Big Data Trends

Rise of the New Age “Data Curators”

- In the coming times, data curator will become a very significant role.
- In simple terms, data curation implies ensuring that people can easily find and use the data now and in the future.
- It can be detailed down to:

1. Identifying the most relevant data sources.
2. Sourcing (getting) the data from data owners.

Rise of the New Age “Data Curators”

3. Fixing any issues with the data such as missing values, unexpected values, and mysterious variables. Also preparing the data if not suitable for analysis;

for example, if the data has too much details or is too granular to be picked up for analysis then maybe summarization or aggregation might help in the analysis.

4. Using annotations, tags, appropriate and crisp documentation, etc. to help make the data easy to find, use, reuse and present.

5. Preserving the data such that it is available for use, reuse, etc.

Rise of the New Age “Data Curators”

CDOs are Stepping Up

- With data becoming the new oil, the clear mandate is to create more and more value from the organization's data.
- Enter the role of CDO (Chief Data Officer) to help with **data leveraging** (use the existing data assets in the best possible way), **data enrichment** (augment the value of data by blending, bringing together internal and external data),
- **data monetization** (exploring newer avenues of earnings and revenues), **data upkeep** (ensuring proper data quality and governance), **data protection** (ensuring security and adequate protection of data), etc.

Few statistics:

- Estimated number of CDOs globally as per Gartner:

  2010: 15

  2014: 400

  2018: 4000+

	CDO	
Year	Yes	No
2012	12%	88%
2017	56%	44%
2018	63%	375

What is Dark Data?

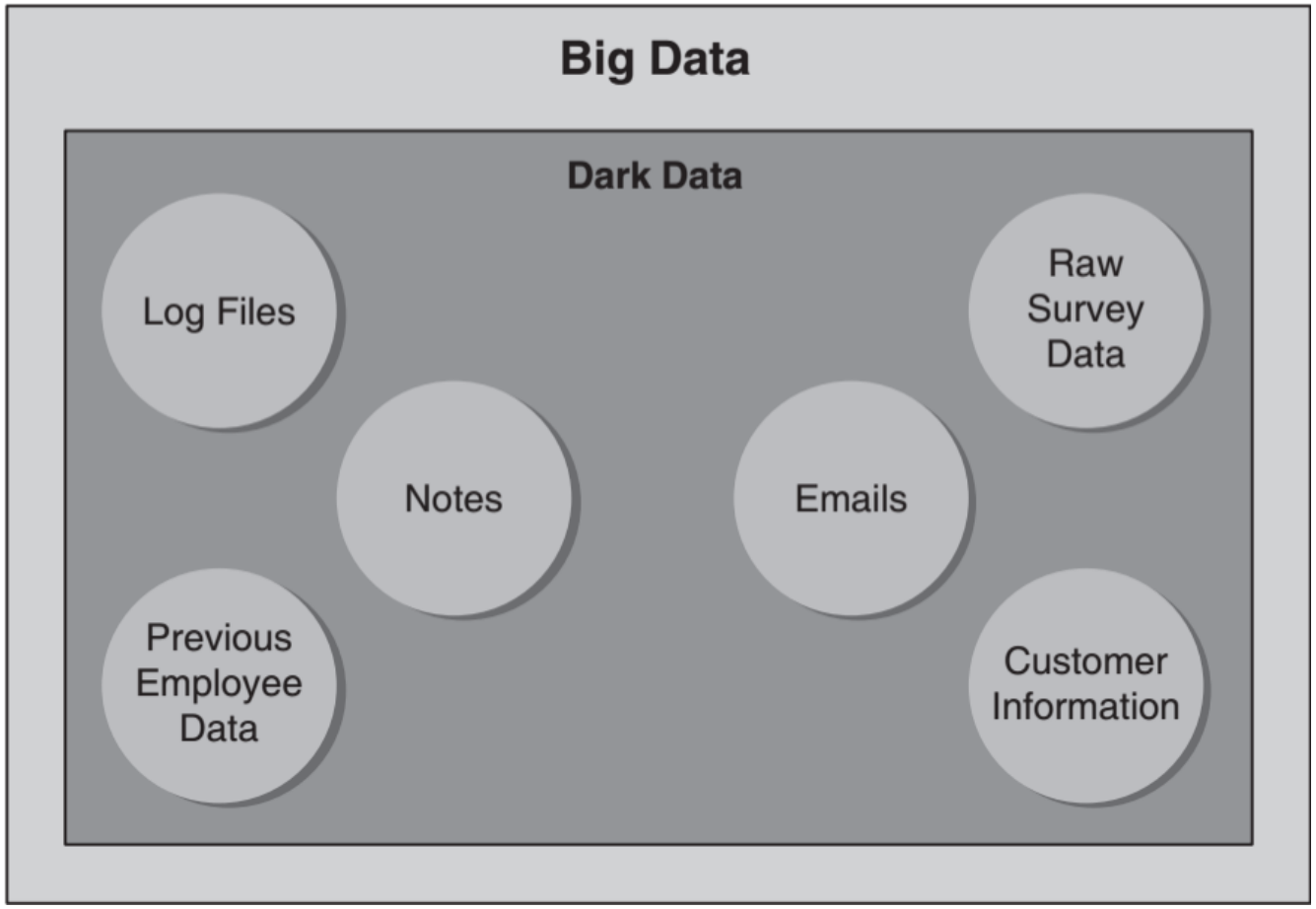
- According to Gartner, dark data is “The information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes”.

