Lesson 5

Data Pipelines

In this lesson we will talk about:

Data pipelines. Origins and Implementations

Data Capture, Transforming & Analysing

Raw Data. Data Aggregation

Pipeline efficiency

Basic statistical functions

Origins

2025. Present time

IBM

"A data pipeline is a method in which raw data is ingested from various data sources, transformed and then ported to a data store, such as a data lake or data warehouse, for analysis."

"Data pipelines act as the piping for data science projects or business intelligence dashboards. Data can be sourced through a wide variety of places: APIs, SQL and NoSQL databases, files."

"During sourcing, data lineage is tracked to document the relationship between enterprise data in various business and IT applications, for example, where data is currently and how its stored in an environment, such as on-premises, in a data lake or in a data warehouse."

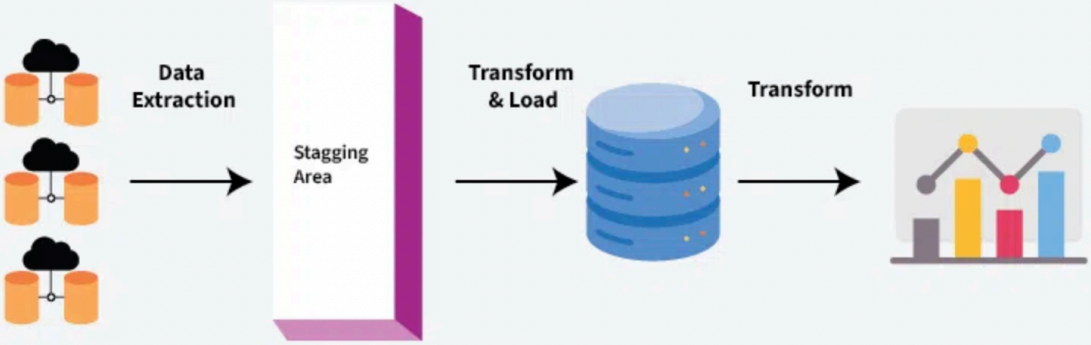
Source: ibm.com

www.geeksforgeeks.org

"Data Pipeline deals with information that is flowing from one end to another. In simple words, we can say collecting the data from various resources than processing it as per requirement and transferring it to the destination by following some sequential activities."

