Statens vegvesen

Verktøy og muligheter i Google-skyen

Arbeidsmøte for datavitere

28.04.2021

Vi skal bruke resten av dagen på å vise dere muligheter i GCP

Statens vegvesen

Bakgrunn

- Google Cloud Platform har mange verktøy som kan være nyttige i analysearbeid ute i fagmiljøene
- Saga ønsker å tilby opplæring og legge til rette for bruk av plattformen i etaten
- Vi ønsker innspill om behov og ønsker knyttet til både verktøy, data og analyser

Arbeidsmøte om GCP

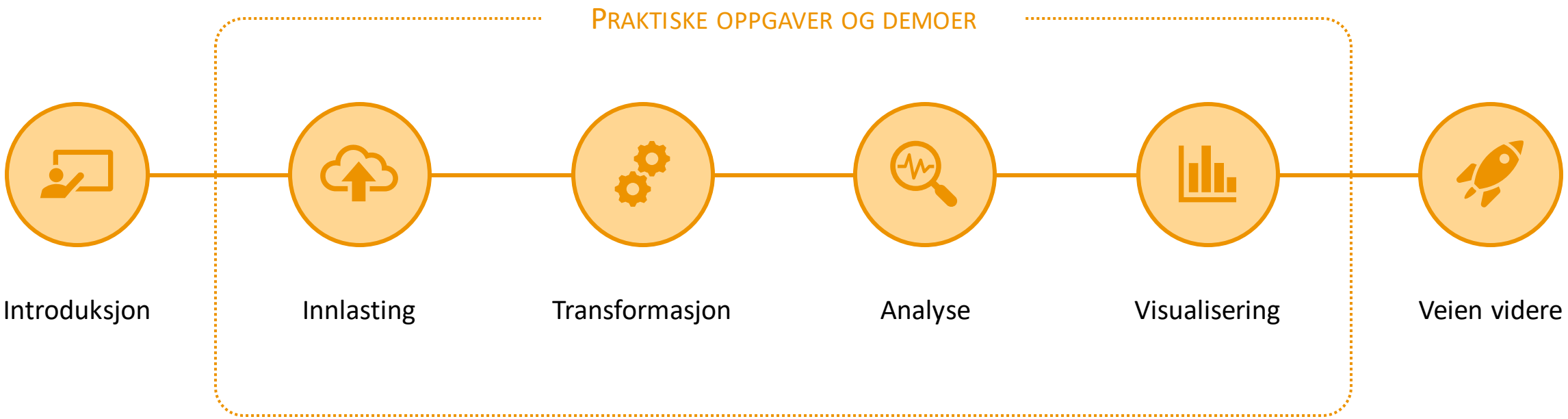
10.00 – 16.00

Mest jobbing med praktiske oppgaver

Lunsj og pauser underveis

Vi vil presentere litt, men mest av alt skal dere få teste verktøyene i praksis

Statens vegvesen



Målet med møtet er at dere skal bli godt nok kjent med verktøyene til å ta dem i bruk

Kurset er laget for deltakere med noe varierende teknisk bakgrunn

Statens vegvesen

Noen kan tenke at vi gir ting inn med teskje

Det er ikke fordi vi undervurderer dere, men fordi vi prøver å få med oss alle.

Andre kan oppleve at dette går over hodet på dem

Da er det vi som har bommet på nivået.
Ikke vær redd for å stille dumme spørsmål!

Men før vi begynner...

Statens vegvesen

... er det noen av dere som vet hva en dataplattform er?

> ”
> A data platform is a centralized system that combines scalable flexibility with distributed data storage and computational power fo acquiring and analyzing large data sets to provide users with reliable and accurate data.

> ”
> En dataplattform er et system for å orkestrere data som kan brukes til analyse eller tredjepartsapplikasjoner.

https://www.tietoevry.com/en/blog/2020/11/the-fundamentals-of-cloud-data-platforms/

https://www2.deloitte.com/content/dam/Deloitte/lu/Documents/technology/lu-next-generation-data-platform.pdf

Hvorfor trenger man så en dataplattform?

Statens vegvesen

- En av de største trendene i organisasjoner er å bli datadrevet. Å bli datadrevet betyr å kunne analysere data og basere beslutninger og verdiskaping på analyser.

- To eksempler:
  - Uber har laget en dataplattform som samler inn og sammenstiller data fra kjøreturer i sanntid. Ved å gjøre sanntidsanalyse på etterspørsel og tilbud, har de laget en modell for dynamisk prising, som er en nøkkelkomponent i deres forretningsmodell.
  - Spotify har laget en dataplattform som kontinuerlig henter inn data fra spilleren – hva slags musikk du spiller av, hvor mye du hører av musikken++ Ved å gjøre analyser av dataene og sammenstille med andre, kan de foreslå nye spillelister.

Hvorfor trenger man så en dataplattform?

Statens vegvesen

- For å kunne analysere dataene må data og analyseverktøy må være tilgjengelige. En typisk situasjon i organisasjoner (som Statens vegvesen) som forhindrer dette, er at dataene ligger i forskjellige systemer, og de er ikke tilgjengelig for analyse. Det kan også være at administrasjonen av dataene skjer manuelt, og det er sakte og kostbart å legge til nye datasett fra til og med offentlig tilgjengelige datakilder.

- Når data ligger spredd rundt vil det ofte resultere i en situasjon der dataanalytikere eller dataforskere mister verdifull tid på å lete etter, rense og forstå dataene. Det sies ofte at dataanalytikere bruker 80% tid på såkalt data wrangling og 20% tid på analyse.

- Og i verste fall, hvis en felles plattform mangler, blir maskinlæringsmodellen som endelig er bygget bare distribuert på dataviterens bærbare datamaskin uten noen mulighet til å utnytte den i et bredere omfang, og uten at noen andre oppdaterer den.

- Her er hvor datadataplattformen spiller inn. Det er et system for å organisere dataene for analyser eller tredjepartsapplikasjoner. Så i stedet for å bruke 20% av tiden på analyse, så vil kapabilitetene til en dataplattform potensielt muliggjøre at dataanalytikere bruker mer tid på den virkelige verdifulle delen.

Hva slags kapabiliteter bringer dataplattformen til torgs?

Statens vegvesen

Fundamentals

- Integrasjonslag som sørger for integrasjoner til kildesystemer.
- Datalagring – i forskjellige formater – filbasert og strukturert.
- Datamodellering – logisk datamodell og modelleringsverktøy.
- Publikasjonslag – gjør det enkelt å tilgjengeliggjøre data og analyser for andre.
- Data pipeline komponenter – brukes til å definer dataflyter fra kildesystemer og ut til publikasjon.
- Overvåkningskapabiliteter – for å overvåke driften av plattformen, at komponentene kjører feilfrie, men også sjekker at datastrømmene og kvaliteten på dataene er som forventet.
- Storskala prosessering av data, både til f.eks rensing av data og analyse av data.

