



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



Trusted IT

Transparency



Relevance



Resilience



Storage + Analysis = *Winning*

- **Store Everything**

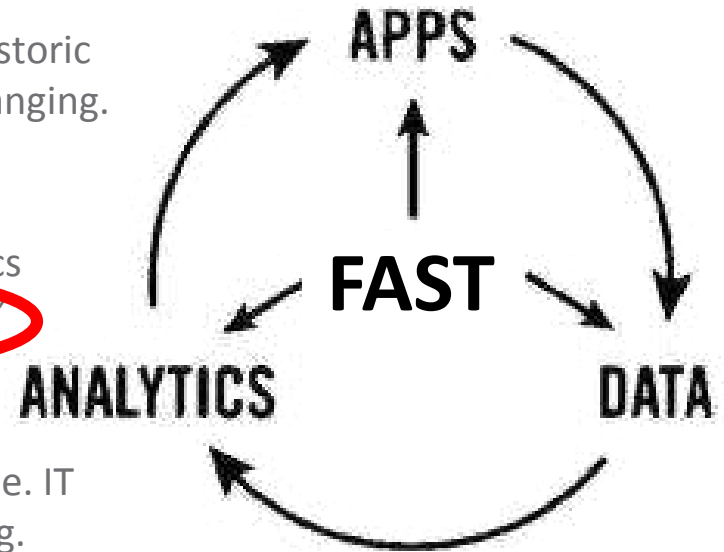
Structured and unstructured. Internal and external. Historic and real-time streamed data. The nature of data is changing.

- **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**”



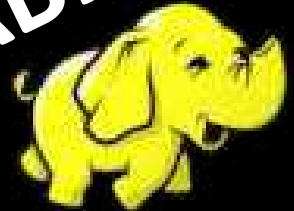
“30 men's heft of grasp in the gripe of his hand”

1996...

“hardware cost for **24-node BOWWOLF** cluster **\$57K** — compared to commercial supercomputers between **\$10m and \$30m**”



LOW LATENCY
FAULT-TOLERANCE
SCALABILITY



...2006

HADOOP

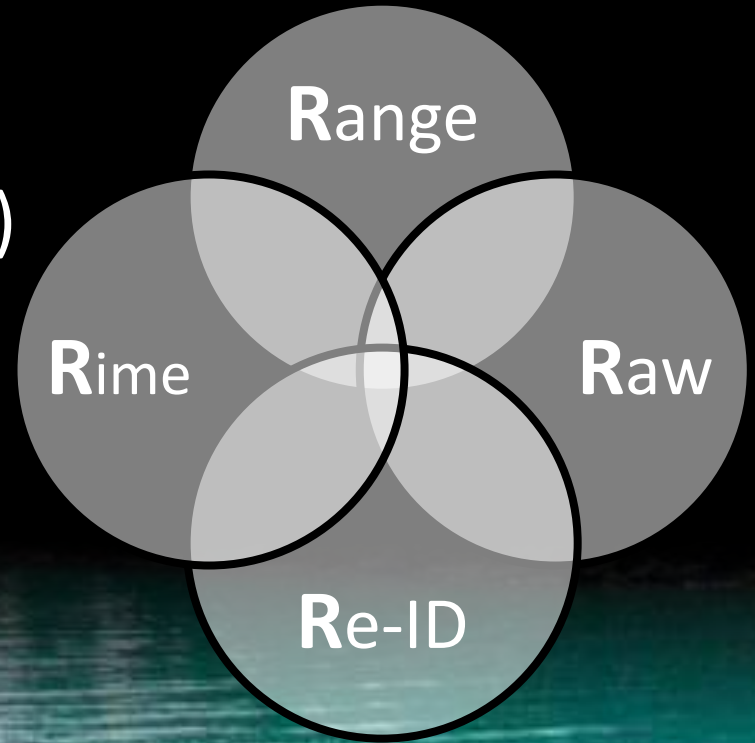


TRUSTED?

2014 “Data Lake” Definition

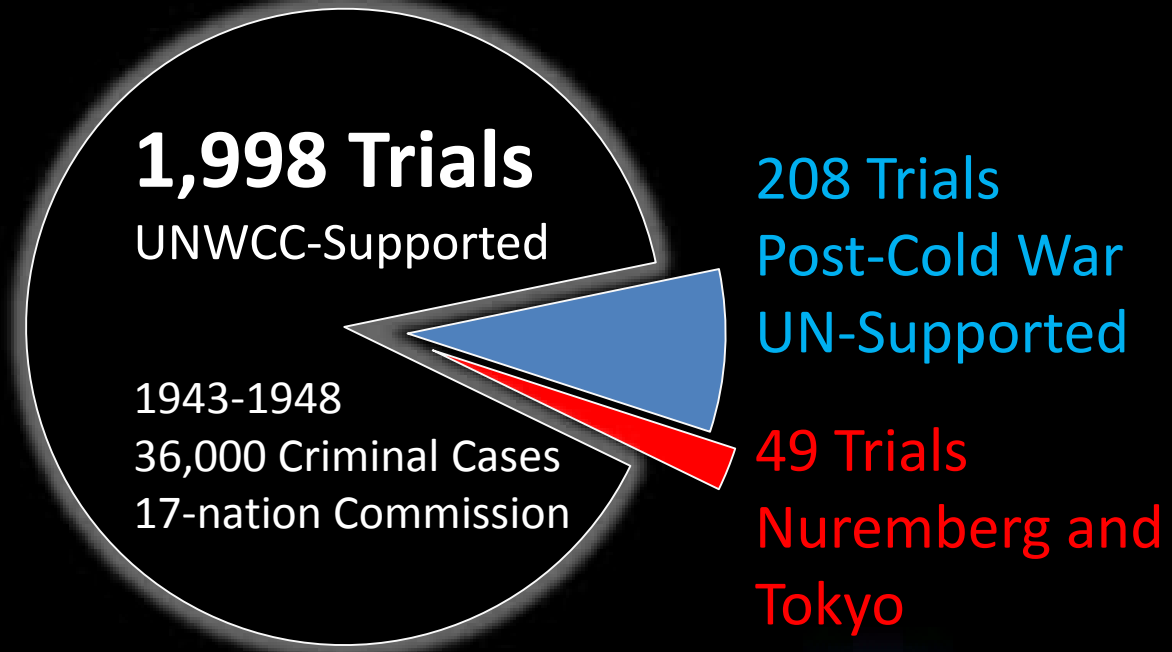
Four RRRRs

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



Mobile Devices + Twitter Use

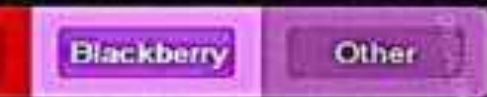
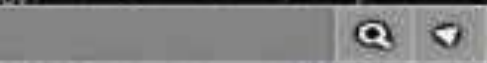
More than 280 million Tweets posted from mobile phones reveal geographic usage patterns in unprecedented detail

