SEGES INNOVATION

NOTAT

Nr. 2415

Udgivet 4. september 2024

SEGES Innovations Data Warehouse

Christian Aastrup, Kevin Sebastin og Gunner Sørensen

SEGES Innovation P/S

STØTTET AF

Svineafgiftsfonden

Hovedkonklusion

Der er udviklet en færdig version af SEGES Innovations Data Warehouse pr. 1. juli 2024 til indlæsning af enkeltdyrsdata fra besætningerne med søer, som anvender managementsoftware fra AgroVision eller Cloudfarms. Herfra læses data til rapporterne om pattegrise- og sooverlevelse i SEGES InSight.

Sammendrag

SEGES Innovations Data Warehouse er et driftssikkert, automatisk system til at hente enkeltdyrsdata fra AgroVision og Cloudfarms datamanagementsoftware på besætningsniveau, når ejeren af besætningen har givet tilladelse til det. Metoden til at opbygge og indlæse data er afsluttet 1. juli 2024 og beskrives i dette notat.

Data lagres i individuelle databaser og herfra udlæses data til rapportgenerering i SEGES InSight, så deltagende griseproducenter modtager et hurtigt, aktuelt og overskueligt overblik over udviklingen i so- og pattegriseoverlevelsen i deres sohold.

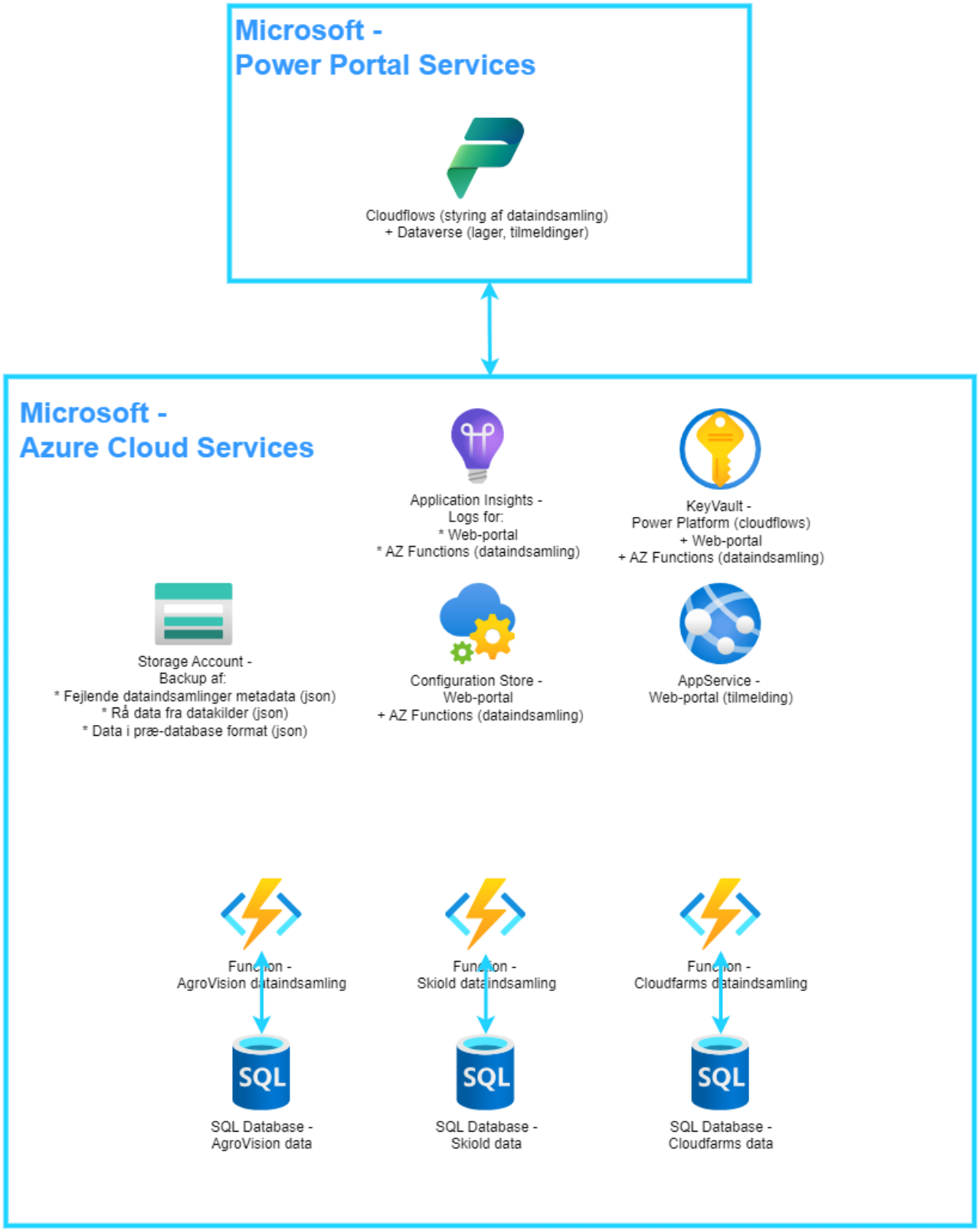
Antallet af deltagende besætninger i SEGES InSight er stigende og pt. indlæses enkeltdyrsdata fra mere end 300 besætninger.

