HPCC SYSTEMS®

Data Analytics in Cyber-Security and Threat Intelligence

RELX Group

2015 Notable Cyber-Security Incidents

PREMERA BLUE CROSS

January: 11.2M Subscribers impacted

Anthem BlueCross

February: 80M patients and employees impacted

PENNSTATE 1855

May: 18,000 people impacted

February: $1B cyberheist impacting as many as 100 banks worldwide

BlueCross BlueShield

May: 1.1M Subscribers impacted

KASPERSKY lab

June: Corporate Intellectual Property stolen

LastPass ****

June: Compromised subscriber data

]HackingTeam[

July: 1M emails published and Intellectual Property stolen

HARVARD UNIVERSITY

July: 8 schools and Administrative offices impacted

ARMY NATIONAL GUARD

July: 850,000 members impacted

T··Mobile·

Experian A world of insight

September: information from 15M T-Mobile customers affected

TalkTalk

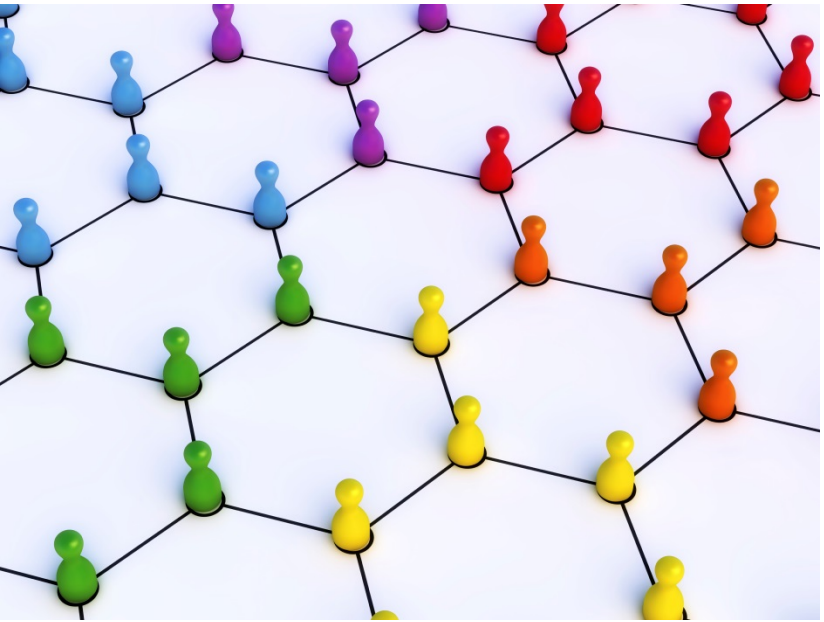
October: information from 5M customers compromised

Big data analytics tools will be the first line of defense, combining machine learning, text mining and ontology modeling to provide holistic and integrated security threat prediction, detection, and deterrence and prevention programs.

- The International Institute of Analytics (IIA)

HPCC SYSTEMS®

Big Data Example: Suspected Iran-Based Hacker Group Creates Network of Fake LinkedIn Profiles



THREAT SUMMARY

Researchers uncovered a network of 25 fake, convincing profiles forming a self-referenced network of seemingly established LinkedIn users

The purpose of this network is to target potential victims through social engineering

Source: Dell SecureWorks Counter Threat Unit™ Threat Intelligence

Big Data Example: Suspected Iran-Based Hacker Group Creates Network of Fake LinkedIn Profiles

DATA DISCOVERY

- Profile photograph linked to multiple identities across numerous websites
- Summary in one profile is identical to a legitimate LinkedIn profile
- Employment history matches a sample résumé downloaded from a recruitment website
- Job descriptions were copied from legitimate job postings

