淺談大數據平台安全現況

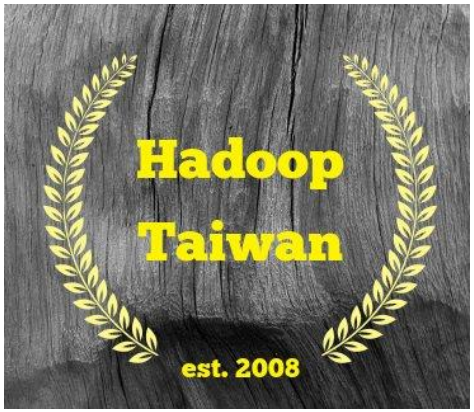
Introduction to Big Data Platform Security

Jazz Yao Tsung Wang

<http://about.me/jazzwang>

About Me

- 王耀聰 **Jazz Wang**
- 資安門外漢 /交大電控碩士
- 前 **Etu Manager** 產品協理
- 現任 **TenMax Data Architect**
- **Hadoop.TW** 共同創辦人
- **HadoopCon** 社群年會總召
- **Hadoop The Definitive Guide** 譯者
- **Hadoop Operations** 譯者
- 自由軟體愛好者 / 推廣者 / 開發者
- http://about.me/jazzwang - slideshare, github, etc.

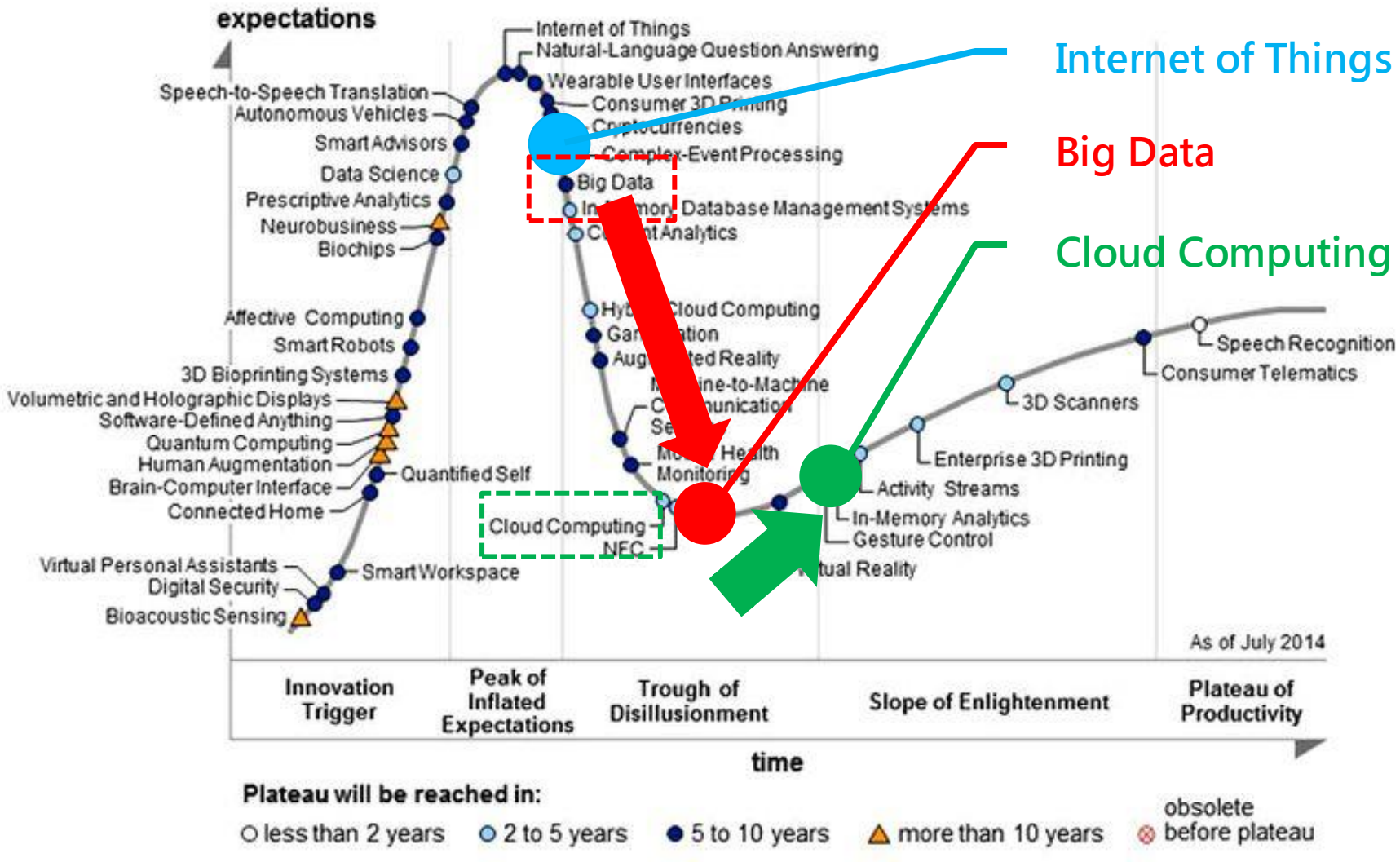


Agenda

- **企業導入大數據的四個階段**
  - 專案規劃、大數據平台建置、**大數據平台資安**、大數據品質管制
- **大數據平台資安範疇與現況**
  - 高可用性 High Availability (HA)
  - 災害復原 Disaster Recovery (DR)
  - 身分認證 Authentication
  - 權限控管 Authorization
  - 存取稽核 Auditing (Accounting)
  - 加密防護 Encryption
- **結語**
  - 題外話：開放原始碼軟體的資安掃描

