DEFINING THE MODERN DATA LANDSCAPE – THE NEED FOR SPEED, SCALE AND COST EFFECTIVENESS

The Rise of the Semantic Layer

By Brian Prascak

Brian Prascak is an expert in the area of data, insights and analytics, with extensive experience helping companies realize the benefits of using AI / ML and BI across a wide variety of industries and functions in the US and globally. Brian is currently Co-founder and Chief Insights Officer at Naratav, Inc. Before starting Naratav, Brian was Director, Advanced Analytics, Platforms and Data Services at Wawa. Brian has worked with many global organizations, including previously at IBM, Diageo, Mastercard, JPMorgan and AC Nielsen.

ATSCALE

Purpose

The purpose of this white paper is to introduce the modern data landscape, including providing perspective on the data and analytics industry - where we've been, where we are now, and where we are heading, linking the catalysts for improvement to the capabilities being sought, the vendors providing them and the investments that are being made. The emphasis will be to define the modern data landscape, frame its purpose and direction, with a focus on the fundamental need for actionable, impactful insights and analytics that are delivered with speed, scale and cost effectiveness. The rise of the semantic layer will be featured, including new research that affirms the value of using a semantic layer to deliver increased speed, scale and success for AI and BI.

Defining the Modern Data Landscape - Catalysts

It's hard to believe that just 15 years ago, big data and cloud technology emerged with the introduction of Hadoop and cloud vendor offerings. Over the past 5 years, most enterprises have moved to the cloud, motivated by the dual need for digital transformation of their business, coupled with embracing cloud-based data platforms and tools to realize the benefits of advanced insights and analytics. Most companies have now implemented migration to the cloud, with most if not all their data available in a cloud-based data lake. Also, given the ever-increasing number of success stories across a wide variety of industries, companies have bought-in to using data and analytics to significantly improve business performance, with many implementing an initial set of use cases and making plans for continued expansion. The resulting investment in big data technology reveals the scope of this transformation: according to research firm International Data Corporation (IDC), worldwide spending on big data and business analytics (BDA) solutions in 2021 was forecast to reach $215.7 billion, an increase of 10.1% over 2020 - with IDC forecasting that BDA spending will gain strength over the next five years as the global economy recovers from the COVID-19 pandemic. The compound annual growth rate (CAGR) for global BDA spending over the 2021-2025 forecast period will be 12.8%, much larger than most every category of IT spend. Per IDC, total BDA spend is expected to be split evenly between services and software solutions.

The Need for Speed, Scale and Cost Effectiveness

Let's start with aligning on the key drivers of value when it comes to implementing data and analytics capabilities and selecting vendors. Fundamentally, the market has moved from an emphasis on basic infrastructure and tools, primarily the book-ends of the modern data stack, representing cloud data lake providers such as AWS, Google and IBM, to basic data wrangling and BI tool providers like Alteryx, PowerBI, Excel, Tableau and Microstrategy and open source resources such as Python and R to additional emphasis on capabilities and vendors that deliver the following:

- Actionable Insights - Fundamentally, data is a means to an end - and that end is an insight that leads to a better answer. And that answer must be actionable in order to deliver impact. Data is nothing if it is not actionable and impactful, so companies are seeking to make their data more actionable, and the way to do that is by addressing the four (4) A's of Actionable Insights - Availability, Accuracy, Actionability and Automation.

THE CORE FOUR A'S OF DATA – ACCELERATED

AVAILABLE	ACCURATE	ACTIONABLE	AUTOMATED
Data are easy to locate and access	Data are accurate and complete	Data address key questions / needs	Data processes are automated
Accelerated – Provide Semantic Layer, Self Service, Governance, COE's			

- Scale - Now that most companies have all of their data in one place, they are seeking to scale the number of data sources, users and use cases they can support. Scale is the most critical challenge that companies face today, and the reward should certainly be worth the effort - as a recent McKinsey study Tipping the Scales in AI indicates, companies that scale insights and analytics achieve on average more than 8% points higher EBIT (3.4x improvement) than companies who have not achieved scale.

