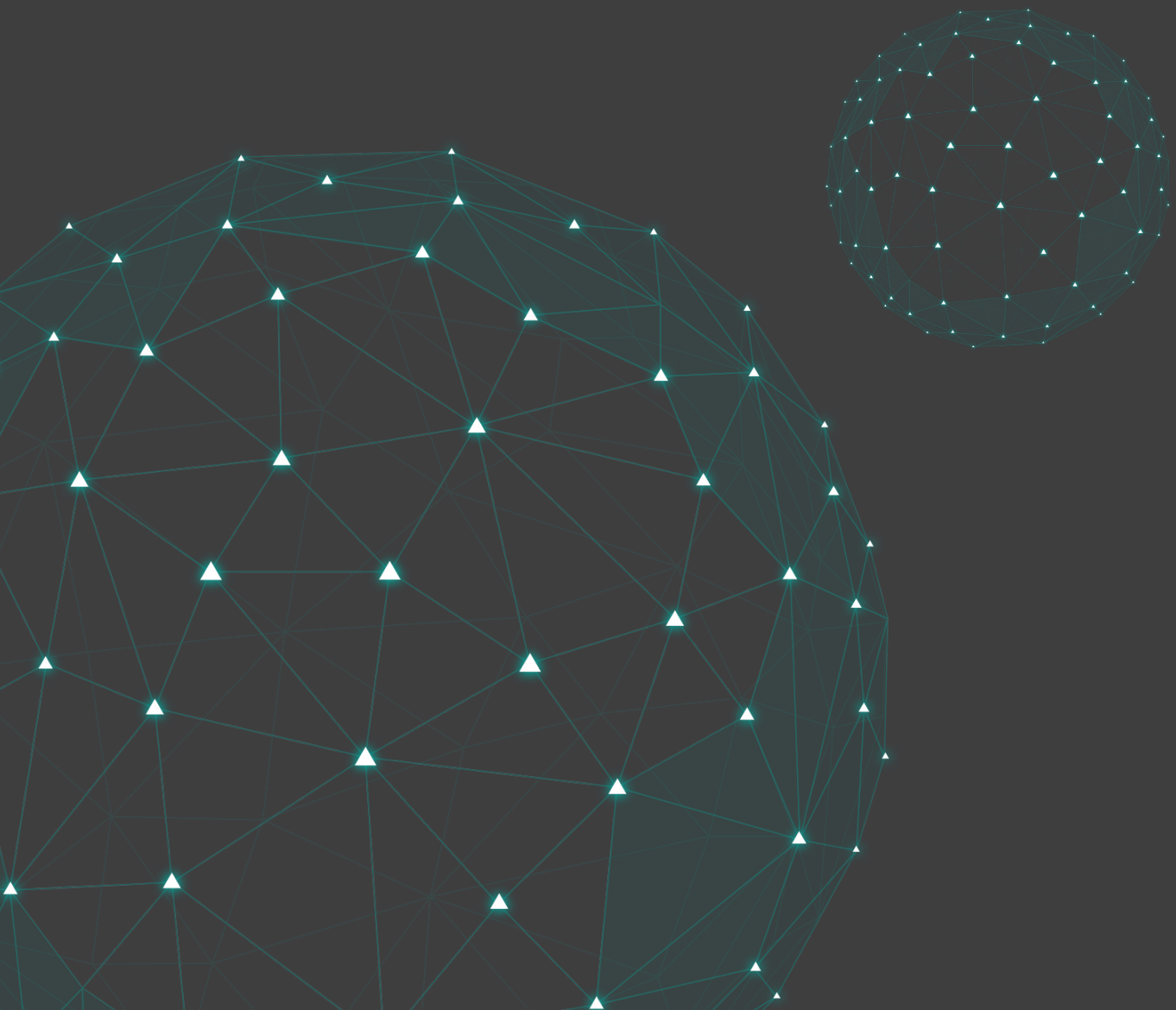
ATSCALE

Modern BI Platforms for Modern Data Demands

Making the Vision of a Single Semantic Layer a Reality



Modern Business Intelligence from AtScale

Over the past few years, more and more enterprises have turned towards Hadoop-based data platforms to support initiatives to centralize data assets. While these projects go by different names—Data Lakes, Data Hubs, Data Reservoirs, Landing Zones—they all have shared objectives around consolidating data, security policies, semantics, storage, and workloads onto a shared platform.

