NC STATE UNIVERSITY Computer Science

Insights to Hadoop Security Threats

Presenter:
Anwesha Das
Peipei Wang

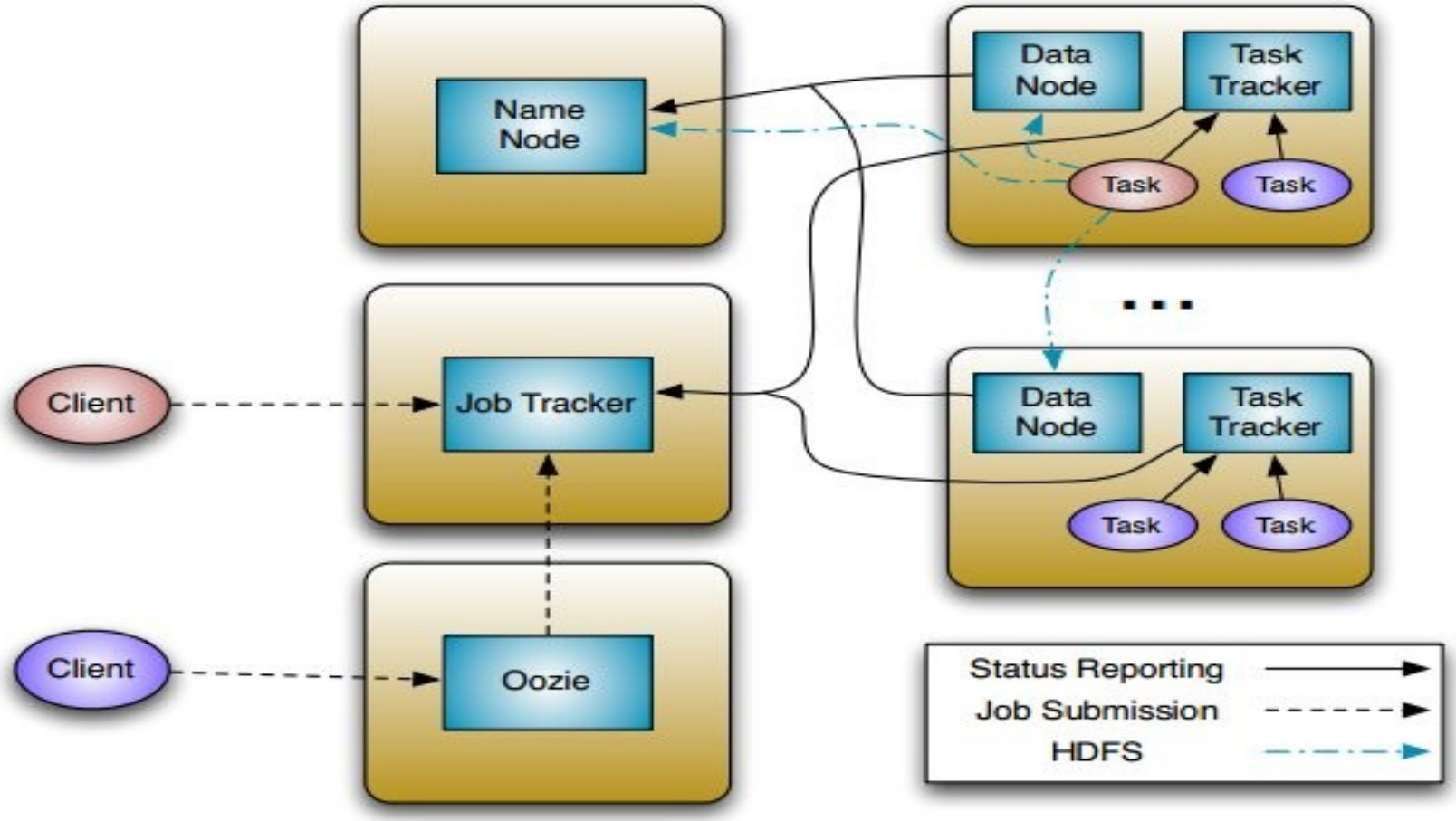
Hadoop

- HDFS – Hadoop Distributed File System
  - Combines cluster's local data into a single namespace
  - Data replicated to multiple machines
  - Locality Information to clients
  - NameNode, DataNodes

