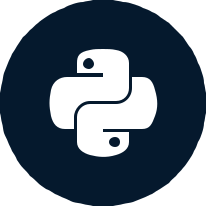
W hat is data

engineering?

INTRO DUCTIO N TO DATA ENGINEERING



Vincent Vankrunkelsven

Data Engineer @ DataC am p



W hat to expect



C hapter 1



W hat is data engineering?



C hapter 2



Tools data engineers use



* hapter 3 Extract



Transform



Load



C hapter 4

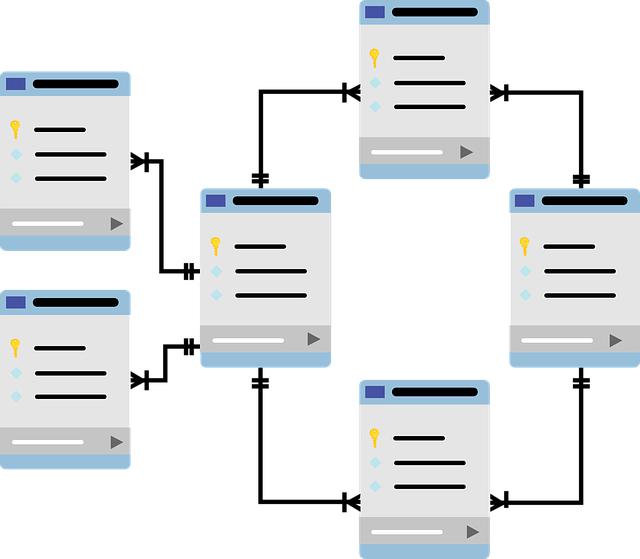


Data engineering at DataC am p



IN TRO D UC TIO N TO DATA EN G IN EERIN G

In com es the data engineer



Data is sca ered



N ot optim ized for analyses



Legacy code is causing corrupt data

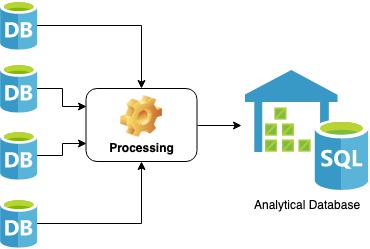


D ata engineer to the rescue!



IN TRO D UC TIO N TO DATA EN G IN EERIN G

D ata engineers: m aking your life easier



G ather data from di erent sources



O ptim ized database for analyses



Rem oved corrupt data

D ata scientist's life got w ay easier!



IN TRO D UC TIO N TO DATA EN G IN EERIN G

D efinition of the job



An engineer that develops, constructs, tests, and m aintains architectures such as databases and large-scale processing system s



Processing large am ounts of data



Use of clusters of m achines



IN TRO D UC TIO N TO DATA EN G IN EERIN G

D ata Engineer vs D ata Scientist

D ata Engineer D ata Scientist



Develop scalable data architecture M ining data for pa erns



Stream line data acquisition Statisticalm odeling



Set up processes to bring together data Predictive m odels using m achine learning



C lean corrupt data M onitor business processes



W ellversed in cloud technology C lean outliers in data



IN TRO D UC TIO N TO DATA EN G IN EERIN G

Let's practice!

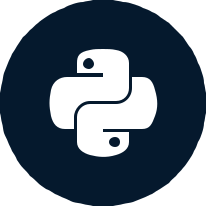
INTRO DUCTIO N TO DATA ENGINEERING



Tools of the data

engineer

INTRO DUCTIO N TO DATA ENGINEERING



Vincent Vankrunkelsven

Data Engineer @ DataC am p



D atabases



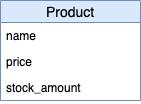
Hold large am ounts of data



Support application



O ther databases are used for analyses



IN TRO D UC TIO N TO DATA EN G IN EERIN G

Processing



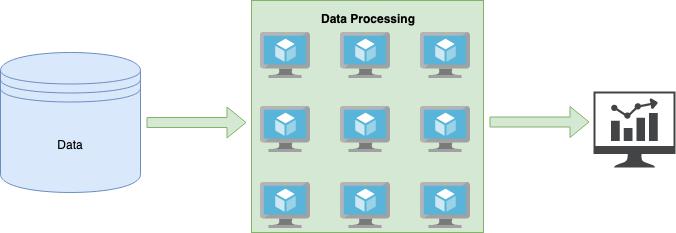
C lean data



Aggregate data

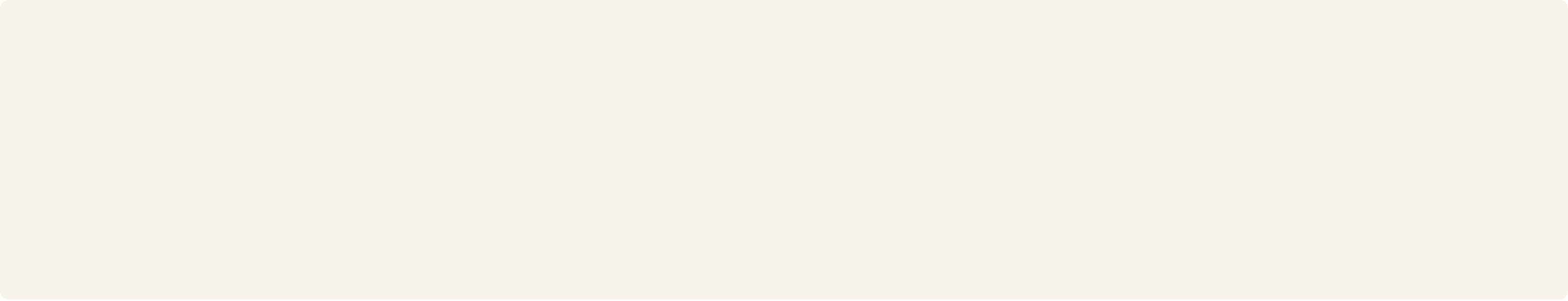


Join data



IN TRO D UC TIO N TO DATA EN G IN EERIN G

Processing: an exam ple



df = spark.read.parquet("users.parquet")

outliers = df.filter(df["age"] > 100)

