Hortonworks Response to State Farm NGIS

February 27, 2015

Table of Contents

1. Executive Summary

Hortonworks is honored and pleased to respond to State Farm's Request for Proposal for the Next Generation Images Services Platform. As you proceed forward with reviewing Hortonworks response we are confident your findings will result in Hortonworks being best suited and equipped to deliver on your objectives of the Hadoop solution you require for your Next Generation Image Services Platform. The executive summary will outline the key elements of our proposal including the primary features included in the solution design for NGIS, the key factors associated with the value Hortonworks offers as a result of building upon our existing relationship to delivery on your requirements as well as a vision and strategy for insuring State Farm's long term success with Hadoop.

To begin with, Hortonworks approach for State Farm begins with establishing a 100% open source based platform based on proven technology that is fault tolerant, performant, and highly scalable and a solution design that has been battle tested for real world use and adoption by some of the largest firms in the world who not only require the scale State Farm is requesting of their NGIS platform but also at a larger scale. Therefore, State Farm stands to benefit from not only the deep level collaboration that has already taken place amongst the Apache open source community as well as that which has occurred from a co-engineering level with experienced practitioners of Hadoop such as AT&T, Verizon, Bloomberg and FINRA that we list as references in the RFP response.

As noted above, this is one of Hortonworks key differentiators. It is based on our commitment and strategy to constantly drive innovation upstream through the Apache Software Foundation by gifting our technology back to ASF. We believe this approach has enabled Hadoop to become enterprise consumable rather than a nifty solution for the large web properties that first adopted Hadoop. Based on this commitment, our customers have benefited from the center of gravity that has been created by Hortonworks to work closely with our partners to co-engineer technology that is constantly being gifted back to the Apache Software Foundation to be made available to the community. These partners include but are not limited to Microsoft, SAP, HP, Pivotal, SAS and RedHat but also the companies as mentioned above who are long-term practitioners of Hadoop and have established co-engineering relationships with Hortonworks and who depend on Hortonworks to steward these innovations back through the Apache Software community process. The end result is technology State Farm can consume without being locked in that future proofs your investment in Hadoop.

As we spend the majority of time in our response detailing our solution design and approach for meeting your requirements for State Farm's NGIS platform it is also important to discuss setting the foundation. The team associated with this effort has also participated in the Future State Architecture discusses had by Hortonworks and State Farm. It is important to briefly sight this in the executive summary because Hortonworks is proposing in our solution design an architecture that completely aligns with our consultation on State Farm's future state architecture. Our approach carries with it the foundation to not only address the requirements

of NGIS but also a reference architecture that can be re-used and re-purposed for other use cases that form at State Farm. The results of the architecture that is implemented and with that the best practices the Hortonworks team delivers by implementing the solution proposed will have long term use for State Farm whether you decide to simply expand this environment for use cases associated with Image Services or you decide to add additional use cases that take advantage of the multi-tenant capabilities of the platform.

Therefore, it is important to revisit the definition of the data lake and how this sets the foundation for the solution that Hortonworks has architected for Image Services.

Data Lake Defined

The shortest definition of a data lake is commodity scaled-out storage with collocated compute. However, a data lake is far more than this. Data lakes are a living, breathing record of the business. They retain all data from all source systems. They are a point of consolidation and collaboration. All teams and all silos store data in the lake together. They are a consistent place to secure data. Data can be encrypted at rest and in motion, in a unique way per dataset and then data can be shared in both encrypted and decrypted form based on user access levels. And data lakes are a place data is made high quality and reliable for all to consume in an automated fashion. Data lakes should come with all the tools to label data as to schema and to send data through processing pipelines that log relationships amongst datasets and health of execution of those pipelines through all stages of transformation and derivation. Data lakes are a gather place where the enterprise comes to meet and work together on data.

Foremost, however, data lakes are not just secured multi-tenant storage but are, in fact, a compute cloud or farm where arbitrary applications can run on data without moving that data out of the lake. It is also important to understand that without multi-tenancy support from the very foundations of the lake, the data lake will fail. Multi-tenancy is defined in two ways:

- Security
- Resource fencing

A secure data lake is one that allows all its tenants to bring data to the infrastructure without risk of data leakage or loss. Data leakage occurs when members of one team or users with inappropriate credentials end up reading data belonging to mission critical or even regulated use cases and users. The aspirational architecture must be secure.

Resource fencing ensures that noisy users on the system do not cause disturbances to other users of the same-shared system. Teams and use cases that are used to a silo-based experience can still feel that level of isolation and throughput guarantee.

Our data lake must be a secure, isolatable environment so that we can ask use cases to pile on without fear of having a degraded user experience relative to their current solution. Only when

users and data come to the lake can we begin to both tear down the walls of isolation where appropriate opportunities arise as well as begin to analyze super or über use cases where cross-dataset analysis may reap new and tremendous predictive value to the firm.

Hortonworks Data Platform as Data Lake

The Hortonworks Data Platform has the capabilities the Data Lake calls for. Specifically, our 5-pillars approach to data management drives the ability to meet the needs of the Data Lake. Those pillars ensure it has solutions around:

- **Apache YARN** as the core resource negotiator. HDFS as the tiered and commodity-based storage solution inside core **Apache Hadoop.**
- Data governance – via Apache **Falcon** – that provides for automation of data pipelines as well as an operational record of those pipeline flows that serves to answer questions around lineage, audit, and more.
- Security – via Apache **Ranger** and Apache **Knox** – that provides access controls, encryption hooks, and perimeter security designed to ensure many teams can share physical storage and compute without seeing each other's data. Further, complete audit trails are recorded giving State Farm the regulatory and consumer-focused protections it demands to ensure privacy and accountability are in place.
- A metadata solution—known as **HCatalog**—that provides schema versioning, schema on read support, as well as an external interface that can be used to both populate and look up data defined and stored in the lake.
- An operations management solution known as Apache **Ambari** that allows State Farm to manage the commodity server-based super computer as one logical entity instead of a loosely couple set of Linux or Windows hosts.

These five pillars of Hadoop enable the key functional interactions that occur in the data lake: Data Movement, Data Management and Data Access.

With the Hortonworks design for State Farm's NGIS platform we are laying down foundation of the core tenants of the data lake by leveraging a 100% open source platform that is endorsed, supported and widely adopted by the Hadoop community. With the platform in place, it's a matter of leveraging the capabilities in the platform for this particular use case. Let's then discuss at a high level the key elements of the solution design.

Everything we do at Hortonworks is to enable a centralized architecture powered by YARN, where YARN becomes the core data operating system for your cluster. Hortonworks' leadership in the community and our aggressive roadmap reflects the commitment to bring all scale-out technologies (both Open-Source and Commercial) to run on YARN. This is the vision we lead by,

where all data technologies including but not limited to ETL, BI, Analytics, Databases, Data-Processing Frameworks that certify on YARN will be well managed services on the cluster from resource (CPU, Memory, Disk, IO, GPU), governance, security and operations standpoint. Following are the key differentiators Hortonworks brings in the context of the solution we propose:

- Hortonworks is heavily invested in State Farm's success. It is as important for Hortonworks as it is for State Farm to see this project come to life. To value our partnership, Hortonworks engineering has offered to provision a 20 node Test Cluster internally to simulate the State Farm workload for this use case. In return, State Farm will benefit from personalized feature enhancements and configurations tailored to unique tuning needs of this project. This is to ensure that State Farm is getting nothing less but the very best from Hortonworks Data Platform.

- Hortonworks ships with Apache Ambari, which provides a powerful feature called Blueprints. With blueprints, you can effectively snapshot the entire deployment configuration of your cluster. This template then can be used to provision a replica of your existing cluster in different environments (Cloud, DR datacenter, etc.) in an automated fashion.

- Hortonworks supports Kafka-Mirroring feature, which will allow you to replicate critical data assets between datacenters in asynchronous fashion.

- Hortonworks packages Apache Phoenix as part of our distribution. HBase is key part of the overall architecture we propose to store Images and Audits. Apache Phoenix provides SQL semantics for HBase. This allows you to run analytics directly on HBase without needing to move data elsewhere.

- Apache Solr is a core component of our design that provides blazing fast search capability at scale. Hortonworks partners with Lucidworks, which is the commercial entity behind Apache Solr and also the leader in search space. Through Lucidworks, Hortonworks provides support for the latest version of Enterprise Solr - 4.10.2 on YARN.

- Hortonworks packages Apache Storm as a key part of the solution design. State Farm's SLA for this use case dictates the need for a sub-second event processing system. Apache Storm is the de-facto industry standard when it comes to real-time data computation at sub-second speeds.

- Finally, a key ingredient that will insure State Farm's success on this project are the resources that Hortonworks will bring to bare for State Farm. You have experienced our engagement model for our strategic customers. The resources we engage on your behalf

permeates through our entire organization and the partners we bring to the table. This is a key differentiator for us. Based on the nature of our business being a support and services model, we believe it is our duty and responsibility to engage our experts at every stage of the project. This means establishing executive sponsorship at the business level, which Herb Cunitz, our President and COO, gladly fulfills. Developing a formidable engineering relationship starting with Ari Zilka, our CTO as your technical executive sponsor. Ari is one of the preeminent thought leaders and practitioners in the big data industry. He has not only provided thought leadership to State Farm, he has also authored white papers on State Farm's behalf as well as worked with our solutions engineering and professional services teams as well as with State Farm to insure our solution designs meet your requirements. State Farm will continue to see Ari Zilka engaged in the success of this project as well as with other projects we are fortunate to embark on with State Farm. In addition, our Solutions Engineering team lead by George Vetticaden and your solutions engineer Rohit Pujari have spent a significant amount of time with your respective teams providing education and architecture guidance based on your requirements to vet out the solution design and architecture. That knowledge exchange has been shared with our professional services organization to scope out the level of effort and resources required. Our implementation services approach to your success is reflected in detail with our response. Our engagement model also includes providing you direct access to the actual committers responsible for leading the Apache projects. These personnel include but are not limited to the following Hortonworks personnel:

- Arun Murthy, the architect of Apache YARN and co-founder of Hortonworks.
- Deveraj Das, one of the preeminent experts in the industry and committer for HBase.
- Owen O'Malley, the architect behind Kerberos and security in Hadoop, Apache committer and Hortonworks founder.
- Sanjay Radia, the architect of HDFS, committer and co-founder of Hortonworks
- The key committers of Lucene with our partner Lucidworks.

With that in mind, Hortonworks is prepared to field a project team that represents the very best Hadoop consultants in the industry to insure State Farm's success with the NGIS platform.

Once again, we are honored to be participating in this Request for Proposal for State Farm's Next Generation Image Services platform. We are prepared to continue to provide the very best Hadoop experts in the industry to insure State Farm's decision to move forward with Hadoop as your new data platform for NGIS results in significant return on your investment that is demonstrative of the value this platform will bring to State Farm as a whole.

We look forward to reviewing our response in more detail and addressing any questions you have. Thank you very much for your consideration and continued support of Hortonworks as your Hadoop partner.

