New In HDP 2.3

$(whoami)

Ajay Singh

Director Technical Channels

About Hortonworks

Customer Momentum

- 556 customers (as of August 5, 2015)
- 119 customers added in Q2 2015
- Publicly traded on NASDAQ: HDP

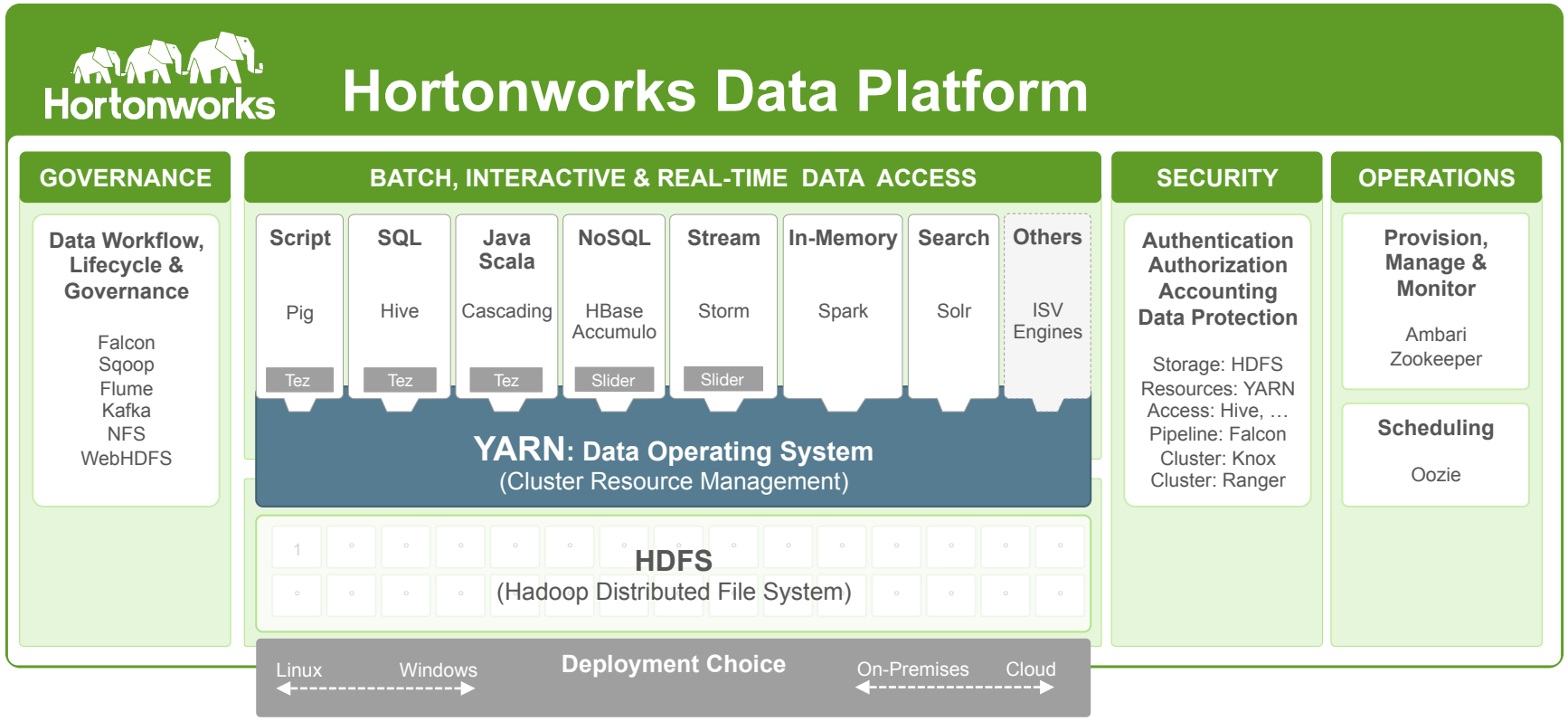
Hortonworks Data Platform

- Completely open multi-tenant platform for any app and any data
- Consistent enterprise services for security, operations, and governance

Partner for Customer Success

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

HDP Is Enterprise Hadoop



YARN
is the architectural center of HDP

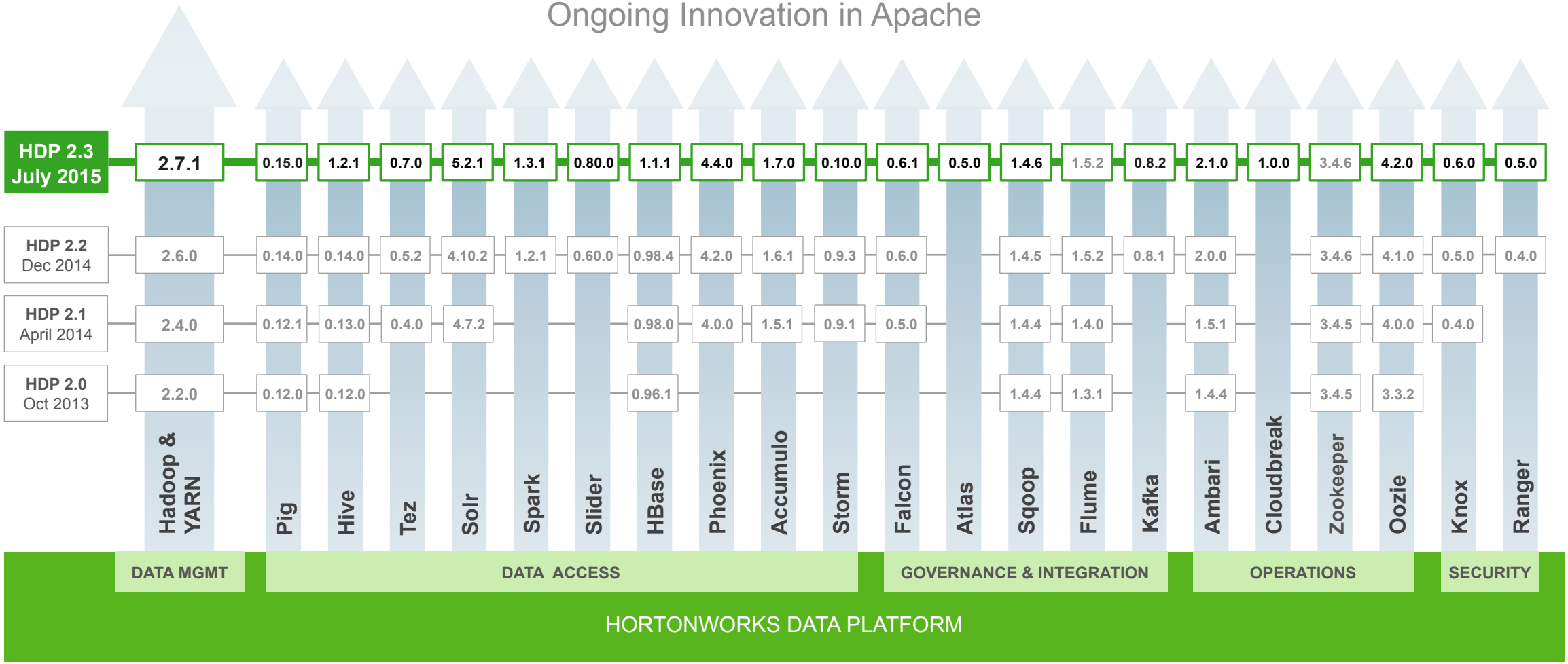
Enables batch, interactive and real-time workloads

Provides comprehensive enterprise capabilities

The widest range of deployment options

Delivered Completely in the OPEN

Hortonworks Data Platform



New Capabilities in Hortonworks Data Platform 2.3

Breakthrough User Experience

Dramatic Improvement in the User Experience
HDP 2.3 eliminates much of the complexity administering Hadoop and improves developer productivity.

Enhanced Security and Data Governance

Enhanced Security and Data Governance
HDP 2.3 delivers new encryption of data at rest, and extends the data governance initiative with Apache™ Atlas.

Proactive Support

Extending the Value of a Hortonworks Subscription
Hortonworks® SmartSense™ adds proactive cluster monitoring, enhancing Hortonworks' award-winning support in key areas.

New In Apache Hadoop

Hortonworks®

HDP Core

User Experience

- Guided Configuration
- Install/Manage/Monitor NFS Gateway
- Customizable Dashboards
- Files View
- Capacity Scheduler

Security

- HDFS Data Encryption at Rest
- Yarn Queue ACLs through Ranger

Workload Management

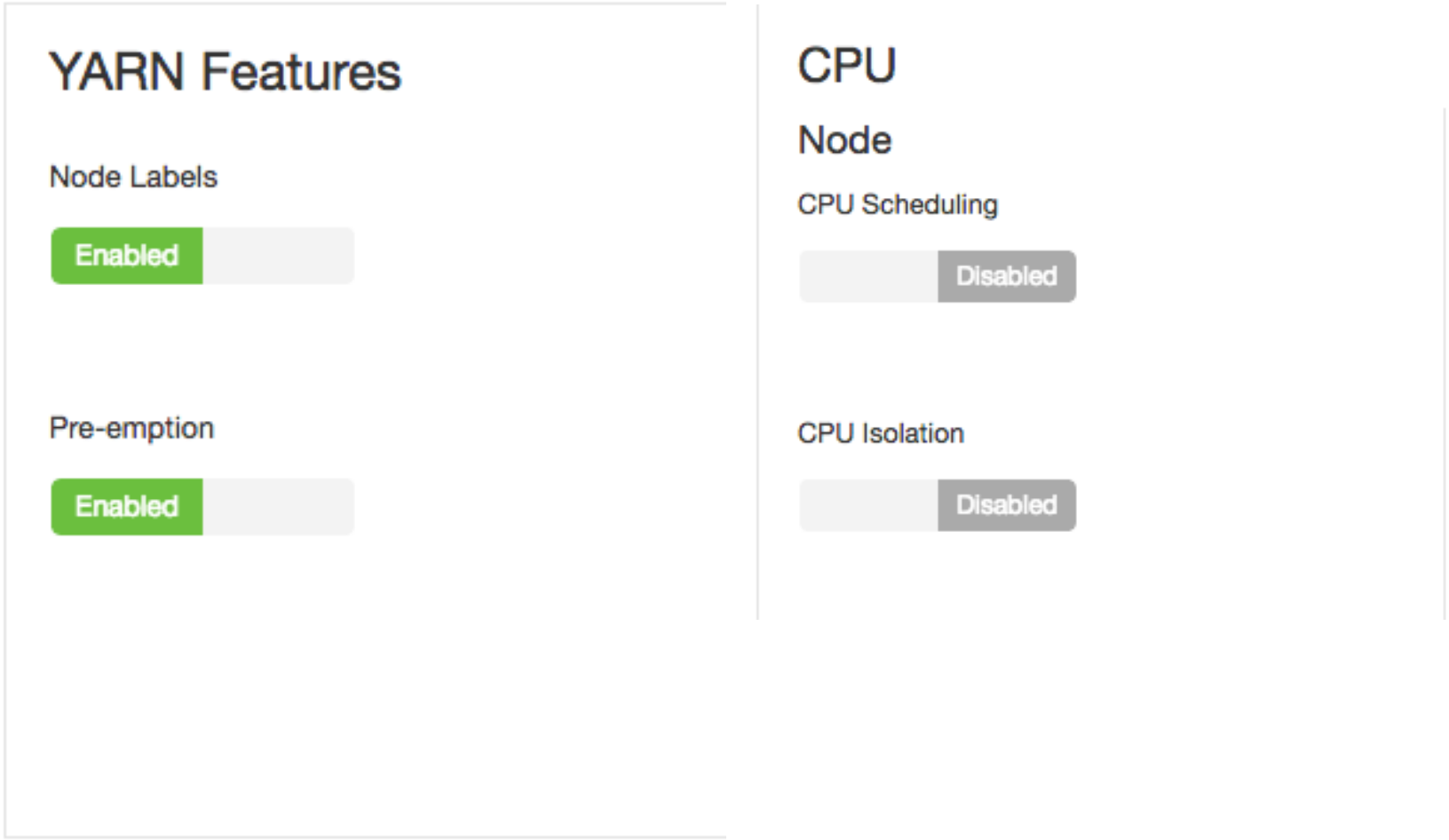
- Non-Exclusive Node Labels
- Fair Sharing Policy
- [TP] Local Disk Isolation

Operations

- Report on Bad Disks
- Enhanced Distcp (using snapshots)
- Quotas for Storage Tiers

Hortonworks®

Simplified Configuration Management



Hortonworks®

Deploy/Manage/Monitor NFS through Ambari

Deploy

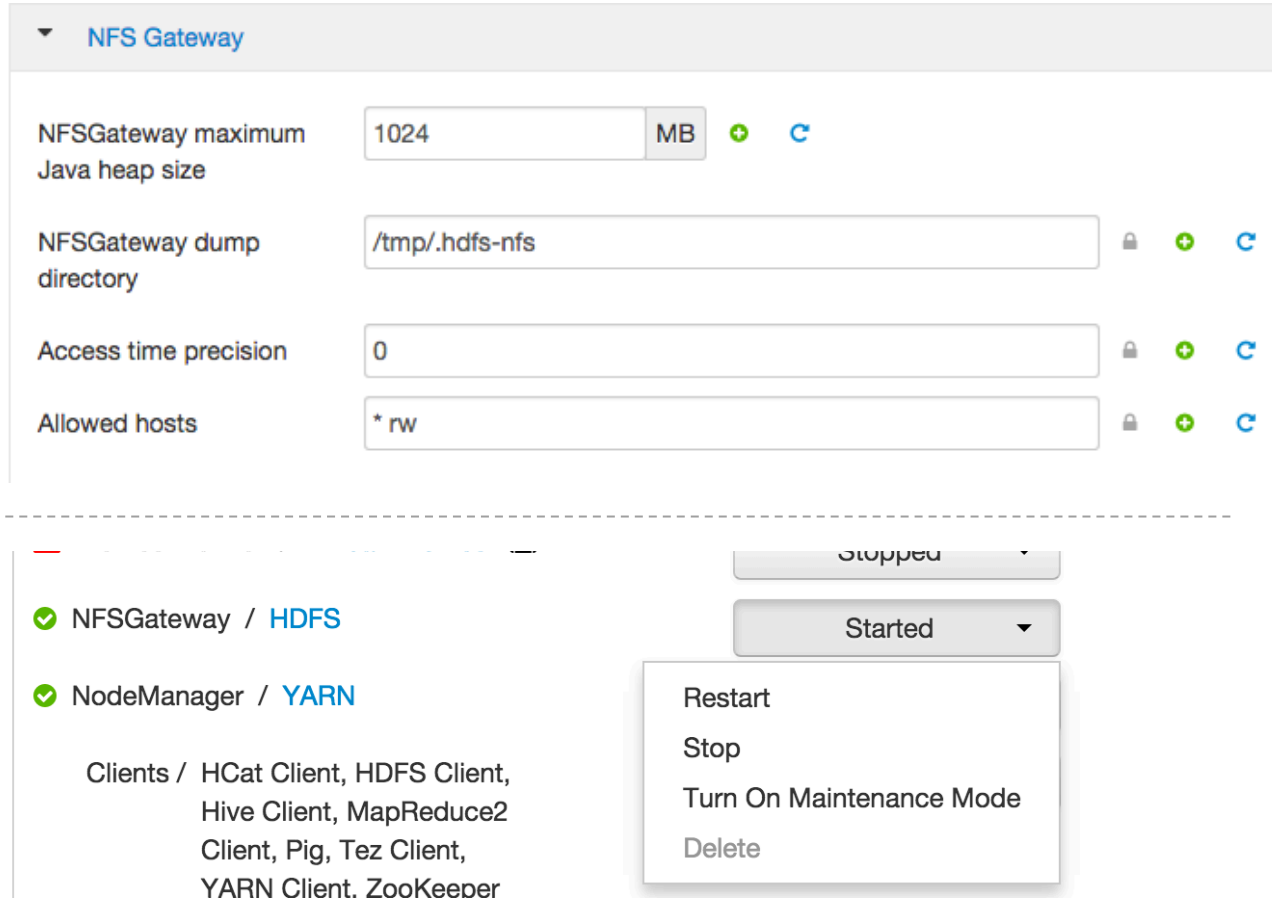
Assign Slaves and Clients

Assign slave and client components to hosts you want to run them on.
Hosts that are assigned master components are shown with ✳.
"Client" will install HDFS Client, MapReduce2 Client, YARN Client, Tez Client and ZooKeeper Client.

