Teradata Basics

Lesson 01: An Overview of Teradata

Capgemini

Module Object

- RDBMS Concepts
- Teradata Overview
- Teradata and Data warehouse
- Components and Architecture
- Teradata Utiltities

RDBMS Concepts What is a Relational Database?

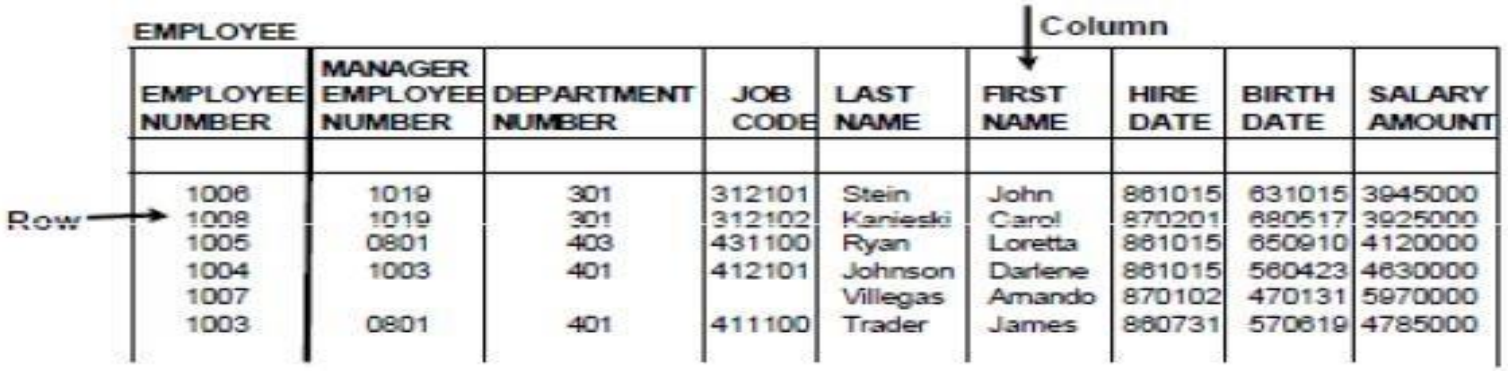
- What is a Relation Database ?
- A database is a collection of permanently stored data that is:

- Logically related → data relates to other data
- Shared → many users may access data.
- Protected → access to data is controlled.
- ☐Managed → data integrity and value is maintained.

- Relational databases are based on the relational model, which is founded on mathematical Set Theory.

RDBMS Concepts Relational Database

- Relational Databases present data as of a set of logically related tables.
- A table is a 2 dimensional representation of data that consists of rows & columns.



EMPLOYEE NUMBER	MANAGER EMPLOYEE NUMBER	DEPARTMENT NUMBER	JOB CODE	LAST NAME	FIRST NAME	HIRE DATE	BIRTH DATE	SALARY AMOUNT
1006	1019	301	312101	Stein	John	861015	631015	3945000
1008	1019	301	312102	Kanieski	Carol	870201	680517	3925000
1005	0801	403	431100	Ryan	Loretta	861015	650910	4120000
1004	1003	401	412101	Johnson	Darlene	861015	560423	4630000
1007				Villegas	Amando	870102	470131	5970000
1003	0801	401	411100	Trader	James	860731	570619	4785000

- **The employee table has:**
  - **Nine columns of data**
  - **Six rows of data—one per employee**
  - **Only one row format for the entire table**
  - **Missing data values represented by nulls**

RDBMS Concepts → Data Modeling

- 3NF (Third Normal Form):-
  - A normalized model includes:-
    - Entities
    - Attributes
    - Relationships
- Rules of Normalization:-
  - 1st NF → Each & every attribute within an entity has 1 & only 1 value.
    No repeating groups are allowed within entities.
  - 2nd NF→ Entity must conform to the 1st NF rules.
  - Every non-key attribute within an entity is fully dependent upon the entire
  - key of the entity & not a subset of the key
  - 3rd NF→ Entity must conform to the 1st NF & 2nd NF rules.

No non-key attributes within an entity is functionally dependent upon another non-key attributewithin the same entity. Star Schema & Snowflake Schema:

- A Star schema model includes:-
  - Facts & Dimensions

RDBMS Concepts → Data Modeling

- Normalization:
  - Process of reducing a complex data structure into a simple,
  - stable model.
  - Involves removing redundant attributes, keys, and relationships from the data model.

- Star Schema:
  - Process of having fewer entities
  - Involves a greater level of denormalization

Example on Normalization

Table: College

StudentName	CourseID1	CourseTitle1	CourseProfessor1	CourseID2	CourseTitle2	CourseProfessor2	StudentAdvisor	StudentID
Tia Carrera	CS123	Perl Regular Expressions	Don Corleone	CS003	Object Oriented Programming 1	Daffy Duck	Fred Flintstone	400
John Wayne	CS456	Socket Programming	DJ Tiesto	CS004	Algorithms	Homer Simpson	Barney Rubble	401
Lara Croft	CS789	OpenGL	Bill Clinton	CS001	Data Structures	Papa Smurf	Seven of Nine	402

1st Normal Form: (Each Column Type is Unique and there are no repeating groups [types] of data)

Table Name: Student Information
StudentID (Primary Key)
StudentName
AdvisorName

Table Name: Course Information
CourseID (Primary Key)
CourseTitle
CourseDescription
CourseProfessor

Table Name: Students and Courses
SnCStudentID
SnCCourseID

2nd Normal Form: (All attributes within the entity should depend solely on the entity's unique identifier)

Table Name: Student Information
StudentID (Primary Key)
StudentName

Table Name: Advisor Information
AdvisorID
AdvisorName

Table Name: Course Information
CourseID (Primary Key)
CourseTitle
CourseDescription
CourseProfessor

Table Name: Students and Courses
SnCStudentID
SnCCourseID

Example on Normalization

3rd Normal Form: (no column entry should be dependent on any other entry (value) other than the key for the table)

Table Name: Student Information
StudentID (Primary Key)
StudentName

Table Name: Advisor Information
AdvisorID
AdvisorName

Table Name: Course Information
CourseID (Primary Key)
CourseTitle
CourseDescription

Table Name: Professor Information
ProfessorID
CourseProfessor

Table Name: Students and Courses
SnCStudentID
SnCCourseID

RDBMS Concepts – Primary Key

- Primary Key (PK) values uniquely identify each row in a table.