Best regards,

Phil Zacharia
Strategic Account Executive

2. Company Profile

D.1 Company Background

Hortonworks® (NASDAQ: HDP) is the leading contributor to and provider of Apache™ Hadoop® for the enterprise, and our mission is to establish Hadoop as the foundational technology of the modern enterprise data architecture. Our solution, the Hortonworks Data Platform (HDP), is an enterprise-grade data management platform that enables a centralized architecture for running batch, interactive and real-time applications simultaneously across a shared dataset. HDP is built on Apache Hadoop, powered by YARN, and supported by a comprehensive set of capabilities that address the core requirements of security, operations and data governance.

Key Hortonworks Facts

Year Founded	In 2011, 24 engineers from the original Hadoop team at Yahoo! spun out to form Hortonworks.
Ticker Symbol	NASDAQ: HDP More info at
Headquarters	Santa Clara, CA
Business Model	Open Source Software Support Subscriptions, Training and Consulting Services
Non-GAAP Billings	Grew from zero to over $125 million on an annualized basis in 10 quarters
Subscription Customers	332 in 10 quarters with 99 added in Q4-2014 alone.
Support	24×7, global web, telephone support
Partners	1000 joint engineering, strategic reseller, technology, and system integrator partners
Employees	As of December 31, 2014, we had 601 full-time employees, including 528 employees in the United States and 73 employees internationally. Our turnover rate is less than 10% over the past year.
Global Operations	17 countries

More public company information can be found on our Investor Relations page at:
http://investors.hortonworks.com/phoenix.zhtml?c=253804&p=irol-irhome

Hortonworks Product Offerings

We generate revenue by selling support subscription offerings and professional services for the Hortonworks Data Platform (HDP).

Support Subscriptions

We provide support under annual or multi-year subscriptions. A support subscription generally entitles a support subscription customer to a specified scope of support, as well as security updates, fixes, functionality enhancements and upgrades to the technology and new versions of the software, if and when available, and compatibility with an ecosystem of certified hardware and software applications.

Support subscription offerings for HDP are designed to support our support subscription customers throughout the entire lifecycle: from development and proof-of-concept, to quality assurance and testing, to production and deployment, and are available in two editions: HDP Enterprise and HDP Enterprise Plus. Both offerings provide support incidents with up to 24x7, one-hour response available from Hortonworks and selected independent software vendor and original equipment manufacturer, or OEM, partners. Support services include but are not limited to remote troubleshooting, advanced knowledgebase, access to upgrades, updates and patches, diagnosis of installation and configuration issues, diagnosis of cluster management and performance issues, diagnosis of data loading, processing and query issues, as well as application development advice.

Professional Services

We offer a range of professional services that are designed to help our customers derive additional value from deploying Hortonworks Data Platform.

- Training. We provide scenario-based Enterprise Hadoop training classes for developers, system administrators and data analysts available in classroom, corporate on-site and online settings, along with examinations that enable individuals to establish themselves as Certified Hadoop Professionals. Our training classes help populate customers with skilled Hadoop professionals who often serve as internal experts and open source advocates, increasing opportunities for successful adoption and use of the Hortonworks Data Platform.

- Consulting. We also provide the services of experienced consultants principally in connection with our technology offerings to assist with the needs of our customers such as deployment assessments, implementations, upgrade planning, platform migrations, and solution integration and application development. By providing consulting services, directly and with our certified system integrator partners, we facilitate adoption of HDP.

We Work within the Community for the Enterprise

Our founding belief is that innovation and adoption of platform technologies like Hadoop is best accomplished through collaborative open source development under the governance

model of the Apache Software Foundation (ASF). Done right, open source helps create an equitable balance of power and a fair exchange of value between vendor and consumer.

Key Community Stats

#1 Apache Hadoop committers	With 28 out of 86 Apache Hadoop committers, Hortonworks employs the largest group of committers under one roof; more than double any other company.
#1 Apache committer seats across the 20+ projects within HDP	With 170 committer seats across the 20+ Apache projects within HDP, Hortonworks employs the largest group of Apache committers focused on the data access, security, operations, and governance needs of the enterprise Hadoop market; more than double any other company.

Source: Committer counts above are as of February 24, 2015 and are obtained from the various Apache project team pages such as http://hadoop.apache.org/who.html

Our Joint Engineering Leadership Accelerates Customer Value

In order to enable a data platform like Hadoop to be easy to use and enterprise-grade, you don't go it alone. Hortonworks has established deep joint engineering relationships with most of the strategic data center players our customers rely on including Microsoft, HP, SAS, SAP, Teradata, EMC, Pivotal, Hitachi Data Systems, Red Hat, Oracle Informatica, and others.

Each of these joint engineering efforts are run through a gauntlet of thousands of certification tests unique to Hortonworks, enabling our joint customers to receive the highest quality, integrated solutions for addressing their modern data architecture needs.

Our World-class Enterprise Open Source Business and Technology Team

In 2011, Rob Bearden partnered with Yahoo! to establish Hortonworks with 24 engineers from the original Hadoop team including founders Alan Gates, Arun Murthy, Devaraj Das, Mahadev Konar, Owen O'Malley, Sanjay Radia, and Suresh Srinivas.

Under the leadership of Greg Pavlik, VP Engineering, and Tim Hall, VP Product Management, this core product team has been enriched with enterprise software talent from the likes of Oracle, IBM, HP, VMware, and others to help ensure that HDP meets the enterprise-grade requirements our customers expect.

Our CEO Rob Bearden has assembled one of the most experienced open source business model teams on the planet that have been instrumental in the success of companies such as JBoss, Red Hat, SpringSource, Zimbra, XenSource, Pentaho, Lucidworks (Solr), Docker, and others. This team includes board of directors Peter Fenton (Benchmark Capital), Paul Cormier (Red Hat), and Martin Fink (HP), along with executive team members Shaun Connolly, Herb Cunitz, Mitch Ferguson, and others.

Key Milestones

Since our founding in April 2011, our partners and we have achieved the following significant milestones:

- October 2011: Announced strategic relationship with Microsoft to deliver Hadoop-based solutions for Windows Server and Windows Azure;
- February 2012: Announced joint development, support and marketing partnership with Teradata;
- June 2012: Hortonworks Data Platform Version 1.0 general availability;
- September 2012: Hortonworks Data Platform Version 1.1 general availability;
- October 2012: Announced joint initiative with Rackspace to deliver OpenStack and Hadoop-based solutions for public and private cloud;
- January 2013: Hortonworks Sandbox general availability and Hortonworks Data Platform Version 1.2 general availability;
- May 2013: Hortonworks Data Platform Version 1.1 for Windows general availability (first release for Windows);
- June 2013: Hortonworks Data Platform Version 1.3 general availability; Teradata Launches Teradata Portfolio for Hadoop (Teradata sells Hortonworks Data Platform in appliance, on commodity hardware, and as a subscription);
- August 2013: Hortonworks Data Platform Version 1.3 for Windows general availability;
- September 2013: Announced reseller agreement with SAP;
- October 2013: Hortonworks Data Platform Version 2.0 with YARN general availability (first general availability release of the Hortonworks Data Platform with YARN), announced reseller agreement with HP and Microsoft launches Windows Azure HDInsight (Azure cloud service built on Hortonworks Data Platform);

- February 2014: Announced deepened strategic alliance with Red Hat to bring enterprise Apache Hadoop to the open hybrid cloud;

- April 2014: Hortonworks Data Platform Version 2.1 for both Linux and Windows general availability (simultaneous distribution of the Hortonworks Data Platform on both platforms);

- May 2014: Hortonworks acquired XA Secure, a data security company. In connection with the acquisition of XA Secure, we acquired developed technology with a fair value of approximately $4.0 million. On August 13, 2014, we contributed the developed technology to the Apache Software Foundation (ASF) and recognized an operating expense equal to the carrying value of the developed technology in our statement of operations at the point in time of the contribution (see discussion in Note 4 of the accompanying notes to the consolidated financial statements);

- June 2014: Launch of Hortonworks YARN Ready Program to accelerate independent software vendor on-boarding onto YARN;

- June 2014: Certified IBM InfoSphere Guardium on HDP 2.1 (heterogeneous solution to meet security and compliance requirements of the enterprise); Announced Apache Spark, an in-memory data processing API and execution engine, is YARN ready and certified to be fully compatible with Hortonworks Data Platform; Introduced HDP Advanced Security (from XA Secure acquisition) for Hortonworks Data Platform Enterprise Plus users;

- July 2014: Announced a strategic partnership with HP that deeply integrated Hortonworks Data Platform with the HP HAVEn big data platform (Hortonworks Data Platform became the Hadoop component of HP HAVEn). The partnership strengthened the existing go-to-market collaboration and brought new integration of engineering strategies (HP Vertica became YARN certified). Additionally, HP made a $50 million equity investment in Hortonworks and joined its board of directors; Announced a collaboration with Pivotal on the Apache Ambari project; Entered into a strategic alliance agreement with Accenture to further build upon its big data and digital capabilities;

- September 2014: Announced integration of Hortonworks Data Platform and Apache Ambari with Cisco's UCS Director Express for Big Data (enabling single pane of glass monitoring and management coupled with automatic deployments of Apache Hadoop);

- October 2014: Hortonworks Data Platform certified on Microsoft Azure Infrastructure as a Service (IaaS) for hosted deployment in a virtual machine (interoperable with on-premise deployments from Hortonworks Data Platform for Windows, managed service deployments from HDInsight, appliance deployments from Microsoft Analytics Platform System); Announced Oracle Data Integrator (ODI) is certified with HDP 2.1; Announced an alliance partnership with Avanade to build new big data solutions and services based on Microsoft technologies (Hortonworks Data Platform on Windows, Azure HDInsight); Expanded

partnership with Databricks built around Apache Spark; Announced SequenceIQ as a new Technology Partner and HDP YARN certified (Docker containers for Hortonworks Data Platform deployments in the cloud)

- November 2014: Expansion of comprehensive Hortonworks Certified Technology Program to include more than 70 technologies that became HDP YARN Ready and introduced new enterprise components (HDP Operations Ready, HDP Security Ready, HDP Governance Ready); and
- December 2014: Hortonworks Data Platform Version 2.2 for general availability (new and improved YARN-ready engines, enterprise SQL at scale, centralized security, improved management, rolling upgrades).
- January 2015: Hortonworks Data Platform became certified and available on the Google Cloud Platform
- January 2015: Data Governance Initiative, a collaboration of industry leaders across key verticals, including Hortonworks, Aetna, Target and Merck launched to address industry-specific and vertical-specific complexities associated with data lifecycle management.
- February 2015: Announced strategic partnership, including joint engineering and support agreement with Pivotal Software, Inc.
- February 2015: Announced strategic partnership, including joint engineering and resale arrangement with Hitachi Data Systems.

