PYSPARK

Data Engineering Bootcamp

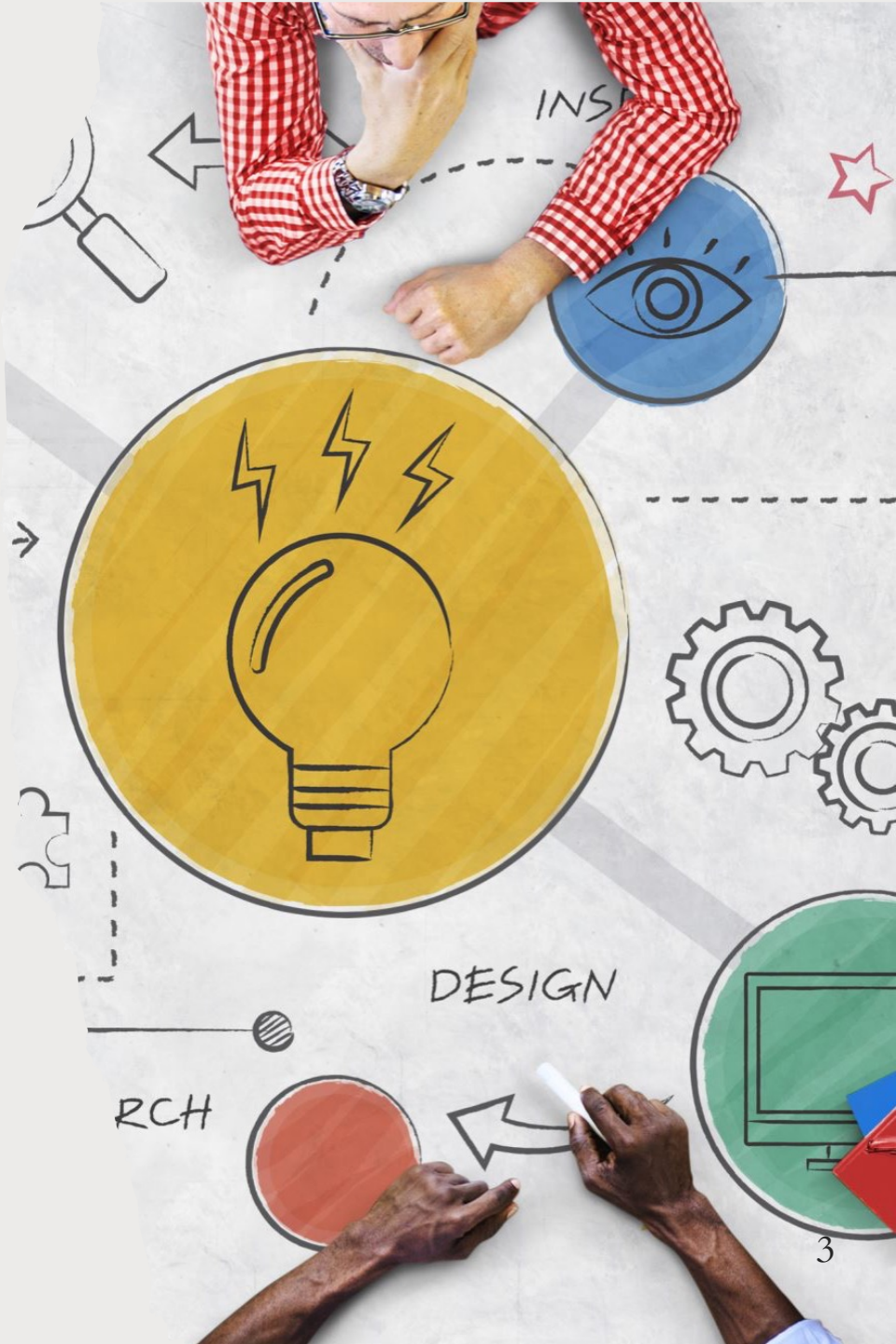
Data management and archiving in the research environment

By Niklas Büchner, Christian Singer, Ahmad Al-Taie, Mike Sickmüller

CONTENT

- Introduction & Presentation ETL Process
- Actions & Basic Transformations
- Advanced Transformations
- DataVault

