Big Data Analytics on Amazon Web Services

Get Started with Big-Data-as-a-Service

bluedata™

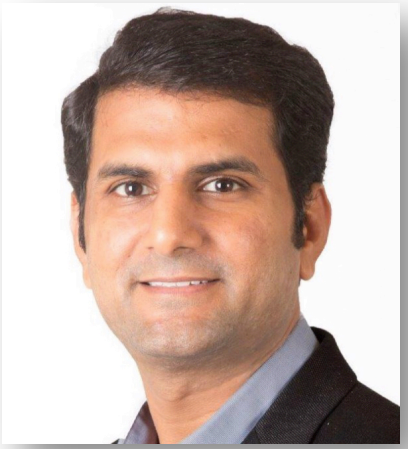
Today's Speakers

Tom Phelan

Co-Founder and Chief Architect

BlueData Software

@tapbluedata



Anant Chintamaneni

Vice President of Products

BlueData Software

@anantcman

bluedata

Agenda

- Big-Data-as-a-Service (BDaaS) Defined
- BDaaS on Amazon Web Services
- Challenges
- Solution Overview
- Customer Examples
- Key Takeaways
- Q & A

< Enter questions at any time in your BrightTalk web client

bluedata

Big-Data-as-a-Service Defined

BDaaS

On-Demand, Self-Service, Elastic
Big Data Infrastructure, Applications, Analytics

"BDaaS basically provides a cloud-based framework that offers end-to-end big data solutions to business organizations."

Source: http://www.marketsandmarkets.com/Market-Reports/big-data-as-a-service-market-4129107.html

On-Premises

Public Cloud

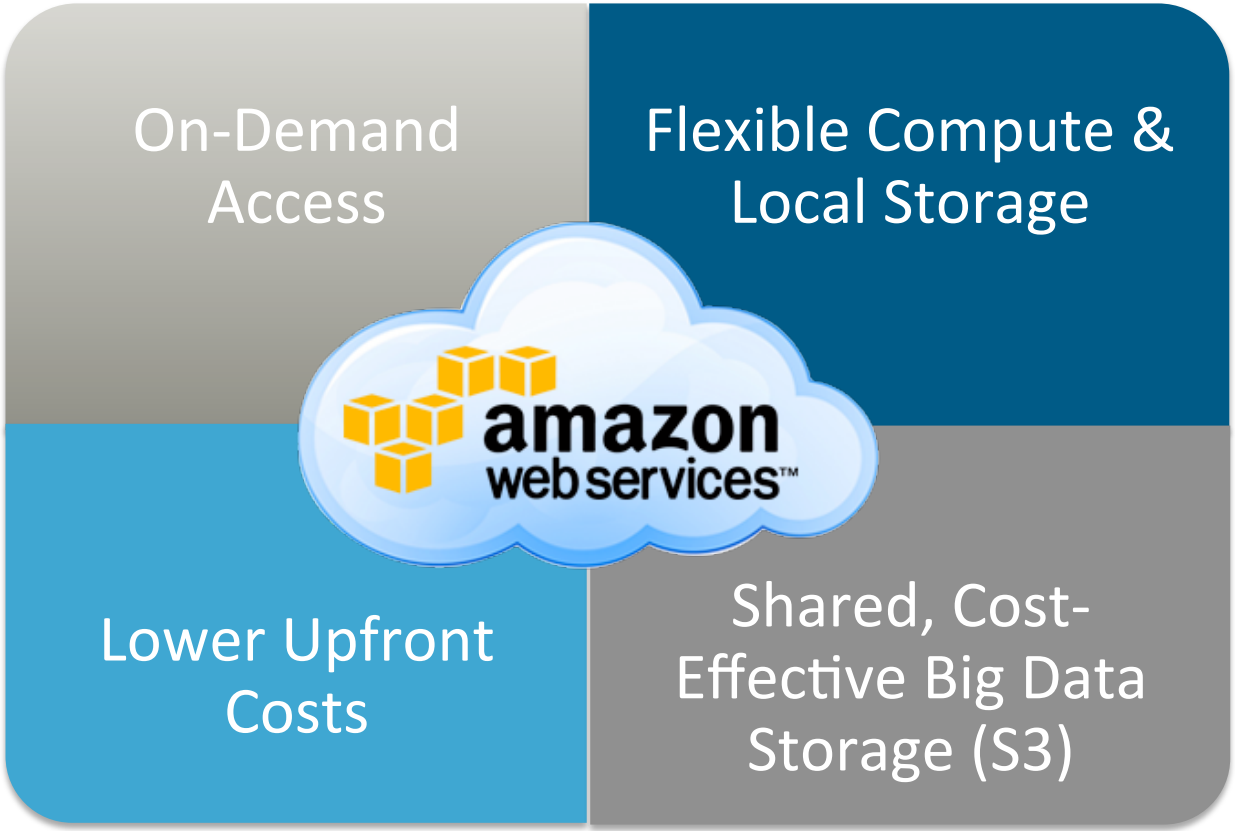
bluedata

Why BDaaS on a Public Cloud?

- Big Data infrastructure has become too complex
  - Few organizations have the skills & resources to manage Big Data infra
  - "Big Data" is no longer just Hadoop. It's a community of interoperating frameworks, each requiring its own cluster
- Increasing importance of agility, flexibility, and expandability
  - The rapid pace of evolution in the Big Data space renders traditional infra platforms obsolete before they can be fully implemented in the enterprise
  - IoT (machine data) is increasingly being stored in the pubic cloud
  - Data Science workflows demand flexibility and bursts in computing power
- Constant need to drive down IT costs

bluedata™

Why BDaaS on AWS?



bluedata™

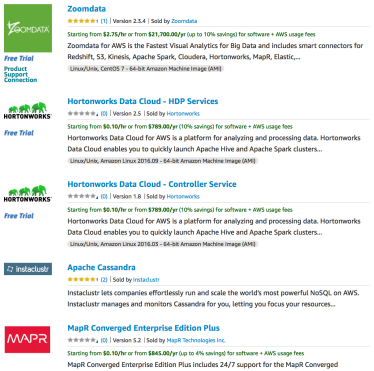
BDaaS Landscape on AWS

Amazon Services

Analytics

- Athena
- EMR
- CloudSearch
- Elasticsearch Service
- Kinesis
- Data Pipeline
- QuickSight

Big Data Products
(e.g. AWS Marketplace)



Managed SaaS

databricks

Qubole
The Big Data SaaS Company

CAZENA
BIG DATA AS A SERVICE

DIY

Simple use case

Product Choices

Product Evaluation

Silo'ed user experience

Support

Hadoop/Spark Only

Flexibility & Choice

Skills, UX, Lock-In

Gap: *Bring your own Big Data application with minimal DevOps overhead & intuitive UX*

bluedata

Challenges with BDaaS on AWS

- Flexibility
- Complexity
- Administration
- Data Gravity
- Security and Auditing
- Multi-Cloud

Big Data Software Flexibility

BI/Integration Tools

Analytics & Machine Learning

App & Workflow Development

Distributed Data Platforms

bluedata

Challenges with BDaaS on AWS

- Flexibility
- Complexity
- Administration
- Data Gravity
- Security and Auditing
- Multi-Cloud

Onboarding Complexity

- Migrating platforms to the Public Cloud (e.g. AWS) is not easy
  - Can all the security and monitoring rules put in place in the data center be maintained in the public cloud?
  - Where is my data?
- Starting in the Public Cloud (e.g. AWS) is not so easy either
  - Why do I need a CloudFormation template?
  - Why do I need to run SSH tunnels and FoxyProxy etc?
  - VP(C, E, N, etc.) what?

Deployment Complexity

- Deploying any specific distributed platform requires specialized skills
- 100's of steps for a "single" data platform. Spark/Hadoop example:
  - Select AMI, instance type, # of instances
  - Add storage, security group, launch instances
  - Set up keys, connect to instance, configure network
  - Copy software, install dependencies, install software
  - Repeat for each instance OR use management software (Cloudera Manager, Ambari)
  - Setup KDC and enforce Kerberos-based authentication, control access to cloud storage
  - Automate using Chef/Puppet or AWS-specific tools
  - Build additional management UI for access control, tools to capture costs, etc.
- Repeat for other data platforms & tools required in your pipeline

Challenges with BDaaS on AWS

- Flexibility
- Complexity
- Administration
- Data Gravity
- Security and Auditing
- Multi-Cloud

bluedata™

Administration

- More than powering on AWS EC2 instances
- More than installing the software bits
- Need to configure the right services on the right instances so the whole cluster works without further human intervention
- Need to manage users, user access, and data sources
- Need to control costs and cost assignment in AWS
- And more …

Software Administration

Provide access to management consoles such as Cloudera Manager, Ambari, MapR Control Center, OR Resource Manager, YARN Job History Server

Spark Master UI ...

Standard Monitoring Tool (e.g. Ops Center)

Gap: *Need to enable easy access to all management consoles from a centralized UI*

bluedata

System Administration

Building a cluster for one or more data platforms is just the start …

User access
(AWS UI or purpose-built UI)

Multi-tenancy
(teams, environments)

Data sources
(S3 storage access)

App/Data Integration
(existing apps, other data platforms)

End User Tools
(gateway/edge node configuration)

Monitoring & Mgmt
(services, log collection)

Elasticity
(scaling clusters up/down)

Cost controls
(resource policies, stop instances)

Cost visibility
(by team, cluster, month, day)

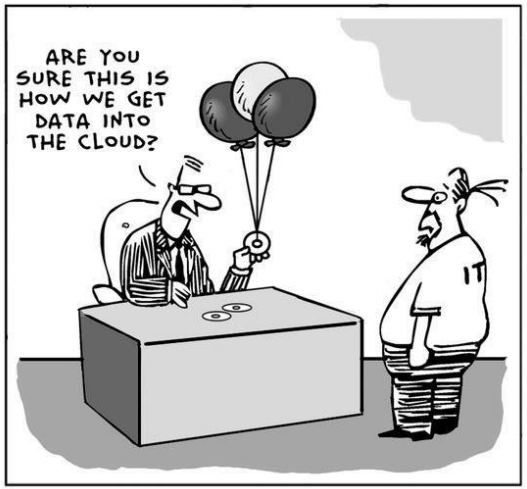
bluedata

Challenges with BDaaS on AWS

- Flexibility
- Complexity
- Administration
- Data Gravity
- Security and Auditing
- Multi-Cloud

Where Is Your Data?

