New Directions in IT Architecture:

Achieving Business Value via New Data, Hadoop, and NoSQL

Philip Russom

TDWI Research Director for Data Management

December 2, 2014

Sponsor

splunk>®

Speakers

Philip Russom
TDWI Research Director,
Data Management

James Hodge
Big Data and Analytics
Technology Expert, Splunk

Agenda

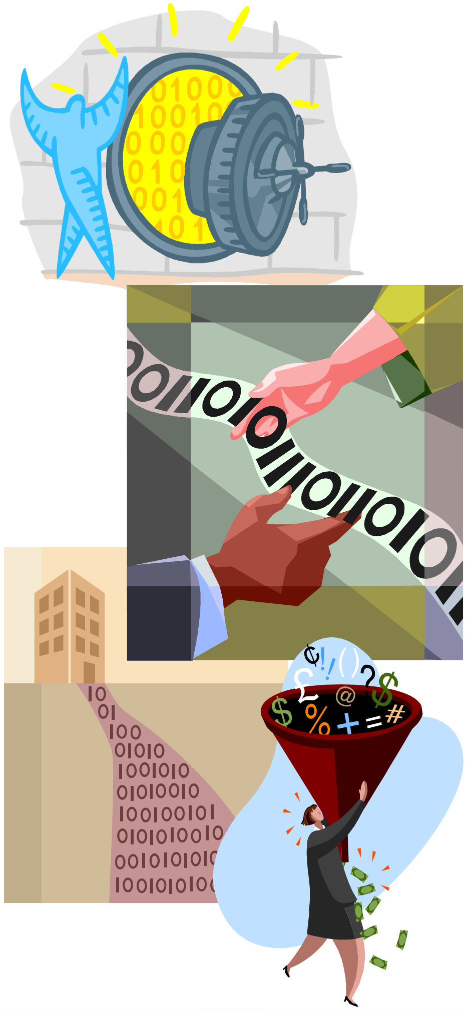
PLEASE TWEET
@pRussom, #TDWI, #DataArchitecture, #Analytics, #RealTime

- Background
  - *Explosion of new big data, most of it from new sources*
  - *Organizations need to leverage new big data*
- Impact on IT Architectures
  - *New big data is diverse*
  - *Some isn’t easily handled by traditional platforms and tools*
- The Solution
  - *New tools and platforms built for diverse new data*
  - *Rising adoption of Hadoop, NoSQL, CEP, analytic, in-memory, real-time functions*
- Recommendations
  - *Update your architectures*
  - *Look for appropriate tools*

Big Data is more than Big

- Big
  - *Multi-terabyte or larger volumes of data*
- Diverse
  - *Many data types*
  - *Many sources*
  - *No schema or evolving*
- New
  - *Types, sources, formats, schema you haven’t tapped much, if at all*
- Fast and furious
  - *Streaming, in real time*
  - *Time sensitive*
- Valuable to the biz
  - *If you handle it well*

Demand Business Value from All Data



- Run the business by the numbers
  - *Requires fresh data, from the best sources, delivered fast, to key people*
- Complete information
  - *Complete customer views, enterprise-scope data, social media, big data…*
- Trusted data
  - *High quality, governed, audit trail*
  - *For reports, analyses, operations, etc.*
- Real-time information
  - *Enables time-sensitive biz practices*
  - *Streaming data sources*
- Business Analytics
  - *predict the future, correlate diverse entities, understand customers, compete on data, etc.*

Use Cases for Big Data Analytics

- Big Data enables exploratory analytics. Discover new:
  - *Customer base segments*
  - *Customer behaviors and their meaning*
  - *Forms of churn and their root causes*
  - *Relationships among customers and products*
- Analyze big data you've hoarded. Finally understand:
  - *Web site visitor behavior*
  - *Product quality based on manufacturing robotic data*
  - *Product movement via RFID in retail*
- Use tools that handle human language for visibility into:
  - *Claims process in insurance*
  - *Medical records in healthcare*
  - *Sentiment analysis in customer-oriented industries*
  - *Call center applications in any industry*
- Big data improves data samples for older analytic apps:
  - *Fraud detection*
  - *Risk management and actuarial calculations*
  - *Anything involving statistics or data mining*
- Big data adds more granular detail to analytic datasets:
  - *Broaden 360-degree views of customers, etc.*
- Streaming big data tells you what just happened:
  - *Plus, what will happen next*
  - *How an event a second ago relates to older events*
  - *New applications in business monitoring, surveillance,*

THE CATCH

Big Data and related practices are Influencing IT Architectures

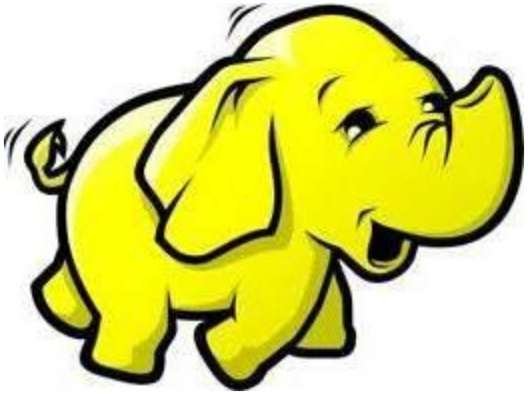
- Achieving the benefits of big data requires change
  - *Changes to existing or upcoming IT architectures and data architectures, plus portfolios of tools and data platforms*
- Why change?
  - *Traditional tools and platforms were designed for structured, relational data at rest*
  - *The same's true of business processes & biz mgt*
- Seems like a problem, but it's an opportunity
  - *Change enables organizations to gain advantages from new, diverse big data*

A COMMON ARCHITECTURAL EVOLUTION FOR BIG DATA

Hadoop integrated with a Relational DBMS

- The strengths of one balance the weaknesses of the other
- A Relational DBMS is good at:
  - *Metadata management*
  - *Complex query optimization*
  - *Query federation*
  - *Table joins, views, keys, etc.*
  - *Security, including roles, directories*
  - *Much more mature development tools*
- HDFS & other Hadoop tools are good at:
  - *Massive scalability*
  - *Lower cost than most DW platforms & analytic DBMSs*
  - *Multi-structured data & no-schema data*
  - *Some ETL functions; late binding; custom code for analytics*
  - *More examples on next slide…*

