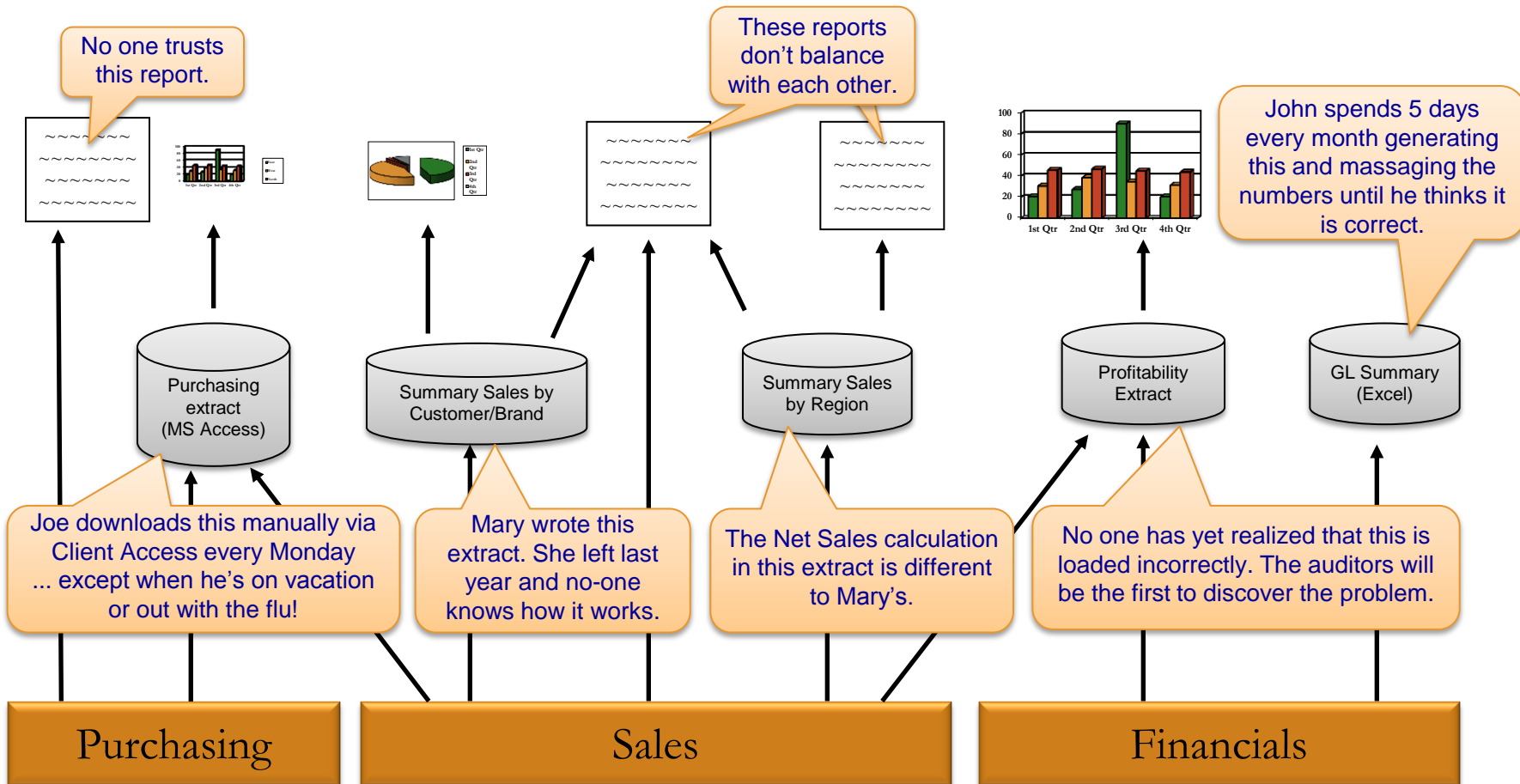
- As each reporting problem occurs, a new extract is written
- No consistent approach
- No documentation produced

Frustrated users create their own “solutions”

- Download data to excel and manipulate it
- Decide on their own rules

The Wrong Solution

The result can be a **Chaotic Reporting Environment!**



Audience Poll

1. Do you own a BI front end tool?
2. Does the previous chart look familiar? Is this your organization?
3. What are the biggest issues you face in effective, reliable BI reporting?
4. Would you say you have a formal BI strategy?
5. Do you have a Data Warehouse?

Simple Implementation

.....

If the simple implementation is not working

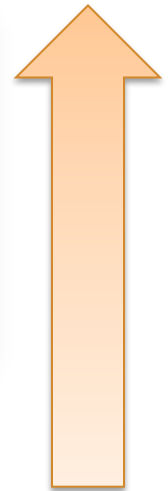
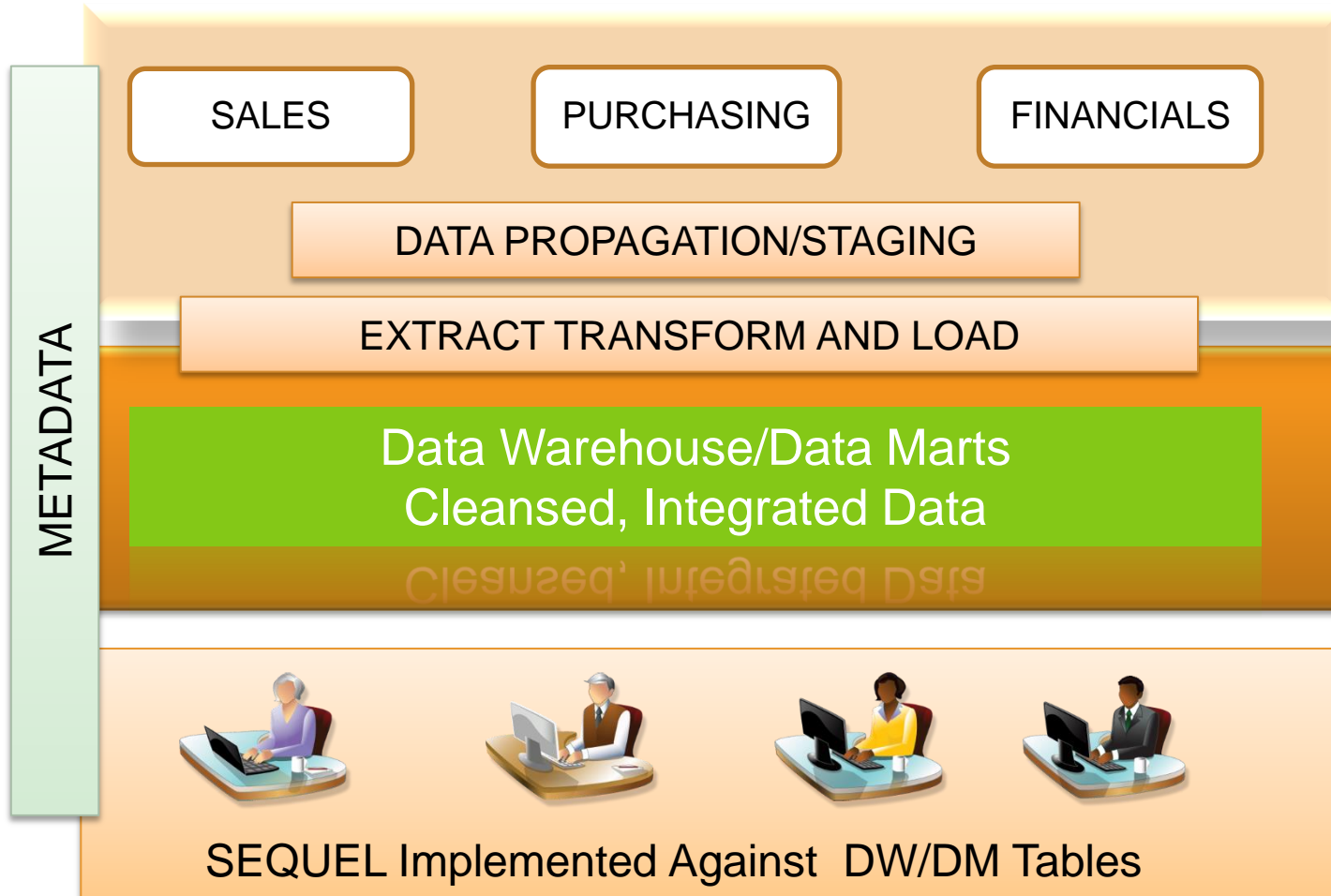
DON'T BLAME THE TOOL

