Introduction to SmartSense

Date of Publish: 2018-07-12

http://docs.hortonworks.com

Contents

Introduction to SmartSense

Hortonworks SmartSense Tool (HST) gives all support subscription customers access to a unique service that analyzes cluster diagnostic data, identifies potential issues, and recommends specific solutions and actions. These analytics proactively identify unseen issues and notify customers of potential problems before they occur.

SmartSense provides cluster diagnostic data collection capabilities, enabling customers to quickly gather configuration, metrics, and logs that they can use to analyze and troubleshoot support cases.

SmartSense:

1. Collects cluster diagnostic information to help you troubleshoot support cases. When filing a support case, a customer should attach a bundle for the respective component for faster case progress.
2. Automatically captures and uploads bundles that are used to produce customized recommendations for your cluster on areas of improvement, such as performance, operational stability, and security.
3. Allows you to automatically apply recommendations (where possible).
4. Reports, analyzes, and visualizes cluster activity.

In HDP, SmartSense is automatically included in Ambari 2.2.x and later. The integration between Ambari and SmartSense is facilitated by the Ambari stack and views extension mechanisms. These extensions enable you to add SmartSense as a native Ambari service, and they automatically deploy an Ambari view, enabling you to quickly capture data using the Ambari web UI.

In HDF, SmartSense is not included by default but SmartSense 1.5.x or newer can be installed on a cluster running HDF 3.2.x or newer. When SmartSense is installed on an HDF cluster, the option to capture for troubleshooting is available.

Hortonworks SmartSense Tool use is subject to written agreement with Hortonworks.

Cluster diagnostic collection

The HST agents capture, anonymize, and encrypt cluster diagnostic data, and then send it to the central HST server to coalesce into a single downloadable file called a "bundle". The HST agent processes are short-lived services that are started only for specific data capture tasks.

To provide the most complete picture of cluster utilization, HST agents must be installed on every node in the cluster. After an HST agent has captured the requested data from the host it is installed on, the process exits.

The following image illustrates the communication between HST agents and the HST server:

SmartSense anonymizes and encrypts the diagnostic information captured in the bundle. You can extend the anonymization process by adding your own rules.

There are two types of bundles: one for ad-hoc troubleshooting of support cases, and the other for proactive analysis and recommendations.

Support case troubleshooting bundles

Bundles captured for troubleshooting contain configuration and metrics for each node in the cluster, and logs for only the subset of services and hosts that you chose before initiating the capture process. Additionally, they may contain application logs if collection is for a YARN application or a Hive query. The purpose of these bundles is to provide support engineers with basic diagnostic information that can help them understand the state of your cluster so that they can troubleshoot and quickly resolve issues.

Proactive analysis bundles

Bundles captured for analysis contain configuration and metrics for each node in the cluster, but do not contain any logs. Their purpose is to produce recommendations for changing your cluster configuration to ensure better security, performance, and operations. These recommendations are available in the SmartSense View in Ambari web UI and in the SmartSense tab on the Hortonworks support portal.

Bundle content

SmartSense collects the following types of data:

- Operating system:
  - Configuration (partition layouts, file system mount options, key service status, network configurations, and so on)
  - Metrics (CPU, memory, I/O statistics, network statistics, and so on)
  - Logs (system messages and driver messages)
- Hortonworks Data Platform (HDP) service:
  - Configuration
  - Metrics (JMX reports and installed packages)
  - Logs (only for support case troubleshooting: not for proactive analysis)
  - Summary of cluster activity

When using SmartSense to capture support case troubleshooting bundles for issues with YARN applications or Hive queries, SmartSense captures additional data.

YARN application capture:

- Job configuration
- Job counters
- Job recommendations
- Job summary
- Job logs
- Task counters
- Task summary

Hive Query capture:

- Query plan
- Explain plain
- set –v output
- HS2 HA znode info
- Hive operations log
- YARN logs

To find out what exactly is captured in your specific environment, navigate to the SmartSense View and from the

menu select What is Captured. The What is Captured page includes a list of all components that are captured by SmartSense.

