Norges miljø- og biovitenskapelige universitet

Masteroppgave 2018 30 studiepoeng
Fakultet for realfag og teknologi
Hovedveileder Sonja Monica Berlijn

Cybersikring av en operasjonell data lake

Cyber securing an operational data lake

Tommy Haugen
Industriell økonomi
Fakultet for realfag og teknologi

Sammendrag

Som systemansvarlig i det norske kraftsystemet vil Statnett SF samle inn store mengder data i årene fremover. For å legge til rette for mulighetene i fremtidens datamengder har Statnett SF kjøpt inn en ny dataplattform som inneholder en data lake. Data laken skal håndtere store og varierte datamengder, og legge til rette for at Statnett SF kan utnytte en større andel av fremtidens data. Samtidig som data laken legger til rette for bedre utnyttelse av Statnett SF sin totale datamengde kan den åpne for et bredere sårbarhetsbildet.

Denne masteroppgaven er initiert av Statnett SF og har tatt utgangspunkt i følgende problemstilling: Hvilke hovedpunkter burde Statnett SF fokusere på med tanke på cybersikring av en operasjonell data lake?

Med hjelp av en risikovurdering basert på litteratur er det kommet frem til fem overordnede cyberrisikoer som er relevante for en operasjonell data lake (se kapittel 4.1.5 for detaljer). Med hjelp av et litteraturstudium er det kommet frem til ni overordnede sikkerhetstiltak som er relevante for en operasjonell data lake (se kapittel 4.2.4 for detaljer).

Statnett SF vil aldri kunne sikre seg 100 % mot uønskede cyberhendelser. Teknologiutviklingen gjør at sårbarhetsbildet i et cyberperspektiv er uoversiktlig, og dermed vanskelig å kartlegge presist.

I denne masteroppgaven konkluderes det med at Statnett SF burde fokusere på å lære opp brukere av data laken i cybersikkerhet. Opplæringen kan gjøres med eksempelvis jevnlig kursing og bevisstgjøringskampanjer. Masteroppgaven konkluderer også med at Statnett SF burde fokusere på å samordne teknologiutvikling og cybersikkerhet for data laken sterkere, og se til at disse utvikles i takt.

Grunnet oppgavens tema er enkelte deler av oppgaven underlagt taushetsplikt etter energiloven § 9-3 jf bfe § 6-2 og unntatt fra innsyn etter offentleglova § 13, 1. ledd. Taushetsbelagt informasjon er plassert i vedlegg C-G.

Abstract

As the transmission system operator in Norway, Statnett SF will collect a great amount of data in the years to come. To facilitate the opportunities on the future data volumes, Statnett SF has purchased a new big data platform, which contains a data lake. The data lake will handle the large amount of data and facilitate Statnett SF to exploit a larger proportion of the data in the future. At the same time, the data lake can increase the vulnerabilities.

This master thesis is initiated by Statnett SF and has been formed around the following research question: What main points should Statnett SF focus on regarding cyber securing an operational data lake?

By doing a risk assessment based on literature, it is identified five overall cyber risks relevant to an operational data lake (se chapter 4.1.5 for details). By doing a literature review, it is identified nine overall security measures that is relevant for cyber securing an operational data lake (se chapter 4.2.4 for details).

Statnett SF will never be able to be 100 % secured against unwanted cyber events. Technology development makes the vulnerabilities in a cyber perspective insignificant, and thus difficule to map accurately.

In this master thesis, it is concluded that Statnett SF should focus on training users of the data lake in cyber security. Training can be done with frequent coursing and awareness campaigns. It is also concluded that Statnett SF should focus on stronger coordination between technology development and cyber security for the data lake, and make sure to develop the two in line.

Due to the topic of this master thesis, parts of the thesis is subject to confidentiality by the Norwegian Energy Act (energiloven) § 9-3 cf. bfe § 6-2 and except from access by the Norwegian Freedom of Information Act (offentleglova) § 13, 1.paragraph. Confidential information is to be found in appendix C-G.

Forord

Denne masteroppgaven markerer slutten på mine fem år som student ved Norges miljø- og biovitenskapelige universitet (NMBU). Jeg ser tilbake på noen lærerike, utfordrende og spennende studieår.

Først vil jeg takke min hovedveileder ved NMBU Sonja Monica Berlijn for veiledning, støtte og motivasjon gjennom hele masterperioden. Videre vil jeg takke Hasse Kristiansen i Statnett SF for hjelp og innspill i tilknytning til masteroppgaven. Jeg vil også takke min biveileder ved NMBU Arne Auen Grimenes for innspill, og alle andre i Statnett SF som har bidratt i arbeidet denne våren.

Avslutningsvis vil jeg takke familie og venner for oppmuntrende ord og gode stunder gjennom hele studietiden.

Oslo, mai 2018

_____

Tommy Haugen

Innholdsfortegnelse

Sammendrag I
Abstract II
Forord III
Innholdsfortegnelse V
Liste over figurer VII
Liste over tabeller VIII
1 Introduksjon 1
1.1 Bakgrunn 1
1.2 Problemstilling 2
1.3 Metode 2
1.4 Avgrensninger 4
2 Big data 5
2.1 Big dataplattform 5
2.2 Data lake 8
3 Cybersikkerhet 11
3.1 Informasjonssikkerhet i Statnett 11
3.1.1 Relevante policyer og instrukser 12
3.1.2 Soneinndeling 13
3.2 Sikkerhet big dataplattform 13
3.3 Automatisk diagnostisering (AutoDig) 14
4 Risikovurdering og sikkerhetstiltak 17
4.1 Risikovurdering 17
4.1.1 Verdier 17
4.1.2 Cybertrusler 17
4.1.3 Cyberangrep 18
4.1.4 Sårbarheter 20
4.1.5 Analyse cyberrisiko 23
4.2 Sikkerhetstiltak 25
4.2.1 Data lake 25
4.2.2 Big datainfrastrukturer 27
4.2.3 Digitale systemer i norsk kritisk infrastruktur 28
4.2.4 Oppsummering sikkerhetstiltak 29

5 Diskusjon og analyse 31
5.1 Cyberrisiko 31
5.2 Sikkerhetstiltak 32
5.3 Sammenstilling 33
5.4 Praktisering av sikkerhetstiltakene 35
5.5 Utfordringer Statnett 36
5.6 Avsluttende betraktninger 37
6 Konklusjon og anbefaling til videre arbeid 39
6.1 Konklusjon 39
6.2 Anbefaling til videre arbeid 39
Referanser i
Vedlegg A – Utsendte spørsmål I
Vedlegg B – Svar på utsendte spørsmål III
Vedlegg C – Risikoer, sikkerhetstiltak og sammenstilling av disse fra ROS-analysen til AutoDig V
Vedlegg D – Verdier fra ROS-analysen til AutoDig XI
Vedlegg E – Diskusjon av cyberrisiko XIII
Vedlegg F – Diskusjon av sikkerhetstiltak XVII
Vedlegg G – Eksempel på sammenstilling XXIII

Liste over figurer

Figur 1: Hovedsonene i verdikjeden for dataene til Statnett sin big dataplattform.. 6
Figur 2: Hvordan data laken er plassert med tanke på omkringliggende soner og Innsikt.....8
Figur 3: Eksempel på hvordan HDFS og MapReduce fungerer.. 10

Liste over tabeller

Tabell 1: Forklaring på de ulike stegene i risikovurderingen.. 3
Tabell 2: Karakteristiske trekk og forklaring av big data. 5
Tabell 3: Hovedsoner for Statnett sin big dataplattform. 7
Tabell 4: Noen av kildene som vil levere data til Statnett sin data lake.. 8
Tabell 5: Relevante sikkerhetspolicyer og -instrukser i Statnett som data laken vil være omfanget av. 12
Tabell 6: Sikkerhetsonene i Statnett. 13
Tabell 7: Sikkerhetskomponentene som tilhører data laken.. 14
Tabell 8: Identifiserte sårbarheter fra litteraturstudiet. 22
Tabell 9: Cyberrisikoer som er funnet gjennom en risikovurdering. 24
Tabell 10: Identifiserte sikkerhetstiltak på bakgrunn av litteraturstudiet med forklaring. 29
Tabell 11: Forklaring på hvordan sammenstillingen er gjennomført. 33
Tabell 12: Sammenstilling av sikkerhetstiltak og cyberrisikoer fra litteraturstudiet og risikovurderingen. 34

1 Introduksjon

1.1 Bakgrunn

Som systemansvarlig i det norske kraftsystemet har Statnett SF (heretter kalt Statnett) ansvar for å sikre strømforsyningen gjennom drift, overvåkning og beredskap på en samfunnsmessig rasjonell måte [1]. I Statnett sin konsernstrategi for 2017 – 2021 er to av de definerte målene at Statnett skal bygge et solid digitalt fundament og bli ledende på digital sikkerhet i bransjen [2].

I fremtiden vil datamengdene Statnett skal behandle ha stort volum, genereres fra mange ulike kilder og i høy hastighet. Fremtidens datamengde kalles big data, og for å utnytte mulighetene som kommer med disse har Statnett kjøpt inn en big dataplattform [9]. Den innkjøpte big dataplattformen inneholder en data lake, som er et lagringssted for ubehandlede data. En data lake klarer å lagre big data, og legger til rette for at data fra ulike format kan kombineres i analyser, som igjen kan forbedre Statnett sitt analysearbeid [3].

Samtidig som data lake legger til rette for fremtidens muligheter i kraftsystemet kan den åpne for nye risikoer. Cybersikkerhet omfatter tiltak for beskyttelse mot reelle og potensielle trusler som kanaliserer via IKT[1]-infrastruktur [4]. Julen 2015 ble Ukraina rammet av et cyberangrep som slo ut strømforsyningen til nærmere 250 000 husstander, og PST trekker frem kritisk infrastruktur i Norge som et utsatt mål for digital overvåkning og sabotasje i sin trusselvurdering for 2018 [5, 6]. De to punktene fra konsernstrategien til Statnett må følgelig sees i sammenheng.

Med dette bakteppet er det derfor ønskelig å undersøke hvilke cyberrisikoer og sikkerhetstiltak litteratur fremhever med tanke på cybersikring av en operasjonell data lake. Målet med denne oppgaven er å vise hva slags cyberrisikoer og sikkerhetstiltak med tanke på cybersikring av en operasjonell data lake, og deretter identifisere hva Statnett burde fokusere på med tanke på cybersikring av sin operasjonelle data lake.

[1] IKT står for informasjons- og kommunikasjonsteknologi.

Grunnet oppgavens tema er deler av oppgaven underlagt taushetsplikt etter energiloven § 9-3 jf bfe § 6-2 og unntatt fra innsyn etter offentleglova § 13. Taushetsplikten har ingen utløpsdato, og masteroppgaven kan klausuleres i kun fem år. For å oppfylle taushetsplikten er sentrale deler av oppgaven plassert i vedlegg. Det gis henvisninger i løpende tekst til vedlegg, og leseren anbefales å lese disse.

1.2 Problemstilling

Denne masteroppgaven har tatt utgangspunkt i følgende problemstilling:

Hvilke hovedpunkter burde Statnett fokusere på med tanke på cybersikring av en operasjonell data lake?

Problemstillingen er delt inn i to forskningsspørsmål:

- Hvilke cyberrisikoer er relevante for en operasjonell data lake?
- Hvilke tilnærminger, konsepter og praksiser er relevante for cybersikring av en operasjonell data lake?

1.3 Metode

For å besvare problemstillingen og forskningsspørsmålene i denne oppgaven er det gjort et litteraturstudie og en risikovurdering.

I risikovurderingen er risiko analysert som en funksjon av verdier, trusler og sårbarheter. Det er i risikovurderingen analysert hendelser som ligger i krysningen mellom disse tre, samt hvilke angrepsformer og motiv trusselaktører har for å gjennomføre angrepet [7]. Risikovurderingen baserer seg på publikasjoner og undersøkelser som er gjennomført i tilknytning til big datainfrastrukturer, informasjonssikkerhet/cybersikkerhet i kraftforsyning og norsk kritisk

infrastruktur. I risikovurderingen er det lagt hovedvekt på å undersøke sårbarheter, og for verdiene i data laken er Statnett sin analyse lagt til grunn.

Risikovurderingen benyttet i denne oppgaven er tilpasset problemstillingen og målet til denne oppgaven, og er en moderert versjon fra Nasjonal Sikkerhetsmyndighet (NSM) sin håndbok Risikovurdering for sikring [7]

Tabell 1: Forklaring på de ulike stegene i risikovurderingen. Risikovurderingen er basert på NSM sin håndbok, men forenklet opp mot oppgavens mål [7].

