WHITEPAPER

HOW ANALYTICS & DATA SCIENCE TEAMS CAN LEVERAGE THE SEMANTIC LAYER

John K. Thompson, Best Selling Author, Analytics Thought Leader and Innovator

John is an international technology executive with over 35 years of experience in the fields of data, advanced analytics, and AI. He is the author of "Analytics Teams: Leveraging Analytics and Artificial Intelligence for Business Improvement" and co-author of "Analytics: How to Win with Intelligence".

ATSCALE

Introduction

Analyzing data has been a unique business activity for at least 5,000 years, and possibly as long ago as 20,000 years. We have come a long way in how we count, summarize, analyze, predict, and prescribe upcoming courses of action in science, business, and all fields of human endeavor. In the past 75 years, the progress in how we analyze data has increased exponentially. We now regularly build analytical models and complete applications that analyze massive amounts of data.

One of the most recent announcements in analytical technology hailed an upcoming Generative Pre-trained Transformer (GPT) model in GPT-4 that will have the capability of managing and maintaining over 100 trillion parameters. The current version of GPT-3, as of April 2022, has a capacity of analyzing 175 billion parameters. GPT-3 was released for general use in July 2020. In the span of less than 2 years, the state of the art in analyzing language has moved from billions of parameters to trillions of parameters. In the US system of math and counting, that is a 1000 times increase; and there is no sign of this trend slowing.

100 trillion parameters is enormous; let's compare that to the human brain. The brain has around 80–100 billion neurons and approximately 100 trillion synapses. The hotly anticipated version of GPT-4 will have as many parameters to manage and tune as the human brain has synapses!

While scaling up to an equivalent number of synapses as in the human brain is impressive, scaling up is not the only dimension of how data analytics has transformed. Scaling out is, in my opinion, even more important to analytics and artificial intelligence (AI) than scaling up. To be clear, scaling up means using more and more data of the same type or source – or the volume of data. Scaling out means including or incorporating more data of various types – or the variety – of data being included in the analytical process.

Scaling out means that we are bringing in a wide variety of data which provides an opportunity to have a better chance of achieving our ultimate goal. In analytics, our ultimate goal is to be able to model the world, or a process, or an activity, or a human behavior that we see and experience in the physical world with increasing accuracy, reliability, and specificity in the computing world.

While scaling out means that we know and have all the tools we need to model the real world with increasing accuracy in the computing world, it also brings significantly more complexity into the process. Complexity in defining data elements, defining the integration of data together, defining the relationship of data elements to another and other elements, and defining, setting, and managing the rate of change of a data element in relation to itself and all other relevant data elements.

Scaling out increases the complexity of our work, and increases the probability of our success, exponentially in both cases.

For teams involved in the field of AI, they, and we, are striving to build models that produce results that are indistinguishable from the results produced in the everyday world. We as analytics professionals are seeking to achieve the modeling reality with unfailing levels of accuracy, maintainability, flexibility, and extensibility. We are seeking to be able to model the world as we see it and live it each and every day.

To be clear, this is very difficult to do. That is why our progress in AI seems to be slow at times. This rate of progress is partially due to the fact that we need to examine each process and activity that we want to model, then attempt to model that activity. Generally, we get it wrong once, twice, or a few times, and then we do get it right and we immediately move into iterative cycles of improvement. All of this takes time, energy, and resources, but we are moving in a direction where we have an understanding of what we need to do in order to accurately model a small part of the complex world that we see and experience with our five senses.

The teams that are making the most progress in this process are teams of professionals focused on advanced analytics and AI. Next, let's outline who these teams are.

Analytics & Data Science Teams

I refer to myself, and the teams that I build and manage, as being comprised of "special snowflakes". I say this because it is evocative of the truth. In addition to being truthful, I say it because the differences we exhibit and embody are a very positive aspect of our personalities and the value that we deliver. The rest of the organization needs to know that analytics & data science teams, and the individuals within those teams, are different, and that difference is, and can be, a source of power, change, and competitive advantage. These differences are not to be managed out or reduced, they are to be understood, nurtured, and employed for the greater good.

Each individual that I have hired over the 30+ years who has turned out to be a brilliant developer, programmer, data scientist, business analyst, system engineer, data engineer, or data architect, has been an unusual or unique person.