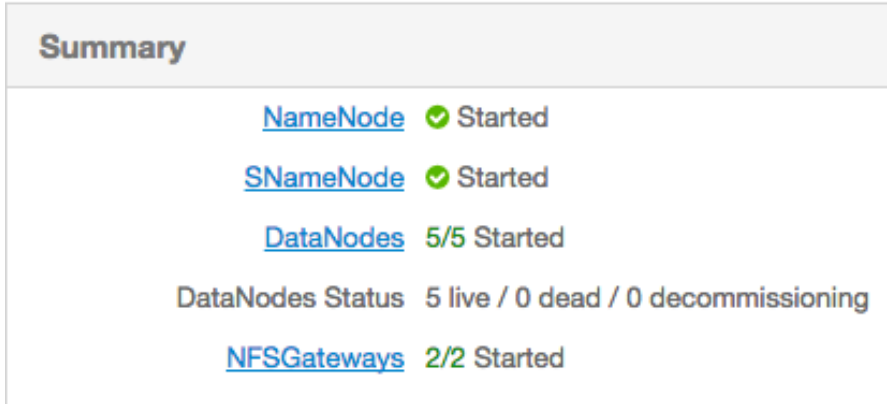
Host	all | none	all | none	all | none	all | none
ambari-nfs-1.c.pramod-th... ✳	DataNode	NFSGateway	NodeManager	Client

Starts ‘portmap’ and ‘nfs’

Manage



Monitor



Hortonworks®

Detect Bad Disks

Detect “bad” disk volumes on a DataNode

In operation

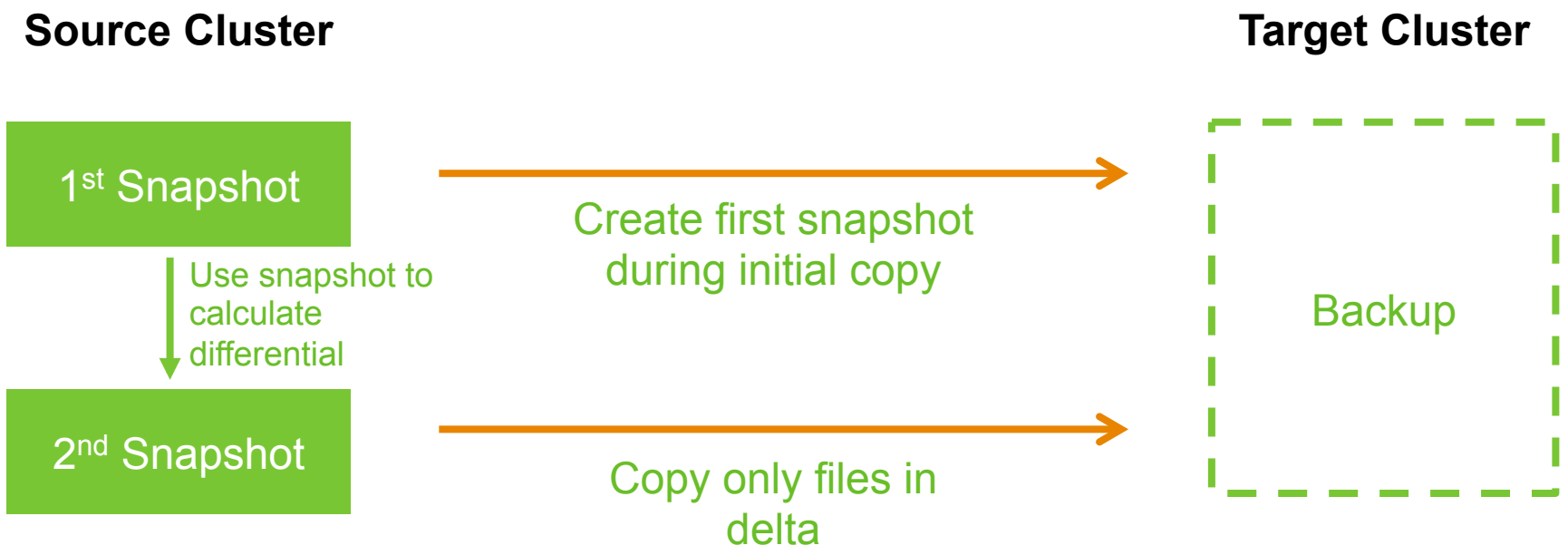
Node	Last contact	Admin State	Capacity	Used	Non DFS Used	Remaining	Blocks	Block pool used	Failed Volumes	Version
rohit-dal-core-4.c.pramod-thangali.internal:50010 (10.240.227.252:50010)	2	In Service	23.48 GB	217.78 MB	1.29 GB	21.98 GB	18	217.78 MB (0.91%)	0	2.7.1.2.3.0.0-2410
rohit-dal-core-5.c.pramod-thangali.internal:50010 (10.240.75.9:50010)	1	In Service	23.48 GB	129.32 MB	1.29 GB	22.06 GB	14	129.32 MB (0.54%)	0	2.7.1.2.3.0.0-2410
rohit-dal-core-8.c.pramod-thangali.internal:50010 (10.240.217.176:50010)	0	In Service	23.48 GB	314.72 MB	1.35 GB	21.82 GB	18	314.72 MB (1.31%)	0	2.7.1.2.3.0.0-2410
rohit-dal-core-7.c.pramod-thangali.internal:50010 (10.240.246.69:50010)	0	In Service	23.48 GB	310.92 MB	1.29 GB	21.89 GB	26	310.92 MB (1.29%)	0	2.7.1.2.3.0.0-2410
rohit-dal-core-6.c.pramod-thangali.internal:50010 (10.240.55.210:50010)	0	In Service	23.48 GB	348.03 MB	1.6 GB	21.54 GB	26	348.03 MB (1.45%)	0	2.7.1.2.3.0.0-2410

HDFS-7604

Enhanced HDFS Mirroring

Efficiency: Snapshot Diff faster than MapReduce based Diff for large directories

Reliability: Snapshots ensure catch any changes to Source directory during Distcp do not disrupt mirror

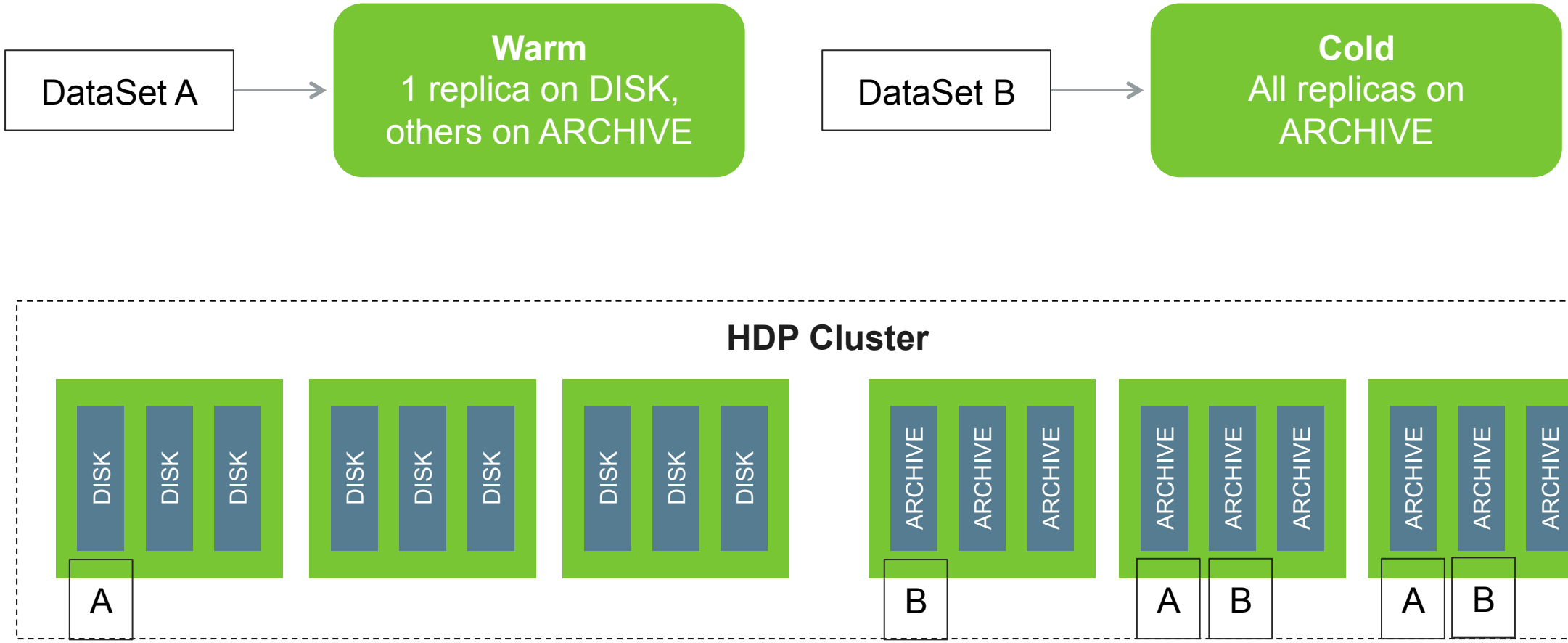


HDFS-7535

Hortonworks®

Quota Management By Storage Tiers

HDP 2.2



Hortonworks®

HDFS Quotas: Extending to Tiered Storage

Quota: Number of files for a directory

```
hdfs dfsadmin –setQuota n <list of directories>
```

Sets total number of files that can be stored in each directory.

Quota: Total disk space for a directory

```
hdfs dfsadmin –setSpaceQuota n <list of directories>
```

Sets total disk space that can be used by each directory.

New in HDP 2.3: Quota by Storage Tier

```
hdfs dfsadmin –setSpaceQuota n [-storageType <type>] <list of directories>
```

Sets total disk space that can be used by each directory.

HDFS-7584

Hortonworks®

Node Labels in YARN

Enable configuration of node partitions

Now with HDP 2.3, two options:

Non-exclusive Node Labels

Exclusive Node Labels

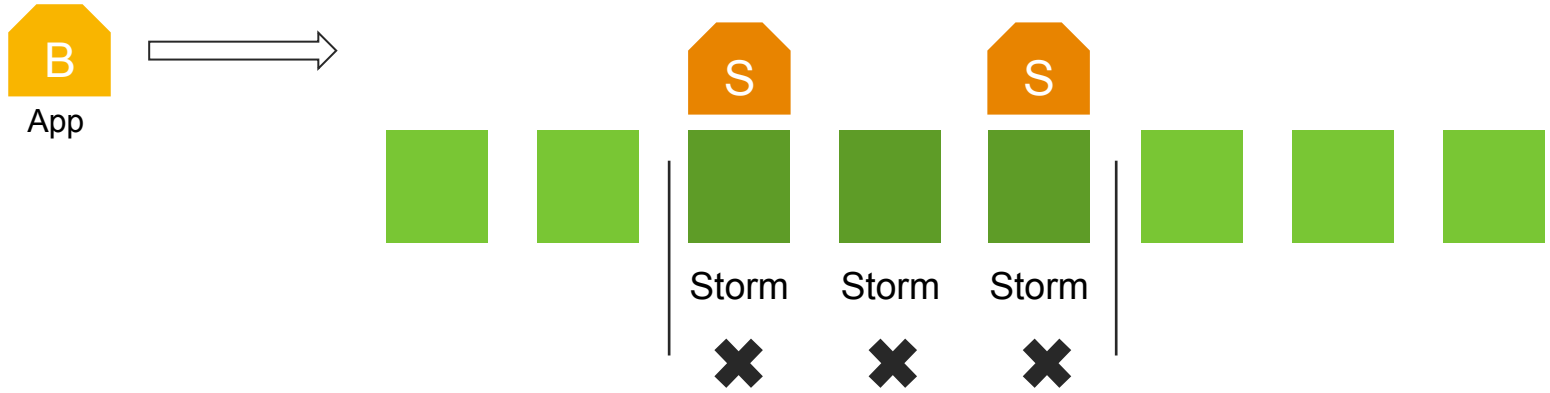
Hortonworks®

Exclusive Node Labels enable Isolated Partitions

HDP 2.2

Configure Partitions

Exclusive Labels enforce Isolation

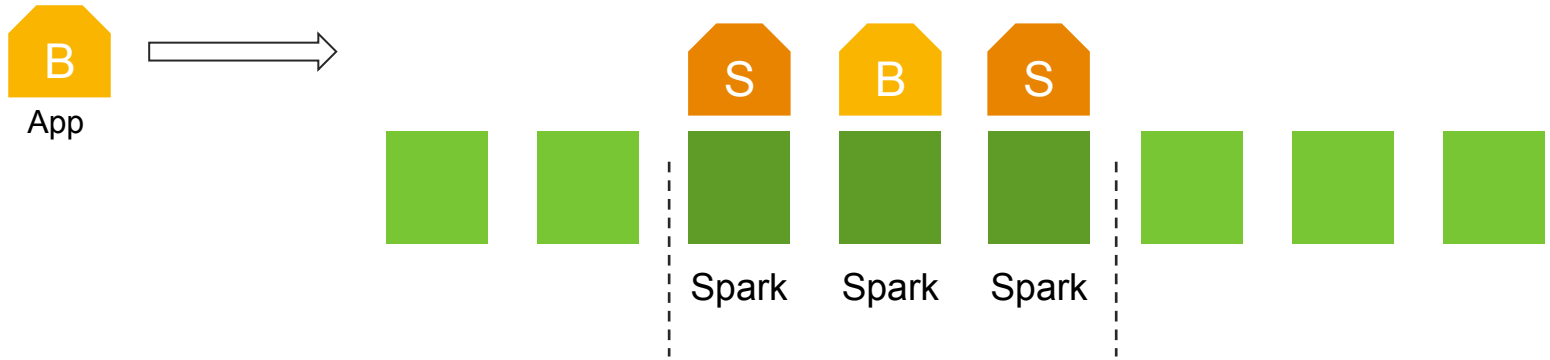


Hortonworks®

Non-Exclusive Node Labels

Configure non-exclusive labels

Schedule if free capacity



YARN-3214

Hortonworks®

Fair Sharing: Pluggable Queue Policies

Choose scheduling policy per leaf queue

FIFO

Application Container requests accommodated on first come first serve basis

Multi-fair weight

Application Container requests accommodated according to:

- Order of least resources used – multiple applications make progress
- (Optional) Size based weight – adjustment to boost large applications making progress

YARN-3318 YARN-3319

New In Apache Hive

Hortonworks®

Hive

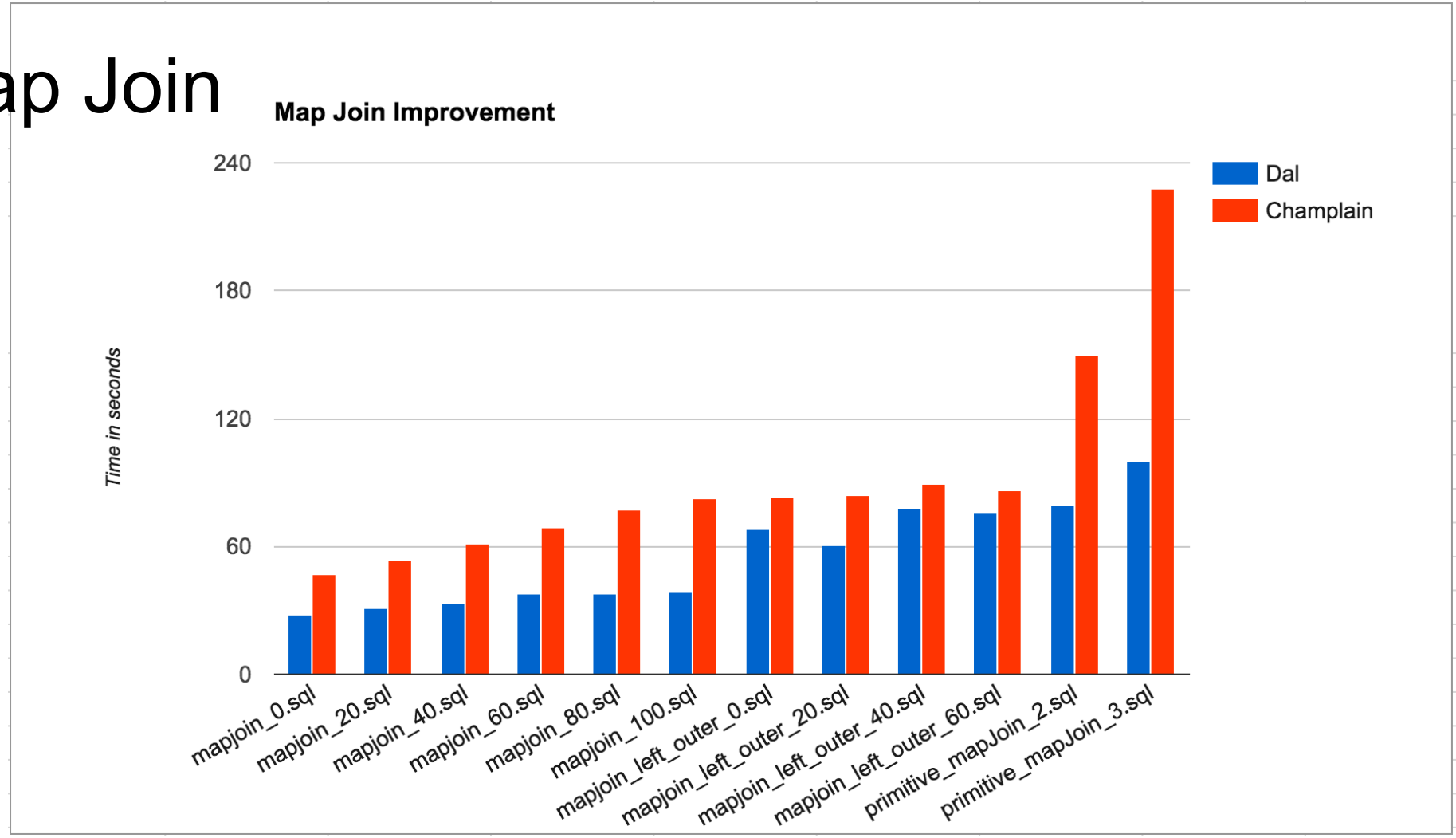
- **Performance**
  - Vectorized Map Joins and other improvements
- **SQL**
  - Union
  - Interval types
  - CURRENT_TIMESTAMP, CURRENT_DATE
- **Usability**
  - Configurations
  - Hive View
  - Tez View

Hortonworks®

Vectorized Map Join

Map Join is up to 5x faster, making the overall query up to 2x faster in HDP 2.3 over Champlain

mapjoin_20.sql means the query had a selectivity of 20 or 20% of rows end up joining




```
SELECT Count(*)
FROM   store_sales
       JOIN customer_demographics2
       ON ss_cdemo_sk = cd_demo_sk
       AND cd_demo_sk2 < 96040
       AND ss_sold_date_sk BETWEEN 2450815 AND 2451697
```

```
SELECT Count(*)
FROM   store_sales
       LEFT OUTER JOIN customer_demographics2
       ON ss_cdemo_sk = cd_demo_sk
       AND cd_demo_sk2 < 96040
       AND ss_sold_date_sk BETWEEN 2450815 AND 2451697
```

New SQL Syntax: Union

```
create table sample_03(name varchar(50), age int, gpa decimal(3, 2));

create table sample_04(name varchar(50), age int, gpa decimal(3, 2));
```

```
insert into table sample_03 values
('aaa', 35, 3.00),
('bbb', 32, 3.00),
('ccc', 32, 3.00),
('ddd', 35, 3.00),
('eee', 32, 3.00);

insert into table sample_04 values
('ccc', 32, 3.00),
('ddd', 35, 3.00),
('eee', 32, 3.00),
('fff', 35, 3.00),
('ggg', 32, 3.00);
```

```
hive> select * from sample_03 UNION select * from sample_04;
Query ID = ambari-qa_20150526023228_198786c5-5c89-4a38-9246-cbba9b903ab4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id
application_1432604373833_0002)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Map 4 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 8.48 s
--------------------------------------------------------------------------------
OK
aaa  35   3
bbb  32   3
ccc  32   3
ddd  35   3
eee  32   3
fff  35   3
ggg  32   3
Time taken: 11.208 seconds, Fetched: 7 row(s)
```

