

A blue-tinted photograph of two women in a server room. They are standing in a hallway lined with server racks, both holding and looking at tablets. The scene is dimly lit with blue light reflecting off the server units and the floor.

Big Data para Cientista de Dados

 Stack Academy

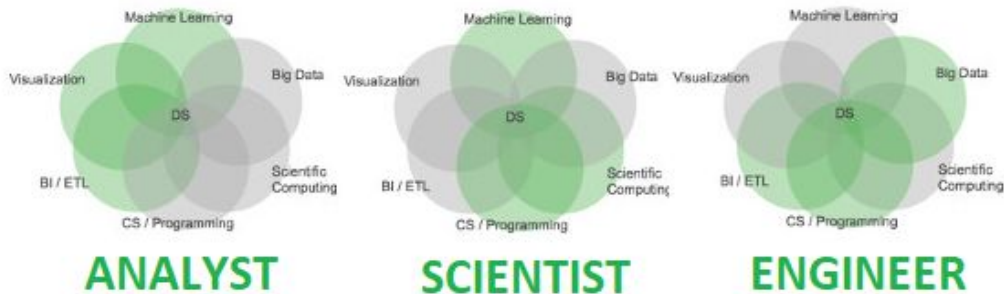
Para quem é esse curso?

1. Cientista de Dados.
2. Analista de Dados.
3. Gestores de Projeto.



Para quem é esse curso?

1. Cientista de Dados.
 2. Analista de Dados.
 3. Gestores de Projeto.
-



Responsabilidades do Cientista de Dados



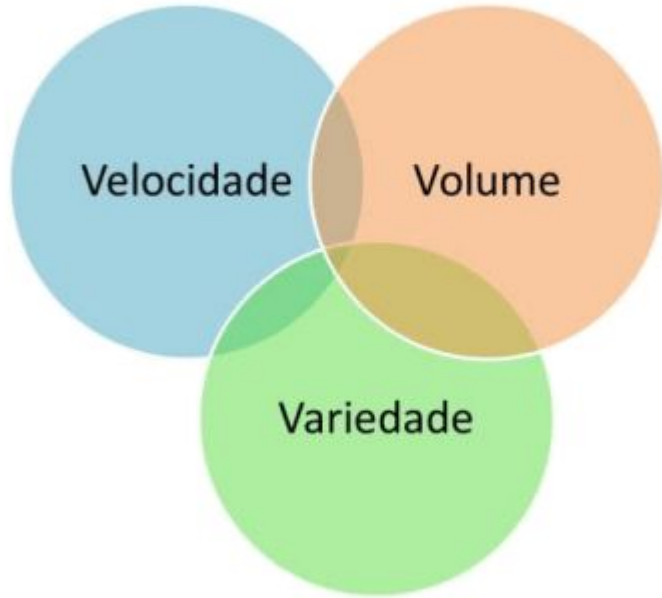
Responsabilidades do Engenheiro de Dados



O que é Big Data?

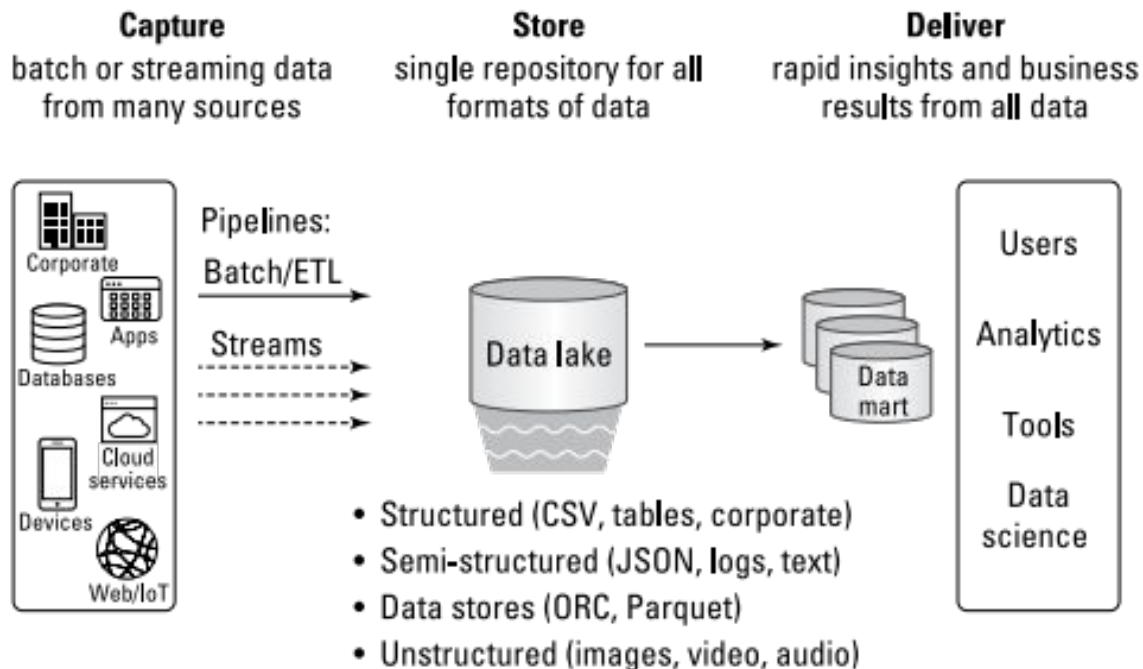


O que é Big Data?

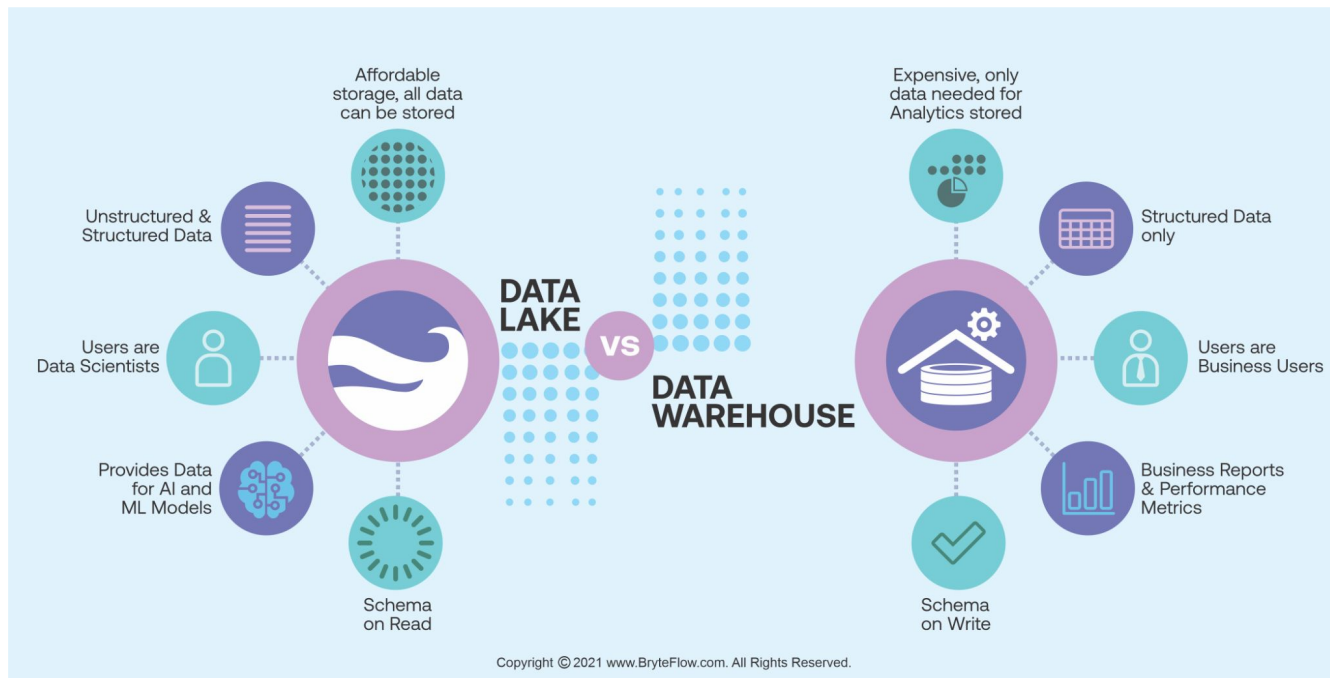


O que são Data Lakes?

