



15TH BUSTECH 2025




**DATA ANALYTICS:
HOW CAN WE GET IT RIGHT?**



ComputationWorld 2025 & DataSys 2025
April 6-10, 2025
Valencia, Spain

Joseph G Vella, Faculty of ICT, UM



**L-Università
ta' Malta**

1



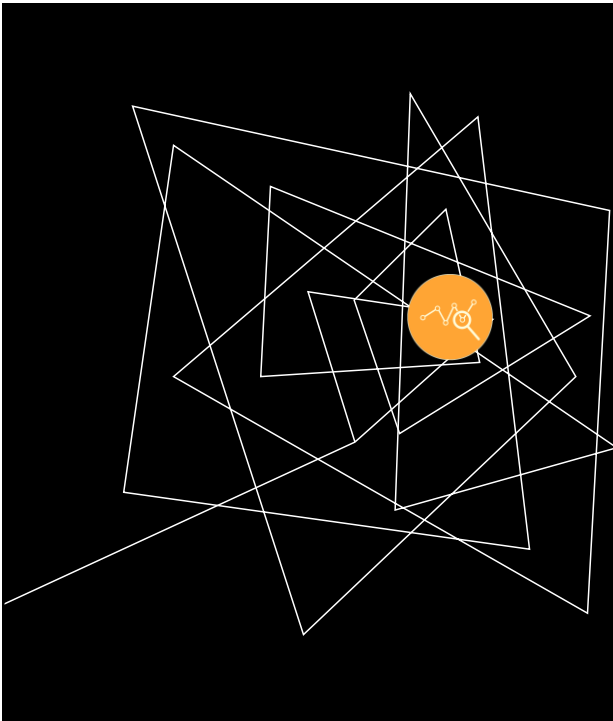
ABOUT



Prof Joseph G Vella is currently an academic member at the University of Malta, emphasising teaching and researching data modelling and DBMS-related technologies. Prof Vella has introduced many teaching units for undergraduate and postgraduate curricula (e.g. Data mining and Warehousing and Digital Forensics). During the same time, Joseph supervised scores of undergraduate projects and a growing number of postgraduate students, authored numerous refereed papers, edited manuscripts, and is asked to review conference and journal submissions.

Prof Vella is involved in several research projects with other academic institutions and industries. The main role in these projects is to apply his expertise in data management to integrate data better and address data quality issues. Another contribution is to advise companies to ensure adequate performance for data-intensive applications running on data servers that are either on-premises or cloud-based.

2



RESEARCH INTERESTS

- Computer Information Systems
- Data and Query Modelling
- DBMS Set-up and Tuning
- Digital Forensics
- Data Science and Analytics

3



DATA ANALYTICS: HOW CAN WE GET IT RIGHT? KEYNOTE'S STRUCTURE

Agenda

History

- Where we are, and from where we came from

Data Analytics

- Definitions, practices, and opportunities

Themes and Variations

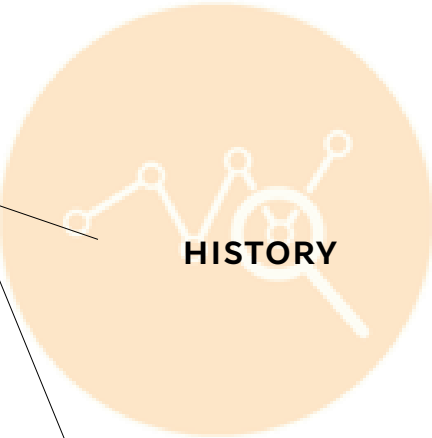
- Two themes for considerations

Take Home

- On team composition, targets, technical savviness, and trade-offs

Data Analytics: How can we get it right? 4

4




HISTORY

Where we are, and from where we came from

Data Analytics: How can we get it right? 5

5



HISTORY OF SIGNIFICANT MILESTONES (BIASED)

Tukey's contributions:
The Future of Data Analysis (1962), Exploratory Data Analysis (1977)

The 1970s DB community

Codd (RDBs, 1970), **Chen** (Conceptual Design, 1977), **Gray** (ACID, 1981 & early work on "massive" datasets use for innovative data processing and analysis), **ISO's** (SQL, 1986)

The 1970s, 1980s and 1990s DB and DWH community


Inmon (Ideas early 1970s, The Book, 1993++), **Kimball** (Dimensional Model and Data Architecture for a DWH, 1996), **Gray** (Data Cube, 1996), **Fayyad & Piatetsky-Shapiro** (data mining conference, 1997, Knowledge Discovery in Databases (KDD) Journal started 2005), **many** contributions to DWH, OLAP, and DM

Gray (Five-minute rule, 1985/1997 & 2007), **Mohan** (DB Recovery and Isolation, 1992), Reed (MVCC, 1978), **Bayer & McCreight** (B-tree, 1970), **Zhuge** (1995) and **Gupta** (1995) View Materialisation in DWH

Great technical strides in data connectivity, data portability, transaction monitors (distributed TP), data and query modelling, web enabled technology

Data Analytics: How can we get it right? 6

6



HISTORY OF SIGNIFICANT MILESTONES (BIASED) CONTINUED

The 2000s DWH & DM

Tracking of ETL processes through a workflow adorned with source data dependences.
Specialise a DWH scope according to usage demand and adopt ad hoc ELT processes.
Data Lakes – keep it raw (2011), and Data Fabrics – end2end integration (cloud) (2017).