New SQL Syntax: Interval Type in Expressions

```
hive> select timestamp '2015-03-08 01:00:00' + interval '1' hour;
OK
2015-03-08 02:00:00
Time taken: 0.136 seconds, Fetched: 1 row(s)
hive> select timestamp '2015-03-08 00:00:00' + interval '23' hour;
OK
2015-03-08 23:00:00
Time taken: 0.057 seconds, Fetched: 1 row(s)
hive> select timestamp '2015-03-08 00:00:00' + interval '24' hour;
OK
2015-03-09 00:00:00
Time taken: 0.149 seconds, Fetched: 1 row(s)
hive> select timestamp '2015-03-08 00:00:00' + interval '1' day;
OK
2015-03-09 00:00:00
Time taken: 0.063 seconds, Fetched: 1 row(s)
hive> select timestamp '2015-02-09 00:00:00' + interval '1' month;
OK
2015-03-09 00:00:00
Time taken: 0.107 seconds, Fetched: 1 row(s)
hive> select current_timestamp - interval '24' hour;
OK
2015-05-25 02:35:13.89
Time taken: 0.181 seconds, Fetched: 1 row(s)
```

```
hive> select current_date;
OK
2015-05-26
Time taken: 0.102 seconds, Fetched: 1 row(s
hive> select current_timestamp;
OK
2015-05-26 02:33:15.428
Time taken: 0.091 seconds, Fetched: 1 row(s
```


Hortonworks®

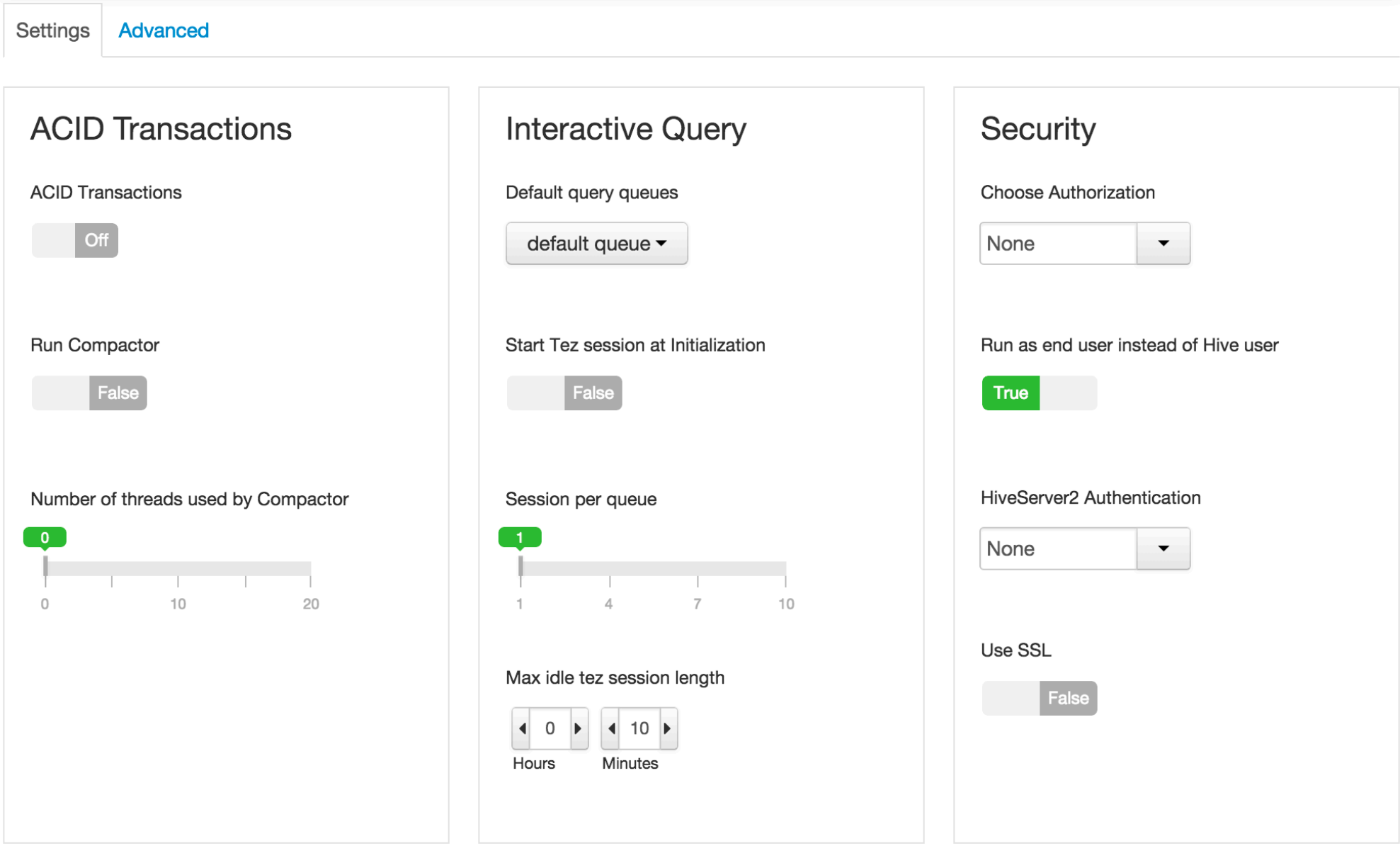
Not Supported: Interval Type in Tables

```
hive> CREATE TABLE t1 (c1 INTERVAL YEAR TO MONTH);
NoViableAltException(142@[])
    at org.apache.hadoop.hive.ql.parse.HiveParser.type(HiveParser.java:38574)
    at org.apache.hadoop.hive.ql.parse.HiveParser.colType(HiveParser.java:38331)
       ...
       at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
FAILED: ParseException line 1:20 cannot recognize input near 'INTERVAL' 'YEAR' 'TO' in column
type
```

```
hive> CREATE TABLE t1 (c1 INTERVAL DAY(5) TO SECOND(3));
NoViableAltException(142@[])
    at org.apache.hadoop.hive.ql.parse.HiveParser.type(HiveParser.java:38574)
    at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
FAILED: ParseException line 1:20 cannot recognize input near 'INTERVAL' 'DAY' '(' in column type
```


Hortonworks®

Simplified Configuration Management



Hortonworks®

New In Apache HBASE

Hortonworks®

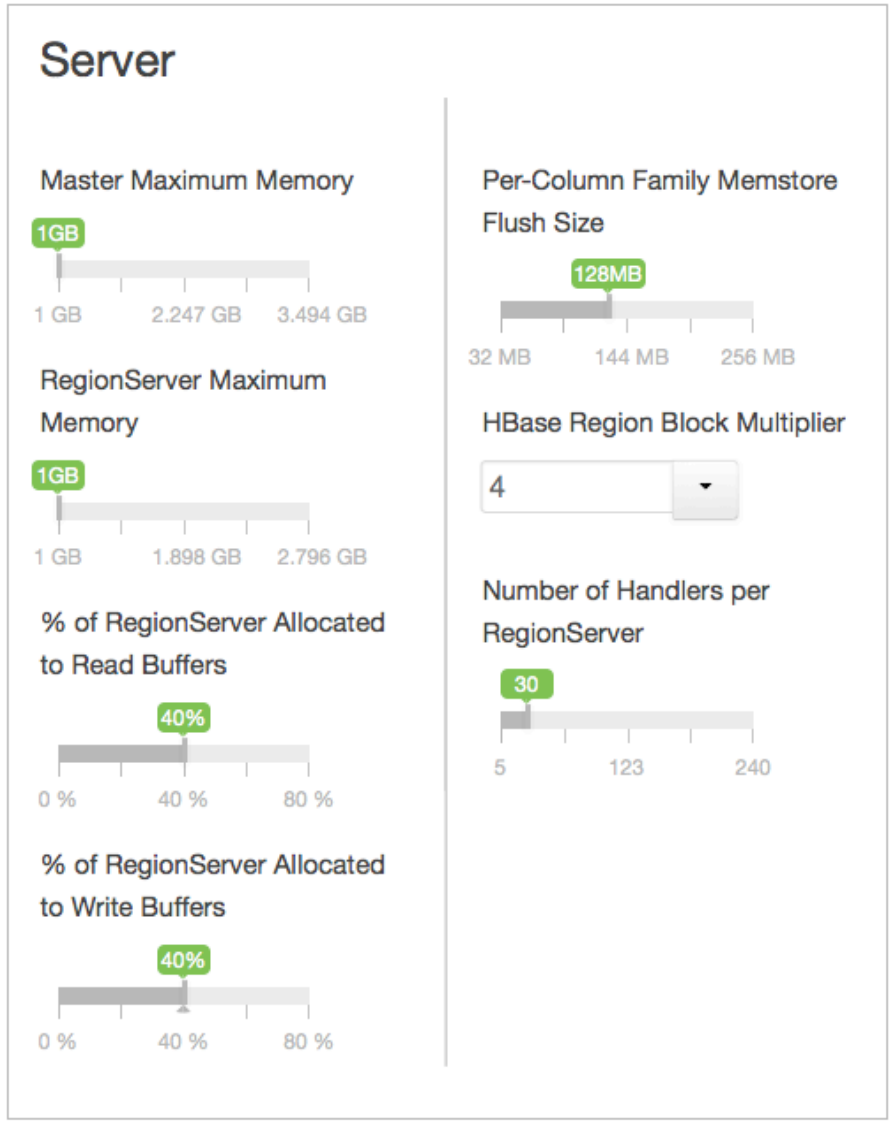
HBase and Phoenix in HDP 2.3

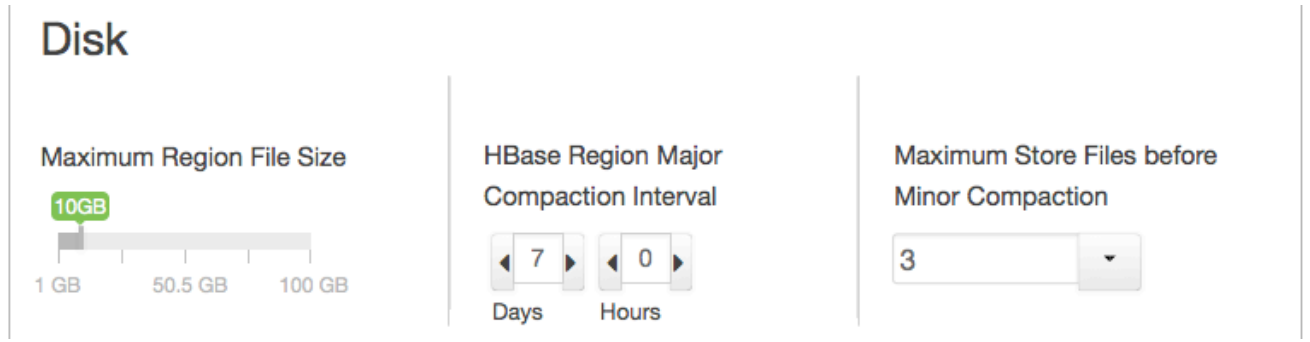
HBase and Phoenix in HDP 2.3			
	Operations	**Scale and Robustness**	**Developer**

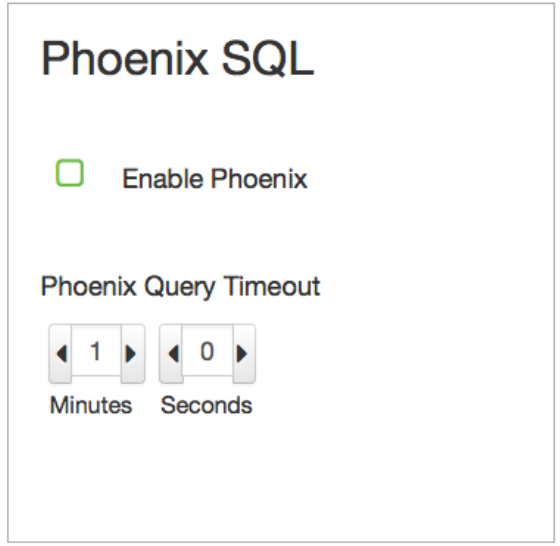
HBase	• Next Generation Ambari UI.
	• Customizable Dashboards.
	• Supported init.d scripts.
		• Improved HMaster Reliability
		• Security:
		  • Namespaces.
		  • Encryption.
		  • Authorization Improvements.
		  • Cell-Level Security.
			• LOB support

Phoenix	• Phoenix Slider Support
	• HBase Read HA Support
		• Functional Indexes
		• Query Tracing
			• Phoenix SQL:
			  • UNION ALL
			  • UDFs
			  • 7 New Date/Time Functions
			• Spark Driver
			• PhoenixServer

Simplified Configuration Management





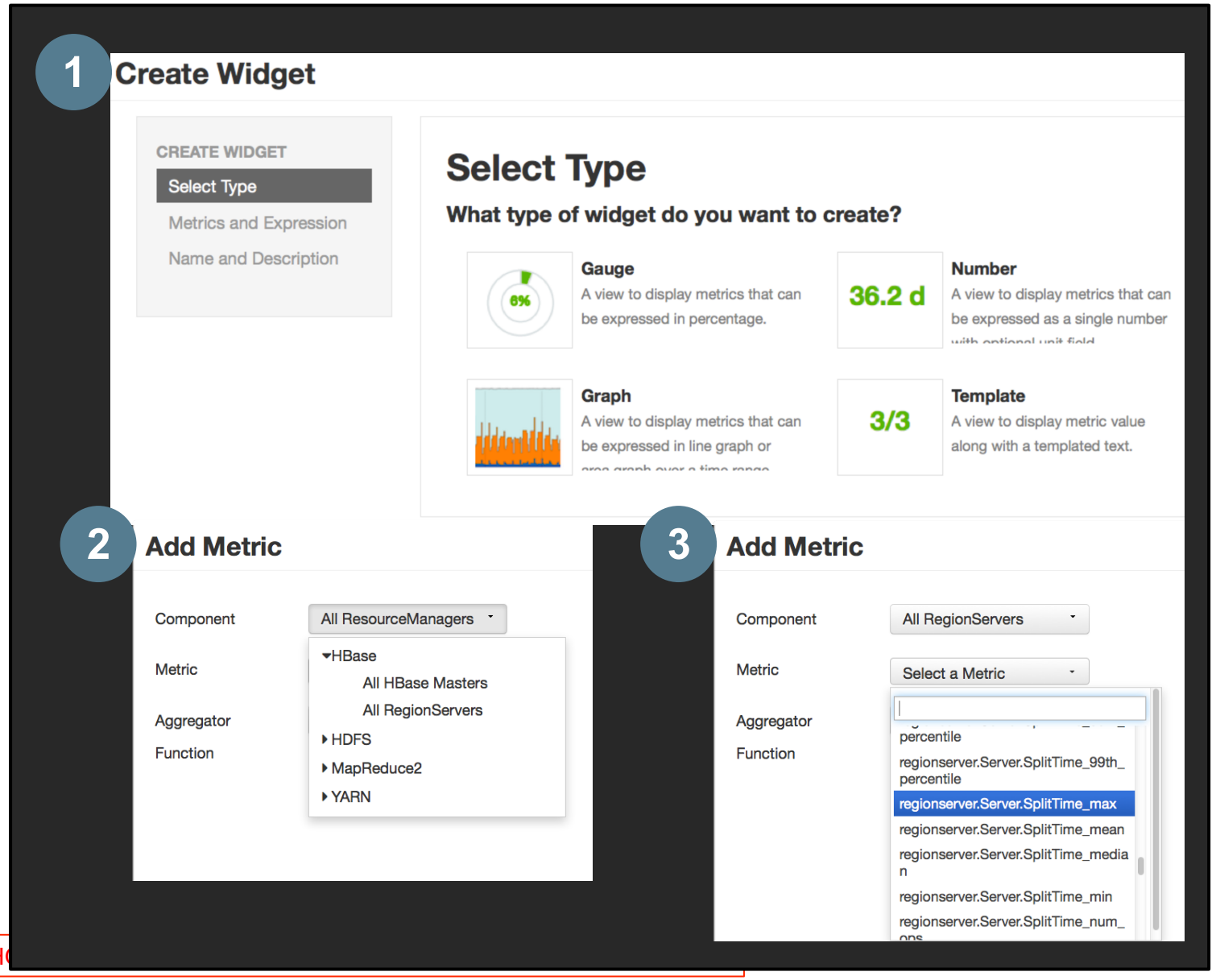


Guide configuration and provide recommendations for the most common settings.

Build Your Own HBase Dashboard

Monitor the metrics that matter to you.

1. Select a pre-defined visualization.
2. Choose from more than > 1000 metrics, ranging from HBase, HDFS, MapReduce2 and YARN.
3. Define custom aggregations for metrics within one component or across components.



Namespaces and Delegated Admin

Namespaces

- Namespaces are like RDBMS schemas.
- Introduced in HBase 0.96.
- Many security gaps until HBase 1.0.

Delegated Administration

- Goal: Create a namespace and hand it over to a DBA.
- People in the namespace can’t do anything outside their namespace.

Security: Namespaces, Tables, Authorizations

Scopes:

- Authorization scopes: Global -> namespace -> table -> column family -> cell.

Access Levels:

- Read, Write, Execute, Create, Admin

Delegated Administration Example

Give a user their own Namespace to play in.

- Step 1: Superuser (e.g. user hbase) creates namespace foo.
  - create_namespace ‘foo’
- Step 2: Admin gives dba-bar full permissions to the namespace:
  - grant ’dba-bar', 'RWXCA', '@foo’
  - Note: namespaces are prefixed by @.
- Step 3: dba-bar creates tables within the namespace:
  - create ’foo:t1', 'f1’
- Step 4: dba-bar hands out permissions to the tables:
  - grant ‘user-x’, ‘RWXCA’, ‘foo:t1’
- Note: All users will be able to see namespaces and tables within namespaces, but not the data.

