Data Quality Management

- Good quality data reduces ~~the~~ risks in projects and aids in effective decision-making.

Good Quality Data: Characteristics

- No duplicates - Consistent/accurate information
- Consistent formatting
- Referential integrity - Up-to-date

Techniques

- Data cleaning: Identify & fix issues in the data.
- Data profiling: Understand the data and its characteristics.
  - ▲ Structure Discovery - verify consistency in formatting & overall data structure. Mathematical/statistical measurements are often used.
  - ▲ Content Discovery - verify values are accure/logical. Check data completion.
  - ▲ Relationship Discovery - Verify Referential Integrity amongst ~~data~~ tables & overall relationships between entities/attributes.
- Data Transformation: Transforming data to become easier to understand & make decisions.
  - ▲ Methods:
    - Splitting - split a column into multiple columns.
    - Generalization - listing detailed data under a more generalized, ~~but still accurate~~ but accurate, category.
    - Translation & Mapping - Matching together the appropriate attributes between different databases to aid in comprehension.
- Data Validation: Ensure consistency, accuracy, completion & complaince w/ business rules.

- Master Data Management : Ensures consistency & exactness.
  - ▲ Master Data – Core, consistent data that represents key entities of a business.
  - ▲ Ensures that the master data is accurate & unified across the business.

Data Governance

- Provides policies, standards & rules to direct the management of data assets.
- Allows for secured sharing of data between depts. of the organization.

Data Integration

- The process of moving data to locations, while the data is cleaned/validated, to be used as needed.
- ETL / Batch Data Integration
  - automated; runs on a schedule
  - Extract , Transform , Load
    - Extract ↓ collect data
    - Transform ↓ transform data
    - Load ↓ load data to a data warehouse
- Streaming
  - Data is processed continuously as it arrives to its location.
  - Real-time processing & application.
- Replication – creates copies of data across applications
  - CDC (Change Data Capture): detects & replicates modifications of the data. ~~to t~~