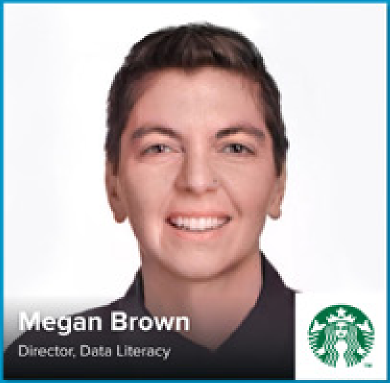
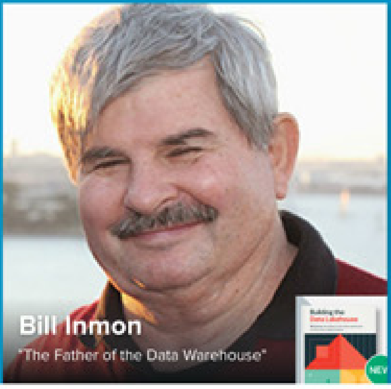
Juan Gorricho
CDO

Ram Kumar
CDAO

David Mariani
CTO


Megan Brown
Director, Data Literacy


Bill Inmon
"The Father of the Data Warehouse"

Greg Mabrito
Director of Data and Analytics

Narendra Narukulla
Senior Engineer, Data Science

Kirk Borne
PhD, Chief Science Officer

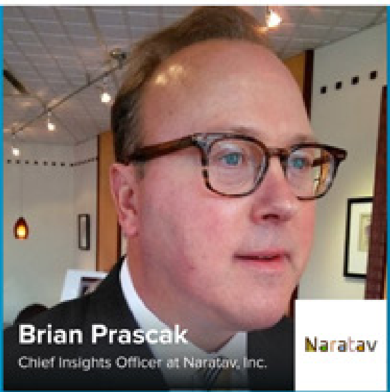
MAKE AI & BI WORK AT SCALE

Get collective advice from 15 thought leaders and industry experts

Michael Gregory
Field CTO Lead

Prashanth Southekal
Managing Principal

Bill Schmarzo
Data Management Incubation


Brian Prascak
Chief Insights Officer at Naratav, Inc.

15 Data Leaders

Anik Bose
Partner

John Thompson
Best Selling Author, Innovator and Thought Leader / Founder of Marketing Sciences

Ganes Kesari
Chief Decision Scientist

Contents

From the Editor’s Desk

The amount of data generated by businesses today is unprecedented. As this data growth continues, so do the opportunities for organizations to derive insights from their Data and Analytics initiatives and derive sustainable competitive advantage. A report from MIT says, digitally mature firms are 26% more profitable than their peers. McKinsey Global Institute indicates that data-driven organizations are 23 times more likely to acquire customers, six times as likely to retain customers and become 19 times more profitable. Overall, Data and Analytics today are the next frontier for innovation and productivity in business.

But achieving a sustainable competitive advantage from Data and Analytics is a complex endeavor and demands a lot of commitment from the organization. Gartner says only 20% of the Data and Analytic solutions deliver business outcomes. A report in VentureBeat says 87% of Data and Analytics projects never make it to production. So, how can organizations get value from Data and Analytics? Specifically, how can enterprises leverage the Semantic Layer to achieve AI (Artificial Intelligence) and BI (Business Intelligence) at scale, given that the Semantic Layer helps business users access data using common business terms?

Against this backdrop, I have worked with numerous Data and Analytics experts and put together a comprehensive book, “Make AI & BI Work At Scale” or rather of a Body of Knowledge (BoK) on how the Semantic Layer can help organizations achieve AI and BI at scale. This book gives a holistic perspective to the Data and Analytics community covering data management, data engineering, data science and decision science to improve the odds of delivering Data and Analytics solutions successfully. The authors in the 16 chapters who have contributed to this book include industry practitioners, subject matter experts (SMEs), and thought leaders who have a stellar track record in leveraging Data and Analytics solutions for improved business performance.

Overall, the Data and Analytics implementation is an evolutionary process just like the business entity itself. The insight needs of the businesses constantly change, the organizational capabilities continuously mature, the data sets grow, improve, and sometimes even degrade, and the technological capabilities to capture, store, and process the data improve over time. The Semantic Layer, if managed well, can play a pivotal role in managing this change and improve business performance. This book will help the community derive value and business results from Data and Analytics initiatives. All the best!

Your Sincerely

Prashanth Southekal, PhD, MBA, MS
Editor of Make AI & BI Work At Scale
Managing Principal, DBP Institute & Professor at IE Business School, Spain

Foreword

Aligning AI & BI to Business Outcomes

VISA

Juan Gorricho

VP of Global Data & Business Intelligence

Juan F Gorricho is the Vice President, Global Data and Business Intelligence at Visa. He leads the efforts for data use at Visa globally, including data acquisition and architecture, data governance and quality, and data consumption for internal decision making, for product development, and for data sharing. Before Visa, he was the Senior Vice President of Data and Analytics at TSYS and Chief Data and Analytics Officer at Partners Federal Credit Union, exclusively serving the Walt Disney Company employees. Gorricho has over 20 years of experience in the data and analytics space, and frequently speaks as a thought leader and an industry expert at data and analytics related conferences and seminars. He holds an industrial engineering degree from Universidad de los Andes in Bogotá, Colombia and a Master's of Business Administration from the Darden Graduate School of Business Administration at the University of Virginia.

There are plenty of books about data and analytics nowadays. If you are reading this eBook – Make AI and BI work AT Scale, you are convinced of the value of data. Most likely, you are trying to figure out ways in which you can make more progress, probably faster, on realizing the value of data. It is possible that you are stuck (or maybe curious?) on looking at technology as the main path forward: cloud, Hadoop, which business intelligence platform, which AI platform and technology to use, etc.

I hope I don't disappoint you with this: technology is usually the easiest part of the data journey. If you are hoping to use this book (or any other data-related book for that matter) to figure out technology as the path to success in data, I can tell you that no book, even the most technical one, will help. As Randy Bean recently published in his annual survey on big data, in response to the question "what is the principal challenge to your organization becoming data-driven?" 92% of the respondents selected "people/ business process/culture" as the problem and 8% selected "technology". There is a similar conclusion in Gartner's annual Chief Data Officer survey. Respondents to both surveys are senior level executives in large organizations across the world. Said differently, senior leaders driving data initiatives.

Don't get me wrong: technology is important. The point I want to make is that you need to think also about other critical aspects such as people and culture. These, as highlighted by the surveys I mention above, will determine how successful your journey in data will be. My advice, as you read through this book, is to always keep in mind how you will leverage what you learn in the context of people and culture as critical success factors.

Another point I want to highlight, which is becoming more relevant (and I know I am going to get a lot of you to roll your eyes) is data governance. Particularly, aspects related to data privacy. Consumers and companies are becoming more aware of the value of data and are becoming more sensitive to privacy. An indicator of this is the number of GDPR and CCPA-like privacy laws being put in place across the world. Like people and culture, good data governance which proactively addresses privacy risks and concerns can make a big difference on what you can do with data and on the impact of your data solutions. Said in another way, the best AI models and the most impressive dashboards developed on the best technology might be completely useless if the underlying data cannot be used or cannot be shared.

I am honored to partner with Dr. Prashanth Southekal and AtScale in writing this foreword as well as being part of this project with such an impressive roster of data and analytics experts. Enjoy!

Chapter 1

Data Analytics and Competitive Advantage

Prashanth Southekal

Managing Principal, DBP Institute and Professor at IE Business School

Dr. Prashanth H Southekal is the Professor and Managing Principal of DBP-Institute, a Data Analytics Consulting and Education company. He brings over 20 years of Information Management experience from over 75 companies such as SAP, Shell, Apple, P&G, and GE. In addition, he has trained over 3000 professionals world over in Data and Analytics, and Enterprise Performance Management (EPM). He is the author of 2 books - Data for Business Performance and Analytics Best Practices and contributes regularly to Forbes.com. He is an Adjunct Professor of Data Analytics at IE Business School (Spain) where he received the teaching excellence award for the 2020-2021 academic year. Dr. Southekal holds a Ph.D. from ESC Lille (FR) and an MBA from Kellogg School of Management (US).

Chapter 1: Data Analytics and Competitive Advantage

Prashanth Southekal
Managing Principal, DBP Institute and Professor at IE Business School

Data analytics is an evolutionary process where the insight needs of the businesses constantly change, the organizational capabilities continuously mature, the data sets grow, and the technological capabilities to process the data improve over time. Working in data analytics projects often feels like shooting at a moving target. As a simple example, if one is looking for one value for the average price of crude oil between 2015 to 2018, it could be represented as mean, median, mode, weighted average, and more. While all these values are factual and correct, each of the above metrics gives a different crude oil price value and each value is subject to varied assumptions and conditions. In other words, the data is the same, but the context is different for each insight or metric.

While the primary purpose of analytics is to gain insights using data, insights can also be derived from intuition – the ability to understand or know something based on feelings rather than facts or data. Human beings are naturally irrational, and irrationality has defined much of human life. In a survey by The **Economist** in 2014, 73% of respondents said they trust their intuition over data when it comes to decision-making [Olavsrud, 2014]. This begs the question of when human intuition should be used and when should data be used for getting insights. Intuition is typically relied upon in the following situations.

- When time is too scarce to collect data, analyze it and derive insights. Personnel in the firefighting, ICU (intensive care unit), and police services rarely have much time to derive data-driven insights. In this scenario, the solution is to respond quickly.
- When there is a well-established or restricted range of actions, for example, when a car has a flat tire, the intuitive decision is to change the tire. You don't need data to make these obvious decisions as the choices are limited.
- Intuition is relied upon when the event is one-time or the first time. This means there is no historical data. For example, when Uber entered the ride-share business, there was little data available to analyze if people would accept this business model.

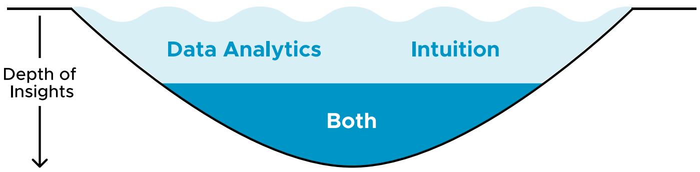
- When little justification is needed on the decision being made, intuition is preferred over data. This typically happens when homogenous stakeholders think and act alike or when the decision maker's authority or power is high.
- Finally, intuition is used when the impact or the repercussion of poor decision-making is low for the stakeholder(s). For a non-loyalty card customer, buying gasoline from Shell or Esso gas stations is almost the same. This situation normally happens when the products and services are highly commoditized with very few differentiating factors.

But insights driven from data apply in the following situations.

- Leveraging or mining data already captured for compliance and operations. Usually, data origination and data capture do not always start with analytics. Data origination and capture mainly start with operations and/or compliance, and analytics is typically pursued when the data volume reaches a critical size.
- When the hypothesis is complex with many interdependent variables, relying on intuition will not necessarily work. Research by George Miller of Harvard University has shown that the number of information pieces the human mind can simultaneously hold and then process is 7 ± 2 [Miller, 1956]. Most adults can store and process between 5 and 9 variables in their short-term memory. So, with complex situations with many interdependent variables, one needs to rely on data and the computing power of machines for running algorithms to derive insights.
- Finally, data-driven insights are needed when there are varied opinions and biases on the course of action. Instead of depending on people's opinions and even biases, the best course of action is to use data based on evidence or facts.

Insights derived from data analytics and intuition need not be mutually exclusive. Data analytics can even validate or test intuition. The combination of intuition and data analytics can offer holistic insights to the business by looking at the issue from various angles. Research by Nobel laureate Daniel Kahneman and his team found that all strategic decisions are evaluated using (1) numerical scores for competing options and/or (2) a yes-no decision on whether to choose a specific path. Insights from data are rational; they are carefully considered and negative outcomes are weighed. Irrational decisions are quick and based on intuitive judgment [Kahneman, 2019]. Hence, the best approach for a business is to combine data analytics and intuition, where data and intuition augment each other, so the insights are reliable and accurate for making good and holistic business decisions.

Delivering Deep Insights



Now, let us look at the fundamental question: Why do insights matter? Benjamin Disraeli, a British statesman and novelist, said: “One who has the most information will have the greatest success.” Louis Pasteur, French biologist and chemist renowned for his discoveries of vaccination, said: “Chance favors the prepared mind.” In this context, businesses are evolving entities that constantly seek insights to better prepare and adapt to the marketplace. And better respond to black swan event. A black swan is an unpredictable event that is beyond what is normally expected of a situation and has potentially severe consequences.

The need for adaptation can come from two scenarios – internal and external. First, from the internal perspective, businesses need insights to learn more about their operations and find avenues to increase revenues, decrease costs, and mitigate risks. Second, from the external perspective, businesses need insights to harness marketplace opportunities, address competition, and ultimately be relevant in the market.

In both scenarios, for the businesses to make the right decisions and be prepared, they need the best insights to take the organization to the desired state with optimal utilization of resources. The next chapter looks at ways Business Intelligence and data science can derive reliable insights and create measurable and monetizable value to the business.

References

Olavsrud, Thor, "Even Data-Driven Businesses Should Cultivate Intuition", https://bit.ly/2vZ8gu6, Jun 2014.
Miller, George, “The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information”, Psychological Review, 1956.
Kahneman, Daniel; Lovallo, Dan; and Sibony, Olivier, "A Structured Approach to Strategic Decisions", MIT Sloan Management Review, Mar 2019

Chapter 2

Business Value Creation with BI and Data Science

Ram Kumar

Chief Data and Analytics Officer, Cigna (International Markets)

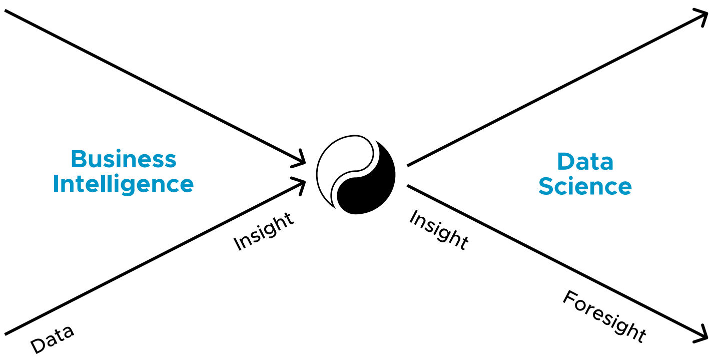
Ram Kumar is the Chief Data and Analytics Officer of Cigna's International Markets. He is responsible for driving data and analytics strategy and its execution for 30+ countries covering the Americas, EMEA, and the Asia Pacific. He has held many executive roles in his 32+ years career, including CEO, Group CTO, CIO, and Group Head of Data and Privacy. Ram has served as a member of the Data Research Advisory Board of MIT Sloan School, published over 150 articles, is a regularly invited keynote speaker in conferences globally and has spoken extensively. Ram holds a Master's degree in Computer Science and Engineering and a Bachelor's degree in Electronics and Communications Engineering with AI as a major in both.

Chapter 2: Business Value Creation with BI and Data Science

Ram Kumar
Chief Data and Analytics Officer, Cigna (International Markets)

There is no universally accepted definition for business intelligence (BI). Business intelligence is a term that could mean different things to different organizations and people. It may be just a set of basic key performance indicators (KPIs) on historical performance for some. But regardless of the methods or tools used, BI or data science provide insights for decision making. The context here is important because the same data sets used to create actionable insights for one business function could create different actionable insights for another business function.

It is imperative you need both BI and data science to create appreciable business value out of your data. Therefore, I call BI and data science the "Yin" and "Yang" of data-driven value creation for the business because they complement each other and each can solve different business problems. In simple terms, Business Intelligence (BI) includes operational reporting and dashboards, descriptive analytics, predictive analytics, and prescriptive analytics. All these types of analytics provide intelligence to the business through data analysis and insights by addressing different business questions and expectations.



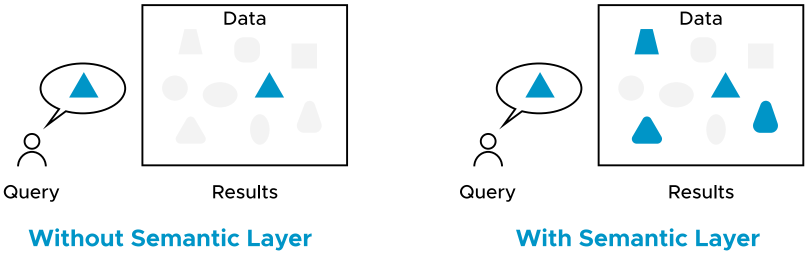
The Yin and Yang of Data-Driven Value Creation

Whether an organization uses BI, data science, or AI, there will be common challenges. Any organization that invests in these areas without a solid data foundation and management strategy and practice will struggle. Any processes or solutions built on poor data infrastructure and data processes will be slow, untrustworthy, poor quality, resource-intensive and expensive. With this backdrop, the following are key initiatives organizations should consider deriving value from BI and data science.

- **Data Democratization:** The organization's data assets should flow seamlessly and interoperate across its business processes and technology systems and reach the hands of knowledge workers. Without data democratization, consumers of data waste time searching for data, accessing the data, and waiting for approval. However, ensure that potential risks such as data ethics, data privacy, and the misuse of data and compliance requirements are managed. Data democratization is fundamental to driving the culture of a data-driven organization. It creates an opportunity for Business intelligence, data science and AI to work together to deliver data-driven value to the business in a collaborative, efficient, effective, consistent and reliable manner.

- **Semantic Layer:** Typically, data assets within organizations are distributed for various reasons – legacy systems, legacy processes, legacy culture, various servers, and more. This has created a massive gap between data sources and business users. Progressive companies are using semantic layers to help bridge this gap. Semantic Layer integrates complex enterprise data assets across the organization to provide a unified, consolidated view of data to provide complete and accurate analysis of data.
  A semantic layer is a business representation of data. It enables users to discover and access data through a self-service facility without relying on IT. It helps users understand the business definition of

On-the-Fly Data Discovery



data (e.g., customer, provider, products, etc.) and its relationships with other data assets. Even more, users could easily discover and integrate on-the-fly data from different datasets that used different descriptors. Perhaps the greatest feature of the semantic layer is it provides data professionals with easy access to the data needed for their specific roles and tasks while also providing data lineage and catalogue information about the data assets.

The semantic layer could be abstracted from the complexities of accessing and querying data from the user by providing a set of Data Service Interfaces (DSI) that could be integrated into business systems, whether internal or customer-facing and operational processes. Teams can work together more effectively with a robust semantic layer in place. They'll be able to collaborate on data value creation, derive better insights and make better decisions in less time. Data science requires close to 80% of the time in preparing data for value creation but these overheads can be reduced through the semantic layer. This layer also improves how the security and governance of the data assets are managed. The semantic layer, therefore, empowers data processes that democratize data.

- **Data Quality:** An organization may have the best business and operational processes: the best technologies, the best data scientists and AI practitioners to generate value, and the best data and digital strategies. Nevertheless, but if the underlying data that touches these people, processes and technologies is of poor quality, then the insights will be poor. The quality of outcomes generated through these initiatives is directly proportional to the quality of data. During the planning stage, organizations need to include "Data Quality of Design" disciplines, including data profiling, cleansing, and enrichment.

- **Data Governance:** It's critical to govern the democratized data, the semantic layer, data quality, and any other data-related risks. Traditional approaches to data governance where an organization tries to govern all its data assets through a big bang approach rarely work nowadays as businesses have neither the patience nor the budget. It is, therefore, important to apply smart governance frameworks by looking at governing the critical data elements that would create value for the business.

- **Data Driven Culture:** To drive the above initiatives and create a positive impact, it is critical for an organization to drive data driven culture across its people and this requires "accountability" to drive

> the same right from the top e.g., at the board level or CEO if the organization truly believes that data is a key strategic and competitive asset. My view is that any employee who touches data has a responsibility to nurture and manage the strategic asset. To drive this responsibility, a solid and well-defined data culture driving KPIs should be established to reward employees at all levels for their contribution to make it happen. It is important to note that building a data driven culture is not a program/project with an end date but is a cultural transformation journey with no end date.