Android iPhone BlackBerry Other

Small Craft = iOS
Ferry = Android

Twitter Use

ets posted from mobile
usage patterns in



Apple Store



Sensors per Turbine

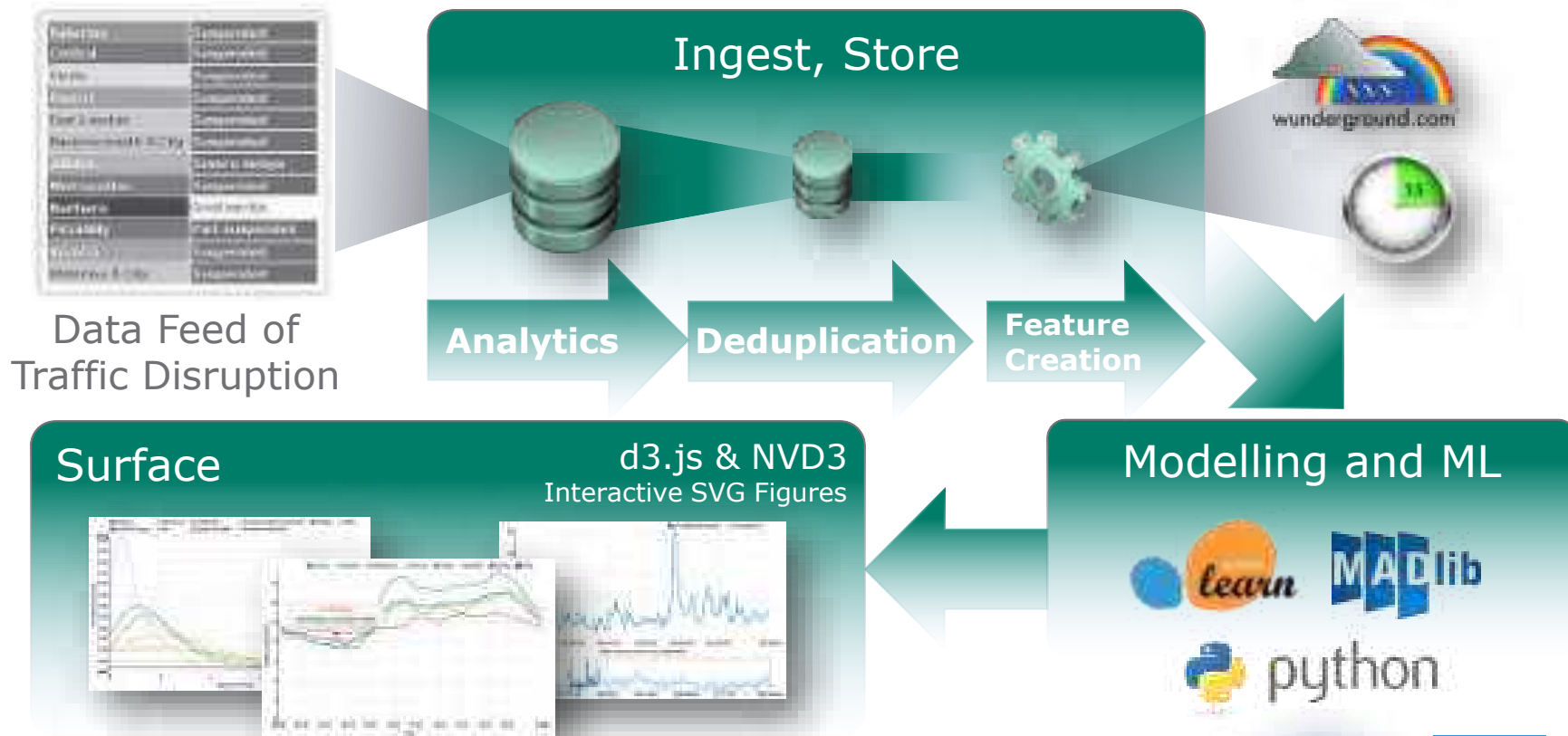
100 Physical, 300 Virtual

Data per Blade

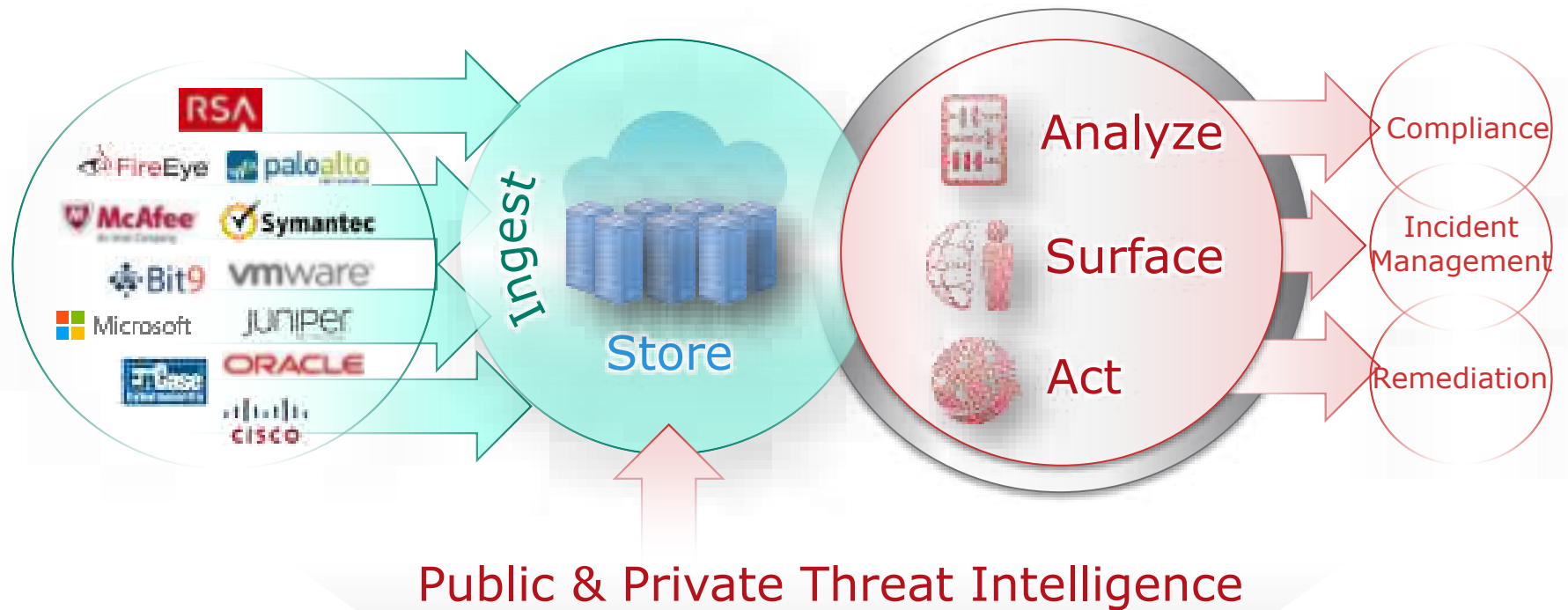
500 GB/day

(1mo = Library of Congress Print Collection)

London Transportation



Continuous Security Monitoring



“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>

© Copyright 2015 EMC Corporation. All rights reserved.



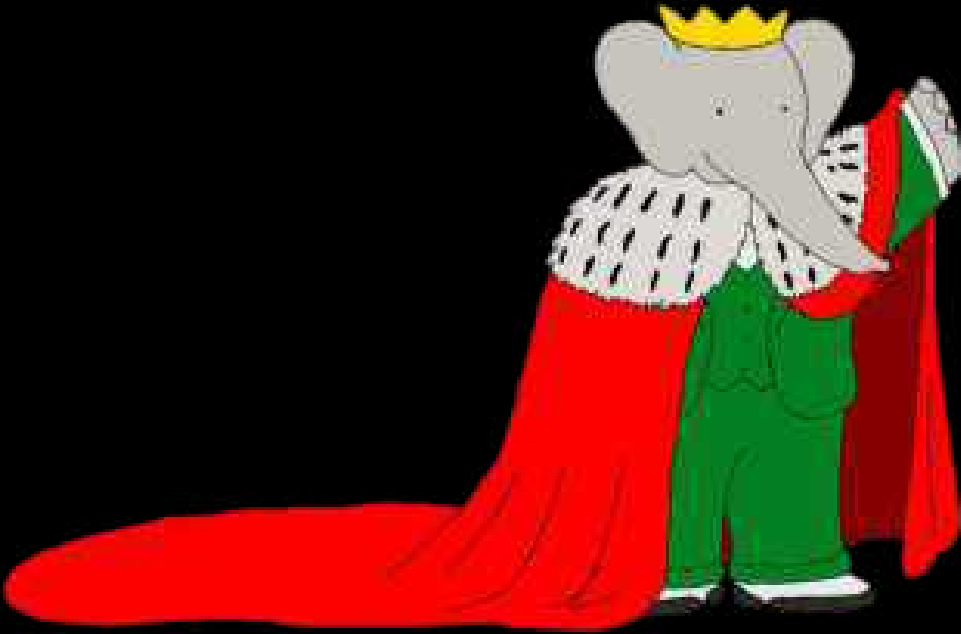
EMC

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

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

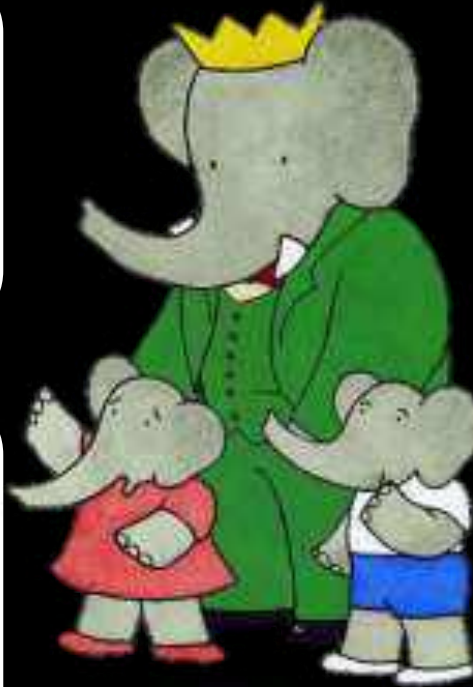
Control Opportunities

Exposure

- Mixed
- Insider
- Outsider

Non-Compliance

- Regulations
- Internal Policy
- External Standards

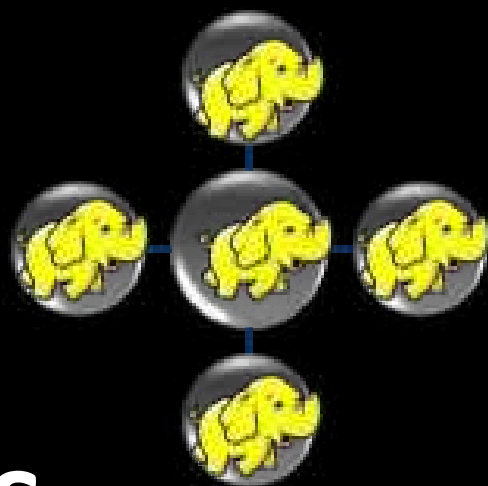
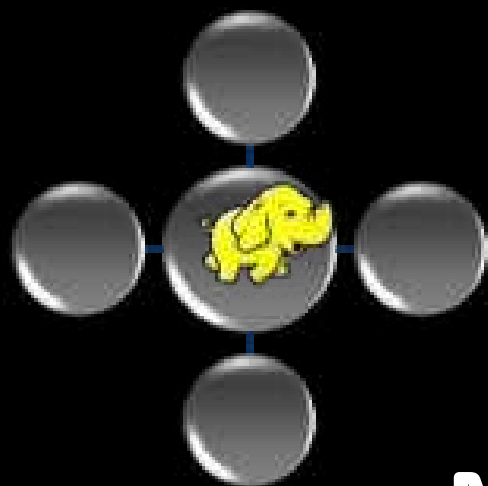


