
HORTONWORKS
Powering the
Future of Data

# Agenda

- **Solution Overview** - *35 mins*
- **Sentiment Analysis Lab Demo** - *15 mins*
- **QA** – *10 mins*
- **Lab** – *120 mins*





HORTONWORKS

# About Hortonworks




HORTONWORKS

# About Hortonworks

## Customer Momentum

- +1 000 customers
- Doubled customer based in 2015
- Publicly traded on NASDAQ: HDP

## The Leader in Connected Data Platforms

- Hortonworks DataFlow for data in motion
- Hortonworks Data Platform for data at rest
- Powering new modern data applications

## Partner for Customer Success

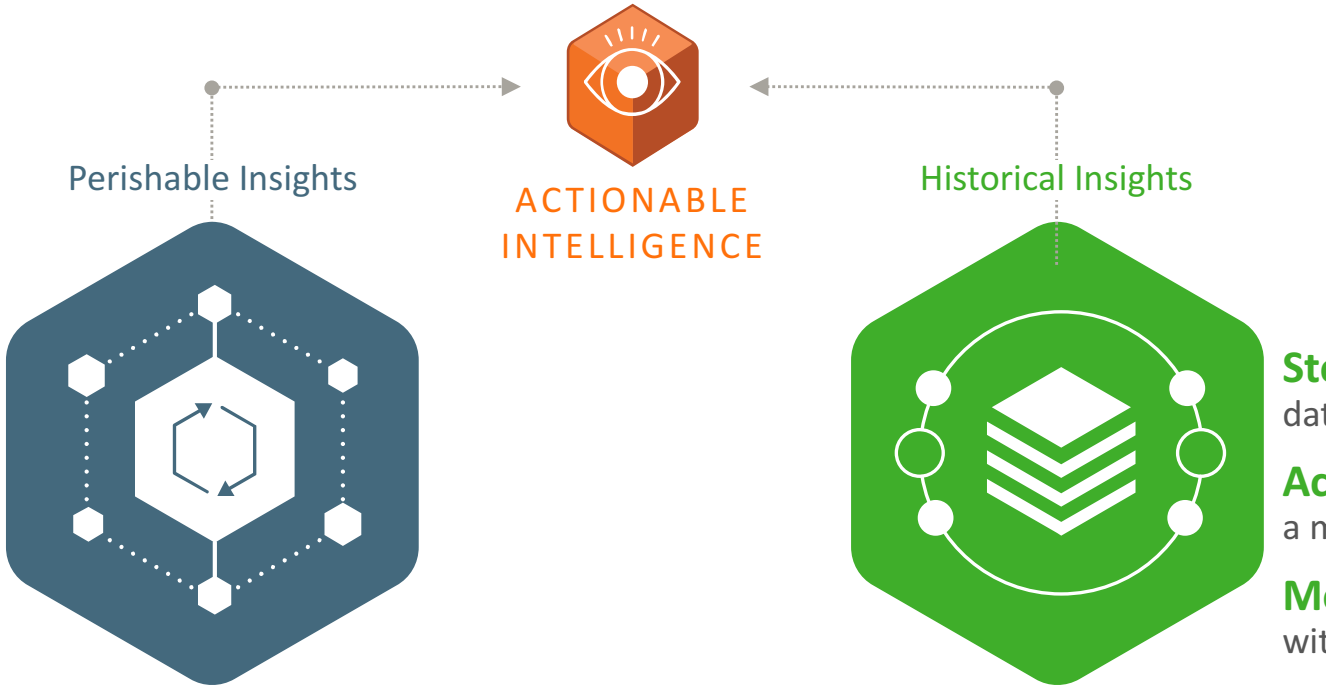
- Leader in open-source community, focused on innovation to meet enterprise needs
- Unrivaled support subscriptions





HORTONWORKS®

# A Connected Data Strategy Solves for All Data







HORTONWORKS

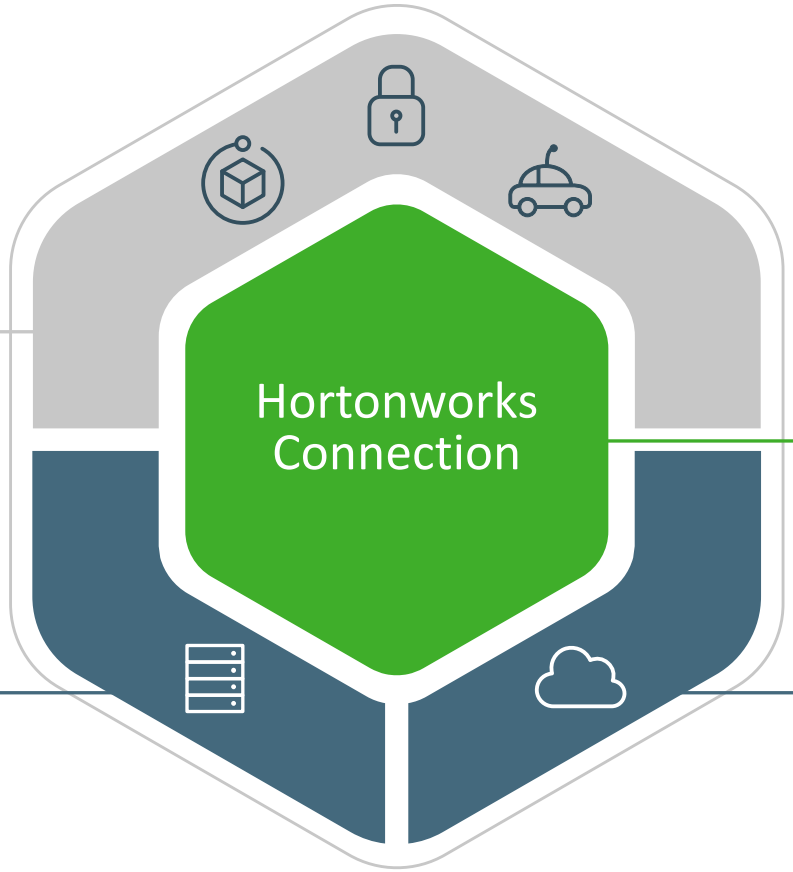
# Hortonworks Connection: Services and Solutions for Your Success

## Hortonworks Solutions

Enterprise Data Warehouse Optimization

Cyber Security and Threat Management

Internet of Things and Streaming Analytics




## Hortonworks Connection

Enablement Subscription

SmartSense™

Premier Operational Support

Educational Services

Professional Services

Community Connection

## Data Center

Hortonworks Data Suite

HDP HDF

## Cloud

Hortonworks Data Cloud

AWS HDInsight

## Data Services




HORTONWORKS

# GA Cloud Offerings: Microsoft Azure and Amazon Web Services


## Azure HDInsight

- Market-leading, feature-rich
- Microsoft branded & supported
- Hourly pricing (with annual option)
- Billed through customer's Azure account
- 24x7 Microsoft support with Azure pricing
- Hortonworks Community Connection (HCC)


## Hortonworks Data Cloud for AWS

- Enterprise security, governance and operations
- Hortonworks branded & supported
- Hourly pricing (with annual option)
- Billed thru customer's AWS account
- 24x7 Hortonworks enablement sold separately
- Hortonworks Community Connection (HCC)



