**MODULE 4- ORGANIZE AND PROTECT DATA**

**BRING DATA TO ORDER**

[**FEEL CONFIDENT IN YOUR DATA**](https://www.coursera.org/learn/data-preparation/lecture/BJn55/feel-confident-in-your-data)

[**LET'S GET ORGANIZED**](https://www.coursera.org/learn/data-preparation/lecture/93jL6/lets-get-organized)

Hey, welcome back. Whether you're organizing your personal data for your own use or organizing project data for work, there are certain procedures you want to follow to make sure your data is easy to find and use. In this video, we'll cover some best organization practices and also check out some different ways project data can be organized.

There are plenty of best practices you can use when organizing data, including naming conventions, foldering, and archiving older files.

We've talked about file naming before, which is also known as **naming conventions**.

These are consistent guidelines that describe the content, date, or version of a file in its name. Basically, this means you want to use logical and descriptive names for your files to make them easier to find and use.

Speaking of easily finding things, organizing your files into folders helps keep project-related files together in one place. **This is called foldering**.

For example, all the files related to your vacation plan might go in the Vacation2025 folder. You might then break that folder down even further by creating subfolders like itinerary or photos, depending on what else you'd like to easily access. It can also be useful to move old projects to a separate location to create an archive and cut down on clutter. It's so much easier to find and use my files when I name them something meaningful and searchable and when I organize them into folders. It makes all my data more accessible and useful.

In addition to these three best practices, there are two more things you'll want to consider when organizing data for work use.

First, the project data you'll be using for work could be accessed and used by multiple people. **It's important to align your naming and storage practices with your team to avoid any confusion**. Your team might also develop metadata practices like creating a file that outlines project naming conventions for easy reference. We'll get to talk more about naming conventions for work files in more detail later.

Secondly, you want to **think about how often you're making copies of data and storing it in different places**. **Most importantly, because if data is stored in lots of different databases or spreadsheets, it can contradict itself and lead to mistakes later on. Also** **storing data in multiple places takes up a lot of space**.

**Relational databases can help you avoid data duplication and store your data more efficiently**. You can use these practices to organize data in different ways according to your project.

Let's look at some examples of data organization. I have some sample project folders here, each organized in a slightly different way. Let's open them up and see what they look like. We'll start with the high-level Finances folder.

The Finances folder has been organized categorically. There are subfolders like budget, invoices, and payroll that represent different categories. Let's click on "Invoices" to see what's in there. In the invoices folder, you can see that we have another set of subfolders labeled by year, 2014, 2015.... Looks like these are in chronological order. Sometimes the way files are organized can tell us how the data within those files is also organized. Let's open a file to see if that's right. In the 2014 subfolder, there's a file with invoices from June. If we open it, we can see that they've been organized by date, just like the folders. There's different ways to organize data depending on what you need it for.

The categorical organization of the subfolders and finances made it easy for me to go straight to the invoices, but the chronological organization of the invoices subfolder can help us find financial data from the exact date we're looking for.

There's other ways to organize data too: **in order of importance or even by location**.

For example, a company might use hierarchical organization so that employee data mirrors the structure of their employee organization. Or a company working with geographical data might choose to organize by location. It's a good idea to take time early on in a project to consider what the best organization methods will be for you and your team to stick to. Here's another way to think about it. Unorganized data is like a messy room. It's overwhelming, hard to find anything in, and gets worse the longer you avoid cleaning it up. But by making sure early on you know where to put your files, you can keep your work data organized, easy to use, and error free. Now that you see how important it is to keep data organized for both personal and work use, we'll take a closer look at file naming conventions and how they carry over into your databases. See you in the next video.

[**FILE ORGANIZATION GUIDELINES**](https://www.coursera.org/learn/data-preparation/supplement/fLKJI/file-organization-guidelines)

Every data analyst’s goal is to conduct efficient data analysis. One way to increase the efficiency of your analyses is to streamline processes that help save time and energy in the long run. Meaningful, **logical, and consistent file names help data analysts organize their data and automate their analysis process**. When you use consistent guidelines to describe the content, date, or version of a file and its name, you’re using file naming conventions.

In this reading, you’ll learn more about best practices for file naming conventions and file organization.



