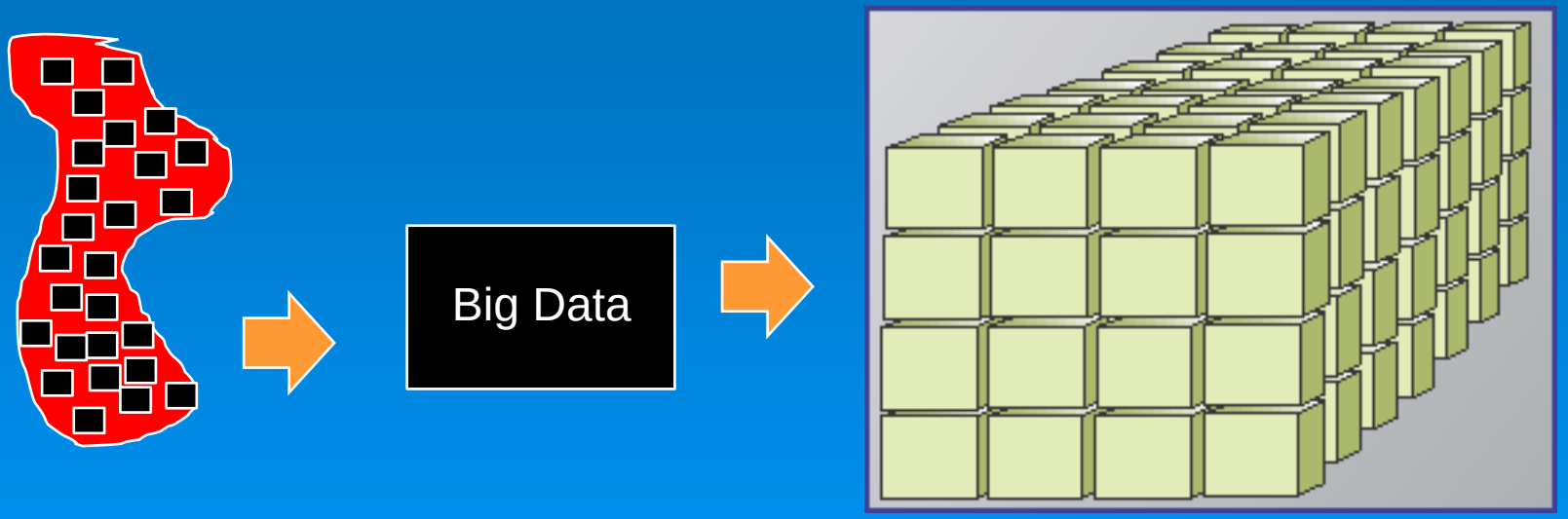
# SWIMMING IN THE DATA LAKE

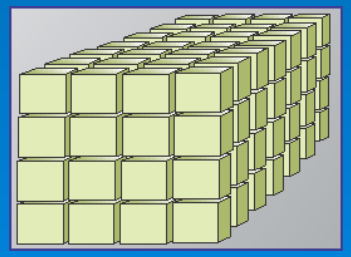
**A presentation by**
**W H Inmon**

FOREST RIM TECHNOLOGY INC




Lots of people are collecting a lot of Big Data

FOREST RIM TECHNOLOGY INC



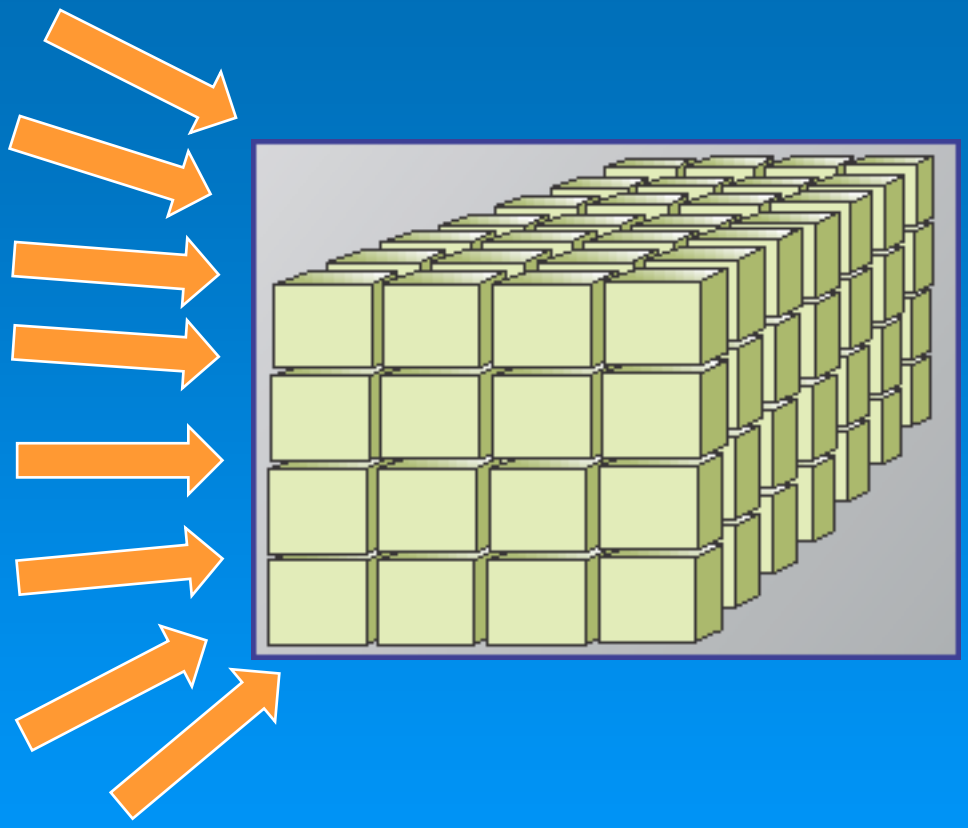
