**Informatica**

**ETL LIFE CYCLE …**

**The Typical real-life ETL cycle consists of the following execution steps:**

1. Cycle initiation

2. Build reference data

3. Extract (from sources)

4. Validate

5. Transform (clean, apply business rules, check for data integrity, create aggregates or disaggregates)

6. Stage (load into staging tables, if used)

7. Audit reports (for example, on compliance with business rules. Also, in case of failure, helps to diagnose/repair)

8. Publish (to target tables)

9. Archive

10. Clean up

**What is Informatica?**

Informatica Power Center is a powerful ETL tool from Informatica Corporation.

**Informatica Corporation products are:**

* Informatica Power Center
* Informatica on demand
* Informatica B2B Data Exchange
* Informatica Data Quality
* Informatica Data Explorer

Informatica Power Center is a single, unified enterprise data integration platform for accessing, discovering, and integrating data from virtually any business system, in any format, and delivering that data throughout the enterprise at any speed.

**Informatica Power Center Editions:**

Because every data integration project is different and includes many variables such as data volumes, latency requirements, IT infrastructure, and methodologies—Informatica offers three Power Center Editions and a suite of Power Center Options to meet your project’s and organization’s specific needs.

* Standard Edition
* Real Time Edition
* Advanced Edition

**Informatica Power Center Standard Edition:**

Power Center Standard Edition is a single, unified enterprise data integration platform for discovering, accessing, and integrating data from virtually any business system, in any format, and delivering that data throughout the enterprise to improve operational efficiency.

Key features include:

* A high-performance data integration server
* A global metadata infrastructure
* Visual tools for development and centralized administration
* Productivity tools to facilitate collaboration among architects, analysts, and developers. [clip_image024](http://lh4.ggpht.com/_MbhSjEtmzI8/Ta7hkhSHWYI/AAAAAAAAAQs/0k5akyjIkdY/s1600-h/clip_image024%5b3%5d.gif)

**Informatica Power Center Real Time Edition:**

Packaged for simplicity and flexibility, Power Center Real Time Edition extends Power Center Standard Edition with additional capabilities for integrating and provisioning transactional or operational data in real-time. Power Center Real Time Edition provides the ideal platform for developing sophisticated data services and delivering timely information as a service, to support all business needs. It provides the perfect real-time data integration complement to service-oriented architectures, application integration approaches, such as enterprise application integration (EAI), enterprise service buses (ESB), and business process management (BPM).

**Key features include:**

* Change data capture for relational data sources
* Integration with messaging systems
* Built-in support for Web services
* Dynamic partitioning with data smart parallelism
* Process orchestration and human workflow capabilities

**Informatica Power Center Real Time Edition:**

Power Center Advanced Edition addresses requirements for organizations that are standardizing data integration at an enterprise level, across a number of projects and departments. It combines all the capabilities of Power Center Standard Edition and features additional capabilities that are ideal for data governance and Integration Competency Centers.

**Key features include:**

* Dynamic partitioning with data smart parallelism
* Powerful metadata analysis capabilities
* Web-based data profiling and reporting capabilities

Power Center includes the following components:

* Power Center domain
* Administration Console
* Power Center repository
* Power Center Client
* Repository Service
* Integration Service
* Web Services Hub
* SAP BW Service
* Data Analyzer
* Metadata Manager
* Power Center Repository Reports

The Power Center Client consists of the following applications that we use to manage the repository, design mappings, Mapplets, and create sessions to load the data:

1. Designer
2. Data Stencil
3. Repository Manager
4. Workflow Manager
5. Workflow Monitor

**1. Designer:**

Use the Designer to create mappings that contain transformation instructions for the Integration Service.

The Designer has the following tools that you use to analyze sources, design target Schemas, and build source-to-target mappings:

* **Source Analyzer**: Import or create source definitions.
* **Target Designer:** Import or create target definitions.
* **Transformation Developer:** Develop transformations to use in mappings.

You can also develop user-defined functions to use in expressions.

* **Mapplet Designer**: Create sets of transformations to use in mappings.
* **Mapping Designer:** Create mappings that the Integration Service uses to Extract, transform, and load data.

**2. Data Stencil**

Use the Data Stencil to create mapping template that can be used to generate multiple mappings. Data Stencil uses the Microsoft Office Visio interface to create mapping templates. Not used by a developer usually.

