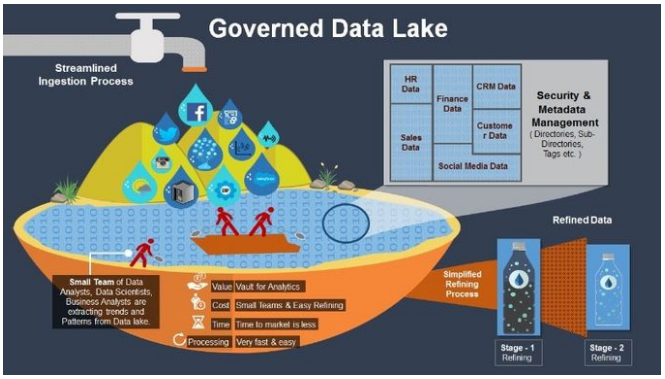
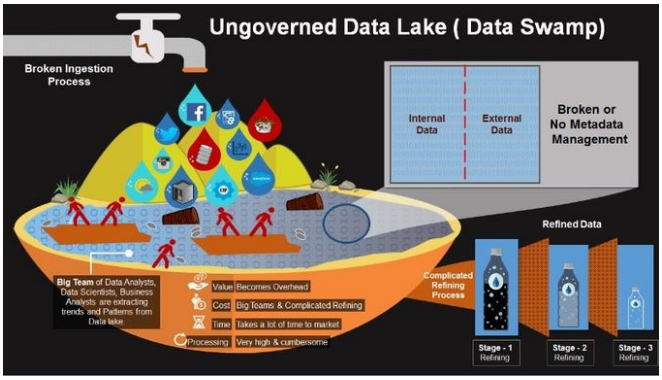
# Preventing a Data Swamp - A Data Governance Process

It is principal to maintain a clean and tidy data lake. In order to prevent a data lake from devolving into a data swamp, there are many steps and considerations we must take. These steps and considerations end up making a set of rules that then become the data governance process that deals with the organization of our data lake. Below are several rules to consider in our data governance process.




**1. <u>Be Selective</u> - Start with data you know you're going to use for a specific project**

Data lakes can hold a substantial amount of data, but it is more important to plan what that data is going to be used for; at least at the beginning. Instead of data dumping, we must build our data lake according to specific needs using clear boundaries. Eventually we look to hold all of our company's data in a data lake, but it starts with specificity and a goal. We must look at what business problem we are trying to address and the goal we are trying to achieve so that we avoid filling our lakes with volumes of unknown and/or unused data.

**2. <u>Load Smart</u> - Load data once and only once, use incremental loads**

A great thing that data lakes can do is load an entire big data file system and its data sets all at once. Then, rather than reloading the data sets when there are changes, we can perform incremental loads that only update whatever changes occurred between the previous load and the new load with in the data sets. This requires identifying just the source data rows that have changed and subsequently merging and synching those changes with existing tables in the data lake.

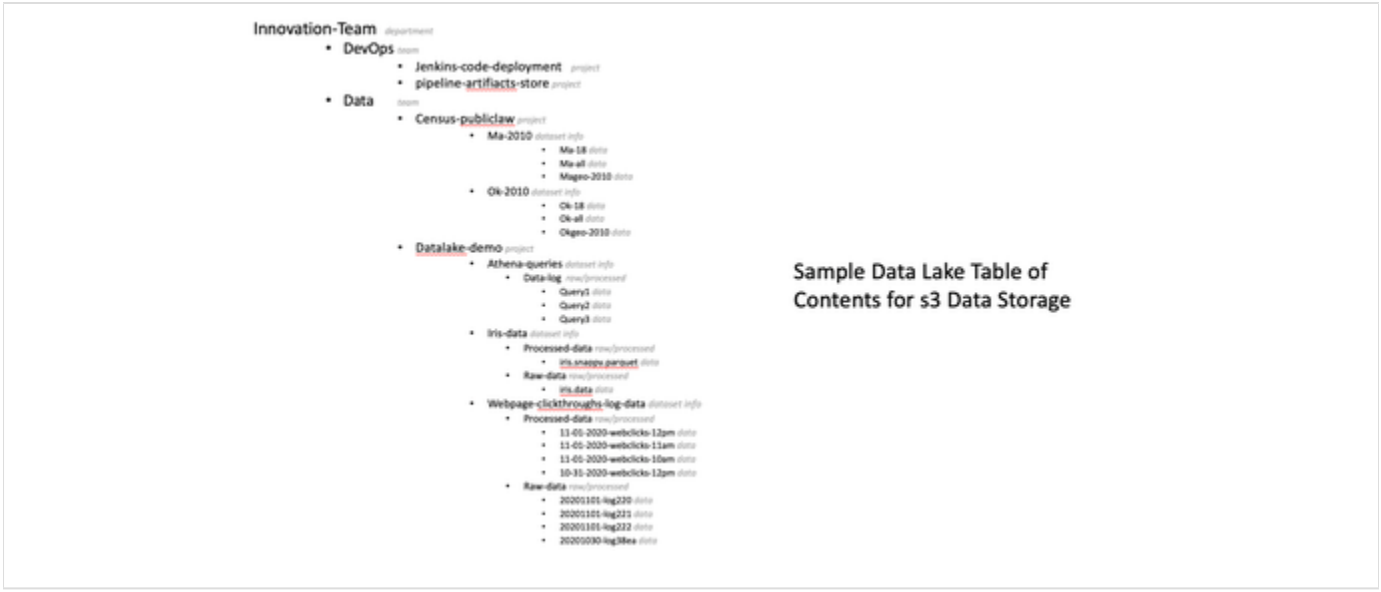
Also, we must make sure a certain team of people are responsible for loading data so that two different people don't load the same data source into different areas of the data lake. This requires strict curation and naming conventions according to a company's size, departments, teams and needs with in the data lake.