However, challenges remain when the information stored in these Data Lakes must be accessed for purposes of interactive business intelligence (BI). While Hadoop-based data sets may be semi-structured, non-curated, and disjointed by design, to handle the variety of data generated today BI users want access to this breadth of data, but want to consume it in ways that match a business-model representation (measures, dimensions, hierarchies).

Additionally, while Hadoop excels at low cost storage and scale-out batch-oriented or longrunning queries, it remains a challenge to support the level of interactivity that BI users require.

The result of this impedance mismatch is that data sets must often be moved off of the big data cluster and into form factors that are easier for BI users to interface with and into data sizes that are more amenable to fast query response times. With data movement the anticipated value of the Data Hub is compromised and a number of additional challenges arise.

In this document we will discuss an approach to supporting BI and Analytics use cases on Big Data assets that:

1. Preserves the benefits of the big data platform
2. Meets the needs of the BI user community

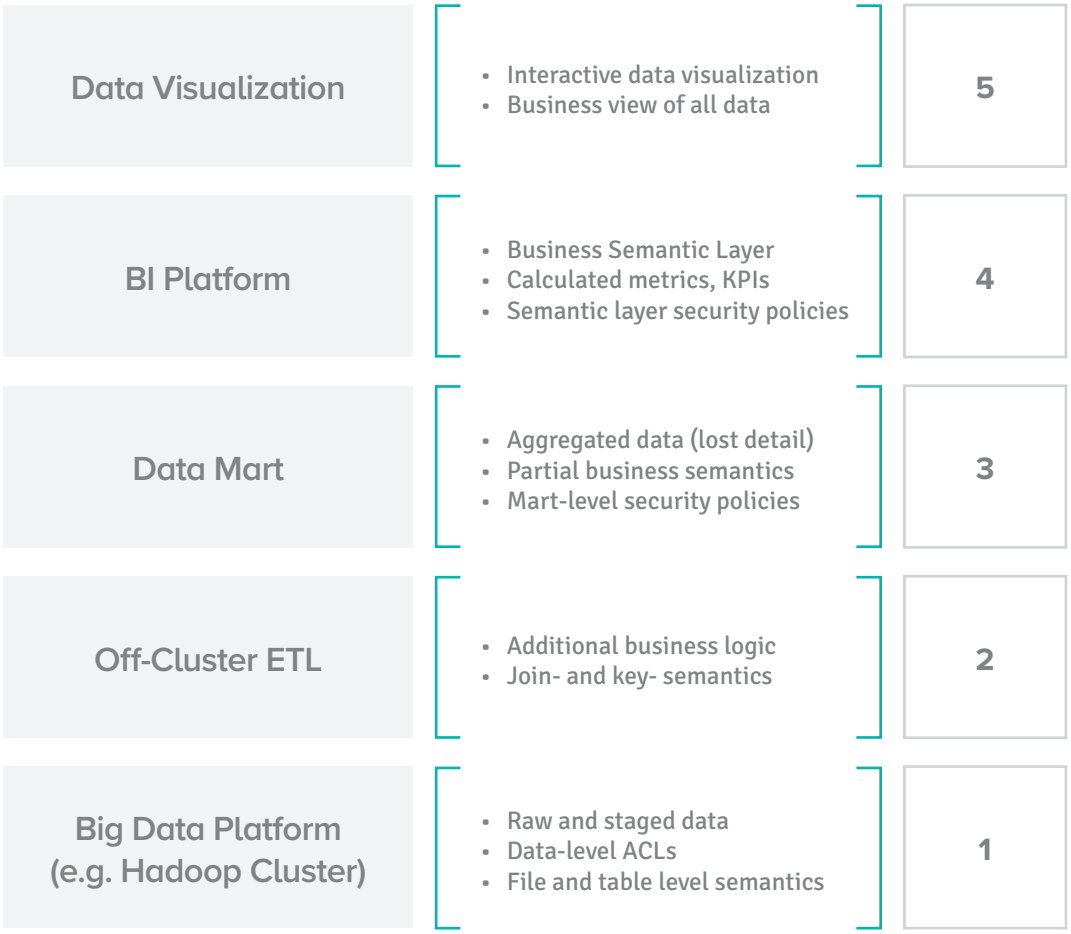
Current State of BI on Big Data

When tasked with supporting BI on Big-Data initiatives today, Enterprise Architects are faced with maintaining a balance between security, control, and consistency for IT and the self-service, interactive access demanded by the business.

