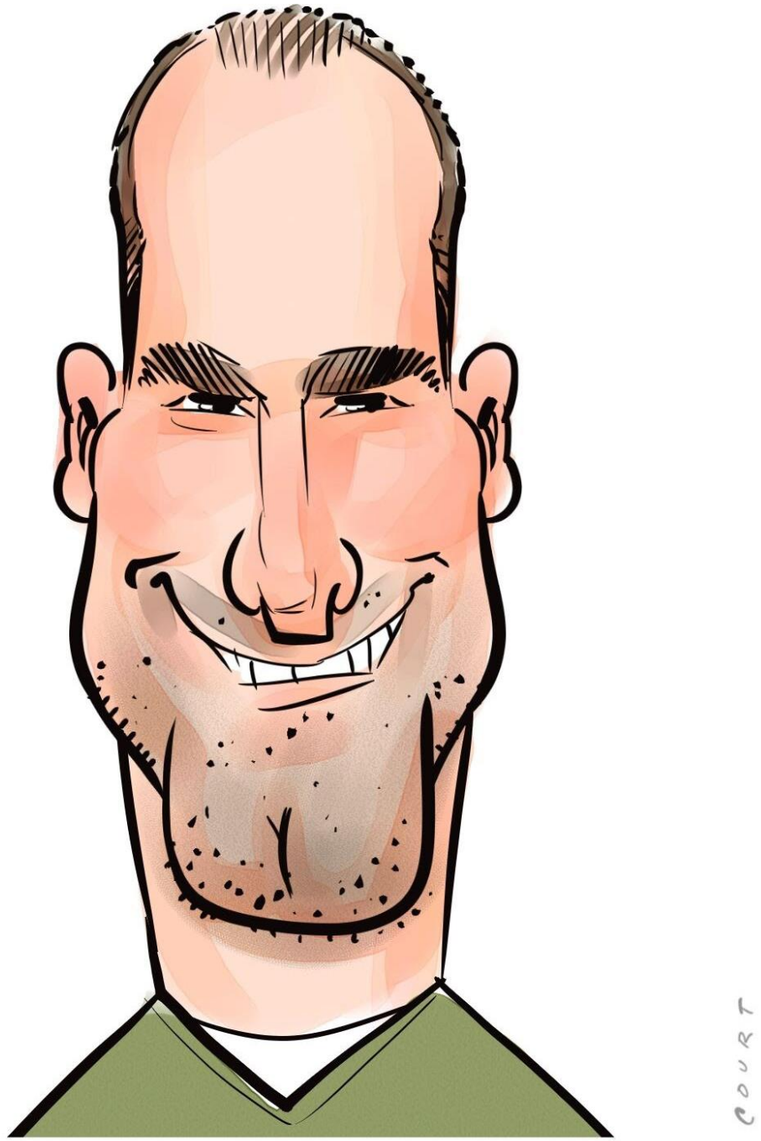
HORTONWORKS

Powering the Future of Data

Henning Kropp
Sr. Systems Architect EMEA

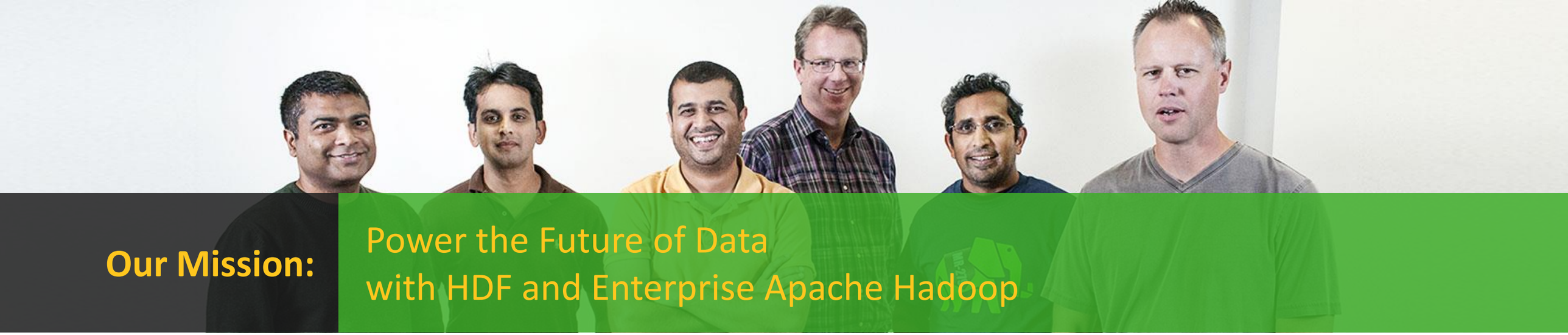
About Me



- Henning Kropp
- Sr. Systems Architect EMEA – Hortonworks
- University Leipzig Graduate
- 5+ Years Hadoop Experience
- 2+ Years with Hortonworks
- Werder Bremen & RB Leipzig Fan

Agenda

- Hortonworks Overview
- Data Trends
- Open Enterprise Hadoop
  - (HDF)
  - HDP
- The Hadoop Ecosystem
- Popular Use Cases
- SQL on Hadoop



Our Mission:

Power the Future of Data with HDF and Enterprise Apache Hadoop

Who we are

June 2011: Original 24 architects, developers, operators of Hadoop from Yahoo!
June 2014: An enterprise software company with 420+ Employees
Oct 2015: Fastest software company to hit $100 M in revenue
Nov 2016: 1000+ customers with 1050+ Employees

Our model

Innovate and deliver Apache Hadoop as a complete enterprise data platform completely in the open, backed by a world class support organization

Key Partners

Microsoft TERADATA SAP hp sas CSC accenture WIPRO Applying Thought Cognizant

HORTONWORKS

Fastest growing Fortune 1000 customer base

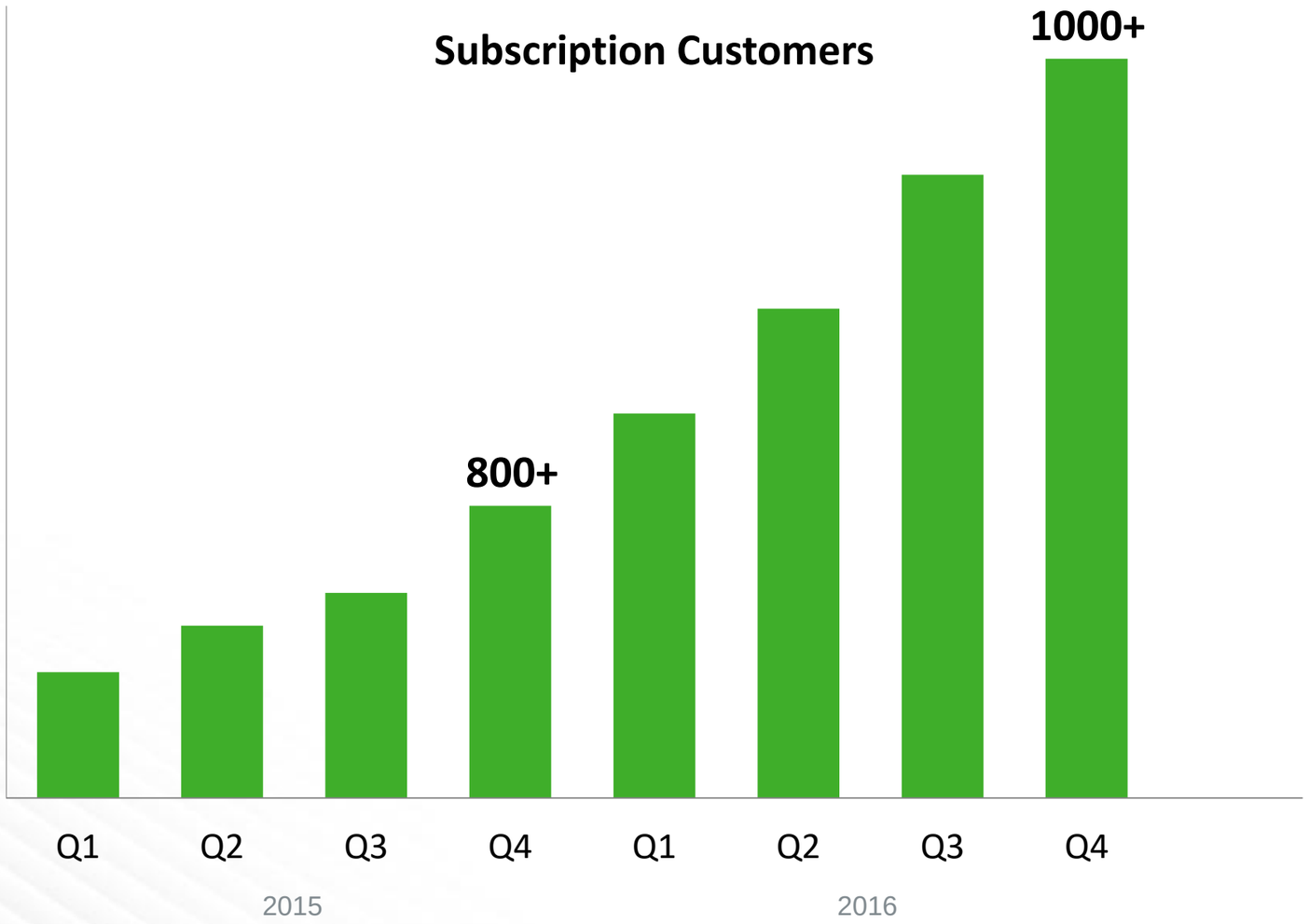
Customer Momentum

- **300+** customers in seven quarters, growing at **75+/quarter**
- Two thirds of customers come from F1000

60+ customers migrated from other distributions

Some notable migrations include many of the early adopters of Hadoop:

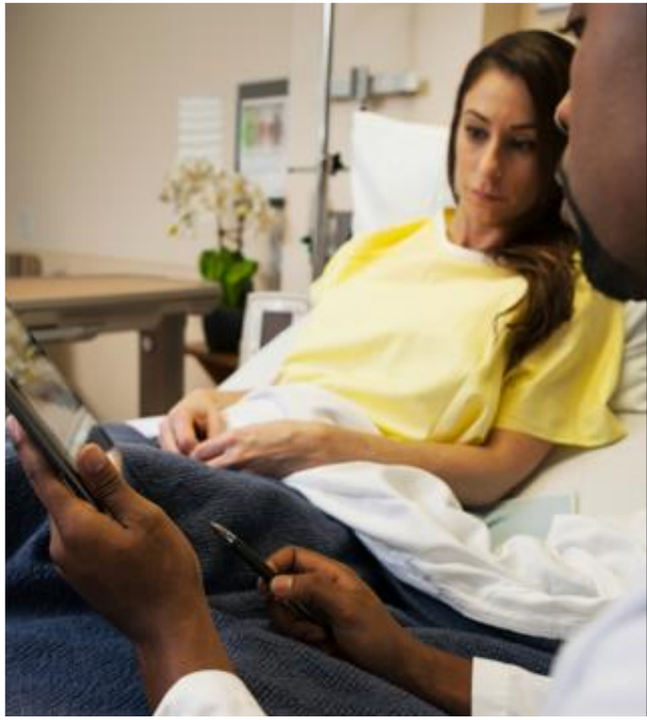
Customer Growth in 2016



Key Highlights

SHOP DIRECT
KOHL'S
DAIMLER
ENTERPRISE HOLDINGS
worldpay
Home Office
Bell Helicopter
Munich RE
ING
Schlumberger
Bloomberg
kpn
neustar
Time Warner Cable
FARMERS
Royal Mail
centrica
PROGRESSIVE
Symantec
MIGROS
YAHOO! JAPAN
ebay
HORTONWORKS

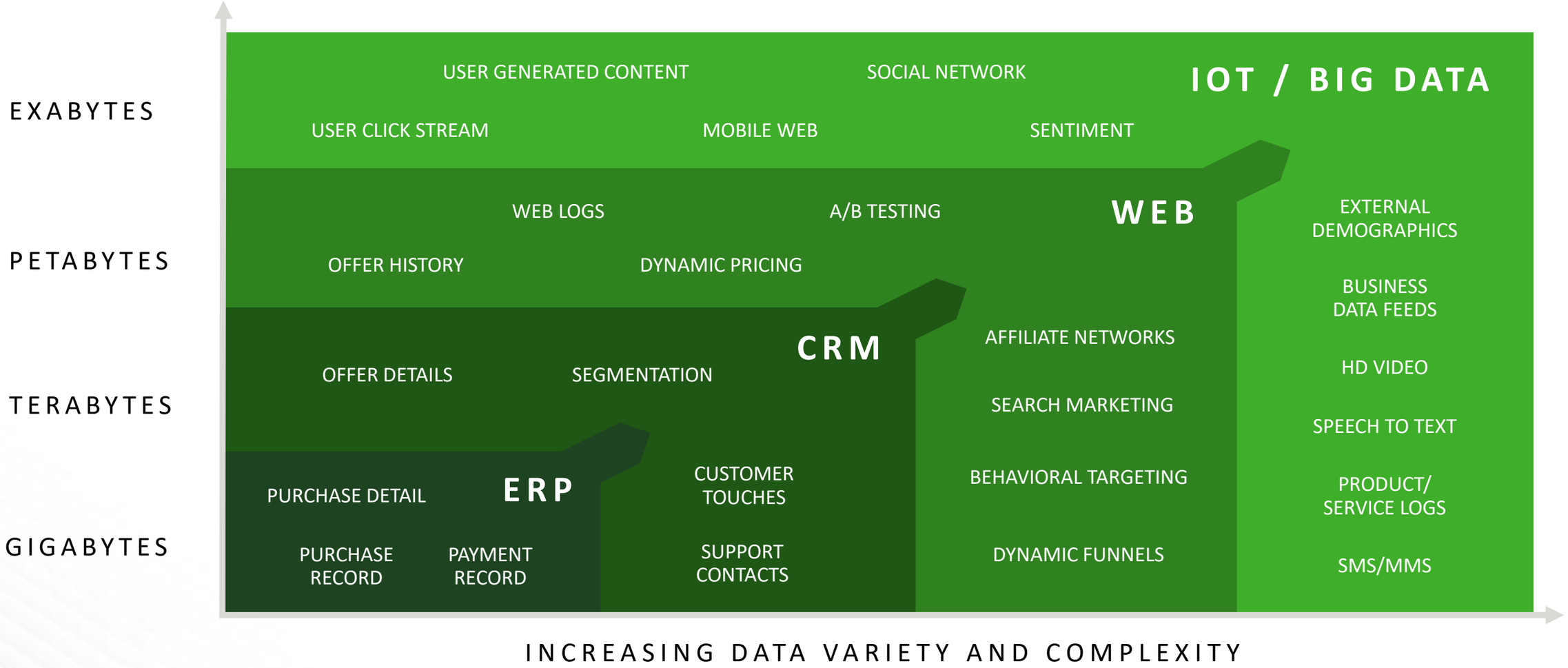
MASTER THE VALUE OF DATA



EVERY BUSINESS IS A DATA BUSINESS

EMBRACE AN OPEN APPROACH

The DATA Ages



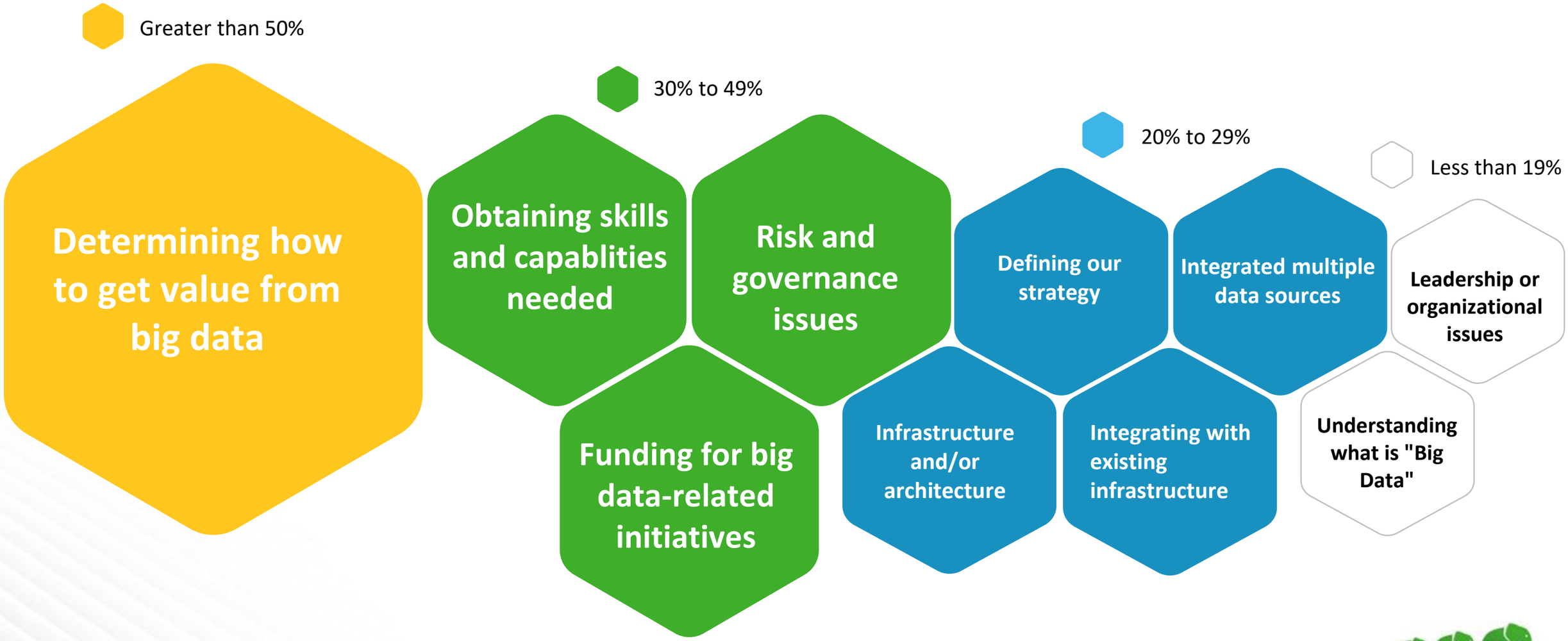
Threats

Existing data architectures make data inaccessible, incomplete, irrelevant, and expensive.