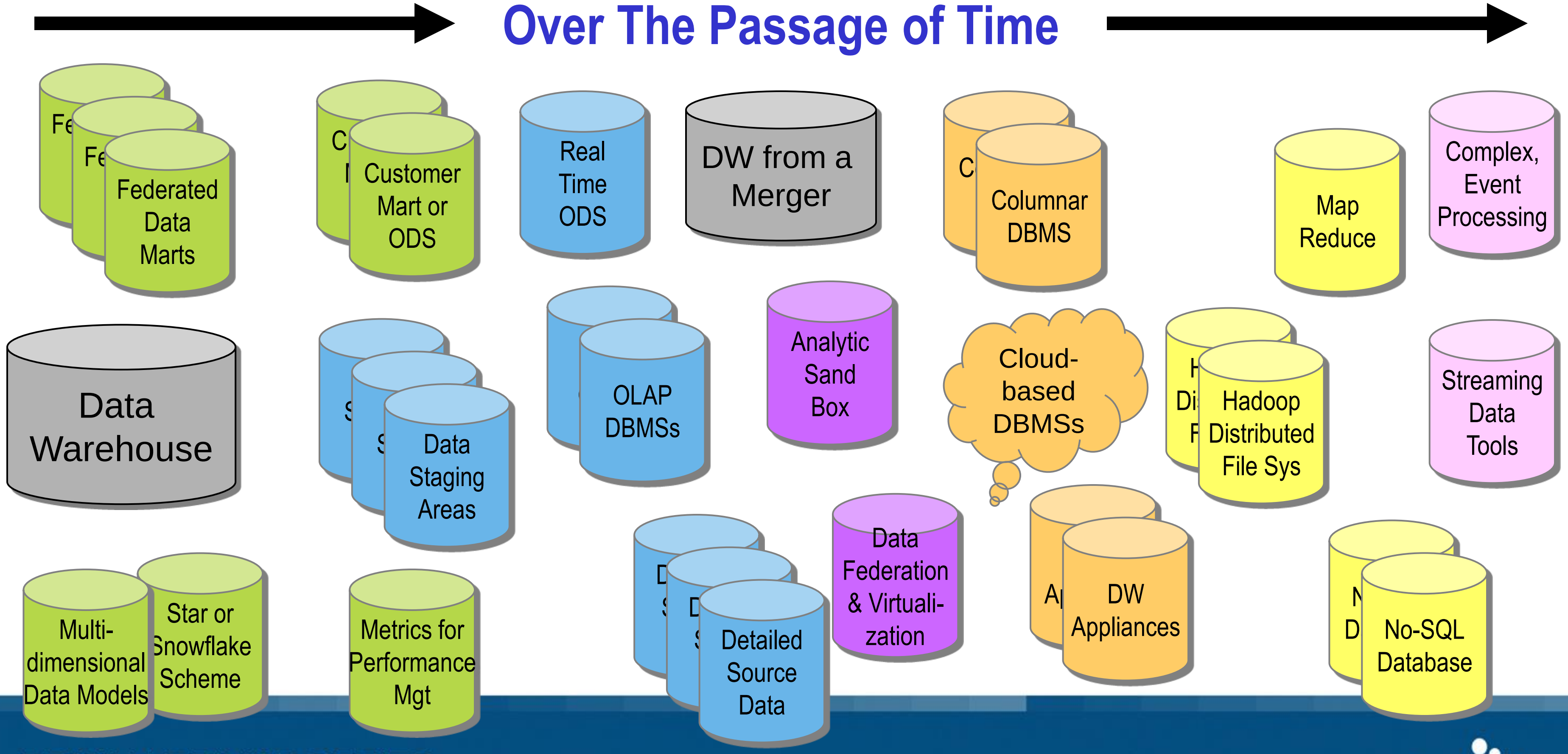
It's not just Data Warehouses. Hadoop has IT Infrastructure uses.



- Data archiving
  - *Most data archives are old and useless*
  - *Hadoop can enable a modern "live archive" that's massively scalable and accessible at any moment by any user*
- Content management
  - *Most "content" is file-based and requires massively scalable search*
  - *Hadoop excels with those, plus adds broad analytics for content*
- Storage as a shared enterprise asset
  - *IT provides SAN/NAS; why not Hadoop?*
  - *Hadoop complements SAN/NAS*

Modern DW System Architectures can be Complex

- The technology stack for DW, BI, analytics, and data integration has always been a multi-platform environment.
- What’s new? The trend toward a portfolio of many data platforms has accelerated. Architecture across them is very important.
- Why do it? More platform types to serve more types of users, data & workloads.



Over The Passage of Time

Federated Data Marts

Customer Mart or ODS

Real Time ODS

DW from a Merger

Columnar DBMS

Map Reduce

Complex, Event Processing

Data Warehouse

Data Staging Areas

OLAP DBMSs

Analytic Sand Box

Cloud-based DBMSs

Hadoop Distributed File Sys

Streaming Data Tools

Multi-dimensional Data Models

Star or Snowflake Scheme

Metrics for Performance Mgt

Detailed Source Data

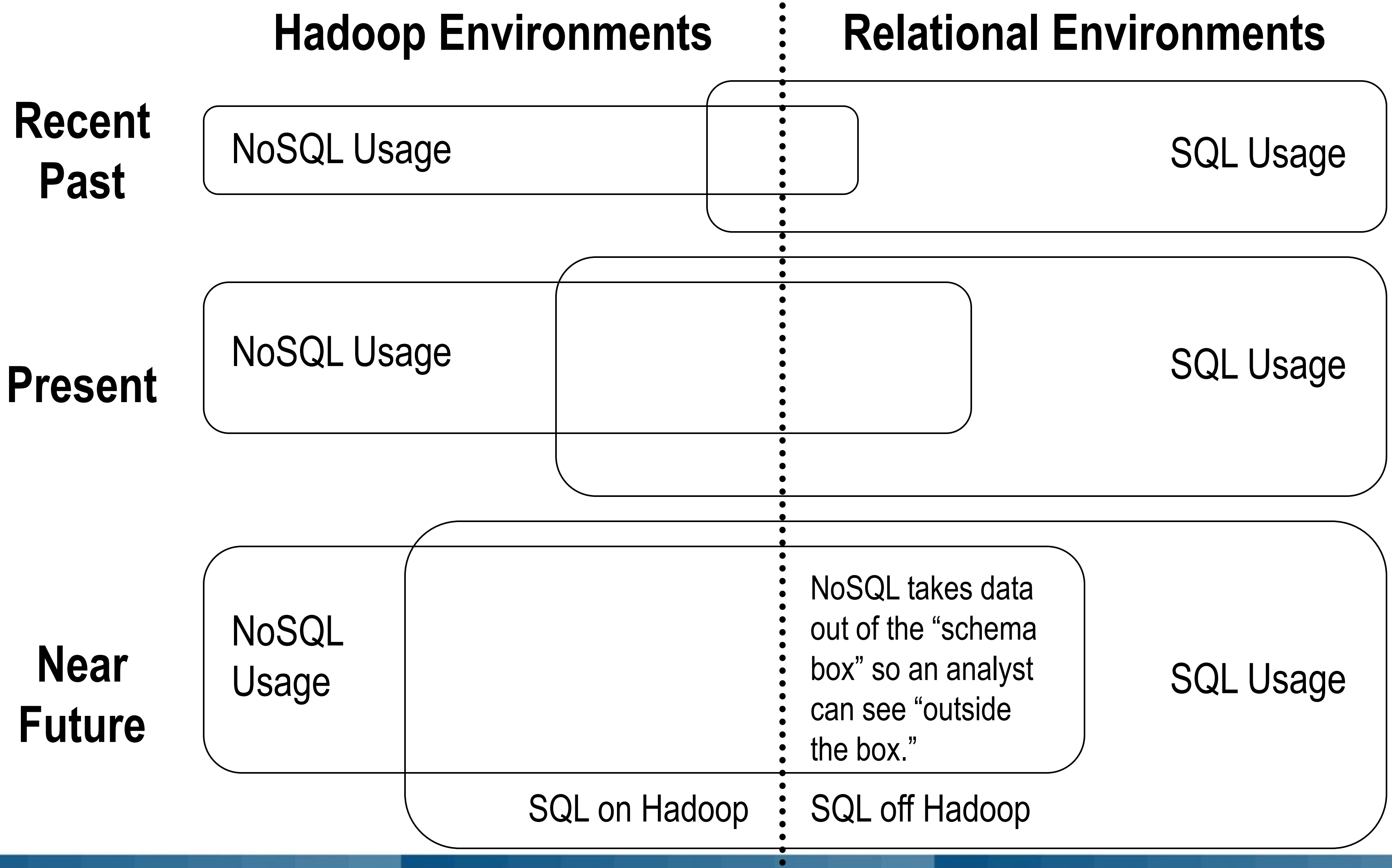
Data Federation & Virtuali-zation

DW Appliances

No-SQL Database

QUERY ARCHITECTURES

NoSQL is joining and complementing SQL



Hadoop Environments

Relational Environments

Recent Past

NoSQL Usage

SQL Usage

Present

NoSQL Usage

SQL Usage

Near Future

NoSQL Usage

NoSQL takes data out of the “schema box” so an analyst can see “outside the box.”

