Auditing Big Data for Privacy, Security and Compliance

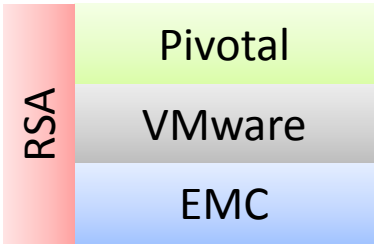
Davi Ottenheimer @daviottenheimer

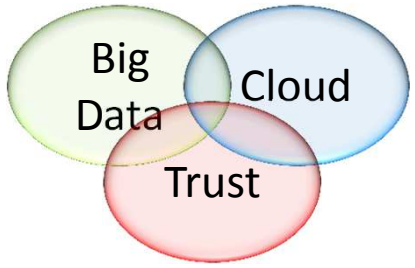
Senior Director of Trust, EMC

In-Depth Seminars – D21

Introduction

- Davi Ottenheimer (@daviottenheimer)
  - 19th Year InfoSec
  - CISM, Platinum ISACA (1997)
  - Ex-Big 5 Auditor, Ex-PCI QSA/PA-QSA
  - Co-Author “Securing the Virtual Environment”
- @EMCTrustedIT





Agenda

- Big Data
- Operations
- Auditing

BIG DATA


ISACA®
Trust in, and value from, information systems
San Francisco Chapter
CRISC
CGEIT
CISM
CISA
2013 Fall Conference – "Sail to Success"

The Big Data Dilemma

"We have massive amounts of data. We know who you are. We know what your history has been on the airline. We can customize our offerings."

- Delta CEO

"Airlines have yet to find the right balance between being helpful and being creepy."

- Associated Press

http://m.apnews.com/ap/db_289563/contentdetail.htm?contentguid=InpMBxrL

Big Data Definitions

- Many Long-Standing Examples
  - Astronomy (“Billions and billions”)
  - Meteorology (Storms)
  - Geology (Quakes)
  - Anatomy (Disease)
  - Economics (Fraud)
  - Espionage (Echelon SIGINT, PRISM…)
- Three V’s (Volume, Variety, Velocity)

Structured + Unstructured Data = Big

Structured in
Relational Databases

Managed, Unmanaged
& Unstructured

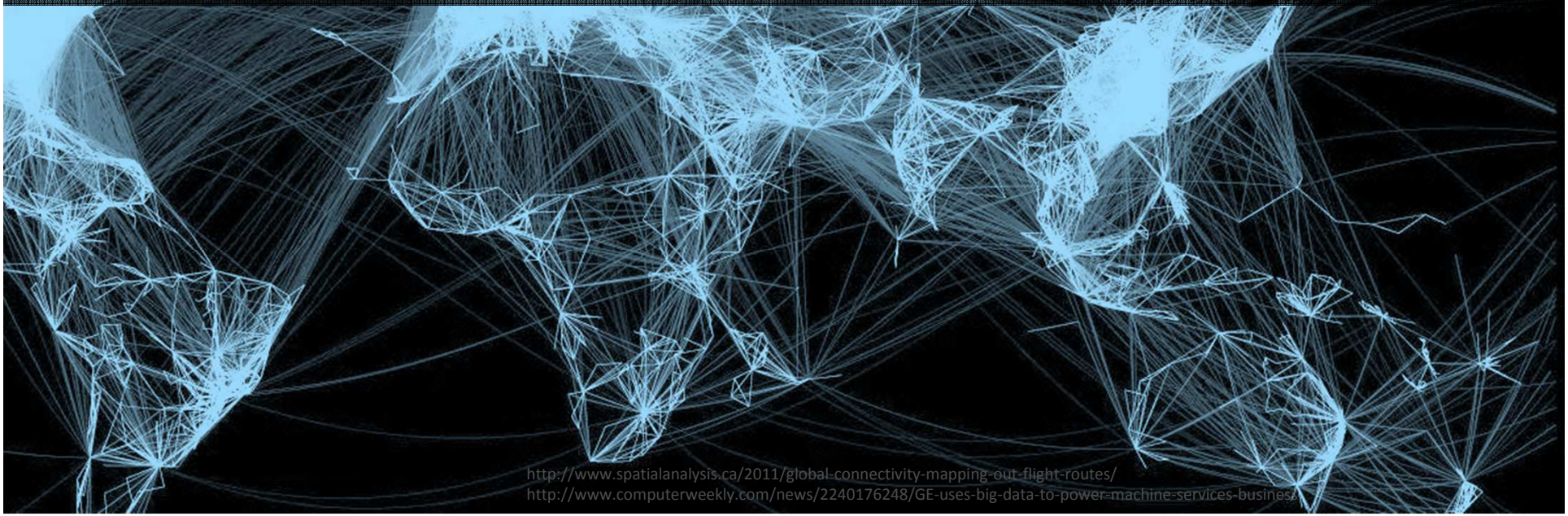
Telemetry, Location-Based, etc.

Internet of Things

Non-Enterprise

Global Flight Analysis

- 60,000 Total Routes
- 1 Tb/day Data Each Gas Turbine Engine
- 400K gal/yr Saved by AA Paperless Pilot (-35lb)

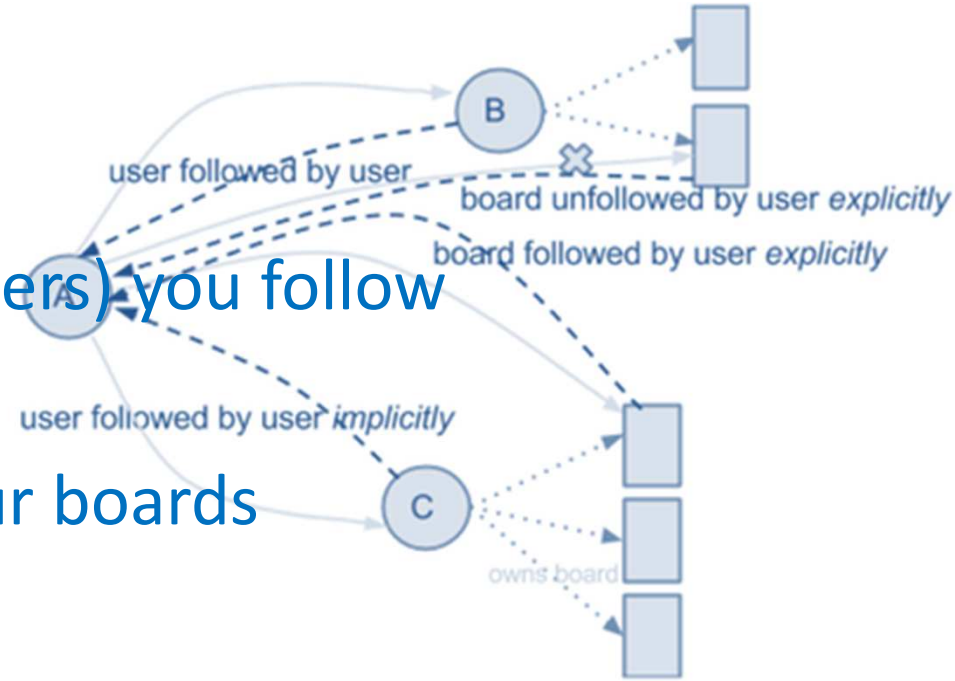


http://www.spatialanalysis.ca/2011/global-connectivity-mapping-out-flight-routes/
http://www.computerweekly.com/news/2240176248/GE-uses-big-data-to-power-machine-services-busines

What is So *Pinterest*-ing?

Stored and Ready on Login for 70m Users

- Lists
  - Users you follow
  - Boards (and related users) you follow
  - Your followers
  - People who follow your boards
  - Boards you follow
  - Boards you unfollowed after following a user
- Followers and unfollowers of each board

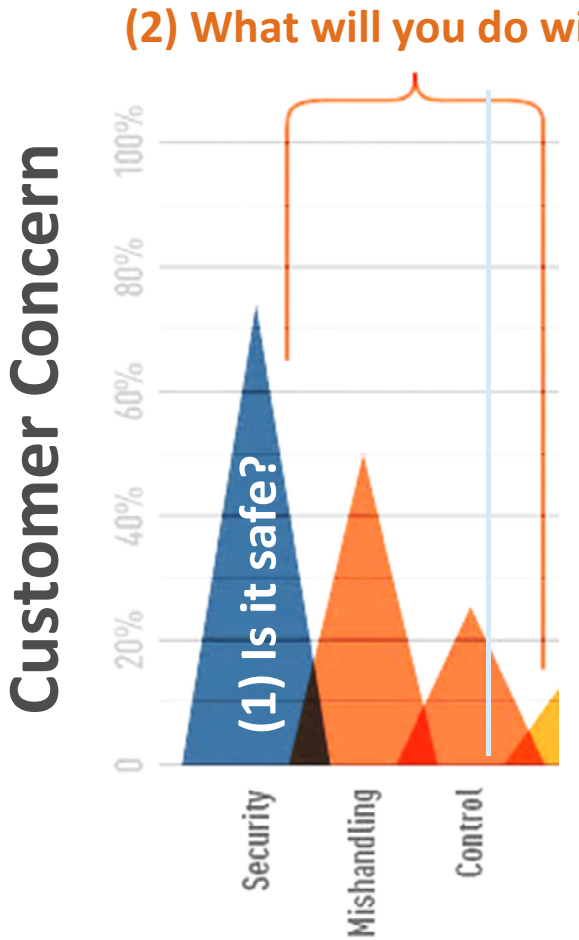


http://blog.gopivotal.com/case-studies-2/using-redis-at-pinterest-for-billions-of-relationships, http://engineering.pinterest.com/post/55272557617/building-a-follower-model-from-scratch

What Makes Data “Ready”?

Where’s the Line for Service and Surveillance?

Should Users Trust?



https://www.unboundid.com/blog/2013/09/05/the-value-of-identity-data-and-identity-etiquette/

Should Users Trust?

