Unit 5

XML and Data Warehousing

XML : XML stands for extensible Markup language. XML was designed to store and transport data. XML is self-descriptive language: → It has sender, information & receiver
→ It has a heading and message body.

XML is just information wrapped in tags.

→ XML stores data in plain text format. This provides a s/w and h/w independent way of storing, transporting and sharing data.

Many computer systems contain data in incompatible formats. Exchanging data between incompatible systems (or upgraded systems) is a time-consuming task for web developers. large amount of data must be converted, and incompatible data is often lost.

XML also makes it easier to expand or upgrade to new operating systems, new app^n, or new browsers without losing data.

→ XML is used in many aspect of web developm.

→ XML does not carry any information about how to be displayed.

with XML, there is full separation b/w data and presentation.

The Extensible Markup language (XML) is a simple text-based format for representing structured information : documents, data, configuration, books, transaction, invoice, and much more.. It was derived from an older standard format called SGML, in Order to be more suitable for web use.

XML features :-

- Excellent for handling data with a complex structure or a typical data
- Data described using markup language
- Text data description
- Human and computer friendly format
- Handles data in a tree structure having one and only one root element
- Excellent for long-term data storage and data reusability.

- Design goals for XML are :

1. XML shall be straightforwardly usable over the internet.
2. XML shall support a wide variety of appⁿ.
3. XML shall be compatible with SGML.
4. It shall be easy to write programs which process XML documents.
5. The number of optional features in XML is to be kept to the absolute minimum, ideally zero.
6. XML document should be human-legible and reasonably clear.
7. The XML design should be prepared quickly.
8. The design of XML shall be formal and concise.
9. XML documents shall be easy to create.

XML	HTML
1. The main purpose is to focus on the transport of data and saving the data.	Focuses on the appearance of data. Enhances the appearance of texts
2. XML is dynamic because it is used in the transport of data.	HTML is static because its main functions is in the display of data.
3. It is case sensitive.	not case sensitive.
we can define tags as per our requirement but closing tags are mandatory.	It has its own pre-defined tags and it is not necessary to have closing tags.
XML can preserve white spaces.	white spaces are not preserves in HTML.
XML is content-driven and not many formatting features are available.	HTML is presentation driven. How the text appears is of utmost importance.
Any error in the code shall not give the final outcome.	Small errors in the coding can be ignored and the outcome can be achieved.
The size of the document may be large.	No lengthy documents. Only the syntax needs to be added for best-formatted output.

XML Syntax :-

```
<?xml version="1.0" encoding="UTF-8"?>
<note>
    <to> Tone </to>
    <from> Jani </from>
    <heading> Reminder </heading>
    <body> Don't forget me this weekend!</body
    </note>
```

→ XML declaration consist of XML version, character encoding or/and standalone status. declaration is optional.

- It has no closing tags.
- Comment are optional. [<!- and ends with ->
- Opening and closing tags [< > content </>]

element

eg: <age> 20 </age>

↳ age is the name of element. (Tag name also refered to as an element or element name)

→ All XML documents must contain a single root element.

→ A tag name can contain letters, digits, hyphens, underscores, and periods.

→ A tag name cannot contain spaces.

→ All elements must be nested properly.

XML Attributes:

Attribute for an element is placed after the tag name in the start tag. We can add more than one attribute for a single element with different attribute name.

```
<company name = "ABC Holdings" location = "London">
      <chairman> Mr. Ravi </chairman>
     <gm> Mr. John </gm>
</company>
```

- There are two attributes in the company element, i.e. name and location.
  name : attribute name
  ABC Holdings : attribute value.

→ company is the root element.

Attributes Rules:

- Attributes value must be within Qoutes.
- An element cannot contain several attributes with the same name.

html :-

```
<html >
<head >
      < title > Document </ title >
</head >
<body >
        <p> Book </p>
      <p> Name : Anna Karenina </p>
      <p> Authors : Leo Tolstoy </p>
      <p> Publisher :The Russian Messenger </p>
</body >
</html>
```

XML :-

```
<? xml version = "1.0" encoding = "UTF-8" ?>
<boot >
      <name > Anna Karenina </name >
      <author > leo Tolstoy </author >
      <Publisher >The Russian Messager </Publisher>
</book>
```

Data warehouse:

A data warehouse is a type of data mangement system that is designed to enable and support business intelligence (BI) activities, especially analytics.

Data warehouses are solely intended to perform queries and analysis and often contain large amounts of historical data.

The data within a data warehouse is usually derived from a wide range of sources such as applications log files and transaction application.

A data warehouse centralizes and consolidates large amount of data from multiple sources. Its analytical capablities allow organizations to derive valuable business insights from their data to improve decision-making.

Over time, it builds a historical data that can be invaluable to data scientists and business analysts.

Data mare includes following elements:

- A relational database to store and manage data.
- An extraction, loading and transformation (ELT) solution for preparing the data for analyses.
- Statistical analysis, reporting and data mining capabilities.
- Client analysis tool for visualizing and presenting data to business users.

Benefits of data warehouse:

- Subject oriented: They can analyze data about a particular subject or functional area (such as sales).
- Integrated: Data warehouse create consistency among different data types from disparate sources.
- Non volatile: Once data is in a data warehouse, it's stable and doesn't change.
- Time - variant: Data warehouse analysis looks at change over time.

A well designed data mare will perform queries very quickly, deliver high data throughput and provide enough flexibility for end users to slice and dice or reduce the volume of data for closer examination to meet a variety of demands - whether at a high level or at a very fine, detailed level.

