Mphasis
The Next Applied

Metadata management on a Hadoop Eco-System

Whitepaper by

Satya Nayak
Project Lead, Mphasis

Jansee Korapati
Module Lead, Mphasis

Introduction

The data lake stores large amount of structured and unstructured data in various varieties at different transformed layers. While the data is growing to terabytes and petabytes, and your data lake is being used by the enterprise, you are likely to come across questions/ challenges, such as what data is available in the data lake, how it is consumed/prepared/transformed, who is using this data, who is contributing to this data, how old is the data… etc.

A well maintained metadata layer can effectively answer these kind of queries and thus im-prove the usability of the data lake. This white paper provides the benefits of an effective metadata layer for a data lake implemented using Hadoop Cluster; information on various metadata Management tools is presented, with their features and architecture.

Benefits and Functions of Metadata Layer

- The metadata layer captures vital information about the data as it enters the data lake and indexes the information so that users can search metadata before they access the data it-self. Capturing metadata is fundamental to make data more accessible and to extract value from the data lake.
- The metadata layer provides significant information about the background and significance of the data stored in the data lake to its users. This is accomplished by intelligently tagging every bit of data as it is ingested.

The Metadata layer

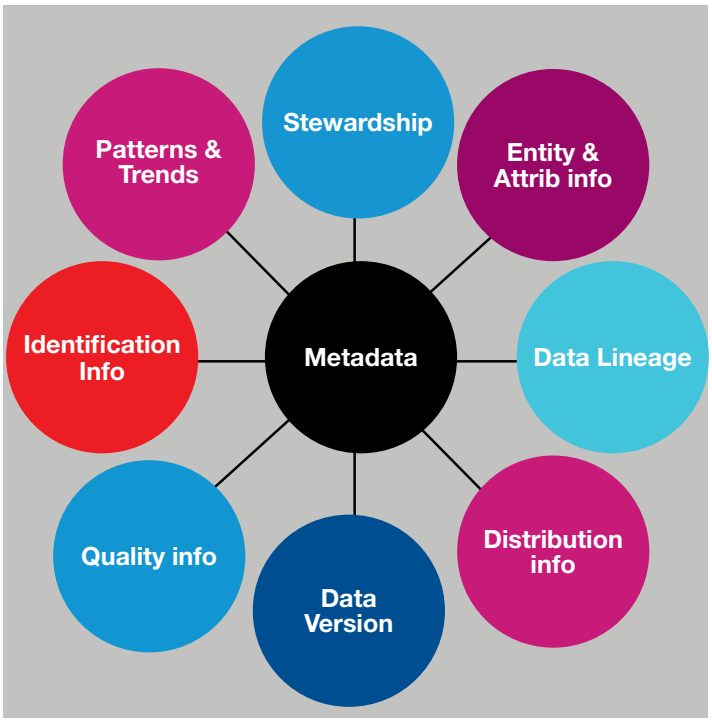

Fig. 1

- A well-built metadata layer will allow organization to harness the potential of data lake and deliver the following mechanisms to the end users to access data and perform analysis:
  - Self Service BI (SSBI)
  - Data-as-a-Service (DaaS)
  - Machine Learning-as-a-Service
  - Data Provisioning (DP)

You can optimize your data lake to the fullest with metadata management.

- The metadata Layer defines the structure of files in Raw Zone and describes the entities in-side the file. Using the base level description, the schema evolution of the file or record is tracked by a versioning schema. This will eventually allow you to create association among several entities and thereby, facilitate browsing and searching the data that the end user is looking for.
- In the consumption layer, it is very convenient to know the source of the data while going through a report as there might be different version of the input data.
- Clarity of relationships in metadata helps in resolving ambiguity and inconsistencies when determining the associations between entities stored throughout data environment.

Where it fits in the System?

Metadata layer is a general purpose component which will be used across different layers and captures data about data from various layers. Let us take an example of typical data flow of a data lake. Data is ingested from disparate source systems like databases, streaming logs, sen-sor data and social media and initially stored in transient zones. It, then, goes through various phases like data validations checks, business rule transformations and finally ends up in the presentation layer. The metadata is captured during all these phases in different layers.

Metadata layer can access data from multiple layers.

Various Metadata Management Tools

Here is the list of different metadata management tools that can capture metadata on a Hadoop cluster. There is no order of preference for this list. These tools are listed in a random order.

- **Cloudera Navigator:** Cloudera Navigator is a data governance solution for Hadoop, offering critical capabilities such as data discovery, continuous optimization, audit, lineage, metadata management, and policy enforcement. As part of ClouderaEnterprise, Cloudera Navigator is critical to enabling high-performance agile analytics, supporting continuous data architecture optimization, and meeting regulatory compliance requirements.
- **Apache Atlas:** Currently in Apache incubator, this is a scalable and extensible component which can create, automate and define relationship on data for metadata in the data lake System. You can also export metadata to third party system from Atlas. It can be used for data discovery and lineage tracking.
- **Apache Falcon:** Falcon is aimed at making the feed processing and feed management on Hadoop clusters easier for their end consumers.
- **HCatlog:** HCatalog is a table and storage management layer for Hadoop that enables users with different data processing tools like Pig and MapReduce to read and write data on the grid more easily.
- **Loom:** Loom provides metadata management and data preparation for Hadoop. The core of Loom is an extensible metadata repository for managing business and technical metadata, including data lineage, for all the data in Hadoop and surrounding systems. Loom's active scan framework automates the generation of metadata for Hadoop data by crawling HDFS to discover and introspect new files.
- **Waterline:** Waterline Data automates the creation and management of an inventory of data assets at the field level, empowering data architects to provide all the data the business needs through secure self-service. It ensures data governance policies are adhered to, by enabling data stewards to audit data lineage, protect sensitive data, and identify compliance issues.
- **Ground Metadata (AMP Lab):** Ground is a data context system, under development at University of California, Berkeley. It is aimed at building a flexible, open source, vendor neu-tral system that enables users to classify about what data they have, where that data is flowing to and from, who is using the data, when the data changed, and why and how the data is changing. Among other things, we believe a data context system is particularly use-ful for data inventory, data usage tracking, model-specific interpretation, reproducibility, interoperability, and collective governance.

