Adatfeldolgozás és – elemzés (Data processing and analysis)

Intelligens Elosztott Rendszerek
http://www.mit.bme.hu/oktatas/targyak/vimiac02

Budapest University of Technology and Economics
Fault Tolerant Systems Research Group
Témata
Outline

- Data collection
- Data processing
- ETL, workflow support
- Data format/representation
- Data storage

→ Data analysis

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data. - @BigDataBorat Twitter
Data science „process”

https://en.wikipedia.org/wiki/Data_science
HOW TO GET THE DATA?
OMG DDS Core notions

Introducing DDS™ – The Proven Data Connectivity Standard for IIoT™
Anatomy of a DDS Application

Domain
(id = 1, 2, ..., n)

Partition (e.g., Telemetry, Marked Data, Analytics, ...)

Topic Instances and Samples

Topic \( x \)

Topic \( y \)

Domain Participant

Topic

Publisher

Subscriber

DataWriter

DataReader
## Technological aspects

<table>
<thead>
<tr>
<th>Transport</th>
<th>Paradigm</th>
<th>Scope</th>
<th>Discovery</th>
<th>Content Awareness</th>
<th>Data Centricity</th>
<th>Security</th>
<th>Data Prioritisation</th>
<th>Fault Tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMQP</td>
<td>TCP/IP</td>
<td>Point-to-Point Message Exchange</td>
<td>D2D, D2C C2C</td>
<td>No</td>
<td>None</td>
<td>Encoding</td>
<td>TLS</td>
<td>None</td>
</tr>
<tr>
<td>CoAP</td>
<td>UDP/IP</td>
<td>Request/Reply (REST)</td>
<td>D2D</td>
<td>Yes</td>
<td>None</td>
<td>Encoding</td>
<td>DTLS</td>
<td>None</td>
</tr>
<tr>
<td>DDS</td>
<td>UDP/IP (unicast + mcast) TCP/IP</td>
<td>Publish/Subscribe Request/Reply</td>
<td>D2D, D2C C2C</td>
<td>Yes</td>
<td>Content-Based Routing, Queries</td>
<td>Encoding, Declaration</td>
<td>TLS, DTLS, DDS Security</td>
<td>Transport Priorities</td>
</tr>
<tr>
<td>MQTT</td>
<td>TCP/IP</td>
<td>Publish/Subscribe</td>
<td>D2C</td>
<td>No</td>
<td>None</td>
<td>Undefined</td>
<td>TLS</td>
<td>None</td>
</tr>
</tbody>
</table>

ETL

- „Extract-Transform-Load”
- Originally: to fill a snowflake/star schema
- In data science: create dataframes

Cleaning tasks
  - Standardization
  - Normalization
  - Deduplication
  - Enrichment
  - Clear/fill NAs
Example data processing workflow (KNIME)

Steps: reading, filtering/aggregation, transformation, plotting, ...

Status of the concrete execution
Measurement processing: RapidMiner

- **Read CSV**
- **Format conversion**
- **Identifying source node**

**Steps:**
- Filter to `cpu.usage.average`
- Calculating averages (interval)
- Add machine information
- Delete unnecessary attribute

**Diagram:**
- ReadMeasure... to ValueToNumb...
- ValueToNumb... to MachineName...
- MachineName... to FilterByMetricId
- FilterByMetricId to AvgValue
- AvgValue to AddMachineInfo...
- AddMachineInfo... to DropTimestamps...
DATA FORMAT
Tidy data

- 3 Simple rules to facilitate statistics and visualization
- One variable – one column
- One observation – one row
- Each type of observational unit – one table
- ... seems to be trivial
- ... not true in most practical cases
- ... and even for statistical tools (e.g. output of R packages)

Data originally: long/wide

https://en.wikipedia.org/wiki/Wide_and_narrow_data
How to use these formats?

