sqrrl

Securely explore your data

BIG DATA INFRASTRUCTURE: LESSONS LEARNED

Adam Fuchs, CTO
Sqrrl Data, Inc.
October 29, 2013


Agriturismo.it
Airfoil Speaker...
AmpliTube FREE
Apple
AutoScout24 Ita...
AutoWeek
BabyFirst's Pla...
Beurs.nl
Bing
Bloomberg
Buienradar.nl
Bushaltes
Chordmaster LE
Concert Vault
Convertbot —...
Currency
De Telefoongid...
De Telegraaf
Denon Remote...
DenONOFF
Dice Match
Dutch Cowboys
Echofon for Twi...
Facebook
FD Mobiel
Fitts-Fitts
Firefox Home
Flickr
foursquare
fring: Video Cal...
FT Mobile
funda
GSMA MWC 2011
The Guardian V...
Guitar Tuner by...
Hipstamatic
HuffingtonPost...
Hyves
Handy Waterpas
Ikcal - zoek lok...
MapMyRun
iWorkout Lite
Kik Messenger
Last.fm
Layar Reality B...
LCD Clock - Cl...
LED Flashlight
LinkedIn
Lyrical
Marktplaats
Mashable!
Metronome
MiniPiano
Mobile Mouse...
NAVIGON Mobi...
NetNewsWire
NOS Journaal
NOS Teletekst
NU
NYTimes
Opera Mini We...
PAC-MAN Lite
PayPal
Ping! Lite
PolyTune
Rabo Bankieren
Rabo SMS Beta...
Radio 1
Shazam
Skype
Spotify
Steel Guitar
Talking Tom Cat
Tap Tap Reven...
TETRIS® [ENG...
Thomson Reut...
TomTom car kit...
TomTom West-...
Trein
TwitterVis
UDID Sender
Uitzending Ge...
Ustream Live B...
Viber - Free Ph...
Volkskrant
vTuner Radio
The Wall Street...
WhatsApp Mes...
3D Brick Break...
2011 Consumer...

RAPID APPLICATION DEVELOPMENT REQUIRES...

