Agenda

- State of Big Data
- Company Background
- Problems We Solve
- How We Are Different
- Our Technology
- Use Cases
Hadoop is one of the most important trends in IT today
### There are some important advantages to Hadoop and NoSQL

<table>
<thead>
<tr>
<th>Flexible Schemas</th>
<th>Resilience + Durability</th>
<th>Cost Effectiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Easily add columns on the fly</td>
<td>• Data is triple replicated</td>
<td>• Designed to run on cheap commodity hardware</td>
</tr>
<tr>
<td>• No initial data modeling or ETL needed</td>
<td>• Failure of any regular node(s) has no impact</td>
<td>• Hardware costs scale linearly with data volumes</td>
</tr>
<tr>
<td>• Accommodates sparse datasets and datasets with evolving schema</td>
<td>• Proven ability to run mission critical applications</td>
<td>• Largely based on open source software</td>
</tr>
</tbody>
</table>
Agenda

• State of Big Data
  • Company Background
• Problems We Solve
• How We Are Different
• Our Technology
• Use Cases
This is what we do

Sqrrl Enterprise is the most secure and scalable platform for building real-time “Big Apps”
There are a lot of players in the Big Data ecosystem...

- **NoSQL / NewSQL**: DataStax, 10gen, Tungsten, Informatica
- **Horizontal Platforms**: Acunu, LucidWorks, Infochimps, Hadoop, Causata
- **Vertical Platforms**: kyruusu, Tresata, Splunk, Explorys
- **BI/Analytics Tools**: Tableau, MicroStrategy, Tableau, MicroStrategy
- **Security**: Gazzang, Virtru, DataGuise, Vormetric
- **Hadoop**: Cloudera, Intel, Hortonworks, EMC2, IBM
- **Data Integration**: Talend, Pentaho, Informatica
- **EDW / SQL**: Teradata Aster, Vertica, Netezza
- **Hardware**: Fusion-io, Dell, Cisco, EMC2, NetApp
- **Services**: Think Big Analytics, Deloitte, Booz Allen Hamilton
- **Cloud Providers**: Amazon Web Services, Eucalyptus, AppScale, GoGrid
...But Sqrrl has some unique capabilities

- Over 20 years of combined Accumulo experience
- Team includes former Technical Director of Accumulo at NSA and 6 committers/contributors

Investors

Google’s BigTable Paper 2006
NSA Builds Accumulo 2008
NSA Open Sources Accumulo 2011
Sqrrl Founded 2012
1st Sqrrl Release and Customers 2013
Agenda

• State of Big Data
• Company Background
• Problems We Solve
• How We Are Different
• Our Technology
• Use Cases
• Demo
Hadoop lacks real-time capabilities and some key enterprise features

February 2013 Survey of Hadoop Users

- Only 24% of Hadoop projects are in production
- However, half of those in production have more than 500 TB
- 37% cited lack of real-time capabilities as the biggest challenge (*top answer*)
- 26% cited the time it takes to reach production
Sqrrl brings secure, real-time apps to Hadoop

**Old Approach (hours to days)**

- **Raw Data**
- **ETL**
- **Cleansed Data**
- **Apps**

**New Approach (seconds)**

- **Raw Data**
- **Sqrrl Enterprise**
- **Real-Time Operational and Exploratory Analytics**
- **Applications**

*Millions of events per day*

*Millions of events per second*
Sqrrl supports both exploratory and operational apps

**Exploratory Apps**

*Seeking Unknown Patterns*

- Search
- Business Intelligence Tools
- Dashboards
- Visualizations
- Drill-Downs

**Operational Apps**

*Using Known Patterns*

- Fraud and Suspicious Behavior Detection
- Recommendations
- Personalization
- Price Setting
- Predictive Analysis
Our technology builds on a decade’s worth of Big Data lessons learned

- **Start small, but design for scalability**
  - One application first, then grow to hundreds
  - One gigabyte first, then grow to petabytes

- **Iterative schema refinement**
  - Initially, let the data define the schema
  - Refine the schema in bulk as you better understand the data
  - Middle ground between flat files and complete ontologies

