Data Management

We Can Do Better

Miki Tebeka

353 SOLUTIONS

LEARN FROM THE EXPERTS

COPYRIGHT 1889 BY THE WERNER COMPANY

Sound familiar?

GIGO

Gartner surveyed a wide range of companies in its study and learned that **data quality costs them over $14 million dollars a year**.

https://blog.kissmetrics.com/bad-data-cost/

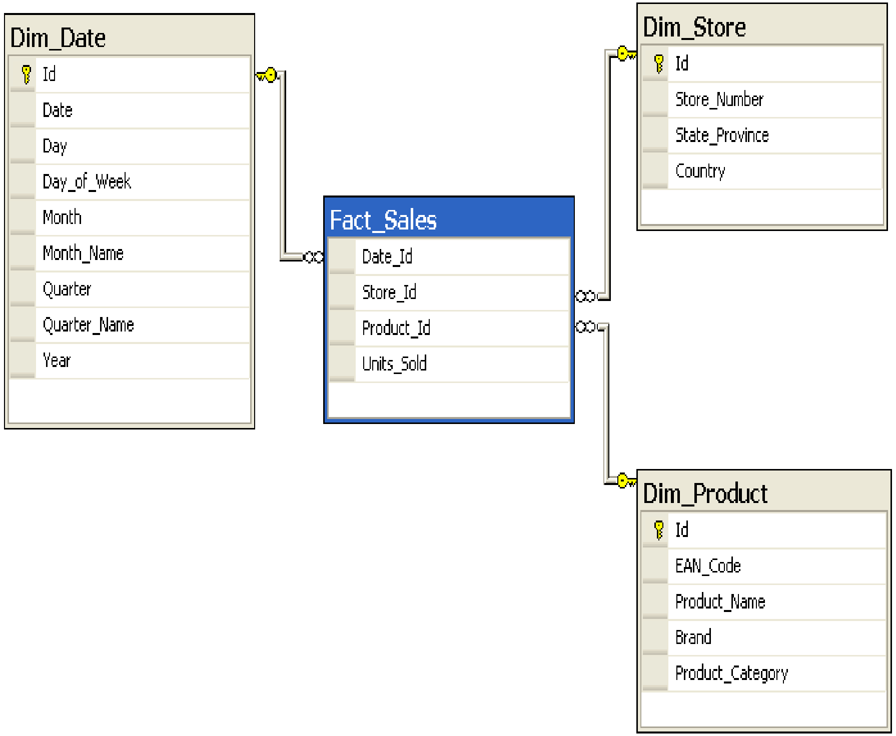
We’ll discuss data quality & organization

Spoiler …

I don’t have the answers :)

I’m hoping together
we’ll find some

Schema
(+ ontology)



Dim_Date
- Id
- Date
- Day
- Day_of_Week
- Month
- Month_Name
- Quarter
- Quarter_Name
- Year

Fact_Sales
- Date_Id
- Store_Id
- Product_Id
- Units_Sold

Dim_Store
- Id
- Store_Number
- State_Province
- Country

Dim_Product
- Id
- EAN_Code
- Product_Name
- Brand
- Product_Category

https://en.wikipedia.org/wiki/Star_schema

Make it explicit
(code, documentation...)

Schema is not enough

Example: NOAA

http://www.ncdc.noaa.gov/cdo-web

```
      DATE  SNOW  TMAX  TMIN  PGTM
2000-01-01     0   100    11  1337
2000-01-02     0   156    61  2313
2000-01-03     0   178   106   320
2000-01-04     0   156    78  1819
2000-01-05     0    83   -17   843
```

```
      DATE  SNOW  TMAX  TMIN  PGTM
2000-01-01     0   100    11  1337
2000-01-02     0   156    61  2313
2000-01-03     0   178   106   320
2000-01-04     0   156    78  1819
2000-01-05     0    83   -17   843


      time   int   int   int   int
      DATE  SNOW  TMAX  TMIN  PGTM
2000-01-01     0   100    11  1337
2000-01-02     0   156    61  2313
2000-01-03     0   178   106   320
2000-01-04     0   156    78  1819
2000-01-05     0    83   -17   843


                mm  c/10  c/10  HHMM
```

```
        DATE  SNOW  TMAX  TMIN  PGTM
  2000-01-01     0   100    11  1337
  2000-01-02     0   156    61  2313
  2000-01-03     0   178   106   320
  2000-01-04     0   156    78  1819
  2000-01-05     0    83   -17   843


  2000-07-16    12   312   245   937
```

https://www.flickr.com/photos/kaykim/3649886264

Data degradation is the gradual **corruption** of computer data due to an accumulation of non-critical failures in a data storage device. The phenomenon is also known as **data decay**, **data rot** or **bit rot**.

