Securing Big Data – Learning and Differences from Cloud Security

Samir Saklikar

RSA, The Security Division of EMC

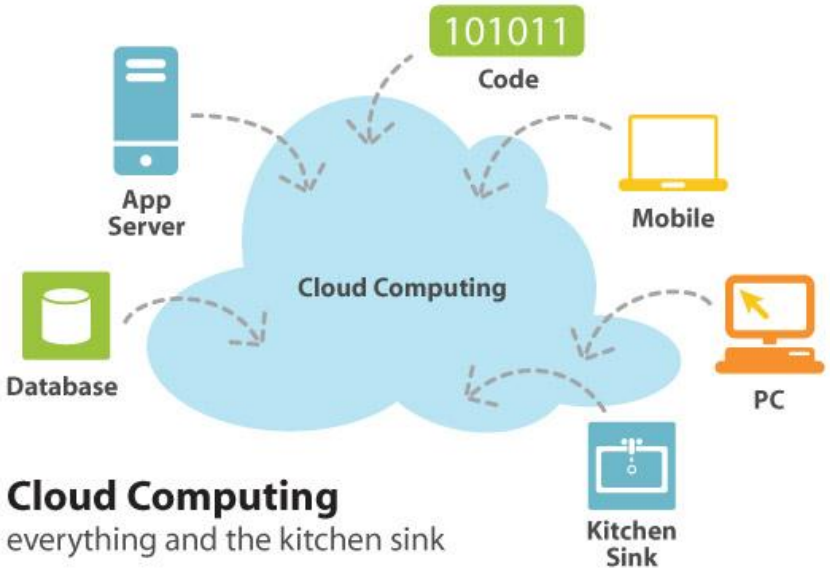
Session ID: DAS-108
Session Classification: Advanced

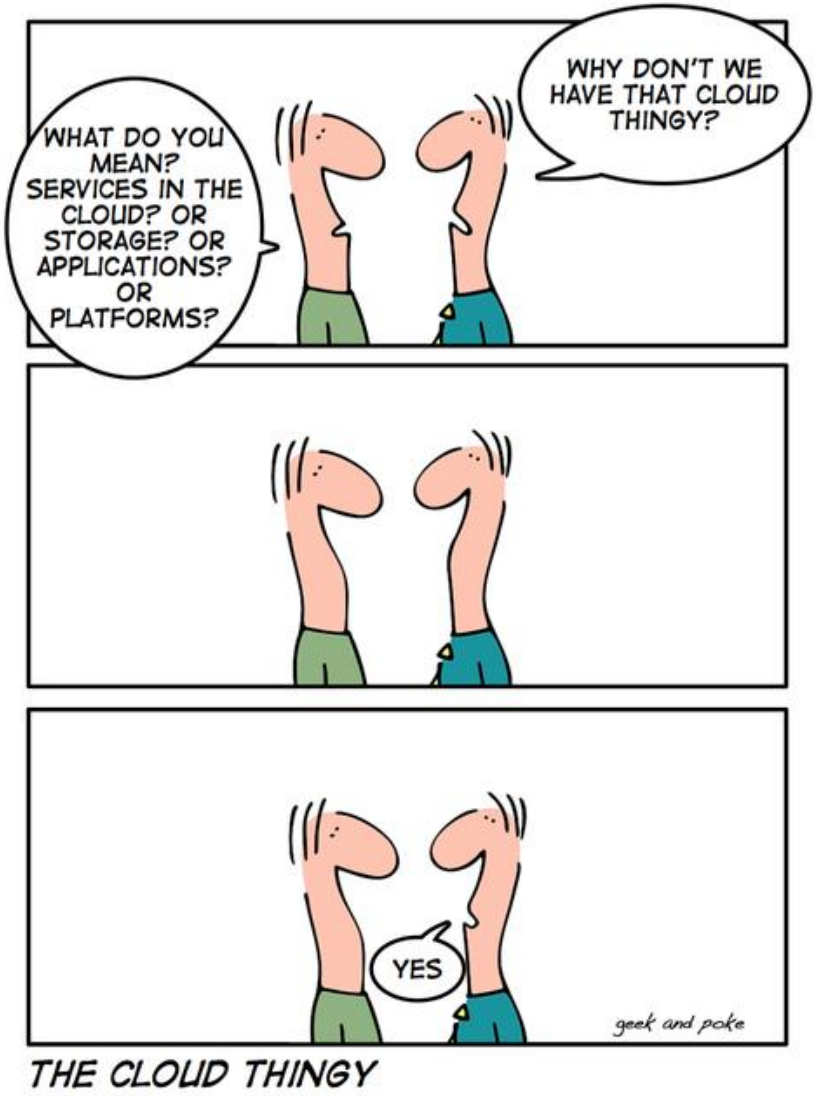
Agenda

- Cloud Computing & Big Data
  - Similarities from a security perspective
- Learning from Cloud Computing Security
- Unique Challenges in Security for Big Data
- Directions for new technology investigation
- Handling PII within Big Data systems
- Conclusion

Cloud Computing and Big Data
Two sides of the same coin?

