epam

CONTENTS

INTRODUCTION TO BIG DATA & DATA GOVERNANCE

Beyond their financial and physical assets, today's enterprises must govern digital assets like Big Data, a new form of value that, when utilized effectively, drives business decisions across all industries in a way that is absolutely unprecedented. The basis for everything from sabermetrics in sports to how self-driving cars get from Point A to Point B, Big Data is the catalyst for improving why and how we make decisions in business and, eventually, every aspect of existence. It sounds dramatic because it is – because until recently, there was no technology powerful enough to uniformly collect, unify, analyze, and take action upon user data in a cost-effective, meaningful way.

Reaching true Big Data enlightenment requires governing it effectively in a way that produces actionable insights – the kind of insights that allow organizations to make the right decisions for themselves as well as their customers. To get there, enterprises must consistently implement the right policies, procedures, and best practices in their data governance strategy.

Fortunately for most organizations, levels of procedures, polices, responsibilities, and systems are already in place to support the timely, secure, and accurate management of data. Consequently, data governance already has many of its building blocks in place. Despite this, however, effective data governance may require a major overhaul due to the following considerations:

1. Defined by an unprecedented Volume, Velocity, and/or Variety of data and subsequent sources, Big Data and its analysis have completely changed the game of governing digital assets.

2. When utilized effectively, Big Data gives enterprises a tremendous competitive advantage as it allows for the application of predictive and prescriptive analytics on huge and continually growing datasets from increasingly varied sources. By implementing novel procedures and technologies as well as overhauling traditional data governance frameworks, it is possible to reveal insights into business processes, market demands, and competitors' strategies. As a result, Big Data is becoming the main driver of innovation within many enterprises. With open-source software like Hadoop, for example, all enterprise data is stored in a central data lake, allowing the organization to have one repository to store and access all assets in their native format and analyze them without the costly ETLs of distributed datasets.

3. Being such a powerful business driver that often transforms an organization's IT toolset, Big Data breeds multiple business challenges, risks, and costs that can be resolved and mitigated by re-thinking data governance frameworks and applying them to Big Data to adhere to changes introduced to the enterprise.

COMMON DATA GOVERNANCE CHALLENGES

- The more data, the more processes needed to keep up with ever-changing, tightening privacy and security requirements
- Without a central data lake, there are multiple points of risk regarding data leakage
- Due to the increase in Volume, Velocity, and Variety as well as the aging of data used for predictions, Big Data introduces more work regarding governance processes and data stewardship
- General data governance frameworks are rigorous and inflexible, and while these strict processes ensure effective data governance, Big Data often results in an enormous amount of time and effort that sometimes exceeds the value of the data
- Data environments, sources, formats, structures, and technology backgrounds are changing at lightning speed, resulting in a lack of control over the data and a feeling that effective management is impossible
- Not all data should be accessible to everyone – in other words, appropriate permissions for each dataset are critical, adding yet another layer of complexity to data governance

OVERCOMING DATA GOVERNANCE CHALLENGES

Given what we've learned about data governance, Big Data, and all of the challenges therein, the task of figuring out where to start seems daunting in itself. While that may very well be true, the task will begin to seem a lot less daunting once we've implemented a proven procedural and philosophical approach. In this case, our approach is two-fold:

1. **Build toward a lightweight, flexible, and cost-effective governance framework.**
   It sounds like a contradiction: There's more data that is of a higher value, and stricter privacy and security regulations and requirements are in place, so how could it possibly be more lightweight, flexible, and cost effective? The solution: We have to invest more in strict data governance.

2. **Control this fast-paced process with agility.**
   We have an environment that is changing at the blink of an eye, and so must our software. We must create technology that automatically adapts to the seeming chaos of Big Data.

Keeping these two approaches in mind, the aim of this whitepaper is to enumerate the domains that need to be addressed to construct an agile, secure, and maintainable Big Data governance framework and give a technology-focused overview on how we support solving data governance problems around the Hadoop eco-system for our clients.

