UNIT - I Data Management: Design Data Architecture and manage the data for analysis, understand various sources of Data like Sensors/Signals/GPS etc. Data Management, Data Quality(noise, outliers, missing values, duplicate data) and Data Processing & Processing.

DESIGN DATA ARCHITECTURE AND MANAGE THE DATA FOR ANALYSIS

Data architecture is composed of models, policies, rules or standards that govern which data is collected, and how it is stored, arranged, integrated, and put to use in data systems and in organizations. Data is usually one of several architecture domains that form the pillars of an enterprise architecture or solution architecture.

Data architecture defines information flows in an organization, and how they are controlled. A data architect is responsible for understanding business objectives and the existing data infrastructure and assets; defining data architecture principles; and shaping the enterprise data architecture to provide greater benefits to the organization.

Various constraints and influences will have an effect on data architecture design. These include enterprise requirements, technology drivers, economics, business policies and data processing needs.

Enterprise requirements:

These will generally include such elements as economical and effective system expansion, acceptable performance levels (especially system access speed), transaction reliability, and transparent data management. In addition, the conversion of raw data such as transaction records and image files into more useful information forms through such features as data warehouses is also a common organizational requirement, since this enables managerial decision making and other organizational processes. One of the architecture techniques is the split between managing transaction data and (master) reference data. Another one is splitting data capture systems from data retrieval systems (as done in a data warehouse).

Technology drivers:

These are usually suggested by the completed data architecture and database architecture designs. In addition, some technology drivers will derive from existing organizational integration frameworks and standards, organizational economics, and existing site resources (e.g. previously purchased software licensing)

Economics:

These are also important factors that must be considered during the data architecture phase. It is possible that some solutions, while optimal in principle, may not be potential candidates due to their cost. External factors such as the business cycle, interest rates, market conditions, and legal considerations could all have an effect on decisions relevant to data architecture.

Business policies

Business policies that also drive data architecture design include internal organizational policies, rules of regulatory bodies, professional standards, and applicable governmental laws that can vary by applicable agency. These policies and rules will help describe the manner in which enterprise wishes to process their data.

Data processing needs

These include accurate and reproducible transactions performed in high volumes, data warehousing for the support of management information systems (and potential data mining), repetitive periodic reporting, ad hoc reporting, and support of various organizational initiatives as required (i.e. annual budgets, new product development).

The General Approach is based on designing the Architecture at three Levels of Specification

- The Logical Level
- The Physical Level
- The Implementation Level

Few basic concepts in data architecture:

Conceptual / business data model—shows data entities such as customer, product and transaction, and their semantics.

Logical/system model—defines the data in as much detail as possible, including relations between data elements, but without considering how data is stored or managed.

Physical/technology data model—defines how the data is represented and stored, for example in a flat file, database, data warehouse, key-value store.

Who creates the data architecture?

The following roles exist to help shape and maintain a modern data architecture:

Data architect (sometimes called big data architects)—defines the data vision based on business requirements, translates it to technology requirements, and defines data standards and principles.

Project manager—oversees projects that modify data flows or create new data flows.

Solution architect—designs data systems to meet business requirements.

Cloud architect or data center engineer—prepares the infrastructure on which data systems will run, including storage solutions.

DBA or data engineer—builds data systems, populates them with data and takes care of data quality.

Data analyst—an end-user of the data architecture, uses it to create reports and manage an ongoing data feed for the business.

Data scientists—also a user of the data architecture, leveraging it to mine organizational data for fresh insights.

Elements typically found in modern data architecture:

Data warehouse: Data warehouses are still important, but are moving to the cloud and interacting with data lakes, traditional databases and unstructured data sources

Relational database: Oracle and SQL Server are still in use, but open source alternatives like MySQL and PostgreSQL are everywhere

NoSQL database: Stores massive amounts of semi-structured and unstructured data. Popular solutions are Redis, MongoDB, CouchDB, Memcached and Cassandra.

Real-time streaming: New tools such as Apache Kafka, Flume and AWS Kinesis help stream large volumes of data from system logs and production systems.

