Transcript for Session 051

Listen to the podcast session, see resources & links:

http://chandoo.org/session51/

Transcript:

Hi and welcome to http://chandoo.org podcast session 51. First up, let me wish you a very happy new year 2016. I know I have been late in sending these wishes your way but this is the first podcast of 2016 and so this is the first opportunity. http://chandoo.org podcast is dedicated to making you awesome in data analysis, charting, dashboards and VBA using Microsoft Excel.

In the 51st episode of our podcast, we are going to talk about **VLOOKUP FAQ's**. We are going to discuss frequently asked questions about VLOOKUP. I run many live classes and online training programs on Excel. Whenever I talk about VLOOKUP people are really happy to discover this function if they haven't used it before. Even if they have been using it for a while, they suddenly see new possibilities for it. At the same time, they also ask me a lot of questions. They want to know what it can do, what it cannot do, what it would do in weird situations etc. So, in this podcast, I am going to discuss the **Top 10 FAQ's about VLOOKUP** and share some resources with you so that you can use VLOOKUP better and become a master of VLOOKUP in your day to day work.

First up, let us start by talking about VLOOKUP. **What is VLOOKUP?** I assume that some of you are familiar with this function of Excel but, for the sake of everybody, let me give a very brief introduction about VLOOKUP. VLOOKUP **provides search functionality into your data** but, instead of manually searching for data and finding information about it, you would use a function so that it becomes automatic. Let us say that you are looking at employee data. You are in the Human Resources department and you are looking at all the employee data and you want to know the date of birth of a certain employee because you need to know the date of birth for some calculation, and based on that, you will declare the bonus amount or annual vacation accrual or whatever. So, you know the employee name and you want to find out when John Doe's date of birth is. How would you do that? In plain real life, if you have the employee records printed on a heap of pages in front of you, you would go and look up the name John Doe in that heap. If the heap is arranged alphabetically, you would immediately jump to the page that contains J and start looking up for John Doe. On the other hand, if the employee records are not arranged alphabetically but by some other order like the date of joining or some random order, you would essentially read from the very first page all the way to the last page until you find John Doe. So, you are trying to find the person John Doe so that you can recover additional information about that person. This exact task is what VLOOKUP does but, instead of doing it very slowly which is what happens when we do it manually, VLOOKUP does the following - let's say you type the name John Doe in

cell A1 and you simply write a formula to VLOOKUP the value in A1 in the range of data that contains the employee details. Let's say the data is in a table then you would simply say:
=VLOOKUP(A1,Employees,4,)

Here, we are assuming that the date of birth data is in the 4th column of the table. VLOOKUP does this job for you and comes back and gives you the answer. This is in a nutshell and a very brief way what VLOOKUP does. It is **looking up the information vertically** because your data is supposed to be arranged in a vertical fashion. Employee names go down in a column. It does a vertical lookup and finds the value that you want to look up and gives you additional data about that value.

There is more to the VLOOKUP function. I have tried to pare it down so that we can talk about introduction to VLOOKUP in a minute or two. If you are keen to learn more about VLOOKUP which I highly recommend - this is a cornerstone function - and for anybody serious about analytics or data visualization or any kind of Excel heavy duty stuff - I highly recommend that you spend some time, maybe a few hours or even a week or two, learning the VLOOKUP function. There is an excellent resource for you podcast listeners as a good start to check out the very **2nd episode** of http://chandoo.org podcast. You can access it at http://chandoo.org/session2/ and that is our very 2nd episode that I entitled **VTALKUP** as a word-play on VLOOKUP. You can listen to that podcast. You can just scroll down the play list of our http://chandoo.org podcast in your podcast app, locate the 2nd episode and play that and get a more thorough introduction to VLOOKUP. So, consider this episode as a sequel to the 2nd episode. In this episode we are going to talk a little bit more about the intricacies of VLOOKUP and some advanced scenarios.

Now, I am assuming that you know what VLOOKUP is. So, let's move on and talk about the top 10 questions that people ask me. These are the questions that I could think of but there are probably more questions. So, at the end of the podcast, I will ask you that if you have additional questions then please submit them as a comment on the podcast page so that I can reply and share my answers and use that as additional material for probably another sequel for the VLOOKUP show that we are doing.

