CHAPTER

Data Management and Visualization

10

INFORMATION IN THIS CHAPTER

- Data
- Information
- Intelligence
- Data management
- Data visualization

INTRODUCTION

Till now we have learned about gathering data using different methods. Generally people think that open source intelligence OSINT means collecting data from different internet-based open source options. But it's not limited to that because if the data we collected from different sources are not categorized properly or we cannot find relations between one another it can be just a huge amount of random data that is of no use. We will discuss later the need of managing data and analyzing its worth, but for the time being let's refresh what we learned so far and how to collect different data using different sources.

From the very beginning we have focused on data extraction using different methods. We started with search engines, where generally a normal user gets answers for all the questions, and we also discussed how that is just a minute part of the web as popular conventional search engines have only a limited amount of internet indexed in their databases. So we learned how to use other specific search engines to get specific data. Some popular features of mainstream search engines that make them unique as compared with other. Further we learned about some interesting tools and techniques to find out data which is openly available. Later we moved on to power searching and learned how to get desired information from the web effectively. Then we moved to metadata and how it can be helpful. We learned how to get all the metadata information and how can we use it for different purposes and last but not the least we covered Deep Web, the part of web which is not directly indexed by conventional search engines. We learned how to access it to get more information.

So for the time being we can say that we learned how to collect data from different sources directly using some well-known solutions and also using some

Hacking Web Intelligence. http://dx.doi.org/10.1016/B978-0-12-801867-5.00010-0

unconventional tools that open even more doors to collect data. Before going any further, let's discuss a bit about what is data, information, and intelligence and how they differ from one another.

DATA

"Data" is one of the most commonly used terms in any domain, especially IT. If we describe data in simple words it means the RAW form of entities. It is the depiction of facts in a basic form. For example, if we get a text file that consists of some kind of names abc.inc, xyz.com, john, 28, info@xyz.com, CTO, etc. We can see there are certain entities we found but have no meaning. This is the raw form. In its original form, data do not have much worth.

INFORMATION

The systematic form of data is called information. When data is categorized based on the characteristics it can be called as information. We can say that aggregated and organized form of data is information. Hence to achieve information we need to process data. Let's take the same example abc.inc is a company name. xyz.com is a domain, john is a username, 28 is age, info@xyz.com is an email address registered with xyz.com, and CTO is a position.

INTELLIGENCE

When we relate different information based on their relations with one another and derive a meaning out of that, then what we get is called intelligence. So we analyze and interpret the information at hand according to the context to achieve Intelligence. From same example we can derive that xyz.com and info@xyz.com belong to same domain. It is quite possible as john is 28 year old and is the CTO of abc.inc. These are primary predictions they may be false positives also, so we need to validate the same later but for the time being the information that we have looks like relative so we can conclude that. John who is 28 years old is the CTO of abc.inc and the domain name of the same company is xyz.com and email id to communicate is info@xyz.com.

To validate we will need to extract information from other sources and we might get to know that the name of the CTO of abc.inc is someone named John and there is a John who works at abc.inc whose email is info@xyz.com and similar information which might correlate to prove our theory right or wrong. Now let's say we are a salesperson and our job is to contact management of different companies, then the validation of this information being right allows us to contact John and craft an email depending upon his age group and other information about him that we might have extracted.

The definition of intelligence may differ from different people. This is our own definition based on our experience, but the bottom line is it's about the usefulness of

the information. Unlike data, which states raw facts, actionable Intelligence allows us to take informed decisions.

As we discussed earlier also, data is the raw form which just contains the entities. Entity means anything tangible or intangible. It may be name, place, character, or anything. If it is just a data it is worthless for us. We do not know what it is about. We can get lots of random data but for using that we must understand what that data is all about. Let's again take another example, we got 1000 random strings, but what to do with that. But if we come to know that those are some usernames or passwords then that 1000 random strings are worth a lot. We can use that as dictionary attack or brute force etc.

It's not the data that is always valuable, it's the information about the data or the information itself that is worth a lot.

Managing data is very important. Managed data can be quickly used to find relationships. Let's say we have got lots of data and we know that the data consists of name, email id, mobile number etc. If we will not manage that systematically in rows and columns, we will lose track of that and later when we need a particular set of data, let's say name, it will be difficult for us to differentiate and fetch from large amount of unmanaged data.

So it's always important to manage the data by its types in a categorized manner so that later we can use the same quite easily. As seen in previous chapters there are various sources of information. Every source is unique in its own way. When all the data from different sources comes together it creates the complete picture and allows us to look at it from a wider perspective.