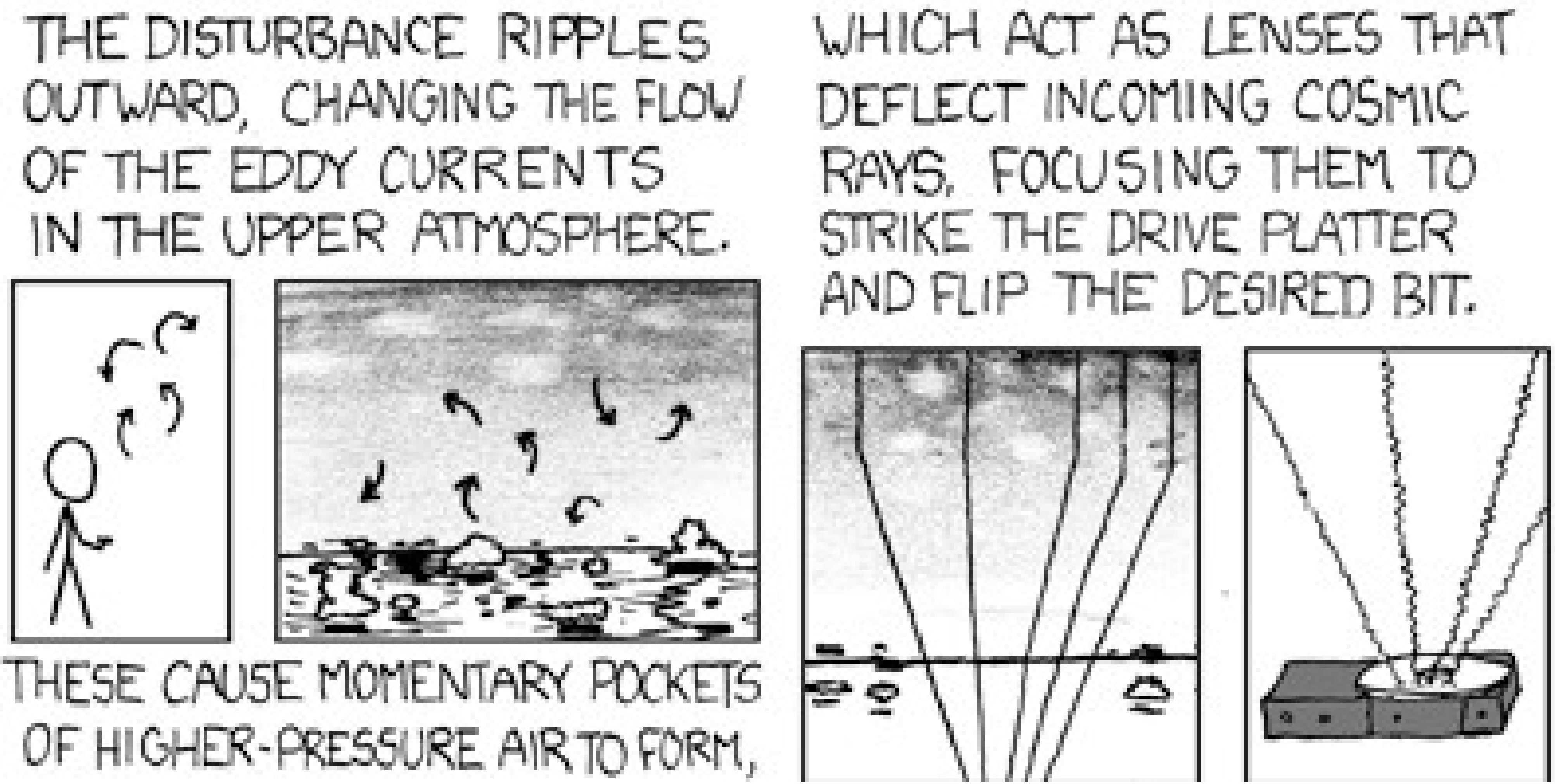
Studies by IBM in the 1990s suggest that computers typically experience about one cosmic-ray-induced error per 256 megabytes of RAM per month.

https://en.wikipedia.org/wiki/Cosmic_ray#Effect_on_electronics

This computer has 32GB of RAM

An error in about 3 hours

SSDs lose data on the shelf
(correlated to temperature)



THE DISTURBANCE RIPPLES OUTWARD, CHANGING THE FLOW OF THE EDDY CURRENTS IN THE UPPER ATMOSPHERE.
THESE CAUSE MOMENTARY POCKETS OF HIGHER-PRESSURE AIR TO FORM,
WHICH ACT AS LENSES THAT DEFLECT INCOMING COSMIC RAYS, FOCUSING THEM TO STRIKE THE DRIVE PLATTER AND FLIP THE DESIRED BIT.

https://xkcd.com/378/

Checksum, MD5, SHA256 …

Also metadata
(e.g. header with number of records)

Computed
Data

Remember the story I told about fixing the wrong code?

