2/12/2024
Data Security, Data Storage
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 or permanent storage?

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
- Data is not included unless the 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 Lakes as a “critique” of a data warehouse

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

Data Security

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Vulnerability assessment – out-of-date software, weak passwords, misconfigurations, where the greatest exposure is

Automate 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 with 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

Bring your own device (BYOD) and mobile
- Require 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 in a database is often exported to a spreadsheet for calculation

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

Well-defined relationships or rules enforce restrictions on the data
Tables can communicate/share data for searching, organization, 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?

- Longer-term projects with lots of data are better in databases
- Spreadsheets use more drive space than a database for an equivalent amount of data
- Spreadsheets are more difficult to search

Editing?

- Easier to editing in a database than in a large spreadsheet
- Especially true if data is stored in multiple spreadsheets
- Databases can backup 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 several spreadsheets
- Databases can operate faster than spreadsheets for similar amounts of data

Data integrity?

- Relational databases follow standardized integrity rules to ensure data is accurate and accessible
- Fields can be restricted to specific data types, formats, length, 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 a database

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

Resources:

1. https://www.dataversity.net/what-is-data-storage/
2. https://www.zdnet.com/article/innovations-in-data-storage-an-executive-guide-to-emerging-technologies-and-trends/
3. https://www.bluegranite.com/blog/bid/402596/top-five-differences-between-data-lakes-and-data-warehouses
4. https://www.ibm.com/topics/data-security
5. https://www.beckershospitalreview.com/cybersecurity/25-most-common-passwords.html
6. https://earthsoft.com/2018/06/07/databases-versus-spreadsheets
7. https://www.oracle.com/database/what-is-a-relational-database
8. https://www.geeksforgeeks.org/definition-and-overview-of-odbms/
9. https://www.c-sharpcorner.com/article/what-are-object-oriented-databases-and-their-advantages2/

Extended commentary:

Data security refers to the protection of data from unauthorized access, use, theft, destruction, or modification. Data security is critical for organizations that collect, process, store, and transfer sensitive and confidential data, such as personal and financial information, intellectual property, and trade secrets.

Data security can be achieved through various means, including:

Access controls: Limiting access to data only to authorized personnel and implementing security protocols to prevent unauthorized access.

Encryption: The process of converting data into a code to prevent unauthorized access or theft of sensitive information.

Backup and recovery: Creating regular backups of data to ensure its availability in the event of data loss or damage.

Firewalls: A network security system designed to prevent unauthorized access to or from a private network.

Anti-virus software: Software designed to detect, prevent, and remove malicious software (malware) from computer systems.

Physical security: Physical measures to protect data, such as security cameras, locks, and access control systems.

Data privacy: Ensuring that personal data is collected, processed, and stored in accordance with applicable data privacy laws and regulations.

Organizations must implement a comprehensive data security strategy that includes technical, administrative, and physical controls to ensure the confidentiality, integrity, and availability of their data. Additionally, organizations must provide adequate training to employees on data security best practices and conduct regular audits and risk assessments to identify vulnerabilities and ensure compliance with data security policies and regulations.

Data storage refers to the process of saving data in a systematic and organized way so that it can be easily accessed and retrieved when needed. Data storage systems can vary greatly depending on the type of data, the volume of data, the required access speed, and other factors.

There are several types of data storage systems available today, including:

Hard disk drives (HDDs): HDDs are the most common form of storage used in computers today. They use rotating disks to store and retrieve data.

Solid-state drives (SSDs): SSDs are similar to HDDs but use flash memory instead of rotating disks. They are faster and more reliable than HDDs but are typically more expensive.

Cloud storage: Cloud storage refers to the use of remote servers to store data. Cloud storage providers offer scalable, pay-as-you-go storage solutions that can be accessed from anywhere with an internet connection.

Network Attached Storage (NAS): NAS is a dedicated storage system that connects to a network, allowing multiple users to access and share files.

Storage Area Network (SAN): SAN is a dedicated high-speed network that provides block-level access to data storage.

Data storage systems must also be secure and protect data from unauthorized access, theft, or damage. This is typically achieved through encryption, access controls, and backups to prevent data loss in the event of a system failure or disaster.

A data warehouse and a data lake are two different approaches to storing and managing data.

