



Data Engineering Solutions

# DataPyro : Data Engineering

*“The world's most valuable resource is no longer oil, but data”*

## Data Platform

We build Data Platforms that allow you to run your ETL/ELT jobs in a scalable fashion. The platform also automates training and deploying Machine Learning models for you. We can help modernizing your outdated ETL tools with Cloud agnostic Data Engineering tools.

## Data Lake

A Data Lake is a central hub for storing and analyzing any structured and unstructured data in **any scale**. We can set up your Data Lake, ingest data from your operational databases, data warehouses or any data sources in both batch or real-time, build models on top of it and make the data accessible to all your organization.

## Clickstream Analytics

Gather information from your website, mobile application or IoT device, track every click and user behaviour. Our Clickstream Analytics solution provides you to collect, store and analyze your massive data in an easy way. You can use the data to improve your service quality or customer satisfaction.

# Solution #1: Data Platform

a Data Platform is an infrastructure that enables organization to build scalable data pipelines.

## Integration

Our Data Platform solution has built-in integration with the most common data sources and it's easy to integrate custom data sources as well.

## Flexibility

The platform infrastructure is flexible, can be built on any cloud or on-premise infrastructures.



## Scalability

Our platform supports various containerization management systems such as ECS, Kubernetes, etc. to scale and run its workloads.

## Cost Effectiveness

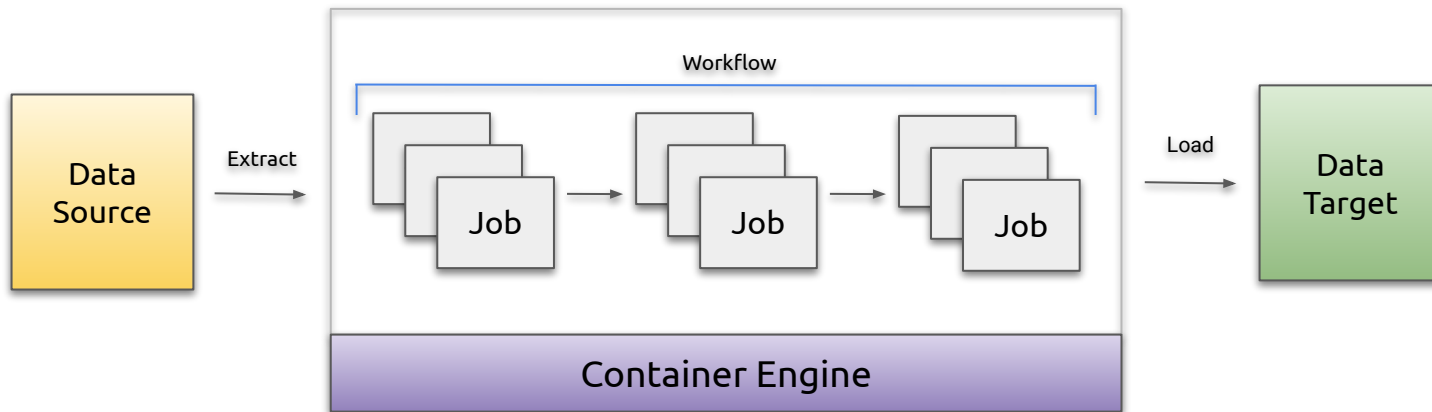
The workloads can run distributed on relatively smaller instances, so they run much more faster and don't rely on expensive hardware.

## Automation

The platform is built with infrastructure-as-code principles. The workloads can be integrated to CI/CD pipelines.