Verdidrivere

- Miljø for å bygge maskinlæringsmodeller, trene de, validere de og produksjonssette de.
- Data science miljø for å kunne gjøre analyser effektivt.
- En datakatalog, der brukere av dataplattformen kan bla gjennom de tilgjengelige dataproduktene, forstår hvordan de blir opprettet, og til og med gir tilbakemelding til andre brukere om dataproduktene

Data katalog

Statens vegvesen

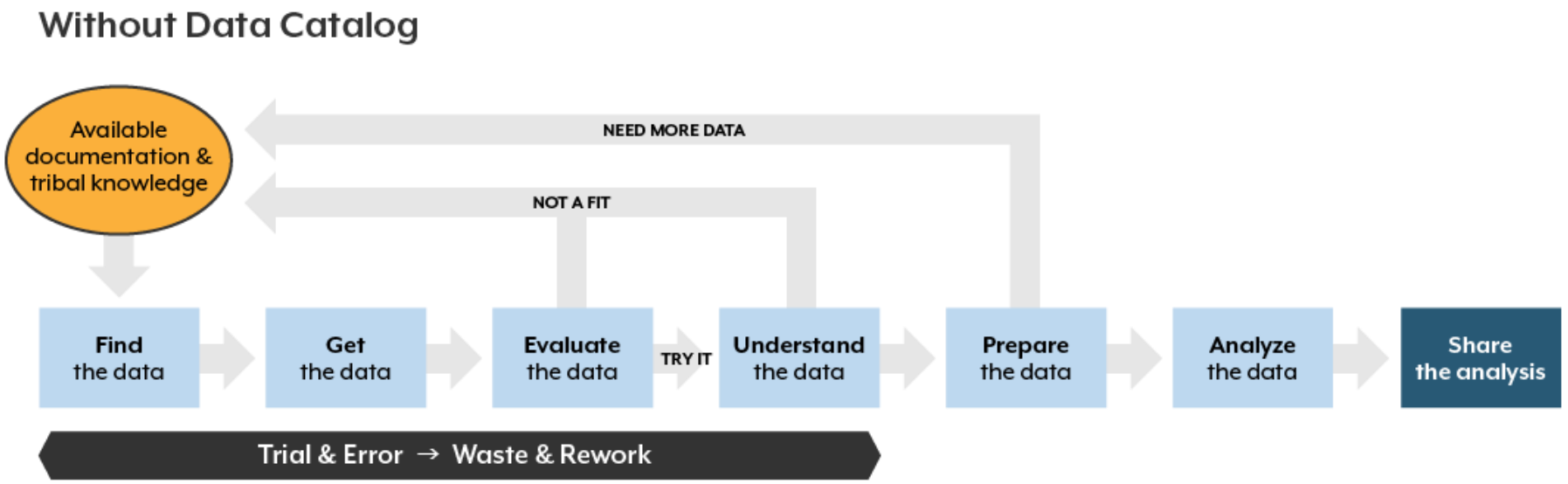
- Forretnings- og dataanalytikere jobber ofte i blinde, uten innsikt i hvilke datasett som finnes, innholdet i disse datasettene, og kvaliteten og nytten av de. De bruker for mye tid på å finne og forstå data, og ofte gjenskape datasett som allerede eksisterer. De jobber ofte med utilstrekkelige datasett, noe som resulterer i utilstrekkelig og feil analyse.

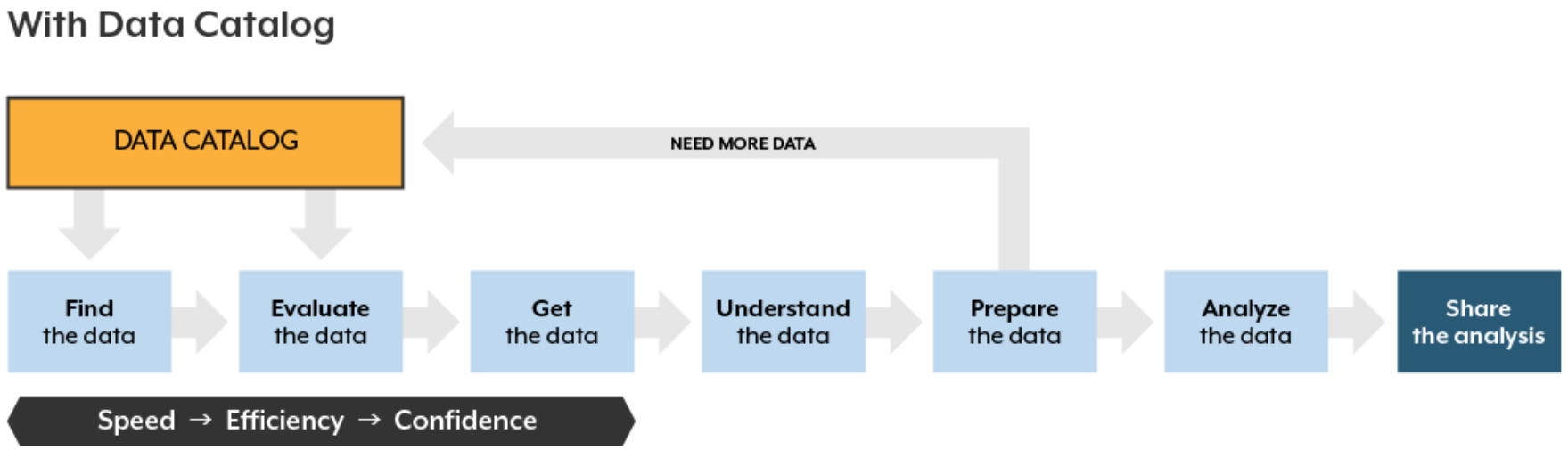
- Uten katalog ser analytikere etter data ved å sortere gjennom dokumentasjon, snakke med kolleger, stole på stammekunnskap eller bare jobbe med kjente datasett fordi de vet om dem. Prosessen er full av prøving og feiling, «waste» og duplikat arbeid og gjentatte søk etter data som ofte fører til å jobbe med "nær nok" data etter hvert som man nærmer seg deadline.

- Med en datakatalog er analytikeren i stand til å søke og finne data raskt, se alle tilgjengelige datasett, evaluere og ta informerte valg for hvilke data som skal brukes, og utføre dataklargjøring og analyse mer effektivt og stole på at man benytter riktige data.

Datakatalog

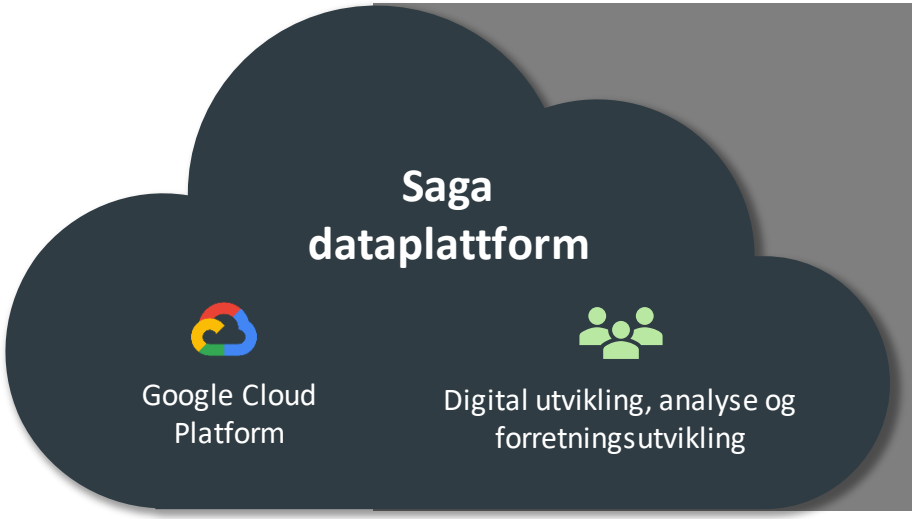
Statens vegvesen





Saga er en dataplattform i skyen som omdanner data til innsikt og handling

Statens vegvesen



Sentrale kapabiliteter

Innhenting av data

Ta inn data fra interne og eksterne kilder, i sanntid og historisk

Sammenstilling og analyse

Kvalitetssikre, sammenstille og analyser Data fra ulike kilder kvalitetssikres, sammenstilles og analyseres