Do you know which version of which script generated the data you’re using?

Can you fix a single bad ETL?
A part of ETL?

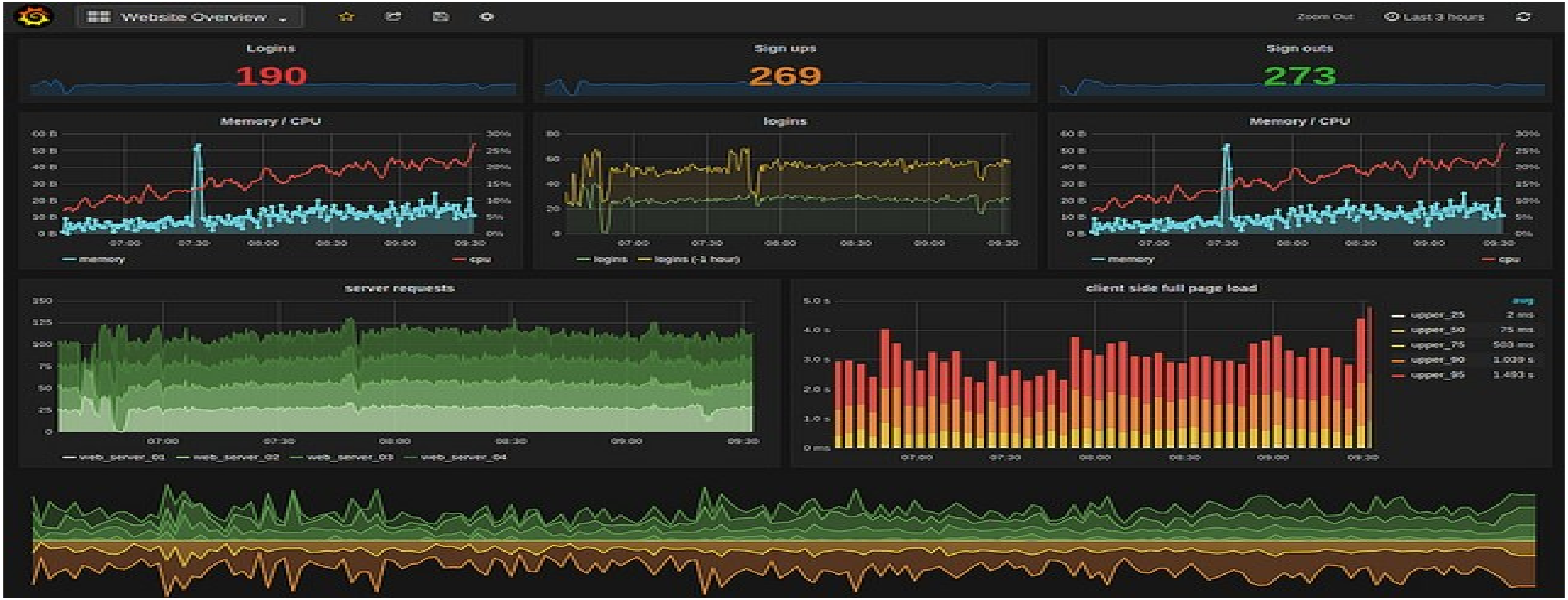
Will you remember to retrain your model after fixing the ETL?

Will you abort ETL on one error?

Will you abort ETL on 1,000 errors?

Do you allow manual editing?
Do you keep an audit trail?

Data KPIs



https://www.flickr.com/photos/xmodulo/24311604930

- Number of errors
- Difference from last ETL
- Anomaly detection (?)
- Number of records / day
- Per source of data
- ...

Slap monitoring & alerting on these KPIs

Recommendation

- Design your data
    - Ontology
    - Schema with units & validation
- Document ETL
    - Track execution history
- Data KPI Monitoring & Alerting

Discussion

- Process
- Tools
- Best practices
- War stories
- ...

Thank You

Miki Tebeka
@tebeka
miki@353solutions.com

353SOLUTIONS
LEARN FROM THE EXPERTS

References

- Pipeline debt
  - Great expectations
- DataFrame validation in Python
- What is Data Quality and How You Measure It for Best Results?