You can choose from metadata management tools widely available, as per your business requirement.

Features of architecture of commonly used Metadata Tools

1. Cloudera Navigator

Cloudera Navigator is a proprietary tool from Cloudera for data management in Hadoop Eco-System. It primarily provides two solutions in the area of Data Governance.

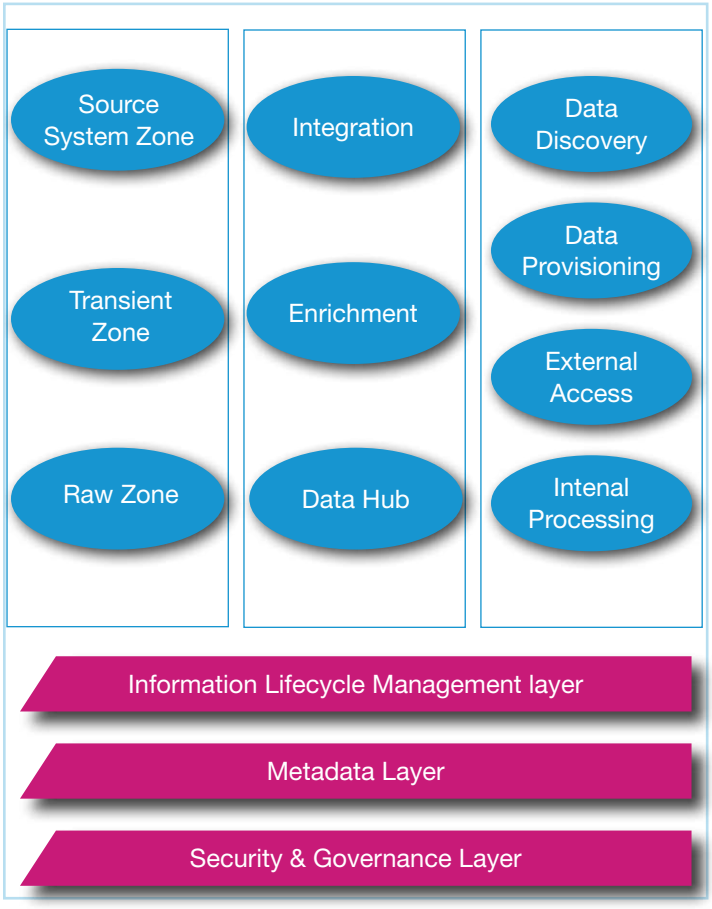

Fig. 2

Data Management

Data management provides visibility into and control over the data residing in Hadoop data stores and the computations performed on that data. The features included here are:

Various Metadata Management Tools

Here is the list of different metadata management tools that can capture metadata on a Hadoop cluster. There is no order of preference for this list. These tools are listed in a random order.

- **Cloudera Navigator:** Cloudera Navigator is a data governance solution for Hadoop, offering critical capabilities such as data discovery, continuous optimization, audit, lineage, metadata management, and policy enforcement. As part of Cloudera Enterprise, Cloudera Navigator is critical to enabling high-performance agile analytics, supporting continuous data architecture optimization, and meeting regulatory compliance requirements.
- **Apache Atlas:** Currently in Apache incubator, this is a scalable and extensible component which can create, automate and define relationship on data for metadata in the data lake System. You can also export metadata to third party system from Atlas. It can be used for data discovery and lineage tracking.
- **Apache Falcon:** Falcon is aimed at making the feed processing and feed management on Hadoop clusters easier for their end consumers.

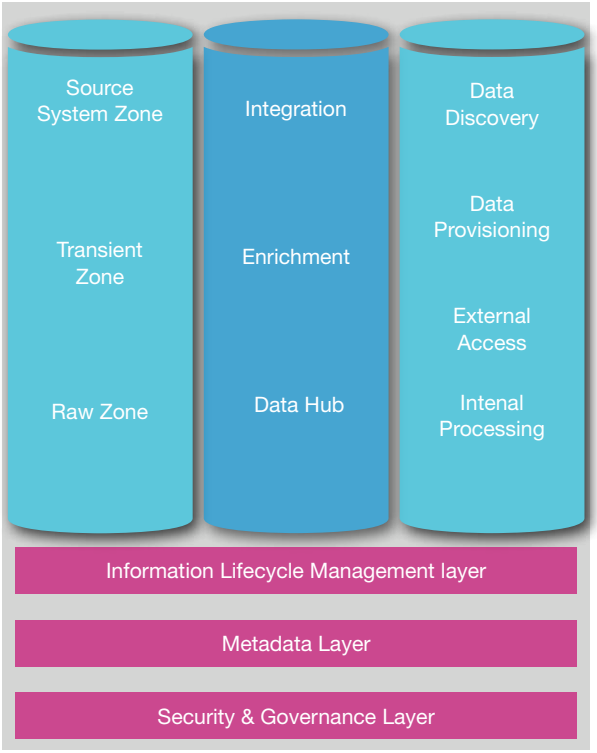

Fig. 2

- **HCatlog:** HCatalog is a table and storage management layer for Hadoop that enables users with different data processing tools like Pig and MapReduce to read and write data on the grid more easily.
- **Loom:** Loom provides metadata management and data preparation for Hadoop. The core of Loom is an extensible metadata repository for managing business and technical metadata, including data lineage, for all the data in Hadoop and surrounding systems. Loom's active scan framework automates the generation of metadata for Hadoop data by crawling HDFS to discover and introspect new files.
- **Waterline:** Waterline Data automates the creation and management of an inventory of data assets at the field level, empowering data architects to provide all the data the business needs through secure self-service. It ensures data governance policies are adhered to, by enabling data stewards to audit data lineage, protect sensitive data, and identify compliance issues.
- **Ground Metadata (AMP Lab):** Ground is a data context system, under development at University of California, Berkeley. It is aimed at building a flexible, open source, vendor neu-tral system that enables users to classify about what data they have, where that data is flowing to and from, who is using the data, when the data changed, and why and how the data is changing. Among other things, we believe a data context system is particularly use-ful for data inventory, data usage tracking, model-specific interpretation, reproducibility, interoperability, and collective governance.

