**DATA ENGINEERING**

S.R.TAANUSRI

B.E -CSE

**DEFINITION**:

To design, scale, build system to organize data for analysis purpose.

**TOOL USED:**

Etl. ETL stands for:

E-Extract

T-Transform

L-Load

**BLUE PRINT:**

**VISUALISE**

* Web ul
* Business Intelligence tools
* Mobile applications

**STORE**

* Big data
* SQL DB

**PROCESSING FRAMEWORK**

* Batch
* Stream

**BUFFER**

* Caches
* Message Queues

**CONNECT**

* API
* Data flow apps

**EXTERNAL**

* Enterprise data warehouse
* SQL Database

Free Transparent Arrow Images, Download Free Transparent ...

**DATA CLASSIFICATION**

1. Raw data

* It is taken directly from the data source
* There is no schema applied

Eg: Rice directly collected by farmers from field.

1. Processed Data

* The schema is applied to raw data.
* It is stored in the destinations.

Eg: Rice being washed in water, soaked in water and then cooking it.

c)Cooked data

* Data being arranged in proper way

Eg: Rice completely cooked ready to eat

**BIG DATA**

To store large amount of data, the data engineering comes into picture.

Properties of big data

1. Velocity

The speed at which data is getting loaded

1. Volume

The amount of data that can be stored

1. Variety

The different types of the data that can be accessed

1. Veracity

The reliability of the data

DATA PROCEESING METHODS

1)Batch Processing

Data can be stored and then analyzed to view results through insight.

2)Stream Processing

Processing data as the data enters the method

MAP REDUCE

* It is done based on the key value pairing
* The big data runs on this concept.
* The data is organized based on key and value
* The values are matched based on the key values.
* The process is done until the final key value pair is obtained.

Eg:

1. Spark
2. Hadoop
3. Samza

DATA STORAGE

1. Relational database(SQL)
2. Document Store(No SQL)

**DATA WAREHOUSING**

**DEFINITION**:

* The subject oriented, integrated, time variant and non volatile collection of data in to support to the management’s system.
* It helps the management for decision making and easy to picturize the business conditions.

**FEATURES**:

1. Subject oriented

* It is modelling and analyzing the data.
* Data is organized based on the subject and not the content.

1. Integrated

* It is modelling and analyzing the data.
* Data is organized based on the subject and not the content.

1. Time variant

* Time period is longer for the data stored in the warehouse

1. Non volatile

* Updates cannot be made.
* The data entered cannot be changed or removed.

**DECISION SUPPORT SYSTEM(DSS)**

**NEEDS:**

* Decision making is a time consuming process but according to business needs decisions should be made faster in a organization
* So the company needs an information system for quick decision making.

**STRUCTURED AND UNSTRUCTURED COMPONENTS:**

|  |  |
| --- | --- |
| **STRUCTURED COMPONENTS** | **UNSTRUCTURED COMPONENTS** |
| * DSS helps directly for decision making. | * DSS needs to be processed and needs some human instructions. |
| * By choosing share market scenario:   i)investment risk  ii)performance of the strategy | * By choosing share market scenario:   i)The list of share that needs to be purchased |

**ARCHITECTURE**

**1.OLTP(Online Transaction Processing):**

* Users can easily access large amounts of data.
* It runs based on the transactions made the amount being deducted and credited.
* It easily helps for the data to be retrieved and entered during transaction processing.

Eg: ATM(Automatic Teller Machine)

**BENEFITS:**

* Simplicity and efficiency
* Fast query processing
* Data Integrity

**PITFALLS:**

* Need to update frequently
* Not suitable for data analysis
* Need to query multiple times

**2.OLAP(Online analytical processing)**

**DATA WAREHOUSE**

* It is a database used for analyzing and reporting the data.
* Data are uploaded from operational systems.
* Data store contains two types of data

a)Business data

b)Business data model

* Business data uses operational and external data that is related with business current trends.

**OPERATIONAL DATA**

* To carry out regular operation of organization.
* Use oltp architecture
* Access data by inserting, updating and deleting data.

**DSS VS OPERATIONAL DATA**

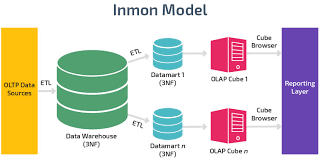
* Time span
* Granularity span
* Dimensionality

Data warehouse using a typic ETL uses:

1. Staging
2. Integration
3. Accessing through data marts

**DATA MARTS**

* Data is stored in a warehouse in form of data mart
* User can access the data in group, from the large amount of data available in data warehouse.

****

**DATA MARTS VS DATA WAREHOUSE**

* Building data marts are easier than data warehouse
* Data marts are flexible and data can be retrieved easily compared to data warehouse.

**DATA WAREHOUSE LIFECYLE**

The four main components are:

* Data warehouse
* Data marts
* Data staging and ETL
* Data source