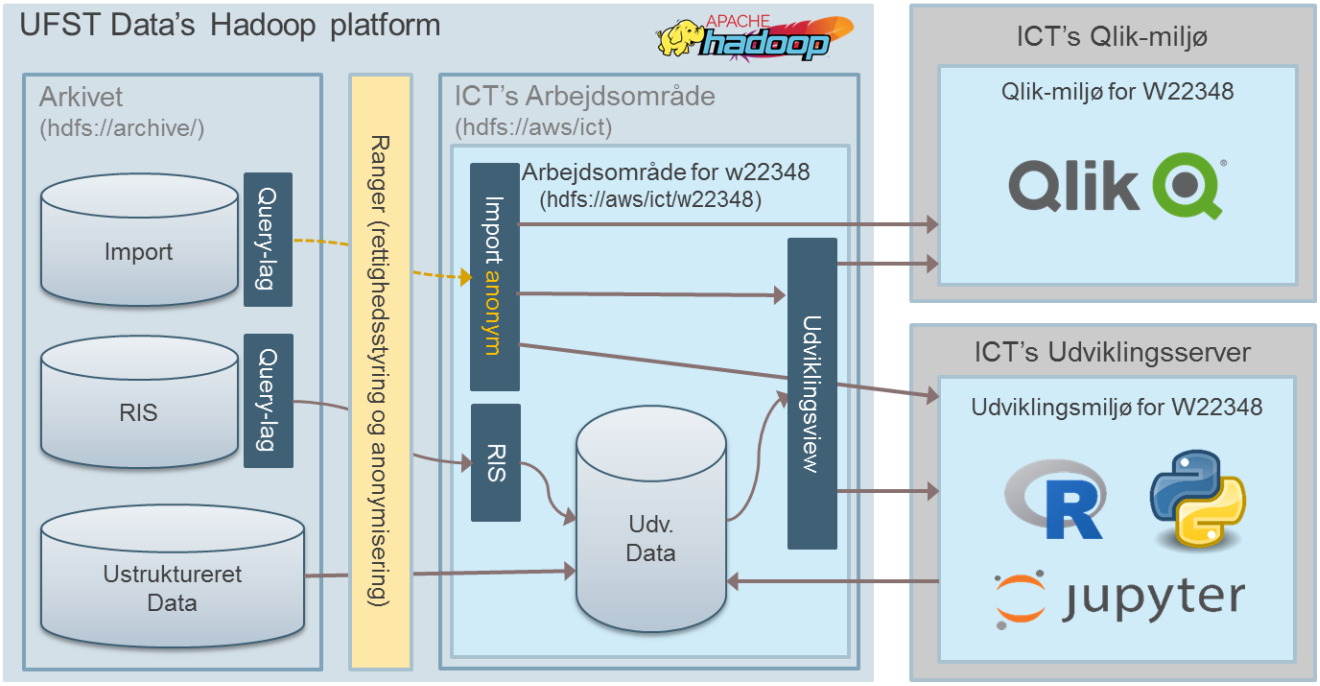
Indhold

Dokument status

Aftalt d.	Aftalt af:	Review d.	Review:
	LGA/TOD		

Beskrivelse af ICT's Analytiske Arbejdsområde



Figur 1 Oversigt over løsningsarkitekturen for ICT's arbejdsområde i Databanken som den er beskrevet i dette dokument.

Teknisk setup med Hadoop og Hive

Der skal oprettes et 'Analytisk arbejdsområde'(AWS) for ICT i Databanken. I praksis er dette en mappe i HDFS, som for eksemplets skyld antages at have følgende sti: 'hdfs:/aws/ict/'. Dette arbejdsområde vil ligge i databankens produktionsmiljø. Når det drejer sig om udviklingsarbejde ift. eksplorative analyser, udvikling af machine learning algoritmer og business intelligence, så foregår det i Produktionsmiljø. De resterende miljøer i Databanken (Udvikling, Test, Pre-prodroduktion) bruges til udvikling af selve Databanken og implementering af ingest/hjemtagelse af data mv.