When enough Big Data is collected it is called a “data lake”

FOREST RIM TECHNOLOGY INC

If you are not careful, you wake up one day and
you find your data lake has turned into a garbage dump

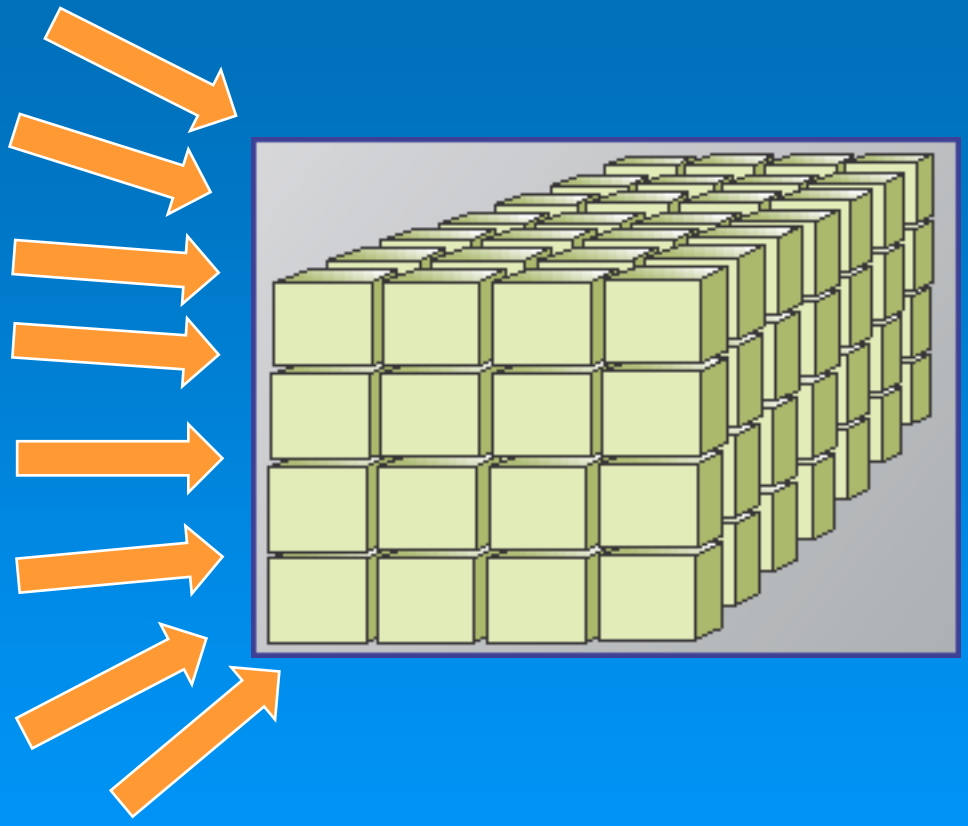
FOREST RIM TECHNOLOGY INC

How did we get to the point of being a garbage dump?



