IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Usability Comparison White Paper: Informatica vs. Ab Initio

Comparative Analysis of ETL Tools in Data Analytics

Rajneesh Shukla
Solution Data Architect, Bengaluru, India

Abstract : This report is an evaluation of two leading Extract, Transform and Load (ETL) tools, Informatica and Ab Initio. In spite of long history of data ware housing, business intelligence and data integration industry, it is not an easy to decide which ETL tool is best suited for the organization needs as both Ab initio and Informatica offers a complete ETL solution and both are leaders in the Market. The purpose of this usability study is to evaluate usability of each solution with respect to various criteria like cost, modularity to support seamless integration with other applications/tools, analytical reporting performance needs, volume and variety of data, loading performance, scalability, ease of use, self-sufficiency, simplicity, compatibility etc. If cost is not a concern Ab initio can be preferred for performance, processing of high volume of data and easy to develop perspective. Informatica provides a broad range of Data Integration products like Informatica PowerCenter, Informatica Data Integrator Hub, Informatica Cloud etc. and hence it boasts of a broader data integration capability. Specific offerings, future scope, ongoing enhancements, data integration capabilities, hardware and operating system requirements, architectural differences, maturity in various market segments etc. are detailed for both solutions.

I. INTRODUCTION

From data ware house to business intelligence now we are thinking one more level above because we are experiencing unexpected growth in structured and unstructured data. Managing such massive information and day to day transformations that too with mindset deliver information quickly. This is the reason the data engineering and analytics industry is heading towards Big Data. But the pain area is Big Data solutions like Hadoop are still emerging and yet to be compelling for heavily relational and SQL based processing. They yet to rely on traditional ETL tools like Ab initio and Informatica to present the processed data in form of analytical data needs for performing transformation, standardization, data cleaning, data validation, data enrichment. Both traditional ETL solutions i.e. Ab initio and Informatica provided diverse portfolio of the tools for data integration and related disciplines, such as data quality, metadata, master data, data replication and data synchronization while available Big Data solutions are extension of traditional ETL tools and need to be emerged further to cope up with broad spectrum of data engineering's needs. Hence there is a need of integrating traditional data ware house with Big Data solutions. It is not easy for organizations to decide which ETL solution is preferred option as both Informatica and Ab initio offers a complete ETL solution and both are leaders in the market. Architecture, offerings, cost and future scope vary between these 2 most popular ETL tools. The objective of this report is to evaluate the usability of both ETL tools and present the comparative analysis to provide the fact full information for decision makers, enabling them for opting right ETL tool based on various factors like their organizations existing data infrastructure, data integration needs, analytics needs, resources, budget, data volume and format, performance needs etc.

II. ARCHITECTURAL COMPARISION

Informatica:
Grew up as a metadata driven, GUI development environment. The engineers and the founders decided it was a good idea to model ANSI-SQL types of commands in a visually driven drag & drop style. They also decided that keeping this information in a metadata repository was a good idea. So they set on their way. Foundationally strong in Metadata, they also made a key decision to put their metadata in to table structures in standard database systems (hence the repository). However, they wanted to protect their metadata repository as "intellectual property" that was owned by the company, so they removed indexes, removed primary and foreign keys, and they named some of the elements the same way. This had its upsides and down sides to the tool – especially being metadata driven, the focus was *not* performance *nor* parallelism. Later they decided to add parallelism, and partitioning, and thus was born the PowerCenter product. Their focus today is still very much metadata (and very strong), but they are now extremely fast and can perform extremely well under the right circumstances.