Steg i risikovurdering	**Forklaring**
Vurdering av verdier	Verdiene i den operasjonelle data laken er identifisert på bakgrunn av Statnett sine analyser.
Vurdering av trusler	Det er søkt i ulik litteratur etter trusler som er relevante for Statnett og deres operasjonelle data lake.
Beskrivelse av scenarier	Det er søkt i ulik litteratur etter mulige angrepsformer en trusselaktør kan bruke for å nå en operasjonell data lake, samt hvilke motiv trusselaktøren har for angrepet.
Vurdering av sårbarheter	Det er søkt i ulik litteratur etter mulige sårbarheter som er relevante for en operasjonell data lake.
Sammenstilling av verdier, trusler, scenarier og sårbarheter	Det er analysert hvilke uønskede cyberhendelser som ligger i krysningen mellom verdier, trusler og sårbarheter for å identifisere hvilke cyberrisikoer en operasjonell data lake er omfattet av.

Det er deretter gjort et litteraturstudium for å identifisere mulige sikkerhetstiltak for å redusere de analyserte cyberrisikoene. Litteraturen innen cybersikkerhet og data lake er svært beskjeden, og det er derfor gått mer analytisk til verks for å finne relevant litteratur. Søkene er gjort for å i størst mulig grad oppfylle tre perspektiver: data lake, cybersikkerhet/informasjonssikkerhet i big datainfrastrukturer og cybersikkerhet/informasjonssikkerhet i digitale systemer i norsk kritisk infrastruktur. For data lake er det søkt etter teknologien data lake bygger på og big datainfrastrukturer, i tillegg til litteratur som retter seg mot digitale systemer i kritisk norsk infrastruktur.

Litteratursøkene er gjennomført i ulike søkemotorer som Google Scholar og Science Direct, i tillegg til generelle Google-søk og undersøkelser i tekstbøker. Litteratursøkene er gjort på både norsk og engelsk.

Cyberrisikoene og sikkerhetstiltakene fra risikovurderingen og litteraturstudiet er sammenlignet med en risiko- og sårbarhetsanalyse (ROS-analyse) gjennomført av Statnett for

deres data lake, samt andre sikkerhetstiltak som er definert i Statnett og som vil fungere på deres operasjonelle data laken.

Det er i tillegg gjennomført en kort undersøkelse blant europeiske sentralnettseiere angående deres forhold til cybersikkerhet og data laker. Spørsmålene som er sendt ut og svarene fra undersøkelsen er vedlagt i vedlegg A og B.

1.4 Avgrensninger

Risikovurderingen og litteraturstudiet er gjennomført med følgende avgrensninger:

- Det er ikke sett på en data lake som er koblet til en skyløsning.
- Det er ikke vurdert sannsynlighet og konsekvens i risikovurderingen.
- Det er sett bort fra litteratur publisert av leverandører.
- Det er ikke sett på IKT-sikkerhet.
- Det er sett bort fra sikkerhetsstandarder.

2 Big data

Stadig flere komponenter kobles til internett og genererer informasjon. Internet of Things (IoT) viser til komponenter som kjøleskap, panelovner o.l. som kan styres via internett. I kraftsystemet er automatiske strømmålere, droner og sensorer eksempel på komponenter som vil generere store mengder informasjon som senere ønskes bruk i analyser.

Big data er et begrep som fanger opp fremtidens datamengder. Begreper omfatter mer enn bare størrelsen på datamengden, og kan forklares ut fra fire underliggende begreper. De fire underliggende begrepene er forklart i Tabell 2 [8]:

Tabell 2: Karakteristiske trekk og forklaring av big data [8].

Karakteristisk trekk	**Forklaring**
Volum	Store datamengder som er stadig økende, og som forventes at øker i fremtiden.
Variasjon	Dataene kommer fra mange ulike kilder, og i strukturert og ustrukturert format. Strukturerte data er data som enkelt kan lagres som vanlige filer, mens ustrukturerte data har sin egen struktur og tilhørende utfordringer med å bli lagret. IoT-komponenter vil gjøre at mengden ustrukturerte data øker i fremtiden.
Hastighet	Dataene blir generert og levert i høy hastighet.
Verdi	Dataene utgjør en stor verdi for bedrifter, og åpner for forbedrede analyser. De forbedrede analysene kommer av at data fra ulike format kan kombineres.

2.1 Big dataplattform

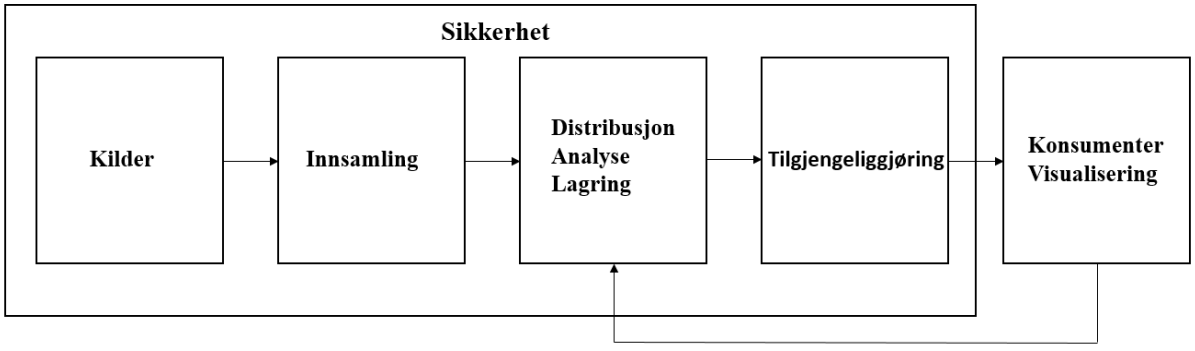
I dag er arbeidet med å lagre alle målingene Statnett gjør så tid- og ressurskrevende at flere av målingene ikke blir riktig håndtert, og dermed ikke analysert. Grunnen til at lagringen av målingene er så tid- og ressurskrevende er blant annet at de kommer i format som Statnett til nå ikke har hatt teknologi til å håndtere på en tilfredsstillende måte. Selv om målingene ikke blir benyttet i dag betyr ikke at de ikke er verdifulle, og ettersom andelen big data forventes at vil øke i fremtiden er det behov for å legge til rette for en ny dataplattform som kan lagre disse dataene [9].

Tradisjonelt har Statnett benyttet datavarehuset *Innsikt* for lagring av data. Et datavarehus er et lagringssted der innkommende data blir plassert i et forhåndsdefinert format, og lagret til senere bruk. For at de innkommende dataene skal kunne lagres må det være mulig å tilpasse dem til det forhåndsdefinerte formatet, og datavarehus er tilpasset hovedsakelig å håndtere strukturerte data [3]. Ved behov for økt lagringskapasitet har Statnett utvidet *Innsikt* ved å anskaffe ekstra kapasitet.

For å møte big data har Statnett iverksatt prosjektet Finbeck med mål om å bygge et digitalt fundament for mer effektiv gjennomføring av selskapets prosesser og oppgaver. Finbeck har identifisert ulike bruksområder til Statnett sin nye big dataplattform. Bruksområdene handler handler om å i større grad utnytte sanntidsdata, bruke automatiske datahåndteringsprosesser og kombinere data i ulike format til analyseformål. Flere av bruksområdene håndteres av ulike deler innad i Statnett i dag, og big dataplattformen skal være med å sentralisere alle bruksområdene i en felles plattform [9].

Statnett har kjøpt inn en big dataplattform av IBM/Hortonworks. Big dataplattformen skal utvikles i tre steg. Første generasjon av dataplattformen skal settes i drift våren 2018, og tredje generasjon dataplattform skal være i drift i 2020 [9].

I Figur 1 er en illustrasjon basert på arkitekturskissen for big dataplattformen til Statnett vist. Illustrasjonen viser hvordan data er tenkt at skal bevege seg fra kilder til konsumenter. Kilder, innsamling, distribusjon/analyse/lagring, tilgjengeliggjøring og konsumenter/visualisering er hovedsoner til arkitekturen for den nye big dataplattformen, og sikkerhet er ment som en overordnet sone.



Figur 1: Hovedsonene for Statnett sin big dataplattform. Kilder, innsamling, distribusjon/analyse/lagring, tilgjengeliggjøring og konsumenter/visualisering er hovedsoner i big dataplattformen. Sikkerhet er ment som en overordnet sone. Pilene viser hvordan data er tenkt at skal flyte. Figuren er illustrert av Tommy Haugen og basert på arkitekturskisse for Statnett sin big dataplattform [9].

De fem hovedsonene for big dataplattformen til Statnett er nærmere forklart i Tabell 3 [9]:

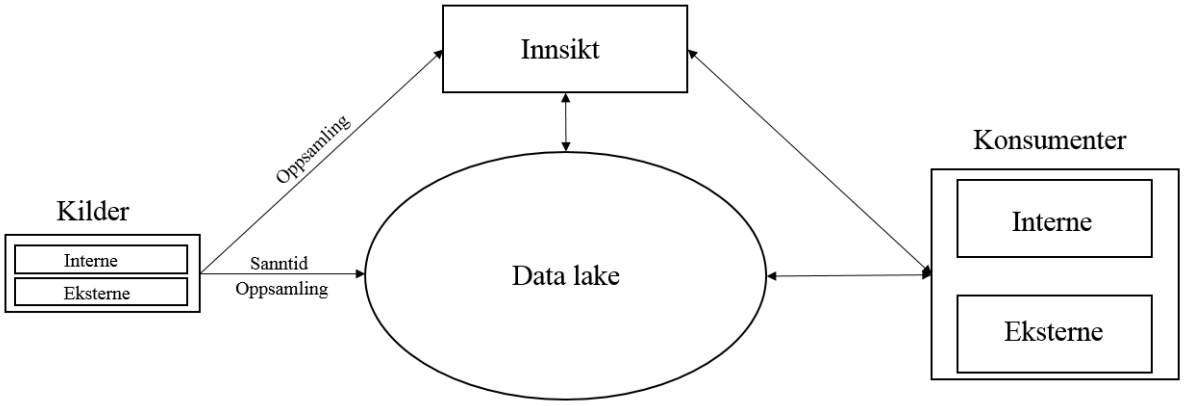
Tabell 3: Hovedsoner for Statnett sin big dataplattform [9].

Hovedsone	**Forklaring**
Kilder	Den nye dataplattformen skal samle data fra ulike kilder. Dataene vil være big data. Kildene vil være både interne og eksterne. Med interne kilder forstås kilder eid av Statnett, mens eksterne er kilder som Statnett ikke eier. Kildene vil gjennomgås nøyere i kapittel 2.2.
Innsamling	Dataene blir samlet inn ved hjelp av ulike komponenter i denne hovedsonen. Innsamlingen vil foregå som oppsamling eller i sanntid. Med oppsamling menes data som samles opp i større enheter før det blir sendt videre, mens sanntidsdata er data som kontinuerlig blir sendt inn til dataplattformen.
Distribusjon Analyse Lagring	I denne hovedsonen vil dataene gå gjennom prosesser for enten distribusjon, analyse eller lagring. Her ligger både datavarehuset *Innsikt* og data laken. Data laken gjennomgås nøyere i kapittel 2.2.
Tilgjengeliggjøring	Dataene gjøres tilgjengelig ved hjelp av ulike komponenter i denne sonen.
Konsumenter Visualisering	I siste sone vil eksterne og interne konsumenter få tilgang til dataene gjennom ulike systemer og komponenter. Det er tenkt at eksterne skal få tilgang til en kopi av dataene, og ikke originaldata. Enkelte bearbeidede data fra denne sonen vil sendes tilbake til distribusjon/analyse/lagring. Konsumentene gjennomgås nøyere i kapittel 2.2.

Sikkerhet tilknyttet big dataplattformen gjennomgås nøyere i kapittel 3.2.

2.2 Data lake

En av komponentene i Statnett sin big dataplattform er en data lake. En data lake er et lagringssted for ubehandlede data, både ustrukturerte og strukturerte [3]. Med ubehandlede data forstås originaldata som lagres som de er. Det finnes ulike leverandører som tilbyr data lake-løsninger, og Statnett har kjøpt sin data lake-løsning av IBM/Hortonworks [9]. Hvordan data laken er plassert med tanke på omkringliggende soner og datavarehuset *Innsikt* i big dataplattformen er vist i Figur 2.



