WHITE PAPER

Common Themes for Data Strategy

Kevin Round
IT Architect Consultant, FIS Global Banking Solutions

July 2018

Common Themes for Data Strategy

Determining an **enterprise data strategy** introduces new business and technology challenges for financial institutions. Whereas data was previously held and managed within distinct silos, an enterprise data strategy focuses on the convergence of all data sources, regardless of source or origin. This results in "crossovers" between existing organizational and technological boundaries. Organizational resistance to this sort of change is natural, and can be significant. Likewise, the technological aspects may seem daunting. It is essential to establish a good working foundation to explore enterprise data strategy requirements for our clients.

When evaluating data strategy and working with various vendors, we at FIS saw the emergence of two common themes that need to be addressed. While they were expressed in different ways, the ideas were consistent. The common themes are **data** and **analytics**. This white paper focuses on these common themes, how they impact data strategy, and ultimately form the basis for defining an effective enterprise data strategy.

It's All about the Data

Today, data drives everything. Understanding customers and potential customers implies knowing more about who they are, what they do (or are doing), and what their goals are. This encompasses everything from determining the next best product offering to detecting suspicious activity. A study by Oracle noted, "Modern engagement technologies that proactively engage customers at the right time across web and mobile … coupled with machine learning and [artificial intelligence], have the ability to transform how customers interact with their brands — delivering meaningful resolutions with the personal touch that customers desire."[1] This requirement for more intelligence, which is based on better analysis and relies on increasing volumes of data, is driving a fundamental shift in the way data is perceived and managed.

Beyond the Data Warehouse

To be clear, we don't believe data warehouse solutions will be disappearing any time soon – not in the next six months, or even in the next six years. Many deliver extremely high value. However, one comment we heard was "The data warehouse is a dinosaur", and others expressed similar sentiments. Why is the data warehouse considered, at least by some, to be outdated? Following are several key factors:

- **Latency and volume of data.** Loading a data warehouse requires extracting data and transforming it into a predefined schema (illustrated in Figure 1). Loading a large data warehouse requires a significant amount of time and processing resources. Along the way, a certain amount of data will be "lost" (i.e., not extracted and subsequently unavailable for analysis). Additionally, loading a data warehouse may require multiple ETL processes (e.g., the Kimball model).

[1] Ginovsky, John. "What do customers really want?" Banking Exchange. April 26, 2018. www.bankingexchange.com/blogs-3/making-sense-of-it-all/item/7525-what-do-customers-really-want

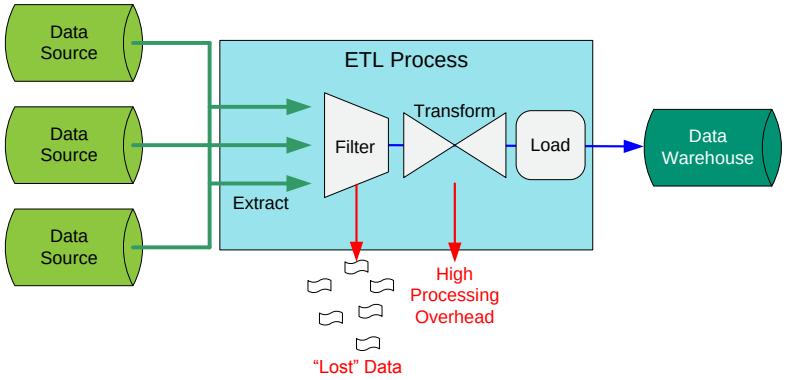

Figure 1. ETL Process for Data Warehouse Creation

- **Span of control.** Data must be understood before it can be extracted and loaded into a data warehouse. Further, management of the warehouse must extend to all applications integrated with it such that any changes within the applications that affect data are communicated to the data warehouse owners. Successful management of the data warehouse implies a centralized management process to oversee any changes to the data structure and its meaning that would impact the warehouse.
- **Breadth of data.** Is all data being loaded, or only a subset? *(see "Lost" Data* in Figure 1*)* Increasingly, the prevailing approach for data capture is to load everything with the understanding that you may not use everything that is loaded.
- **New data sources.** Financial institutional data is increasing at an extreme rate due to the volume of existing data sources and the number of new data sources. Ingesting new data sources into an existing data warehouse can be difficult, especially if the FI has no control over the format of the data or if certain elements are missing; it may be impossible if the data is unstructured (e.g., text, social media, or video data).

Storing enterprisewide and external data in a data warehouse forces the business to conform to the data requirements and technology restrictions of the data warehouse. Simultaneously, this approach imposes additional processing overhead for data that may not have immediate value and limits data availability when choosing to extract only selected data.

For modern data strategies, the direction is for the technology to conform to the business, instead of the other way around. A business may not even know what data it wants until it needs it: The enterprise data strategy needs to allow for this reality.

Next Generation Data Sources

Why is the data warehouse suddenly not enough? In the past, much analysis relied on data created and managed within the FI itself. Typically, this predominantly consisted of data collected from core and channel applications. Little or no attention was given to new, often external, nontraditional sources of data. Developing a modern data strategy requires taking stock of additional sources of data – far beyond what is readily available within a single organization.

Data can be classified as *internal data* and *external data*.

