



Acceleration Academy 2010

Welcome

Textual ETL



New Webinar!
Feb 26, 2010

Guest Speaker -
Bill Inmon

Register Now for Acceleration Academy
First 20 to Register/Attend Get Free Inmon Book!

The complex block contains a promotional graphic for a webinar. It features a whiteboard on the left with a diagram of data flow represented by 'X' and 'O' characters and arrows. To the right of the whiteboard, the text 'Textual ETL' is at the top, followed by 'New Webinar! Feb 26, 2010'. Below that, it says 'Guest Speaker - Bill Inmon'. At the bottom, a dark grey banner contains the text 'Register Now for Acceleration Academy' and 'First 20 to Register/Attend Get Free Inmon Book!'.

Please Type in Questions Along the Way

Time for Q/A the end

Please fill out survey at the end.

If we can't get to your question, we will
respond via email

Held last Friday of each month

www.xtremedata.com/accelerationacademy

KXEN – The Data Mining Automation Company™

Laura Squier - VP of Technical Services

Game-Changing Data Mining - KXEN's Automation Revolutionizes Modeling Productivity

Agenda:

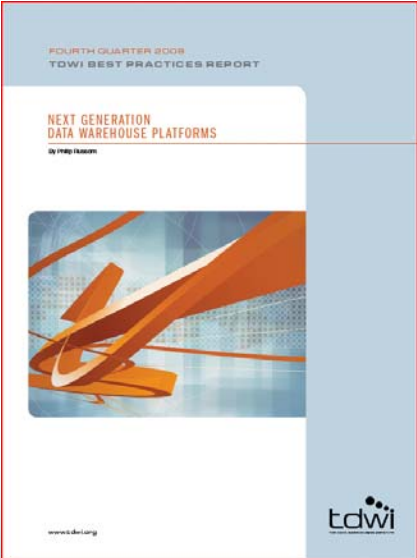
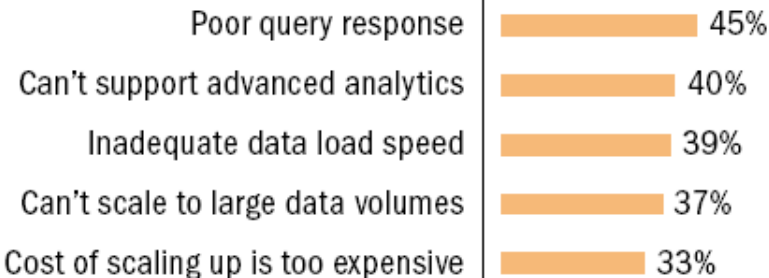
Social Network Analysis

Text Analytics

How to produce models in 1/10th to 1/100th of the time of traditional data mining tools, while not sacrificing model accuracy or robustness

- <https://www1.gotomeeting.com/register/885618120>
- Get details at www.twitter.com/xtremedata

What problems will eventually drive you to replace your current primary data warehouse platform?
(Select nine or fewer.)



Reference TDWI –
Q4 2009 TDWI Best Practices Report –
Next Generation Data Warehouse Platforms by Philip Russom

http://www.tdwi.org/articles/2009/10/08/tdwi-announces-new-best-practices-report-next-generation-data-warehouse-platforms.aspx?sc_lang=en

- 4 years in the making
- Based in Chicago-area
- Deeply experienced leadership team
- Our solution
 - Database appliance - Scalable
 - Operating range 1TB – 3.8 PB User Data
 - Ad hoc/unconstrained access
 - Fast query & load/unload performance
 - Low TCO
 - Energy efficient
 - \$20K/TB Uncompressed – Industry Leading

A large, stylized red arrow pointing to the right, with a white grid pattern on its tail. The text is positioned inside the arrow's body.

Competing on Analytics

Ad-Hoc Data
Exploration

- “Unrestricted” ad hoc access to data
 - Explore data to unearth patterns, relationships, so forth without having to modify/tune underlying database schema
- Scale to support petabytes of data
 - Deep and wide time series, archived and sub-transactional data
- High performance
 - Compress cycle time for analytics studies - from hypothesis thru validated strategy
- Low cost - energy efficient
 - Hardware/software licensing, support/maintenance, data center, power, cooling
- Easy to use
 - ANSI SQL - standard tools and methods



