ESG Lab Review

Efficiently Offloading and Optimizing BI Workloads to Hadoop with Cloudera Navigator Optimizer

Date: May 2017 **Author:** Mike Leone, Senior Lab Analyst

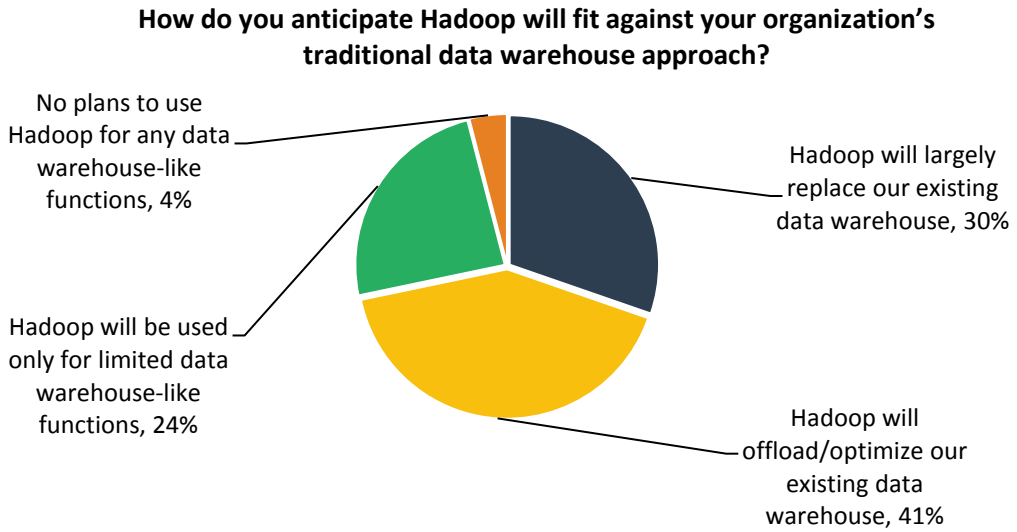
Abstract

This ESG Lab Review documents the recent analysis of Cloudera Navigator Optimizer with a goal of validating its ability to enable organizations to identify and offload SQL workloads from legacy data marts and costly enterprise data warehouses (EDWs) to a modern analytic database built on Hadoop, while optimizing existing workloads already on Hadoop. The report also highlights MicroStrategy, a comprehensive business intelligence platform integrated with Navigator Optimizer, which can seamlessly support this migration without impacting existing BI reports, dashboards, or applications. While the focus of this report is on BI offload and optimization, Navigator Optimizer also provides similar capabilities for ETL workload migration.

The Challenges

Enterprise data warehouses have long been the norm to handle analytic processing, but limitations related to cost-effective scalability, data flexibility, and accessibility have nearly forced organizations to consider complementary offerings and platforms. Hadoop-based platforms are one such option and with their rapid adoption and maturation in the big data and analytics segment, organizations are modernizing their infrastructures to better meet their business needs. By combining the power of traditional EDWs with the scalability, flexibility, and open architecture of Hadoop, organizations can extend the value of their data warehouse landscape and do more within their existing footprint. Further, with Hadoop specifically, the potential to converge data silos and data marts onto a single, scalable platform has organizations excited about the potential reduction in workload inefficiencies, management complexity, and cost. In fact, ESG research shows that there is a willingness to pursue an approach that leverages both legacy and modern platforms for data warehousing and business intelligence. As shown in Figure 1, 30% of enterprises plan to largely replace their existing data warehouses with Hadoop, while an additional 41% are looking to offload and optimize their existing EDW workloads by leveraging Hadoop as a complementary platform.[1]

Figure 1. How Current Users Anticipate Hadoop Affecting Traditional Data Warehouse Approach



Source: Enterprise Strategy Group, 2017

[1] Source: ESG Research Report, *Enterprise Big Data, Business Intelligence, and Analytics Trends: Redux*, July 2016.

Though accessibility improvements and optimizations to existing analytic workloads on a cost-effective, flexible platform are the hopeful outcomes, transitioning from an EDW to Hadoop does not come without its challenges. Which workloads are ideally suited to each platform? How much of a workload can be offloaded easily and what changes are needed to offload more complex portions? Are there existing commonalities across queries that can be leveraged? What are the development resources needed to make them compatible with the new platform? Is there an easy way to prioritize existing workloads to be offloaded? Answering these questions becomes even harder once you factor in the total number of queries being run on a regular basis, which can fall in the millions for many large organizations. In the end, organizations are looking for ways to simplify and automate the process of offloading and optimizing BI workloads, enabling use of the right platform for the right workloads to gain better insights faster and get more value from their data warehouse landscape.

