White Paper

The Compelling Advantages of a Cloud Data Lake

By Nik Rouda, ESG Senior Analyst

April 2017

This ESG White Paper was commissioned by Amazon Web Services and is distributed under license from ESG.

Contents

Defining the Data Lake

"Big data" is an idea as much as a particular methodology or technology, yet it's an idea that is enabling powerful insights, faster and better decisions, and even business transformations across many industries. In general, big data can be characterized as an approach to extracting insights from very large quantities of structured and unstructured data from varied sources at a speed that is immediate (enough) for the particular analytics use case. Enabling big data are new economics that make the value of these insights worth more than the cost of building and running a system to capture data and extract the insights at a scale that traditional approaches cannot handle. These new technologies often include scale-out distributed storage and processing resources on generic hardware, supporting open source software for a wide range of parallelized analytics techniques. Hadoop and Spark are good examples of big-data-enabling technologies, though the ecosystem incorporates many other distinct and complementary pieces.

One particularly interesting big data concept is the "data lake." A data lake is a specific architectural approach designed to create a centralized repository of all potentially relevant data available from enterprise and public sources, which can then be organized, discovered, analyzed, understood, and leveraged by the business. Again, cost-effectiveness is important here, as the utility of all that data may be initially unknown; but the expectation is that different groups will find new and valuable ways to leverage it over time. Business analysts and data scientists might join data in new combinations to get richer insights, find new questions to answer, or bring new analytics techniques to look at it in innovative ways. A data scientist may wish to use graph analytics or machine learning, while a business analyst might prefer to utilize a more familiar ad hoc query language like search or SQL or simply interact via her existing business intelligence (BI) dashboard. Developers and DevOps teams may want to build entirely new business applications that leverage the data lake. Clearly, data lakes are intended to not just handle large scales, high speeds, and diversity of data, but to also provide a range of agility and versatility for analytics that empowers all interested knowledge workers in the enterprise. Cloud services offer similar flexibility and even better economies of scale that can also help enable data lakes.

Of course, the business will likely have existing technologies that serve some of these analytics functions. Relational databases, enterprise data warehouses, data marts, and BI applications are very common for businesses. The problem is that legacy architectures have a number of limitations. Common issues with traditional approaches include scale limitations, high costs, performance gaps, accessibility restrictions, and rigid designs. Usually organizations must start with a specific use case in mind, and work backwards with a fixed analytics model using data that has been extracted from a particular source, transformed according to narrow definitions, and loaded into a fixed hardware platform. Any changes to this system are usually difficult and time-consuming to make. Business users begin to assume they can only refer to specific existing reports and metrics, and the pain of requesting something new from IT often hinders their creativity. A lot of data is simply ignored or discarded. Big data, and especially data lakes, aim to turn this whole insights workflow on its head. The environment should be built to have *all* data discoverable to anyone who comes up with an idea or intriguing question, and let them analyze that data in any way they like to get an immediate answer. Ideally, this provokes more and better questions, and the business becomes far more curious and nimble.

