Definition of a Data Engineer

Definition of what is a data engineer based on the webpage visited of IBM cognitive class

https://cognitiveclass.ai/blog/data-scientist-vs-data-engineer

Data Engineer

Data engineer is the transformation of war data to a form suitable for analytics, major contributions has been done by the big data community Apache spark (), Apache finck, apache twister and Hadoop, qre considered widely used systems for data engineer

Data Engineers are the data professionals who prepare the “big data” infrastructure to be analyzed by Data Scientists.

Data engineers are software engineers who design, build, integrate data from various resources, and manage big data. with a goal for the optimization and performance of the accessible big data ecosystems. With vital roles responsible for managing, optimizing, overseeing, and monitoring data retrieval, storage, and distribution throughout the general process.

A **data engineer** is a technical person who’s in charge of architecting, building, testing, and maintaining the data platform as a whole. Depending on the project, they can focus on a specific part of the system or be an architect making strategic decisions. In the case of a small team, engineers and scientists are often the same people. But as a separate role, data engineers implement infrastructure for data processing, analysis, monitoring applied models, and fine-tuning algorithm calculations.

Data engineers develop the data infrastructure and interfaces necessary to collect data from different sources. They also design the systems used to transform that information into clean data sets that data analysts and data scientists can sort in ways that lead to useful conclusions. That data transformation makes everything that happens in the data science world possible.

This is mostly a technical position that combines knowledge and skills of computer science, engineering, and databases.

Responsibilities

Scaling your data science team. Here's a general recommendation: When your team of data specialists reaches the point when there is nobody to carry technical infrastructure, a data engineer might be a good choice in terms of a general specialist.

Big data projects. Currently, data engineering shifts towards projects that aim at processing big data, managing data lakes, and building expansive data integration pipelines for noSQL storages. In this case, a dedicated team of data engineers with allocated roles by infrastructure components is optimal.

Requiring custom data flows. Even for medium-sized corporate platforms, there may be the need for custom data engineering. Extract, Transform, Load is just one of the main principles applied mostly to automated BI platforms. In practice, a company might leverage different types of storages and processes for multiple data types. This involves a large technological infrastructure that can be architected and managed only by a diverse data specialist. A data engineer in this case is much more suitable than any other role in the data domain.

What are a data engineer's key responsibilities?

Simply put, data engineers manage data. They construct and oversee database architecture. They determine how data is collected and stored. They prepare that data for analysis by creating the pipelines that transform raw data into useful formats. And sometimes, they even create the systems through which people who are not part of the data science or data engineering team can access essential data (via, for instance, a custom real-time analytics dashboard).

Data related expertise. Data engineers would closely work with data scientists. Strong understanding of data modeling, algorithms, and data transformation techniques are the basics to work with data platforms. Data engineers will be in charge of building ETL (data extraction, transformation, and loading), storages, and analytical tools. So, experience with the existing ETL and BI solutions is a must.

What does a data engineer's in business analytics framework do ?

Data engineers spend their days:

- Building custom data pipelines based on business logic
- Collaborating with a database administrator to create data stores
- Collecting data from various sources
- Creating new data validation methods
- Developing frameworks to serve data
- Evaluating, parsing, and cleaning data sets
- Gathering requirements for data models
- Identifying and analyzing new data sources
- Maintaining computer clusters
- Making sure data is secure
- Preparing data as part of ETL (extract, transform, and load) processes
- Managing real-time data processing
- Sorting through raw data
- Stitching data from various sources together
- Writing ETL logic
- Writing queries that deliver accurate results

For Data Warehouse

Fixing issues and adding in any missing features required to keep a data warehouse well-stocked might keep our hypothetical data engineer busy until lunchtime. Next, they will turn their attention to whatever big project (e.g., creating a new ETL pipeline) is currently in their inbox. If they have already mapped out an implementation plan, they will start coding.

Architecture design. In its core, data engineering entails designing the architecture of a data platform.

