Intellias > Blog > Data & Analytics > Data Engineering: An Essential Element of Your Big Data Strategy

Blog post

Data Engineering: An Essential Element of Your Big Data Strategy

In a world reliant on big data, its collection and storing has become vital for businesses striving to stay ahead of the curve

Updated: June 09, 2023 • 6 mins read | Published: October 06, 2020

Gone are the days when virtually all data management practices could be squeezed into a single database administrator's job description. Over the past decade or so, databases have migrated to the cloud, gained never-before-seen performance and complexity, and evolved into data warehouses and data lakes to address the growing need for ultra-fast data aggregation and instant availability.

In this new reality, the role of former database administrators has also changed dramatically. According to big data best practices, the process now requires these roles:

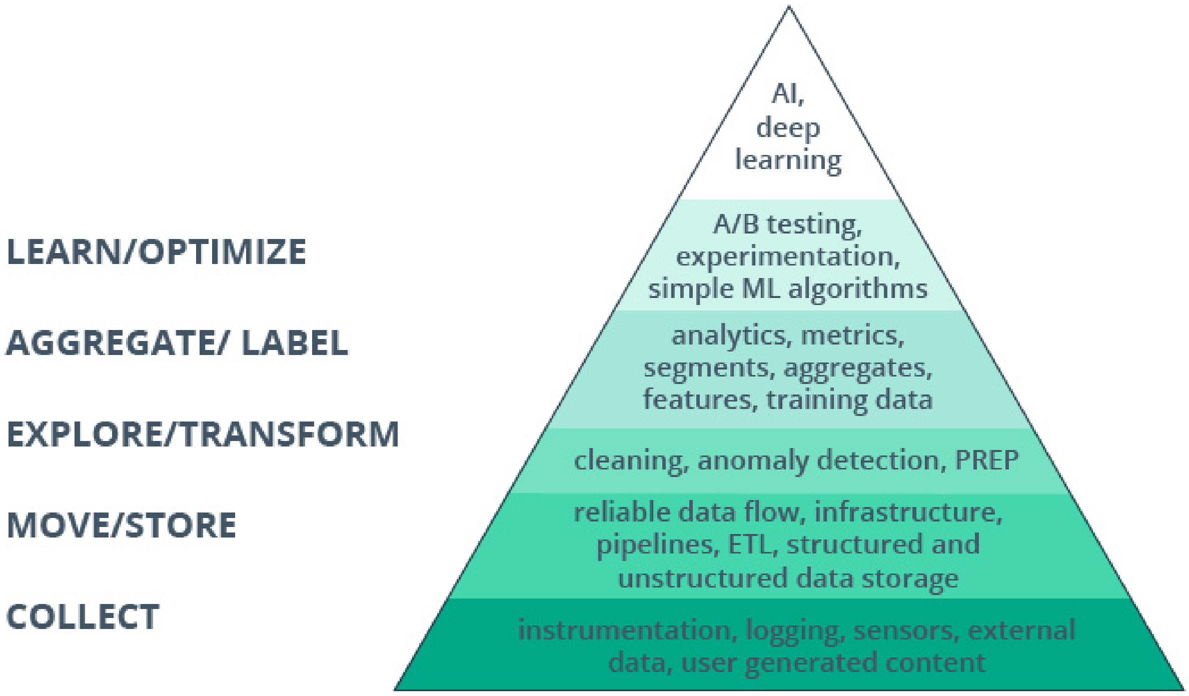
- Data engineers
- Data analysts
- Data scientists

Contrary to popular belief, these roles have very distinct boundaries, and while a data engineer may pack some analytical skills, a data scientist is very unlikely to do a data

Data engineering in a nutshell

To answer the question above, let’s take a look at the data science stack. At the very top of the food chain are data scientists, people who apply their extensive knowledge of artificial intelligence and deep learning frameworks to draw advanced insights from layers and streams of data and build efficient, self-sustaining forecasting mechanisms.

The data science hierarchy of needs


AI, deep learning
LEARN/OPTIMIZE
A/B testing, experimentation, simple ML algorithms
AGGREGATE/ LABEL
analytics, metrics, segments, aggregates, features, training data
EXPLORE/TRANSFORM
cleaning, anomaly detection, PREP
MOVE/STORE
reliable data flow, infrastructure, pipelines, ETL, structured and unstructured data storage
COLLECT
instrumentation, logging, sensors, external data, user generated content

Source: Monica Rogati

One tier below them are analysts, experts in data manipulation. They perform a number of fairly complex business intelligence (BI) operations:

- Selection of relevant data sets
- Preparation of selected data sets for analysis (clean-up, sorting, etc.)
- Search for data patterns in data sets
- Data visualisation
- Business analysis
- Reporting and creation of the optimization roadmap

Finally, the foundation of the pyramid rests on data engineers. These are the people that make all of the above possible. In many ways, their contribution to the process is even

Data engineers design, build, and maintain the vast and complex infrastructure required for data collection and storage. They are responsible for creating and maintaining the data pipeline that enables organization to capture unstructured, raw data from multiple disparate inputs, process it accordingly, and store it in a way that guarantees free, fast, and unobstructed access for data analysts and scientists.

