Data Lakes:

Home for Unstructured Data and Unlimited Insights

Sravanthi Prattipati

Ajinkya Bhavik

Professor Jensen

BUS 247

San Jose State University

Table of Contents

What is a Data Lake? **2**

How do Data Lakes Work? **2**

Why use Data Lakes? **2**

Who uses Data Lakes? **2**

The Origin and the Evolution of Data Lakes **3**

Data Lakes VS Data Swamps VS Data Warehouses **3**

Types of Data Lakes (Data Lake Providers) **4**

Amazon Web Services: 4

GCP: 5

Snowflake: 5

Where are Data Lakes used? (Success Stories) **5**

Woot : 5

Innovaccer : 6

Data Lakes Future/Vision **7**

WorksCited **8**

What is a Data Lake?

In today's world, there is something that is more valuable than the money itself which is none other than data. Companies are spending money on everything from the collection of the data to the analysis of the data to make better decisions. In 2020 alone there will be 40 trillion gigabytes of data and 90% of all the data was created in the past two years, based on an IBM source (Petrov, 2020). Where is this data stored and why is it important to be discussed?

Data Lake is a place where data of all sorts (structures, semi structured, and unstructured) can be gathered, stored, and analyzed. However, it is not as simple, there are many other places where data can be gathered, stored, and analyzed ("Data Lakes and the Data Lake Market," 2020). For a data lake, the massive amount of data is in its raw and native format which takes the possibilities of what we can do with the data to infinity.

How do Data Lakes Work?

Data Lakes gives us an opportunity to use the data for various uses to derive relevant conclusions. There is a process for data lakes to work the way they do. Just like its title, data lake can be compared to a water body. The concept can be compared to a water system where the water flows in , fills up a reservoir and then the water flows out. The incoming flow of water represents multiple raw data archives which can be in different formats. The data can be unstructured ranging from an email, pdf file, images, audio, social media content, spreadsheet, video, etc. ("Data Lakes and the Data Lake Market," 2020). The reservoir then gets filled with the data which makes it a dataset where one can run analytics on all the data. The outflow of the water can be referenced to the analyzed data. Then through this process, one will be able to gain key business insights.

James Dixon, who was CTO at Pentaho, promoted the concept of data which can be a better alternative for data mart or a data warehouse. In 2011, Dixon explained this novel concept as "If you think of a data mart 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" (iscoop). The natural state of data lakes is what makes it unique and resourceful and there is no end to the findings that can come out of it.

Why use Data Lakes?

There are several benefits to using Data Lakes. Data Lakes give an opportunity for companies to build applications, provide flexibility and accessibility, retain data authenticity, provide speed , and give an ability to explore and analyze massive data. Data Lakes serve as the backbone for applications that can be used to get the data and the development of the data-driven applications that the company really needs. Data Lakes are very flexible since they will take the data in any format which gives the company access to analyze data which they were not able to analyze before. Clean and structured data can be desirable, but data lakes let the company maintain the authenticity of the data that is used for companies to make key decisions. Despite giving unlimited opportunities for the data to be analyzed, the data lakes have the ability to sift through the immense quantities of data quickly. Data Lakes help reveal complex business issues and build predictive models to address the issues that the company might be facing. Small startups to big corporations across several industries use the data lakes' solutions to solve their business issues and make the data lake analytics part of their everyday process.

Who uses Data Lakes?

It is given that the companies use data lakes to find data backed solutions to their complex business problems. In these companies there are three main groups of people who are closely interacting with the data lakes. Firstly, business and data analysts analyze the repairs on a specific date to provide the organizations insights which were gained from the data. Data Architects are relevant to manage data lakes as they are ones who are responsible for everything from "designing, creating, deploying, and managing an organization's data architecture" ("Data Lakes and the Data Lake Market," 2020). Data scientists and app developers use the data lakes as well to perform statistical analysis on big data so they can identify trends, solve business problems, which can ultimately optimize the performance of the company.

