

Case Study: Addressing Barriers to Big Data

UIW PMBA 6330 – Applied Data Analytics

Case Study #2

### Case Study: Addressing Barriers to Big Data

Big Data is defined in the case study by three elements. The first requires the data to be in a large volume. Vast amounts of data must be stored and retrievable to use in an analytic manner. The second element of big data is variety. The data is compiled from many different points, which corresponds to the third element, velocity. The velocity of data to be categorized as big data must be ever present, a daily, or continuous increase in the volume. These three elements present unique challenges to modern businesses which seek to make use of the data that is being captured, accumulated, and stored.

#### **Background**

The ever-expanding complexity and use of technology now allows businesses to expand their analytic collection ability. Examples provided by the case study include airways, theme parks, city councils, auto manufacturers, and parcel services. All collect and analyze big data to increase business effectiveness and ultimately profitability.

Business IP traffic in North America is growing at a rate of 23% per annum. In 2021, it is estimated that there will be 14 exabytes of data (14 billion gigabytes) per month of business IP traffic in North America (The Zettabyte Era, 2017). Big data is growing at a steady rate. The ability to harness this data can result in a competitive advantage to organizations.

#### **Problem Statement**

There are multiple barriers for organizations to be able to use big data. If we look back at the definition of big data, we can find the overarching issues there. The first is the amount of data that is being stored, coupled with the increasing amount of data being generated. Businesses have a huge burden to upgrade storage capacity to keep up with the amount of data that it needs to

maintain. Storing data costs money. Even with a decrease in data storage per terabyte, there is an ever-increasing cost related to the exponential growth of data (Lyng, 2017). One other related barrier is the ability to estimate the amount of data that will need to be stored (Lyng, 2017). If an organization is budgeting on a yearly basis for storage needs, but underestimates the amount of data that will need to be stored, there can be consequences. These can lead to budget overruns, loss of old data, or loss of new data. One other factor that is commonly overlooked in data storage is regulation (Adshead, 2016). Innocuous regulation that does not appear to be inherently related to data storage can often contain provisions that outline requirements. For example, HIPAA (Health Insurance Portability and Accountability Act) outlines requirements for maintaining the health and personal information of customers (How to Remain HIPAA Compliant In Data Storage, 2017).

Data accessibility can be another challenge. Organizations with older data can have thousands of reels of data stored on tape in storage facilities. If the data is not accessible, it cannot be used. United States federal regulation dictate the length of time an employer must maintain employment records of employees. It can be anywhere from 1 to 6 years post termination (Recordkeeping Policy: Record Maintenance, Retention and Destruction, 2017). For smaller organizations, this can amount to boxes stacked in a storage facility. For larger organizations, it is important to consider the size and scope of document retention requirements and the variety of them to ensure the right records are being maintained over time. There are 13 different retention requirements for employment data alone (Recordkeeping Policy: Record Maintenance, Retention and Destruction, 2017). Industry specific record retention can be even more complicated when customer data includes highly confidential information, such as investment and banking data.

Along the same lines of accessibility, data organization and formats can also be a barrier. Data storage from many different systems can take on many different formats. Being able to analyze this data in a uniform way is often difficult. This harkens back to the Zurich job catalog case study. The barrier to data analytics from that perspective was lack of uniformity of job data and standardization which created additional work and gaps in the data analysis.

As stated previously, regulation rules from HIPAA or FINRA, or any other regulatory body that applies regulation to an organization's data collection and storage, dictate how data should be stored and how well it is kept private. FINRA rules go so far as to dictate how information is shared through social media (FINRA.org, 2017). Data privacy is a concern that is increasingly important as customer data becomes compiled and stored in a single place. One of the more troubling data breaches to date is the very recent breach of Equifax. The reason why it is so important is that it contained the names, Social Security numbers, birth dates, addresses and driver's license numbers of those citizens of potentially 143 million consumers (John, 2017). Data breaches and other malware can also corrupt data and make it inaccessible for use.

The talent and skill to mine and make use of big data is at a deficit. Currently, 40% of organizations are having difficulty with finding the talent needed to analyze data (A Shortage of Talent in Data Analysis, 2017). The cost of storing, maintaining, and securing data is worthless if there are not analysts that can make use of it. One other big barrier of this kind is the ability to translate and interpret the findings of the analysis to make business use of it. One study found 60% of businesses feel their analytic community needs to increase their skill in translating the data (Harris, 2014).

One other related barrier to talent and skill is the culture of the organization. If there is not strategic direction nor emphasis placed on overcoming the barriers of big data, there may be limited resources available to tackle them.

### **Recommendations**

The barriers to big data appear overwhelming when laid out together. One of the biggest and longest dwelled on is storage and maintenance of big data. One of the biggest trends in the last decade has been migration to cloud storage. Cloud storage is remote storage of data that is often scalable to need. It is accessed via the internet and most often maintained by a third party organization (What is Cloud Storage?, n.d.). Private cloud storage also exists. Third party storage comes with benefits, like scalability on demand, reduced cost, and security expertise (Nielsen, 2017). With the service being managed by a third party there is reduced cost in infrastructure and labor. Experts are managing the security of the data and it is scalable. As the data grows, most cloud storage will grow as needed.

Integration of legacy systems is the least expensive way to begin accumulating and tying data together to analyze it. XML (extensible markup language) can be used to standardize data into formats that are readable by data analysis programs (Myers, 2016). For example, using XML to format banking data into measurable employee analytics can be done to inform an incentive compensation system. The legacy banking system may not be using the same data or file formatting as the incentive compensation, but the XML can translate and input the data as needed into the newer system.

