saama

# Analytics Elements to Establish Semantic Consistency in Data Lakes

Business Outcomes from Analytics

March 10, 2015

# Summary

The complexity of data supply chains and the technical infrastructures to manage them are growing every minute. Businesses are forced to be at the top of their game to leverage data to better the business and offer solutions that deliver results based on their understanding of data. Over the past decade, we have mastered building the plumbing infrastructure, but have also realized that this is just not sufficient anymore to tackle today's business problems. The hidden signals that are mined when seemingly unrelated data elements are brought together need to be orchestrated so that the business can make intelligent, data-driven decisions. However, understanding signals or correlations across time series data and dimensions is not a simple job. In most organizations, neither IT nor business has the capabilities to search for these needles in the haystack, let alone discover the correlations. Technology, math and business skills need to be combined to solve this problem. Technical solutions that utilize data lakes solve the data access problem, but leave the crucial understanding of the issues to only a few advanced users and data scientists. In this article, we discuss a method of defining and using data assets as a foundation block for faster and flexible analytics solutions.

## A Short History of Data Semantics

### Complex SQL Queries

Twenty years ago when a CFO wanted a simple quarterly summary report of Product Revenues, what did we do? We wrote complex SQL queries, extracting data from multiple tables, then grouping, cleaning, and presenting the information. Currency conversions, reconciliations and revenue recognition rules across geographies added additional tables to join and the complexity of the SQL queries kept increasing. It made for plenty of job satisfaction (or job security) for IT folks, but if you were the CFO then, you were left frustrated and at the mercy of your IT team.

### The Semantic Layer and Business Intelligence

Fortunately, the semantic layer on top of raw data reduced the complexity. The concept of providing a business semantic layer that hid the complexity of data joins was a true innovation that changed the landscape of structured data. In this new world, CFOs (or their business analysts) were able to drag the Revenue business objects, apply filters, and generate the ad-hoc reports that they needed. Subsequently, a whole slew of innovations and technologies followed and formed a new industry called Business Intelligence (BI).

With BI, programmers and business analysts worked together to build enterprise data warehouses (EDW), marts and dashboards to respond to ever-changing business needs.

### Big Data and Data Lakes

The advent of Big Data and the technical advances to access it via data lakes [1] is again redefining the data landscape. Unlike an enterprise data mart, a data lake is data in its natural state for users to examine and access. It retains all the attributes of the data, as no expectation is made on the scope and usage of data.  Data lakes provide the capability to have information that is not related by function, form, semantics or structure to co-exist. Data lakes are thus, a beginning of a fact-based hypothesis, which is derived from data lakes by discovering relationships within the facts using mathematical models, instead of rudimentary models, like joins and schema design. By their inherent definition, data lakes are not curated with governance, semantic consistency and security.

### The Challenge Today

In view of Big Data opportunities that offer broader information bases and more flexible insights, data lakes are clearly becoming the norm. So businesses now have an issue to handle. Enterprise structured data in warehouses is still the source of truth, but it is no longer complete or sufficient to generate insights. Data lakes potentially have all the answers, but are too low-level for business users to consume.

In our view, businesses have to deal with two questions:

> i) How to identify the joins to answer business questions that couldn’t previously be answered or indicate questions that couldn’t previously be asked when using a single database?
>
> ii) How to package this newfound understanding of data into data products or Analytics Apps, thereby providing a variety of business users with a quick and reliable way to solve business problems?

## A Smarter Data Lake with Analytics Elements

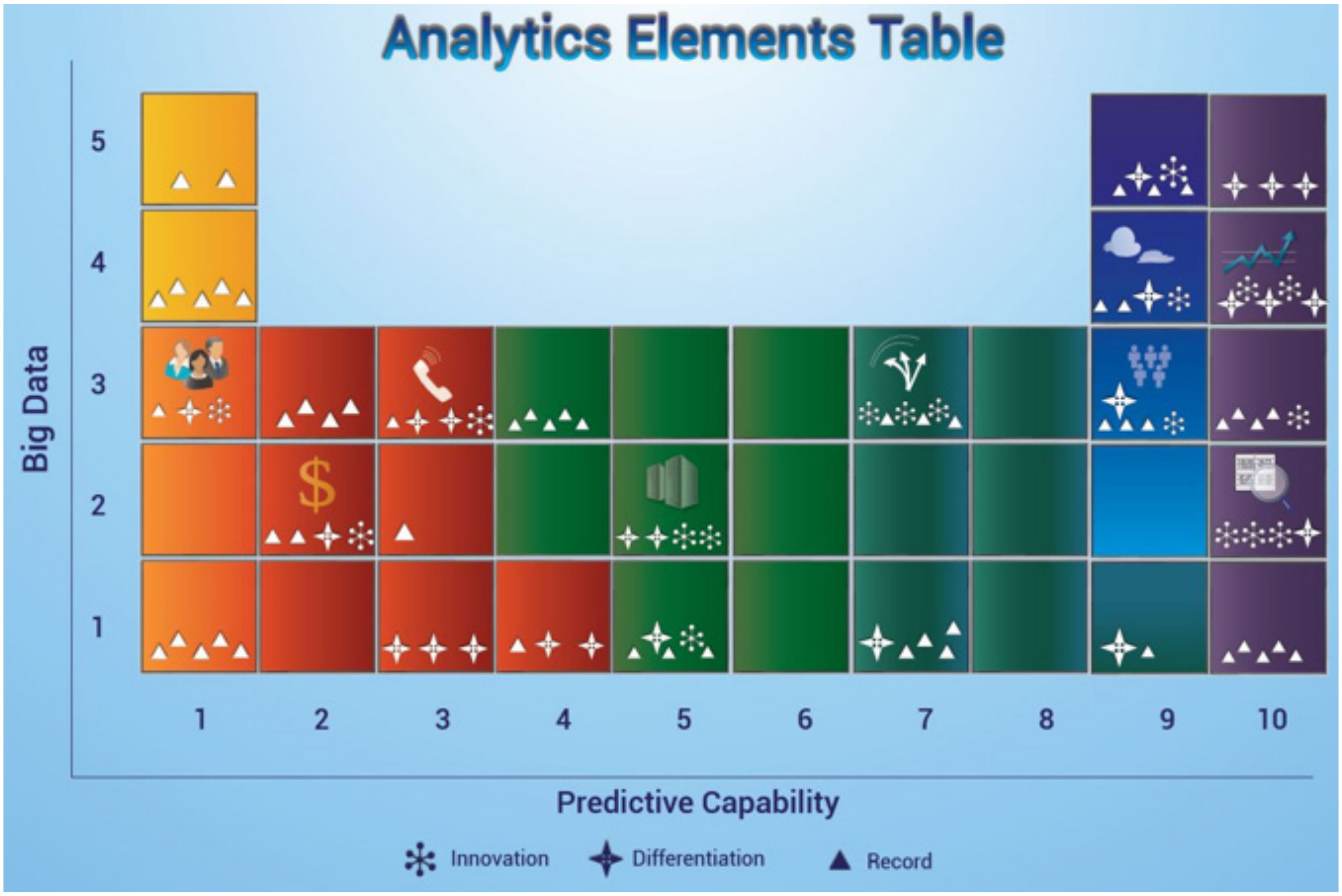
Our point of view is that a smarter data universe can be built with Analytics Elements to address the above questions. Analytics Elements (AE) embed insights by joining disparate sets and types of data, allowing businesses to understand known unknowns and unknown unknowns. An AE is a set of contextual data with defined data science models and an outcome. It is a fully functional object providing specific analytics insights around the chosen data elements.