Turning Authorization On

Turn Authorization On in Non-Kerberized (test) Clusters:

- Set `hbase.security.authorization` = true
- Set `hbase.coprocessor.master.classes` = `org.apache.hadoop.hbase.security.access.AccessController`
- Set `hbase.coprocessor.region.classes` = `org.apache.hadoop.hbase.security.access.AccessController`
- Set `hbase.coprocessor.regionserver.classes` = `org.apache.hadoop.hbase.security.access.AccessController`

Authorization in Kerberized Clusters:

- `hbase.coprocessor.region.classes` should have both `org.apache.hadoop.hbase.security.token.TokenProvider` and `org.apache.hadoop.hbase.security.access.AccessController`

SQL in Phoenix / HDP 2.3

UNION ALL

Date / Time Functions

- now(), year, month, week, dayofmonth, curdate
- hour, minute, second

Custom UDFs

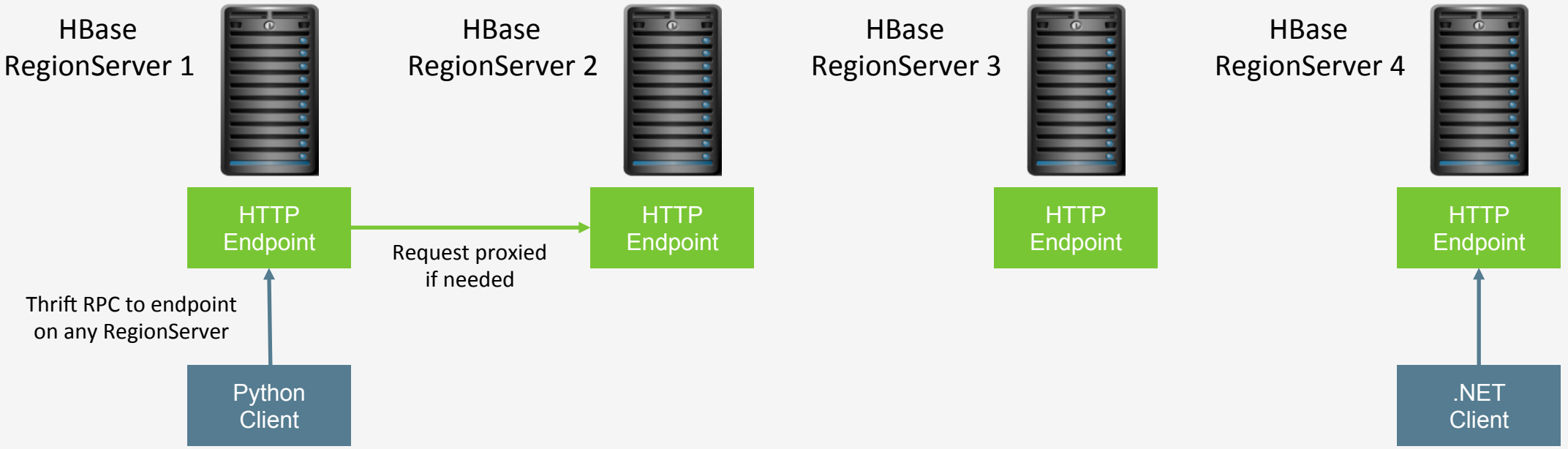
- Row-level UDFs.

Tracing

- Trace a query to pinpoint bottlenecks.

Hortonworks®

Phoenix Query Server: Supporting Non-Java Drivers



1. Endpoints colocated with RegionServers. No Single-Point-of-Failure. Optional loadbalancer.
2. Endpoints can proxy requests or perform local aggregations

Hortonworks®

Using Phoenix Query Server

Client Side:

- Thin JDBC Driver: /usr/hdp/current/phoenix/phoenix-thin-client.jar (1.7mb versus 44mb)
- Does not require Zookeeper access.
- Wrapper Script: sqlline-thin.py
- sqlline-thin.py https://host:8765

Server Side:

- Ambari Install and Management: Yes
- Port: Default = 8765

HTTP Example:

- `curl -XPOST -H 'request: {"request":"prepareAndExecute","connectionId":"aaaaaaaa-aaaa-aaaa-aaaa-aaaaaaaaaaaa","sql":"select count(*) from PRICES","maxRowCount":-1}'`

Phoenix / Spark integration in HDP 2.3

Phoenix / Spark Connector

- Load Phoenix tables / views into RDDs or DataFrames.
- Integrate with Spark, Spark Streaming and SparkSQL.

Hortonworks®

New In Apache Storm

Hortonworks®

Stream Processing Ready For Mainstream Adoption

Stream analysis, scalable across the cluster

Nimbus High Availability

No single point of failure for stream processing job management

Ease of Deployment

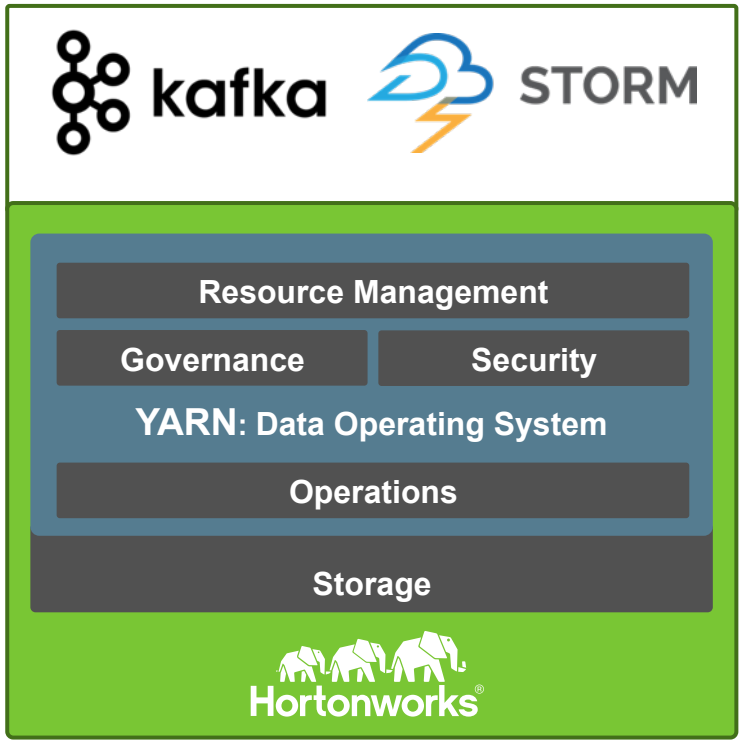
Quickly create stream processing pipelines via Flux

Rolling Upgrades

Update Storm to newer versions, with zero downtime

Enhanced Security for Kafka

Authorization via Ranger and authentication via Kerberos



Hortonworks

Connectivity Enhancements

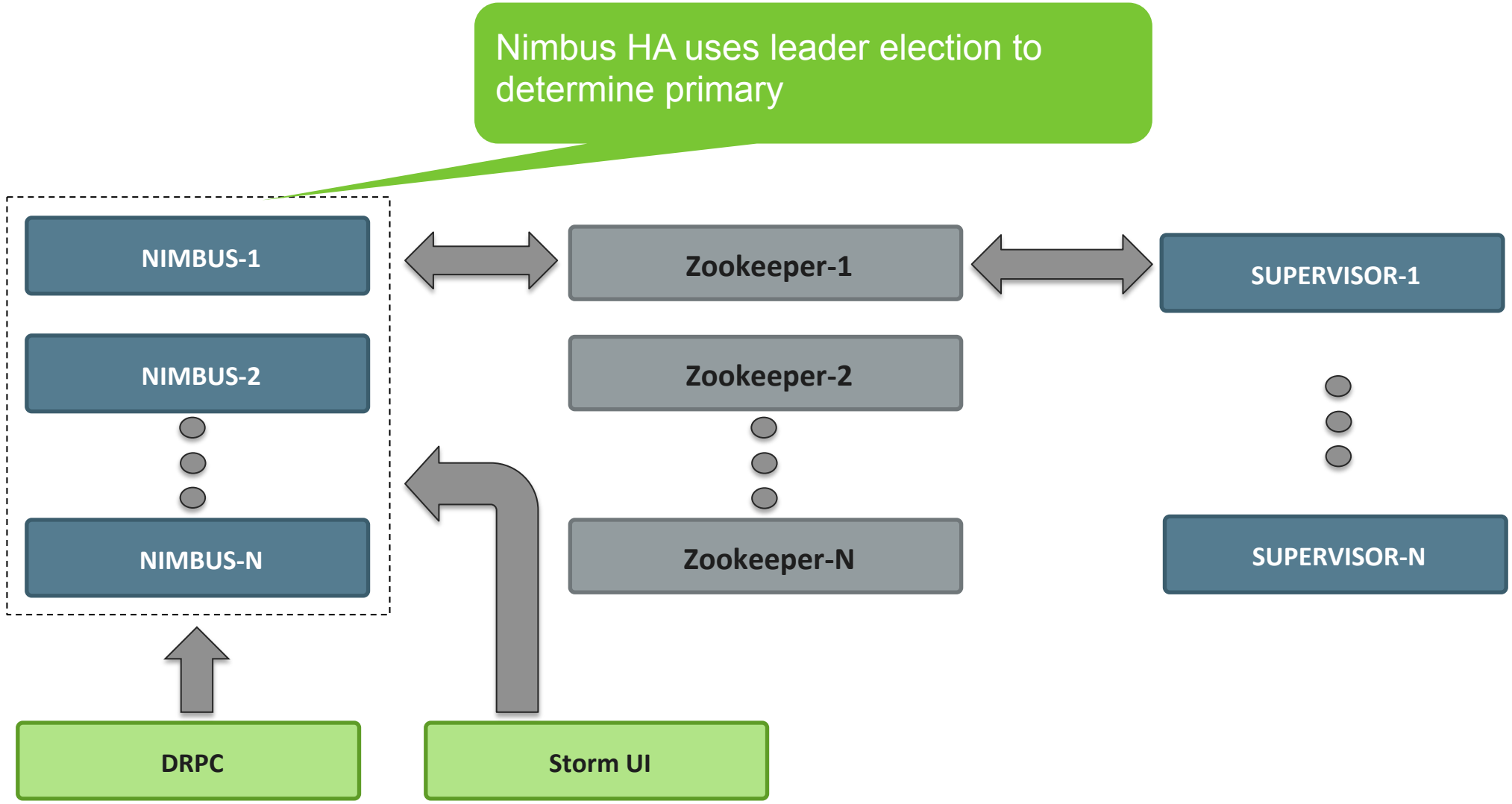
Apache Storm 0.10.0

- Microsoft Azure Event Hubs Integration
- Redis Support
- JDBC/RDBMS Integration
- Solr 5.2.1 -- Storm Bolt: some assembly required

Kafka 0.8.2

- Flume Integration (originally released in HDP 2.2) – not supported when Kafka Security is activated

Storm Nimbus High Availability



Productivity

Partial Key Groupings

- **The stream can be partitioned by fields specified in the grouping, like the Fields grouping, but in this case are load balanced between two downstream bolts, which provides better utilization of resources when the incoming data is skewed.**

Reduced Dependency Conflicts with shaded JARs

- **This enhancement provides clear separation between the Storm engine and supporting code from the topology code provided by developers.**

Productivity

Declarative Topology Wiring with Flux

- **Define Storm Core API (Spouts/Bolts) using a flexible YAML DSL**
- **YAML DSL support for most Storm components (storm-kafka, storm-hdfs, storm-hbase, etc.)**
- **Convenient support for multi-lang components**
- **External property substitution/filtering for easily switching between configurations/environments (similar to Maven-style ${variable.name} substitution)**

Examples

https://github.com/apache/storm/tree/master/external/flux/flux-examples

Hortonworks®

Security

- **Storm**
  - User Impersonation
  - SSL Support for Storm UI, Log Viewer, and DRPC (Distributed Remote Procedure Call)
  - Automatic credential renewal

- **Kafka**
  - Kerberos-based Authentication
  - Pluggable Authorization and Apache Ranger Integration

New In HDP Search

Hortonworks®

HDP Search 2.3

	HDP Search 2.2	HDP Search 2.3	
Package	jar	RPM	Solr, SiLK (Banana), Connectors all in one package
Solr	4.10.2	5.2.1	Latest stable release version of Solr (Included with package)
HDFS	2.5	2.7.1	Batch Indexing from HDFS (Included with package)
Hive	0.14.0	1.2.1	Batch indexing from Hive tables (Included with package)
Pig	0.14.0	0.15.0	Batch indexing from pig jobs (Included with package)
Storm	X	0.10.0	Streaming data real-time index (access from https://github.com/LucidWorks/storm-solr)
Spark Streaming	X	1.3.1	Streaming data real-time index (Included with package)
Security	X	Included in Solr 5.2.1	Kerberos and Ranger support (Included with Solr)
HBase	X	1.1.1	1. Near Real time indexing of data from HBase tables 2. Batch indexing from HBase tables (Included with package)
Ranger	X	0.5.0	Extend Ranger security configuration to HDP Search

HDP Search: Packaging and Access

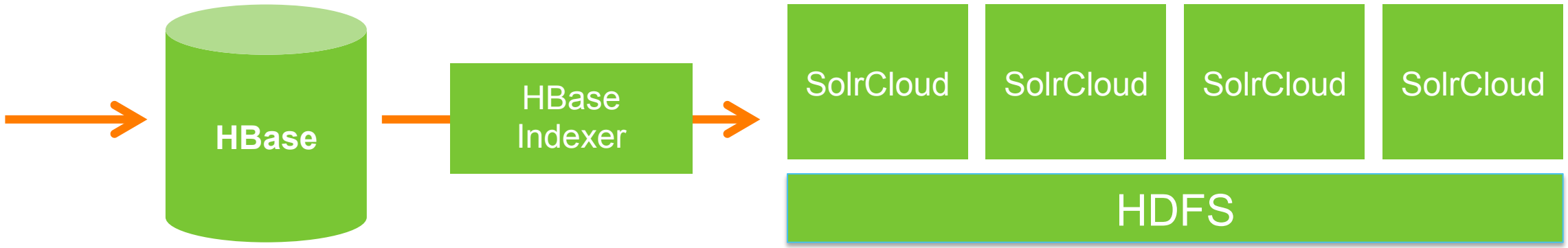
Available as RPM package

Downloadable from HDP-UTILS repo

yum install “lucidworks-hdp-search”

Hortonworks®

HBase Near Real Time Indexing into Solr



Indexer for table to collection

Asynch. replication from row update to document insert into index

Hortonworks®

Hbase Indexer

Hbase Realtime Indexer:

- The HBase Indexer provides the ability to stream events from HBase to Solr for near real time searching.
- HBase indexer is included with Lucidworks HDPSearch as an additional service
- The indexer works by acting as an HBase replication sink.
- As updates are written to HBase, the events are asynchronously replicated to the HBase Indexer processes, which in turn creates Solr documents and pushes them to Solr.

Bulk Indexing:

- Run a batch indexing job that will index data already contained within an HBase table.
- The batch indexing tool operates with the same indexing semantics as the near-real-time indexer, and it is run as a MapReduce job.
- The batch indexing can be run as multiple indexers that run over HBase regions and write data directly to Solr
- Indexing shards can be generated offline and then merged into a running SolrCloud cluster using the --go-live flag

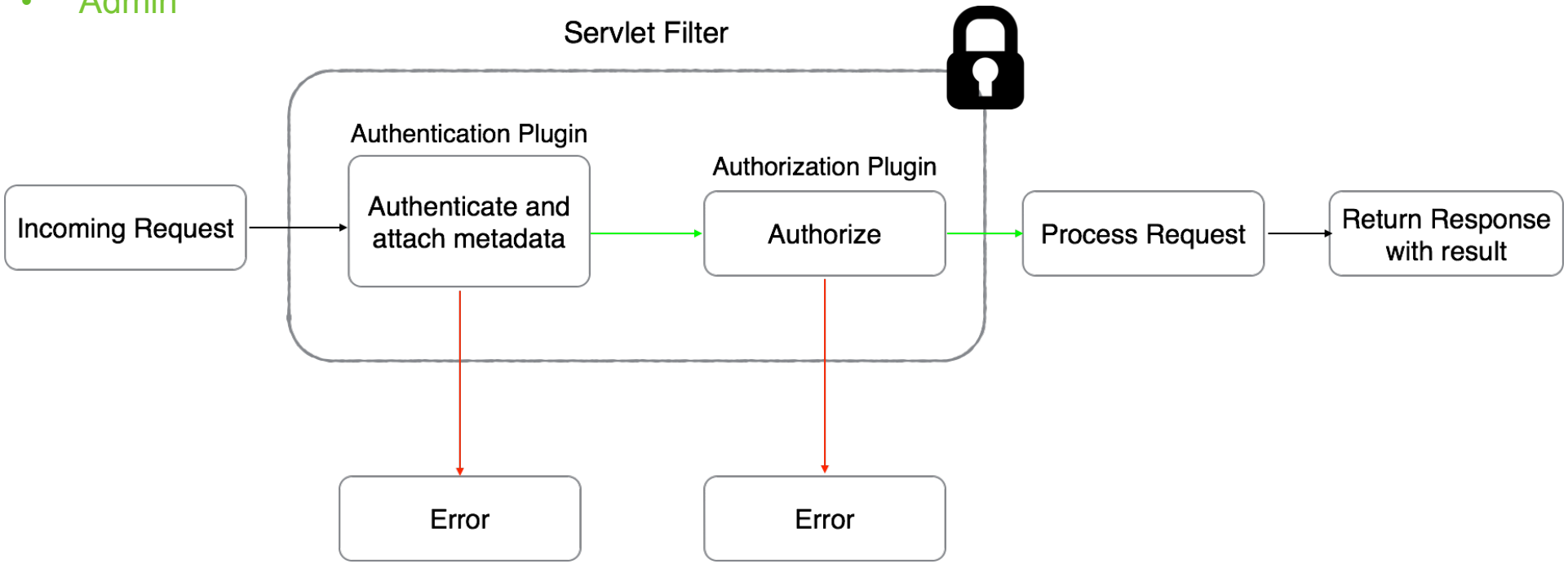
Thread is a parameter and can parallelize the indexing process

HDP Search Security

- Apache Solr supports authentication using Kerberos
- Apache Solr supports ACLs for authorization for a collection
- Following permissions are supported through Ranger, at a collection and core level
- Query
- Update
- Admin

Why is it important?