Using algorithms, as the case study suggests, like Hadoop can assist with managing the overwhelming amount of data that needs to be analyzed. This functionality allows organizations to refine its data mining. Adding nodes to a data storage system also limits downtime as failure

of a node can be replaced by another. The cost is limited to the infrastructure as applications/system software like Hadoop are open source, or free for use (What is Hadoop?, n.d.).

Privacy of data is not often looked at as a people problem. If the solution that an organization is looking for is related to only the software that is used for firewalls and encryption, they are missing half of the pieces of the puzzle. Human manipulation is often a large cause of data breaches. An organization can protect from human related data breaches by training its staff on what to look for via phishing campaigns. Phishing is a method used to target employees by presenting realistic looking websites or emails that contain malware. The data gained is often keys to the door of the employer's data stores. Education is key to preventing phishing from working on an employee. Having adequate locked facilities for server hardware helps prevent inappropriate access to hardware systems where malware can be inserted or other manual manipulation can occur. Running background checks is often overlooked as a data security protocol. An employer can unwittingly hire criminals looking to target customer or employer data if they are not performing background checks for criminal or financial wrongdoing prior to hire and access to data (Marks, 2014).

Lack of skill and talent can be resolved by teaming up with universities. Universities will often create internship partnerships with organizations to provide training and talent both ways. For example, a small food distribution company, Labatt Food Service LLC, has a partnership with St. Mary's university to provide interns for data analytics from its Industrial Organization program. This is a small local to San Antonio, TX scale example. On the larger side, organizations like Google will farm candidates from Stanford University (Oremus, 2014). If an organization is short handed the talent that is needed, big or small, they can create a partnership

with a local university to develop the talent that they need. By providing internships for that program, the organization can train the talent in a more focused and direct fashion to the exact needs they are looking for.

One of the biggest takeaways from the recommendations should be to have a plan. If an organization is not yet on route to big data analytics and they desire to be, having a plan can help facilitate getting there. If this requires a change in organizational culture, then adding culture change to the roadmap can help. This helps layout for not only executive level management where to allocate resources, it also helps facilitate the understanding of where each part of the organization needs to be. Analyzing data analytics is not just standing up an analytic team. It includes many parts of the organization to get to the end of the road. These include human resources, corporate security, software development, network engineering, and finance to coordinate and define needs.

### References

Adshead, A. (2016). Top five challenges of compliance in storage and backup. Retrieved September 24, 2017, from <http://www.computerweekly.com/podcast/Top-five-challenges-of-compliance-in-storage-and-backup>

A Shortage of Talent in Data Analysis. (2017, February 22). Retrieved August 31, 2017, from <https://www.business.com/articles/big-data-big-problem-coping-with-shortage-of-talent-in-data-analysis/>

FINRA.org. (2017). Finra.org. Retrieved 25 September 2017, from <http://www.finra.org/industry/books-records>

Harris, J. (2014, August 07). Data Is Useless Without the Skills to Analyze It. Retrieved August 31, 2017, from <https://hbr.org/2012/09/data-is-useless-without-the-skills>

How to Remain HIPAA Compliant In Data Storage. (2017). Miami and Broward Colocation | Volico Data Center. Retrieved 25 September 2017, from <https://www.volico.com/how-to-remain-hipaa-compliant-in-data-storage-and-accessibility/>

John, A. (2017, September 21). Equifax Data Breach: What Consumers Need to Know. Retrieved September 24, 2017, from <https://www.consumerreports.org/privacy/what-consumers-need-to-know-about-the-equifax-data-breach/>

Lyng, G. (2017, March 15). Eight Storage Challenges to Overcome in 2017. Retrieved September 24, 2017, from <https://itblog.sandisk.com/eight-storage-challenges-overcome-2017/>

Marks, G. (2014, June 23). 7 Ways To Protect Yourself Against A Data Breach. Retrieved September 24, 2017, from <https://www.forbes.com/sites/quickerbetteertech/2013/12/31/7-ways-to-protect-yourself-against-a-data-breach/#391f30392873>

Myer, T. (2016, February 01). A Really, Really, Really Good Introduction to XML — SitePoint. Retrieved September 24, 2017, from <https://www.sitepoint.com/really-good-introduction-xml/>

Nielsen, K. (2017). Top Ten Advantages of Using Online Storage Services. Retrieved September 24, 2017, from <http://www.toptenreviews.com/services/articles/top-ten-advantages-of-using-online-storage-services/>

Oremus, W. (2014, May 23). Googlers Are From Stanford, Applers Are From San Jose State. Retrieved August 31, 2017, from [http://www.slate.com/blogs/future\\_tense/2014/05/23/tech\\_company\\_feeder\\_schools\\_stanford\\_to\\_google\\_washington\\_to\\_microsoft\\_sjsu.html](http://www.slate.com/blogs/future_tense/2014/05/23/tech_company_feeder_schools_stanford_to_google_washington_to_microsoft_sjsu.html)

Recordkeeping Policy: Record Maintenance, Retention and Destruction. (2017). SHRM. Retrieved 25 September 2017, from [https://www.shrm.org/resourcesandtools/tools-and-samples/policies/pages/cms\\_017186.aspx](https://www.shrm.org/resourcesandtools/tools-and-samples/policies/pages/cms_017186.aspx)

The Zettabyte Era: Trends and Analysis. (2017, August 01). Retrieved September 24, 2017, from <https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html>

What is Cloud Storage? - Definition from Techopedia. (n.d.). Retrieved September 24, 2017, from <https://www.techopedia.com/definition/26535/cloud-storage>

What is Hadoop? (n.d.). Retrieved September 24, 2017, from [https://www.sas.com/en\\_us/insights/big-data/hadoop.html](https://www.sas.com/en_us/insights/big-data/hadoop.html)