Babar-ians at the Gate: Data Protection at Massive Scale

Davi Ottenheimer (@daviottenheimer)
Senior Director of Trust, EMC

INTRODUCTION

@daviottenheimer

- 94-04 DFIR / Audit / Security Ops
- 2004 UCSC, UARC (NASA)
- 2005 West Marine
- 2006 Yahoo!
- 2007 BGI
- 2008 ArcSight
- 2009 `flyingpenguin`
- 2013 EMC


```
    __
  -'  o)=-
 /.-.'
//  |\
||  |'
_,:(_/_
```

Trusted IT

Transparency

Relevance

Resilience

Storage + Analysis = *Winning*

- Store Everything

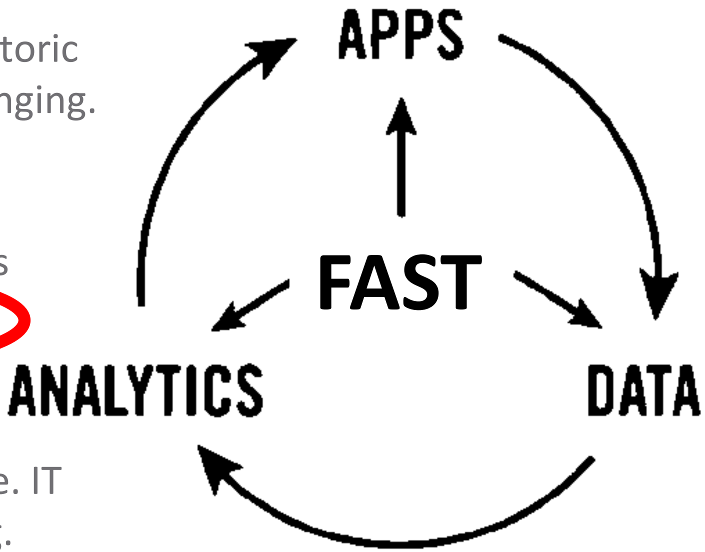
  Structured and unstructured. Internal and external. Historic and real-time streamed data. The nature of data is changing.

- Analyze Anything

  Analytics forward looking and predictive. Data analytics complements business intelligence "rear-view mirror"

- Build the Right Thing

  Apps take big data insights and turn into business value. IT has role in rapid development, deployment and scaling.



“Predictive Analysis Tool”
Rear View Mirror

Ray Harroun’s Marmon Wasp, Inaugural Indy 500 Winner, 1911


SO MANY HORSES!

1996…

"hardware cost for **24**-node **BEOWULF** cluster **$57K** — compared to commercial supercomputers between **$10m and $30m**"

"**30** men's heft of grasp in the gripe of his hand"

Parveen S. Thampi, "Linux Operating System: A Freeware and the Net War," *Computers Today*, October 31, 1998

1996…

"hardware cost for **24**-node **BEOWULF** cluster **$57K** — compared to commercial supercomputers between **$10m and $30m**"

LOW LATENCY
FAULT-TOLERANCE
SCALABILITY

…2006

HADOOP

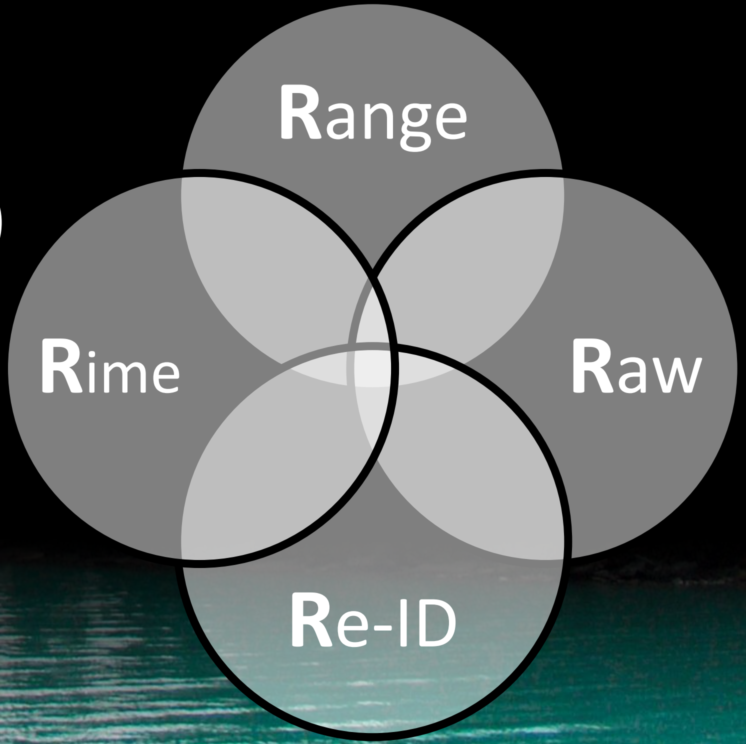
Parveen S. Thampi, "Linux Operating System: A Freeware and the Net War," *Computers Today*, October 31, 1998


TRUSTED?

2014 “Data Lake” Definition

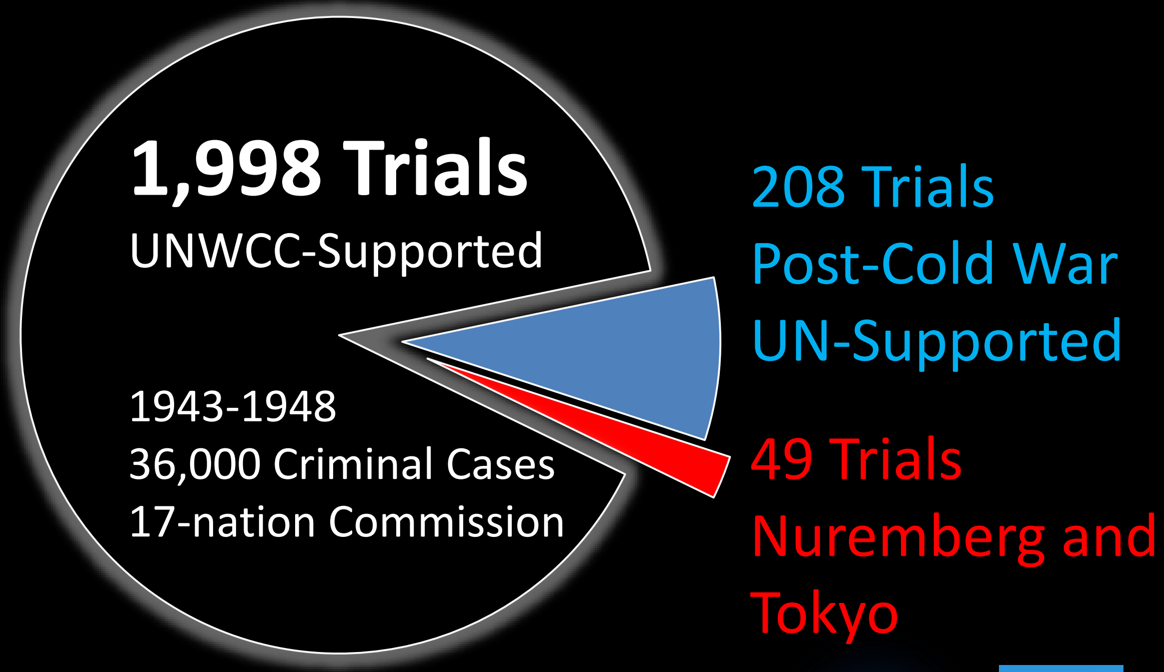
Four RRRRs

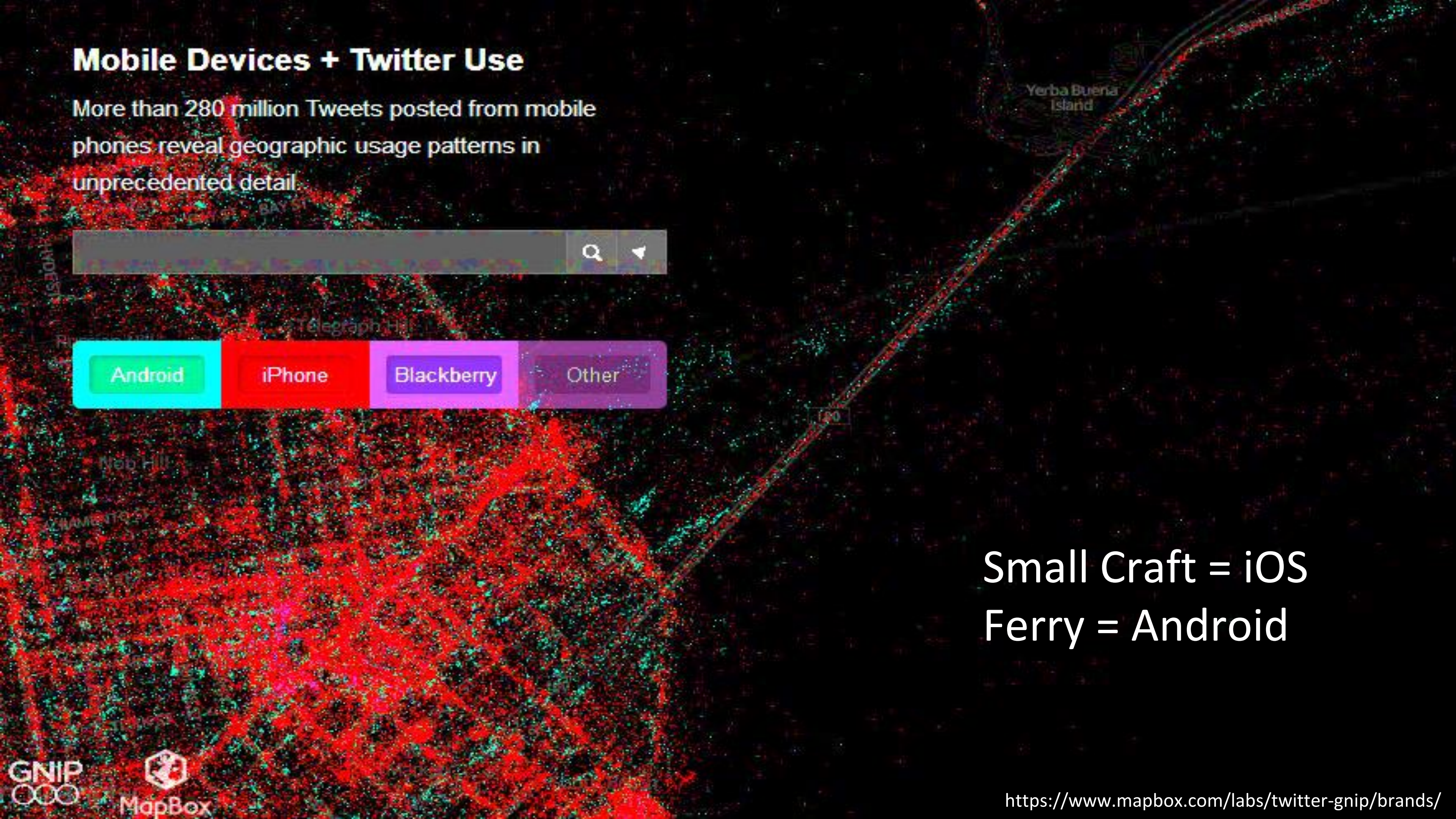
1. Range of Sources (Volume)
2. Raw Formats (Variety)
3. Real-Time (Velocity)
4. Re-ID (Vulnerability)