The 2000s Machine Learning & AI (2000s)


Python's (**van Rossum**, 1991) resurgence since 2003
Extensive libraries available – e.g., Scikit-Learn (**Cournapeau**, 2007).
GPUs and Deep learning CNN came together (e.g., an early trend setter was DanNet (**Cirensen**, 2011) visual pattern recognition).

The 2000s DB

ISO's SQLIII standard issue (1999) – various updates done.
Mainstream DBMS – different transaction models (Serializable, Causal – see **Ahamad**, 1995 + **Lampport**, 1978, Eventual consistency – **Vogels** 2008), sophisticated data indexing. Main Memory DBMS – VoltDB (changed name a some months ago).
CAP theorem by **Brewer** 2000
NoSQL movement with non-SQL rationale (data placement regimes, newer/revamped data models – but niche applicability). Later, pitch changed to “Not Ony SQL”

Data Analytics: How can we get it right? 7

7



PAIN POINTS (AN INCOMPLETE LIST)

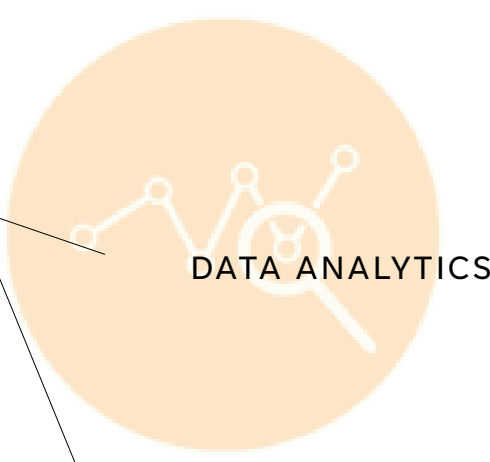
DBMS **Cost** of application development, cost of DBMS (and related software), cost of hardware (+ networking), and operational costs are significant. **Expertise** is in limited supply.
Issues with meeting **scalability** and **availability** requirements.
Security exposure risk, e.g., application centric data security – “Who can see what and when?”

DWH Very **expensive** endeavour, and requires your **best team** and resources!, no guarantee of breakeven.
Very hard **adaptability**, e.g., the “opaque” ETL processes are very susceptible to “breaking”!
This also exuberates the “**data silo**” syndrome.
Disruptive technology, e.g., big data's push, have added complications rather than tenable and sustainable progress.
Almost real-time analysis is limited by the latest ETL process, i.e., sometimes hours.
Little aggregating of “older data” – e.g., convert sales entries of five years prior, into daily tallies.

Query Processing **Performance issues** – query response time might be longer than indicated/expected/required. This is due to: data volume, complexity of query, data sources might be dipartite (different instances).
Metadata on the provenance of the data is limited and confusing, making adoption untenable.

Data Analytics: How can we get it right? 8

8




DATA ANALYTICS

Definitions, practices, and opportunities

Data Analytics: How can we get it right? 9

9



WHAT IS DATA ANALYTICS?

“**data analytics** is the act of examining datasets to extract value and find answers to specific questions” G Mathur (Mat23)

The **extraction** from an enterprise data to valuable indications is undertaken, this this session, by computer information systems that **integrate, consolidate, and transform, select/restrict, recode** data.

The **transformed data model** is then used as a basis on which questions and pattern matching requests are unleashed on.

Enterprises that adopt **data analytics** have the following **expectations**:


- Management, and tactical **decisions are data driven**.
- Making and reaching solutions faster**, than your competitors, is an advantage.
- The **actualisation** of a solution is faster.
- Good decisions lead to **favourable financial results**.

Which enterprises can benefit **data analytics**?

Large, small enterprises! Any area.

Data Analytics: How can we get it right? 10

10




WHAT CAN DATA ANALYTICS DO?

Data Analytics can help address an enterprise **tactical** and **strategic** directions:

- DESCRIPTIVE** **What just happened?**
The right information is on tap, easily gotten to, reflects the very latest, it is accurate, can easily be rehashed into various mediums (e.g., just the right visualisation).
- DIAGNOSTIC** **Why did that happen?**
We map a situation into a “clause and effect” model. Also, the clause and effect are elaborated with more details (ability to dig deeper).
Is it likely that this situation creates a cascade of similar situations?
- PREDICTIVE** **What and when will a target happen?**
Are the models generated to reach a target, doable?
Which data (and patterns) were used to build the target? Are these reasonable, relevant?
What are models exactly predicting? (The answer helps to monitor progress)

Data Analytics: How can we get it right? 11

11




WHAT CAN DATA ANALYTICS DO? (CONTINUED)

Data Analytics can help address an enterprise **tactical** and **strategic** directions:

- PRESCRIPTIVE** **How can we reach a set target?**
Identify target and actionable jobs/processes and recommendations to reach it.
Identify the team best fit to execute these jobs. It is best to supplement the target with scaled down versions of the project, to ensure the target implementation can scale out in full and with more confidence.
For recommendations, other than a lead team, we need milestones to finely tune whilst addressing and implementing these recommendations!

Data Analytics: How can we get it right? 12

12




A DATA ANALYTICS PROJECT METHODOLOGY

How is a data analytics project **executed**?

- ✓ **Set out** the requirements
- ✓ **Assemble** an eclectic team, and prove the requirements
- ✓ **Refine and focus** the requirements into focal areas (e.g., data requirements)
- ✓ **Survey** the enterprise data infrastructure for sources (and identify external ones)
- ✓ **Start data exploratory analysis** and shift to building with basic models
- ✓ **Elaborate** models and evaluate against the set of requirements
- ✓ **Fine tune model** and ensure adequate visualisation and dissemination material is at hand
- ✓ **Revert results/models** to the data owners for their evaluation and analysis of how to make the model relevant to their enterprise goal's

Data Analytics: How can we get it right? 13

13




WHAT IS THE PAIN?

