CONTRIBUTOR INSIGHTS

# Data Pipelines for Engineered Decision Intelligence

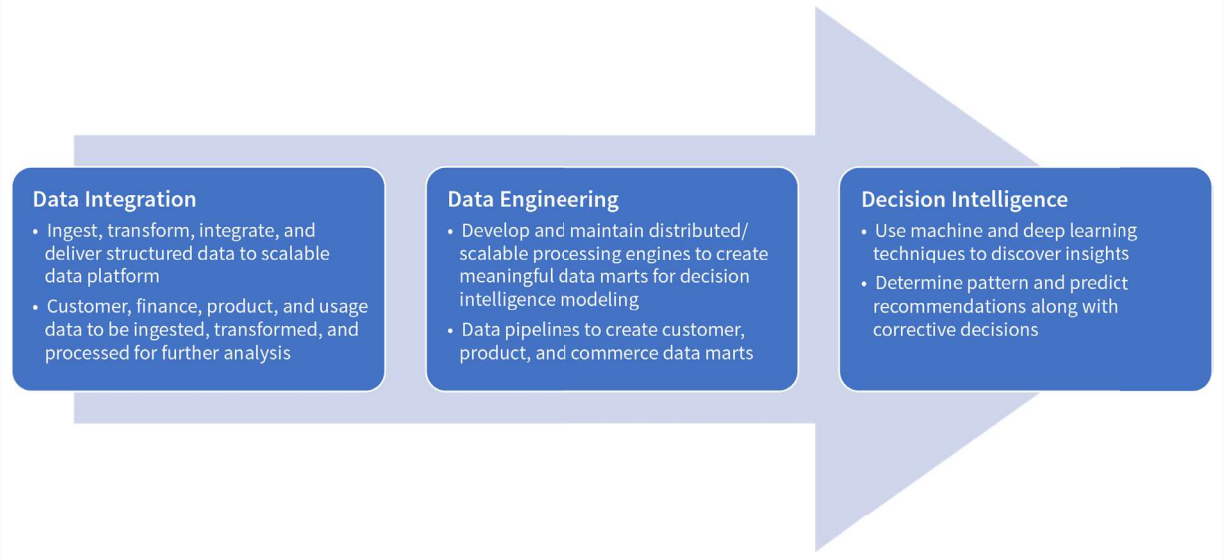
**By Dr. Tuhin Chattopadhyay, Founder & CEO of Tuhin AI Advisory**

Data science has reached its peak through automation. All the phases of a data science project — like data cleaning, model development, model comparison, model validation, and deployment — are fully automated and can be executed in minutes, which earlier would have taken months. Machine learning (ML) continuously works to tweak the model to improve predictions. It's extremely critical to set up the right data pipeline to have a continuous flow of new data for all your data science, artificial intelligence (AI), ML, and decision intelligence projects. Decision intelligence (DI) is the next major data-driven decision-making technique for disruptive innovation after data science. It is:

- **Futuristic** – Models ML outcomes to predict social, environmental, and business impact.
- **Holistic** – Meaningfully integrates both managerial and behavioral perspectives.
- **Realistic** – Models all contextual variables and real-life constraints.

So it's more important for DI projects to have a robust data pipeline. They need a continuous inflow of the right data with the right velocity to get stored in the right container and subsequently processed correctly for model development to generate actionable insights.