BLAME THE DATA!

... and implement a data warehouse

Data Warehouse Architecture

OPERATIONAL SYSTEMS



Issues and complexity pushed to the back-end

Definitions

Data Warehouse

- A centralized repository of mostly historical information, built from operational data sources
- Usually contains several different subject areas
- A single version of the truth
- Always in open database tables
- Always detailed level information
 - To allow creation of new data marts, or re-creation of existing ones
- Rarely queried directly by users
 - Everyone but *power users* will usually access the data marts

Definitions

Data Mart

- Built from the data warehouse to support a specific business reporting requirement
- Often summarized, but may be detailed
- Updated (or re-built) on a regular basis from the data warehouse
- May be in a proprietary format –
 - i.e. multi-dimensional structures (cubes)
- If in database tables, often a *star schema* structure
- A key element of **Dimensional Modeling**

Definitions

Operational Data Store

- A reporting database containing the 'current' view of the operations of the business
- Contains little or no historical data
- Contains incomplete or in-progress entities (e.g. sales orders not yet fulfilled)
- Usually completely re-built on a regular (usually daily) basis

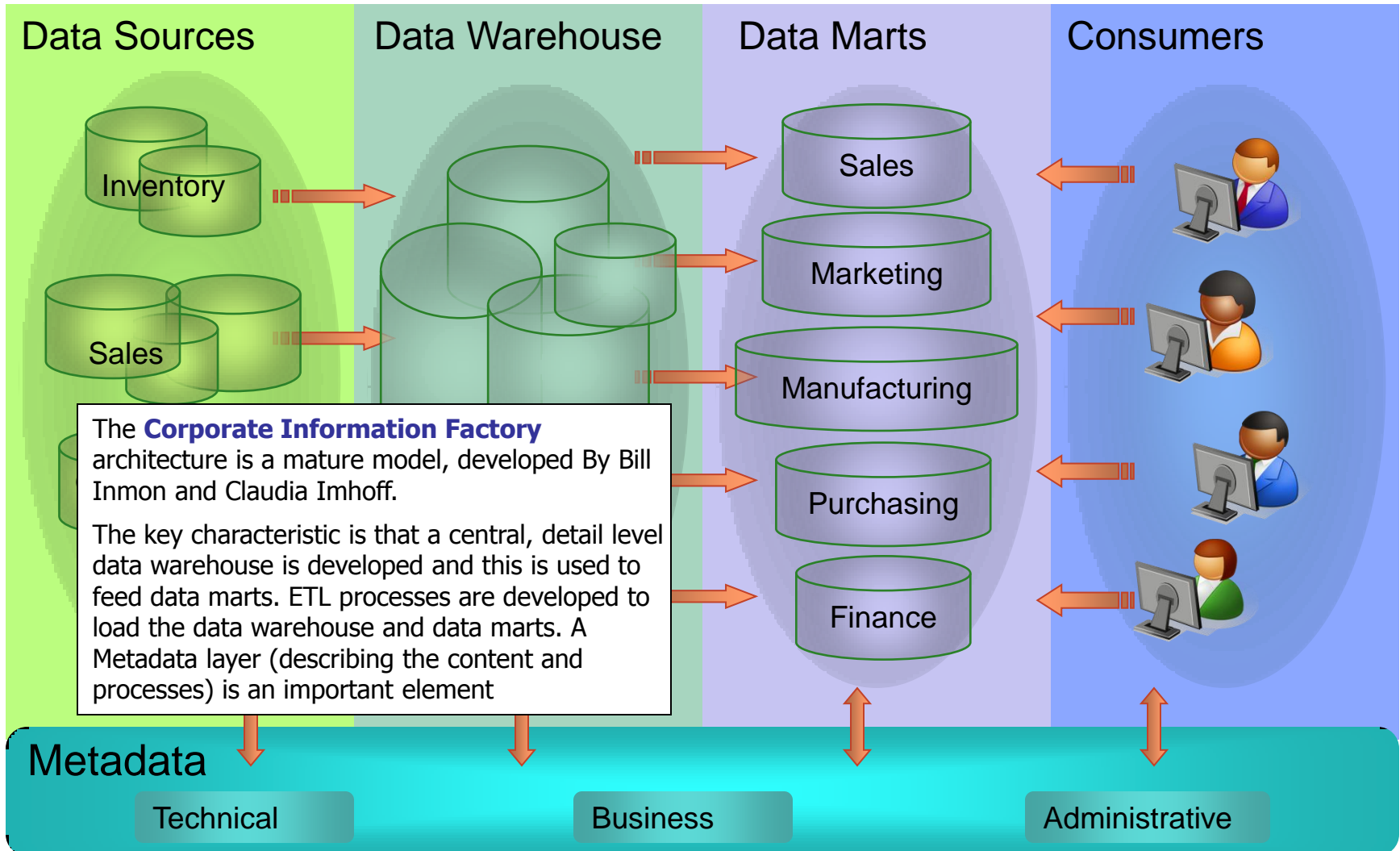
Definitions

Metadata

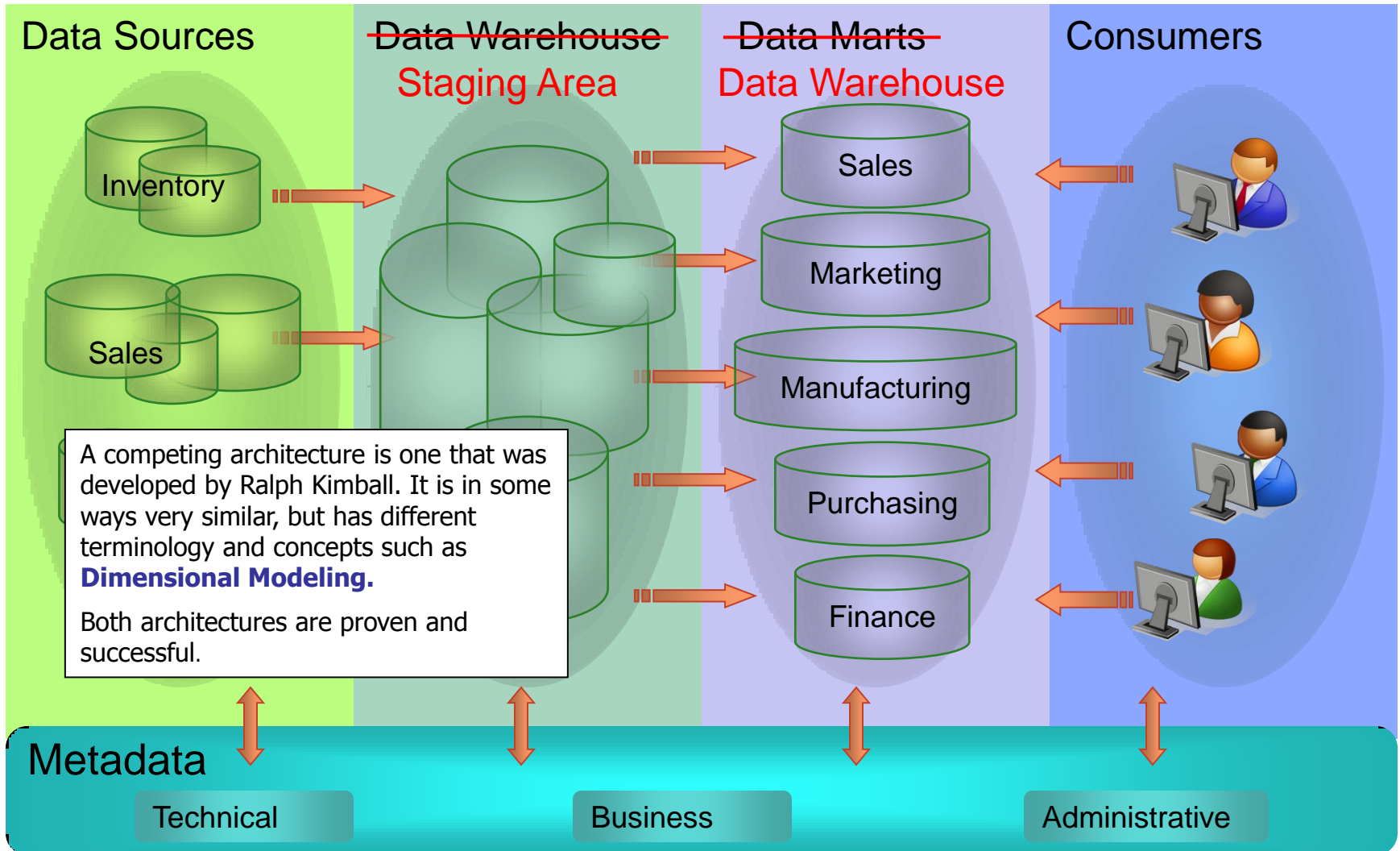
- “Data that describes data”
- Technical metadata
 - Table and column names, length, data type, decimals
- Business metadata
 - Validation rules, transformation rules, source/target relationships
- Administrative metadata
 - Users, authorities, size, usage, performance and data quality statistics, change history



3-Tier Architecture



3-Tier Architecture

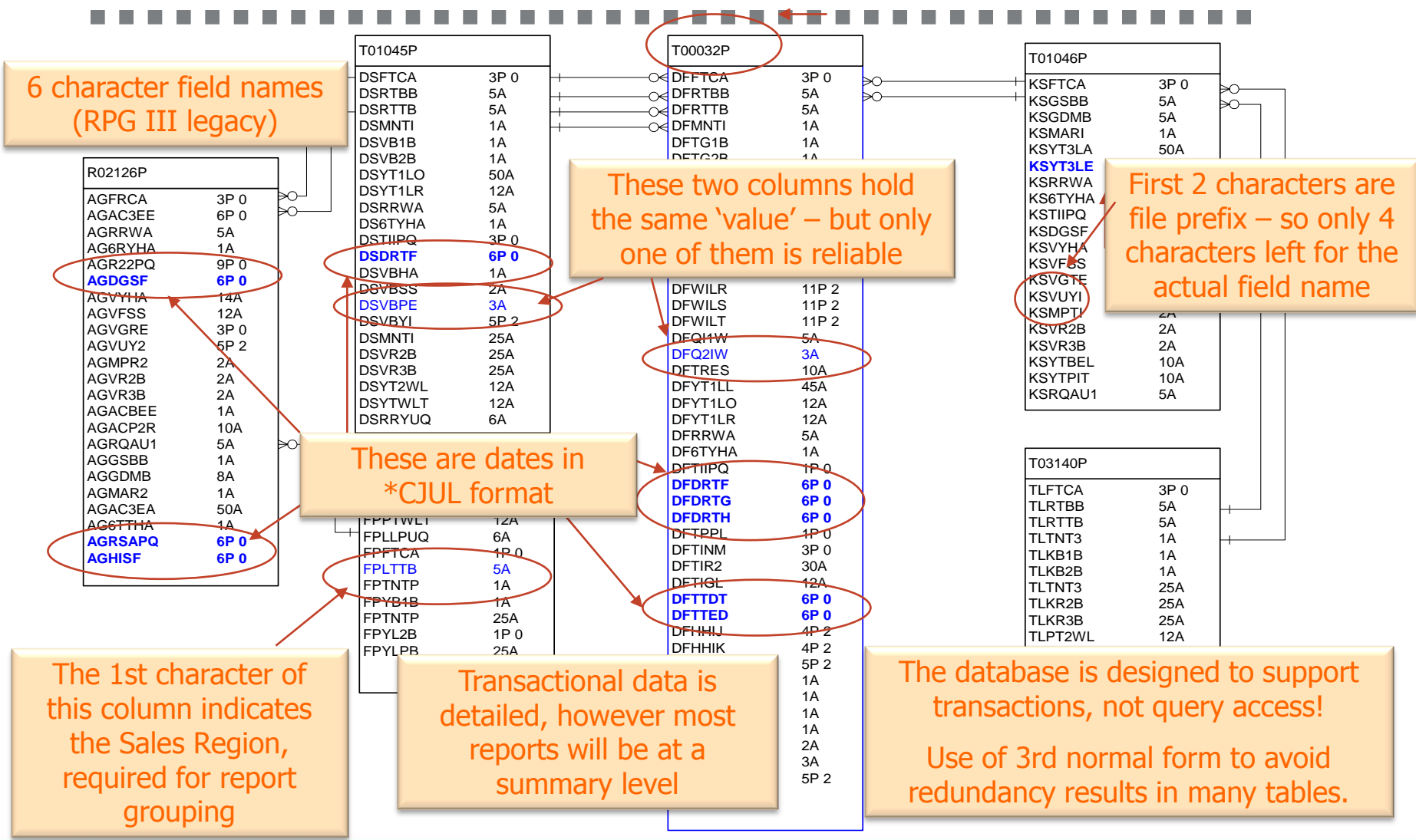


Why a Data Warehouse

Issues Review:

Operational data can be complex and difficult to understand

Operational Data Example



Data Mart Example

De-normalized design reduced to only a few tables

PRODUCTS	
PRODUCT_NUMBER	5P 0
PRODUCT_DESCRIPTION	42A
BRAND_CODE	5A
BRAND_DESCRIPTION	20A
ORIGIN_CODE	5A
ORIGIN_DESCRIPTION	20A
FAMILY_CODE	5A
FAMILY_DESCRIPTION	20A
COST	9P 2
BASE_PRICE	9P 2
PRODUCT_WEIGHT	9P 4
PRODUCT_VOLUME	9P 4
LOAD_DATE	DATE
LAST_CHANGE_TIME	TSTP
STATUS_FLAG	1A

Only includes the columns we care about

INVOICE_LINES	
INVOICE_NUMBER	7P 0
INVOICE_LINE_NUMBER	3P 0
PRODUCT_NUMBER	5P 0
CUSTOMER_NUMBER	10A
SELLING_COMPANY	5A
SUPPLY_WAREHOUSE	5A
QUANTITY_ORDERED	11P 0
QUANTITY_SHIPPED	11P 0
TOTAL_DISCOUNT	9P 2
NET_PRICE	9P 2
BASE_PRICE	9P 2
UNIT_COST	9P 2
EXTENDED_COST	11P 2
EXTENDED_PRICE	11P 2
MARGIN	11P 2
SALES_REP	5A
COMMISSION_VALUE	7P 2
INVOICE_DATE	DATE
SHIP_DATE	DATE
DELIVERY_DATE	DATE
INVOICE_TIME	TIME
MONTH_NUMBER	2P 0
WEEK_NUMBER	2P 0
LOAD_DATE	(DATE)

Meaningful table and column names

Complex calculations already done

Dates are true date columns

CUSTOMERS	
CUSTOMER_NUMBER	10A
CUSTOMER_NAME	35A
ADDRESS_LINE_1	35A
ADDRESS_LINE_2	35A
CITY	35A
STATE_CODE	2A
ZIP_CODE	10A
PHONE_NUMBER	35A
PHONE_EXTENSION	15A
ACCOUNT_DEFAULT	5A
ACCOUNT_CATEGORY	5A
CUSTOMER_CLASS	5A
REGION_CODE	5A
LOAD_DATE	DATE
LAST_CHANGE_TIME	TMSTP
STATUS_FLAG	1A

Why a Data Warehouse

Issues Review:

Questionable or unknown data quality

Perform validation and error management in the load of the data warehouse

- Build data quality rules
- Set aside and report on bad data

Why a Data Warehouse

Issues Review:

Incorrect use due to lack of understanding

Data Warehouse team provides information in the form of metadata

- Available to report authors and other consumers of the data
- Part of a data governance initiative

Why a Data Warehouse

Issues Review:

Different applications/databases/platforms

The disparate data is transformed and conformed in the data warehouse

- Report authors don't need to deal with different databases and applications
- Reports that were difficult or impossible before are now routine

Why a Data Warehouse

Issues Review:

Poor performance

Data Marts are created and loaded at the ideal summary level for various reports

- No need to aggregate millions of rows of data for a dashboard or report
- Increased productivity
- Reduced load on system

Why a Data Warehouse

Issues Review:

Multiple versions of the truth

All calculations, transformations and aggregations are performed in a standard way based on the same conformed, cleansed (validated) data

- Reports now agree with each other
- More confidence in their accuracy

Data Warehouse Technologies

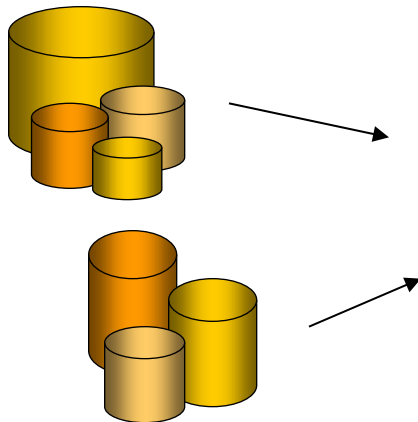
Extract, Transform & Load (ETL)

E.T.L.