ACHIEVING SCALE WITH DATA AND ANALYTICS - CORE ELEMENTS

DATA SOURCES	USES	USERS	CONSISTENCY
Expand number of integrated data sources available for analysis	Expand number of use cases that can be addressed: AI and BI	Expand the number of insights creators and consumers, including self-service users	Implement semantic layer, data catalogs and feature stores for reusability

Data Governance - Data Access, Pipelines, Master Data, Data Products, Metrics/Features

- Speed - Speed and Scale are really two sides of the same success coin: scale means nothing without speed, but speed means nothing without scale. The most important characteristic of actionable insights is that they are relevant - and relevant means timely. All too often, companies take way too long to make desired data sources available, and even longer to make that data actionable for AI and BI. Why? The process for turning data into actionable insights - what we call the "Last Mile" - can take as many as seven steps (accessing, profiling, preparing, integrating, analyzing, synthesizing, presenting) - often these steps are done manually with multiple resources requiring multiple handoffs and reviews with frequent refinement loops - most companies say that it takes an average of 4 months or more to launch a new data source. Recent research reveals that using a semantic layer can reduce this time by 1/4th to just 4 weeks or less.

ACHIEVING SPEED-TO-INSIGHTS WITH DATA AND ANALYTICS - CORE ELEMENTS

RAPID ACCESS	RAPID PREPARATION	RAPID REFINEMENT	PUBLISHING/SHARING
Enable rapid, governed access to analysis-ready data	Provide self-service data preparation and modeling tools	Use semantic layer and feature stores for consistency and reuse	Use Semantic Layer to automate data product publishing

Capabilities - Data Strategy, Capability Roadmap, Tools, Skills, Literacy, Delivery

Achieving Success - Transformational Capabilities

Before we dig into the modern data landscape, let's also review the major areas that are being transformed to deliver actionable insights for AI and BI with the necessary speed, scale and cost effectiveness. These address the people and process aspects. Fundamentally, we see the following capabilities being implemented by companies that are successfully implementing data and analytics at scale.

DATA AND ANALYTICS TRANSFORMATIONAL CAPABILITIES

DATA LITERACY	DATA AS PRODUCT	DATA DEMOCRATIZATION	DECISION INTELLIGENCE
Understanding how to improve using data and analytics	Managing data as a product across the enterprise	Decentralizing insights and analytics with central support	Understanding how decisions can be improved and scaled

Technology - Data Lakes, Virtualization, Semantic Layer, Catalogs, Feature/Metric Store

Defining the Modern Data Landscape

Now that we have defined what we want to achieve and what we want to transform, let's define the modern data landscape, including key vendors. We will also take a look at the areas that are rapidly emerging, including those most suited to support achieving increased speed, scale and cost savings for AI and BI.

Modern Data Landscape - Capability Areas

The modern data landscape consists of seven (7) major capability areas, representing fifteen (15) individual capability components. Let's briefly review each of the capability areas:

1. Raw Data - Raw data represent sources and storage of raw data sources. There are two major categories of raw data:

   - Data Lakes managed by cloud providers
   - SaaS applications where data is managed by the vendor for clients, who can access via web protocols, including API's.

2. Data Preparation, Integration, Workflow (DPIW) - The DPIW capability area enables data to be extracted and prepared: cleaned / transformed and made available as a ready-to-use set of data. There are two major categories of DPIW:

   - Data Transformation and Preparation Tools - These are tools to profile the data, assess it, cleanse it and do basic transformations to it to make it ready for analysis, including as a single data source or integrated with other data prepared data sources. Often, these tools create data pipelines and automate the process of data preparation.
   - Customer Data Platforms / Event Tracking - With the increased maturity and confluence of ecomm and digital marketing, companies now must use a plethora of data sources and vendors to manage customer data across a myriad of channels. As a result, Customer Data Platforms, which offer purpose-built capabilities to manage customer identification, hygiene as well as rapid access and integration of data between multiple marketing data vendors and channels have increased exponentially in popularity. According to IDC, the worldwide customer data platform software market will grow at 19.5% CAGR from $1.3 billion in 2020 to $3.2 billion in 2025.

3. Data API - The Data API Layer is rapidly emerging as another accelerator for companies to more rapidly access data from source systems, including data warehouses in the cloud and process it at the source (rather than move it). There are three (3) major categories of Data API vendors:

   - Cloud Data Warehouse - A cloud data warehouse is a database stored as a managed service in a public cloud and optimized for scalable BI and analytics. Cloud data warehouses typically offer three major services: secure access, compute or query processing and storage.
   - Data Lake Engines - A data lake engine is an open source software solution or cloud service that provides critical capabilities for a wide range of data sources for analytical workloads through a unified set of APIs and data model. Data lake engines address key needs in terms of simplifying access, accelerating analytical processing, securing and masking data, curating datasets, and providing a unified catalog of data across all sources. Data lake engines simplify these challenges by allowing companies to leave data where it is already managed, and to provide fast access for data consumers, regardless of the tool they use.
   - SaaS API's - These providers offer rapid, software-as-a-service (SaaS) data integration service for companies to extract, load and transform (ELT) data from different sources into data warehouses. Often these providers create a standardized data model and framework to move data from standardized sources, including other SaaS-based data providers / sources.

