CELEBRATING 12 YEARS

Quality Thought®

JOB ORIENTED JOIP INTENSIVE PROGRAM

GCP

Cloud Data Engineer

CELEBRATING 12 YEARS

Quality Thought®

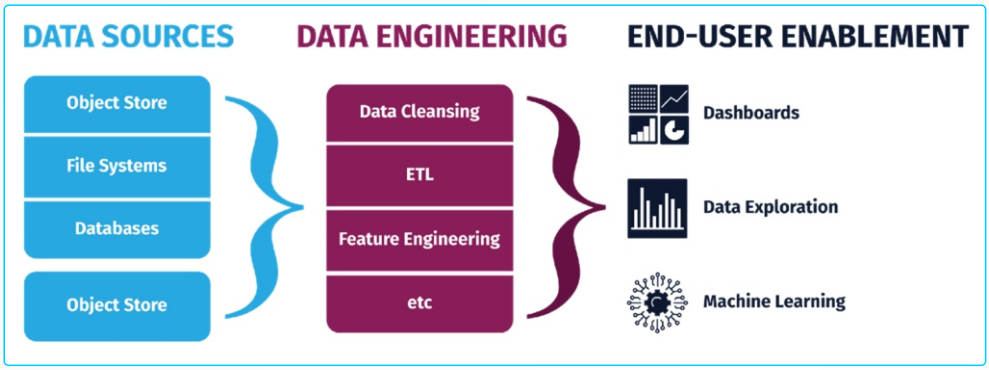
Cloud Data Engineer - GCP

Course Duration
45 Days

total sessions hours
300 Hrs

What Is a Cloud DATA ENGINEER?

- A cloud data engineer is like a swiss army knife in the data space; there are many roles and responsibilities that data engineers are capable of, depending on the particular needs of the organization.
- In short, data engineers set up and operate the organization's data infrastructure preparing it for further analysis by data analysts and scientists

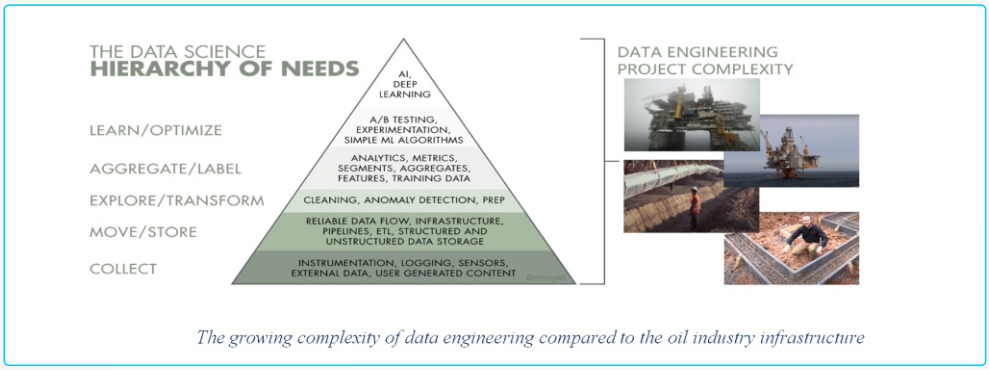


CLOUD DATA ENGINEER JOB DESCRIPTION

Specific responsibilities expected of a cloud data engineer can include any or all of the following:

- Migrating on-premises corporate applications and related data to the cloud
- Designing and deploying new applications directly in the cloud
- Identifying best practices for cloud services monitoring and management and promoting these best practices across the corporation
- Researching and implementing cloud services to support cloud apps and maintain cloud services
- Monitoring cloud app performance for potential bottlenecks and resolving performance issues
- Identifying and implementing cost-saving strategies to reduce ongoing cloud expenses
- Automating key services and tasks across cloud systems to increase efficiency and further reduce cloud costs
- Formulating a recovery plan and executing the plan in the event of cloud downtime or failure.

If you look at the Data Science Hierarchy of Needs, you can grasp a simple idea: The more advanced technologies like machine learning or artificial intelligence are involved, the more complex and resource-heavy data platforms become.