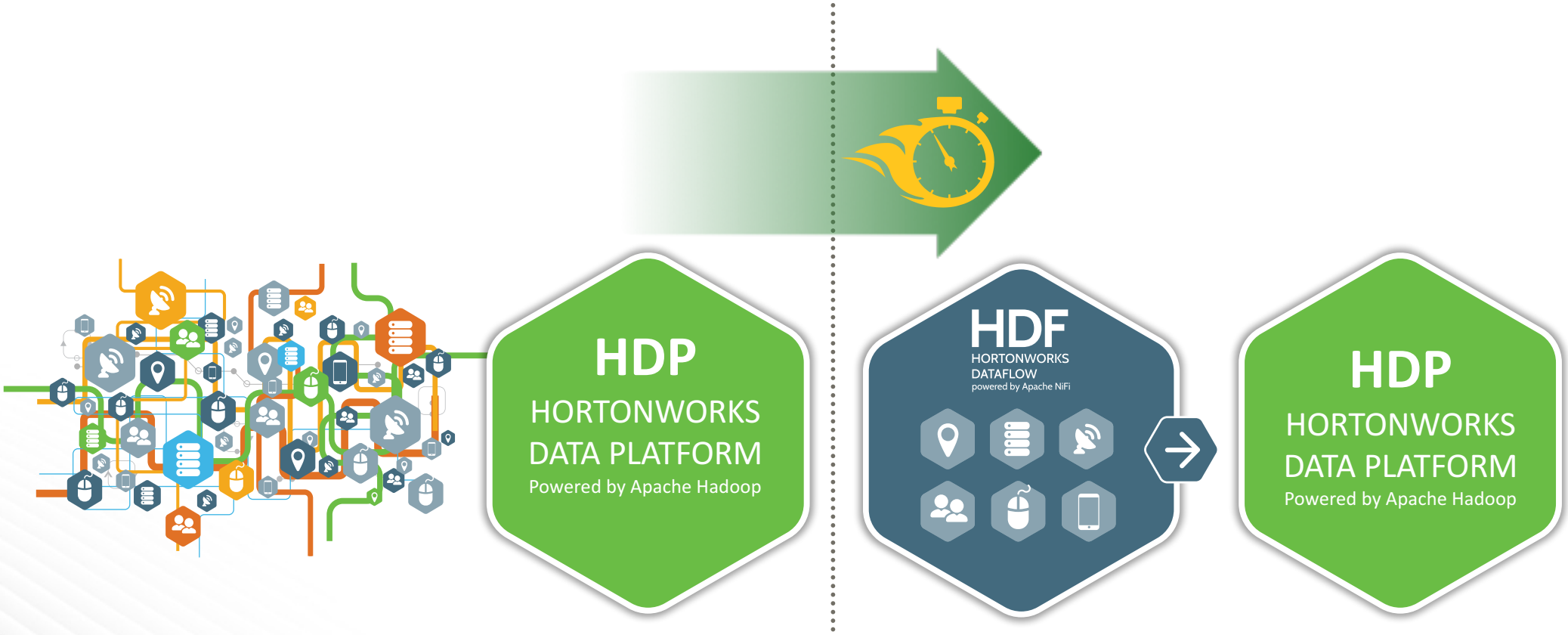

HORTONWORKS

# Hortonworks Data Flow Overview




HORTONWORKS

# HDF Makes Big Data Ingest Easy

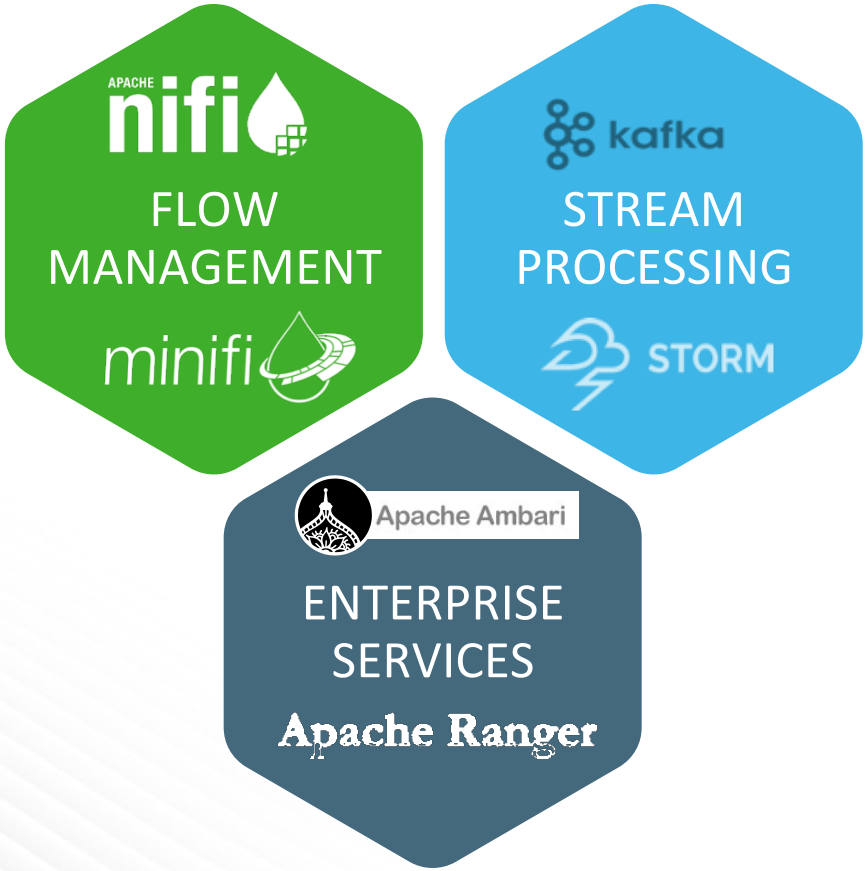






HORTONWORKS

# Hortonworks DataFlow Powered by Apache NiFi, Storm & Kafka




**Streaming Analytics**
Capture perishable insights from data-in-motion

**Visual Control Over Data Flows**
to manage who can see and touch data in transit

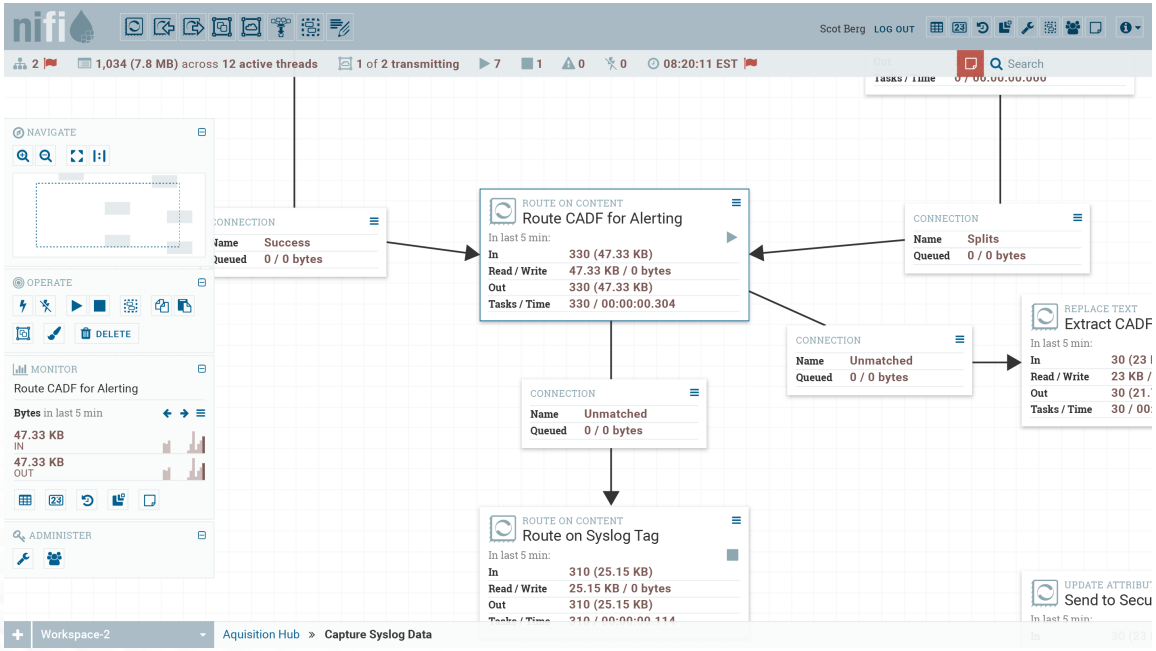
**End-to-End Security**
to encrypt, decrypt, and filter data on its journey

**Real-Time Traceability**
Rich metadata and contextual detail helps troubleshoot security issues



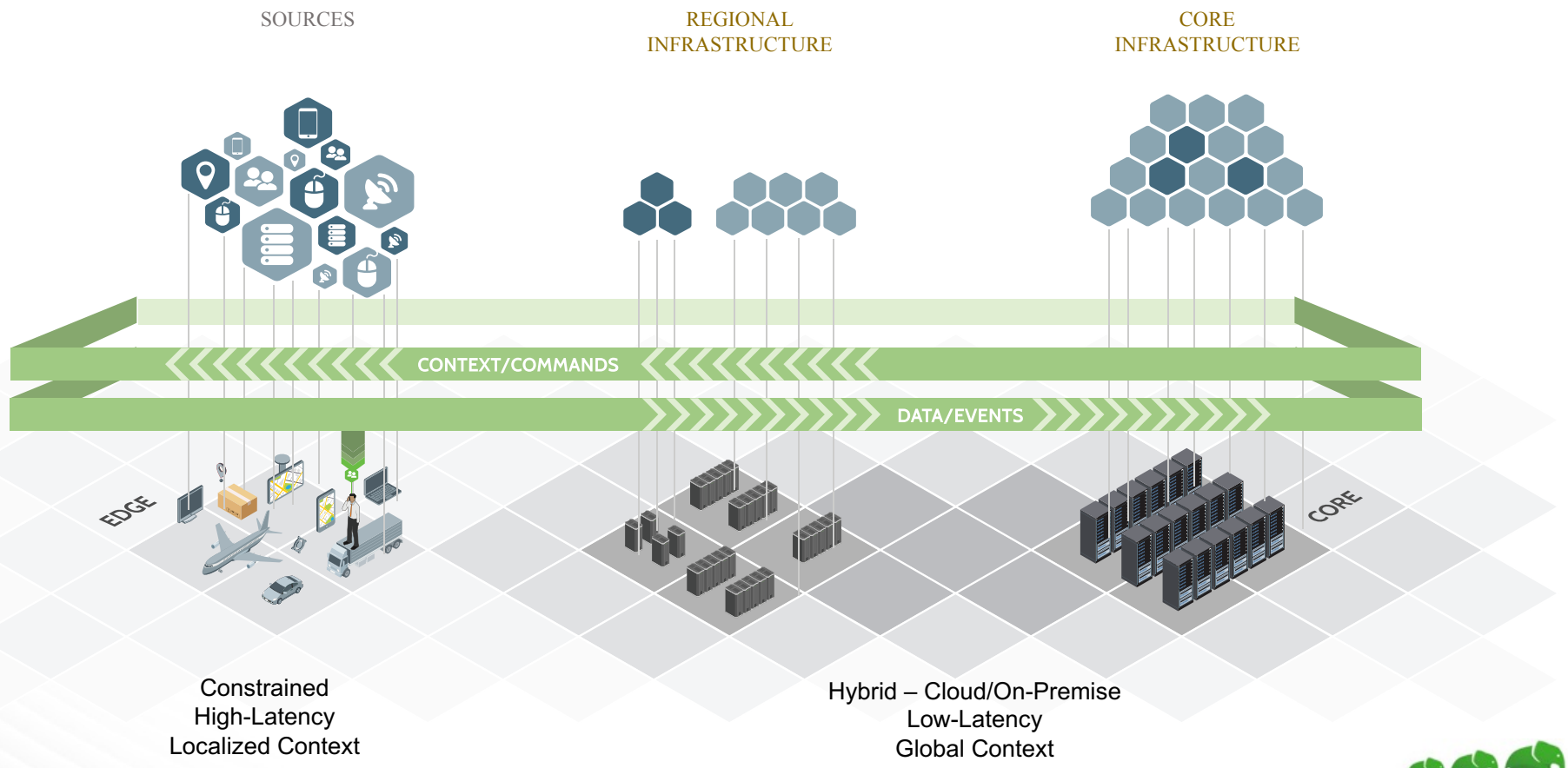

HORTONWORKS

# Apache NiFi
## Key Features




- Visual User Interface
  - Drag and drop for efficient, agile operations
- Immediate Feedback
  - Start, stop, tune, replay data flows in real-time
- Adaptive to Volume and Bandwidth
  - Any data, big or small
- Event Level Data Provenance
  - Governance, compliance & data evaluation
- Secure Data Acquisition & Transport
  - Fine grained encryption for controlled data sharing and selective data democratization

# HDF Manages Dataflow







HORTONWORKS

# Hortonworks Data Platform Overview




HORTONWORKS

# What is Apache Hadoop?

***Hadoop was designed for Big Data***

HDFS – Hadoop Distributed File System
- Data broken into blocks and replicated 3x
- Automatically replaces lost data / computers
- Very high bandwidth, not IOPs optimized