The Origin and the Evolution of Data Lakes

When companies started to realize the need for managing data, they used a relational database which is able to collect, store, and analyze data. Relational databases gave companies an opportunity to store highly structured data using Structured Query Language (SQL). Relational databases or also called as Relational Database Management Systems (RDBMes) addressed companies needs as the data used to be small, structures, and simple ("History and Evolution of Data Lakes," 2020). If companies deal with data which is relatively structured and small, relational databases would be a good choice for companies.

As the dot com boom happened and the use of the internet started to rise, the companies are receiving data in all formats and in unimagnale quantities where the relational database is not addressing their needs. Companies in the industry combatted this by building multiple databases. However this led to massive amounts of data which was split into disorganized data bases ("History and Evolution of Data Lakes," 2020). Although the abundance of the data helps the companies make better decisions, this also led to data silos. Data silos are "decentralized, fragmented stores of data across the organization" ("History and Evolution of Data Lakes," 2020). Although there is a lot of data which could provide insightful information for the organization, without a way to synthesize and centralize the data many organizations failed to make the best use of that data that is available to them. Eventually to alleviate themselves from this pain, data warehouses were brought into the data management picture.

Data warehouses, just like what the name suggests, emerged to "unite companies' structured data under one roof" ("History and Evolution of Data Lakes," 2020). Data warehouses' purpose is to bring a company's sets of relational databases under one roof with a mission to unite the disconnected database across the firm. Initially, data warehouses were run by expensive hardware from vendors such as Vertica and Teradat, but eventually they became available in the cloud. In the late 90's major organizations have made a shift to having data warehouses which were able to integrate many data sources, optimized the data for easy access, able to run quick ad hoc queries, and able to do data governance and audit. However, data warehouses were not able to store the raw and unstructured data. It was expensive to deal with its hardware and software and faced trouble in scaling due to limited storage and compute power.

Early 2000's gave birth to big data which has tested the organization's ability to take in the massive and different forms of data and to be able to get the insightful results. Organizations need a way to make use of unstructured data to not miss out on the key insights. Apache Hadoop rose as an open source distributed data processing technology. This was huge for many organizations for two reasons: 1) The organization can now stay away from the expensive in house computing clusters and can switch to the open source Hadoop. 2) Now, companies are able to analyze massive amounts of unstructured data which can make the possibilities of finding useful insights unlimited ("History and Evolution of Data Lakes," 2020). Since the companies have the capability to analyze the raw data, this set up the stage for the modern data lake. Although many data lake architectures have changed the paths from Hadoop to running Spark in the cloud, many people associate the term "data lake" with Hadoop since it was the first framework which gave birth to the concept of data lake. In that case, the data of an organization is uploaded to Hadoop's platform and then the necessary data cleansing and analysis is done to the data where the data resides on Hadoop's cluster ("Data Lake," 2015). Apache Spark became more popular as its interactive model made it easy to use for the data analysts. Today, many companies rely on data lakes to make better decisions for their company by using as muchas data as possible since the dta lakes have helped lift the restrictions on what can or cannot enter the data lake.

Data Lakes VS Data Swamps VS Data Warehouses

Companies collecting huge amounts of data and storing it may risk the creation of dataswamps. Data Swamps are highly disorganized and hampers the basic tasks of retrieving and utilization of data effectively. A data lake on the other hand is a data repository which is highly organized and helps in effectively retrieving and utilizing the same data (Data Lake VS Data Storage, 2019).

Both the data warehouse and data lake fulfill the purpose of storing high level data yet they are a lot different from each other. Traditionally data warehouses carry out advanced querying and analytics in structured databases while data lakes work as a single point scalable storage repository which holds data in its native format (Matthews, 2020).