**Best practices for naming files**

File-naming conventions help you organize, access, process, and analyze data because they act as quick reference points to identify what’s in a file. One important practice is to decide on file naming conventions—as a team or company—early in a project. This will prevent you from spending time updating file names later, which can be a time-consuming process. In addition, you should align your project’s file names with your team’s or company’s existing file-naming conventions. You don’t want to spend time learning a new file-naming convention each time you look up a file in a new project!

**It's also critical to ensure that file names are meaningful, consistent, and easy-to-read.**

File names should include:

The project’s name

The file creation date

Revision version

Consistent style and order

Further, file-naming conventions should act as quick reference points to identify what is in the file. Because of this, they should be short and to the point.

In the following sections, you’ll explore each part of a sales report file name that follows an established naming convention, SalesReport\_20231125\_v02. This example will help you understand the key parts of a strong file name and why they’re important.

**Name**

Giving a file a meaningful name to describe its contents makes searching for it straightforward. It also makes it easy to understand the type of data the file contains.

In the example, the file name includes the text SalesReport, a succinct description of what the file contains: a sales report.

**Creation date**

Knowing when a file was created can help you understand if it is relevant to your current analysis. For example, you might want to analyze only data from 2023.

In the example, the year is described as 20231125. This reads as the sales report from November 25, 2023 following the year, month, and day (YYYYMMDD) format of the international date standard. Keep in mind that different countries follow different date conventions, so make sure you know the date standard your company follows.

**Revision version**

Including a revision version helps ensure you’re working with the correct file. You wouldn’t want to make edits to an old version of a file without realizing it! When you include revision numbers in a file name, lead with a zero. This way, if your team reaches more than nine rounds of revisions, double digits are already built into your convention.

In the example, the version is described as v02. The v is short for the version of the file, and the number following the v indicates which round of revisions the file is currently in.

**Consistent order and style**

Make sure the information you include in a file name follows a consistent order. For example, you wouldn’t want version three of the sales report in the example to be titled 20231125\_v03\_SalesReport. It would be difficult to find and compare multiple documents.

When you use spaces and special characters in a file name, software may not be able to recognize them, which causes problems and errors in some applications. An alternative is to use hyphens, underscores, and capital letters. The example includes underscores between each piece of information, but your team could choose to use hyphens between year, month, and date, too: SalesReport\_2023\_11\_25\_v02.

Ensure team consistency

To ensure all team members use the agreed-upon file naming conventions, create a text file as a sample that includes all of the naming conventions on a project. This can benefit new team members to help them quickly get up to speed or a current team member who just needs a refresher on the file naming conventions.

File organization

To keep your files organized, create folders and subfolders—in a logical hierarchy—to ensure related files are stored together and can be found easily later. A hierarchy is a way of organizing files and folders. Broader-topic folders are located at the top of the hierarchy, and more specific subfolders and files are contained within those folders. Each folder can contain other folders and files. This allows you to group related files together and makes it easier to find the files you need. In addition, it’s a best practice to store completed files separately from in-progress files so the files you need are easy to find. Archive older files in a separate folder or in an external storage location.

Key takeaways

Use consistent, meaningful file-naming conventions throughout your project to save you and your team time by making data easy to find and use. File-naming conventions should be agreed upon by all team members before starting a project and should describe the project by including its name, the date, and the revision version. Document this information in a location that team members can access.

[**EFFECTIVE FILE-NAMING AND ORGANIZATION METHODS**](https://www.coursera.org/learn/data-preparation/ungradedWidget/PIoLM/effective-file-naming-and-organization-methods)

**SECURE DATA**

[**SECURITY FEATURES IN SPREADSHEETS**](https://www.coursera.org/learn/data-preparation/lecture/HO5QC/security-features-in-spreadsheets)

Okay, now that our data's organized and easy to find, it's time to start thinking about how to protect it. The good news is that spreadsheets come with **security features** already built in.

We'll look at different spreadsheet programs and how their security features, like sheet protections and access control, are similar. When I say "**security features**," you might be imagining ways to protect data from other people. But that's just one kind of security.

Security features can be designed to keep unauthorized users from viewing certain files, or just lock your worksheets so that you don't accidentally break your formulas. This is called **data security**. Data security means protecting data from unauthorized access or corruption by adopting safety measures. Whatever spreadsheet program you're using will have similar security measures built in.