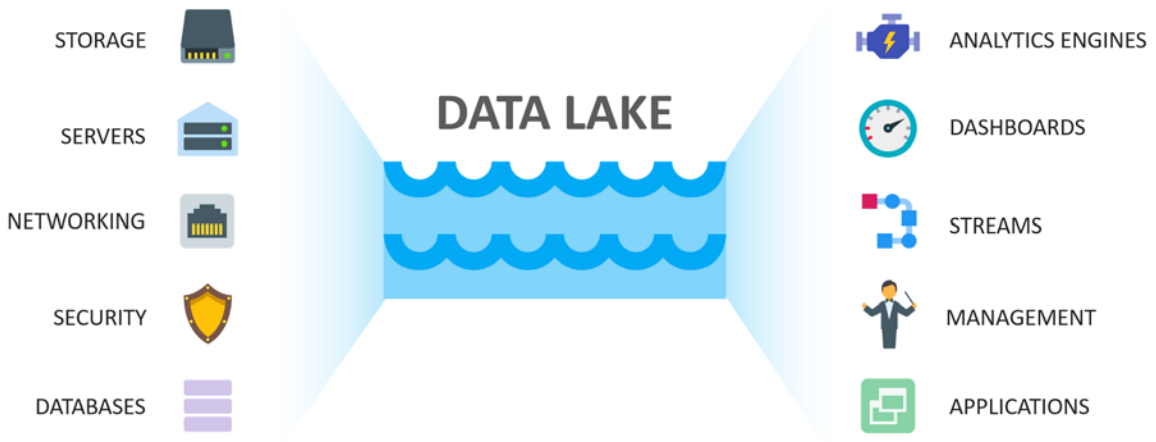
The Essential Elements of a Data Lake

Despite the breathy editorials of glossy business publications, big data is neither magic nor a single ultra-powerful device. A data lake is typically assembled from a number of discrete technology building blocks, all of which are readily available to any business. This includes:

- Choices of storage to contain all that data, plus file systems or objects to locate it.
    - Some storage may need to be focused on cost-efficiency for extreme volumes of cold data, or data that is not used very frequently.

  - Some storage may need to be high performance for intensive workloads or for handling fast-streaming data.
- Servers to manipulate the data in their memory and processors.
- Networking to connect adjacent resources and give access to remote workers.
- Security to protect the data from threats, both internal and external, including authentication, authorization, and encryption services.
- Databases to organize data more clearly.
- Advanced engines to query and analyze data; and build, test, and run models in a variety of ways, including machine learning and AI.
- Applications and dashboards for data visualization and collaboration around insights.
- Streams to move and capture the data, sometimes also to process and analyze it in the pipeline.
- Management frameworks to govern the data, including moving, transforming, and cataloging data.
- Application services to facilitate development and deployment.

Figure 1. The Essential Elements of a Data Lake



Source: Enterprise Strategy Group, 2017

Interesting and important is the extent to which each of the components and functions above can be decoupled from one another. Unlike traditional database and data warehouse technology stacks, a data lake should enable the architect to define each element—compute, storage, networking, and more—individually and in fact mix, match, customize, and grow as desired. This decoupling brings much needed flexibility to the environment, and further enables blends of varying resources to deliver the needed balance of scale, performance, versatility, and costs. Another advantage of decoupling is that new technologies can be swapped in and out as innovation continues. An example here could be migrating to use Apache Spark instead of Apache Hive or Elastic MapReduce.

The Value of a Data Lake

Clearly a data lake is a powerful architectural approach to finding insights from untapped data, which brings new agility to the business. The ability to harness more data from more sources in less time will directly lead to a smarter organization making better business decisions, faster. The newfound capabilities to collect, store, process, analyze, and visualize high volumes of a wide variety of data, drive value in many ways.

There is a comparison to be made to the traditional enterprise data warehouse. Legacy data warehouses have served well (and still do), but most of them lack the scalability, flexibility, and cost profile of a data lake. A traditional data warehouse works well for strictly defined analytics and use cases, and can support high performance and high concurrency for tightly structured data sets. Yet as an organization moves to save (and potentially analyze) all enterprise data, traditional data warehouses can be complemented by data lakes. Data lakes can more cost-efficiently perform ancillary functions like extract, transform, and load (ETL), and flexibly support a wider range of analytics approaches. ESG research shows that 36% of organizations would like to offload and/or optimize their data warehouse with Hadoop.[1] The most common scenario then is likely to be complementary between data warehouses and data lakes. Either way, data lake architectures are proving themselves to be more feasible in supporting fluid needs at scale by utilizing more economical resources. Again, a big part of the value of a data lake approach is that much enterprise data may be idle or of unknown utility, but the potential must be amassed first if it is to be leveraged later. Every organization will need to be able to accommodate a continuous series of new questions from the various lines of business. Machine learning models will be constantly built, trained, and updated to focus on different topics and changing data over time. One might think of the data lake as making all of an organization's data readily available for analytics.