Figur 2: Hvordan data laken er plassert med tanke på omkringliggende soner og Innsikt. Pilene viser hvordan data er tenkt at skal bevege seg i tilknytning til data laken. Pilene fra kilder til data lake/Innsikt er også ment som sonen innsamling, og pilene fra data lake/Innsikt til konsumenter er ment som tilgjengeliggjøring. Illustrasjon er laget av Tommy Haugen.

I Figur 2 er kildene tegnet inn som enten interne eller eksterne, og de leverer data i form av oppsamling eller sanntid til data laken. Merk at av data laken og Innsikt er det kun førstnevnte som skal behandle sanntidsdata. I hovedsak vil data laken inneholde sensordata (eksempelvis fra PDC, se tabell 4), værdata, bilde/video og behandlede datasett [10]. Et utvalg av kildene som vil levere data til Statnett sin data lake er forklart i Tabell 4 [9]. Merk at Tabell 4 ikke er en fullstendig liste over alle kildene til data laken.

Tabell 4: Noen av kildene som vil levere data til Statnett sin data lake. Det er også gitt en forklaring på kildene, og vist om kildene er interne eller eksterne [9].

Type kilde	Navn på kilde	Forklaring
Intern	PDC	Data som kommer fra Phasor Measurement Units (PMU).
	DFR	Feilskriverdata fra vern i kraftnettet.
	SCADA	Gir informasjon fra overvåking og drift av kraftnettet.
Ekstern	Met.no	Lynmeldinger og værdata fra Meteorologisk institutt.
	Drone fleet data capture	Dronebilder.

Statnett har estimert at deres data lake vil inneholde datamengder tilsvarende 10 PB[2], men antar at mengden vil være større [9].

Konsumentene til data laken er enten interne eller eksterne brukere eller systemer. Enkelte data skal bearbeides av eksterne eller interne konsumenter før de lagres i data laken eller *Innsikt*. Noe data vil kun prosesseres ved hjelp av data laken, og leveres direkte til konsumenter uten lagring i data laken eller *Innsikt*. I Figur 2 er dette illustrert med piler i begge retninger mellom data laken og konsumenter [9].

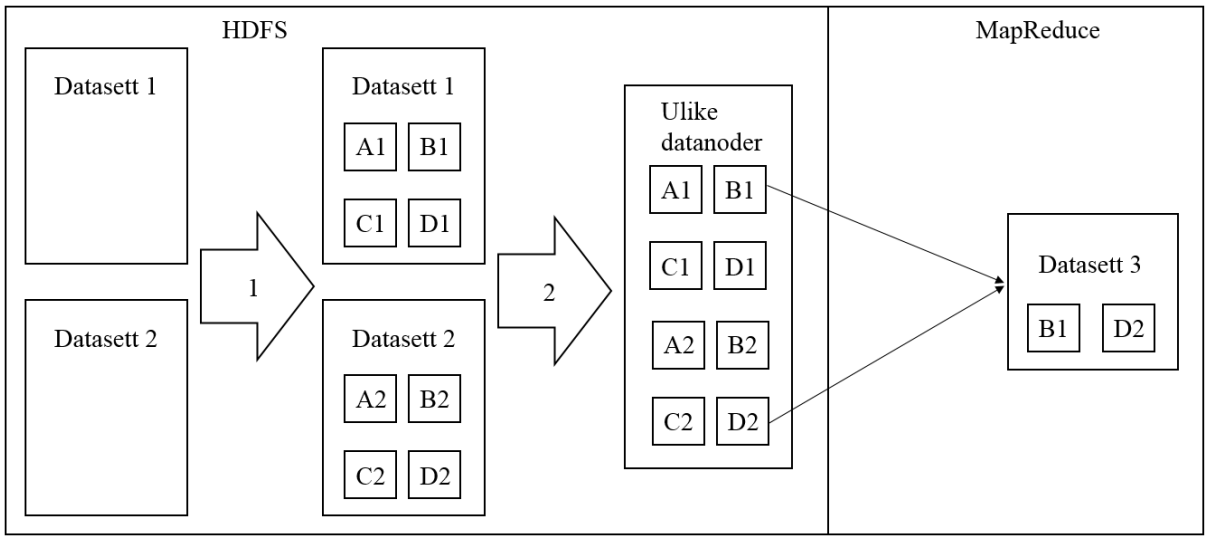
Hvilke komponenter en data lake består av avhenger av hvilken data lake-løsning som er kjøpt inn. Felles for de fleste data lake-løsninger er at de bygger på konseptet Highly Archived Distributed Object Oriented Programming (Hadoop). Hadoop inneholder ulike teknologikonsepter for databehandling ved hjelp av distribuerte datasystemer, og røttene til Hadoop kan spores tilbake til Google sitt distribuerte filsystem [11, 12]. Distribuerte datasystemer er systemer som består av ulike programvarer som er plassert på ulike datamaskiner, og som til sammen utgjør et felles system [3].

Sentralt i en Hadoop-løsning er konseptene Hadoop distributed file system (HDFS) og MapReduce. HDFS er et distribuert fillagringssystem, og det er i HDFS innkommende data til data laken blir lagret. HDFS tar store datasett, splitter datasettene i mindre datasett og lagrer de mindre datasettene på ulike datamaskiner i Hadoop-løsningen. Datamaskinene i Hadoop-løsningen kalles datanoder. For å ha en oversikt over de lagrede dataene blir informasjon om hvilke datanoder dataene er lagret på plassert i en namenode. Informasjonen som lagres i namenode kalles metadata. [13].

Ved analyser er det nødvendig å samle informasjon fra ulike datanoder i HDFS. For å samle informasjonen benytter data lake konseptet MapReduce. I MapReduce gis kommandoer i programmeringsspråket Java for å finne igjen informasjon i HDFS, og MapReduce setter informasjonen fra ulike datanoder sammen til et felles datasett som kan benyttes til analyser [13].

[2] B (byte) er en målenhet for datamengde.

I Figur 3 er et eksempel på HDFS og MapReduce vist.



Figur 3: Eksempel på hvordan HDFS og MapReduce fungerer. Datasett 1 og 2 splittet opp i mindre datasett, og lagret de på ulike datanoder. MapReduce får kommandoer på hvilke data den skal hente, og i tilfellet i figur 3 setter den sammen datasett 3 bestående av B1 og D2. Illustrasjon laget av Tommy Haugen.

Datasett 1 og 2 er store datasett som ankommer HDFS for lagring. I steg 1 vil HDFS splitte datasettene opp i mindre datasett. I Figur 3 er dette vist som at datasett 1 blir delt opp i A1, B1, C1 og D1, og tilsvarende for datasett 2. I steg 2 vil HDFS lagre de mindre datasettene på forskjellige datanoder i fillagringssystemet. MapReduce har i Figur 3 fått kommando om å hente ut datasett B1 og D2, og leverer disse dataene som et eget datasett – datasett 3.

Bedrifter som skal samle inn big data vil ha størst fordel av å gå over til en data lake. I et fillagringssystem som data laken kan data lagres uten å bli tilpasset til et forhåndsdefinert format, og følgelig vil ulempen med datavarehus for big data forsvinne [14]. I tillegg har data lake konseptet MapReduce, som gjør det mulig å kombinere data fra ulike datakilder med ulike dataformat i analyser.

I energibransjen har data lake-løsninger noe popularitet. Ni spurte sentralnettseiere i Europa har fem vurdert å implementere en data lake, og i tillegg har ENTSO-E[3] vurdert å implementere en data lake-løsning (se vedlegg B). Merk at Statnett er inkludert i de ni spurte sentralnettseierne.

[3] ENTSO-E er en organisasjon for sentralnettseiere i Europa [15].

3 Cybersikkerhet

Sikkerhet handler om å sikre verdier mot uønskede hendelser, og tradisjonell informasjonssikkerhet handler om å sikre at korrekt informasjon er tilgjengelig for rettmessig personell til enhver tid. Kjennetegnene til informasjonssikkerhet kan oppsummeres i begrepene konfidensialitet, integritet og tilgjengelighet. Konfidensialitet viser til at kun rettmessig personell skal ha tilgang til informasjon, og integritet handler om å sikre at informasjonen er korrekt. Tilgjengelighet handler om at informasjonen skal være tilgjengelig til enhver tid [16]

Det finnes ulike definisjoner av cybersikkerhet. PST definerer cybersikkerhet som tiltak for beskyttelse mot reelle og potensielle trusler som kanaliserer via IKT-infrastruktur [4].

Med tanke på Statnett sin data lake vil informasjonssikkerhet i cybersikkerhetsperspektiv være det mest sentrale, da det er informasjonsverdier som skal være lagret i data laken. Sammenlignet med tradisjonell informasjonssikkerhet viser cybersikkerhet kun til sikring av verdier i det digitale rom, mens informasjonssikkerhet fanger opp sikring av fysiske verdier [17].

Cyberrisiko fokuserer på trusler som kan nå verdier via IKT-infrastruktur. Cyberrisiko omfatter følgelig ikke risiko som kan utløses av fysiske fenomener som flom eller brann. For cyberrisiko er den potensielle angrepsflaten stor, som gjør cyberrisikobildet uoversiktlig og gjør det vanskelig å kartlegge det totale cyberrisikobildet. I tillegg endres cyberrisikobildet i takt med teknologiutviklingen, som ytterligere vanskeliggjør fullstendig kartlegging av cyberrisikobildet [18].

3.1 Informasjonssikkerhet i Statnett

Knyttet til informasjonssikkerhet er Statnett blant annet underlagt Energiloven og Beredskapsforskriften som setter krav til sikring av informasjonsverdier. I Statnett er informasjonssikkerhet en egen avdeling med organisatorisk tilhørighet under divisjon for informasjons- og kommunikasjonsteknologi (IKT). Avdeling informasjonssikkerhet er en støttefunksjon til resten av Statnett, og hjelper til med saker som angår informasjonssikkerhet [19].

Brukere som trenger tilgang til Statnett sine systemer får automatisk et brukernavn og passord som gir tilganger til systemer og informasjon utfra hvilket behov for tilganger brukeren har. Behovet for tilganger er forankret i brukerens arbeidsbeskrivelse. I Statnett er administreringen av brukernavn og passord til brukere lagt til et identity and access management-system (IAM-system) [9].

3.1.1 Relevante policyer og instrukser

Statnett har forankret informasjonssikkerhet i ulike dokumenter i Statnett. Disse dokumentene vil være relevante for Statnett sin innkjøpte data lake, og ansatte i Statnett er pålagt å sette seg inn i disse. I Tabell 5 er et utvalg av Statnett sine interne dokumenter forklart. Dokumentene i Tabell 5 er en liste over dokument som vil være relevante for data laken, men ikke en fullstendig liste over alle sikkerhetsdokumenter knyttet til informasjonssikkerhet i Statnett. Det er også viktig å merke seg at disse sikkerhetstiltakene ikke er direkte rettet mot cybersikkerhet, men vil virke forsterkende for cybersikkerheten.

Tabell 5: Relevante sikkerhetspolicyer og -instrukser i Statnett som data laken vil være omfanget av. Det er også gitt en kort forklaring på relevant innholdet i de ulike policyene og instruksene.

Dokument i Statnett	**Forklaring**
Instruks for tilgangsstyring [20]	Detaljert instruks for tilgangsstyring, der brukere skal få tilgang til informasjon og IKT-systemer utfra tjenstlig behov.
Instruks for informasjonssikkerhet [21]	Detaljert instruks for informasjonssikkerhet, som blant annet stiller krav til hvordan IKT-utstyr og informasjon skal behandles av brukere.
Instruks for verdivurdering av informasjonsprodukter [22]	Detaljert instruks som forklarer hvordan informasjon skal markeres og verdien av informasjonen skal bestemmes. Verdien skal bestemmes av informasjonseier.
Instruks - Krav til behandling av informasjon [23]	Detaljerte krav til hvordan informasjon skal behandles.
Funksjonspolicy for sikkerhet og beredskap [24]	Detaljert instruks for forebygging og håndtering av uønskede hendelser som kan påvirke forsyningssikkerheten. Herunder krav til sikkerhetsøvelser og etablerte beredskapsløsninger dersom informasjon blir utilgjengelig.
Funksjonspolicy for IKT-styring og digital sikkerhet [25]	Overordnet dokument som stiller krav til sikkerhetsarbeidet i hele organisasjonen.

3.1.2 Soneinndeling

Statnett har et sonesystem for sine IKT-systemer. Soneinndelingen i Statnett definerer hvilken sikkerhetsgradering systemet skal ha, og sikkerhetsgraderingen er stigende fra 0 til 4. Sikkerhetssonene gjelder også informasjon som blir generert av systemene i de ulike sonene. Sikkerhetssonene og en kort forklaring på disse er gitt i Tabell 6 [26].