The “one way” data lake – data only goes into the lake but never comes out

FOREST RIM TECHNOLOGY INC



Given enough time, your “one way” data lake starts to “smell”

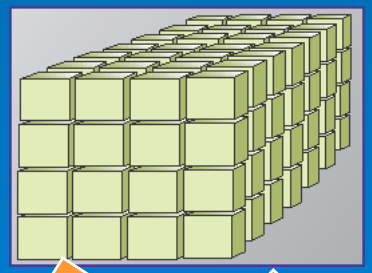
FOREST RIM TECHNOLOGY INC

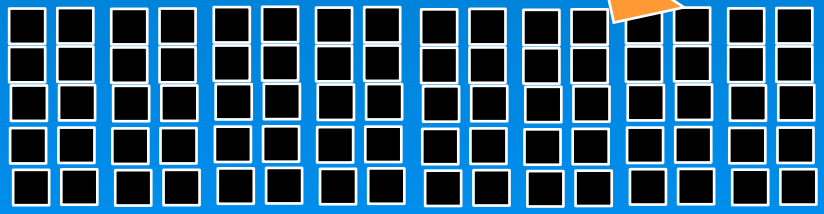
The reason why your data lake turns into a garbage dump

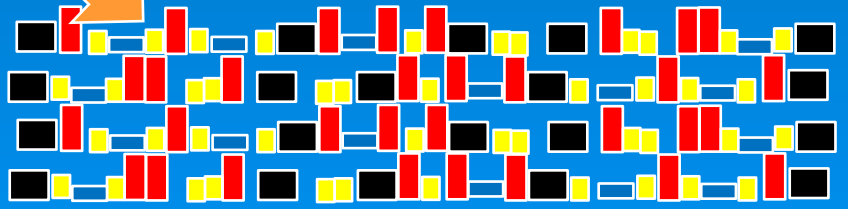
“I can’t find anything in my data lake”

FOREST RIM TECHNOLOGY INC

What's going on here?







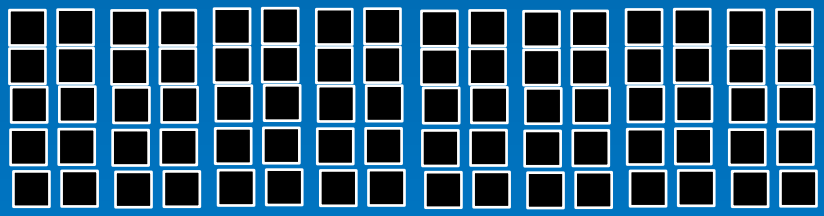
Repetitive data

Non repetitive data

There are two kinds of data in the data lake

FOREST RIM TECHNOLOGY INC

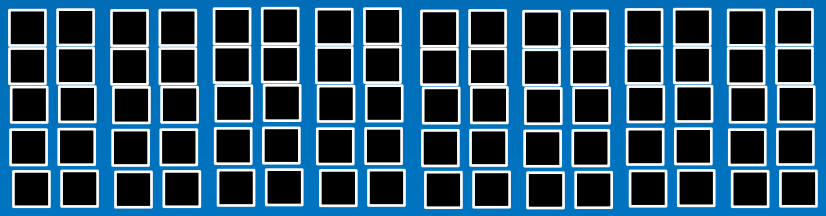
Repetitive data



Repetitive data
- telephone call record detail
- metering data
- click stream data
- log tape data
- meteorological survey data
- and so forth

FOREST RIM TECHNOLOGY INC

Repetitive data

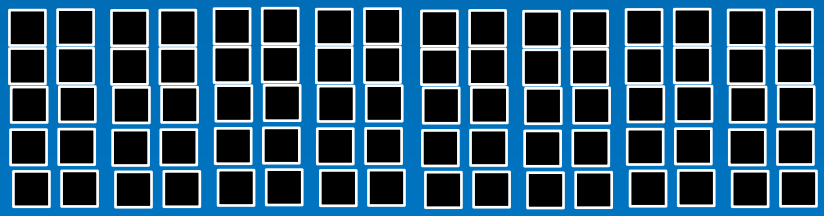


1st requirement

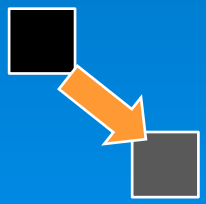
You need metadata to tell you about
- attributes
- definitions of attributes
- records
- keys
- indexes
- sources of data
- refreshment schedule of data
- and so forth