- Prototyping with operational data


+

+

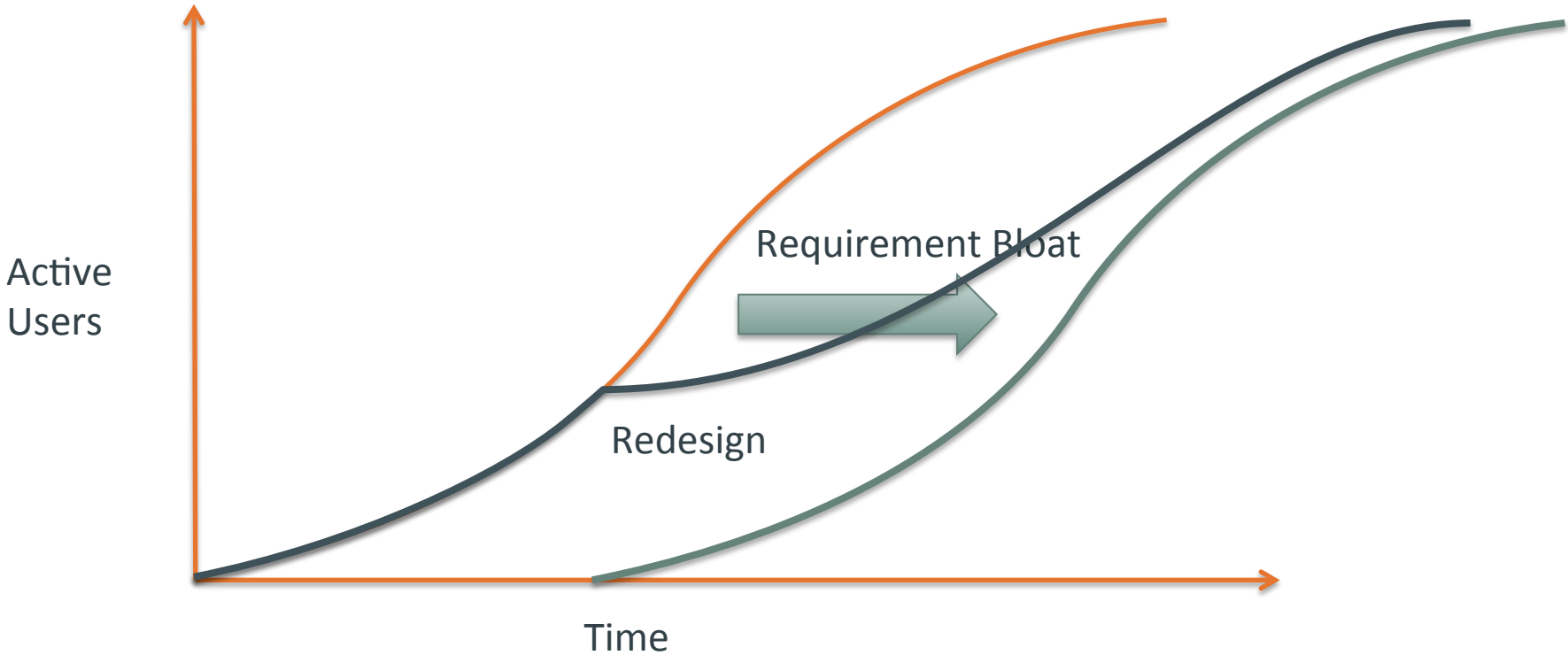
=

?

- Smooth transition from dev to production
  - Scaling users, data
  - Achieving durability, consistency, availability
  - Security and compliance

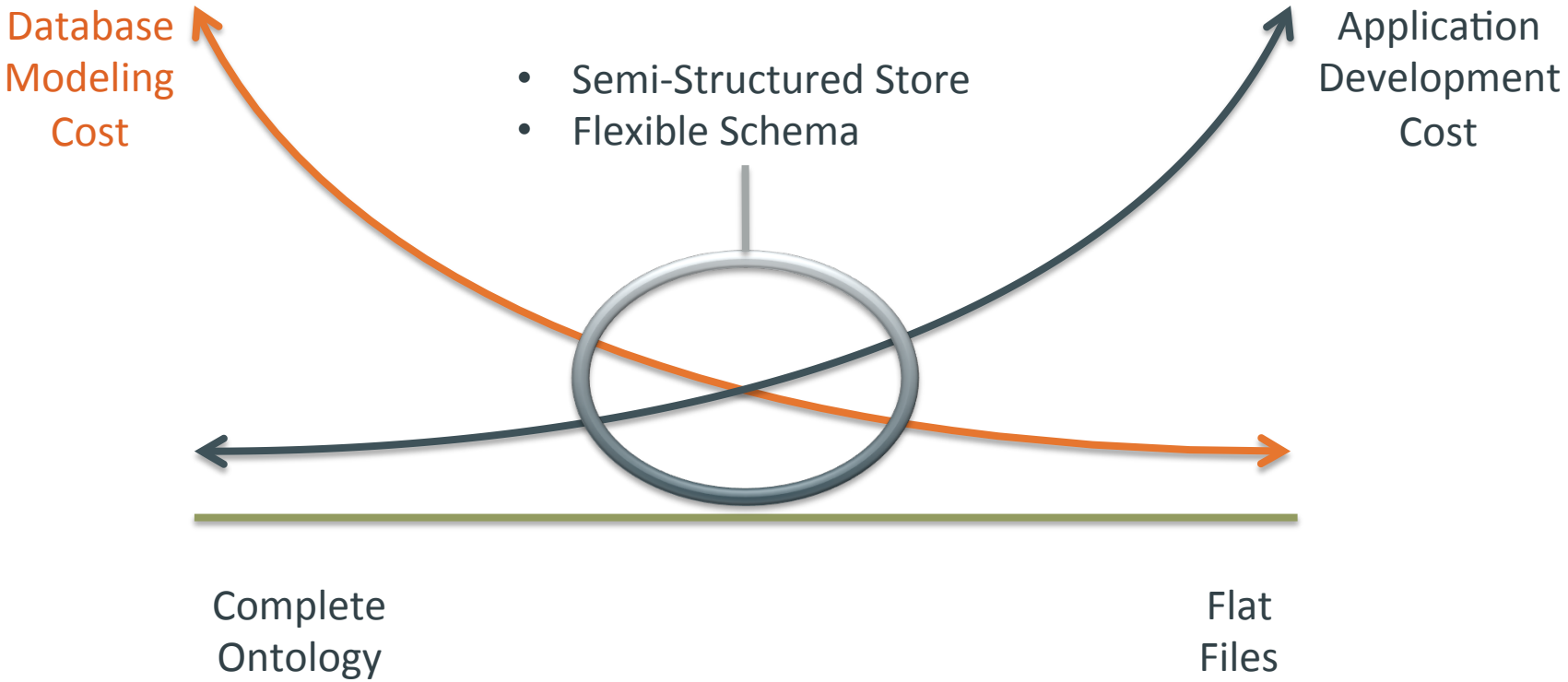
ADOPTION CURVES

Start small, but design for scalability



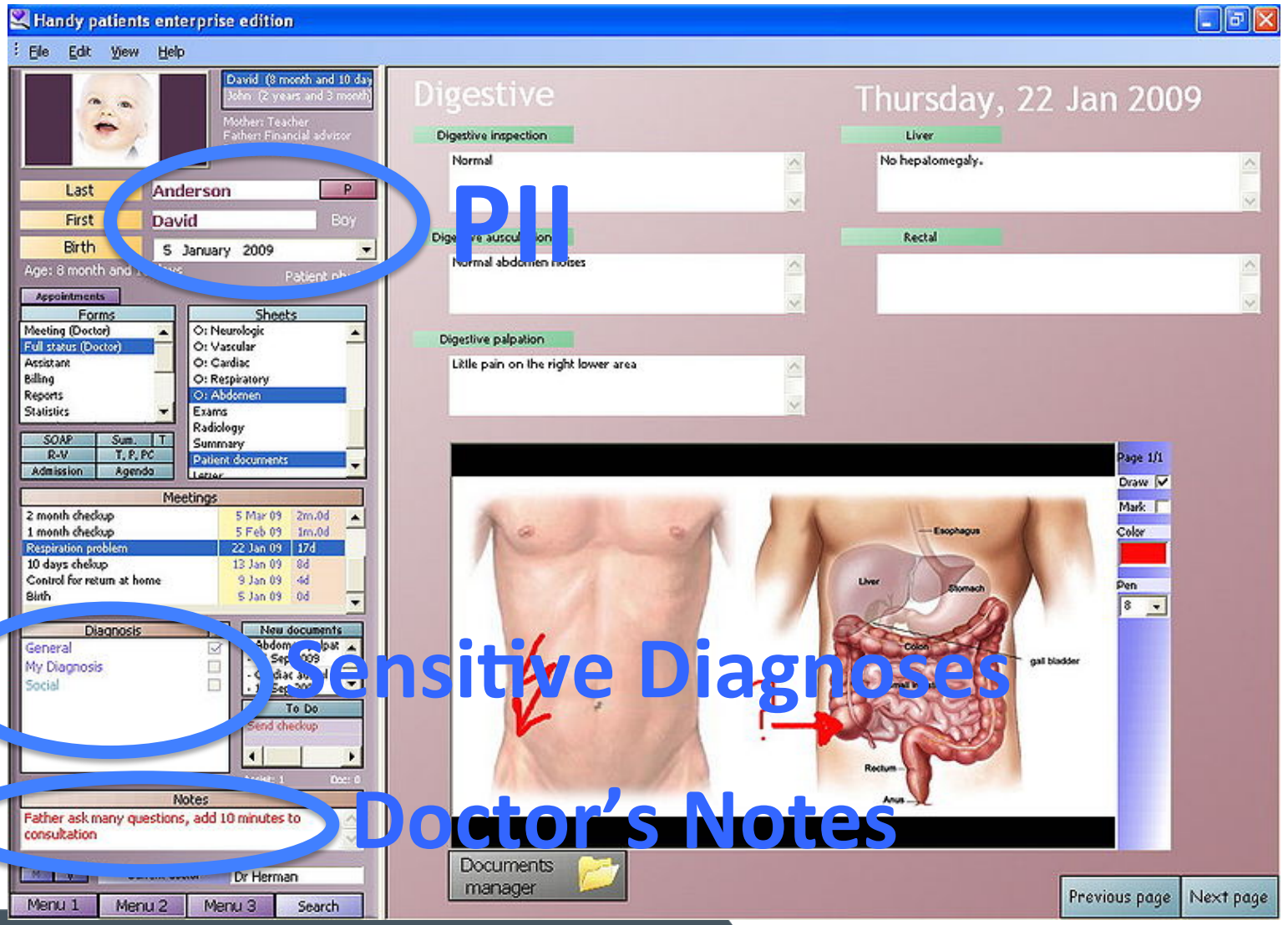
DATA MODELING COSTS

Seek a happy medium...



DATA-CENTRIC SECURITY

Simplifying complex security models



DISCOVERY ANALYTICS

Building blocks for secure, scalable, operational apps.

Common solutions for operational requirements, like:

- Scalability
- Security
- Schema Adaptation
- Multi-tenancy

In domains like:

- "Data Lake"
  - Indexed Search
  - Discovery
  - Big-Picture Views
  - Bulk slicing
- "Operational Apps"
  - High concurrency, low-latency
  - Real-time
  - Denormalized

DATA LAKE

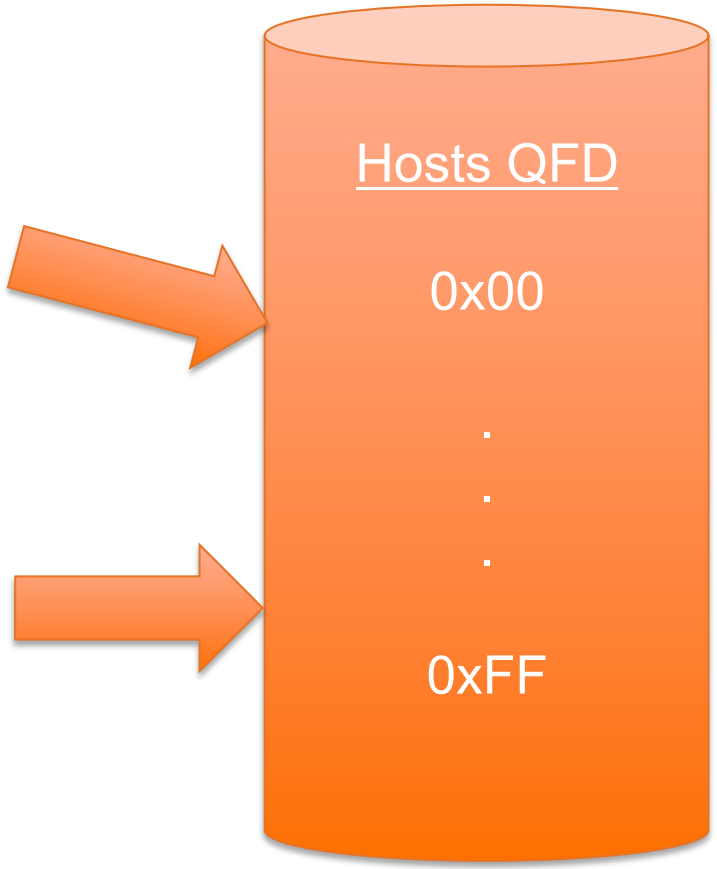
+Data-Centric Security

+Operational, Real-Time

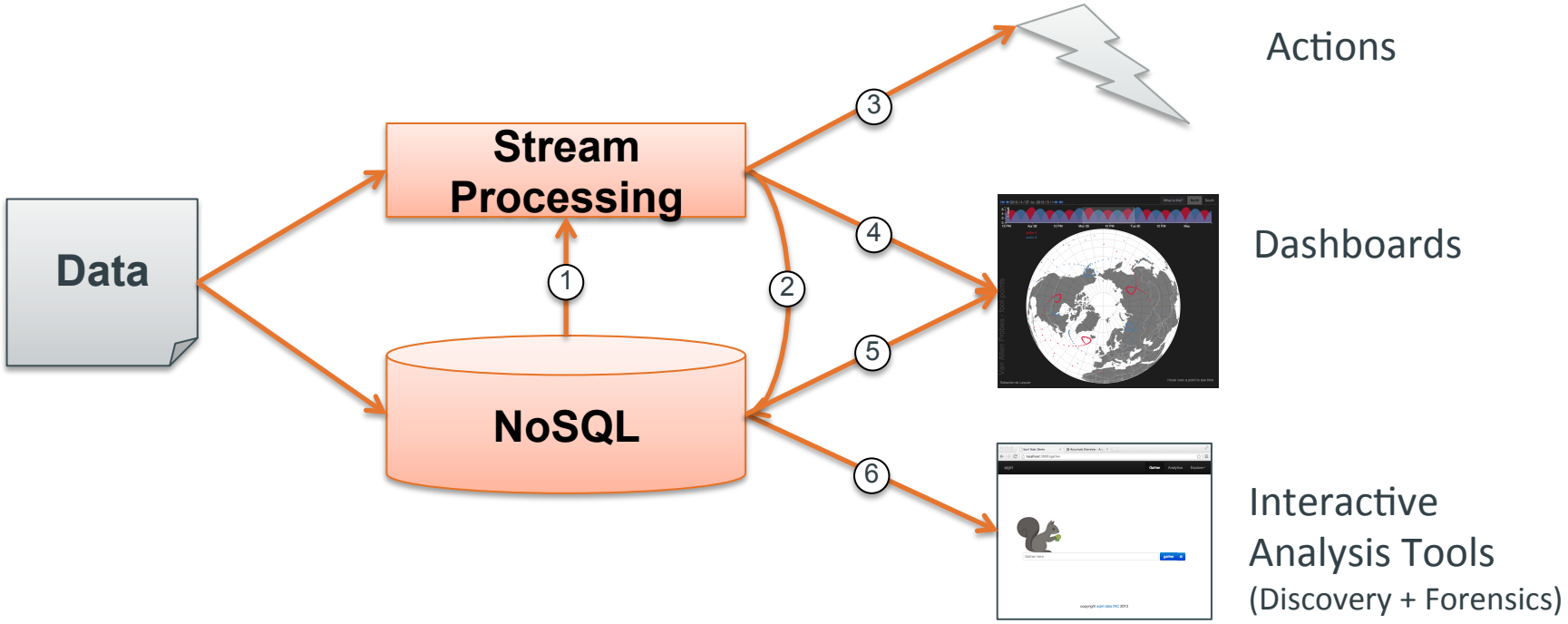
QUESTION-FOCUSED DATASETS

Source	Destination	Port	Bytes In	Bytes Out	Protocol
10.1.2.3	10.9.8.7	80	73,824	15,632	http

Key	-> Value
10.1.2.3, Bytes In	-> +73,824
10.1.2.3, Bytes Out	-> +15,632
10.1.2.3, Ports Used	-> +{80}
10.1.2.3, Protos Used	-> +{http}
10.9.8.7, Bytes In	-> +15,632
10.9.8.7, Bytes Out	-> +73,824
10.9.8.7, Ports Hosted	-> +{80}
10.9.8.7, Protos Hosted	-> +{http}



NOSQL: REAL-TIME
Real-Time Ecosystem



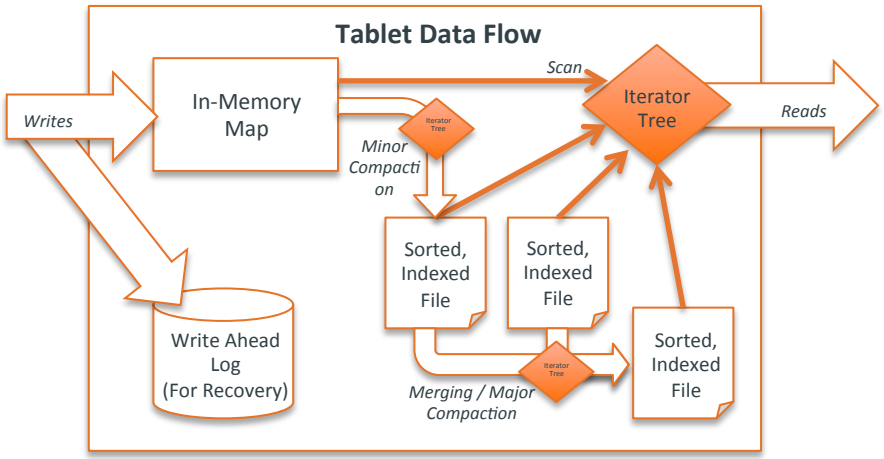
1. SPE queries NoSQL to enrich streaming data
2. SPE persists results in NoSQL for future query
3. SPE takes action automatically
4. SPE issues data-driven alerts
5. Sqrrl provides context for dashboards
6. Analysis tools query use Sqrrl to search and manipulate historical data

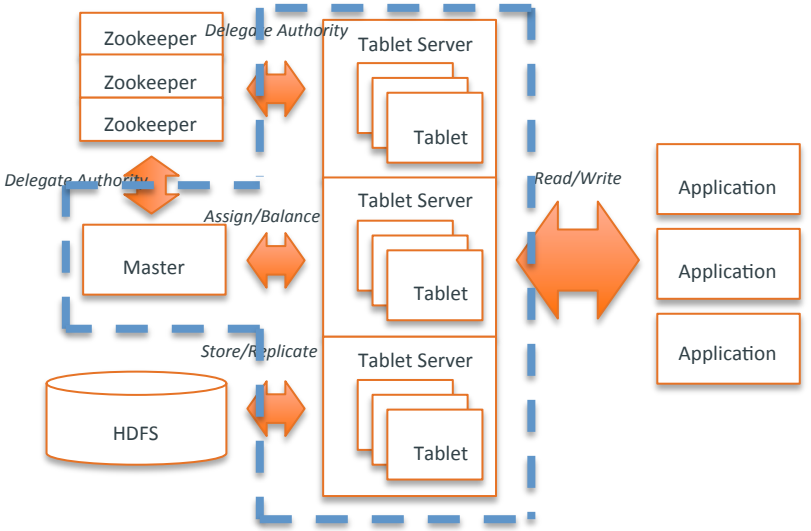
APACHE ACCUMULO

- Founded in 2008 (internal NSA)
- Modeled after Bigtable
- Open-sourced to ASF in 2011

Strengths

- ***Shared-Nothing*** => Scalability
- Micro-Batching for Efficient ***Random I/O***
- ***Iterators*** for local, server-side analytics on sorted data
- ***High Concurrency, Low Latency*** for Denormalized Data
- Sparse, ***Flexible Schema*** supports dynamic and diverse data models
- ***Cell-level Security*** promotes sharing





SECURITY AT ACCUMULO'S CORE

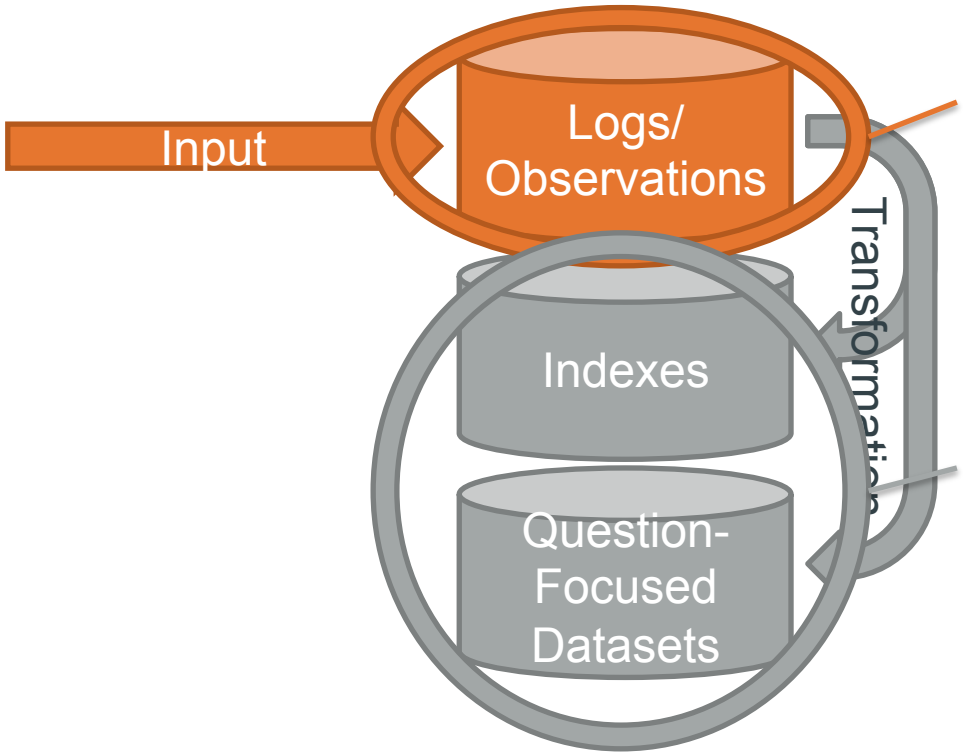
An Accumulo key is a 5-tuple, consisting of:

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

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 Example

SECURITY IN INDEXES AND QFDS

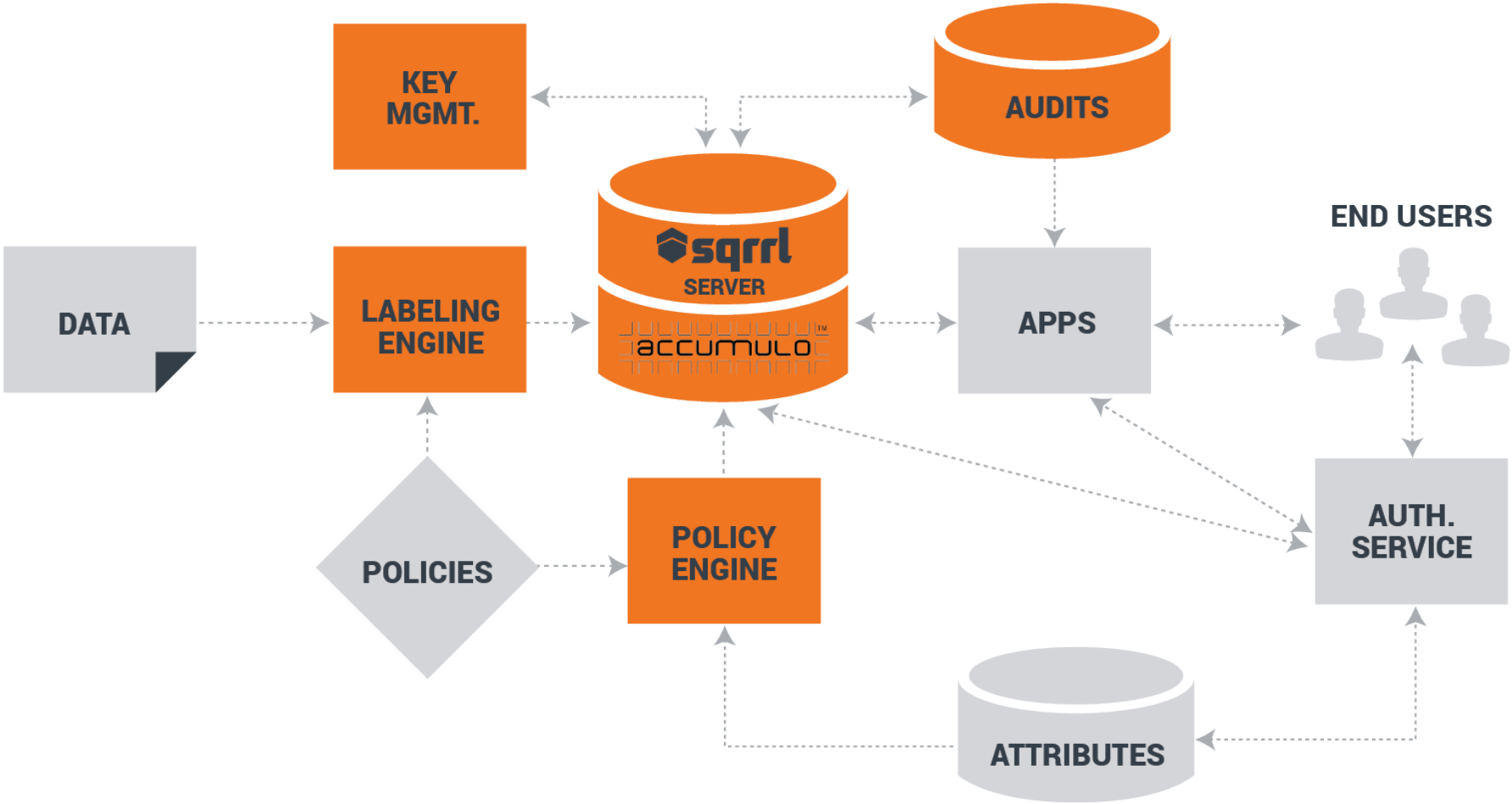


- **Simple Provenance**
- ***Simple Security Model***

- **Complex Provenance**
- ***Complex Security Model***

DATA-CENTRIC SECURITY

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



LAYERED ARCHITECTURE

Turtles all the way down...

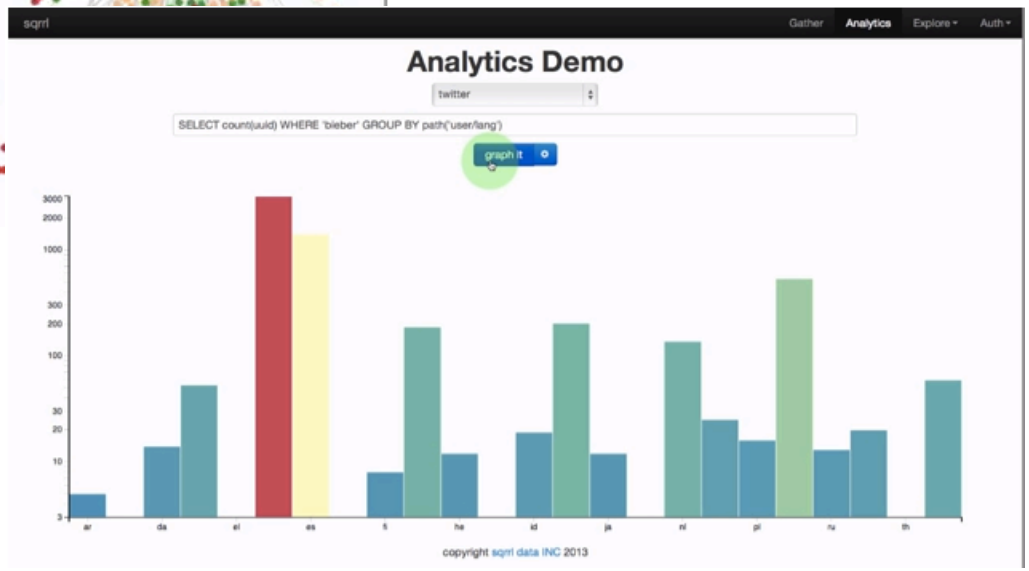
Sqrrl API over Apache Thrift RPC
(JSON, Graph, Aggregation, Search, etc.)

Accumulo RPC
(Sorted Key/Value I/O)

Hadoop RPC
(File I/O)

REAL-TIME OPERATIONAL APPS

Innovation through rapid development



sqrrl

Securely explore your data

QUESTIONS?

Adam Fuchs, CTO
Sqrrl Data, Inc.
October 29, 2013