- Primary Key Rules
  - A Primary Key is required for every table.
  - Only one Primary Key is allowed in a table.
  - Primary Keys may consist of one or more columns.
  - Primary Keys cannot have duplicate values (ND).
  - Primary Keys cannot be null (NN).
  - Primary Keys are considered “non-changing” values (NC).

RDBMS Concepts – Primary Key

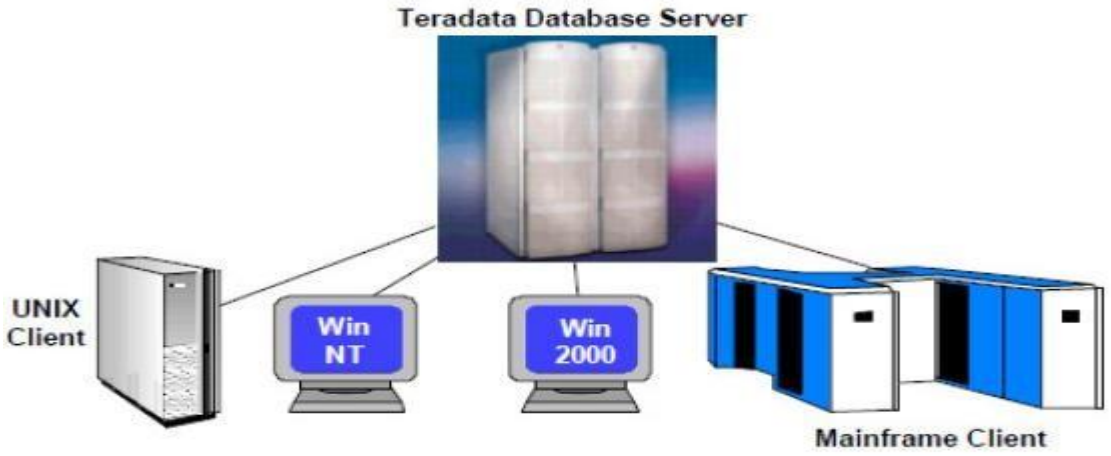
- Foreign Key (FK) values model relationships.
- FKs are optional.
- More than one FK is allowed per table.
- FKs can be made up of more than one column.
- Duplicate values may be allowed.
- Missing (null) FK values may be allowed.
- Changes to FKs are allowed.
- Each FK value must exist somewhere as a PK value (i.e.
- have referential integrity).

RDBMS Concepts – Relational Advantages

- Advantages of a Relational Database compared to other database methodologies include:
- More flexible than other types
- Allowing businesses to quickly respond to changing conditions
- Being data-driven vs. application driven
- Modeling the business, not the processes
- Makes applications easier to build because the data does more of the work
- Being easy to understand
- Supporting trend toward end-user computing
- No need to know the access path
- Solidly founded in set theory

Teradata Overview

- Teradata is a Relational Database Management System (RDBMS) that drives an company's data warehouse.
  - An open system, compliant with industry ANSI standards.
  - Capable of supporting many concurrent users from various client platforms (over a TCP/IP or IBM channel connection).
  - Runs on various OS like Novell SUSE Linux, MS Windows Enterprise Server & other traditional OS. Hence it is considered an open architecture
  - Built on a parallel architecture.



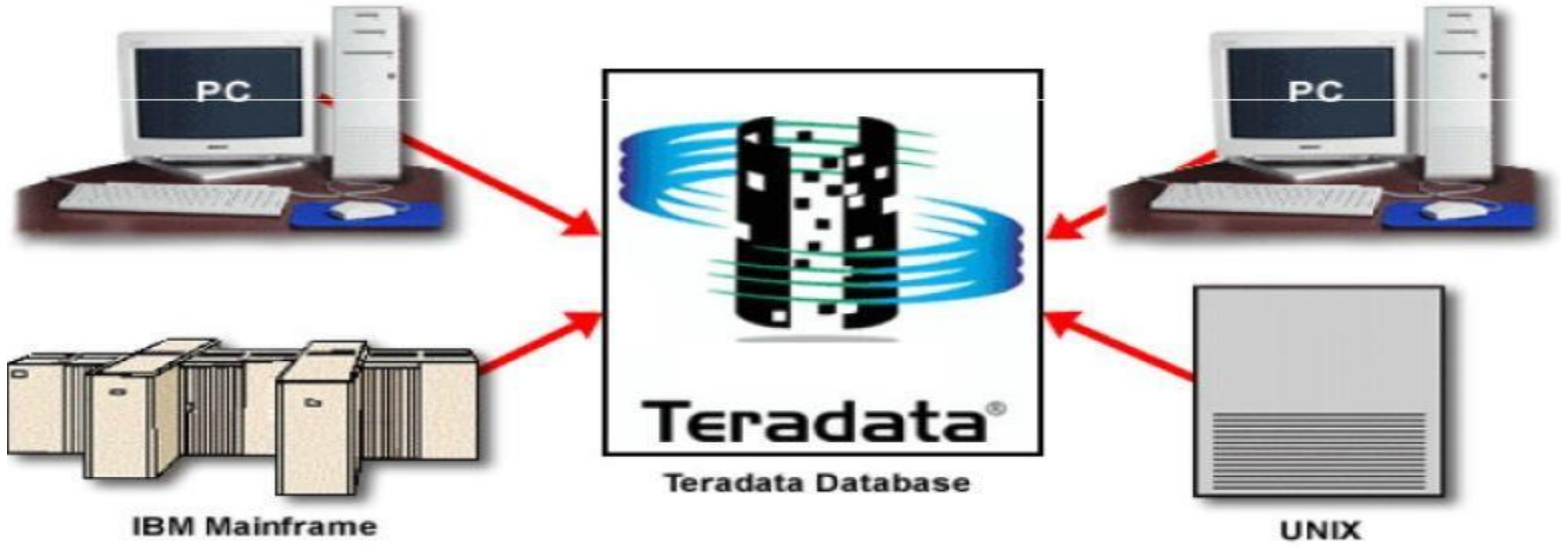
Teradata Overview – Why Teradata?