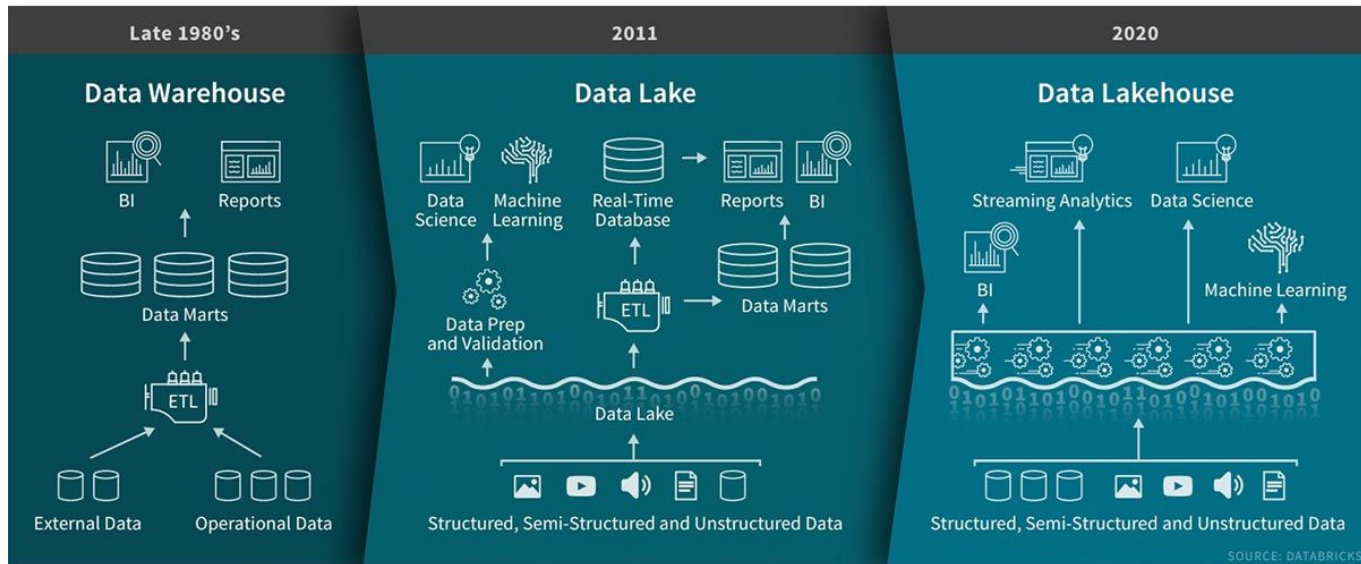
The Traditional Data Lake



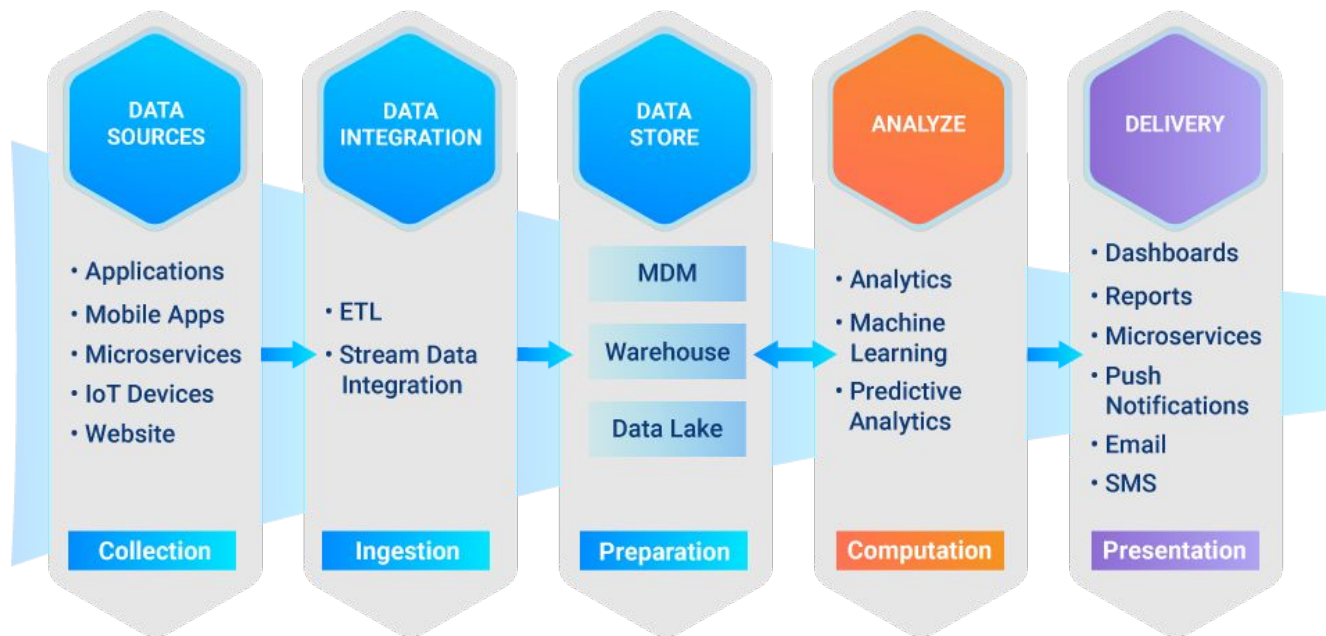
Data Warehouses Vs Data Lakes



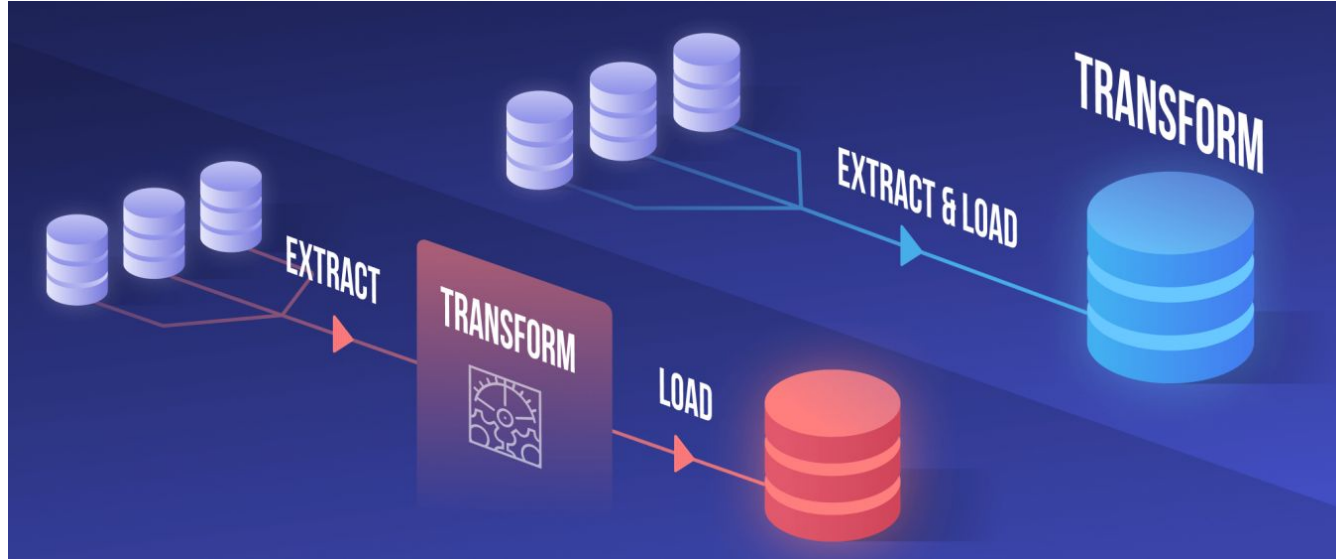
Data Warehouse vs Data Lake vs Data Lakehouse