FOREST RIM TECHNOLOGY INC

Repetitive data



2nd requirement

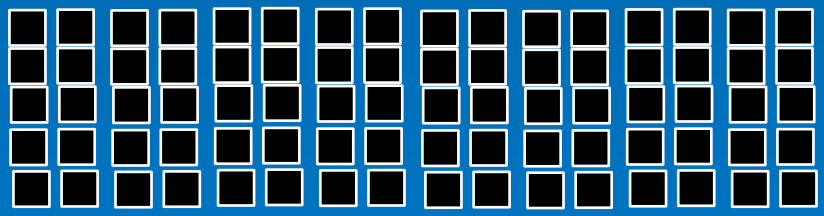


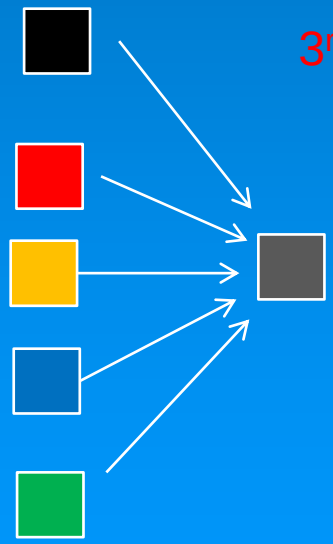
Metadata changes over time

You need to carefully track those changes over time

FOREST RIM TECHNOLOGY INC

Repetitive data





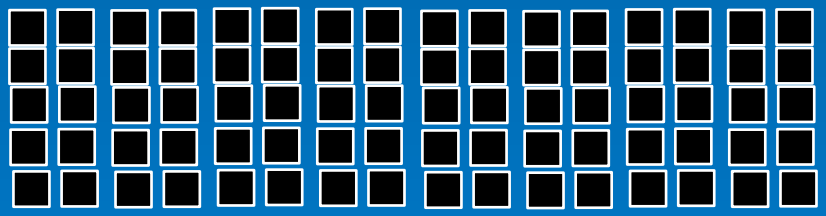
3rd requirement

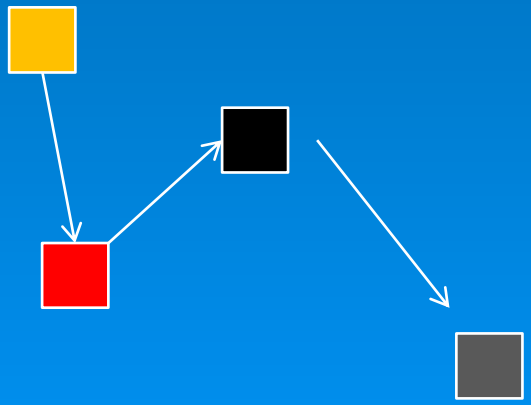
You need metadata transformation rules
in order to see how data needs to be integrated

In data warehousing these were called
“transformation mapping” rules

FOREST RIM TECHNOLOGY INC

Repetitive data



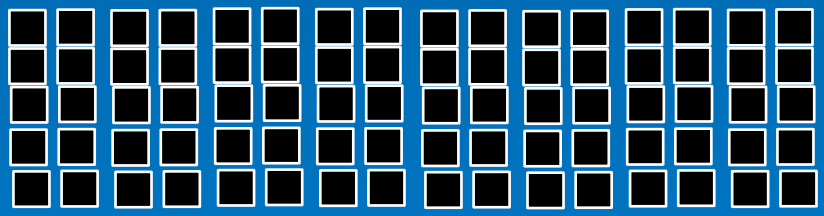


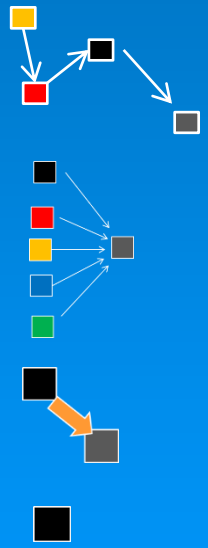
4th requirement

You need to know the “lineage” of the data as it arrived in your data lake

FOREST RIM TECHNOLOGY INC

Repetitive data





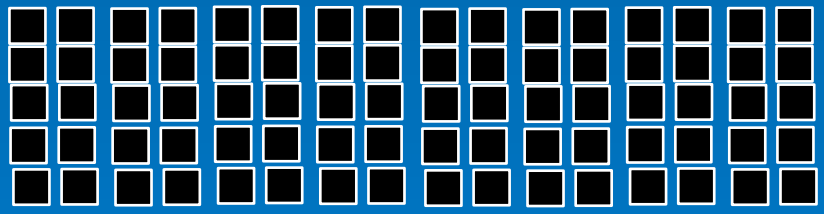
These are called the “transformation” rules

They first appeared in data warehousing.