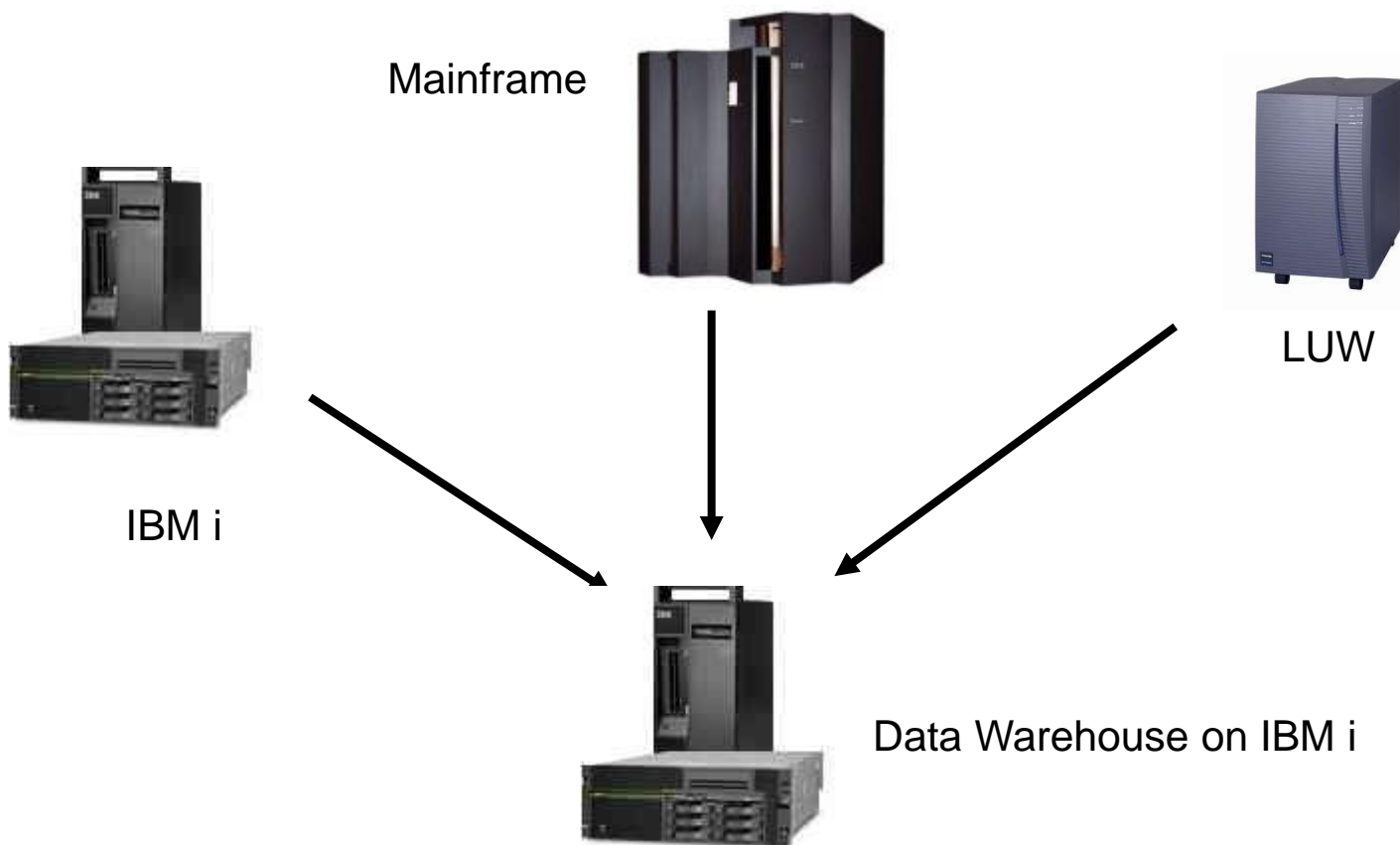
Extract Data from Sources

- Database tables (IBM i)
- Remote databases (e.g. DB2, MS SQL Server, Oracle)
- Text/delimited files
- Change Data Capture from journal images



Accessing Remote Data

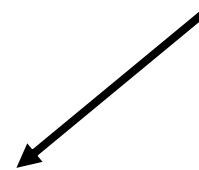
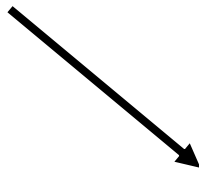
RDBMS Data Sources



Accessing Remote Data

.....

Non database sources can be more of a challenge



Data Warehouse on IBM i

Accessing Remote Data

In all cases, it is recommended to STAGE remote data prior to ETL

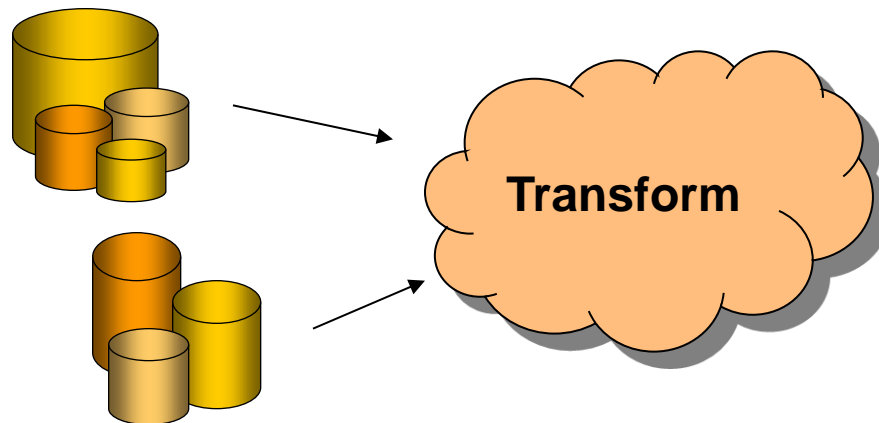
- If not in DB2 format, convert it to DB2 at this point
- The staged data becomes your local copy
- Don't correct or enhance it
- You now own it and can re-use it as needed
- Can now be easily joined to other data in the ETL process
- Simplifies the actual ETL process – fewer points of failure
- Allows for re-run if an ETL step fails

E.T.L.



Transform the Extracted Data

- Arithmetic calculations
- String operations
- Lookup/replace
- Date/time conversions and calculations



Transformations

Examples

- Convert a legacy date in the format *cyymmdd* into a true date
 - Need to also manage errors and exceptions
 - date value of zero or all 9's may be handled as special cases
 - but a value of 1140230 (February 30th) is an error!

Transformations

Examples

- Convert meaningless codes and values
 - e.g. Gender Code:

Source Value	Replace with	Or
'1'	'M'	'MALE'
'2'	'F'	'FEMALE'
' '	'U'	'UNKNOWN'

Transformations

Examples

- Create values/attributes from complex relationships:

Derived Attribute ***SALE TYPE***

When CUSTYP = '11' and TRFYD <> Blank
= 'INTERNAL'

When CUSTYP = '08' or '09' and TRFYD = 'TR'
= 'TRANSFER'

Otherwise
= 'NORMAL'

Transformations

Examples

- Standardize formatting:
 - Format all telephone numbers using a mask
 - Remove commas etc from address lines
- Scan & Replace
 - Change Mens Polo Shirt Sz 12, Wht
 - To Mens Polo Shirt Size 12 White
- Justification
 - For example, many codes in JDE are right justified!

Transformations

A recent customer example

I have a contact name field that has the person's name in it and I need to parse them out into separate first and last name columns.

Some of the names have a middle initial and others are just first name last name.

e.g.

John Smith

Susan B DeMille

E.T.L.