The very first question that people ask when they learn about VLOOKUP is - **what happens when the value is not found.** We are talking about John Doe and we are assuming that John Doe is in our employee list. But, what if the name that I am looking up is not there? What does VLOOKUP do in that case? Well, what would you do in real life when you try to find something and you don't find it? In real life we are more graceful and so we look up the name John Doe in the heap of papers and when we can't find it, we simply come back and say that we can't find the value. Now, this is more or less similar to what Excel does, in most default scenarios, if the value you are looking up is not there in the data. VLOOKUP comes back and throws an error. Errors are the computer's way of saying that we are asking it to do something that it doesn't understand. In this case, you are asking Excel to lookup John Doe but the name John Doe is not even there in your data and so Excel does not know what to do and it comes back and throws an error. This error is shown as #N/A. This is what you get when you try to do a lookup and

VLOOKUP does not find the answer. This is a very simple answer to this question. At the end of this FAQ, we will talk a little bit more about what to do in the case of such errors.

The #N/A error will only occur when you ask VLOOKUP to lookup the data in a certain way. VLOOKUP has two behaviors. The default behavior for VLOOKUP is that it assumes that your data is sorted. It could be alphabetic or numerical or whatever but VLOOKUP assumes that your data is sorted. So, VLOOKUP has four parameters - the value you want to lookup, where your data is, the column from which you want to retrieve additional data. In this case, John Doe will be the first parameter, the employees name table (Employees) is the second parameter, and column 4 which contains the date of birth is the third parameter. VLOOKUP also has an additional parameter - the fourth parameter - which can be true or false. By default, it is assumed that the data is arranged alphabetically. So, if you don't specify the fourth parameter, Excel assumes it to be true, and Excel tries to look up John Doe under the assumption that your data is sorted.

What happens when the assumption is sorted? If I give you a heap of papers and tell you too look up John Doe and I also tell you that the heap is arranged alphabetically, you would be very fast. You know what to do in case of alphabetical listing. You don't have to start from A or B or Z. You can straightaway jump to J and start looking up John Doe. It is the same thing with Excel. Excel assumes that your data is sorted, and so it speeds up VLOOKUP dramatically, and it gets to the point where John Doe should be and only looks there. If it cannot find the John Doe value within the bucket where all the J values should go (I am simplifying the entire thing here), it will come back and throw an error.

So, VLOOKUP has two behaviors. The first behavior is that, by default, it assumes that your data is sorted. Because, many times in real life, business data is not assumed to be sorted, we provide the last parameter as FALSE, which means that we tell VLOOKUP that we don't know what order the data is in and ask it to go and look up each and every item until you find John Doe. It then vertically looks up the data from the very first cell all the way to the bottom until it finds John Doe and comes back and gives the answer to you. When you provide the last parameter as FALSE, VLOOKUP tries to find John Doe and if it cannot find John Doe, it comes back and gives an error. On the other hand, if you don't provide the last parameter and tell Excel that the data is sorted, Excel will go and try to find John Doe, and if it cannot, it will give you an approximate value. For example, it might give you something like John Dish. It cannot find John Doe but it finds John Dish and gives you his date of birth. These two things in more plain English terms are called **approximate match and exact match**. In the first case when you provide TRUE as a parameter or when you omit it, Excel will give you an approximate match. In the second scenario, Excel will give you an exact match if it finds it. And, because you are asking for an exact match, if it cannot find an exact match, it will come back and give you an error. That's a very long answer to the question about what happens if the value is not found. What happens depends on how you are asking the question. If you asking it to give you an approximate match, it will give you some approximate answer even if the value is not found. On the other hand, if it is not found and you are looking for an

exact match, VLOOKUP will give you an error. This is the first question that people ask me and this is what happens - when VLOOKUP cannot find an answer - it will give you an error in most cases.