If you are serious about doing analytical processing in a data lake you have to have transformation rules

FOREST RIM TECHNOLOGY INC

Repetitive data

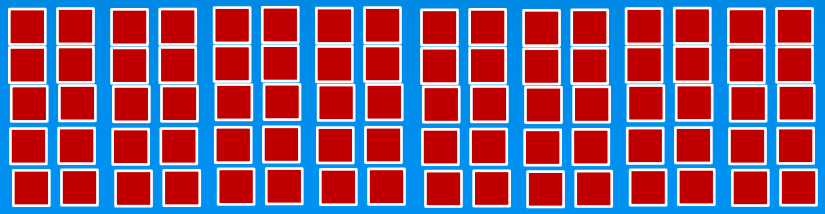


metadata

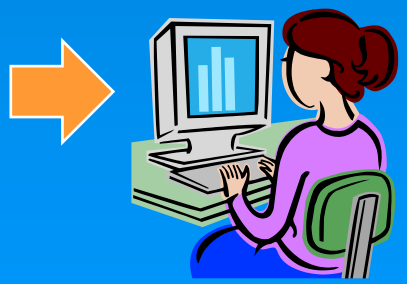
Transformation mapping

transformation

Once repetitive data has been transformed, you can do analysis on it

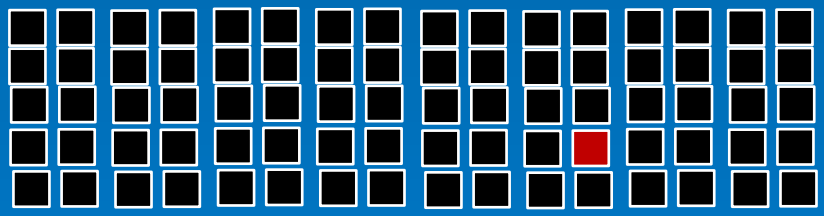


Transformed repetitive data



FOREST RIM TECHNOLOGY INC

Repetitive data



But there is another reason why finding business value from analysis in the repetitive environment is so difficult

often times the data with business value simply isn't there

or…

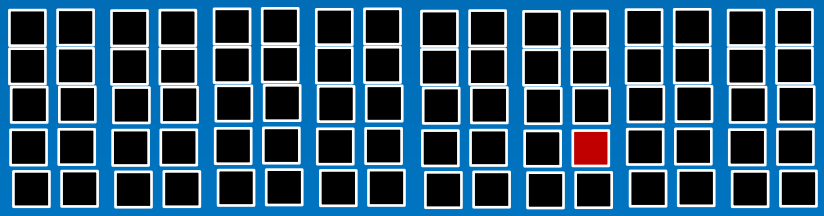
if the data with business value is there it is really hard to find

or…

if there is data with business value there, there just isn't much of it at all

FOREST RIM TECHNOLOGY INC

Repetitive data

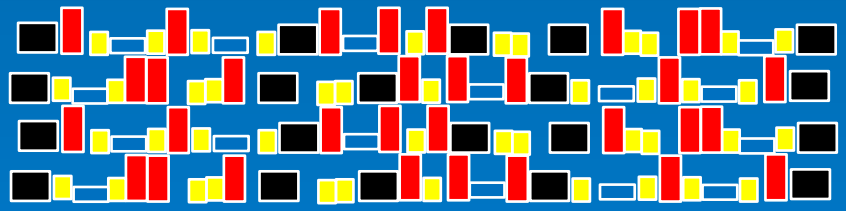


Why is finding business value in repetitive data so difficult?

There simply is no real business value in repetitive data or there is so limited business value in repetitive data that it is not worth finding

FOREST RIM TECHNOLOGY INC

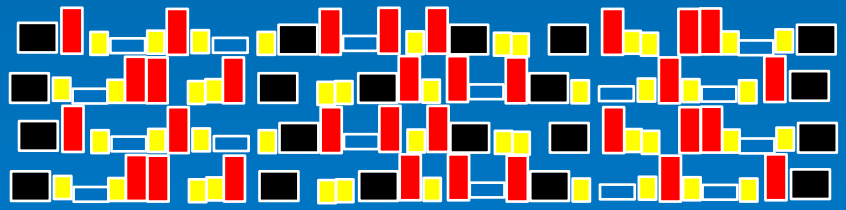
Non repetitive data



Non repetitive data –

- emails
- call center conversations
- customer feedback – restaurants, hotels
- corporate contracts
- medical records
- and so forth

Non repetitive data



Textual disambiguation

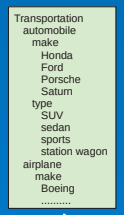
Textual disambiguation is the contextualization of text into a standard data base format.