Opportunities

Apache™ Hadoop® transforms your business, making Big Data easily accessible for advanced analytic applications.

Challenges adopting big data according to Gartner



A Connected Data Strategy Solves for All Data

HDF*

DATA IN MOTION

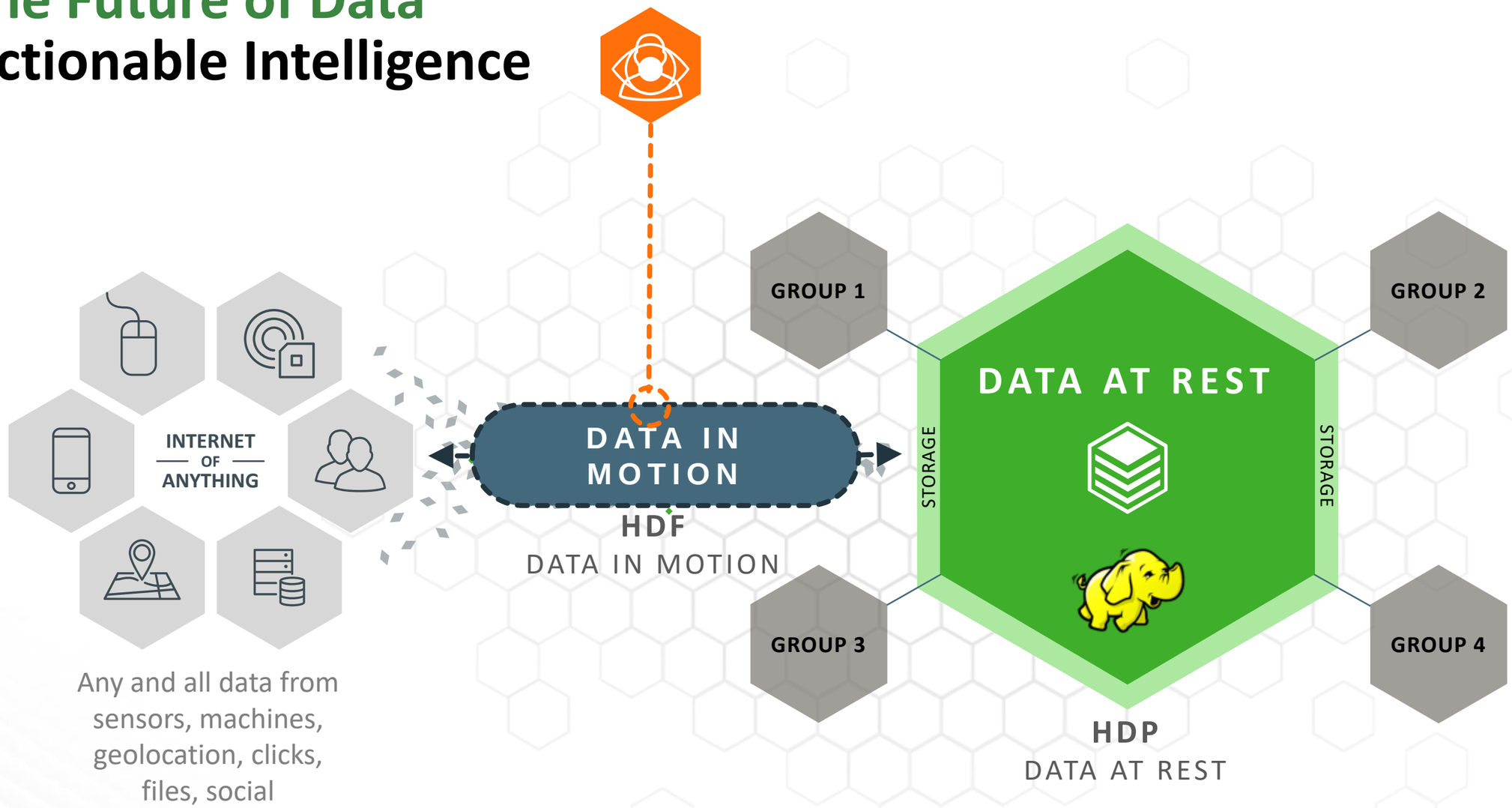
HDP**

DATA AT REST

* Hortoworks Data Flow

** Hortonworks Data Platform

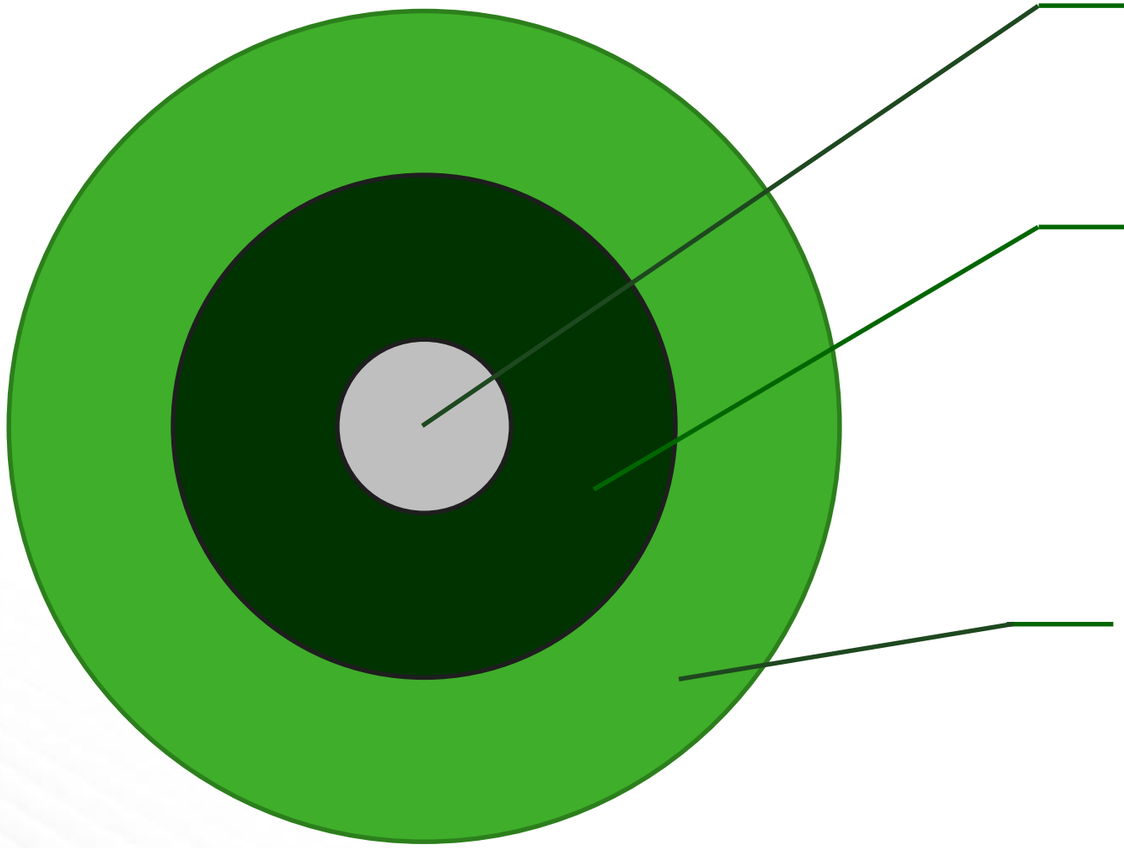
The Future of Data
Actionable Intelligence



Open Enterprise Hadoop

- Open
- Central
- Interoperable
- Ready

How To Think About “Open”



Proprietary IP

This is unique intellectual property that is important to keep proprietary and protected for business reasons.

Open source, closed/controlled community

This is about intellectual property that you are OK sharing with others but still want to retain strict controls over.
Ex. public Github repos where your engineers are the only ones who can checkin code / approve pull requests. Others may “Github fork” the code, however, to use for their own purposes.

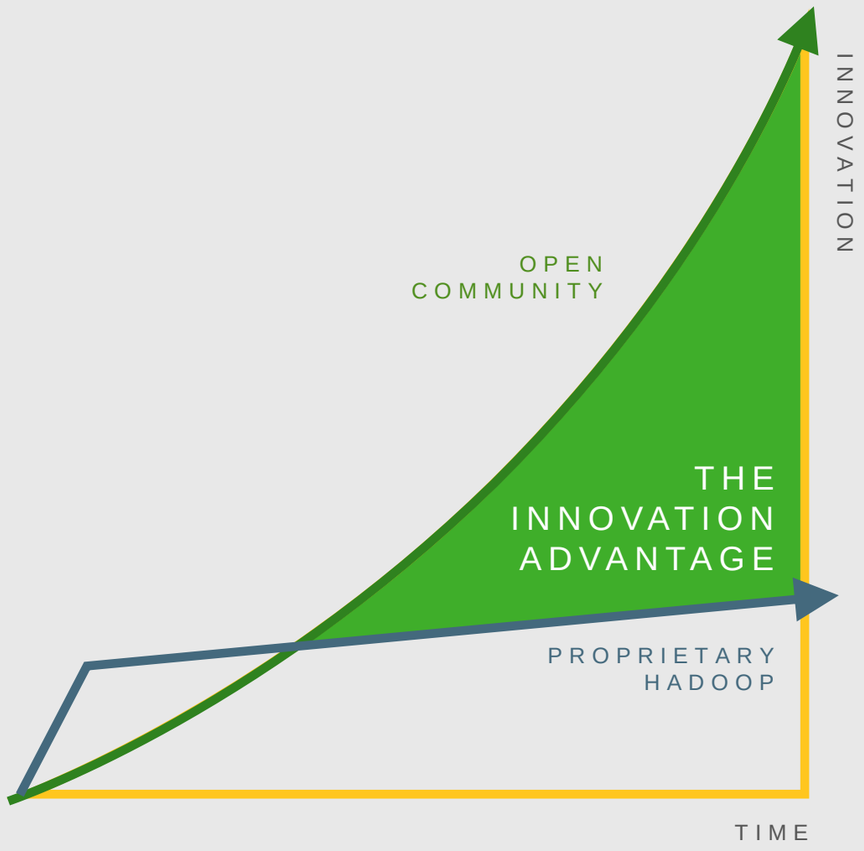
Open source, open community

This is about intellectual property that you are OK sharing with others and want to participate in and/or drive forward as part of a broader community that has a defined and consistent governance model.
Ex. Apache Software Foundation projects are an example of this.

Open Source Is The Norm: The Github Generation

NEW SOFTWARE DEVELOPER GRADS

- GitHub houses their public “portfolio”
- They want companies that embrace open source



MAXIMUM COMMUNITY INNOVATION

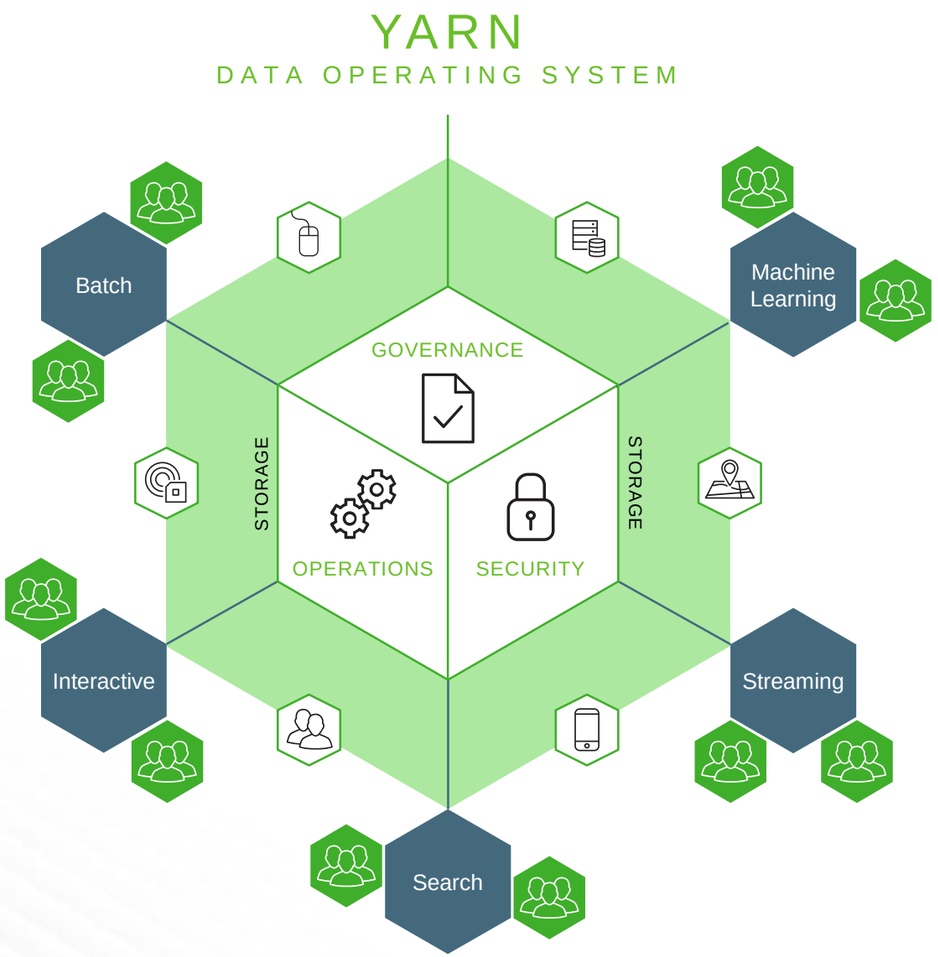
Hortonworks Data Platform Is Genuinely Open

- Eliminates Risk
  - of vendor lock-in by delivering 100% Apache open source technology
- Maximizes Community Innovation
  - with hundreds of developers across hundreds of companies
  - **Integrates Seamlessly**
  - through committed co-engineering partnerships with other leading technologies