- Single Data Store
- Scalability
- Unconditional Parallelism (parallel architecture)
- Parallel aware optimizer

Teradata – Single Data Store

- Teradata DB acts as a single data store, with multiple client applications making inquiries against it concurrently.

- Instead of replicating a database for different purpose, with the Teradata DB you store the data once & use it for many applications.



Teradata – Scalability

- "Linear Scalability" means that as you add components to the system, the performance increase is linear. Linear Scalability enables the system to grow to support more users/data/queries/complexity of queries without experiencing performance degradation.
- Teradata DB was the 1st commercial database system to scale & support a trillion bytes of data.
- Teradata DB can scale from 100 GB's to over 100+ petabytes of data.

The chart below lists the meaning of the prefixes:

Prefix	Exponent	Meaning
kilo-	10^3	1,000 (thousand)
mega-	10^6	1,000,000 (million)
giga-	10^9	1,000,000,000 (billion)
tera-	**10^{12}**	**1,000,000,000,000 (trillion)**
peta-	10^{15}	1,000,000,000,000,000 (quadrillion)
exa-	10^{18}	1,000,000,000,000,000,000 (quintillion)

Teradata Overview – Designed for Today's Business

- Teradata meets the business needs of today & tomorrow with:
  - Relational model - standard for database design.
  - Huge capacity (billions of rows, terabytes of data).
  - High performance parallel processing
  - Single database server for multiple clients ("Single Version of the Truth").
  - Network and mainframe connectivity.
  - Industry standard access language (SQL).
  - Manageable growth through modularity.
  - Fault tolerance at all levels of hardware and software.
  - Data integrity and reliability.

Teradata Overview – Review

- Designed to process large quantities of detail data.
- Ideal for data warehouse applications.
- Parallelism makes easy access to very large tables possible.
- Open architecture - uses industry standard components.
- Scalability - Performance increase is linear as components are added.
- Runs as a database server to client applications.
- Runs on multiple hardware platforms.

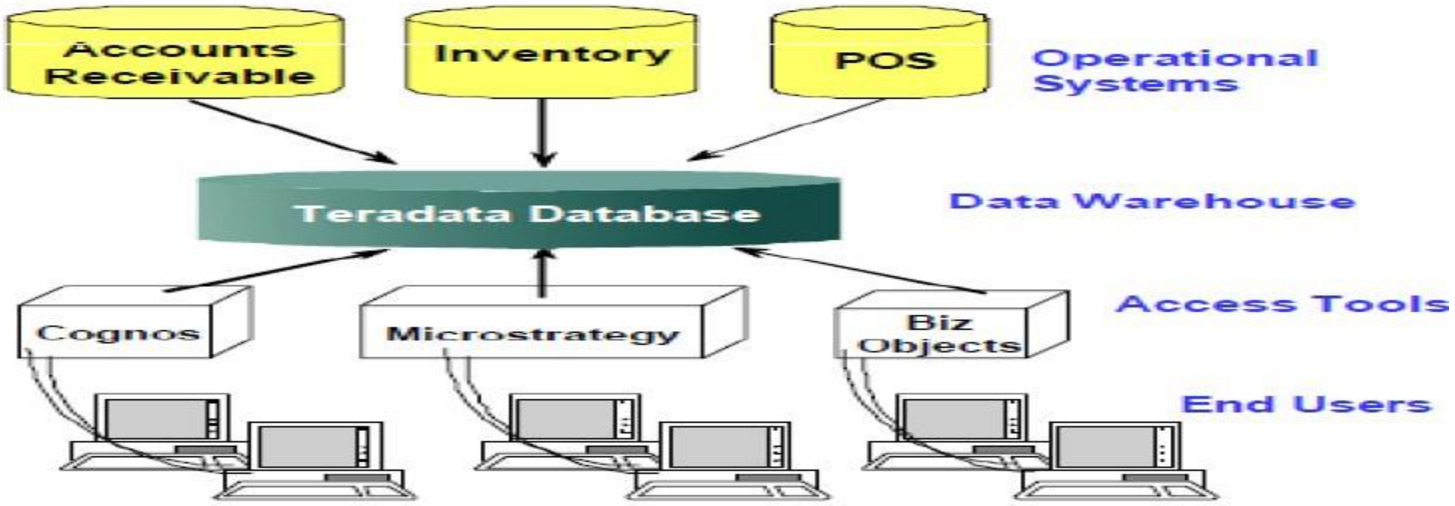
Teradata and Data Warehouse

Evolution of Data Processing

	Type	Examples	Number of Rows Accessed	Response Time
TRADITIONAL	OLTP	Update a checking account to reflect a deposit. Debit transaction takes place against current balance to reflect amount of money withdrawn at ATM.	Small	Seconds
TRADITIONAL	DSS	How many child size blue jeans were sold across all of our Eastern stores in the month of March? What were the monthly sales of shoes for retailer X?	Large	Seconds or minutes
TODAY	OLCP	Instant credit—How much credit can be extended to this person? What interest rate for a loan can be given to this customer?	Small to moderate against multiple databases	Minutes
TODAY	OLAP	Show the top ten selling items across all stores for 1997. Show a comparison of sales from this week to last week.	Large of detail rows or moderate of summary rows	Seconds or minutes

