6 Responsibilities of a Data Engineer

Introduction

Data engineering is a relatively new field, and as such, there is a huge variance in the actual job responsibilities across different companies. If you are a student, analyst, engineer, or new to the data space and

> Unclear with data engineers' job responsibilities

> Believe that the current state of a data engineer's job description is messy

Then this post is for you. In this post, we cover the 6 key responsibilities of a data engineer.

Responsibilities of a data engineer

1. Move data between systems

This represents the main responsibility of a data engineer. It usually involves

1. **Extract**: Extracting data from any number of sources. The source can be an `external API, cloud storage, databases, static files` , etc
2. **Transform**: This step involves transforming the data. Some common transformations are `mapping, filtering, enrichment, changing the structure of the data(eg denormalizing data), and aggregating` .
3. **Load**: This is the step where the data is loaded into the destination system. The destination system can be a `cloud storage file system, data warehouse, and/or cache database, etc` .

Common tools/frameworks: Pandas, Spark, Dask, Flink, Beam, Debezium, Kafka, Docker, Kubernetes

2. Manage data warehouse

More often than not, most of the company's data lands within the data warehouse. The responsibilities of a data engineer in this context are

1. **Warehouse data modeling**: Model the data for analytical queries, which are typically aggregation queries on large tables. Modeling here involves applying appropriate `partitions, handling fact and dimension tables, etc` .
2. **Warehouse performance**: Make sure the queries are fast and the warehouse can scale as needed.
3. **Data Quality**: Ensuring data quality within the data warehouse.

Common modeling techniques: Kimball modeling, Data Vault, Data Lake

Common frameworks: Great expectations, dbt for data quality

Common warehouses: Snowflake, Redshift, Bigquery, Clickhouse, Postgres

3. Schedule, execute, and monitor data pipelines

Data engineers are also responsible for scheduling the ETL pipelines, making sure they run without any issue, and monitoring them.

1. **Scheduling** data pipelines to be run at a certain schedule or in response to some event.
2. **Executing** data pipelines and ensuring that they can scale, have the right permissions, etc.
3. **Monitoring** data pipelines for failures, deadlocks, and long-running tasks.
4. **Managing metadata** such as time of the run, end to end time taken, failure reasons, etc

Common frameworks: Airflow, dbt, Prefect, Dagster, AWS Glue, AWS Lambda, Streaming pipeline using Flink/Spark/Beam

Common databases: MySQL, Postgres, Elastic search and data warehouses

Common storage systems: AWS S3, GCP cloud store

Common monitoring systems: Datadog, Newrelic

4. Serve data to the end-users

Once you have the data available in your data warehouse, it's time to serve it to the end-user. The end-user can be analysts, an application, external clients, etc. Depending on the end-user you may have to set up

1. **Data visualization/Dashboard tool**: Tool used by humans to analyze the data and create pretty charts that can be shared easily.
2. **Permissions for the data**: If it's a table, then granting correct permissions to your applications or end-users. If it's in cloud storage, granting cloud users appropriate permissions, etc.
3. **Data endpoints(API)**: Some application/external clients may need API-

based access to your data. In such cases, a server to send data via an API endpoint will need to be set up.
4. **Data dumps for clients**: Some clients may require data dumps from your system. In such cases, you will have to set up a data pipeline to facilitate this.

Common tools/languages: Looker, Tableau, Metabase, Superset, role-based permissions(for your system), Python/Scala/Java/Go for API endpoints, pipeline tools for client data dumps

5. Data strategy for the company

Data engineers are involved in coming up with the data strategy for the company. This involves

1. Deciding what data to collect, how to collect it, and store it securely.
2. Evolving data architecture for custom data needs.
3. Educating end users on how to use data effectively.
4. Deciding what data(if any) to share with external clients.

Common tools/frameworks: Confluence, google docs, RFC documents, brainstormings, meetings

6. Deploy ML models to production

Data scientists and analysts develop sophisticated models that closely model the working of a specific business process. When it's time to deploy these models, data engineers are usually the ones who optimize them to be used in a production environment.

1. **Optimizing training and inference**: Setting up a batch/online learning pipelines. Ensuring the model is appropriately sized.
2. **Setting up monitoring**: Setting up monitoring and logging systems for the ML model.

Common frameworks: Seldon core, AWS MLOps

Conclusion

Hope this article gives you a good understanding of the different responsibilities that a data engineer may take on. The number of responsibilities that you may have depends on the company, team structure, and workload. The **main objective** of the data engineering team(s) is to **enable company-wide use of data for decision making**.

Usually, the bigger the company the more narrow and deep your responsibilities get. You can use this as a list to identify your areas of interest and make sure that your job responsibilities match them. Please leave any questions or comments in the comment section below.

Further reading

1. 10 Key skills, to help you become a data engineer
2. What is a Data Warehouse?
3. A proven approach to land a Data Engineering job

If you found **this article** helpful, share it with a friend or colleague using one of the socials below!

Build effective data systems equipped with core data engineering principles!

Subscribe to my newsletter for:

1. **Core Data Concepts** to master the tools and frameworks
2. **Career Growth Tips** to align with business needs
3. **Design Patterns** for smarter data pipeline strategies

No sponsors, no agenda—just pure, actionable, and reliable content. Get on the list now and future-proof your data engineering career.

Your name

Your email address

Get actionable data engineering content!

We only send useful and actionable content. Unsubscribe at any time.

Previous Post
6 Key Concepts, to Master Window Functions
Next Post
How to improve at SQL as a data engineer

Login

Add a comment

M↓ MARKDOWN COMMENT ANONYMOUSLY ADD COMMENT

Upvotes Newest Oldest

J **Justin Wong**
1 point · 3 years ago

This is a fantastic summary. Very high quality write-ups as always.

? **Anonymous**
0 points · 3 years ago

I think "Data strategy for the company" should consider good soft skills. I would sugest an article to explore more theese soft skills required to make good data strategies for the company.

E **Elie Kawerk**
0 points · 3 years ago

Isn't data governance (technical aspects) also a data engineer's responsibility? Caring about metadata and proper documentation are really crucial.

> J **Joseph Machado** MODERATOR
> **0 points** · 3 years ago
>
> That is correct. I have managing metadata as part of point 3. Documentation should be explicitly added as part of dev process as well.

D **Dewei Zhai**
0 points · 20 months ago

As a data engineer with 6 years of experience, I like this summary because it focused on something that doesn't change often: the problem you are hired to solve. Nice summaries!

> J **Joseph Kevin Machado** MODERATOR
> **0 points** · 20 months ago
>
> Thank you Dewei

I **Ioan Simion Belbe**
0 points · 17 months ago

Great article, although I think the last responsibility (deploy ML models to production) is quickly becoming and de facto is the responsibility of an ML Engineer/ML DevOps.

K **Karan**

> **0 points** · 2 years ago
>
> Postgres is a Transactional Database whereas SQL Server is more data warehouse oriented. I would recommend to add SQL Server within the list as it is on of the powerful database out there.