model futures
www.modelfutures.com

presents:

IntegratedEA

STRATEGY • OPERATIONS • TECHNOLOGY

www: http://www.integrated-ea.com
HashTag: #IEA13
Twitter: @IntegratedEA

Ministry of Defence

PA Consulting Group

de&S

mega

BMT Hi-Q Sigma
"Where will our knowledge take you?"

LOCKHEED MARTIN

GCHQ

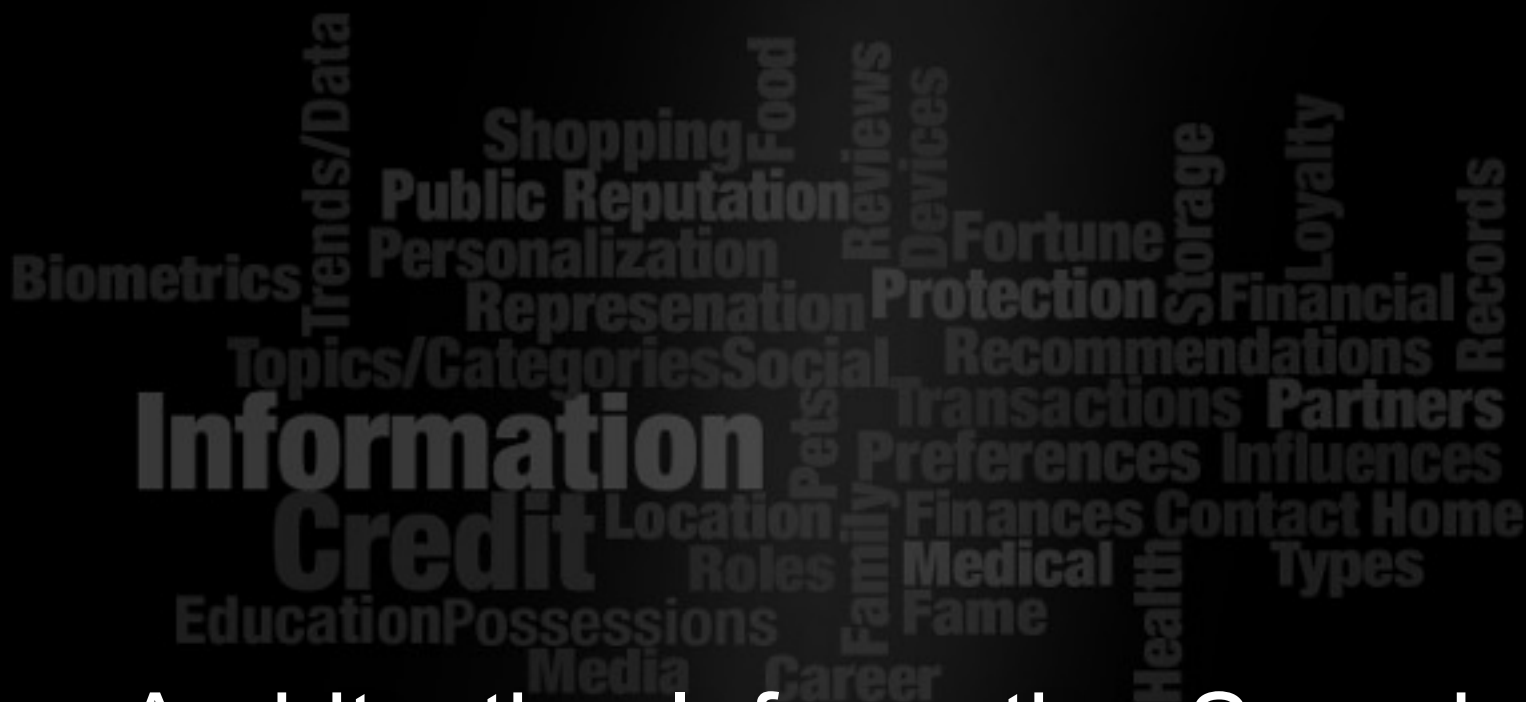
SyntheSys

JCSys

ATKINS

SCJ SecurityCleared Jobs.com

UKCeB



Architecting Information Superiority in the Fight Against Cyber-Crime

Steve Harris
Experian Ltd.

steve.w.harris@experian.com

Introduction

- Background
- About the technology
- Advantages
- Challenges and issues
- Lessons learned