Teradata & Data Warehouse

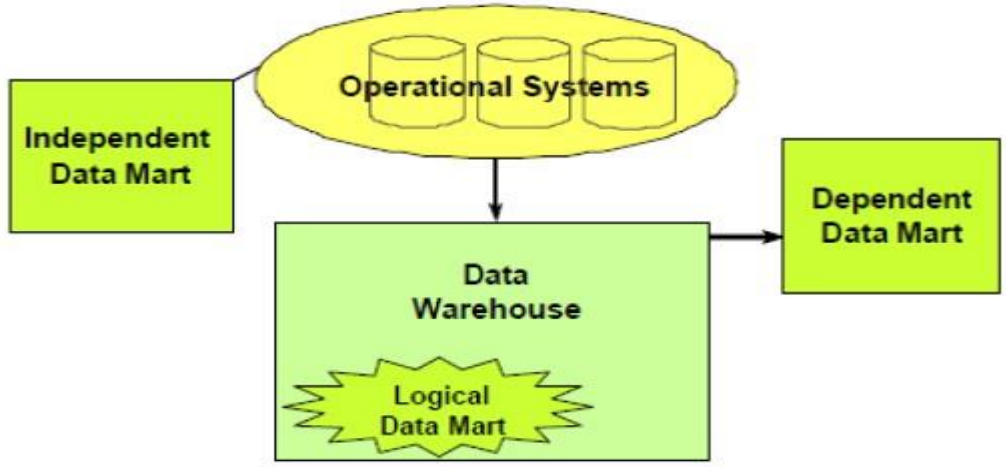
- Data Warehouse:
- A central, enterprise-wide database that contains information extracted from operational systems.
  - Based on enterprise wide model
  - Can begin small but may grow large rapidly
  - Populated by extraction/loading of data from operational systems
  - Responds to end-user “what if” queries



Teradata & Data Warehouse

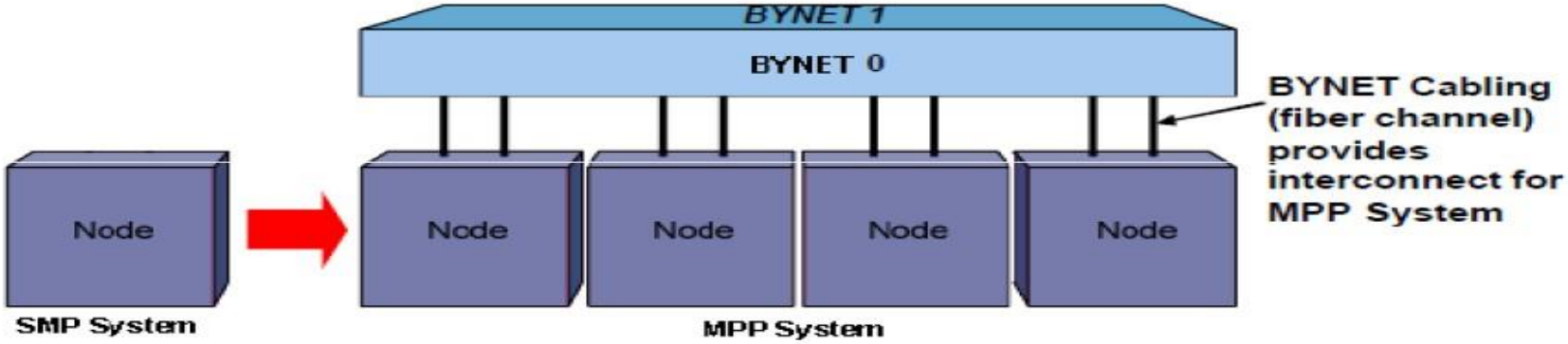
- Data marts:
- A data mart is a special purpose subset of enterprise data for a particular function or application. It may contain detail or summary data or both.

- Data mart types:
  - Independent—created directly from operational systems to a separate physical datastore.
  - Logical—exists as a subset of existing data warehouse.
  - Dependent—created from data warehouse to a separate physical data store.



Teradata → SMP v/s MPP

- SMP → Symmetric Multiprocessing Platforms manage terabytes of data to support an entry-level data warehousing system.
- MPP → Massively Parallel Processing systems can manage hundreds of petabytes of data. You can start with a couple of nodes



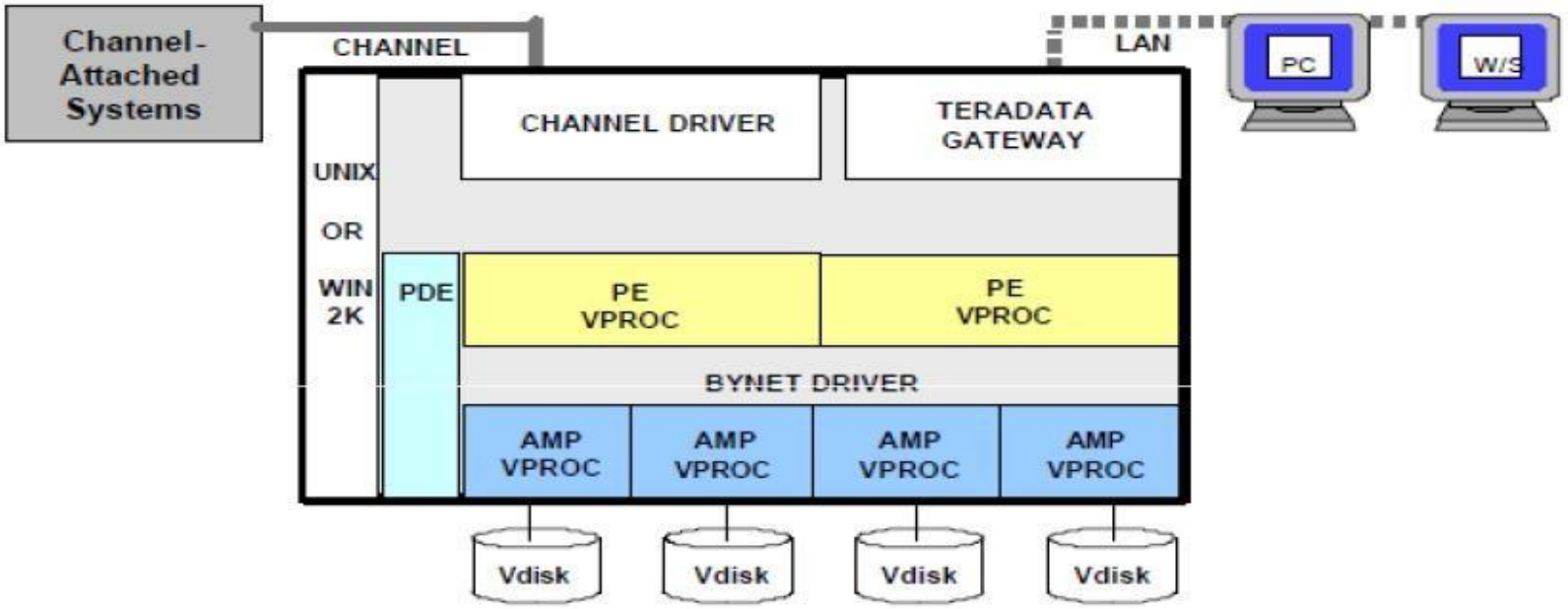
- Multiple nodes are configured into a Massively Parallel Processing (MPP) system.
- A physical message-passing layer called the BYNET is used to interconnect multiple nodes.
- Teradata is linearly expandable—as your database grows, additional nodes may be added.
- The BYNET can support 512 nodes.

