WHITEPAPER

DEMYSTIFYING THE SEMANTIC LAYER

FOR SMARTER, FASTER AI AND BI

The What, So What, and Now What

A Perspective From Legendary Best-selling Author Prashanth Southekal, PhD, MBA

Dr. Southekal has consulted for over 75 organizations including P&G, GE, Shell, Apple, and SAP. is the author of two books — “Data for Business Performance” and "Analytics Best Practices”

ATSCALE

DEMYSTIFYING THE SEMANTIC LAYER

The What, So What, and Now What

The data economy is increasingly embraced worldwide in every industry. Data has enabled firms like Netflix, Facebook, Google and Uber to have a distinct competitive advantage.

In 2021, the market capitalization of Amazon ($1.7 Trillion), a data company, was more than the combined GDP (Gross Domestic Product) of two big G20 countries -Turkey ($780 Billion) and Saudi Arabia ($700 Billion). Companies that are data-driven demonstrate improved business performance. A report from MIT says digitally mature firms are 26% more profitable than their peers [MIT, 2013]. McKinsey Global Institute indicates that data-driven organizations are 23 times more likely to acquire customers, six times as likely to retain customers, and 19 times more profitable [Bokman et al., 2014]. Overall, data and analytics, when deployed at scale, can generate a 5% to 10% uplift in revenue and 3 to 6 percentage point increase in EBITDA margin [CGT, 2021]. Today, every company is leveraging data and analytics for improved business performance.

However, most organizations struggle to use data for improved business performance and one reason is poor data quality. According to Experian Data Quality, a boutique data management company, inaccurate data affects the bottom line of 88% of organizations and impacts up to 12% of revenues [Experian, 2015].

According to IBM research, in the U.S. alone, businesses lose $3.1 trillion annually due to poor data quality [IBM, 2020].

According to Mckinsey, an average user spends 2 hours a day looking for the right data. A report by Harvard Business Review says just 3% of the data in a business enterprise meets quality standards. A joint study by IBM and Carnegie Melon University found that over 90% of the data in a company is unused [Southekal, 2020]. All these studies point out that poor data quality affects the firm's financial performance, growth, reputation, and branding.

But what is data quality? How do you define data quality? Data is of high quality if they are fit for use in operations, compliance, and decision-making, leveraging the 12 different data quality dimensions.

These data quality dimensions are based on good data definitions [Southekal, 2017]. Unfortunately, many enterprises have challenges even in defining the data. Why? How? A data definition is a descriptor for the attributes (also known as features and the labels in data science and machine learning) of the data object. A comprehensive and consistent data definition is [Stanford, 2022]:

1. **Concise:** Described succinctly and clearly.
2. **Precise:** Described using unambiguous words when possible.
3. **Non-Circular:** The term being described should not be used in the definition.
4. **Distinct:** Described so it differentiates this data element, data entity or concept from others.
5. **Unencumbered:** The definition should not refer to a physical location or how it is created.

Against this backdrop, this whitepaper is written as a reflection paper (WHAT? SO WHAT? NOW WHAT?) to thoroughly understand the data definition problem and guide the implementation of the solution. It is based on thoughts and analysis I have seen from a practical viewpoint. Specifically, this whitepaper looks at three main "What, So What, Now What" elements: WHAT is the problem? SO, WHAT is the impact of this problem? Finally, NOW WHAT solves this problem.

Let us start first the discussion by looking at the problem. What is the problem in defining data? Why does defining the data well matter for improved business performance? Data attributes can be defined from both technical and functional perspectives. The technical data definition includes the format, type, length, etc. These are the metadata characteristics. However, the real problem is in defining the data object or the attributes in the data objects from a semantic or functional or business view because context plays a big role in how business users access, communicate, interpret, and consume data, especially in a fast-paced distributed working environment.

In today's big data world, this context is severely magnified due to the volume, velocity, and variety of data that is getting ingested into the IT systems.