This whole Cloud Thingy..
(now including the kitchen sink)

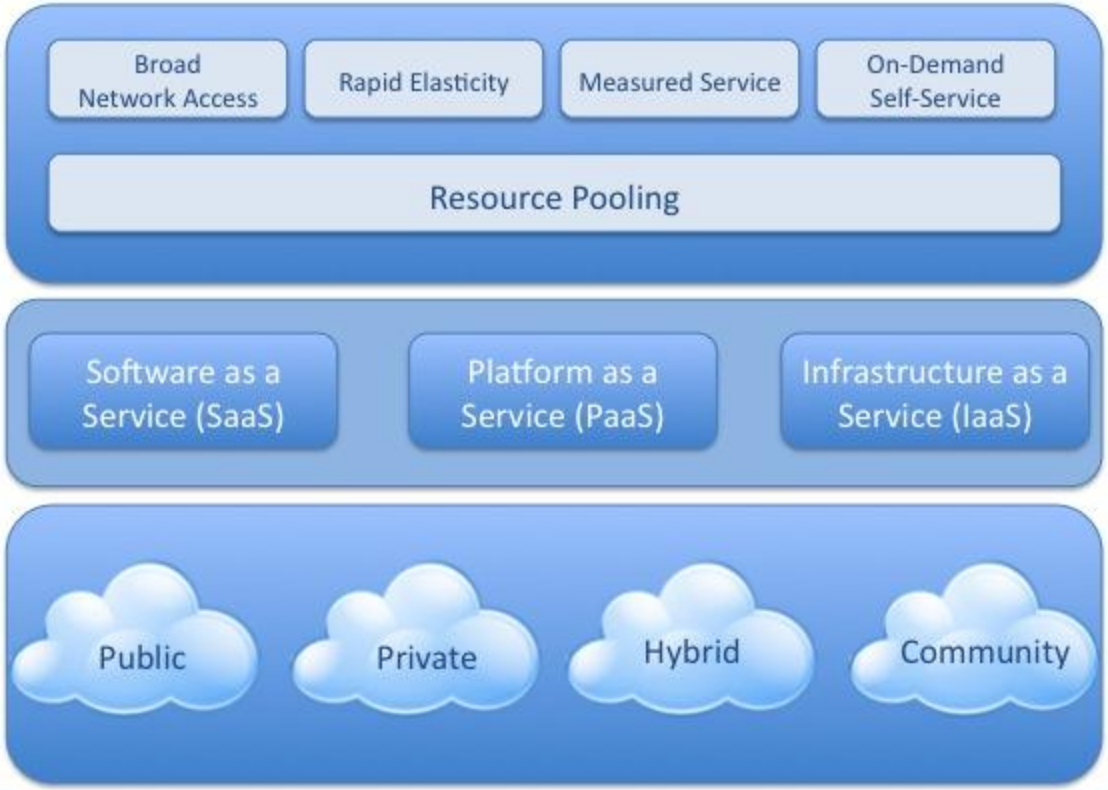




.. with all it's avatars..