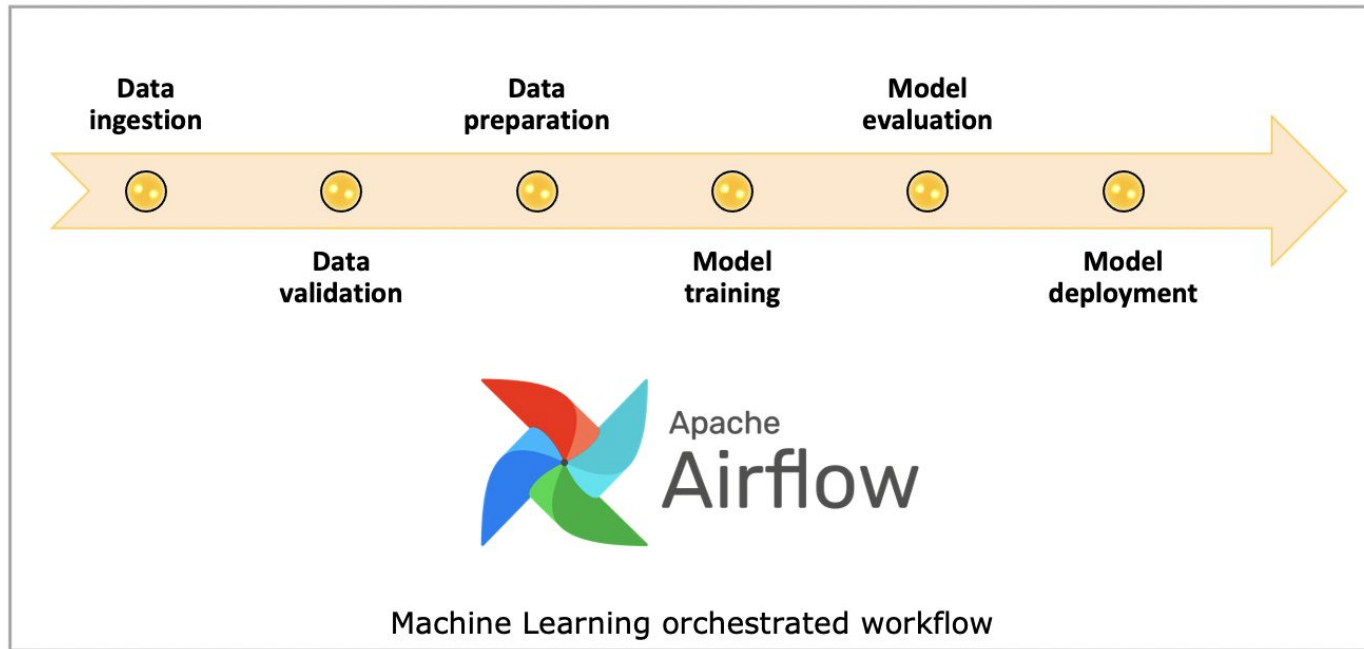
BIG DATA PIPELINE



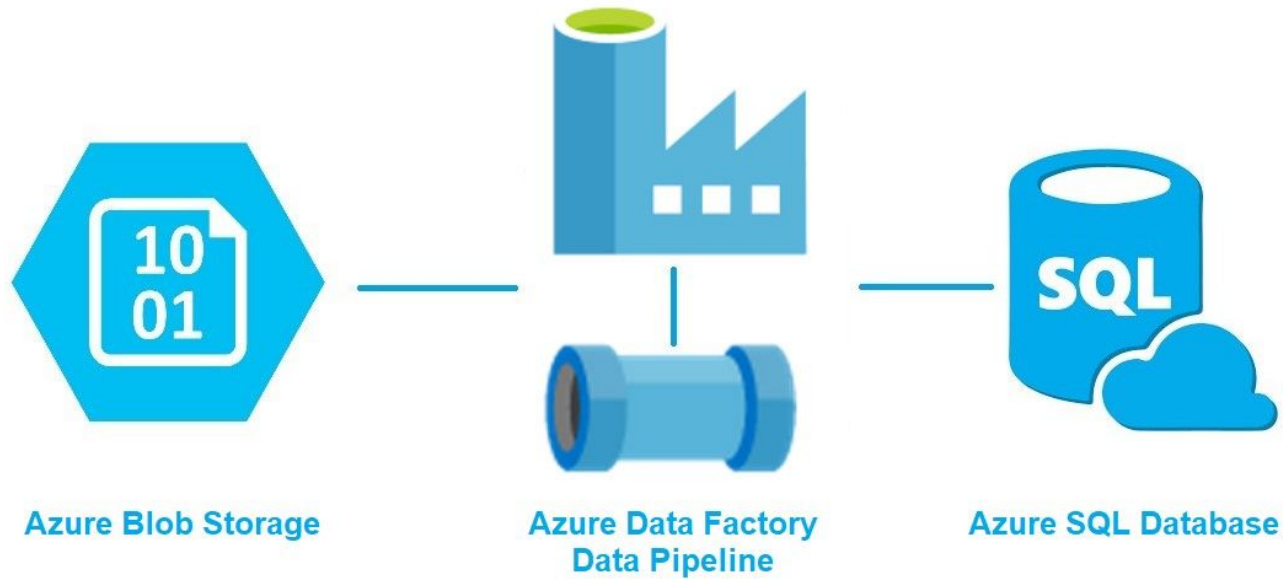
ETL vs ELT



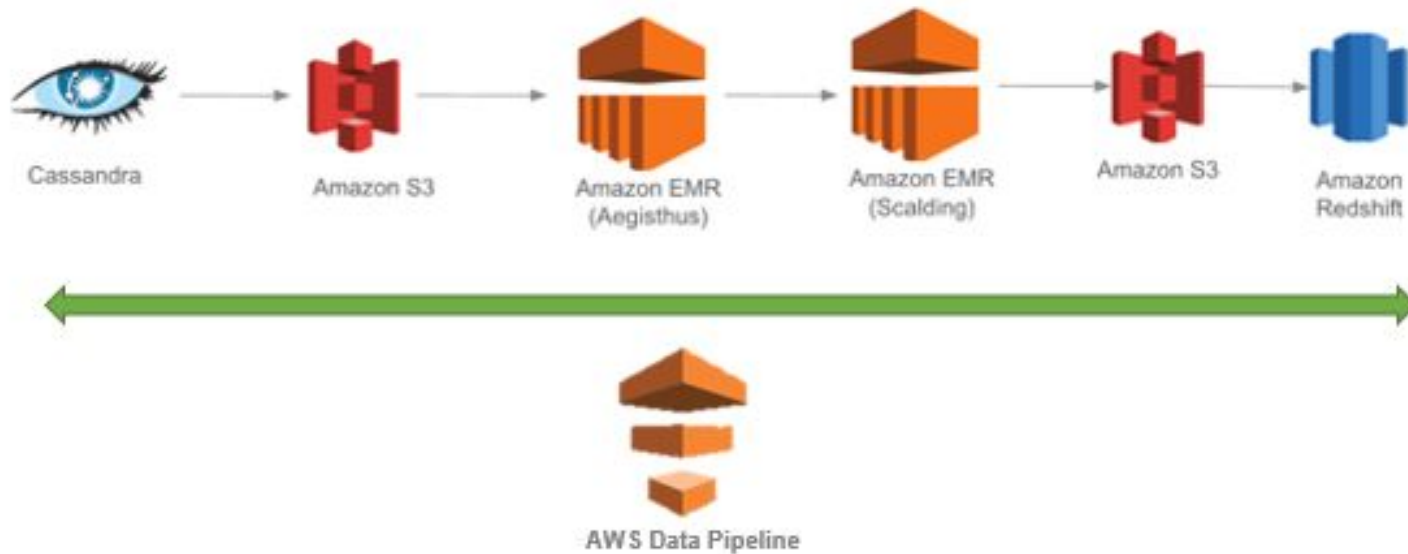
Soluções para Data Pipelines



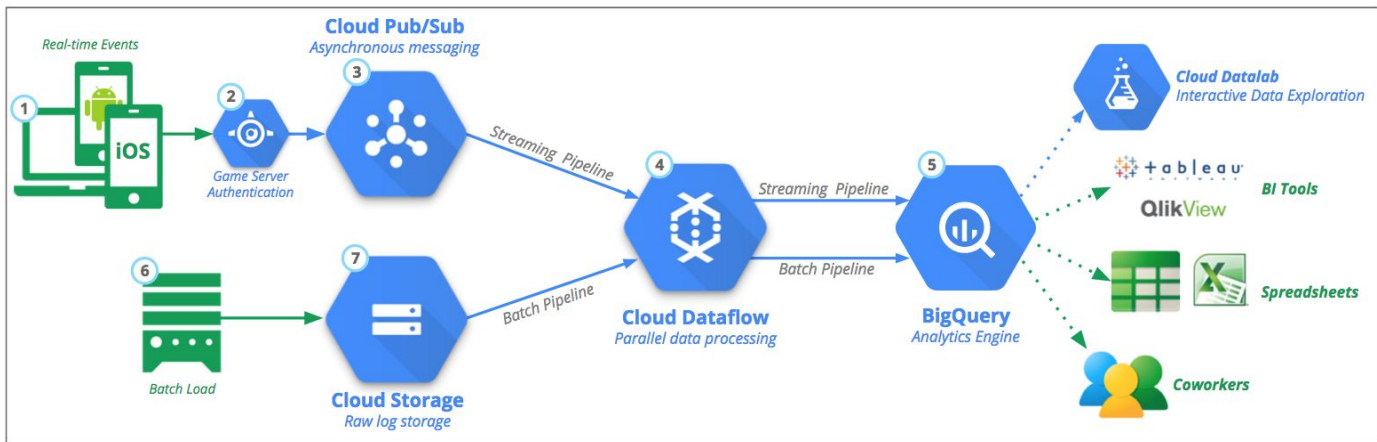
Soluções para Data Pipelines