Features of Architecture of Commonly Used Metadata Tools

1. Cloudera Navigator

Cloudera Navigator is a proprietary tool from Cloudera for data management in Hadoop Eco-System. It primarily provides two solutions in the area of Data Governance.

Data Management

Data management provides visibility into and control over the data residing in Hadoop data stores and the computations performed on that data. The features included here are:

- **Auditing** data access and verifying access privileges: The goal of auditing is to capture a complete and immutable record of all the activities within a system. Cloudera Navigator auditing features add secured, real-time audit components to key data and access frame-works. Cloudera Navigator allows compliance groups to configure, collect, and view audit events, and understand who accessed what data and how.
- **Searching** metadata and visualizing lineage - Cloudera Navigator metadata management features allow DBAs, data stewards, business analysts, and data scientists to define, search, and amend the properties, and tag data entities and view relationships between datasets.
- **Policies** - Cloudera Navigator policy features enable data stewards to specify automated ac-tions based on data access or on a schedule to add metadata, create alerts, and move or purge data.
- **Analytics** - Cloudera Navigator analytics features enable Hadoop administrators to examine data usage patterns and create policies based on those patterns.

Data Encryption

Data encryption and key management provide a critical layer of protection against potential threats by malicious actors on the network or in the datacenter. Encryption and key manage-ment are also required for meeting key compliance initiatives and ensuring the integrity of your enterprise data. The following Cloudera Navigator components enable compliance groups to manage encryption:

- Cloudera Navigator Encrypt transparently encrypts and secures data at rest without requiring changes to your applications and ensures there is minimal performance lag in the encryption or decryption process.
- Cloudera Navigator Key Trustee Server is an enterprise-grade virtual safe-deposit box that stores and manages cryptographic keys and other security artifacts.
- Cloudera Navigator Key HSM allows Cloudera Navigator Key Trustee Server to seamlessly integrate with a hardware security module (HSM).

