IDC
Analyze the Future

Directions 2013

March 5 | Santa Clara, CA
March 13 | Boston, MA

High-Performance Data Analysis: HPC Meets Big Data

Steve Conway
IDC Research VP, HPC
sconway@idc.com

Chirag DeKate
IDC Research Manager, HPC
cdekate@idc.com

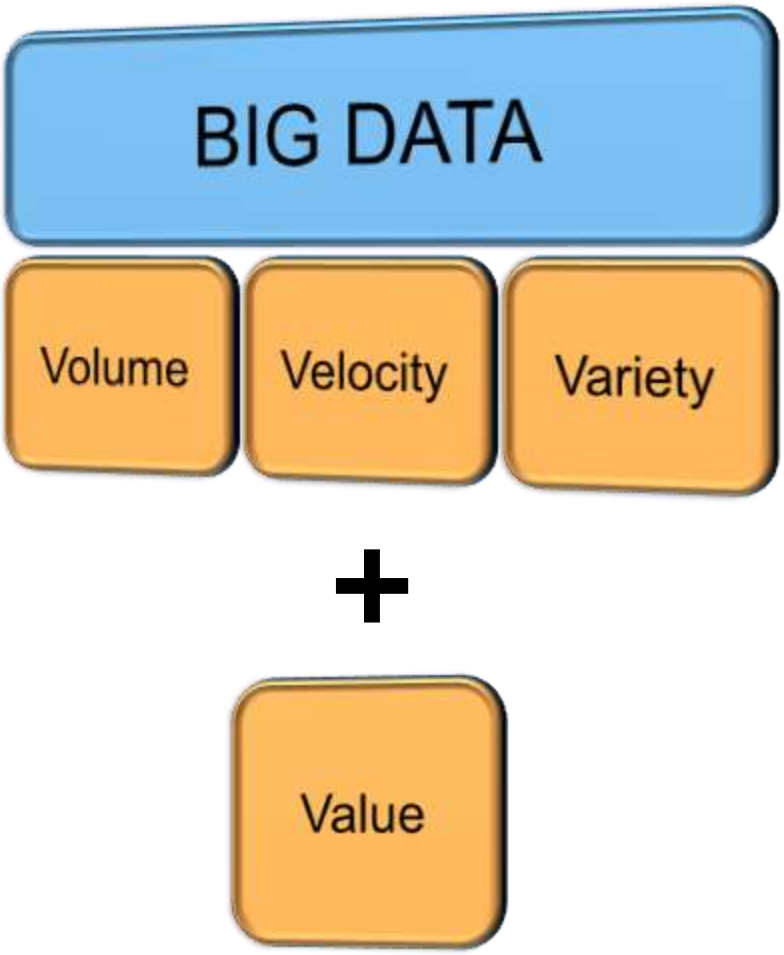
#directions13

HPDA User Talks: HPC User Forums, UK, Germany, France, China, U.S.

- HPC in Evolutionary Biology, Andrew Meade, University of Reading
- HPC in Pharmaceutical Research: From Virtual Screening to All-Atom Simulations of Biomolecules, Jan Kriegl, Boehringer-Ingelheim
- European Exascale Software Initiative, Jean-Yves Berthou, Electricite de France
- Real-time Rendering in the Automotive Industry, Cornelia Denk, RTT-Munich
- Data Analysis and Visualization for the DoD HPCMP, Paul Adams, ERDC
- Why HPCs Hate Biologists, and What We're Doing About It, Titus Brown, Michigan State University
- Scalable Data Mining and Archiving in the Era of the Square Kilometre Array, the Square Kilometre Array Telescope Project, Chris Mattmann, NASA/JPL
- Big Data and Analytics in HPC: Leveraging HPC and Enterprise Architectures for Large Scale Inline Transactional Analytics in Fraud Detection at PayPal, Arno Kolster, PayPal, an eBay Company
- Big Data and Analytics Vendor Panel: How Vendors See Big Data Impacting the Markets and Their Products/Services, Panel Moderator: Chirag Dekate, IDC
- Data Analysis and Visualization of Very Large Data, David Pugmire, ORNL
- The Impact of HPC and Data-Centric Computing in Cancer Research, Jack Collins, National Cancer Institute
- Urban Analytics: Big Cities and Big Data, Paul Muzio, City University of New York
- Stampede: Intel MIC And Data-Intensive Computing, Jay Boisseau, Texas Advanced Computing Center
- Big Data Approaches at Convey, John Leidel
- Cray Technical Perspective On Data-Intensive Computing, Amar Shan
- Data-intensive Computing Research At PNNL, John Feo, Pacific Northwest National Laboratory
- Trends in High Performance Analytics, David Pope, SAS
- Processing Large Volumes of Experimental Data, Shane Canon, LBNL
- SGI Technical Perspective On Data-Intensive Computing, Eng Lim Goh, SGI
- Big Data and PLFS: A Checkpoint File System For Parallel Applications, John Bent, EMC
- HPC Data-intensive Computing Technologies, Scott Campbell, Platform/IBM
- The CEA-GENCI-Intel-UVSQ Exascale Computing Research Centre, Marie-Christine Sawley, Intel

WHAT IS HPDA?