Overall, measurement is an integral part of modern science. In today's VUCA centric world, performance measurement with BI, data science, and AI holds the key for predicting and explaining real-world phenomena. Effective performance measurement on the right transactional data using the right goals, hypotheses, questions, and KPIs will help companies identify their capabilities, set benchmarks, and adapt to the changing needs of the business. Though there are overlaps between BI and data science, the easiest way to differentiate them is to think of data science in terms of the future and BI in the past. Data science uses predictive and prescriptive analysis, while BI uses descriptive analysis. The next chapter from Bill Schmarzo discusses the salient differences between BI and data science.

Chapter 3

Difference Between BI and Data Science

Bill Schmarzo

Customer Advocate, Data Management Incubation, Dell Technologies

Bill Schmarzo “Dean of Big Data” is a recognized global innovator, educator, and practitioner in the areas of Big Data, Data Science, Design Thinking, and Data Monetization. He is currently part of Dell Technology’s core data management leadership team, where he is responsible for spearheading customer co-creation engagement to identify and prioritize the customers' key data management, data science, and data monetization requirements. Prior to Dell, he was the Chief Innovation Officer at Hitachi Vantara, CTO at Dell EMC, and VP of Analytics at Yahoo. Bill Schmarzo is the author of “Big Data: Understanding How Data Powers Big Business” and “Big Data MBA: Driving Business Strategies with Data Science”. He is currently an Adjunct Professor at Menlo College, an Honorary Professor at the University of Ireland – Galway, and an Executive Fellow at the University of San Francisco, School of Management. Bill holds an MBA from University of Iowa and a BS degree from Coe College.

Chapter 3: Difference Between BI and Data Science

Bill Schmarzo
Customer Advocate, Data Management Incubation, Dell Technologies

In the previous chapters, we said a lot about BI (business intelligence) and data science. But what exactly is the difference between the two? Defining data science and business intelligence and the relationship between them has led to a lot of discussions and arguments. Although these two terms are related, a good understanding of the salient characteristics and concepts can significantly impact how we realize value from data and analytics. A client recently asked me to explain to his management team the difference between BI) and data science. I frequently hear this question and typically resort to showing the figure below.

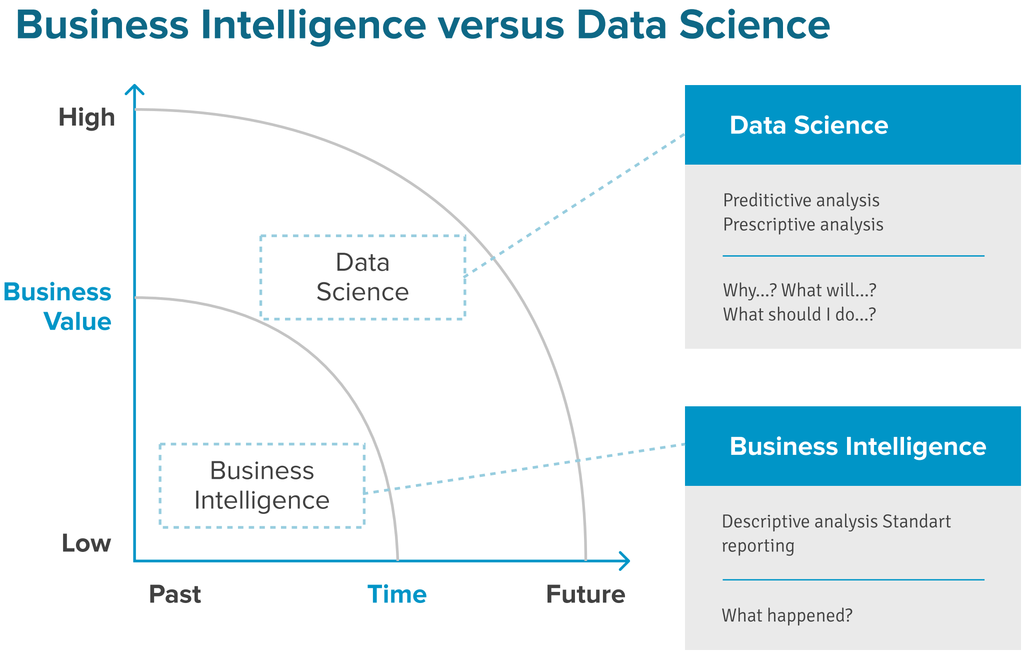
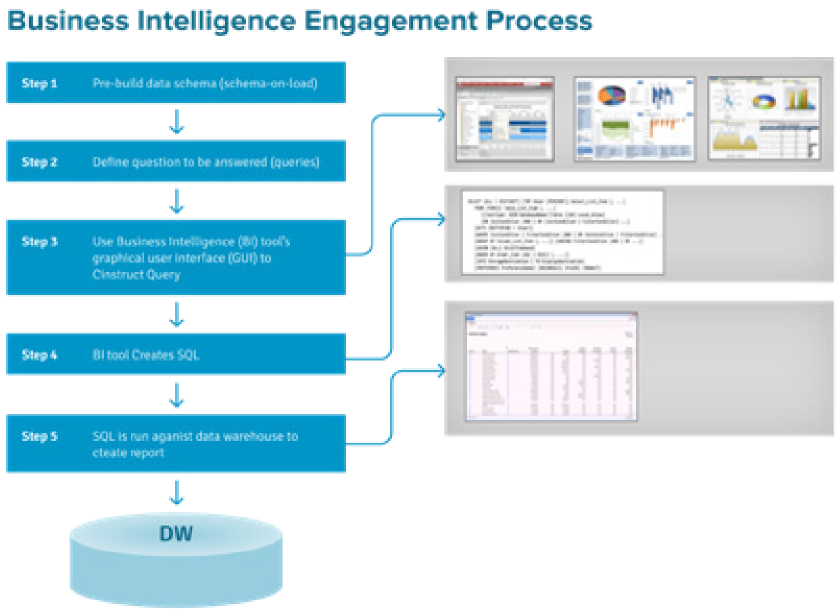


Figure 1: Business Intelligence vs. Data Science

To summarize, the key to understanding the differences between BI and data science lies in goals, tools, techniques and approaches. The figure below outlines the high-level analytics process that a typical BI analyst uses when engaging with business users.


Business Intelligence Engagement Process
Step 1
Pre-build data schema (schema-on-load)
Step 2
Define question to be answered (queries)
Step 3
Use Business Intelligence (BI) tool's graphical user interface (GUI) to Cinstruct Query
Step 4
BI tool Creates SQL
Step 5
SQL is run aganist data warehouse to cteate report
DW

Source: "Scientific Method: Embrace the Art of Failure", University of San Francisco School of Management Big Data MBA
Source: Bill Schmarzo "Big Data MBA" Course Curriculum

Figure 2: Business Intelligence Engagement Process

Step 1: Build the Data Model. The process starts by building the underlying data model. Whether you use a data warehouse or data mart or hub-and-spoke approach, or whether you use a star schema, snowflake schema, or third normal form schema, the BI analyst must go through a formal requirement gathering process with business users to identify pertinent questions that the business users want to answer. In these requirements gathering processes, the BI analyst must identify the first and second-level questions the business users want to address to build a robust and scalable data warehouse.

Once the key questions are formulated, the BI analyst then works closely with the data warehouse team to define and build the underlying data models that support the questions being asked.

> **Note:** The data warehouse is a "schema-on-load" approach because the data schema must be defined and built before loading data into the data warehouse. Without an underlying data model, the BI tools will not work.

Step 2: Define the Report. Once the analytics requirements have been transcribed into a data model, step two of the process occurs. This is where the BI analyst uses a BI product – SAP Business Objects, MicroStrategy, Cognos, QlikView, Pentaho, etc. – to create the SQL-based query for the desired questions (see Figure 3).

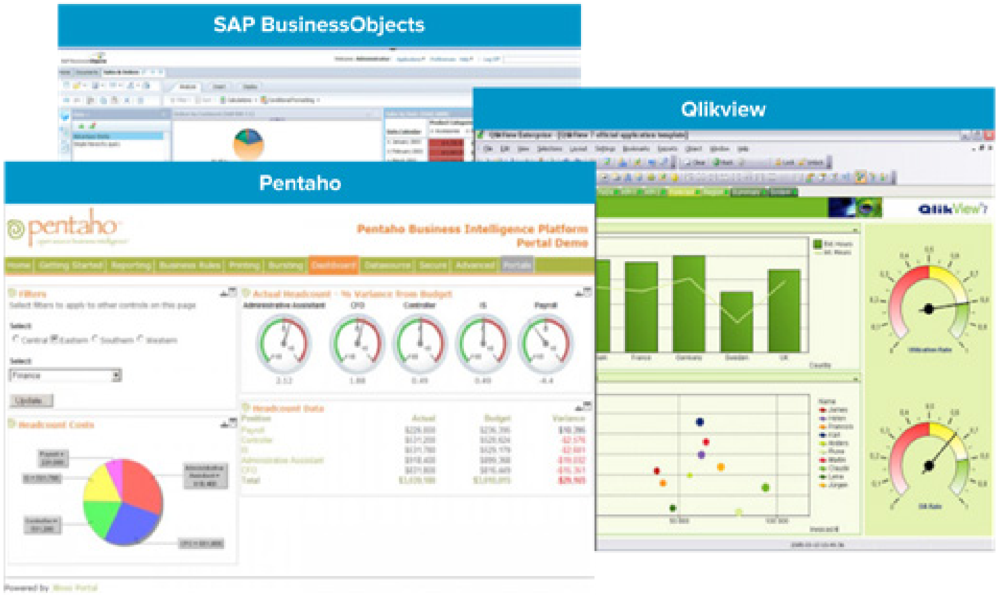


Figure 3: Business Intelligence (BI) Tools

The BI analyst will use the BI tool's graphical user interface (GUI) to create the SQL query by selecting: the measures and dimensions; the page, column and page descriptors; and the specifying constraints, subtotals and totals, creating special calculations (mean, moving average, rank, share of) and selecting sort criteria. The BI GUI hides much of the complexity of creating the SQL.

Step 3: Generate SQL commands. Once the BI analyst or the business user has defined the desired report or query request, the BI tool creates the SQL commands. Sometimes, the BI analyst will modify the SQL commands generated by the BI tool to include unique SQL commands that may not be supported by the BI tool.

Step 4: Create Report. In step 4, the BI tool issues the SQL commands against the data warehouse and creates the corresponding report or dashboard widget. This is a highly iterative process where the BI analyst will tweak the SQL (either using the GUI or hand-coding the SQL statement) to fine-tune the SQL request. The BI analyst can also specify graphical rendering options (bar charts, line charts, pie charts) until they get the exact report and/or graphic they want (see Figure 4).

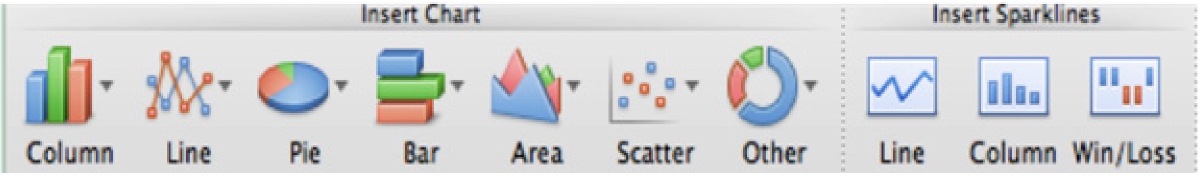
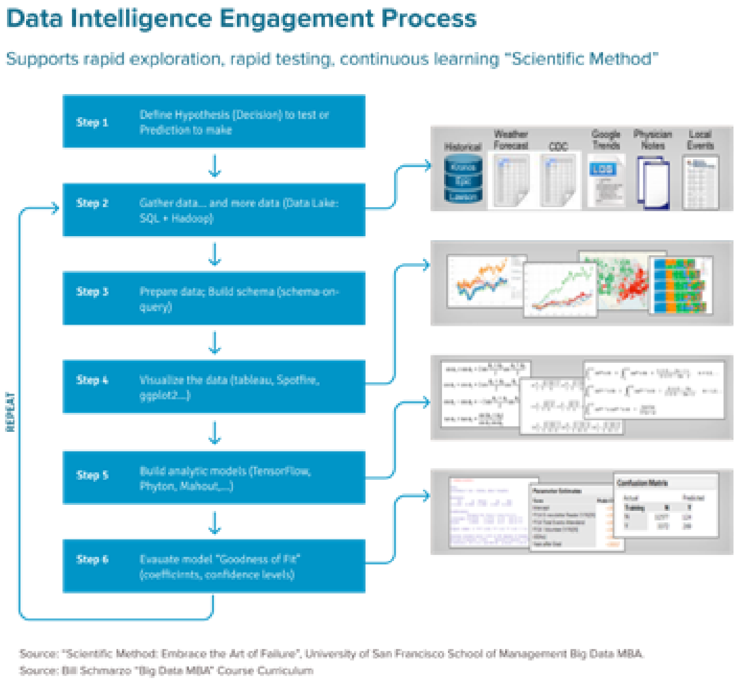


Figure 4: Typical BI Tool Graphic Options

This traditional schema-on-load approach removes much of the underlying data complexity from the business users who can then use the GUI BI tools to more easily interact and explore the data (think self-service BI).

In summary, the BI approach leans heavily on the pre-built data warehouse (schema-on-load), which enables users to quickly and easily ask questions – as long as the data they need is already in the data warehouse. If the data is not in the data warehouse, adding data to an existing warehouse (and creating all the supporting ETL processes) can take months. Now let's look at the management of the data science processes. The figure below shows the data science engagement process.



Data Intelligence Engagement Process

Supports rapid exploration, rapid testing, continuous learning "Scientific Method"

Step 1 Define Hypothesis (Decision) to test or Prediction to make

Step 2 Gather data... and more data (Data Lake: SQL + Hadoop)

Step 3 Prepare data; Build schema (schema-on-query)

Step 4 Visualize the data (tableau, Spotfire, ggplot2...)

Step 5 Build analytic models (TensorFlow, Phyton, Mahout,...)

Step 6 Evauate model "Goodness of Fit" (coefficirnts, confidence levels)

REPEAT

Source: "Scientific Method: Embrace the Art of Failure", University of San Francisco School of Management Big Data MBA.
Source: Bill Schmarzo "Big Data MBA" Course Curriculum

Figure 5: Data Science Engagement Process

Step 1: Define Hypothesis To Test. Step 1 of the data science process starts with the data scientist identifying the hypothesis to be tested. Again, this results from collaborating with the business users to understand the key sources of business differentiation (e.g., how the organization delivers value) and then brainstorming data and variables that might yield better predictors of performance. A Vision Workshop can enhance the collaboration between the business users and the data scientists to identify data sources that may help improve the predictive value.

Step 2: Gather Data. Step 2 of the data science process is where the data scientist gathers relevant and/or interesting data from multiple sources – ideally both internal and external to the organization. The data lake is a great approach for this process, as the data scientist can grab any data they want, test it, ascertain its value given the hypothesis or prediction, and then decide whether to include that data in the predictive model or throw it away.

Step 3: Build Data Model. Step 3 is where the data scientist defines and builds the schema to address the hypothesis being tested. The data scientist can't define the schema until they know the hypothesis they are testing AND know what data sources they will be using to build their analytic models.

> **Note:** This "schema on query" process is notably different than the traditional data warehouse "schema on load" process. The data scientist doesn't spend months integrating all the data sources into a formal data model first. Instead, the data scientist will define the schema as needed based upon the data being used in the analysis. The data scientist will likely iterate through several versions of the schema until finding a schema (and analytic model) that sufficiently answers the hypothesis being tested.

Step 4: Explore The Data. Step 4 of the data science process leverages the outstanding data visualization tools to uncover correlations and outliers of interest in the data. Data visualization tools like Tableau, Spotfire, and AtScale are great data scientist tools for exploring the data and identifying variables they might want to test (see Figure 6).

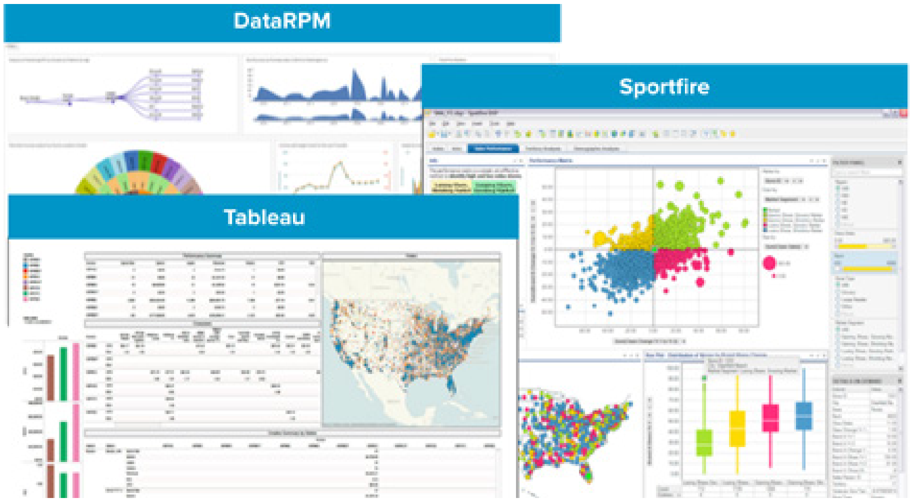
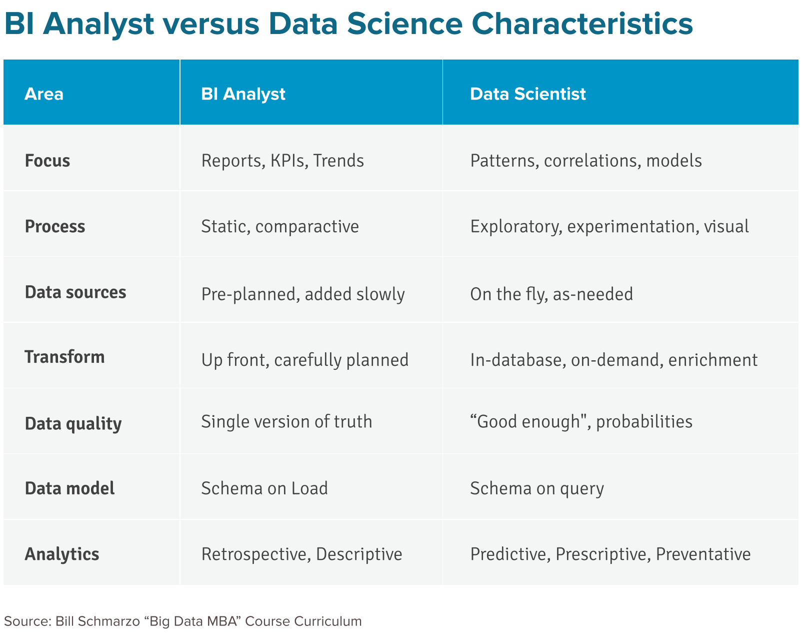


Figure 6: Sample Data Visualization Tools

Step 5: Build and Refine Analytic Models. Step four is where the real data science work begins and where the data scientist uses tools like SAS, Python, PowerBI and R, to build analytic models. The data scientist will explore different analytic techniques and algorithms such as Markov chain, genetic algorithm, geofencing, individualized modeling, propensity analysis, neural network, Bayesian reasoning, principal component analysis, singular value decomposition, optimization, and linear and non-linear programming.

Step 6: Ascertain Goodness of Fit. Step five in the data science process is where the data scientist will ascertain the model's goodness of fit. The goodness of fit of a statistical model describes how well the model fits a set of observations. A number of different analytic techniques will determine the goodness of fit, including Kolmogorov–Smirnov test, Pearson's chi-squared test, analysis of variance (ANOVA) and confusion (or error) matrix.

The table below summarizes the differences between BI (performed by the BI analyst) and data science (performed by the data scientist).



