Data: A Lenovo Solutions Perspective
Positioning Information

Digital transformation is impacting all areas of IT and business, having profound effects. One key recurring theme is the increasing need to gain insights from exponentially growing volumes of data. This data can be structured or unstructured and come from sources as diverse as traditional IT systems, social media or sensors embedded in a smart infrastructure. Whatever the source, businesses are increasingly reliant on mining its content, often in near real time. It is for this reason that data is often being compared to oil as a core natural resource fueling our lives and businesses.

Ingesting, storing and analyzing the data is placing unprecedented strains on the underlying IT systems. Its pressure is restructuring IT from marketplace dynamics to the way customers procure, deploy and use IT systems.

This white paper provides an overview of these changes and illustrates the different approaches being adopted. The paper is structured in five sections:

- An overview of how data is evolving into this new world
- Data centricity and its impact on all aspects of IT
- The key changes in the marketplace
- Analytics. This is perhaps the area of greatest change. How is this evolving to meet the business requirements?
- Emerging technologies to watch

It is the first in a series of white papers and is intended to set the scene and review the changes from a high level. Subsequent papers will look in more detail at specific areas of data or software ecosystems and how they deliver the needs of business within this context.

Data Evolution

The increasing digitization of the world is often called digital transformation. Naturally data is at the heart of this, effectively the lifeblood of the change. Today companies are leveraging data to improve customer experience, open new markets, make employees and processes more productive or create new methods of business. All of this to build competitive advantage. This is a very different environment from that which set our expectations of data in traditional IT.

Data is evolving:

- Growth is exponential
- New types of data are coming from new sources
- We can no longer assume classic data lifecycles

These aspects are outlined below and underpin the new paradigms and approaches discussed later in the paper.

Growth

Data growth is perhaps the most commonly discussed change in our industry. However you measure it, the growth is enormous and never ending. IDC and Seagate publish trends for the global datasphere (data stored in all formats) and their prediction is it will grow from 45ZB to 17ZB between 2019 and 2025[1]. A Zettabyte is hard to imagine, but it is 1000 million TB. Perhaps more simply in these six years almost 4x more data will be created than in history to date.

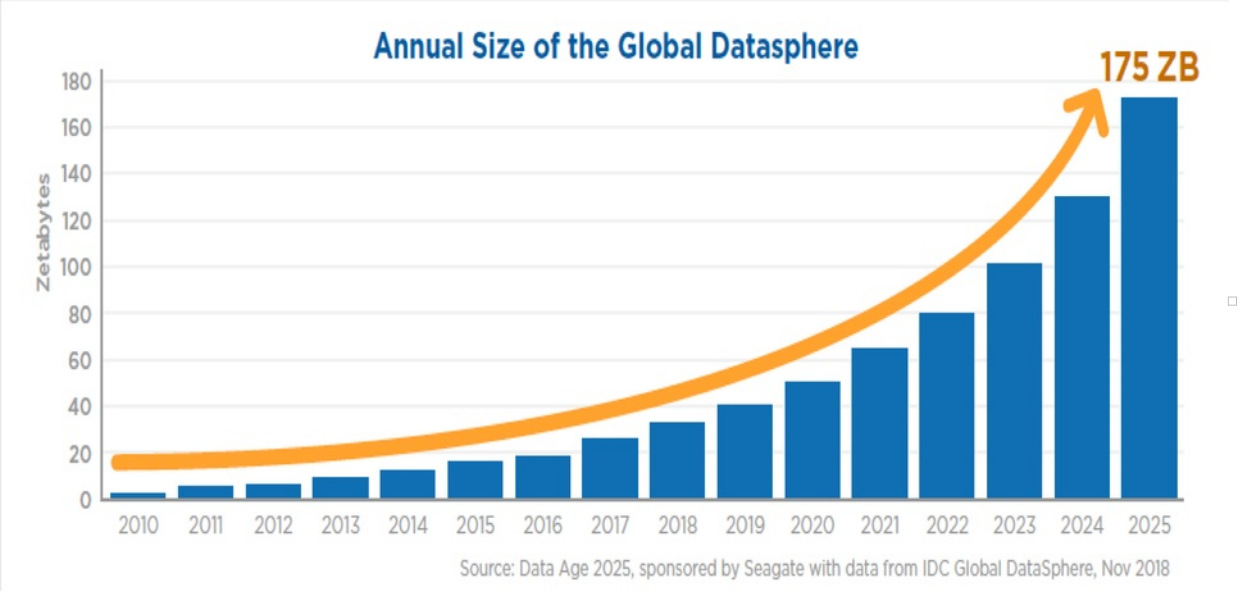

Figure 1. Annual Size of the Global Datasphere

So where is all this data coming from and why is it being stored? Modern business is always looking for new ways to differentiate and gain an edge. This can come from driving efficiency in business processes or greater personalization of customer experience. All this requires analysis of ever larger amounts of data, such as records of customer interactions, data from social media or the billions of smart devices being deployed. Customers are expecting this new personalized approach, which is forming the new baseline.

The next question is how is all of this data managed and how do companies ensure compliance with increasing data regulations? Whilst much of the new data is being generated at the edge from smart devices, it is controlled and managed from the core (either corporate data center or public cloud). This in itself presents unique challenges.

Of course, the other dimension is how will it be processed and where? This is discussed later.

New Types and Sources

Wherever it is stored much of this data growth is from outside the data center. Sources can be smart end user devices, social media, remote systems or devices at the edge. Instead of classic structured data, this is often in new formats such as video, images, GPS co-ordinates etc. This raises several new challenges.