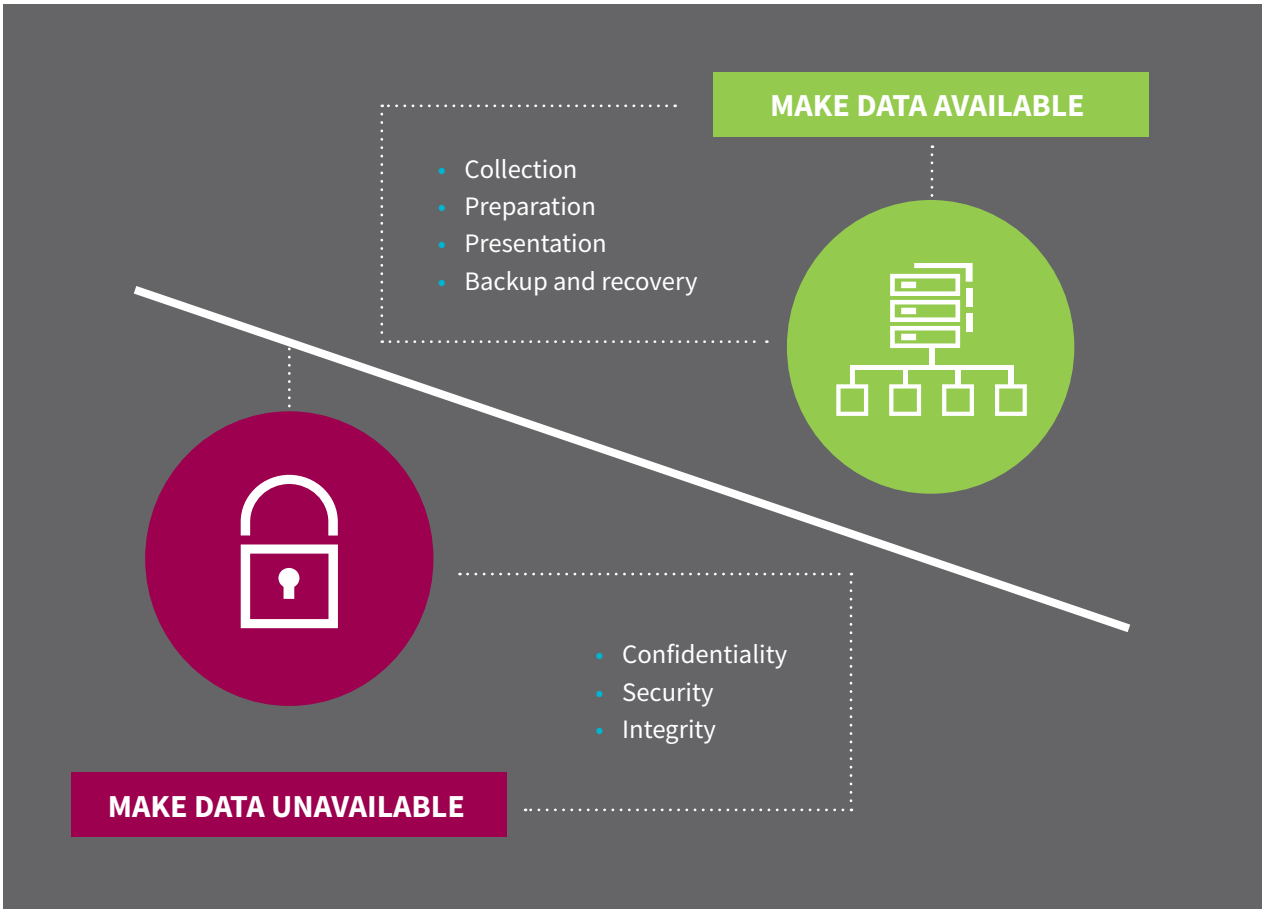

Figure 1 – Challenge: making data available and unavailable at the same time

DATA GOVERNANCE STRUCTURE

Before the advent of Big Data, data governance frameworks contained many elements applicable to the Big Data world. While these legacy frameworks were helpful in constructing modern data governance frameworks, they failed to take into account the sheer velocity at which datasets can grow and change.

For years, enterprises have been handling data through established processes and working infrastructures, meaning that procedures, policies, and best practices for data governance are widely understood among IT professionals. Still, to comply with the internal rules of data management, cleansing, security, and permissions while providing a communication structure for personnel, companies must approach data governance for Big Data in a more agile, fluid manner.

By definition, data governance aims to create a strategy for data management; set up the organizational structure for tasks; define, approve, and publish policies, procedures, and standards; define what metrics to monitor against which baselines; outline a general technological background; and communicate all results, rules, and changes.

