DTCC

第十四届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA

数智赋能 共筑未来

IT168.com ChinaUnix ITPUB

北京国际会议中心 | 2023/8/16-18

合规下的B站大数据安全

bilibili 大数据安全负责人
郭跃鹏

About me

- Apache Griffin PMC Chair
- 历任eBay数据架构师
- 历任IBM，HP企业解决方案架构师

日程

- 数据安全背景
- 解决方案
- 安全收益
- Take away
- Q&A

法律法规背景

- 数据安全法
- 个人信息保护法
- 网络安全法
- SOX审计法案
- 等保
- 本地法律法规
- 公司规定

法律法规背景

第三条　本法所称数据，是指任何以电子或者其他方式对信息的记录。

数据处理，包括数据的收集、存储、使用、加工、传输、提供、公开等。

数据安全，是指通过采取必要措施，确保数据处于有效保护和合法利用的状态，以及具备保障持续安全状态的能力。

重要数据的处理者应当明确数据安全负责人和管理机构，落实数据安全保护责任。

解决方案

分类分级

数据发现 → 存储加密 → 访问控制 → 审计控制 → 风险控制

KMS

Kerberos

异常检测

分类分级

✓分类分级的管理

- C4 绝密数据
- C3 机密数据
- C2 内部数据
- C1 公开数据

敏感字段类型	类目名称	安全等级	规则类型	启用状态	操作
手机号	个人基本资料	C4	自定义规则	已生效	
邮件地址	个人基本资料	C4	自定义规则	已生效	
真实姓名	个人基本资料	C4	自定义规则	已生效	
生日年龄	个人基本资料	C4	自定义规则	已生效	
性别	个人基本资料	C4	自定义规则	已生效	
民族	个人基本资料	C4	自定义规则	已生效	