Teradata and Data Warehouse

- Active Data Warehousing
- Performance
  - Response time within seconds.
- Scalability
  - Support for large amounts of detailed data, mixed workloads (both tactical and strategic queries) for mission critical applications, and concurrent users.
- Availability and Reliability
  - – 7 x 24
- Data Freshness
  - – Accurate, up to the minute, data including access to operational data store level information.

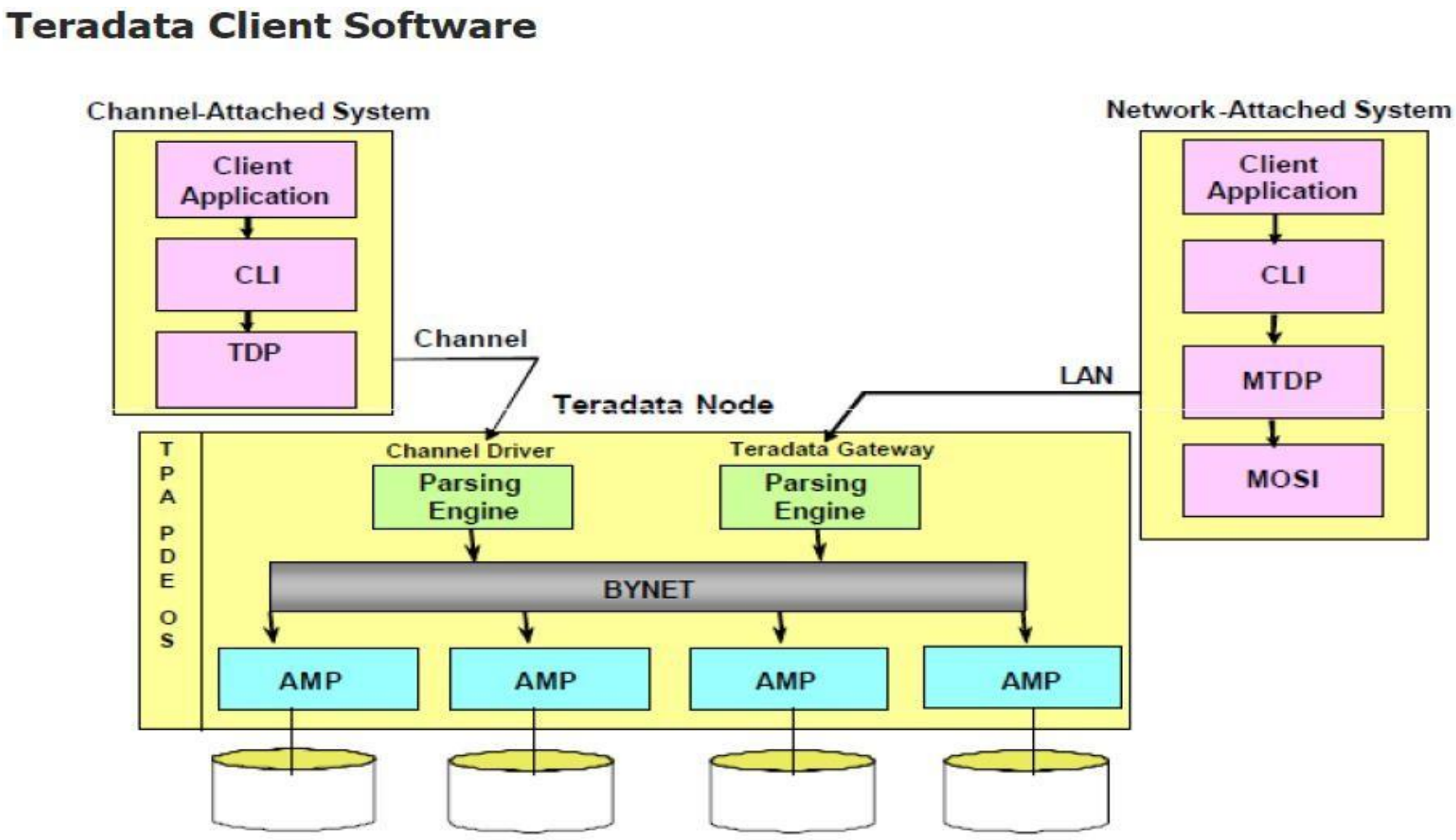
Teradata → Components and Architecture

Node (SMP)