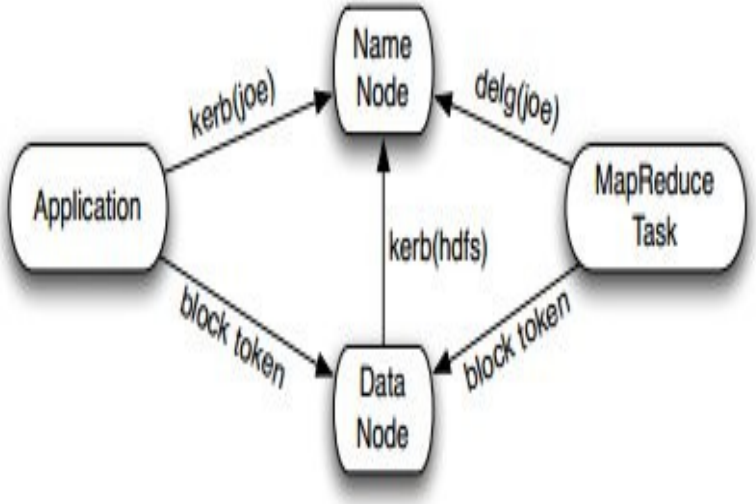
- MapReduce Engine
  - Batch Computation Framework
  - Tasks re-executed on failure
  - Optimizes for data locality of input
  - JobTracker, TaskTracker

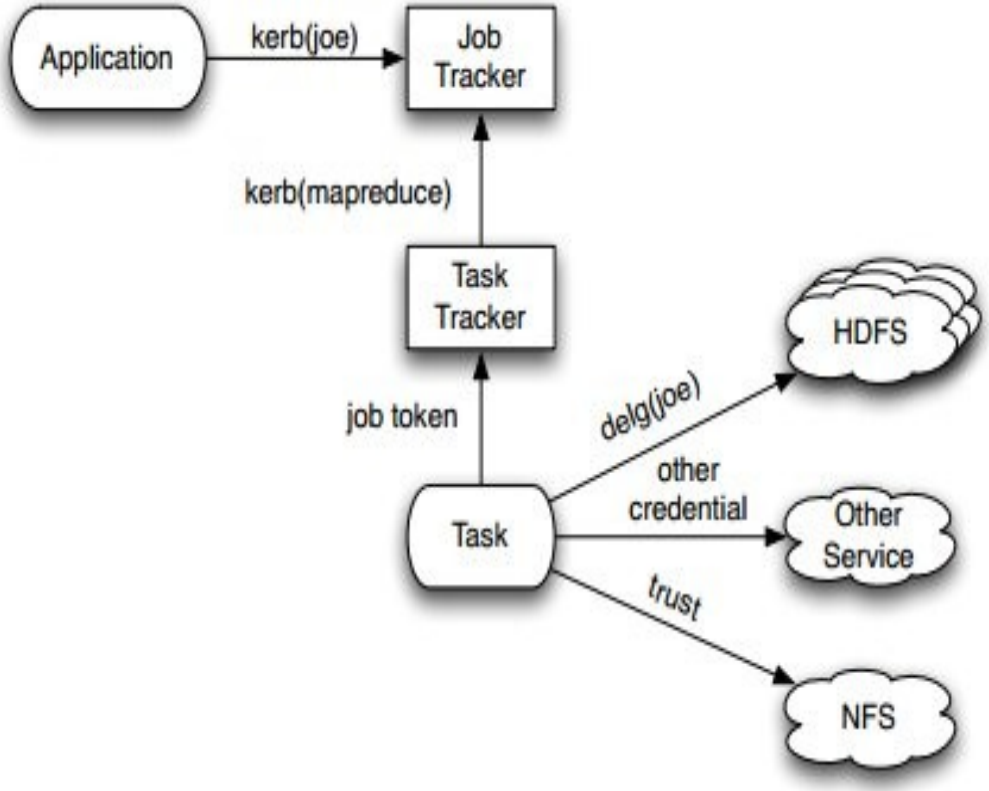
- Oozie Workflow system – Auxillary Service

Hadoop High Level Overview



Hadoop High Level Overview





Security Loopholes

LoopHoles

- Poor default SASL (Simple Authentication Security Layer)
- quality of protection
- Incomplete authentication
- Lack of data security that flows between the nodes

