CoGrammar

Introduction to Data Engineering

SKILLS FOR LIFE
SKILLS BOOTCAMPS

Department for Education

Data Science Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly. **(FBV: Mutual Respect.)**
- No question is daft or silly - **ask them!**
- There are **Q&A sessions** midway and at the end of the session, should you wish to ask any follow-up questions. Moderators are going to be answering questions as the session progresses as well.
- If you have any questions outside of this lecture, or that are not answered during this lecture, please do submit these for upcoming Open Classes. You can submit these questions here: **Open Class Questions**

Data Science Lecture Housekeeping cont.

- For all **non-academic questions**, please submit a query: [www.hyperiondev.com/support]

- Report a **safeguarding** incident: [www.hyperiondev.com/safeguardreporting]

- We would love your **feedback** on lectures: [Feedback on Lectures]

Data Engineering

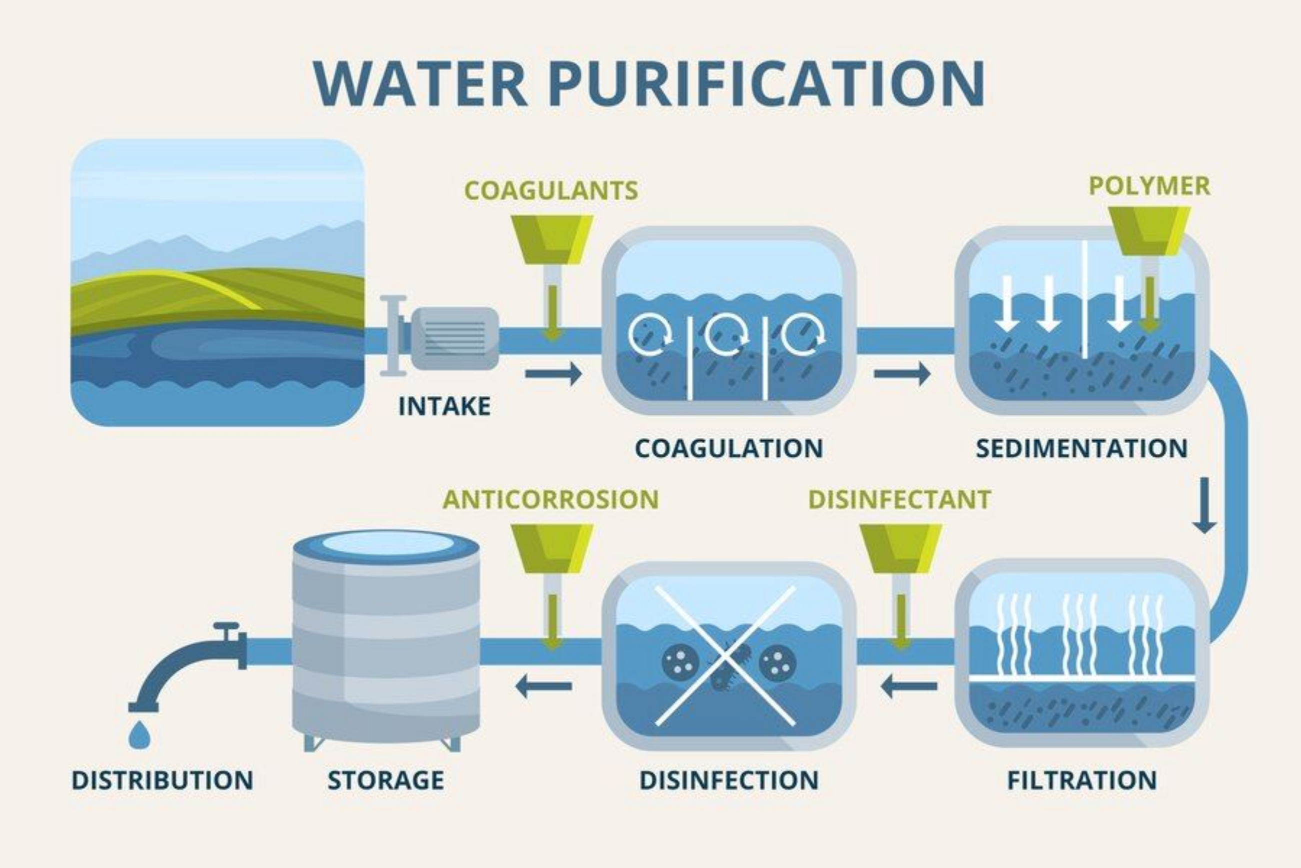
Water Process

Tap Water

- You can open our tap and water flows
- You can use this water for drinking, cleaning, watering your plants ...
- You can trust that the water is clean.

We never have to think about how our water gets to us, but if you ever thought about it, something probably went wrong.

Water Process



Data Engineering

What is it

- The process of taking raw data from multiple sources and making it fit for business operations.

Importance

- **Better decision making**, allows for clean, accurate and up-to-date data allowing for **data-driven decision making**
- **Enhanced Efficiency,** Data scientists and Machine learning engineers can focus on what they are good at
- **Data quality and consistency,** If the process is well design, the data will always be of good quality.
- **Auditing,** Makes it easier to stay compliant with regulations and keep track of business activities.

