IST722: Unit 09 Participation Questions

**This is an individual assignment.**

**Before you begin, please make sure you’ve read and understand 1) our class honor code, 2) course policies on late work and 3) participation policies as posted on the syllabus. “I didn’t know” is not an excuse.**

**You should cite your sources in a standard format like MPA or APA and include a list of works cited.**

|  |  |
| --- | --- |
| Your Name: | Harsh Rajesh Darji |
| Your Email: | hrdarji@syr.edu |

# Instructions

Answer each of the following questions as concisely as possible. More is not necessarily better. Please justify your answer by citing your sources from the assigned readings from our textbooks, our class lectures, or online if directed to do so. Be sure to cite in text and include a list of works cited. Place your answer below each question. When you’re finished, print out this document and bring it to class as part of your participation grade.

# Questions

1. Explain the difference between ETL and ELT? Be sure to cover the advantages and disadvantages of each.

Extract, Transform, and Load (ETL) is a process that involves extracting data from outside sources, transforming it to fit operational needs (sometimes using staging tables), then loading it into the end target database or data warehouse.  This approach is reasonable as long as many different databases are involved in your data warehouse landscape.  In this scenario you have to transport data from one place to another anyway, so it’s a legitimate way to do the transformation work in a separate specialized engine.  Typically, the entire process is done in SSIS using data flow.

Extract, Load, transform (ELT) is a process where data is extracted for the source, then loaded into a staging table in the database, transforming it where it sits in the database and then loading it into the target database or data warehouse.  When the data is extracted from the source into the staging tables, it is a raw copy, meaning you keep the column names the same as in the source database and you don’t convert data, calculate new data fields, etc.

ELT has the benefit of minimizing the processing on the source since no transforming is being done, which can be extremely important if the source is a production system where you could be impacting the user experience as opposed to a copy of the source (via replication, database snapshot, etc).  The negative of this approach is it may take longer to get the data into the data warehouse as with the staging tables you have an extra step in the process, and you will need more disk space for the staging tables.

1. What are the advantages and disadvantages of a graphical ETL tool like SSIS? Research another ETL product on the market today. It is graphical or does it use a different implementation?

**Advantages**

* SSIS can handle data from heterogeneous data sources at a same package. We say Data sources can be diverse, including custom or scripted adapters.
* SSIS consumes data which are difficult like FTP, HTTP,MSMQ, and Analysis services etc.…
* SSIS provides transformation functionality.
* Easier to maintain and package configuration
* Tightly integrated with Microsoft Visual Studio and SQL Server.
* Use the SQL Server Destination instead of OLE DB; which allows you to load data into SQL faster.
* Remove network as a bottleneck for insertion of data by SSIS into SQL.
* Better for complex transformations, multi-step operations, aggregating data from different data sources or types, and structured exception handling.
* Data can be loaded in parallel to many varied destinations

**Disadvantages**

* To see package execution report need Management Studio rather than being published to reporting services or other way.
* If multiple packages are available that need to run parallel then you have a trouble. SSIS memory usage is high and it conflicts with SQL.
* In case of CPU allocation it also a problematic case when you have more packages to run parallel. You need to ensure that processer allocation between SQL and SSIS is done properly otherwise SQL have upper hand in it and due to that SSIS run very slow.

Informatica PowerCenter is another ETL tool which is graphical. PowerCenter forms the foundation for all your data integration initiatives, including analytics and data warehousing, application migration, or consolidation and data governance.

Key Features:

* Business and IT collaboration
* Reusability, automation, and ease of use
* Scalability, performance, and zero downtime
* Rapid prototyping, profiling, and validation

WORKS CITED:

<https://www.sarjen.com/2016/08/23/ssis-advantages-disadvantages/>

<https://www.jamesserra.com/archive/2012/01/difference-between-etl-and-elt/>

<https://www.informatica.com/products/data-integration/powercenter.html#fbid=Y5q6L7zgxRs>