INTRODUCTION &PRESENTATION ETL PROCESS

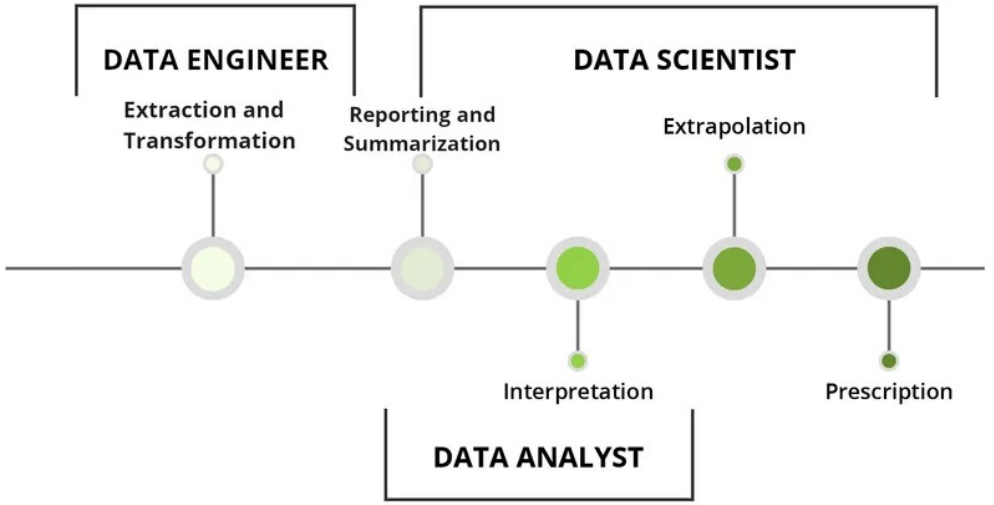


GOAL OF THE LEARNING SECTIONS

- Create a basic understanding for Data Engineering
- Understand how an ETL workflow is structured
- The Role of Spark in such a contex

WHAT IS DATA ENGINEERING ?

- Developing and building systems for collecting, storing and analyzing data
- Provide data for evaluation and optimization performance of enterprises
- Data engineers manage data resources
- Data analysts use data to gain insights



WHAT IS ETL ?

Extraction

- Raw data is copied or exported from a variety of data sources
- These can be structured or unstructured

