**SQL** ( i/ˈɛs kjuː ˈɛl/, or i/ˈsiːkwəl**/;** Structured Query Language) is a special-purpose programming language designed for managing data held in a relational database management system (RDBMS), or for stream processing in a relational data stream management system (RDSMS**).**

**front end** :(of a device or program) directly accessed by the user and allowing access to further devices, programs, or databases**.**

**Front End Development**, also known as Client Side Development is the practice of producing HTML, CSS and Javascript for a website or web application so that a user can see and interact with them directly.

**A front-end developer** architects and develops websites and applications using web technologies (i.e. HTML, CSS, and JavaScript) which run natively in a web browser or act as compilation input for non-web browser environments

**Middleware**is the software that connects software components or enterprise applications. **Middleware** is the software layer that lies between the operating system and the applications on each side of a distributed computer network.

**back** end: denoting a subordinate processor or program, not directly accessed by the user, which performs a specialized function on behalf of a main processor or software system.

VARCHAR is variable-length.

CHAR is fixed length.

In [database management](http://www.webopedia.com/TERM/D/database_management_system_DBMS.html), a field that is allowed to have no values is called **nullable.** Depending on the [application](http://www.webopedia.com/TERM/A/application.html), nullable may also be called a null reference or null object.

A **varchar** or Variable Character Field is a set of character data of indeterminate length. The term varchar refers to a data type of a field (or column) in a database management system which can hold letters and numbers.

**int** i = 123; When an integer literal has no suffix, its type is the first of these types in which its value can be represented: int, uint, long, ulong. In this example, it is of the type int.

A **decimal** number, or just decimal, refers to any number written in decimal notation, although it is more commonly used to refer to numbers that have a fractional part separated from the integer part with a decimal separator

* general a double has 15 to 16 decimal digits of precision, while float only has 7.
* Float - 7 digits (32 bit)
* Double-15-16 digits (64 bit)
* **Decimal** -28-29 significant digits (128 bit)
* **tinyint**: 1 byte, -128 to +127 / 0 to 255 (unsigned)
* **smallint**: 2 bytes, -32,768 to +32,767 / 0 to 65,535 (unsigned)
* **mediumint**: 3 bytes, -8,388,608 to 8,388,607 / 0 to 16,777,215 (unsigned)
* **int/integer**: 4 bytes, -2,147,483,648 to +2,147,483,647 / 0 to 4,294,967,295 (unsigned)
* bigint: 8 bytes, -9,223,372,036,854,775,808 to 9,223,372,036,854,775,807 / 0 to 18,446,744,073,709,551,615 (unsigned)

**DATE** and **TIMESTAMP** have the same size (7 bytes). Those bytes are used to store century, decade, year, month, day, hour, minute and seconds. But TIMESTAMP allows to store additional info such as fractional seconds (11 bytes) and fractional seconds with timezone (13 bytes).

An **index** is used to speed up the performance of queries. **Indexes** are optional structures associated with tables. Indexes can be created to increase the performance of data retrieval. Just as the index in this manual helps you quickly locate specific information, an Oracle index provides an access path to table data. It does this by reducing the number of database data pages that have to be visited/scanned. In SQL Server, a clustered index determines the physical order of data in a table. There can be only one clustered index per table (the clustered index IS the table

**Views** are customized presentations of data in one or more tables or other views. A view can also be considered a stored query. Views do not actually contain data. Rather, they derive their data from the tables on which they are based, referred to as the **base tables** of the views.

**Clusters** are groups of one or more tables physically stored together because they share common columns and are often used together. Because related rows are physically stored together, disk access time improves.

A **process** is a "thread of control" or a mechanism in an operating system that can run a series of steps. Some operating systems use the terms job or task. A process generally has its own private memory area in which it runs.

Keys describe the relationships between the different tables and columns of a relational database.

A **primary key**, also called a primary keyword, is a key in a relational database that is unique for each record. It is a unique identifier, such as a driver license number, telephone number (including area code), or vehicle identification number (VIN). A relational database must always have one and only one primary key.

In the context of relational databases, a **foreign key** is a field (or collection of fields) in one table that uniquely identifies a row of another table. In simpler words, the foreign key is defined in a second table, but it refers to the primary key in the first table.

In relational model database design, a **natural key** is a key that is formed of attributes that already exist in the real world. For example, a US citizen's social security number could be used as a natural key.

**serv·er** :a computer or computer program that manages access to a centralized resource or service in a network.

**types of index**

[**https://technet.microsoft.com/en-us/library/ms175049(v=sql.110).aspx**](https://technet.microsoft.com/en-us/library/ms175049(v=sql.110).aspx)

A table without a clustered-index is called a *“heap table”.* A heap table has not its data sorted. The SQL server has to scan the entire table in order to locate the data, in a process called a *“scan”.*

*Table scan*

<https://www.giantstride.gr/sql-indexing-part2/>

In the case of a clustered index, the data are sorted on the key values (columns) of the index. The SQL server is now able to locate the data by navigating down from the root node, to the branch and finally to the leaf nodes of the B-tree structure of the index. This process called a *“seek”.* The later approach is much faster, when you want to filter or sort the data you want to retrieve.

A **clustered index** sorts and stores the data rows of the table or view in order based on the clustered index key. The clustered index is implemented as a B-tree index structure that supports fast retrieval of the rows

**nonclustered index** can be defined on a table or view with a clustered index or on a heap. Each index row in the nonclustered index contains the nonclustered key value and a row locator. This locator points to the data row in the clustered index or heap having the key value. The rows in the index are stored in the order of the index key values, but the data rows are not guaranteed to be in any particular order unless a clustered index is created on the table.

A **unique index** ensures that the index key contains no duplicate values and therefore every row in the table or view is in some way unique.

Uniqueness can be a property of both clustered and nonclustered indexes.

An xVelocity memory optimized **columnstore index** based on vertical partitioning of the data by columns, stored as large objects (LOB)

**Index with included columns**:A nonclustered index that is extended to include nonkey columns in addition to the key columns.

**Index on computed column** :An index on a column that is derived from the value of one or more other columns, or certain deterministic inputs.

**Filtered**An optimized nonclustered index, especially suited to cover queries that select from a well-defined subset of data. It uses a filter predicate to index a portion of rows in the table. A well-designed filtered index can improve query performance, reduce index maintenance costs, and reduce index storage costs compared with full-table indexes.

A **spatial index** provides the ability to perform certain operations more efficiently on spatial objects (spatial data) in a column of the geometry data type. The spatial index reduces the number of objects on which relatively costly spatial operations need to be applied.