- Data that is big in volume, variety and velocity. Not all the big data that is collected by the organizations are processed, analyzed or used. This data that is collected and stored by the organizations thinking that it will be used for some analysis sometime in future but is never put to any use is “dark data”.

What is Dark Data?

- Dark data is a subset of big data. It also happens to constitute the biggest portion of the large volume of big data that organizations collect and store.
- As per IDC (International Data corporation), 90% of unstructured data is “dark data”.

What is Dark Data?



Why is it important then to consider dark data?

- Primarily because this dark data could mean opportunity or opportunities lost for an organization.

- It may have untapped, undiscovered useful insights which could spell success for the organization.

- Another reason why it becomes important is if the dark unanalyzed data is not handled well, it can result in a lot of problems such as legal and security problems.

Why the Dark Data has not been handled the way it should have been?

1. Not knowing what to do with the data:

Picture this: A bank receives online applications for credit cards. Their focus is mainly on collecting and analyzing customer details and eligibility. They attach little or no priority to finding out how the customer came to the application page. If this data was collected and analyzed, it could have provided useful insights about the usage of the bank website or could have helped improve the application.

Why the Dark Data has not been handled the way it should have been?

2.Disconnect among departments:

Typically in large organizations, departments operate in silos. They have their own data collection and storage processes. They may or may not reveal these processes to other departments. So, there is a good chance of data lying unused even though the possibility is that some or all of this data may be relevant/useful to some other department.

Why the Dark Data has not been handled the way it should have been?

3.Technology and tool constraint:

If we take the case of a large organization, there is a high possibility that all applications do not use the same technology and tools for data collection, storage, etc. Sometime the integration of all this data becomes a problem and at times impossible. There could be integration issue, data quality issues, data governance issues, data ownership issues, etc.

What problems can dark data cause?

- **Opportunity lost:**

This is the data that is as yet untapped. It may have useful insights which can help the company surge ahead of competition.

- **Legal and regulatory issues:**

A lot of this data is lying around sometimes secured, sometimes unguarded. Any inadvertent disclosures could lead to intelligence risk, legal liabilities and even loss of reputation for the firm.

What is the way forward when it comes to handling dark data?

- Properly structuring or categorizing the data will go a long way in ensuring that the process of searching for the data later can be eased out.

- Securing the data by way of encryption, etc.

- Having clearly defined policies for dark data retention as well as disposal.

Streaming the IoT for Machine Learning

- Machine Learning uses “stored data” for training in a “controlled” learning environment.

- With the rise of IoT (Internet of Things), the need of the hour is to use “streaming data in real time” in a “much less controlled environment”.

- This will help to provide more flexible, more appropriate responses to a variety of situations including communicating with humans.

Streaming the IoT for Machine Learning

- Today, IoT data, streaming analytics, machine learning, and distributed computing have come together to offer a very powerful, yet an inexpensive proposition to store and analyze big volume and varied types of digital data.

- Some examples of IoT, Big Data, and Machine Learning working together include:

1. Healthcare: Continuous monitoring of chronic diseases.
2. Smart Cities: Traffic patterns and congestion management.
3. Transportation: Optimizing routes and fuel consumption.
4. Automobile: Smart cars.
5. Retail: Location-based advertising.