RSA CONFERENCE 2014
ASIA PACIFIC & JAPAN

Share.
Learn.
Secure.
Capitalizing on
Collective Intelligence

How to Hadoop Without the Worry: Protecting Big Data at Scale

SESSION ID: CDS-W06

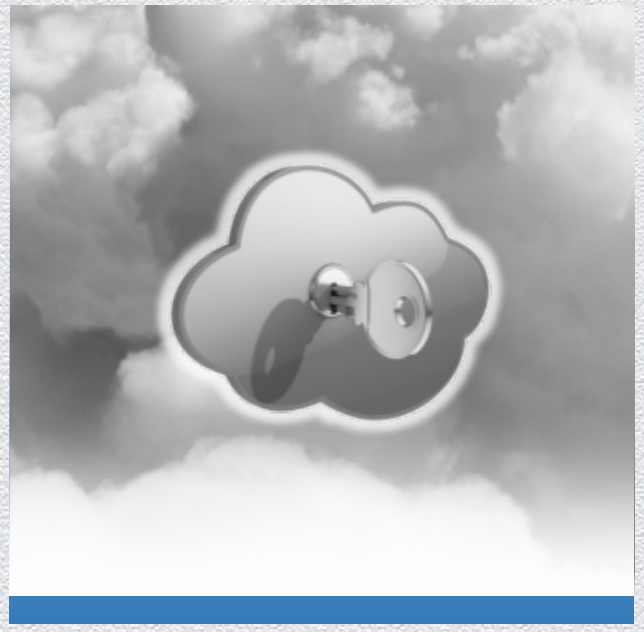
Davi Ottenheimer

Senior Director of Trust
EMC Corporation
@daviottenheimer

Big Data Trust. Redefined

Transparency

Relevance



Resilience

http://gardenofeaden.blogspot.com/2012/12/the-starling.html

http://weedworld.blogspot.com/2012/09/starling-count-eldernell.html

EMC²

#RSAC

RSACONFERENCE2014
ASIA PACIFIC & JAPAN

スマートフォンリモコン ※1,2

Bluetooth® ※3 無線技術でお持ちのスマートフォンからリモコン操作。
トイレ本体と同期して今までのリモコンでは実現できなかった機能を搭載。

リモコン機能

シャワートイレの個人設定や、スマートフォンに保存している音楽をトイレ本体のスピーカーで再生できます。

トイレ日記

日々の排便状況をカレンダー上に記録して、健康管理に活用いただけます。

Bluetooth

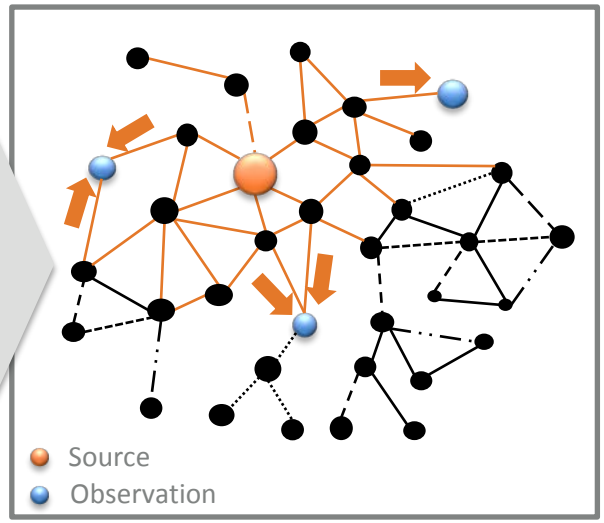
※1 スマートフォンリモコンはアンドロイドのみに対応します。
※2 専用アプリ「My SATIS」のサービス内容、画面デザインは予告なく変更する場合があります。
※3 Bluetooth®は、米国Bluetooth SIG Inc. の登録商標です。