**XML**:shredded, and persisted, representation of the XML binary large objects (BLOBs) in the xml data type column

**Full-text**A special type of token-based functional index that is built and maintained by the Microsoft Full-Text Engine for SQL Server. It provides efficient support for sophisticated word searches in character string data.

In data warehousing, a **dimension table** is one of the set of companion tables to a fact table. The fact table contains business facts (or measures), and foreign keys which refer to candidate keys (normally primary keys) in the dimension tables. A dimension table is a table in a star schema of a data warehouse. A dimension table stores attributes, or dimensions, that describe the objects in a fact table.

A **fact table** is the central table in a star schema of a data warehouse. A fact table stores quantitative information for analysis and is often denormalized

**Dimensional modeling** (DM) names a set of techniques and concepts used in data warehouse design. It is considered to be different from entity-relationship modeling (ER). Dimensional Modeling does not necessarily involve a relational database. Dimensional modeling is one of the methods of data modeling, that help us store the data in such a way that it is relatively easy to retrieve the data from the database.

**Data modeling** is a process used to define and analyze data requirements needed to support the business processes within the scope of corresponding information systems in organization

In computing, a **snowflake schema** is a logical arrangement of tables in a multidimensional database such that the entity relationship diagram resembles a snowflake shape. The snowflake schema is represented by centralized fact tables which are connected to multiple dimensions. [citation needed]. In data warehousing, snowflaking is a form of dimensional modeling where dimensions are stored in multiple related dimension tables. The snowflake schema is an extension of the star schemaThe main advantage of the snowflake schema is the improvement in query performance due to minimized disk storage requirements and joining smaller lookup tables. The main disadvantage of the snowflake schema is the additional maintenance efforts needed due to the increase number of lookup tables. In a star schema, each dimension is represented by a single dimensional table, whereas in a snowflake schema, that dimensional table is normalized into multiple lookup tables, each representing a level in the dimensional hierarchy.

**Precision** is the number of digits in a number. **Scale** is the number of digits to the right of the decimal point in a number. For example, the number 123.45 has a precision of 5 and a scale of 2.

In SQL, we have the following constraints:

**NOT NULL** - Indicates that a column cannot store NULL value

UNIQUE - Ensures that each row for a column must have a unique value

**PRIMARY KEY** - A combination of a NOT NULL and UNIQUE. Ensures that a column (or combination of two or more columns) have an unique identity which helps to find a particular record in a table more easily and quickly

FOREIGN KEY - Ensure the referential integrity of the data in one table to match values in another table

**CHECK** - Ensures that the value in a column meets a specific condition

**DEFAULT** - Specifies a default value when specified none for this column

For example, the **mantissa** of the number 12.345 is .345

There are following **six** phases in every **Software development life cycle** model:

1. Requirement gathering and analysis
2. Design
3. Implementation or coding
4. Testing
5. Deployment
6. Maintenance

<http://istqbexamcertification.com/what-are-the-software-development-life-cycle-sdlc-phases/>

The **waterfall model** is a sequential design process, used in software development processes, in which progress is seen as flowing steadily downwards (like a waterfall) through the phases of conception, initiation, analysis, design, construction, testing, production/implementation and maintenance.

**Agile software development** is a group of software development methods in which requirements and solutions evolve through collaboration between self-organizing, cross-functional teams. It promotes adaptive planning, evolutionary development, early delivery, continuous improvement, and encourages rapid and flexible response to change.[1]

<http://www.base36.com/2012/12/agile-waterfall-methodologies-a-side-by-side-comparison/>

A **data definition language or data description language (DDL)** is a syntax similar to a computer programming language for defining data structures, especially database schemas.

**data manipulation language (DML)** is a family of syntax elements similar to a computer programming language used for selecting, inserting, deleting and updating data in a database. Performing read-only queries of data is sometimes also considered a component of DML

DDL

Data Definition Language (DDL) statements are used to define the database structure or schema. Some examples:

CREATE - to create objects in the database

ALTER - alters the structure of the database

DROP - delete objects from the database

TRUNCATE - remove all records from a table, including all spaces allocated for the records are removed

COMMENT - add comments to the data dictionary

RENAME - rename an object

DML

Data Manipulation Language (DML) statements are used for managing data within schema objects. Some examples:

SELECT - retrieve data from the a database

INSERT - insert data into a table

UPDATE - updates existing data within a table

DELETE - deletes all records from a table, the space for the records remain

MERGE - UPSERT operation (insert or update)

CALL - call a PL/SQL or Java subprogram

EXPLAIN PLAN - explain access path to data

LOCK TABLE - control concurrency

TCL

-----

Transaction Control Language commands are used to manage transactions in the database. These are used to manage the changes made by DML-statements. It also allows statements to be grouped together into logical transactions

COMMIT: Commit command is used to permanently save any transaction

into the database.

ROLLBACK: This command restores the database to last committed state.

It is also used with savepoint command to jump to a savepoint

in a transaction.

SAVEPOINT: Savepoint command is used to temporarily save a transaction so

that you can rollback to that point whenever necessary.

DCL

----

A Data Control Language is a syntax similar to a computer programming language used to control access to data stored in a database (Authorization). In particular, it is a component of Structured Query Language (SQL).

Grant

revoke

***System Variables***

IBM® InfoSphere® DataStage® provides a set of variables containing useful system information that you can access from a transform or routine. System variables are read-only

<http://www-01.ibm.com/support/knowledgecenter/SSZJPZ_8.7.0/com.ibm.swg.im.iis.ds.serverjob.dev.doc/topics/r_dsvjbref_System_Variables.html>

DSHostName

This macro will give you the Host name of DataStage server.

DSJobController

It returns the name of the job that started the current one (if any).

DSJobInvocationId

This will return the InvocationId of job if job is multiple instance enables

DSJobName

As name says, return the Job name

DSJobStartDate

Return Job Start Date

DSJobStartTime

Return Job Start Time

DSJobStartTimestamp

Return Job Start timestamp

DSJobWaveNo

This is kind of Sequence no generated for each successful job run.