The Solution: Cloudera Navigator Optimizer

Cloudera Navigator Optimizer is a software service included with Cloudera's Hadoop-based analytic database to help organizations quickly, confidently, and securely gain more valuable insights from their SQL workloads. Navigator Optimizer profiles existing BI queries and provides analysis, guidance, and understanding into those workloads. This enables key personnel like big data architects to identify and offload queries from EDWs and data marts to Hadoop, and, for ETL and BI workloads already running on Hadoop through Hive or Impala, Navigator Optimizer can help database administrators stay ahead of optimization needs so these workloads run more efficiently.

By providing Navigator Optimizer with query text and other necessary information extracted from query log files from database or BI tools, query workloads are quickly analyzed. The outcome is rapid insight into a various key areas: duplicate queries, sets of similar queries, common query patterns, join complexity, and query syntax compatibility with Hadoop-based SQL processing and analytic query engines, Hive and Impala. More importantly, offload risk analysis gives users a clear picture of what access patterns can be risky for the new platform if not fixed. Optimization recommendations such as schema design and query simplification gives users actionable solutions to run the workload safely on the new platform. Further, once those queries have been offloaded, Navigator Optimizer provides performance optimizations and recommendations based on Hive and Impala best practices.

Navigator Optimizer enables big data architects commonly responsible for the overall infrastructures and what goes where to improve operational efficiency by providing insight and prioritized guidance into what workloads and underlying data sets are best to offload. Risk assessments highlight the risk of offloading workloads to Hive or Impala based on compatibility and complexity, as well as offer advice to the architects on suggested improvements across both the physical infrastructure and business workflows. Once the workloads are migrated, the management responsibility shifts to the DBAs for ongoing customization and optimization of the newly migrated workloads, ensuring everything runs as efficiently as possible. Altogether, big data personnel gain peace of mind knowing Navigator Optimizer can provide the information they need to make their day-to-day tasks easier as the business takes advantage of these big data platforms, enabling more time to prioritize more strategic initiatives within the organization.

MicroStrategy

MicroStrategy is a comprehensive business intelligence platform for enterprise analytics and mobility applications at scale. From data discovery to operational reporting, interactive dashboards, predictive analytics, automated distribution and mobile productivity apps, MicroStrategy supports the full breadth of business intelligence capabilities with a single comprehensive platform that is available both on-premises and in the cloud.

With MicroStrategy, any user in the organization can access, blend, visualize, and analyze information from a variety of sources (including relational sources, Hadoop, cloud systems, personal spreadsheets, and others), and make actionable insights in minutes. Partnering with Cloudera, MicroStrategy delivers access to the modern Hadoop-based platform with native and optimized connectors, as well as built-in workflows for data preparation and data blending. With Impala as the underlying SQL analytics engine, analysts can easily perform complex data analytics and visualization with the click of a button, all with their accustomed speed and concurrency. By combining the scalability and flexibility of Cloudera's platform, and the power of MicroStrategy's enterprise analytics, organizations gain a robust infrastructure and set of services to meet even the most sophisticated business intelligence needs of large organizations.

With this integration, Navigator Optimizer and MicroStrategy can together be leveraged to migrate workloads from legacy systems to a modern analytic database, without disrupting the business. Navigator Optimizer provides the insights and recommendations for prioritized migrations, and MicroStrategy allows organizations to easily rewire the BI application to the new Hadoop-based data source. This ensures these reports and dashboards continue to run seamlessly and analysts maintain productivity, regardless of the underlying platform.

Migrating BI Workloads to Hadoop

It is important to plan the migration of BI workloads from EDWs to Hadoop before completing the task of offloading and optimizing. ESG reviewed the Cloudera-recommended process for a successful migration, which consisted of four phases: evaluate, plan, offload, and optimize.

- Evaluate – Identify the use cases and objectives for needing to offload, whether it's related to cost, scale, performance, or modernization. Then identify the scope of what should be offloaded, based on workload activity, access, characteristics, and proofs of concept to understand this process.
- Plan – Establish a prioritized project plan through impact analysis to understand the required development changes, effort, and capacity.
- Offload – Upload query workloads, run risk analysis based on offload unit (report, application, or user), design the schema for the platform, implement, and then validate success.
- Optimize – Fine-tune the data models of offloaded BI workloads for optimized performance, and validate the ongoing cost savings and return on investment based on initial objectives.

