SAS® UND HADOOP® MIT KERBEROS-AUTHENTIFIZIERUNG

HANS-JOACHIM EDERT, SOLUTION ARCHITECT
SAS DEUTSCHLAND

EINE FRAGE VORWEG …

Würden Sie von diesem Automobil erwarten, dass es mit einem abschließbaren Tankdeckel ausgestattet ist?

Wohl eher nicht! Zu dieser Zeit ging es vorrangig darum, die Basisfunktionalität bereit zu stellen.

EINFÜHRUNG … UND WIE STEHT ES HIERMIT?

Würden Sie von einem state-of-the-art Framework zur parallelen Verarbeitung großer Datenmengen erwarten, dass es über ein robustes Authentifizierungsverfahren verfügt?

Definitiv! Vor allem, wenn es mit dem Anspruch antritt, als Data Lake oder als Enterprise Data Hub eingesetzt zu werden.

HADOOP® IN DER UNTERNEHMENS IT-INFRASTRUKTUR

- Die Einsicht, dass Security-Vorgaben für Enterprise Systeme auch für Hadoop®-Cluster gelten müssen, ist noch nicht alt.
- Analog zum Automobil wurde lange Jahre an der Basisfunktionalität gefeilt:
  - Historie: Offene Cluster, Laborumgebungen für Data Scientists.
  - Auch heute noch weit verbreitet (die wenigsten Kunden haben Hadoop bereits produktiv im Einsatz).
- Fehlende Security Compliance ist mit einer der Hauptgründe für den ausbleibenden Schritt in Richtung Produktion.

HADOOP® IN DER UNTERNEHMENS IT-INFRASTRUKTUR

- Umdenken bei den grossen Hadoop® Distributoren (Cloudera / Hortonworks).
- Beide Distributoren propagieren eine geschlossene Security-Architektur in ihren Hadoop® Blueprints.
- Kerberos als zentrales Authentifizierungssystem steht im Mittelpunkt.

SECURITY-ARCHITEKTUREN BEI HORTONWORKS / CLOUDERA

Comprehensive Approach To Security

Hortonworks

Administration Centrally manage consistent security	APACHE RANGER
Authentication Prove the identity of systems and users	KERBEROS & APACHE KNOX
Authorization Provide secure access to data	APACHE RANGER
Audit Maintain a record of data access events	APACHE RANGER & APAC
Data Protection Safeguard data at rest and in motion	HDFS TDE with RANGER K

Cloudera

Result: Cloudera is the most secure Hadoop pla

Perimeter	Access	Visibility	Data
Guarding access to the cluster itself	Defining what users and applications can do with data	Reporting on where data came from and how it's being used	Protecting data in the cluster from unauthorized visibility
Technical Concepts: Authentication Network isolation	**Technical Concepts:** Permissions Authorization	**Technical Concepts:** Auditing Lineage	**Technical Concepts:** Encryption, Tokenization, Data masking
Kerberos | AD/LDAP	Rhino | Sentry	Cloudera Navigator	Encrypt | Key Trustee