2012 RSAC Breach Data Lessons

1854 GHOST MAP

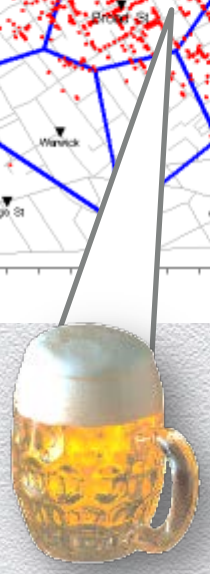
CHOLERA VORONOI

NETWORK BREACHES



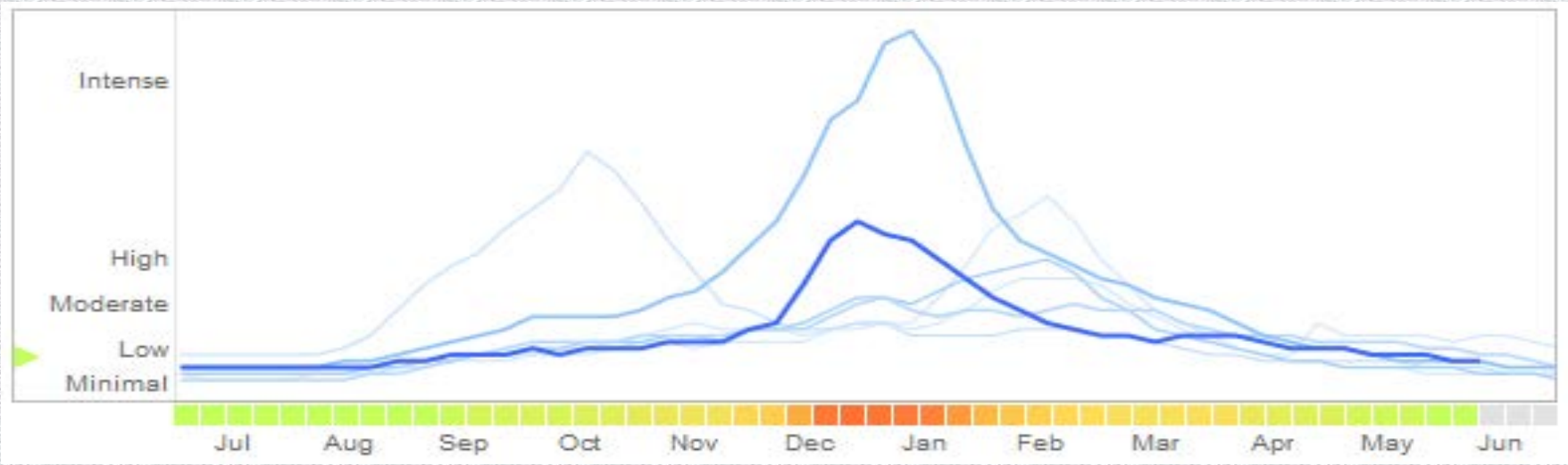
http://www.flyingpenguin.com/?p=18259

Dr. John Snow
1813-1858



Yet…Google Wrong 3rd Year in a Row

"Algorithmic accountability one of the biggest problems of our time"

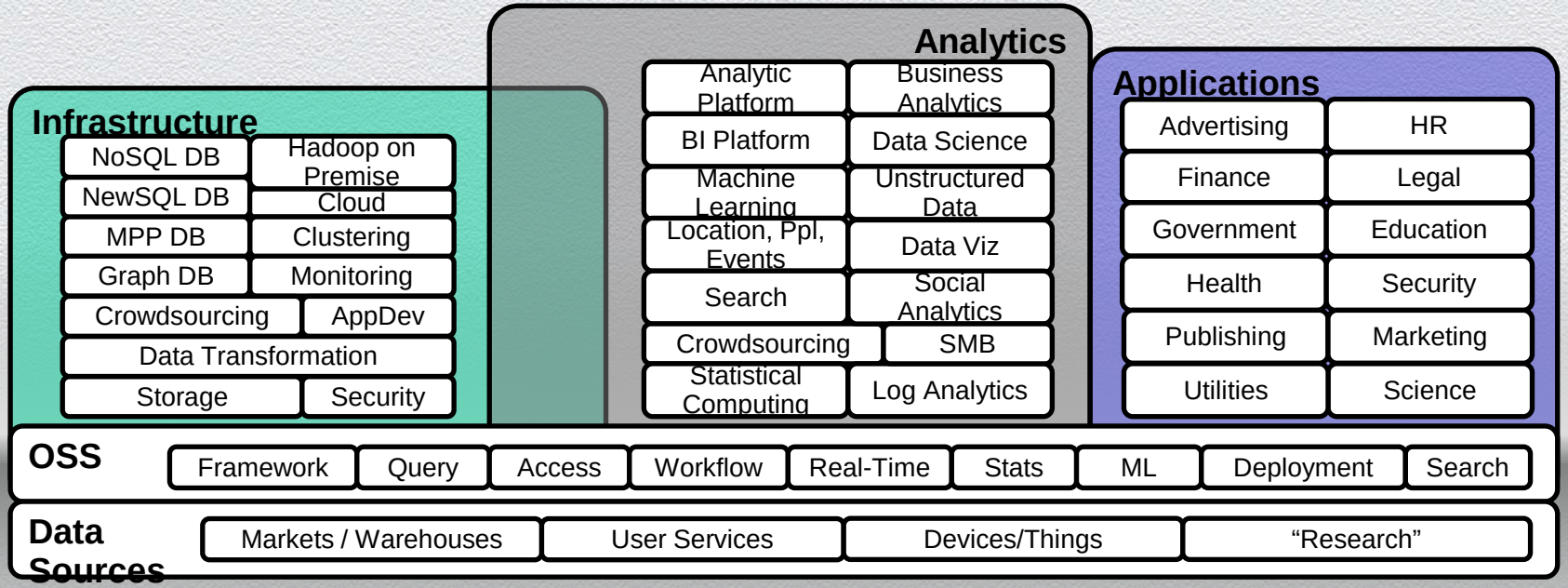


"…system consistently overestimated flu-related visits over the past 3 years...

