DATA ENGINEERING FOR ENTERPRISE

UNIT – I

DEFINITION OF DATA ENGINEERING

- Data engineering is the practice of designing and building systems for collecting, storing, and analyzing data at scale.
- It is a broad field with applications in just about every industry.
- Organizations have the ability to collect massive amounts of data and they need the right people and technology to ensure it is in a highly usable state by the time it reaches data scientists and analysts.

WHAT IS DATA ENGINEERING?

- The key to understanding what data engineering lies in the "engineering" part.
- Engineers design and build things. "Data" engineers design and build pipelines that transform and transport data into a format wherein, by the time it reaches the Data Scientists or other end users, it is in a highly usable state.
- These pipelines must take data from many disparate sources and collect them into a single warehouse that represents the data

uniformly as a single source of truth.

WHY DATA ENGINEERING IS IMPORTANT?

- Data engineering is important because **it allows businesses to optimize data towards usability**.
- For example, data engineering plays a large role in the following pursuits: Finding the best practices for refining your software development life cycle.
- Tightening information security and protecting your business from cyberattacks
- Increasing your understanding of business domain knowledge •

Bringing data together into one place via data integration tools

DATA ENGINEERING DRIVING FACTORS

Data engineering facilitates organization data management and data analytics platform

Data engineering Acts as a base for Predictive analytes or Data Science on demand

THE BUILDING BLOCKS OF DATA ENGINEERING ARE

Data Collection/Data Extraction

Data Processing

Data Transformation

Business Intelligence

Data Governance

Data Privacy

Data Collection: Date comes from a variety of sources. They must be collected, integrated, transformed, and loaded into the target systems. This process is called Data Collection.

Data Processing: The Storage and the processing layer is the fundamental base of any architectural data platform. There are three types of frameworks in use.

Data Warehouses: Data warehouse is the linchpin of any data platform architecture. They are selected by choice because of their power and computing speed.

Data Lakes :Data lake is an architectural design with the data processing done in the cloud with 53 buckens and integrated with other external tools like the apache spark for distributed processing These architectural frameworks are the cheapest data storage made for storing a large amount of raw data.

Data transformation in an organization, there are typically twe implementation strategies. The companies with data warehouse architecture rely on tools dbt. The other implementation is with the tools like airflow.

Business Intelligence: Standalone Business intelligence like the dashboats and the reports have been there for ages, but now the BI tools with inbuilt into the data platform. The in-built tools allow plotting graphs visuals and more explanation.

Data Catalog and governance: In general, the stata platform is modest,

but when it comes to data discovery, trust has great challenges.

Data Privacy: As the data platform spectrum grows, so do the privacy and access control artifacts. It might be too early but there are industrial players who have the capacity to apply data privacy and access controls across the data stack.

DATA ENGINEERING ADVANCEMENT

The role of a data engineer will take an uplift especially now, with the inclusion of Artificial Intelligence and Machine learning data, engineers are more focused on building the data pipeline, predictive analytics and extracting hidden insights from Data

Many automation tools in the market will surely decrease the amount of data engineer requirements. Nevertheless, data engineering will play a significant role in data analytics.

Data Engineering is set to improve how to transform, manage and analyse data. As a result the companies will need to reformulate their gears and needs in data analysis journey.

Data Engineering Skills and Expertise

To become a data engineer, one must possess the following critical skills.

Programming Skills

Data engineering requires one to understand fundamental programming concepts like algorithms, date structure, etc. Python is the most commonly used library data engineering tasks Hence one must have good python skills. Another important programming language gaining popularity is the Scala.

Data Skills

Data engineering requires one to have good hands-on data-related skill They must be able to collect data from various sources integrate them process and store them; hence a Data Engineer must be comfortable working with structured, Unstructured datasets and processing them efficiently.

Cloud Skills

A data engineer must also be good at working with cloud services. They must be comfortable working with cloud services like the google cloud, AWS and Microsoft Azure, in addition, they must be able to process data in distributed way using cloud services.