In a world where departmental budgets rule and business initiatives, understandably, trump IT desires, the result is often that IT organizations build solution stacks oriented towards solving the needs of demanding business analysts. A typical stack required to support interactive visualization of data that is persisted in a Hadoop cluster often ends up looking like Figure 1 on the following page.

FIGURE 1: TYPICAL BI-ON-HADOOP STACK (SIMPLIFIED)

TIP: Read figure 1 bottom to top (1–5). Note the recurrent data movement required to support business intelligence, analysis, reporting and visualization.



This current approach to business-semantics-oriented and interactive BI on Hadoop data today has some inherent costs.

- There is significant data movement involved
  - First moving data off the cluster
  - Then moving it into BI-oriented data marts.
- Additional hardware is required to host ETL, Data Mart, and BI Platform applications.
- These same applications then drive additional license and maintenance fees as well.

All of these items lead to significant direct, and indirect, costs in terms of time, hardware, and software.

Hidden Cost of Change

An additional hidden (but significant) challenge with this approach is the cost of making changes. For example, propagating a newly collected data element from the Big Data layer through to the visualization layer for consumption involves 4 distinct but related steps.

1. The creation of data parsing jobs
2. Data transfer steps between layers
3. Multiple instances of aggregation logic
4. Additional changes to the semantic layer

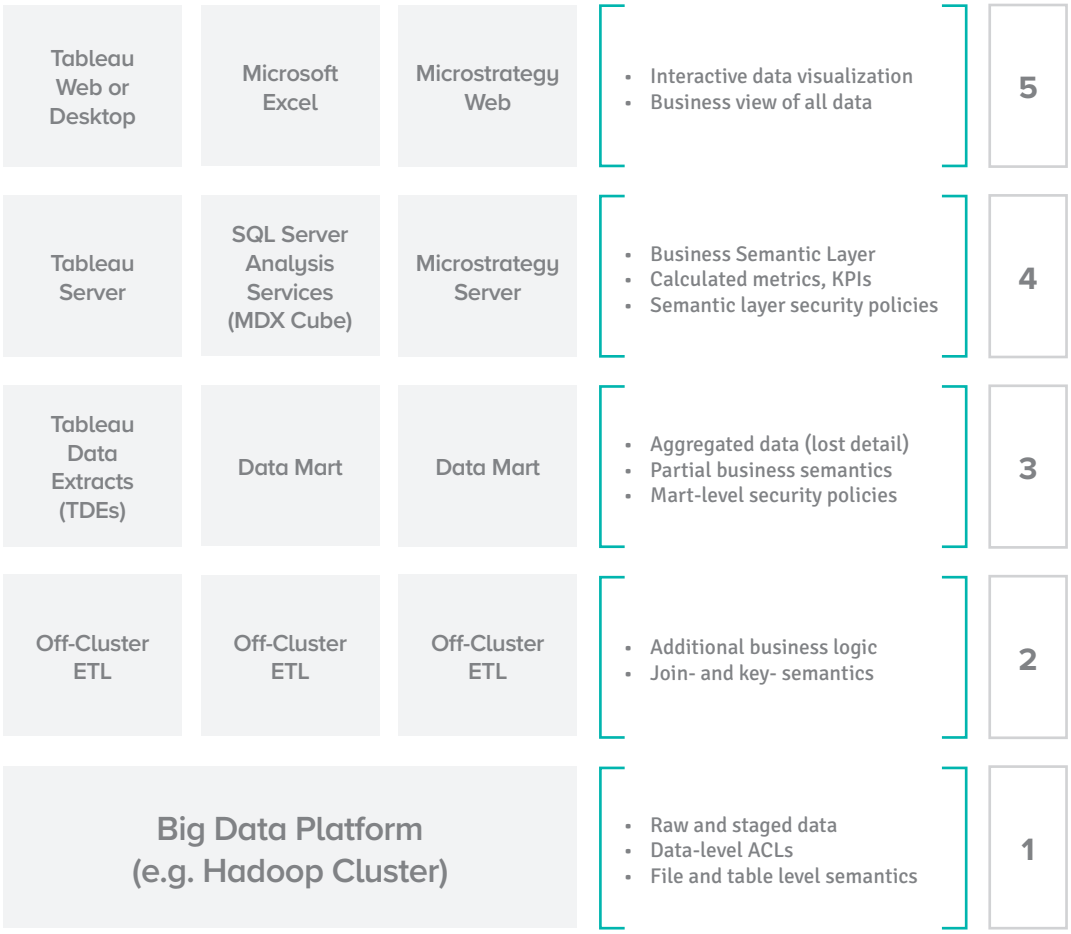
It's been estimated that in a mature IT organization the labor-related costs of making such a change can be roughly $100,000. And the time it takes to make such a change can range between 6 to 12 weeks in a large enterprise. Every time a new element is added, the same costs are incurred. To make matters worse, the opportunity cost of the time taken to make such changes can significantly hamper business agility and decision making.