D.3 References

Hortonworks is honored to be selected as one of State **Farm's** chosen Hadoop vendors and as such we enjoy a mutually beneficial working and strategic relationship. For the purposes of this RFP, we are also happy to share with you these references who are extensive and prolific users of Hadoop and the Hortonworks Data Platform. Particularly around the solution design Hortonworks is proposing for the State Farm NGIS platform.

Company Name: FINRA – Financial Industry Regulatory Organization

Contact: Scott Donaldson, Sr Director

Overview: FINRA have been working with Hadoop for over four years. They are skilled practitioners of HBase/Phoenix/Hive and have made significant contributions back to the Apache Software Foundation by developing a co-engineering relationship with Hortonworks. Their cluster is scaled to support the running of over 2 trillion rows on Phoenix/HBase with a two second response time. FINRA also leverages the cloud for bursting when capacity requirements dictate a temporary need for more compute. Ari Zilka, our CTO, as well as our professional services organization have been directly involved in the design, development and implementation of FINRA's Hadoop environment.

Company Name: United Health Group

Contact: Adam Waldal

Overview: UHG is utilizing the entire HDP stack and plan to be at 1,000 nodes by the end of 2015. Hortonworks CTO, solutions engineering and professional services have been directly involved with the design, development, implementation of HDP as well as the operationalization of their data science team.

Company Name: Verizon Wireless

Contact: Ashok Prasad

Overview: Verizon is running an HDP cluster of over 800 nodes and growing leveraging HBase, Kafka, Storm and Ranger. Hortonworks solutions engineering and professional services have been directly involved with the design, development and implementation of Verizon's Hadoop environment.

Company Name: Bloomberg

Contact: Bloomberg is featured in speaking engagements at HBaseCon as well as Hadoop Summit.

Overview: Bloomberg is one of the largest practitioners of HBase and Hadoop. Hortonworks and Bloomberg work closely in a co-engineering partnership that has resulted in the advancement of HBase and numerous innovations gifted back to the Apache Software Foundation under Hortonworks stewardship.

Company Name: AT&T

Contact: John Yovanovich

Overview: AT&T has over 1,000 nodes in production for their internal use leveraging all aspects of the HDP platform and running batch, interactive in real-time use cases leveraging a multi-tenant design. Ari Zilka, our CTO, as well as our solutions engineering and professional services organization have been directly involved in the design, development and implementation of AT&T's Hadoop environment.

3. Solution Recommendation

The architecture we propose leverages key mix technologies such as HBase, SolrCloud, Kafka and Storm to deliver functional and nonfunctional requirements of the solution.

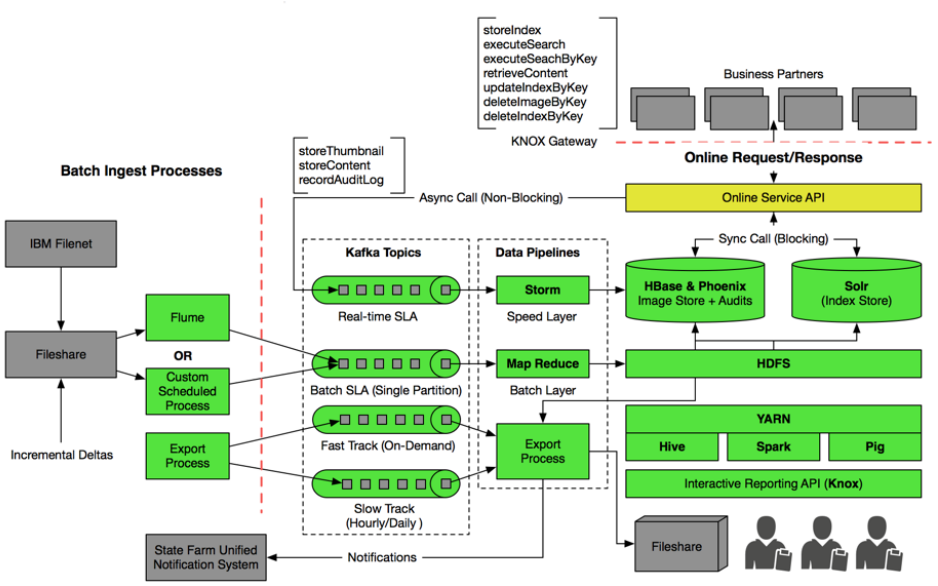
HBase: Distributed Column Family oriented database on Hadoop

SolrCloud: Highly reliable and scalable Search Engine

Kafka: High throughput distributed publish-subscribe broker

Storm: Distributed real-time (sub-second) computation system.

The solution will use HBase as system-of-record to store images (Image Store) and uses SolrCloud as system-of-record to store document metadata (Index Store). The figure below describes the overall architecture where components in green represent the technologies and processes that are part of HDP solution stack. The components in gray represents some of the key State Farm services Hadoop will be interacting with. The Online Service API highlighted in yellow is the service layer which will be a joint effort between Hortonworks and State Farm teams.



The first goal of Online Service API is to provide sub-second response times for search and content retrieval requests, and the second goal is to provide a secure and standardized way to interact with Hadoop services such as HBase, Solr, HDFS, etc. The Online Service Interface will be designed to handle both Synchronous (blocking) and Asynchronous (non-blocking) requests. Synchronous requests will directly interact with HBase and Solr APIs, whereas asynchronous requests will be passed through Kafka topic and Storm in real-time.

In contrast to Online Service API, Batch processes will have relaxed SLAs. Batch processes will be primarily used to handle insert, update, and delete requests (for both index and image store) that come through non-online medium. There are multiple options when it comes to ingesting data into hadoop in batch fashion. Below are the few options we recommend you explore.

1. Custom process that is made up of one or more jobs can read the data from source system and land it into Hadoop. This process can be scheduled using Oozie/Falcon to run periodically.

2. Apache Flume provides a way to efficiently used collecting, aggregating, and moving large amounts of data into HDFS. It has a simple and flexible architecture based on streaming data flows. Apache flume provides out of the box connectivity to wide variety of sources. You can find more info here https://flume.apache.org/FlumeUserGuide.html#flume-sources

3. The Hadoop connector for Solr provides way to ingest data into Apache Solr in batch mode. The Hadoop Connector is developed by the Lucidworks open source team and is certified to work with HDP. The connector will ingest documents, using built-in map-reduce jobs, in CSV, Microsoft Office files, Grok (log data), Zip,Solr XML,Seq files, and WARC. The connector also supports pig scripts to perform large scale pre-processing and joining of data sending the resulting dataset to Lucene/Solr.

4. Design Recommendations

F.2 Information Lifecycle Management

1. Please recommend options for achieving cost reduction and optimum performance. 2. Explain the architecture implications of moving the content between "hot" and "cold" storage spaces. 3. Please provide the cost saving analysis of this solution reflecting the architecture/maintenance overhead of the proposed ILM implementation.

Option 1 - Archive to Passive Standby

HBase allows setting up replication stream from active cluster to one or more geographically distributed clusters. The main idea here is to keep live data in active environment and use passive environment to store both live and archived data. This option takes advantage of near real-time log-shipping feature in HBase to keep passive cluster/s in sync. Based on ILM policy per business area, HBase entities can be configured with TTL (Time-to-Live). Upon reaching contractual expiry period, aged documents will be automatically cleared away from active cluster. This approach requires no explicit backup process, which makes it architecturally simple and attractive option.

When the data needs to be restored back, it is readily available in ILM or passive cluster and can be recovered in relatively short time.

Option 2 - Archive to Passive Standby

The main idea here is to take periodic extracts of aged data from HBase and move them to denser (cold) portion of the cluster. Cold-tier is generally referred to the portion of a cluster

that is made up of low-cost high-density nodes (anywhere from 48-60TB/node). Heterogeneous storage feature allows you to mark specific directories in HDFS as Cold, Hot, Warm depending on SLA needs.

This option will use HBase export utility or a custom export process to extract the contents of HBase table into sequence file, which will then be stored in HDFS Cold-tier. Exporting data to Hadoop sequence files has merits for data backup, because the Hadoop SequenceFile format supports several compression types and algorithms. With that you can choose the best compression options to fit our environment. Export utility can be configured with a start and an end timestamp, so that only the data within a specific time frame will be exported. Architecturally, this option requires creating, maintaining and scheduling separate archival process.

Restoring data will require running HBase import utility or custom import processes, which will essentially take the docIDs, requested and reconstruct HBase records from archived sequence file.

HP DL380 G9 2660v3	Estimated Cost
Server	$8,475
Drives (12 X 2TB) - HBase Node	$5,508
Total	$13,983
Cost/TB	$583

HP DL380 G9 2660v3	Estimated Cost
Server	$8,475
Drives (12 X 4TB) - Cold Tier	$9,468
Total	$17,943

Cost/TB	$374

Example:	Data (TB)	Cost	Savings
Live Data in HBase	1000	$582,625	
Cold Tier - Data archived in HBase with additional 10% Compression	**900**	**$336,431**	**$246,194**
Cold Tier - Data archived in Sequence file with 30% Compression	**700**	**$261,669**	**$320,956**

F.4 Reporting

Discuss ability to provide a report/matrix of performance and utilization i.e.: How is the system performing, what percentage of documents stored are pdf, tiff, jpg, etc; total documents stored; number of documents retrieved for a given interval; what is the growth rate in the number of documents for a class, entire system; what is the growth rate in storage consumed

Given the high-volume transactional nature of this application, we recommend maintaining separate Audit table in HBase to effectively capture ongoing system activity. This means audit table will store detail log and timestamp of every Insert/Update/Delete action performed on Index store as well as Image store. Having this information readily available in HBase allows you to run queries against Audit table to gain insights into various metrics around document type, usage, growth projections, etc. To bring HBase to mainstream reporting, Hortonworks supports Apache Phoenix, which allows users to interact with HBase using free-form SQL queries and it also provides integration with enterprise dash-boarding tools that are JDBC source compatible.

F.5 Questions and General Discussion

1. **Can Hadoop meet the Index Store (iSEIT) requirements described in Sections E and F, especially given high volume random index search and update index requirements? Describe in detail how we can do the update.**

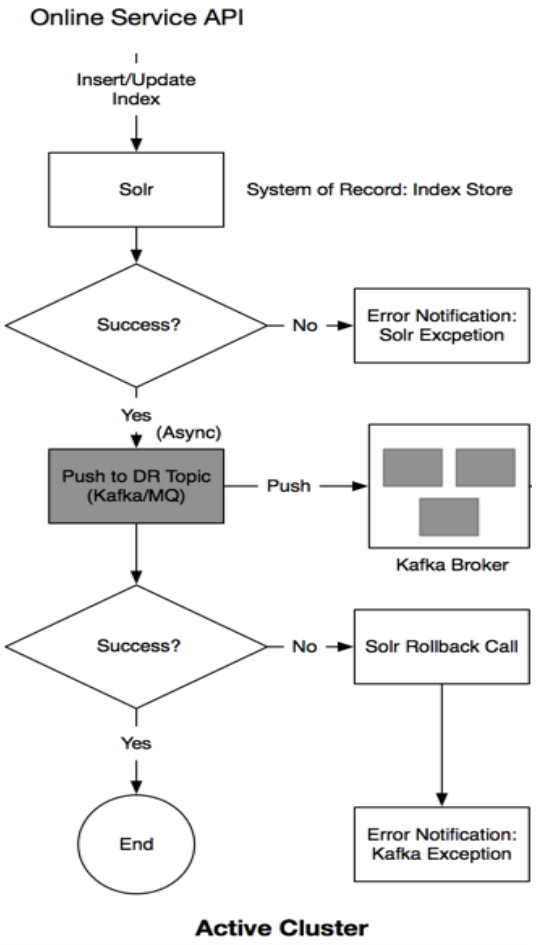
Yes, the solution we propose can meet the high volume random index search and update requirements.