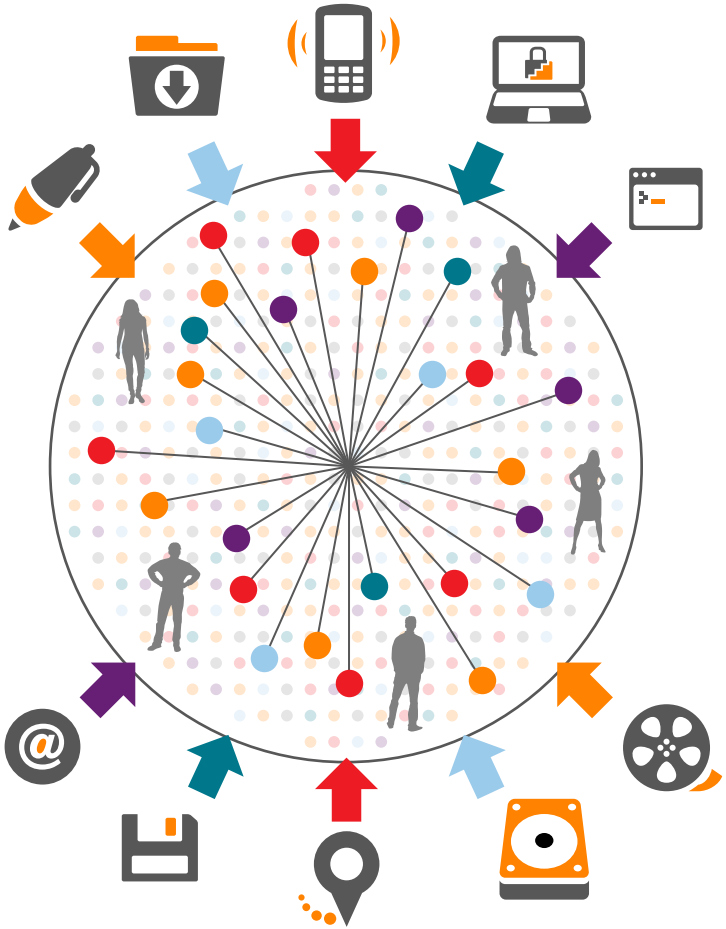
THE OUTCOME:

Linked to documented threat actors using malware disguised as a résumé application

HPCC SYSTEMS®

The Data Problem

- Significant volume of data
- Varying formats
- High dimensional
- Non-linear classification (at least not in reasonable high dimensional spaces)
- Requires near real-time event processing and continuous data integration
- Multi-dimensional anomaly detection can be computationally expensive
- When you add social media, NLP comes into play



Tackling the Trivial Pieces

Data adaptation layer

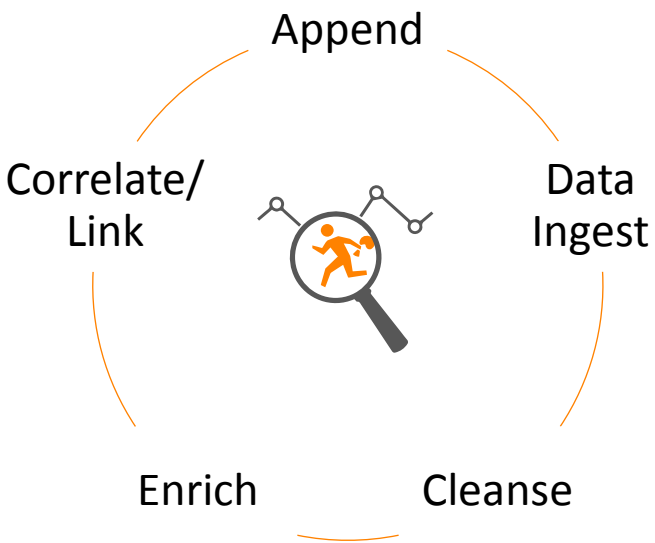
- Parsing and structuring data (50-100 different log formats to deal with)
- Data enrichment (geo-location, name resolution, tagging, etc.)

Stream/event processing

- Feature/Fact extraction
- Heuristics based event processing and lightweight correlation
- Classification/anomaly detection