Dataindsamling

Beregninger på enkeltdyrsdata fra besætninger med søer er grundlaget for SEGES InSight. Det er essentielt for aktiviteten at udvikle et driftssikkert system til at indlæse, lagre og kvalitetssikre data fra AgroVision og Cloudfarms, som beskrevet i det efterfølgende. Der arbejdes på en lignende løsning for Skiold angivet i diagrammet figur 1.

SEGES Innovation Data Warehouse indlæser via web-API'er fra AgroVision og Cloudfarms. Web-API'errne er forskellige for de to selskaber. Indlæsningen foregår ved hjælp af automatiserede processer i Microsofts Azure cloudtjenester samt i Microsofts Power tjenester. Data gemmes i Microsoft SQL Server databaser (én database pr. selskab). Den enkelte databases datamodel afhænger af datakilden selv – data normaliseres under den senere analyse.

Dataindsamlingen styres fra cloud flows i Microsofts Power services og der er etableret et flow pr. kilde. Hvert flow er tidsudløst, det vil sige, at der indsamles data på faste tidspunkter af døgnet. Der indsamles data alle årets dage.



Figur 1: Oversigtsdiagrammer for dataflow.

Det er kun medarbejdere i SEGES Innovation P/S, Digital, som har adgang til samtlige ressourcer i ovenstående oversigtsdiagram. Der er tildelt adgang til dele af ressourcerne for medarbejdere, som

udfører analyserne i SEGES InSight. Der er ligeledes adgang fra SEGES Innovation P/S Github miljø til SEGES InSight (kun Azure ressourcer), idet test og deployment faciliteres gennem GitHub.
Der er oprettet et testmiljø både i Azure samt Power (og GitHub i form af GitHub Environment), der er identisk med produktionsmiljøet (dog ikke med produktionsdata).

Datamodel – AgroVision

Data fra AgroVision (API kan ses her: Swagger UI (agrovision.com)) udstilles pr. datatype, f.eks. dyr, faringer mv. Der er derudover mulighed for at hente data*ændringer*, der kan repræsentere en hvilken som helst af de øvrige datatyper. Alle data fra AgroVision registreres med et unikt ID (en "global unique identifier (GUID)"), ExternalId, som benyttes som nøgle i databasetabellerne. Data*ændringer* er identificeret unikt, men kan relatere til samme hændelse (unikke ID); kun seneste ændring gemmes, da AgroVisions API ikke bibeholder komplet historik.

Der hentes følgende data fra AgroVision API (medmindre andet er anført, er alle hentede dataposter altid associeret med en bedrift (gård)):