BI Analyst versus Data Science Characteristics

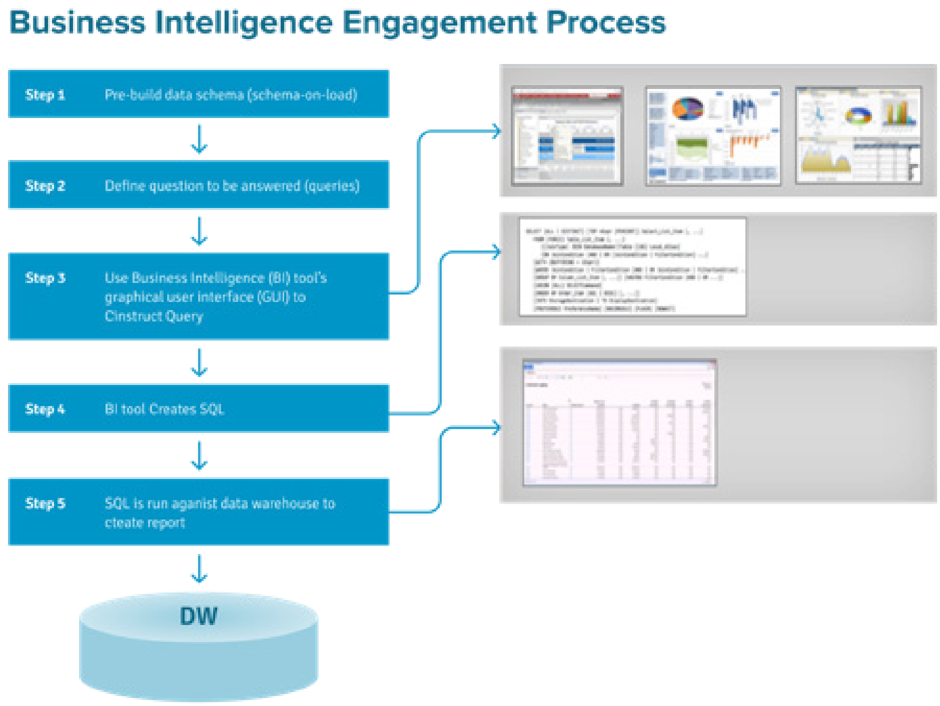
Area	BI Analyst	Data Scientist
Focus	Reports, KPIs, Trends	Patterns, correlations, models
Process	Static, comparactive	Exploratory, experimentation, visual
Data sources	Pre-planned, added slowly	On the fly, as-needed
Transform	Up front, carefully planned	In-database, on-demand, enrichment
Data quality	Single version of truth	"Good enough", probabilities
Data model	Schema on Load	Schema on query
Analytics	Retrospective, Descriptive	Predictive, Prescriptive, Preventative

Source: Bill Schmarzo "Big Data MBA" Course Curriculum

Figure 7: BI Analyst V/s Data Scientist

Basically, business intelligence and schema-on-load, and data science and schema-on-query, have specific use-cases where they solve a specific business need. They address different questions. They are different approaches, intended for different environments, and used at different stages in the analysis process. In the BI process, the schema must be built first and must be built to support a wide variety of questions across a wide range of business functions. So, the data model must be extensible and scalable, so it is heavily engineered. Think production quality. In the data science process, the schema is built to support only the hypothesis being tested so the data model can be done more quickly and with less overhead.

Think ad hoc quality. Figure 6 below depicts the relationship between BI and data science.



Business Intelligence Engagement Process

Step 1 Pre-build data schema (schema-on-load)

Step 2 Define question to be answered (queries)

Step 3 Use Business Intelligence (BI) tool's graphical user interface (GUI) to Cinstruct Query

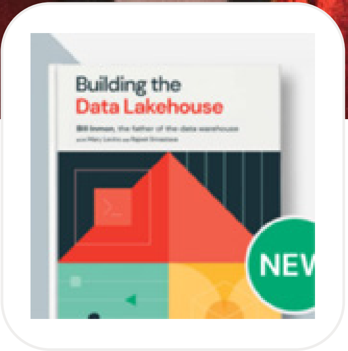
Step 4 BI tool Creates SQL

Step 5 SQL is run aganist data warehouse to cteate report

Source: "Scientific Method: Embrace the Art of Failure", University of San Francisco School of Management Big Data MBA
Source: Bill Schmarzo "Big Data MBA" Course Curriculum

Figure 8: Sample Data Visualization Tools

The data science process is highly collaborative; the more subject matter experts involved, the better the resulting model. And maybe even more importantly, business users should be involved throughout the process to ensure that data scientists focus on uncovering analytic insights that pass the S.A.M. test – Strategic (to the business), Actionable (insights that the organization can act on), and Material (where the value of acting on the insights is greater than the cost of acting on the insights). Both data science and BI rely on a canonical data source to derive insights. This canonical data source could be a data mart, data warehouse, data lake or even a Data Lakehouse. The next chapter from Bill Inmon, "Father of Data Warehouse," looks at how a Data Lakehouse can solve some of the big challenges in the world of data and analytics.



Chapter 4

Evolution to the Data Lakehouse

Bill Inmon

“Father of Data Warehouse”
Founder & CEO, Forest Rim Technologies

Bill Inmon – the “father of data warehouse” – has written 60 books published in nine languages. Bill’s latest adventure is building technology known as textual disambiguation (textual ETL) – technology that reads the raw text and allows the text to be placed in a conventional database so it can be analyzed by standard analytical technology. ComputerWorld named Bill as one of the ten most influential people in the history of the computer profession. Bill lives in Denver, Colorado. For more information about textual disambiguation (textual ETL), refer to www.forestrimtech.com. Three of Bill’s latest books are DATA ARCHITECTURE: SECOND EDITION, Elsevier press, HEARING THE VOICE OF THE CUSTOMER, Technics Publications, and TURNING TEXT INTO GOLD, Technics Publications.

Chapter 4: Evolution to the Data Lakehouse

Bill Inmon "Father of Data Warehouse"
Founder & CEO, Forest Rim Technologies

With the proliferation of IT applications came the problem of data integrity. The problem with the advent of large numbers of applications was that the same data appeared in many places with different values. The user had to find out WHICH version of the data was the right one to use among the many applications to make a decision. If the user did not find and use the right version of data, the organization might make incorrect decisions. Organizations needed a different architectural approach to find the right data for decision-making. Thus, the data warehouse was born. The data warehouse caused disparate application data to be placed in a separate physical location. The designer had to build entirely new infrastructure around the data warehouse.

So, what is a data warehouse? A data warehouse is a database that pulls together data from many source systems within an organization for reporting and analysis. They store historical data in one place used for creating analytical reports throughout the enterprise. The analytical infrastructure surrounding the data warehouse contains such things as:

- Metadata – A guide to what data is located where.
- A data model – An abstraction of the data found in the data warehouse.
- Data lineage – The tale of the origins and transformations of data in the warehouse.
- Summarization – A description of the algorithmic work designed to create the data.
- KPIs – Where the key performance indicators are found.
- ETL – Enables application data to be transformed into corporate data

The limitations of data warehouses became evident with the increasing variety of data in the enterprise (text, IoT, images, audio, videos, etc.). In addition, the rise of machine learning (ML) and AI introduced iterative algorithms that required direct data access and were not based on SQL.

As important and useful as data warehouses are, mostly, data warehouses center on structured data. But now, there are many other data types in a corporation, including structured, text or other forms of unstructured data. Structured data is typically transaction-based data generated by an organization to conduct day-to-day business activities. Textual data is data generated by letters, email and conversations inside the corporation. Other unstructured data is data with other sources, such as IoT data, image, video and analog-based data.

The second most common canonical database is the data lake. The data lake is an amalgamation of all the different kinds of data found in the corporation. It has become the place where enterprises offload all their data because of its low-cost storage systems with a file API that hold data in generic and open file formats, such as Apache Parquet and ORC. Using open formats also made data lake data directly accessible to a wide range of other analytics engines such as machine learning systems.

When the data lake was first conceived, the assumption was that all that was required was that data should be extracted and placed in the data lake. Once in the data lake, the end-user could just dive in and find data and do analysis. However, corporations quickly discovered that using the data in the data lake was different from merely having the data placed in the lake. Many promises of data lakes have not been realized due to the lack of critical features such as no support for transactions, data quality or governance enforcement, and poor performance optimizations. Most of the data lakes in the enterprise have become data swamps.

So, what are the challenges with current data architecture? Due to the limitations of data lakes and warehouses, a common approach is to use multiple systems – a data lake, several data warehouses, and other specialized systems, resulting in three common problems:

1. **Lack of openness:** Data warehouses lock data into proprietary formats that increase the cost of migrating data or workloads to other systems. Data warehouses primarily provide SQL-only access, so it is hard to run any other analytics engines, such as machine learning systems. It is very expensive and slow to directly access data in the warehouse with SQL, making integrations with other technologies difficult.

2. **Limited support for machine learning:** Despite much research on the confluence of ML and data management, none of the leading machine learning systems, such as TensorFlow, PyTorch and XGBoost, work well on top of warehouses. Unlike BI, which extracts a small amount of data, ML systems process large datasets using complex non-SQL code. For these use cases, warehouse vendors recommend exporting data to files, which further increases complexity and staleness.

3. **Forced trade-off between data lakes and data warehouses:** over 90% of enterprise data is stored in data lakes due to their flexibility from open direct access to files and their low cost (data lakes use cheap storage). To overcome the data lake's lack of performance and quality issues, enterprises ETLed a small subset of data in the data lake to a downstream data warehouse for the most important decision support and BI applications. This dual system architecture requires continuous engineering to ETL data between the lake and warehouse. Each ETL step risks incurring failures or introducing bugs that reduce data quality; meanwhile, keeping the data lake and warehouse consistent is difficult and costly. Besides paying for continuous ETL, users pay double the storage cost for data copied to a warehouse.

These issues led to the emergence of the Data Lakehouse, which is enabled by a new open and standardized system design: implementing similar data structures and data management features to those in a data warehouse, directly on the kind of low-cost storage used for data lakes.

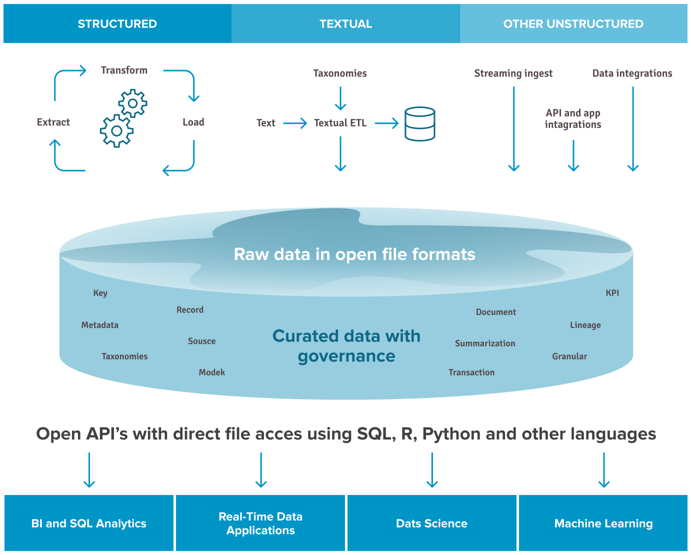


Figure 1: Data Lakehouse

Data Lakehouse architecture addresses the key challenges of current data architectures discussed in the previous section by enabling open direct access by using open formats, such as Apache Parquet. This provides native class support for data science and machine learning, offering best-in-class performance and reliability on low-cost storage. Here are the various key benefits of a Data Lakehouse architecture:

Openness:

1. Open File Formats: Built on open and standardized file formats, such as Apache Parquet and ORC.
2. Open API: Provides an open API that can efficiently access the data directly with no proprietary engines and vendor lock-in.
3. Language Support: Supports SQL access and a variety of other tools and engines, including machine learning and Python/R libraries.

Machine learning support:

1. Support for diverse data types: Store, refine, analyze and access data for many new applications, including images, video, audio, semi-structured data and text.
2. Efficient non-SQL direct reads: Direct efficient access of large volumes of data for running machine learning experiments using R and Python libraries.
3. Support for DataFrame API: Built-in declarative DataFrame API with query optimizations for data access in ML workloads since ML systems such as TensorFlow, PyTorch and XGBoost have adopted DataFrames as the main abstraction for manipulating data.
4. Data Versioning for ML experiments: Providing data snapshots, enabling data science and machine learning teams to access and revert to earlier versions of data for audits and rollbacks or to reproduce ML experiments.

Best-in-class performance and reliability at low cost:

1. Performance optimizations: Supports various optimization techniques -- caching, multi-dimensional clustering and data skipping -- by leveraging file statistics and data compaction to right-size the files.
2. Schema enforcement and governance: Supports DW schema architectures like star/snowflake-schemas and provides robust governance and auditing mechanisms.
3. Transaction support: Leverages ACID transactions to ensure consistency as multiple parties concurrently read or write data, typically using SQL.

4. Low-cost storage: Lakehouse architecture is built using low-cost object storage such as Amazon S3, Azure Blob Storage or Google Cloud Storage.

The table below compares data warehouse, data lake, and Data Lakehouse features.

	Data warehouse	Data lake	Data lakehouse
Data format	Closed, proprietary format	Open format	Open format
Types of data	Structured data, with limited support for semi-structured data	All types: Structured data, semi-structured data, textual data, unstructured (raw) data	All types: Structured data, semi-structured data, textual data, unstructured (raw) data
Data access	SQL-only, no direct access to file	Open APIs for direct access to files with SQL, R, Python and other languages	Open APIs for direct access to files with SQL, R, Python and other languages
Reliability	High quality, reliable data with ACID transactions	Low quality, data swamp	High quality, reliable data with ACID transactions
Governance and security	Fine-grained security and governance for row and columnar level for tables	Poor governance as security needs to be applied to files	Fine-grained security and governance for row/ columnar level for tables
Performance	High	Low	High
Scalability	Scaling becomes exponentially more expensive	Scales to hold any amount of data at low cost, regardless of type	Scales to hold any amount of data at low cost, regardless of type
Use case support	Limited to BI, SQL applications and decision support	Limited to machine learning	One data architecture for BI, SQL and machine learning

So, what is the business impact of the Data Lakehouse? The Data Lakehouse architecture presents an opportunity comparable to the one we saw during the early years of the data warehouse market. Data Lakehouses have the potential to unlock incredible value for organizations due to their unique ability to manage data in an open environment, blend varieties of data from all parts of the enterprise and combine the data science focus of data lakes with the end-user analytics of data warehouses.

Chapter 5

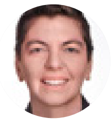
Data Literacy and Business Intelligence Drive AI/ML

Megan Brown

Megan Brown, PhD, Director, Knowledge Management & Data Literacy, Starbucks

Megan C. Brown (they/them) is the Director of Knowledge Management and Data Literacy at Starbucks. They build high performing, inclusive teams that deliver forward-thinking data products, build trusted relationships across the business, and remove obstacles to making decisions with data and analytics. Prior to their current role, they were a data scientist in the people and marketing spaces. They are a quant research psychologist (with experience in experimentation, inferential statistics, econometrics, and ML) who enjoys teaching, writing, and speaking about data literacy, data science, and how to use analytics to make data-informed decisions.

Chapter 5: Data Literacy and Business Intelligence Drive AI/ML



Megan Brown, PhD
Director, Knowledge Management & Data Literacy, Starbucks

Business intelligence (BI) and data literacy have a symbiotic relationship: one can exist without the other, but your business will work best when they are both present. If you have a data-literate organization and few BI tools, you likely have a frustrated workforce. Here, your employees want to use data to make decisions, but can't. If you have BI tools but lack data literacy, you have limited adoption of your data, analytics, research insights, and visualizations. Your employees could make data-informed decisions, but they don't. In both situations, your employees aren't using the insights from data to make decisions.

Let's imagine an individual contributor in your marketing department. If you asked them to name your top five business KPIs, could they? (Hopefully, they could!) When they rattle them off accurately, can they state the enterprise definition for those KPIs? If they cannot, do they know where to find that information? Do they know how to access those KPIs via the most reliable, updated, governed source? Are they comfortable discussing those KPIs with leaders? (You can ask the same questions about departmental KPIs and program performance metrics.)

Individual data literacy is the ability to read, work with, analyze, and communicate with data in context [Baykoucheva, 2015]. Generally, employees aren't confident in their data literacy skills - 21% of employees reported being confident in their data literacy [Vohra & Morrow, 2020]. Of course, this matters because "Data-driven decisions markedly improve business performance," [Bersin & Zao-Sanders, 2020].

When an organization has weak BI tools and/or weak data literacy, they come to rely on their data and analytics teams for low-level requests. This can be trouble - it's not possible for an analytics team to scale enough to make up for limited tools and data skills. Often (expensive) analytics employees would be best applied to higher-order products, rather than support of small-scale decisions. In fact, limited data literacy is one of the top-3 barriers to building effective, strategic data and analytics teams [Goasduff, 2020]. Further, your business's analytics professionals may be too removed from business decisions to put the insights into use -- they depend on their business stakeholders to understand and apply the findings. If the stakeholders aren't data literate, the analytics team's effort can go unused.

How do you build that level of data literacy? How can businesses leverage the power of data and analytics? Organizational data literacy requires data literate employees. These employees create pressure for new employees to become data literate quickly and help skill up the folks around them. Data literate organizations also have easily accessible, well-governed BI tools connected to searchable metadata (including governance information). There should also be a pathway for employees to request more advanced analytics or suggest interesting questions that can be addressed with analytics.

Further, the organization needs to support data literacy in two ways: education and leadership advocacy. To build on the expectation that employees become data literate, organizations should create educational programs relevant to their department's work. Specifically, your business should teach employees (ideally in their first months) how to find, access, and summarize your business data using existing BI tools. Bonus points go to organizations that embed links to educational offerings on dashboards and in metadata search. Leaders who advocate for their teams' data literacy growth and commit resources to the endeavor will make progress faster.

There are many risks of not achieving organizational data literacy. Your employees will trust neither your data nor your BI tools. This could lead to several poor behaviors: hesitation to use data in presentations, rejection of reliable insights, and low adoption of new BI views. Here, decisions are made without including data and analytics in the portfolio, leading to suboptimal business decisions. You would likely have overburdened and underutilized data and analytics teams, as business employees attempt to outsource their data efforts. Your analytics employees will not have the resources to dig into root causes and complicated problems. Worst of all, your employees' imaginations will be limited about the growth and disruption possible with the use of higher-level statistical methods, machine learning, and artificial intelligence.

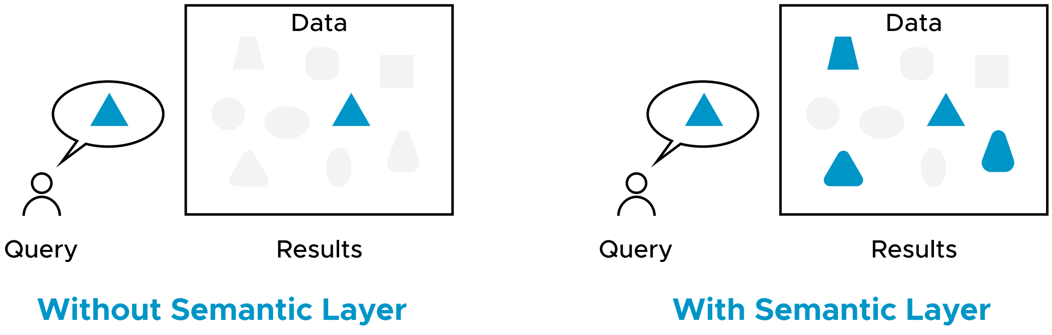
The benefits of organizational data literacy will have the opposite effect: insights from data and analytics become a part of your business language. Presentations without data are challenged and conflicts between data and experience are directly addressed. Employees are confident that the data they've woven into their pitches are accurate and reliable. Data, analytics, and advanced analytics are a part of a portfolio of information to make strategic decisions. Your stakeholders and analytics teams work together to resolve hard problems and the insights are used immediately to drive better decisions.

Organizational data literacy is fundamental for building employee curiosity about and trust in more advanced analytics methods. To create that true competitive advantage, your employees need to know enough about advanced analytics, experimentation, machine learning, and artificial intelligence to do the following: identify opportunities; advocate for the most impactful AI/ML solutions; collaborate with the tech teams building the solutions; and speak intelligently about the benefits, risks, and costs of building AI/ML.

Even if most of your employees aren't setting the strategic roadmap for AI/ML at your company, their knowledge of AI/ML supports business change. First, your employees close to the heart of your business will see the greatest opportunities for AI/ML solutions to support their work. They can likely speak to rote tasks, simple decisions, or places where they'd like a recommendation based on prior successes. They'll only be able to view those situations as an opportunity for AI/ML if they know what's possible. Second, they'll be more supportive of the changes created by AI/ML, especially if it helps them deliver their best work.