Open Enterprise Hadoop

- Open
- Central
- Interoperable
- Ready

Centralized Platform with YARN-Based Architecture



Centralized Platform

for operations, governance and security

Diverse Applications

run simultaneously on a single cluster

Maximum Data Ingest

including existing and new sources, regardless of raw format

Shared Big Data Assets

across business groups, functions and users

Open Enterprise Hadoop

- Open
- Central
- Interoperable
- Ready

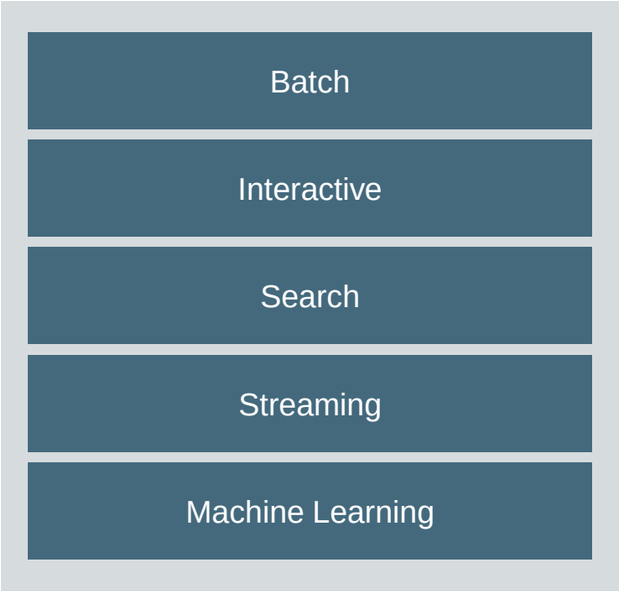
Offering You the Most Flexibility

ANY DATA

Existing and new datasets

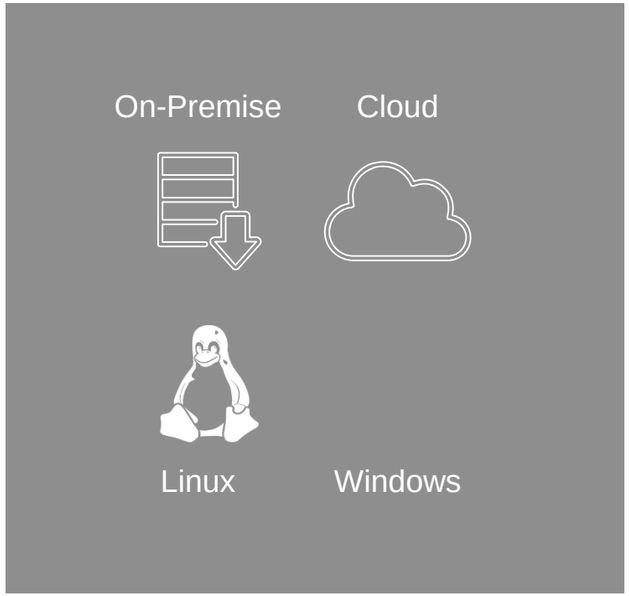
ANY APPLICATION

Multiple engines for data analysis



ANYWHERE

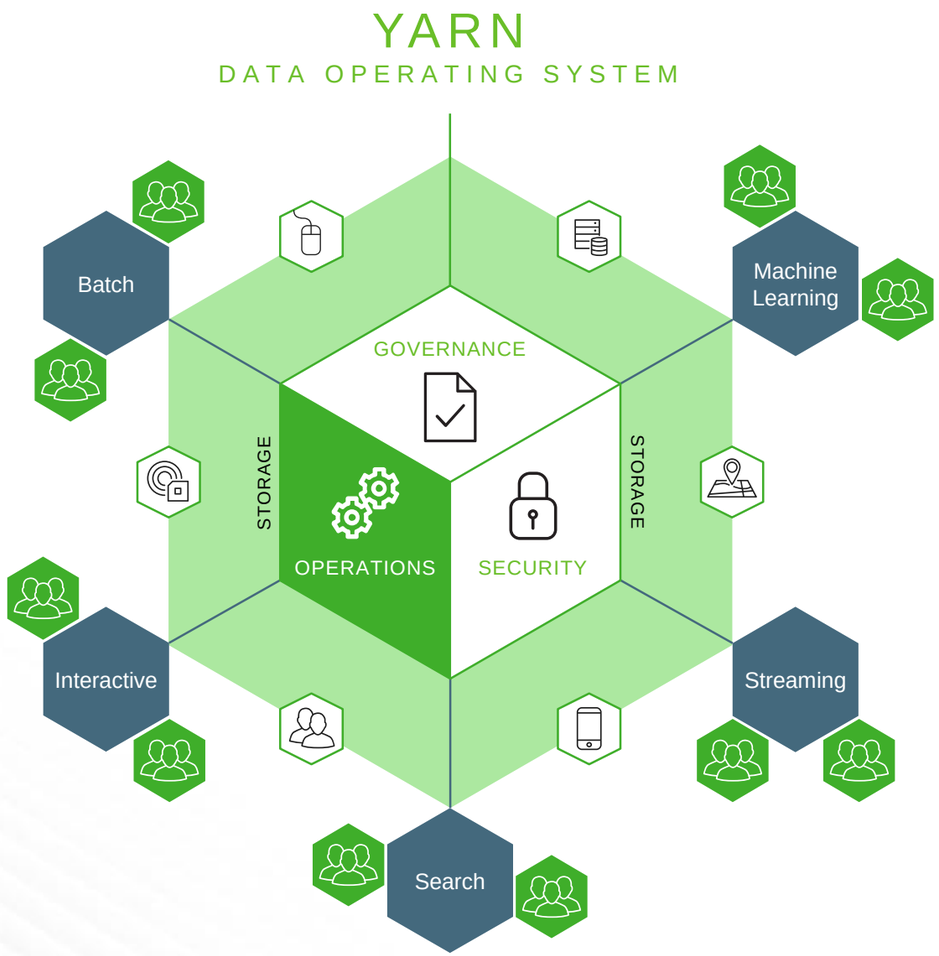
Complete range of deployment options



Open Enterprise Hadoop

- Open
- Central
- Interoperable
- Ready

Provides Consistent Operations



Centralized

management and monitoring of Hadoop clusters

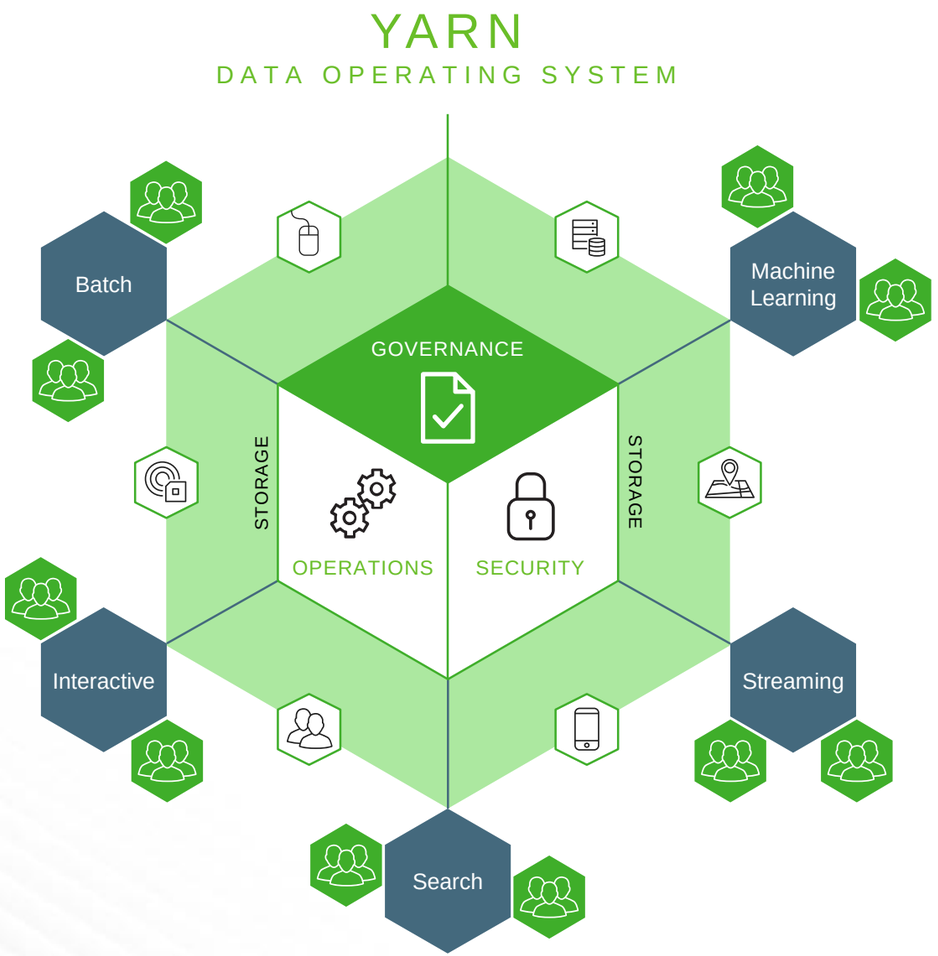
Automated Provisioning

either on-premises or in the cloud with the Cloudbreak API for clusters in minutes

Managed Services

for high availability and consistent lifecycle controls, with dashboards and alerts

Enables Trusted Governance



Data Management

along the entire data lifecycle

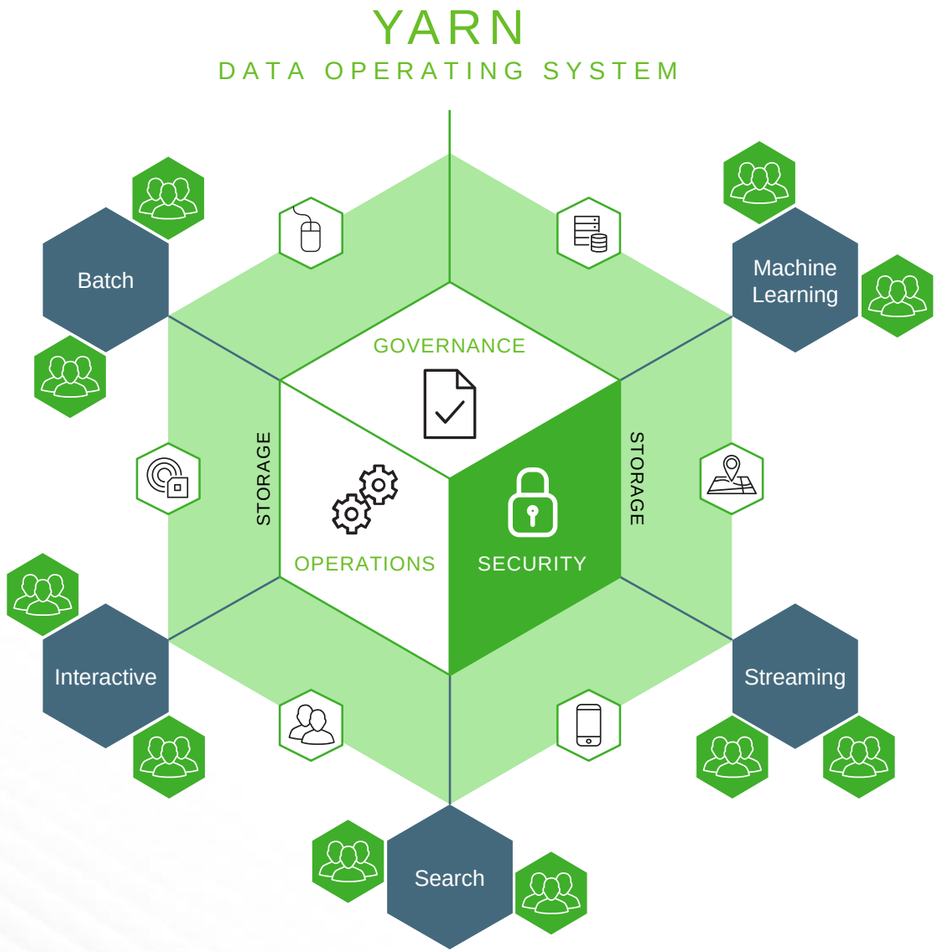
Modeling with Metadata

enables comprehensive data lineage through a hybrid approach

Interoperable Solutions

across the Hadoop ecosystem, through a common metadata store

Ensures Comprehensive Security



Comprehensive Security

through a platform approach

Encrypted Data

at rest and in motion

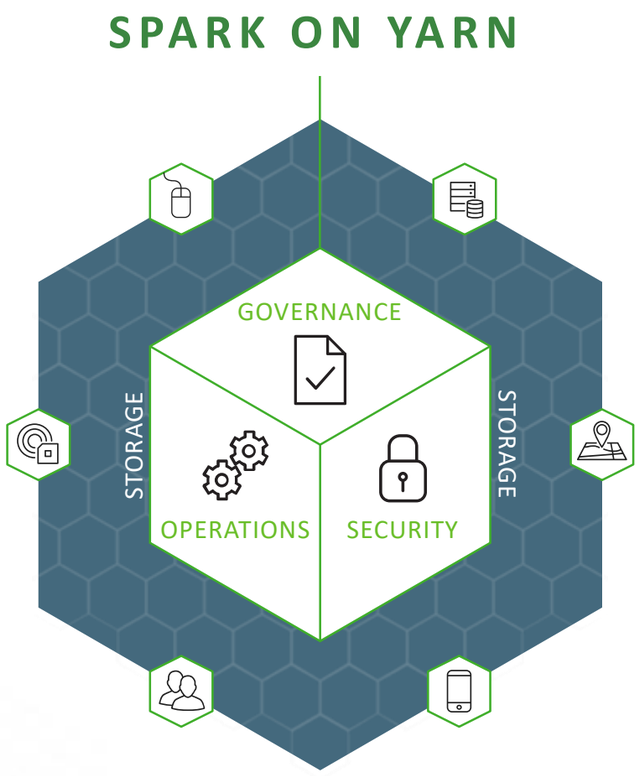
Centralized Administration

of security policies and user authentication

Fine-Grain Authorization

for data access control

Agile Analytics with Enterprise Spark at Scale



Powering Agile Analytics

via data science notebooks and automation for most common analytics (including geospatial and entity resolution)

Seamless Data Access

across as many data types as possible

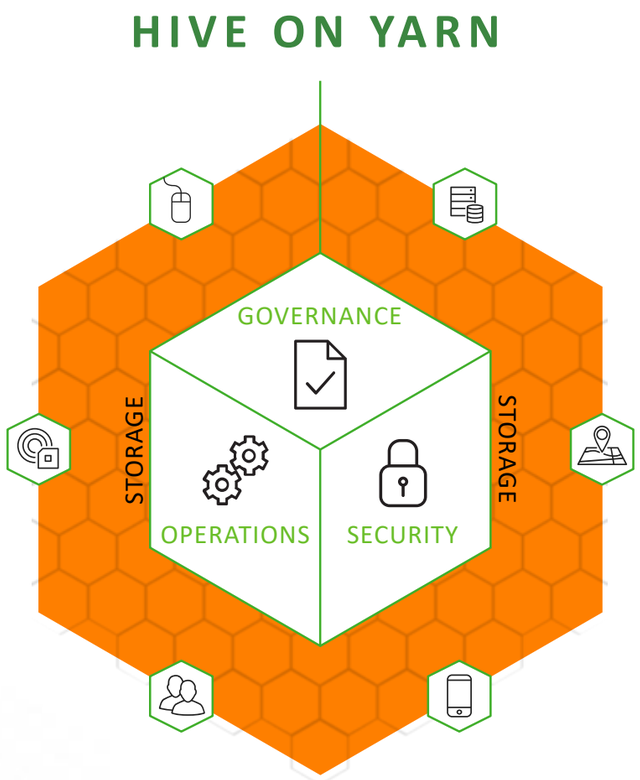
Unmatched Economics

Combining in-memory processing speed with HDP's cost efficiencies at scale

Ready for the Enterprise

with robust security, governance and operations coordinated centrally by Apache Hadoop and YARN

Fast SQL with Apache Hive



Pluggable Architecture

supports Apache Hive, Pivotal HAWQ and other leading SQL engines

Familiar SQL Query Semantics

enable transactions and SQL:2011 Analytics for rich reporting

Unprecedented Speed at Extreme Scale

returns query results in interactive time, even as data sets grow to petabytes

The Hadoop Ecosystem

HORTONWORKS®
POWERING THE FUTURE OF DATA™

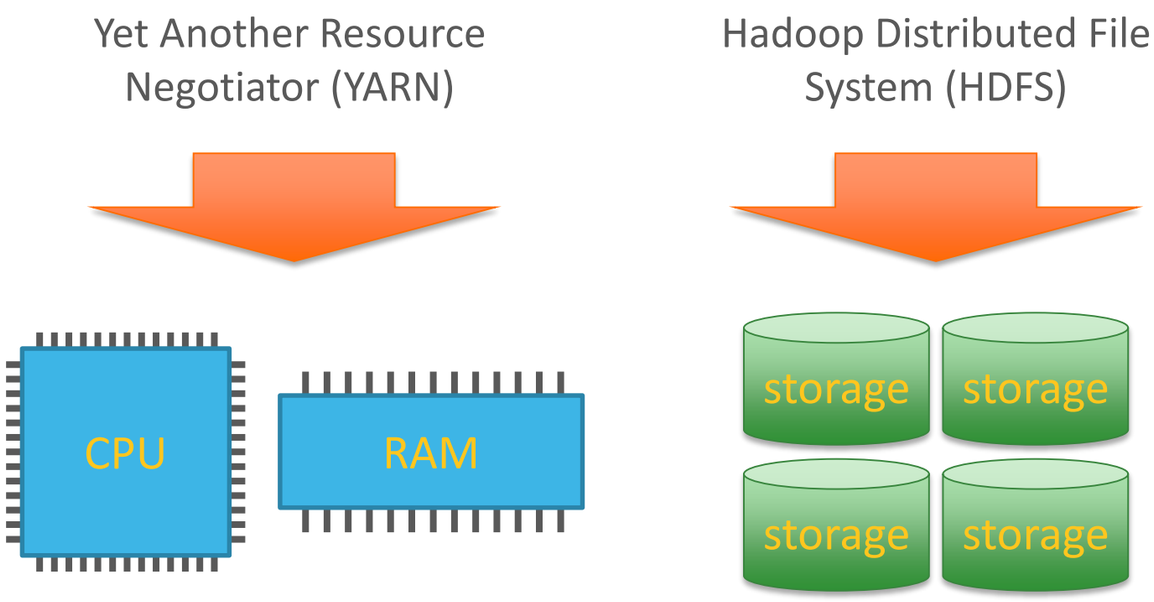
What is Apache Hadoop?