Our initial approach was to use HBase as “ultimate system of record" for both Index and Image store. After learning about transactional requirements around concurrency and locking, we realized that separation of responsibilities between different data stores will be the most effective way to address these requirements. Apache Solr fits very well to the needs of State Farm around update process. When Solr is used as a system of record, you no longer need to worry about managing concurrent updates, optimistic locking, rollbacks, etc. Solr handles this out of the box. So taken the revised approach, we would recommend using –

Hbase: "System of Record" for Images, Thumbnails, Document Properties and other non-editable metadata.

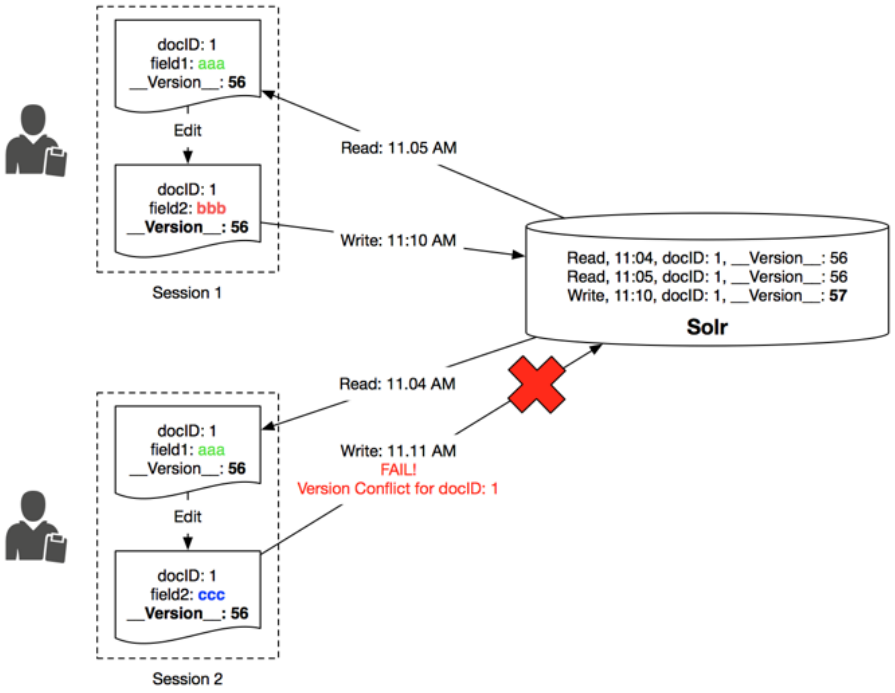
SolrCloud: “System of Record” for Index Store and other searchable metadata.

This architecture will entail that all index insert/update requests will exclusively go to Solr. Hbase/HDFS will only be accessed when user performs read/upload document request based on rowkey. The figure below describes end-to-end process to Insert/Update Index through online service API



Atomic updates feature in Solr allows you to send updates to only the fields you want to change. This brings Solr more in line with how database updates work. Solr will still delete and create a new document, but this is transparent to your client application code.

It's conceivable that two users will attempt to update the same document at the same time. To avoid conflicts, Solr supports optimistic concurrency control using a special version-tracking field. When a new document is added, Solr assigns a unique _version_ number automatically. When you need to guard against concurrent updates, you include the exact version the update is based on in the update request.



When Solr processes this update, it will compare the _version_ value in the request with the latest version of the document, pulled from either the index or the transaction log. If they match, Solr applies the update. If they don't match, the update request fails and an error is returned to the user. A client application can handle the error response to let the user know another user already modified the document. This approach is called "optimistic" because it assumes that most updates will work on the initial attempt and that conflicts are rare.

2. **Can Hadoop index design participate in external transaction management for index update? If not, what are some ideas to overcome the limitation? (Example: Hadoop index update should rollback when external IMS transaction issues rollback.)**

Hadoop will not participate in external transaction management. It will be programmatically handled. Please refer to 1

3. **Can Hadoop send out notification for success or failure of Store, Update, Export and Delete services?**

Yes, In case of failure the solution will produce a system/business notification, which can also be pushed out to State Farm Unified Notification System.

4. **Can you use a unique key provided by State Farm as a key for image file or index data?**

Yes.

5. **Can we design logical separation or partitions in Hadoop as detailed in F.1 Partitioning Requirements section above?**

We recommend using HBase for *Image Store* and SolrCloud for *Index Store*. Both HBase and Solr are built with scalability as a first class concern and are inherently to scalable to meet your growing data needs. This allows you to focus less on scaling heuristics but more on the actual use case.

HBase will host *Image Store* and will also be a "System of Record" for all documents/Images. This means HBase will store the actual image/doc, thumbnail, document properties, etc. In order to meet regulatory requirements and institute a strong security model, we recommend creating table-based logical separation for business areas. For example:

- Table 1: Auto, Fire, SFPP, Life, Health and Health Claims
- Table 2: P&C Claims (ECS)
- Table 3: ECC Documents

Solr will host *Index Store* to provide sophisticated search at sub-second speeds. SolrCloud has a concept of "Collections", which essentially is an instance of Lucene index. In SolrCloud you can create multiple collections (just like in database you create tables). A collection is typically used to separate documents that belong to different areas and/or have different indexing requirements. We recommend creating collection-based logical separation for each business area. For example:

- Collection 1: Auto, Fire, SFPP, Life, Health and Health Claims
- Collection 2: P&C Claims (ECS)
- Collection 3: ECC Documents

Both HBase and Solr are distributed systems, which expose set of APIs for users and applications to interact with it. Thus accessing particular table or a collection is simply a matter of specifying required table/collection name as parameter in an API call. This parameter will be "LOB" or artifactID.

6. **Can Hadoop use ArtifactID or LOB to logically partition the document store and index?**

Please refer to 5.

7. **Can Hadoop export files and write them to an external file share? Can this be a configuration setting and if so how?**

Yes, it is a 3-step process, which can be scripted with configuration options

1. Export Process will export requested set of Images to HDFS
2. Copy/Move data from HDFS to Edge Node using HDFS client library command
3. Copy/Move data from Edge Node to ftp

8. **Thumbnail management – thumbnails will not likely be available during initial document store or index insert processing. Can the Update Index service be used to add thumbnails or design separate services to insert and retrieve the thumbnail?**

Inserting and retrieving thumbnail logically seems like a separate unit of work. Designing separate service will be a best practice approach.

9. **Audit data management – audit data will be collected each time the index store is accessed or modified. Can the Update Index service be used or would we need to design a separate service to insert and retrieve audit data?**

Audit logs are essentially set of activities ordered by time. Time-series data is an apt use case for HBase. We recommend using separate HBase table to capture all audit related information. Based on the functional requirements described in section E2, best way to perform audits would be to do insert/update/deletes to Index Store on a main thread and record audit related information on a background thread. This can be accomplished by asynchronously calling "Record Audit" Service in conjunction with "Update/Insert Index" service.

To retrieve audit logs, we recommend creating a separate service.

10. **In what format does index data need to be stored in Hadoop XML and key-value pair? How does the index data need to be fed into Hadoop?**

The proposed solution will leverage SolrCloud to store index data. It will be stored as Lucene segments for blazing fast search performance. Solr has Java Client API libraries like Solrj, Spring Data Solr that makes it easy for developers to construct index payload using familiar Java semantics. For batch feeds, we recommend using Data Import Handler (DIH), which integrates with various data sources (including databases) to index data into Solr. You can also use MR/Spark job scheduled via Oozie/Falcon to periodically ingest data into Solr. For online high-frequency requests, we recommend creating standardized Service API layer, which can load balance and broker insert/update request in real-time.

11. **How many Index field variables can be supported in Hadoop search, given the TPS in Appendix E (Business Value Metrics and Service Levels)? (Example: Account =111, Receive Date: 3/3/2001, Amount= 1000, office code = 08 etc.)**

You can add as many field variables as you want to filter your search results without impacting performance SLA. There is no theoretical limit on number of filter fields allowed in a search query.

12. Can more index types be dynamically added to index store? How do we make the addition seamless? (Example: Add account holder, first name, and last name to bank documents in addition to account number, check amount etc.)

One of the key advantages of using Solr is that it has a flexible schema. This means, the documents in a search index don't need to have a uniform structure. Solr provides a simple, declarative way to define fields in your index; and how you want those fields to be represented in a configuration document named schema.xml. This gives you an ability to add new fields on the fly with no downtime.

13. How are the iSEIT data types (set forth in Appendix F) supported in Hadoop? We would like to simplify Search service by promoting native primitive data types like integer, float etc. for index.

Solr supports all iSEIT data types. For more information, please refer to link below. https://cwiki.apache.org/confluence/display/solr/Field+Types+Included+with+Solr

14. Explain the solution to automate cluster creation and expansion, deployment of software and services within Hadoop box in Appendices A and B attached hereto.

With Ambari Blueprints, system administrators and dev-ops engineers can expedite the process of provisioning a cluster. Once defined, Blueprints can be re-used, which facilitates easy configuration and automation for each successive cluster creation. Blueprints provide scripted instantiation of a Hadoop cluster quickly and without requiring manual user intervention. And because Blueprints contain knowledge around service component layout, they preserve best practices across different environments. Blueprints ensure that those best practices for service component layout and configuration are consistently applied across clusters in multiple environments (dev, test, prod) and in multiple data centers.

15. With respect to Hadoop, in general:

a. Can Oozie/Falcon be installed on an edge node?

Yes

b. Can multiple concurrent instances of Oozie/Falcon be run on the same edge node? Yes

c. Can Oozie/Falcon orchestrate plan Java programs and Perl Scripts? If so, can Oozie/Falcon take output from one Java process and feed it as an input to the subsequent process?

Yes, this is typically accomplished using file channel where output of first process is written to a file that will then be used as input to subsequent process.

d. Can Hadoop store the images on the WORM complaint hardware mounted as network file share ("NFS") mount? Hadoop jobs can treat a WORM complaint hardware mount just as any other source or destination.

We recently announced an engineering partnership with EMC where we certify running HDP on EMC Isilon Scale-out Network Attached Storage. EMC Isilon comes bundled with SmartLock feature, which is software-based approach to Write Once Read Many (WORM) data protection. You can store SmartLock-protected data alongside other data types in your Isilon scale-out storage environment.