- Secure users using Solr
- Apply security policies for Solr Query
- Audit Solr Queries



SiLK: Visualize Bigdata Insights

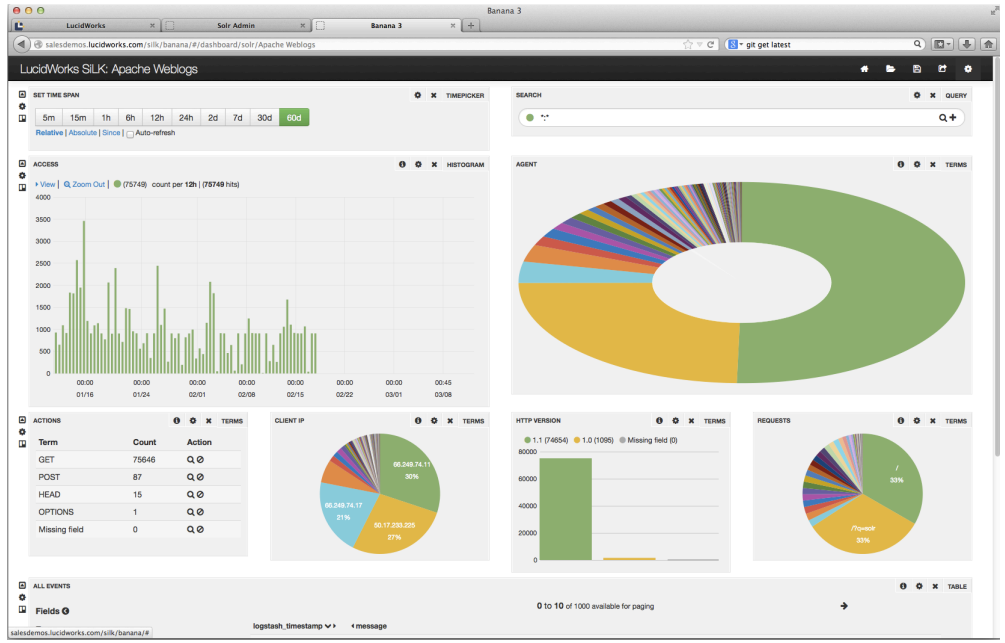
- **Bundled with HDP Search RPM package**

- **Real time interactive analytics**
  - Dashboards display real time users interaction
  - Integration will deliver pre-defined dashboards with most common analytics
  - Drill down into the analytics data all the way to a single event or user interaction
  - Create time-series to understand patterns and anomalies over time

- **Configure personalized dashboards**
  - Administration interface to build new dashboards with minimal effort
  - Create personalized dashboard views based on business unit or job role
  - Admin can setup dashboards per their business requirements to enable real time analysis of their products and user activity

- **Proactive alerts (Fusion only)**
  - Configure alerts to notify new events
  - Realtime proactive alerts help businesses react in real time

- **Security:**
  - No authentication or authorization support for SilK with HDP Search
  - Use Lucidworks Fusion to secure SilK as well



New In Apache Spark

Hortonworks®

Spark In HDP

Made for Data Science
All apps need to get predictive at scale and fine granularity

Democratizes Machine Learning
Spark is doing to ML on Hadoop what Hive did for SQL on Hadoop

Elegant Developer APIs
DataFrames, Machine Learning, and SQL

Realize Value of Data Operating System
A key tool in the Hadoop toolbox

Community
Broad developer, customer and partner interest

Hortonworks

HDP 2.3 Includes Spark 1.3.1

- **DataFrame API – (Alpha)**
  - SchemaRDD has become DataFrame API
- **New ML algorithms:**
  - LDA (Latent Dirichlet Allocation),
  - GMM (Gaussian Mixture Model)
  - & others
- **ML Pipeline API in PySpark**
- **Spark Streaming support for Direct Kafka API gives exactly-once delivery w/o WAL**
  - Python Kafka API

DataFrames: Represents Tabular Data

- RDD is a low level abstraction
- DataFrames attach schema to RDDs
- Allows us to perform aggressive query optimizations
- Brings the power of SQL to RDDs!

DataFrames are Intuitive

dept	name	age
Bio	H Smith	48
CS	A Turing	54
Bio	B Jones	43
Phys	E Witten	61

RDD

```
data.map(lambda x: (x.dept, [x.age, 1])) \
    .reduceByKey(lambda x, y: [x[0] + y[0], x[1] + y[1]]) \
    .map(lambda x: [x[0], x[1][0]/ x[1][1]]) \
    .collect()
```

DataFrame

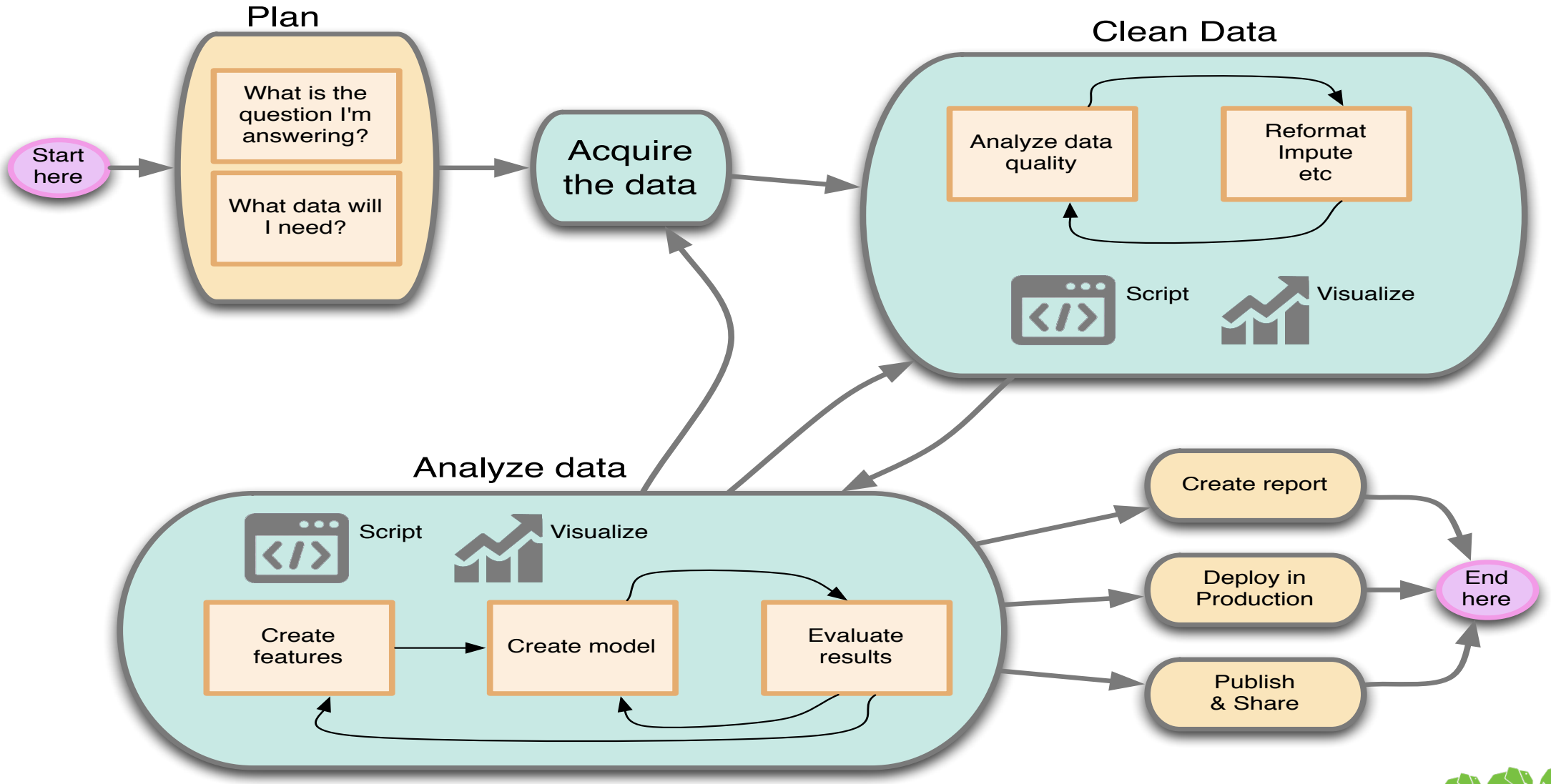
```
data.groupBy("dept").avg("age")
```


Hortonworks®

DataFrame

- Select, withColumn, filter etc.
- Explode
- groupBy
- Agg
- Join
- Window Functions

The Data Science Workflow Are Complex



ML Pipelines

Transformer

Transforms one dataset into another.

Estimator

Fits model to data.

Pipeline

Sequence of stages, consisting of estimators or transformers.

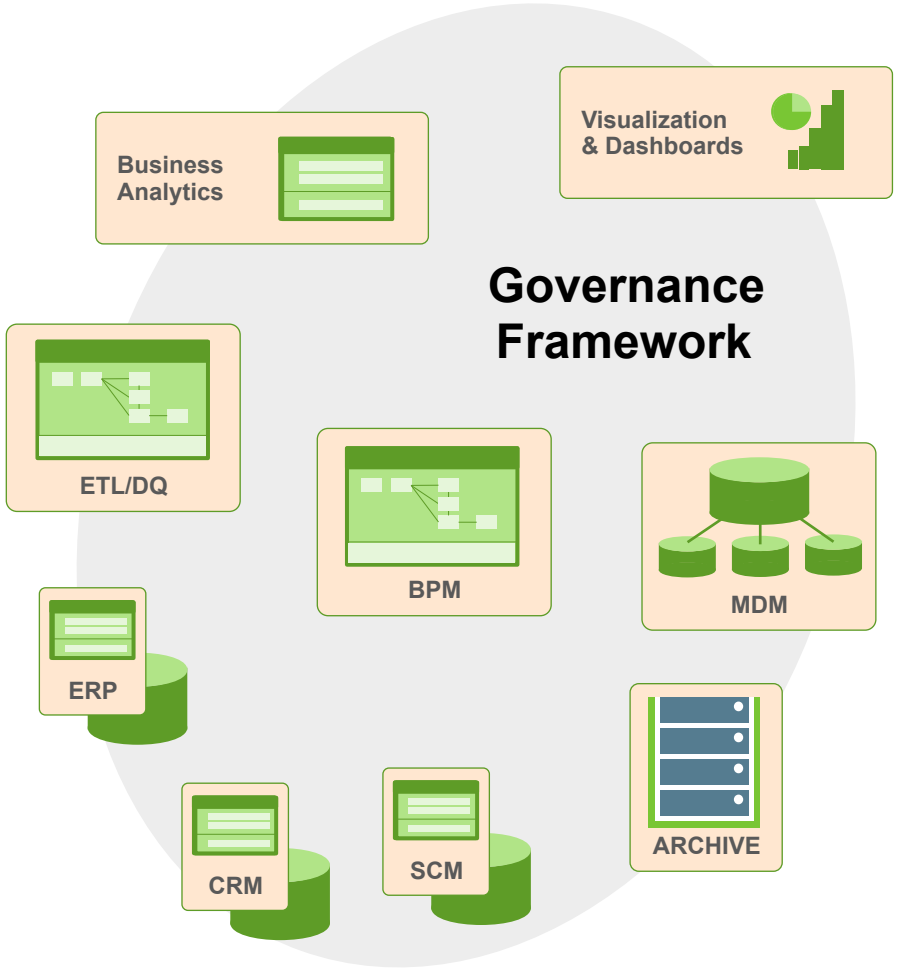
Tools for Data Science with Spark

- **DataFrame – intuitive manipulation of tabular data**
- **ML Pipeline API – construct ML workflows**
- **ML algorithms**
- **Notebooks (iPython, Zeppelin) – Data Exploration, Visualization, Code**

Apache Atlas

Hortonworks®

Enterprise Data Governance Goals

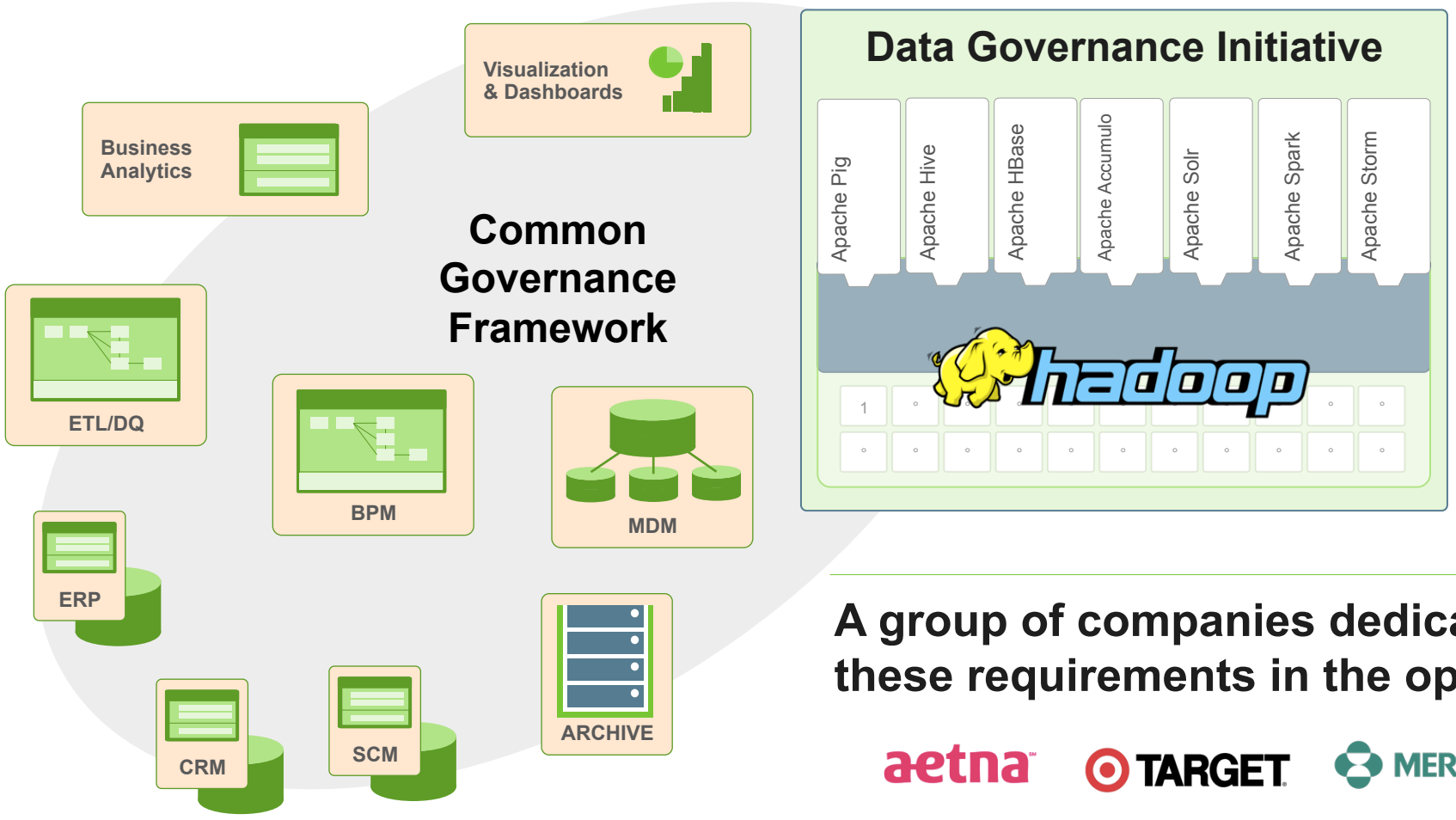


GOALS: Provide a common approach to data governance across all systems and data within the organization

- **Transparent**
  Governance standards & protocols must be clearly defined and available to all
- **Reproducible**
  Recreate the relevant data landscape at a point in time
- **Auditable**
  All relevant events and assets but be traceable with appropriate historical lineage
- **Consistent**
  Compliance practices must be consistent

Hortonworks®

DGI becomes Apache Atlas



TWO Requirements

1. Hadoop must snap in to the existing frameworks and be a ***good citizen***
2. Hadoop must also provide governance within its own stack of technologies

A group of companies dedicated to meeting these requirements in the open

aetna TARGET MERCK *Be well* Hortonworks

SAP *Major Bank* Schlumberger sas

Hadoop Data Governance for the Data Steward

Responsibilities include:

- Ensuring Data Integrity & Quality
- Creating Data Standards
- Ensure Data Lineage

Resolve issues before they occur

Scalable Metadata Service

Business modeling with industry-specific vocabulary

Extend visibility into HDFS path

REST API

Hive Integration

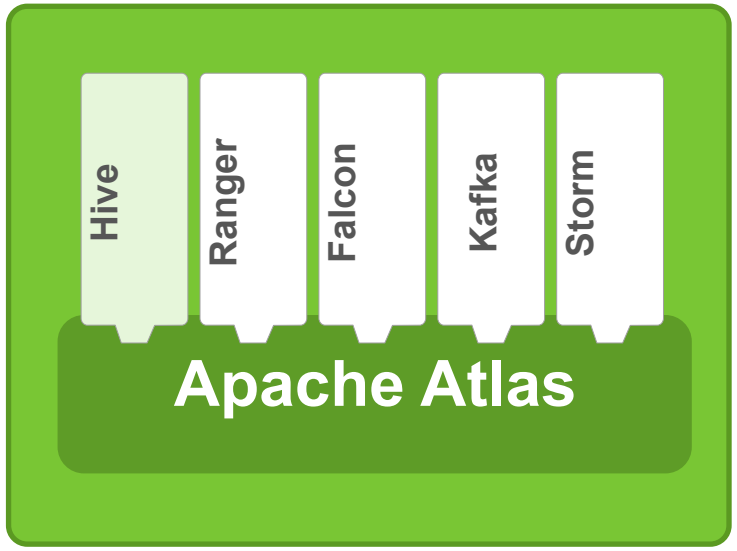
Leverage existing metadata with import/ export capability

Enhanced User Interface

Hive table lineage and Search DSL

Hortonworks®

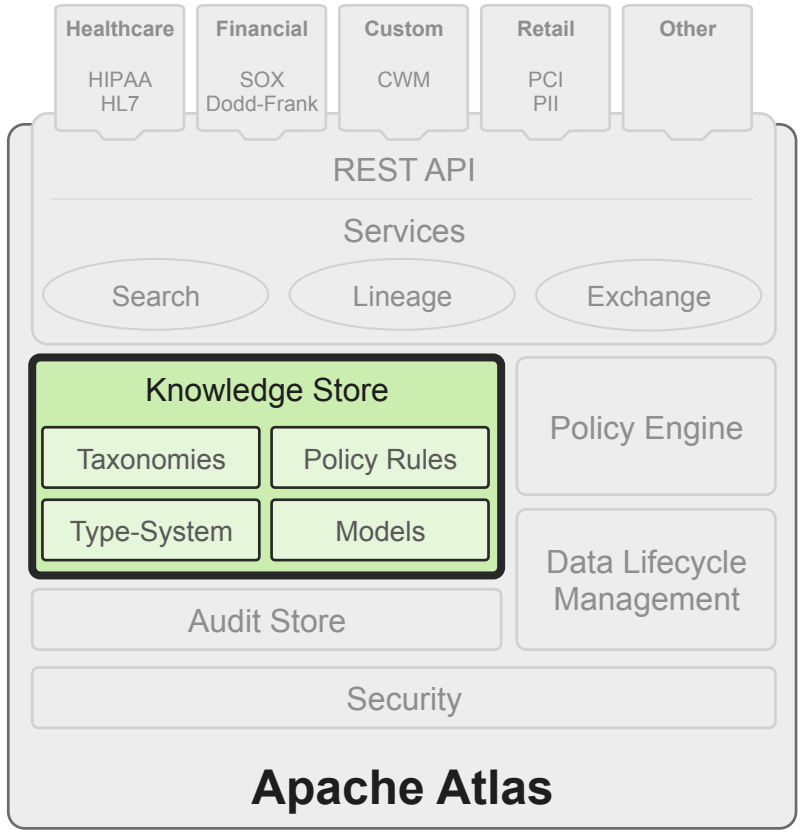
Apache Atlas



Metadata Services

- Business Taxonomy - classification
- Operational Data – Model for Hive: DB, Tables, Col,
- Centralized location for all metadata **inside HDP**
- Single Interface point for Metadata Exchange with platforms **outside of HDP.**
- Search & Prescriptive Lineage – Model and Audit