Who Benefits From Data Engineers

Management and analysts

- Non-technical people, they are not able to manually extract data
- Need to get business insight

Machine Learning Scientists

- Technical people, they are able to work with raw data
- Working with raw data would be time consuming
- Need to have clean data for their operations

Data Scientists

- Technical people, they are able to work with raw data
- They prefer to work with raw data
- Need help acquiring data.

Understanding Data

As the name suggests, data engineers need to know about data. There are a few key things that we need to keep in mind.

- **Formats**
- **Sources**
- **Storage**

Formats

Data comes in many different shapes and forms , data engineers need to be able to understand these forms of data and be able to work with them

- **Structured**
  - Data that has a rigid structure
  - Tabular Databases (SQL)
  - Spreadsheets
- **Semi-Structured**
  - Data that can have some structure, but are more dynamic
  - NoSQL (MongoDB)
- **Unstructured**
  - Seemingly random data or artifacts
  - Media (mp4, png)

Sources

If something contains data, it can be a source of data. As long as it is accessible (both logically and legally), it can be used in your data engineering process

- **Company database**
- **System logs**
- **Social media**
- **APIs**
- **Sensors**
- **Surveys**
- **Biometric**
- **Websites**

Storage

Once we have gathered our data, we need to keep it somewhere. It can be very expensive to run a pipeline for individual queries, so we need to persist our data.

We have a few options

- **Normal Databases**
  - SQL, MongoDB, Neo4j
- **Data Warehouse**
  - RedShift, BigQuery, Snowflake
- **Data Lake**
  - Amazon S3, Azure Blob Storage, Google Cloud Storage

Storage: Database

Databases like SQL or MongoDB can be used to store certain types of data. They are usually a good option if you will continuously store very small data sets.

Pros

- Affordable
- Easy to set up
- Can store Semi-structured data and structured data.

Cons

- Not efficient at querying large datasets

Storage: Data Warehouse

If you are working with extremely large datasets, you will need something that can efficiently query the data that you are working with.

Pros

- Optimized for querying large datasets
- Built with business in mind and integrates with BI tools

Cons

- Very expensive
- Cost per gigabyte is higher than other options.
- Only stores structured data

Storage: Data Lake

Keeping in mind that we might have unstructured data, we need a resource that can allow us to store this type of data.

Pros

- Stores all types of data,
    - Structured, semi-structured, unstructured
- Affordable storage option

Cons

- Slow at performing queries
- Requires extensive management and planning to keep the data useable.

Big Data Problem

Big data is a common problem that data engineers will need to work around. Big data refers to the large amount of data that an organization needs to work with in order to perform their operations.

Three Vs

- **Volume**
  - How much data is being produced and stored
- **Velocity**
  - How often is the data being generated
- **Variety**
  - What are the different formats of data

Sorting Date Well

Without going into the tools that can make working with Big Data, we can take a look at the important considerations that you can apply.

Data Strategy

- Artifacts go to data lake
- Current business data goes to data warehouse
- Infrequently accessed or old data goes to data lake

Data Engineering Workflow

We know how we can collect our data and how we can store our data, but the core part of being a data engineer is performing transformations on the data.

Thoughts

- Who will be working with the data
- Approach for the pipeline
- Pipeline frequency

Data Recipient

In some cases, data engineers need to tailor the output of their pipelines to work for certain business functions.

Business

- Provide key business data
- Data needs to be clean without the need for further processing
- Can be provided through a BI tool providing a simple interface for querying and visualizing data.

Machine Learning

- Requires clean data that directly relates to the operations required
- Can be provided through an API, Database or files.

Data Scientists

- Need to perform their own operations
- Might require raw data and perform their own cleaning operations

Approaching the Pipeline

Once the requirements of the user are known, the approach for the pipeline can be decided.

ETL (Extract, Transform, Load)

- Get the data from our source
- Perform our operations on the data
- Store the data in our storage solution.

ELT (Extract, Load, Transform)

- Get the data
- Store the data in our storage solution
- Transform the data as it is being used

Pipeline Frequency

One important thing we need to know when creating our pipeline is how frequently we need to run our operations.

Batch Processing

- The pipeline will run on a certain schedule (hourly, daily, weekly ...).
- The most common approach as it is cost effective
- Works well when the data we are accessing doesn't frequently change and our business operations can work with data that might be a little bit outdated.

Streaming

- Our pipeline will run constantly
- Used when we need our data to be processed in real time
- More expensive as we need to have multiple servers running 24/7

Becoming a Data Engineer

Technical Skills

- **Programming -** Python, Scala, Java
- **Databases -** SQL (Mandatory), NoSQL
- **Cloud -** AWS, Azure, GCP
- **DevOps Tools -** Docker, Git

Tools

- **Workflow Management -** Airflow, Luigi
- **Streaming -** Kafka

Good to know

- BI Tool (Power BI, Looker Studio)
- Dbt
- Snowflake
- Hadoop

Things to Practice

- Writing Efficient Code
- Data Cleaning
- Database modeling and design

CoGrammar

Thank you for joining!