Data lakes are used to capture new insights for almost any line of business operation, including:

- Customer interactions - Sales, marketing, billing, and support are all core elements of almost any business. It is a truism to say that the better you understand your customers, the happier they will be (assuming you act on that information). Every contact with a customer is an opportunity to learn about his motivations and preferred style of engagement. A data lake will help in a number of ways here. Critically, a data lake can break down the silos between types of customer interactions. Combining customer data from a customer relationship management (CRM) platform with a marketing platform that includes buying history and incident tickets helps show the range of possible outcomes. Further, it empowers business analysts to diagnose the causes of those outcomes. Data science can therefore identify how to market to specific cohorts of customers most effectively, which groups are likely to be profitable, what events are likely to cause churn, and what promotions or rewards will engender loyalty (and at what cost). Predictive analytics on a time-series of events can model and demonstrate these relationships over time. Graph analytics can help identify brand influencers for social marketing. Looking at any one type of interaction cannot give a full or nuanced picture, but improving understanding together allows for mass personalization, and subsequently maximization of customer value.

- Research and development - Building better products and services, and getting them to market faster is critical to successful competition in a market. Sometimes this is accomplished via incremental improvements, sometimes via radical innovation, but either way it requires an in-depth understanding of how the offering is meant to perform and how to deliver that functionality most effectively. Machine learning on data lakes can be leveraged to analyze data during product development and over a lifetime of use in the field. For example, a new pharmaceutical drug may be discovered through extensive modeling of physiological processes and biochemical interactions. This may involve extremely large data sets around DNA proteins to determine applicability to a

[1] Source: ESG Research Report, *Enterprise Big Data, Business Intelligence, and Analytics Trends: Redux*, July 2016. All ESG research references and charts in this white paper have been taken from this report.

particular class of diseases and types of patients. Often, this is an iterative process of hypothesis, test, refine, test again, etc., before a discovery is made and validated, much less proven safe for regulatory approval. Longitudinal studies can amass extreme amounts of data that may be needed for further analysis in the future. A data lake can enable and accelerate this work, bringing new drugs to market faster, which not only increases profits but can potentially save lives, too.

- Operational processes and events - Business success is often a matter of driving efficiency everywhere, along with smoothing or eliminating problems and risk. While many businesses are simply reactive to issues as they arise, smarter organizations measure and monitor as many activities as possible, and look for ways to continuously improve. Supply chain efficiency is a well-known example, with six sigma and just-in-time production as illustrative strategies. The Internet of Things (IoT) era is now introducing many more ways to pervasively collect data on processes like manufacturing, with instrumentation and automation of activities like assembly and quality control. Each tool and step can generate data that may be used to find idle resources, wastage, defects, and breakdowns. A data lake can aggregate all this machine-generated IoT data, and analytics will lead to opportunities for enhancement throughout the process, reducing operational costs and increasing quality.

What all of these scenarios have in common is that a data lake can offer more value than traditional approaches. The absolute scale of storage for all data enables a comprehensive viewpoint. The flexibility of sources and variety of data enriches the analysis. The flexibility of distributed resources like processors and memory keeps down the cost of analysis at scale, while still enabling high performance. The ability to work with unstructured data improves agility, allowing quick analysis in many different ways. Ad hoc and exploratory analytics enables creativity and quick validation of ideas for analytics. Not least, analytics can be robustly operationalized for production in place, directly within the data lake, not requiring export to another system. Together all of these qualities make the modern data lake ideal for delivering new business value.

The Misconceptions and Challenges of Data Lakes

One of the biggest issues with the data lake is the confusion over what exactly it is. Simply loading raw data into Hadoop is not the same as building the modern data lake. Without some organizing frameworks, this approach threatens to become yet another data silo, or the dreaded “data swamp,” where investments are seen as failing to realize the promised value. To be successful, a data lake needs to have defined mechanisms to collect, store, catalog, secure, and analyze data, plus aid collaboration and sharing of findings.

In fact, the qualities needed for a successful big data initiative are startlingly similar to traditional enterprise IT priorities. Figure 2 shows security, performance, reliability, economics, availability, integration, scalability, and agility as common requirements for evaluating technologies for big data and analytics. To be fully accepted by the enterprise, a data lake then must also support these qualities, no less so than a traditional data warehouse.

