Big data for the intelligence community

Contents

Summary

Data volumes are evolving from petascale to exascale, and this evolution brings huge challenges. The solutions for dealing with burgeoning amounts of information are changing from "the analyst finding the data" to "the data finding the analyst." Building on the architectural concepts of *acquire*, *apply* and *store*, IBM big data capabilities are incorporating existing tools with new technology innovations. IBM is making ongoing, substantial research and development (R&D) investments in big data solutions that are highly applicable to intelligence agency requirements in order to help agencies extract timely and efficient insights from both structured and unstructured data.

Objectives of this paper

Based on our experiences with commercial markets and government agencies, big data can impact the technology and business aspects of analytics systems. With these aspects in mind, this paper has the following objectives:

- Explain how big data requirements drive emerging flexible and integrated analytics platforms by combining new and traditional solutions
- Highlight the capabilities of big data solutions that are required for the intelligence community to address needs for ingestion, analytics and data persistence
- Describe a context for capabilities in which the government can iteratively build on the capabilities and technologies it already has in place

- Present a high-level understanding of big data and analytic systems technologies from an IBM perspective
- Provide insight into current and future big data solutions from IBM, and show their organizational and operational impact on core processes, workforce, and information and communication systems

The big data challenge

The world generates over 2.5 quintillion bytes—that is 2.5 billion billion bytes—of data every day. Ninety percent of the world's data has been generated in the last two years alone[1], and the proliferation of human- and machine-generated data sources will continue. Figure 1 shows the projected explosion in social media and sensor data volumes through 2015.

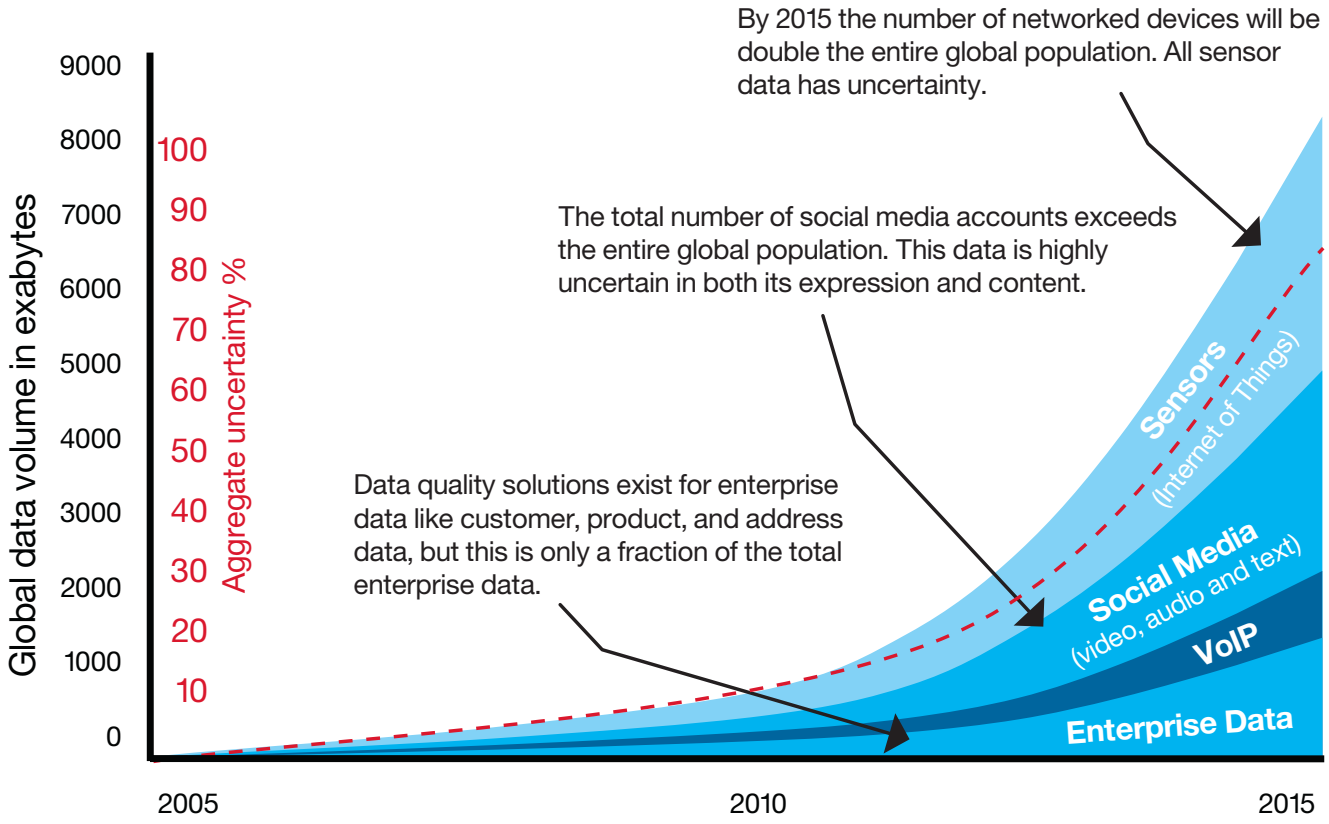

Figure 1. The global data explosion

This increase in data is driving a clear distinction between data that can be considered in batch (data at rest) and data that needs to be analyzed in real time (data in motion). To prepare for the future, intelligence agencies must cope with more and varied information—both structured and unstructured—to ensure an effective, timely and traceable "information-to-intelligence" transformation.

Data and information interpretation and management are critical to support the rapidly changing needs and targets of law enforcement and intelligence agencies. An example is to provide adaptable intelligence analytics by using a big data storage management infrastructure with agency-specific controls on the data and its security.

The iterative process of data acquisition, application and storage is central to the daily operation of these organizations, as is the focus of their business missions on creating actionable intelligence from raw data. At different levels within an intelligence organization, the following important questions must be addressed:

- How is exponential growth in data handled?
- How does an organization capitalize on the increasing range of data that is now available and, in particular, on the surge in unstructured information?
- How does an organization take advantage of this data in a timely manner to ensure relevance and value, while maintaining the traceability of intelligence products back to the source data?
- How should an organization share data and types of data across people and organizations to break down the silos that keep key insights hidden and isolated?
- How does an organization sustain its current capabilities while keeping in line with big data solutions?
- How does an organization deal with both cost and compliance constraints when considering big data solutions?