Tabell 6: Sikkerhetsonene i Statnett. Det er også gitt en forklaring på relevante systemer og informasjon som omfattes av sikkerhetssonen. Sikkerhetsgraderingen er økende fra 0 til 4 [26].

Sone	Forklaring
0	Sone som er utenfor Statnett sin kontroll, eksempelvis internett.
1	Sone for datautveksling med eksterne brukere. E-post er eksempel på system som ligger i denne sonen.
2	Sone for administrative systemer i Statnett. Det er her alle Statnett ansatte jobber til vanlig. Statnett sin data lake er plassert i denne sonen.
3	Sone for intern datautveksling. Her utveksles data mellom sone 2 og 4.
4	Den sikreste sonen, hvor driftsstøttesystemer er plassert. SCADA er eksempel på system som ligger i denne sonen, og er én av kildene til data laken.

3.2 Sikkerhet big dataplattform

For big dataplattformen er sikkerhet tenkt som en overordnet sone for sonene kilder, innsamling, distribusjon/analyse/lagring og tilgjengeliggjøring. Konsumenter/visualisering er ikke påvirket av de sikkerhetskomponentene som Statnett har kjøpt inn med big dataplattformen [9].

For hele big dataplattformen skal IBM Big SQL brukes som en overordnet komponent for å gi tilganger innad i hele big dataplattformen, og IBM Governance Catalog skal logge hvem som har gjort hvilke endringer i dataene. Det skal også benyttes brannmur for å beskytte hele dataplattformen [9].

Innkjøpet av den nye dataplattformen til Statnett omfattet også enkelte sikkerhetskomponenter som er direkte rettet mot data laken. De innkjøpte sikkerhetskomponentene er forklart ytterligere i Tabell 7 [9].

Tabell 7: Sikkerhetskomponentene som tilhører data laken. Det er også gitt en forklaring på hva sikkerhetskomponentene gjør [9]. Forklaringene er ikke hentet fra [9].

Sikkerhetskomponent	Forklaring
Kerberos	Gir brukere tilgang til selve dataplattformen utfra hvilke rettigheter de har [27].
Ranger	Gir brukere tilgang til ulike deler av dataplattformen avhengig av hvilke rettigheter de har [28].
Knox	Gir brukere tilgang til dataene ved å validere at tilgangene Kerberos og Ranger har gitt er korrekt. Valideringen skal foregå opp mot Statnett sitt IAM-systemet [29].
Ambari	Komponent som overvåker hele Hadoop-teknologien. I Ambari kan det leses hvordan aktiviteten i Hadoop-teknologien er [30].

I Finbeck-rapporten kommenteres det at sikring av sensorer lokalt er et usikkerhetsmoment, men at kryptering[4] av dataene ved oversendelse fra kilde til data laken er tilfredsstillende. Finbeck-rapporten kommenterer også at det kan settes opp kryptering av data i ro og i bevegelse i data laken. Med data i ro menes data som er lagret, og med bevegelse menes data som beveger seg fra et punkt til et annet [9].

3.3 Automatisk diagnostisering (AutoDig)

Automatisk diagnostisering (AutoDig) er et prosjekt i avdeling Feilanalyse i Statnett som vil ta i bruk Statnett sin big dataplattform våren 2018. Prosjektet omfatter den innkjøpte data laken og en applikasjon for analyse, diagnose og prediksjon. AutoDig har som mål å legge til rette for proaktiv forebygging ved hjelp av effektiv bruk av big data med moderne dataverktøy. Prosjektet vil hente data fra data laken og *Innsikt* [8].

I tilknytning til AutoDig-prosjektet ble det gjennomført en risiko- og sårbarhetsanalyse (ROS-analyse) i 2016 for daglig drift og bruk av den ferdige løsningen for AutoDig. I ROS-analysen er det gjennomført en verdivurdering, identifikasjon av risikoer, analyse av risikoer og tiltaksplanlegging. I analyse av risiko er det sett på sannsynlighet og konsekvens av risikoene, og det er sett på hvilken av sikkerhetsparameterne konfidensialitet, integritet og tilgjengelighet risikoene påvirker [31].

[4] Metode for å sikre at data ikke kan leses av andre enn de som har tilgang til nøkler for å lese dataene (krypteringsnøkler). Bygger ofte på matematiske modeller [49].

Hverken de identifiserte risikoene eller de foreslåtte sikkerhetstiltakene fra ROS-analysen til AutoDig er direkte rettet mot cyberperspektivet.

De identifiserte risikoene, foreslåtte sikkerhetstiltakene og sammenstillingen av disse fra ROS-analysen til AutoDig er gjengitt i vedlegg C.

4 Risikovurdering og sikkerhetstiltak

Det er viktig å merke seg at både cybersikkerhet og data lake er begreper uten entydige definisjoner. I dette kapittelet er det brukt samme begrep i teksten som i litteraturen det henvises til.

4.1 Risikovurdering

4.1.1 Verdier

Som systemansvarlig i det norske kraftsystemet er Statnett ansvarlig for å drifte omlag 11 000 kilometer med høyspentledninger og 150 stasjoner [1]. Oppdraget til Statnett innebærer å sikre den norske strømforsyningen, og for å gjennomføre sitt samfunnsoppdrag samler Statnett inn detaljert informasjon om kraftsystemet. Deler av informasjonen som Statnett samler inn vil bli lagret i Statnett sin innkjøpte data lake [9].

Ettersom AutoDig vil være konsument av data laken vil verdien som er bestemt for AutoDig også være relevant for den totale data laken.

En gjengivelse av verdiene som Statnett har analysert at vil være i data laken er gjengitt i vedlegg D.

4.1.2 Cybertrusler

PST sine trusselvurderinger fra 2015 frem til og med 2018 trekker frem kritisk infrastruktur som et mål for blant annet sabotasje eller etterretning [32, 33, 34, 6]. Med kritisk infrastruktur forstås blant annet energiforsyningen, og dermed Statnett sine oppgaver og informasjon. For 2018 trekker både PST og Etterretningstjenesten (E-tjenesten) frem Russland og Kina som statlige trusselaktører mot Norge [6, 35]. NSM støtter denne vurderingen, og påpeker at også ikke-statlige aktører kan være aktuelle trusler for kritisk infrastruktur i Norge [36].

Trusselaktører kan deles inn i interne og eksterne. Interne trusler er ansatte, innleide eller leverandører som har tilgang til systemer/informasjon. De interne truslene kan utføre ondsinnede eller ikke villede handlinger som kan medføre skade på informasjonsverdiene [36].

Eksterne trusler kan for eksempel være statlige og ikke-statlige aktører som forsøker å få seg tilgang til systemer/informasjon de ikke har eller skal ha tilgang til. Eksterne trusler kan bli interne ved at de rekrutterer innsidere. Tendensen med rekruttering av innsidere forventes at øker i 2018 [36].

4.1.3 Cyberangrep

I 2016 gjennomførte Næringslivets Sikkerhetsråd (NSR) en undersøkelse blant 1 500 norske bedrifter med tanke på informasjonssikkerhet og datakriminalitet kalt Mørketallsundersøkelsen. Mørketallsundersøkelsen viser at 27 % av de spurte bedriftene har opplevd uønskede sikkerhetshendelser [37]. Som en oppfølging av Mørketallsundersøkelsen gjennomførte NVE en tilsvarende undersøkelse for bedrifter i energibransjen, og NVE sin rapport viser at nærmere 70 % av de spurte bedriftene har hatt uønskede sikkerhetshendelser. NVE fremhever bedrageri, virus og/eller malwareinfeksjon, forsøk på datainnbrudd/hacking, dataskadeverk og hendelse forårsaket av bedriftens ansatte som de mest sentrale uønskede sikkerhetshendelsene [38]. Bak tallene fra både Mørketallsundersøkelsen og NVE sin oppfølging antas det at det skjules store mørketall [37,38].

Cyberangrep kan foregå på ulike måter, og enkelte angrepsformer går igjen. NSM påpeker at trusselaktører forsøker å få tilgang til systemer de ikke skal ha tilgang til ved eksempelvis å benytte seg av brukeres innloggingsdetaljer [36]. For å få tak i en brukers innloggingsdetaljer er to angrepsformer vanlige: spydfiske og vannhullangrep. I forkant av spydfiske kartlegger trusselaktøren offentlig kjent informasjon om bedriftens ansatte. Deretter sender trusselaktøren målrettede e-poster som inneholder ondsinnede koblinger til den ansatte. Når den ansatte åpner de ondsinnede koblingene vil trusselaktøren kunne klare å få tilgang til innloggingsdetaljene til brukeren. Vannhullangrep er en lignende angrepsform, men her lures den ansatte til å klikke på en ondsinnet kobling på et nettsted som den ansatte ofte besøker. Etter å ha skaffet seg innloggingsdetaljer har trusselaktøren tilgang til systemer og informasjon på lik linje med en betrodd ansatt [38].

Tjenestenektangrep er en annen vanlig cyberangrepsform i energibransjen. I et tjenestenektangrep forsøker trusselaktører å sette digitale systemer ut av spill ved å eksempelvis sende store mengder nettverkstrafikk mot det digitale systemet. Den store nettverkstrafikken gjør at systemet kollapser [38].

Julen 2015 ble Ukraina utsatt for et cyberangrep, som begynte med et spydfiskeangrep våren 2015. Etter å ha kartlagt systemet i nærmere et halvt år gjennomførte angriperne fysiske utkoblinger i kraftsystemet julen 2015. I forbindelse med angrepet ble det også gjennomført et tjenestenektangrep mot telefondisken til nettselskaper som ble berørt av angrepet for å hindre strømkunder i å ringe inn avbrudd [5].

Spredning av virus er også en angrepsform som også forekommer i energibransjen. NVE viser til angrep rettet mot ulike selskaper innen energibransjen, der oppdatering i et regnskapsprogram som selskapene benyttet førte til at maskiner og systemer ble gjort utilgjengelige [38].

Målet og motivet for et cyberangrep kan være ulikt. Etterretningstjenesten sin trusselvurdering for 2018 fremhever at angrep rettet mot kritisk infrastruktur har som mål å samle inn sensitiv informasjon som senere kan benyttes til sabotasje [35]. Nasjonal sikkerhetsmyndighet (NSM) støtter Etterretningstjenesten sin vurdering om at spionasje er en betydelig risiko mot kritisk infrastruktur i en digital tid [36]. I angrepene mot Ukraina observeres det at angrepet først kartla ulike deler av kraftsystemet, og deretter brukte informasjonen til å gjennomføre fysisk sabotasje i kraftsystemet [5].

4.1.4 Sårbarheter

NSM fremhever ulike sårbarheter for digitale systemer i kritiske samfunnsfunksjoner. Uten noen plan for sikkerhetsarbeidet kan sårbarheter forbli uten å bli oppdaget og tettet. Videre påpeker NSM at lange digitale verdikjeder er sårbare, da hvert steg i verdikjeden er avhengig av det forrige steget. Cyberangrep kan rettes mot steder i den digitale verdikjeden som er tilstøtende til hovedmålet, som betyr at sårbarheter som er knyttet til hele verdikjeden for dataene er viktige å kartlegge. NSM påpeker videre at ny teknologi kan ha manglende sikkerhetsmekanismer, og at de mange komponentene som til sammen utgjør det digitale systemet kan gjøre det vanskelig med en helhetlig sikkerhetstilnærming. Utfordringer med sikkerhet i teknologiske komponenter fremheves også med tanke på datakilder, og NSM fremhever IoT-komponenter spesielt. IoT-komponenter er nye i forbindelse med datainnhenting, og kan ha manglende eller lite tilfredsstillende sikkerhetsmekanismer. NSM fremhever også viktigheten av at teknologiutviklingen må sees i sammenheng med sårbarheter, og at sårbarhetsreduserende tiltak må iverksettes tidlig. NSM fremhever også mennesket som en sårbarhet, da det kan utnyttes [36].