A Real Picture of BI on Big Data

In reality, IT organizations that are attempting to deliver BI-on-Hadoop capabilities often need to support multiple user types and multiple BI and visualization tools. For example, Financial Planners may access data in Excel, Marketing Analysts might be extensive users of Tableau, and Enterprise Reporting functions could be satisfied using Microstrategy. Because each of these BI front ends has slightly different consumption patterns and interfaces, the resulting "real-life" BI-on-Hadoop stack at any given enterprise might look more like the image shown in Figure 2.

FIGURE 2: A MORE REALISTIC BI-ON-HADOOP STACK (REAL-LIFE SCENARIO)

TIP: Once again read the image from bottom to top (1 to 5).



You'll notice that with the addition of multiple data consumption personae and interfaces, the cost and complexity of maintaining a functional BI-on-Hadoop operation increases exponentially. On top of the hardware, storage, and software costs discussed above, the introduction of multiple purpose-built silos creates additional challenges.

Silo Challenges Deconstructed

1. Multiple Language & Interactivity Requirements

One such challenge is the deployment and maintenance of dedicated physical stacks. Different business intelligence front ends often speak different query languages (for example SQL in Tableau vs. MDX in Excel) and have different requirements in terms of interactivity; as a result, IT organizations must often create vertical stacks to satisfy the specific language and interactivity requirements of each front end.

2. Non-Centralized Business Logic

Compounding the issue with the multiple stack approach is the development of non-centralized business logic: in the "dedicated BI stack" model (Figure 2), the logic required to transform and represent underlying data sets as business consumable elements (Measures and Dimensions) must be created and maintained within each stack. The result is that different BI front ends may produce different results for values that are logically and semantically the same. One consequence is a significant organizational cost in the form of data reconciliation (time spent explaining and investigating why different systems produce different results for the same metric). Additionally, it leads to organizational distrust of data - as demonstrated in a recent survey revealing that only 30% of CFOs trust the numbers that they receive from internal BI systems.

3. Security Implications

A third, less obvious, but equally (or even more) important concern with the traditional approach to BI on top of Big Data platforms is the need for multiple security implementations. As enterprises endeavor to populate their data lake with comprehensive data sets, the security implications increase significantly. With personally identifiable information (PII), detailed financial data, and other types of proprietary and protected data sets CIOs must be increasingly mindful of security policies, data access controls, auditing requirements, and enforcement capabilities of their big data platforms. The traditional, multiple-stack approach to BI on Hadoop results in the need to implement multiple security controls, often using different systems, in a multitude of data platforms. This introduces security risk exposure, auditing difficulties, and a significant data governance challenge. Increasingly enterprises who adopt big data platforms like Hadoop are mandating that ALL queries against big data assets be individually identifiable by the user id of the end consumer of the data. With off-cluster BI stacks this becomes essentially impossible.

Overall, the current approach to supporting enterprise Business Intelligence use cases on big data platforms is rife with challenges, ranging from high cost, lack of agility, and significant security risk. Even with these challenges many IT organizations continue to use the traditional approaches described above, simply because they are unaware of better alternatives.