As we covered Maltego earlier, we have seen how it collects data from different sources, but even then there are many other sources. The thing to focus here is that it's not about running a tool and collecting all the data. It's about running transformations one by one to get desired final result. To extract data from different sources, correlate it and interpret it according to our needs. it's not possible in most cases that we will be able to get all the data we want from a single source. So we need to collect different data from multiple sources to complete the picture.

For example, let's take a condition that we need to collect all the data about a person called John. How to do that? As John is quite common name, it is very difficult to get all the information. Let's start with some primary information. If we can identify the picture of John then we might start with a simple Google search to check the images, we might or might not get his picture, if we get the picture visit the page from where Google fetched this picture to get to know more about John but if not then simply try social networking sites like Facebook or LinkedIn, there is a chance that we can get the picture as well as the profile of John in one of the social network sites or all. If we get the profile then we can get more further information like his email id, company name, position, social status, current city, permanent residence.

After getting those details we can use the email id to check what other places it is used, such as any other sites, blogs, forums etc. There are different online

sources which can help us get these details and we have discussed some in previous chapters. Then we can manually visit those sites to gather more information. This is just an example how to collect different data from different steps by taking one output as another step input and collect all the data to complete the picture.

As from the above example it is clear that the enumeration or data collection process is a stepwise process which includes number of different sources and also the result of one process is generally used as the point of enumeration for other process. So the relationship between the data enumerated or collected is very difficult to track, if we won't work on it from the very beginning. If we are looking for some specific data flow without sorting it in a structural manner it will be very difficult to search for relations among the data. Then it will be worthless to get all the required data about a single entity without knowing what are those all about. So hence we need to manage data in structural manner.

There are different ways to structure the data collected but the best way is to sort them according to parent and child entities. Let's say we found email from name, then name will be parent entity and email is it's child. If we get something from email, like domain name then that email will become parent entity, likewise we can organize the whole data collected.

Data can be structured using rows and columns in a spreadsheet quite easily but the tracking process will become little difficult. For getting the every parent node we need to search for the shell number. And the text form of data is not that easy to remember. That is the only reason there are graphs, flowcharts, smart arts used to define complex processes or statistics so that it will be easy to remember as well as understand. And this is also a major reason why Maltego is popular in its segment. As Maltego provides easy to understand visualized output, it proves the importance of visualized data that can help in easy analysis of the relations and crystal clear about the established relations between entities.

DATA MANAGEMENT AND ANALYSIS TOOLS

EXCEL SHEET

Before jumping into data analysis and visualization let's not avoid the fact that there are other simpler ways to categorize the data and are used in industry for a long time. One of such tools is excel sheet. Some also call it as spreadsheet. The easy user interface differentiating rows and columns in a tabular way is a great way to categorize the data. It's also very handy when we just have static values for different entities such as a user detail. Let's say we want to manage data that consists of username, email id, organization, and position. For every user in a row we will add all the details in columns. And for this type of job, excel or spreadsheet is the best feasible option.

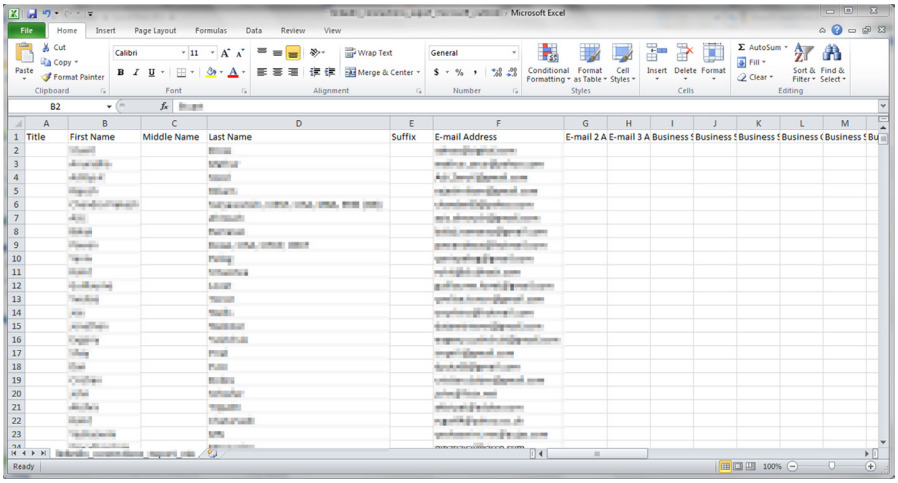

FIGURE 10.1

A sample excel sheet.

