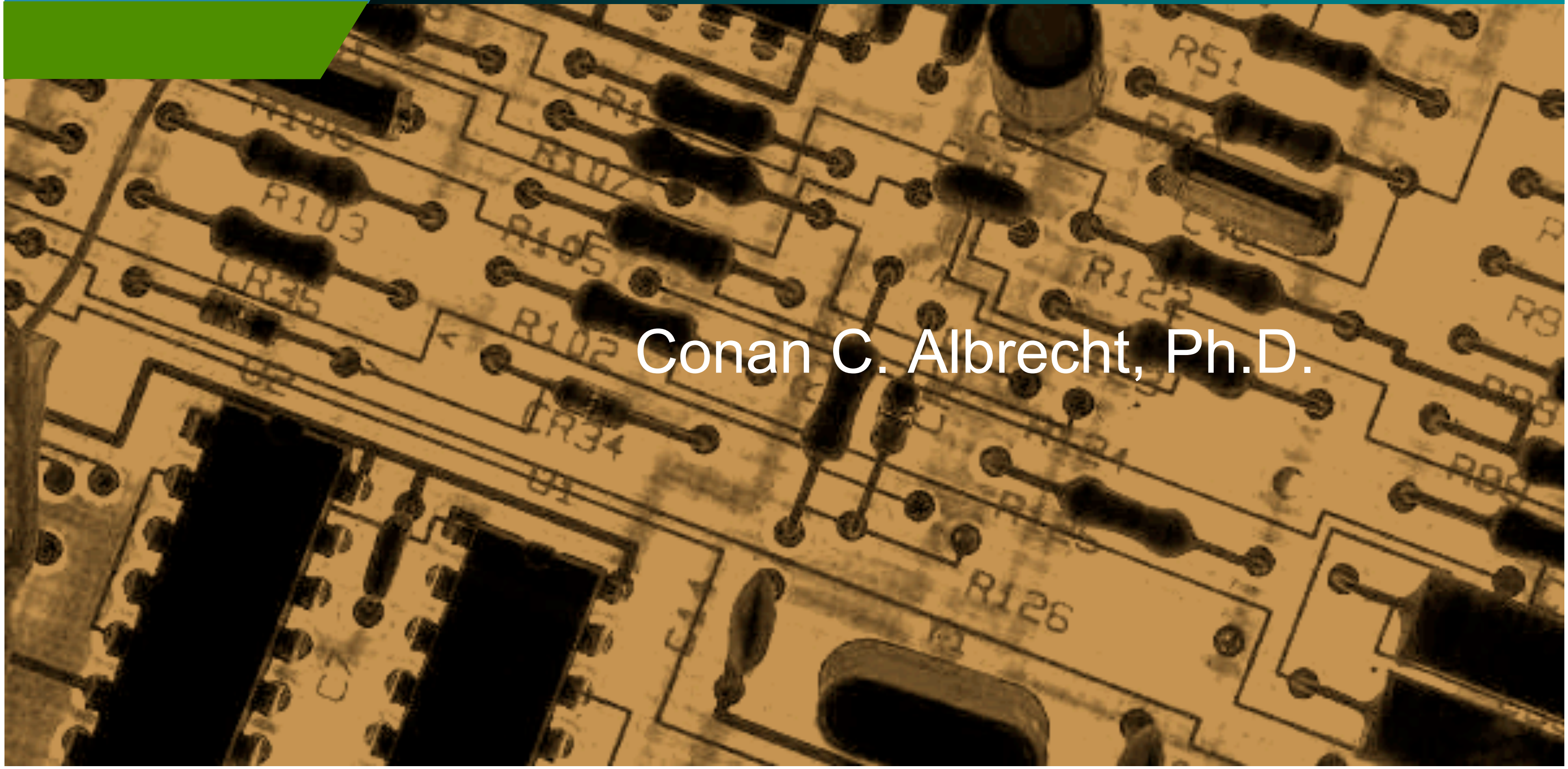
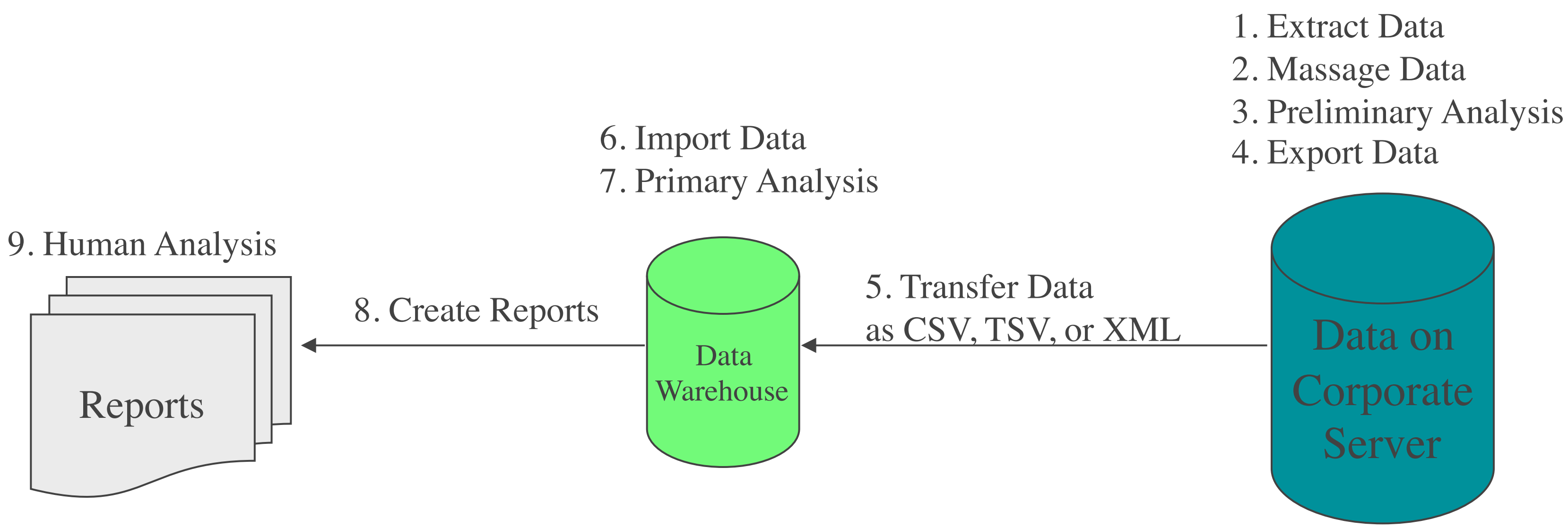
Technology Foundations