The Apache Hadoop project describes the technology as a software framework that:

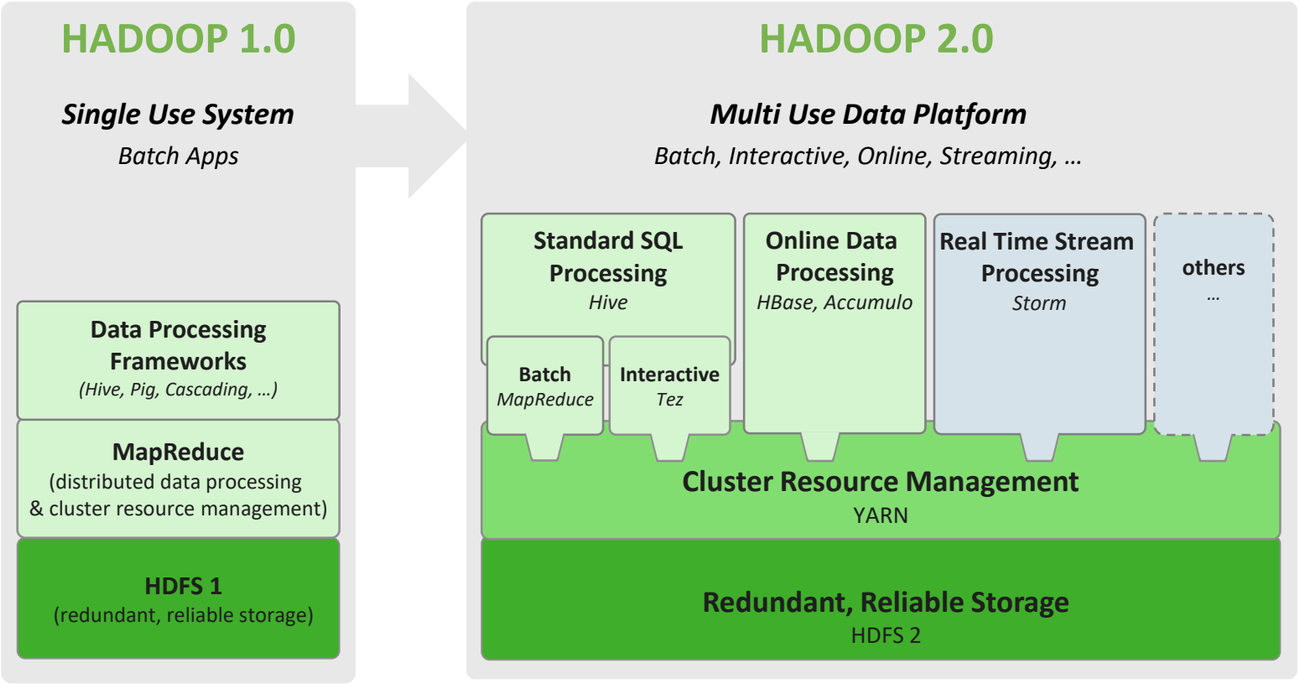
- Allows for the **distributed processing** of **large data sets across clusters of computers** using simple programming models
- Is designed to **scale up from single servers to thousands of machines**, each offering local computation and storage
- Does **not rely on hardware to deliver high-availability**, but rather the library itself is designed to detect and handle failures at the application layer
- Delivers a **highly-available service** on top of a cluster of computers, each of which may be prone to failures

Source: http://hadoop.apache.org

Hadoop Core = Storage + Compute



YARN Enables Multiple Workloads



Interact with all data in multiple ways simultaneously

Architectural components

Hadoop under pins everything, ensuring data is always available and that resources are effectively managed

Nifi ensures that data can be reliably acquired and transmitted regardless of source or destination

HBase provides low latency storage for time series as they are generated and accessed most frequency.

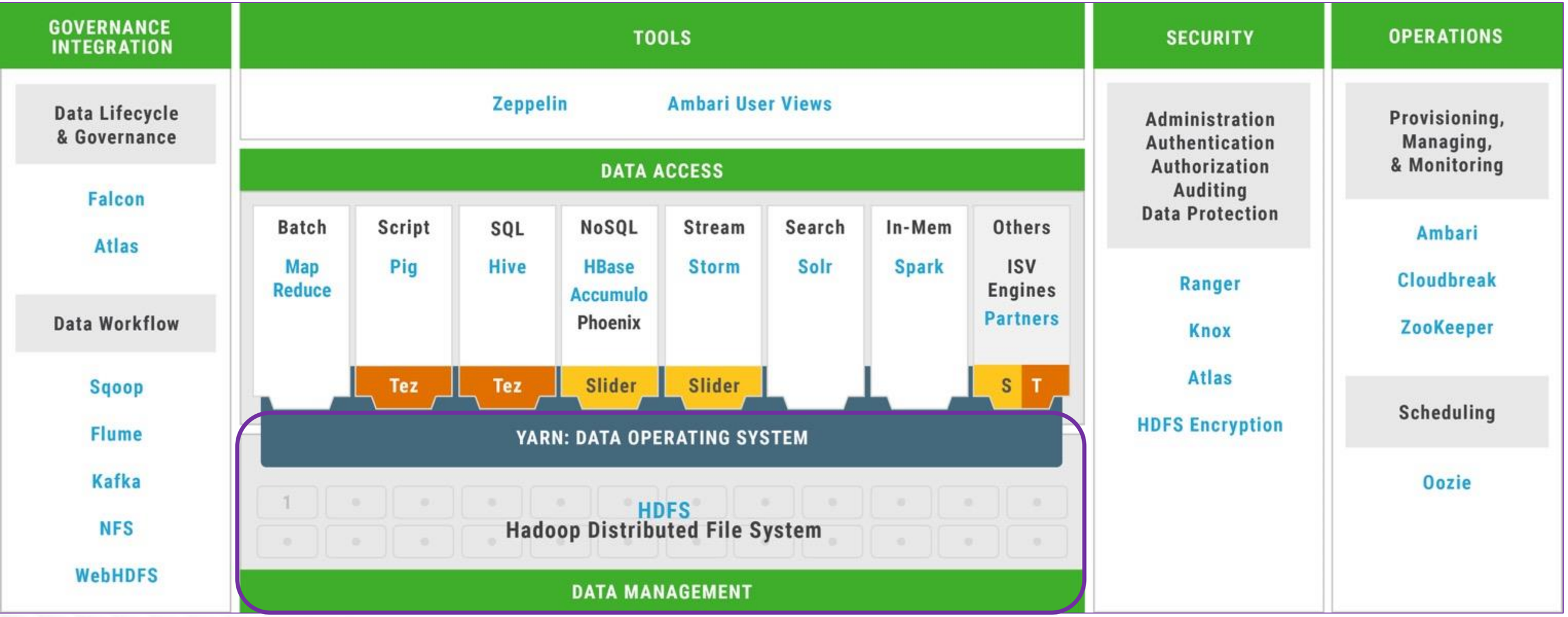
Spark provides libraries for data manipulation, data access and machine learning. It can serve as a one stop language

ORC and Parquet provide file formats for sharing data within the Hadoop ecosystem

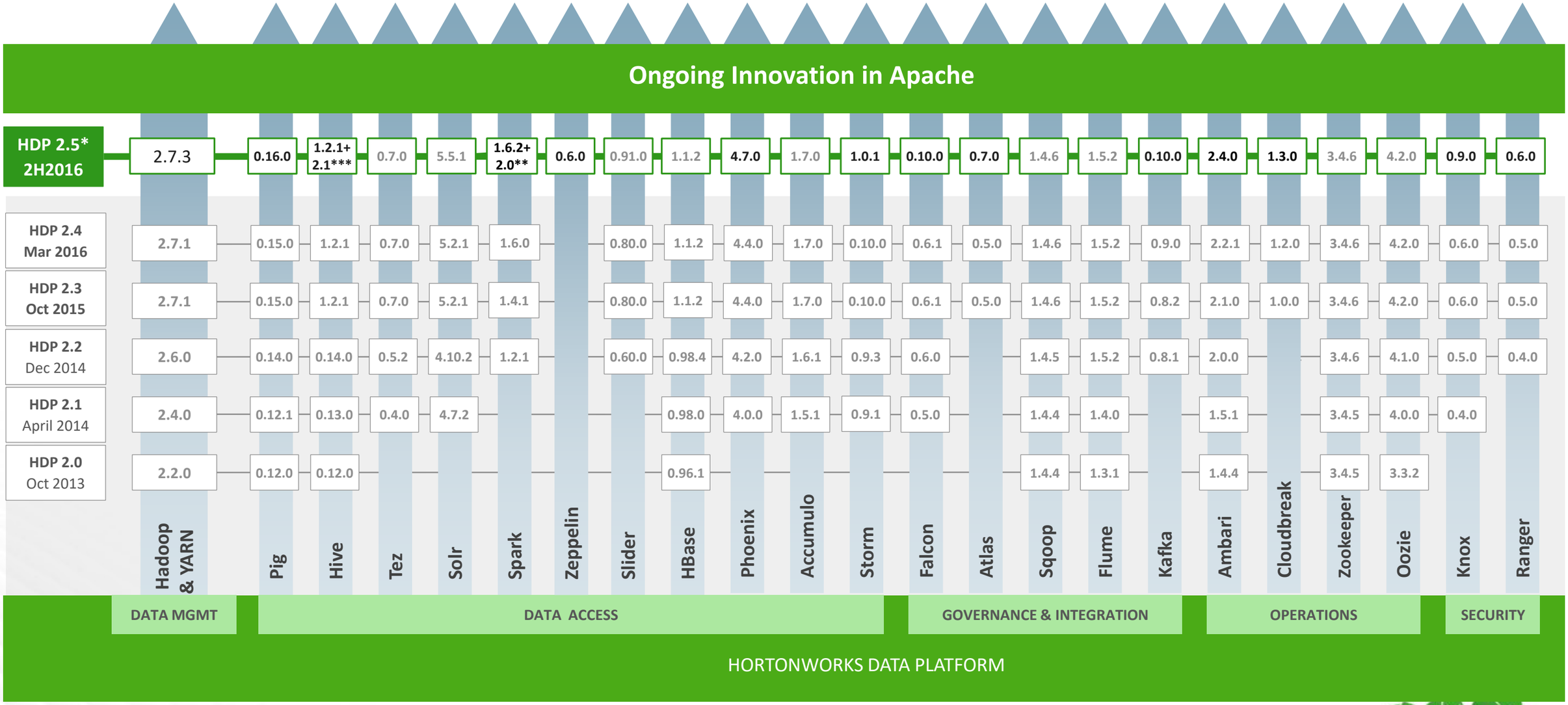
Hive provides a data ware house where data can be served to general consumers who do not require real time access

Zeppelin provide a rich UI to manage time series as they flow the HaaH architecture. The notebook style environment also is a natural home for data scientist who leverage the historical data for insights

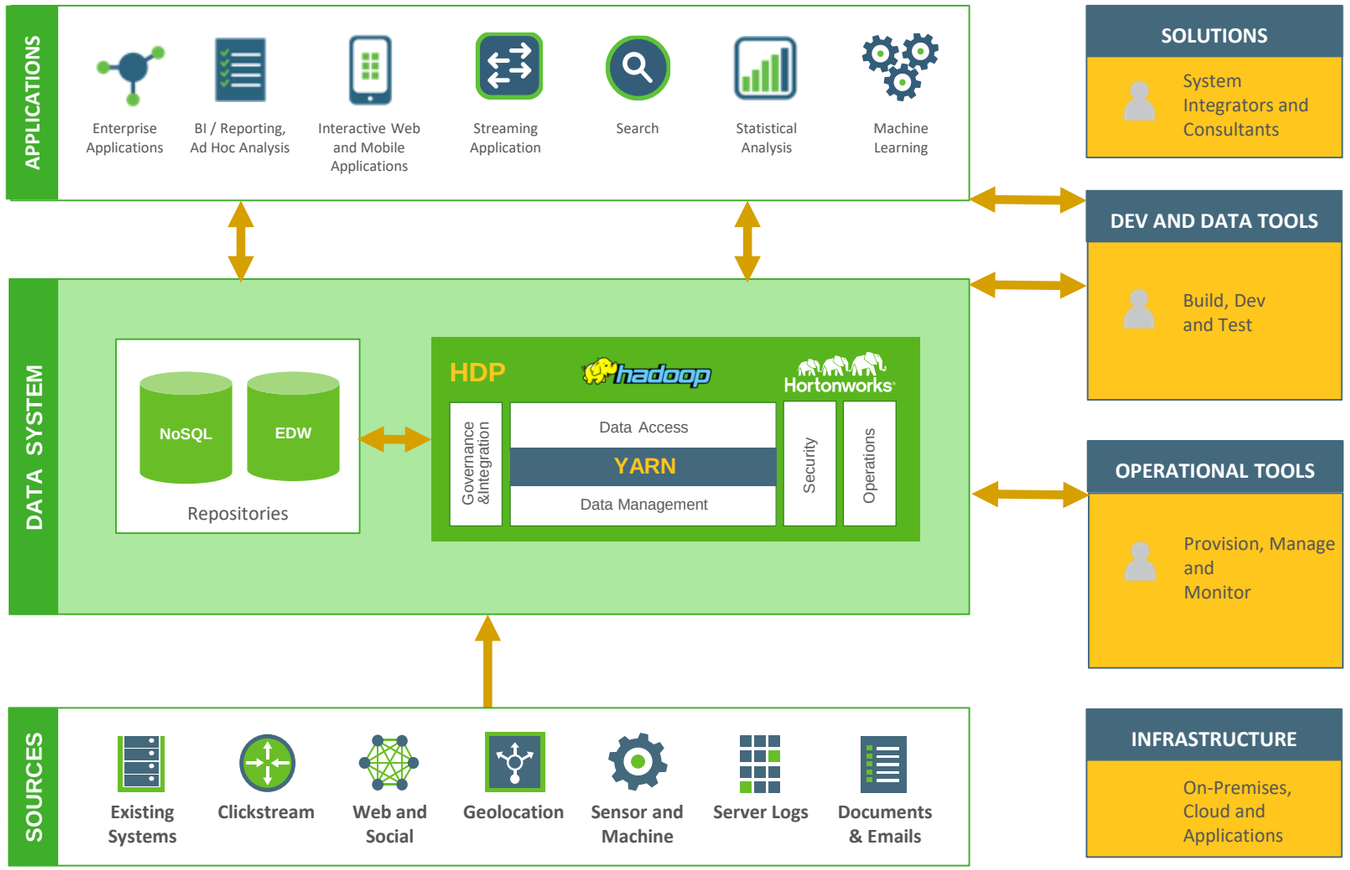
Hortonworks Data Platform Architecture – Data Management