Containers: Platforms like Docker and Kubernetes help spin up and deploy data infrastructure at the click of a button, and orchestrate complex systems in a flexible and scalable manner.

Micro services and Server less computing: Data systems built using microservices or functions as a service (FaaS) are independent units that expose a standard interface, allowing data architects to compose and arrange data environments to suit business needs.

UNDERSTAND VARIOUS SOURCES OF DATA LIKE SENSORS/SIGNALS/GPS ETC

Data can be generated from two types of sources namely Primary and Secondary Sources of Primary Data

Sources of Primary Data

The sources of generating primary data are -

- Observation Method
- Survey Method
- Experimental Method
- Experimental Method

There are number of experimental designs that are used in carrying out and experiment. However, Market researchers have used 4 experimental designs most frequently.

- CRD - Completely Randomized Design

- RBD - Randomized Block Design
- LSD - Latin Square Design
- FD - Factorial Designs

CRD - Completely Randomized Design:

- Simplest design to use.
- Design can be used when experimental units are essentially homogeneous, because of the homogeneity requirement, it may be difficult to use this design for field experiments.
- The CRD is best suited for experiments with a small number of treatments.

Advantages

1. Very flexible design (i.e. number of treatments and replicates is only limited by the available number of experimental units).
2. Statistical analysis is simple compared to other designs.
3. Loss of information due to missing data is small compared to other designs due to the larger number of degrees of freedom for the error source of variation.

Disadvantages

1. If experimental units are not homogeneous and you fail to minimize this variation using blocking, there may be a loss of precision.
2. Usually the least efficient design unless experimental units are homogeneous.
3. Not suited for a large number of treatments.

RBD - Randomized Block Design:

The term Randomized Block Design has originated from agricultural research. In this design several treatments of variables are applied to different blocks of land to ascertain their effect on the yield of the crop. Blocks are formed in such a manner that each block contains as many plots as a number of treatments so that one plot from each is selected at random for each treatment. The production of each plot is measured after the treatment is given. These data are then interpreted and inferences are drawn by using the analysis of Variance

Technique so as to know the effect of various treatments like different dozes of fertilizers, different types of irrigation etc.

LSD - Latin Square Design:

A Latin square is one of the experimental designs which has a balanced two way classification scheme say for example - 4 X 4 arrangement. In this scheme each letter from A to D occurs only once in each row and also only once in each column. The balance arrangement, it may be noted that, will not get disturbed if any row gets changed with the other.

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

The balance arrangement achieved in a Latin Square is its main strength. In this design, the comparisons among treatments will be free from both differences between rows and columns. Thus the magnitude of error will be smaller than any other design.

FD - Factorial Designs:

This design allows the experimenter to test two or more variables simultaneously. It also measures interaction effects of the variables and analyses the impacts of each of the variables. In a true experiment, randomization is essential so that the experimenter can infer cause and effect without any bias

Sources of Secondary Data

While primary data can be collected through questionnaires, depth interview, focus group interviews, case studies, experimentation and observation; The secondary data can be obtained through

- Internal Sources - These are within the organization
- External Sources - These are outside the organization
- Internal Sources of Data

The Internal Sources Include

Accounting resources- This gives so much information which can be used by the marketing researcher. They give information about internal factors.

Sales Force Report- It gives information about the sale of a product. The information provided is of outside the organization.

Internal Experts- These are people who are heading the various departments. They can give an idea of how a particular thing is working

Miscellaneous Reports- These are what information you are getting from operational reports. If the data available within the organization are unsuitable or inadequate, the marketer should extend the search to external secondary data sources

External Sources of Data External Sources are sources which are outside the company in a larger environment. Collection of external data is more difficult because the data have much greater variety and the sources are much more numerous.

External data can be divided into following classes.

<u>Government Publications:</u>

Government sources provide an extremely rich pool of data for the researchers. In addition, many of these data are available free of cost on internet websites. There are number of government agencies generating data, these are

Registrar General of India It is an office which generates demographic data. It includes details of gender, age, occupation etc.