The Unrestricted Ad Hoc Data Warehousing Appliance



DBX™

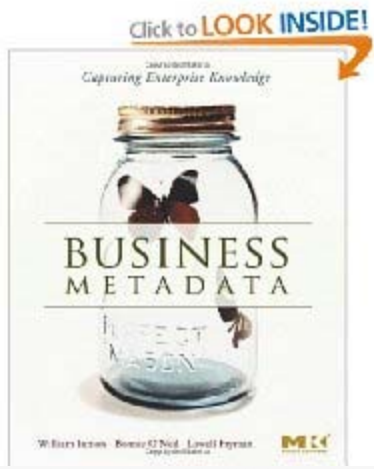
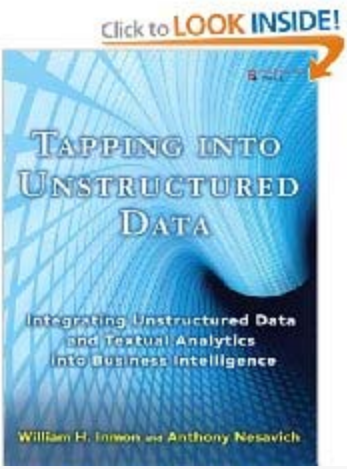
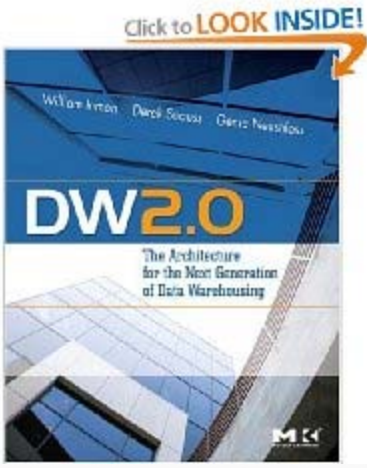
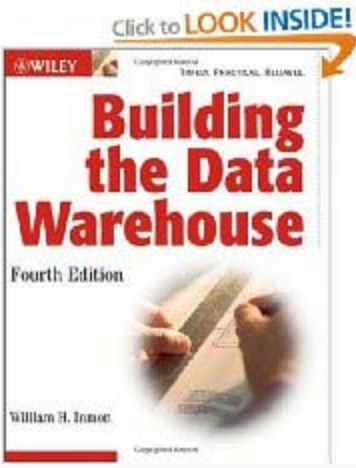


- Fully integrated database appliance
 - Computation + Storage + Interconnect + Database Software
- Full-featured, ACID compliant database
 - ANSI SQL
 - Standard interfaces - ODBC/JDBC
- Designed for unconstrained access
 - Fast execution of ad-hoc/analytic queries
 - Fast load/unload
- Optimal mix of technology
 - Patented IP + Commodity technology
 - Scales from 1TB to 3.8 PB of user data
- Low TCO
 - Industry leading \$20k/TB of user data uncompressed
 - Small carbon footprint

Bill Inmon

The Father of Data Warehousing

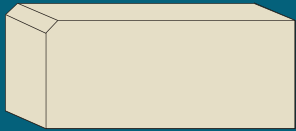
President of Forest Rim Technology Inc.



VOLUMES OF DATA

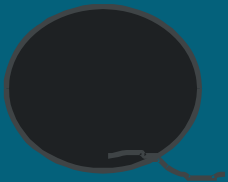
A presentation by
W H Inmon





small volumes

accts payable/receivable;
HR



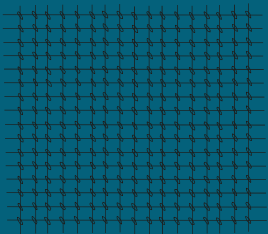
modest volumes

inventory control;
claims management



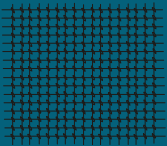
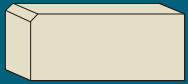
large volumes

bank teller; ATM;
airline reservation;
manufacturing control



really large
volumes

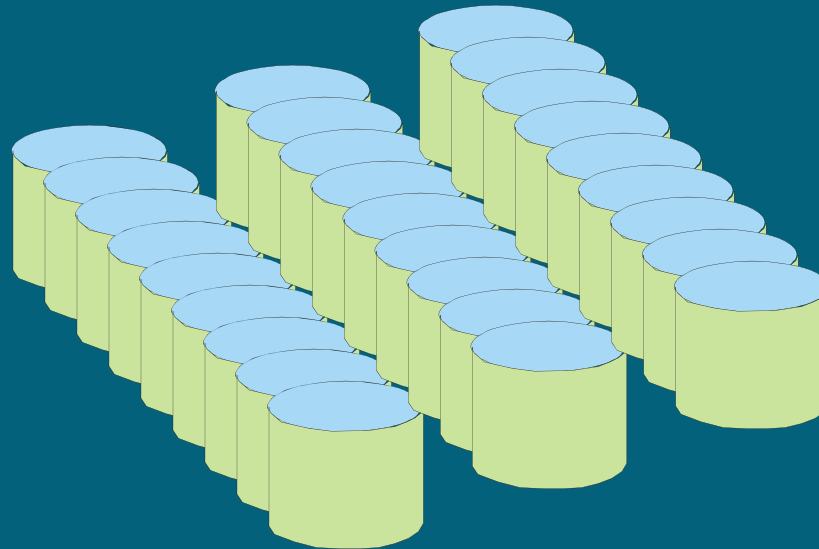
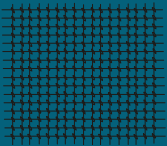
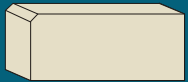
analytic processing;
retail management;
historical trending



