Data Quality

Defintion #1: “Fitness for Use”

- *Degree to which data can be used for its intended purpose*

Definition #2: “Real-World Representation”

- *Degree to which data accurately represents the real world*

We can try to satisfy BOTH definitions!

Data Quality: Measures

Completeness: Do we have all the data we expect to have?

- Are all events captured?
- Are all attributes of an event captured?
- Are all values of reference data accounted for?
- Uniqueness: Are single events captured only once?

Accuracy: Is the data a true representation of the idea it's trying to capture?

- Are numbers and string values correct?
- Are timestamps and other attributes correctly captured?
- Consistency: Do I capture the same data the same way every time? Do I capture it the same in each place?

Data Quality: Measures

Conformance / Validity: Does stored data conform to syntax, encoding, or other model specifications?

- Are data formats correct?
- Are codes as expected?
- Are naming conventions adhered to?

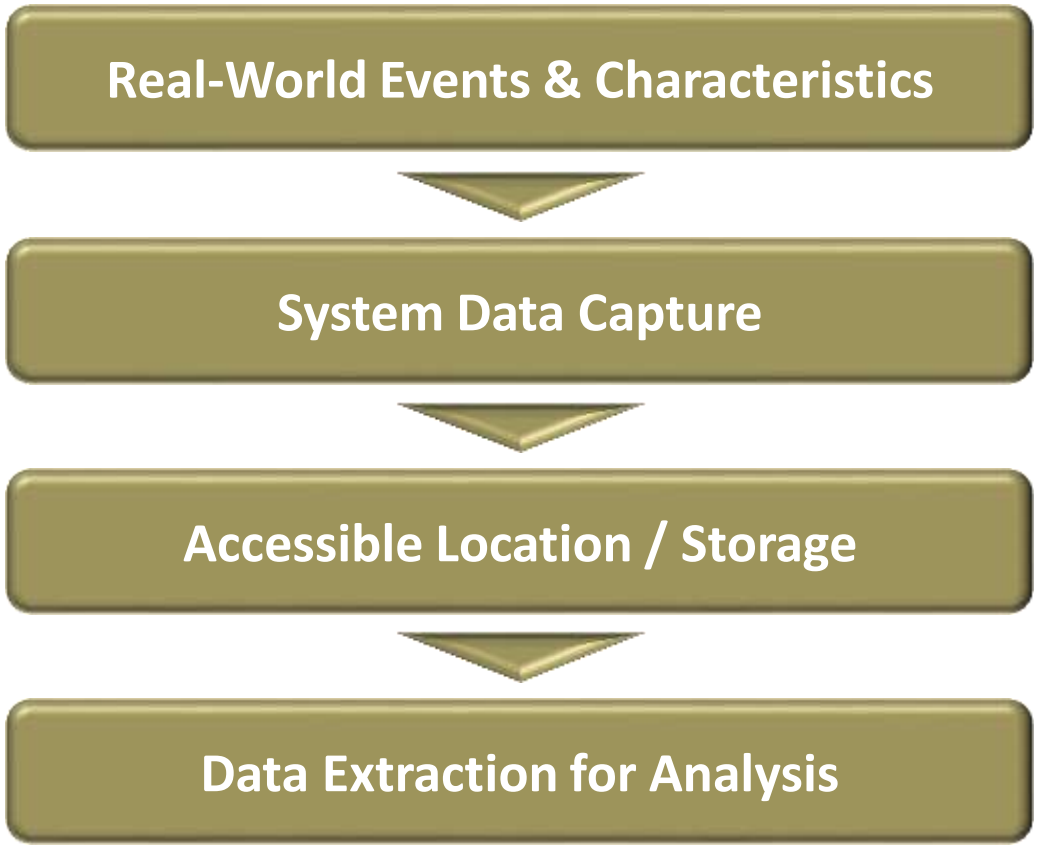
Timeliness: Is data available by the time it is needed?