- You can't run your Big Data jobs without the data
- On-premises storage
  - Requires migration/ingestion into a local HDFS file system
- In Amazon S3
  - Requires access setup
- In a database (e.g. RDS)
  - Requires network connectivity
- Somewhere else
  - ?



bluedata

Challenges with BDaaS on AWS

- Flexibility
- Complexity
- Administration
- Data Gravity
- Security and Auditing
- Multi-Cloud

bluedata™

Security and Auditing

- LDAP/AD
  - Authorization and authentication
- Kerberos
  - Secure HDFS file-level data access
- Amazon S3/Public Object Store
  - Authorization and ACLs (requires cloud-specific expertise)
- Knox, Ranger
  - Secure column level data access, data access auditing
- Logging

Challenges with BDaaS on AWS

- Flexibility
- Complexity
- Administration
- Data Gravity
- Security and Auditing
- Multi-Cloud

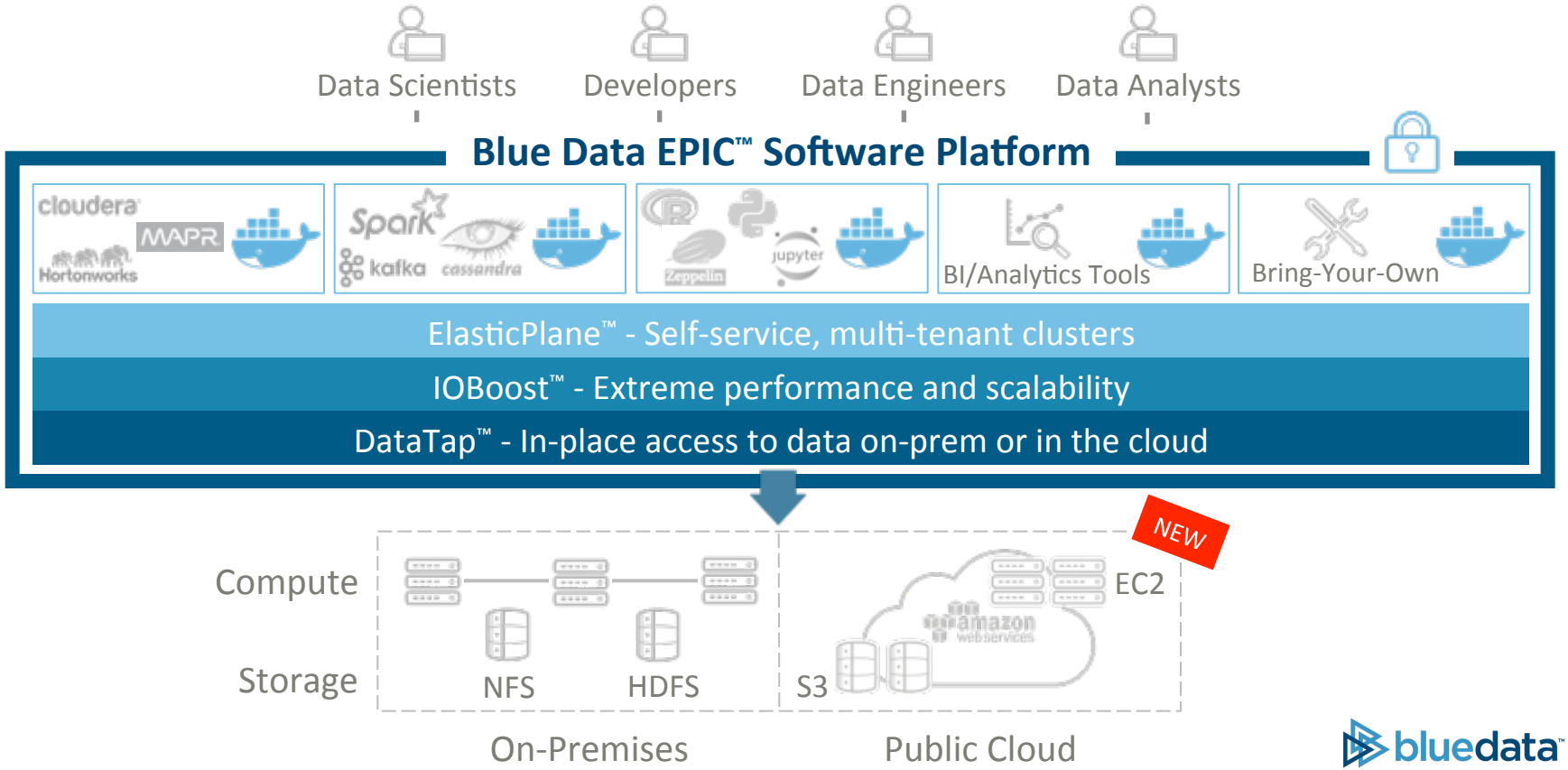
bluedata™

AWS + 1 and Multi-Cloud Strategies

Google Cloud Platform

bluedata

BlueData EPIC: On-Prem or on AWS



bluedata

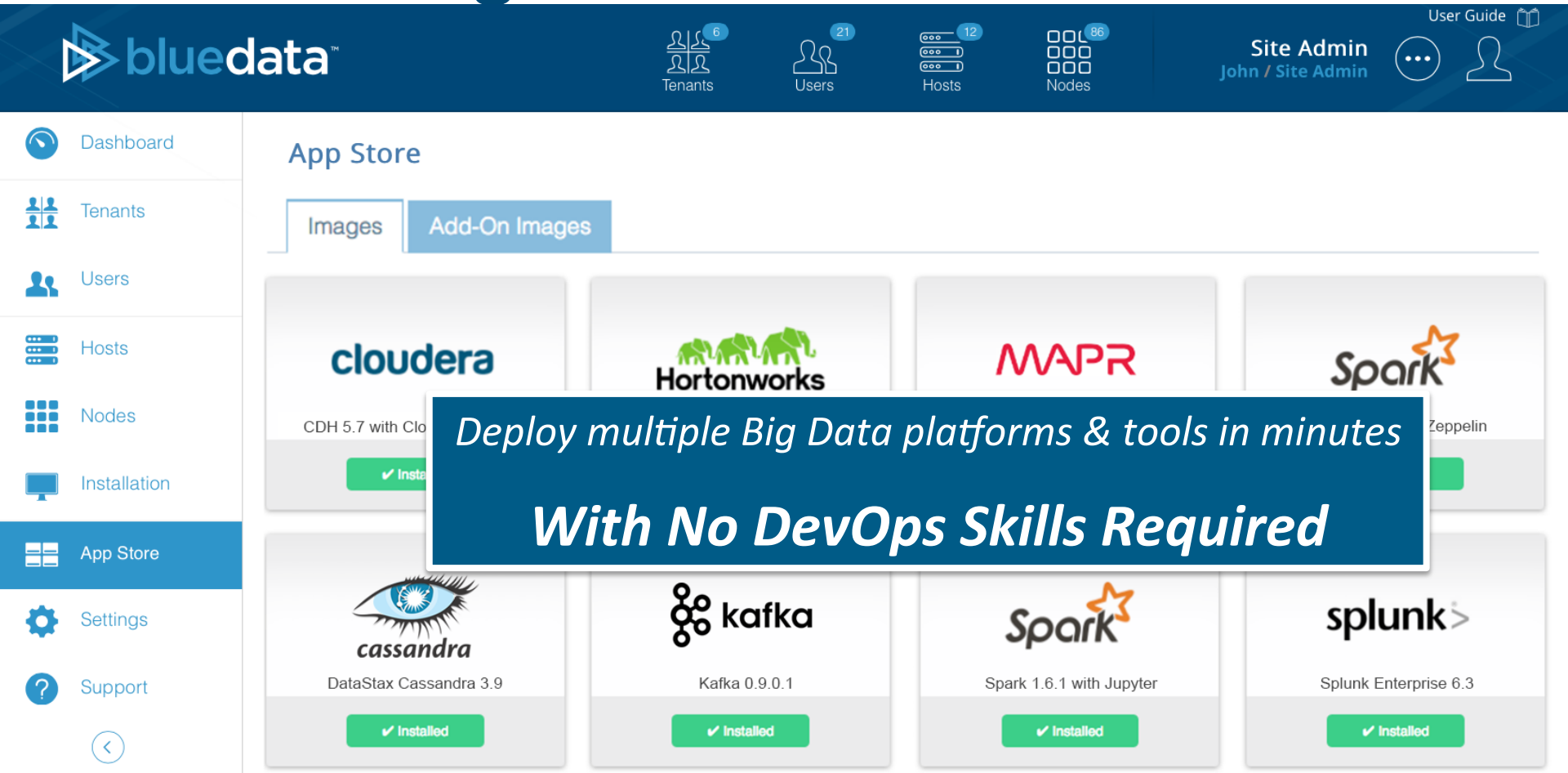
Introducing BlueData EPIC on AWS

BDaaS software platform deployed on a single Amazon EC2 instance that delivers …

on

amazon web services

Introducing BlueData EPIC on AWS

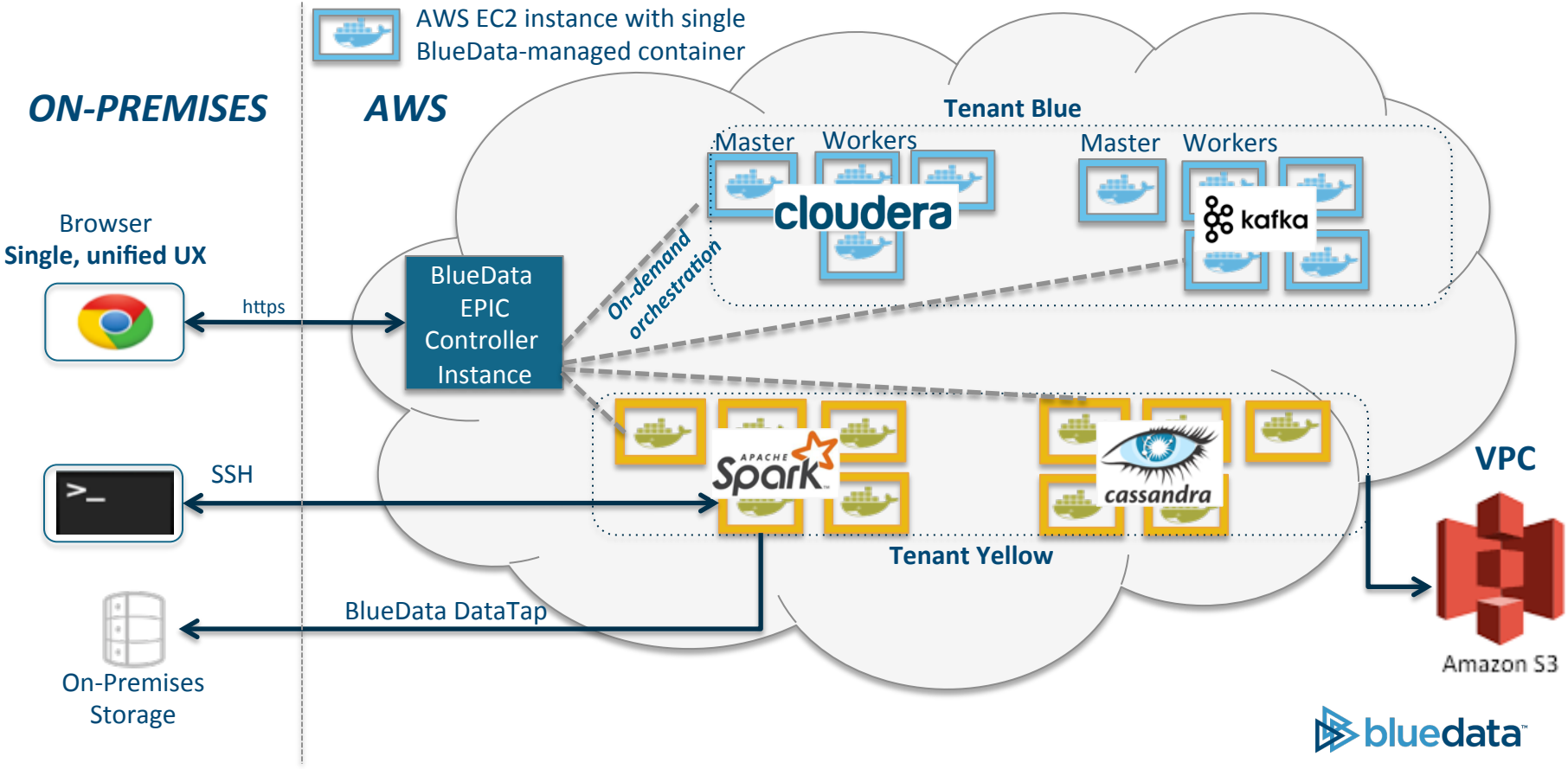