Central Statistical Organization This organization publishes the national accounts statistics. It contains estimates of national income for several years, growth rate, and rate of major economic activities. Annual survey of Industries is also published by the CSO. It gives information about the total number of workers employed, production units, material used and value added by the manufacturer.

Director General of Commercial Intelligence This office operates from Kolkata. It gives information about foreign trade i.e. import and export. These figures are provided region-wise and country-wise.

Ministry of Commerce and Industries This ministry through the office of economic advisor provides information on wholesale price index. These indices may be related to a number of sectors like food, fuel, power, food grains etc. It also generates All India Consumer Price Index numbers for industrial workers, urban, non manual employees and cultural labourers. **Planning Commission** It provides the basic statistics of Indian Economy.

Reserve Bank of India This provides information on Banking Savings and investment. RBI also prepares currency and finance reports.

Labour Bureau It provides information on skilled, unskilled, white collared jobs etc.

National Sample Survey This is done by the Ministry of Planning and it provides social, economic, demographic, industrial and agricultural statistics.

Department of Economic Affairs It conducts economic survey and it also generates information on income, consumption, expenditure, investment, savings and foreign trade.

State Statistical Abstract This gives information on various types of activities related to the state like - commercial activities, education, occupation etc.

Non Government Publications:

These includes publications of various industrial and trade associations, such as The Indian Cotton Mill Association, Various chambers of commerce, The Bombay Stock Exchange, Various Associations of Press Media, Export Promotion Council, Confederation of Indian Industries (CII) Small Industries Development Board of India Different Mills like - Woollen mills, Textile mills etc

The only disadvantage of the above sources is that the data may be biased. They are likely to colour their negative points.

Syndicate Services

These services are provided by certain organizations which collect and tabulate the marketing information on a regular basis for a number of clients who are the subscribers to these services. So the services are designed in such a way that the information suits the subscriber. These services are useful in television viewing, movement of consumer goods etc. These syndicate services provide information data from both household as well as institution.

In collecting data from household they use three approaches

- **Survey** They conduct surveys regarding - lifestyle, sociographic, general topics.
- **Mail Diary Panel** It may be related to 2 fields - Purchase and Media.
- **Electronic Scanner Services** These are used to generate data on volume.

They collect data for Institutions from

- Whole sellers Retailers, and Industrial Firms

The importance of Syndicate services are becoming popular since the constraints of decision making are changing and we need more of specific decision-making in the light of changing environment.

The disadvantage of syndicate services is information provided is not exclusive and a number of research agencies provide customized services which suits the requirement of each individual organization

International Organization

These includes

The International Labour Organization (ILO)

It publishes data on the total and active population, employment, Unemployment, wages and consumer prices

The Organization for Economic Co-operation and development (OECD)

It publishes data on foreign trade, industry, food, transport, and science and technology.

The International Monetary Fund (IMA)

It publishes reports on national and international foreign exchange regulations

DATA MANAGEMENT

Data management is an administrative process that includes acquiring, validating, storing, protecting, and processing required data to ensure the accessibility, reliability, and timeliness of the data for its users. Organizations and enterprises are making use of Big Data more than ever before to inform business decisions and gain deep insights into customer behavior, trends, and opportunities for creating extraordinary customer experiences.

Data management is the process of ingesting, storing, organizing and maintaining the data created and collected by an organization. Effective data management is a crucial piece of deploying the IT systems that run business applications and provide analytical information to help drive operational decision-making and strategic planning by corporate executives, business managers and other end users. The data management process includes a combination of different functions that collectively aim to make sure that the data in corporate systems is accurate, available and accessible.

To make sense of the vast quantities of data that enterprises are gathering, analysing, and storing today, companies turn to data management solutions and platforms. Data management solutions make processing, validation, and other essential functions simpler and less time-intensive.

Leading data management platforms allow enterprises to leverage Big Data from all data sources, in real-time, to allow for more effective engagement with customers, and for increased customer lifetime value (CLV). Data management software is essential, as we are creating and consuming data at unprecedented rates. Top data management platforms give enterprises and organizations a 360-degree view of their customers and the complete visibility needed to gain deep, critical insights into consumer behavior that give brands a competitive edge.