22

cloudera

KERBEROS 101

KERBEROS | EINLEITUNG

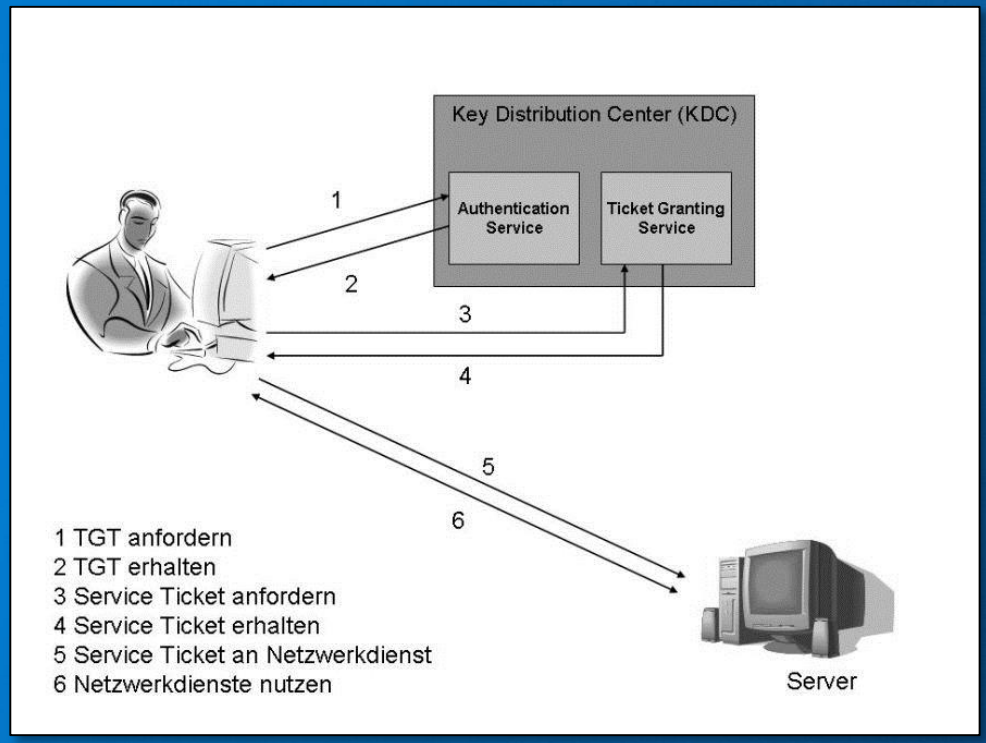
- Kerberos = Höllenhund aus der griechischen Mythologie, bewacht den Eingang zum Hades.
- Drei Köpfe = die drei zentralen Komponenten in einer Kerberos Umgebung.
- Verteilter Authentifizierungsdienst, entwickelt am MIT.
- Liegt seit 1993 in der (immer noch aktuellen) Version 5 vor.
- Verfügbar als Stand-Alone Implementierung und als Bestandteil der Active Directory Services.

KERBEROS | EINLEITUNG

- Verfolgt einen Trusted-Third-Party Ansatz (Dritte Partei übernimmt Authentifizierung, nicht der angefragte Dienst selbst).
- Basiert auf dem Austausch von Schlüsseln (es werden keine Passwörter übermittelt).
- Unterstützt Single-Sign-On (einmalige Authentifizierung ausreichend für die Nutzung der weiteren Dienste).

KERBEROS | ABLAUF DER AUTHENTIFIZIERUNG

- Key Distribution Center (KDC):
  - Authentication Service:
    - Führt die Authentifizierung durch,
    - Erteilt das Ticket Granting Ticket (TGT),
  - Ticket Granting Service:
    - Erteilt auf Vorlage des TGTs die benötigten Service Tickets.
- Client / Server Kommunikation:
  - Client legt das Service Ticket dem Netzwerkdienst vor, den er nutzen möchte.



Quelle: http://www.tecchannel.de/server/windows/461645/die_funktionsweise_von_kerberos/