Soluções para Data Pipelines



Soluções para Data Pipelines



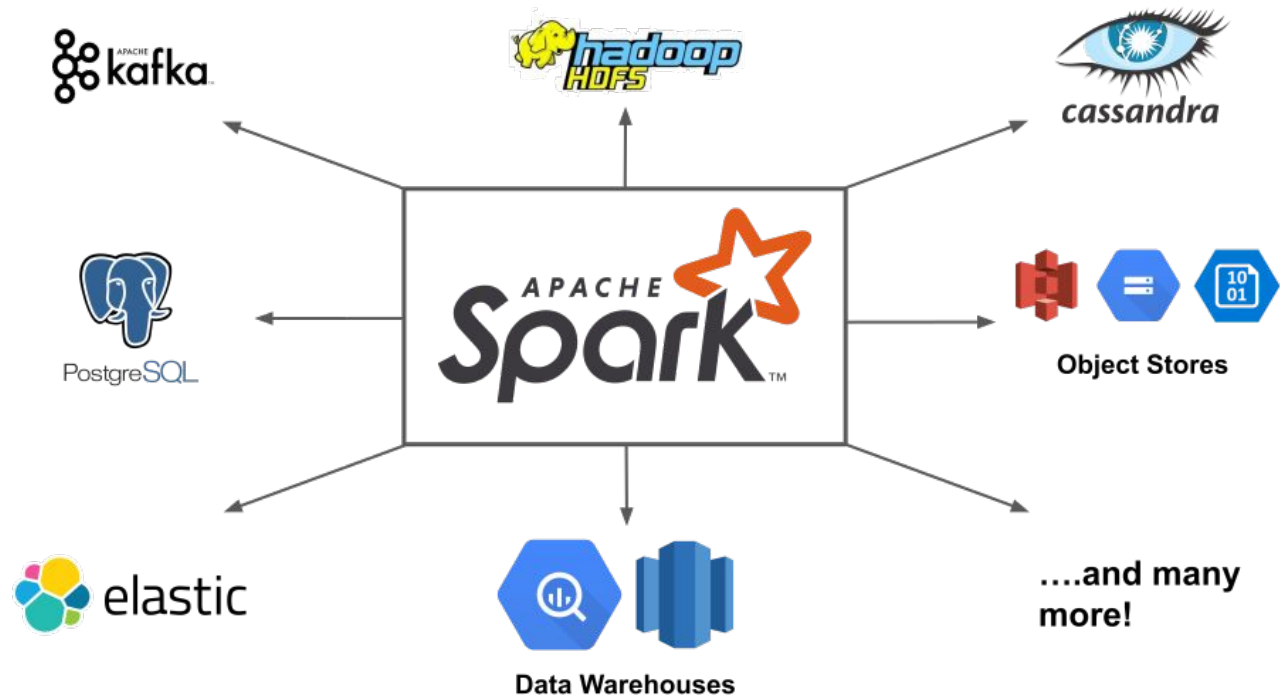
O que é o Apache Spark?



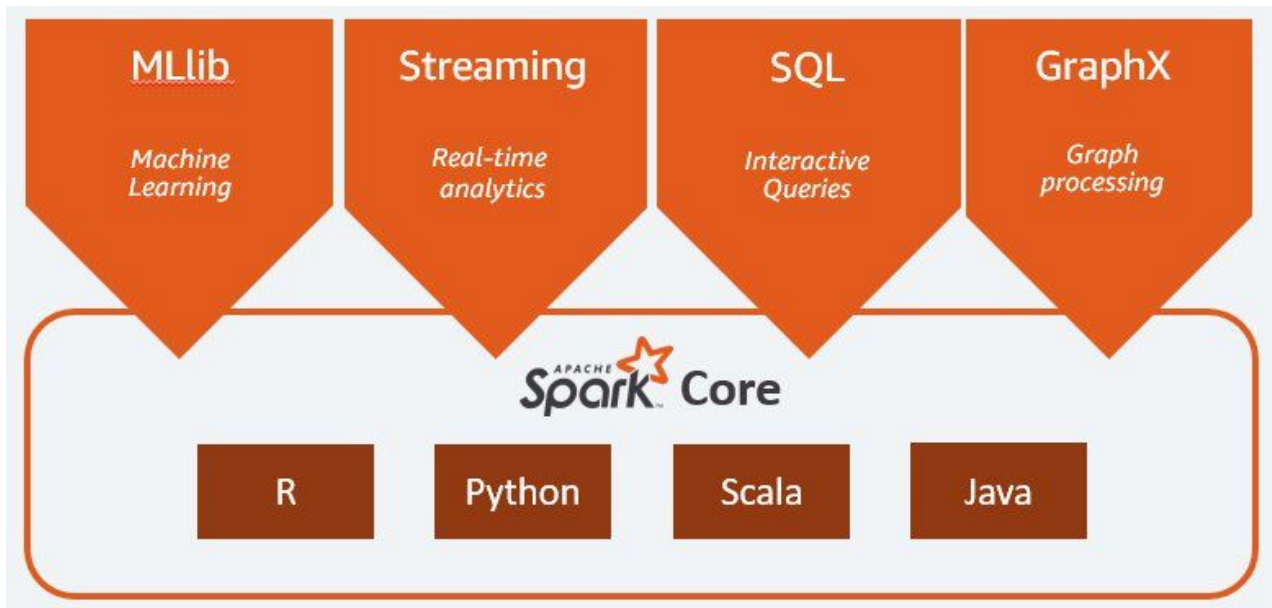
O que é o Cluster?



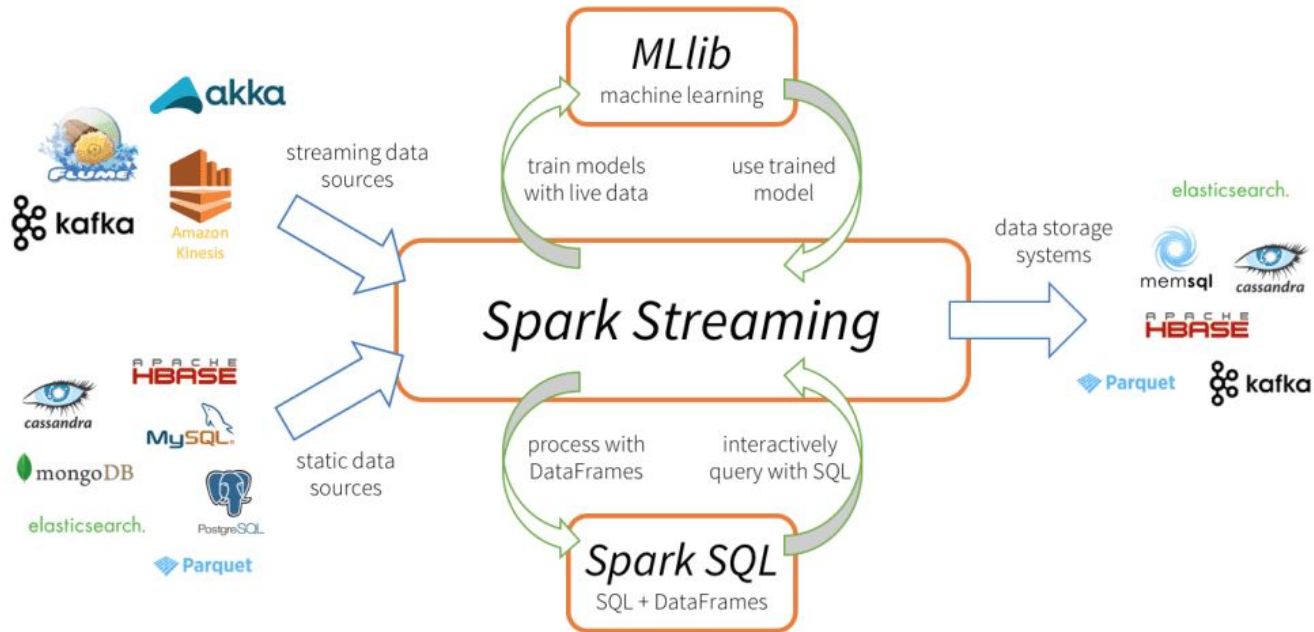
O que é o Apache Spark?



