Outline

• Introduction to Sqrrl and Accumulo
• Security In The Wild
• Sqrrl and Accumulo Technology
• The Data-Centric Security Ecosystem
• How to Learn More
Poll

• What is your level of familiarity with Accumulo?
  a. Only heard of it
  b. Well read on it
  c. Downloaded and experimented with it
  d. Currently using it
Who We Are

Google’s BigTable Paper 2006
NSA Builds Accumulo 2008
NSA Open Sources Accumulo 2011
Sqrrl Founded 2012
3rd Sqrrl Release September 2013

Investors

ATLAS VENTURE
MATRIX PARTNERS
The Value of Sqrrl and Accumulo

Security

Sqrrl Enterprise
and Apache Accumulo

Scalability

Adaptivity
Outline

• Introduction to Sqrrl and Accumulo
  • Security In The Wild
• Sqrrl and Accumulo Technology
• The Data-Centric Security Ecosystem
• How to Learn More
Secure Data Lake Concept

Non-sensitive data
Sensitive data
Highly sensitive data
Sensitive data
Non-sensitive data
Non-sensitive data

Real-Time Apps

Sqrrl Enterprise
Healthcare Security Requirements

Sensitive Diagnoses
Doctor’s Notes
Outline

• Introduction to Sqrrl and Accumulo
• Security In The Wild
  • Sqrrl and Accumulo Technology
• The Data-Centric Security Ecosystem
• How to Learn More
Sqrrl Enterprise Built on Apache Accumulo

Bulk Processing Integration

Exploratory / Operational Apps

Graph + Document I/O

Sqrrl proprietary
Automated indexing
Custom iterators
Lucene integration
Security extensions

Sqrrl API over Apache Thrift RPC
(JSON, Graph, Aggregation, Search, etc.)

Accumulo RPC
(Sorted Key/Value I/O)

Hadoop RPC
(File I/O)

Open source (including Sqrrl contributions)

Open source or commercial distributions

Sqrrl Data, Inc. All Rights Reserved
# Accumulo Data Format

**An Accumulo key is a 5-tuple, consisting of:**

- **Row:** Controls Atomicity
- **Column Family:** Controls Locality
- **Column Qualifier:** Controls Uniqueness
- **Visibility Label:** Controls Access
- **Timestamp:** Controls Versioning

<table>
<thead>
<tr>
<th>Row</th>
<th>Col. Fam.</th>
<th>Col. Qual.</th>
<th>Visibility</th>
<th>Timestamp</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Doe</td>
<td>Notes</td>
<td>PCP</td>
<td>PCP_JD</td>
<td>20120912</td>
<td>Patient suffers from an acute ...</td>
</tr>
<tr>
<td>John Doe</td>
<td>Test Results</td>
<td>Cholesterol</td>
<td>JD</td>
<td>PCP_JD</td>
<td>20120912</td>
</tr>
<tr>
<td>John Doe</td>
<td>Test Results</td>
<td>Mental Health</td>
<td>JD</td>
<td>PSYCH_JD</td>
<td>20120801</td>
</tr>
<tr>
<td>John Doe</td>
<td>Test Results</td>
<td>X-Ray</td>
<td>JD</td>
<td>PHYS_JD</td>
<td>20120513</td>
</tr>
</tbody>
</table>

**Accumulo Key/Value Example**
Accumulo Technology

Tablet Data Flow

- In-Memory Map
- Iterator Tree
- Sorted, Indexed File
- Write Ahead Log (For Recovery)
- Sorted, Indexed File
- Table Data Flow
  - Iterator Tree
  - Sorted, Indexed File
  - Merging / Major Compaction

Tablet States

- Assigned
- Defined
- Splitting
- Splitting 1
- Splitting 2
- Split
- Offline
- Failed
- Deleted

Fate States

- Start Op
- Seed Op
- Wait For Op
- Delete Op
- New
- In Progress
- Failed
- Executing Top Op
- Save Exceptions
- Mark Success
- Mark Failed
- Mark Failure
- Mark Failure In Progress
- Entry Points