"…a 26-year-old mother of two teenagers…is just about biologically impossible"

"Audience Operating System"

acxiom

		RIGHT OR WRONG?
Age Range	Age 26 - 27	✓
Gender	Female	✓
Ethnicity Based on Surname	American	✓
Education	Completed College	✓
College Graduate	True	✓
Marital Status	Married	✗
Presence of Children	Children Present	✗
Number of Children	2 Children	✗
Children's Age	14 Years old,17 Years old	✗
Children's Gender by Age	Unknown Gender 11 - 15,Unknown Gender 16 - 17	✗

FROM: ABOUTTHEDATA.COM

http://money.cnn.com/2013/09/05/pf/acxiom-consumer-data/index.html

Finding Value in Data

Chicago and suburbs = 1.5b gal/day wastewater

- Disease
- Drugs
- Environmental Risk

"…we know the estimated numbers of people being served by each waste water treatment plant, we can back-calculate daily [drug] loads…"

- Dr Kasprzyk-Hordern

http://gizmodo.com/5844925/chicagos-stickney-wastewater-treatment-plant-is-the-crappiest-place-on-earth, http://planetearth.nerc.ac.uk/news/story.aspx?id=1185&cookieConsent=A
http://www.treehugger.com/natural-sciences/fish-near-water-treatment-plants-are-harmed-by-human-drugs.html

Finding Errors in Data

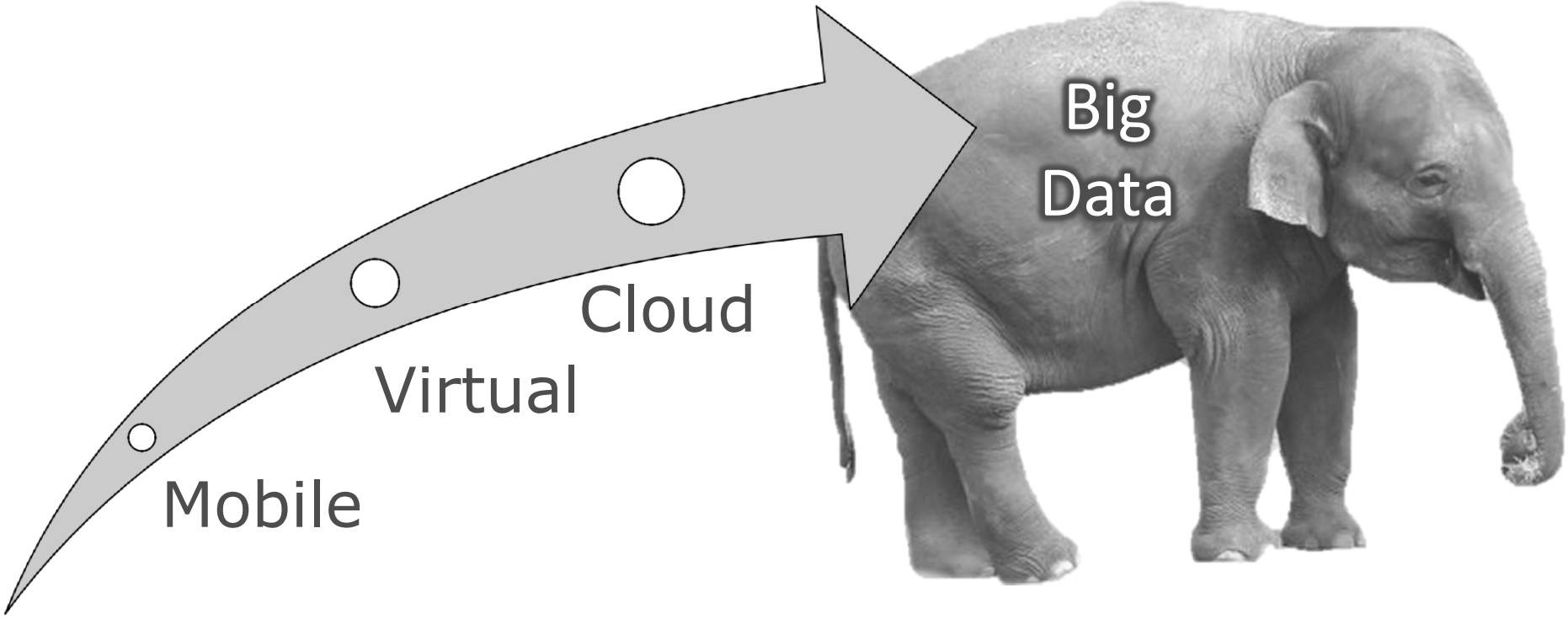
of Internet users have taken steps online to remove or mask their digital footprints

http://pewinternet.org/Reports/2013/Anonymity-online.aspx, http://www.connecture.com/the-connecture-difference/

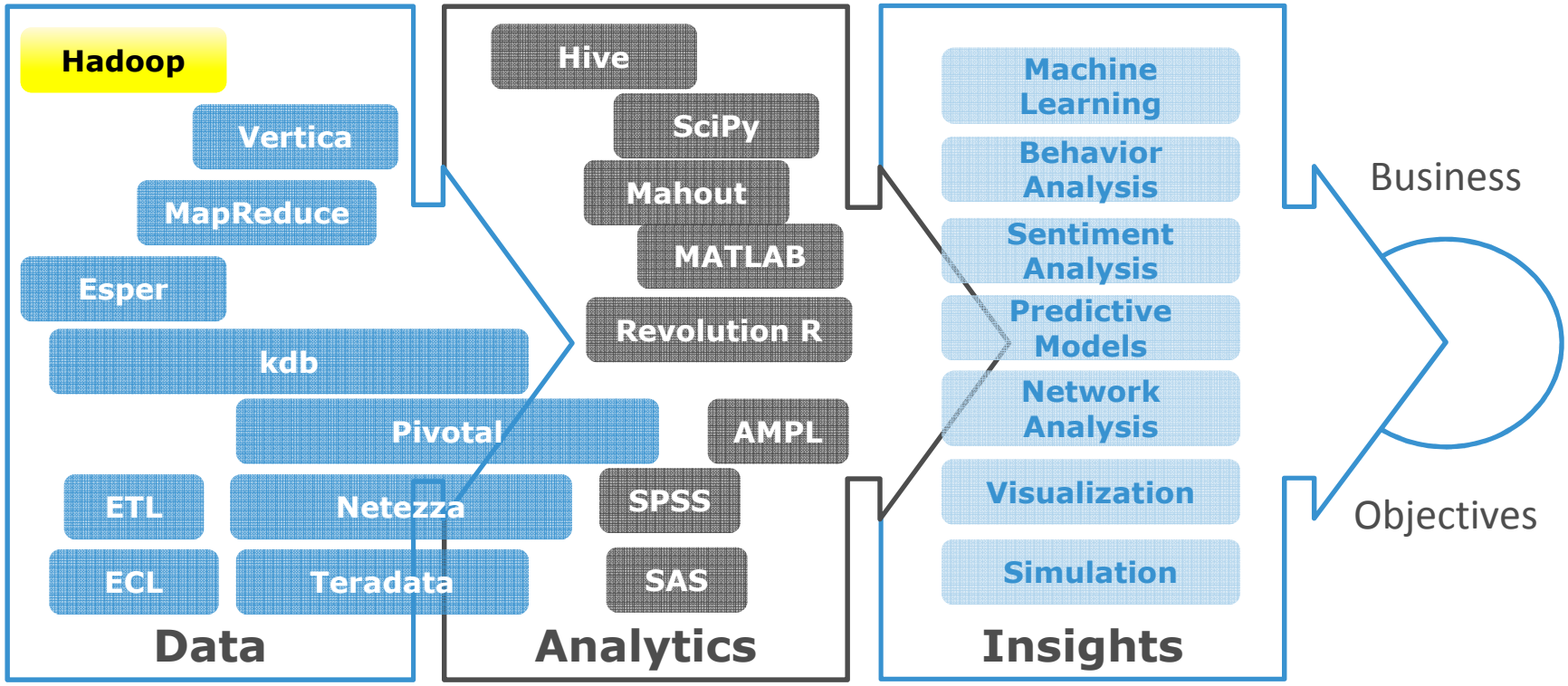
OPERATIONS


ISACA®
Trust in, and value from, information systems
San Francisco Chapter
CRISC
CGEIT
CISM
CISA
2013 Fall Conference – "Sail to Success"

Transformation of IT

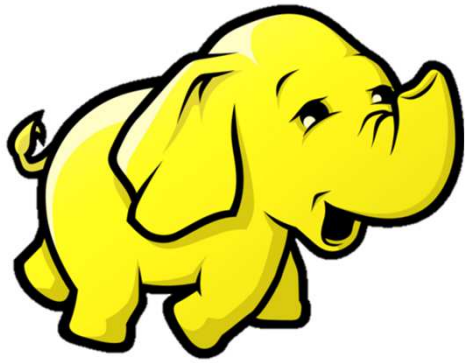


New Wave of Big Data Technology



(Over?) Emphasis on Performance

- Nodes Distributed
- Data Shared
- Access Controls Open
- Networks Open
- Clients Unauthenticated
- Web Services Open



