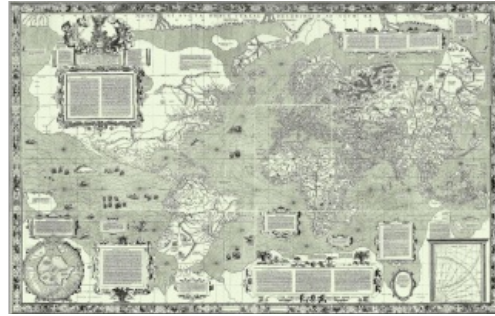


# (EXTRACT, TRANSFORM, LOAD PARADIGM)

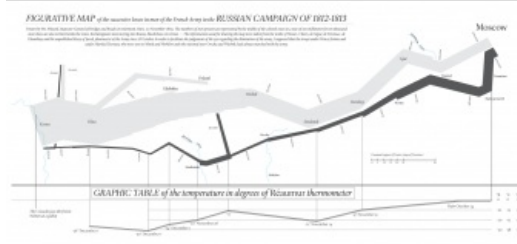
## The Death of the ETL Paradigm



Around three thousand years ago, someone living near the Euphrates River drew lines on a clay tablet to explain where they were in relation to others. This was the birth of cartography. However, it is only during the Renaissance period that maps became tools which could be used to safely navigate the world with the creation of the Mercator projection. This map has shaped the way we perceive the world nowadays, and still allows today's sailors to chart a constant heading course as a straight line on a map.

*By Marc Melvitz, CEO, Luciad*

Maps evolved again in 1869, when a French engineer named Charles Joseph Minard published a map depicting Napoleon's army losses during the disastrous Russian campaign of 1812.



This was the first map used to represent additional variables, such as temperatures and losses of human life over time and distance. Minard's contemporaries realized the innovative nature of his work, with photography pioneer Étienne-Jules Marey commenting that it "defies the pen of the historian in its brutal eloquence".

The advent of aircraft and submarines in the 20<sup>th</sup> century forced mapping to better account for the third dimension. At the same time computers made the rapid processing of large quantities of data possible, vastly increasing the potential power of both mapping and data analysis. The tools used to practice geography and cartography also changed; surveyors replaced yardsticks with lasers and geographers replaced pencils with keyboards and mice. However, the architecture of computerized systems remained essentially the same as those of paper systems. Just as all variables have to be represented at the same time as on paper, digital information has to be extracted from its original storage, transformed in order to become compatible with the base map and loaded in a geodatabase.

Today humans and machines produce data at a pace never before seen; although a lot of this data includes location information, location is not the main key by which the information is characterized. For example, a plane's engines can produce up to 1 Terabyte of data per flight, and we know or can compute the location of the engine when each parameter was recorded. However, the engineer analyzing the performance data of the engine will only be interested in location if location can help explain the data he is looking at.

It therefore does not make sense to extract plane engine performance data from the plane's computers, transform the data into geospatially indexed information and then load the data in a geodatabase; the time to process and the cost of duplicating the data storage are prohibitive. Yet this remains the paradigm for geospatial solutions, tying many current systems to an expensive and inefficient system for dealing with geospatial data. This is not just an issue of financial cost, either. The temporal cost of utilizing data in this fashion can seriously affect the time it takes for the users of geospatial systems to complete tasks. This slows down the tempo of the work geospatial analysts perform and, in certain settings such as defense, can affect the success of operations and the lives of those carrying them out.

However, this solution to this cannot just be selectively loading data into a database. After all, the data must sometimes be available for visualization and analysis in a geospatial context. For example, the engineer in our example may need to combine weather information with engine performance. The paradigm must therefore shift to one in which the data that analysts need can be loaded and accessed at will with systems that can respond to the changing needs of users.

The answer [Luciad](#) proposes is based on two concepts: the data lake and data virtualization. The data lake is the virtual place where data is first produced and stored, often in raw format. This can take the form of streams or files, SQL or noSQL databases, centralized or distributed, on-premise or in the cloud.

Data virtualization leaves the original data in place and makes it accessible to services performing further processing, visualization and analysis. Because it is impossible to predict which combinations of data and what transformations will be required, Luciad uses an in-

memory model of the data to perform data abstraction, federation, and transformation on the fly.

Add advanced visual analytics capabilities, and you have all the building blocks to create real time situational awareness systems, that are capable of helping the end user uncover unknown unknowns.

---

<https://www.gim-international.com/content/article/the-death-of-the-etl-paradigm>

---