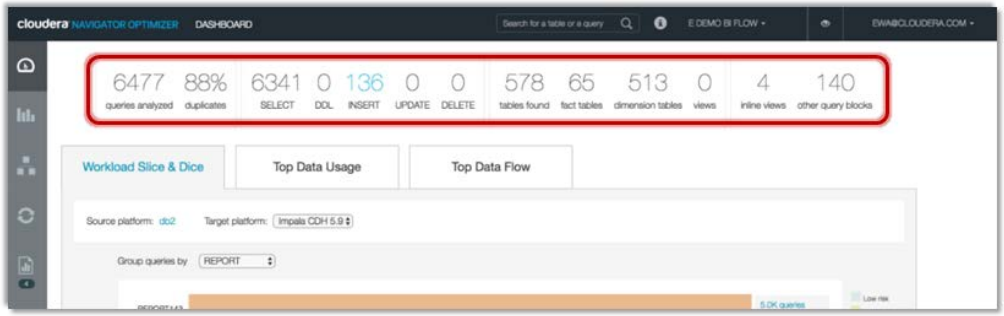
Though Navigator Optimizer is not meant to serve as an end-to-end tool that covers all phases of the migration process, it is meant to effectively and efficiently provide the right insights and recommendations in key areas of the overall migration process.

Analyzing BI Workloads and Providing Instant Insight

ESG walked through a typical interaction with Navigator Optimizer to understand the efficient process of uploading and analyzing BI workloads. It should be noted that the process of analyzing BI workloads to identify either what to offload from EDW to Hadoop or what to optimize based on what is already running on Hadoop uses a similar process. For this section, ESG focused on analyzing MicroStrategy workloads running on a legacy system with a goal of prioritizing specific queries to offload to Hadoop. Upon logging into the Navigator Optimizer dashboard, ESG was presented with a user-friendly interface customized to a sample user. An existing workload was already uploaded. Workloads serve as a container for analyzing a group of queries together. As shown in Figure 2, the workload contained thousands of queries from daily logs of all reports. It is made up of mostly SELECT statements across hundreds of tables.

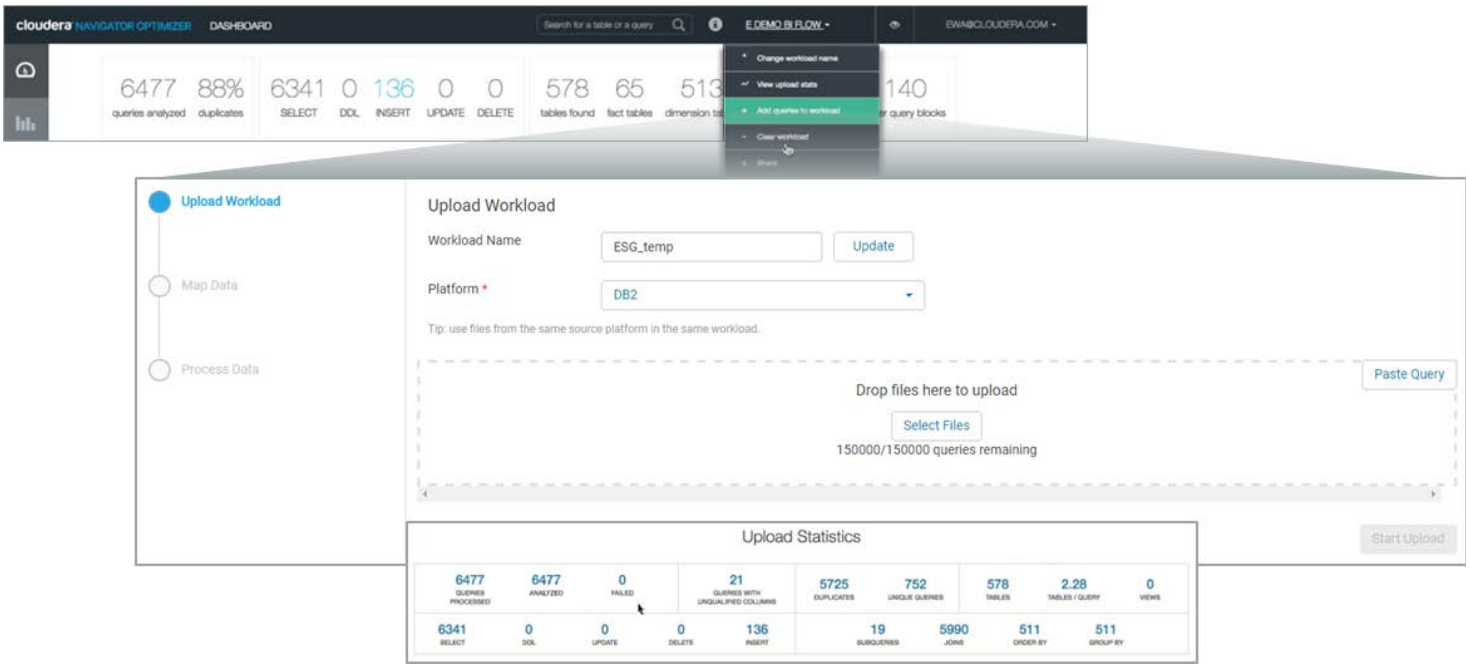
Though data was uploaded into Navigator Optimizer prior to the demonstration, ESG viewed the interface used to add new workloads and queries to analyze in Navigator Optimizer. As shown in Figure 3, an intuitive interface was displayed that guides users through the workload upload process. The guidance includes sample scripts that simplify the extraction of query workloads as well as database statistics in the proper format to be uploaded to Navigator Optimizer. This includes selecting the source platform, such as Teradata, Oracle, and Netezza when offloading, or Impala and Hive when optimizing. Cloudera uses a step-by-step wizard to make the upload process simple. Finally, once the upload process has completed, statistics are available based on the uploaded data set, including the number of queries, statement types, number of tables, etc.