Zookeeper

Master

HDFS

Application

Delegation Authority

Tablet Server

Tablet

Delegate Authority

Assign/Balance

Read/Write

Sqrrl Data, Inc. All Rights Reserved
Table Design Patterns

**Forward Index**
- Row: <UUID>
- Column Family: <Type>
- Column Qualifier: <Field>
- Value: <Term>

**Inverted Index**
- Row: <Term>
- Column Family: <Type> + <Field>
- Column Qualifier: <UUID>
- Value: <Digest of Event>

**Ingest Process**
- Map UUID to Event
- Find UUIDs and Digests for Term
- Map Terms to UUID
- Inverted Index

**Query Process**
- Event
- Term

**Shard Table**
- Row: <Partition ID>
- Column Family: "Docs" -> "Inv. Index" -> "Field Index" -> "Geo"
- Column Qualifier (Tuples): <Field> -> <UUID> -> <Term> -> <Field:Term> -> <Hash> -> <uuid>
- Value: <Value>

**Ingest Process**
- Event Columns
- Index Columns

**Query Process**
- Event Columns
- Index Columns
Outline

• Introduction to Sqrrl and Accumulo
• Security In The Wild
• Sqrrl and Accumulo Technology
  • The Data-Centric Security Ecosystem
• How to Learn More
**Data-Centric Security**

**Definition:** Data carries with it information that is required to make policy decisions on its releasability.

<table>
<thead>
<tr>
<th>Row</th>
<th>Col</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Name</td>
<td>Anon1</td>
</tr>
<tr>
<td>1</td>
<td>Sales</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>Name</td>
<td>Smith</td>
</tr>
<tr>
<td>2</td>
<td>Sales</td>
<td>350</td>
</tr>
<tr>
<td>2</td>
<td>Age</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>Quota</td>
<td>1000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Row</th>
<th>Col</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Name</td>
<td>Jones</td>
</tr>
<tr>
<td>1</td>
<td>Sales</td>
<td>100</td>
</tr>
<tr>
<td>1</td>
<td>Age</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>Name</td>
<td>Smith</td>
</tr>
<tr>
<td>2</td>
<td>Sales</td>
<td>350</td>
</tr>
<tr>
<td>2</td>
<td>Age</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>Quota</td>
<td>1000</td>
</tr>
</tbody>
</table>
Security for Transformed Data

Simple Provenance:
Row + Column Security OK

Complex Provenance:
Cell-Level Security Needed
## Security

### Example Accumulo Key/Value Pairs

<table>
<thead>
<tr>
<th>Row</th>
<th>Col. Fam.</th>
<th>Col. Qual.</th>
<th>Visibility</th>
<th>Timestamp</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Doe</td>
<td>Notes</td>
<td>PCP</td>
<td>PCP_JD</td>
<td>20120912</td>
<td>Patient suffers from an acute ...</td>
</tr>
<tr>
<td>John Doe</td>
<td>Test Results</td>
<td>Cholesterol</td>
<td>JD</td>
<td>PCP_JD</td>
<td>20120912</td>
</tr>
<tr>
<td>John Doe</td>
<td>Test Results</td>
<td>Mental Health</td>
<td>JD</td>
<td>PSYCH_JD</td>
<td>20120801</td>
</tr>
<tr>
<td>John Doe</td>
<td>Test Results</td>
<td>X-Ray</td>
<td>JD</td>
<td>PHYS_JD</td>
<td>20120513</td>
</tr>
</tbody>
</table>

*Accumulo is the only NoSQL database with cell-level access controls*
JSON Document Security Labels

- JSON maps to document and index entries, all of which preserve the security labels
- Labels follow the document hierarchy
- The label is part of the name of the field (affects uniqueness)
Data-Centric Security Ecosystem

Data → Key Mgmt → Sqrrl Enterprise → Apps → End Users

Data → Labeler → Sqrrl Enterprise → Audits → End Users

Policies → Sqrrl Enterprise → Auth. Service

User Attributes → Sqrrl Enterprise → Auth. Service
Outline

• Introduction to Sqrrl and Accumulo
• Security In The Wild
• Sqrrl and Accumulo Technology
• The Data-Centric Security Ecosystem
• How to Learn More
Accumulo Community Growth

http://accumulo.apache.org

Current Developer Base:
• 18 Committers
• 39 Additional Recognized Contributors

Contributing Organizations:
How To Learn More

• Download our White Paper
  – www.sqrrl.com/whitepaper

• Watch a video
  – www.sqrrl.com/downloads#videos

• Request a demo or one-on-one workshop
  – www.sqrrl.com/contact

• Come meet us
  – Strata Rx (Sept 25-27, Boston)
  – Hadoop World (Oct 28 – 30, New York)
Thank you

Thanks for attending!

To keep up to date with Sqrrl, check out or social media sites:

www.twitter.com/sqrrl_inc
www.linkedin.com/company/sqrrl