- **Internal data** is data from sources inside an enterprise. This may include systems of record (files and databases), data warehouses, data lakes, etc.

  There are two general approaches when working with internal sources when they are managed in multiple organizational silos. The first is to access the data via existing APIs. This has the advantage of allowing the owner of the data to continue managing it without having to extract and load the data into a local data store. This works well if the internal source provides an interface that is easily adapted for use with the data strategy. Where problems could surface using this approach is when the interface cannot support high-volume queries (e.g., an interface to a transactional system).

  Data extraction from internal sources is another option, one which can provide several challenges and opportunities. One of the bigger challenges is that the enterprise data strategy ultimately has political and organizational consequences (for example, parts of the FI organization will perceive a loss of control, and employees may fear that jobs will be eliminated). However, the data strategy may also provide opportunities for new jobs, while likely providing more valuable information to those same organizations that are sensing a loss of control. Enacted properly, the data strategy offers value to be gained by the FI's organizations and its employees.

- **External data** comes in many forms and has its own unique set of challenges and opportunities. External data may be purchased data, textual data (e.g., captured from social media), logs, feeds from devices (Internet of Things or IoT), graphics, video, and more. Whereas internal data is closely curated, external data may be subject to questions surrounding validity and accuracy. Purchased data is often anonymized, so it can't be directly integrated with existing customer and account information; there may be missing elements; most of it will be unstructured. The expanding variety and number of both internal and external data sources introduces new challenges to the acquisition and management of data. Two components have moved to center stage in this regard: the **data lake** and the **data catalog**.

Emergence of the Data Lake

> *A **data lake** is a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as-is, without having to first structure the data and run different types of analytics — from dashboards and visualizations to big data processing, real-time analytics, and machine learning — to guide better decisions.*[2]

Amazon Web Services (AWS) points out that most organizations will have both a data warehouse and a data lake, each servicing different needs. This contrasts with some web articles, which suggest an either/or decision. In our discussions with vendors, several factors have emerged:

[2] "What is a data lake?" AWS. June 6, 2018. aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/.

- Many data warehouses today serve the purpose for which they were designed – to provide reporting and business analytics for specific applications.
- An enterprise data strategy should focus on extending value and not trying to replace something that already works.

In the same article, AWS further distinguishes between a data warehouse and a data lake based on six different characteristics: the nature of the data stored, the use of schemas, price versus performance, data quality, users, and the type of analytics usually created from each. These are outlined in Table 1.

In general terms, a data lake supports a broader range of data (structured and unstructured). This data may be largely unmanaged, meaning that it could be un-curated, incomplete (missing elements), etc. Part of the challenge is getting such into a usable form. A data lake solution typically leverages less expensive storage and is used for more advanced analytics. While the table indicates that business analysts can only access the data that is curated, we are seeing the advent of tools that open up the data to broader access by automating some machine learning capabilities.

Characteristics	Data Warehouse	Data Lake
Data	Relational from transactional systems, operational databases, and line of business applications	Non-relational and relational from IoT devices, websites, mobile apps, social media, and corporate applications
Schema	Designed prior to the Data Warehouse implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
Price/Performance	Fastest query results using higher cost storage	Query results getting faster using low-cost storage
Data Quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (i.e., raw data)
Users	Business analysts	Data scientists, data developers, and business analysts (using curated data)
Analytics	Batch reporting, Business Intelligence, and visualizations	Machine learning, predictive analytics, data discovery and profiling

Table 1. AWS Comparison[3] of Data Warehouse and Data Lake Characteristics

[3] Ibid, 2.

The inherent value of a data lake is its capability to receive and store a wide range of data and make it available for immediate retrieval. James Serra, a solution architect with Microsoft, outlines several advantages of a data lake over previous solutions, such as RDBMS.[4] These include the capability to:

- Quickly ingest and store unlimited amounts of data long-term at a much lower cost than traditional solutions
- Collect all data, including data that you might not use initially (just in case)
- Integrate structured (e.g., RDBMS), semi-structured (e.g., XML), unstructured (e.g., text), and machine (e.g., sensor or IoT) data
- Define schema on either Write or Read – the advantage of the latter being that the data requirements are not defined until the data is needed
- Support a wide range of analytical processes

What value do corporations gain from implementing a data lake strategy? A study conducted by Aberdeen[5] evaluated the use of data lakes, as differentiated between two groups of companies – those that had implemented data lakes (leaders) and those that were planning to implement (followers). These companies reported that they were managing an average of 33 unique data sources. The graph in Figure 2 shows the improved value observed by the leaders when leveraging data lakes.