Visual Model Of NIST Working Definition Of Cloud Computing

http://www.csrc.nist.gov/groups/SNS/cloud-computing/index.html

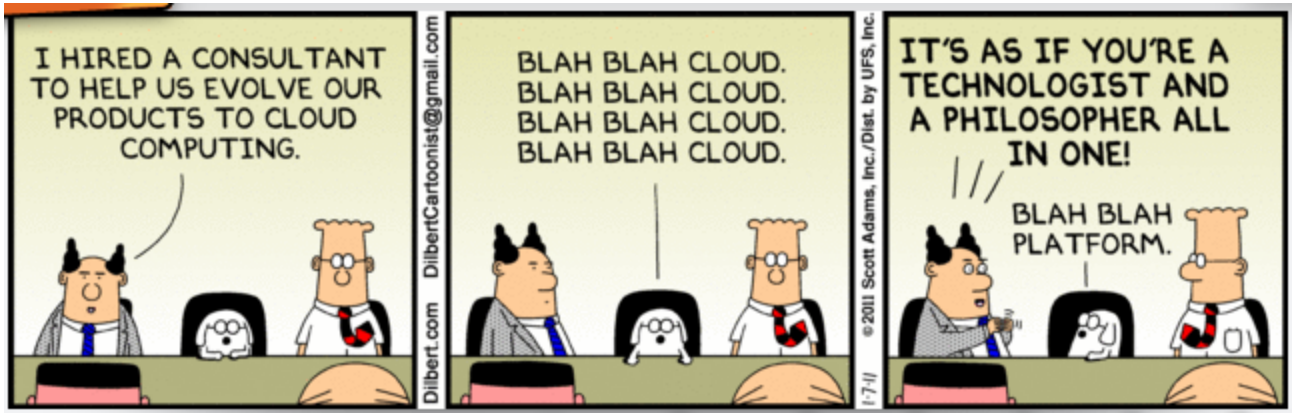


Essential Characteristics

Service Models

Deployment Models

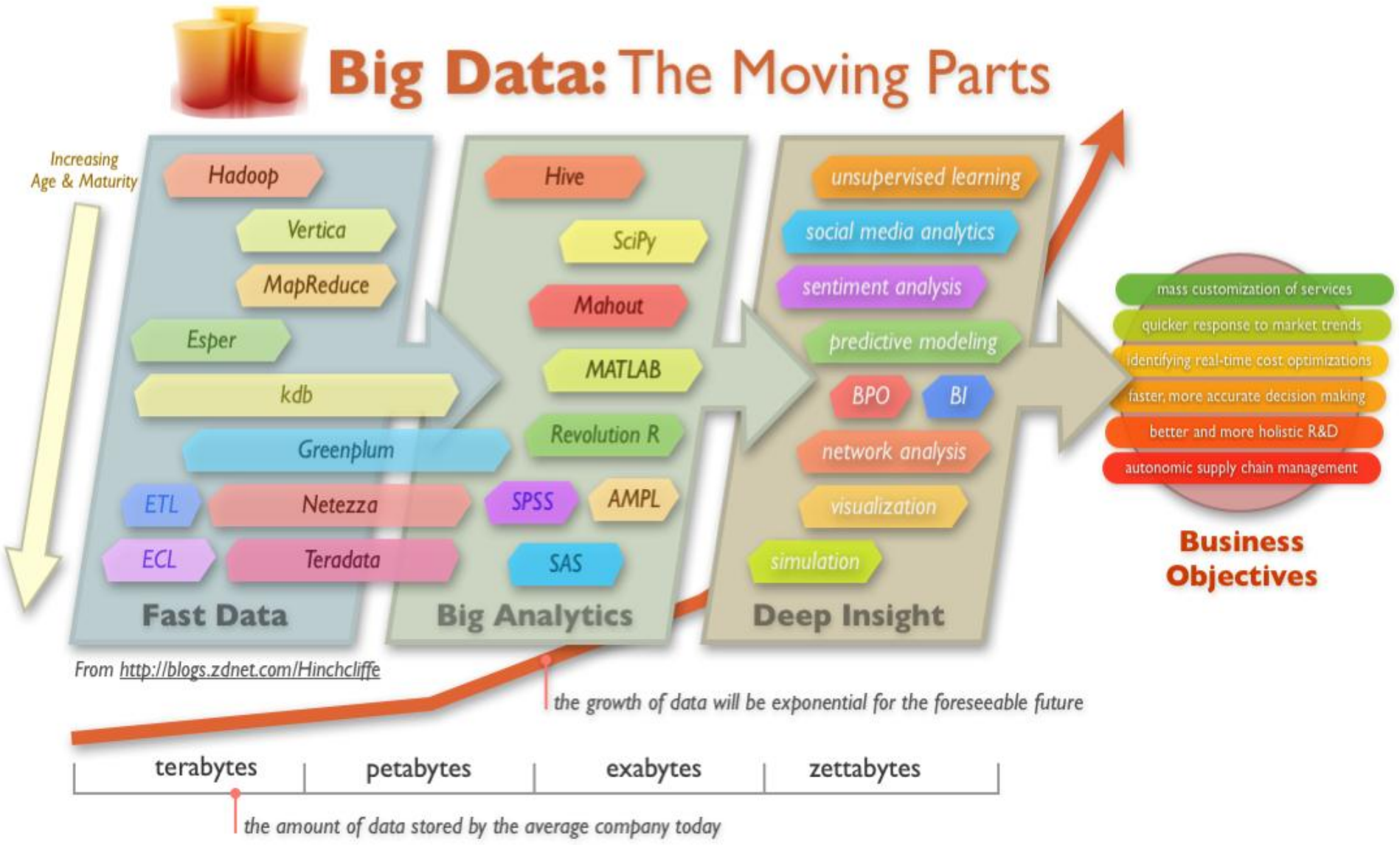
.. seems ‘so yesterday’!



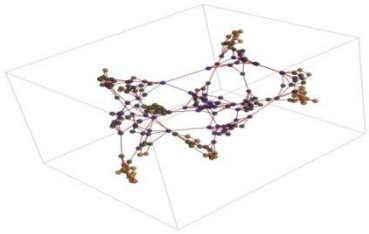
Jan 7, 2011

July 29, 2012

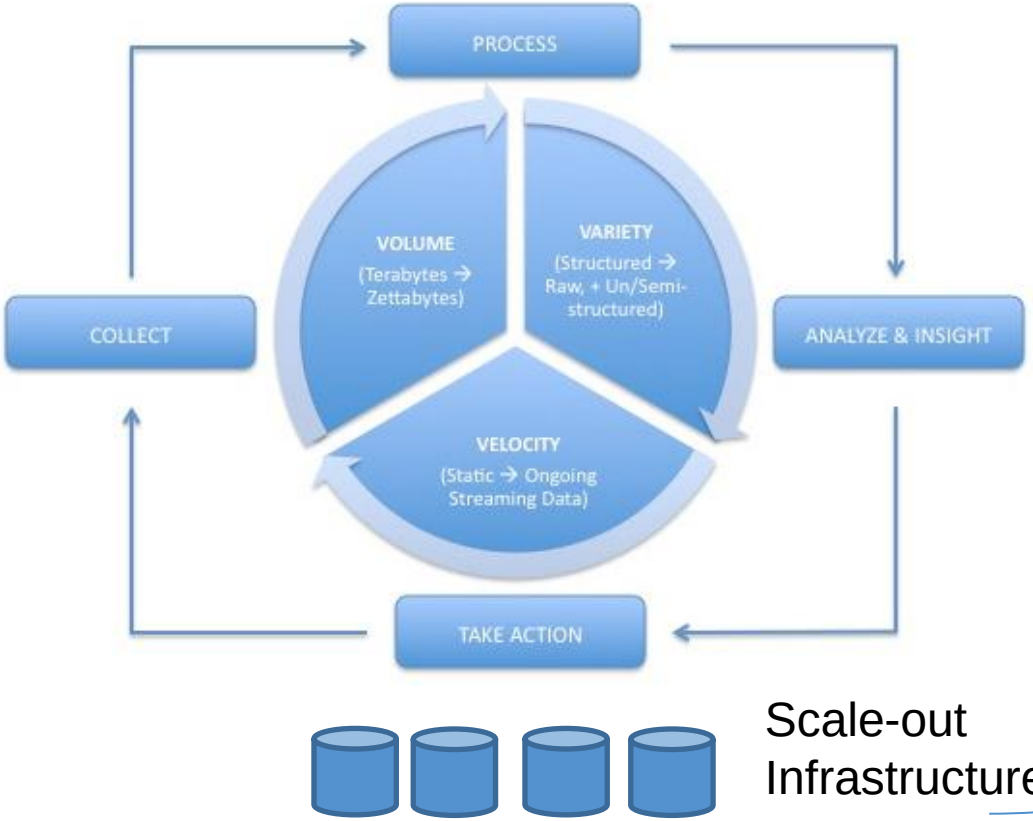
Meet the next Big Thingy..