Development of data related instruments/instances. As a data engineer is a developer role in the first place, these specialists use programming skills to develop, customize and manage integration tools, databases, warehouses, and analytical systems.

Data pipeline maintenance/testing. During the development phase, data engineers would test the reliability and performance of each part of a system. Or they can cooperate with the testing team.

Warehouse-centric. Historically, the data engineer had a role responsible for using SQL databases to construct data storages. This is still true today, but warehouses themselves became much more diverse. So, there may be multiple data engineers, and some of them may solely focus on architecting a warehouse. The warehouse-centric data engineers may also cover different types of storages (noSQL, SQL), tools to work with big data (Hadoop, Kafka), and integration tools to connect sources or other databases.

Data engineer in an analytic role

- Data engineers help make raw data more useful to the enterprise.
- Are responsible for building algorithms to help give easier access to raw data.

The Data engineer role

Write the different roles of a data engineer in different disciplines like

1) Business analytics.
2) Data warehouse maintenance.
3) Information security.
4) Y otro.

Data engineer in an analytic role

- Data engineers help make raw data more useful to the enterprise.
- Are responsible for building algorithms to help give easier access to raw data.

Business Analytics

Data engineers are tasked with managing and organizing data, while also keeping an eye out for trends or inconsistencies that will impact business goals. It's a highly technical position, requiring experience and skills in areas like programming, mathematics and computer science. But data engineers also need soft skills to communicate data trends

to others in the organization and to help the business make use of the data it collects. Some of the most common responsibilities for a data engineer include:

- Develop, construct, test and maintain architectures
- Align architecture with business requirements
- Data acquisition
- Develop data set processes
- Use programming language and tools
- Identify ways to improve data reliability, efficiency and quality
- Conduct research for industry and business questions
- Use large data sets to address business issues
- Deploy sophisticated analytics programs, machine learning and statistical methods
- Prepare data for predictive and prescriptive modeling

- **Machine learning algorithm deployment**. Machine learning models are designed by data scientists. Data engineers are responsible for deploying those into production environments. This entails providing the model with data stored in a warehouse or coming directly from sources, configuring data attributes, managing computing resources, setting up monitoring tools, etc.

- More specific expertise is required to take part in big data projects that utilize dedicated instruments like Kafka or Hadoop. If the project is connected with machine learning and artificial intelligence, data engineers must have experience with ML libraries and frameworks (TensorFlow, Spark, PyTorch, mlpack).

Data ingest

A data ingest refers to the extraction of data from different sources. During the extraction process, the data engineer needs to pay close attention to the formats and protocols that apply to the situation—all while extracting the data swiftly and seamlessly.

Configuring Business Intelligence Systems

After storing the data, data scientists establish the important connections between information sources. These sources could be data warehouses, data marts, data lakes, and applications. Establishing connections between sources could involve exposing the company's data to advanced machine-learning algorithms for business intelligence. Data engineers must understand how this process works to support data scientists in their jobs

9) Building Dashboards to Display Insights and Analytics

Many business intelligence and machine learning platforms allow users to develop beautiful, interactive dashboards. These dashboards showcase the results of queries, AI forecasting, and more. Creating dashboards is, usually, the responsibility of data scientists. However, data engineers may assist the data scientists in this process. Many BI platforms and RDBMS solutions allow users to create dashboards via a drag-and-drop interface. Knowledge of SQL, R, and Python can come in handy, though. It allows a data engineer to assist the data scientist in setting up dashboards that fit their needs.

Machine Learning deployment

Machine learning is, primarily, the domain of data scientists. However, because data engineers are the ones who build the data infrastructures that support machine learning systems, it's important that they feel comfortable with statistics and data modeling. Moreover, not all organizations will have a data scientist. Therefore, it's good to understand how to set up BI dashboards, deploy machine learning algorithms, and extract deep insights independently.