HPCC SYSTEMS®

The Cool Parts



Human-machine collaboration is a good opportunity for active learning

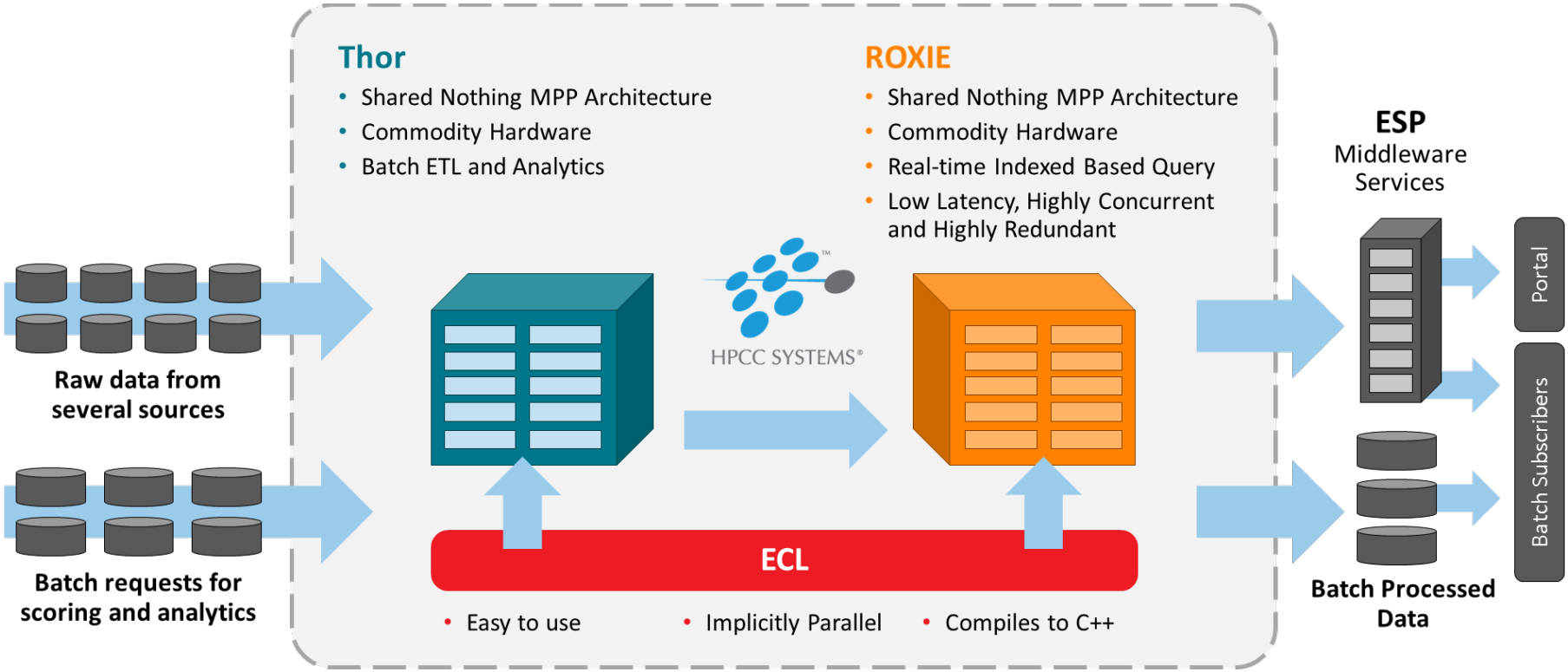
- Deterministic/probabilistic data linkage
- Distribution analysis for anomaly detection
- Build/append to entity relationship model
- Rank anomalies/events and bubble them up for the analysts to review and label
- Learn, rinse, repeat