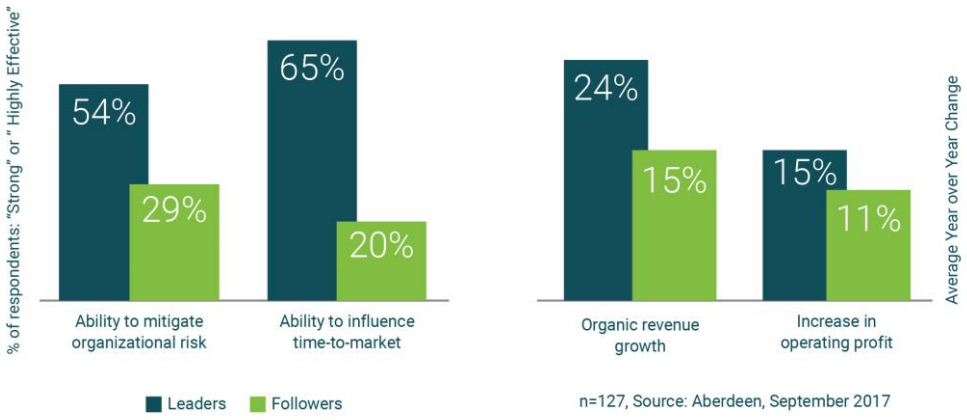

Figure 2. Value of Data Lakes to Integrate Multiple Sources of Data (Aberdeen Study)

Aberdeen also asked the reason(s) why the corporations were investing in a data lake strategy, with the following results:

- 43% to increase operational efficiency
- 32% to make data available from various silos (organizational and platform)
- 27% to lower transactional costs

[4] Serra, James. "What is a data lake?" James Serra's Blog. April 8, 2015. www.jamesserra.com/archive/2015/04/what-is-a-data-lake/
[5] Lock, Michael. "Angling for insight in today's data lake". Aberdeen. October 2017. s3-ap-southeast-1.amazonaws.com/mktg-apac/Big+Data+Refresh+Q4+Campaign/Aberdeen+Research+-+Angling+for+Insights+in+Today's+Data+Lake.pdf

- 26% to offload capacity from either the mainframe or data warehouse (reducing mainframe MIPS was a primary driver behind a recent proof of concept with a large client investigating IBM's Db2 Analytics Accelerator)

A valid concern is how a data lake addresses the ingestion and storage of unstructured data. An important distinction is that the data lake has the option of using ELT (extract-load-transform) as opposed to ETL (extract-transform-load), as shown in Figure 3. The differences are significant. First, data is extracted and loaded directly into the data lake. Ideally, there is no filtering. This addresses the previously mentioned "just in case" scenario. Second, a data schema is applied when the data is read (i.e., transform on read). Once the data is read, it can be directly processed by analytics or it can be stored as selected data.

Because the transform overhead occurs with the reading of the data, some analysts have noted that this could result in extremely high overhead when analytics are run repeatedly – such as when the analyst is tweaking the results. One way to overcome this is to selectively read the data, transform it, and store it. The selected data can then be read into the analytics process repeatedly without repeating the transformation process.

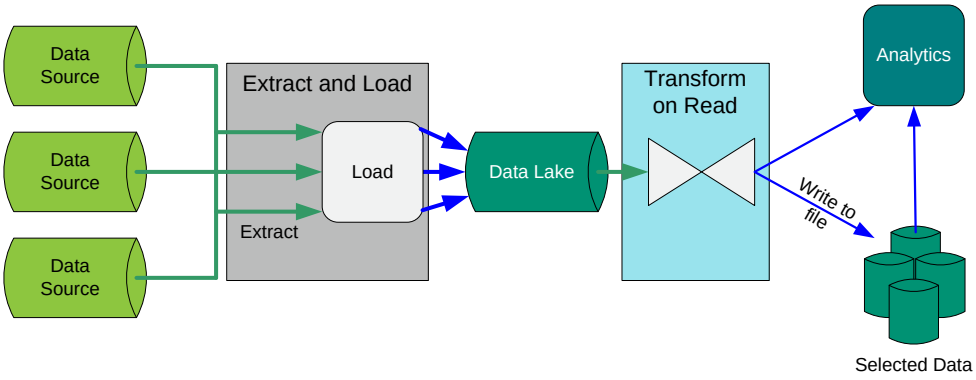

Figure 3. ELT Process – Transformation Occurs When Data is Read

The Data Catalog

The second major component of a modern data strategy is the data catalog. For enterprise data to be of value, it must be accessible. Otherwise, it's just taking up space and its value remains hidden. The data catalog provides the instrumentation that connects the data user with data sources in a meaningful fashion.

> *"A **data catalog** that lists what data the organization has, what it's called, where it's stored, who's responsible for it, and other key metadata can easily be the most valuable information offering that an IT group can create."* [6]

[6] Held, J., Stonebraker, M., Davenport, T. H., Ilyas, I., Brodie, M. L., Palmer, A., & Markarian, J. (2015). Getting Data Right: Tackling the Challenges of Big Data Volume and Variety(Preview), p. 17. O'Reilly. www.tamr.com/wp-content/uploads/2015/11/Getting_Data_Right_Preview_Edition_Nov2015.pdf.

As mentioned previously, a data lake may consist of several sources – data warehouses, HDFS™, data access APIs, etc. The challenge is how to find them. Jen Underwood, founder of Impact Analytix, points out that for many years reporting was a complex process that required a significant level of technical expertise. This complexity simplified management of the reporting infrastructure, including data sources, accessibility, and data knowledge. There simply weren't that many people who could write reports. However, reporting has evolved during this time. Modern business intelligence tools make it possible for novice analysts to create meaningful reports and graphics. Now creating reports is easy, but the management of the data underlying those reports has become much more complex.[7]

According to Gal Ziton, CTO and cofounder of Octopai, the major work effort of the data scientist is cleansing and organizing data for analysis, accounting for approximately 60% of their work time (see Figure 4).[8] This is where the data catalog comes to the forefront.