Navn, SEGES InSight	Navn, AgroVision	Beskrivelse
Abortions	Reproduction/Abortions	Aborter. Hændelser i tid bundet til individdyr.
Animals	IndividualAnimal/Animals	Enkeltregistrerede dyr bundet til et unik gård-ID. Gemmes som eneste datapunkt 10 år tilbage.
BackfatMeasurements	IndividualAnimal/BackfatMeasurements	Rygspækmålinger. Hændelser i tid bundet til individdyr.
Causes	FarmSettings/Causes	Liste over mulige årsager (f.eks. afgangsårsager) i bedriften.
DeadPigletGroups	Reproduction/DeadPigletGroups	Pattegris død. Hændelser i tid bundet til en gruppering af dyr.
DeadPiglets	Reproduction/DeadPiglets	Pattegris død. Hændelser i tid bundet til individdyr.
Entries	IndividualAnimal/Entries	Indsættelser i bedriften. Hændelser i tid bundet til individdyr.
Exits	IndividualAnimal/Exits	Afgang (inkl. død) fra bedrift. Hændelser i tid bundet til individdyr.
FarmAnimalCause	IndividualAnimal/FarmAnimalCause	Ikke-døds ændringer i individdyrs status/tilstand. Hændelser bundet i tid til individdyr.
Farms	Farms	Bedrifter.
Farrowings	Reproduction/Farrowings	Faringer. Hændelser bundet i tid til individdyr.
Feeds	FarmSettings/Feeds	Fodertyper anvendt i bedriften.

HeatObservations	Reproduction/HeatObservations	Registreringer af drægtighed. Hændelser bundet i tid til individdyr.
Herds	Breeding/Herds	Grupperinger i bedriften pr. race.
HerdEntries	Breeding/HerdEntries	Indsættelser i gruppering. Hændelser bundet i tid til individdyr.
Locations	Farms/Locations	Lokationer i bedriften.
PigletCauses	Reproduction/PigletCauses	Ikke-døds ændringer i status/tilstand for et antal pattegrise pr. individdyr.
PigletTransfers	IndividualAnimal/PigletTransfers	Flytninger af pattegrise. Hændelser bundet i tid til individdyr.
PregnancyTests	Reproduction/PregnancyTests	Graviditetstests. Hændelser bundet i tid til individdyr.
Services	Reproduction/Services	Insemineringer. Hændelser bundet i tid til individdyr.
Transfers	IndividualAnimal/Transfers	Flytninger. Hændelser bundet i tid til individdyr.
Weanings	Reproduction/Weanings	Fravænninger. Hændelser bundet i tid til individdyr.
Weights	IndividualAnimal/Weights	Vægtmålinger. Hændelser bundet i tid til individdyr.

Data fra AgroVision gemmes uden indbyrdes relationer (dvs. ingen afhængigheder); data kan bindes sammen vha. de unikke ID-felter for bedrift, dyr og hændelser (ExternalId).

Der foretages filtrering af data hentet fra API'et inden lagring. Filtreringen foregår både på feltet ExternalId i form af sikring af unik forekomst pr. datapost samt en filtrering ift. plausibilitet på data, f.eks. multiple afgange for samme individdyr eller lignende biologisk umulige dataposter.

Data fra AgroVision genindlæses i sin helhed for hver dataindlæsning og de eksisterende data fjernes. Dermed giver de i SEGES InSight-databasen lagrede data med maksimum 12 timers forsinkelse et billede af datakildens tilstand.

Datamodel – Cloudfarms

Data fra Cloudfarms hentes i sin helhed pr. bedrift og fordeles herefter i tabeller pr. datatype. I lighed med AgroVision er alle data relateret til en bedrift (udover bedriften selv). Data fra Cloudfarms identificeres ved et fortløbende unikt id.

Navn, SEGES InSight	Navn, Cloudfarms	Beskrivelse
BackFatSows	BackFatSow	Rygspækmålinger. Hændelser i tid pr. individdyr.
Farms	Farm	Bedrifter.
Farrowings	Farrowings	Faringer. Hændelser i tid pr. individdyr.
Matings	Matings	Insemineringer. Hændelser i tid pr. individdyr.
MedicineUsage	Medicineusage	Medicineringer. Hændelser i tid pr. individdyr.
PigletDeaths	PigletsDeath	Pattegrisedød. Hændelser i tid pr. individdyr.
Sows	Sows	Søer. Enkeltregistrerede dyr bundet til et unik gård-ID.
Weanings	Weanings	Fravænninger. Hændelser i tid pr. individdyr.

Data fra Cloudfarms genindlæses i sin helhed for hver dataindlæsning og eksisterende data fjernes ligeledes inden genindlæsning.