Here is where the differences between core Hadoop and a proper data lake become clear. Security is often added as a distinct bolt-on function (like Apache Ranger or Sentry). Performance is tied to the hardware capabilities and/or cleverness of the analytics algorithm. Reliability requires more than just three redundant copies of data on servers. Return on investment (ROI) and total cost of ownership (TCO) will include the whole technology stack and staff, not just “free” open source Hadoop software. Availability needs to span beyond a single data center. Integration ties beyond Hadoop to the whole analytics and application ecosystem. Scalability needs to be elastic. The architecture needs to decouple storage from compute resources, so you can scale resources independently of one another. Elasticity, scalability, and decoupled architecture are key enablers for data lakes, and fundamental benefits of cloud deployments. Agility is dependent on how easily manageable and changeable the environment is for both users and administrators. Hadoop alone does not solve all these issues, but a data lake must address them to be successful.

Figure 2. Important Considerations for Big Data and Analytics Solutions

Source: Enterprise Strategy Group, 2017

Interestingly, a number of these issues also stem from the complexity of building an entirely new data platform architecture on-premises. That is to say, infrastructure and operations processes represent most of the challenge in getting up to speed with a data lake in a typical company's data center. Investing the time and effort to build a big data environment tends to be more expensive and slower. Even just the typical hardware procurement and provisioning process—including selection, evaluation, purchasing, receiving, testing, and integration—is an issue. This can cause significant delay and upfront capital expense. There will be a big learning curve for staff, too, with false starts and lessons learned the hard way. Indeed, 35% of enterprises say it will take 7-12 months before they see value from a big data initiative, and 42% say it could be more than a year. Then patching and maintaining all the hardware can cause ongoing headaches. Many are looking for a better way. Managed services (often cloud services) significantly accelerate the learning curve and eliminate a lot of the undifferentiated heavy lifting of building and running a data lake environment.

Why the Cloud Makes Data Lakes Better

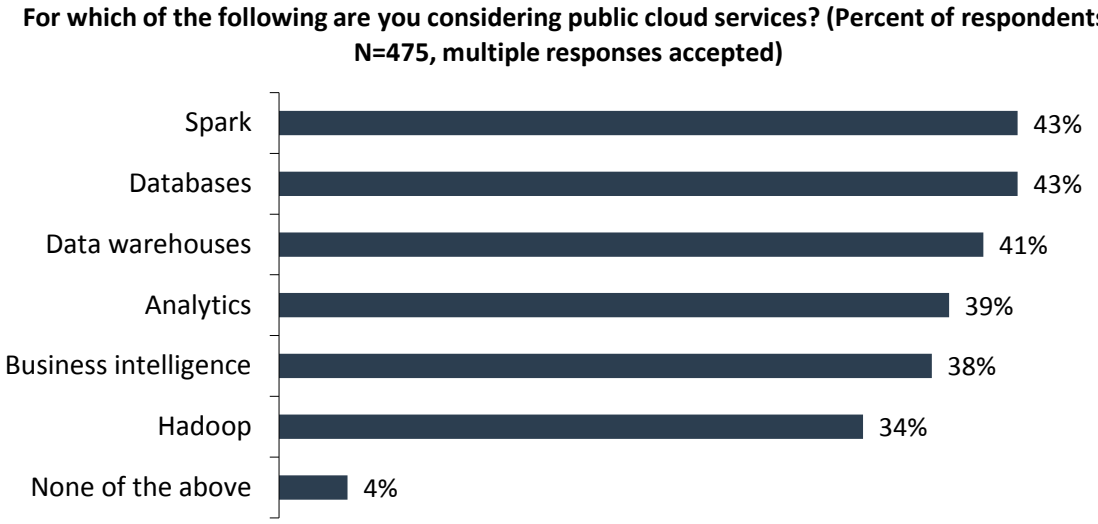
ESG research shows roughly 35-45% of organizations are actively considering cloud for functions like Hadoop, Spark, data bases, data warehouses, and analytics applications. It makes sense to build your data lake in the cloud for a number of reasons. Some of the key benefits include:

- Pervasive security - A cloud service provider incorporates all the aggregated knowledge and best practices of thousands of organizations, learning from each customer's requirements.
- Performance and scalability - Cloud providers offer practically infinite resources for scale-out performance, and a wide selection of configurations for memory, processors, and storage.

