Summary: Data Quality: Definition, Causes & Measuring

Definition of Data Quality:
High-quality data is free from errors and artifacts. Errors are missing or irrecoverable data, while artifacts arise during data gathering, processing, or cleaning. Poor data can result from individual errors or systemic issues, often amplified in the context of big data (characterized by Volume, Variety, Velocity, and Veracity).

Consequences of Poor Data Quality:
Bad data leads to severe economic and operational issues, exemplified by costly errors such as NASA's Mars Climate Orbiter failure. Poor data quality impacts decision-making, increases costs, and undermines system reliability.

Types of Data Errors:

1. **Formatting Issues:** Irregular formats, NULL handling inconsistencies, and incompatible delimiters.
2. **Content Issues:** Duplicates, missing data, outliers, or inconsistent units (e.g., metric vs imperial).

Errors in the Data Pipeline:

1. **Gathering Phase:** Transmission problems, manual entry mistakes, and flawed survey designs. Solutions include preemptive integrity checks and retrospective cleaning.
2. **Storage Phase:** Format conversion errors, lack of metadata, and poor version control. Addressed by preemptive metadata documentation and retrospective smell tests.
3. **Processing Phase:** Errors from integrating heterogeneous data sources or legacy systems. Solutions involve data browsing, exploration, and using commercial integration tools.
4. **Analytics Phase:** Issues like scale problems, lack of domain expertise, and biased models. Addressed by continuous analysis, accountability, and sanity checks.

Measuring Data Quality:
Key attributes include **accuracy, completeness, uniqueness, timeliness, consistency, and credibility.** These measures are often difficult to quantify due to context-dependence and vague definitions. Practical approaches include schema adherence, process constraints, and proxy measures like customer complaints.

Key Insights and Tools:
Ensuring data quality requires:

- Proactive measures (integrity checks, reliable protocols).
- Retrospective measures (cleaning, exploration).
- Building intuition through domain expertise and practice.

Conclusion:

Data quality is critical for reliable analytics, but challenges persist in defining, measuring, and addressing errors. A robust mix of strategies, tools, and domain expertise is essential to minimize errors and artifacts in the data lifecycle.