International and national law enforcement and intelligence agencies have overarching imperatives to ensure that data value is not diluted through inefficient management. This approach helps agencies operate within appropriate legislative frameworks and adhere to data and records management policies. In today's economic climate, intelligence organizations also need to meet fiscal constraints that are likely to continue for some time. Each intelligence organization must balance the threats and opportunities that are available in today's big data, while driving efficiency within the existing information technology (IT) infrastructure.

Big data perspective

To maximize operational effectiveness and minimize cost, the vast amount of information being generated today cannot be processed and taken advantage of by using only traditional methods, algorithms and tools. The variety, volume and velocity of today's data require new ways of handling information. Big data is often defined by these three distinguishing characteristics, and when all are present, they necessitate a change in the architectural approach needed for handling that information:

- **Volume.** Exponential growth in data must be accommodated in a cost-effective, highly available and high performing way. The traditional methods of data processing and storage place an unrealistic burden on infrastructure when working with big data.
- **Variety.** Handling a wide variety of structured and unstructured data, whether human-generated, machine-generated or sensor-generated, cannot be done by traditional relational databases.
- **Velocity.** Both historic and near-real-time processing and analysis must be accommodated to ensure that data is used while it still has value. Traditional systems are designed for bulk or relatively low-speed acquisition, and can be used only for post-event application.

Successful initiatives tend to follow three patterns of deployment, underpinned by the selection of one big data entry point that corresponds to one of these key characteristics of big data.[2]

Beyond these three initial dimensions of approaching big data challenges, organizations are also normally challenged by additional characteristics:

- **Veracity.** Data-based decisions normally require traceability and justification. Applying techniques to address data uncertainty and the inherent truthfulness of data—whether in nuanced text, sensor precision, model approximation or process uncertainty—is important. Understanding and using provenance and pedigree information might be required if data is merged from upstream company or agency sources.
- **Volatility.** It is necessary to reduce the latency between the acquisition of data and the derivation of actionable insights, ensuring that time-sensitive and high-value information is used as rapidly as possible.
- **Value.** Depending on the nature of the application or the acquisition mechanism, some data might have more intelligence worth than others. The value of information can increase over time, and historical data can provide additional insights in support of a changing threat picture.

To address the challenges of what are called "the six V characteristics," the platform on which future big data solutions are built must evolve, extend and embrace new capabilities while building on current capabilities. New architectural models and system designs for working with big data must:

- Integrate and process data in real time from an increasing array of available sources.
- Handle the diversity of large, complex data formats.
- Enable fast access to commodity petabyte-scale storage hardware.
- Establish massively parallel computing clusters that can support distributed jobs by processing chunks of big data, with inherent resilience and fault tolerance.
- Manage the lifecycle of enterprise and government data that is approaching critical mass.
- Support industry standards and norms for information exchange.
- Support various IT deployment models, including services delivered using the cloud. This approach is important for big data clusters that must share resources with other clusters, be provisioned on demand with self service means, or be measured and metered for allocating costs and providing security between multiple tenants.
- Coexist with high performance computing (HPC) initiatives and strategies within the agency—especially those initiatives and strategies that are designed to produce a workload-optimized solution that makes cost-effective use of system resources and human resources to significantly reduce space, power and cooling in a target compute environment.

Differing big data solutions

Intelligence organizations each have a unique focus for data acquisition, processing and use. Each agency's needs vary and require different data types and techniques. To address these differences. As shown in Figure 2, IBM breaks down big data solutions into three segments—*acquire*, *apply* and *store*. These segments describe the end-to-end big data process, from availability, to the creation of an intelligence product, to retention or disposal.

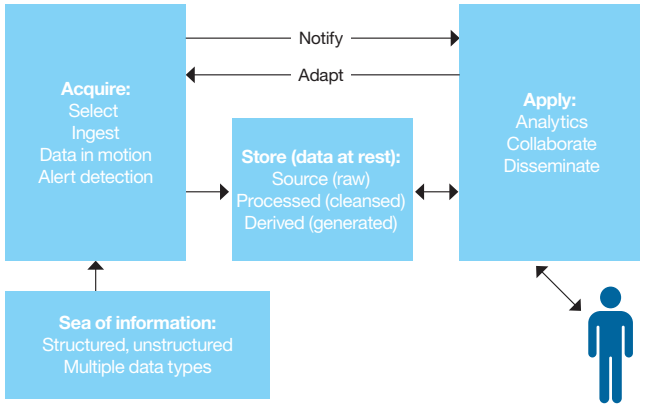

Figure 2. Acquire, apply and store

The logical groupings of *acquire*, *apply* and *store* represent capabilities that are used in combination to progressively enhance and generate intelligence from source data. The underlying processes are iterative rather than linear.

- **Acquire** considers when data becomes available to the organization and the subsequent processes as data arrives at the organizational boundary.
- **Apply** considers how data can be used, from analysis and sharing through dissemination.
- **Store** covers all data that is stored in a raw or processed form, in addition to newly generated data.

Within each of these logical groupings, IBM recognizes the need for greater collaboration across intelligence agencies to improve efficiency and effectiveness, while driving down costs to help address government budget deficits.

Connecting these three logical groupings of capabilities are technologies and infrastructure to provide integration as data with high volume and variety flows through a big data system at high velocity. Integration between new big data capabilities and existing IT investments is often necessary for deploying big data solutions. Information integration and governance allow agencies to understand, cleanse, transform, manage and deliver trusted information to critical business initiatives.[3]

Acquire

Information exploitation has historically focused on using high-quality data that was previously gathered, cleansed and stored. The explosive growth of source data is driving a shift towards new governance for data acquisition and subsequent applications. In addition to the retention of critical business data, there is a need to perform data triage, based on relevance and time to value.

A sea of information is now flowing into intelligence organizations through various paths, as shown in Figure 3.