Styring og tilgjengeliggjøring

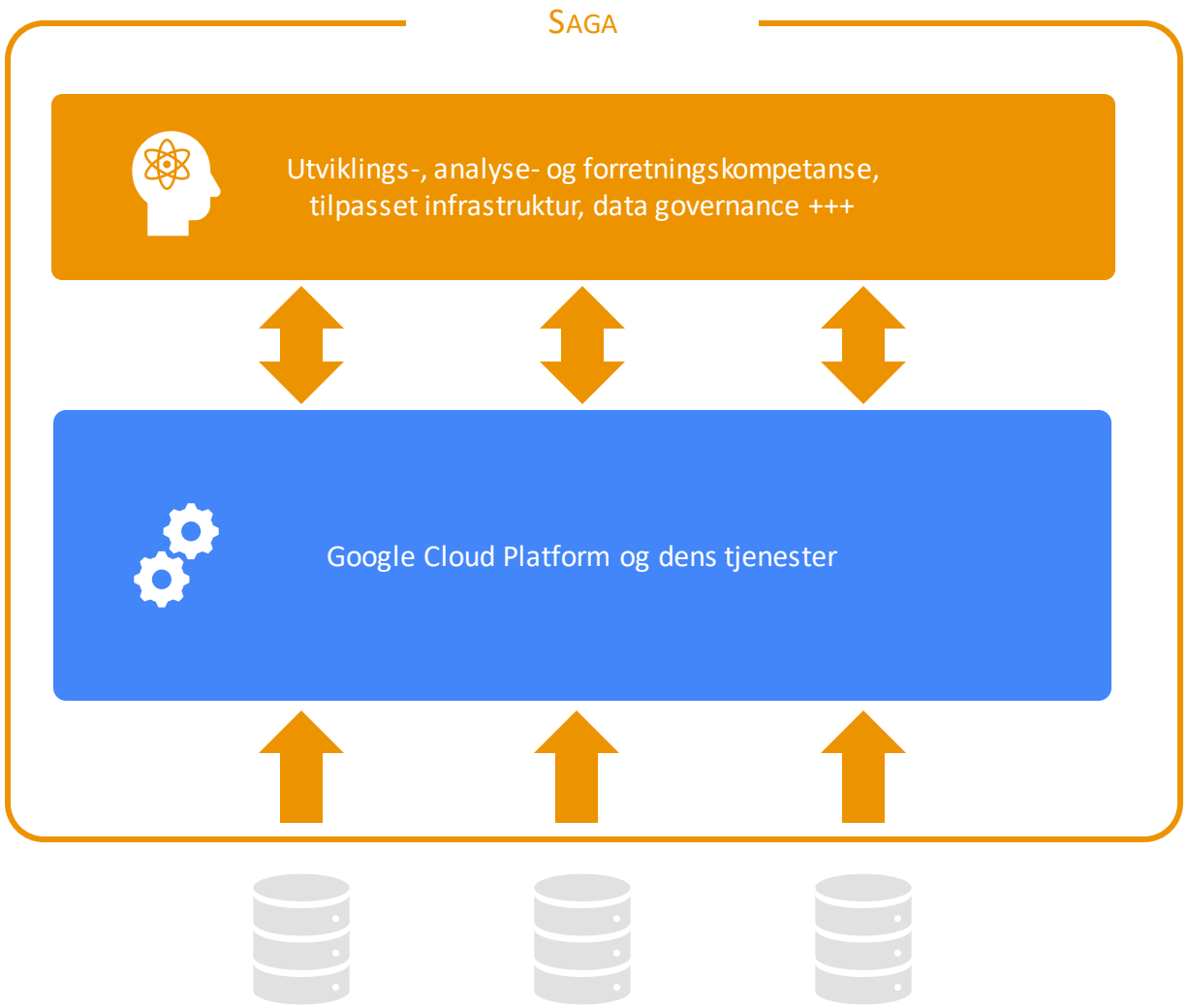
Data, innsikt og resultater styres gjennom sitt livsløp, og tilgjengeliggjøres brukere i egne verktøy eller i plattformen

Statens vegvesen

Demo

Gjennom Saga vil etaten få et tilpasset GCP-miljø med viktige data klare til bruk

Statens vegvesen



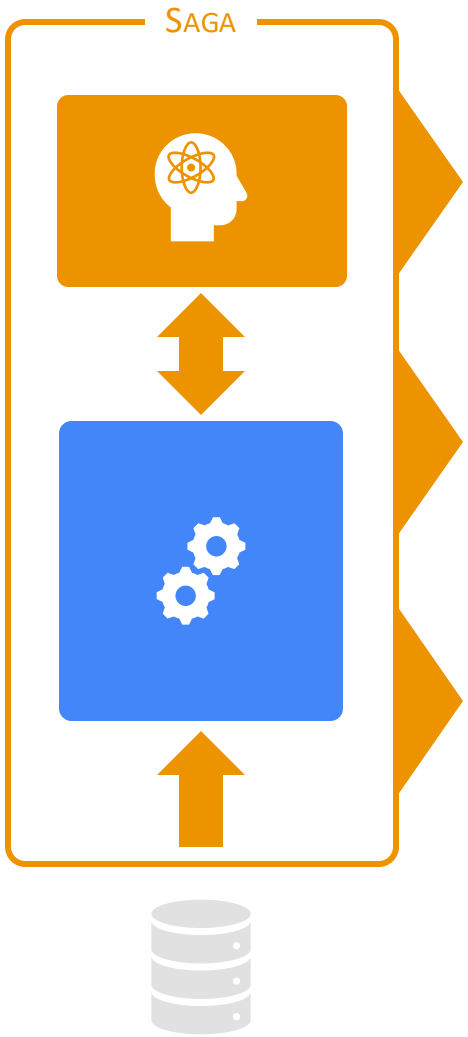
Saga er både en teknisk plattform og et kompetansemiljø med kapasitet til å utvikle plattformen og gjøre analyser på bestilling.

GCP er kjernen i plattformen, men den har også egenutviklede komponenter.

Plattformen inneholder flere datasett som er klare til bruk, og på sikt er ambisjonen å være SVVs primære lagringssted for stordata.

