NoSQL: Real-Life Experience

Vít Bukač
Václav Lorenc

Agenda

- High-Level Data Processing
  - Engineering vs Analysis vs Presentation
  - Time-based DBs
- Data Onboarding
- User Experience
- Operational Issues

Data Processing

Data Processing

Engineering

Collection

Validation

Normalization

Transformations/Extractions

Enrichment

Production Servers

Analysis

Filtering

Processing/Querying

Modeling

Confirmation Bias

Presentation

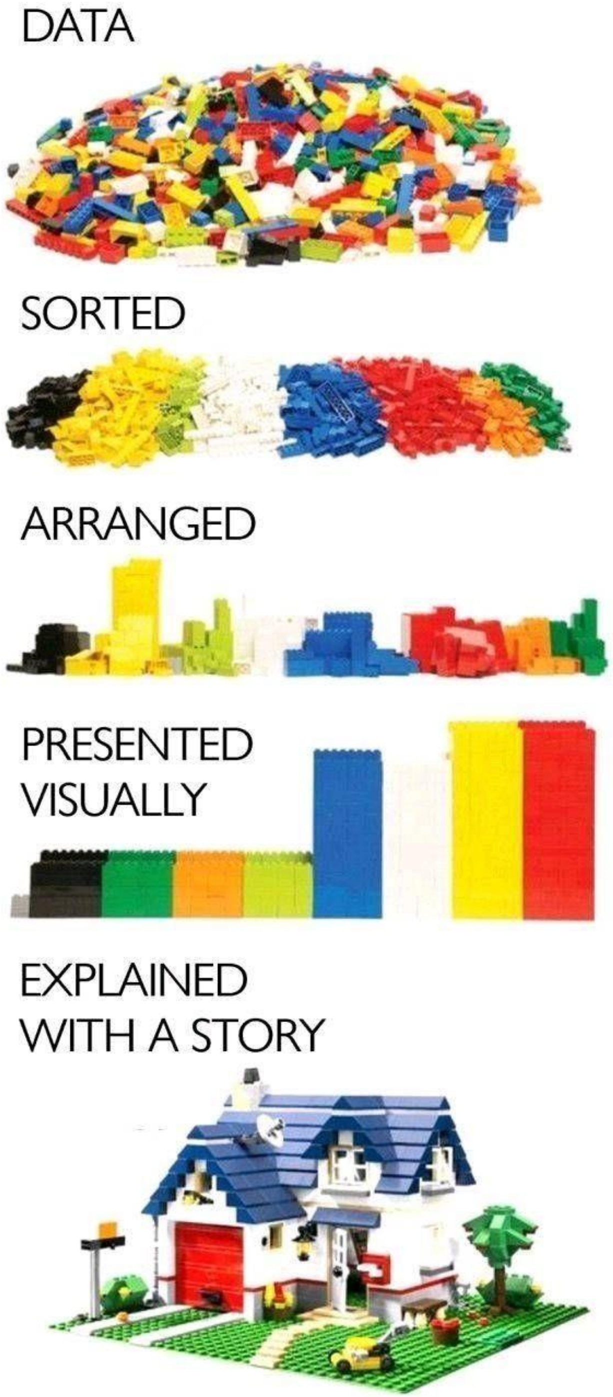
Visualization + Reporting

Storytelling

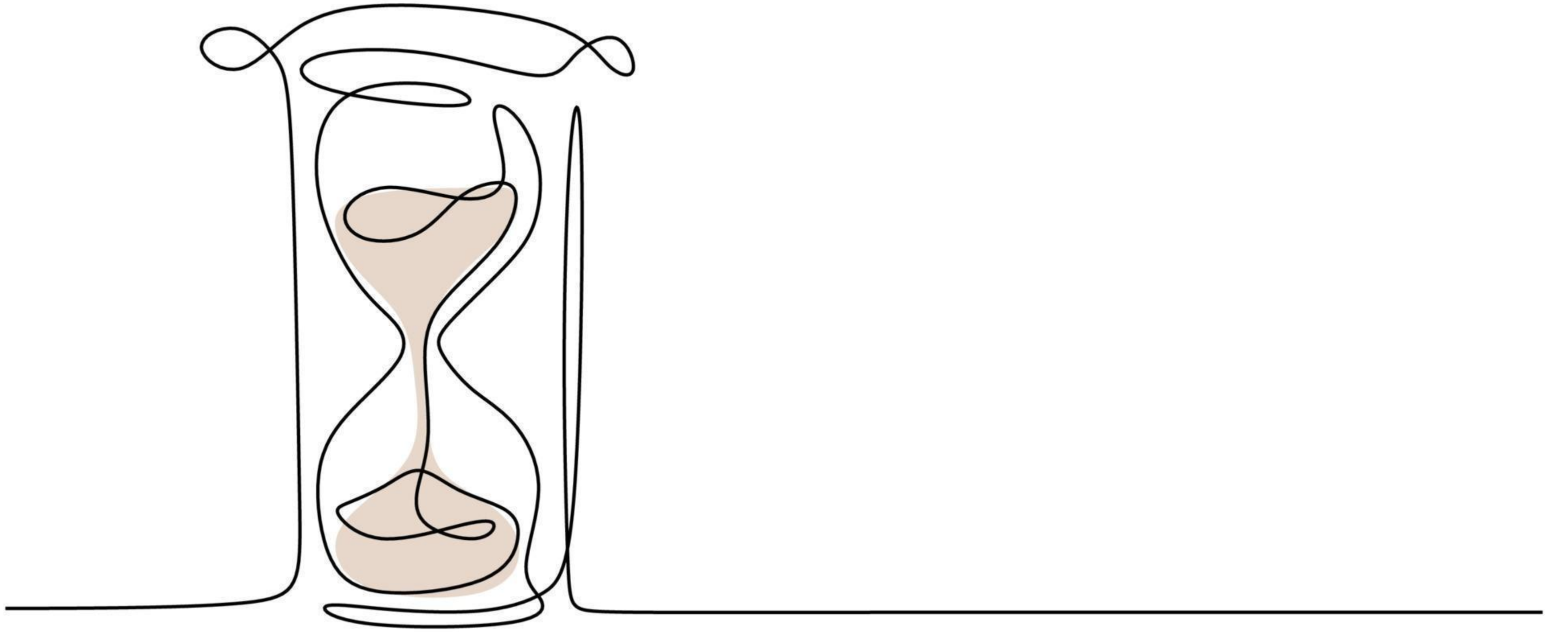
Psychology

Data Processing

"No numbers without stories,
no stories without numbers."
(a quote from internet)



Time-Series Databases



Time-Series Databases

- Time-oriented document databases
  - e.g. useful for security analyses, log aggregation
  - Better aggregation features
- Basic architecture
  - Buckets of data based by time-range
  - Accessing recent data efficiently
- Examples:
  - MongoDB, Splunk, ElasticSearch
  - ...

Data Onboarding

Data Onboarding

- Problem
  - People don't agree on one format
  - Input data sources may not be flexible
  - Multiple systems providing "same" data
- Data Normalization
  - Yeah, even in NoSQL world
  - Splunk Common Information Model (CIM)
  - Elastic Common Schema (ECS)
- Data Query Languages
  - Sigma query meta-language

Data Onboarding

- Timestamps(!)
  - No timezones? Bad timezones? Daylight saving time issues?
  - Ambiguous formats? (*05-02-2020)*
  - iso8601 is a recommended standard
- Value standardization
  - "yes/no", "true/false", "1/0", etc.
- Encoding
  - Separator characters - ,;|"'
  - CRLF vs LF
- Field names
  - Synonyms (`ip`, `ip_addr`, `src_ip`), capitalization (`DomainName`, `domainName` etc.)
  - And stability over time (fields renaming vs aliases)
- JSON/XML paths

Data Onboarding

- Monitoring over time
  - Format changes
  - Volume changes
- DB Return on Investment
  - SaaS? IaaS?
  - Budget