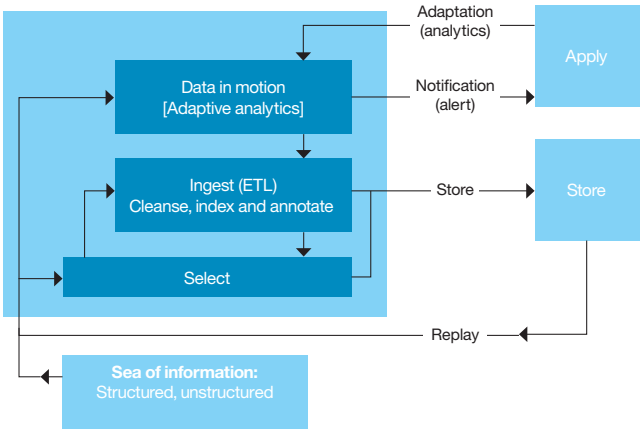

Figure 3. Acquire flow

Traditional acquisition

Large-scale data acquisition has traditionally required significant data preparation before it becomes suitable for exploitation. This preparation effort occurs during the acquisition phase of data processing. It must be able to address the needs of the target data storage environments through transformation or filtering, or be prepared during acquisition for semantically neutral treatment for a range of consumers. Data is extracted from sources, transformed by being cleansed, formatted and augmented, and loaded into a destination system for subsequent use, which is a process commonly known as extract, transform and load (ETL).

Relational data stores, a common destination for acquired data, require a significant degree of design to correctly structure the data repository. The design serves to normalize the data while expressing and indexing it in ways that support its potential uses. Entities or objects are defined, attributes are associated with these entities, and relationships are established among them. Modification or change to the data schema, after the data is loaded, can be difficult and, therefore, is typically avoided.

Unstructured data complicates this process. Before the data can be processed, it is necessary to extract structured features, such as metadata, from the unstructured data, so that it can be normalized, deduplicated, and fused or aligned with data from other sources. The same governance and rigor associated with the ETL of traditional structured data still applies.

Big data technologies allow for the delay of the ETL development process, with much of the transformation happening at the point of exploitation. The process evolves from transformation that is ready for exploitation, toward transformation as needed during exploitation.

The velocity dimension of big data is highlighted when attempting to address cases that require a high rate of data acquisition for data processing and real-time decision-making using that data during acquisition. Variety affects the challenges in transforming that data for storage downstream, as does volume in terms of the capacity of infrastructure to support the data in motion.

Data in motion

Technological advances in speed and capacity, combined with improvements in analytical techniques, have enabled a new class of data exploitation, known as real-time analytics. Moving beyond transactional data or discrete or atomic messages, IT solutions can now consume and process real-time, continuous streaming data, such as high-resolution video imagery and electromagnetic or acoustic signals.

A characteristic of many of today's data sources is the decaying value of the data over time. Data that is used to derive useful information and actionable intelligence at one moment might be irrelevant the next moment. When recognizing this time sensitivity during data acquisition, it is critical to rapidly pass this data to downstream exploitation systems.

In 2003, IBM Research had already recognized the importance of stream computing, and partnered with the US government on an early big data project. It focused on analysis, enabling high-speed, scalable and complex analytics of heterogeneous data streams in motion, and it ultimately became the IBM® InfoSphere® Streams product. According to US government representatives, the impact has been so positive that multiple installations are planned and in progress.

The InfoSphere Streams big data analysis targets include data feeds, such as video surveillance, wire taps, communications and call records. The application identifies critical patterns

and relationships among these diverse data streams, analyzing millions of messages per second. Moreover, InfoSphere Streams allows user-developed applications to rapidly acquire, analyze and correlate information as it arrives from thousands of real-time sources. The applications can be further modified over time to tune the analysis process. Data can be dynamically filtered based on user-specified criteria and then enriched with reference data, transformed and correlated with other diverse feeds. This general set of capabilities can be done today with other products and technology. However, InfoSphere Streams enables intelligence organizations to apply these capabilities much earlier in the acquisition cycle, while the data is still in motion and before it has been placed into a repository.

A benefit of timely exploitation is the ability to allow for near-real-time changes in the acquisition process to support analytics to be performed as the data is acquired. For some analysis requirements, intelligence agencies need to react or respond in the moment. They need the ability to potentially influence or change the outcome of an event while it is still unfolding. Effective and timely change can greatly improve the quality and relevance of the intelligence product.

The main business benefit is the improved enablement of decision-making because critical data is pushed to the analyst. Analysts can avoid menial and time-intensive queries to pull data from historical systems and shift to high-value decision-making as soon as prioritized alerts are pushed to them.

Triage: The selection of data

Source data is often sparse in nature with a low value per byte. In isolation, it has little or no exploitation value. Through the combination of data sources by fusion or co-registration coupled with sufficient data volume, intelligence agencies have the opportunity to extract significant value. “Panning for gold” is frequently used as a metaphor to describe these processes.

Data triage must balance cost per compute versus the cost of retention or storage. Intelligence agencies should opt for solutions that couple low exploitation cost with high value per byte. Furthermore, they need to optimize what is stored long term and be selective based on data value and relevance.

Triage includes data cleansing and de-duplication at the point of acquisition to further refine the accuracy, resolution and fidelity of the acquired data, addressing the volatility dimension.

The process of entity integration is a good example of data condensation for entities such as people or organizations:

1. Extract relevant structured records from unstructured documents.
2. Link these records with entity resolution and disambiguation when they refer to the same real-world entity.
3. Amass all the facts about the same entity into one common object with entity population, mapping and fusion.

The future of acquire

Big data makes data movement more difficult, which means it makes sense to co-locate the data with the computing resources needed for exploitation. Capabilities to support acquisition continue to evolve in the following areas:

- Current processing techniques of pipelining and parallelism, which are typically applied to massive scale analytical tasks, will be applied to the process of data acquisition. Breaking down a data source across many processing nodes, each tasked with acquiring a subset of the total data, will enable significant improvements in sourcing and presentation to downstream analytical capabilities.

- In addition to the volume-based improvement in data acquisition, IBM expects to see significant improvement in the ability to acquire data from various sources and structures and support increasing velocity in that context. Data acquisition adapters will become commodity plug-ins across analytic platforms.
- To further build context around a data set, IBM expects to see systems that demonstrate the ability to correlate and corroborate across self-discovered data sources.

Apply

Big data does not create value until it is put to use in solving important business challenges. This effort requires access to more and different kinds of data, in addition to strong analytics capabilities that include software tools and the skills required to use them.

Customer analytics most commonly drive big data initiatives. Other big data applications that are frequently mentioned include operational optimization, financial risk management, employee collaboration and the enablement of new business models. These applications can each use powerful advances in new information management and business analytics technologies.

Data exploitation will become more challenging as the proportion of unstructured source data grows, as illustrated in Figure 4.

