Modernizing Business Intelligence for Big Data

An Ovum white paper

Publication Date: June 2016

Information Management Practice, Ovum

Summary

Catalyst

When it comes to providing insights to business users, time-to-value has always been the challenge – and promise – of business intelligence (BI) systems. The challenge is magnified when working with big data, which traditionally has required specially trained practitioners to write complex MapReduce programs to find the signals and ask the right questions. But until recently, access to analytics on this data was restricted to programmatic approaches. As Hadoop has evolved into a multi-purpose platform that can handle multiple types of analytics, it has been opened up to the large base of practitioners with SQL skills.

That triggers a clear question: Could Hadoop become more accessible using the BI tools that have become a staple of analytics for many enterprises, and could it become accessible in a manner that delivers faster time-to-benefit?

Ovum view

The addition of interactive SQL access to Hadoop promises to open the floodgates to big data analytics that is accessed with common, off-the-shelf BI tools. Hadoop is now 10 years old; it is Ovum's view that the technology's maturity, and its use as an operational tool in enterprises, is accelerating rapidly.

Hadoop could help business users gain better perspective when answering familiar questions surrounding use cases such as Customer 360, Internet of Things, process efficiency, and risk management by enriching it with new data sources such as that coming from social media, logs, and connected devices.

Ovum believes that pairing Hadoop with these types of tools will be low-hanging fruit for BI users as advanced visualizations will aid in finding patterns in larger and more varied data sets. And with technologies such as OLAP (online analytical processing), which helps improve the analysis performance and can provide a business semantic layer in front of often complex data, Hadoop could become even more accessible.

This paper is intended to help guide readers through the key concepts that will help them consider ways to modernize and optimize their current BI investments for big data on Hadoop.

Key messages

- Business intelligence approaches need to be updated in order to take advantage of the breadth of insight offered by big data.
- SQL is rapidly becoming the preferred way for experienced BI users to access Hadoop, because BI tools and users are well versed with SQL.
- Achieving time-to-benefit for experienced BI users will rely on providing superior access to big data via the tools to which they are well accustomed, for example Excel, Tableau, or Qlik, among others.

The state of BI today

BI first emerged in the early 1990s with the promise to make decision-making smarter and easier. But during the last twenty years, BI's limitations have become apparent. BI works really well when the business requirements are understood and reoccurring, and when the data sets are modest and come from enterprise applications such as ERP, CRM, or other transactional sources, where the universe of data is known and well bounded. These sources can readily be modeled and cleansed into structured schemas (e.g. tables and rows) that can be loaded into a data warehouse (DW) which can be used for common BI deliverables such as reporting, dashboards, OLAP, and visualizations.

The appeal of SQL

SQL databases proved a natural target for BI tools as they provided a much more logical, intuitive means of access compared to legacy data sources. Unlike legacy data stores, where data was tied to the application and underlying hardware and storage (requiring programs that specified the physical location of data), relational databases were theoretically machine-independent. Data could be accessed declaratively (e.g. the user did not have to specify how to access data, just what data should be accessed and in which form), based on the data's logical structure. Also, with time, vendors widely adopted SQL as a standard in their products, spurring an ecosystem of practitioners and analysts.

The changing market for BI

BI has, traditionally, been procured by IT, sometimes in conjunction with either the finance or marketing department – the equation is changing and an ever-greater proportion of solution acquisition is being driven by the business.

- **Business buyers take the lead –** This is the latest manifestation of a long-term trend. While IT continues its role of "keeping the lights on" and seeking means for lowering running costs to support more innovation, value-added solutions that add to the top or bottom line are those that get greenlighted.
- **Adapt governance to handle more diverse sources of data –** Despite new buying habits and the need for more information faster and to more users, the same requirements apply. Providing "freedom within a framework," i.e. self-service with governance, is an essential component for success. Businesses require a flexible governance model to support self-service and need to tackle governance in a much more modular fashion than ever before. From a strategic perspective, this will require organizations to adopt a "divide and conquer" strategy with data to achieve some degree of control. Instead of providing heavily guarded access to one curated store of data, new approaches have to work with multiple data sources and stores and apply some measure of governance (access control, data quality, profiling) at the source – in near-real time, while importing the data, or inside the analytics solution itself.
- **Does the combination of self-service and business buyers equal "bring your own BI"? –** If lines of business are selecting their own analytical tools, then the role of IT must change to allow access to data sources – big and small – to those tools, without compromising standards of governance and control. Bring your own BI (BYOBI) is a term favored by some in the industry as an expression of this, descriptive of the changing BI landscape and an attempt to ease the long-standing tension between the business and IT.