Load the Transformed Data

- Into one or more target tables
- Detail or summary level
- Insert or update



Loading the transformed data

Examples

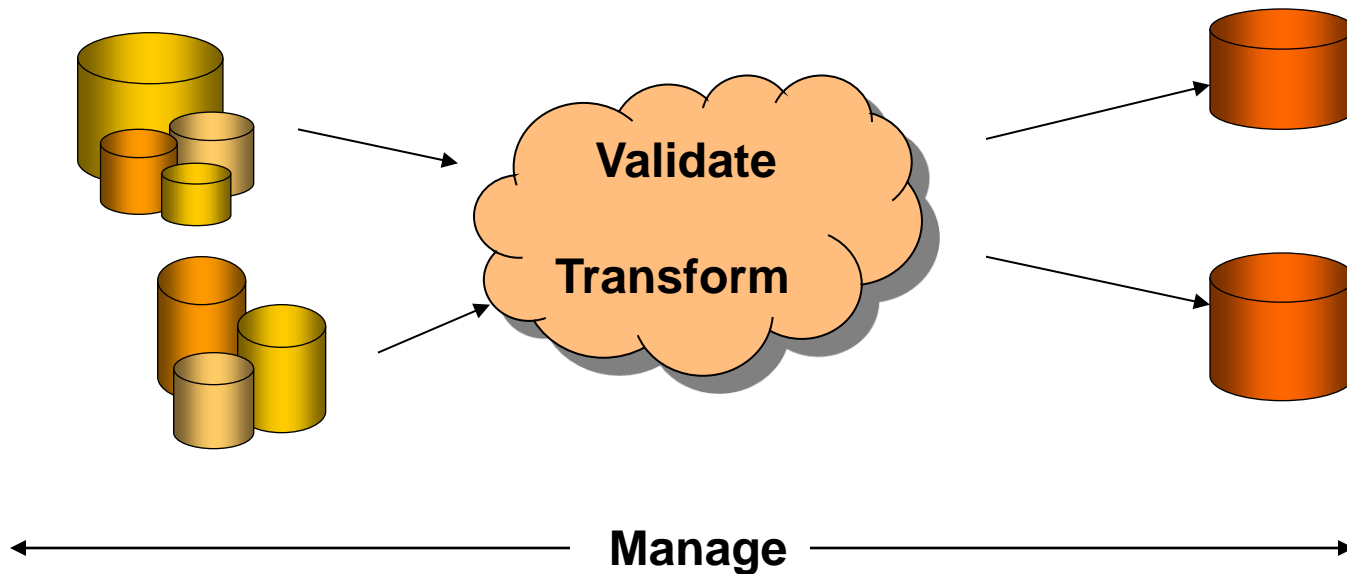
- Set **LOAD_DATE** on insert, and never update it
- Set **CHANGE_DATE** on insert and update
- On update, accumulate the **TOTAL_SALES** value
- On update, replace **LAST_INV_AMOUNT**
- Only update **HIGHEST_ACCT_BAL** if it is a new maximum value
- **Load DEBITS into table A and CREDITS into table B**

E.T.L.



There are Two VITAL Additional Requirements

- Validate – define business rules
- Manage – data errors
- the overall environment



Extract, Transform and Load (ETL)

Other Requirements of ETL

- Provide real-time load option
- Allow for re-run if ETL fails
- Provide audit trails
- Provide comprehensive error management and reporting
- Provide metadata support
- Manage changes to data sources
- Provide security layer (only allow authorized users)
- Provide excellent performance

Data Warehouse Technologies

Change Data Capture

Change Data Capture

Change Data Capture

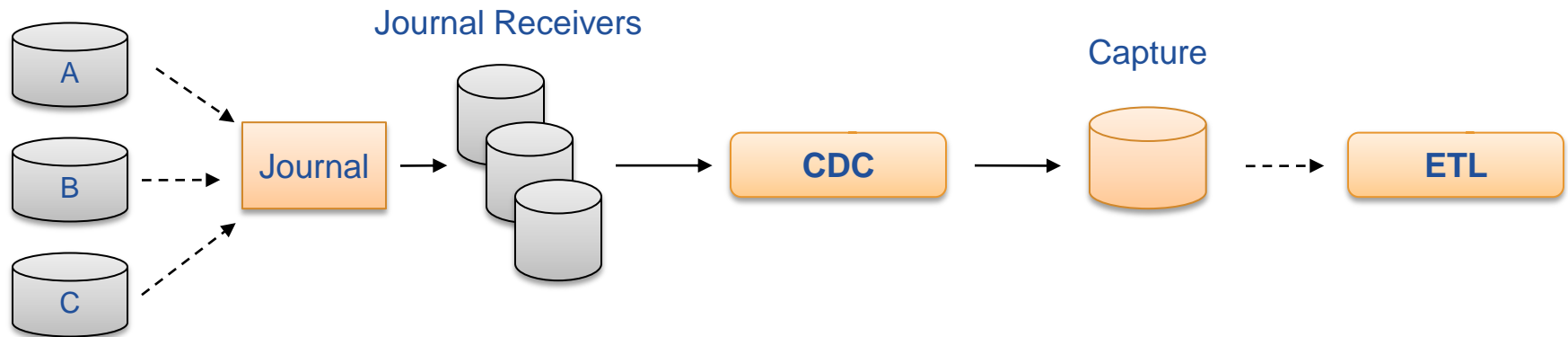
The process of selecting new or changed data based on journal entries

Non-intrusive on source systems

The 'output' of CDC is the input to the ETL process

Only useful/recommended in certain situations

Application Tables



When to use CDC ?

CDC is not recommended for use when

- ✗ Loading Transaction data that has a reliable date or timestamp.
- ✗ Master tables that have a reliable change date or timestamp
- ✗ Small tables that take seconds or minutes to fully load/replace
- ✗ The source data is not in DB2 for i tables
- ✗ The source data is in a DB2 for i table that is not journaled

CDC may be a good option when

- ✓ There is no reliable date or timestamp to select the required data
- ✓ Re-loading all data on a regular basis would be take a lot of time
- ✓ An audit trail of all changes to a row in a table is required
- ✓ Real-time load is required

Change Data Capture

Real-Time Load

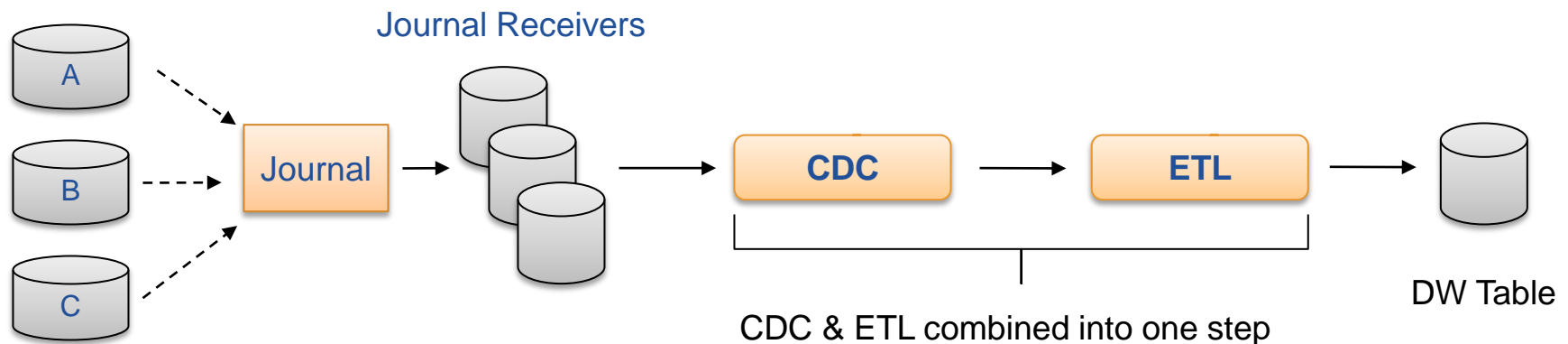
First of all, why would you want to do this?

