Hadoop im Unternehmenseinsatz, aber sicher

Oliver Gehlert
Ventum Consulting & Co KG
München

Schlüsselworte
Hadop, Security, Ranger, Kerberos, Knox, Sentry

Einleitung

Hadoop ist die Lösung für die Verarbeitung großer, multistrukurierter Daten. Performance und Features standen bei der Entwicklung an oberster Stelle. Zugriffschutz, Verschlüsselung und Auditing waren aber nicht die Top-Prioritäten bei den Entwicklern, sondern die Auswertung über den kompletten Datenbestand hinweg.

Mit dem Erfolg von Hadoop im Unternehmen, wird der Cluster für unterschiedliche Use-Cases und von verschiedenen Abteilungen benutzt. Aus Sicherheitsgründen und aufgrund von regulatorischen Vorgaben, möchte man nicht, dass alle User Vollzugriff auf alle Daten bekommen.

Was muss man beachten, um einen Cluster effizient abzusichern und welche Hilfsmittel stehen einem hierbei zur Verfügung? Welche Initiativen und Frameworks gibt es aktuell und welche Anforderungen decken diese ab?

Was heißt Sicherheit

Sicherheit in der IT beruht auf 5 Säulen

- Zentrale Administration
- Authentifizierung
- Autorisierung
- Datensicherheit
- Auditing

Anhand dieser Punkte zeigen wir den aktuellen Stand und die Möglichkeiten der Absicherung auf

Ausgangssituation

Hadoop ist kein Tool, sondern ein komplettes Ökosystem aus vielen einzelnen Komponenten. Diese müssen einzeln administriert und auditiert werden und es gibt keine zentrale Administration und Übersicht. Zusätzlich bietet die Basiskonfiguration eines Hadoop-Clusters bietet keinerlei effektive Sicherheit und ermöglicht einfachen Zugriff mit privilegierten Rechten.
Diese Ausgangssituation ist erstmal erschreckend, aber der Vortrag wird zeigen, dass Sicherheit in einem Hadop-Cluster kein Hexenwerk ist.

Authentifizierung und Zugriff auf Clusterresourcen

Kernpunkt im Rahmen der Sicherheit ist die Authentifizierung von Benutzern. Hier muss die Authentifizierungsmethode unbedingt von „None“ auf Kerberos umgestellt werden. Sonst kann sich jeder Benutzer innerhalb weniger Sekunden zum Superuser machen.

Die Konfiguration der Kerberos-Authentifizierung mit Kerberos ist manuell nicht trivial, aber inwischen gut dokumentiert bzw. die Basiskonfiguration kann über Tools wie Cloudera Manager oder Ambari erfolgen.

Zusätzlich sollte man den Zugriff auf den Hadoop-Cluster einschränken. Hierzu kann man entweder sogenannte Edge-Nodes verwenden, oder aber ein Framework wie Apache Knox.

Edge-Nodes oder Gateways sind Server, die zu zwei Netzwerksegmenten gehören, einerseits sind sie aus dem internen Netzwerk erreichbar, andererseits gehören sie zum selben Netzwerkbereich wie der Hadoop-Cluster. Auf Edge-Nodes sind die klassichen Hadoop-Tools, wie Hive, Pig, Impala, … installiert und die Knoten werden auch zum Staging von Daten verwendet.

Apache Know dient ebenfalls als Gateway zum Cluster, bietet aber andere Services, als ein Edge-Node:

- Zentraler Zugriffspunkt für Hadoop REST API Zugriffe
- Zentralisierte Authentifizierung und Autorisierung für Hadoop REST services
- LDAP/AD Authentifizierung und Autorisierung
- Übergreifendes Auditing
- Kapselung der Hadoop-Architektur
- Zentrale SSL Endpunkt

Apache Knox unterstützt aktuell folgende Komponenten

- Hive
- Hbase
- HDFS
- OOZIE
- HCat

Damit sind bereits viele Use-Cases abgedeckt. Apache Knox wird aktiv weiterentwickelt und wird insbesondere von Hortonworks gesponsort. Weitere Komponenten wie Pig, Storm oder Kafka werden aktuell nicht unterstützt und daher ist Knox nicht für alle Fälle ausreichend.