4. Logical Data Models - This capability is critical to ensuring that data is consistently available to the consumption layer for AI and BI applications. With the number of applications consuming data, it is critical to ensure that the data is consistently defined, modeled, aggregated and optimized for presentation and rapid query response. There are three (3) major categories of Logical Data Model providers:

   - Semantic Layer - The Semantic Layer improves the time to insights for AI and BI by simplifying, automating, standardizing, and optimizing how data products are created, consumed, and queried for AI and BI. Semantic Layer leaders like AtScale offer a comprehensive set of capability components, including Consumption Integration, Semantic Modeling, Data Preparation Virtualization, Multi-Dimensional Calculation Engine, Performance Optimization, Analytics Governance and Data Integration.
   - Metric / Feature Stores - Another fast-growing area within the data landscape is the use of metric and feature stores. Metric stores are typically used to support business intelligence whereas feature stores support data science uses. Both metric stores and feature stores address common needs and provide common benefits - namely to support the consistent definition of metrics and features, and provide a single, centralized source for consistent reuse across the enterprise.
   - Data Virtualization - Data virtualization provides a logical data layer that presents, and enables integration of data that may be siloed across the disparate systems, manages the unified data for centralized security and governance, and delivers it to business users without having to physically move the data. Data Virtualization is often used in conjunction with a semantic layer, where the data virtualization speeds access to the data whereas the semantic layer speeds the ability to access the data consistently (and refine it) through multiple AI and BI consumption tools without creating multiple versions for each tool.
   - Data Governance - Data governance (DG) is the process of managing the availability, usability, integrity and security of the data in enterprise systems, based on internal data standards and policies that also control data usage. As the number of data sources, users, uses and consumption tools increase for both the data, but also the data products (refined data sets created by the semantic layer models and metric / feature stores), data governance becomes increasingly important. Please note that companies like AtScale provider governance capabilities built into the semantic layer to govern data as a product.

5. Data Consumption - This capability is critical to ensuring that data is structured and presented effectively for business intelligence as well as analytics. There are two major categories of data consumption vendors:

   - BI Tools - Business intelligence (BI) tools are types of application software that collect and process large amounts of data from internal and external systems, including books, journals, documents, health records, images, files, email, video, and other business sources. BI tools provide a way of amassing data to find information primarily through queries. These tools also help prepare data for analysis so that you can create reports, dashboards, and data visualizations. The results give both employees and managers the power to accelerate and improve decision making, increase operational efficiency, pinpoint new revenue potentials, identify market trends, report genuine KPIs, and identify new business opportunities.
   - AI /ML Tools - These are tools designed to speed up the process of creating AI / ML models. Often they offer workflow automation, data preparation, access to models / algorithms and support training and operationalization.

6. Data Catalogs - A data catalog is an organized inventory of data assets available for access within the enterprise. Data Catalog uses metadata to help organizations manage access to their data, including collecting, organizing, accessing, and enriching metadata to support data discovery and governance.

7. Data Observability - Emerging as a newer area within the modern data landscape, Data Observability refers to an organization's ability to fully understand the health and reliability of the data in their system. Traditionally, data teams have relied on data testing alone to ensure that pipelines are resilient. However, as companies ingest ever-increasing volumes of data and the data pipelines become more complex, testing during deployment is no longer sufficient. Continuous monitoring of data to determine if changes are taking place is critical to ensuring tracking of data quality, lineage, consistency, usage, governance, and refinements across the entire ecosystem - all part of what is now being called "data operations" - ensuring that all data sourced, created, transformed, synthesized, summarized and consumed used to support multiple applications are consistently defined and delivered as needed.

Modern Data Landscape - Fast Growing Areas

As the modern data landscape continues to evolve, focusing on delivering actionable insights and analytics via improved speed, scale and cost savings, we see the following areas accelerating growth in investment, customers and market coverage:

Semantic Layer - Although AtScale was the first to introduce a standalone Semantic Layer Platform that maintains a semantic model independent of any BI platform or data store more than 10 years ago, client interest and investment have accelerated in the past 3 years as companies have migrated to the modern data platforms, and have realized the importance of achieving speed, scale and cost savings at the actionable insights level. As more companies move more of their data to the modern data platforms, the importance of the semantic layer becomes even more important - recent research by Ventana Research reveals organizations that have successfully implemented a semantic model are more than twice as likely to report satisfaction with analytics (77%) compared with a 33% overall satisfaction rate.

