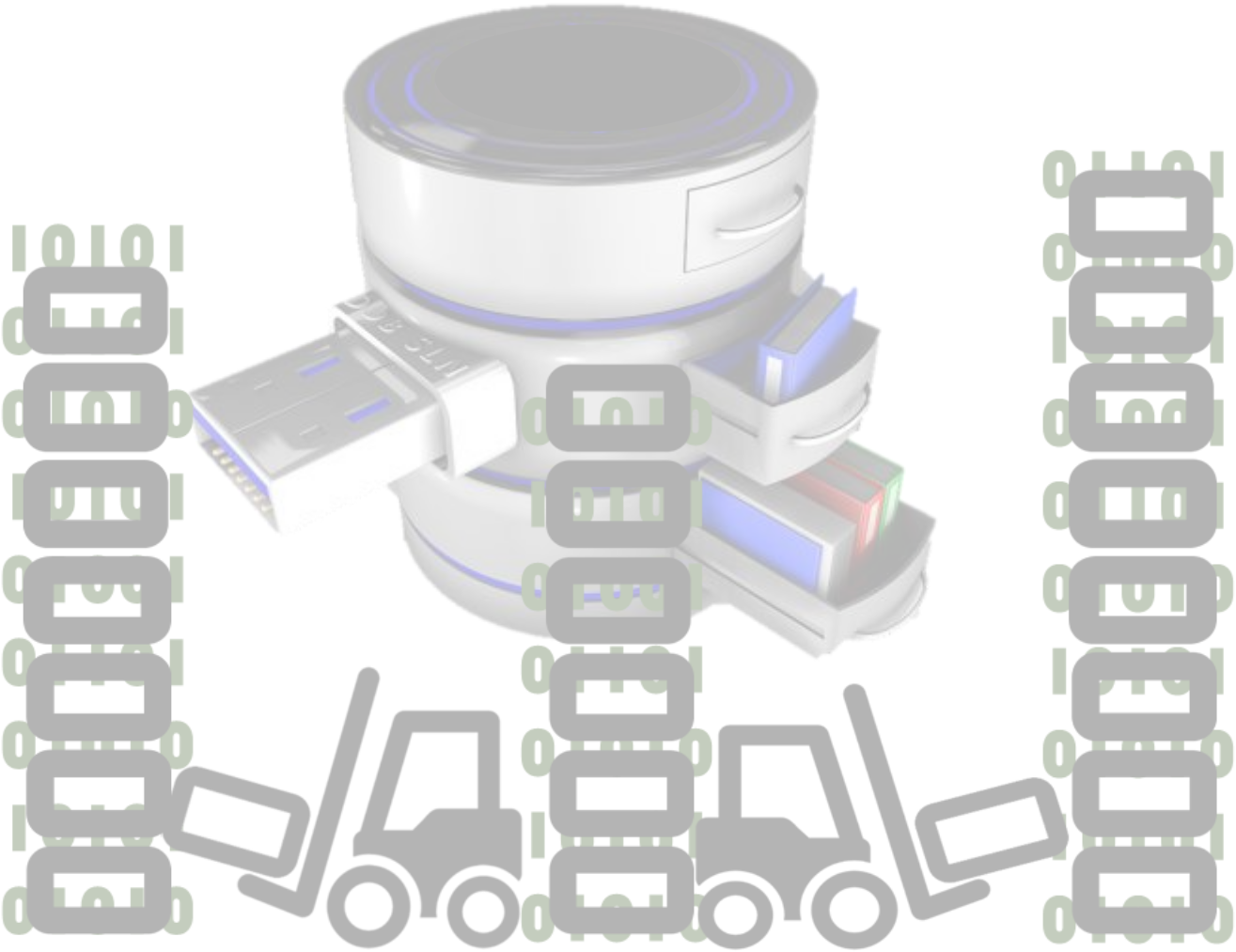
TRANSORG ANALYTICS



Data warehousing and big data

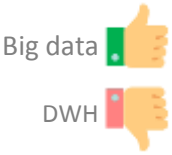
Vivek Jain
VP - Technology, Big Data Platform & Analytics
Vivek.jain@transorg.com