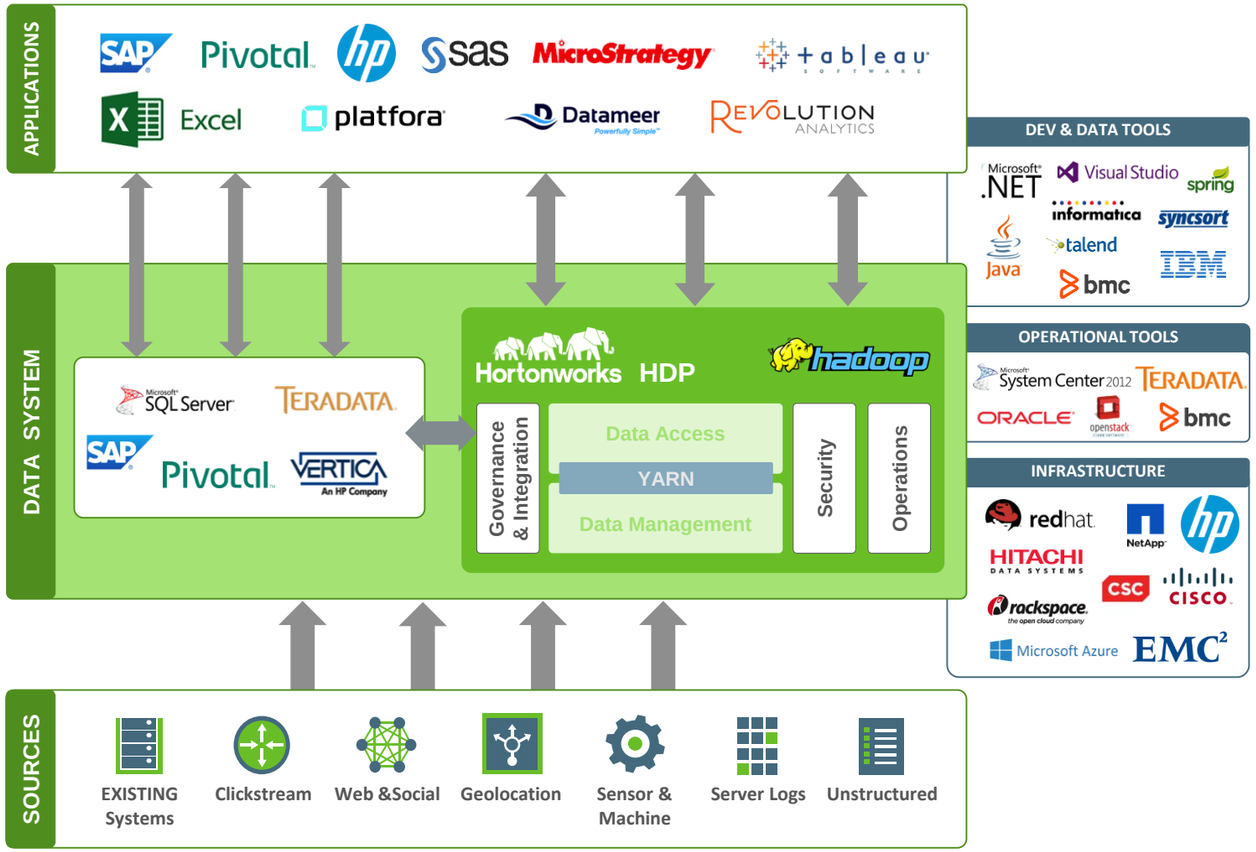
HDP 2.5



Connected Data Platforms



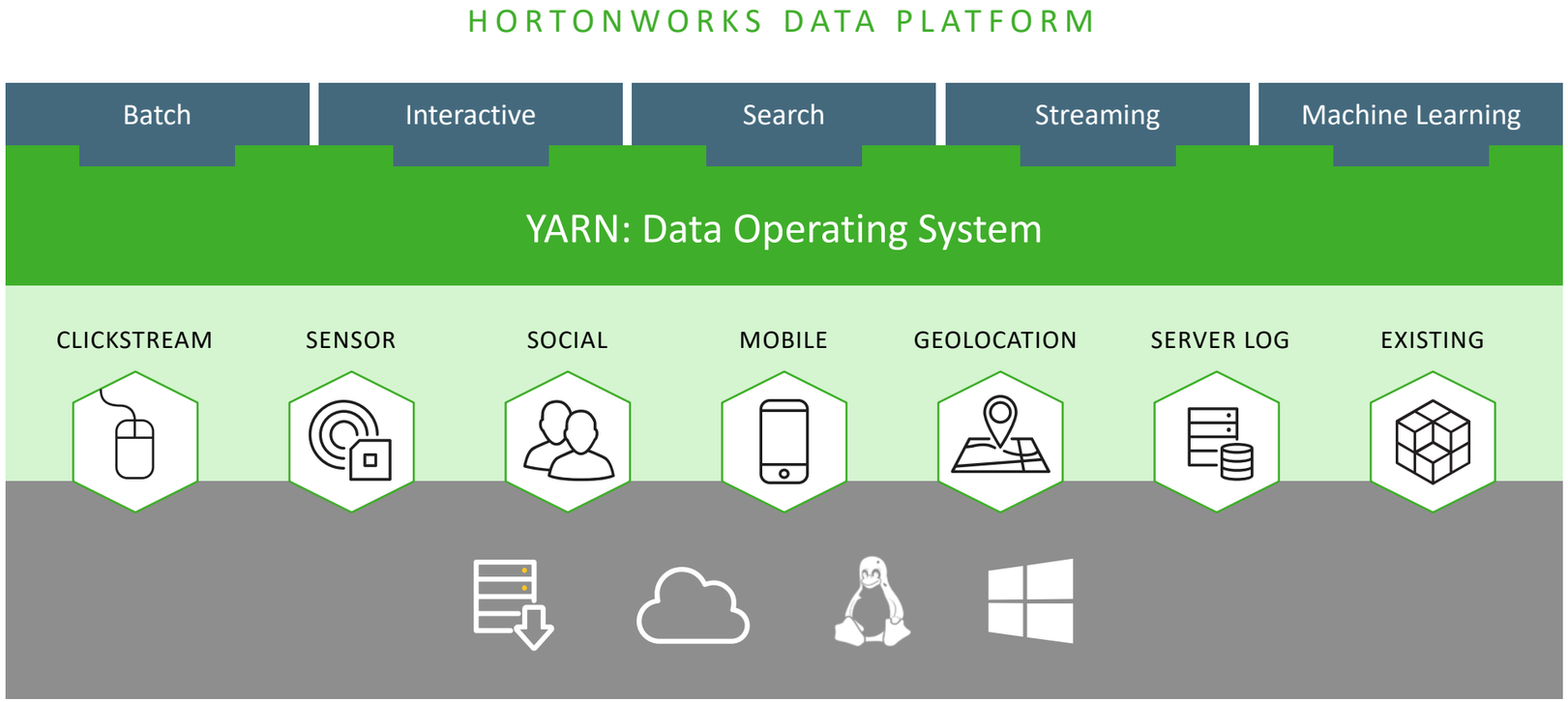
Hadoop as a +1 Architecture



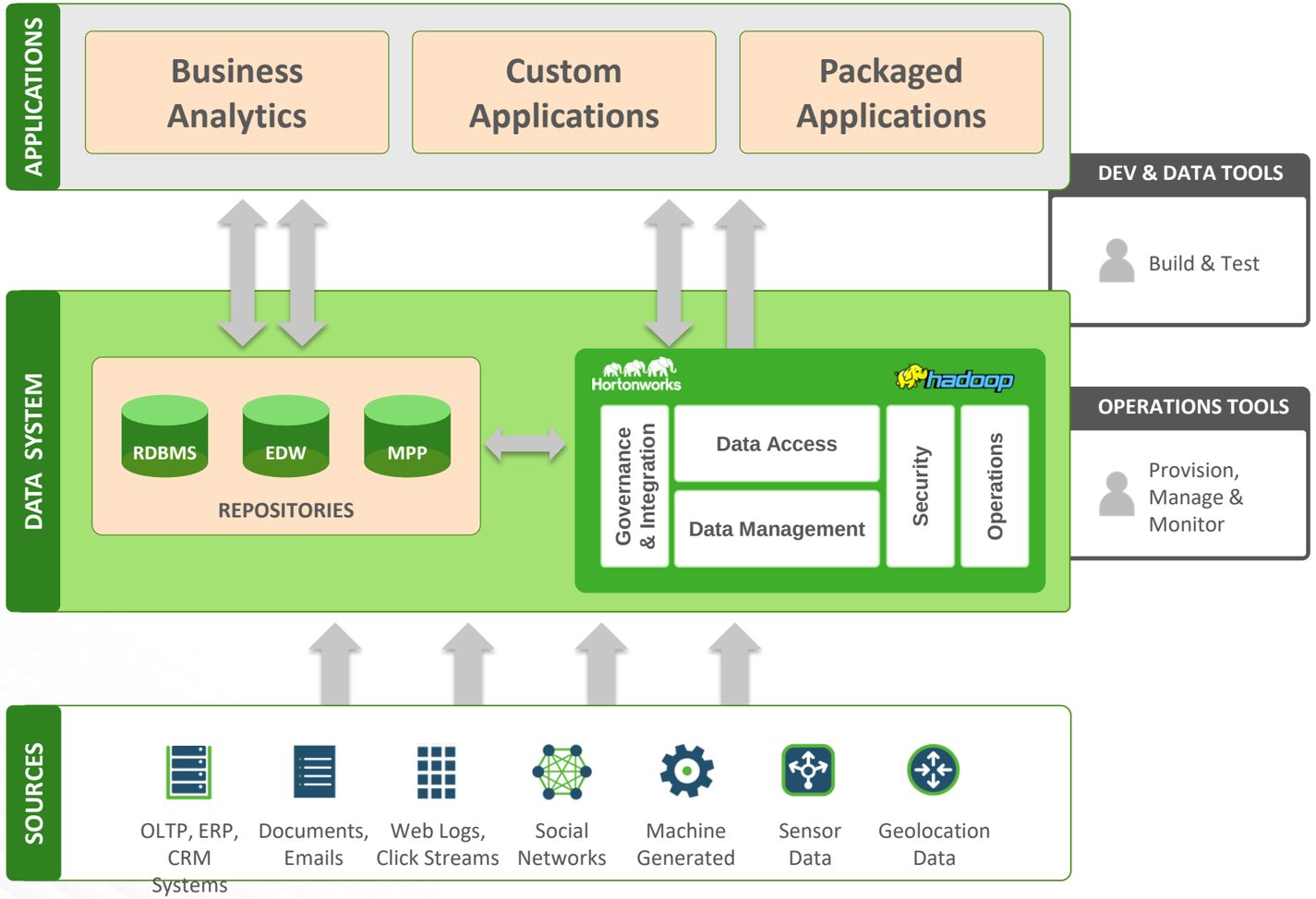
Popular Use Cases

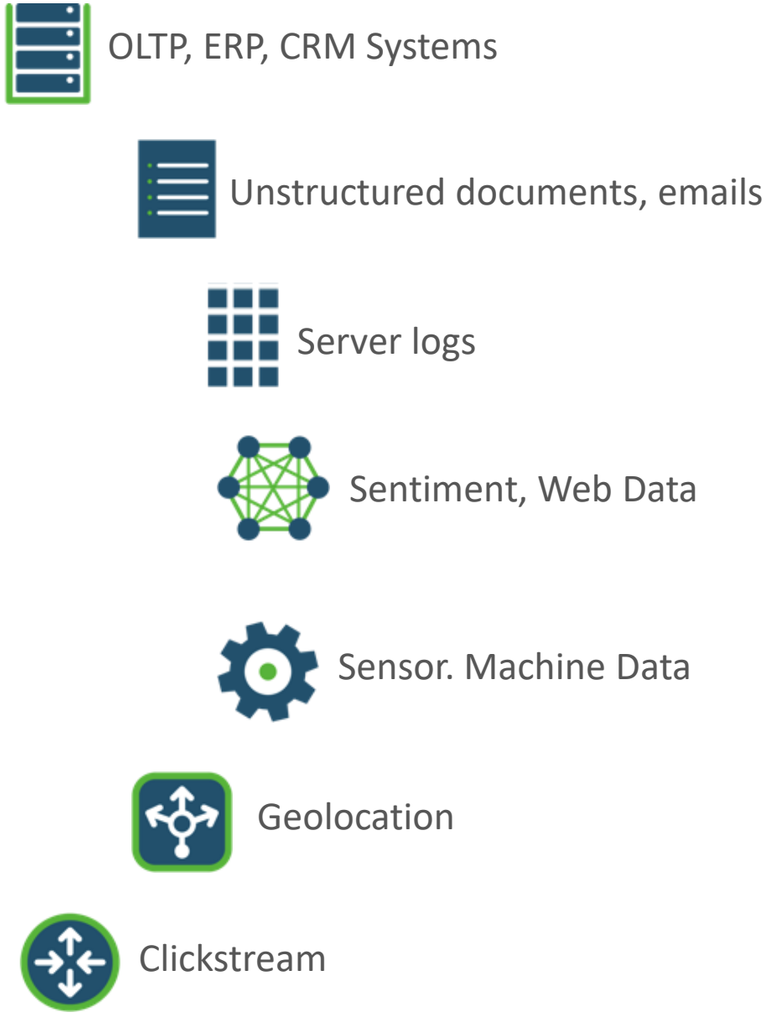
HORTONWORKS®
POWERING THE FUTURE OF DATA™

Hortonworks Delivers Open Enterprise Hadoop



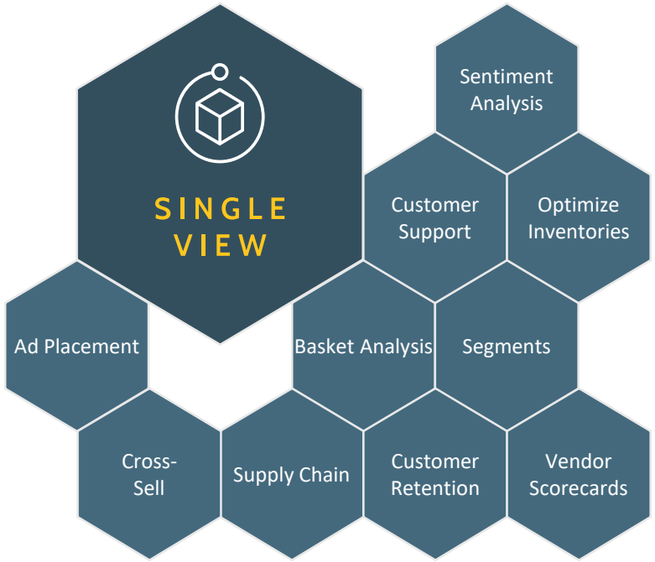
Modern Data Architecture





New Analytic Applications from New Types of Data

INDUSTRY	USE CASE	Sentiment & Web	Clickstream & Behavior	Machine & Sensor	Geographic	Server Logs	Structured & Unstructured
Financial Services	New Account Risk Screens					✓	✓
	Trading Risk					✓	
	Insurance Underwriting			✓	✓		✓
Telecom	Call Detail Records (CDR)			✓	✓		
	Infrastructure Investment			✓		✓	
	Real-time Bandwidth Allocation	✓				✓	✓
Retail	360° View of the Customer		✓			✓	✓
	Localized, Personalized Promotions				✓		
	Website Optimization		✓				
Manufacturing	Supply Chain and Logistics			✓			
	Assembly Line Quality Assurance			✓			
	Crowd-sourced Quality Assurance	✓					
Healthcare	Use Genomic Data in Medial Trials			✓		✓	✓
	Monitor Patient Vitals in Real-Time						
Pharmaceuticals	Recruit and Retain Patients for Drug Trials	✓	✓				
	Improve Prescription Adherence	✓			✓	✓	✓
Oil & Gas	Unify Exploration & Production Data			✓	✓	✓	✓
	Monitor Rig Safety in Real-Time			✓		✓	✓
Government	ETL Offload/Federal Budgetary Pressures					✓	✓
	Sentiment Analysis for Government Programs	✓					

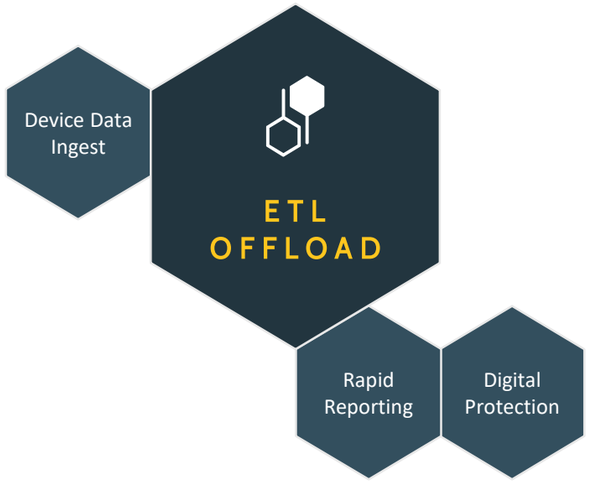


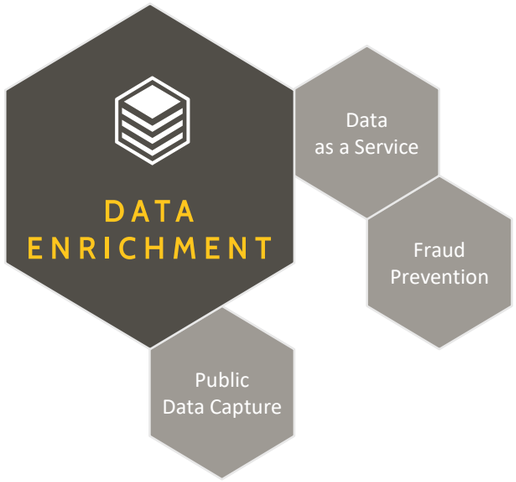
BUSINESS OUTCOMES

Business executives are driving transformational outcomes with next-generation applications that empower new uses of Big Data including: data discovery,
a single view of the customer and predictive analytics.

COST SAVINGS

IT executives are delivering substantial reductions in operating costs by modernizing their data architectures with Open Enterprise Hadoop. These cost saving innovations include active archive of cold data, offloading ETL processes and enriching existing data.





CUSTOMER JOURNEY

Hortonworks® customers leverage our technology to transform their businesses, either by achieving new business objectives or by reducing costs. The journey typically involves both of those goals in combination, across many use cases.

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

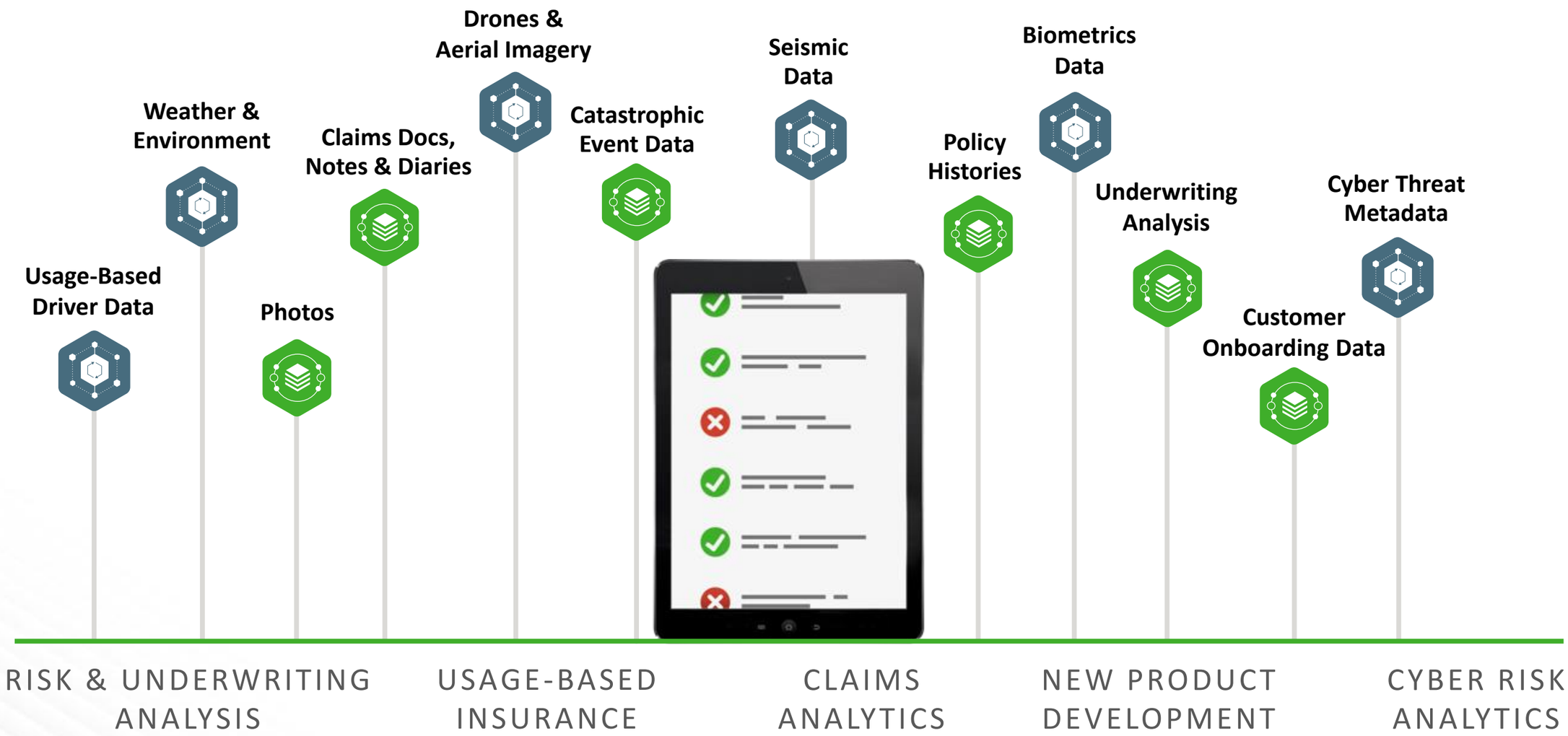
- Analyze public sentiment
- Protect critical networks
- Prevent fraud and waste
- Crowdsource reporting for repairs to infrastructure
- Fulfill open records requests