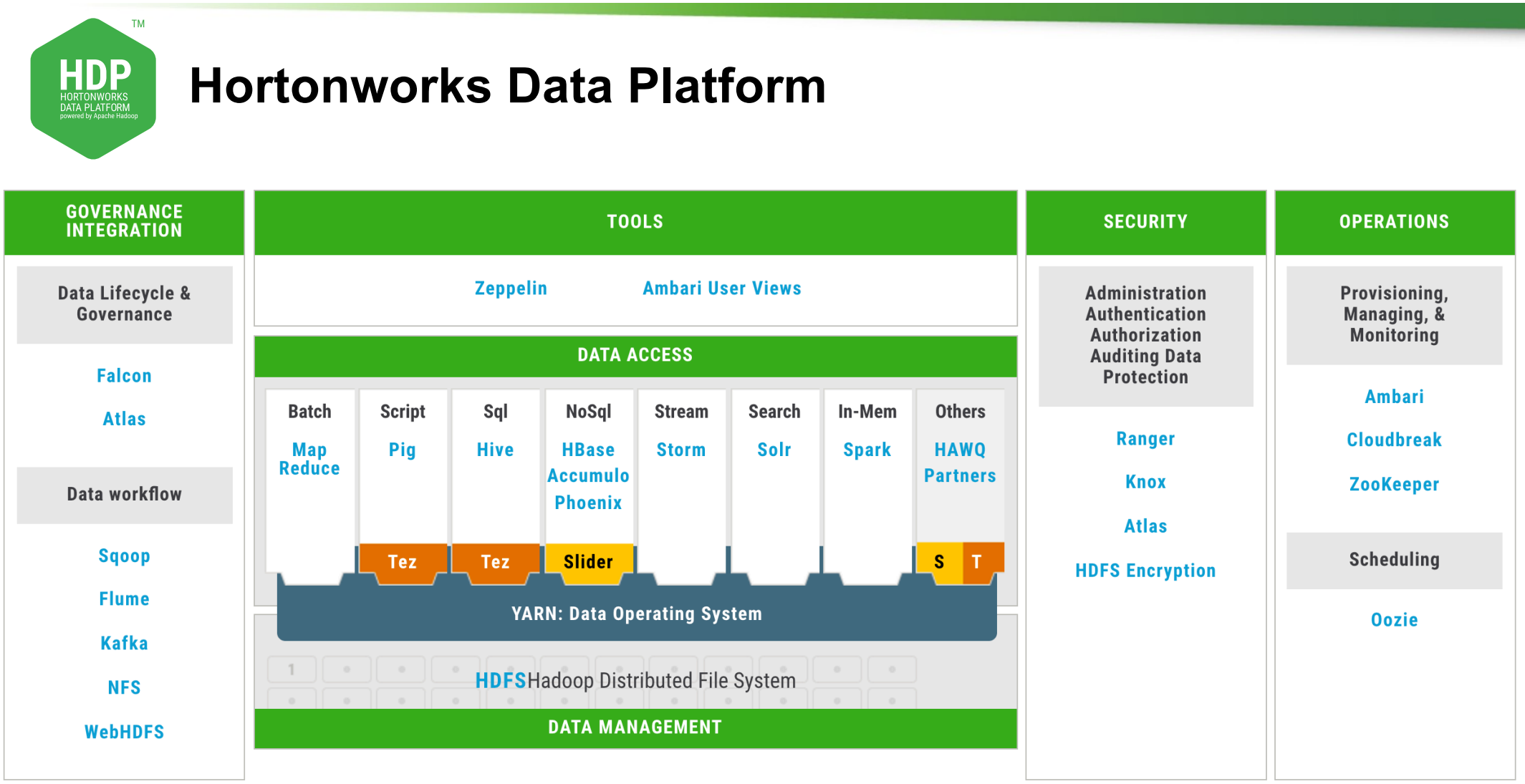
YARN – Distributed Computation Layer
- Distributed execution on HDFS
- Many programing models
  - MapReduce, SQL, Streaming, ML…
- Multi-users, with queues, priorities, etc…

- **Cost Effective**
  - Commodity hardware
  - Open source software
- **Scalable**
  - Efficiently store and process petabytes of data
  - Grows linearly by adding commodity computers
- **Reliable**
  - Self healing as hardware fails or is added
- **Flexible**
  - Store all types of data in many formats
  - Schema-on-read




HORTONWORKS

# Hortonworks Data Platform



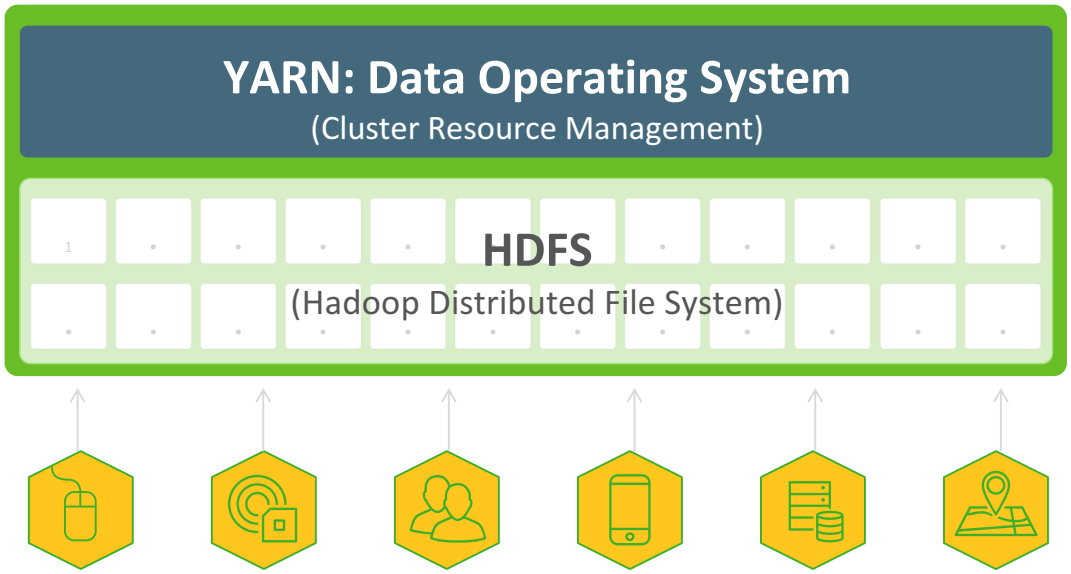
# Core Hadoop: HDFS + YARN

Distributed Computing of Storage, CPU and Memory

**Centralized Architecture**

**Highly Scalable**

**Cost Effective**







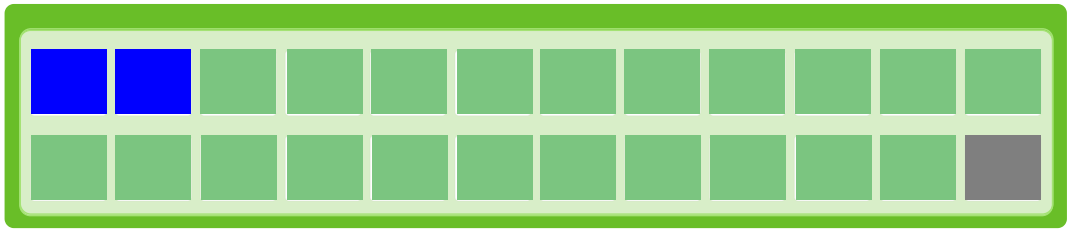
HORTONWORKS

# HDFS: Name Nodes & Data Nodes

Data and its Namespace

## Name Nodes

- Data Catalog, a “Namespace”
- Primary & Secondary Nodes for HA
- Manual or Automated Failover
- Quorum Journaling (States & Edits)



## Data Nodes

- “Where the data lives”
- Data stored in replicas of 3 (default)
- Rack aware
- Heterogeneous Storage Options
- Large block sizes (64MB default)

### Edge & Utility Nodes

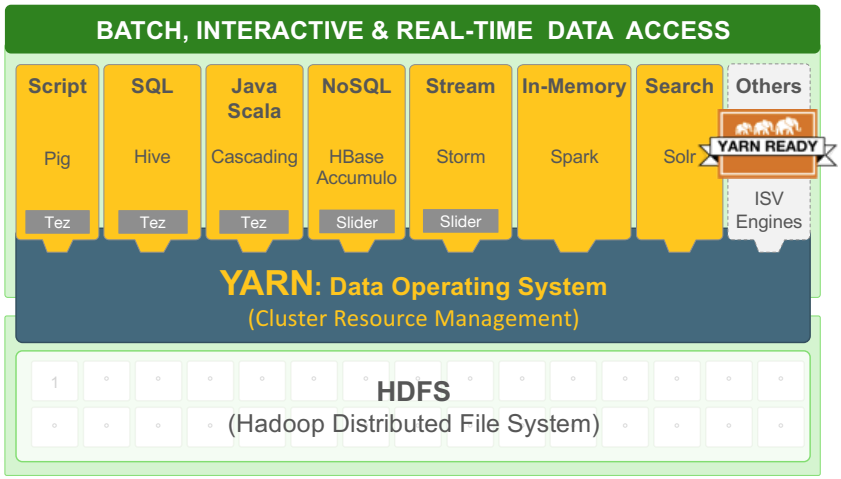
- Cluster Administration
- 3rd Party Tools




HORTONWORKS

# YARN: A Data Operating System

*Yet Another Resource Negotiator*




## YARN
### The Architectural Center of Hadoop

- Common data platform, many applications
- Support multi-tenant access & processing
- Batch, interactive & real-time use cases
- Supports 3rd-party ISV tools

  (ex. SAS, Syncsort, Actian, etc.)

