# Growth of Data: Data lakes, Machine learning, Data Warehouses

The amount of data that is generated and recorded in the world has been growing exponentially in the past 10 years and will continue to grow in the coming future. In 2010 the amount of data in the world was estimated to be 2 Zettabytes, and by May 2020 this has grown to 59 Zetabytes. Models predict that the exponential growth shall continue and reach about 150 Zetabytes by 2024. The main concern with the growth of data storage is that compute resources need to scale also. Businesses should no longer need to make large upfront hardware investments or maintain unutilized infrastructure. The physical space and administrative overhead that is budgeted per year may leave a business vulnerable to overspending or under provisioning resources. This balance is going to be more difficult to maintain as technology scales to incorporate and process more data. Network infrastructure transforms to meet the demands of the technology that is being utilized, the datacenter itself is evolving past the sum of its parts. What has evolved from a single server has scaled to a server stack, which has itself scaled to a cluster hypervisor serving compartmentalized instances of runtime environments. The technology has dramatically changed in the past 10 years and will continue to do so in the future.

Businesses will begin to shift towards a consumption based model for processing and storage, the companies that will provide cloud solutions will help their clients benefit from the massive economies of scale. Thus resources will be cheaper to utilize when they are reduced to a more granular scale of consumption. This micro services model allows businesses to stop guessing about capacity and launch working prototypes in minutes without risking massive capital expenditure. This flexible model allows growth to match utilization and accounting teams can closely relate expenditure with productivity.

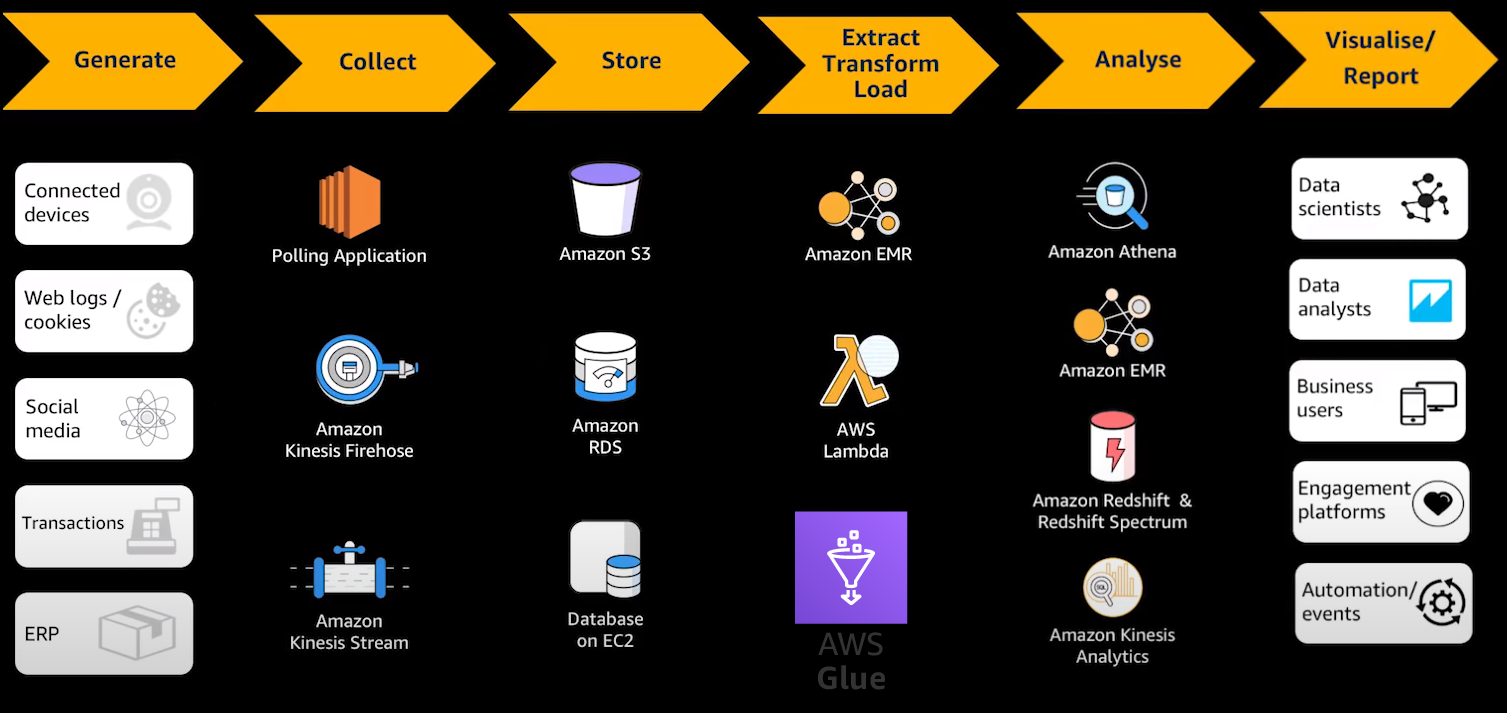
The technology behind ingestion of data evolves over time. Simple Query Language (SQL) databases evolved and changed into multiple other kinds of metadata or document style databases. The ingestion and processing of data would evolve over time also. Eventually the ingestion of data needed to evolve to consume the entire stream of data from multiple sources in a business. This gave rise to Data lakes and Data warehouses. OLAP (Online Analytical Processing) would allow multidimensional comparison of data tables however the issue of converting this data and governance would need to be resolved first.