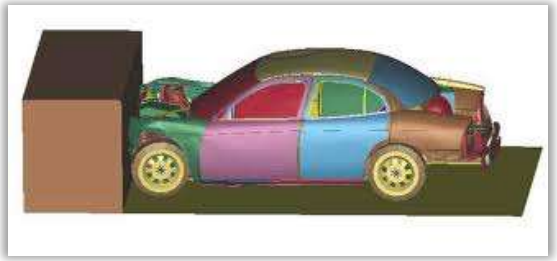
Big Data: General Definition



- Lots of data
- Time critical
- Multiple types
  (e.g., numbers, text, video)
- Worth something to someone

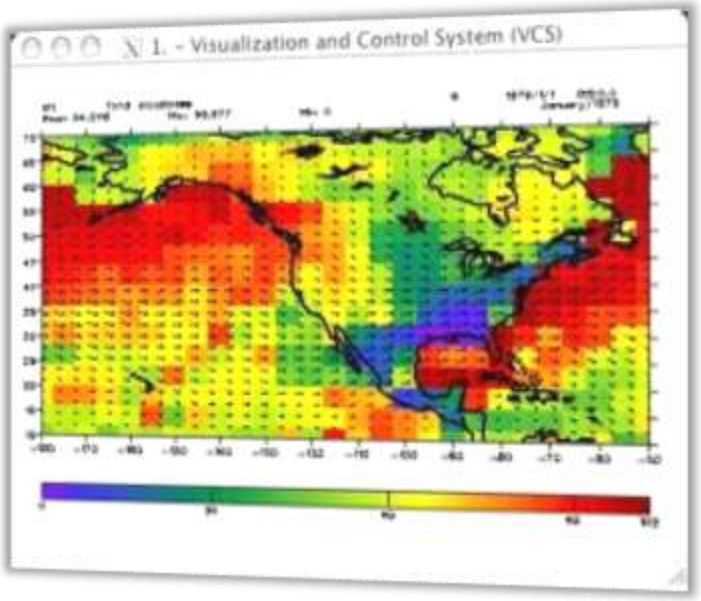
HPDA: Data-Intensive Simulation and Analytics

- HPDA = tasks involving sufficient data volumes and algorithmic complexity to require HPC resources
- Established (simulation) or newer (analytics) methods
- Structured data, unstructured data, or both
- Regular (e.g., Hadoop) or irregular (e.g., graph) patterns
- Government, industry, or academia
- Upward extensions of commercial business problems
- Accumulated results of iterative problem-solving methods (e.g., stochastic modeling, parametric modeling).



Simulation: An Original Big Data Market

- Simulation-based data-intensive computing is a long-standing HPC workload category
- Examples:
  - Weather forecasting/climate modeling
  - Parametric modeling for product design
  - FSI: portfolio optimization, risk analysis, pricing exotics
  - Life sciences: genomics, drug discovery and more
  - National security: signal intelligence
  - Etc…