Cloud Security Alliance (CSA) har gjennomført intervju av medlemmer i tilknytning til cyberangrep og datalekkasje for big datainfrastrukturer. CSA fremhever at de store datamengdene kan utgjøre en sårbarhet, da det kan være vanskelig å se hvem som har gjort hva med dataene. CSA påpeker også at de store datamengdene kan gjøre det vanskelig å oppdage feil eller mangler ved informasjonen. For nye datakilder poengterer CSA at de kan ha manglende sikkerhetsmekanismer, og for den digitale infrastrukturen påpeker CSA at sikkerhetsmekanismene kan være tillagt infrastrukturen i ettertid. CSA fremhever også at den digitale infrastrukturen kan ha problemer med å legge til rette for at brukere kun har tilgang til nødvendig informasjon, ettersom big datainfrastrukturer generelt har utfordringer med detaljert tilgangsstyring. Hvordan administreringen av så detaljert tilgangsstyring skal gjøres i praksis fremheves også som en utfordring. I tillegg poengterer CSA at overvåking av en big datainfrastruktur kan være vanskelig, da omfanget av infrastrukturen sammen med de store datamengdene kan gjøre infrastrukturen uoversiktlig [39].

European union agency for network and information security (ENISA) har gjennomført analyser knyttet til informasjonssikkerhet og big data for blant annet kraftforsyningen. ENISA fremhever at i big datainfrastrukturer er det potensielt mange brukere og komponenter som skal ha tilgang til systemet på likt, og med et distribuert system kan dette medføre utfordringer med tanke på å holde oversikt. For big datainfrastrukturen kommenterer ENISA at sikkerheten burde være forankret i behovsanalysen til selve teknologien. ENISA kommenterer at de ulike kildene kan medføre en sårbarhet til systemet, da de befinner seg ute og er utsatt for ytre elementer som bedriften nødvendigvis ikke har kontroll over. Sensorer trekkes spesielt frem som en utfordring med tanke på dette. ENISA kommenterer videre at big datainfrastrukturen består av ulike komponenter, og hevder at dette kan øke sannsynligheten for en sikkerhetsbrist. Med tanke på teknologien kommenterer også ENISA at infrastrukturen som samler inn dataene kan få problemer med å håndtere de store datamengdene, og feil kan være vanskelig å oppdage [40].

I big dataperspektiv har National institute of standards and technologies (NIST) kommet frem til syv forskjeller mellom sikring av tradisjonelle data og big data. NIST fremhever at big datainfrastrukturer inneholder komponenter som ikke nødvendigvis er ment til å kommunisere med hverandre, og som fordrer sikkerhet tenkt på komponentnivå. Videre påpeker NIST at innhenting av informasjon fra nye datakilder kan ha manglende sikkerhetsmekanismer, og at de relevante sikkerhetsmekanismene er for lite brukervennlig. Med lite brukervennlig fremhever NIST at sikkerhetsmekanismene kan redusere big datainfrastrukturens evne til å håndtere de store datamengdene. I tillegg fremhever NIST at den totale datamengden kan være skadelig for datakvaliteten, da det er vanskeligheter med å ha kontroll på hvilke data som hentes inn og brukes. NIST påpeker også at mengden data kan gjøre at data bevisst eller ubevisst endres eller feiltolkes av brukere. Verdisettingen av big data er også utfordrende, ettersom verdien av de enkelte datasettene må sees i sammenheng med alle dataene i big datainfrastrukturen. To datasett som hver for seg ikke avslører noe sensitiv informasjon kan gjøre det dersom de settes sammen, og på den måten er ikke verdien av de enkelte datasettene det mest sentrale. NIST fremhever også viktigheten av å se sikkerheten i lys av dataenes levetid. Med stadig teknologiutvikling kan sikkerheten bli svekket med årene [41].

Flere av sårbarhetene er nevnt i flere av kildene. I Tabell 8 er sårbarhetene samlet til overordnede punkter og gitt en forklaring. Tabell 8 har også henvisning til hvilken kilde sårbarheten er funnet. Det er viktig å merke seg at sårbarhetene ikke nødvendigvis er ordrett nevnt i kilden de er funnet, men at kombinasjonen av de ulike kildene gir den nevnte forklaringen. Sårbarhetene er satt sammen, da de supplerer hverandre.

Tabell 8: Identifiserte sårbarheter fra litteraturstudiet. Det er også gitt en forklaring på sårbarheten, og henvisning til hvilken kilde de er funnet. Det gjøres oppmerksom på at sårbarheten ikke nødvendigvis er ordrett funnet i samme kilde.

Sårbarhet	**Forklaring**	**Kilde**
Datakilde	Manglende eller lite tilfredsstillende sikkerhetsmekanismer hos nye datakilder.	[36,39,40,41]
Mennesker	Manglende forståelse for den digitale infrastrukturen, dens innhold og sikkerhet. Mennesker kan også feiltolke og lekke informasjon.	[36]
Big datainfrastrukturer	Teknologikomponentene som utgjør den digitale infrastrukturen kan ha manglende sikkerhetsmekanismer, og mangfoldet av teknologikomponenter kan gjøre det vanskelig med en helhetlig sikkerhetstanke. Vanskelig med overvåking.	[39,41,36]
Administrering	Manglende sikkerhetsrutiner og oppfølging av sikkerhetsarbeidet er en sårbarhet.	[36]
Datamengde	Datamengdene som skal håndteres er store, og dette kan gjøre det vanskelig å oppdage feil eller mangler. Dataene kan også ha utfordringer med å bruke tradisjonelle sikkerhetsmetoder. Vanskelig med overvåking.	[39,41]

4.1.5 Analyse cyberrisiko

På bakgrunn av verdiene, cybertruslene, mulige cyberangrep og sårbarheter kan det analyseres enkelte cyberrisikoer en operasjonell data lake står ovenfor. Informasjonen som er lagret i data laken er av slik karakter at den kan antas å være et attraktivt mål for trusselaktører. Det antas også at trusselaktørene Statnett står ovenfor har potensiale og motivasjon til å gjennomføre cyberangrep.

Data fra kildene er råstoffet i data laken, og danner fundamentet for analyser og beslutninger. Med dårlig sikring av kildene kan angripere sende manipulerte data til data laken. De manipulerte dataene kan videre bli benyttet i analyser, og dermed bidra til at gale beslutninger fattes. Videre kan de store datamengdene gjøre det vanskelig å oppdage feil eller mangler ved dataene [39,41].

Mennesker har direkte tilgang til data og systemer med sine innloggingsdetaljer. Dersom innloggingsdetaljene havner uvedkommende i hende kan potensielt uvedkommende få tak i samme informasjon som en betrodd ansatt. Innloggingsdetaljer kan bli stjålet av trusselaktører ved hjelp av spydfiske eller vannhullangrep [38]. Med innloggingsdetaljene kan en angriper samle informasjon om den digitale infrastrukturen, som kan brukes til sabotasje på et senere tidspunkt [35,36,5]. Mennesket kan også bevisst eller ubevisst slette eller feiltolke data, eller lekke informasjon dersom de er en innsider [36].

Big datainfrastrukturen gjør det vanskelig med overvåking [39]. Utfordringen med overvåking gjelder fordi big datainfrastrukturen inneholder mange ulike teknologikomponenter, og spesielt kildene svært ulike. Omfanget av big datainfrastrukturen kan gjøre det vanskelig å avdekke om det er feil eller mangler ved dataene, uvanlig dataaktivitet eller en systemsvikt er på vei. Sammen med manglende sikkerhetsrutiner og oppfølging av sikkerhetsarbeidet, kan sårbarbarheter forbli uoppdaget [36].

På bakgrunn av ovennevnte analyse kan punktene oppsummeres i fem overordnede cyberrisikoer. De fem overordnede cyberrisikoene er forklart nærmere i Tabell 9. Tabell 9 viser også hvilke sårbarheter som kan utløse cyberrisikoen.

Tabell 9: Cyberrisikoer som er funnet gjennom en risikovurdering. Det er også vist hvilken sårbarhet som er mest relevant for cyberrisikoen, og det er en forklaring på hva som menes med cyberrisikoen.

Cyberrisiko	**Forklaring**	**Sårbarhet**
Innhenting av manipulerte data	Innhenting av manipulerte data kan føre til at gale beslutninger fattes.	Datakilder, administrering.
Bruk av manipulerte data	Bruk av manipulerte data kan føre til at gale beslutninger fattes.	Big datainfrastrukturer, datamengde og mennesker
Tap av sensitive data	Tap av sensitive data gjør at uvedkommende kan få tilgang til dem.	Datamengde og mennesker
Feil brukertilgang	Feil bruker med feil tilgang gjør at feil bruker kan gjøre ting med dataene som den ikke har tillatelse til.	Datamengde, big datainfrastrukturer og mennesker
Systemsvikt	Systemsvikt kan gjøre at beslutninger fattes på for sent tidspunkt.	Datakilder, big datainfrastrukturer og administrering.

Innhenting og bruk av manipulerte data påvirker dataenes integritet, da de ikke er korrekte. Innhenting av manipulerte data, brukertilgang på avveie og systemsvikt påvirker alle tre sikkerhetsparameteren tilgjengelighet. Fra kildene kan en cybertrussel gjennomføre tjenestenektangrep, som kan sette hele systemet ut av spill. Tap av kritisk informasjon gjør at informasjonen ikke lenger er tilgjengelig til rettmessige brukere, og brukertilgang på avveie kan legge til rette for eksempelvis tjenestenektangrep. Systemsvikt påvirker naturlig tilgjengeligheten til dataene. For konfidensialitet er tap av kritisk informasjon og brukertilganger på avveie de relevante cyberrisikoene.

4.2 Sikkerhetstiltak

4.2.1 Data lake

Search Technologies (eid av Accenture) påpeker fire hovedområder for sikring av en data lake: plattformtilgang, nettverksisolering, databeskyttelse og datatilgang. Plattformtilgang handler om å sikre at riktige brukere har tilgang til riktig del av data laken, og riktige funksjoner for å utføre sine arbeidsoppgaver. Lese- eller redigeringstilganger er eksempel på ulike funksjoner en bruker kan ha tilgang til. Konkrete sikkerhetstiltak for å imøtekomme dette er gode og detaljerte rutiner for tilgangsstyring. Nettverksisolering handler om sikring av den fysiske data laken med eksempelvis brannmur. For databeskyttelse fremhever Search Technologies kryptering av data i ro og i bevegelse som relevant sikkerhetstiltak. Datatilgang skal gis i samsvar med hvilke data brukere skal ha tilgang til [42].

Deloitte fremhever de samme sikkerhetstiltakene som Search Technologies, og legger til viktigheten av overvåking og gjenoppretting. Tilgangsstyringen mener Deloitte burde være helt ned på filnivå, og følge bedriften sine eksisterende rutiner på området. Teknologikomponenter som er relevante for tilgangsstyring er Ranger, Knox og Kerberos. Kerberos burde samkjøres med eksempelvis overordnede systemer for tilgangsstyring. For databeskyttelse mener Deloitte at kryptering av data i bevegelse og i ro er viktig. Videre fremhever Deloitte overvåking av data laken med komponenten Ambari som et sikkerhetstiltak. Overvåkingen må være slik at personell kan finne ut hvem som har gjort hvilke endringer. Analysene i tilknytning til overvåkningen burde gjennomføres av kompetent personell på et tidlig tidspunkt. Hva som menes med kompetent personell eller tidlig tidspunkt er ikke videre definert. Mest mulig av aktiviteten i data laken burde fanges opp av overvåkingen. Avslutningsvis kommenterer Deloitte at dersom uønskede sikkerhetshendelser oppstår er det viktig med rutiner for gjenoppretting for å oppnå normal drift. For at gjenopprettingen skal være mest mulig vellykket burde data være markert slik den viktigste informasjonen blir gjenopprettet først [43].

Hadoop

Det er forsket noe på generell sikkerhet knyttet til Hadoop-teknologi. En gjennomgang av ulike sikkerhetsaspekter knyttet til Hadoop publisert i International Journal of Computer Science and Information Technologies (IJCSIT) støtter Deloitte og Search Technologies sine sikkerhetstiltak, og understreker at disse sikkerhetstiltakene må implementeres på hver enkelt teknologikomponent som utgjør Hadoop-teknologien [44].

En analyse gjennomført med tanke på sikkerheten knyttet til HDFS publisert i International Research Journal of Engineering and Technology (IRJET) konkluderer med at autentisering[5], autoriseringen[6] og kryptering er sentrale sikkerhetstiltak for å sikre informasjonen som lagres i HDFS. Artikkelen i IRJET fremhever at krypteringen burde gjennomføres på filnivå [45].

For å redusere muligheten for feil brukertilgang i HDFS, foreslår en artikkel ulike metoder for tilgangsstyring som kan brukes for å sikre at rettmessige brukere har rettmessig tilgang. Motivasjonen er å hindre at data blir manipulert eller slettet. Analysen konkluderer blant annet med at Kerberos er én metode for å sikre tilgangsstyring [46].