DATA ANALYTICS OVERVIEW

- Data analytics is the science of analyzing raw datasets in order to derive a conclusion regarding the information they hold. It enables us to discover patterns in the raw data and draw valuable information from them.
- Data analytics processes and techniques may use applications incorporating machine learning algorithms, simulation, and automated systems. The systems and algorithms work on the unstructured data for human use.
- These findings are interpreted and used to help organizations understand their clients better, analyze their promotional campaigns, customize content, create content strategies, and develop products. Data analytics help organizations to maximize market efficiency and improve their earnings.

DATA ANALYSIS VS DATA ANALYTICS

- Data analysis is a process involving the collection, manipulation, and examination of data for getting a deep insight. Data analytics is taking the analyzed data and working on it in a meaningful and useful way to make well-versed business decisions.
- Data analysis helps design a strong business plan for businesses, using its historical data that tell about what worked, what did not, and what was expected from a product or service. Data analytics helps businesses in utilizing the potential of the past data and in turn identifying new opportunities that would help them plan future strategies. It helps in business growth by reducing risks, costs, and making the right decisions.

- Tools used for data analysis are Open Refine, Rapid Miner, KNIME, Google Fusion Tables, Node XL, Wolfram Alpha, Tableau Public, etc. Tools used in Data analytics are Python, Tableau Public, SAS, Apache Spark, Excel, etc.

- Data analytics is more extensive in its scope and encompasses data analysis as a sub-component. The life cycle of data analytics also comprises data analysis as one of the significant steps.

- Data analytics and data analysis, both are essential to understand the data as the first one is useful in estimating future demands and the

second one is necessary for gaining insight by analyzing the details of the past data. Data analysis is actually studying past data to understand ‘what happened?’ Whereas data analytics predicts ‘what will happen next or what is going to be next?’

DATA ANALYTICS ADVANCEMENT

- Data analytics is a field which witnesses a continuous revolution. Since data is becoming increasingly valuable with each passing time, it has been now treated with great care and concern.

- To cope up with the constant changes in the industries and societies as a whole, new tools, techniques, theories, and trends and always introduced in the data analytics sector.

DATA WAREHOUSING : OVERVIEW

Businesses need a central repostory that houses all the business-critical information integrated from multiple data sources. This central pository is called Data Warehouse. From the Data Warehouse, one can make data for data analysis. A Datawarehouse possesses data from a varied or diverse set of sources ut msource. A Datawarehouse is used to analyse multidimensional data

There are multiple departments across the operation of a business or a company

Data warehouses can create customed reports for the users that greatly help decision making.

Building Blocks of Data Warehouse

The architecture of a data warehouse posseses all the components used together. Thes architectural paradigm's focus is delivering the constant data, so the design is more centred and transformation rather than just data preparation and data storage. Hence, the architecture is a framework composed of components used to develop and deploy a data warehouse. The building blocks of a data warehousing architecture are

Storage of Data

Delivery of information

Management and control

Date about Data

Source Data

The Source data coming from the data house architectures is of four types. They are

Production data

Internal data

External Data

Archived data

Production data comes from various data systems like the database, ERP and CRM systems and Supply management systems etc. Internal data is the company's internal data. External data comprises data from external source websites. Archived data is the periodically taken data stored in an archive

Staging Data

The data is collected from multi data systems to the staging area. The major process of Extraction, Transformation and loading happen at the staging area. Hence, in other words, the staging area is a location where data cleaning, data preparation, and deduplication happens

Data Storage

We can make out that huge historical stata get stored in the database from the name. The data stored has some structure and usually, the data warehouse follows a structure of the relational database management systems, but some of the data warehouses also follow a

multidimensional modelling structure

Information Delivery

Once going through the STL process Extract Transform and load the data reaches the information delivery area. This area supplies the data for the queries and the reports.

Management and Control

All the components of a data warehouse architecture need to support one another by the managers and control system. The information and the control components sit over the other components and work with the services of the data warehouse

DATA WAREHOUSING CHALLENGES

Below are some of the data warehouse challenges

Lack of Governance:

A data warehousing solution must integrate data from different source systems and functions, bringing the business into a tricky situation, and they may not be keeping an eye on the data vernance and compliance factors

Access Control:

Organisations must control the access to data warehouses and pre access to the data warehouse to only those departments who actually need to use it. It is an absolute must for the organization and some fail to implement it.

