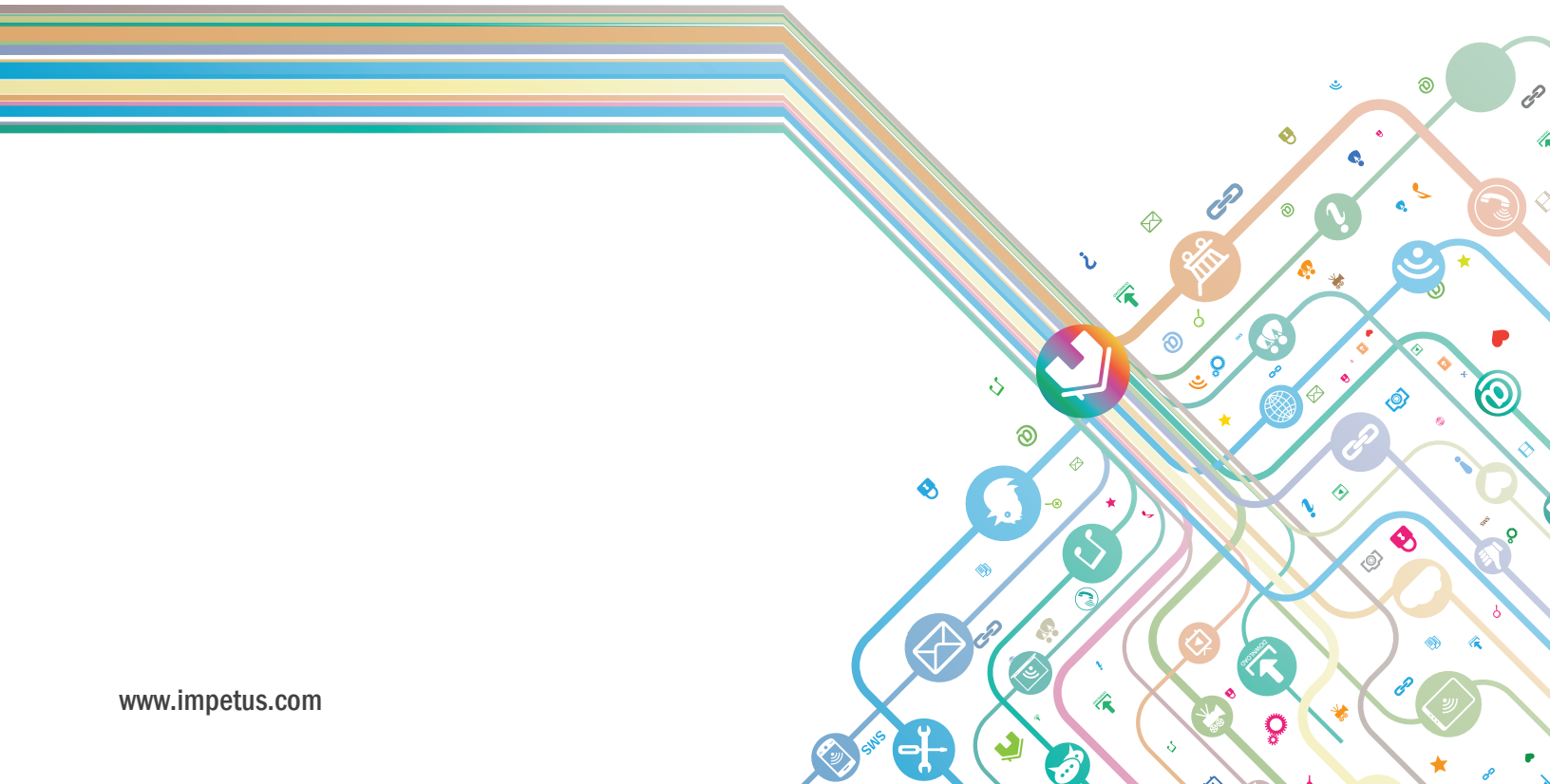
IMPETUS

FIVE STEPS TO IMPLEMENT AN ENTERPRISE DATA LAKE



www.impetus.com

Data Lake Initiation & Execution

This guide is designed to help you determine the emerging importance, significant value and long-term benefits of the adoption of a Data Lake - a pioneering idea for comprehensive data access and management.

It has been created with the guidance of relevant whitepapers, point-of-view articles and the additional expertise of subject matter experts from a variety of related areas, such as technology trends, information management, data security, big data utilities and advanced analytics.