15-Nov-2017

Introduction

The purpose of this paper to upraise the audience on the use cases that can be solved using big data technologies and the ones that fit data warehousing (DWH). In this paper, we will also discuss the key differences between Data Warehousing solutions vs. Big Data Solutions. This paper should answer for you whether to deploy Data warehousing solution or Big Data technology stack or both. If you have already deployed Data Warehousing solution this paper should also help you decide whether you want to augment the same with Big Data technology stack.

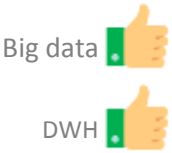
Data Sources



Data warehousing solutions are typically backed by a relational or database, thus making it an ideal solution for analysis of columnar structured data. Big Data technologies is a bouquet of technologies making it suitable for analysing structured as well as Semi-structured (web click stream, IoT devices etc.) & Unstructured (Web & Social Media like Facebook, Twitter etc.) data.

Supported Use Cases

Data warehousing solutions provide a good interface to data scientists to perform Descriptive analytics and business owners to create Business Intelligence dashboards. It must be integrated with other solutions to perform Predictive Analytics. On the other hand, Big data technologies provide a holistic echo system that provides tools for not only Descriptive Analytics but also integrates tools needed for Predictive analytics and Streaming Analytics. As far as Descriptive analytics is concerned



a) The response time in Data warehousing solutions is better as compared to Big Data technologies though serious work is happening in Big data technologies to improve on this aspect.

b) As Big Data technologies support semi-structured and unstructured data, descriptive analysis can be enriched from these type of data sources to provide better insights

Support for Graph computing

Spark has inherent support for graph computation using GraphX module whereas Data warehousing solutions do not support graph computation. Graph computations can help solve several use cases for fraud.

Data integrity & Data quality

As the underlying layer for data warehousing solution is a database, concepts like referential integrity checks and data constraints can be easily leveraged to provide better data quality. On the other hand, in big data technologies, filters have to be applied at query time to ensure bad data is filtered out.

Authentication & Authorization

Authentication & Authorization used to be the Achilles heels for Big Data technologies which made the adoption for these technologies difficult in security sensitive domains ex banking, insurance etc. Huge strides have been made in this area in last couple of years. Today Big Data technologies can be installed in compliance with PCI-DSS and HIPAA. Data warehousing had always been very strong on this aspect.

Cost of ownership & maintenance

Data warehousing solutions are based on expensive proprietary technologies vs Big data technologies which are mostly open sources. Given it is open source it is technically free to use, but there are too many open source components to integrate. As each of them have their own road map and release cycles it is a complex task. Fortunately, there are integrators who provide enterprise support taking away this pain at a fraction of the cost as compared to Data warehousing solutions.

Licensing model

Most of the data warehousing solutions license their product based on amount of data to be stored. This typically forces the users to sacrifice old data to maintain cost budgets. Whereas the support of Big Data technologies is based on the number of servers deployed. This helps the users retain old data which may be useful in some cases. Thus, users pay for computational requirement instead of storage requirements.

Development Resources

Data warehousing has been around for decades; thus, it is very easy to find resources for developing solutions on top of a Data warehouse. As compared to Data warehouse, Big data technologies are still evolving which makes hiring of resources with this skill set difficult.

Extendibility

Data warehousing is typically build on proprietary technology with SQL interface to insert/query data as compared to Big data technology stack which is a bouquet of open source components with several open interfaces for data ingestion and processing. This makes Big Data technology stack an ideal system for advanced computing and analytics like deep learning & text analytics.

Contact us

Shuchita Jain
Head – Client Development & Marketing
shuchita.jain@transorg.com
Mob: +91 98112 60911

TRANSORG ANALYTICS

www.transorg.com | Tel: +91 124 4231894