Cloudera Navigator Metadata Architecture

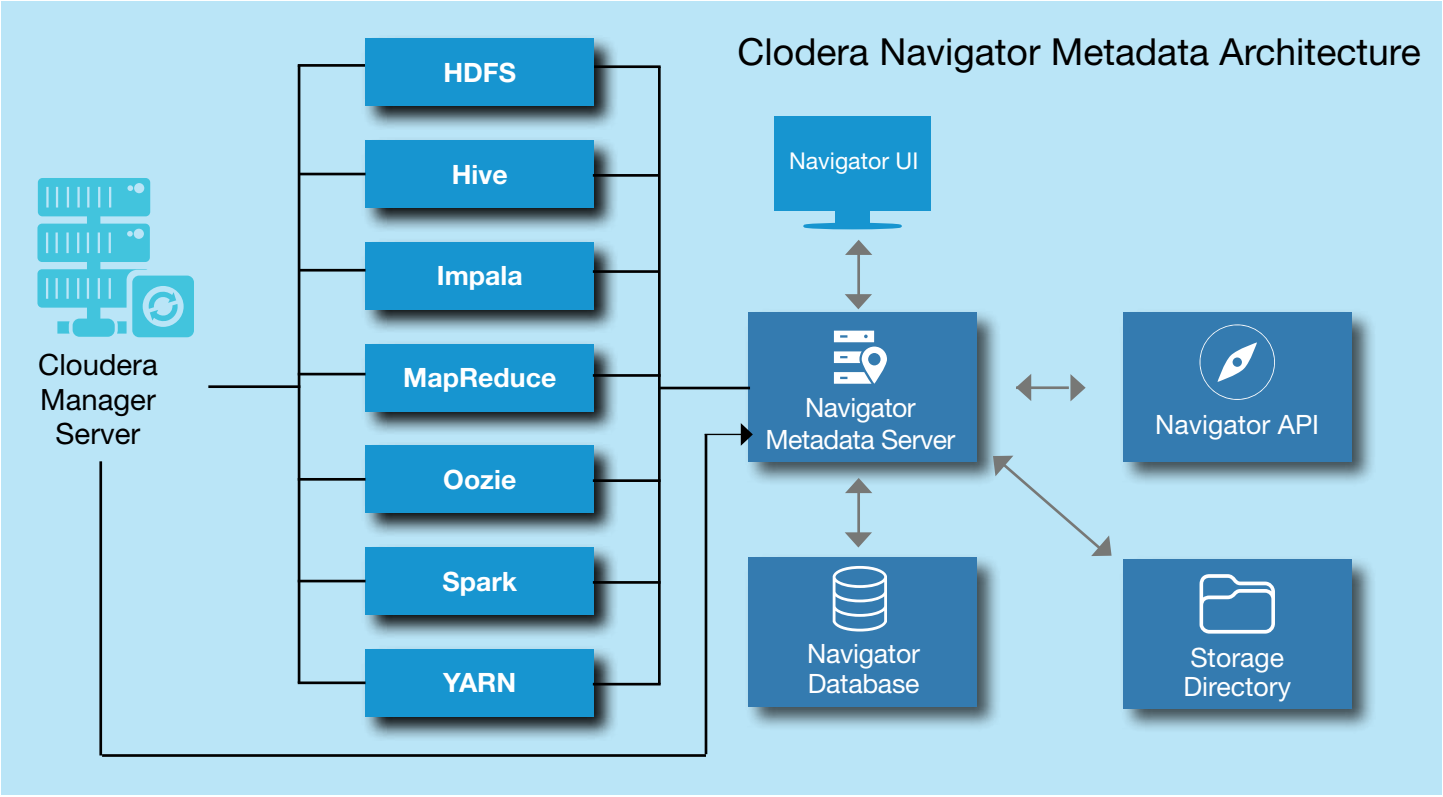

Fig. 3

The **Navigator Metadata Server** performs the following functions:

- Obtains connection information about CDH services from the Cloudera Manager Server
- At periodic intervals, extracts metadata for the entities managed by those services
- Manages and applies metadata extraction policies during metadata extraction
- Indexes and stores entity metadata
- Manages authorization data for Navigator users
- Manages audit report metadata
- Generates metadata and audit analytics
- Implements the Navigator UI and API

2. Apache Atlas

Atlas is a Data Governance initiative from Hortonworks on Hadoop cluster. It was initially started from Hortonworks and then taken over to Apache as a top level project. Atlas is a scal-able and extensible set of core foundational governance services – enabling enterprises to ef-fectively and efficiently meet their compliance requirements within Hadoop and allows integra-tion with the whole enterprise data ecosystem.

Features

Data Classification

- Import or define taxonomy business-oriented annotations for data
- Define, annotate, and automate capture of relationships between datasets and underly-ing elements including source, target, and derivation processes
- Export metadata to third-party systems

Centralized Auditing

- Capture security access information for every application, process, and interaction with data
- Capture the operational information for execution, steps, and activities

Search & Lineage (Browse)

- Pre-defined navigation paths to explore the data classification and audit information
- Text-based search features locate relevant data and audit event across data lake quickly and accurately
- Browse visualization of data set lineage allowing users to drill-down into operational, security, and provenance related information

Security & Policy Engine

- Rationalize compliance policy at runtime based on data classification schemes, attrib-utes and roles.
- Advanced definition of policies for preventing data derivation based on classification (i.e. re-identification) – Prohibitions
- Column and Row level masking based on cell values and attributes.