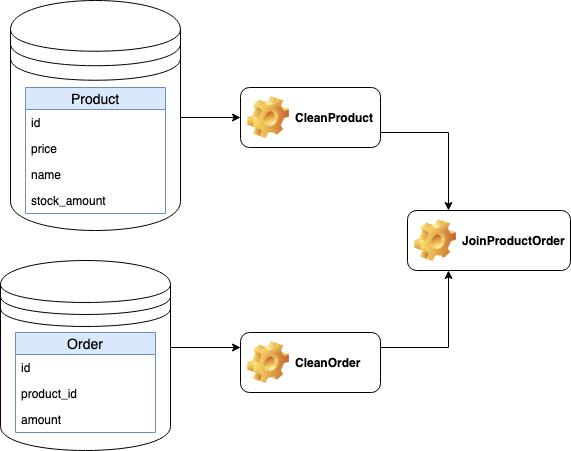
print(outliers.count())

D ata engineer understands the abstractions.



IN TRO D UC TIO N TO DATA EN G IN EERIN G

Scheduling



Plan jobs w ith speci c intervals



Resolve dependency requirem ents of jobs

JoinProductOrder needs to run a er

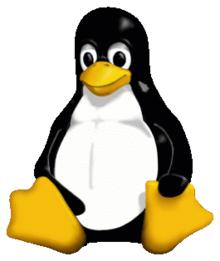
CleanProduct and CleanOrder



IN TRO D UC TIO N TO DATA EN G IN EERIN G

Existing tools

D atabases Scheduling

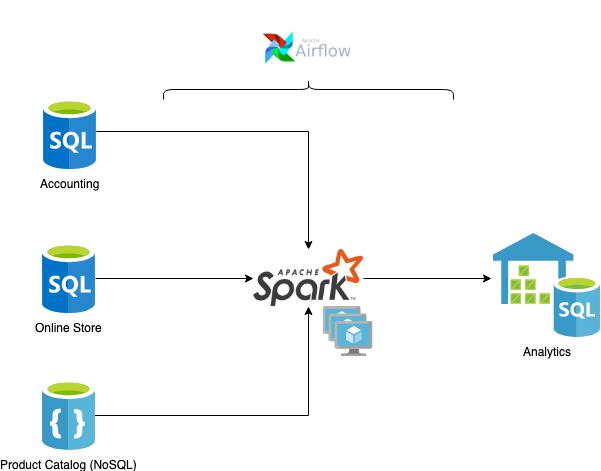


Processing



IN TRO D UC TIO N TO DATA EN G IN EERIN G

A data pipeline



IN TRO D UC TIO N TO DATA EN G IN EERIN G

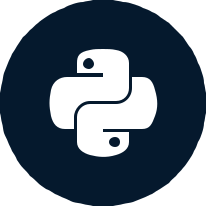
Let's practice!

INTRO DUCTIO N TO DATA ENGINEERING



C loud providers

INTRO DUCTIO N TO DATA ENGINEERING



Vincent Vankrunkelsven

Data Engineer @ DataC am p



D ata processing in the cloud



C lusters of m achines required

Problem : self-host data-center



C over electricaland m aintenance costs Peaks vs. quiet m om ents: hard to optim ize



Solution: use the cloud



IN TRO D UC TIO N TO DATA EN G IN EERIN G

D ata storage in the cloud



Reliability is required

Problem : self-host data-center



D isaster w illstrike



N eed di erent geographicallocations

Solution: use the cloud



IN TRO D UC TIO N TO DATA EN G IN EERIN G

The big three: AW S, A zure and G oogle



|  |  |  |  |
| --- | --- | --- | --- |
| 32% | m arket share in 20 18 | Storage |  |
|  |  | C om putation |  |
| 17% | m arket share in 20 18 | Databases. |  |
|  |  |  |



10 % m arket share in 20 18



IN TRO D UC TIO N TO DATA EN G IN EERIN G

Storage

Upload les, e.g. storing product im ages

Services



AWSS3



Azure Blob Storage



G oogle C loud Storage



IN TRO D UC TIO N TO DATA EN G IN EERIN G

C om putation

Perform calculations, e.g. hosting a w eb server

Services



AW S EC2



Azure VirtualM achines



G oogle C om pute Engine



IN TRO D UC TIO N TO DATA EN G IN EERIN G

D atabases

Hold structured inform ation

Services



AW S RDS



Azure SQ L Database



G oogle C loud SQ L



IN TRO D UC TIO N TO DATA EN G IN EERIN G

Let's practice!

INTRO DUCTIO N TO DATA ENGINEERING