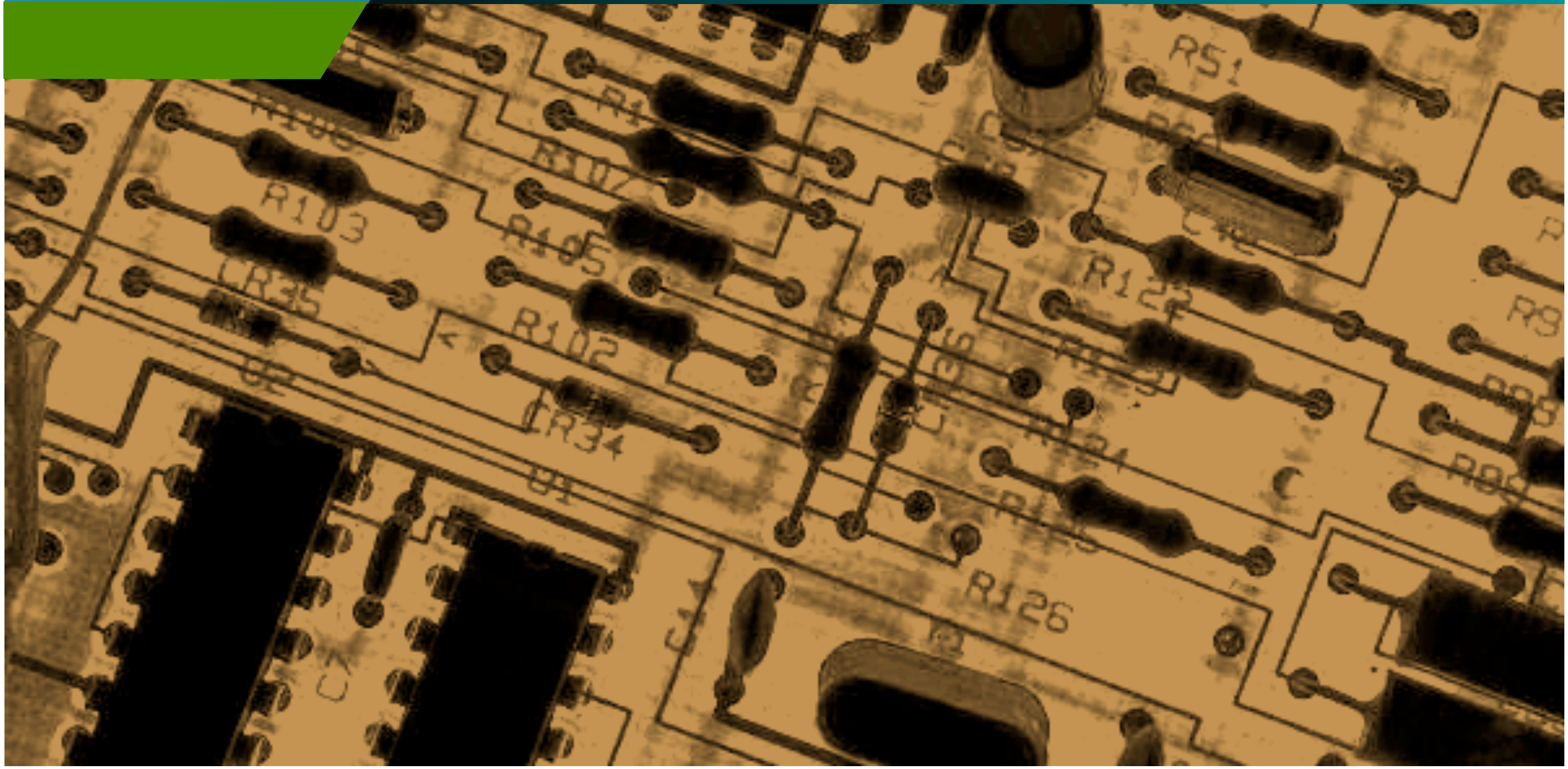


Conan C. Albrecht, Ph.D.

Overview


1. Extract Data
2. Massage Data
3. Preliminary Analysis
4. Export Data
6. Import Data
7. Primary Analysis
9. Human Analysis
5. Transfer Data
as CSV, TSV, or XML
8. Create Reports
Data
Warehouse
Data on
Corporate
Server
Reports

Background Principles



Sampling

- Why do we sample?
- What is the end goal of sampling?

Risks of Sampling

- Why do we sample?
  - Efficiency: we can’t review all records
- What’s the end goal of sampling?
  - To extrapolate to a population
- Computers don’t need to sample
- Fraud detection is *not* about extrapolation to the entire population
  - We’re only interested in the 2-3 bad records!
- Rather than sample, create scripts to do your analyses on the entire population
  - Some sampling can be done to check the script, but not to check the data