Ab Initio:

Grew up as a speed monger. They did not start with metadata first, their whole purpose in life was bent on making the base OS: a) fault-tolerant b) fast c) partitionable flat file access.– that's right, it's all 100% flat-file driven. So what they did was create something called the Co-Operating system. The co- operating system basically is a parasite. It sits on top of your existing OS, and plugs in to the I/O interrupt level – it assumes and overrides 100% of all file operations on the machine. Even files that aren't "running" through Ab-Initio go through the AB-I interrupts at the file system level. They then, constructed their own "data description and manipulation language" known as ab-initio scripts (much the same way C-Shell, or K-Shell acts in Unix). They interpret these scripts at the start of a run-request (compile them really), and then figure out mathematically what operations can be split across which resources. Their only notion of a shared repository (at that time) was the systems-performance metadata. They gather this metadata from each registered system upon install to make a profile of the machine's capabilities. This makes heterogeneous activity possible, and load-balancing mathematically easy. It's also why when you upgrade your OS, or upgrade your hardware that you're asked to re-run the machine profile, and re-register it with the "central brain". The Co-Operating system is parasitic. It renders the rest of the OS (for all intensive purposes) useless, even though they'd like you to believe it's seamless and harmless. Yes, some things still work properly – but others give spurious errors (unknown and untraceable errors). They then developed a GUI that allowed designers to build "data flows" – graphs they call them. These graphs were then "exported" to the co-operating system (produced if you will). What that process really does is generate Co-Operating system scripting code. The unfortunate part of this, is: programmers decided they could take the generated code and make it better/faster so they tweak it. Once they tweak it, it's disconnected from the metadata graph in the GUI and bang, your maintenance costs rise again – and you've got uncontrollable code sprawling across your enterprise. They built their own metadata database (an internal binary format). Anyhow, binary or proprietary metadata is a serious problem in the industry.

Table1:

Informatica PowerCenter	Ab Initio
• Informatica is an engine based ETL tool, the power this tool is in its transformation engine and the code that it generates after development cannot be seen or modified.	• Ab Initio is a code based ETL tool, it generates ksh or bat etc. code, which can be modified to achieve the goals, if any that cannot be taken care through the ETL tool itself. The script is human readable but not intended to be messed with. Any change must be done inside Graphical Development Environment (GDE) and script be re-deployed.