En annen artikkel viser hvordan databeskyttelsen kan forsterkes i Hadoop-teknologi, og fokuserer på hvordan autentisering og kryptering kan bedre sikkerheten for dataene i Hadoop. Artikkelen konkluderer med at kryptering, autentisering og autorisering er gode hjelpemiddel for å sikre data i Hadoop [47].

Kryptering nevnes også av en annen artikkel. Denne artikkelen konkluderer med at kryptering i Hadoop kan gjøre prosesseringen av data noe mindre effektiv, men fremhever at kryptering likevel er et sentralt sikkerhetstiltak for Hadoop-teknologi [48].

[5] Autentisering viser til prosessen hvor en bruker må bevise dens identitet for å få tilgang til et system [49].
[6] Autorisasjon viser til prosessen hvor en bruker må benytte brukernavn og passord for å få tilgang til konkrete data [49].

4.2.2 Big datainfrastrukturer

ENISA fremhever syv sikkerhetstiltak som sentrale knyttet til informasjonssikkerhet og big data for blant annet kraftforsyningen [40]:

- Kryptering av data som er i ro og i bevegelse.
- Jevnlig sikkerhetstesting av systemer og komponenter. Med systemer forstås hele big datainfrastrukturen.
- Bruk av sikkerhetssertifiserte teknologikomponenter i big datainfrastrukturen, og sikkerhetsstandarder i overordnet sikkerhetsarbeid.
- Kildefiltrering for å sikre at informasjonen som hentes inn er korrekt.
- Tilgangskontroll og autentisering for alt som er tilkoblet dataplattformen, inkludert IoT-komponenter.
- Overvåking og logging av endringer i systemer og informasjon.

ENISA konkluderer med at jevnlig sikkerhetstesting er det mest sentrale sikkerhetstiltaket. Tilgangskontroll og autentisering, samt bruk av sikkerhetssertifiserte teknologikomponenter følger som de neste sentrale sikkerhetstiltakene [40].

Tilgangskontroll for alt som er koblet til big dataplattformen blir trukket frem av en annen artikkel også. Artikkelen fremhever at alle dataene som skal samles inn i big dataplattformen må kartlegges, og tilganger til dataene må gis utfra hvilket behov brukeren til enhver tid har. Tilgangsstyringen er følgelig å forstå som en kontinuerlig jobb. I tillegg må dataene klassifiseres utfra sensitivitet, merkes med hvor lenge de skal være lagret og hvor i den big datainfrastrukturen de skal være plassert. Overvåking kan samordnes med allerede eksisterende overvåkingsmekanismer i bedriften. Artikkelen fremhever også at data burde beskyttes med eksempelvis kryptering, og at nøklene for å få lest dataene må gis til rettmessig personell [50].

4.2.3 Digitale systemer i norsk kritisk infrastruktur

Mørketallsundersøkelsen gjennomført i 2016 anbefaler sikkerhetskopiering, styrking av medarbeideres sikkerhetsbevissthet, deteksjon av uønskede sikkerhetshendelser og systematisk sikkerhetsarbeid som viktige sikkerhetstiltak. Sikkerhetskopiering trekkes frem, da utestengelse fra IKT-systemer er en stadig økende sikkerhetsutfordring. Mørketallsundersøkelsen trekker konkret frem god sikkerhetsopplæring av personell som et sentralt sikkerhetstiltak, og rutiner for å oppdage og håndtere uønskede sikkerhetshendelser. Rutiner for oppdaging av uønskede sikkerhetshendelser kan være overvåking, og håndtering handler om å følge opp de alarmene som fremkommer under overvåkingen. Mørketallsundersøkelsen fremhever også at sikkerhetsarbeidet burde følge standarder, og at uønskede sikkerhetshendelser blir fulgt opp og registrert [37].

For å kunne reagere på sikkerhetshendelser må de avdekkes. I NVE sin oppfølging av Mørketallsundersøkelsen rapporteres det at nærmere 27 % av de spurte bedriftene oppdaget en uønsket sikkerhetshendelse ved en tilfeldighet. 42 % av de spurte bedriftene visste ikke hva som var den medvirkende årsaken til at hendelsen oppstod. NVE fremhever at sikkerhetsoppdateringer må gjennomføres så fort de dukker opp, og rutiner for autorisasjon og sikkerhetsovervåkning er sentrale sikkerhetstiltak for beskyttelse av innloggingsdetaljer. For sikkerhetsovervåking trekker NVE også frem overvåking av barrierene som skal sikre verdiene, og jevnlige sikkerhetstester for å kartlegge om sikkerhetsbarrierene og overvåkingen fungerer. NVE fremhever også viktigheten av å lære av disse. For å øke læringsutbytte fra den uønskede sikkerhetshendelser fremhever NVE at årsaken til den uønskede sikkerhetshendelsen burde avdekkes. NVE nevner også at det burde bygges god sikkerhetskultur innad i bedriften, og at bevisstgjøring er et tiltak som kan være med å bidra til dette. I tillegg anbefaler NVE at bedrifter innfører styringssystem for informasjonssikkerhet, og at hendelser rapporteres til sentralt organ for sikkerhet [38].

NSM fremhever at risikoreduserende tiltak handler om kontinuerlig sikkerhetsarbeid, styrket bevissthet innen sikkerhet blant ansatte og oppdaging og håndtering av uønskede sikkerhetshendelser knyttet til IKT som sentrale sikkerhetstiltak. NSM fremhever spesielt at sikkerhet og teknologiutvikling må foregå i takt [36].

4.2.4 Oppsummering sikkerhetstiltak

Flere av de fremhevede sikkerhetstiltakene går igjen i ulik litteratur. I Tabell 10 er lignende sikkerhetstiltak satt sammen, da det vil oppleves repeterende å nevne dem flere ganger. Tabell 10 viser også hvilken kilde tiltaket er funnet i. Det er viktig å merke seg at tiltakene nødvendigvis ikke er nevnt ordrett i samme kilde, men at sikkerhetstiltakene er så pass lignende at de kun trenger å bli nevnt en gang. Sikkerhetstiltakene definert i tabell 13 vil bli vist til sikkerhetstiltak funnet i litteratur videre.

Tabell 10: Identifiserte sikkerhetstiltak på bakgrunn av litteraturstudiet med forklaring. Det er også vist til hvilken kilde sikkerhetstiltaket er funnet. Merk at sikkerhetstiltaket ikke nødvendigvis er ordrett funnet i den oppgitt kilden.

Navn	**Sikkerhetstiltak fra litteraturstudiet**	**Kilde**
Detaljert tilgangsstyring	Detaljert tilgangsstyring som følger brukeres arbeidsoppgaver.	[38,40,42,43,50 44,45,46,47]
Robust kilde- og dataverifisering	Kontinuerlig gjennomgang av kilder og innkommende data.	[40,50]
Dataverdsettelse	Klassifisering av dataene.	[50]
Databeskyttelse	Beskytte data som er i ro og i bevegelse. Nøkler for å låse opp beskyttelsen må holdes adskilt fra dataene.	[40,42,43,44,45, 47,48,50]
Overvåking	Overvåking og logging av endringer i system og informasjon i hele big datainfrastrukturen.	[36,37,38,40 43,44,50]
Kontinuerlig sikkerhetsarbeid	Jevnlige sikkerhetsoppdateringer og –tester av komponenter og system. Rapportering av uønskede sikkerhetshendelser.	[36,37,38,40]
Teknologi	Bruk av sikkerhetssertifiserte komponenter i teknologiarkitekturen, og bruk av brannmur.	[40,42,43]
Opplæring	Sikkerhetsopplæring av alt personell.	[36,37,38]
Beredskap	Plan for håndtering av uønskede sikkerhetshendelser, og bedre evne til å reagere på hendelser. Backupløsninger for sentral informasjon og IKT.	[36,37,38,43]

5 Diskusjon og analyse

5.1 Cyberrisiko

For å se om de identifiserte cyberrisikoene fra risikovurderingen har noen relevans for en operasjonell data lake, er cyberrisikoene fra risikovurderingen sammenlignet med risikoene fra ROS-analysen til AutoDig. I denne diskusjonen er risikoene fra ROS-analysen til AutoDig sett i et cyberperspektiv. Med det menes at det antas at truslene kan komme gjennom sikringer i digitale komponenter. Diskusjonen viser til cyberrisikoene fra risikovurderingen som cyberrisikoer og risikoene fra ROS-analysen til AutoDig som risiko.

Se vedlegg E for en detaljert diskusjon av risikoene fra ROS-analysen opp mot cyberrisikoene fra risikovurderingen.

Som følge av diskusjonen i vedlegg E kan det tolkes en sterk likhet mellom cyberrisikoene funnet i litteratur og de risikoene som er definert i ROS-analysen til AutoDig, og samtlige risikoer fra ROS-analysen til AutoDig kan aggregeres til de overordnede cyberrisikoene. At cyberrisikoene kan omfatte risikoene fra ROS-analysen kan komme av at cyberrisikoene er overordnede og brede, og mange underordnede risikoer kan plasseres innunder disse. Likevel er det rimelig å påpeke at også risikoene fra ROS-analysen til AutoDig er overordnede og åpenbare, og forklaringene for risikoene i ROS-analysen eller som ligger bak cyberrisikoene bryter ikke med hverandre.

Risikovurderingen gjennomført i denne masteroppgaven er annerledes gjennomført enn ROS-analysen for AutoDig. Risikovurderingen ser på angrepsformer som kan utføres ved å utnytte sårbarheter for å nå verdier. I ROS-analysen til AutoDig blir ikke angrepsformen eller trusselaktøren kommentert. I diskusjonen i vedlegg E er derfor risikoene fra ROS-analysen sett i et cyberperspektiv ved at det er lagt til grunn at angrepsformene fra risikovurderingen i denne masteroppgaven kan være årsak til at risikoene i ROS-analysen utløses. Denne antakelsen beror på at cybertruslene Statnett står ovenfor har nok kapasitet og ressurser til å gjennomføre cyberangrep. For andre enn Statnett er det viktig å se risikovurderingen i lys av trusselbildet de står ovenfor.

Basert på denne argumentasjonen kan cyberrisikoene fra risikovurderingen fremheves som relevante med tanke på en operasjonell data lake. For liste over cyberrisikoene, se kapittel 4.1.5 tabell 9.

5.2 Sikkerhetstiltak

For å diskutere de identifiserte sikkerhetstiltakene fra litteraturstudiet er de sammenlignet med sikkerhetstiltak definert i Statnett sine dokumenter, samt ROS-analysen til AutoDig. Det er også sett på hvilke sikkerhetskomponenter og –systemer i Statnett som vil være relevante med tanke på sikkerhetstiltakene.

Se vedlegg F for en detaljert sammenligning av sikkerhetstiltakene fra Statnett sine dokumenter, sikkerhetskomponenter og –systemer samt ROS-analysen til AutoDig.

Basert på diskusjonen i vedlegg F kan det sees klare likheter mellom de foreslåtte sikkerhetstiltakene fra litteraturstudiet og ROS-analysen til AutoDig. Likhetstrekkene kan forklares av at sikkerhetstiltakene funnet i litteratur er overordnede, og på den måten kan forsvare mange underliggende sikkerhetstiltak. Likevel er det rimelig å påpeke at samtlige sikkerhetstiltak funnet i ROS-analysen kan aggregeres til de overordnede sikkerhetstiltakene, og bryter ikke med underliggende argumenter fra litteraturstudiet. ROS-analysen kan heller tenkes at tilfører sikkerhetstiltakene detaljer som er relevante for Statnett, som er logisk med tanke på at ROS-analysen er gjennomført av Statnett.

Diskusjonen i vedlegg F viser også klare likheter mellom Statnett sine overordnede dokumenter og sikkerhetskomponenter og –systemer. Likhetstrekkene kan komme av at resultatene fra litteraturstudiet og dokumentene til Statnett er overordnede og generelle, og kan tolkes som beste praktisering av sikkerhet. For sikkerhetstiltakene databeskyttelse, dataverdsettelse og kontinuerlig sikkerhetsarbeid er det viktig å merke seg at ROS-analysen til AutoDig ikke har noen tilsvarende sikkerhetstiltak, som betyr at det kun er i overordnede dokumenter tiltakene står definert. Det er likevel antatt at disse er relevante også med tanke på data laken, da de overordnede dokumentene skal fungere også på den.

Basert på denne argumentasjonen kan sikkerhetstiltakene i litteraturstudiet fremheves som relevante med tanke på cybersikring av en operasjonell data lake. For liste over sikkerhetstiltakene, se kapittel 4.2.4 tabell 10.