If you would like to see the exact data and files that are captured, perform a capture and then download the unencrypted bundle. To see a step-by-step example of how to do this, refer to the Hortonworks Community Connection post on how to inspect SmartSense bundle contents. If any files contain information that you would like to remove, replace, or anonymize, refer to the documentation for configuring data anonymization rules.

Related Information

How to inspect SmartSense bundle contents (HCC)

Configuring data anonymization rules

Bundle security

Hortonworks takes security seriously. Multiple levels of provisions ensure that sensitive data is protected:

- Anonymization and exclusions:
  - IP addresses and host names are always anonymized.
  - Passwords are not collected.
- Encryption:
  - SmartSense analysis bundles are encrypted using AES-256 and RSA-2048 encryption.
- Further customizations:
  - You can configure custom anonymization rules to include environment-specific patterns.
  - By default, all IP addresses, the domain component of fully qualified domain names, and S3 and WASB access keys are anonymized.
  - You can add custom configuration to exclude files and from collection.

Bundles sent to the Hortonworks SmartSense analysis environment are stored in their original anonymized and encrypted form for 90 days before being removed. Specific metadata, such as Apache Ambari and HDP stack version, node count, and amount of storage available and used, are stored for trending rules analysis. Recommendations generated for each bundle are available through the Hortonworks support portal and are stored for feedback purposes and used to improve future recommendations.

Services available for troubleshooting capture

The following services can be captured for troubleshooting:

HDP cluster

- Accumulo
- Ambari Metrics
- Atlas
- Data Analytics Studio
- Druid
- Falcon
- HBase
- HDFS
- Hive
- Kafka
- Knox
- Log Search
- MapReduce2
- NiFi
- Oozie
- Pig
- Ranger
- Ranger KMS
- SmartSense
- Solr Infra
- Spark
- Spark2
- Sqoop
- Storm
- Superset
- Tez
- YARN
- Zeppelin Notebook
- ZooKeeper

HDF cluster

- Ambari Metrics
- NiFi
- NiFi Registry
- Kafka
- Knox
- Ranger
- Streaming Analytics Manager
- Schema Registry
- SmartSense
- Solr Infra
- Storm
- Streamline
- ZooKeeper

Bundle transport

After a bundle has been captured, there are three ways to upload that bundle to Hortonworks:

- Automated bundle upload:
  - HST server
  - SmartSense gateway
- Manual bundle upload

Automated bundle upload

Depending on the availability of outbound internet access, you have two choices for automated bundle upload. If the HST server host has outbound internet access, you can configure it to automatically upload captured bundles to Hortonworks.

In this case, bundles are uploaded automatically over HTTPS from the HST server to the externally hosted Hortonworks SmartSense environment. If the HST server does not have outbound internet access, you can deploy a standalone SmartSense gateway to forward bundles to the hosted Hortonworks SmartSense environment.

HST server

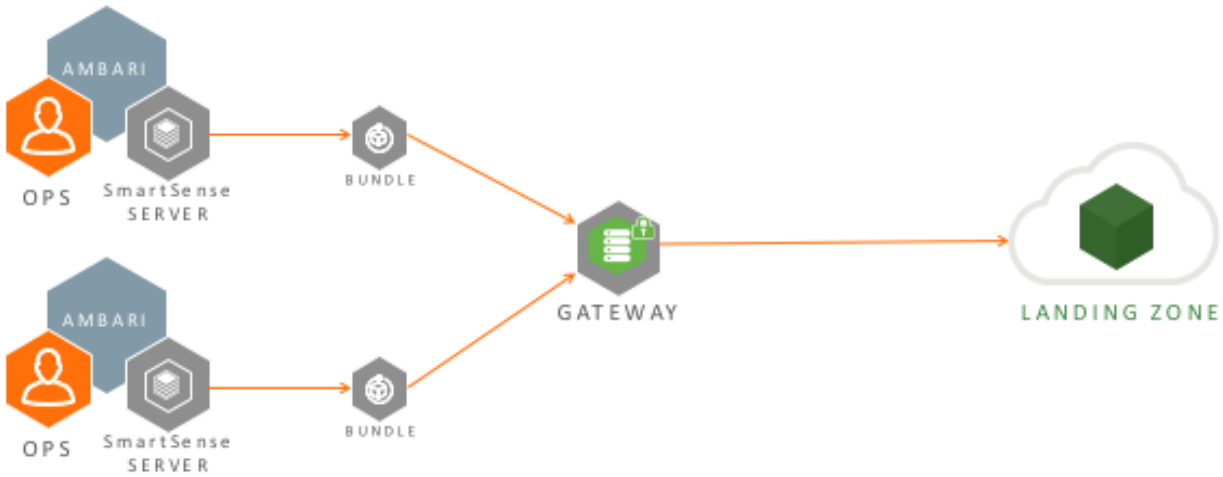
After a bundle has been captured, the HST server attempts to upload bundles to the Hortonworks hosted environment over HTTPS by default. This upload succeeds if your HST server host has outbound internet access. If your HST server host does not have outbound internet access, you have two options. If the HST server host can use a corporate HTTP proxy to upload bundles, you can configure your HST server host to do so by configuring bundle upload, or you can use the SmartSense gateway.