Big Data: More than the 3 V's (Volume, Velocity, Variety)



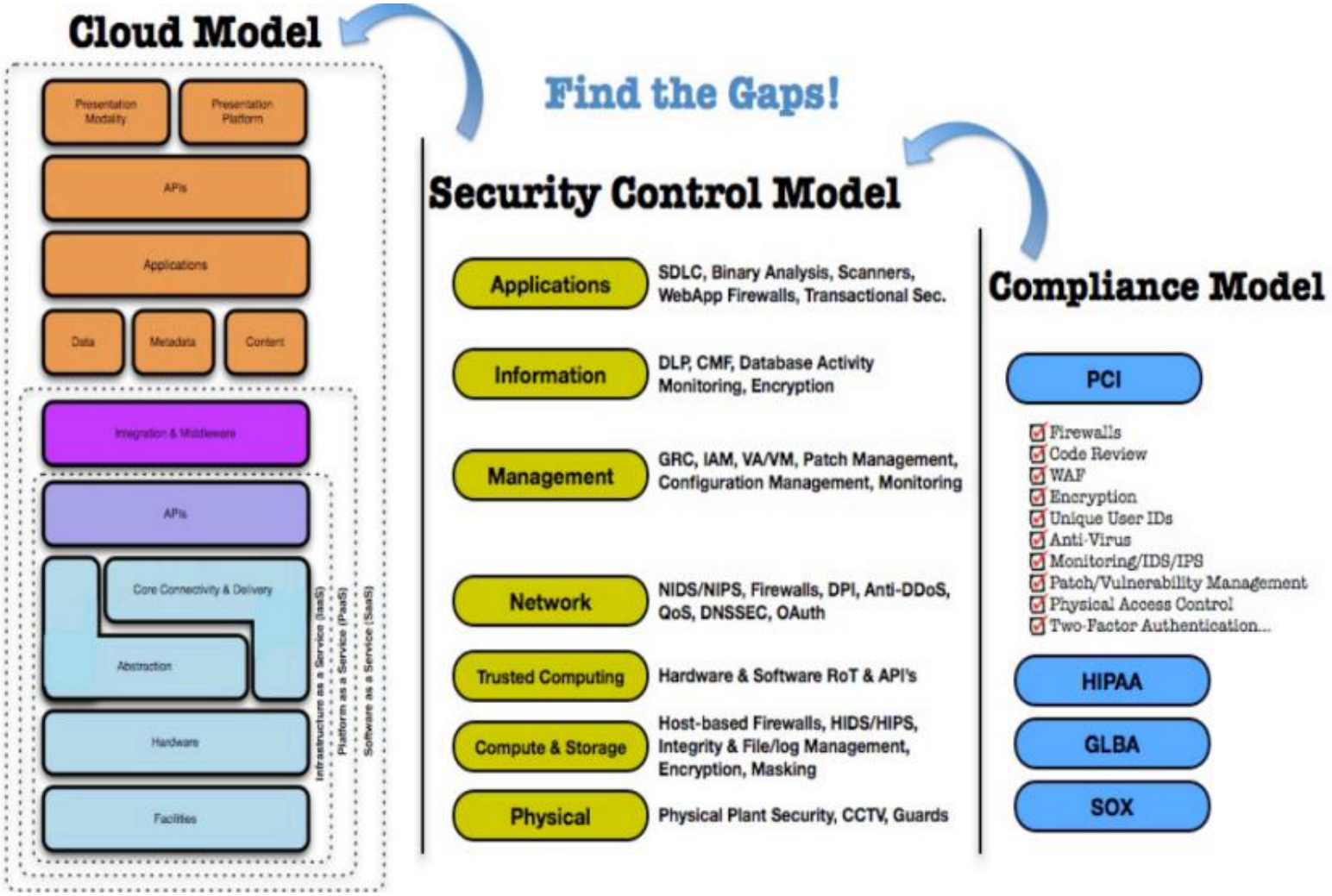
Analytics



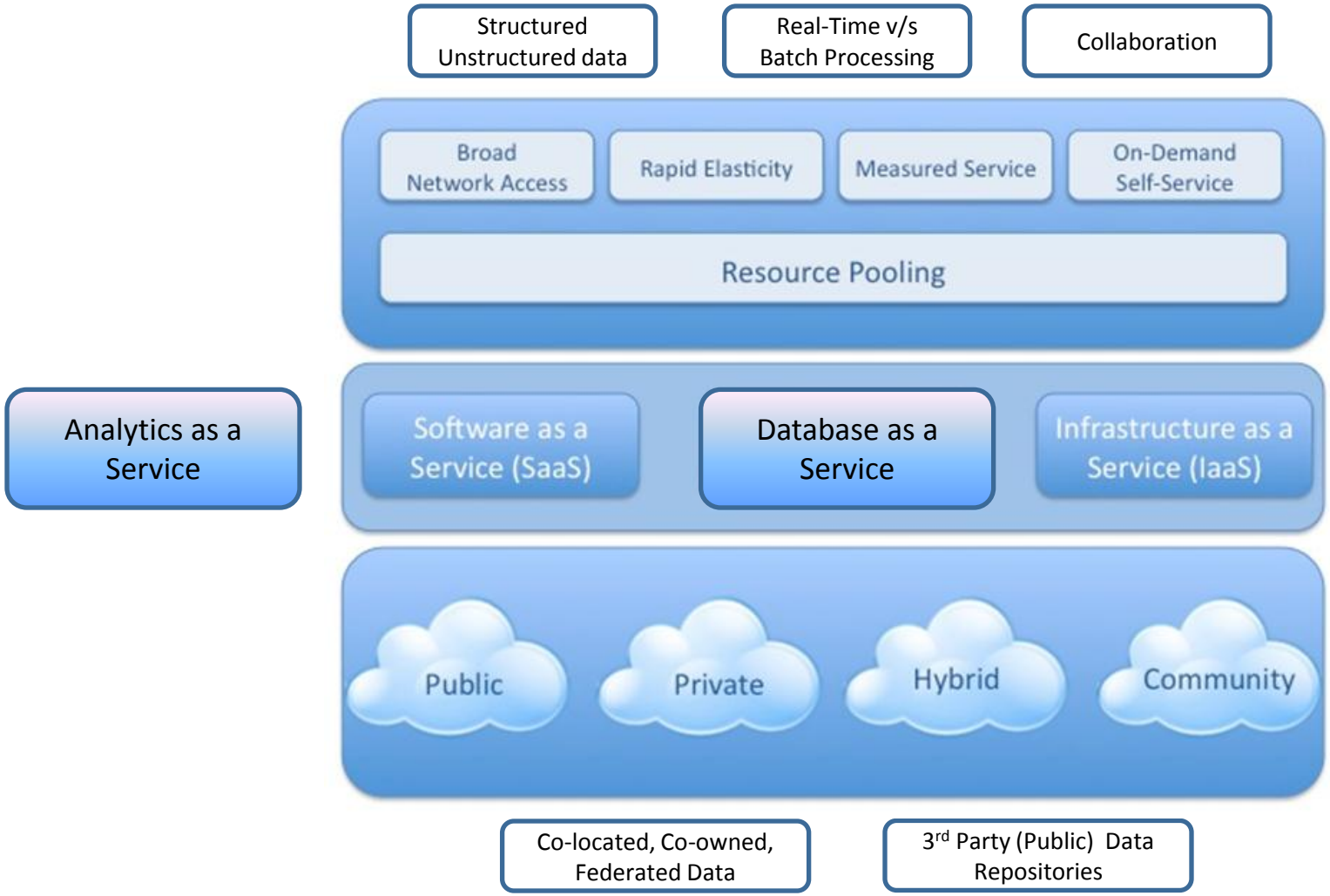
Scale-out Infrastructure

An Enterprise Setting

But wait.. Didn't we just do this for Cloud Computing?



Cloud Computing and Big Data - The Similarities



Big Data-specific capabilities

Big Data has similar computing requirements as Cloud Computing

Different modes of offering a Big Data processing capability

Big Data deployment strategies

Data Ownership, Location and Jurisdiction is crucial

Big Data Ecosystem - Assets & Concerns

Assets

- *Profile Data (user preferences)*
- *Behavioral Data (usage/consumption patterns)*
- *Characterizing information (endpoint identifiers)*

Concerns

- *Leakage of PII → Identity Theft & Loss of Privacy*
- *Malicious use of sensitive information*
- *Mis-profiling leading to wrong service personalization*
- *Lack of Control over Data Portability Management*
- *Very less (if any) stake in the data's monetization*

Assets

- *Data Analytics capability*
- *Inherent Data Semantics*

Derived Data

Needs and Concerns

- *Leakage of Analytics capability, resulting in loss of IP*
- *Leakage of PII via analytics, resulting in legal liabilities*
- *Corruption of Data, resulting in incorrect results*
- *Seamless access to rich and varied sources of Data*

Data Corpus

Assets

- *Large Corpus of Data – direct or supporting Business functions*
  - *Users/Employee, Intellectual Property, Business functions, Information Technology*

Concerns

- *Leakage and/or misuse of data, resulting in legal liabilities*
- *Leakage /Corruption of data, resulting in loss of business*

Assets

- *Data management infrastructure*
- *Data analytics infrastructure*

Concerns

- *Leakage and/or misuse of data, resulting in legal liabilities*
- *Corruption of data, resulting in loss of business*

Distributed Data Storage

Securing Big Data
Learning from Cloud Computing

Cloud Security Alliance – 13 Domains of Critical Focus

- Governance and Enterprise Risk Management
- Legal Issues
- Compliance and Audit
- Information Management and Data Security
- Portability and Interoperability