When compiled – 0

Each successful run – last+1

Each un-successful run – last+0

DSProjectName

Return Project Name of running job

<http://www-01.ibm.com/support/knowledgecenter/SSZJPZ_9.1.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/c_parjdev_Slowly_Changing_Dimension_Stage.html?lang=en>

If RCP is enabled for a project, you can just define the columns you are interested in using in a job, but ask InfoSphere DataStage to propagate the other columns through the various stages. So such columns can be extracted from the data source and end up on your data target without explicitly being operated on in between. Sequential files, unlike most other data sources, do not have inherent column definitions, and so InfoSphere DataStage cannot always tell where there are extra columns that need propagating. You can only use RCP on sequential files if you have used the Schema File property to specify a schema which describes all the columns in the sequential file. You need to specify the same schema file for any similar stages in the job where you want to propagate columns. Stages that will require a schema file are:

Sequential File

File Set

External Source

External Target

Column Import

Column Export

Types of Parallelism

There are two basic types of parallelism

1) Pipeline parallelism

2) Partitioning parallelism

àIn practice we will be combining pipeline and partition parallelism to achieve the greater performance gains.

àTo run a pipeline parallelism at least three processors are required, (1-source, 1-target, 1-business logic)

Pipeline Parallelism

àIf we run a job on a system with at least three processors the stage reading would start on one processor and start filling a pipeline with the data it had read.

à The transformation stage would start running on second processor as soon as there was a data in a pipeline, process it and start filling another pipeline.

à The target stage would start running on 3rd processor as soon as there was data in pipeline all three stages are operating simultaneously.

It is the ability for a downstream stage to begin processing a row as soon as an upstream stage has finished processing that row (rather than processing one row completely through the job before beginning the next row). In Parallel jobs, it is managed automatically.

Partitioning Parallelism:

Using Partitioning Parallelism the same job would effectively be run by several processors simultaneously

Each processor handles a separate subset of total data.

The following are the types of data partitioning methods.

1. auto
2. Round robin
3. same
4. db2
5. hash
6. modulus
7. random
8. range
9. entire

Partitioning parallelism means that entire record set is partitioned into small sets and processed on different nodes. That is, several processors can run the same job simultaneously, each handling a separate subset of the total data.

For example if there are 100 records, then if there are 4 logical nodes then each node would process 25 records each. This enhances the speed at which loading takes place.

How to run a Shell Script within the scope of a Data stage job?

Ans:- By using "ExcecSH" command at Before/After job properties.

How will you run data stage from Unix? What’s the command?

Dsjob – status –project name – job name

A **dataset** is a file/stage where the data can be **read**

**directly** by the **DataStage**, whereas a **file set** needs to be

**converted** into **DataStage** readable format (which happens

internally)

A **mainframe computer** is a very large computer capable of handling and processing very large amounts of data quickly. They are used by large institutions, such as government agencies and large corporations.

**Mainframe jobs**. These are available only if you have Enterprise MVS Editioninstalled. A mainframe job is compiled and run on the mainframe. Data extractedby such jobs is then loaded into the data warehouse

**Combinality** is a property of DataStage in which it combines two or   
more Operators into a single Process.  
You can explicitely enable/disable it with the help of   
APT\_DISABLE\_COMBINATION environment variable at the job   
level or at the Project level as well.  
At stage level also you can enable/disable it:  
In Stage properties>> Advance>> Operator Combinality.  
By default it's Auto, so that datastage will decide whether   
to combine or not.  
  
You can see this combinality behavior by using the   
following Environment Variable:  
APT\_DUMP\_SCORE

In datastage both peek and copy stages are called stub stages.

**Copy stage** allows us to copy the input records into multiple o/p links, includes file stages or processing stages.

**peek stage** is one of the debug stage, which allows us to view

the sample no of records in log or to output file, but not to

any other processing stages.

**Information technology (IT)** is the application of computers and telecommunications equipment to store, retrieve, transmit and manipulate data,[1] often in the context of a business or other enterprise

**Computer Science** is the study of computers. You typically **learn** about **hardware** and **operating systems**. Things like **registers,** multitasking **kernels**, data buses, address buses. It's real techy stuff.   
**Information Technology** is the more about the study of the **use** of **computers** to **solve human/business problems**. You study some **programming, databases, applications, etc**

**IBM Infosphere Information Server Version 8.5**

The term **"fields**" refers to **columns, or vertical categories of data**; the term **"records"** refers to **rows, or horizontal groupings of unique field data**

In creating a database, **normalization** is the process of organizing it into tables in such a way that the results of using the database are always unambiguous and as intended.

**performance tuning** :monitoring the current performance of the **DataStage** jobs. And implementing process to improve the performance of jobs

The **Transformer stage** is a **processing stage.** It appears under the **processing** category in the **tool palette**. Transformer stages allow you to create **transformations** to apply to your data. These transformations can be simple or complex and can be applied to individual columns in your data. Transformations are specified using a set of functions.

Transformer stages can have a **single** input and **any** number of **outputs**. It can also have **a reject link** that takes any rows which have not been written to any of the

**IBM InfoSphere DataStage** is an **ETL tool** and part of the **IBM Information Platforms Solutions suite and IBM InfoSphere.** It uses a **graphical notation** to construct **data integration solutions** and is available in various versions such as the **Server** Edition, the **Enterprise** Edition, and the **MVS Edition** outputs links by reason of a write failure or expression evaluation failure. Data stage allows **scheduling**, **monitoring** and **running** the jobs

**Infosphere** is a neologism composed of information and sphere. The word refers to an environment, like a biosphere, that is populated by informational entities called inforgs. While an example of the sphere of information is cyberspace, infospheres are not limited to purely online environments.

Using the AutoSys **Job Information Language** (**JIL**) through a command-line interface.

**AutoSys** is an automated job control system for scheduling, monitoring, and reporting. **AutoSys** is an automated job control system for scheduling, monitoring, and reporting. These jobs can be a UNIX script, java program or any other program which can be invoked from shell. An AutoSys job is any single command, executable, script, or Unix batch file. Each AutoSys job definition contains a variety of qualifying attributes, including the conditions specifying when and where a job should be run.

##### 2.2.3 JIL Sub Commands

When writing a JIL script, you must follow the syntax rules. JIL sub-commands are used to create, modify, override, or delete a job definition. These sub-commands are listed below:

         insert\_job    :  Add a new job to AutoSys.

         update\_job  : Edit fields on an existing job.

         delete\_job   :  Delete an existing job from the AutoSys database.

         box\_name   :  Add a new box job

         delete\_box  :  Delete an existing Box job, and also delete all the jobs which are contained in the box.

         override\_job:  Apply overrides on indicated job attributes for the next run. It change the behavior of a job for the next time the job runs

**Oracle 11i** refers to the **Oracle** ERP Application Suite (often called**Oracle** E-Business Suite)

**Enterprise resource planning** **(ERP**) is business management software—typically a suite of integrated applications—that a company can use to collect, store, manage and interpret data from many business activities, including: Product planning, cost. Manufacturing or service delivery. Marketing and sales.