**Figure 1: Enterprise decision intelligence**

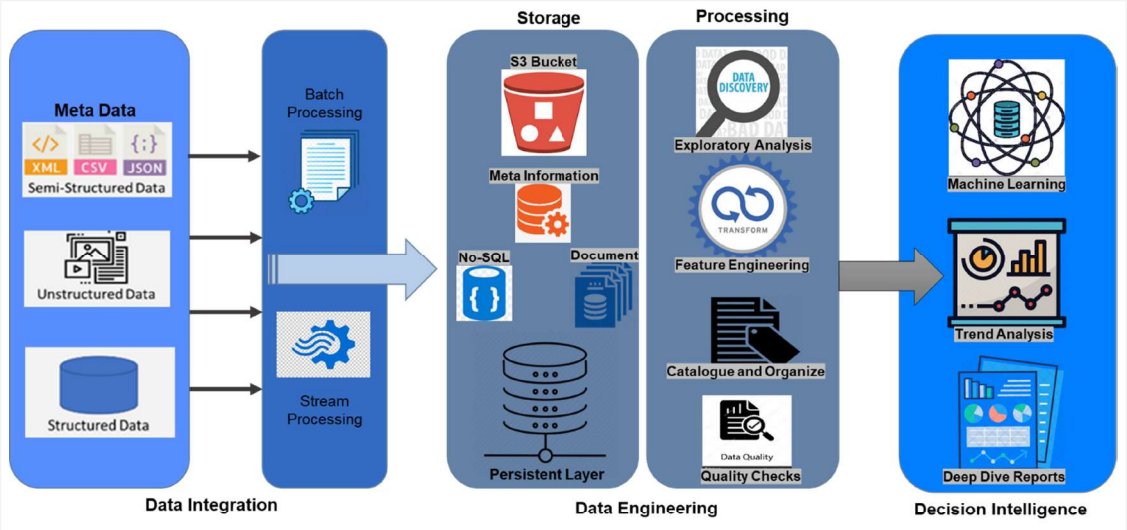



## Developing a Data Pipeline

The first phase of developing a data pipeline is **data integration**: ingests various customer, product, and usage data for processing and analysis. There are two stages of data integration. The most fundamental step is to identify the right sources of both internal data — comprising IoT, CRM, ERP, OLAP, Excel reports, etc. — and external data, like Facebook, Twitter, and statistical databases. The second step of data integration is to gather semi-structured, unstructured, and structured data through batch processing and stream processing.

After obtaining the data through integration, the next phase is **data engineering**, which involves storing and processing data for further model development. The objects comprising the files and metadata can be uploaded in any container that can store diverse unstructured, hierarchical, and structured data. Processing stored data includes data sanitization, feature engineering, and splitting the data for training and testing before sending it for model development through ML, deep learning, and natural language processing techniques.

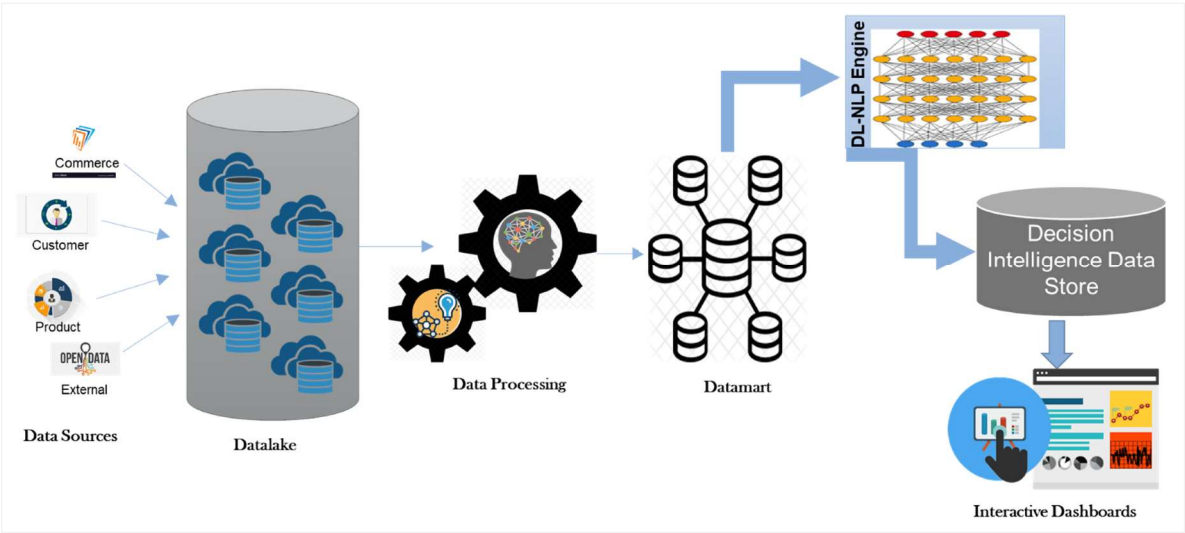
**Figure 2: Data engineering framework for decision intelligence**




In the last stage, data is sent for **decision intelligence** model development. Some of the most popular modeling techniques for DI include decision modeling and simulation, optimization and game theory, system dynamics and systems modeling, sensitivity and scenario analyses, knowledge management, hidden Markov models, and Markov Chain Monte Carlo. Advanced modeling techniques like quantum Bayesian networks (QBNs) with directed acyclic graphs, data-driven predictions of the Lorenz attractor, and intelligence augmentation work on top of ML outcomes to figure out the decision impact on society, business, and environment. Last but not least, the final outcome can be presented through interactive dashboards that can easily be used for managerial decisions.