We focus below on the different parameters which are important for understanding the different facets of data storage :

Parameters	Data Lakes	Data Swamps	Data Warehouses
Data Structure	Raw	Processed	Processed
Data Relevancy	Relevant Data	Irrelevant Data	Relevant Data
Data Governance	Maintains high level of data quality due to high and fine grained data governance.	Lacks Data Governance	Medium to high level of Data Governance.
Metadata tags	Metadata Tags Acting as Tiered Structure	No Metadata tags	Contains Metadata
Accessibility	Highly Accessible & Quick Updates	More complicated and not easily updatable	More complicated and costly to make changes
Users	Anyone in the organisation	Business Professionals	Data Professionals

Types of Data Lakes (Data Lake Providers)

Data Lakes act as a primary source for major organizations as it does the main task of generating value using tools and technologies which analyze data. Thus, there are many data lake services providers/vendors in the market whose services act as the competitive differentiator for many organizations.

1) Amazon Web Services:

AWS provides Data Lake services for many major organisations such as Netflix, Twitch, Linkedin and Facebook. AWS Data Lakes are known to help find the fastest insights and answers from the data to all their users. Some of the AWS Data Lake perks are that AWS data lakes are easiest to build as well as analyze. They provide a very broad, open and comprehensive array of analytical tools and technologies which puts AWS with the highest scalability. AWS data lakes are cost effective and provide a secure architecture for analytics. These lakes also provide fine grained access and control of data. They use features like Amazon Macie which helps find sensitive data or wrongly stored data as well as Amazon Insector to spot errors in configuration to avoid data breaches (Data Lakes on AWS, 2020).

2) GCP:

Google being one of the last few entering the cloud and data storage domain did things a lot differently than conventional data lake service providers. The main component of GCP's data lake is its Google Cloud Storage which acts as storage spaces for Data Lakes. In a Data Lake, GCP is used for unstructured data while CloudSQL, Big Query are used for structured data. One can store and save all types and process all sizes of data in this Data Lake. There are multiple tools for all phases of data processing in data lakes. Those are Cloud Transfer Service, BigQuery API [Ingestion], Cloud Storage, Bigtable [Store], Dataprep, Data flow [Process], Big Query [Analyse], Data Studio [Reporting] (Data Lake Modernization, 2020).

3) Snowflake:

Snowflake's philosophy centers on eliminating data silos and running multiple sets of workload on a single platform. It acts as a governed and safe access to all data along multiple levels and types into one single platform. Some of the key features of using snowflakes data lakes are that they provide consolidated data, fast scalable analytics on that data. The data lake is also very simple and cost effective while at the same time pretty secured and governed. Snowflake provides exceptional fast query performance, integrated and extensible data pipelines and secure governed collaboration with users. It transforms data efficiently using ANSI SQL. Granular level access control for precise accessibility is something which renders snowflake an advantage in certain sectors as well (Snowflake, 2020).

Where are Data Lakes used? (Success Stories)

The deluge of data in all its forms from different sources hold valuable insights for an organisation. Usage of Data Lakes in organisations have brought about a lot of change and flexibility in the data analysis within the organisations. Below we discuss some of the examples of how many organisations shifted their systems into using data lakes and the transformations brought about within these organisations.

Woot:

The team at Woot was focused on designing a cloud native data lake as a replacement for their legacy data warehouse system which was based on relational databases. They didn't go for utilizing the simplest migration method of lift and shift migration from one RDBMS (Relational Database) to another. Instead they focused on their primary concern of shifting into the usage of cloud native technologies to take them to their end state. The company wanted to decouple their original oracle database into a system where they could use the right tools for the right jobs within the system ("AWS Big Data Blog", 2019).