## A relational database management system (RDBMS) is a program that lets you create, update, and administer a relational database.A relational database management system (RDBMS) is a database management system (DBMS) that is based on the relational modelMany popular databases MS SQL Server, IBM DB2, Oracle, MySQL, and Microsoft Access.currently in use are based on the relational database model.

# IBM DB2 database software

## DB2 is a family of relational database management system (RDBMS) products from IBM that serve a number of different operating system platform .DB2 product is extended with the support of Object-Oriented features and non-relational structures with XML.

***Teradata*** is a Relational Database Management System (RDBMS) for the world’s largest commercial databases.  Teradata can store data upto Teradata bytes in size. This makes the Teradata as a market leader in data warehousing applications.Through the concept of parallelism, Teradata obtains the ability to manage terabytes of data.

**Siebel Relationship Management Warehouse**(RMW)

**Siebel** Customer Relationship Management (**CRM**)

 The Siebel Data Warehouse is also referred to as the Siebel Relationship Management Warehouse (RMW).

**Siebel CRM Systems, Inc.** was a [software](https://en.wikipedia.org/wiki/Computer_software) company principally engaged in the design, development, marketing, and support of[customer relationship management](https://en.wikipedia.org/wiki/Customer_relationship_management) (CRM) applications.

1. **ERwin** is a popular data modeling tool used by a number of major companies in Omaha and throughout the world. The product is currently owned, developed, and marketed by Computer Associates, a leading software developer.

## Microsoft Office is an [office suite](https://en.wikipedia.org/wiki/Office_suite) of applications, servers and services. contained Microsoft Word, Microsoft Excel etc

## Solaris is a Unix operating system originally developed by Sun Microsystems. It superseded their earlier SunOS in 1993. Oracle Solaris, as it is now known, has been owned by Oracle Corporation since Oracle's acquisition of Sun in January 2010.

## IBM AIX Power System software - Offers the highest level of performance, security and reliability of any UNIX operating system

1. **HP**-**UX** (Hewlett-Packard **UniX**) is Hewlett-Packard's proprietary implementation of the **Unix** operating system, based on **UNIX** System V (initially System III) and first released in 1984.

## CA-7 is a job scheduling / workflow automation software package sold by CA Technologies (formerly CA, Inc. and Computer Associates International, Inc.). It is commonly used by banks and other large enterprises with IBM mainframe IT computing platforms. http://www.ca.com/us/~/media/Images/Hero-Images/hero-wla-gov.png

## Data Mining is an analytic process designed to explore data (usually large amounts of data - typically business or market related - also known as "big data") in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications. The process of data mining consists of three stages: (1) the initial exploration, (2) model building or pattern identification with validation/verification, and (3) deployment (i.e., the application of the model to new data in order to generate predictions). http://documents.software.dell.com/Statistics/Textbook/Data-Mining-Techniques

## Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

## command line interface (CLI)

## the dsjob command provided by IBM InfoSphere DataStage to invoke and monitor jobs manually and shows how to use it in automation scripts. A sample automation script will help you to use the dsjob command to run a job, view or reset job status, stop the job, and also retrieve the log for the latest job run.

## dsjobs" can be automated by using Shell scripts in UNIX system

## We can call Datastage Batch Job from Command prompt using 'dsjob'. We can also pass all the parameters from command prompt.

## Then call this shell script in any of the market available schedulers.

## The 2nd option is schedule these jobs using Data Stage director.

## Run your Datastage jobs in batch process through command line

## dsjob -server IP:Port -user username -password passwd -run -jobstatus -mode RESET Project\_name Job\_name

## Irrespective what scheduler you use (AutoSys, SeeBeyond, ControlM, at, cron, to name a few) use the command line interface dsjob to specify what you want DataStage to do. The dsjob command is doc .

## Unicenter Autosys Job Management is a workload automation ou use the command line interface ( dsjob ) from CA Unicenter to start your DataStage jobs. Use the -jobstatus option to have it report the job status

## PL/SQL (Procedural Language/Structured Query Language) is Oracle Corporation's procedural extension for SQL and the Oracle relational database. PL/SQL is available in Oracle Database

## LOOKUP use RAM to perform operations where as JOIN use buffer to perform operations

## Open Database Connectivity API (ODBC)

## Change Data Capture(cdc)

## The FTP Enterprise stage transfers multiple files in parallel. This stage invokes an FTP client and transfers the files to and from a remote host.

## <https://www-01.ibm.com/support/knowledgecenter/SSZJPZ_11.3.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/c_deeref_FTP_Enterprise_Stage.html>

## A container, as its name indicates, is used to group stages and links.

## Local containers. These are created within a job and are only accessible by that job

## Shared containers. These are created separately and are stored in the Repository in the same way that jobs are. There are two types of shared container: (1.) Server shared container. Used in server jobs (can also be used in parallel jobs). (2.) Parallel shared container. Used in parallel jobs. You can also include server shared containers in parallel jobs as a way of incorporating server job functionality into a parallel stage

## service level agreements (SLAs)

## A Data warehouse is a repository of an organization's electronically stored data. Data warehouses are designed to facilitate reporting and analysis.

## Omni-Patient is an enterprise master data application that combines an enterprise master patient index (EMPI) with pre-packaged models to provide full patient identity management, and easily achieve a 360-degree view of key entities, with a single golden record for each patient, provider, payer, workforce and facility.

### CA Workload Automation

* Simplify workload scheduling and management to lower costs.
* Manage batch workloads across diverse computing platforms and technologies
* Seamlessly integrate with common business applications
* Optimize resources across physical, virtual and cloud environments
* Improve service level compliance via critical path analysis and forecasting
* Avoid errors with simulation and forecasting
* Reduce learning curve with streamlined user interfaces and web enabled access

Reducing the cost and complexities of managing application workloads across physical, virtualized and cloud environments. You can do it all with **Workload Automation**solutions from CA Technologies.

## CA-7 is a [job scheduling](https://en.wikipedia.org/wiki/Job_scheduling) / [workflow automation](https://en.wikipedia.org/wiki/Workflow_automation) software package sold by [CA Technologies](https://en.wikipedia.org/wiki/CA_Technologies) (formerly CA, Inc. and Computer Associates International, Inc.).[[1]](https://en.wikipedia.org/wiki/CA-7_(software)#cite_note-sysprog-guide-1) It is commonly used by banks[[2]](https://en.wikipedia.org/wiki/CA-7_(software)" \l "cite_note-arthur-2012-06-25-2) and other large enterprises with [IBM mainframe](https://en.wikipedia.org/wiki/IBM_mainframe) [IT](https://en.wikipedia.org/wiki/Information_technology) [computing platforms](https://en.wikipedia.org/wiki/Computing_platform).