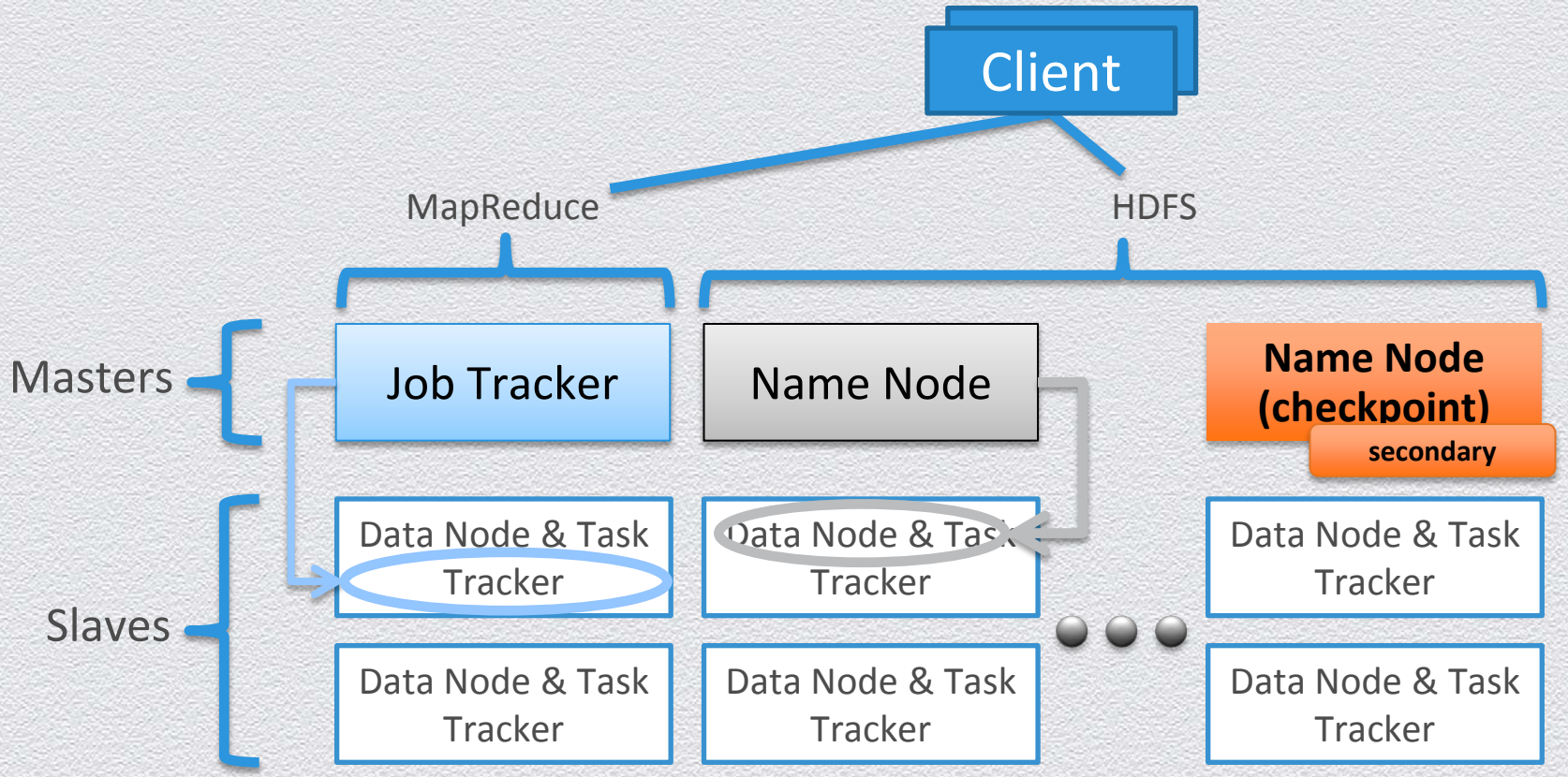
...especially inaccurate around the peak of flu season – when such data is most useful."

http://www.google.org/flutrends/us/
http://www.newscientist.com/article/dn25217-google-flu-trends-gets-it-wrong-three-years-running.html

