A NetApp + APY + LEXISTEMS. white paper

Leveraging enterprise data with automatic Artificial Intelligence and Natural Language Processing

Data has become every organization’s main asset. In a global society where digitalization is now a given, data is the fuel on which all exchanges run and the basis on which most value is created.

Yet, data is still considered a problem. Its intangible nature and exponential growth require management from many perspectives – human, technical, organizational… – so that it is generally associated to the notions of cost, disruption and headache.

It doesn’t have to be. In a recent usecase, LEXISTEMS, NetApp and APY tackled a business problem typical of today’s document-rich organizations by automating AI-based data processing in natural language.

The key actors: LEXISTEMS’ SensibleData solution and NetApp’s DataFabric AI-based storage combined in a global AI-bound solution designed and built by APY’s AI Lab.

The result: data and documents become a sustainable source of profit, performance and customer satisfaction -- providing smarter competitive insights and actively helping users zero in on unrealized business opportunities.

The enterprise data challenge

The numbers on the growth of data production are staggering. Of the 40 trillion gigabytes (i.e. zettabytes or 10^21 bytes) in storage today, 90% have been created in the last two years. That’s the equivalent of every person on the planet producing 1.7 megabytes every second, while Internet use only generates 2.5 exabytes (10^18) each and every day. Based on the current trend, the treasure of worldwide data will reach 175 ZB in 2025.

Useful vs dark data – How much of this is useful? According to IDC, information with business value has jumped from 22% in 2012 to 37% in 2020. However, IDC also estimates that only 0.5% of it is analyzed (3% being merely tagged), while TRUE Global Intelligence locates “dark” data – information that remains unquantified and untapped – north of 50% in average. In every organization – including yours. If, according to The Economist, data has replaced oil as the world most valuable source, that’s a lot of waste.

The problem of silos – In addition to the sheer mass problem, data availability is often hampered by silo structures. Employees in organizations have different data needs and requirements, and so have customers. Everyone generally agrees that a single knowledge base of data and documents would be beneficial to all, but there seems to be a dark force against bridging organizational gaps, fostering synergies and factorizing information value. When data is not available, it just cannot be consumed.

An infinity of formats – From legacy to mobile, data is diverse in structure and formats, which makes it difficult to unify. The more sources and applications – corporate, business, production, emails, social networks… – the bigger the problem, as demonstrated by the difficulties reported by administrations worldwide in managing healthcare data during the recent Covid-19 outbreak. Ideally, data should be accessible by content, regardless of container or other infrastructure technicalities.

Public and open data – The same applies to open data. Governments and public agencies are generally chartered with making their data publicly available. That’s a wealth of information waiting to be combined to private corporate data and transformed into valuable business insights. In 2020, interoperability, lifecycles and languages of expression should not be blocking.

Although well documented, these pain points still characterize the data landscape in many organizations, resulting in missed opportunities and lost revenues. For its potential to be fully realized, data must be made totally findable, accessible, interoperable and re-usable. By humans and machines alike.

A few of **LEXISTEMS' SensibleData®** use cases, here in mobile versions.

Meaning is the new keywords

In every organization, data comes as either structured (typically databases) or unstructured (typically PDF and Office files, emails, multimedia content). For users, this should not matter any more than where and how it is stored. This data exists so everyone should be able to easily find it, access it and work with it. Achieving true findability requires applications capable of understanding both users queries and the addressed data. That is where the age-old keywords don't cut it anymore.

Keywords are obsolete – The data industry was built on keywords but today's data challenges make them obsolete. Keywords are meaningless, fixed, monolingual sequences of characters. From a data usability standpoint, that means:

- no spelling variations or imprecisions
- mediocre results with vocal interfaces
- no results on synonyms or semantic equivalents
- no complex questions as several keywords generally overlap or eliminate each other
- no adaptability to multilingual information.

Keywords-based applications are therefore limited by nature to very basic data processing: they just manipulate exact clusters of letters, regardless of meaning and context. These technical limitations narrow operational possibilities, nurtures users' frustration and prevents easy value creation.

Meaning reveal data – In meaning-based applications, these limitations disappear. Meaning-based applications work by ideas and concepts, like we humans do, on existing and new data alike. They support the infinite variety of our expression, including complex, conceptual, imprecise or fuzzy data queries. Meaningful queries delivering meaningful results (i.e. richer and inherently contextual), the possibilities in data processing and value extraction become several orders of magnitude broader. By allowing organizations to expose and consume information in natural language, in different languages if necessary, meaning bridges the gap between users and data.