Inefficient Data Strategy:

Again, the organizations must put a well-deed day to succeed in the market. Without a well-planned data strategies the business will struggle to adapt to the data warehousing setup and will not be able realise the potential benefits of a data warehousing

Data Warehousing Limitations

The Date warehouse house set up is different from the databases because the data warehouses consolidate data from multiple sources located in a particular physical location

1. The Cost-Benefit Ratio

2. Data Ownership

3. Data Flexibility

4. ETL Processing Time
5. Hidden Issues

6. Not Collecting the Required Data

DATA WAREHOUSING INDUSTRY NEEDS AND MARKET DEMAND

- Data warehousing market will reach at an estimated value of USD 43.45 billion by 2028 and grow at a CAGR of 8.50% in the forecast period of 2021 to 2028. **Rise in the demand for storage system for growing volume of data acts as an essential factor driving the data warehousing market**.
- Data warehousing is defined as a process which is used to store large amount of data and information by various sources of business and organizations. It uses different data and component so that they can use this data efficiently. They are specially designed for query and analysis rather than transaction process. Extractions, statistical analysis, data mining among others are some of the

common type of offering offered by the data warehousing. The main benefit of the data warehousing is that they separate analytical process from operational process which increase the operational system.

- Rise in the need for real- time view and analytics on real data on operational data and rise in the applications of AI in data warehouse are the major factors among others boosting the data warehousing market. Moreover, rise in the demand from emerging economies will further create new opportunities for data warehousing market in the forecast period mentioned above.

DIFFERENCE BETWEEN DATA ENGINEERING AND DATA WAREHOUSING

- Data engineers are engineers that handle data transformation and storage activities for any applications, while data warehouse engineers handle data transformation and storage activities associated with building a data warehouse.

- **Data Warehousing Engineers** are responsible to build the data warehouse applications to support business intelligence requirements of a company. There are several niche roles within

the umbrella of data warehousing engineers.

- Data Warehousing engineers are high profile engineers and generally work for very large global companies, they sit in the ivory tower headquarters of these companies or offshore development centers and use high end and expensive technologies like informatica, cognos, business objects, oracle, teradata etc.

Data Engineers

- The term “Data engineer” is relatively new, and has become popular recently due to the various new applications within the “big data analytics” area. Due to various advancements in technology and digital revolution, data is now being used for many other applications, apart from just business intelligence.
- Various digital products are acquiring and using data for improving efficiency, transparency etc. and understanding customer behavior patterns.Smart cities are using data to become smarter. Data is being used for automation, artificial intelligence, and swarm intelligence.
- Machines are generating data, internet is generating data and all sorts of data is now available via APIs.

DATA ENGINEERING BUILDING BLOCKS

They use data and analytics to understand how their customers are actually using their products and services, how their operations and supply chain are performing, how to manage their workforce and how to identify key risks – and then they act upon these insights.

DATABASE AND NO SQL DATABASE

Before we dive into the concepts of Relational Databases and Net Databases, let us understand a few things

Data

Data is a group of small pieces of information, and data can be in any form such as numbers, strings, images or wide audio. We can store data anywhere on a piece of paper, hard disk drives or databases. The data must be translated or converted into a different form for processing

Database

A Database is a place where the collection of small pieces of data in an organized form is stored to be managed. retrieved and processed easily.The main abjective is to operate a huge amount of data by staring

and processing R

Examples of a Database

Oracle

Mysql

IBM DB2
Microsoft SQL Server

NoSQL Database

NoSQL expanded as Not only SQL is a form of database that is non tubuler. The methods of staring date in NaSQL databases are different in NoSQL Databases. The NoSQL databases are various types such as keyvalues and graphs. The main advantage of NoSQL database 48/145 flexible schema types can be scaled with no effort.

Examples of a NoSQL Databases

MongoDB

Cassandra

Hbase

Big Table

Data Model, ERD model, MDM model, Star scheme, Extended Star Schema, Snowflake Schema

DATA MODELS

The Data models define how we can model a database. The data model establishes a connection between the master and the transactional tables and how the data inside is processed.

Data models are **visual representations of an enterprise's data elements and the connections between them**. By helping to define and structure data in the context of relevant business processes, models support the development of effective information systems.