We hope this guide helps you in making the best decisions for your enterprise and in initiating a new IT culture mapped to your business goals.

Introduction

In today's world of fluctuating customer-market-policy dynamics, data is of the greatest essence — data clarifies, revitalizes and reinforces your organization's ability to stay ahead of the competition. It is therefore, a giant value generator, and its maintenance, management, utilization and storage is pivotal to creating a blueprint for the future.

However, despite the advances in technology, managing data is still an arduous task — ensuring its continued relevance, storing and securing it without a glitch and using that voluminous information to your advantage is difficult at times, and requires a smoothened and streamlined process flowchart.

- So, how do you go about creating an efficient data reservoir?
- What makes data more useful to you?
- How do you benefit from its infinite possibilities?
- What are the cutting-edge tools/devices/applications you need, to keep your enterprise future-ready?

To answer these questions we come to the notion of a 'Data Lake' — an idea that is altering the way data is handled across the globe. It is helping organizations hold raw, disparate data in its native format without the need to formulate how or when the data will be used, governed, defined or secured.

So what is a Data Lake?

Step 1: Understand the Changing Data Environment

Before we discuss Data Lakes, let us try and analyze the transitioning 'data-universe' around us.

'Unstructured data' is a term we often hear nowadays — imminent in your organization, even as you imbibe and assimilate huge chunks of information — 'unstructured data' is a continuous presence in your system. Broadly defined, 'unstructured data' refers to information that either does not have a pre-defined data model or is not organized in a pre-defined manner.

The vast majority of information captured from nontraditional sources contributes towards unstructured data — and this data has immense potential. That said, it is also difficult to interpret, and its processing is a laborious, time-consuming task.

Nearly all data existing beyond the realms of a database is unstructured, unrefined and largely indeterminate. What is equally worrying is the sheer volume of unstructured data: At least 80 percent of all digital material operable in our daily lives. Mining this data for insights can give your company a huge competitive advantage.

So how do you manage this avalanche of data?

Enter the data lake.

What is a Data Lake?

TWO COMMON DEFINITIONS OF 'DATA LAKE' ARE:

- A MASSIVE, EASILY ACCESSIBLE, FLEXIBLE AND SCALABLE DATA REPOSITORY
- AN ENTERPRISE-WIDE DATA MANAGEMENT PLATFORM FOR ANALYZING DISPARATE SOURCES OF DATA IN THEIR NATIVE FORMAT

Whichever way you look at it and whatever you may choose to call it (big data repository, unified data architecture, modern data architecture), what is evident is its consolidating and integrating facility — absorbing torrents of 'unstructured data' and creating an accessible, flexible and scalable storage facility, ready for interpretation and analysis. When the data is not specialized and categorized, it can be manipulated in a variety of ways. It is not limited by specific, static structures.

With the wealth of data coming in every moment from different data warehouses (Enterprise Resource Planning, Customer Relationship Management, Human Resource Management) it is important to synergize these diverse silos into a composite whole.

What makes the Data Lake a unique and differentiated repository framework is its ability to unify and connect. It helps you access your entire body of data at the same time, unleashing the true power of big data — a correlated and collaborative output of superior insights and analysis. It presents you with a dynamic scenario where one can dictate a variety of need-based analysis made possible by this unstructured repository.

What are the challenges for your organization's storage portfolio?

Traditionally, an Enterprise Data Warehouse (EDW) has served as the foundation for business intelligence and data discovery. It organized and defined your data according to its quality.
However, EDWs today represent a world consistently falling out-of-sync with the times — predictable, restricted in scope and ability, cost-intensive and unable to handle data complexities.

WHAT YOUR ORGANIZATION NEEDS TODAY IS:

- **STRATEGIC, ADAPTIVE AND INTUITIVE DATA ANALYSIS**
- **REAL-TIME VELOCITY MAINTENANCE AND RESULT SHARING**
- **RESPONSIVE DATA CLEANSING, REFINING AND RECALIBRATING**

The fast, robust, aggressive and effective matrix of big data forces clearly demands a more developed and state-of-the-art data management service. Discovering unique insights is only possible when your data is not limited by structures.

This then, is the point: A newer and vibrant system is needed that can meet the multiple requirements of a razor-sharp business paradigm.