Sparse Screening for Exact Data Reduction. Jieping Ye, Arizona State University
Examples for tidy data

R dataframe representation:

http://garrettgman.github.io/tidying/
```
R: spread(data,key,value)
```
„tidying”

R: spread(data,key,value)

table4

country | 1999 | 2000  
--- | --- | ---  
Afghanistan | 745 | 2666  
Brazil | 37737 | 80488  
China | 212258 | 213766  

table4

country | 1999 | 2000  
--- | --- | ---  
Afghanistan | 19987071 | 20595360  
Brazil | 172006362 | 174504898  
China | 1272915272 | 1280428583  

table5

Generalization?

http://garrettgman.github.io/tidying/
Data restructuring examples (in R)

With Aggregation

- cast(md, id=variable, mean)

<table>
<thead>
<tr>
<th>ID</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>5.5</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>2.5</td>
</tr>
</tbody>
</table>

- cast(md, time=variable, mean)

<table>
<thead>
<tr>
<th>Time</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.5</td>
<td>3.5</td>
</tr>
<tr>
<td>2</td>
<td>2.5</td>
<td>4.5</td>
</tr>
</tbody>
</table>

- cast(md, id-time, mean)

<table>
<thead>
<tr>
<th>ID</th>
<th>Time1</th>
<th>Time2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.5</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>3.5</td>
<td>3</td>
</tr>
</tbody>
</table>

Without Aggregation

- cast(md, id+time~variable)

<table>
<thead>
<tr>
<th>ID</th>
<th>Time</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

- cast(md, id+variable~time)

<table>
<thead>
<tr>
<th>ID</th>
<th>Variable</th>
<th>Time1</th>
<th>Time2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X1</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>X2</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>X1</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>X2</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

- cast(md, id=variable+time)

<table>
<thead>
<tr>
<th>ID</th>
<th>X1</th>
<th>X1 Time1</th>
<th>X1 Time2</th>
<th>X2</th>
<th>X2 Time1</th>
<th>X2 Time2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

DATA STORAGE
## Reminder: Tabular Representation

- **Rows of the table** = Model elements
- **Columns of the table** = Properties

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Size (kB)</th>
<th>Last modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents</td>
<td>directory</td>
<td></td>
<td>2016.02.02</td>
</tr>
<tr>
<td>Contracts.pdf</td>
<td>file</td>
<td>569</td>
<td>2015.11.09</td>
</tr>
<tr>
<td>Pictures</td>
<td>directory</td>
<td></td>
<td>2016.02.02</td>
</tr>
<tr>
<td>Logo.png</td>
<td>file</td>
<td>92</td>
<td>2015.03.06</td>
</tr>
<tr>
<td>Groundplot.jpg</td>
<td>file</td>
<td>1226</td>
<td>2016.02.02</td>
</tr>
</tbody>
</table>

- Data analysis languages (e.g. R, Python): **dataframe**
  - One row: one measurement/observation
  - Columns have their own **Types**
Common data storage techniques

- .CSV
  - Majority of inputs
  - Length? Header? Encoding?
- DB with a schema (in memory?)
- Graph databases, ontologies, RDF...
- Key-value stores (redis)
- Time series databases (openTSDB, influxDB)
  - Time series + metadata
- „Data in motion”
  - Streams as input for processing/analysis
**Time series example: influxDB**

- **Data:** measurement
  - Fields, tags, timestamp

<table>
<thead>
<tr>
<th>AGGREGATIONS</th>
<th>SELECTORS</th>
<th>TRANSFORMATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>COUNT()</td>
<td>BOTTOM()</td>
<td>CEILING()</td>
</tr>
<tr>
<td>DISTINCT()</td>
<td>FIRST()</td>
<td>DERIVATIVE()</td>
</tr>
<tr>
<td>INTEGRAL()</td>
<td>LAST()</td>
<td>DIFFERENCE()</td>
</tr>
<tr>
<td>MEAN()</td>
<td>MAX()</td>
<td>FLOOR()</td>
</tr>
<tr>
<td>MEDIAN()</td>
<td>MIN()</td>
<td>HISTOGRAM()</td>
</tr>
<tr>
<td>SPREAD()</td>
<td>PERCENTILE()</td>
<td>NON_NEGATIVE_DERIVATIVE()</td>
</tr>
<tr>
<td>SUM()</td>
<td>TOP()</td>
<td>STDDEV()</td>
</tr>
</tbody>
</table>
Dashboards... (e.g. Grafana)

https://grafana.com/dashboards/1443
Data mining „brickstones‟

- Clustering
- Classification
- Association rules
- Regression
Clustering

Estimated number of clusters: 3
Association rules

MARKET BASKET ANALYSIS

98% of people who purchased items A and B also purchased item C
Classification

- is sex male?
  - yes
  - is age > 9.5?
    - died (0.17, 61%)
  - no
    - survived (0.73, 36%)
    - is sibsp > 2.5?
      - died (0.05, 2%)
      - survived (0.89, 2%)
THE SYSTEMATIC WAY:
EXPLORATORY DATA ANALYSIS

Look and see
„About half of our sensory neurons are dedicated to vision, endowing us with a remarkable pattern-recognition ability.”