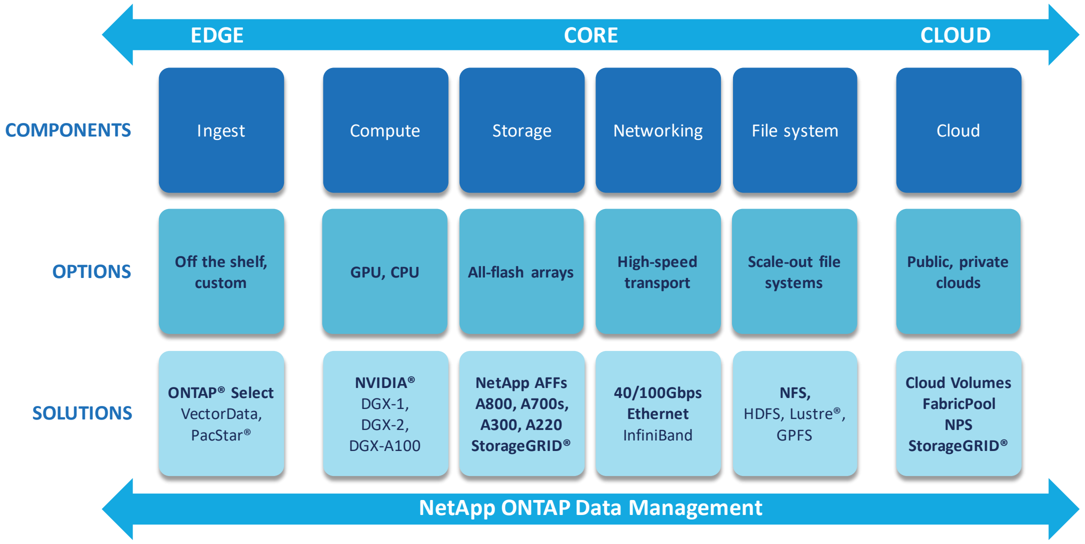
Learning on steroids – Another important benefit of meaning is that it gives applications a much higher ability to learn. With meaning, learning algorithms finely adapt to the nature and evolution of the target data, both in concepts and expression. That is why meaning-based applications are so good at absorbing business, corporate and consumer-originated vocabularies. And that is why they're so good at making new data consumable automatically, as soon as it is produced.

The value of information depends on what consumers can make of it. Unlike keywords, meaning enables applications to understand users and data, and therefore delivers meaningful, contextual, compounded results from any kind of business query. As soon as they know where databases and documents are stored, SensibleData applications can initiate a continuous cycle of real-time AI learning that makes information consumable and monetizable like never before.

LEXISTEMS' SensibleData®: The "Data by meaning" solution

For LEXISTEMS, keywords and Artificial intelligence are contradictory. There lies the motivation for Sensible.ai®, a very innovative technology for processing data and piloting systems by meaning, regardless of language. Based on 10+ years of R&D at the highest level, Sensible.ai is at the core of a number of usage-specific enterprise solutions, among which LEXISTEMS' SensibleData®.

SensibleData lets anyone search, process and connect information by meaning instead of keywords. From any application, voice or text. That's what makes it way smarter than the most famous search engines or data platforms. And contrary to these, SensibleData is applicable to organizations' private data and documents with full GDPR and CCPA compliance. Read more...



	EDGE	CORE				CLOUD
COMPONENTS	Ingest	Compute	Storage	Networking	File system	Cloud
OPTIONS	Off the shelf, custom	GPU, CPU	All-flash arrays	High-speed transport	Scale-out file systems	Public, private clouds
SOLUTIONS	ONTAP® Select VectorData, PacStar®	NVIDIA® DGX-1, DGX-2, DGX-A100	NetApp AFFs A800, A700s, A300, A220 StorageGRID®	40/100Gbps Ethernet InfiniBand	NFS, HDFS, Lustre®, GPFS	Cloud Volumes FabricPool NPS StorageGRID®

NetApp ONTAP Data Management

NetApp's Data Fabric offering and implementation.