the basis of the processing is structured data

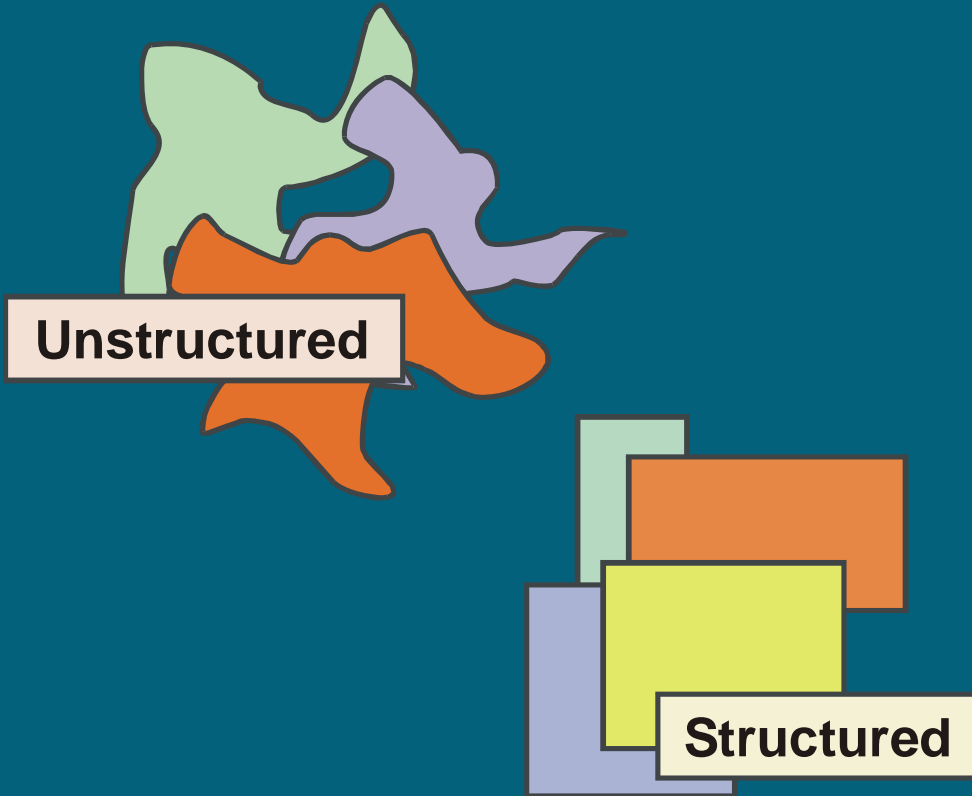
structured data –

- repetitive
- transaction based
- form and structure do not change;
only contents of records change

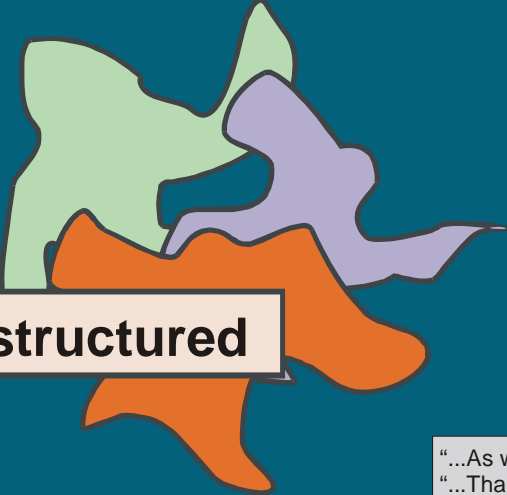


with structured data some large
collections of data have been created –

- data warehouses
- archives
- transaction processing systems



but there is another type of data that exists –
- unstructured data



Unstructured

“...As we see it, the e
 “...Thanks for the sale
 “...I think you ought to
 “...I went down the str
 “...I want to return my
 “...I found a stain on t
 “...Let me tell you hov
 “...Your salesperson i
 “...Yo
 “...I v

“We went for a walk down the park lane. It went by the river and there was a bridge to cross the creek at one point. She was talking so intently that she never even noticed the bridge. She was wrapped up in her thoughts. First there had been the loss at graduate school. Then there was the Derek. And Tally. She just couldn't take it any more. There had to be another answer. Maybe she needed a change. Maybe a change of

The offering includes stock option rights exercisable at .10
 The dealership offered the latest model. Of course you can order off of the Internet and pick up the car at the dealership. When you do it this way you get to choose all the options you want. But the price is not negotiable and you are responsible for selling your car...”

