Summary

Technology and Infrastructure

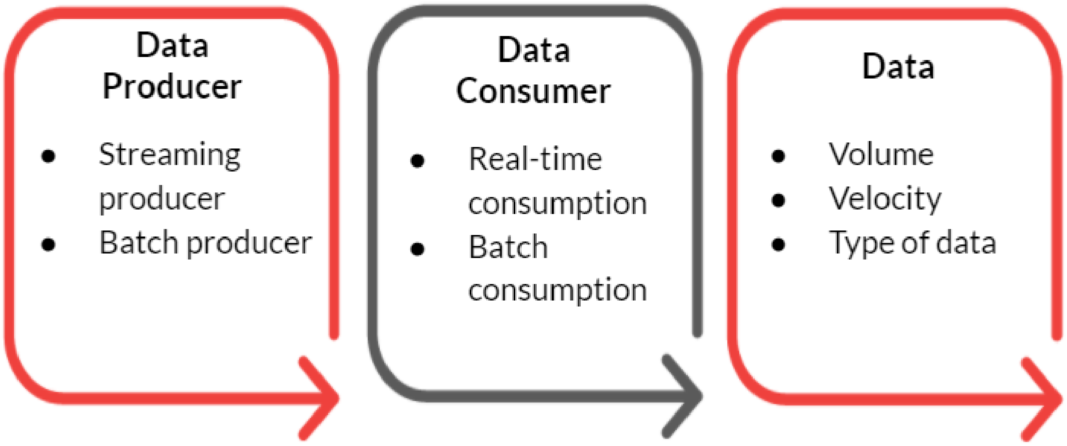
As a Manager, you might not need to make the decisions regarding which technology or tool to use, but having an understanding of the different technologies and tools available to support your AI/ML strategy would definitely aid in better communication with data engineers and data scientists, help in understanding their requirements and aligning the stakeholders

Technology and infrastructure are needed for the two driving forces of an AI/ML initiative:

1. Data (Developing and implementing AI/ML solution)
2. AI/ML solution (Developing and implementing AI/ML solution.

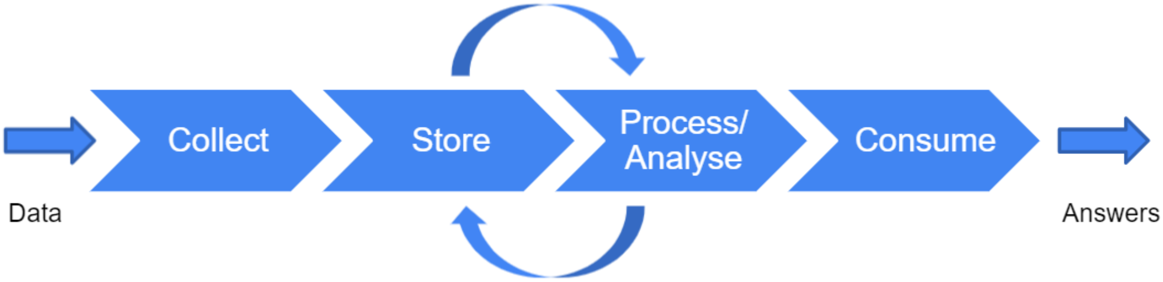
Technology for Data Engineering

Data engineering is a combination of the characteristics of data producers and data consumers along with the characteristics of data itself.



A data pipeline refers to setting up a path through which you can store and process the data you need to collect based on the type of data, the consumer characteristics and the producer characteristics - to extract meaningful insights.

The components of the data pipeline are:



Phase I : Collect

The choice of tools to be used for collecting data must always start with taking the nature of the data into account.

Data can be classified into the following three categories according to its temperature.

Temperature	Volume	Latency	Request rate	Cost
Hot	MB-GB	ms or <ms	Very high	Very high
Warm	GB-TB	ms, s	High	High
Cold	PB-EB	Min, hrs	Low	Low

Data can also be classified into the following types according to their nature.

Type	Source
Transactional Data	Applications such as web apps, mobile apps, data centres, etc.
File/Object data	Logging, search, etc

Events/Streaming data	IoT devices, sensors, messages, etc.

Phase II : Store

For each class of data, you will need different types of tools to store it. It is essential to consider an appropriate storage option based on the data temperature in order to optimise not only for latency but also for cost as well.

Based on the temperature of data:

Type	Tools available (AWS)
Hot data	Amazon S3
Warm data	Amazon S3
Cold data	Amazon Glacier

Based on nature of data:

Type	Tools available (AWS)
Transactional Data	Amazon ElastiCache (Cache) Amazon DynamoDB (Unstructured - NoSQL) Amazon RDS (Structured - SQL)
File/Object Data	Amazon S3 (Logging) Amazon ElastiSearch (Search)
Events/Streaming data	AWS SQS (messaging) AWS Kinesis (streaming data)

Combining the temperature and nature of data:

Temperature	Structure	Nature	Choice of tool (AWS)
Cold data	Unstructured file Storage	File data	Amazon Glacier
Warm	Semi or Unstructured file Storage	Logging data	Amazon S3
Warm	Semi-Structured Search	Logging and Searching data	Amazon ElastiSearch
Warm	Structured Data	Transactional Data	Amazon RDS
Hot	Unstructured data	Transactional data	Amazon DynamoDB
Hot	Unstructured Data	Transactional data	Amazon ElastiCache

Phase III : Process

Based on the input and the output of the process, most processes can be divided into the following two classes:

Processing	Description	Tools available (AWS)
Batch processing	This method is used to run high value and repetitive data tasks. It runs in the background and requires no human interaction.	AWS Batch AWS EMR
Streaming processes	This method focuses on the real-time processing of a continuous stream of data.	AWS Kinesis

There are two other broad categories as well:

1. Interactive processing: This method requires user interaction.
2. Predictive processing: This method involves making predictions using an AI/ML model.

Both of these encompasses multiple latencies and, depending on the use case, could be either batch or streaming.

Phase IV: Consume

The two sets of tools available based on the consumers are:

1. Data scientists and engineers can use lower-level tools for the development and implementation of AI/ML solutions, such as Anaconda and Jupyter.
2. Software engineers and business analysts need higher-level tools for data visualisation, such as Tableau, Amazon and QuickSign.

Technology for Implementing AI/ML Solution

Some key aspects to consider while focussing on technology and infrastructure for the implementation of the AI/ML model

1. Real-time vs Batch implementation
2. Fault tolerance and high availability
3. Load balancing
4. Model governance and tracking

Disclaimer: All content and material on the upGrad website is copyrighted material, belonging to either upGrad or its bona fide contributors, and is purely for the dissemination of education. You are permitted to access, print, and download extracts from this site purely for your own education only and on the following basis:

- *You can download this document from the website for self-use only.*
- *Any copies of this document, in part or full, saved to disc or to any other storage medium, may be used for subsequent, self-viewing purposes or to print an individual extract or copy for non-commercial personal use only.*
- *Any further dissemination, distribution, reproduction, and copying of the content of the document herein, or the uploading thereof on other websites, or use of the content for any other commercial/unauthorized purposes in any way that could infringe the intellectual property rights of upGrad or its contributors is strictly prohibited.*