So, what is the business impact of poor data definitions? Why does semantically defining the data matter? Semantically defining the data is based on the context in which the business users consume the data to run the business based on objectives, questions, and metrics. This context in business can come in three main flavours - stakeholder views, value chain impact, and business process differences.

Let's look at the impact of the semantic definition based on the above three main flavors using some common and simple business examples.

1. **Stakeholder views.** Finance and procurement often have diverse views on managing vendor relationships. While procurement sees the vendor as a service provider, finance looks at the same vendor from the costing and budgeting perspective. A low payment term (say net 30 days) is desirable for procurement as it is seen by the vendor as reward and recognition of this service. This improves the service levels of the vendor. However, this low payment term affects the cash flow which is often not supported by the finance department. So, are vendor payment terms a service element or a cost element? Here is another example from the Retail industry. Marketing needs a good amount of inventory to serve the customers, but finance believes more inventory increases the carrying cost. So, is inventory bad or good? Who defines this?
2. **Value chain impact.** Is the customer a prospect or an account (who pays for the invoice)? If a vendor gets paid for providing goods and services, can an employee be defined as a vendor given that the employee provides services and gets paid for the work? So, unless one defines the customer, the vendor, or the employee semantically based on their impact on the business value chain, there will be misunderstandings on the use of data.
3. **Business Process differences.** Let us take an example of a financial services company. Is the start time for processing the credit application when the adjudicator receives the file or is it when the processing of the previous credit application is completed by the adjudicator? Unless the start time is clearly defined, there could be multiple interpretations of these start times. Another common example is using telephone and fax numbers to derive the jurisdiction and tax rates. While the telephone or fax numbers are not meant for tax calculation, the business circumstance or even the limitations in the data model force the business to use the available resources.

To address the above contextual and circumstantial constraints and issues, we need to clearly and holistically define the data - technically and functionally. Overall, while the technical or metadata aspects are relatively easy to define, the business or functional or semantic aspects are challenging as the definition is formulated based on business context. There are four main ways to handle this data definition problem: Master Data Management (MDM), Data Integration Methods, Data Wrangling, and Semantic Layer. Let's quickly discuss these four solution options from the data definition perspective.

According to Gartner, MDM is a technology-enabled discipline in which business and IT work together to ensure the uniformity, accuracy, stewardship, consistency, and accountability of the enterprise's critical data assets [Gartner, 2022]. These critical data assets could be customers, products, vendors, factories/plants, currencies, general ledgers, and more. The goal of MDM is to provide a trusted, single version of the truth (SVOT) so organizations do not use multiple and inconsistent versions and definitions of the same data in different systems. The MDM initiative starts early in the data lifecycle (DLC) and includes defining the data, formulating the business rules, setting up the workflows, roles mapping, formulating the governance policies, processes, procedures, standards, nomenclature, taxonomies and so on.

The second possible solution to fix the inconsistent versions and definitions of the data in different IT systems is with data integration tools. The data integration process (such as EAi, ESB, Message Queue and so on) happens in the DLC. The selection of these data integration tools and practices to address inconsistent data definitions is based on three key factors.

1. **Capabilities of APIs** (REST, SOAP, RPC, GraphQL and more) and their request-response dependencies.
2. **Number of transactional systems** in scope with inconsistent data definitions that need to be integrated.
3. **Sequence of Transfer, Transpose and Orchestration (TTO)** in the data integration process.

Data Wrangling, especially cleansing the data in the canonical system like the data warehouse or data mart is also a potential option. Technically Data Wrangling is formatting, de-duping, renaming, correcting, improving accuracy, populating empty data attributes, aggregating, blending and any other data remediation activities that help to improve the data quality. Most of the data cleaning work is manual, even though stored procedures (set of SOL statements reused and shared) and automated routines are often used to support this manual labour.