1. A **shell script** is a computer program designed to be run by the Unix**shell**, a command line interpreter. The various dialects of **shell scripts**are considered to be **scripting** languages.

## In Unix, the shell is a program that interprets commands and acts as an intermediary between the user and the inner workings of the operating system

## IBM-AIX, Sun Solaris, HP-UX:versions of unix;

# TOAD - Tools for Oracle Application Development

## Toad is a [software application](https://en.wikipedia.org/wiki/Software_application) from [Dell Software](https://en.wikipedia.org/wiki/Dell_Software) that [database](https://en.wikipedia.org/wiki/Database) developers, database administrators and data analysts use to manage both relational and non-relational databases using [SQL](https://en.wikipedia.org/wiki/SQL).[[](https://en.wikipedia.org/wiki/Toad_(software)#cite_note-1)

1. In computing, Oracle **SQL Developer** (internally often: "**sqldeveloper**") is an Integrated development environment (IDE) for working with **SQL** in Oracle databases. Oracle Corporation provides this product free; it uses the Java Development Kit.

## Quality Center is quality management software

## starTeam is a single source of truth for managing change throughout the software development life-cycle, regardless of methodology

## StarTeam's software configuration management manages and tracks source code changes – including change requests, defects, tasks, requirements

## Xsd-xml source

## Xml extensible markup language

# *WebSphere MQ File Transfer Edition (wmqfte)*

 wmqfte.properties file specifies the name of your default set of configuration options

1. In software development, user acceptance testing (**UAT**) - also called beta testing, application testing, and end user testing - is a phase of software development in which the software is tested in the "real world" by the intended audience.

## A delta load, by definition, is loading incremental changes to the data. When doing a delta load to a fact table, for example, you perform inserts only... appending the change data to the existing table

| *able 1. Supported LOBs and InfoSphere DataStage data types for them* | | |
| --- | --- | --- |
| **LOB** | **Description** | **InfoSphere DataStagedata type** |
| Binary LOB (BLOB) | Represents a LOB in binary form | LongVarBinary |
| Character LOB (CLOB) | Represents a LOB that is held in character form. The database defines the character set. | LongVarChar |
| NLS Character LOB (NCLOB) | Represents a LOB that is held in the national language set (NLS) for the database. Used for multibyte character sets. |  |

1. enterprise data warehouse
2. In computing, a data warehouse (DW or DWH), also known as an **enterprise data warehouse**(EDW), is a system used for reporting and data analysis. DWs are central repositories of integrated data from one or more disparate sources.

|  |  |
| --- | --- |
| Datastage7.5 | Datastage8.1 |
| 1) in ds 7.5.2we have manager as client. | 1)in 8.0.1 we dont have any manager client. the manager client is embeded in designer client. |
| 2)in 7.5.2 quality stage has seperate designer | 2)in 8.0.1 quality stage is integrated in designer. |
| 3)in 7.5.2 code and metadata is stored in file based system. | 3)in 8.0.1 code is a file based system where as metadata is stored in database |
| 4)in 7.5.2 we required operating system authentications. | 4)in8.0.1 we requiree operating system authentications and datastage authentications |
| 5)in 7..5.2 a single join stage can't support multiple references. | 5)in 8.0.1 a single join stage can support multiple references. |
| 6)in 7.5.2 we dont have range lookup. | 6)in8.0.1 we have range lookup. and scd stage |
| 7)in 7.5.2 , when a developer opens a particular job, and another developer wants to open the same job , that job can't be opend. | 7)in 8.0. it can be possible when a developer opens a particular job and another cdeveloper wants to open the same job then it can be opend as read only job |
| 8.) 7.5.2 it is not possible | 8)in 8.0.1 a compare utility is avilable to compare 2 jobs , one in development another is in production. |
| 9) in 7.5.2 not avilable | 9)in 8.0.1 quick find and advance find features are avilable |
| 10)in 7.5.2 first time one job is run and surogate key s generated from initial to n value. next time the same job is compile and run again surrogate key is generated from initial to n. automatic increment of surrogate key is not in 7.5.2 | 10)but in 8.0.1 surrogate key is incremented automatically.a state file is used to store the maximum value of surrogate key. |
| 11)In 7.5 there is no parameterset option available where as | 11)in 8.0.1 version onwards there was parameterset . |

## The main difference is that DataStage has gone from being a standalone ETL tool to part of the Information Server. So you cannot install DataStage without including the Metadata Server and components that go with it.

## Input: Marketing\_person    City-1         city-2       city-3 James Bond         Washington     Newyork      Newjersey Output will be Marketing\_person     City James Bond         Washington James Bond          New York James Bond           New Jersey This achieved  with the new system  variable @ ITERATION

**WHAT ARE THE CLIENT COMPONENTS IN DATASTAGE 7.5X2 VERSION**

In Datastage 7.5X2 Version, they are 4 client Components. They are  
  
1) Datastage Designer  
2) Datastage Director  
3) Datastage Manager  
4) Datastage Admin  
 **In Datastage Designer, We**  
Create the Jobs  
Compile the Jobs  
Run the Jobs  
  
**In Director, We can**  
View the Jobs  
View the Logs  
Batch Jobs  
Unlock Jobs  
Scheduling Jobs  
Monitor the JOBS  
Message Handling  
  
  
**In Manager , We can**  
Import & Export the Jobs  
Node Configuration  
  
**And by using Admin , We can**  
Create the Projects  
Organize the Projects  
Delete the Projects

## DataStage 8.1 to DataStage 8.5

## 1. DataStage Designer performance improvement By changing the Metadata algorythm, copy/delete/save jobs got faster about 30-40%.

## 2. Parallel Engine Performance and Resource improvements Resource usage is about 5% smaller than 8.1, for T-Sort, Windows desktop heap size has been decreased 94%.

## 3. Transformer enhancements Key break support

## LastRowInGroup() function is added. This will return true for the last record of the group.

## LastRow() will return the last record of input.

## Output looping :: Allows multiple output records to be created per single input record.

## Input looping :: Allows aggregation of input records so that aggregated data can be included with the original input data. ( like adding average column to the original input is now possible. ( 2 pass.... calculation. )

## New Null handling This is pretty complicate and need more verification by myself to explain clearly. But this is the description I got.

