How Do prepare for first Cloud Data Engineering interview

Ganesh R

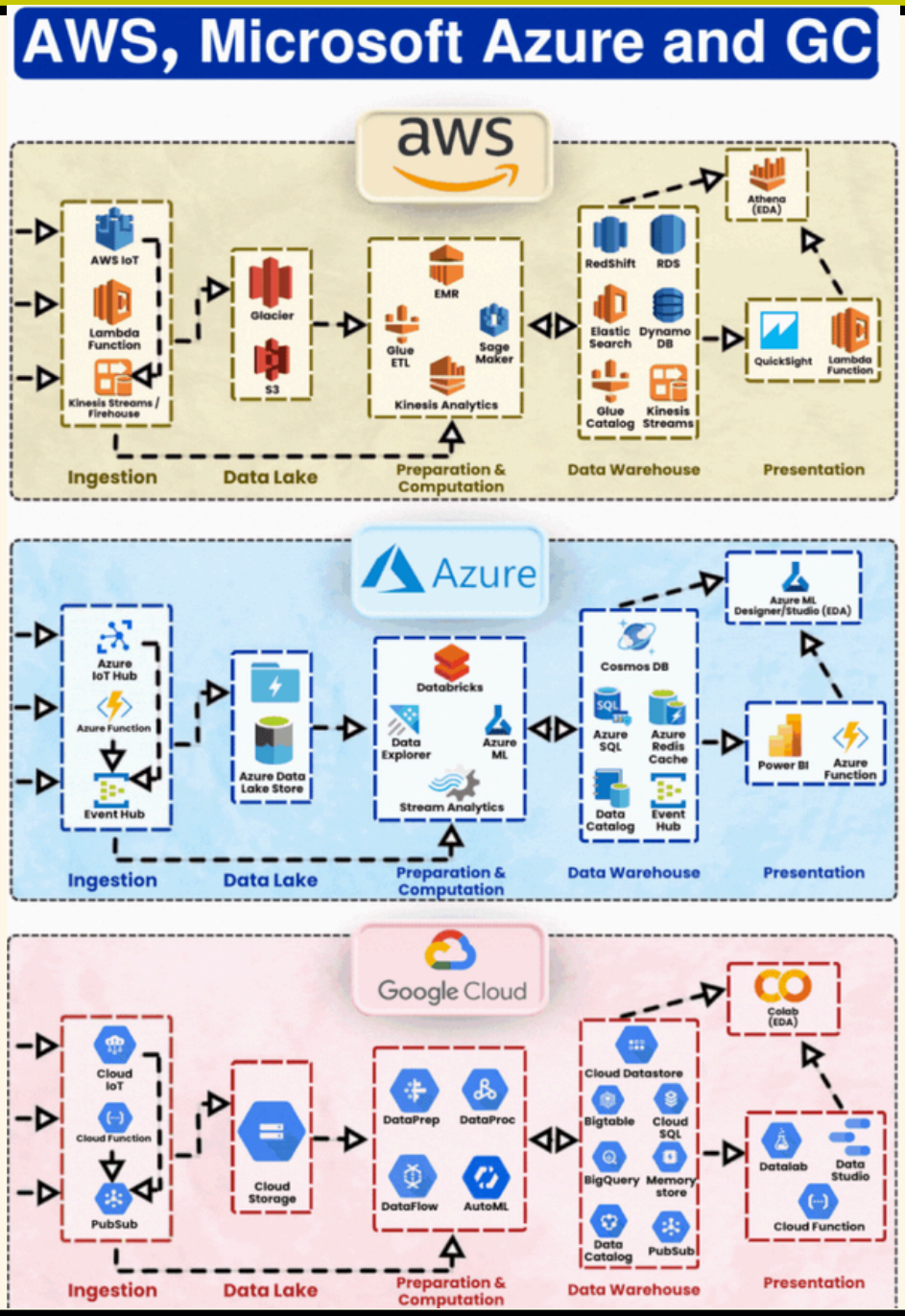
Azure Data Engineer

I recently given guidelines to crack interview for an intern cloud Data Engineering position and later got an offer. I wanted to share my experience preparing for the technical interview and hopefully give you some insight to how to prepare for yours. Here's a quick rundown:

- Python essentials
- Database fundamentals
- Essential SQL
- Database Design and optimization
- Data Processing and
- ETL Data pipeline design
- Data warehouse and Lakes
- Essential tools and technologies

But before we get into the details, it’s worth pointing out that every Data Engineering role is different. This is because, in practice, the data engineering lifecycle cuts across many domains of responsibility. You will wear multiple hats, juggling various tasks crucial to the success of data-driven initiatives within an organization:

Not all companies will have people in all of the above roles. So you might have to do some Data Science or BI in your job as well. Make sure you read the job description and understand exactly what you’re expected to do, and if you’re unsure, be sure to clarify at the end of your interview. Now let’s get into it!