- **Does more visualization mean more insights? –** Not necessarily. Visualization brings many benefits to analytics – easy-to-consume information and greater accessibility being the prime examples. However, it must come with a warning – good-looking visualization of data does not always equal actionable insight. Visualization must be considered a valuable enabler to greater insight, in other words, a means to an end, not the end itself.

Why BI on Hadoop?

Why should you care about BI on big data?

The value of delivering insight from data – the heart of BI – is as applicable to big data sources as it is the well-understood, structured sources most organizations are accustomed to working with. Several trends are driving an interest in expanding BI's remit to incorporate those new sources:

- **Time-to-value is shortening –** The need to interact with customers and partners with greater immediacy, acting on opportunity as it occurs rather than observing it after the fact, is a driving force in all industries.
- **New sources of data –** Adding Hadoop to the BI stack enables BI users to go beyond aggregated and structured data sets from transactional sources, enabling new use cases and generating more insights. Ranging from Customer 360 to risk mitigation, fraud detection, or operational efficiency, these uses can be driven by nontraditional sources such as mobile and device data, log files, messaging, and social networks – in addition to traditional transactional data sources, and without having to move the data from its Hadoop store.
- **Greater scale and depth of data can grow insight –** There are solid arguments to be made that suggest greater volume of relevant data can deliver better insights for the enterprise.

Where does Hadoop enter the big data debate?

As an analytic/data warehousing platform, Hadoop provides the scale, flexibility, and economic storage that picks up where traditional relational data warehouse platforms leave off. Specifically, Hadoop delivers near-linear performance, documented up to thousands of nodes, along with the cost advantages of leveraging commodity processor, disk, and network interconnects. Yet, the flexibility of the Hadoop platform could yield its most profound benefits, as it can be utilized for multiple purposes. Some of the key benefits include:

- **Schema flexibility –** While Hadoop does not eliminate the need to model or structure data, it provides the freedom for organizations to take advantage of "schema on read." And that frees users of business intelligence tools to generate insights on new forms of data that would otherwise not be easily accommodated inside traditional data warehouses.
- **Economics –** Hadoop's reliance on commodity infrastructure makes it well-suited as a target for data consolidation. It can service multiple audiences – from business analysts running BI reports to power users running exploratory queries and data scientists who can run highly complex analytic problems involving very large, heterogeneous sets of data.
- **Extensibility –** Storage is inexpensive enough that modeled and raw data can sit side by side.
- **Active archiving –** With inexpensive storage and compute cycles, it becomes feasible from a cost perspective to keep aging data online and available for analytics.

- **New analytic options –** As Hadoop becomes a multifunctional platform, it allows business users and data scientists alike to take advantage of the languages and tools that they prefer.

Hadoop adoption is a growing and disruptive force in information management – based on our estimates, the Hadoop market is growing at approximately 50% year on year. Ovum sees the Hadoop installed base beginning to grow past the classic early-adopter base, driven by emerging, commercial use cases and the factors outlined above.

Why change your approach to using BI on Hadoop?

Ovum considers several key trends shaping the world of data and analytics as being indicative of a need to bring long-standing BI capabilities to emerging Hadoop-based capabilities:

- **Data discovery and the big data pipeline –** We are living in the golden age of data exploration and discovery. More users are demanding access to exploratory analytics than ever before. The new data sources that are becoming available to enterprises must be made accessible, not only to data scientists, but to business users who can operationalize insights from new sources that can give them the bigger picture.
- **Business groups' insatiable appetite for self-service –** The success of self-service BI visualization tools revealed not only that self-service is feasible, but that there is huge latent demand for it. Business users are chomping at the bit to have access to information without having to constantly wait for IT to prepare materialized views for their changing needs.
- **When traditional BI tools are used directly on Hadoop, they may reach their limits –** These tools, which rely on static schema, cannot keep pace as data is flowing from new sources, and analytic needs grow more volatile. New approaches are essential for providing the freedom to explore data, as problems – and data sets – change.