KERBEROS | VORBEDINGUNGEN

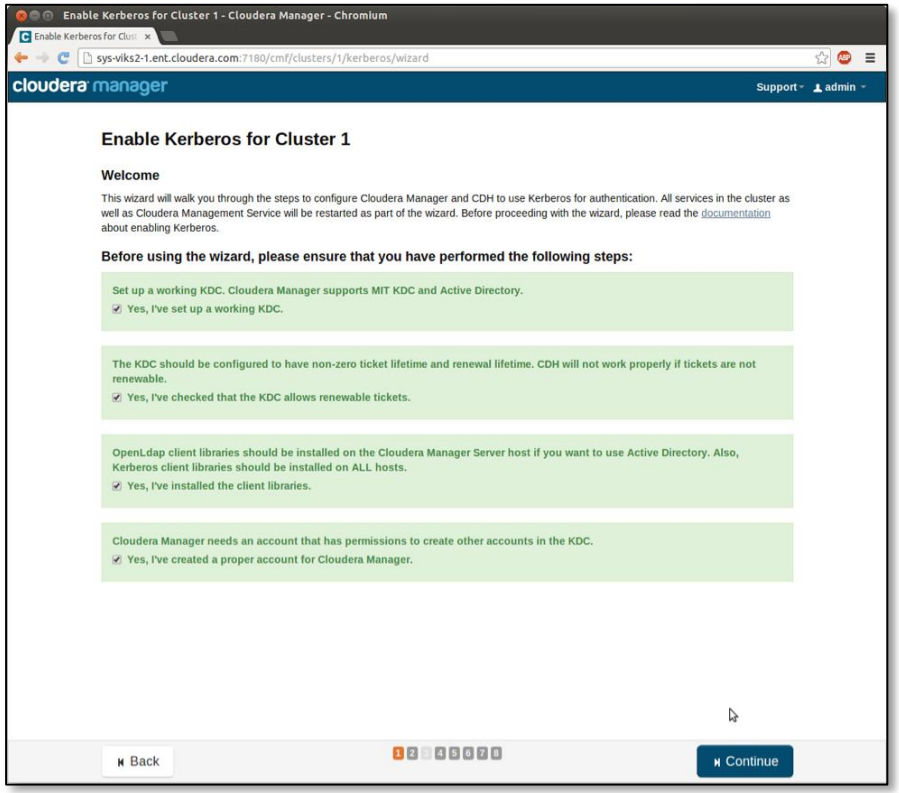
- TGTs und Service Tickets besitzen grundsätzlich nur eine beschränkte Lebensdauer (z.B. einige Stunden).
  - Synchronisierung der Uhrzeiten auf allen beteiligten Systemen ist essentiell (z.B. Linux: ntpd).
- Kerberos verwendet verschiedene Schlüsselstärken, die teilweise US Export-Restriktionen unterliegen (256 Bit):
  - Alle Java Runtime Environments auf Client und Server müssen mit den Java Cryptography Extensions (JCE) Unlimited Strength Policy Files gepatcht werden.
  - In manchen Fällen muss auch das Betriebssystem gepatcht werden (z.B. Solaris benötigt das SUNWcry package).