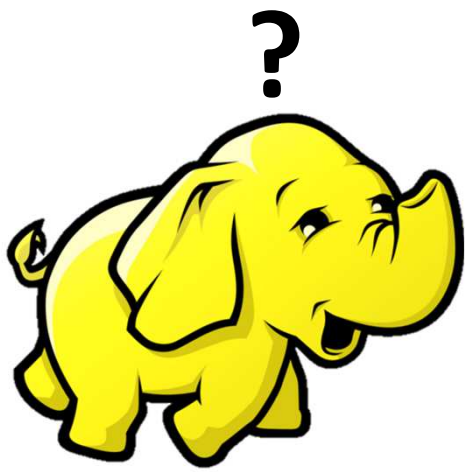
Hadoop Machine Roles

1. Clients
2. Masters
3. Slaves

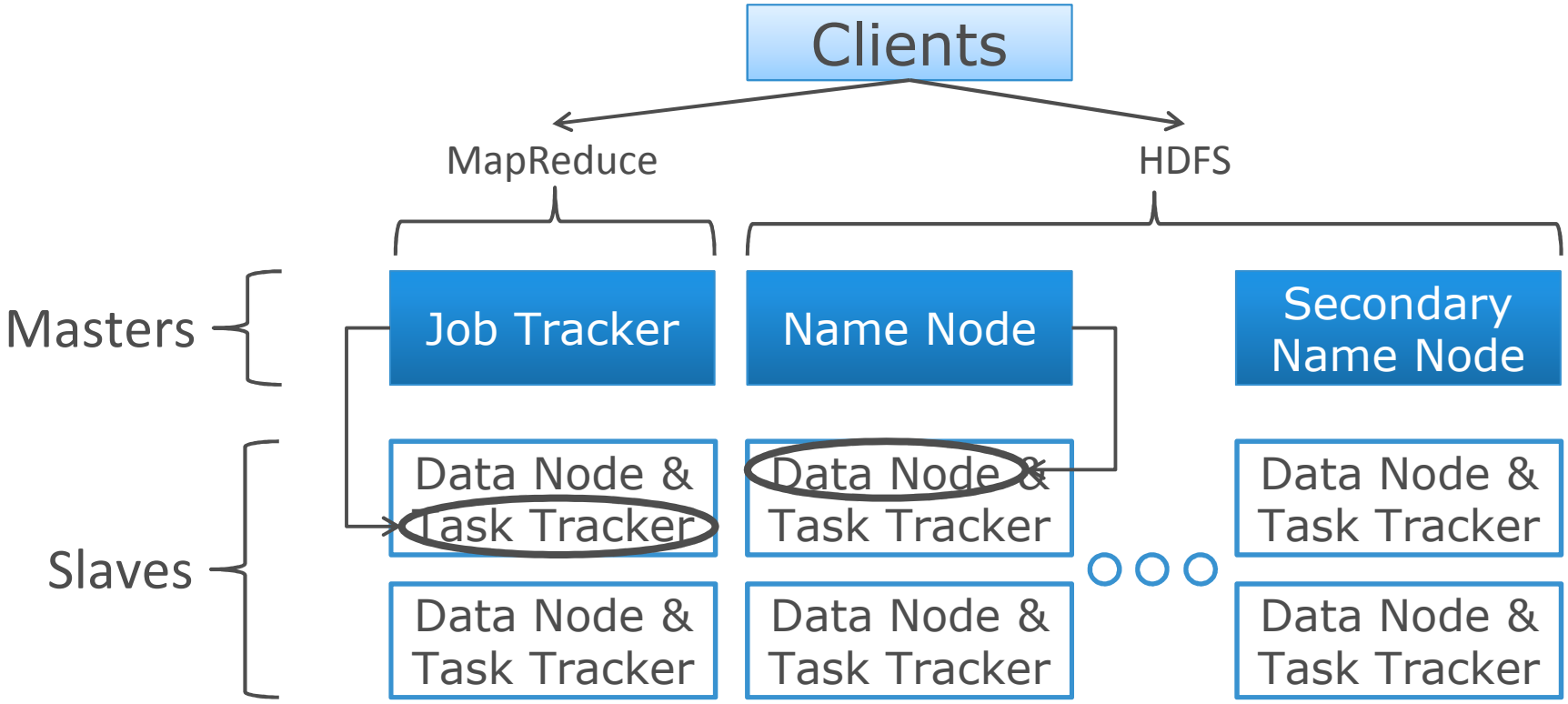
1. Load Data to Cluster
2. Submit MapReduce Jobs
3. View Results

1. Storage Management (Name Node -> HDFS)
2. Compute Management (Job Tracker -> MapReduce)

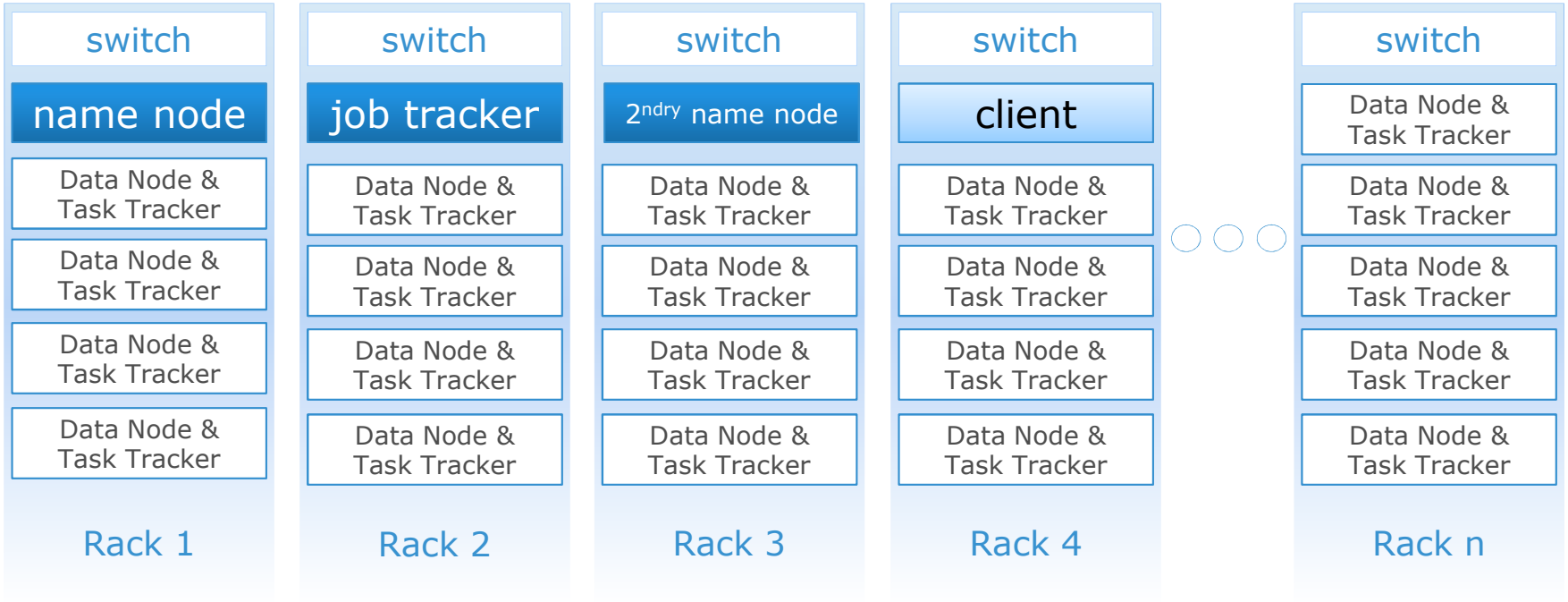
1. Storage (Data Node)
2. Compute (Task Tracker)