Key considerations for BI on Hadoop

In preparing to conduct analytics on Hadoop, organizations need to be aware that several key design guidelines depart from those of traditional data warehousing:

- For starters, the volume of data, and the variety of data types and structures, will outstrip what would normally be populated into a data warehouse. Compared to data warehouses, where schema is built into the core design, Hadoop can accept data without having to structure it in advance; instead, schema is generated when the data is required for analysis (schema on read). Once the analytics for that data is operational, it becomes appropriate to transform data into a consistent schema.
- Another key consideration is minimizing or avoiding the movement of data; when contending with data at terabyte scales, the overhead of data movement becomes prohibitive. Instead, the common design practice is to move the analytics to the data.
- The benefits of OLAP come in more than one flavor: OLAP makes data easy for business users to consume, and offers interactive performance for the types of queries that the BI tools generate. There are arguments to be made for multidimensional OLAP (MOLAP) and relational OLAP (ROLAP) approaches. Some suggest MOLAP does not scale well, making it challenging to use in the big data world; ROLAP does not, it is argued, share this challenge.

Others will note that MOLAP outperforms ROLAP. Careful consideration is required when exploring these approaches.

- Semantic layers and one version of the truth – The likelihood is that the data architecture will contain several semantic layers, with big data and data lakes being yet another layer. In this heterogeneous environment, the value of having a unified semantic layer for BI and Hadoop will be an important consideration – tying together analysis on different types and sources of data.

How Hadoop augments BI

Using EDW skills to work with Hadoop

Regardless if companies use an EDW (enterprise data warehouse) or a Hadoop data lake in exclusivity or in concert, enterprises will still need to collect, process, and create structured data for BI's many processes. Every organization needs a way to supply clean, standardized, and aggregated data to BI to generate reports, dashboards, and OLAP cubes successfully.

Because Hadoop can store data in virtually any form, it can supplement (and enrich) aggregated transaction data of EDWs with data from a multitude of sources (e.g. messaging, social media, clickstreams, and videos, along with location and other device data). These "other" data feeds can enrich the transactions that traditional EDWs store.

Getting better context

More data – and more diverse sources of data – improves context. For example, if a sales manager wants to compare sales of a seasonal product – e.g., ice cream – globally, he or she can start with aggregated and analyzed transactional data in a traditional BI environment. The sales manager can see what specific region did well, and which did not, and take action accordingly. The problem, however, is that this approach explains the "what," but not as much the "why." That is, why did ice cream sales perform the way they did?

To expand the analysis, the sales manager might want to include weather data and social media sentiment to better understand how both factors affected sales, whether there is a correlation anywhere, and what actions the organization can take to improve sales tactics in the future. For instance, social media can help pinpoint not only consumer sentiment, but also provide hints of external events that might impact sales, such as local events (e.g. local sports teams' championships or local appearances by pop stars) that might not otherwise be captured by internal transaction systems. Hadoop can readily accommodate data from a variety of sources, from weather sensors to social networks or other external data feeds, to produce a more complete analysis.

This type of analysis, which hinges on Hadoop being part of the BI stack and acting as an additional data source for BI, will require organizations to make architectural modifications.

Getting a longer view

In Hadoop, the costs of keeping aging data live are relatively marginal because the open source software does not have licensing fees (clients typically pay annual subscriptions for support) and the commodity hardware used is relatively inexpensive to add. Keeping archived data online is often termed "active archiving." The chief benefit of active archiving is that it allows analysts to extend the period of time covered by analytics.

Getting a collaborative view

Using the multifunctional capabilities of the Hadoop platform, domain experts, data scientists, and business analysts can collaborate on ferreting out signals from the noise. As data scientists explore patterns in the data, they can share their insights with domain/subject-matter experts and business stakeholders who can zero in on the relevant questions. The result can be operationalized reports that can be widely accessed and consumed by business analysts.

BI tools and techniques to help frame these questions

To ask these types of questions, an analyst needs to understand what analytics tools are available and how they might play a part. The insights gained can then be operationalized with materialized views optimized for reporting.