Using unconstrained data in lakes, we can define a set of AEs, which may or may not be constrained by context and time. There could be AEs that are very basic and applicable for all applications and at any time. On the other hand, there could be AEs that are specific to a subject area (context) and valid for only a specific time. No two or more AEs can be combined if there is a fundamental mismatch in any of these two dimensions (validity/ time and context/subject area).

This semantic consistency only applies within the subject area and capability of AEs. Hence, a more advanced user who knows the data really well and knows the meaning of techniques applied to create AEs can most effectively combine them to create more complex elements or applications.

**The Analytics Elements Table**

One can think of these Analytics Elements as similar to elements in a Periodic Table, categorized into periods and groups based on their atomic properties. See figure below.



In this case, the rows or periods indicate the number of distinct databases that an organization is able to join for meaningful business results. Elements in higher rows demonstrate technical maturity of the organization in Big Data management. The columns or groups indicate the predictive capabilities that the organization has built using its data assets. Elements in the far right columns indicate the advanced maturity of an organization to predict business outcomes using data science.

The colors and shapes indicate the type of each of these Analytics Elements. Diamond-shaped Analytics Elements in the right and top sections of the table can be viewed as noble elements with the highest level of business value.  These elements provide predictive insights to take prescriptive actions for specific business outcomes. They operate upon a variety of high volume data assets, often in real-time and at high velocity. The color-coding indicates the analytics maturity of the organization. Purple is most advanced and red, least advanced.

A quick look at the table and distribution of elements indicates the hidden value of data assets that the organization is able to understand and consume.

## A Practical Example

A lumber supplier's CMO wants to analyze the revenue of one of their lumber products. A search for the "Total Revenue" Analytics Element in the new paradigm will provide some or all of the following:

- total revenue of lumber products over last 4 quarters
- most likely prediction of revenue from lumber products over the next 90 days
- potential segments among which revenues from lumber are going to decrease or increase
- competitive analysis and trends in lumber sales
- correlations from social data to indicate predicators that are potentially impacting lumber revenues

A data universe filled with semantically consistent Analytics Elements is a great start for business users. Advanced users will either consume the available AEs or build their own. And just as programmers and business analysts cooperated in earlier times, we will see data scientists and business analysts joining hands. Analytics Elements will be the constructs that they will be operating upon.

Additional Reading:
1. "The Data Lake Fallacy: All Water and Little Substance", Gartner Research Note
2. "Gartner gets the 'data lake' concept all wrong", Andrew C. Oliver, InfoWorld

## About the Authors

*Suresh founded Saama Technologies, Inc. with a vision to turn raw data into actionable insights that enable enterprise leaders to make timely and reliable decisions for critical business discovery. As the CEO of Saama, Suresh is responsible for the strategy, vision and the execution of the company's long-term plan. Suresh applied his lifelong love of mathematics to solve complex data analysis well before the term "big data" became ubiquitous. Today, Saama is made up of intelligent, helpful people with a passion for math, data, and analytics.*

**Suresh Katta**
Founder & CEO

*Sagar Anisingaraju is the Chief Strategy Officer of Saama Technologies. His recent contributions include architecting a repeatable advanced analytics engine and solutions using the same for a variety of industries. He is passionate about helping customers derive business outcomes and a differentiating analytics advantage with their data assets. He is the recipient of the Chief Strategy Officer of the Year award from Innovation Enterprise. Prior to joining Saama, Sagar ran InfoSTEP Inc. for 11 years as CEO until its acquisition by Saama.*

**Sagar Anisingaraju**
Chief Strategy Officer

A Technical Solution Paper from Saama

saama

900 East Hamilton Avenue · Campbell · California · 95008 www.saama.com