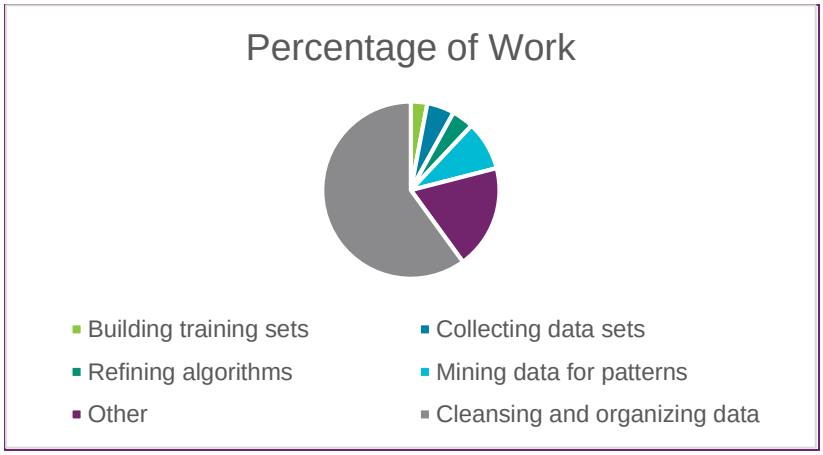

Figure 4. Percentage of Time Data Scientists Spend on Various Tasks

The data catalog plays several key roles in the management of enterprise data which typically includes security, data governance, and data lineage, although the specific capabilities can vary by solution. A pre-eminent role is security (i.e., only allowing access to the metadata for the data sources for which a user is authorized). While the data catalog also supports data governance and lineage, its most important role is supporting data scientists and analysts.

An effective data catalog should provide:

- Search capabilities to look for data sources and identify access methods
- Resynchronization, including identifying updated versions of data sources and changes in metadata
- Scoring to determine the relative value of different data sources, including usage information and data profiling
- Identification of data stewards and subject matter experts

[7] Underwood, Jen. "Why You Need a Data Catalog and How To Select One". Analytics Industry Highlights. August 30, 2017. www.jenunderwood.com/2017/08/30/need-data-catalog-select-one/
[8] Ziton, Gal & Yarmoluk, D. "Metadata Management as a Strategic Imperative". Towards Data Science. November 1, 2017. www.towardsdatascience.com/metadata-management-as-a-strategic-imperative-88a16c6ec731

- Support of user tagging and notes

Modern data catalogs have integrated artificial intelligence and machine learning. This reduces the human effort of maintaining the catalog, and allows the catalog to create metadata by analyzing the data defined to it. Recognizing a telephone number or a state field in data that is poorly defined and subsequently associating the field with the correct business term is a simplistic example of managing a data catalog using AI.

Next Generation Analytics

The other essential part of the enterprise data strategy is analytics. The focus today is on artificial intelligence (AI) and data science. AI may sound like science fiction to some, but people should be aware that we have been interacting with AI solutions for quite some time. Examples include Google dynamically determining which ads to display; Netflix guessing the next movie you might like; or having customer interactions with a bank's intelligent chatbot, such as utilized by Bank of America and Wells Fargo.[9]

> *"AI is the latest and most promising technology so far since it aims to mimic and enhance human abilities. It can perform fast computations, pattern detection and replicate understanding of spoken language. It is expected to replace or at least help employees on different levels, ranging from call center agents to portfolio managers and stock brokers."* [10]

> *"AI is a powerful tool for banks thanks to its ability to harness vast quantities of data to learn more about customer patterns and behaviors. ... As AI continues to mature, we expect to expand the insightful, personalized experiences and solutions we can offer customers and team members. We believe AI will touch nearly every piece of our business in some way." Steve Ellis, Wells Fargo* [11]

AI has existed for more than 60 years and uses several approaches to analyzing data. Figure 5 provides an overview of the evolution of AI. Early work in AI involved programming computers to play games, such as tic-tac-toe and checkers. It also saw the development of rules engines, which are used heavily with solutions such as process choreography, workflow, and decision trees. The significance is that until recently software developers still had to code the business rules and decision trees. Over time, better statistical algorithms emerged as well as much faster processors to support them; this in turn paved the way for machine learning and deep learning.