On a side note - HDFS as a "file system" has characteristics of WORM. A file once created, written and closed will not be changed.

16. Please provide feedback comments, and recommendations with respect to the following

a. **Technology for an initial cluster build and configuration.**
b. **Technology for cluster expansion.**
c. **Technology for daily cluster operations management.**
d. **Technology for meeting governance and security requirements.**
e. **Technology for audit of the cluster.**

Hortonworks Data Platform (HDP) includes Apache Ambari - A completely open operational framework for provisioning, managing and monitoring Apache Hadoop clusters. Ambari includes an intuitive Web interface that allows you to easily build, configure and deploy all Hadoop services. No matter what the size of your Hadoop cluster is today or expansion plans for tomorrow, the deployment and maintenance of hosts is simplified using Ambari. Ambari also provides the powerful Ambari Blueprints API for automating cluster installations without user intervention. The Web interface allows you to control the lifecycle of Hadoop services and components, modify configurations and manage the ongoing growth of your cluster. Through single pane of glass you can gain instant insight into the health (CPU, disk, IO, memory, threads, job-progress, capacity, java heap, network activity) of your cluster. Ambari pre-configures alerts for watching Hadoop services and visualizes cluster operational data in a simple Web interface.

HDP ships with Apache Falcon - A framework for simplifying and orchestrating data management and pipeline processing in Hadoop. It enables automation of data movement and processing for ingest, pipelines, replication and compliance use cases. Falcon simplifies the development and management of data processing pipelines with a higher layer of abstraction, taking the complex coding out of data processing applications by providing out-of-the-box data management services. This simplifies the configuration and orchestration of data motion, disaster recovery and data retention workflows.

HDP delivers comprehensive approach to security through Apache Ranger, Knox and Kerberos. It provides central security policy administration across the core enterprise security requirements of authentication authorization, accounting, data protection and perimeter security. Apache Ranger offers a centralized security framework to manage fine-grained access

control over Hadoop Components. Using the Apache Ranger web console, security administrators can easily manage policies for access to files, folders, databases, tables, or columns. These policies can be set for individual users or groups and then enforced within Hadoop. Security administrators can also use Apache Ranger to manage audit tracking and policy analytics for deeper control of the environment.

17. **General Discussion: State Farm uses a logical document consisting of individual pages referenced by a single key. Would there be extra complexity/issues in developing services dealing with logical documents versus individual files for each of the services (store, retrieval, export, delete)? This would include management of input and output collections and persisting logical document order and sequence information. Should Hadoop be aware of Logical document concept so it can store the pages in a sequential order to gain the efficiencies during retrieval?**

Rows in HBase are sorted lexicographically by rowkey. Requirements state that logical document is made up of one or more pages (physical document). Rowkey design plays an important role in simplifying overall service design and in incorporating concept of logical document. As long as rowkey is designed in a way that logical document follows "docID+pageNumber" syntax, HBase by its very nature will store individual pages of a given logical document in sequential order. This rowkey design is optimal in the context of your needs, allowing you to store related rows, or rows that will be read together, near each other. During document retrieval, you can request all the pages of the document by of supplying only the first part of the key (docID). You can also request individual pages by passing set of docID+pageNumber.

18. **General discussion about iSEIT about iSEIT index data hierarchy. iSEIT will be able to support the data hierarchy today in the future design. Hadoop need not be aware of the hierarchy of the data, just need to ensure index data includes LOB, Document Class, Document type and Sub Document codes so they can filter that data**

To efficiently handle the operation of finding matching documents, Solr makes use of 'fq' (filter query). 'fq' serves a single purpose: to limit your results to a set of matching documents. One important characteristic to note about filter queries is that you can add as many fq parameters as you want to your Solr request. Thus you can include LOB, Document Class, Document Type, etc. as part filter query and Solr will make sure to serve only the results that match a given hierarchy.

19. **Discuss monitoring capability and tool to ensure the health of the system. Ability to monitor system state via tools and automation. Ability to monitor for free space for the storage of documents. Provide access to system specifics for troubleshooting,**

administering, etc. Needs to include automation to generate incident records, see note on Monitoring in F.4 Additional Non-functional Requirements.

Hortonworks Data Platform ships Apache Ambari to perform operational and administrative tasks on the cluster. Ambari provides a dashboard for monitoring health and status of different services that run on your cluster. Ambari leverages Ganglia for metrics collection and Nagios for system alerts to notify when your attention is needed (e.g., a node goes down, remaining disk space is low, etc.).

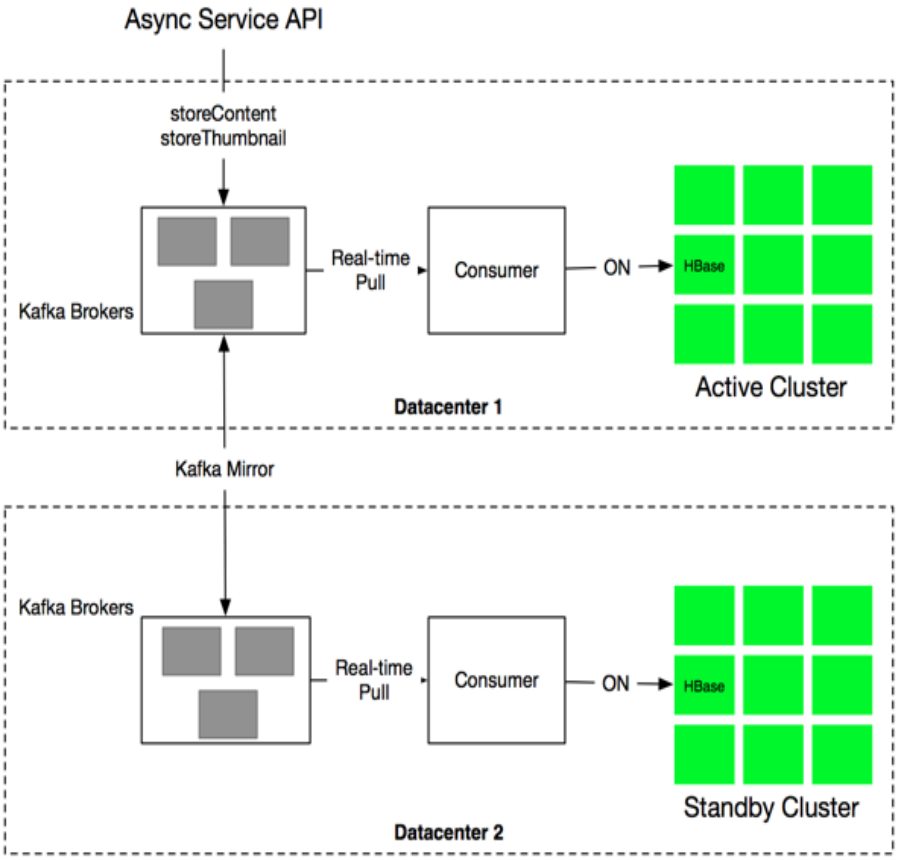
Ganglia (http://ganglia.sourceforge.net/) is a distributed monitoring framework designed to monitor clusters. It was developed at UC Berkeley and is the de facto solution to monitor Hadoop clusters. Ganglia captures metrics to explain cluster's behavior related to the system load, network statistics, RPCs, JVM heap, and JVM threads, etc. Most of the Hadoop services expose metrics via JMX. Several open source tools such as Cacti and OpenTSDB can be used to collect metrics via JMX.

20. **Discuss replication and disaster recovery. Assume we have two distributed clusters, one active and one passive. If we failover to the passive cluster, what are the implications for switching replication back the other direction? Are there any tools automating both the failover and reversing replication? What happens if the old active cluster that failed is offline for 2 weeks? How does replication catch up? During the switchover from Active to Passive if the passive is now online and being used as the active cluster for a few months, can the previously active cluster become the passive cluster? If so is this just a move of the incremental that occurred during the switchover a full push of the content? Some of our assumptions are that the index information is in HBase and the image files are in HDFS**

DR for HBase:

Option 1 - Kafka Mirror

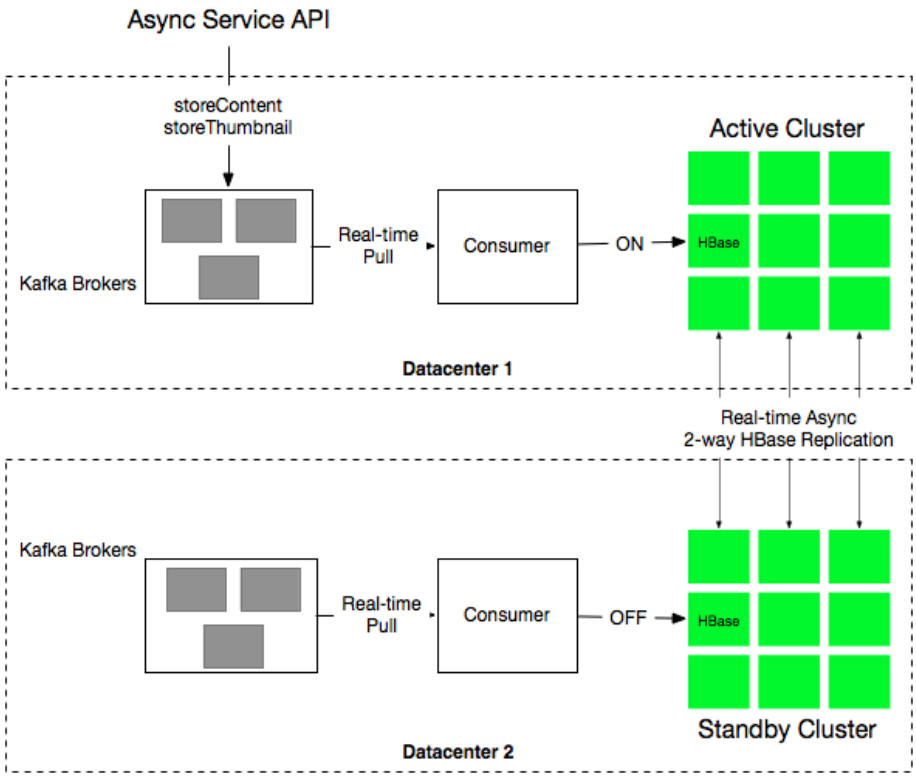
Kafka serves as a transient landing zone for Images/Thumbnails. Kafka Mirroring is process of replicating data between Kafka clusters. This approach does not depend on HBase to do the replication and hence frees up the HBase resources for online reads/writes. Since the consumer feeds both data centers at the same time, both clusters stand a better chance of staying closely synced.



Option 1: HBase High Availability

Option 2 - HBase Replication functionality to set-up 2-way replication.

