ACCELERATE DATA DISCOVERY
# CONTENTS

1. Accelerate Data Discovery
2. Profile
3. What’s Out There?
4. Turning Data into Business Value
5. Data Profiling and Semantic Analysis
6. Entity Extraction
7. Entity Co-occurrence and Relationship Detection
8. Temporal Analysis
9. Classification Models
10. Structured Data Analysis
11. Human Tagging and Active Learning
12. Reduce Time and Cost
13. Identify
14. Finding the Answer
15. The Power of Natural Language Search
16. Delivering True Data Self Service
17. Unify
18. Understanding the Data
19. Graph Analysis
20. Dynamic Modeling
21. Model Management
22. Provisioning
23. Take the Final Step to Data Self-Service
24. About Attivio
ACCELERATE DATA DISCOVERY

As enterprises capture more and more data of all types—structured, semi-structured, and unstructured—data discovery requirements for business intelligence (BI), big data, and predictive analytics initiatives have grown more complex. There’s a widening gap between what BI solutions and analytic engines have access to via traditional, almost exclusively manual approaches to finding the right data and what’s theoretically available from all the data sources across the enterprise. This gap prevents organizations from delivering data self-service to business users, data engineers, and data scientists and maximizing ROI by treating information as a strategic enterprise asset.

Many BI vendors have tried to close the gap by adding better data mashup and integration features to their products. But this doesn’t address the root problem. If an organization can only “see” 10 percent of its data, better tools simply make that 10 percent easier to work with.

What about the other 90 percent—the “dark data” that “organizations collect, process and store during regular business activities, but generally fail to use for other purposes (for example, analytics, business relationships and direct monetizing).” ¹ If businesses want to compete on analytics, that 90 percent is the key to the kingdom. And they need tools that can handle legacy data warehouses, NoSQL databases, Hadoop file systems, data lakes of raw data, and other unstructured and semi-structured systems such as SharePoint and Outlook.

¹ www.gartner.com/it-glossary/dark-data
Figure 1. Traditional data discovery tools leave a gap between data sources and analytic tools.

Closing the data source discovery gap and accelerating data discovery comprises three steps: profile, identify, and unify. This white paper discusses how the Attivio platform executes those steps, the pain points each one addresses, and the value Attivio provides to advanced analytics and business intelligence (BI) initiatives.

Attivio feeds BI tools and Unified Information applications. These tools are only as good as the information they manage.

Attivio profiles, identifies, and unifies information to address the bottleneck in agile Business Intelligence.

Figure 2. The Attivio Data Source Discovery and Unified Information Application platform harnesses the vast amount of information in the enterprise.
PROFILE

Decades ago—who knows how many—a business analyst needed a report. She made her request very specific and submitted it to an IT data analyst. Then she waited. When she saw the results, they were very helpful. But they triggered additional questions. She needed another report so she submitted another request. Then she waited some more. For far too long, that’s how it’s been for every business analyst. Request. Wait. Analyze. Repeat. In recent years, that manual and time-intensive process has begun to change.

In the last 10 years, more BI tools with greater agility have emerged to solve that problem. Sort of. Now instead of asking IT for reports, business analysts and data scientists ask IT for a specific data set, which they can explore on their own with these new tools. They can query their data in real time and generate reports—known today as “visualizations.”

This more self-service approach to data discovery has been a major step forward. It’s democratized information, making it far more accessible and shortening the time it takes for line-of-business executives and knowledge workers to make data-driven decisions. But it hasn’t addressed the root problem, which is discovering all the right information out of all the data that the organization possesses—not just the known ten percent.

WHAT’S OUT THERE?

In terms of data, that question has hounded IT and everyone who depends on IT for data for a very long time. Even before Big Data became big news, large enterprises had a lot of data. It existed in silos. So if you were a data analyst, finding out the size and content of your data universe could take a very long time. You could not ask IT for a master list of all the data sources available. It didn’t exist. Besides, IT always had more pressing matters to deal with.
There were databases, data marts, and unstructured repositories like SharePoint. And don't forget email and various archives, including those with a significant compliance layer. Then Hadoop comes along and we now have data lakes—a giant file system with very little traditional information architecture. Let's throw everything in the data lake—structured, unstructured, semi-structured. The good news: it's all there, in its raw form. The bad news: what's all there?

TURNING DATA INTO BUSINESS VALUE

When executives don't know what's in the company's data stores, they make business decisions on incomplete information. Or, equally as troubling, they don't make critical decisions because they know they have incomplete information. For IT, the problem is tactical. How can IT provide visibility into all of an organization's data, make it accessible to everyone that needs it, and remove itself as the bottleneck and gatekeeper?