How should all this data be stored? A common answer is a large object store or data lake. This certainly can provide a useful repository, but then it needs to be analyzed. Are different structures useful for this? The answer is perhaps, and it is giving rise to many new data management tools such as non-relational databases, graph or document databases, time series databases, etc. These all have a place, but within the core (i.e. data center or cloud) the classic relational database will still be the dominant approach for many years.

Whilst the core will control data management, data will be stored across the enterprise. The location will significantly affect how and where it is processed. This is outlined in the next section.

One new question raised by much of this data is whether or not it is accurate? Classic enterprise data was carefully controlled and managed and its veracity was certain. Data captured from beyond the data center, be it sensor readings or comments on social media does not have the same pedigree. One of the first tasks before using it is to verify the data – usually correlating it with other related information or known data points. This in itself adds significantly to the processing required to gain insights from the data.

Lifecycle and Lifetime

All of these new sources of data are changing lifecycles and lifetimes. In 2016 15% of the datasphere was real time, by 2025 it will be almost 25%. This means that by 2025 there will be more real time data than exists in total today. Such data has a very different lifecycle. It needs processing immediately and then often is no longer of value. In many cases organizations just keep the last 'n' minutes or hours of data. No additional needs to be stored. Equally it does not need to be backed up or archived. This trend extends well beyond real time data. Much of the customer-related data, especially that harvested from social media, has a limited lifetime during which it is valuable.

Cost is another key aspect. This is a classic balancing act. The more data stored about a business process or a customer, the better decisions can be taken. However, all this comes at a cost and there is an inflection point where more data or a longer history does not significantly improve the quality of the outcome. Judging this balance is another new discipline relating to data.

Of course, business value is not the only determinant of data lifetime. Increasingly, privacy and compliance regulation is limiting what can be stored and where. Customers can withdraw consent for businesses to store their data (or anything identifying them) at any point under new regulatory regimes such as GDPR (General Data Protection Regulation) with the European Union. On the other side, audit and compliance regulation can require business to keep data securely for many years. These highly varied lifetimes are another complexity which needs to be managed and relates closely to the data source and its business use.

The data lifecycle is also much more varied. This is referring to where it may be stored, usually based on the service levels required. In today’s agile business environment, data becomes cold much more quickly. Critical data one minute may no longer be needed the next. In many cases this is driving far more tiering of data both within the data center and offsite. Equally, new challenges such as ransomware are bringing back concepts of offsite and even offline storage for key data. In general, business now needs to have lifecycles and compliance regimes tailored to specific classes of data across the enterprise. No one size fits all. This in itself is impacting how and where data is stored and processed which will be discussed in the next section.

Location Location Location

There has long been a saying about property values that it's all about location, location, location. Increasingly the same is true of data. Where your data is captured, sourced and stored is a key factor in determining how and where it is processed. IT architectures are evolving to center on data. It is also one of the core factors in the choice between public, private and hybrid cloud deployments – where is the center of gravity of the data?

This section outlines the main principles behind this, which are fueling the cloud and driving the dramatic growth of computing at the edge.

Data Centricity

Traditionally IT has been application centric both physically and logically. Infrastructure was provisioned around delivering a specific application or function. Physically, this manifests itself as one or more servers with storage attached. Logically, only the data required by the specific application or function was stored and the structure within it was what was efficient for that process. As we look across the enterprise this leads to a complex web of systems. Data is physically and logically siloed. It can be hard to integrate, and parts of the data may be duplicated. The structure is very inflexible. As processes change, new approaches needed to address this complexity leads to the inertia and costs of change.

The business need for ever faster insights places the data centerstage. It becomes the core of the IT system and applications are brought to it. For evolving cloud native or microservice applications, an analogy would be installing apps on your phone. After a short time, you could uninstall one and add another, with all using the data in your phone. More traditionally we are seeing increasing application function moving into the data storage layers. Hadoop flipped the paradigm for big data repositories and now increasingly analytics are embedded in the Spark framework or brought into traditional databases through AI stored procedures, etc.

Data Gravity and the Edge

There is a physical manifestation of this data centricity as well. In the modern virtualized world, applications can move between servers or networks and be reconfigured almost instantaneously. However, data exists somewhere on physical media. It costs time and money to move it and so it has a natural tendency to stay where it is. This could be seen as gravity holding a weight in place. As volumes increase the effect of gravity is only more pronounced.

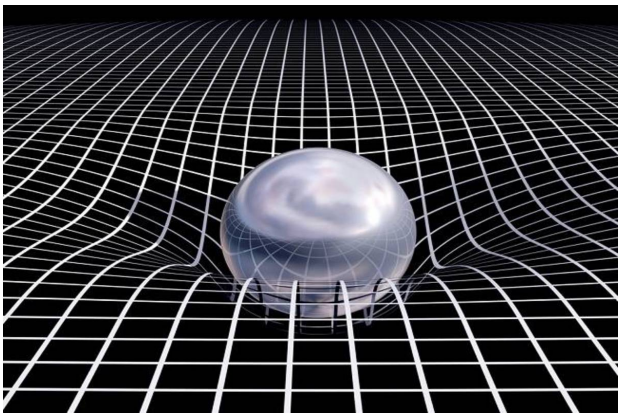

Figure 2. Data Gravity

The effect is that data is increasingly processed at or close to the point of capture. As was discussed in the previous section, this will increasingly be outside of the data center. Such local processing is essential for latency sensitive applications (e.g., self-driving cars or controlling a manufacturing line) and is one of the fundamental factors driving the exponential growth in the “edge”. It is either impossible or uneconomic to move all the data to the data center before processing. In fact, depending on its lifetime, it may never be moved there.