- Traditional Security, Business Continuity and Disaster Recovery
- Data Center Operations
- Incident Response, Notification and Remediation
- Application Security
- Encryption and Key Management
- Identity and Access Management
- Virtualization
- Security as a Service

Governance and Enterprise Risk Management

- Risk of correlating published enterprise data with public or 3rd Party data sources resulting in unintended de-anonymization
  - *Netflix data-sets de-anonymization*
- Risk due to information leakage via previously un-seen data patterns
  - Un-seen or novel data patterns may not have right access control policies
- Risk of insufficient protection of sensitive data at big-data scale
  - Insufficient or incorrect data classification is a problem even at non-big data scale
- Risk due to multi-tenant Big Data infrastructures

Compliance and Audit

- Need to define and publish how and what compliance policies are relevant to different data-sets
  - How do the policies change when working with federated data sources, or with derived data ?
- How are the compliance policies adapted and responsibilities shared across Data Owners, Managers and Consumers?
  - Need to identify and document via SLA, relevant compliance policy requirements on different stakeholders
- Audit to verify data sources used in big data analytics results
  - Can the consumer prove that they have not tapped into un-intended data sources to compromise data subject privacy?
- Audit to verify data analytics platform integrity
  - How can the consumer protect their analytics capability but still share semantics about the data sources?

Information and Data Security (1)

- What is your Big Data Security lifecycle?
  - Create, Store, Use, Share, Archive, Destroy(?)
- Where is Big Data located?
  - What are the different geographical jurisdiction policies?
  - What will be the jurisdiction on derived data, given your data is correlated with public information?
- Access Control?
  - Is static/role-based access control sufficient when dealing with un-structured and/or novel data?
  - Do Content-based Access control policies scale for you?
  - How will your Access Control policies adapt with data transformations?