**3. Repository Manager**

Use the Repository Manager to administer repositories. You can navigate through multiple folders and repositories, and complete the following tasks:

* **Manage users and groups:** Create, edit, and delete repository users and User groups. We can assign and revoke repository privileges and folder Permissions.
* **Perform folder functions: Create**, edit, copy, and delete folders. Work we perform in the Designer and Workflow Manager is stored in folders. If we want to share metadata, you can configure a folder to be shared.
* **View metadata:** Analyze sources, targets, mappings, and shortcut dependencies, search by keyword, and view the properties of repository Objects. We create repository objects using the Designer and Workflow Manager Client tools.

We can view the following objects in the Navigator window of the Repository Manager:

* **Source definitions:** Definitions of database objects (tables, views, synonyms) or Files that provide source data.
* **Target definitions:** Definitions of database objects or files that contain the target data.
* **Mappings:** A set of source and target definitions along with transformations containing business logic that you build into the transformation. These are the instructions that the Integration Service uses to transform and move data.
* **Reusable transformations:** Transformations that we use in multiple mappings.
* **Mapplets:** A set of transformations that you use in multiple mappings.
* **Sessions and workflows:** Sessions and workflows store information about how and When the Integration Service moves data. A workflow is a set of instructions that Describes how and when to run tasks related to extracting, transforming, and loading Data. A session is a type of task that you can put in a workflow. Each session Corresponds to a single mapping.

**4. Workflow Manager:**

Use the Workflow Manager to create, schedule, and run workflows. A workflow is a set of instructions that describes how and when to run tasks related to extracting, transforming, and loading data.

The Workflow Manager has the following tools to help us develop a workflow:

* **Task Developer:** Create tasks we want to accomplish in the workflow.
* **Work let Designer**: Create a worklet in the Worklet Designer. A worklet is an object that groups a set of tasks. A worklet is similar to a workflow, but without scheduling information. We can nest worklets inside a workflow.
* **Workflow Designer:** Create a workflow by connecting tasks with links in the Workflow Designer. You can also create tasks in the Workflow Designer as you develop the workflow.

When we create a workflow in the Workflow Designer, we add tasks to the workflow. The Workflow Manager includes tasks, such as the Session task, the Command task, and the Email task so you can design a workflow. The Session task is based on a mapping we build in the Designer.

We then connect tasks with links to specify the order of execution for the tasks we created. Use conditional links and workflow variables to create branches in the workflow.

**5. Workflow Monitor**

Use the Workflow Monitor to monitor scheduled and running workflows for each Integration Service. We can view details about a workflow or task in Gantt chart view or Task view. We Can run, stop, abort, and resume workflows from the Workflow Monitor. We can view Sessions and workflow log events in the Workflow Monitor Log Viewer.

The Workflow Monitor displays workflows that have run at least once. The Workflow Monitor continuously receives information from the Integration Service and Repository Service. It also fetches information from the repository to display historic Information.

**Informatica interview questions:**

1. **E**xplain your Project?
2. **W**hat are your Daily routines?
3. **H**ow many mapping have you created all together in your project?
4. **I**n which account does your Project Fall?
5. **W**hat is your Reporting Hierarchy?
6. **H**ow many Complex Mapping’s have you created? Could you please me the situation for which you have developed that Complex mapping?
7. What is your Involvement in Performance tuning of your Project?
8. What is the Schema of your Project? And why did you opt for that particular schema?
9. What are your Roles in this project?
10. Can I have one situation which you have adopted by which performance has improved dramatically?
11. Where you Involve in more than two projects simultaneously?
12. Do you have any experience in the Production support?
13. What kinds of Testing have you done on your Project (Unit or Integration or System or UAT)? And Enhancements were done after testing?
14. How many Dimension table are there in your Project and how are they linked to the fact table?
15. How do we do the Fact Load?
16. How did you implement CDC in your project?
17. How does your Mapping in File to Load look like?
18. How does your Mapping in Load to Stage look like?
19. How does your Mapping in Stage to ODS look like?
20. What is the size of your Data warehouse?
21. What is your Daily feed size and weekly feed size?
22. Which Approach (Top down or Bottom Up) was used in building your project?
23. How do you access your source’s (are they Flat files or Relational)?
24. Have you developed any Stored Procedure or triggers in this project? How did you use them and in which situation?
25. Did your Project go live? What are the issues that you have faced while moving your project from the Test Environment to the Production Environment?
26. What is the biggest Challenge that you encountered in this project?
27. What is the scheduler tool you have used in this project? How did you schedule jobs using it?