There are several topics that need to be addressed to build AI/ML literacy. Types of AI/ML algorithms should be presented in an easy-to-digest manner. Ongoing conversations about how your business plans to use AI/ML can allay fears employees have about "being replaced" by robots. Access to AI/ML metadata (e.g., algorithm type, predictors, outputs, purpose, limitations) will improve transparency and build trust. Much like data governance, your company's approach to AI ethics is a necessary topic of discussion. Where are the ethical boundaries for your company? How will you confirm that you are building fair models? Addressing these concerns outright puts your company in a leading position.

On-the-Fly Data Discovery



Without Semantic Layer

With Semantic Layer

Creating the right educational opportunities is the key. Formal courses and workshops are likely the best paths to teach employees different AI/ML algorithms. AI ethics is likely best handled in a more informal, conversational setting. Existing AI/ML solutions can be presented in videos, large team meetings, and departmental quarterly business reviews. Employees taking these opportunities will a) meet the folks who build that AI/ML solutions (for future collaboration) b) learn categories of algorithms (e.g., prediction, recommendation, forecasting, grouping...) and what sorts of questions they answer; c) discuss the risks, rewards, and limitations of different algorithms; d) learn the process for proposing and building AI/ML solutions; and e) learn what AI/ML solutions are already in use at your company. There is also power in creating business-analytics-tech groups of AI-minded leaders who can share ideas, evaluate opportunities, and decide where and when to invest.

To summarize, individual data literacy and solid BI tools are necessary to support organizational data literacy. Leadership advocacy and well-designed educational opportunities will accelerate the growth of organizational data literacy. However, data literacy is only the beginning. To craft an aggressive move into AI/ML, businesses must build AI/ML literacy on top of organizational data literacy. AI/ML literacy builds employee trust and helps them feel comfortable with AI/ML solutions. It creates the right environment for an organization to identify opportunities for AI/ML, evaluate those opportunities for impact and cost, invest in the best opportunities, and then put those solutions into practice to create a competitive advantage for your business.

References

Citation: (Baykoucheva, Svetla (2015). Managing Scientific Information and Research Data. Waltham, MA: Chandos Publishing. p. 80. ISBN 9780081001950)
Citation: (Vohra & Morrow, 2020, The Human Impact of Data Literacy, https://www.accenture.com/us-en/insights/technology/human-impact-data-literacy)
Citation: (Bersin & Zao-Sanders, 2020, Boost Your Team's Data Literacy, https://hbr.org/2020/02/boost-your-teams-data-literacy)
Citation: (Goasduff, 2020, Avoid 5 Pitfalls When Building Data and Analytics Teams, https://www.gartner.com/smarterwithgartner/avoid-5-pitfalls-when-building-data-and-analytics-teams)

Chapter 6

Delivering Actionable Insights with Modern Data Platforms

Naratav

Brian Prascak

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

Brian Prascak is Co-founder, Chief Insights Officer at Naratav, Inc. Naratav is an analytics software company that utilizes AI to provide Autonomous Insights and Automated Data Storytelling that are more relevant and personalized. Brian has helped many organizations become smarter faster using data, insights and analytics, particularly customer and marketing analytics. Brian has held leadership positions in business, technology, analytics and consulting, including at IBM, JPMorgan, Diageo, Mastercard, ACNielsen and Wawa.

Chapter 6: Delivering Actionable Insights with Modern Data Platforms

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

Reaping the benefits from utilizing modern data platforms to deliver timely, actionable data, insights and analytics (DIA) for the enterprise requires establishing well-defined strategies, capabilities and priorities aligned to targeted opportunities, intended performance, scale and maturity. The promise is to be "smarter faster" – creating actionable insights – delivered through modern data platforms that improve the confidence and clarity of the answers we seek, the decisions we make and the actions we take; delivered rapidly, flexibly and cost-effectively.

The explosive growth in using modern cloud-based data platforms with DIA to improve business performance is driven by five (5) major catalysts:

1. Big Data – Data is the fuel for becoming smarter faster. Data provide the insights, including feedback required to improve all facets of understanding, planning, deciding and acting. Key to realizing the promise of using "Big" data (volume, variety, velocity,) to be smarter faster, is having a data source strategy and capabilities to ensure data is available, accurate, actionable and automated.
2. Cloud Technology – Seldom does a technological transformation address so many needs than the Cloud. Cloud enables data access, storage, and processing of large amounts of diverse data in a central location in a very efficient manner.
3. Advanced Analytics – Advances in algorithmic-based analytical models, supported by the practice of data science, particularly machine learning, enables rapid, automated learning from data.
4. Modern Data Platforms – Commensurate with the increased maturity of and migration to the cloud, data platforms such Enterprise Data Platform, Customer Data Platform, Master Data, Data Governance and Data Science and Business Intelligence have grown exponentially. Each of these platforms are crucial to realizing the benefits of becoming smarter faster, because they are offered as a service in the cloud, which makes it easier to access, utilize and operate, including typically more cost-effective.
5. Data Virtualization –Companies realize the need to harness their data and how their data is being

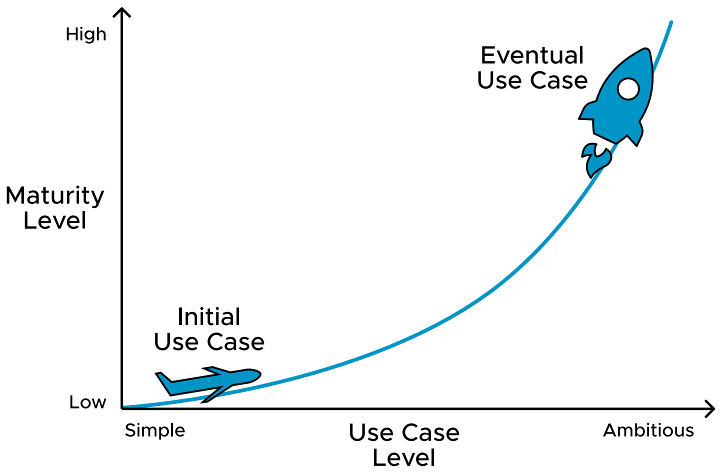
> defined, structured, and analyzed. This has created the need to create a common “semantic layer” for describing data attributes and metrics (often called “features”) consistently using common business terms enables these features to be understood and reused. Data virtualization is also called the semantic layer, as it offers the capability for data to be viewed and searched quickly across the organization.

Against this backdrop, critical to making a case for using the modern data platforms, including modernizing data science and business intelligence capabilities, is identifying the major use cases and benefits, including financial benefits that will support enabling use case realization supported by capability transformation. Given that these platforms support multiple use cases and the initial investment to stand up and operate the new platforms, it is advised that companies link investment to a major set of company-wide transformations that support core enterprise strategic growth and operational improvement objectives. Although a core set of strategic initiatives should make a case for moving to the modern data platforms, there will be ongoing benefits from serving the entire enterprise – work to ensure that Data, Insights and Analytics modernization effort doesn’t get pigeonholed into and associated with a couple of initial strategic initiatives as there should be many more initiatives to that will benefit from modernizing DIA capabilities and support.

Critical to the strategy is the support from C Suite and key functional leads. The success of any major enterprise capability transformation requires transparent commitment from the C-suite and key functional leads, and the transformation to being smarter faster through DIA is no exception. Perhaps more so, since for many organizations that face the need to embrace digital transformation, the need to embrace DIA transformation is just as critical. Many organizations that undertake digital transformation also have a team addressing DIA transformation, as the two transformations complement each other.

Once the organization has determined that it is ready to begin the journey to smarter faster, the first step is to take stock of capabilities and establish a target state maturity level guided by an initial set of use cases, investment and benefit case and 3-year roadmap. Establishing an enterprise maturity model for data, insights and analytics is critical to establishing the agenda for DIA, including use cases, roadmap, investments and impacts. Further, the maturity levels provide a guidepost for tethering use cases and initiatives to demonstrate progress and assist with data platform and capability mobilization priorities.

The 'Smarter Faster' Journey



The most critical piece of the effort is identifying the initial set of use cases and the broader agenda that will fund and fuel the transition to the modern data platforms that enable becoming smarter faster. Ideally, use cases represent major business improvement areas where DIA can provide the greatest impact. Example use cases could support digital transformation, supply chain modernization, pricing yield and promotion lift optimization, customer relationship management, and personalization of product recommendations and communications. It is important to lay out the use cases aligned to the maturity level required, so the business case is string and resonates well with the stakeholders.

The importance of establishing a strategy and plan for how data sources are acquired and utilized across the enterprise is often overlooked. Organizations with low maturity have many data sources in disparate locations and conditions of readiness for insights delivery. Further, data science teams depend on having rapid access to a multitude of data sources to develop effective analytical models. Finally, moving data for analysis to the enterprise data lake should be done with prioritization in mind so the cost to set up automated data pipelines, data structure and reporting from newly integrated data sets can be aligned to the agreed-upon use cases. An effective data source strategy is crucial to ensuring that the improvement in both breadth and depth of insights is aligned to the data, and that all data sources be seen as being available for enterprise use. An effective, aligned data source strategy typically consists of the following:

- Identify Use Cases and Data Source Subjects
- List Key Questions to be addressed for each use case
- Explain how data source(s) address the questions
- Determine data source availability and readiness
- Establish a roadmap for data source acquisition and distribution

Overall, the path to smarter faster is certainly a journey! Fortunately, it is a journey proven to deliver winning when it counts, including incremental revenue and customer growth, market share, customer loyalty and profitability across a multitude of industries applied to a wide variety of use cases. Today, more than ever, the available advice, experience, platforms, tools, capabilities, skills and resources are available to enable a successful journey, whether you are just beginning or ready to jump to the next level of maturity. We will always seek more accurate, actionable answers to our questions, and harvesting data, insights and analytics is the substance we need to succeed to ensure that we do not drown in data and become starved for meaningful answers, insights and feedback.

Chapter 7

AI for BI – Bridging the Gaps

Michael Gregory

Field CTO lead, Data Science and ML, Snowflake

Michael Gregory is the tech lead for the machine learning team in the Field CTO office at Snowflake, where he helps organizations leverage the power of Snowflake to build mature, enterprise-grade data science and machine learning capabilities. Before joining Snowflake, Michael held leadership positions in AI and ML go-to-market at Amazon Web Services and Cloudera and has deep experience in bringing ML products to market and building high-powered technical sales teams. Michael leverages over 20 years of experience (including Sun Microsystems, Oracle, and Teradata) in designing, building, selling, and architecting data and analytics solutions to help organizations leverage the power of data to build communities.

Chapter 7: AI for BI – Bridging the Gaps

Michael Gregory
Field CTO lead, Data Science and ML, Snowflake

Artificial Intelligence and Machine Learning have spread to every industry, across geographies, and throughout nearly every organizational function. In the last few years, the question has moved from

IF there is value in using AI and ML to HOW to best use it. Across both the public and private sector, organizations are looking to move beyond experimentation and prototyping to weaving these technologies into the fabric of their processes, products, and services. They are, rightly so, working steadfastly to create standards and processes to manage risk and build consistently.

However, most organizations encounter challenges when implementing AI and BI solutions. There are many reasons for these challenges, including formulating a strong use-case, having very few standard approaches, inconsistent definitions, the lack of best practices, and more. It’s no wonder that many AI and ML leaders feel like they’re pushing a rope uphill as they strive to build standard platforms and repeatable processes across technologies, personas, skill sets and departments.

At the heart of this discussion is the fact that AI and ML are being built into many scenarios, from self-driving cars, virtual-reality video games, and security operations tools to business intelligence dashboards, marketing segmentation, policing anti-money laundering, and payment fraud. While the same mathematics and many of the same algorithms are underlying all these use cases, they differ in how they are implemented, the risk factors involved, and the governance structure, skillsets and infrastructure needed. These can all be considered AI projects, but without a framework to differentiate them, it becomes very difficult for organizations to create consistency and standards supporting scale and efficiency.

This chapter will discuss a framework to consider these different use-case areas to drive meaningful conversations, effective processes and value-added infrastructure.

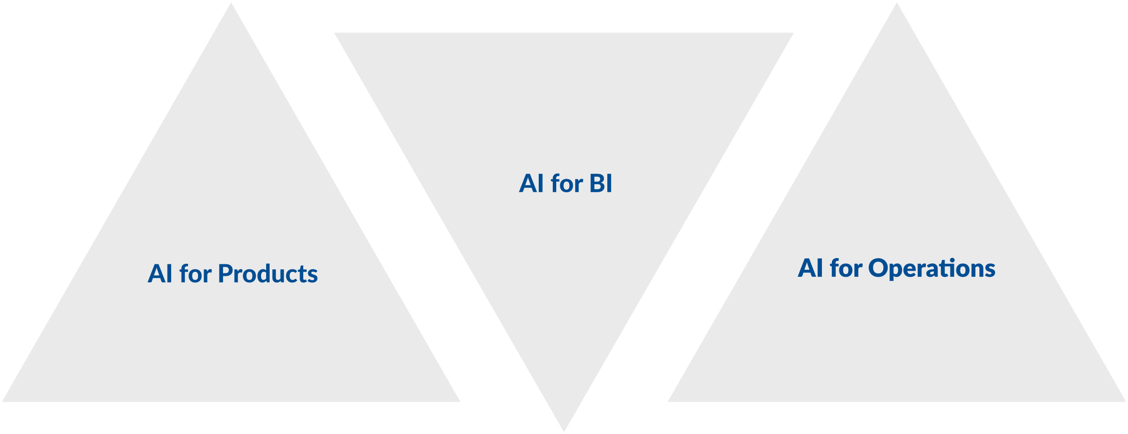


Figure 1: Frameworks for realize business value from AI

AI for Products are customer-facing capabilities where the AI data products directly empower greater customer experiences. Self-driving cars are an obvious example, but recommendation engines, facial recognition for identification and authorization, chatbots and even things like intrusion detection and spam filters are all examples of building AI into products.

The second framework is AI for Operations. This encompasses those back-office functions within an organization that can leverage the power of high-quality predictions to run operations more efficiently. Some examples include supply chain operations, security operations centers (SOCs), asset maintenance, fraud detection, customer account takeover, expense report generation, and payment automation. And, again, while the underlying mathematics are the same as for AI for Products, the risk, ROI and compliance aspects in this framework may be very different.

The third framework, AI for BI, uses advanced AI and ML technologies to enhance what many organizations have been doing in their business intelligence functions. Although there is some overlap between AI for BI and AI for Operations, generally these use-cases focus on leveraging more data and advanced algorithms to create faster, more accurate and more precise predictions and business decisions. To really build scale and efficiency in AI for BI, organizations need to be able to bridge the following gaps:

The Platform Gap

Many gaps between AI (and the data science teams building, optimizing and maintaining them over time) and the BI teams that will leverage it to develop sustainable business outcomes begin with the platform gap. Most people in this space are familiar with Google's 2015 paper "Hidden Technical Debt in Machine Learning Systems. This paper highlights the challenges of building long-term sustainable data products with ML. It specifically discusses how the ML models themselves are a relatively small amount of work compared to the complexity of things like collecting, processing, storing and managing data. Teams need to have very good, mature practices for data governance and semantic understanding, and many other tasks relating to the data itself, and somewhat orthogonal to the upstream use of that data in training and inference.

In Monica Rogati's The AI Hierarchy of Needs, she takes this a step further by saying that it's not just an unordered set of things that are needed to build effective AI data products. But rather there is a hierarchy. As a prerequisite for building scale, an organization needs to have very mature practices for data collection, ingest, exploration, and transformation. And the more time spent building maturity at the bottom of this pyramid of needs, the less time, money and complexity required at the top.

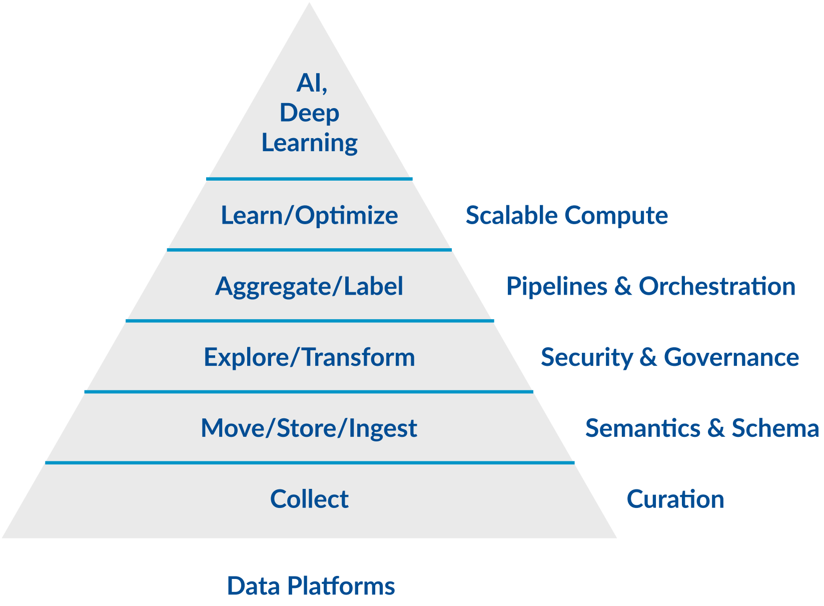


Figure 2: AI Hierarchy for Needs

To put a different spin on it, looking at it from a platform perspective, many data platforms have been built to support the things that Monica calls out at the bottom of the pyramid. Traditional data platforms are great for curation, semantics, security and governance, with powerful compute for building complex, scalable data pipelines. However, as you move up the pyramid, the ability of data platforms to generalize across use-cases limits the ability to meet specific needs in AI, machine learning, and data science. The relationship is shown below.

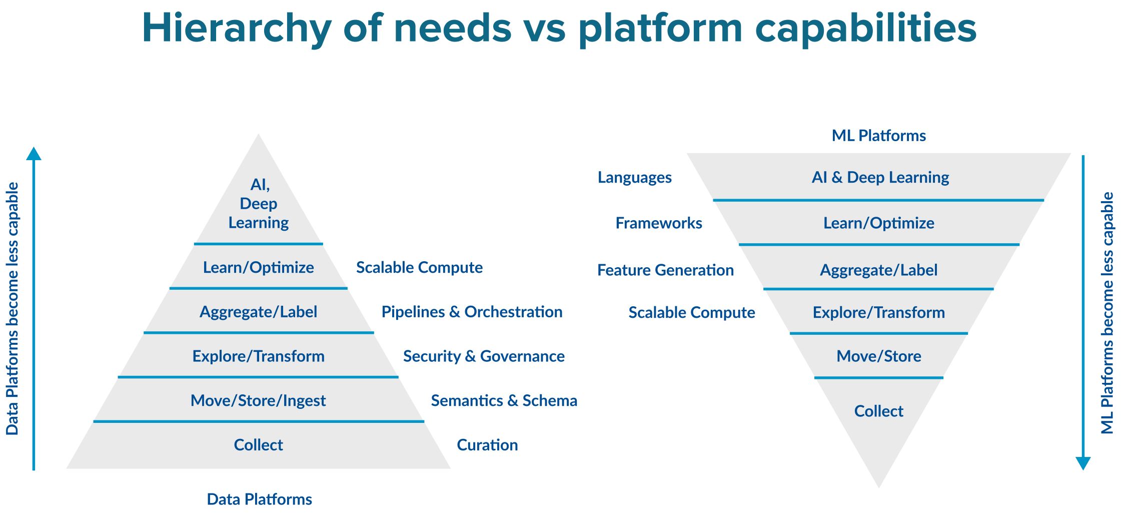


Figure 3: Integration of Data Platforms and ML Platforms

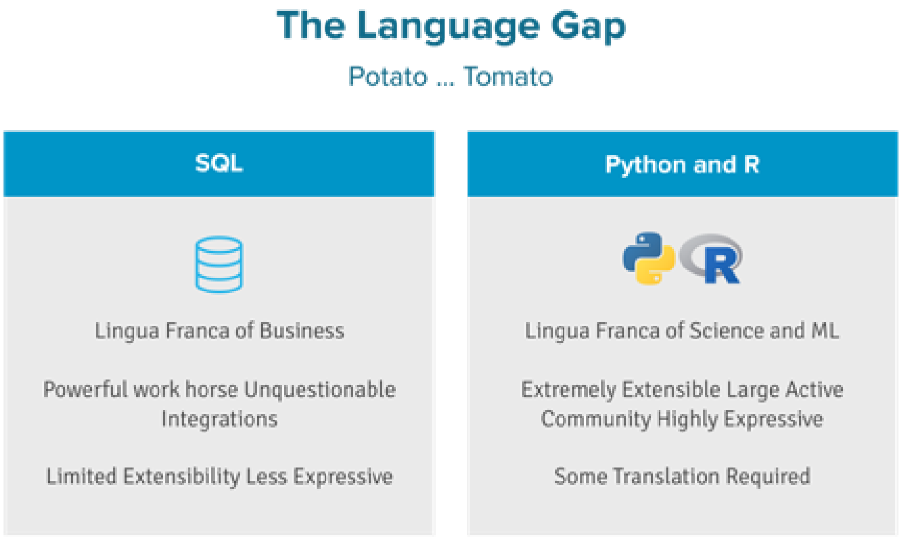
Data science teams are glad to have the data in one place, but without the languages and frameworks they need in the data platform, they are forced to build parallel platforms with specific capabilities for learning, optimization, experiment tracking, and collaboration. In doing so, they build an island, separate from the rest of the business, and try, over months and years, to recreate the governance, security, lineage, auditing, semantic and other capabilities they left behind in the core data platform. These platforms lose capabilities as they go down the pyramid, creating a gap at the interface layer between the ML platform and the core data platform.