Security Threats

- Unauthorized Release of information
- Unauthorized Modification of information
- DDOS – denial of services, eavesdropping, replay attack

Solutions

Authentication

- Kerberos

Token (Secret Key)

- Delegation, Block access, Job

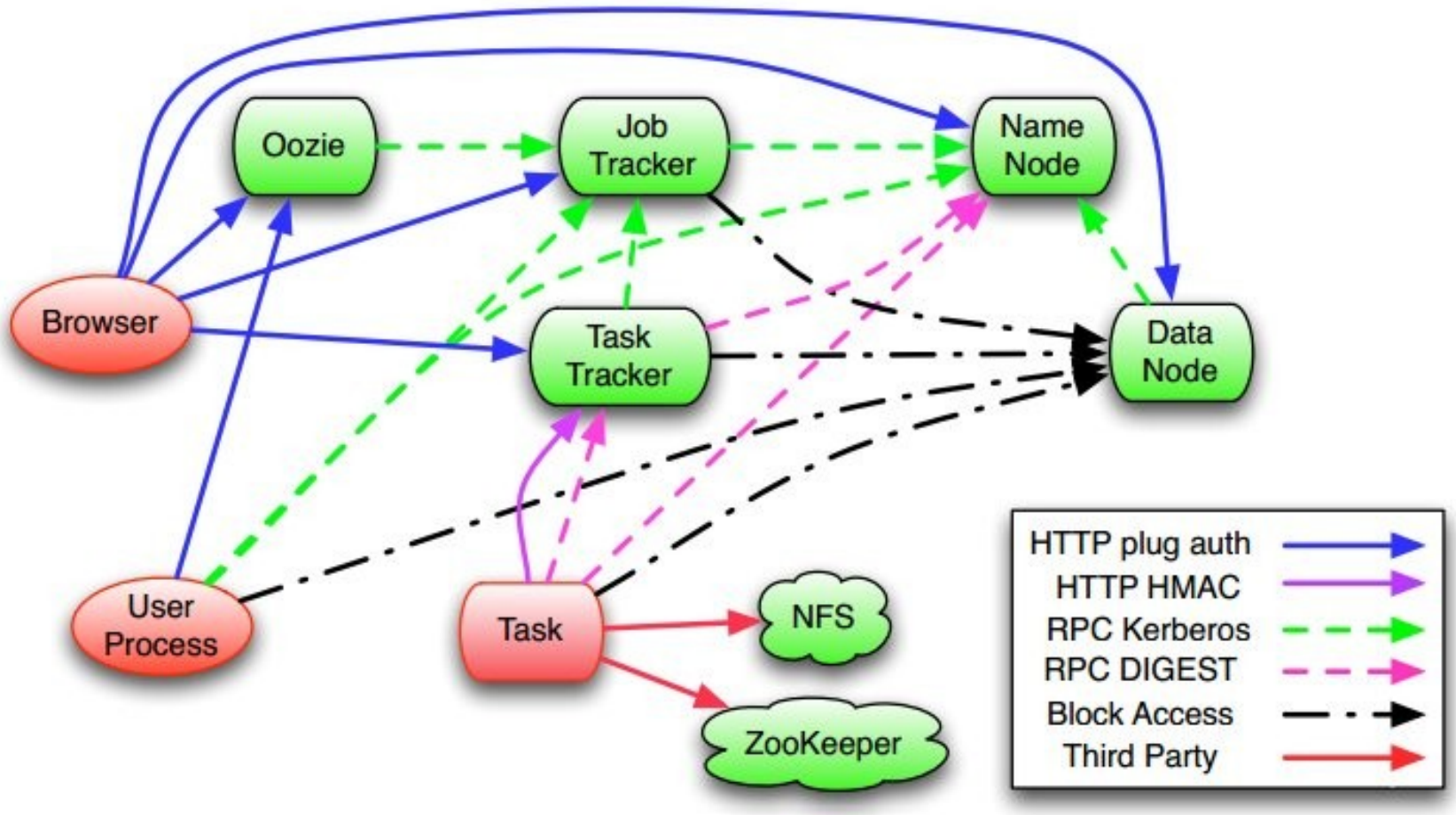
Security Protocol

- RPC, Data Transfer Protocol
- Rate Limit

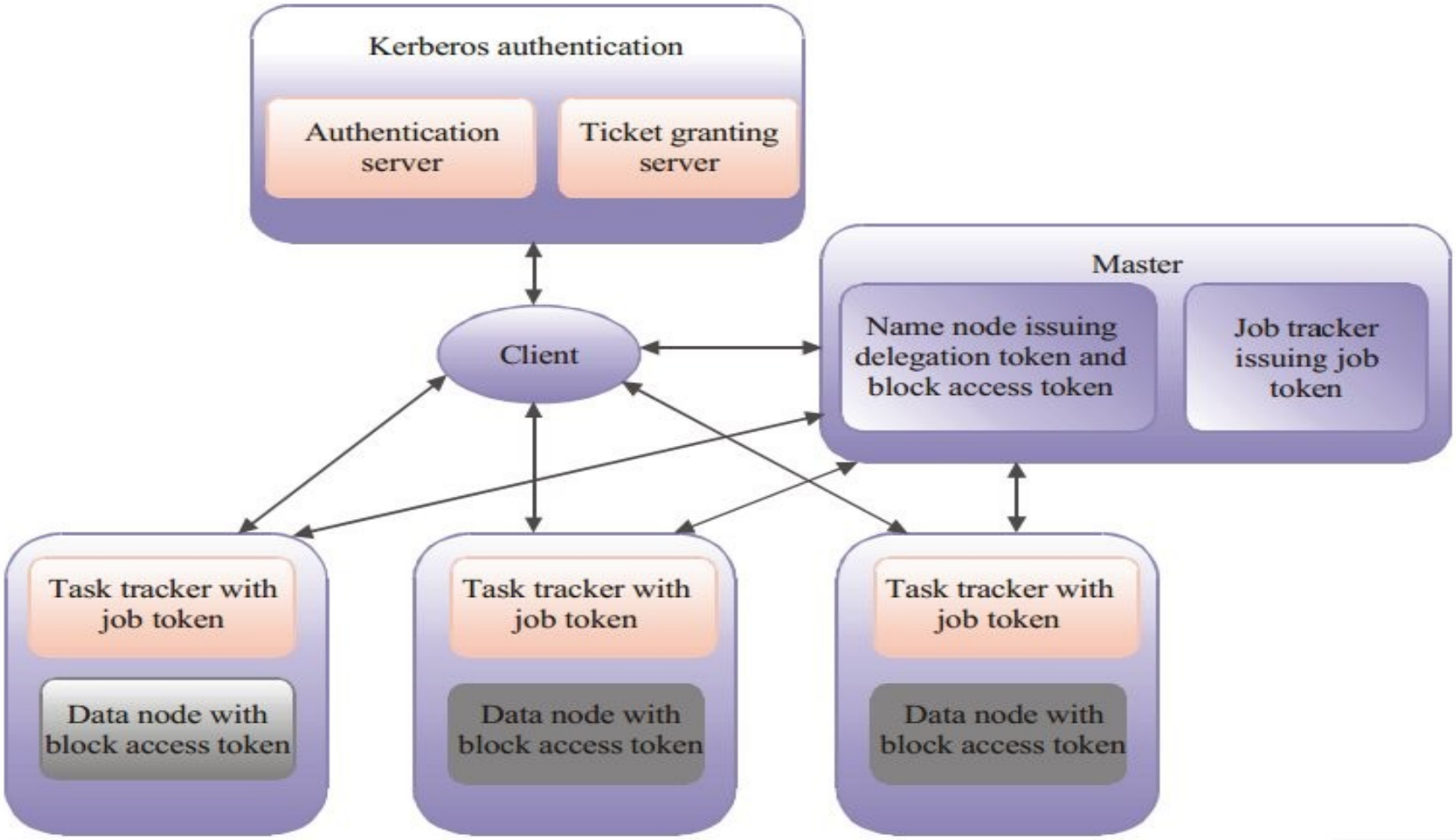
Tokens

- SecretKey generated by NameNode, JobTracker
- Shared with clients, TaskTracker, DataNodes
- BlockAccess, Job token time-stamped for validity
- Assume secure channel token passing (Trusted n/w)

Authentication Flow



Token Flow


Kerberos authentication
Authentication server
Ticket granting server
Master
Name node issuing delegation token and block access token
Job tracker issuing job token
Client
Task tracker with job token
Data node with block access token
Task tracker with job token
Data node with block access token
Task tracker with job token
Data node with block access token

Attacks ??

