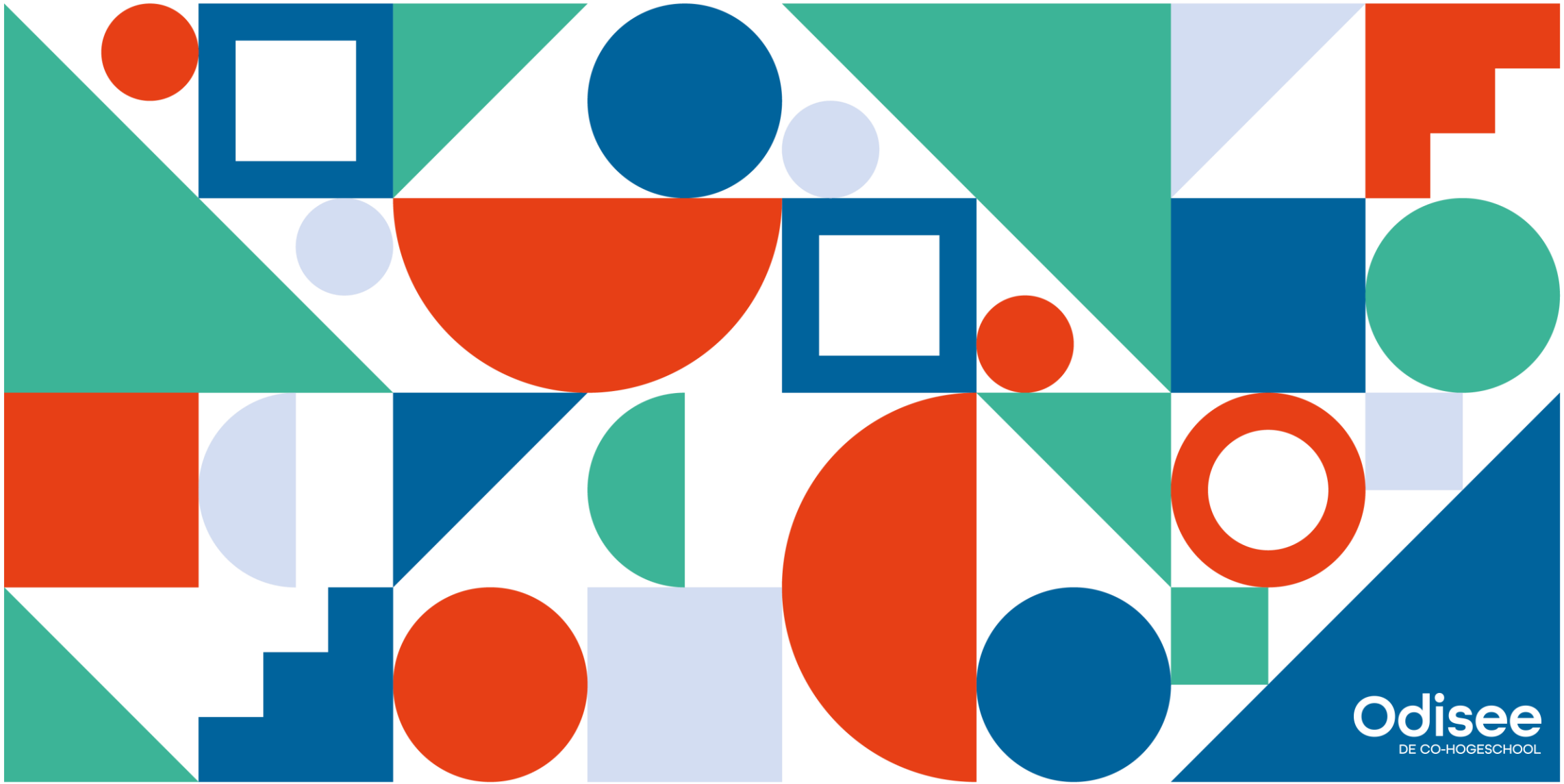
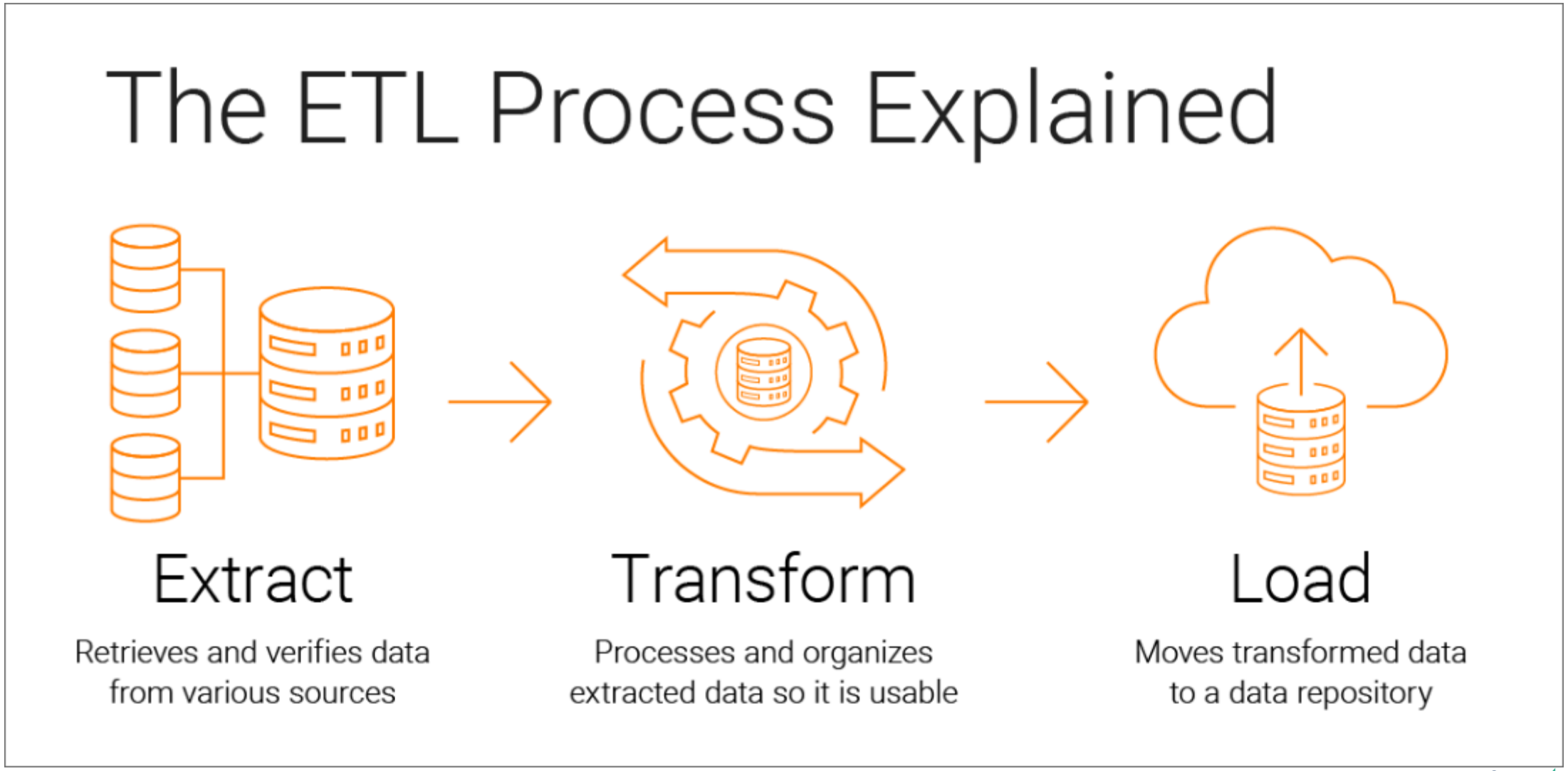