Data Loss/Corruption

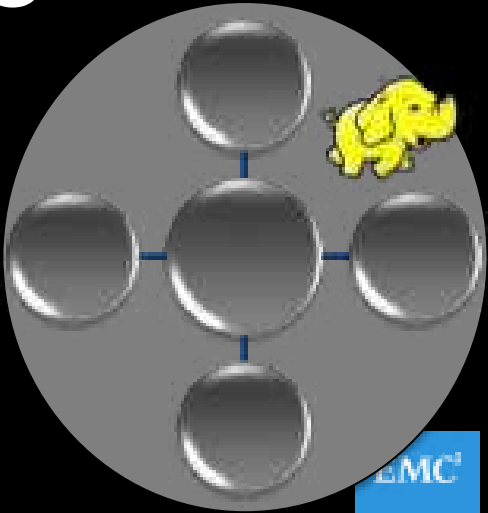
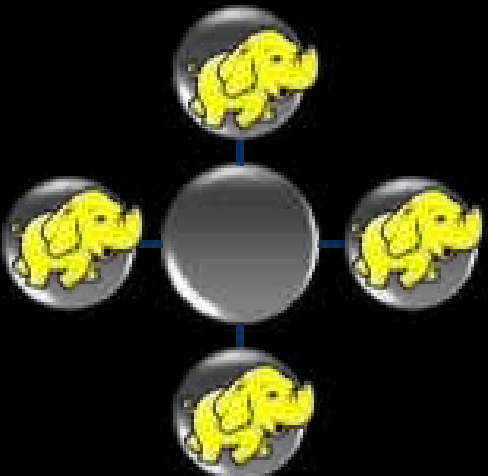
- Backup
- Restore
- Snapshot/Rollback

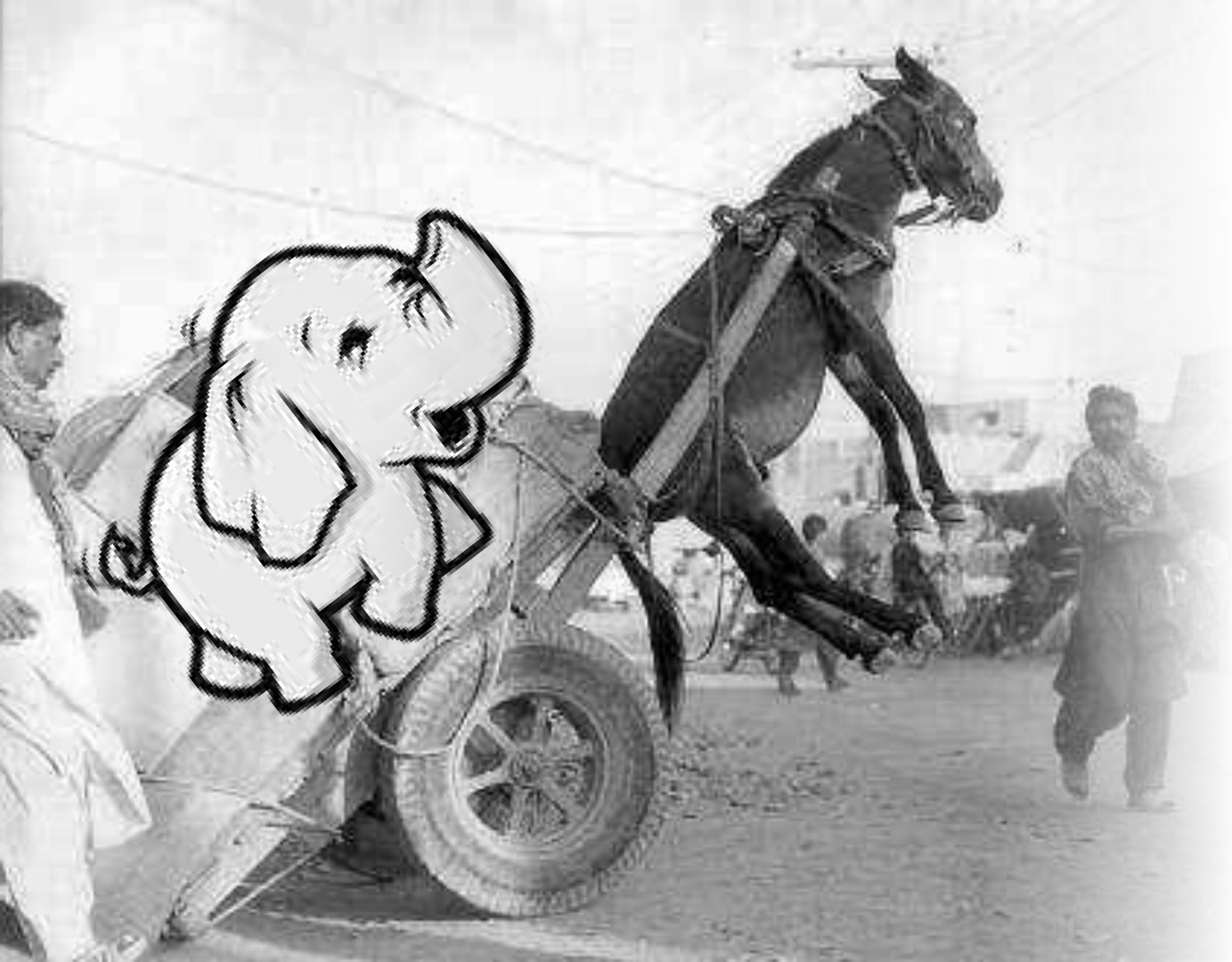
Outages

- Repairs
- Changes
- Disasters



MANY MODELS





NONE

LED BY

IT

HADOOP CRITICISMS

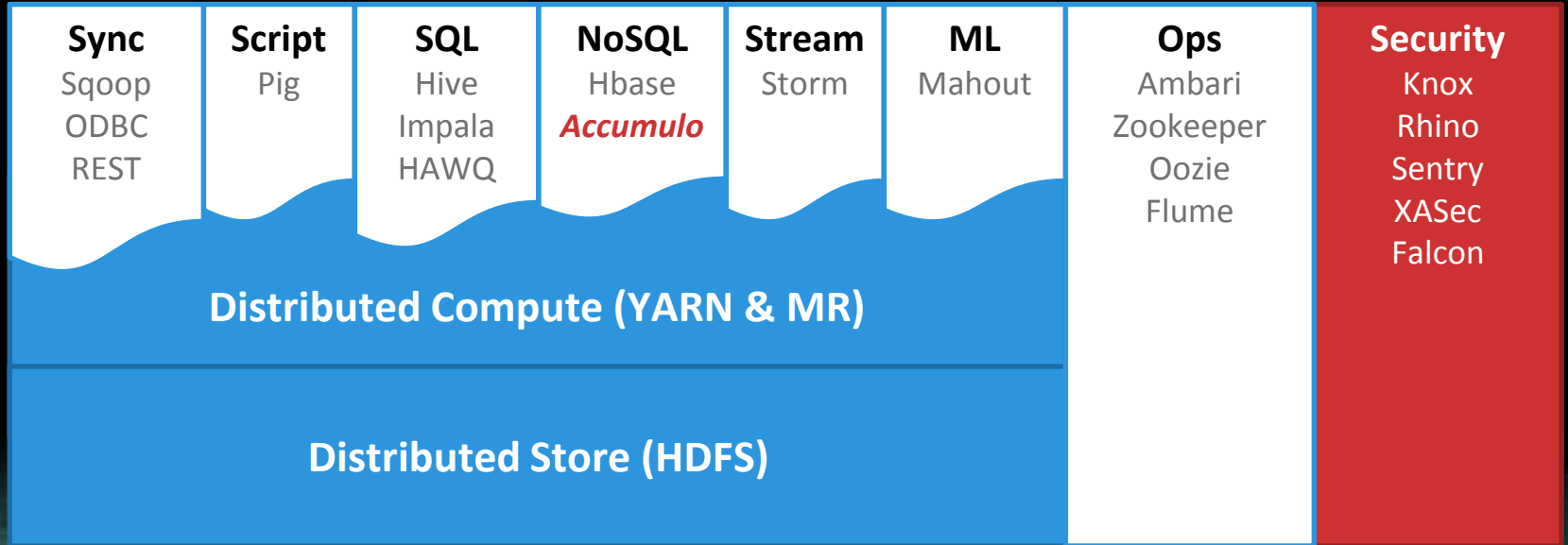
- Lack of Maturity (Security, Audit...)
 - File Immutability By Design Yet Mutability in Reality
 - Resiliency By Design Yet Outages/Latency in Reality
 - Security Mode By Design Yet Insecure in Reality
- Fractured and Complex
- Non-Technical Usability



