



DATA PREPROCESSING

Pronalaženje skrivenog znanja

Bojan Furlan



WHY DO WE NEED TO PREPROCESS THE DATA?

- Raw data contained in databases is unpreprocessed, incomplete, and noisy.
- For example, the databases may contain:
 - Missing values
 - Outliers
 - Data in a form not suitable for data mining models
 - Fields that are redundant
 - Values not consistent with policy or common sense.
- Database preprocessing:
data cleaning and data transformation.
- Objective: minimize the “garbage” that gets into model,
⇒ minimize the amount of garbage that models give out.
(GIGO)
- Data preparation takes 60% of the data mining process.

DATA CLEANING

Customer ID	Zip	Gender	Income	Age	Marital Status	Transaction Amount
1001	10048	M	75000	C	M	5000
1002	J2S7K7	F	-40000	40	W	4000
1003	90210		10000000	45	S	7000
1004	6269	M	50000	0	S	1000
1005	55101	F	99999	30	D	3000

- Customer 1002 has strange (to American eyes) zip code of J2S7K7.
 - Classify this unusual value as an error and toss it out will be wrong.
 - This is the zip code of St. Hyacinthe, Quebec, Canada, real data from a real customer.
 - Ready to expect unusual values in fields such as zip codes, which vary from country to country.
- Customer 1004 has zip code 6269. Zip codes for the New England states begin with the numeral 0.
 - Zip code field is defined to be numeric and not character (text), the software will probably chop off the leading zero
- Gender contains a missing value for customer 1003.
- customer 1003 is having an income of \$10,000,000 per year.
 - possible when considering the customer's zip code (90210, Beverly Hills), this value of income is outlier, an extreme data value.
- 1004 reported income is negative of -\$40,000 that must be an error.



DATA CLEANING

- Age field has problems. 1001 “age” of C probably reflects an earlier categorization.
 - The data mining software will definitely not like this categorical value in an otherwise numerical field.
- 1004 age of 0?
 - Perhaps there is a newborn male who has made a transaction of \$1000.
 - The age of this person is probably missing and was coded as 0 to indicate this or some other anomalous condition.
- It is better to keep date-type fields (such as birthdate) than age fields, since these are constant and may be transformed into ages.
- The marital status field. The problem lies in the meaning behind these symbols.
 - Does the S for customers 1003 and 1004 stand for single or separated?