Sagas kapasitet avhenger av *hvordan* vi bistår resten av etaten

Statens vegvesen



Analyser
Saga kan gjennomføre analyser og utvikle beslutningsstøtteverktøy på bestilling fra fagmiljøene

Teknisk og analytisk støtte
Saga kan bidra med spisskompetanse og veiledning i fagmiljøenes arbeid med data og analyse

Selvbetjeningsløsninger og opplæring
Saga kan tilby datakatalog, verktøy, anbefalte arbeidsmåter og opplæring som gjør fagmiljøene bedre i stand til å *drive sitt eget analysearbeid*

Det er mange fordeler med å bruke GCP til databehandling og analyse

Statens vegvesen

Kraftige verktøy

GCP tilbyr mange ulike verktøy, uendelig skalering og enkelt oppsett av nye ressurser

Tilrettelagt for samarbeid

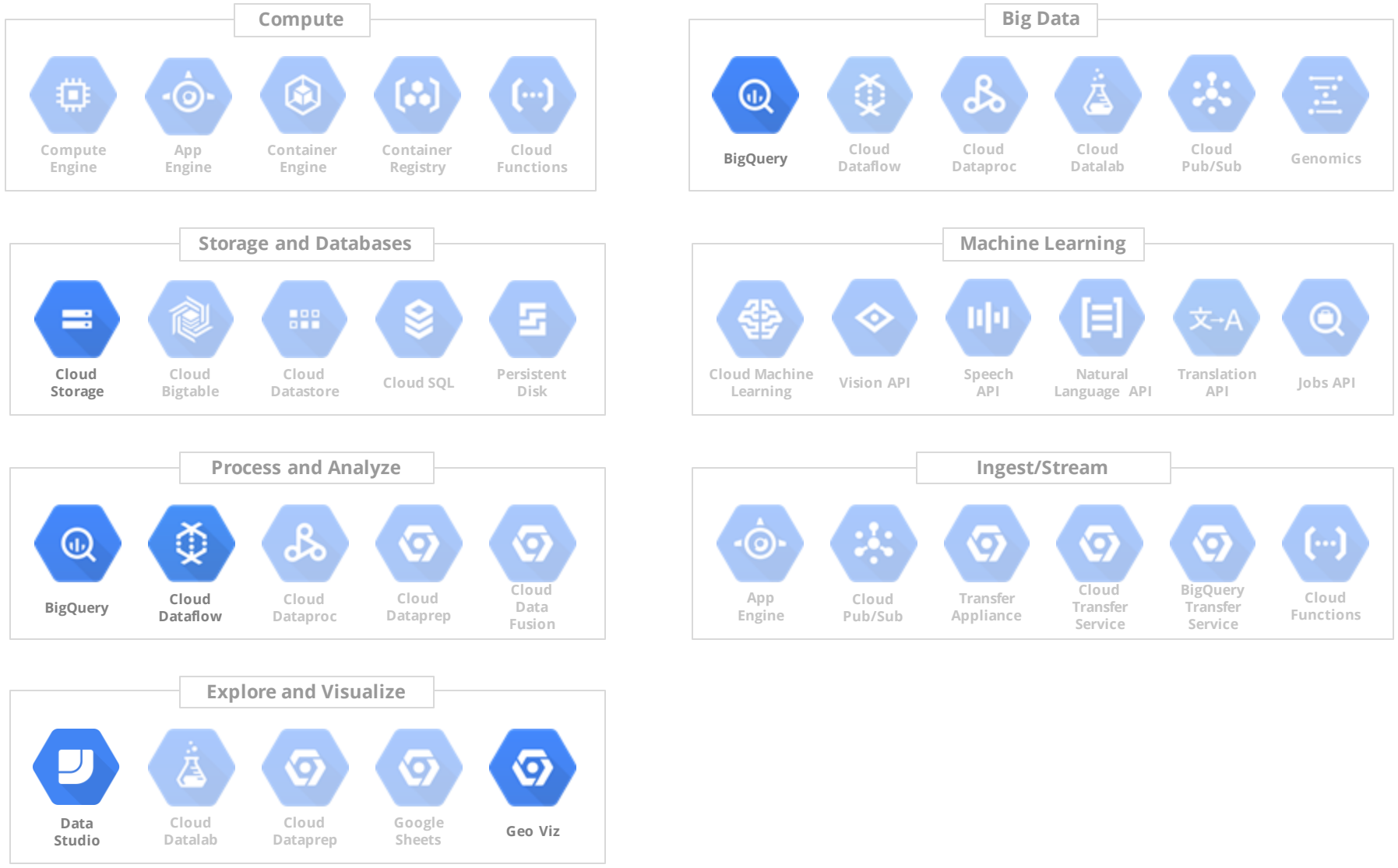
Skylagring og felles verktøy legger til rette for gjenbruk, samarbeid og erfaringsdeling

Oppdatert og vedlikeholdt

Google håndterer installasjoner, oppdateringer og kompabilitet, så vi kan bruke tiden vår på analyse