**Data Warehouse Interview Questions:**

**1. W**hat is a data-warehouse?

**2. W**hat are Data Marts?

**3. W**hat is ER Diagram?

**4. W**hat is a Star Schema?

**5. W**hat is Dimensional Modelling?

**6. W**hat Snow Flake Schema?

**7. W**hat are the Different methods of loading Dimension tables?

**8. W**hat are Aggregate tables?

**9. W**hat is the Difference between OLTP and OLAP?

**10. W**hat is ETL?

**11. W**hat are the various ETL tools in the Market?

**12. W**hat are the various Reporting tools in the Market?

**13. W**hat is Fact table?

**14. W**hat is a dimension table?

**15. W**hat is a lookup table?

**16. W**hat is a general purpose scheduling tool? Name some of them?

**17. W**hat are modeling tools available in the Market? Name some of them?

**18. W**hat is real time data-warehousing?

**19. W**hat is data mining?

**20. W**hat is Normalization? First Normal Form, Second Normal Form, Third Normal Form?

**21. W**hat is ODS?

**22. W**hat type of Indexing mechanism do we need to use for a typical

Data warehouse?

**23. W**hich columns go to the fact table and which columns go the dimension table? (My user needs to see <data element<data element broken by <data element<data element>

All elements before broken = Fact Measures

All elements after broken = Dimension Elements

**24. W**hat is a level of Granularity of a fact table? What does this signify?(Weekly level summarization there is no need to have Invoice Number in the fact table anymore)

**25. H**ow are the Dimension tables designed? De-Normalized, Wide, Short, Use Surrogate Keys, Contain Additional date fields and flags.

**26. W**hat are slowly changing dimensions?

**27. W**hat are non-additive facts? (Inventory, Account balances in bank)

**28. W**hat are conformed dimensions?

**29. W**hat is VLDB? (Database is too large to back up in a time frame then it's a VLDB)

**30. W**hat are SCD1, SCD2 and SCD3?

**New Interview Questions:**

### What are the differences between Connected and Unconnected Lookup?

The differences are illustrated in the below table

|  |  |
| --- | --- |
| **Connected Lookup** | **Unconnected Lookup** |
| Connected lookup participates in dataflow and receives input directly from the pipeline | Unconnected lookup receives input values from the result of a LKP: expression in another transformation |
| Connected lookup can use both dynamic and static cache | Unconnected Lookup cache can NOT be dynamic |
| Connected lookup can return more than one column value ( output port ) | Unconnected Lookup can return only one column value i.e. output port |
| Connected lookup caches all lookup columns | Unconnected lookup caches only the lookup output ports in the lookup conditions and the return port |
| Supports user-defined default values (i.e. value to return when lookup conditions are not satisfied) | Does not support user defined default values |

### What is meant by active and passive transformation?

An active transformation is the one that performs any of the following actions:

1) Change the number of rows between transformation input and output. Example: Filter transformation.

2) Change the transaction boundary by defining commit or rollback points., example transaction control transformation.

3) Change the row type, example Update strategy is active because it flags the rows for insert, delete, update or reject.

On the other hand a passive transformation is the one which does not change the number of rows that pass through it. Example: Expression transformation.

**Active Transformation -**An active transformation changes the number of rows that pass through the mapping.

1. **Source Qualifier Transformation**
2. **Sorter Transformations**
3. **Aggregator Transformations**
4. **Filter Transformation**
5. **Union Transformation**
6. **Joiner Transformation**
7. **Normalizer Transformation**
8. **Rank Transformation**
9. **Router Transformation**
10. **Update Strategy Transformation**
11. **Advanced External Procedure Transformation**

**Passive Transformation - Passive transformations do not change the number of rows that pass through the mapping.**

1. **Expression Transformation**
2. **Sequence Generator Transformation**
3. **Lookup Transformation**
4. **Stored Procedure Transformation**
5. **XML Source Qualifier Transformation**
6. **External Procedure Transformation**
7. **Input Transformation(Mapplet)**
8. **Output Transformation(Mapplet)**