- **Reliability and availability** - Cloud providers have developed many layers of redundancy throughout the entire technology stack, and perfected processes to avoid any interruption of service, even spanning geographic zones.
- **Economics** - Cloud providers enjoy massive economies of scale, and can offer resources and management of the same data for far less than most businesses could do on their own.
- **Integration** - Cloud providers have worked hard to offer and link together a wide range of services around analytics and applications, and made these often "one-click" compatible.
- **Agility** - Cloud users are unhampered by the burdens of procurement and management of resources that face a typical enterprise, and can adapt quickly to changing demands and enhancements.

Accordingly, many organizations are now looking toward cloud-based solutions. ESG research found 34% of survey respondents considering cloud as the primary deployment model for Hadoop, 39% for analytics, 41% for data warehouse, 43% for databases, and 43% for Spark (see Figure 3). As enterprises progress on their journey to the cloud, data management and analytics functions will inevitably follow.

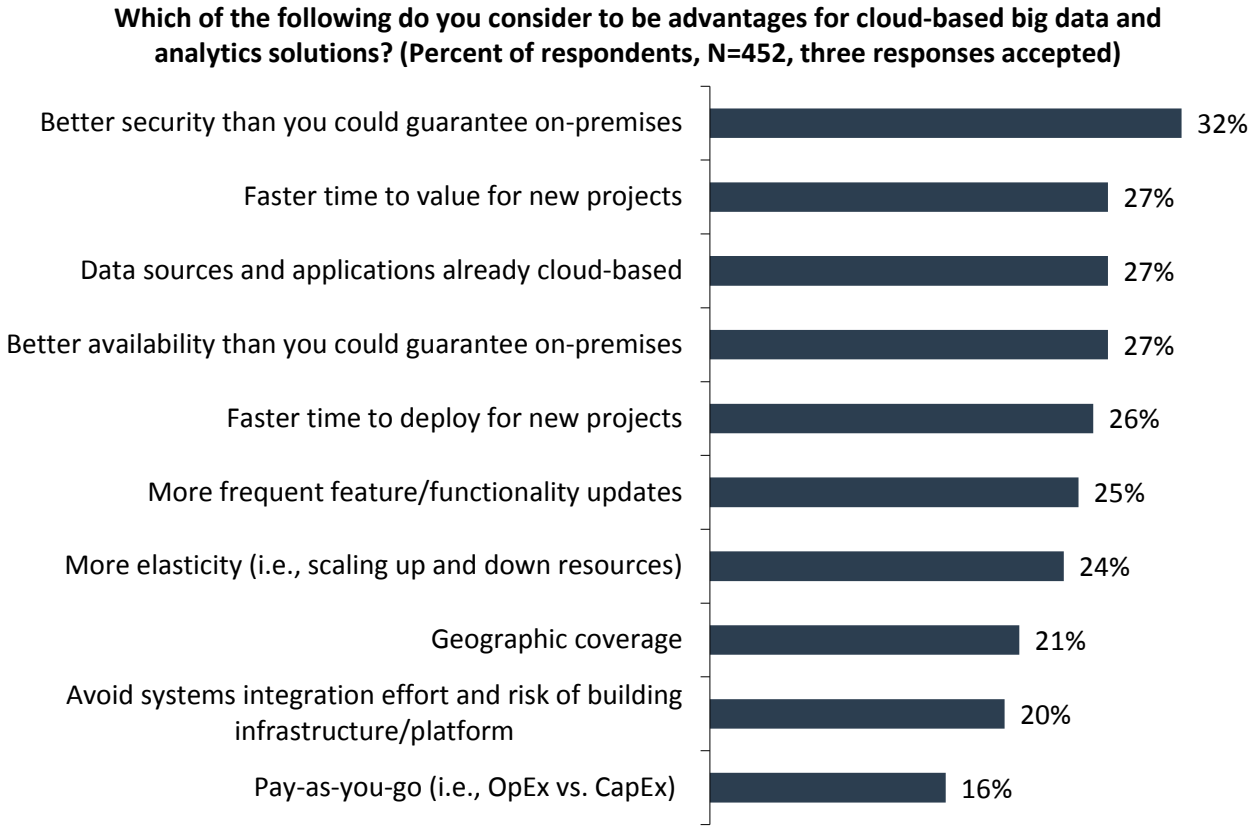
Figure 3. Migration to Cloud for Data Management Services