- Do you really care what happened 5 minutes ago?
- Constantly changing data can really mess with your analytics

How can you achieve it?

- Hard coding – not a sensible option
- Triggers – can have major impact on performance
- CDC - is best option

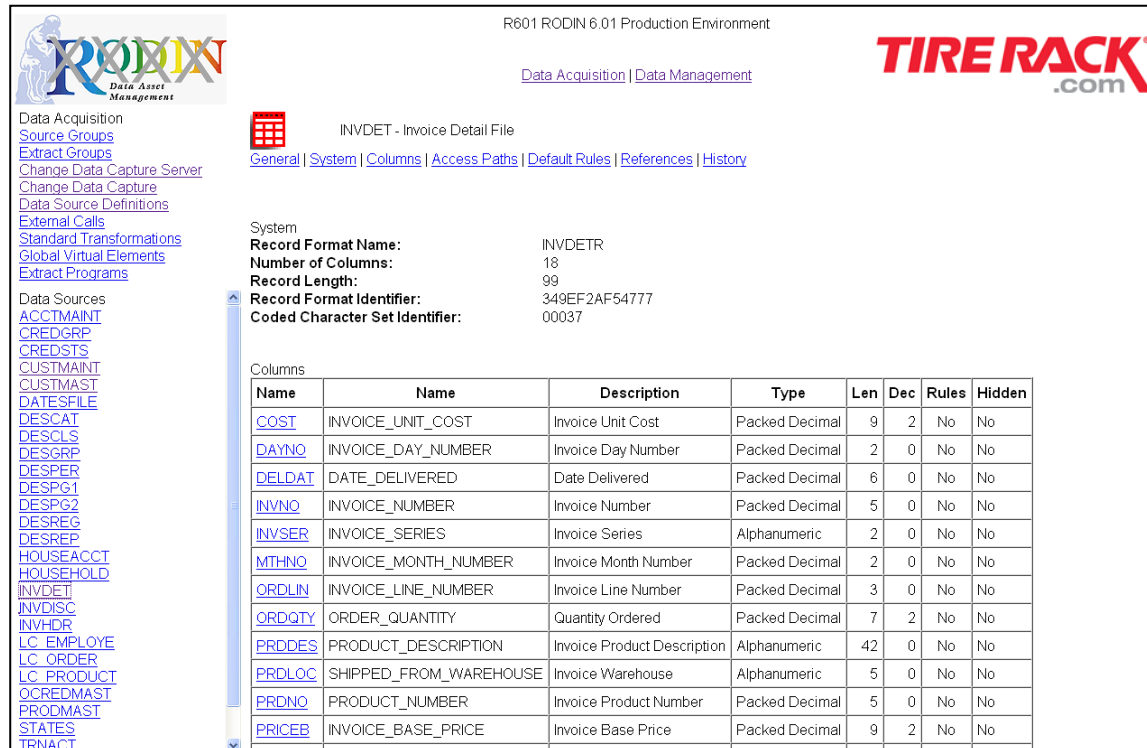
Application Tables



Delivering Metadata

Metadata should be made available to everyone

- Not tool dependent
- Not printed
- Browser is the ideal interface



R601 RODIN 6.01 Production Environment

RODIN Data Asset Management

TIRE RACK.com

Data Acquisition | Data Management

INVDET - Invoice Detail File

General | System | Columns | Access Paths | Default Rules | References | History

System
Record Format Name: INVDETR
Number of Columns: 18
Record Length: 99
Record Format Identifier: 349EF2AF54777
Coded Character Set Identifier: 00037

Columns

Name	Name	Description	Type	Len	Dec	Rules	Hidden
COST	INVOICE_UNIT_COST	Invoice Unit Cost	Packed Decimal	9	2	No	No
DAYNO	INVOICE_DAY_NUMBER	Invoice Day Number	Packed Decimal	2	0	No	No
DELDAT	DATE_DELIVERED	Date Delivered	Packed Decimal	6	0	No	No
INVNO	INVOICE_NUMBER	Invoice Number	Packed Decimal	5	0	No	No
INVSER	INVOICE_SERIES	Invoice Series	Alphanumeric	2	0	No	No
MTHNO	INVOICE_MONTH_NUMBER	Invoice Month Number	Packed Decimal	2	0	No	No
ORLDLN	INVOICE_LINE_NUMBER	Invoice Line Number	Packed Decimal	3	0	No	No
ORDQTY	ORDER_QUANTITY	Quantity Ordered	Packed Decimal	7	2	No	No
PRDDES	PRODUCT_DESCRIPTION	Invoice Product Description	Alphanumeric	42	0	No	No
PRDLOC	SHIPPED_FROM_WAREHOUSE	Invoice Warehouse	Alphanumeric	5	0	No	No
PRDNO	PRODUCT_NUMBER	Invoice Product Number	Packed Decimal	5	0	No	No
PRICEB	INVOICE_BASE_PRICE	Invoice Base Price	Packed Decimal	9	2	No	No

Building a Data Warehouse



Building a Data Warehouse

Recognize that it is a journey, not a destination

It will evolve, grow and change over time, responding to your changing business requirements

Think of the larger picture, but build in small steps

Don't try to complete everything you want in one project

Focus on critical needs first

Get value as early as possible

Involve end-users

But identify their real needs

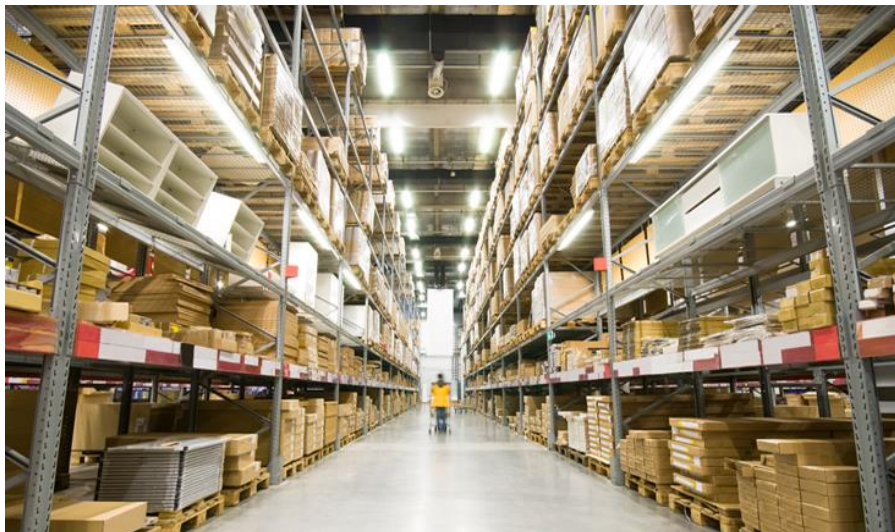
- As Henry Ford said "If I'd asked people what they wanted, they would have said faster horses"