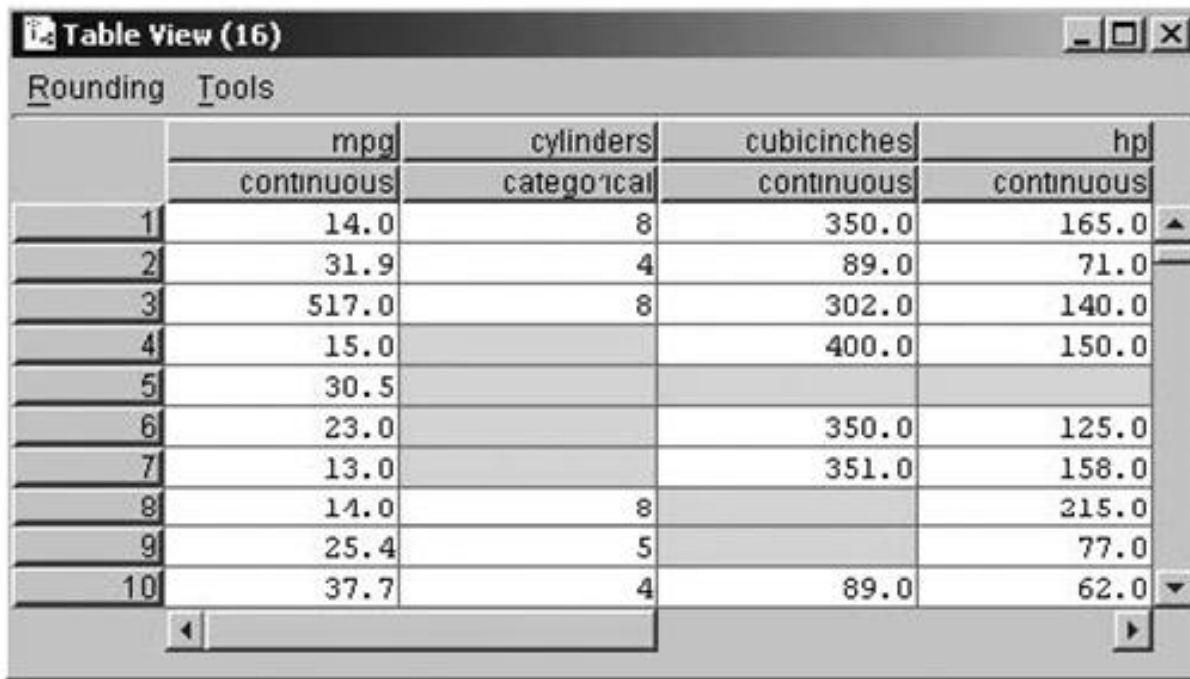


HANDLING MISSING DATA

- A common method of handling missing values: omit that data!
- Dangerous! The pattern of missing values may in fact be systematic:
 - May lead to a biased subset of the data.
 - waste to omit the information in all the other fields, because one field is missing.
- Choice of replacement values for missing data:
 - constant
 - mean (for numerical variables)
or mode (for categorical variables)
 - mode - most frequent appearing value in a data set
 - value generated at random from the variable distribution.

HANDLING MISSING DATA (EXAMPLE)

- Cars data set consists of gas mileage, number of cylinders, cubic inches, horsepower.



The screenshot shows a window titled "Table View (16)" with a menu bar containing "Rounding" and "Tools". Below the menu bar is a table with the following structure:

	mpg	cylinders	cubicinches	hp
	continuous	categorical	continuous	continuous
1	14.0	8	350.0	165.0
2	31.9	4	89.0	71.0
3	517.0	8	302.0	140.0
4	15.0		400.0	150.0
5	30.5			
6	23.0		350.0	125.0
7	13.0		351.0	158.0
8	14.0	8		215.0
9	25.4	5		77.0
10	37.7	4	89.0	62.0

Some of our field values are missing!

EXAMPLE

- for the numerical - constant 0.00
- for the categorical - label *Missing*.

+	mpg	cylinders	cubicinches	hp
	continuous	categorical	continuous	continuous
1	14.00	8	350.00	165.00
2	31.90	4	89.00	71.00
3	517.00	8	302.00	140.00
4	15.00	Missing	400.00	150.00
5	30.50	Missing	0.00	0.00
6	23.00	Missing	350.00	125.00
7	13.00	Missing	351.00	158.00
8	14.00	8	0.00	215.00
9	25.40	5	0.00	77.00
10	37.70	4	89.00	62.00

HANDLING MISSING DATA (EXAMPLE)

	mpg	cylinders	cubicinches	hp
+	continuous	categorical	continuous	continuous
1	14.00	8	350.00	165.00
2	31.90	4	89.00	71.00
3	517.00	8	302.00	140.00
4	15.00	8	400.00	150.00
5	30.50	4	144.15	116.55
6	23.00	4	350.00	125.00
7	13.00	6	351.00	158.00
8	14.00	8	323.45	215.00
9	25.40	5	81.84	77.00
10	37.70	4	89.00	62.00

	mpg	cylinders	cubicinches	hp
	continuous	categorical	continuous	continuous
1	14.00	8	350.00	165.00
2	31.90	4	89.00	71.00
3	517.00	8	302.00	140.00
4	15.00	4	400.00	150.00
5	30.50	4	200.65	106.53
6	23.00	4	350.00	125.00
7	13.00	4	351.00	158.00
8	14.00	8	200.65	215.00
9	25.40	5	200.65	77.00
10	37.70	4	89.00	62.00

- The same value (right) – mean and mode
- Various values (left) - drawn proportionally random from the distribution of attribute values.
 - distribution should remain closer to the original.



HANDLING MISSING DATA

- BUT for both cases there is no guarantee that the resulting records would make sense.
 - E.G. Some record could have drawn:
cylinders = 8 with cubicinches = 82,
a strange engine! (**1 cubic inches = 16.387064 cubic centimeters**)
- Replacing missing values is a gamble:
weight benefits against the possible invalidity
of the results

IDENTIFYING MISCLASSIFICATIONS

Level Name	Count
USA	1
France	1
US	156
Europe	46
Japan	51

```
select origin, count(origin)
from ...
group by origin
```

- The frequency distribution shows five classes: USA, France, US, Europe, and Japan.
- Two of the records have been classified inconsistently: USA ⇒ US, France ⇒ Europe.