# Solution #1: Data Platform

## Batch Data Processing

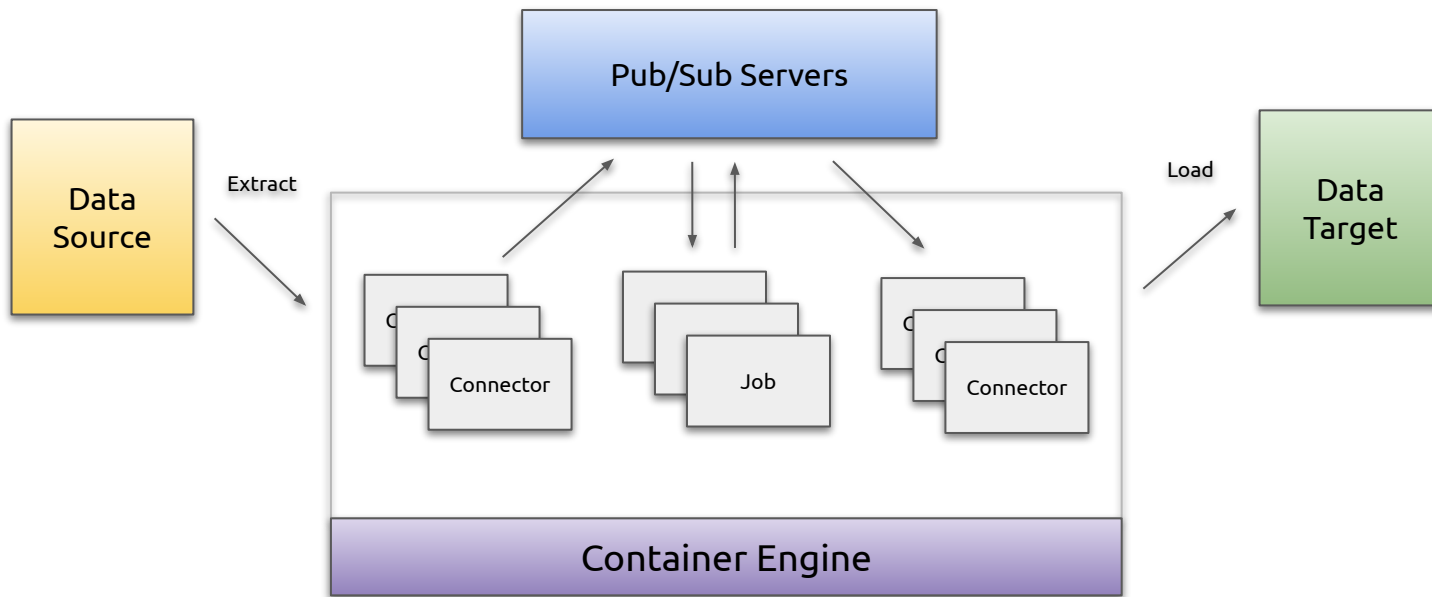


**Batch data processing is efficient and cost effective for most use-cases.**

The platform can process billions of rows of data every day.

# Solution #1: Data Platform


















## Real-time Data Processing



Real-time data processing is crucial for some specific use-cases like Finance.

# Solution #1: Data Platform

## Built-in Supported Data Sources

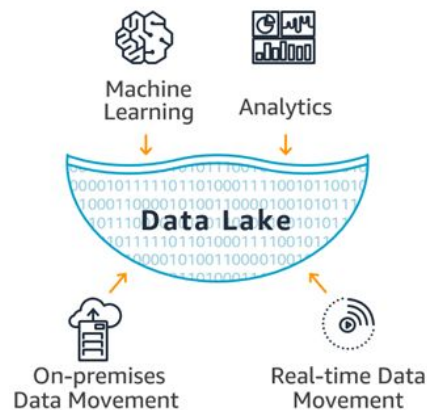
RDBMS	NoSQL	DWH / DL	Web / FTP
 Microsoft SQL Server	 mongoDB  cassandra	 Amazon Redshift	 REST API
 PostgreSQL	 Amazon DynamoDB  APACHE HBASE	 Amazon S3	 FTP
 MySQL	 redis  Amazon DocumentDB	 hadoop HDFS	
 ORACLE DATABASE	 Apache Solr  elastic		

# Solution #2: Data Lake

a **Data Lake** is a central hub for storing and analyzing any structured and unstructured data in any scale