5.3 Sammenstilling

For å se at sikkerhetstiltakene fra litteraturstudiet har relevans for cyberrisikoene fra risikovurderingen er det gjennomført en sammenstilling av sikkerhetstiltakene og cyberrisikoene. For å gjennomføre sammenstillingen har det vært nødvendig å gjennomføre en avansert analyse for å beholde logikken fra ROS-analysen til AutoDig. Hovedstegene i sammenstillingen er forklart i Tabell 11. Et eksempel for å illustrere fremgangsmåten er forklart i vedlegg G.

Tabell 11: Forklaring på hvordan sammenstillingen er gjennomført.

Steg	Forklaring
1	Alle risikoene fra ROS-analysen er sammenfallende med cyberrisikoene fra risikovurderingen. Risikoene fra ROS-analysen som underbygger cyberrisikoene er funnet igjen tabell 1 i vedlegg E.
2	Risikoene fra ROS-analysen har egne sikkerhetstiltak definert fra ROS-analysen. I tabell 3 fra vedlegg C er de foreslåtte sikkerhetstiltakene funnet igjen.
3	Diskusjonen av sikkerhetstiltak har aggregert alle de foreslåtte sikkerhetstiltakene fra ROS-analysen til overordnede sikkerhetstiltak funnet i litteratur. Sikkerhetstiltaket fra steg 2 er derfor funnet igjen som forklaring på overordnede sikkerhetstiltak i tabell 1 og/eller tabell 2 i vedlegg F.
4	Det overordnede sikkerhetstiltaket fra steg 3 er sammenstilt med cyberrisikoen som var utgangspunktet i steg 1.

Resultatet fra sammenstillingen er presentert i Tabell 12. Kryssene indikerer hvilket sikkerhetstiltak som er analyser at vil fungere på hvilken cyberrisiko.

Tabell 12: Sammenstilling av sikkerhetstiltak og cyberrisikoer fra litteraturstudiet og risikovurderingen. Sammenstillingen er basert på ROS-analysen til AutoDig [31]. Kryssene indikerer hvilke sikkerhetstiltak den identifiserte cyberrisikoen fra litteraturstudiet er tenkt at vil påvirke.

Sikkerhetstiltak fra litteratur	**Cyberrisiko**				
	Innhenting av manipulerte data	**Bruk av manipulerte data**	**Tap av sensitive data**	**Feil brukertilgang**	**Systemsvikt**
Detaljert tilgangsstyring	X		X	X	X
Robust kilde- og dataverifisering	X		X		X
Dataverdsettelse					
Databeskyttelse					
Overvåking	X				X
Kontinuerlig sikkerhetsarbeid					
Teknologi					X
Opplæring	X	X			
Beredskap					X

I Tabell 12 sees det at de identifiserte sikkerhetstiltakene fra litteraturstudiet har ulik grad av relevans for cyberrisikoene fra litteraturstudiet. Det er viktig å merke seg at ingen av de identifiserte sikkerhetstiltakene har noen definert styrke, som betyr at selv om detaljert tilgangsstyring og robust kilde- og dataverifisering påvirker størst antall cyberrisikoer, er de ikke nødvendigvis mest virkningsfulle.

I sammenstillingen er dataverdsettelse, databeskyttelse og kontinuerlig sikkerhetsarbeid tre sikkerhetstiltak uten definert påvirkning med tanke på cyberrisikoer. Dette kommer av at disse sikkerhetstiltakene ikke har noen tilsvarende tiltak definert i ROS-analysen til AutoDig, og at de kun fanges opp av overordnede dokumenter i Statnett. Denne sammenstillingen er kun basert på ROS-analysen til AutoDig, så hvordan disse tre sikkerhetstiltakene vil fungere på de ulike cyberrisikoene er ikke videre analysert.

Sammenstillingen av sikkerhetstiltakene fra litteraturstudiet og cyberrisikoene kan følgelig indikere at det er en sammenheng mellom enkelte av disse.

5.4 Praktisering av sikkerhetstiltakene

I risikovurderingen ble det kommentert ulike aspekter ved sikkerhetstiltakene som kan være utfordrende ved praktiseringen av dem.

Detaljert tilgangsstyring fremheves som utfordrende, grunnet de store datamengdene og den uoversiktlige big datainfrastrukturen [39,40]. Av samme grunn er også robust kilde- og informasjonsverifisering fremhevet som utfordrende [40]. Det kommenteres også at det kan bli ressurskrevende å drifte en big dataplattform med detaljert tilgangsstyring [39].

Dataverdsettelse fremheves som utfordrende, da verdien av dataene må bestemmes av dataenes totale verdi [41]. Dette kan gå på bekostning av data lakens ønskede fleksibilitet, ved at informasjon er strengere sikret enn hva den behøver å være.

I Statnett skal verdien av dataene bestemmes av informasjonseier, og for data laken er verdien av informasjonen allerede bestemt gjennom AutoDig-prosjektet [25,31]. Databeskyttelse fremheves også som utfordrende, da det kan redusere tiden det tar for data å bli prosessert [48].

For å kunne reagere på sikkerhetshendelser må de avdekkes. NVE sine analyser av informasjonssikkerhetstilstanden i kraftforsyningen viser at avdekking av uønskede sikkerhetshendelser er utfordrende [38]. Avdekking av uønskede hendelser gjør det også vanskelig å forebygge uønskede cyberhendelser, spesielt da omfanget av cyberrisiko er omfattende [18]. For overvåking av en big datainfrastruktur fremheves det spesielt at det kan være vanskelig å til enhver tid vite hvem som har tilgang til hva, ettersom big datainfrastrukturen er omfattende [40].

De praktiske utfordringene knyttet dreier seg hovedsakelig om at data lake består av ulike teknologikomponenter i et distribuert system, og at data laken skal behandle big data. Disse to punktene i kombinasjon gjør data laken uoversiktlig, og det gjør det spesielt vanskelig å avdekke uønsket aktivitet og endringer i systemet.

For sikkerhetstiltakene detaljert tilgangsstyring, robust kilde- og dataverifisering, dataverdsettelse, databeskyttelse, overvåking, teknologi og beredskap identifisert i litteraturstudiet kan denne diskusjonen indikere at sikkerhetstiltakene har utfordringer med å bli praktisert.

5.5 Utfordringer Statnett

Det er poengtert at cyberrisikoene fra risikovurderingen og sikkerhetstiltakene fra litteraturstudiet er av relevans for cybersikring av en operasjonell data lake. I diskusjonene sammenlignes resultatene fra risikovurderingen og litteraturstudiet hovedsakelig med en ROS-analysen gjennomført av Statnett. Enkelte av de fremhevede sikkerhetstiltakene fra litteraturstudiet er også nevnt i Statnett sine overordnede dokumenter. For Statnett betyr dette at de allerede er klar over cyberrisikoene, og at de har de relevante sikkerhetstiltakene for å håndtere cyberrisikobildet til den operasjonelle data laken.

Litteratur fremhever ulike praktiske utfordringer knyttet til praktisering av de enkelte av sikkerhetstiltakene. Ettersom Statnett allerede praktiserer disse sikkerhetstiltakene er det antatt at sikkerhetstiltakene er innført i den grad det er mulig å praktisere dem.

For Statnett sin del kan likevel enkelte utfordringer som strekker seg utover sikkerhetstiltakene observeres. Fra litteraturstudiet omfatter sikkerhetstiltaket litteratur opplæring innen sikkerhetsarbeid. I Statnett er sikkerhetsopplæringen plassert i overordnede dokumenter, og opplæringen som nevnes i ROS-analysen til AutoDig er ikke eksplisitt fremhevet at er rettet mot sikkerhet. Sikkerhetstiltakene innen opplæring ble i diskusjonen rundt cyberrisiko (se vedlegg E) hovedsakelig plassert innunder sikkerhetstiltaket opplæring, men det er rimelig å påpeke at opplæringen som poengteres i sikkerhetstiltaket fra litteraturstudiet er konkret rettet mot sikkerhet.

I risikovurderingen er tap av sensitive data og feil brukertilgang fremhevet som cyberrisikoer med mennesket som sårbarhet, men fra sammenstillingen i kapittel 5.3 er ikke opplæring et relevant sikkerhetstiltak for disse cyberrisikoene. Dette forsterker tanken om at opplæringen fra ROS-analysen er mer rettet mot opplæring innen arbeidsoppgaver, og ikke i like stor grad rettet mot sikkerhet. Opplæringen innen sikkerhet er dermed å oppfatte som forankret i overordnede dokumenter i Statnett, og ikke direkte rettet mot data laken.

For Statnett sin operasjonelle data lake er organiseringen av sikkerhetsarbeidet også rimelig å kommentere. Finbeck-prosjektet er drivende for hele big dataplattformen inkludert data laken, og i rapporten fra Finbeck-prosjektet er sikkerhet kommentert [9]. Likevel er hovedfokus i rapporten å tolke som bygging av big dataplattformen, der data laken kun er en del av den totale plattformen. ROS-analysen som omfatter data laken er gjennomført for AutoDig-prosjektet – et prosjekt som er i tilknytning til big dataplattformen [31]. Denne organiseringen av

sikkerhetsarbeidet kan indikere at helheten i sikkerhetsarbeidet knyttet til data laken kan ha mangler, og det virker som at teknologiutviklingen og sikkerhetsarbeidet for data laken ikke har utviklet seg i takt. NSM fremhever viktigheten av at nettopp teknologiutviklingen og sikkerheten utvikles i takt [36]. Grunnen til at teknologiutviklingen og sikkerhet utvikles i utakt kan komme av at informasjonssikkerhet er en støttefunksjon til resten av Statnett. Likevel skal data laken utvikles frem til 2020, så det er fortsatt tid til å kombinere teknologiutviklingen med sikkerhet.

At sikkerhetsopplæring er plassert i overordnede dokumenter, og opplæring innen arbeidsoppgaver er plassert i AutoDig-prosjektet, gjør at disse to blir adskilt på en måte som kan være uheldig for data laken. Ettersom mennesket er i daglig kontakt med dataene og data laken, er det rimelig å fremheve mennesket som en ressurs i sikkerhetsarbeid og ikke bare en sårbarhet. NVE fremhever sikkerhetskultur i bedriften som et relevant sikkerhetstiltak [38].

5.6 Avsluttende betraktninger

I risikovurderingen og litteraturstudiet er det benyttet et kilder med ulik kvalitet, som er generell og overordnet. Kildene retter seg enkelte ganger ikke direkte mot kraftforsyningen, og heller ikke mot data lake. Cyberrisikoene fra risikovurderingen og sikkerhetstiltakene fra litteraturstudiet kan dermed i beste fall anses for å være overordnede cyberrisikoer og sikkerhetstiltak som er av relevans for en operasjonell data lake. For å møte den varierende kvaliteten er cyberrisikoene og sikkerhetstiltakene aggregert for å se på helheten heller enn å se på detaljene i litteraturen, og sammenligningen mellom funnene fra risikovurderingen og litteraturstudiet er gjennomført på en kritisk måte.

Ved å sammenligne funnene fra risikovurderingen og litteraturstudiet kan alle cyberrisikoene og sikkerhetstiltakene verifiseres. Rett nok beror verifiseringen på antakelsen om at risikoene fra ROS-analysen kan sees i et cyberperspektiv, ved at cybertrusler kan bryte de sikringene som ROS-analysen fremhever som mangelfulle. Det er rimelig å anta at mangelfull sikring av digitale komponenter kan brytes av cybertrusler sett i lys av hvor ressurssterke truslene Statnett står ovenfor er. Det er også viktig å merke seg tolkningen av risikoene og sikkerhetstiltakene fra ROS-analysen til AutoDig, da de er sentrale for aggregeringen. For andre enn Statnett er

det viktig å se detaljene i risikovurderingen gjennomført i denne oppgaven spesielt i lys av hvilket trusselbilde de står ovenfor og hvilken bransje de operer innen.

I diskusjonene er funnene fra risikovurderingen og litteraturstudiet sammenlignet i hovedsak opp mot én ROS-analysen gjennomført av Statnett. Hvilke komponenter en data lake består av avhenger av hva slags data lake som er kjøpt inn. I litteraturstudiet og risikovurderingen er HDFS, kilder og big datainfrastrukturer trukket frem, som ikke er å anse som spesielle for Statnett sin data lake. Resultatene er overordnet, og kan tenkes at er relevante for flere data laker. Det er likevel rimelig å påpeke at variasjoner grunnet ulikheter i data laken vil kunne forekomme.