Source: Enterprise Strategy Group, 2017

These environmental qualities lead directly into realizing some of the top perceived advantages of cloud-based data lakes, as shown in Figure 4. These include better security, faster time to value, easy utilization of data in place in the clouds, better availability, faster deployment, more current software, more elasticity, more geographic coverage, reduced risk, and costs linked to actual utilization. Anyone wanting to run an enterprise data lake should be interested in these outcomes, and they will be easily justified to the business decision makers, not only the IT department. Accordingly, a cloud service provider that delivers against these promises is extremely well suited for a modern data lake.

Figure 4. Advantages of Cloud-based Big Data



Source: Enterprise Strategy Group, 2017

Selecting the Best Cloud-based Data Lake Ecosystem

As already discussed, a number of fundamental building blocks are needed for sophisticated big data applications. Architects should look for ways to utilize the best set of resources anywhere, and have an open mind before choosing any particular distribution, vendor, or service provider. As part of designing a data lake, it is important to identify services to make the desired architecture approach possible and practical for the enterprise, business users, and data scientists alike.

While much marketing may make this exercise sound trivially easy, before making a decision, one should think about how to bring it all together and create that ideal operating data lake environment. This is still a very complex technology stack, but simplifying the number of components and vendors required will help. The whole data lake should be much more than the sum of its parts. Revisiting the architectural components may be best done with some concrete examples of each. In each case, Amazon Web Services (AWS) offerings illustrate the points perfectly. These include:

- Storage - If nothing else, data lake storage needs to be capable of holding extreme amounts of structured and unstructured data. While Hadoop's HDFS is capable, cloud-based object storage may be a better choice for data redundancy distributed not only across nodes but also across facilities—without 3x the cost for the resources. AWS offers both Amazon Simple Storage Service (S3) for reliable, secure, and scalable object storage and Amazon Glacier, which has similar qualities for extremely low-cost, long-term archiving and backup. These services can clearly scale for virtually any data type, with minimal administrative overhead, and further offer choice of cost profiles based on how actively the data might be used. Reliability, security, and analytics improve the storage layer itself. And the storage serves as the foundation for the analytics clusters and data engines above it.

- **Compute** - Different analytics workloads will have different compute resource requirements. For example, streaming analytics will need high throughput, while batch may be processor-intensive. Apache Spark can require a lot of memory, while AI may work best on GPUs. Here, AWS again offers significant flexibility compared to other cloud providers as well as on-premises Hadoop, which ties storage directly to compute in each node. Options for compute-optimized, memory-optimized, or storage-optimized instances avoid the "one-size-fits-all (sort of)" problem. Again, decoupling compute from storage enables the flexibility to utilize specialized resources efficiently.

- **Analytics** - A data lake virtue is how it enables the same data to be analyzed in many different ways for many different use cases. For AWS, all common analytics approaches are covered, including Amazon Redshift as a data warehouse; Amazon Athena for SQL querying on demand; Amazon EMR for running popular open source frameworks such as Hadoop, Spark, Presto, Flink, and HBase; Amazon QuickSight for business intelligence; and Amazon Elasticsearch Service for logs and text. There is no need for data migration to different operating environments, or the accompanying overhead, costs, effort, or delays.

- **Databases** - Not all data lake data is unstructured. Often, it makes sense to have tighter organization for both transaction and analytics processing. Amazon Relational Database Service (RDS) lets people choose from popular commercial and open source databases, including Amazon Aurora for MySQL. Additional services are Amazon DynamoDB for NoSQL models like key-value and document, and Amazon ElastiCache for in-memory performance with Redis and Memcached. Again, this provides the versatility to meet the needs of many data lake applications.