## YARN Ready Applications

Facilitates ongoing innovation and enterprise adoption via ecosystem of new and existing "YARN Ready" solutions

## YARN in Production

- Yahoo: ~40,000 nodes, multiple clusters running YARN across over 365PB of data
- Spotify, Progressive, Kohls, UHG, Sprint, JPMC, Target, AIG, Samsung

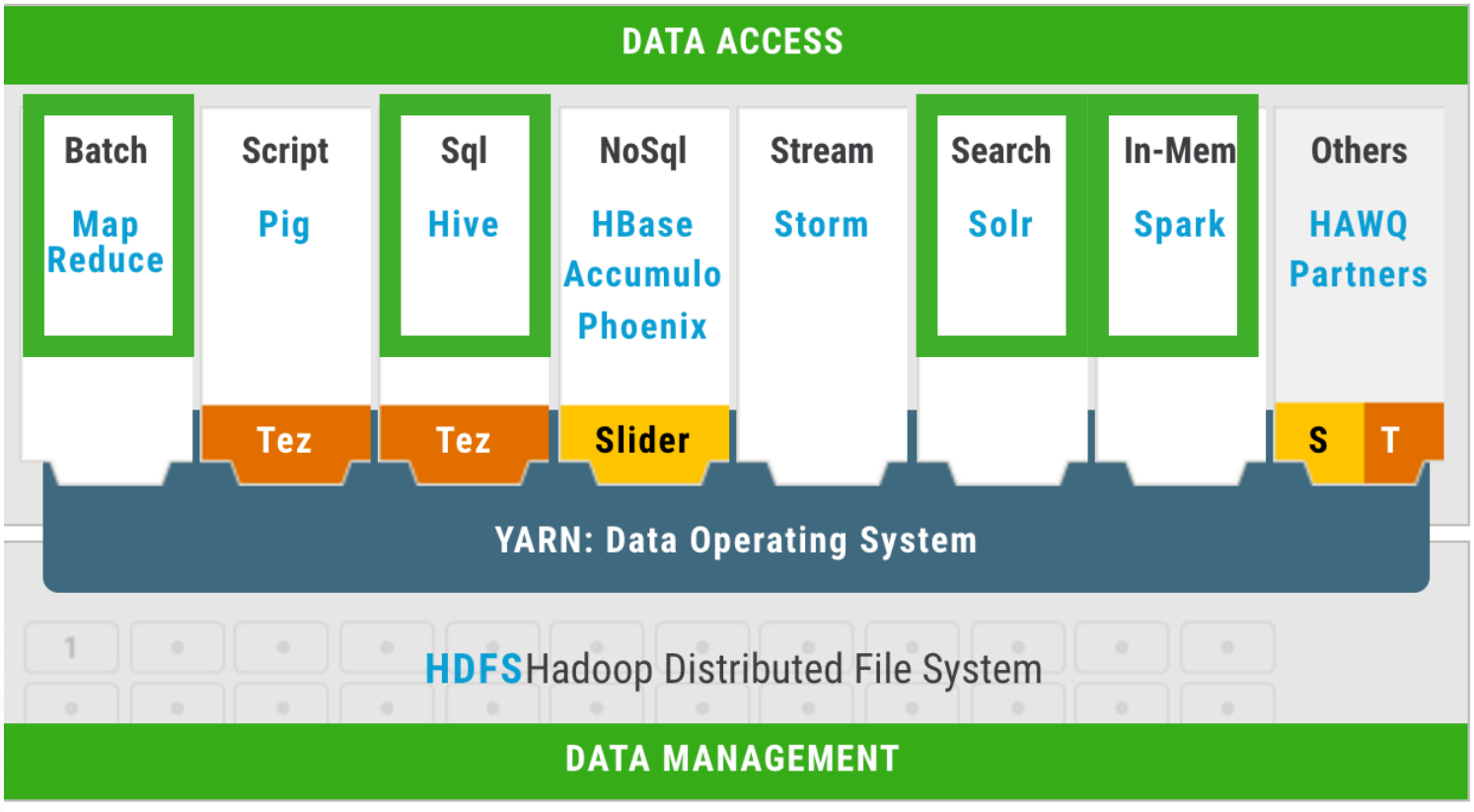



HORTONWORKS

# Data Access on Hadoop

## Hortonworks Data Platform

MapReduce, Hive, Spark, Solr

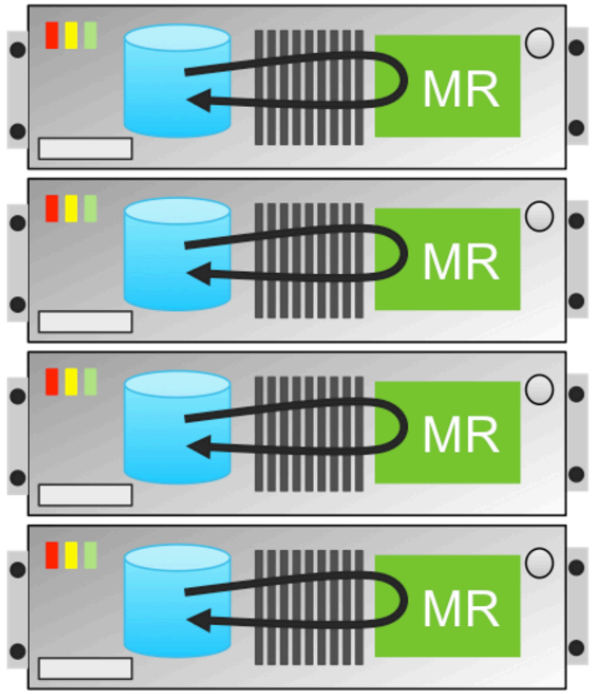



HORTONWORKS

# MapReduce

## Hadoop architecture is based on a cluster of machines:

- Scale-out, shared nothing architecture, distributed processing.
- Requires distributed processing programs.

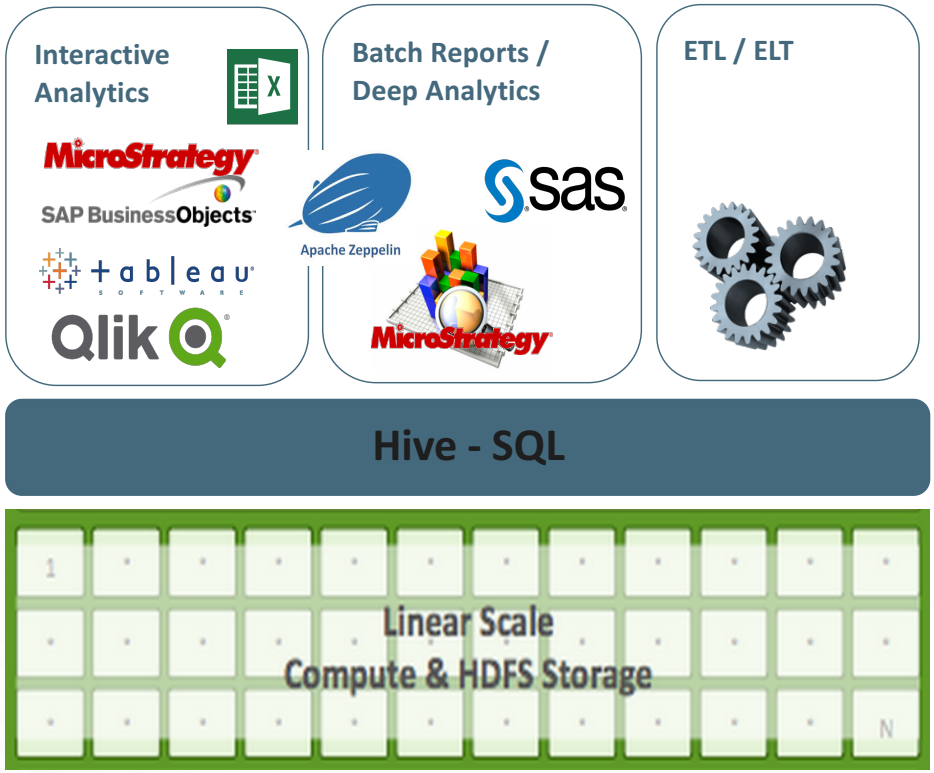


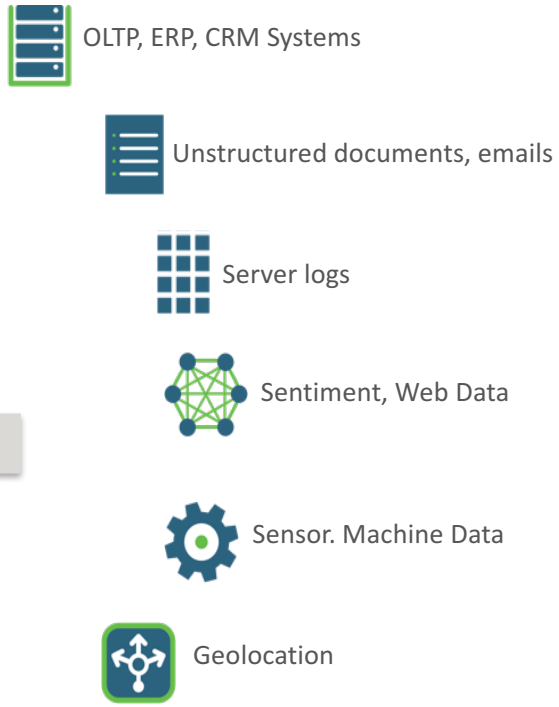

## MapReduce:

- MapReduce Framework offers APIs to write distributed processing programs:
  - 3 Tasks/APIs: ***Map, Shuffle, Reduce***
- Highly Scalable: Principle of Data Locality
- Best suited for batch processing.

HORTONWORKS

# Hive – Single tool for all SQL use cases

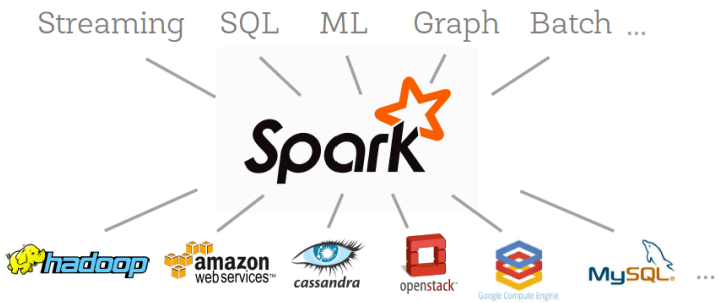


# Apache Spark

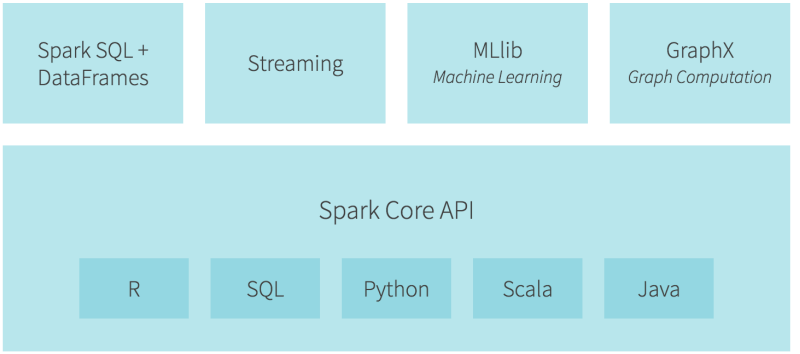
## What is Apache Spark ?