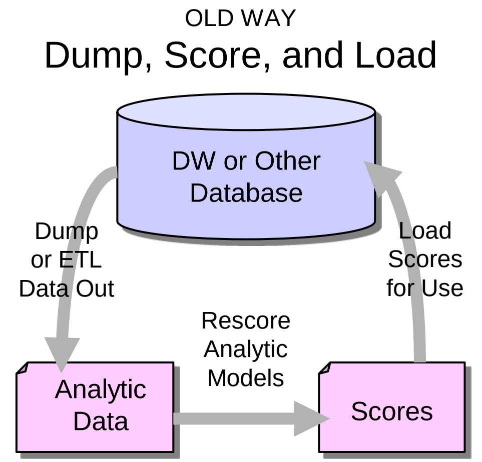
SQL Usage

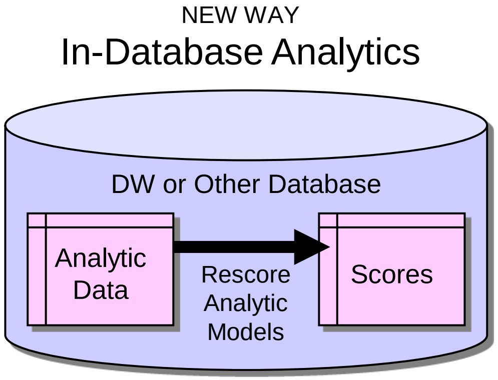
SQL on Hadoop

SQL off Hadoop

AN INFLUENTIAL TREND IN ANALYTIC ARCHITECTURES

Misc Forms of In-Database Analytics





- Old way
  - *Take the data to the analytic algorithm*
  - *Common with ETL & data warehousing*
- New way
  - *Take the analytic algorithm to the data*
  - *As seen in the following examples…*
- In-Database Analytics
  - *Algorithm runs as a UDF or stored procedure in a relation database mgt system (RDBMS)*
  - *Algorithm runs in field programmable gate array (FPGA) in a storage subsystem*
  - *MapReduce job or other algorithm runs on the node(s) of a Hadoop cluster closest to the data that needs processing*
- Big data needs this analytic architecture
  - *Data is now too big to move*

Enterprise Architectures include Clouds

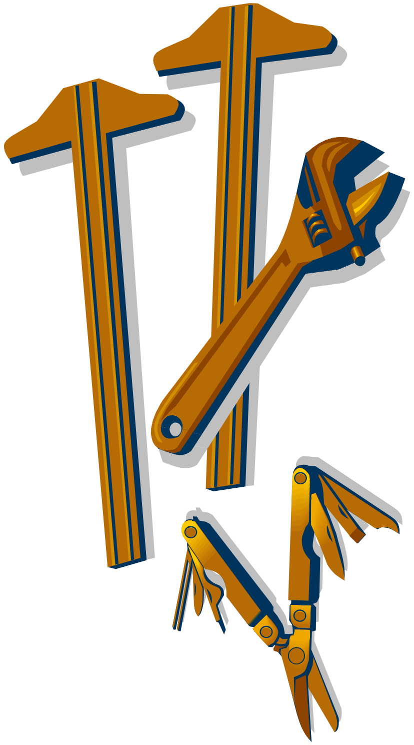
- TDWI sees clouds becoming quite common
  - *Clouds take many forms: public, private, third-party, on premises, as well as combinations of these*
  - *All these are established enterprise or departmental platforms today*
- Benefits of a cloud
  - *More nimble to address changes in the business, evolving capacity requirements, seasonality of data use.*
  - *Cost reductions: less in-house man power for admin & development; Reduces capital expenditures*
  - *Infusion of domain expertise from cloud's professional services*
- Cloud-based data tools and platforms are common today
  - *Data warehouse platforms*
  - *Analytic databases*
  - *Advanced analytics tools*
  - *Text analytics tools*
  - *Hadoop-as-a-service*
  - *Clouds for streaming data*
  - *Wide range of enterprise applications*
  - *Many outsourced data centers are on clouds*

Common Use Cases for Real-Time Technologies

- Enhance complete views of customers with real-time data, BI, and analytics.
- Combine real-time data with historic data from data warehouses and BI systems.
- Understand customer behavior in real-time across multiple channels.
- Evaluate sales performance in near time.
- See a product recurring in abandoned shopping carts on an eCommerce Web site.
- Identify a new social media sentiment or pattern.
- Spot potentially fraudulent activity, even as it's being perpetrated.
- Take logistics to a new level of accuracy, efficiency, and customer service.
- Monitor the performance of interconnected infrastructures, such as utility grids, computer networks, and manufacturing facilities.
- Let software take action automatically.

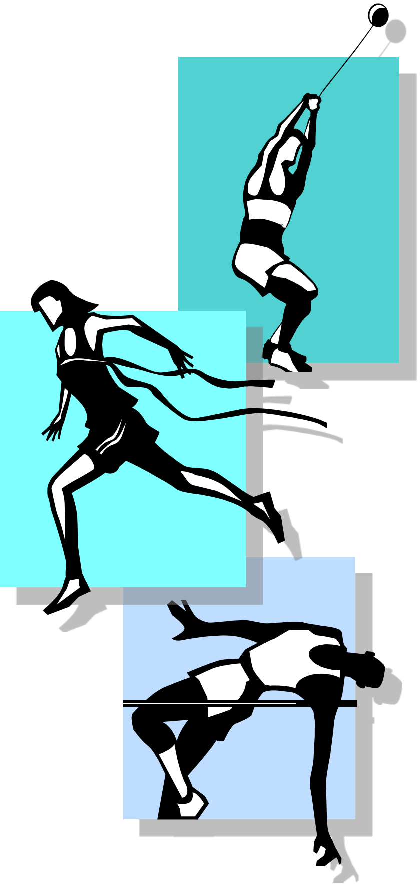
USERS ARE ADJUSTING MISC ARCHITECTURES TO INCLUDE

Technologies for Real-Time Operations



- Data federation and virtualization
- Data replication and data synchronization
- Batch and micro batch
- Complex event processing (CEP)
- In-database analytics
- In-memory databases
- Columnar DBMS
- Hadoop, Spark, Stinger, Storm, etc.
- Cloud-based real-time solutions
- Massively parallel processing (MPP)
- Solid-state drives & other hardware upgrades

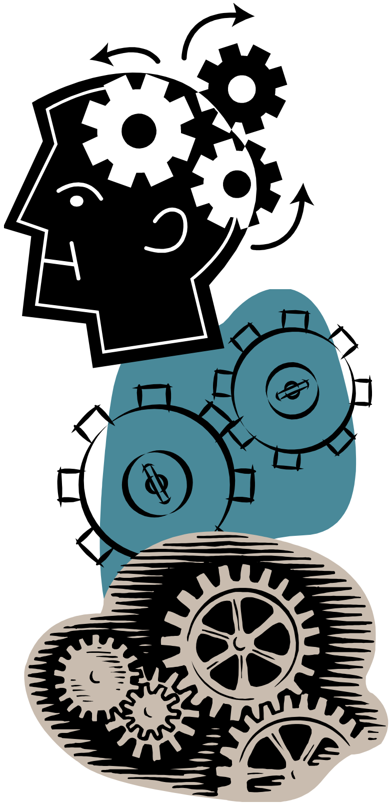
Complex Data Ecosystems can be Agile



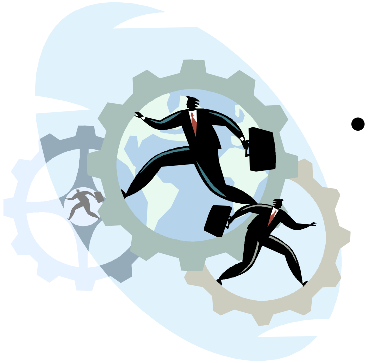
- New tools automate the whole analytic process:
  - *Explore, discover, analyze, and visualize data in one seamless user interface*
- New tools enable new data modeling practices
  - *Nowadays, you can't take 90 days to model a dataset*
  - *Today, build the model as you explore and discover data*
  - *Retain raw source data, then build models as needed*
  - *Work with data that lacks fixed schema, but discover schema, build schema on the fly, and/or let the tool suggest schema and metadata to you*
- Cloud and SaaS
  - *Short time to use: Hadoop, analytics, DBMSs…*
  - *Automatic maintenance and no system integration*
  - *Frees up time for more development*