The entire equity was to be a distr

there are volumes of text in the corporation





80-90%



10-20%

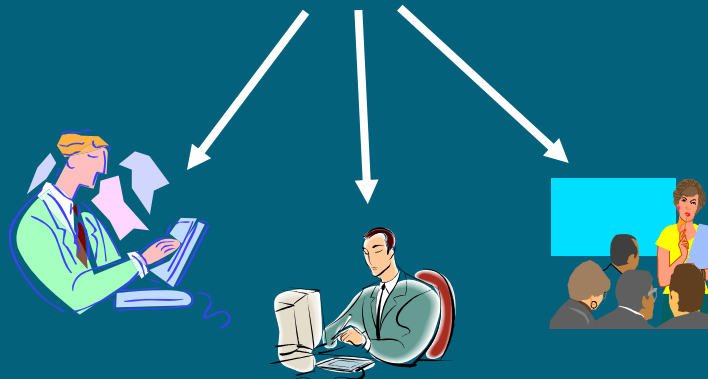
it comes as a surprise to many people that there is far more unstructured data in the corporation than there is structured data



80-90%



10-20%



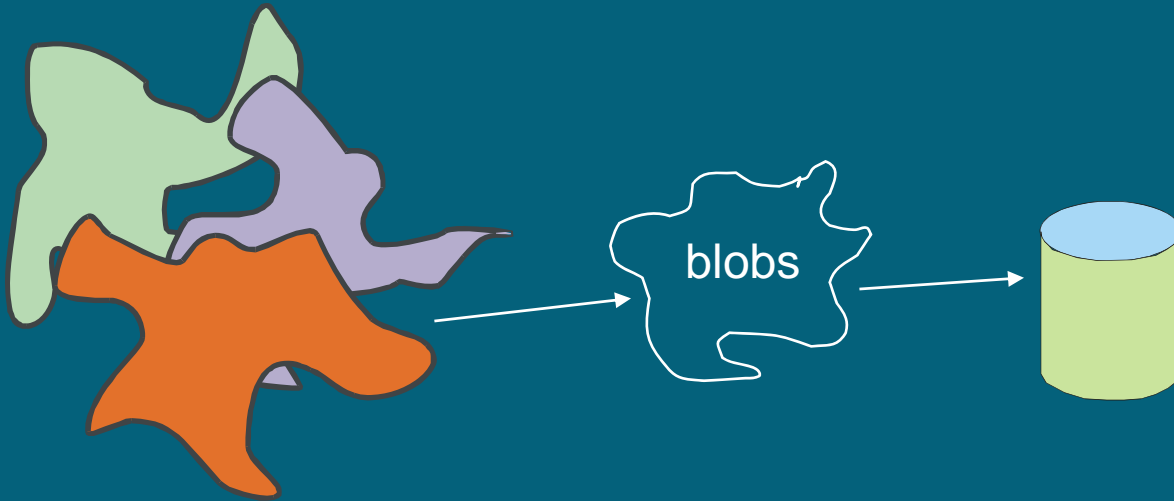
yet 99% of the decision making in the corporation
is done based on structured data