### What is the difference between Router and Filter?

Following differences can be noted,

|  |  |
| --- | --- |
| **Router** | **Filter** |
| Router transformation divides the incoming records into multiple groups based on some condition. Such groups can be mutually inclusive (Different groups may contain same record) | Filter transformation restricts or blocks the incoming record set based on one given condition. |
| Router transformation itself does not block any record. If a certain record does not match any of the routing conditions, the record is routed to default group | Filter transformation does not have a default group. If one record does not match filter condition, the record is blocked |
| Router acts like CASE.. WHEN statement in SQL (Or Switch().. Case statement in C) | Filter acts like WHERE condition is SQL. |

### What can we do to improve the performance of Informatica Aggregator Transformation?

Aggregator performance improves dramatically if records are sorted before passing to the aggregator and "sorted input" option under aggregator properties is checked. The record set should be sorted on those columns that are used in Group By operation.

It is often a good idea to sort the record set in database level ([click here to see why?](http://www.dwbiconcepts.com/etl/14-etl-informatica/28-informatica-oracle-sort-performance-test.html)) e.g. inside a source qualifier transformation, unless there is a chance that already sorted records from source qualifier can again become unsorted before reaching aggregator

You may also read [this article](http://www.dwbiconcepts.com/etl/14-etl-informatica/47-tuning-informatica-aggregator.html) to know how to tune the performance of aggregator transformation

### What are the different lookup cache(s)?

Informatica Lookups can be cached or un-cached (No cache). And Cached lookup can be either static or dynamic. A **static cache** is one which does not modify the cache once it is built and it remains same during the session run. On the other hand, A [**dynamic cache**](http://www.dwbiconcepts.com/etl/14-etl-informatica/16-dynamic-lookup-cache.html)is refreshed during the session run by inserting or updating the records in cache based on the incoming source data. By default, Informatica cache is static cache.

A lookup cache can also be divided as [**persistent** or **non-persistent**](http://www.dwbiconcepts.com/etl/14-etl-informatica/9-implementing-informatica-persistent-cache.html) based on whether Informatica retains the cache even after the completion of session run or deletes it

### How can we update a record in target table without using Update strategy?

A target table can be updated without using 'Update Strategy'. For this, we need to define the key in the target table in Informatica level and then we need to connect the key and the field we want to update in the mapping Target. In the session level, we should set the target property as "Update as Update" and check the "Update" check-box.

Let's assume we have a target table "Customer" with fields as "Customer ID", "Customer Name" and "Customer Address". Suppose we want to update "Customer Address" without an Update Strategy. Then we have to define "Customer ID" as primary key in Informatica level and we will have to connect Customer ID and Customer Address fields in the mapping. If the session properties are set correctly as described above, then the mapping will only update the customer address field for all matching customer IDs.

### Under what condition selecting Sorted Input in aggregator may fail the session?

* If the input data is not sorted correctly, the session will fail.
* Also if the input data is properly sorted, the session may fail if the sort order by ports and the group by ports of the aggregator are not in the same order.

### Why is Sorter an Active Transformation?

This is because we can select the "distinct" option in the sorter property.

When the Sorter transformation is configured to treat output rows as distinct, it assigns all ports as part of the sort key. The Integration Service discards duplicate rows compared during the sort operation. The number of Input Rows will vary as compared with the Output rows and hence it is an Active transformation.

### Is lookup an active or passive transformation?

From Informatica 9x, Lookup transformation can be configured as as "Active" transformation.

Find out [How to configure lookup as active transformation](http://www.dwbiconcepts.com/etl/14-etl-informatica/90-active-lookup-transformation.html)

However, in the older versions of Informatica, lookup is a passive transformation

### What is the difference between Static and Dynamic Lookup Cache?

We can configure a Lookup transformation to cache the underlying lookup table. In case of static or read-only lookup cache the Integration Service caches the lookup table at the beginning of the session and does not update the lookup cache while it processes the Lookup transformation.

In case of dynamic lookup cache the Integration Service dynamically inserts or updates data in the lookup cache and passes the data to the target. The dynamic cache is synchronized with the target.

