DATA WAREHOUSE USING IBM CLOUD
DB2 WAREHOUSE

TEAM MEMBER

730321104306:R.SURYA PRAKASH

PHASE:2 SUBMISSION DOCUMENT

PROJECT:DATA WAREHOUSE USING IBM CLOUD DB2 WAREHOUSE

INTRODUCTION:

A Data Warehouse is Built by combining data from multiple diverse sources that support analytical reporting, structured and unstructured queries, and decision making for the organization, and Data Warehousing is a step-by-step approach for constructing and using a Data Warehouse. Many data scientists get their data in raw formats from various sources of data and information. But, for many data scientists also as business decision-makers, particularly in big enterprises, the main sources of data and information are corporate data warehouses. A data warehouse holds data from multiple sources, including internal databases and Software (SaaS) platforms. After the data is loaded, it often cleansed,

transformed, and checked for quality before it is used for analytics reporting, data science, machine learning, or anything.

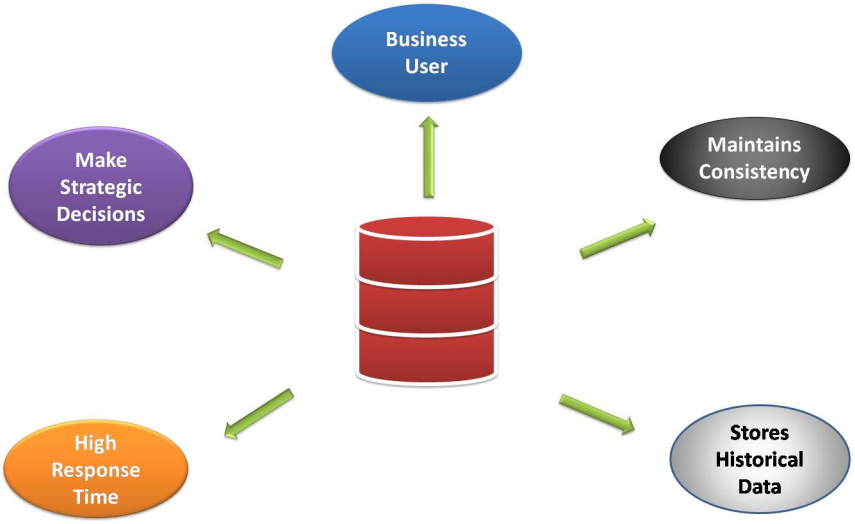
- **Business User:** Business users or customers need a data warehouse to look at summarized data from the past.

  Since these people are coming from a non-technical background also, the data may be represented to them in an uncomplicated way.

- **Maintains consistency**: Data warehouses are programmed in such a way that they can be applied in a regular format

  to all collected data from different sources, which makes it effortless for company decision-makers to analyze and share data insights with their

  colleagues around the globe. By standardizing the data, the risk of error in interpretation is also reduced and improves overall accuracy.



- **tore historical data**: Data

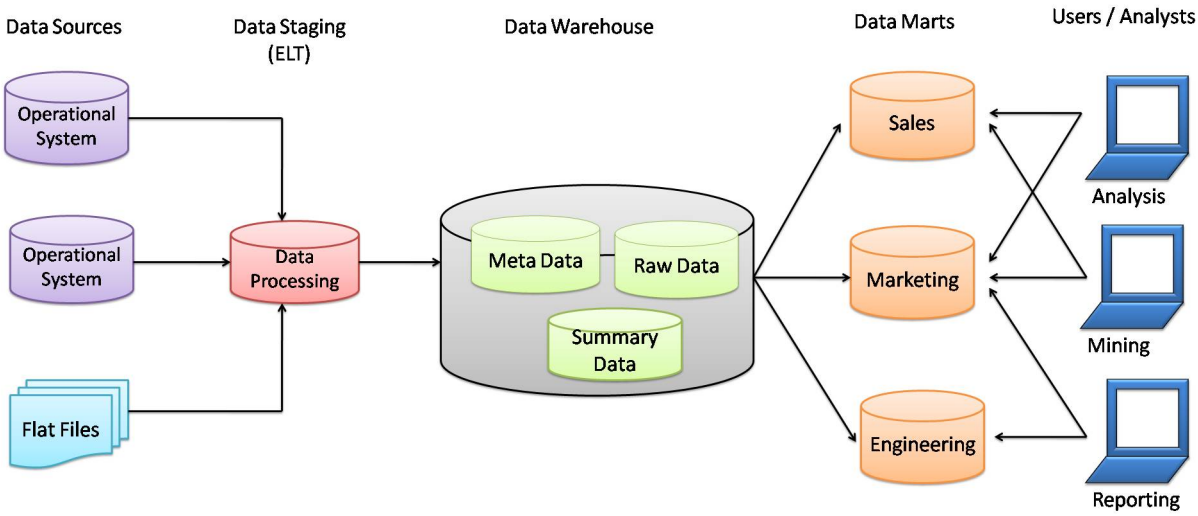
  Warehouses are also used to store historical data that means, variable data from the past and this input can be used for various purposes.