The second consequential question is - **should you sort your list all the time**? If it is up to you, and you have lots of data, I highly recommend that you sort the data. But, this is probably not practical in many day to day situations. Let us say that your employee data is not maintained in an Excel worksheet but is part of some sort of Enterprise Resource Planning system, or some such corporate database. So, the data is sitting somewhere, and you bring that data to Excel for analysis purposes and then analyze it. So, when you do something like that, naturally you don't own the data. You are just brining it from somewhere like a dump or snapshot of the data. When you do that, bringing the data is usually automated. You set up a connection and refresh the connection, and get the data into Excel. But, making sure that the data is then sorted is going to introduce some manual step in your automated process. So, every time that you have new data, you have to go and sort and re-sort the data, and that takes quite a bit of time, and you might sometimes forget. So, in case, if you don't own the data and manually maintain it, I suggest that you don't bother with sorting. Just go and do the exact matches. On the other hand, if you are doing some sort of manual steps anyway, then it makes a little sense to do the sorting. Why bother sorting the data? When you sort the data, the lookups will be faster. So, purely from a performance point of view, if you have sorted data, then it makes VLOOKUP faster. Should you sort your data? It is up to you. In case you have a manual process going on then just add a manual step there. In case you have written a macro to do some of these steps, you can also add an additional step in your macro that will sort the data. In case you are bringing data from external sources and you cannot trust the sources to send sorted data, you'll probably want to use the exact match parameter of VLOOKUP and live with that.

The third question that people ask me is a bit controversial. People often have very serious loyalty towards this. **People ask me if VLOOKUP is slower than INDEX+MATCH**. INDEX+MATCH is another variation of doing VLOOKUP. Using INDEX+MATCH you will be able to get the same answer as VLOOKUP but using two different functions and you will go in a different way. Explaining what INDEX+MATCH is a little beyond the scope of this podcast but in the resources section on the show notes page, I am going to link to an article that explains INDEX+MATCH. Let's go back to the question - which one is faster? Here is the simple answer which is based on my experience of working with lots of data in lots of different situations, and using the data for building dashboards and advanced analysis reports and what not. Based on what I have seen and learnt and taught people, to me it seems that both VLOOKUP and INDEX+MATCH are practically similar in terms of performance. There is a big if condition at the end of this statement, but for most day to day situations where you analyze data and are using the data analysis and feeding that into your reports, dashboards or whatever, you would not find significant difference in performance between VLOOKUP and INDEX+MATCH. This is the first part of the statement.

There is a big if at the end. This is because for certain specific scenarios and certain specific situations, you might find that the INDEX+MATCH approach is faster. Likewise, for certain other scenarios, you

might find VLOOKUP to be faster. What those things are varies a little bit and depends on the kind of data that you have and the kind of report you are generating and how many look ups you are doing etc. But, from my point of view, when I am creating reports, some of them are very complex and depend on lots of data. I have found that irrespective of VLOOKUP or INDEX+MATCH, the performance seemed to be similar. There is no significant difference between these.

Let us say that VLOOKUP takes a fraction of a millisecond. Imagine what a second is. A second is so small that we don't even notice it. A millisecond is one-thousandth of a second. So, we can't really spot the millisecond fragment of it. A millisecond is a really small fraction. VLOOKUP usually runs in the fraction of a millisecond and so it is even smaller than a millisecond, and it is very difficult for us to imagine what that looks like for most big picture kind of levels. If one VLOOKUP takes a fraction of a millisecond and you are running two thousand VLOOKUPS in your report, at best, the entire calculation will be done in one second. And, this is fairly not discernible. You enter your data, press F9 or you navigate to the cell, and Excel instantly calculates it. The good thing with computers is that they are getting faster every year. You have new processors, memory, internal circuitry and hard disks are getting faster. So, the time it takes to calculate and present the results on the screen is shrinking. From Excel 2003 to now, you would have noticed that most of the calculations that took ages in the earlier versions of Excel are now done in fractions of seconds. So, if this is the case, and if you have for some reason figured out that the INDEX+MATCH version of the same formula is 2% faster than the VLOOKUP version of the formula that 2% will translate in to 1.02 seconds. Now, can you tell the difference between 1 second and 1.02 seconds? For me it seems like we cannot tell them apart. If they are presented to you in two separate halves of a report, you would not be able to tell them apart, given the fact that you only have 2000 calculations going on.