Apache Atlas Overview



Taxonomy

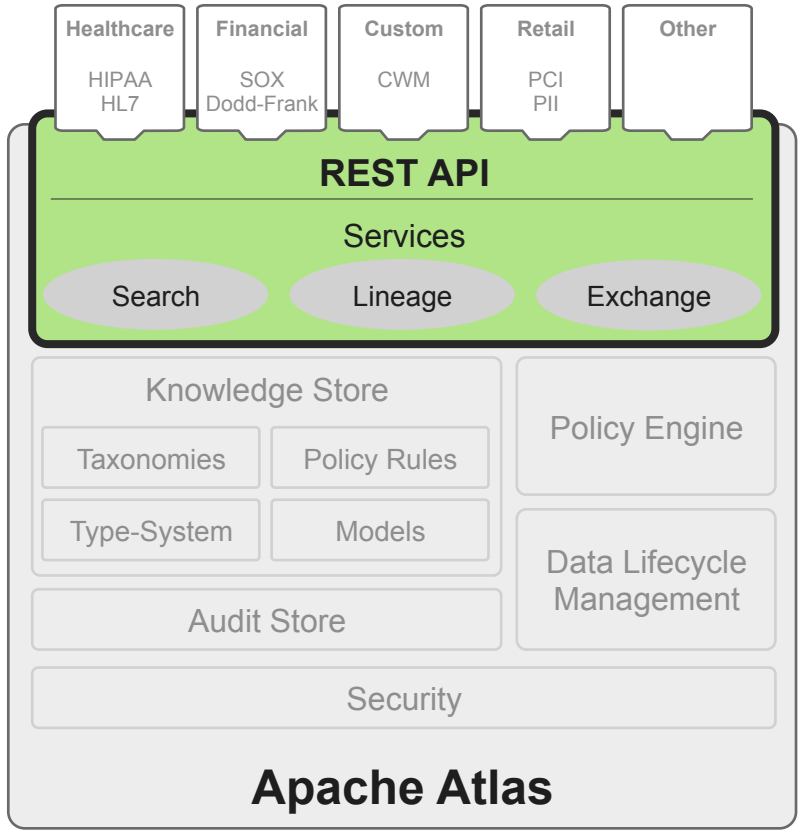
Knowledge store categorized with appropriate business-oriented taxonomy

- Data sets & objects
- Tables / Columns
- Logical context
- Source, destination

Support exchange of metadata between foundation components and third-party applications/governance tools

Leverages existing Hadoop metastores

Apache Atlas

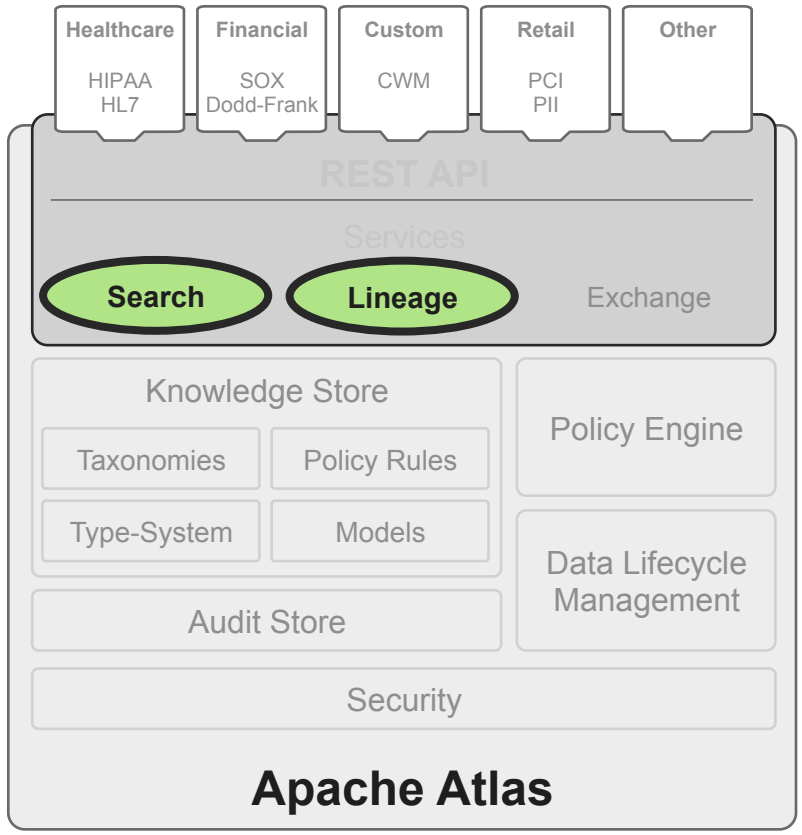


RESTful interface

- Extensible enterprise classification of data assets, relationships and policies organized in a meaningful way -- aligned to business organization.
- Supports exploration via user interface
- Supports extensibility via API and CLI exposure

Hortonworks®

Apache Atlas Overview



Search & Lineage (Browse)

- Pre-defined navigation paths to explore the data classification and audit information
- Text-based search features locates relevant data and audit event across Data Lake quickly and accurately
- Browse visualization of data set lineage allowing users to drill-down into operational, security, and provenance related information
- SQL like DSL – domain specific language

New In Apache Ambari 2.1

Hortonworks®

New in Ambari 2.1

- **Core Platform**
  - Guided Configs (AMBARI-9794)
  - Customizable Dashboards (AMBARI-9792)
  - Manual Kerberos Setup (AMBARI-9783)
  - Rack Awareness (AMBARI-6646)
- **Stack Support**
  - NFS Gateway, Atlas, Accumulo, others…
  - Storm Nimbus HA (AMBARI-10457)
  - Ranger HA (AMBARI-10281, AMBARI-10863)
- **User Views**
  - Hive, Pig, Files, Capacity Scheduler
- **Ambari Platform**
  - New OS: RHEL/CentOS 7 (AMBARI-9791)
  - New JDKs: Oracle 1.8 (AMBARI-9784)
- **Blueprints API**
  - Host Discovery (AMBARI-10750)
- **Views Framework**
  - Auto-Cluster Configuration (AMBARI-10306)
  - Auto-Create Instance (AMBARI-10424)

Ambari 2.1 HDP Stack Support Matrix

Support for HDP 2.3 and HDP 2.2

Deprecated Support for HDP 2.1 and HDP 2.0

- Plan to remove support for HDP 2.1 and HDP 2.0 in **NEXT** Ambari release

	HDP 2.3	HDP 2.2	HDP 2.1	HDP 2.0
Ambari 2.1	✓	✓	✓ deprecated	✓ deprecated
Ambari 2.0		✓	✓	✓
Ambari 1.7		✓	✓	✓

Hortonworks®

Ambari 2.1 HDP Stack Components

	HDP 2.3	HDP 2.2	HDP 2.1	HDP 2.0
HDFS, YARN, MapReduce, Hive, HBase, Pig, ZooKeeper, Oozie, Sqoop	✓	✓	✓	✓
Tez, Storm, Falcon, Flume	✓	✓	✓	
Knox, Slider, Kafka	✓	✓		
Ranger, Spark, Phoenix	✓	✓		
Accumulo, NFS Gateway, Mahout, DataFu, Atlas	**NEW! Ambari 2.1**			

Hortonworks®

Ambari 2.1 HDP Stack High Availability

	HDP Stack	Mode	Ambari 2.0	Ambari 2.1
HDFS: NameNode	HDP 2.0+	Active/ Standby	✔	✔
YARN: ResourceManager	HDP 2.1+	Active/ Standby	✔	✔
HBase: HBaseMaster	HDP 2.1+	Multi-master	✔	✔
Hive: HiveServer2	HDP 2.1+	Multi-instance	✔	✔
Hive: Hive Metastore	HDP 2.1+	Multi-instance	✔	✔
Hive: WebHCat Server	HDP 2.1+	Multi-instance		✔
Oozie: Oozie Server	HDP 2.1+	Multi-instance	✔	✔
Storm: Nimbus Server	HDP 2.3	Multi-instance		✔
Ranger: AdminServer	HDP 2.3	Multi-instance		✔

Ambari 2.1 JDK Support

	HDP 2.3	HDP 2.2	HDP 2.1	HDP 2.0
JDK 1.8	✓			
JDK 1.7	✓	✓	✓	✓
JDK 1.6 *		✓	✓	✓

Important: If you plan on installing HDP 2.2 or earlier with Ambari 2.1, be sure to use JDK 1.7.

Important: If you are using JDK 1.6, you **must** switch to JDK 1.7 **before** upgrading to Ambari 2.1

Ambari 2.1 Platform Support

- **Add RHEL/CentOS/Oracle Linux 7 support**
- **Removed RHEL/CentOS/Oracle Linux 5 support**
- **Ubuntu + Debian NOT AVAILABLE until first Ambari 2.1 and HDP 2.3 maint. releases!!!**

	RHEL 7	RHEL 6	RHEL 5	SLES 11	Ubuntu 12	Ubuntu 14	Debian 7
Ambari 2.1 M10	✓	✓		✓	✓	✓	✓
Ambari 2.1 GA	✓	✓		✓			
Ambari 2.0		✓	deprecated	✓	✓		

Ambari 2.1 Database Support

Ambari 2.1 + HDP 2.3 added support for Oracle 12c

Ambari 2.1 DB: SQL Server *Tech Preview*

```
Enter advanced database configuration [y/n] (n)? y
Configuring database...
========================================================================
Choose one of the following options:
[1] - PostgreSQL (Embedded)
[2] - Oracle
[3] - MySQL
[4] - PostgreSQL
[5] - Microsoft SQL Server (Tech Preview)
========================================================================
Enter choice (1):
```

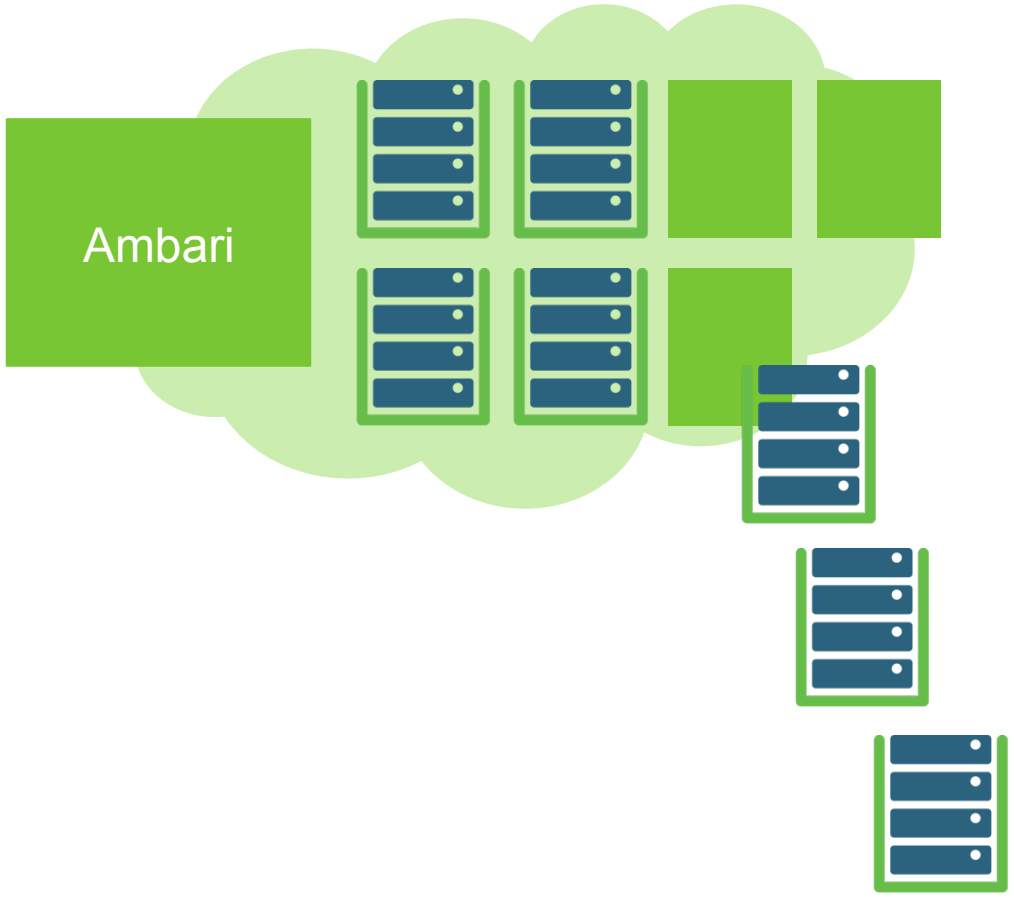

Hortonworks®

Blueprints Challenge Today

- **Today: Blueprints need ALL VMs available to provision cluster**
- This can be a challenge when trying to build a large cluster, especially in Cloud environments

- **Blueprints Host Discovery feature allows you to provision cluster with all, some or no hosts**
- When Hosts come online and Agents register with Ambari, Blueprints will automatically put the hosts into the cluster

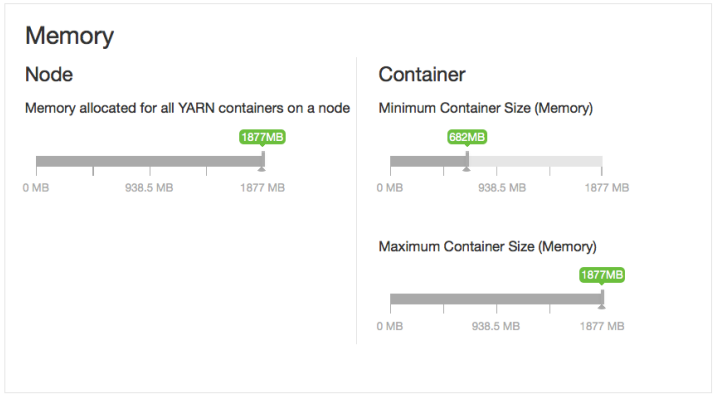
Blueprints Host Discovery (AMBARI-10750)

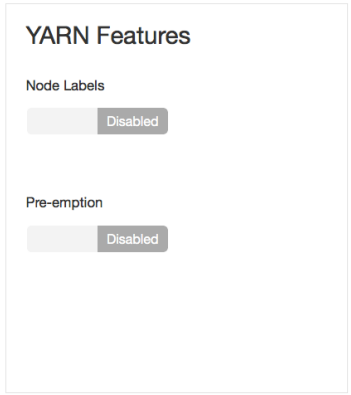


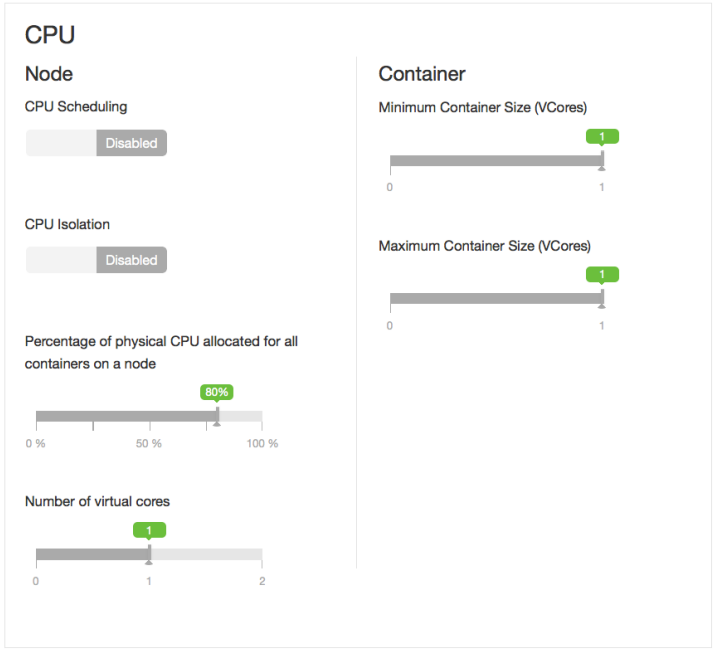

```
POST /api/v1/clusters/MyCluster/hosts

[
  {
    "blueprint" : "single-node-hdfs-test2",
    "host_groups" :[
      {
      "host_group" : "slave",
      "host_count" : 3,
      "host_predicate" : "Hosts/cpu_count>1”
      }
    ]
  }
]
```

Guided Configurations







- **Improved layout and grouping of configurations**
- **New UI controls to make it easier to set values**
- **Better recommendations and cross-service dependency checks**
- **Implemented for HDFS, YARN, HBase and Hive**
- **Driven by Stack definition**

Hortonworks®

Alert Changes

Alerts Log (AMBARI-10249)

- **Alert state changes are written to `/var/log/ambari-server/ambari-alerts.log`**

```
2015-07-13 14:58:03,744 [OK] [ZOOKEEPER] [zookeeper_server_process] (ZooKeeper Server
Process) TCP OK - 0.000s response on port 2181
2015-07-13 14:58:03,768 [OK] [HDFS] [datanode_process_percent] (Percent DataNodes Available)
affected: [0], total: [1]
```

Script-based Alert Notifications (AMBARI-9919)

- **Define a custom script-based notification dispatcher**
- **Executed on alert state changes**
- **Only available via API**

