Apache Sentry

Prasad Mujumdar
prasadm@apache.org
prasadm@cloudera.com
Agenda

- Various aspects of data security
- Apache Sentry for authorization
- Key concepts of Apache Sentry
- Sentry features
- Sentry architecture
- Integration with Hadoop ecosystem
- Sentry administration
- Future plans
- Demo
- Questions
Who am I

• Software engineer at Cloudera
• Committer and PPMC member of Apache Sentry
• also for Apache Hive and Apache Flume
• Part of the the original team that started Sentry work
Aspects of security

Perimeter
- Authentication
  - Kerberos, LDAP/AD

Access
- Authorization
  - what user can do with data

Visibility
- Audit, Lineage
  - data origin, usage

Data
- Encryption, Masking
Data access

- Provide user access to data
- Manage access policies
- Provide role based access
Agenda

● Various aspects of data security
● Apache Sentry for authorization
● Key concepts of Apache Sentry
● Sentry features
● Sentry architecture
● Integration with Hadoop ecosystem
● Sentry administration
● Future plans
● Demo
● Questions
Apache Sentry (Incubating)

**Unified Authorization module for Hadoop**

Unlocks Key RBAC Requirements

- Secure, fine-grained, role-based authorization
- Multi-tenant administration

Enforce a common set of policies across multiple data access path in Hadoop.
Key Capabilities of Sentry

Fine-Grained Authorization
Permissions on object hierarchies. E.g., Database, Table, Columns

Role-Based Authorization
Support for role templates to manage authorization for a large set of users and data objects

Multi-Tenant Administration
Ability to delegate admin responsibilities for a subset of resources
Project history and status

• Started at Cloudera

• Entered incubation in 2013

• Growing community
  • Committers from Cloudera, IBM, Intel, Oracle, ...
  • Three releases from incubation

• Widely adopted by industry
  • Part of multiple commercial Hadoop distros
Agenda

● Various aspects of data security
● Apache Sentry for authorization
● Key concepts of Apache Sentry
● Sentry features
● Sentry architecture
● Integration with Hadoop ecosystem
● Sentry administration
● Future plans
● Demo
● Questions
Key Concepts in Sentry

• Global concepts
  • User, Group, Role, Privilege

• Authorization Models
  • SQL
    • Server, Database, Table, URI
  • Search Model
    • Collection
Global Concept: User

• Individual person
  • Runs SQL, SOLR queries

• Identified by authentication provider
  • Kerberos, LDAP etc

• Just a string for Sentry
  • Not enforcing existence
  • Sentry is NOT an authentication system
Global Concept: Group

• Set of users
  • Same needs/privileges

• Plugable group mapping
  • Using Hadoop Groups
  • OS, LDAP, Active Directory
Global Concept: Privilege

- Unit of data access
- Tuple
  - Object
  - Action
- Always positive

- READ TABLE logs
- READ DATABASE prod
- WRITE and READ TABLE logs
- QUERY COLLECTION logs
- UPDATE COLLECTION admin
Global Concept: Role

• Set of privileges
  • Functional template

• Unit of grant
Global Concepts: Relations

- Groups have multiple users
- Role have multiple privileges
- Roles are assigned to groups
  - Sentry does not support direct grants to user
- No jumping
  - User to role, group to privilege, ...

User — Group — Role — Privilege
Agenda

- Various aspects of data security
- Apache Sentry for authorization
- Key concepts of Apache Sentry
- Sentry features
- Sentry architecture
- Integration with Hadoop ecosystem
- Sentry administration
- Future plans
- Demo
- Questions
Sentry features – Fine grain authorization

• Privileges at various levels for resource hierarch
  • Eg Database, Table and Column for SQL model
  • Read or Select access on Database implicitly grant access on child tables

• Supports different actions on resources
  • Eg, Select, Insert, Create, Alter in the SQL model
  • Query, Update in Search model
  • …
Sentry Features – Role based

• Supports Role as collection of permission
  • Template for a functional access rules
  • Eg, Analyst role → Read table sales, Read table customer, Admin of Sandbox
  • Makes auth administration manageable in large and complex deployments

• Allows granting roles to groups
  • A role can be granted to a large set of users in a single operation
  • Easier integration with existing identity management systems like AD

• Onboarding and removing users is lot simpler with roles and groups
Sentry features – misc