Arbejdsområder

Alle ICT's analyse-medarbejdere der skal arbejde med data, får sin egen mappe i ICT's aws. Eksempelvis har medarbejderen med w-nummer w12345 sit eget arbejdsområde i ICT's aws, i form af mappen 'hdfs:/aws/ict/w12345'. Medarbejderen er den eneste, foruden administratorer, der har rettighed til at tilgå denne mappe, dvs. det er et isoleret miljø (alá G-drevet på PC) hvor medarbejderen har fuld læse/skrive rettighed. Der vil være en Hive database tilknyttet dette arbejdsområde, som medarbejderen også har læse/skrive/oprette rettigheder til. I arbejdsområdet kan hver enkelt medarbejder manipulere med data efter behov, eksempelvis:

- Oprettelse af tabeller i Hive baseret på filer som er lagt op i arbejdsområdet

- Oprettelse af Views i Hive. Det kan være views baseret på tabeller fra eget arbejdsområde/database, views baseret på data fra arkivet, som er stillet til rådighed for medarbejderen via BRAS, eller views der blander egne data med data fra arkivet.
- Upload af data som eksempelvis er generet ifm. udviklingsarbejdet. Det kan være 'mellem-regninger' i form af data-transformationer som gemmes midlertidigt, eller som skal samkøres med andre data i arbejdsområde.

Der er tilknyttet en BRAS/AD-gruppe til hvert enkelt medarbejders område, dvs. én BRAS-adgang som tildeles én bruger. BRAS-adgangene for hvert enkelt arbejdsområde konfigureres i Ranger af UFST Data. Medarbejdere i ICT, som har et udviklingsmiljø på ICT's udviklingsserver, kan tilgå arbejdsområdet i Databanken fra deres udviklingsmiljø og dermed hente/gemme data fra udviklingsmiljøet. Udviklingsmiljøet samt Ambari bliver de primære måder at tilgå Hive/HDFS med.

Arbejdsområder Udestående:

- *Samspillet mellem Qlik og arbejdsområderne.*
- *Mulighed for 'fælles arbejdsområder' hvor eksempelvis et projekt bestående af flere medarbejdere har adgang.*

Arkivet

I arkivet udstilles data til ICT's medarbejdere via HDFS/Hive. Herfra kan data benyttes i Arbejdsområdet og ICT's udviklingsmiljø. Udstillingen rettighedsstyres via BRAS og der tildeles som minimum BRAS-grupper på system-niveau, *hvilket konfigureres i Ranger af UFST Data*. Dette er den mest grovmaskede rettighedsinddeling, men Ranger kan sagtens definere adgange med højere granularitet hvis behovet opstår. Disse behov kan opstå løbende og når dette sker, skal rettighederne oprettes i BRAS, defineres/konfigureres i Ranger og tildeles de relevante brugere.

Arkivet Udestående:

- *Skal der være en SLA for behandlingstiden ift. at få en BRAS-rettighed opfyldt?*
- *Der skal være en proces-beskrivelse for tildeling af rettigheder samt oprettelse af nye, mere granulerede, rettigheder.*

Værktøjer

I ovenstående tekniske løsning på Hadoop-platformen er det tiltænkt at der skal være følgende teknologiske værktøjer til rådighed for medarbejderne:

- HDFS: Distribueret filsystem opbevaring af data
- Hive: Kan bruges til at afvikle SQL-queries data fra arkivet samt egne data i medarbejdernes arbejdsområder.
- Ambari: En browser-baseret grafisk brugergrænseflade. Via Ambari kan udstillede data fra arkivet og arbejdsområdet tilgås via HDFS/Hive.

- (Py)Spark: Det vil være muligt at eksekvere distribuerede beregninger i Databanken vha. Spark. Dette kan ske ved at give adgang for ICT's medarbejde til en af Hadoop clusterens Edge-noder.