Building a Data Warehouse

Identify the Data Items that are Required

- These will become the columns in the DW and DM tables

Design & Create the Tables



Building a Data Warehouse

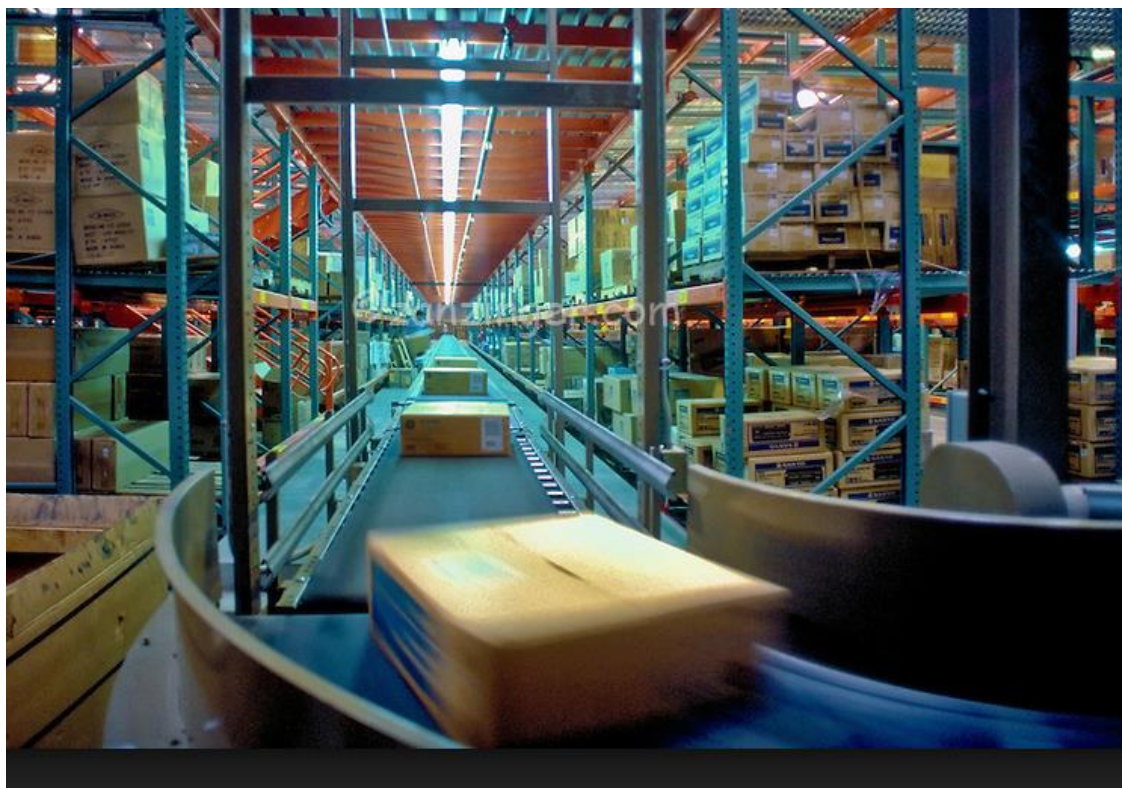
Develop the Load Processes



Building a Data Warehouse

.....

Above all, consider how to efficiently get data out!



Building a Data Warehouse

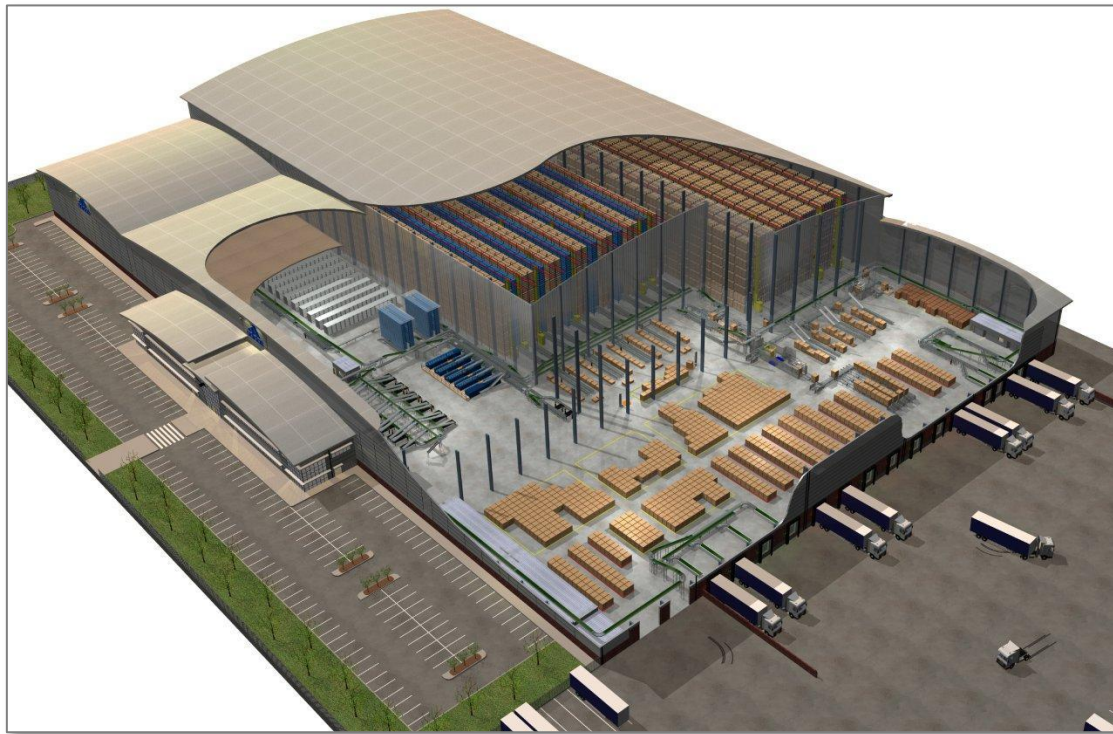
Document the data warehouse using metadata

- Essential to success
- It is the roadmap to what data is available, where to find it, and understanding what form it is in.



Building a Data Warehouse

The overall design needs to consider all these needs



Building a Data Warehouse

Can you do all of this by hand-coding?

- You would not even consider writing your own query and reporting tool



- Why 'roll your own' when it comes to a Data Warehouse, ETL and Metadata?

Building a Data Warehouse

Can you do all of this by hand-coding?

- Many organizations completely under-estimate the scope of work in performing the ETL. It is often 50% or more of the total effort, yet it is often allocated just a small fraction of the overall project budget.
- By the time this error is discovered and the true ETL effort is recognized, the project can be in serious trouble.
- It is then very difficult to request additional funds or resources.
- You end up cutting corners.
 - You deliver ETL processes are inadequate and provide little or no data quality management.
 - Metadata is non-existent.
- It becomes a nightmare to maintain.

Questions & Answers



Stay tuned for Part 2!

Dimensional Modeling