In case you are wondering why we need to make lookup cache dynamic, read this article on [dynamic lookup](http://www.dwbiconcepts.com/etl/14-etl-informatica/16-dynamic-lookup-cache.html)

### What is the difference between STOP and ABORT options in Workflow Monitor?

When we issue the STOP command on the executing session task, the Integration Service stops reading data from source. It continues processing, writing and committing the data to targets. If the Integration Service cannot finish processing and committing data, we can issue the abort command.

In contrast ABORT command has a timeout period of 60 seconds. If the Integration Service cannot finish processing and committing data within the timeout period, it kills the DTM process and terminates the session.

### What are the new features of Informatica 9.x in developer level?

From a developer's perspective, some of the new features in Informatica 9.x are as follows:

* Now Lookup can be configured as an active transformation - it can return multiple rows on successful match
* Now you can write SQL override on un-cached lookup also. Previously you could do it only on cached lookup
* You can control the size of your session log. In a real-time environment you can control the session log file size or time
* Database deadlock resilience feature - this will ensure that your session does not immediately fail if it encounters any database deadlock, it will now retry the operation again. You can configure number of retry attempts.

### Informatica 9 Fetchers:

These are the new features in informatica 9.

* Informatica 9 supports data integration for the cloud as well as on premise. You can integrate the data in cloud applications, as well as run Informatica 9 on cloud infrastructure.
* Informatica analyst is a new tool available in Informatica 9.
* There is architectural difference in Informatica 9 compared to previous version.
* Browser based tool for business analyst is a new feature.
* Data steward is a new feature.
* Allows unified administration with a new admin console that enables you to manage power center and power exchange from the same console.
* Powerful new capabilities for data quality.
* Single admin console for data quality, power center, power exchange and data services.
* In Informatica 9, Informatica data quality (IDQ) has been further integrated with the Informatica Platform and performance, manageability and reusability have all been significantly enhanced.
* The mappings rules are shared between the browser based tool for analysts and the eclipse based development leveraging unified metadata underneath.
* The data services capabilities in Informatica 9 both over SQL and web services can be used for real time dash boarding.
* Informatica data quality provides worldwide address validation support with integrated geocoding.
* The ability to define rules and view and run profiles is available in both the Informatica developer (Thick client) and Informatica analyst (browser based tool-Thin client).these tools sit on a unified metadata infrastructure. Both tools incorporate security features like authentication and authorization ensuring.
* The developer tool is now eclipse based and supports both data integration and data quality for enhanced productivity. It provides browser based tool for analysts to support the types of tasks they engage in, such as profiling data, specifying and validating rules & monitoring data quality.
* There will a velocity methodology. Soon it’s going to introduce on I9.
* Informatica has the capability to pull data from IMS, DB2 on series and series and from other several other legacy systems (Mainframe) environment like VSAM, Data com, and IDMS etc.
* There are separate tools available for different roles. The Mapping architect for Vision tool is designed for architects and developers to create templates for common data integration patterns saving developer’s tremendous amount of time.
* Informatica 9 does not include ESB infrastructure.
* Informatica supports open interfaces such as web services and can integrate with other tools that support these as well including BPM tool.
* Informatica 9 complements existing BI architectures by providing immediate access to data through data virtualization, which can supplement the data in existing data warehouse and operational data store.
* Informatica 9 supports profiling of Mainframe data. Leveraging the Informatica platform’s connectivity to Mainframe sources.
* Informatica 9 will continue support feature of running the same workflow simultaneously.
* Browser based tool is a fully functional interface for business analysts.
* Dashboards are designed for business executives.
* There are 3 interfaces through which these capabilities can be accessed. Analyst tool is a browsed tool for analyst and stewards. Developers can use the eclipse based developer tool. Line of business managers can view data quality scorecards

* [MAPPLETS](http://informaticatutorials-naveen.blogspot.in/2011/04/mapplets.html" \o "Permanent Link to MAPPLETS)
* A mapplet is a reusable object that we create in the Mapplet Designer.
* It contains a set of transformations and lets us reuse that transformation logic in multiple mappings.
* Created in Mapplet Designer in Designer Tool.

We need to use same set of 5 transformations in say 10 mappings. So instead of making 5 transformations in every 10 mapping, we create a mapplet of these 5 transformations. Now we use this mapplet in all 10 mappings. Example: To create a surrogate key in target. We create a mapplet using a stored procedure to create Primary key for target table. We give target table name and key column name as input to mapplet and get the Surrogate key as output.