Diagram1: Ab initio overview

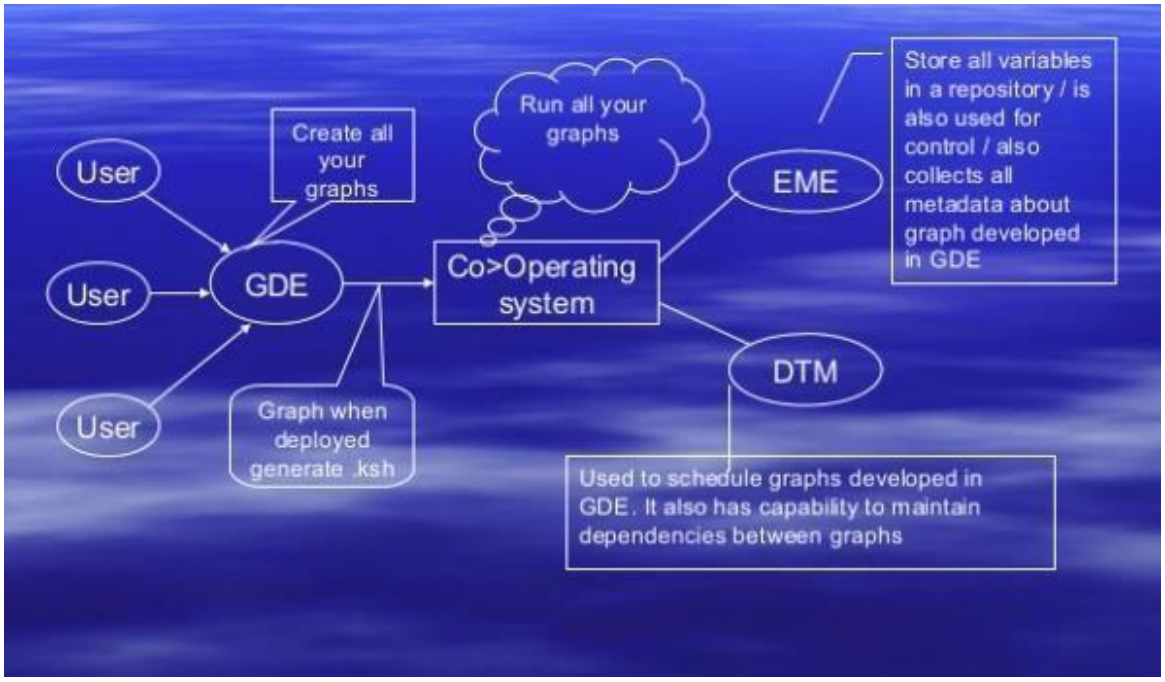


Diagram 2: Ab Initio Architecture:

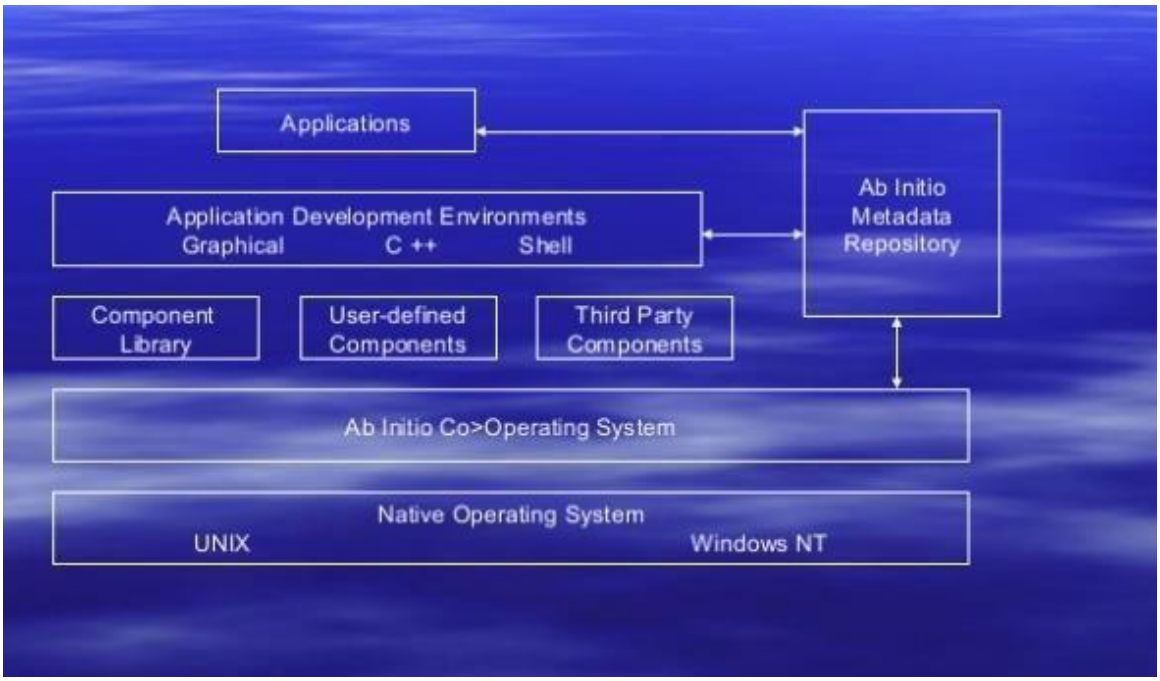
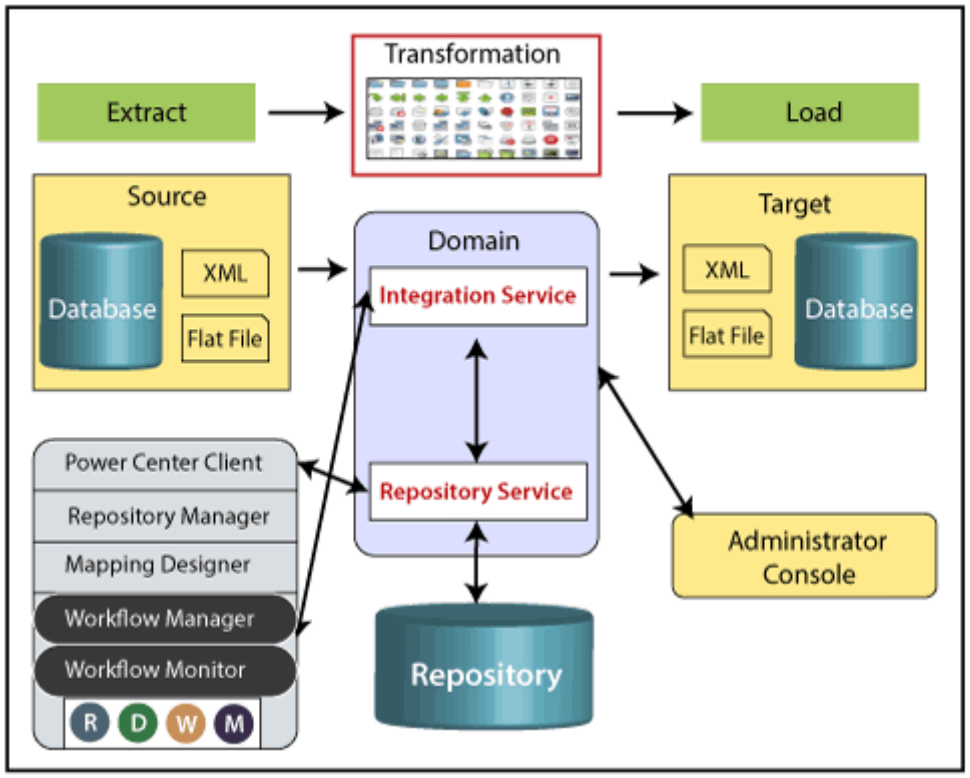
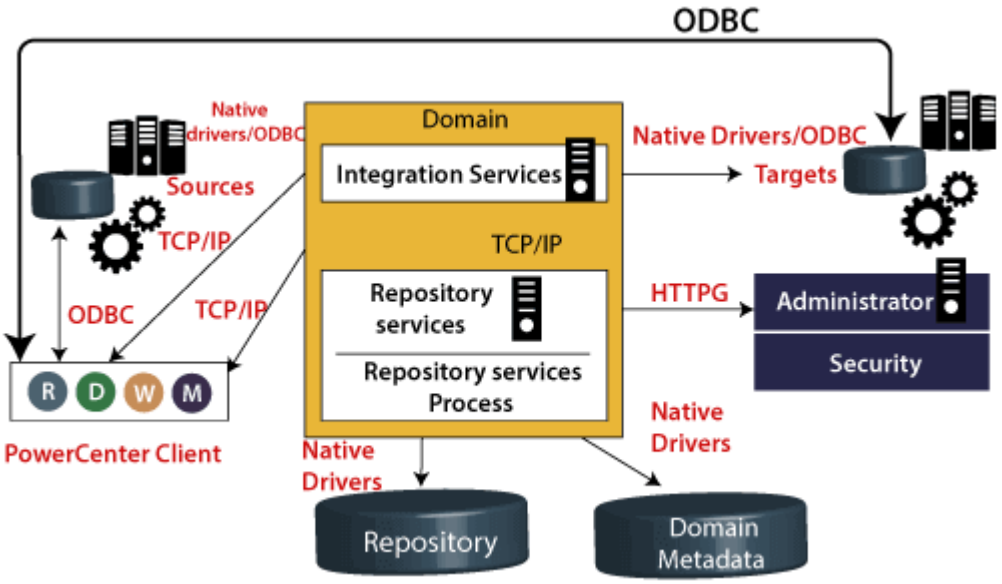