The **extraction process** context is **complicated** by various combinations of:

- Voluminous data sources** (including external datasets we have little control on accept access rights)
- Variety of data formats** (e.g., relational, graph, tree)
- Data Arrival frequency** and/or rate
- Departure frequency**, rate, response time (as in, how long to compute), based on which data (e.g., last week's, yesterday's, few moments before)
- Data quality assurance** is nether complete nor sound; + source unclear

Data Analytics: How can we get it right? 14

14



DATA THROUGHPUT, TRANSACTION RATE, FRESHNESS


Data throughput is the rate a DBMS can sustain READs and WRITEs
Specifically, the **Operations Per Second (OPS)** is often quoted.
Also used to describe the number of atomic transaction executed – we use transaction throughput in that case.

Latency is the time required to serve a request
This applies to READ, WRITE, or TRANSACTION requests.
It is related to throughput, but it measures queue time too.

Transaction rate per second (TPS) is logical work per second
We presume that these transactions are abiding by ACID principles.
TPS depends on data throughput, but also on **data contention** from user's data requests. I.e., no data contention between users, the quicker is the TP scheduling.
If **data is distributed**, and ACID is still required, then **synchronisation** is required between the distributed copies of the data. Expect the TPS rates to become an issue.

Data Analytics: How can we get it right? 15

15



DATA THROUGHPUT, TRANSACTION RATE, FRESHNESS (CONTINUED)


Data Freshness measures the time lag a committed event is visible
For example, a client submitted an order in an OLTP, when is his order visible for OLAP? From the OLAP perspective, once the order is accepted by the OLTP, the client's data in the OLAP is **stale**, until it is refreshed with the new update.

This measure is becoming very important when an OLTP requires a verification from the OLAP; e.g., The order processing system ask the data analytics system, "is this order a fraud?"

Decision is data driven, on fresh data, and with very quick latency
Remember,
some decisions are affected by **timeliness**;
increasing the **dwell time** for a web user correlates with losing the sale!

Data Analytics: How can we get it right? 16

16



DATA SOURCES CHARACTERISTICS FOR DATA ANALYTICS

ONLINE TRANSACTION PROCESSING (OLTP)

Typical based on a **relational model**

Continuously arriving, **short lived** and **ACID** transactions

Updates involve **limited data scope** (compared to total storage)

Database Activity Latency sensitive

Elastic scalability for off-prem set-ups


Tight budget.

DBMS is asked to support a spectrum of **workloads: from mixed READ/WRITE to WRITE heavy**

Typically, an enterprise has a **few OLTPs** (related to its various business processes)

Data Analytics: How can we get it right? 17

17



DATA SOURCES CHARACTERISTICS FOR DATA ANALYTICS

ONLINE ANALYTIC PROCESSING (OLAP)

Typically a **cube data model** (multi dimensional data structure)

Few updates, and these do not occur continuously

? In the last decade, this is becoming less so:

1. Adoption of ELT instead of ETL;
2. Data freshness requirements.

Throughput sensitive

Scalability is expected


Processing requires **large input and output data** chunks

DBMS is asked to **support heavy read workload** (including serial data traversal) with very demanding latency

The **budget** is a concern, but enterprises are usually amiable to increasing support.

Data Analytics: How can we get it right? 18

18




DATA SOURCES CHARACTERISTICS FOR DATA ANALYTICS

RAW DATA FILES (+ META-DATA?)

- Structured** (e.g., fixed record CSV), **semi-structure** formats (e.g., JSON, Parquet, email box), and **binary** (e.g., images)
- Rarely does the content **change** in these files
- Processing requires **large input and output data** chunks
- Workload is **read intensive**
- Early implementations that processes such datasets include **Hadoop** and later **Apache Spark** (emphasis on batch or serial processing)
- Integration with an enterprise datasets are based on ad hoc solutions** (i.e., no systematic, nor orchestrated ETL)
- Closely associated with Cloud based hosting and Cloud computing

Data Analytics: How can we get it right? 19

19



MODALITIES TO MOVE DATA TOWARD THE DATA ANALYTIC

[OLTP|RAW DATA FILES]

Freshness: OLTP OK, Raw data source depends on its nature / provenance.

Query latency, at OLAP, might be an issue as the OLTP has limited resources and tuned to optimise OLTP operations.

[OLTP|RAW] asynchronously feed OLAP

Freshness: OLTP becomes an issue, Raw data source depends on its nature / provenance.

! The ETL process is the pipeline between data and OLAP availability; this can get contrived.

[OLTP|RAW]_{first} OLAP_{second}


Basically, they are trying to address the ETL bottleneck (both in the logic and computation requirements) by forcing updates directly to the OLAP.

Freshness at OLAP improves

! A heavier computational load at both ends is expected to meet throughput and latency.

Data Analytics: How can we get it right? 20

20



MODALITIES TO MOVE DATA TOWARD THE DATA ANALYTIC (CONTINUED)

OLTP|RAW DATA FILES
[OLTP|RAW] asynchronously feed OLAP
[OLTP|RAW]_{first} OLAP_{second}

OLTP & RAW & OLAP working synchronously - HTAP


We are trying to address **freshness concerns** at each end of the **pipeline** – get **almost real time processing** on **fresh data** and with very **demanding latency**.