As a data analyst, you'll run into Google Sheets and Excel a lot. Let's talk about what they have in common.

First, both programs have features that let you protect your spreadsheets or parts of your spreadsheets from being edited, from the entire worksheet down to single cells in a table. If you're collaborating with other users, you can easily lock down your formulas so that they aren't accidentally broken.

Speaking of collaborating, Excel and Google Sheets both have **access control features** like password protection and user permissions. This gives you more control over who can do what to your spreadsheet. Because these programs are located in different places, these features are slightly different.

For **Excel spreadsheets**, you can encrypt files and worksheets with passwords before emailing them to other users. In Google Sheets, these settings are found under the sharing menu, which allows you to control who can see or edit the sheet online.

**Google Sheets** can also be copied so that users can work with that data without altering the original.

Tabs can also be hidden and unhidden in Sheets and Excel, allowing you to change what data is being viewed. But remember, even hidden tabs can be unhidden by someone else, so be sure you're okay with those tabs still being accessible.

**As a data analyst, data security will be a priority**. But no matter which program you use to create spreadsheets, there's security features to help you keep your work safe and secure. There are some other basic best practices you can take to keep your data more secure overall, which we'll cover later in a reading.

You've made it to the end of this module. Congrats. In these videos, we've covered strategies for organizing data for personal and work use, how to develop functional file naming conventions, and some security measures you can take advantage of in spreadsheets.

Before you move on to the next step in the data analysis lifecycle. It's important that you make sure your data is prepared, and that includes organizing and securing it. As usual after this video, you'll have your weekly challenge. I know you've got this. Then after the weekly challenge, there's some optional material all about connecting to the online data community.

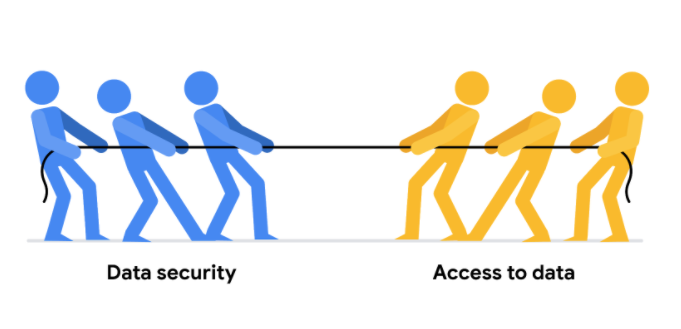
As you start building your career in data analytics, it'll be really valuable to connect with others, learn about new trends in the field and share your own work. I think you'll get a lot out of those videos. That'll help you develop a professional online presence and find ways to communicate with people in your field, which is key as networking becomes more and more online and remote work opportunities become the norm. But if you feel pretty confident about your online presence, you can move into the course challenge instead. Good luck on this weekly challenge, and I'll see you soon!

[**BALANCE SECURITY AND ANALYTICS**](https://www.coursera.org/learn/data-preparation/supplement/CmRzN/balance-security-and-analytics)

**Data security** means protecting data from unauthorized access or corruption by putting safety measures in place. Usually the purpose of data security is to keep unauthorized users from accessing or viewing sensitive data.

**Data analysts have to find a way to balance data security with their actual analysis needs**. **This can be tricky**-- we want to keep our data safe and secure, but we also want to use it as soon as possible so that we can make meaningful and timely observations.

In order to do this, companies need to find ways to balance their data security measures with their data access needs.



Luckily, there are a few security measures that can help companies do just that. The two we will talk about here are **encryption** and **tokenization**.

**Encryption** uses a unique algorithm to alter data and make it unusable by users and applications that don’t know the algorithm. This algorithm is saved as a “key” which can be used to reverse the encryption; so if you have the key, you can still use the data in its original form.

**Tokenization** replaces the data elements you want to protect with randomly generated data referred to as a “token.” The original data is stored in a separate location and mapped to the tokens. To access the complete original data, the user or application needs to have permission to use the tokenized data and the token mapping. This means that even if the tokenized data is hacked, the original data is still safe and secure in a separate location.

Encryption and tokenization are just some of the data security options out there. There are a lot of others, like using authentication devices for AI technology.