The Anchors

Compliance with data sovereignty and privacy laws is further limiting data movement. They can be seen as “anchors” ever more firmly holding the data in place. Such laws are often a political matter. Instead of tariffs for goods there are restrictions on data. In Australia, all health records must be stored inside the country. Similarly, Russia requires personal data to be stored domestically. In some other countries, access to data from outside the country can be controlled, especially when it is seen to impact the government or national attitudes.

Data privacy can also impact movement, certainly outside the enterprise. Regulations such as GDPR place tight legal limits on the time to notify a user of a data breach (72 hours) or in which all data identifying a user must be deleted following a formal request. This creates its own pressures and means enterprises need to keep control of certain classes of data within their IT infrastructure. Similarly the classic application-centric data structures can lead to challenges even identifying everything stored related to an individual.

Compliance to these rules is a catalyst to data centricity. Managing this ever more complex landscape is one of the factors driving the evolving discipline or DataOps discussed in the final section.

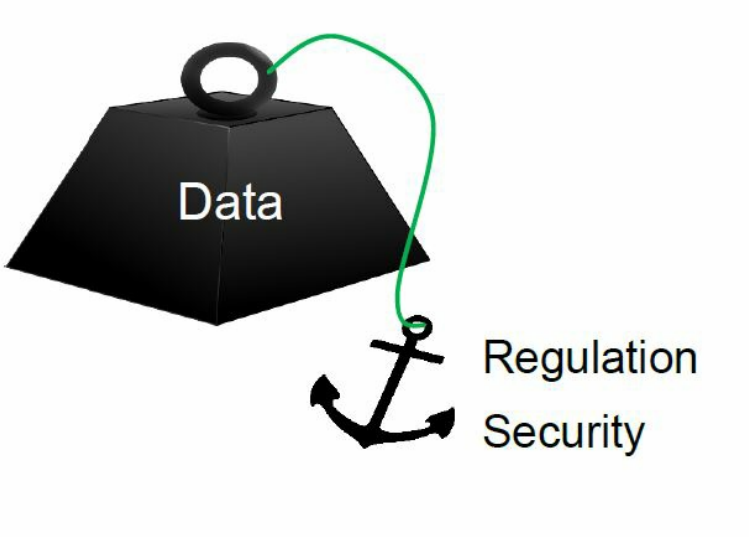

Figure 3. Data Anchors

Data Location

Returning to the introduction of this section, data location is becoming increasingly important. If we combine data centricity and gravity, then new IT infrastructure will increasingly develop around data. This is one of the key factors determining where processing is now being performed. If data is stored at the edge then, at least initially, it is likely to be processed there.

Similarly, the choice between enterprise data centers and public cloud deployment often revolves around data. If your data is stored in the public cloud, it is likely to be processed there. Alternatively, if the data is stored in the data center that is where it will be processed.

The common thread is that the core (be that an inhouse data center or public cloud) will be the center for command and control of your data combined with deeper analytics.

The Changing Marketplace for Data Offerings

The evolution of data and impact of its location discussed in the previous sections is fundamentally reshaping the marketplace when combined with the business need for ever deeper and more agile insights. Whilst the impacts are wide ranging some key trends are appearing. This section discusses the key aspects shaping the response to meet customer needs.

Big Data → Data

Originally Big Data referred to Hadoop-based systems which formed large data repositories on which some analytics such as MapReduce were performed. This was the start of data centricity discussed in the previous section. With the exponential growth in data being stored across the enterprise, it can be argued that all data is becoming “big”. Big Data systems are evolving along with the marketplace.

Increasingly there is a need to embed ever more complex analytics into data storage systems and with Spark 3.0, this framework is moving strongly to focus on analytics rather than simply data storage in HDFS (Hadoop Distributed FileSystem). Increasingly classic query languages such as SQL are being embedded along with the de facto standards for AI/ML (Artificial Intelligence/Machine Learning). These include frameworks such as TensorFlow or OpenBLAS. This much greater analytics capability is stretching current processing systems and so Spark 3.0 also includes significant GPU acceleration though NVIDIA’s RAPIDS framework.

Convergence and Federation

Until recently data management software (both structured and unstructured) have been separated into specific silos: SQL based relational databases (RDBMS), No SQL and optimized (e.g. graph) databases, Hadoop (or Big Data) and analytics. As with other application centric approaches, the data repository was selected to match the processing need and more often than not was an RDBMS.

Moving into the data centric world is driving a convergence of database, big data and analytics. All offerings are moving to support structured and unstructured data in various forms, classic SQL-based queries and deeper AI-based analytics. The first step in this could be seen by the emergence of HTAP (Hybrid Transactional/Analytic Processing) a few years ago when the previously separated OLTP (OnLine Transaction Processing) and OLAP (OnLine Analytic Processing) began to merge.

The business driver is to provide access to increasing amounts and variety of data and use the appropriate analytics to gain insights ever close to real time. Two examples of this convergence following very different paradigms are:

- Spark 3.0 as mentioned above. Here, the Hadoop/Spark framework which is becoming increasingly popular as a data lake is embedding complex accelerated analytics. At the same time it is incorporating ubiquitous interfaces, such as SQL to be able to exploit that application base and skills.
- Microsoft SQL Server has taken a different approach. For several years increasing levels of analytics have been embedded in SQL Server, started with Machine Learning (ML) stored procedures in SQL Server 2017. With the latest version, Microsoft adds the Big Data Cluster (BDC) to its SQL Server product. This provides HDFS as the underlying layer to support both SQL processing and Spark functions. Then it uses its Polybase technology to allow it to query many external data sources. Effectively this builds analytics in a federated manner, allowing the data to stay in its original location, respecting data gravity and also reducing copy management issues.

