

**IN THE UNITED STATES DISTRICT COURT  
FOR THE WESTERN DISTRICT OF TEXAS  
AUSTIN DIVISION**

**BYTEWEAVR, LLC,**

**Plaintiff,**

**v.**

**CLOUDERA, INC.,**

**Defendant.**

§  
§  
§  
§  
§  
§  
§  
§  
§  
§  
§  
§  
§

**JURY TRIAL DEMANDED**

**CIVIL ACTION NO. 1:24-cv-261**

**PLAINTIFF’S ORIGINAL COMPLAINT FOR PATENT INFRINGEMENT**

Plaintiff BYTEWEAVR, LLC files this Complaint in the Western District of Texas (the “District”) against Defendant Cloudera, Inc. for infringement of U.S. Patent No. 6,839,733 (the “733 patent”), U.S. Patent No. 7,949,752 (the “752 patent”), U.S. Patent No. 6,862,488 (the “488 patent”), U.S. Patent No. 6,965,897 (the “897 patent”), U.S. Patent No. 6,999,961 (“961 patent”), U.S. Patent No. 7,082,474, (“474 patent”), U.S. Patent No. 8,275,827 (the “827 patent”), and U.S. Reissued Patent No. RE42153 (the “153 patent”) (collectively referred to as the “Asserted Patents”).

**THE PARTIES**

1. BYTEWEAVR, LLC (“BYTEWEAVR” or “Plaintiff”) is a Texas limited liability company, with registered address at 17350 State Hwy 249, Suite 220, Houston, Texas 77064.

2. On information and belief, Defendant Cloudera, Inc. (“Cloudera” or “Defendant”) is a corporation formed and organized under the laws of Delaware with its principal executive offices and corporate headquarters located at 5470 Great America Parkway, Santa Clara, CA 955054. Cloudera is registered to do business in Texas. *See* TEXAS SECRETARY OF STATE,

<https://direct.sos.state.tx.us/> at Filing No. 802113671 (showing that Cloudera has been registered since 2014 as a foreign corporation in Texas) (last visited Oct. 9, 2023). Cloudera’s registered agent in Texas is Corporation Service Company located at 211 E. 7<sup>th</sup> Street, Suite 620, Austin, TX 78701-3128.

3. Cloudera was founded in 2008 and was publicly traded in the New York Stock Exchange under the symbol “CLDR.” In October of 2021, Cloudera was acquired by investment firms Clayton, Dubilier & Rice (“CD&R”) and KKR “in an all cash transaction valued at approximately \$5.3 billion.” *See Cloudera Completes Agreement To Become a Private Company*, CLOUDERA, <https://investors.cloudera.com/home/default.aspx>. As a result, Cloudera ceased trading its common stock and is no longer listed on the NYSE.

4. On information and belief, Cloudera provides data management and analytics by providing “data warehouse, data science, data engineering, and operational database workloads together on a single integrated platform,” referred to at least as the “Cloudera Enterprise.” *See Overview of Cloudera and the Cloudera Documentation Set*, CLOUDERA, <https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/introduction.html> (last visited Oct. 9, 2023). Since at least 2014, Cloudera offered to its customers products and services including the Cloudera Distributed Hadoop (or “CDH”), as a component of at least the Cloudera Enterprise platform, to “meet [the] enterprise demands” of its customers. *See CDH Components*, CLOUDERA, <https://www.cloudera.com/products/open-source/apache-hadoop/key-cdh-components.html> (last visited Oct. 9, 2023); *see also CDH Version and Packaging Information*, CLOUDERA, [https://docs.cloudera.com/documentation/enterprise/release-notes/topics/rg\\_cdh\\_vd.html](https://docs.cloudera.com/documentation/enterprise/release-notes/topics/rg_cdh_vd.html) (last visited Oct. 10, 2023).