Diagram 3: Informatica PowerCenter Architecture:





III.PARAMETER BASED COMPARISION

Table2:

Parameter Name	Informatica PowerCenter	Ab initio
Performance	Using similar hardware as Ab initio, how Informatica will perform will depend heavily on how the code was developed. In general, Informatica does not offer similar performance as Ab Initio. Informatica does not offer parallelism by default. You need to develop your workflow in such a way so that it runs in parallel. If you develop multiple transformations within a single mapping, they will run in serial mode!	Due to the inherent way Ab Initio handles parallelism, it works brilliantly with massive data volumes. Ab Initio supports 3 types of parallelism from the ground up. These are – component, pipeline, and data parallelism. Even if you don't know much about parallelism, you can be developing Ab Initio graphs using parallelism.
Volume of Data	Informatica is more of a mainstream product preferred by small to medium companies.	Ab ignition is preferred by large companies which require very large amount of data handling.
Integration with Big Data	Informatica® PowerCenter® Big Data Edition is the safe on-ramp to big data that works with both emerging technologies and traditional data management infrastructures.	3.2+ versions of Abinitio handles the Hadoop also via Hadoop connector to process data from Hadoop Distributed File System (HDFS) for ETL process
Pricing	Far cheaper.	Too expensive. So much so that some large companies are now looking for alternatives.
Maintainability	Less Maintainable.	The maintainability of Ab Initio is much easier than of Informatica
Version Control	it is tightly controlled in Informatica	The version control of Ab Initio is moderate
Tool Bugs	Informatica are known to carry more tool bugs than the Ab initio tool.	in Ab Initio, tool bugs are minimal.
Overall Market Leader	Informatica ETL tool is the market leader in data integration and data quality services. Informatica is successful ETL and EAI tool with significant industry coverage.	This is also leading tool but comparatively less market capture.
Push Down Mechanism	Informatica has push down optimization concept, where it can generate SQL statements from the workflow/mapping which can be directly executed on database	push down optimization concept.
Data Integration Capability	Informatica boasts of a broader data integration capability which includes many related functions such as profiling and data quality. Also, it is more mainstream in nature than the Ab Initio.	Comparatively lesser range of Data Integration Products.

Parameter Name	Informatica PowerCenter	Ab initio
Company strategy	Informatica is normal in this aspect. You can download a scaled down free version of their software and plenty of free documents available on Informatica on internet.	Ab Initio has some stupid strategies like Non-Disclosure Agreement (NDA). Ab Initio does not offer any free version or not even documents to explain their products. Their strategy is to create a buzz in known circle and leverage that. Since last few years, this strategy has backfired, especially when there is fierce competition in ETL market and customers are less interested in a product which they can't research upon without signing. an NDA to start with!
Repository	Informatica requires a repository which must be a relational database (e.g. Oracle). It also installs some services and Power Center tools. To open an Informatica code in another computer, one needs to export those as XMLs and import in another instance of Informatica installations.	Although Ab Initio does offer a repository (EME as mentioned earlier) it is not mandatory. Even if you use it, it does not require any 3rd party database like Oracle to store repository information. As a bare minimum you need to install the Co- operating system (which runs deployed script) and GDE. A graph is a single .mp file which can be opened in another computer (with GDE installed) without anything else.
Debugging	Debugging is easy in Ab Initio. It is a much complex procedure in Informatica (through debugger).	Very easy – just place some watchers on required places and intermediate data will be saved in temporary files for easy viewing.
Maintainability	Less Maintainable.	The maintainability of Ab Initio is much easier than of Informatica
Version Control	it is tightly controlled in Informatica	The version control of Ab Initio is moderate
Tool Bugs	Informatica are known to carry more tool bugs than the Ab initio tool.	in Ab Initio, tool bugs are minimal.
Overall Market Leader	Informatica ETL tool is the market leader in data integration and data quality services. Informatica is successful ETL and EAI tool with significant industry coverage.	This is also leading tool but comparatively less market capture.
Job Scheduling	In built Scheduler.	Scheduling is also more cumbersome in Ab Initio because there's no scheduler unlike in the case of Informatica. Hence, you need to run a script or enter a schedule manually if you're using the Ab Initio.