UN War Crimes Commission

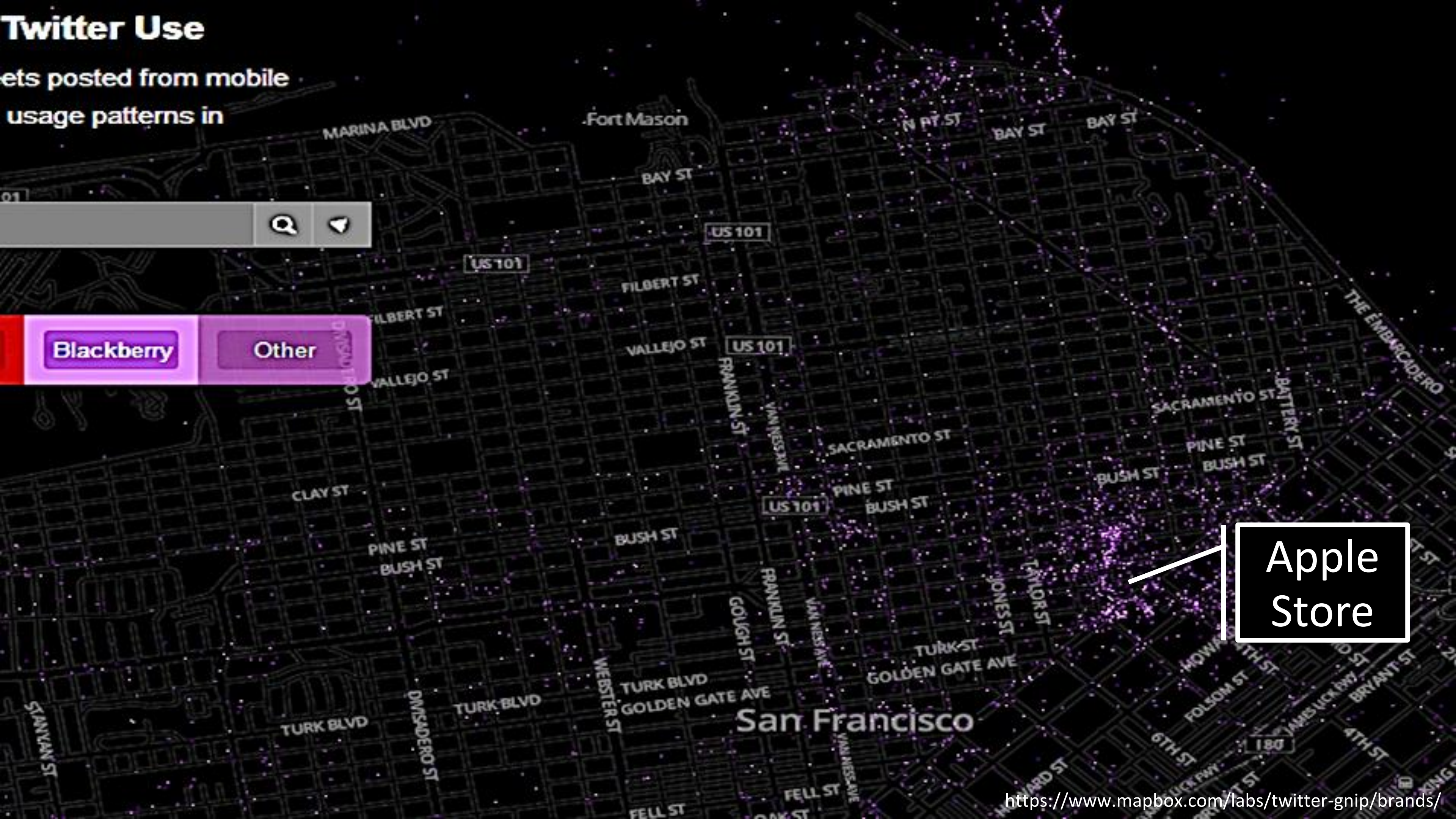
"War Crimes Tried" Plesch-Sattler, Aug 2014





Small Craft = iOS
Ferry = Android

https://www.mapbox.com/labs/twitter-gnip/brands/


Twitter Use
ets posted from mobile
usage patterns in
Blackberry
Other
MARINA BLVD
Fort Mason
BAY ST
US 101
FILBERT ST
VALLEJO ST
SACRAMENTO ST
PINE ST
BUSH ST
CLAY ST
TURK BLVD
GOLDEN GATE AVE
FRANKLIN ST
VAN NESS AVE
GOUGH ST
WEBSTER ST
DIVISADERO ST
STANYAN ST
JONES ST
TAYLOR ST
BATTERY ST
THE EMBARCADERO
FOLSOM ST
BRYANT ST
4TH ST
6TH ST
FELL ST
San Francisco
Apple Store
https://www.mapbox.com/labs/twitter-gnip/brands/

Sensors per Turbine

100 Physical, 300 Virtual

Data per Blade

500 GB/day

(1mo = Library of Congress Print Collection)

London Transportation

Bakerloo	Suspended
Central	Suspended
Circle	Suspended
District	Suspended
East London	Suspended
Hammersmith & City	Suspended
Jubilee	Severe delays
Metropolitan	Suspended
Northern	Good service
Piccadilly	Part suspended
Victoria	Suspended
Waterloo & City	Suspended

Data Feed of Traffic Disruption

wunderground.com

Modelling and ML

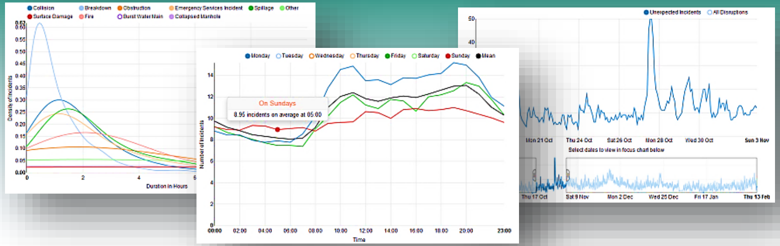
scikits learn

MADlib

python

Surface

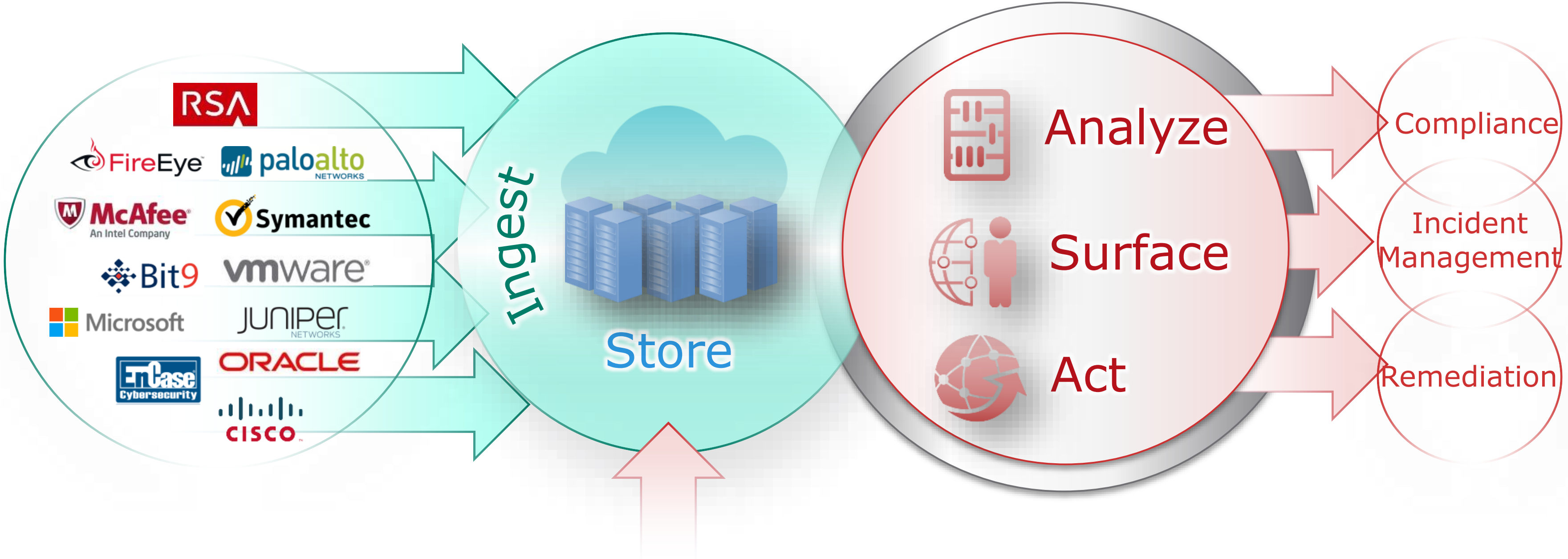
d3.js & NVD3
Interactive SVG Figures



Pivotal

blackhat

Continuous Security Monitoring



Public & Private Threat Intelligence

“Security Capabilities of Central Data Lake Technologies Still Embryonic…”