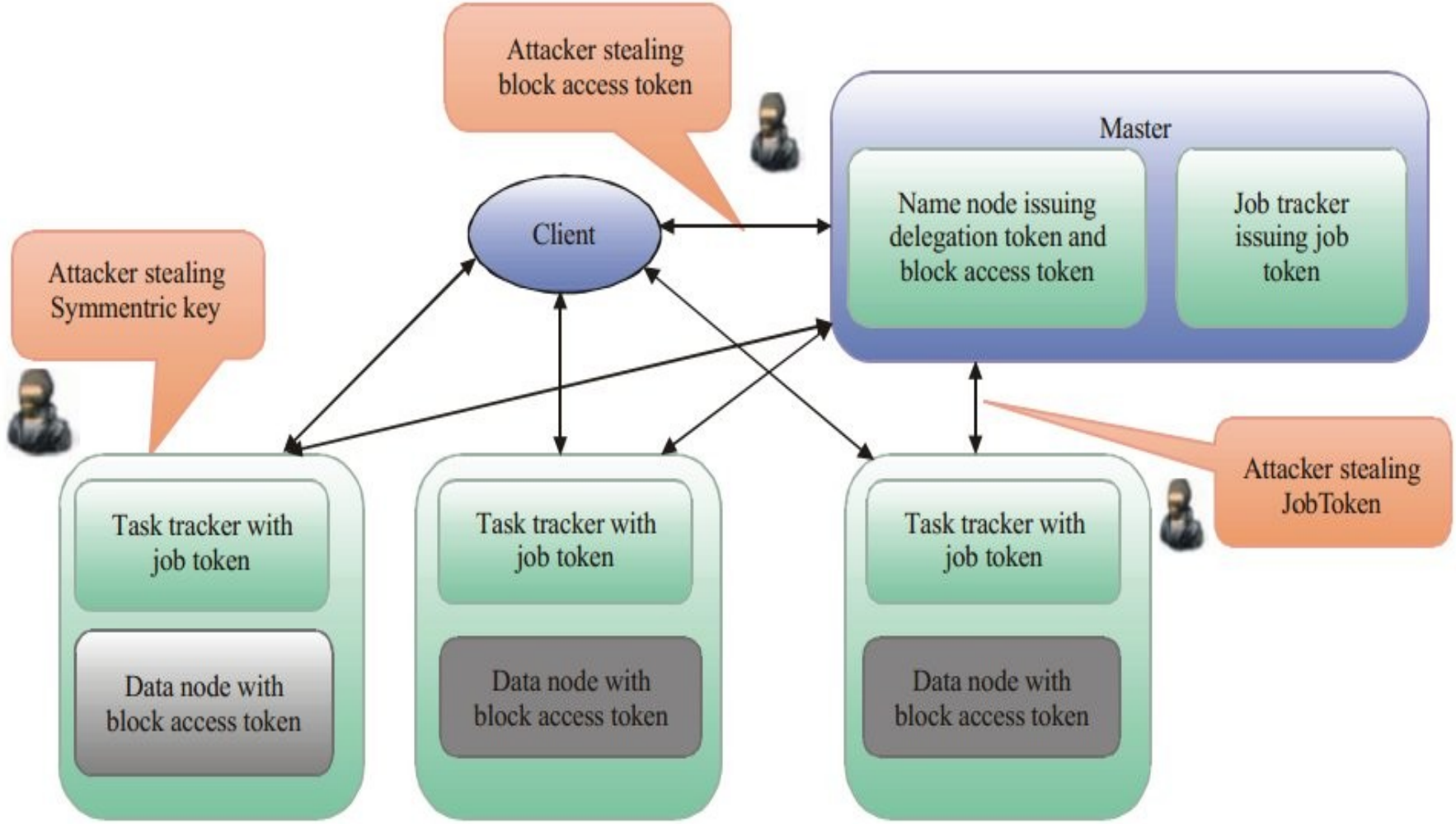
DataNodes donot enforce access control

- Anyone with block_id can access data blocks
- BlockAccess Token shared between all DataNodes
- Symmetric Hashing HMAC-SHA1 technique
- Attacker with shared key can affect all DataNodes
- Affect Integrity

Insecure Channel

- Intruder access JobTokens
- Modify results of map or reduce tasks
- Affect confidentiality

Attack ??


