How streaming is dramatically changing our analytic world

With Hortonworks DataFlow (HDF)

Kabir Paul - Solutions Engineer

Keith Symons - Account Enterprise Representative

Agenda

- Overview of Hortonworks
- Hortonworks Connected Data Architecture
- Streaming Solutions using Hortonworks Data Flow (HDF)
- HDF Customer Stories & Use Cases
- Q&A

HORTONWORKS®
POWERING THE FUTURE OF DATA™

The Hortonworks Opportunity

At the core of ~$1.9T in market opportunity over the next 5 years

Big Data
~$210 B

Cloud
~$410 B

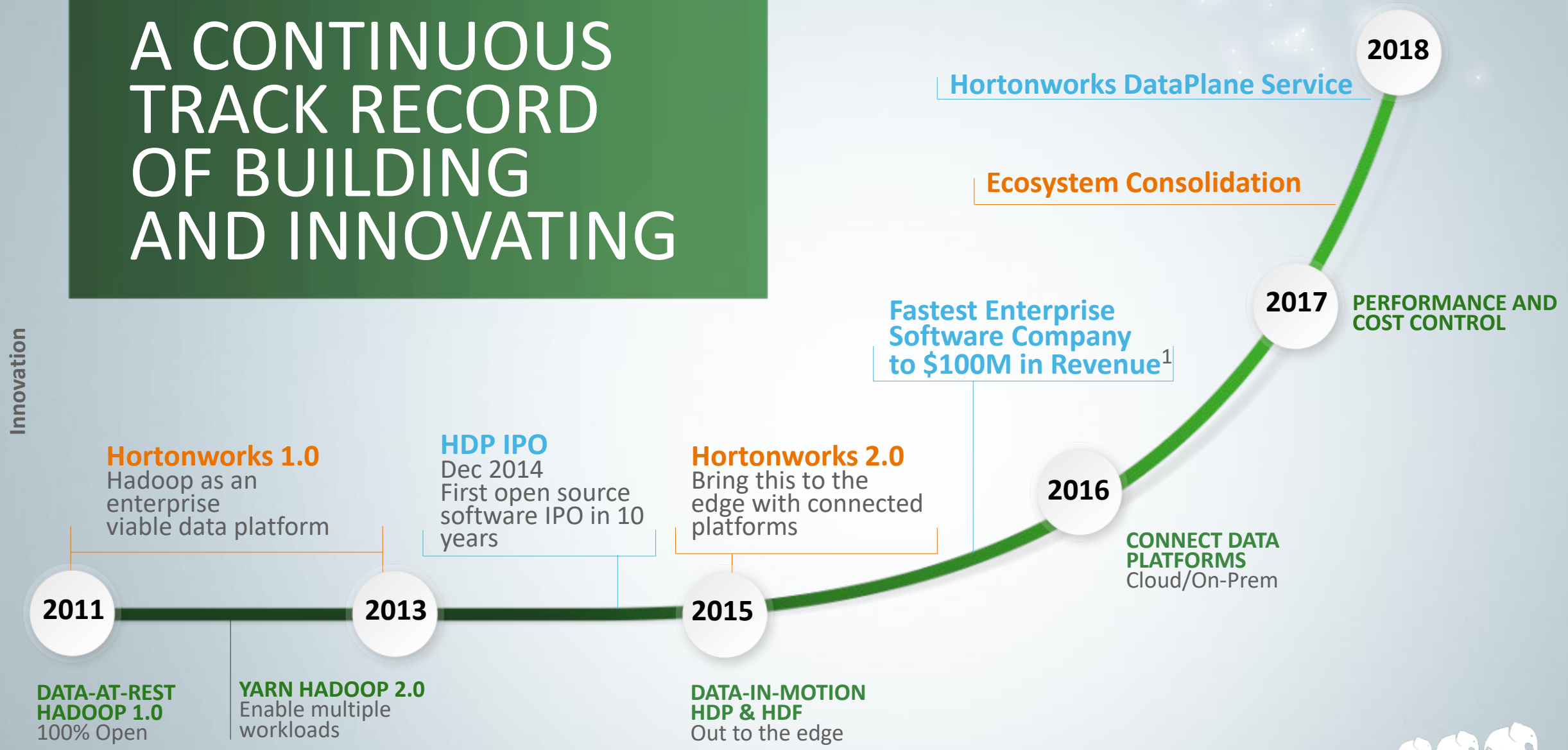
IoT
~$1.1 T

Data Science
~$180 B

Streaming
~$1.65 B

Sources: IDC Worldwide Big Data and Analytics Software Forecast, 2017-2021, Forecasts Continuous/Streaming Analytics revenue to be $1.65B by 2021, July, 2017; Data Science Platform market size to reach $183.7B by 2023, Allied Market Research, Data Science Platform Market by Type and End User: Global Opportunity and Forecast, 2017-2023; IDC Worldwide Semiannual Big Data and Analytics Spending Guide Update, Forecasts Big Data & Business Analytics revenues to be $210B by 2020, Press Release March 2017; Gartner Worldwide Public Cloud Services Revenue, Forecasts Public Cloud Services Revenue to be $411.4B by 2020, Press Release October 2017; IDC Worldwide Semiannual IoT Spending Guide Update, Forecasts Worldwide IoT Spending forecast to be ~$1.1T by 2021, Press Release December 2017.

A CONTINUOUS TRACK RECORD OF BUILDING AND INNOVATING

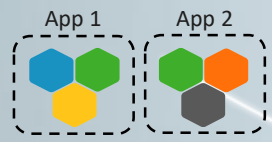


[1] Source: Barclays Big Data Data Report, July 10, 2015

HORTONWORKS

Market Drivers

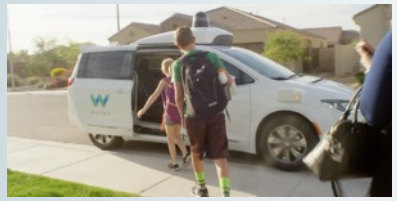
FASTER



Faster time to deployment
(Containerized Micro-Services)

Release Agility
(De-coupled HDP Components)

BIGGER



Infinitely Scalable
(Billions of files, Exabytes)

Low TCO
(Less Storage Overhead)

SMARTER

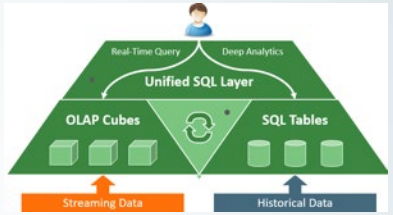
Deep Learning frameworks
(TensorFlow, Caffe)

GPU Pooling/Isolation

TRUSTED

Data Swamp->Data Lake

REAL-TIME SQL



One SQL Layer
(Across Historical, Real-time)