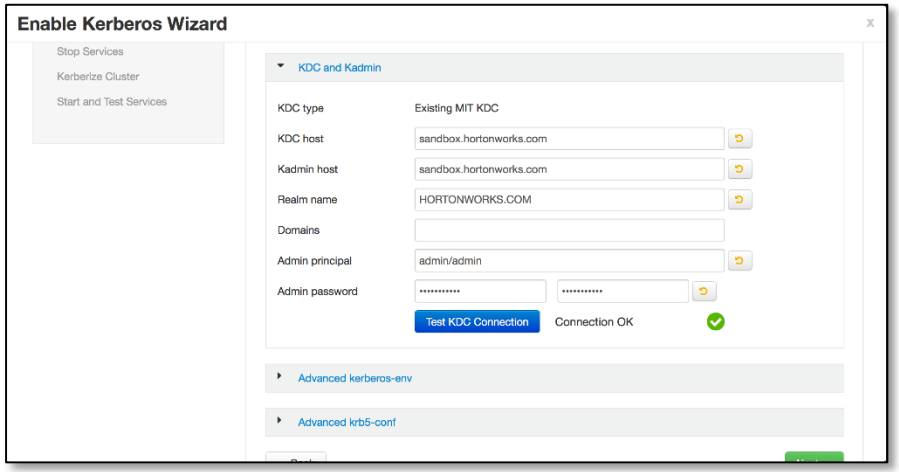
HADOOP® & KERBEROS

HADOOP® & KERBEROS

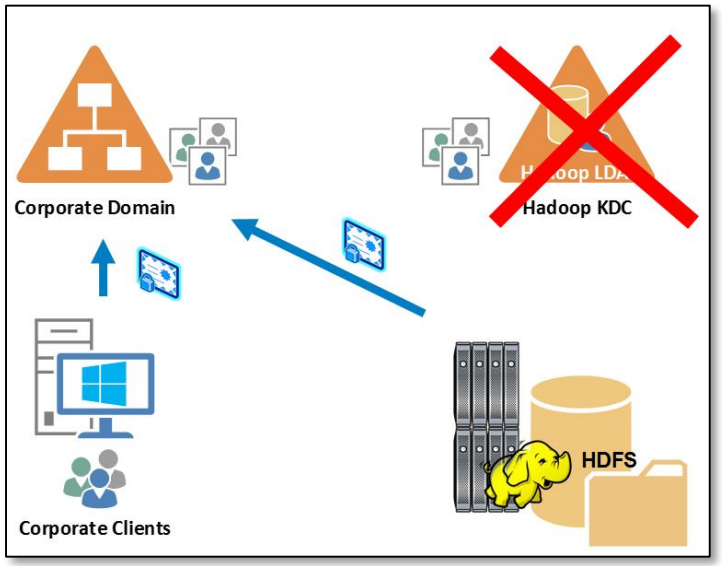
EINLEITUNG

- Hadoop® ist ein Ökosystem und besteht aus zahlreichen separaten Projekten:
  - HDFS, Hive, Yarn, MapReduce, HBase, Solr, Storm…
- Jedes Projekt muss für Kerberos konfiguriert werden:
  - Dienste benötigen einen Service Principal und ein Keytab File.
  - Verteilte Dienste (HDFS) benötigen einen Service Principal für *jeden DataNode*, auf denen sie laufen.
  - Die Dienst-Konfigurationen (XML) müssen angepasst werden.
- Tools von Hortonworks (Ambari) und Cloudera (Cloudera Manager) automatisieren diesen Vorgang, benötigen dafür aber Schreibrechte für den KDC / das AD.



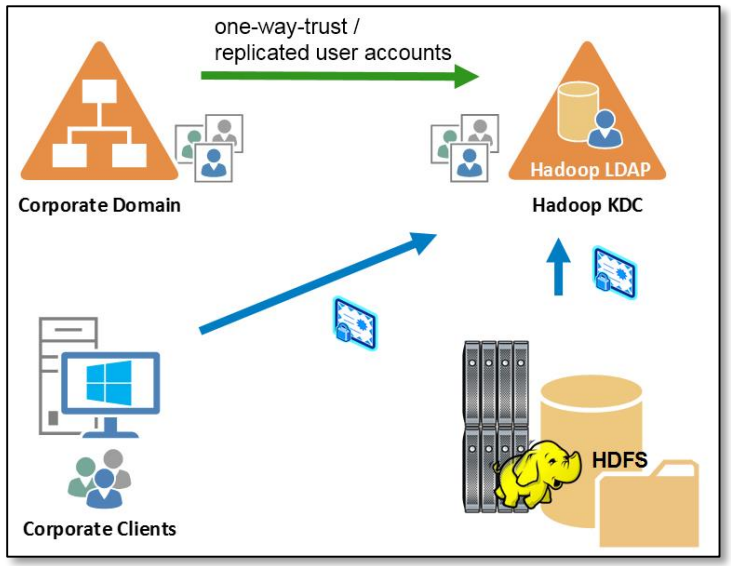


HADOOP® & KERBEROS – SECURITY INFRASTRUKTUR



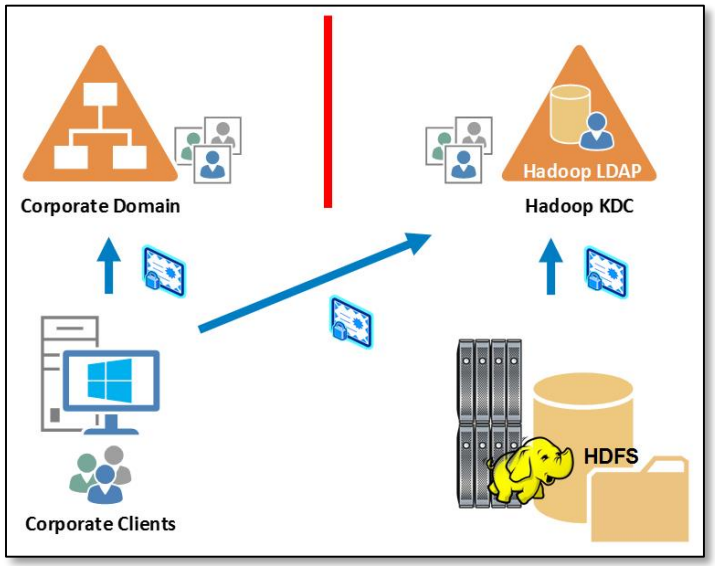
„AD ONLY“ - HADOOP® CLUSTER AUTHENTIFIZIERT GEGEN DAS CORPORATE AD

- Unterstützt von Cloudera Manager und Ambari.
- User Principals und Service Principals im AD gespeichert.
- Hadoop® Tools benötigen Schreibrechte für AD.