Google Cloud Platform har et bredt utvalg verktøy og tjenester

Statens vegvesen



De to mest brukte datalagrene

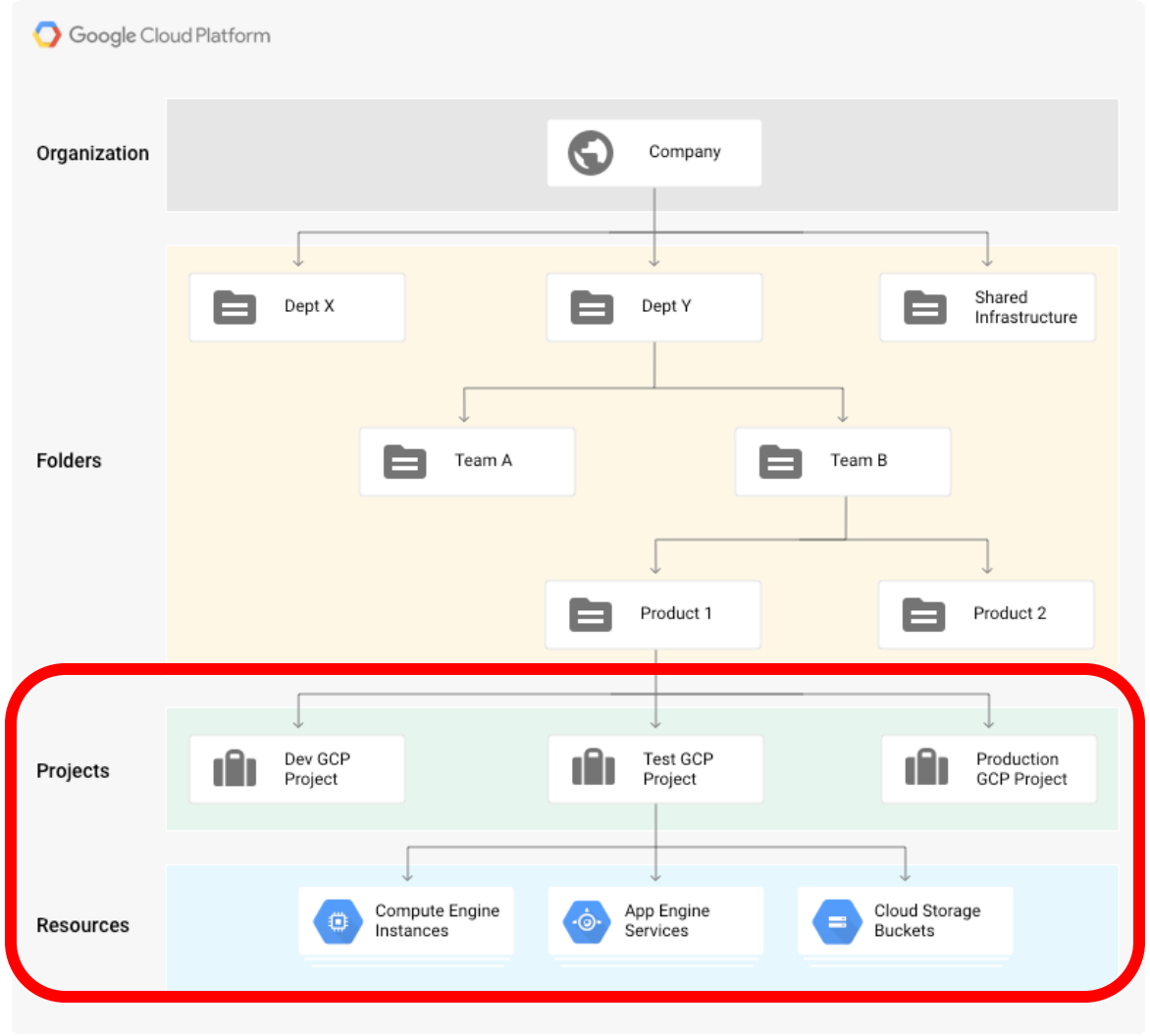
Statens vegvesen

- Google Cloud Storage (GCS)
  - Brukes til lagring av "blobs" - altså filer og lignende
  - Relativt billig å lagre store mengder data (mellom 4 og 26 USD per TB/mnd)
  - Hva betaler man for?
    - Antall TB som lagres
    - Å hente data ut fra GCS til egen maskin
- BigQuery
  - Serverless, superskalerbart datavarehus
  - Kan lagre og gjøre SQL-spørringer mot flere **petabyte** med data
  - Hva betaler man for?
    - Antall TB som lagres (cirka samme pris som GCS)
    - Antall TB som prosesseres i spørringer. Første TB er gratis, så 5 USD per TB/mnd.

Prosjekter og ressurser

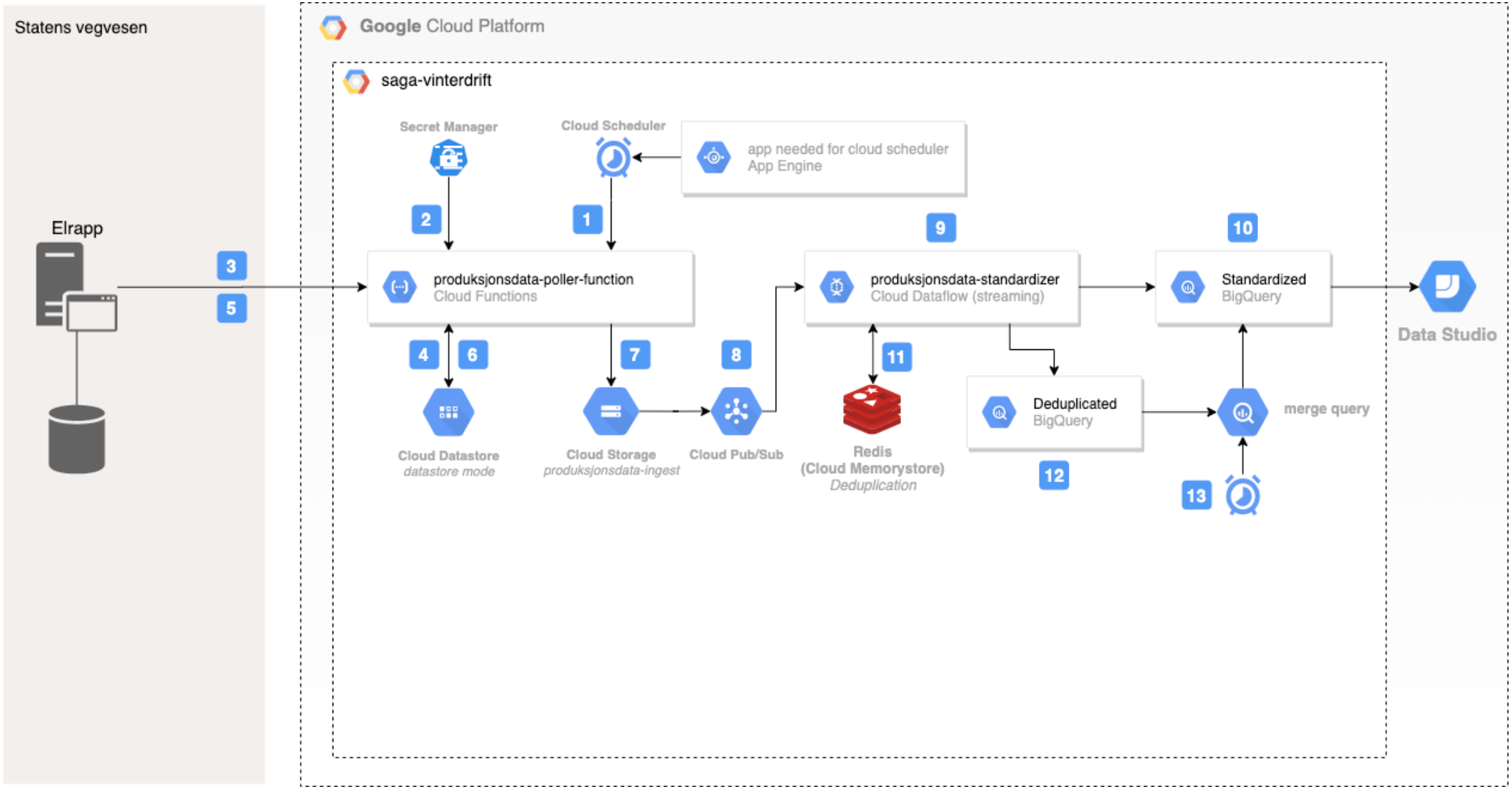
Statens vegvesen

- Alle ressurser i GCP må ligge i et prosjekt
- Til i dag har vi laget et eget GCP-prosjekt til hver av dere
- Dere skal bruke dette prosjektet under workshoppen
- Vi kommer til å slette dette prosjektet i slutten av mai. Det medfører at alle ressursene i prosjektet blir ryddet opp, slik at vi slipper å ha løpende kostnader på ressursene som blir opprettet i dag.
- Vil du fortsette å utforske GCP? Vi har også opprettet et prosjekt til alle deltakerne som kan brukes til videre utforsking