[9] Marous, Jim. "Meet 11 of the Most Interesting Chatbots in Banking". The Financial Brand. March 14, 2018. www.thefinancialbrand.com/71251/chatbots-banking-trends-ai-cx/
[10] Brooke, Sophia. "Is AI The Right Fuel For Fintech?" Finextra. May 13, 2018. www.finextra.com/blogposting/15353/is-ai-the-right-fuel-for-fintech.
[11] Yurcan, Bryan. "The top tech priorities for banks in 2018". American Banker. December 19, 2017. www.americanbanker.com/news/ai-development-top-of-2018-list-for-many-banks

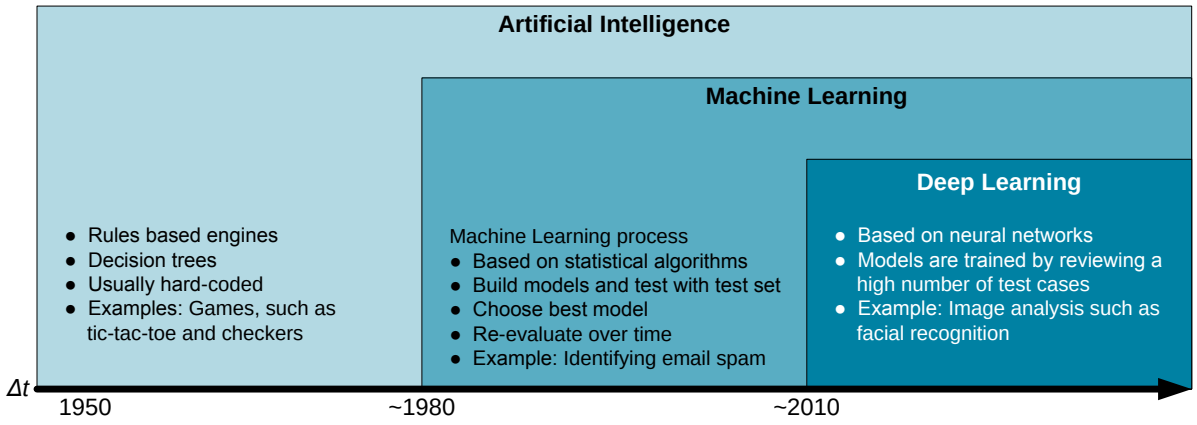

Figure 5. Evolution of AI [12]

Machine Learning and Deep Learning

The implementation of the new algorithms and technology led to machine learning (ML), sometimes called predictive analytics. With machine learning, a data scientist[13] could use a set of sample data to determine which statistical model resulted in the best analysis (i.e., the analysis of the sample data resulted in the expected results). The model is then tested using known test data to see if the proper results are generated. This requires repetitive resting to fine tune the model. Fine tuning might involve tweaking the algorithm and/or incorporating additional data elements. An important aspect of ML is that the environment can evolve over time, so it is necessary to periodically reevaluate the model and possibly re-tune it. The goal of ML is that, given similar data for an unknown sample, the model could produce a reliable prediction of the outcome. For example, is a transaction potentially fraudulent, or is a customer a good candidate for a particular offering?

Deep learning (DL) is based on computerized neural networks which simulate, to some degree, the complexity of the animal brain. Data is received in an input layer. This is passed to a series of hidden layers and ultimately reaches an outcome layer. In the hidden layers, discrete data is analyzed. Weights and biases are applied by the data scientist, which tune the network to generate the proper results. Typically, a very large volume of data is required for training the neural network, along with significant processing power. Once trained, the model can be deployed on smaller processors to support day-to-day operations.

Future of Analytics

Working with AI traditionally required significant knowledge of statistics to be done well. However, we are now seeing an emergence of tooling that makes it possible for the “citizen data scientist” – a person who is not a data scientist but has significant business knowledge and is familiar with analytics – to build and deploy models. Gartner subsequently introduced the concept of augmented analytics which they define as "a next-generation

[12] Copeland, Michael. "What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?" The Nvidia Blog. July 29, 2016. blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

[13] A data scientist is an individual with advanced training, frequently a PhD in statistics, and significant experience with data analysis.

data and analytics paradigm that uses machine learning to automate data preparation, insight discovery and insight sharing for a broad range of business users, operational workers and citizen data scientists."[14]

> *"Central to this development is the use of machine-learning automation to augment human intelligence and contextual awareness across the entire data and analytics workflow — from data to insight, to action, to impact the entire data management, BI and analytics, and data science and machine learning analytic workflow."* [15]

Looking Ahead

Where do we go in developing an enterprise data strategy that can be accepted and adopted by our clients? The starting point is to focus on what we need to deliver.

What Do We at FIS Deliver?

Figure 6 shows a high-level hierarchical analytics technology stack, as presented by Dell.[16] Moving from the bottom up, Infrastructure as a Service (IaaS) is the hardware environment. It includes abstraction and optimization of computer, storage, and memory resources. Data as a Service (DaaS) is the abstraction of the data layer. It contains the data topology and supports standard means for accessing and managing data. Platform as a Service (PaaS) includes virtualization environments, such as VMWare and Docker, as well as tooling to support the analytics work. The top layer is Data Science as a Service (DSaaS) which is where support for the data scientist exists – including sandbox creation and destruction, Jupyter Notebooks, ML tools, DL tools, etc. To the degree that the use of some tools may be automated – for example, machine learning – they could be rationally defined as both PaaS and DSaaS.