Prof. Alfred Inselberg
Exploratory data analysis (EDA)

- Summary of the **main** characteristics of a data set
  - Identification of **outliers, trends, other patterns**
  - Often with **visual** methods.
  - A statistical model can be used or not,

- „For seeing what the data can tell”
  - beyond formal modeling or hypothesis testing
  - **hypotheses → new data** collection and experiments
Approach - visual exploratory analytics

Resources

- 120.000.000 sensors
- $10^{10}$ processors

Process based on interactivity

1. **Graphical presentation** of the data
   - multiple diagrams
2. **Visual evaluation**
   - exploiting human overview
3. **Visual selection, manipulation** – multiple diagrams
4. **Interpretation, correlation with other models, evaluation** (like architecture etc.)
Visualisation in Everyday Life

Trend Analysis and Forecast

Correlation Analysis

Time Series Analysis

Analysis of Spatial Data
Additional knowledge

- „People buying coffee often buy milk”
- „There is a significant difference in salaries depending on gender”
- „The memory consumption of a software grows exponentially wrt. number of requests in queue”.
- „The population follows a $N(100, 15)$ distribution”
- „BME students fall into 3 main different groups (according to their grades)”
Data analysis

- Modell
- Data
- Cleaning
  - Exploratory
  - Confirmatory

Additional knowledge
Data analysis

Exploratory analysis
- **Goal:** formulate hypotheses
- Know the data/domain
- Highly ad-hoc
- Mainly descriptive statistics + data mining + visualization

Confirmatory analysis
- **Goal:** test hypotheses
- Validate
- Mainly statistical tests + statistical inference
E.g. distribution analysis

Exploratory

Hypothesis: variable $x$ follows normal distribution

Confirmatory

Variable $x$ follows $N(12, 4)$ distribution
Exploratory Data Analysis

- Goal: hypothesis formulation
- Pattern recognition
- Early validation
- „Sensors of type X are sensitive to high temperature”
- „Application of Type Y is sensitive to CPU load”
- Interactive, human expert needed
- Later: automated support (IBM Watson Analytics)
Validation for automated methods

For all cases:

- **Means:**
  \[ M[x] = 9 \]
  \[ M[y] \sim 7.5 \]

- **Variance:**
  \[ \sigma[x] = 11 \]
  \[ \sigma[y] \sim 4.12 \]

- **Correlation:**
  \[ C(x,y) \sim 0.816 \]

- **Regression:**
  \[ y \sim 3 + 0.5x \]

Avoid false assumptions
Use human intuition (1973)
... and some more

https://www.autodeskresearch.com/publications/samestats
Distribution vs summary

Same summary != same story

https://www.autodeskresearch.com/publications/samestats
VISUAL EDA EXAMPLES
EDA example

Imre Kocsis: Measurement-Driven Resilience Design of Cloud-Based Cyber-Physical Systems
EDA example2
EDA example 3: pairwise correlation

- **Variables**
  - Blue: positive
  - Red: negative
  - Strong color: strong correlation

- **Correlating variables**
- **Independent variables**

**R „corrgram package”**

**Pearson linear correlation**

Scatterplot matrix

Goal: identify correlating variables, outliers
- Dimension reduction
- Feature selection
Example: Visualisation of State Spaces

https://www3.hhu.de/stups/prob/index.php/State_space_visualization_examples
Example: State Space of the CAN Bus

https://www3.hhu.de/stups/prob/index.php/State_space_visualization_examples
Parallel Coordinates

- Multi-dimensional visualization
- Compact, scalable
- Axis order?

```
[2, 4, 4, 5, 6]
[3, 6, 6, 4, 5]
[1, 2, 2, 3, 3]
```

```
[3, 4, 6, 5, 6]
[2, 5, 4, 6, 4]
[1, 3, 2, 3, 2]
```
Parallel Coordinates: Analysis of the Test Cases

1 test case: 1 broken line

The variables appear on the $x$-axis
Parallel Coordinates: Analysis of the Test Cases

The ones detecting an error did not even come to the actual computation.