Figure 2. Navigator Optimizer Dashboard



Source: Enterprise Strategy Group, 2017

Figure 3. Uploading Workloads to Navigator Optimizer

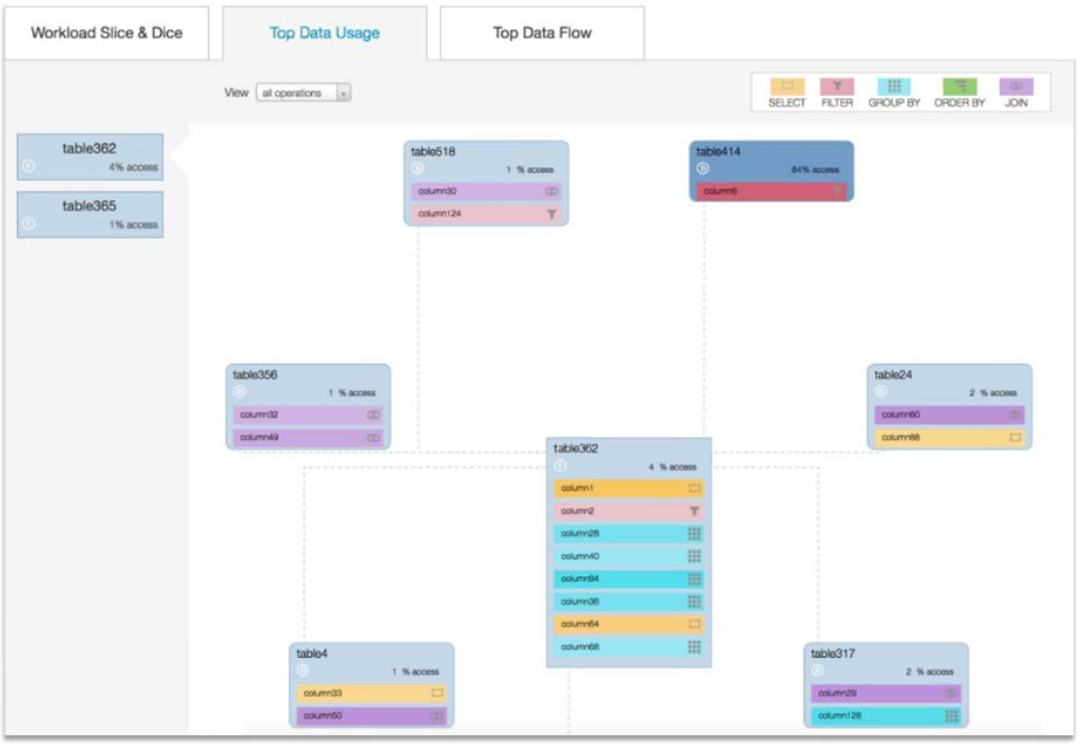


Source: Enterprise Strategy Group, 2017

Once the data is uploaded, Navigator Optimizer automatically completes its analysis and provides instantaneous insight into the workload. From the main dashboard, the *Workload Slice & Dice* tab displays risk assessment of the workload. The *Top Data Flow* tab takes all the queries and displays a top-down data flow based on the selected group of queries.

As shown in Figure 4, the *Top Data Usage* tab leverages the data usage information inside of queries to construct a multi-dimensional E-R diagram without the need for DDL. Further, specific columns within each of the mapped tables are color-coded based on query type—select, filter, group by, order by, and join—to provide organizations with a granular view of not only how the tables and columns are queried, but also how frequently they are accessed. This level of insight allows organizations to quickly and easily prioritize workloads purely based on usage.