New Architectures Present Challenges

- Lots of moving parts:
  - *Complexity is high, in multi-platform data ecosystems*
    - Numerous platforms, tools, datasets; assume considerable data movement among these
  - *Heterogeneity is high*
    - Many types of platforms: relational, row stores, columnar, appliances, file systems, Hadoop, NoSQL…
    - Multiple vendors, open source, home grown
  - *Data is remarkably diverse*
    - Structured, relational, records, mixed structure, hierarchical, evolving schema, no schema, text, streaming, machine data…
  - *Processing is also diverse*
    - Relational analytics, non-relational analytics, reporting, ETL, ELT, federation, virtualization, query, search, data exploration and discovery, visualization…



New Tools for Complex Data Ecosystems



- Look for tool that focus on:
  - *Designing and architecting a “big picture”*
  - *Interoperability among diverse systems and data types*
  - *Data operations optimized across multiple platforms*
- Features that help with complex architectures:
  - *Distributed queries and search*
  - *Distributed administration and management*
  - *Easy ingestion of new data, whether streaming or static*
  - *Monitor and alert for more automation*
  - *Real-time indexing, to keep pace with data ingestion*
  - *High performance, even with multiple platforms*
  - *High availability – one down system affects many multi-system processes*
  - *Single-sign-on security, despite multiple systems*

Recommendations

- Prepare to leverage big data, analytics, real time, and other opportunities by…
  - *Adjusting your architectures for IT, hardware, data, data warehousing, analytics, interoperability and integration…*
- Multiple architectures need attention
  - *Big data, Analytics, Real-time*
  - *Queries, computing, in-memory, in-database, IT infrastructure*
  - *Include new platforms, like Hadoop, NoSQL, CEP, clouds, SaaS…*
- Look for tools that make developers agile, despite complexity:
  - *Explore, discover, analyze, and visualize in one user interface*
  - *Model data on-the-fly, instead of weeks of offline work*
  - *Cloud and SaaS tools or platforms for quick time to use, low admin*
- Look for tool features that assist with multi-platform ecosystems
  - *Top of the list: Distributed query, search, and admin; support both data in motion and at rest; high performance & availability*

New Directions in IT Architecture:

Achieving Business Value via New Data, Hadoop, and NoSQL

James Hodge

Big Data and Analytics Technology Expert

Splunk Company Overview

Company

- Global HQs:
  - San Francisco
  - London
  - Hong Kong
- 1,200 employees globally
- Annual Revenue: $302.6M (YoY +52%)
- NASDAQ: SPLK

Products

- Free trial to massive scale
- Splunk products:
  - Splunk Enterprise
  - Splunk Cloud
  - Hunk
  - Splunk MINT
  - Premium Apps

Customers

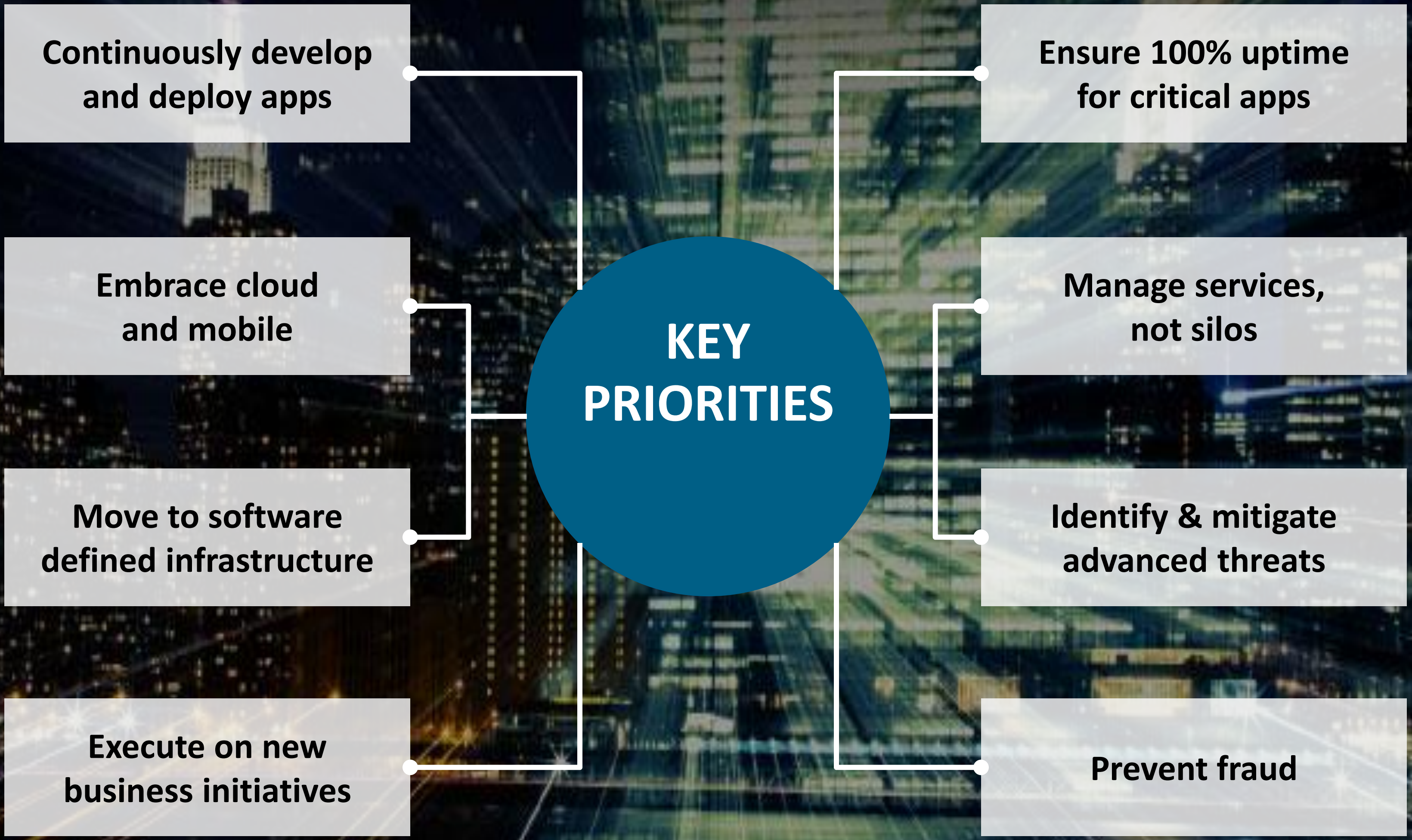
- 8,400+ customers
- Across 100 countries
- Small to large organizations
- 70+ of the Fortune 100
- Largest license:
  - 400+ Terabytes/day

Big Data Comes from Machines

Volume | Velocity | Variety | Variability

GPS,
RFID,
Hypervisor,
Web Servers,
Email, Messaging,
Clickstreams, Mobile,
Telephony, IVR, Databases,
Sensors, Telematics, Storage,
Servers, Security Devices, Desktops

Insights Deliver Competitive Advantage



KEY PRIORITIES

Continuously develop and deploy apps

Embrace cloud and mobile

Move to software defined infrastructure

Execute on new business initiatives

Ensure 100% uptime for critical apps

Manage services, not silos

Identify & mitigate advanced threats

Prevent fraud

splunk>

Make machine data accessible, usable and valuable to everyone.