HYBRID

S3, ADLS/WASB, GCS with Truly Incremental Replication

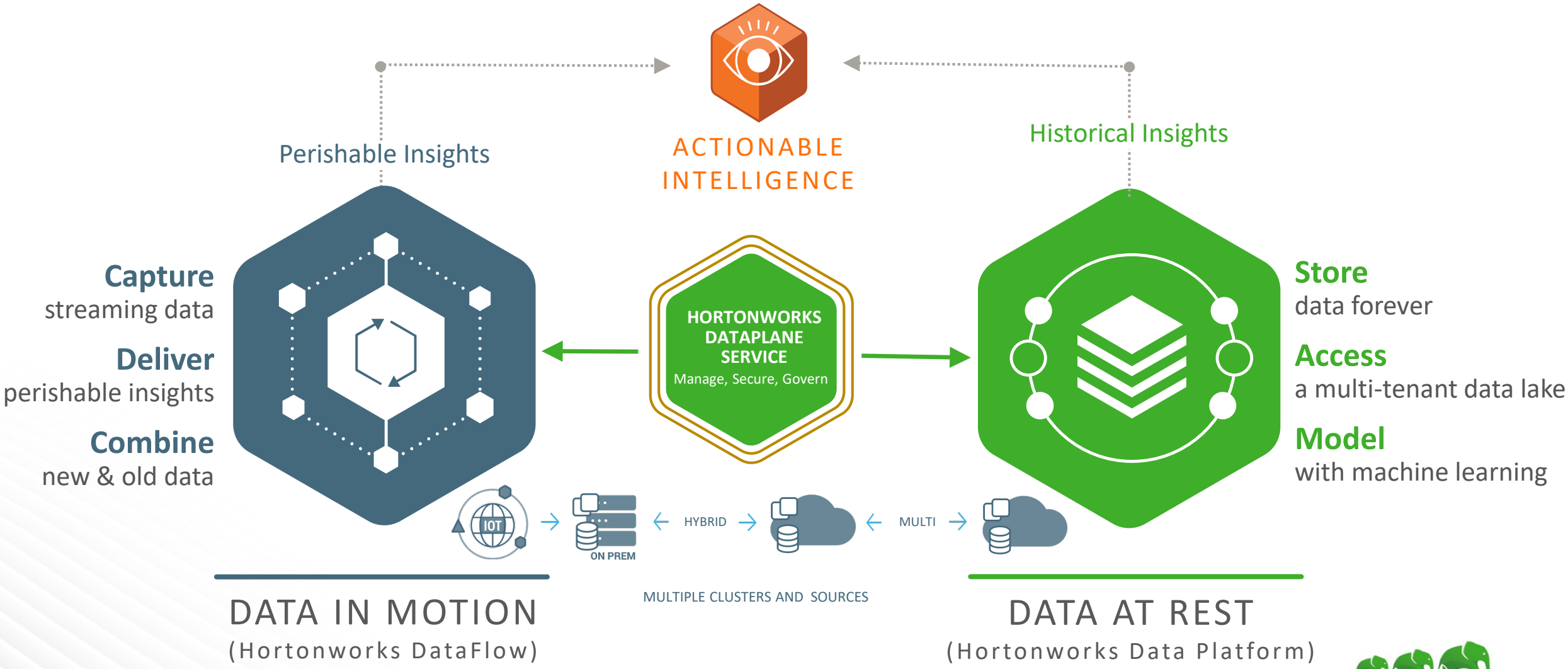
HORTONWORKS®

Agenda

- Overview of Hortonworks
- Hortonworks Connected Data Architecture
- Streaming Solutions using Hortonworks Data Flow (HDF)
- HDF Customer Stories & Use Cases
- Q&A