Figure 4. Uploading Workloads to Navigator Optimizer



Source: Enterprise Strategy Group, 2017

Why This Matters

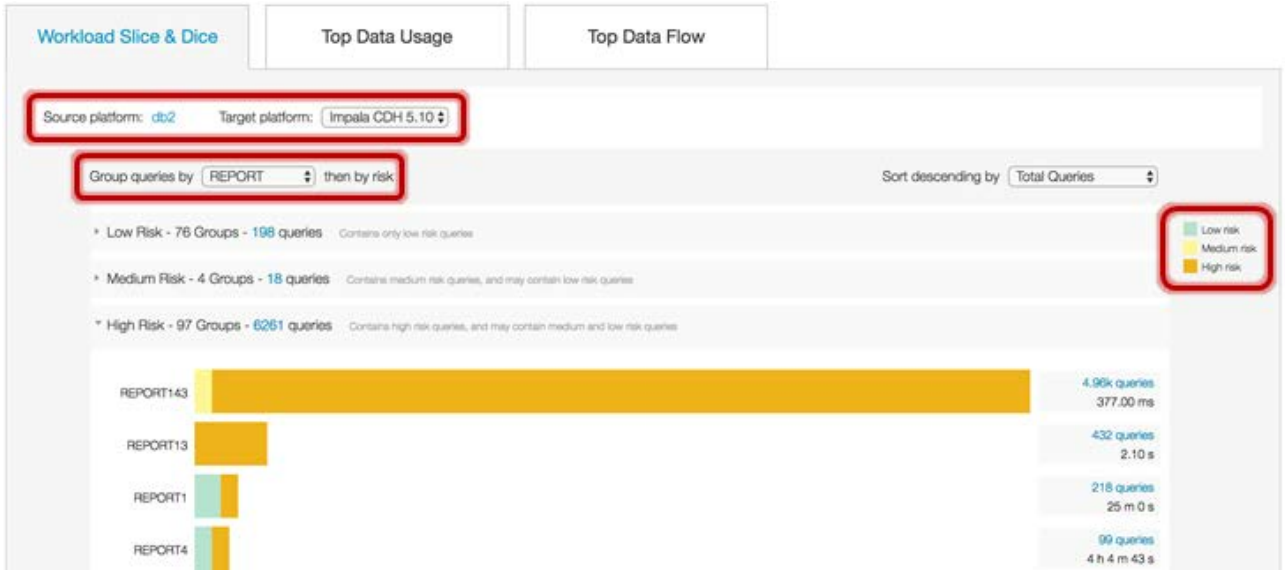
Manually analyzing a single query is easy. Trying to sift through hundreds of thousands of queries to identify and prioritize what to offload to Hadoop is nearly impossible, especially in a timely manner—never mind the inability to interconnect queries to show common access patterns and usability in an easy-to-understand format. Organizations need an easy and fast way to gain insight into the usage of their BI platform(s) to not only help extend the value of their infrastructure usage, but also to improve business processes and workflows, as well as to continue modernization to not get left behind.

ESG validated that Cloudera Navigator Optimizer provides detailed analysis of BI workloads within minutes. Organizations gain rapid insight to help identify and understand key information about the BI environment, such as granular database usage details down to the column of a specific table, the number and types of queries accessing it, and the level of query duplication that exists throughout an organization. Together, this enables organizations to balance tradeoffs and focus their efforts based on business impact, making the offload process from an EDW to Hadoop as effective and efficient as possible.

Using Risk Assessment to Prioritize BI Workload Offloads

The next phase of ESG's validation of Cloudera Navigator Optimizer focused on analyzing BI workloads using risk assessment. Using the *Workload Slice & Dice* tab in the main dashboard, ESG used a DB2 source platform consisting of numerous MicroStrategy reports. The workload was grouped by MicroStrategy report, and each horizontal bar represented the level of risk each query within each report posed if it were to get offloaded as-is to the destination platform, which in this example was Impala. This is all shown in Figure 5.