Economies of Scale and aaS

Economies of scale can be seen driving acquisitions across the IT industry. Increasingly companies are merging or being acquired to allow vendors to offer a unified portfolio to customers. Such scale is also required to fund the innovation that is increasingly required to keep pace with the ever greater function and customer business need.

This consolidation can be seen clearly in the data management software marketplace. If we look at Hadoop/Spark, mergers and acquisitions have reduced the four main vendors to one, Cloudera. Equally if you look at the operational database market, then increasingly this is dominated by the major cloud vendors. Microsoft is the only major software vendor to grow its revenue and market share powered by its Azure public cloud. Gartner's revenue estimates would show only Microsoft and Amazon topped $1B in database software revenue for 2019. These clouds now offer a different type of "one stop shop". A broad range of organic data and analytics offerings and also major ISVs from across the industry. Effectively multi-offering instead of multi-cloud. This data marketplace also provides a platform for specialized offerings such as graph DBs optimized for specific processing.

The are two other factors driving the remarkable acceleration of cloud-based data offerings:

- Data as a Service is increasingly the paradigm being adopted by enterprises. Gartner predicts that 75% of operational database software revenue will be DBaaS by 2023. This is a phenomenal growth from 25% in 2019 and is fueling the dominance of these vendors.
- Data will increasingly be stored in the cloud. IDC predicts that by 2025 over 50% of the entire Dataspace will be in the cloud. As can be seen from the graph below this is mainly a change in paradigm from data being stored in consumer devices. Enterprise data centers stay reasonable constant. In the data centric era, as data is increasingly stored in the cloud, the processing moves there as well.

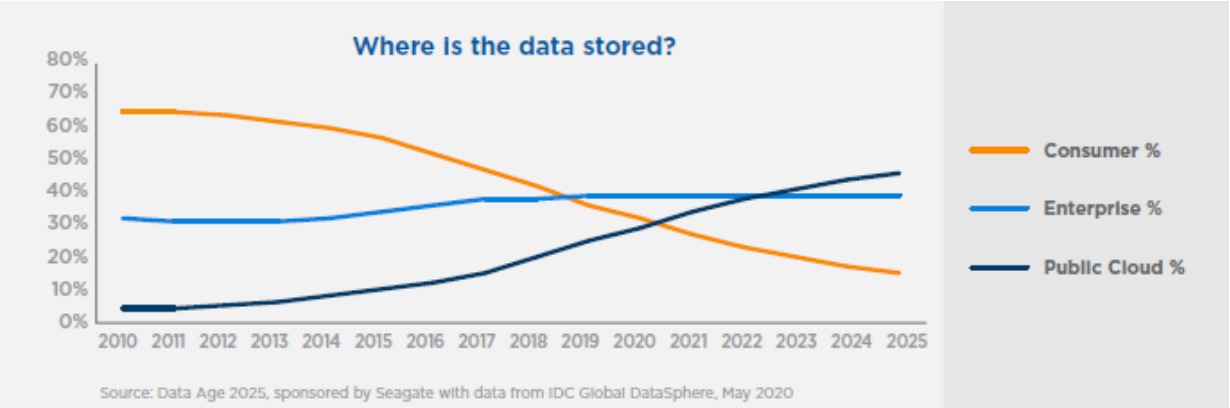

Figure 4. Where Data is Stored

The Rise of Open Source and New Approaches

The other major trend alongside the consolidation of the major vendors is the increasing growth of both open source and specialized vendors. Gartner predicts that by 2022, 25% of database software revenue will come from open source-based offerings[2], such as MySQL or EnterpriseDB. These are the only vendors growing except for Microsoft and the major cloud providers. This illustrates a growing acceptance of open source offerings for data within the enterprise, especially in specific industries. Increasingly the collaborative innovation and lack of license fees is proving an attractive approach. Of course the Apache Spark framework is the best known example.

Alongside the growth of open source is the increased presence of specialized technologies. Optimized databases such as graph databases have been in the market for many years, but the large cloud vendors are effectively providing a “marketplace” which can act as a platform for many start-ups as well as the CSP's own products. These specialized offerings are usually related to a type of data or a method of processing. Some examples include:

- Document databases (e.g., Amazon DynamoDB or DataStax Enterprise)
- Graph processing databases (e.g., Neo4j or Amazon Neptune)
- Time series databases (e.g., InfluxDB or TimescaleDB)
- Real-time databases (e.g., Google Firebase)

Analytics – Creating Essential Insights

With all this data becoming available, the quest for extracting useful insights from it is top of mind for everyone. There is a growing expectation that all one needs to do is point a suitable analytics tool at all that data and insights will come pouring out. The reality, not surprisingly, can be very different from this somewhat simplistic view.

This section describes broad categories of analytics methodologies that need to be selected based on the type of insights being sought. Designing and implementing the appropriate analytics tool-kit requires handling data from an end-to-end perspective, starting with ingesting data from the various sources, storing them for processing, preparing specific datasets for training and evaluating models that can be used for the targeted analytics use-case, and finally, using the trained model with real-life data to drive decisions. The section concludes with a discussion of incorporating AI techniques in all manner of data processing and management scenarios. AI is indeed being required everywhere, but a fruitful use of AI techniques requires following a methodical approach to prepare the requisite data and set up the optimized models that can help address the targeted business challenges.

