Data as a Service

Your Data Advantage
Starts Here

Data as a Service

Insight-driven organizations grow 30% faster than their less informed peers[1]. You must already know that otherwise, you wouldn't be reading this paper. Your data is your advantage. Used effectively, it can help direct the decisions that transform how your business operates and competes in the future.

Your challenge is finding a way to make better use of your diverse and growing data landscape. You need a faster way to integrate new sources and make data available for self-service analytics across the different data science and BI tools used within your organization. It's a difficult problem to solve and it's only getting harder.

In the not too distant past, the only type of data you had to manage was well understood relational data. Then Big Data and the nature of data changed forever. Today's data is messy, but so much more interesting. It includes images, video, geospatial data, multi-structured nested arrays and, of course, relational data.

New storage technologies, namely NoSQL data stores, have emerged as an optimal way to store this new type of data. However, as soon as organizations added NoSQL technologies to their data stacks, they quickly discovered analyzing this data using their existing BI tools is not possible without a lot of help.

As a result, to enable analytics on diverse data stacks that include NoSQL data stores, organizations implemented one or more of these common approaches:

ETL & Custom Coding

Custom coding: Data is pulled out of a NoSQL source and filtered, transformed, and aggregated by handwritten code.

ETL: Data is pulled out of a NoSQL source, transformed, flattened to a simpler relational data model, and loaded into a data warehouse. This approach is relatively common primarily among larger companies who have advanced analytical needs and a heavy investment in legacy relational analytics tooling.

---

[1] Forrester, *Insights-Driven Businesses Set The Pace For Global Growth*, October 2018

Hadoop Data Lakes

As a general purpose computing platform, Hadoop's Map/Reduce framework has supported analytics on NoSQL data from the beginning. However, a couple of issues have arisen for teams who have adopted Hadoop for analytics. First, there was usually a lack of strategy related to how to use the data. Many started with the idea of putting everything in the lake and figuring it out later. As a result, they ended up with data swamps that did little to increase data access for business teams while adding significant data governance issues and cost. The second issue was the data could generally only be leveraged by skilled big data engineers using Map/Reduce, PIG or Hive or other API-based frameworks. Data requests, long development backlogs, and data latency plague Data Lakes implementations to the point that, more often than not, business teams moved on to find other ways to access their data.

Relational Model Adapters

As NoSQL data systems entered the mainstream, there has been a growing demand for the ability to connect relational analytics tools (such as Tableau, Qlik, Cognos, MicroStrategy, and Looker) to these NoSQL systems. Using 20-year-old ODBC or JDBC protocols, relational adapters were developed to expose a virtual relational model on top of a NoSQL data system. These drivers are not without use, but customer satisfaction is low. The inherent mismatch between relational and NoSQL data models require schemas to be defined and data to be moved, transformed and flattened. The underlying mechanics of how these adapters work creates run-time bottlenecks that cause a number of performance issues.

At the end of the day, most organizations find these solutions too rigid to completely address this dynamic problem.

Why the Old Approach Doesn't Work Anymore

Increasing Business Demands for Self-Service Analytics on Any Data, Anywhere

Business leaders understand that data and analytics have the power to transform how their business operates and is the cornerstone for digital transformation initiatives. Making a fact-based decision can only happen if an organization successfully enables everyone with the analytics and the data agility they need to experiment with data to expose new insights.

The diverse nature of most enterprise data stacks puts additional strain on traditional self-service analytics solutions because they only understand relational data. To prepare non-traditional data for use in these tools,
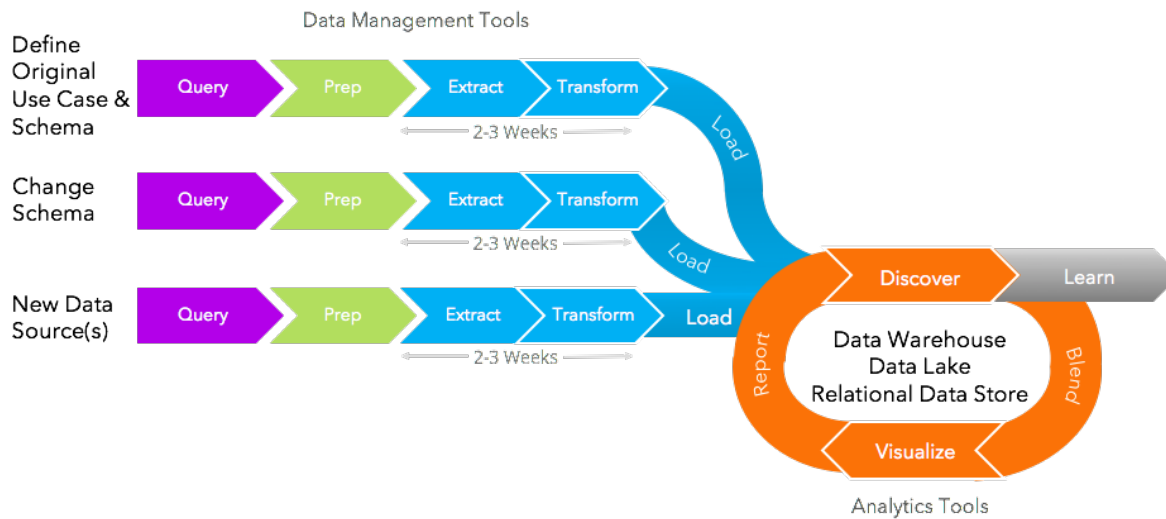
Figure 1: Old Way of Doing Analytics