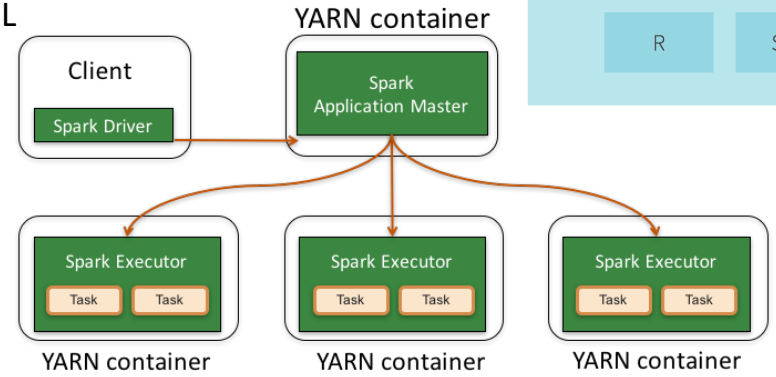
- **Apache open source project** originally developed at AMPLab (University of California Berkeley)
- **Unified data processing engine** that operates across varied data workloads and platforms




## Why Spark ?

- **Elegant Developer APIs**
  - Single environment for data munging and Machine Learning (ML)
- **In-memory computation model – Fast!**
  - Effective for iterative computations and ML
- **Machine Learning**
  - Implementation of distributed ML algorithms
- **YARN Ready**







HORTONWORKS

# Apache Solr

## What is Lucene?

- Apache Lucene is a high-performance, **full-featured text search engine library**
- **Provides API** to add search and indexing to your applications
- Provides scalable, High-Performance Indexing
- Efficient **Search Algorithms** (scoring, phrase, wildcard, sorting, geospatial, faceting….)
- When using Lucene you need to write code to do all of this

## What is Solr?

- **Search server built on top of Apache Lucene**
  - Provides API to access Lucene over HTTP
  - Ability to add more features on top of Lucene
- Majority of programming tasks in Lucene are XML configurations in Solr
- Provides **SolrCloud** (Scalability, Replication, Load Balancing)
- Provides **features** like Faceting, Clustering, Data Import Handlers, Multiple Language support, Rich document support

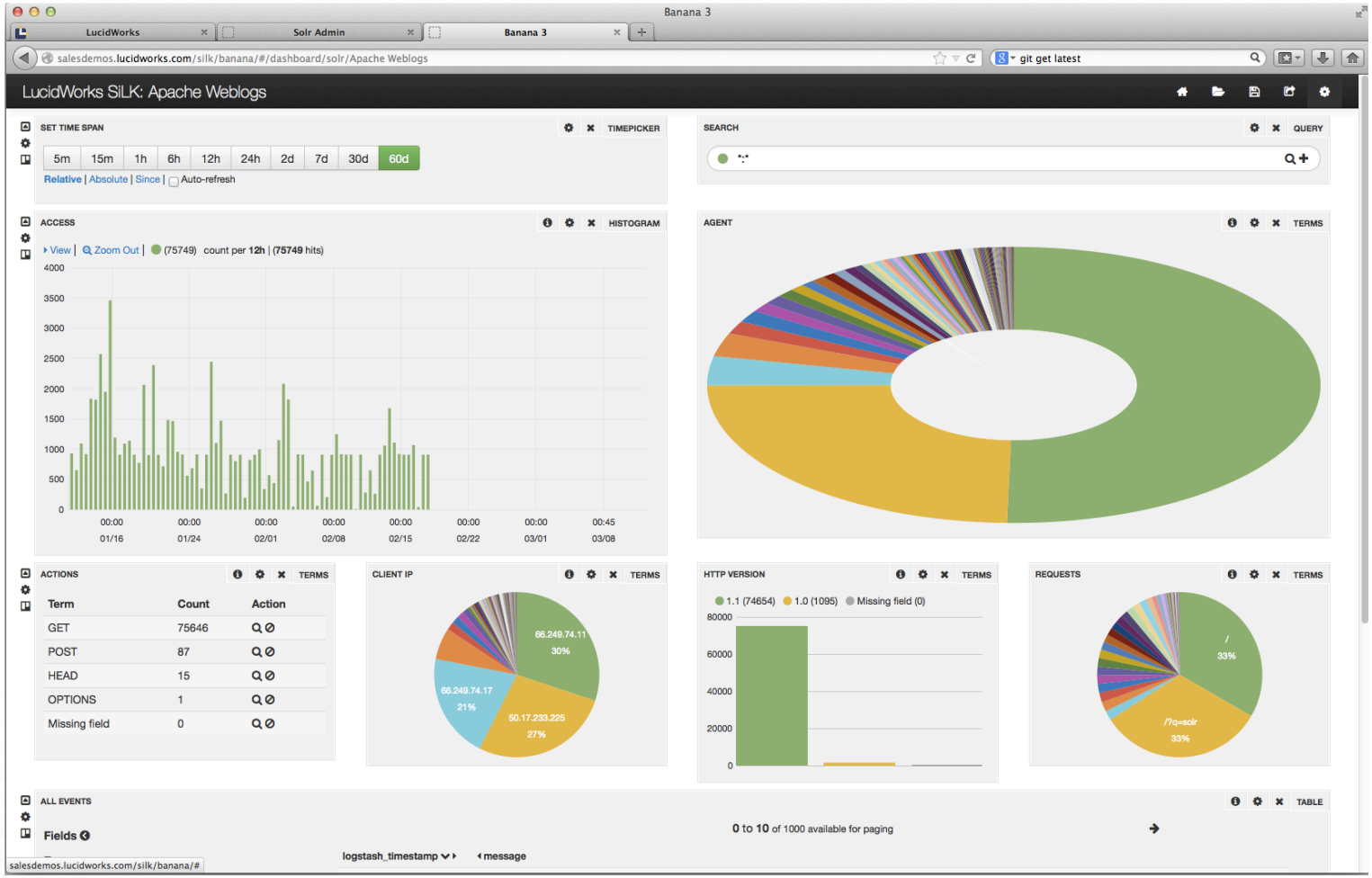

## Key Features

- **Advanced Full-Text Search Capabilities**
- **Optimized for High Volume Traffic**
- **Standards based open interfaces: XML, JSON, HTTP**
- **Comprehensive Administration interface**
- **Near Real-Time, Batch and Real-Time Indexing capabilities**
- **Linearly scalable, highly available and extensible plugin architecture**
- **Deep Hadoop integration**




HORTONWORKS

# Banana: Visualization and Reporting





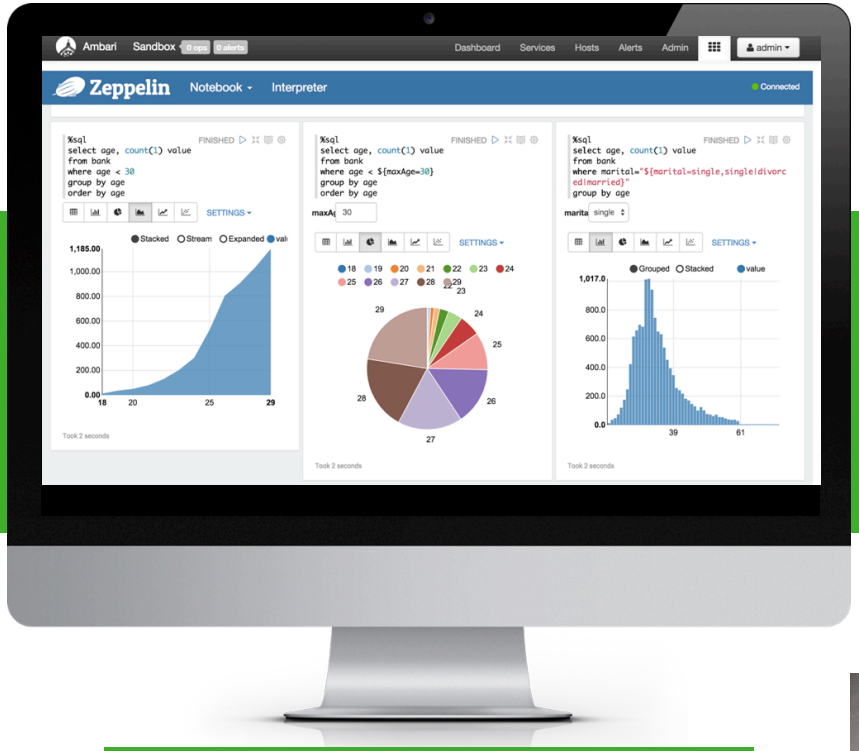


HORTONWORKS

# Apache Zeppelin GA: The Data Science Notebook

Spark SQL Scala python HIVE

Apache Zeppelin




## Features

- Ad-hoc experimentation
- Deeply integrated with Spark + Hadoop
- Interactive data ingestion, data exploration, visualization, sharing and collaboration

## Use Cases

- Data exploration and discovery
- Visualization
- Interactive snippet-at-a-time experience
- “Modern Data Science Studio”

Data engineers, data analysts and data scientists

HORTONWORKS

# Security, Governance & Operations

## Hortonworks Data Platform

# HDP Security: Comprehensive, Complete, Extensible

**Administration**

Central management and consistent security

Single administrative console to set policy across the entire cluster: **Apache Ranger**

**Authentication**

Authenticate users and systems

Authentication for perimeter and cluster; integrates with existing Active Directory and LDAP solutions: **Kerberos | Apache Knox**

**Authorization**

Provision access to data

Consistent authorization controls across all Apache components within HDP: **Apache Ranger**

**Audit**

Maintain a record of data access

Record of data access events across all components that is consistent and accessible: **Apache Ranger**