Introducing BlueData EPIC on AWS

- **Simplified UI** abstracts users from AWS-specific details
- **Reduces AWS onboarding** / development (DevOps) costs
- **Flexibility and choice** for Big Data apps, distributions, and tools
- **Reduces AWS costs** with fine-grained resource quotas, start/stop controls, and cost reporting integrations (via tags)
- **Faster time to insight** with in-place analytics against on-premises data via DataTap
- **Improves data governance** and lowers risk with pre-built integrations to S3 and VPC (incl. site-to-site VPN)

bluedata™

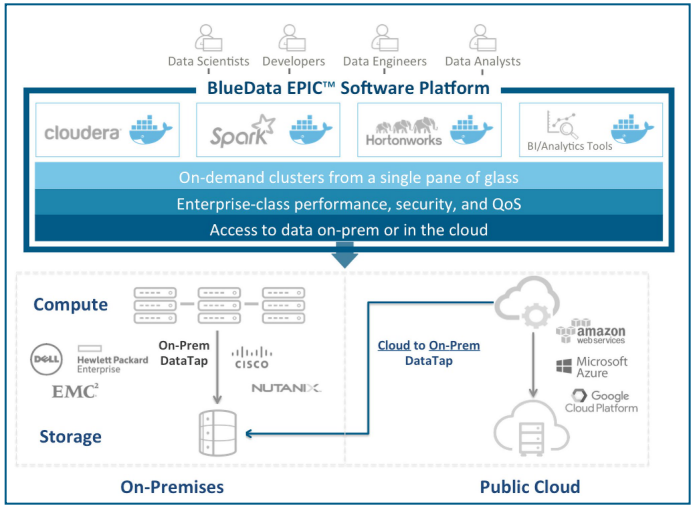
BlueData EPIC on AWS Architecture



BlueData for Big-Data-as-a-Service

Unified BDaaS software platform for Big Data whether on-prem or in the public cloud

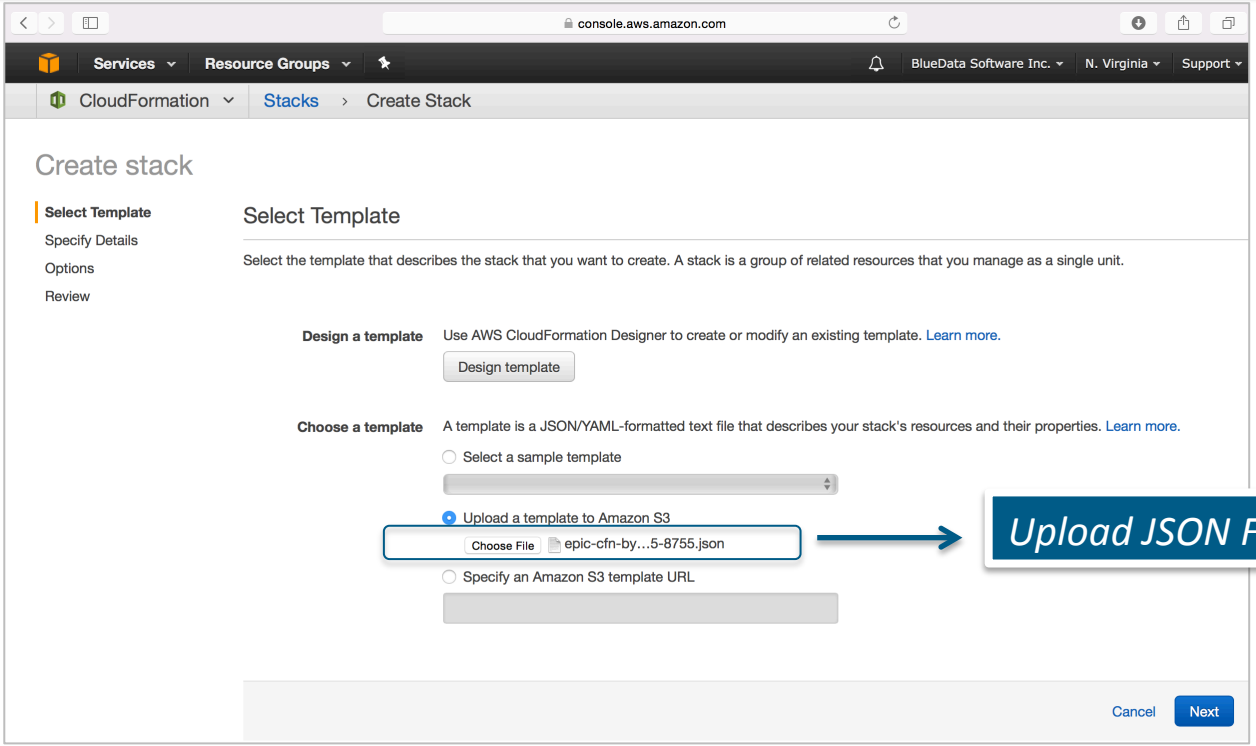
- Self-service creation of Big Data clusters running on Docker containers
- Ultimate flexibility and choice for Big Data frameworks, distributions, applications
- Tap into both on-prem and cloud storage
- Any server, any storage, any public cloud
- Enterprise-class security and control



bluedata™

Deploy BlueData EPIC Controller in AWS

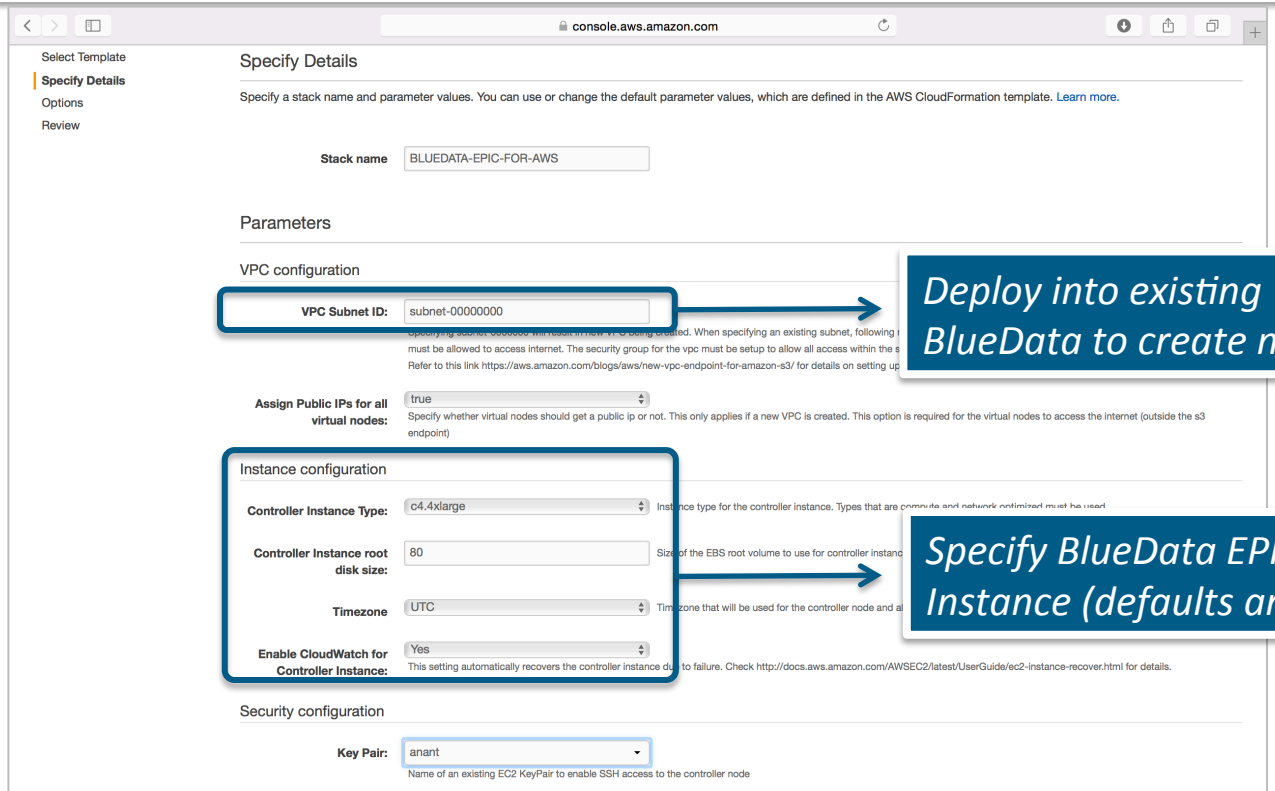
Obtain BlueData EPIC CloudFormation JSON File & Follow Wizard ➔ 20 minutes