The growing complexity of data engineering compared to the oil industry infrastructure

CELEBRATING 12 YEARS

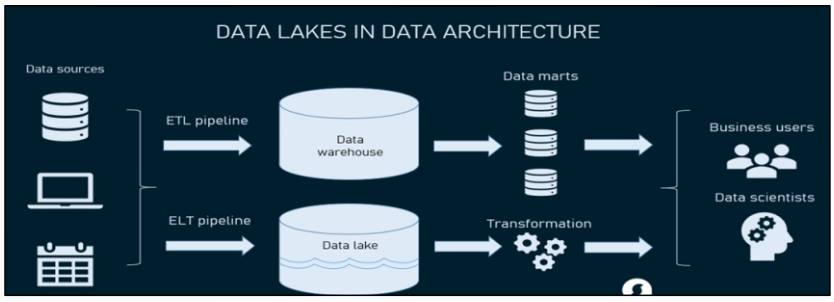
QualityThought®

ELT data pipeline and big data engineering

- Speaking about data engineering, we can't ignore the big data concept. Grounded in the four Vs – volume, velocity, variety, and veracity – big data usually floods large technology companies like YouTube, Amazon, or Instagram. Big data engineering is about building massive reservoirs and highly scalable and fault-tolerant distributed systems able to inherently store and process data.
- Big data architecture differs from conventional data handling, as here we're talking about such massive volumes of rapidly changing information streams that a data warehouse isn't able to accommodate. The architecture that can handle such an amount of data is a data lake.

Data Lake

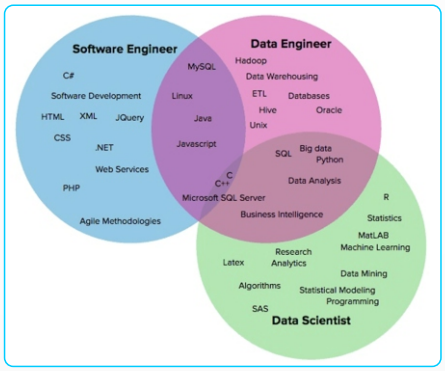
- A Data lake is a vast pool for saving data in its native, unprocessed form. A data lake stands out for its high agility as it isn't limited to a warehouse's fixed configuration.
- In contrast to the ETL architecture we described above, a data lake uses the ELT approach swapping transform and load operations. Supporting large storage and scalable computing, a data lake starts data loading immediately after extracting it, handling raw — often unstructured — data.
- You can check our detailed comparison of ETL and ETL approaches , but in a nutshell, ELT is a more advanced method as it allows for significantly increasing volumes of data to be processed. It also expedites information processing (since transformation happens only on-demand) and requires less maintenance.
- A data lake is worth building in those projects that are going to scale and would need a more advanced architecture. Besides, they are very convenient, for instance, when the purpose of data hasn't been determined yet since you can load data quickly, store it, and then modify it as necessary. Once you need data, you can apply such data processing tools as Apache or MapReduce to transform it during retrieval and analysis.
- Data lakes are also a powerful tool for data scientists and ML engineers, who would use raw data to prepare it for predictive analytics and machine learning.



Hadoop

- So, Hadoop is a large-scale data processing framework based on Java. This software project is capable of structuring various big data types for further analysis. The platform allows for splitting data analysis jobs across various computers and processing them in parallel.

Skills and qualifications



Overlapping skills of the software engineer, data engineer, and data scientist

CELEBRATING 12 YEARS

Quality Thought®

Data-related skills

- "A data engineer should have knowledge of multiple kinds of databases (SQL and NoSQL), data platforms, concepts such as MapReduce, batch and stream processing, and even some basic theory of data itself, e.g., data types, and descriptive statistics," underlines Juan.

Systems creation skills

- Data engineers need to have experience with various data storage technologies and frameworks they can combine to build data pipelines.

Toolkit

- Data engineering process involves using different data storage and manipulation tools together. So a data engineer should have a deep understanding of many data technologies to be able to choose the right ones for a certain job.