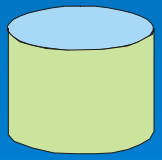
With textual disambiguation you –

- organize text into a form that is suitable for a data base
- identify the context of the text that will be analyzed

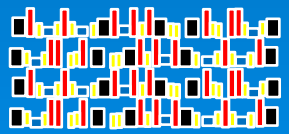
Non repetitive data

taxonomy



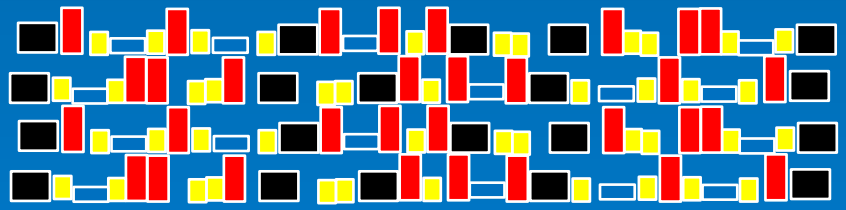



Standard data base format



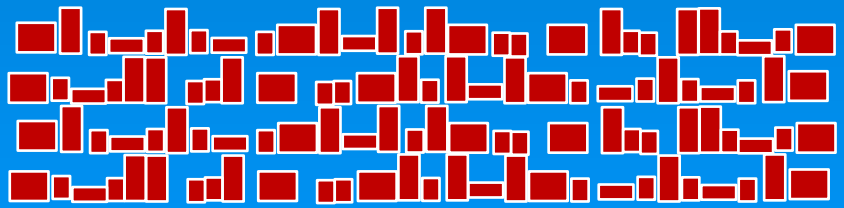
**textual ETL**

Non repetitive data

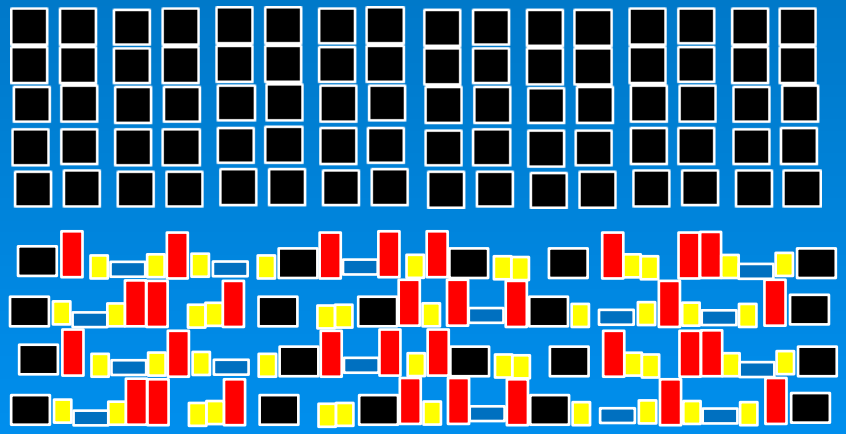


In order to make sense of unstructured, non repetitive data it is necessary to transform the data

