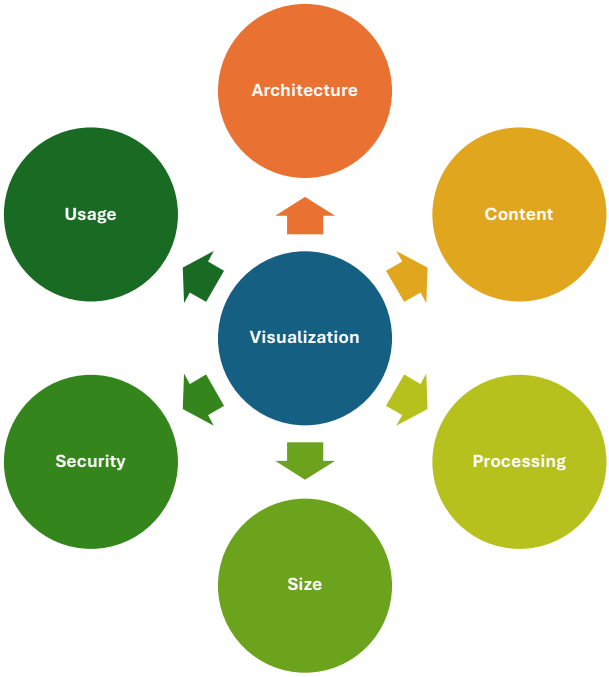
Data Testing Analysis Heuristics

This document lists a set of heuristics that guide analyzing data as used by an application or system for sake of shaping a testing investigation. The heuristics are presented as a series of guiding questions that prompt further exploration. They are organized into the following categories:



Visualization: Ways of helping you see or interpret the data.

Architectural Structure: Ways the data is arranged relative to parts of the system.

Data Content: Information about the data itself, the specific structure, properties, and values and attributes of those values.

Data Processing: Ways the data is handled, transformed, moved, and used by the system.

Size and Scale: How the amount of data may affect system behavior.

Security: Thinking about data that is sensitive, vulnerable, or that must be protected from exploit.

Data Usage: Different ways data is used affects our understanding of the data.

Visualization

What can I notice if I draw the schema?

What can I notice if I draw the relationships

- What will I see if I break relationships into repeated patterns and shapes?
- What are some common shapes I use to understand the data?
    - Trees
    - Loops (size: 1,2,3,N – bugs differentiate across some of these sizes)
    - Linear
    - Single and Two-directional links
    - 1 to 1, 1 to many, many to 1, many to many

Will I see something if I try different visualization types?

- Heatmaps, scatter plots, line graphs, state models, flow charts...

What kinds of patterns and trends matter or exist in the data?

- Recurrence and repeats
- Fluctuations across time
- Correlations between variables
- Spikes
- Outliers and Anomalies
- Cluster Analysis
- Time Series Analysis

Will looking at the data with different granularity and aggregation reveal something?

- Aggregate to fine granularity
- Fine granularity to aggregation of higher level
- Different aggregations on different pivots, clusters and values

What sorts of comparisons might reveal something?

- different time periods
- demographics and other dimensions in the data
- value ranges on specific columns or properties

Are there gaps and missing data?

- Are there classes of missing expected data or values?
- Are there gaps in time?

- Is anything missing from specific sources, services, or subsystems?
- Are there gaps between certain users, use cases, geographies?

Are we able to examine cause-and-effects or correlations in data relationships?

- Does one factor appear to influence another
- Do some values change together?

How does data distribution affect our understanding of the system?

- Pareto principle: are there large impacts from small proportions? (80/20 rule)
- How do values distribute across the population? Normal? Skewed? Pareto? No pattern? Long tail? Exponential growth?

Architectural Structure

Where is the data kept and processed?

- Does the data ever cross boundaries?
- Are there multiple tiers or different types of storage?
- How does data cross from one part of the system to another?

Is there some form of centralization and control, or is everything decentralized?

- Where is the “definition of truth” and how is it kept in sync?
- Does anything control how to find data or resources centrally? Central “lookup” or index or directory of storage?
- Is anything in charge of special value assignments – unique id, sequence numbering, lock and access control?
- If distributed, how does that affect data integrity, update and in-sync mechanisms? How do we check status on distributed data state?

What kinds of technology and platforms are used for data handling and storage?

- File system, relational database, No-SQL data, data lake, data warehouse
- What types of query and retrieval systems?
- What kind of tools for display and analysis?
- What sorts of data movement and pipelines are in use?
- How does storage technology/platform affect data and schema design, and vice versa?

Data Content

What is the definition of invalid or disallowed data?

- Are there external standards and rules?
- Are there operations which cannot succeed or will fail if data does not conform?
- Are there factors such as time and sequence or data received which might invalidate the data?

Is there anything schematically allowed, syntactically, but semantically invalid?