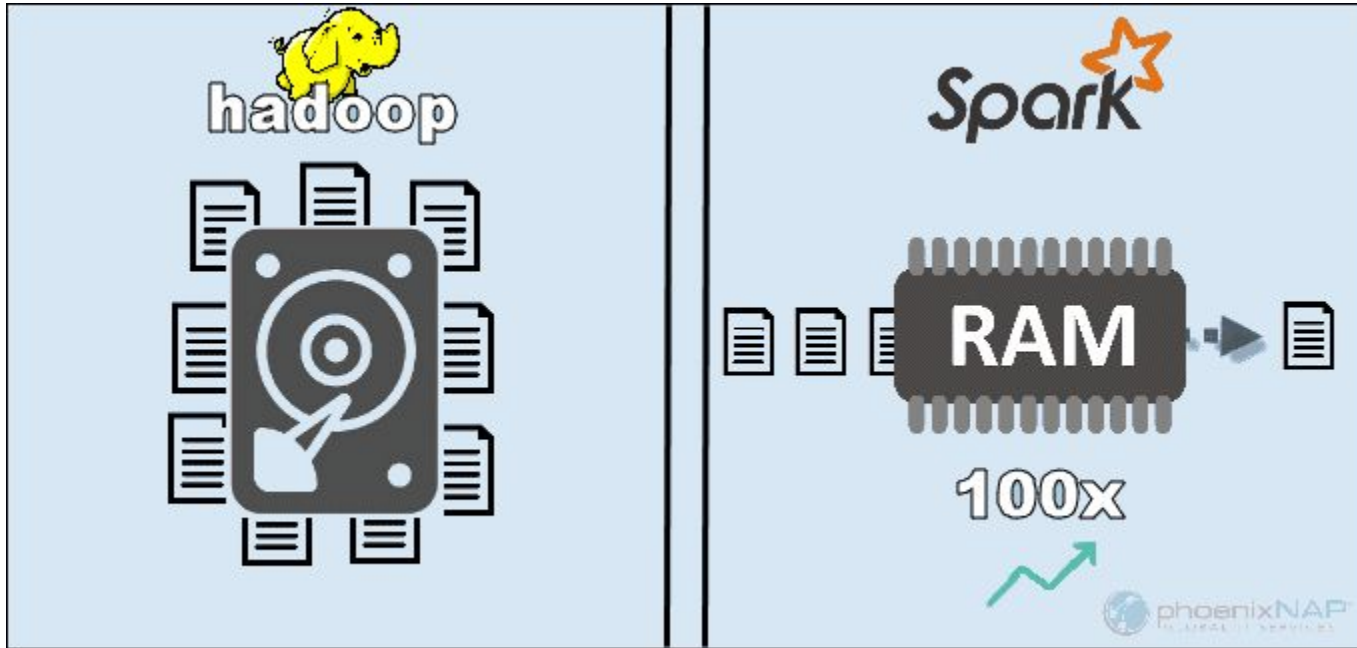
O que é o Apache Spark?



O que é o Apache Spark?



Spark Vs Hadoop



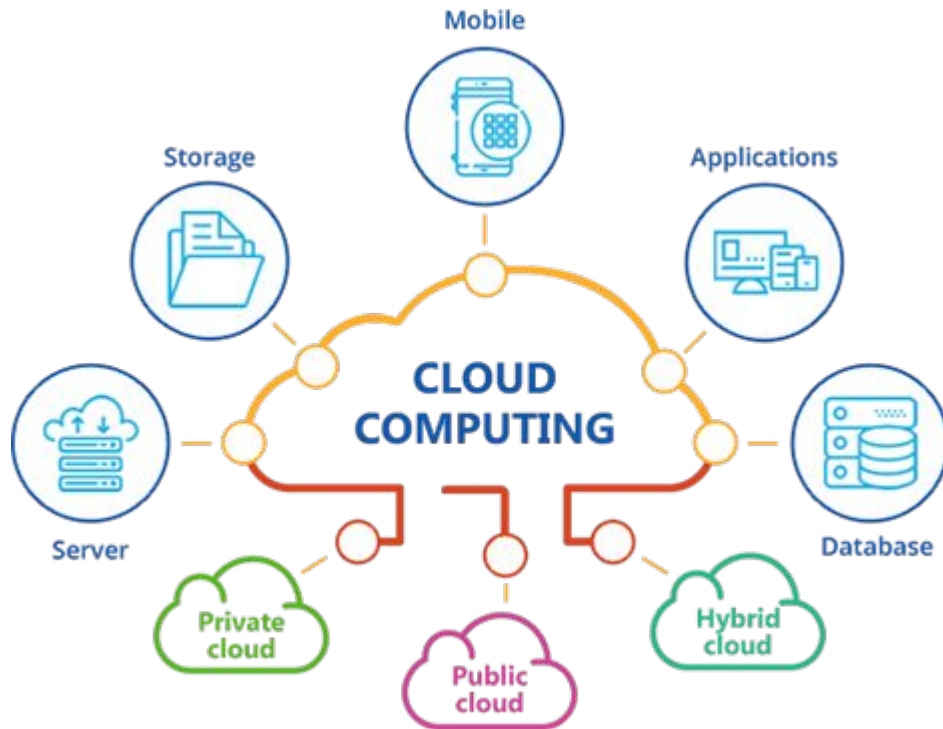
Spark Vs Hadoop

Factors	Spark	Hadoop MapReduce
Speed	100x times than MapReduce	Faster than traditional system
Written In	Scala	Java
Data Processing	Batch / real-time / iterative / interactive / graph	Batch processing
Ease of Use	Compact & easier than Hadoop	Complex & lengthy
Caching	Caches the data in-memory & enhances the system performance	Doesn't support caching of data

O que é Databricks?



Por que Databricks?



Por que Databricks?