Importance of data management

Data increasingly is seen as a corporate asset that can be used to make more-informed business decisions, improve marketing campaigns, optimize business operations and reduce costs, all with the goal of increasing revenue and profits. But a lack of proper data management can saddle organizations with incompatible data silos, inconsistent data sets and data quality problems that limit their ability to run business intelligence (BI) and analytics applications -- or, worse, lead to faulty findings.

Data management has also grown in importance as businesses are subjected to an increasing number of regulatory compliance requirements, including data privacy and protection laws such as GDPR and the California Consumer Privacy Act. In addition, companies are capturing ever-larger volumes of data and a wider variety of data types, both hallmarks of the big data systems many have deployed. Without good data management, such environments can become unwieldy and hard to navigate.

Development of a data architecture is often the first step, particularly in large organizations with lots of data to manage. An architecture provides a blueprint for the databases and other data platforms that will be deployed, including specific technologies to fit individual applications.

Databases are the most common platform used to hold corporate data; they contain a collection of data that's organized so it can be accessed, updated and managed. They're used in both transaction processing systems that create operational data, such as customer records and sales orders, and data warehouses, which store consolidated data sets from business systems for BI and analytics

Database administration is a core data management function. Once databases have been set up, performance monitoring and tuning must be done to maintain acceptable response times on database queries that users run to get information from the data stored in them. Other administrative tasks include database design, configuration, installation and updates; data security; database backup and recovery; and application of software upgrades and security patches.

Better Ways to Manage Data

- **Focus on the information, not the device or data center** Focus on building an information infrastructure that optimizes the ability of your organization to find, access and consume critical business information. Key technologies include virtualization, cloud computing and mobile devices and applications.
- **Gain a complete understanding:** Know your information and recognize that not all information is equal. Many organizations lack basic knowledge like who owns specific information, how important the data is or even whether it is personal or business in nature. You need to map and classify information to discover its relative value. Once you've done this, you can more easily prioritize security, protection and management resources for the information that really matters.

- **Be efficient:** Use deduplication and archiving technologies to protect more while storing less. Only store what you really need.
- **Set consistent policies:** It's essential to set consistent policies for information that can be enforced wherever the information resides, whether in physical, virtual or cloud environments. This unifies information classification, automates discover of who owns and uses specific information, controls access and distribution, automates information retention and deletion and speeds the process of eDiscovery.
- **Stay agile:** Plan for future information needs by implementing a flexible infrastructure that supports continued growth

DATA QUALITY (NOISE, OUTLIERS, MISSING VALUES, DUPLICATE DATA):

Data have quality if they satisfy the requirements of the intended use. There are many factors comprising data quality, including accuracy, completeness, consistency, timeliness, believability, and interpretability.

There are many possible reasons for inaccurate data (i.e. having incorrect attribute values). The data collection instruments used may be faulty. There may have been human or computer errors occurring at data entry. This is known as disguised missing data.

Timeliness also affects data quality.

Two other factors affecting data quality are **believability** and **interpretability**. Believability reflects how much the data are trusted by users, while interpretability reflects how easy the data are understood.

Data cleaning routines work to “clean” the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.

Data reduction obtains a reduced representation of the data set that is much smaller in volume, yet produces the almost same analytical results. Data reduction strategies include dimensionality reduction and numerosity reduction.

Noisy Data

Noisy data is data with a large amount of additional meaningless information in it called noise. This includes data corruption and the term is often used as a synonym for corrupt data. It also includes any data that a user system cannot understand and interpreted correctly. Noise is a random error or variance in a measured variable.

The following are the techniques helps to remove noise.

Binning:

Binning methods smooth a sorted data value by consulting its “neighbourhood,” that is, the values around it. The sorted values are distributed into a number of “buckets,” or bins. Because binning methods consult the neighbourhood of values, they perform local smoothing

In smoothing by bin means, each value in a bin is replaced by the mean value of the bin.

In smoothing by bin medians can be employed, in which each bin value is replaced by the bin median.

In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value. In general the larger the width, the greater the effect of the smoothing.

