Aletheum

open-source business-systems consultancy

Part 0: Data Foundations

Data Analysis with Python (and Google Sheets Foundations)

Hisham Ahmed

Objective

Build a shared vocabulary for data concepts before touching any tool.

- Every later module reuses the terms defined here.
- Precision now avoids ambiguity later — in a formula, a schema, or a chart.

ALETHEUM

What Is Data?

Data, defined

ISO/IEC 2382:2015, term 2121272

Data: a reinterpretable representation of information in a formalized manner, suitable for communication, interpretation, or processing.

ISO/IEC 2382:2015, term 2121273

Information: knowledge concerning objects — facts, events, things, processes, or ideas — that, within a context, has a particular meaning.[1]

[1] ISO/IEC 2015.

Data vs. information

- Data is the formalised representation: a cell value, a byte string, a row.
- Information is the interpreted meaning: what that value tells someone, in context.
- The same datum (42) carries no information until a schema says whether it is an age, a row count, or an error code.

Data → Information → Knowledge → Wisdom

Ackoff's hierarchy: data organised becomes information; information understood in relation becomes knowledge; knowledge applied with judgement becomes wisdom.[2]

[2] Ackoff 1989.

ALETHEUM

Data Types

Data types: levels of measurement

Stevens' four scales classify any variable by which operations are meaningful on it.[3]

Scale	Admits	Course term
Nominal	equality ($=$)	categorical, boolean
Ordinal	ordering ($<$)	categorical (ranked)
Interval	difference	dates
Ratio	true zero, ratio	numerical

A boolean is nominal, restricted to two values. Dates are interval-scaled: differences are meaningful, but zero is arbitrary. Prices are ratio-scaled: zero means none of the quantity, so $20 is genuinely twice $10.

[3] Stevens 1946.

ALETHEUM

Data Structures

Data structures

Standard definitions, independent of any language.[4]

Array fixed-size, contiguous, homogeneous — indexed access in constant time.

List ordered sequence, dynamic size — a spreadsheet column is a list of values.

Table (relation) a set of tuples over a fixed schema — a spreadsheet, formally.

[4] Cormen et al. 2009.

ALETHEUM

Data Modelling

The relational model

Codd (1970)

Data is represented as *relations* (tables); it is manipulated by relational operations, independently of how it is physically stored.[5]

- A relation is a table: a fixed set of named, typed columns and a set of rows.
- Structure (schema) is separated from content (rows) — the schema is what the modelling module is about.

[5] Codd 1970.

Primary and foreign keys

Primary key a minimal set of attributes that uniquely identifies each row of a relation.[6]

Foreign key an attribute set in one relation that references the primary key of another, enforcing *referential integrity*.[7]

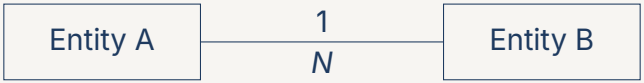
[6] Date 2003.
[7] Codd 1985.

ALETHEUM

Relationships and Cardinality

Chen (1976)

Data is modelled as *entities* and the *relationships* between them, each relationship carrying a *cardinality*.[8]



One-to-one, one-to-many, many-to-many — the cardinality determines which join operations are safe.

[8] Chen 1976.

Cardinality problems in joins

Joining across the wrong cardinality silently multiplies or drops rows — textbook *connection traps*.[9]

Fan trap a one-to-many join fanned out through a second one-to-many relationship, inflating aggregates.

Chasm trap a missing relationship between two entities that are only connected via an optional intermediary, silently dropping rows.

A join is only as trustworthy as the cardinality it assumes.

[9] Connolly and Begg 2014.

ALETHEUM

Data Categories

Dimensional vs. transactional data

Dimensional (reference)

Slowly changing descriptive attributes: a customer, a product, a calendar date.[a]

[a] Kimball and Ross 2013.

Transactional (event-based)

Immutable records of things that happened: an order, a payment, a login.[a]

[a] Inmon 2005.

	Dimensional	Transactional
Changes	rarely	continuously
Volume	low	high, growing
Role in a join	the “lookup” side	the “fact” side

ALETHEUM

Hands-on Exercise

Exercise: map a familiar process

Pick a process you know well — ordering coffee, booking a room.

1. List its entities (customer, order, room, booking…).
2. For each pair of entities, name the relationship and its cardinality (one-to-one, one-to-many, many-to-many).
3. Mark which entities are dimensional (reference) and which are transactional (event-based).

This is the conceptual data map every later module will keep returning to.

References I

- Ackoff, R. L. (1989). “From Data to Wisdom”. In: *Journal of Applied Systems Analysis* 16.1, pp. 3–9.
- Chen, P. P.-S. (1976). “The Entity-Relationship Model—Toward a Unified View of Data”. In: *ACM Transactions on Database Systems* 1.1, pp. 9–36.
- Codd, E. F. (1970). “A Relational Model of Data for Large Shared Data Banks”. In: *Communications of the ACM* 13.6, pp. 377–387.
- — (1985). “Is Your DBMS Really Relational?” In: *ComputerWorld*. Codd’s twelve rules; rule 10 covers referential integrity.
- Connolly, T. and C. Begg (2014). *Database Systems: A Practical Approach to Design, Implementation, and Management*. 6th. Fan and chasm connection traps. Pearson.
- Cormen, T. H. et al. (2009). *Introduction to Algorithms*. 3rd. MIT Press.

References II

- Date, C. J. (2003). *An Introduction to Database Systems*. 8th. Addison-Wesley.
- Inmon, W. H. (2005). *Building the Data Warehouse*. 4th. Wiley.
- ISO/IEC (2015). “ISO/IEC 2382:2015 Information technology — Vocabulary”. In: Terms 2121272 (data) and 2121273 (information).
- Kimball, R. and M. Ross (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. 3rd. Wiley.
- Stevens, S. S. (1946). “On the Theory of Scales of Measurement”. In: *Science* 103.2684, pp. 677–680.