“The Data Lake Fallacy: All Water and Little Substance”, Nick Heudecker and Andrew White, Gartner, July 28, 2014
https://www.gartner.com/newsroom/id/2809117

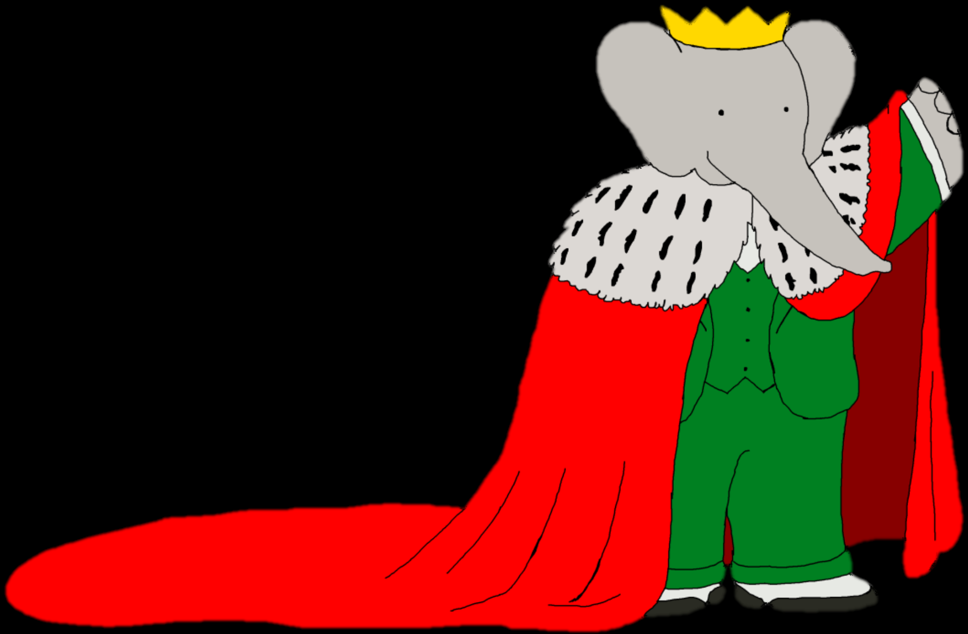
Babar-ians at the Gate?

1931 French Colonialism

- Leave Behind Ignorance of Instincts
- Backward Versus Developed People
- Transformation to Polite and Decent Human Beings
- New Rulers Must Be Outsiders, Instructed in New Ways
- Rationalize Some Countries Have All, Others Nothing

https://www.goodreads.com/book/show/163331.The_Empire_s_Old_Clothes

Emperor’s New Controls



“By 2017, Big Data will be the norm for data management…”

Top Emerging Technologies To Watch: Now Through 2018, Brian Hopkins and Frank E. Gillett, Forrester, Feb 7, 2013 http://blogs.forrester.com/brian_hopkins/13-02-07-forresters_top_15_emerging_technologies_to_watch_now_to_2018

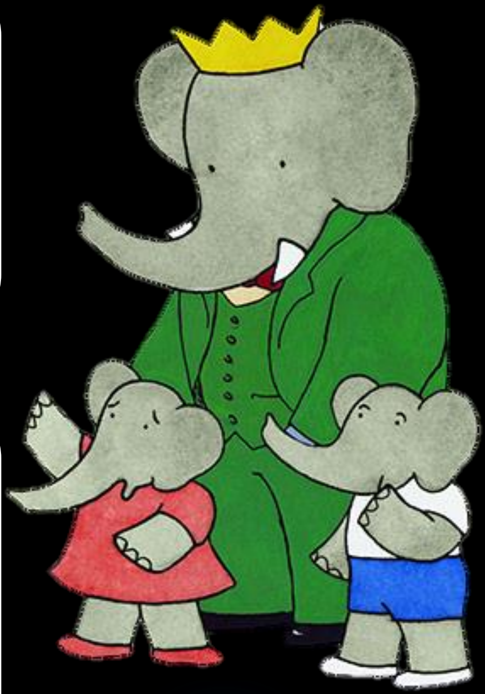
Control Opportunities

C

Exposure
- Mixed
- Insider
- Outsider

I

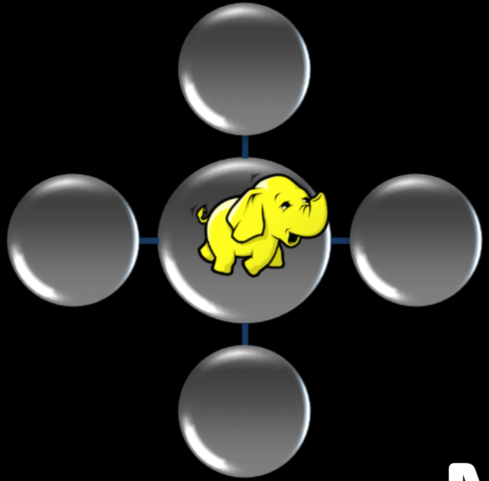
Data Loss/Corruption
- Backup
- Restore
- Snapshot/Rollback

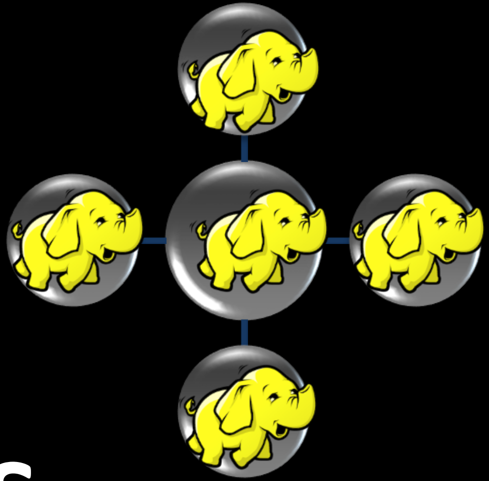


A

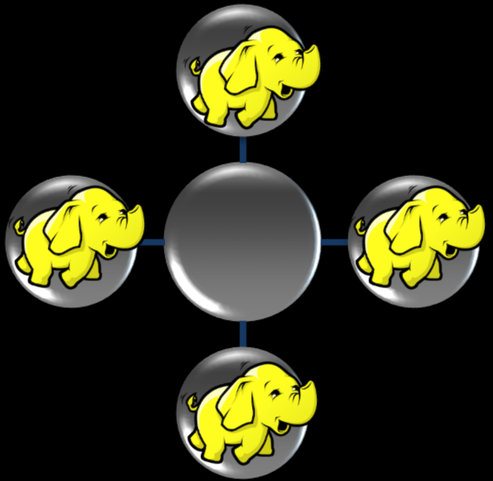
Non-Compliance
- Regulations
- Internal Policy
- External Standards