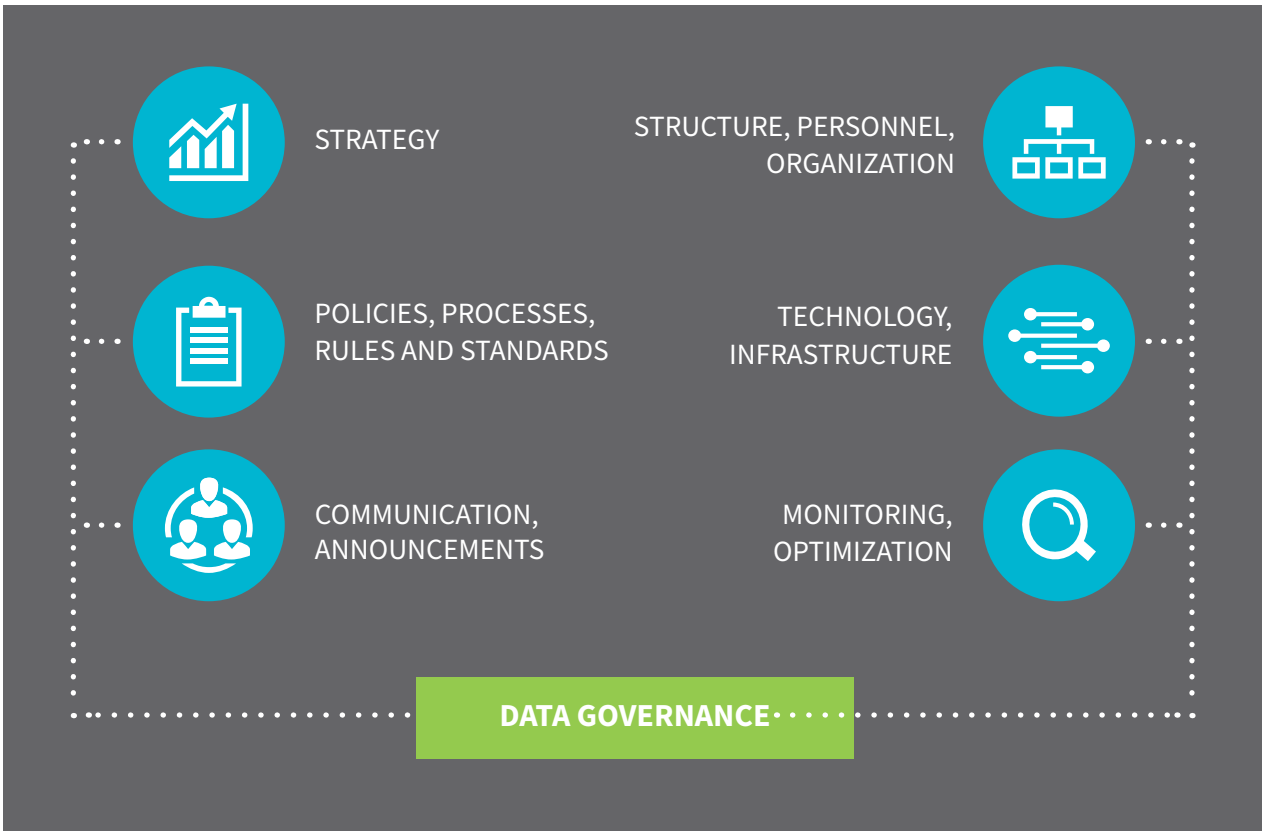

Figure 2 - Data Governance to support business in value generation and strategic decisions

The basis for this structure is to identify data objects and sources by importance (value) in a way that meets all security requirements and regulations. In turn, a glossary is established for use in communication between all technical and non-technical stakeholders.

DATA GOVERNANCE & THE CIA TRIAD OF DATA

In a refined data governance framework, data complies with the CIA triad (Perrin). This means that data is managed, stored, and transferred in such a way that the following properties remain intact:

Confidentiality, Integrity, Availability

To guarantee that these traits remain intact, organizations must do their due diligence to ensure the following:

- Compliance is established with internal and external rules and regulations, making sure that lawful requirements are established.
- Responsibilities of IT are defined so that they can design the systems for timely, accurate, and cost-effective data management that supports the business with complete reporting within the framework.
  - IT sets up a secure, non-redundant, high-performing, and maintainable data management solution.
- Risks of data accuracy, integrity, security, and costs are known and mitigated.
  - As concerns grow over privacy, failing to implement a data governance framework with a focus on data security could result in high costs, business continuity outage, reputation loss, or even bankruptcy.

Figure 3 - CIA Triad of Data

DATA GOVERNANCE REALIZATION BEST PRACTICES

Implementing a world-class data governance framework requires the development of many complex processes, and diving in can often result in a number of high-cost projects. Be careful not to do it all at once with a waterfall approach, though, where failures might only surface at the end of the project. Instead, take a systematic, end-to-end approach following these guidelines for best practices:

- Quantify, communicate, measure, and present the value of the data governance project
- Create a corporate data dictionary, and define the data governance building blocks based on the dictionary:
  - Enumerate data object types
  - Prioritize data assets based on their value, availability, and integrity requirements along with the risks of those getting imposed due to security threats
- Change is only possible if we know what to change and have an understanding of the environment, so be sure to define which metrics to measure, set up monitoring, and act based on objective measurements
- Enhance communication between the business and IT, ensure that common goals are well known for all parties, and obtain sponsorship of management for all activities
- Define (or refine) data governance in the following order
    1. Data strategy
    2. Organization
    3. Processes
    4. Technology
- Take a holistic approach, but make step-by-step progress - data governance implementation is an ever-changing process that needs the continuous alignment of multiple stakeholders, processes, and IT systems

For more information on best practices for data governance frameworks, please see the STANDARDS & REGULATIONS portion of the Appendix

MAKING BIG DATA WORK FOR YOUR ORGANIZATION

For an organization to produce the most useful analytics and ensure optimal decision making, it's critical that you establish a Big Data governance framework where all datasets comply with the CIA triad. By following this method, you avoid misleading or bad-quality data input and get the most out of your Big Data.

To avoid issues with implementing a Big Data governance framework, EPAM suggests introducing Big Data to the organization in an iterative, agile way (EPAM Systems). Doing so effectively means following these essential guidelines:

- Know the value that you want to achieve by utilizing Big Data
- Become aware of the properties of your data: 3V, sources, structure, schema, peaks, value
- Train personnel on data objects and Big Data paradigms
- Identify privacy and security requirements
- Maintain data quality and consistency
- Log actions and analyze them via automation (always automate everything that can be automated)
- Define a simplified data dictionary with roles and responsibilities assigned to data objects
- Simplify the data dictionary by including only the essential metadata with light timeframes on updates based on priority/security
  - Define a retention policy based on the value of data
- Use proof-of-concept projects delivered on a per-iteration basis to ensure there is an effective architecture in place

As opposed to a waterfall approach, organizations should employ a continuous transition when implementing Big Data solutions. In the first phase, for example, proprietary systems should be operated in parallel with the new multi-node systems, often resulting in additional implementation and operation costs as the organization finances more governance and ETL projects that utilize a mix of casual and novel IT architecture elements. Further, be prepared to carve out new engineering roles such as "Big Data Scientists" and "Data Steward."

While there are many new costs and changes to be wary of as we make the shift from traditional data warehouses to the Big Data stack, remember that business value generated by the competitive advantage brought on by Big Data outweighs risks, leading to cost effectiveness in the medium and long term. For more significant long-term cost savings, explore vendor-independent implementation and utilize mainstream operational hardware when possible.