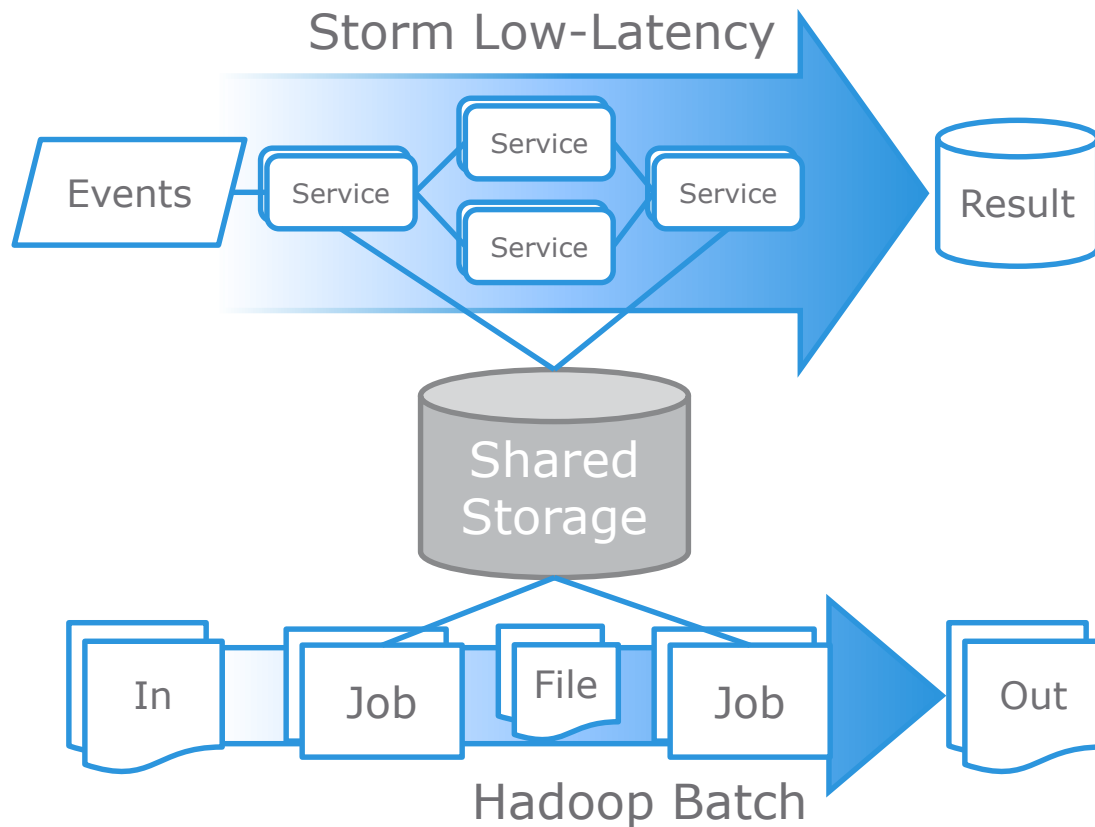
Threat
Assumptions

No MitM
Users not root
No NIC control

HADOOP ECOSYSTEMS

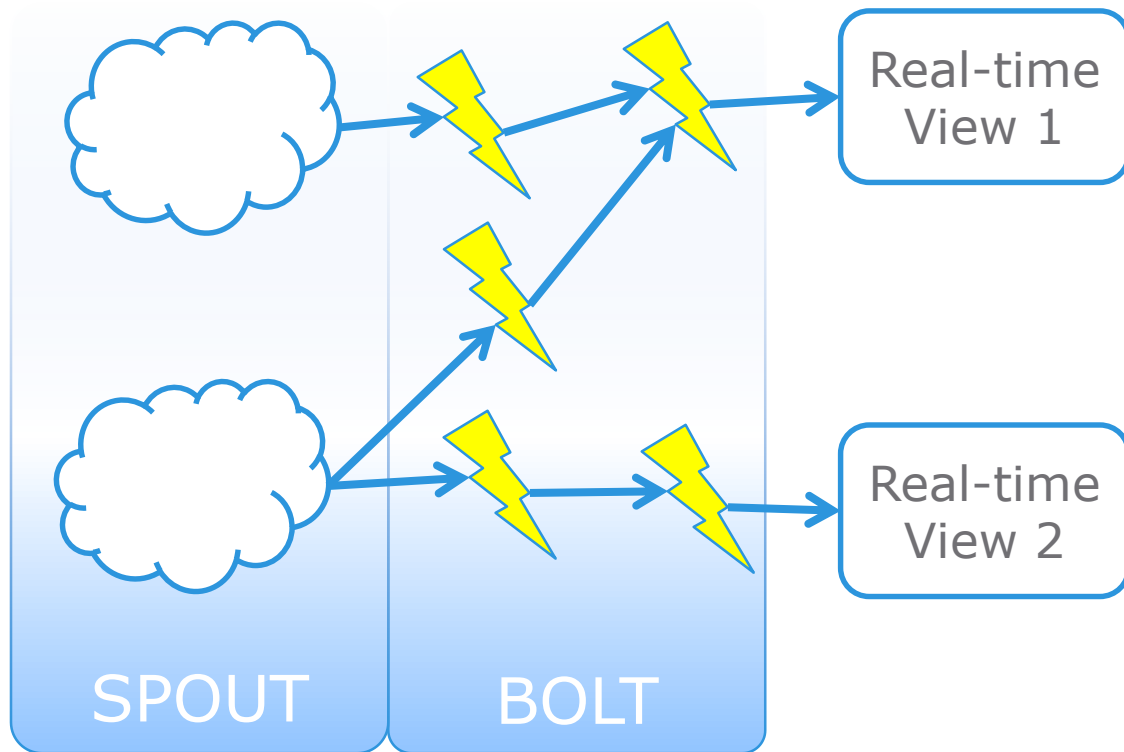


Example: Storm



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

Example: Storm



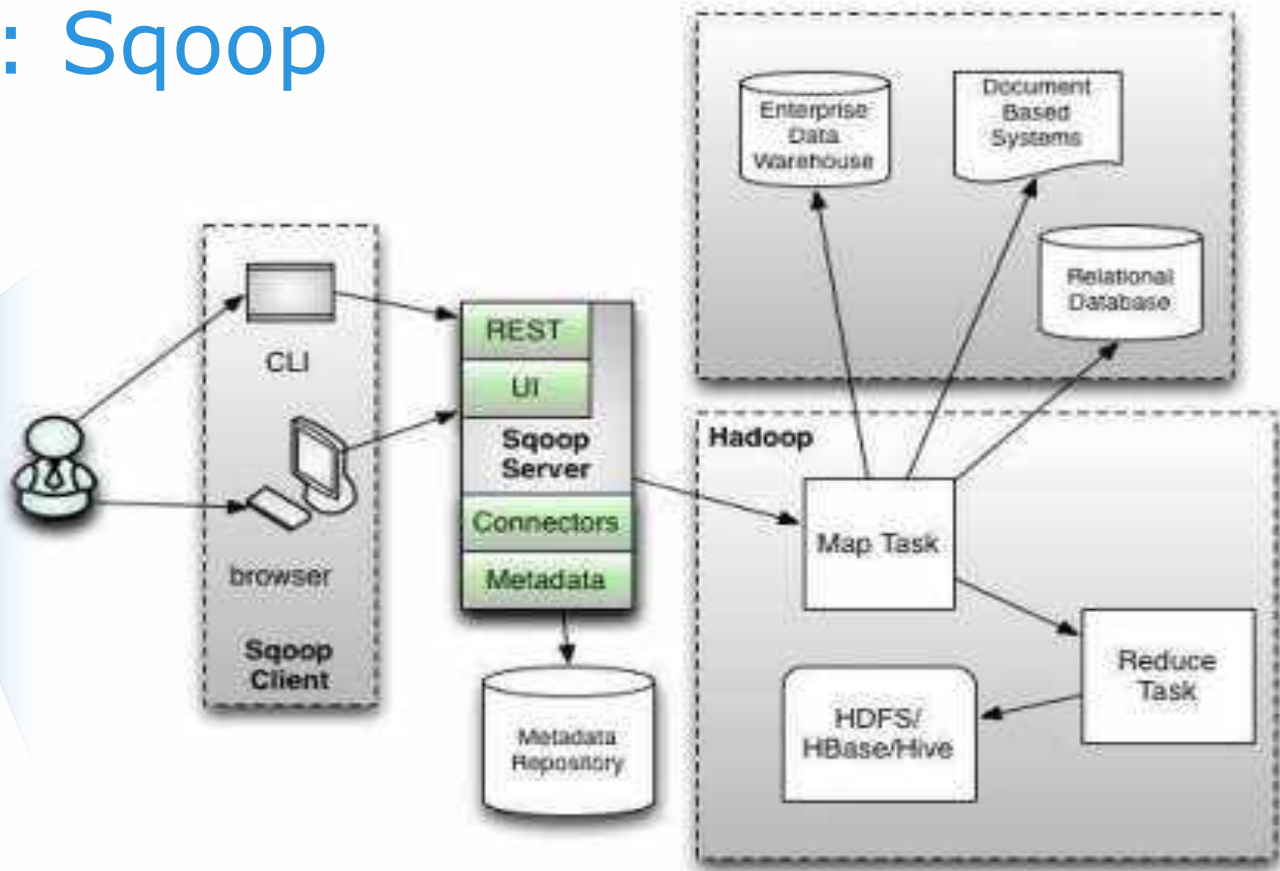
“Storm will hang when you do weird things with it”