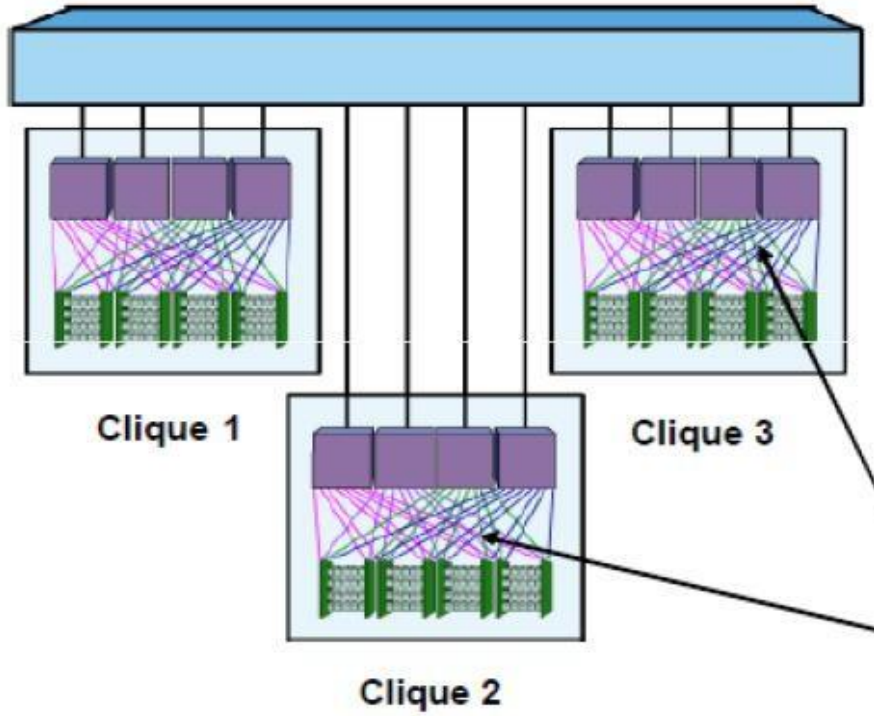
- Teradata software, the LAN gateway, and channel-driver software run as processes.
- AMPs and PEs are virtual processors (vprocs) which run under Parallel Database Extensions (PDE).
- AMPs are associated with virtual disks (vdisks).
- A single node is called a Symmetric Multi-Processor (SMP).

Teradata → Components and Architecture



Components and Architecture

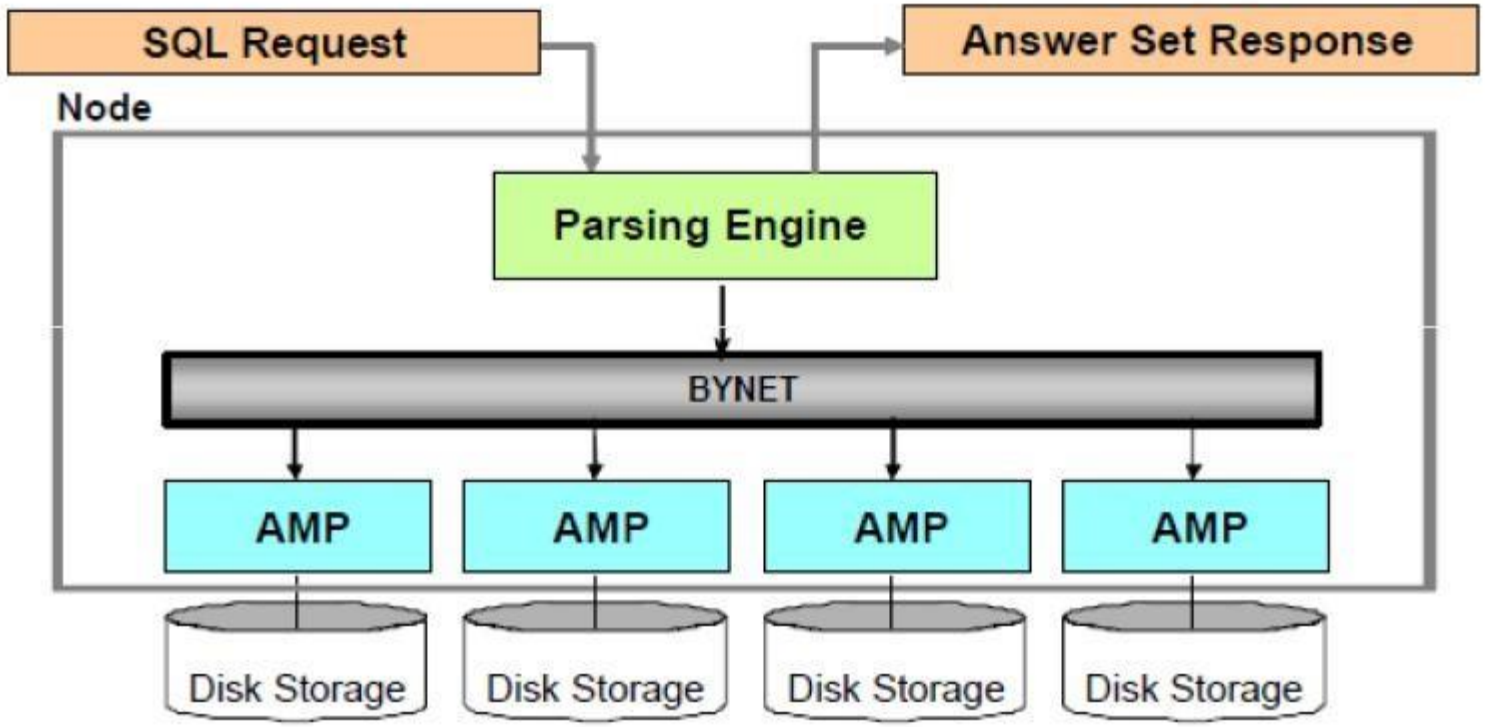
Cliques



- A clique is a defined set of nodes with failover capability.
- All nodes in a clique are able to access the vdisks of all AMPs in the clique.
- If a node fails, its vprocs will migrate to the remaining nodes in the clique.
- Each node can support 128 vprocs.

Components and Architecture

Processing flow of SQL Request in Teradata system



Components and Architecture

- Major Components of a Teradata System

- Parsing Engine (PE)
  - The Parsing Engine (PE) is a component that interprets SQL requests, receives input records, and passes data. To do that it sends the messages through the BYNET to the AMPs.

- BYNET
  - The BYNET is the message-passing layer. It determines which AMP(s) (Access Module Processor) should receive a message.

Components and Architecture

- Access Module Processor (AMP)
  - The AMP is a virtual processor (vproc) designed for and dedicated to managing a portion of the entire database. It performs all database management functions such as sorting, aggregating, and formatting data. The AMP receives data from the PE, formats rows, and distributes them to the disk storage units it controls. The AMP also retrieves the rows requested by the Parsing Engine.