More importantly than the overhead and duplication of effort, this platform gap results in additional risks and governance challenges, delays and complexity in getting data to the ML platforms and hinders ML teams bringing their AI capabilities back to the BI users.

To bridge this platform gap, it is important to architect the platforms with integration in mind. It will be some time before the industry standardizes and commoditizes both data and ML platforms. Until then, many challenges involved with this gap can be overcome by choosing a data platform with lots of options for integration and support for a large ecosystem of ML tooling choices. The data platform should have many options for push-down, to allow upstream tools to bring code to the data without moving and copying data.

The Language Gap

Another gap between AI and BI is in the languages they use. This is related to the platform gap insofar as the platform features either widen or narrow this gap. But focussing specifically on languages themselves, SQL is still the lingua franca of business. It is a ubiquitous and highly standardized language with a large addressable labor market. SQL is a powerful work horse for large-scale data preparation and analytics functions. However, it does suffer in terms of extensibility (beyond analytics use-cases) and expressiveness.



The Language Gap

Potato ... Tomato

SQL	Python and R
Lingua Franca of Business	Lingua Franca of Science and ML
Powerful work horse Unquestionable Integrations	Extremely Extensible Large Active Community Highly Expressive
Limited Extensibility Less Expressive	Some Translation Required

Figure 4: The Language Gap

Many data scientists, however, rely heavily on Python and R languages. Both languages are highly extensible, with large, active, open-source communities providing powerful AI frameworks. The nature of the language gap arises when both the producers of the data needed to build AI capabilities (business data and business users) as well as the consumers who will generate business value from those capabilities use SQL-based systems or record. Data scientists and ML teams, on the other hand, need systems where they can work with business data in Python and “deploy” models to users who do not use Python.

To bridge this gap, standardize on platforms that support both ANSI standard SQL and native Python on top of the same data. The platform should allow for broad extension with any Python library or framework. It should also support building consistent, reusable, auditable pipelines and automation with both Python and SQL.

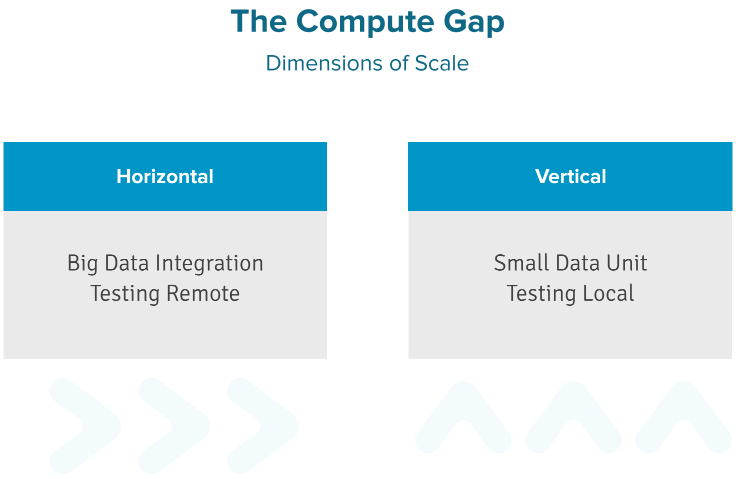

The Compute Gap
Dimensions of Scale
Horizontal
Big Data Integration Testing Remote
Vertical
Small Data Unit Testing Local

Figure 5: The Compute Gap

Strongly related to both platform and language gaps, there are often unrecognized gaps in the type and scale of compute resources needed. Many BI use cases build on data engineering functions across very large datasets with large timescales. The requirements lend themselves very nicely to horizontally scalable compute infrastructure and can often benefit from near-linear scalability in today’s cloud-native data platforms leveraging massively parallel processing (MPP).

In contrast, most ML frameworks are not designed for, and cannot leverage, horizontally scalable infrastructure for things like model training and inference on large data sets. With only a few exceptions, current ML frameworks require symmetric multiprocessing (SMP), usually with a single-node, vertically-scaled infrastructure including specialized compute like Graphics Processing Units (GPU). Data scientists dealing with large datasets are challenged with things like exploratory data analysis, experimentation and optimization to fit in memory on their SMP systems. To deal with large datasets without exploding infrastructure costs, data scientists will use subsets and aggregates of data that will fit into memory on small systems. Therefore, there is a constant back-and-forth between the bulk data from horizontally scaled data platforms and the subsets of data in vertically scaled ML platforms or “local” interactive computer resources.

This often results in developing data products in one platform optimized for cost and ease of use, training in another platform optimized for vertical scale, and deploying in yet another platform optimized for horizontal scale. Software developers have similar challenges in unit testing versus integration testing across platforms. The result is longer development cycles, duplication of cost, difficulty in debugging, and overall opacity of processes.

To bridge this gap, the platforms selected must have the ability to scale both vertically and horizontally as needed with consistent auditing, governance, security and transparency. The platform should support the current (but limited) ecosystem of distributed ML frameworks as well as non-distributed frameworks. And, most importantly, the user experience should be as simple and consistent as possible with little or no obvious friction points between user interfaces in either dimension of scalability for both training and inference with the data in one place.

Finally, how do all these frameworks work together? There are many ways to create integration points at different layers in the stack. Focussing on the semantic layer has consistently proven to drive not only business value and quick ROI but also bring people and communities together. Having a platform that provides consistency across users, skill sets, languages, and compute needs is a great start, but there is also a need for a common understanding of the data. This often materializes as the semantic consistency layer. Companies working at the semantic layer for many years know that it is an ideal layer of the stack to bring people together and bridge gaps. Data science, BI and analytics are all about mining “Semantic Capital,” as Oxford researcher Luciano Floridi discusses. Combining a powerful data platform with a large ecosystem and integrating across users, languages and data types at the semantic layer is an ideal way to tear down boundaries, bridge gaps and build economies of scale for all three frameworks: AI for Products, AI for Operations, and AI for BI.

References

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips , “Hidden Technical Debt in Machine Learning Systems”, NIPS, December 2015
Rogati, Monica, “The AI Hierarchy of Needs”, https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007 , June 2017

Chapter 8

The Semantic Layer: The Link Between Data and Insights

Kirk Borne, PhD

Chief Science Officer, DataPrime, Inc.

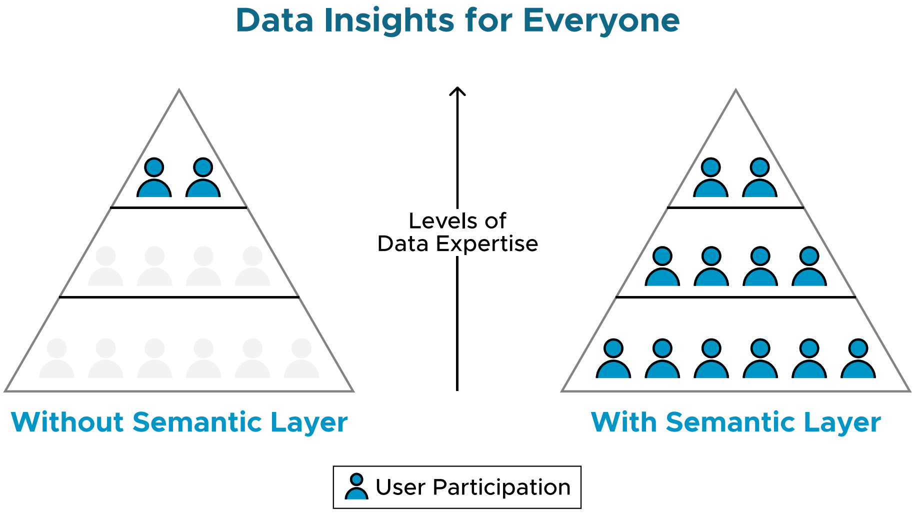
Dr. Kirk Borne is the Chief Science Officer at DataPrime. Before that, he was the Principal Data Scientist at Booz Allen Hamilton. Before that, he worked 18 years on various NASA contracts -- as a research scientist, as a manager on a large science data system contract, and as the Hubble Telescope Data Archive Project Scientist. Prior to working at Booz Allen Hamilton, Kirk was Professor of Astrophysics and Computational Science in the graduate and undergraduate Data Science degree programs at George Mason University for 12 years, where he taught courses, advised students, and did research. He has applied his expertise in science and large data systems as a consultant and advisor to numerous agencies and firms. He is also a blogger (rocketdatascience.org) and actively promotes data literacy for everyone by where he has been named consistently since 2013 among the top worldwide influencers in big data and data science. His PhD is in Astronomy from Caltech.

Chapter 8: The Semantic Layer: The Link Between Data and Insights

Kirk Borne, PhD
Chief Science Officer, DataPrime, Inc.

The semantic layer has been mentioned in almost every chapter so far. But what is a semantic layer? That's a good question, but let's first explain semantics. In the early days of web search engines, those engines were primarily keyword search engines. If you knew the right keywords to search and if the content providers also used the same keywords on their website, then you could type the words into your favorite search engine and find the content you needed. So, I asked my students what results they would expect from such a search engine if I typed these words into the search box: "How many cows are there in Texas?" My students were smart. They realized that the search results would probably not provide an answer to my question, but the results would simply list websites that included my words on the page or in the metadata tags: "Texas," "Cows," "How," etc. Then, I explained to my students that a semantic-enabled search engine (with a semantic meta-layer, including ontologies and similar semantic tools) could interpret my question's meaning and then map that meaning to websites that can answer the question.

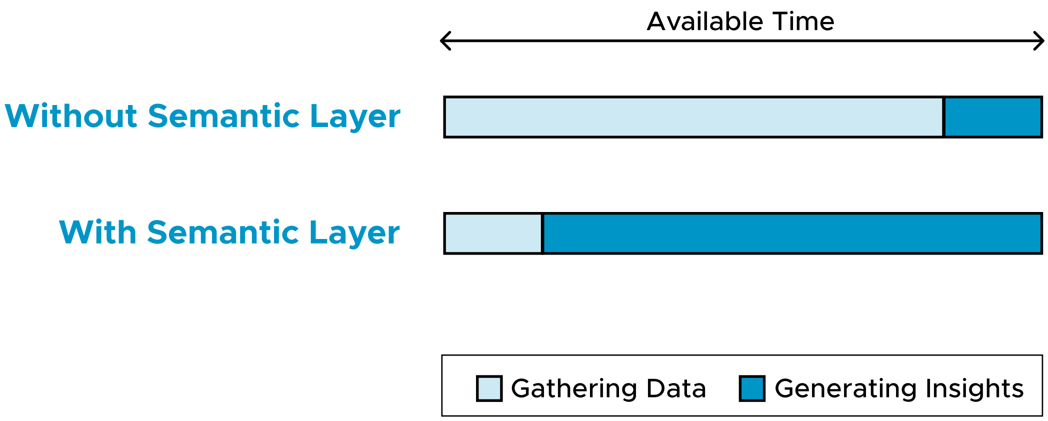
This was a good introduction to the wonderful world of semantics for my students. I brought them deeper into the world by pointing out how much more effective and efficient the data professionals' life would be if our data repositories had a similar semantic meta-layer. We could go far beyond searching for correctly spelled column headings in databases or specific keywords in data documentation to find the data we needed (assuming we even knew the correct labels, metatags, and keywords used by the dataset creators). We could search for data with common business terminology, regardless of the specific choice or spelling of the data descriptors in the dataset. Even more, we could easily discover and integrate, on-the-fly, data from different datasets that used different descriptors. For example, if I am searching for customer sales numbers, different datasets may label that "sales," or "revenue," or "customer_sales," or "Cust_sales," or any number of other such unique identifiers. What a nightmare that would be! But what a dream the semantic layer becomes!



When I was teaching those students so many years ago, the semantic layer itself was just a dream. Now it's a reality. We can now achieve the benefits, efficiencies, and data superhero powers we previously could only imagine. But wait! There's more.

Perhaps the greatest achievement of the semantic layer is to provide different data professionals with easy access to the data needed for their specific roles and tasks. The semantic layer represents data that helps different business end-users discover and access the right data efficiently, effectively, and effortlessly using common business terms. The data scientists need to find the right data as inputs for their models — they also need a place to write-back the outputs of their models to the data repository for other users to access. The BI (business intelligence) analysts need to find the right data for their visualization packages, business questions, and decision support tools — they also need the outputs from the data scientists' models, such as forecasts, alerts, classifications, and more. The semantic layer achieves this by mapping heterogeneously labeled data into familiar business terms, providing a unified, consolidated view of data.

Removing Bottleneck to Discovery and Access



The semantic layer delivers data insights discovery and usability across the whole enterprise, with each business user empowered to use the terminology and tools specific to their role. How data are stored, labeled, and meta-tagged in the data cloud is no longer a bottleneck to discovery and access. The decision-makers and data science modelers can fluidly share inputs and outputs with one another to inform their role-specific tasks and improve their effectiveness. The semantic layer takes the user-specific results out of being a “one-off” solution on a user’s laptop to becoming an enterprise analytics accelerant, enabling business answer discovery at the speed of business questions.

Everyone can discover insights. The semantic layer becomes the arbiter (multi-lingual data translator) for insights discovery between and among all business users of data within the tools that they’re already using. The data science team may be focused on feature importance metrics, feature engineering, predictive modeling, model explainability, and model monitoring. The BI team may be focused on KPIs, forecasts, trends, and decision-support insights. The data science team needs to know about and use the data that the BI team considers most important. The BI team needs to know which trends, patterns, segments, and anomalies are being found in those data by the data science team. Sharing and integrating such important data streams has never been such a dream.

The semantic layer bridges the gaps between the data cloud, the decision-makers, and the data science modelers. The key results from the data science modelers can be written back to the semantic layer and be sent directly to consumers of those results in the executive suite and on the BI team. Data scientists can focus on their tools; the BI users and executives can focus on their tools and the data engineers can focus on their tools. The enterprise data science, analytics, and BI functions have never been so enterprisey. (Is "enterprisey" a word? I don't know, but I'm sure you get my semantic meaning.)

That's empowering. That's data democratization. That's insights democratization. That's data fluency/literacy-building across the enterprise. That's enterprise-wide agile curiosity, question-asking, hypothesizing, testing/experimenting, and continuous learning. That's data insights for everyone.

Chapter 9

Using the Semantic Layer to Make Smarter Data-Driven Decisions

Greg Mabrito

Director of Data and Analytics, Slickdeals

Greg Mabrito, Director of Data and Analytics at Slickdeals where his team delivers analysis-ready data to support business-critical decisions. His expertise is in Data Integration. Electronic Transactional Messaging. Middleware Technologies. Integration Brokers. EAI, and ETL. Greg is a senior Data and Analytics leader and is regularly invited to speak and write on data and analytics in several IT and Data related forums. Before joining Slickdeals, Greg worked for HEB and USAA. Greg holds a BS in Computer science from Texas A&M University-Corpus Christi.

Chapter 9: Using the Semantic Layer to Make Smarter Data-Driven Decisions

Greg Mabrito,
Director of Data and Analytics, Slickdeals

The previous chapter from Dr. Kirk Borne looked at the fundamentals of the Semantic Layer. This chapter looks at how the Semantic Layer was implemented to derive insights and guide data-driven decisions in Slickdeals, the leading online deal community site dedicated to sharing, rating, and reviewing deals and coupons. Slickdeals consistently ranks in the top 100 most visited sites in the U.S.

Big data is a blessing and a curse. As the director of data and analytics at a large, user-driven deal-sharing company, my team has a steep hill to climb. We started in 2012 with an on-prem server to manage our data. Today, it's a very different situation. We have continually added new tools to our stack to better parse and leverage the data our teams need to make critical business decisions, and in 2020 we made major changes to fully modernize our data architecture. Here's what we've gleaned on our journey and the key changes we've made over the last year to empower users across the business with access to fresh and verified data.

Like many organizations, we deal with a massive amount of data every day. We're talking about one billion visits annually and 12 million unique monthly users across our services to give you a sense of scale. Slickdeals is the top external traffic referrer for Amazon, eBay, and Walmart. We ingest data from many sources across the internet, and we are continually updating the deal information we provide our users with. To them, up-to-date data is the only valuable data.

This is also true internally. Our goal has always been for employees across the company to easily access data and gain business insights without needing to learn to code or forcing them to use a specific tool. From Excel to Power BI to Tableau, we've always made it our mission to make data accessible to people when and how they need it.

As you might imagine, a constant theme for us has been handling ever-growing volumes of data. By 2019, we had effectively pushed the limits of SQL Server Analysis Services (SSAS). Trying to store huge volumes of data via SSAS while enabling users to create pivot tables without filters led to serious slowdowns and user experience problems.

In 2020, we realized it was time to move to the cloud. However, rather than approaching this as a pure “lift and shift” exercise, we recognized an opportunity to rethink our data infrastructure and strategy from the ground up. Our goal was to modernize our entire data strategy. We studied what we needed for our business and, from there, customized the stack to meet our needs. There were a few keys to success here, which I’ll cover below.

Event-Driven Architecture: Connecting Loosely Coupled Systems

One of the most important software architectures is event-driven architecture. This enables us to use any relevant or significant change in data to push updates to decoupled or loosely coupled systems. Kafka has been a large part of our strategy and stack here. This allows us to publish events out to the business and ingest them in Snowflake (our data warehouse), increasing speed, scalability, and resilience across our entire data stack.

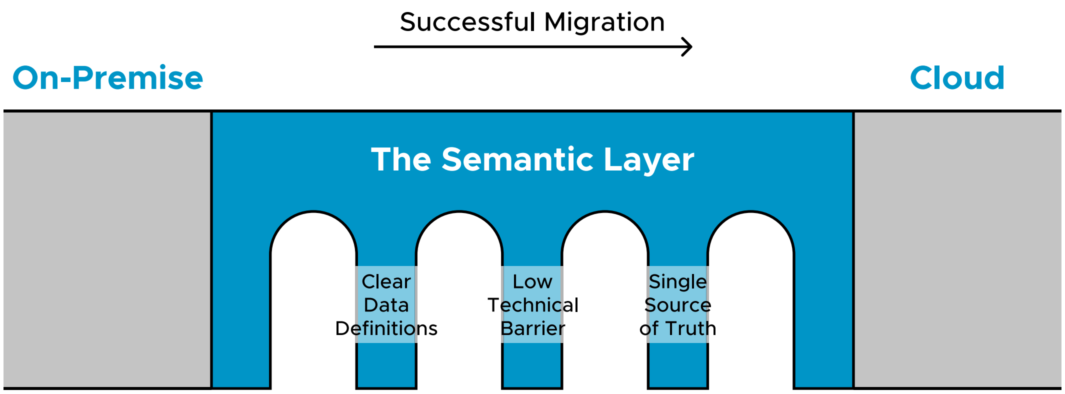
The Semantic Layer: A Single Source Of Truth

Data is complex, in part because every department in a business speaks a different “language.” A simple term like “employee” can have different meanings to different people. Does “employee” include contractors? Does it include part-time employees? How do you accurately calculate employee count with these nuances in play?

This is why a semantic layer is so important for empowering all business users to extract value from data. A semantic layer pre-defines specific types of data so everyone across the organization can rely on a single definition. A semantic layer also makes it, so those who need access to data do not need in-depth knowledge of how calculations work; they simply need to understand their business domains.