Information and Data Security (2)

- What is the right scale-out Data Protection technique?
  - Encryption ?
    - At a Storage Volume-level?, Database-level?, Object-level?
    - Scale-out encryption ?
    - Key Management with Owners, Managers and Consumers?
  - Tokenization?
  - Information Dispersion
    - Protection via fragmentation
    - Share fragmentation mapping with validated consumers?
  - Mix of the above?
- How to do scale-out Data Leakage Prevention?
  - Database activity monitoring?
  - Validated and pre-approved Data Consumer Analytics?
    - Leverage secure containerization techniques?

Application Security

- Analytics Protection
  - Protection for Data Consumers against leakage of internal analytics semantics?
  - Protection for Data Consumers against cross-stack attacks in a multi-tenant Big Data architecture
- Protection against rogue Analytics
  - Secure Containerization to protect against malicious analytics
  - Analytics Model Validation against malicious behavior
- Authorization and Access Control
  - Scalable and Granular Access Control over data

Unique Challenges in Security for Big Data

- Defense against un-intended usages of Big Data
  - Existing diversity and anonymization techniques not sufficient
  - Need visibility, accountability and actionable insight into usage
- Scale-out approach for embedding Security and Trust primitives within Big Data
  - Ability to adapt known and tested security primitives to Big Data's scale-out nature
- Need for secure collaboration across Data Scientists
  - Multi-Tenancy cannot be solved via just secure containerization
  - Big Data inertia requires secure analytics exchange channels

Securing Big Data
New Technology Directions

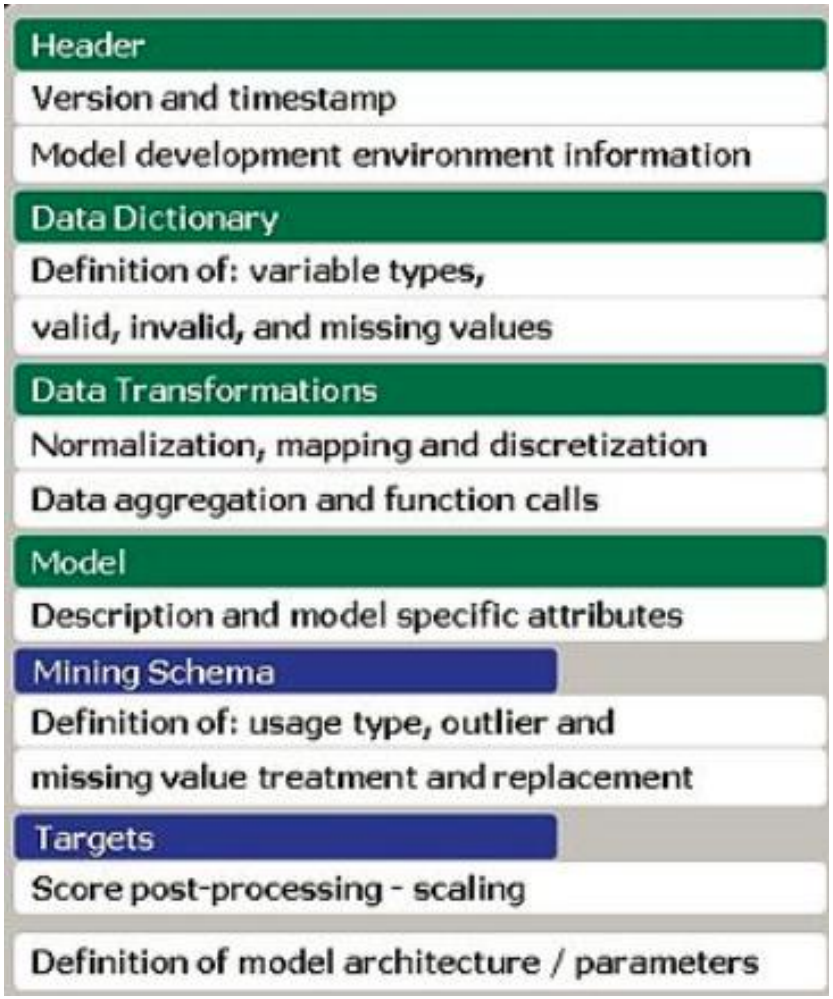
1. Defense against un-intended usages of Big Data

- Support Auditing via Analytics Meta-Data Description
  - Describe Analytics Application
    - Inputs
    - Outputs
    - Application behavior
    - Data Normalization, Transformation, Mining Models etc
  - Meta-Data enables applications to publish behavior in a privacy-aware manner
  - Can be consumed by auditor to validate published application behavior against intended usage
  - Can be used for typed filtering within the Data Managers platform against malicious Data Consumers

Standardized Representation of mining models and data

- Predictive Modeling Markup Language (PMML)
- Encompasses the various stages in a typical data-mining/analytics task
  - Data Dictionary definition
  - Data Transformations
  - Handling missing or outlier data values
  - Model Definition
  - Outputs
  - Post-Processing steps
  - Model Explanation
  - Model Verification
- Supported by leading Data analytics tools vendors (commercial and open-source likewise)

PMML – Existing and Proposed Extensions



- Proposed Extensions
  - Allow incomplete data and mining models for privacy reasons
  - Allow wild-carded/pattern-matched data-model and mining-model representations
    - Enabling easier audit checks
  - Enable versioning of the shared data and mining-model
    - Enable sharing of analytics updates over time
  - Allow Model Filter templates

2. Scale-out approach for Security & Trust primitives

- Need to adapt the Big Data processing framework for embedding security primitives
- Needs to work at the scale of Big Data
- Needs to follow the Big Data programming model and data flow
- Needs to be extensible

- Very similar to Operating System “hooks” for embedding security processing!