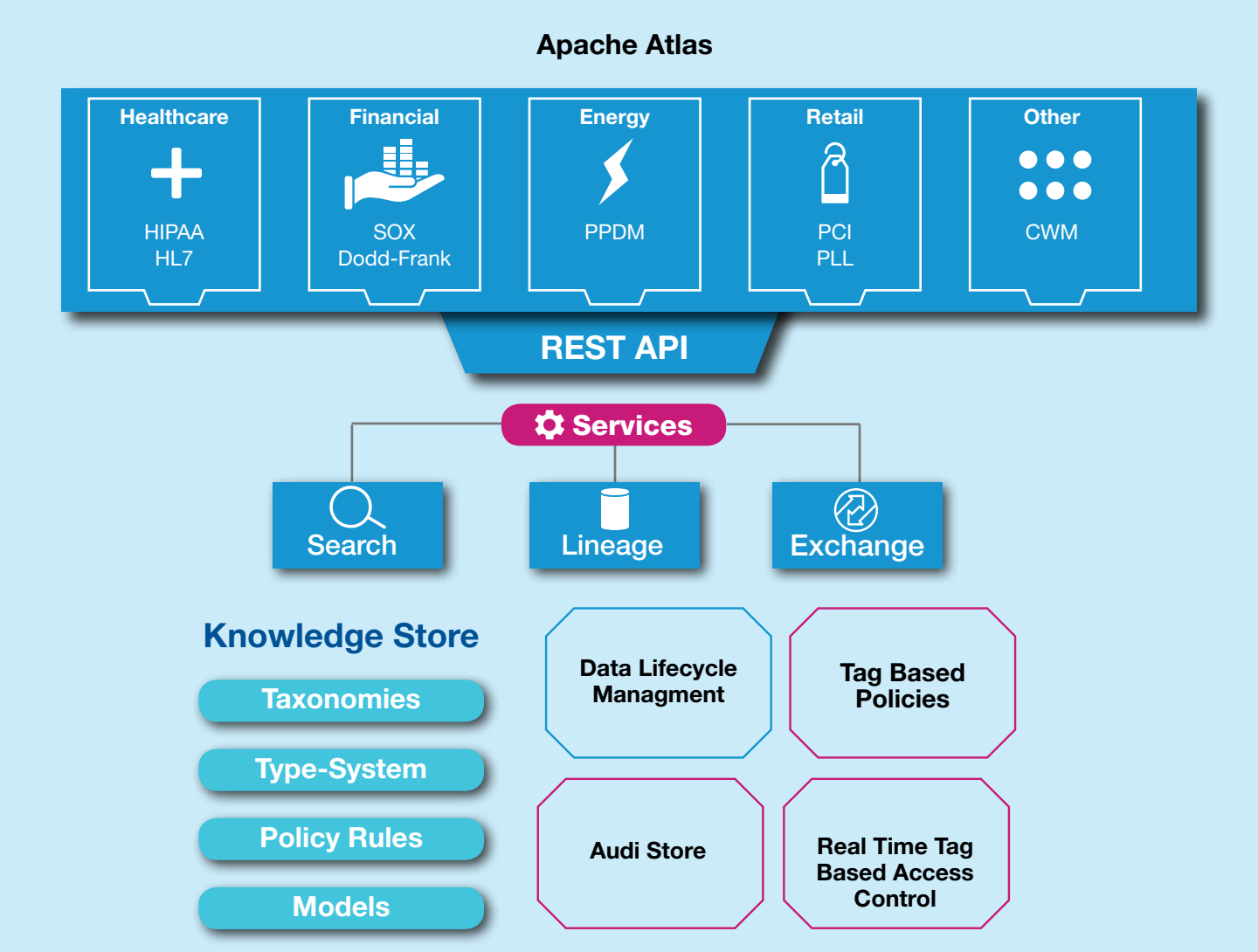

Fig. 4

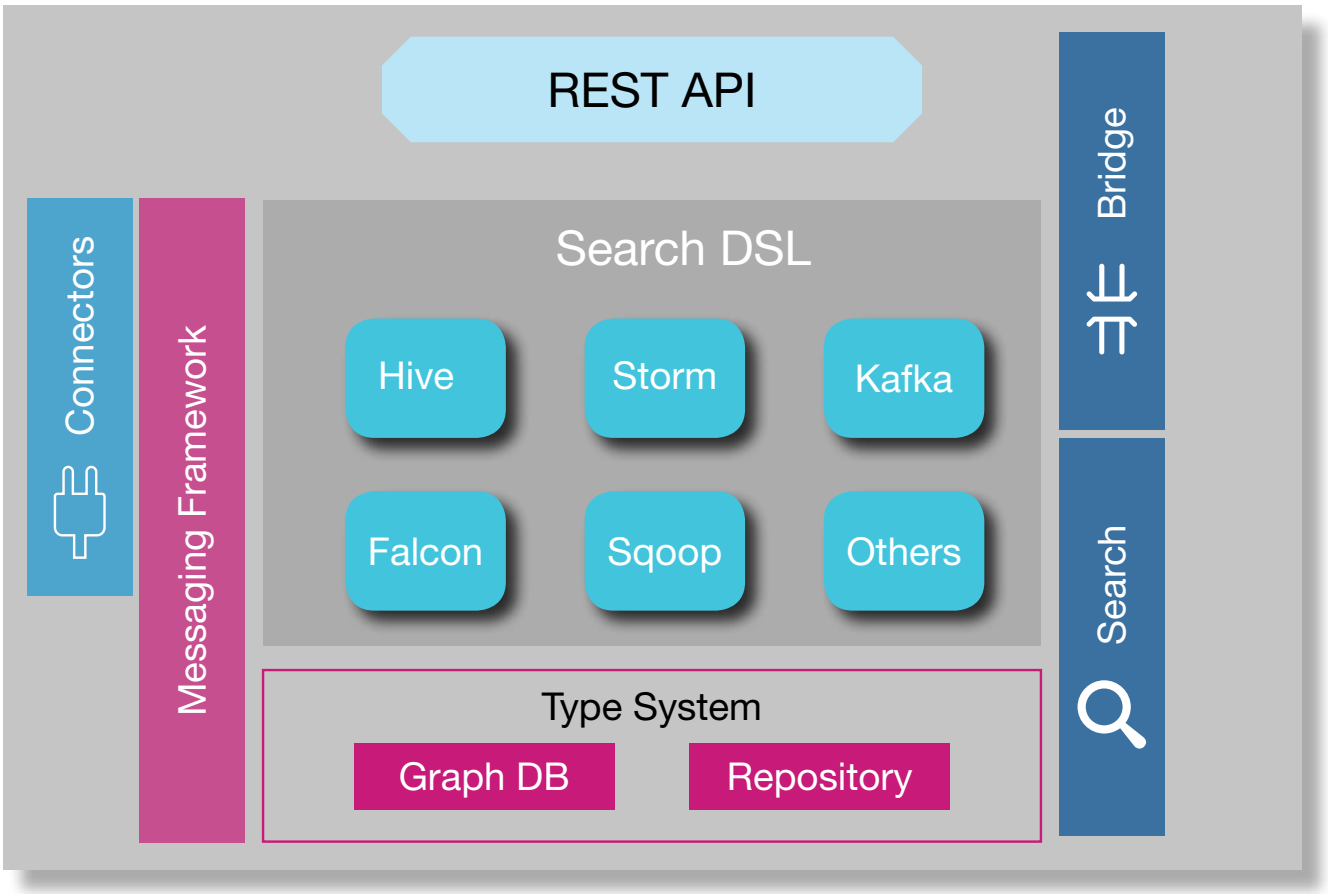

Fig. 5

In terms of implementation, Atlas has the following components to accomplish the design.

- **Web service:** This exposes RESTful APIs and a web user interface to create, update and query metadata.
- **Metadata store:** Metadata is modeled using a graph, implemented using the Graph database Titan. Titan has options for a variety of backing stores for persisting the graph, including an embedded Berkeley DB, Apache HBase and Apache Cassandra. The choice of the backing store determines the level of service availability.
- **Index store:** For powering full text searches on metadata, Atlas also indexes the metadata, again via Titan. For the full text search feature, it can use backend systems like Elastic Search or Apache Solr.
- **Bridges/Hooks:** To add metadata to Atlas, libraries called 'hooks' are enabled in various systems like Apache Hive, Apache Falcon and Apache Sqoop, which capture metadata events in the respective systems and propagate them to Atlas. The Atlas server consumes these events and updates its stores.
- **Metadata notification events:** Any updates to metadata in Atlas, either via the Hooks or the API, are propagated from Atlas to downstream systems via events. Systems like Apache Ranger consume these events and allow administrators to act on them, for e.g. to configure policies for Access control.
- **Notification server:** Atlas uses Apache Kafka as a notification server for communica-tion between hooks and downstream consumers of metadata notification events. Events are written by the hooks and Atlas to different Kafka topics. Kafka enables a loosely coupled integration between these disparate systems.