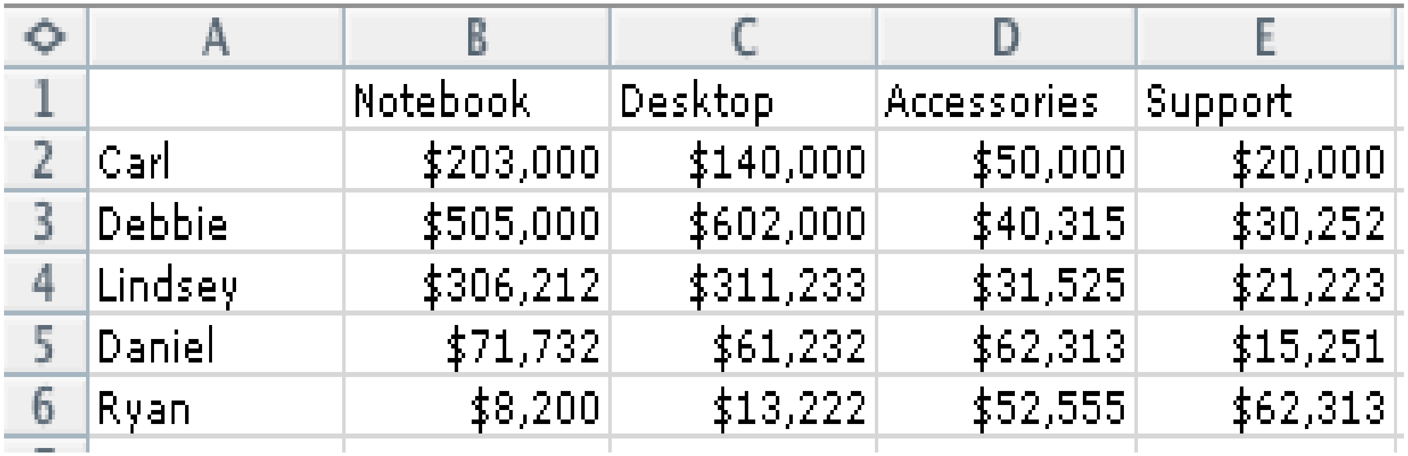
Why a database primer?

- Most corporate data is stored in large databases
  - Oracle, DB2, MS-SQL Server, MySQL
- But that’s what geeks are for!?!
  - A basic knowledge empowers you to guide and direct IT personnel
  - Can you imagine doing fraud examination without basic (or even advanced) accounting knowledge?
  - Future CFEs will need to know more and more “geek stuff”

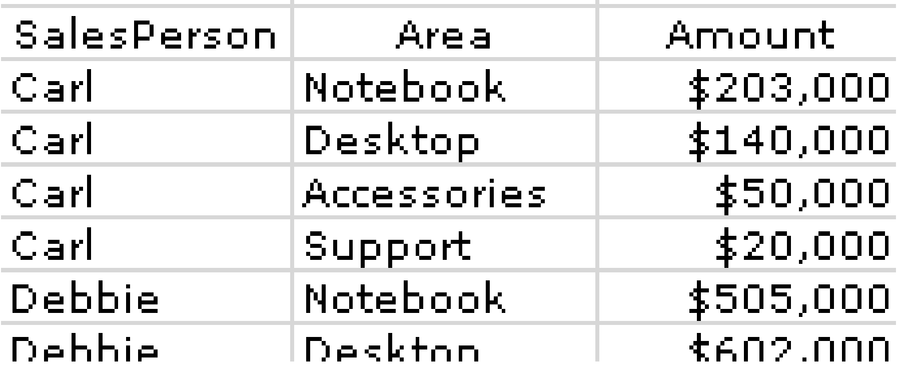
Spreadsheets vs. Databases

- Design a spreadsheet to store:
  - Salesperson, type of sale, sales amount
  - Now add:
    - Region
    - Returns
    - Customer

Spreadsheets and Databases Represent Data Differently



◇	A	B	C	D	E
1		Notebook	Desktop	Accessories	Support
2	Carl	$203,000	$140,000	$50,000	$20,000
3	Debbie	$505,000	$602,000	$40,315	$30,252
4	Lindsey	$306,212	$311,233	$31,525	$21,223
5	Daniel	$71,732	$61,232	$62,313	$15,251
6	Ryan	$8,200	$13,222	$52,555	$62,313



SalesPerson	Area	Amount
Carl	Notebook	$203,000
Carl	Desktop	$140,000
Carl	Accessories	$50,000
Carl	Support	$20,000
Debbie	Notebook	$505,000
Debbie	Desktop	$602,000

Cross-Tabulation

- A *crosstab* is a conversion from database format to spreadsheet format
- It is necessary for spreadsheet analyses of data
- Applications that perform crosstabs:
  - Access, Excel, ACL, IDEA, Picalo

Table: chargessmall
Rows: Vendor
Cols: Purchaser
Data: sum(Amount)

Spreadsheets

- About 1.4M+ rows
- Cells are often calculations of other cells
- Columns are predefined (A, B, C, …)
- Limited searching ability
- Many blank cells (null values)
- Spreadsheets are wonderful for complex mathematical data storage
  - Loan amortization
  - Stock ratio analysis

Databases

- Virtually unlimited numbers of rows
- Custom-defined columns
- Limited calculations on cell values
- Empty cells (null values) are rare
- Extensive searching capabilities
- Databases are wonderful for data storage
  - Employee records
  - Transaction records

Types of Databases

- Relational
  - Most databases you’ll encounter are thankfully relational
  - Stores data in two-dimensional tables
  - Tables are related to one another
- SAP, PeopleSoft
- Other
  - Hierarchical
  - Object
  - Hash
  - Lotus Notes