Værktøjer Udestående:

- *Der mangler afklaring ift. hvordan PySpark kan bruges til at eksekvere store parallelle jobs. Kan dette eksekveres fra udviklingsmiljøet, eksempelvis? Der kan være brug for en PoT her.*
- *Skal vi have SLA på mængde af hardware ressourcer vi vil have?*

Sikkerhed

Brugerrettigheder

Rettigheder administreres primært i to komponenter:

- **AD/BRAS**: Her defineres de forskellige rolle-typer, eksempelvis:
  a) En rolle per arbejdsområde per medarbejder.
  b) En rolle per data-adgang, eksempelvis en rolle der giver adgang til Import-systemets data i arkivet.
  c) Special-designede roller kan også oprettes og gives. Eksempelvis en rettighed på tabel X, kolonne y, z.
  d) En rolle til at få adgang til Databankens Hadoop edge-node som giver mulighed for at afvikle Spark-jobs på clusteren.

  AD/BRAS roller skal oprettes af UFST data (hvis de ikke allerede er, som f.eks. roller til arbejdsområderne) mens tildelingen til den/de enkelte medarbejdere skal ske af ICT's BRAS-ansvarlige med godkendelse af KC eller anden ansvarshavende.

- **Apache Ranger**: Det er i Ranger at en given AD-gruppe konfigureres ift. bruger-adgang til arbejdsområder samt adgang til arkivet i Databanken. I Ranger kan det angives hvilke mapper (og eventuelle undermapper) som en given gruppe kan have adgang til i HDFS. Ligeledes kan AD-grupper sammenkobles med adgang til specifikke tabeller + kolonner + rækker i Hive. *Konfigurationen af AD-grupper i Ranger foretages af UFST Data.*

Brugerrettigheder udestående:

- *Der er behov for en SLA omkring behandlingstider for tildeling af disse adgange i Ranger.*
- *Der er behov for en SLA og procesbeskrivelse for oprettelse af special-designede roller.*

Anonymisering

Den nuværende løsning på anonymisering bliver også via Ranger. Udover angivelse af kolonner/tabeller er det også muligt at angive en bruger-defineret funktion (UDF) som alle data skal behandles af, før de præsenteres for en given bruger. Dette giver mulighed for at implementere anonymisering on-the-fly, så det undgås at data skal kopieres ifm. anonymiseringsprocessen. Det er tiltænkt at relevante tabeller/kolonner og anonymiseringsteknikker skal gemmes i en separat tabel, som så kan opdateres og konfigureres

løbende efterhånden som sensitive oplysninger bliver afdækket. *Denne opdatering og konfiguration skal kunne foretages af visse ICT-medarbejdere* som har de nødvendige rettigheder, mens *opsætningen i Ranger foretages af UFST Data*.

Angivelsen af en UDF til anonymisering giver også muligt for at have to roller til hver data-adgang: Den ikke-anonymiserede og den anonymiserede. Det er hensigten at de fleste medarbejdere kun har adgang til den anonymiserede data-adgang.

Anonymisering Udestående:

- *Der er behov for en SLA ift. behandlingstiden på disse konfigurationsopgaver.*
- *Der er behov for at tryk teste denne løsning via en PoT for at se performance samt få fuld forståelse for løsningens omfang og muligheder.*

Governance og Processer ifm. ICT's arbejdsområde

Som skitseret ovenfor, er der en klar rolle- og rettighedsfordeling i forhold til ICT's analytiske arbejdsområde:

- Hver enkelt ICT-medarbejder har fuld råderet over sit arbejdsområde (læse, skrive, oprette).
- Hver enkelt ICT-medarbejders data-adgange til Arkivet tildeles via BRAS (med den dertilhørende proces for godkendelser der hører til).
- UFST Data står for at oprette arbejdsområderne.
- UFST Data står for at lave AD-grupper/BRAS-rettigheder og konfigurerer disse i Ranger. AD-grupper på specifikke data kan ICT bestille hos UFST Data.
- UFST Data konfigurerer anonymiseringsfunktioner i Ranger, men ICT udvikler algoritmerne og konfigurerer hvilke kolonner der skal anonymiseres og hvordan.