## Null values can now be included in any expression.

## -> Null values no longer need to be explicitly handled.

## A null value in an expression will return a null value result. As long as the target column is nullable, records will not be dropped. Stage variables are now always nullable.

## APT\_TRANSFORM\_COMPILE\_OLD\_NULL\_HANDLING is prepared to support backward compatibility.

## New Transformer Functions

## Various packed decimal conversions

## Not much change in comparison between 8.7 and 9.1 except in 9.1 we have Balanced Optimization for Hadoop.

## New Transformation Expressions has been added. EREPLACE :  Function to replace substring in expression with another substring. If not specified occurrence, then each occurrence of substring will be replaced.

**Server components**

**DS server:** runs executable server jobs, under the control of the DS director, that extract,transform, and load data into a DWH.

**DS Package installer:** A user interface used to install packaged DS jobs and plug-in;

**Repository or project:** a central store that contains all the information required to build DWH or data mart.

1. xmeta :will have information about the project and DataStage software

With the recent versions of Datastage (7.5, 8, 8.1, 8.5, 9.1), IBM does not release any updates to Datastage Server Edition (however it is still available in Datastage 8)

**OSH (Orchestrate Shell script language)**.

## Differences between Datastage Enterprise and Server Edition

1. The major difference between Infosphere Datastage Enterprise and Server edition is that Enterprise Edition (EE) introduces **Parallel jobs**. Parallel jobs support a completely new set of stages, which implement the scalable and parallel data processing mechanisms. In most cases parallel jobs and stages look similiar to the Datastage Server objects, however their capababilities are way different.   
   In rough outline:
   * Parallel jobs are executable datastage programs, managed and controlled by Datastage Server runtime environment
   * Parallel jobs have a built-in mechanism for Pipelining, Partitioning and Parallelism. In most cases no manual intervention is needed to implement optimally those techniques.
   * Parallel jobs are a lot faster in such ETL tasks like sorting, filtering, aggregating
2. Datastage EE jobs are compiled into **OSH (Orchestrate Shell script language)**.  
   OSH executes operators - instances of executable C++ classes, pre-built components representing stages used in Datastage jobs.   
   Server Jobs are compiled into Basic which is an interpreted pseudo-code. This is why parallel jobs run faster, even if processed on one CPU.
3. Datastage Enterprise Edition adds functionality to the traditional server stages, for instance record and column level format properties.
4. Datastage EE brings also completely new stages implementing the parallel concept, for example:
   * Enterprise Database Connectors for Oracle, Teradata & DB2
   * Development and Debug stages - Peek, Column Generator, Row Generator, Head, Tail, Sample ...
   * Data set, File set, Complex flat file, Lookup File Set ...
   * Join, Merge, Funnel, Copy, Modify, Remove Duplicates ...
5. When processing large data volumes Datastage EE jobs would be the right choice, however when dealing with smaller data environment, using Server jobs might be just easier to develop, understand and manage.   
   When a company has both Server and Enterprise licenses, both types of jobs can be used.
6. **Sequence jobs** are the same in Datastage EE and Server editions.

Datastage Enterprise Edition, formerly known as Datastage PX (parallel extender) has become recently a part of IBM InfoSphere Information Server and its official name is **IBM InfoSphere DataStage**.

<https://www-01.ibm.com/support/knowledgecenter/SSZJPZ_11.3.0/com.ibm.swg.im.iis.productization.iisinfsv.install.doc/topics/wsisinst_pln_engscalabilityparallel.html>

SMP - Symmetric Multi-Processing. In a symmetrical multi-  
processing environment, the CPU's share the same memory,   
and as a result code running in one CPU can affect the   
memory used by another.  
MPP - Massively Parallel Processing. computer system with   
many independent arithmetic units or entire   
microprocessors, that run in parallel.

 Teradata v2r5  
2. Teradata v2r6  
3. Teradata 12  
4. Teradata 13  
5. Teradata 13.10  
6. Teradata 14

## Some of the ETL tools which are commonly used in Teradata are DataStage, Informatica,

## UPSERT basically stands for Update Else Insert. This option is available only in Teradata.

## Fastload has two phases: acquisition phase and application phase. mload has 5 phases(Note: however there is no acquisition phase for mload delete). mload phases: Preliminary,DML Transaction,Acquisition,Application,Cleanup.

**Teradata FastLoad** is a highly reliable, parallel-load utility designed to move large volumes of data — collected from data sources on channel and network-attached clients — into empty tables in the Teradata Database. Execute Teradata FastLoad from any client platform, mainframe or load server — for automatic parallel data allocation, conversion, movement and loading that achieves higher performance levels.

**Teradata MultiLoad** is a time-tested, highly reliable, parallel-load utility used to create and maintain Teradata Databases. MultiLoad optimizes operations that rapidly acquire, process and apply data to tables in a Teradata Database. For data maintenance, MultiLoad updates, inserts, upserts and deletes large volumes of data in empty or populated tables and works at the data block level for higher performance. MultiLoad runs on a variety of client platforms, operates in a fail-safe mode and is fully recoverable.

Fast Load: to load empty tables at high speed.

MLoad: to insert,update and delete up to 5 different tables or views

**Load, update and delete large tables in Teradata in a bulk mode**

Fastload is used to load empty tables and is very fast, can

load one table at a time.

Multiload can load at max 5 tbls at a time and can also

update and delete the data.

Fastload can be used only for inserting data, not updating

and deleting.

## In terms of speed, fastload would be suitable than mload for loading an empty table. But when it comes to multiset tables, mload would be a better option.  Mload allows use of aggregated values whereas fastload does not.

## Teradata RDBMS is a complete relational database management system. The system is based on off-the-shelf Symmetric Multiprocessing (SMP) technology combined with a communication network connecting the SMP systems to form a Massively Parallel Processing (MMP) system

## Oracle database is a collection of data treated as a unit. The purpose of a database is to store and retrieve related information

**What's the difference between a primary key and a unique key?**

Both primary key and unique key enforces uniqueness of the column on which they are defined. But by default primary key creates a clustered index on the column, where are unique creates a nonclustered index by default. Another major difference is that, primary key doesn't allow NULLs, but unique key allows one NULL only.