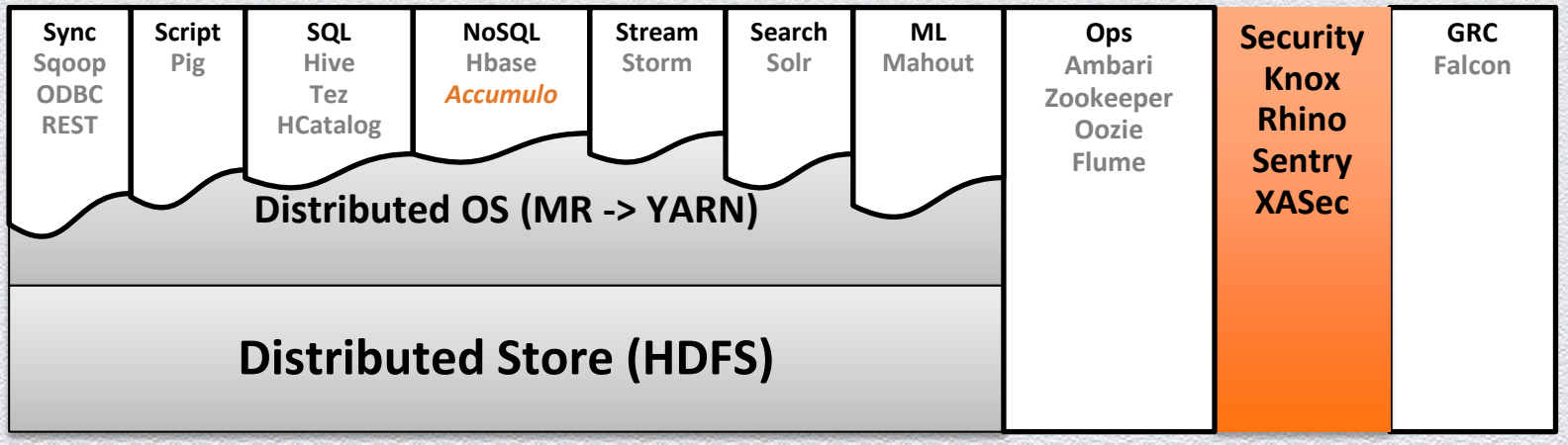
Therefore…A “Build-Your Own” Market



BYO Scale-Out Compute Environment



BYO Scale-Out Compute Environment



BYO Scale-Out Compute Environment

Authentication	Kerberos + “Tokens”
Authorization	Some ACLs
Audit	Scattered Logs
Confidentiality	???

Security
Knox
Rhino
Sentry
XASec

Hadoop Distribution #1 Access, Data, **Perimeter**, Visibility	Hadoop Distribution #2 Policy, Audit, Access, Encryption

BYO Scale-Out Compute Environment

Authentication

Daemons run as single user (hadoop)

user	process
hdfs	namenode, datanode, secondary namenode
mapred	jobtracker, tasktracker, child tasks

group	users
hadoop	hdfs, mapred

```
sudo -u hdfs hadoop fs -rmr /
```

BYO Scale-Out Compute Realities

1. Data Shared
2. Networks Open
3. Nodes Distributed
4. Web Services Open
5. Access Controls Open
6. Clients Unauthenticated

BYO Scale-Out Compute Risk Symptoms

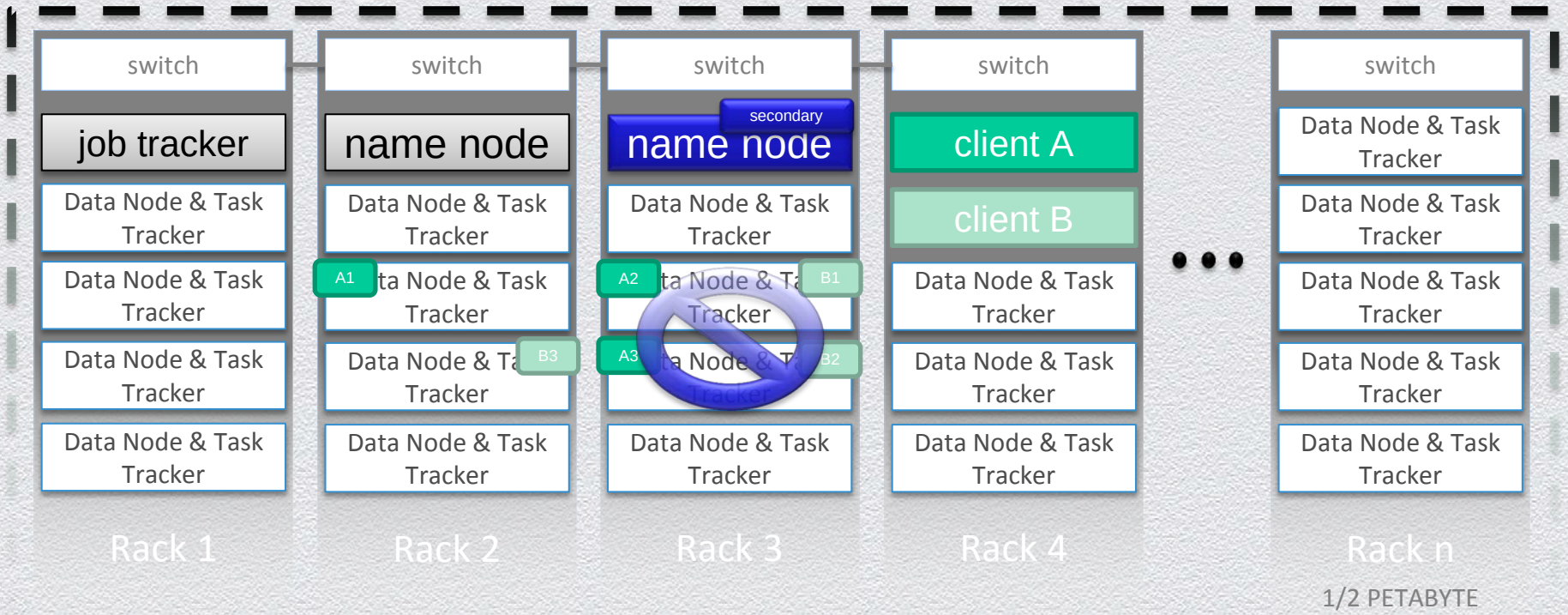
- Kerberos (Randomness, Scalability)
- Job Ticket / Service Delegation
- Data Node Authority (non-ACL)
- API Lack of Multi-Tenancy Awareness
- Local Disk Map Output Access via HTTP Service

