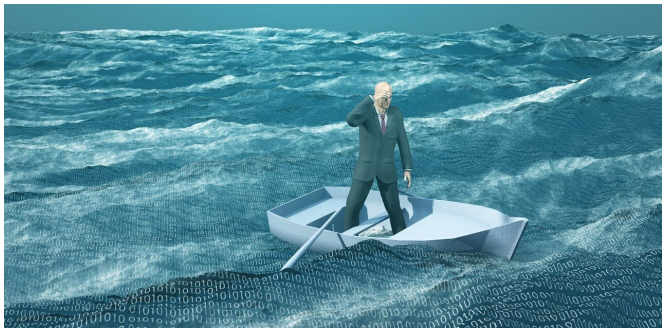
Data lakes: Where big businesses dump their excess data, and hackers have a field day

22 October 2019, by Mohiuddin Ahmed



Unlike purpose-built data storage systems, a data lake can be used to dump data in its original form. This data usually remains unsupervised. Credit: Shutterstock.com

Machines and the internet are woven into the fabric of our society. A growing number of users, devices and applications work together to produce what we now call "big data". And this data helps drive many of the everyday services we access, such as banking.

A comparison of internet snapshots from 2018 and 2019 sheds light on the increasing rate at which digital information is exchanged daily. The challenge of safely capturing and storing data is becoming more complicated with time.

This is where data warehouses and data lakes are relevant. Both are online spaces used by businesses for internal data processing and storage.

Unfortunately, since the concept of data lakes originated in 2010, not enough has been done to address issues of cyber security.

These valuable repositories remain exposed to an increasing number of cyber attacks and data breaches.

A proposed panacea for big data problems

The traditional approach used by service providers is to store data in a "data warehouse"—a single repository that can be used to analyze data, create reports, and consolidate information.

However, data going into a warehouse needs to be preprocessed. With zettabytes of data in cyberspace, this isn't an easy task. Pre-processing requires a hefty amount of computation done by high-end supercomputers, and costs time and money.

Data lakes were proposed to solve this. Unlike warehouses, they can store raw data of any type. Data lakes are often considered a panacea for big data problems, and have been embraced by many organizations trying to drive innovation and new services for users.

James Dixon, the US data technician who reputedly coined the term, describes data lakes thus: "If you think of a datamart as a store of bottled water—cleansed and packaged and structured for easy consumption—the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples."

Be careful swimming in a data lake

Although data lakes create opportunities for data crunchers, their digital doors remain unguarded, and solving cybersafety issues remains an afterthought.

Our ability to analyze and extract intelligence from

data lakes is threatened in the realms of cyber space. This is evident through the high number of recent data breaches and cyber attacks worldwide.

With technological advances, we become even more prone to cyber attacks. Confronting malicious cyber activity should be a priority in the current digital climate.

While research into this has flourished in recent years, a strong connection between effective cybersecurity and data lakes is yet to be made.

Not uncommon to be compromised

Due to advances in malicious software, specifically in malware obfuscation, it's easy for hackers to hide a dangerous virus within a harmless-looking file.

False data injection attacks have increased over the past decade.

The attack happens when a cyber criminal exploits freely available tools to compromise a system connected to the internet, to inject it with false data.

The foreign data injected gains unauthorized access to the data lake and manipulates the stored data to mislead users. There are many potential motivators behind such an attack.

Components of data lakes

Data lake architecture can be divided into three components: data ingestion, data storage and data analytics.

Data ingestion refers to data coming into the lake from a diverse range of sources. This usually happens with no legitimate security policies in place. When incoming data is not checked for security threats, a golden opportunity is presented for cyber criminals to inject false data.

The second component is data storage, which is where all the raw data gets dumped. Again, this happens without any sizable cyber safety considerations.

The most important component of data lakes is data analytics, which combines the expertise of analysts, scientists and data officers. The objective of data analytics is to design and develop modeling algorithms which can use raw data to produce meaningful insights.

For instance, data analytics is how Netflix learns about its subscribers' viewing habits.

Challenges ahead for data experts

The slightest change or manipulation in data lakes can hugely mislead data crunchers and have widespread impact.

For instance, compromised data lakes have huge implications for healthcare, because any deviation in data can lead to a wrong diagnosis, or even casualties.

Also, government agencies using compromised data lakes may face mayhem in international affairs and trade situations. The defence, finance, governance and educational sectors are also vulnerable to data lake attacks.

Considering the volume of data stored in data lakes, the consequences of cyberattacks are far from trivial.

And since generating huge amounts of data in today's world is inevitable, it's crucial that data lake architects try harder to ensure these at-risk data depots are correctly looked after.

Provided by The Conversation