Security and Privacy in a Big Data World

Dr. Flavio Villanustre, CISSP, LexisNexis Risk Solutions
VP of Information Security & Lead for the HPCC Systems open source initiative

28 January 2013

LexisNexis®

Risk Solutions

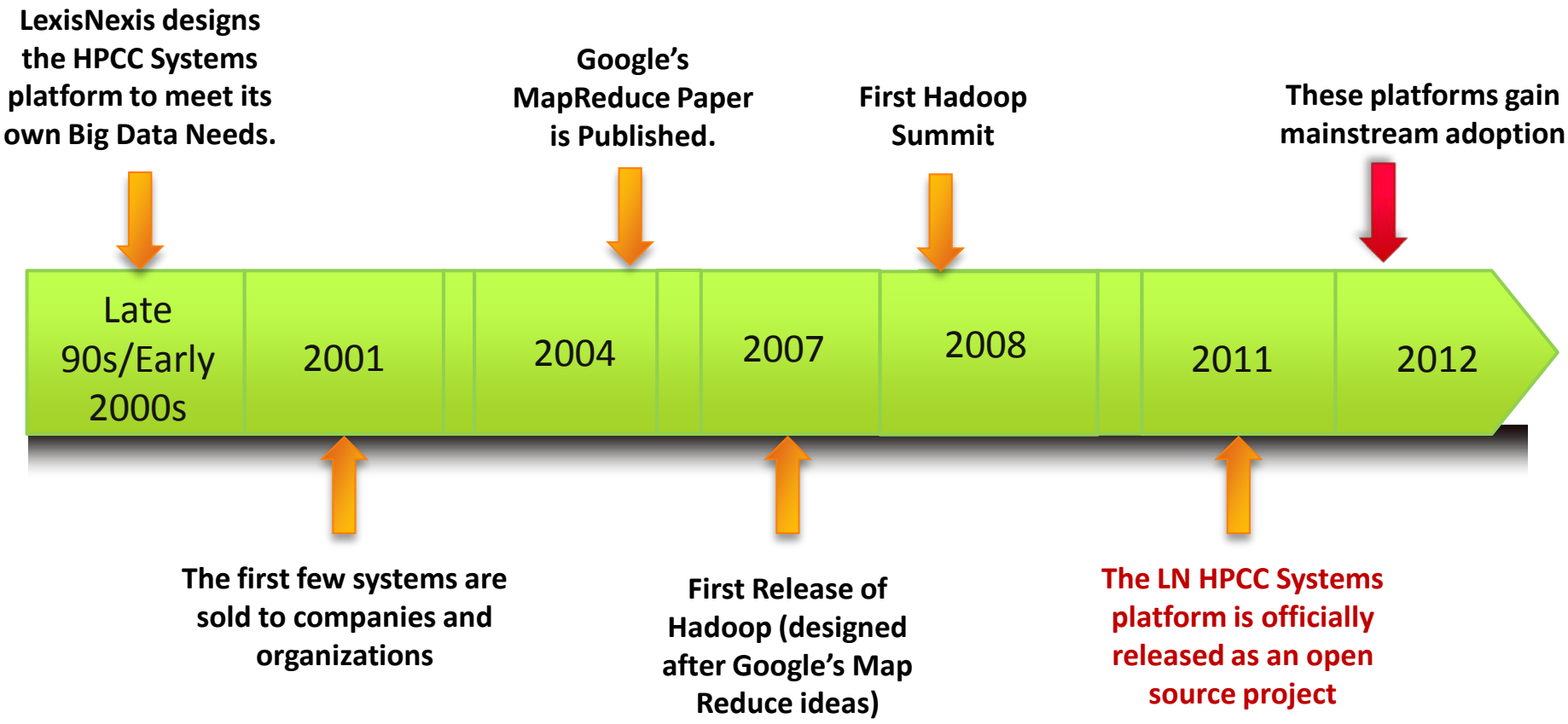
But what is Big Data?