Scale-Out Integrity Concerns

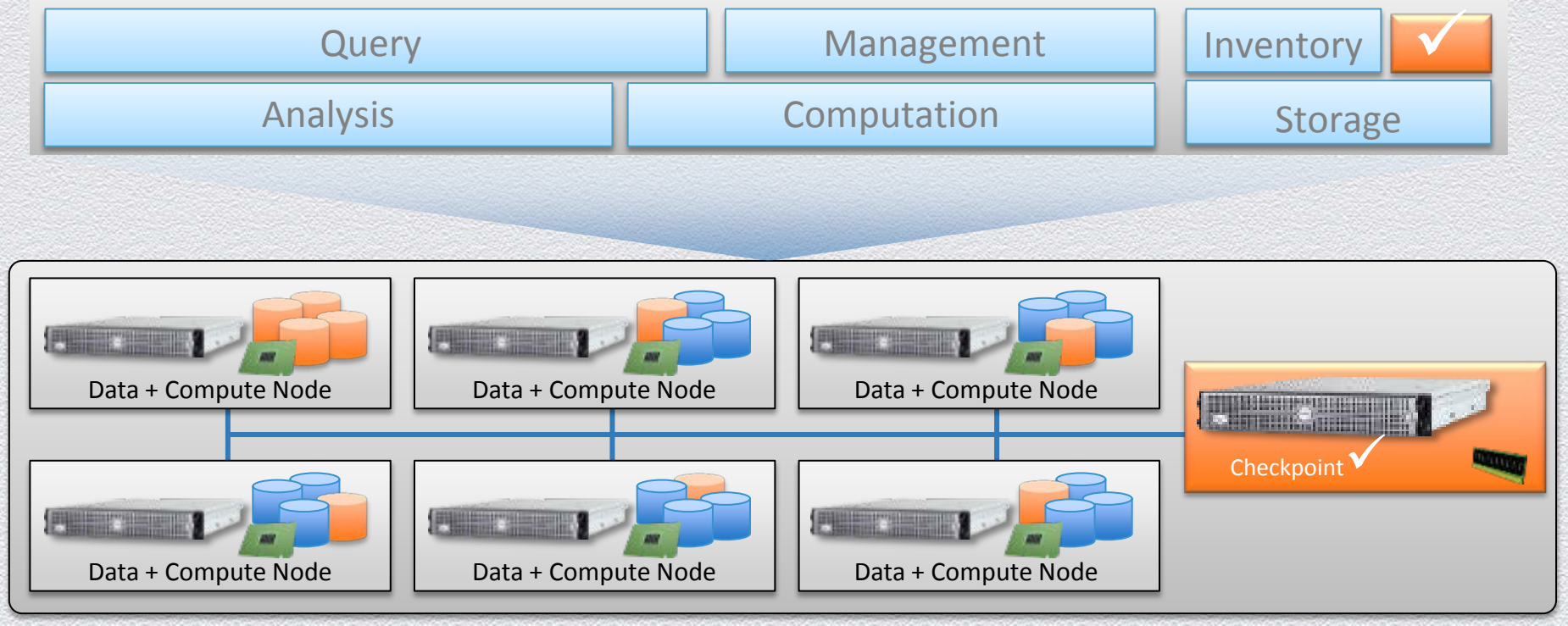
- Tweet Errors
- Balance Sheet Errors

Scale-Out Availability Concerns

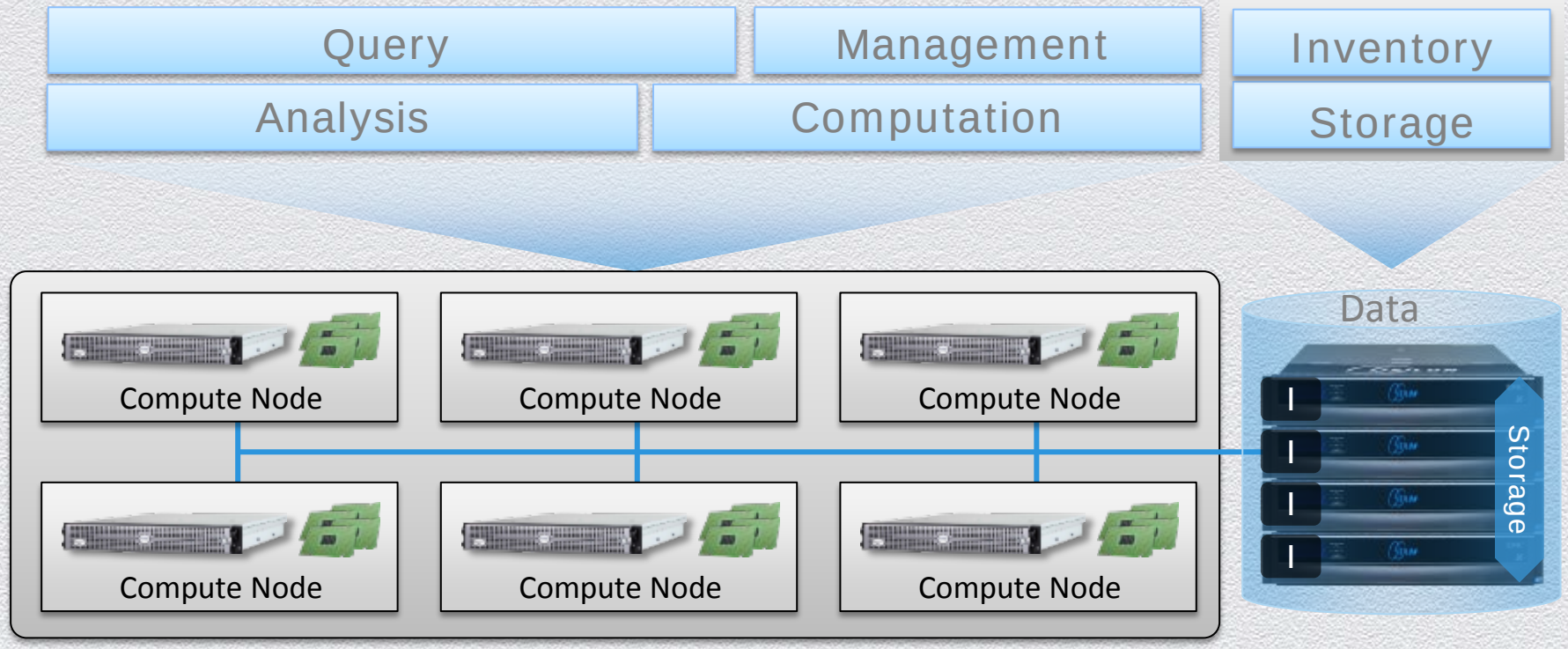
1 Admin : 30,000+ Nodes



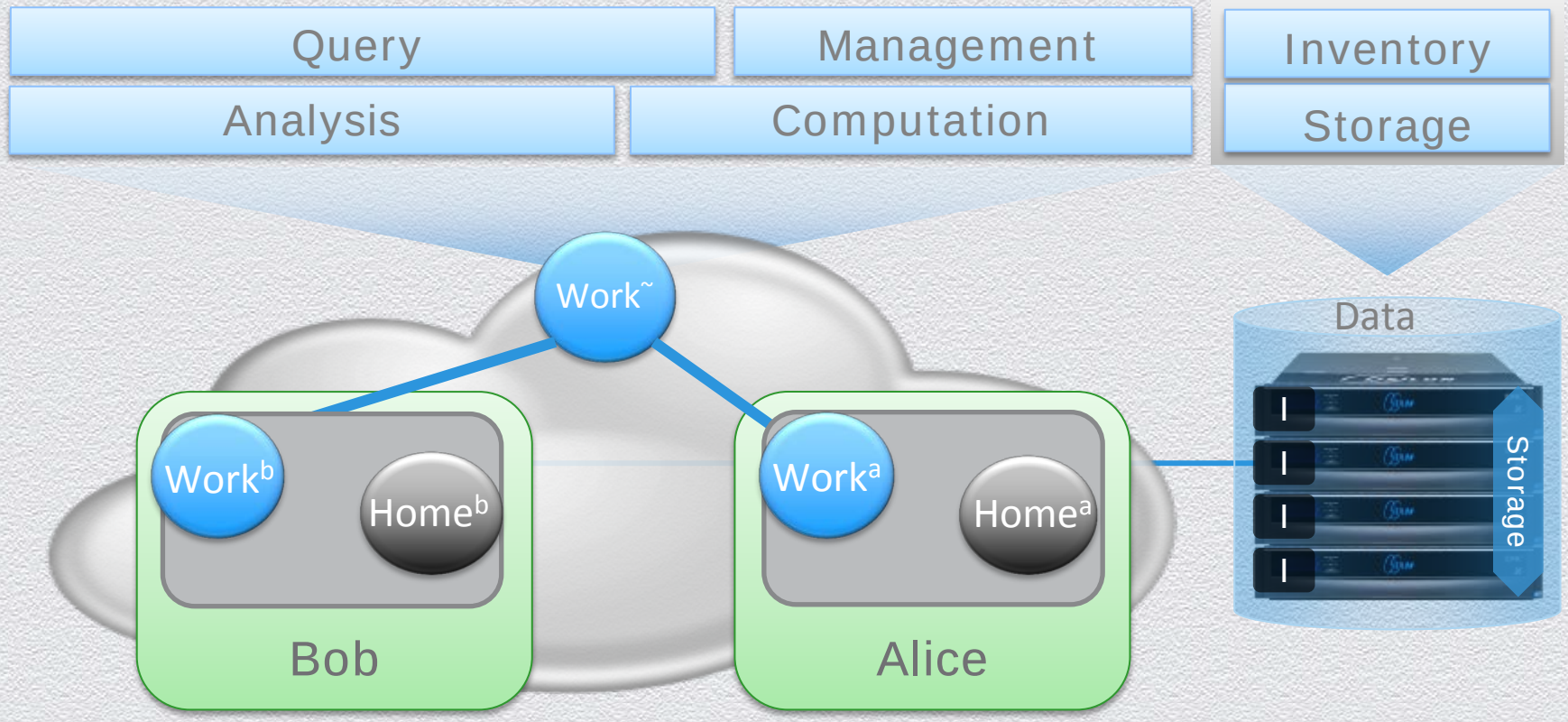
Phase One – Scale-Out



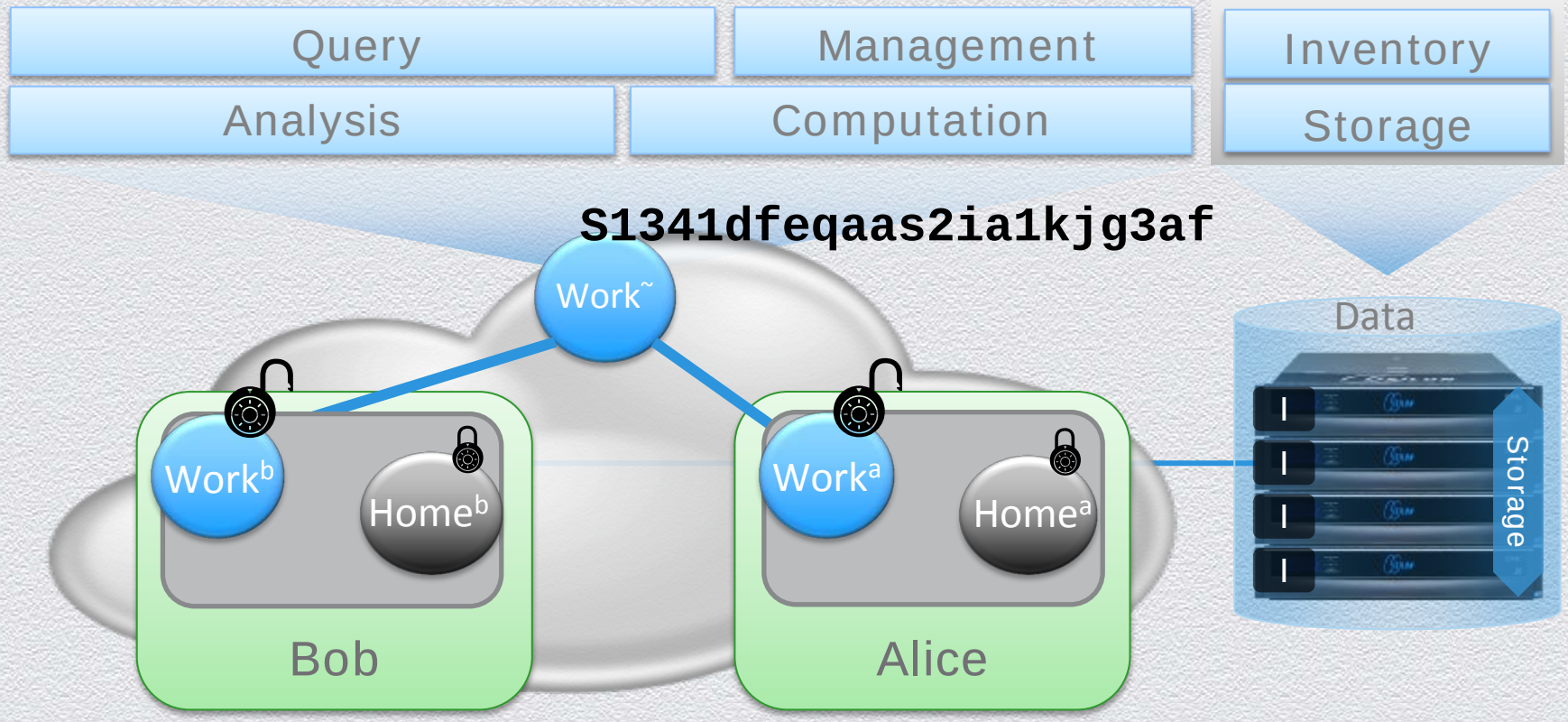
Phase Two – Resilience