maybe there's not any useful data in the unstructured pile of data

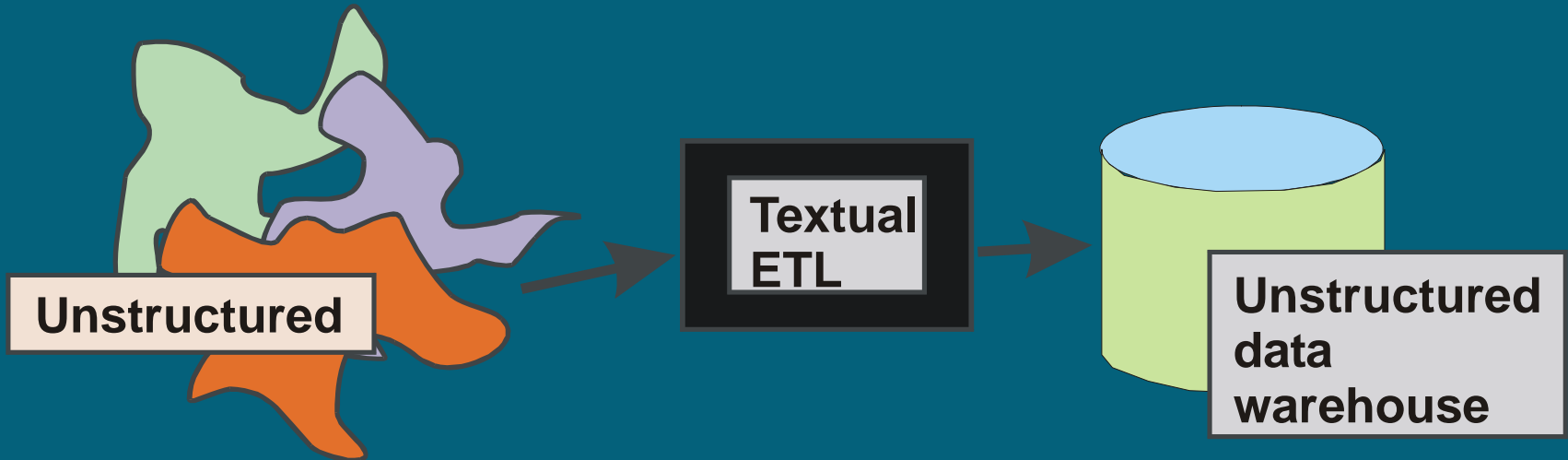
but what about –

- emails
- medical records
- corporate contracts
- legal documents
- loans applications

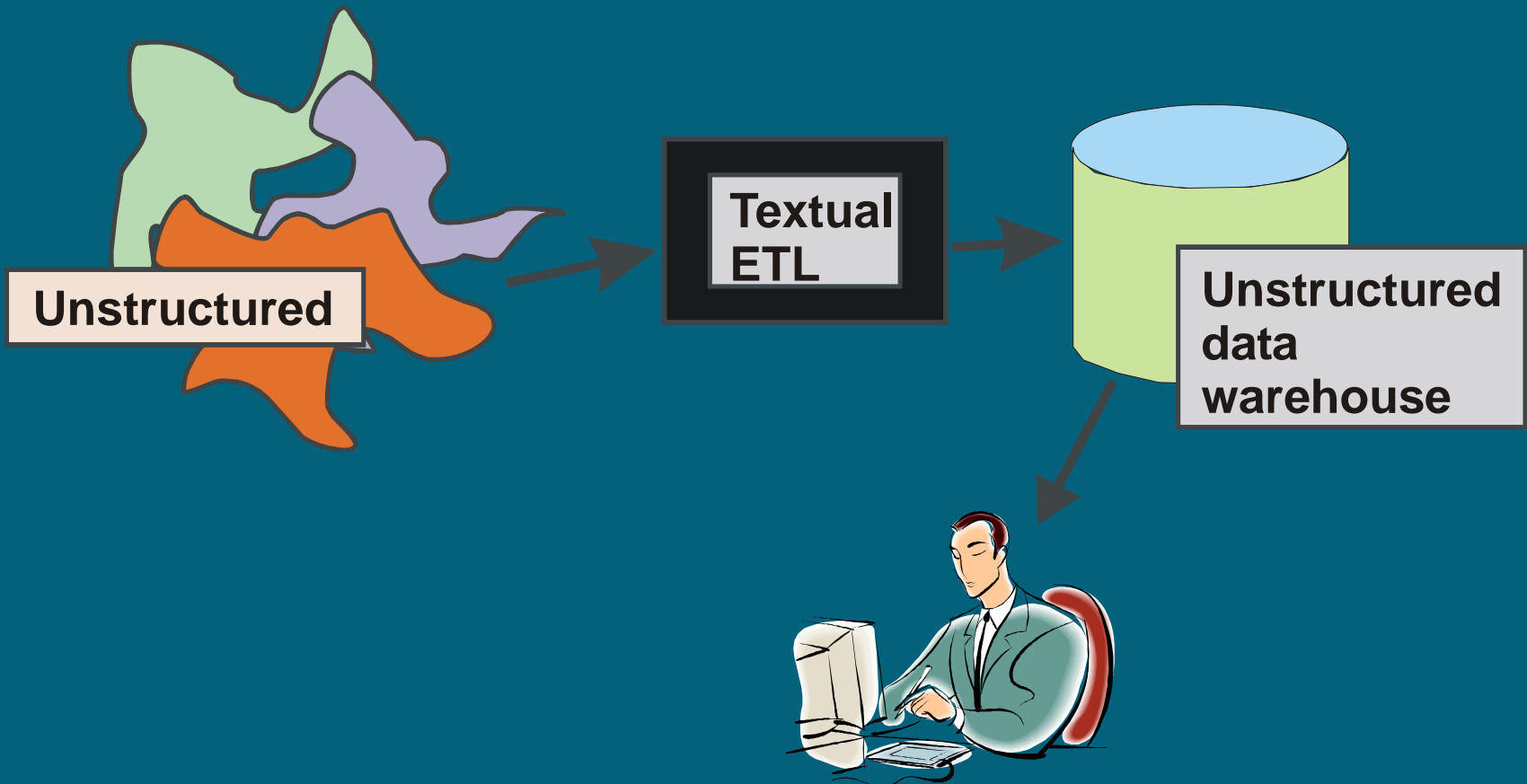


anybody can put blobs of data in a data base.....