- Also referred to as “Data Latency”

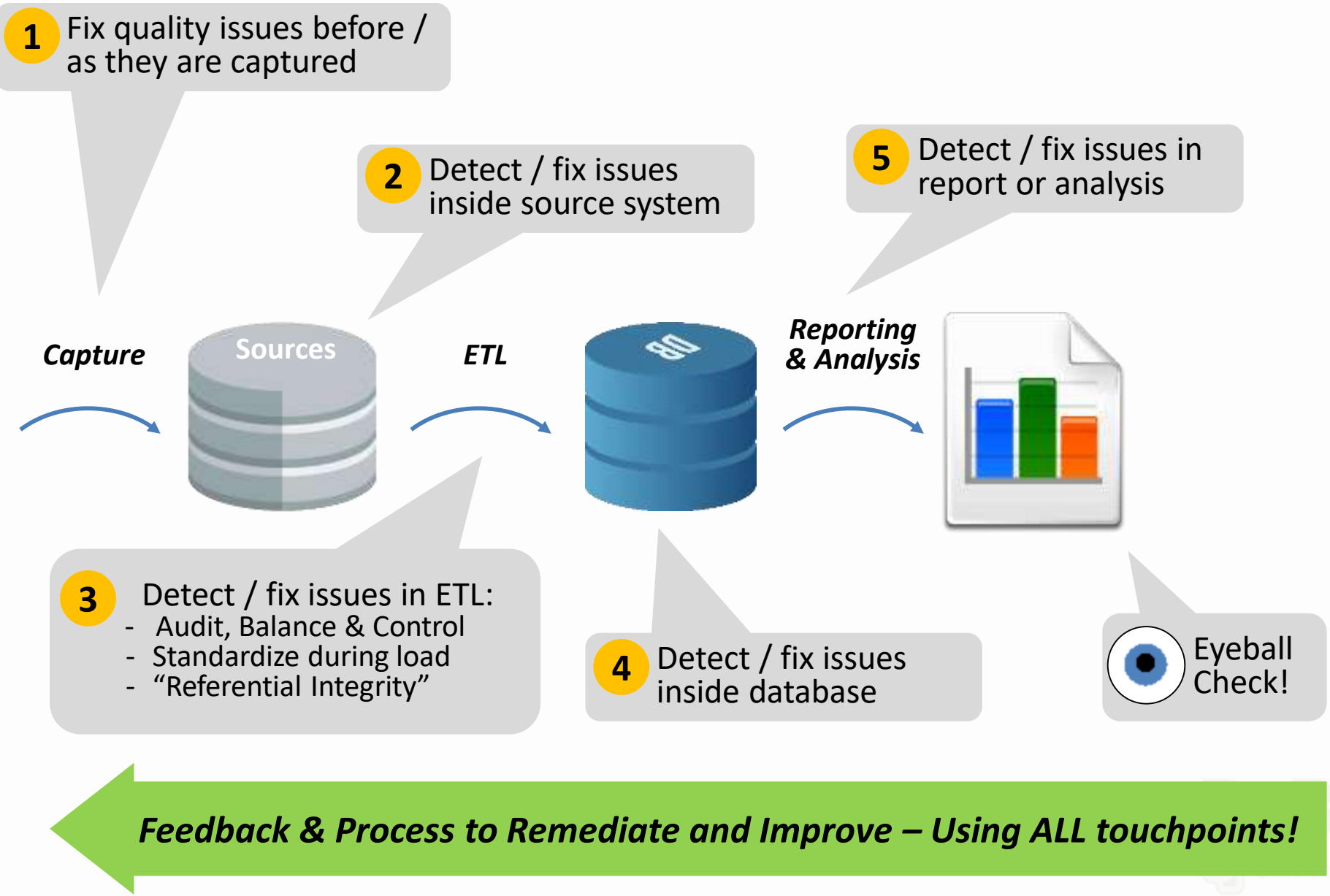
Provenance: Do we have visibility into the origins of the data?

- How much confidence do we have that the data is real and accurate?

The Information-Action Value Chain



Data Quality: Process



Recap

“Fitness for Use” and “Real-World Representation” Definitions

Measures:

- Completeness / Uniqueness
- Accuracy / Consistency
- Conformance / Validity
- Timeliness
- Provenance

Data Quality Process