Learn how we upgraded a client's network of retail data acquisition services

Let us now dive deeper into the specifics of data engineering practices and explain their paramount importance to any company executing a big data roadmap.

The importance of data engineering

Experts estimate the global big data implementation and data engineering market to hit the $77.37 billion mark by 2023. The wide proliferation of intelligent platforms, such as high-frequency trading systems and global eCommerce platforms, calls for the implementation of big data analytics systems meeting the most stringent performance and resilience requirements.

However, it's not just about cutting-edge solutions for large enterprises. Today, even small businesses may be consuming vast amounts of data coming from external systems, users, field teams, sensor arrays and other sources. As companies grow and the number of sources and data types multiply, the task of processing these streams without delays and data loss becomes extremely challenging.

Without data engineering, implementation of big data initiatives and fulfillment of the overall big data strategy is impossible:

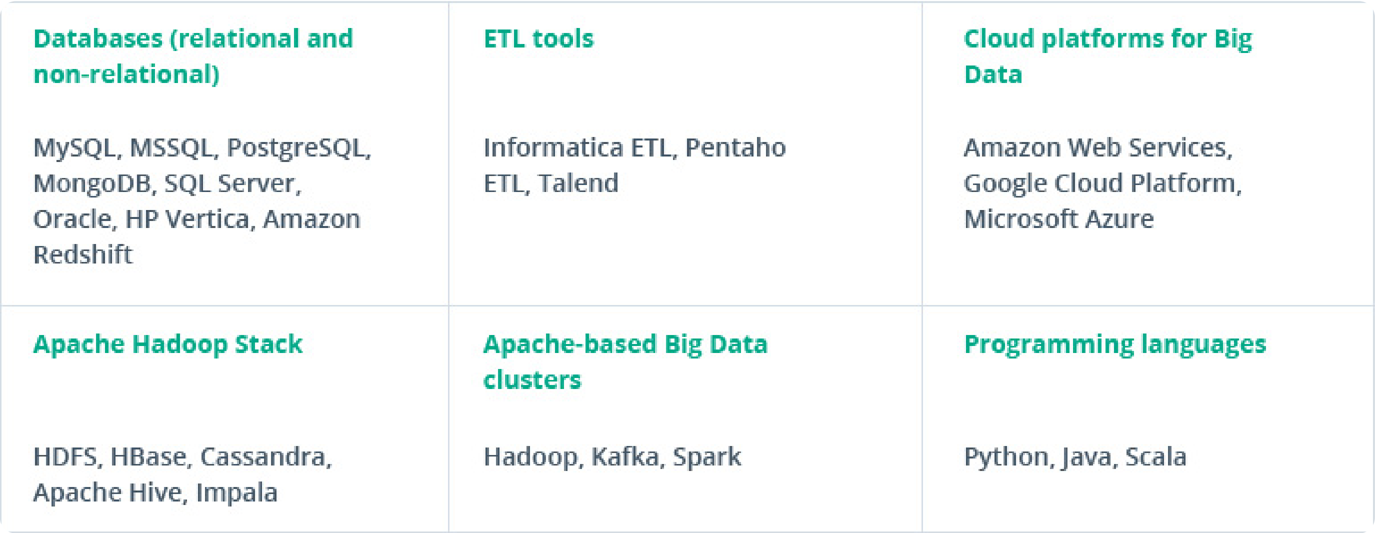
No data means no data analytics and no data science.

Delayed data translates into wrong, untimely decisions.

Fragmented data results in inaccurate measurements, flawed data models, and low-quality forecasting.

rules, and ultimately send it to designated storage destinations. From this point onward, the ball is in the court of analysts and scientists.