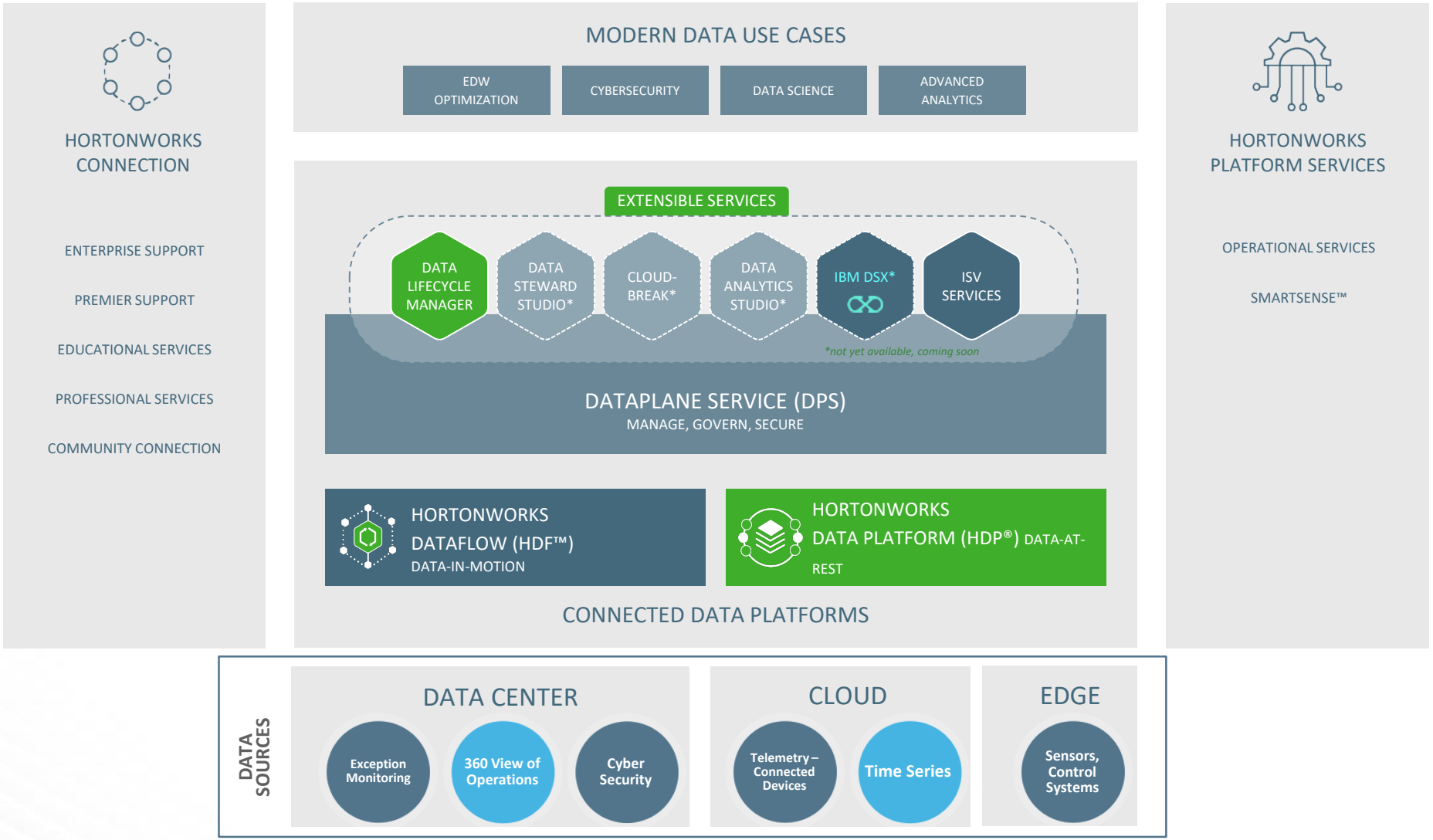
HORTONWORKS®
POWERING THE FUTURE OF DATA™

A Connected Data Strategy Solves for All Data



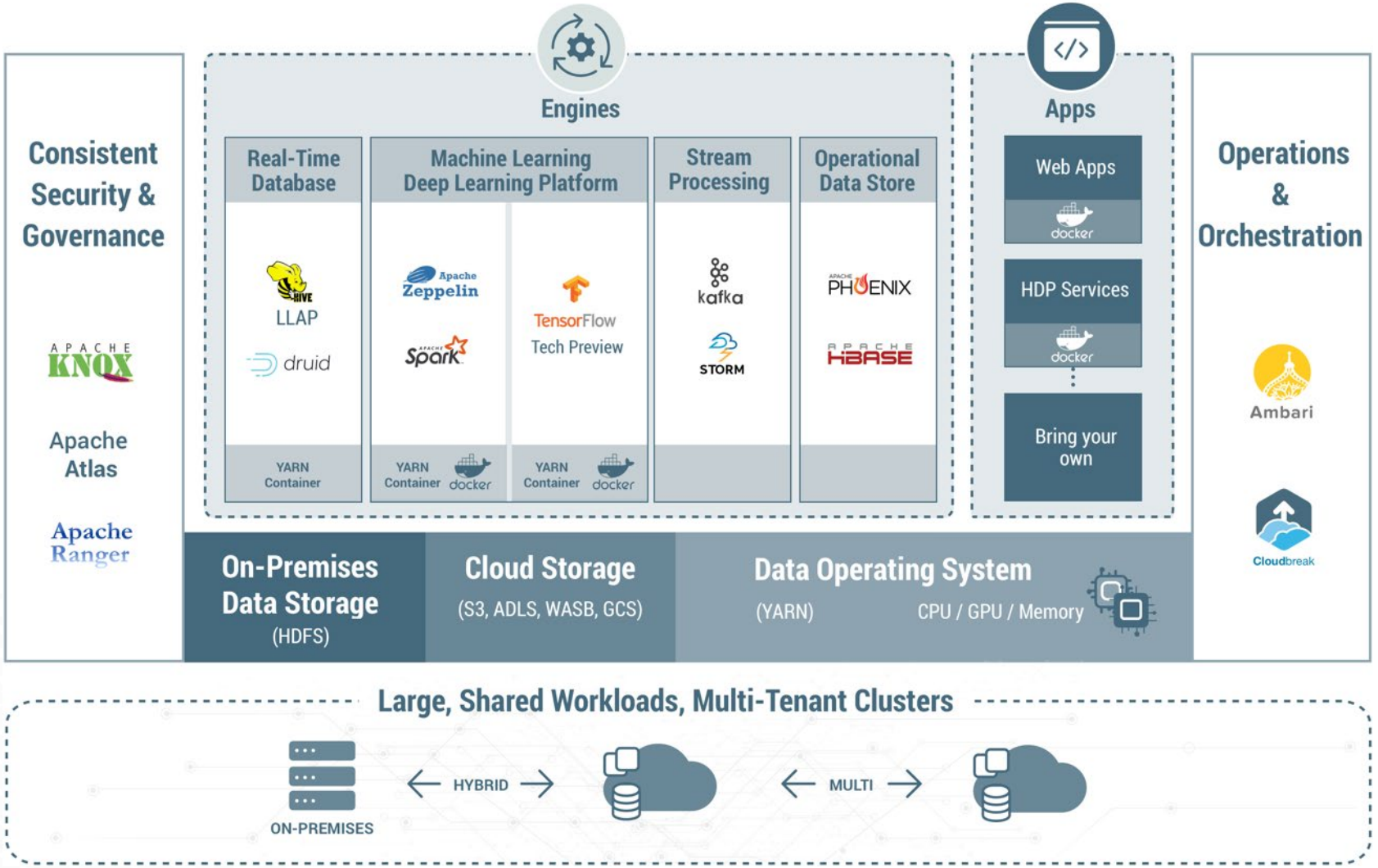
Global Data Management With Hortonworks

Globally Manage, Secure, Govern, Consume



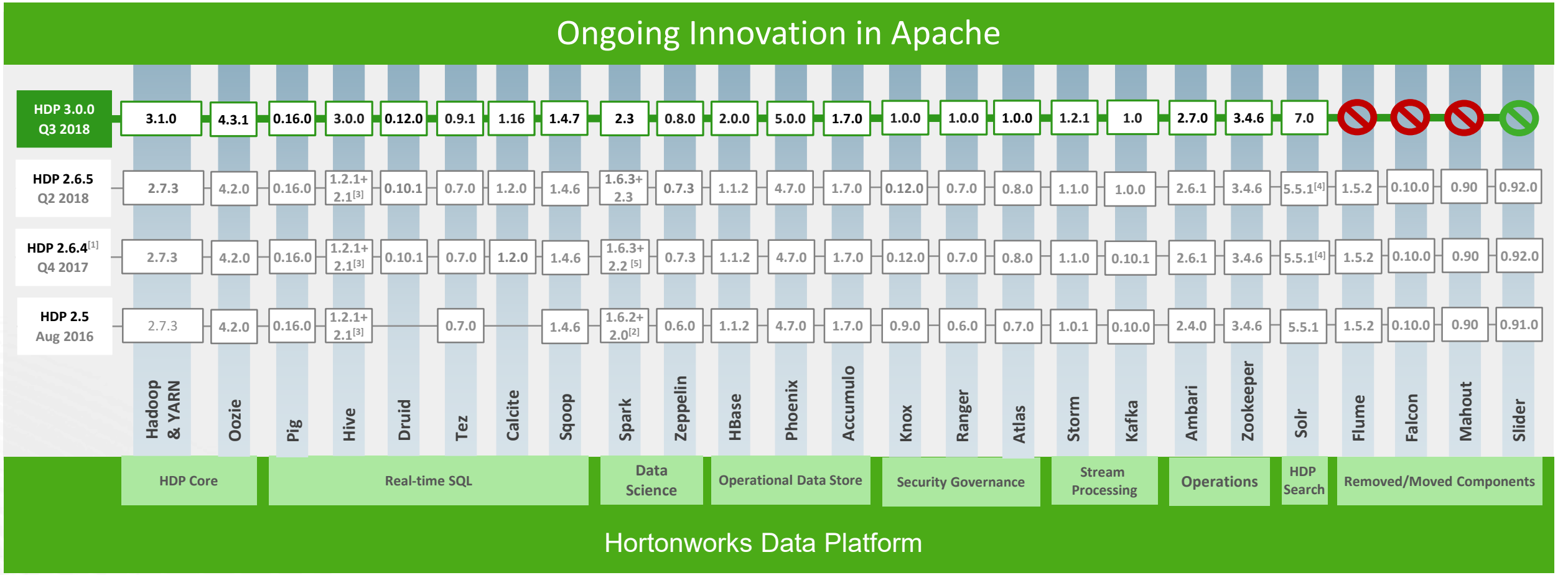
HORTONWORKS®

HDP Hybrid Architecture



HORTONWORKS

Major Changes Across Big Data Eco-System



[1] HDP 2.6 – Shows current Apache branches being used. Final component version subject to change based on Apache release process.
[2] Spark 1.6.3+ Spark 2.1 – HDP 2.6 supports both Spark 1.6.3 and Spark 2.1 as GA.
[3] Hive 2.1 is GA within HDP 2.6.
[4] Apache Solr is available as an add-on product HDP Search.
[5] Spark 2.2 is GA