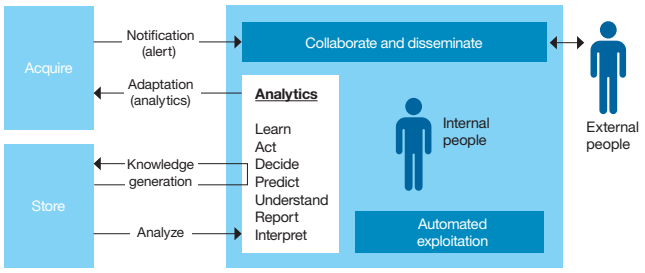

Figure 4. Apply flow

Four key changes are happening in the apply grouping of capabilities:

1. Alterations in analytic techniques
2. Shifts in data volume and real-time processing
3. Broadening of context in analysis
4. Changes in analytic techniques

Alterations in analytics techniques

The methods of applying data are changing because of new data types that represent a significant increase in the variety of data, which, when combined with current data types, leads to new insights and increased value. Data application combines real-time and historic exploitation through a spectrum of analytical techniques: from a simple single data type to complex multiple data types to multisource data types. This application of data is iterative and adaptive, involving human intervention and semi-automated or automated analytics processes. New techniques are also required to cope with the increasing prevalence of social media.

Successful interpretation of data, whether structured or unstructured, is not a single technique or algorithm, but a real-time and historical application of a spectrum of techniques, technologies and algorithms. The analytics can range from simple to complex, including queries, statistical and predictive modeling, and traditional targeted analyses. Examples include digital signal processing for acoustic data, image and video co-registration and fusion analysis, unstructured text feature extraction and entity identification, and geospatial and temporal data analysis.

The following view shows how analytics and support capabilities might be delivered and organized for an agency:

- A team supporting business intelligence and performance management capabilities might report outcomes of business processes and programs, automate management dashboards and scorecards, and create planning, budgeting and forecasting tools.
- A team supporting advanced analytics and optimization capabilities might apply advanced statistical and regression analysis upon historical data for predictive decision making, and integrate optimization algorithms and technology into operations.
- Ensuring that robust and trusted data is available when needed and is easy to consume might be handled by an enterprise information management (EIM) team. This team could provide a consolidated and efficient information platform to support analytics and initiatives in an optimized way.
- Managing structured and unstructured content for providing efficiency and transparency to complicated workflows might be handled by an enterprise content management (ECM) team.

Future analytic techniques include advanced processing across multiple data sources and types. Structured and unstructured data are analyzed in a holistic way, avoiding disconnects, such as silos, and enabling the timely compilation of a richer total picture, including:

- Combining the power of machine-generated and sensor-generated data in motion with historic information, accommodating multimedia sources.
- Providing automated exploitation on data in motion and enabling real-time adaptive analytics to the data, creating a faster "learn and adapt" culture and improving the quality of the intelligence product.
- Handling imagery historically and in real time with such techniques as real-time identification and tracking of defined objects, automatic classification of images using textual and digital metadata, or derived features of the images using classifiers. This classification can then be used to trigger a range of activities from deduplication of data by making strategic decisions and taking action based on real-time situational awareness.
- Incorporating machine learning approaches, such as image recognition, to problems.

Few intelligence agencies have the will, time or resources to create and maintain a single unified platform. Therefore, any framework must support the fusion of analytic outputs, whether in the context of the multiple approaches or techniques that exist today, or to support new techniques or data sources that will emerge in the future. Recognizing the need for this type of search and access, IBM recently added the Vivisimo Velocity information optimization platform to its product offerings. Now called the InfoSphere Data Explorer, the platform allows intelligence organizations to access, navigate and analyze the full variety, velocity and volume of structured and unstructured data without migrating to a single system.

Shifts in data volume and real-time processing

The increasing volume of data will necessitate a shift from human-based analysis to automated analysis. The latencies associated with the sequential application of *acquire*, *apply* and *store* are incompatible with the rapidly changing threat picture when coupled with explosive growth in data volumes. New data and information must be examined at the time of acquisition for early threat or incident identification and exploitation. For example, to address cyber security challenges, agencies must focus on large-velocity support for data analysis during ingestion to enable immediate and automated follow-up and remediation.

In the big data context, automated application exploitation has two purposes. It allows an agency to handle significant volumes of data in an efficient and consistent way, freeing analysts from mundane and repetitive tasks. It also enables data users to focus on areas where human intervention is needed or most effective. Through the automation of accumulated, fused, historical and real-time data analytics, items of significance are detected and communicated to the user faster than with a manual process, while incorporating larger and broader data sets for analysis.

The recently announced InfoSphere Sensemaking system provides automated capabilities for extracting non-obvious relationships from massive amounts of structured and unstructured data, while considering previous historical decisions. The technology is based on techniques and capabilities developed for commercial industries, such as gaming and financial markets. In the context of automated intelligence, this system involves:

- Making assertions about observations.
- Using new observations to reverse earlier assertions.
- Drawing on the accumulated context for higher quality relevance detection and selection of a next best action, if any.

Broadening of context in analysis

An intelligence agency cannot act effectively without understanding the larger context of what is seen from analysis. In the current environment, organizational requirements and procedural mandates, including security requirements, encourage analysts to work in isolation. Two analysts might be working independently on separate problems that consider a common aspect of a person, place or thing. How are collaboration and communication encouraged across these isolated domains when new, relevant data becomes available?

Through collaboration technology, tools can track individual work scopes, preferences and the profile of analytic requests and search parameters. Collaboration tools, such as wikis, chats, tweets and emails, can notify analysts about areas of intersecting interest and value. These tools and techniques can work within an organization when creating insights and when communicating discovered insights to external organizations.

Such collaboration tools in themselves can create organizational knowledge. For example, mining content and exploring the implied network can lead to a better understanding of skills and expertise in the organization. Data can then be pushed to people who might be interested based on previous usage matters. Search can also be enhanced to take into account a person's network and interests. Techniques for masking certain information can extend this collaboration. For instance, sensitive information such as credit card and social security numbers can be obfuscated, allowing users to access and understand the broader data context without disclosing sensitive information.

The 2012 Federal Big Data Commission Report gave examples pertinent to the intelligence community to support cyber security by exploiting data using analytics. It stated that "deep forensics, critical infrastructure protections,

Supervisory Control and Data Acquisition (SCADA) security, and insider threat protection are areas of focus for big data cyber security capabilities."[4]

IBM is seeing the emergence of collaborative intelligence analysis through tools such as IBM i2® iBase IntelliShare, which was announced in 2012. i2 iBase IntelliShare was extended to support access to intelligence data across the enterprise, empowering collaboration and enhancing situational awareness for law enforcement, defense, national security and private sector intelligence teams. By using a browser-based working environment, users can contribute to, question, explore and receive on-demand, priority-driven intelligence from a shared repository.