Phase Three – Classification



Phase Four – Least Authority



Phase Three – Classification

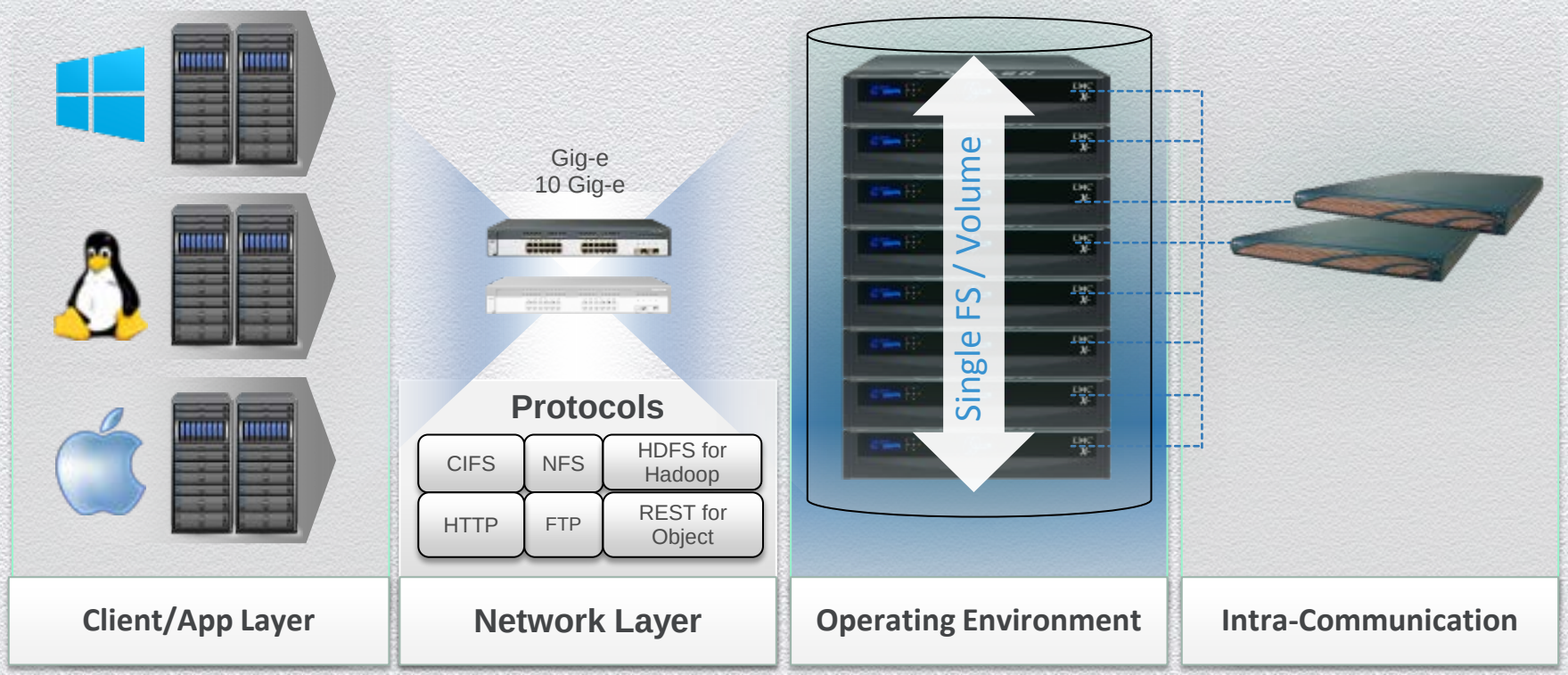
Accumulo Cell-Level Access – Dist Key/Values

- NoSQL (HBase) Performance
- Evolution

2006 – Google BigTable
2008 – NSA [redacted]
2011 – accumulo.apache.org

```
// specify which visibilities we are allowed to see