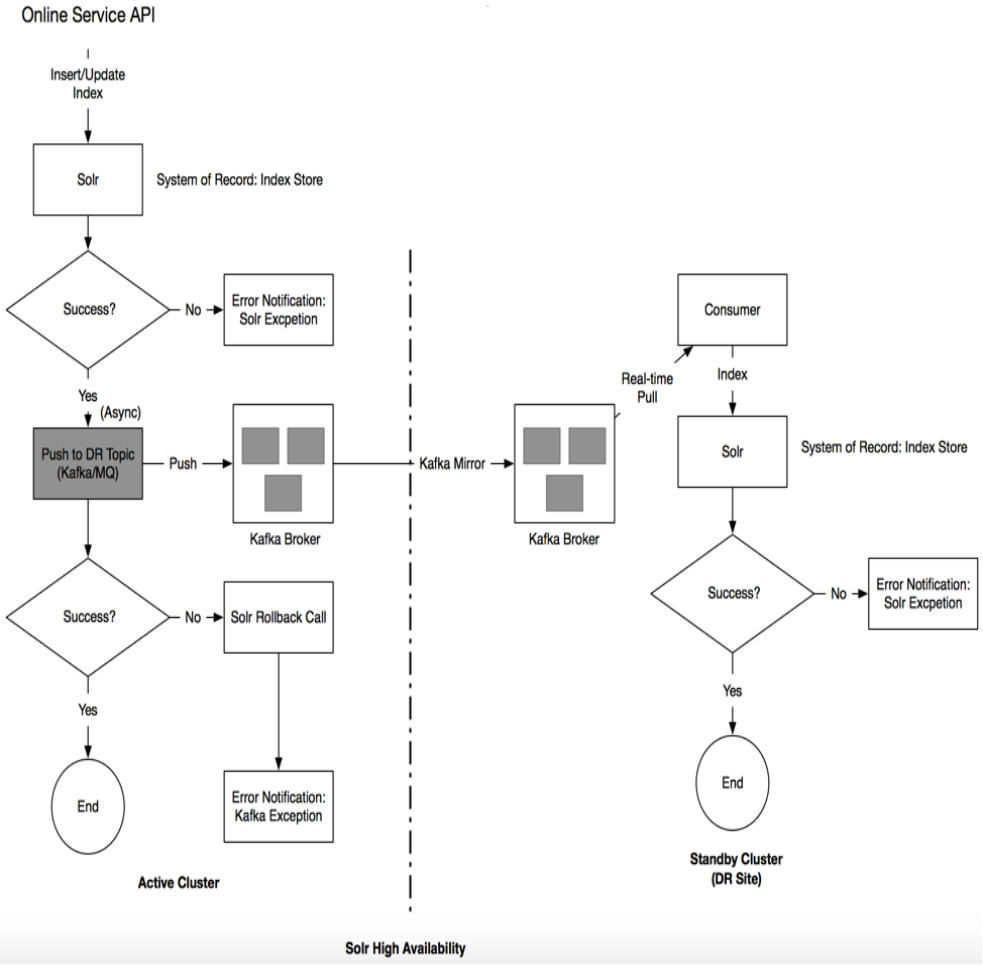
HBase provides a cluster 2-way replication mechanism that allows you to keep one cluster's state synchronized with that of another cluster, using the write-ahead log (WAL) of the source cluster to propagate the changes. An HBase cluster can be a source (also called active, meaning that it is the originator of new data), a destination (also called passive, meaning that it receives data via replication), or can fulfill both roles at once. Replication is asynchronous near real-time, and the goal of replication is eventual consistency. When the source receives an edit to a column family with replication enabled, that edit is propagated to all destination clusters using the WAL for that for that column family on the RegionServer managing the relevant region.



Option 2: HBase High Availability

DR for Solr

SolrCloud will be used as system-of-record for searchable document metadata. All inserts/updates to index store will be done using Solr API calls. Solr will piggyback on Kafka for inter-datacenter replication. Each cluster will maintain a topic called “DR”, which will capture the sequence of operations as write-ahead-log. With Kafka mirroring enabled, data from DR topic will be replicated to standby cluster. The cluster acting as Standby will then be able to apply these operations to its Solr Cloud instance (Index Store) in the order they were received on the Active site.



In case of a disaster failover can be executed in 2 ways -

1. Manual:

This requires IT personnel to change your application's public DNS to point to the secondary data center. The DNS change will take some time to propagate, but as end users get the updated DNS record, they'll see the "Under Maintenance" page (if you created one), rather than an error in their browser. In the meantime you will be re-routing your batch ingest/export processes performing sanity checks.

2. Automatic:

If your public DNS provider supports it, you could even set up an automatic failover upon loss of connectivity to your primary data center. In this scenario, rather than simple "Under Maintenance" messaging, you could have a portion of the application environment in place to deliver critical online services search/update/retrieval until it is augmented with batch ingestion/export pipelines to handle the full application load and act as current active.

HBase allows you to take a snapshot of a table. Snapshot doesn't involve data copying and is only a representation of a table during a window of time. The amount of time the snapshot operation will take to reach each Region Server may vary from a few seconds to a minute, depending on the resource load and speed of the hardware or network, among other factors.

Rebuilding a failed cluster will begin with importing a HBase snapshot from current active. This operation copies the snapshot data and metadata to another cluster. The operation only involves HDFS so there's no communication with the Master or the Region Servers. In addition to snapshots, custom processes will be used to sync HDFS and Solr index. We see this as a one-time effort, which then can be automated through scripts. To catch up on incremental deltas and sync both sites to the exact same state, short planned maintenance window may be required.

OCR/Analytics Framework for systematic mining of the image documents

Although tools such as Tesseract, Apache Tika, HIPI, etc. are known OCR, Parsing, Image processing frameworks that work well with Hadoop, we recognize no one tool or technology will practically cover all your requirements and needs. Hortonworks is more than happy to work with State Farm in this area to help find avenues to do OCR and related analytics at scale.

5. Project Requirements – Training

Please provide the following information with respect to training:

1. A list of all general Hadoop training offerings, and for each offering, please provide the following information:

a. **Description of the training course;**
b. **Location of training course (e.g., online, classroom, onsite, etc.);**
c. **Limit on number of attendees (if applicable);**
d. **List of materials provided to State Farm as part of the training course;**
e. **Number of offerings per year;**
f. **Duration of the training course;**
g. **Notice required to schedule the training course (on average); and**
h. **Any other information with respect to the training course you'd like to note.**

Hortonworks provides a robust set of tutorials via the "Getting Started" section of the Hortonworks web site http://hortonworks.com/tutorials/ that provide a general and detailed overview of Hadoop based on what kind of information you are seeking. This training is all on line, there is no limit in terms of access, the tutorials can be watched on line and the sandbox environment can be downloaded to try hands on exercises. The duration of the tutorials will vary from 30 minutes to several hours depending on how long it takes someone to go through the tutorials. No notice is required to schedule access to these tutorials. This training is free and available anyone interested in Hadoop, this is separate from the formal product training that is offered through Hortonworks University.

2. A list of all training offerings specific to your technology, and for each offering, please provide the following information:

a. **Description of the training course;**
b. **Location of training course (e.g., online, classroom, onsite, etc.);**
c. **Limit on number of attendees (if applicable);**
d. **List of materials provided to State Farm as part of the training course;**
e. **Number of offerings per year;**
f. **Duration of the training course;**
g. **Notice required to schedule the training course (on average); and**
h. **Any other information with respect to the training course you'd like to note.**

Below is a table detailing the Hortonworks University Training offerings. All Hortonworks training courses are specific to the Hortonworks Data Platform (HDP) and focus on teaching students how to use HDP.

Course delivery is available in the following ways:

a. Instructor Lead delivery at the client location (Onsite Training). This class is dedicated to the customer attendees only, the client does lab set up in advance and instructor comes to your location to deliver the course. Most students bring their own device or use a preconfigured training room

b. Instructor Lead Virtual delivery (Virtual Onsite Training) this is a dedicated virtual delivery of a class for only the customer attendees. All attendees and the instructor connect to the class virtually. The instructor does not come on site but lab set up by each of the attendees is still required in advance of the class.

c. Instructor Lead delivery at a Public Training Facility. Hortonworks has a network of global public training center where students can attend Hortonworks classes either in person or, in North America, they can join a virtual remote session where they are broadcast in to the class and the instructor can see them and they can see the instructor and other students.

d. Self Paced Learning Library: Named support contacts for customers with Enterprise Support agreements are provided free access to the Hortonworks University Self Paced Learning Library. This library of learning contains all current course books and in many cases, recorded labs from the courses for the students to study and practice at their own pace. Access is available for the duration of the support relationship. Access for additional students is available for purchase. The Self Paced Learning Library is always growing in content and provides good on-going support to Hortonworks customers.

e. Certification Exams: Hortonworks certifications establish that you are a trusted and valuable Apache Hadoop professional. These qualifications have been designed and developed by the leaders and core committers of Apache Hadoop so that you and the industry can be assured that the highest standards in the industry have been obtained. Currently certification exams are available for Hadoop Administration (Certified Hortonworks Administrator) and for Hadoop Developers in both Java and Pig & Hive. Hortonworks currently offers certification exams through our testing vendor Kryterion but is in the process of converting all exams to Performance Based Exams that students can take from any computer at any location with a remote proctor monitoring the exam.

f. Tailored training: State Farm is able to select topics from a variety of classes and have them delivered in one class. Topics would be mutually agreed upon between Hortonworks and State Farm.

g. Custom training: Hortonworks University staff would work with State Farm to build a class that is custom to State Farm's use cases and environment. Work is done as education consulting and a daily development rate is charged along with the delivery fees for the training class. State Farm would own the training materials for future distribution at the conclusion of the engagement.