5. In 2019, Cloudera “introduced Cloudera Data Platform (CDP), [its] cloud-native data platform for the enterprise data cloud built on open source software,” which incorporated functionality and capabilities of the Cloudera Enterprise and CDH products and services. *See 2021 Form 10-K Annual Report*, CLOUDERA, INC., available at <https://investors.cloudera.com/financials-and-filings/sec-filings/>, at page 5 (cited as “*2021 Cloudera Annual Report*”). CDP is “offered as Public Cloud services and Private Cloud software subscriptions.” *Id.* Cloudera “license[s its] products under a primarily open source licensing model based on the Apache Software License (ASL) and the Affero General Public License (AGPL).” *Id.* Cloudera also offers other “traditional on-premises data management and analytics offerings” that include Cloudera DataFlow (CDF), Cloudera Enterprise Data Hub (EDH), Cloudera Data Science and Engineering, and Cloudera SDX. *Id.*

6. On information and belief, CDH “delivers the core elements of Hadoop – scalable storage and distributed computing – along with a Web-based user interface and vital enterprise capabilities.” *See CDH Overview*, CLOUDERA, [https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/cdh\\_intro.html](https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/cdh_intro.html) (last visited Oct. 9, 2023). CDH “is Apache-licensed open source and is the only Hadoop solution to offer unified batch processing, interactive SQL and interactive search, and role-based access controls.” *Id.* CDH allows users to “[s]tore any type of data and manipulate it with a variety of different computation frameworks including batch processing, interactive SQL, free text search, machine learning and statistical computation.” *Id.* The components included with CDH provide to the Cloudera Enterprise various features and functionalities, including, a “[w]orkflow scheduler to manage Hadoop jobs” via the Apache Oozie component, “[j]ob scheduling and cluster resource management” via the YARN component, and an “SQL workbench for data warehouses” via the Hue component. *See Platform*

*Features*, CLOUDERA, <https://www.cloudera.com/products/pricing/product-features.html> (last visited Oct. 9, 2023). Also, Hadoop supports data compression and compression formats, including using and compression of Apache Avro Data files with CDH. *See Data Compression*, CLOUDERA, [https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/introduction\\_compression.html#concept\\_wlk\\_hgy\\_pv](https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/introduction_compression.html#concept_wlk_hgy_pv) (last visited Oct. 30, 2023).

7. On information and belief, “Cloudera Manager provides unified and centralized management and monitoring for Cloudera Runtime and Cloudera Search.” *See What is Cloudera Search*, CLOUDERA, <https://docs.cloudera.com/cdp-private-cloud-base/7.1.8/search-overview/topics/search-introducing.html> (last visited Dec. 14, 2023). Cloudera Runtime provides, as a component, the Cloudera Search service as an “integrated part of CDH and supported with Cloudera Enterprise.” *See Cloudera Search*, CLOUDERA, <https://www.cloudera.com/products/open-source/apache-hadoop/apache-solr.html> (last visited Dec. 14, 2023). Cloudera Search is powered by Apache Solr which “makes Apache Hadoop accessible to everyone via integrated full-text search.” *Id.*

8. On information and belief, the Cloudera Platforms provide “customers a very secure, efficient, and easy way to traverse data back and forth between the different environments they have in many other locations.” *See Apache NiFi – the data movement enabler in a hybrid cloud environment*, CLOUDERA BLOG, <https://blog.cloudera.com/apache-nifi-the-data-movement-enabler-in-a-hybrid-cloud-environment/> (last visited Oct. 12, 2023). The Cloudera Platforms include, within the Cloudera Shared Data Experience (SDX), “Cloudera Flow Management powered by Apache NiFi...[to] move data back and forth between your environments, while ensuring the proper level of security, resilience, auditability, and governance.” *Id.* Apache NiFi

“provides a wide range of processors to interact with the native managed services of the cloud providers.” *Id.*