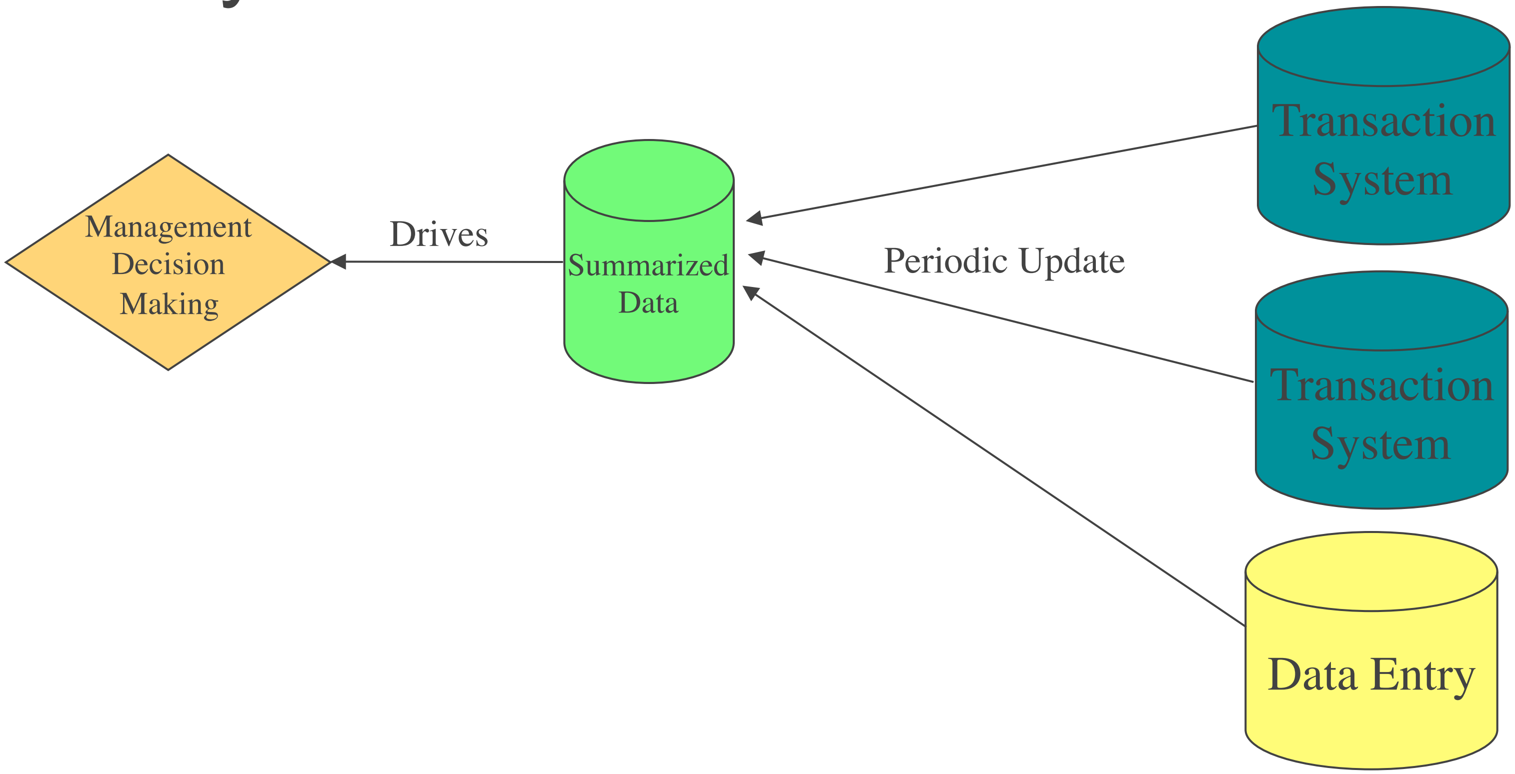
Relational Databases

- Made up of tables (called *relations*)
- Each table has a *primary key*
- Have only as many columns as are defined
- Relatively unlimited number of rows
- Each row usually represents some real world ‘thing’, such as a timecard entry, employee, or purchase
- Cell values are *atomic*
- Columns have data types

HomeMart Database in
MS Access

Data Warehousing

- Data Warehousing is a fancy term for databases specifically designed for analysis



Data Warehousing

- Fraud Detection Data Warehouses
  - Temporary (usually)
  - Data is reloaded each time
  - Designed to highlight transactions, employees, and companies that have symptoms of specific frauds

Databases Used In Warehousing

- MS Excel
  - Easy to use for small data sets
  - Record limits
- MS Access
  - Easy to use
  - Record limits
- Production Databases
  - SQL Server, PostgreSQL, MySQL
  - Harder to use, handles significantly more data
- Corporate Databases
  - Oracle, DB2
  - Normally too heavy for fraud warehouses
- My Recommendation
  - Production Database with MS Access/IDEA/ACL/Picalo front end

PgAdmin III

Getting Data

- Option 1: Query yourself with a direct link
- Option 2: Have someone else (IT dept) query and send you the results

- Which is better?

- Which is possible?