The Data Journey to Safe Roads

PROGRESSIVE®

HORTONWORKS®
POWERING THE FUTURE OF DATA™

Actionable Intelligence Is Shaping the Modern Insurance Industry



Progressive Rewards Safe Drivers and Improves Traffic Safety

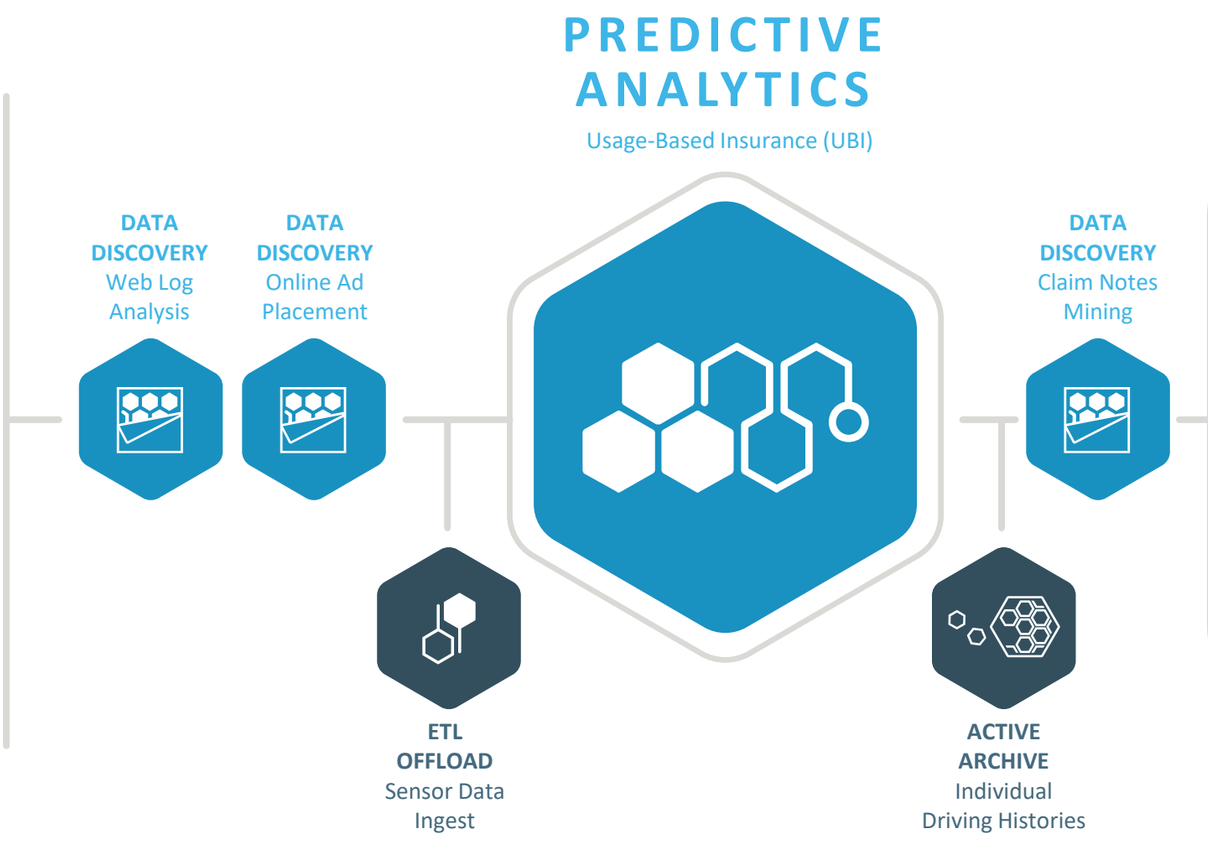
SITUATION

Usage-Based "Snapshot" Insurance Program

In-Car Sensor Captures IoT Data

Existing Data Systems Did Not Scale Efficiently

~7 Days to Transform Only 25% of UBI Data



PROGRESSIVE

100% in 2-3 Days	driving detail captured from Snapshot, in HDF
+12 Billion	miles driven stored
Web App-Enabled	customers see driving detail and improve safety
$2.6 Billion	in 2014 Premiums

"We're looking at datasets that we never dreamed we could look at…It's joining dots that in the past we didn't even know we could join." Pawan Divakarla, Data & Analytics Business Leader

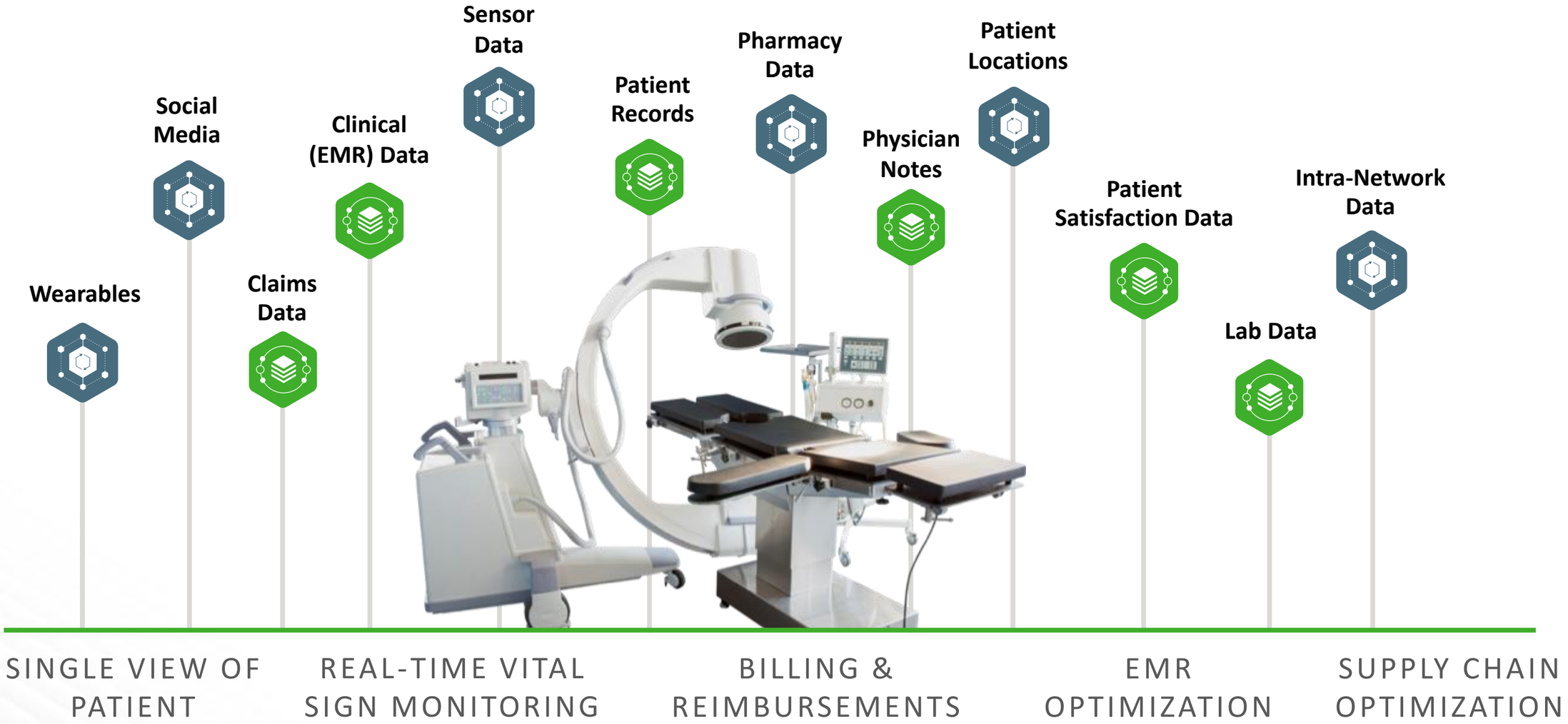
HORTONWORKS

The Data Journey to Better Health

Mercy

HORTONWORKS®
POWERING THE FUTURE OF DATA™

Actionable Intelligence Makes Healthcare Precise and Personal



Mercy Transforms Healthcare Through "One Patient, One Record"

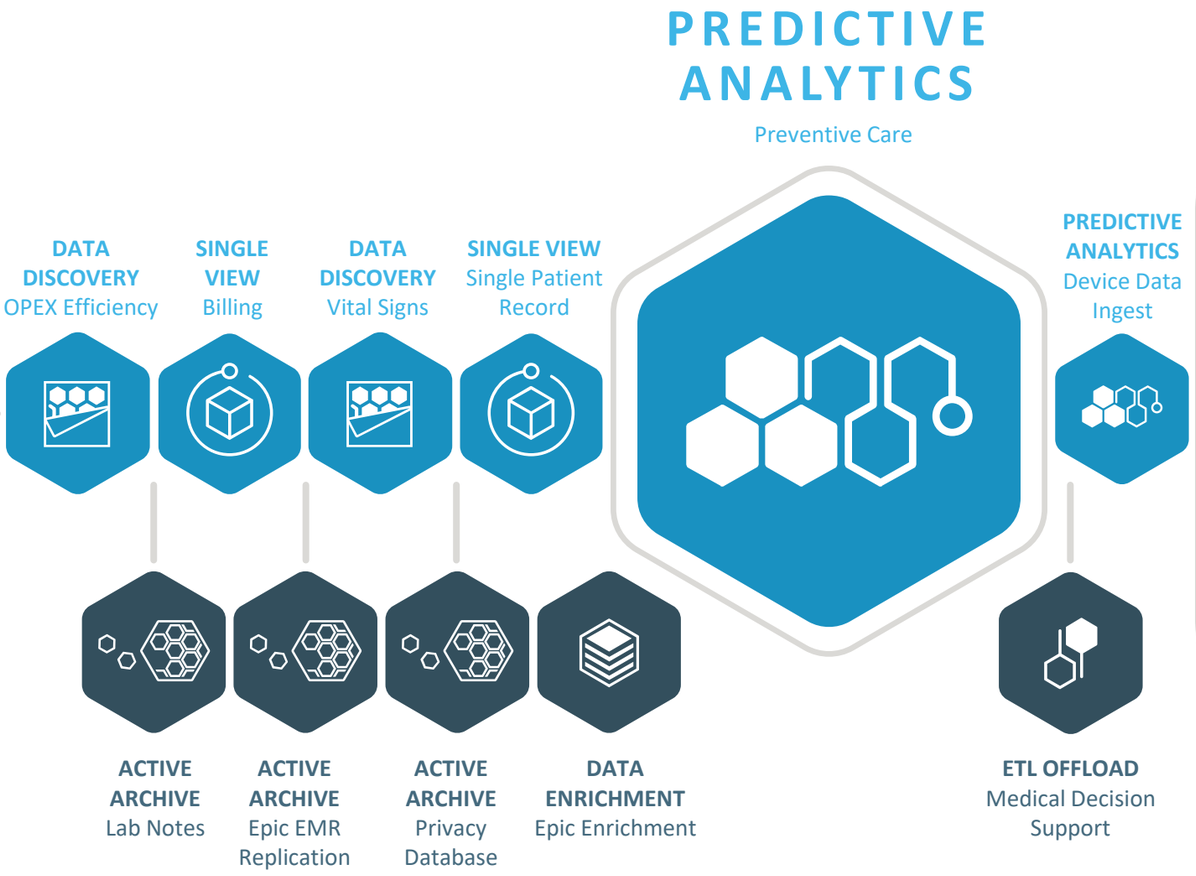
SITUATION

Existing platform impeded goals

Data enrichment needed for 1 million patients

Move to Clarity wouldn't enable real-time analytics

Extracting data from Epic to Clarity took a day



Mercy

3-5 Minutes	move data off Epic to Clarity with HDP
$1M Additional Annual Revenue	from improved billing process
From "Never" to "Seconds"	accelerated researcher insight
900x more data	ingest of ICU vital signs

"[HDP] provides us a place and way to leverage our Epic data in addition to other data that comes from outside of Epic." Paul Boal, Director of Data Management

HORTONWORKS

Webtrends

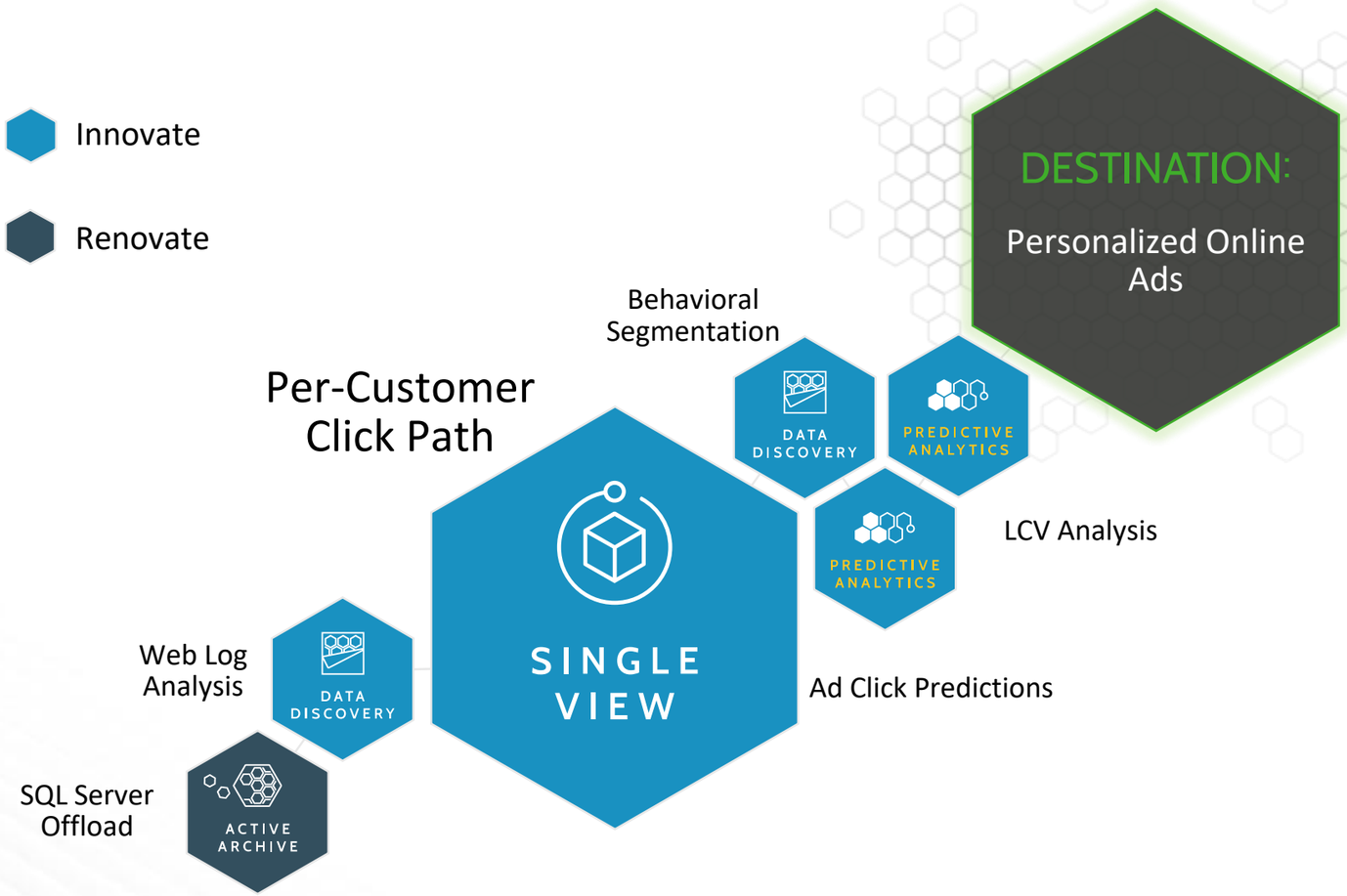
The Data Journey Towards Personalized Online Ads

Massive Volumes of Weblogs Fueled Webtrends Growth—and also its Skyrocketing Storage Costs

Webtrends' Journey

- Webtrends provides digital marketing solutions for more than 2,000 companies in 60 countries – processing 13 billion daily online events
- Data used to be processed in relational databases, stored on large NAS appliances, which were not economical at scale
- Processing occurred on-premises, without cloud-based capabilities
- Diseconomies of scale hampered the company objective to help its customers predict optimal online ad placement

Webtrends' Journey



"We're able to…look at this data set and process it and do predictions, behavioral analysis. We can do things that allow us to determine ROI for different actions and behavioral patterns."