Self-service is a well-understood concept in the world of BI; its application to BI on Hadoop should come as no surprise. Much of the debate surrounding the relative pros and cons of self-service has been more recently focused on the issue of governance versus the value of freeing users to work with data. There are two key points to consider:

- **Taking the burden off of IT while providing proper governance –** This is a collaborative process between the business and IT, where end users collaboratively manage data sets through data preparation approaches, while IT ensures that data sets are properly stored and secured. This is essential because IT will not have the bandwidth to constantly prepare data each time an end user or group of end users wish to explore a new dimension.
- **Enabling business users to ask more questions faster, while managing growing data complexity –** As business end users are given new options for exploration, it is not realistic to expect them to possess database administration skills for creating schema. Business users must be enabled to create new views without the need to rely on IT.

Data discovery and search can be the first analytic step

In a traditional BI environment, where the data is coming from transactional sources such as ERP or CRM, the data is typically well understood. But as data sets grow and become more complex, it will become more difficult to actually understand what data is actually available. That is where Hadoop comes in; it provides the platform for exploring new forms of data that can enrich traditional BI/query and reporting views.

Not surprisingly, one of the hottest areas in BI during the last couple of years has been data discovery. Data discovery allows users to see patterns in the data that might otherwise not be caught if kept in a tabular format, for example. Data discovery and Hadoop are destined to work together. Data discovery tools turn large sets of data into consumable visualizations that can be explored and interacted with rather intuitively. In particular, data discovery can be a great first step for a user to sift through a slew of data to better understand:

- What kind of data does the organization possess?
- Are there any patterns worth noting?
- Any important outliers or anomalies?

Who can benefit from BI on Hadoop?

BI systems are not used, built, or monitored by just one type of worker. Instead, a variety of users with different roles, capabilities, and requirements work within the BI ecosystem. If Hadoop is to extend into enterprises as part of the accepted toolbox of information management capabilities, it is essential that the needs of these different types of users can be accommodated. Organizations need to not only understand how these different types of users interact with the traditional BI ecosystem; they also need to understand how they might interact with a BI system that includes Hadoop. In many cases, particularly for those non-expert business users, "hiding" the inherent complexity of big data will be essential for success.

Business end user

The priority for most casual users is to get their jobs done. In order to benefit from BI on Hadoop, business users should not be expected to learn specialized languages such as Java, R, or Python for Scala, or to design Spark or MapReduce programs. Regular business users must be shielded from the complexity of the underlying technology, freeing them to work with all available data types. Both traditional and newer BI solutions are starting to be better integrated into the Hadoop ecosystem, a trend that Ovum expects to continue into the near future. More importantly, the ability to access Hadoop is becoming increasingly simplified through the explosion of interactive-SQL-on-Hadoop.

Power user and data scientist

BI power users and data scientists are the traditional prime targets for Hadoop. The power user in a Hadoop context can be thought of as a data curator. He or she is accustomed to asking questions that make extensive demands on the analytical capabilities of tools by using more advanced and sophisticated modeling and analytic techniques. They will be comfortable coding with statistical and query programming languages such as R and SQL, but in some cases might not be comfortable with MapReduce. This person must possess multidisciplinary skills that encompass application design, some database design, and an awareness of the complementary roles that Hadoop and data warehouses can play. This person can help create assets, such as materialized views of dimensions, for consumption by the greater population of business end users.

While data scientists are not typically considered users or consumers of BI tools, they play an important complementary role upstream in discovering the signals, and identifying the right data sets and the right questions to ask, which can subsequently be operationalized for business end users.

Developer

Traditionally, the main role for developers has been setting up the managed environments to make query and reporting accessible to less technical users. Hadoop increases the need for developers skilled in SQL along with the emerging languages of data scientists, such as Python, R, and Java, and the ability to work with JSON. Ultimately, tools will emerge to simplify developing such queries, just as SQL-based query and reporting tools emerged to make BI accessible back in the 1990s. Nonetheless, while tools may ease access to such information, designing the environments will require practitioners who understand the nature of the data and how to build queries against it.