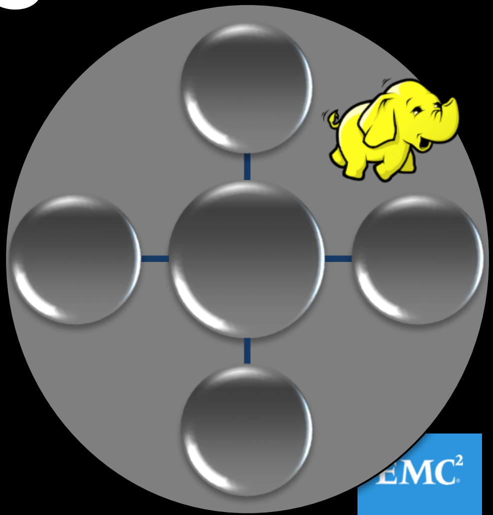
Outages
- Repairs
- Changes
- Disasters

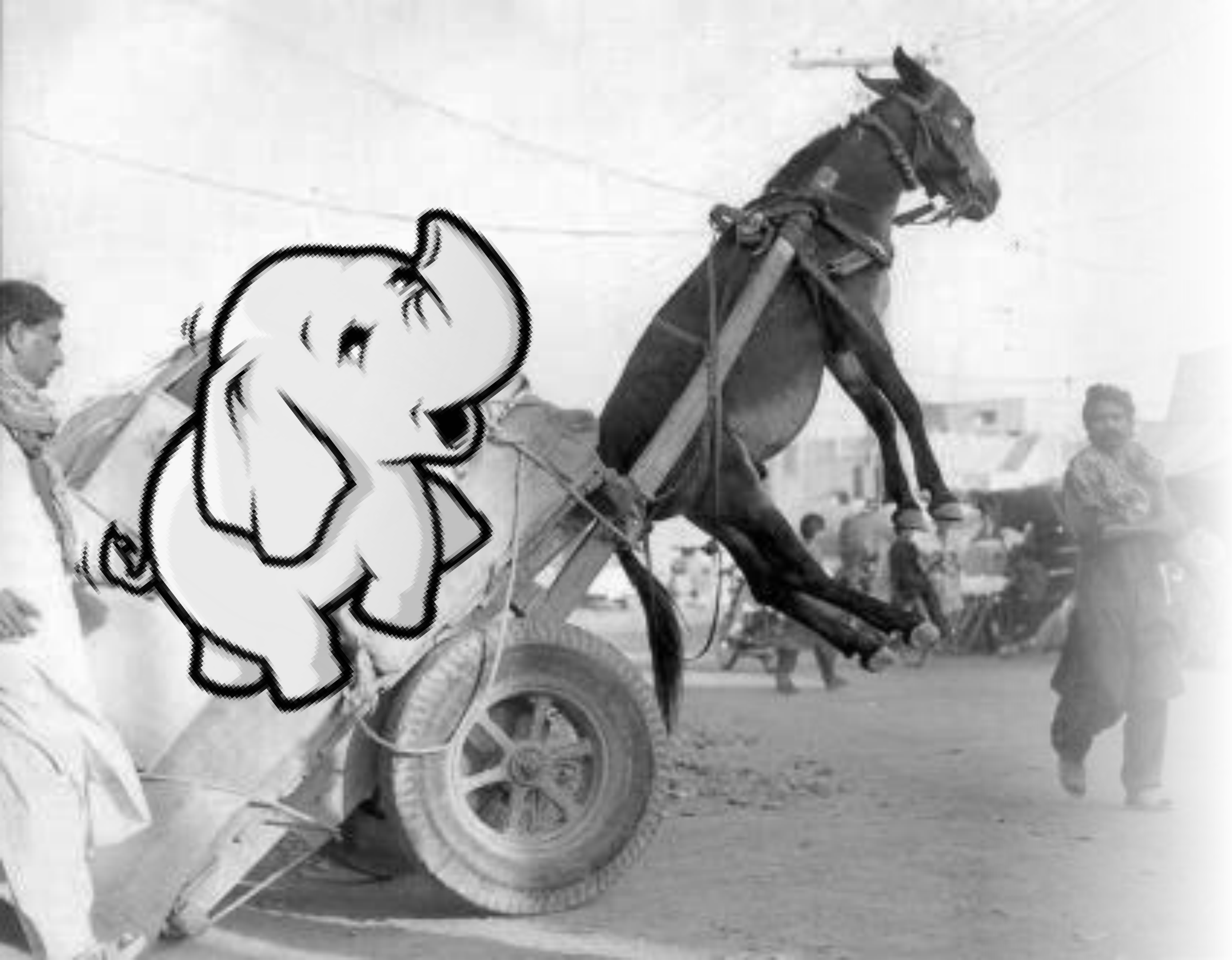




MANY MODELS







NONE

LED BY

IT

HADOOP CRITICISMS

- Lack of Maturity (Security, Audit…)
  – File Immutability By Design Yet Mutability in Reality
  – Resiliency By Design Yet Outages/Latency in Reality
  – Security Mode By Design Yet Insecure in Reality
- Fractured and Complex
- Non-Technical Usability

Threat Assumptions

No MitM
Users not root
No NIC control

HADOOP ECOSYSTEMS

Sync
Sqoop
ODBC
REST

Script
Pig

SQL
Hive
Impala
HAWQ

NoSQL
Hbase
Accumulo

Stream
Storm

ML
Mahout

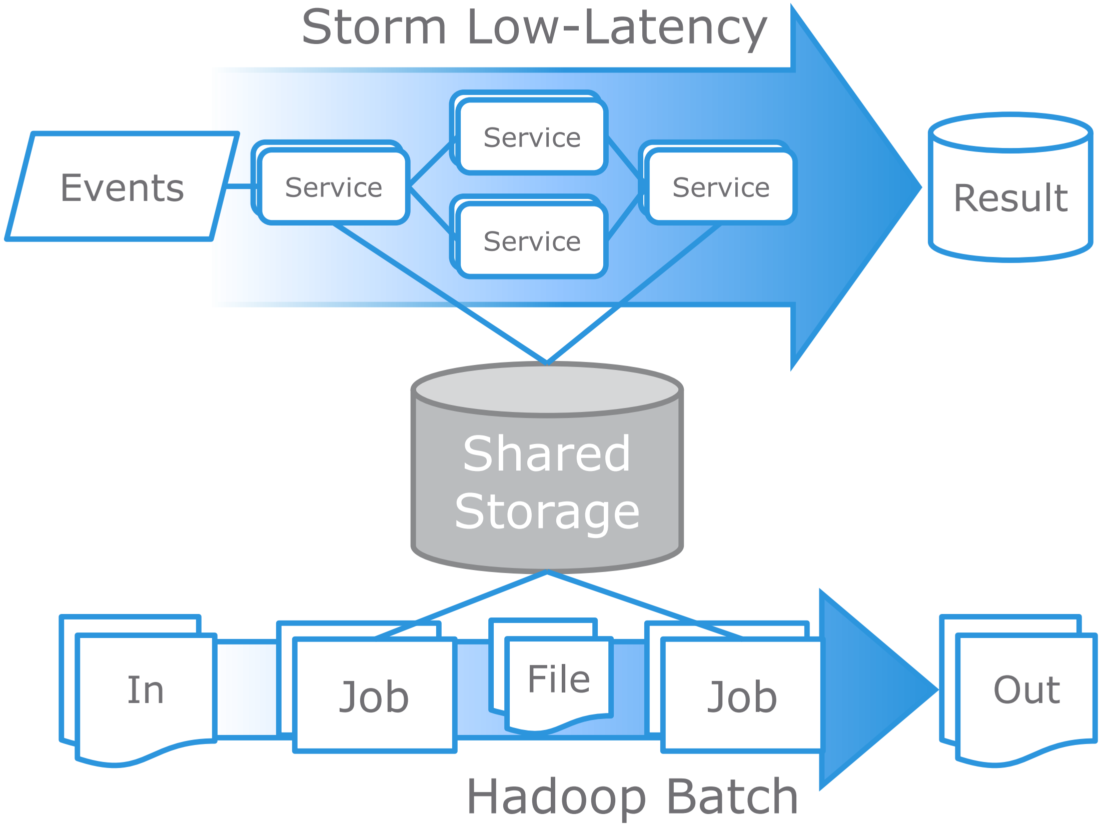
Ops
Ambari
Zookeeper
Oozie
Flume

Security
Knox
Rhino
Sentry
XASec
Falcon

Distributed Compute (YARN & MR)

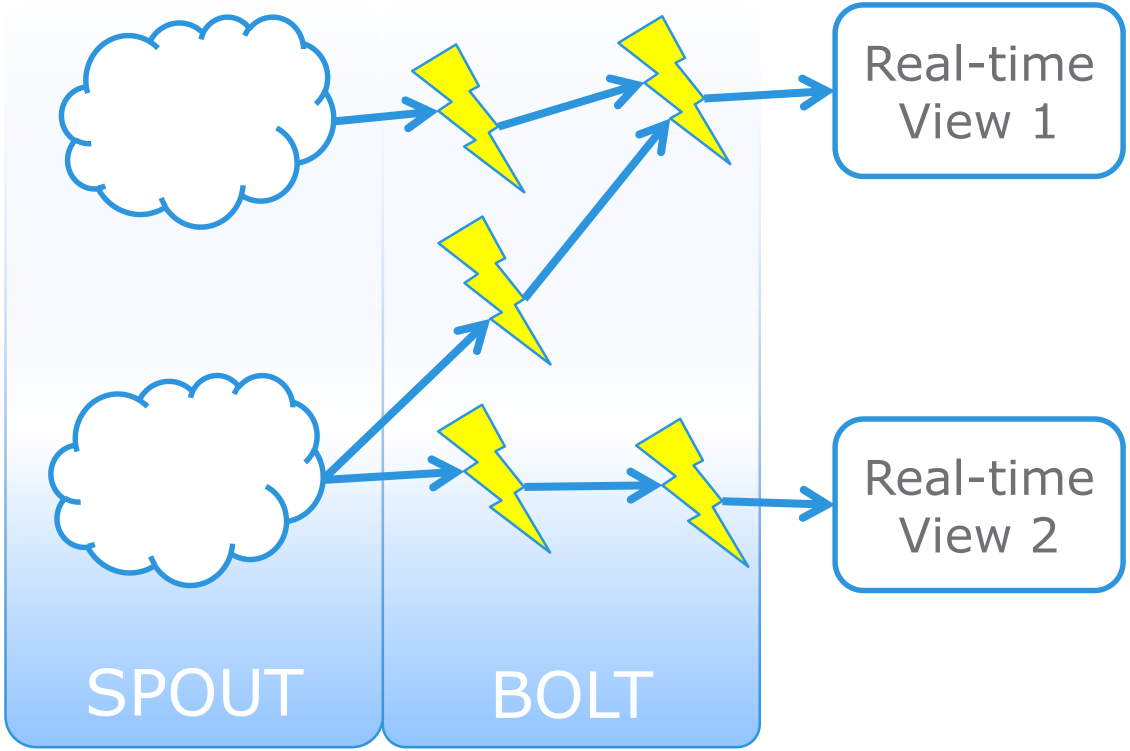
Distributed Store (HDFS)

Example: Storm



- Stream data (spout) to execution agents (bolts)
- Process over 1m tuples/second per node
- Scalable, fault-tolerant

Example: Storm



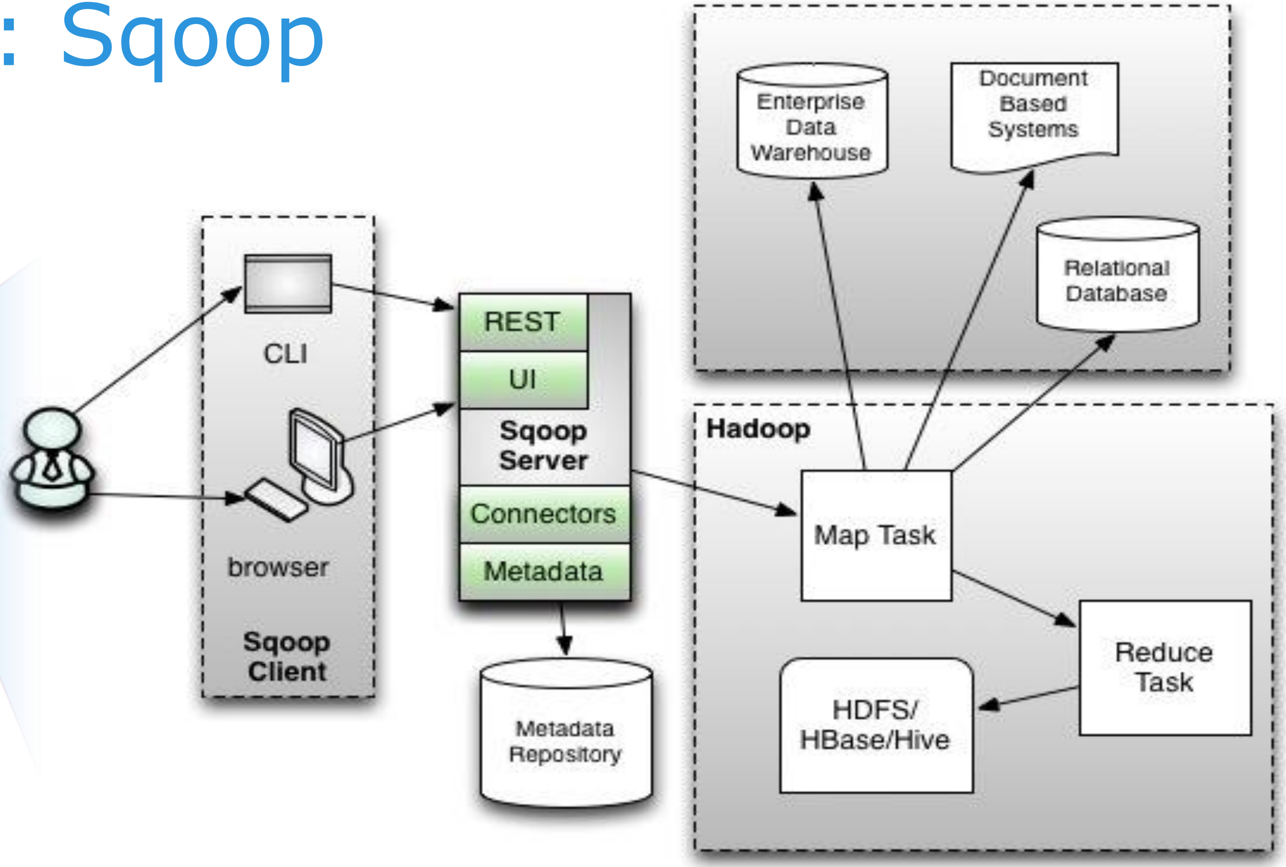
"Storm will hang when you do weird things with it"