Peter Crossley, Chief Architect

Petabytes of Weblogs Analyzed with Spark at Scale

- Data streams from a vast array of desktop and mobile devices
- 13 billion daily events collected in fewer than milliseconds per event
- No data cleansing necessary prior to analysis with Apache Spark
- 2 clusters consolidated into 1 YARN-based HDP cluster
- Launched new product Webtrends Explore™ – powered by HDP

The Data Journey for Cyber Security

Symantec™

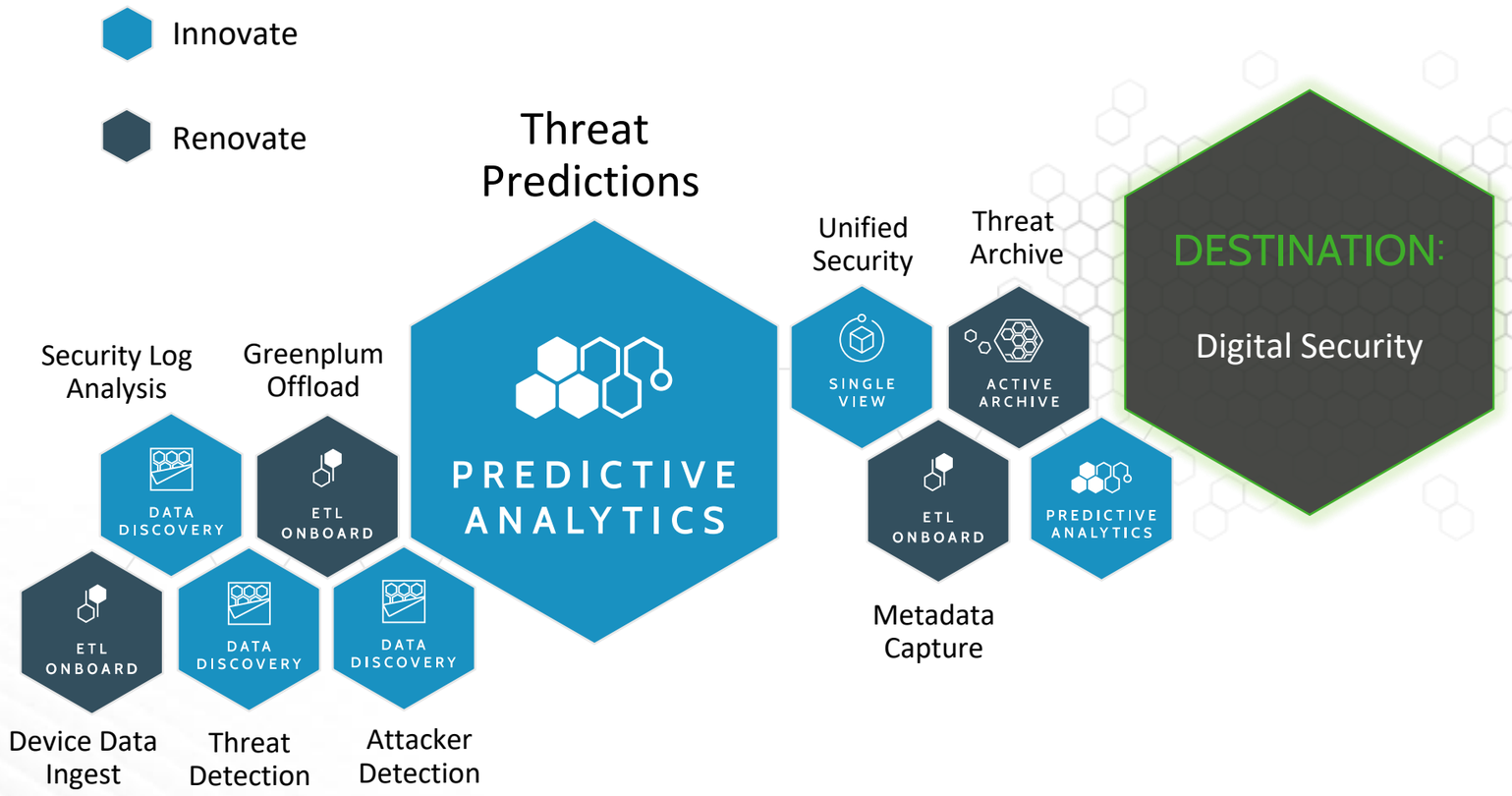
HORTONWORKS®
POWERING THE FUTURE OF DATA™

Symantec's Journey

Analyzing Streaming Threat Data to Increase Velocity for Time to Protection

- The Symantec™ Global Intelligence Network includes more than 57 million attack sensors over 157 countries
- Data streams from 75 million users on 120 million devices
- Legacy platforms created 3-4 hour processing latencies to analyze logs files for digital threats
- Attackers could exploit those processing time windows

Symantec's Journey



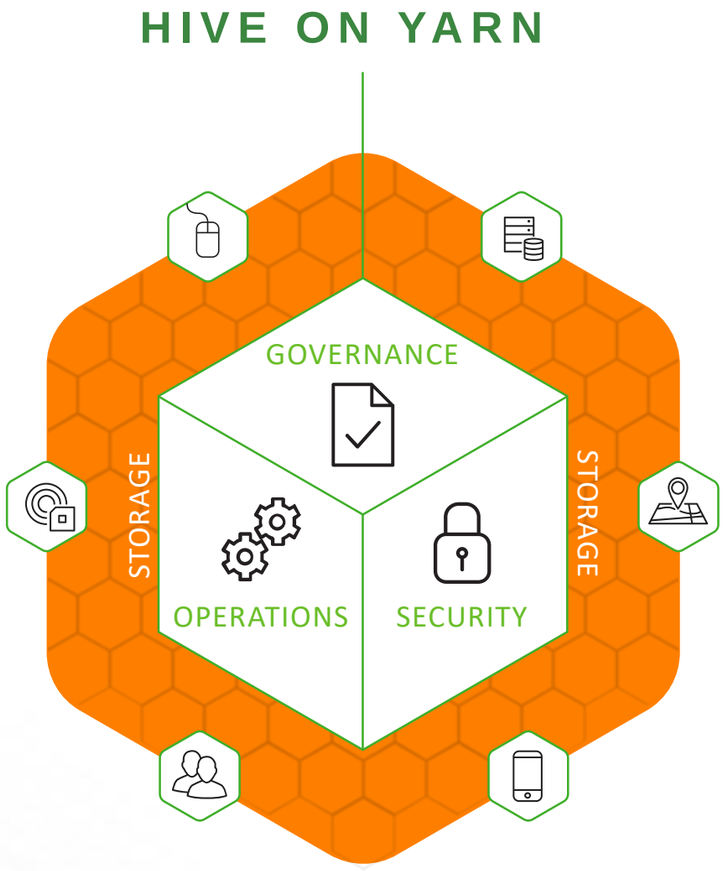
Data Science Speeds Time to Protection

- Threat detection latency reduced from 4 hours to 2 seconds
- Time to protection improved 5000x
- Machine learning over tens of petabytes of historical data predicts threats to customers
- Cloud team uses Ambari and Cloudbreak for dynamic clusters to meet peak workloads

Hive LLAP

HORTONWORKS®
POWERING THE FUTURE OF DATA™

Fast SQL with Apache Hive at Scale



Pluggable Architecture
supports Apache Hive, Pivotal HAWQ and other leading SQL engines

Familiar SQL Query Semantics
enable transactions and SQL:2011 Analytics for rich reporting

Unprecedented Speed at Extreme Scale
returns query results in interactive time, even as data sets grow to petabytes

Apache Hive: Fast Facts

Most Queries Per Hour

100,000 Queries Per Hour
(Yahoo Japan)

Analytics Performance

100 Million rows/s Per Node
(with Hive LLAP)

Largest Hive Warehouse

300+ PB Raw Storage
(Facebook)

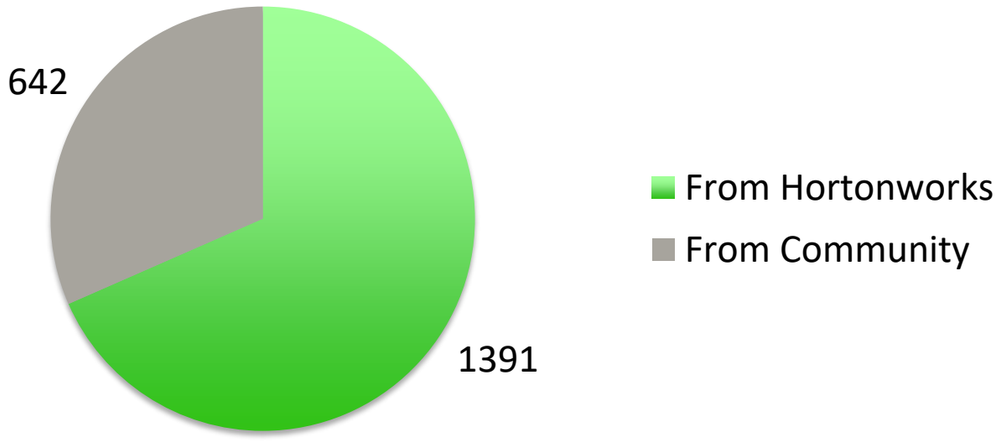
Largest Cluster

4,500+ Nodes
(Yahoo)

HDP 2.5 is a Major Milestone for Hive

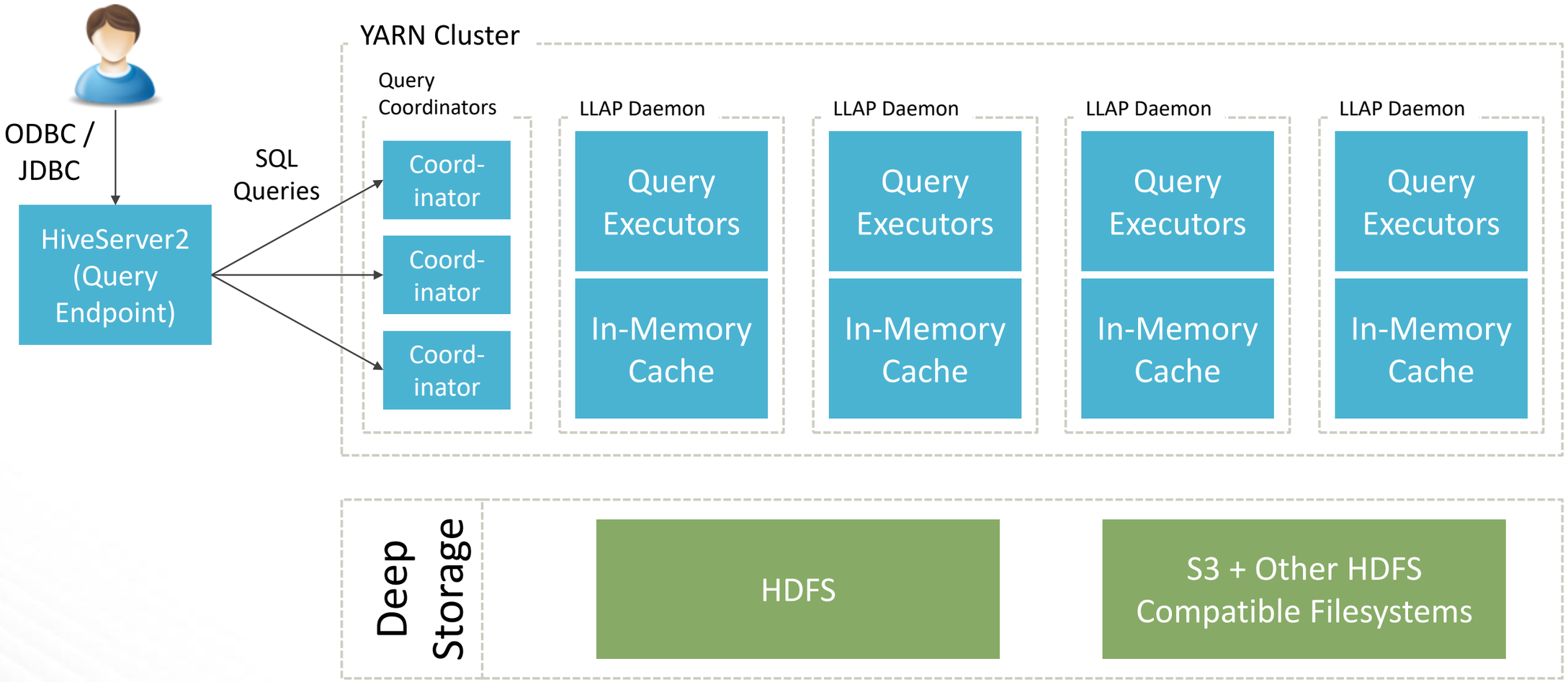
- At a High Level:
  - 2000+ features, improvements and bug fixes in Hive since HDP 2.4.
  - 600+ of these from outside of Hortonworks.
- Major Improvements:
  - Preview: Hive LLAP: Persistent query servers with intelligent in-memory caching.
  - ACID Ready for Production Use: Hardened and proven at scale.
  - Expanded SQL Compliance: More capable integration with BI tools.
  - Performance: Interactive query, 2x faster ETL.
  - Security: Row / Column security extending to views, Column level security for Spark.
  - Operations: LLAP integration in Ambari, new Grafana dashboards.

Hive 2 Improvements

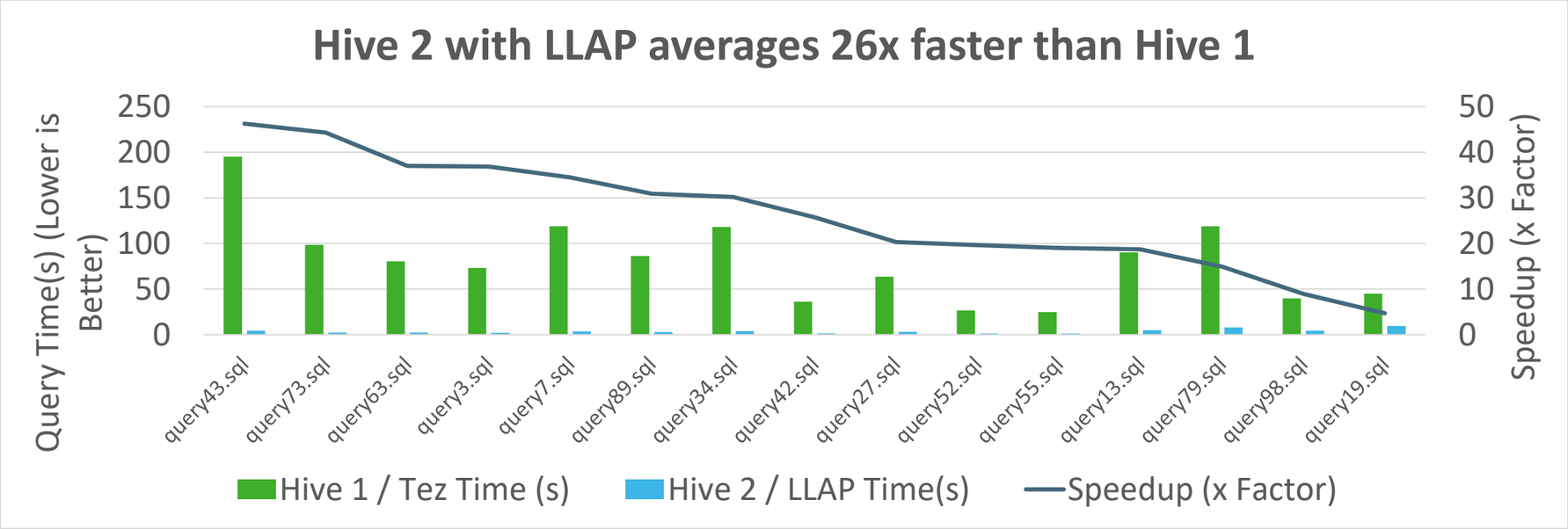


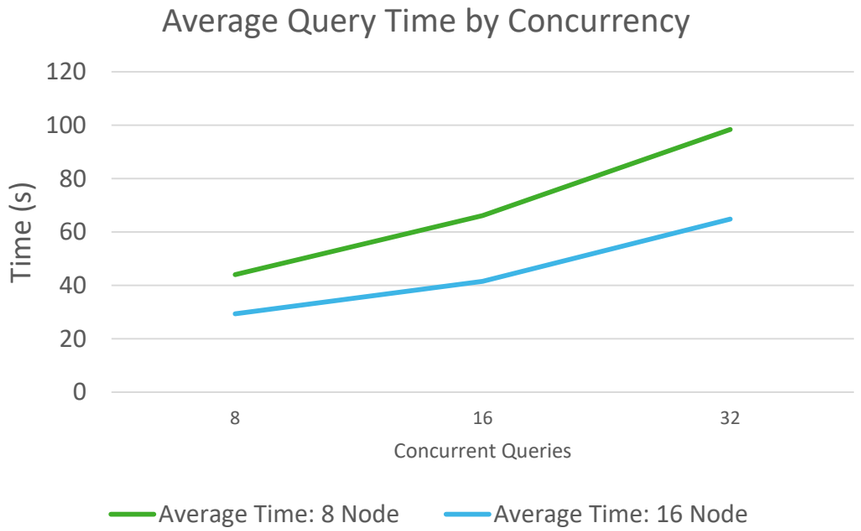
- Interactive Query with Hive LLAP
- SQL ACID Fully Supported
- 2x Faster ETL

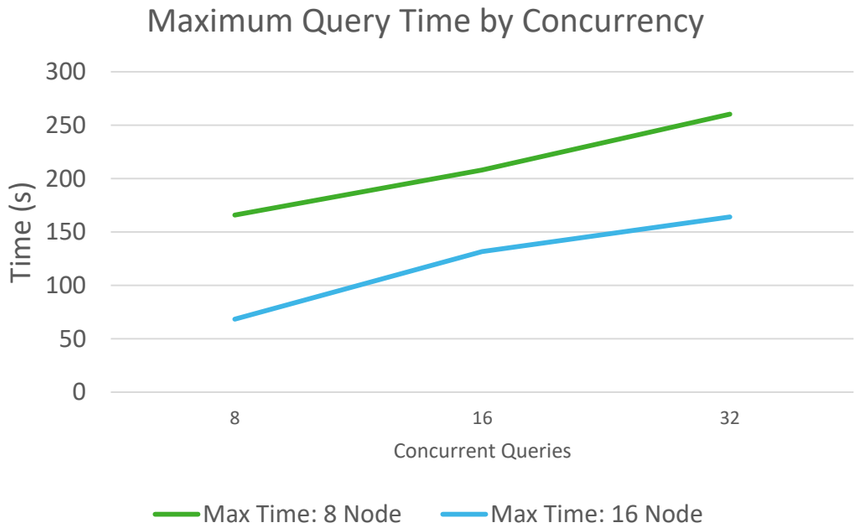
Hive 2 with LLAP: Architecture Overview



Hive 2 with LLAP: Performance and Concurrency







Apache Hive: Journey to SQL:2011 Analytics

Data Types
Numeric
FLOAT/DOUBLE
DECIMAL
INT/TINYINT/SMALLINT/BIGINT
BOOLEAN
String
CHAR / VARCHAR
STRING
BINARY
Date, Time
DATE
TIMESTAMP
Interval Types
Complex Types
ARRAY
MAP
STRUCT
UNION

SQL Features
Core SQL Features
Date, Time and Arithmetical Functions
INNER, OUTER, CROSS and SEMI Joins
Derived Table Subqueries
Correlated + Uncorrelated Subqueries
UNION ALL
UDFs, UDAFs, UDTFs
Common Table Expressions
UNION DISTINCT
Advanced Analytics
OLAP and Windowing Functions
CUBE and Grouping Sets
Nested Data Analytics
Nested Data Traversal
Lateral Views
ACID Transactions
INSERT / UPDATE / DELETE
MERGE

File Formats
Columnar
ORCFile
Parquet
Text
CSV
Logfile
Nested / Complex
Avro
JSON
XML
Custom Formats
Other Features
XPath Analytics

Futures
Procedural Extensions (PL/SQL)
Primary Key / Foreign Key
Non-Equijoin
Scalable Cross Product
Enhanced OLAP
ACID MERGE
Multi Subquery
Comparison to sub-select
INTERSECT and EXCEPT

Legend

- Existing
- Projected: HDP 2.5
- Projected: HDP 3.0

Track Hive **SQL:2011 Complete:** HIVE-13554

Hive SQL:2011 Complete Initiative

- Objective: Provide complete ANSI compliance for SQL:2011 analytical capabilities.
- Identified 9 specific improvements needed to reach this goal.
- Tracked in HIVE-13554.
- To be delivered in HDP 3.