Textual disambiguation



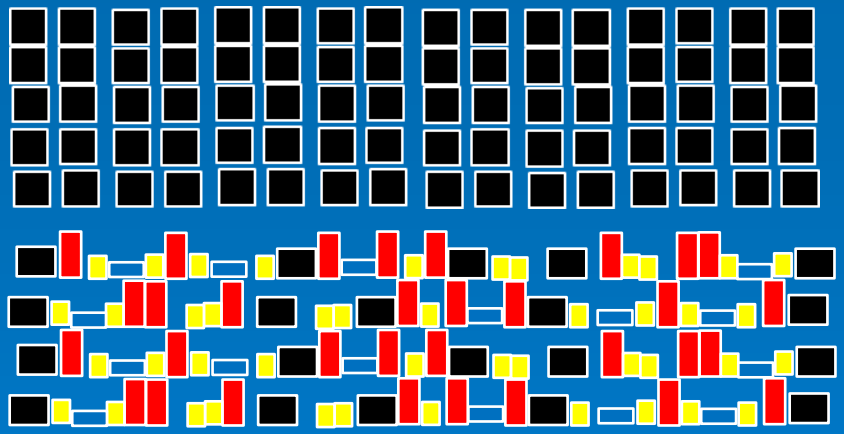
Transformed non repetitive data



Raw Big Data

Access tool

The vision of the Big Data vendor

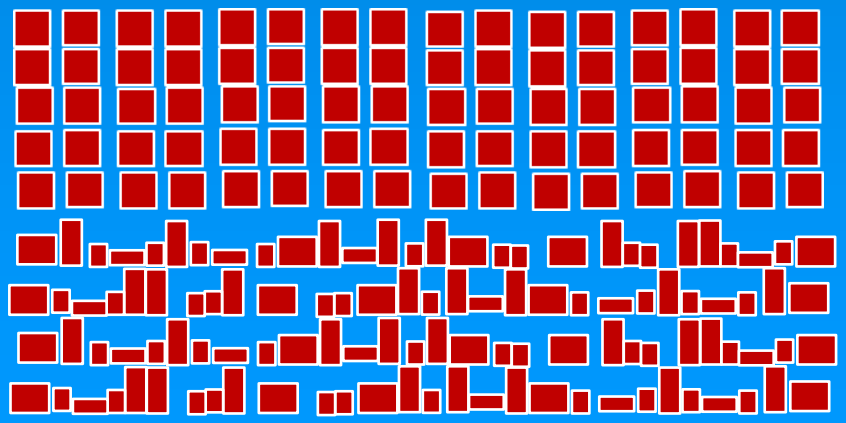


The vision of the data architect
who is serious about doing analysis

Textual
disambiguation

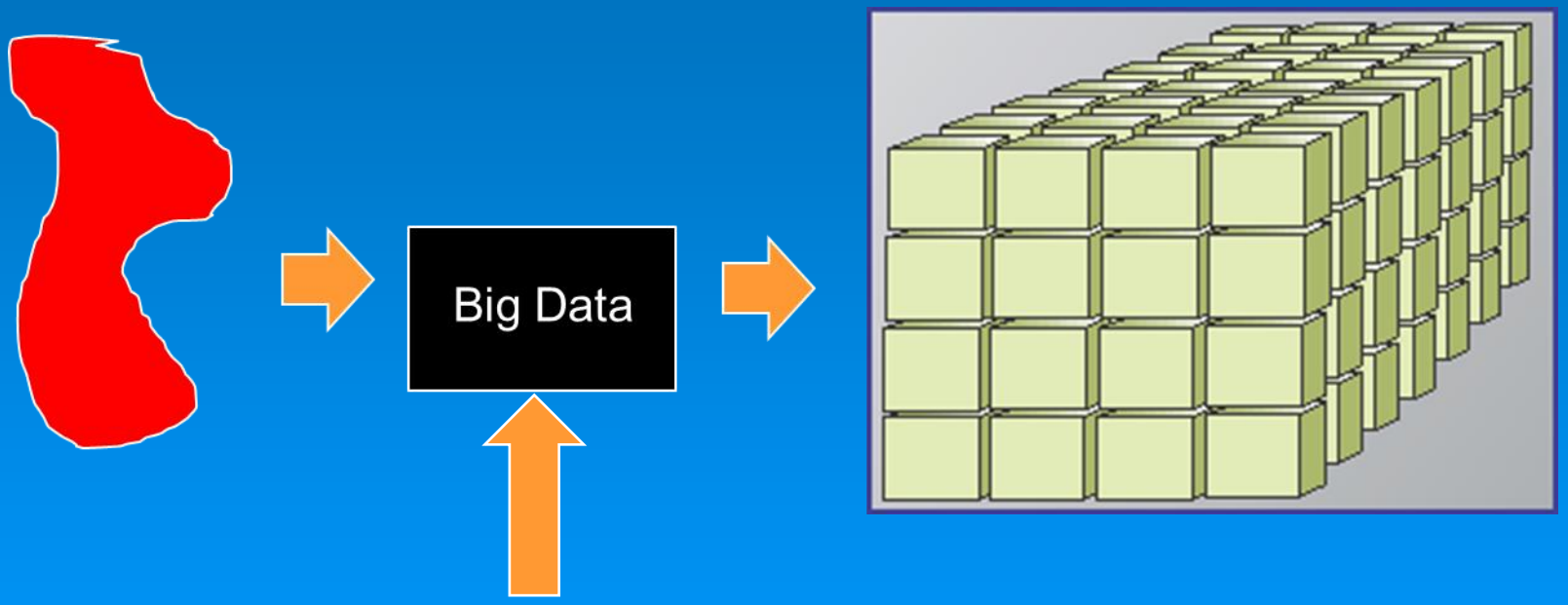
Transformation

Transformation
mapping