**Data Protection**

Protect data at rest and in motion

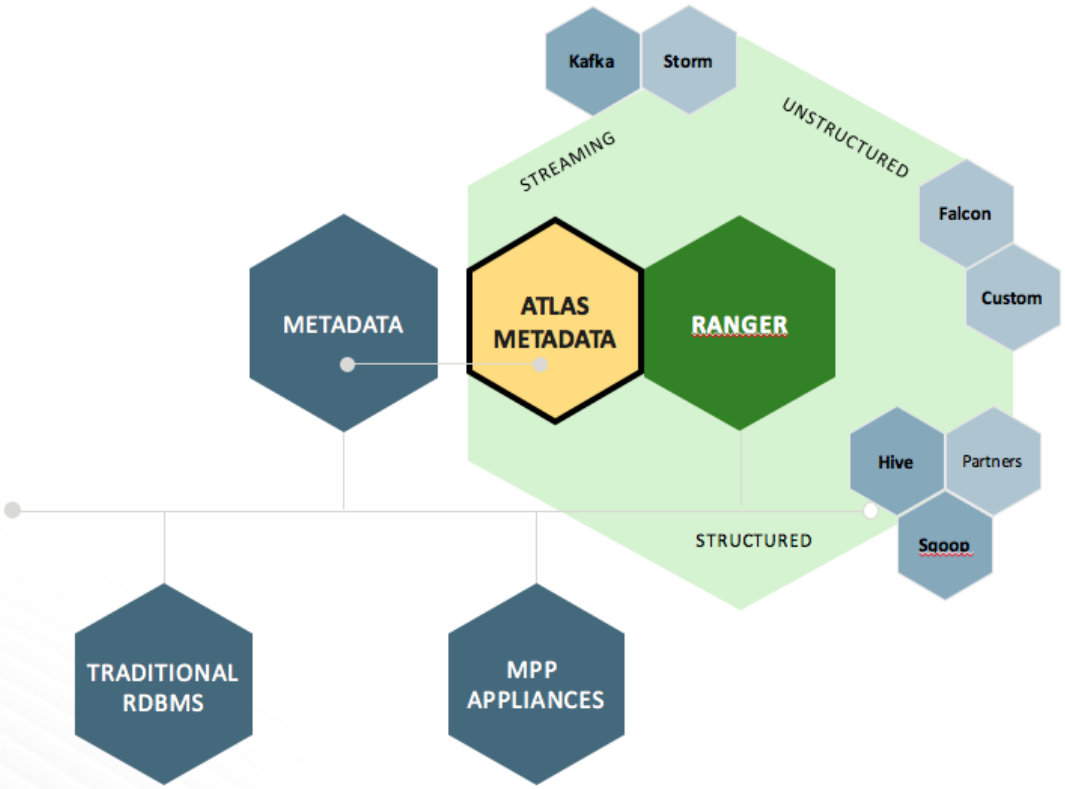
Secure data in motion and data at rest: **HDFS TDE w/ Ranger KMS + HSM, Ranger Data Masking + Row Filtering, Wire encryption + Partner Solutions**




HORTONWORKS

# Apache Atlas

Atlas is a scalable and extensible set of core foundational governance services – enabling enterprises to effectively and efficiently meet their compliance requirements within Hadoop and allows integration with the whole enterprise data ecosystem.

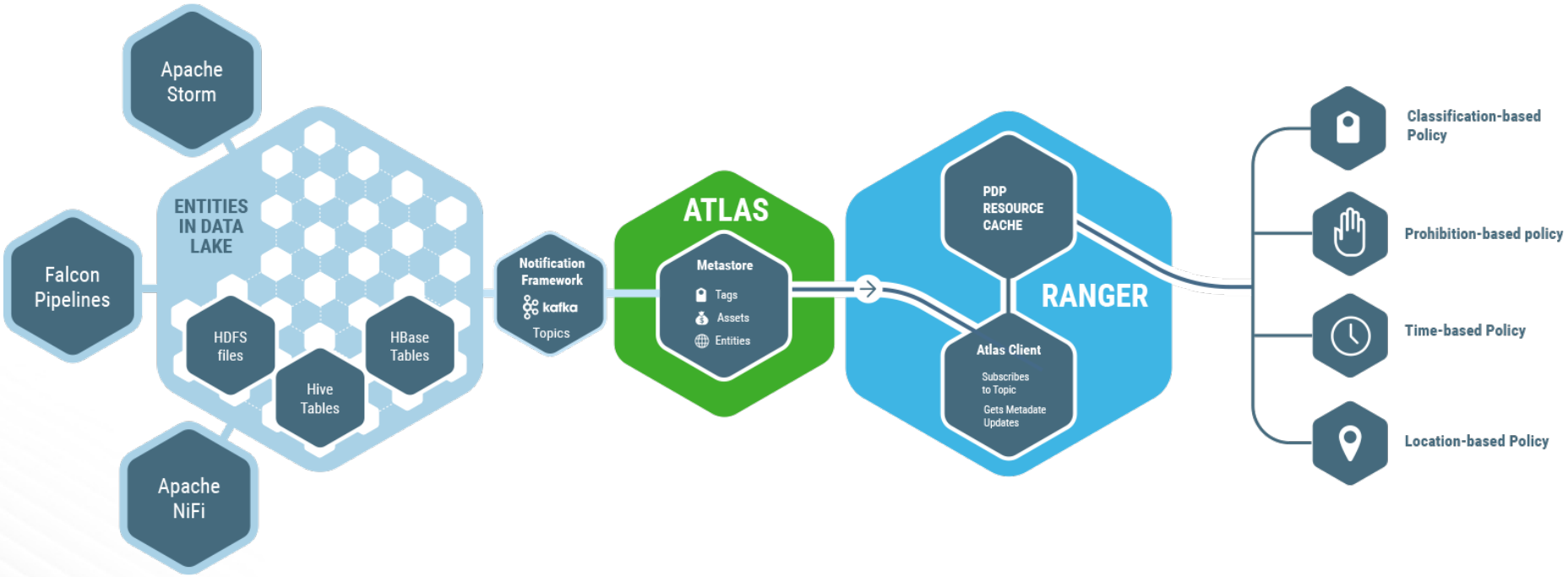



## Key features

- Data Classification (taxonomy)
- Security and policies engine
- Data Lineage and Search over metadata
- Centralized Auditing
- Metadata Exchange

HORTONWORKS

# Atlas + Ranger: Metadata driven security model



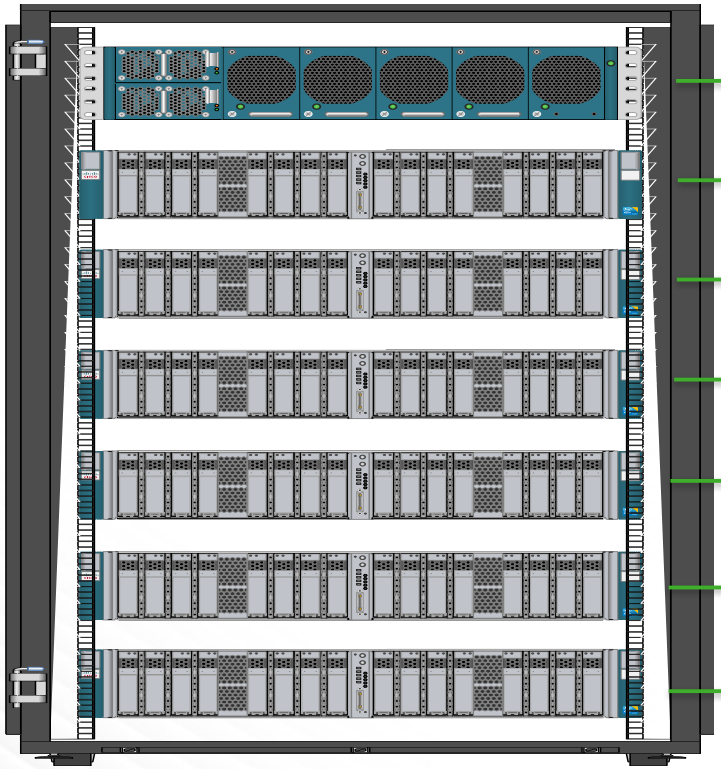




HORTONWORKS

# Typical Hadoop Rack Topology

**Hadoop is deployed as a "Rack Aware" platform so that the data is securely replicated across multiple racks and the Hadoop services are also deployed across multiple racks, preventing cluster outages and job failures due to asingle or multiple rack failures**



**Top of Rack Switch:** The Top of Rack Switch is responsible for the management of physical network traffic on the Hadoop cluster

**Master Node:** The Master Node is responsible for the Management Oversight of all MapReduce and HBase processes

**Edge Node:** The gateway node is responsible for receiving all Job requests from outside of cluster and submitting for processing

**Management Node:** The management node is responsible for all processes related to management and operation of a Hadoop cluster

**Worker Node:** The worker node is responsible for data storage along with batch and real time data processing

**Worker Node:** The worker node is responsible for data storage along with batch and real time data processing

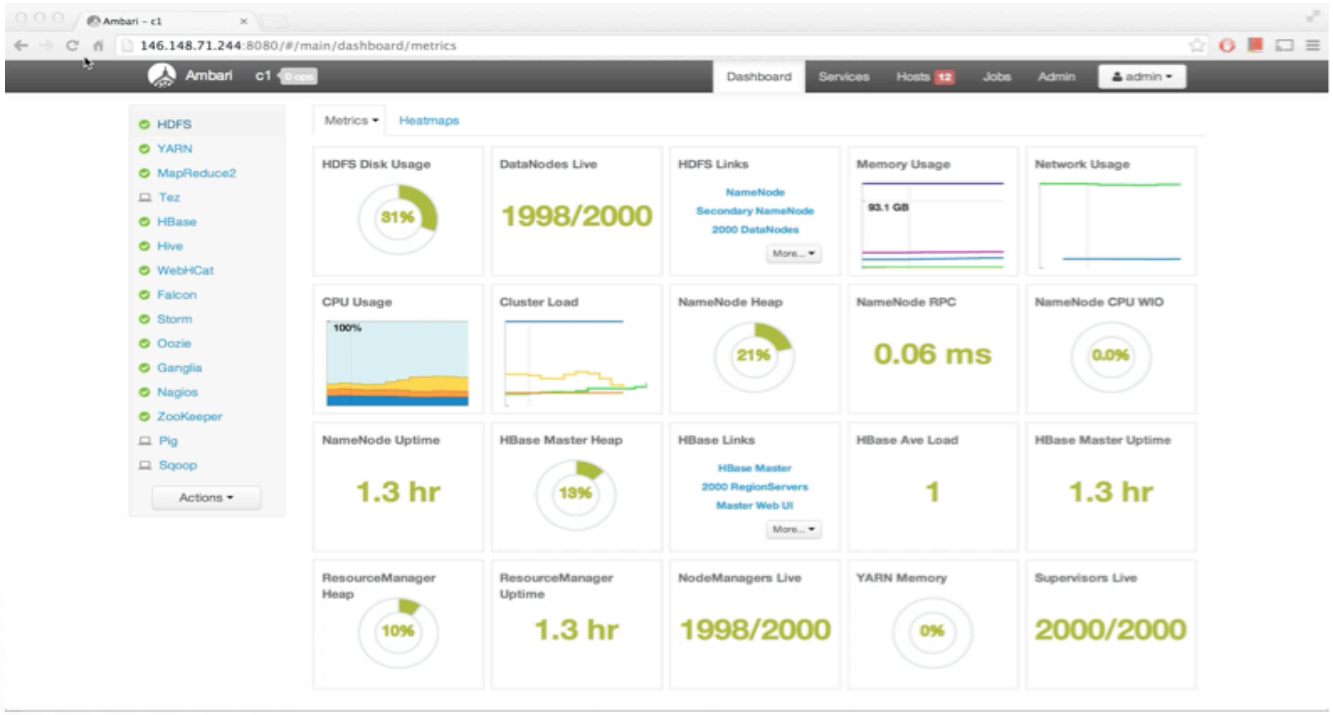
**Worker Node:** The worker node is responsible for data storage along with batch and real time data processing






