**Experiment No. 6**

**Title: Case Study on Data Lake**

**Batch: B2 Roll No: 1914078 Experiment No.:6**

### Aim: To prepare a case study on Data lake

Resources needed: MS Office

### Theory:

 A data lake is a central storage repository that holds [big data](https://www.talend.com/resources/future-big-data/) from many sources in a raw, granular format. It can store structured, semi-structured, or [unstructured data](https://help.talend.com/reader/DDfDbx8_eDYPoPp_19Gqag/JVLZEPwmPaYt43XC7vH7aQ), which means data can be kept in a more flexible format for future use. When storing data, a data lake associates it with identifiers and metadata tags for faster retrieval.

Coined by James Dixon, CTO of Pentaho, the term “data lake” refers to the ad hoc nature of data in a data lake, as opposed to the clean and processed data stored in traditional [data warehouse](https://www.talend.com/resources/what-is-data-warehouse/) systems.

[Data lakes](https://www.stitchdata.com/resources/what-is-data-lake/) are usually configured on a cluster of inexpensive and scalable commodity hardware. This allows data to be dumped in the lake in case there is a need for it later without having to worry about storage capacity. The clusters could either exist on-premises or in the cloud.

**Benefits of a data lake:**

A data lake works on a principle called *schema-on-read*. This means that there is no predefined schema into which data needs to be fitted before storage. Only when the data is read during processing is it parsed and adapted into a schema as needed. This feature saves a lot of time that’s usually spent on defining a schema. This also enables data to be stored as is, in any format.

Data Scientists can access, prepare, and analyze data faster and with more accuracy using data lakes. For analytics experts, this vast pool of data — available in various non-traditional formats — provides the opportunity to access the data for a variety of use cases like sentiment analysis or fraud detection.

### Activities:

### Challenges faced by organization before using Data lake

### Implementation of Data lake

### Architecture/Framework used

### Features

### Usages

### 3. Solutions provided by Data lake

### Result:

### Data Lake Architecture | Components & Purpose of Data Lake in Business

**THE CHALLENGE**

The fund has an initiative to make both internal and external data sources available to the entire organization to speed core services execution and to use data for performing data analytics by investment teams. The approach to data the fund took historically was to implement local solutions that solve individual business needs or manual efforts by each independent team, resulting in a fragmented approach and a large number of silo’d data ponds. Implementation of a data strategy began a few years prior centered around an on-premise approach to centralizing in-house data around a third party product. Subsequently, a canvassing of data related needs across investment departments and core services groups identified a need for a technology capability that provided visibility and timely access to all existing data and allowing for introduction of new datasets with relative ease, along with a data operating model and governance that supported the sharing of data across the fund.

#### The challenges of data lakes in managing data

##### Issues with security and governance.

Data lakes are an open-source of knowledge designed to streamline the analytics pipelines. However, the open nature of the lake makes it difficult to implement security standards. The open nature of the lake and the rate data is inputted, makes it difficult to regulate the data coming in. To eliminate this problem, data lake designers should work with data security teams to set access control measures and secure data without compromising loading processes or governance efforts.

However, it’s not just security that’s causing problems with data lakes. It’s also an issue of quality. Data lakes collect data from different sources and pool it in a single location, but the process makes it difficult to check data quality. It is problematic because it leads to inaccurate results when the data is used for business operations. When the data is inaccurate, the findings will be inaccurate, causing a loss of confidence in the data lake and even in the organisation. To resolve this problem, there needs to be more collaboration between data governance teams and data stewards so that data can be profiled, quality policies implemented and have action taken to improve quality.

##### Meta management becomes impossible

[Metadata management](https://whatis.techtarget.com/definition/metadata-management) is one of the most important parts of data management. Without metadata, data stewards (those who are responsible for working with the data) would have little choice but to use non-automated tools like Word and Excel. Moreover, data stewards spend most of the time working with metadata, as opposed to actual data. However, metadata is not implemented on data lakes, which is a problem, in terms of data management. The absence of metadata makes it difficult to perform vital big data management functions like validating it or implementing organisational standards. Since there is no metadata management, it becomes less reliable, hurting its value to the organisation.