„ONE WAY TRUST“ ODER „REPLICATED USER ACCOUNTS“

- „one way trust“ - User Accounts verbleiben im AD, können aber über KDC authentifiziert werden.
- „replicated user accounts“ - User Accounts werden auch im Hadoop® KDC angelegt.



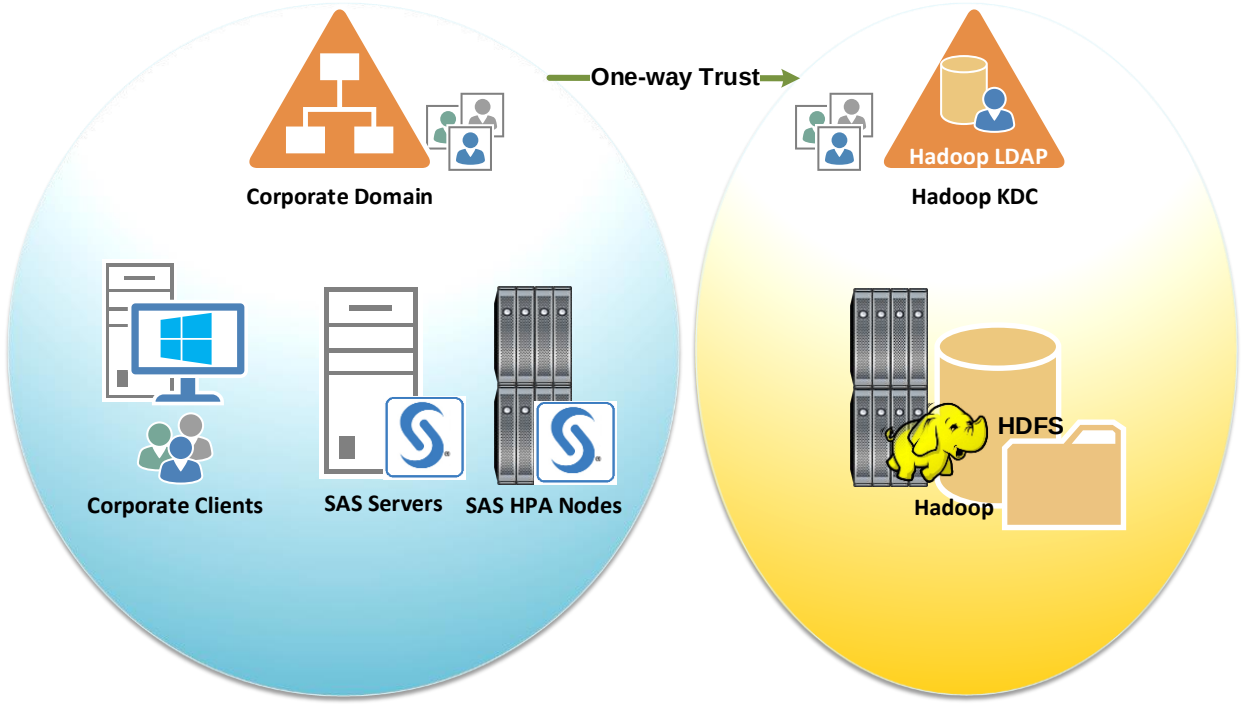
KEINE BEZIEHUNG ZWISCHEN AD UND KDC

- User Accounts werden manuell im KDC angelegt.
- User Accounts sind technische User bzw. Service Accounts:
  - alles andere wäre zu aufwändig.

SAS®, HADOOP® & KERBEROS

SECURITY INFRASTRUKTUR

- Ausgangssituation für SAS®:
  - Gewählte Security Infrastruktur für Hadoop® ist zumeist gegebene Rahmenbedingung.
  - Authentifizierungsverfahren für existierende SAS® Umgebungen ebenfalls zumeist bereits gesetzt.
  - In den meisten Fällen ist die SAS® Umgebung in die Corporate Domain eingebunden (und nicht in den Hadoop® Kerberos Realm).
  - Ggf. Ausnahmen bei Neuinstallationen.



SECURITY INFRASTRUKTUR

Was benötigt ein SAS® Anwender für den Zugriff auf Hadoop® Ressourcen (mit aktivem Kerberos)?