Data engineer in an analytic role

- Data engineers help make raw data more useful to the enterprise.
- Are responsible for building algorithms to help give easier access to raw data.

A **business intelligence developer** is a specific engineering role that exists within a business intelligence project. *Business intelligence* (BI) is a subcategory of data science that focuses on applying data analytics to historical data for business use. While a data engineer and ETL developer work with the inner infrastructure, a BI developer is in charge of

- defining reporting standards,
- developing reporting tools and data access tools,
- constructing interactive dashboards,
- developing data visualization tools,
- implementing OLAP cubes,
- testing warehouse architecture,
- validating data,
- testing user interface, and

- testing data querying process.

So, theoretically the roles are clearly distinguishable. In practice, the responsibilities can be mixed: Each organization defines the role for the specialist on its own. Everything depends on the project requirements, the goals, and the data science/platform team structure. The bigger the project, and the more team members there are — the clearer responsibility division would be. And vice versa, smaller data platforms require specialists performing more general tasks.

Information Security

Data engineers have different roles from a cyber security engineer.

Manage data and meta-data. The data can be stored in a warehouse either in a structured or unstructured way. Additional storage may contain meta-data (exploratory data about data). A data engineer is in charge of managing the data stored and structuring it properly via database management systems.

Connectors

Data engineers develop essential data pathways that connect various information systems. Therefore, data engineers should have a good understanding of data pipelines. They should know how they help different parts of an information network communicate with each other. For example, they should be able to work with REST, SOAP, FTP, HTTP, and ODBC—and understand strategies for connecting one information system or application to another as efficiently as possible.

Pipeline-centric data engineers would take care of data integration tools that connect sources to a data warehouse. These tools can either just load information from one place to another or carry more specific tasks. For example, they may include data staging areas, where data arrives prior to transformation. Managing this layer of the ecosystem would be the focus of a pipeline-centric data engineer.

Information systems

Provide data-access tools. In some cases, such tools are not required, as warehouse types like data-lakes can be used by data scientists to pull data right from storage. However, if an organization requires business intelligence for analysts and other non-technical users, data engineers are responsible for setting up tools to view data, generate reports, and create visuals.

Track pipeline stability. Monitoring the overall performance and stability of the system is really important as long as the warehouse needs to be cleaned from time to time. The automated parts of a pipeline should also be monitored and modified since data/models/requirements can change.

Data Warehouse Maintenance

Fixing issues and adding in any missing features required to keep a data warehouse well-stocked might keep our hypothetical data engineer busy until lunchtime. Next, they will turn their attention to whatever big project (e.g., creating a new ETL pipeline) is currently in their inbox. If they have already mapped out an implementation plan, they will start coding.

After extracting information from various business systems, data engineers may need to prepare the information for integrating it with a data warehouse system. Data integration is crucial if they want to query it for deep insights. This could involve transforming the data with an ETL tool like Xplenty.

Cloud-based data warehouses form the backbone of most advanced business intelligence data systems. Data engineers should understand how to set up a cloud-based data warehouse. They should be adept at connecting a wide variety of data types to it, and optimizing those connections for speed and efficiency.

Data Lakes

Data warehouses can only work with structured information, such as information in a relational database. Relational database systems store data in clearly-identified columns and rows. Meanwhile, data lakes can work with any type of data. This includes unstructured information, such as streaming data. BI solutions can hook up to data lakes to derive valuable insights. For this reason, many companies are incorporating data lakes into their information infrastructures.

For applying machine learning algorithms to unstructured data, it is important to know how to integrate data and connect it to a business intelligence platform.

Process followed by a data engineer

The important process that develops a data engineer referred as the ETL process, which stands for Extract, Transform and Load, which are developed on top of big datasets and create big data warehouses that can be used for reporting or analysis by data scientists. Beyond that because Data Engineers focus more on the design and architecture.