so what? putting blobs of data in a data base does not lead to anything productive



with Textual ETL it is possible to create a standard relational data base that is populated with unstructured data



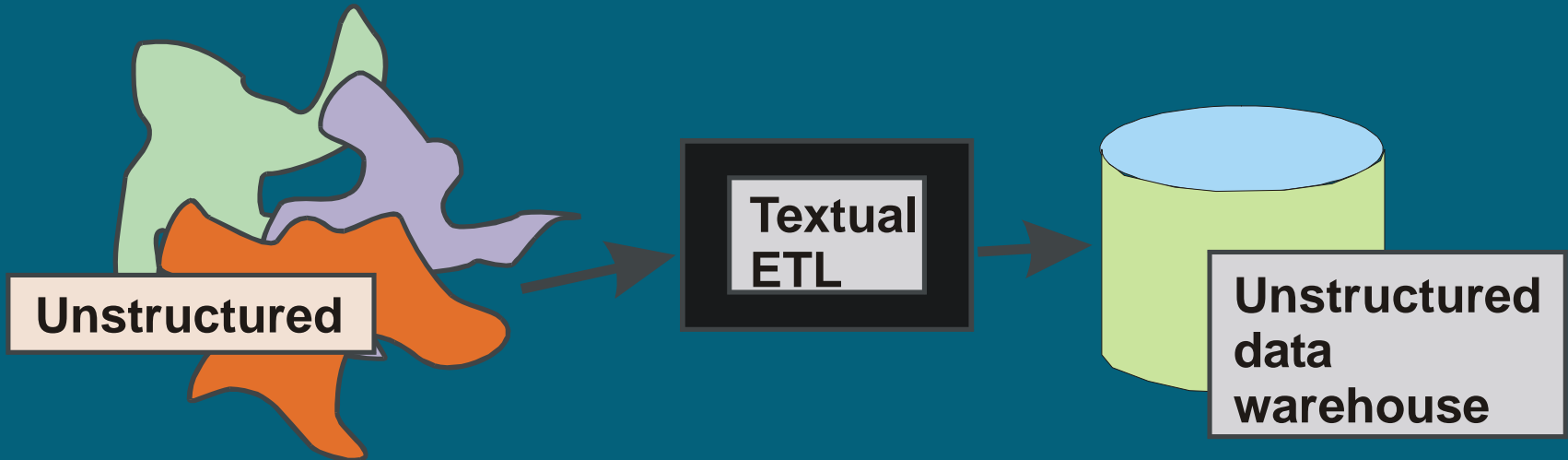
after text has been integrated into a data warehouse,
then and only then can textual analytics be done effectively



- .doc
- .docs
- .xls
- .txt
- many extension types

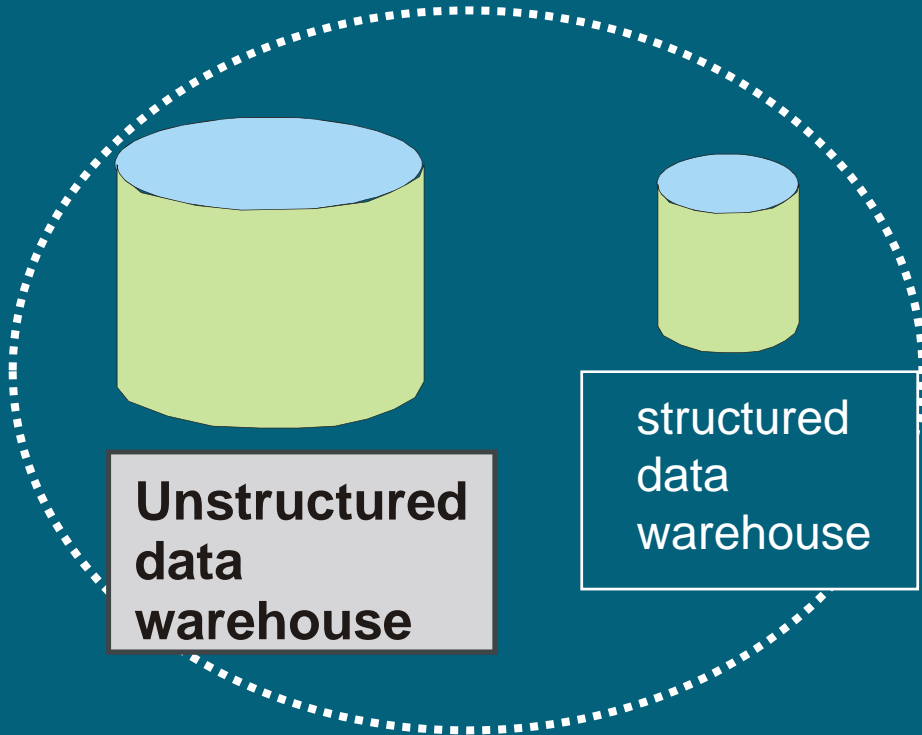
- stop word processing
- homographic resolution
- date standardization
- named value indexing
- document fracturing
- multiple foreign languages
- alternate spelling
- many more forms of transformation