A data warehouse is a central repository of data that has been extracted from various sources, transformed into a consistent format, and loaded into a relational database for analysis and reporting. The data is typically structured and organized according to a predefined schema or data model. A data warehouse is designed to support business intelligence and decision-making by providing a consistent and reliable view of the data. It is often used for reporting and analysis of historical data, and it is optimized for complex queries and ad-hoc analysis.

A data lake, on the other hand, is a large, centralized repository that stores all types of data in its native format, including structured, semi-structured, and unstructured data. Unlike a data warehouse, a data lake is not limited to a specific schema or data model. It is designed to store raw, unprocessed data from various sources, such as social media, sensors, and log files, and make it available for analysis and exploration by data scientists, analysts, and other users. A data lake is often used for advanced analytics, such as machine learning and artificial intelligence, as well as for data discovery and exploration.

The advantages of a data warehouse include:

Consistent and reliable data: Data is transformed and organized according to a predefined schema, which ensures that it is consistent and reliable for reporting and analysis.

Optimized for complex queries: A data warehouse is optimized for complex queries and ad-hoc analysis, which makes it ideal for business intelligence and decision-making.

Data security and governance: Data warehouses typically have robust security and governance features to ensure that data is protected and managed according to organizational policies.

The advantages of a data lake include:

Scalability and flexibility: Data lakes can store and process large volumes of data in its native format, which makes it scalable and flexible for different types of data and analysis.

Agility and speed: Data lakes allow data scientists and analysts to quickly and easily explore and analyze data without the need for a predefined schema or data model, which enables faster time-to-insight.

Advanced analytics: Data lakes are ideal for advanced analytics, such as machine learning and artificial intelligence, which require large volumes of data and a flexible data environment.

The disadvantages of a data warehouse include:

Cost: Data warehouses can be expensive to implement and maintain, especially for small to mid-sized organizations.

Complexity: Data warehouses require a significant amount of planning, design, and implementation to ensure that the data is consistent and reliable for reporting and analysis.

Data silos: Data warehouses can create data silos if data is not properly integrated from different sources.

The disadvantages of a data lake include:

Lack of governance: Data lakes can be prone to data quality issues if proper governance and management processes are not in place.

Complexity: Data lakes can be complex to implement and manage, especially for organizations without the necessary data engineering and data science expertise.

Security: Data lakes can be vulnerable to security breaches if proper security measures are not in place.

Spreadsheets and databases are both commonly used tools for managing and analyzing data. Here are some advantages and disadvantages of each:

Advantages of spreadsheets:

Easy to use: Spreadsheets have a familiar interface and are relatively easy to use, making them accessible to a wide range of users.

Flexible: Spreadsheets can be used to store and analyze many types of data, and users can easily manipulate the data with built-in functions and formulas.

Quick and simple: For smaller datasets, spreadsheets can be a quick and simple solution for data management and analysis.

Inexpensive: Many spreadsheet applications are available for free or at low cost.

Disadvantages of spreadsheets:

Limited scalability: Spreadsheets are generally not well-suited for managing large datasets or handling complex relationships between data.

Prone to errors: Spreadsheets can be prone to errors, particularly when they are used to store and manipulate large amounts of data or when multiple users are involved.

Limited security: Spreadsheets may not offer the same level of security as databases, particularly if they are shared or stored on unsecured devices.

Difficult to track changes: It can be difficult to track changes to a spreadsheet, particularly if multiple users are making changes simultaneously.

Advantages of databases:

Scalability: Databases are designed to handle large amounts of data and can be scaled up as needed.

Robust: Databases offer more robust features for managing and analyzing data, including the ability to handle complex relationships between data and support for multiple users.

Secure: Databases can offer better security features than spreadsheets, including user authentication and access controls.

Easy to track changes: Many databases offer built-in auditing and logging features that make it easy to track changes to data over time.

Disadvantages of databases:

More complex: Databases can be more complex to set up and manage than spreadsheets, particularly for users without a technical background.

More expensive: Database software and hardware can be more expensive than spreadsheet software.

Requires specialized knowledge: Databases require more specialized knowledge to use effectively, including knowledge of SQL or other query languages.

May be overkill for smaller datasets: For smaller datasets, databases may be more complex than necessary.

Overall, spreadsheets and databases both have their place in data management and analysis, and the choice between them will depend on the specific needs of the user and the dataset in question.