Types of Analytics

The growing importance of data analytics is manifested in its adoption across many different parts of an organization, each with their specific needs and desired outcomes. Accordingly, analytics techniques can be grouped into categories that are distinguished based on the desired outcome.

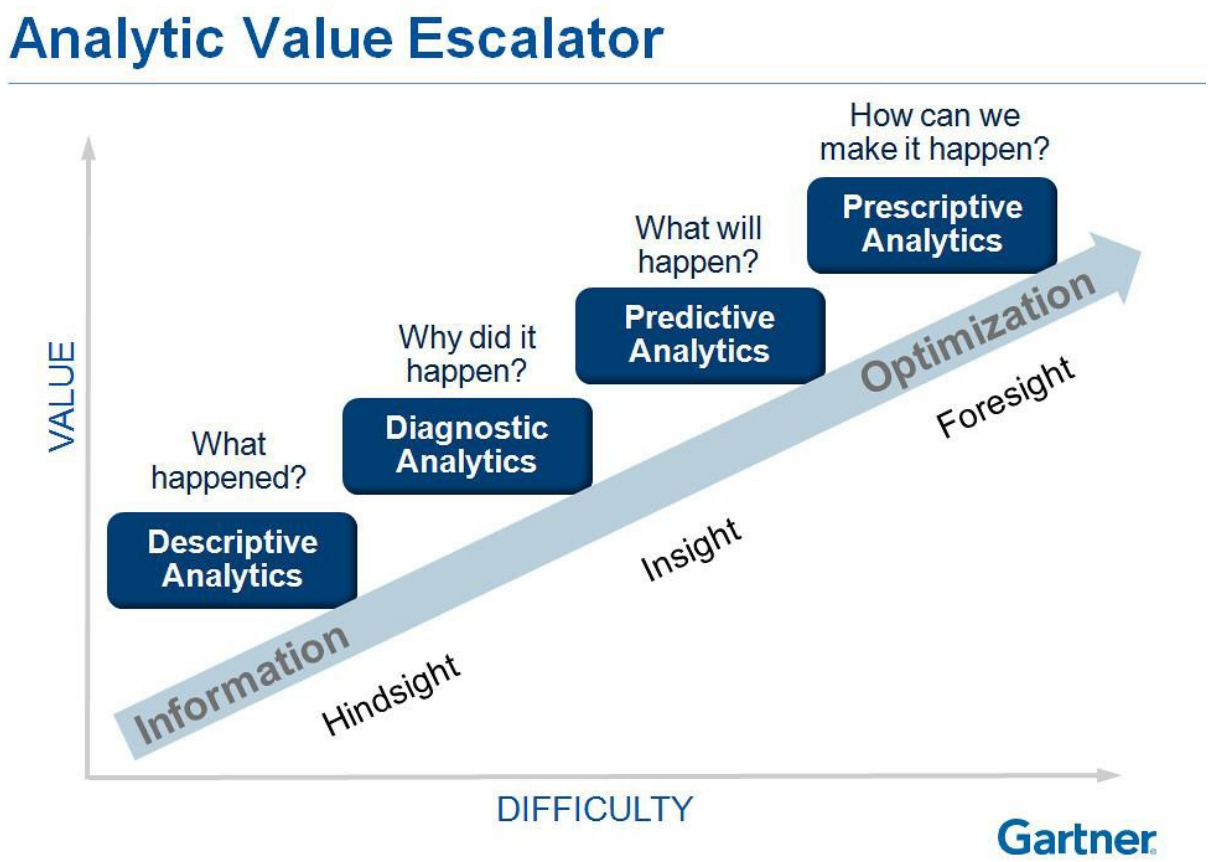

Figure 5. Analytic Value Escalator

Most common type of analytics are “descriptive” in nature, where large amounts of data are summarized to explain outcomes and exhibit trends that were observed. The key facet here is that the focus is on the past and the reports describing the events of interest provide a summary. A good example is the annual report of a business enterprise where the business performance is summarized and explained.

A somewhat more valuable question is about understanding why some phenomenon happened. Data is analyzed to extract correlations among disparate trends and a root-cause analysis is undertaken to arrive at plausible explanations. This is the domain of “diagnostic” analytics.

Moving towards the future, a lot of attention is being put these days on “predictive” analytics. In this case, the available data is used to not only extract past but also future trends, based on which future actions can be predicted. Use of AI and ML techniques is gaining popularity, driving further adoption of real-time recommendations based on the use of predictive algorithms. A common example is predicting the next-best action that a customer or other subjects of interest will undertake, given the insights obtained from analyzing their past actions. Demand for increasingly more accurate predictive systems continues to skyrocket.

Predictions are useful and necessary but not sufficient when it comes to force a specific action to take place. For example, data analytics helping an autonomous vehicle can predict obstacles along its path but that alone will not get the vehicle to successfully navigate. This is where “prescriptive” analytics comes into play. A specific action needs to be selected and executed from among the list of potential options as recommended by a predictive step.

The Data Flow

The traditional data flow is shown in the upper half of the graphic below. It is aimed mainly at text and numeric data sources that are ingested, transformed to fit a schema and then stored in data warehouses or databases. This is the so-called “Extract, Transform, Load” or ETL process. Once stored in a repository, the data is analyzed in batches, using software that employs rule-based reasoning. The output from this step is then compiled into a report or a dashboard for consumption by humans for decision making.