one of the middleware layers described above must be implemented. It's a bit ironic that to use modern data sources for analytics, business teams are forced to wait for IT. It's a step back especially when this data potentially holds the key to propelling the business forward.

Lower Tolerance for Stale Data

Business leaders championing digital transformation initiatives want decisions driven by data, not intuition. But the traditional method of using ETL to move data into a central repository adds latency. These processes usually run in batch at off-peak times so analytics users typically analyze data that is now hours or days old. If data is stale or incomplete, teams are unlikely to rely on it to make immediate decisions and instead go with their gut.

Creates Unintended Bias

A by-product of building data pipelines to extract, move, and transform your modern data is that it adds layers of unintended bias. Each data hop has the potential to add bias to your data by the simple fact that you have to decide what data to extract, how to transform it and what aggregations to apply. You have to know the question you want to ask so the various teams involved can pull the data you need to answer that question. However, in doing so they are also determining what data relationships and correlations exist. The net result is the elimination of potentially unseen interrelationships that may change the answer or maybe even change the question.

What is Knowi's Data-as-a-Service

Knowi's Data-as-a-Service allows access to a wide variety of data sources, including SQL databases, NoSQL databases, SaaS applications, and files (such as CSV files, removing the restriction to rely solely on SQL

databases. Unlike other options that copy data and force unstructured and semi-structured data back into a tabular structure using ETL and pre-defined schemas, Knowi's Data-as-a-Service keeps data in its raw form and original source. By directly accessing data sources and raw data, a Data-as-a-Service approach enables data engineers and business analysts to collaborate in an agile manner to develop the widest possible set of ready-to-use data and make it available to anyone via a self-service model.

Make your teams 100x more efficient so they can deliver 10x more analytics projects

Benefits to Data Scientists and Data Engineers

Knowi's Data-as-a-Service solution is a unique combination of integrated capabilities that delivers the self-service capabilities modern companies need from a data platform that powers the future of their business.

Direct Access to Any Data, Anywhere: Today's enterprise data stacks are a mix of diverse data sources that are constantly evolving. Knowi natively integrates to a wide range of these sources including SQL, NoSQL, APIs, and files, to provide data engineers and data scientists secure access to any data instantly. Native integration is especially important for NoSQL sources as it eliminates the need to build complex ETL processes to move data back into relational structures shaving weeks, if not months, off analytics projects.

Dynamic Data Blending: With storage optimized for specific data types and use cases, data is rarely sourced from a single database anymore. Knowi allows users to blend data across multiple sources, on the fly, including munging structured and unstructured data.

Future-Proof: There will always be new applications and data sources to connect. With Knowi, new applications, tables, fields can be added to datasets in a matter of minutes. There is no need to rerun data loading jobs or update tables. As soon as the data is added, it is available for use.

Securely Curated Datasets: While you want data access to be flexible, it also must be secure and governed to ensure people are accessing the right data and have the authority to do so. Knowi includes a number of enterprise-level data security features that enable admins to govern access and curate datasets so everyone is using the same data for similar use cases.

Self-Service: With a Data-as-a-Service solution, Data engineers create virtual datasets by querying and blending data from source systems to create the widest possible dataset. These virtual dataset are exposed to the business users, data scientist, or downstream systems via an API. Users can start slicing and dicing the

data any way they want using their favorite BI or Data Science tools, such as Tableau, Looker, or R. Alternatively, business users can interact with a virtual dataset via Natural Language queries in applications like Slack.

Figure 2: Knowi Data-as-a-Service Architecture

The ability to transition to an agile development model and experiment rapidly with any data is the real power of Knowi's Data-as-a-Service. The cumulative effect of the platform's benefits makes it different than other approaches. Other solutions require multiple solutions to be put in place that ultimately result in a platform that is too rigid, too expensive, and too hard to use for self-service. With Knowi, everything you need is integrated into a single solution so you can meet your goal of making data accessible to anyone using their favorite data science or BI tools.

About Knowi

Knowi is an augmented analytics platform that instantly connects to any data, anywhere, and exposes previously unseen insights through AI to accelerate your business success. Knowi combines AI & NLP to create the intelligence required for anyone to make data-driven decisions. With customers around the globe and of all sizes from Fortune 50 to startups, Knowi is changing everything about how companies do analytics.

https://www.knowi.com/data-as-a-service