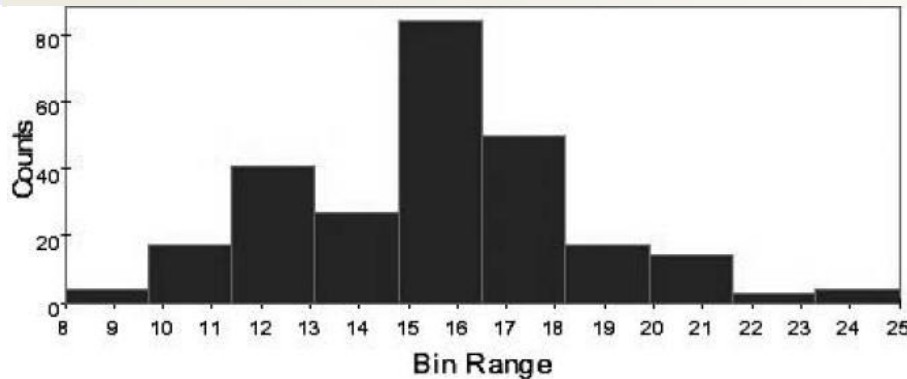
DATA TRANSFORMATION

- Variables tend to have ranges that vary greatly from each other. E.g.– age and incomes
- For some data mining algorithms differences in the ranges will lead to a tendency for the variable with greater range to have bigger influence on the results.
- \Rightarrow normalize numerical variables:
standardize the scale of each variable

- **Min-Max Normalization**
how much greater the field value is than the minimum value $\min(X)$ and scale this difference by the range:

$$X^* = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

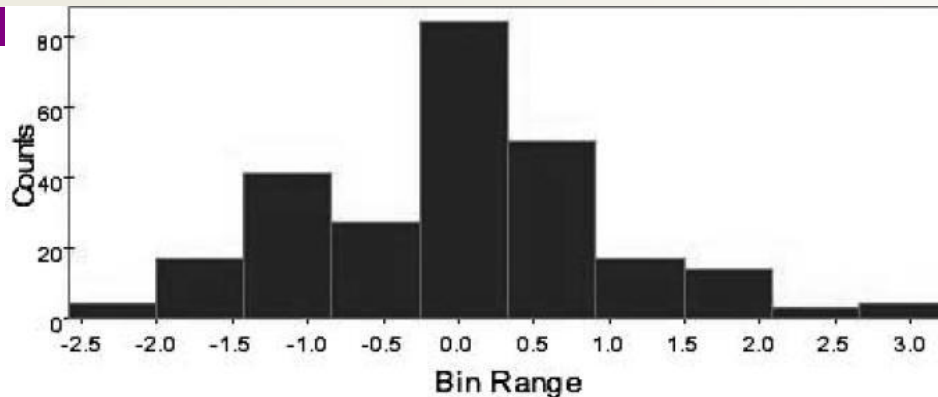
Min-Max Normalization



Count:	261
Missing:	0
Max:	25.0
Min:	8.0
Mean:	15.548
Std dev:	2.911

- Time-to-60 variable from the cars data set: measures how long (in seconds) each automobile takes to reach 60 miles per hour.
- Minmax normalization for three automobiles having times-to-60: 8, 15.548, and 25 seconds:
- Vehicle which takes only 8 seconds (the field minimum)
$$X^* = \frac{X - \min(X)}{\max(X) - \min(X)} = \frac{8 - 8}{25 - 8} = 0$$
- Vehicle (if any), which takes exactly 15.548 seconds (the variable average):
$$X^* = \frac{X - \min(X)}{\max(X) - \min(X)} = \frac{15.548 - 8}{25 - 8} = 0.444$$
- Vehicle which takes 25 seconds (the variable maximum)
$$X^* = \frac{X - \min(X)}{\max(X) - \min(X)} = \frac{25 - 8}{25 - 8} = 1.0$$

Z-Score Standardization



Count:	261
Missing:	0
Max:	3.247
Min:	-2.593
Mean:	0.0
Std dev:	1.0

- Z-score standardization: take the difference between the field value and the field mean value and scale this difference by the standard deviation of the field values.

$$X^* = \frac{X - \text{mean}(X)}{\text{SD}(X)} \quad \sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$$