**3. <u>Catalog and Curate</u> - Make your data searchable, findable and accessible after ingestion**

One way to help prevent loading the same data more than once, while also making it easy for analysts to find that data, is to catalog it immediately after loading or transforming it. Waiting to do so later is a huge mistake and leads to a cluttered outline and storage of data. This is a very important step in the data governance process of preventing a data swamp.

It is even more important to curate the data by properly gathering and organizing relevant information to specific areas with in the data lake. It may seem grueling at first, but once a curation process is in place, it'll be extremely easy to source and locate data within the data lake. It is crucial to:

- Have strict controls as to what data can go into the data lake and who can access which data
- Specify who manages and develops the processes to sort, describe and catalog ingested data
- Create metadata essential to finding relevant information quickly for insights and applications
- Create department > team > project > data > ingestion method > process taken - specific s3 buckets and folders to keep stored data clean, do the same with other storage tools within the data lake
    - *view sample data lake table of contents image below*
- Implement a naming convention to files that separates raw from processed (can be two separate folders) and sorts by dates and times, especially if automated
- Automate the process, when possible, so that changes are prevented and optimization occurs. This is done by pointing automation tools to specific buckets, folders and storage areas for every single job




**4. <u>Automate the Process</u> - Use the best tools for each type of job**

There are many tools that can automatically describe and tag data as it enters the data lake. Setting up these tools requires a company to invest in training and developing *data curation and metadata managers* in order to keep an optimal and tight pipeline functioning correctly.

- Different processes require different tools
- Different data types require different pipelines
- Different goals require different strategies
- Different users require different outputs

This all goes back to having someone or a team that can properly curate and manage the data lake processes. More process automation means less likelihood of a data swamp.

**5. <u>Have Data Lineage</u> - Document the lineage of data through your governance process**

Data lineage goes back to proper curation and creation of buckets, folders, storage areas and the naming convention of our data. Often other people have already cleaned or integrated the data sets we want to use or work on. Only looking for the raw data will start to create repeat data. Documenting the data lineage, especially through proper curation will help avoid this problem. As mentioned previously, it is important to make note of what actions other people have taken when ingesting and transforming data. That is why we include an ingestion method and process taken folder within our storage areas and have proper naming conventions to know what data we are dealing with and where it came from.

**6. <u>Keep a Good Balance</u> - Find the right balance between speed and quality**

Depending on the size of the company or the needs of the data lake users, some processes may not need to be completely automated, while others can stay raw. Taking in log data from IoT's will take time to set up but needs automation in order to stay optimized because it is dynamic data. Once that data is processed, it can sit in its area of the data lake until a use is found, or it can immediately go to updating dashboards or generating reports. But static, persistent or even incomplete data may not always need automation or immediate ETL and can sit in certain areas with in the data lake until an analyst finds a need for it. It is not the ideal situation, but this is what makes a data lake great; we can store data that we have no immediate use for.

***NOTE*:**

*It is important to consider not naming anything under 'AWS' or even 'A' when creating a bucket because AWS creates automatic buckets from processes and tools for storage or temporary storage with this naming convention.*

---

Sources:


- Four Basic Steps to Prevent Your Data Lake from Becoming a Data Swamp
- Data Lakes - Tips to Avoid a Data Swamp
- How to Stop Data Lakes From Getting Swamped
- 6 Steps To Clean Up Your Data Swamp