bluedata™

Deploy BlueData EPIC Controller in AWS

Obtain BlueData EPIC CloudFormation JSON File & Follow Wizard ➔ 20 minutes



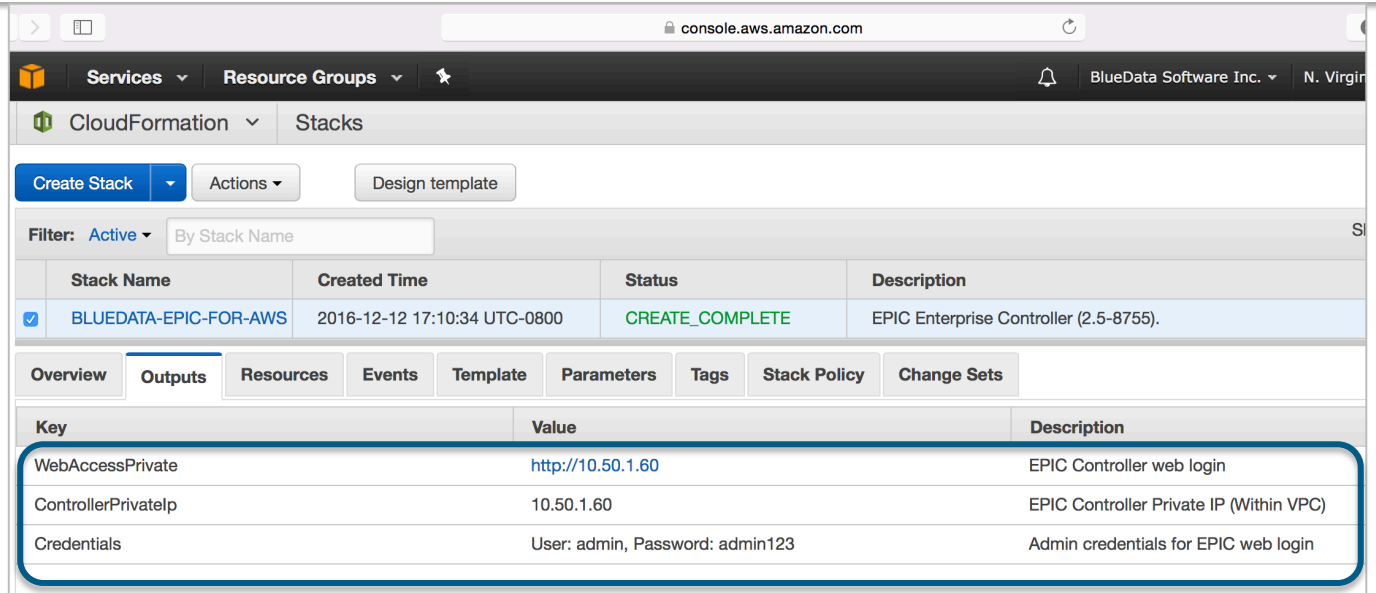
Deploy into existing VPC or allow BlueData to create new VPC

Specify BlueData EPIC Controller Instance (defaults are perfect!)

bluedata™

Deploy BlueData EPIC Controller in AWS

Obtain BlueData EPIC CloudFormation JSON File & Follow Wizard ➔ 20 minutes



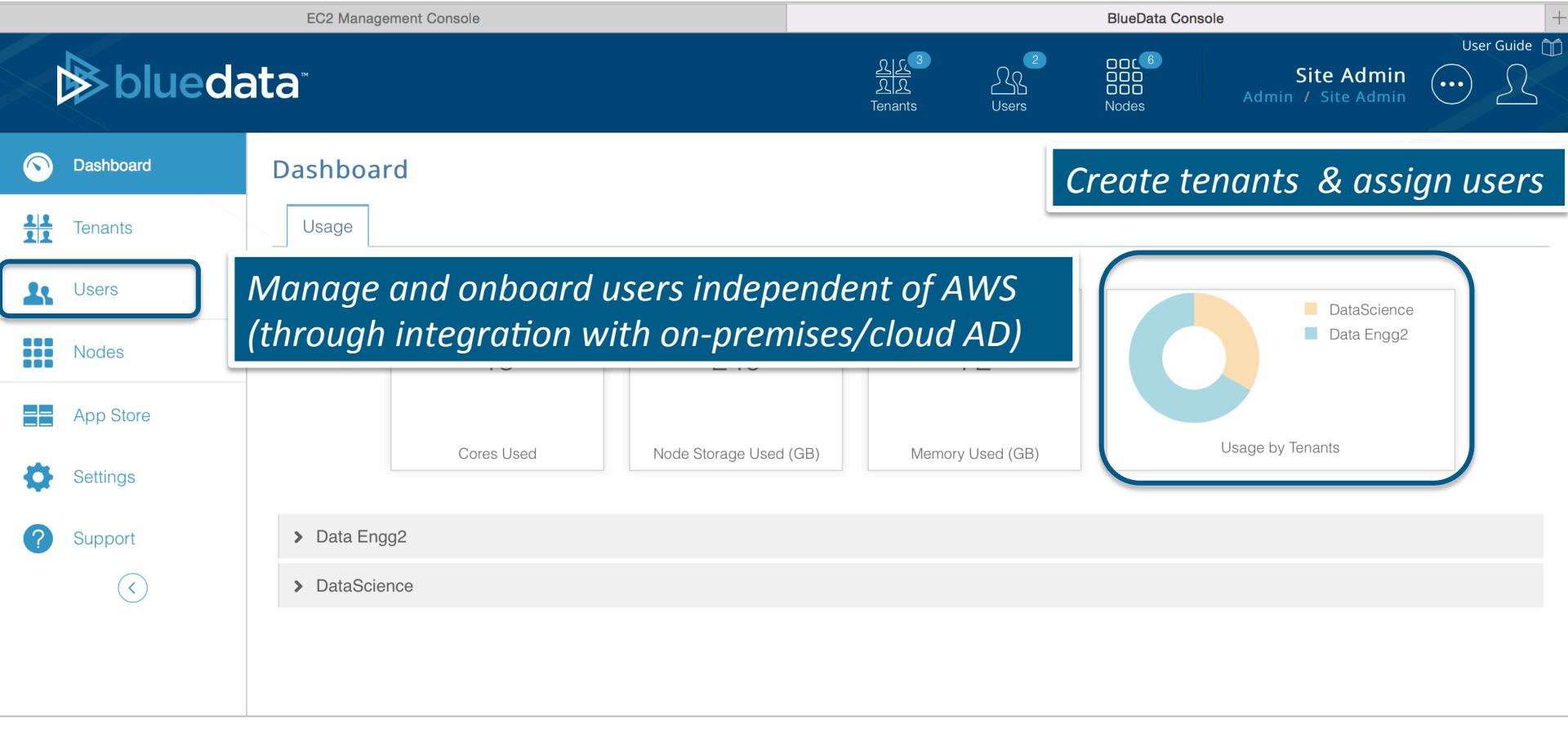
Ready-to-use multi-tenant BDaaS App

(Users never see AWS UI from here on …)

bluedata™

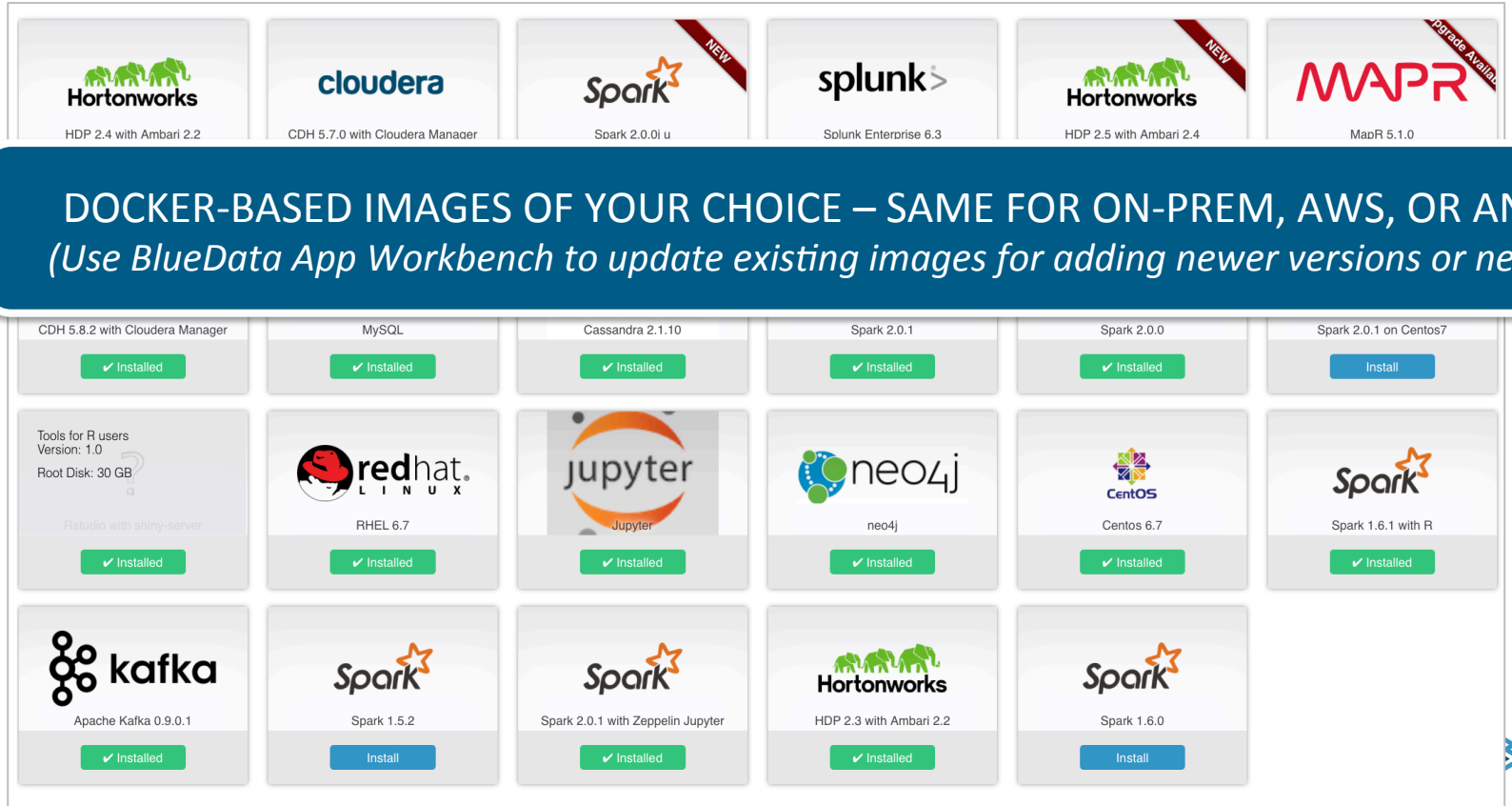
Simplified UI for End Users

AWS UI abstracted from end users (incl. admins) and is “under the hood”