Source: www.geeksforgeeks.org

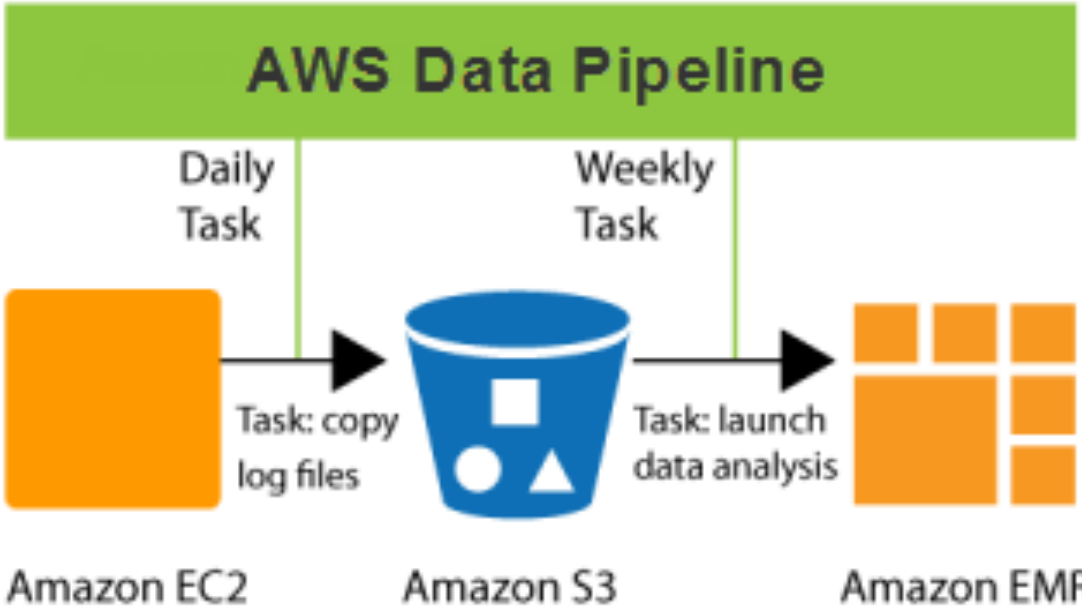
Data Pipeline Overview



AMAZON

"A data pipeline is a series of processing steps to prepare enterprise data for analysis. Organizations have a large volume of data from various sources like applications, Internet of Things (IoT) devices, and other digital channels. However, raw data is useless; it must be moved, sorted, filtered, reformatted, and analyzed for business intelligence. A data pipeline includes various technologies to verify, summarize, and find patterns in data to inform business decisions."

"Just like a water pipeline moves water from the reservoir to your taps, a data pipeline moves data from the collection point to storage. A data pipeline extracts data from a source, makes changes, then saves it in a specific destination. We explain the critical components of data pipeline architecture below."

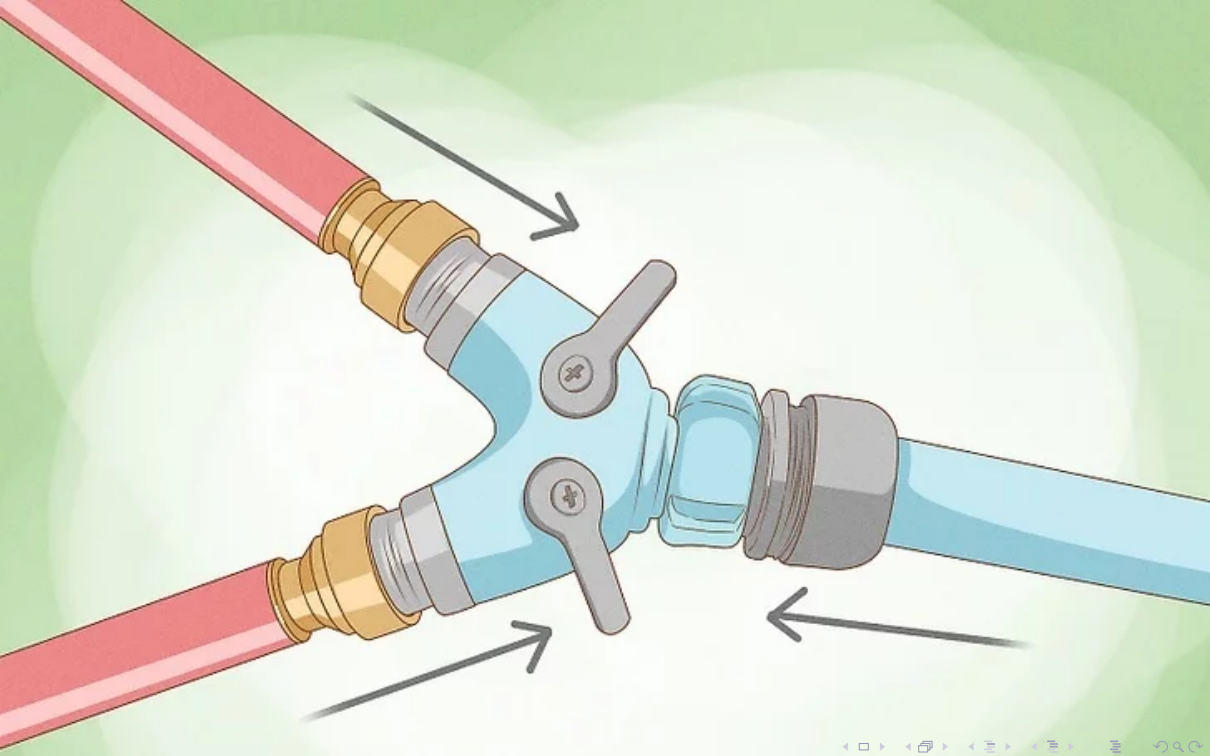

AWS Data Pipeline
Daily Task
Weekly Task
Task: copy log files
Task: launch data analysis
Amazon EC2
Amazon S3
Amazon EMR

Data pipeline = a program or what?
Lets rollback time...