Bridges/Hook

External components like hive/sqoop/storm/falcon should model their taxonomy using type system and register the types with Atlas. For every entity created in this external component, the corresponding entity should be registered in Atlas as well. This is typically done in a hook, which runs in the external component and is called for every entity operation. Hook generally processes the entity asynchronously using a thread pool to avoid adding latency to the main operation.

Atlas exposes notification interface and can be used for reliable entity registration by hook as well. The hook can send notification message containing the list of entities to be registered. Atlas service contains hook consumer that listens to these messages and registers the entities.

Notification Server Design:

Notification is used for reliable entity registration from hooks and for entity/type change notifications. Atlas, by default, provides Kafka integration, but it's possible to provide other implementations as well. Atlas service starts embedded Kafka server by default. Atlas also provides NotificationHookConsumer that runs in Atlas Service and listens to messages from hook and registers the entities in Atlas.

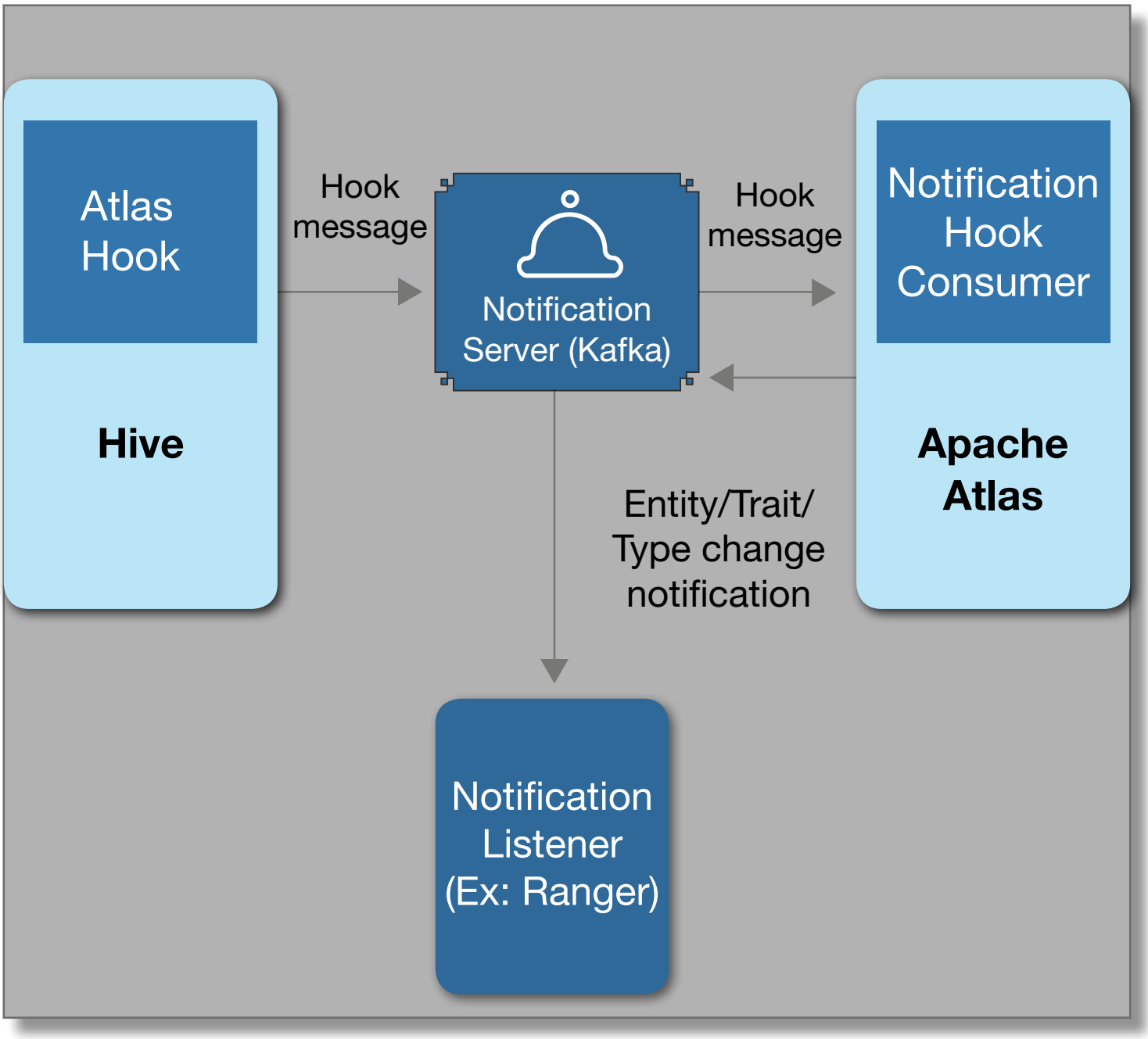

Fig. 6

3. Apache Falcon

Apache Falcon addresses enterprise challenges related to Hadoop data replication, business continuity, and lineage tracking by deploying a framework for data management and process-ing. Falcon centrally manages the data lifecycle, facilitates quick data replication for business continuity and disaster recovery and provides a foundation for audit and compliance by track-ing entity lineage and collection of audit logs.

Apache Falcon is a framework to simplify data pipeline processing and management on Hadoop clusters. It makes on-boarding new workflows/pipelines simpler, with support for late data handling and retry policies. It allows user to easily define relationships between various data and processing elements and integrate with metastore/catalog such as Apache Hive/HCatalog. Finally it also captures the lineage information for feeds and processes.

Following is the high level architecture of Apache Falcon.

