**Solution**

This article provides more than 50 questions that can be used within an SSIS job interview. The questions are classified into four categories.

General Knowledge Questions

What is ETL?

ETL stands for Extract, Transform, and Load. It is a three-phase process that extracts data from different sources, applies some transformations, and loads it into a destination repository (database, data warehouse…).

What is SSIS?

SSIS stands for SQL Server Integration Services. It is the official ETL tool in SQL Server.

What is the difference between a database, a data warehouse, and a data lake?

A database is a repository that stores structured data used for transaction and retrieval operations. A data warehouse is a structured data repository used for analytical purposes. A data lake is a data repository where data is stored in its raw format and can later be loaded into a data warehouse or searched using advanced techniques.

*Datawarehouse is a location where we can store business related data for our analysis purpose. Read data and dump here.*

*Location means whether data is stored in SQL Server or Oracle.*

*Location where data is dumped and it is huge in size and we get different data sources*.

*Can you explain what a data lake is and how it differs from a data warehouse? Answer: A data lake stores raw, unstructured data in its native format, whereas a data warehouse stores processed, structured data optimized for querying*.

When are ETL jobs usually executed?

After working hours / at midnight.

How are ETL jobs scheduled in SQL Server?

Using SQL Server Agent Jobs.

------------------------------------------------------------------------------------------------------------------

What is the difference between ELT and ELT?

**ETL transforms data before loading it on the server, while ELT transforms it afterward. ELT is more suitable for larger volumes of data.**

ETL : we need to extract data from multiple sources, transform data and load it in different destinations. but if data volume is large this process/transformation takes long time

ELT: EXTRACT data, load data into staging table and then we perform transformation operation.

Eg: if data volume is large, amazon if want to perform transformation in SSIS LEVel, it is huge .in that case amazon extract data from source, then load data into staging area (sql server) without any transformation.After data is available in sql server then perform transformation and load data into final table(Sql server)

eg: we have Date time field, in final table we need only date

Datetime | Firstname |Lastname

Final table : we need only date and full name

We do Data conversion and derived column in SSIS.

if data volume is large, run one by one .

Insert into final (Date,FullName)

Select CONVERT(Date,DateTime), Firstname + '' + LastName from Staging

If transformation took long time, we can create indexing on staging table.

Staging area means before data warehouse.--------where data is dumped.

Eg: Truck ---------- multiple goods ------------- final output in ordered manner in proper section

In ETL/ELT/ SSIS we always talk about structured data. Order of column is changed.

------------------------------------------------------------------------------------------------------------------

What is the difference between relational and non-relational databases?

In a relational database, data is structured and stored within tables. Moreover, relations are modeled using primary/foreign key relations. In contrast, non-relational databases support semi-structured and unstructured data.

Relational and non-relational databases differ mainly in how they store, structure, and manage data:

**Relational Databases (RDBMS):**

* **Data Structure**: Store data in structured formats, using tables with rows and columns (like a spreadsheet).
* **Schema**: Require a predefined schema that defines the structure of the data (e.g., columns, data types).
* **Relationships**: Use primary and foreign keys to establish relationships between tables.
* **Query Language**: Use SQL (Structured Query Language) for managing and querying data.
* **ACID Properties**: Support strong consistency with ACID (Atomicity, Consistency, Isolation, Durability) transactions, ensuring data integrity.
* **Examples**: SQL Server, MySQL, PostgreSQL, Oracle.

**Non-Relational Databases (NoSQL):**

* **Data Structure**: Can store semi-structured, unstructured, or structured data. They often use collections, documents, key-value pairs, or graphs instead of tables.
* **Schema**: Flexible schema, allowing data to be stored without predefined structures.
* **Relationships**: Do not rely on rigid foreign key relationships. Relationships may be managed at the application level or through different indexing techniques.
* **Query Language**: Vary by database; some use their own query language (e.g., MongoDB uses a query language similar to JSON).
* **CAP Theorem**: Focus on the CAP theorem (Consistency, Availability, Partition Tolerance), often offering scalability and high availability over strict consistency.
* **Examples**: MongoDB, Cassandra, Couchbase, Redis.

What are lookup tables?

Lookup tables are commonly used for standardizing and unifying values within a text column, ensuring data integrity and improving data load performance. The values within a**data** table are replaced with a code that points to the value stored within a lookup table.