Data lakes would act as the repository of data streams for a business. This source of vast unprocessed data would need to be extracted, transformed, and loaded into a standardized data warehouse that would act as an immutable source of applicable data. These sources of data would be the fundamental building blocks for business intelligence.

In a Cloud based architecture we would need to begin this process by aggregating and collecting data. The raw form of ungoverned data would be gathered from sources such as: Databases, Spreadsheets, IOT sensors, Log files, Applications, and Block storage like S3. Then all would need to be funneled through an ETL service. The Growth of the storage layer would need to be matched by the Processing layer. Open source ETL solutions like Spark/Hadoop would need to be run on dedicated cluster solutions. These solutions can be hosted on cloud infrastructure however because of the growth of Data, this infrastructure needs to scale dynamically. Business run a risk of overprovisioning systems or missing SLA constraints. Amazon EMR allows dynamic scaling of storage and compute resources, data can be stored in Block storage S3, and compute instances can be managed by AWS EC2. Clusters can be elastic and easily scalable to meet demand. This type of infrastructure is good for a predictable growth model; however, an unpredictable dynamic growth model would benefit from a more granular form of ETL processing which is found in Serverless AWS Glue.

Combining a consumption model and serverless architecture, services like AWS Glue can allow dynamic data integration and processing. This allows easier prototyping and batch processing for data that has unpredictable growth and utilization cycles. This model is good for lower administrative costs as well. I would highly recommend AWS Glue for startup companies that need to process data but do not have the administrative overhead to provision large datacenter or compute cluster resources. The infrastructure and software is managed by AWS.

# AWS ETL PIPELINE



1. Generate Data: Data can be generated from multiple sources, in the past transactional data would be the biggest data stream but with the growth of IOT and Web logging systems we can see a growth in more granular data collection from various sources.
2. Collection of Data: Polling applications would need to gather batches of data over time, real-time streaming of data would be accomplished by systems like Kinesis.
3. Storage of Data: Cloud block storage of data can be used to synchronize on premise data solutions and be an almost limitless source of storage. Other database storage services can be Amazon RDS or Databases hosted on EC2 compute infrastructure.
4. ETL of Data: Extraction transformation and loading of data. Data in its raw form needs to be extracted from varied sources and be transformed to be usable to create models. Data governance is very important when considering things like legacy and disparate data. Because of the sheer volume of data ingestion we also need to consider the infrastructure of data processing tools and will need to implement a cluster (EMR) or serverless solution (AWS GLUE). Once data is processed correctly and efficiently, the data will be prepared for machine learning.
   1. Disparate data: Data may be changed due to licensing, usage of different systems, legacy systems, legacy formats etc.
   2. We will want to process and create a unified data set. Target Schemas may change so a process to transform and load data is absolutely necessary.
5. Analyze Data: Different programs can analyze the processed data
   1. Athena: Query engine against data in S3
   2. Redshift: Data warehouse for OLAP analysis of Multidimensional data
   3. Kinesis Analytics: Analysis of real-time data
6. Application of data: Data Aggregation can allow models to be produced that will allow machine learning to be applied to real world scenarios. For example the processing of various medical radiology records can allow machine learning AI to utilize learning algorithms to act as a diagnostic aid to help radiologists and oncologists in cancer imaging. A more accurate form of cancer diagnosis driven by machine learning can have an epidemiological effect on cancer treatment.