Hortonworks®

HDFS Topology Script + Host Mappings

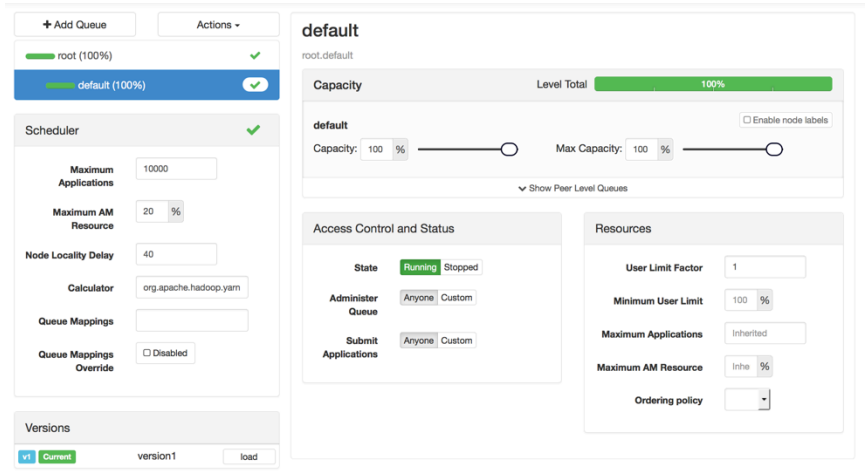
Set Rack ID from Ambari

Ambari generates + distributes topology script with mappings file

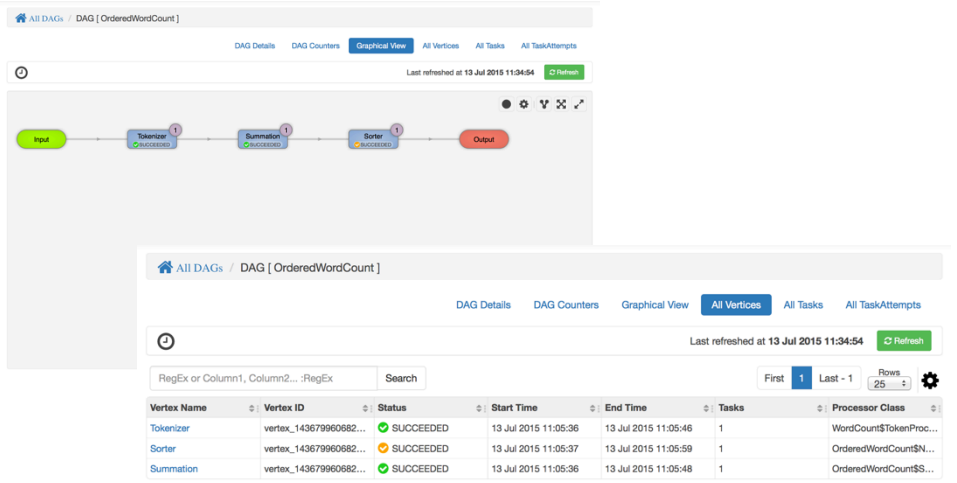
Sets core-site “net.topology.script.file.name” property

If you modify Rack ID HDFS, YARN

New User Views



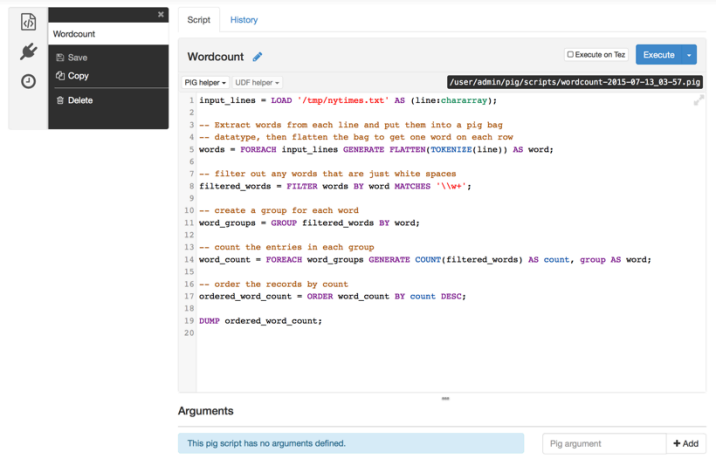
Capacity Scheduler View
Browse + manage YARN queues



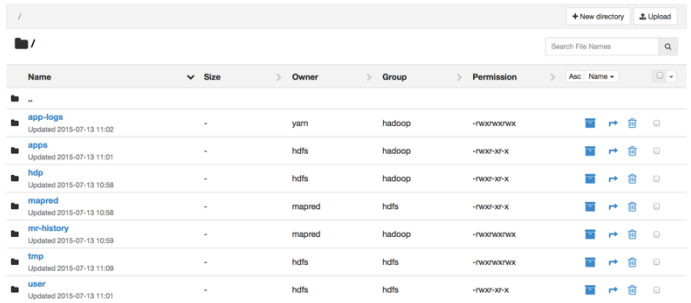
Tez View
View information related to Tez jobs that are executing on the cluster.

Hortonworks®

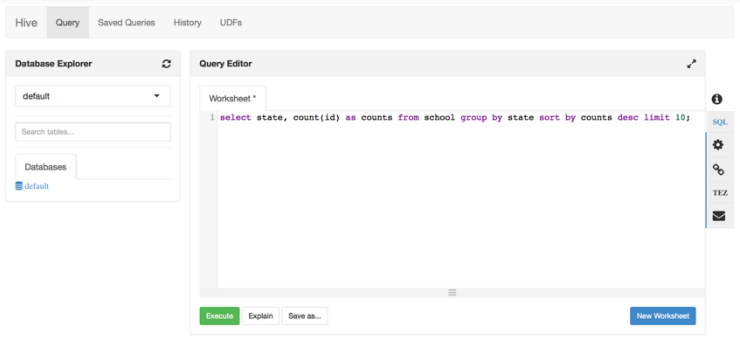
New User Views



Pig View
Author and execute Pig Scripts.



Files View
Browse HDFS file system.



Hive View
Author, execute and debug Hive queries.

Separate Ambari Servers

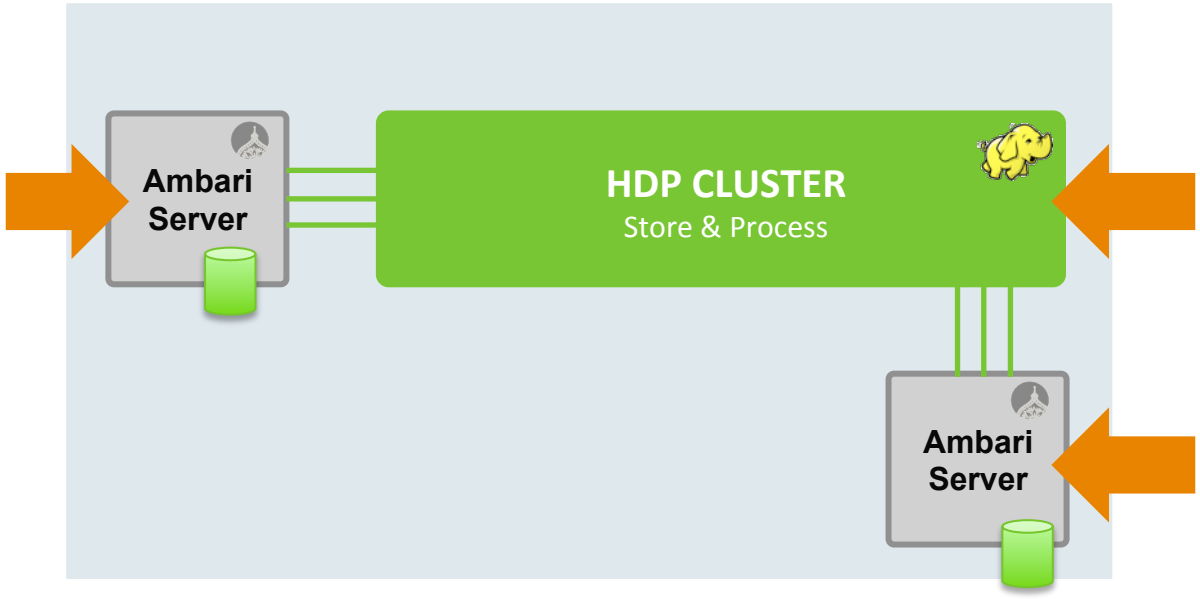
- **For Hadoop Operators:**

  Deploy Views in an Ambari Server that is managing a Hadoop cluster

- **For Data Workers:**

  Run Views in a “standalone” Ambari Server

Operators manage the cluster, **may have Views** deployed

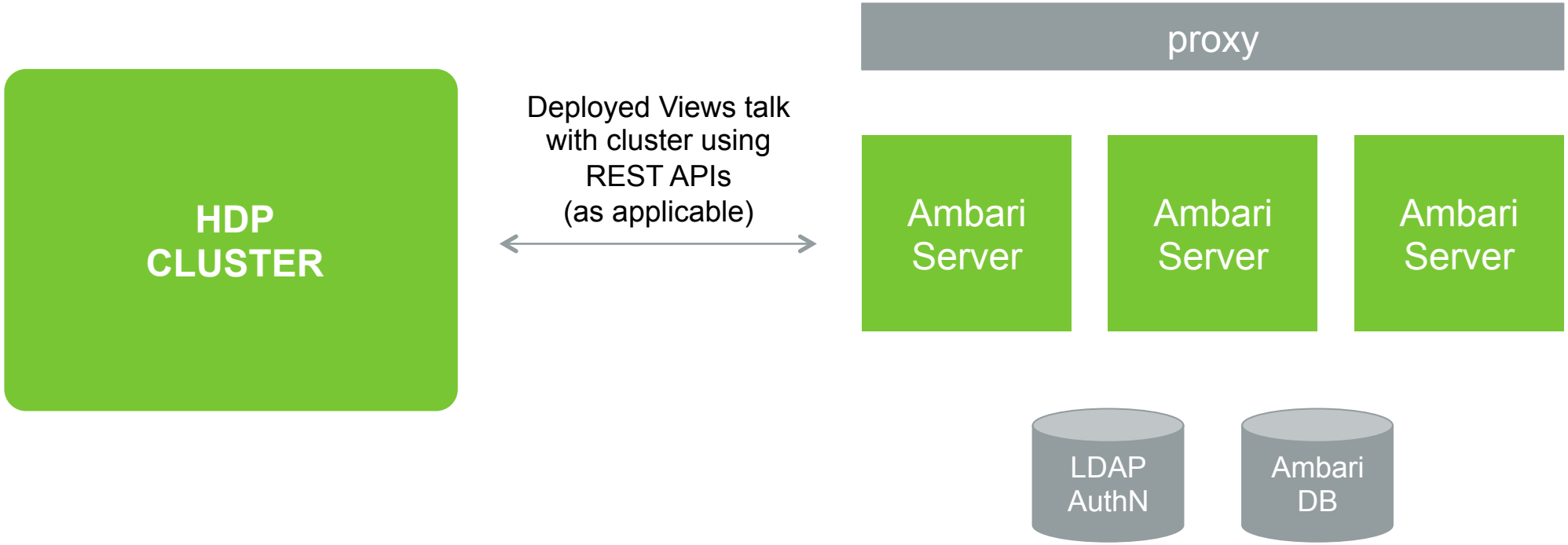


Data Workers use the cluster and use a “standalone” Ambari Server **for Views**

Hortonworks®

Views <-> Cluster Communications

Important: It is NOT a requirement to operate your cluster with Ambari to use Views with your cluster.



Upgrading Ambari 2.1

Preparing	Stop Ambari	Upgrade Ambari Server + Agents	Upgrade Ambari Schema	Complete + Start
Perform the preparation steps, which include making backups of critical cluster metadata.	On all hosts in the cluster, stop the Ambari Server and Ambari Agents.	On the host running Ambari Server, upgrade the Ambari Server. On all hosts in the cluster, upgrade the Ambari Agent.	On the host running Ambari Server, upgrade the Ambari Server database schema.	Complete any post-upgrade tasks (such as LDAP setup, database driver setup).

Ambari Upgrade Tips

- **After Ambari upgrade, you will see prompts to restart services. Because of all new guided configurations, Ambari has added the new configurations to Services.**
- Review the changes by comparing config versions.
- Use the config filter to identify any config issues.

- **Do not change to JDK 1.8 until you are running HDP 2.3.**
- HDP 2.3 is the ONLY version of HDP that is certified and supported with JDK 1.8.

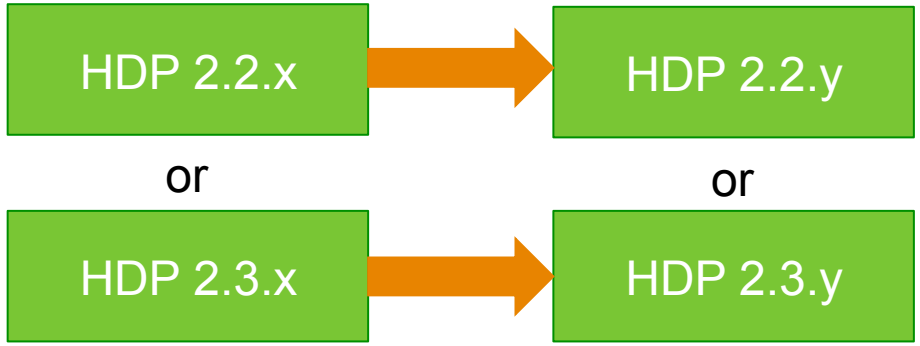
- **Before upgrading to HDP 2.3, you must upgrade to Ambari 2.1 first.**
- Be sure your cluster has landed on Ambari 2.1 cleanly and is working properly.
- Recommendation: schedule Ambari upgrade separate from HDP upgrade

HDP Upgrade Options

2.2 -> 2.3 MINOR UPGRADE

Rolling Upgrade
OR
Manual "Stop the World"

MAINTENANCE UPGRADE



Rolling Upgrade
OR
Manual "Stop the World"

2.0/2.1 -> 2.3 MINOR UPGRADE

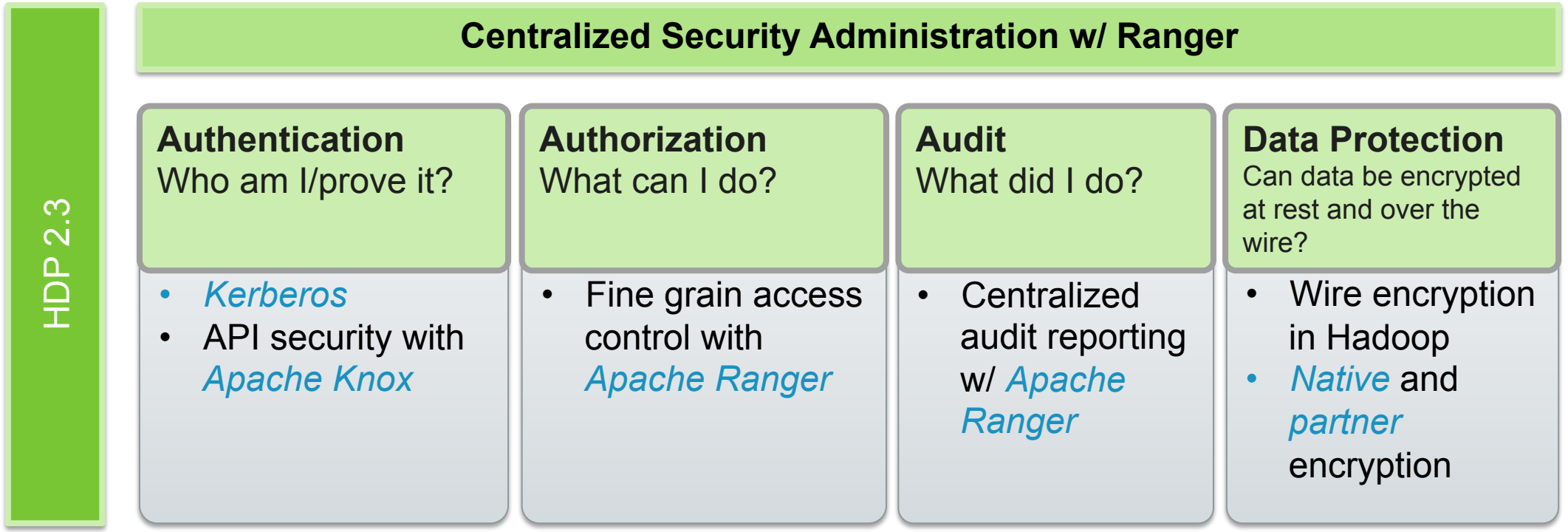
Manual "Stop the World"
(not available at GA)
(must go HDP 2.2 FIRST)

Hortonworks®

New In Apache Ranger

Hortonworks®

Security today in HDP



Security items planned in HDP 2.3

New Components Support

- Ranger to support authorization and auditing for Solr, Kafka and Yarn

Extending Security

- Hooks for creating dynamic policy conditions
- Protect metadata in Hive
- Introduce Ranger KMS to support HDFS Transparent Encryption
  - UI to manage policies for key management

Auditing changes

- Ranger to support queries for audit stored in HDFS using Solr
- Optimization of auditing at source

Security items planned in HDP 2.3

Extensible Architecture

- Pluggable architecture for Ranger – Ranger Stacks
- Config driven new components addition – Knox Stacks

Enterprise Readiness

- Knox to support LDAP caching
- Knox to support 2 way SSL queries
- Ranger to support PostGres and MS-SQL DB for storing policy data
- Ranger permission changes

Kafka Security

- **Kafka now supports authentication using Kerberos**
- **Kafka also supports ACLs for authorization for a topic per user/group**
- **Following permissions are supported through Ranger**
  - *Publish*
  - *Consume*
  - *Create*
  - *Delete*
  - *Configure*
  - *Describe*
  - *Replicate*
  - *Connect*

Solr Security

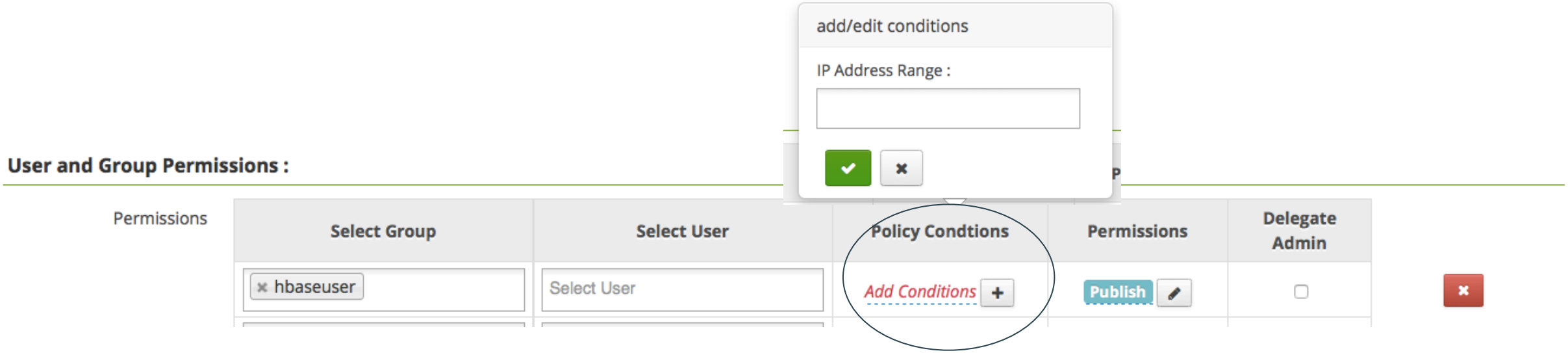
- **Apache Solr now supports authentication using Kerberos**
- **Apsche Solr also supports ACLs for authorization for a collection**
- **Following permissions are supported through Ranger, at a collection level**
  - Query
  - Update
  - Admin