Deploying a Single Semantic Layer for BI on Big Data

In the past few years, new technological approaches to supporting enterprise BI on big data platforms like Hadoop have emerged. The AtScale Intelligence Platform, based on decade's worth of experience in the business intelligence and big data spaces, incorporates an innovative architecture that uniquely addresses the numerous challenges that traditional "BI-on-Big-Data" approaches suffer from. The AtScale solution has been designed from the ground up to deliver the performance and user-friendly interfaces that BI users demand while at the same time providing the consistency and security controls that enterprise IT organizations require. All of this is done in a way that vastly reduces cost and improves agility.

FIGURE 3: A SHARED SEMANTIC LAYER APPROACH FOR BI ON BIG DATA

TIP: Reading from bottom to top, take note of the reduction in data moves and siloed BI stacks.

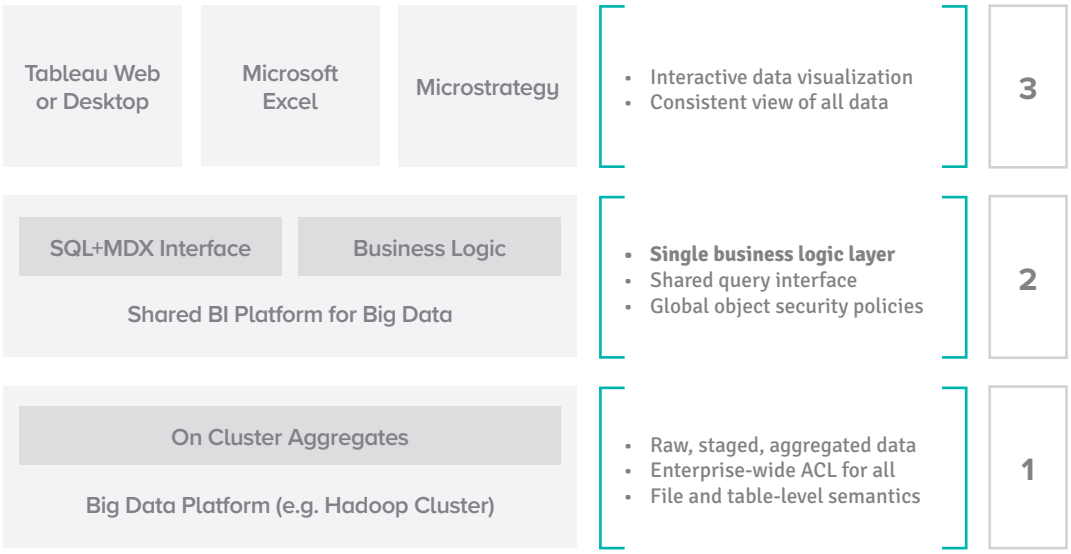


Figure 3, above, shows what an enterprise BI solution for Big Data looks like when implemented using the AtScale Intelligence platform (single semantic layer approach). As the diagram clearly indicates, this approach vastly simplifies and improves on the traditional approach to BI on Big Data. The sections below outline how this is achieved along with the significant benefits that are delivered via this modern technology stack.

Shared Business Logic

At the core of the AtScale solution is semantic layer that represents the core business elements that require analysis and can be shared across multiple consumption interfaces (BI tools). AtScale leverages the traditional 'concepts' of OLAP (Online Analytical Processing) by exposing these business entities as measures and dimensions. The benefit of this approach is that the semantics of OLAP (hierarchies, measures, dimensions) are very well understood by the business users of enterprise data visualization tools as well as the languages that these visualization tools use to communicate with data providers. As industry experts have pointed out, the core elements of OLAP - support for many-to-many relationships, semiadditive metrics, and period-to-period comparisons - are both valuable and complex to support. The AtScale model supports these concepts as a shared interface for BI.

As a result, the AtScale semantic layer (referred to as a Virtual Cube) can be used by virtually any enterprise BI tool. Unlike the traditional approach to BI on Big Data (where semantic logic and associated acceleration approaches) exist at the BI tool layer, AtScale enables the semantic layer to exist directly on top of the big data cluster and does not require the development of tool-specific, or siloed, semantics.