## Database version history

## The major Oracle versions, with their latest patch-sets are:

## [Oracle 7](http://www.orafaq.com/wiki/Oracle_7): 7.3.4.5

## [Oracle 8](http://www.orafaq.com/wiki/Oracle_8): 8.0.3 - 8.0.6

## [Oracle 8i](http://www.orafaq.com/wiki/Oracle_8i): 8.1.5.0 - 8.1.7.4

## [Oracle 9i](http://www.orafaq.com/wiki/Oracle_9i) (Release 1): 9.0.1.0 - 9.0.1.4

## [Oracle 9i](http://www.orafaq.com/wiki/Oracle_9i) (Release 2): 9.2.0.1 - 9.2.0.8

## [Oracle 10g](http://www.orafaq.com/wiki/Oracle_10g) (Release 1): 10.1.0.2 - 10.1.0.5

## [Oracle 10g](http://www.orafaq.com/wiki/Oracle_10g) (Release 2): 10.2.0.1 - 10.2.0.5

## [Oracle 11g](http://www.orafaq.com/wiki/Oracle_11g) (Release 1): 11.1.0.6 - 11.1.0.7

## [Oracle 11g](http://www.orafaq.com/wiki/Oracle_11g) (Release 2): 11.2.0.1 - 11.2.0.3

## [Oracle 12c](http://www.orafaq.com/wiki/Oracle_12c) (Release 1): 12.1.0.1 - 12.1.0.2.0

## Microsoft SQL Server Internal Database Versions

|  |  |  |
| --- | --- | --- |
| [SQL Server 2014](http://sqlserverbuilds.blogspot.com/#sql2014) |  |  |
| [SQL Server 2012](http://sqlserverbuilds.blogspot.com/#sql2012) |  |  |
| SQL Server 2012 CTP1 (a.k.a. SQL Server 2011 Denali) |  |  |
| [SQL Server 2008 R2](http://sqlserverbuilds.blogspot.com/#sql2008r2) |  |  |
| [SQL Server 2008](http://sqlserverbuilds.blogspot.com/#sql2008) |  |  |
| [SQL Server 2005 SP2+ with VarDecimal enabled](http://sqlserverbuilds.blogspot.com/#sql2005) |  |  |
| [SQL Server 2005](http://sqlserverbuilds.blogspot.com/#sql2005) |  |  |
| [SQL Server 2000](http://sqlserverbuilds.blogspot.com/#sql2000) |  |  |
| [SQL Server 7.0](http://sqlserverbuilds.blogspot.com/#sql7) |  |  |
| [SQL Server 6.5](http://sqlserverbuilds.blogspot.com/2006/05/sql-server-6-5.html) |  |  |
| [SQL Server 6.0](http://sqlserverbuilds.blogspot.com/2006/01/sql-server-6.html) |  |  |

## Db2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Version 10.5 |  |  |  |  |
| Version 10.1 |  |  |  |  |
|  | | | | | |
| Version 9.8 |  |  |  |  |  |
| Version 9.7 |  |  |  |  |  |
| Version 9.5 |  |  |  |  |  |
| Version 9.1 |  |  |  |  |  |
|  | | | | | |
| Version 8.1 and 8.2 |  |  |  |  |  |
|  | | | | | |
| Version 7.2 |  |  |  |  |  |
| Version 7.1 |  |  |  |  |  |
| Version 7.1 |  |  |  |  |  |

* Sybase SQL Server 3.0 – the first publicly released version of the product released in 1988.
* Sybase SQL Server 4.0 – released in 1990.
* Sybase SQL Server 4.2 – released in 1991.
* Sybase SQL Server 4.8 – released in 1992.
* Sybase SQL Server 4.9 – also released in 1992, along with version 4.9.1; version 4.9.2 was released in 1993.
* Sybase System X, aka *Sybase SQL Server 10.0* – released in 1993.
* Sybase SQL Server 11.0 – released in 1995.
* Adaptive Server Enterprise 11.5 – released in 1997.
* Adaptive Server Enterprise 11.9 – released in 1998. It does not appear that version 11.9.0 was ever publicly released; instead 11.9.2 was the released version, and 11.9.3 was the first 64 bit version made available.
* Adaptive Server Enterprise 12.0 – released in 1999.
* Adaptive Server Enterprise 12.5 – released in 2001.
* Adaptive Server Enterprise 12.5.0.1 – released in 2002, along with version 12.5.0.2.
* Adaptive Server Enterprise 12.5.0.3 – released in 2003, along with version 12.5.1.
* Adaptive Server Enterprise 12.5.2 – released in 2004, along with version 12.5.3.
* Adaptive Server Enterprise 12.5.3a – released in 2004/2005.
* Adaptive Server Enterprise 12.5.4 – released in 2006.
* Adaptive Server Enterprise 15.0 – released in 2005, along with version 12.5.3.
* Adaptive Server Enterprise 15.0.1 – released in 2006.
* Adaptive Server Enterprise 15.0.2 – due for release in 2007.

## Win·dows

## a computer operating system with a graphical user interface.

## U·nix

## a widely used multiuser operating system.

## Information technology (IT) is the application of [computers](https://en.wikipedia.org/wiki/Computer) and [telecommunications equipment](https://en.wikipedia.org/wiki/Telecommunications_equipment) to store, retrieve, transmit and manipulate data,[[1]](https://en.wikipedia.org/wiki/Information_technology" \l "cite_note-DOP-1) often in the context of a business or other enterprise

difference between varchar and varchar2

 VARCHAR is going to be replaced by VARCHAR2 in next version. So, Oracle suggests the use VARCHAR2 instead of VARCHAR while declaring datatype.

2. VARCHAR can store up to 2000 bytes of characters while VARCHAR2 can store up to 4000 bytes of characters.

3. If we declare datatype as VARCHAR then it will occupy space for NULL values, In case of VARCHAR2 datatype it will not occupy any space.

## *WebSphere is a suite of* [software](https://en.wikipedia.org/wiki/Computer_software) products

## Some of the Performance Tuning techniques

## Select suitable configurations file (nodes depending on data volume)     2. Select buffer memory correctly   3. Select proper partition   4. Turn off Runtime Column propagation wherever it’s not required   5. Taking care about sorting of the data.  6. Handling null values (use modify instead of transformer)    7. Try to decrease the use of transformer. (Use copy, filter, modify)    8. Use dataset instead of sequential file in the middle of the vast jobs   9. Take maximum 20 stages for a job for best performance.  10. Select Join or Lookup or Merge (depending on data volume)

## 11.Use parametrized jobs and sequence with the proper parallel/sequential mapping flow.

## 12. Stop propagation of unnecessary metadata between the stages

1. A **stored procedure** is a group of SQL statements that form a logical unit and perform a particular task, and they are used to encapsulate a set of operations or queries to execute on a database server.

## Service-Oriented Architecture (SOA)

1. **Unit testing** is a software development process in which the smallest testable parts of an application, called units, are individually and independently scrutinized for proper operation. **Unit testing** is often automated but it can also be done manually.