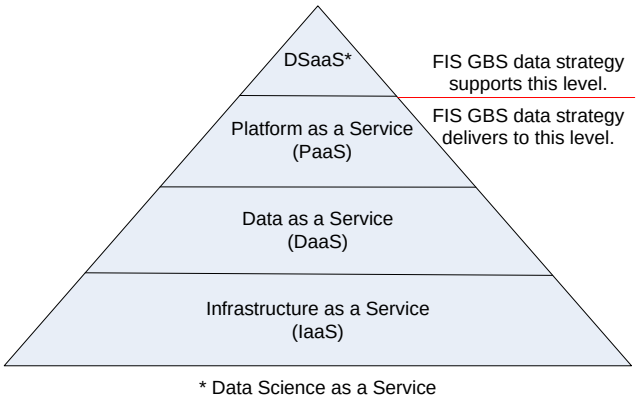

Figure 6. Dell Hierarchical Analytics Topology Stack / FIS GBS Data Strategy

[14] Sallam, Rita L., Cindi Howson, and Carlie J. Idoine. "Augmented Analytics Is the Future of Data and Analytics". Gartner: G00326012. Stamford, CT. July 27, 2017.
[15] Ibid, 11
[16] Burke, Brian. "Data Science as a Service: Technical Overview". Dell EMC Presentation. Round Rock, TX. May 30, 2018.

An important delineation in the stack is where we believe FIS GBS has a role, particularly regarding DSaaS. Our goal is to deliver function and capability through the PaaS layer. However, while we recognize that we need to support DSaaS, the actual work is likely to be through a services arrangement with a vendor. In time, as GBS develops its own resources, we may incorporate DSaaS as part of our delivery strategy.

Focus on Immediate Value and Fast Delivery

A repeated warning from the vendors we interviewed was “Don’t try to boil the ocean.” In other words, a common mistake in implementing an enterprise data strategy is trying to deliver everything at once. The better approach is what AWS refers to as a “hero project” – the first project that delivers observable high-value. Using it as a base, we define a limited set of iterative projects, each building on the other, quickly delivering value to the business. Each project must deliver an immediate, high value, limited, marketable product – also called a minimally marketable product (MMP).

The first project must extend beyond existing capabilities. For example, if an FI already has a data warehouse, simply building out more reporting or analytical capabilities using the same data source won’t prove the value of the data strategy. However, if we deliver machine learning or deep learning using the existing data sources, we are providing additional value. Ideally, the goal is to deepen the analytics using ML or DL while simultaneously expanding the number of data sources used. Subsequent iterations take what was created in the first project and build on it, adding more analytics and more data sources. Figure 7 illustrates this type of approach.

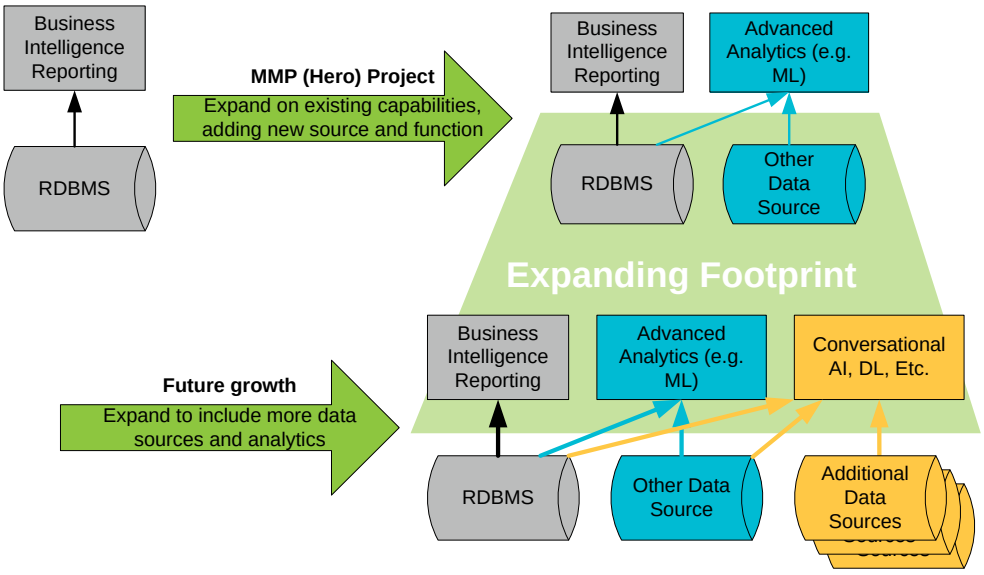

Figure 7. Building on the Success of the MMP

We envision the following set of minimally marketable products for FIS Global Banking Solutions. These are based on the Profile 8 next generation core banking ecosystem. Note that the "Base" provides foundational BI and data support capabilities including the data store and the essential BI components (i.e., data ingestion, governance, basic reporting and graphics capabilities, and the data API marketplace).

Product	Description	Value Add
Base	Data marketplace with operational and BI reports. Includes data ingestion, data governance, custom sandboxes, and a data API marketplace.	Creates the foundational analytic services. Defines the processes for expanding support for additional MMPs. Introduces data governance and data catalog.
Chatbots	AI-enabled chatbots using a third-party vendor. Supports alerts and sophisticated customer interaction.	Improves customer satisfaction while reducing operational costs. Integrates additional capabilities including natural language.
Customer Analytics	Customer analytics across the customer life cycle. Addresses acquisition, onboarding, retention, cross-selling, win back, and loyalty.	Leverages Pinpoint Marketing. Incorporates additional data sources, including external.
Regulatory and Crimes	Financial crimes, information security, regulatory reporting, and compliance (includes Hexanika).	Regulatory reporting, AML look-back, suspicious activity, analytics, FASB.
Risk	Financial and credit risk mitigation.	May include integration with other FIS data sources.
RPA	Robotic Process Automation to automate specific processes such as back office, forms, application submissions, etc.	RPA eliminates or reduces human involvement in specific functions while simultaneously increasing process speed and reducing errors.