Before you can know what data you want to work with, you need to know what you have. It's the first step in turning data into business value. At Attivio, we call that step profiling. Profiling is an automated process of accelerated data discovery.

During profiling, Attivio crawls your defined network finding all the sources where data resides and samples the contents to discover what they contain. It distinguishes between database records (customer name, location, phone number, etc.), unstructured documents or emails, and semi-structured information such as log files.

Attivio extracts the information, along with whatever metadata it contains, and appends additional metadata to all the objects it finds. It creates a unified semantic understanding of the information and makes it accessible via a universal index. If the information resides in an unstructured repository like Hadoop, Attivio also suggests where the information originated and whether it may be a subset of a larger store.

This automated process repeats on a schedule that can be adjusted, ensuring that new information and data sources are swiftly and automatically included in the index. Moreover, when users interact with information or make changes to metadata, such as reclassifying a data element, the index immediately reflects these actions.
DATA PROFILING AND SEMANTIC ANALYSIS

Attivio employs different advanced analytical techniques depending on the information it encounters.

1. Entity Extraction

By classifying elements in text into categories such as person, company and location, Attivio builds semantic metadata around unstructured data and uncovers the relationships between structured, semi-structured, and unstructured information. Attivio uses four classification methods. **Attivio uses three classification methods.**

A: STATISTICAL

Attivio applies machine learning to identify entities. It identifies entities without using a predefined list and disambiguates entities that have different meanings based on context. For example, Attivio can distinguish between when “Washington” as a location and a person. To increase contextual awareness when classifying entities, Attivio leverages conditional random fields (CRFs), which take into account the “data neighborhood.” Attivio understands the relationships between the data entities, providing greater accuracy and better handling of ambiguity.

B: RULES-BASED

In some cases, the relationships between entities can be expressed as a set of rules. For example, if you are looking for compliance violations, you might apply a rule that says if “free tickets” are mentioned in the same sentence as “customer name,” tag this as a potential violation for review. Another approach uses regular expressions — a sequence of characters that form a search pattern — to match substrings in data. Attivio ships with a set of pattern matchers to identify known patterns, such as social security and phone numbers. Customers can also add their own regular expressions, such as product IDs, to match patterns in their data.

C: DICTIONARY-BASED

Typically, profiling uses dictionaries to identify a known and controlled list of words or phrases, such product names. Business users can easily manage dictionaries or they can be automated through integration into other systems.
2. Entity Co-occurrence and Relationship Detection

By identifying and categorizing entities, and storing this data in the universal index, Attivio can detect relationships between entities, documents, and data that were not previously known. Using statistical graphing analysis, Attivio maps the co-occurrence of entities across the entire collection of objects in the index. This generates unique insights into the relationships between entities and the correlation of information sources across silos. Co-occurrence analysis also provides a semantic map of how the business uses entities, creating an inferred ontology of the business vocabulary.

3. Temporal Analysis

Attivio also tracks when an entity appears in the data collection as well its frequency. Temporal analysis provides insight into relationships between entities that change over time. For example, Attivio may detect a relationship between an entity for “CEO of Apple” and “Steve Jobs.” Over time, this relationship will change as the entity “Tim Cook” becomes associated with the entity “CEO of Apple.” This technique also allows Attivio to detect an increase in the frequency of entities, such as references to a particular product in customer communications.

4. Classification Models

Attivio can automatically classify unstructured content, such as emails or word documents, into categories. This provides a way to evaluate a large volume of documents without having to review each one. The groupings make it simple to filter a set of documents for deeper analysis without the noise of unrelated documents. Attivio employs a document classifier that learns based on training data. Leveraging stochastic gradient descent, Attivio automatically builds a model that can evaluate documents in the universal index and classify them according to the categories.

5. Structured Data Analysis

Attivio applies a different set of techniques to structured data typically stored in tables and rows. This allows Attivio to detect column types and find related columns. For structured data sets, Attivio not only evaluates patterns, but also looks at ranges, sparseness, outliers, cardinality, primary foreign keys, and referential integrity.
6. Human Tagging and Active Learning

Attivio enables experts to tag information to improve semantic mapping. Attivio uses the work of experts to infer new relationships among data sets. This active learning approach refines the semantic understanding of the information as people use the system, evolving as the business evolves.

REDUCE TIME AND COST

Automated profiling can perform in minutes the data normalization and preparation that would take months to do manually—if it was done at all. Without an automated solution, these tasks would consume 60 to 80 percent of the total time analysts and data scientists invest in any analytics initiative.