For the most part, individuals who are adept at building analytical environments possess or exhibit the following characteristics:

- Optimistic, yet skeptical
- Intensely curious
- Mostly introverted
- Logical
- A combination of left and right brain orientation at the same time
- Intelligent
- Self-critical
- Prone to perfection
- Social, but reserved
- In some cases, they appear to exhibit a lack of focus or possibly too much focus

Managing a high performing analytics team is a unique endeavor. The teams need solid guidance, but in general, do not react well to micromanagement.

Advanced analytics projects are not, for the most part, linear. They are iterative, marked by exhilarating successes and punctuated by dead ends, missteps, and disproved theories. Most information technology professionals, while intelligent and mildly curious, do not have the intestinal fortitude for the iterative or recursive nature of advanced analytics projects.

Advanced analytics & data science teams, at least high-performing eams, are seeking to solve challenging problems. The pursuit of the solution is the goal, not adherence to a budget number or delivery according to a preset date. The optimal solution, the source of competitive advantage, is the objective.

The most successful advanced analytics and artificial intelligence teams are creative groups staffed with talented, motivated, and curious people who can convert business discussions with subject matter experts into analytical applications and solutions that can drive operational change on a daily basis. The analytical teams that realize the most success have wide-ranging mandates to drive practical and pragmatic change resulting in competitive advantage.

What does it mean to have a high performing advanced analytics and artificial intelligence team?

It means that the team is staffed with the highest caliber team members that you can attract, afford, and retain. The team is cohesive and collaborative and willing to review the projects of each other and sub teams. The team members are willing to work together for the greater good of the whole team.

The advanced analytics & data science teams present results with confidence and receive feedback and input from internal and external parties to improve data quality, model results, and the fit of the applications built for use by end users,business analysts, and other data scientists outside. Projects are scoped, described, undertaken, and completed and the groups move on to execute subsequent projects with enthusiasm and engagement.

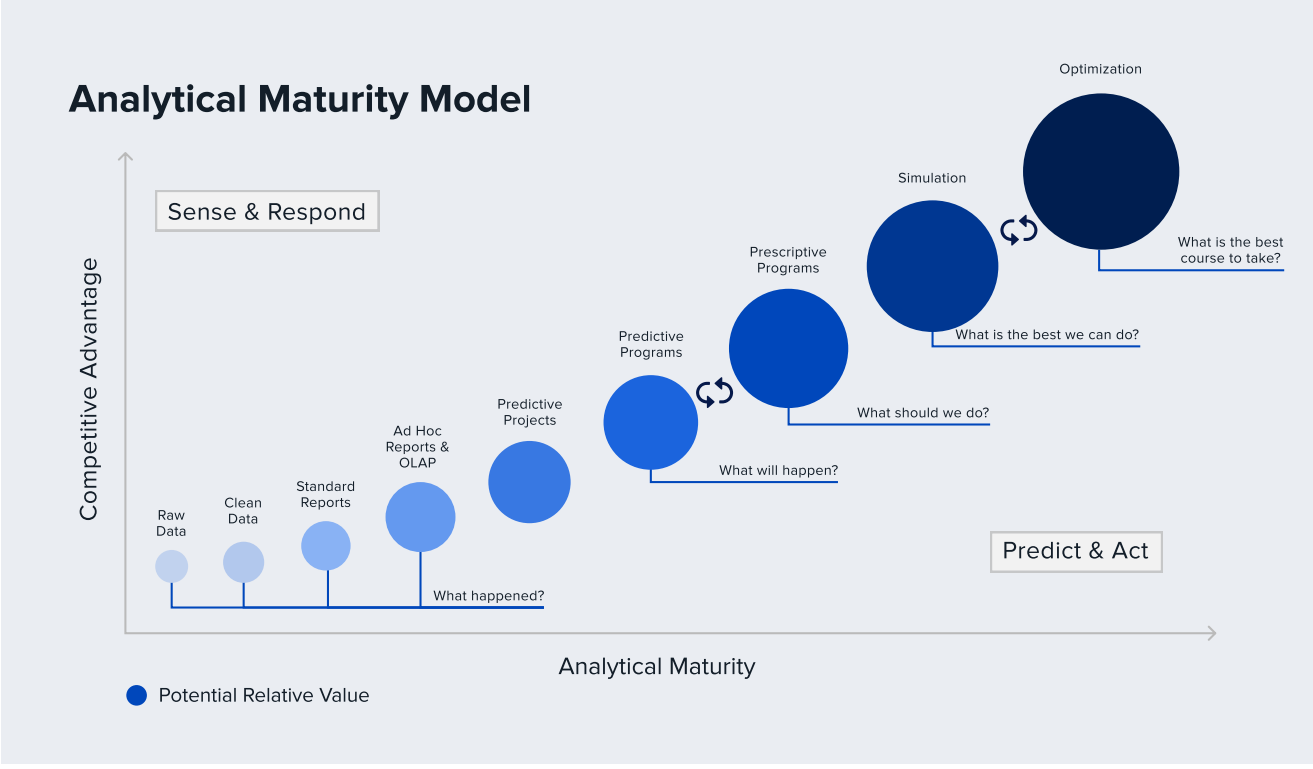
When these foundational team dynamics are in place and improving over time, you have achieved the establishment of a highly functioning advanced analytics capability for your team and organization.