The following image illustrates bundle upload using the HST server:

SmartSense gateway

For those whose HST server hosts do not have outbound internet access, Hortonworks created the SmartSense gateway, which simplifies uploading bundles to Hortonworks. You can deploy a single gateway that supports multiple internal HST server deployments. In this deployment scenario, you do not need direct outbound internet access from the HST server to upload bundles. You need access only from the HST server to the gateway, and the gateway uploads all bundles to Hortonworks Support or to the SmartSense environment for SmartSense analysis.

The following image illustrates bundle upload using the SmartSense gateway:



Related Information

Configuring bundle upload

Manual bundle upload

If you are just getting started with SmartSense, you might still be waiting on your security or network operations resources to provide the necessary access for the HST server or the SmartSense gateway to send bundles. If you are in this situation, you can manually upload bundles via HTTPS.

After a bundle has been captured, you can go to SmartSense view in Ambari and download the bundle onto your desktop. You can then navigate to https://smartsense.hortonworks.com and log.

Related Information

Downloading and uploading bundles

SmartSense Portal

Activity analysis

Activity analyzer and activity explorer provide job utilization metric aggregation, reporting, and visualization for YARN-based workloads.

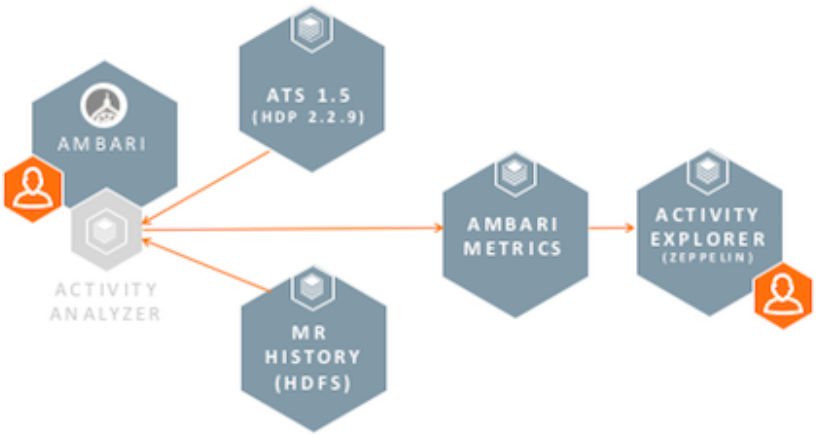
Activity Analyzer

Activity Analyzer communicates with YARN Application Timeline Server v1.5 and later, and with Hadoop Distributed File System (HDFS) to consume MapReduce history data. It aggregates and transforms this data, and stores it in the Ambari Metrics Collector.

Activity Explorer

Activity explorer includes an embedded instance of Apache Zeppelin, which hosts prebuilt notebooks that visualize cluster utilization data for YARN, Apache Hive or Apache Tez, and MapReduce workloads. Specifically, activity explorer includes data related to user, queue, job duration, and job resource consumption.

The following image illustrates how activity analyzer sends aggregated job history data to Ambari Metrics Collector, which makes that data available to activity explorer:


AMBARI
ATS 1.5
(HDP 2.2.9)
ACTIVITY
ANALYZER
MR
HISTORY
(HDFS)
AMBARI
METRICS
ACTIVITY
EXPLORER
(ZEPPELIN)