Attacker stealing block access token
Master
Name node issuing delegation token and block access token
Job tracker issuing job token
Client
Attacker stealing Symmentric key
Attacker stealing JobToken
Task tracker with job token
Data node with block access token
Task tracker with job token
Data node with block access token
Task tracker with job token
Data node with block access token

Circumvention

Solution

- Replace HMAC-SHA – Cryptographic Token Encryption
- Asymmetric Encryption – Public-Private Key(RSA)
- Symmetric Encryption – Private Keys
  - Refined HMAC-SHA1 technique
  - Share same secret key, unique secret for every node instead of the same key

Hadoop Security Work

- Hadoop Storage Security
  - encryption
  - data-at-rest encryption
  - encryption on the wire
  - data in transit-protection
  - isolation
  - SELinux

example: secHDFS,VPN-Cubed

Hadoop Security Work

- Hadoop Computation Security
  - authentication
  - customized mapper and reducer
  - differential privacy
  - trusted computing

example:Airavat, TPM-based attestation

Our Plan of Work

- Secure Deployment of Source code version 0.21x
- Identify security breaches ??
  - DOS – resource exhaustion
  - DOS – Node availability
  - Issues with usage of tokens, access control, rate-limit?
- Help from bug repository !!
- Does the attack exist across multiple versions of Hadoop?
- Evaluation

Relevance of our research

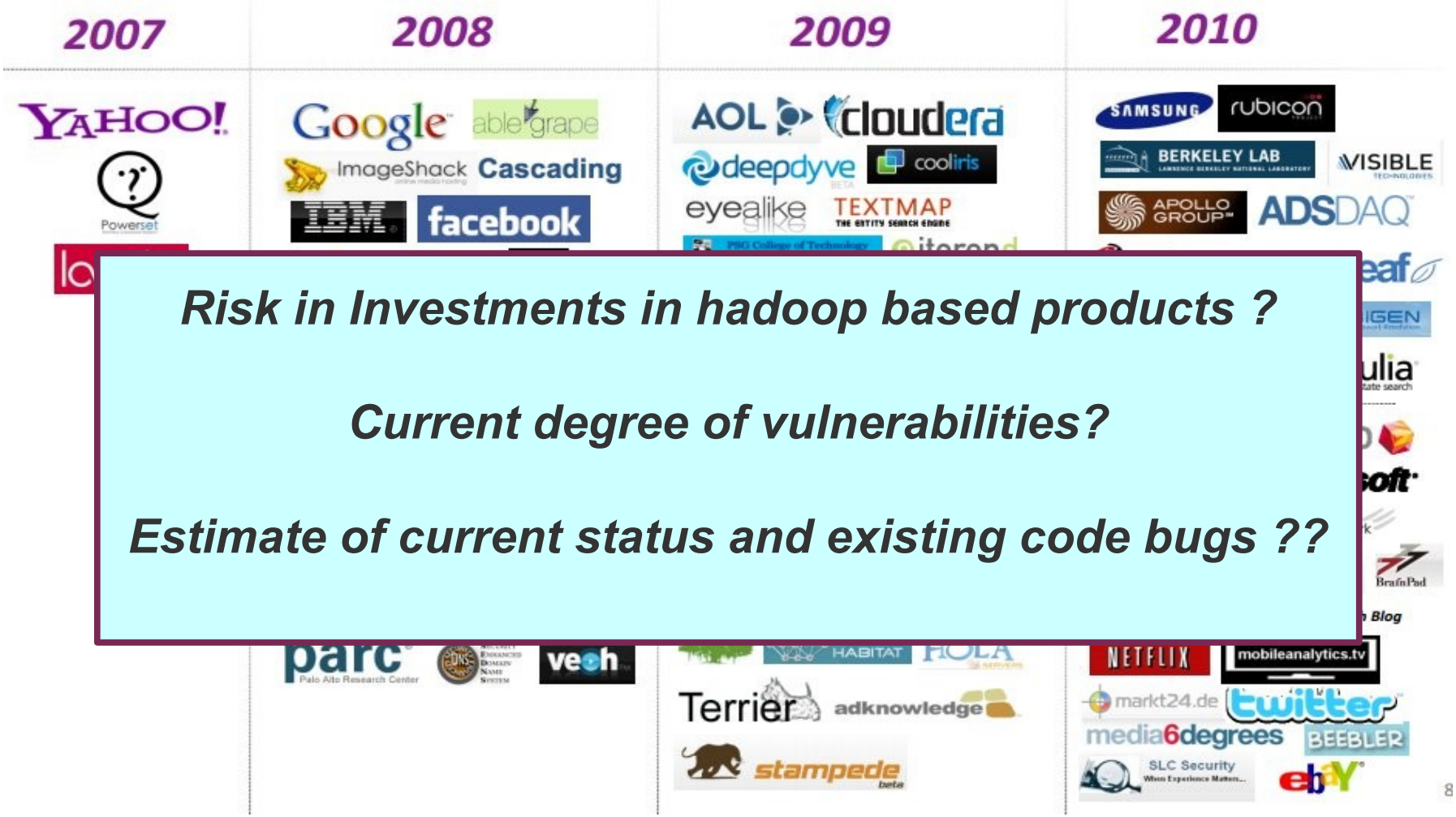
- Encryption hot topic in hadoop
- Third party innovation for security integration
- Lot more to be done, future Roadmap exists !!
- Security least (secondary) concern but booming usage