Smart storage enables smart data

Applying Artificial Intelligence to enterprise data – as LEXISTEMS' SensibleData does – also requires a robust workflow to handle operations associated with consolidation and preparation, building and training models, deployment in production environment and monitoring results in real-time. With data anywhere and everywhere, in every form imaginable, and growing by the minute, automating AI processing, at both developers modeling and users consumption levels is a real technical challenge. Hence the need for seamless and actively smart storage across all possible applicative environments.

The concept of a data fabric – Unifying data located in public clouds, private clouds and on premises takes a new, non-disruptive approach in storage implementation and management. Imagine, on the one hand, an environment-agnostic API implementing the same software-defined services wherever it is executed, regardless of systems vendors or cloud stacks. And on the other hand, a single user interface allowing immediate data discovery, integration, optimization, protection and migration to/from any cloud. With simplicity and robustness, this architecture aligns with IT priorities as it makes clouds efficient and transparent extensions of any data center, at will.

Optimizing the Edge to Core to Cloud pipeline – The deployment of a data-bound AI model consists of three stages through which data travels: edge (data ingestion), core (training clusters, data lake), and cloud (data archival). This movement of data is very typical in today's applications, where data spans all three phases of the pipeline. Thanks to the data fabric approach, all necessary components are easily integrated to provide end-to-end workload optimization at every level, including local preparation, farmed compute, file system(s) architecture, networking topology and cloud provisioning.

Immediate benefits – Once the first data fabric elements are in place, the automatic AI pipeline starts producing results. At the edge, secured data traffic benefits from low latency in predominantly sequential large files. At the core, federated multi-ecosystem data lakes (NFS, HDFS, S3, Hadoop, Splunk, Lustre, GPFS…) maximize the performance of data exchanges according to data types (from PDF or Office files to knowledge graphs, logs, time series, multimedia…). In the end, DevOps-style repositories make the trained models eligible to continuous integration and deployment, allowing them to be absorbed for production or iteratively refined until qualification. IT and Data Science people's time has been significantly saved and global costs dramatically reduced.

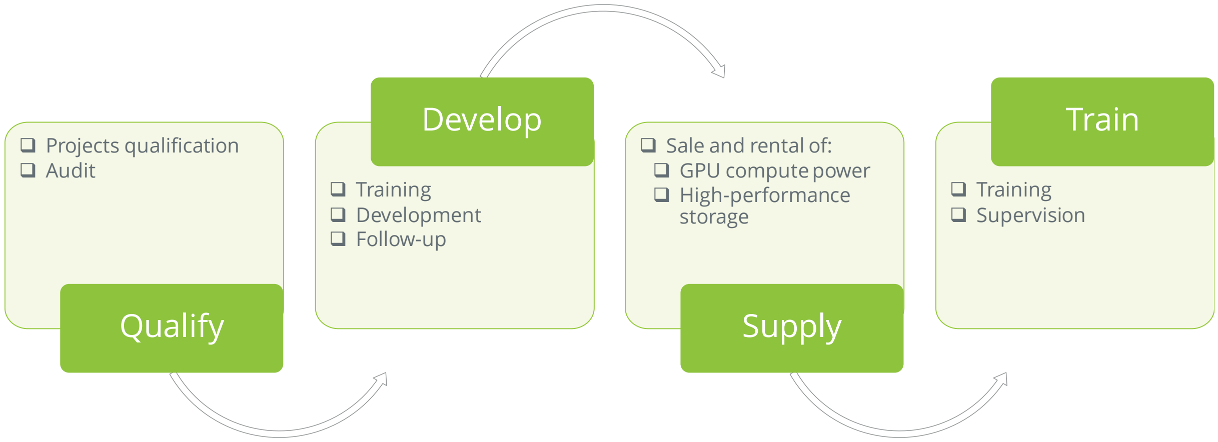
Storage is no longer just the persistence of files, however complex that has become. In contexts of multisource and multi-repositories applications, with increasingly distributed and hierarchized data, NetApp's Data Fabric smart storage brings facilitation and enablement. By unifying and optimizing data access across all enterprise environments, they make solutions like LEXISTEMS' SensibleData capable of producing AI-based information automatically.