Why Splunk?

FAST TIME-TO-VALUE

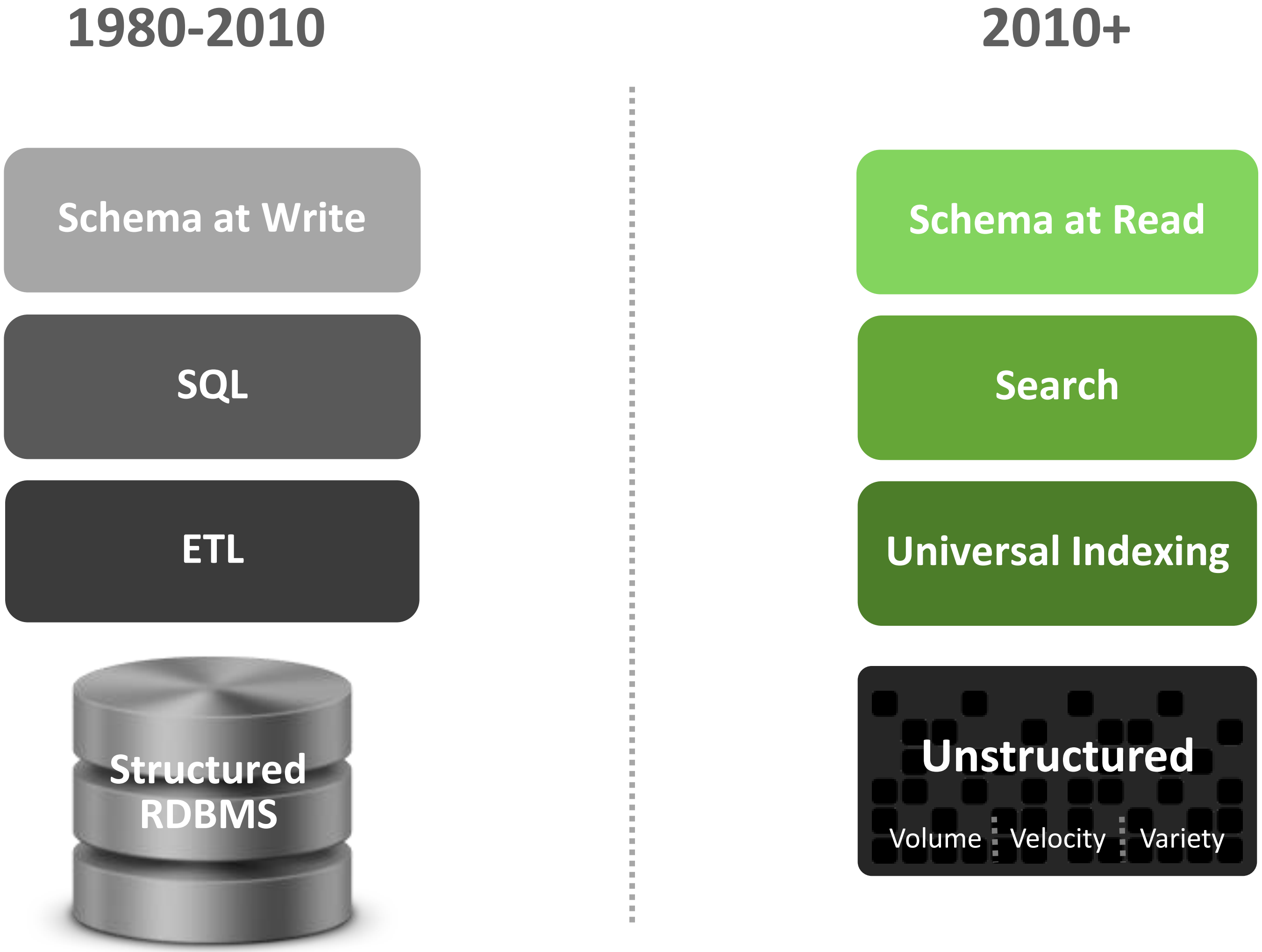
ONE PLATFORM, MULTIPLE USE CASES

VISIBILITY ACROSS STACK, NOT JUST SILOS

ASK ANY QUESTION OF DATA

ANY DATA, ANY SOURCE

Disruptive Approach to Unstructured Data



1980-2010

Schema at Write

SQL

ETL

Structured RDBMS

2010+

Schema at Read

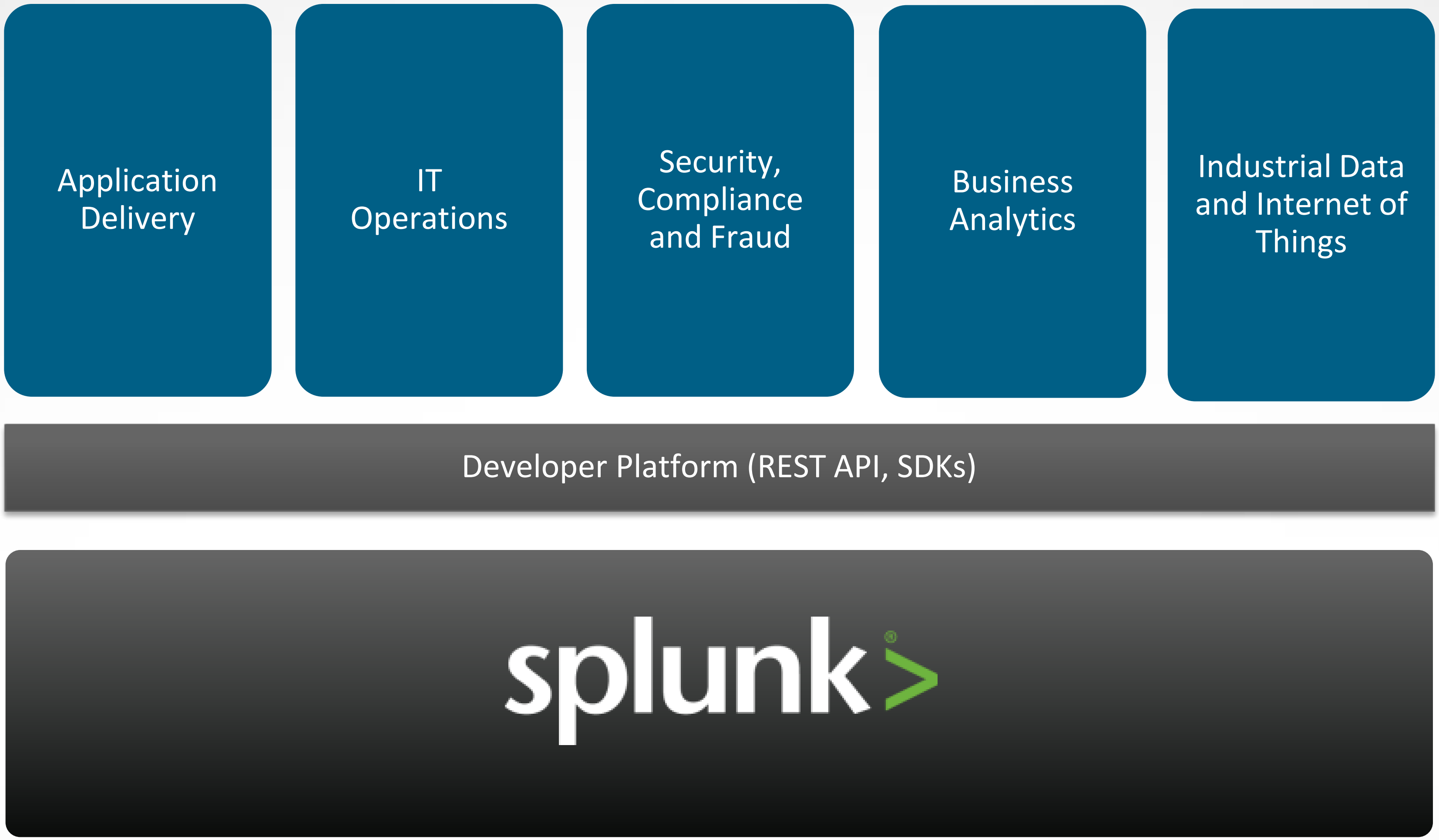