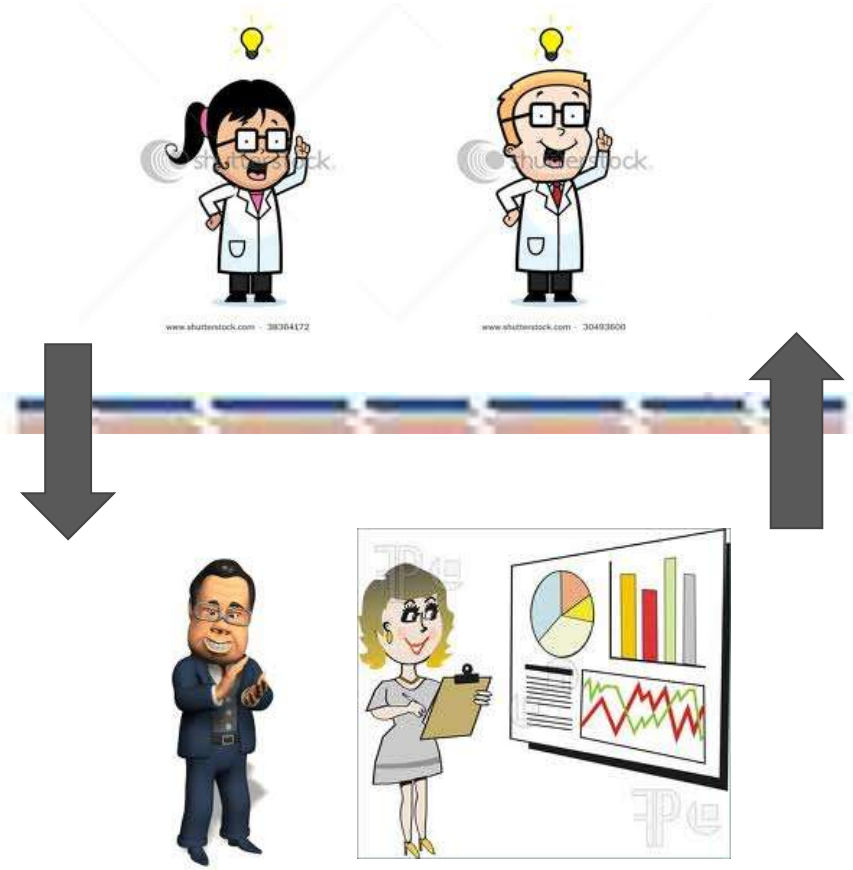
Analytics: A Newer, Complementary Big Data Market

- Analytics methods applied to established HPC domains in industry, government, academia:

- High-end commercial analytics pushing up into HPC

- The journey from science to industry/ commerce can be relatively short:

The Boundaries Between Data-Intensive HPC and High-End Commercial Analytics Are Dissolving



- HPC vendors are targeting commercial markets, driven by opportunity.
- Commercial vendors are moving up to HPC, driven by customers.

HPDA Market Drivers

- More input data (ingestion)
  - More powerful scientific instruments/sensor networks
  - More transactions/higher scrutiny (fraud, terrorism)
- More output data for integration/analysis
  - More powerful computers
  - More realism
  - More iterations in available time
- The need to pose more intelligent questions
  - Smarter mathematical models and algorithms
- Real time, near-real time requirements
  - Catch fraud before it hits credit cards
  - Catch terrorists before they strike
  - Diagnose patients before they leave the office
  - Provide insurance quotes before callers leave the phone

Data Movement Is Expensive: Energy and Time-to-Solution

Energy Consumption

- 1MW ≈ $1 million
- Computing 1 calculation ≈ 1 picojoule
- Moving 1 calculation = up to 100 picojoules

Time

- "Sneakernet" lives on!

Strategies

- Accelerate data movement (bandwidth, latency)

- Minimize data movement (e.g., data reduction, in-memory compute)

Different Systems for Different Jobs

Partitionable Big Data Work

- Most jobs are here!
- Goal: search
- Regular access patterns (locality)
- Global memory not important
- Standard clusters + Hadoop, Cassandra, etc.

Non-Partitionable Work

- Toughest jobs (e.g., graphing)
- Goal: discovery
- Irregular access patterns
- Global memory very important
- Systems turbo-charged for data movement +graphing

versus

HPC architectures today are compute-centric (FLOPS vs. IOPS)

IDC HPDA Server Forecast

- Fast growth from a small starting point
- In 2015, conservatively approaching $1B (£600M+)

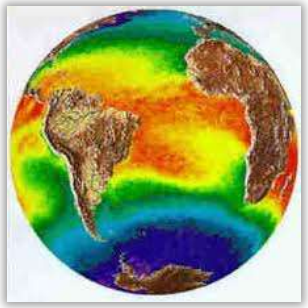
TABLE 2

IDC Worldwide Data Intensive (Big Data) Focused HPC Server Revenues

($ Millions)

	2009	2010	2011	2012	2013	2014	2015	CAGR '10-'15
WW HPC Server Sales	8,637	9,504	10,034	10,564	11,397	12,371	13,485	7.2%
Big Data Workloads	535	603	655	708	786	881	989	10.4%
Big Data in HPC Portion	6.2%	6.3%	6.5%	6.7%	6.9%	7.1%	7.3%	3.0%

Source: IDC 2012



Use Cases: HPDA

Some Major Use Cases for HPDA

- Fraud/error detection across massive databases
  - A horizontal use – applicable in many domains
- National security/crime-fighting
  - SIGINT/anomaly detection/anti-hacking
  - Anti-terrorism (including evacuation planning)/anti-crime
- Health care/medical informatics
  - Drug design, personalized medicine
  - Outcomes-based diagnosis & treatment planning
  - Epidemiology
  - Systems biology
- Customer acquisition/retention
- Smart electrical grids
- Design of social network architectures

Apollo Group/University of Phoenix: Student Recruitment and Retention

- University of Phoenix: 280,000 students and growing
- Must target millions of students to produce this yield.
- Also tracks student performance for early identification of potential dropouts – “churn” is very expensive
- Solution: sophisticated, cluster-based Big Data models

APOLLO GROUP

GEICO: Real-Time Insurance Quotes

- **Problem:** Need accurate automated phone quotes in 100ms
- **Solution:** Each weekend, use HPC cluster to pre-calculate quotes for every American adult and household (60 hours)

Use Case: PayPal
Fraud Detection / Internet Commerce

Slides and permission provided by PayPal, an eBay company

The Problem

Detecting fraud in 'real time' as millions of transactions are processed between disparate systems at volume.