- SAS® benötigt immer ein TGT um weitere Service Tickets anfordern zu können:
  - Hierfür ist eine vorherige Authentifizierung notwendig.
- Darüber hinaus - abhängig von den genutzten SAS® Produkten:
  - SAS® /Access to Hadoop® / Impala + Embedded Process (Accelerators):
    - Service Tickets für Hive und HDFS.
    - Service Ticket für Impala (Cloudera).
  - SAS® In-Memory Analytics Infrastructure (TKGrid):
    - Host Service Tickets, um passwordless SSH via Kerberos zu ermöglichen (GSSAPI Authentication).

ERWERB EINES TGTS

- Falls Quest Authentication Services (QAS)* verfügbar sind:
  - Das TGT wird automatisch angelegt, wenn der Anwender sich auf dem SAS® Server authentifiziert (PAM).
  - Erlaubt Single-Sign On mit IWA.
- Ohne QAS:
  - Anwender führt interaktiv ein `kinit` auf dem SAS® Server aus (generiert / aktualisiert den Ticket Cache in `/tmp/krb5cc_<uid>_<rand>`) - keine praktikable Lösung.
  - Anpassungen in der `WorkspaceServer_usermods.sh` automatisieren diesen Vorgang:
    - Es wird ein vorab hinterlegtes Keytab File benötigt, darum vorrangig ein Ansatz für Service Accounts.
    - Denkbar auch für Endanwender Accounts, allerdings mit erhöhtem Verwaltungsaufwand.
- SAS® wertet die Umgebungsvariable KRB5CCNAME aus, die auf einen gültigen Ticket Cache zeigen muss (automatisch gesetzt).

* oder allgemein (seit 9.4 M2): „a *shared library that implements the GSSAPI with Kerberos 5 extensions*”

SAS®, HADOOP® & KERBEROS

BEISPIEL #1: AUF HADOOP® SOLL MIT EINEM SERVICE ACCOUNT FÜR SAS® ZUGEGRIFFEN WERDEN

- Schritt 1: SAS® Konfiguration für Kerberos anpassen.

```
[lasradm@gertesthadoop1 ~]$ cat /opt/sas/config/Lev1/SASApp/sasv9_usermods.cfg
-JREOPTIONS (    -Djava.security.krb5.conf=/etc/krb5.conf    )
```

- Schritt 2: Beim Start der Workspace Server Session wird das TGT angefordert:
  - Ein Keytab File wurde zuvor erzeugt und auf dem SAS® Server gespeichert.

```
[lasradm@gertesthadoop1 ~]$ cat /opt/sas/config/Lev1/SASApp/WorkspaceServer/WorkspaceServer_usermods.sh
kinit -k -t /opt/sas/Lev1/lasradm.keytab lasradm
```

- Schritt 3: Kontrolle in SAS® Session.

```
%let krb5env=%sysget(KRB5CCNAME);
%put &krb5env.;
```

BEISPIEL #2: AUF HADOOP® SOLL MIT ANWENDER ACCOUNTS ZUGEGRIFFEN WERDEN

- Beim Start der Workspace Server Session wird das TGT angefordert:
  - Ein Keytab File wurde zuvor für jeden Anwender erzeugt und auf dem SAS® Server in dessen Home-Verzeichnis gespeichert.

```
[lasradm@gertesthadoop1 ~]$ cat /opt/sas/config/Lev1/SASApp/WorkspaceServer/WorkspaceServer_usermods.sh
workspace_user=$(whoami)
kinit -k -t ~/hadoop.keytab $workspace_user

workspace_user_ccaches=$(find /tmp -maxdepth 1 -user ${workspace_user} -type f -name "krb5cc_*" -printf
'%T@ %p\n' | sort -k 1nr | sed 's/^[^ ]* //' | head -n 1)
if test ! -z "$workspace_user_ccaches"; then
  echo "Most recent krb5 ccache found for '${workspace_user}' at '${workspace_user_ccaches}'."
  export KRB5CCNAME=$workspace_user_ccaches
  echo "KRB5CCNAME has been set to ${KRB5CCNAME}."
else
  echo "No krb5 credentials caches were found in /tmp for '${workspace_user}'."
fi
```