Excel has great formatting features where we can apply formulas, filter data, add comment, create dropdown buttons, and many more. The final thing is that excel is a great tool to categorize the data if we are categorizing all for a primary entity, but we might face problem when we have more than one primary entity. In that case either we need to create different tables in a single sheet or have to go for a new sheet for another table. Then we have to manually track down the relationships between data by switching the tables or sheets. This is definitely a problem for huge amount of data which contains more than one primary entity.

SQL DATABASES

SQL also known as structured query language designed for managing data. Using SQL databases we can store data in tabular format in a system. It allows us to insert, query, delete, and update database elements. Though it's not just what it looks like. SQL databases have great features of data management and are widely used in industry to store lots and lots of data. The only reason it is being popular in the industry is that using simple queries we can able to manage the database.

We discussed about the problem in above paragraph that when there are multiple tables then it's very difficult to relate and manually fetch data for a particular entity. Here SQL comes as a savior. In SQL by writing simple queries we can fetch data of a particular entity from multiple tables quite easily. There are lots of DBMS or database management softwares are available. Some are open source and some are not, but some popular DBMS are MySQL, MSSQL, SQL Server, Oracle etc. Apart from the SQL-based databases there are also some NoSQL databases which allow to store more than just text.

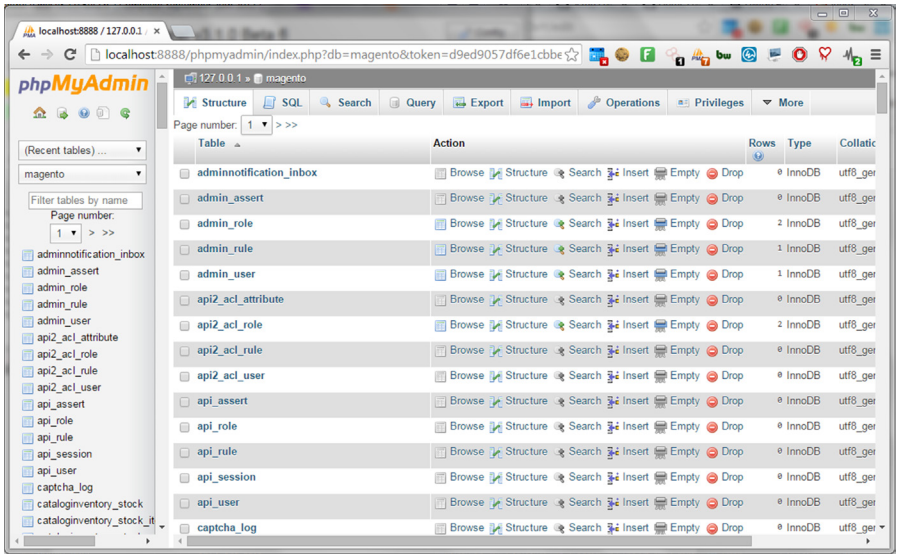

FIGURE 10.2

A sample MySQL database accessed through phpMyAdmin.

The only disadvantage is that, it cannot be used by an average user. Though it is not very difficult to install and configure yet it requires a bit of technical knowledge and also SQL query knowledge to manage the database properly. Though there are certain tools and frameworks available which provide features like command auto-complete, syntax correction etc., for providing ease of use to the user who just do not want to open a dark black screen whether that is a terminal or a command prompt or any kind of console but still the understanding of language limits its use to corporate sector. They are actually designed to hold and work with large amount of data and hence are not much used by people for their normal data storage needs.

FLOWCHARTS

As excel sheet and SQL databases store data in text form or we can say just text form which only categorizes the data, but flowchart adds graphics to the data. There are certain chart symbols available for different types of data and it allows a user to not only manage data graphically but also providing feature to easily show their relations by different symbols and arrows.

It is a type of diagram that represents a set of data, workflow, or process, showing the steps as boxes of various kinds, and their relations by connecting them with arrows. This graphical representation can be very helpful as we discussed earlier for ease of understanding and keeping things remembered. Flowcharts are used to analyze and manage different data and it has special boxes for different purposes. Some examples are given below.