“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
<code>security.containermanagement.protocol.acl</code>	ApplicationMasters communicate with NodeManager	"*"
<code>security.applicationmaster.protocol.acl</code>	ApplicationMasters communicate with ResourceManager	"*"
<code>security.job.task.protocol.acl</code> (MR1 uses <code>security.task.umbilical.protocol.acl</code>)	MR2 tasks report task progress	"*"

[hadoop-policy.xml](#)

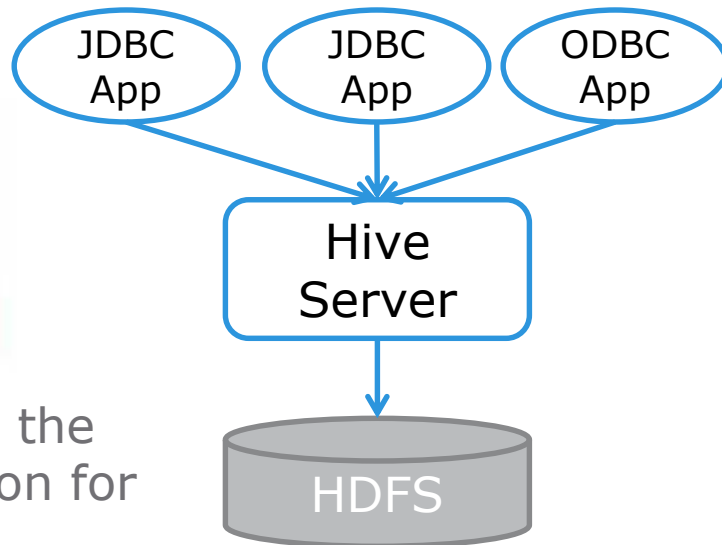
Other Controls Still Needed!



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

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

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

INGEST



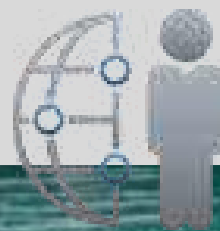
STORE



ANALYZE



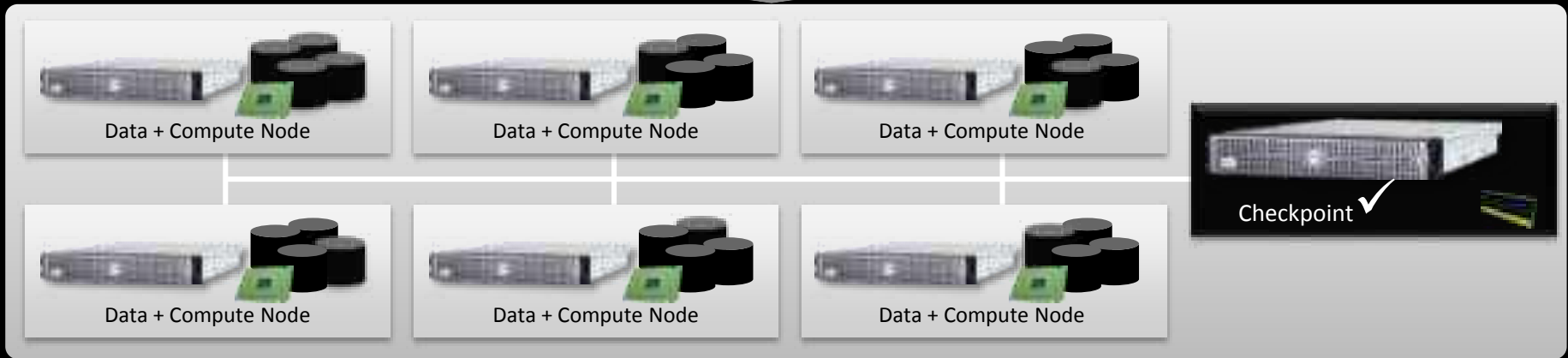
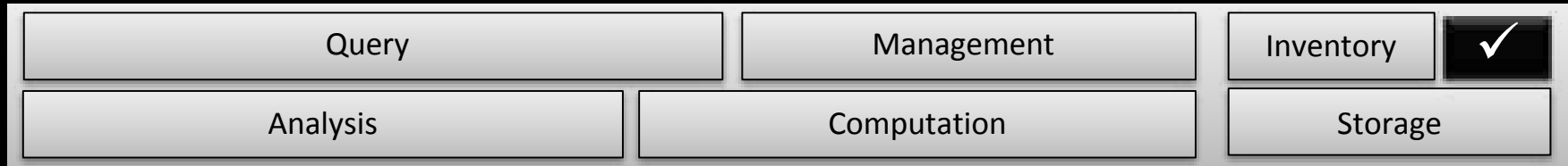
SURFACE