- **Discovery analytics as application building blocks**
  - Universal search: structured and unstructured data, across data sets, low latency
  - Basic statistics: aggregations of query results, parallelized, low latency, to support big picture analysis
  - Graphs: scalable graph analytics for analyzing how everything is connected

- **Data-centric security**
  - Separate modeling of security and analysis
  - Simplifies multi-tenancy and application accreditation
Agenda

- State of Big Data
- Company Background
- Problems We Solve
- How We Are Different
- Our Technology
- Use Cases
Sqrrl extends open-source Accumulo

- Automated indexing
- SQL-like query language with full-text search and statistics capabilities
- Graph search
- Identity and access management and encryption plug-ins
- JSON documents
- Streaming ingest

- Open source / Hadoop integration
- Scalability to tens of petabytes
- Millions to billions of reads & writes per sec.
- Fine-grained access controls for multitenancy
- Flexible schemas
- Strong consistency
- Highly resilient to failure and extreme durability
- Scales elastically on commodity hardware
Sqrrl capabilities cut across the NoSQL landscape

<table>
<thead>
<tr>
<th>Type</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key-Value Store</td>
<td>redis, riak</td>
</tr>
<tr>
<td>Column Store</td>
<td>HBase, Cassandra</td>
</tr>
<tr>
<td>Document Store</td>
<td>mongoDB, CouchDB</td>
</tr>
<tr>
<td>Graph Store</td>
<td>Neo4j, the graph database</td>
</tr>
</tbody>
</table>

(Column Store) + Document Store and Graph Store Functionality
Agenda

• State of Big Data
• Company Background
• Problems We Solve
• How We Are Different
  • Our Technology
• Use Cases
Sqrrl Enterprise has an “open core” architecture

- Sqrrl proprietary
- Automated indexing
- Custom iterators
- Lucene integration
- Security extensions

Sqrrl Server

Sqrrl API over Apache Thrift RPC (JSON, Graph, SQL, Lucene, and more)

Accumulo RPC (Sorted Key/Value I/O)

Hadoop RPC (File I/O)

- Open source (including Sqrrl contributions)
- Open source or commercial distributions
Accumulo’s scale, security, and flexibility drive many of our advantages

- Sorted, distributed, secure, low-latency NoSQL database for multi-structured data
- Fine-grained access controls, massive scalability, iterators for real-time processing, flexible schemas
- User growth expanding exponentially in both federal and commercial sectors

What is Largest Known Cluster Running Stably on a Single Instance of the Software?

- Cassandra: 400 nodes
- Hbase: 500 nodes
- Accumulo: Classified (1000s of nodes)
We simplify things by converting data to JSON documents

Bulk or Streaming Ingest

Raw Data

Accumulo Key/Value Pairs

Sqrrl Enterprise
JSON Documents +
Graph Relationships

<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>JDCPC_JD</td>
<td>20120912</td>
<td>183</td>
</tr>
<tr>
<td>John Doe</td>
<td>Test Results</td>
<td>Mental Health</td>
<td>JDPSYCH_JD</td>
<td>20120801</td>
<td>Pass</td>
</tr>
<tr>
<td>John Doe</td>
<td>Test Results</td>
<td>X-Ray</td>
<td>JDPHYS_JD</td>
<td>20120513</td>
<td>10101101101100...</td>
</tr>
</tbody>
</table>
We have a best-in-class security model...
...That enables fine-grained access controls

Sqrrl’s data-centric security approach allows all the data to be stored on a single platform and only authorized data is returned to the user.

Pushing security to the data-level, simplifies application development and enables more powerful queries.
We are building integrations with stream processing for additional real-time analysis.

1. SPE queries Sqrrl to enrich streaming data
2. SPE persists results in Sqrrl for future query
3. SPE issues data-driven alerts
4. Sqrrl provides context for dashboards
5. Analysis tools query use Sqrrl to search and manipulate historical data
Our real-time analytics include SQL, statistics, full-text search...