HORTONWORKS

Agenda

- Overview of Hortonworks
- Hortonworks Connected Data Architecture
- Streaming Solutions using Hortonworks Data Flow (HDF)
- HDF Customer Stories & Use Cases
- Q&A

HORTONWORKS®
POWERING THE FUTURE OF DATA™

HDF Data-In-Motion Platform

Flow Management

Data acquisition and delivery
Simple transformation and data routing
Simple event processing
Edge to Enterprise data lineage and provenance
Edge device connectivity and IoT data ingestion

nifi

minifi C++ Agent Java Agent

Stream Processing

Scalable data broker for streaming apps
Scale out streaming computation engine

kafka

STORM

Stream Analytics

Pattern Matching
Prescriptive & Predictive Stream Analytics
Complex Event Processing
Continuous Insights

STREAMING ANALYTICS MANAGER

Enterprise Services

Provisioning, Management, Monitoring,
Security, Audit, Compliance, Governance,
Multi-tenancy

APACHE KNOX

SCHEMA REGISTRY

APACHE NIFI registry

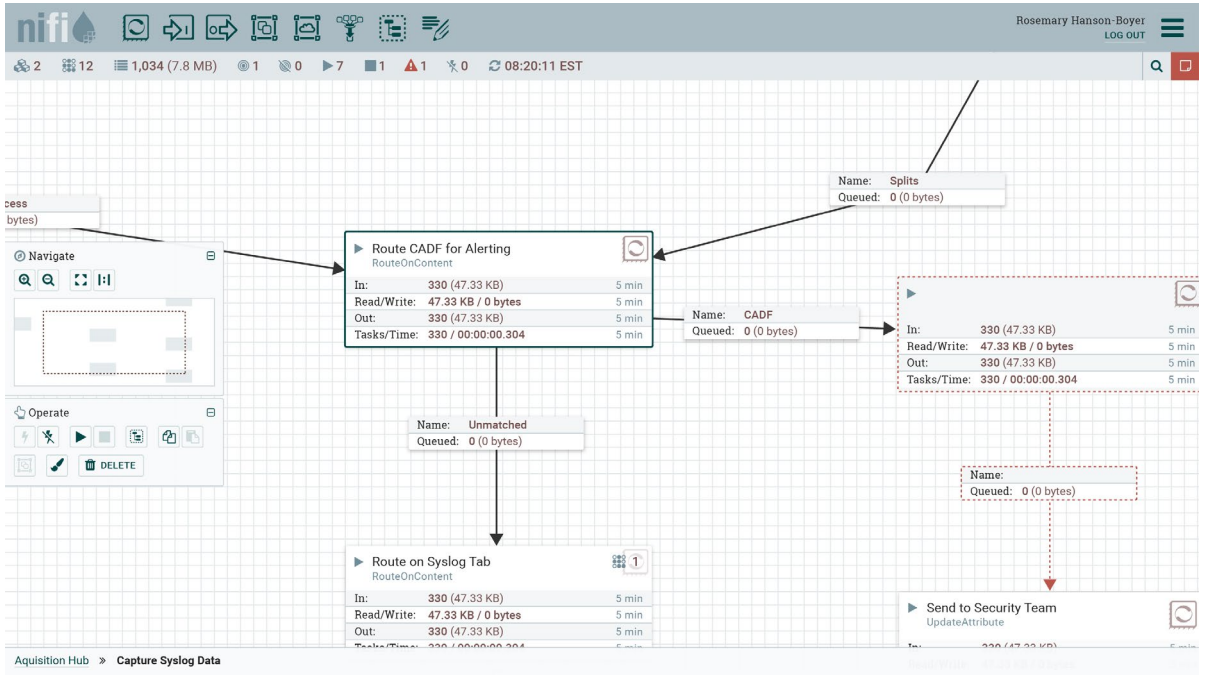
Ambari

Apache Ranger

Flow Management with Apache NiFi

Apache NiFi High Level Capabilities

- Web-based user interface
  - Design, control, feedback & monitoring
- Highly configurable
  - Loss tolerant vs guaranteed delivery
  - Low latency vs high throughput
  - Dynamic prioritization
  - Flow can be modified at runtime
  - Back pressure
- Data provenance
  - Track dataflow from beginning to end
- Designed for extension
  - Build your own processors
- Secure
  - SSL, SSH, HTTPS, etc.



HORTONWORKS

HDF - Flow Management powered by Apache NiFi

- Ingestion: connectors to read/write data from/to several data sources
- Transformation:
  - Format conversion
  - Compression/decompression, Merge, Split, encryption, etc
- Data enrichment
  - Attribute, content, rules, etc
- Routing
  - Priority, dynamic/static, based on content or metadata, etc
- Parsing

HORTONWORKS

260+ Processors for Deeper Ecosystem Integration

FTP
SFTP
HL7
UDP
XML
HTTP
WebSocket
Email
HTML
Image
Syslog
AMQP

Hash | Encrypt | GeoEnrich
Merge | Tail | Scan
Extract | Evaluate | Replace
Duplicate | Execute | Translate
Split | Fetch | Convert