Search

Universal Indexing

Unstructured

Volume | Velocity | Variety

Delivers Value Across IT and the Business



Application Delivery

IT Operations

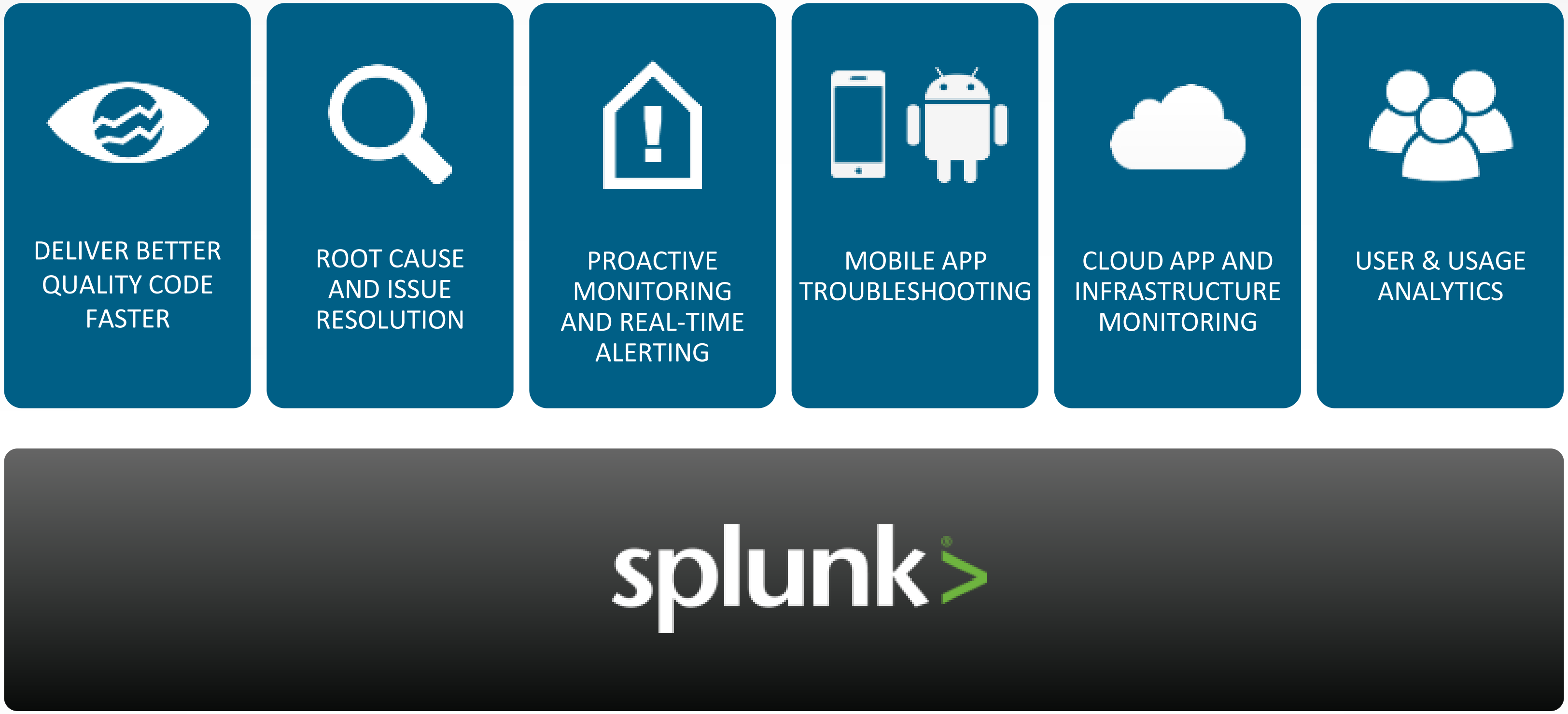
Security, Compliance and Fraud

Business Analytics

Industrial Data and Internet of Things

Developer Platform (REST API, SDKs)