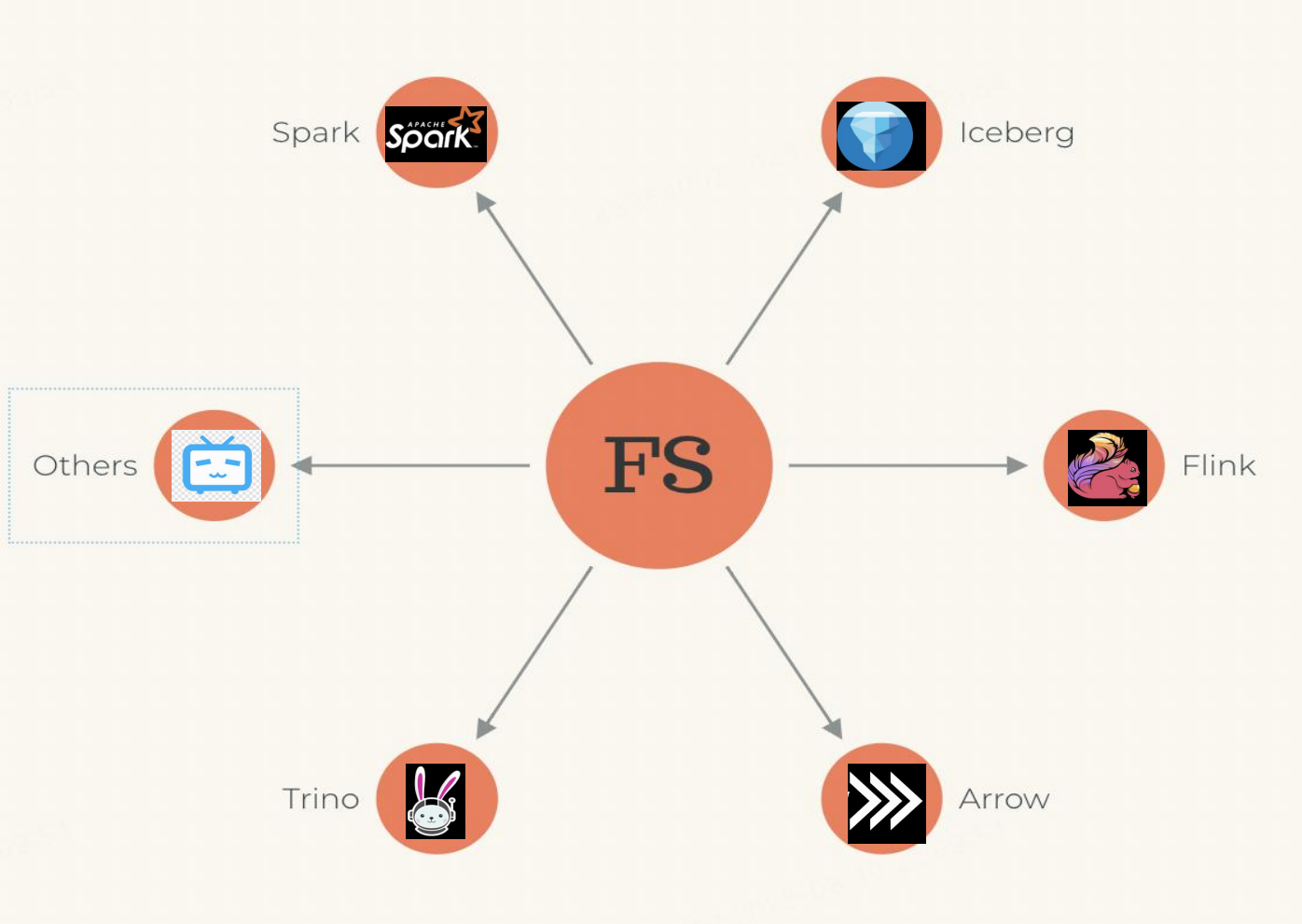
分类分级

- 基于规则的识别算法
- 基于NLP的文本分类算法
- 人工标注
- 人工审核

数据发现

- 数据资产
  - 静态资产
  - 流动资产

存储加密



存储加密

- 来自于分类分级的存储要求细则
- 我司C2 及以上的数据需要加密存储。
- 相应的访问控制，需要通过相应的key来访问。

Name	Age	City
C4	C4	C1
[redacted]	28	上海
[redacted]	22	北京
[redacted]g	35	成都
[redacted]ai	30	深圳
[redacted]gchao	14	深圳

Name/Age加密

City 不加密

```
00000000  50 41 52 31 80 00 00 00  28 d3 00 f6 5a 8a 1c 9d  |PAR1....(...Z...|
00000010  42 e0 c1 30 a1 22 bd ae  bf 72 3f 6e 83 88 ad 03  |B..0."...r?n....|
00000020  de 8e 17 bf 57 9c 60 1a  4f 39 40 ab e6 64 fe cc  |....W.`.O9@..d..|
00000030  7e dc 98 9a ad 3b 96 53  40 97 67 92 a0 a1 0e 14  |~....;.S@.g.....|
00000040  04 99 bd 85 95 55 14 82  51 d6 04 81 81 3c 70 f1  |.....U..Q....<p.|
00000050  72 d8 86 bd b3 bd 08 e2  58 e3 49 c6 3b ec cb d0  |r.......X.I.;...|
00000060  4f 2c f6 2e 38 10 7e 24  c2 26 7b 14 c1 70 f2 cd  |O,..8.~$.&{..p..|
00000070  ee 18 c7 6e d4 8c 3b d1  fd ef 73 32 38 0d 63 2b  |...n..;...s28.c+|
00000080  4f 2e d6 c0 dc 8f 24 8d  57 00 00 00 6c e6 c1 4f  |O.....$.W...l..O|
00000090  8d 61 fe 38 46 61 b7 17  73 49 ca f3 e3 72 6a 02  |.a.8Fa..sI...rj.|
000000a0  e5 3d 84 af 41 6c 35 63  b5 1a d1 6e 89 73 d0 e3  |.=..Al5c...n.s..|
000000b0  e9 16 38 76 d0 18 c7 f1  dd 3a 69 cd 76 b9 e5 31  |..8v.....:i.v..1|
000000c0  39 7a 8e f7 58 60 79 0b  83 64 64 bd a3 c2 32 02  |9z..X`y..dd...2.|
000000d0  e0 fe c5 3c 6f a0 39 c4  d7 9d 4e ec 01 98 7f 33  |...<o.9...N....3|
000000e0  90 cd 95 80 00 00 00 94  3a 1e 8c 15 83 22 20 70  |........:....." p|
000000f0  c7 2a 02 5f 33 44 d7 a4  30 37 81 00 68 c7 df 7f  |.*._3D..07..h...|
00000100  41 5f 18 80 9d b9 53 4e  3e 66 c9 df d1 86 77 33  |A_....SN>f....w3|
00000110  74 2e 61 a9 4a 49 38 24  0e c9 27 be 2a 12 99 8d  |t.a.JI8$..'.*...|
00000120  f7 36 e0 d6 26 b8 7d 57  a7 7b 0a f3 d7 be 9f 1f  |.6..&.}W.{......|
00000130  cb e5 f6 12 a3 cd 14 ac  6d 51 f8 ef 74 87 b3 9a  |........mQ..t...|
00000140  63 bf c2 8f 77 b1 71 12  6e 54 84 56 d9 5b 55 17  |c...w.q.nT.V.[U.|
00000150  01 c9 62 6a 2f f4 db 95  ce 06 49 d1 6c 57 5d b3  |..bj/.....I.lW].|
00000160  5f e1 6e df 97 1f 76 38  00 00 00 8e 93 5e ba 4f  |_.n...v8.....^.O|
00000170  9d 66 9b 2d 95 fc fc 0e  4e 18 6a 6a e0 74 ab db  |.f.-....N.jj.t..|
00000180  14 e7 74 f4 e5 8f b5 35  5e 52 24 ef 2a 2c f5 28  |..t....5^R$.*,.(|
00000190  45 58 e7 45 20 51 9b 06  01 23 80 e0 ca 81 7c 1a  |EX.E Q...#....|.|
000001a0  35 3a 76 15 04 15 50 15  50 15 d3 9b b8 f5 06 3c  |5:v...P.P......<|
000001b0  15 08 15 04 00 00 28 24  06 00 00 00 e4 b8 8a e6  |......($........|
000001c0  b5 b7 01 0a 14 e5 8c 97  e4 ba ac 01 0a 3c e6 88  |.............<..|
```