SAS®, HADOOP® & KERBEROS

SAS® ZUGRIFF AUF HIVE, IMPALA UND MAPREDUCE UNTER KERBEROS

- SAS®/Access to Hadoop®:
  - Warum der HDFS_PRINCIPAL? Beim Anlegen von Hive Tabellen werden die Daten temporär ins HDFS geschrieben und von dort dann nach Hive kopiert.

```
libname HIVE hadoop server="gatecdh01.gatehadoop.com" subprotocol=hive2
   HDFS_PRINCIPAL="hdfs/_HOST@GATEHADOOP.COM" HIVE_PRINCIPAL="hive/_HOST@GATEHADOOP.COM";
```

- Hadoop® Accelerators (Embedded Process):
  - Keine besondere Konfiguration in SAS® - Informationen werden aus der mapred-site.xml ausgelesen.

- SAS® /Access to Impala:
  - Keine besondere Konfiguration in SAS® - Informationen werden in der odbc.ini gesetzt.

```
KrbRealm=Realm;
KrbFQDN=DomainName;
KrbServiceName=ServiceName
```

STARTEN EINES TKGRID PROZESSES UNTER KERBEROS

- TKGrid (~ „In-Memory Grid“) - die Basiskomponentefür sowohl VA als auch HPA.
- TKGrid benötigt eine Form von „passwordless SSH“ - normalerweise via Public Key:
  - Unter Kerberos wird umgestellt auf GSSAPI Authentication.

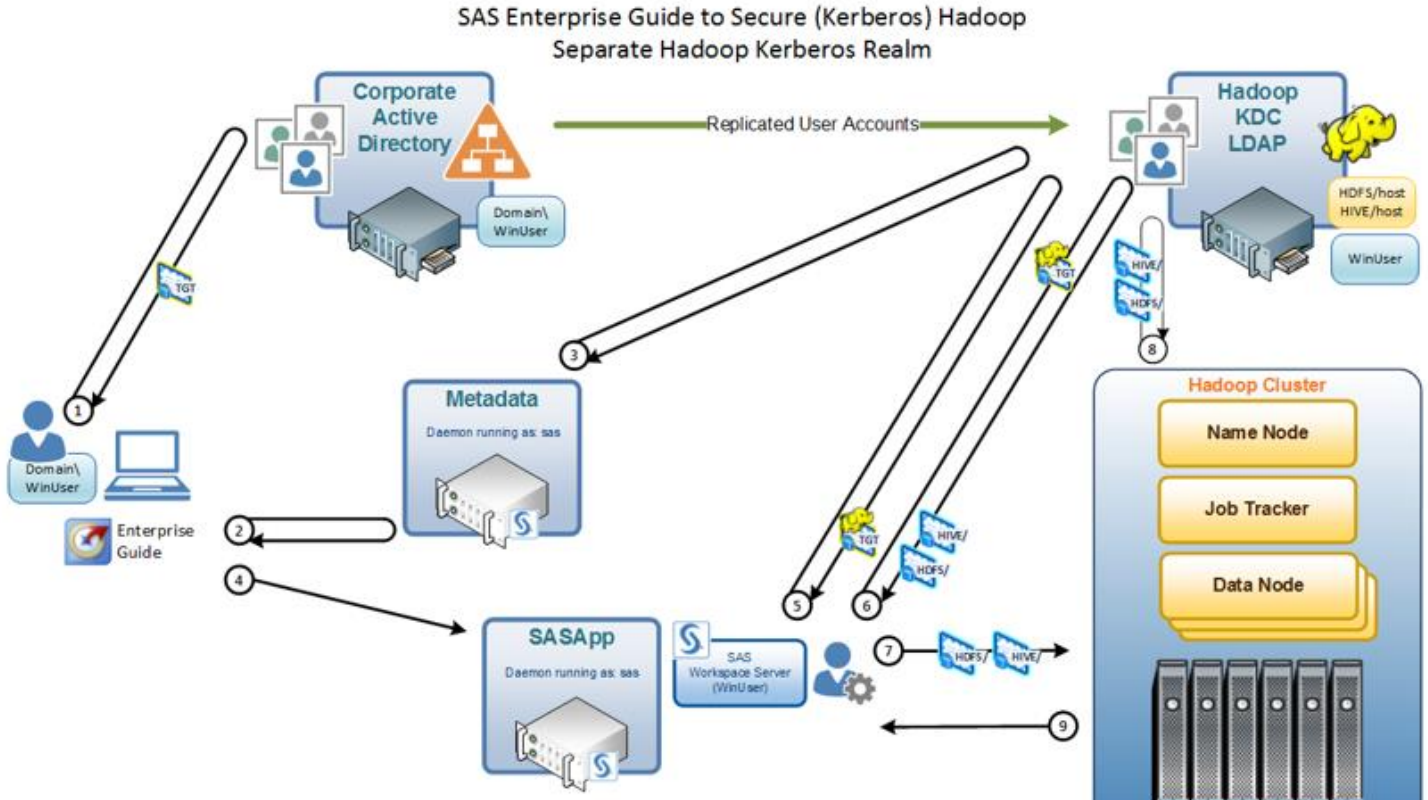
- Konfiguration #1: Auslesen des Ticket Caches beim Start von TKGrid.

```
[root@gertesthadoop1 ~]$ cat /opt/TKGrid/tkmpirsh.sh
export MPI_OPTIONS="$MPI_OPTIONS -genv DISPLAY=$DISPLAY -genvlist `env | sed -e s/=.*/,/ | sed /KRB5CCNAME/d | tr -d
'\n'`TKPATH,LD_LIBRARY_PATH“
```