Authorizations auths = new Authorizations("public"); Scanner scan =
    conn.createScanner("table", auths);
scan.setRange(new Range("user100","user200"));
scan.fetchFamily("attributes");
for(Entry<Key,Value> entry : scan) {
    String row = entry.getKey().getRow(); Value value = entry.getValue();
}
```

Protecting Big Data at Scale



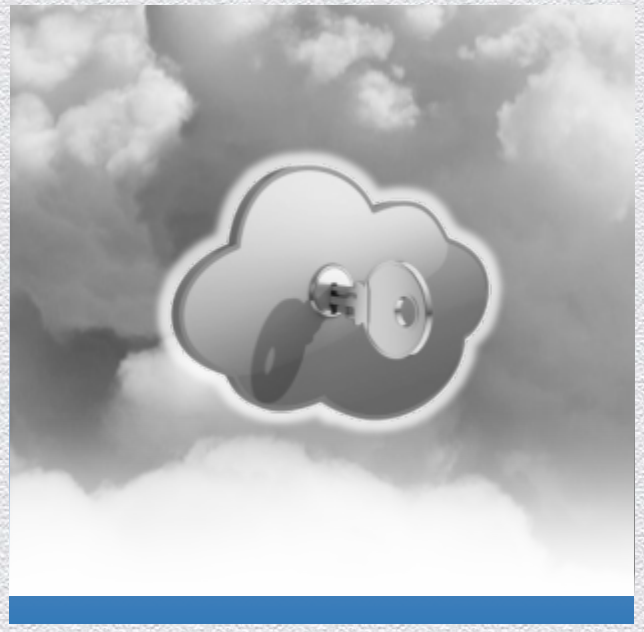
Protecting Big Data at Scale

- Multi-Tenancy Aware
- Full-ACL File Systems
- Kerberos Authentication
- High-Resilience Architecture
  - Name Node Continuous Availability
  - Data Protection (BC/DR, Snapshots, etc.)
  - SEC 17a-4 Compliant WORM

Big Data Trust. Redefined

Transparency

Relevance



Resilience

Thank You!

@daviottenheimer