- **Real-time streaming processing** - Not all data is simply stored in the data lake and analyzed later. Often, there is a need to collect, store, process, and even analyze real-time data in motion. AWS has Amazon Kinesis, a platform for streaming data on AWS, offering powerful services to collect, store, and analyze streaming data as well as the ability to build custom streaming data applications for specialized needs.

- **Artificial intelligence** - Increasingly, artificial intelligence and machine learning are becoming popular tools for building smart applications such as predictive analytics and deep learning. To make it more accessible, Amazon Machine Learning abstracts away from the algorithms with wizards, APIs, and guidance. Newer AI services include Amazon Polly for text-to-speech, Amazon Lex for natural language processing and conversational bots, and Amazon Rekognition for image identification and classification.

- **Security services** - As shown, security, privacy, and governance are essential elements for sensitive data to be trusted to a cloud data lake. AWS has a number of ready-to-roll services here, including AWS Identity and Access Management (IAM) for roles, Amazon Cloudwatch to monitor and respond to events, AWS Cloudtrail for audits and logging, AWS Key Management Service (KMS) to create and control the encryption keys used to encrypt your data, a security token service, and Amazon Cognito for mobile/web user account management, authentication, and user data sync. AWS environments are continuously and strictly audited for certifications such as ISO 27001, FedRAMP, DoD SRG, and PCI DSS. AWS also has assurance programs to help customers prove compliance against 20+ standards, including HIPAA, FISMA, and more—all from one vendor.

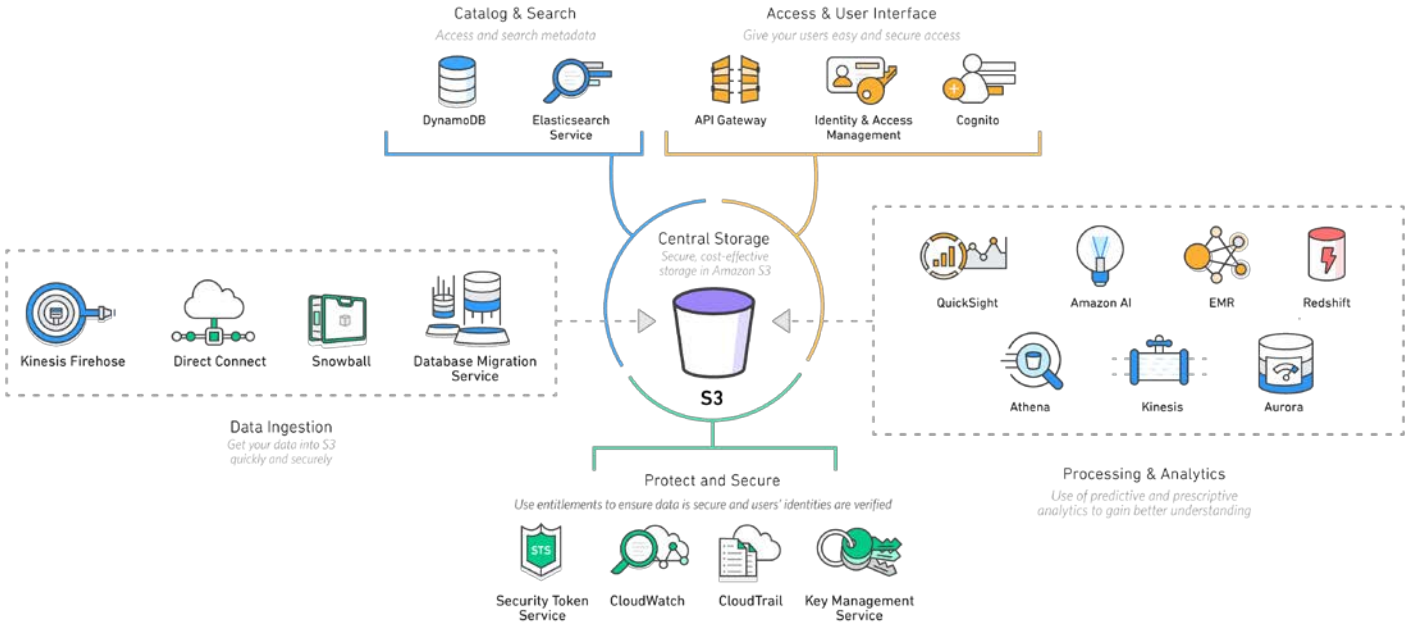
- **Data management services -** As data is used in different platforms, ETL is an important function to ensure that it is moved and understood properly. AWS is introducing AWS Glue as an ETL engine to easily understand data sources, prepare the data, and load it reliably to data stores.

