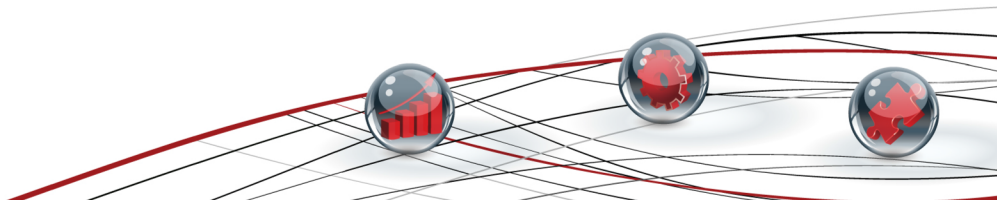
Full Orbit

Oefeningen Introductietraining Spark

- Hoofdstuk 2 RDD's
  - Oefening 1
    Maak een RDD van deze steden in Nederland: Amsterdam, Rotterdam, Den Haag, Utrecht, Groningen, Maastricht, Leeuwarden, Nijmegen.
    Gebruik de actie count() om te testen of het aanmaken van de RDD gelukt is (lazy execution).
  - Oefening 2
    Maak een RDD van van het bestand sales.csv in de map bestanden.
    Gebruik de actie count() om te testen of het aanmaken van de RDD gelukt is (lazy execution).
- Hoofstuk 3 Acties
  - Oefening 1
    Maak een RDD van van het bestand sales.csv in de map bestanden.
    Laat het eerste element van de RDD zien
  - Oefening 2
    Maak een RDD van van het bestand sales.csv in de map bestanden.
    Onderzoek of er dubbele rijen inzitten (hiervoor kun je transformatie distinct() gebruiken)
  - Oefening 3
    Maak een RDD van deze steden in Nederland: Amsterdam, Rotterdam, Den Haag, Utrecht, Groningen, Maastricht, Leeuwarden, Nijmegen.
    Sla de RDD op als tekst bestand
- Hoofdstuk 4 Transformaties
  - Oefening 1
    Maak een RDD van van het bestand sales.csv in de map bestanden.
    Maak alle tekst uppercase
  - Oefening 2
    Maak een RDD van van het bestand sales.csv in de map bestanden.
    Maak een lijst van alle genres
  - Oefening 3
    Maak een RDD van van het bestand sales.csv in de map bestanden.
    Tel het aantal woorden
  - Oefening 4
    Maak een RDD van van het bestand sales.csv in de map bestanden.
    Maak een RDD zonder header regel
  - Oefening 5
    Maak een RDD van van het bestand sales.csv in de map bestanden.
    Maak een RDD met alleen de gegevens van 2017
  - Oefening 6
    Maak een RDD van van het bestand sales.csv in de map bestanden.
    Toon 2010 per genre het aantal records



- Oefening 7
  Lees de bestanden landen1.csv en landen2.csv in en maak één RDD voor beide bestanden
- Hoofdstuk 5 Groeperen en sorteren
  - Oefening 1
    Maak een pairRdd van onderstaande sales data.
    Jaar Sales (mln)
    2001 11.6
    2002 12.3
    2003 12.4
    2004 13.0
    2005 13.4
    Tel het aantal elementen
  - Oefening 2
    Maak een RDD van van het bestand sales.csv in de map bestanden
    Maak een pairRDD met de sales in EU per jaar en toon de eerste 25 rijen
  - Oefening 3
    Maak een RDD van van het bestand sales.csv in de map bestanden
    Maak een pairRDD met de sales in EU, NA, JP en overig per jaar en toon de eerste 25 rijen
  - Oefening 4
    Maak een RDD van van het bestand sales.csv in de map bestanden
    Maak een pairRDD met de sales in EU per jaar en groepeer per jaar
    Hoeveel elememten hou je over
    Toon het eerste element, wat valt op?
  - Oefening 5
    Maak een RDD van van het bestand sales.csv in de map bestanden
    Maak een gesorteerde lijst van de jaartallen in het bestand
  - Oefening 6
    Maak een RDD van van het bestand sales.csv in de map bestanden
    Maak een lijst van de jaartallen in het bestand met het aantal sales records per jaar
  - Oefening 7
    Maak een RDD van van het bestand sales.csv in de map bestanden
    Maak een lijst van de jaartallen in het bestand met de totaal van sales in EU
  - Oefening 8
    Maak een RDD van van het bestand sales.csv in de map bestanden
    Maak een lijst de 3 jaartallen met de hoogste omzet
- Hoofdstuk 6 Joins
  - Oefening 1
    Lees de bestanden steden.csv en landen1.csv in en koppel de bestanden aan elkaar