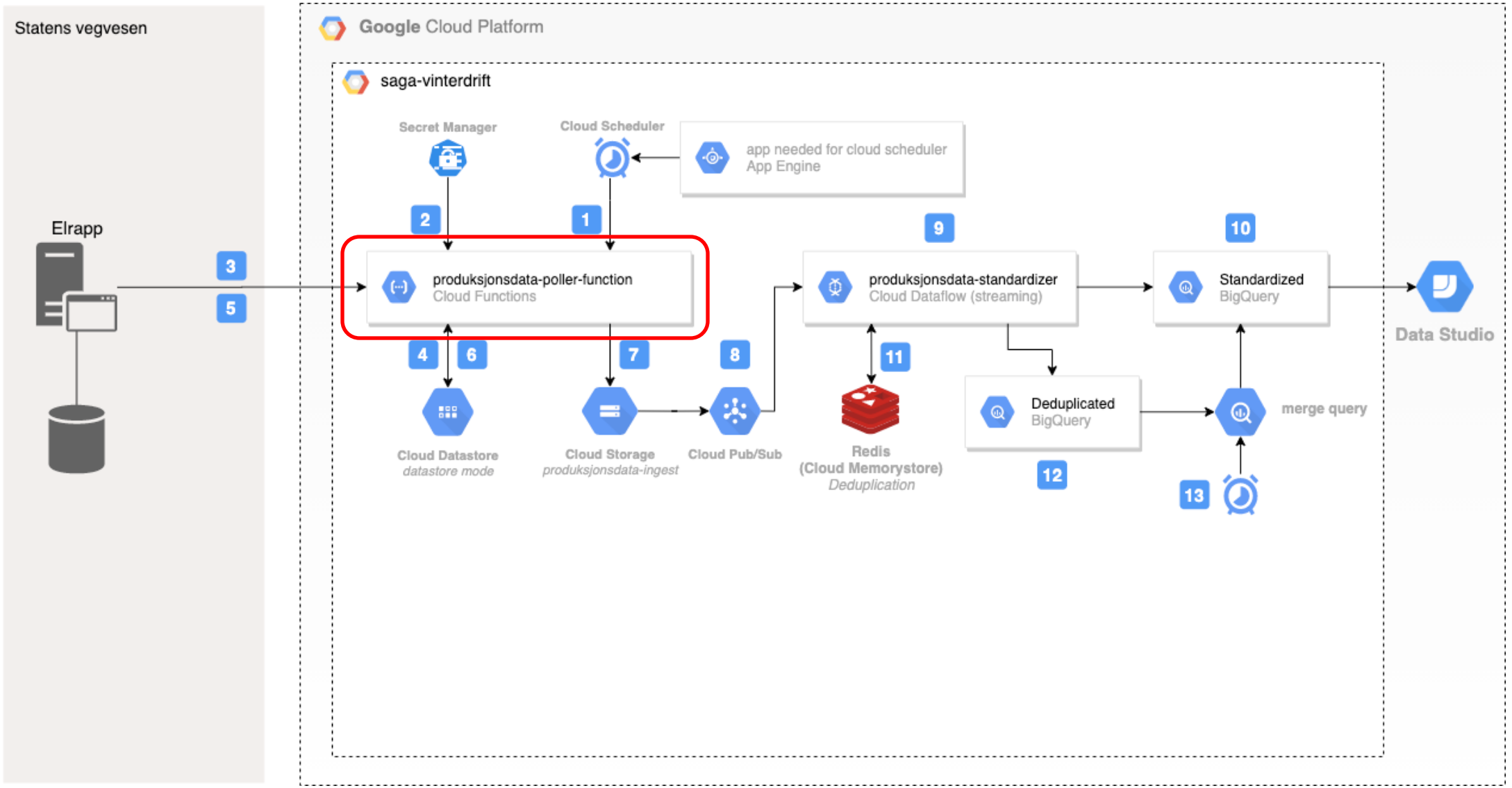
Eksempel på en produksjonspipeline på GCP - vinterdrift

Statens vegvesen



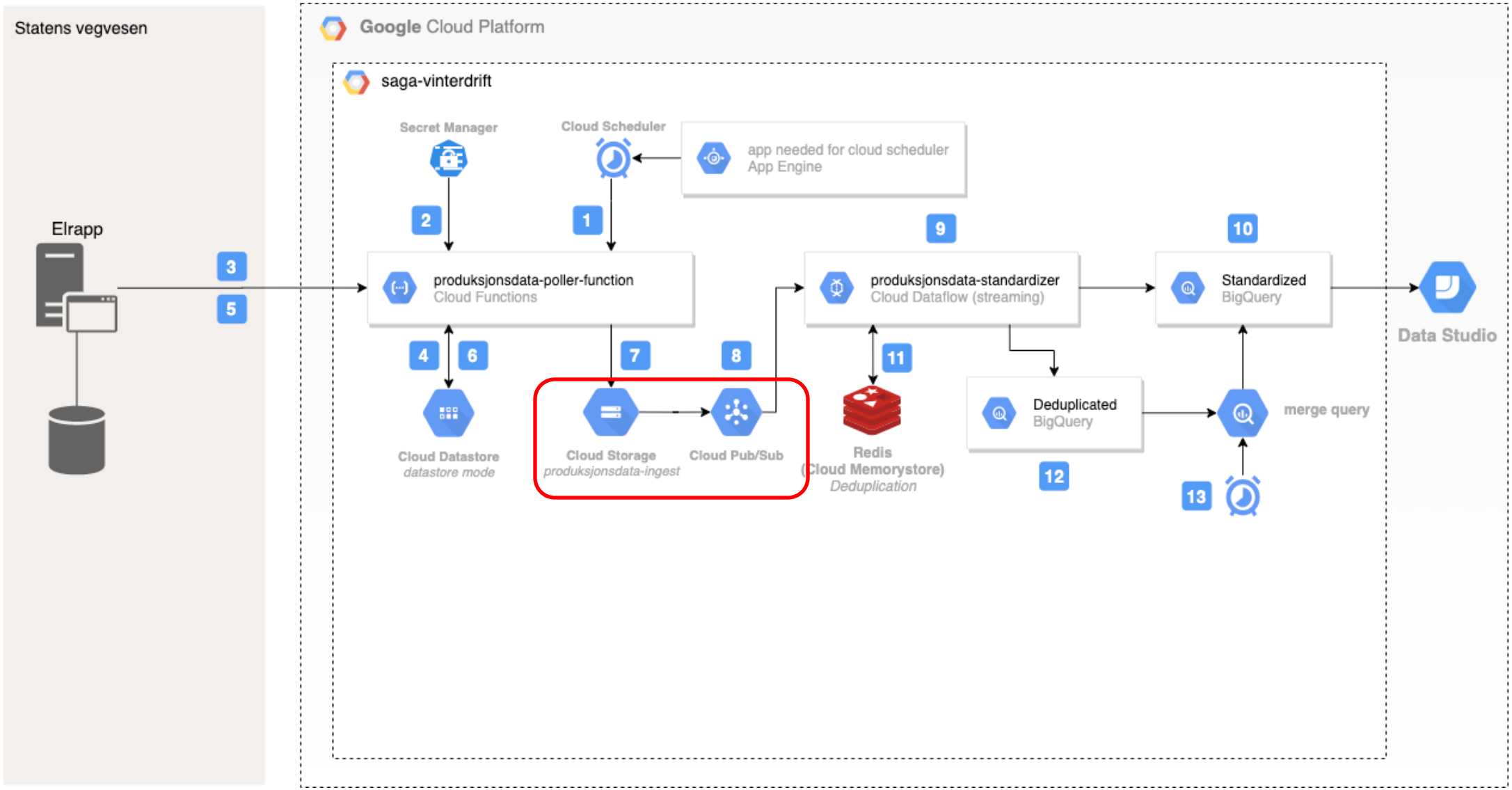
Eksempel på en produksjonspipeline på GCP - vinterdrift