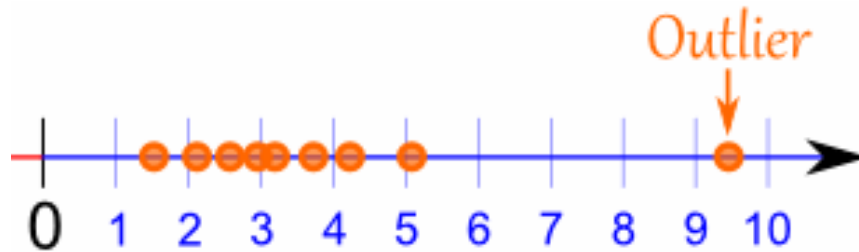
$$X^* = \frac{X - \text{mean}(X)}{\text{SD}(X)} = \frac{15.548 - 15.548}{2.911} = 0$$

$$X^* = \frac{X - \text{mean}(X)}{\text{SD}(X)} = \frac{8 - 15.548}{2.911} = -2.593$$

$$X^* = \frac{X - \text{mean}(X)}{\text{SD}(X)} = \frac{25 - 15.548}{2.911} = 3.247$$

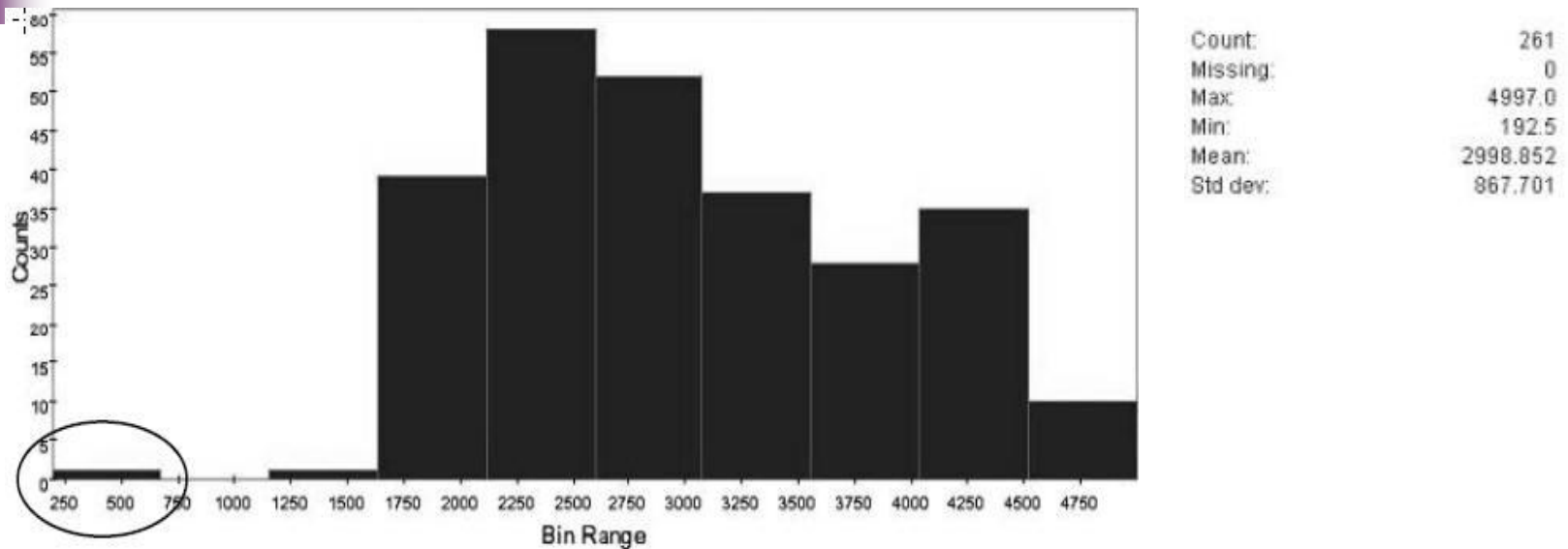
- Z-score standardization values will usually range between -4 and 4, with the mean value having a Z-score standardization of zero.