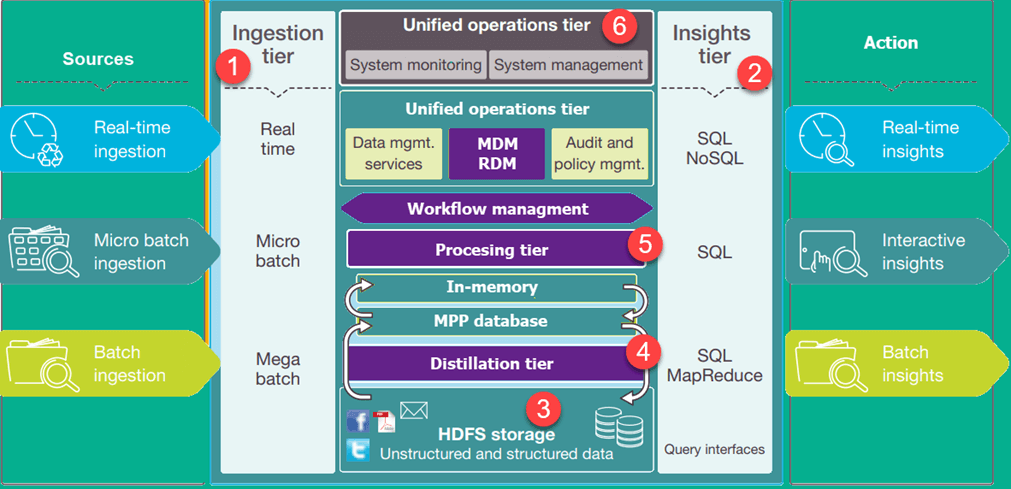
##### Conflict in the organisation hinders full value

Data lakes are incredibly useful, but they are not immune to clashes within the organisation. If the organisation’s structure is plagued with red tape and internal politics, then little value can be derived from the lake. For example, if data analysts cannot access the data without obtaining permission, then it holds up the process and hurts productivity. Different departments might also have rules for the same data set, leading to differences in rules, policies and standards. This situation can be somewhat mitigated by having a robust data governance policy in place to ensure consistent data standards across the whole organisation. While there is no denying the value of data lakes, there need to be better governance standards to improve management and transparency.

##### Identifying data sources is difficult

Identifying data sources in a data lake is not often done, which is a problem in big data management. Categorising and labelling data sources is crucial because it prevents several problems like duplication of data. Yet, this is not done regularly, which is problematic. At the very least, the source of metadata should be recorded and available to users.

## Data Lake Architecture

[](https://www.guru99.com/images/1/022218_0459_WhatisDataL2.png)

The figure shows the architecture of a Business Data Lake. The lower levels represent data that is mostly at rest while the upper levels show real-time transactional data. This data flow through the system with no or little latency. Following are important tiers in Data Lake Architecture:

1. **Ingestion Tier**: The tiers on the left side depict the data sources. The data could be loaded into the data lake in batches or in real-time
2. **Insights Tier:**The tiers on the right represent the research side where insights from the system are used. SQL, NoSQL queries, or even excel could be used for data analysis.
3. **HDFS**is a cost-effective solution for both structured and unstructured data. It is a landing zone for all data that is at rest in the system.
4. **Distillation tier** takes data from the storage tire and converts it to structured data for easier analysis.
5. **Processing tier** run analytical algorithms and users queries with varying real time, interactive, batch to generate structured data for easier analysis.
6. **Unified operations tier** governs system management and monitoring. It includes auditing and proficiency management, data management, workflow management.

## Key Data Lake Concepts

### Data Ingestion

Data Ingestion allows connectors to get data from a different data sources and load into the Data lake.

Data Ingestion supports:

* All types of Structured, Semi-Structured, and Unstructured data.
* Multiple ingestions like Batch, Real-Time, One-time load.
* Many types of data sources like Databases, Webservers, Emails, IoT, and FTP.

### Data Storage

Data storage should be scalable, offers cost-effective storage and allow fast access to data exploration. It should support various data formats.