Gartner Hype Cycle 2014

萌芽期　夢幻期　幻滅期　平原期　高原期



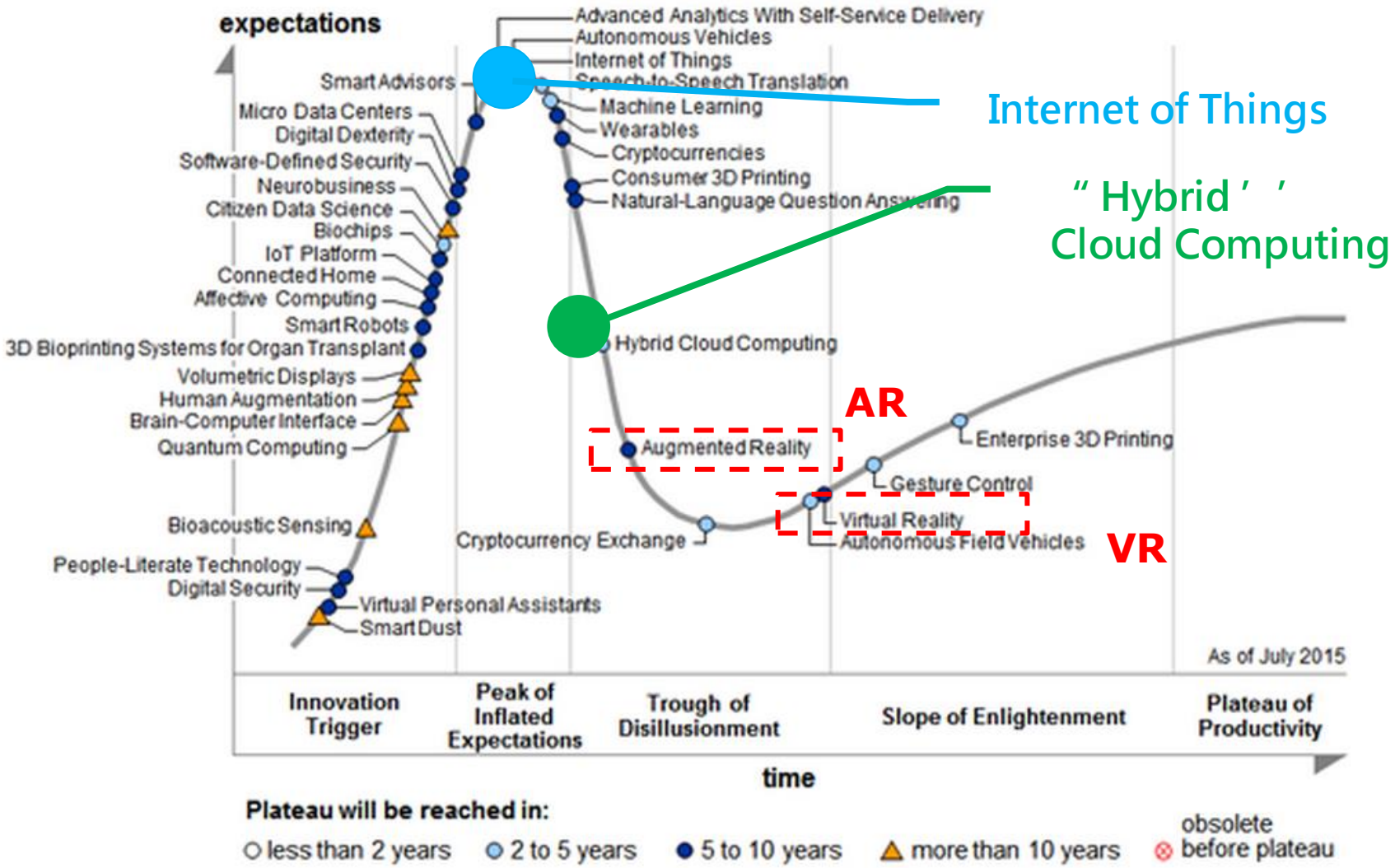
Internet of Things

Big Data

Cloud Computing

Gartner Hype Cycle 2015

萌芽期 夢幻期 幻滅期 平原期 高原期



Big Data ~~退燒~~畢業了!!