"A focus for upcoming releases will be security and multi-tenancy"

doing a 'rm -rf /hadoop/storm' worked for me

Example: Sqoop

Sqoop
ODBC
REST



"Sqoop2 Does Not Support Security"

Yarn Authorization

- Jobs and Queues
  - Hadoop, MR/Yarn, Pig, Oozie, Hue…
- Data
  - HDFS, Hbase, Hive Metastore, Zookeeper
- Queries
  - HAWQ, Impala, Drill

Level	Value
Everyone	"*"
No One	" "
Users and Groups	"user1, user2 group"
Users no Groups	"user1, user2 "
Groups no Users	" group"

https://hadoop.apache.org/docs/stable/Secure_Impersonation.html

Yarn Authorization

hadoop.security.authorization Set to True in core-site.xml
Default Setting = "*"

Property	Description	Value
security.containermanagement.protocol.acl	ApplicationMasters communicate with NodeManager	"*"
security.applicationmaster.protocol.acl	ApplicationMasters communicate with ResourceManager	"*"
security.job.task.protocol.acl (MR1 uses security.task.umbilical.protocol.acl)	MR2 tasks report task progress	"*"

hadoop-policy.xml

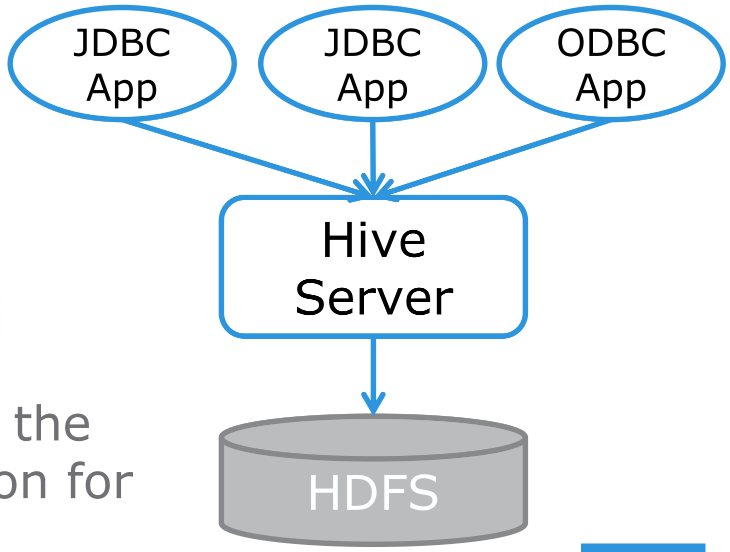
Other Controls Still Needed!

SQL on Hadoop

MPP SQL query execution against HDFS

- Hive jdbc:hive2://<host>:<port>/<db>;principal=<Server_Principal_of_HiveServer2>
- Stinger
- Apache Drill
- Cloudera Impala
- Pivotal GEMFIRE, HAWQ

"Currently native client only supports the NULL cipher with mutual authentication for SSL socket communications."



Example: Authentication

Delegation Token Exposure

- Delegation tokens in URLs may end up in log files while still valid (passed as partial referrer)
- Users can pass links including delegation tokens

Solution: Use header instead of URL parameters

https://issues.apache.org/jira/browse/HADOOP-10799

Example: Authentication

Password and Secrets Exposure

“Need ability to eliminate passwords and secrets in clear text within configuration files or code”

Solution: API with “CredentialProviders” URLs

https://issues.apache.org/jira/browse/HADOOP-10607

Danger Signs in Data Lakes

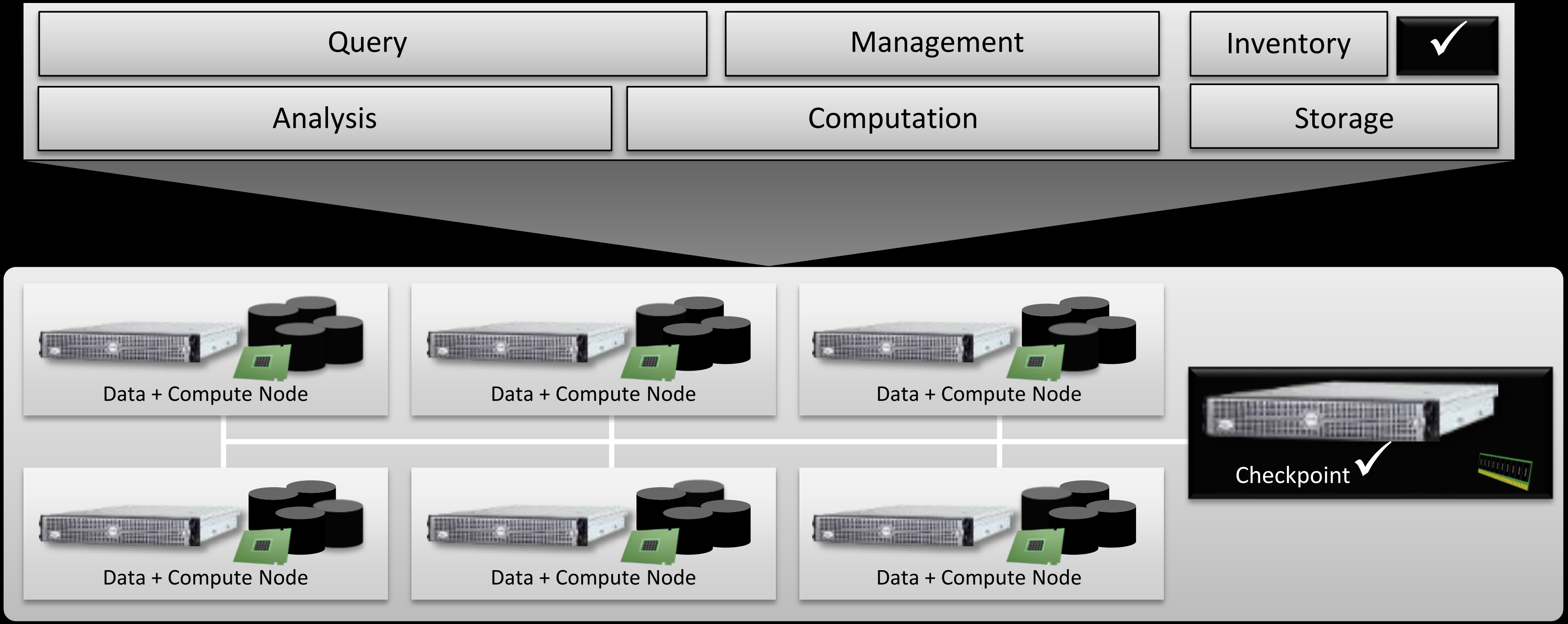
- HTTP Services (Local Disk Map Output)
- Job Ticket / Service Delegation
- Data Node Authority (non-ACL)
- API Abuse (Lack of Multi-Tenancy Awareness)
- Kerberos (Randomness*, Performance)

* http://www.cryptopp.com/docs/ref/class_auto_seeded_random_pool.html#_details

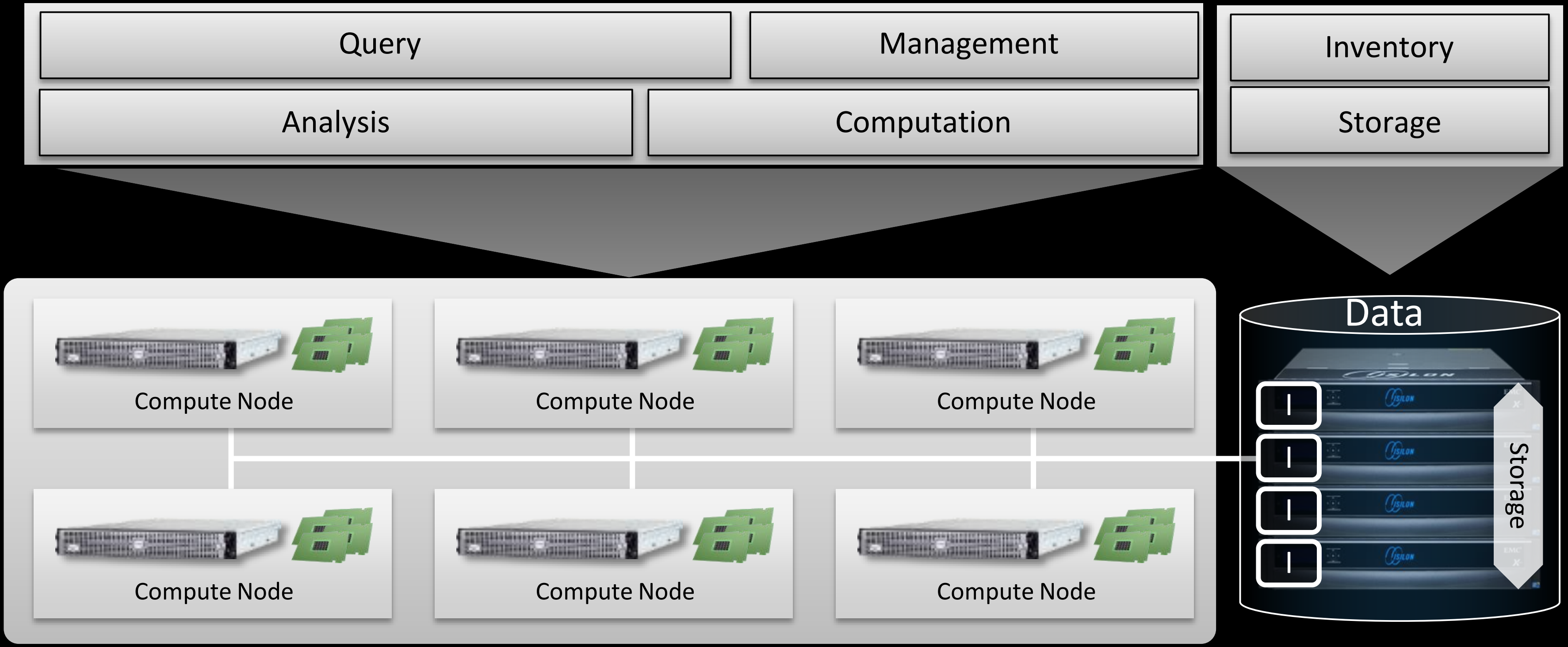
7. Governance, Risk, Compliance (GRC)