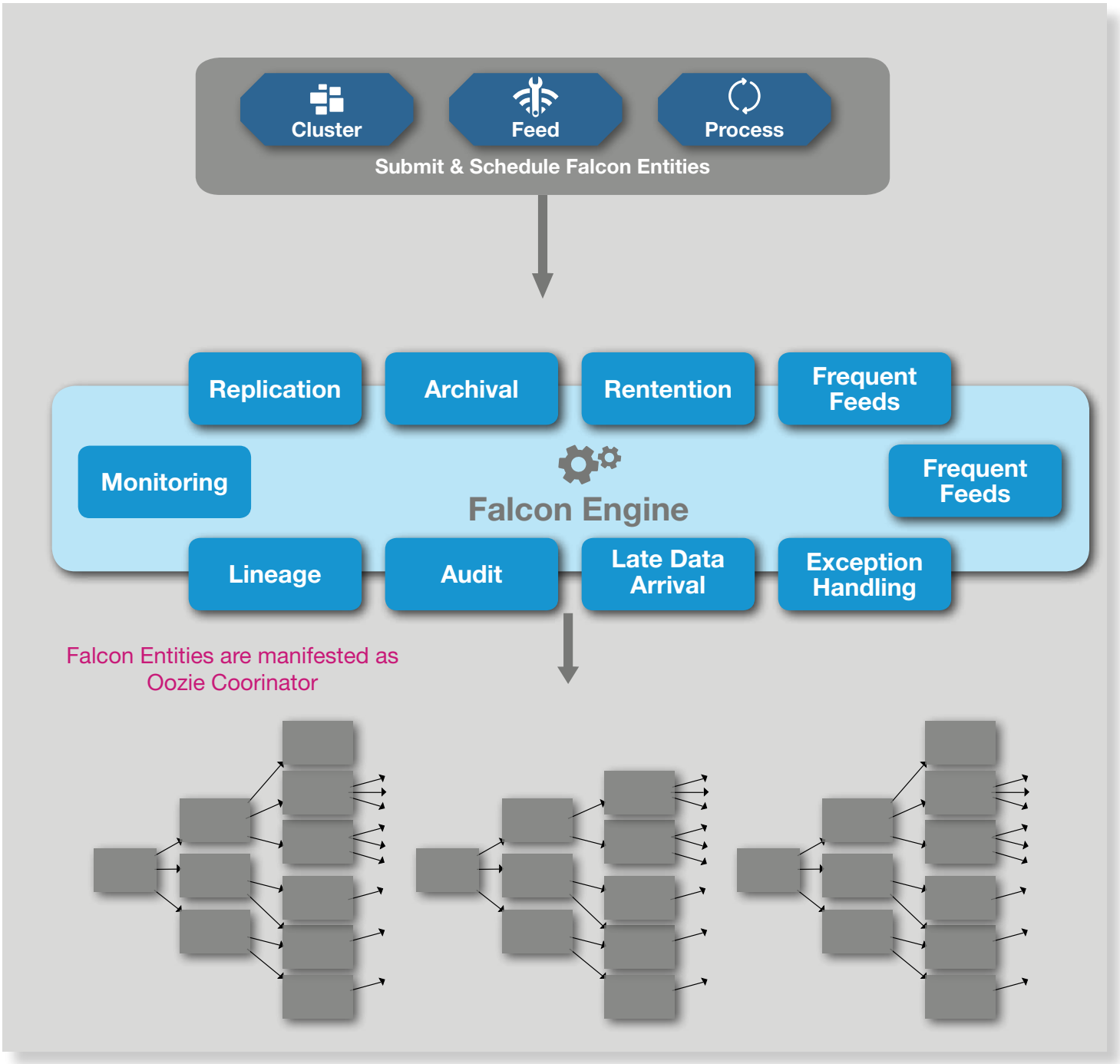

Fig. 7

4. Waterline Data Inventory

Waterline Data is a data marketplace platform provider, combining an automated data inventory with a self-service data catalog. The inspiration for “Waterline” came from the meta-phor of the data lake where data is hidden below the waterline. 80% of business value is cre-ated from Big Data by discovering data, and 80% of data is discovered by finding and under-standing trusted data. The mission of Waterline Data is to accelerate time to value for data discovery by helping business analysts and data scientists find, understand, and provision in-formation assets through self-service – without having to “dive” for data – and by helping data stewards provide agile data governance with automated and crowd-sourced business seman-tics and data lineage.