- Absence of comprehensive evaluation of threats !!

Is our work Important ??

2007	2008	2009	2010
YAHOO!	Google	AOL	SAMSUNG
Powerset	ableGrape	cloudera	rubicon
last·fm	ImageShack	deepdyve	BERKELEY LAB
	Cascading	cooliris	VISIBLE
	IBM	eyealike	APOLLO GROUP
	facebook	TEXTMAP	ADSDAQ
	ENORMO	PSG College of Technology	rackspace HOSTING
	A9	iterend	RapLeaf
	krugle	tailsweep	wordnik
	rackspace HOSTING	hulu	MOBIGEN
	Lookery Control freaks welcome	RapLeaf	comSCORE
	The New York Times	USCMS	trulia real estate search
	joost	Ning	quantcast
	zvents Discover Things To Do	amazon web services	Accela COMMUNICATIONS
	FORMATION CIENCES INSTITUTE	pressflip	Forward3D
	News Corporation	detikSearch	Linked in
	Cornell University Computing and Information Science	WorldLingo	Microsoft
	VISIBLE MEASURES	Systems@ETH Zürich	Infochimps Find the world's data
	LOTAME	VK SOLUTIONS Global Solutions Provider	Pharm2Phork
	NetSeer	TARAGARA Innovation + Quality + Simplicity	ADMELD
	parc Palo Alto Research Center	HOSTING HABITAT	gumgum
	SECURITY ENHANCED DOMAIN NAME SYSTEM	HOLA	BrainPad
	veoh	Terrier	Pronux The Datagraph Blog
		adknowledge	NETFLIX
		stampede beta	mobileanalytics.tv
			markt24.de
			twitter
			media6degrees
			BEEBLER
			SLC Security When Experience Matters...
			ebay

Is our work Important ??



Risk in Investments in hadoop based products ?

Current degree of vulnerabilities?

Estimate of current status and existing code bugs ??

References

- A New Solution of Data Security Accessing for Hadoop Based on CP-ABE
- A security framework in G-Hadoop for big data computing across distributed Cloud data centres: http://www.sciencedirect.com/science/article/pii/S002200001400018X
- A survey on security issues in service delivery models of cloud computing: http://www.sciencedirect.com/science/article/pii/S1084804510001281
- Access Control for Sensitive Data in Hadoop Distributed File Systems: www.thinkmind.org/download.php?articleid=infocomp_2013_4_10_10050.pdf
- Access Security on Cloud Computing Implemented in Hadoop System: http://www.slideshare.net/jgabriellima/access-security-on-cloud-computing-implemented-in-hadoop-system
- Addressing cloud computing security issues: http://www.sciencedirect.com/science/article/pii/S0167739X10002554
- Airavat: Security and Privacy for MapReduce. http://static.usenix.org/events/nsdi10/tech/full_papers/roy.pdf
- Security in Data Intensive Computing Systems: http://link.springer.com/chapter/10.1007%2F978-1-4614-1415-5_16