The fourth option to fix the inconsistent data definitions is using the Semantic Layer. A Semantic Layer is a business representation of data that helps users access data using common business terms. A Semantic Layer maps business data into familiar business terms to offer a unified, consolidated view of data across the organization. Implementing the Semantic Layer process happens in the end of the DLC and is generally considered as part of "last mile analytics" - the key piece that connects insights to business results. In simple words, the Semantic Layer creates the context for actionable analytics.

All these four solutions (MDM, Data Integration, Data Wrangling, and Semantic Layer) that can help in fixing the inconsistent data definitions depend on data mapping. The data mapping creates data element linkages between data attributes in two distinct data models. Overall, the MDM and Data Integration methods are more suitable for compliance and operations. But if the use case is on deriving insights, then the Semantic Layer is an attractive option. In terms of data and analytics, the Semantic Layer manages the relationships between the various data attributes in the database to create a simple and unified business view that can be used for querying and deriving insights. But more importantly, each of these four methods depends on specific use cases and the control one needs on data quality in the data lifecycle (DLC). A simplified and generic DLC is shown below.

Figure 1: Simplified and Generic DLC

This brings us to the third part of the whitepaper. Now, what is the solution from the data and analytics perspective? Specifically, how to implement the Semantic Layer? Implementing the Semantic Layer requires some preparation and leadership. As Bill Gates, Microsoft's founder, said - "The first rule of any technology used in a business is that automation applied to an efficient operation will magnify the efficiency. The second is that automation applied to an inefficient operation will magnify the inefficiency." Against this backdrop, how can an organization prepare itself for the successful implementation of the semantic layer platform?

Step 1: Identify the use cases.

If the data is in one system format, you do not need the Semantic Layer. However, that is rarely the case in most enterprises today, given the variety, volume, and velocity in capturing and ingesting data into the data landscape. For example, one client, a large and global engineering conglomerate, has 17 systems. The Semantic Layer is effective if the data is distributed in multiple systems (in diverse types and formats). This is because the distributed landscape with diverse data models often creates a situation of multiple data definitions. If there are data silos in the company with multiple definitions for the same data object, then the Semantic Layer is a strong solution on the table. While most use cases describe the system's needs, meaningful use cases also identify the problem or opportunity owner, potential risks, and the business benefits in monetary terms. Also critical to success is active subject matter expert (SME) engagement to ensure proper representation of the business knowledge and understanding/use of the data.

Step 2: Identify the business KPIs and the Ownership.

Every meaningful initiative starts with a purpose that can be objectively measured and owned. Management guru Peter Drucker once said - "You cannot manage what you cannot measure". When it comes to ownership the selection of the business KPI (Key Performance Indicator) is based on the strategy and business objective. For instance, if the business objective is to improve the firm's liquidity, it is prudent to have the cash conversion cycle (CCC) as one of the KPIs. Also, it is always advisable to have a leader very close to the business and data to own the KPI. For instance, to reduce the inventory carrying cost, it is better to assign the KPI ownership to the Sales manager than to the Finance manager. This is because the sales manager has more variables under his control such as demand variability, forecast accuracy, service levels, order sizes, etc.

The granularity of the insights from the KPI also matters. If the KPI owner is a C-level executive, the KPI will be very different from that of a manager. Once we have the KPIs and ownership identified, the data objects must be identified. This process will help us define the data from the right stakeholder views. For example, if the KPIs are focused on reducing expenses, then the definition of data from the finance is more important than that of marketing with inventory management.

Step 3: Build data literacy in the Enterprise.

Taking ownership of any initiative for success requires a strong commitment and one effective way to bring data ownership is with good education or awareness. Data literacy is the ability to understand and communicate data and insights. Data literacy is to the 21st century what literacy was in the past century given that over 93% of the high-value business process today are digital and data-centric [Hurst, 2018]. The digitization and data capture rate will continue to grow in the coming years. The figure below is the 10 key data literacy competencies.