COMPARISON OF TECHNICAL GOVERNANCE BUILDING BLOCKS FOR TRADITIONAL EDW & BIG DATA

GOVERNANCE	TRADITIONAL EDW	BIG DATA
Data confidentiality	Granular access settings available for enterprise architecture, encrypted transfer, and timely data disposal have well-known workflows and IT solutions.	In geo-distributed storage, failure to comply with different privacy regulations of multiple countries is a possible risk. The more data, the higher the chance of exposing data objects due to irresponsibility. The security and permission management framework is often immature, so custom development and thorough testing are critical. Default behavior of BD architecture includes suboptimal permission definition and unencrypted data transfer.
Data integrity	Commercial solutions are available for maintaining an audit trail on data changes and keeping track of transaction logs. ACID compliancy ensures complete record management.	Out-of-the-box ACID compliant software stack is scarcely available. Change tracking, transaction logging, data lineage establishment, and logical integrity enforcement each require custom solutions.
Data availability	The software stack is commercially supported and has specific, already proven SLAs. Backup and disaster recovery solutions are available out-of-the-box.	Big Data clusters are built with data redundancy in mind, although the high-availability settings of a cluster require trained professionals. Maintenance of backup solutions for disaster recovery is costly due to high volumes. RPOs and RTOs are to be defined, implemented, and thoroughly tested for high-priority data objects quantified in the Data Dictionary.
Data sources	Mainly organization-generated internal data.	Organization-generated internal & 3rd party data from external brokers. Intentionally or unintentionally shared social, sensor, smartphone and/or system-generated data.
Data quality	Organization-generated internal data is assumed to be of high quality. ACID compliant software stack enforces completeness and accuracy.	One data source's bad quality could ruin the validity of the analytic results used for business decision support. Links might be missing among various sources and incomplete. Checking abbreviations by rules requires custom code and can lead to inaccuracy.
Data architecture	Multiple enterprise data system best practices and standards are already established.	Problem-specific, complex architecture could fail to comply with requirements due to lack of expertise on the design and sizing.
Monitoring systems	Proven commercial RDBMS, EDW, or DMS systems include logging, monitoring, and reporting solutions out-of-the box. Third-party SW also available.	Best practices on setting logging and monitoring are to be elaborated and implemented based on the custom purpose of the solution. Big Data solutions themselves often generate a high volume of logs that are to be automatically analyzed by the Big Data cluster itself.
Information management roles	CIO	CIO and distributed data management roles of data generators and users such as marketing for social media, supply chain for manufacturing logs, HR for employee-generated browsing logs, legal for data privacy, billing for call records, MD for health records, etc.
Information ownership	Data stewards and Responsibility Assignment Matrix (RACI) (PMI, 2013) defined for different schema structures, IT systems already supporting and guiding stewardship processes	Often data is collected but not used ("Do not throw data away"). Based on the data volume, source, and request frequency, significantly more data stewards are needed. IT systems must be customized, and additional systems must be developed as no stable, out-of-the-box auditing and security systems are available for Big Data solutions.
Technical personnel	Data steward, Data analyst, Database architect, IT operations	Increase in headcount and expertise of data stewards. Big Data steward, Data analyst, Data scientist, Big Data architect, IT operations. Training on novel technologies needed as experienced technologists are hard to find in the market.
Data Dictionary (DD)	Conventional process: best practices, experience, and supporting tools available for a small set of data	A wider dataset being stored requires disciplined, well-trained data stewards who maintain the simplified DD on the data with minimal possible frequency and full coverage to remain up-to-date.

Table 1 - Differences of Data Governance and Big Data Governance Common Risks and Mitigation Strategy

DATA GOVERNANCE WITH HADOOP AS A DATA LAKE

Hadoop is a free, Java-based software framework that supports data storage and analytics on a commodity hardware stack. As an open-source tool with multiple integrators providing enterprise support, Hadoop is an excellent framework to support the data lake concept – a large data storage repository with immense processing power.

As a data lake, Hadoop ensures distributed storage in a reliable and scalable way, making parallel data processing and analytics possible with a simple programming model (MapReduce). Hadoop consists of a redundant file system (HDFS) as well as a cluster resource negotiator (YARN) that optimizes each workload with multiple applications built upon these entities to facilitate analysis, NoSQL, graph databases, in-memory processing, and cluster administration.

EPAM engineers contribute code to Hadoop projects.