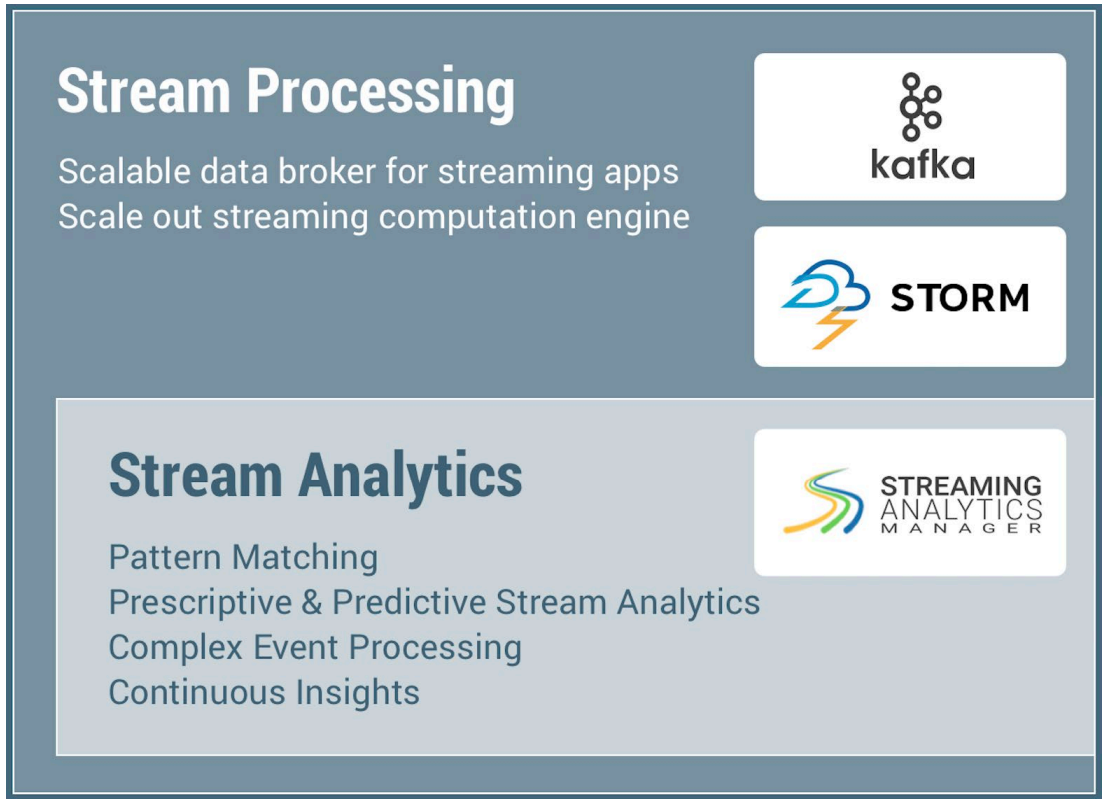
Route Text | Distribute Load
Route Content | Generate Table Fetch
Route Context | Jolt Transform JSON
Control Rate | Prioritized Delivery

HORTONWORKS

Stream Processing

HDF Stream Processing – Streaming Analytics Manager (SAM)

- A product module in the HDF stack to design, develop, deploy and manage streaming analytics app with drag-and-drop ease
  - Build streaming analytics applications that do event correlation, context enrichment , complex pattern matching, analytical aggregations and creation of alerts/notifications when insights are discovered.
  - Supports multiple streaming substrates (e.g: Storm, Spark Streaming, Flink)
  - Extensibility is a first class citizen (add custom sinks, processors, spouts, etc..)

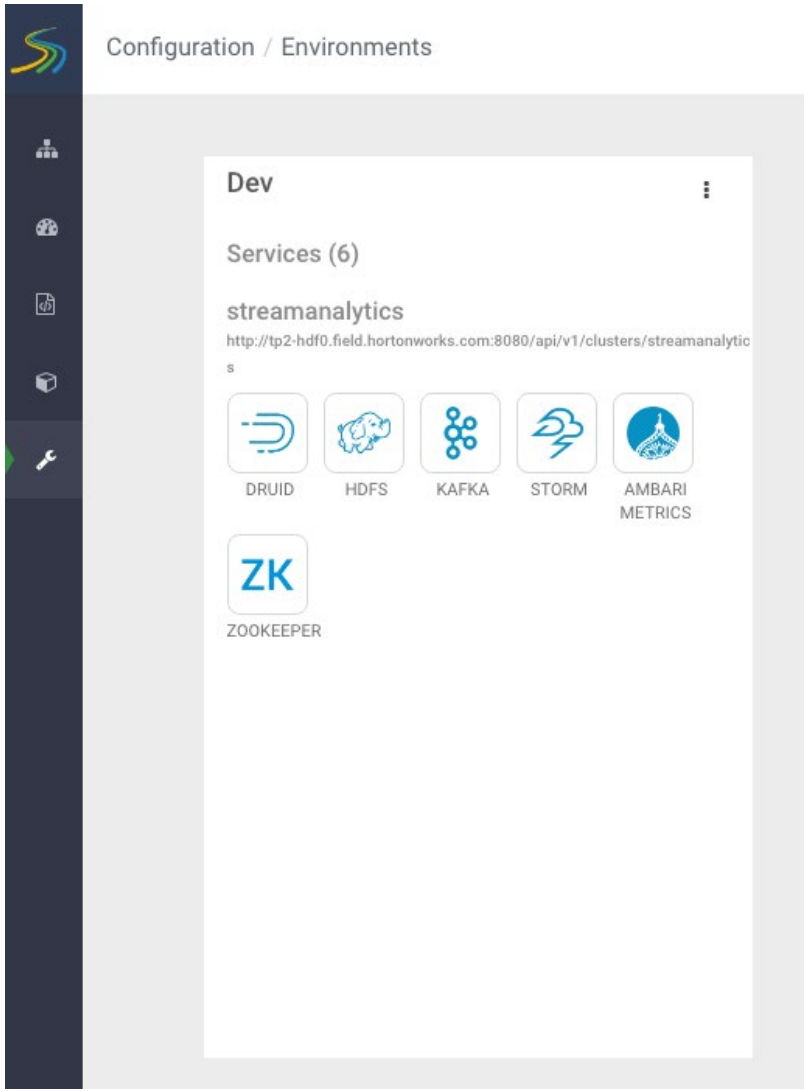


HORTONWORKS

Who Uses SAM?