Option 1: Get It Yourself

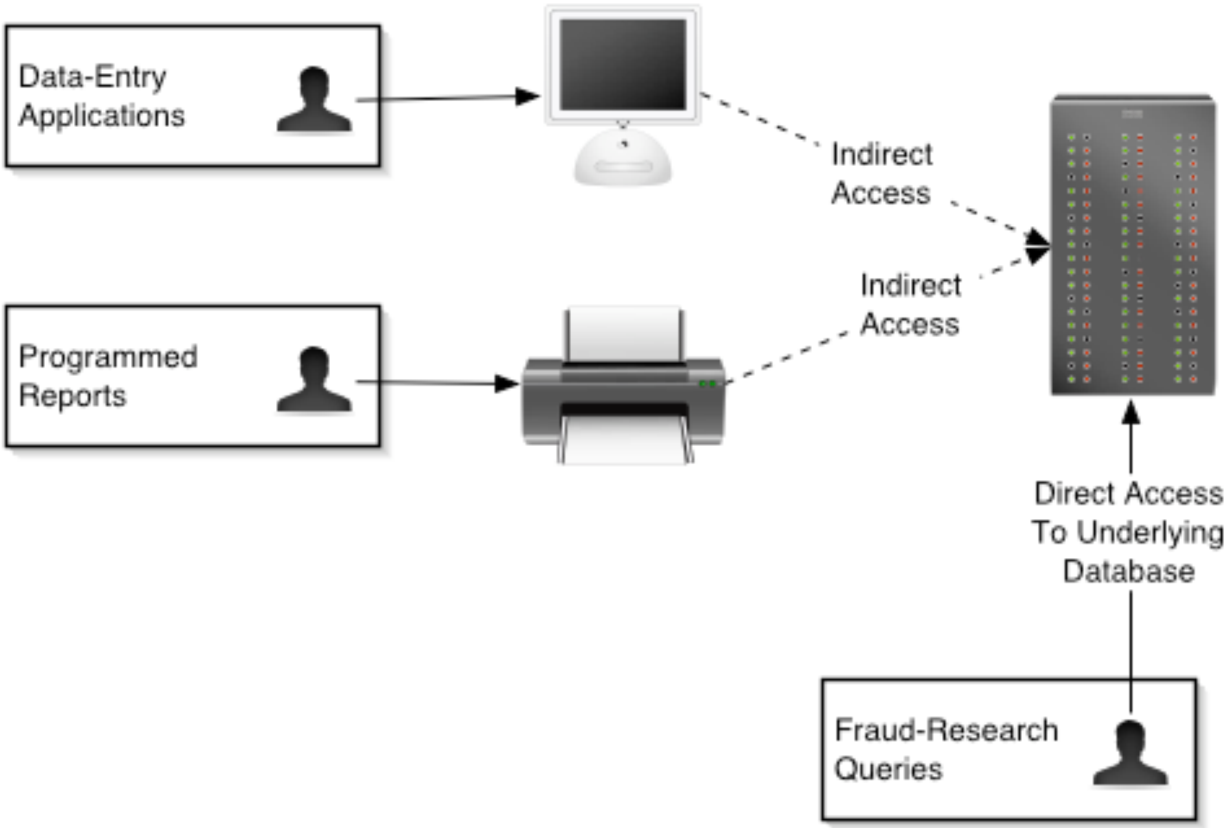
Advantages of Getting it Yourself

- Corporate servers are made to handle the amount of transactions companies have
- Laptops/desktops do not usually have the processing power, memory, or disk space to massage and analyze large amounts of data
- Most corporate servers have unused cycles (at night or on weekends) you can harness to do your analyses for you

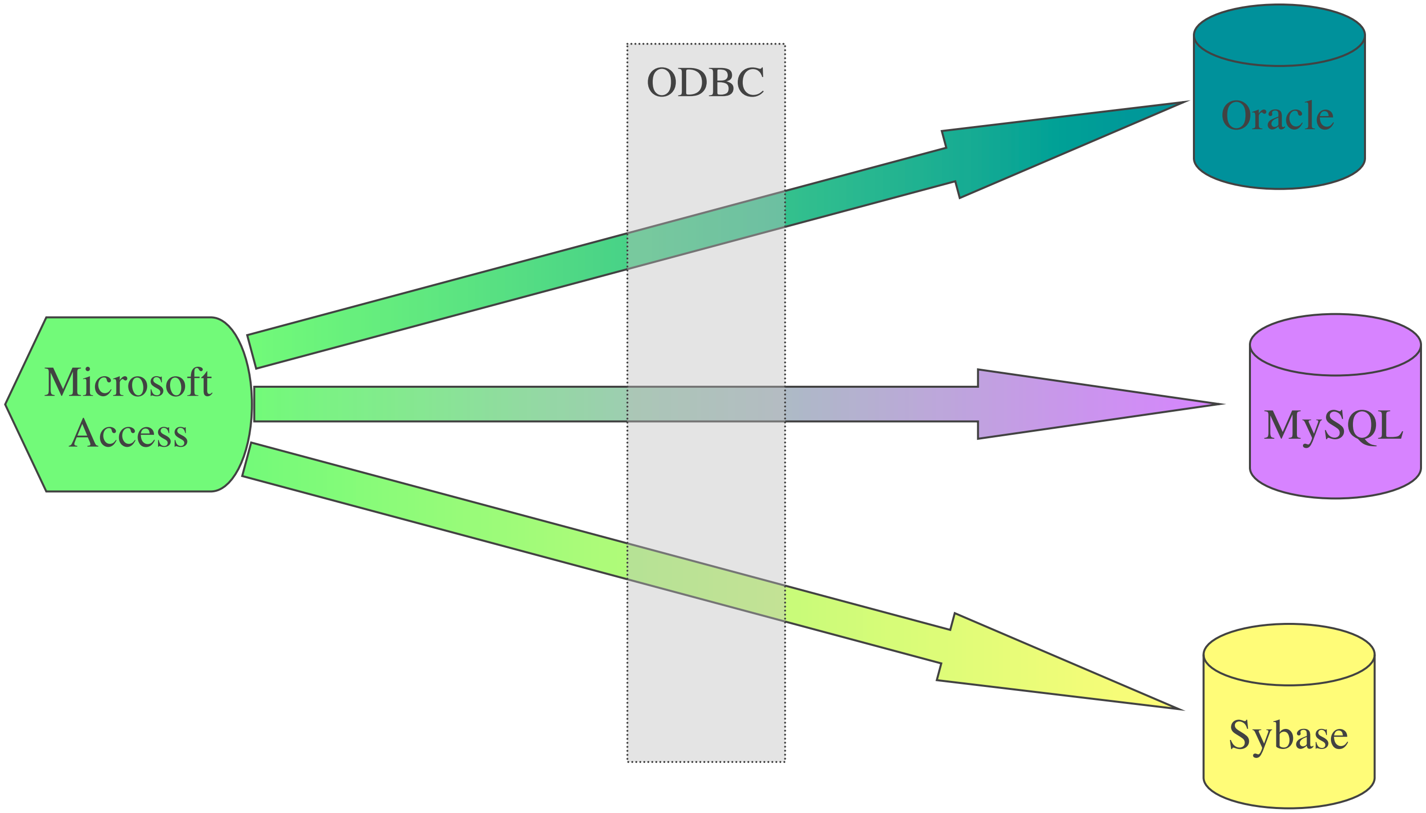
Open Database Connectivity (ODBC)

- Since there are thousands of relational databases, Microsoft developed the ODBC standard
- Provides a standard way of connecting to DB
- ODBC is your friend. Learn to use it!

