Y-Point Perspective...

Performance Issues in Analytics

You are working on an analytics project and you are continuously plagued by performance issues. No matter what patch you apply these issues don't seem to go away. When you fix one part of your application, problems creep up in other. Not just that, your existing analytic applications that used to perform, are now slowing down. Business demand for real time data is growing, but your analytic applications seem to slow down as the data that you use grows.

Based on years of our experience in turning around failed analytic projects we have identified that some of the key factors that impact performance in analytic projects can be distilled down to following

- Lack of Precision in Business Requirements (Data Elements, Latency, Grain)
- Lack of Data Architecture Discipline
- Lack of ETL Design

Lack of Precision in Business Requirements

Lack of precision in business requirement is the major culprit. During the requirements gathering process most teams forget to create a conceptual data model of ***data elements*** and don't attempt to understand how various data elements are related from a business's perspective. Lack of precision in gathering data elements and their relationships results in development team going after more information than what is really needed by the business. Information technology department ends up sourcing lot more data than needed, and many more data elements than needed.

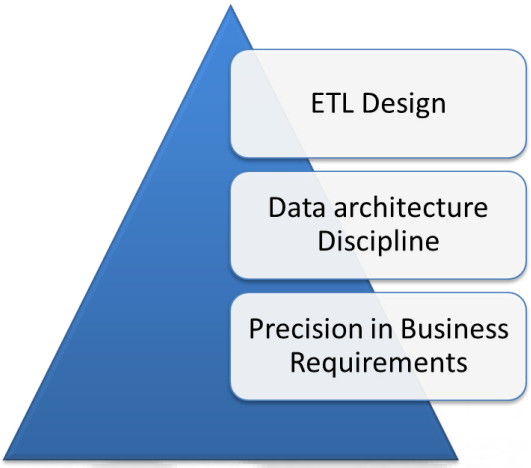
This results in significant increase in project complexity, time to completion and eventual failure.

Typically ***complexity*** of an analytic application grows exponentially to the number of ***data elements***. Typically integration complexity is a function of square of number of data elements. It increases at a faster pace than the increase in your data elements. These increases in complexity results in creation of complex ETL code that lacks performance and has to deal with added volume and complexity of no value add data elements.

Wrong Latency: During the requirements gathering process it is critical that the actual data latency is appropriately captured. If this question is asked during a casual conversation, every business user would always want all the information in real time. An expert business analyst realizes that and digs deep enough to understand the real business need and how they tie to available data. No data can ever be real time. ***We can only get as close to real time as possible***. And the way real time data is integrated is very different from the way batch data is integrated. Real time data integration along with complex data transformation and integration is guaranteed to cause performance issues. We have yet to come across an

analytic project that has no complex data integration or data transformation needs. The need to get complex data in real time is the next reason why many analytic projects fail to meet the performance expectations of the business.

Wrong Grain: Grain is the aggregation level at which the business needs the data. It is also reflected in various paths a user can take while navigating the OLAP cubes. An atomic fact table can be aggregated at a higher grain by reducing a few dimensions it is associated to. The metrics in the fact table get rolled up (aggregated) along the selected dimensions. Business does not always need data at the most atomic grain. Even though the best practice is to keep data available at the most atomic grain, knowing the exact grain at which business consumes the ***information goes a long way in helping performance***. Knowing the exact grain allows data architects to create aggregated marts. If an aggregate aware BI tool is used, it can point its roll up queries to appropriate aggregate tables significantly boosting the performance. Remember that each BI tool has its own attribute and table naming standards that allows them to do that. So the data architect should be aware of this capability.



Lack of the Data Architecture Discipline

We come across several projects where, when we ask for names and contact information of data architects, we are given the names of database administrators. In most large companies the database administrator manages the administrative aspects, backup, restore, performance tuning, security, installation set up, configuration of databases. But they are usually not the architects who understand the relationships between all the data elements, the business impact of changes to data structures, and have an understanding of the data sources to assess the impact on the data supply chain when source systems change.

Data Architects are also responsible for enforcing standards that are required by the BI tool. These are standards such as Aggregate Navigation, Naming standards, Attribute precision, and length. They also maintain various versions of models and release models with changes tracked to previous version to all the stakeholders of analytic projects. Lack of data architecture as a discipline from the word go, results in patchwork data model over time. Loading and reading data from this patchwork creates a significant performance overhead on both your ETL tools and analytic applications.

Lack of ETL / Data Integration Design

A good ETL design is like a Vulcan mild meld. It collects information from Data Architects, Business Analysts, Data Stewards and projects it as a data flow needed to complete development. A good ETL design always stays in sync with code.

We have yet to come across complex ETL development that was delivered in a single iteration. ***It takes several iterations to get complex data integration right.*** This is because there are several data quality, integration, missing data issues that need to be addressed. A good data integration design significantly cuts down the number of iterations needed to get complex integration right. We see several projects that just don't get the concept of ***Design***. There is always the urgency to

jump in and start coding. And, places where we have seen the design discipline, we have noted that they overdo designs. They incorporate so much information into the design that it is almost impossible to continuously ***update the Design to the match code***. Very quickly the design and code go off sync, making the design useless and creating 100% reliance on code. This again breaks down the design process. Getting the design right is an art. At Y-Point we have mastered it, and maintaining designs requires a continuous improvement mid-set that is very hard to find.

Conclusion

There are several reasons why performance suffers in analytic projects; these range from not using the right kind of database - Columnar vs. Row based, vs. node based to incorrect partitioning of the tables and/or ETL code.

At Y-Point we strongly believe that the above three factors are the primary and the most ignored ones and usually give you the most bang for your buck. Let's engage.

About Us

Y-point Solutions was founded by IITB (IIT Bombay) alumni and a set of young entrepreneurs who saw the utility of data far beyond its transactional manifestation. Data to them was much more than transactions, it was a fact with many dimensions. And these dimensions when connected together offered unparalleled insight into an enterprise's health.

There are two broad objectives of an analytical system. Understand the dynamics of the ecosystem and be able to anticipate the future challenges to a business or future disruptions. Generating data as useful information is thus the key to simplifying complexity.

Contact Us
Y-point Consultants
13702 Maple Sugar Lane, Herndon, VA, 20171
Phone – 703-880-4128
Email – info@ypointanalytics.com

Call us at: +1-703-880-4128
Mon-Fri 7am-1am EST, Sat 7am-11:30pm EST, Sun 9am-10pm EST