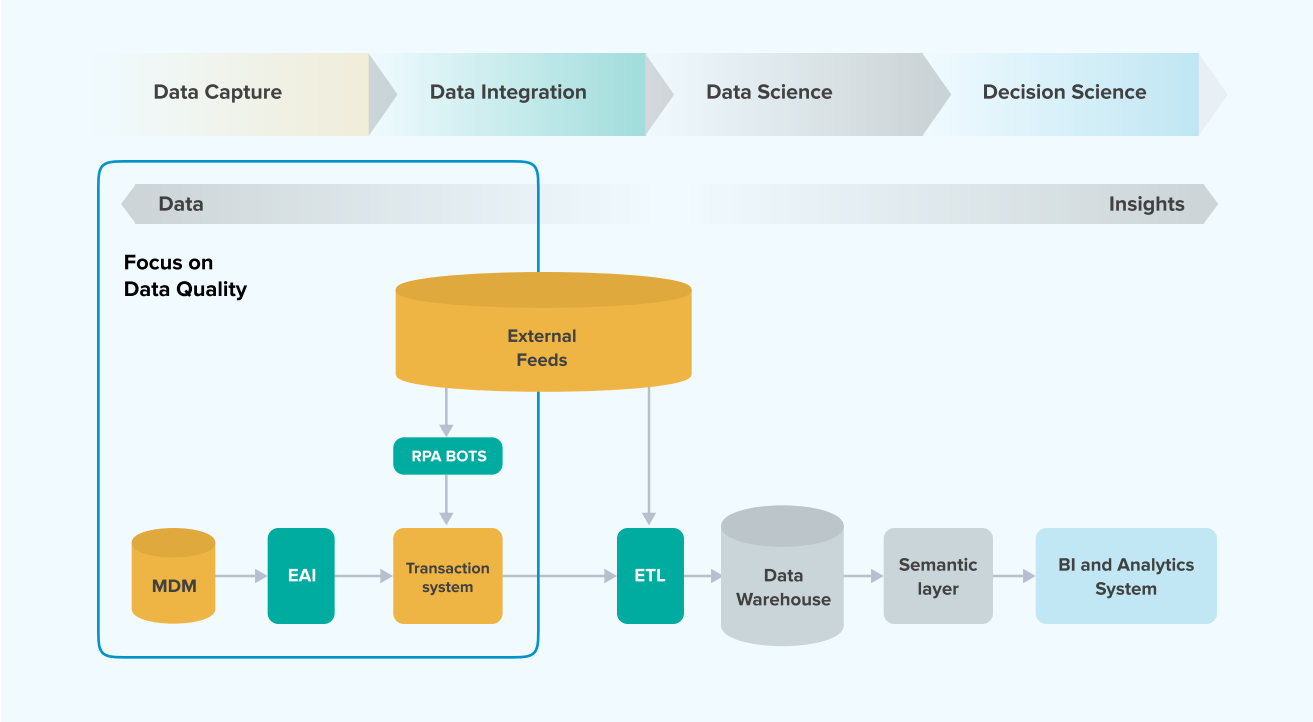

Figure 2: Data Literacy Competencies

Step 4: Define the data attributes.

With the above four steps, one can define the data attributes technically and semantically. The technical data definition includes information such as format, type, length, etc. These are the metadata characteristics. The semantic or functional view defines the data attributes from a business viewpoint, which is very challenging. Given the context can come from the KPIs and the ownership, defining the data attribute from the functional or semantic perspective at this stage should not be very difficult.

Step 5: Implement the Semantic Layer Platform

With the strong foundation built in the first four steps, you are now ready to deploy the Semantic Layer platform. The Semantic Layer platform links the analytics consumption platform with the data platforms using the facts (data values), dimensions (data attributes) and hierarchies (i.e. taxonomies) in the Data Warehouse (DWH) or any other canonical data platforms such as the data lakes or data marts or lake houses. The consumption or analytics tools can be Power BI, Tableau, Python, Business Objects. Looker, Jupyter Notebook, and even Microsoft Excel. The queries from the business users could be in SOL, DAX, MDX, etc., using the tool-specific native protocols such as XMLA, JDBC, ODBC, SOAP, and REST interfaces. By abstracting the physical form and location of data, the Semantic Layer platform makes data stored in the canonical data platforms accessible with one consistent and secure interface for the business users.

