## **Data Ingestion**: A process of obtaining and importing data for immediate use or storage in a database. Data can be streamed in real time or ingested in batches.

## **Data Lake**: Data lake is a system of insights. It’s a collection of data platforms that receive data from many different systems across the organization. Its cataloged together to enable different people in the organization to find the data they need for advanced analytics.

## **Data Profiling**: Data profiling is the process of examining data available from an existing information source (e.g. a database or a file) and collecting statistics or informative summaries about that data.

## **Data Wrangling**: A process of cleaning and unifying messy and complex data sets for easy access & analysis. This process typically includes manually converting/mapping data from one raw form to another to allow convenient consumption & organization

**Data governance** : It refers to the overall management of the availability, usability, integrity, and security of the data employed in an enterprise. A sound data governance program includes a governing body or council, a defined set of procedures, and a plan to execute those procedures.

## **Data Lineage –** A process of tracking the life cycle of data from the point where the it originated and where it is transforming to over a period of time. It also describes what happens to data as it goes through the diverse process. It helps provide visibility into data analytics pipeline and simplifies tracking errors back to their sources

## **Edge Node / Gateway Node:** An interface between Hadoop cluster and outside network. They are used to run client applications and administration tools. They are often used as a staging area for data being transferred into Hadoop cluster. As such **Oozie, Pig, Sqoop, Hue & Ambari** will run in Edge Node. Hardware requirements of Edge Node is different from Master & Salve nodes.

**Kerberos:** A network authentication protocol, it is designed to provide strong authentication for client/server application by using secret-key cryptography.

## **Micro Batching**: Micro-batching is a technique that allows a process or task to treat a stream as a sequence of small batches or chunks of data. For incoming streams, the events can be packaged into small batches and delivered to a batch system for processing. Apache Storm can be used for Micro Batching

**Data Curation:** Data curation is a broad term used to indicate processes and activities related to the organization and integration of data collected from various sources, annotation of the data, and publication and presentation of the data such that the value of the data is maintained over time, and the data remains available for reuse and preservation. Data curation includes "all the processes needed for principled and controlled data creation, maintenance, and management, together with the capacity to add value to data".[1] In science, data curation may indicate the process of extraction of important information from scientific texts, such as research articles by experts, to be converted into an electronic format, such as an entry of a biological database.[2]

The term is also used in historical uses and the humanities,[4] where increasing cultural and scholarly data from digital humanities projects requires the expertise and analytical practices of data curation.[5] In broad terms, curation means a range of activities and processes done to create, manage, maintain, and validate a component

**Jypyter:** The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, machine learning and much more.

## **RDF**: Resource Description Framework

## **RDF Triples**: An RDF triple contains three components:

## the subject , which is an RDF URI reference or a blank node.

## the predicate , which is an RDF URI reference.

## the object , which is an RDF URI reference, a literal or a blank node.