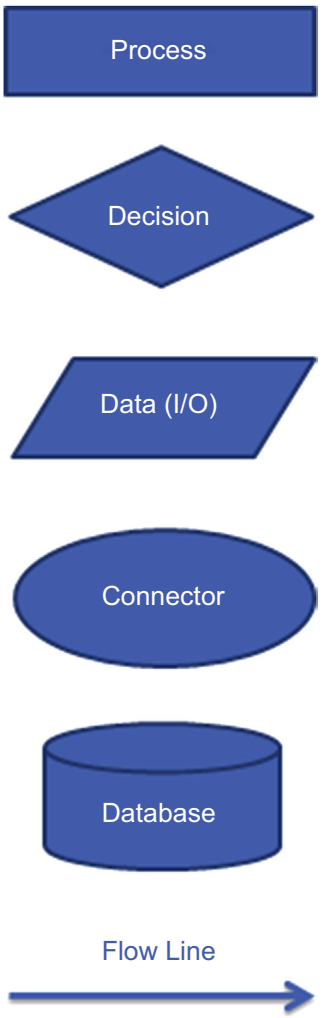

FIGURE 10.3

Some commonly used flowchart symbols.

We discussed a bit about the methods which are usually used for data storage and/or management. Now let's move on to learn about something different than the usual stuff and see what other great options are available out there which can help us with our data management and analysis needs.

MALTEGO

Any open source assessment is not complete without the use of Maltego. It's an integral part of OSINT. We already discussed a lot about this tool earlier and discussed how to

utilize it to extract data using it. The reason Maltego is very popular and widely used is not just it's cool features but also its data representation. Maltego has different set of views such as main view, bubble view, and entity view. And we can also change the type of view. The result looks very easy to understand as different types of icons are used for different types of entities and their relations are well expressed by the arrows.

Maltego represents all the information is a nice and easy to understand entity–relationship model. Apart from extracting data using various transforms and machines we can also take the data we have found from other sources and include it into the graph to create a bigger picture. For that we simply need to take the appropriate entity type from the left entity bar and bring it into the graph, then insert the data we have found and simply connect it to the relevant entity or entities. If we don't find the appropriate entity for the datatype, Maltego allows us to create a new entity and use it according to our needs. This makes it very easy to take the advantage of the data mining feature and further extend it for data analysis purpose.

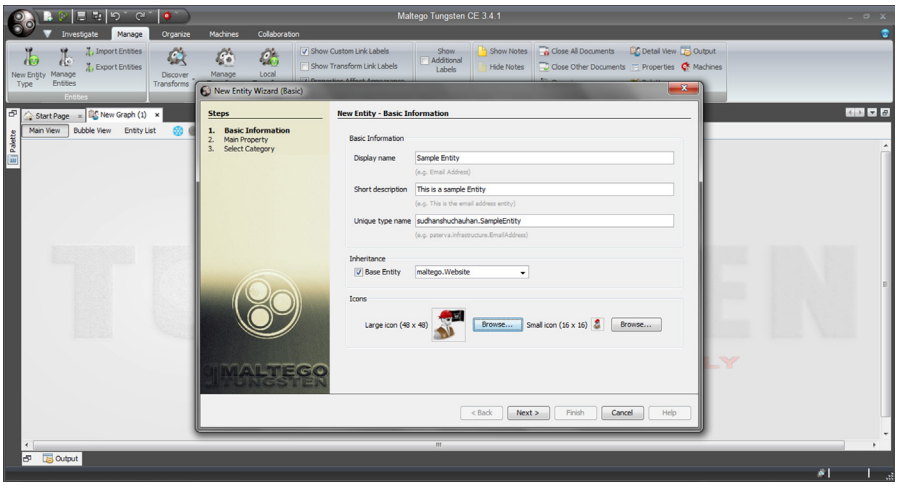

FIGURE 10.4

Maltego entity creation.

CASEFILE

CaseFile is another great tool from Paterva which can be found on https://www.paterva.com/web6/products/casefile.php. Similar to an original CaseFile used by investigators to collect all the information related to any case at one place, Maltego CaseFile also helps us to build an information map which connects all this information in one place along with the connection between them. It provides us simple interface to add, link, and analyze data quickly and effectively. The motive behind creating this tool is to provide offline data analysis. There are certain jobs which include lots of field work and real-life information gathering such as investigators

etc. So to help them which something as cool as Maltego visualized data representation, Paterva came up with this tool which is worth using.