Tools for writing ETL/ELT pipelines

- **Airflow.**
  This Python-based workflow management system was initially developed by Airbnb to rearchitect their data pipelines. Migrating to Airflow, the company reduced their experimentation reporting framework (ERF) run-time from 24+ hours to about 45 minutes. Airflow's key feature is automating scripts to perform tasks. Among the Airflow's pros, Juan highlights its operators: "They allow us to execute bash commands, run a SQL query or even send an email". Juan also stresses Airflow's ability to send Slack notifications, complete and rich UI, and the overall maturity of the project. On the contrary, Juan dislikes that Airflow only allows for writing jobs in Python.
- **Cloud Dataflow.**
  A cloud-based data processing service, Dataflow is aimed at large-scale data ingestion and low-latency processing through fast parallel execution of the analytics pipelines. Dataflow has a benefit over Airflow as it supports multiple languages like Java, Python, SQL, and engines like Flink and Spark. It is also well maintained by Google Cloud. However, Juan warns that Dataflow's high cost might be a disadvantage for some.
- **Kafka.**
  From a messaging queue to a full-fledged event streaming platform, Apache Kafka distributes data across multiple nodes for a highly available deployment within a single data center or across multiple availability zones. As an abstraction of a distributed commit log, it provides durable storage.
  Other popular ETL and data solutions are the Stitch platform for rapidly moving data and Blendo, a tool for syncing data from various sources to a data warehouse.
- **Warehouse solutions.**
  Widely used on-premise data warehouse tools include Teradata Data Warehouse, SAP Data Warehouse, IBM db2, and Oracle Exadata. Most popular cloud-based data warehouse solutions are Amazon Redshift and Google BigQuery. Be sure to check our detailed comparison of the top cloud warehouse software.
- **Big data tools.**
  Big data technologies that a data engineer should be able to utilize (or at least know of) are Hadoop, distributed file systems such as HDFS, search engines like Elasticsearch, ETL and data platforms: Apache Spark analytics engine for large-scale data processing, Apache Drill SQL query engine with big data execution capabilities, Apache Beam model and software development kit for constructing and running pipelines on distributed processing backends in parallel.

CONCLUSION

- A data engineer needs to enjoy finding patterns, identify new ways to create complex systems that work and keep the big picture in mind. Keep up to date with trends: As data science is a dynamic field, data engineers must constantly upskill so they can work well with data scientists, analysts, and architects.

CELEBRATING 12 YEARS

Quality Thought®

- Understanding Data
- Understanding Data Engineering role
- Cloud Computing
- OLTP vs. OLAP
- ETL to ELT
- Data store vs. Data Lake vs. Data Warehouse (DWH)
- Big Data Fundamentals
- Big Data Architecture
- Hadoop Fundamentals
- Data Ingestion/ Data Pipelines
- Python Basics (Scripting)
- RDBMS (SQL)
- No SQL
- Data Bricks with PySpark

Google Cloud Platform (GCP)

- Google Cloud Platform
- Understanding GCP components
- Compute Engine

GCS – Google Cloud Components

- Cloud Functions
- Batch and streaming data
  - a) Cloud Dataflow
  - b) Cloud Dataproc/Hadoop Ecosystem
  - c) Pub/Sub
  - d) Apache Kafka
- Data Publishing and Visualization
  - a) BigQuery
- Job Automation and Orchestration
  - a) Cloud Composer
  - b) AirFlow
- Cloud Data Fusion
- Big Table

Overview of AWS & Azure

- Understand Azure Data Factory
- Explain the data factory process
- Azure Delta Lake
- Create datasets
- Create data factory activities and pipelines
- S3, Lambda Functions
- Redshift overview

CELEBRATING 12 YEARS

Quality Thought®

Google Cloud Certified PROFESSIONAL DATA ENGINEER

Other Responsibilities

- Agile Process (JIRA, Scrum, Sprint)
- GIT process – Code/Scripts
- Confluence - Documents
- Requirements Understanding
- Go Live/Prod deployment process
- End to End Use cases
- RESUME & Interview PREPARATION

Prerequisites

- Basic SQL knowledge
- Any Basic programing Knowledge (Java/Python/C)

Who can attempt this course?

- Database Engineers
- BigData/Hadoop Engineers
- ETL/Data Warehouse Engineers
- Any Application Programmers
- Test Engineers
- Data Analysts

QualityThought®
Transforming Dreams! Redefining Future!

CELEBRATING 12 YEARS

STUDENT TRANSFORMATION STAGES

Student's Enrollment

8Hours JOIP Training

Mock Interview

Resume Preparation

Interview Schedule

3-6 Months Internship

JOB

Interview Schedule

JOB

wipro
LARSEN & TOUBRO
Apollo HOSPITALS
Cognizant
Microsoft
servicenow
amazon
S&P Global
Aadvi Tech Solutions
Medtronic
meslova
{actlogica} Actuarial | Finance | Technology
Sunseaz Technologies Pvt Ltd
CAPRUS IT UNLOCKING SMART SOLUTIONS
SOLUGENIX
Collabera Value. Accelerated.
i Prism Technologies
NOOR
Nendrasys Accelerating Business Transformation
magnasoft
Innominds Powering the Digital Next
High Noon Consulting. Possible Is Everything
JRD SYSTEMS
ValueLabs
ENVISION