Det er også viktig å fremheve at sikkerhetsarbeidet er en kontinuerlig prosess [36]. Sikkerhetsarbeidet er en kontinuerlig prosess, og resultatene fra denne masteroppgaven må dermed ikke leses som en absolutt vurdering av cyberrisiko med tilhørende sikkerhetstiltak for data lakens totale levetid. Teknologiutviklingen må sees i sammenheng med sårbarhetene, men det er vanskelig med en presis kartlegging av det totale risikobildet i et cyberperspektiv [36,18].

For Statnett kan sammenligningen mellom funnene i denne masteroppgaven og ROS-analysen til AutoDig også vise at Statnett har oversikt over cyberrisikoene de står ovenfor, og praktiserer de sikkerhetstiltakene som behøves for å redusere cyberrisikoene. For de overordnede cyberrisikoene fra risikovurderingen og sikkerhetstiltakene fra litteraturstudiet kan detaljene fra ROS-analysen til AutoDig være med å tilføre detaljer til funnene, og sette dem i sterkere virksomhetskontekst for Statnett.

6 Konklusjon og anbefaling til videre arbeid

6.1 Konklusjon

Denne masteroppgaven har tatt utgangspunkt i problemstillingen: Hvilke hovedpunkter burde Statnett fokusere på med tanke på cybersikring av en operasjonell data lake?

I denne masteroppgaven konkluderes det med at Statnett SF burde fokusere på å lære opp brukere av data laken i cybersikkerhet. Opplæringen kan gjøres med eksempelvis jevnlig kursing og bevisstgjøringskampanjer. Masteroppgaven konkluderer også med at Statnett SF burde fokusere på å samordne teknologiutvikling og cybersikkerhet for data laken sterkere, og se til at disse utvikles i takt.

Målet med denne masteroppgaven var å undersøke hva litteratur fremhever som viktig med tanke på cybersikring av en operasjonell data lake. Gjennom arbeidet med masteroppgaven er det identifisert lite litteratur som retter seg konkret mot cybersikring og data lake. Det er derfor analysert litteratur som i hovedsak retter seg mot cybersikkerhet og big data, samt generell cybersikkerhet i kritisk infrastruktur. Resultatene er overordnede, men kan verifiseres som relevante ved hjelp av en ROS-analyse gjennomført av Statnett. Denne masteroppgaven har klart å presentere en total risikovurdering, og identifisere sikkerhetstiltak som er relevante med tanke på cybersikring av en operasjonell data lake. I tillegg har oppgaven klart å identifisere hovedpunkter som Statnett burde fokusere på med tanke på cybersikring av en operasjonell data lake.

6.2 Anbefaling til videre arbeid

På generelt grunnlag anbefales det å forske mer på cybersikkerhet og data lake. For Statnett er det i forbindelse med denne oppgaven kommet frem til seks relevante forskningspartnere. De relevante forskningspartnerne er: ELES, Swissgrid, Elia system operator, Transelectrica og Amprion (se vedlegg B).

Arbeidet med sikkerhet er en kontinuerlig prosess, så det anbefales Statnett å opprettholde og videreutvikle dagens sikkerhetsmiljø. Det anbefales også å gjennomføre ytterligere risikovurdering for Statnett sin data lake.

Referanser

[1] Statnett. *Om Statnett.* Tilgjengelig på: http://www.statnett.no/Om-Statnett/ [lest: 17.01.18].

[2] Statnett. *Konsernstrategi 2017-2021.* [upublisert dokument, Statnett].

[3] Lahke, B. (2016). *Practical Hadoop Migration (s.151-188).* USA: Apress.

[4] Koordineringsgruppen for IKT-risikobildet (2010). *Bakgrunnsnotat Cybersikkerhet 2010-06-01.* Tilgjengelig på: https://www.regjeringen.no/contentassets/252f869fdfac46648e41e6ca5fb0600a/cybersikkerhet_svar-med-merknader_nsm-pst-etterretningstjenesten.pdf [lest: 10.05.18].

[5] Lee, M. R., Assante, J. M., Conway, T. (2016). *Analysis of the Cyber Attack on the Ukrainian Power Grid*. Washington: Electricity Information Sharing and Analysis Center (E-ISAC).

[6] Politiets sikkerhetstjeneste (PST). *Trusselvurdering 2018.* Tilgjengelig på: https://www.pst.no/trusselvurdering-2018/ [lest: 03.02.18].

[7] Nasjonal sikkerhetsmyndighet (2016). *Håndbok: Risikovurdering for sikring.* Tilgjengelig på: https://www.nsm.stat.no/globalassets/dokumenter/handboker/risikovurdering_nsm_handbok_mars2016.pdf [lest: 04.02.18].

[8] H., Mohanty et al. (eds.) (2015). *Big Data, Studies in Big Data (s. 1-7).* India: Springer.

[9] Schjelderup, K., Andersen, M. L. (2018) *Finbeck, Road-map for IKT-arkitektur, fremtidig analyseplattform. Sluttrapport fase 1.* [upublisert dokument, Statnett].

[10] Statnett. *AutoDig, Fremtidens analyse- og diagnoseplattform.* [upublisert dokument, Statnett].

[11] Apache Hadoop. *What is Apache Hadoop?* Tilgjengelig på: http://hadoop.apache.org/ [Lest: 05.02.18].

[12] Ghemawat, S. et al. (2003). *The Google File System.* USA: Google.

[13] Lahke, B. (2016) *Practical Hadoop Security (s. 19-35).* USA: Apress.

[14] TechTarget Network. *Data lake governance: A big data do or die.* Tilgjengelig på: https://searchcio.techtarget.com/feature/Data-lake-governance-A-big-data-do-or-die [lest: 20.01.18].

[15] ENTSO-E. *Who is ENTSO-E?* Tilgjengelig på: https://www.entsoe.eu/about/inside-entsoe/objectives/ [lest: 31.04.18].

[16] Stamp, M. (2011). *Information Security, principles and practice, 2. ed. (s. 1-14)*. New Jersey: Wiley.

[17] Niekerk, van J., Rossouw, von S. (2013) *From information security to cyber security.* Elsevier Ltd.

[18] Refsdal. A. et al. (2015). *Cyber-Risk Management (s.25-47).* Springer Cham Heidelberg New York Dordrecht London.

[19] Statnett. *Organisasjonsoversikt.* [upublisert dokument, Statnett].

[20] Statnett (godkjent 04.04.17). *Instruks for tilgangsstyring.* [upublisert dokument, Statnett].

[21] Statnett (godkjent 20.02.18). *Instruks for informasjonssikkerhet.* [upublisert dokument, Statnett].

[22] Statnett (godkjent 12.06.17). *Instruks for verdivurdering av informasjonsprodukter.* [upublisert dokument, Statnett].

[23] Statnett (godkjent 12.06.17). *Instruks – Krav til behandling av informasjon.* [upublisert dokument, Statnett].

[24] Statnett (godkjent 12.01.18). *Funksjonspolicy for sikkerhet og beredskap.* [upublisert dokument, Statnett].

[25] Statnett (godkjent 11.04.18). *Funksjonspolicy IKT-styring og Digital Sikkerhet*. [upublisert dokument, Statnett].

[26] Statnett (2013). *Sonemodell, vedtatt sonemodell 2013*. [upublisert dokument, Statnett].

[27] TechTarget Network. Kerberos. Tilgjengelig på: https://searchsecurity.techtarget.com/definition/Kerberos [lest: 22.01.18].

[28] Hortonworks. *Apache Ranger.* Tilgjengelig på: https://hortonworks.com/apache/ranger/ [lest: 22.01.18].

[29] Hortonworks. *Apache Knox Gateway.* Tilgjengelig på: https://hortonworks.com/apache/knox-gateway/ [lest: 22.01.18].

[30] Hortonworks. *Apache Ambari.* Tilgjengelig på: https://hortonworks.com/apache/ambari/ [lest: 22.01.18].

[31] Lervik, R. (2016). *Rapport fra ROS-analyse av informasjonssikkerhet, versjon 1.0. ROS-analyse: Nytt AutoDig.* [upublisert dokument, Statnett].

[32] Politiets sikkerhetstjeneste (2015). *Trusselvurdering 2015.* Tilgjengelig på: https://www.pst.no/alle-artikler/trusselvurderinger/trusselvurdering-2015/ [lest: 10.02.18].

[33] Politiets sikkerhetstjeneste (2016). *Trusselvurdering 2016.* Tilgjengelig på: https://www.pst.no/alle-artikler/trusselvurderinger/trusselvurdering-2016/ [lest: 10.02.18].

[34] Politiets sikkerhetstjeneste (2017). *Trusselvurdering 2017.* Tilgjengelig på: https://www.pst.no/trusselvurdering-2017/ [lest: 10.02.18].

[35] Etterretningstjenesten (2018). *Fokus 2018. Etterretningstjenestens vurdering av aktuelle sikkerhetsutfordringer.*

[36] Nasjonal Sikkerhetsmyndighet (2017). *Risiko 2017. Risiko og sårbarheter i en ny tid. En vurdering av sårbarheter og risiko i Norge.* Sandvika: Nasjonal Sikkerhetsmyndighet.

[37] Næringslivets sikkerhetsråd (2016). *Mørketallsundersøkelsen 2016.* Oslo: Næringslivets sikkerhetsråd.

[38] Norges vassdrags- og energidirektorat (2017). *Informasjonssikkerhetstilstanden i energiforsyningen.* Oslo: Norges vassdrags- og energidirektorat.

[39] Cloud Security Alliance (2012). *Expanded Top Ten Big Data Security and Privacy Challenges.*

[40] European Union Agency For Network And Information Security (2015). *Big Data Security. God Practices and Recommendations on the Security of Big Data Systems.*

[41] National Institute of Standards and Technologies (2015). *NIST Big Data Interoperability Framework Volume 4, Security and Privacy.*

[42] Search Technologies. *Data lake security.* Tilgjengelig på: https://www.searchtechnologies.com/blog/data-lake-security [lest: 19.01.18].

[43] Deloitte Consulting GmbH (2017). *Five key principles to secure the enterprise Big Data platform.*

[44] Sharma, P.P., Chandrakant, P. N. (2014). *Securing Big Data Hadoop: A review of Security Issues, Threats and Solution.* International Journal of Computer Science and Information Technologies.

[45] Inamdar, Y. S. et al. (2016). *Data Security in Hadoop Distributed File System.* Studentrapport. International Research Journal of Engineering and Technology.

[46] Saraladevi, B. et al. (2015). *Big Data and Hadoop – A study in Security Perspective.* Elsevier B.V.

[47] Patil, N.,V., (2015). *Securing Hadoop using OAuth 2.0 and Real Time Encryption Algorithm.* India: International Journal on Recent and Innovation Trends in Computing and Communication.

[48] Park, J.J. et al. (eds) (2013). *Secure Hadoop with Encrypted HDFS.* Berlin: Springer-Verlag Berlin Heidelberg.

[49] BU TechWeb. *Understanding Autentication, Authorisation and Encryption.* Tilgjengelig på: https://www.bu.edu/tech/about/security-resources/bestpractice/auth/ [lest: 04.05.18].

[50] Tankard, C. (2012). *Big data security.* Elsevier Ltd.

Vedlegg A – Utsendte spørsmål

I dette vedlegget er spørsmålene som ble sendt ut til europeiske sentralnettseiere (TSOer) gjengitt. Avkrysningsspørsmål er markert med en A og tekstsvar er markert med en T. Spørsmål markert med * var obligatorisk å svare på.

Information to the person answering: *This survey is conducted as part of a master thesis regarding cyber security and data lakes. For any questions regarding the survery or thesis, please contact Tommy Haugen (tommy.haugen@statnett.no). Please note the definition of a data lake before answering the questions: A data lake is a digital infrastructure for storing and prosessing big data. The data lake is often build on Hadoop-technology.*

Question 1 (A): What are you representing?*

TSO	
Other	

Question 2 (T): Please note the name of what you are representing.

Question 3 (A): Have you considered implementing a data lake?*

Yes	
No	

Information to the person answering: *If you have considered implementing a data lake, we would appreciate your answers on the following questions. Every answer will be treated confidentially.*

Question 4 (A): If you have implemented a data lake: Have you had any cyber security issues using the data lake?

Yes	
No	

Question 5 (A): If you have rejected implementing a data lake: Did you reject the data lake because of cyber security issues?

Yes	
No	

Question 6 (T): If you have had cyber security issues using a data lake, please note what kind of security issues you have had.

Question 7 (T): If you have rejected implementing a data lake because of cyber security issues, please note what kind of cyber security issues that contributed to the decision.

Question 8 (T): Are you interested in cooperation with Statnett SF regarding research on cyber security issues in data lakes? Please note your contact information.

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway