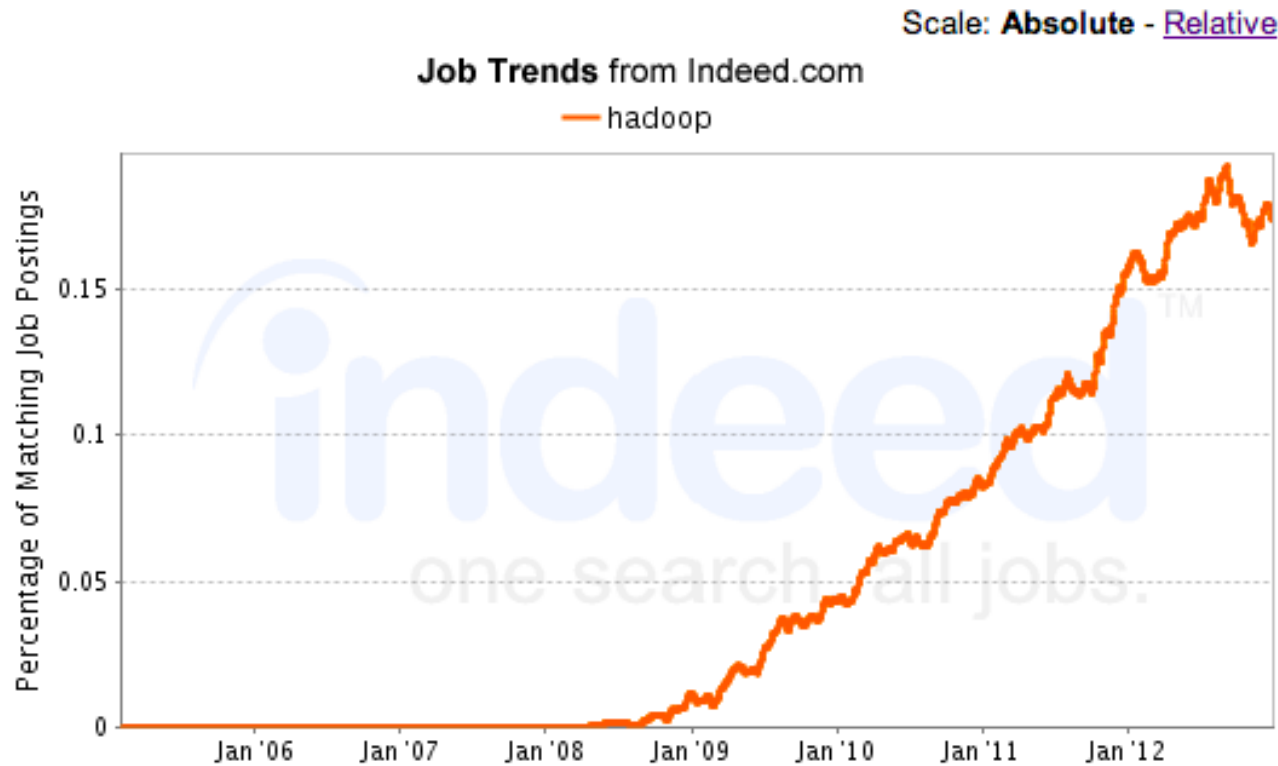# sqrrl

Secure. Scale. Adapt.

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

hadoop Job Trends

Scale: **Absolute** - Relative

▶ **Email to a friend**

▶ **Post on your blog/website**

Job Trends from Indeed.com
— hadoop

Indeed.com searches millions of jobs from thousands of job sites.
This job trends graph shows the percentage of jobs we find that contain your search terms.

Top Job Trends

1. HTML5
2. MongoDB
3. iOS
4. Android
5. Mobile app
6. Puppet
7. Hadoop
8. jQuery
9. PaaS
10. Social Media

3

There are some important advantages to Hadoop and NoSQL

Flexible Schemas

- Easily add columns on the fly
- No initial data modeling or ETL needed
- Accommodates sparse datasets and datasets with evolving schema

Resilience + Durability

- Data is triple replicated
- Failure of any regular node(s) has no impact
- Proven ability to run mission critical applications

Cost Effectiveness

- Designed to run on cheap commodity hardware
- Hardware costs scale linearly with data volumes
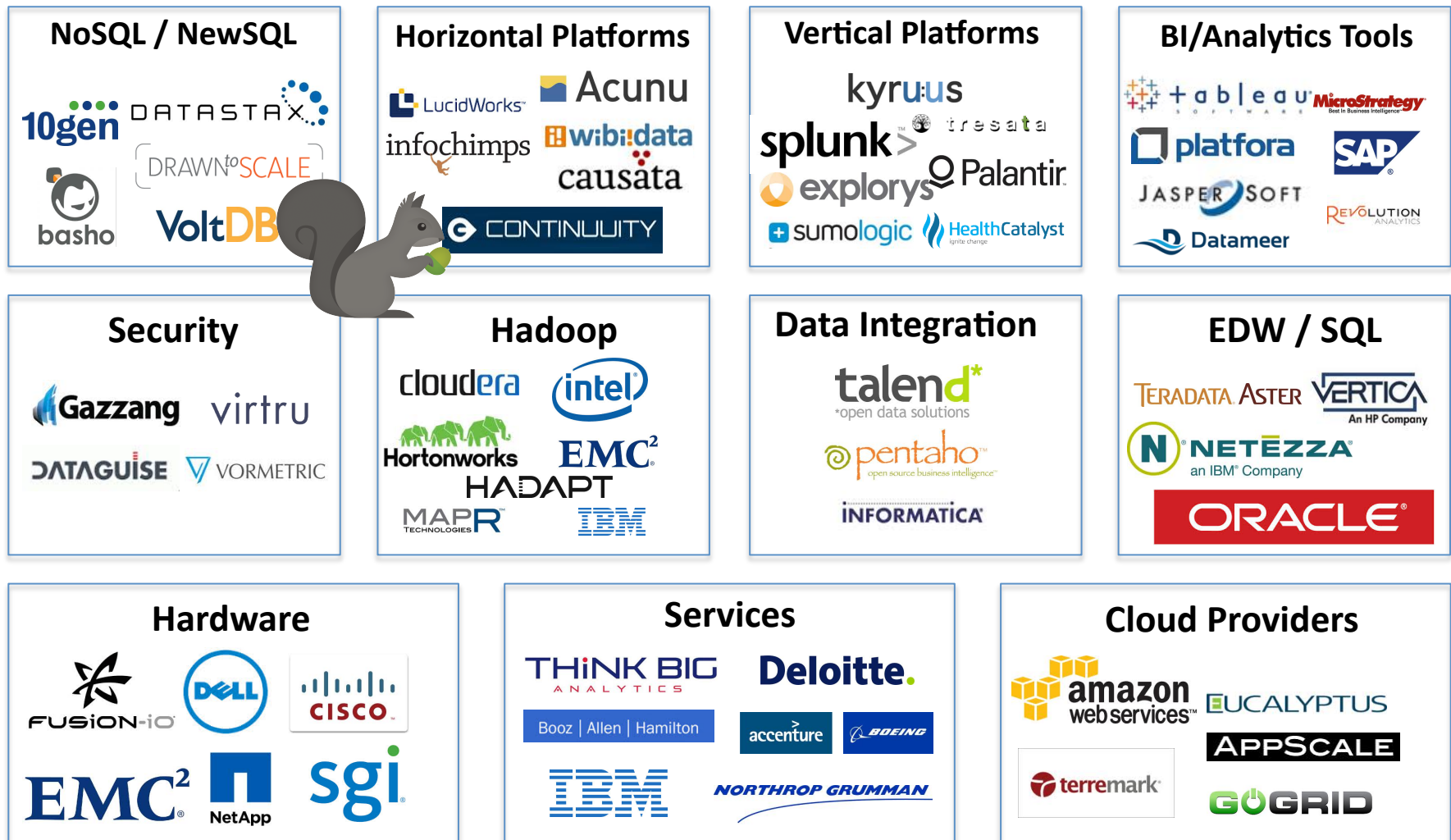- Largely based on open source software

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...

...But Sqrrl has some unique capabilities

accumulo ™

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

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

Investors

ATLASVENTURE

matrix PARTNERS

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
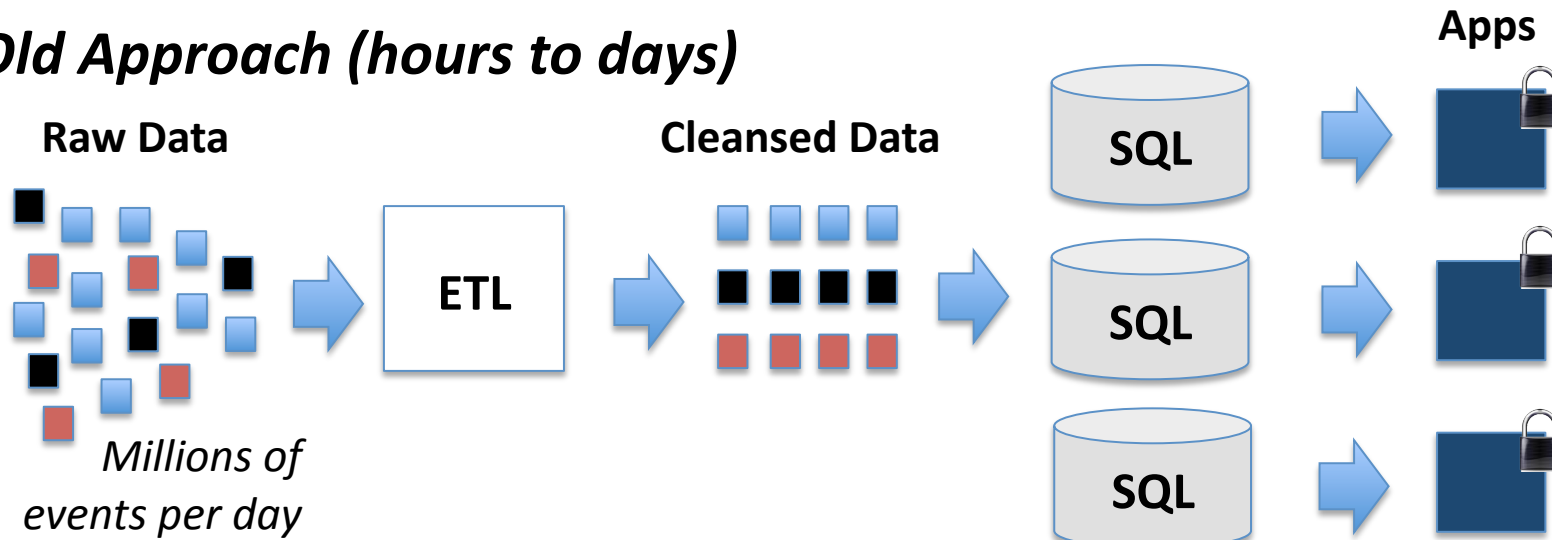
From the front-lines of...

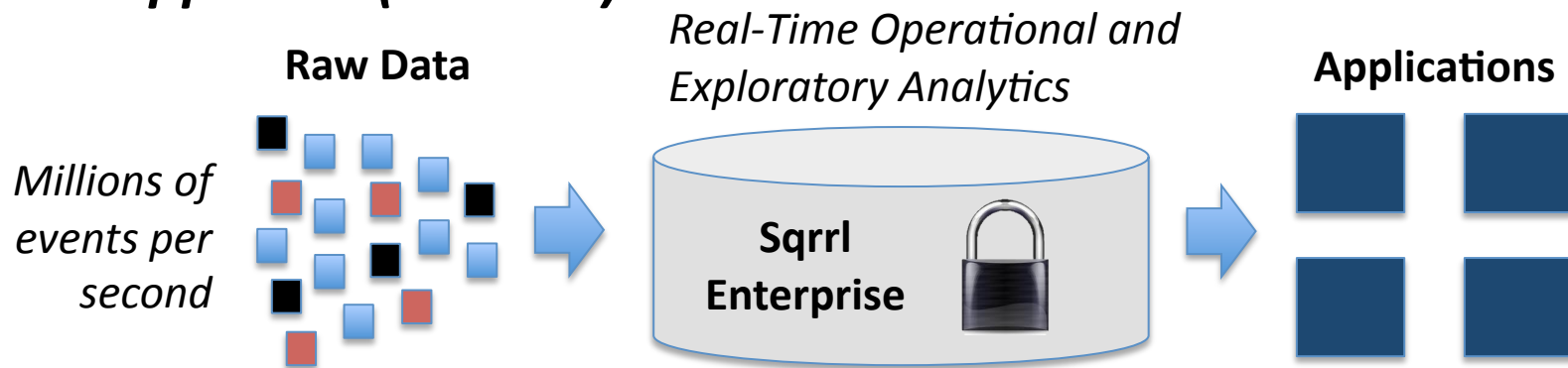O'REILLY®

Strata
CONFERENCE

- 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
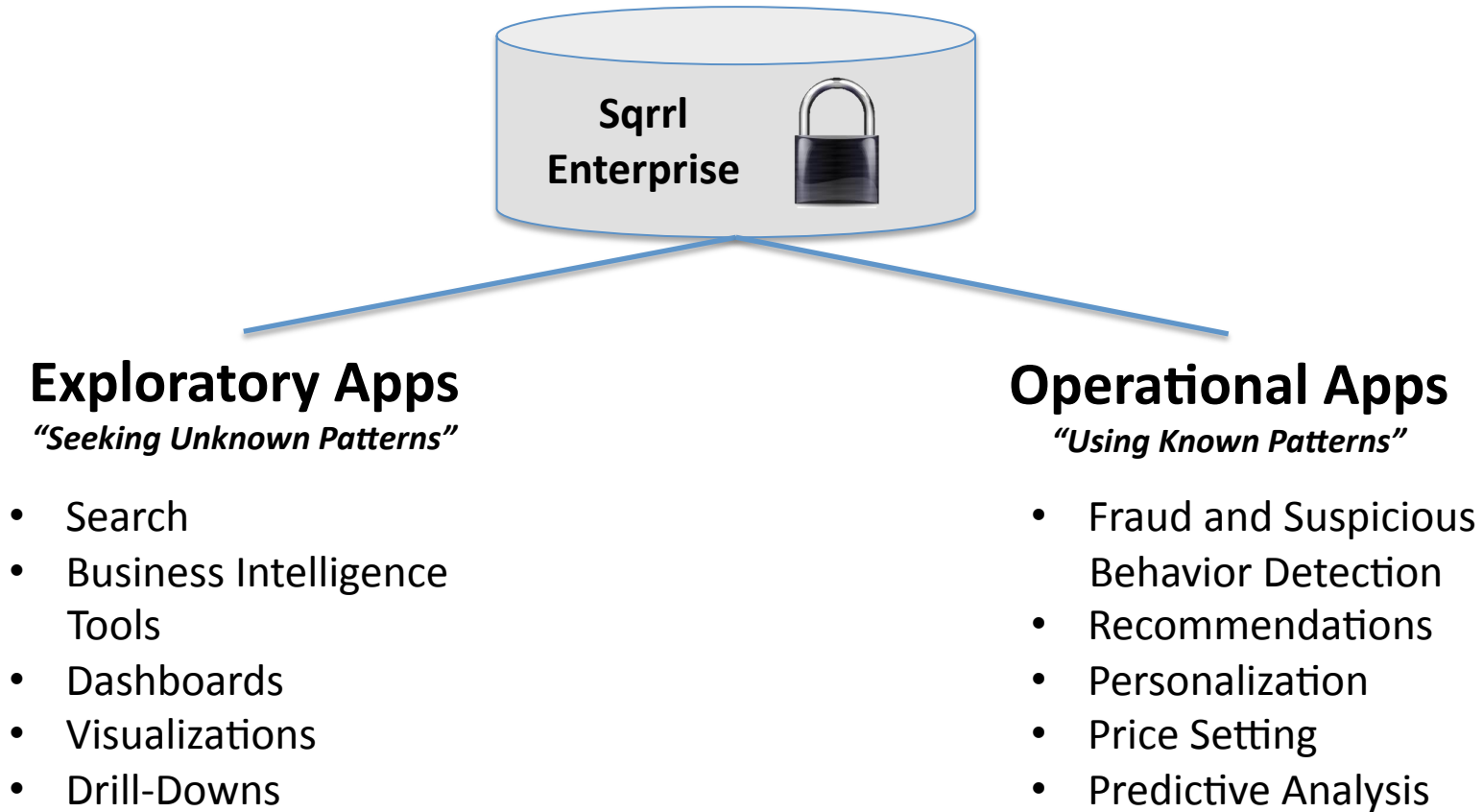
Sqrrl brings secure, real-time apps to Hadoop

Old Approach (hours to days)

Apps

Raw Data

Cleansed Data

SQL

SQL

SQL

ETL

Millions of events per day

New Approach (seconds)

Raw Data

Real-Time Operational and Exploratory Analytics

Applications

Millions of events per second

Sqrrl Enterprise

Sqrrl supports both exploratory and operational apps

Sqrrl
Enterprise

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

Type	Examples
Key-Value Store	
Column Store	
Document Store	
Graph Store	

<----- Increasing Data Complexity

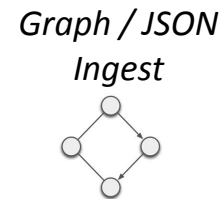*Increasing Scalability ----->*

 =

(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

Hadoop Analytic Integration

Exploratory / Operational Apps

Graph / JSON Ingest

Data Export

Sqrrl API over Apache Thrift RPC
(JSON, Graph, SQL, Lucene, and more)
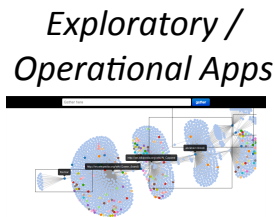
- **Sqrrl proprietary**
- **Automated indexing**
- **Custom iterators**
- **Lucene integration**
- **Security extensions**

Sqrrl Server

Accumulo RPC
(Sorted Key/Value I/O)

- **Open source (including Sqrrl contributions)**

Hadoop RPC
(File I/O)

- **Open source or commercial distributions**

Accumulo's scale, security, and flexibility drive many of our advantages

accumulo™

- 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 <u>Stably</u> on a Single Instance of the Software?

Classified (1000s of nodes)

500 nodes

400 nodes

| Cassandra | Hbase | Accumulo |
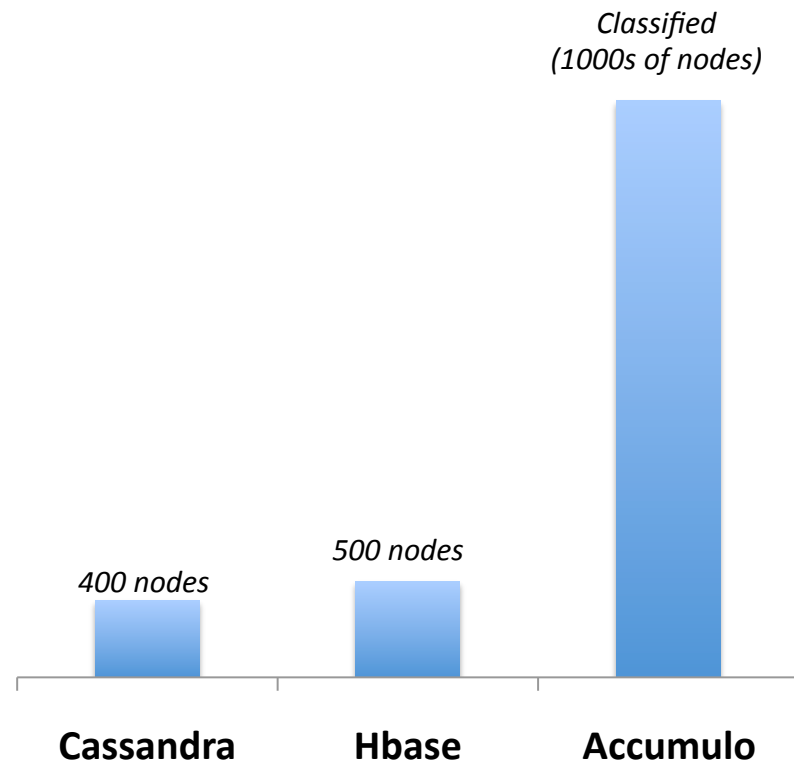
We simplify things by converting data to JSON documents

XML

Flat Files

CSV

Logs

Images

Tables

Raw Data

Bulk or Streaming Ingest

Row	Col. Fam.	Col. Qual.	Visibility	Timestamp	Value
John Doe	Notes	PCP	PCP_JD	20120912	Patient suffers from an acute ...
John Doe	Test Results	Cholesterol	JD|PCP_JD	20120912	183
John Doe	Test Results	Mental Health	JD|PSYCH_JD	20120801	Pass
John Doe	Test Results	X-Ray	JD|PHYS_JD	20120513	1010110110100...

Accumulo Key/Value Pairs

```
- Album: {
      Name: "Thriller",
      Position: 4,
      ReleaseDate: "2001"
   },
- Artist: {
      Name: "Michael Jackson",
      - SimilarArtists: [
          - {
                Name: "The Jackson 5"
          },
          - {
                Name: "The Jacksons"
```

Sqrrl Enterprise JSON Documents + Graph Relationships

We have a best-in-class security model...

Provided by Sqrrl

Provided by Customer

21
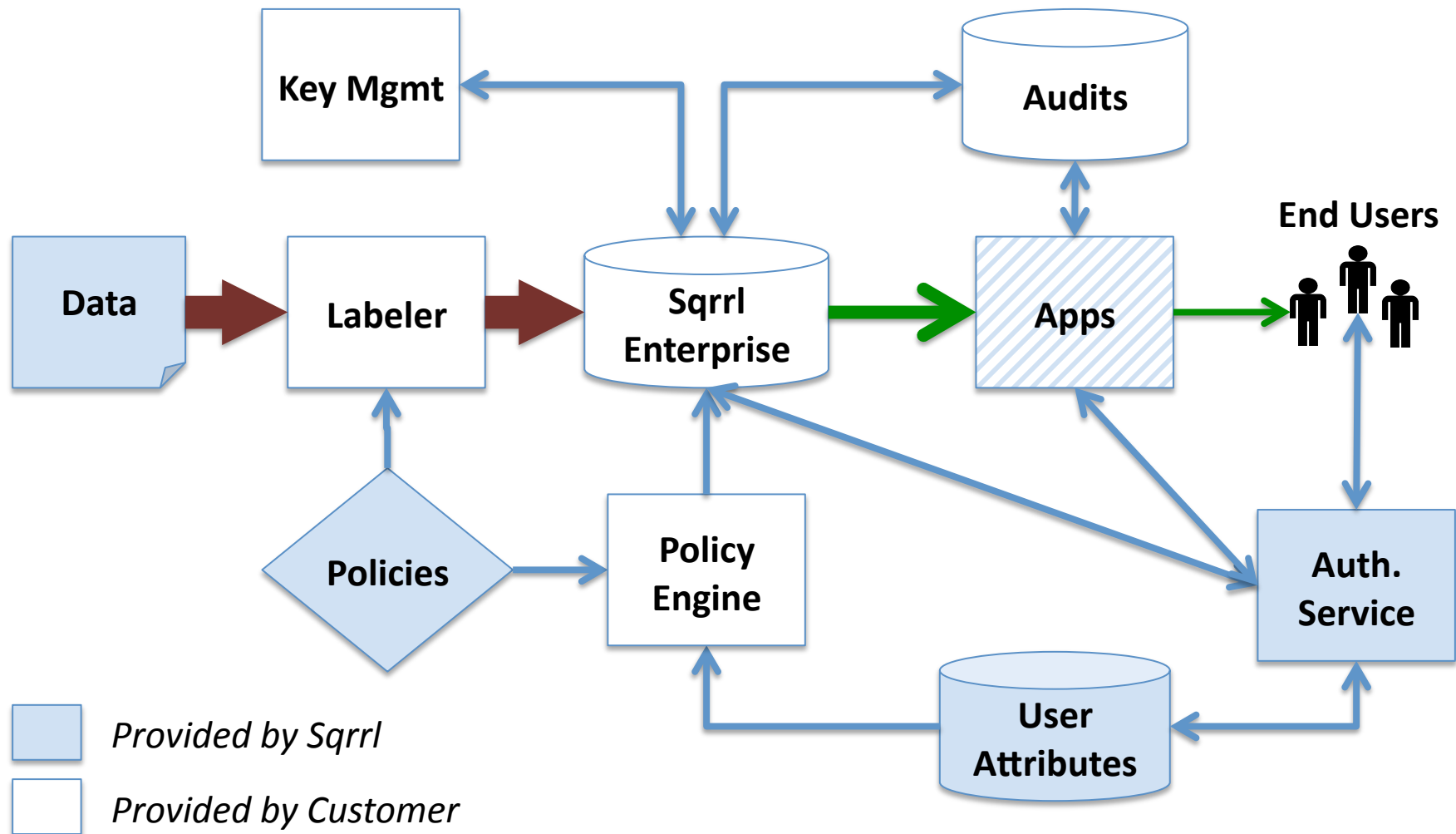
...That enables fine-grained access controls

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

Row	Col	Value
1	Name	Anon1
1	Sales	100
2	Name	Smith
2	Sales	350
2	Quota	1000

User 1

Sqrrl Enterprise

User 2

Row	Col	Value
1	Name	Jones
1	Sales	100
1	Age	28
2	Name	Smith
2	Sales	350
2	Age	25
2	Quota	1000
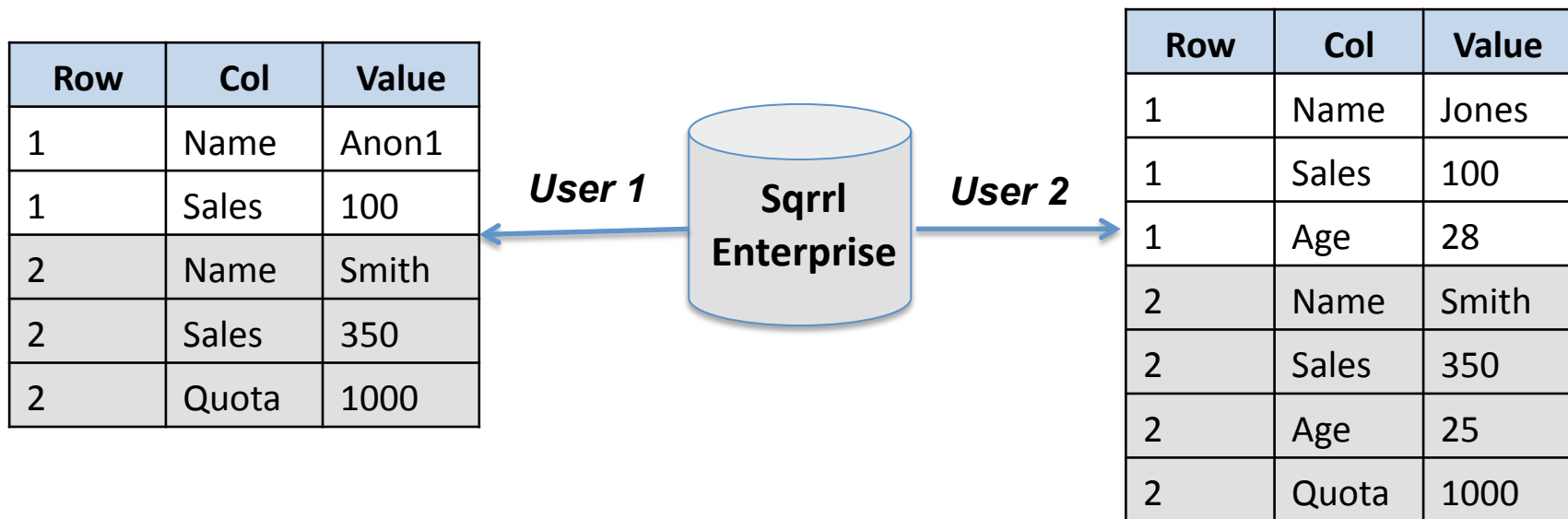
Pushing security to the data-level, simplifies application development and enables more powerful queries

We are building integrations with stream processing for additional real-time analysis
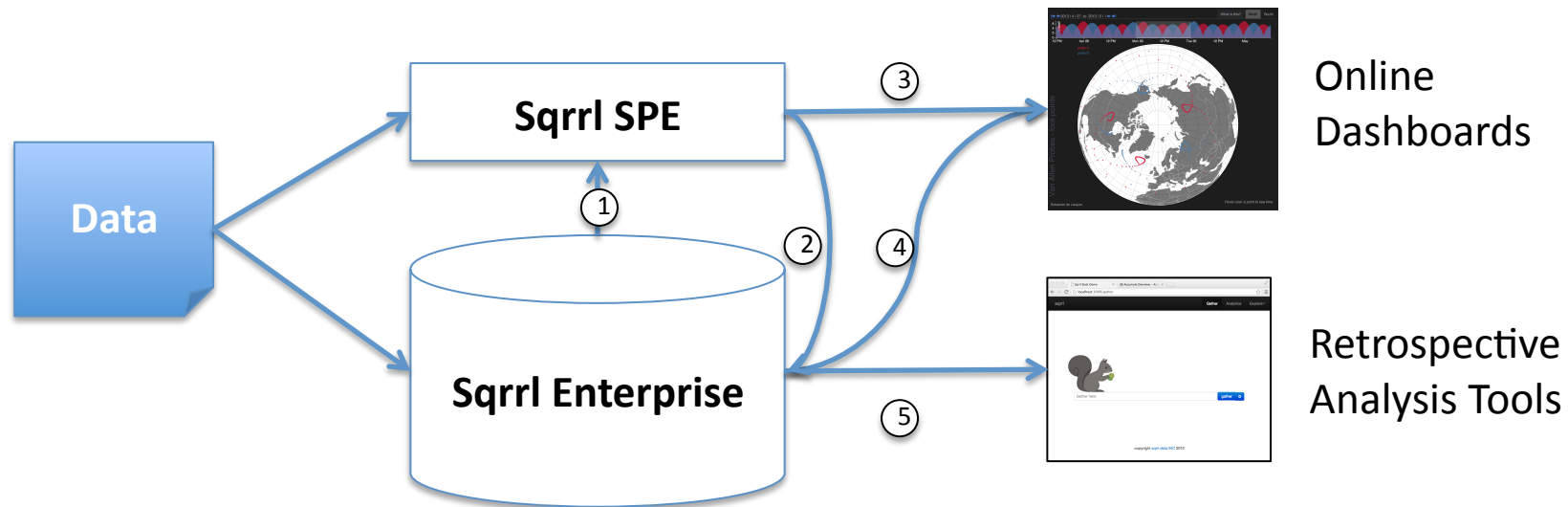
1. SPE queries Sqrrl to enrich streaming data
2. SPE persists results in Sqrrl for future query
3. SPE issues data-driven alerts
4. Sqrrl provides context for dashboards
5. Analysis tools query use Sqrrl to search and manipulate historical data

Our real-time analytics include SQL, statistics, full-text search...

SQL	Statistics	Full-Text Search

SQL
- SELECT
- FROM
- WHERE
- LIMIT
- GROUP BY

Statistics
- Aggregate F(x)s
  - SUM
  - COUNT
  - AVG
  - MAX, MIN
  - FIRST, LAST
- Scalar F(x)s
  - +,-,*,/,mod,div
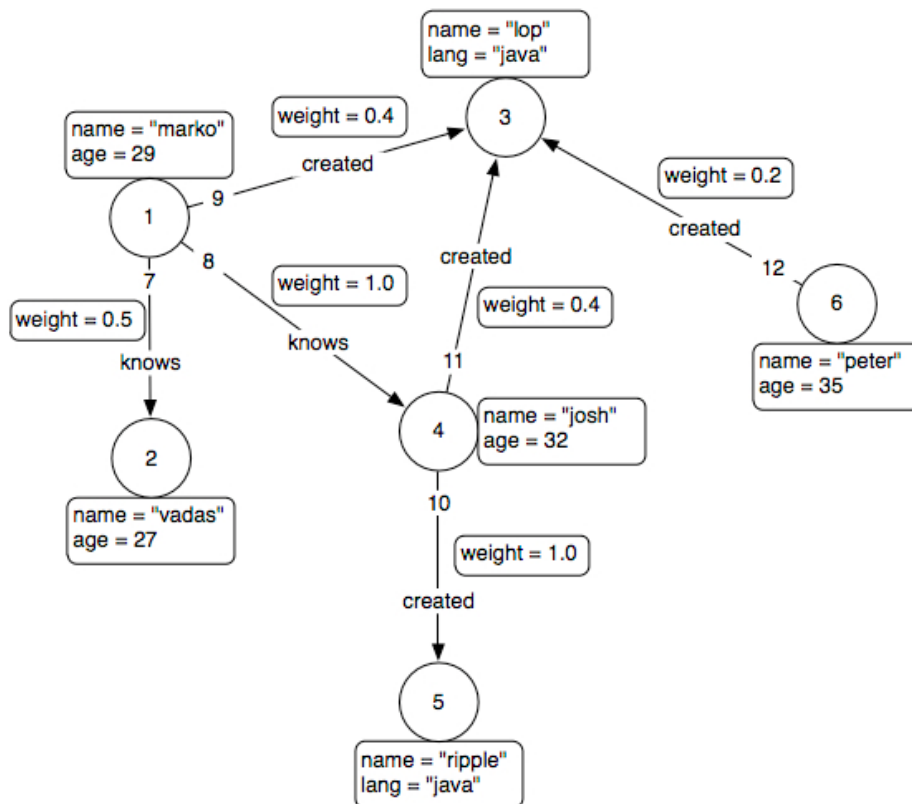 - =,<,>,<=,>=,<>
  - date(),time()
  - LOCATE

Full-Text Search
- Lucene 4.0 Syntax
- Custom indexing
- Fielded/Unfielded
- Phrase search
- Range Queries
- AND, OR
- Wildcards
- Full regex

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

Use Case	Description
Platform for building real-time Big Data apps	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
Combining real-time analytics with transactions	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
Secure search and information sharing	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
Combining datasets and collapsing data siloes	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

Cybersecurity Use Case

Big Data Platform For Cybersecurity
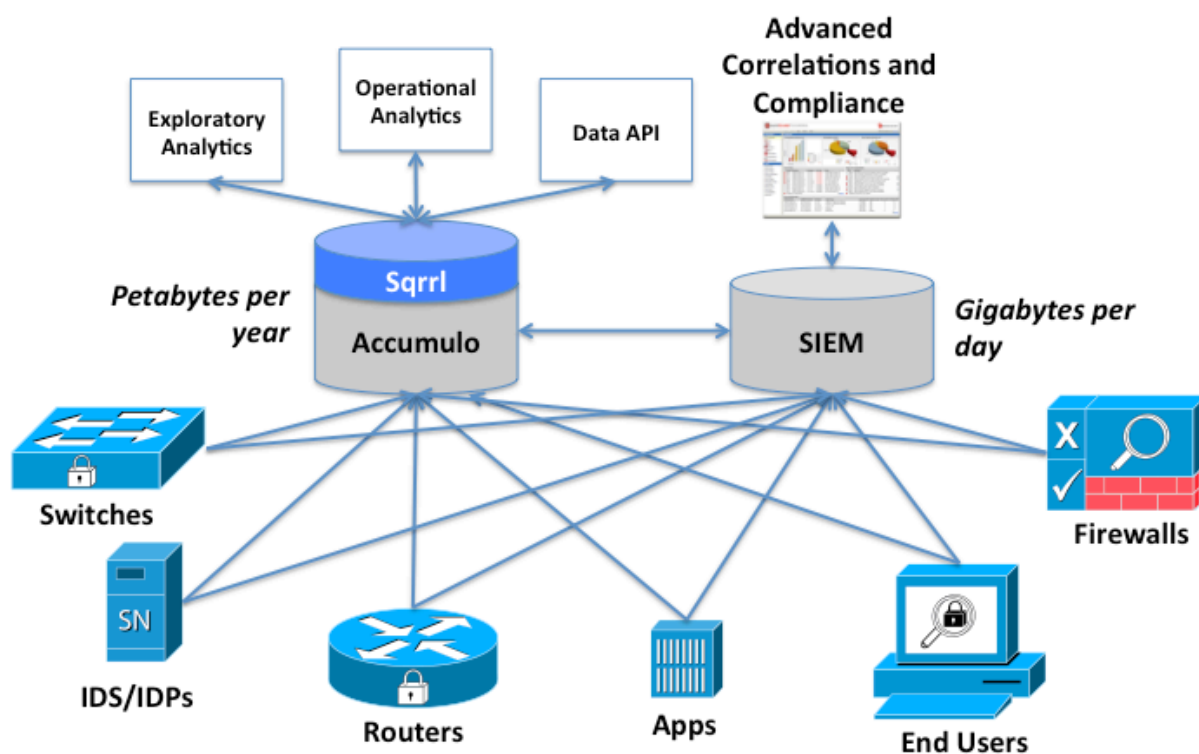
Key Differentiators

- **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-structued 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