**As a junior data analyst, you probably won’t be responsible for building out these systems**. A lot of companies have entire teams dedicated to data security or hire third party companies that specialize in data security to create these systems. But it is important to know that all companies have a responsibility to keep their data secure, and to understand some of the potential systems your future employer might use.

However, one thing you absolutely can do to help strike the right balance is to use **version control** best practices.

**Version control** enables all collaborators within a file to track changes over time. You can understand who made what changes to a file, when they were made, and why.

Here's a simple example: Perhaps you're working on a project with a team of other people. You are all collaborating within the same set of files, but each person is responsible for a different part of the project. Without version control, it would be very difficult to keep track of who made what changes to the files and when. This would lead to confusion and, even worse, people accidentally overwriting each other's work! Version control is essential for data analytics professionals because it allows users to effectively collaborate with others and experiment with new ideas without fear of losing their work.

[**SELF-REFLECTION: PROTECT YOUR RESOURCES**](https://www.coursera.org/learn/data-preparation/quiz/ZUcQW/self-reflection-protect-your-resources)

**Activity Overview**

Empty alt text.

Now that you have learned about the importance of data security, you can pause for a moment and think about what you are learning. In this self-reflection, you will consider your thoughts about data privacy, collaboration, and version control, then respond to brief questions.

This self-reflection will help you develop insights into your own learning and prepare you to apply your knowledge of data privacy to your experience with Kaggle. The data privacy skills you’re developing will apply to many of the tools you’ll use throughout your career as a data analytics professional.

**This is an example of a transferable skill**: your understanding of data privacy in Kaggle can be applied to other tools in the future.

As you answer questions—and come up with questions of your own—you will consider concepts, practices, and principles to help refine your understanding and reinforce your learning. You’ve done the hard work, so make sure to get the most out of it: This reflection will help your knowledge stick!

Step-By-Step Instructions

Follow the steps to learn how Kaggle helps you manage your data. Then answer the question at the end of the activity before going to the next course item.

**Step1: Maintain data privacy**

On Kaggle, you can upload your own datasets and keep them private. This means that they are visible and accessible by only you. You also have the option to add collaborators to your dataset, whom you can add as viewers or editors. Viewers are able to see your private dataset and editors are able to make changes to your private dataset.

You can share the link to your private dataset so anyone with the link is able to view it. If you don’t want this feature, [you can disable it in the Settings tab of your dataset](https://www.kaggle.com/product-feedback/120243).

Note: **If you have a private dataset on Kaggle and you choose to make it public, you will not be able to make the dataset private again**. The only option you would have is to delete the dataset from Kaggle completely.

**Step 2: Connect with collaborators**

**Any notebooks that you create on Kaggle are private by default**. Like in datasets, you can add collaborators as viewers or editors. You can also make a notebook public, which will share it with the entire Kaggle community.

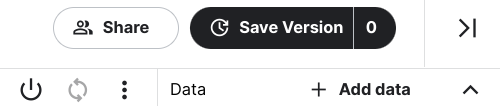
If you add collaborators to your Kaggle notebook, they can make changes to it. You want to make sure you communicate and coordinate with your collaborators because the last person who saves the notebook will overwrite all of the previous work. If you’d like more fine-grained control of changes to your code, a system like GitHub provides more version control.