As it is focused for offline data analysis, it contains lots of daily life entities that can be found mostly during information gathering exercise. Some of the entity categories are devices, locations, tracking, weapons, infrastructure etc. It also provides user the freedom to add customized entities. Another exciting feature of CaseFile is that it can be used to visualize data stored in excel sheets or CSV files, which make it easier to make sense of the data we have acquired in those forms.

The interface is very similar to the Maltego and as it is very popular in OSINT segment and widely used, if any Maltego user wants to try CaseFile he/she can use it without any problem. The options to create visualizations are also quite the same.

Similar to Maltego, to create a new graph in CaseFile we need to take all the entities we have information about, or add new entities and input relevant data into them, then further make the connections among those entities to create a complete picture. Though CaseFile does not have any data extraction feature yet its data visualization feature proves to be very helpful when we have data from a variety of sources and needs to be connected together to make sense out of it.

CaseFile contains of three tabs placed at the top of the interface named Investigate, Manage, and Organize. Under the Investigate tab there are functions like cut, copy, paste, entity selection, and graph zooming. It allows to quickly work with the graph and play with the entities around. Another tab is the Manage tab which allows to add new entities and to manage existing ones. We can also add notes and work with features related to the CaseFile window. The last tab is the Organize tab which provides features like managing the layout of the graph into different structures and also performs the alignment according to the requirement.

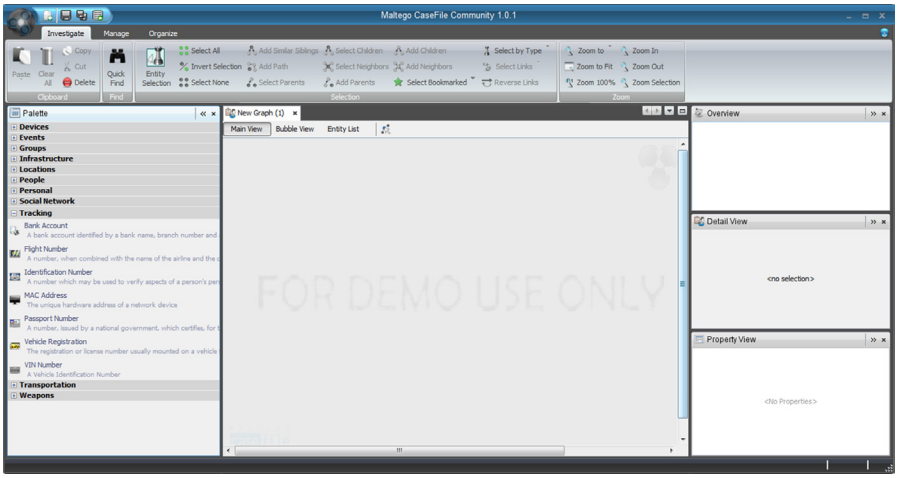

FIGURE 10.5

CaseFile entity palette.

So this is all about Maltego CaseFile that we can use for different purposes starting from investigation to data analysis. Like Maltego, CaseFile also comes in two forms, one is community or the free version and another is the commercial version. Both the versions can be found on https://www.paterva.com/web6/products/download2.php. It also supports operating systems like Windows, Mac, and Linux. The installation process is quite easy and similar to Maltego.

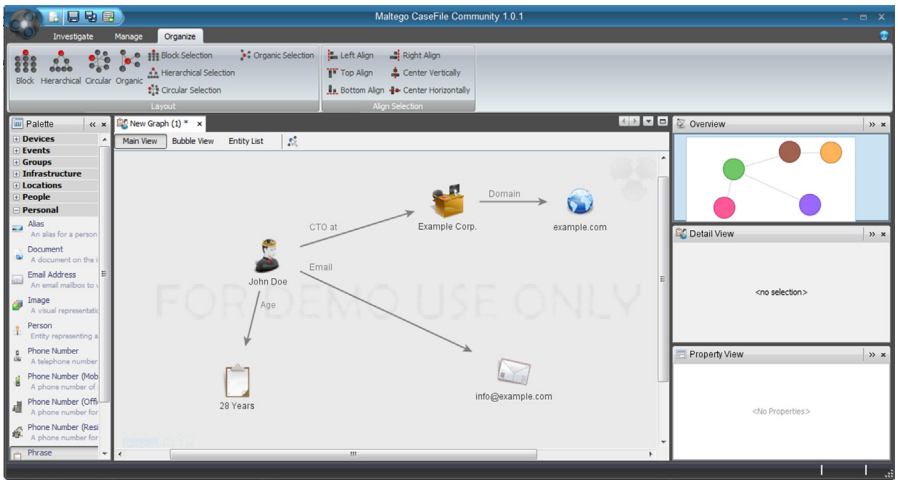

FIGURE 10.6

A sample CaseFile graph.