Finding suspicious patterns that we don't even know exist in related data sets.

Ability to create and deploy new fraud models into event flows quickly and with minimal effort.

Provide environment for fraud modeling, analytics, visualization, M/R, dimensioning and further processing.

What Kind of Volume?

10 million+ logins / day

13 million financial transactions / day

300 variables calculated per event for some models

~4 Billion inserts / day

~8 Billion selects / day

Where Are We Using HPC?

Infiniband on all internal Trinity network (Mellanox QDR 40Gb dual plane)

SGI InifiniteStorage IS4600 for EFL databases.

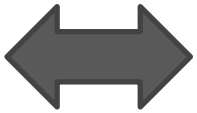
3 SGI Altix ICE 8200/8400 clusters for all 120+ EFL memory based apps – no disk i/o overhead.

MPI "like" apps. MPP features with scale out and affinity processing.

R&D with Lustre on Hadoop cluster and POC of columnar based database.

Global Courier Service: Fraud/Error Detection

- Check 1 billion-plus packages per hour in central sorting facility
- Benchmark won by unannounced vendor with a turbo-charged interconnect and memory system



CMS: Government Health Care Fraud

- 5 separate databases for the big USG health care programs under Centers for Medicare and Medicaid Services (CMS)
- Estimated fraud: $150B-$450B (£95B-£280B). <$1B caught today)
- ORNL, SDSC have won evaluation contracts to unify the databases and perform fraud detection on various architectures.