Table 2. Proposed BI and Analytics Projects

Each subsequent product builds on top of the base, delivering additional functionality and value. As we proceed, FIS will continue to look for opportunities to leverage existing solutions. Examples already identified include Pinpoint Marketing (part of Customer Analytics) and Hexanika (for regulatory reporting).

Conclusion

This white paper began with the assertion that "*It's all about the data*." The components that comprise a data strategy are common themes throughout the industry. As Burke indicated, the focus in the past was to deliver siloed and geographically diverse products.[17] This resulted in having collections of products, many with their own set of semantics, some of which are extremely specialized. The challenge is how to make the collective data available and meaningful – yielding the most value for the enterprise.

We believe the FIS strategy addresses this issue. First, we realistically acknowledge that pulling all data together in one location under one set of schemas is impractical. Instead, we focus on technologies that support the integration of multiple data sources using data lake and data catalog capabilities, with appropriate governance and securitization of data. FIS GBS anticipates using open solutions where practical, specifically those solutions

[17] Ibid, 6, p. 16

that have achieved industry-wide acceptance. At the same time, we intend to leverage existing vendor relationships, and we intend to keep the solution open and adaptable to our clients.

As Tom Blomfield, founder of the British bank Monzo, told *The Guardian*, "There is just so much valuable data sitting in these bank accounts. I don't think banks are doing it malevolently, they are not using it for nefarious purposes … They just can't get it out. And if consumers are able to unlock that data and use it for their own benefit it will create entire new industries."[18] Using recent banking regulatory changes in Britain, Monzo is positioned to be an intermediary between the customer and the traditional bank. They will accomplish this, in part, by using the bank's own data to enrich the customer experience.

The path forward is to pursue an encompassing enterprise data strategy using a realistic approach. This can be accomplished through the delivery of a series of specifically targeted, high-value solutions, each expanding on its predecessor.

In the coming weeks and months, we plan to further elaborate on the proposed products listed in Table 2. We will also continue writing about the technology stack, why artificial intelligence has become so important, and how the proposed strategy is intended to address all FIS clients.

[18] Lewis, Tim. "Is Monzo the Facebook of banking?" The Guardian, US Edition. December 17, 2017. www.theguardian.com/technology/2017/dec/17/monzo-facebook-of-banking

Glossary of Terms

Analytics	A general reference to solutions provided by Business Intelligence (BI) tools, including reports, dashboards, graphics, etc.
AML	Anti-money laundering.
API	Application program interface.
Artificial intelligence (AI)	A field of computer science that emerged in the 1950s and continues to evolve today. The goal is to emulate human thinking. Early approaches produced very modest outcomes. However, recent improvements in hardware and statistical algorithms have made AI a sophisticated technology.
Augmented analytics	A concept developed by Gartner wherein it is realized that the implementation of AI is neither a human nor a machine solution. Rather, it is the better facilitation of work by using AI to support human effort.
AWS	Amazon Web Services.
Business Intelligence (BI)	Refers to the technologies, applications, and methodologies that are utilized to collect, integrate, analyze, and present business information for the purpose of supporting the business and making better business decisions.
Chatbot	Chatbots support the interaction of a consumer (or other user) with an institution using a traditional chat session. With the incorporation of natural language and AI in this technology, along with proper training and orientation, it is possible for many chat sessions to be managed by an AI product.
Citizen data scientist	A term developed by Gartner that describes the fact that modern tooling makes it possible for knowledgeable business analysts who do not have advanced training in statistics to perform data scientist tasks. A recent Gartner report defines a citizen data scientist as "a person who creates or generates models that leverage predictive or prescriptive analytics but whose primary job function is outside of the field of statistics and analytics."
Data analyst	In the context of this white paper, a data analyst is a person who uses data and BI tools to perform Data Analysis tasks (e.g., to inspect, cleanse, transform, and model data to discover useful information, make informed conclusions, and support decision making).
Data catalog	A solution that supports data governance, stewardship, and security while providing data analysts with efficient and clear access to data sources. A data catalog is critical to the support of an enterprisewide data lake.
Data lake	A centralized repository for storing all data (structured and unstructured). In the context of this white paper, this pertains to a combination of HDFS, IoT data, and structured data (e.g., data warehouse), plus the use of APIs to access remote data.
Data scientist	A person with advanced training in statistics who is familiar with machine and deep learning algorithms and processes, including the identification and evaluation of test data. Several years of