Hadoop Machine Roles



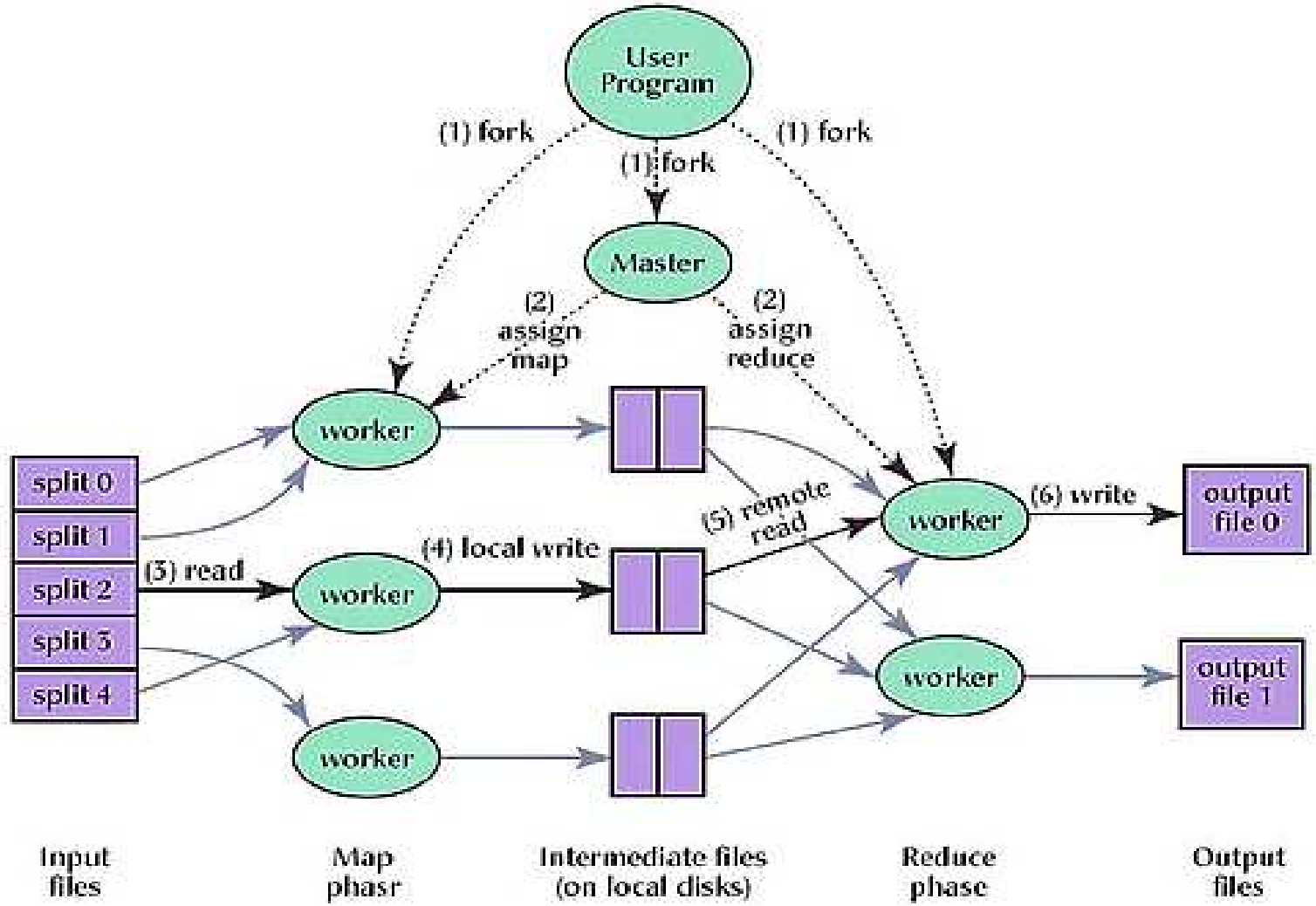
Hadoop Cluster



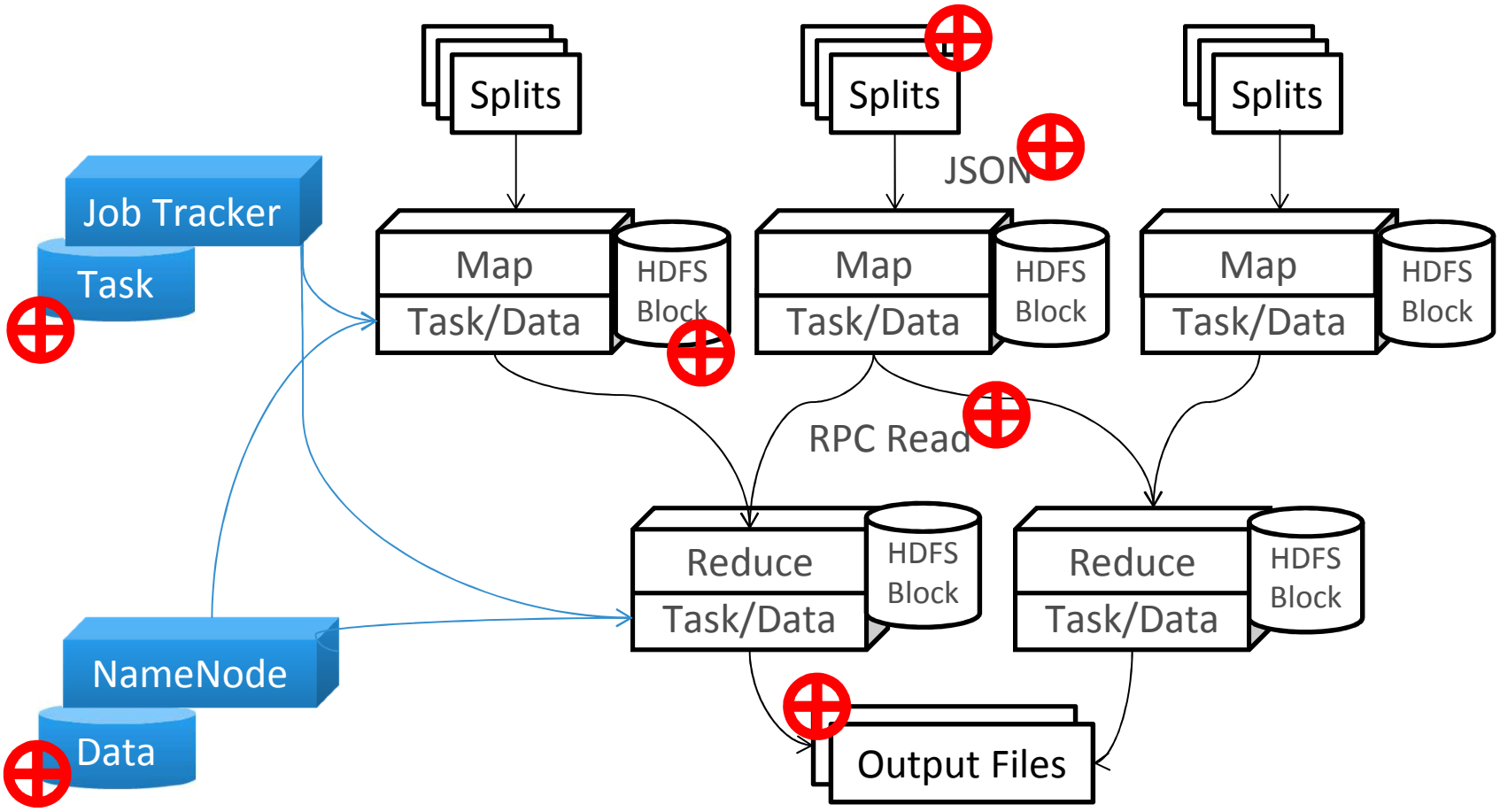
Typical Cluster Environment

- Petabytes
- 10,000s Slots Per Cluster
- Shared or Dedicated
- Production *and* Non-Production
- Mixed Software and Uses

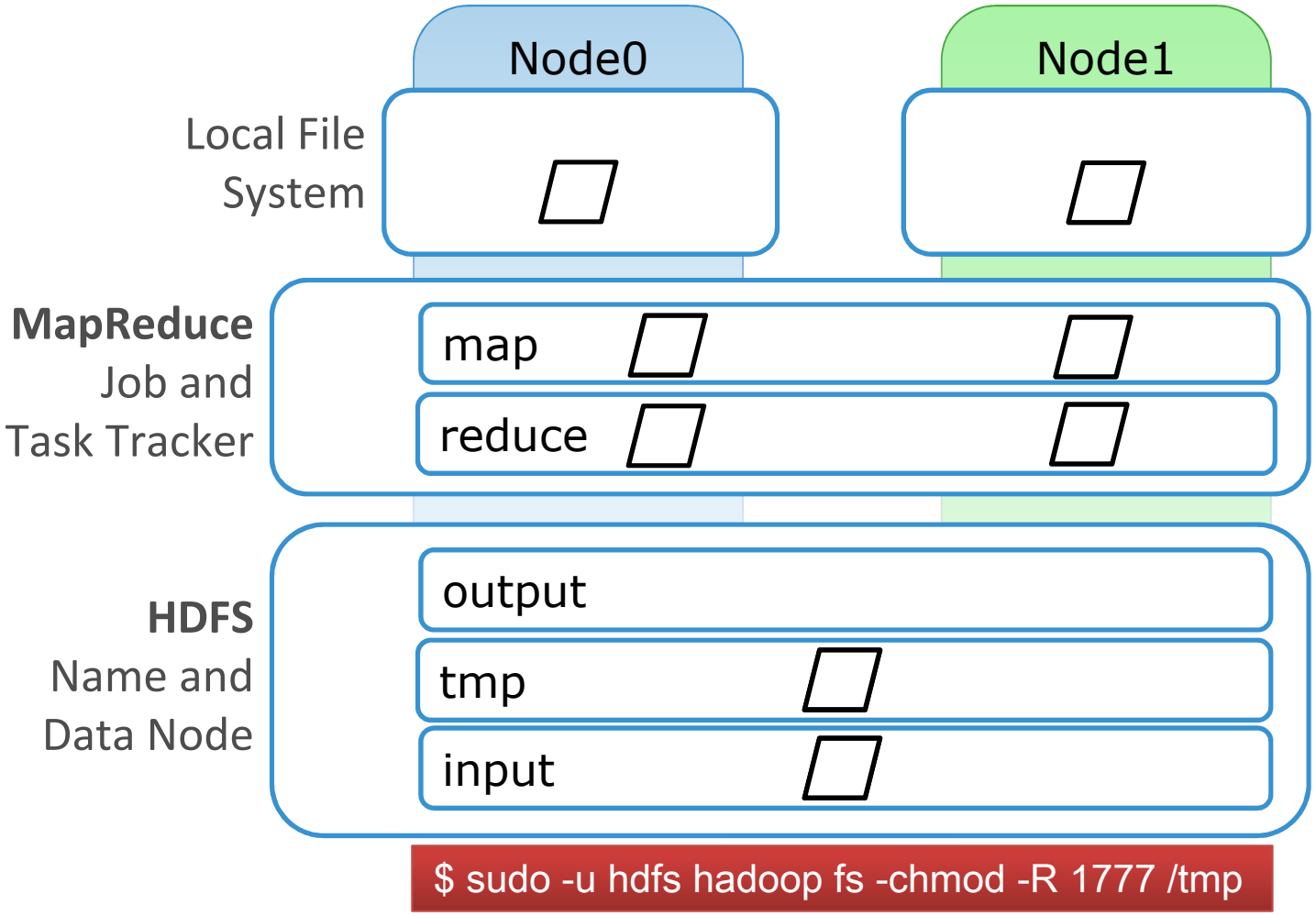
Google MapReduce: 2004



MapReduce Today



MapReduce Today



dir	owner	perm
name	hdfs	700
data	hdfs	700

dir	owner	perm
HDFS	hdfs	775
MAPRED	mapred	775

```
$ sudo -u hdfs hadoop fs -chmod -R 1777 /tmp
```