Moreover, IBM sees enormous potential in the combination of big data analytics, social monitoring and collaboration. For example, IBM works closely with IBM Business Partners HMS Technologies to deliver their TACTrend solution to several government agencies. TACTrend monitors and analyzes social media content, and IBM Connections social software disseminates the "needle in a haystack" to the required intelligence teams.

The future of apply

Much of the insights used by the intelligence community today comes from a skilled human analyst who selects the right tool for the job and uses data that is readily available. In the future, analytical systems will move toward a more declarative model for problem solving. In this model, a user focuses more on describing the problem and less on worrying about the details of how the problem will be solved. Increasing cases will occur where the data finds the analyst rather than the analyst spending significant time finding the data.

The following methods are being considered:

- A declarative approach will allow analytical problem solving that applies the most appropriate analytical technique and selects the most appropriate runtime environment. This abstraction from the underlying technology platform and runtime architecture assists the user in determining the best approach to apply.
- IBM expects to see improvements in the methods available for big data visualization. New visualization techniques will allow iterative inspection of complex and massive data types.

The promise of achieving significant, measurable business value from big data can be realized only if organizations put into place an information foundation that supports the rapidly growing volume, variety and velocity of data.

Store

Store is all about repositories: how information is retained, accessed and managed across its lifecycle in the context of the big data paradigm. Traditional IT information management solutions cannot scale to meet the combined storage, network and computational needs of the big data infrastructure. This challenge is seen when the volume and variety of data required for downstream parallel retrieval en masse must be stored and distributed to a cluster in high velocity.

At some point, the combination of data volume for storage, the velocity at which it is changing, and the variety of the data causes a need for an architectural change to support these three dimensions. Content management systems, HPC systems and enterprise databases were each designed to handle one or two of these characteristics at a time. However, the three together create the need to have extremely high network bandwidth, storage capacity, distributed processing and high input/output to the storage from the parallel processing capability of the cluster.

The big data requirements of emerging intelligence demand an interconnected infrastructure to provide responsive and high-performing capacity in a cost-effective manner. The volume of big data increases the focus on access to stored data across a large compute infrastructure. The velocity characteristic leads to architecture decisions that support large bandwidth across the cluster boundary of the solution. The variety dimension challenges the techniques that are used for parsing, storing and retrieving data with high volume and velocity.

Generally, traditional and standard infrastructures are not optimized to support the special and severe patterns of resource usage that big data workloads require. For example, traditional storage area networks (SANs) do not have the read/write bandwidth that is needed to perform deep analytics against a massive, unstructured data repository. However, it is not realistic to create an inflexible, single-purpose big data infrastructure that is optimized for one application of the technology. The reason is that you must consider the needs of other applications, legacy environments, and finite resources and finances.

There should be an agency-wide interest in considering the trade-offs of building a reusable standard infrastructure for big data or an infrastructure optimized for a class of big data workloads directed at total cost and performance. No one-size-fits-all approach will do, nor can every configuration be suited perfectly for each workload. In addition, workload usage requirements change over time, and your underlying infrastructures must be flexible and adaptable. Therefore, the choice of any big data infrastructure must be geared to the best approach for the current problem. This includes hybrid operations, which combine optimized environments for storing and accessing information in analytic applications with a general-purpose IT infrastructure.

Figure 5 shows the iterative, internal flow of general *store* processing and its relationship with *acquire* and *apply*.

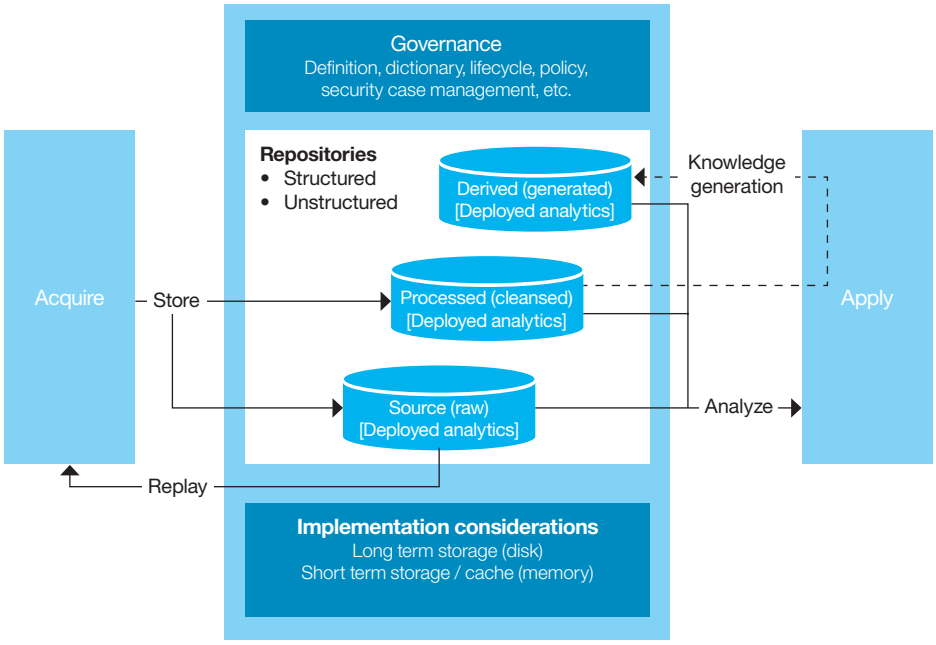

Figure 5. Store flow

Repositories and implementation considerations

Big data is fundamentally about the ability to address the combination of structured and unstructured data. Even if all of the source data is unstructured, the output of analytic operations, such as feature extraction, will be structured results. Data progresses along a lifecycle in the repositories, from the initial raw information to processed information, which is cleansed, normalized, enhanced and eventually combined with other data to derive new insights. This data evolution takes place by applying analytic algorithms and methods.

Repositories serve as the source for input and the storage location for the analytic results. They often act as the framework for running analytics. Rather than exporting the data to an external system and then importing the results, the analytics are sent to the repository and run co-located with the input information.

Traditional network-based storage topologies, such as SAN and network-attached storage (NAS), are optimized for shared access to storage, with centralized management across many connected computational units. By contrast, emerging IT techniques for big data take a data set, divide the data and run it in parallel over multiple nodes. Proximity of the data to the processing unit is key to efficient and timely completion of each unit of work.

Structured and unstructured big data repositories have massive amounts of stored data combined with a high volume of data access. The combined effect will overload traditional IT infrastructures that have separate compute and storage environments. As a result, big data environments are increasingly created by using co-located storage and processors to allow for effective access to mass amounts of data for intensive analytics and queries. This type of architecture is most notably associated with the Hadoop approach, but has also been applied to structured data, for example with the IBM Netezza® solution.

Hadoop style stores that are typically associated with unstructured data can also support applications for structured data, such as NoSQL implementations like Hbase and Accumulo.

This approach is more than just setting up large server farms of commodity hardware loaded with locally attached storage. Without more refinement, these systems can have daily hardware failures because of standard failure rates that can result in data loss if not addressed. They can also incur tremendous power and space cooling costs.

However, intelligence agencies can apply commercially available techniques and capabilities. Advanced processor capabilities, such as field-programmable gate arrays (FPGA), allow for denser storage and processor combinations to reduce power, space and cooling. In some cases, advanced RAID techniques can minimize data loss. To achieve significant space and power reductions for a target big data environment, some dense platforms that are normally used for HPC workloads can be used for consolidation with critical data center resources. The platforms can be interconnected with sufficient bandwidth and performance to fully use its capacity. Agencies can also federate traditional data warehouses and file systems with big data environments to allow for unified access and transparent migration of information between environments when needed. For example:

- Relational database management system (RDMS) structured data warehouses can be accessed using federation layers. This way, users can seamlessly query short-term data in traditional RDMS environments and long-term data that has been migrated to a big data RDMS warehouse.
- Unstructured data can be accessed using file systems that support locally attached storage and SAN storage and that manage transparent migration of information between these environments based on policy, usage and lifecycle.

As previously discussed, the InfoSphere Data Explorer platform supports the virtual federation of structured and unstructured data without having to migrate to a common storage environment.

Governance

Within the intelligence enterprise, data governance is a necessary and critical discipline to support operations. This discipline covers the full lifecycle for data: storing, accessing, protecting in disaster recovery and security scenarios, monitoring, changing and eventually disposing of data and information. The emergence of unstructured source data increases the requirement for strong data governance. Without effective governance of the massive volumes of structured and unstructured data, the accuracy and completeness of the official record is diluted, and the phrase "garbage in, garbage out" becomes a reality.

The combined increase in storage capacity and the drop in storage costs have allowed intelligence organizations to retain and access massive amounts of readily available information that they previously discarded or moved to inaccessible archives. Beyond access, they need strong governance. Examples of effective governance include:

- The preservation of key information even if there are IT failures, outages and or disasters to ensure that an organization has access to all necessary information for business decisions, operations and regulatory compliance.
- Optimization of data replication and access across separate big data clusters in multiple locations to ensure effective, efficient and transparent multisite or worldwide access.
- Elimination and reconciliation of duplicate or conflicting information to ensure that decisions across an organization are made in a consistent way.
- Intelligent, policy-driven data storage and distribution that allows for the efficient delivery of data according to required quality of service as judged by policy and demand.

Security and privacy

Source and derived information in a big data environment must be protected by security and privacy mechanisms to ensure that access and change is restricted to appropriate users. It should allow for a single analytic environment that supports multiple data sources, data types and analytic domains, such as:

- Provenance and pedigree, defining both the information source and reliability. This can be critical to address the confidence needed for offering or using shared data. This sharing is often required for an increasing number of big data analytics agency opportunities to use the variety of harvestable data.
- Data classification on individual sections and in overall documents.
- Data security, and access and update control.
- Audit and compliance for information access and updating.
- Privacy and data masking to limit visibility of sensitive fields and attributes while still allowing broader analytics and access.
- Adopting the practice of due process regarding privacy, deletion and disclosure.

IBM has decades of experience in developing and accrediting product and production solutions that meet the stringent security and privacy concerns of defense and intelligence customers. These solutions include:

- Accredited operating systems, such as IBM z/OS®, IBM AIX® and Security-Enhanced Linux (SELinux)
- Databases and warehouses with built-in row and column access controls, such as IBM DB2® and Netezza
- Industry-leading audit and compliance tamper-proof appliances, such as IBM Guardium®

Security and privacy are part of the DNA of IBM solutions and are never an add-on.

The future of store

Future storage architectures for big data should:

- Provide transparent access to data at the application layer where the analytics take place.
- Intelligently adjust to deliver data to widely distributed locations.
- Reflect policy-driven tiering and distribution for efficient delivery of data according to quality-of-service-based on-demand monitoring.
- Incorporate new storage mediums and architectures that break from the standard controller spinning disk model. These new architectures will offer significantly higher storage capacity with higher reliability, lower costs and reduced power, space and cooling.

IBM Research drives the future of big data

The explosion of unstructured data from sensors and social media creates new challenges in all aspects of the big data process. For example, decisions need to be made while the data is acquired. Consequently, analytics needs to move into the acquisition phase to drive some of this decision-making.

This approach might require new architectures at the data acquisition level and new systems to facilitate this up-front decision-making.

New architectures that bring analytics closer to the data can change the way the intelligence community traditionally views computing, from a collection of processing units with NAS, to a collection of compute and store nodes.

When taking advantage of new data types, intelligence agencies need new algorithms to handle data at speed and data at scale. Specifically, social media analytics requires algorithms that can sift through large data sets and find the nuggets of important information to handle the relatively low information to noise ratios. Algorithms are required to understand the networks that are derived from this information. Visualization of large data sets and methods to explore the data also become crucial because part of the big data challenge is finding answers to questions you did not even know you should ask.

Figure 6 illustrates the flow of an analytics system, from acquisition to filtering and extraction validation, to core analytic algorithms, to composition, packaging and deployment.

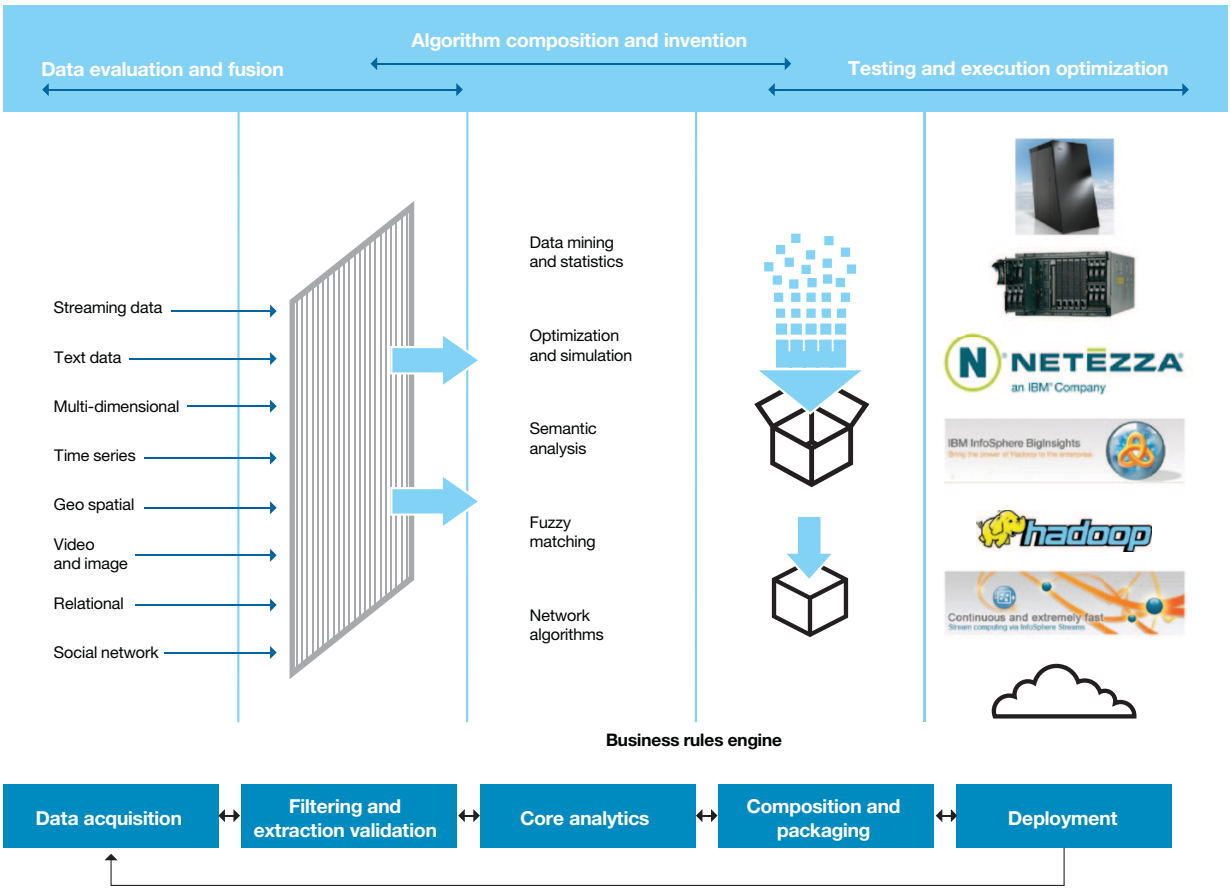

Figure 6. The evolution of analytic systems

IBM Research has long realized the upcoming significance and potential of big data and made it a major focus of research beginning in 2010. The research teams have been addressing all layers of big data solutions. Several key cases have driven innovation to lower levels of the system stack and into standard off-the-shelf products. Some of these innovations are available today in InfoSphere BigInsights™, a portfolio of software and services for the analysis and visualization of big data, and include:

- IBM BigSheets, a spreadsheet-like tool that allows data scientists and business users to explore InfoSphere BigInsights collections and discover new insights without writing any code.

- Jaql, a high-level user-centric query language that is designed to analyze any type of structured and non-traditional big data, enabling the analyst to construct queries and analyses without having to use programming languages. Jaql was developed by IBM and donated to the open source community.
- Advanced analytics of unstructured data using Unstructured Information Management Architecture (UIMA), which was developed by IBM and donated to the open source community.
- Performance optimization with adaptive MapReduce that automatically adjusts systems parameters based on workloads and by using a workload scheduler that provides for optimization and job control based user-selected metrics.
- General parallel file system-shared nothing cluster (GPFS-SNC), an extension to the proven and scalable IBM General Parallel File System (GPFS™) distributed file system that supports the Hadoop environment with better performance, less duplication and POSIX compliance.

As new data types emerge, such as sensor data, new filtering and extraction tools must be created to process them, and new analysis algorithms might be required to interpret them. IBM Research might be looking at data mining and statistics tools, optimization heuristics, fuzzy matching or network algorithms. Sometimes, the novelty of the data can imply creating something new, beyond stringing together existing pieces to solve the problem. Therefore, leaving room for new capabilities is important. IBM Research is also developing novel analytics for new data types that are typically handled by big data systems that specifically apply to the intelligence community. These analytics include:

- A rich set of classifiers for image and video data.
- Algorithms for processing large-scale graphs to locate communities, leaders and people of influence.
- A framework for processing textual data and tools that enable customization to various domains.
- Technology to index large-scale textual data.
- Tools for real-time processing of spatio-temporal data.

There are a growing variety of data types within social media. The intelligence solutions of the future must be able to process this information and combine it with more traditional sources of information in a timely manner. IBM Research is developing novel approaches to acquire and analyze social media data. These approaches include the social media micro-segmentation and real-time correlation (SMARC) framework that uses previously gathered data in combination with social media data sources to improve the timeliness and quality of the decision-making process. SMARC uses previously gathered data combined with text analytics and statistical analysis of new data to achieve entity integration. The goal is to create 360-degree intelligence profiles of targets in real time for rapid decision making. SMARC was developed for commercial clients, but can be readily repurposed to support intelligence agency requirements.

Based on the accelerating demand for data storage capacity, IBM Research is investigating a new high-capacity recording device that is designed for low-cost, reliable and robust archival storage of data for 50 - 100 years. The device provides a self-contained, mechanically robust, sealed environment, eliminating possible head-media incompatibility. The system uses current state-of-the-art magnetic recording technology and provides an order of magnitude advancement in track following accuracy, allowing recording capacities of 250 terabytes or a quarter of a petabyte in a single unit. Four of these units would fit within a single 1U high, 19-inch rack enclosure to store one petabyte.