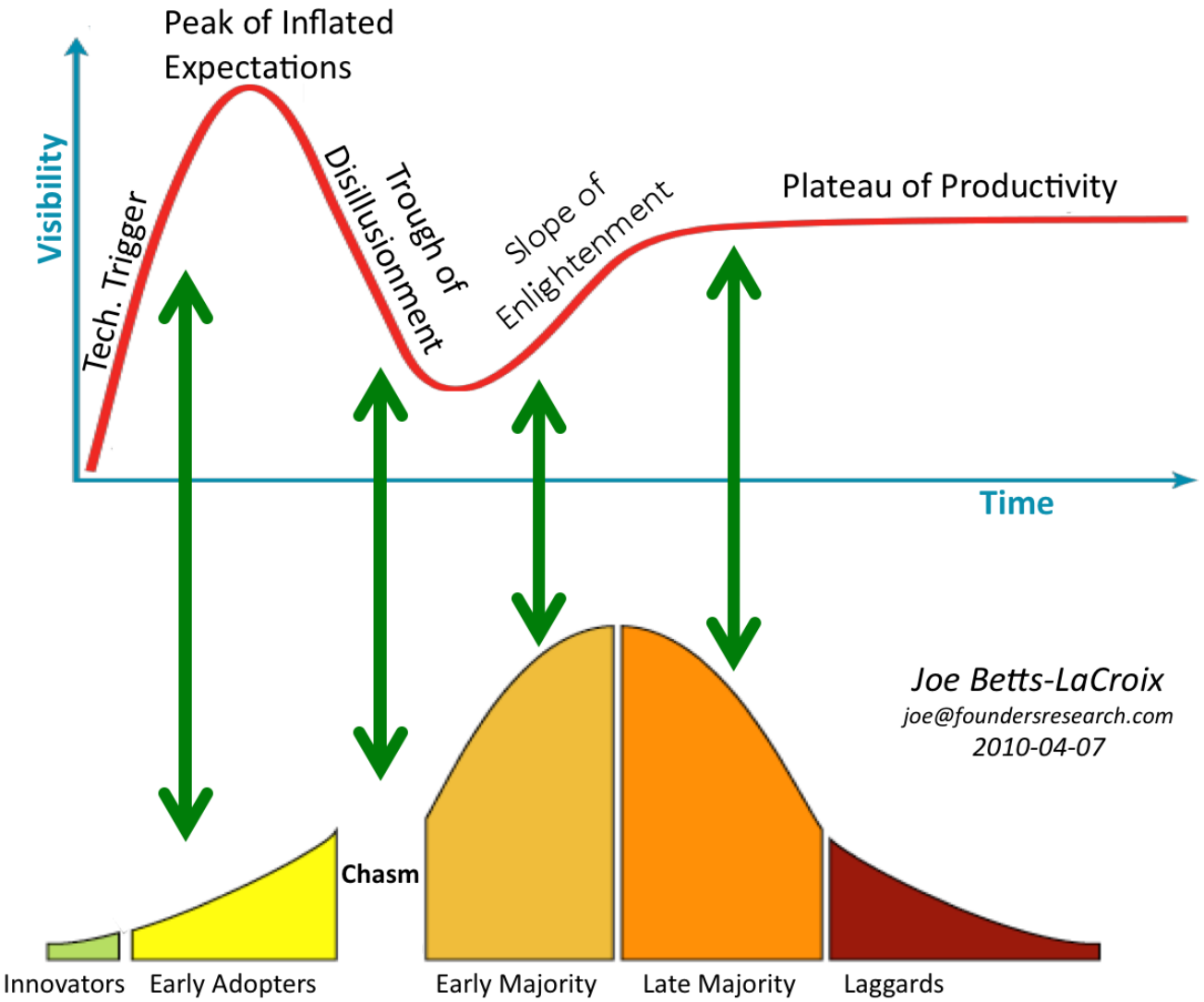
隱身進入以下領域：

- Internet of Things (物聯網)
- Business Intelligence and Analytics (商業智慧)
- Enterprise Architecture
- Web-Scale IT
- Digital Banking Transformation (數位金融)
- Utility Industry IT
- CRM Customer Service and Customer Engagement
- CRM Marketing Applications
- Digital Commerce (電子商務)

導入 Big Data 到底是想要？還是需要？

國際金價	**提供給客戶的價值**	產品通路	1
開採成本	總擁有成本	軟硬體投資	6
提煉廠	分析平台與工具軟體	SMAQ	5
含金度	資料鑑價？	**商業模式**	4
開採權	分析資料的合法性	個資法	3
金礦	資料集	Open Data	2

2016 年剛好是"跨越鴻溝"的時間點



企業導入 Big Data 的四個階段

今天談大數據平台資安只是
預告後面的路還很崎嶇.....

台灣只有極少數
創新者在這個位置
Innovators

台灣只有少數
先行者在這個位置
Early Adopters

台灣開始有一些追隨者在這個位置
Early Majority
但往往問題是“剛開始蒐集數據”或
“剛開始思考如何讓數據產生價值”

Big Data 專案規劃的六頂思考帽

專案規劃

- 問題一：組織想要解決什麼商業問題 ?? (Value)
  可以用資料解決嗎 ?? (降低成本 or 增加收益)
- 問題二：這些資料哪些是內部資料 ?? 哪些是外部資料??
  該如何獲得 ?? 有哪些型態 ?? (Variety)
- 問題三：分析這些資料是否合乎法規需求 ??
  有無需要事先聲明的保護條款 ?? (Legality)
- 問題四：驗證答案真的在這堆資料裡 ?? 資料是否可靠 ??
  需要多少資料才能找到答案 ?? (Volume , Veracity)
- 問題五：挑選合理的資料處理/分析平台 – 人、流程、技術
  定義多快找到答案才能解決商業問題 (Velocity)
- 問題六：定義效益評量指標 (怎麼算 ROI ?? 或 KPI 是什麼 ??)
  持續改善的時程藍圖 (Validation , Roadmap)