Google Cloud

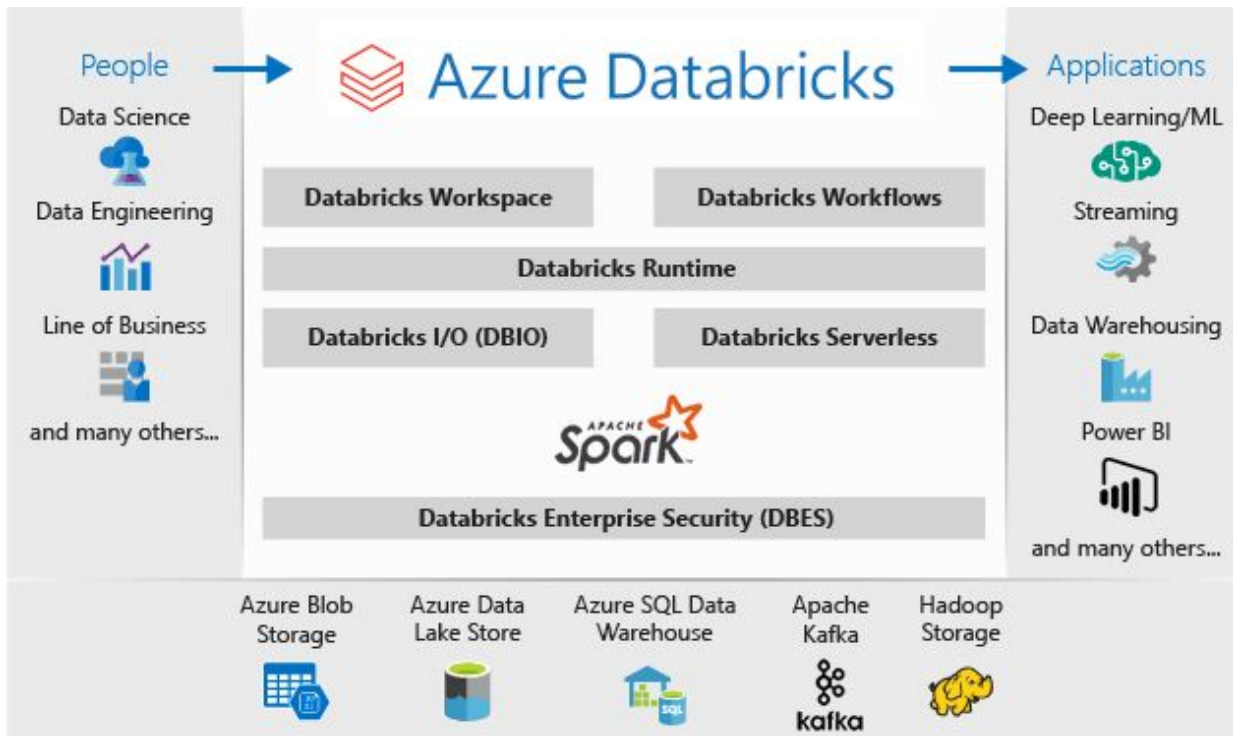


databricks



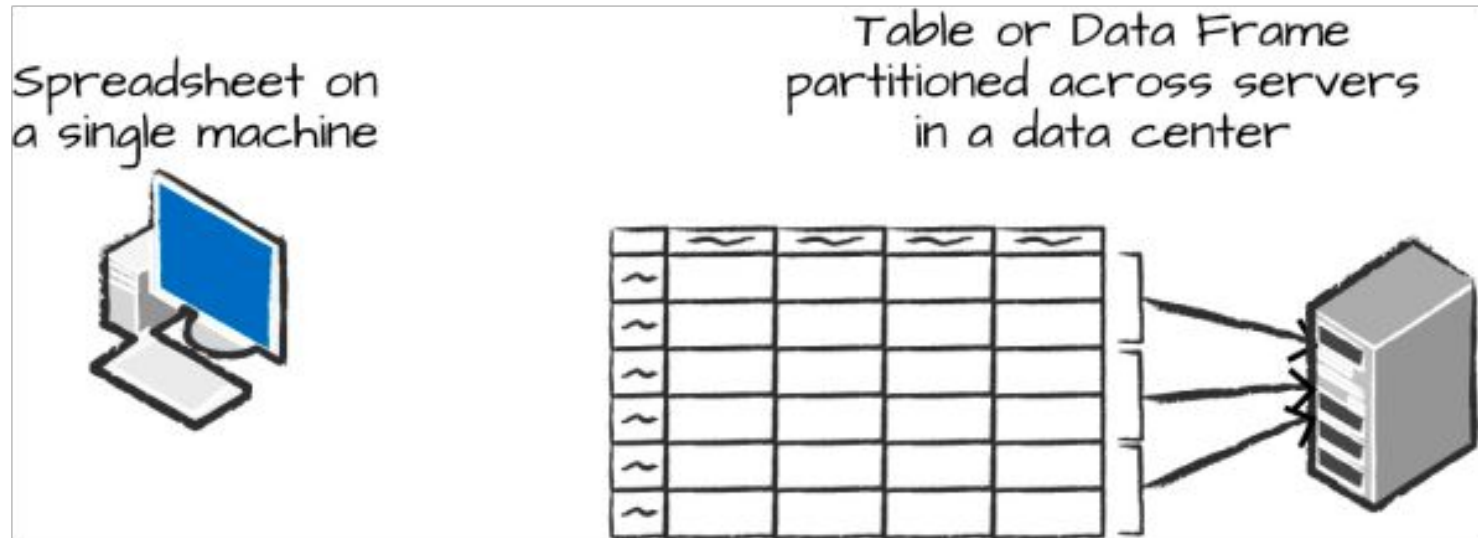
Stack

Por que Databricks?



Hands on!

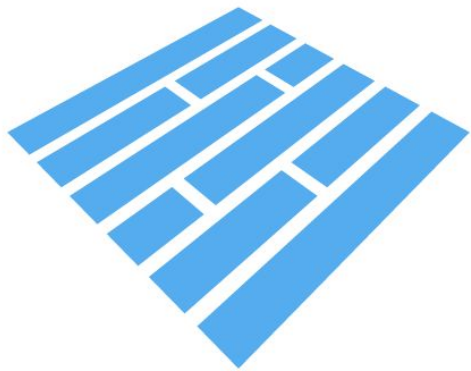
Dataframes e Partições no Spark



SQL vs Python vs Scala



Apache Parquet



Parquet

Apache Parquet

Data In Columns On Disk

Title	Date	Chart

Row-Oriented data on disk

Led Zeppelin IV	11/08/1971	1	Houses of the Holy	03/28/1973	1	Physical Graffiti	02/24/1975	1
-----------------	------------	---	--------------------	------------	---	-------------------	------------	---

Column-Oriented data on disk

Led Zeppelin IV	Houses of the Holy	Physical Graffiti	11/08/1971	03/28/1973	02/24/1975	1	1	1
-----------------	--------------------	-------------------	------------	------------	------------	---	---	---

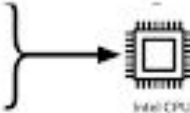


Apache Parquet

	day	location	product	sale
row 1	2017-01-01	l1	p1	300
row 2	2017-01-01	l1	p2	40
row 3	2017-01-01	l2	p1	44
row 4	2017-02-01	l1	p1	200

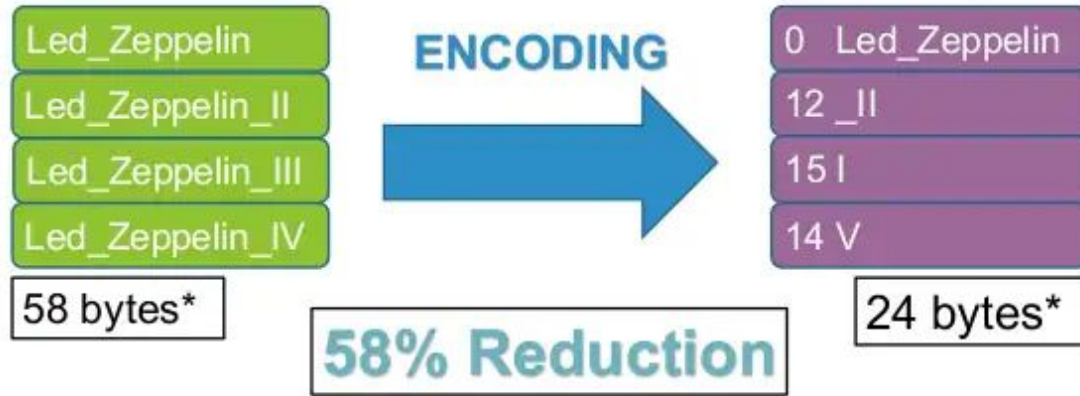
Traditional Memory Buffer	
row 1	2017-01-01
	l1
	p1
	300
row 2	2017-01-01
	l1
	p2
row 3	40
	2017-01-01
	l2
	p1
	44

Columnar Storage	
day	2017-01-01
	2017-01-01
	2017-01-01
	2017-01-02
location	l1
	l1
	l2
	l1
product	p1
	p2
	p1
	p1



Apache Parquet

Encoding: Incremental Encoding

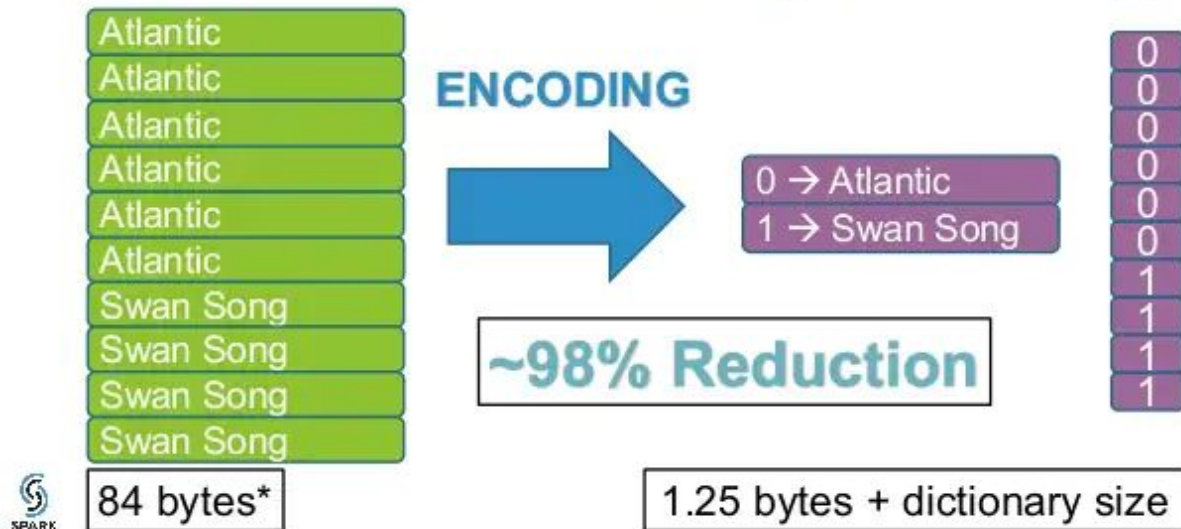


*not counting delimiters



Apache Parquet

Encoding: Dictionary Encoding



Apache Parquet

Partitioning

```
dataFrame
  .write
  .partitionBy("Whatever", "Columns", "You", "Want")
  .parquet(outputFile)

// For a common example
dataFrame
  .write
  .partitionBy("Year", "Month", "Day", "Hour")
  .parquet(outputFile)
```



Apache Parquet



VS



- Não existe estatísticas dos dados no arquivo.
- É preciso ler todo o arquivo para definição dos data types.
- Não permite compressão por coluna.
- Ocupa mais espaço e mais processamento.