There were many architectural and design concerns that were present in the original data warehouse system. The design points which state those concerns are discussed below:

a) Customer Experience : Due to working backwards on identifying and fulfilling customer needs, the organisation's data warehouse is used widely in multiple teams by multiple users across the organization. The ability to garner insights into operations by all of these users was something focused on ("AWS Big Data Blog", 2019).

b) Minimal Architectural Maintenance : To remove the cumbersome heavy lifting of the managing infrastructure as per changing demands and evolving technologies. Therefore one of the important design points was to focus on the utilization of Data lake Services to allow us to use cloud native technologies.

c) Data Source Change Responsiveness : In the existing warehouse, updates to ETL jobs and tables were required to be done manually in the warehouse. Since the response time was very significant to the stakeholders, a high end performance architecture is selected. ("AWS Big Data Blog", 2019).

To meet the above requirements, data lakes were introduced architecturally and in an operational context as well. A shared responsibility was introduced for data ingestion and serverless model was preferred over traditional RDBM models. This shifted the responsibility of the data ingestion team and customized pipelines towards services to push their data ahead. This led to the development team having control over their services data. Setting up this data lakes thus increased accessibility which was brought about by using AWS technologies like Amazon Kinesis Data Firehose for data ingestion, Amazon S3 for data storage, AWS Lambda and AWS Glue for data processing, AWS Data Migration Service (AWS DMS) and AWS Glue for data migration, AWS Glue for orchestration and metadata management & Amazon Athena and Amazon QuickSight for querying and data visualization. ("AWS Big Data Blog", 2019) Customer Services also had levelled up since users now didn't have to download sql client, request user name password and learn sql to get data out. Now they could just directly user queries with Athena or even use Amazon Quicksight for accounts managed through pre existing directories.

After adopting AWS Data lakes into their system, operation costs have fallen down by almost 90% as stated by the organisation. Since AWS data lake enabled a serverless infrastructure gaining access data and migrating it became an easier task. The S3 based architecture also enabled the organisation to facilitate experimentation with newer technologies which integrated seamlessly with core services like Amazon EMR, Sagemakes and Lambda. ("AWS Big Data Blog", 2019).

Innovaccer:

The healthcare sector has been dealing with a lot of outdated data ingestion, aggregation and analysis methods. Innovacer decided to develop a data activation platform [Big Data Platform] based on the need of healthcare organisations to make powerful, data driven decisions ("AWS Startups Blog", 2019).

The organisation started by first moving its development load from Google Cloud to Amazon Web Services. It leveraged EC2 servers [Elastic Cloud Compute] for hosting its web-based applications and data warehouses. The deployment provided developers in the organisation with better accessibility, more fine grained control and improved performances ("AWS Startups Blog", 2019)

To undergo the fine tuning of the organisation's data platform, the organisation re-evaluated its architecture and developed a data lake platform which focused on fulfilling 4 specific parameters which were high performance for huge data, low maintenance costs, cost-effectiveness, ease of scalability. Working on fulfilling these parameters the organisation was able to find the perfect system to go with and decided to transition to Amazon Redshift. This shift helped Innovaccer in achieving a higher rated time to value in garnering insights and analyzing data and developing ROI's. Usage of drag and drop modules helps it build and run pipelines thus integrating data across multiple sources. This was done within 70% less cost and half time as compared to the current industry standards ("AWS Startups Blog", 2019).

To accommodate a lot of data and process it in real time, Innovacer started using Hbase, a NoSql database which allowed live time-series data from HDFS (Hadoop Data File System). The system also began using spark to process large Gigabytes of Data in batch jobs from multiple sources within less than one hour. To ensure avoiding time managing clusters and troubleshooting issues with the increasing load, Innovacer leveraged the use of AWS managed services to scale operations up and down dynamically and efficiently ("AWS Startups Blog", 2019)

Being deployed on EC2 instead of inhouse clusters also enabled focus on improving customer service and reducing costs as well as data security concerns. This also led to customers being enabled to perform date analytical operations on large data sets in a few hours. Innovacer now provides an scalable EMR for using Spark jobs for complex models and predictive analytics. This facilitates on demand analytics for customers thus helping them achieve results within hours as opposed to days ("AWS Startups Blog", 2019)

Innovaccer since moving into serverless architecture is now using AWS lambda for its non production services and has now become a leader in helping healthcare professionals and organisations activate their healthcare data, garnering insights into ways care can be provided more efficiently. Using

AWS Data Lakes helped Innovacer drive data into powerful insights and informed decisions rapidly ("AWS Startups Blog", 2019).