HORTONWORKS

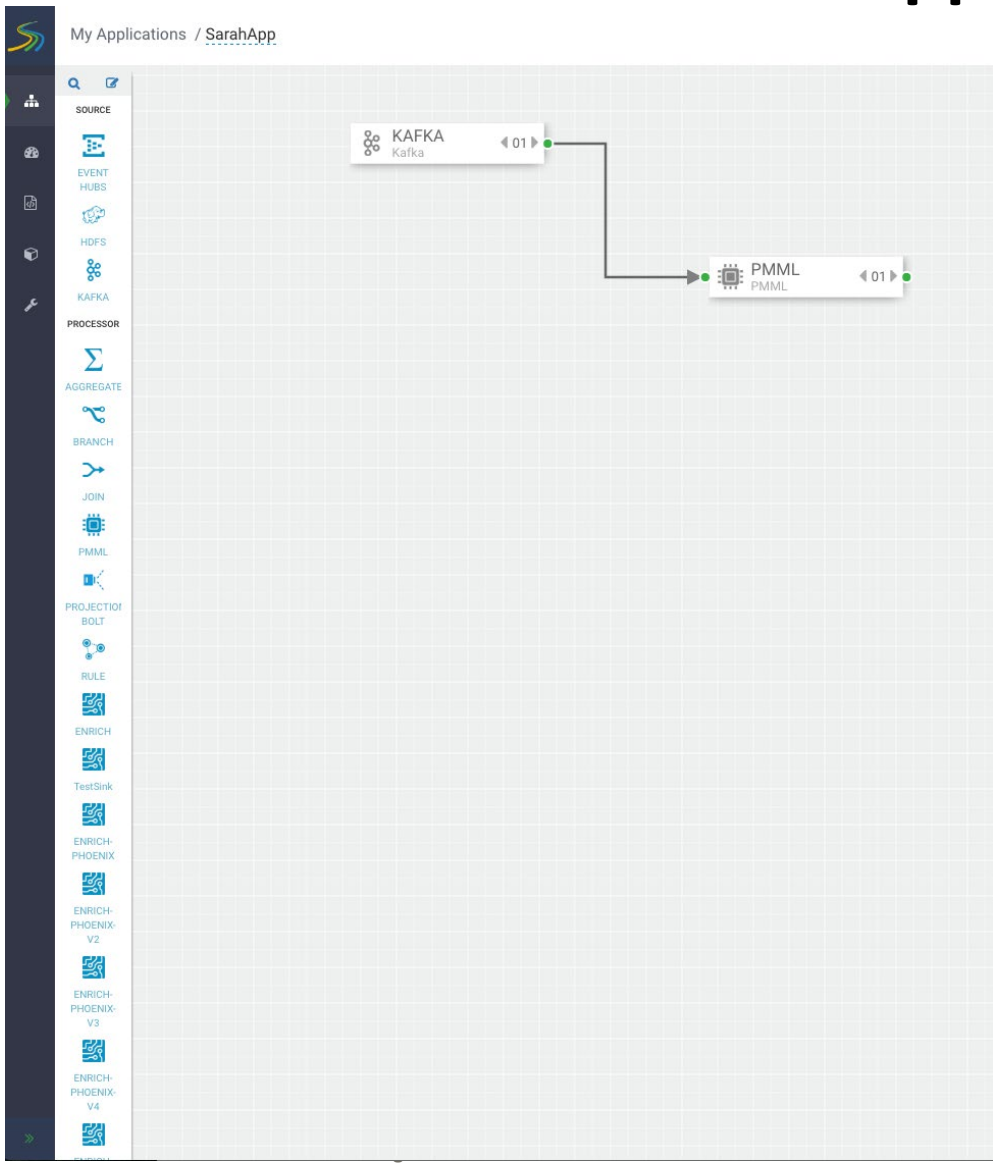
Stream Ops Module for IT Operations

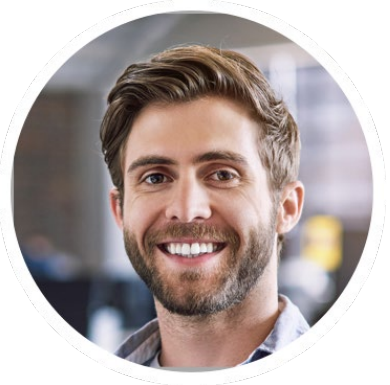


- Service Pool Abstraction
- Create and manage different environments in which individual streaming applications will be built
- Environments consists of services such as HDFS, Kafka, Storm from different service pools
- Save time and reduce operational overhead with same drag and drop paradigm as the stream build module
- SAM takes away the complexity of deploying secure streaming analtyics on kerborized cluster

HORTONWORKS

Stream Builder Module for App Developers

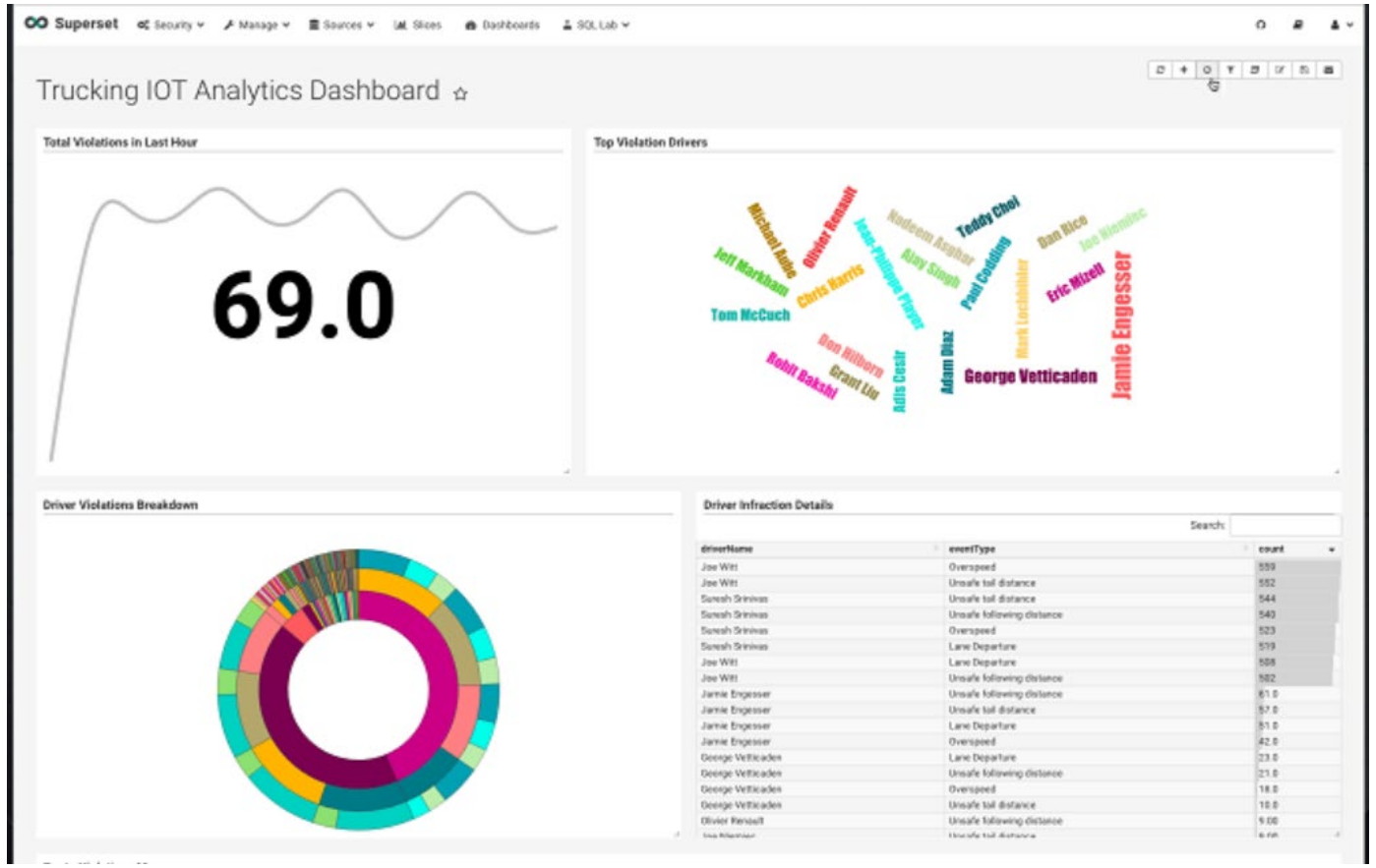




- Builder components, shown on the canvas palette, are the building blocks used by the app developer to build streaming apps.
- Drag and drop to build a working streaming application without writing a single line of code.
- 4 Types of Components: Sources, Processors, Sinks and Custom

HORTONWORKS

Stream Insight Module for Business Analysts



- A tool to create time-series and real-time analytics dashboards, charts and graphs
- 30+ visualization charts out of the box with customization capability
- Druid is the Analytics Engine that powers the Stream Insight Module.