Multiple Query Language Support

One of the challenges of providing a unified semantic layer to a diverse set of analytic, business intelligence, and data visualization clients is that not every tool utilizes the same query language. Some front-ends, like Tableau, are most effective when querying relational data structures using SQL. Other tools (like Microsoft Excel) can only do dynamic querying of data sources using MDX (multi-dimensional expression language); Even other platforms, like Microstrategy, can either consume relational data (through the creation of proprietary Microstrategy metadata) or by inheriting multidimensional schemas through the XMLA (XML for Analysis) standard.

With AtScale's Hybrid Query service, all of these distinct query and catalog use cases can be satisfied with a single Virtual Cube. Our Hybrid Query Service can expose an AtScale Virtual Cube as if it were a relational source that can then be easily queried using SQL (using ODBC or JDBC), and also as an XMLA-compliant data source that can be queried using MDX. As a result, all BI client tools - from Microsoft Excel to Tableau to others - can query using the same set of business logic. AtScale even allows traditional "Full Stack" enterprise BI tools to, such as Microstrategy and Cognos, to inherit AtScale Virtual Cube semantics (via XMLA), allowing organizations to leverage existing investments in enterprise report distribution and analysis.

On Cluster Aggregates for Performance

One of the main reasons that traditional approaches to BI on Big Data require separate storage and query systems is to be able to deliver acceptable end user performance; query response times on the order of several seconds or less. To achieve this level of performance target data sets are often summarized, aggregated, or indexed using systems that are, once again, 'separate' from the Hadoop cluster. Some examples of off-cluster acceleration include:

- Multidimensional OLAP (MOLAP) cubes using Microsoft's SQL Server Analysis Services (SSAS)
- Data extracts such as Tableau Data Extracts (TDEs)
- Separate indexes, created and stored, via off-cluster post processing
- Separate aggregated data marts in traditional relational databases

AtScale achieves the performance benefits of aggregations using its Adaptive Cache technology. The Adaptive Cache engine continuously analyzes end user query patterns and determines an appropriate set of aggregate tables that can be used to most optimally satisfy future versions of similar queries. These aggregates are created and stored directly on the Big Data cluster, using the scale-out data processing and low-cost storage that platforms like Hadoop provide. Additionally, the Adaptive Cache manages the complete lifecycle of these aggregate tables, from creation, to updates, to end-of-useful-life.

The benefits of the AtScale on-cluster aggregate approach are significant.

1. No data movement or additional software/hardware is required to achieve the required level of interactive query results.
2. The labor and maintenance efforts related to data movement and aggregation jobs is eliminated.
3. Automation of the formerly time consuming process of query analysis required to identify new aggregates and to identify and remove non-valuable aggregates.

A Unified Approach to Secure BI on Big Data

By moving BI workloads on Big Data to an on-cluster solution, enterprises vastly reduce risk and exposure related to secure data access. User and group level data access controls can be implemented directly on cluster level data assets and enforced at query time by the AtScale query engine. No additional administrative effort is required to re-implement data ACLs in downstream systems. Additionally, AtScale's patent-pending approach to supporting delegated queries against cache level data means that a shared query cache can be used to satisfy end-user queries while still respecting the ACLs defined against the raw data. For Chief Security Officers (CSO) this approach to integrated on-cluster security results a significant reduction in risk along with the ability to continue to meet demanding business-user data access requirements.

In Conclusion

Enterprises around the globe are diving headfirst into the era of big data and making significant platform decisions regarding their next generation data architectures. Big data platforms, such as Hadoop, are key components of these architectures. It is critical that in addition to re-architected their core data storage and processing infrastructure that CIOs and CTOs consider the architecture for their business intelligence solutions as well. With the increasing centralization of data assets combined with the explosion in self-service data consumers, a new approach to BI is required, along with the deployment of a single semantic layer that meets the needs of the Business and IT. The benefits of the right architecture are significant and wide-ranging, and include reduced capital expenditure, reduced operating costs, increased agility, and improved data governance. With the AtScale Intelligence Platform, these benefits can be achieved as part of the broader journey to successful business intelligence on big data.

To find out more, visit atscale.com

ATSCALE

400 S El Camino Real Suite 800, San Mateo, CA 94402