Statens vegvesen



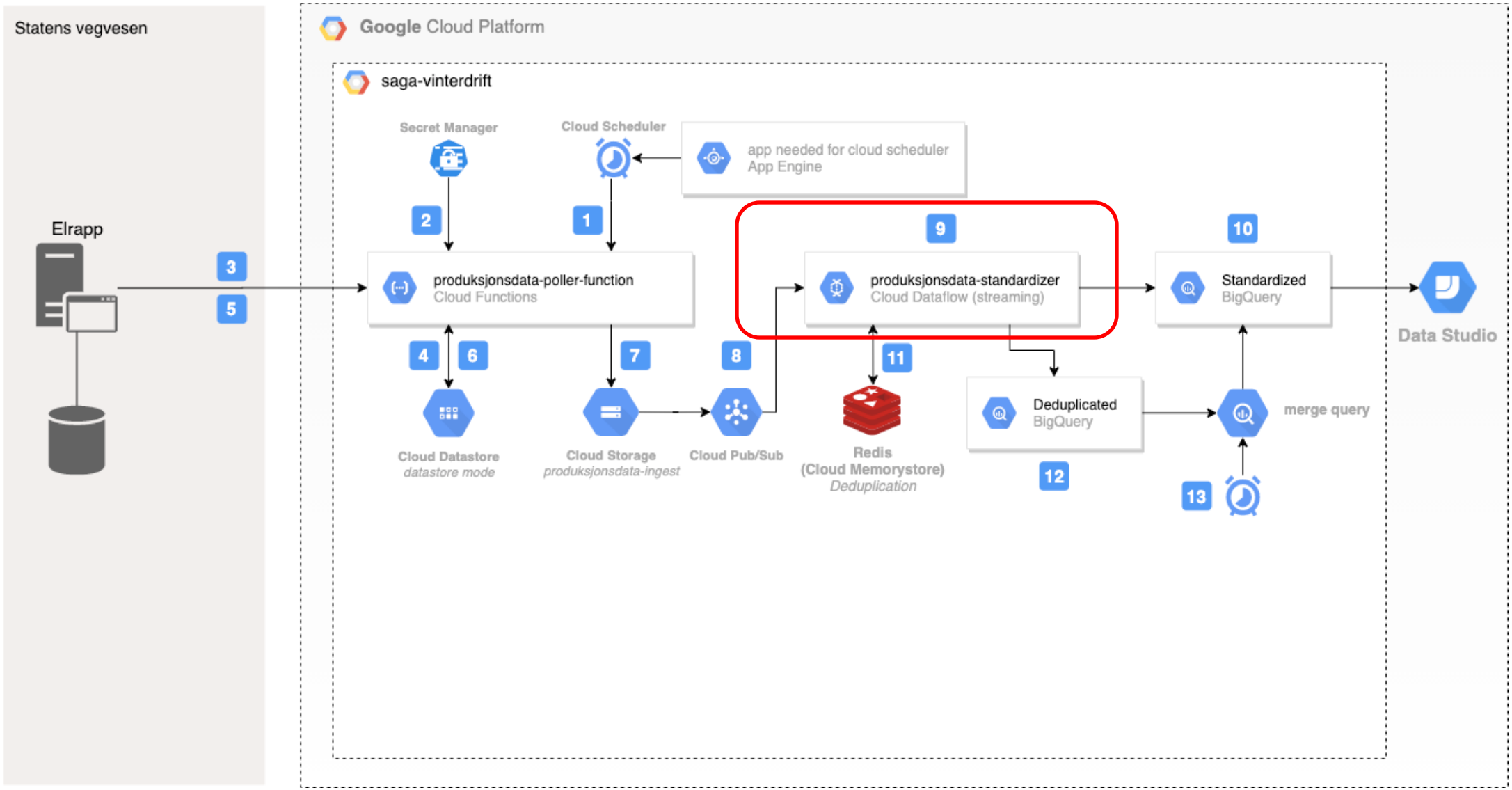
Eksempel på en produksjonspipeline på GCP - vinterdrift

Statens vegvesen



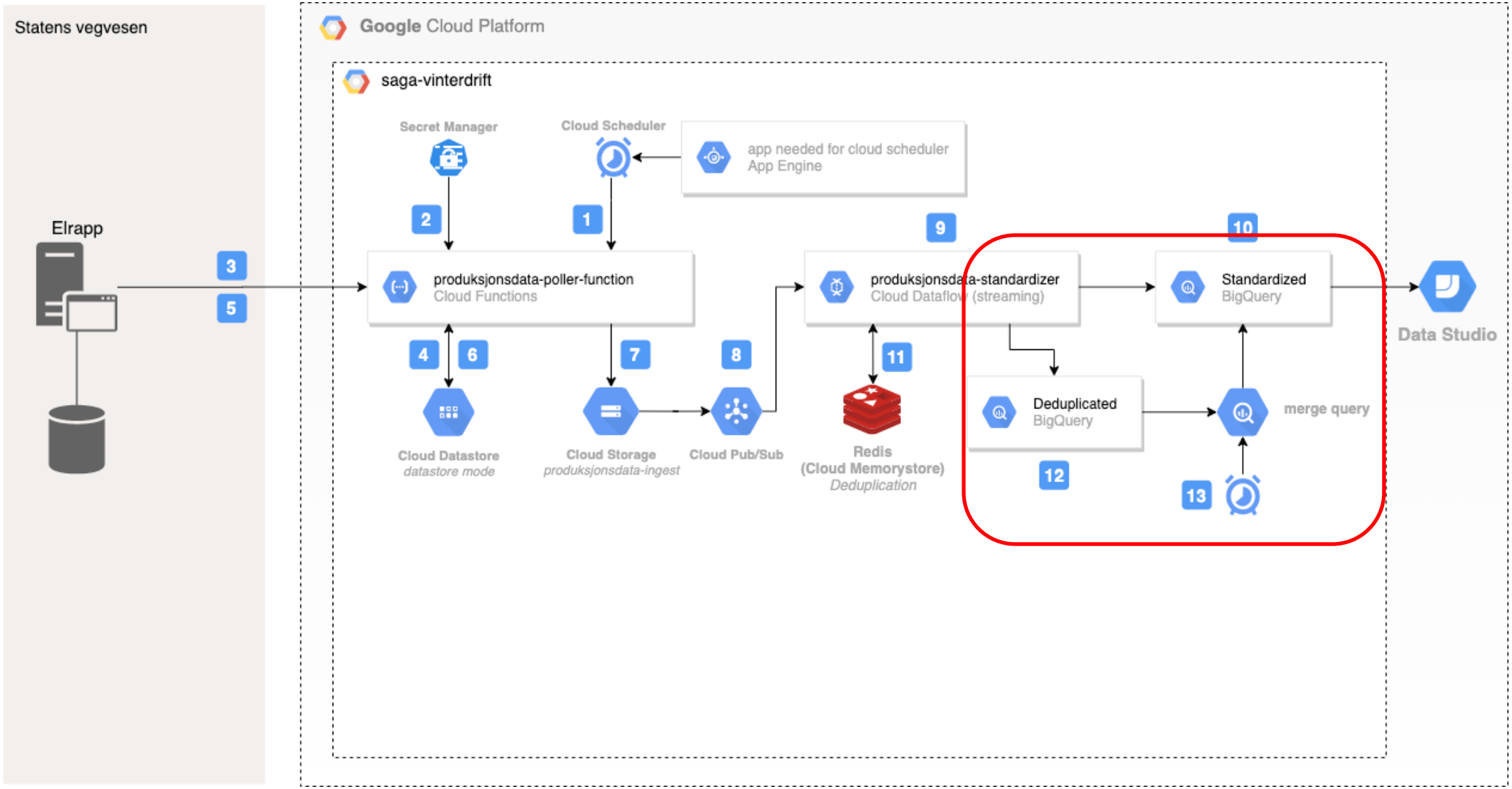
Eksempel på en produksjonspipeline på GCP - vinterdrift