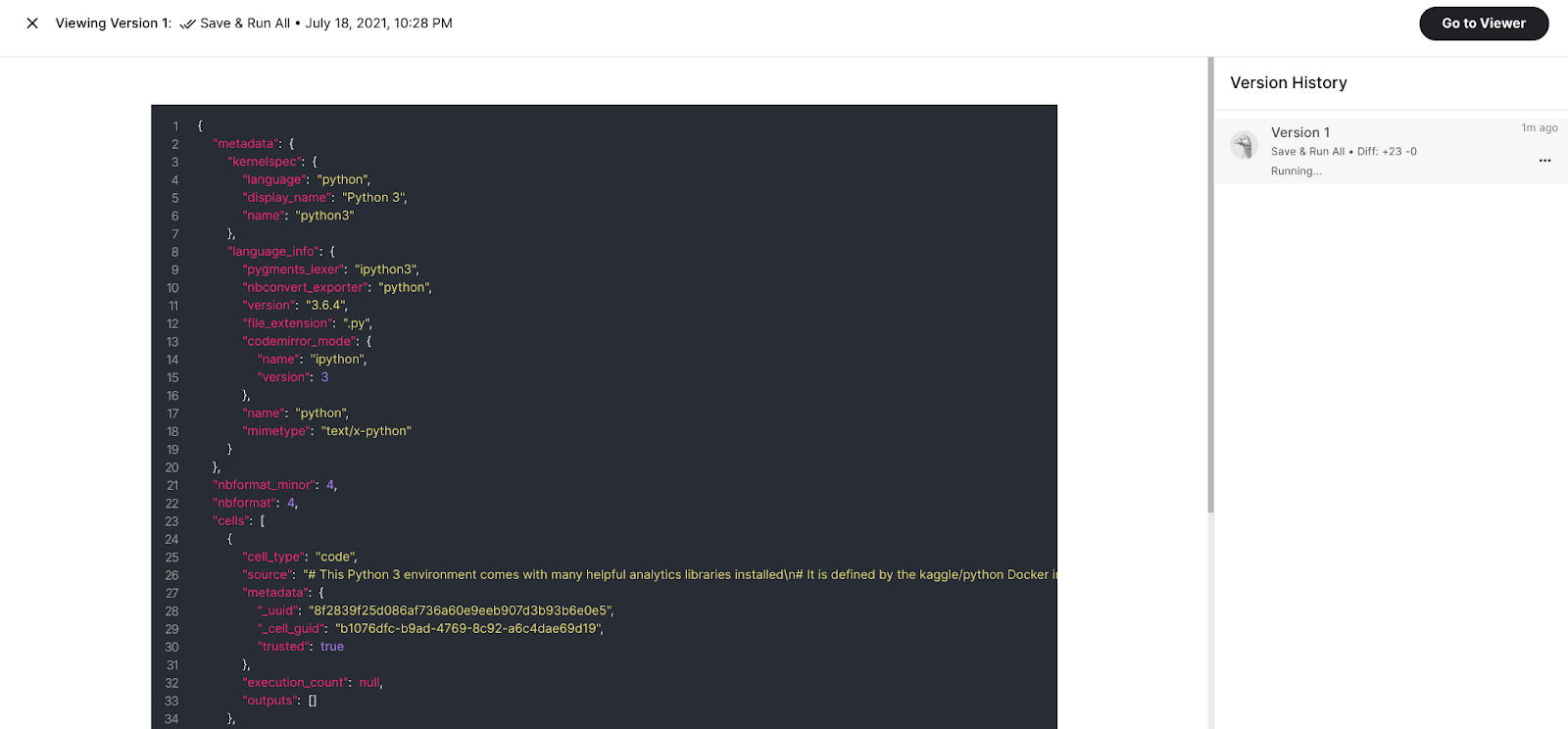
**Step 3: Maintain version control**

As for version control, Kaggle has its own style of letting you keep records of your progress. You can read all of the details [in this post](https://www.kaggle.com/product-feedback/139884), but think back to when you’ve done some work in a Kaggle notebook and clicked on the Save Version button.