If you are running 2 million VLOOKUPS and 2 million INDEX+MATCH, then in such a case, we have essentially multiplied by 1000, and so the calculation that is supposed to take 1 second is now taking up to 1000 seconds. On the other hand, it will take a little less than 1000 seconds because it is a little faster. In such a case you will notice the difference because now the difference will be maybe 0.5 to 1 minute, and so you will see the results of the INDEX+MATCH coming up there, and then after 15 to 30 seconds, you will see the result of the VLOOKUP.

When you have lots of data and have a set up and scenario that makes the INDEX+MATCH faster, in such cases, you will notice that INDEX+MATCH is faster. But, again, from my observation, we don't usually run 2 million calculations in Excel. It is very rare. Most business analysis and data happens in thousands and probably up to hundred thousand calculations and that seems to be the cap end. At that level, VLOOKUP is really fast. You don't need to spend a lot of time debating these things; you can stick to whichever approach works best for you and that you are comfortable with. For most scenarios, there is really no difference in the speed of VLOOKUP versus INDEX+MATCH. Go with whatever works best for you and whatever you feel comfortable with.

The fourth question is - **what if my list has multiple matches**? There isn't just one John Doe but there are multiple John Doe's. What would VLOOKUP do in such a case? It will find the first John Doe and it will stop there. It won't care what happens after John Doe is found. It stops there.

A follow up question that people often ask is - **if VLOOKUP cannot tell me that there are multiple John Doe's then how am I going to get the second occurrence of John Doe or the third occurrence of John Doe**, i.e. what if I want to know the date of birth of the third John Doe and not the first John Doe. This is where VLOOKUP is not the function to use. You can still use VLOOKUP but you would need to be a little creative and come up with a variation of VLOOKUP that would source the answer for you. It is possible to get the second or third occurrence of John Doe and the date of birth for that person but that requires a little bit of trickery. Explaining those techniques in the podcast is probably not conducive. So, I am going to leave you with a resource that will help you do this. Visit http://chandoo.org/session51/ which is the link for this podcast and you will find information on how to get the second or third match of a VLOOKUP there.

Another follow up question that people often ask is that **they don't want just the second or third John Doe but instead they want all the John Doe's and their dates of birth**. Let's say that you analyzing some invoices against some vendors and you have multiple vendors who are sending you invoices. And, you want to find all the invoices against a particular vendor, i.e. how many invoices have Microsoft sent you etc. So, you enter the value Microsoft in A1 and you are doing a VLOOKUP. Since Microsoft has sent you multiple invoices, you will only find one of them and not all of them. You will probably find the very first one. So, how are you going to get all the details for Microsoft and add up the total amount due or get additional insights based on it? In such cases, you can again use VLOOKUP but you would have to do some extra steps like setting up helper columns or writing array formulas or something like that. Alternatively you can use other formulas that can do this kind of work. Let's say that you just want to sum up the total invoice amount against Microsoft. Instead of using VLOOKUP, you can use the SUMIFS formula and simply ask it to sum up the invoice column wherever the vendor name is Microsoft. In case you want to add additional conditions like if you want to find out the total amount of invoices due against Microsoft that are still not paid then in this case you are writing two conditions - the vendor name has to be Microsoft and the invoice status is unpaid. So, you are doing VLOOKUP but on two conditions and this is what we call **multi-condition VLOOKUP**. In such scenarios, you can again use SUMIFS and put up multiple conditions or you can use the multi-condition VLOOKUP techniques. There are quite a few of them. I am going to post a link in the resources section that talks more about multi-conditions VLOOKUPS and how to use them and how to set them up along with an example workbook. So, check out the resources page where you can learn about those things.

The follow up questions that we talked about are - how do you get the second John Doe if there are multiple matches, how to get all John Doe's and how to get multiple conditions lookups. All of these are

standard questions that people ask about VLOOKUP and all of these are very easy to do. The only thing is that if you just use VLOOKUP alone then you won't be able to get them. You would have to go a little further and you would have to imagine a little more or at least know the technique and after a couple of times you would kind of know them as second nature and you won't have to think a lot about these problems anymore.