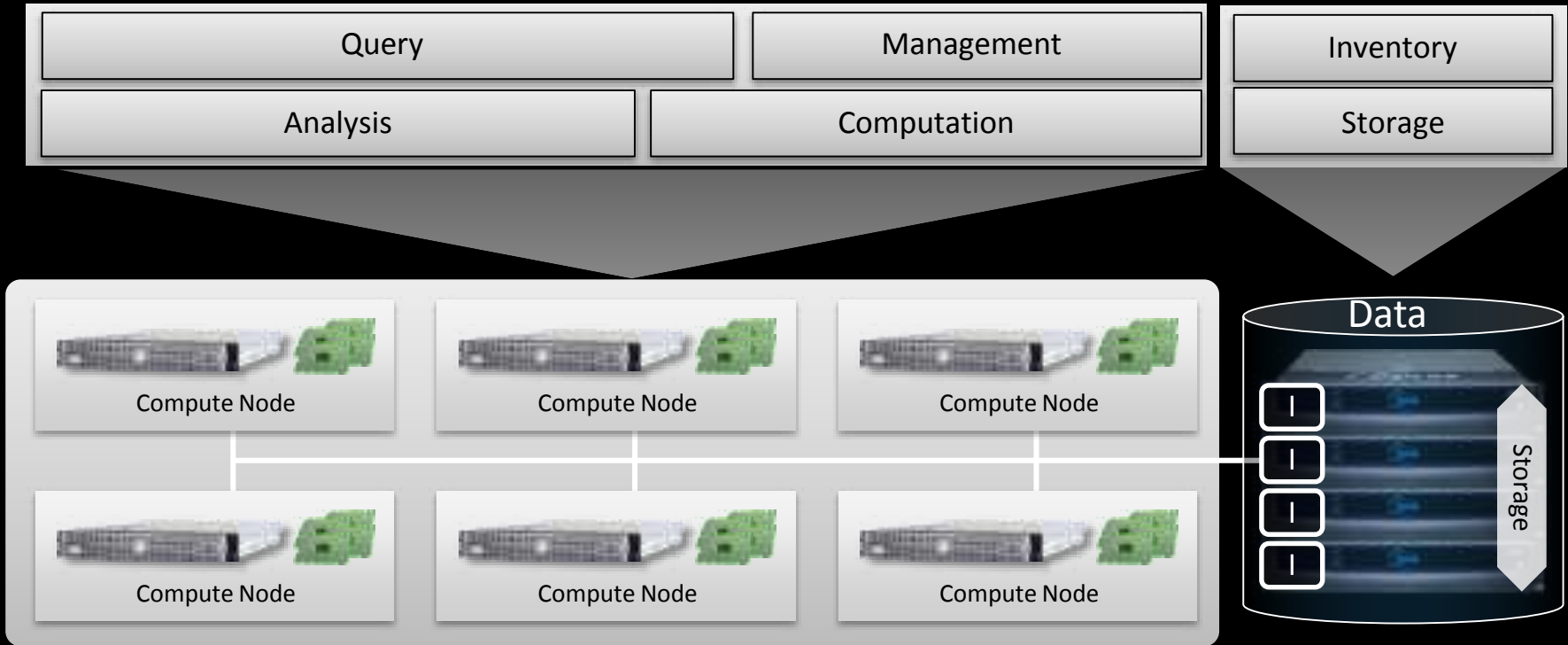
ACT



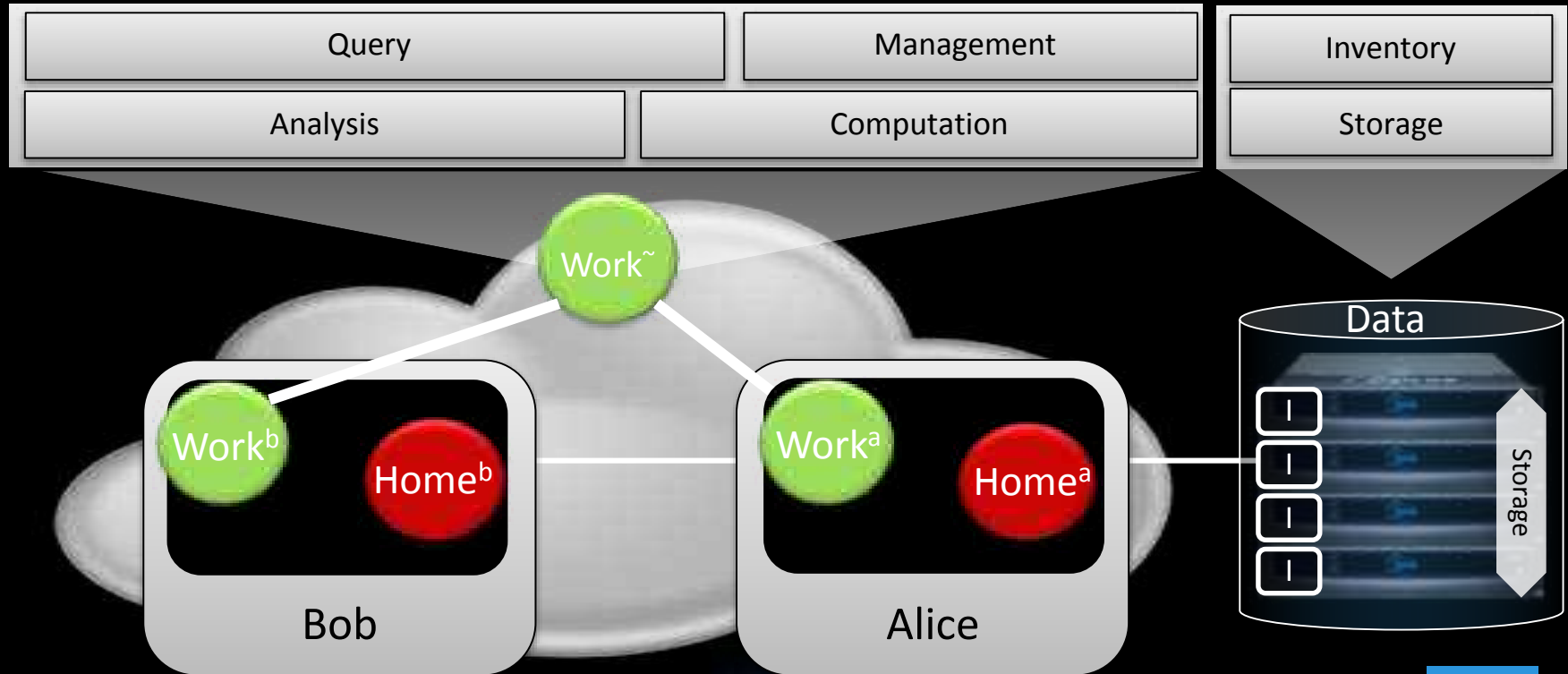
Phase One - Scoping



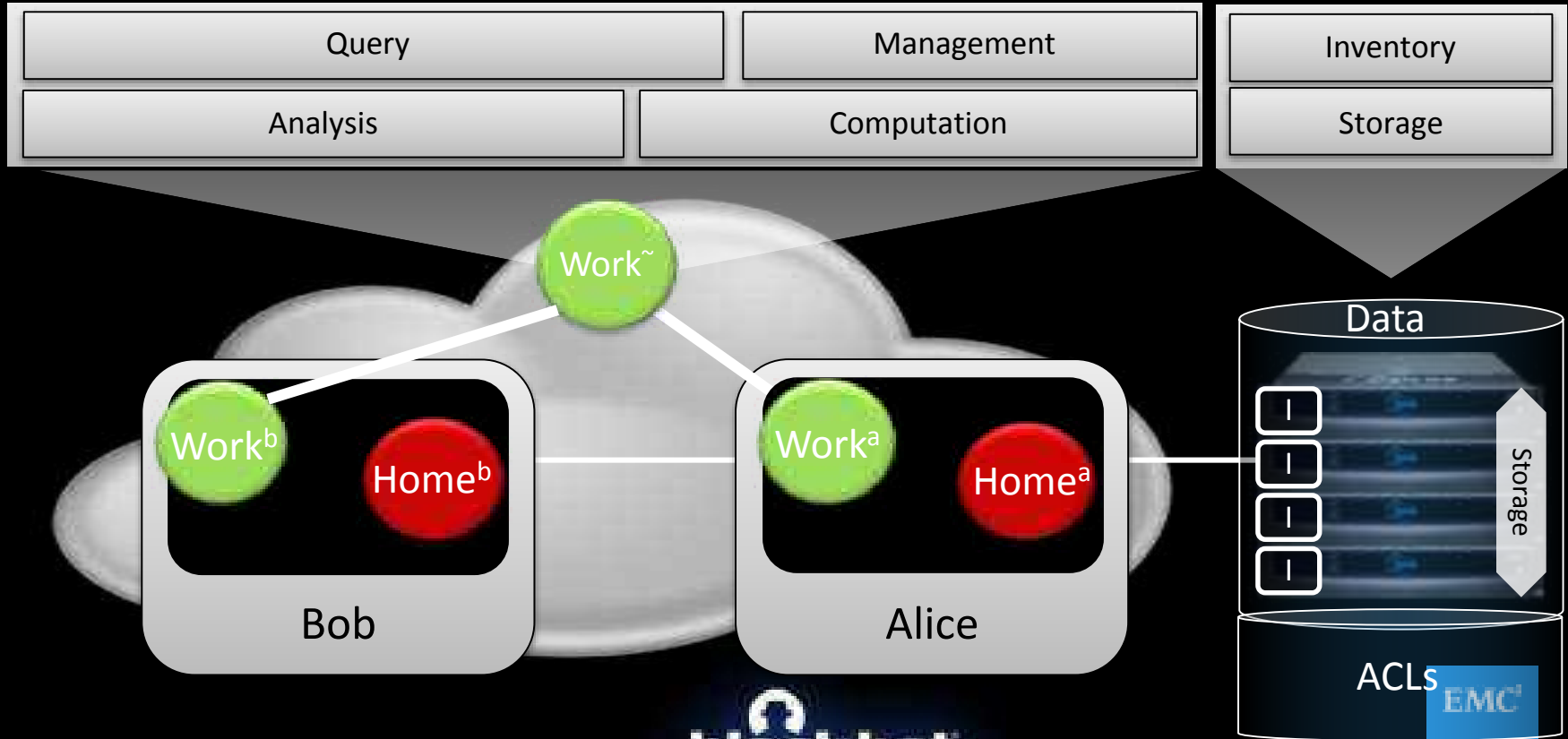
Phase Two - Availability



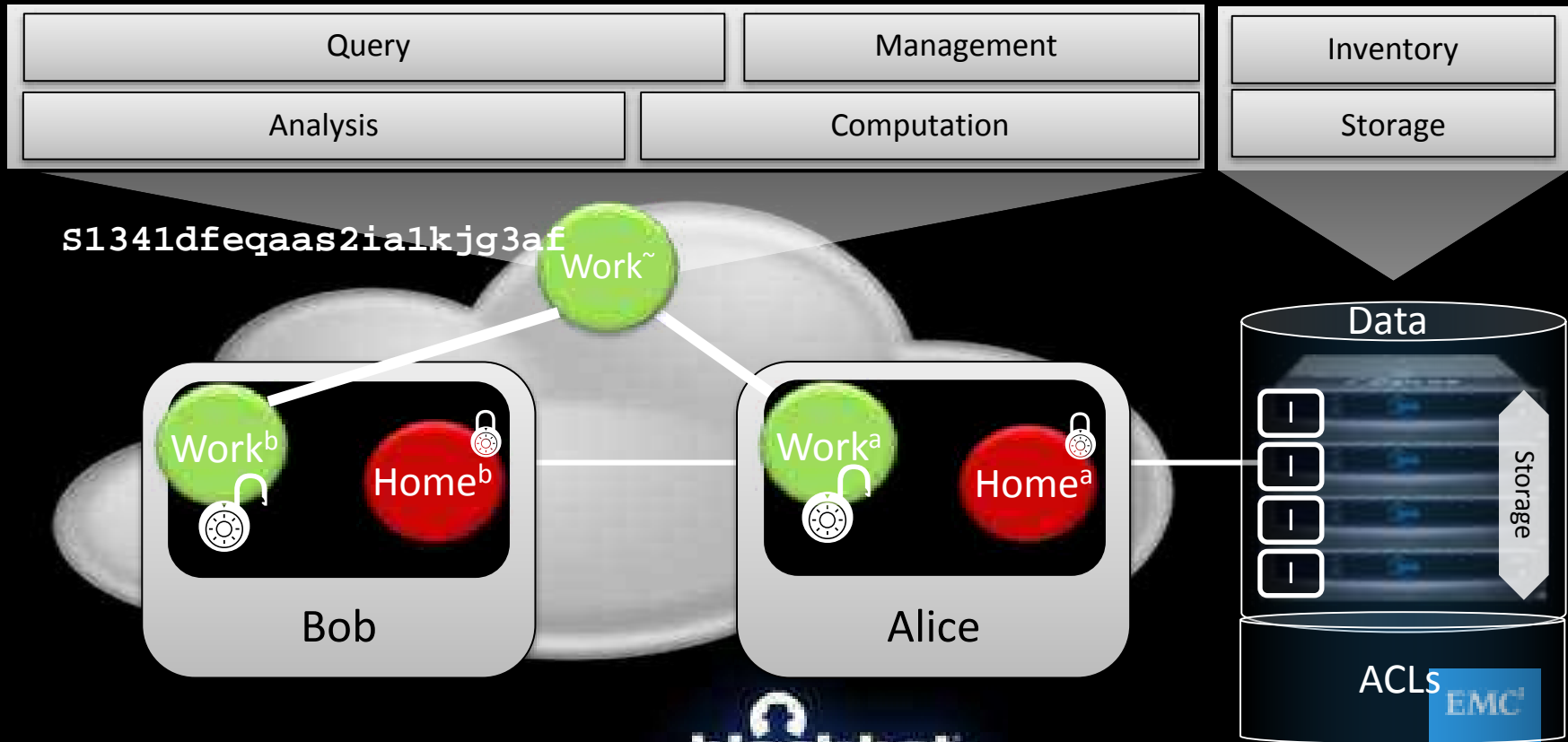
Phase Three – Classification



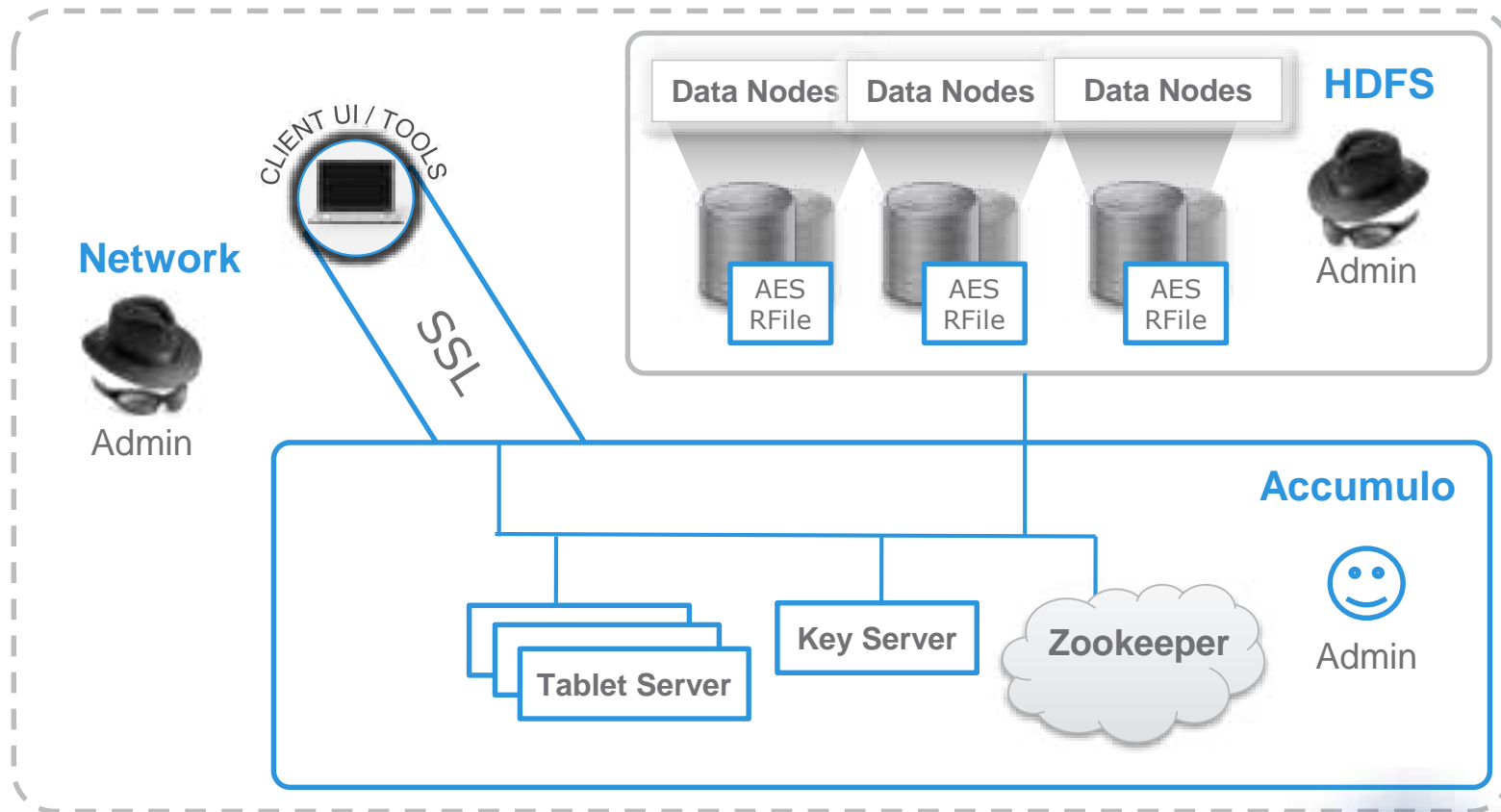
Phase Four – Authority