Security Advancements

- Dynamic Column-Level Masking:
  - Value masking or consistent hashing that permits joins.
  - Supports per-user policies.
- Dynamic Row Filtering:
  - Filter rows for security policy compliance.
  - Supports per-user policies.
- Identical Policies for Hive and Spark.

Example: Per-User Row Filtering by Region in Hive/SparkSQL

Spark User 1 (West Region)

Original Query:
SELECT * from CUSTOMERS
WHERE total_spend > 10000

Spark User 2 (East Region)

Dynamic Rewrite:
SELECT * from CUSTOMERS
WHERE total_spend > 10000
AND region = “west”

Query Rewrites based on
Dynamic Ranger Policies

Dynamic Rewrite:
SELECT * from CUSTOMERS
WHERE total_spend > 10000
AND region = “east”

LLAP Data Access

User ID	Region	Total Spend
1	East	5,131
2	East	27,828
3	West	55,493
4	West	7,193
5	East	18,193

Hive Ambari View v1.5.0: More Robust, More Secure

- Robustness: 87 fixes in Q2, 121 fixes in 2016.
- JDBC Support:
  - Enables security / SSL.
  - HTTP / Knox integration.
  - All JDBC options now possible in Hive View.
- Works with Hive 1, Hive 2 and Hive LLAP.

Hortonworks DataFlow for Data in Motion

Powered by Apache NiFi

TM

Real-time

Integrated

Secure

Adaptive

HDF
HORTONWORKS
DATAFLOW
powered by Apache NiFi

HORTONWORKS

HDF Use Cases

Data Ingestion

- Optimize Log Collection & Analysis
  Optimize log analytics such as Splunk with HDF for content based routing from the edge and HDP for lower cost storage options
- Ingest telemetry for Cyber Security
  Integrated, easy and secure telemetry collection for real-time data analytics and threat detection
- Capture IoT Data
  Transport disparate and often remote IoT data in real time, despite any limitations in device footprint, power or connectivity— avoiding data loss

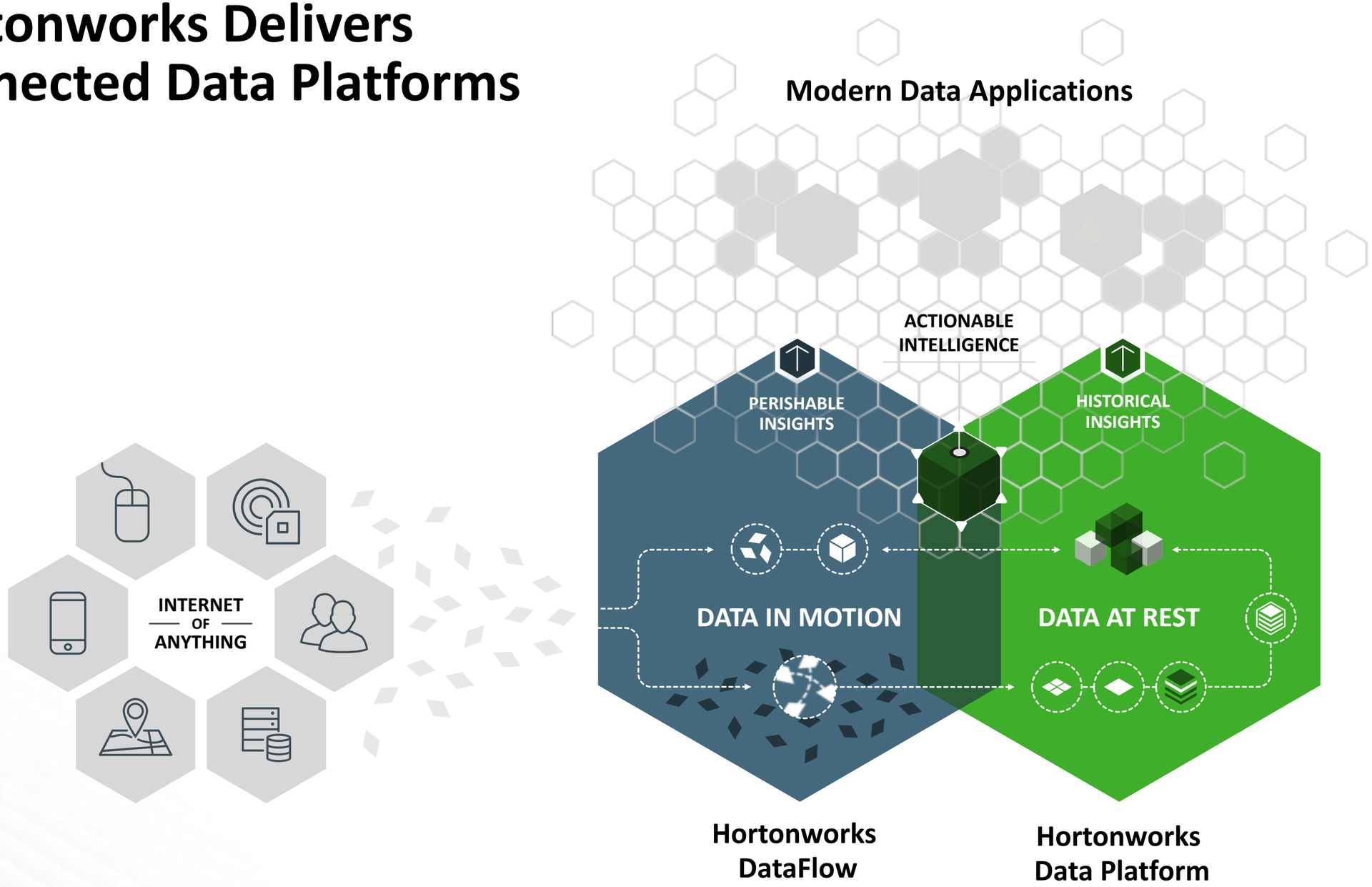
Data Movement

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

Streaming Analytics

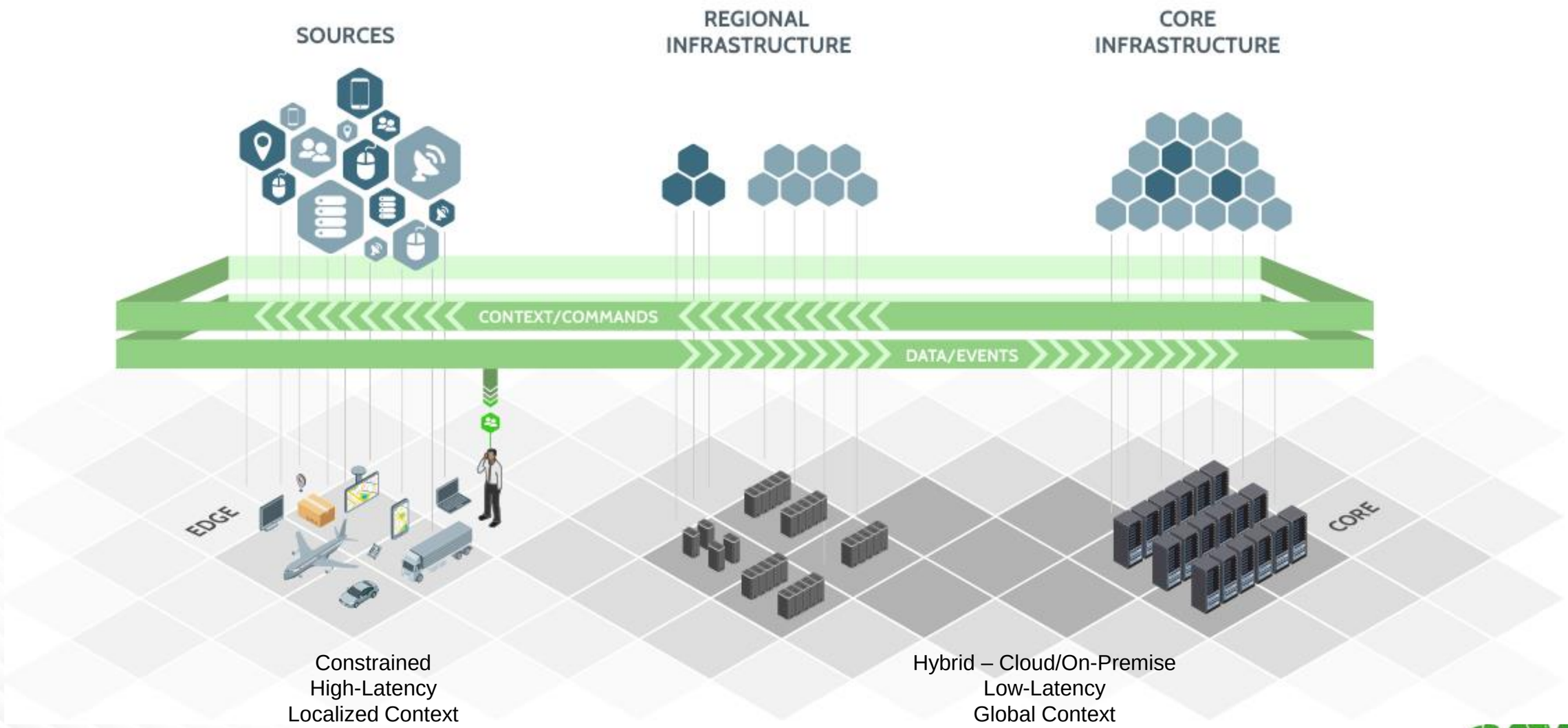
Accelerate big data ROI by moving streaming data into analytics systems such as Apache Storm or Spark Streaming faster & easier

Hortonworks Delivers Connected Data Platforms



HDF Provides “Data Plan of Control” by Managing IoT Dataflows

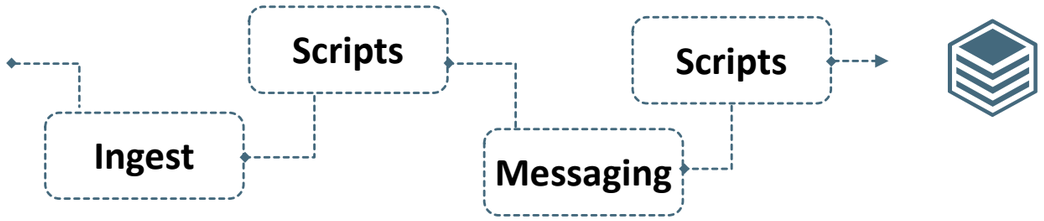
Data source agnostic collection of data across heterogeneous environments



Integrated Processes and Control

COMMON ARCHITECTURE

WITHOUT HORTONWORKS DATAFLOW

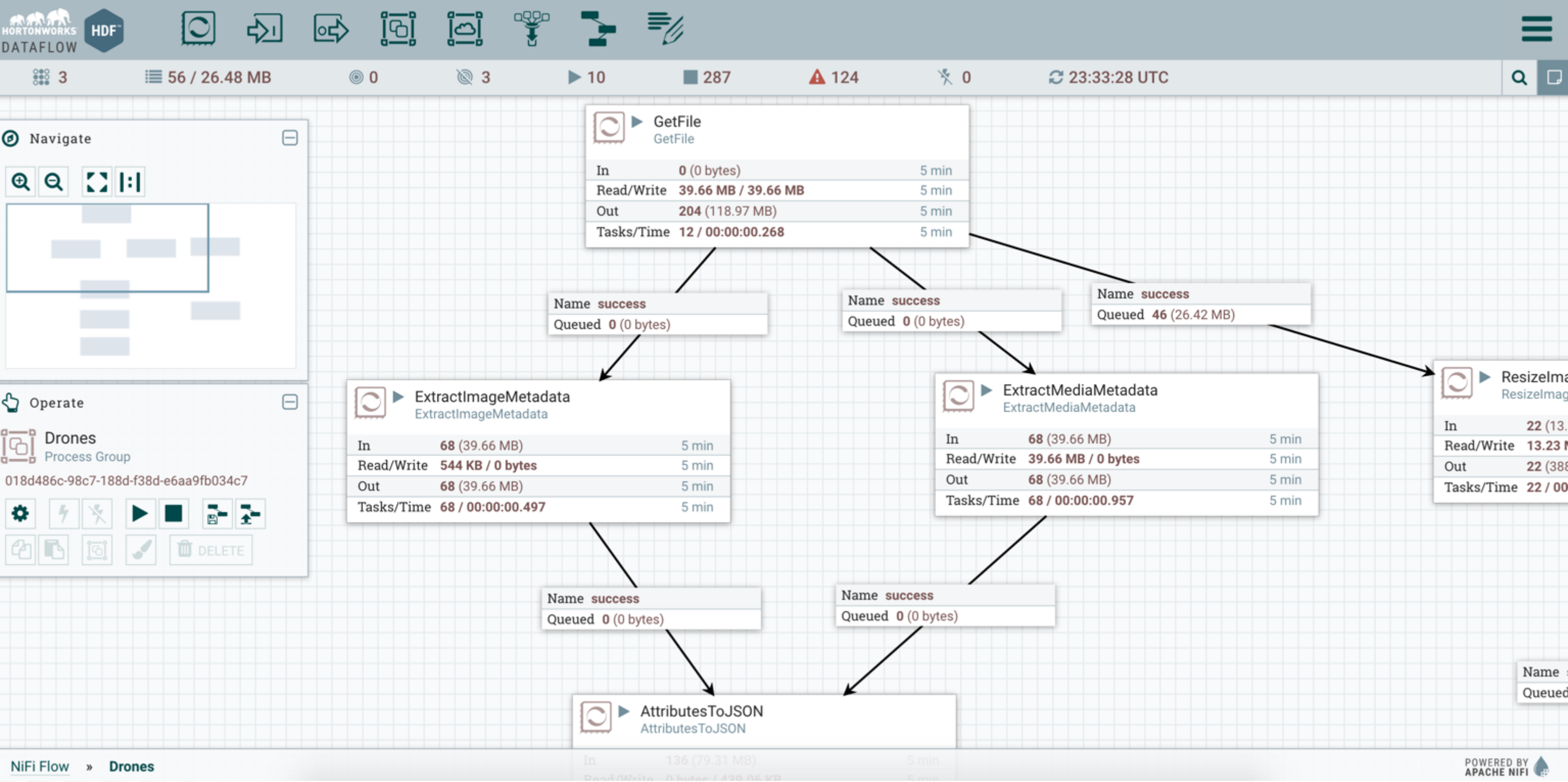


WITH HORTONWORKS DATAFLOW

Optimize the Architecture
Reduce cost an complexity with the most efficient data collection technologies

Assure Efficient Operations
Via real-time control of data inputs, outputs, transportation and transformations

Rely on a Common Foundation
Eliminating dependence on multiple customized systems



Data Ingest

- Demo: HDF NiFi pulls in BeBop 2 Drone images
- Demo: HDF NiFi routes and parses metadata from drone images including geodata
- Demo: HDF NiFi uses TensorFlow Inception v3 to recognize objects in image
- Demo: HDF stores images, metadata and enriched data in HDP.
- Demo: HDF NiFi calls ML Vision REST APIs from vendors
- Kafka Sends Job to Spark Streaming for Anomaly Detection and supervised machine learning
- Kafka Sends Job to Spark for VORA connection to SAP HANA