Attivio solves the problem of analysts who only work with 10 tables out of 150,000 because those are the ones they know exist—and they know what's in them. The other 149,990 are invisible. And it allows the analyst to follow leads that may be dead ends without sending IT on a costly, time-consuming wild goose chase. With accelerated data profiling, Attivio provides immediate visibility into all enterprise information and speeds the time-to-value of any analytics initiative.

IDENTIFY

Data scientists possess the business knowledge, technical acumen, and computational prowess to detect patterns, trends, and outliers in massive data sets. Unfortunately, they spend the majority of their time hunting for the right data to analyze. Once they locate an initial set, they must dig into supporting data and correlate their findings with related data sets. As a result, data scientists spend significantly more time preparing data for analysis than they do analyzing it. For them, identifying and unifying data for analysis consumes a disproportionate share of time and resources.
Similarly, when business analysts need data to test a hypothesis, they often inadvertently send IT on a wild goose chase looking for data that in the end might deliver no tangible benefit. A few too many of those false starts, and IT stops answering their calls. What these data-driven decision makers need is a way to quickly find all the relevant data sets on their own. And, if IT has profiled all its data sources with Attivio, they can do exactly that.

As described above, data profiling with Attivio is an automated step that crawls every data source in an organization, profiles the data, enriches it with metadata, and creates a common, universal index that IT can provide to end users. The next step in that process is identify. Recipients of the universal index can browse it—much as they would shop on an eCommerce site—for relevant data sets. Not only does the universal index provide a master list of data sources, but it will also provide additional guidelines to help users narrow their search.

“Attivio is unique in its ability to integrate customer information across multiple file types, making it all easily findable and searchable”

- Universum Inkasso GmbH

FINDING THE ANSWER

Data source identification serves a broad group. Some may simply want to know if the answer to a question already exists. The data analyst may be looking for a specific column in a set of tables. And the data scientist might want to pull a large volume of raw data from which they'll create data sets for further analysis or use by others. Regardless of their roles in an organization, they all want answers — and the quicker the better.
Through the Attivio interface, they can describe what they're looking for regardless of their level of data sophistication. Think Amazon.com for data. For example, suppose a vice president of production planning at a manufacturing company wants to find its top producing plants sorted by geography and the age of machinery on the plant floor. If this analysis already exists in the universal index, Attivio will flag it along with all the data sources that support it. Otherwise, Attivio will present the most relevant information for that analysis, including reports, emails, and columns in a database, ranked by relevancy. Business users, whose decisions are driven by time to market, find this familiar eCommerce “shopping” experience vastly more efficient and effective.

Figure 3. Attivio simplifies finding the right data
THE POWER OF NATURAL LANGUAGE SEARCH

When data scientists, analysts, or line-of-business knowledge workers search for data, they may not know exactly what they’re looking for or the extent of the data that exists. They need help.

Through the Attivio interface, they can use plain language to initiate a query. Attivio will search the titles, descriptions, values, and metadata that profiling generated from the organization’s data sources. The query needn’t be exact. Attivio can invoke approximate or fuzzy string matching, which will generate partial matches using the terms it has. It also applies relevancy algorithms to search results, returning an ordered list.

Likewise, once a user selects a useful data set, Attivio references that input in subsequent searches to “recommend” additional sources that relate to that set.

Natural language search is not just a single activity. It’s a set of activities that happen simultaneously, some of which include:

Recommendations
uncover relationships between data sets to automatically identify the best data sources for an analysis. While there may be multiple sources that match a simple query, the data sources that most closely link to the data you’re analyzing will appear at the top of the list.

Semantic search
uses a variety of signals to understand the user’s intent and handle ambiguity. For example, semantic search understands that when a user searches for “profit,” she would also want to find data sets that reference “net income.” Similarly, if a user searches for “NJ,” they would also return entries for “New Jersey”.

Autocomplete
displays relevant and popular search suggestions as users type, saving time and frustration. Attivio bases suggestions on user activity and data analysis, making the autocomplete suggestions highly targeted.
Spelling correction
increases the speed and efficiency of search. When an exact match isn't available, Attivio identifies the closest logical alternative.

Lemmatization
uses variations on words such as plurals, tenses, genders, hyphenated forms, and more. This allows for more intelligent matches against sources, for example a search for “running” would return matches for “runs” or “ran.” The vocabulary of the user does not always match the vocabulary of the information. Lemmatization overcomes this issue.

Faceted search
groups items returned from a query into the most relevant sub-categories. Users can refine their search by drilling down into a particular group or facet. As a search continues, Attivio generates new facets automatically.

Advanced syntax
increases precision through techniques such as phrase search, fielded search, Boolean matching, and proximity search.

Fuzzy matching
increases recall and allows for looser matching. Substring and approximate matches allow users to find data sources when they only have partial information or even incorrect information.