The ODBC Approach



ODBC Architecture



ODBC Clients

- MS Access is a great ODBC front-end
- ACL and IDEA
- Picalo

MS Access - link tables to “conan” database

Option 2: Get Data From The IT Dept.

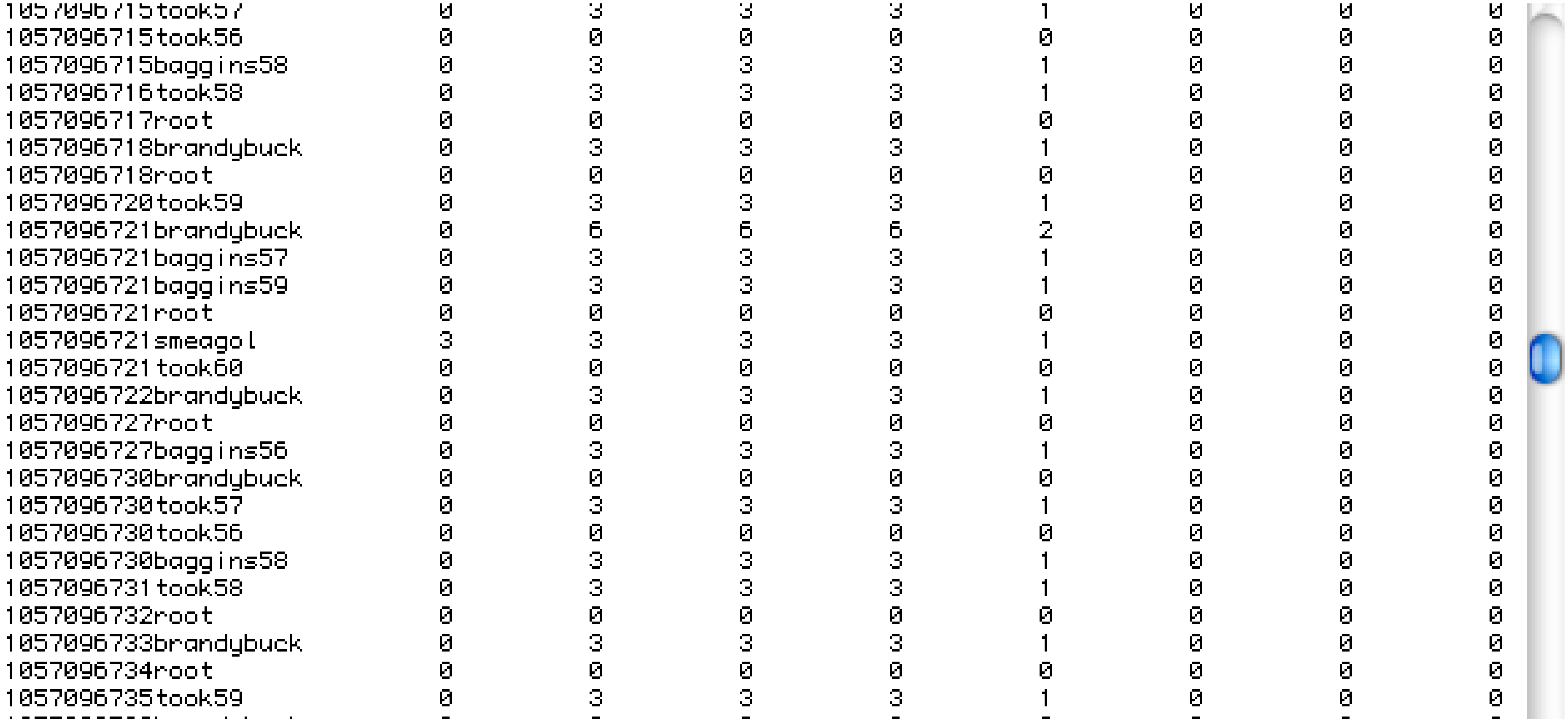
Drawbacks of Using IT

- IT personnel are not trained in fraud detection principles
- IT personnel usually take too long
- IT personnel usually send data via CD or Zip disk, which limits the amount of data that can be sent
- Queries must be run 10 to 15 times to get them property honed to individual data set “personalities”
  - This processs is infinitely more efficient if done directly
- If you want it done right, do it yourself. :)
  - Significant data processing and prepration occurs during and after query runs. It is important that CFE’s are involved in every step of this process