AWS, Microsoft Azure and GC
aws
AWS IoT
Lambda Function
Kinesis Streams / Firehouse
Glacier
S3
EMR
Glue ETL
Sage Maker
Kinesis Analytics
RedShift
RDS
Elastic Search
Dynamo DB
Glue Catalog
Kinesis Streams
Athena (EDA)
QuickSight
Lambda Function
Ingestion
Data Lake
Preparation & Computation
Data Warehouse
Presentation
Azure
Azure IoT Hub
Azure Function
Event Hub
Azure Data Lake Store
Databricks
Data Explorer
Azure ML
Stream Analytics
Cosmos DB
Azure SQL
Azure Redis Cache
Data Catalog
Event Hub
Azure ML Designer/Studio (EDA)
Power BI
Azure Function
Ingestion
Data Lake
Preparation & Computation
Data Warehouse
Presentation
Google Cloud
Cloud IoT
Cloud Function
PubSub
Cloud Storage
DataPrep
DataProc
DataFlow
AutoML
Cloud Datastore
Bigtable
Cloud SQL
BigQuery
Memory store
Data Catalog
PubSub
Colab (EDA)
Datalab
Data Studio
Cloud Function
Ingestion
Data Lake
Preparation & Computation
Data Warehouse
Presentation

Python Essentials

Python is definitely a must-have for Data Engineers. A good place to refresh your skills is this page on geeksforgeeks. Here's a list of things you should know:

1. How to use data structures such as lists, tuples, sets and dictionaries.
2. File input/output (I/O) operations — crucial for processing data.
3. Python's data manipulation and analysis libraries, such as NumPy and Pandas.
4. Working with regular expressions (Regexes)
5. Recursion — can be helpful for tasks involving nested or hierarchical data structures,

such as JSON objects, XML files, or database schemas that resemble trees.

Database fundamentals

The database is your workstation. Here are a list of things to prepare for:

1. The characteristics and use cases of relational databases.
2. Different types of keys, normalizations and constraints.
3. ACID properties and their advantages.
4. Difference of NoSQL databases and their use cases.
5. CAP theorem and scenarios of tradeoff.
6. OLAP vs OLTP databases.
7. Types of Data models.

Essential SQL

This is a no-brainer. Any company hiring at any level of DE will have an SQL round. Going beyond the syntax, you will need to know:

1. The SQL order of operations 2. Data types (INT, VARCHAR, DATE, etc.) 3. Types of operators (arithmetic, comparison, logical, etc.) 4. DDL, DML and DCL commands 5. Different types of joins 6. Aggregation functions 7. Subqueries and Views

Apart from these, you will need to know some advanced concepts such as different types of triggers, Stored procedures, CTEs and Window functions. I recommend practicing a few SQL questions on DataLemur before facing your interview.

Database Design and Optimization

Apart from database theory and SQL, you need to have a sound understanding of database design. This involves things like ER modelling, understanding tradeoffs in normalization, schema design, when and how to use different types of indexing, data partitioning best practices, etc.

It would also be beneficial to know Database design software such as MySQL Workbench, Microsoft SQL Server Management Studio, or Oracle SQL Developer.

Knowledge on optimizing database performance is definitely a plus! Measures for decreasing latency and increasing query speed may be asked during the interview. Knowing different replication strategies when scaling databases is also crucial. The Designing data intensive applications book was extremely helpful.

Essential Tools

At the core of the DE tech stack are programming languages such as Python, Java and Scala. You should be able to write clean and reproducible code. For Python users, the PEP-8 standard is a great guide. You should also be comfortable with OOP.

SQL databases such as Oracle, MySQL, PostgreSQL and Microsoft SQL Server are frequently used as on-premise storage solutions for structured data because of their ACID properties. NoSQL databases, such as MongoDB, Cassandra, and Redis, are popular options for unstructured or semi- structured data.

Data Processing

Data Processing is a key part of a Data Engineers role. A data pipeline typically has 4 pillars — data sources, DPUs, data sinks, and orchestration. You will need a sound understanding of the lifecycle of an ETL job, steps for designing an ETL pipeline, best practices when implementing a pipeline, and finally, how to optimize it.

Each type and flavor of data pipeline has its own benefits and disadvantages. In general, they can be classified according to the processing method (batch vs stream) and the data flow pattern (ETL vs ELT).

Data warehouses and Data Lakes

Data warehouses are the backbone of modern data analytics. Data is frequently arranged using well-known schemas such as star or snowflake to enhance query performance.

You will need to know the different types of transformations used in the ETL process, such as cleaning, standardizing, enrichment, reformatting and aggregation, and also the 2 types of loading strategies: full load and incremental load. In contrast, data lake architecture encompasses different zones involving different steps for processing as well.

A vital component of data management is security and governance. The main components of data governance include Data quality, Data lineage, Privacy policies and Metadata management. On the other hand, Data security involves access control (RBAC and ABAC), encryption, data masking and security monitoring. You will need to have a sound understanding of how to implement these in real-world data-driven scenarios.