Simplified Cloud Onboarding

Eliminate extensive "DevOps" FTE investment



DOCKER-BASED IMAGES OF YOUR CHOICE – SAME FOR ON-PREM, AWS, OR ANY CLOUD

(Use BlueData App Workbench to update existing images for adding newer versions or net new apps)

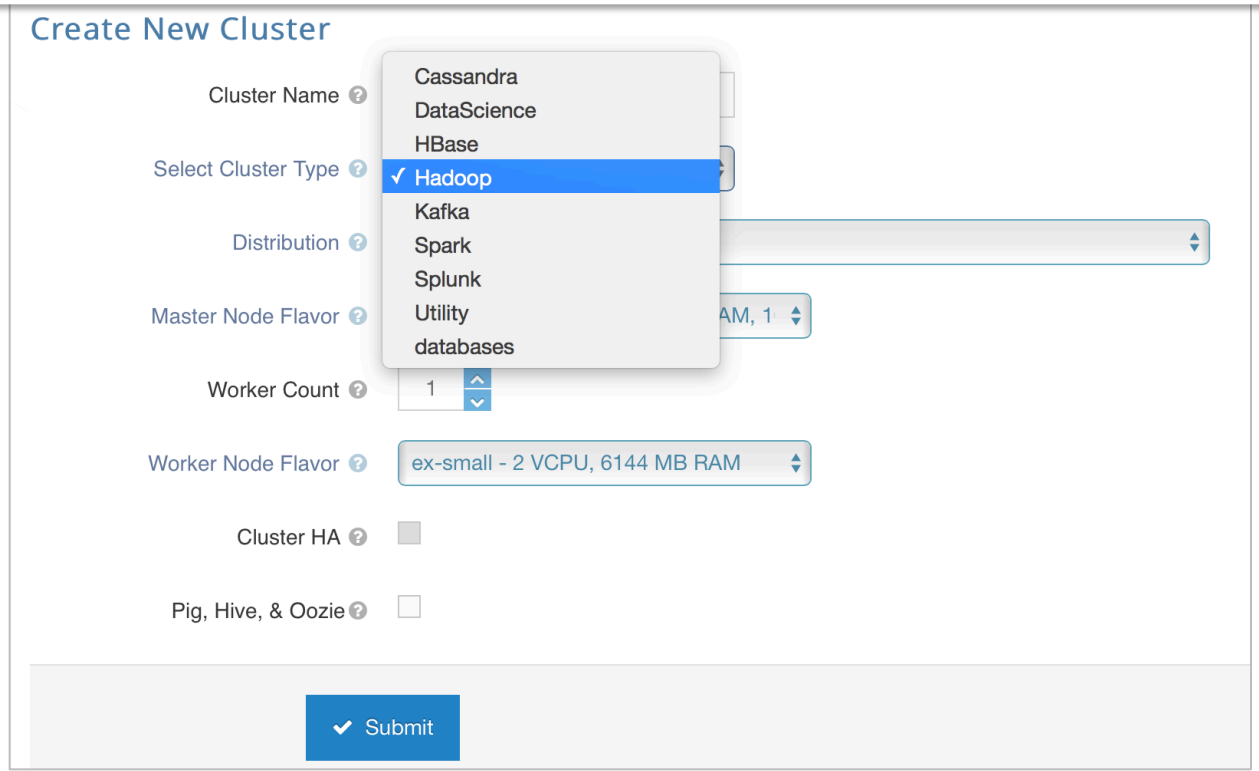
bluedata™

Simplified Cloud Onboarding

Eliminate extensive "DevOps" FTE investment

POINT & CLICK CLUSTER CREATION – DATA-DRIVEN BASED ON IMAGE DEFINITION

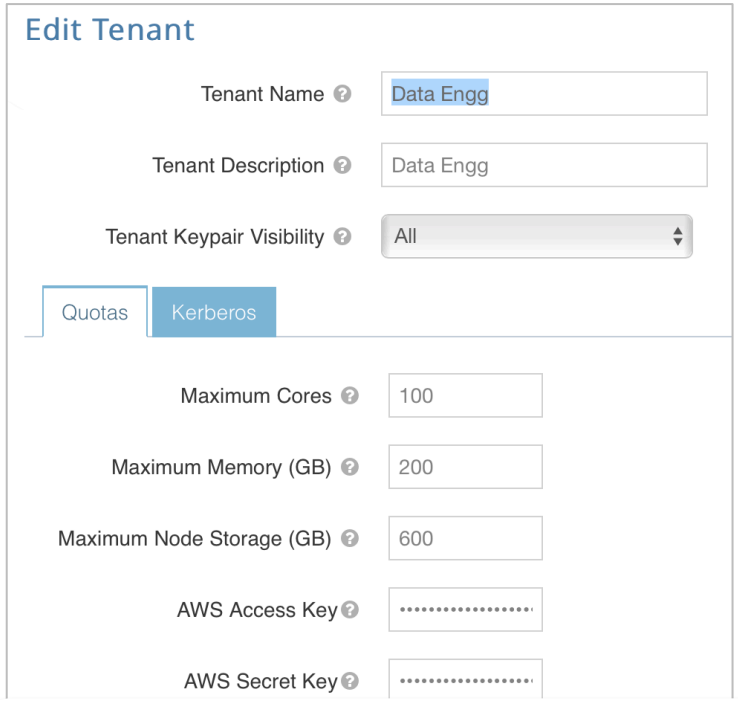
Fully automated orchestration, AWS resource controls, security, and elasticity (scale up/down)



bluedata

Resource and Cost Controls

Onboard multiple users/groups with fine-grained quotas & visibility

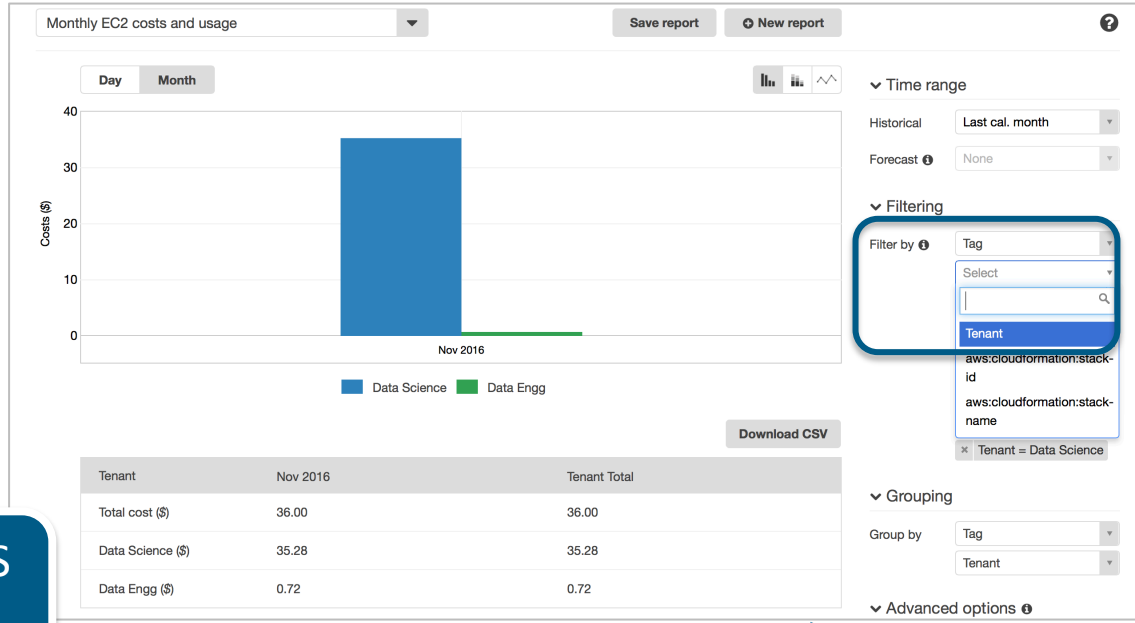


MULTI-TENANT RESOURCE CONTROLS

# of EC2 instances that can be created by a user are limited by the quotas, thereby reducing costs

COST VISIBILITY

Admins can view costs by tenants



Tenant	Nov 2016	Tenant Total
Total cost ($)	36.00	36.00
Data Science ($)	35.28	35.28
Data Engg ($)	0.72	0.72