Analytical
access
tool

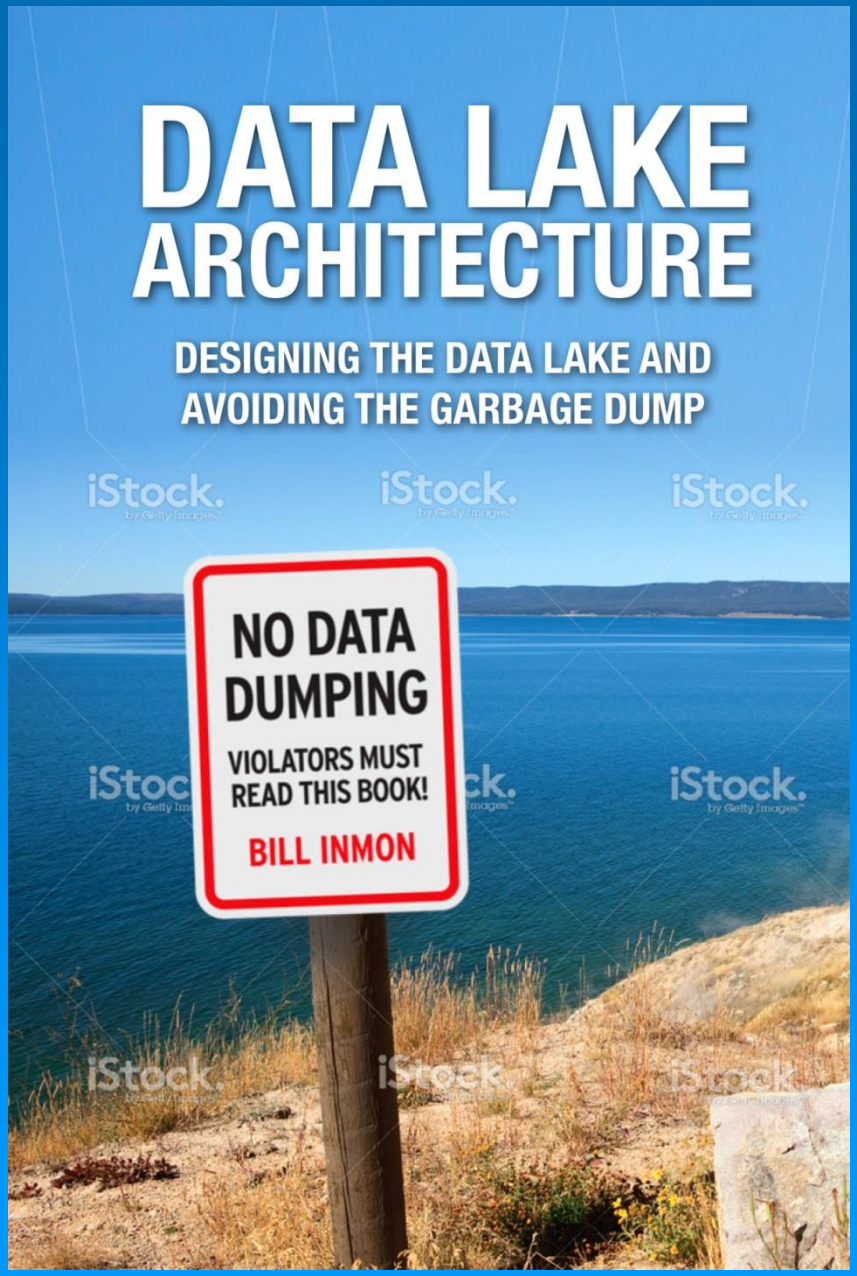
The problem with Big Data and data lakes




The vendors spend 100% of their time and effort on getting the data into the data lake and then tell you all you have to do is to access the data

The problem with Big Data and data lakes

That is why there are so many garbage dumps out there




Bill Inmon's new book