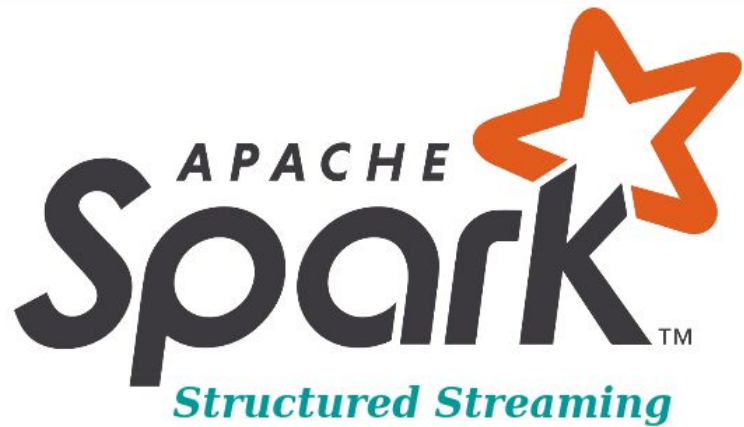
Apache Parquet

The following table compares the savings as well as the speedup obtained by converting data into Parquet from CSV.

Dataset	Size on Amazon S3	Query Run Time	Data Scanned	Cost
Data stored as CSV files	1 TB	236 seconds	1.15 TB	\$5.75
Data stored in Apache Parquet Format	130 GB	6.78 seconds	2.51 GB	\$0.01
Savings	87% less when using Parquet	34x faster	99% less data scanned	99.7% savings

Hands on!

Spark  Streaming



Transações com cartões de crédito



Clicks em um site..



Internet das coisas.



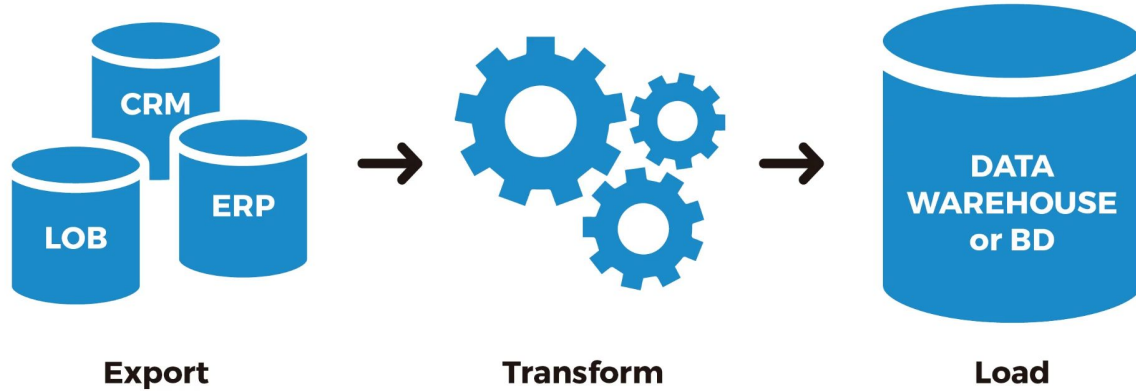
Notificações



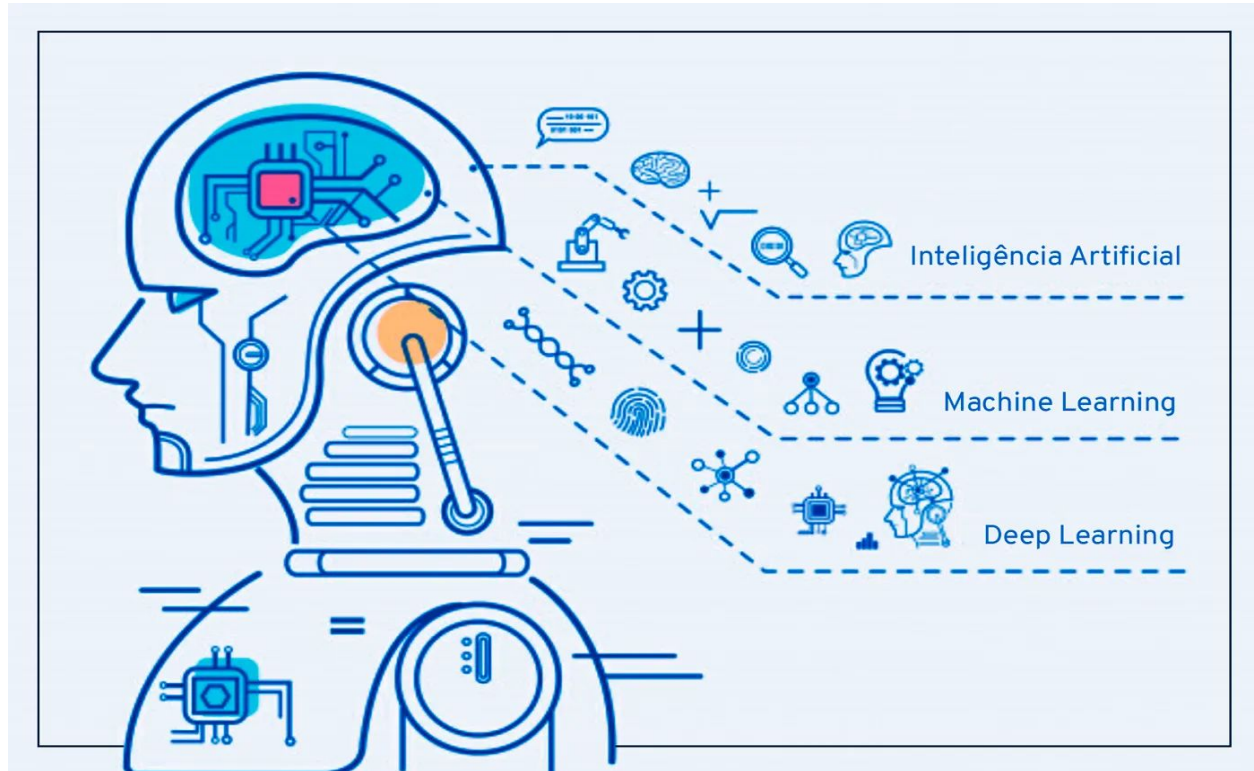
Dashboards



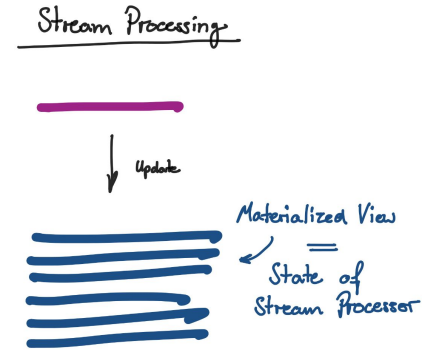
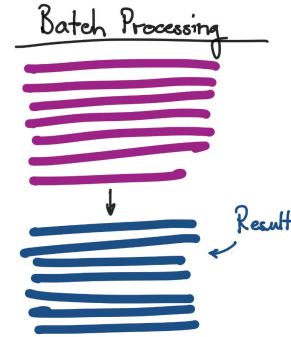
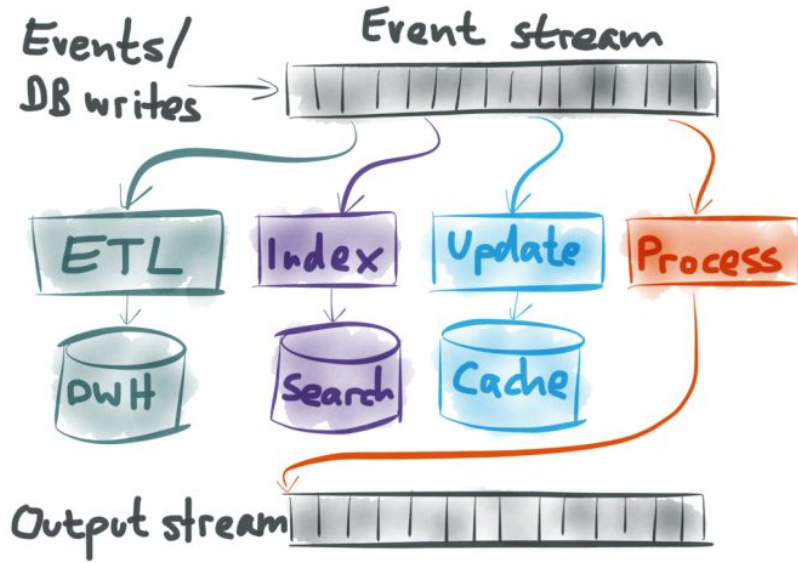
Incremental ETL