Data Formats

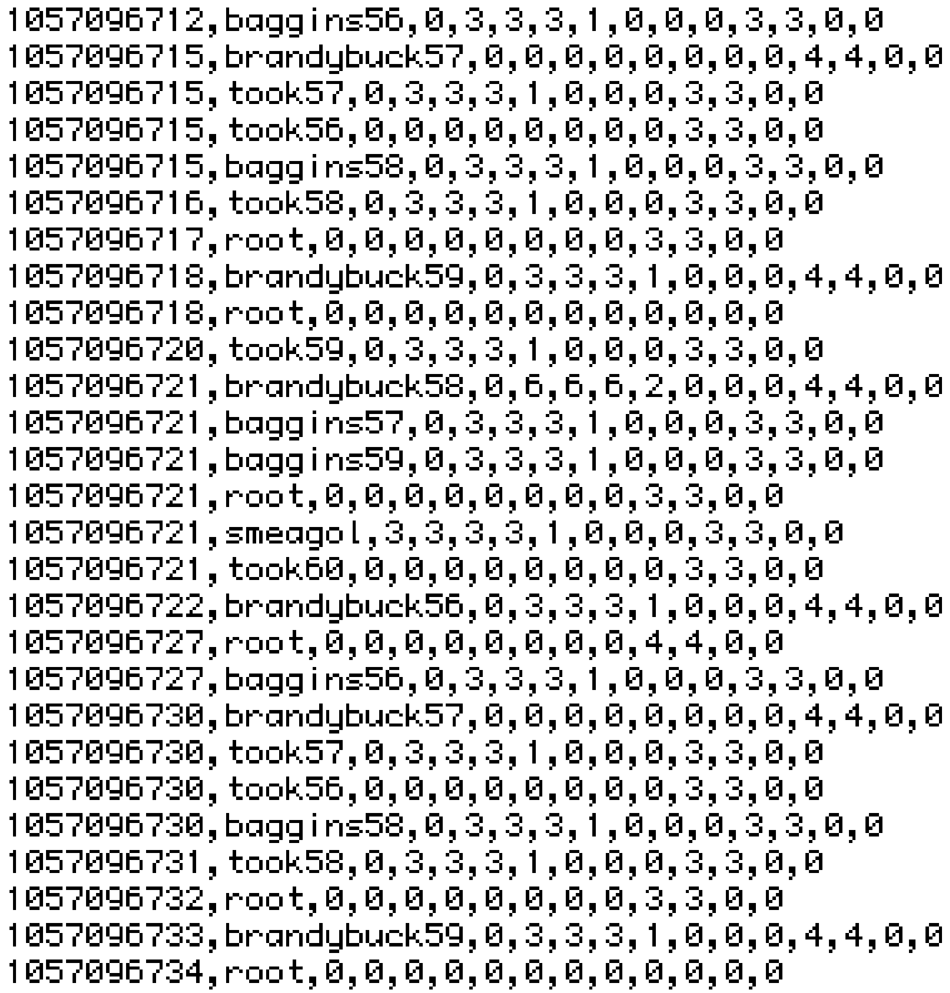
- Data are stored in the computer as 1’s and 0’s
  - A data format is the way these 1’s and 0’s are organized (for example, how fields are delimited)
- Proprietary formats
  - .doc (Word) .xls (Excel) .mdb (Access)
- Open formats
  - .csv (Comma Separated Values)
  - .tsv (Tab Separated Values)
  - .xml (eXtensible Markup Language)
- Most corporate servers will only export data in open formats such as fixed width, delimited text (CSV & TSV), or XML
- Excel and Access can import fixed and delimited text easily

Fixed Width Files



```
1057096715took57          0    3    3    3    1    0    0    0
1057096715took56          0    0    0    0    0    0    0    0
1057096715baggins58       0    3    3    3    1    0    0    0
1057096716took58          0    3    3    3    1    0    0    0
1057096717root            0    0    0    0    0    0    0    0
1057096718brandybuck      0    3    3    3    1    0    0    0
1057096718root            0    0    0    0    0    0    0    0
1057096720took59          0    3    3    3    1    0    0    0
1057096721brandybuck      0    6    6    6    2    0    0    0
1057096721baggins57       0    3    3    3    1    0    0    0
1057096721baggins59       0    3    3    3    1    0    0    0
1057096721root            0    0    0    0    0    0    0    0
1057096721smeagol         3    3    3    3    1    0    0    0
1057096721took60          0    0    0    0    0    0    0    0
1057096722brandybuck      0    3    3    3    1    0    0    0
1057096727root            0    0    0    0    0    0    0    0
1057096727baggins56       0    3    3    3    1    0    0    0
1057096730brandybuck      0    0    0    0    0    0    0    0
1057096730took57          0    3    3    3    1    0    0    0
1057096730took56          0    0    0    0    0    0    0    0
1057096730baggins58       0    3    3    3    1    0    0    0
1057096731took58          0    3    3    3    1    0    0    0
1057096732root            0    0    0    0    0    0    0    0
1057096733brandybuck      0    3    3    3    1    0    0    0
1057096734root            0    0    0    0    0    0    0    0
1057096735took59          0    3    3    3    1    0    0    0
```

CSV (and TSV)



```
1057096712,baggins56,0,3,3,3,1,0,0,0,3,3,0,0
1057096715,brandybuck57,0,0,0,0,0,0,0,0,4,4,0,0
1057096715,took57,0,3,3,3,1,0,0,0,3,3,0,0
1057096715,took56,0,0,0,0,0,0,0,0,3,3,0,0
1057096715,baggins58,0,3,3,3,1,0,0,0,3,3,0,0
1057096716,took58,0,3,3,3,1,0,0,0,3,3,0,0
1057096717,root,0,0,0,0,0,0,0,0,3,3,0,0
1057096718,brandybuck59,0,3,3,3,1,0,0,0,4,4,0,0
1057096718,root,0,0,0,0,0,0,0,0,0,0,0,0
1057096720,took59,0,3,3,3,1,0,0,0,3,3,0,0
1057096721,brandybuck58,0,6,6,6,2,0,0,0,4,4,0,0
1057096721,baggins57,0,3,3,3,1,0,0,0,3,3,0,0
1057096721,baggins59,0,3,3,3,1,0,0,0,3,3,0,0
1057096721,root,0,0,0,0,0,0,0,0,3,3,0,0
1057096721,smeagol,3,3,3,3,1,0,0,0,3,3,0,0
1057096721,took60,0,0,0,0,0,0,0,0,3,3,0,0
1057096722,brandybuck56,0,3,3,3,1,0,0,0,4,4,0,0
1057096727,root,0,0,0,0,0,0,0,0,4,4,0,0
1057096727,baggins56,0,3,3,3,1,0,0,0,3,3,0,0
1057096730,brandybuck57,0,0,0,0,0,0,0,0,4,4,0,0
1057096730,took57,0,3,3,3,1,0,0,0,3,3,0,0
1057096730,took56,0,0,0,0,0,0,0,0,3,3,0,0
1057096730,baggins58,0,3,3,3,1,0,0,0,3,3,0,0
1057096731,took58,0,3,3,3,1,0,0,0,3,3,0,0
1057096732,root,0,0,0,0,0,0,0,0,3,3,0,0
1057096733,brandybuck59,0,3,3,3,1,0,0,0,4,4,0,0
1057096734,root,0,0,0,0,0,0,0,0,0,0,0,0
```

XML

- XML is a powerful markup language
  - More exact than TSV/CSV, but not supported by most products yet
- XML’s strength is in cross-platform data transfer
  - It is wonderful for import and export

