2/22/2021

Data Security, Storage of Data
Spreadsheets vs. Databases

Data Storage

- What type of data?
- Where is the data storage?
- How does the software store the data?
- How does the system delete data?
- How is the data backed up?
- Temporary storage or permanent?

Old Methods:

- Punch cards
- Floppy Disks
- Tape Drives
- Hard Drives (Solid State, Flash Memory)
- Memory Cards
- CDs/DVDs

Where:
On-site storage, or cloud storage (server farms)

Data warehouse vs. Data Lake

Data warehouse

- More organized
- Abstract picture of a business categorized by subject
- Highly transformed and structured
- Not included unless a use for the data is well-defined
- Follows a set methodology

Data Lake

- No data is turned away
- Data is stored in a largely untransformed state
- Data is only processed when it is analyzed

Data Lake is a “critique” of data warehouse

1. Data lakes retain all data
2. Data lakes support all data types
3. Data lakes support all users
4. Data lakes adapt easily to change
5. Data lakes provide faster insights

Data Security

- Encryption
- Data Erasure – overwrite the old data to ensure non-recoverable
- Data masking – hides personally identifiable information (PII)
- Data resiliency – ability to recover from any type of failure

Data discovery/classification – identify sensitive information to help remediate any vulnerability

Data file and activity monitoring – see who is accessing data and spot anomalies and identify risks

Vulnerability assessment – out-of-date software, weak passwords, misconfigurations, where is the greatest exposure

Automated compliance reporting – audit trails

Physical security of servers and devices
Access management controls
Application security patching
Backups
Employee education
Network and endpoint security (monitoring and controls)

Future trends –

- Artificial intelligence
- Multi-cloud security – applications, data, and processes running on public and private clouds
- Quantum computing

Things to think about when considering data security:

- Adopt a risk-based approach
- Consider regulatory requirements and stakeholders
- Identify the most sensitive information
- Extend best practices
- Cloud network security is different than in-house
- Need classification tools, and ongoing monitoring

BYOD (bring your own device) and mobile

- Requires specific security measures for access
- Strong passwords
- Multi-factor authentication
- Regular software updates
- Device backups
- Encryption

Spreadsheets vs. Databases

Spreadsheets are simpler and easier to understand
Information from a database is often exported to a spreadsheet for calculation

(relational database) stores data in table: columns = fields, and rows = records
Like a dataframe
The relational database has connection between the tables
    Linked and cross-referenced

Well-defined relationships or rules enforce restrictions on the data
Tables can communicate/share data for searching, organizations and reporting

Databases have no formatting

Many databases have forms that make data entry and retrieval easier

Which is better?

What is the purpose of the data being collected?

- Spreadsheet – low volume, numeric or text data
- Database – better for images and documents, higher volumes, including dataloggers, GPS devices, cameras, drones, etc.

Data volume?

- Long-term projects with lots of data are better for databases
- Spreadsheets use more drive space for an equivalent amount of data
- Spreadsheets are more difficult to search

Editing?

- Easier to edit in a database than in a large spreadsheet
  Especially true if data is stored in multiple spreadsheets
- Databases can update records in bulk

Data accessibility and speed

- Databases can accommodate complex aggregation functions
- Reports and queries can be automated
- Hard to compare data in separate spreadsheets
- Can't enforce the same quality standards across spreadsheets
- Databases can operate faster than spreadsheets for similar amounts of data

Data integrity

- Relational databases follow standardizes integrity rules to ensure the data is accurate and accessible
- Fields can be restricted to specific data types, formats, lengths, etc.
- Referential integrity establishes relationships between tables
- Further value restrictions can prevent data-entry errors

Redundancy

- Database structures avoid redundancy (normalization)
- Databases reduce the need for version control vs. spreadsheets or data stored in multiple locations

Error Proliferation

- Preventing and identifying data errors in spreadsheets is a challenge
- Easier to prevent overwriting/deleting data in databases

User access and security

- Databases are designed for multiple users
- Centralized data storage
- User permissions can be granted to view data, edit data, or restrict access entirely