Timeout?

Run time and memory usage seem to be in a positive relation (if the test is successful)
Parallel Coordinates: the Alternatives
SOME BACKGROUND... OLAP
On-Line Analytical Processing

- Business intelligence approach
- Extensively used since early 2000s
  - Still! (although not that popular as it was – at least in academic research)
- Features
  - Multi-dimensional analysis
  - Fast query execution
  - Exploratory analysis of data
    - Support ad-hoc queries
  - Report generation
  - (Visualization)
Central concept: OLAP cube

- Multi-dimensional array:
  - set of separate data
    - Dimensionality >3
    - technically a hypercube
    - ~ a multi-dimensional spreadsheet

- Slicer: dimension held constant
  - For a given query (e.g. sales in a particular year)
OLAP process (img: Pranav Joshi)

Data Warehouse

OLAP Server

End User Tools

ACCESS

- Analytics Applications (Business Objects, Cognos)
- Business Reporting Applications (Dashboard Manager)
- Data Mining (SQL Server Mining, SAS Mining)
- Business Modeling

Physical Multi-Dimensional Cubes

Load

Access of Detailed Data

Access
OLAP operations

- Operations
  - Slicing & dicing
  - Drill up & down
  - Pivoting

- Easy to visualize by the cube itself
Dicing (img: Wikipedia)
Drill up & down

(img: Wikipedia)
OLAP vs. “regular/modern” data analysis

- **OLAP cube**: like a set of spreadsheets
  - multi-dimensional
  - interlinked

- **Modern data analysis**: “flat” data frames
  - Modern machine learning algorithms:
    - require (?) single dataframes

- **Operations**: basically the same (slicing, dicing, drill up & down, pivoting)
DATA VISUALIZATION GOALS
Aims of visualization

- Assumption formulation
- Communication of results
- Share research in a reproducible manner
- Give a tool to browse the data
- Create a high level report
- Evaluate a system/solution
- Tell a story
- …
Visualization types

- Reporting
- EDA
- Data tour
- Reproducible research
- Dashboard
- Self-Service BI
- Storytelling
- Data journalism
- Infographics
Cloud-based services

- Telecommunication services in the cloud
- Application MUST work...
  - What to measure? Under which conditions?
- Migration, fault tolerance
  - ...under environmental changes
- Is the system „UP”?
  - Critical services should work
- Rare events
Example: System Model $\rightarrow$ Performance Model

- **Modelling**
  - System model
  - Qualitative model
  - Performance model

- **Design of the Analysis**
  - Expectations
  - Parameters to measure
  - Design of the experiments

- **Data Collection**
  - Measurements
  - Benchmarks
  - Simulation

- **Analysis**
  - Exploratory Analysis
  - Hypothesis testing
Example: performance analysis

Outliers/background operations

Deviations?

Linearity

Unpredictable behavior
Recent research

Similarity?

Feedback?

Coverage?

Iterative refinement?

HA → coverage of rare phenomena?

Benchmarked object

Measurement results

Cluster classes

QR model

Evaluation

Reusable, parametrizable

Statistical / business (call flow) / platform (action-driven)

Data acquisition

Effect of change?

Representative set?

Representative metrics

Selection:

Business/platform QoS-SLA
Minimal variable set
Proactive control support
Technical measurability

Technical measurability

Monitoring and mitigation

Model check (variable selection)

Model recognition

Resolution

Xtract

Resource

Synthetic/realistic?