9. The Cloudera Enterprise and/or CDP (collectively the “Cloudera Platforms”) and their components are utilized by customers of Cloudera across industries, including Technology, Financial Services, Telecommunications, Business Services, and Healthcare and Life Sciences, among many others. *See Customers: Unleashing Hidden Data Treasures for Customers*, Cloudera, <https://www.cloudera.com/about/customers.html?industry=Financial%20Services> (providing a drop-down to access customer stories in various industries). The Cloudera Platforms are offered for “public cloud consumption and on-premises private cloud software subscription.” *See Cloudera Pricing*, CLOUDERA, <https://www.cloudera.com/products/pricing.html> (last visited Oct. 9, 2023). On information and belief, Cloudera collects revenues and profits from the installation, licensing, and use of the Cloudera Platforms. *See id.* Cloudera, for example, charges public cloud platform customers “per Cloudera Compute Unit (CCU) which is a combination of Core and Memory” usage and charges private cloud platform customers via an annual subscription model with CCU, node cap, and storage limits. *See id.*

10. On information and belief, Defendant Cloudera on its own and/or via subsidiaries, distributors, and affiliates maintains a corporate and commercial presence in the United States, including in Texas and this District. Defendant maintains its business presence in the U.S. and Texas via at least the following activities: 1) distributing and providing its Cloudera Platforms, among other products and services of Cloudera, to customers; 2) maintaining an online presence (<https://www.cloudera.com>) that solicits sales and sales inquiries and provides customer support for Cloudera products and services; 3) registering to do business in Texas; 4) employing persons across the world who support the development of products and services and provide customer support to

U.S. residents and companies, and 5) employing persons in the United States, including residents of Texas and this District. For example, Defendant employs Texas residents in at least one location in the Austin, Texas area at 515 Congress, Suite 1300, Austin, TX 78701. *See, e.g., North America, CLOUDERA*, <https://www.cloudera.com/about/locations.html> (showing Cloudera locations in the U.S. and Texas). Thus, Defendant Cloudera does business in the United States, the state of Texas, and in the Western District of Texas.

### **JURISDICTION AND VENUE**

11. This action arises under the patent laws of the United States, namely 35 U.S.C. §§ 271, 281, and 284-285, among others.

12. This Court has subject matter jurisdiction pursuant to 28 U.S.C. §§ 1331 and 1338(a).

13. On information and belief, Defendant Cloudera is subject to this Court's specific and general personal jurisdiction pursuant to due process and/or the Texas Long Arm Statute, due at least to its substantial business in this State and this District, including: (A) at least part of its infringing activities alleged herein, including its registration to do business in Texas, which purposefully avail the Defendant of the privilege of conducting those activities in this state and this District and, thus, submits itself to the jurisdiction of this Court; and (B) regularly doing or soliciting business, engaging in other persistent conduct targeting residents of Texas and this District, and/or deriving substantial revenue from infringing goods offered for sale, sold, and imported and services provided to and targeting Texas residents and residents of this District vicariously through and/or in concert with its alter egos, intermediaries, agents, distributors, importers, partners, customers, subsidiaries, affiliates, and/or consumers.

14. For example, Cloudera has corporate offices in the United States, including in Texas. Cloudera owns or leases a corporate office in this District at 515 Congress Ave., Austin, Texas. *See Property Search*, TRAVIS COUNTY CENTRAL APPRAISAL DISTRICT, <https://stage.travis.prodigycad.com/property-search> (Search results for “Cloudera” as owner) (last visited Oct. 9, 2023). Importantly, Cloudera maintains its own employees or agents at this office to conduct its business of at least distribution of Cloudera products and services. *See, e.g., Cloudera Careers*, CLOUDERA, [https://cloudera.wd5.myworkdayjobs.com/en-US/External\\_Career/job/Cloud-Solution-Specialist\\_230270-1?locations=099bd8052f77105bfed69a9cf552387f](https://cloudera.wd5.myworkdayjobs.com/en-US/External_Career/job/Cloud-Solution-Specialist_230270-1?locations=099bd8052f77105bfed69a9cf552387f) (showing a “Cloud Solution Specialist” position open in Texas) (last visited Oct. 9, 2023).

15. Such a corporate and commercial presence by Defendant Cloudera furthers the development, design, manufacture, importation, distribution, sale, offering for sale, and use of Defendant’s infringing data management and analytics products and services in Texas, including in this District. Through utilization of its business segments and partners, Cloudera has committed acts of direct and/or indirect patent infringement within Texas, this District, and elsewhere in the United States, giving rise to this action and/or has established minimum contacts with Texas such that personal jurisdiction over Cloudera would not offend traditional notions of fair play and substantial justice.

