Secure Your Hadoop Cluster With Apache Sentry (Incubating)

Xuefu Zhang | Software Engineer, Cloudera
April 07, 2014
Outline

• Introduction
• Hadoop security primer
  • Authentication
  • Authorization
  • Data Protection
  • Governance and Auditing
• Introducing Apache Sentry
  • What's Sentry
  • Sentry Architecture
  • Sentry Internal
• Future work
• Q&A
Introduction

• **Hadoop gets bigger ...**
  • Hadoop has been enjoying an increasing adoption rate
  • More and more data on Hadoop Cluster
  • More and more access to the data
  • Data warehouse offload is the most common use case
  • Apache Hive, Apache Drill, Cloudera Impala
  • SQL on Hadoop is phenomenon
Introduction (cont'd)

• But more encumbrance ...
  • Enterprises wants to protect sensitive data
  • Government regulations, compliance, like HIPPA, PII, FISMA
  • Existing security problems with Hadoop has hindered the adoption
  • Security has become the top priority
Introduction (cont'd)

• Reality is
  • Different components, different security mechanisms
  • Multiple components may access the same dataset
  • Hadoop was born out of trust, not security
  • Thinking of Windows
Outline

• Introduction
• **Hadoop security primer**
  • Authentication
  • Authorization
  • Data Protection
  • Governance and Auditing
• Introducing Apache Sentry
  • What's Sentry
  • Sentry Architecture
  • Sentry Internal
• Future work
• Q&A
Hadoop Security Primer

• Authentication
  • Identify *who* you are
  • Untrusted users has no access to the cluster network
  • Trusted network, every one is good citizen
  • Who you are is determined by client host
Hadoop Security Primer

• Strong Authentication
  • Kerberos
  • LDAP, ActiveDirectory
  • LDAP, AD integrated with Kerberos, establishing a single point of truth
  • Single point of truth
Hadoop Security Primer (cont'd)

- Kerberos
  - Strong authentication
  - Provides mutual authentication
  - Protects against eavesdropping and replay attacks
  - Every user and service has a Kerberos “principal”
  - Credentials: keytabs (service), password (user)
Hadoop Security Primer (cont'd)

• Authorization
  • HDFS Posix style permission R/W/E for O/G/O, coarse-grained
  • Other components have authorization
  • MR job queue
  • HBase ACLs on table and column family.
  • Accumulo provides cell-level access control
  • Impersonation
Hadoop Security Primer (cont'd)

• Data Protection
  • Data at rest and in transit
  • Hadoop provides encryption on data in transit: DTP, HTTP, RPC, JDBC/ODBC
  • Hadoop has no native encryption on data at rest
  • Relying on OS-level encryption
Hadoop Security Primer (cont'd)

• Governance and auditing
  • Again, component to component
  • DFS and MapReduce provide base audit support
  • Apache Hive metastore records audit (who/when) information for Hive interactions.
  • Apache Oozie provides audit trail for services
Outline

• Introduction
• Hadoop security primer
  • Authentication
  • Authorization
  • Data Protection
  • Governance and Auditing
• Introducing Apache Sentry
  • What's Sentry
  • Sentry Architecture
  • Sentry Internal
• Future work
• Q&A
Introducing Apache Sentry

• Hadoop Authorization
  • Existing authorization is fragmented, coarse-grained, and manual
  • A lot of times data is just unprotected for simplicity
  • Enterprises need a centralized authorization component that work across components with ease of use, fine-grained, role based
Introducing Apache Sentry (cont'd)

• What's Sentry
  • Sentry is an authorization module for Hive, Search, Impala, and beyond
  • It unlocks Key RBAC Requirements: secure, fine-grained, role-based authorization, multi-tenant administration
  • Open Source, Apache Incubator project
  • Ecosystem Support: Apache SOLR, HiveServer2, & Impala 1.1+
Introducing Apache Sentry (cont'd)

• Key Benefits
  • Store Sensitive Data in Hadoop
  • Extend Hadoop to More Users
  • Comply with Regulations
Introducing Apache Sentry (cont'd)

• Key Capabilities
  • Fine-Grained: SERVERS, DATABASES, TABLES & VIEWS; INDEXES, COLLECTIONS
  • Role-Based: role including privileges such as SELECT, INSERT, ALL; UPDATE, QUERY
  • Multi-Tenant administration
  • Separate policies for each database/schema
  • Can be maintained by separate admins
Introducing Apache Sentry (cont'd)

Sentry Architecture

- Binding Layer:
  - Impala
  - Hive
  - Search
  - Pig
  - SOLR

- Authorization Provider:
  - Policy Engine
  - Policy Provider
  - File
  - Database
  - Local FS/HDFS

- Impala
- HiveServ er2

Cloudera®
Ask Bigger Questions
Introducing Apache Sentry (cont'd)

1. Parse SQL
2. Build
3. Check
   - First check: Authorization
4. Plan
5. Validate SQL grammar
6. Construct statement tree
7. Validate statement objects
   - *First check*: Authorization
8. Forward to execution planner
Introducing Apache Sentry (cont'd)

• Actors
  • User
  • User group membership
  • Resources
  • Privilege
  • Role
Introducing Apache Sentry (cont'd)

• User
  • User authenticated
  • User identity obtained from session context
Introducing Apache Sentry (cont'd)

• User group membership
  • Defined outside sentry policy
  • Obtained from user directory (LDAP, AD, HDFS)
  • Maybe available from session context
Introducing Apache Sentry (cont'd)

• Resources
  • Data to be protected
  • File or directory on HDFS
  • Table or views in Hive
  • URI
  • Resource can be hierarchical
Introducing Apache Sentry (cont'd)

• Privilege
  • Action or operation on a resource
  • Exists in a role only
  • SELECT on a given TABLE or VIEW
  • CREATE a TABLE or VIEW
  • QUERY on a search COLLECTION
  • DELETE a FILE or DIRECTORY
  • Example

    collection=customerCol->action=query
Introducing Apache Sentry (cont'd)

• Roles
  • A collection of privileges
  • Defined in Sentry policy
  • Example

[roles]
ana_query_role = collection=sentryColl->action=query
ana_update_role = collection=sentryColl->action=update
test_role = collection=testColl->action=update
full_admin_role = collection=*
Introducing Apache Sentry (cont'd)

• (Group, Role) mapping
  • Defined in policy
  • One-to-Many
  • Example

[groups]
analysts = ana_query_role, ana_update_role
admins = full_admin_role
testgroup = test_role
hbase = full_admin_role
Introducing Apache Sentry (cont'd)

• Rule evaluation
  • Who's the user?
  • Which group(s) does the user belong to?
  • What resource to be accessed?
  • How the resource is accessed (READ, SELECT, etc.)?
  • Does any of the user's groups have a role, which has the right privilege?
    • Yes – great! Go head!
    • No – sorry! No sufficient privilege!
Outline

• Introduction
• Hadoop security primer
  • Authentication
  • Authorization
  • Data Protection
  • Governance and Auditing
• Introducing Apache Sentry
  • What's Sentry
  • Sentry Architecture
  • Sentry Internal
• Future work
• Q&A
Future Work

• Introduce Sentry to more Hadoop components for their authorization needs
• Centralized policy store aiming for the whole enterprise
• Grant/Revoke
• Centralized authorization service for all protected resources including metadata

• We need your contribution or support