CS –32 Data Warehousing with SQL Server 2012

Unit -1: Introduction to Data Warehousing

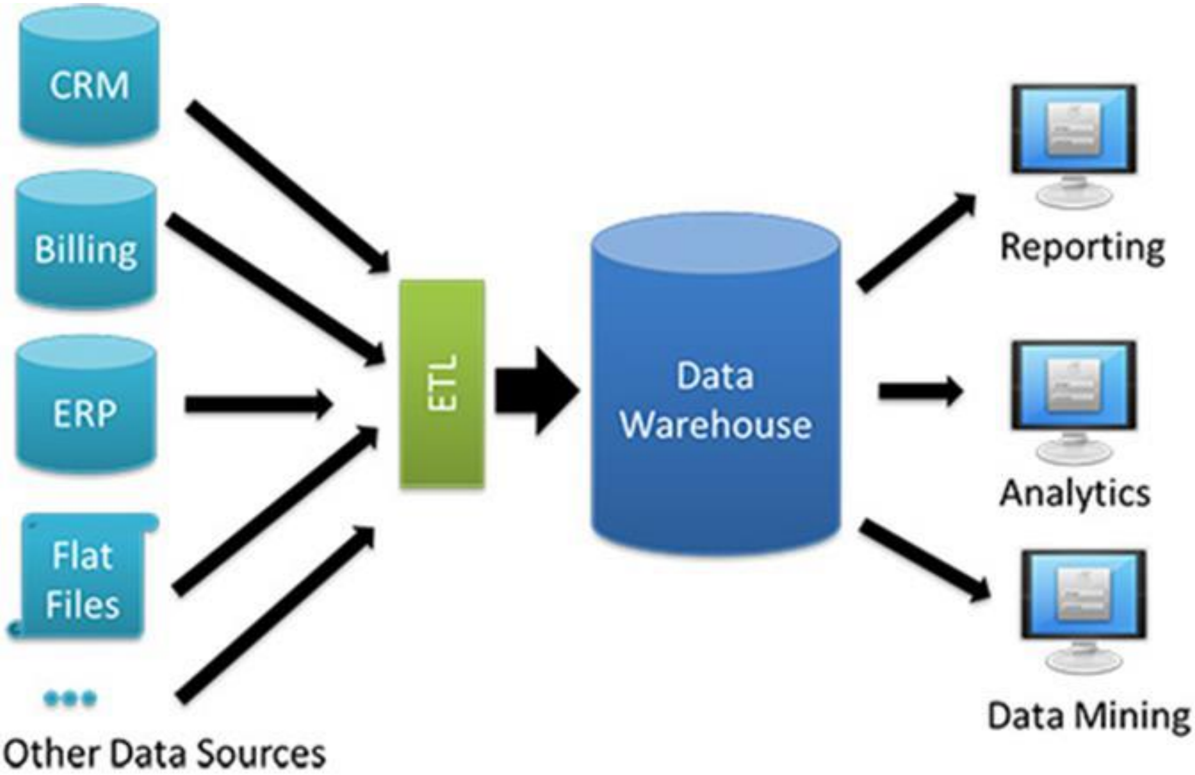
What Is a Data Warehouse?

In computing, a data warehouse, also known as an enterprise data warehouse, is a system used for reporting and data analysis, and is considered a core component of business intelligence. DWs are central repositories of integrated data from one or more disparate sources.

Data Warehousing Today

Top 5 data warehouses on the market today

In this day of rapid scale growth in Big Data, predictive analytics, and real time processing platforms like Hadoop, a fair question may arise what value is the traditional data warehouse? It's a fair question because before the iPhone, Facebook, Twitter, and Xbox, there was well the data warehouse. For the last 30 odd years the data warehouse has been, what one articles describes, as "the business-insights workhorse of enterprise computing." And despite many transformations over the past 5 years in the area of cloud, mobile, and information technologies, data warehousing has stayed relevant. Yes, there are more options on the table today for data storage, analysis, and indexing, but data warehouses have remained as timely as ever.



To be sure we're clear on definitions, a data warehouse is "a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other sources. It separates analysis workload from transaction workload and enables an organization to consolidate data from several sources."

Oracle, a well-known player in the market, last year identified the top 10 trends in data warehousing, including such things as real-time analytics, better customer experience capabilities, in-memory technologies, and more. In the words of one analysis, the data warehousing landscape is comprised of “a new generation of data warehouses that are bigger, better, and faster than ever before, transforming data into information and information into actionable insights, enabling businesses to forge ahead with unprecedented speed and agility.”

So, with these points in mind let’s review in more detail the state of the date warehouse market by surveying the top 5 vendors. Here’s a review of the major players you’ll want to pay attention to if you’re looking to get started in or upgrade to a data warehouse in 2015.