OUTLIERS



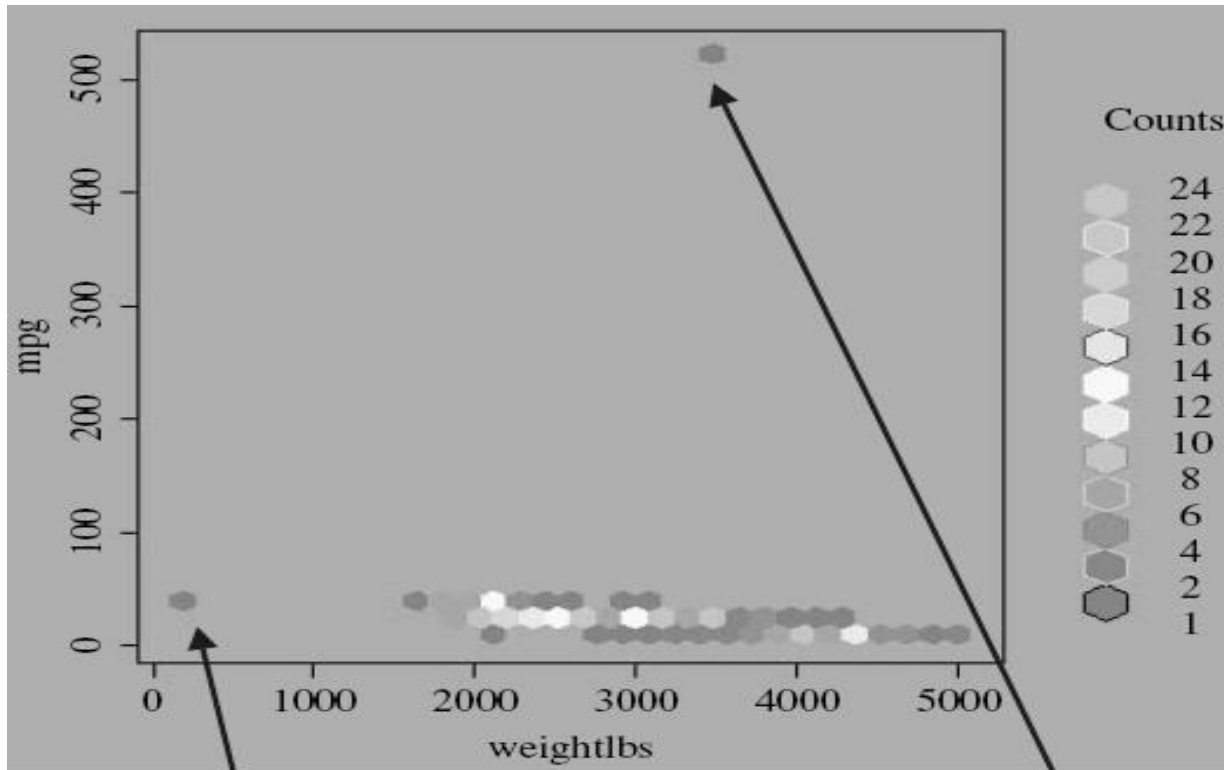
- Outliers are extreme values that lie near the limits of the data range or go against the trend of the remaining data.
- May represent errors or a valid data points.
- Even if it is a valid data point and not error, it may corrupt data analysis:
 - certain statistical methods are sensitive to the presence of outliers and may deliver unstable results.
- How to deal: graphical or numerical methods

GRAPHICAL METHODS FOR IDENTIFYING OUTLIERS



- Histogram generated of the vehicle weights from the cars data set.
- 192.5 pounds is a little light for an automobile (**1 pound = 0.45359237 kilograms**)
- Perhaps the weight was originally 1925 pounds, with the decimal inserted somewhere along the line.

GRAPHICAL METHODS FOR IDENTIFYING OUTLIERS



- Two-dimensional scatter plots: for outliers in more than one variable.
- (left down) the same vehicle as that identified in histogram weighing only 192.5 pounds.
- (upper right corner) a car that gets over 500 miles per gallon!



OUTLIERS: Min-Max Normalization

- Min-max normalization values range is $[0,1]$
 - unless new data values lie outside the original range.
- Outliers in the original data set will make that most of the values lie near extreme (0 or 1)
 - e.g. Car with times-to-60 = 200s makes other values be normalized near 0.
- Problem – Min-max normalization depends on range.



OUTLIERS: Z-score standardization

- An outlier can be identified because it has a value that is: less than -3 or greater than 3
- PROBLEM: if outliers are in the original data set
 - the mean and standard deviation, both part of the formula for the *Z-score*, are sensitive to the presence of outliers.

Interquartile range

- The first quartile (Q1) is the 25th percentile.
- The second quartile (Q2) is the 50th percentile.
- The third quartile (Q3) is the 75th percentile.

- The IQR is calculated as $IQR = Q3 - Q1$ and may be interpreted to represent the range of the middle 50% of the data.
- A data value is an outlier if:
 - It is located $1.5 * IQR$ or more below Q1
 - It is located $1.5 * IQR$ or more above Q3.