So, how does the end state with the Semantic Layer look once implemented? A holistic SL platform meets these five features: connect to any data source, support modeling, governance, security, and performance [Thuma, 2019]. Again, even if there is real-time data ingestion (say from Kafka Pubsub) or batch data ingestion (say from files), the Semantic Layer is a viable and strong solution only when multiple data definitions exist. Essentially, the Semantic Layer works as middleware between the data sources and the analytics platforms by providing virtualized connectivity, modelling, and other data management features. As all the data required to derive insights from analytics data is filtered through the Semantic Layer, the data scientists and the business users see the same data in one consistent way resulting in a single version of the truth with the same measures and dimensions.

In his backdrop, below are five key value propositions or reasons for the business to implement the Semantic Layer.

Value #1: Democratization of Data Analytics and Machine Learning (ML)

As data analytics have spread more within organizations, relying on one monolithic BI (Business Intelligence) or ML (Machine Learning) platform to meet everyone's needs is becoming less realistic. A Semantic Layer platform is needed to connect and work with diverse data platforms, protocols and consumption tools. This will decouple the data from consumption, enabling the democratization of data analytics and ML in the enterprise.

Value #2: Seamless Model development and Sharing

While Data scientists rely on raw and granular data for deriving insights from their models, this raw data has little business value from the data and analytics perspective. Businesses need insights to make decisions and not the raw data per se. But adding a data model to the raw data makes it very valuable because data models create a visual description of the business for analyzing, understanding, and clarifying the data and the associated relationships The Semantic Layer, with its data modeling capabilities, enables easy authoring, sharing, and collaborating of data models and insights.

Value #3: Improved query performance and reduced computing costs

The limited scalability and the higher up-gradation costs of on-premise data warehouses are forcing companies to leverage the power of the cloud to offer enhanced scalability, flexibility, and elasticity. While cloud computing, including cloud data warehouses, offers many benefits, these benefits come at the expense of performance and costs. We have often heard stories like the $50,000 query in the cloud [Lynch, 2020]. A good Semantic Layer platform includes a comprehensive performance management system beyond simple caching techniques in today's big data environment. At the core, the Semantic Layer facilitates improved query performance (and faster time to insights) and reduced computing costs.

Value #4: Reduced Data Cleaning Effort

Studies have shown that over 70% of the effort in data and analytics projects is on data cleansing [Southekal, 2020]. A common and consistent data definition using the governance-enabled Semantic Layer will help business analysts, data analysts, and data scientists have the same definition and context on the data. In addition, the Semantic Layer offers pre-built controls for managing data access, integration and feature creation. All this will not only reduce the data cleaning effort but will also produce reliable insights. In addition, the Semantic Layer provides a logical schema with views, stored procedures, functions, and more.

Value #5: Better Security and Governance

As the Semantic Layer sits between the data platform and the analytics tools, it secures the digital infrastructures with the right levels of authentication and authorization. The Semantic Layer can authenticate users with single sign-on solutions through Active Directory, LDAP (Lightweight Directory Access Protocol), OAuth, or any other user authentication platforms. Secondly, the semantic layer offers R BAC (Role-Based Access Control), including the ability to protect sensitive data attributes, limit data access as per user's business roles, and more.