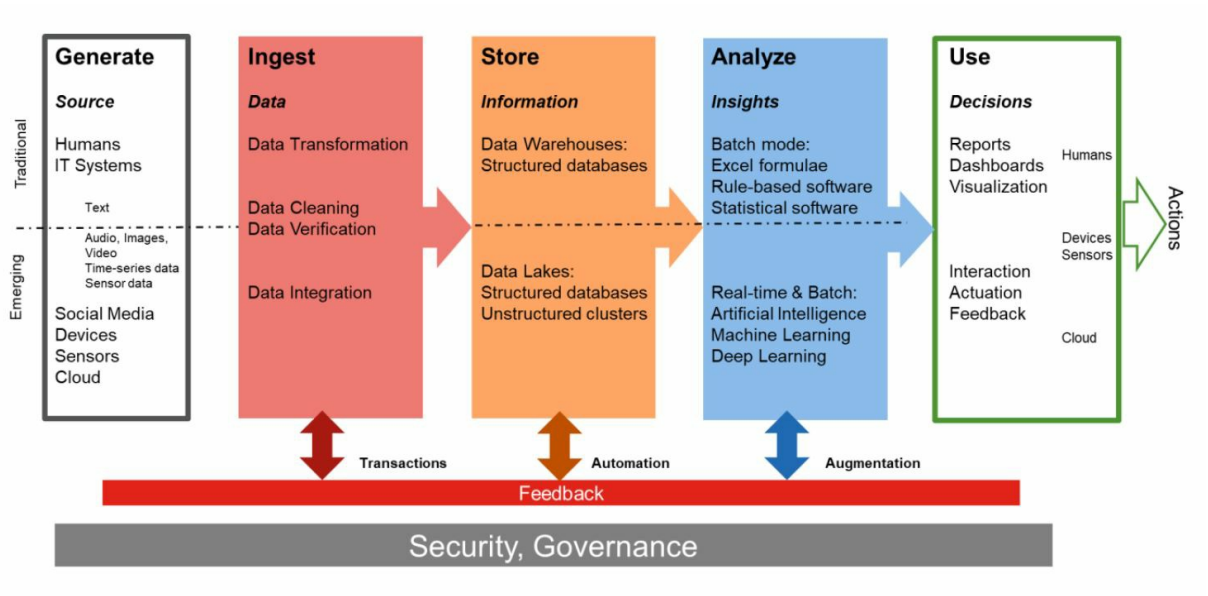

Figure 6. Traditional and Emerging Data Flow

Over the past decade or two, new data types have emerged, including images, video, audio, sensor data, time-series data and so on. Sources of such data are social media platforms, devices and sensors that make up the IoT infrastructure, as well as a host of cloud-based collections available in as-a-Service offerings. Ingesting data from this variety of sources require further processing steps such as data verification, cleaning and integration prior to storage and processing. Many of the new data types are so-called unstructured or semi-structured data that do not fit into rows and columns of a traditional database. In such cases, the entire data set has to be stored, in distributed storage architectures like Hadoop clusters and object stores. Collectively, storage of structured and unstructured data has come to be known as “Data Lakes,” alluding to the volume of data flowing in. Furthermore, much of this data is generated and consumed in real-time, requiring mechanisms to handle its velocity, in addition to the aforementioned volume and variety attributes.

Data captured in such data lakes is processed using emerging analytics techniques from the field of artificial intelligence (AI), including machine learning (ML) and deep learning (DL) models and algorithms. The computing power and memory plus data transmission requirements for AI-based analytics can be significantly higher, requiring a totally new class of system architectures. Finally, the output from the AI-based analytics is often used to augment real-time decision making rather than static reports or dashboards. Equally important is the fact that humans are no longer the only consumers of the insights coming out of AI-based analytics. Many of these consumers are the machines or systems themselves and the decisions are feedback into the operation of systems to achieve a better outcome. Need for such analytics-driven actuation has facilitated significant advances in automation across a multitude of industries and sectors.

At a high-level, each stage of the end-to-end data flow poses a set of challenges that must be overcome to arrive at a data platform design that meets the target requirements.

Ingest

- Cost
- Fan out and parallelism
- New technologies

Store

- Databases
- Data warehouses
- Data marts
- Data Lakes
- Capacity
- Throughput
- Cost of storage
- Tiering of data storage

Analyze

- Rule-based, statistical methods
- Traditional machine learning (ML)
- Next-gen AI, including ML and deep learning

Use

- Reports
- Beyond reports, actuation

The implementation of end-to-end data processing flow as described in this section must incorporate the associated security and governance aspects at a foundational level. Protecting the data from malicious or unintentional misuse a fundamental requirement in the design on any data platform. Achieving this on an ongoing basis requires governance models and practices that must get incorporated at the architecture and also the practical implementation levels.

AI is Everywhere

As articulated in the previous section, emergence of new types of data sources have resulted in growing adoption of analytics techniques from the domain of AI and ML. It is fair to say that AI is the new BI – business intelligence is now increasingly being based on AI methodologies.

In this sense, AI is following the lead of database management systems. Initially, DBMSs were standalone systems, deployed separately from application software that implemented various functions within a large business enterprise. Over time, databases became embedded inside these applications software systems, such as CRM, ERP and SCM. With the growth of cloud computing, databases have also now become available on “as-a-Service” basis. AI systems are very likely to follow these modalities of deployment, with the only difference that all three of these scenarios will be adoption in parallel rather than the serial manner in which DBMSs were adopted. Some AI systems are being set up as standalone systems, to be used by data scientists and AI engineers. Simultaneously, AI is being embedded inside application software, including databases and data warehouses. Finally, AI systems are also being offered as-a-Service on almost all cloud computing platforms today.

