Introduction

As more organizations execute on digital transformation strategies, they shift away from focusing on traditional physical assets and instead make the production and use of “data assets” the primary driver of return on investment. Just as having optimized production lines for physical assets was true in the old economy, streamlining the production and use of data assets in the digital economy becomes the benchmark for efficiency and a key driver of faster execution.

The modern data lake has become the central production line and repository for data assets that enable digital transformation. Agile production of data assets to feed the information-hungry organization requires a re-thinking of how to approach the problem. Recreating traditional methods used with the enterprise data warehouse (EDW) on the data lake is simply applying an old world approach to a new age problem – a complete mismatch.

The new age approach requires a tighter partnership between IT, who manages the data, and the business, who ultimately determines the value and usefulness of the data. Ultimately, this necessitates integrative data exploration to bring the business analyst into the production process.

Datameer’s next-generation platform with the new Visual Explorer completes the last mile linking your business analysts to the data lake and creating a full cooperative curation process to deliver data assets across the entire organization. Visual Explorer breaks down the inefficiencies and slow response times that plague multi-level stacks of data preparation and OLAP-on-Hadoop products, providing a truly interactive data exploration environment that further speeds time to insight.
Evolution of the Data Lake

The data lake has evolved from an experimental analytic sandbox to become the centerpiece of an architecture that creates new data assets in an agile fashion to drive digital transformation.

As organizations learn more, we are seeing an evolution from Data Lake 1.0 to Data Lake 2.0 with three major differences:

- While early iterations of data lake focused on filling those lakes with data that could possibly be used, the second edition of the data lake will focus on how the data is curated from raw form into useful data that can be readily consumed by the business.

- A number of experimental projects and early high impact use cases were the use patterns for data lake 1.0, while data lake 2.0 will strive to be used across a much greater variety of business use cases with data readily used.

- To date, the process of operating a data lake was extremely people intensive with expensive custom coding efforts, while in the new version, organizations will focus on optimizing processes to create new data assets and deliver these to the business.

<table>
<thead>
<tr>
<th>Focus</th>
<th>Data Lake 1.0</th>
<th>Data Lake 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creating the data lake:</td>
<td>Ingesting data</td>
<td>Using the data lake: Curating and consuming data</td>
</tr>
<tr>
<td>Use</td>
<td>Experimental, ad-hoc analysis and data science</td>
<td>Purposeful use cases for consumed data assets</td>
</tr>
<tr>
<td>Processes</td>
<td>Highly people and cost intensive</td>
<td>Productive, optimized data delivery processes</td>
</tr>
</tbody>
</table>

The process optimization aspect is important to help organizations drive more value from their data lakes. The enterprise data warehouse went through its own pains, but was able to grow as best practices evolved and software products adjusted to meet those best practices.

The next evolutionary phase of the data lake will center around optimizing the process of creating, curating and consuming data assets that will drive digital transformation throughout various aspects of the business.
Creating a Cooperative Process

In their 1.0 versions, many data lakes were created based on expected need, either by the data teams within the IT organization or IT-savvy data teams within the business units. While there were some successes, there was also failure.

In both cases, organizations are learning from their experiences to create a new round of best practices that will help drive more value from their data lakes and ultimately greater success. At the heart of these best practices is the process of creating, curating and consuming data assets out of the vast pools in the data lake.

Inherently, data is a joint asset that is managed by an IT or data organization and value is driven and derived by the business. One cannot be successful without the other. IT doesn’t have a role if the business doesn’t need data. The business isn’t successful if IT doesn’t manage the data effectively.

Digital transformation is dramatically altering the time to value equation for data assets. Successful digital transformation is requiring organizations to create agile processes that optimize how new data assets get delivered and dramatically shrink the time to value for those assets.

One approach taken by some firms was to replicate an EDW stack on Hadoop, with some of the components changing a bit (see figure 1). Inherently, these approaches had multiple flaws:

- The lengthy stack created “latency” in moving the data from one end to the other
- Data was often duplicated consuming additional system and storage resources
- Data duplication also complicated data governance and security, with each tool having their own models and approaches
- Each tool had different execution and processing models, making operationalizing pipelines impossible

Additionally, there were technology issues. SQL on Hadoop was slow and inherently non-interactive, frustrating business analysts who connected their BI/discovery tools. OLAP on Hadoop created pre-fixed models, limiting the ability of the analyst to effectively explore the data in any direction.
These stacks also created process level inefficiencies that slowed down the delivery of the final data assets. Business analysts would describe their data needs to the data engineer, but there was often guess-work performed. A data set would get delivered to the analysts, who would explore it and find it did not meet their needs. It often took multiple passes until the final data set was curated properly.