- Gartner told us that it's defined by its four dimensions:
  - **Volume**
  - **Velocity**
  - **Variety**
  - **Complexity**
- Driven by the proliferation of social media, sensors, the Internet of Things and the such (a lot of the latter)
- Became accessible thanks to open source distributed data intensive platforms (for example, the open source LexisNexis HPCC Systems platform and Hadoop)
- Made popular by consumer oriented services such as recommendation systems and search engines

Big Data platforms: key design principles

- Distributed local store
  - Move algorithm to the data – exploit locality
- Many data problems are embarrassingly parallel
  - Leverage massive resource aggregation:
    - Thousands of execution cores
    - Hundreds of disk controllers
    - Hundreds of network interfaces
    - Terabytes of memory and massive memory bandwidth
- Storage is cheap
- Moving data into the system takes time (hence keep the data around, if possible)
- It's fine (and encouraged) to perform iterative exploration
- In the end: It's just and all about the data

A timeline of the main open source Big Data platforms



Just when we thought that we knew data security…

Big Data is not your dad's data

a. More data sources (beware! data + data > 2 * information)
b. Boiling the ocean is at the reach of your hand
c. Public clouds offer scalability but introduce risks
d. Distributed data stores can blur boundaries
e. Leveraging diverse skills implies more people accessing the data

Old tricks may not work

a. Tokenization only goes so far
b. Encryption at rest just protects against misplaced hardware
c. Tracking data provenance is hard
d. Enforcing data access controls can quickly get unwieldy
e. Conveying policy information across multiple systems is hard

The ever present challenges

- Security
  - Keeping the bad guys out
  - Making sure the good guys are good and stay good
  - Preventing mistakes
  - Disposing of unnecessary/expired data
  - Enforcing “least privilege” and “need to know” basis
- Privacy
  - Statistics safer than aggregates
  - Aggregates safer than tokenized samples
  - Tokenized samples safer than individuals
  - Don’t underestimate the power of de-anonymization
  - Mistakes in privacy are irreversible
- Security <> Privacy

Be wary of “tokenization in a box”

- Tokenized dataset * Tokenized dataset ~= identifiable data
- The problem of eliminating inference is NP-complete
- Several examples in the last decade:
  - The “Netflix case”
  - The “Hospital discharge data case”
  - The “AOL case”

Common sense to the rescue

- Track data provenance, permissible use and expiration through metadata (data labels and RCM)
- Enforce fine granular access controls through code/data encapsulation (data access wrappers)
- Utilize homogeneous platforms that allow for end-to-end policy preservation and enforcement
- Deploy (and properly configure) Network and host based Data Loss Prevention
- A comprehensive data governance process is king
- Use proven controls (Homomorphic encryption and PIR are, so far, only theoretical concepts)

And always keep in mind that…

- Access to the hardware ~= Access to the data
  - Encryption at rest only mitigates the risk for the isolated hard drive, but NOT if the decryption key goes with it
  - Compromise of a running system is NOT mitigated by encryption of data at rest
- Virtualization may increase efficiency, but…
  - In virtual environments, s/he who has access to your VMM/Hypervisor, also has access to your data
  - Cross-VM side-channel attacks are not just theoretical

Remember

- Data exhibits its own form of **quantum entanglement**: once a copy is compromised, all other copies instantly are
- **The closer to the source** the [filtering, grouping, tokenization] is applied, **the lower the risk**
- **YCLWYDH!** You can't lose what you don't have (data destruction)

Useful Links

- LexisNexis Risk Solutions: http://lexisnexis.com/risk
- LexisNexis Open Source HPCC Systems Platform: http://hpccsystems.com
- Cross-VM side-channel attacks: http://blog.cryptographyengineering.com/2012/10/attack-of-week-cross-vm-timing-attacks.html
- K-Anonymity: a model for protecting privacy: http://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity.pdf
- Robust de-anonymization of Large Sparse Datasets (the Netflix case): http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf
- Broken Promises of Privacy: Responding to the surprising failure of anonymization: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006
- Tamper detection and Relationship Context Metadata: http://blogs.gartner.com/ian-glazer/2011/08/19/follow-up-from-catalyst-2011-tamper-detection-and-relationship-context-metadata/

Questions?

Email: info@hpccsystems.com