- **Make strategic decisions**: Data warehouses contribute to making better strategic decisions. Some business

  strategies may be depending upon the data stored within the data warehouses.

- **High response time**: Data warehouse has got to be prepared for somewhat sudden masses and type of queries that

  demands a major degree of flexibility and fast latency.



DATA WAREHOUSE ARCHITECTURE

1.Source Data Component:

- In the Data Warehouse, the source data comes from different places. They are group into four categories:

- ★ External Data: For data gathering, most of the executives and data analysts rely on information coming from external sources for a numerous amount of the information they use. They use statistical features associated with their organization that is brought out by some external sources and department.
- ★ Internal Data: In every organization, the consumer keeps their “private” spreadsheets, reports, client profiles, and generally even department databases. This is often the interior information, a part that might be helpful in every data warehouse.
- ★ Operational System data: Operational systems are principally meant to run the business. In each operation system, we periodically take the old data and store it in achieved files.
- ★ Flat files: A flat file is nothing but a text database that stores data in a plain text format. Flat files generally are text files that have all data processing and structure markup removed. A flat file contains a table with a single record per line.

2. Data Staging:

After the data is extracted from various sources, now it’s time to prepare the data files for storing in the data warehouse. The extracted data collected from various sources must be transformed and made ready in a format that is suitable to be saved in the data warehouse for querying and analysis. The data staging contains three primary functions that take place in this part:

- **Data Extraction**: This stage handles various data sources. Data analysts should employ suitable techniques for every data source.
- **Data Transformation:** As we all know, information for a knowledge warehouse comes from many alternative sources. If information extraction for a data warehouse posture huge challenges, information transformation gifts even important challenges. We tend to perform many individual tasks as a part of information transformation. First, we tend to clean the info extracted from every source of data. Standardization of information elements forms an outsized part of data transformation. Data transformation contains several kinds of combining items of information from totally different sources. Information transformation additionally contains purging supply information that's not helpful and separating outsourced records into new mixtures. Once the data transformation performs ends, we've got a set of integrated information that's clean, standardized, and summarized.
- **Data Loading**: When we complete the structure and construction of the data warehouse and go live for the first time, we do the initial loading of the data into the data warehouse storage. The initial load moves high volumes of data consuming a considerable amount of time.

3. Data Storage in Warehouse:

Data storage for data warehousing is split into multiple repositories. These data repositories contain structured data in a very highly

normalized form for fast and efficient processing.

- Metadata: Metadata means data about data i.e. it summarizes basic details regarding data, creating findings & operating with explicit instances of data. Metadata is generated by an additional correction or automatically and can contain basic information about data.
- Raw Data: Raw data is a set of data and information that has not yet been processed and was delivered from a particular data entity to the data supplier and hasn't been processed nonetheless by machine or human. This data is gathered out from online sources to deliver deep insight into users' online behavior.
- Summary Data or Data summary: Data summary is an easy term for a brief conclusion of an enormous theory or a paragraph. This is often one thing where analysts write the code and in the end, they declare the ultimate end in the form of summarizing data. Data summary is the most essential thing in data mining and processing.

4. Data Marts:

Data marts are also the part of storage component in a data warehouse. It can store the information of a specific function of an organization that is handled by a

single authority. There may be any number of data marts in a particular organization depending upon the functions. In short, data marts contain subsets of the data stored in data warehouse