38 types of indexes

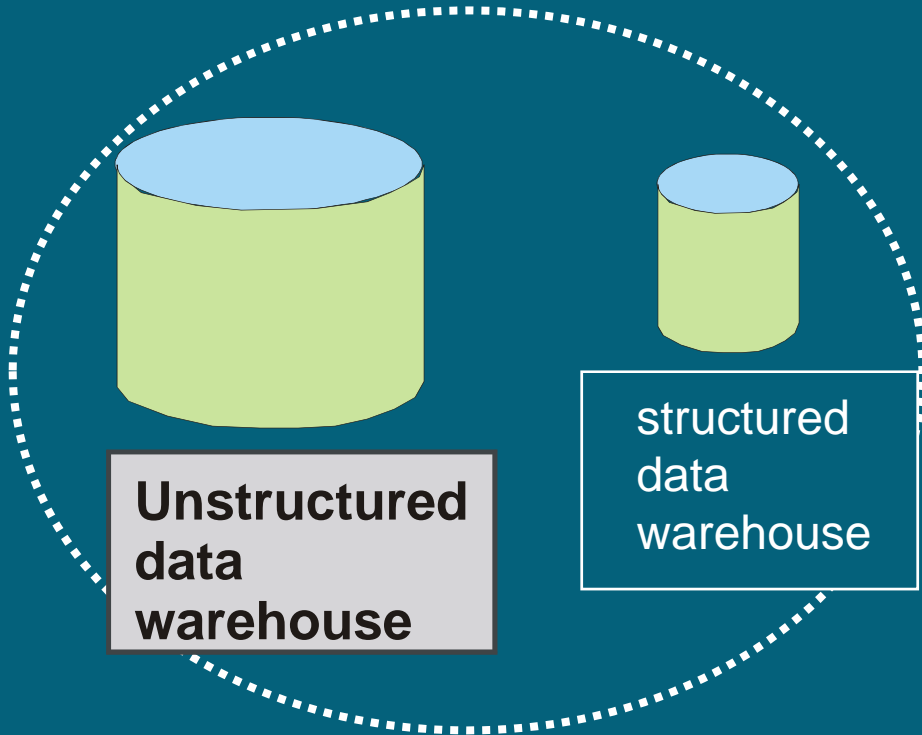


applications –

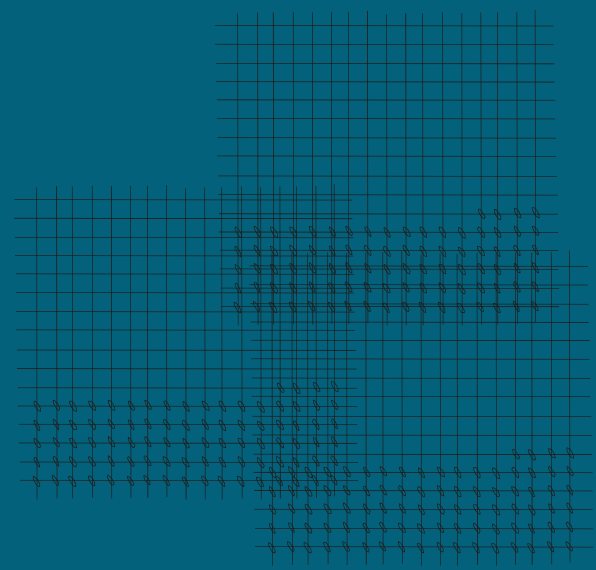
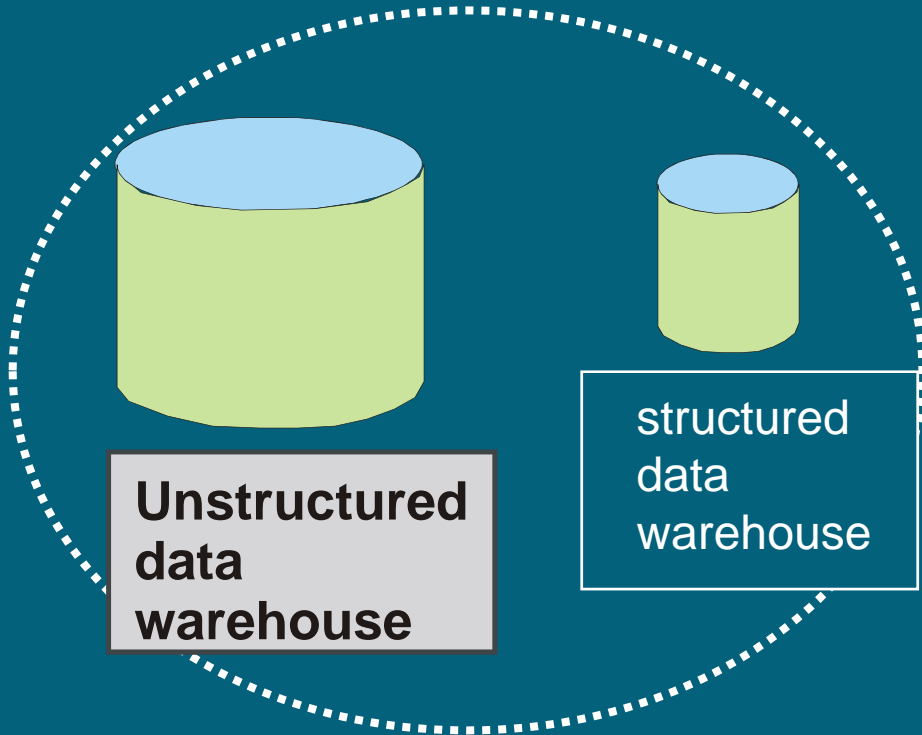
- corporate contracts
- litigation support
- health/medical records
- loans portfolio analysis
- safety systems