Interquartile range - Example

- Set of 12 values
 $\{65, 69, 70, 71, 73, 75, 77, 79, 80, 85, 88, 100\}$
- the 25th percentile was $Q1 = 70$
- the 75th percentile was $Q3 = 80$,
- Interquartile range:
 $IQR = 80 - 70 = 10$.
- half of all the test scores are between 70 and 80.
- A test score would be identified as an outlier if:
 - It is lower than $Q1 - 1.5 * IQR = 70 - 1.5 * 10 = 55$
 - It is higher than $Q3 + 1.5 * IQR = 80 + 1.5 * 10 = 95$.



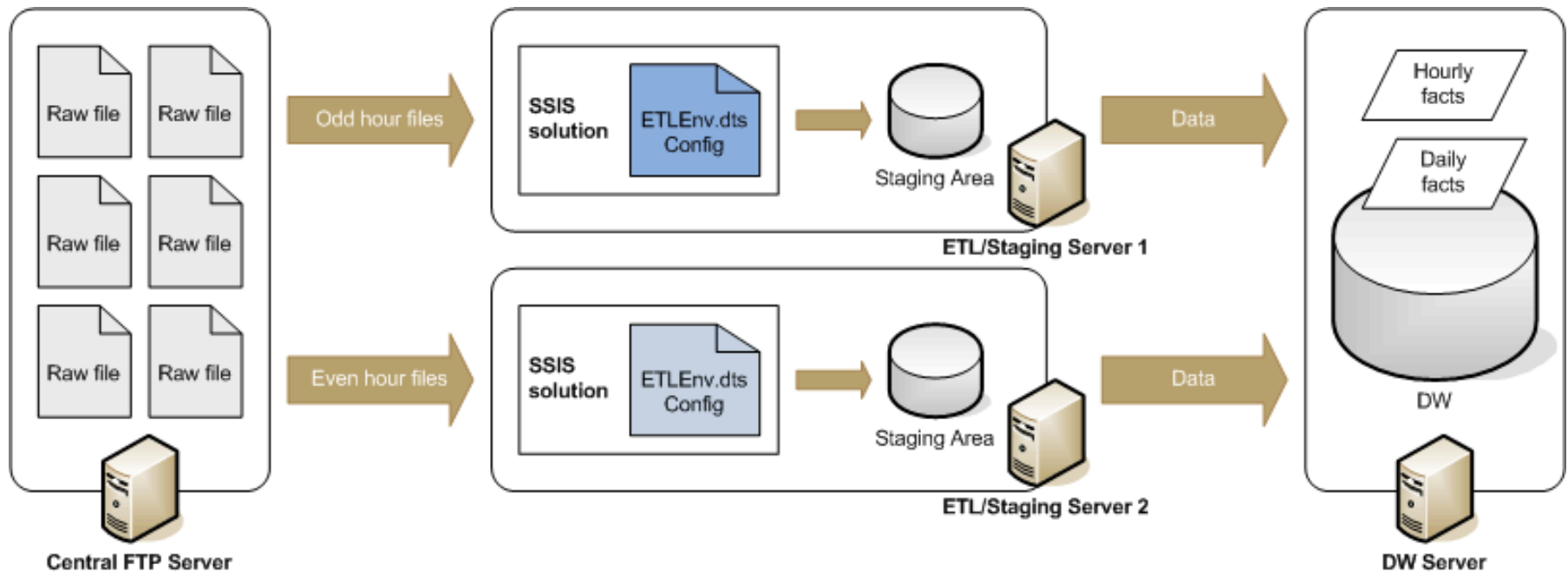
SQL SERVER INTEGRATION SERVICES (SSIS)



Challenges of Data Integration

- Multiple sources with different formats.
- Structured, semi-structured, and unstructured data.
- Data feeds from source systems arriving at different times.
- Huge data volumes.
- Data quality.
- Transforming the data into a format that is meaningful to business analysts.

ETL – extract, transform, load





SSIS

- BI Design Studio installed with SQL Server.
- One or more packages can be part of an Integration Services Project.
- Layers
 - *Connections to data sources.*
 - Data Flow
 - Control Flow
 - Event Handlers



SSIS tutorials

- Introduction to SQL Server 2008 Integration Services,
http://www.accelebrate.com/sql_training/ssis_2008_tutorial.htm
- Tutorial: Creating a Simple ETL Package
<http://msdn.microsoft.com/en-us/library/ms169917%28v=SQL.100%29.aspx>