Job Perimeter

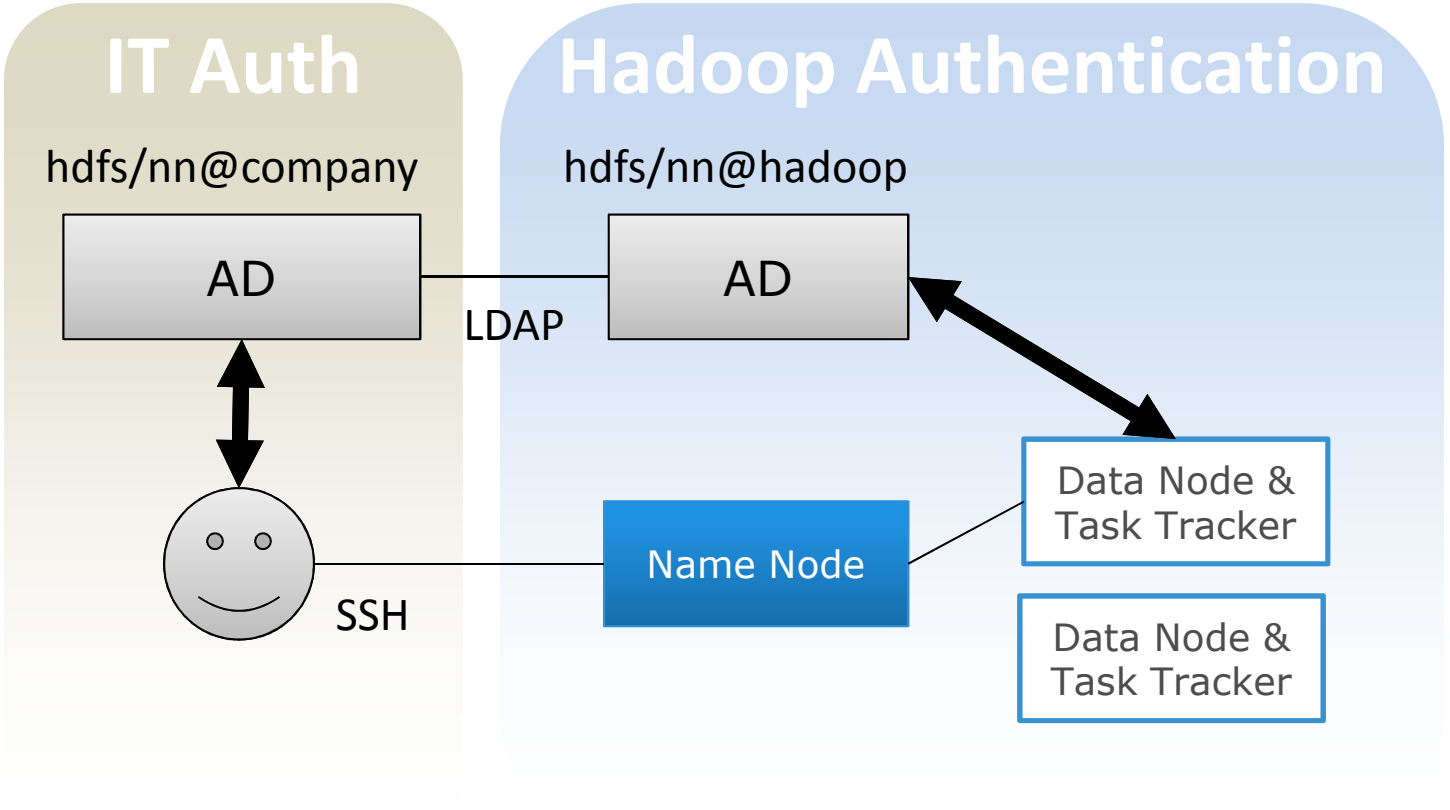
Ad-Hoc

- User Accounts
- Ticket at Login
- Tickets Expire

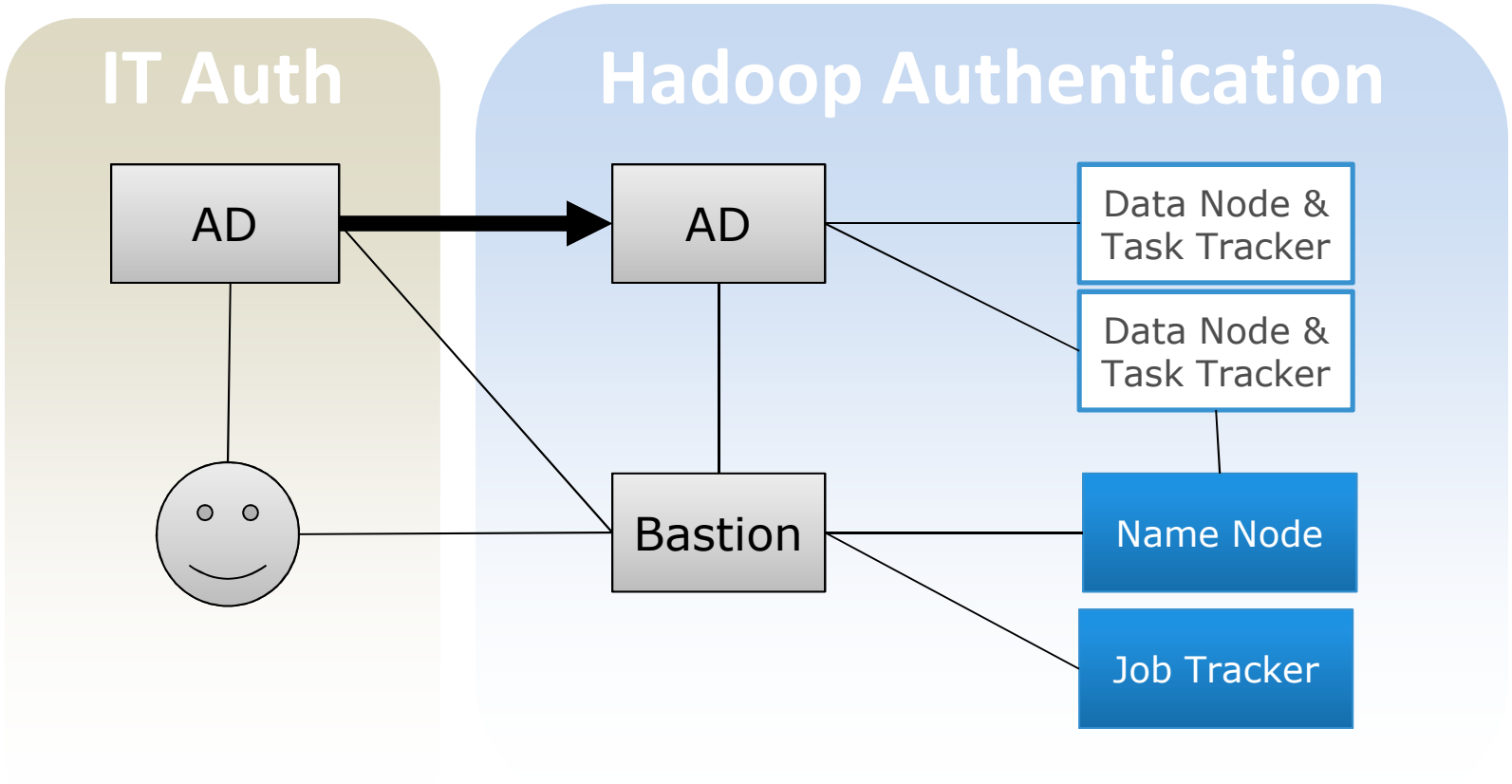
Production

- Batch Accounts
- Tickets from Keytab
- Tickets AutoRenew

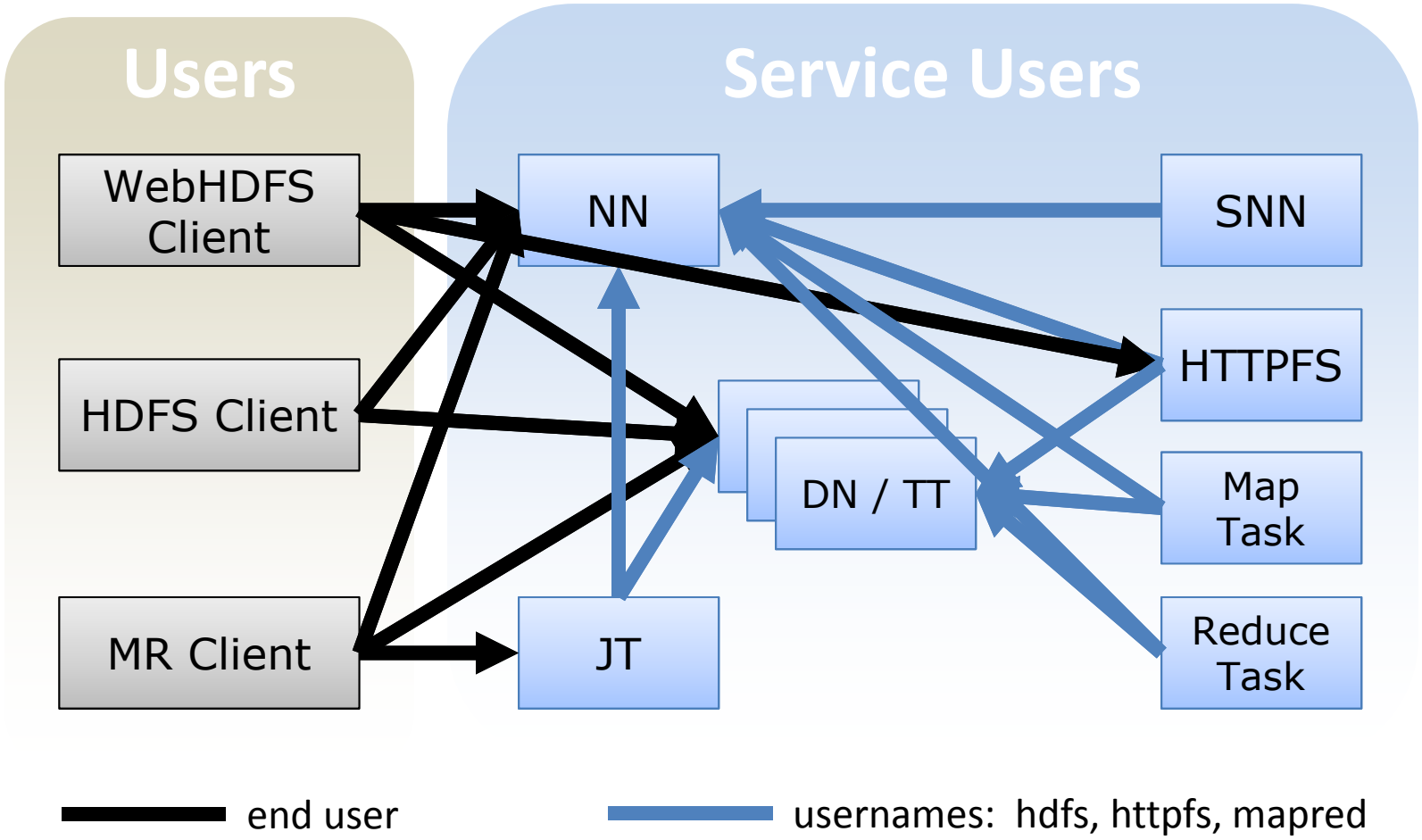
Domain Perimeter



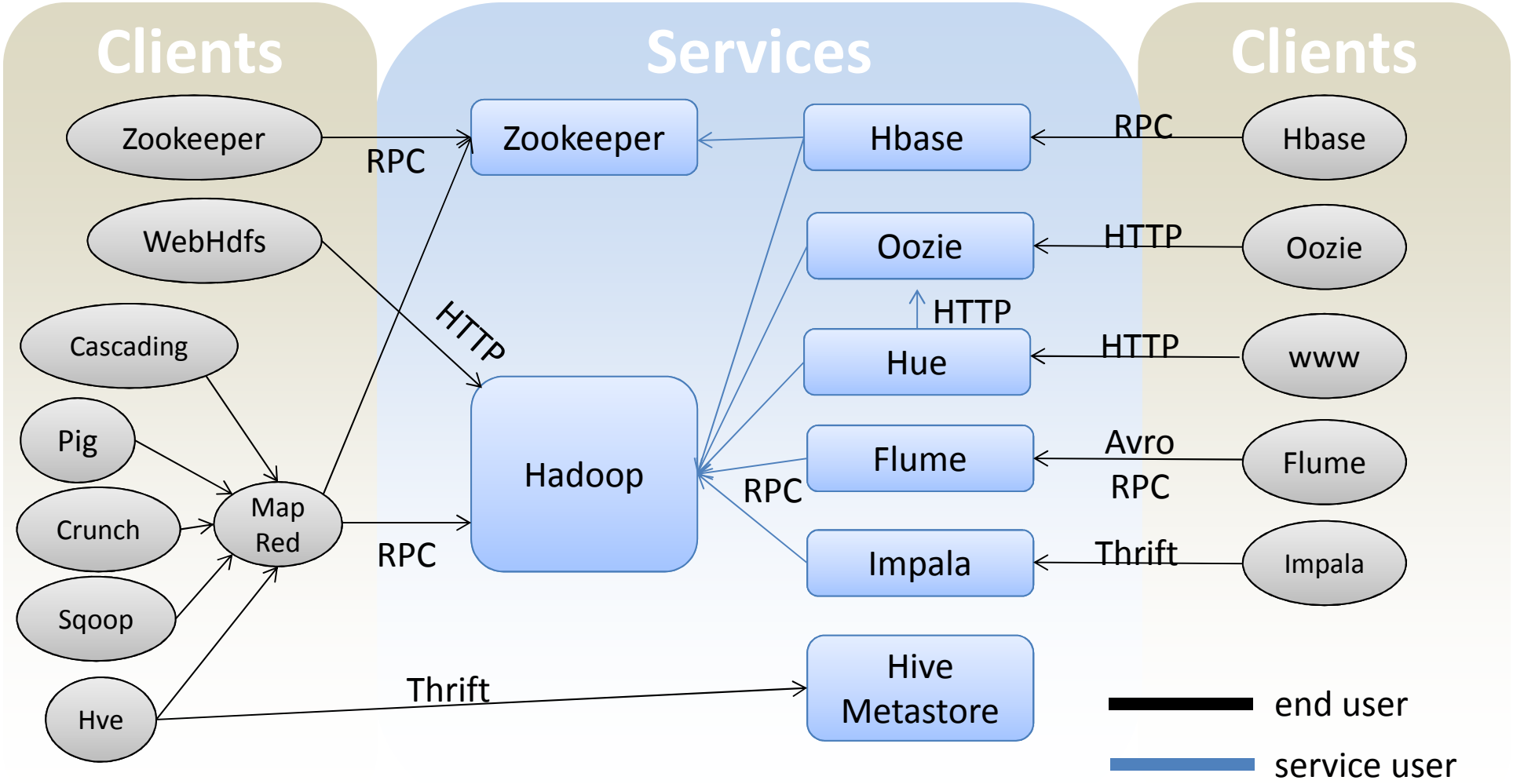
Domain Perimeters



Authentication Perimeters?



Authentication Perimeters?

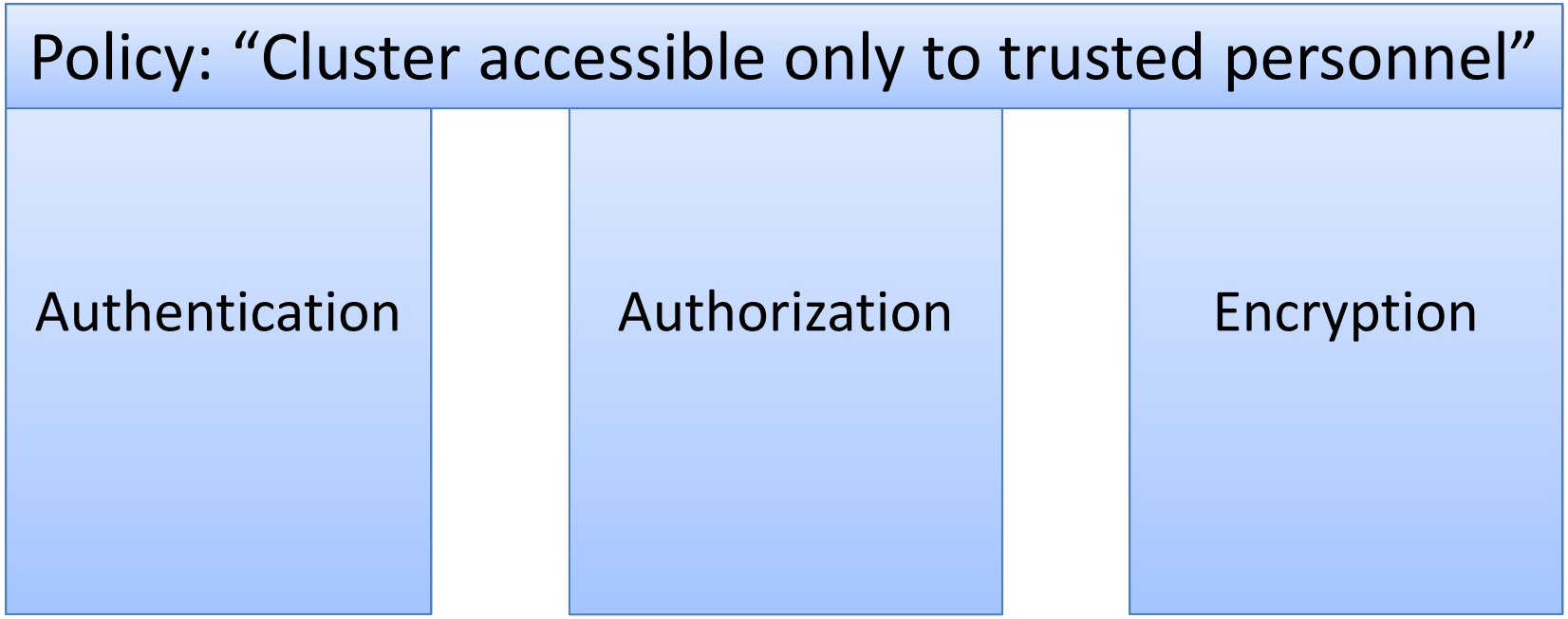


Moment of Reflection

AUDITING


ISACA
Trust in, and value from, information systems
San Francisco Chapter
CRISC
CGEIT
CISM
CISA
2013 Fall Conference – "Sail to Success"

Big Data Security Today



Caveats

- All or Nothing: Nodes without security cannot communicate with secure nodes
- Rolling upgrades to enable cluster security are impossible

Authentication

- End User to Service
- Service to Service
- Service to Service, for a User
- Job Task to Service, for a User

> "Big Data Security Configuration is a PITA. Do only what you really need."
>
> -- 2013 Hadoop Summit

Authorization

- Data

  HDFS, Hbase, Hive Metastore, Zookeeper

- Jobs

  Hadoop, MR, Pig, Oozie, Hue…

- Queries

  Impala, Drill

Secure Impersonation?

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

Encryption

- Stored Data (Emerging)
  - Data Nodes Only
  - Keys Stored Separate from Data Node Storage
  - Keys for “Secure Process” Not Users
- Transmitted Data
  - RPC Supports Only SASL
  - HTTPS

Typical “Security” Setup

Access Controls

1. Authentication (Kerberos)
2. Role Based Authorization (ACLs)
3. Identity Management (AD or LDAP)

Strong Authentication?