Proposed Introspection Framework for Map Reduce

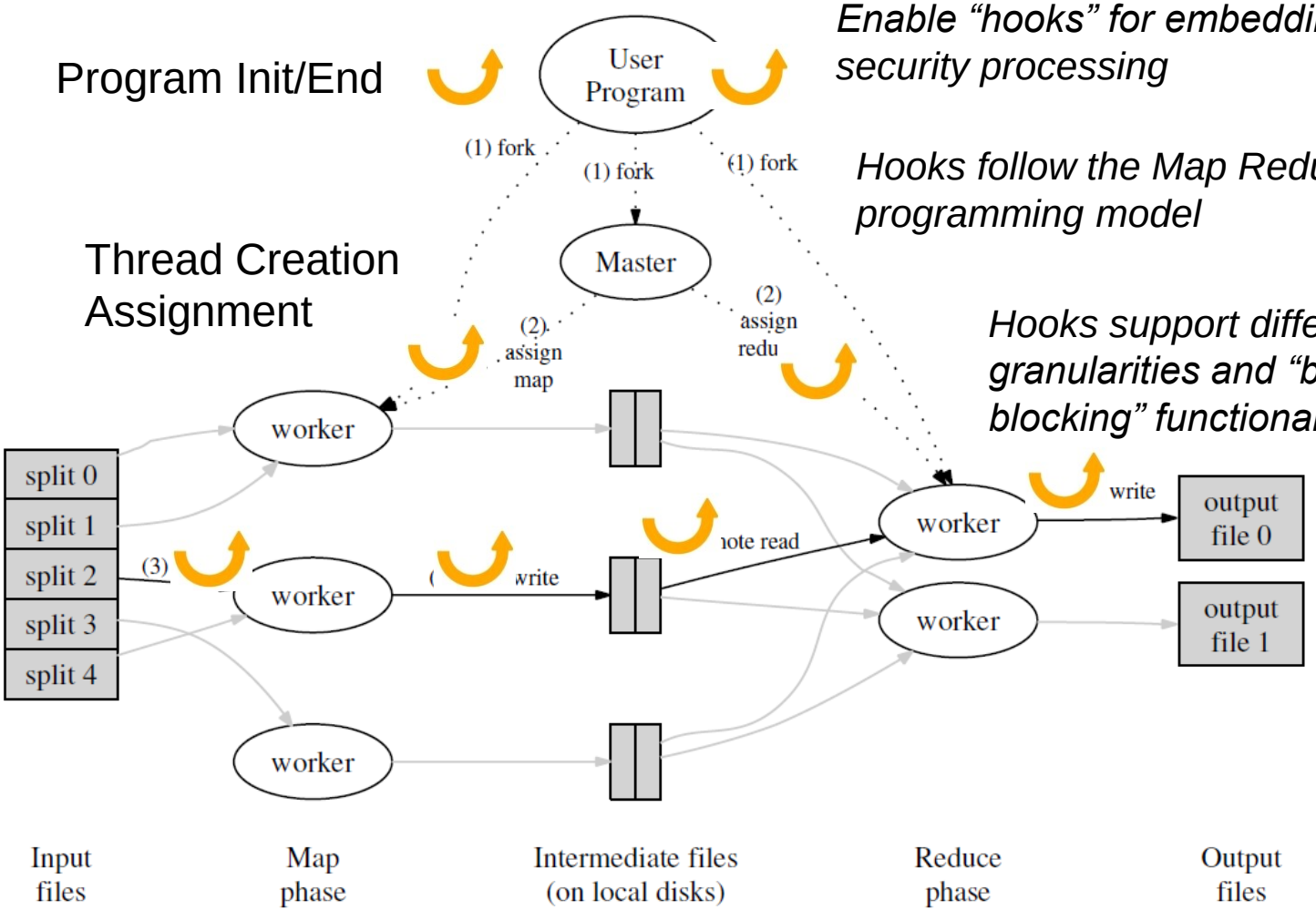
Program Init/End

Thread Creation Assignment

Enable "hooks" for embedding security processing

Hooks follow the Map Reduce programming model

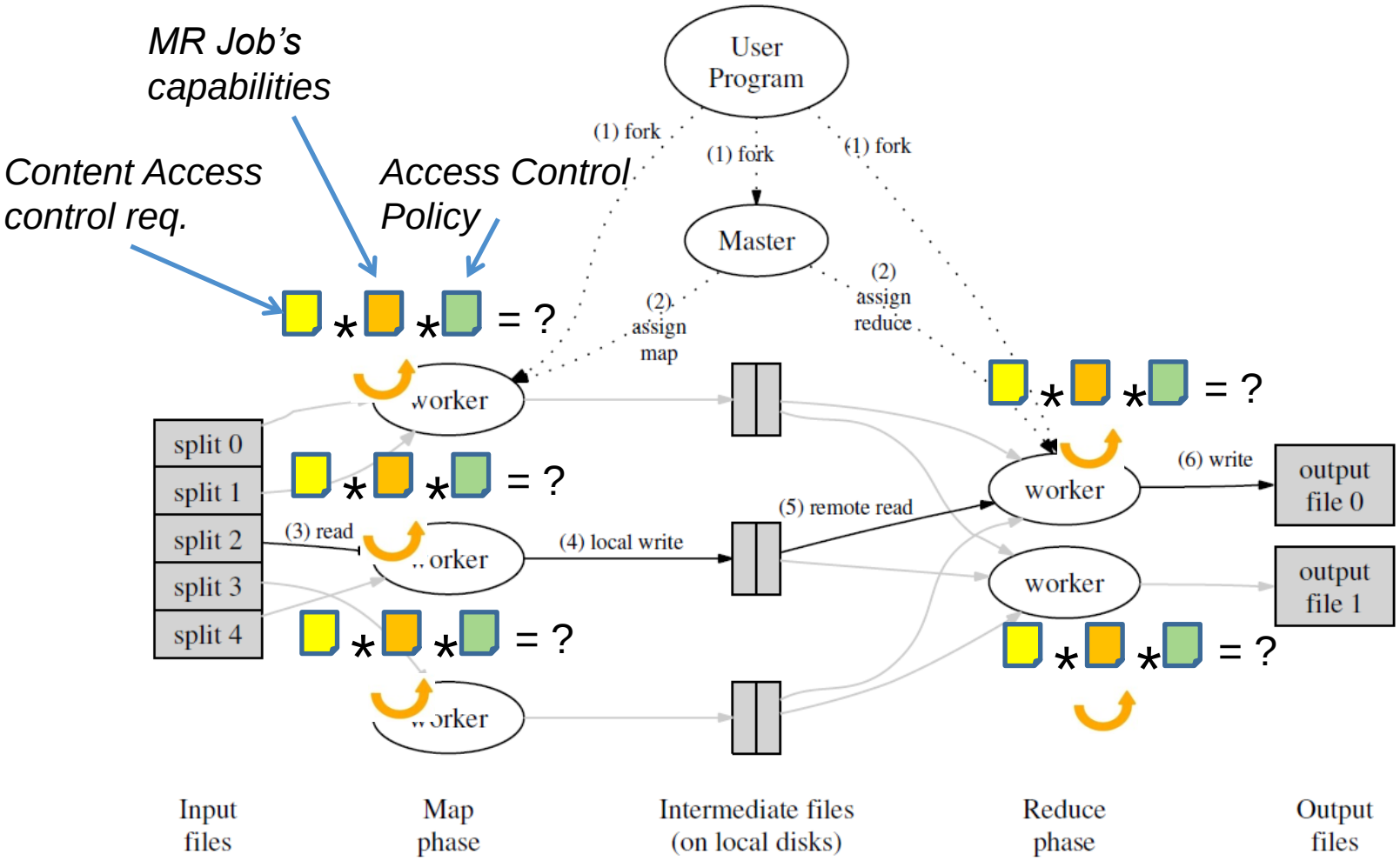
Hooks support different granularities and "blocking"/"non-blocking" functionality



Scale-out Dynamic Access Control

- Policy Delivery
  - Global policy broken into policy fragments and delivered to decision points (Map/Reduce tasks)
- Capability Delivery
  - MR Job's capability delivery to decision points
- Access Control attributes
  - Content-based attributes identified based on data-type and value
- Execution
  - Hook into Map-Reduce programming to execute decision points
- Output
  - Policy enforcement via results modification
  - Identifying attributes for output data (inputs to next round of access control decisions)

Scale-Out Access Control Policy Decision



1 + 2: Analytics and Data Audit Capability

- Audit analytics application's meta-data about inputs, outputs and analytics behavior
  - Published and verified meta-data can be used to customize the approved data path in/out of the application
- Use Introspection to hook into application's data path
  - Introspected data-path can be used to define the applications meta-data
- Correlate expected v/s actual data path for compliance

Security for Big Data, Defending against Big Data
How to apply our discussions?

PII in the face of Big Data

- Privacy implications of Big Data
  - Between 63% and 87% of the US population is re-identifiable based on widely available demographics
    - Benitez, K, Malin B., J Am Med Inform Assoc 2010; 17: 169-177
    - Sweeney L. Uniqueness of simple demographics in the U.S. population Working paper LIDAP-WP4. Pittsburgh, PA: Data Privacy Lab, Carnegie Mellon University, 2000.