once the unstructured data warehouse is created, it can be combined with the classical structured data warehouse



whole new sophisticated forms of analysis can now be done that were never possible before



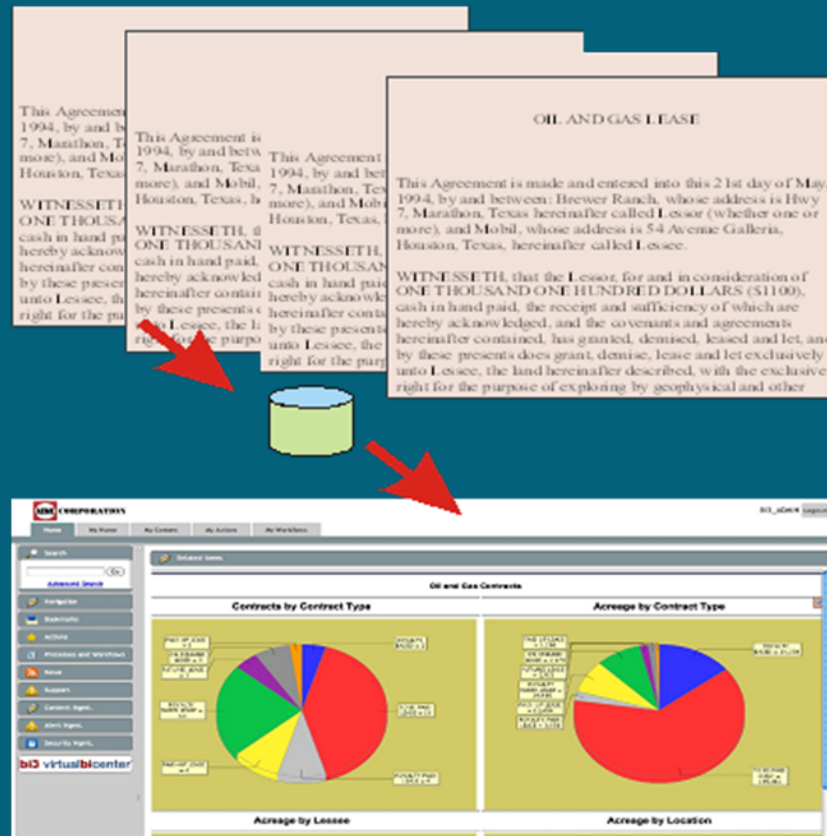
with the volumes of storage that are needed, and the way that queries are done, whole new technologies are needed

Support must include:

- 1) Scalability to 100's of TB
- 2) Extremely Low Price Point
- 3) Unrestricted Ad Hoc Access via SQL



Example of Textual ETL



Summary

- Textual ETL
 - Creating the textual analytical relational data base automatically
- For more information
 - Contact: Heather McClary
 - 303-798-4818
 - heather.mcclary@forestrimtech.com

Thank you..



twitter.com/xtremedata

- First 20 to register and attend today will receive a prize
 - Bill's "Tapping into Unstructured Data" Book
 - Emails will be sent to the winners on Monday March 1st.
 - Books should arrive in a few weeks.
- Article: <http://tdwi.org/Articles/2010/02/03/Next-Wave-DW.aspx>
- Webinar is being Recorded and can be found at www.xtremedata.com/accelerationacademy
- Please fill out the feedback forms – We read each and everyone of them.



Analytics. The wait is over.