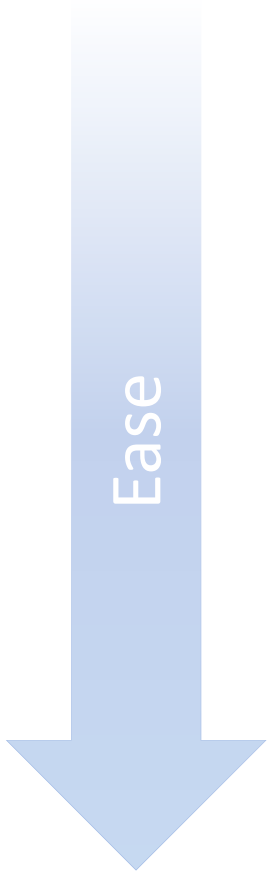
Older versions ran all daemons as single user (hadoop)

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

group	users
hadoop	hdfs, mapred

```
$ sudo -u hdfs hadoop fs -chmod -R 1777 /tmp
```

Trying to Get to Compliance



1. Node Access (Authentication)
2. Node API Authentication
3. Store Data (Disk Encryption)
4. Transmit Data (Net Encryption)
5. RBAC (Job ACLs)
6. Logs

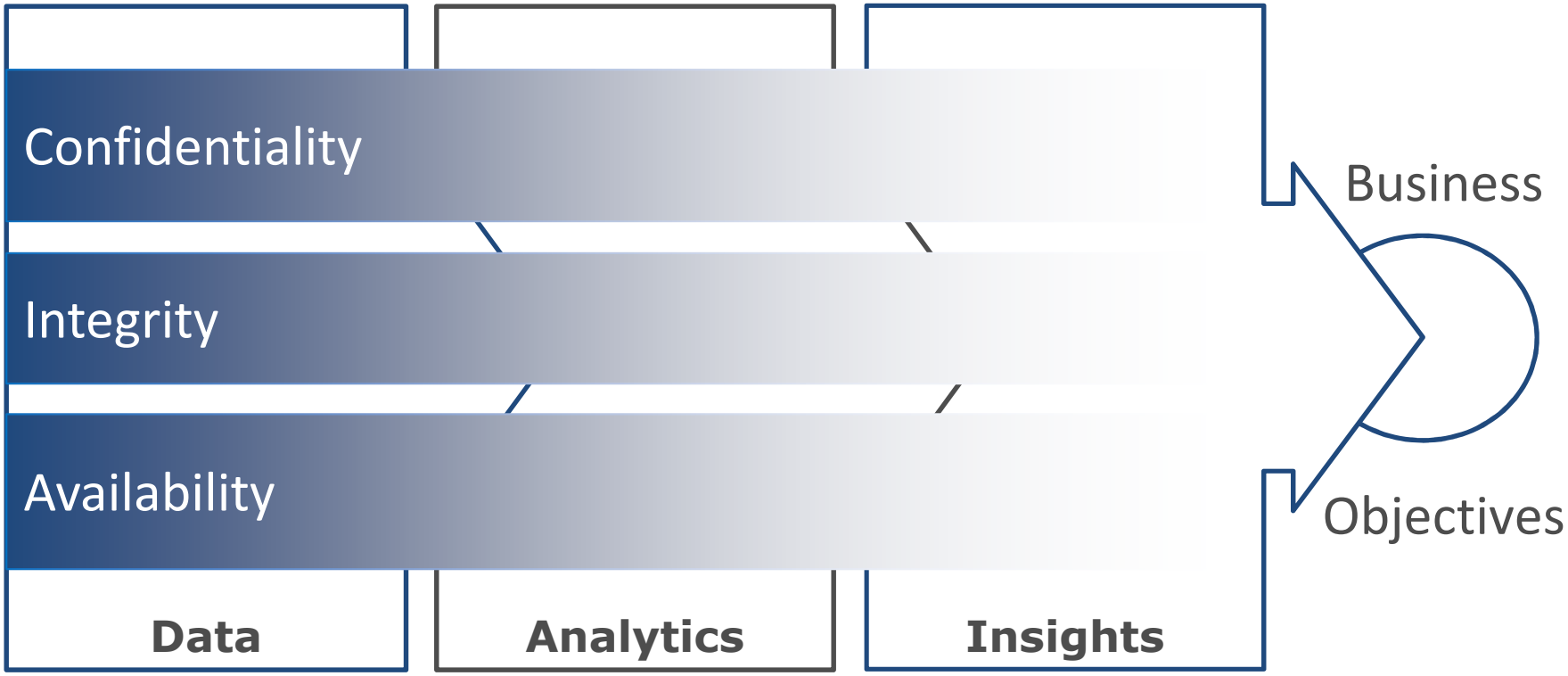
MitM?
root?
NIC control?

sudo -u hdfs hadoop fs -rmr /

Threat Model Thoughts

- Confidentiality
  - Regulated Data
  - Intellectual Property
- Integrity
  - Analysis
  - Output
- Availability
  - Data Load
  - Jobs, Analysis

Future


Confidentiality
Integrity
Availability
Data
Analytics
Insights
Business
Objectives

Regulated Data Example

- Confidentiality
  - Stored
  - Transmitted
- Integrity
  - Backup
  - Restore
- Availability

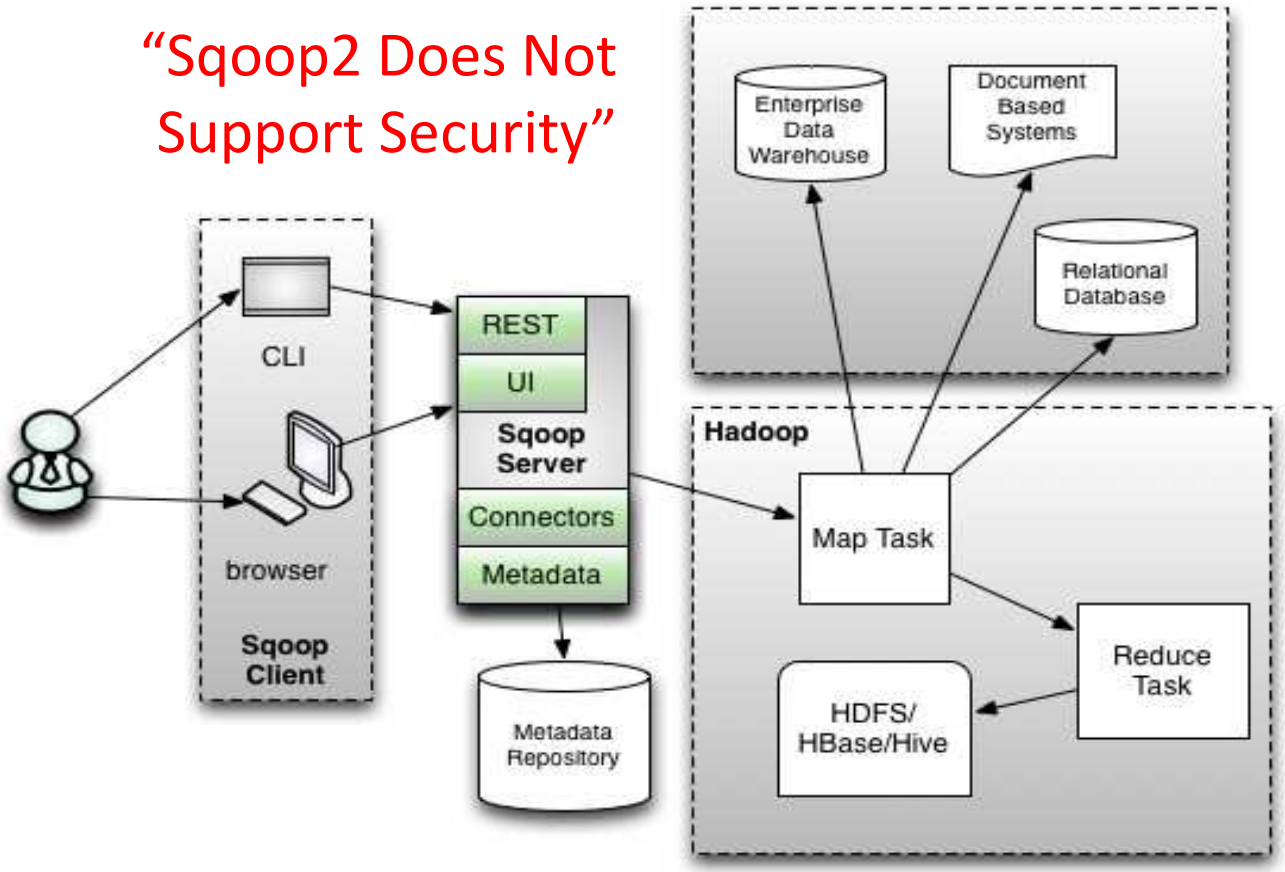
Cost Calculations

- Error
- Load Time
- Drop Time
- Re-Load Time

Regulated Data Example: Sqoop2

Bulk Data Transfer Tool – Import/Export

"Sqoop2 Does Not Support Security"