Depending on the specifics of a particular business application and big data project plan, engineers can use a variety of technologies and tools:



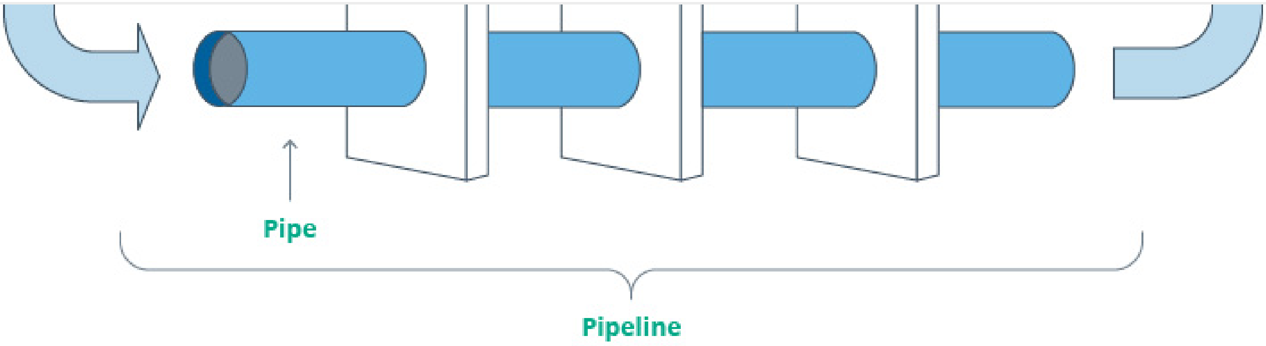
Databases (relational and non-relational)	ETL tools	Cloud platforms for Big Data
MySQL, MSSQL, PostgreSQL, MongoDB, SQL Server, Oracle, HP Vertica, Amazon Redshift	Informatica ETL, Pentaho ETL, Talend	Amazon Web Services, Google Cloud Platform, Microsoft Azure
Apache Hadoop Stack	**Apache-based Big Data clusters**	**Programming languages**
HDFS, HBase, Cassandra, Apache Hive, Impala	Hadoop, Kafka, Spark	Python, Java, Scala

It’s important to understand that tools alone don’t get the job done. Ensuring an uninterrupted flow of data, its automatic conversion and transformation requires a wide outlook on the business needs of the company, a thorough understanding of its infrastructure, and an ability to construct a flexible and scalable framework feeding perfectly structured, clean data outside. In addition, it is typically assumed that data engineers are responsible for data security, integrity, and the overall support and maintenance of the pipeline.

All of the above, combined, makes the job of a data engineer a vital element of any company’s big data implementation plan, as demonstrated by a recent LinkedIn job market report, which placed Data Engineers as 8th on the list of the most popular emerging jobs.

Setting up your data pipeline

In the most generic sense, a data pipeline is a multi-component software system composed of various automated ETL tools, scripts, and programs that receives data from one or more inputs, processes it, and then sends it to a corresponding destination. The main purposes of a data pipeline are the uninterrupted flow of data and its 24/7 availability for users (as a rule, analysts and data scientists).



Data pipelines can be built in a multitude of different ways and with a varying degree of custom code. Some pipelines can be set up using off-the-shelf workflow automation tools like Apache Airflow or Azkaban, while others may require a more custom approach. They might require an enormous amount of programming in Python using such frameworks as Luigi, for example.

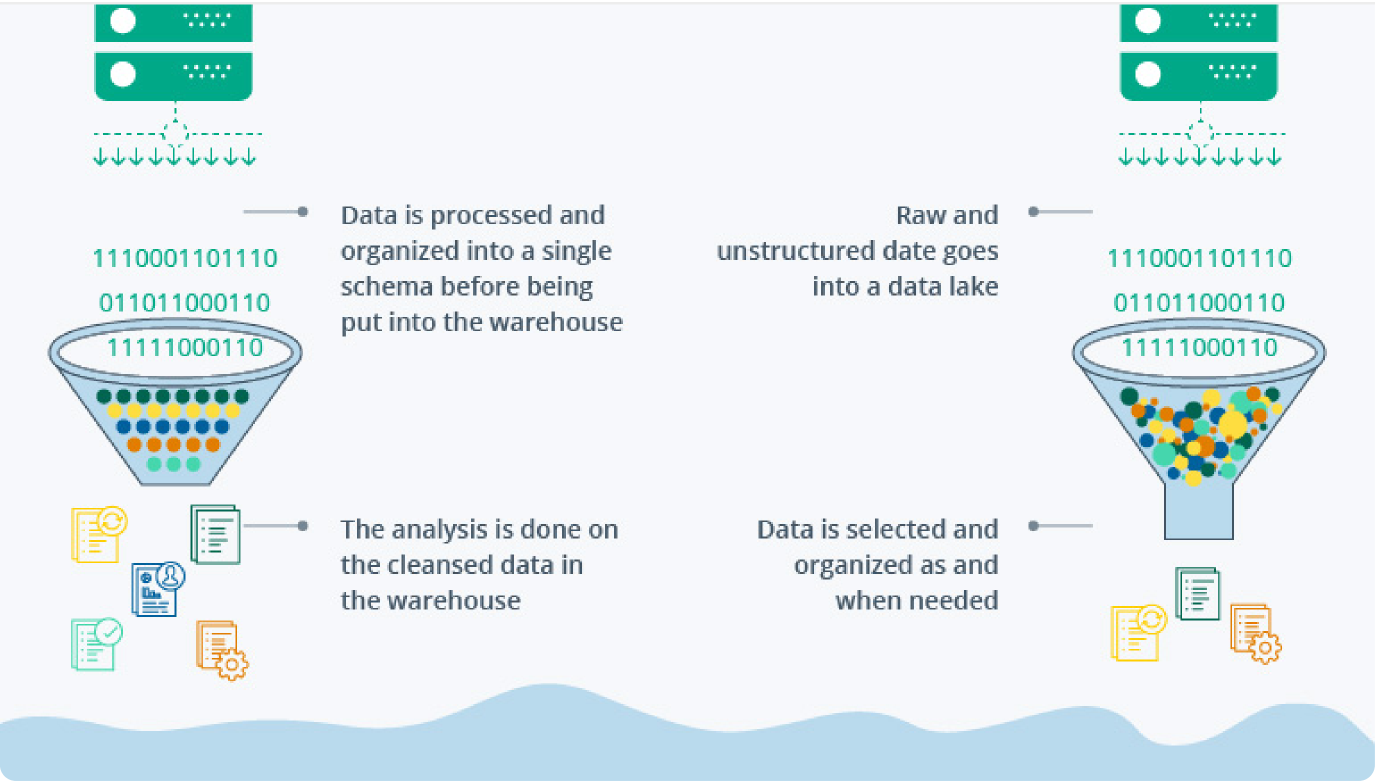
Whichever case you choose, be prepared to deal with a number of fairly non-trivial tasks:

- Coming up with a way to monitor data channels and capture incoming data in various formats
- Converting and transforming data captured from individual sources into the format and schema of corresponding destinations
- Saving the data to the target database/data warehouse/data lake
- Creating a reliable mechanism for modifying fields and schemas to reflect changes in the business logic
- Maintaining and optimizing the data pipeline

Implementation of big data pipelines requires a fair amount of effort, so if your company relies on data collection and analysis on a permanent basis, it makes perfect sense to hire a skilled data engineer or find a trusted vendor who is well-versed in big data best practices and offers professional support services.

Data warehouses vs. data lakes

Data warehouses and data lakes represent two different approaches to storing and using data. Warehouses represent a more conventional approach where data is stored in a centralized enterprise repository used primarily for reporting and data analysis. New data is written to the repository by operational (source) systems in strict accordance with


1110001101110
011011000110
11111000110
Data is processed and organized into a single schema before being put into the warehouse
The analysis is done on the cleansed data in the warehouse
Raw and unstructured date goes into a data lake
Data is selected and organized as and when needed
1110001101110
011011000110
11111000110

In contrast, the poetically named data lakes store data in highly scalable cloud storages in a completely unstructured, raw form. By utilizing the Schema-on-Read approach (as opposed to Schema-on-Write in data warehouses), they offer outstanding flexibility to any user or system accessing the data, since the knowledge of existing database schemas is no longer required.

In addition, since data is stored in the native format, no transformation/conversion is needed, which makes the work of analysts and data scientists easier. To top it off, cloud storages are completely decoupled from compute resources, which means that users with serious storage needs can balance their spending and not pay for the CPU time that they have no interest in or immediate need of.

Check out how we built a custom data lake management platform for a rapidly growing FinTech platform

Conclusion

no doubt that as the complexity of data processing systems grows, we will be seeing more and more solutions for the streamlining of ETL operations and solving of the most challenging data engineering riddles.

We fully recognize the fundamental role of data engineering in today's data-driven world and offer a slew of corresponding services. Contact us today for a guided tour around our diverse portfolio and an overview of our capabilities.

Rate this article

Tags

Data & Analytics

You may also like

Case study

Automated Data Acquisition for a PropTech Innovator

Blog post

Entering the Future: Top Big Data Trends to Define Upcoming Years

How can we help you?

Get in touch with us. We'd love to hear from you.

Name *

Email *

Phone

Message *

Upload or drag & drop files (5mb)

□ I give consent to the processing of my personal data given in the contact form above as well as receiving commercial and marketing communications under the terms and conditions of Intellias Privacy Policy.

intellias
Global Technology Partner

Send

intellias
Global Technology Partner

Chicago

500 West Madison Street,
Suite 1000, Chicago, IL 60661

+1 857 444 0442

info-chicago@intellias.com

Munich

Wappenhalle Business Center,
Konrad-Zuse-Platz 8, 81829

info-munich@intellias.com

info@intellias.com

| Privacy Policy | Cookie Policy | Security | Impressum | Sitemap