Workload

Goals
MEASUREMENT SETUP
Experimental environment

- Dedicated cores
- Dedicated memory
- Capped IOPS and packet rate
Collecting metrics

SERVICE + COMPONENTS + OS + HYPERVISOR

SIPp
Proxy
Router
Auth DB
MMTel DB
ESXi

Monitoring
Types of gathered metrics

SIPp
- Ongoing calls
- Successful calls
- Call failure rate
- Op. latencies

Component SNMP
- rejects (overload)
- Latency stats
- # responses
- # connections
- Latency stats

Proxy
- OS metrics
- VM + Hypervisor metrics

- SNMP: ~ same across components
- Sampling: 1 sec
- Temporal resolution: 5 Sec
Experiment design space
Workload profile

- Can be any possible profile, but generally:
  - Increasing load
  - Plateau (stable plateau)
  - Decreasing load
  - Cooling down
- Plus optional additive noise (sinus, burst)

Representativeness? ... could be any telco-specific
Background:
Motivation

Enterprise cloud
Purchased CPU time

Private university cloud

Lab
Our VCL cloud

- Maintained by our research group
- 5 semesters
  - 2 courses/semester
- 9 hosts
- ~20 000 reservations
  - Only 22 rejected
Reservation Workflow in VCL

- Request
  - VM type
  - Length
  - Immediately or later

- Hard reservation limit

Load time
Capacity Planning

Can I start solving my homework now?

Do we have spare capacity for my research next week?

I am responsible for a course with 250 students next September. Can we handle this workload?
Capacity Planning

Support for hard limit estimation

Spare capacity prediction

Long-term capacity planning/scheduling
The Available Dataset

deadlines, #students

reservation type, time to load, etc.

cpu usage, memory usage, etc.

cpu usage, memory usage, etc.
Data Analysis Steps

Workload prediction

Resource util. pred.
Workflow

Capacity planning

Use case 1
- VM
- VM
- ?

Host

Use case 2
- VM
- VM
- VM

?

Use case 3

?
Workload prediction

Lab 1

Reservation count

02/25 Tue | 02/27 Thu | 03/01 Sat | 03/03 Mon | 03/05 Wed | 03/07 Fri

Course 1 (HW 1)

Reservation count

04/03 Thu | 04/05 Sat | 04/07 Mon | 04/09 Wed | 04/11 Fri | 04/13 Sun | 04/15 Tue

Course 2 (HW 1)

Reservation count

11/10 Sun | 11/12 Tue | 11/14 Thu | 11/16 Sat | 11/18 Mon

Deadline
Workload prediction

- Daily workload follows a Gaussian-like distribution
Model fitting
Workload prediction

- Daily workload follows a Gaussian-like distribution
- Exponential increase in peak numbers
- Maximum location between 7 PM and 11 PM
- ~4 hours as standard deviation
Workload prediction

Students work even in the night

They have lunch and dinner
Workload prediction

Changes in students’ behavior?
Resource Utilization Prediction

\[ f(\text{workload})?? \]

System workload

CPU utilization, %
Challenges

- It is a cloud
  - Statistical multiplexing 😊

---

2012/2013/2  2013/2014/1  2013/2014/2  2014/2015/1
Challenges

- It is a cloud
- Hosts show different behavior
  - Warm spare
  - Different user behavior
  - ???
Resource utilization analysis: memory

- Linear model
  - $\text{Mem}(VM_1) + \text{Mem}(VM_2) + \ldots + \text{Mem}(mgmt)$
  - Weighted by the workload

Very good at following drastic changes
Within 5% by the 97% of time
Resource utilization analysis: memory

- Linear model
  - $Mem(VM_1) + Mem(VM_2) + ... + Mem(mgmt)$
  - Weighted by the workload
Resource utilization analysis: CPU

- Linear model
  - \( CPU(VM_1) + CPU(VM_2) + \ldots + CPU(mgmt) \)
  - Weighted by the workload

CPU is much more sensitive than memory
Resource utilization analysis: CPU

The students use the CPU more intensively before the deadline.
Resource utilization analysis: CPU

- Linear model
  - $CPU(VM_1, wl) + CPU(VM_2, wl) + ...$
  - Weighted by the workload
Summary

- Data-driven static capacity planning
  - „user behavior” analysis
  - resource fingerprint estimation

- Conclusions:
  - student behavior can be modelled
  - resource allocation were sometimes (too) strict

- Dynamic capacity planning?
  - Long loading time → failed reservations soon
  - When to burst out to a public cloud?
References

- Garrett Grolemund, Hadley Wickham. „R for Data Science” (O’Reilly, 2017) http://r4ds.had.co.nz/