Background

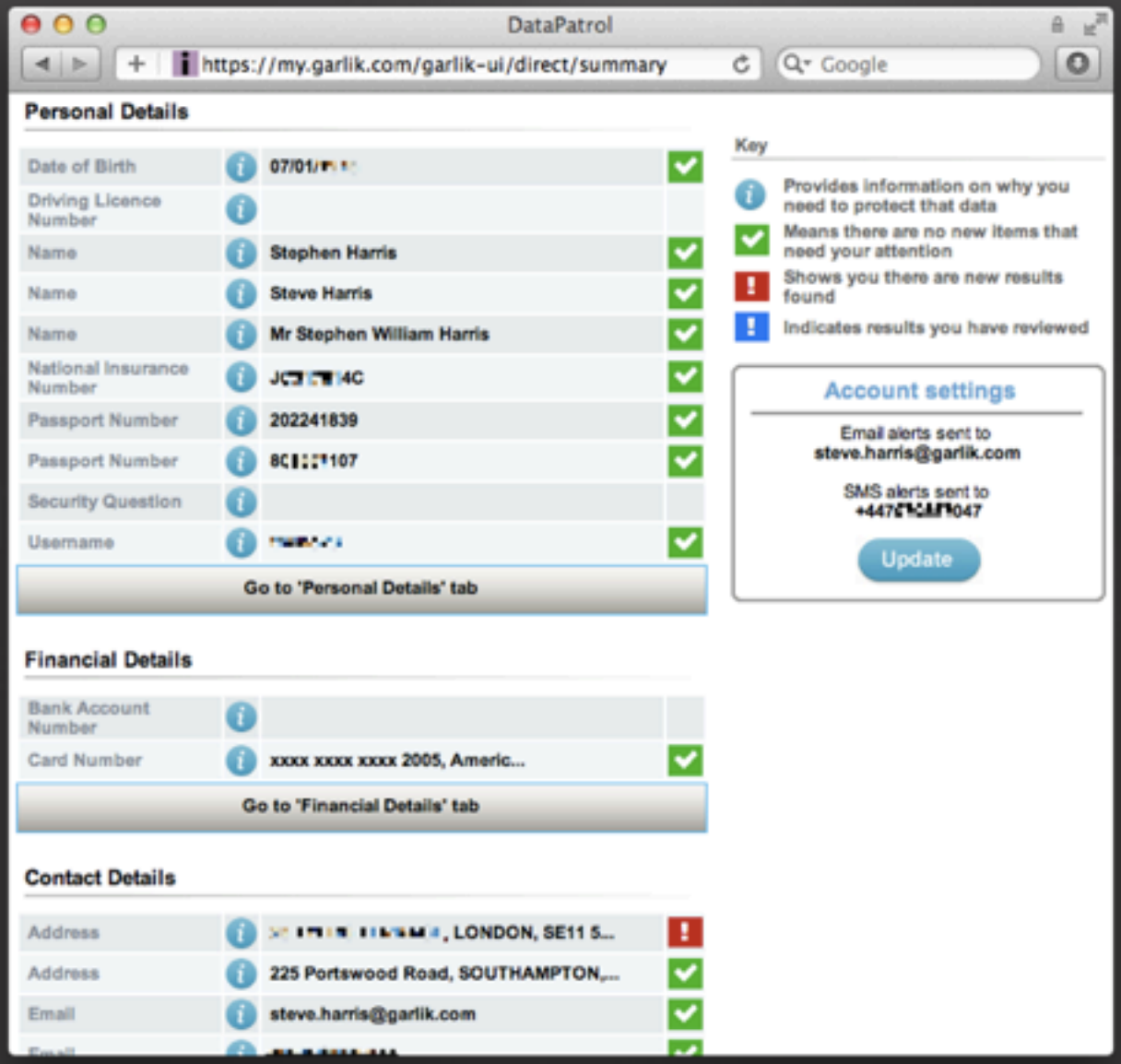
- Experian is a global information services company, providing data and analytical tools to clients around the world
- Acquired this technology through acquisition
- Garlik was a startup, founded in 2005, focussed on personal data
- Acquired by Experian in late 2011
- Provides advance warning of online fraud by analysing various data sources

garlik™
a part of Experian

The DataPatrol Service

- **Web application or Web Service**
  - B2B2C channel, or
  - Embedded in customers online banking system
- **Monitors users data in**
  - Web sites
  - Social networks
  - Hacker forums
  - IRC channels
  - Public databases
- **Distributed mainly though Financial Services**
  - Also ISPs, AV vendors, etc.
- **Live in 11 countries**
  - Australia, Canada, France, Germany, India, Ireland, Italy, Spain, Turkey, UK, US