1. Net and System Security
2. Data Protection
3. Vulnerability Mgmt
4. Access Control
5. Monitoring
6. Policies

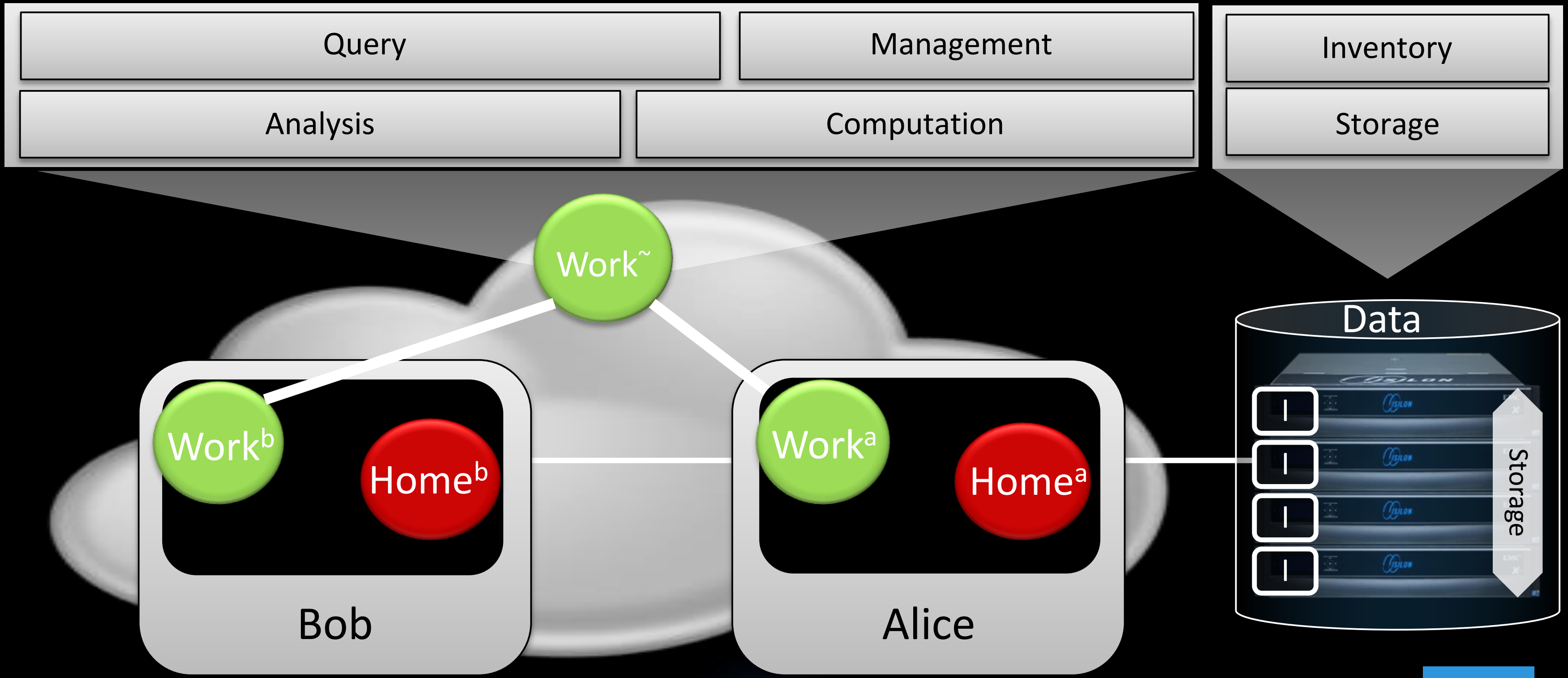
Phase One - Scoping



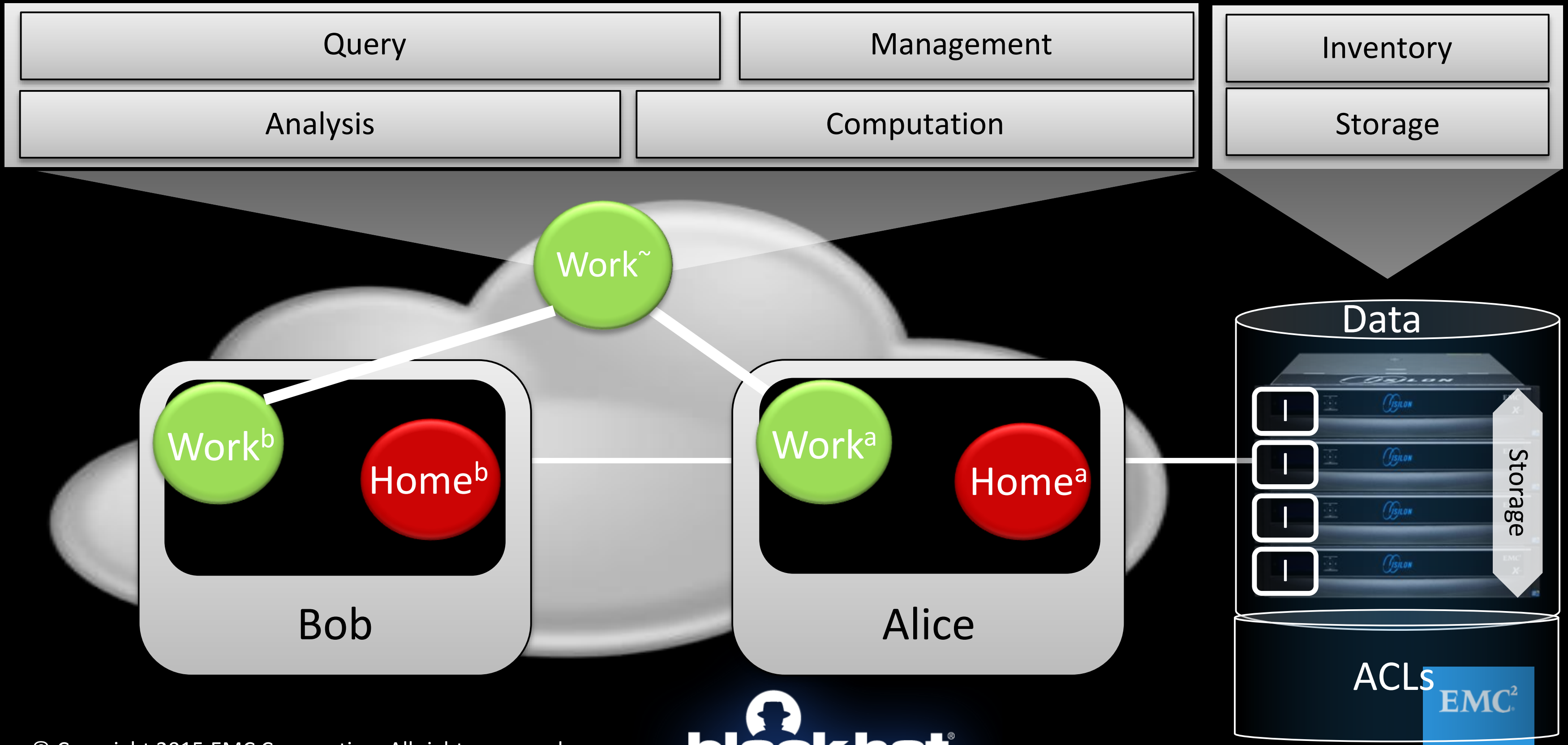
Phase Two - Availability



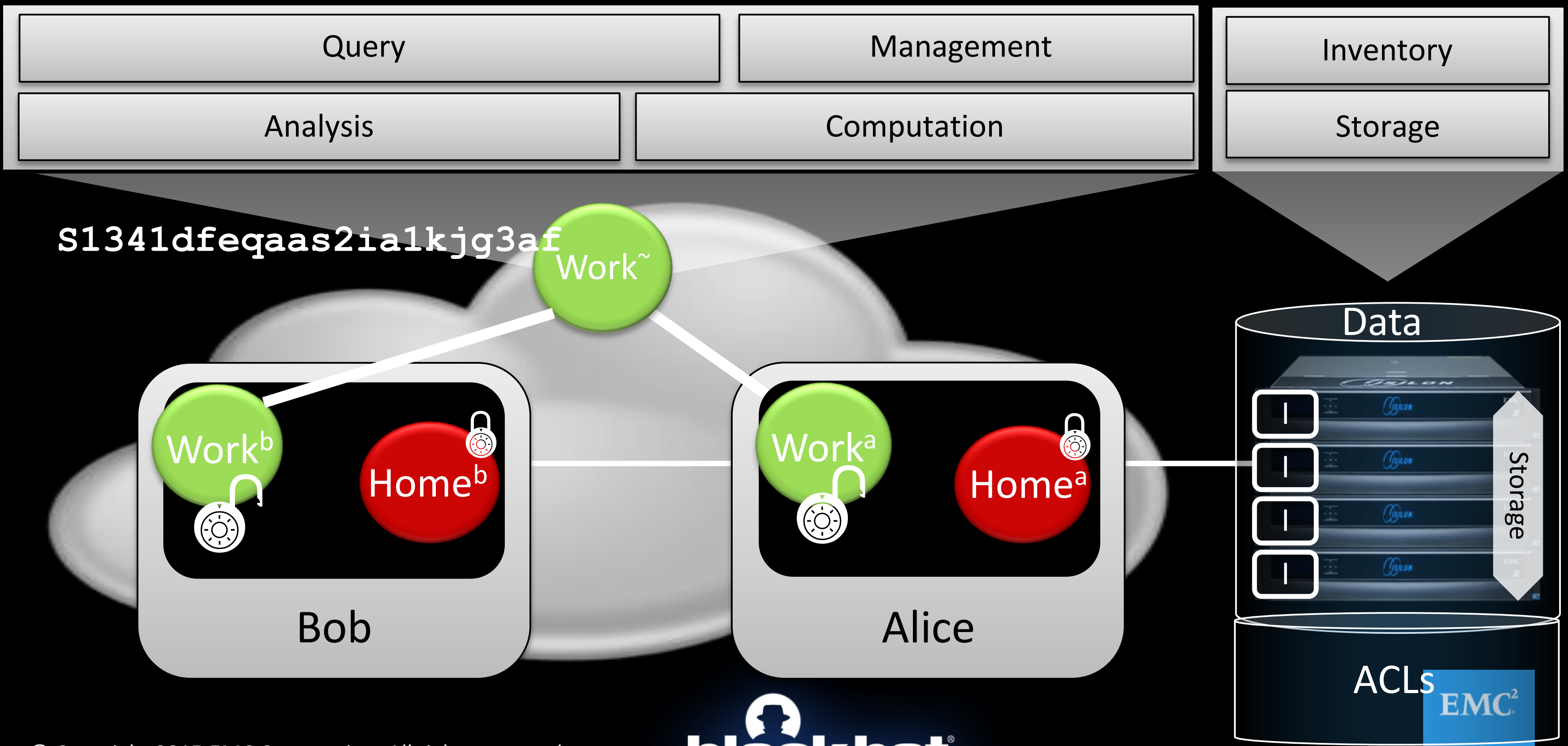
Phase Three – Classification



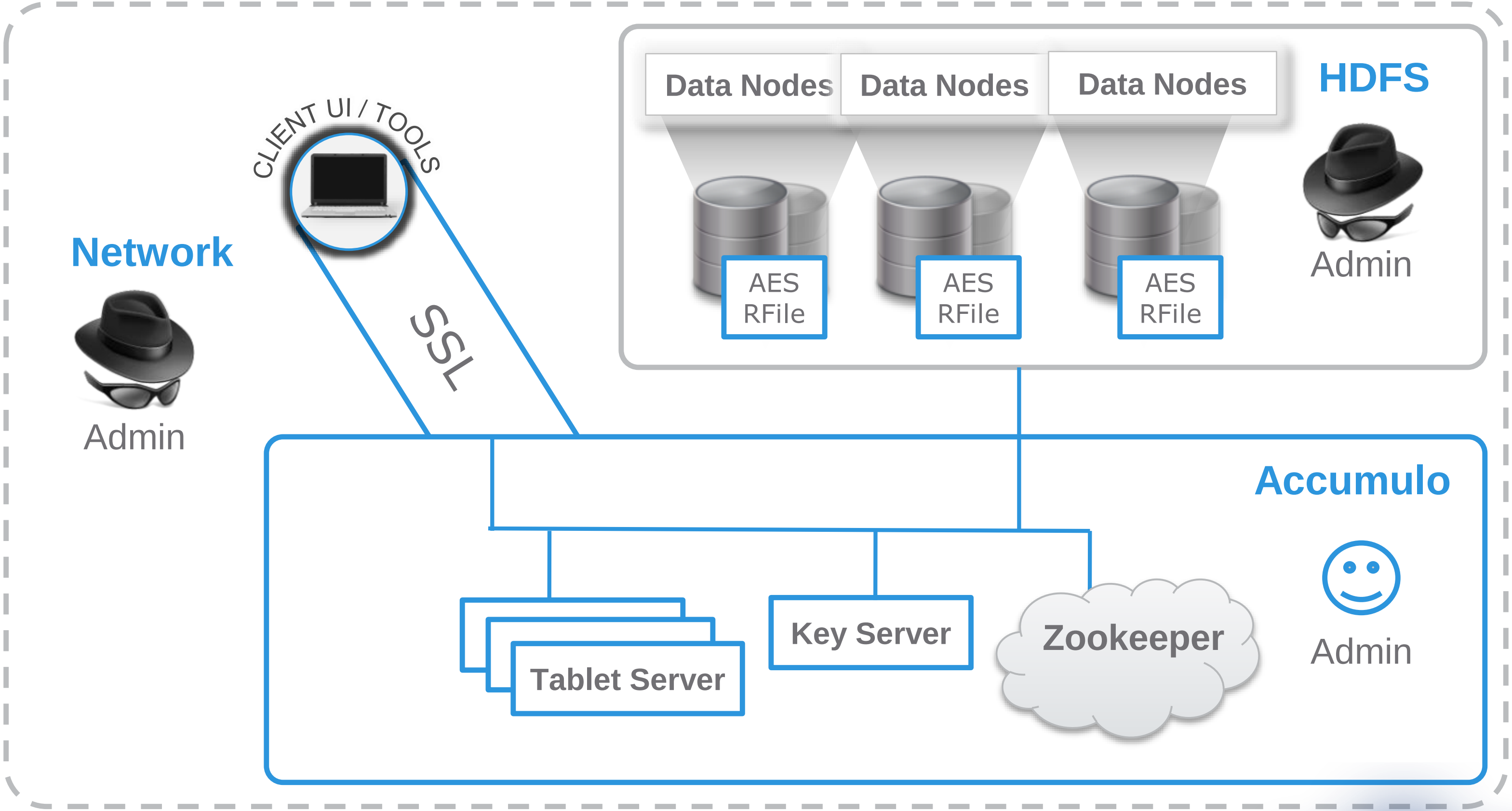