* **Definition**: A **lookup table** is a reference table used to store a set of predefined values or codes to standardize and unify data across a database.
* **Purpose**: Ensure data consistency by mapping values in the main table to a standardized set of codes or descriptions in the lookup table.
* **Normalization**: Lookup tables contribute to database normalization by reducing data redundancy and improving data integrity. Reduce redundancy by replacing repetitive text values with codes that reference the lookup table.
* **Benefits**:
  + **Data Integrity**: Ensures consistent values across the database.
  + **Performance**: Improves query performance by using shorter codes instead of lengthy text.
  + **Ease of Maintenance**: Centralizes updates to a single table.
* **Example**: Country codes in a Customers table refer to the full country names in a CountryLookup table.

What does data staging mean?

It is a location where data is stored before being modified and exported to the destination data repository (database/data warehouse)

What is data cleaning?

It is a set of processes and transformations to remove erroneous values from the data to make it more readable and consumable by the end-user by guaranteeing an acceptable data quality level.

What is data wrangling?

Data wrangling focuses on changing the data format to make it usable within an analytical context.

SSIS Development Questions

What is the difference between a control flow and a data flow?

Control flow and data flow are two execution levels in SSIS. A data flow defines a flow from a data source to a destination. In contrast, control flow defines a higher level of workflow of tasks to be executed. A control flow can contain one or more data flow tasks.

*Control flow task says which task run first and which task run last, how execution control flow from one task to another but do not talk about how data flow from source to destination. We can run parallel tasks in control flow. if we did not define order, both will run parallelly. if two data are not dependent on each other , then this is helpful. Everything in the toolbox is called task in Control flow.*

*CF always talks about task which is running paralleley also.*

*DF talks about source, destination and different types of transformations. We cannot run parallel tasks. We always choose one source*

What is the difference between an SSIS task and a component?

The SSIS task is executed on the package control flow level, while the SSIS component is executed on the data flow level.

What SSIS objects are required to connect to a data source?

A connection manager is required to establish a connection to a data source.

What are SSIS containers?

Containers are objects that provide structure to packages and services to tasks. It enables the repetition of control flows in packages and the grouping of tasks and containers into meaningful work units. Containers can include other containers in addition to tasks.

*SSIS containers are components in SQL Server Integration Services (SSIS) that help structure and organize the control flow in an SSIS package. They enable the grouping of tasks and containers into logical work units and can provide looping and conditional processing functionalities.*

There are four types of containers in SSIS:

1. For loop container
2. For each loop container
3. Sequence container
4. Task host container

Official documentation: [Integration Services Containers](https://learn.microsoft.com/en-us/sql/integration-services/control-flow/integration-services-containers?view=sql-server-ver16)

What is a precedence constraint?

As part of the control flow, precedence constraints link executables, containers, and tasks in a package and specify conditions for executing executables.

Official documentation: [Precedence Constraints](https://learn.microsoft.com/en-us/sql/integration-services/control-flow/precedence-constraints?view=sql-server-ver16)

What are the data flow components types?

There are three components available in the SSIS data flow:

1. Source components
2. Transformation components
3. Destination components

What are SSIS expressions?

Expressions can be used to define conditional statements, create and update values in data columns, assign values to variables, update or populate properties at run time, define constraints in precedence constraints, and provide the expressions used by containers.

Official documentation: [Integration Services (SSIS) Expressions](https://learn.microsoft.com/en-us/sql/integration-services/expressions/integration-services-ssis-expressions?view=sql-server-ver16)

What is the difference between SSIS variables and parameters?

Parameters store a value that should be configured before the package execution. They are read-only during the execution of a package. In contrast, variables can be modified during the package execution.

For more information: [SSIS Parameters vs. Variables](https://www.mssqltips.com/sqlservertip/4272/getting-started-with-sql-server-integration-services-ssis-parameters-vs-variables/)

How do you add a computed column into the data pipeline?

We should use a derived column transformation.

What are the SSIS objects that can handle XML data?

Two main SSIS objects can handle XML data type:

1. XML task: used to perform several operations unrelated to the data import process, such as schema validation.
2. XML source: used to read data from an XML file.

How do you convert a column's data type in SSIS?

There are several ways to convert a columns data type:

1. Changing the data type from the source advanced editor
2. Using a data conversion transformation
3. Using a derived column transformation

How do you extend SSIS functionalities?

Two objects allow extending the SSIS functionalities by writing customized C# or VB.NET scripts:

1. Script Task: execute a script on the control flow level.
2. Script Component: execute a script within a data flow task.

Is SSIS able to continue the package execution from the last successful step before a failure?

Yes, this can be done using Checkpoint.

Official documentation: [Restart Packages by Using Checkpoints](https://learn.microsoft.com/en-us/sql/integration-services/packages/restart-packages-by-using-checkpoints?view=sql-server-ver16)

What is the difference between the merge and merge join transformations?