### Data Governance

Data governance is a process of managing availability, usability, security, and integrity of data used in an organization.

### Security

Security needs to be implemented in every layer of the Data lake. It starts with Storage, Unearthing, and Consumption. The basic need is to stop access for unauthorized users. It should support different tools to access data with easy to navigate GUI and Dashboards.

Authentication, Accounting, Authorization and Data Protection are some important features of data lake security.

### Data Quality:

Data quality is an essential component of Data Lake architecture. Data is used to exact business value. Extracting insights from poor quality data will lead to poor quality insights..

### Data Auditing

Two major Data auditing tasks are tracking changes to the key dataset.

1. Tracking changes to important dataset elements
2. Captures how/ when/ and who changes to these elements.

Data auditing helps to evaluate risk and compliance.

### Data Exploration

It is the beginning stage of data analysis. It helps to identify right dataset is vital before starting Data Exploration.

All given components need to work together to play an important part in Data lake building easily evolve and explore the environment.

**Need of Data Lake :**

1. Scalability : The data lake analogy was conceived to help bring a common and visual understanding to the benefits of distributed computing systems able to handle multiple types of data, in their native formats, with a high degree of flexibility and scalability. While the analogy might not be perfect, the goal of a data lake is certainly well-aligned with the challenges so many companies struggle with today.
2. Performance and quality : Not only is the complexion of data changing in today’s business environment, but the face of the traditional data user is evolving as well. Decision makers today, in a variety of different job roles, have heightened expectations when it comes to the data that fuels their most important decisions. At the same time though, quality and latency issues plague many of today’s organizations. When comparing similar companies in the market for a data lake, those who have deployed the technology are more likely to report user satisfaction when it comes to critical metrics like quality and timeliness.

**THE SOLUTION**

The solution was to create a data ecosystem (data lake and data pipeline) that would be central to providing investment departments and service groups across the fund with access to core data domains and new sources of data. Based upon the technical needs to enable the data lake, we envisioned a cloud-based platform that would accommodate different types of data and compute needs that were most relevant to the business – a data pipeline that ingests and normalizes data to a data lake which acts as a central repository, discovery through a data catalog, data access methods to support disparate needs integrated with a cloud compute environment for applications and analytics.

Amazon Web Services (AWS) was chosen as the cloud provider. IData was engaged to drive the process to build and support the AWS technology foundation, security implementation, continuous integration/continuous delivery (CI/CD) pipeline, and create the application stack to support the data lake and ETL ingestion pipeline.

Features of the solution provided to the company are as follows:

* **Cloud-based Account / Project Structure**

The proper root account / sub-account / project structure was implemented to achieve huge gains in productivity, innovation, and cost reduction as the pension fund migrated to the AWS cloud. There are a variety of services and features that allow for flexible control of cloud computing resources and also of the AWS account(s) managing those resources. On the account level, these options are designed to help provide proper cost allocation, agility, and security. A project-based mapping one-to-one to a sub-account structure was implemented. Creating a security relationship between sub-accounts was a key element added to assess the security of cloud-based deployments, centralize security monitoring and management, manage identity and access, and provide audit and compliance monitoring services.

* **Project-based Implementation with Infrastructure as Code (IaC)**

Infrastructure as Code (IaC) was implemented as a method to provision and manage IT infrastructure through the use of source code, rather than through standard operating procedures and manual processes. IaC helps the fund to automate the infrastructure deployment process in a repeatable, consistent manner, also providing the benefit to easily deploy standard infrastructure environments in other regions where the cloud provider operates so they can be used for backup and disaster recovery.

* **Serverless Compute and Storage**

By employing cloud serverless compute and storage, such as Lambdas and object-stores (S3), the fund leverages the ability to build and run applications and services and with infinite elasticity without using physical hardware. In addition, all existing costs associated with managing servers and containers (operating system updates, maintenance updates, image snapshots, backups, restarts, etc.) largely disappeared.

