Milepæl 2 - Big data

Ola Wethal Petersen

September 2020

1 Teoridel

Hva er en Dataframe?

En dataframe er et dataset organisert i kolonner. Et dataset kan "puttes" inn i en dataframe og vi kan så definere de forskjellige kolonnene som tilhører dataen. Basert på språket som implementerer dataframen og datasettet kan det også hende at innholdet kan inferes, altså at dataframen klarer å finne ut hvilke kolonner som har hvilket navn selv.

Hva er forskjellen på et dataset og en dataframe?

Et dataset er da dataen selv, som er umodifisert. Dette kan bety at dataen kan være vanskelig å jobbe med, uten at vi setter opp en schema for dataen. Om dette er et problem eller ikke spørs på hvordan dataen er satt opp og om den er strukturert eller ikke.

Å legge dataen inn i en frame gjør det rett og slett enklere for oss å utføre arbeid med dataen. Når vi har en frame kan vi enkelt definere kolonner og datatyper. Ved å legge det inn i en frame kan vi også kanskje utføre jobber med dataen raskere enn vi kanskje vill gjort originalt. Dette spørs selvfølgelig helt på hvor stor datamendgen faktisk er og hva slags arbeid som skal utføres.

Hva er en partisjon?

Har partisjonering noe å si hvis vi bare har en maskin?

2 Programmering