DELIVERING TRUE DATA SELF-SERVICE
When business users need to ask IT for data, it can take months for their requests to be fulfilled. Moreover, Gartner predicts that by 2017 ninety percent of information assets will be siloed and inaccessible across multiple business processes. So the data source discovery problem will become more critical, not less.

Attivio enables users of business data to identify the most relevant data for their predictive analytics and BI projects. Through Attivio’s natural language, eCommerce-like interface, they can:

- Enjoy true data self-service
- Search from a universal index of an organization’s structured, semi-structured, and unstructured information
- Browse search results ranked by relevancy
- Leverage recommendations based on Attivio’s contextual understanding of the query
- Identify existing analyses that match a query
- Accelerate decision making and time to market

UNIFY

Unify is the last step to full data self-service agility. The two previous steps—profile and identify—show business analysts and data scientists where data resides and what it contains. The output of those steps would enable them to make a more targeted request to IT. But it doesn’t provide full self-service. For that, users of data need to understand the connections and relationships between data elements.

UNDERSTANDING THE DATA

During the unify phase, Attivio graphs all of an organization’s information sources to understand how they’re related. It creates a highly visual, unified model or map that describes the sources and their linkages. Attivio exposes previously unknown relationships between disparate data sets and automatically recommends logical models to accelerate analytics initiatives. These relationships provide unique insight, including which data sets should be analyzed and how they should be used.

WITH ATTIVIO, USERS CAN TAKE ADVANTAGE OF:

Graph Analysis

Attivio builds a graph of the fields and columns in data sets and evaluates how they are related based on factors including overlapping values, similarities in column type and naming, and co-occurrence of related entities. Attivio’s graph does not rely on referential integrity (though it will use it when available) to understand the relationships between data.
Attivio produces the graph using probabilistic data structures and infers foreign and primary keys by analyzing the features of the columns and looking at intersections, column types, and data cardinality. Attivio graphs relationships across databases and data types. For example, Attivio can find references to locations and customers in unstructured content and graph their relationship to structured data sources such as customer orders.

**Dynamic Modeling**

Once data sources are selected, Attivio evaluates the relationships between them and identifies missing links through sources not currently included in the set. Functioning like a GPS for data, Attivio finds the shortest and best path to connect all the data in a group. Attivio selects subsets of the graph and generates the optimal relationship tree for the sources in the graph.

![Attivio-graph](image)

*Figure 4. Attivio creates a visual model of data and data relationships*
Model Management
While Attivio’s graph analysis and dynamic modeling produces an optimal model, it's possible that the user will have other ideas on how the data should be put together. For this reason, Attivio provides a user interface that makes it simple to edit the model. The user interface exposes all connections between data elements and allows even non-technical users to change the model, editing and removing links and defining new links.

Provisioning
The final step provisions data for data discovery and advanced analytic tools, or search-based unified information applications. Attivio’s SQL interface can provision data in a flattened data structure or as relational tables, whichever is required. All analytic tools “speak” SQL. It’s the lingua franca for interrogating, exchanging and understanding data. Attivio generates the necessary SQL statements, with no coding required by the user. Because Attivio puts structure around unstructured data, any analytics tool that looks at an Attivio data model will see a database, even if the model contains a combination of structured, semi-structured, and unstructured data. For data not indexed in the Attivio engine, Attivio can federate the query to the source systems.

TAKE THE FINAL STEP TO DATA SELF-SERVICE
Attivio delivers the power to understand, correlate, and model all structured, semi-structured, and unstructured sources to speed data discovery and accelerate analysis. It ensures that the answers you need don’t remain hidden in the dark data that typically includes up to 90 percent of enterprise information. And Attivio surfaces trends and insights that depend on the inclusion of unstructured content in analytic data sets.
WITH ATTIVIO, YOU CAN:

- Discover all the relationships in your information
- Avoid the cost of bad decisions based on incomplete or incorrect information
- Discover and act quickly on new opportunities

Attivio data source discovery and integration accelerates BI initiatives and delivers a true 360-degree of your business.

> Attivio accelerates data discovery in three steps: profile, identify, and unify. Get started on a free trial to see for yourself how Attivio reduces the data gathering process from months to minutes.

Attivio is the leading provider of self-service data discovery acceleration and unified information applications aimed at helping companies solve their big data challenges. Powering business innovation and productivity for some of the world’s top brands, Attivio enables enterprises to gain immediate visibility into all their information. The company’s solutions dramatically reduce time-to-insight so business intelligence and IT professionals can empower decision makers to act with certainty and achieve global impact.

Connect with Attivio on Twitter and LinkedIn, or visit www.attivio.com to learn more.