NetApp's Data Fabric: Enabling automatic Artificial Intelligence across any storage environment

Whereas storage vendors focus on either workload or data delivery optimization, NetApp unifies both with Data Fabric, a foundational architecture that provisions, manages and runs production, development or test applications where it makes the most sense at any given time. Thanks to a cloud-agnostic API, customers control, orchestrate, optimize and secure everything from a single interface.

This seamless data accessibility is what makes applying automatic Artificial Intelligence to data possible. From ingestion to training to production to archival, data processing is accelerated and availability is guaranteed, to both developers and users, across edge environments, on-premises data centers and any cloud on the planet. Read more...

Ai Lab by APY's
360° AI offering,
from project
to production.



Qualify
□ Projects qualification
□ Audit

Develop
□ Training
□ Development
□ Follow-up

Supply
□ Sale and rental of:
□ GPU compute power
□ High-performance storage

Train
□ Training
□ Supervision

From idea to results with usecase-specific implementations

Once the contours and applicative features of your AI-based data project start taking shape, its overall implementation requires careful consideration. As AI veterans will confirm, technical architecture and execution are keys to success. Hence the importance of teaming up with an AI-specialized partner capable of guaranteeing both best practices and best-in-class configurations.

Learning with GPUs – Future-proof data projects must be able to learn in order to self-improve and deliver better results over time. That's where Artificial Intelligence comes into play. But what kind of learning exactly? It depends on the nature of the underlying data. "Simple" Machine Learning gives excellent results with structured data - typically database records or JSON datastore documents. For unstructured data – typically PDF or Office files – Deep Learning is better suited because the many layers that compose Deep Learning's neural networks make up for the lack of data structuration. In both cases, NVIDIA GPUs are the hardware components of choice.

Inferring with multinode servers – The process of learning results in models. Trained models are loaded by AI-based data applications in order to allow inference, i.e. answering data queries based precisely on the loaded model. Compared to learning, inference requires much less computing power. Accordingly, multinode servers featuring enterprise-grade Intel® Xeon® CPUs are the recommended way to cost-effectively scale up to the targeted number of users -- now and in the future. Depending on the specifics of the usecase and the size of the models, inference CPUs can also be seconded by memory capacities in the order of the TeraByte, thanks notably to Intel's new affordable Optane® memory technology and the combination of operating systems add-ons and software libraries optimized for inferencing.

Implementation and execution – The size and specifications of the GPUs and CPUs host servers – either ready-made or built to order based on project requirements – mostly depend on the volumetry of the data and user base. In this respect, just like no data project is identical, there is no such thing as a universal dimensioning formula. Performance, however, is a function of the proximity between users, data sources and processing servers. The closer their locality, the less latency between questions and answers. A partner fitting the description above will propose rational alternatives including combinations of private and hybrid clouds hosting, in order to achieve the best possible continuous delivery of models to your application and results to your users.

As ambitious as it may sound, leveraging enterprise data with automatic Artificial Intelligence and Natural Language Processing is now possible. By combining LEXISTEMS' meaning-based data processing and NetApps AI-based smart storage within a scalable infrastructure, APY's AI Lab created a uniquely data-centric solution that is far more profitable than the sum of its parts, ensures long-lasting value creation and yields highly positive user satisfaction metrics.

APY Group's AI Lab: A one-stop shop to simplify the adoption of Artificial Intelligence

Created in 1998 and firmly established in Europe and North America, APY is a recognized builder of bespoke professional computing solutions for the Media & Entertainment and artificial intelligence industries. Within the company, APY's AI Lab helps businesses build the right solutions for their specific processes and assist them all along project lifecycle, including change management.

From idea to industrialization to users training, the AI Lab experts also rely on an ecosystem of preferred partners in the Machine and Deep Learning fields. For APY's customers, this alliance guarantees best-in-class results, in the form of dedicated solutions built on battle-trusted software stacks and running on certified storage and compute equipment. Read more...

© 2020 NetApp, Inc, APY MICRO SAS, LEXISTEMS SAS. All Rights Reserved. NETAPP, the NETAPP logo and the marks listed at netapp.com/TM are trademarks of NetApp, Inc. APY, Ai lab by APY and the APY logo are trademarks of APY MICRO SAS. LEXISTEMS, the LEXISTEMS logo and the marks listed at lexistems.com/tm are trademarks of LEXISTEMS SAS. Other company and product names may be trademarks of their respective owners.