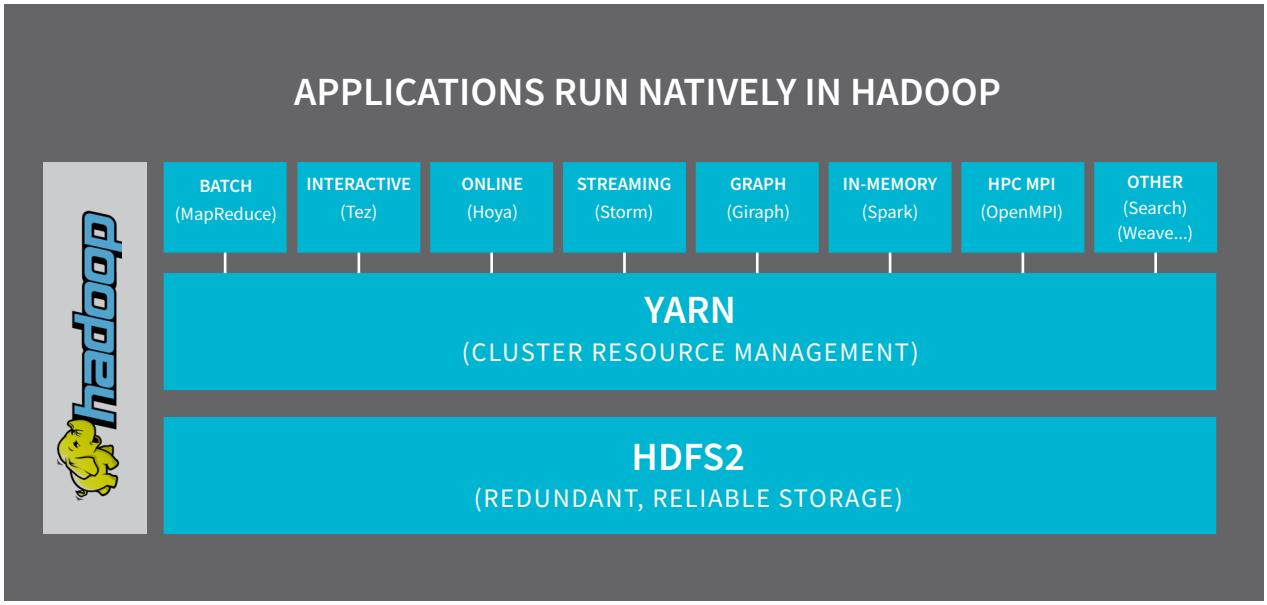

Figure 6 - Hadoop 2's architecture (O'Reilly, 2014)

BIG DATA SOLUTION: THE POWER OF HADOOP FOR DATA GOVERNANCE

To get the most out of Hadoop as a Big Data governance framework, understanding its storage schema and concepts is essential. Hadoop's powerful storage capabilities make it an effective tool for the following:

- Importing a wide variety of data through multiple interfaces
- Storing data to ensure cost-effective, software-based redundancy on commodity hardware in various formats such as:
  - HDFS: redundant file storage with Linux-like user- and group-based ownership and permissions management that can store any file type (structured or unstructured; binary data from images and videos; etc.)
  - Hive / HCat: structured, table-based storage
  - HBase: Key-Value NoSQL storage with cell-based access rules
  - Graphs
- Keeping up with the numerous fast-evolving components for data analytics

When utilized as a robust Big Data governance solution, Hadoop eases operation and maintenance of the infrastructure and data governance processes in the following ways:

- It provides a single point for storing and analyzing any type of data
- It facilitates a cost-effective data storage and/or backup solution via HDFS

COMMON HADOOP DATA GOVERNANCE CHALLENGES

No exercise in data governance is without its own unique challenges. To create relationships for all of the data types above and uncover trends, Hadoop-specific challenges are as follows:

1. The security stack is immature as granular permission management tools are either not available or are new and unproven.
2. There are lots of components with multiple applications, creating a necessity to understand a multitude of components and their relations.
3. There is no full-fledged metadata management system available. The best existing system is HCatalog, which maintains a very limited set of metadata for a limited set of use cases.
4. Auditing and logging data requires a lot of fine-tuning and manual analysis
   - The tools are easy to use, but not always traceable or auditable
   - Data lineage is not maintained automatically

COMMON RISKS & MITIGATION TACTICS FOR BIG DATA

RISK DOMAIN	MITIGATION
Data confidentiality Data loss or theft results in reputation and financial losses	• Identify privacy and security requirements for all data objects • Ensure secure, encrypted storage, transfer and wipe of the data in the EO-lifecycle • Enforce granular pre-defined permission settings • Define data stewards and RACI for all data objects • Test, monitor, and alert security incidents • Make policies, procedures, and RACI available for security breaches • Create a documented and controlled procedure for requests of data
Data integrity For data-driven companies: the risk of bad quality data that is the basis of the business decisions results in business losses	• Access controls in place for data protected against intentional corruption • Test data protected against unintentional corruption from an IT architecture and process perspective • Enforce automated data verification and validation by code • Keep an audit trail for each state change of the data objects • Allow for revertible state changes by versioning the data/transaction logging • Have data stewards and automated systems regularly analyze change logs to alert upon unusual events and unauthorized changes • Always make change of data a documented and controlled procedure
Data availability Competitive advantage loss due to the data not being available when and where it is needed	• Define SLAs for each process stage for each data object and role • Upkeep of controlled IT architecture sizing through POC projects and thorough, high-coverage testing of system performance • Continuously monitor against SLAs and alert upon failure to meet those • Establish a disaster recovery plan and efficient backup system that meets defined RPOs and RTOs for each data object
Failure to deliver value to the organization Costly data governance and management projects fail to meet the main goal: deliver business value	• Consult with experts on how to define business issues, business cases, and opportunities while taking into account all costs for knowledge generation to answer questions • Be aware of the complexity of Big Data projects and the possible lack of experience within the organization • Acquire knowledge on data analytics, data science, and Big Data architecture • Control the beginning stages of projects by using proven methodologies • Utilize IT balanced scorecards on governance implementation and data management (customer focus, internal processes, innovation, costs, etc.) • Implement a monitoring and control framework for measuring the fulfillment of pre-defined goals
Cost overrun Data governance and management projects are overrunning their budget	• Go step-by-step, in an agile way, as you measure and communicate the progress of each iteration • Utilize a vendor-independent IT architecture in order to avoid lock-in • Calculate and measure the ROI of the Big Data stack to quantify the competitive advantage achieved through knowledge generation • Define the budget and be aware that Big Data is not a complete alternative to current systems and involves implementation costs that will return only in the mid to long term • Monitor and quantify the maintainability of the complex IT architecture to avoid redundancy and high long-term costs

Table 2 – Big Data Governance – Risks and Their Mitigation

RISK MITIGATED DATA GOVERNANCE CHECKLIST

Start the project with the following checklist, which has been proven to mitigate the above risks:

- Roles and responsibilities are defined, including new roles associated with Big Data
- Big Data Governance policies are set and enforced
- Process and environmental building blocks are monitored, measured, and controlled
- A Data Dictionary is created for the essentials and is (not too frequently) maintained
  - Requirements for CIA, value, priority, owner (RACI) of data objects are defined
- Data sources are trusted and are providing high quality data (original format or normalized)
- The quality of the data is maintained throughout its lifecycle
- The collection of specific data objects complies with the legal and regulatory requirements of all regions of operation
- The data and information mined are protected from theft and leakage
- All data- and IT-related risks influencing the enterprise and business are identified, managed, and do not exceed the organization's risk appetite
- The infrastructure is built and operated in a cost effective manner
- IT is a trusted partner and actively minimizes business risks

DATA GOVERNANCE WITH HADOOP – EPAM CASE STUDIES

TRAVEL & HOSPITALITY – CASE STUDY I.	
Client	• A major, global online travel services company
Project	• Enhance data processing pace and store large volumes of ever-growing data – The traditional RDBMS system was not capable of cost-effectively managing a large volume of stored binaries (e.g. videos) – Data analysis by conventional means took 10 hours on a data set that changed on an hourly basis – business losses were high due to the incapability of uncovering trends of a high volume and velocity dataset
Strategy	• Create a central Hadoop data lake that contains up-to-date information on all travel itineraries, booking records, customer interaction, clickstreams, and system logs
Scope	• Big Data governance and management, information mining
Timeline	• > 4 years
Team	• > 100 engineers – Teams of 20-30 engineers are responsible for maintaining scheduled reporting processes with often-changing data inputs – Smaller teams are assigned to the ad-hoc querying of specific data objects
Tech. stack	• Hadoop, Hive, Python, various proprietary Hadoop add-ons
Cluster	• Various large clusters, the largest being a 50-node setup with 3 PBs of stored data
Confidentiality	• Multi-layered security: OS-level authentication and disk encryption; Hadoop-level HDFS and Hive table permission settings • Encrypted data flow when importing and moving data among cluster nodes • Compressed, CSV-based clickstream input data used to arrive via FTP, later changed to SCP due its encrypted communication means • Fulfilling statutory requirements: personal details of clients are purged after a year • Data depersonalization module for creating development datasets: generate mock data and replace data columns that are eligible to trace the IDs of customers • Distributed teams use a managed (virtualized) environment; environment access log analytics with the Big Data stack
Integrity	• Automated, software-based provisioning on input data: XML analysis (semi-automated structuring tool) warns of broken format rules as the tool manages automatic corrections and ensures schema adherence to standards • Manual provisioning: data stewards assigned to data objects • Audit review process in place for audit logs • Change log analytics with the Big Data stack • Strict and enforced rules on extending and changing ETL code: the code can only be submitted to production after a multi-stage verification and validation process, including security and performance testing • Mandatory training on Hadoop: permission management systems are connected with the training inventory – only eligible personnel who took Hadoop-related training could submit jobs to production
Availability	• Hadoop and HDFS configured with rack awareness where racks are globally distributed • Business-essential data is regularly backed up to tapes (compressed) • Monitoring: cluster utilization logs analyzed with the Big Data stack, bottlenecks, hotspots identified • Secondary cluster available for (currently) obsolete marked data that might have value for later analytics
Data quality	• Rule engine-based cleanser and validator for the data: the XMLs and CSVs of various sources often contain changed orders of structure/columns, invalid header descriptors, or invalid records. – When there is a rule violation and an action available for resolving it, Hadoop with Cascading automatically makes corrections through scheduled Oozie workflows – In case a violation on validity is found with no rule available to make a correction, an automated digest email on the problem and involved records is sent, allowing for fast manual resolution and automated rule creation • In the case of a system failure, mechanisms to revert data in motion (e.g. during a converting process) are available in the form of transaction logging
Data stewardship	• Assigning data stewards to data objects is a strict rule – there are personnel who – Develop great experience for working with the data structure – Identify possible integrity problems based on the nature of the information contained within data objects – Ensure that processes are in place to firstly avoid, and secondly identify, data leakage

TRAVEL & HOSPITALITY – CASE STUDY I. (CONTINUED)	
Data dictionary	• Maintained by data stewards on KB pages as a manual process • After cleansing and automating structuring, a significant percentage of the data is added to fixed schema Hive tables where headers describe the stored data – Table descriptions are maintained in corporate Data Management Systems by a person assigned to that • Due to fast-paced changes, metadata is often obsolete, and therefore a third-party review and notification procedure is implemented to warn stewards of outdated knowledge of information assets
Monitoring	• Baselines for monitoring both the performance of the service desks on the Big Data stack (e.g. response times, number of requests) and technical system metrics are defined, operation and ETL teams receive alerts when those are exceeded • By using the monitoring results, auditing became a regular activity, be it checking the processes of data governance or the conformance to security standards regarding statutory laws on personal data management • Hadoop is used to analyze the data management environment, therefore providing valuable inputs for the data governance processes – the results are synthetized by analyzing the Big Data stack, with the Big Data stack
Lessons learned	• There is an inevitable need for all team members to know the technical environment's building blocks and the limitations of those, the data, its structure and possible deviations, and the business questions to be answered • Using the agile approach was the key to success: to create such a mature environment, small iterations must be constantly implemented to lead to the continuous evolution of the data governance framework
ENERGY, OIL AND GAS – CASE STUDY II.	
Client	• A major company in the energy sector
Project	• Build a central storage for data coming from multiple, separate relational databases to one central location • Report the data to various consumers, including data scientists conducting ad-hoc analysis on the dataset with Hive, custom-built reporting dashboards, and BI tools
Strategy	• Implement a Hadoop data lake for storage and transformation
Scope	• Big Data governance and management, information mining
Timeline	• >1 year
Team	• 3 engineers, 1 QA at near shore location I • 4 engineers, 2 QA at near shore location II • 3 engineers, 1 project manager at the client's site
Tech. stack	• Hadoop and HBase - Cloudera CDH 5.4 with Cloudera Navigator
Cluster	• The main cluster is being supplemented with multiple additional nodes as the data store scales – it is currently a small one with 6 nodes, 24 cores, 256 GB RAM and with 20 TB capacity • The secondary cluster is a replica of the first one at a far distant geographical location, and it is not purely for data redundancy and emergency recovery purposes as the data scientist teams will distribute the jobs submitted for a balanced data execution • Geo-distribution of the clusters is in progress • Both clusters are built on the client's private cloud
Confidentiality	• Both input and in-data lake data is accessed through managed permissions – for Hadoop it is defined on a file level, for HBase on a cell level • The communication flow is encrypted • Authentication is done via Kerberos • The cluster can only be accessed via SSH channels • Statutory law mandates that selected data objects cannot leave the borders of the country of operation, therefore a separate team is dealing with depersonalization, dataset mocking and sandbox development environment maintenance
Integrity	• Cloudera Navigator is responsible for tracking changes and maintaining an audit log • Cloudera Navigator, with its SolR based search, is used for data cataloging, tagging and event tracking – It logs all HDFS and Hive events – All HDFS files and tables are assigned with searchable metadata – Oozie workflow documentation is maintained – Data lineage is kept track of • For the comprehensibility of transformations and storage, standard UML diagrams are used for data flow and use-case documentation – Artifacts are made mandatory throughout the software development lifecycle – Pre-commit hook enabled for the version control system – updated diagrams must be attached when pushing code into the repository
Availability	• Geographically distributed primary and secondary Hadoop clusters • Policies include retention and ETL job requirements for backup management • The most valuable information assets (tagged with high priorities) are synced to NAS devices via encrypted channels

ENERGY, OIL AND GAS – CASE STUDY II. (CONTINUED)	
Data quality	• As the data lake is filled from multiple sources, including publicly available datasets, many of the sources tend to provide bad-quality data – The quality of input data is always assumed to be low • Data cleansing can be done on the data before adding it to the cluster or after leaving the cluster – In case the former is preferred by data stewards and business users, IT implements rules for cleansing – In the case of the latter, analysts cleanse the data
Data stewardship	• A two-level data stewardship organization was set up – Level 1 for input data: RDBMS-original, row-level owners who could be coming from multiple divisions – Level 2 for in-data lake data: a data steward is responsible for acquiring authorization from the original owners and granting authorization to the transformed data of Hadoop
Data dictionary	• As structured inputs and outputs are generated, data dictionary creation was automated by extracting table headers to knowledge base documents • Client-side data stewards are responsible for the maintenance of the tables' schema and headers, and the same goes for the management of the data dictionary generation scripts
Monitoring	• Aggregated results of logs are shown on a dashboard and reviewed daily • Information is provided on cluster usage, scaling, and warnings • Error-prone events are defined and a notification system alarms either IT personnel or data stewards in the case of system failures or data policy breaches
Lessons learned	• Security compliance is a major part of the challenge; security requirements have to be elaborated before the planning of an architecture blueprint and the procurement of the infrastructure begin • Agile execution enabled successful scaling of the project from a POC to a production data lake of geo-distributed clusters

APPENDIX

DEFINITIONS

Data Governance

1. The definition, enforcement, monitoring, and reporting of an internal structure of roles and responsibilities, processes, and IT systems to provide value to the organization through regulated data storage and information mining while minimizing the risks of operation failures and security breaches
2. The formal orchestration of people, processes, and technology to enable an organization to leverage data as an enterprise asset (MDM Institute)
3. Data governance is the formal execution and enforcement of authority over the management of data and data-related assets (Seiner)
4. Data Governance is a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods (Data Governance Institute)

Big Data

Defining Big Data with its "3V" properties:

1. Volume: the amount of the data to be stored or processed
2. Velocity: the frequency of the data to be processed by IT systems
3. Variety: the types, sources and structure of the data

We talk about Big Data when any of the above properties becomes the challenge – when no widespread, conventional technologies or paradigms are capable of solving problems resulting from the unusually high value of any of the above properties.