Growing ubiquity of AI systems brings with it the need to incorporate steps such as data preparation, model training and evaluation, and ultimately model publication for use by applications. Each of these steps require different types of computing and storage requirements. A growing realization among the data engineers today is the imbalance in amount of computation needed at each stage of the data processing flow. This is also impacting the investments needed to set up such systems. In some cases, each stage of the data flow will be in constant use, justifying the cost of implementation. In other scenarios, some stages will be used infrequently, leaving the highly specialized computing systems sitting idle. While there are algorithmic techniques like reinforcement learning and transfer learning that keep the underlying computing systems occupied, a careful consideration of all the AI processing needs of an organization is necessary to achieve a cost-effective design and implementation.

New and Evolving Architectures

As discussed in the previous section, the quest for useful insights requires a methodical approach to collecting, storing, processing and using the available data. In this section, we address the question of setting up the necessary infrastructure for each of the stages in the data processing flow.

With the growing adoption of cloud computing, business enterprises and other organizations face a wide variety of choices in defining and deploying the necessary IT infrastructure. This includes on-premises dedicated installations, on-premises private cloud, public and hosted clouds and a combination thereof. Application software aimed at doing specific business tasks can now be deployed across such a hybrid landscape. However, making the necessary data available requires careful considerations along a number of dimensions. The as-a-service deployment model for data flow is taking hold and brings a need for proper governance and security along with it. Traditional data centers are expanding to include edge infrastructures, whether they are in branch offices, remote locations or on mobile assets owned by an organization.

The ultimate success in creating a data-centric architecture that enables an organization to become data-driven is based on an ongoing convergence between the disciplines of high-performance computing (HPC), AI and Big Data. In this section, we provide a broad overview of the resulting new and evolving data-centric architectures.

Everything as a Service

Cloud computing is now a given when it comes to the enterprise IT architecture design for all types of organizations, big and small, across industries and geographical locations. The key benefits demonstrated by cloud computing are as follows:

- As-a-Service business models, enabling pay-as-you-use consumption of IT infrastructure
- Metering of usage
- Ability to scale up and tear down IT infrastructure based on needs
- Speed of getting access to IT infrastructure resources
- Faster prototyping and validation
- Production-ready enterprise grade service level agreements to let major business processes run entirely based on the cloud.

While these features are a compelling set of drivers for universal cloud adoption, there is still a need to retain and maintain IT infrastructure on premises of an organization, based on factors such as security, privacy, ownership, regulations and complete control of key business process steps. These factors become more important when it comes to data ownership and governance requirements. Many countries around the world have laws requiring local storage and processing of data which often leads to the need for on-premises installations for data processing and management. Even in the cases where cloud computing can be used, a hybrid model has emerged, with a private cloud replacing the traditional on-premises infrastructure. The goal of such hybrid systems is to allow use of common tools and practices for managing data and other IT resources independent of where they reside.

The exponential growth in the amount of data has also facilitated the hybrid model. Data is best processed where it is generated because this avoids the cost of moving large amounts of data around. This notion of "data gravity" is forcing design of data platforms to enable hybrid capabilities as an architectural element.

DataOps

As discussed briefly in the previous section, the security and governance aspects of data processing and management are foundational and must be designed into the overall implementation. In the application software domain, the notion of DevOps has taken hold, wherein Development of software and its Operational use are tied together in a common iterative process so as to reduce the time it takes to bring out new features. Traditionally these disciplines have been separated and their objectives are often at odds with each other. Developers want to keep adding new features, while the Operations managers what to avoid unnecessary changes to a tested and qualified software system that is being actively used in production. But bridging these two worlds using common tools that enable continuous integration and continuous delivery has been shown to provide significant benefits.

Data processing is poised to benefit from such coming together of the early and later stages of the end-to-end data flow. Bringing in a variety of data and establishing integration and storage mechanisms for further use is one side of the story. This leads to faster development of new analytics capabilities and software tools. On the other hand, using models trained on prepared, curated datasets in production applications enables continuous business benefits, and interrupting this usage to update the models based on new data can incur significant costs and loss of business and trust. Bridging these two sides of the data flow requires careful planning and new tools and techniques. This is an emerging area of research and development.

The Edge

With the explosion of IoT technologies, the traditional data centers used by enterprises are undergoing a significant transformation. Edge infrastructure is fast becoming an integral part of an enterprise's IT architecture. Much of this change is fueled by data that can now be collected at the edge by using sensors and devices installed in various environments and usage scenarios. Analyzing this data can help a business gain significant advantage in meeting customer needs. This in turn means that the analytics stage of the data flow also needs to extend to the edge infrastructure.

Incorporation of local data staging and analytics at the edge is an emerging trend that is driving the design of edge infrastructure components. Ability to implement AI-based ML and DL techniques inside the edge infrastructure is a key driver for emerging low-powered designs.

Convergence of HPC, Analytics and AI

All the details and descriptions of the end-to-end data processing flow here has highlighted the fact that implementing comprehensive data platform requires:

- **Big Data and Analytics:** Collecting and storing raw data, processing and then preparing data sets of training
- **HPC:** Model training, evaluation and selection can require significant computing power. These are best characterized as HPC problems
- **AI:** Use of trained model for inference on real-time data as an end-goal is enabled by having the complete architecture shown below in place.