The concept of a semantic layer has been around for a long time (originally patented in 1991), and it has evolved along with the challenges of managing Big Data. Data is increasing in magnitude, speed, and diversity. This makes it very challenging to establish certainty. A semantic layer can provide a single source of truth, enabling business users to agree upon definitions for terms like “employee,” “customer,” or “net sales.” A semantic layer provides a single definition so that when data is queried and an answer returned, it can be trusted as the truth. Applying AtScale’s semantic layer to our big data allowed us to accomplish exactly that.

Migration Enabled by The Semantic Layer



Self-Service Analytics: No Coding Required

I want to go back to my earlier point about enabling all users to extract value from data to drive the point home on why this semantic layer matters so much. Having a semantic layer in place means IT and development are not standing in the way of access to data.

Users across all departments can ask questions and get answers from data without bottlenecks or slowdowns. This gives them independence from the tech team while still making them feel confident that the answers they derive are accurate, truthful, and meaningful—and that they will translate across lines of business and job functions.

Empowering The Business To Make Smarter Data-Driven Decisions

As data scientists and engineers, it's our job to make data more accessible—not less. Investing in flexible, extensible, and user-friendly big data architectures makes this possible. At Slickdeals, every department, from customer success to marketing to engineering, will soon be able to access the data they need with no need to understand data architectures or code.

We continue to evolve our approach and hope to build a single universal cube this year and retire Microsoft SQL Server Analysis Services (SSAS). Refactoring our pipelines for an event-driven architecture and implementing a semantic layer has enabled us to tap into fresher data faster and more flexibly to meet the needs of our entire business better.

Chapter 10

Empowering Decision Makers with BI and Data Science

Narendra Narukulla

Senior Data Science Engineer, The Wendy's Company

Narendra Narukulla is a Senior Data Science Engineer at The Wendy's Company. He is an expert in providing data-driven solutions with 10+ years of experience in Business Intelligence and Data Science roles, including building, optimizing, and maintaining data pipelines to enable analytics applications. He has broad experience in analytics in Supply Chain, Operations, Sales, Marketing, and Finance functions. Before joining Wendy’s, he worked in the CPG industry as a Data Scientist and consulted for Telecom companies in various BI roles.

Chapter 11: Empowering Decision Makers with BI and Data Science

Narendra Narukulla
Senior Data Science Engineer, The Wendy's Company

How can we enable decision-makers and key stakeholders by bridging business intelligence (BI) and data science teams together under one layer of data functionalities and capabilities? This chapter brings light to the importance of creating a data-centric culture for your company. This relates to teams across multiple analytical departments, whether in the supply chain, operations, sales, marketing, or finance. This chapter looks at the differences and overlaps between BI and DS regarding their prioritization to the data, how we can align them, how we can use BI as a guide, and finally, how we can bridge the needs of data science with BI data.

As data professionals, we try to uncover truths, facts, and share these insights with decision-makers or business stakeholders, but in different ways and forms. BI has been helping business stakeholders and decision-makers for a long time, and data science elevates BI to the next level. BI is about informing what is happening, whereas data science informs the relationship, associations, inferences, and predictions.

The role of BI teams is to focus on the historical performance of the business. However, businesses need insights on future performance, and that is where data science comes in. The reason for bringing up this distinction between data science and BI groups is simply to share their different views and prioritize data.

BI teams present their results in a cleaned-up fashion via dashboards; therefore, how many data science projects can be done just using the data from BI? About 40% of enterprise data science projects can be done just by using the data ready for BI. When a data science project is based on BI data and uses the same terms as BI, reports will be consistent and easily understandable to stakeholders. There are multiple stages involved before raw data will be useful for generating reports or insights as seen in the figure below.

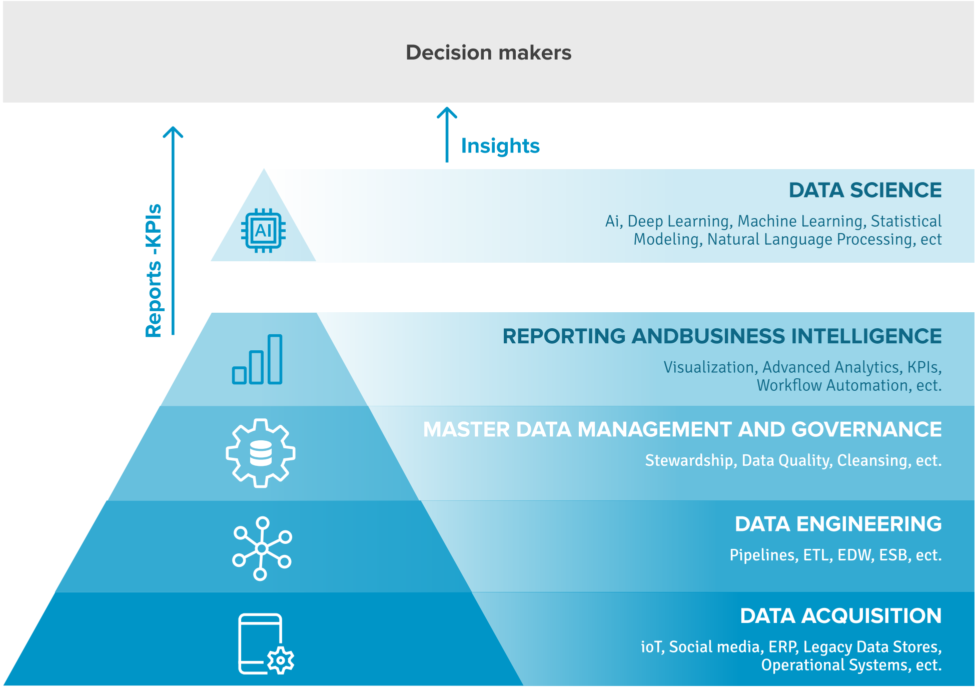

Decision makers
Insights
Reports -KPIs
AI
DATA SCIENCE
Ai, Deep Learning, Machine Learning, Statistical Modeling, Natural Language Processing, ect
REPORTING ANDBUSINESS INTELLIGENCE
Visualization, Advanced Analytics, KPIs, Workflow Automation, ect.
MASTER DATA MANAGEMENT AND GOVERNANCE
Stewardship, Data Quality, Cleansing, ect.
DATA ENGINEERING
Pipelines, ETL, EDW, ESB, ect.
DATA ACQUISITION
ioT, Social media, ERP, Legacy Data Stores, Operational Systems, ect.

Figure 1: Insights Continuum

Despite the differences between BI and data science, there are a lot of similarities. First, both teams often depend on the same set of data. While the BI team needs aggregated data and the data science team needs the most granular or atomic data, the aggregated data for BI is derived from the granular or atomic data. Second, not all teams rely on one canonical data source – data warehouse, data lake or Data Lakehouse. Thirdly, both teams must align on the metrics that they use -- the BI team offers insights on the lagging metrics, while the data science team offers insights on the leafing metrics. For example, for a forecasting project, some metrics-related questions would be: Is it annual company sales? What are the product sales? Do we do sales in dollars? Sales in quantity? What will be the sales in the next quarter? Which is the most profitable customer segment? As you can see, both BI and data science projects will be about aligning on KPIs – leading and lagging.

Managing BI and data science projects requires a solid understanding of the business problem, the KPIs (or calculations behind the KPIs) and the consumers of these KPIs or insights. Often, there's some misunderstanding about those calculations. For example, one of your KPIs may be calculating forecast errors in monthly sales. Is it calculated by average monthly sales? Every three months? Or within a year? Is it the average of errors from sales forecasted by brand or just the average of all brands together? One solution is to identify the data needed to be transformed into metrics. If the KPIs are not reliable, then change management will be much more difficult, and the insights won't be useful in helping decision-makers understand the cause-and-effect relationships within the business. Having access to reliable and easily understandable metrics will increase the chances of transforming insights into business results.

Overall, data science needs to leverage BI efforts as it saves time and avoids costly mistakes from using different ways to calculate the same metrics. But how do we bridge the gap and be accessible for data science? Sending data scientists to work in BI projects could give them more exposure to the business. Also, both teams must create guidelines for accessing business technology and definitions. One key enabler is to leverage semantic layer products such as AtScale to help both the BI and data science teams. According to Gartner, unprecedented levels of data scale and distribution are making it almost impossible for organizations to effectively exploit their data assets. Data and analytics leaders must adopt a semantic approach to their enterprise data assets or face losing the battle for competitive advantage.

However, the biggest downside of the semantic layer is you must build, maintain and manage it. The semantic layer must be kept in sync with any database changes. Therefore, focus on data governance to ensure all people across the community use the same data definitions in the enterprise when leveraging the semantic layer.

Chapter 11

Model Drift in Data Analytics: What Is It? So What? Now What?

DBP

Data For Business Performance

Prashanth Southekal

Managing Principal, DBP Institute and Professor at IE Business School

Dr. Prashanth H Southekal is the Professor and Managing Principal of DBP-Institute, a Data Analytics Consulting and Education company. He brings over 20 years of Information Management experience from over 75 companies such as SAP, Shell, Apple, P&G, and GE. In addition, he has trained over 3000 professionals world over in Data and Analytics, and Enterprise Performance Management (EPM). He is the author of 2 books - Data for Business Performance and Analytics Best Practices and contributes regularly to Forbes.com. He is an Adjunct Professor of Data Analytics at IE Business School (Spain) where he received the teaching excellence award for the 2020-2021 academic year. Dr. Southekal holds a Ph.D. from ESC Lille (FR) and an MBA from Kellogg School of Management (US).

Chapter 11: Model Drift in Data Analytics: What Is It? So What? Now What?

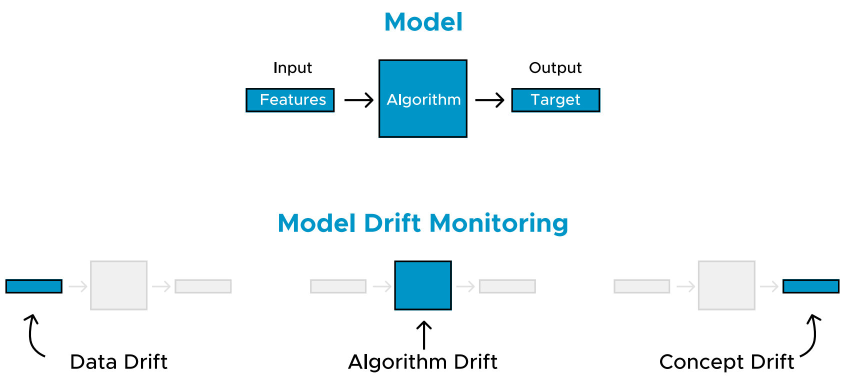
Prashanth Southekal,
Managing Principal, DBP Institute & Professor at IE Business School

Previous chapters were primarily focused on deriving reliable insights from data using the semantic layer as the input to data science models. So once the data science models are developed, how can businesses use those models to derive good insights? For some context, a report from MIT said that digitally mature firms are 26% more profitable than their peers. Also, according to Forrester, data-driven companies grow at an average of over 30% annually. Despite the potential of data to improve business performance, data analytics projects have a poor success rate. Only 20% of analytic solutions deliver business outcomes, according to Gartner. A report in VentureBeat said that 87% of data analytics projects never make it to production.

There are many reasons for this poor success rate. From the technical side, one problem is model drift in data analytics. What is model drift? Model drift is the degradation of data analytics model performance due to changes in data and relationships between data variables. Model drift occurs when the accuracy of insights, especially from predictive analytics, differs significantly from the insights derived during the model's training and deployment periods. Specifically, there are three main sources or symptoms of model drift.

1. **Data Drift**: When the characteristics of the independent, feature or predictor variables change.
2. **Concept Drift**: When the characteristics of the dependent, label or target variables change.
3. **Algorithms Drift**: When the algorithms, including assumptions, lose relevance due to changes in business needs.

Managing Model Drift



So, what can businesses do to address the problem of model drift? What are the root causes of these three main sources or symptoms of model drift? The primary reason for model drift is a change in business. Business strategies and objectives change due to mergers, acquisitions and divestitures (MAD), new product introduction, new laws and regulations, entry into new markets and more. Basically, a business is a constantly evolving entity. All these disruptions will change the way original data analytics models are used by the business. Knowing the sources of model drift will help you identify the right remediation measures you will need to get the model back to an acceptable or desired level of performance.

Before we go further: Why does model drift matter? What is the business impact of model drift? Today, data analytics models are increasingly becoming the major drivers of business decisions and performance. This trend will continue at a much faster pace, given the rate at which data is captured and the increasing maturity of machine learning (ML) platforms. In this reality, managing model drift is critical to ensuring the accuracy of insights or predictions. Reducing or eliminating model drift will enhance the trust you can place in models, promoting the adoption of data and analytics across your organization.

So, how can organizations reduce or eliminate model drift? At its core, model drift is not a technology management problem; it is a change management problem. Change management issues arise when dealing with people and the change in people. People have their own beliefs and values, and we all must understand change and believe in this. If people don’t see the benefit and vision, they will not change and barriers will be created. This change in data and analytics can be effectively managed by implementing the following three strategies.

First, data reflects reality, and often, the degradation of data results in the degradation of model and business performance. Thus, you need to manage data drift with effective data governance practices. We all know the fundamental principle of data processing is “garbage in, garbage out.” So, identify the variables in your hypothesis, define your data quality KPIs, set targets and thresholds and track these KPIs continually to stay up to date with changes in data quality.

Second, continuously assess your business dynamics and constantly review the relevance of the existing data analytics models with your stakeholders. While talking to your stakeholders, ask these questions:

1. Why do you want to have insights? How much do you want to know? What is the value of knowing and not knowing these insights?
2. Who owns the insights coming out of our models? Who is accountable when it comes to transforming insights into decisions and actions?
3. What are the relevant data attributes required for the model to derive accurate and timely insights?

Last, integrate ModelOps and DataOps practices to enable the quick and ethical replacement of the deployed analytics model with another if the business circumstances change. MLOps is about managing reproducible and automated algorithms, while DataOps is an automated and process-oriented methodology to supply data for data analytics. There is absolutely an interplay between MLOps and DataOps because insights = data + algorithms. So, to produce insights and recommendations in near-real time, you need timely ML algorithms coming from MLOps and timely data coming from DataOps. Data is the fuel on which the models run; models have practically no business utility without data. The sound integration of ModelOps and DataOps practices helps quickly progress analytics models from the lab to production.

Overall, the best way to manage model drift is by continuously governing and monitoring the model performance with the right KPIs. While deploying data analytics models is important, what really matters are the models that are consumable by the business for improved business performance. As they say, change is the only constant in life, and businesses change and evolve to stay relevant. Involving business stakeholders early on, reviewing any change with KPIs and continuously adjusting for improvements is critical in managing drift in data analytics models.

Chapter 12

Four Key Signals That Indicate a Data Culture Is Thriving in Your Organization

Gramener
Insights as Stories

Ganes Kesari

Chief Decision Scientist, Gramener

Ganes Kesari is an entrepreneur, AI thought leader, author, and TEDx speaker. He co-founded Gramener, where he heads Data Science Advisory and Innovation. He advises executives on data-driven decision-making. He helps apply data science to solve organizational challenges and in building teams to promote a culture of data. Ganes contributes articles to leading magazines such as Forbes and Entrepreneur. He teaches guest lectures on data science in schools such as Princeton University and runs corporate training on data-driven leadership.

Chapter 12: Four Key Signals That Indicate A Data Culture Is Thriving In Your Organization

Ganes Kesari,
Chief Decision Scientist, Gramener

The main objective of data and analytics is to offer insights that enable data-driven decision-making. However, data-driven decision-making is closely related to the data culture. But what does this term "data culture" really mean? It is often tossed around to explain failures in delivering value from Artificial Intelligence (AI), or to describe vague goals with analytics. Is data culture just another buzzword? Why should leaders pay attention, and what is its significance in an organization's journey to becoming data-driven? What are the practical challenges in fostering a data culture, and what are examples of businesses that have done it right?

Data-driven culture was the theme of a recent panel discussion of industry leaders at the MIT CDOIQ (Chief Data Officer and Information Quality) Symposium. I spoke to Robert Audet, Director in Advanced Solutions at Guidehouse, who moderated the panel of chief data officers (CDOs). This chapter answers the above questions using insights from our conversation.

So, what is data culture? An organization's culture is often said to be what employees do when no one is watching. Culture is the collection of shared values and practices that motivate all team members. Similarly, data culture is the collective behavior and beliefs of people in how they use (or don't use) data for decision-making. "To make sense of data culture, we need to understand how it fits into the overall corporate culture," said Katherine Tom, CDO at the Federal Reserve Board of Governors. Data is often meaningless on its own with decision-making. What matters is how it enables business goals and operations. "The essence of data culture is in figuring out the real purpose of data," she adds. Data has been collected, analyzed, and reported in organizations for decades, but conversations around data-driven culture are just a few years old. What's driving this trend?

"Today, 83% of executives look at becoming data-driven as a way to maintain their competitive edge," says Audet, referring to a recent survey by International Data Corporation (IDC). Over the last two decades, digitalization has led to an explosion of data. Information about critical stakeholders such as customers, employees, competitors, or the larger community is readily available today.

Given the opportunity to turn this information into a competitive advantage, the survey found that organizations are no longer content being just data-aware, but they aspire to become data-leading. If culture is the subconscious behavior of an organization, infusing data-driven decisions into this subconsciousness is easier said than done. Peter Drucker famously said that culture eats strategy for breakfast. Bringing about a culture change is tough and time-consuming.

What are the roadblocks to internalizing the use of data analytics for decision-making? “Data is often perceived as something purely technical,” said Samir Boualla, CDO of ING Bank France. The biggest challenge is to shift stakeholders’ perspectives of data and analytics from a technology tool into an integral business component.

Surprisingly, the key hurdles in making this shift aren’t technical but semantic. “The language we use to communicate data is often our biggest inhibitor,” explained Tom. She shared an example of how business teams in her organization struggled to understand the importance of data governance. When technical jargon and Google-search definitions didn’t help, she explained data governance simply as “the group of activities or business procedures to make the management of data easier and better.” Tom shared how this layperson-focused explanation addressed confusion and helped onboard the business teams.

If data culture can be a crucial enabler for business success, how do you cultivate it? Four key traits emerge for every organization that has successfully forged a culture of data-driven decision-making.

1. **Executive leadership owns and drives the use of data**

Executives at data-leading organizations don’t just sponsor data and analytics initiatives; they own them. “If you don’t have leadership buy-in, then you don’t really stand a chance,” said Elizabeth Puchek, CDO of U.S. Citizenship & Immigration Services (USCIS). “If a director looks at a dashboard or an analytics product frequently and uses that to ask questions of the operations, then the adoption of this product will easily get propagated down the chain.”

IDC’s survey found that executives in data-leading companies are eight times more likely to use data actively in their own work than data-aware companies. “Another indicator is the presence of data leadership roles, such as CDO as part of the governing board,” added Puchek. To bring about change, start at the top.

2. Data champions break silos between teams and promote collaboration

While leaders can start the fire, change agents carry the torches across the organization. Identify people in departments you consider data-driven and make them data champions, recommends IDC.

These champions evangelize the use of data among their groups and demonstrate best practices. They often turn into interpreters of data and help build bridges between teams. Pick champions across levels, skills, and backgrounds for the greatest impact. Endorse their evangelization efforts and create a conducive environment to sustain the change.

3. Data is trusted, easily accessible, and freely shared

For people to use data for decisions, they must first trust it. And data and insights must be readily accessible to every employee within the organization. Contrary to conventional opinion, data must be shared both internally and externally. Gartner found that sharing data externally generates three times more measurable economic benefits for organizations.