An **ETL developer** is a specific engineering role within a data platform that mainly focuses on building and managing tools for Extract, Transform, and Load stages. So, the border between a data engineer and ETL developer is kind of blurred. However, an ETL developer is a narrower specialist rarely taking architect/tech lead roles. These tasks typically go to an ETL developer.

- ETL process management
- Data warehouse architecting
- Data pipeline (ETL tools) development
- ETL testing
- Data flow monitoring

Data Engineers Skills

- Data engineers create data pipelines that connect data from one system to another. They are also responsible for transforming data from one format to
- Tools needed in certain topics and programming languages.
- SQL statement to perform a specific action.
- MapReduce when analyzing a large data set featuring a parallel, distributed algorithm on a cluster.

Programming knowledge

Data engineers need expertise in the following programming languages

- SQL: In order to set up, make query, and manage database systems. Also, with RDBMS. Good knowledge of the relational database systems. Relational And Non-Relational Database Systems

Data engineers need to know how to work with a wide variety of data platforms. SQL-based relational database systems (RDBMSs) like MySQL, PostgreSQL (a hybrid SQL and NoSQL database), and Microsoft SQL Server are particularly important For example, they should feel comfortable using SQL to build and set up database systems. Data engineers should also develop skills working with NoSQL databases such as MongoDB, Cassandra, Couchbase, and others.

- **Python:** In order to create data pipelines, write ETL scripts, and to set up statistical models and perform analysis.

- **R Language:** In order to analyze data, and set up statistical models, dashboards, and visual displays.

Knowledge of these scripting languages allows data engineers to troubleshoot and improve the database systems. It also allows them to optimize business insights tools, and machine-learning systems they're working with. Data engineers could also benefit from being familiar with Java, NoSQL, Julia, Scala, MATLAB, and TensorFlow.

Programming tools

- **Python** is a very popular general-purpose language. Widely used for statistical analysis tasks, it could be called the lingua franca of data science. Fluency in Python (along with SQL) appears as a requirement in over two-thirds of data engineer job listings.
- **R** is a unique language with features that other programming languages lack. This vector language is finding use cases across multiple data science categories, from financial applications to genetics and medicine.
- **Java**, because of its high execution speeds, is the language of choice for building large-scale data systems. It is the foundation for the data engineering efforts of companies such as Facebook and Twitter. Hadoop is written mostly in Java.
- **Scala** is an extension of Java that is particularly suited for use with Apache Spark. In fact, Spark is written in Scala. Although Scala runs on JVM (Java Virtual Machine), Scala code is cleaner and more concise than the Java equivalent.
- **Julia** is an up-and-coming general-purpose programming language that is very easy to learn. Its speed is on par with C or Fortran, which allows it to be used as the single language in data projects that formerly required two languages. For example, Python may have been used for prototyping, with re-implementation in Java or C++ to meet production performance requirements. Now, with its speed and ease of use, Julia can be used for both prototyping and production.

Data engineer tools

- **Apache Hadoop** is a foundational data engineering framework for storing and analyzing massive amounts of information in a distributed processing environment. Rather than being a single entity, Hadoop is a collection of open-source tools such as HDFS (Hadoop Distributed File System) and the MapReduce distributed processing engine. Precisely Connect is a highly scalable and easy-to-use data integration environment for implementing ETL with Hadoop.
- **Apache Spark** is a Hadoop-compatible data processing platform that, unlike MapReduce, can be used for real-time stream processing as well as batch processing. It is up to 100 times faster than MapReduce and seems to be in the process of displacing it in the Hadoop ecosystem. Spark features APIs for Python, Java, Scala, and R, and can run as a stand-alone platform independent of Hadoop.
- **Apache Kafka** is today's most widely used data collection and ingestion tool. Easy to set up and use, Kafka is a high-performance platform that can stream large amounts of data into a target like Hadoop very quickly.
- **Apache Cassandra** is widely used to manage large amounts of data with lower latency for users and automatic replication to multiple nodes for fault-tolerance.