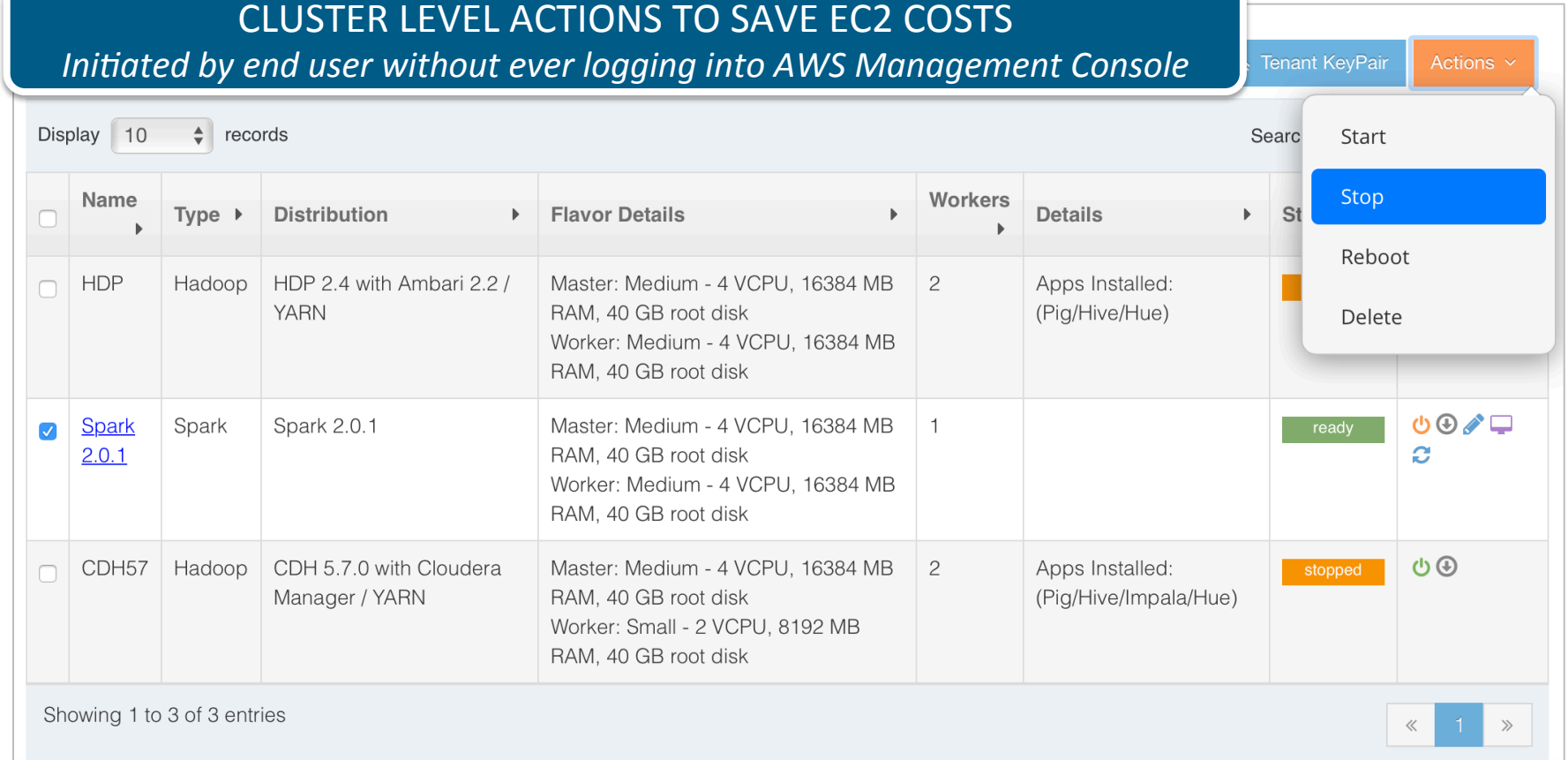
bluedata™

Resource and Cost Controls

Stop clusters & re-start with full state retained

CLUSTER LEVEL ACTIONS TO SAVE EC2 COSTS

Initiated by end user without ever logging into AWS Management Console



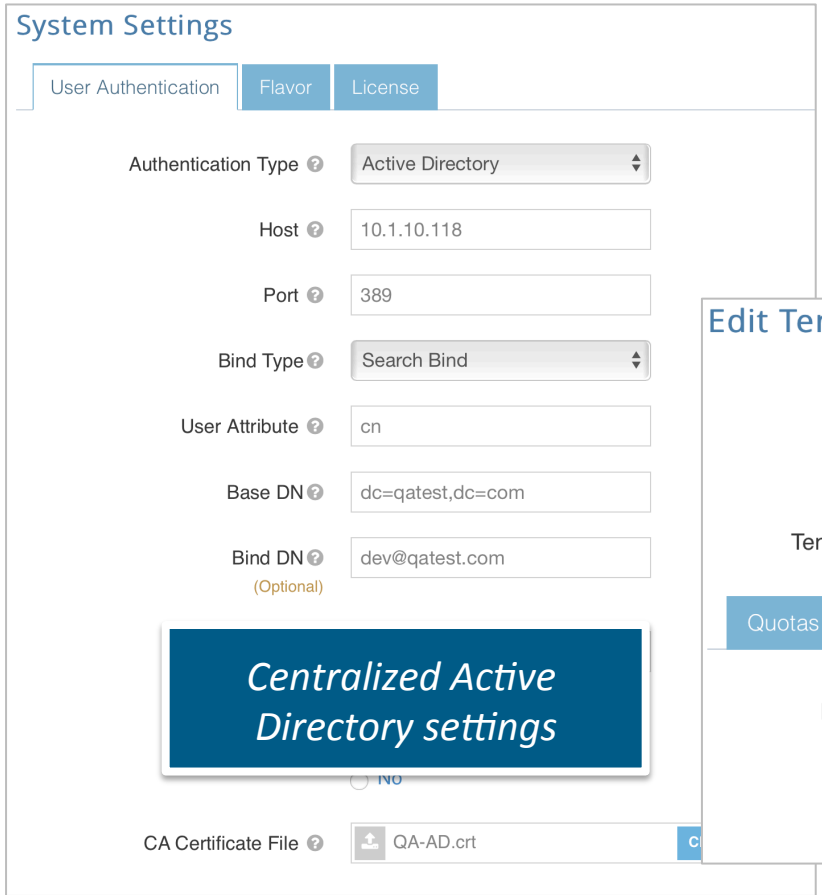
	Name	Type	Distribution	Flavor Details	Workers	Details	St	
☐	HDP	Hadoop	HDP 2.4 with Ambari 2.2 / YARN	Master: Medium - 4 VCPU, 16384 MB RAM, 40 GB root disk Worker: Medium - 4 VCPU, 16384 MB RAM, 40 GB root disk	2	Apps Installed: (Pig/Hive/Hue)		
☑	Spark 2.0.1	Spark	Spark 2.0.1	Master: Medium - 4 VCPU, 16384 MB RAM, 40 GB root disk Worker: Medium - 4 VCPU, 16384 MB RAM, 40 GB root disk	1		ready	
☐	CDH57	Hadoop	CDH 5.7.0 with Cloudera Manager / YARN	Master: Medium - 4 VCPU, 16384 MB RAM, 40 GB root disk Worker: Small - 2 VCPU, 8192 MB RAM, 40 GB root disk	2	Apps Installed: (Pig/Hive/Impala/Hue)	stopped	

Showing 1 to 3 of 3 entries

« 1 »