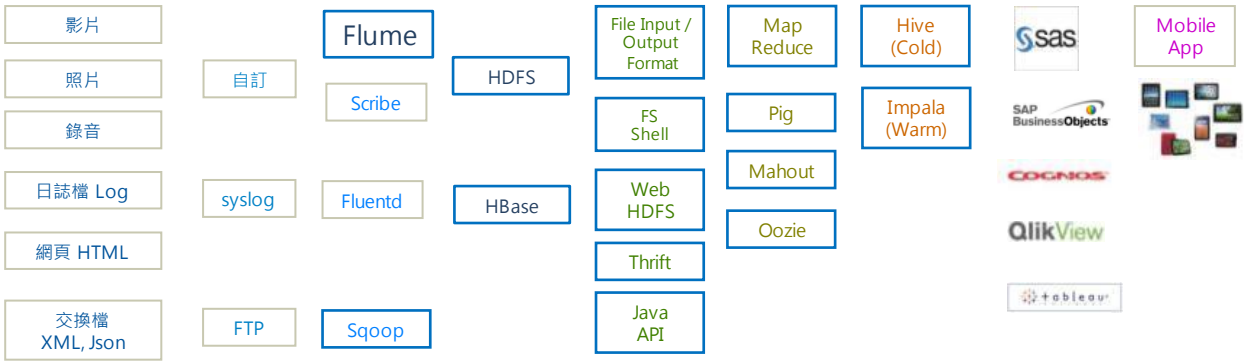
大數據平台建置的三個面向

企業內部的人力資源盤點 People 1

處理巨量資料的常見流程 Process 2

生	流	蒐	存	取	算	析	用	看	變
資料源	網路協定	前處理	儲存方式	存取方式	資料處理	資料分析	視覺化	解讀	行動

處理巨量資料的技術盤點 Technology 3



如果您對以上內容有興趣...

http://www.ithome.com.tw/news/101577

Agenda

大數據
平台資安

- 企業導入大數據的四個階段
  - 專案規劃、大數據平台建置、**大數據平台資安**、大數據品質管制
- **大數據平台資安範疇與現況**
  - 高可用性 High Availability (HA)
  - 災害復原 Disaster Recovery (DR)
  - 身分認證 Authentication
  - 權限控管 Authorization
  - 存取稽核 Auditing (Accounting)
  - 加密防護 Encryption
- **結語**
  - 題外話：開放原始碼軟體的資安掃描

高可用性 High Availability (HA)

- 架一座 Big Data Platform 的叢集，其實同時買了很多不同功能的元件！
- 國際大廠對於各元件高可用性的支援還持續隨著時間，正在慢慢增加中

		CDH 4.7	CDH 5.2	CDH 5.3	CDH 5.4	CDH 5.7
	文件日期	2014/12	2015/9	2015/10	2015/11	2016/06
管理者介面	Cloudera Manager				V	V
金鑰管理	Key Trustee KMS					V
稽核者介面	Cloudera Navigator Key Trustee Server				V	V
使用者介面	Hue				V	V
查詢引擎	Llama / Impala		V	V	V	沒寫?
ODBC 接口	HiveServer2					V
Schema	Hive Metastore		V	V	V	V
工作流程	Oozie		V	V	V	V
索引引擎	Solr (Search)		V	V	V	V
運算引擎	MRv1 / YARN	V	V	V	V	V
快速查表	HBase		V	V	V	V
儲存層	HDFS	V	V	V	V	V

高可用性 High Availability (HA)

- **現狀：**
  - 大數據平台的賣點是划算--分散儲存、分散運算、平行查詢一次購足
  - 缺點是潛藏的維運成本 -- 請不要過度期待高可用性的支援是完整的！
    (10 歲的童工 vs 38 歲的老員工，成本不同，強項不同，互補非取代)
  - 分散式系統難解的耦合性：
    如果您想要支援 AAA 與 Encryption 就會隨之增加高可用性的挑戰!
- **建議：**
  - 麻煩先根據組織的需求，由需求往回推估最小功能元件集合
  - 再根據最小功能元件集合，逐一驗證每個元件的高可用性支援程度
  - 寧可分階段依商務問題的急迫性，逐一增加元件的複雜度；
    千萬別想一次到位，所有功能元件都想馬上用得上。

注意!! 並不是所有角色都可以跑兩個

- **以下角色，一座叢集只能跑一個！**
  - HDFS Balancer
  - YARN JobHistory Server
  - Impala StateStored
  - Impala Catalog Server
  - Spark History Server
- **真的不幸那台掛掉，只好手動進管理介面改派給別台!!（或者寫好自動化隔離腳本靠 API 來達成）**