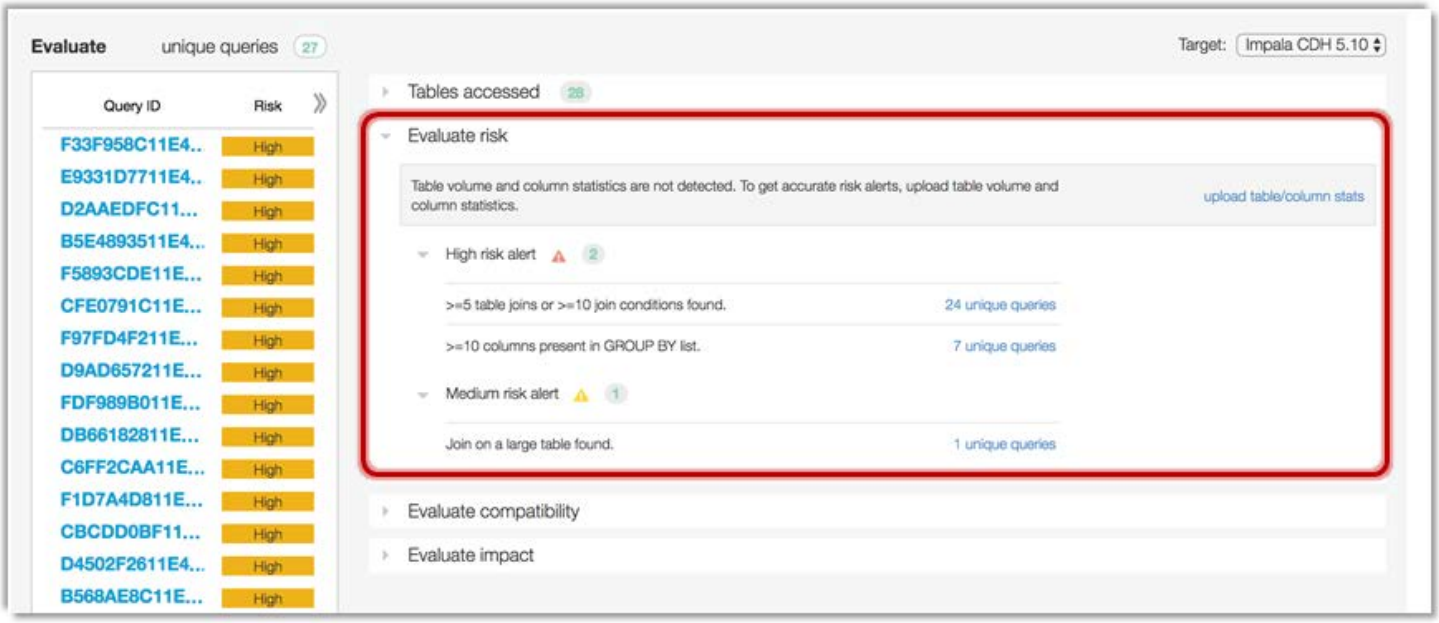
Figure 5. Offloading BI Workloads by Risk Factor



Source: Enterprise Strategy Group, 2017

ESG navigated to the design tab and viewed the group of queries from Report143 to further assess the risk associated with offloading them. As shown in Figure 6, 27 unique queries were identified as part of the group, all of which had a high-risk designation. By navigating to the *Evaluate Risk* dropdown, ESG identified the exact reason for these queries being high risk. Twenty four of the queries contained a large number of table joins and seven of them contained a large GROUP BY statement. Both conditions are predicted to execute poorly on Impala. By scrolling over the two alerts, Navigator Optimizer offers recommendations to help fix the queries. For the queries with many table joins, Optimizer suggests denormalizing the tables, while the large GROUP BY statements can be optimized by evaluating the memory requirements of each query. Additionally, one medium risk alert was displayed due to a large table join being discovered.

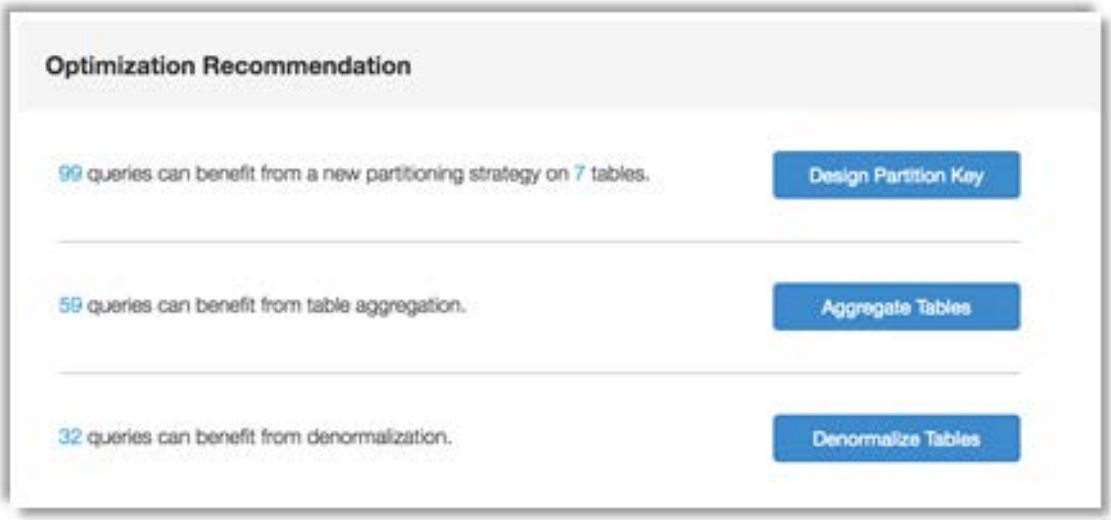
Figure 6. Evaluating Risk of Queries



Source: Enterprise Strategy Group, 2017

Based on the recommendations, ESG continued down the path of understanding the optimization steps required to offload specific workloads with higher risk. This is a very important aspect of the platform to highlight because Cloudera Navigator Optimizer does not just provide insight, but can also provide a granular view and detailed optimization recommendations on which to act. Common recommendations include denormalizing tables, aggregating tables, and partition key design. ESG explored an optimization recommendation for both a high-risk and low-risk report. A view of how the optimization recommendations appear in the interface is shown in Figure 7.