For a data lake to rapidly produce valuable data assets, there needs to be an integrated cooperative curation process that seamlessly brings together the data engineer and the business analyst that is both resource and process efficient. It requires a process where the two personas meet in the middle, each employing their own strengths, yet using an integrated platform that optimizes resources, operationalization and governance at an enterprise-scale.
Bringing True Interactivity to the Data Lake

The “last mile” to bridge the data lake to the business, and the data engineer with the business analyst, is a truly interactive data exploration capability that is integrated into a unified big data platform that enables the entire data lifecycle – ingestion, preparation, exploration and consumption. Such an experience will draw the business analyst into the data lake to do the last mile of refinement of datasets through interactive exploration in a common, friendly visual metaphor.

To deliver on the promise of big data exploration, a solution needs to fulfill the following definition of “interactive”:

- **Unconstrained** — enables exploration in any direction and path without restrictive pre-defined schemas;
- **Visual** — offers a fully visual metaphor to explore data and create refined datasets that meeting specific analytic needs;
- **Fast and Responsive** — has sub-second response time, even with large datasets, and provides fast visual exploration methods;
- **Scale and Drill to Detail** — works over datasets in billions of records and enables drill-down to extremely granular detail;
- **Auto-generate Datasets** — captures the entire exploration path parameters and allows rapid generation of the resulting dataset;

In addition, the interactive visual exploration needs to work in concert with a comprehensive big data platform that:

- Covers the complete data lifecycle
- Eliminates data duplication
- Has comprehensive governance
- Offers strong security
- Scales to enterprise needs
Datameer Visual Explorer

Datameer Visual Explorer combines a fully visual data exploration metaphor with a unique architecture to be the industry’s first solution to truly deliver interactive data exploration at scale. It bridges the last mile between your data lake and business analysts and works with the remainder of the Datameer platform to create an efficient, scalable, full lifecycle platform that creates an agile, cooperative curation process.

The unique schema-less architecture frees the analysts from the complex modeling, rigid multi-dimensional schemas, and pre-aggregation processing of traditional OLAP solutions. This lets your business analysts focus on what’s important – the data and the process of extracting real value from it for the business.

Unconstrained Exploration

The questions analysts answer in the era of big data are often unbounded. They have a general idea of where to explore, but the objective is to let the data led you down the right paths. Pre-defined schemas limit the ability of the analyst to explore the data, revealing only pre-determined paths that could obscure the true answer to a question.

With Visual Explorer, business analysts are no longer constrained by pre-defined models and materialized cubes, giving them the freedom to explore the data at the speed of thought. The cube-less architecture and dynamic search-based indexing allows an analyst to slice the data using any attribute and value, in any direction they want to explore and see results instantaneously.

Visual Metaphor

The advent of desktop data discovery tools has made the modern business analyst much more comfortable with visual approaches to data exploration. Yet, this metaphor has not reached the big data world to speed exploration of the large amounts of data inside the data lake – until now.
Datameer’s Visual Explorer makes it easy to explore, shape and refine in a fully visual manner that any business analyst can use. Create graphs interactively, drill into sections of the data, add or adjust attributes or aggregations – with sub-second response time on millions or even billions of rows.

**Lightning Fast Response**

Hadoop has traditionally offered a batch-oriented processing model, requiring the analyst to wait and see the results of their data exploration. The advent of faster engines such as Spark alleviated this problem somewhat, but still did not provide the interactive response time that analysts need for data exploration on extremely large datasets – in the billions of rows.

With Visual Explorer you get lightning fast response times exploring up to billions of records interactively. You don’t have to wait for batch processes, or to pre-compute and materialize cubes and indexes. Visual Explorer’s unique architecture rapidly generates optimized indexes on the fly based on what fields the analyst is exploring. This eliminates loading and navigating heavyweight indexes and requires no pre-compute and extra storage.
Work at scale and drill to detail

No one will argue that datasets inside of a data lake are extremely large. But it is important to recognize this size comes in three dimensions:

- **Long** – billions of rows
- **Wide** – hundreds of attributes and metrics
- **Deep** – large numbers of values for each attribute

This creates extremely high cardinality on the dataset and could exponentially increase the size of the indexes used to represent the data. At these extremes, SQL and OLAP on Hadoop engines simply give out.