The Merge transformation is similar to the UNION statement in SQL. In comparison, the Merge Join transformation is similar to the JOIN statement.

What is the difference between the execute SQL task and execute T-SQL statement task?

Execute SQL Task allows the user to execute statements over SQL Server and other supported sources. In contrast, the Execute T-SQL statement task is only dedicated to SQL Server (ADO.net) connections.

What is the difference between the OLE DB destination and the SQL Server destination?

OLE DB destination supports more DBMS than SQL Server and can insert data into a remote SQL Server. In contrast, SQL Server destination only supports local SQL Server database and has more restrictions.

For more information: [SQL Server Destination vs. OLE DB Destination - Stack Overflow](https://stackoverflow.com/questions/47388925/sql-server-destination-vs-ole-db-destination)

When should we use raw files?

Raw files are binary files supported in SSIS. They are meant to increase the performance of data staging operations compared to flat files.

For more information: [SSIS Raw File Source and Destination Example](https://www.mssqltips.com/sqlservertip/6217/ssis-raw-file-source-and-destination-example/)

Does SSIS have the ability to connect to the Hadoop ecosystem?

There are several components added in SQL Server 2016 that allow SSIS to read data and execute commands on a Hadoop cluster:

* Hadoop connection manager
* HDFS source and destination
* Pig Task
* Hive Task

Official documentation: [Hadoop and HDFS Support in Integration Services (SSIS)](https://learn.microsoft.com/en-us/sql/integration-services/hadoop-and-hdfs-support-in-integration-services-ssis?view=sql-server-ver16)

How can SSIS be integrated with Azure?

The SQL Server Integration Services (SSIS) Feature Pack for Azure provides several components that give the ability to connect to Azure services, transfer data between Azure and on-premises data sources, and process data stored in Azure.

Official documentation: [Azure Feature Pack for Integration Services (SSIS)](https://learn.microsoft.com/en-us/sql/integration-services/azure-feature-pack-for-integration-services-ssis?view=sql-server-ver16)

How do you distribute a data pipeline?

We can distribute the data pipeline using the Balanced Data Distributor component.

Official documentation: [Balanced Data Distributor Transformation](https://learn.microsoft.com/en-us/sql/integration-services/data-flow/transformations/balanced-data-distributor-transformation?view=sql-server-ver16)

What are event handlers?

SSIS Event handlers can be considered a separate package that operates asynchronously once an event occurs.

Official documentation: [Integration Services (SSIS) Event Handlers](https://learn.microsoft.com/en-us/sql/integration-services/integration-services-ssis-event-handlers?view=sql-server-ver16)

How do you log SSIS events?

SSIS includes several log providers that allow storing the package logs within an external file or a SQL database.

Official documentation: [Integration Services (SSIS) Logging](https://learn.microsoft.com/en-us/sql/integration-services/performance/integration-services-ssis-logging?view=sql-server-ver16)

What is the difference between OLE DB, ODBC, and ADO objects in SSIS?

In brief, OLE DB, ODBC, and ADO.NET are data access providers that allow SSIS to connect with different data sources. In addition to the SQL syntax differences, ADO.NET and OLE DB can access the same data sources. In contrast, ODBC can only access relational database systems, and they cannot access XML-based sources, Microsoft Exchange Server, and others.

How do you load data as batches from a SQL database?

We can use a for-loop container to specify the row count per batch and use an OLE DB Source with a SQL Command data access mode to read the data using an SQL OFFSET FETCH command.

For more information: [Reading Huge volume of data from SQLite to SQL Server fails at pre-execute](https://stackoverflow.com/questions/54168322/reading-huge-volume-of-data-from-sqlite-to-sql-server-fails-at-pre-execute/54172707#54172707)

How do you loop over values retrieved from a SQL table?

We can use an Execute SQL Task to store the SQL query result within a result set object. Then we can loop over the resultset rows using a for each loop container.

How do you loop over files stored in a specific directory?

We can use a for each loop container to loop over text files stored within a directory.

What components are used to read open data?

We can use the SSIS ODATA connection manager and ODATA source to read data from an open data provider.

For more information: [Using the OData Source in SSIS](https://www.mssqltips.com/sqlservertip/3220/using-the-odata-source-in-sql-server-integration-services/)

When should we use bulk insert task instead of data flow task to import data into SQL Server?

Bulk Insert tasks are more performant than data flow tasks and are more suitable for ELT pipelines. But they only support text files and requires some additional permission. In comparison, data Flow tasks are more suitable for ETL pipelines.