An analytics team works, typically, in an organization. Each organization is moving through an Analytical Maturity Model (AMM). A number of AMMs have been offered and used in the market. I have extended Gartner's base model to be more representative of what I have seen and experienced in enterprise class companies around the world.

Let's take a look at a version of the AMM.

An Analytical Maturity Model

All companies are on a journey to move forward through the stages and phases of the Analytical Maturity Model. All companies move through the stages in a linear manner. No stages can be skipped, and all stages are required to move to the subsequent stage.



This version of the AMM has been developed from my experience in working across 20 different industries over 37 years and having delivered over 60 predictive analytical applications.

The Sense & Respond section is the phase that all companies have experienced when building their data warehouses, business intelligence, and data lake environments. This phase enables historical reporting, descriptive statistics, and the beginning of advanced analytics.

The Predict & Act phase is where advanced analytics and AI are established and evolved in all organizations. The recursive loops between the stages of Predictive/Prescriptive and Simulation/Optimization denote the next level of development that is possible for leading firms when they have mastered building and managing predictive and simulation applications.

In both cases, predictive and Simulation applications, leading organizations realize that by saving every prediction and simulation scenario, that data becomes the database which is the foundation for prescriptive and optimization applications. Only when companies come to this realization and build the environment to produce, save and manage these analytical outcomes can they move up to the highest level of achievement in analytics.

In 2022, only the most advanced firms are working at the Predictive/Simulation levels and only the top 1% of all firms are executing at the Prescriptive/Optimization levels.

One foundational element that helps all companies across both phases and all stages of the AMM is to have a well-established, highly-functioning semantic layer for all staff members to access and leverage as they are using data and analytics to execute their jobs more effectively and efficiently.

Let's discuss what a semantic layer is and why it is critical for enterprises today.

The Semantic Layer

What is the Semantic Layer?

A semantic layer provides a single, consistent definition of corporate data that also enables operational improvement in the creation of insights and analytics (e.g., BI and AI), including autonomous data access and rapid data product creation – combining rapid data modeling, virtualized data pipelining, automated data aggregation, and optimized query performance.

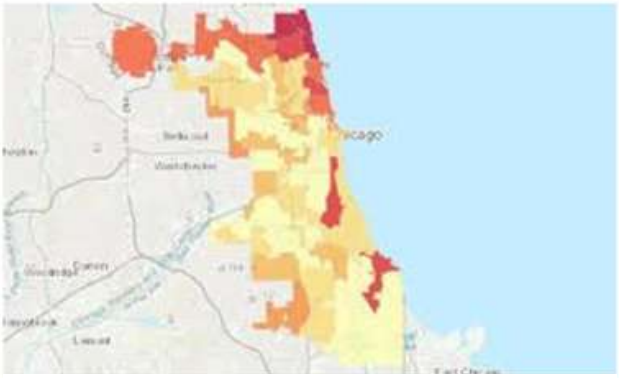
Why is the semantic layer important, useful, and valuable to analytics & data science teams in their journey to model and understand our physical world?

To model and understand a process, an activity, a cycle of human behavior, and any other phenomena that we are attempting to analyze, we need to be working from a common definition of what we are examining, modeling, predicting, and prescribing.

One of the most challenging problems faced by analytics & data science teams in all the years that I have been involved in the advanced analytics and AI field is arriving at a common definition of data describing data elements, relationships between data elements, people, processes, time, geography, models, rates of change, and concepts.