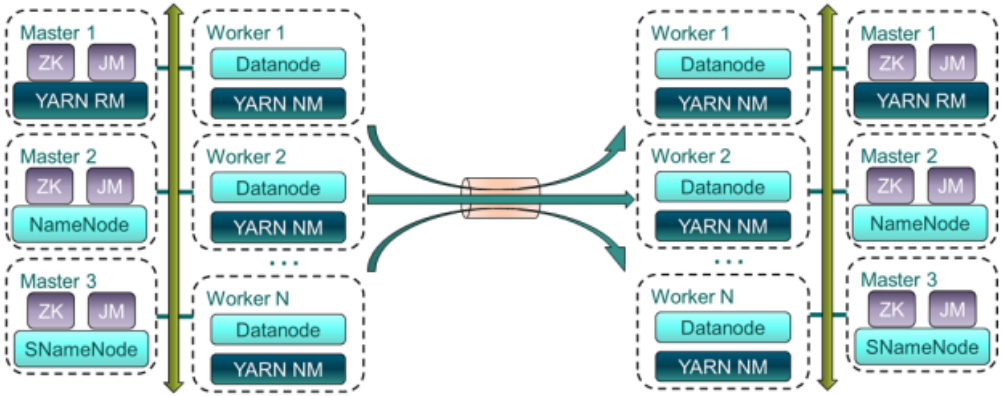
（謎之音：前提是管理介面還活著！或許這就是為何基本軟體授權最低台數從 5 台漲到 10 台，就是要把服務拆散）

災害復原 Disaster Recovery (DR)

- **高可用性是用在臨時有一台機器暫時故障(還有救)的時候**
- **災害復原是用在臨時有一台機器完全救不回來的時候**
- **現況：複雜的 State Machine !!**
  - HDFS 靠 Journal Node 所以可以從另外兩台救回
  - MRv1/YARN 因為狀態存在 HDFS 所以裝一台新的也沒關係
  - HBase 因為狀態存在 HDFS 所以重裝一台也沒關係
  - 但其他的呢?? Hive Metastore 背後的資料庫 ??
    Cloudera Manager 背後的資料庫?? Oozie 的 Metastore ??
- **建議:**
  - 有關聯式資料庫的地方儘量維持兩台 Active-Standby 或 A-A 副本
  - 裝機後至少做一次全系統備份（把當時的狀態存起來）
  - 行有餘力，別忘了做遞增備份

災害復原 Disaster Recovery (DR)

- 本錢夠粗的可以考慮 異地備援 (架兩座叢集做同步)
- 一些解決方案：
  - Cloudera Backup and Disaster Recovery (BDR)
  - 把 HDP 備份到 Azure
- BDR 主要備份的對象是
  - Metastore (有用到外部關聯式資料庫的部分)
  - HDFS 的內容



https://0x0fff.com/hadoop-cluster-backup/

Hadoop 剛滿十歲，後繼者還在追趕進度

- **Hadoop Security 的四大範疇：**
  - **Authentication** – 帳號密碼認證
  - **Authorization** – 基於帳號身分，管理讀寫權限
  - **Auditing** – 稽核讀寫的紀錄
  - **Encryption** – 資料的加密、通訊的加密 (運算過程的加密?)
- **那 Spark 呢?? 還在經歷生長痛中....**
  - **Authentication** – 1.3 剛支援 Kerberos
    https://issues.apache.org/jira/browse/SPARK-5493
  - **Authorization** – 目標做到 Spark SQL column-level 管控
  - **Auditing** – 是否有工具??還在找
  - **Encryption** – 進行中
    https://issues.apache.org/jira/browse/SPARK-5682

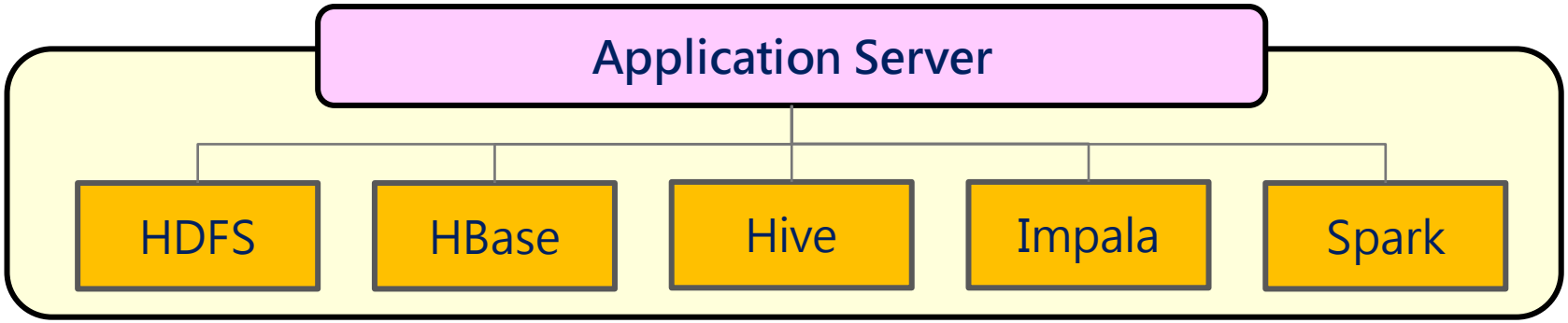
身分認證 Authentication

- 現況：支援度最廣的還是 Kerberos

	CM	CN	HDFS	MRv1	YARN	Flume	HBase	HCatalog	Hive Server 2	HiveServer	Hue	Impala	Llama	Oozie	ZooKeeper
	一個以上	擇一	擇一	擇一	擇一	擇一	擇一	擇一	擇一	擇一	擇一	雙重*	擇一	擇一	擇一
simple			V	V	V	V	V	V	V	V		V	V	V	V
Database	V	V									V				
Open LDAP	V	V							V		V	V*			
AD	V	V							V		V	V*			
LDAPS	V	V							V			V*			
Kerberos	V		V	V	V	V	V	V	V	V	V	V	V	V	V
External Program	V								CLASS						
SAML	V										V				
OpenID											V				
Oauth											V				