HORTONWORKS


Enterprise Services
HORTONWORKS

HDF Enterprise Services

- Schema Registry
- Apache NiFi Registry
- Ambari
- Apache Ranger
- Apache Knox
- SmartSense

HORTONWORKS

Agenda

- Overview of Hortonworks
- Hortonworks Connected Data Architecture
- Streaming Solutions using Hortonworks Data Flow (HDF)
- HDF Customer Stories & Use Cases
- Q&A

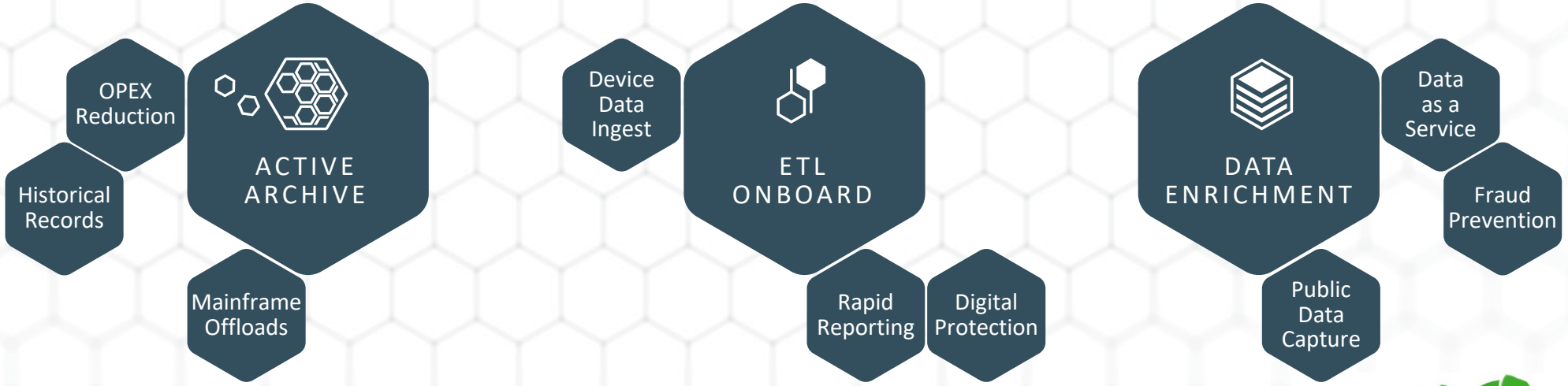
HORTONWORKS
POWERING THE FUTURE OF DATA™

INSURANCE

EXPLORE | OPTIMIZE | TRANSFORM

INNOVATE

RENOVATE



HORTONWORKS

HDF Use Cases

Data Movement

Optimize resource utilization by moving data between data centers or between on-premises infrastructure and cloud infrastructure

Optimize Log Collection & Analysis

Optimize log analytics solutions such as Splunk by using HDF as a single platform to collect and deliver multiple data sources and using HDP for lower cost storage options

Gain key insights with Streaming Analytics

Accelerate big data ROI by analyzing streaming data for patterns, comparing with ML models and delivering actionable intelligence

Single view / 360° view of customer

Ingest, transform and combine customer data from multiple sources into a single data view / lake

Stream Processing

Combine multiple streams of data in real-time, enrich the data and route it to different end points based on rules

Capture IoT Data

Ingest sensor data from IoT devices and stream it for further processing and comprehensive analysis

HORTONWORKS

New analytic applications for new types of data

Financial Services

- New Account Risk Screens
- Fraud Prevention
- Trading Risk
- Maximize Deposit Spread
- Insurance Underwriting
- Accelerate Loan Processing

Retail

- 360° View of the Customer
- Analyze Brand Sentiment
- Localized, Personalized Promotions
- Website Optimization
- Optimal Store Layout

Telecom

- Call Detail Records (CDRs)
- Infrastructure Investment
- Next Product to Buy (NPTB)
- Real-time Bandwidth Allocation
- New Product Development

Manufacturing

- Supplier Consolidation
- Supply Chain and Logistics
- Assembly Line Quality Assurance
- Proactive Maintenance
- Crowdsourced Quality Assurance

Healthcare

- Genomic data for medical trials
- Monitor patient vitals
- Reduce re-admittance rates
- Store medical research data
- Recruit cohorts for pharmaceutical trials

Utilities, Oil & Gas

- Smart meter stream analysis
- Slow oil well decline curves
- Optimize lease bidding
- Compliance reporting
- Proactive equipment repair
- Seismic image processing

Public Sector

- Analyze public sentiment
- Protect critical networks
- Prevent fraud and waste
- Crowdsource reporting for repairs to infrastructure
- Fulfill open records requests

Hortonworks


How Our Customers
Rely On Hortonworks for Hadoop
HORTONWORKS

EDW Optimization

Active Archive, ETL Offload and Data Enrichment

Meeting Deadlines for Time-Sensitive Employment Reports

Government

US federal government labor agency

Why Hadoop?

ETL Offload

Problem: Federal agency had only 9 days to prepare monthly report

- Monthly employment report moves financial markets
- State agencies report unemployment data to federal office by first Friday of month
- Total data set is hundreds of millions of rows in 30 comma-separated files
- Final report must be published by the third Friday of the month, time is precious

Solution: HDP speeds processing and improves confidence in findings

- Easy POC pilot: processing one of thirty files on HDP/Amazon Cloud solution
- Processing time reduced from 18 hours to less than 1 hour
- Absolutely no disruption to existing systems or operations
- Cloud cluster runs on “as needed” basis, shut down remotely when not needed

Hortonworks

Capturing Consulting Revenue for Government ETL Offload

Government

Professional service provider consulting on federal projects

Why Hadoop?

Active Archive

Problem: Federal consulting firm inundated with ETL backlog driven by budget sequestration standoff

- Sequestration budget cuts created demand for ETL from SAS to reduce expense
- Millions of consulting dollars available and at risk from projects to offload older, infrequently accessed data from SAS at 20 fed civilian agencies
- After offload, all data still needed to be easily accessible (i.e. not stored to tape)

Solution: Rationalized offload of ETL processes earned consultant revenue and saved taxpayers money

- Federal civilian agencies reduced ongoing data storage costs
- There was no loss of data or disruption to ongoing operations
- Apache Hive connects out-of-the-box with Base SAS and SAS/ACCESS for connectivity between SAS and Hadoop

Hortonworks

Telco Clickstream Data Offload, Projected Savings > $1 Million

Telecom

Major telco

Why Hadoop?

Active Archive

Problem: System ingests millions of call detail records per second, unable to economically retain such large volumes

- Netezza EDW operating near capacity sand storing exhaust data not required for intended reporting and analytics, leading to unnecessary expense
- Enterprise IT maintained redundant data stores
- Unable to store clickstream data to meet goal of enriching consumer intelligence

Solution: Lower storage costs and schema-on-read architecture drive improvement in consumer intelligence

- HDP recovers Teradata cycles, currently used ETL and data movement
- Projected costs savings of >$1M by offloading exhaust data
- Analysis of clickstream adds new dimension to customer view
- Improved service efficiency with customer bill processing & reporting

Hortonworks

Storage Efficiencies for 100x the Data with $3M in Savings

Telecom

Telco information and analytics vendor

Why Hadoop?