MagicTree

This tool is basically for pentesters who need to manage the data they get during the security testing. MagicTree http://www.gremwell.com/what_is_magictree is made to solve some problems that every pentester faces in their job and that is finding details from a large set of data generated from different tools. During a penetration testing exercise, pentesters look for loopholes, or possible risks, or vulnerabilities in any application or network so that they can be patched. In this process, lots of tools are used to automate the testing process and the result of those tools is huge depending on the scope, size, and number of loopholes found. Generally in any network penetration testing, pentester faces problem because the scope is always large and the tools used always provide a huge amount of result. In that case MagicTree comes as a savior. It supports popular network penetration testing tools like Nmap and Nessus, and it allows its users to import this data. Later the same data can be queried, analyzed, or used to generate a report using MagicTree.

To use MagicTree first we need to download and install it. We can download the same from following URL: http://www.gremwell.com/download. This is a jar file so can be used in any operating system but with java installed. So to start working with

MagicTree simply open it, add some network address or host address to the scope so that MagicTree will able to build a data tree for the same. The advantage of storing data in tree form is that if later we want to add some other data it will not affect the tree, we just need to create a new tree. It stores the data in tabular or list form and uses XPath expression to extract data. There are many report templates that can be customized and used for report generation.

The only limiting feature of this tool is that it only supports import option for xml. So we cannot add tools which generate text output. Although it is limitation but still this tool is pretty helpful for workflow automation for data retrieval from any tool, and also highly recommended for pentesters.

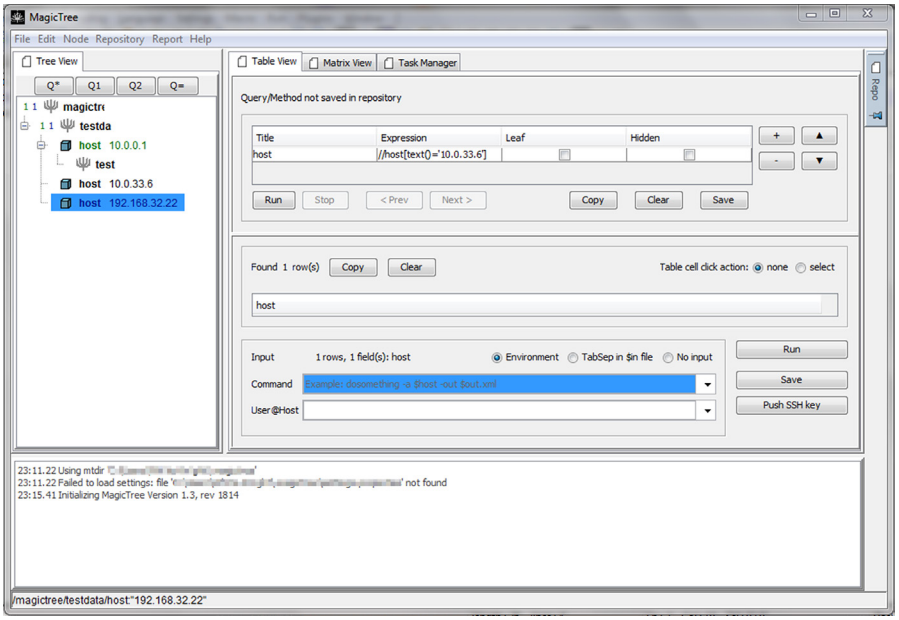

FIGURE 10.7

MagicTree interface.

KeepNote

As the name suggests KeepNote is a note taking application. It is a cross-platform application which can be downloaded from http://keepnote.org/. Unlike traditional tools for note making such as Notepad, KeepNote contains of various features which make note taking more efficient and allows to include multiple media into it.

To start note taking using KeepNote we need to first create a new notebook from the File option. Once a notebook has been created we can add new pages and sub-pages into the notebook. Now in these pages we can keep our notes and keep them categorized. We can simply write the text into the bottom right part of the interface.

Apart from that we can also include different supporting media such as images to make our notes more informative.