Phase Four – Authority



Phase Five – Least Authority

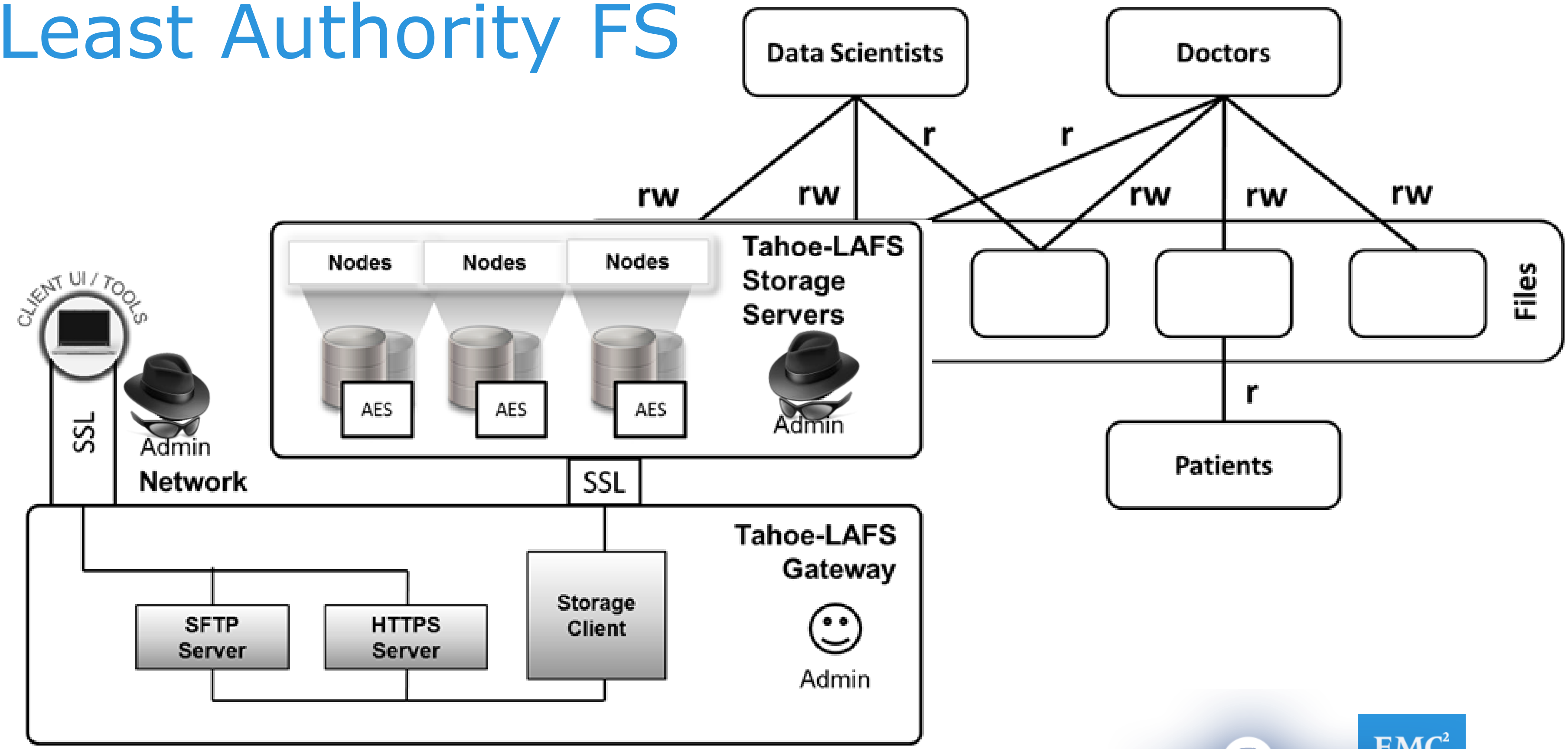


ACCUMULO

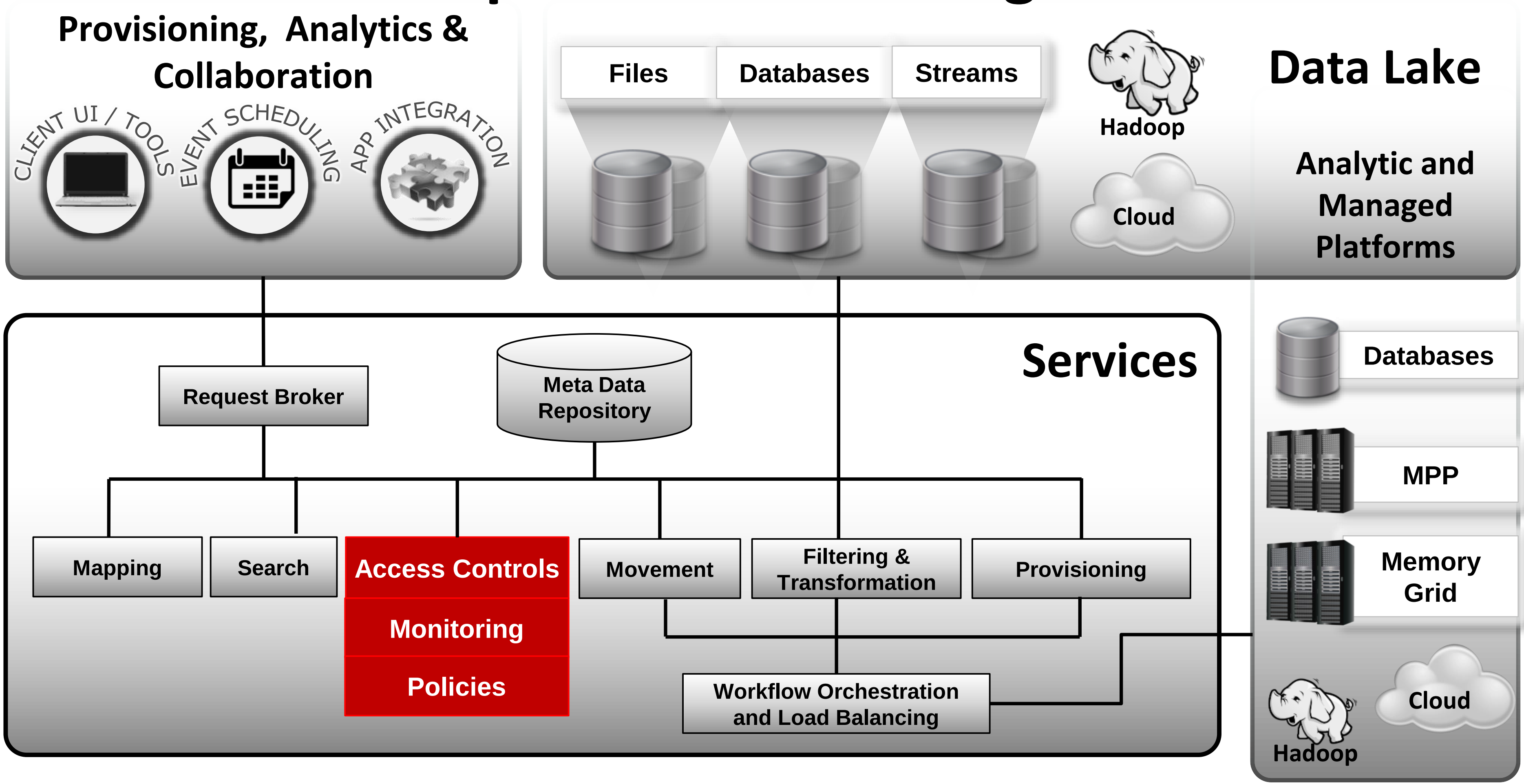


Data Nodes
Data Nodes
Data Nodes
HDFS
AES RFile
AES RFile
AES RFile
Admin
CLIENT UI / TOOLS
Network
Admin
SSL
Accumulo
Admin
Tablet Server
Key Server
Zookeeper