Metric / Feature Stores - Supporting the accelerating interest using a Semantic Layer for AI and BI, enterprises are also embracing the complimentary capability of using centralized metric and feature stores to ensure consistent definition of metrics and features, and provide a single, centralized source for consistent reuse across the enterprise. Companies embracing the use of the Semantic Layer typically also embrace the use of metric and feature stores to ensure that both existing and new datasets / data products are consistently defined and productively shared.

Data Governance, Data Catalogs and Data Observability - As companies embrace the use of cloud-based data platforms, and as data sources and applications that consume data expand, companies are embracing the use of the Semantic Layer and Metric / Feature Stores supported by the increased emphasis on Data Governance to manage data privacy, access and usage, Data Catalogs to support data discovery and data observability to monitor data moving through the entire data ecosystem.

AI / ML Tools - As companies increase their embrace of AI / ML, they are also looking to improve the productivity of their data science teams, including analysts using AI / ML automation tools. As more companies do more with AI / ML, interest in tools to increase productivity through workflow and automation, including self-service increases, as does improvements in their capabilities to support self-service for both data scientists and analysts.

Customer Data Platforms (CDP) - As mentioned earlier, investment in CDP's is growing exponentially due to the combination of cookies going away (companies having to manage customer data more directly with explicit permissions), digital transformation and use of analytics to improve customer engagement and deliver more personalized recommendations.

Data API Layer - All of the capabilities within the Data API layer are growing rapidly as companies seek faster ways to access, integrate and compute data from multiple sources within the cloud, including many new data sources from existing sources (not analyzed before) and new sources (new vendors).

Semantic Layer Rising

As mentioned, over the past year there has been a tremendous resurgence in the Semantic Layer among large enterprises. This traces to their recent experience migrating to modern data platforms and now experiencing the need to improve speed, scale and cost savings for AI and BI - being able to generate actionable insight from newly available data sources for many new users and use cases. The good news is that recent research affirms the value of using a semantic layer.
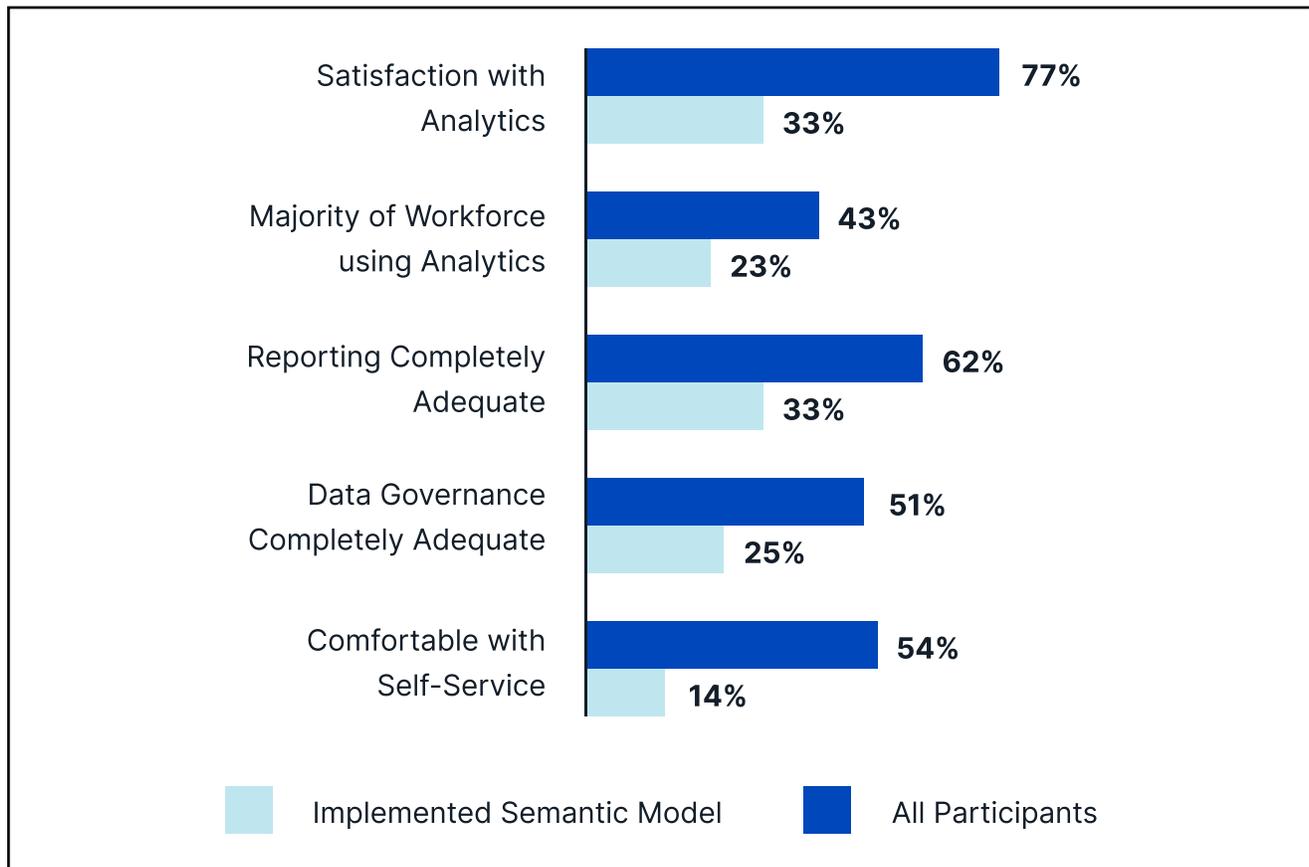
The research points to companies realizing the promise of successful, impactful data and analytics programs using a semantic layer - and in stark contrast to those that don't use a semantic layer.