Mapplets help simplify mappings in the following ways:

* Include source definitions: Use multiple source definitions and source qualifiers to provide source data for a mapping.
* Accept data from sources in a mapping
* Include multiple transformations: As many transformations as we need.
* Pass data to multiple transformations: We can create a mapplet to feed data to multiple transformations. Each Output transformation in a mapplet represents one output group in a mapplet.
* Contain unused ports: We do not have to connect all mapplet input and output ports in a mapping.

**Mapplet Input:**

Mapplet input can originate from a source definition and/or from an Input transformation in the mapplet. We can create multiple pipelines in a mapplet.

* We use Mapplet Input transformation to give input to mapplet.
* Use of Mapplet Input transformation is optional.

**Mapplet Output:**

The output of a mapplet is not connected to any target table.

* We must use Mapplet Output transformation to store mapplet output.
* A mapplet must contain at least one Output transformation with at least one connected port in the mapplet.
  1. **What are the types of loading in Informatica?**

**Ans:** There are two types of loading, normal loading and bulk loading. In normal loading, it loads record by record and writes log for that. It takes comparatively a longer time to load data to the target in normal loading. But in bulk loading, it loads number of records at a time to target database. It takes less time to load data to target.  
  
**2) What is aggregate cache in aggregator transformation?**

**Ans:** The aggregator stores data in the aggregate cache until it completes aggregate calculations. When you run a session that uses an aggregator transformation, the Informatica server creates index and data caches in memory to process the transformation. If the Informatica server requires more space, it stores overflow values in cache files.  
  
**3) Define Informatica repository?**

**Ans:** Infromatica Repository: The Informatica repository is at the center of the Informatica suite. You create a set of metadata tables within the repository database that the Informatica application and tools access. The Informatica client and server access the repository to save and retrieve metadata.  
  
**4) What are the tasks that source qualifier performs?**

**Ans:** Join data originating from same source data base. Filter records when the informatica server reads source data. Specify an outer join rather than the default inner join specify sorted records. Select only distinct values from the source. Creating custom query to issue a special SELECT statement for the informatica server to read source data.  
  
**5) What is the target load order?**

**Ans:** Specify the target load order based on source qualifiers in a mapping. If you have the multiple source qualifiers connected to the multiple targets you can designate the order in which informatica server loads data into the targets.  
  
**6) How Union Transformation is used?**

**Ans:** The union transformation is a multiple input group transformation that can be used to merge data from various sources (or pipelines). This transformation works just like UNION ALL statement in SQL that is used to combine result set of two SELECT statements.  
  
**7) Can two flat files be joined with Joiner Transformation?**

**Ans:** Yes, joiner transformation can be used to join data from two flat file sources.  
  
**8) What is a look up transformation?**

**Ans:** This transformation is used to lookup data in a flat file or a relational table, view or synonym. It compares lookup transformation ports (input ports) to the source column values based on the lookup condition. Later returned values can be passed to other transformations.  
  
**9) Can a lookup be done on Flat Files?**

**Ans:** Yes.  
  
**10) Which transformation should you need while using the COBOL sources as source definitions?**

**Ans:** Normalizer transformation which is used to normalize the data. Since Cobol sources are often consists of Denormailzed data.  
  
**11) What is the difference between a connected look up and unconnected look up?**

**Ans:** Connected lookup takes input values directly from other transformations in the pipeline.  
Unconnected lookup doesn’t take inputs directly from any other transformation, but it can be used in any transformation (like expression) and can be invoked as a function using: LKP expression. So, an unconnected lookup can be called multiple times in a mapping.  
  
**12) What are the types of data that passes between informatica server and stored procedure?**

**Ans:** 3 types of data  
• Input/output parameters  
• Return Values  
• Status code.  
  
**13) What is the status code?**

**Ans:** Status code provides error handling for the informatica server during the session. The stored procedure issues a status code that notifies whether or not stored procedure completed successfully. This value cannot see by the user. It only used by the informatica server to determine whether to continue running the session or stop.  
  
**14) What is source qualifier transformation?**

**Ans:** When you add a relational or a flat file source definition to a mapping, you need to connect it to a source qualifier transformation. The source qualifier transformation represents the records that the informatica server reads when it runs a session.