Ansvarsfordelingen ift. den skitserede løsningsbeskrivelse for arbejdsområderne er følgende:

Ansvarsområde	**Ansvarshaver**
Drift af Hadoop platformen, inkl. de nævnte værktøjer og ICT's analytiske arbejdsområde.	UFST Data
Arkivet, inkl. ingest og udstilling	UFST Data
Data i ICT's arbejdsområde, inkl. udstilling i forbindelse med udvikling.	ICT
ICT's udviklingsmiljø v. UFST Drift- og Udviklingscenter	ICT

Governance og Processer ifm. ICT's arbejdsområde udestående:

- Hvor går grænsen mellem hvad ICT og UFST Data skal gøre og har ansvar for?
- Hvordan bliver processen for idriftsættelse af anonymisering?
- Hvordan bliver processen for at gemme onetime-dumps af data i Databanken?
- Hvordan bliver processen for at integrere systemer med Databanken (eksempelvis DMS)?

- Hvordan bliver processer for hvordan en opgave omkring oprettelse af nye BRAS-rettigheder til et specifikt udsnit af data indmeldes? Eksempelvis specifikke tabeller og kolonner.

Andet Udestående

Andre udeståender som kræver yderligere nedbrydning:

- Beskrivelse af muligheder for idriftsættelse af ICT's datatransformationer (WIP)
- Kravsstillelse ift. behandlingstid fra at (CDC) data rammer Databanken til det kan tilgås fra ICT's analytiske arbejdsområde.
- Samarbejdsform ifm. samarbejde om fælles aktiviteter (opsætning, udvikling mv.): Oprettelse af scrum-teams med folk fra både ICT og UFST Data.
- Rækkefølge på data-kilder der skal hjemtages til Databanken og ICT's analytiske arbejdsområde.
- Paralleldrift ift. DMS og legacy systemer. Hvordan påvirker det Databanken? Hvordan skal det håndteres.
- Databanken og bilag 2 af DMS-udbudet
- Integration til Databank ift. at hente Masterdata (DMS?).

Ordliste

- **Hadoop**: Den grundlæggende teknologi-stak som udgør fundamentet for arkivet i Databanken.
- **Hadoop Edgenode**: En node i en Hadoop cluster som tilgås eksternt og som har de nødvendige applikationer konfigureret til at kommunikere med Hadoop-clusteret via eksempelvis Ambari og Spark.
- **HDFS**: Distribueret filsystem som er en af grundstenene i Hadoop-stakken. Med distribueret menes at flere computere bruges til at drive det samme filsystem.
- **Hive**: En teknologi der gør muliggør SQL i HDFS. Vha. af Hive er det muligt at bruge SQL-queries på strukturerede filer i HDFS.
- **UDF / UDAF**: User Defined Function / User Defined Aggregate Function. Hentyder til en funktionalitet i Hive, hvor det er muligt at skrive sine egne ETL-funktioner som behandler data inden de hentes ud af tabellen/view'et i Hive.
- **Ranger**: En komponent i Hadoop-stakken som bruges til rettighedsstyring af HDFS/Hive mv. I Ranger kan der til en given AD-gruppe gives nøje specificeret adgang til eksempelvis specifikke kolonner og rækker i en Hive tabel. Eller angives læse/skrive rettigheder til specifikke mapper i HDFS.
- **Ambari**: En web-baseret brugergrænseflade til administration af Hadoop. F.eks. kan HDFS-kataloger gennemses i HDFS og queries kan eksekveres og gennemses i Ambari.
- **(Py)Spark**: Spark er en teknologi til distribueret beregninger baseret på data fra eksempelvis HDFS. Dertil findes PySpark, et API til Python.
- **Ingest/hjemtagelse**: Indlæsning af data i Databankens arkiv i Hadoop. Det primære værktøj til dette er Nifi, men det er også muligt at benytte en API-gateway.
- **Qlik**: Et business intelligence værktøj og platform. Qlik bruges til at lave dashboards mv. til visualisering af data, eksempelvis kpi'er eller andre nøgletal for forretningen.