Therefore, Big Data is not only about lots of data. Big Data could be of a small volume, but if this small volume contains structured and unstructured data of many sources and should be transferred and processed over the limits of our conventional systems' CPU, memory, I/O, or network bandwidths, a Big Data technology stack and associated governance strategy are to be utilized to overcome the limitations of our organization and architecture.

Hadoop

Hadoop is an open-source software framework that boasts massive storage capabilities for all types of data. It also provides tremendous processing power and is capable of running a nearly limitless number of applications or jobs at any given moment.

Data Lake

A data lake is a large virtual storage repository that holds and processes data maintaining all of its native attributes. Oftentimes, data lakes are built on commodity hardware and designed to support Big Data storage.

STANDARDS, GUIDELINES

- ISO 15489-1 - Information and documentation - Records management - Incorporates guidelines, policies, procedures, systems and processes description on record management of organizations (International Organization for Standardization (ISO), 2001)
- ISO/IEC 20000 - Information technology - Service management - Detail requirements for the lifecycle of IT services (International Organization for Standardization (ISO), 2011)
- ISO 27000 family contains standards for secure information assets
  - ISO/IEC 27001 - Information technology - Security techniques - Information security management - Providing details on Information Security Management System (ISMS) (International Organization for Standardization (ISO), 2013)
- ISO/IEC 38500 - Corporate governance of information technology - Details high-level advice on the role of the governing body and standards for those activities (International Organization for Standardization (ISO), 2015)
- IT Infrastructure Library (ITIL) - Describes an approach to IT service management to realize business change, transformation and growth (Information Technology Infrastructure Library (ITIL), 2011)
- Control Objectives for Information and Related Technology (COBIT) version 5 - A framework for governance and management of enterprise IT (Information Systems Audit and Control Association (ISACA), 2012)
- Country-specific data confidentiality regulations, e.g. UK Data Protection Act - Act for the regulation of the processing of information relating to individuals, including the obtaining, holding, use or disclosure of such information. (UK Act 1998 CHAPTER 29, 1998)
- Domain-specific data confidentiality regulations, e.g. PCI Data Security Standard (PCI DSS) - to enhance payment card data security (PCI Security Standards Council, 2013)

REFERENCES

- CMMI Institute. (n.d.). Data Management Maturity (DMM). Retrieved from http://cmmiinstitute.com/data-management-maturity
- DAMA. (n.d.). DAMA Guide to the Data Management Body of Knowledge. Retrieved from https://technicspub.com/dmbok/
- Data Governance Institute. (n.d.). Retrieved from http://www.datagovernance.com/
- EPAM Systems. (n.d.). EPAM Agile Practices. Retrieved from http://www.epam.com/strengths/agile.html
- IBM. (2007). IBM Data Governance Council Maturity Model. Retrieved from https://www-935.ibm.com/services/uk/cio/pdf/leverage_wp_data_gov_council_maturity_model.pdf
- Information Technology Infrastructure Library (ITIL). (2011). ITIL - IT Service Management (ITSM). Retrieved from https://www.axelos.com/itil
- International Organization for Standardization (ISO). (2001). ISO 15489-1.
- International Organization for Standardization (ISO). (2008). ISO 38500. Retrieved from http://www.38500.org/
- International Organization for Standardization (ISO). (2011). ISO/IEC 20000-1. Retrieved from http://www.iso.org/iso/catalogue_detail?csnumber=51986
- International Organization for Standardization (ISO). (2013). ISO/IEC 27001. Retrieved from http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=54534
- International Organization for Standardization (ISO). (2013). ISO/IEC 27002. Retrieved from http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=54533
- International Organization for Standardization (ISO). (2015). ISO/IEC 38500. Retrieved from http://www.38500.org/
- MDM Institute. (n.d.). http://www.tcdii.com/.
- PCI Security Standards Council. (2013). PCI SSC Data Security Standards Overview - PCI DSS. Retrieved from https://www.pcisecuritystandards.org/security_standards/
- Perrin, C. (2008). The CIA Triad. Retrieved from http://www.techrepublic.com/blog/it-security/the-cia-triad/
- PMI. (2013). PMBOK. Retrieved from http://www.pmi.org/PMBOK-Guide-and-Standards.aspx
- RACI. (n.d.). Responsibility assignment matrix. Retrieved from http://en.wikipedia.org/wiki/Responsibility_assignment_matrix
- Seiner, B. (2014). Non-Invasive Data Governance.
- UK Act 1998 CHAPTER 29. (1998). Data Protection Act 1998. Retrieved from http://www.legislation.gov.uk/ukpga/1998/29/introduction

Established in 1993, EPAM Systems, Inc. (NYSE: EPAM) is recognized as a leader in software product development by independent research agencies. Headquartered in the United States, EPAM serves clients worldwide utilizing its award-winning global delivery platform and its locations in 19 countries across North America, Europe, Asia and Australia. EPAM was ranked #6 in 2013 America's 25 Fastest-Growing Tech Companies and #3 in 2014 America's Best Small Companies lists by Forbes Magazine.

FOR MORE INFORMATION, PLEASE VISIT EPAM.COM

AUTHOR

BIG DATA COMPETENCY CENTER | EPAM SYSTEMS

Peter Kortvelyesi
Solution Architect
Technical Informatics (MSc), Economics (Postgrad.)
CISA, ITIL, ISTQB, IREB, HDPCA

SALES CONTACTS

Please contact us for any sales inquiries or general information on EPAM.

41 University Drive, Suite 202
Newtown, PA 18940, USA
P: +1-267-759-9000