  - Prioritization of data sources
  - Data retention principles

Data Searching

Data Searching

- User Experience
  1. Fulltext & wildcard searches to familiarize with data type
     - Data sampling!
  2. Identify key fields, learn structure
  3. Use advanced searches, correlate with (other) sources, statistical searches
  4. Visualizations, scheduled searches
  5. Find and report anomalies

Data Searching

- Data Correlations
  - (Missing) Joins
  - "Transactions" (or grouped data)
    - Not in terms of atomic transactions
- Statistics
  - Count, # distinct values, list values
  - Statistics over time
- Anomalies, outliers, trendlines

Operational Challenges

Operational Challenges

- Healthchecks
  - Missing or duplicate data
- Backups
  - And restores :)
- Upgrades
- Mixed Technologies
  - SQL combined with NoSQL, legacy, ...
- Production Deployment
  - Security (MongoDB),
  - Reliability,
  - AAA: Authentication, Authorization, Auditing…

Operational/Legal Challenges

- Privacy
    - "Store everything" mentality
    - International, cross-region applications
    - GDPR
        - The right of access
        - The right to be forgotten
        - The right to restrict processing
    - Data security
        - Encryption at rest/in transit
        - Encryption of DB nodes, caching nodes, backups etc.
        - ACLs

(Lack of) Expertise

(Lack of) Expertise

- Unsuitable NoSQL Tech
  - Write-once, read-many-times
- Good-old SQL may work
  - Better than many custom SQL workarounds
- DB users won't be DB experts => communication problems
- Plus all the operational/architectural/legal issues

DB Incident Response

Incident Response in DB World

- GDPR
  - What actions are required?
- Audits, incident scope
  - Where are the data?
  - Was it encrypted? (in-transit, at-rest)
- Logs quality
  - Available?
  - Ready-to-use?
  - Useful?

Questions & Answers

Links & Resources

Resources

- (see notes, turn them into resources)

Placeholder

DB Incident Response