Online Machine Learning



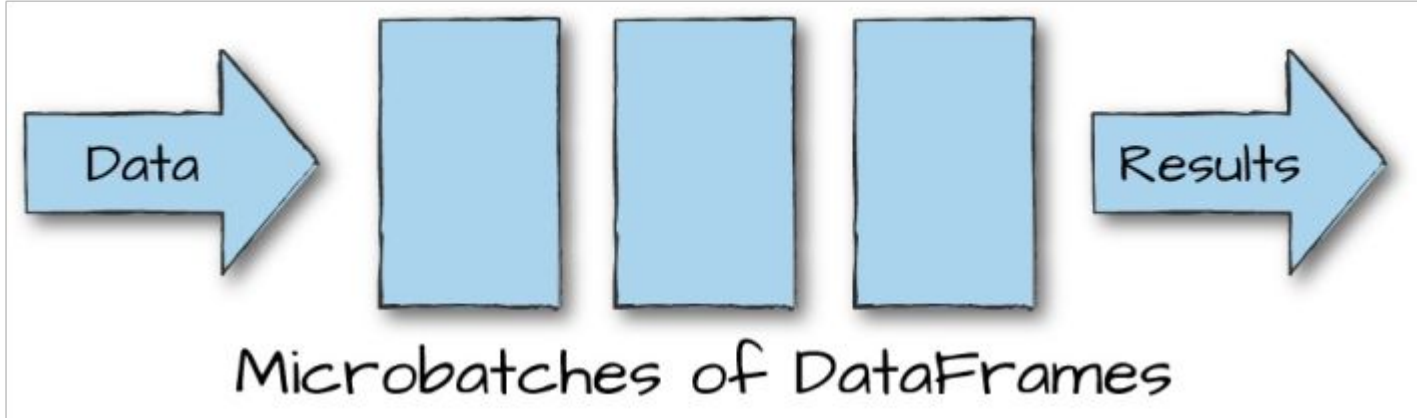
Como funciona o Streaming?



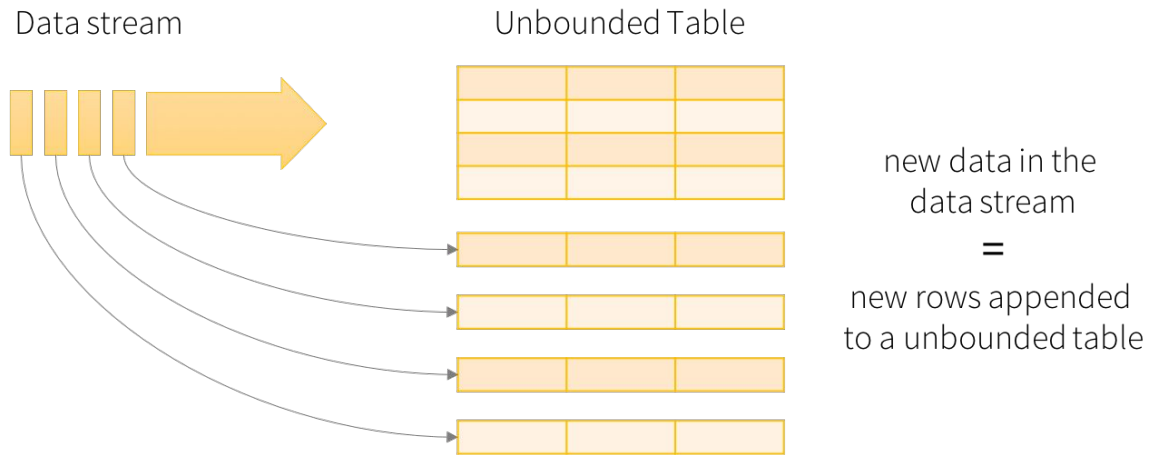
Spark Streaming



Spark Streaming

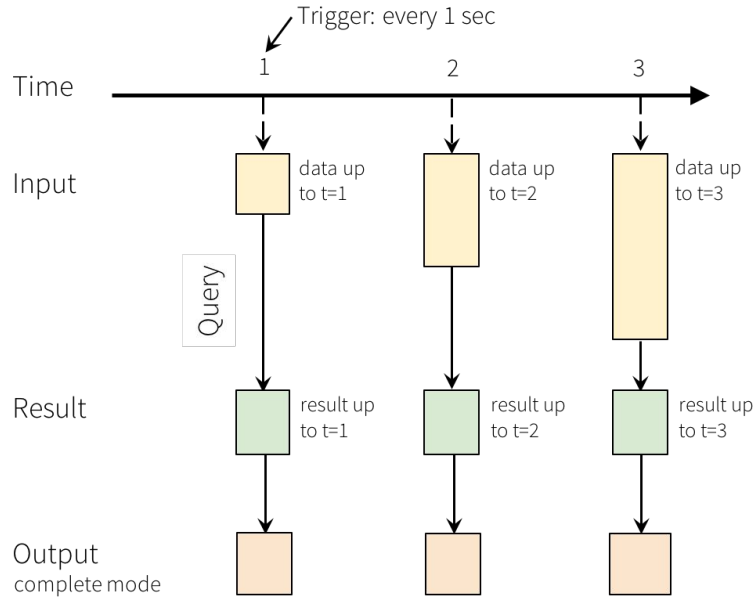


Spark Streaming



Data stream as an unbounded table

Spark Streaming



Programming Model for Structured Streaming

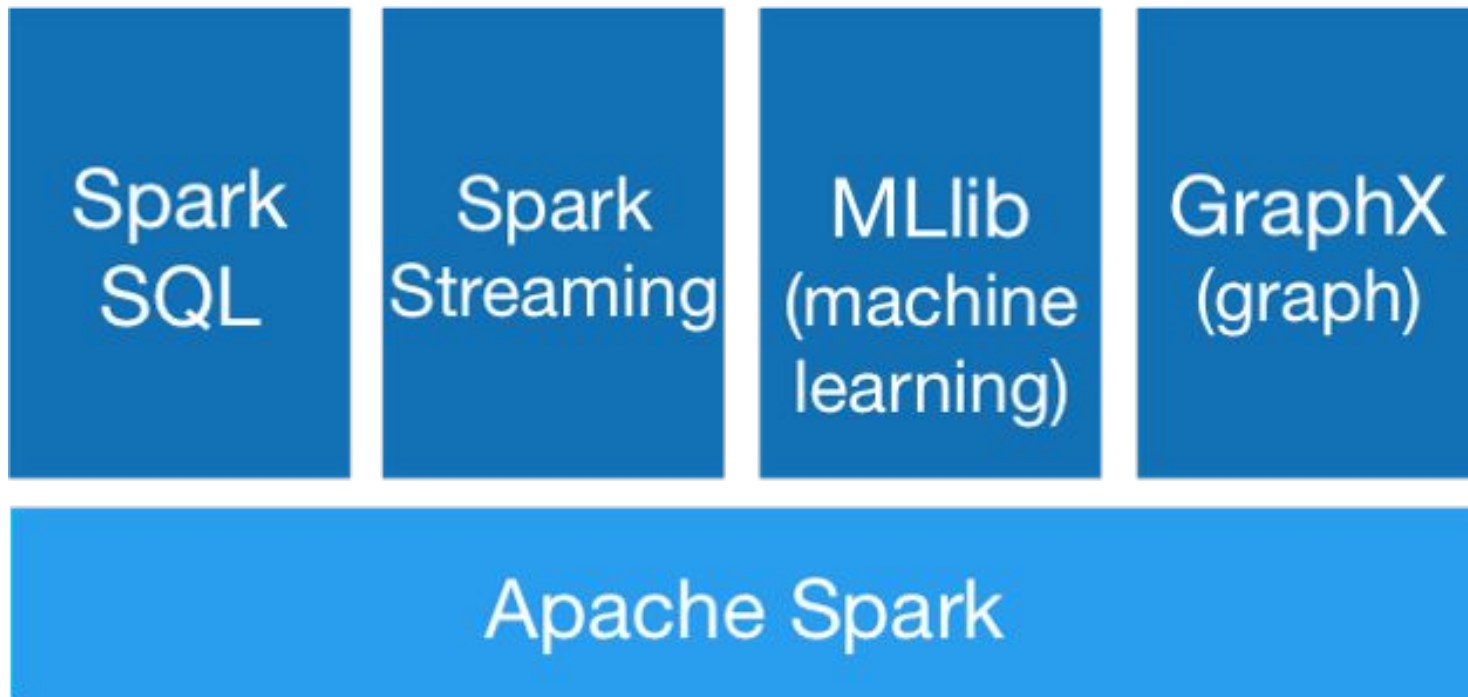
Hands on!

MINIO



MINIO

Spark MLlib



Hands on!