<table>
<thead>
<tr>
<th>SQL</th>
<th>Statistics</th>
<th>Full-Text Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>• SELECT</td>
<td>• Aggregate F(x)s</td>
<td>• Lucene 4.0 Syntax</td>
</tr>
<tr>
<td>• FROM</td>
<td>• SUM</td>
<td>• Custom indexing</td>
</tr>
<tr>
<td>• WHERE</td>
<td>• COUNT</td>
<td>• Fielded/Unfielded</td>
</tr>
<tr>
<td>• LIMIT</td>
<td>• AVG</td>
<td>• Phrase search</td>
</tr>
<tr>
<td>• GROUP BY</td>
<td>• MAX, MIN</td>
<td>• Range Queries</td>
</tr>
<tr>
<td></td>
<td>• FIRST, LAST</td>
<td>• AND, OR</td>
</tr>
<tr>
<td></td>
<td>• Scalar F(x)s</td>
<td>• Wildcards</td>
</tr>
<tr>
<td></td>
<td>• +,-,*,,mod,div</td>
<td>• Full regex</td>
</tr>
<tr>
<td></td>
<td>• =,&lt;,&gt;,&lt;=,=,&lt;,&gt;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• date(),time()</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• LOCATE</td>
<td></td>
</tr>
</tbody>
</table>

*Example:* SELECT count(uuid) WHERE ‘love’ GROUP BY doc(‘user/followers_count’) LIMIT 500
...And graph analysis

Graph Example

Sqrrl Graph API Capabilities

- Massively scalable with billions or more edges
- Cell-level security
- Neighbor search
- Path search and traversals
- Building integrations with Blueprints, Jung, and Gremlin
- Use cases include recommendation engines, clustering, network flows
Agenda

• State of Big Data
• Company Background
• Problems We Solve
• How We Are Different
• Our Technology
  • Use Cases
We have a variety of general use cases

<table>
<thead>
<tr>
<th>Use Case</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platform for building real-time Big Data apps</td>
<td>Sqrrl Enterprise is the NoSQL platform with the most diverse set of real-time analytic capabilities; it can serve as the platform for building a wide variety of apps and extends Hadoop to make it real-time</td>
</tr>
<tr>
<td>Combining real-time analytics with transactions</td>
<td>Sqrrl Enterprise breaks down the walls between OLAP and OLTP; now you can build apps that do analytics in real-time and layers that analysis on top of transactions</td>
</tr>
<tr>
<td>Secure search and information sharing</td>
<td>Sqrrl Enterprise’s technology is based on the same technology that Google uses to power its search engine; securely search across all your data in real-time</td>
</tr>
<tr>
<td>Combining datasets and collapsing data siloes</td>
<td>Sqrrl Enterprise is the only NoSQL database with cell-level access controls (data-centric security); use these fine-grained access controls to bring together sensitive datasets into a single Big Data platform</td>
</tr>
</tbody>
</table>
Cybersecurity Use Case

Big Data Platform For Cybersecurity

Key Differentiators

- **Security**: Fine-grained access controls enable multitenancy and secure access to diverse data sets.
- **Scalability**: A large SOC can store and search petabytes of log files, emails, etc. in real-time.
- **Adaptivity**: Diverse analytic capabilities such as real-time graph, full-text search, SQL, and statistics support a wide variety of apps.
Homeland Security Use Case

Secure “Data Playground”

• Create a “data playground” for homeland security analysts to explore a variety of immigration, intelligence, and benefits data sets
• Uncover hidden patterns in the data using exploratory analysis tools; export patterns into operational systems

Key Differentiators

• **Security:** Enable access to sensitive data for individuals both inside and outside the organization using fine-grained access controls
• **Scalability:** Ingest tera- and petabytes of multi-structured datasets in all different formats
• **Adaptivity:** Begin with simple search and build more complex apps using graph and statistics capabilities
Contact

sqrrl data, inc.
275 Third St.
Cambridge, MA 02142

617-902-0784
www.sqrrl.com
@sqrrl_inc
info@sqrrl.com