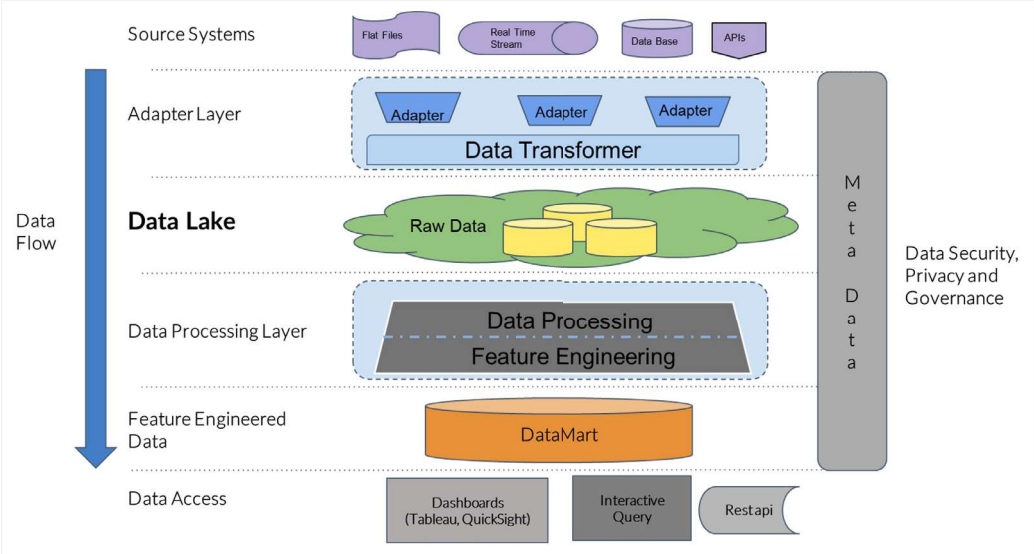
The data architecture is critical in setting up the data pipeline for DI. Traditionally, structured data is stored in a data warehouse for data discovery and querying. With the advent of semi-structured clickstream data, a data lake became the natural choice to hold vast amounts of raw data. A data lakehouse is a hybrid approach, in which a warehouse layer resides on the top of a data lake to store both structured and unstructured data. After processing the data, the feature-engineered data gets stored in a data mart before it finally flows to the DI engine for model development.

**Figure 3: Data architecture**



Below is an instance of platform architecture that demonstrates an increased level of abstraction for applied solution development and usage scenarios.

**Figure 4: Data platform architecture**




To illustrate this architecture in the cloud, Figure 5 represents a data pipeline over an AWS ecosystem, which shows the applicable AWS components in every step from data extraction to dashboarding. The dashboarding can be accomplished through any non-AWS tools like Tableau.

**Figure 5: Key data engineering components in AWS**


## Considerations for Data Quality, Governance, Privacy and Security

While setting up the data pipeline for any DI project, it's critical to prioritize data quality, data governance, data privacy, and security issues.

### DATA QUALITY

Reliable and consistent input data is critical for eliminating errors and biases. Therefore, impeccable quality of the data is sacrosanct for any DI project.

The following checkpoints can be used to measure data quality:

- **Completeness** – Is all necessary data available and accessible?
- **Consistency** – How consistent is the data across different systems that hold instances of the data?
- **Validity –** Measures whether a value conforms to a pre-defined standard.
- **Accuracy** – How correctly and accurately is the data presented?
- **Uniqueness –** A discrete measure of duplication of identified data items.
- **Timeliness** – A measure of time between when data is expected against when data is made available.

### DATA GOVERNANCE

The cornerstone of any successful DI project is collaboration in correctly framing the problem and estimating the implications of actionable insights. A data governance framework enables this by assigning responsibilities to people, processes, contributors, and technology that make decision-making easier. As far as the command and control for data governance is concerned, the framework designates a few employees as data stewards. Their responsibilities include determining answers to the following questions:

- **Where** is the data?
- Who should **access** it?
- What does the data **contain**?
- What is the data **quality**?
- How can the data adhere to **compliance**?
- How **secure** is the data?

### DATA PRIVACY AND SECURITY

Data privacy is critical for any DI project, and businesses must ensure compliance with all relevant data protection laws and regulations, such as PDP and GDPR (intended to protect security, privacy, PII, etc.). Privacy should also be guaranteed across data collection, processing, sharing, and deletion. Data security issues can be addressed through:

- **Authorization** – Implement measures to prevent any unauthorized access by third parties.
- **Encryption –** Encrypt data at flight and at rest, masked for PII.
- **Penalty** – Adopt and enforce huge penalties for breaches of data security.

## Parting Thoughts

As we move forward, decision intelligence will connect the ML outcomes of data science projects with businesses at first, and then with society at large. Data integration and data engineering are the key components of an enterprise DI project. Both the data lake and data lakehouse have become the industry's natural choice as they can store semi-structured and unstructured data that is easy to retrieve and model. Besides the traditional ML model, many sophisticated optimization techniques are used for developing DI models. Cloud-native computing seamlessly drives the entire operations from data integration and model development to interactive dashboards for visualization and decision-making.

Ultimately, the key to a successful decision intelligence project boils down to ensuring data quality, governance, privacy, and security remain priorities at every step of the process.


**Dr. Tuhin Chattopadhyay, Founder & CEO of Tuhin AI Advisory**

@tuhinc on DZone | @tuhinai on LinkedIn | tuhin.ai

Tuhin spent the first 10 years of his career in academia and research, teaching business statistics, analytics, and technology at several reputed B-Schools of India. As a corporate practitioner, Tuhin has a proven record of accomplishment as a transformational leader in organizations like The Nielsen Company. Currently, he runs his own consultancy for providing a full suite of AI, analytics, CTO/CAO-as-a-Service, and digital transformation services to clients.