Platform for Application Delivery and IT Operations



DELIVER BETTER QUALITY CODE FASTER

ROOT CAUSE AND ISSUE RESOLUTION

PROACTIVE MONITORING AND REAL-TIME ALERTING

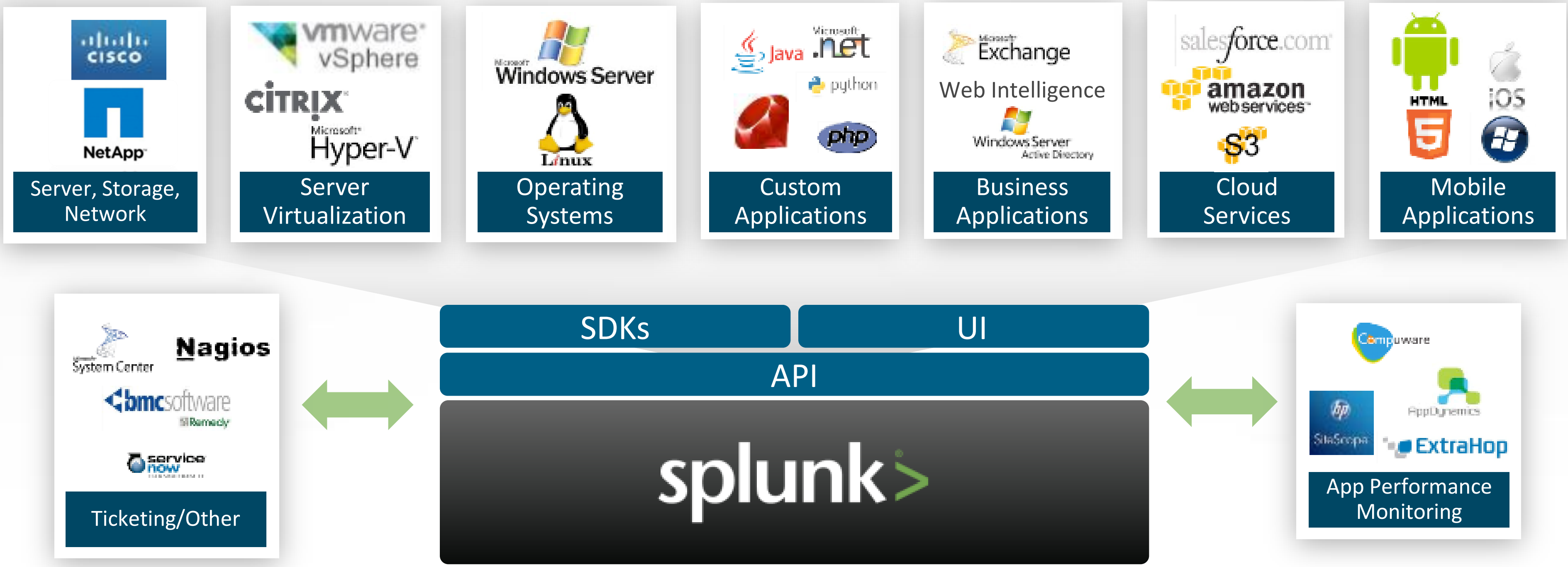
MOBILE APP TROUBLESHOOTING

CLOUD APP AND INFRASTRUCTURE MONITORING

USER & USAGE ANALYTICS

splunk>

Application Delivery & IT Ops Landscape



Server, Storage, Network

Server Virtualization

Operating Systems

Custom Applications

Business Applications

Cloud Services

Mobile Applications

SDKs

UI

API

Ticketing/Other

App Performance Monitoring

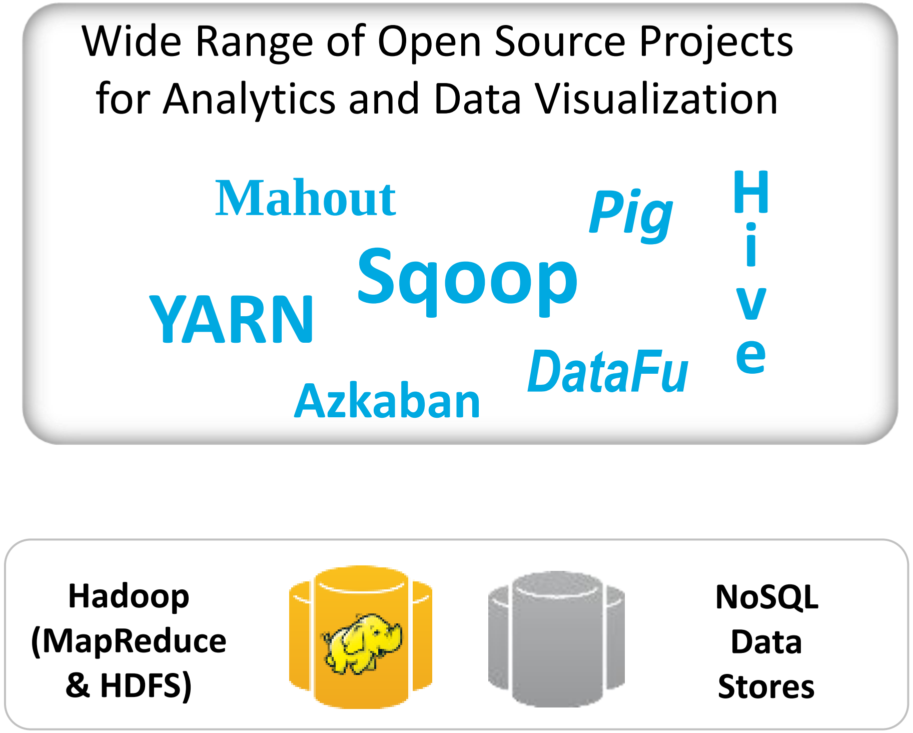
Better Code, Faster Development and Migration to Cloud

- Reduced error rates by 2 orders of magnitude in a couple of weeks
- Rapidly found and fixed one line of code responsible for 30,000+ errors
- Real-time dashboards on error rates and production impact
- In-depth visibility as they strategically migrate apps to AWS Cloud

Bring the algorithm to the data

It’s Hard to Turn Raw Data into Refined Insights

- **Open source offers simple storage but hard analytics:** difficult to explore, analyze, visualize
- **Hard-to-staff skills:** require months of labor by specialists with rare and expensive skill sets
- **Inflexible approaches:** must predefine fixed schemas or program MapReduce jobs



Wide Range of Open Source Projects for Analytics and Data Visualization

Mahout

Pig

Hive

Sqoop

YARN

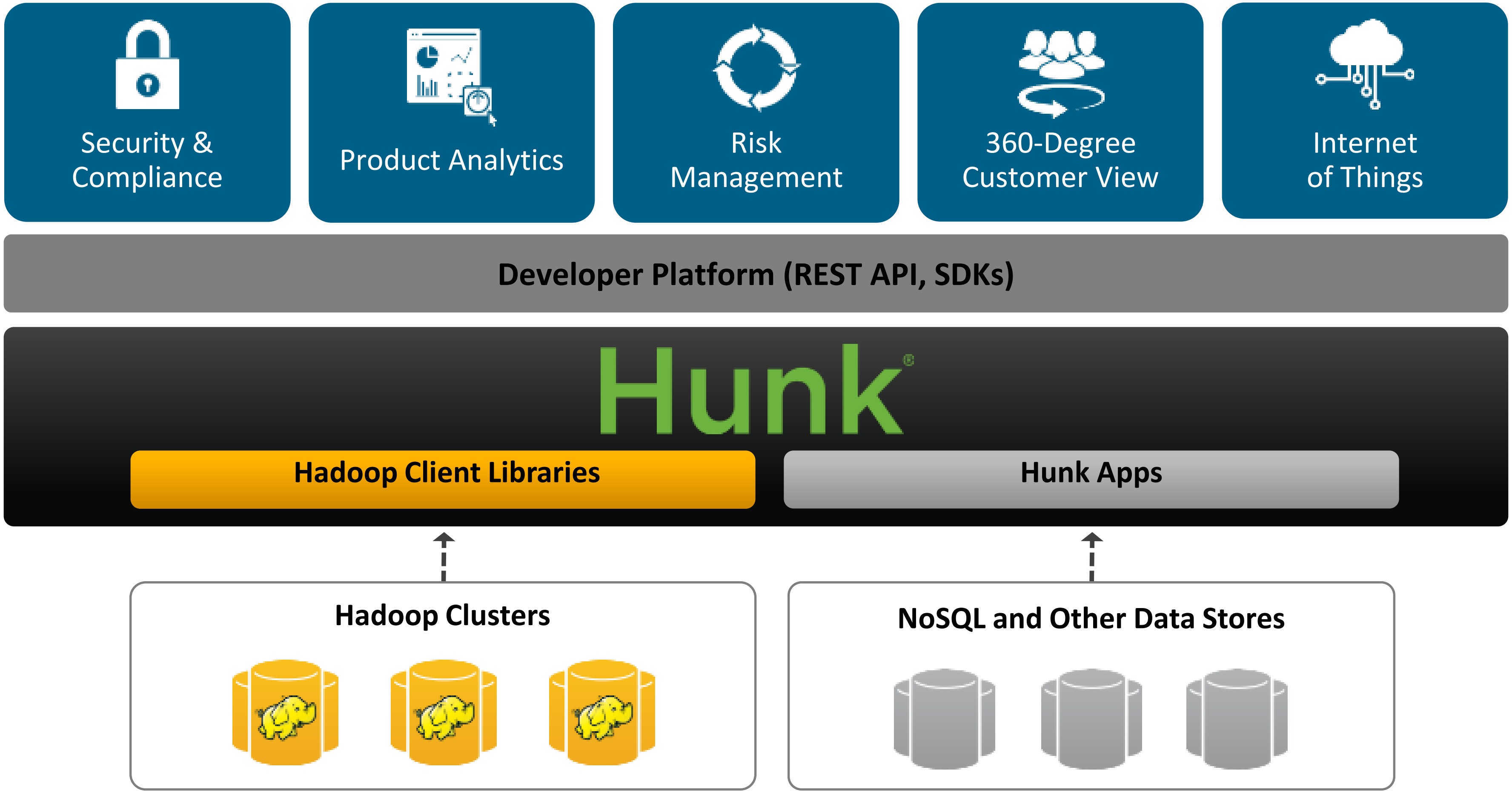
DataFu

Azkaban

Hadoop (MapReduce & HDFS)