    - Golle P. Revisiting the uniqueness of simple demographics in the US population. In: Proc 5th ACM Workshop on Privacy in Electronic Society. 2006:77-80.
- The breaches are coming!
  - Over 250M records (Non-Hacking or Non-Credit Card) breached over 7 years (http://www.privacyrights.org/)
  - Almost all are “Non-Big Data” scenarios (Access-Controlled DBs, Legacy security monitoring techniques, Regulatory Audits, Required Compliance etc.)
    - What about Big-Data enabled providers?

Step 1 – Protect your Data

- Establish, Document and Enforce a Big Data Security Lifecycle
- Understand where your data lives and moves
- Ask for scale-out and granular authentication and dynamic access control
- Identify the best Data Protection strategy
  - Encryption – Scale-out Challenges?
  - Dispersion – Hiding data in Big Data
  - Data Leakage Prevention
- Adopt learning from your Cloud security program drawing analogies to Big Data deployment

Step 2 – Audit your systems

- Auditing the Platform
  - Ask for and enforce platform hardening guidelines
  - Ask for Data movement patterns – Leverage Big Data Analytics to find anomalies!
- Auditing Data Consumers
  - Ask for Data Model and Mining Model representations. Establish a chain of custody
  - Verify against Data Inputs and Outputs types for discrepancies
- Continuous auditing of suitability/need of published Data streams
- Establish a Remediation procedure
  - What happens when in the case of un-intended data leakage
  - How quickly can the data stream be disabled without harming functionality?

Step 3 – Make your Data Resilient

- Investigate in the right tokenization and redaction strategy
  - Scale-out Tokenization can leverage proposed introspection framework
- Is application-specific tokenization required?
- Investigate into Form Preserving Encryption techniques or simpler alternatives
- Do you need Active Correlation (a la “Active Defense” from Incident Response) to identify data leakage?

Thank You and Questions?

samir.saklikar@rsa.com