Hadoop Essentials Course Description: An overview of Apache Hadoop and Hortonworks Data Platform for decision makers and business users. Offered as a self paced on line class only, there is no charge for this class and has approximately 40-60 hours of content that can be studied on line.
Developer Classes
HDP DEVELOPER: PIG & HIVE (4 DAYS) This training course is designed for developers who need to create applications to analyze Big Data stored in Apache Hadoop using Pig and Hive. Topics include: Hadoop, YARN, HDFS, MapReduce, data ingestion, workflow definition and using Pig and Hive to perform data analytics on Big Data. Labs are executed on a 7-node HDP cluster
HDP DEVELOPER: JAVA (4 DAYS) This advanced course provides Java programmers a deep-dive into Hadoop application development. Students will learn how to design and develop efficient and effective MapReduce applications for Hadoop using the Hortonworks Data Platform, including how to implement combiners, partitions, secondary sorts, custom input and output formats, joining large datasets, unit testing, and developing UDFs for Pig and Hive. Labs are run on a 7-node HDP 2.1 cluster running in a virtual machine that students can keep for use after the training.
HDP DEVELOPER: WINDOWS (4 DAYS) This course is designed for developers who create applications and analyze Big Data in Apache Hadoop on Windows using Pig and Hive. Topics include: Hadoop, YARN, the Hadoop Distributed File System (HDFS), MapReduce, Sqoop and the Hive ODBC Driver
HDP DEVELOPER: DEVELOPING APPLICATION ON YARN (2 DAYS) This course is designed for developers who want to create custom YARN applications for Apache Hadoop. It will include:

the YARN architecture, YARN development steps, writing a YARN client and Application Master, and launching Containers. The course uses Eclipse and Gradle connected remotely to a 7-node HDP cluster running in a virtual machine.
OPERATIONS & ADMINISTRATION CLASSES
HDP OPERATIONS: INSTALL & MANAGE WITH APACHE AMBARI (4 DAYS) This course is designed for administrators who will be managing the Hortonworks Data Platform (HDP). It covers installation, configuration, maintenance, security and Performance topics.
HDP OPERATIONS: MIGRATING TO THE HOROTNWORKS DATA PLATFORM (2 DAYS) This course is designed for administrators who are familiar with administering other Hadoop distributions and are migrating to the Hortonworks Data Platform (HDP). It covers installation, configuration, maintenance, security and performance topics.
DATA ANALYST TRAINING COURSES
HDP ANALYST: DATA SCIENCE (3 DAYS) This course Provides instruction on the processes and practice of data science, including machine learning and natural language processing. Included are: tools and programming languages (Python, IPython, Mahout, Pig, NumPy, Pandas, SciPy, Scikit-learn), the Natural Language Toolkit (NLTK), and SparkMLlib.
HDP ANALYST: HBASE (2 DAY WORKSHOP) This participatory workshop introduces HBase basics, structure and operations in an intensely hands-on experience guided by an HBase expert

3. A list of all other training courses you offer, and for each course, please provide the following information:

a. **Description of the training course;**
b. **Location of training course (e.g., online, classroom, onsite, etc.);**
c. **Limit on number of attendees (if applicable);**
d. **List of materials provided to State Farm as part of the training course;**
e. **Number of offerings per year;**
f. **Duration of the training course;**

g. Notice required to schedule the training course (on average); and
h. Any other information with respect to the training course you'd like to note.

Hortonworks University focuses on HDP training therefore additional training courses outside of the HDP platform are not offered. However, Hortonworks University does have partnerships with certified training partners that cover a wide range of training courses on other technology, if needed, an introduction can be made.

6. Agreements

APPENDIX C – STATE FARM MASTER TRAINING AGREEMENT

HW Legal Cho 2/27/15 5:24 PM
Comment [1]: @SF: We will need to attach our Training Services Policies re: cancellation. It is attached for your review.

MASTER TRAINING AGREEMENT

This is an agreement (the "Agreement") between **Error! Unknown document property name.** ("**Error! Unknown document property name.**"), located at _____, _____, _____, and State Farm Mutual Automobile Insurance Company, on behalf of itself, its subsidiaries and affiliates ("STATE FARM"), located at One State Farm Plaza, Bloomington, Illinois 61710.

RECITALS

WHEREAS, it is the desire of the parties that, in accordance with the terms and conditions hereof, the **Error! Unknown document property name.** shall, from time to time, develop and present a seminar to STATE FARM; and

WHEREAS, both parties wish to enter into an agreement which will govern the relationship of the parties.

ACCORDINGLY, the parties agree as follows:

EFFECTIVE DATE. This Agreement shall become effective on the date the second of the two parties executes this Agreement below.

TIME IS OF THE ESSENCE. The parties agree that time is of the essence, and that the transaction set out hereunder shall be completed by the date specified in Exhibit A.

HW Legal Cho 2/27/15 5:24 PM
Comment [2]: @SF: Unless it is mutually agreed to by the parties, we cannot agree to a time of the essence provision.

SERVICES PROVIDED. STATE FARM shall from time to time request **Error! Unknown document property name.** to present a seminar (the "Seminar") to instruct on certain topics. STATE FARM shall make its request on a work order, the form of which is attached hereto and incorporated herein as Exhibit A (the "Work Order"). Each Work Order shall be in the format prescribed in Exhibit A and shall set forth, among other things, the topics, the Seminar objectives, the Seminar dates, the Seminar location, the type and description of materials to be provided, and the fee to be paid to **Error! Unknown document property name.**.

DATES/LOCATIONS. Upon execution of the Work Order by both parties, **Error! Unknown document property name.** shall present the Seminar to such STATE FARM employees and/or independent contractors as STATE FARM shall determine, on the date(s) and location(s) set forth in the particular Work Order.

HW Legal Cho 2/27/15 5:24 PM
Comment [3]: @SF: Needs to be mutually agreed to.

MATERIALS. **Error! Unknown document property name.** shall provide in a quantity equal to the number of participants attending the Seminar, as set forth in Work Order, all necessary educational materials, including, but not limited to manuals, books, overheads, graphs, graphics. Unless otherwise specified in Work Order:

HW Legal Cho 2/27/15 5:24 PM
Comment [4]: @SF: Training Materials are already covered in the previously negotiated agreement.

All student course materials distributed to STATE FARM as part of the Seminar shall be the property of STATE FARM; and

All the materials in printed, audio, visual or machine readable form originated and prepared by the **Error! Unknown document property name.** prior to the effective date of this Agreement shall belong exclusively to the **Error! Unknown document property name.** and must be returned to the **Error! Unknown document property name.** upon completion of the Seminar; provided, however, that STATE FARM may incorporate any part of all of the Seminar material into any material distributed internally to STATE FARM; and provided further that STATE FARM may create or copy screen prints from any materials distributed, furnished, used or produced under this Agreement; and

The parties agree that any materials created or developed hereunder shall be considered a work made for hire. **Error! Unknown document property name.** further transfers, assigns, sells, and conveys to STATE FARM all rights to such materials created or produced hereunder, including but not limited to copyright, trademark, trade secret, and patent rights. Nothing in this section limits **Error! Unknown document property name.**'s rights in intangible know-how, ideas and concepts it learns or develops during the course of its performance under this Agreement.

INFRINGEMENT INDEMNIFICATION. **Error! Unknown document property name.** at its own expense shall defend and hold STATE FARM fully harmless against any action asserted against STATE FARM (and specifically including costs and reasonable attorneys' fees associated with any such action) to the extent it is based on a claim that use of any materials or other things besides services provided to STATE FARM under this Agreement (collectively, the "Deliverables") or a Seminar or other services provided by **Error! Unknown document property name.** (collectively, the "Services") under this Agreement infringes any patent, copyright, license or other proprietary right of any third party. STATE FARM shall promptly notify **Error! Unknown document property name.** in writing of any such claim. If as a result of any claim of infringement against any patent, copyright, license or other proprietary right of any third party, STATE FARM is enjoined from using the Deliverables or Services, or if **Error! Unknown document property name.** believes that the Deliverables or Services are likely to become the subject of a claim of infringement, **Error! Unknown document property name.** at its option and expense will procure the right for STATE FARM to continue to use the Deliverables or Services, or replace or modify the Deliverables or Services so as to make them non-infringing.

HW Legal Cho 2/27/15 5:24 PM
Comment [5]: @SF: Already covered in the previously negotiated agreement.

FEES. In consideration of the services and materials provided hereunder by **Error! Unknown document property name.**, STATE FARM shall pay to **Error! Unknown document property name.** the fee set forth in the particular Work Order. Such fee shall be due and payable at the later of (i) thirty (30) days following the presentation of the Seminar or (ii) thirty (30) days following STATE FARM's receipt of an accurate invoice.

CONFIDENTIALITY.

HW Legal Cho 2/27/15 5:24 PM
Comment [6]: @SF: Already covered in the previously negotiated agreement.

HW Legal Cho 2/27/15 5:24 PM
Comment [7]: @SF: Already covered in previously negotiated Agreement.

The parties expressly acknowledge that in the course of their performance hereunder, they may learn or have access to certain confidential, patent, copyright, business, trade secret, proprietary or other like information or products of the other party or of third parties, including but not limited to the other party's vendors, consultants, suppliers or customers (the "Information"). Anything in the Agreement to the contrary notwithstanding, the parties expressly agree that they will keep strictly confidential any such Information.

STATE FARM and **Error! Unknown document property name.** agree that, for the purposes of the Agreement, third parties whose duties for STATE FARM, or as a subcontractor for **Error! Unknown document property name.** in performing **Error! Unknown document property name.**'s duties under this Agreement, require access to the Information provided under the Agreement shall have access to the Information as required by such duties, provided that: (i) such third parties have agreed in writing with either STATE FARM or **Error! Unknown document property name.**, in terms no less protective than the confidentiality obligations of the Agreement, to keep confidential the Information; (ii) such third parties have agreed in writing with either STATE FARM or **Error! Unknown document property name.** not to use the Information for their own benefit or the benefit of any person or entity besides STATE FARM; and (iii) STATE FARM, when allowing such third parties access to **Error! Unknown document property name.**'s Information, will not exceed the license or use restrictions in the Agreement.

Error! Unknown document property name. agrees not to use STATE FARM's or a STATE FARM third party's Information for its own benefit or the benefit of any person besides STATE FARM.

The term "Disclosing Party" shall refer to the party to the Agreement providing the Information to the other party, and the term "Receiving Party" shall refer to the party receiving the Information in the course of its performance under the Agreement. The term "Information" shall not include products or information that: (i) are in the public domain or in the possession of the Receiving Party without restriction at the time of receipt under the Agreement; (ii) are used or released with the prior written approval of the Disclosing Party; or (iii) are independently developed by the Receiving Party.

It shall not be a violation of this Confidentiality section for the Receiving Party to disclose Information if a court of competent jurisdiction or appropriate regulatory authority demands such disclosure. In such case, prior to disclosing the Information, the Receiving Party will: (i) notify the Disclosing Party immediately; and (ii) will cooperate with the Disclosing Party in asserting a confidential or protected status for the Information.

Each party expressly further agrees that, at the sole discretion and request of the Disclosing Party, it shall either return to the Disclosing Party or destroy any such Information and copies thereof; and it will certify any such destruction in writing to the Disclosing Party.

PERFORMANCE STANDARDS. **Error! Unknown document property name.** represents and warrants that it has the ability and expertise to perform its responsibilities hereunder and in doing so shall use the highest standards. STATE FARM shall have the right to reject any of **Error! Unknown document property name.**'s employees or agents whose qualifications, in STATE FARM's judgment, do not meet the standards hereunder. **Error! Unknown document property name.** agrees that STATE FARM may at any time and for any reason, other than an unlawful reason, terminate the assignment on work involving STATE FARM of an individual employee or agent of TRAINER.

CLICK-THROUGH AGREEMENTS/ON-LINE TERMS. No terms and conditions related to the subject matter of this Agreement and presented at any time in a "click-through" or "click-wrap" agreement or web site shall apply to such subject matter.

USE OF STATE FARM NAME. **Error! Unknown document property name.** expressly agrees that it shall not disclose or otherwise identify STATE FARM orally or in any of its advertising, publications, or other media that are displayed or disseminated to its customers or other parties.

COOPERATION. **Error! Unknown document property name.** and STATE FARM each acknowledge that it shall be necessary to cooperate in order to carry out this Agreement, and each agrees to reasonably cooperate with the other.

BACKGROUND CHECKS/RESTRICTIONS.