Metadata

For alle dataindlæsninger for alle datakilder gemmes forskellige metadata. Disse kan opdeles i to kategorier:

- Kontrol af indlæsning (kan også bruges som styring af dataindlæsning), herunder evt. fejl
- Kontrol af data (rådata fra kilden samt *snapshot* af data umiddelbart før lagring

Metadata gemmes i kildens tilhørende database og i et centralt lager i Azure (Storage Account, Blobstorage).

Til kontrol af indlæsning gemmes eventuelle fejl, tidspunkter for indlæsning pr. bedrift, kildens seneste opdateringstidspunkt samt seneste data i databasen. Ved fejl sendes metadata til yderligere inspektion, f.eks. bedrifts-ID, dyr-ID, hændelses-ID, hændelsestype mv.

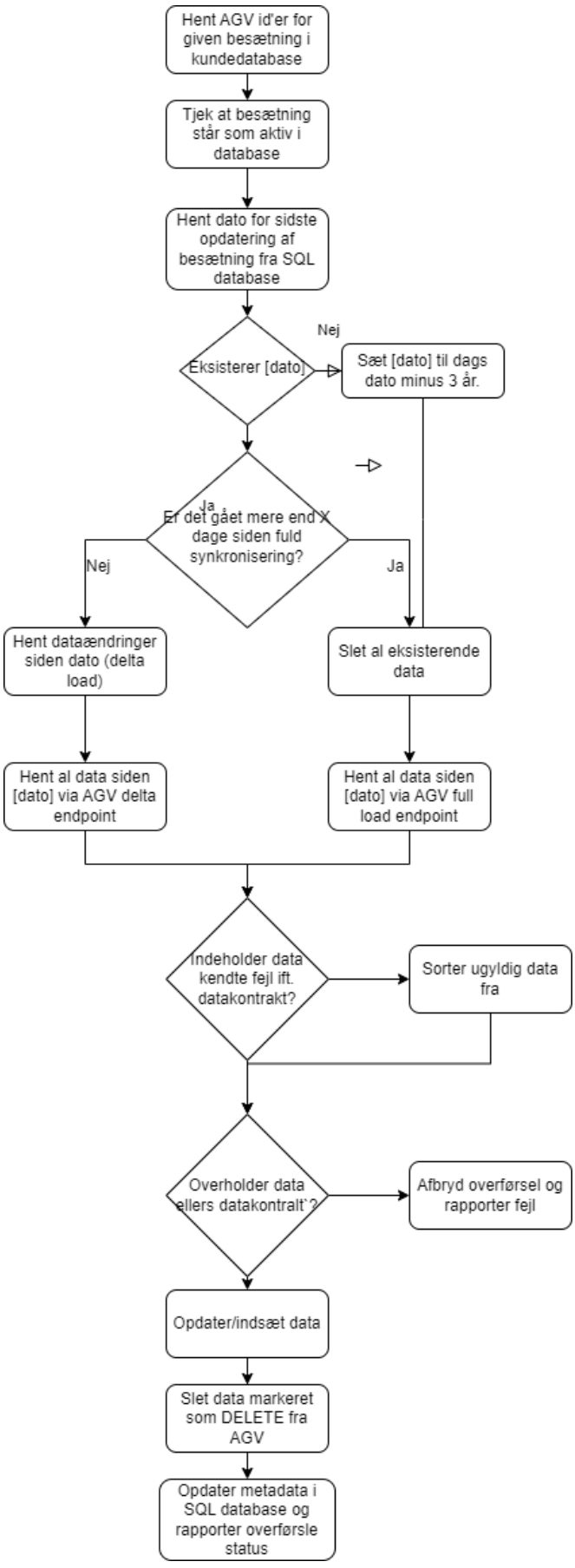
Til kontrol af data gemmes kildens rådata (json format) i et blob-storage (Azure) kategoriseret efter dato og bedrift samt (for AgroVision) datapunkt. Ligeledes gemmes data, konverteret til json, i fortolket format efter filtrering. Derudover lagres alle metadata (inkl. fejl) også i et særskilt blob-storage (Azure) fordelt pr. kilde. Alle dataindlæsninger logger både interne handlinger samt alle handlinger ud og ind af de enkelte dataindlæsningssystemer.