- Konfiguration #2: Internes SAS-SSH durch System-SSH ersetzen.

```
option set=GRIDRSHCOMMAND="/usr/bin/ssh -o StrictHostKeyChecking=no -o PasswordAuthentication=no -o PubkeyAuthentication=no";
```

SAS®, HADOOP® & KERBEROS - WITH SECURED HADOOP®



1. User logs into Windows and obtains a TGT
2. User logs into SAS Enterprise Guide, authenticated by SAS Metadata Server
3. SAS Metadata Server uses Operating System to validate user, which in turn uses PAM
4. SAS Enterprise Guide connects to the SAS Object Spawner and requests a SAS Workspace Server, passing credentials (WinUser)
5. SAS Object Spawner runs a shell script as the user and spawns a continuously running SAS Workspace Server. A Kerberos TGT is generated by PAM
6. SAS Code executes: LIBNAME hivelib HADOOP; specifying Kerberos security principals for HIVE & HDFS. The Hadoop client libraries use the TGT to request Service Tickets for HIVE & HDFS
7. SAS connects to HIVE & HDFS using Service Tickets
8. User authenticated using the Service Ticket & Service Key
9. Process HIVE/Map Reduce request & send data back through SAS Workspace Server to client

FAZIT

- **Basisbaustein**: Ohne Kerberos kein Enterprise Hadoop®!
- **Wasserdicht**: Ist das Cluster kerberisiert, müssen Clients für jeden Hadoop® Dienst, den sie nutzen wollen, ein Service Ticket vorweisen.
- **Knackpunkt TGT**: Wie kommt der SAS® Anwender an ein Ticket Granting Ticket?
  - Mittels Quest o.ä. (automatisch beim Logon via PAM) – endanwendertauglich.
  - Mittels Workspaceserver_usermods.sh (`kinit`) - primär für Service Accounts tauglich.
- **Tickets please**:
  - SAS®/Access to Hadoop® und die Hadoop® Accelerators benötigen Service Tickets für Hive und HDFS.
  - TKGrid Applikationen (VA/VS, HPA) benötigen zusätzlich Host Service Tickets.

WEITERFÜHRENDE LINKS

- SAS® & Hadoop®:
  - http://support.sas.com/resources/thirdpartysupport/v94/hadoop/index.html
- Insbesondere für Hadoop® und Kerberos:
  - Hadoop with Kerberos: Architecture Considerations (PDF)
    Read how to ensure interoperability between secure Hadoop environments and SAS®. Although this paper was written for SAS 9.4 M2, consider the content as best practices for later releases.
  - Hadoop with Kerberos: Deployment Considerations (PDF)
    Learn about deploying secure Hadoop® environments with SAS®. Although this paper was written for SAS 9.4 M2, consider the content as best practices for later releases.

VIELEN DANK FÜR IHRE AUFMERKSAMKEIT!

sas | THE POWER TO KNOW®