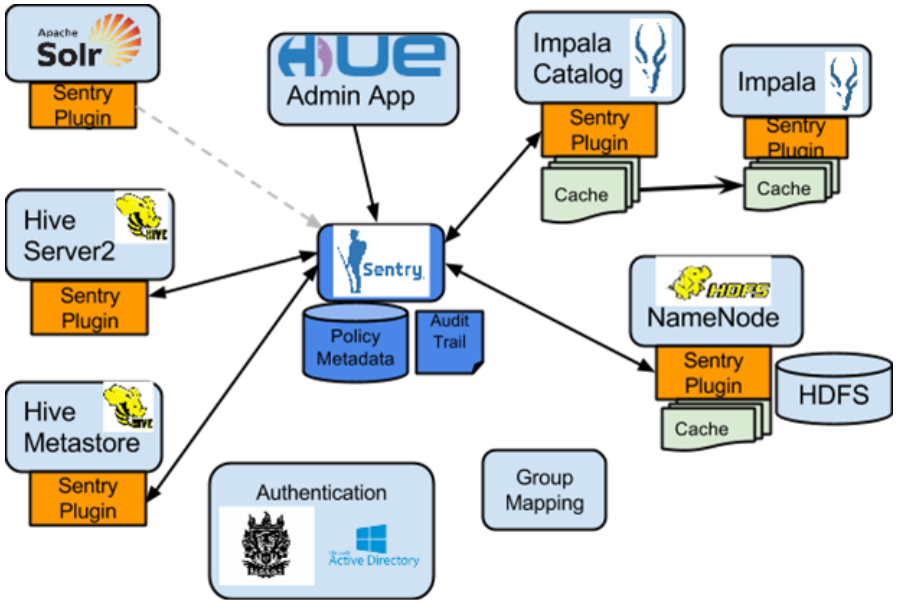
身分認證 Authentication

- **導入前的自我檢驗項目：**
  - 組織內部有沒有合適的系統管理者可以協助 Kerberos 問題排查
  - 組織內部的網管能否協助 Kerberos 跨網段傳輸的問題排查
  - 叢集成長到一定數量時，能否接受 Kerberos 認證會影響效能
- **折衷選擇一：AD/LDAP**
  - 如果目前組織的需求跟 Data Warehouse Offload 有關
    只會用到 Hive / Impala 等 SQL on Hadoop 的元件
- **折衷選擇二：透過 API Server 做隔離層**



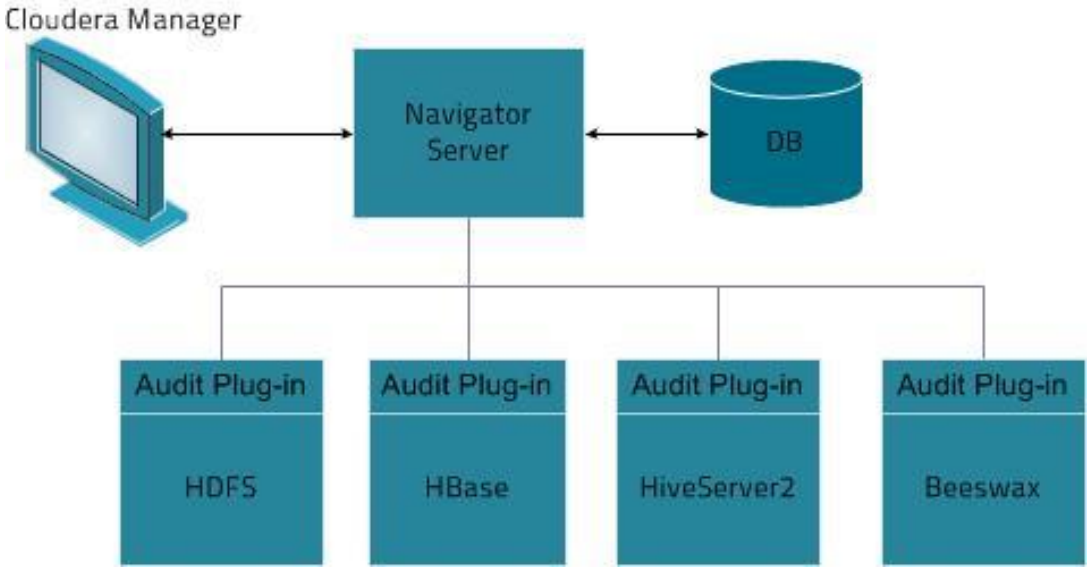
權限控管 Authorization

- 現況：
  - Apache **Sentry** or Apache **Ranger**
  - Role-Based Access Control
  - Fine-grained access control – 目標是 column-based 權限控管
  - Unified Authorization



存取稽核 Auditing（ Accounting ）

- **現況：**
  - 土砲手工打造每個元件的Audit Log 蒐集系統
  - Cloudera Navigator
  - Apache Ranger Audit Framework