Using the Semantic Layer for creating the context and deriving insights from data analytics is promising. To remain competitive in today's market, Toyota, the multinational automotive manufacturer, empowered its teams to work with data and analytics more independently using AtScale's Semantic Layer platform. Toyota has achieved a 2100% reduction in the insights derivation cycle time and reduced the IT infrastructure by over 60% using the Semantic Layer platform. Home Depot, the largest home improvement retailer in the United States and Canada, deployed the Semantic Layer solution from AtScale by working directly in Google's Memory cloud data warehouse, i.e. BigQuery and reduced the cost of a query by 91%. And the company realized efficiencies to increase data retention from 3 months to 3 years (a 1200% increase). This enabled Home Depot to support over 17,000 queries per day executed by internal and external users on a real-time basis. In addition, companies like Cardinal Health (health care services company), Wayfair (e-commerce furniture company), Tyson Foods (a food company) and many more have implemented the Semantic Layer with minimal disruption to their team's working style while accelerating their efforts to derive insights for better business results at a much lower cost. A generic system architecture with the AtScale Semantic Layer is shown below.

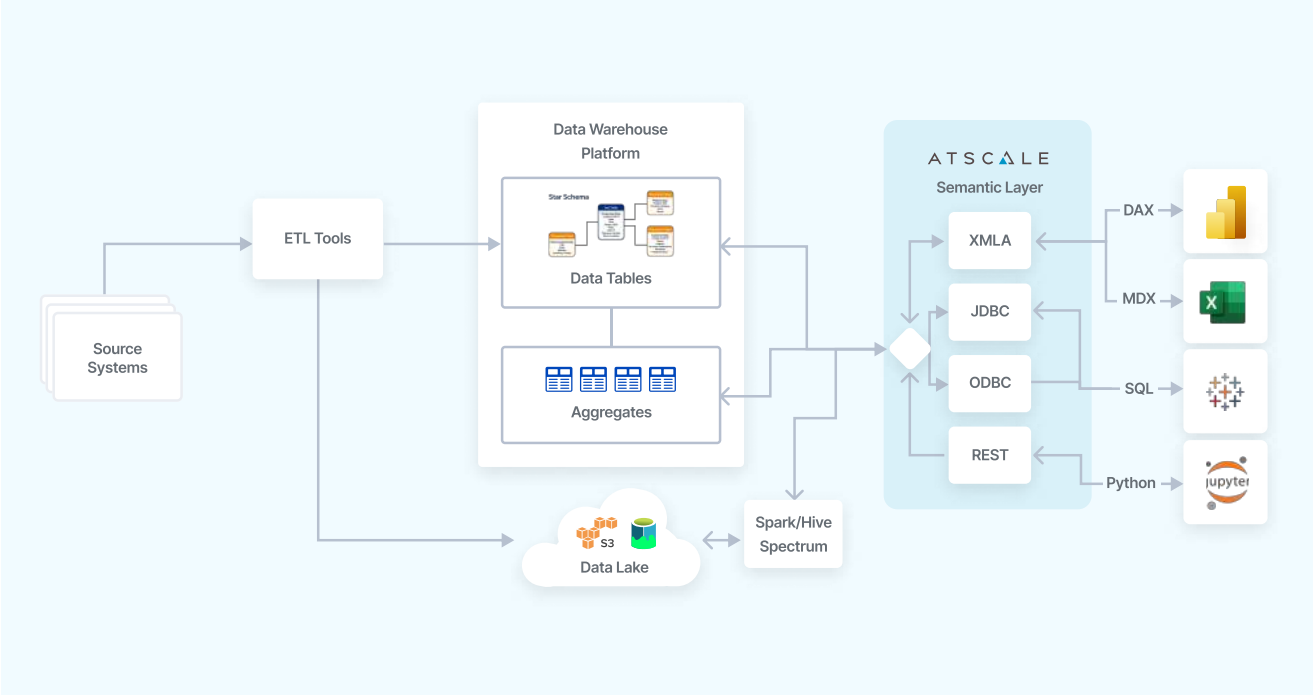

Figure 3: Semantic Layer Based System Architecture

Though enterprises have been using Semantic Layer tools to manage data for a long time, the data landscape has changed significantly in the last few years due to the increased adoption of big data, cloud data warehouses, self-serve analytics, and more. Companies need quicker and better insights in today's VUCA (volatility, uncertainty, complexity, and ambiguity) world of sudden and unpredictable change. Sadly, many of these companies have deployed numerous data and analytics solutions across diverse cloud and on-prem data platforms, resulting in data and insight silos. In addition, this distributed set-up has created challenges in data quality, literacy, adoption, and ultimately, business performance. The Semantic Layer makes data accessible to business users while hiding the complexities with data definition, manipulation, reading, and mapping. The Semantic Layer creates actionable data!