ETL Offload

Problem: Changing business model required a new data architecture

- Telco started in 1990s as neutral intermediary for telco networks
- Network management market matured, CEO challenged company to build business for telco network data analysis and information services
- Netezza capacity limited to 20TB—1% of data available—retained for only 60 days

Solution: HDP stores more data for longer while saving $3 million

- New HDP solution avoided $3M annual Netezza expense
- Now 100% of network data is captured, stored and retained for two years
- Larger data set supports new, accurate information products, spurring new growth in the business
- Improved data access across the company drives greater enterprise productivity

Hortonworks


Advanced Analytic Applications:
A Single View
Deeper Insight into Customers, Products and Networks
HORTONWORKS

Sentiment Analysis for Responsive Government

Government

European national government

Why Hadoop?

Single View

Problem: Ministry of Education felt distant from public sentiment on obesity reduction programs

- In-person events lacked reach to many citizens and persistence over time
- Dedicated analysts pored over social media and provided daily manual reports
- IT team sought path to better sentiment analysis to members of parliament

Solution: Powerful daily sentiment analysis improves government accountability and citizen engagement

- Team produces daily memos on public sentiment, now with:
  - Reach: includes opinions from broader base of the citizenry
  - Confidence: more data corresponds to more confidence in opinion analysis
  - Frequency: daily reads show policy-makers changes over time
- Social media leaders now invited to in-person meetings with government ministers

Hortonworks

HDP Combined with Data Warehouse for Single View of Patient

New Analytic Applications
Sensor & Structured Data

Healthcare

Catholic healthcare system

Why Hadoop?

Single View

Problem: Current monitoring of patient vitals via Epic EHR and Clarity data warehouse limits data capture to 15 minute windows

- Electronic health record (EHR) from Epic with analysis on Oracle's Clarity EDW limits data capture to 1/900 of total data and requires manual intervention
- In ICU, devices capture patient vitals once per second. Every 15 minutes a nurse selects on reading as representative of patient and selects that for retention.
- Limited data capture hampers analysis (e.g. effectiveness of medication)

Solution: HDP captures vitals for each ICU patient every second, leading to 900X improvement in data capture and greater patient detail

- Data lake captures far more data, with benefits including: data offload from Clarity (with metadata tags), Hive queries with SQL semantics familiar to existing analysts, more granular updates hourly (rather than daily)

Hortonworks

Single View of Customer Improves Call Center Recommendations

Retail

IT solution and equipment reseller

Why Hadoop?

Single View

Problem: Call center inside reps unable to recommend the best products to buy next

- 2000+ product lines represent too much information to commit to memory
- Multiple customer interaction channels (web, Salesforce, face-to-face, phone) cloud the company's understanding of its customers
- Poor visibility causes sales reps to miss opportunities, customer satisfaction suffers

Solution: HDP improves cross-sell and up-sell with next product to buy (NPTB) recommendations

- Recommendation engine predicts the next best product for each customer
- Call center reps feel more confident and productive which improves turnover
- Natural language analysis of emails identifies best response language and coaching opportunities for struggling reps

Hortonworks

Predictive analytics and proactive maintenance for military aircraft

New Analytic Applications
Sensor, Structured and Unstructured Data

Government

Branch of US military

Why Hadoop?

Predictive Analytics

Problem: Arm of the US military had limited analytical capabilities to tie aircraft maintenance records to performance and safety

- Condition-based maintenance (CBM) manages aircraft lifecycles, by optimizing maintenance resources and making component end-of-life decisions
- CBM requires merging maintenance records with sensor data and detailed aircraft usage data, correlated over time
- Existing platform could only analyze a single flight, with no same-craft history
- Each aircraft model was managed by a separate program with a different data repo

Solution: Service is building an HDP data lake with the goal of ingesting flight data from every flight ever flown

- Superior analysis improves aviator safety, maintenance efficiency and IT efficiency
- Now the team has the ability to conduct historical trend analysis
- Individual programs now contribute data to Hadoop, driving collaboration
- Previously impossible queries now return results in seconds

Hortonworks

Improve Patient Treatment with Real-time Monitoring of Vital Signs

New Analytic Applications
Sensor, Social Data
& ETL Offload

Healthcare

Public university teaching hospital

Why Hadoop?

Predictive Analytics

Problem: Inability to store and access sufficient data for medical decision support in real time

- 9 million patient records on a legacy system were not searchable nor retrievable
- Cohort selection for research projects was slow, despite abundance of data
- Clinicians had minimal access to historical data gathered across all patients

Solution: Unified data lake improves patient health, speeds research

- Legacy system retired immediately, saving $500K in annual recurring expense
- Records stored with patient identification for clinical use, same data presented anonymously to researchers for cohort selection
- Wireless patches transmit vital signs, algorithms notify doctors of high risk patterns
- Heart patients weigh themselves from home, algorithms notify doctors about unsafe weight changes and recommend a visit to the clinic

Hortonworks

Searchable Data Lake for Next-Product-To-Buy Recommendations

New Analytic Applications
Sensor, Structured, Server Log & Geo-location Data

Telecom

Telco vendor specializing in VOIP

Why Hadoop?

Predictive Analytics

Problem: Data storage costs limit the amount and types of data available for analysis of CDRs and CRM records

- Teradata and Vertica used for data storage, ideal for certain data workloads, but unsuited for less structured types of data
- Limited retention of call detail records (CDRs)
- Limits unified analysis of call logs, CRM records & customer acquisition models

Solution: HDP data lake for ETL offload, ad hoc data exploration and next product to buy (NPTB) recommendations

- Partners Teradata, HP and Impetus partnered with Hortonworks to create integrated solution for modern data architecture
- HDP retains CDRs for longer, improving data visibility and analysis
- Customer retention data correlated to service quality for efforts to reduce attrition
- Integrated search powers real-time NPTB recommendations

Hortonworks

Data-Driven Romantic Recommendations Improve Dating Site

Online Community

Online dating site

Why Hadoop?

Data Discovery

Problem: Newer types of data unavailable to matchmaking algorithms

- Unable to store clickstream data and user-entered content at sufficient scale
- Other types of data were only retained for seven days, due to storage costs
- Recommendations would help users craft attractive profiles, improving satisfaction
- Relational data platform did not fulfill their requirements for scale or cost

Solution: HDP cluster used for A/B testing to improve romantic matchmaking recommendations

- A/B testing driven by consolidated email & clickstream data from SQL databases
- Deeper understanding of use behavior across devices, browsers and applications
- Mine user-created text (profile language and user-to-user communications) for recommendation engine that improves the likelihood of successful matches
- Longer data retention uncovers subtle trends over longer time window

Hortonworks


Questions?
HORTONWORKS