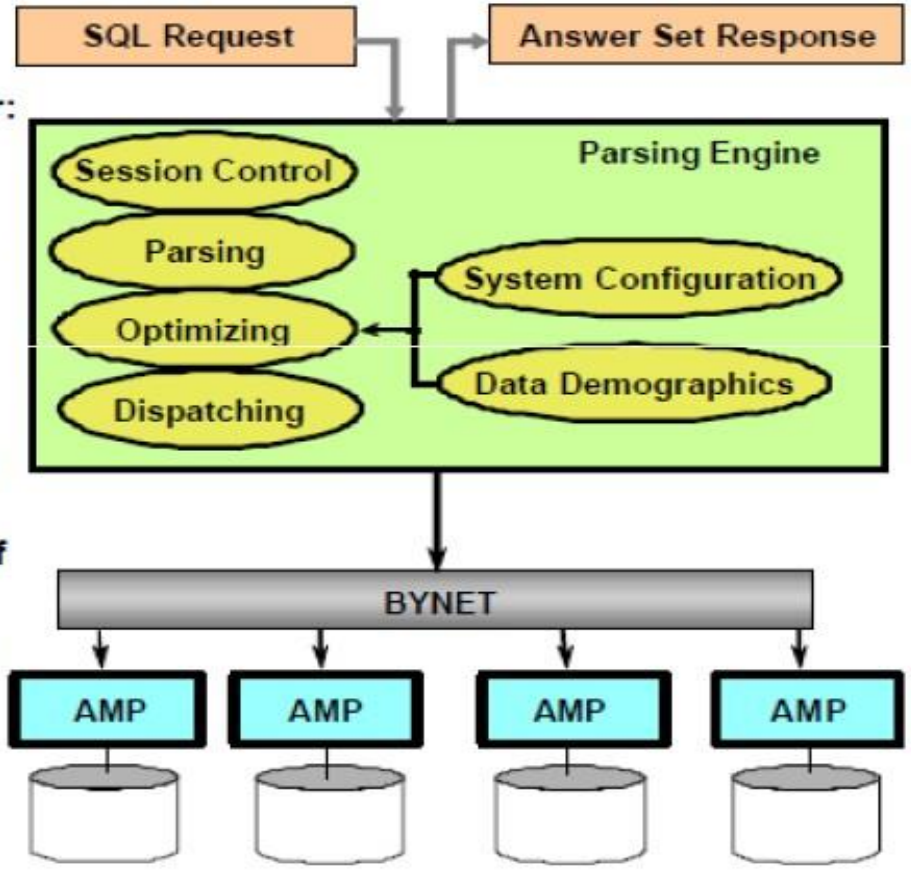
- Disks
  - Disks are disk drives associated with an AMP that store the data rows. On current systems, they are implemented using a disk array

Components and Architecture

The Parsing Engine (PE)

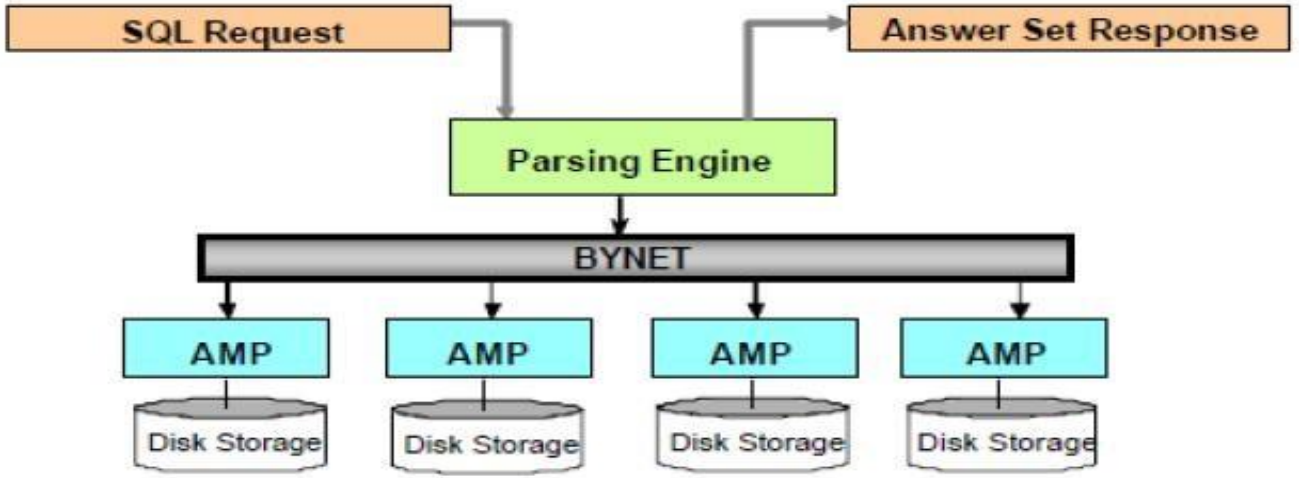
The Parsing Engine is responsible for:

- Managing individual sessions (up to 120 sessions per PE)
- Parsing and optimizing your SQL requests
- Building query plans with the parallel-aware, cost-based, intelligent Optimizer
- Dispatching the optimized plan to the AMPs
- EBCDIC/ASCII input conversion (if necessary)
- Sending the answer set response back to the requesting client



Components and Architecture

BYNET



Dual redundant, fault-tolerant, bi-directional interconnect network that enables:

- **Automatic load balancing of message traffic.**
- **Automatic reconfiguration after fault detection.**
- **Scalable bandwidth as nodes are added.**

The BYNET connects all the AMPs on the system:

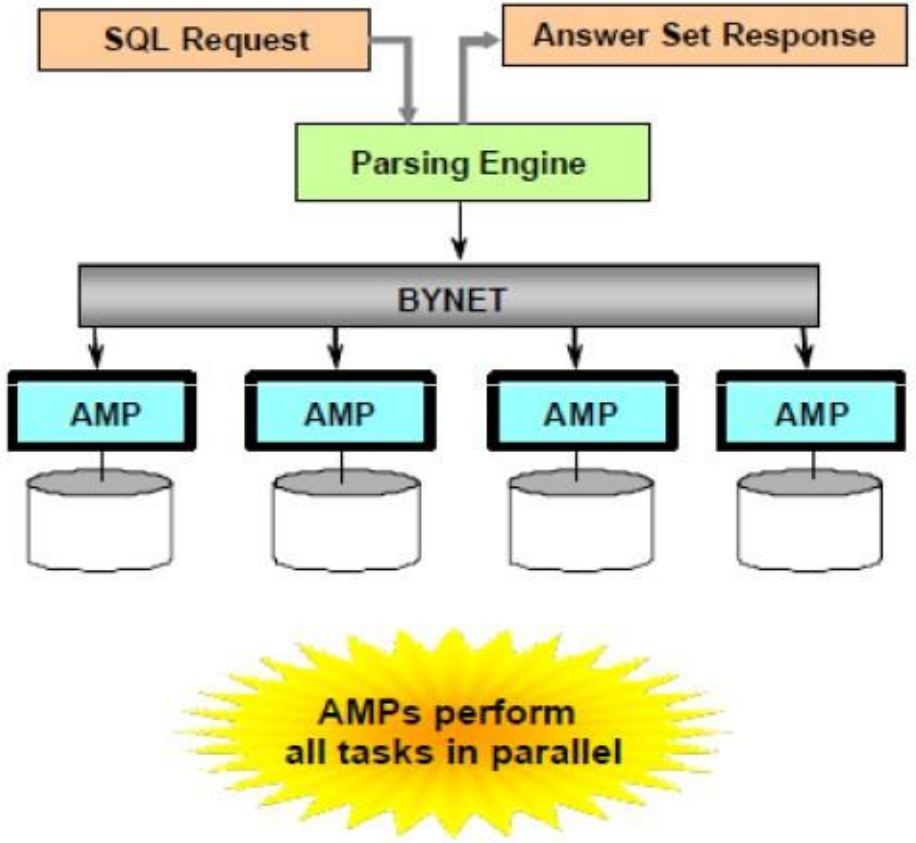
- **Between nodes, the BYNET hardware carries broadcast and point-to-point communications.**
- **On a node, BYNET software and PDE together control which AMPs receive a multicast communication.**

Components and Architecture

The Access Module Processor (AMP)

AMPs are responsible for:

- Storing and retrieving rows to and from disks
- Lock management
- Sorting rows and Aggregating columns
- Join processing
- Output conversion and formatting (ASCII, EBCDIC)
- Creating answer sets for clients
- Disk space management and Accounting
- Special utility protocols
- Recovery processing



Teradata Utilities

- Teradata utilities leverage the Teradata Database's high performance capabilities & are fully parallel & scalable. They run on both SMP & MPP systems.
- Utilities which are commonly used by all Teradata Developer':-

- **Query Submitting Utilities**
  - BTEQ
  - Teradata SQL Assistant
- **Load and Unload Utilities**
  - FastLoad
  - MultiLoad
  - TPump
  - FastExport
  - Teradata Parallel Transporter (TPT)

Teradata Utilities

- Utilities which are commonly used by all Teradata DBA's:-

  - **Administrative Utilities**
    - Teradata Manager
    - Teradata Dynamic Workload Manager (TDWM)
    - Priority Scheduler
    - Database Query Log (DBQL)
    - Teradata Workload Analyzer
    - Performance Monitor (PMON)
    - Teradata Active Systems Management (TASM)
    - Teradata Analyst Pack
  - **Archive Utilities**
    - Archive Recovery Facility (ARC)
    - NetVault (third party)
    - NetBackup (third party)
  - **Tools that support a Multi-System Environment**
    - Viewpoint
    - Unity
    - Teradata Multi-Systems Manager (TMSM)
    - Data Mover

Review Questions

- Match each term with its definition below:
  - 1. Database
  - 2. Table
  - 3. Relational database
  - 4. Primary Key
  - 5. Null
  - 6. Foreign Key

- a - A set of columns which uniquely identify a row. b - A set of logically related tables.
- c - One or more columns that are a PK somewhere in the database. d - The absence of a value.
- e - A two-dimensional array of rows and columns.
- f - A collection of permanently stored data.

Answers

- Match each term with its definition below:
  - _f_ 1. Database
  - _e_ 2. Table
  - _b_ 3. Relational database
  - _a_ 4. Primary Key
  - _d_ 5. Null
  - _c_ 6. Foreign Key

- a - A set of columns which uniquely identify a row
- b - A set of logically related tables
- c - One or more columns that are a PK somewhere in the database d - The absence of a value
- e - A two-dimensional array of rows and columns f - A collection of permanently stored data

QUIZ

- 1. Which type of database is built around both set processing and row-at-a-time processing?
  - A. network
  - B. data mart
  - C. relational
  - D. hierarchical
  - E. inverted list

Option C

- 2.Which three are terms that are associated with star schema and dimensional modeling? (Choose three.)
  - A. facts
  - B. entities
  - C. attributes
  - D. snowflakes
  - E. dimensions
  - F. relationships

Option A,D,E

QUIZ

- 3.What is a characteristic of a dependent data mart?
  - A. contains unique data
  - B. is the system of record
  - C. sourced from operational systems
  - D. subset of data often used by a department

Option D

- 4.What is used by PEs and AMPs for internal communications?
  - A. the YNET
  - B. the BYNET
  - C. a proprietary LAN
  - D. an inter-node WAN

Option B

Module 1: Lab Exercise

- Create a copy of the table employee_sales.employee using the subquery form with a "SELECT *", populating it with data. Show the definition of it to check primary index. Then perform a SELECT * against it.

- Use a subquery form to create a table, with data, from a left outer join against Employee, Department and Job. Project last name, first name, department name and job description. Verify the results.

Module 1: Lab Exercise

- Using the subquery form of CREATE TABLE AS, create a new Job table and assign the column for hourly billing rate a default of zero. Leave all column names the same as in the Job table. Insert the default into the new table from the subquery and use the WITH DATA option. Select from the table to confirm the result.

- Optional: Perform the following and follow it up by selecting from the table and showing the definition.

- CREATE TABLE y_me AS (SELECT 1 a, 'abc' d, 1e6 j) WITH DATA;

Summary includes all differences between 13 and 14

- The Release Notes are part of the Documentation, available at

  - http://www.info.teradata.com/edownload.cfm?itemid=102320028
  - http://www.info.teradata.com/edownload.cfm?itemid=113480006 http://www.info.teradata.com