Autorisierung

Nicht jeder authentifizierte Benutzer darf auf alle Daten des Clusters zugreifen. Hierzu müssen die Zugriffsrechte für die einzelnen Gruppen oder User auf Toolebene gesetzt werden. Abhängigkeiten zwischen den einzelnen Hadoop-Komponenten müssen dabei explizit berücksichtigt werden. Es genügt z. B. nicht einem User Leserechte auf Tabellen in Hive zu entziehen, wenn der User noch Leserechte auf die zugrundeliegenden Dateien im HDFS hat.

Frameworks wie Apache Ranger, Sentry oder Recordservice unterstützen hier bei der effizienten Vergabe von Rechten. Im Vortrag werden die Features der Frameworks genauer miteinander verglichen.

Datensicherheit

Die Datensicherheit muss an mehreren Punkten gegeben sein. Einerseits bei der Speicherung der Daten und andererseits während des Transports.

Der Transport der Daten lässt sich gut per SSL absichern, hierbei ist zu prüfen, ob man den Datenverkehr innerhalb des Clusters per SSL absichern muss, oder ob es genügt, den Datenverkehr nach draußen zu sichern. Welches Vorgehen man wählt, hängt von der Klassifikation der vorhandenen Daten ab und sollte zusammen mit dem Information Security Officer bzw. der Compliance Abteilung abgestimmt werden. Apache Knox unterstützt bei der Absicherung des Datenverkehrs und minimiert den Aufwand, da nur für das Know-Gateway SSL Zertifikate gemanaged werden müssen.

Für die Datenspeicherung müssen zwei Aspekte berücksichtigt werden, einmal die komplette Festplattenverschlüsselung, die wahlweise direkt durch die Festplatte oder durch extra Software ermöglicht wird. Dies schützt die Daten nach einem Ausbau der Festplatten vor einem unbefugtem Auslesen und erleichtert die Entsorgung.

Andererseits schützen verschlüsselte Festplatten nicht durch Zugriff durch Benutzer, die Zugriff auf das System haben. Um sensible Daten vor dem Zugriff durch Administratoren zu schützen, gibt es analog zu Datenbanken, die Möglichkeit der Transparenten Datenverschlüsselung (TDE). Die Verwaltung der Schlüssel ist relativ komplex, hierbei helfen Apache Ranger oder Cloudera Navigator.

Auditing

Security ohne Auditing bietet keine Sicherheit. Unterlässt man die Überwachung der Logfiles, so fallen unzulässige Zugriffe gar nicht, oder zu spät auf. Hadoop erleichtert das Auditing nicht, da jede Komponente eigene Logfiles schreibt. Apache Ranger ermöglicht die Logs zu zentralisieren, in dem diese in ein RDBMS, nach HDFS oder per log4j geschrieben werden. Apache Know bietet ebenfalls ein zentrales Auditing aller Zugriffe, die über Apache Knox erfolgen.

Nutzt man ein RDBMS, so kann man die Auditdaten zentral auswerten und hält diese unabhängig vom Cluster.

Zentrale Administration

Die zentrale Administration aller Komponenten des Hadoop Ökosystems ist out-of-the-box nicht möglich. Die großen Hadoop-Distributionen liefern Webkonsolen mit, über die die meisten Komponenten administriert und überwacht werden können. Die Konsolen erleichtern die Administration, das Aufsetzen und das Monitoring ganzer Cluster, aber im Bereich Security sind die

Möglichkeiten noch eingeschränkt. Neue Frameworks, wie Apache KnoX, Apache Ranger oder Sentry können über diese Konsolen aufgerufen und konfiguriert werden.

Ohne diese genannten Frameworks ist eine zentrale Security-Konfiguration und Überwachung nur sehr mühselig umzusetzen.

Im Vortrag vergleichen wir die 3 Frameworks noch detaillierter und zeigen ihre Unterschiede auf.

Fazit
Security und Hadoop sind keine Gegensätze mehr. Mit den neuen Frameworks und insbesondere der Erkenntnis bei den Entwicklern, dass Security ein zentraler Punkt für die Anwender von Hadoop ist, lässt sich eine sichere Hadoop-Umgebung aufsetzen.
Da Hadoop ein komplette Ökosystem ist und nicht ein “Tool”, wird die Verwaltung immer komplexer bleiben, als bei einer Datenbank.

Kontaktadresse:
Oliver Gehlert
Ventum Consulting GmbH & Co KG
Infanteriestraße 11a
D-80797 München

Telefon:	+49 (0) 89-122 219 64 - 2
Fax:	+49 (0) 122 219 64 - 25
E-Mail	oliver.gehlert@ventum.de
Internet:	www.ventum.de