Finally, IBM Research is developing several novel applications of big data, some of which are especially relevant to the intelligence community, including:

- Anomaly detection to decrease insider threats, such as data leakage, sabotage and malware propagation.
- A solution to enable cities and law enforcement agencies to detect incidents from crowd sourced data to obtain key performance indicators (KPIs) and evaluate trends in multiple areas, including public safety, transportation, water and energy usage.

The future of IBM Watson

In 2007, IBM began a quest to push the boundaries of automated open-domain question answering. We sought to develop a computer that could understand and answer natural language questions over a broad range of topics by analyzing and learning from large volumes of text. In 2011, IBM introduced IBM Watson™, a computer system capable of precisely answering natural language questions with accurate confidence estimation and speed.

The direction for future Watson systems is to provide interactive, evidenced-based decision support, directly using the knowledge in large volumes of natural language text. Watson will be enhanced to explain its results, delivering rich evidence profiles that explain the current set of possible answers. As Watson tackles new knowledge domains, a new continuous learning process will be created to efficiently adapt and train it.

For example, in healthcare we see an important opportunity to help IBM clients better convert large volumes of unstructured content into actionable knowledge, providing profound social and business significance. In considering possible answers and their evidence, Watson 2.0 will identify gaps in its understanding and generate learning questions. If answered by a medical expert, these questions will help Watson fill in knowledge about the language and meaning it needs to improve its performance. The essential point in this example is that Watson is learning to alter its answers and its confidence in those answers by directly reading the evidence. It will provide intelligent decisions based on data originally written by humans for humans, like natural language text, rather than by experts crafting hundreds of thousands of formal rules about medicine.

IBM sees the ready transference of the advances of Watson to directly benefit government customers as its methodology is applied to intelligence domains. The Watson system shares many of the same technical and capabilities challenges as big data solutions. Therefore, many of the components of a big data solution will benefit from advances demonstrated in Watson, which is especially true for the broad variety of data and semantics that the Watson system corpus needs to contain and process. Many of the innovative language processing and analytics patterns and techniques will be applicable as well.

IBM enables big data solutions

- Considering the interest that big data analysis has generated, developers and researchers worldwide are urgently working to deliver solutions. IBM's many recent acquisitions and R&D projects demonstrate that IBM is ahead of this curve. Through investments in open standards such as UIMA, the ability to build from silicon to systems, and the IBM system integration experience, IBM has created and will continue to create highly agile and powerful systems for unstructured analysis to address the impending data explosion.

IBM is uniquely positioned to deliver a single, integrated big data platform, drawing on a range of different tools and techniques for acquisition, storage and application in the new big data paradigm. Circling back to the six V characteristics of big data discussed earlier in the paper, following are examples of the current capabilities from IBM that support many aspects of each one:

- **Volume.** Netezza industry-leading multipetabyte data and analytic data warehouse appliances and GPFS enable a highly reliable and POSIX-compliant alternative to the HADOOP HDFS.
- **Variety.** The UIMA standard, created by IBM, promoted to the open source community and incorporated in such products as IBM LanguageWare® and IBM SystemT, provides the foundational capabilities for using unstructured data.
- **Velocity.** Proven acquisition and storage capabilities with the necessary bandwidth can handle the rate of structured and unstructured data by using such products as InfoSphere Streams, InfoSphere BigInsights, Netezza and DB2.

- **Veracity.** Modeling, scoring and hypothesis generation are provided by such products as IBM SPSS® and InfoSphere Sensemaking.
- **Volatility.** InfoSphere Streams products reduce the latency between the acquisition of data and the derivation of actionable insights. They also help ensure that time-sensitive and high-value information is used as rapidly as possible. However, fully handling volatility assumes that the acquisition components are in control of the solution that ultimately acquires it. If the data is transmitted from a source that transmits inaccurate or imprecise time information, addressing volatility requirements might be outside the control of the solution that is using it.
- **Value.** Such capabilities and products as Watson and InfoSphere Sensemaking provide the analysis, hypothesis generation and evaluation, and significance generation and detection to support the intelligence, investigation and analytical community.

IBM anticipates that analytics computing will heavily influence future system designs. Considering the six "V" characteristics, IBM is in a unique position, based on our decades-long experience in system stack integration, to optimize designs across operating systems, processors and peripherals for analytics success.

Moreover, to help customers manage their cost to value, IBM emphasizes ease of installation, integration and system modification as needs change. IBM considers this ease a key differentiator because of the depth of IBM experience and the breadth of the IBM customer base.

Further IBM strengths include deep expertise in the management of high volumes of data, data center optimization and secure web services, as well as complex analytics and cloud computing offerings. In the future, innovative IBM solutions built by our R&D teams will provide advanced information correlation and collaboration services for the intelligence analyst.

IBM is currently building complete platforms for big data analysis across the spectrum of customer requirements. IBM will continue to adapt its big data and analytics portfolio, meeting product integration challenges while maintaining a vision of creating a Smarter Planet®.

Conclusion

Facing an explosion of information that coincides with budgetary constraints, the intelligence community is preparing to undergo historic changes. Business challenges that have big data characteristics can be addressed by new architectural approaches, using flexible and integrated analytics capabilities and a range of new and existing architectural patterns for applying these capabilities. They can be used to support the full range of big data challenges, from acquisition to data persistence to analytics. Intelligence community application and infrastructure service providers can address these challenges iteratively with the capabilities and technologies it already has in place. They support the mission owners' use of shared big data services to gain economies of scale and enterprise class features.

The risks associated with the complexity of the intelligence community infrastructure and mission processes can be mitigated with the intelligent application of such technical capabilities. Through various commercial, IBM internal, and intelligence community initiatives, IBM has firmly demonstrated that transformation can be achieved that helps leads to cost reduction, improved mission support, and speed to deployment of new capabilities.

For more information

To learn more about big data and the intelligence community, please contact your IBM representative or IBM Business Partner, or visit: **ibm.com**/government

IBM Corporation
Sales and Distribution
Route 100
Somers, NY 10589

Produced in the United States of America
August 2013

[1] As accessed at: **ibm.com**/big-data

[2] *Demystifying Big Data: A Practical Guide To Transforming The Business of Government*, *2012*

[3] *Demystifying Big Data: A Practical Guide To Transforming The Business of Government*, *2012*

[4] *Demystifying Big Data: A Practical Guide To Transforming The Business of Government*, *2012*

GOW03010-USEN-00