Data Lakes Future/Vision

Data Lake a few years ago seemed like a dream, but now every major organisation has been working and implementing data lakes into their reality. On this journey of connecting enterprise data together, access to organisational data from a secure, well managed and controlled single point will soon become a necessity. This would lead to all major organisations opting for data solutions like Data Lakes which contain all organizational level data together and would integrate seamlessly with Data Tools and technologies as well (Future of Data Lakes, 2019).

Data Lakes offer better analytical capabilities for organisations wrestling with management, storage and processing of ever growing data volume. The speed of a data retrieval in data lakes is considerably higher than contemporary data warehouses thus major operations like necessary extraction/transformation/load (ETL), data cleaning, and exploration will be done on Data Lakes. The adoptions of IOT in many work areas will also enforce the usage of Data lakes to fulfil the purpose of Data Proliferation as well (Future of Data Lakes, 2019).

Businesses lately are more focused on data driven insights based decision making. Digitalization of business gives rise to an enormous quantity of data. Data Lakes will thus overtime emerge as a practical approach to finding solutions for increasing data as organisations will need advanced predictive and analytical data capabilities. The processing of data on cloud also acts as a catalyst to the growth of data lakes in future markets. Further work on using Data Lakes in organisations if focused towards Artificial Intelligence will help in achieving better business outcomes (Future of Data Lakes, 2019).

Organisations are basically looking for a simplistic and efficient way to facilitate the benefits of data lakes. Their upcoming requirement would be of an approach which utilizes a data approach without having to replace the existing system setup or retrain the whole staff. The system should be able to leverage the latest AI and ML technologies and be able to stay up and running for hours and days rather than weeks and months (Future of Data Lakes, 2019).

WorksCited

Data Lake. (2015). Retrieved December 01, 2020, from https://searchaws.techtarget.com/definition/data-lake

Data Lakes and the Data Lake Market: the What, Why and How. (2020). Retrieved December 01, 2020, from https://www.i-scoop.eu/big-data-action-value-context/data-lakes/

Data lakes on AWS. Retrieved December 01, 2020, from https://aws.amazon.com/solutions/implementations/data-lake-solution/#:~:text=AWS%20offers%20a%20data%20lake,or%20with%20other%20external%20users.

Data Lakes vs Data Warehouses - Snowflake (2020). Retrieved December 01, 2020, from https://www.snowflake.com/trending/data-lake-vs-data-warehouse

The Future of Data Lakes AI Martin. Mar 13, 2019. Retrieved December 01,2020, from https://medium.com/ibm-data-ai/the-future-of-data-lakes-33dada28fed6

History and Evolution of Data Lakes. (2020). Retrieved December 01, 2020, from https://databricks.com/discover/data-lakes/history

How Innovaccer is Helping Healthcare Organizations Activate Their Healthcare Data. 6 August 2019 Retrieved December 02,2020 from https://aws.amazon.com/blogs/startups/how-innovaccer-helps-organizations-activate-health-data/

Matthews, K. (2019). The difference between a data swamp and a data lake? Retrieved December 01, 2020, from https://www.information-age.com/data-swamp-data-lake-123481597/

Our data lake story: How Woot.com built a serverless data lake on AWS. 22 January 2019 Retrieved December 01, 2020, from https://aws.amazon.com/blogs/big-data/our-data-lake-story-how-woot-com-built-a-serverless-data-lake-on-aws/

Petrov, C. (2020, September 10). 25+ Impressive Big Data Statistics for 2020. Retrieved December 01, 2020, from https://techjury.net/blog/big-data-statistics/#gref