Step 2: Realize Data Lake Repository Benefits

Broadly, a Data Lake blueprint can be demystified into four major capability pillars:

- **The Active Archive**
  An active archive provides a unified data storage system to store all your data, in any format, at any volume, indefinitely. It helps you to address compliance requirements and deliver data on demand to satisfy internal and external regulatory demands.

- **Self-Service Exploratory Business Intelligence**
  Often enterprises need to access a plethora of data for reporting, exploration, and analysis. An enterprise Data Lake allows users to explore data with full security by using traditional interactive business intelligence (BI) tools via SQL and keyword search.

- **Advanced Analytics**
  A fundamental element for effective analysis is the capacity to search or analyze data on a large scale and at a microscopic or granular level. It helps to keep your company's vision and understanding of core functionalities at an optimum. Data lakes can deliver this level of fidelity.

- **Workload Optimization & Transition Management**
  Most companies run their Extract, Transform and Load (ETL) workloads on expensive systems. These can now be migrated to the enterprise Data Lake, therefore harnessing costs, driving efficiency, and ensuring timely delivery.

Step 3: Prepare for the Shift from Schema-on-write to Schema-on-read

Pivotal questions to ask before navigating to a Data Lake setup:

- Are you working with a growing **amount of unstructured data?**
- Are your lines of business **demanding even more unstructured data?**
- Do you need to leverage **big data across your operations, offerings, new products?**
- Does your organization **need a unified view of information?**
- Do you need to be able to **perform real-time analysis on source data?**
- Is your organization moving toward **a culture of democratized data access?**
- Do you need seamless and **simultaneous access to data, analytics and applications?**
- Would your organization benefit from **elasticity of scale?**

If the answers to the above point to a 'must-change' scenario for your company, you need to formulate a phased transformational process.

The exponential growth, velocity and volume of information can be harnessed for sharper, clearer business insights via your data lake. However, bear in mind, your decision to switch to a Data Lake must be prompted by a larger intellectual shift: ***The will to transform your business entity into a more comprehensive, empirical, and data science-driven model.***

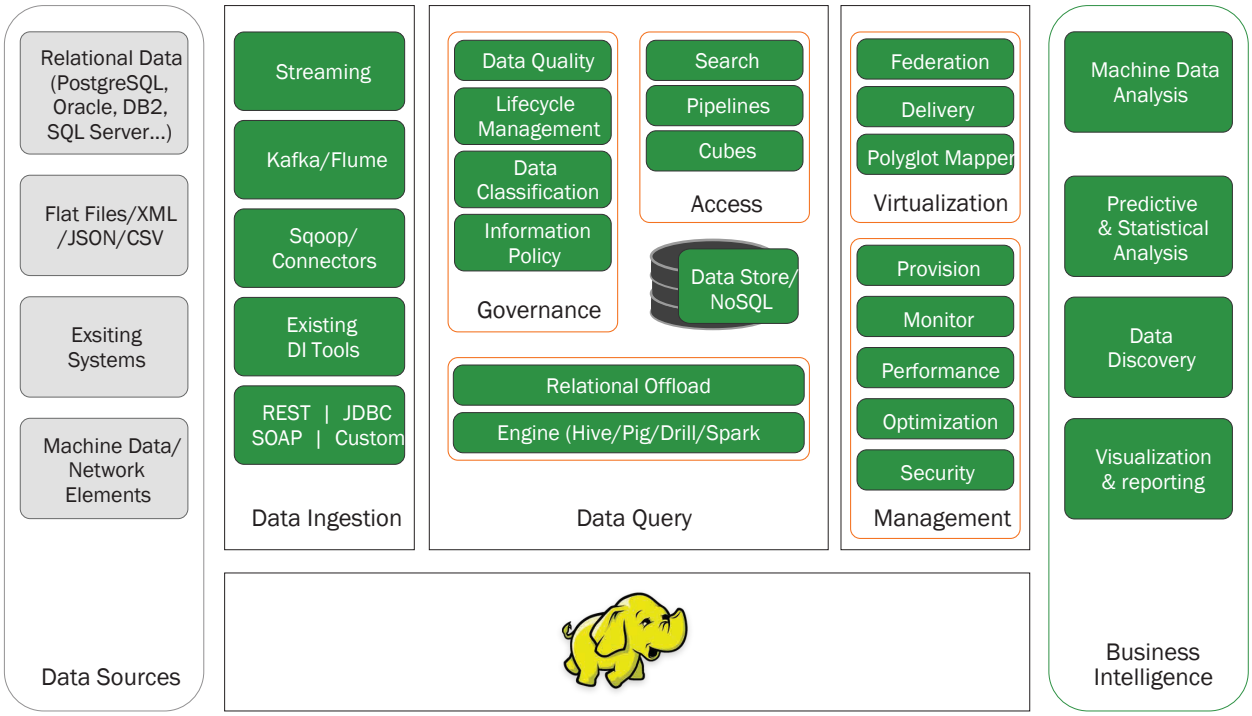
KEY DESIGN/STRUCTURING PRINCIPLES:

- **DISCOVERY WITHOUT LIMITATIONS**
- **LOW LATENCY AT ANY SCALE**
- **MOVEMENT FROM A REACTIVE MODEL TO PREDICTIVE MODEL**
- **ELASTICITY IN INFRASTRUCTURE**
- **AFFORDABILITY**

Step 4: Putting Together the Infrastructure — Inside the Data Lake Matrix

The successful installation of a Data Lake, requires persistence, detailing and attention to the many facets that must be kept in mind. Here are the key drivers, accelerators and tool-boxes.

We will begin with a diagram listing the major components of a big data warehouse:



Big Data: *warehouse reference architecture by Impetus*

A Data Lake structure basically consists of:

Data sources

- Relational database (such as Oracle, DB2, PostgreSQL, SQL Server and the like)
- Multiple disparate, unstructured and semi-structured data sources in formats such as flat files, XML, JSON or CSV
- Integration data in EDI or other B2B exchange formats
- Machine data and network elements for voluminous data generation

Hadoop distribution

- Most popular choice for big data today, Hadoop is available in open source Apache and commercial distribution packages
- Consists of a file system called HDFS (Hadoop distributed file system) — the key data storage layer of the big data warehouse

Data ingestion

- For data assimilation from various sources to the Hadoop file system — reliable and scalable data ingestion mechanisms
- For connecting relational database — Sqoop and database-specific connectors

- For data streaming — Apache Kafka and Flume
- For topology creation of streaming source, ingestion, in flight transformation and data persistence — common CEP (Complex Event Processing) or streaming engines such as Apache Storm or StreamAnalytix
- For existing Data Integration (DI) connector creation — custom scripts for integration using REST, SOAP or JDBC components

Data query:

- For the data resident in HDFS, a multitude of query engines such as Pig, Hive, Apache Drill and Spark SQL utilized
- Tools and solutions (from organizations like Impetus Technologies) to help enterprises offload expensive computing from relational data warehouses to big data warehouses

Data stores

- Data store coupling or NoSQL database like HBase, Cassandra in the big data warehouse for additional functions and capabilities

Access

- Complex access requirement management — includes features from the traditional world like search and cubes functions
- New tools to manage complex job pipelines (output of one query may be fed as input into another)

Governance

- Data quality maintenance
- Data quality regulation adherence (ensuring data lake does not turn into a data swamp)
- Data lifecycle management
- Comprehensive data classification
- Enterprise level information policy definition

Virtualization

- For consistent results with appropriate polyglot querying, data and delivery mechanism federation maintenance

Management

- Cluster monitoring - via tools and dashboards for cluster management
- Optimal speed and minimal resource consumption - via MapReduce jobs and query performance diagnosis

KEY POINTS TO REMEMBER:

- DATA SOURCE INTEGRATION
- HIGH VELOCITY HADOOP STORAGE DATA INGESTION
- MULTI-FORMAT DATA

KEY POINTS TO REMEMBER:

- ADVANCED QUERYING ENGINE EXPLORATION (STARTING WITH MAPREDUCE, AND MOVING ONTO APACHE SPARK, FLINK)
- USE CASE BUILDING FOR BOTH BATCH AND REAL TIME PROCESSING (USING STREAMING SOLUTIONS LIKE APACHE STORM AND STREAMANALYTIX)
- ANALYTIC APPLICATIONS FOR ENTERPRISE ADOPTION, EXPLORATION, DISCOVERY AND PREDICTION

KEY POINTS TO REMEMBER:

- DATA AND SPECIFIC CAPABILITY MANAGEMENT
- EDW AND DATA LAKE SYNERGY
- SPECIALIZED SOLUTIONS LIKE IMPETUS RELATIONAL OFFLOAD SOLUTION AID IN COST OPTIMIZATION

Business intelligence

- Various visualization and reporting tool deployment
- Data pattern discovery using predictive/statistical algorithms and machine data analytics

Step 5: Chart the Four Essentials Stages for Data Lake Creation

Building a robust Data Lake is a gradual movement. With the right tools, a clearly-planned platform, a strong and uniform vision and a quest for innovation, your organization can architect an integrated, rationalized and rigorous Data Lake repository. You can drive data potential, no matter how unwieldy it appears.