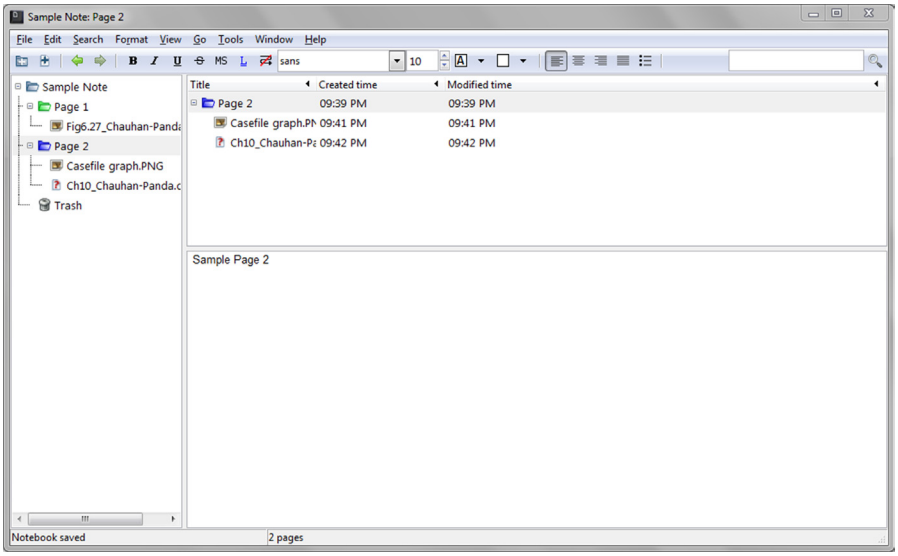

FIGURE 10.8

A Sample MySQL database accessed through phpMyAdmin.

Some of the useful features of KeepNote which other similar applications lack are organized hierarchy, spell check, media attachment, interlinking etc. Apart from these the application also has many extensions which allow to enhance the already rich features of it. Extensions can be found at http://keepnote.org/extensions.shtml.

LUMIFY

One of the best options available for data visualization and analysis in the open source domain is Lumify. As Lumify is open source, its code is available at https://github.com/lumifyio/lumify. There is an easier way to use and test Lumify and that is through the preconfigured virtual machines, which can be found at https://github.com/lumifyio/lumify/blob/master/docs/prebuilt-vm.md.

Lumify presents us with an easy to use web-based interface. Based on the graph-based model we can aggregate our data into the interface and perform analytical operations over it. There are several entities with data field which we can use to represent the information we have. Apart from this, it also provides advanced features like Map integration, using which we can represent the data over world map; live shared workspace, which allows to share information with other team members in real time and work in a collaborative manner; customization, which makes it easier to create own data model etc.

The wide range of features and ease of use makes Lumify a great choice for our data visualization and analysis requirements. Some good examples of Lumify usage can be found at their homepage http://lumify.io/.

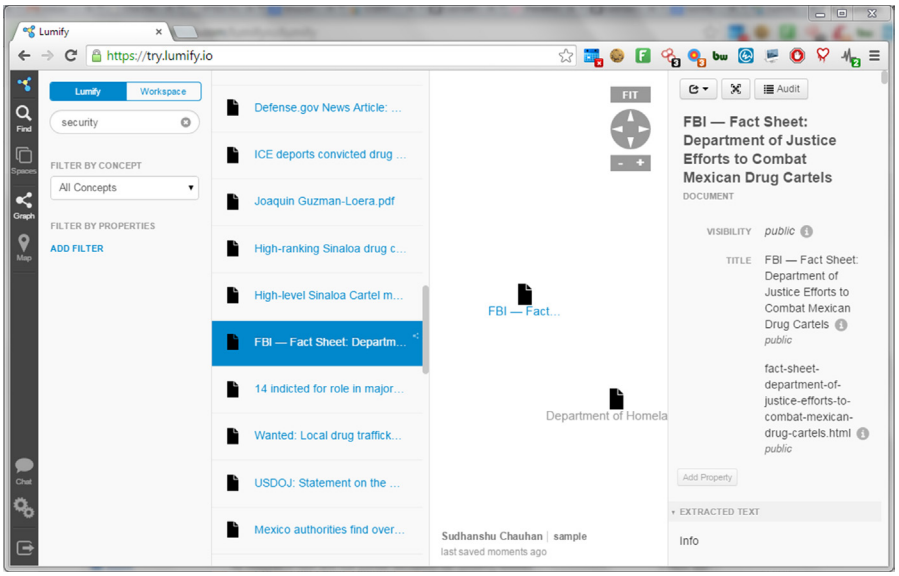

FIGURE 10.9

Lumify interface.