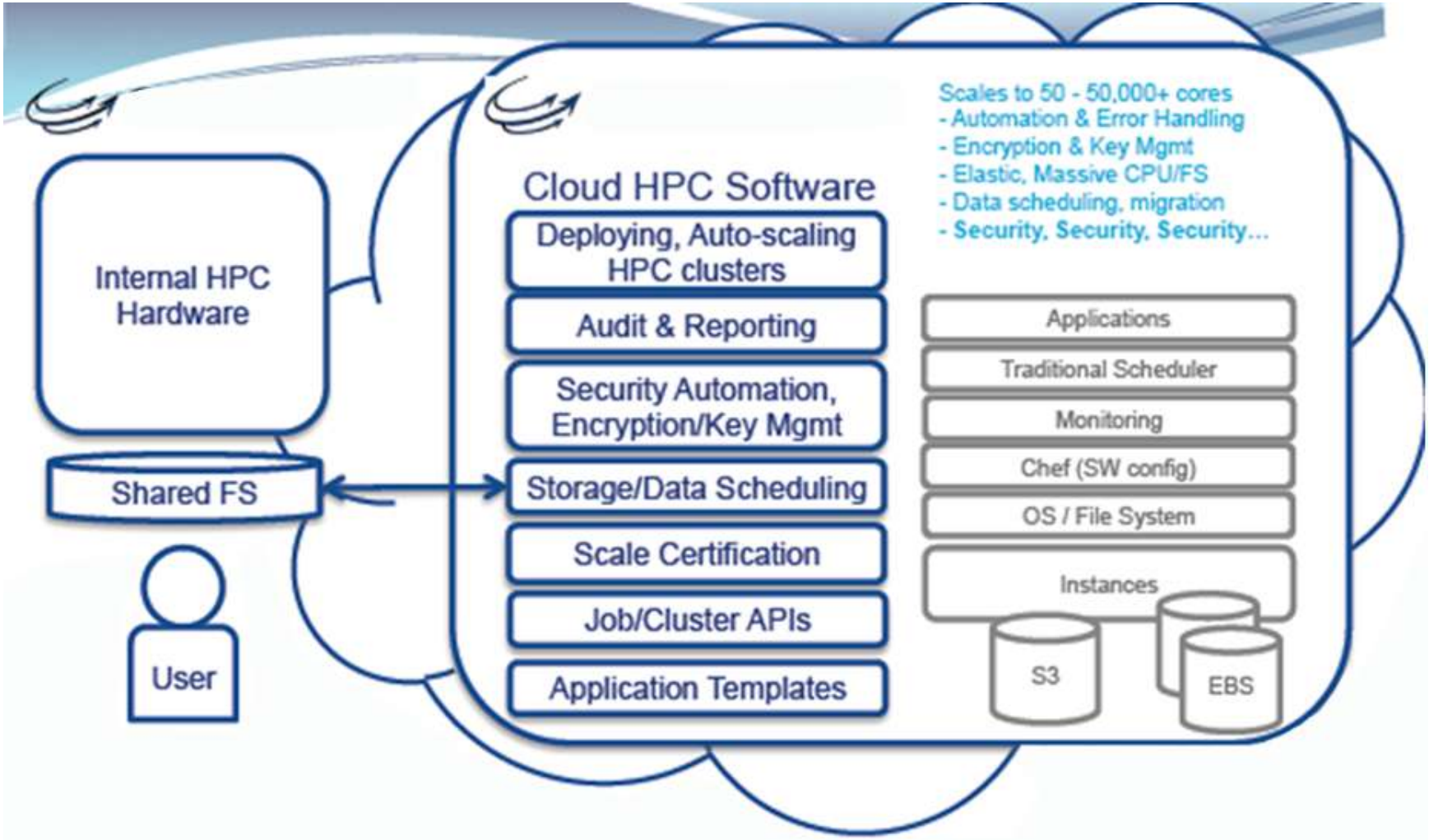
SCHRÖDINGER

Protein
Target

Drug
Candidates

Drug design = finding the few good candidates from millions of compounds

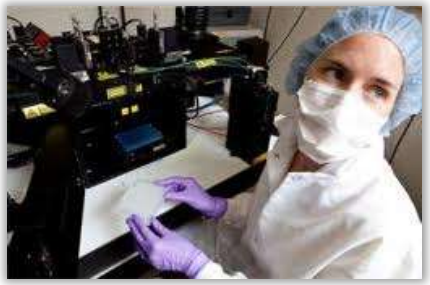
Architecture



Genome Sequencing/Analysis

Shotgun genomics

- Collect samples
- Extract DNA
- Feed into sequencer
- Computationally analyze
- A lab can generate ~100 Gbp in ~1 week for $10k (£6.3K)



Real-World Use Cases

- Real-time pathogen analysis
- Cancer genome analysis => diagnosis & treatment
- Drug resistance in HIV
- Gene expression analysis in agricultural animals
- Microbial community change in response to agriculture or global climate change
- Gene discovery & genome sequencing in non-model organisms

Courtesy Titus Brown, Michigan State University

Medical Diagnosis / Treatment Planning

Typical Medical Practice

- 20-30 providers
- 200,000 patients
- 5 years of data
- 90% of data unstructured
- 4TB scanned images & notes, coded for payment
- 4TB text documents
- 2TB structured documents