- Multi Tenant administration
  - Ability to delegate admin access for a subset of resources
  - Eg. A user can be an admin of his/her own sandbox database

- Plugable architecture
  - A new authorization model can be implemented with little code changes
  - Can easily integration with new identity management systems for groups
  - Supports various callbacks for custom monitoring
Agenda

• Various aspects of data security
• Apache Sentry for authorization
• Key concepts of Apache Sentry
• Sentry features
• Sentry architecture
• Integration with Hadoop ecosystem
• Sentry administration
• Future plans
• Demo
• Questions
Apache Sentry conceptual overview
Apache Sentry conceptual overview

• Policy Provider
  • Abstraction for loading and manipulating privilege metadata
  • Support for external DB backed storage (default)
  • Also support local or HDFS file storage (deprecated)

• Policy Engine
  • Makes the authorization decision
  • Reads the metadata from policy provider

• Binding
  • Bridging layer between the downstream service and Sentry
  • Handles translating the native access request into Sentry APIs
Sentry Service Architecture

- Data Engine, eg Hive
- Sentry plugin
- Sentry RPC server
- Policy metadata store
Sentry Service

- RPC Service to manage metadata
  - Apache Thrift RCP implementation
  - Java client
  - Secured with kerberos
- API to retrieve and manipulate policies
- Metadata stored in external backend DB
  - Supports Derby, MySQL, Postgres, Oracle and DB2
Sentry Service HA

ZooKeeper

Sentry Service Client

Sentry Policy Store | Config
---|---
Sentry Service

Sentry Policy Store | Config
---|---
Sentry Service

Sentry Policy Store | Config
---|---
Sentry Service

DB
Sentry Service HA

- Active/Active HA
- Each service registers with ZK
- Client first retrieves service address for ZK
- User Apache Curator framework
File based privileged metadata

- Policy information can be stored in local or HDFS files
- Deprecated in newer releases in favor of DB based policies
- ini format property file

```ini
# group to role mapping
[groups]
manager = analyst_role, junior_analyst_role
analyst = analyst_role
admin = admin_role

# role to privilege mapping
[roles]
analyst_role = server=server1->db=analyst1, 
server=server1->db=jranalyst1->table=*->action=select, 
server=server1->db=default->table=tab2
```
Sentry Client Plugin

● Client side piece of Sentry
  ● Integrates via the authorization interfaces
● Responsible for authorization decision
  ● Receives requested resources and user from caller
  ● Retrieves relevant privileges from Sentry service
● Evaluates the request
Auditing

- Sentry service generates audit trails
- Policy changes are audited
  - Eg granting privileges, create/drop roles
- Audit JSON format audit log
  - Easier for processing by audit reporting tools
- Client side auditing handled by client’s auth auditing mechanism
  - Eg Hive and Impala
  - Sentry support client callbacks which can be used for customization
Agenda

- Various aspects of data security
- Apache Sentry for authorization
- Key concepts of Apache Sentry
- Sentry features
- Sentry architecture
- Integration with Hadoop ecosystem
- Sentry administration
- Future plans
- Demo
- Questions
Integration with Hadoop Ecosystem

- Hive Server2
- Hive Metastore
- Impala Catalog
- Impala
- NameNode
- HDFS
- sqoop2
- HDFS Policy Meta data
- Policy Meta data
- Audit Trail
- Group Mapping
- Authentication
- Active Directory
Unified authorization for Hadoop ecosystem

• Single source of truth

• Other projects don’t have to implement their own auth

• Same set of roles and group available across tools
  • Makes the authorization administration lot simpler

• Same privileges enforced irrespective of the access path
SQL on Hadoop

- Data on HDFS, owned by Hive
- Metadata in Hive Metastore
- Auth policies in Sentry

Diagram:

- Hadoop
  - Hive Server2
  - Impala
- Hive Metastore
  - Schema Metadata
  - Privileges
- Sentry
- Data

MR, Yarn, ..
Sentry with Apache Hive

SQL
Parse
Build
Check
Sentry
Plan
MR
Query

Validates access to SQL entities before executing the query.
Sentry with Apache Hive

- Requires HiveServer2, not supported with thick hive client
- SQL model with fine grained authorization
  - DB objects - Database, Table and Column
  - DB Actions - SELECT, INSERT, CREATE, ALTER, ..
- Support managed and external tables
  - Special handling of external path specification via URI level privilege
- Authorization administration via SQL
  - grant, revoke, create/drop role etc.
Sentry with Impala

Validates access to SQL entities before executing the query using the cached privileges.
Sentry with Impala

- Uses same SQL model as Hive
- Fine grained authorization
  - DB objects - Database, Table, Column
  - DB Actions - SELECT, INSERT, CREATE, ALTER, ..
- Support managed and external tables
  - Special handling of external path specification via URI level privilege
- Impala engine caches privilege metadata for faster access
View level privileges for SQL authorization

• Views are essentially queries defined on one or more tables
  • Eg CREATE VIEW v1 AS SELECT tab1.col1, tab2.col2 FROM tab1, tab2 ...