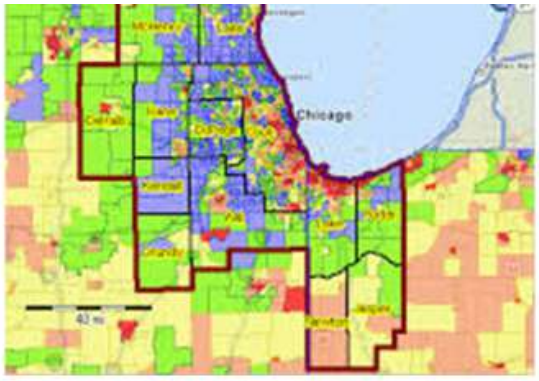
When talking about gender, geography, or any other dimension or description of data that we will use in our modeling work, what definitions are we to employ? We need to start with a common understanding of the base data and the base model that we are building.

Let's take Chicago as an example. Are we talking about the legal definition of Chicago, where the city boundaries are drawn? Are we talking about the Chicago metro area that encompasses the surrounding counties? It makes a significant difference due to differences in geographic size, population, ethnic and racial composition, and more.



Is this Chicago?

Or is this Chicago?



If you have been in a meeting where basic performance measurements are being discussed, then you know from firsthand experience that, on many occasions, the discussion starts as a debate about who has the "right" data. Who has the data that will form the basis of the discussion? This has been problematic for decades if not millennia.

The semantic layer is a solution to this problem. The semantic layer is where we as collaborators and companies as a whole decide, define, and document an agreement of definitions of data, processes, people, analytical models, and more.

As a start, we can use the semantic layer to define basic concepts like Chicago as an entity that we can all use and agree upon. And it is not that we need one definition of an entity, that is simple, easy, and would make life and our work much easier, but in many cases, organizations need multiple definitions of concepts, data, entities, and more.

Let's use Chicago again. Perhaps for legal analyses and definitions of work for the local, state, or federal governmental projects, Chicago is defined in the legal sense. The semantic layer would hold that definition of the city, boundaries, population, and related statistics. Perhaps for marketing projects and purposes, Chicago needs to be defined as the metro area including the collar counties, which includes parts of the neighboring state of Indiana. The semantic layer can hold and maintain both definitions. The semantic layer is the repository for all definitions of all objects, simple to complex.

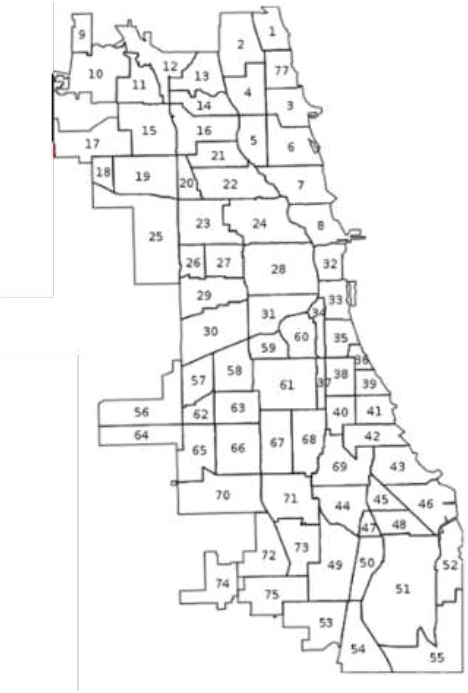
The semantic layer holds definitions of all possible objects, concepts, and elements that we need to have defined and also holds definitions of objects beyond data. The semantic layer is where we create and maintain definitions for models, processes, organizations, concepts, and more.

In advanced analytics and AI, as AI models and environments become increasingly more common and complex, it will be imperative that those models and environments are built upon definitions that are documented, understood, and are capable of being explained to executives, managers, oversight committees, federal and state regulators, and more.