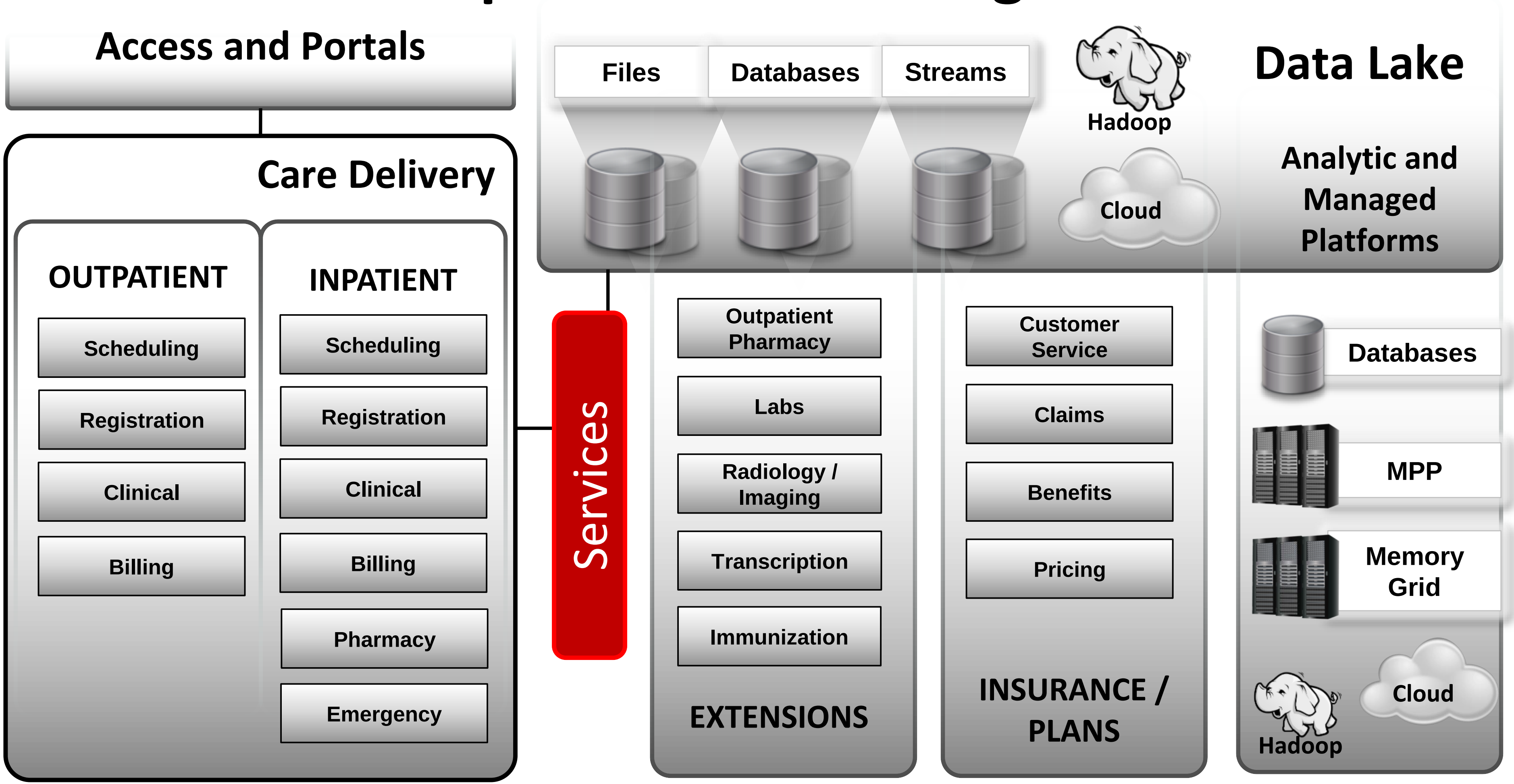
TAHOE
Least Authority FS



Example: HealthCare Big Data



Example: HealthCare Big Data



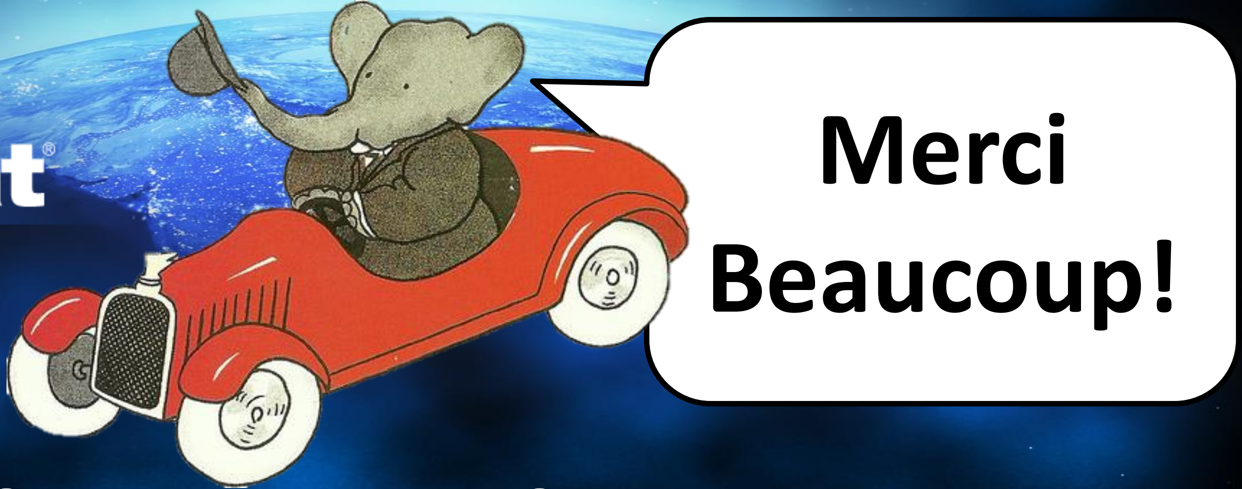
Trusted IT

Transparency

Relevance

Resilience

black hat®



Babar-ians at the Gate: Data Protection at Massive Scale

Davi Ottenheimer (@daviottenheimer)
Senior Director of Trust, EMC