According to recent research from Ventana Research, based on 300 respondents:

Organizations that have successfully implemented a semantic model/layer:

- Are significantly more satisfied with analytics (77% compared with 33% overall)
- Have more of the workforce engaged in analytics (43% compared with 23% have more than one-half the workforce using analytics)
- Find analytics capabilities completely adequate (62% vs. 33% overall)
- Say data governance capabilities are completely adequate (51% vs. 25% overall)
- Are more comfortable with self-service: (54% very comfortable vs. 14% overall)

VALUE OF SEMANTIC MODELS

Satisfaction with Analytics: 77% / 33%
Majority of Workforce using Analytics: 43% / 23%
Reporting Completely Adequate: 62% / 33%
Data Governance Completely Adequate: 51% / 25%
Comfortable with Self-Service: 54% / 14%

Legend: Implemented Semantic Model | All Participants

Further recent research from DBP Institute, over 100 respondents cited the following benefits using a semantic layer:

Companies using a Semantic Layer cite a 4.2x improvement (i.e., a magnitude of 4.2 times improvement over the base level of performance from not using a semantic layer) in performance with less than half the effort required (e.g. savings in both number of resources, hours, project time / duration, and overall cost). This is a significant order-of-magnitude improvement in performance as well as a reduction in effort and cost. It means that a typical project taking 4 months to complete could be done in just 4 weeks using a Semantic Layer! Performance improvement was significant and consistent across every measure.

- 4.4x improvement in Time-to-Insights (e.g., insights and analytics development)
- 4.4x improvement in number of self-service users, data sources, metrics consistency
- 4.2x improvement in Cloud Analytics performance
- 3.7x improvement in cost savings

AtScale Semantic Layer: Enabling Actionable Insights for Everyone

AtScale provides a Semantic Layer, which sits between the Data Source Layer and the Insights Consumption Layer (e.g., AI, BI and Applications). The Semantic Layer converts data into actionable insights via Automation (self-service data access, preparation, modeling, and publishing), Alignment (centralized data product management and governance with a single, consistent metric store) and Acceleration (cloud analytics optimization - BI query speed optimization, multidimensional OLAP in the cloud, AI-based data connectors, and automated PDM tuning). This supports insights and analytics creators, enablers and consumers without requiring data movement, coding, or waiting.

Providing Automation + Alignment + Advancement With No Data Movement, No Coding and No Waiting		
AUTOMATION	**ALIGNMENT**	**ADVANCEMENT**
Self-service data access, preparation, modeling, publishing for AI & BI	Centralized Data Product Management with Single Enterprise Metric Store	10X Increase in Query Performance, Automated Tuning, Cloud OLAP

Brian Prascak is an expert in the area of data, insights and analytics, with extensive experience helping companies realize the benefits of using AI / ML and BI across a wide variety of industries and functions in the US and globally. Brian is currently Co-founder and Chief Insights Officer at Naratav, Inc. Before starting Naratav, Brian was Director, Advanced Analytics, Platforms and Data Services at Wawa. Brian has worked with many global organizations, including previously at IBM, Diageo, Mastercard, JPMorgan and AC Nielsen.

Brian Prascak, Co-founder, Chief Insights Officer, Naratav, Inc.