Xmind

We have seen some tools to organize data for better understanding. There is actually a term used for visually organizing information and it is called as mind map. As the name suggests a mind map is map of our ideas or thoughts about any topic. Usually a mind map is started around a central idea and then expanded from it further. The central idea remains the theme of the diagram and all the information revolves around it. It is actually chain of associated ideas and information spreading in different directions. Using branches the subcategories are created out of the main branches and similarly expanded further. Mind maps include different forms to represent the information or idea such as images, text, colors, shapes etc.

One of the most famous and efficient tools to create mind maps is Xmind. The download link can be found at http://www.xmind.net/download/. Once we open up the interface of Xmind, it provides us with a huge list of templates and themes from which we can choose the one which best suits our needs. Once we have made the selection, we can start by editing the data fields, changing or adding new ones. Xmind allows us to insert information in the form of text,

image, marker, summary, attachment, audio notes etc. The variety of data types allowed by Xmind makes it very easy and effective to create a mind map which can actually translate our ideas into a visual representation. We can create mind maps to create diagrams for project management, planning, decision making etc.

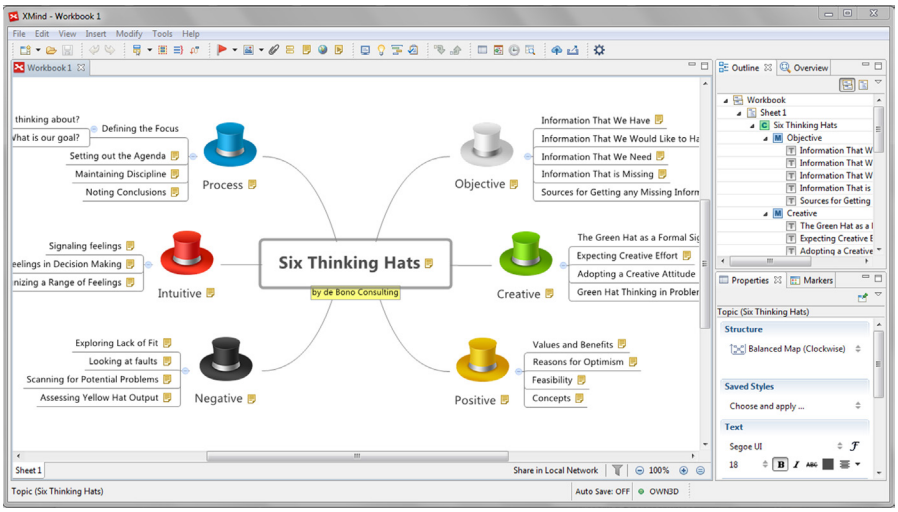

FIGURE 10.10

Xmind sample template.

Though the free version of Xmind has some limitations as compared to the pro version, yet it provides ample ways to visualize our ideas in a creative and effective manner.

There are various models and methodologies which are used in different domains for data analysis process. Some are generic and some only fit to certain industries. Here we are giving a basic approach which applies generically and can be modified according to the specific needs:

- Objective: Decide what is the question that needs to be answered.
- Identify sources: Identify and list down the sources which can provide data related to our objective.
- Collection: Collect data using different methods from all the possible sources.
- Cleaning: From the data collected, anything that is irrelevant needs to be removed and the gaps present need to be filled.
- Data organization: The cleaned data needs to be organized into a manner which allows easy and fast access.
- Data modeling: Performing the modeling using different techniques such as visualization, statistical analysis, and other data analysis methods.
- Putting in context: After we have performed the analysis of data we need to interpret it according to the context and then take decision based on it.

Unlike other chapters where we focused on data gathering, here we focused around data management and visualization. Collecting data is important but managing it and representing it into a form which makes the process of analysis easy is quite important. As we learned earlier in this chapter that raw data is not of much use, and we need to organize it and analyze it to convert it into a form which is actionable, the tools mentioned in this chapter assist in that process. Once we have analyzed the data and put it in context, we will achieve intelligence which helps us to take decisions.

Moving forward in the next chapter we will discuss about online security. Day by day cyberspace is becoming more insecure. New malwares keep on surfacing every now and then, attack techniques are advancing, scammers are developing new techniques to trick people etc. With all this around there is so much that we need to safeguard ourselves from. We will discuss about tools and techniques to shrink this gap in our security and will learn how to minimize our risk.