Audit Logging

In 0.18 and later, one can enable audit logging from
number of events emitted from that interface (see e

Format

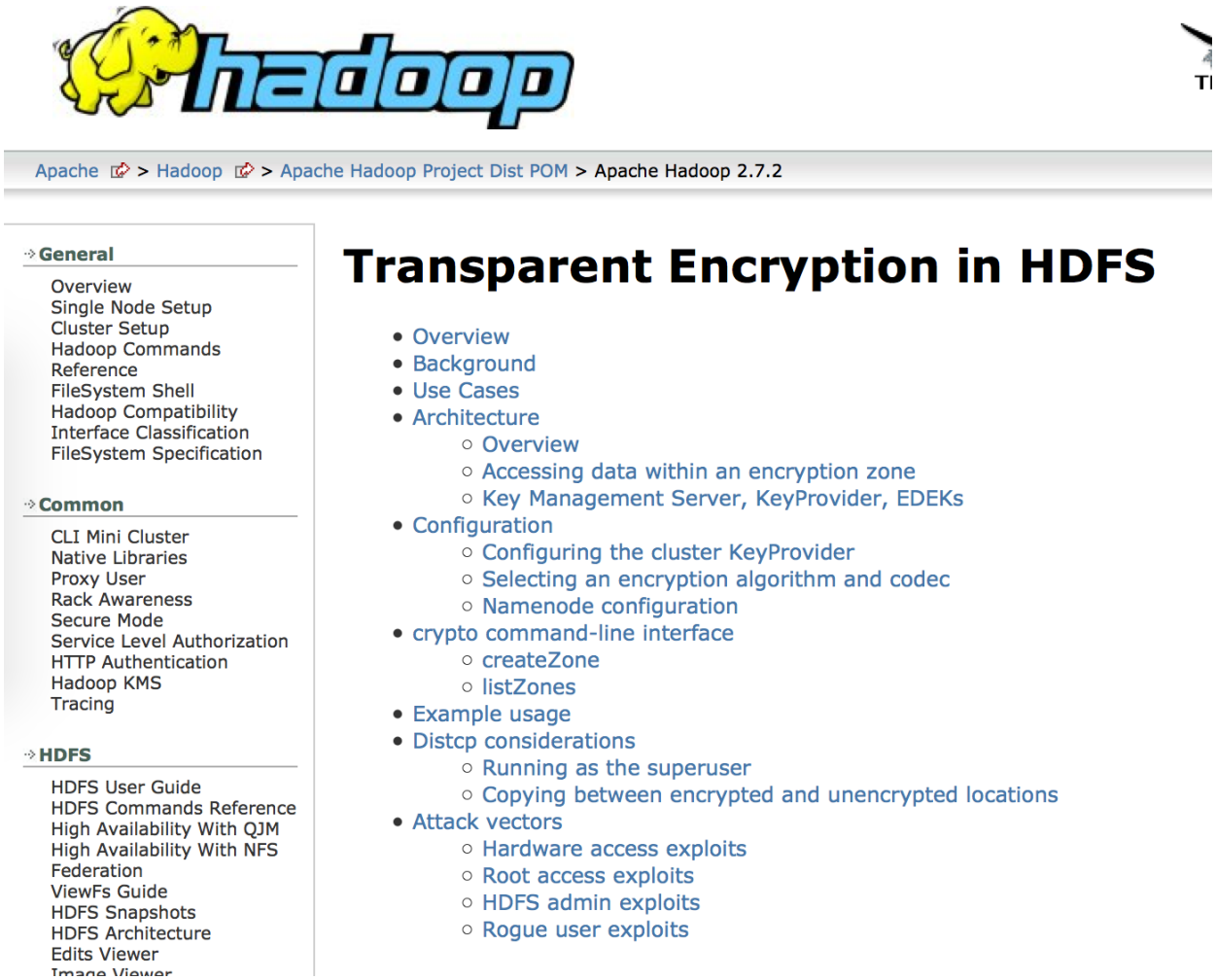
key	value
ugi	<user>,<group>[,<group>]*
ip	<client ip address>
cmd	(open|create|delete|rename|mkdirs|listSta
src	<path>
dst	(<path>|"null")
perm	(<user>:<group>:<perm mask>|"null")

Sample line of audit output:

```
<log4j header> ugi=wsmith,users,staff
```

加密防護 Encryption

- 傳輸過程的加密
- 儲存資料的加密



Agenda

- **企業導入大數據的四個階段**
  - 專案規劃、大數據平台建置、**大數據平台資安**、大數據品質管制
- **大數據平台資安範疇與現況**
  - 高可用性 High Availability (HA)
  - 災害復原 Disaster Recovery (DR)
  - 身分認證 Authentication
  - 權限控管 Authorization
  - 存取稽核 Auditing (Accounting)
  - 加密防護 Encryption
- **結語**
  - 題外話：開放原始碼軟體的資安掃描

現在進行式：Data Governance

- 全球頂尖的極極少數創新者已經走到「數據品管」的階段!!
- Data Governance

```
Data governance is a control that ensures that the data
entry by an operations team member or by an automated
process meets precise standards, such as a business rule,
a data definition and data integrity constraints in the
data model.
```