Building the Semantic Layer consists of many solutions, ranging from the organizational data itself to data models that support object or context-oriented design, semantic standards to guide machine understanding, and tools and technologies to enable and facilitate implementation and scale for interoperability and governance [Tesfaye, 2020]. But once built, business users can access the data as per the business terminology. This will reduce the complexity/costs, improve security, and accelerate and streamline reporting for the business users in today's complex data environments.
Importantly all these can happen using the data and analytics tools the users already have expertise in. This will ultimately increase the odds of better analytics adoption and improved business performance.

References

- Bokman, Alec; Fiedler, Ars, Perrey, Jesko; Pickersgill, Andrew, "Five facts: How customer analytics boosts corporate performance", https://mck.co/2Ju0xYo, Jul 2014
- CGT," Learn How Tyson Foods' Appetite for Data is Customer-Driven", https://consumergoods.com/learn-how-tyson-foods-appetite-data-customer-driven Sept 2021
- Experian, "Is Dirty Data Costing you?", https://www.xperience-group.com/the-cost-of-dirty-data/, 2015
- Gartner Glossary, "Master Data Management (MDM)", https://www.gartner.com/en/information technology/glossary/master-data-management-mdm, Feb 2022
- Hurst, Heather, "5 Systems of Record Every Modern Enterprise Needs", https://www.workfront.com/blog/systems-of-record, Nov 2018
- IBM, "Spreadsheets vs. Watson Studio Desktop", IBM Research, Jan 2020
- Lynch, Christopher, "How to Avoid the Not So Mythical $50,000 Query in the Cloud", https://www.dataversity.net/how-to-avoid-the-not-so-mythical-50000-query-in-the-cloud/, Oct 2020
- MIT, "Digitally Mature Firms are 26% More Profitable Than Their Peers", https://bit.ly/2xBTPNe, Aug 2013.
- Southekal, Prashanth, "Data for Business Performance", Technics Publications, April 2017
- Southekal, Prashanth, "Analytics Best Practices", Technics Publications, April 2020
- Stanford, "Data Definitions Best Practices", http://web.stanford.edu/dept/pres-provost/cgi-bin/dg/wordpress/, 2022
- Tesfaye, Lulit, "What is a Semantic Architecture and How do I Build One?", https://enterprise knowledge.com/what-is-a-semantic-architecture-and-how-do-i-build-one/,April 2020
- Thuma, John, "Five Things That Make a Great Universal Semantic Layer", https://www.arcadiadata.com/blog/five-things-that-make-a-great-universal-semantic-layer/, Feb 2019

Prashanth Southekal, PhD, MBA

Prashanth Southekal is the Managing Principal of DBP Institute (www.dbp institute.com), a data and analytics consulting and education firm. He is a Consultant, Author, and Professor. He has consulted for over 75 organizations including P&G, GE, Shell, Apple, and SAP. Dr. Southekal is the author of two books - "Data for Business Performance" and "Analytics Best Practices" - and writes regularly on data, analytics, and machine learning in Forbes.com, FP&A Trends, and CFO.University. Apart from his consulting pursuits, he has trained over 3,000 professionals worldwide in Data and Analytics. Dr. Southekal is also an Adjunct Professor of Data and Analytics at IE Business School (Madrid, Spain). COO Magazine included him in the top 75 global academic data leaders of 2022. He holds a Ph.D. from ESC Lille (FR) and an MBA from Kellogg School of Management (U.S.). He lives in Calgary, Canada with his wife, two children, and a high-energy Goldendoodle dog. Outside work, he loves juggling and cricket.