1. Teradata

Teradata is a market leader in the data warehousing space that brings more than 30 years of history to the table. It appears as the leader in Gartner’s 2014 Magic Quadrant for Data Warehouse Database Management Systems and has been so consistently for the past 15 years. The company is leading the charge with new tools, innovations, and capabilities, including all the latest in Hadoop-based technologies. Teradata’s EDW (enterprise data warehouse) platform provides businesses with robust, scalable hybrid-storage capabilities and analytics from mounds of unstructured and structured data leading to real-time business intelligence insights, trends, and opportunities. Teradata also offers a cloud-based DBMS solution via its Aster Database platform. Gartner reports that Teradata counts more than 1200 customers.

2. Oracle

Oracle is basically the household name in relational databases and data warehousing and has been so for decades. Oracle 12c Database is the industry standard for high performance scalable, optimized data warehousing. The company’s specialized platform for the data warehousing side is the Oracle Exadata Machine. There are an estimated 390,000 Oracle DBMS customers worldwide, and Gartner estimates about 4,000 Exadata data warehousing appliances have been sold. This state-of-the-art platform provides such advanced features as Flash Storage for low I/O overhead and Hybrid Columnar Compression (HCC), which enables high level compression of data for reduced I/O especially for analytics.

3. Amazon Web Services (AWS)

The whole shift in data storage and warehousing to the cloud over the last several years has been momentous and Amazon has been a market leader in that whole paradigm. Amazon offers a whole ecosystem of data storage tools and resources that complement its cloud services platform. For example, there is Amazon Redshift, a fast, fully managed, petabyte-scale data warehouse cloud solution; AWS Data Pipeline, a web service designed for transporting data between existing AWS data services; and Elastic MapReduce, which provides an easily managed Hadoop solution on top of the AWS services platform. According to Gartner, Amazon was the overall leader in data warehousing customer satisfaction and experience in last year's survey.

4. Cloudera

Cloudera has emerged in recent years as a major enterprise provider of Hadoop-based data storage and processing solutions. Cloudera offers an Enterprise Data Hub (EDH) for its variety of operational data store, or data warehouse. The EDH is Cloudera's proprietary framework for the "information-driven enterprise" and focuses on "batch processing, interactive SQL, enterprise search, and advanced analytics—together with the robust security, governance, data protection, and management that enterprises require." Cloudera's data warehouse is based on CDH, which is Cloudera's version of Apache Hadoop and the world's largest distribution at that. The organization offers a number of different bundles of its Hadoop-based services, including Cloudera Express and Cloudera Enterprise. Gartner reports high customer satisfaction and confidence in Cloudera's personnel and their skills in deploying Hadoop as a data processing and management system.

5. MarkLogic

MarkLogic is a Silicon Valley-based private software firm founded in 2001 that offers an enterprise NoSQL database platform. MarkLogic was included in Gartner's Magic Quadrant on Data Warehouse Database Management Systems for the first time in 2014. This inclusion also reflects a broader shift in the data warehousing market as organizations are seeing NoSQL and other alternative forms of storage and processing as the new reality for architecting their datacentre infrastructures and minimizing data complexity. In 2013 MarkLogic released a new semantics platform which provides the capability of storing billions of RDF triples that can queried with SPARQL (a semantic query language for the RDF platform) to provide richer, deeper insights to data in ways not possible within relational models. The inclusion of semantics-based technologies, along with what we've seen already of cloud and Hadoop, represents yet another level of innovation that will continue to keep data warehouses scalable and adaptable to the fast-paced needs of the digital era.

Future Trends in Data Warehousing.

Following are the future aspects of data warehousing.

- As we have seen that the size of the open database has grown approximately double its magnitude in the last few years, it shows the significant value that it contains.
- As the size of the databases grow, the estimates of what constitutes a very large database continues to grow.

- The hardware and software that are available today do not allow to keep a large amount of data online. For example, a Telco call record requires 10TB of data to be kept online, which is just a size of one month's record. If it requires to keep records of sales, marketing customer, employees, etc., then the size will be more than 100 TB.
- The record contains textual information and some multimedia data. Multimedia data cannot be easily manipulated as text data. Searching the multimedia data is not an easy task, whereas textual information can be retrieved by the relational software available today.
- Apart from size planning, it is complex to build and run data warehouse systems that are ever increasing in size. As the number of users increases, the size of the data warehouse also increases. These users will also require to access the system.
- With the growth of the Internet, there is a requirement of users to access data online.

Hence the future shape of data warehouse will be very different from what is being created today.

Data Warehouse Architecture

Data Warehouse Architecture is complex as it's an information system that contains historical and commutative data from multiple sources. There are 3 approaches for constructing Data Warehouse layers: Single Tier, Two tier and Three tier. This 3-tier architecture of Data Warehouse is explained as below.

Single-tier architecture

The objective of a single layer is to minimize the amount of data stored. This goal is to remove data redundancy. This architecture is not frequently used in practice.

Two-tier architecture

Two-layer architecture is one of the Data Warehouse layers which separates physically available sources and data warehouse. This architecture is not expandable and also not supporting a large number of end-users. It also has connectivity problems because of network limitations.

Three-Tier Data Warehouse Architecture

This is the most widely used Architecture of Data Warehouse.
It consists of the Top, Middle and Bottom Tier.