Outcomes-Based Medical Diagnosis and Treatment Planning

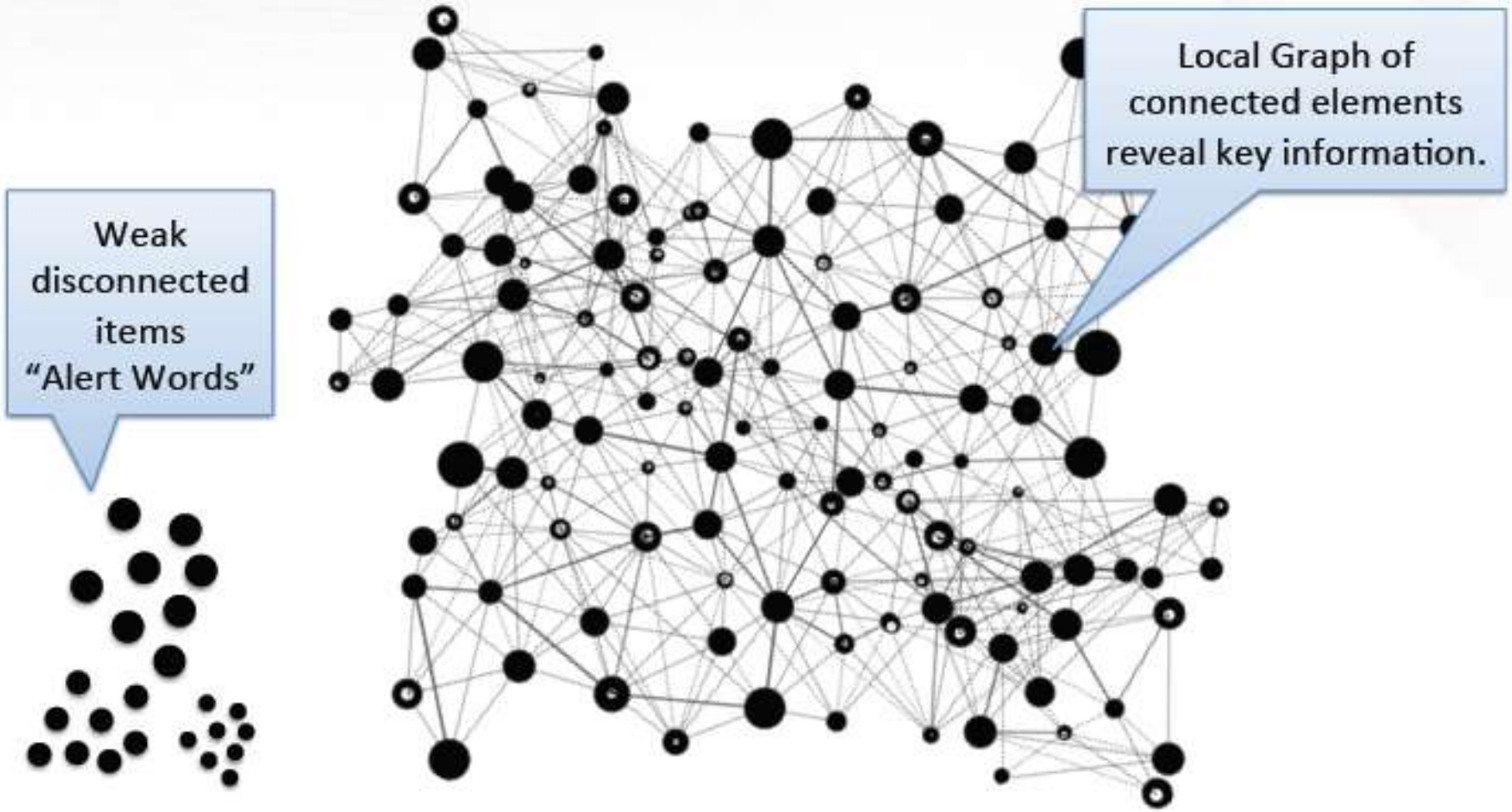
- Enter the patient's history and symptomology.
- While patient is still in the office, sift through millions of archived patient records for relevant outcomes.
- Provider considers the efficacies of various treatments for "similar" patients (but is not bound by the findings).
- Ergo, this functions as a powerful decision-support tool.
- Benefits: better outcomes + rein in costly outlier practices

Optum + Mayo Initiative to Move Past Procedures-Based Healthcare

- New Optum Labs nonprofit research center (Cambridge, MA) to improve healthcare quality and lower costs.
- Data: 100M United Health Group claims (20 years) + 5M Mayo Clinic archived patient records. Option for genomic data.
- Findings will be published.
- Goal: outcomes-based care

Automated Rank & Focus on Key Patterns

- Emcien measures, ranks and organizes the valuable connections to focus on the key nodes and edges



emcien
Converting Data into Value

Use Case: Banking

Fraud Detection

480,000 Items (accounts, codes, locations, etc.)

2.94 million Transactions

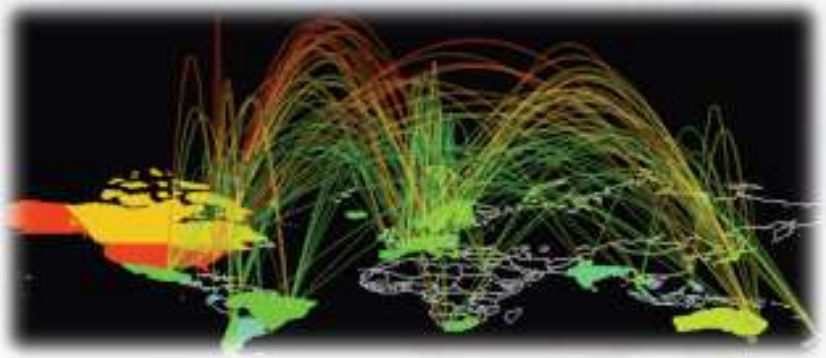
Revealing top suspicious transaction patterns (account numbers, transaction types,..)

1	Acct# 16303	→	Cash Deposit	Trancode 111	9 Strong	75%
4	Trancode 151	→	Check Deposit	Acct# 63286	6 Strong	66.7%

Revealing ID fraud with fuzzy match

emcien
Converting Data into Value

Use Case: Network Security



Network Intrusion Detection

32,000 Items (src & dest. IP address,ports, days, times, activities, etc.)