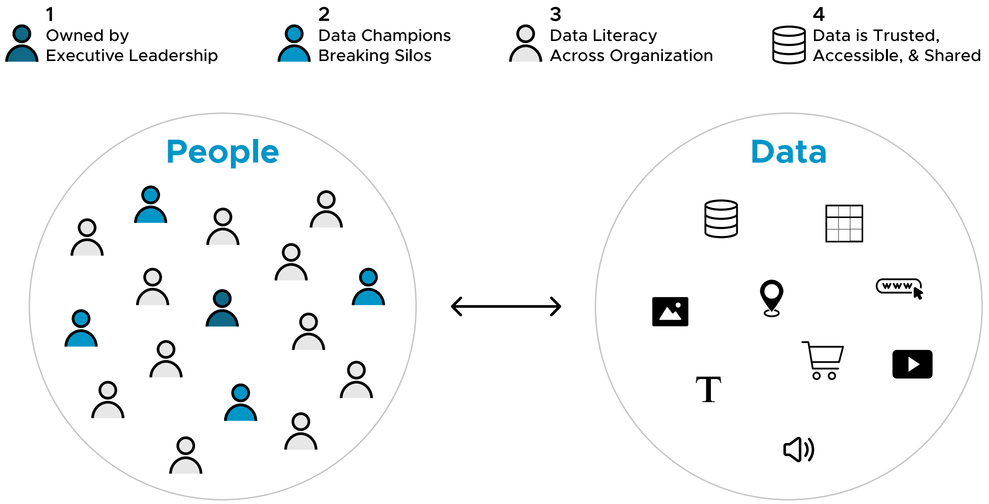
“When the CDO office was first set up at USCIS, we embarked on listening tours to figure out how to connect data to the mission,” said Puchek. Her team found that one of the service centers lacked access to another agency’s data, affecting their ability to do daily checks. This issue was swiftly resolved by talking to the agency’s point of contact. As a result, the cycle time dropped from six hours to just 12 minutes. Such small wins demonstrated the power of data democratization, earned credibility, and built momentum toward data-driven decisions.

4. Data literacy is considered a critical skill for every role

Data-driven organizations see understanding and communicating with data as critical skills for every employee. Data literacy is not relegated to just data and analytics teams. With a common language for data, people across business and technology teams can freely exchange ideas so it is enabling rather than inhibiting.

Leaders encourage data exploration and promote a healthy curiosity around data. “You know an organization is investing in data literacy when it’s a part of the hiring process, shows up in ongoing data skills assessments, and is discussed in performance reviews,” says Audet. IDC found that data-leading companies are three times more likely to require new hires to know how to present data persuasively when arguing a point than data-aware organizations.

Four Key Signals of a Thriving Data Culture



So, what are the benefits of building a data-driven company? Given the time, commitment, and resources needed to foster a data-driven culture, is it worth all the investment? IDC found a clear 46.2% improvement in financial, customer, and employee metrics at data-leading organizations. Thus, the business benefits of data culture are tangible. Often, organizations get intimidated by the magnitude of change needed and the subtle behavioral aspects that must shift. The key thing is to start small, secure, easy wins, and continue building momentum. Begin by figuring out how data supports the mission. Onboard executive support, identify data champions, democratize data, and make data literacy much more than a training item.

Chapter 13

Building a Competent Data and Analytics Team

John K. Thompson

Best Selling Author, Innovator and Thought Leader /
Founder of Marketing Sciences

John K. Thompson is an international technology executive with over 35 years of experience in the business intelligence and advanced analytics. He currently leads the AI and Advanced Analytics team at a leading biopharmaceutical company. Previously, John was an Executive Partner at Gartner, and was responsible for the advanced analytics business unit of the Dell Software Group. John is also the author of the book – Analytics Teams: Leveraging analytics and artificial intelligence for business improvement and the co-author of the bestselling book – Analytics: How to win with Intelligence. Thompson holds a Bachelor of Science degree in Computer Science from Ferris State University and an MBA in Marketing from DePaul University.

Chapter 13: Building a Competent Data and Analytics Team

John Thompson
Best Selling Author, Innovator and Thought Leader / Founder of Marketing Sciences

Whether the company is a big corporation with thousands of employees or a small company with just twenty employees, high-performing teams inevitably offer superior business results.

“No matter how brilliant your mind or strategy is, if you are playing a solo game, you will always lose out to a team.” That’s how Reid Hoffman, LinkedIn co-founder, sums it up. Successful analytics initiatives are no exception and depend on high-performing teams to provide good business insights to the insight consumers. As data analytics in business enterprises today is the new language of business communication, the data analytics teams should ensure that the data and insights are in the hands of competent leaders and front-line knowledge workers who will use data and insights to drive better business results.

This chapter is intended for senior managers and executives contemplating or have committed to hiring and managing a team. In this chapter, we will walk through succeeding in the analytics journey from beginning to end. We will assume that you are just considering building a team. You are selecting projects, succeeding and failing at those projects, and learning from those failures. You are moving from a development and test mode environment to a production mode, evolving from manual or traditional digital processes to data-driven ones. These analytically enriched processes improve with time and undertaking the resulting organizational change management initiatives required to actualize the value you planned for when you started this journey. You are managing the executive expectations of the scale of investment, the scope of change, the speed at which the return on investment will be realized, and the realities associated with becoming an analytically driven organization. This long set of steps makes up an organization’s “macro” process becoming analytically driven.

One of the mindset changes -- and the organizational process changes -- that’s required to succeed in this journey is that by becoming a data and analytically driven organization, you at some point realize that the organizational change you seek is never “done.” This process is evergreen and ever-changing.

Along with the “macro” process of organizational and mindset change, there is a “micro” process of evolving in response to the needs, wants, and desires of customers, patients, the market, the environment, suppliers, investors, stakeholders, and competitors. From the perspective of the middle of the processes as described, these processes at the execution level are usually described, organized, and discussed as projects. The larger overall process is typically made up of projects that focus on a specific objective or goal, but the overall process is dynamic and ever-changing. If you, your leadership, and your organization want to be part of this continual evolution, then your organization, data operations, and analytical models need to be set up and organized so it accounts for and reacts to the constant collection of data inputs. You also need to update and test the models being developed and deployed, monitor execution and performance and refresh the analytical contents and models at the appropriate time and cycle.

Advanced analytics models are trained on data. The data represents the world or the subject area when the data was collected. Once the model is trained and that model is accurately "predicting" the characteristics of the subject area as represented by the training data, the model is "locked." By locking the model, we end the training phase and move the model into production.

The model ingests and examines data in the operational world and predicts the information we are interested in. But we all know that the world changes and so does the data that is the by-product of those activities. The models must be updated or retrained using current data to ensure that the models generate predictions based on data that is as close to the current state of the world as possible. We "unlock" the model and train it again using new data. The model now predicts based on the new frame of information. The model continues to track the evolution of changes in the operational world through these cycles of training and production through its operational life cycle.

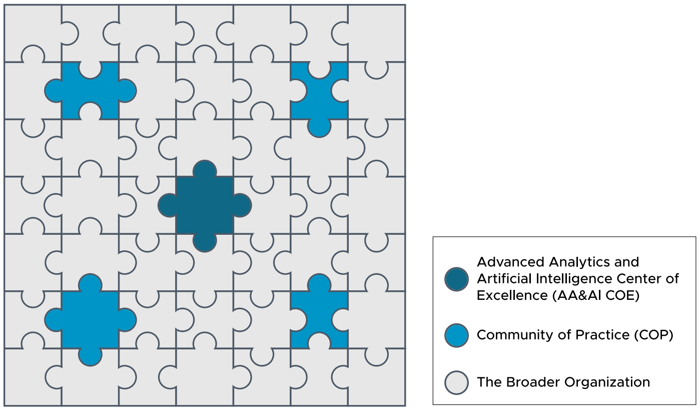
People are like analytical models. But sadly, the great majority of people lock their mental models and rarely, selectively, or simply do not open those mental models to update their views to align with the new reality of the world. This is one of the primary reasons people are called "out of touch" or clueless.

Let's, as a group, keep in mind that the world and all the phenomena in it evolve and change. Just because we were exposed to a set of norms, technologies, cultural constructs, and other conditioning when we were growing up, that does not mean we cannot retrain our models to include new norms and activities. We do not need to throw away the old, but we can include the new and have a richer and more inclusive world view. Let's unlock and retrain our mental models frequently. The essence of being analytical and data-driven is to be dynamic and guided by empirical evidence and measurable factors.

Building an analytics team and an environment depends on collaboration. You will need to hire a leader. You will need to empower and fund that leader to hire and lead a team. I refer to this team as the Advanced Analytics and Artificial Intelligence Center of Excellence (AA&AI COE). The COE leader and team will need to learn the broader organization and to build a network of collaborators, sponsors, and allies. This new network, which we will call the Community of Practice, or COP, will span the entire organization and global operations. I traveled to every continent and spent much of my time on the road and in discussions with executives, managers, and people who should be involved. The primary objectives of those meetings were to:

- Communicate and convince executives, senior managers, managers, and individual contributors they should collaborate with the AA&AI COE team and the other staff members in the COP.
- Communicate that the AA&AI COE team helped them understand and develop analytics to drive their operations forward in the manner they dictate.
- Empower their staff members to join the COP and to join projects with the AA&AI COE staff to develop analytical applications, models, and new ways of reaching higher levels of effectiveness and efficiency.
- Let them know that we were not there to judge their ideas and current state of operating but to help them see how data and analytics will help them reach and exceed their operating goals.
- Improve employee engagement and remove the tedious parts of staff members' duties to focus on the more creative aspects of their work that leverages their experience and expertise.

Building an Analytics Team


● Advanced Analytics and Artificial Intelligence Center of Excellence (AA&AI COE)
● Community of Practice (COP)
○ The Broader Organization

Let us now define the taxonomy and naming of the collaborators we will discuss and describe in our work and on our analytics journey. I will be calling executives sponsors because they typically set the direction and control funding and staffing for their organization.

Also, I will be calling managers stakeholders, as they typically own the headcount needed to collaborate with the AA&AI COE staff. And finally, I will be referring to the staff members of the operational departments as subject matter experts. The AA&AI COE leadership, data scientists, data engineers, data visualization experts, and others cannot be successful without the full-throated support of sponsors, stakeholders, and subject matter experts. A transparent and trust-based relationship between all parties is crucial to our joint success.

To close the discussion of organizational dynamics and politics, if the organization numbers in the thousands, spans the globe, and is committed to improving through using data and analytics, your team of a handful of people cannot deliver on every possible area of improvement. You should support and actively promote two areas of augmentation of your capabilities.

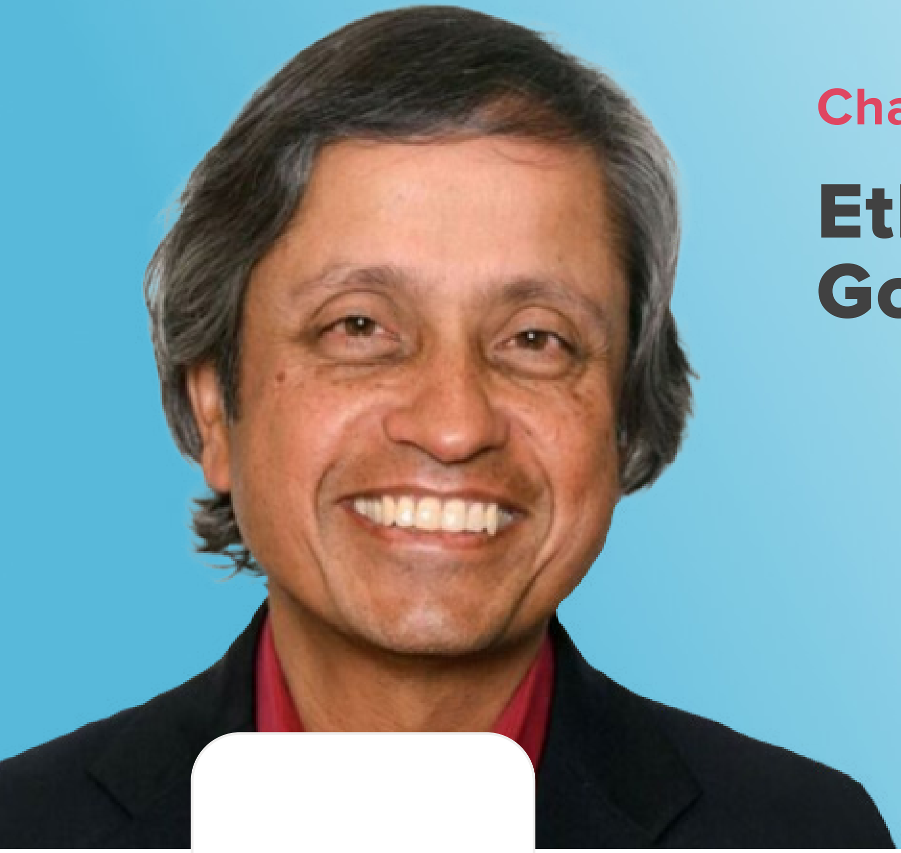
How can a data analytics team support the company's competitive advantage or simply help the company stay competitive? First, consider outsourcing certain projects to competent, capable, and proven third parties. The projects to be considered are those that others in your industry have completed and are now considered table stakes to be competitive at the new level of efficiency that the industry operates at or that the market, customers, suppliers, and patients' demand. Projects such as inventory management, supply chain efficiency, or designing servicing maps for optimal territory coverage by a sales team are projects successfully executed across numerous industries with well-known and published success stories.

Second, support the operational areas that want to invest in business intelligence, descriptive statistics, and small-scale predictive analytics. Your team cannot do it all, but your team can help these functional areas hire the right staff members, undertake initial projects, and consult on predictive and prescriptive applications. Helping the organization build a broader and deeper capability is part of your responsibility and it will build a network of supportive team members who aspire to grow in their skills and abilities and may be good candidates to join the AA&AI COE. It can only help your cause and ease your journey to build an ecosystem of talent that the organization pays for and nurtures to become your team's future talent.

After you've developed a common understanding of the area to be analyzed and improved through data and analytics, the functional team and the AA&AI COE team can gather data, build pilot environments, and discern whether the hypothesis developed to this point in the process is possible and probable. There is a non-trivial chance at this point that the hypothesis will be proven to be incorrect and/or the data required will not be available in the quantities and historical depth needed to build reliable models. There are several reasons the pilot could prove this path will not drive measurable or significant business value or not be technically feasible.

Once the prototype has been built and tested, then it is time to talk about how to implement the next iteration of the data flows, analytical models, governance systems, process changes, and end-user interaction systems to realize the benefit of all the work completed to date. It is very difficult to state this next fact and even harder for most readers to believe it, but this is where most efforts fail.

I hope that after reading this chapter, your interest has been sparked and that you're intrigued about how to execute and drive the analytics process forward in your profession and in your organization. The process will not be quick, it will not be easy, and in some parts of the process, you will not enjoy it. But it won't be dull or boring, and you will always be learning. If you are a lifelong learner -- and the chances are good that if you have read this far, you are -- a career in analytics is one of the most fulfilling careers you can choose.



Chapter 14

Ethical AI Governance

Anik Bose

Partner, BGV Ventures

Anik Bose has 15 years of active venture capital and corporate development experience, emphasizing transaction structuring and strategic planning, including 7 years as SVP, Corporate Development at 3Com Corporation, and 10 years as General Partner at BGV. In his role at BGV, Anik is very active in every portfolio company where he spearheaded the investment, and serves as a board member of Cyberinc, WebScale and Blue Cedar Networks. Earlier in his career, Anik served as a Partner in the High Technology Strategy Practice of Deloitte in San Francisco. In his management consulting roles, Mr. Bose worked with high-tech companies such as LSI Logic, Advanced Micro Devices and Hewlett Packard.

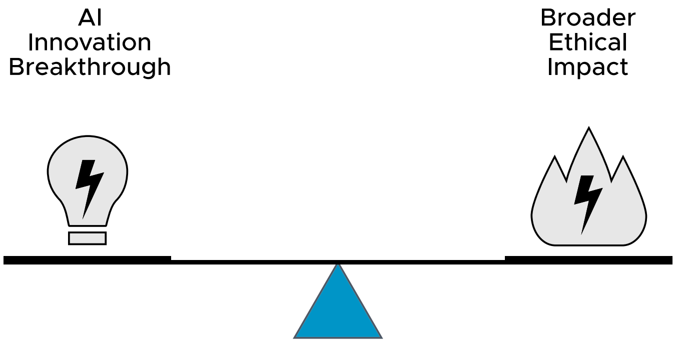
Chapter 14: Ethical AI Governance

Anik Bose
Partner, BGV Ventures

Artificial intelligence (AI) is consistently cited as one of the top macro-trends poised to dominate the next decade of innovation. Chatbots, conversational systems, banking apps, smartphones, self-driving cars and even home devices like Alexa have rapidly become part of our daily lives. As intelligent automation technology becomes increasingly pervasive across all industries, McKinsey Research predicts it could unlock $3.5 – $5.8 trillion in annual value, roughly equivalent to 40% of all data analytics techniques.

The promise of AI, however, presents a double-edged sword. The mere mention of AI and automation can conjure images of The Terminator and a dystopian future dominated by menacing cyborgs overtaking our planet. Consumer attitudes toward developing high-level machine intelligence are marked by fear, doubt and skepticism. The public is mostly split on whether this is a positive or negative development for humanity or whether these technologies should even be developed. Confidence falls much further for trusting major tech corporations like Facebook, Amazon, Microsoft, Google, or Apple to develop and stewardship these innovations. While Skynet remains the stuff of sci-fi novelty, the contemporary mainstreaming of AI across a variety of industries, from healthcare and financial services to autonomous vehicles and manufacturing, is now driving unique, legitimate concerns around ethical AI governance.

Ethical AI Governance



Whether it's Google search, Netflix recommendations, auto email correction or virtual assistants, enterprise AI is already entrenched, invisible and pervasive in our daily lives. A back-of-the-napkin calculation could conceivably claim that today's modern enterprises are running 25+ AI models at once. That number will quickly scale to 50, then 75, and onward and upward along a hyperbolic trajectory. As more enterprises leverage AI-based systems, the integrity of those models and the impact of their decisions and recommendations become increasingly consequential in both breadth (cutting across an incredibly wide array of industries) and in depth (building automated data-driven decisions at an exponential rate without human intervention).

In this brave new world, it's critical to ensure that AI-powered engines provide the same level of accuracy and fairness regarding protected populations and customer sub-groups (i.e., young/old, rural/urban, customers from different geographies, etc.). It's equally important to avoid using biased variables in models (like racial or ethnic origin, political opinions, etc.) and to ensure ongoing monitoring for verification, explainability and stability.

This is particularly crucial in a regulated industry like insurance, where legal compliance is required. A good example of this is the Insurance Distribution Directive (IDD) in Europe. Introducing AI into loan origination and financing paints a clearer picture of what's at stake. While the rise of algorithmic lending is helping companies raise loan production without raising delinquency or default rates, a bias in the AI governance model risks hurting, rather than helping, low-income and minority borrowers.

"African Americans may find themselves the subject of higher-interest credit cards simply because a computer has inferred their race," claims Nicol Turner Lee, a fellow with the Brookings Institution. Machine learning underwriting models have so many data points to infer that race, socio-economic background, and a host of other variables could influence a loan decision without the lenders even knowing. If the system's operators cannot explain the set of signals used to derive the model's outcome, the algorithm cannot be "fixed." Bias in the outcomes may not even be the algorithm's fault, but of those developing the algorithm or those inputting the data into the model, to begin with. AI cannot be truly trustworthy if the model remains a black box.

Digging deeper, an AI engine's output is only as valuable as the data that feeds it. Therefore, the integrity of datasets presents a vulnerability to the system. Datasets may contribute to bias due to unbalanced, unrepresentative or non-representative data in an AI model's training set. If the data is more representative of some classes of people than others, the model's predictions may be systematically worse for the underrepresented classes. And any learnings from the prediction data in deployed models can also amplify the bias.

To aggravate the predicament, a learning algorithm may drift over time. It is possible that after a model has been deployed, the distribution of the data that the deployed model sees differs from the data distribution in the training dataset. This change could introduce bias in a model. In these cases, the AI engine operators need to continuously monitor the bias metrics of a deployed model and adjust for this bias drift. There are a number of techniques to mitigate bias and improve fairness (the details of which are outside the scope of this paper). However, it's clear that there's an imperative for a consistent methodology backed by an AI governance framework.

Against this backdrop, it is incumbent upon AI practitioners, data analysts and executives to focus on ethical AI governance and lay out a shared set of principles, frameworks and standards to apply to AI technology development, deployment and maintenance to mitigate and address these concerns. The path forward requires a multi-pronged approach from a broad set of actors that includes the following:

A. **ENTERPRISE EXEMPLARS** – Large enterprise technology companies have appointed chief privacy officers, chief ethics officers, and similar roles to ensure data privacy and open-sourcing AI fairness tools. IBM is an exemplar, having formed an entire Ethical AI advisory board alongside VP's and directors across the organization. Other tech players like Microsoft, Google and SFDC are also setting powerful examples by implementing AI Responsibility or AI Ethics Officer roles to oversee AI adoption.