1. **Bottom Tier:** The database of the Datawarehouse servers as the bottom tier. It is usually a relational database system. Data is cleansed, transformed, and loaded into this layer using back-end tools.
2. **Middle Tier:** The middle tier in Data warehouse is an OLAP server which is implemented using either ROLAP or MOLAP model. For a user, this application tier presents an abstracted view of the database. This layer also acts as a mediator between the end-user and the database.
3. **Top-Tier:** The top tier is a front-end client layer. Top tier is the tools and API that you connect and get data out from the data warehouse. It could be Query tools, reporting tools, managed query tools, Analysis tools and Data mining tools.

Data Flow Architecture

- Data Flow Architecture is transformed input data by a series of computational or manipulative components into output data.
- It is a computer architecture which do not have a program counter and therefore the execution is unpredictable which means behaviour is indeterministic.
- Data flow architecture is a part of Von-Neumann model of computation which consists of a single program counter, sequential execution and control flow which determines fetch, execution, commit order.

- This architecture has been successfully implemented.
- Data flow architecture reduces development time and can move easily between design and implementation.
- It has main objective is to achieve the qualities of reuse and modifiability.
- In data flow architecture, the data can be flow in the graph topology with cycles or in a linear structure without cycles.

There are three types of execution sequences between modules:

1. Batch Sequential
2. Pipe and Filter
3. Process Control

1. Batch Sequential

- Batch sequential compilation was regarded as a sequential process in 1970.
- In Batch sequential, separate programs are executed in order and the data is passed as an aggregate from one program to the next.
- It is a classical data processing model.

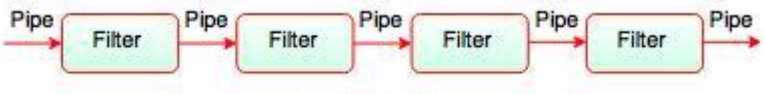


Fig. Batch Sequential

The above diagram shows the flow of batch sequential architecture. It provides simpler divisions on subsystems and each subsystem can be an independent program working on input data and produces output data.

- The main disadvantage of batch sequential architecture is that, it does not provide concurrency and interactive interface. It provides high latency and low throughput.

2. Pipe and Filter

What is meant by Pipe?

- Pipe is a connector which passes the data from one filter to the next.
- Pipe is a directional stream of data implemented by a data buffer to store all data, until the next filter has time to process it.
- It transfers the data from one data source to one data sink.
- Pipes are the stateless data stream.

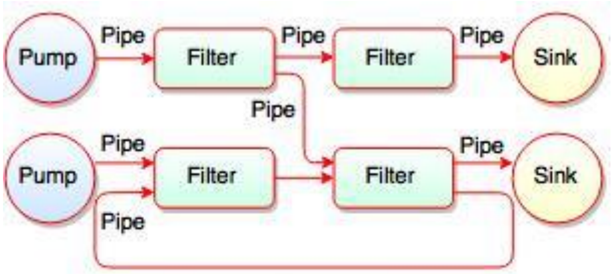


Fig. Pipes and Filters

The above figure shows the pipe-filter sequence. All filters are the processes that run at the same time, it means that they can run as different threads, coroutines or be located on different machines entirely.

Each pipe is connected to a filter and has its own role in the function of the filter. The filters are robust where pipes can be added and removed at runtime.

Filter reads the data from its input pipes and performs its function on this data and places the result on all output pipes. If there is insufficient data in the input pipes, the filter simply waits.

What are the Filters?

- Filter is a component.
- It has interfaces from which a set of inputs can flow in and a set of outputs can flow out.
- It transforms and refines input data.
- Filters are the independent entities.

There are two strategies to construct a filter:

1. Active Filter
2. Passive Filter

1. Active filter derives the data flow on the pipes.
2. Passive filter is driven by the data flow on the pipes.
Filter does not share state with other filters.

- They don't know the identity to upstream and downstream filters.
- Filters are implemented by separate threads. These may be either hardware or software threads or coroutines.

Advantages of Pipes and Filters

- Pipe-filter provides concurrency and high throughput for excessive data processing.
- It simplifies the system maintenance and provides reusability.
- It has low coupling between filters and flexibility by supporting both sequential and parallel execution.

Disadvantages of Pipe and Filter

- Pipe and Filter are not suitable for dynamic interactions.
- It needs low common denominator for transmission of data in ASCII format.
- It is difficult to configure Pipe-filter architecture dynamically.

3. Process Control

- Process Control Architecture is a type of Data Flow Architecture, where data is neither batch sequential nor pipe stream.
- In process control architecture, the flow of data comes from a set of variables which controls the execution of process.
- This architecture decomposes the entire system into subsystems or modules and connects them.
- Process control architecture is suitable in the embedded system software design, where the system is manipulated by process control variable data and in the Real time system software, process control architecture is used to control automobile anti-lock brakes, nuclear power plants etc.
- This architecture is applicable for car-cruise control and building temperature control system.