2.57 million Transactions

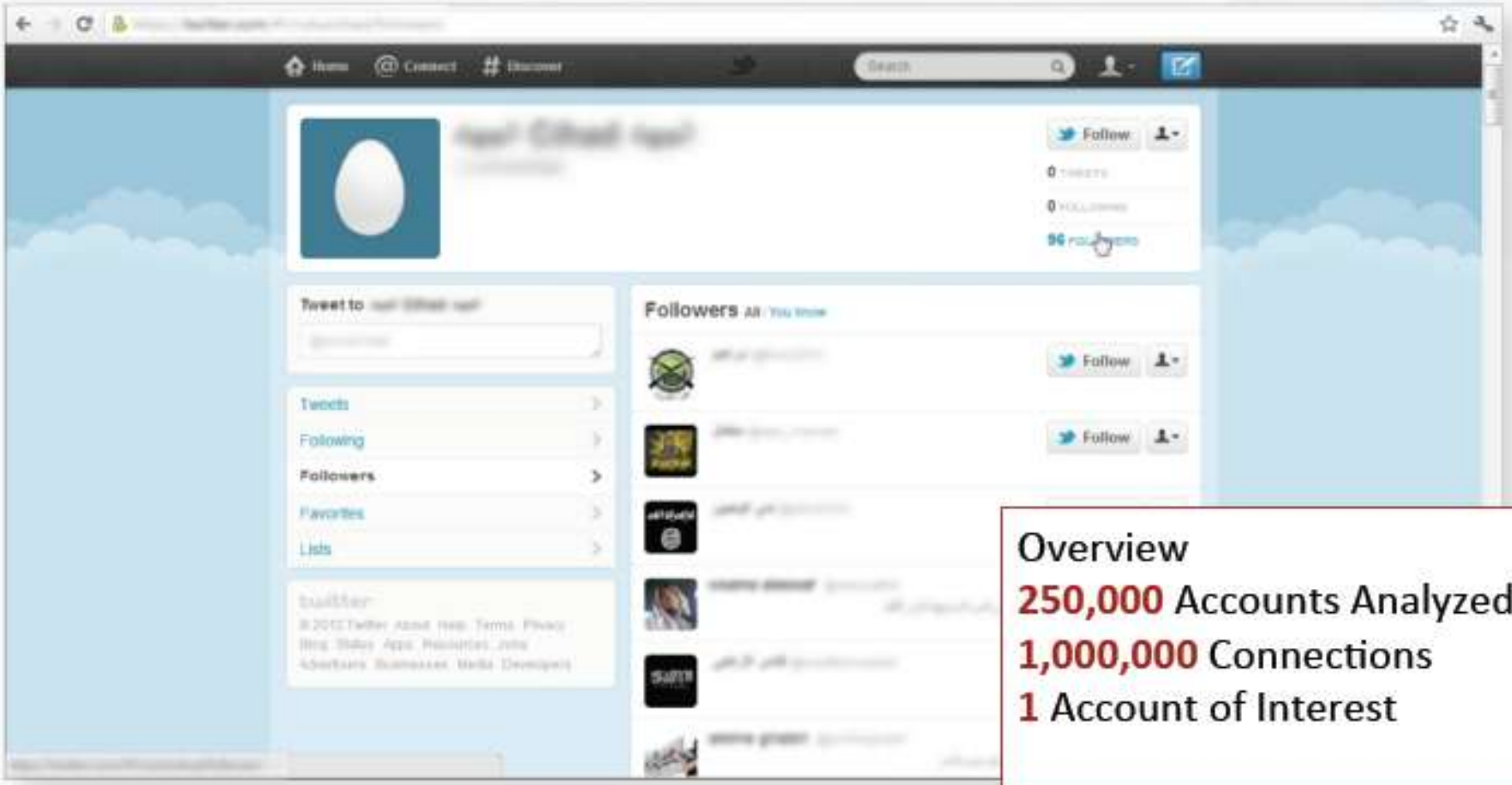
No rules or queries required. Auto-detect intrusion patterns and surface suspicious activity.

1	Src-10.2.197.245 →	Start_hour:9	Thu	Dst-80	6,806	Strong	100%
		Dst-154.241.88.201					
11	Src-10.2.197.245 →	Priority:1	Dst-80	Dst-154.241.88.201	6,804	Strong	100%