16. On information and belief, Cloudera has placed and continues to place infringing data management and analytics products and services, including the Cloudera Platforms and their components into the U.S. stream of commerce. Cloudera has placed such products and services into the stream of commerce with the knowledge and understanding that such products and services are, will be, and continue to be sold, offered for sale, and/or imported into the State of Texas and this

District. *See Litecubes, LLC v. Northern Light Products, Inc.*, 523 F.3d 1353, 1369-70 (Fed. Cir. 2008) (“[T]he sale [for purposes of § 271] occurred at the location of the buyer.”); *see also Semcon IP Inc. v. Kyocera Corporation*, No. 2:18-cv-00197-JRG, 2019 WL 1979930, at \*3 (E.D. Tex. May 3, 2019) (denying accused infringer’s motion to dismiss because plaintiff sufficiently plead that purchases of infringing products outside of the United States for importation into and sales to end users in the U.S. may constitute an offer to sell under § 271(a)).

17. On information and belief, Defendant Cloudera also purposefully places infringing data management and analytics products and services in established distribution channels in the stream of commerce by contracting with “partners” who distribute Cloudera’s products in the U.S. via license, to users affiliated with those partners, including in Texas and this District. *See Become a Cloudera Partner*, CLOUDERA, <https://www.cloudera.com/partners/cloudera-partner-network-program.html> (stating “[p]artner with Cloudera, and your customers will never think about data the same way again”) (last visited Oct. 9, 2023). Cloudera contracts with these partner companies with the knowledge and expectation that Cloudera’s data management and analytics products and services will be imported, distributed, advertised, offered for sale, sold, and used in the U.S. market, including to users affiliated with such partners. Such partner types include “Cloudera Resellers,” “Distributors,” “Hardware Vendor,” “Software Vendor,” “System Integrator,” and “Training Reseller,” among others. *See Find a partner*, CLOUDERA, <https://www.cloudera.com/partners/partners-listing.html> (last visited Oct. 9, 2023). Moreover, “Cloudera partners with federal, state and local, and higher education institutions to support data security and governance mandates, modernize data architectures across any platform, and meet the zero-trust mandate related to data flow.” *See Government Solutions: We Move Your Data*, Cloudera, <https://www.cloudera.com/solutions/public-sector.html> (last visited Oct. 11, 2023).



Each of these partners (among many more), on information and belief, have a significant business presence in the U.S. and Texas and serve as a distribution channel for Cloudera's products of services into this District.

18. Based on Defendant Cloudera's physical and virtual presence and connections and relationships with its distributors, resellers, vendors, contractors, dealers, installers, trainers, and other partners, Cloudera knows that Texas is a termination point of the established distribution channel for the sale and use of Cloudera data management and analytics products and services, including the Cloudera Enterprise platform(s) to customers and other users in Texas. Cloudera, therefore, has purposefully directed its activities at Texas, and should reasonably anticipate being brought in this Court, at least on this basis. *See Icon Health & Fitness, Inc. v. Horizon Fitness, Inc.*, 2009 WL 1025467, at (E.D. Tex. 2009) (finding that "[a]s a result of contracting to manufacture products for sale in" national retailers' stores, the defendant "could have expected that it could be brought into court in the states where [the national retailers] are located").

19. Venue is proper in this District pursuant to 28 U.S.C. §§ 1391(c) and 1400(b). As alleged herein, Defendant Cloudera has committed acts of infringement in this District. As further alleged herein, Defendant Cloudera, via its own operations and employees located there, has a regular and established place of business in this District. Cloudera's regular and established place of business is at least at 515 Congress Ave., Austin, Texas 78701, which according to publicly available records is located in Travis County. Accordingly, Cloudera may be sued in this district under 28 U.S.C. § 1400(b).

20. On information and belief, Defendant Cloudera has significant ties to, and presence in, the State of Texas and the Western District of Texas, making venue in this District both proper and convenient for this action.