HPCC SYSTEMS®

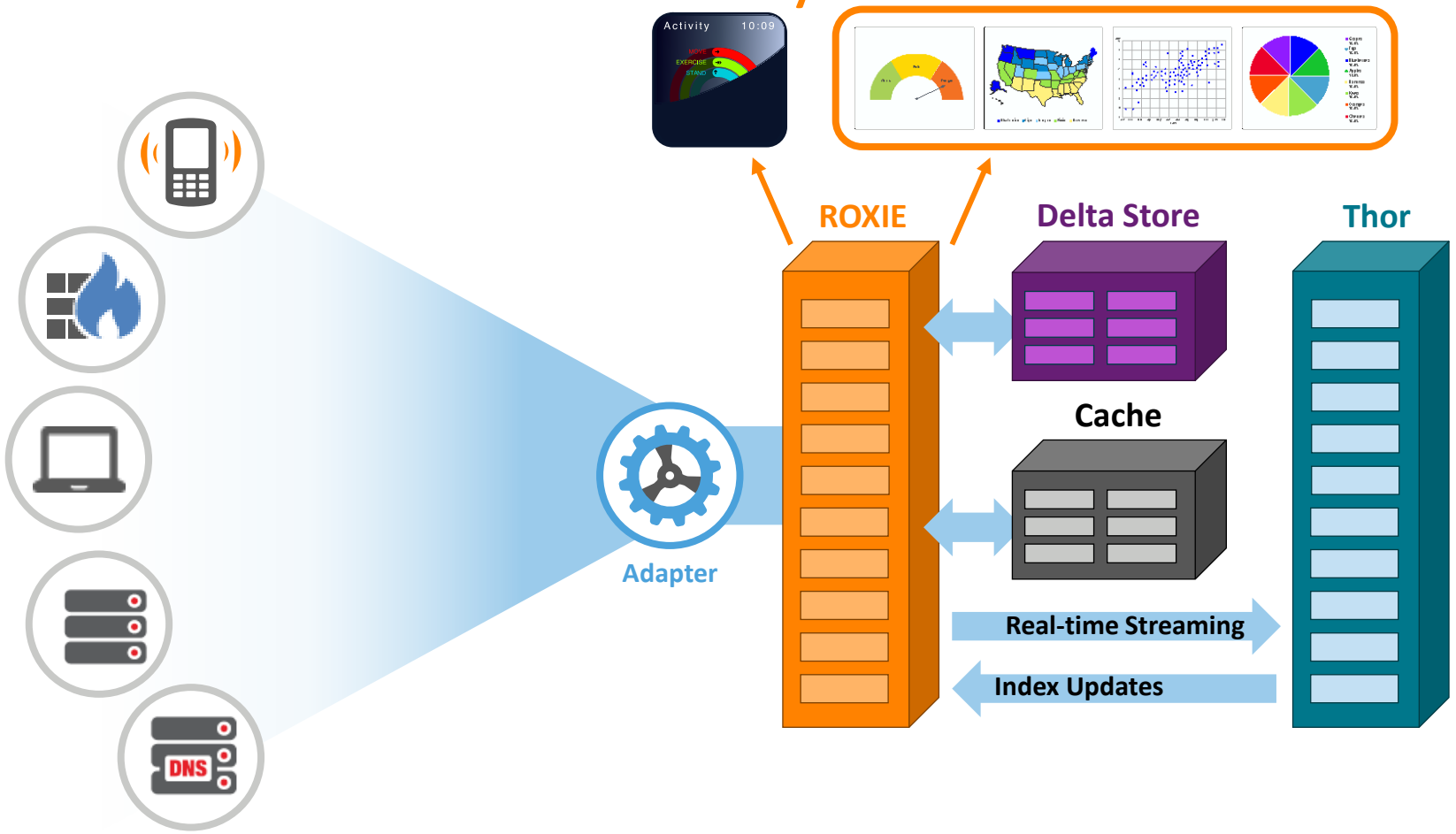
The Real Challenge

- Limited analyst bandwidth
- Manual exploration takes long and requires significant expertise
- Feature extraction and feature creation is difficult and error prone (hierarchical learning of higher order features shows promise)
- Important information may not be available to the system and/or available only on demand
- Cyber-security events are, hopefully, anomalies, so don't expect a training set large enough to offset dimensionality