emcien
Converting Data into Value

Use Case: Surfacing Sleeper Cell for Intel

The silent signal – Automatically detecting a sleeper cell



Overview

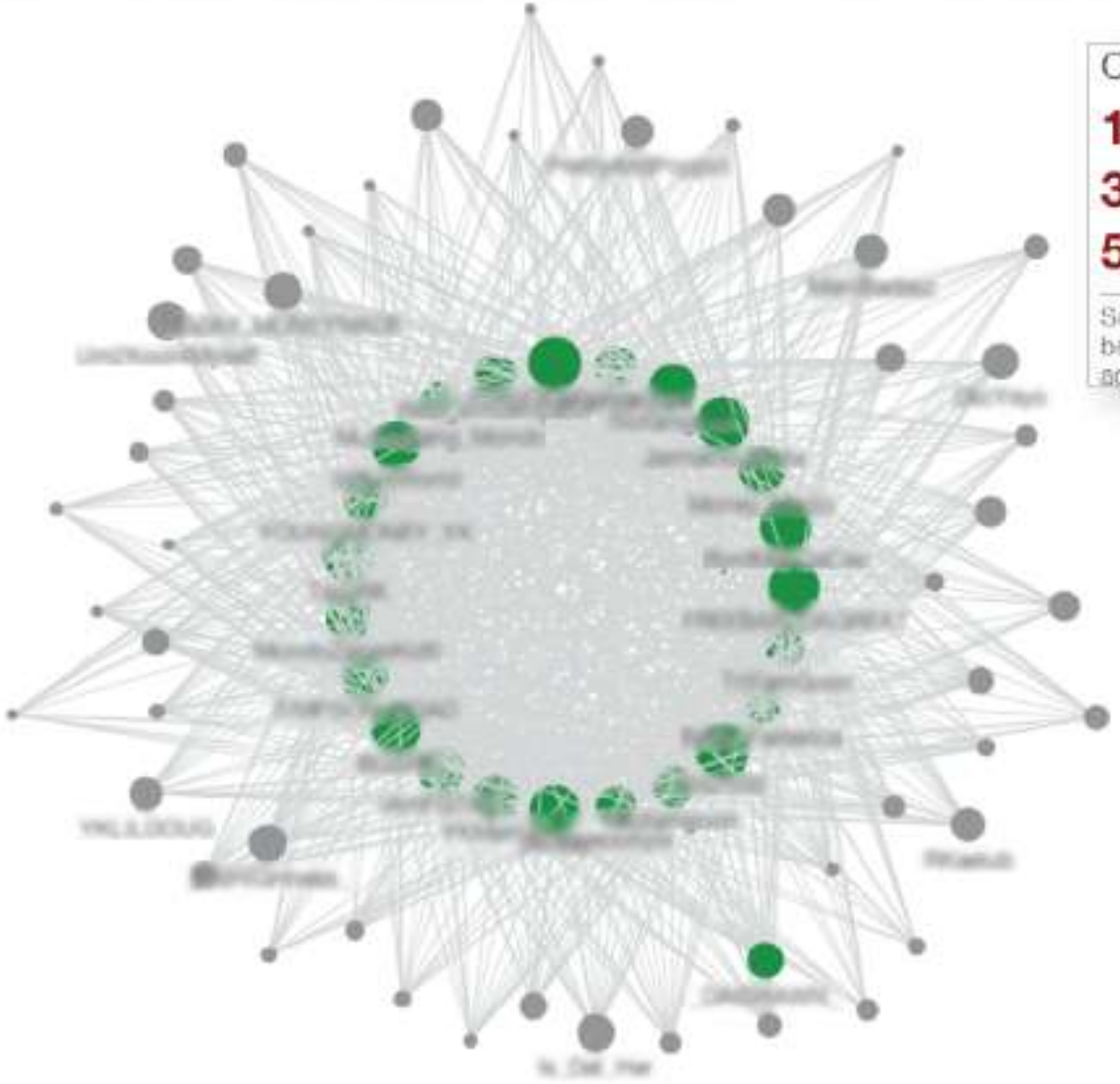
250,000 Accounts Analyzed

1,000,000 Connections

1 Account of Interest

emcien
Converting Data into Value

Use Case: Atlanta Gang Members Network Analysis



Overview

13,941 Accounts Analyzed

39,318 Connections Detected

50 Accounts of Interest

Scout started with **23 Seeds** to detect 39,318 connections between 13,941 accounts. After analysis, Scout found 50 accounts of interest.

Most important accounts: View All | Seeds

23 Seeds:

emcien
Converting Data Into Value

Some Important Challenges

- Identifying important market opportunities
- Data storage (how much to store, where)
- Data movement
- Metadata management
- Tools for large-scale data integration and analysis
- Public databases – who has right to do what?
- Data security (including in the cloud)

Summary: HPDA Market Opportunity

- HPDA: simulation + newer high-performance analytics
  - IDC predicts fast growth from a small starting point
- HPC and high-end commercial analytics are converging.
  - Algorithmic complexity is the common denominator
- Economically important use cases are emerging
  - Which ones will become attractive markets?
- No single HPC solution is best for all problems.
  - Clusters with MR/Hadoop will handle most but not all work (e.g., graph analysis)
- IDC believes our growth estimates could be conservative.