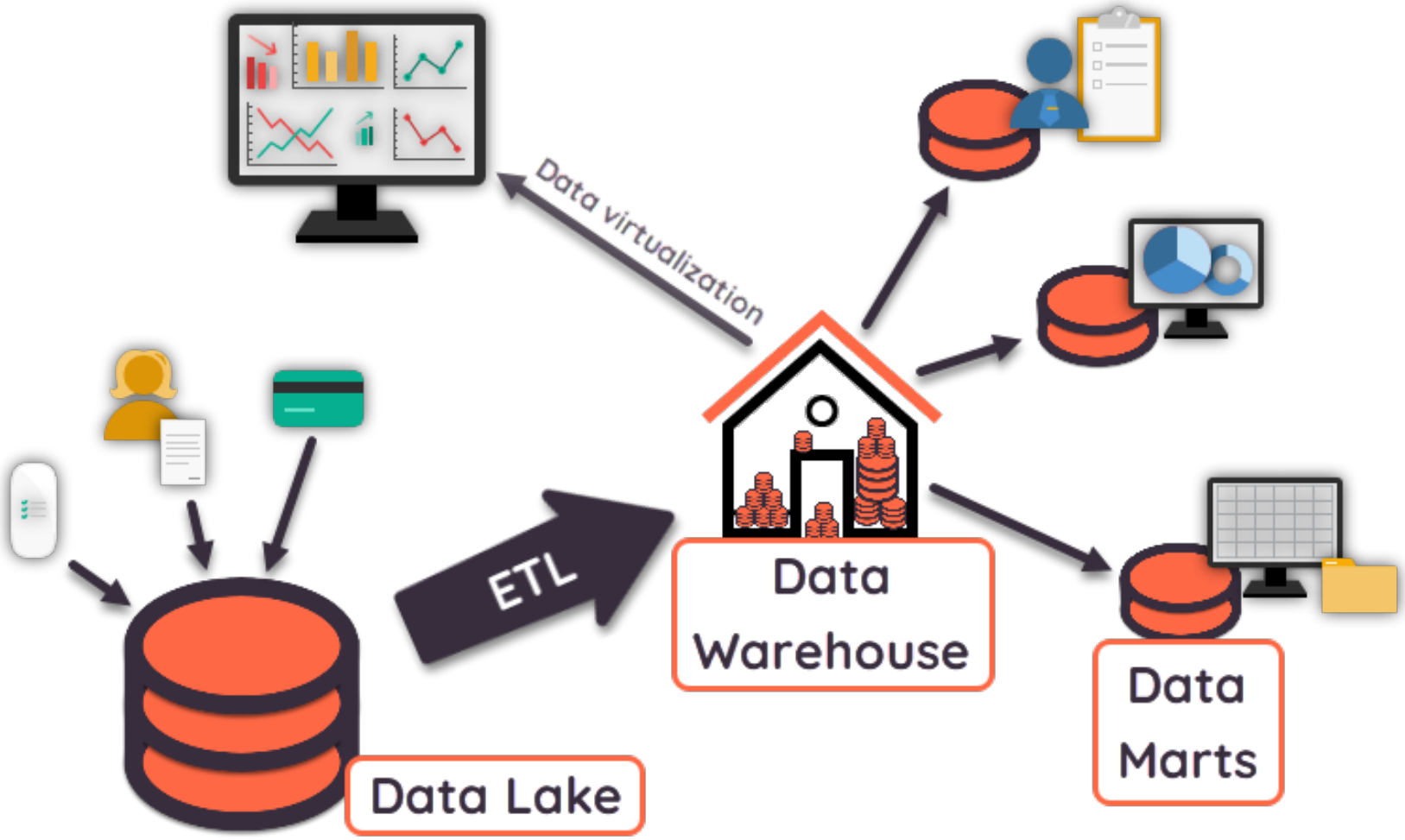
Transformation

- Collected raw data is consolidated for the intended use case
- During transformation, data is deduplicated, translated or summarized
- Adapt data to the Data Warehouse schema

Loading

- Load transformed data into the Target-Data-Warehouse

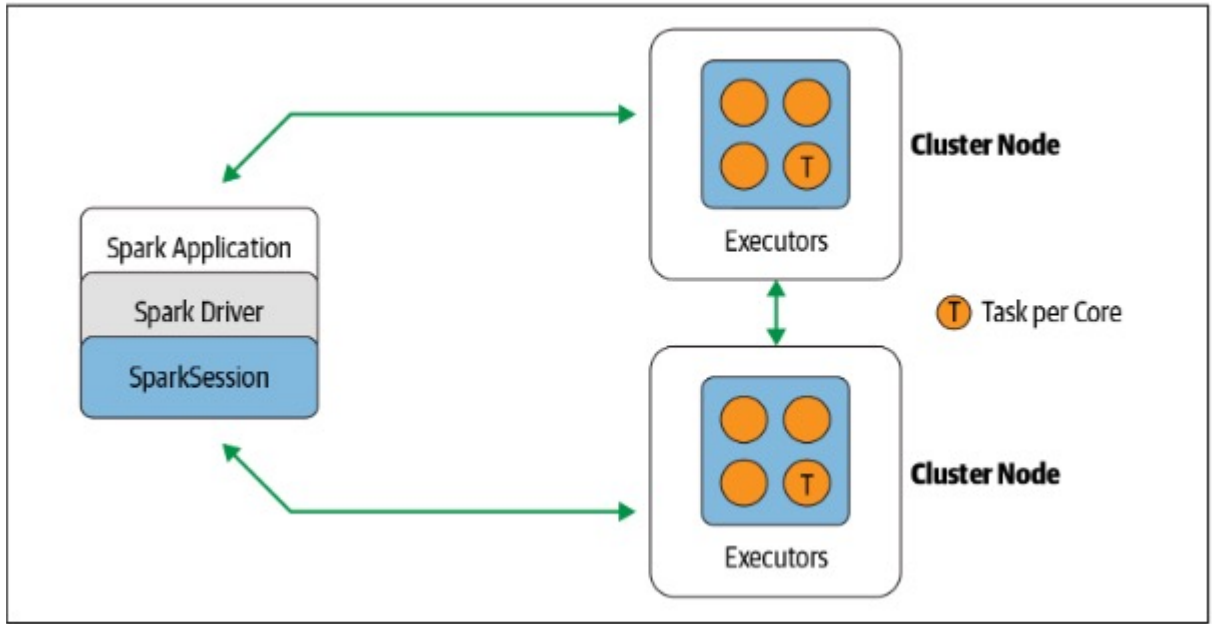
WHAT IS ETL ?


Data virtualization
ETL
Data Warehouse
Data Lake
Data Marts

WHAT IS SPARK ?

Apache Spark/PySpark

- Multilingual engine for data engineering execution
- Outsourcing of the data development workflow to a number of servers
- Processing Big Data through parallelization
- Spark is written in Scala
- Python functions are available via Python-based wrapper PySpark



SPARKSQL

- Basic Data Structure Resilient Distributed Dataset (RDD)
- Tutorial focus on SparkSQL model, DataFrame
- DataFrame has great advantages over RDD
  - powerful optimization engine
  - Data Science module working with DataFrames