1969. UNIX

The very first data pipelines

As early as 1964 Douglas McIlroy thought about a mechanism how we could connect programs same way like we connect and combine garden hoses together.



Connecting Garden Hoses

- it should connect different hoses
- as efficiently as possible
- with no leaks
- covering all the garden

Pipelines are everywhere

- transport fresh water pipelines
- collection pipelines from ground gas and oil to refineries
- wastewater pipelines
- distribution pipelines

The idea was to reuse the model from other industries to CS

"A pipeline is a mechanism for connecting the output of one program directly and conveniently into the input of another program."

— *Brian Kernighan, UNIX father*

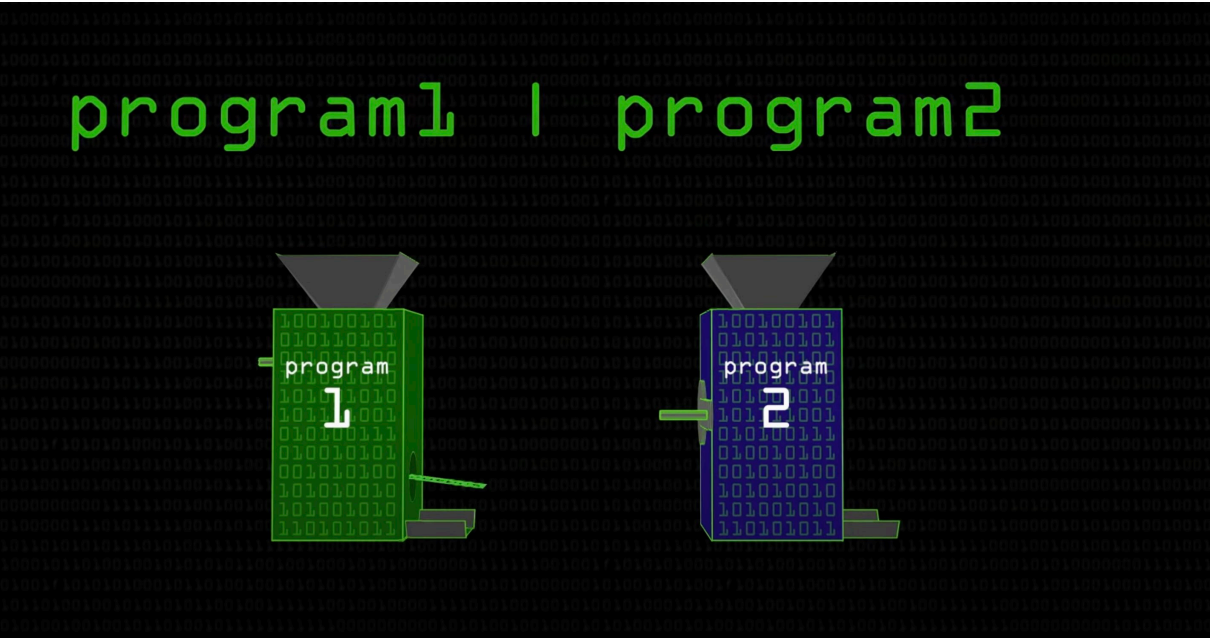
The very first models of data pipelines were defined, introduced and implemented in the operating systems.

Enter UNIX

UNIX pipelines

The very first data pipeline was implemented for UNIX based operating systems. The ideas was very simple: combine one or many programs using the **pipe** operator |

program | program | ...


program1 | program2
program
1
program
2

The pipeline: is connecting one or many programs together! It is not a single monolithic program.