Til hver kilde er udviklet en Microsoft Azure Function (durable) i C#. Hver Function har sit eget repository i GitHub (SEGES Innovation P/S egen), hvorfra også test og deployment af nye versioner udføres. Hver Function har adgang til en database til lagring af kildedata samt adgang til fælles lager af nøgler (KeyVault) – de enkelte dataindlæsninger benytter *eget lager*. Hver Function kan kun tilgås med en unik nøgle, der er lagret i Azure (KeyVault), hvortil de styrende flows har adgang (se afsnit om Styring af indlæsning).

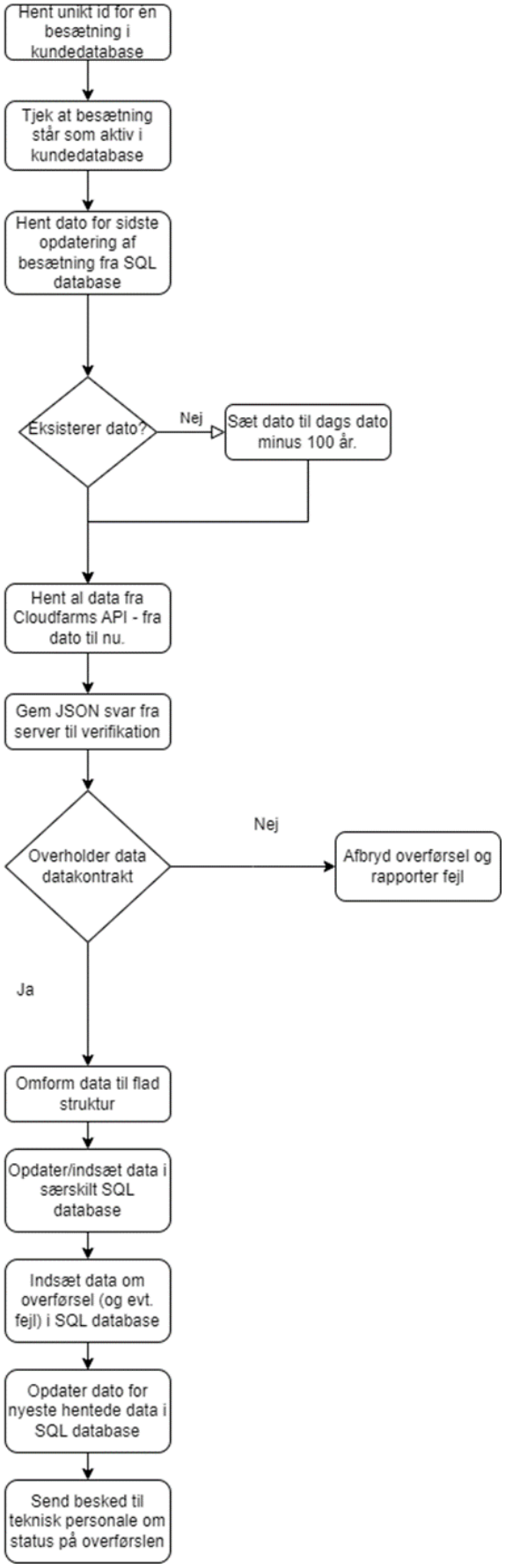
Styring af indlæsning

Til styring af dataindlæsning benyttes Microsoft Power platforms Cloud Flows (Power Automate), der læser fra SEGES InSight kundelager og aktiverer dataindlæsningsfunktionerne (Azure Functions). Der er oprettet et flow pr. datakilde, som aktiveres to gange dagligt med 12 timers mellemrum.

Flow for dataindlæsning (AgroVision):



Flow for dataindlæsning (Cloudfarms):



Konklusion

SEGES Innovations Data Warehouse er et driftssikkert, automatiseret system til at hente enkeltdyrsdata fra AgroVision og Cloudfarms managementsoftware på besætningsniveau. Systemet udvides til også at omhandle Skiold Digital Solutions og kan derefter indsamle enkeltdyrsdata fra alle udbydere af managementsoftware til besætninger med søer.

Fra SEGES Innovations Data Warehouse kan der udlæses data til rapportgenerering i SEGES InSight, så deltagende producenter modtager et hurtigt, aktuelt og overskueligt overblik over udviklingen i so- og pattegriseoverlevelsen i deres sohold.

Deltagere

Christian Aastrup og Kevin Sebastin

Afprøvning nr. 1832

NAV nr.: 101453

//JAHP//