1. A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as-is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions.

Organizations that successfully generate business value from their data, will outperform their peers. An [Aberdeen survey](https://s3-ap-southeast-1.amazonaws.com/mktg-apac/Big+Data+Refresh+Q4+Campaign/Aberdeen+Research+-+Angling+for+Insights+in+Today's+Data+Lake.pdf) saw organizations who implemented a Data Lake outperforming similar companies by 9% in organic revenue growth. These leaders were able to do new types of analytics like machine learning over new sources like log files, data from click-streams, social media, and internet connected devices stored in the data lake. This helped them to identify, and act upon opportunities for business growth faster by attracting and retaining customers, boosting productivity, proactively maintaining devices, and making informed decisions.

2. There are multiple elements that constitutes a data lake. Data ingestion allows connectors to get data from different data sources and load data into the data lake. various examples of ingestion sources are database servers, web servers, emails, IoT, and FTP servers. Open source ingestion tools include Sqoop, Flume, Kafka, NiFi and other third-party tools.

When it comes to storage, data storage should be scalable, cost-effective, and allow fast access to data. The storage should support various data formats. HDFS is one mechanism to store data at scale. There are many other cloud-based solutions which offers large-scale data storage as well.

Once this data is ingested and stored, we need a mechanism to process this data at scale. Various tools such as MapReduce, Pig, Hive, spark, and many other third-party tools will come in to process this data at scale.

The next is data governance and security considerations. Data governance is a process of managing availability, usability, security, and integrity of the data in an organization. Data governance includes monitoring, logging, tracking the lineage, and various operations on the data. Data governance will become very crucial at one in building a data lake.

Security needs to be implemented in every single layer of the data lake. It start with storage and consumption. The basic needs to stop access for unauthorized users. It should support different tools to access the data with easy to navigate GUI and dashboards. Authentication, accounting, authorization, and data protection are some important features of data like security.

3. Depending on the requirements, a typical organization will require both a data warehouse and a data lake as they serve different needs, and use cases.

A data warehouse is a database optimized to analyze relational data coming from transactional systems and line of business applications. The data structure, and schema are defined in advance to optimize for fast SQL queries, where the results are typically used for operational reporting and analysis. Data is cleaned, enriched, and transformed so it can act as the “single source of truth” that users can trust.

A data lake is different, because it stores relational data from line of business applications, and non-relational data from mobile apps, IoT devices, and social media. The structure of the data or schema is not defined when data is captured. This means you can store all of your data without careful design or the need to know what questions you might need answers for in the future. Different types of analytics on your data like SQL queries, big data analytics, full text search, real-time analytics, and machine learning can be used to uncover insights.

As organizations with data warehouses see the benefits of data lakes, they are evolving their warehouse to include data lakes, and enable diverse query capabilities, data science use-cases, and advanced capabilities for discovering new information models. Gartner names this evolution the “Data Management Solution for Analytics” or “[DMSA](https://www.gartner.com/doc/3614317/magic-quadrant-data-management-solutions).”

4. Data lake breaks Data Silos. A company often has more than one product. Different products are handled by different departments or organizations. Oftentimes, these products are added into the product length of a company by doing mergers and acquisitions. Each product uses its own storage and database systems to store the required data. This makes it incredibly hard to analyze the customer from a 360 degree perspective. You really want to understand the customer behavior across all of your product lines and not just one, because customers may purchase multiple products. Since each organization manage

each product and held their own system of storing the data, this creates Data Silos. Data Lake solves this problem by bringing all the crucial enterprise data under one centralized system. This makes it easy for

different organizations within a company to collaborate and analyze the data.

In a large enterprise, the most powerful impact of a Data Lake is the enablement of innovation. With the help of Data Lake, you can now easily perform exploratory as well as advanced analytics which was not possible before, enabling many many new possibilities.

In today's business environment, data consumption requirements and use cases emerge extremely rapidly. The entire philosophy of a Data Lake revolves around being ready for an unknown use case. When the source data is in one central lake with no single controlling structure or schema embedded within it, supporting a new additional use case can be much more straightforward.

Historically, it takes weeks or months within a company to enable access to a dataset for the data analyst within the company. With help of a Data Lake, you can build a self-service culture within the company. Allow the business people to access whatever slice of the data they need, letting them develop the reports they want using any of the wide range tools available. IT becomes the custodian of the infrastructure and data on the Cloud while business takes responsibility for exploring and mining it.

Data Lake can generate a huge amount of value for the organizations. Organizations that successfully generate business value from their data will outperform their peers. It enables the leaders within the organization, new types of analytics like machine learning over new sources like log data, click streams, social media, and internet connected devices. All of those data stored in one central Data Lake. This helps leaders identify opportunities for business growth by retaining the customers, boosting productivity, or proactively maintaining the devices, making much smarter, informed, and productive decisions.

5. Deploying Data Lakes in the cloud

Data Lakes are an ideal workload to be deployed in the cloud, because the cloud provides performance, scalability, reliability, availability, a diverse set of analytic engines, and massive economies of scale. [ESG research](https://s3-ap-southeast-1.amazonaws.com/mktg-apac/Big+Data+Refresh+Q4+Campaign/ESG-White-Paper-AWS-Apr-2017+(FINAL).pdf) found 39% of respondents considering cloud as their primary deployment for analytics, 41% for data warehouses, and 43% for Spark. The top reasons customers perceived the cloud as an advantage for Data Lakes are better security, faster time to deployment, better availability, more frequent feature/functionality updates, more elasticity, more geographic coverage, and costs linked to actual utilization.

Build your Data Lakes in the cloud on AWS

AWS provides the most secure, scalable, comprehensive, and cost-effective portfolio of services that enable customers to build their data lake in the cloud, analyze all their data, with a variety of analytical approaches including machine learning. As a result, there are more organizations running their data lakes and analytics on AWS than anywhere else. Because SQL-based access is needed, Redshift is added to the storage layer. Redshift is a fast and fully managed petabyte-scale data warehouse that costs less than $1,000 per terabyte per year.

A data lake on AWS leverages S3 for secure, cost-effective, durable, and scalable storage. Amazon S3 also offers an extensive set of features to provide strong security for the data lake, including access controls and policies, data transfer over SSL, encryption for data at rest and in motion, logging and monitoring, and more.

Data can be quickly and easily ingested into Amazon S3 from the SQL Server, FTP server and APIs by Direct Connect, AWS Snowball, and Amazon Kinesis.

For processing and analyzing the data stored in Amazon S3, AWS provides fast access to flexible and low-cost services, like Amazon EMR, Amazon Redshift with Redshift Spectrum, Amazon Athena, and Amazon AI services, so any analytical solution can be rapidly scaled to power any big data applications, meet demand, and improve innovation.

For governing and securing the data, AWS Glue, Amazon DynamoDB, and Amazon ElasticSearch can be leveraged to catalog and index the data in Amazon S3. Using AWS Lambda functions that are directly triggered by Amazon S3 in response to events such as new data being uploaded, the catalog can be easily kept up-to-date. With Amazon API Gateway, API can be created that acts as a “front door” for applications to access data quickly and securely by authorizing access via AWS Identity and Access Management (IAM) and Amazon Cognito.