OLAP gets high data freshness, the OLTP gets analytic replies from the OLAP with good latency and results based on fresh computation on fresh data from the OLAP.

⚡ Most challenging, now, this is a very intense research and industry interest!

Data Analytics: How can we get it right? 21

21



A SYSTEM TO SUPPORT OLTP & OLAP WORKLOADS

Hybrid Transactional/Analytical Processing HTAP (Gartner, 2014)

Is a computerised system that **supports workloads found in OLTPs and OLAPs** through a common data architecture.

HTAPs are built to support intensive TP systems with embedded analytical processing with **near real time latency** and **use of very fresh data** when executing data analytics.

The HTAP data architecture is **devoid of ETL** processes, and it is simpler.

There is **no clear definition** on what an HTAP should be, nor what targets it must meet.

Goals of HTAP (other than those of OLTPs and OLAPs)

Performance Isolation requirement:

1. OLTP's workload does not interfere with OLAP latency and freshness requirements
2. OLAP workload does not interfere with OLTP throughput and latency requirements


Clearly: **freshness, throughput and latency measures** are important KPIs to monitor (for implementers)

The pain

Hype, plus lack of standards. **Resource** trade-offs. **Expertise** requirements. Continuous **monitoring** of its performance (and consequent rebuilds).

Data Analytics: How can we get it right? 22

22



BUILDING AN HTAP DATA ARCHITECTURE

The **data and query model** choice is usually accommodating
 Relational, document (e.g., JSON), key-value pairs, columnar, and graph based are possible
 We usually managed add specialised data modelling too; for example:
 Spatial, Temporal, Spatial-temporal, Dimensional

The **physical data models'** choices are a few, this has some serious implications

- ✓ Row (and in heaps)
- ✓ Columnar
- ✓ Multi-dimensional structures (R-trees)
- ✓ Partitioned data (e.g., rows by month)
- ✓ Distribute physical data across multiple hosting sites (e.g., component databases)

Boost these with
main memory
caches systems!


The **compute model** could be on-prem or cloud based

For example, Google offer AlloyDB (reconfigured PostgreSQL) with decoupled compute and storage facilities.

Another option is **main memory DBMS**

Data Analytics: How can we get it right? 23

23

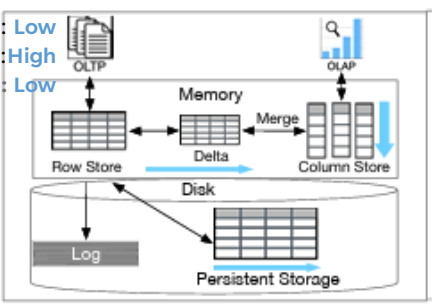


A QUICK DIGRESSION: REPORTED HTAP SETUPS

The paper “**HTAPdatabases: What is New and What is Next**” by Li & Zhang, 2022 @ SIGMOD22 gave four categories (as follows).

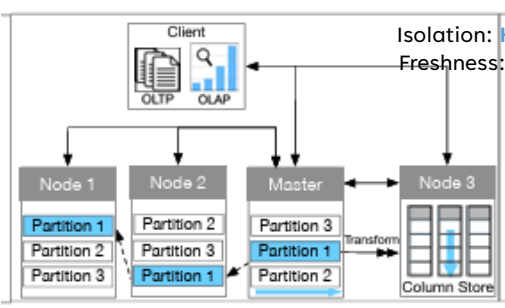
Primary Row Store + Ram Column Store

Isolation: **Low**
 Freshness: **High**
 AP scale: **Low**



Distributed Row Store + Column Store Replica (node #3)

Isolation: **High**
 Freshness: **Low**



Data Analytics: How can we get it right? 24

24

**A QUICK DIGRESSION:
REPORTED HTAP SETUPS**

The paper “**HTAP Databases: What is New and What is Next**” by Li & Zhang, 2022 @ SIGMOD22 gave four categories - (as follows) CONTINUED

**Disk Row Store
+ Distributed Column Store**

Isolation: **High**
Freshness: **Medium**

**Primary Column Store
+ Delta Row Store**

Isolation: **Low**
Freshness: **High**
TP scale: **Low**

Data Analytics: How can we get it right? 25

25

**A QUICK DIGRESSION:
REPORTED HTAP SETUPS**

GRIDGAIN APACHE IGNITE

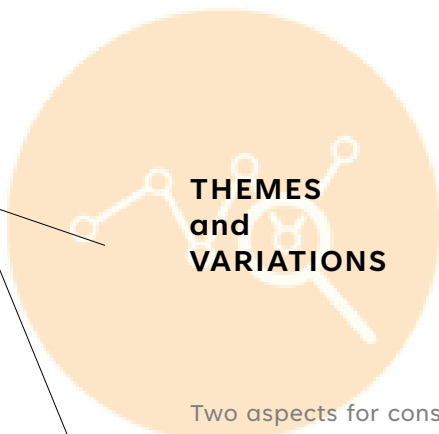
Ignite is a computing platform offers **in-memory data processing** with aggressive horizontal scalability, and latching to other technologies (e.g., streaming, AI learning frameworks).

Also, it supports **SQL III**, **ACID** support for TP, and **massive parallel processing (MPP)**.