For objects, noise is an extraneous object

For attributes, noise refers to modification of original values

Origins of noise

Outliers -- values seemingly out of the normal range of data

Duplicate records -- good database design should minimize this (use DISTINCT on SQL retrievals)

Incorrect attribute values -- again good db design and integrity constraints should minimize this

Numeric only, deal with rogue strings or characters where numbers should be.

How to locate and treat outliers (values seemingly out of the norm)

Null handling for attributes (nulls=missing values)

<u>Outliers</u>

An outlier is a value that escapes normality and can (and probably will) cause anomalies in the results obtained through algorithms and analytical systems. There, they always need some degrees of attention.

Understanding the outliers is critical in analyzing data for at least two aspects:

- The outliers may negatively bias the entire result of an analysis.
- The behavior of outliers may be precisely what is being sought.

Outliers can be classified into three categories, namely global outliers, contextual (or conditional) outliers, and collective outliers.

In a given data set, a data object is a global outlier if it deviates significantly from the rest of the data set. Global outliers are sometimes called point anomalies, and are the simplest type of outliers. Most outlier detection methods are aimed at finding global outliers.

In a given data set, a data object is a contextual outlier if it deviates significantly with respect to a specific context of the object. Contextual outliers are also known as conditional outliers because they are conditional on the selected context. Therefore, in contextual outlier detection, the context has to be specified as part of the problem definition. Generally, in contextual outlier detection, the attributes of the data objects in question are divided into two groups

Contextual attributes: The contextual attributes of a data object define the object's context. In the temperature example, the contextual attributes may be date and location.

Behavioural attributes: These define the object's characteristics, and are used to evaluate whether the object is an outlier in the context to which it belongs. In the temperature example, the behavioural attributes may be the temperature, humidity, and pressure.

In a data set, a subset of data objects forms a collective outlier if the objects as a whole deviate significantly from the entire data set. Importantly, the individual data objects may not be outliers.

Missing Values

It is very much usual to have missing values in your dataset. It may have happened during data collection, or maybe due to some data validation rule, but regardless missing values must be taken into consideration.

Eliminate rows with missing data:

Simple and sometimes effective strategy. Fails if many objects have missing values. If a feature has mostly missing values, then that feature itself can also be eliminated.

Estimate missing values:

If only a reasonable percentage of values are missing, then we can also run simple interpolation methods to fill in those values. However, most common method of dealing with missing values is by filling them in with the mean, median or mode value of the respective feature.
The following are the methods

Ignore the tuple: This is usually done when the class label is missing (assuming the mining task involves classification).This method is not very effective, unless the tuple containsseveralattributeswithmissingvalues.Itisespeciallypoorwhenthepercentage of missing values per attribute varies considerably. By ignoring the tuple, we do not make use of the remaining attributes' values in the tuple. Such data could have been useful to the task at hand.

Fill in the missing value manually

In general, this approach is time consuming and may not be feasible given a large data set with many missing values.

Use a global constant to fill in the missing value

Replace all missing attribute values by the same constant such as a label like "Unknown" or −∞. If missing values are replaced by, say, "Unknown," then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common—that of "Unknown." Hence, although this method is simple, it is not foolproof.

Use a measure of central tendency for the attribute (e.g., the mean or median)

To fill in the missing value Measures of central tendency, which indicate the "middle" value of a data distribution. For normal (symmetric) data distributions, the mean can be used, while skewed data distribution should employ the median. For example, suppose that the data distribution regarding the income of All Electronics customers is symmetric and that the mean income is $56,000. Use this value to replace the missing value for income.

Use the attribute mean or median for all samples belonging to the same class as the given tuple

For example, if classifying customers according to credit risk, we may replace the missing value with the mean income value for customers in the same credit risk category as that of the given tuple. If the data distribution for a given class is skewed, the median value is a better choice.

Use the most probable value to fill in the missing value

This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction. For example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for income.