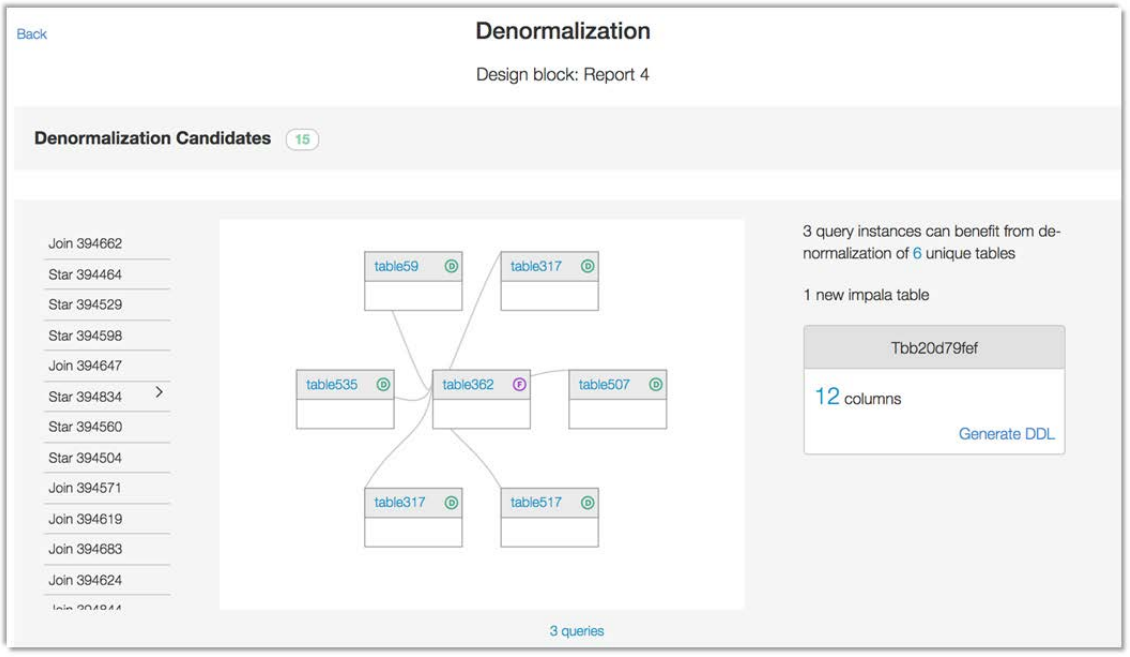
Figure 7. Optimization Recommendation View



Source: Enterprise Strategy Group, 2017

By clicking on one of the recommendation options, a specific module based on that recommendation is displayed. For the high-risk report, due to there being a large number of table joins, Navigator Optimizer recommended denormalizing the tables. Details related to the candidates and query instances that could benefit from denormalization as well as a diagram of the unique tables that would be denormalized were shown. Further, with the simple click of a button, new DDL could be generated from directly within the interface. A view of the denormalization module is shown in Figure 8.

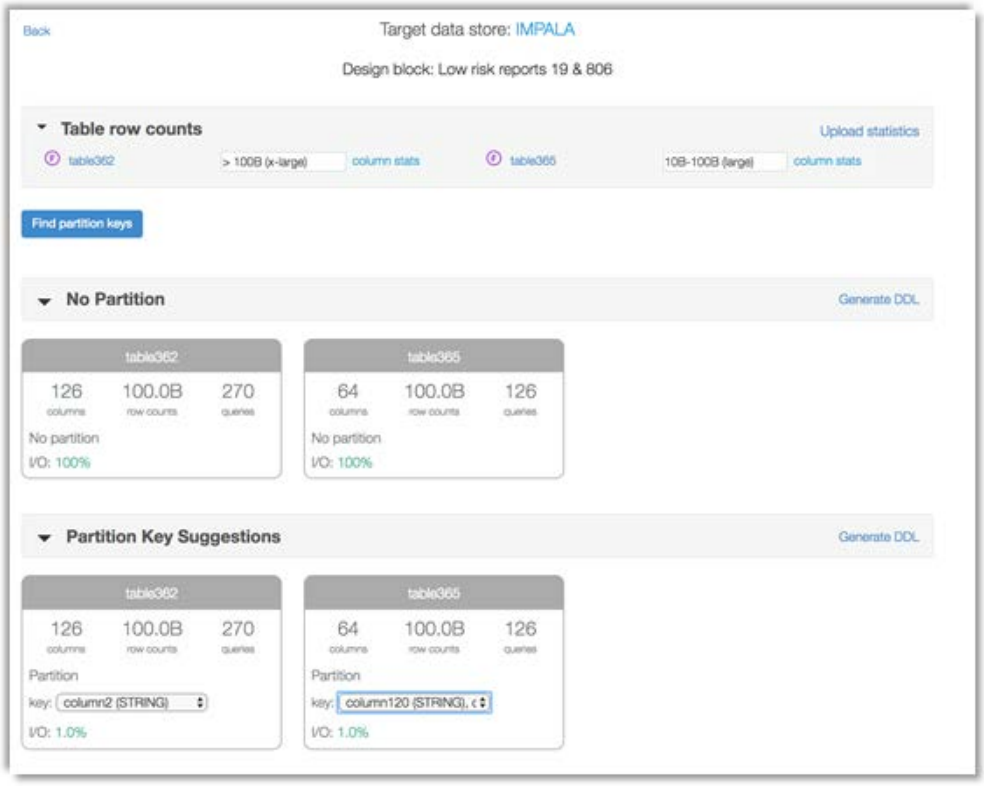
Figure 8. Optimization Recommendations – Denormalization



