The Future of Data Storage – Data Warehouse or Data Lake?

The differences and advantages of both data warehouses and data lakes are frequently being debated. Is the data warehouse obsolete? Will the data lake replace the data warehouse or is it just another new technology that will fade into obscurity? This article looks at the pros and cons of both technologies and their suitability for the clinical trials arena.

Defining Data Lakes

The term data lake was first credited to James Dixon, chief technology officer for Pentaho in 2008[1]. Dixon was looking for a way to describe the way in which unstructured data could be sorted and settled upon a metaphor using water. The terms 'data mart' and 'data warehouse' were both in use and Dixon began to think about thirsty people getting bottled water from a mart, a mart getting cases from the warehouse and the warehouse obtaining and bottling it from the wild source – the lake. Today we still use the three different terminologies – data marts refer to the access layers in a data warehouse, a data warehouse stores data accumulated from a wide range of sources and a data lake takes the raw data and transforms it ready to be used.

What are the Differences Between Data Lakes and Data Warehouses?

The table below highlights the main differences between the data warehouse and the data lake. These differences are split into the main areas that users are looking for when it comes to making an appropriate selection of the storage of data.

DATA WAREHOUSE	DATA LAKE
Data retention Data incorporated in a warehouse will be used to answer a specific business question or be used in a defined report. Data is stored in traditional relational databases which can be expensive to scale. Most warehouses are built using structured transactional data as a source.	Data lakes retain all data, kept indefinitely so that users can go back to conduct analysis at any time. Built on technologies such as Hadoop (open-source framework used for processing big data) and NoSQL (for storage and retrieval of data) databases. Can be scaled horizontally and performs well in cloud-based infrastructure.
Ability to handle varied data types Designed to process structured data, based on a fixed schema model of columns and rows that can be manipulated by SQL to establish relationships. Raw data must be put through a cleaning and structuring process called ETL (extract, transform and load) before it's stored. Schema must be defined up front before loading data resulting in less flexibility.	Single store of data all in one place, ranging from raw to transformed data which can be used for various tasks including reporting, analytics and machine learning. Data can include structured, semi-structured and unstructured data and even binary data in the form of images, audio and video files. Built on ELT (extract load transform) creating a centralised data store accommodating all data that can be manipulated into meaningful data sets.
Support for the user A suitable tool for those wanting to conduct further analysis of data, creating new reports leveraging BI tools or spreadsheets with data stored in the warehouse. Allow review of data with well-defined key performance metrics using reports and dashboards.	Allows data scientists to go beyond capabilities of data warehouses, use the data to come up with new questions to answer and use advanced analytical tools and capabilities like statistical analysis and predictive modelling. Extraction of data is more flexible, either into a standard format (e.g. a spreadsheet) or something more bespoke.
Future-proofing Relatively less flexible when set up and can be difficult and time-consuming to change. Can be adapted to suit user requirements but will require more extensive developer time and insight to make changes.	Data lakes are highly adaptive. Source data is stored in its raw format in HFDS (Hadoop Distributed File System) or NoSQL databases requiring only changes in the aggregations.
Speed of insight Conforms to a pre-defined enterprise data model and is capable of answering a limited number of questions.	Big data, data lake approach enables faster insight into information. More access to information across all types of data, stored in the raw format. Raw source data can be processed and queried to provide faster insights across existing and new sources of data.

Which Approach is Best?

On face value, there is no right or wrong solution, only a solution that is not fit for the purpose that it is required for. Ultimately, organisations need to spend time considering what data they want to store and with what purpose in order to determine which is the best possible solution. In fact, there could be times when it is sensible to have a data lake working alongside a data warehouse, with the warehouse allowing access to data in the way it has already been done, with the data lake providing a repository for new data sources. As a data warehouse ages, a data lake could be used in its place to ensure long-term accessibility.

Why Use Data Lakes in the Clinical Sphere?

Electronic health records (EHRs) are having an increasing influence on the clinical trials sphere. This information provides access to scalable cloud-based platforms of data that can positively influence clinical trials. Not only can these platforms help with suitable patient identification for the trial, they can also provide a wealth of relevant unstructured data about the selected participants. This unstructured data (from clinicians' notes, social media, wearable technology, printed academic papers, articles and other sources) can be easily processed using data lakes.

The data lakes can process huge volumes of data from a variety of sources including those drawn from EHRs, and combine it to provide meaningful insights with the use of big data technologies.

Clinical trials can truly benefit from the prevalence of big data. It can help researchers to pick the most appropriate subjects for their studies and monitor them more closely throughout the trial. Big data can personalise processes during the trial, which can help to save millions of dollars in getting drugs to market sooner. The relative ease with which data lakes can transform the variety of data originating from wearables, medical devices, structured clinical and operational data, safety data, etc. can help to achieve greater insights in terms of patient safety and improved trial performance.

The advantage of smart data lakes is that they provide the sort of flexibility and versatility that business users need to monetise the data. Smart data lakes are designed to facilitate the use of data in a way that is very similar to basic human thought – it becomes more straightforward to use the data in a way that is intuitive. This represents a triumph of accessible, end-user flexibility over the conventional constraints of information technology. The volume and variety of clinical study data that needs to be managed by sponsors, sites and CROs make a data lake an ideal model.

Smart data technologies will help to substantially reduce the complexity of data lake implementation and additionally help to accelerate the time-to-value ratio. In the clinical trials space, this means that disparate sources can now be merged both in structured and unstructured forms, which had previously required lengthy manual inputting. The machine learning also enables users to access near real-time analysis of the data.

Some of the challenges associated with smart data lakes are associated with governance of the information contained within it and the lack of management tools because of the speed at which the technology has been adopted. Organisations must learn to be agile enough to develop the processes and data-governance protocols as the lake fills up.

Data Lakes – The Future Approach

Data warehouses will not disappear overnight because of the substantial investments that have been made into this type of data

storage. Data warehouses and data lakes are optimised for different purposes and the goal should always be to use each one for what they have been designed to do. Both technologies have a role to play and can happily co-exist. In the longer term, the smart data lake will slowly become the go-to solution to afford the end user the flexibility and scalability that is required to run a successful trial.

For more information about data lakes and how ThoughtSphere solutions can integrate, please visit: www.thoughtsphere.com

REFERENCES

1. James Dixon, 2014. Pentaho, Hadoop, and Data Lakes (Internet) https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/ [Accessed 16/11/2017]

Pankaj Manon

Founder and CTO of ThoughtSphere. Prior to founding ThoughtSphere, Pankaj was the global CTMS lead for the largest CRO in the world, Quintiles, handling customer demand and solution delivery across all product verticals. He also has experience in large pharmaceutical, life sciences, and biotechnology companies, as well as project management and solution architecture roles with Oracle's North America consulting organisation. Pankaj has a Masters in IT Business Administration and advanced systems diploma from National Institute of Information Technology (NIIT).

Email: pankaj.manon@thoughtsphere.com