NoSQL Data Stores

Explore, Analyze, Visualize Data in Hadoop, NoSQL



Security & Compliance

Product Analytics

Risk Management

360-Degree Customer View

Internet of Things

Developer Platform (REST API, SDKs)

Hunk®

Hadoop Client Libraries

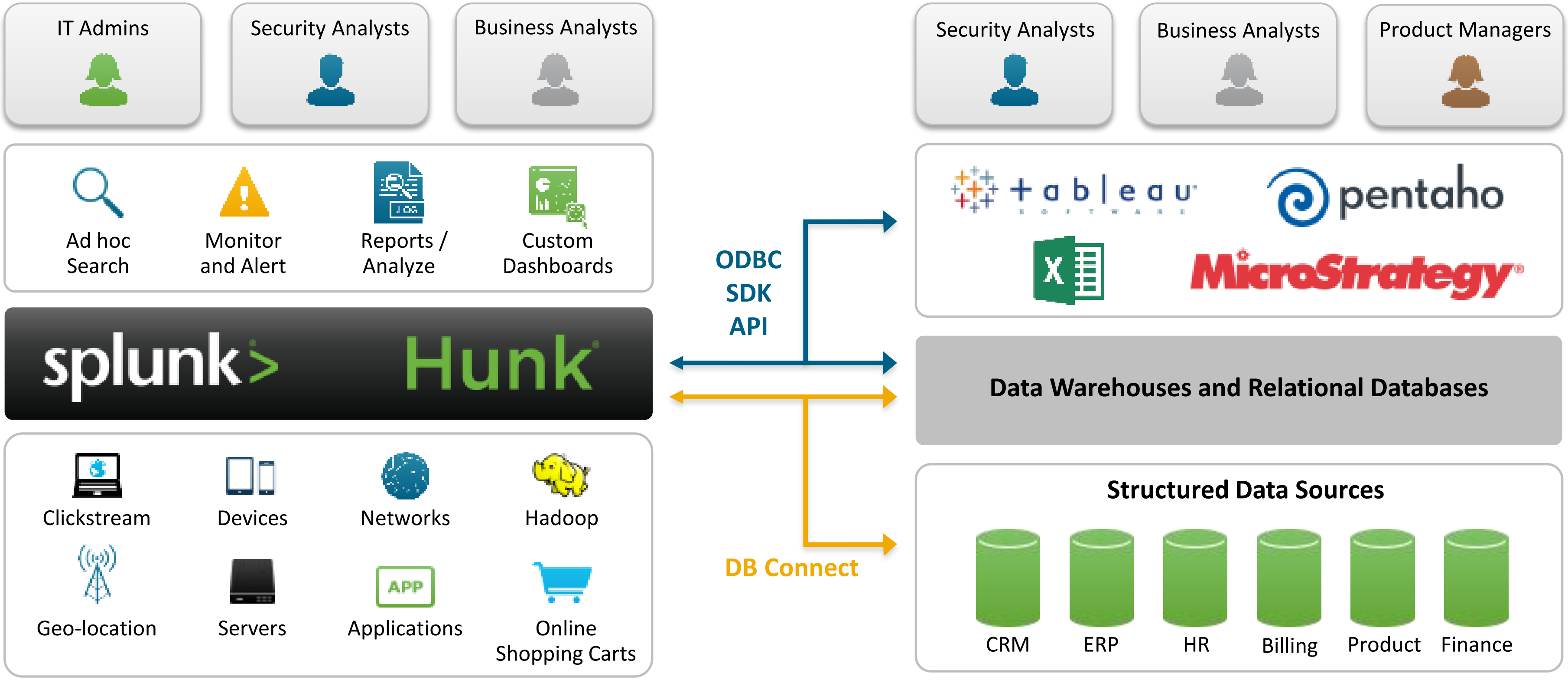
Hunk Apps

Hadoop Clusters

NoSQL and Other Data Stores

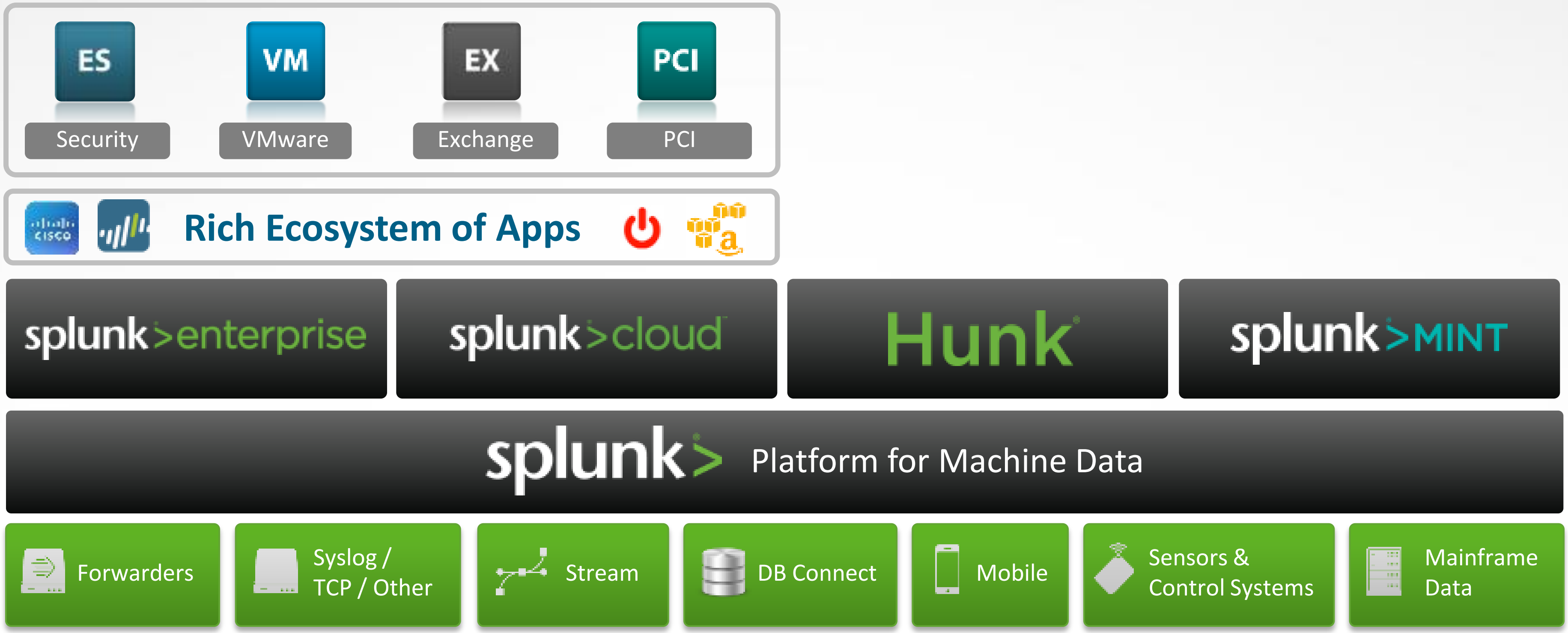
Data Platform in Enterprise Architecture

Machine Data Platform in Enterprise Architecture



Easy to Adopt Splunk

Across Data Sources, Use Cases & Consumption Models



Thank you

Questions?

Contact Information

If you have further questions or comments:

Philip Russom, TDWI
prussom@tdwi.org

James Hodge, Splunk
jhodge@splunk.com