HORTONWORKS

# Apache Ambari: Cluster Operations




*A completely open source management platform for provisioning, managing, monitoring and securing Apache Hadoop clusters. Apache Ambari takes the guesswork out of operating Hadoop.*

Simplified Installation, Configuration and Management

*Highly Extensible and Customizable*

*Full Visibility into Cluster Health*

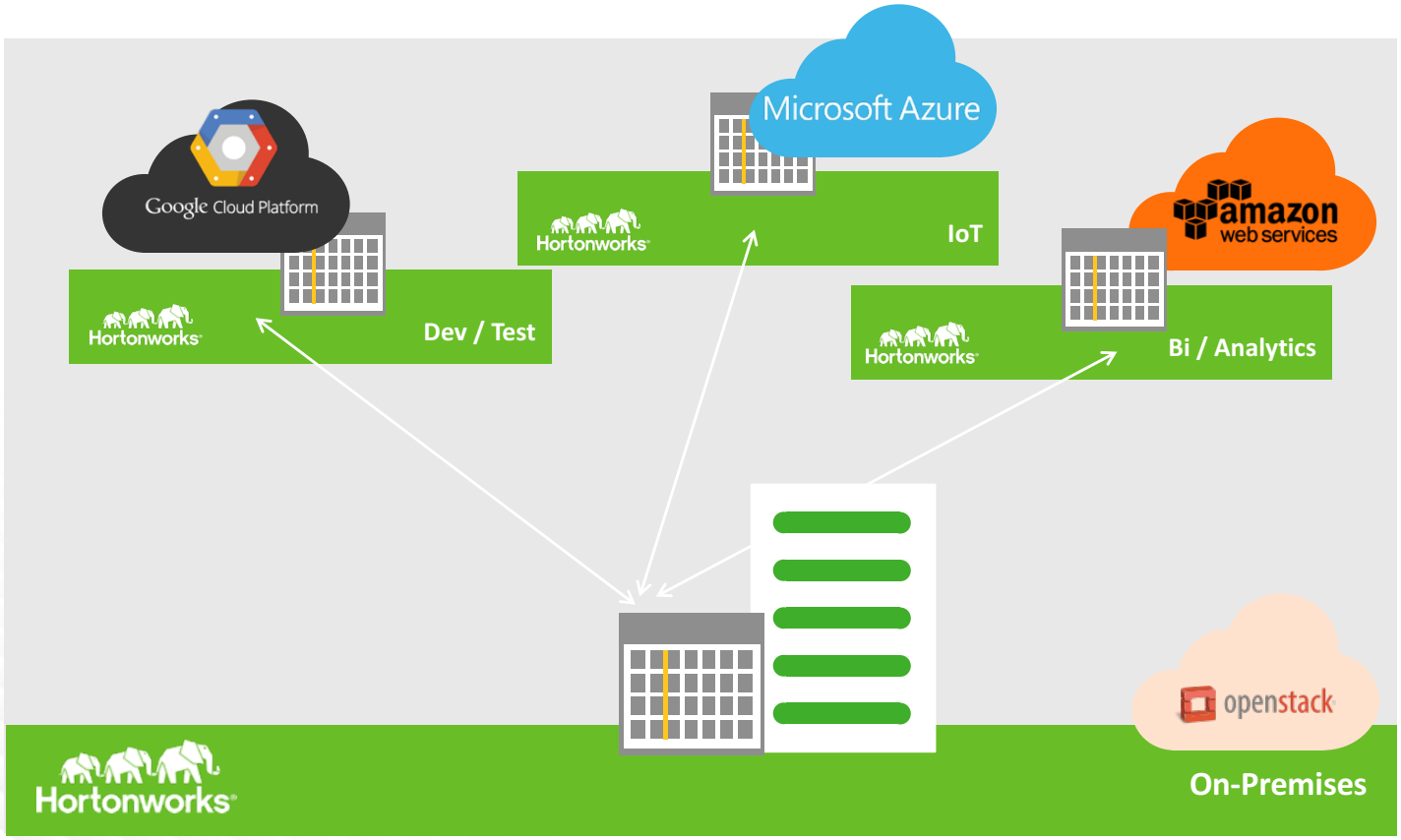
*Centralized Security Setup*




HORTONWORKS

# Cloudbreak: Cloud operations

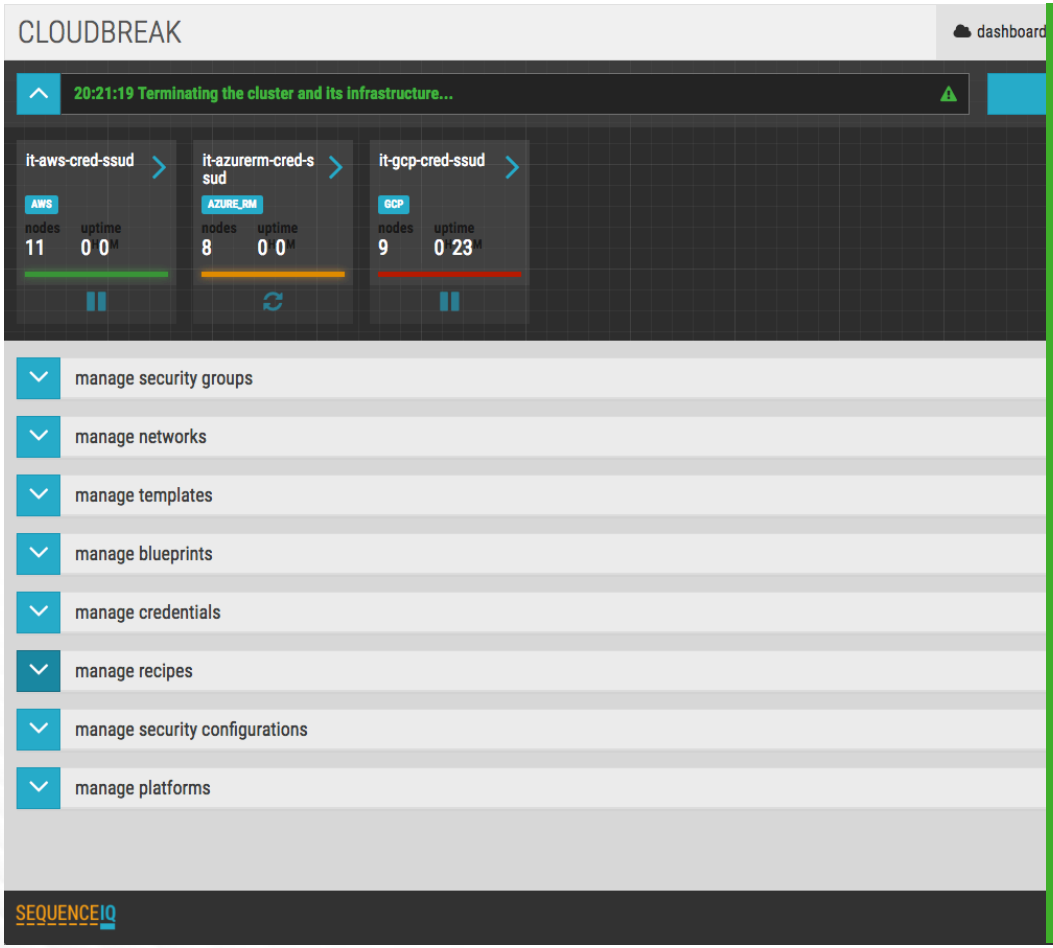
Easily Launch HDP on Any Cloud with Cloudbreak




Cloudbreak is a tool for provisioning HDP clusters on cloud infrastructure. Cloudbreak allows enterprises to simplify the provisioning of clusters in the cloud and optimize their use of cloud resources as workloads change.

HORTONWORKS

# Cloudbreak: Cloud operations




## Primary Use Cases

### Deploy on Public or Private Clouds

Dynamically configure and manage clusters on public or private clouds (Amazon Web Services, Microsoft Azure, Google Cloud Platform and OpenStack)

### Automated Scaling

Seamlessly manage elasticity requirements as cluster workloads change

### Secured Cluster Access

Supports configuration for Kerberos, defining network boundaries and configuring security groups