[HTTPS://WWW.GRIDGAIN.COM/DEVELOPER](https://www.gridgain.com/developer)

Data Analytics: How can we get it right? 26

26



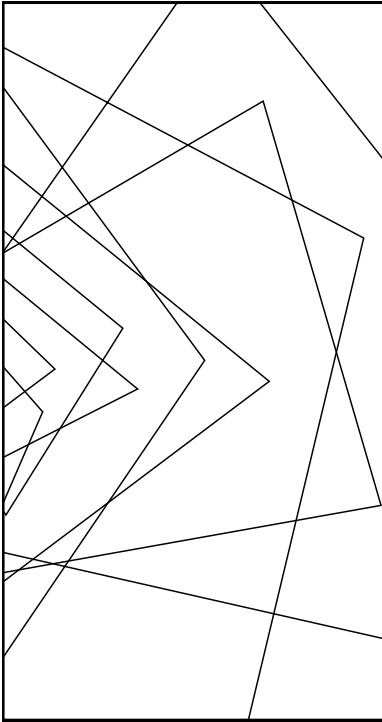
**THEMES
and
VARIATIONS**

Two aspects for considerations:

- How to ensure **adequate performance** (e.g., freshness, latency) before going live?
- How to massage your data into **better quality**?

Data Analytics: How can we get it right? 27

27




BENCHMARKING FOR ADEQUATE PERFORMANCE

A BENCHMARKING MIND IS AN “INTERPRETER”
MIND WHICH CAN BRIDGE THE GAPS.
PEARL ZHU

Data Analytics: How can we get it right? 28

28



DATABASE BENCHMARKING

Database benchmarking are meant to compute performance metrics for on standardised datasets and workload.

Database benchmarking has a long history and through time numerous benchmarking suites have been published. The best known are those from the TP Performance Council (www.tpc.org).


The Council provides benchmark suites for OLTP, OLAP, and DSS. Each benchmark includes:

- ✓ Schema and Data generation at a scale factor;
- ✓ Workload for database change of state operations
- ✓ A set of queries (in SQL) with runtime binding of literals.
- ✓ A strict reporting procedure (including specific metrics to compute and publish).

Historical note: Jim's Gray description of database benchmarking (1992)
"This quantitative comparison starts with the definition of a benchmark or workload. The benchmark is run on several different systems, and the performance and price of each system is measured and recorded. Performance is typically a throughput metric (work/second), and price is typically a five-year cost-of-ownership metric. Together, they give a price/performance ratio."

Data Analytics: How can we get it right? 29

29




HOW IS BENCHMARKING USED?

1. Select a DBMS and a hardware setup, run the benchmark suite and report the absolute performance metrics (through audited results – see Council's website).
2. Run the same benchmark suite multiple times, on same hardware and software, once as a baseline and the others with changes in the set-up or techniques for executing the benchmark.
For example, baseline uses B-trees indexes, the other uses R-trees indexes).
The metrics show the relative differences in performance between set-up.
3. Run a benchmark suite and add other workload (i.e., on top of the specified) to test the performance of an added workload – this is sometimes called micro benchmarking.
For example, run an OLTP suite and run an OLAP query as the added workload concurrently on the same schema.
The metrics show the differential effect of running two different workload modalities in interleaved fashion.

Data Analytics: How can we get it right? 30

30



BENCHMARKING IN DATA ANALYTICS SYSTEMS: WHY?


INSIGHTS:

- There is a good number of variations in design options – these multiply.
- Once adopted, some of the variations are not easily undone.
- There are various processes (e.g., pre-processing, statistical analysis requirements, data quality processes) whose workload computational requirements are not easily modelled.
- Home datasets develop data distributions, not necessarily growth oriented, that disrupt the computational requirement of a workload.

We have used benchmarking suites, for relative and differential comparisons, to help decide on the better data architecture to adopt that, meets a specific data requirements of a data analytics project.

Data Analytics: How can we get it right? 31

31



SPECIFIC BENCHMARKING SOLUTIONS FOR HTAP

HTAP system benchmarking requirements

HARNES EG. 1

- Suite:** TPC-H (DSS)
- Add:** workload of adding raw data, and specific query
- Measure:** query response times (of TPC-H queries and specific query), and data freshness

Run a few different schema sizes and two different data placement options (e.g., heap vs columnar).

For each combination, compute the query set performance, metric measures and build a simple model.

HARNES EG. 2

- Suite:** TPC-DS (DSS)
- Add:** TPC-C (OLTP)
- Measure:** query response times (of TPC-DS queries and TPC-C), and data freshness

Run a few different schema sizes and two different data indexing options (e.g., B-tree vs R-tree).

For each combination, compute the query set performance, metric measures and build a simple model.

Data Analytics: How can we get it right? 32

32




DATA QUALITY MAINTENANCE

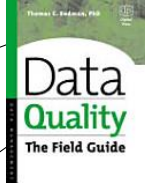
QUALITY IS NEVER AN ACCIDENT; IT IS ALWAYS THE RESULT OF INTELLIGENT EFFORTS.
JOHN RUSKIN

Data Analytics: How can we get it right? 33

33



DATA QUALITY



We say our data has a **high quality** if it is "fit for [its] intended uses in operations, decision making and planning"
wikipedia.org/Data_quality citing the work of TC Redman "The Field Guide" (2008)

Characteristics of data quality include:

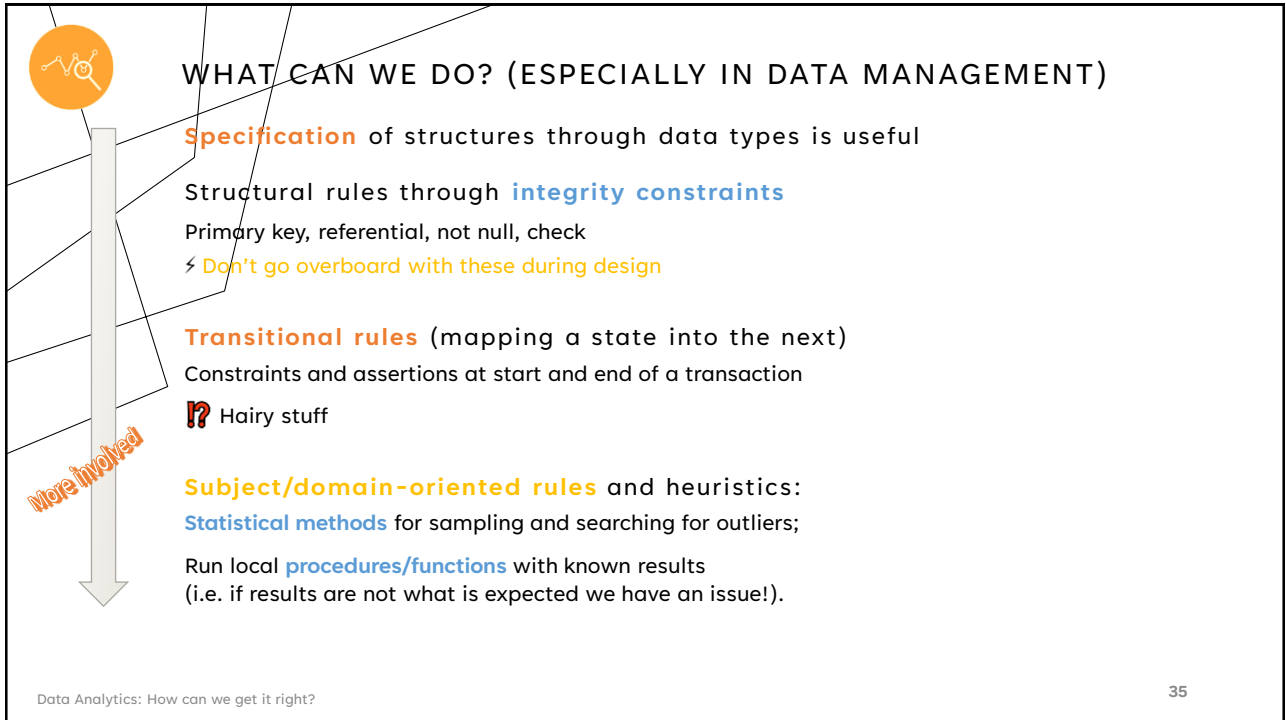
- **Completeness** of data instance;
- **Conformity** within the attribute's range;
- **Consistency** between an instance attributes;
- **Referential** integrity between records;
- **Accuracy** of record when compared to current reality.

Origin or onset of **bad data** includes:

- ☒ **Misreads and Missing** data;
- ☒ **Wrong import** (movement of data from one application to another);
- ☒ **Application errors** (in design, assignment, and runt time checks).

Data Analytics: How can we get it right? 34

34



WHAT CAN WE DO? (ESPECIALLY IN DATA MANAGEMENT)

Specification of structures through data types is useful

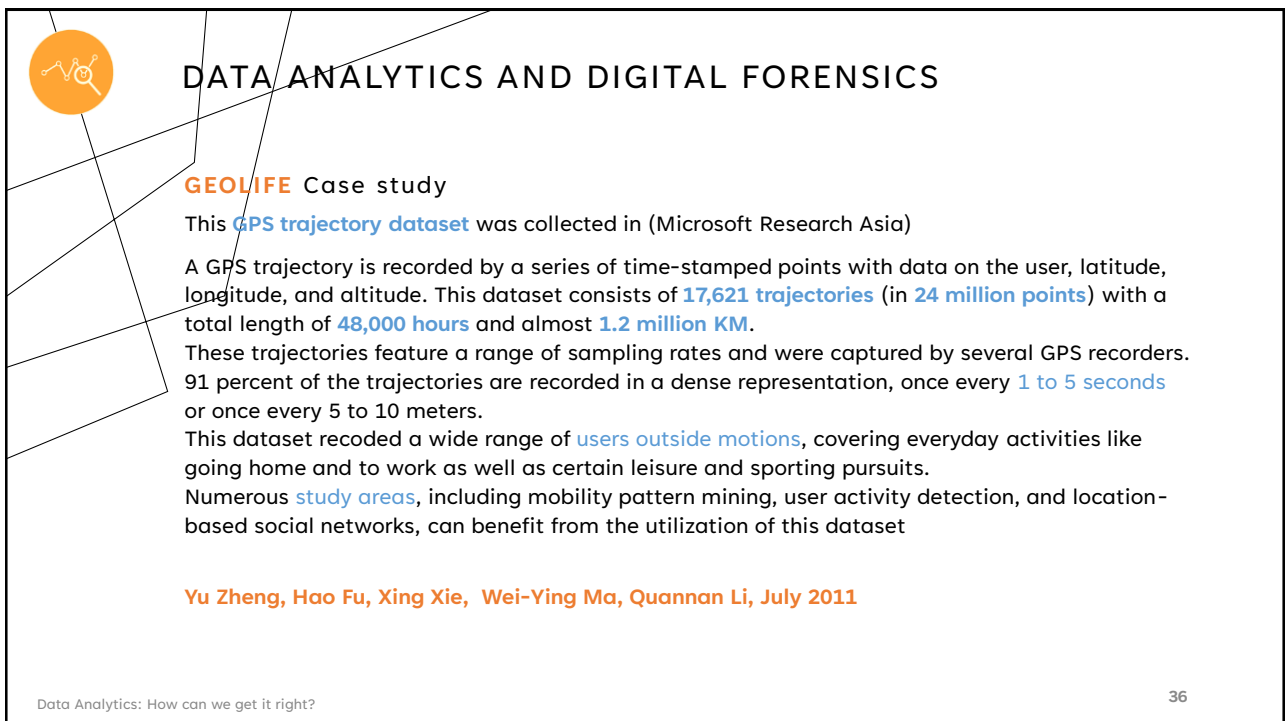
Structural rules through **integrity constraints**
Primary key, referential, not null, check
⚡ **Don't go overboard with these during design**