	experience are often implied as well. The data scientist is currently considered one of the most lucrative jobs in IT.
Data source	Any data that might be fed into a data analysis solution, which is almost everything. Data sources are typically defined as internal (usually curated and highly secured) and external (often unstructured and sometimes of questionable validity).
Data warehouse	A tightly managed database designed to support analysis. It typically requires the transformation of data from local semantics into a set of common semantics. While this level of management results in highly reliable data, it also constrains how much data can be ingested by the data warehouse due to the effort required to define the transformation rules and the processing overhead of extracting, transforming, and loading data into the data warehouse. (*See also* ETL)
Decision tree	In computer programming, decisions trees are represented by if-then-else or case logic. Decision trees are often captured in rules engines, mapping possible outcomes based on a series of choices.
Deep learning (DL)	For our purposes, deep learning is considered part of machine learning (though some writers insist on distinguishing between the two). It typically involves a computerized neural network, which consists of an input layer, an output layer, and some number of intermediary layers. Certain calculations are made at each layer. Rather than adjusting the algorithm directly, the data scientist adjusts the neural network by setting various bias and weight values within each layer. An excellent overview of neural networks is available from 3Blue1Brown at www.youtube.com/watch?v=aircAruvnKk.
ELT (Extract, Load, Transform)	Represents a modern way of acquiring data for analytics purposes. Data is read in and written out, without augmentation, to a data store (e.g., HDFS). When the data is subsequently read from the data store, it is transformed into the appropriate format and values.
ETL (Extract, Transform, Load)	Refers to the traditional method of loading data into a data warehouse. Data is read in (extracted) and modified (transformed) to meet the requirements of a previously defined schema. Once the data is in the proper format, it is loaded into the target data store, usually a data warehouse or data mart.
FASB	Financial Accounting Standards Board.
FI	Financial institution.
GBS	FIS Global Banking Solutions.
HDFS™	Hadoop Distributed File System.
Hexanika™	The FIS next generation regulatory reporting component is provided via a licensed reseller agreement with Hexanika™, a company that provides regulatory reporting based on the Apache Spark technology stack.
Internet of Things (IoT)	Refers to the connection of devices (beyond computers and smartphones) to the internet. Any stand-alone internet-connected device that can be monitored and/or controlled from a remote location is considered an IoT device. Examples abound, including kitchen appliances, medical devices, home electronics, and much more.

JSON	JavaScript Object Notation.
Kimball model	A common process for creating data warehouses which typically requires more than one ETL process to properly normalize, format, and load data.
Latency	The amount of time that elapses between when something is requested and when it becomes available. For example, the amount of time required to provision a data warehouse with the source data.
Machine learning (ML)	A method of data analysis that automates the building of analytical models. It is a branch of AI which is based on the premise that systems can learn from data, identify patterns, and make decisions with minimal human intervention. The ML process consists of a collection of algorithms that are used by data scientists to create predictive models. The process includes selecting an appropriate algorithm or algorithms, training the model using a known set of data and results, adjusting as needed, and testing using an additional set of known data and results to determine accuracy. Once trained, an ML model can be used in production to analyze either live or stored data for insights.
MIPS	Million Instructions Per Second (computer processing metric).
MMP	Minimally marketable product.
Neural networks	A computing system made up of multiple highly interconnected processing elements, which process information by their dynamic state response to external inputs. Neural networks simulate, to some degree, the complexity of the animal brain. Data is received in an input layer. This is passed to a series of hidden layers and ultimately reaches an outcome layer. In the hidden layers, discrete data is analyzed. Weights and biases are applied by the data scientist, which tune the network to generate the proper results. Typically, a very large volume of data is required for training the neural network, along with significant processing power. Once trained, the model can be deployed on smaller processors to support day-to-day operations.
Nontraditional (data) sources	Data not previously (traditionally) used by an institution for analysis. Nontraditional data includes XML, JSON, sensor data, social media, generally available data (e.g., government data), and purchased data.
RDBMS	Relational Database Management System.
Robotic Process Automation (RPA)	The automated processing of typically labor-intensive processes. The idea is that RPA tasks can be completed faster, with fewer errors, and without human involvement. An example in the banking industry is the process flow of loan request, approval, and onboarding. RPA is similar in nature to an earlier technology called *process choreography*, and it is different from *workflow* in that it attempts to avoid human intervention. While primarily rules-based, AI enters into RPA as a way to support better decision making and alternative processing.
Schema on read	The concept that data is stored in its original form and isn't transformed into a usable structure until it is read for processing. Used with data lakes.

Schema on write	The concept that data is transformed into a desired format before being stored for subsequent analytical use. Used with traditional data warehouses.
Test data	Data used to train and evaluate machine learning (ML) and deep learning (DL) models. Typically, a set of data, consisting of values and expected results, is divided into two sets. The first set is used to train the model, allowing the data scientist to make tweaks along the way. Once the training data achieves the desired results, the second set of data is used to evaluate the model.
Unstructured data	Data that is not in a relational schema format. Arguably, some data is semi-structured, such as XML and JSON. The significance in this white paper is noting the distinction between data that is readily consumable for analysis versus data that requires some degree of pre-processing before use. Human-generated unstructured data includes text files, email, social media, website contributions, mobile data, communications (phone recordings, text messages), media files (photo, audio, and video files), software application data, and more. Machine-generated unstructured data includes satellite imagery, scientific data, digital surveillance, sensor data, and more.
XML	Extensible Markup Language.

Contact Us

For further information, visit our website at www.fisglobal.com.