References

- Danger in the clouds: http://www.sciencedirect.com/science/article/pii/S1353485808701405
- HIGH LEVEL VIEW OF CLOUD SECURITY: ISSUES AND SOLUTIONS: http://airccj.org/CSCP/vol4/csit42005.pdf
- Horus: Fine-Grained Encryption-Based Security for Large-Scale Storage: https://www.usenix.org/conference/fast13/technical-sessions/presentation/li_yan
- Implement a reliable and secure cloud distributed file system:
- Implementation of identity based distributed cloud storage encryption scheme using PHP and C for Hadoop File System:
- Managing Data Access on Clouds: A Generic Framework for Enforcing Security Policies:
- Privacy in Map Reduce Based Systems: A Review: http://ijcsmc.com/docs/papers/February2014/V3I2201475.pdf
- Secure Hadoop with Encrypted HDFS: http://link.springer.com/chapter/10.1007%2F978-3-642-38027-3_14
- Security and privacy for storage and computation in cloud computing: http://www.sciencedirect.com/science/article/pii/S0020025513003320
- Security and privacy in cloud computing: A survey

References

- SpongeFiles: Mitigating Data Skew in Mapreduce Using Distributed Memory: http://doi.acm.org/10.1145/2588555.2595634
- The security of cloud computing system enabled by trusted computing technology:
- Toward Data Confidentiality via Integrating Hybrid Encryption Schemes and Hadoop Distributed File System
- Towards a More Secure Apache Hadoop HDFS Infrastructure: http://link.springer.com/chapter/10.1007%2F978-3-642-38631-2_64
- Towards a trusted {HDFS} storage platform: Mitigating threats to Hadoop infrastructures using hardware-accelerated encryption with TPM-rooted key protection: http://www.sciencedirect.com/science/article/pii/S2214212614000155

References

- Hadoop Security Design http://carfield.com.hk:8080/document/distributed/hadoop-security-design.pdf
- Adding Security to Apache Hadoop http://br.hortonworks.com/wp-content/uploads/2011/10/security-design_withCover-1.pdf
- Improving Security of Parallel Algorithm Using Key Encryption Technique: http://scialert.net/qredirect.php?doi=itj.2013.2398.2404&linkid=pdf
- Integrating Hadoop with Kerberos: http://www.kerberos.org/events/2010conf/2010slides/2010kerberos_owen_omalley.pdf
- Securing Big Data Hadoop: http://www.ijcsit.com/docs/Volume%205/vol5issue02/ijcsit20140502263.pdf
- Hadoop Based Defense Solution to Handle Distributed Denial of Service (DDoS) Attacks: http://file.scirp.org/Html/4-7800161_34629.htm
- HDFS Encryption: http://blog.cloudera.com/blog/2014/06/project-rhino-goal-at-rest-encryption/
- Oozie: http://dl.acm.org/citation.cfm?id=2443420

References

- Just Add Kerberos? Really? https://media.blackhat.com/bh-us-10/whitepapers/Becherer/BlackHat-USA-2010-Becherer-Andrew-Hadoop-Security-wp.pdfl
- Big Data Security: The Evolution of Hadoop Security Model: http://www.infoq.com/articles/HadoopSecurityModel
- Hadoop CDH – Cloudera: http://www.cloudera.com/content/cloudera/en/products-and-services/cdh.html
- Hadoop: https://hadoop.apache.org/docs/r2.4.1/hadoop-project-dist/hadoop-hdfs/HdfsUserGu
- Big Data Security Gap: Protecting the Hadoop Cluster: http://www.zettaset.com/wp-content/uploads/2014/04/zettaset_wp_security_0413.pdf
- Security Implementation in Hadoop: http://search.iiit.ac.in/cloud/presentations/28.pdf
- Hadoop Security Today and Tomorrow: http://hortonworks.com/blog/hadoop-security-today-and-tomorrow/
- LoopHoles in Hadoop: http://readwrite.com/2014/08/13/hadoop-slow-security-issues-still-popular

References

- Towards a More Secure Apache Hadoop HDFS Infrastructure: http://link.springer.com/chapter/10.1007/978-3-642-38631-2_64
- Big Data Security: http://www.sciencedirect.com/science/article/pii/S1353485812700636
- A Novel Authentication Service for Hadoop in Cloud Environment: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6354591
- Cloud Security in MapReduce: http://hackedexistence.com/downloads/Cloud_Security_in_Map_Reduce.pdf
- Taking hadoop Security to the next level: http://www.securityweek.com/bigger-data-smaller-problems-taking-hadoop-security-next-level

Thank you