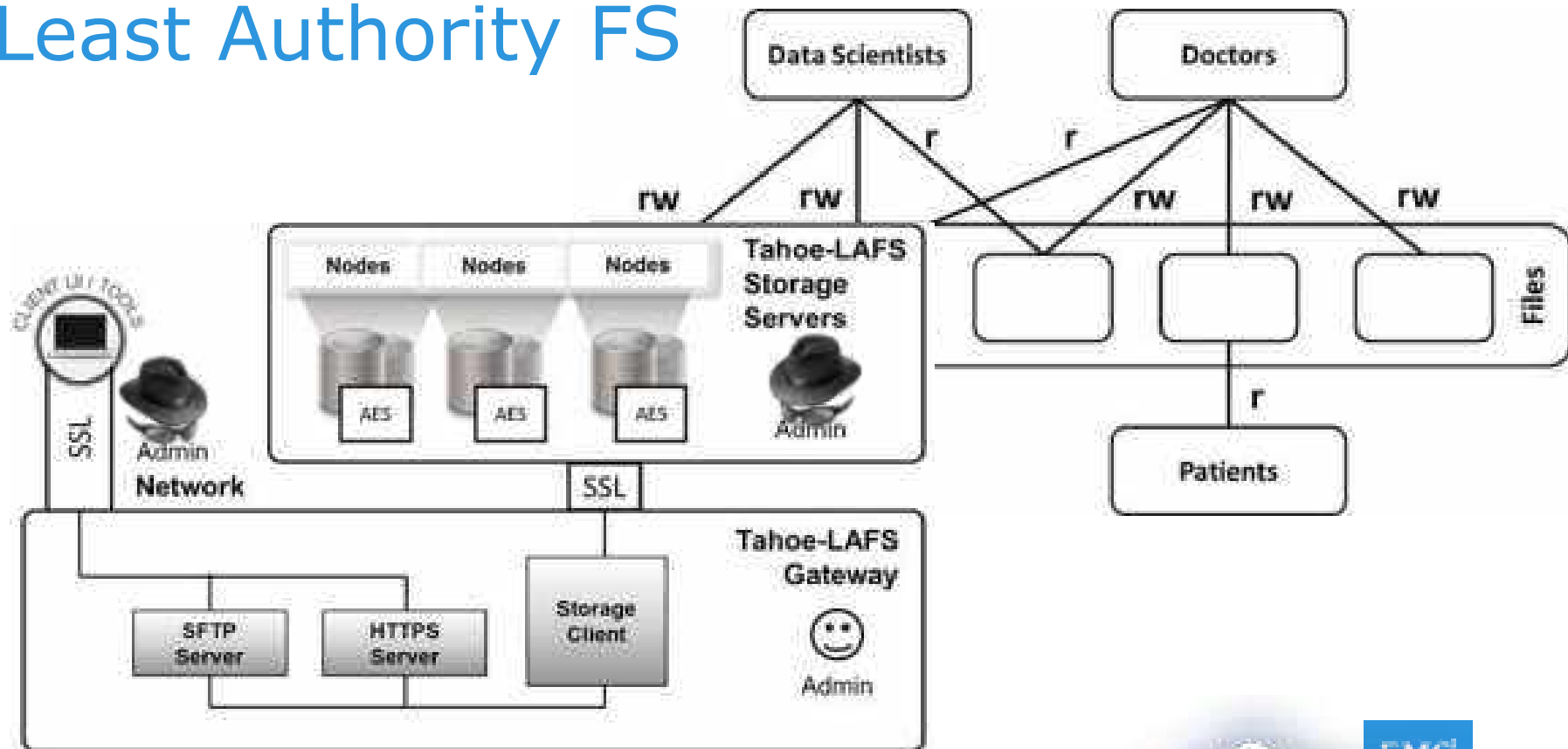
Phase Five – Least Authority



ACCUMULO

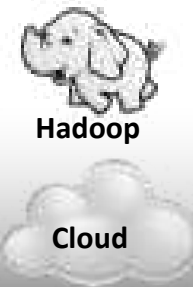


TAHOE Least Authority FS



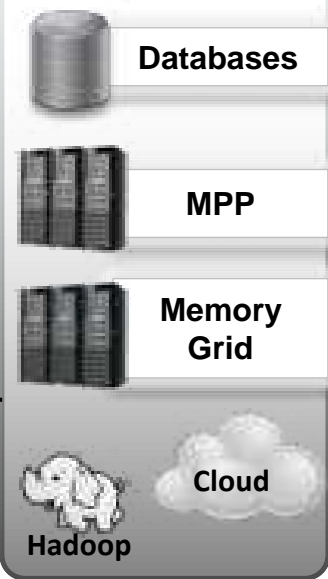
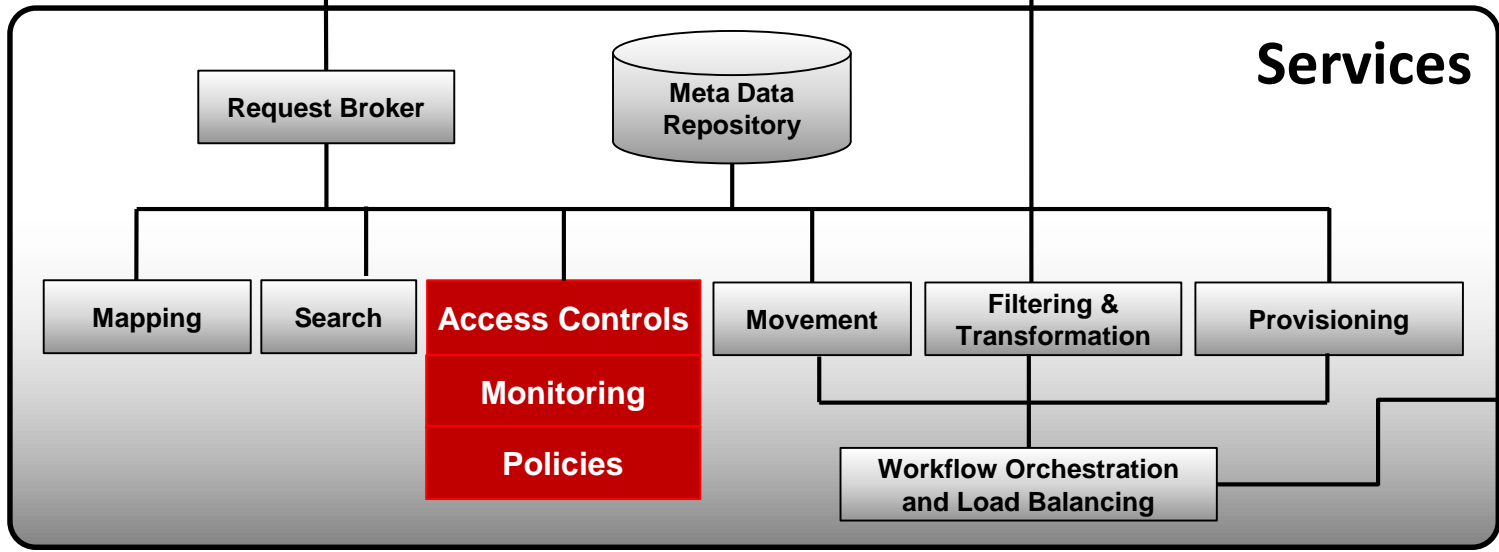
Example: HealthCare Big Data

Provisioning, Analytics & Collaboration

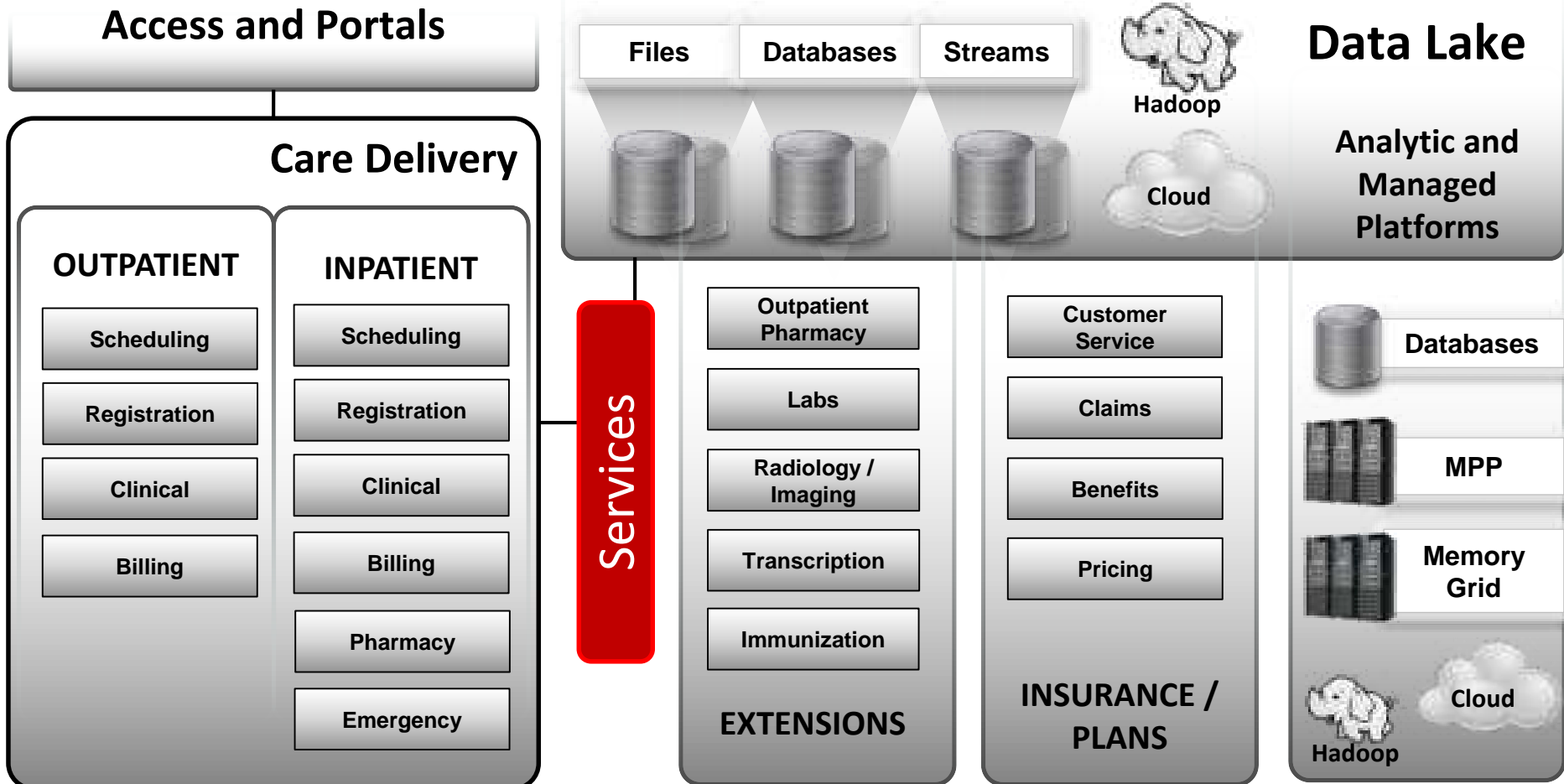