Odisee
DE CO-HOGESCHOOL

Big Data - ETL

Jens Baetens

Extract Transfer Load



ETL Life Cycle

- Extract
  - Validate hoort hier ook bij
- Transform
- Load / Stage
- Audit
- Publish

Voorbeeldtoepassingen

- Bundelen en groeperen van verschillende databronnen op 1 plaats
- Verplaatsen van data van 1 opslagplaats naar een andere
  - Bijvoorbeeld van een lokale database naar de cloud
- Omdat de drie stappen elk tijd vragen worden deze vaak in parallel uitgevoerd

Extract

Extract

- Haal data uit verschillende databronnen
  - Corporate databases
  - Online API's (twitter, facebook, …)
  - Website scraping
- Verschillende bronsystemen
- Verschillende formaten
- Omvat ook een data validatiestap
  - Data dat hierbij faalt wordt best gerapporteerd voor verdere analyse

Functionaliteiten binnen Spark

- Hiervoor gebruiken we dus vooral de inleesfuncties (.read.csv(), …) in de SQL module
- Extract kan ofwel
  - Eenmalig zijn: zelf gestart via code, programma dat data verwerkt en dan stopt
  - Continue zijn: programma blijft actief om continu binnenkomende data te transformeren en te bewaren
    - Streaming

Transform

Transform

- Transformeer de geextraheerde data naar het gewenste dataformaat
- Data Cleaning valt hier ook onder
- Veel gebruikte transformaties
  - Select
  - Vertalen gecodeerde waarden (0/1 vs true/false , male/female vs M/F)
  - Encoderen van kolommen
  - Samenvoegen/splitsen
  - Pivoteren / transponeren
  - Berekeningen

Functionaliteiten binnen Spark

- De Extract stap levert een DataFrame aan
  - Alle functionaliteiten voor een DataFrame kunnen gebruikt worden

- Speciale aandacht vereist in geval van streaming omdat data in stukjes binnenkomt
  - Werken met windows

Load

Load

- Bewaren van de data op een finale opslagplaats
  - In een bestand, database, datawarehouse of datalake
- Data is vaak maar relevant binnen een bepaald tijdsvenster
  - Verwijder of archiveer te oude data

Uitdagingen

- Het opvangen van de verscheidene datastructuren/formaten is complex
  - Een robust ETL-process is essentieel om bruikbare data over te houden
- Volume van data kan zeer groot worden dus is schaalbaarheid belangrijk
  - Zowel op vlak van opslag als op rekencapaciteit

ELT – Extract Load Transform

- Variant waarbij ruwe, onverwerkte data bewaard wordt
- Transformatie / verwerking pas wanneer het nodig is
- Kan vooral handig zijn bij Cloud-toepassingen die heel schaalbaar zijn
  - Afweging kost opslag - verwerkingstijd

- ETL eerder bij data integratie
- ELT popular bij data warehouse/lake toepassingen

https://youtu.be/6kEGUCrBEU0?list=RDCMUCKWaEZ-_VweaEx1j62do_vQ