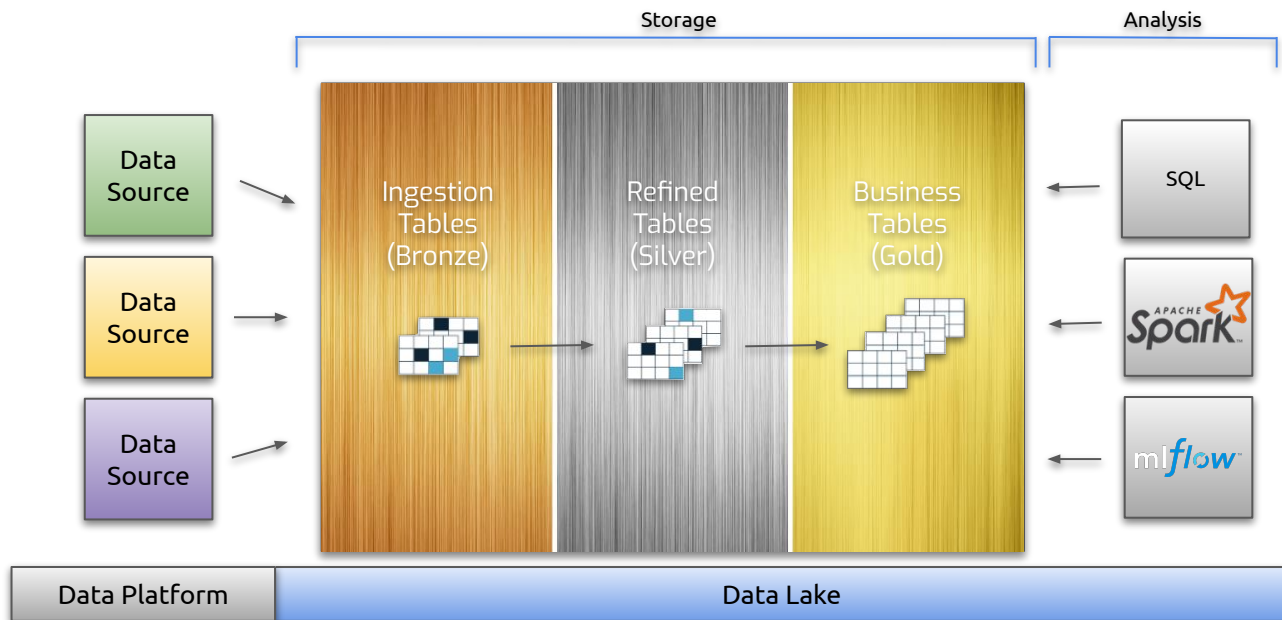
## Key advantages of a Data Lake

- **Data Modeling:** Store and analyze TB/PB of data, build models on top of it without need to change the actual data
- **Scalability:** Can scale both storage and processing separately based on demand.
- **Availability:** Cloud technologies allow you to start building a Data Lake within minutes. Failed resources will be replaced automatically.
- **Maintainability:** Anything can be automated using infrastructure as code technologies.
- **Low Cost:** No need to pay for licences, no need for expensive storage hardware, etc.



# Solution #2: Data Lake

## Data Lake Architecture





# Solution #2: Data Lake

## Data Warehouse vs Data Lake

	Data Warehouse	Data Lake
<b>Data</b>	Relational from transactional systems, operational databases, and line of business applications	Non-relational and relational from IoT devices, web sites, mobile apps, social media, and corporate applications
<b>Schema</b>	Designed prior to the DW implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
<b>Price / Performance</b>	Fastest query results using higher cost storage	Query results getting faster using low-cost storage
<b>Data Quality</b>	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (ie. raw data)
<b>Users</b>	Business analysts	Data scientists, Data developers, and Business analysts (using curated data)

# Solution #3: Clickstream Analytics

**Clickstream Analytics** allows organization to get more insight about the customers and enables new opportunities.



## **Know your customer**

Clickstream Analytics allows you to get to know your customers better. Analyze, measure and improve your products, engage better.

## **See the new opportunities**

Understand the trends in near real-time, build strategies.

## **Build campaigns**

Customize your campaigns, target user segments or even personalize them, improve conversion.

# Tech Stack

Powered by Open Source Data Engineering Tools



# References

*MODA*  
*CRUZ*

**C O T Y**  
BEAUTY, LIBERATED

 **OakNorth**

 **Streetbees**



**DataPyro Ltd**

105, Thornsbeach Road, London, SE6 1EY, UK

Email: [info@datapyro.co.uk](mailto:info@datapyro.co.uk)

<https://www.datapyro.co.uk>