Yarn Integration

- **Yarn supports ACL for queue submission**
- **Ranger now integrated with Yarn RM to manage these permissions from Ranger**
- **Following permissions are supported through Ranger**
- Submit-app
- Admin-queue

Dynamic Policy Conditions

- Currently Ranger supports static “role” based policy controls
- Users are looking for dynamic attributes such as geo, time and data attributes to drive policy decisions
- Ranger has introduced for hooks for these dynamic conditions



Dynamic Policy Hooks - Config

- Conditions can be added as part of service definition

```
99          "contextEnrichers":
100         [
101
102         ],
103
104         "policyConditions":
105         [
106                 {
107                         "id": 1,
108                         "name": "ip-range",
109                         "evaluator": "org.apache.ranger.plugin.conditionevaluator.RangerIpMatcher",
110                         "evaluatorOptions": { },
111                         "validationRegEx":"",
112                         "validationMessage": "",
113                         "uiHint":"",
114                         "label": "IP Address Range",
115                         "description": "IP Address Range"
116                 }
117         ]
118 }
```

Conditions can vary by service (HDFS, Hive etc)

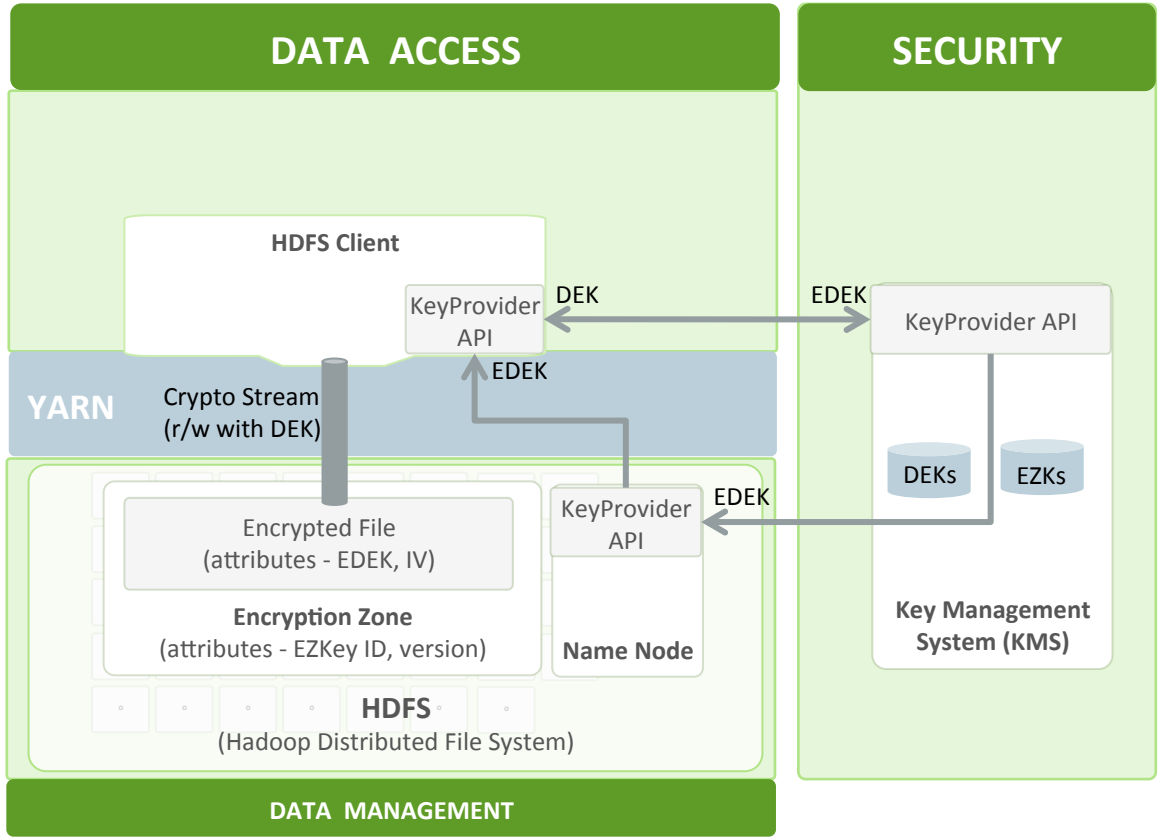
Hortonworks®

Protect Metadata in Hiveserver2

- In Hive, metadata listing can be protected by underlying permissions
- Following commands are protected
  - Show Databases
  - Show Tables
  - Describe table
  - Show Columns

HDFS Transparent Encryption

HDP 2.2



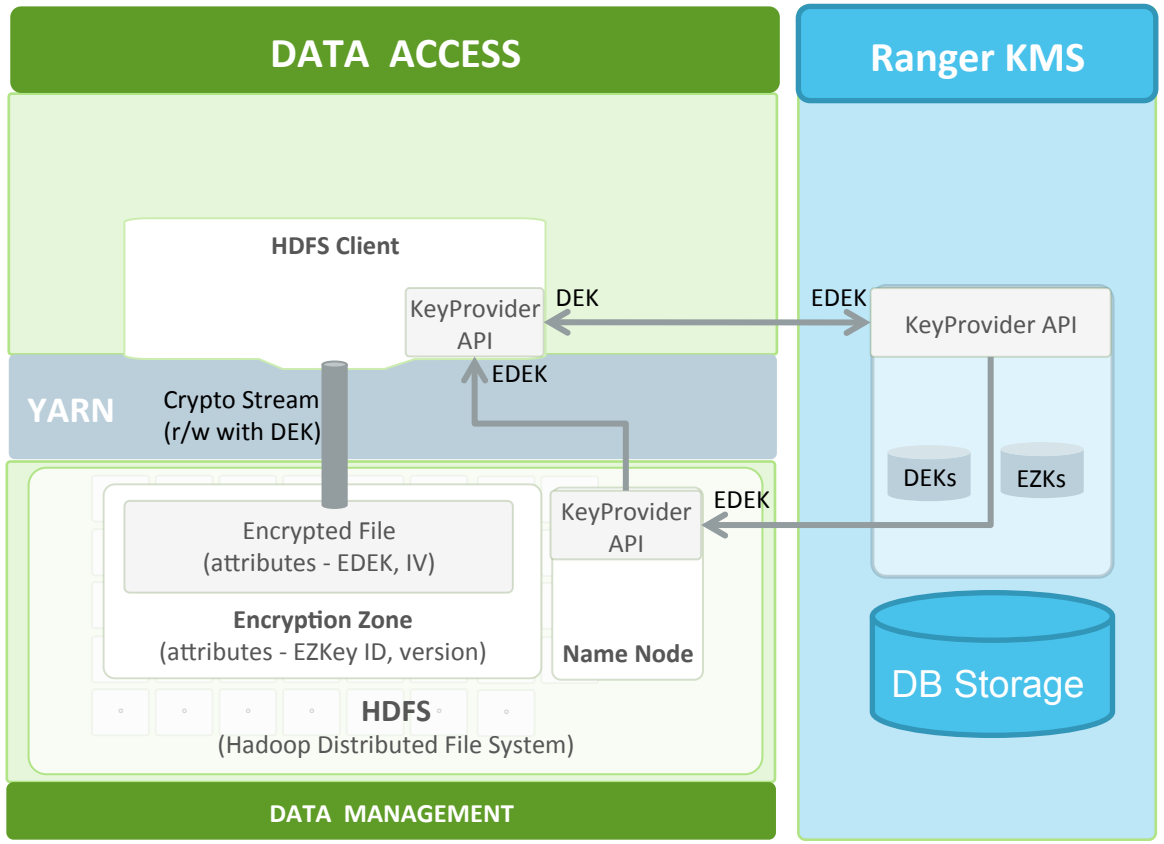
Acronym	Description
EZ	Encryption Zone (an HDFS directory)
EZK	Encryption Zone Key; master key associated with all files in an EZ
DEK	Data Encryption Key, unique key associated with each file. EZ Key used to generate DEK
EDEK	Encrypted DEK, Name Node only has access to encrypted DEK.
IV	Initialization Vector

Open source KMS based on file level storage.

Hortonworks®

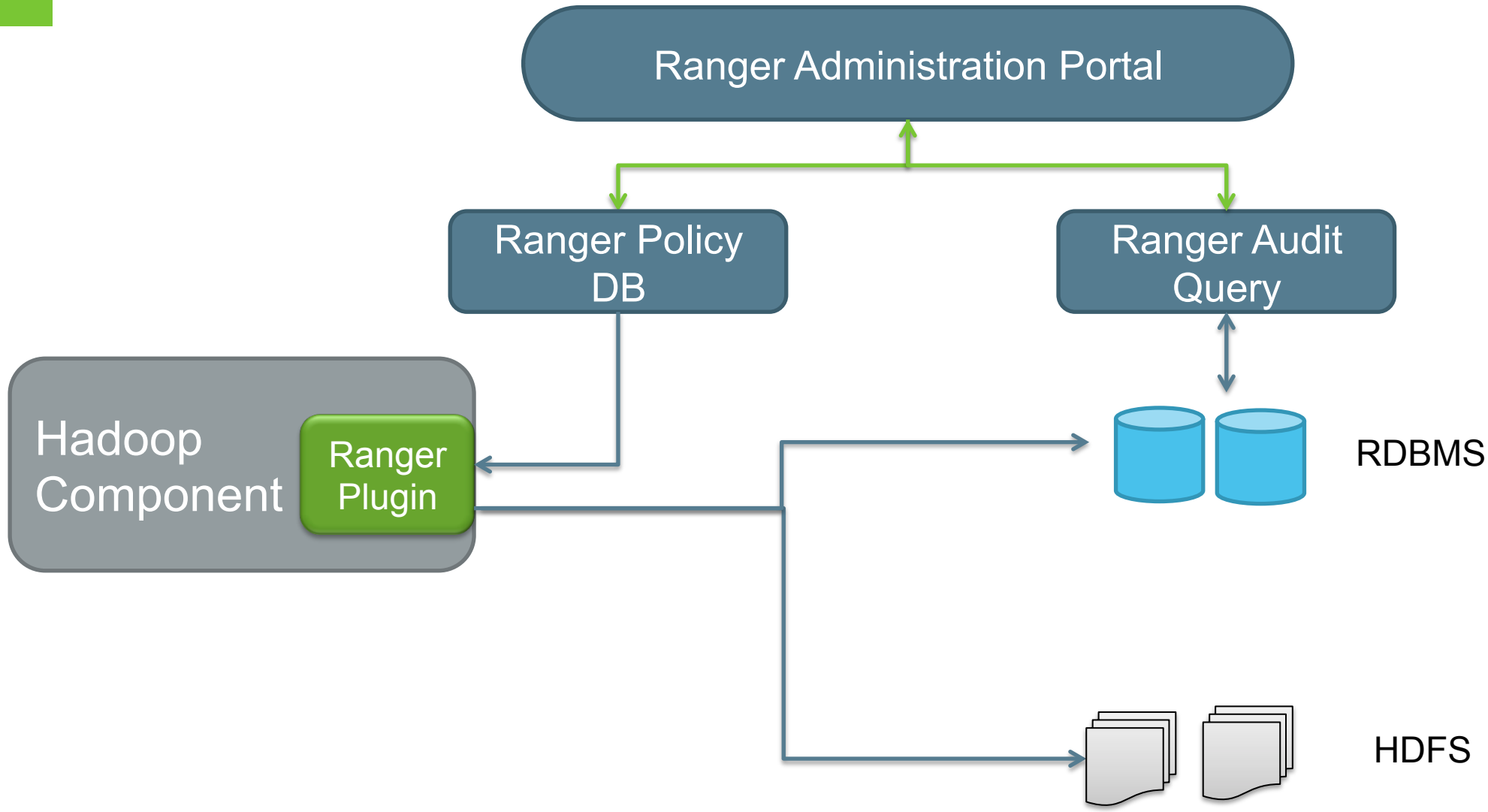
HDFS Encryption in HDP 2.3

HDP 2.3

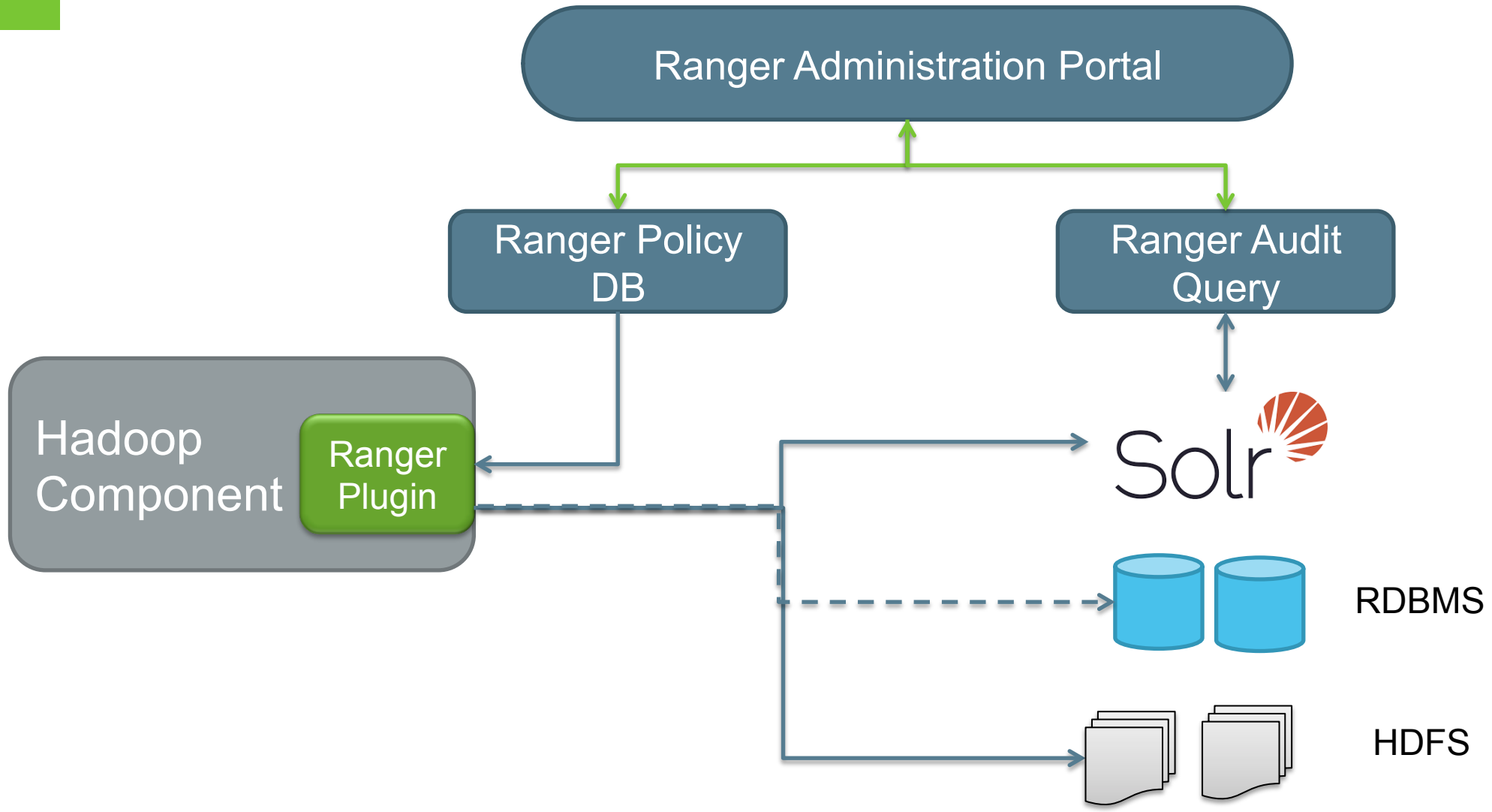


Hortonworks®

Audit setup in HDP 2.2 – Simplified View



Audit setup in HDP 2.3 – Solr Based Query



Why is it important?

- Scalable approach
- Remove dependency on DB for audit
- Ability to use banana for dashboards

Lab

https://github.com/abajwa-hw/hdp22-hive-streaming/blob/master/LAB-STEPS.md

Lab Overview

- **Tenants**
  - Groups - IT & Marketing
  - Users – it1 (IT) & mktg1 (Marketing)

- **Responsibility**
  - IT – Onboard Data & Manage Security
  - Marketing – Analyze Data

- **Lab Environment**
  - Using HDP 2.3 Sandbox
  - Linux and Ranger users it1 and mktg1 pre-created
  - Global Allow policy set in Ranger

Lab Steps

- **Step 1**
  - Create hdfs directories for users it1 and mktg1
- **Step 2**
  - Disable Ranger Global Allow Policy
  - Enable hdfs & hive permissions for it1
- **Step 3**
  - Create interactive and batch queues in YARN
  - Assign user it1 to batch queue and mktg1 to default queue
- **Step 4**
  - Create ambari users it1 and mktg1 and enable hive views
- **Step 5**
  - Load data at it1
- **Step 6**
  - Enable table access for mkt1
- **Step 7**
  - Query Data as mkt1
- --

Thank You

Hortonworks®

This presentation contains forward-looking statements involving risks and uncertainties. Such forward-looking statements in this presentation generally relate to future events, our ability to increase the number of support subscription customers, the growth in usage of the Hadoop framework, our ability to innovate and develop the various open source projects that will enhance the capabilities of the Hortonworks Data Platform, anticipated customer benefits and general business outlook. In some cases, you can identify forward-looking statements because they contain words such as “may,” “will,” “should,” “expects,” “plans,” “anticipates,” “could,” “intends,” “target,” “projects,” “contemplates,” “believes,” “estimates,” “predicts,” “potential” or “continue” or similar terms or expressions that concern our expectations, strategy, plans or intentions. You should not rely upon forward-looking statements as predictions of future events. We have based the forward-looking statements contained in this presentation primarily on our current expectations and projections about future events and trends that we believe may affect our business, financial condition and prospects. We cannot assure you that the results, events and circumstances reflected in the forward-looking statements will be achieved or occur, and actual results, events, or circumstances could differ materially from those described in the forward-looking statements.

The forward-looking statements made in this prospectus relate only to events as of the date on which the statements are made and we undertake no obligation to update any of the information in this presentation.

Trademarks

Hortonworks is a trademark of Hortonworks, Inc. in the United States and other jurisdictions. Other names used herein may be trademarks of their respective owners.

Hortonworks®