Here are the four key steps:

ONE: Scalable data handling and ingestion

This first stage involves creating a basic building block — putting the architecture together and learning to acquire and transform data at scale. At this stage, the analytics are simple, consisting of simple transformations; however, it's an important step in discovering how to make Hadoop work for your organization.

TWO: Analytical ability enhancement

The second stage focuses on enhancing data analysis and interpretation. Leveraging the Data Lake, enterprises employ several tools and frameworks to gradually combine and amalgamate the EDW and the Data Lake.

THREE: EDW and Data Lake collaboration

In the third stage, a wave of democratization takes the organization. A porous, all-encompassing data pool is created, allowing analytics and intelligence to flow freely across the company. This is also the stage where we witness a complete and seamless synergy between the EDW and the Hadoop-based Data Lake, absorbing the strengths of each architecture.

FOUR: End-to-end adoption and maturity acquisition

The fourth stage is the final and highest stage of maturity: This ties together enterprise capabilities and large-scale unification covering information governance, compliance, security, auditing, metadata management and information lifecycle management capabilities.

KEY POINTS TO REMEMBER:

- TOOLS AND UTILITIES INTRODUCTION
- CONSISTENT DATA DISCOVERY
- INSIGHT MANAGEMENT
- INFORMATION LIFECYCLE MANAGEMENT

BREAK DOWN INFORMATION SILOS:

- TRANSITION FROM SCHEMA-ON-WRITE TO SCHEMA-ON-READ TO MATCH THE CHALLENGES POSITED BY DIVERSE AND COMPLEX DATA SETS
- STORE AND UTILIZE AND LEVERAGE DATA WITHOUT FORMATTING AND STRUCTURING
- ELIMINATE THE UPFRONT COSTS OF DATA INGESTION
- BENEFIT FROM THE DATA LAKE'S CAPACITY TO EXPAND EASILY
- REJECT THE LIMITATIONS IMPOSED BY DATA STRUCTURING, MODELING AND NORMALIZATION AND EDW REQUIREMENTS

Summary

Organizations are increasingly attempting to innovate and digitally refine processes, driving heightened service excellence and delivery quality. While the traditional EDW continues to occupy an important position within the scheme of things — faster, simpler and leaner ideas are the obvious next step. Data Lakes represent a smarter opportunity for effective data management, maintenance and usage.

Concluding Thoughts: Capitalize on Big Data Opportunities

As new technology replaces the old, a buzz of anticipation, excitement, uncertainty and even apprehension surrounds each shift.

In recent times, disruptive big data technologies, the Hadoop eco-system and the widening Data Lake storage facility have been some of the most talked about tech-innovations — each unlocking a minefield of potential and possibility. However, every act of evolution is a gradual process, a motion defined by the strength of a company's belief system, its ability to challenge the status quo, and its stringent commitment to planning and preparation.

Intelligent, effective and conscientious data lake development can be a true driver of value and capability — tapping the incredible power of information to envision real insight.

How We Help

Impetus Technologies specializes in this specific area of tech-advancement and has delivered unique and customized solutions to many customers, cementing its place at the top of this evolving ecosystem. With our robust product experience and service expertise available at every stage of the big data transformational curve, you can now calibrate your organization-wide change mechanism.

We help you turn information into a veritable goldmine, a new-age competitive asset, a secret ingredient, propelling revenue and growth like never before. We enable you to:

- WORK WITH UNSTRUCTURED DATA
- INITIATE DEMOCRATIZED DATA ACCESS
- CONTAIN COSTS WHILE CONTINUING TO DO MORE WITHMORE DATA

About Impetus

Impetus is focused on creating big business impact through Big Data Solutions for Fortune 1000 enterprises across multiple verticals. The company brings together a unique mix of software products, consulting services, Data Science capabilities and technology expertise. It offers full life-cycle services for Big Data implementations and real-time streaming analytics, including technology strategy, solution architecture, proof of concept, production implementation and on-going support to its clients.

Visit www.impetus.com or write to us at bigdata@impetus.com