SIMPLIFYING
ENTERPRISE DATA MANAGEMENT

Sachin Rane
Executive Vice President & Head – Software Solutions

Enterpise data management today is an important subject. All our customers that we meet across the world are looking at enterprise data management strategy.

In fact people are saying that can you help us in consulting, in having a proper enterprise data management strategy. What's happening is there are silos of information or data across organization. And our customers are looking at how do I make a cohesive strategy. Now enterprise data management let me simplify. What is it? It basically is a strategy as to how you manage your data, how you control your data, how you secure data, how you ensure that who accesses your data and when. How the data is stored, how the records will be maintained. So that is basically right from data capture to data storage. You completely manage the cycle of the data. That's basically enterprise

data management. Now enterprise data management has different aspects. You've look at how to capture the data. How do I transform the data. How do I qualify the data, how to store the data. How do I visualize the data. How do I analyze the data. And how do we ensure that the entire access is secured. Now that is what the entire lifecycle of our typical data is. So your idea is to get an intelligence out of data that's basically the perspective so that information is stored over the data repository in a typical enterprise data management is in three forms. Data warehouse, data mart and a data like. Data lake or big data, both are synonyms.

Let me explain the concept of data warehousing in the simplest form as much as I can. Data warehouse typically is a qualified data which is organized in such a way that you can use it for queries.

Now the queries typically are time series analysis.

Basically you want to go back in time and you look at historical perspective of your data. That's where the data warehouse comes in picture. Then the obvious question is what is data mart. Now I'm going to take an example of the simplest example which is actually pointed out by a CTO for Pentaho. Pentaho is again a tool which is used in business intelligence. I have a couple of bottles here. Now these are basically water bottles but sealed water bottle which basically means that this is a pure water. Now this I can say for data warehouse for example. What it means is that it has pure data. This bottle has pure water but I can in analogy it basically pure data. Now this is a small bottle. Again it's a pure water. But is the smallest form of pure data. I

mean if you look at analogy this is the smallest form of qualified data which is a smaller version of our data warehouse. Which means that I derive a data mart from a data warehouse. This is data warehouse and this is data mart which is again the smallest formal data warehouse but it is meant for a specific purpose. Right. This is basically this bottle is used for one person. This perhaps can satisfy a couple of people or three more people. So data warehouse is meant for many.

Data mart is meant for a specific purpose. So that's what it means in its simplest form of form of both data warehouse and data mart. Then the next question obviously is what is data lake? Because people have heard this concept and people have been talking about data lake and big data so are they two different things? Well obviously the answer is no. Big data, data lake, they are all synonyms.

What I mean by that is in the simplest form the difference between data warehouse and data lake, data warehouse data is meant for only structured data. Data lake whereas is both structured, unstructured and

sem-structured. What I mean by unstructured? Typically if we look at your e-mails, if you look at the logs, if look at voice, if you look at the audio, these are basically unstructured data. Again excel sheet for example. semi-structured data. Logs of a of a particular event. These are again, structured, semi-structured data. So a data lake is pretty much an information store that can take both structured as well as unstructured data put together.

Now the difference between a warehouse and a data lake is data warehouse it's only meant for a structured data whereas a data lake is again meant for data store but both structured as well as unstructured.

I hope I'm making it very clear. To summarize, data warehouse and data lake both are information store but meant for different purposes.

One is for structured one for structure as well as unstructured and semi-structured. So now that we've seen both what's the basic difference between this information types of information store. Let me tell

you what's happening in the market. And when we meet several customers, customers are seeing now data lake is becoming popular and it obviously can serve structured data as well as you're saying. So why not I just know dump data warehouse and go to Data lake. Is data like a substitute for data warehouse? Well my answer is clear no.

Both are meant for different purposes and both have different use cases. So if you are looking at for example data lake to be one answer for all your information need or query needs or data needs. Perhaps you should look at it differently. Many organizations now are looking at I have unstructured data, I have logs, I have Facebook I have social media. I need data collation from all the or all the sources. And I need a facility which basically can sustain me for next few years.

Answer perhaps is yes. Data is the way that you should go for but that doesn't mean that you should perhaps junk your data warehouse. It is made for a specific purpose and that purpose still remains even today.