HORTONWORKS

# Wrap-up

## Hortonworks Data Platform

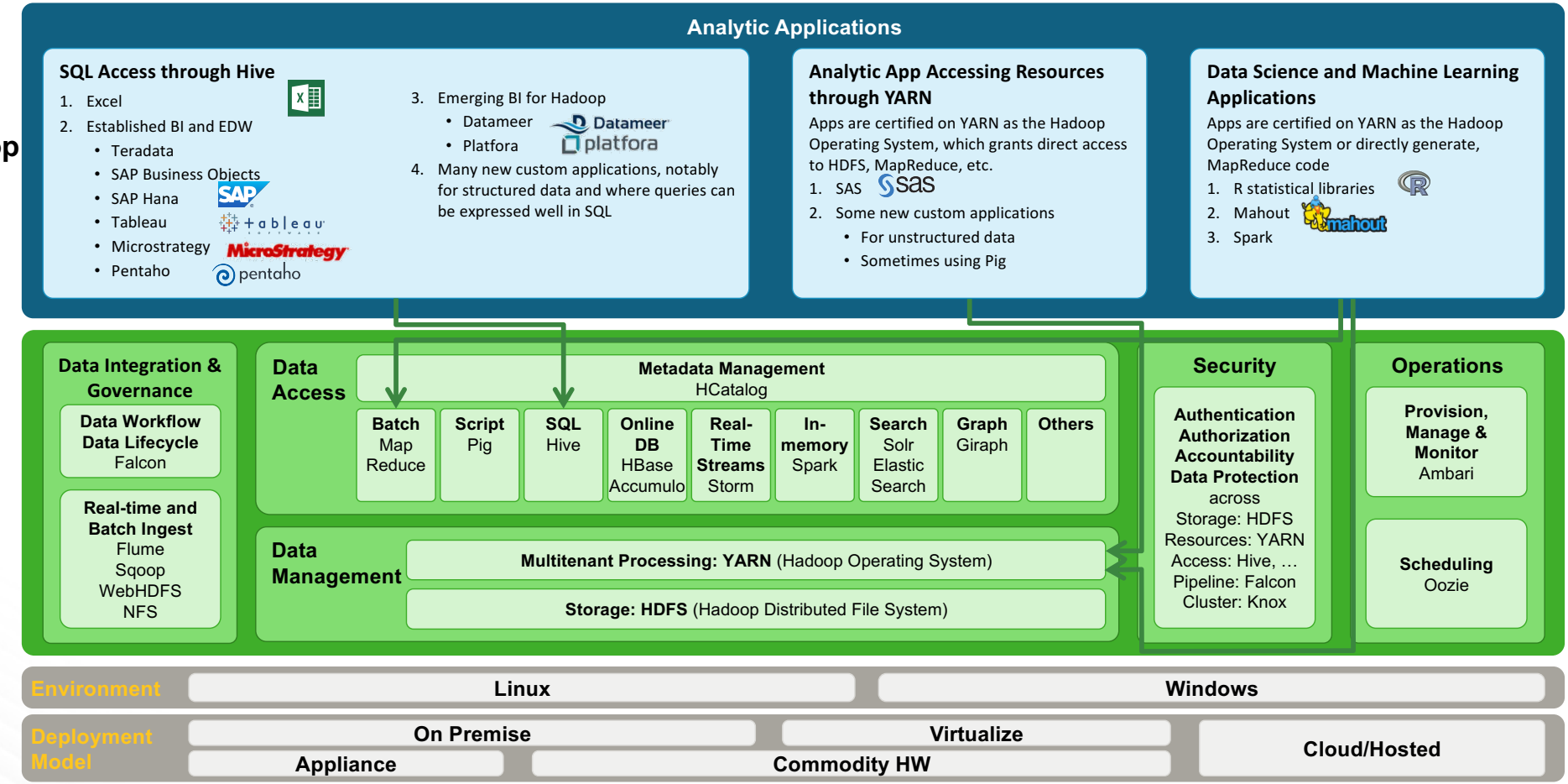
# The big picture: Ways of Consuming Hadoop

= Hadoop

**How is Hadoop consumed?**

**What is Hadoop?**

**Where does Hadoop run?**



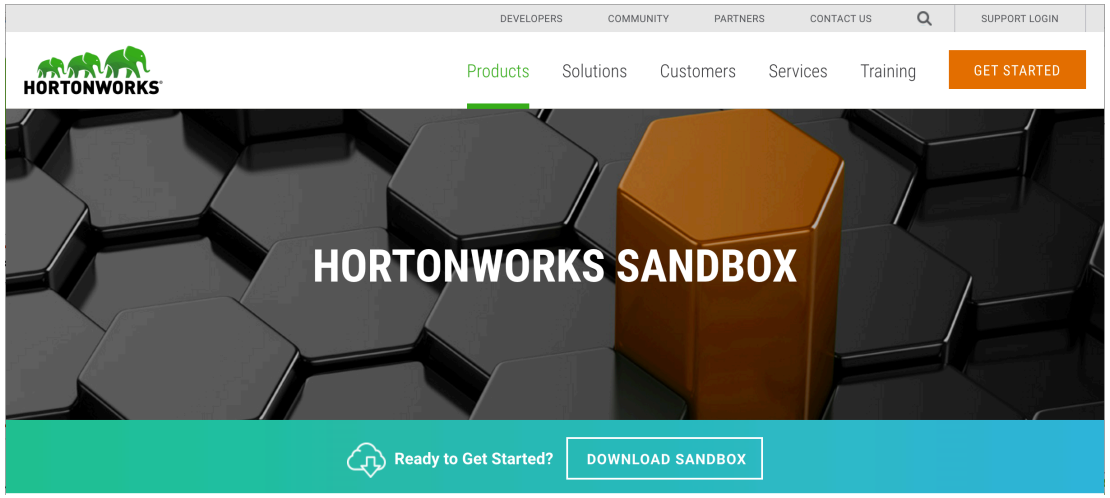


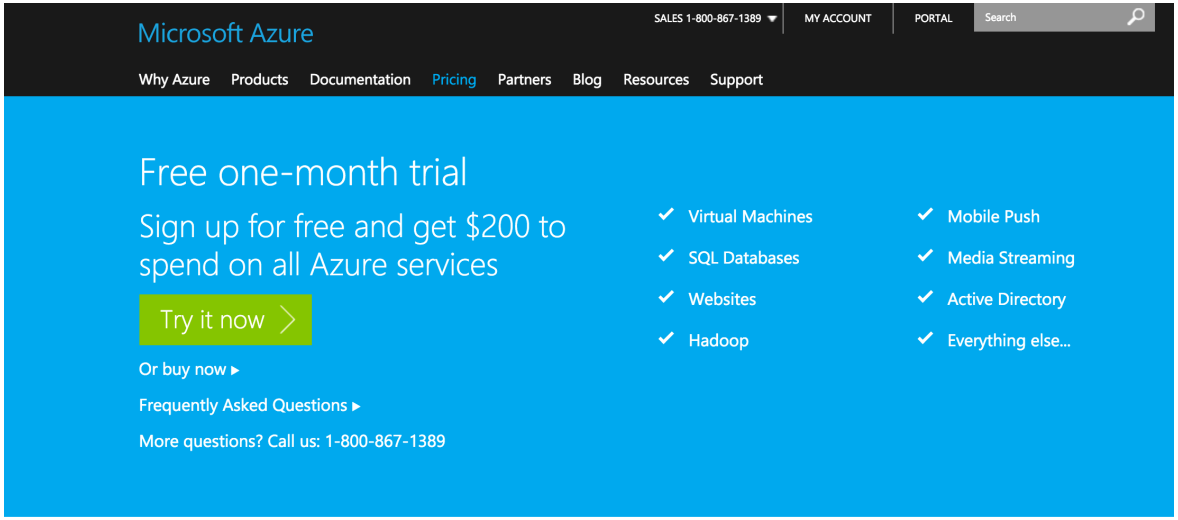


HORTONWORKS

# Next Steps: Try Hortonworks Data Platform

- Download HDP Sandbox:
  http://hortonworks.com/products/sandbox/
- Deploy the HDP Sandbox on Azure:
  http://hortonworks.com/hadoop-tutorial/deploying-hortonworks-sandbox-on-microsoft-azure/
- Learning the ropes:
  http://hortonworks.com/hadoop-tutorial/learning-the-ropes -of-the-hortonworks-sandbox



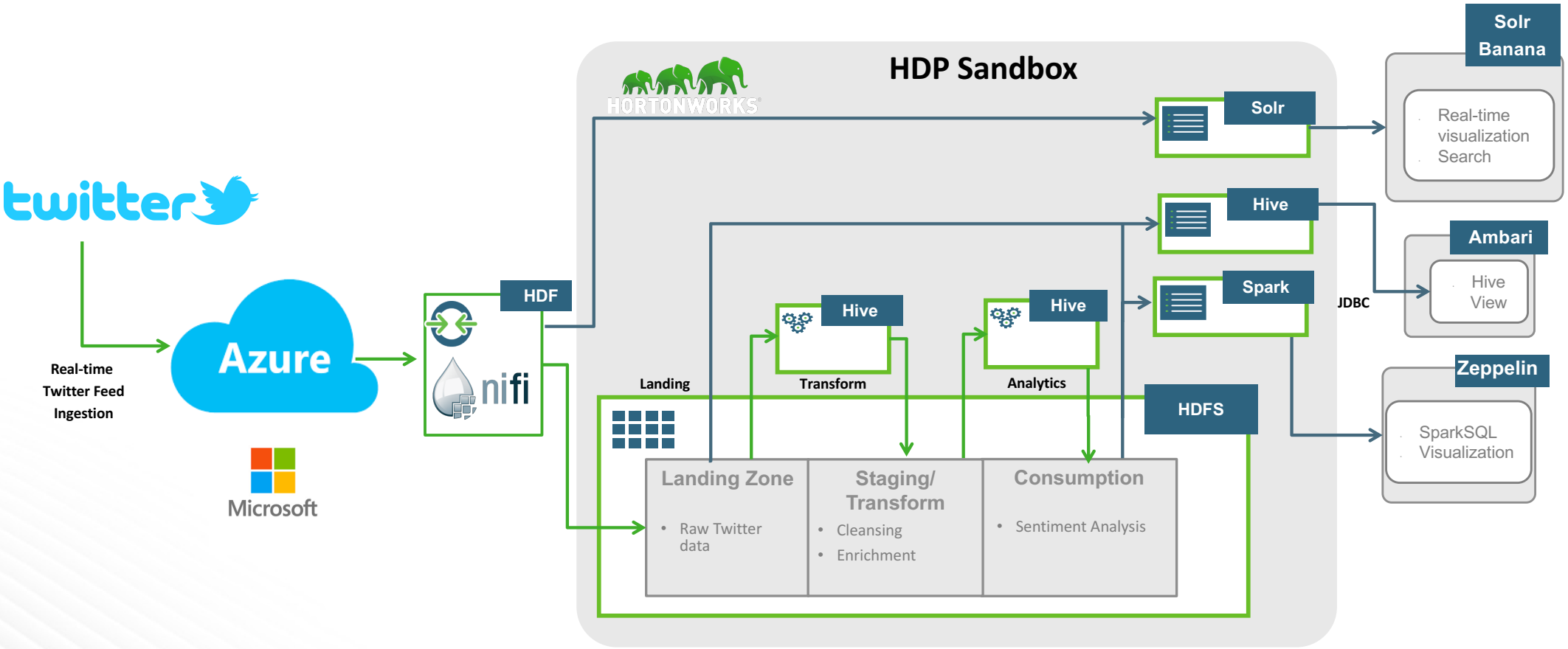

DATA WORKS SUMMIT
Presented by Hortonworks and Yahoo!
HADOOP SUMMIT
DATAWORKS SUMMIT AND HADOOP SUMMIT | SAN JOSE
JUNE 13–15 2017
CALL FOR ABSTRACTS IS NOW OPEN
https://dataworkssummit.com/san-jose-2017/abstracts/submit-abstract
CALL FOR ABSTRACTS CLOSES FEBRUARY 10
https://dataworkssummit.com/san-jose-2017/agenda/#tracks

# Demo of HDP and HDF

## Sentiment Analysis on Twitter

HORTONWORKS®
POWERING THE FUTURE OF DATA™

# Sentiment Analysis Demo Overview







HORTONWORKS


Hortonworks
Thank You !