Statens vegvesen



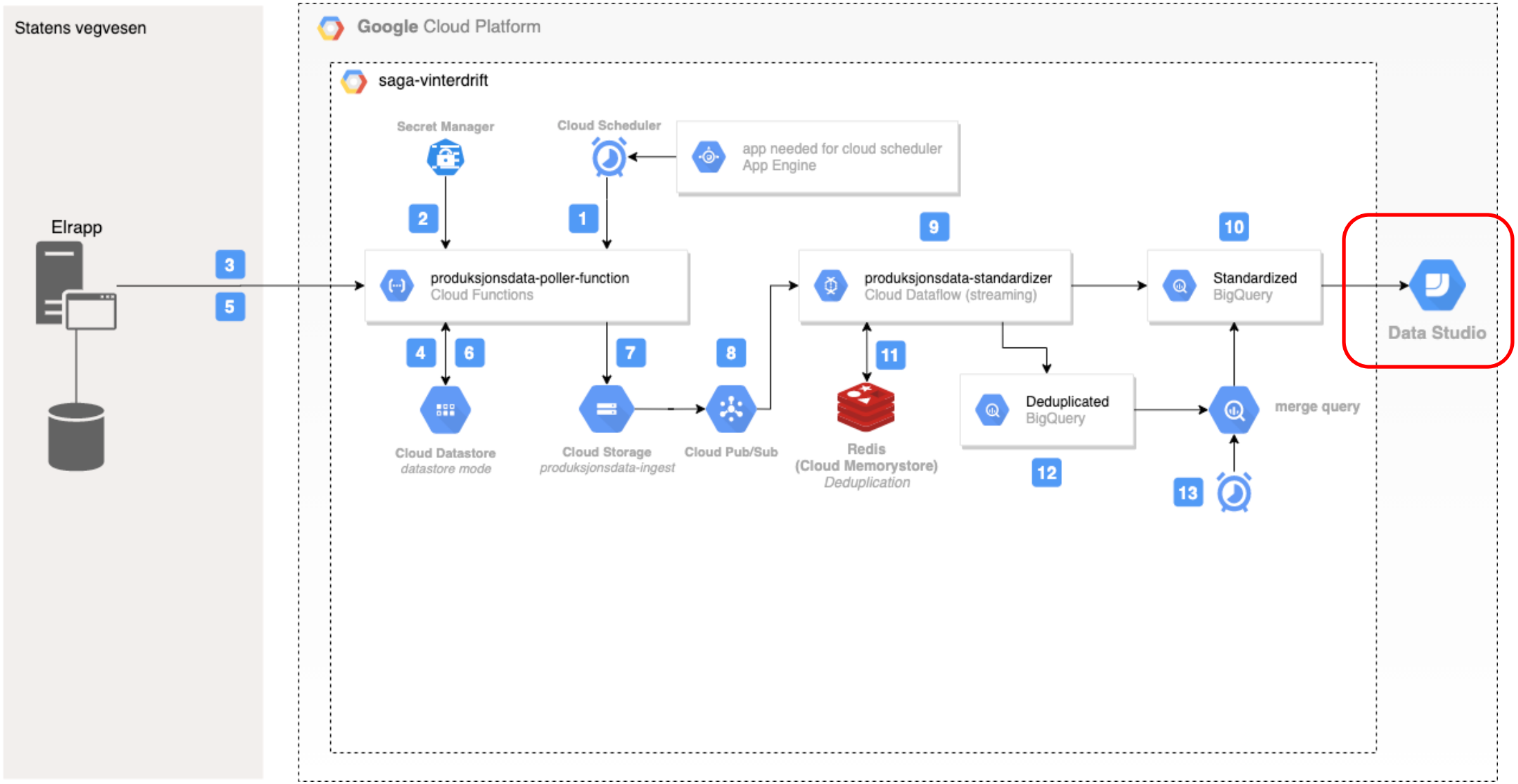
Eksempel på en produksjonspipeline på GCP - vinterdrift

Statens vegvesen



Eksempel på en produksjonspipeline på GCP - vinterdrift

Statens vegvesen



Målet for dagen er at dere skal kunne…

Statens vegvesen

- Laste opp egne datasett til GCS og BigQuery
- Hente data gjennom API og Data Catalog
- Gjøre enkle transformasjoner med FME på data som ligger i GCS og BigQuery
- Bruke BigQuery som et analyseverktøy
- Forstå BigQuerys sentrale styrker og begrensinger
- Bruke Data Studio og Geo Viz til utforskende analyse og enkel visualisering

Greit å vite

Statens vegvesen

- Dere vil snart bli delt inn i grupper som skal sitte sammen under workshoppen
  - Vi sprer oss også utover i gruppene som fasilitatorer
- Spør gruppa eller oss dersom dere står fast
- Lav takhøyde - det er greit å ikke forstå!

- **De fleste sesjonene har flere oppgaver enn dere rekker å utføre. Det er helt greit, dere trenger ikke å komme gjennom alt!**
  - Dersom dere synes at temaene er interessante kan dere alltids utføre flere oppgaver senere

- **Det er mange interessante temaer vi ikke rekker å gå gjennom i dag. Diskusjon om veien videre mot slutten av workshoppen.**
  - **Noter ned kommentarer, spørsmål, ting dere savner og det som var bra i workshoppen. Vi tar gjerne i mot feedback mot slutten av dagen.**

Tidsplan

Statens vegvesen

- 10:00 – 10:30: Felles intropresentasjon
- 10:30 – 11:30: Interaktiv sesjon 1 – Innlasting av data til GCS og BigQuery
- 11:30 – 12:15: Lunsjpause
- 12:15 – 13:10: Interaktiv sesjon 2 – Datatransformering med FME
- 13:10 – 13:20: Pause
- 13:20 – 14:30: Interaktiv sesjon 3 – Analyse med BigQuery og GeoViz
- 14:30 – 14:40: Pause
- 14:40 – 15:30: Interaktiv sesjon 4 – Visualisering med Data Studio
- 15:30 – 16:00: Oppsummering og vegen videre

Grupper og lenke til workshop-materialet

Statens vegvesen

URL til alt workshop-materiale: www.tinyurl.com/saga-workshop

Martin og Peder

- Tomas Levin
- Snorre Hansen
- Kenneth Sørensen
- Lars Meisingset

Safi

- Jan Kristian Jensen
- Bjor Grønnevet
- Finn Tore Johansen

Sondre og Gunnar

- Joakim Møyholm
- Jorunn Riddervold Levy
- Harald Stensholt
- Christian Berg Skjetne

Statens vegvesen

Verktøy og muligheter i Google-skyen

Oppsummering

Målet for dagen var at dere skal kunne…

Statens vegvesen

- Laste opp egne datasett til GCS og BigQuery
- Hente data gjennom API og Data Catalog
- Gjøre enkle transformasjoner med FME på data som ligger i GCS og BigQuery
- Bruke BigQuery som et analyseverktøy
- Forstå BigQuerys sentrale styrker og begrensinger
- Bruke Data Studio og Geo Viz til utforskende analyse og enkel visualisering

Har vi lært dette?

Var opplegget lett, vanskelig eller akkurat passe?

Er det noe dere kunne tenke dere å lære mer om?

Veien videre

Statens vegvesen

- Hvordan kan Saga sikre god oppfølging i etterkant av workshopen? Hva har dere behov for?
- Hvordan kan vi opprettholde kontakt og dele erfaringer?
- Hva er aktuelle temaer for fremtidige, kortere workshops?
- Hvilke kapabiliteter har dere behov for i dataplattformen?