Visual Explorer can work on billions of rows and hundreds of attributes with response times still in the seconds, and continues to scale out the number of concurrent threads. Then, drill into any granular level of detail to find the exact set of data needed for the analysis at hand.
Auto-Generates Dataset Refinements

Once you’ve explored the data, then what do you do? You need to capture the data and leverage it for analysis wherever its needed. Analysts can stay within Datameer for further analysis and visualization, or, bring the prepared dataset to their favorite local BI and discovery tools to explain the results to the business teams, and perhaps even share it with other analyst colleagues.

Visual Explorer captures the entire trail while you explore and refine the data – attributes, filters, drill paths, etc. Once you’ve gotten to the data set you need, simply click a button and a new sheet containing the refined dataset is created. From there you can:

- Stay within Datameer’s unified platform and leverage the same spreadsheet interface for analysis plus apply advanced algorithms for clustering, data dependencies, decision trees, and more, or,
- Export this dataset into your favorite BI and discovery tool such as Tableau or PowerBI to create dashboards and explain the results to the business
- Operationalize the resulting data pipeline to continually send fresh new results to yourself and business teams
- Enable the workbook and resulting data to be shared with colleagues while having the artifact managed and governed inside your data lake by Datameer
Optimal Blend of Search and Column Storage

OLAP-on-Hadoop suffers from one major problem – they rely on data from underlying SQL engines, typically Hive, which is extremely slow. That dependency, with its high latency, has a ripple effect through OLAP on Hadoop engines, preventing them from being able to dynamically generate indexes and do fast scans.

This requires the typical OLAP on Hadoop engines to pre-build and compute cubes, with highly bloated indexes that cover all possible dimensions and values, and aggregations that are materialized. Caching these materialized cubes helps performance, but the bloated indexes and pre-defined models limit the ability to explore the data and force analysts down pre-determined paths.

The dynamic indexing architecture of Visual Explorer delivers the scale, speed and efficiency needed to enable a large number of concurrent users explore large datasets and drill into the details. Highly focused indexes are generated on the fly based on the data being explored for maximum efficiency and rapid sub-second response.

The secret to the speed, efficiency and free-form exploration capabilities of Visual Explorer comes from a highly integrated combination of:

- A highly tuned distributed search technology that focuses on optimizing the generation of search indexes on data
- The use of optimized columnar data management that makes it extremely fast and efficient to scan the data in real time, even at scale

The two features complement each other. The distributed search technology enables highly focused, easily navigated indices that can be generated on the fly as the user explores certain areas of the data. Because the “queries” are highly focused, the engine can then do “micro-scans” of the data, grabbing and aggregating only the data needed for that exploration path.
These two extremely efficient approaches enable the Visual Explorer Dynamic Indexing Server to deliver results with sub-second response times on extremely large datasets. But it also contributes to the free-form exploration as the indexing on the fly and extremely fast micro-scans enable agility to explore in any direction, on any attribute, value and metric in the dataset.

**Hadoop Integration for Scale**

The Visual Explorer back-end engine runs as a container on Hadoop to gain the compute and storage scalability it provides. Execution requests are routed via YARN, providing access to the vast resources available on the cluster, while allowing it to properly share resources with other applications on multi-tenant clusters.
Integrated Platform

The final secret to delivering successful interactive data exploration at scale is an integrated platform that eliminates the resource, process and data flow inefficiencies of a multi-product stack.

Through its integration with the Datameer platform, Visual Explorer provides the last cog in a comprehensive big data preparation, curation, and exploration big data platform that:

- Manages your full data lifecycle from ingestion through consumption to streamline the creation and delivery of data assets;
- Uses optimal data retention policies to eliminate the duplication of data and the accompanying wasted resources;
- Offers a single governance and security model that simplifies the process of securing the data while fostering collaboration and sharing;
- Delivers a single execution and operationalization model that optimizes the delivery of information through data pipelines.

The combination of Visual Explorer, the easy spreadsheet-style preparation paradigm, and enterprise-grade capabilities of Datameer provides a single platform that can transform your raw, complex data into business ready data assets for your entire organization.

Conclusion

Are you ready to explore your data at the speed of thought on the industry’s first interactive visual exploration platform that can scale across even your largest data needs? Rather than bring the data lake to your analysts, bring your analysts to the data lake, letting them explore all your data at will and to derive greater value from your data and solve tough analytics problems faster and easier. To learn more, please visit www.datameer.com/visualexplorer.