

Data Engineer - 25

Programmering inom Data Platform

Labboration 1 (IG/G)

“Data Ingestion, manipulation & workflow”

-BONUS delar är ej krav-

Kunskap

ID	Beskrivning
KUN9	Redogöra för typiska komponenter och teknologier i en dataplattform
KUN10	Redogöra för begreppet ETL pipelines

Färdigheter

ID	Beskrivning
F10	Med hjälp av Python kunna läsa, ändra och skapa data i olika dataformat och databaser

ID	Format för kunskapskontroll	Betygsskala
KUN9; KUN10, F10	Inlämningsuppgift	IG/G

Get Started

1. **Ladda ner** products.json (*Laboration 1 - Learnpoint (sti)*)
2. **Läs in Filen med Pandas** read_from() metod (lägg in den i projektet först)
3. **Transformera** innehållet så att det är användbart (Lektion #7 & #8)
4. **Flagga** möjliga problem såsom:
 1. Saknad av 'currency'
 2. produkter med extremt höga priser (kan vara 'luxury products')
 3. Kostar 0 (var den gratis? Eller problem?)
5. **Avvisa** omöjliga värden i kolumner
6. **När du är klar** genererar du följande filer med följande kolumner:

analytics_summary.csv

- snittpris
- medianpris
- antal produkter
- antal produkter med saknat pris

price_analysis.csv (bonus)

- top 10 dyraste produkter
- top 10 produkter med mest avvikande pris

rejected_products.csv (bonus)

- Inkludera alla produkter som är omöjliga

Studerande ska kunna:

- peka ut **ingest** —> **storage** —> **transform** —> **access** (*teori*)
- känna igen teknologityper (Pyscopg3, Pandas) (*teori*) **Bonus:** pydantic
- förklara **Extract** —> **Transform** —> **Load** (*teori*)
- Få ut information från en produkt. (Average mm.. mest sold osv..)