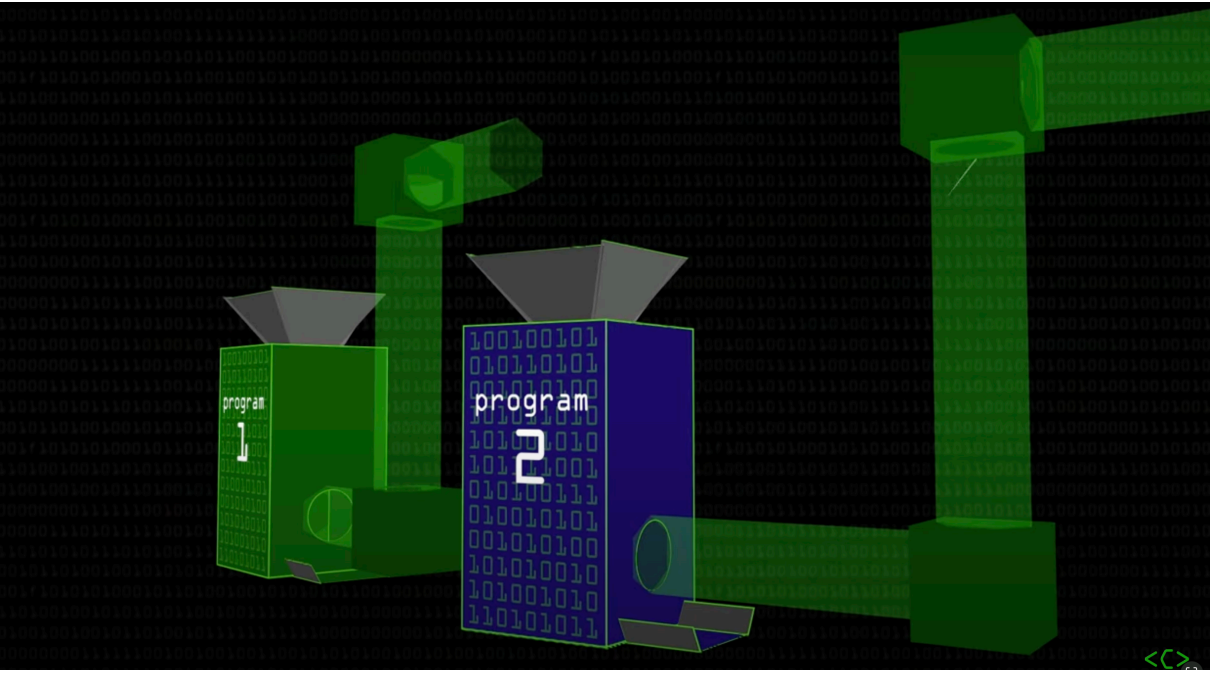
UNIX pipelines

But why connect different programs instead of having a single big program to handle all the input data?

Few reasons behind... have small, **modular** programs which can connect together to solve a problem quickly and as modular as possible

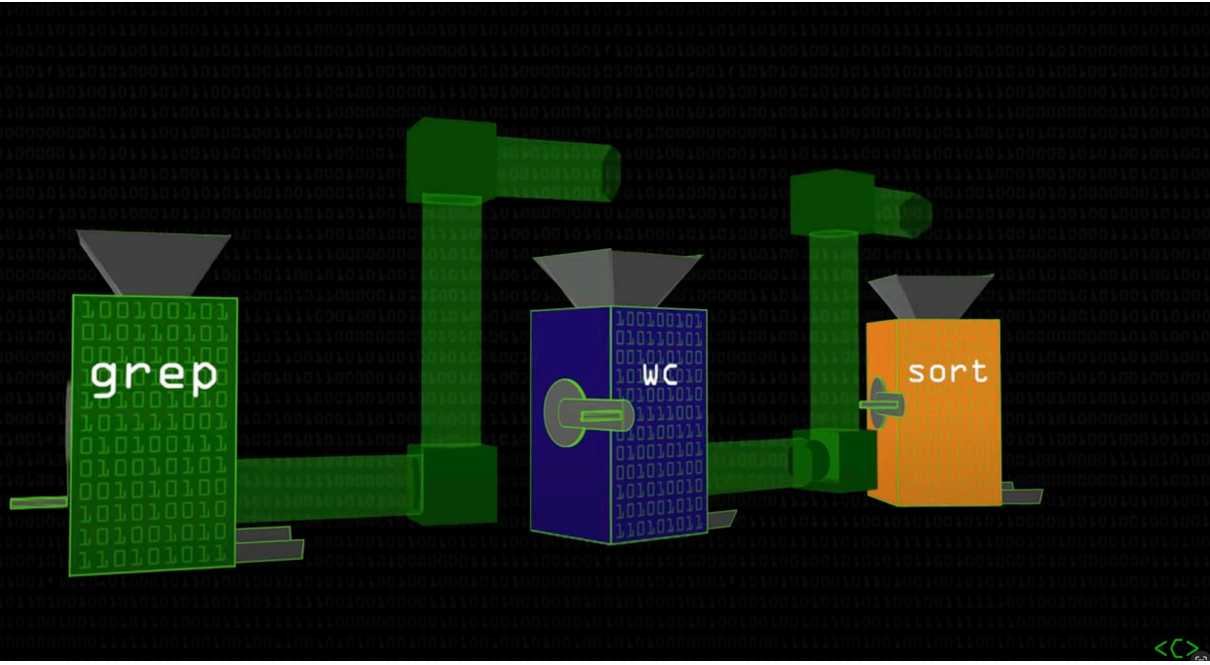
And enforce a strict order of execution!!!