Another question that people often ask is **how to speed up VLOOKUP**. This is similar to the scenario where you have 2 million calculations as against 2000. Let's say that you have a very big workbook with lots of data and you happen to have lots of VLOOKUPS or other lookup formulas like HLOOKUP, INDEX-MATCH etc. and you want to speed things up as it is obviously going to be slow. So, the standard approach to reduce a roadblock or bottleneck is as follows - **see if you can altogether avoid it**. Let's say that you are running an airport and the security clearance where they check your bags and person is a bottleneck and everybody feels that from the time you enter the airport till the time you catch the flight, the biggest bottleneck is the security clearance. How do you clean this up? The easiest, really simple, common sense answer is to see if there is way to avoid it. Is there a way where the bottleneck can be completely removed because then everything will flow smoothly? Obviously, in case of air travel, maybe you can't avoid the security check but, in real life, in the case of our VLOOKUP problem, you can avoid it. So, if possible, avoid VLOOKUPS altogether. Let's say that you are looking up some information between one table and another table then the easiest way to avoid this would be by using new features of Excel. For example, Excel 2013 has a feature where you can connect one table with another just like you do in databases. You can specify that employees here are connected to the employee details there. So, if you can set up relationships like that, you can avoid VLOOKUPS and you can do a lot of reporting without having to worry about VLOOKUPS. I usually find that when you have millions of VLOOKUPS going on it is because you have disconnected data. You have separate data sets and you need to bring everything together in order to make a report or do some analysis. You don't have to bring everything together; you can keep them separate them and connect them and still create reports using pivot tables or Power Pivot and do the analysis. So, when you have lots of VLOOKUPS, that's probably a place where you want to avoid VLOOKUPS altogether and set up connections and use powerful tools like Power Pivot or pivot tables that are meant to process such large volumes of data. So, avoid it.

But, let's say that you cannot avoid it and you have to live with it. What would you do then? The second best technique in such cases is to **sort your data**. Even if it has to be done manually, sort the data. When you sort the data, VLOOKUPS will be naturally faster. How fast will they be? That depends on the volume of data and how many lookups you are doing but, usually, you will find speed improvements to the magnitude of 100 to 1000. It could be even more. But, there is a basic logic to it. If your list is not sorted and if you have n elements in the list (n=2 million here because you have 2 million items and you are looking up a particular item), the time it takes to find any item in that list is usually n. If we go a little further, it is basically called n/2. That means that the average time it takes to find that item in that 2 million is 1 million. 1 million multiplied by the time is the time it takes to lookup each item. Usually, it is in the order of a million, but, when it comes to execution time, it will be far less. It will probably be 10 seconds or something. If the list is sorted, in such a case, the time is no longer a multiple of n but it is

usually log n, i.e. logarithmic value of n and this is significantly less than n. So, how much smaller will log n be compared to n? That depends on how big the value of n is. So, that's why I said that for most large volumes of data, you would see speed improvements in the multiples of thousands and sometimes even bigger than that. So, sort the data and that will significantly improve the performance of our VLOOKUPS. Obviously, when you are sorting, you won't be able to trap the errors. When the list is sorted and you are doing an approximate lookup, if John Doe cannot be found, VLOOKUP will give some other garbage value that you don't want to see in your reports. So, how are you going to tackle that problem? This is where it gets a little trickier but there is an excellent idea that you can use. I'll leave a resource in the show notes page but this is how it goes. We first sort the list and then we do an approximate look up but we don't get the date of birth. We get the day. We do a look up on the name column for John Doe and get the name back again. So, if the return name is not John Doe then it means that we've found an approximate match. Originally, we are looking up John Doe, and we get the value again from the first column which contains the name. If John Doe is there, the return value of the original VLOOKUP will be John Doe again. If John Doe is not there then the return value will be something else. It could be James Doe or whatever. In such a case, we manually print an error by using an IF formula. We say that if A1 is equal to the VLOOKUP of A1 in Employees then we do the date of birth lookup otherwise we print an error. I know that imagining this particular technique in your mind with an audio only format is a little hard. So, there is a link that I will provide in the show notes page that talks a little more about this particular technique. Check out http://chandoo.org/session51/ for that. So, if possible, avoid VLOOKUPS. If that cannot be done then sort your data.

The third technique for speeding up VLOOKUPS is **paste replace**. So, do all your 2 million lookups and replace the VLOOKUPS with values. This is going to be even faster than the 'avoid' or 'sort' approaches. But, that's another technique. This way you are **replacing the formulas with values**. Again, you lose a little bit of flexibility. In case your data changes, you have to again do the VLOOKUP and replace. So, this is something that adds an extra bit of manual step but in case you are looking at waiting for 15 minutes for the VLOOKUPS to calculate, pasting replacing is a viable technique. It only takes one minute to do it and you can prepare a process document or something like that so that you can just follow those steps every time that you have to do it.