* **ETL and the Data Pipeline**

The data pipeline acts as a utility – a standard suite of data tools that enabled the fund to automate the sourcing, processing, and entitlement of data. Automation of these processes allows data sources to be quickly added and the approach for the cloud data lake then extracted, transformed, combined, validated and loaded (ETL) for further use. The data pipeline is able to simultaneously process multiple data sources at once.

* **Enterprise Data Lake**

The introduction of an enterprise data lake provided a central data repository and access to analytics tools that maximized the value of the data. The enterprise data lake is a centralized repository that allows storage of structured and unstructured data at any scale. Data can be stored as-is, without having to first structure the data, and run different types of analytics – from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions.

* **Data Catalog and Discover-ability**

The introduction of a data catalog provided a single searchable glossary of data that is available to the organization, including the data source, definition and entitlements. The data catalog built on top of the data lake allows users to find the data they need and then use it in the tools that they prefer along with ensuring information boundaries and data contracts are not violated.

* **Application Programming Interface**

An Application Program Interface (API) is a set of software routines that allow programs to interact. The use of an API at the pension fund allowed data from the enterprise data lake to be accessed by upstream applications that rely on it. Additionally, the API enabled end users access to data for their individual analytics and modelling needs.

* **Analyze and Visualize Data**

Through self-discovery of data resident in the enterprise data lake through the data catalog, individuals are able to access data based on their entitlements. For low technology use cases, end users are able to upload datasets into Excel or tools such as Tableau, or alternatively, the API allows data to be integrated with programs coded in Python, Scala, Java, R, etc.

* **Sandbox for Experimentation**

To support innovation, the data lake includes a sandbox environment that provides the functionality of the enterprise data lake but allows one to easily introduce new datasets and technologies for experimentation.

* **Flexibility of Architecture**

Unlike traditional technical approaches for data warehousing which are inflexible in terms of data schemas, technical capabilities and tools, the cloud based approach allows flexibility on all of these fronts. Fit for purpose cloud based data tools and technologies can be incorporated with relative ease as needs get identified.

## Benefits of using Data Lake:

Here are some major benefits in using a Data Lake:

* Helps fully with product ionizing & advanced analytics
* Offers cost-effective scalability and flexibility
* Offers value from unlimited data types
* Reduces long-term cost of ownership
* Allows economic storage of files
* Quickly adaptable to changes
* The main advantage of data lake is the **centralization**of different content sources
* Users, from various departments, may be scattered around the globe can have **flexible access** to the data
* The creation of an enterprise data lake had substantial benefits. Historically, the fund had data that existed in individual business teams or systems. Providing access to this data required point-to-point solutions and significant time was spent preparing and reconciling data by each team who uses it. Additionally, teams were not aware of data that existed within the organization. The enterprise data lake enabled fund teams to capitalize on the value in data, by bringing together internal and external datasets in a single place as well as eliminating redundant reconciliations by using the same dataset. This value grows as new data sets are added. This easy access to a broad set of data empowers users to innovate the way the fund constructs and measures portfolios, manages assets, and thinks about risk.
* By leveraging the data catalog, investment groups at the fund can now easily discover, access and perform analytics on hundreds of datasets. After discovery, analysts can access data directly via an API into their environment, or they can leverage sophisticated scalable cloud based analytics software, such as Spark based services, to perform intense algorithms against multiple datasets. This greatly speeds analytics and increases its use in investment decision making.
* The implementation of a Continuous Integration/Continuous Deliver (CI/CD) approach to cloud development and deployment, the fund can deliver new software features in hours or days instead of months. Smaller code changes are simpler (more atomic) and have fewer unintended consequences. Upgrades introduce smaller units of change and are less disruptive. The products improve rapidly through fast feature introduction and fast turn-around on feature changes. End-user involvement and feedback during continuous development leads to usability improvements. You can add new requirements based on customer’s needs on a daily basis.

### Outcome:

### Understanding of data warehouse and its multi-dimensional modeling

**Conclusion:** (Conclusion to be based on the objectives and outcomes achieved)

Hereby,we have learned the uses and benfits of data lake in corporate world.

**Grade: AA / AB / BB / BC / CC / CD /DD**

**Signature of faculty in-charge with date**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**References:**

* 