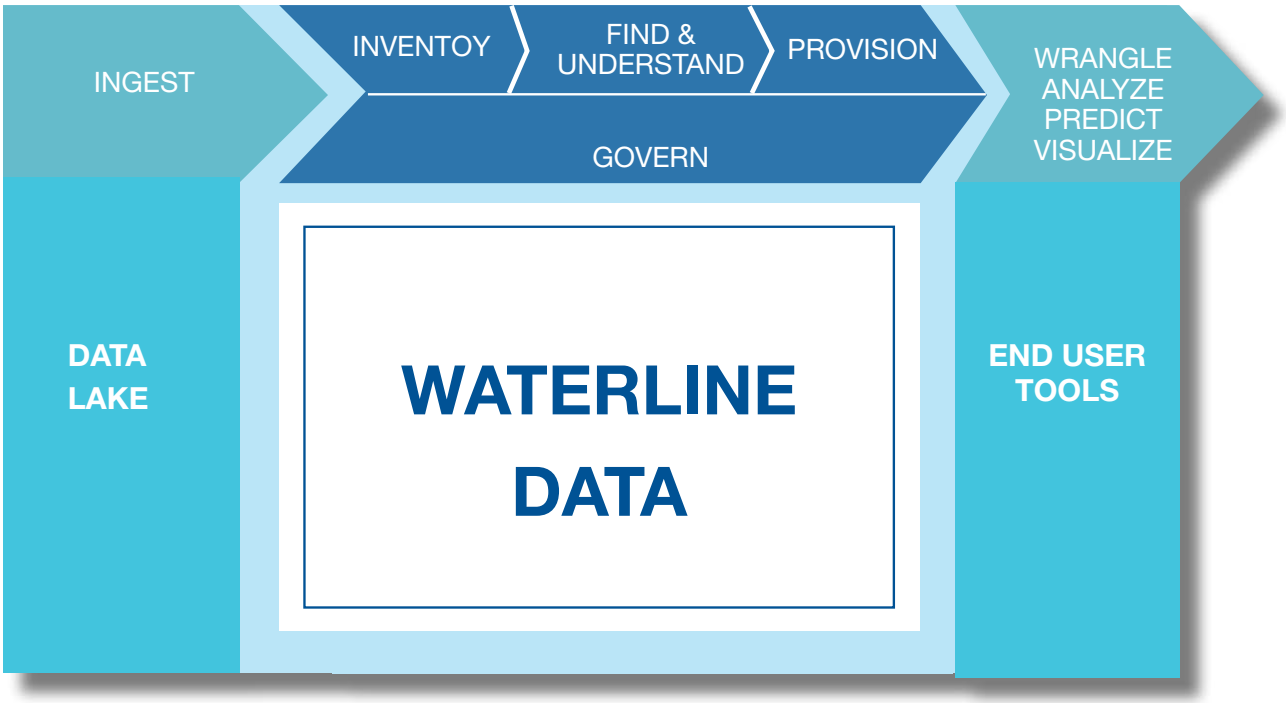

Fig. 8

Enterprises are evolving their data lakes to encompass all enterprise information as-sets, creating a “data marketplace” for the business users-–a logical data layer and catalog over physical data assets, to find, understand, and provision data assets at the speed of the business. Waterline Data is pioneering the “data marketplace” platform by providing smart data discovery to discover and crowd-source business metadata, data lineage, and infonomics, underlying a self-service business data catalog across on-premise and cloud data sources.

Features Summary

The below table lists the top features and the tools that support them.

Feature	Cloudera Navigator	Atlas	Falcon
Data Discovery	Yes	Yes	Yes
Continuous Optimization	Yes	-	-
Audit	Yes	Yes	Yes
Lineage	Yes	Yes	Yes
Security & Policy Enforcement	Yes	Yes	-
Export Metadata to thirdparty system	-	Yes	-
Provenance	Yes	Yes	Yes
Multi-Cluster Replication	-	-	Yes
Dataset Retention/Archival	Yes	Yes	Yes
Late data Handling	-	-	Yes
Automation	Yes	Yes	Yes
Feed Management	-	-	Yes
Data Classifications	-	Yes	-
Text Based Search	Yes	Yes	-

Conclusion

In this digital age, IT industry is heavily dependent on data and its innovative usage. As we are able to capture enormous data and do analytics on them, there is a lot of inherent values getting realized every day in the business. As we keep storing data from various sources, managing data gets difficult day-by-day. Since data goes into more complexity, Metadata Management is very critical for companies. Community development is helping to create vari-ety of tools to address this challenge.

References

1. https://www.cloudera.com/products/cloudera-navigator.html
2. http://atlas.incubator.apache.org/
3. http://hortonworks.com/apache/atlas/
4. https://falcon.apache.org/
5. http://hortonworks.com/apache/falcon/
6. http://hortonworks.com/apache/ranger/
7. http://www.waterlinedata.com/

Satya Nayak

Project Lead, Mphasis

Satya Nayak has 11 years of experience in enterprise applications and delivering software solutions. Satya has provided variety of solutions to various industry domains in his career. He has been working on Big Data and related technologies for 2.5 years. Currently he is working as a Project Lead, which involves creating designs and making innovative solutions for client requirements in Big Data Space.

Satya enjoys exploring new domains and taking up challenging roles. Satya is Cloudera Certified Spark and Hadoop Developer (CCA-175) and MapR Certified Spark Developer (MCSD).

Jansee Korapati

Module Lead, Mphasis

Jansee Korapati has 7+ years of experience in IT services, which includes 2 years in implementing Data Lakes using Big Data technologies like Hadoop, Spark, Sqoop, Hive, Impala and Flume. She is an expert in assessment and performance tuning of the long running queries in Hive with massive data.

Jansee has been working for Mphasis for 6 months as Module Lead. She is responsible for metadata management, ingestion and transformation layers of the current project.

About Mphasis

Mphasis (BSE: 526299; NSE: MPHASIS) applies next-generation technology to help enterprises transform businesses globally. Customer centricity is foundational to Mphasis and is reflected in the Mphasis' Front2Back™ Transformation approach. Front2Back™ uses the exponential power of cloud and cognitive to provide hyper-personalized ($C = X2C^{2}_{TM} = 1$) digital experience to clients and their end customers. Mphasis' Service Transformation approach helps 'shrink the core' through the application of digital technologies across legacy environments within an enterprise, enabling businesses to stay ahead in a changing world. Mphasis' core reference architectures and tools, speed and innovation with domain expertise and specialization are key to building strong relationships with marquee clients. To know more, please visit www.mphasis.com

For more information, contact: marketinginfo@mphasis.com

USA
460 Park Avenue South
Suite #1101
New York, NY 10016, USA
Tel.: +1 212 686 6655
Fax: +1 212 683 1690

UK
88 Wood Street
London EC2V 7RS, UK
Tel.: +44 20 8528 1000
Fax: +44 20 8528 1001

INDIA
Bagmane World Technology Center
Marathahalli Ring Road
Doddanakundhi Village
Mahadevapura
Bangalore 560 048, India
Tel.: +91 80 3352 5000
Fax: +91 80 6695 9942

VAL 18/01/19 US Letter BASIL 4188

www.mphasis.com