DataPatrol

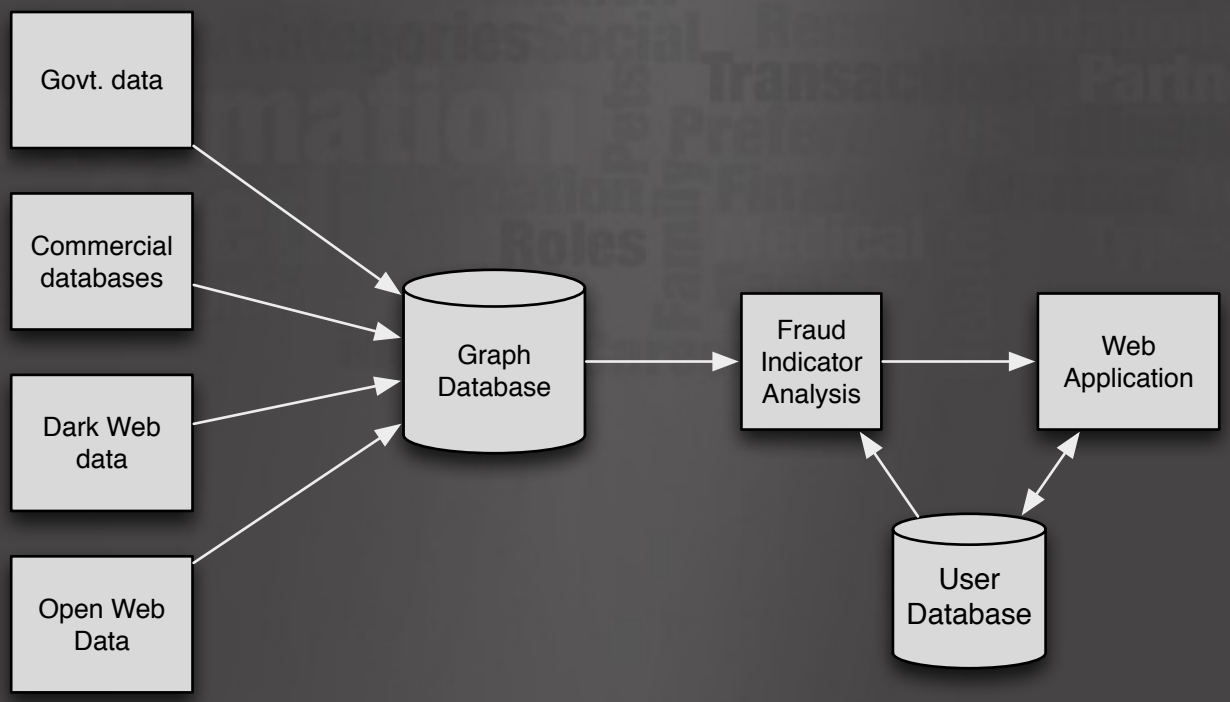

DataPatrol
https://my.garlik.com/garlik-ui/direct/summary
Google
Personal Details
Date of Birth
07/01/
Driving Licence Number
Name
Stephen Harris
Name
Steve Harris
Name
Mr Stephen William Harris
National Insurance Number
4C
Passport Number
202241839
Passport Number
107
Security Question
Username
Go to 'Personal Details' tab
Financial Details
Bank Account Number
Card Number
xxxx xxxx xxxx 2005, Americ...
Go to 'Financial Details' tab
Contact Details
Address
, LONDON, SE11 5...
Address
225 Portswood Road, SOUTHAMPTON,...
Email
steve.harris@garlik.com
Key
Provides information on why you need to protect that data
Means there are no new items that need your attention
Shows you there are new results found
Indicates results you have reviewed
Account settings
Email alerts sent to
steve.harris@garlik.com
SMS alerts sent to
047
Update

The Requirements

- **Large amounts of data**
  - Over 2TB per day
- **Complex analysis and interlinking**
  - Mine 1Bn relationships per day
  - Spot relatively complex patterns that indicate potential fraud
- **Flexible to new data patterns, sources, and types**
  - Currently identifies around 30 different types of PII, and financial identifiers
- **Respond quickly to new data arriving**
  - Responsiveness is key to providing protection
- **High uptime requirements**
  - Financial services sector has strict SLA needs