For more information: [SSIS Load Staging Data with Data Flow Task or Bulk Insert Task](https://www.mssqltips.com/sqlservertip/7216/ssis-load-staging-data-flow-task-bulk-insert-task/)

What feature allows SSIS tasks to be reused over several packages?

SSIS Package Parts.

Official documentation: [Reuse Control Flow across Packages by Using Control Flow Package Parts](https://learn.microsoft.com/en-us/sql/integration-services/reuse-control-flow-across-packages-by-using-control-flow-package-parts?view=sql-server-ver16)

How do you adjust a data pipeline buffer size?

There are three data flow task properties used to adjust the data pipeline buffer size:

* DefaultBufferMaxRows
* DefaultBufferSize
* AutoAdjustBufferSize

For more information: [SSIS Performance Tuning When Designing ETL Pipelines](https://www.mssqltips.com/sqlservertip/7475/ssis-performance-tuning-design-etl-pipelines/)

SSIS Administration Questions

What is SSISDB?

The SSISDB catalog is the central point for executing and managing Integration Services (SSIS) projects you've deployed to the Integration Services server.

Official documentation: [SSIS Catalog - SQL Server Integration Services (SSIS)](https://learn.microsoft.com/en-us/sql/integration-services/catalog/ssis-catalog?view=sql-server-ver16)

What are the different options to save an SSIS package?

There are three options:

1. Standalone .dtsx file
2. SQL Server msdb database
3. SSIS catalog

What are the different SSIS package deployment configurations?

There are two deployment configurations:

1. Single package deployment
2. Project deployment

Can the SSIS package be deployed to a previous version of SQL Server?

Yes, by adjusting the TargetServerVersion property in the SSIS project configuration.

Can SSIS packages be scaled over several servers?

Yes, Using SSIS Scale Out.

Official documentation: [SQL Server Integration Services (SSIS) Scale Out](https://learn.microsoft.com/en-us/sql/integration-services/scale-out/integration-services-ssis-scale-out?view=sql-server-ver16).

Are there any SSIS package development automation tools?

Several tools can automate the SSIS package creation. For example:

* Integration Services Object model
* EzApi class library
* [Biml](https://www.mssqltips.com/sqlservertutorial/9089/introduction-to-the-biml-language/)

For more information: [SQL Server Integration Services packages automation tools](https://medium.com/munchy-bytes/automating-ssis-package-creation-manageddts-vs-ezapi-vs-biml-vs-etlgen-d0dca92bf416)

Are there any SSIS DevOps services?

Microsoft provides two SSIS DevOps tools:

* SSISBuild: automates building SSIS packages
* SSISDeploy: automates the deployment of the packages.

Official documentation: [Standalone SQL Server Integration Services (SSIS) DevOps Tools](https://learn.microsoft.com/en-us/sql/integration-services/devops/ssis-devops-standalone?view=sql-server-ver16)

What is a .ispac file?

The .ispac is the deployment file extension of an SSIS project.

SSIS Online Communities Questions

While many people may find the question below out of the scope of an SSIS interview, being active in an online SSIS community should be a required skill of the job applicant.

What are the active SSIS online communities?

Three of the most popular communities are:

* [Stack Overflow SSIS tag](https://stackoverflow.com/questions/tagged/ssis)
* [The SSIS official blog](https://techcommunity.microsoft.com/t5/sql-server-integration-services/bg-p/SSIS)
* [The official MSDN SSIS forum](https://social.msdn.microsoft.com/Forums/en-US/home?forum=sqlintegrationservices)

What are the popular free SSIS extensions?

An example of popular free SSIS extensions is:

* [SSIS Community Samples](https://github.com/HadiFadl/mssql-ssis-community-samples): This project contains Integration Services samples and components created by the product team.
* [BI developer extensions](https://bideveloperextensions.github.io/): A Visual Studio extension that enhances development functionality in Business Intelligence Development Studio (BIDS) and SQL Server Data Tools (SSDT).
* [SSIS JSON Source component](https://github.com/albertogeniola/SSIS-Json-Source-Component): An open-source SSIS component that allows reading data from JSON.
* [SSIS RabbitMQ component](https://github.com/HadiFadl/SSISRabbitMQ): Custom SSIS Components for RabbitMQ

What are the leading SSIS components providers in the Visual Studio marketplace?

As an example of the leading companies that develop custom SSIS tasks and components:

* [Melissa](https://www.mssqltips.com/sqlservertip/6244/microsoft-sql-server-data-warehouse-data-quality-cleansing-verification-and-matching/)
* COZYROC
* KingsWaySoft
* [CDATA](https://www.mssqltips.com/sqlservertip/7171/universal-data-integration-tools-cdata/)
* ZappySYS