Parameter Name	Informatica PowerCenter	Ab initio
File Structure	Informatica is only able to read record with single type of delimiter.	Ab Initio is a friendlier system than the Informatica. It is because it can process varying text files so you can read or browse through them even if each file is structured differently. Also, Ab initio is better known as a tool which can read record with multiple type of delimiter.
Go Live Rate	High "GO Live" success rate. Informatica claims the highest ration of successful deployment which it says is near 100%. Product rate of renewal and customer loyalty (94% and 92%) is significant from industry average.	Comparatively less Go Live Success as developer used to modify. ksh code after taking Graphs export so both become out of sync.
Go Live Rate	High "GO Live" success rate. Informatica claims the highest ration of successful deployment which it says is near 100%. Product rate of renewal and customer loyalty (94% and 92%) is significant from industry average.	Comparatively less Go Live Success as developer used to modify. ksh code after taking Graphs export so both become out of sync.
Training and Tool availability	Easy training and tool availability has made easy resource availability for software industry. This definitely helps the organization in reducing training costs. Moreover, forming a new team for this tool is not as challenging.	Comparatively expensive.
Expert Opinions	Rob Karel of Forrester Research says "Informatica: A Leader In Enterprise ETL, With A Pure-Play Solution.	Best Suited for companies who can pay high.
Operating Systems	The PowerCenter Server is supported on various flavors of NT, Unix, Linux, and the mainframe. PowerCenter client is supported on various flavors of Windows.	Ab initio runs on many operating systems. Compaq Tru64 UNIX Digital Unix HP-UX IBM AIX NCR MP-RAS Red Hat Linux IBM/Sequent DYNIX/pt Siemens Pyramid Reliant UNIX Silicon Graphics IRIX Sun Solaris Windows OS
Hardware Platforms	Hardware Platforms requirement for Informatica and Ab initio is almost same.	No Problem is too big or too small for Ab initio. Ab initio runs on a few processors or few hundred processors. Abinitio runs on every kind of hardware: Symmetric Multiprocessor System. Massively Parallel Processor System. Clusters. PCs

IV. CONCLUSION

The study report shows that while both Informatica and Ab initio offer a wide range of ETL features, Ab initio provides users with significant advantage over Informatica in terms of integration with Big Data, processing performance, high volume data, ease of use & implementation. Informatica boasts of a broader data integration capability which includes many related functions such as profiling and data quality. Also, it is more mainstream in nature than the Ab Initio. Informatica ETL tool is the market leader in data integration and data quality services. Informatica is successful ETL and EAI tool with significant industry coverage. If money is no object, then Ab Initio offers some advantages and is easier to use. However, in real life, money is often the most important factor. Thus, Informatica has a greater market share than Ab Initio. Overall, both the Ab Initio and Informatica are different tools whose appropriateness depends on the company's need, the nature of data integration needed, amount of data to be handled and the overall infrastructures available.

V. RRFERENCES

[1] http://etllabs.com/
[2] http://dwhnotes.com/
[3] http://www.wikidifference.com/
[4] http://www.differencebetween.net/
[5] https://www.itcentralstation.com/
[6] http://empoweredholdings.com/
[7] https://www.informatica.com/in/products/data-integration/
[8] https://www.tmforum.org/resources/best-practice/gb979-big-data-analytics-solution-suite-r16-5-1/
[9] https://www.tmforum.org/resources/standard/guidebook_integrationframework-conceptsandprinciples-v1-0-2-2/