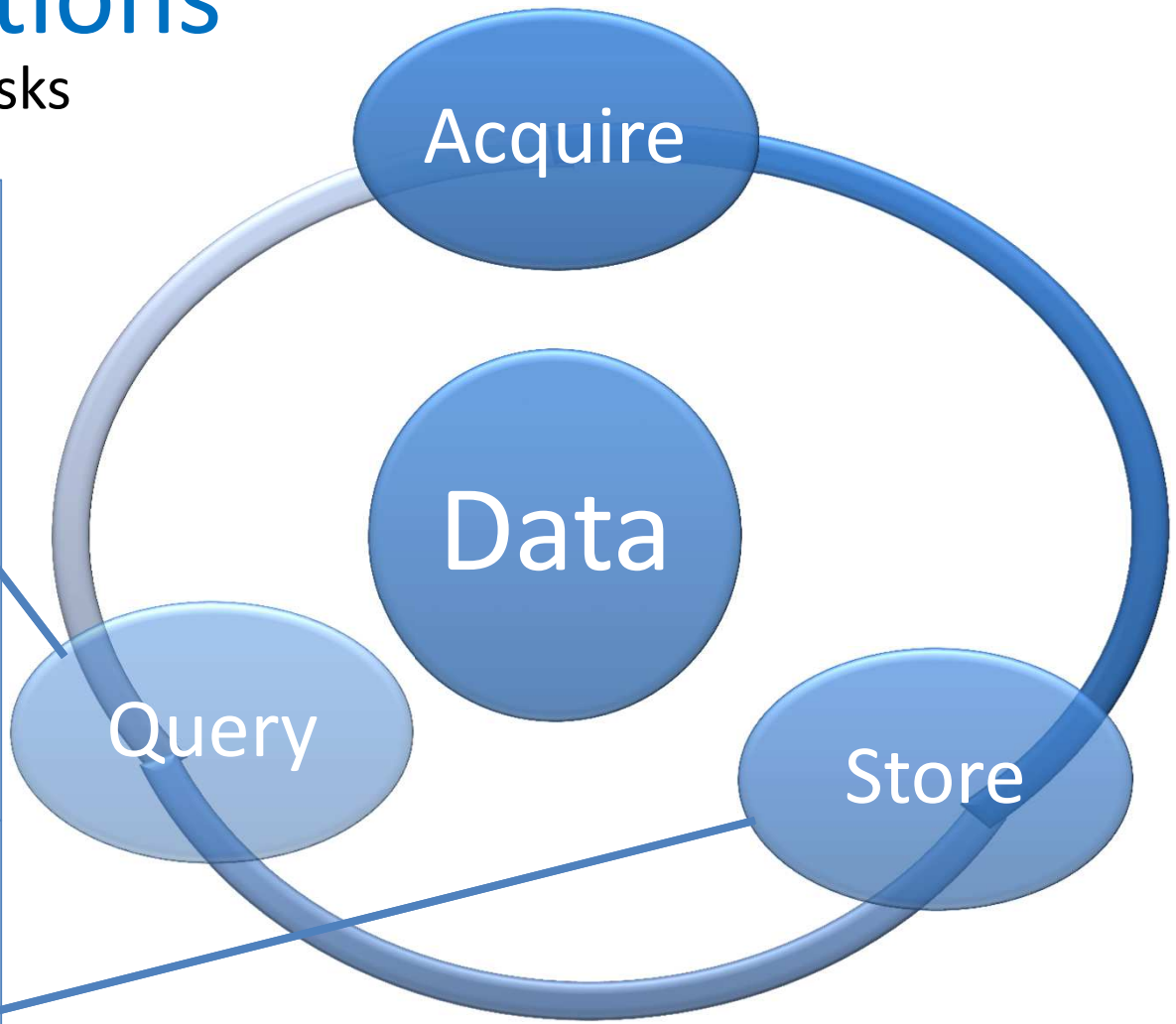
https://blogs.apache.org/sqoop/entry/apache_sqoop_highlights_of_sqoop

Future Directions

Building Controls for Risks

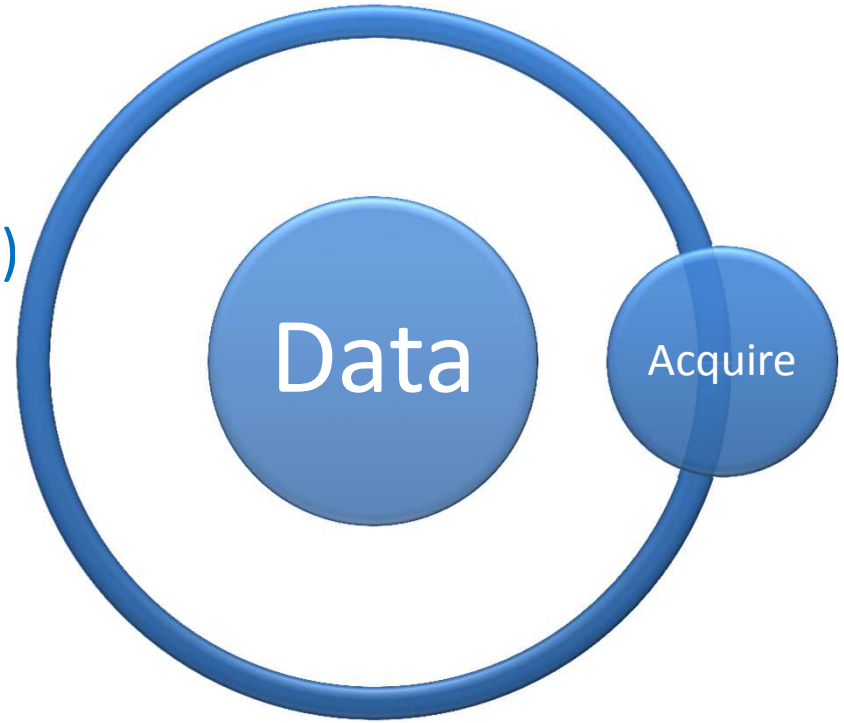
- MapReduce
- Pig (simple query language)
- Hive (SQL queries)
- Cascading (workflow)
- Mahout (machine learning)
- Hama (scientific compute)
- Drill (ad-hoc query)

- Hbase (column-orient DB)
- Zookeeper (coordination)
- HDFS



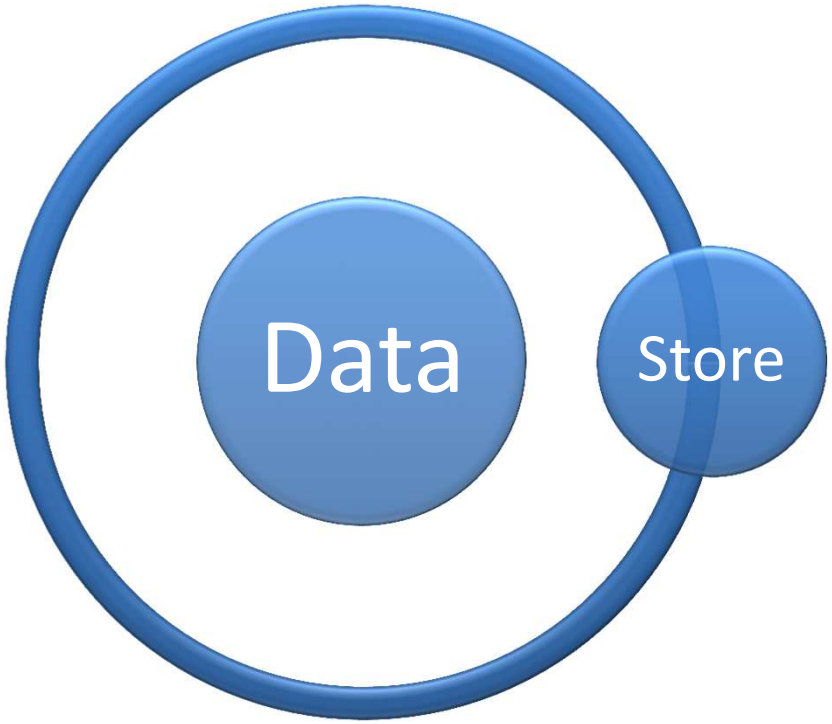
Acquire

- Data Validation
- Aggregation
  - DOB
  - Age Group (18-24, 25-36, etc.)
- Salt (Noise)
- Imitation (Synthetic)
- Replacement (Swapping)
- Wiping (Suppression)



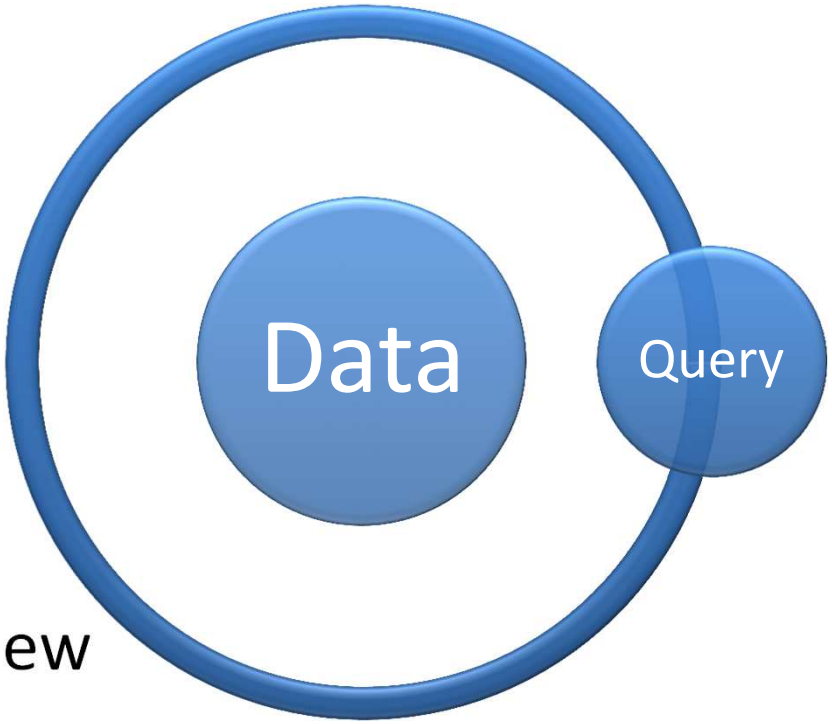
Store

- Encryption
- Unique Users (RBAC)
- API Authentication
- “Hardened Clusters”
- Logs
- Forensics



Query

- Encryption
- Unique Users (RBAC)
- Load Restrictions
- Add, Remove, Change Jobs
- Job Alerts
- Scheduling
- Secure Code Practices / Review



CSA "Securing Big Data"

SQL Security – Old	NoSQL Security –
Config Management	Cell-Level Access Labels
Multi-Factor Authentication	Keberos-Based Authentication
Data Classification	ACLs
Data Encryption	
Consolidated Audit/Report	
Database Firewall	
Vulnerability-Scanner	

Example: Accumulo

Cell-Level Access – Distributed Key/Value Store

- NoSQL (HBase) perf…with security
- Timeline
  - 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("alexander","snowden"));
scan.fetchFamily("attributes");
for(Entry<Key,Value> entry : scan) {
    String row = entry.getKey().getRow(); Value
    value = entry.getValue();
}
```


https://accumulo.apache.org/1.5/accumulo_user_manual.html

Conclusions

- Big Data
  - Performance Pressure
  - Customer Concern / Backlash
- Operations Risks
  - Perimeter Holes
  - Business Logic Error
  - Weak Policies
- Auditing
  - Trusted Personnel
  - Perimeters

THANK YOU!

Auditing Big Data for Privacy, Security and Compliance

Davi Ottenheimer @daviottenheimer

Senior Director of Trust, EMC

In-Depth Seminars – D21