存储加密 – 访问权限控制

➢没有加密列权限的用户访问加密列

- 没有加密列权限的用户访问非加密列
- 有加密列权限的用户访问加密列

```
scala> val result = spark.sql("SELECT City  FROM my_table WHERE Age > 25")
result: org.apache.spark.sql.DataFrame = [City: string]

scala> result.show
23/08/08 18:53:13 ERROR Executor: Exception in task 0.0 in stage 14.0 (TID 13)
org.apache.parquet.crypto.ParquetCryptoRuntimeException: [Age]. Null File Decryptor
        at org.apache.parquet.hadoop.metadata.EncryptedColumnChunkMetaData.decryptIf
        at org.apache.parquet.hadoop.metadata.EncryptedColumnChunkMetaData.getStatis
        at org.apache.parquet.filter2.statisticslevel.StatisticsFilter.visit(Statist
        at org.apache.parquet.filter2.statisticslevel.StatisticsFilter.visit(Statist
        at org.apache.parquet.filter2.predicate.Operators$NotEq.accept(Operators.jav
```

存储加密 – 访问权限控制

- 没有加密列权限的用户访问加密列

➢没有加密列权限的用户访问非加密列

- 有加密列权限的用户访问加密列

```
scala> val result = spark.sql("SELECT City  FROM my_table")
result: org.apache.spark.sql.DataFrame = [City: string]

scala> result.show
+----+
|City|
+----+
|上海|
|北京|
|成都|
|深圳|
|深圳|
```