<u>XHTML</u> : extensible Hyper Text Markeup language

XHTML was developed by world wide web Consortium (W3C). It helps web developers to make the transition from HTML to XML.

XHTML documents contains three parts :

- DOCTYPE : It is used to declare a DTD
- head: The head section is used to declare the title and other attributes.
- body: The body tag contains the content of web page. It consist many tags,

→ Creating a XHTML web page, it is necessary to include DTD (Document Type definition) declaration.

Three types of DTD :-

1. <u>Transitional DTD</u> : It is supported by the older browsers which does not have inbuilt cascading style sheets supports. There are several attributes enclosing the body tag which are not allowed in strict DTD.

2. <u>Strict DTD</u> : Strict DTD is used when XHTML page contains only markeup language. Strict DTD is used together with cascading style

sheets, because this attribute does not allow CSS property in body tag.

3. Frameset DTD: The frameset DTD is used when XHTML page contains frames.

Why use XHTML:

- XHTML documents are validated with standard XML tools.
- It is used to define the quality standard of web pages.
- XHTML is official standard, website becomes more compatible and accurate with many browsers.

Benefits:

- XHTML documents are lean which means they use less band width. This reduces cost Particularly if web site has 1000 of Pages.
- XHTML works in association with CSS to create web pages that can easily be updated.

Data Mart :

A data mart is a scaled-down version of a data warehouse aimed at meeting the information needs of a homogenous small group of end users such as a department or business unit (marketing, finance, logistics or human resources).

It typically contains some form of aggregated data and is used as the primary source for report generation and analysis by this end user group.

~~There are various reasons~~ of In data mart, we can decide, who can access the data set by design and build database tables.

→ Data marts can be divided into two types:

- The first one is independent data mart, the ETL architecture, and the source of the database belong to a single entity.
- The second one is dependent data mart, in this type of data mart the incoming data arrived from other sources mainly from the data warehouse.

→ Data mart improve team efficiency, reduce costs and facilitate smarter tactical business decision making in enterprises.

There are various reasons for setting up datama

- They provide focused content, such as financ Sales or accounting information in a format tailo to user groups at hand.
- They improve query ~~per~~ performance by offload complex queries.
- Data marts can be located closer to the end user alleviating heavy network traffic and giving t more control.

Operational data stores :-

An ODS can be considered a staging area that Provides querying facilities. A normal staging area is only meant for receiving the operational data from the transactional sources for the sake of transforming the data and loading it into the data warehouse.

Analysis tools that need data that is closer to real-time can query the ODS data as it is received from the respective source systems, before time-consuming transformation and loading operations.

The ODS then only provides access to the current, fine grained and non-aggregated data, which can be queried in an integrated manner without burdening the transactional systems.

- ODS ~~cannot~~ do not host large amounts of historical data, and thus cannot handle huge data transactions.
- ODS systems host configurable, easily accessible and fast real-time comprehensive data.
- ODS occupy less space due to the compression of data and operations.
- An ODS system makes the creation of back-ups and recovery processes effortlessly since the size

of the data is small.

- ODS is also known as On-Line Transfer Processing (OLTP) Database Management System where data is stored and processed in real-time.
- An operational Data Store contains atomic or indivisible data, such as prices and transactions that are captured in real-time, and they has a limited history.

☒ The main purpose of an ODS is to integrate data from diverse source systems into a single entity, through technologies such as Extract, Transform and Load (ETL), Data federation or Data Virtualization.

→ In creation of an ~~OD~~ ODS, multiple data sources can be integrated. Each data source system must have to the following principles to qualify:

- Subject-Oriented: The design of the ODS should be built based on the business's functional requirements, especially in regard to a specific area.
- Integrated: All the data from diverse sources must undergo the ETL process, which involves cleaning junk for redundancy, data transformation into a single format, and loading of the dataset into the ODS as indicated by the business policies for control and regularity of data.
- Up-to Date: ODS data should be current and updated
- Detailed: As the rules are implemented, it is crucial to maintain the business's comprehensive detailing level for the proper execution of respective functions.

Identifying web presence Goals:

- Business always create a presence in the physical world by building stores and office buildings.
- On the web, business have the luxury of intentionally creating a space that creates a distinctive presence.
- A website can perform many image-creation tasks very effectively, including:
  - Serving as a sales brouchere
  - Serving as a product showroom
  - Showing a financial report
  - posting an employment ad
  - Serving as a customer contact point

- web Presence should include:
  - History
  - Mission statement
  - Financial and product information
  - Method of contacting the organization

Meeting the needs of web site Visitors;

why visitors come to web sites :

- To learn about or buy a company's products or services
- Get product support for products already bought
- obtain financial or general product information about a company.
- communicate with the company or identify who manage it.

Achieving web presence goals:

Goals associated with effective web sites include;

- Attracting visitors
- Making the site interesting to explore.
- Creating a Positive image consistent with the company's desloas.
- Building trusting relationship with visitors
- Encourage to visitors to return

Site Adhesion:

- Content
- format
- Access

Maintaining a website:

- Convey an integrated image of the organisation

- Offers eastly accessible facts about the Organisation
- Allow visitors to experience the site in different ways and at different levels.
- Sustain visitor attention and encourage return visits.
- Offer easily accessible information about products and services and how to use them.

Metrics to measure web Service Performance:

- Metrices approximate the end-to-end response time observed by the client for web page download. Additionally, we provide a means to calculate the breakdown between server processing and networking portions of overall response time.
- metrices evaluating the caching efficiency for a given web page by computing the server file hit ratio and server byte hit ratio for the web page.
- metrices relating the end-to-end performance of aborted web pages to the QoS.