B. **STARTUP INNOVATION** – Startup innovation has emerged to provide foundational tools to enterprises implementing AI innovation to avoid unintended consequences. Sustainable AI and Credo have developed ethical AI governance tools, while TruEra has built AI monitoring and development systems. "AI First" startups with an ambition to become Fortune 1000 companies will also need to implement AI governance and data privacy standards into their own products and processes to ensure market adoption and to mitigate risks that could become obstacles to going public or being acquired.

C. **REGULATION** – As AI tools are introduced into highly regulated fields such as financial services and health care that already require compliance with existing laws, there will be a degree of transparency and explainability on these high-risk use cases. While it's still early days for governmental regulation of AI, the EU commission has laid out seven areas for the development of ethical AI, including

1) Human agency and oversight, 2) Safety, 3) Privacy, 4) Data governance, 5) Transparency and traceability, 6) Non-discrimination and fairness, and 7) societal and environmental responsibility. It is only a matter of time before regulators in Washington, Brussels, Beijing and other major capitals follow.

D. **Grassroots Communities** – The largest corporations deploying AI solutions (FAMGA) suffer a credibility gap in the public's eyes, and many have failed to establish public trust or to champion industry best practices. While regulation will play a role, AI practitioners from the industry must step forward to share lessons learned, provide practical insights, and frame the broader discussion. The formation of a diverse grassroots community of AI practitioners (like EAIGG) to democratize the development of ethical AI governance represents a positive step in the right direction.

The imperative for ethical AI governance is clear. However, it will take a concerted effort by the entire ecosystem to fully realize AI's promise and mitigate the risks.

In this backdrop, Ethical Intelligence and BGV recently collaborated on a research project to develop a cutting-edge maturity framework (AI Ethics Maturity Continuum) designed to guide enterprises especially the startups on their journey towards building billion $ responsible AI companies. This was developed based on widely accepted industry best practices, in-depth enterprise and academic research, and AI policy guidelines. The development was supported by input from the recently formed Ethical AI Governance Group (EAIGG), a grass roots community comprised of startups, big tech exemplars and academics coming together to democratize the development of ethical AI governance.

The Maturity Framework provides a granular approach to ensure AI first startups can implement best practices around the dimensions of accountability, intentional design, fairness, social impact, trust and transparency into their own products and processes, crucial to ensure market adoption and scaling of their companies. Successfully operationalizing these best practices will yield increased brand loyalty, customer engagement, market adoption, enhance regulatory readiness, attract talent and eventually eliminate any obstacles to going public or being acquired. The framework is based on 5 core principles namely: Accountability, Intentional Design, Fairness, Social Impact, and Trust and Transparency.

The AI Ethics Maturity Continuum has been designed to quickly assess a company’s level of ethics maturity and identify areas for improvement. It prioritizes agility and action, enabling users to build concrete strategies for sustainable AI systems and track development overtime. Most importantly, it empowers startups to embed ethics from the very beginning, resulting in stronger products, happier customers and more favorable exits. To learn more please visit www.eaigg.org.

Chapter 15

Speed of Thought for Insights

Dave Mariani

Founder and CTO, AtScale

Dave Mariani is the co-founder and CTO of AtScale. Before AtScale, he was VP of Engineering at Klout and at Yahoo! where he built the world's largest multi-dimensional cube for BI on Hadoop. He is a hands-on technology executive with over 25 years of experience in delivering Big Data, consumer Internet, Internet advertising and hosted services platforms, creating nearly $800M in company exits. David founded AtScale in 2013 to provide access to Big Data to everyone who needs it, regardless of data format, size or the type of tool business users want to use. Enterprises like JPMC, Visa, Bloomberg, United Healthcare, Cigna, Kohls, Home Depot, Wayfair and Toyota all use AtScale to democratize access to data for their employees and partners.

Chapter 15: Speed of Thought for Insights

Dave Mariani
Founder and CTO, AtScale

Moving beyond simple historical analysis to generating accurate predictions is an important step in the data maturity of an organization. In the past, business analysts focused on historical analysis while data science teams attempted to surface interesting insights about the future. Today, with the advent of the semantic layer, these two siloed worlds are coming together. Enterprises that merge these two disciplines can deliver augmented analytics, helping everyone in the organization better understand the past and predict the future.

Organizations leverage analytics to help them understand and improve their business operations and customer satisfaction. Before we go further, let's define the four types of analysis we typically see in an organization, each with increasing levels of sophistication.

Analysis Type	Description	Typical User Persona
Descriptive	Answers the question, "what happened"? This type of analysis focuses on a historical view of events helping users understand what happened in the past.	BI Users
Diagnostic	Answers the question, "why did it happen"? This type of analysis explains why historical events happened the way they did.	BI Users
Predictive	Answers the question "what will happen"? This type of analysis helps to predict the future based on building data science models.	Data Scientists
Prescriptive	Answers the question, "what will we do about it"? This type of analysis operationalizes predictive analysis to act on predictions	Data Scientists

Table 1: Types of Analytics

As you can see in the table above, business users typically focus on historical analysis while data scientists work to predict the future. Business users make better decisions if they can anticipate the future. It's also obvious that data scientists build better models to compare their predictions to what happened. Historical analysis and predictive analysis relate to both teams but rarely do the two meet. How do we make this happen? One key enabler is the semantic layer.

What is a semantic layer? While a lot has been discussed in this book on the value of the semantic layer, a semantic layer is a business representation of data that makes it easier for end-users to access data using common, business-friendly terms. A semantic layer maps complex data relationships into easy-to-understand business terms to deliver a unified, consolidated view of data across the organization. A semantic layer provides the following four key benefits.

1. **Usability**
   One of the biggest complaints from the business is that it takes way too long for IT to build or deliver reports for them. Users want to control their own destiny and subject matter experts (not IT) are best suited to applying data to improve the business. A well-designed semantic layer hides the complexity of data's physical form and location while translating data into understandable business constructs. A semantic layer frees business users and data scientists from dependency on IT and data experts by making data easy to use.

2. **Security and Governance**
   Currently, enterprises have strong and sometimes regulatory requirements to track "who" saw "which" data and "when". A modern semantic layer allows users to appear as themselves to the underlying data platforms from any consumer tool. A semantic layer ensures that data is consistent regardless of consumption style and ensures everyone plays by the same (governance) rules.

3. **Agility**
   Analytics agility, also thought of as "time to insight" is how long after data lands that it can be used to make decisions. BI tools that require data imports, extracts or cube building take from minutes for small data to days/weeks for large data before data can be accessed. A modern semantic layer leverages data virtualization to enable new data landing in your data warehouse to be immediately queryable by your BI tool, regardless of size.

4. **Performance and Scale**
   Cubes and data extracts were introduced to overcome the performance issues of analytics and data platforms. This approach introduces data copies, adds complexity, destroys agility and introduces

latency. A modern semantic layer improves performance regardless of the underlying data model, whether it's a snowflake, a star, or purely OLTP schema. By automatically creating and managing aggregates or materialized views inside the underlying data platform, a semantic layer learns from user query patterns and optimizes the data platform's performance and cost without data movement.

Overall, the Semantic Layer is the Unifying Thread. With a semantic layer, you can bridge the gap between BI users and data science teams. This enables your teams to work transparently and cooperatively with the same information and goals.

A semantic layer abstracts away the complexity of underlying raw data using a business model, allowing any data consumer to access quantitative metrics, attributes, features, predictions, business hierarchies, and complex calculations in an intuitive, easy-to-understand interface. A semantic layer solution presents this consumer-friendly interface in the "language" of their tooling (SQL, MDX, DAX, JDBC, ODBC, REST or Python), translating queries into the dialect of the underlying cloud platform. With a common set of business terms, both teams can interact with the same data, with the same governance rules, with the same results, using the tooling of their choice.

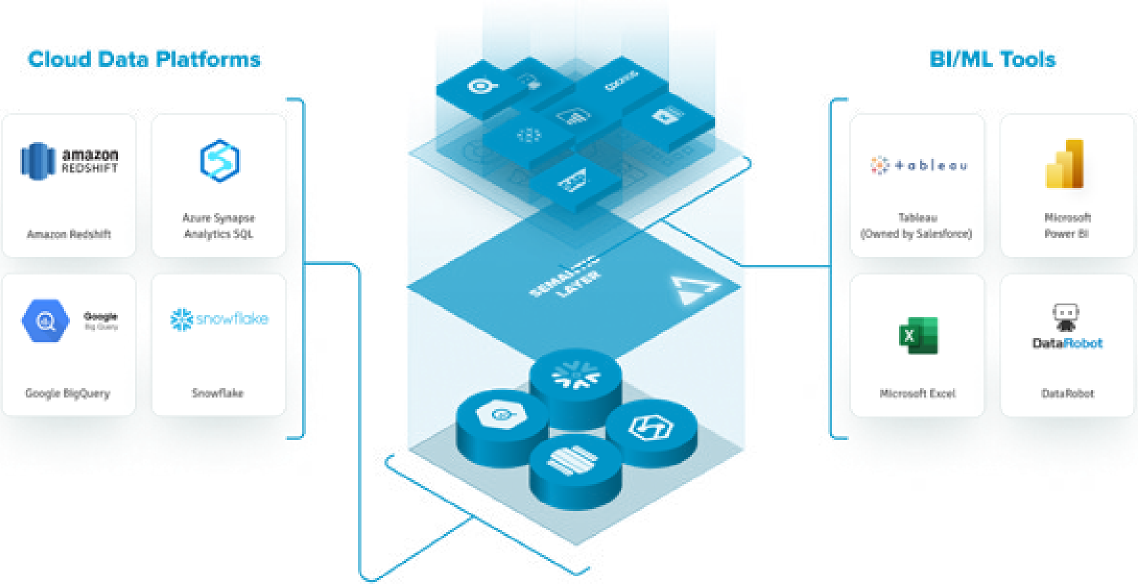


Figure 1: How a Semantic Layer Unifies BI and AI

With both teams working on the same semantic layer solution, data scientists can share (or publish) their generated features and predictions with business analysts, while business analysts provide feedback to data science teams on the quality of their predictions and model drift. Once data becomes highly accessible, teams can collaborate not just within their four walls but blend data from second and third-party data sources to unlock the power of data and analytics for everyone. Closing the gap between business intelligence and data science teams is the key to achieving a high level of data analytics maturity and applying all types of analytics at scale, and this is where Augment Intelligence comes into the picture.

So, what is Augmented Intelligence? When business and data science teams collaborate using a semantic layer, they enhance their historical data with predictive insights. Closing the gap between business intelligence and data science teams provides more visibility into the output of data science initiatives throughout the organization and enables organizations to leverage their data for predictive and prescriptive analytics. Augmented intelligence (also called augmented analytics or decision intelligence) brings AI-generated insights into traditional business intelligence workflows to improve data-driven decisions.

When most people think of augmented intelligence, they think about specific features that may appear in AI-enhanced business intelligence tools. For example, some BI tools add natural language query (NLQ) or outlier analysis to help users ask better questions or find the needle in the haystack. These are valuable features, but they are confined to the particular tool and may work differently across different tools.

Augmenting data through BI and data science unification adds AI-enhanced data to the semantic layer, providing the same insights across the consumer spectrum, regardless of the tool used. Essentially, a semantic layer amplifies the effect of the data science team by sharing their work with a wider audience and providing that audience with the ability to deliver feedback on the quality and utility of their predictions – a win-win.

Augmented intelligence has the power to transform businesses into data-driven organizations. This starts with implementing the right processes and tools to democratize data and empower individuals to utilize data through self-service analytics. The AtScale semantic layer solution is the unifying fabric that delivers all flavors of analytics, from descriptive to prescriptive, by breaking down business and data science silos. The AtScale semantic layer combines human-driven signals with machine learning to deliver an augmented data model.

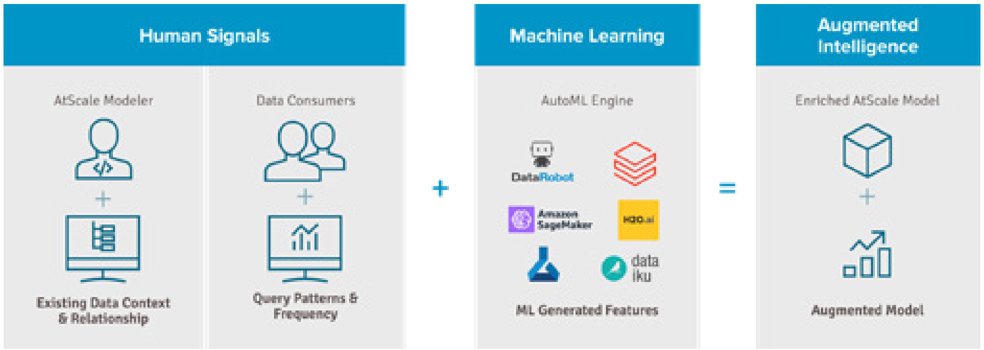


Figure 2: How AtScale Delivers Augmented Intelligence

As you can see in the diagram above, the AtScale modeler leverages their subject matter expertise to define data relationships and context in an AtScale data model. Next, data consumers query those data models and AtScale captures their interests and query patterns. Finally, data science teams augment the model using the same semantic layer by writing back features and predictions to the AtScale model, delivering augmented intelligence – all in a universal semantic layer.

Ultimately, every organization wants to empower every individual to make data-driven decisions. A semantic layer can become the vehicle for delivering augmented intelligence to a broader audience by publishing the results of data science programs through existing BI channels. Your organization can capture benefits beyond just historical analysis by feeding data science model results back into the semantic layer. Decision-makers can consume predictive insights alongside historical data. They can also use the same governed data to reliably “drill down” into the details of a prediction. Your organization can foster more self-service and greater data science literacy and generate a better return on data science investments.

Glossary

- **Aggregation:** Searching, gathering and presenting data.
- **Algorithm:** A mathematical formula or statistical process used to analyze data.
- **API (Application Program Interface):** A set of programming standards and instructions for accessing or building web-based software applications.
- **Artificial Intelligence:** The ability of a machine to apply information gained from previous experience accurately to new situations in a way that a human would.
- **Big Data:** Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze. Big data sets are characterized by 3Vs, i.e., volume, velocity, and variety.
- **Black Swan Event:** A black swan is an unpredictable event that is beyond what is normally expected of a situation and has potentially severe consequences.
- **Business Intelligence:** The general term used to identify, extract, and analyze multi-dimensional data.
- **Change Management:** Change management is the discipline that guides how we prepare, equip and support individuals to successfully adopt change to drive organizational success and outcomes.
- **Cloud Computing:** A distributed computing system hosted and running on remote servers and accessible from anywhere on the internet.
- **Correlation.** Correlation is a statistical technique that shows how strongly two variables are related. For example, height and weight are correlated; taller people are heavier than shorter people.
- **Cube.** A data structure in OLAP systems. It is a method of storing data in a multidimensional form, generally for reporting. In OLAP cubes, data (measures) are categorized by dimensions. OLAP cubes are often pre-summarized across dimensions to drastically improve query time over relational databases.
- **Dashboard:** A graphical representation of KPIs and Visuals.
- **Data:** Data is a set of fields with quantitative or qualitative values in a specific format.
- **Data Analytics:** Answering business questions using data. Businesses typically use the three types of analytics: Descriptive, Predictive and Prescriptive Analytics.
- **Data Architecture:** It is the mechanism in which data is collected and how it is stored, arranged,

integrated, and used in data systems and in organizations.

- **Data Governance:** A set of processes or rules that ensure data integrity and that data management best practices are met.
- **Data Integration:** The process of combining data from different sources and presenting it in a single view.
- **Data Lake:** A large repository of enterprise-wide data in raw format – structured and unstructured data.
- **Data Mart:** The access layer of a data warehouse used to provide data to users.
- **Data Mining:** Finding meaningful patterns and deriving insights in large sets of data using sophisticated pattern recognition techniques. Data miners use statistics, machine learning algorithms, and artificial intelligence techniques to derive meaningful patterns.
- **Data Product:** A data product is the application of data for improving business performance; it is usually an output of the data science activity.
- **Data Science:** A discipline that incorporates statistics, data visualization, computer programming, data mining, machine learning and database engineering to solve complex problems.
- **Data Warehouse:** A repository for enterprise-wide data but in a structured format after cleaning and integrating with other sources. Data warehouses are typically used for conventional data (but not exclusively).
- **Database:** A digital collection of data and the structure around which the data is organized. The data is typically entered into and accessed via a database management system.
- **Descriptive Analytics:** Condensing big numbers into smaller pieces of information. This is like summarizing the data story. Rather than listing every single number and detail, there is a general thrust and narrative.
- **ETL (Extract, Transform and Load):** Extracting raw data, transforming it by cleaning/enriching the data to fit operational needs and loading it into the appropriate repository for the system's use.
- **Hypothesis.** A hypothesis is an assumption, an idea, or a gut feeling proposed for validation so it can be tested to see if it might be true.
- **Insight.** It is the understanding of a specific cause and effect within a specific context. In this book, the

terms insight and information are used interchangeably.

- **KPI.** A key performance indicator (KPI) is a measurable value that demonstrates how effectively the entity achieves key objectives or targets.
- **Machine Learning:** A method of designing systems that can learn, adjust and improve based on the data fed to them. Using statistical algorithms fed to these machines, they learn and continually zero in on "correct" behavior and insights and they keep improving as more data flows through the system.
- **Metadata.** Any data used to describe other data — for example, a data file's size or date of creation.
- **Multicollinearity.** It is a state of very high intercorrelations among the independent variables, indicating duplicate or redundant variables in the analysis. It is, therefore, a type of disturbance in the data, and if present in the dataset, the insights derived may not be reliable.
- **Online analytical processing (OLAP).** Analyzing multidimensional data using three operations: consolidation (the aggregation of available data?), drill-down (the ability for users to see the underlying details), and slice and dice (the ability for users to select subsets and view them from different perspectives). OLAP systems are used in BI reports.
- **Online transactional processing (OLTP).** Providing users with access to large amounts of transactional data so that they can derive meaning from it. OLTP systems are used in transactional reports
- **Predictive Analytics:** Using statistical functions on one or more data sets to predict trends or future events.
- **Prescriptive Analytics:** Prescriptive analytics builds on predictive analytics by including actions and making data-driven decisions by looking at the impacts of various actions.
- **Regression Analysis:** A modeling technique used to define the association between variables. It assumes a one-way causal effect from predictor variables (independent variables) to a response of another variable (dependent variable). Regression can explain the past and predict future events.
- **SQL (Structured Query Language):** A programming language for retrieving data from a relational database.
- **Semantic Layer:** The semantic layer represents data that helps different business end-users discover and access the right data efficiently, effectively, and effortlessly using common business terms.

- **Systems of Insight (SoI).** It is the system used to perform data analysis from the data combined from the SoR or transactional systems.
- **System of Record (SoR).** The authoritative data source for a data element. To ensure data integrity in the enterprise, there must be one — and only one — system of record for a data element.
- **Stakeholder:** Individuals and organizations who are actively involved in the initiative, or whose interests may be positively or negatively affected because of execution or successful completion of the initiative.
- **Structured Data:** Data organized according to a predetermined structure.
- **Unstructured Data:** Data that has no identifiable structure, such as email, social media posts, documents, audio files, images, videos, etc.

Acronyms and Abbreviations

- **3DM** – Data-Driven Decision Making
- **AI** – Artificial Intelligence
- **API** - Application Programming Interface
- **BI** - Business Intelligence
- **CDO** - Chief Data Officer
- **D&A** - Data and Analytics
- **DAX** - Data Analysis Expressions
- **DLC** – Data Lifecycle
- **EDG** - Enterprise Data Governance
- **IT** - Information Technology
- **KPI** - Key Performance Indicator
- **LoB** - Line of Business
- **MDX** - Multidimensional Expressions
- **ML** - Machine Learning
- **MLR** – Multiple Linear Regression
- **NLP** – Natural Language Processing
- **OLTP** - Online Transaction Processing
- **OLAP** - Online Analytical Processing
- **PII** - Personally Identifiable Information
- **RBAC** – Role-Based Access Control
- **SaaS** - Software-as-a-Service
- **SoI** - System of Insights
- **SoR** - System of Record
- **SQL** - Structured Query Language
- **SSA** – Self Serve Analytics