Missing Data Handling

Many causes: malfunctioning equipment, changes in experimental design, collation of different data sources, measurement not possible. People may wish to not supply information. Information is not applicable (children don't have annual income)

- **Discard** records with missing values
- **Ordinal-continuous** data, could **replace with attribute means**
- **Substitute** with a value from a similar instance
- **Ignore** missing values, i.e., just proceed and let the tools deals with them
- **Treat** missing values **as equals** (all share the same missing value code)
- **Treat** missing values **as unequal values**

Missing completely at random (MCAR)

- Missingness of a value is independent of attributes
- Fill in values based on the attribute as suggested above (e.g. attribute mean)
- Analysis may be unbiased overall

Missing at Random (MAR)

- Missingness is related to other variables
- Fill in values based other values (e.g., from similar instances)
- Almost always produces a bias in the analysis

Missing Not at Random (MNAR)

- Missingness is related to unobserved measurements
- Informative or non-ignorable missing ness

Not possible to know the situation from the data. You need to know the context, application field, data collection process, etc.

Duplicate Data

A dataset may include data objects which are duplicates of one another. It may happen when say the same person submits a form more than once. The term deduplication is often used to refer to the process of dealing with duplicates.

In most cases, the duplicates are removed so as to not give that particular data object an advantage or bias, when running machine learning algorithms.

DATA PROCESSING

Data processing occurs when data is collected and translated into usable information. Usually performed by a data scientist or team of data scientists, it is important for data processing to be done correctly as not to negatively affect the end product, or data output.

Data processing starts with data in its raw form and converts it into a more readable format (graphs, documents, etc.), giving it the form and context necessary to be interpreted by computers and utilized by employees throughout an organization.

There are six stages in data processing

1. Data collection

Collecting data is the first step in data processing. Data is pulled from available sources, including data lakes and data warehouses. It is important that the data sources available are trustworthy and well-built so the data collected (and later used as information) is of the highest possible quality.

2. Data preparation

Once the data is collected, it then enters the data preparation stage. Data preparation, often referred to as “pre-processing” is the stage at which raw data is cleaned up and organized for the following stage of data processing. During preparation, raw data is diligently checked for any errors. The purpose of this step is to eliminate bad data (redundant, incomplete, or incorrect data) and begin to create high-quality data for the best business intelligence.

3. Data input

The clean data is then entered into its destination (perhaps a CRM like Salesforce or a data warehouse like Redshift), and translated into a language that it can understand. Data input is the first stage in which raw data begins to take the form of usable information.

4. Processing

During this stage, the data inputted to the computer in the previous stage is actually processed for interpretation. Processing is done using machine learning algorithms, though the process itself may vary slightly depending on the source of data being processed (data lakes, social networks, connected devices etc.) and its intended use (examining advertising patterns, medical diagnosis from connected devices, determining customer needs, etc.).

5. Data output/interpretation

The output/interpretation stage is the stage at which data is finally usable to non-data scientists. It is translated, readable, and often in the form of graphs, videos, images, plain text, etc.). Members of the company or institution can now begin to self-serve the data for their own data analytics projects.

6. Data storage

The final stage of data processing is storage. After all of the data is processed, it is then stored for future use. While some information may be put to use immediately, much of it will serve a purpose later on. Plus, properly stored data is a necessity for compliance with data protection legislation like GDPR. When data is properly stored, it can be quickly and easily accessed by members of the organization when needed.

The future of data processing lies in the cloud. Cloud technology builds on the convenience of current electronic data processing methods and accelerates its speed and effectiveness. Faster, higher-quality data means more data for each organization to utilize and more valuable insights to extract.

As big data migrates to the cloud, companies are realizing huge benefits. Big data cloud technologies allow for companies to combine all of their platforms into one easily-adaptable system. As software changes and updates, cloud technology seamlessly integrates the new with the old.

The benefits of cloud data processing are in no way limited to large corporations. In fact, small companies can reap major benefits of their own. Cloud platforms can be inexpensive and offer the flexibility to grow and expand capabilities as the company grows.

Big data is changing how all of us do business. Today, remaining agile and competitive depends on having a clear, effective data processing strategy. While the six steps of data processing won't change, the cloud has driven huge advances in technology that deliver the most advanced, cost-effective, and fastest data processing methods to date.