## A test case is a document, which has a set of test data, preconditions, expected results and postconditions, developed for a particular test scenario in order to verify compliance against a specific requirement.

## sCD Type 1

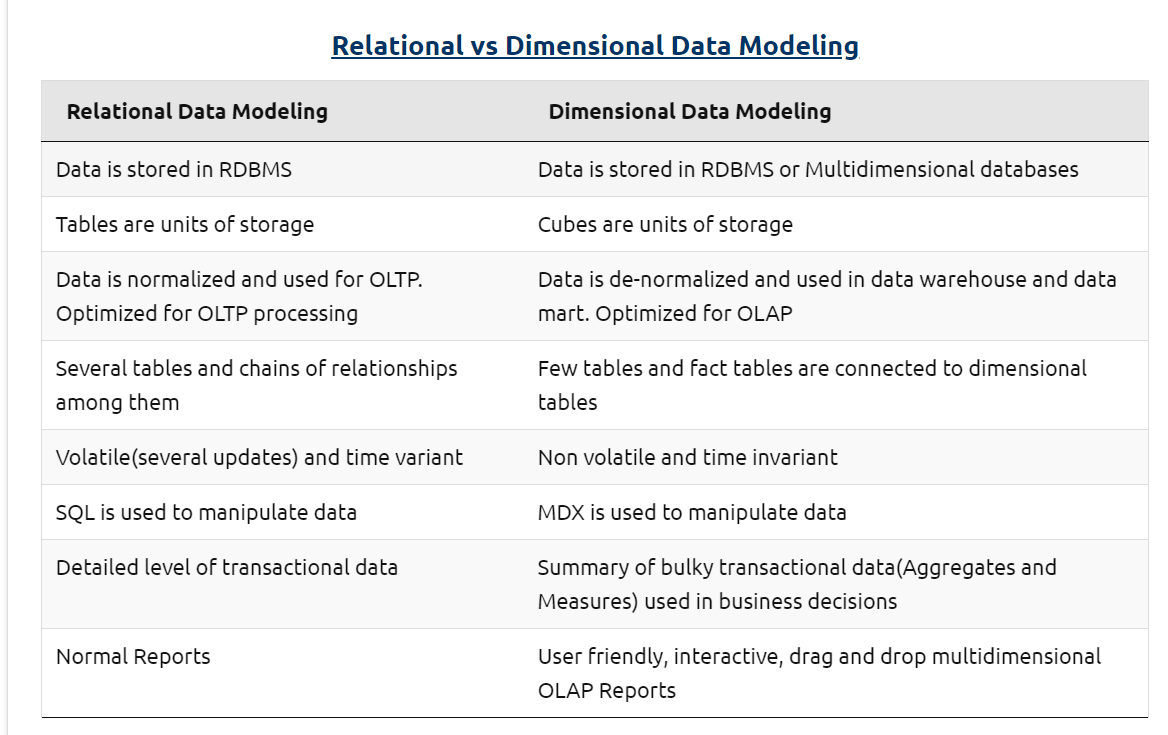
## Type 1 Slowly Changing Dimension in which no history is kept in the database.

## Slowly changing dimension Type 2 is a model where the whole history is stored in the database.

## Business Intelligence is the process of collecting raw data or business data and turning it into information that is useful and more meaningful.

What is the difference between data mining and data warehousing?

Data warehousing comes before the mining process. This is the act of gathering data from various exterior sources and organizing it into one specific location: the warehouse. Data mining is when that data is analyzed and used as information for making decisions.



* What is data mart?
* A **data mart** is the access layer of the data warehouse environment that is used to get data out to the users

Data mart vs data warehouse[[edit](https://en.wikipedia.org/w/index.php?title=Data_mart&action=edit&section=1" \o "Edit section: Data mart vs data warehouse)]

|  |  |
| --- | --- |
| Data warehouse: | * datamart |
| Holds multiple subject areas  1) | * Often holds only one subject area- for example, Finance, or Sales |
| * Holds very detailed information * 2) | * May hold more summarized data (although many hold full detail) |
| * . * Works to integrate all data sources * 3) | * Concentrates on integrating information from a given subject area or set of source systems * 3) |
| * 4)Does not necessarily use a [dimensional model](https://en.wikipedia.org/wiki/Dimensional_modeling) but feeds dimensional models | * 4)Is built focused on a dimensional model using a star schema |
| * 5)100 GB-TB+size | * <100gb * 5) |

## What are OLTP and OLAP

* What is Master data management?

## Master data management (MDM) is a comprehensive method of enabling an enterprise to link all of its critical data to one file, called amaster file, that provides a common point of reference.

* **SELECT** 1+1
* **FROM** DUAL;
* **SELECT** 1
* **FROM** DUAL;
* **SELECT** **USER**
* **FROM** DUAL;
* **SELECT** SYSDATE
* **FROM** DUAL;
* **SELECT** \*
* **FROM** DUAL

## *The DUAL table is a pseudo table, not a real table. The DUAL table has only one column named DUMMY*

## SELECT COUNT(\*) FROM DUAL;

1. **DELETE** **FROM** DUAL;
2. **TRUNCATE** **TABLE** DUAL

Dual is a table that is created by Oracle together with data dictionary. It consists of exactly one column named “dummy”, and one record. The value of that record is X.

Dual is a virtual table ...it do not eists .

It is used to query fro pseudo columns..

Select NVL(max(s\_key),0) as maxskey,1 as dummy from table;

## Control table perform a small verification before running job, Control table will help avoid running query number of times when we are expecting to run single time in a day. Control table contains fields like JOBID,JOB\_NAME,PREVIESE\_RUN\_DATE,

## <http://pr3systems.com/blog/information-management/load-multiple-database-tables-with-a-single-connector-stage/>

## When designing a job, you can specify a subroutine to run before or after the job, or before or after an active stage. Server routines are stored in the repository, where you can create, view, or edit them using the Routine dialog box

## Before/After subroutines. When designing a job, you can specify a subroutine to run before or after the job, or before or after an active stage.

## IBM InfoSphere Streams is an advanced analytic platform that allows user-developed applications to quickly ingest, analyze and correlate information as it arrives from thousands of real-time sources. The solution can handle very high data throughput rates, up to millions of events or messages per second.

## <http://www.slideshare.net/IBMInfoSphereUGFR/infosphere-streams-technical-overview-use-cases-big-data-jerome-chailloux>

## Early Arriving Fact

## <https://www.ateam-oracle.com/implementing-early-arriving-facts-in-odi-part-i-proof-of-concept-overview>