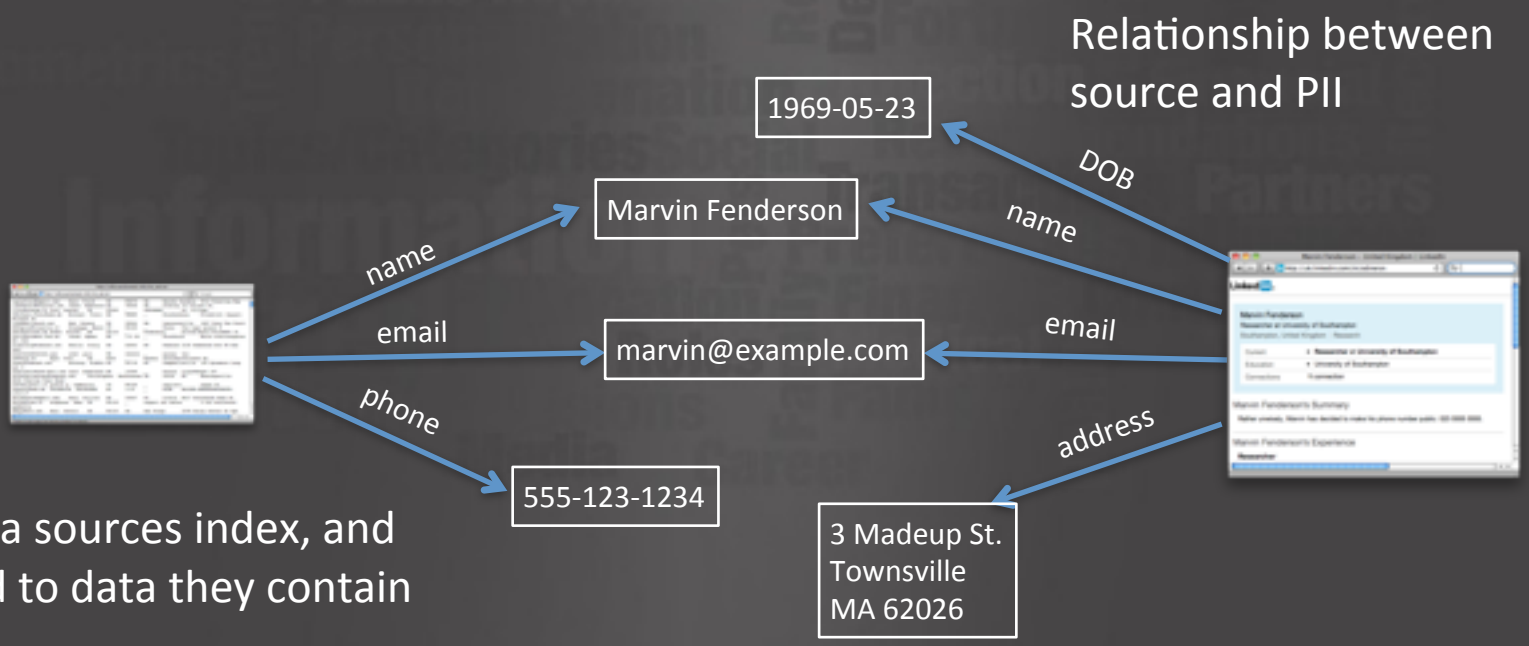
System Architecture

- SoA - for ease of deployment and resilience
- Query answering services - for modularity
- Scheduling and prioritisation services - to ensure responsiveness of key operations
- Modular data conversion system - for ease of maintenance



Data Modelling

Uses graphs / networks to represent PII occurrences



Benefits of graph data storage

- **Strong standards, defined by W3C**
  - RDF - graph data
  - SPARQL - query language
  - Good support for unicode, and URL/URIs

- **Very flexible, lightweight schema**
  - Schema can be extended trivially
  - Requires no downtime
  - Doesn't affect existing queries

- **Scales well**
  - Hundreds of billions of relationships

- **Sophisticated graph-based queries**
  - Match patterns
  - Extract relationships

- **Intrinsic provenance tracking**
  - Commercial reasons
  - Legal reasons
  - Operational benefits

- **RESTful SoA**
  - Efficiently query over HTTP

The Challenges

- **Quantity of IO**
  - More than a high performance SAN can sustain - use local SSD for backing store
- **Backups**
  - Size is an issue - backup to high performance SAN/NAS devices
- **Finding developers with graph database experience**
  - We don't try anymore - train up skilled software engineers
  - Basics can be picked up quickly
  - Similar challenge with tuple stores - people out there, but not many
- **Relative immaturity of tools**
  - Similar to tuple stores
  - Support tools and optimisers not up to SQL standard

Lessons learned

- **There are lots of technologies that can be layered on RDF & SPARQL**
  - We don't use any of them
  - Appropriate for some cases
  - Hurt simplicity and performance, cost is high
- **It's practical for enterprise applications**
  - Even in demanding industries
- **Even large, dynamic data sets can be processed**
  - Like Big Data, but with complex joins

Conclusions

- **Another tool in the database storage toolkit**
  - Not a replacement for SQL or Tuple stores
  - All have their strengths
  - A blend of Big Data and SQL capabilities - has some strengths and weaknesses of both
- **Particularly appropriate for**
  - Complex provenance requirements
  - Complex/partial data
  - Data with variable cardinality
  - Environments where data needs change often
  - System which require topographically complex queries
- **Not appropriate for**
  - Datasets where data is highly rectangular
  - Systems which require very simple queries
  - Datasets that require complex, hand-specified indexes