Transitional rules (mapping a state into the next)
Constraints and assertions at start and end of a transaction
!? Hairy stuff

Subject/domain-oriented rules and heuristics:
Statistical methods for sampling and searching for outliers;
Run local **procedures/functions** with known results
(i.e. if results are not what is expected we have an issue!).

Data Analytics: How can we get it right? 35

35



DATA ANALYTICS AND DIGITAL FORENSICS

GEOLIFE Case study

This **GPS trajectory dataset** was collected in (Microsoft Research Asia)

A GPS trajectory is recorded by a series of time-stamped points with data on the user, latitude, longitude, and altitude. This dataset consists of **17,621 trajectories** (in **24 million points**) with a total length of **48,000 hours** and almost **1.2 million KM**.

These trajectories feature a range of sampling rates and were captured by several GPS recorders. 91 percent of the trajectories are recorded in a dense representation, once every **1 to 5 seconds** or once every 5 to 10 meters.

This dataset recorded a wide range of **users outside motions**, covering everyday activities like going home and to work as well as certain leisure and sporting pursuits.

Numerous **study areas**, including mobility pattern mining, user activity detection, and location-based social networks, can benefit from the utilization of this dataset

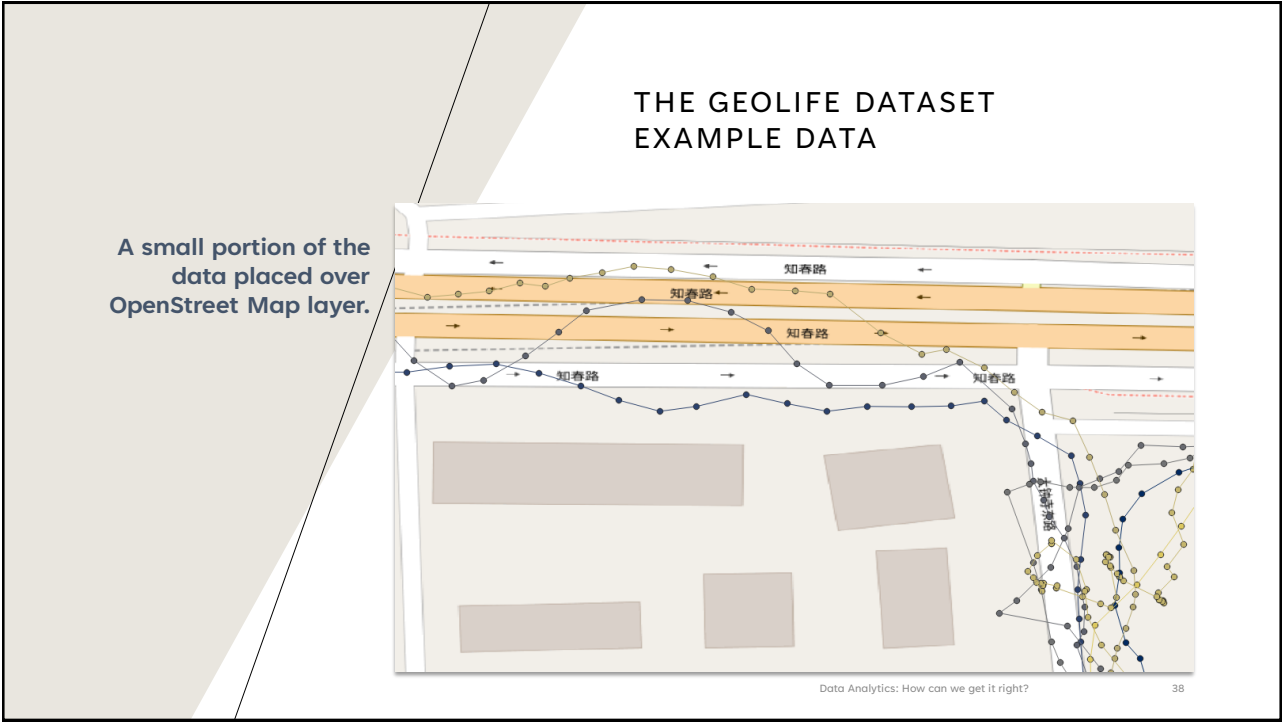
Yu Zheng, Hao Fu, Xing Xie, Wei-Ying Ma, Quannan Li, July 2011