存储加密 – 访问权限控制

- 没有加密列权限的用户访问加密列
- 没有加密列权限的用户访问非加密列

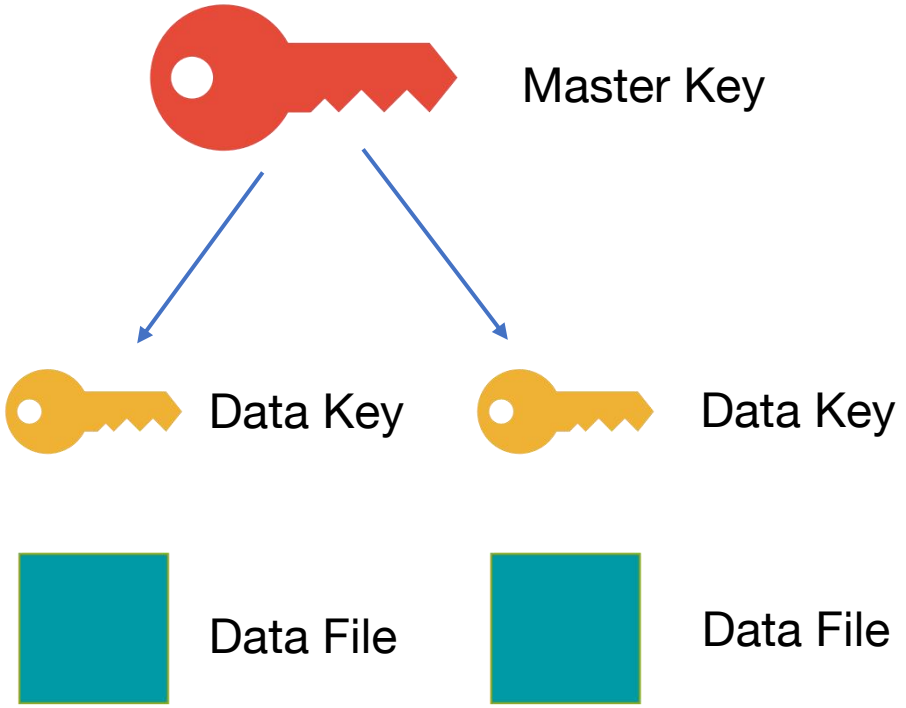
➢有机密权限的用户访问加密列

```
scala> val result = spark.sql("SELECT * FROM my_table")
result: org.apache.spark.sql.DataFrame = [Name: string, Age: int ... 1 more field]

scala> result.show
+-------+---+-----+
|   Name|Age| City|
+-------+---+-----+
|[redacted]| 28|上海|
|[redacted]| 22|北京|
|[redacted]| 35|成都|
|[redacted]| 30|深圳|
|[redacted]| 14|深圳|
```

存储加密 – Key

- Envelope encryption
- KMS
- Key TTL问题(key rotation)
- 权限与审计



访问控制 – 身份认证

- Kerberos
  - HA问题
  - 时钟同步问题

访问控制 – ABAC

- ABAC VS RBAC
  - RBAC:
    - 简单
    - 预定义预计算
    - 角色爆炸
    - 管理员瓶颈
  - ABAC:
    - 用户和访问内容分离
    - 运行时计算
    - 可扩展的策略管理