- smaller programs are easy to develop
- and are simple and cheaper to maintain
- as well the amount of data might not fit one program
- no temporary files or data in between


program
1
program
2

UNIX pipelines

Connects different UNIX programs: like ls, sort, wc, grep to solve a problem using the | pipe character


grep
wc
sort

UNIX pipelines, example

```
$ ls -lrt Images/*.png | wc -l
      140
```

Sort all images which were produced during Feb, example

```
$ ls -l Images/ | awk '{print $6,$7,$9}' | grep Feb | sort
Feb 26 treevsgraph.png
Feb 26 treevsgraph2.png
Feb 26 tsp.png
Feb 3 Alan_turing.jpg
Feb 3 Bombe.jpg
Feb 4 Queue.png
...
Feb 4 factorial2.png
Feb 4 recursion.png
```

Sort all images which were produced during Feb, example

```
$ ls -l Images/ | awk '{print $6,$7,$9}' | grep Feb | sort -V
Feb 19 hash_function4.png
Feb 19 hash_function5.png
Feb 25 binary-tree.png
Feb 26 btree.png
Feb 26 depth-height.png
Feb 26 directed-graph.png
Feb 26 graph.png
Feb 26 selfbalanced-tree.png
Feb 26 treevsgraph.png
...
```

But how about the pipeline's speed and overall performance? Should we care about?

Search for all png images measuring the elapsed time, example

```
$ time find $HOME -type f | grep png
...
/myhome/feynman_path/examples/no-entanglement.png
/myhome/feynman_path/examples/no-interference-circuit.png
/myhome/feynman_path/examples/entanglement.png
/myhome/feynman_path/examples/no-entanglement-circuit.png

real 0m9.167s
user 0m0.450s
sys 0m3.147s
```

Search and sort for all jpg and png images measuring the elapsed time, example

```
$ time find $HOME -type f | egrep 'png|jpg' | sort -V
 ...
/myhome/.vscode/extensions/sumneko.lua-3.6.3-darwin-arm64/client
/myhome/.vscode/extensions/sumneko.lua-3.6.3-darwin-arm64/client
/myhome/.vscode/extensions/sumneko.lua-3.6.3-darwin-arm64/client
/myhome/.vscode/extensions/sumneko.lua-3.6.3-darwin-arm64/image
/myhome/.vscode/extensions/vadimcn.vscode-lldb-1.8.1/images/lldb.p

real 0m9.240s
user 0m0.773s
sys 0m3.034s
```

Performance matters! Measuring pipeline execution time matters! Take a look...

Search and sort for all jpg and png images measuring the elapsed time under Windows, example

$ time find /myhome −**type** f | grep png

...

/myhome/.vscode/extensions/osi−certified−72x60.png
/myhome/.vscode/extensions/osi−certified−72x60.png
/myhome/.vscode/extensions/osi−certified−72x60.png
/myhome/.vscode/extensions/images/logo.png
/myhome/.vscode/extensions/images/lldb.png

real	2m20.786s
user	0m0.421s
sys	0m1.686s