Data Lake

Analytic and Managed Platforms



Example: HealthCare Big Data



Trusted IT

Transparency

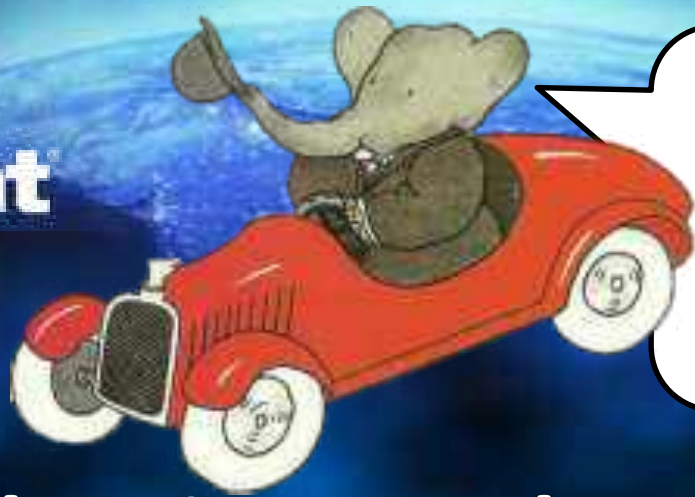


Relevance



Resilience





**Merci
Beaucoup!**

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

Davi Ottenheimer (@daviottenheimer)

Senior Director of Trust, EMC