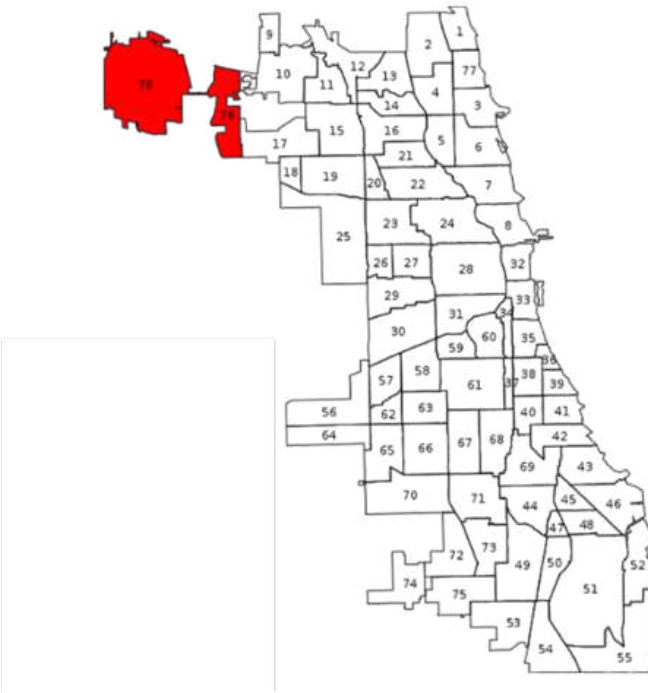
The semantic layer is where these definitions and the evolution of these definitions will be held, managed, maintained, and used for internal and external purposes.

Time also plays a role in this discussion because definitions do change. They evolve and morph to fit the current time and purpose. Let's go back to our example of Chicago. In recent history, Chicago grew by annexing the land where O'Hare Airport operates. As you can see from the maps below, Chicago expanded, and the legal boundaries of the city changed. These definitions and the evolution from one state to the next needs to be documented and maintained. The semantic layer is the tool to maintain this information and knowledge.

Chicago before O'Hare Airport Annexation



Chicago After O'Hare Airport Annexation



Cities change, they grow, and they contract in some cases. Similarly, models change constantly. The semantic layer is the place where all these definitions, over time, can reside and provide documentation to the evolutions of data, definitions, models, and more.

In advanced analytics and AI, it is crucial that companies document the changes in their analytical approaches and models. Also, as companies include more and more data sources in their scaling out of their modeling and analytics work, the relationship of each data source and data elements in those sources needs to be understood, documented, and maintained. The semantic layer is the perfect place to document these definitions, relationships, versions, changes, and the evolutionary process.

Explainable AI (XAI) is an important development in the field of AI. XAI enables the development of human readable documentation that explains how a model made all of its adjustments and decisions. One of the primary reasons why this is important is that we can now use our most powerful AI tools on the most intractable problems. Especially where regulatory bodies and governments mandate that companies be able to explain and describe how their models work. In industries such as pharmaceuticals and finance it is required by law that the companies in those industries be able to clearly explain how their AI models operate. The sematic layer is the most appropriate place to store and manage the output of XAI modules of AI environments.

The semantic layer is an enabling technology that forms a crucial part of the data and analytics infrastructure of any company. As analytical environments grow more and more detailed, complicated, and intricate, the need for a semantic layer will increase exponentially.

Summary

The majority of companies are interested in at least understanding advanced analytics and AI.

All leading companies are building AI environments, and leveraging data and analytics to extend their competitive advantage in their chosen markets. AI, data, and analytics are complex endeavors that require intelligence, resources, investment, vision, and fortitude. Not all companies have these attributes.

In addition to these tangible and intangible attributes, companies need to invest in the infrastructure that makes AI possible. Most people are aware that they need servers, software, databases, analytical tools, and more, but not all firms are aware of what the leading firms are using to build a solid, extensible, flexible, and valuable foundation that their AI operations can stand on and grow from.

The journey into the world of AI is not a single project or program, it is not a one-time excursion. Engaging the market and the world through AI is a mindset. It is a way of operating that does not have an end, and it is a way of working. To engage in the world of AI, your team needs to be world class in how they find, define, leverage, and employ data. Not just a single source of data, but many sources of data and not data on a stand-alone basis, but massive amounts of various data sources all interrelated and integrated in many different manners and schemes.

This AI and data environment needs to be understood, documented, and managed. How do leading firms accomplish this nearly Herculean task?

One way is with a semantic layer.

The semantic layer becomes the enabling technology that empowers the organization to not only use data, but to leverage data for measurable and repeatable competitive advantage.

Thank you for taking the time to read our white paper on analytics & data science teams and the semantic Layer.

More about the semantic layer and AtScale can be found at - https://www.atscale.com/

John K. Thompson, Best Selling Author, Analytics Thought Leader and Innovator

John is an international technology executive with over 35 years of experience in the fields of data, advanced analytics, and AI. He is the author of “Analytics Teams: Leveraging Analytics and Artificial Intelligence for Business Improvement” and co-author of “Analytics: How to Win with Intelligence”.