HW Legal Cho 2/27/15 5:24 PM
Comment [8]: @SF: Already covered in previously negotiated agreement.

Error! Unknown document property name. agrees that prior to providing its employees or agents to STATE FARM under this Agreement, it will complete an appropriate background check, including: (i) a review of all felony and misdemeanor convictions in the county in which the employee or agent has lived longest or where he/she currently resides; (ii) an employment history for all employers of the employee or agent for the past five years or the three most recent employers; and (iii) if the employee or agent will be operating a STATE FARM vehicle, a review of his/her driving record and driver's license.

In no event will **Error! Unknown document property name.**, in the performance of this Agreement, use the services of an employee or agent who has been convicted of a felony involving dishonesty or a breach of trust without first obtaining the written consent of the Illinois Department of Insurance or other appropriate regulatory authority in order to comply with the provisions of 18 U.S. Code Section 1033.

INDEPENDENT CONTRACTOR. The parties expressly agree that **Error! Unknown document property name.** shall be an independent contractor for all purposes in the performance of this Agreement, and that none of its employees or agents shall be considered an employee of STATE FARM for any purpose. **Error! Unknown document property name.** shall be responsible for compliance with all tax, workers compensation and other applicable laws or regulations. **Error! Unknown document property name.** accepts exclusive liability for all contributions and payroll taxes payable under federal and state laws and regulations governing social security or old age benefits, unemployment insurance, and workers compensation as to persons performing this Agreement.

INSURANCE. a. **Error! Unknown document property name.** shall secure, pay the premium for, and keep in force until the termination of this Agreement, the following insurance: (i) Workers' Compensation at statutory limits for occupational disease and injury coverage, including Employer's Liability coverage at a limit of not less than $500,000; and (ii) Commercial General Liability, including Premises and Operations coverage and Products and Completed Operations coverage at not less than a combined single limit of $1,000,000 per occurrence.

HW Legal Cho 2/27/15 5:24 PM
Comment [9]: @SF: From Insurance and on, these provisions have been previously addressed in our agreement.

b. Prior to performing under this Agreement, or upon STATE FARM's request, **Error! Unknown document property name.** shall provide STATE FARM a Certificate(s) of Insurance verifying that it has these insurance coverages. **Error! Unknown document property name.** must provide at least a 30-day notice to STATE FARM of a material change to or termination of a policy. State Farm Mutual Automobile Insurance Company will be listed on the Certificate as an additional insured under **Error! Unknown document property name.**'s Commercial General Liability policy.

APPLICABILITY TO SUBCONTRACTORS. **Error! Unknown document property name.** shall ensure that its subcontractors performing hereunder also adhere to the applicable provisions of this Agreement.

HOLD HARMLESS. Anything in the Agreement to the contrary notwithstanding, **Error! Unknown document property name.** shall indemnify and hold STATE FARM fully harmless against any loss, damages, claims, penalties, or expenses of any kind whatsoever (including costs and reasonable attorneys' fees), sustained or incurred by a third party as a result of the negligent or intentional acts or omissions of **Error! Unknown document property name.**, and for which recovery is sought against STATE FARM by that third party. **Error! Unknown document property name.** also shall indemnify STATE FARM for any costs and reasonable attorneys'

fees sustained or incurred by STATE FARM in the defense of any such third party claim.

LIMITATION OF LIABILITY. ANYTHING IN THE AGREEMENT TO THE CONTRARY NOTWITHSTANDING, UNDER NO CIRCUMSTANCES WHATSOEVER SHALL STATE FARM BE LIABLE TO **Error! Unknown document property name.** FOR ANY SPECIAL, CONSEQUENTIAL, PUNITIVE, INDIRECT, OR INCIDENTAL DAMAGES OF ANY KIND WHATSOEVER. IN NO EVENT WHATSOEVER SHALL STATE FARM'S TOTAL LIABILITY TO **Error! Unknown document property name.** FOR ANY OTHER DAMAGES WHATSOEVER EXCEED IN THE AGGREGATE THE SUM OF TWENTY-FIVE-THOUSAND DOLLARS DOLLARS ($25,000.00).

TERMINATION. If either party neglects or fails to perform any of its obligations under this Agreement and such failure continues for a period of ten (10) days after written notice thereof, the other party shall have the right to terminate this Agreement. If **Error! Unknown document property name.** assigns the Agreement to another entity without receiving STATE FARM's prior written approval, or if there is an affiliation, merger, or acquisition involving **Error! Unknown document property name.**, STATE FARM may terminate the Agreement upon providing thirty (30) days' prior written notice.

NO WAIVER OF BREACH. If either party on any occasion breaches any term of this Agreement, and the other party does not enforce that term, the failure to enforce on that occasion shall not prevent enforcement on any other occasion.

ASSIGNMENT. Anything in the Agreement to the contrary notwithstanding, **Error! Unknown document property name.** may not assign the Agreement to any other entity, including an entity that affiliates or merges with or acquires **Error! Unknown document property name.**, except when such assignment is approved in advance by STATE FARM in writing, which approval STATE FARM may in its sole discretion grant or deny.

FORCE MAJEURE. Neither party shall be liable for any delays in performance hereunder due to circumstances beyond its control including, but not limited to, acts of nature, acts of governments, delays in transportation, and delays in delivery or inability of suppliers to deliver. STATE FARM shall have the option to terminate any and all obligations under this Agreement as amended by so notifying the **Error! Unknown document property name.** in writing if the delay in performance exceeds ten (10) calendar days from the originally agreed upon performance date.

DISPUTE RESOLUTION.

Any dispute arising out of or relating to the Agreement shall be submitted to arbitration under the International Institute for Conflict Prevention and Resolution Rules for Non-

Administered Arbitration (except it shall be non-binding) by three arbitrators. Each party shall appoint one arbitrator, and those two arbitrators shall choose the third arbitrator. The United States Arbitration Act, 9 USC §§1 et seq. shall govern the arbitration, and judgment may be entered by any court having jurisdiction thereof. The place of arbitration shall be Chicago, Illinois. The arbitrators are empowered to award damages in accordance with the Limitation of Liability section of this Agreement, and may not award punitive damages.

After the non-binding arbitration decision is rendered, either party may initiate litigation upon thirty (30) days' written notice to the other party; provided, however, that either party may seek injunctive relief at any time.

CHOICE OF LAW. This Agreement shall be governed by the laws of the State of Illinois without regard to its conflict of law rules.

WORK ORDERS. The parties agree that whether or not a Work Order has been signed, they shall nevertheless adhere to the provisions of this Agreement.

SURVIVAL. The following Sections shall survive termination of this Agreement: Materials, Infringement Indemnification, Confidentiality, Use of State Farm Name, Independent Contractor, Applicability to Subcontractors, Hold Harmless, Limitation of Liability, Dispute Resolution, Choice of Law and Survival.

This Agreement (and any attachments, addenda, and supplements thereto) shall be the complete and exclusive statement of the agreement between the parties as to the subject matter of this Agreement, and shall be binding upon each of the parties hereto, their respective successors, and to the extent permitted their assigns. This Agreement cannot be amended or otherwise modified, except as agreed to in writing by each of the parties hereto.

Error! Unknown document property name.	State Farm Mutual Automobile Insurance Company
_____	_____
Signature	Signature
_____	_____
Printed or Typed Name	Printed or Typed Name
_____	_____
Title	Title
_____	_____
Date	Date

EXHIBIT A

WORK ORDER

This Work Order defines certain seminars or training to be presented by **Error! Unknown document property name.** ("**Error! Unknown document property name.**") in accordance with the terms and conditions of that certain Master Training Agreement #_____ between State Farm Mutual Automobile Insurance Company, on behalf of itself, its subsidiaries and affiliates, ("STATE FARM"), located at One State Farm Plaza, Bloomington, Illinois 61710, and **Error! Unknown document property name.**, located at **Error! Unknown document property name.**, **Error! Unknown document property name.**, **Error! Unknown document property name. Error! Unknown document property name.**.

Job Number:
#_____

Services:

Seminar Name:

Seminar Date(s):

Seminar Location(s):

Number Attending Seminar:

Seminar Topics:

Seminar Objectives:

Seminar Schedule:

Type and Description of Materials to be developed/provided:

Compensation:

Daily Rate(s) to be charged:

Travel Expenses:

Total Fee:

Invoicing:

All invoices submitted under this Work Order should be sent to the following address, and should reference Work Order Number _____:

State Farm Mutual Automobile Insurance Company
Attn: Payment Team
One State Farm Plaza, A-1
Bloomington, IL 61710

Error! Unknown document property name. and STATE FARM agree that the referenced Agreement and this Work Order are the complete and exclusive statement of the agreement between the parties, superseding all proposals or prior agreements, oral or written, and all other communications between the parties relating to the subject matter hereof. In the event of any conflicts between this Work Order and the Agreement, the terms of the Agreement shall prevail. This Work Order shall be effective on the date the second of the two parties hereto signs below.

Error! Unknown document property name.	State Farm Mutual Automobile Insurance Company
_____	_____
Signature	Signature
_____	_____
Printed or Typed Name	Printed or Typed Name
_____	_____
Title	Title
_____	_____
Date	Date

7. Risks & Resolutions

Based on our understanding of the project, the table below summarizes few potential risks and mitigation plan. The Hortonworks project manager will work with the State Farm project manager to review and monitor identified risks and deploy the appropriate mitigation plan throughout the engagement.

Brief Description	Severity	Mitigation
Cluster available prior to engagement commencement	H	Ensure that cluster pre-requisites are delivered to State Farm and reviewed with Infrastructure team prior to engagement commencement
Complex service and business logic	H	Ensure State Farm SME availability throughout project lifecycle for clarification and detail
Projects within State Farm competing for resources/services	H	Working with State Farm Project Manager, identify any State Farm projects that could impact timeline, work streams, and resources
Increased overheads for State Farm operational team due to potential challenges in technology and service	M	Ensure that operational considerations are accounted for and links to Hortonworks online documentation is distributed early on
Several integration points (heavy dependency on API) could be a challenge, particularly while implementing security controls	M	Begin unit and integration testing early in the implementation based to resolve potential issues in time
Unclear rules of engagement across different teams and vendors	L	Clearly define communication framework and responsibilities
Pending decisions – what tool to use	L	Address early during Architecture
Compliance with State Farm	L	Review State Farm guidance &

Processes and Standards		standards early on
Hortonworks proposed solution would leverage Apache Solr for Index Store. Apache Solr as pure open source offering has little to no built-in security. The solution we recommend will enforce security model and role-based access control at the application layer. In our assessment, this may be sufficient to fulfill perceived security needs. But if we identify corner cases where we absolutely can't address security concerns at application layer, then we need to explore options from our partner ecosystem.	L	Apache Solr is backed by commercial entity called Lucidworks. Lucidworks is a Hortonworks certified partner, which ships commercial offering called Fusion to extend any Apache Solr deployment with enterprise-grade security.
Patches and upgrade may potentially break the solution if not planned well, due to several moving parts	L	Automation and testing of system deployment process