- **Application services** - While the data lake can be an invaluable resource in its own right, it really comes alive when integrated with higher-level applications. AWS has fully capable utilities for IoT use cases, for mobile applications,

and for API calls to anything else. AWS Lambda enables users to run code without provisioning or managing servers on demand, and with very granular consumption billing.

A basic premise of the data lake is adaptability to a wide range of analytics and analytics-oriented applications and users, and clearly AWS has an enormous range of services to match any need (see Figure 5). Many engines are available for many specific analytics and data platform functions. And all the additional enterprise needs are covered with services like security, access control, and compliance frameworks and utilities.

Figure 5. Amazon Web Services Data Lake Framework



Source: Enterprise Strategy Group, 2017

Contrary to some opinions, there is little "lock-in" concern either, as AWS approaches many data lake components from an ecosystem perspective. Anyone can easily bring their choice of software distribution for big data tools like Hadoop, Spark, and Kafka, or bring and tie in traditional and newer databases, data warehouses, and BI packages as desired. An enterprise will find many easy ways to integrate existing tools with an AWS data lake. Capabilities like these mean the difference between success and frustration with data lakes, and all the "advantages of cloud" noted above are in effect with a cloud provider like Amazon Web Services.

The Bigger Truth

The modern data lake is foundational for the modern enterprise. If set up properly, a data lake will draw people to naturally gravitate there with ideas and come away with useful insights. Much of the discussion above has been about why people should care about data lakes and how better approaches yield better outcomes. The pillars of a data lake include scalable and durable storage of data, mechanisms to collect and organize that data, and tools to process and analyze the data and share the findings. This is a critical point: A data lake, like the wider world of big data, is as much about architecture as it is about processes. With the right tools and best practices, an organization can use all its data, making it accessible to more users and fueling better business decisions. Many paths will deliver value, suitable for any organization, and the ongoing flexibility is a big part of the value.

Just as a cloud-based data lake is versatile from an architectural perspective, it serves many corporate audiences, including IT applications, infrastructure, and operations teams, and even line of business groups. This technology won't be successful if it remains the rarefied realm of only a few data scientists. With adequate security in place, data lakes should be made easily accessible to a wide range of users, and their efforts in implementing and supporting core applications, for any line of business or function.

In any use case, the technical decision makers defining and designing a data lake need to be able to articulate their choices. This paper has outlined many of the common considerations, challenges, and requirements for building an enterprise-ready operational data lake. Along the way, it has also perhaps shown AWS as offering the broadest and deepest set of managed services for big data and analytics. All these services must be tightly integrated, without compromising the ability to pick and choose what suits your specific needs for each component of the data lake. The built-in advantages of cloud for cost, scale, performance, ease of use, and security should clearly be recognized for their impact on the overall data lake initiative and outcomes.

AWS is helping businesses of all sizes build productive data lakes. Customers such as Nasdaq, FINRA, Capital One, and GE Oil & Gas have proven AWS offers an enterprise-class solution. Still, everyone might not be ready to make a transition from an on-premises approach to a cloud-based data lake. Whether you start your journey to the data lake with a hybrid architecture as a bridge or choose to go "all in" on cloud now, AWS is very well positioned to help you on your way.

Enterprise Strategy Group is an IT analyst, research, validation, and strategy firm that provides actionable insight and intelligence to the global IT community.

 www.esg-global.com

 contact@esg-global.com

 P. 508.482.0188