Where OLAP fits in

As Hadoop has evolved to a multi-purpose engine, thanks to the emergence of YARN, which allows allocations of clusters to support multiple workloads, the vendor and open source community have busily created an abundance of compute engines and frameworks for delivering analytics. Among them are the dozen-plus variants of interactive-SQL-on-Hadoop. These SQL engines have significantly improved the performance of Hadoop in providing an alternative to classic batch-style queries.

But, as Hadoop was not designed as a database, these interactive SQL frameworks in and of themselves will not achieve performance parity with traditional SQL relational data warehouses. This is where innovations such as OLAP on Hadoop come in.

OLAP emerged as a means for optimizing the storage of data for high-performance reporting; familiar "dimensions" can present multiple views of common data that are optimized for the production of reports for specific attributes, such as profitability by customer region, or market share of product.

Over the past couple of years, OLAP has come to Hadoop. There are multiple flavors of OLAP; two in particular – MOLAP and ROLAP – require examination in the context of this analysis:

- MOLAP (multidimensional online analytical processing) is online analytical processing that indexes directly into a multidimensional database. MOLAP typically precomputes and stores every single possible intersection across a model's dimensions. It is very fast but does not scale as well as other methods because it materializes every possible answer. This means storage is required to materialize results as more data is available. It also means that models need to be recomputed when new data is available so new information can be accessible.
- ROLAP (relational online analytical processing) is a form of online analytical processing that performs dynamic, multidimensional analysis of data stored in a relational database rather than in a multidimensional database. ROLAP uses a relational structure and aggregates. This means ROLAP will scale better than MOLAP, although aggregates will need to be managed and refreshed.

New technologies like virtual ROLAP have emerged and can help enterprises get the benefit of ROLAP technologies without having to manage or maintain aggregates.

Preparing for the future without forgetting the past

There can be little doubt that Hadoop is set to play a growing role in the data architecture of many enterprises, a journey it is just setting out on as organizations look to realize use cases that are enabled by big data. This transition will not be without some pain, as the tools and techniques required to capture, store, manage, and analyze new and unfamiliar big data sources are still developing. Equally, organizations should not view the emergence of Hadoop as necessitating a "rip and replace" of existing analytics technologies – this paper has explored how existing skills and investments in familiar BI capabilities that leverage well-understood SQL and OLAP can be a bridge that helps manage the transition.

Hadoop is not the only piece of the puzzle, however, as other macro trends are driving interest in accessing data, notably as discussed in this paper, self-service analytics. A common complaint of IT had been the implications this can have for governance. Governance is an equal challenge in big data and one that necessitates a more modular approach. There is likely a "renegotiation" of the relationship between the business and IT, one where IT enables business users to bring their analytical tools of choice to the data, without compromising governance of that data. A final thought: If big data and Hadoop, along with its ecosystem of technology tools, are to become a regular part of the enterprise information management toolbox, they must make data available to the growing audience of non-expert business users who are interested in generating insight, irrespective of data source (big or small). Ovum believes successful enterprises will adopt tooling that hides the inherent complexity of big data technology from those users, with IT providing access to it, without those users needing to learn a whole new skill set and without compromising governance and security.

Appendix

Authors

Information Management Practice, Ovum

Ovum Consulting

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Ovum's consulting team may be able to help you. For more information about Ovum's consulting capabilities, please contact us directly at consulting@ovum.com.

Copyright notice and disclaimer

Whilst reasonable efforts have been made to ensure that the information and content of this product was correct as at the date of first publication, neither Informa Telecoms and Media Limited nor any person engaged or employed by Informa Telecoms and Media Limited accepts any liability for any errors, omissions or other inaccuracies. Readers should independently verify any facts and figures as no liability can be accepted in this regard – readers assume full responsibility and risk accordingly for their use of such information and content.

Any views and/or opinions expressed in this product by individual authors or contributors are their personal views and/or opinions and do not necessarily reflect the views and/or opinions of Informa Telecoms and Media Limited.

Ovum
TMT intelligence | informa

CONTACT US

www.ovum.com

analystsupport@ovum.com

INTERNATIONAL OFFICES

Beijing
Dubai
Hong Kong
Hyderabad
Johannesburg
London
Melbourne
New York
San Francisco
Sao Paulo
Tokyo