Creating an impressive portfolio

It would be very beneficial to have a portfolio of projects that reflect your knowledge on these areas. Here are a few things you can do to enhance your projects:

Ingesting from different data sources

- local files: Familiarity with CSV, TXT or Excel

- Web pages: Building web scrapers using libraries such as BeautifulSoup, Selenium, Requests, and Urllib.
- APIs and JSON files

Data storage Practice staging your data in separate zones based on transformation levels:

- Raw and unprocessed Cleaned and transformed Curated views for dashboarding and reporting

Your portfolio will stand out if you can show that you are capable of managing a variety of data sources, including flat files and APIs. Be sure to highlight certain best practices, including error handling, data validation, and efficiency.

Data transformation

In data processing, data quality and integrity are also evaluated. Missing values are handled, outliers are examined, and data types are cast appropriately to ensure that the data is reliable and ready for analytics. Here are some aspects of the data processing portion that you want to highlight in your projects:

- Converting data types: You should be able to convert your data types as necessary to allow for optimized memory and easier-to-use formats.
- Handling missing values: Apply necessary strategies to handle missing data.
- Removing duplicate values: Ensure all duplicate values are removed.
- Error handling and debugging: To create reproducibility, implement blocks of code to handle anticipated errors and bugs.
- Data validation and quality checks: Implement blocks of code to ensure processed data matches the source of truth.

The Data Engineering Tech stack

Cloud technologies Cloud technologies provide the fundamental framework for a wide range of DE tasks in today’s data-driven world.

Among the top cloud service providers is AWS. The following are some Services:

S3: Raw or processed data can be stored using

S3 Glue: An entirely managed ETL solution

Redshift: A solution for data warehousing

Kinesis: Data streaming in real time

Microsoft Azure is definitely a more popular option providing the following services:

Blob Storage: Scalable object storage for unstructured data

Azure Data Factory: A service for data integration and ETL.

Data Warehouse: Azure SQL a fully managed data warehouse.

Event Hubs: Ingestion of data in real time

Google Cloud Provider (GCP) is also provides a range of widely used services:

Cloud Storage: AWS S3-like object storage solution

Dataflow: Processing of data in batches and streams.

BigQuery: A highly scalable, serverless data warehouse

AWS, Azure, GCP: Databricks, Snowflake and confluent Kafka

Ingestion, processing, and storage tools Gaining an understanding of the features, advantages, and disadvantages of these tools will enable you to design scalable and effective data pipelines, making you an excellent prospect in any DE interview and a priceless asset in practical situations:

Apache Kafka: For real-time data ingestion. Kafka has established itself as the industry standard
Apache Flume: Another popular tool for data ingestion, specially designed for ingesting data into Hadoop environments.

Apache Spark is undoubtedly the most popular in-memory data processing engine for big data tasks. Hadoop is a dependable option for batch processing large datasets and is frequently combined with other tools such as Hive and Pig.

For data storage, some popular options are HDFS, Amazon S3, Google Cloud Storage and Azure Data Storage.

Scheduling tools

Coordinating all of the above elements into a smooth, automated workflow is crucial after you've set up your environment. Some of the most widely used scheduling tools, include Luigi, Cron Jobs, and Apache Airflow. Among these Apache Airflow has become the de facto standard across a wide range of industries.

At its core, Airflow's architecture consists of several components such as the scheduler, worker nodes, metastore database and web server. Since Apache Airflow is widely used in industry and is often asked about in interviews, learning about it would definitely be benificial.

Beyond just carrying out tasks at predetermined times, scheduling serves other purposes as well. It entails intricate workflow orchestration in which jobs are interdependent and must be successfully completed before starting another.

These are some of the things I became familiar with before my first interview. However, different positions may have different degrees of difficulty. There's a lot more things you can learn to stand out from the crowd and be a valuable Data Engineer. These include knowledge of:

- Unit testing and automation CI/CD Data Security and Privacy
- Data quality monitoring
- Pipeline catchup and recovery
- Infrastructure as Code
- Business Intelligence and visualization

Resources

A couple good resources you can refer to are Ace the Data Engineering interview and Cracking the Data Engineering Interview by Kadeisha Bryan.

There's a wonderful Medium article by Nisha Sreedharan that links to different resources to learn about all of these things. Don't forget to prepare for Behavioral interviews and brain teasers!

These are some of the things I became familiar with before my first interview. However, different positions may have different degrees of difficulty. There's a lot more things you can learn to stand out from the crowd and be a valuable Data Engineer. These include knowledge of:

- Unit testing and automation CI/CD Data Security and Privacy
- Data quality monitoring
- Pipeline catchup and recovery
- Infrastructure as Code
- Business Intelligence and visualization

Resources

A couple good resources you can refer to are Ace the Data Engineering interview and Cracking the Data Engineering Interview by Kadeisha Bryan.

There's a wonderful Medium article to different resources to learn about all of these things. Don't forget to prepare for Behavioral interviews and brain teasers!

Follow for more content like this Azure Cloud for Data Engineering

Ganesh R

Azure Data Engineer

https://www.linkedin.com/in/rganesh0203/