- Duplicate entries, are they allowed? What do they mean?
- Are there references to other items in the data and are they enforced?
- Are certain values meaningful if empty or null?
- Are certain values restricted to specific values and is that restriction enforced?

Does the data use loose or strict contract expression and enforcement?

- Foreign key enforcement or not
- Uniqueness constraints
- Character pattern limits
- Data types – do they limit to only valid values or permit non-valid?
- Non-null/empty enforcement
- Upper/Lower case sensitive, insensitive, enforced or not
- Similar spellings (non-enforced sameness in terms)
- Negative versus non-negative
- Integer versus floating point
- Is the contract described in an explicit way, or is there a lot not described?

Are date and time relevant within the data content?

- Formatting
- Time zone and daylight savings sensitivity
- Server versus client representations

Are multi-valued and multi-row data types used? How do they affect usage, representation?

Does order matter?

- Are two lists with items of same value but different order the same or different?
- Is order increasing, decreasing, random, non-changing?
- Is data changing at different slopes and rates, gaps, clusters of order?

- Are consumers going to expect data to be ordered? How does that affect things like paginated retrieval?

How do relationships affect the system?

- Do certain data points and values travel together?
- Are certain data exclusive, should never be together?

Are special characters relevant to the system?

- Which command and control content should we worry about?
    - Does the content it affect processing decisions in parts of the system?
    - Is it a command for something outside the business logic?
        - SQL
        - Command line
        - JavaScript, HTML
- Are there any disallowed characters? How is that enforced?
- Are there paired characters and values? Are there recursive or hierarchical relationships?
    - Left/Right parentheses, braces
    - Quotes
    - XML & HTML Tags
- Do we have to deal with character encoding?
    - Canonicalization, Unicode, multi-byte switching, code pages, upper/lower ASCII, escape characters and values
    - Control characters confused with regular multi-byte characters
    - What is the level and point where encoding is handled?

Is data direction relevant?

- RTL vs. LTR display
- RTL vs. LTR processing
- Byte ordering (old-school testing problems on different platforms)

Are there interesting character and value sequences?

- Must follow or go together
- Must not follow or go together
- Imply special meaning

Data processing

What is the data lifecycle?

- Does the data age?
- Is there temporary data used for transitory operation?
- Is there data that must be archived?
- Is there data that must be deleted?
- What is the cost of data processing and storage?

How does data move through the system?

- Does data need to pass between components?
- How is data interpreted between shared components?
- Are there tiers of storage (hot/cold, cached)?
- Is the data transformed when moved?

Is the data ever transformed, altered, or processed?

- Is data changed or altered on ingestion?
- Does time affect data in terms of order of processing?

Do we control all the data processing?

- Does data coming from other systems match our expectations?
- What is impact on other systems that consume our data?
- Are there rules about data format, content, or relationships dictated by systems other than ours?

Size and Scale

How big does the data get?

How does the data grow, what values drive it? Do certain values/variables grow together or compound on each other?

What is the impact of size on processing/write/retrieval? does cost grow as size increases, or is it always incremental, one piece at a time?

How is the system impacted by data scale?

- What is the driver of scale – which data?
- How do we add more capacity?
- Do current resources have maximum limits?

- How much do our resources cost the business?

Security

What policies or regulations are relevant to this data?

How is the data secured?

What roles, groups, actions or other authorization of data exists?

What is the level of granularity on authorization and restrictions: whole thing, db level, table, row, column, groups of data, tags – other?

Is there any sensitive data representing PII, special interest or highly valuable/confidential data?

What mechanisms/process/tier decides access and permission?

Data Usage

Are there ways values in usage affect the system?

- Is there anything that is nonsense?
  - E.g. brand mismatch – "Volkswagen Ranger", geographical "Seattle, Oregon"
  - Common vs. Rare values
    - "Johnson", "Srinivasan" – common names regionally
    - "Ziniewicz" – uncommon in the US
  - Significant values in real life
    - Locations in different time zones from each other
    - Location distances apart from each other
    - Very large things (semi-truck) vs. small (tricycle)
    - Odd combinations: e.g. "2-ton tricycle"
  - Data calculated from other data: e.g. vin implies year, make, model, country…
  - Two pieces of data that must be in sync
  - Data ranges and order that make no sense together: e.g. birth date after first year of school

How does data affect the business use case?

- Is there certain data that matters more than other data?

- What are some of the things end users will do with this data, and is the data sufficient to satisfy that use?
- What happens during usage when certain data is different than expected, missing, invalid?

How does domain expertise affect the data?

- Does data match domain specific expectations and hypotheses?
- Are there things which are invalid or nonsense within a domain?
- Are there gaps that working with a domain need filled?

Can I test hypotheses with this data?

- Usage patterns
- Defect discovery
- Growth and scale rates
- Load patterns