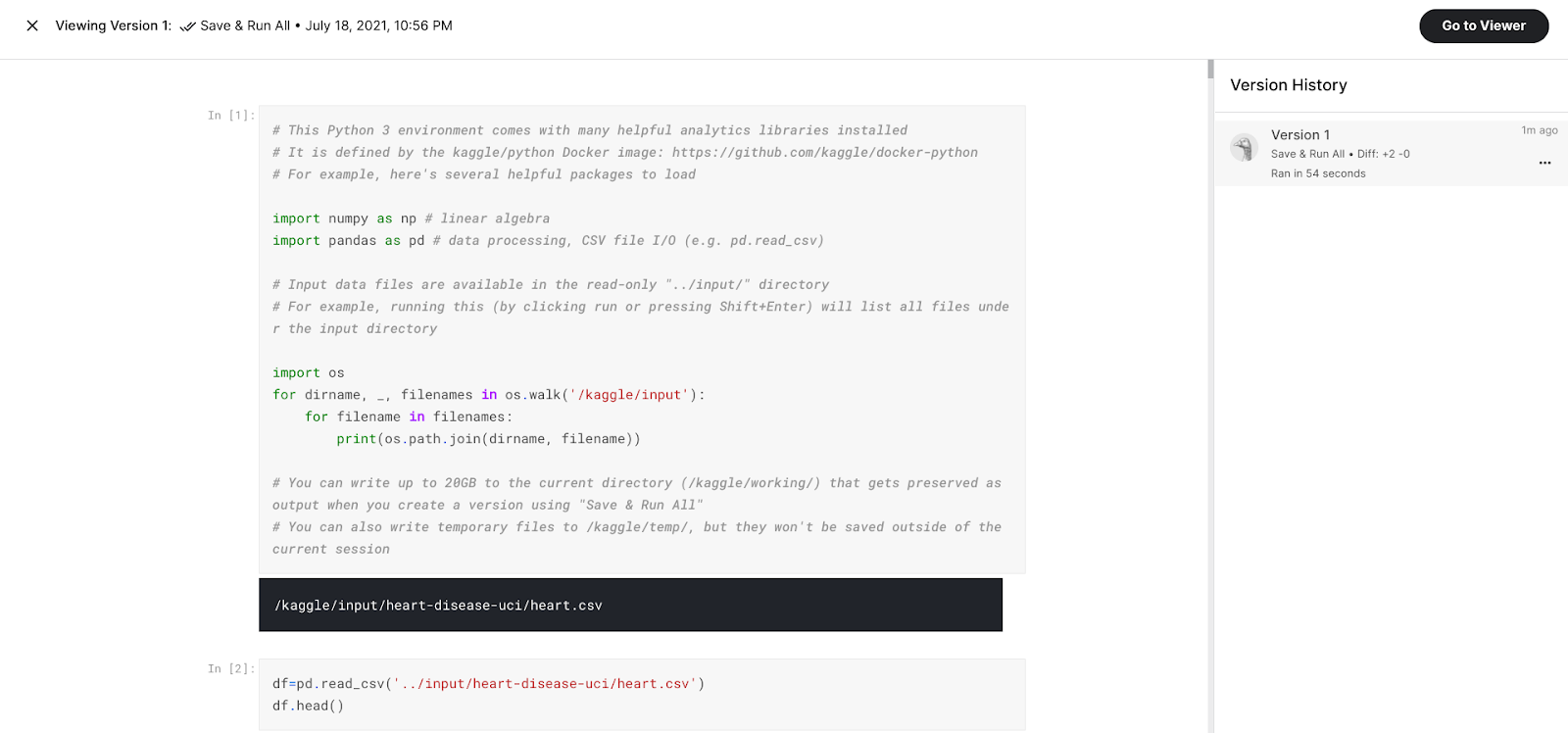
When you clicked this button then clicked Save, you did it without changing anything. But you also have the option to add a short descriptive note about what changes you’ve made.

This can be helpful when you’ve made changes to your notebook but want to go back to an earlier version. To do this, go to Edit mode and click on the number next to the Save Version text at the top of your notebook.



This will open a navigation bar on the right side of the screen and list out all of the versions of your notebook. When you click on different versions of your notebook, the left side of the screen will populate with the code and text from that version.

Then, once the version has run, your screen will appear like this:



From this screen you can also open the version in Viewer mode, pin a version as the default, or even change the version name. Pinning a version as the default can be helpful when you have a working version of your notebook available to the Kaggle community, but want to make changes and updates that might not work the first time you implement them. This allows you to safely make changes behind the scenes while sharing with the Kaggle community the most recent working version of your notebook.

**Reflection**

Empty alt text.

Consider what you learned about data security in Kaggle:

What are some cases in which you should use the privacy, collaboration, and version control features on Kaggle?

What other scenarios can you think of where you might want to pin a different version of your notebook other than the most recent version?

**[TEST YOUR KNOWLEDGE ON SECURING DATA](https://www.coursera.org/learn/data-preparation/quiz/konVV/test-your-knowledge-on-securing-data)**

**MODULE 4 CHALLENGE**

[**GLOSSARY TERMS FROM MODULE 4**](https://www.coursera.org/learn/data-preparation/supplement/BUNI2/glossary-terms-from-module-4)

**Terms and definitions for Course 3, Module 4**

**Access control:** Features such as password protection, user permissions, and encryption that are used to protect a spreadsheet

**Data security:** Protecting data from unauthorized access or corruption by adopting safety measures

**Inbox:** Electronic storage where emails received by an individual are held