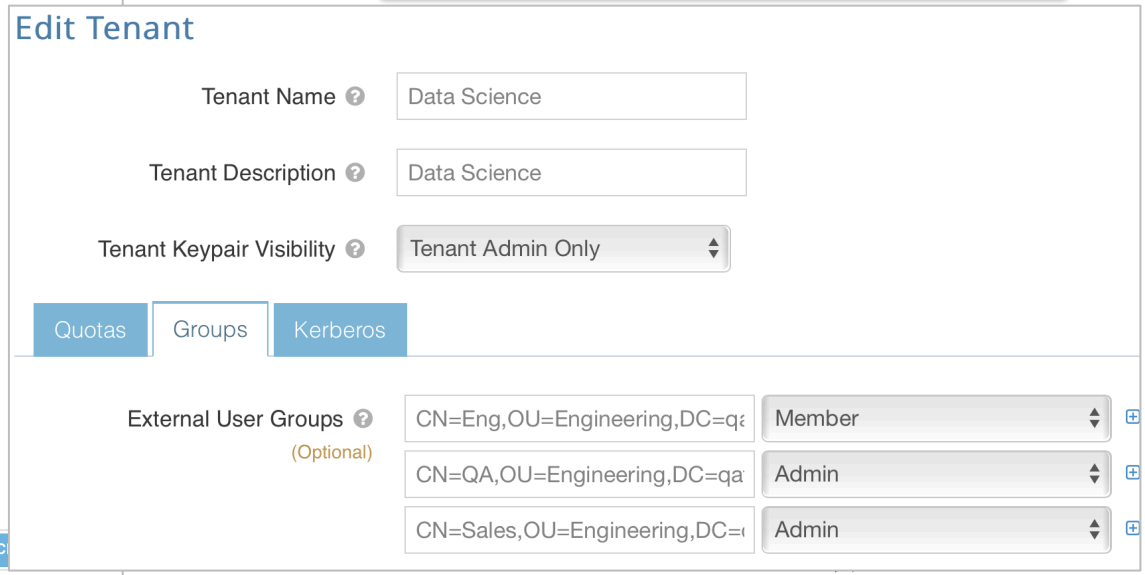
Pre-built Security Integrations

LDAP/AD integration to BlueData EPIC UI and tenants



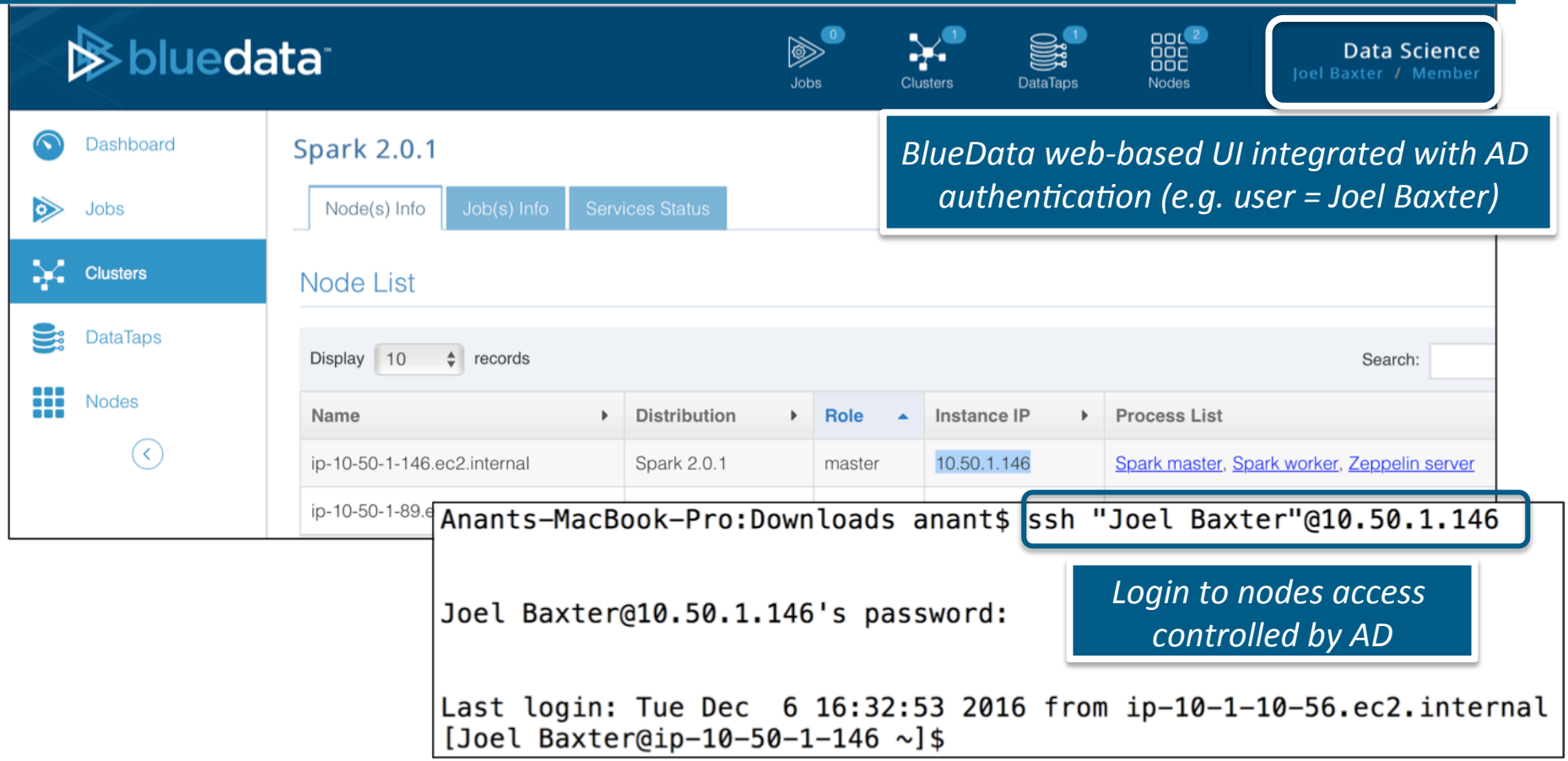
Centralized Active Directory settings

Tenant-level AD/LDAP group mappings



Pre-built Security Integrations

LDAP/AD integration to BlueData EPIC UI and Big Data nodes

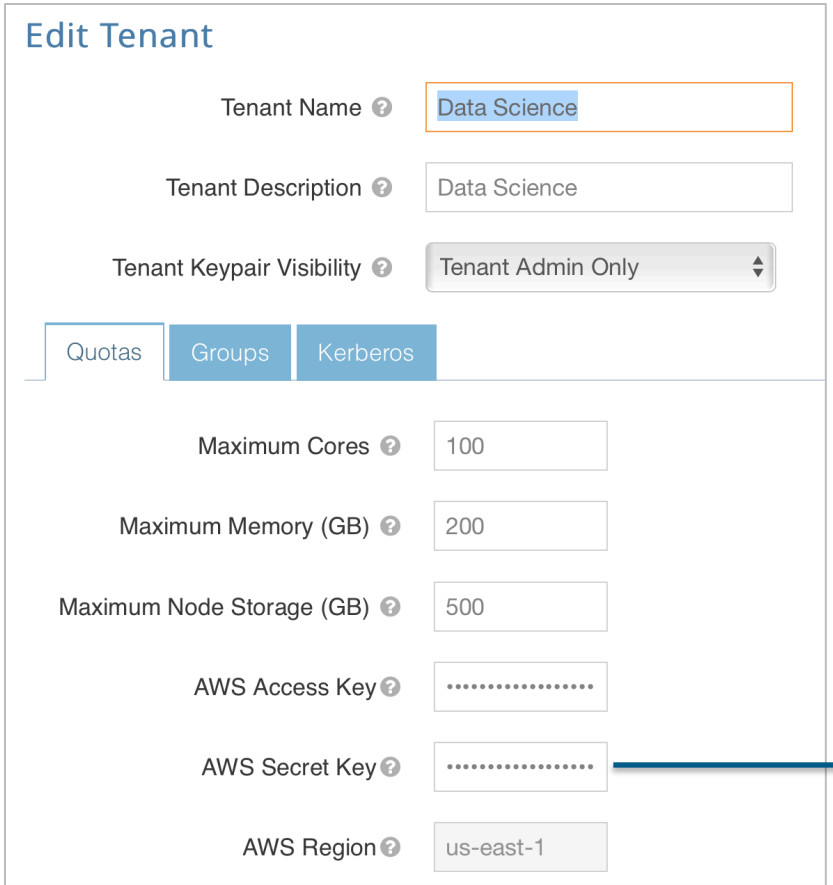


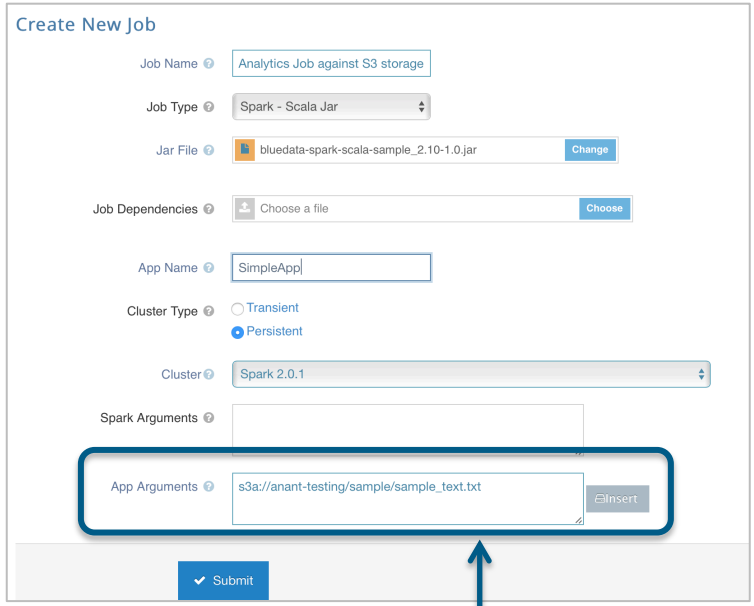
BlueData web-based UI integrated with AD authentication (e.g. user = Joel Baxter)

Login to nodes access controlled by AD

Pre-built Security Integrations

Automated S3a integration for tenant-level credentials



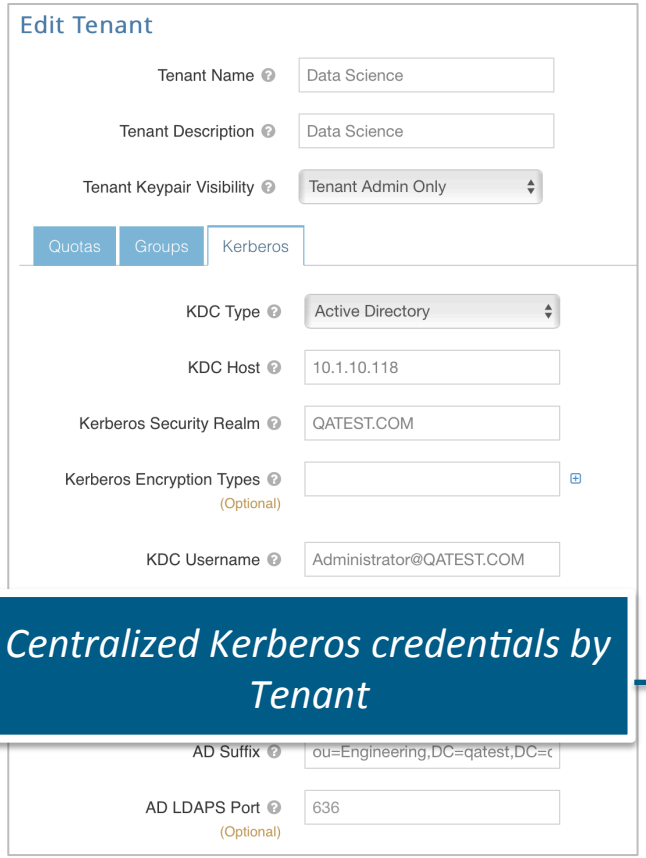


Injected into every Spark and Hadoop cluster. Access via s3a://....