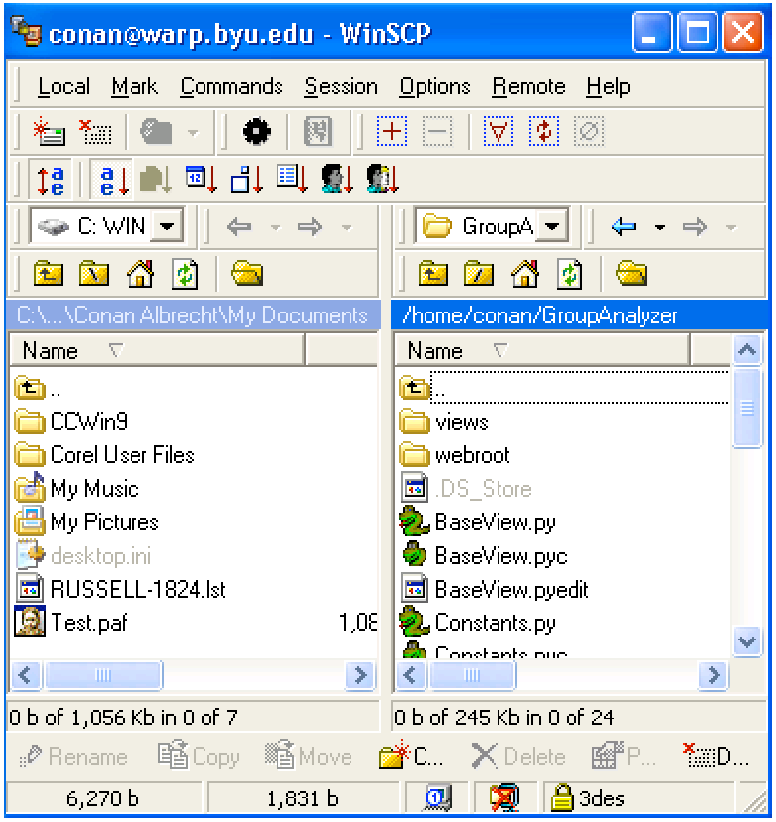
```
<data>
  <employee id=“123456”>
    <FirstName>Louis</FirstName>
    <LastName>Sampsonite</LastName>
    <Salary>24000</Salary>
    ...
  </employee>
</data>
```

Data Transfer

- Best options
  - ODBC -- transfer is part of the connection
  - Internet -- transfers virtually unlimited amounts of data
    - Compress files before transfer using Zip or GZip
    - File Transfer Protocol (FTP), Secure Copy (SCP/SFTP)
- Next best options -- physical transfer
  - DVDs (4800 MB)
  - CDs (800 MB)
  - Zip disks (100 or 250 MB)
- Poor options
  - Email -- Must convert to text, most mail servers will stop anyway

Transferring Files Over the Internet

- File Transfer Protocol
  - Very old, reliable transfer mechanism
  - Many Windows clients exist
    - Internet Explorer with ftp:// prefix in url
- Secure FTP
  - Newer, encrypted version of FTP
  - Compresses automatically
  - WinSCP 3: Windows client
- Skype
  - Encrypted IM peer to peer
  - Send files to coworkers securely



C:\...\Conan Albrecht\My Documents

..
CCWin9
Corel User Files
My Music
My Pictures
desktop.ini
RUSSELL-1824.lst
Test.paf 1,08

0 b of 1,056 Kb in 0 of 7

/home/conan/GroupAnalyzer

..
views
webroot
.DS_Store
BaseView.py
BaseView.pyc
BaseView.pyedit
Constants.py
Constants.pyc

0 b of 245 Kb in 0 of 24

6,270 b 1,831 b 3des

Importing Data

- Software that will link to data
  - MS Access
  - Picalo
- Software that will import data
  - MS Access
  - Picalo
  - MS Excel
  - ACL
  - IDEA

Table: chargessmall
Access, IDEA, ACL, Picalo

First Steps To Perform

(now that you have your data)

Verify Data Types

- Computers must type data columns to know what operations can be performed
- String
  - VARCHAR, CHAR, etc.
  - Most data can be typed as a string
- Number
  - Integer (int, long): no decimal point
    - Take less memory than decimal numbers
  - Decimal (float, double, money): decimal point
- Date
  - Databases, cultures, time zones, countries have a wide variance in formatting

Expression converting in Picalo, chargessmall.tsv

Type Conversion

- Data normally need type conversion after import
- Data scales must be consistent
- Fixed and delimited text files have *no* typing information
  - “1” + “1” = “11”
- Dates
  - “Standard” SQL date isn’t very standard
  - Unix standard: milliseconds since epoch

Massage Data For Consistency

- Massaging data is ensuring the data is consistent and ready for analysis
  - Computers must have consistent data
  - Real world data is noisy and inconsistent
- Examples
  - Convert all dollar amounts to the same base unit (millions, thousands, etc.)
  - Removing extraneous data
  - Filling in blank values
  - Calculation of new columns
- Real-world Example
  - Invoice dollars with zero amounts would cause errors when found in the denominator

Your data is now ready

for core analysis

Common Pitfalls

- Improperly imported data field formats (numbers imported as text)
- Running calculations on fields that have incompatible types
- Calculating ranges/dates on fields with incompatible types
- Comparing numbers that have different scales (date ranges are notorious)
- Trusting an analysis routine just because it doesn’t throw errors

Adapted from Tammi Johnson, 2004 Annual Fraud Conference

Common Pitfalls

- Not using control totals
- Trying to accomplish too much in a single analysis
- Not creating (and checking) analyses step by step
- Dealing with large data sets
- Not creating “tick” columns to mirror date fields

Common Pitfalls

- Sharing database results with other people (who may take them too far)
- Using technology that is not up to the analysis being done
- Using technology that is far beyond the analysis being done
- Not spot checking analysis output with expected output
- Not understanding the schema of the database you are accessing