访问控制 – ABAC

- 资产
  - 分类
  - 分级
  - owner
  - 工作空间
  - 其他

账号
- 部门
  - 个人
- 工作空间
  - 服务账号
- 其他

访问控制 – Policy

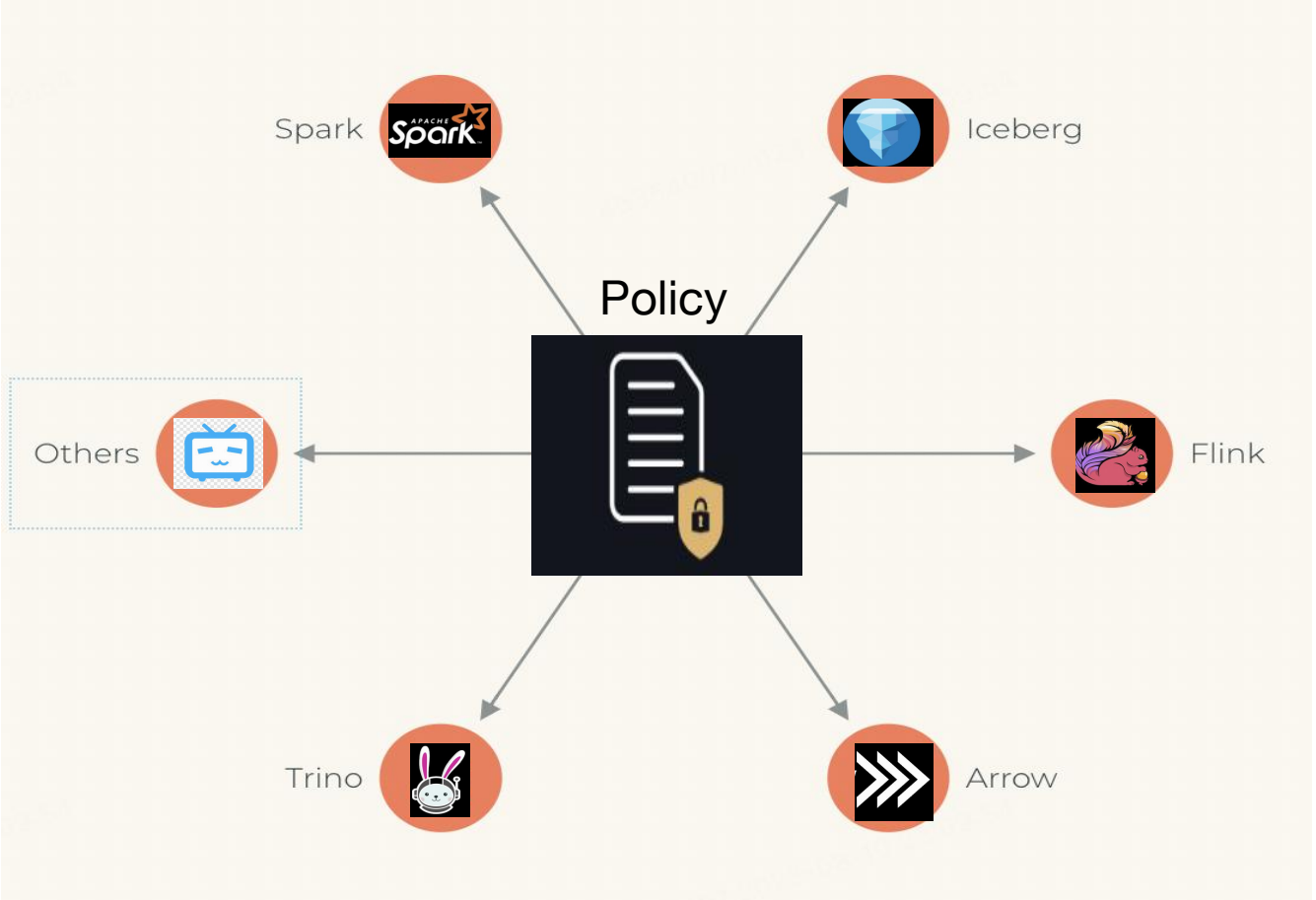
- Policy DSL
  - **`Allow access to a tables where user.BU is in table.BU`**
  - 行列权限

```
Allow access to a tables where user.BU is in table.BU
Filter accessible rows where Users.Region =
table.column.Region
```

访问控制 – Policy

- Policy 管理
  - ➢全局 Policy
    - ➢法律 Policy
    - ➢法规 Policy
  - ➢部门Policy
  - ➢其他Policy

- Define policy once, enforce everywhere.



审计控制

- Who
- Where
- When
- What
- Why

昵称	用户名	部门	敏感类型	访问次数	下载次数
[redacted]	[redacted]	[redacted]	邮件地址	3	0
[redacted]	[redacted]	[redacted]	邮件地址	6	0
[redacted]	[redacted]	[redacted]	邮件地址	10	5
[redacted]	[redacted]	[redacted]	邮件地址	18	9
[redacted]	[redacted]	[redacted]	邮件地址	1	0
[redacted]	[redacted]	[redacted]	邮件地址	2	0

访问时间	表	字段	sql	风险行为
2023-08-02 16:42:00	[redacted] trad_cntc [redacted]	order_email	select * from [redacted] ct_a_d [redacted] log_date = '20230801' and id = 2106	访问

风控

- 基于规则的风控
- 基于模型的异常行为风控

敏感数据审计　风险行为审计　用户行为审计

访问时间	风险行为	命中原因	sql	场景
2023-08-15 21:05:54	按mid查询	mi[redacted]	[redacted] ods.roo[redacted]y_du int_l_[redacted] here log_date = '20230805' and [redacted] and buvid = [redacted]326b 05[redacted] and...	adhoc
2023-08-15 21:01:49	按mid查询	mid	[redacted] _play_du [redacted] _hr where log_date = '20230805' anc[redacted] buvid [redacted]02 304[redacted]UkF[redacted]...	adhoc

安全收益

- 某公司罚被80.26亿元
- Meta被罚12亿欧
- 0 OR 1

Take away

- 分类分级为纲
- 存储加密为本
- 灵活的ABAC权限控制

Q&A

Q&A

请关注哔哩哔哩技术

THANKS