The last one is **'live with it'**. Because you are doing so many lookups and they are necessary for your process or workflow, you can't do anything about it. It's got to take 15 minutes, and just like the airport security check, it is annoying, frustrating and time consuming but everybody lives with it. So, maybe you can live with it and consider that to be part of the cycle.

This is how you can speed up VLOOKUPS.

The ninth question that people often ask is - **why doesn't my VLOOKUP work?** I am expecting to find John Doe. I know John Doe is in my list. Why doesn't VLOOKUP find the answer? This is because

sometimes when you get data and copy paste into Excel, you might get things like spaces or invisible spaces. As such space itself is not invisible but there are some other **invisible characters** also. So, it might look like John Doe to plain eyes but if you go and edit the cell there might be an extra space at the end of Doe, or there might be a space at the beginning of the word John, or there might be some extra spaces between the words John and Doe. In all such cases, VLOOKUP cannot consider that as John Doe. The value you are looking up is John Doe but the value that your database contains could be John Doe with an extra space at the end, or John Doe with an enter character at the end. These are all invisible characters. When your look up data or the value that you are looking up has some extra spaces or special characters that cannot be perceived by the human eye, you will get into a scenario where you are frustrated and you ask why your VLOOKUP doesn't work. The most likely answer is that you are looking up but your data is not what it looks like. So, you need to clean up your data and remove those extra spaces and invisible characters. There are some functions to do that like the **TRIM** function which will remove all the spaces at the beginning or end that are unnecessary. There is a **CLEAN** function that removes some of the hidden characters, or you can use a **VBA macro** or **Power Query** or something like that to **clean up the data**.

The last question that people ask me is - **why is my VLOOKUP giving an error?** The obvious answer is that it cannot find the value. But, then the follow up question is - **how do I fix this error?** This is where you can use the formula **IFERROR to suppress or trap the errors** of VLOOKUP. You simply write IFERROR, VLOOKUP, comma and then write whatever value you want in case of an error like 'value not found' or 'zero' etc. In case your VLOOKUP gives an error, the IFERROR will elegantly suppress that error and give the other value that you want to see in the report.

These are the top 10 most frequently asked questions about VLOOKUP. I know I might have missed some of the other burning questions about VLOOKUP that you have in your mind. So, in case you want to know more about VLOOKUP and in case you want to get answers to those questions, visit http://chandoo.org/session51/, scroll down to the comments section on that page and post your question as a comment. I will be watching that page for the next few days, and as and when possible, I will answer your questions.

Thank you so much for listening to the VLOOKUP FAQ's podcast. I have a few resources for you; I am going to link to them in the VLOOKUP show notes page. The most important two resources that I want to highlight are - our **VLOOKUP comprehensive guide page** on http://chandoo.org - it contains various examples and scenarios of VLOOKUP and how to handle them. I highly recommend that you check it out. The second thing that I recommend is my own book. I have written a book about VLOOKUP called **The VLOOKUP Book**. It talks about VLOOKUP, INDEX-MATCH, SUMIFS, SUMPRODUCT, COUNTIFS etc. and all these additional techniques and how to use them to solve many day to day practical VLOOKUP problems that you face in your work. Check that out. The book is available on Amazon. You can buy it for your Kindle. You can also buy the PDF version from our website. There is a video book available as well. So, it is the book plus a bunch of videos that explain VLOOKUP more. You can choose whichever one you

want. It is a very simple book but very powerful, in depth and advanced. So, it doesn't take a lot of time, and you can learn more about VLOOKUP using that book.

There you go; these are the VLOOKUP FAQ's. I hope you enjoyed this podcast. In case you have any feedback or suggestions for me, drop by at http://chandoo.org/session51/ and leave your feedback there. Also, if you have a minute, check out our iTunes page and provide your review and rating there so that I can learn more about how you are feeling about this podcast and how I can help you better.

Thank you once again and I wish you a very happy new year 2016. Bye.