As shown in the graphic below, various roles associated with the component disciplines need to be employed to achieve the full extent of desired outcome. The Data Engineer needs to put together the data ingest, repository and analytics infrastructure. The Data Scientist needs to then use the underlying data platform for model training and selection. Finally, the Data Analyst uses the trained models to extract insights, facilitate making of decisions and drive actions based on the decisions for realizing the business benefits.

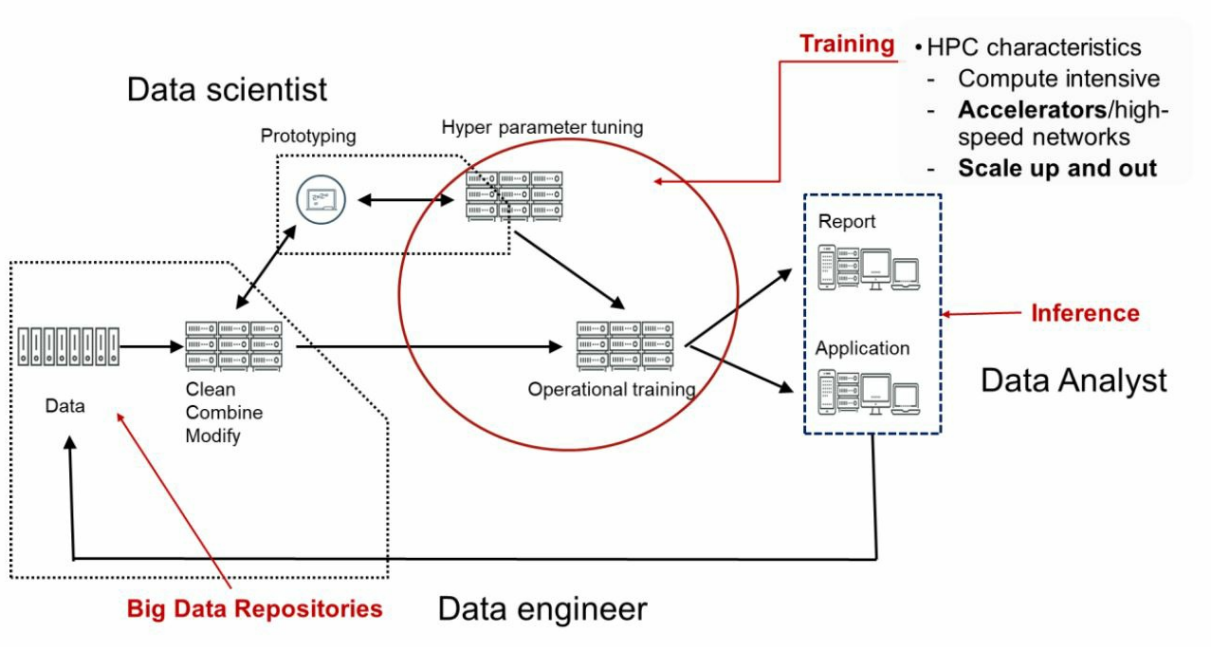

Figure 7. Data Roles

Adoption of AI in the edge infrastructure is seen as a key step in becoming data-driven. In particular, the DL models have two major phases – training and inferencing. The training phase is where the model learns the weights for the neural network that is being trained. The end-to-end process of training a complete neural network can take anywhere from hours to days. Due to this long training time, a lot of effort has been placed into reducing these times either algorithmically or distributing the training across multiple processors or accelerators both within a system or across multiple systems in a cluster. The inference phase is distinctly different. Here, the model is not learning new weights, but just computing them via forward propagation. The inferencing thus takes a sample (e.g. image, phrase) and performs the calculations necessary to classify it.

In practice, inferencing is more complicated than training. It requires a server-like infrastructure and multiple key parameters such as latency, throughput, and efficiency need to be balanced. For example, latency can be improved by performing inference at the edge, but that limits computational power.

With inferencing there are also multiple scenarios to consider, such the AI workload or how queries are sent to the server. The most common workload is computer vision, which relies on convolutional neural networks, and includes image classification and object detection. The next most common workload is natural language processing, which traditionally involved the use of recurrent neural networks, but now increasingly incorporates transformer-based models. These tasks have different computational characteristics and with multiple patterns of inference requests, the number possible scenarios is considerable.

Summary

In this paper, we have covered the dramatic changes in data and its impact on the IT industry. As we move to a data centered world, computing will become more hybrid both architecturally and physically. At the same time new disciplines and approaches are needed as AI becomes embedded in all that we do. The needs of business for differentiation and the expectations of their customers for customized real time experiences will continue to fuel this revolution.

Moving forwards in this hybrid world requires partnerships: between data scientists and IT departments, between customers and vendors and between hardware and software vendors. With its open partnership approach and strong alliance/customer relationships, Lenovo is ideally positioned to work with customers in this new world.

The next paper in this series will look at SQL Server across the enterprise and how this ecosystem is evolving into the data centered world.

Notes:
[1] Data Age 2025. The Digitization of the World from Edge to Core, November 2018. Data refreshed in May 2020. An IDC White Paper sponsored by Seagate.
[2] Gartner Magic Quadrant for Operational Database Management Systems, November 2019

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION ”AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document, LP1367, was created or updated on October 15, 2020.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1367
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1367.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:
Lenovo®

The following terms are trademarks of other companies:

Microsoft®, Azure®, and SQL Server® are trademarks of Microsoft Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.