Source: Enterprise Strategy Group, 2017

For the low-risk report, Navigator Optimizer recommended designing a new partitioning key. The partition key design module was launched that, similar to the denormalization module, showed all the details related to the selected workload, including all filters that were detected by Navigator Optimizer, as well as the number of queries that could benefit from the use of a new partitioning key on the selected table. Further, by clicking on ***Recommendations*** from within this module, ESG could import saved table statistics and then actual partitioning keys would be recommended based on those table and column stats, as well as popular filters. A button could then be clicked to generate a new DDL, which consists of a new table with the optimal partitioning key, enabling organizations to immediately use the new table to benefit from the optimization. Figure 9 highlights the partitioning key recommendation interface to upload table statistics.

Figure 9. Optimization Recommendations – New Partition Key



Source: Enterprise Strategy Group, 2017

Why This Matters

Identifying BI workloads better suited for a modern platform like Hadoop is the first step for big data architects. Next comes the work of offloading and optimizing that workload to best utilize the destination platform. Not only must each query be compatible to ensure seamless productivity for analysts, but as usage and data needs change, it's critical to actively optimize data models to meet the business needs. Are there too many table joins? How large are the GROUP BYs? These are just two of the many questions administrators must have answers to before opening up access to analysts, or their inboxes will be littered with requests for help.

ESG confirmed that Cloudera Navigator Optimizer enables the offloading and optimization of BI workloads to efficiently run on Hadoop with Hive or Impala. Whether assessing risk, compatibility, or impact, Navigator Optimizer provides detailed insights to administrators to ensure workloads are offloaded, optimized, and ready to perform as expected on the platform of choice. Further, for queries that have high levels of risk associated with them, Navigator Optimizer provides detailed recommendations on ways to optimize them and in some cases, directly take action from within the user interface.

The Bigger Truth

As organizations look to extend the value of their data warehouse landscape to better address self-service business needs, it's critical for them to understand how a modernized platform can relieve pressure from traditional EDWs, while also consolidating workloads from the proliferation of data silos. However, identifying and prioritizing the workloads to migrate can be a massive undertaking. Where should an organization begin when developing its offload strategy? Which workloads are better suited for the new platform? How can complex analytic workloads be offloaded? Being able to efficiently and automatically analyze BI, analytic, and ETL workloads for instant insight into all queries and database interactions down to the column is essential for a successful and predictable migration, as well as platform adoption and continued success within the new modernized landscape.

Cloudera Navigator Optimizer enables organizations to relieve the operational and cost pressure from legacy EDWs, disparate database silos, and expensive data marts by quickly and easily identifying workloads best suited for a Hadoop-based platform. And, with integrations with partner technologies such as MicroStrategy, organizations can rest assured that there'll be a seamless transition of existing reports, workloads, and skills so the same tools can be used across complementary systems. Cloudera's technology enables big data architects and DBAs to eliminate the task of manually sifting through query logs by allowing the import of the entire SQL workloads. Insight is achieved nearly instantaneously, with Navigator Optimizer creating workload flow charts, E-R diagrams, and granular table and column interactivity statistics to efficiently identify, offload, and optimize queries based on usage, priority, and compatibility. Workloads are also analyzed by risk, and Navigator Optimizer provides detailed recommendations on ways to adjust them based on best practices, whether it is Hive for flexible ETL/data preparation or Impala for high-performance BI and SQL analytics. DBAs can take this advice a step further, and act based on those recommendations directly within the software's user interface, whether denormalizing tables or creating new partitioning keys for more efficient query execution.

Organizations no longer have to put off upgrading their legacy infrastructures for fear of downtime, inefficiency, or incompatibility by moving to a new platform. Let Cloudera Navigator Optimizer do the heavy lifting, by providing insights and assistance throughout the offload process, so organizations can start getting more value from their technologies, faster.