Data Analytics: How can we get it right? 36

36



37



38

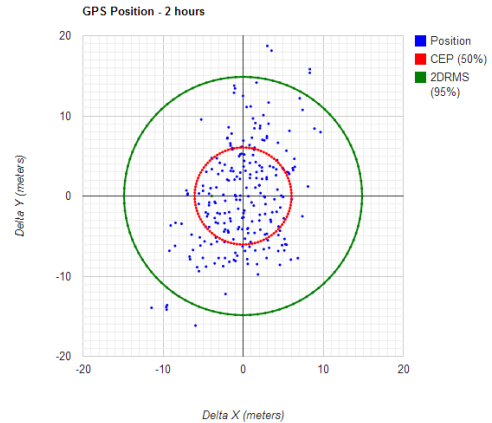
ASIDE: GPS ACCURACY

Position of satellites

The unit features (e.g. frequency bands it supports)

Environment factors (from satellite to receiver)

Surrounding of the receiver (e.g. getting a signal after a reflection)



M Coyle

<https://blog.oplopanax.ca/2012/11/calculating-gps-accuracy/>

NOTE – GPS TIME IS VERY ACCURATE!

Data Analytics: How can we get it right?

39

39



DATA QUALITY ISSUES WITH GEOLIFE

Data cleaning and visualisation.
Some issues include:


- Data is **out of Beijing** – like Rome and Seattle.
- Successive points show spikes in speed.
- Journey **points are interrupted** (e.g. tube, tunnels, GPS signal obfuscation and or multipath issues).
- **Gaps** between a user's journey (e.g. turned off).

In general, the data is getting old!

Data Analytics: How can we get it right?

40

40



BRIEF: INVESTIGATE IF TWO PERSONS TRACK HAVE AFFINITY

Affinity?
In time, space, and travel mode

In case you are asking!
An investigation is being undertaken to determine any stalking of a target.

Very Important: Any real data of the target and suspects can only be obtained by court order.

Main tasks vis data quality

Striaghten and mimimise the data required to represent a traversal in a mode of transport

Ensure the data falls within strict limits relative to the mode of transport (e.g. in walking one cannot have Olympic speed records);

Ensure networks of different transport modes are available;


Overlap with known tracks of mode of transport with user tracks;

Minimise the size, in terms of data reduction, of the "corrected" trajectory.

Data Analytics: How can we get it right? 41

41

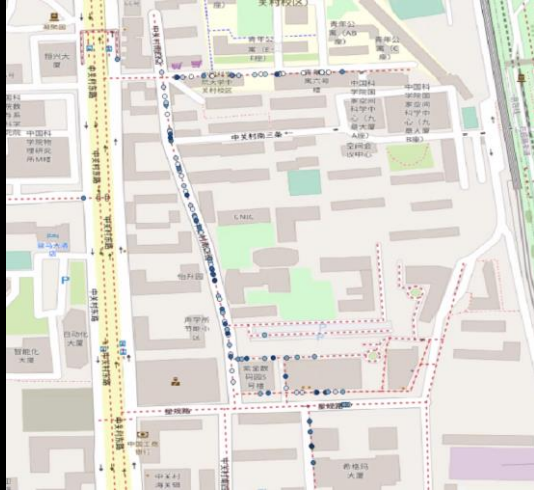
**ONE PERSON – 5 DAYS WALKING
WHITE TO DARK BLUE IS THE TIME PROCESSING**



Data Analytics: How can we get it right? 42

42

ONE PERSON – 5 DAYS WALKING
ADJUSTED POINTS
(WHITE TO DARK BLUE IS THE TIME PROGRESSION)



Data Analytics: How can we get it right?


43

ANIMATION OF RAW AND ADJUSTED DATA



Data Analytics: How can we get it right?

44



DATA QUALITY & REDUCTION TECHNIQUES

Reduce the number of points, but remain loyal to the data.
Sometimes, we need to offload data off too!?

Technique one:
Douglas-Peucker algorithm (e.g., ST_SIMPLIFY() is a Postgis Function)
results not clear as algorithm requires localised tuning

Technique two:
Approximate entropy (it works for a timeseries data fragments)
It helps breaking a track into “regular” parts. Promising results and is under testing!

Technique three:
Classification of a track by speed and direction changes to determine type of mobility (walking, car, train)

Technique four:
Better placement of a track on a road infrastructure (use of Hidden Markov Chains *).

(*) Newson and Krumm. 2009. Hidden Markov map matching through noise and sparseness. <https://doi.org/10.1145/1653771.1653818>

Data Analytics: How can we get it right? 45

45




TAKE HOME

On team composition, targets, technical savviness, and trade-offs

Data Analytics: How can we get it right? 46

46



TAKE AWAYS

Team formation and progress
ECLECTIC * DYNAMIC * DELEGATE WORK * KEEP EYES ON THE BALL

Target setting
DATA OWNER SETS A TARGET * TEAM SUPPORT (i.e., financial backing) * CLEAR FEEDBACK ON A TEAM'S FEEDBACK! * WATCH OUT FOR THE BIGGER HAPPENINGS

Technical prowess
EXTENSIVE * RE-USE WHAT IS SENSIBLE * DATA MODELS AND QUERY MODELS SHOULD BE CHOSEN ON THEIR ABILITY TO BE COHESIVE WITH OTHERS * TRANSACTION PROCESSING MODELLING IS VERY IMPORTANT * WATCH OUT FOR CHOKE AND KINKS IN PERFORMANCE GRAPHS

Trade-offs
DATA QUALITY VS RESULTS * ENSURE PERFORMANCE METRICS ARE KNOWN – SET LOW AND HIGH-WATER MARKS FOR INTERVENTION

Data Analytics: How can we get it right? 47

47



**THANK YOU!
ANY QUESTIONS?**

48