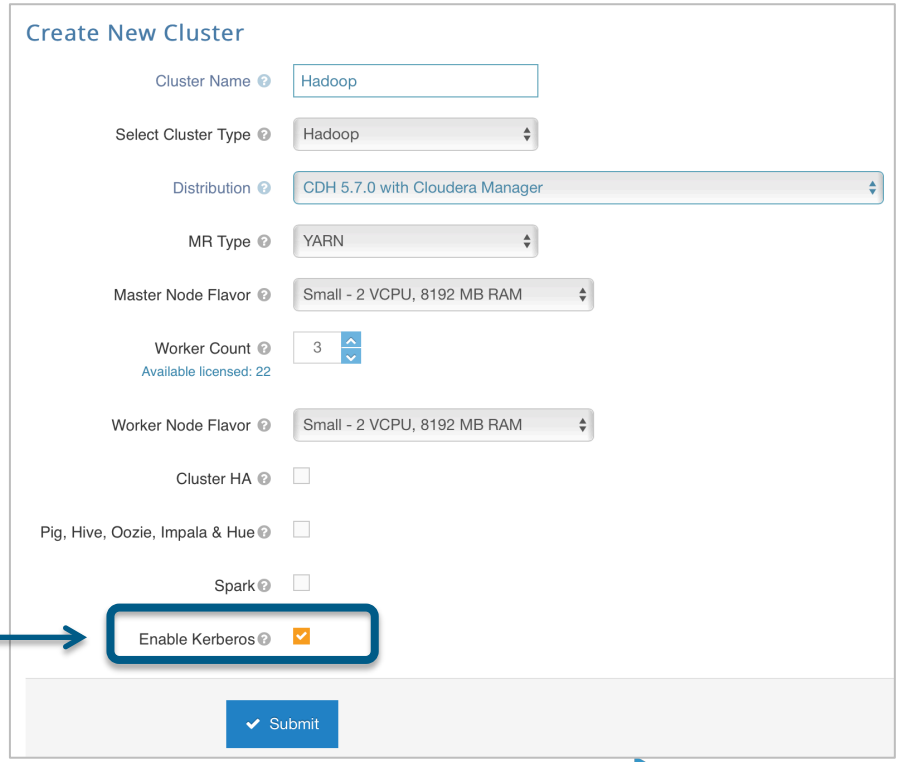
bluedata

Pre-built Security Integrations

Automated Kerberos



Centralized Kerberos credentials by Tenant



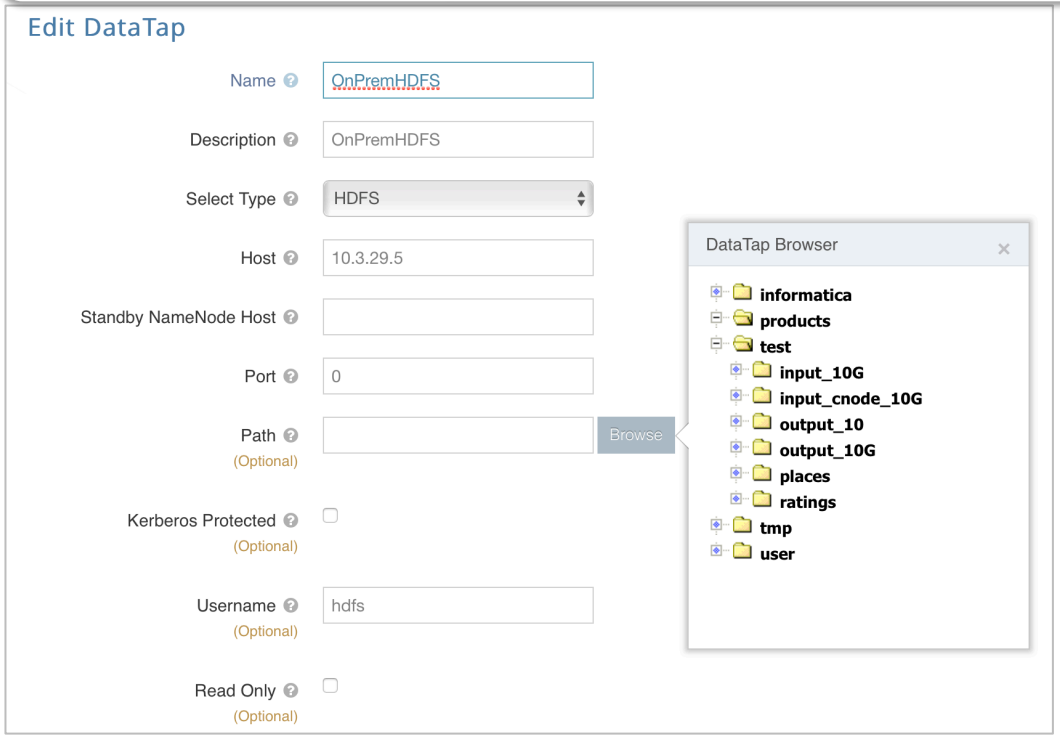
bluedata

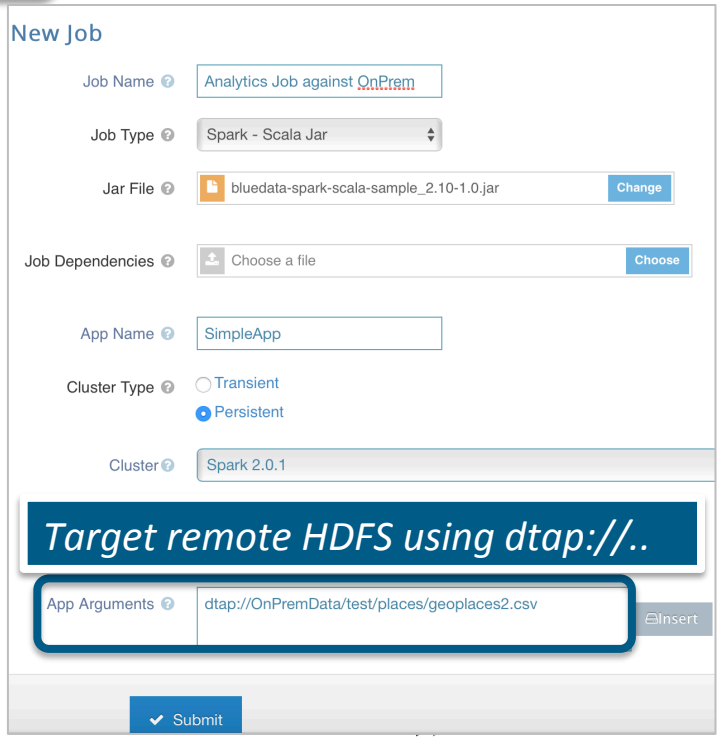
On-Prem Data Connectivity

Accelerate time-to-insight with security and control

CONNECT TO ANY ON-PREM HDFS OR NFS SYSTEM

** Site to site VPN and/or AWS Direct Connect recommended*





Customer Example: Fintech

Spark with Jupyter Notebooks and Cassandra on AWS

- Financial firm with proprietary research and trading technology
- Multiple developer and data science teams
- Custom data analytics pipelines
- App serving with Cassandra as backend
- Dev, Test, and Production environments
- Fully self-sufficient without AWS expertise

> “With BlueData, we can improve the productivity of our developers, shorten our development cycles, and control AWS costs.”

bluedata™

Customer Example: ISV

Cloudera, Hortonworks, & MapR on AWS

- Data integration ISV testing their product with multiple Hadoop distributions on AWS
- Multiple QA engineers & teams
- Quick self-service access to Hadoop clusters
- Allow DevOps teams to focus on other core competencies
- Create images of new versions of Hadoop with product artifacts as necessary

> “With BlueData, we can be self-sufficient and accelerate our testing cycles by using reproducible Hadoop clusters integrated with our product.”

BlueData EPIC on AWS

Big Data App Choice and Flexibility
(vs. opinionated stack with other BDaaS)

Simplifies Big Data DevOps on AWS
(vs. DIY initiatives on AWS)

Speed & agility

Security & control

Efficiency & lower cost

bluedata

Why BlueData for BDaaS?

Flexible software solution (Your BDaaS) vs. rigid managed service (Their BDaaS)

- Bring your own distro/application, provide application and cloud flexibility

In-place analytics against on-premises data and/or cloud storage

- Leverage DataTap to connect Hadoop/Spark compute with remote storage

Vendor-agnostic, Docker container orchestration provides maximum flexibility

- Same Docker images for on-premises and cloud deployments

Unique hybrid and multi-cloud model for BDaaS *[Next: Azure and GCP support]*

- Start on-prem and extend to cloud, or focus on cloud-first and multi-cloud

bluedata™

Key Takeaways: Big Data on AWS

- Know your “use case” and map to the right type of BDaaS
  - AWS EMR may be good fit if the use case is singular (e.g. Hadoop/Spark)
  - Fully-managed SaaS if you care more about support than choice & flexibility
  - Flexible platform with choice of frameworks, apps, versions for DevOps agility
- Big-Data-as-a-Service is more than Hadoop or Spark-as-a-Service
  - Modern analytics include real-time streams, NoSQL, ML runtimes, notebooks, etc.
  - Streamlined interface with common tool chain is key for faster time-to-value
  - Future-proof your implementations to deploy on-prem and/or multiple clouds
- DIY on AWS will be DOA … leverage a purpose-built platform
  - Focus on your Big Data applications, not the AWS infrastructure and plumbing
  - BlueData EPIC is a turnkey, flexible BDaaS platform – for on-prem and cloud

bluedata™

Tom Phelan

@tapbluedata

Anant Chintamaneni

@anantcman

bluedata

Thank You

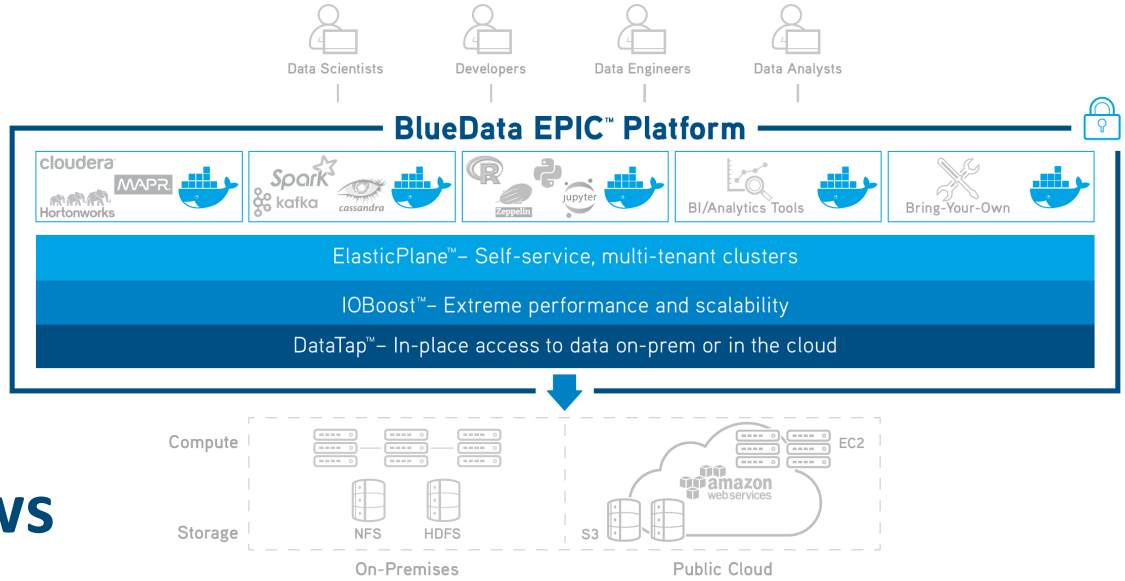
For more information:

www.bluedata.com

sales@bluedata.com

www.bluedata.com/aws

TRY BLUEDATA EPIC ON AWS



bluedata™