How We Do It

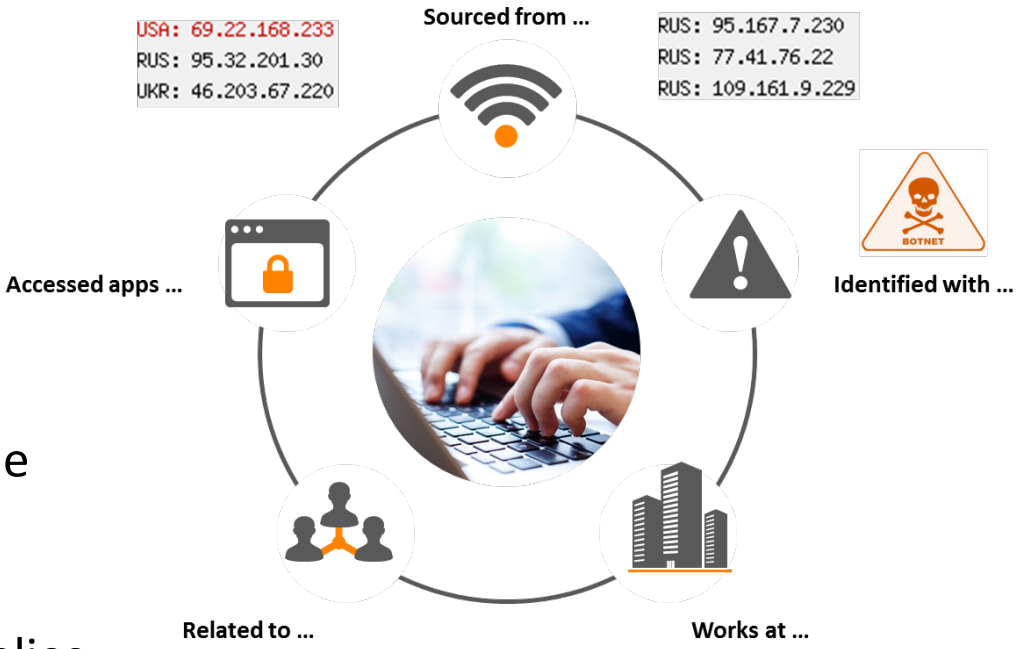

Thor
• Shared Nothing MPP Architecture
• Commodity Hardware
• Batch ETL and Analytics
ROXIE
• Shared Nothing MPP Architecture
• Commodity Hardware
• Real-time Indexed Based Query
• Low Latency, Highly Concurrent and Highly Redundant
HPCC SYSTEMS®
ESP
Middleware
Services
Portal
Batch Subscribers
Raw data from several sources
Batch requests for scoring and analytics
ECL
• Easy to use
• Implicitly Parallel
• Compiles to C++
Batch Processed Data

Real-Time Collection and Analytics Platform



In a Nutshell

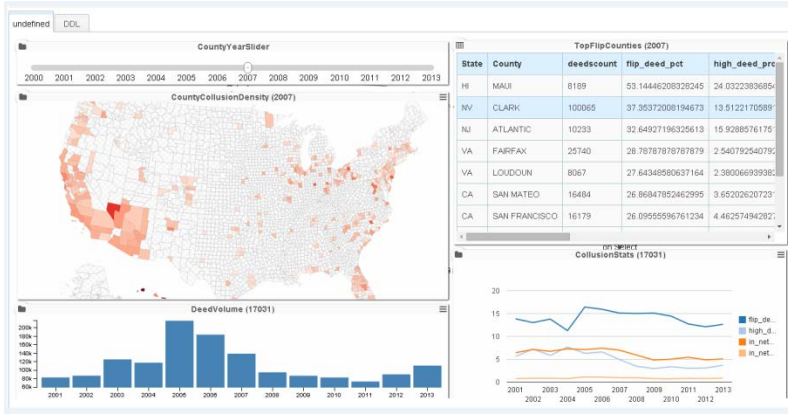
- Millions of application transaction log records per day
- Combine information from network routers, DNS servers, etc.
- Use third party sources to identify blacklisted information
- Apply learning techniques to match the digital footprint to identify the virtual identities
- Prevent intrusions by detecting anomalies and actively block known bad actors

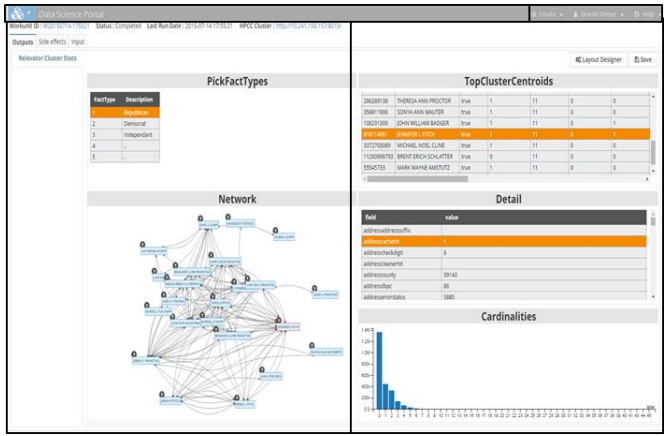


HPCC SYSTEMS®

How does it look?

- Build complex relationship graphs easily using a **drag and drop visual interface**
- Gain insight into hidden relationships using **state of the art visualization techniques**





What's next?

- Accelerated stream processing with hybrid computing
  - Micron in-memory Finite State Machines
  - FPGA
  - Co-processors
- Tighter integration with other systems
  - Mitre CRITS
  - Google GRR
  - Sandboxing, malware detection, etc.
- Leverage social media feeds and non-traditional threat intelligence sources
- Take over the World **No, not really…..**

HPCC SYSTEMS®

Resources

- Open Source HPCC Systems Platform: http://hpccsystems.com
- Source Code on GitHub: https://github.com/hpcc-systems
- Free Online Training: http://learn.lexisnexis.com/hpcc
- HPCC Community Forums: http://hpccsystems.com/bb
- ECL-WLAM: https://github.com/hpcc-systems/ECL-WLAM
- Mitre CRITS: http://www.mitre.org/publications/project-stories/cyber-intelligence-gets-even-smarter-with-crits
- Google GRR: https://github.com/google/grr

HPCC SYSTEMS®

Thank You Very Much!

Questions?