- Apache Atlas

http://atlas.incubator.apache.org/

Apache Atlas

題外話：開放原始碼軟體的資安掃描

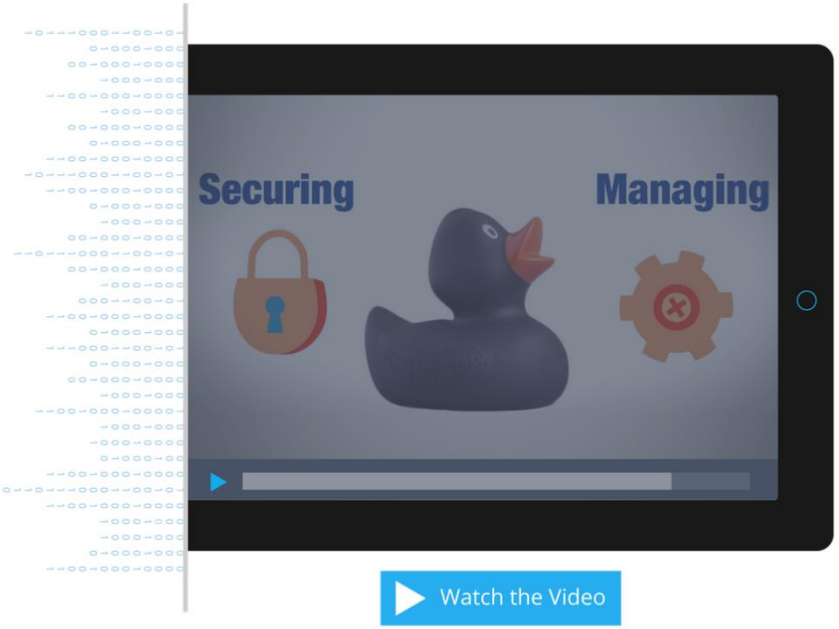
當 Open Source 變成一種商業策略

企業該如何確保所採用的開源軟體是安全的呢？

縱使掃過原始碼，又能確保多少安全性呢？

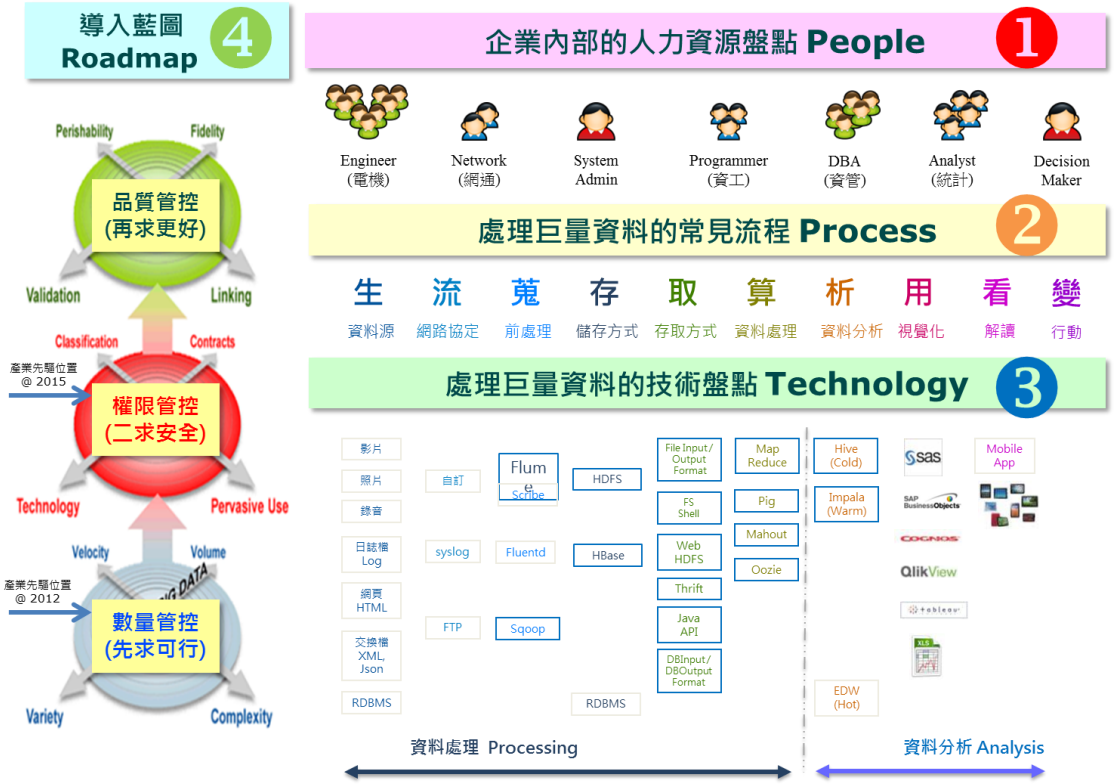
Black Duck Hub Open Source Security Management

- Automatically inventory open source in your code
- Map to known vulnerabilities
- Manage remediation activities
- Monitor and alert when new threats are reported



結語

- **企業對「資訊安全」的需求成為大數據平台的獲利模式**
  - 對各種 Security 需求的支援完整性也象徵著Hadoop 正式進入 Enterprise Software !!
- **先求有，二求安全，再求品質**



Q & A

JAZZWANG.TW 老鼠 GMAIL 點 COM