• Privileges on views are independent of the base tables

• This enables row/cell level privileges

• Requires data files to be owned and access by Hive user
URI level privilege

• Hive SQL supports file URI leading to security loopholes
  • Alternate storage path for tables
    • Create table
    • Alter table
    • External table
  • One can specify the path of a different table and bypass authorization
    • `ALTER TABLE sandbox.sales SET LOCATION '/user/hive/warehouse/production/sales'`

• Hive UDFs using jars with untrusted/unauthorized static code

• URI resource privilege to can be used to prevent this
  • A file URI can only be used if you have explicit grant to use it
Sentry with Metastore

Metastore RPC clients can read/write metadata directly. Sentry enforces the same privileges on metadata.
Sentry with Metastore

• Enforces the same policies for metadata access
• Prevents unauthorized schema changes
• Hides metadata from unauthorized users
• Works for all Metastore RPC clients
  • Apache Pig with Hcatalog
  • Hadoop jobs
  • Third party applications
HDFS applies Sentry privileges as ACLs for files/directories that are part of Hive data to enable non-SQL clients.
HDFS ACL sync for non-sql clients

- Apply Sentry privileges as HDFS ACLs
  - Requires HDFS extended ACLs enabled
- Namenode maintains a cache of privileges
- Currently supported for Hive data only
- Enables same granularity of access to files for non-sql clients
- *Hadoop side changes is recently committed only available in trunk*
Sentry with Apache Solr

Validate access to Solr collection and documents.
Sentry with Apache Solr

- Fine grained authorization
  - Collection
  - Documents
  - Index
- Support query and update access on the resources
Sentry with Apache Sqoop

- Authorization of various sqoop resources
  - connector, link, jobs
- Fine grained authorization of actions
  - Create, Enable, Start/Stop, List etc.
- Under development
  - SENTRY-612 being reviewed
Agenda

- Various aspects of data security
- Apache Sentry for authorization
- Key concepts of Apache Sentry
- Sentry features
- Sentry architecture
- Integration with Hadoop ecosystem
- Sentry administration
- Future plans
- Demo
- Questions
Sentry Administration

● Privileges managed natively by downstream app
  ● Auth SQL statements
  ● Application APIs

● Hue UI
  ● Sentry App for policy administration

● Plugable groups mapping
  ● By default same as Hadoop (OS or LDAP/AD)
Sentry App in Hue
Sentry App in Hue
Setting up Sentry in Hadoop cluster

• Should have strong authentication like Kerberos or LDAP

• Setup sentry service
  • Setup metadata DB
  • Configure and run the service

• Setup data services to use Sentry
  • Configure auth plugins
  • Setup sentry client configuration to use sentry service

• Create roles and privileges
  • Hue UI app is super useful
Agenda

- Various aspects of data security
- Apache Sentry for authorization
- Key concepts of Apache Sentry
- Sentry features
- Sentry architecture
- Integration with Hadoop ecosystem
- Sentry administration
- Future plans
- Demo
- Questions
Future plans

• Integration with more Hadoop ecosystem components
  • Hbase
  • Kafka, Flume ..

• Attribute based access control

• Row/cell level authorization for HDFS
References

• Project page
  • https://sentry.incubator.apache.org

• Wiki
  • https://cwiki.apache.org/confluence/display/SENTRY/Home

• Source
  • git clone http://git-wip-us.apache.org/repos/asf/incubator-sentry.git

• Downloads
  • https://sentry.incubator.apache.org/general/downloads.html

• Jira
  • https://issues.apache.org/jira/browse/Sentry

• How to contribute
  • https://cwiki.apache.org/confluence/display/SENTRY/How+to+Contribute

• Mailing list
  • dev@sentry.incubator.apache.org
Demo Time

Thank You!

Contributions are welcome!