**THE ASSERTED PATENTS AND TECHNOLOGY**

21. The Asserted Patents cover various aspects of network systems extensible by users as subscribers to a network service. Such extensibility by users of network services includes interaction with the network by creating, copying, modifying, editing, and deleting agents. Such agents are invoked by users to consume service resources. Such network systems further include automation of validation of equipment and/or processes via a user interface and validation processing engine. Moreover, such network systems include server systems with network connected distributed client systems to provide workload processing. Such workload processing includes indexing of the location of data required to process workloads and processing of search results via a content aggregator. Data stored in such network systems, may be arranged in data files in a mixed format physical layout divided into fixed-sized fields and variable sized fields and compressed.

22. The '733 patent involves at least admitting a user to a network system wherein at least one agent is operable to consume a service resource (e.g., CPU, memory resource, etc.) while utilizing a service to perform a task for the user. The user is allowed to create, modify, or delete the agent within the network system.

23. The '752 patent involves at least receiving, using a computing device, data for creating a network-based agent. An execution of the network-based agent is invoked in response to receiving a URL that defines a type of event and identifies the agent. Invoking execution of the network-based agent uses a service and a service resource that is consumed by the network-based agent for performing the invoking operation. The result of the operation is communicated over a network communication link.

24. The '488 patent involves at least automating, in a computing environment, the validation of equipment and/or processes for use, for example, in a pharmaceutical and/or biotechnology manufacturing facility. A user interface is provided that accepts and/or displays data representative of validation processing and/or validation workflow management information. A validation processing engine is provided that comprises a processing rule that operates to produce validation protocol information.

25. The '897 patent involves at least arranging data in a data file on a mixed format physical layout. This layout has a plurality of fixed-sized fields, a plurality of variable-sized fields, and a plurality of offset slots. The fixed-sized fields are of a first size and the offset slots are of a second size. The data on the mixed format physical layout is divided into the fixed-sized fields and the variable sized fields. The data of the variable sized fields and the fixed-sized fields is compressed.

26. The '961 patent involves at least accessing a content aggregator and transmitting a search query to the content aggregator. The search query is transmitted to a plurality of remote agents located on one of a plurality of distinct networks. Each network is searched for content responsive to the query. A search result is transmitted from the remote agents to the content aggregator. The search results are processed via the content aggregator, wherein processing includes applying rules and standards designated by a client. And processed information is transmitted from the content aggregator to the client.

27. The '474 patent involves at least receiving client requests from server systems to use a distributed processing system to process a workload. The first workload is sent to a host distributed device. An index defining a location of data required to process the first workload is sent to the host distributed device. The data is accessed from a first data address in the index. And the index is

updated to include a storage address of storage coupled to the host distributed device as a location of the data.

28. The '827 patent involves at least configuring a distributed processing system with distributed devices coupled to a network. The devices include client agents that process workloads for the system. The client agents have software-based network attached storage (NAS) components that assess unused or underutilized storage resources in distributed devices. The NAS devices have storage resources related to the unused or underutilized storage resources. The system processes data storage or access workloads and enables the distributed devices to store location information associated with data stored by the distributed devices through the use of client agents. At least one of the distributed devices is enabled to function as a stand-alone dedicated NAS device through the use of the client agents.

29. The '153 patent involves at least a server system coupled to a network with network-connected distributed client systems having under-utilized capabilities. The client systems run a client agent program to provide workload processing for a project of a distributed computing platform. The server system distributes project workloads to the client systems and distributes initial project and poll parameters to the client systems. Poll communications are received from the client systems during the processing of project workloads and a dynamic snapshot information of a current project status is provided based on the poll communications. The poll communications are analyzed to determine whether to modify the initial project and poll parameters, which indicate how many client systems are active in the project. If fewer client systems are desired, including within a polling response communications, the number of actively participating client systems is reduced. And if a greater number of client systems is desired, then client systems are added to active participation in the project. The poll response communications are sent to the client systems to modify the initial

project and poll parameters, depending on the analysis of the poll communications. The steps of receiving and analyzing poll communications and sending poll response communications are repeated to dynamically coordinate project activities of the client systems during project operations.

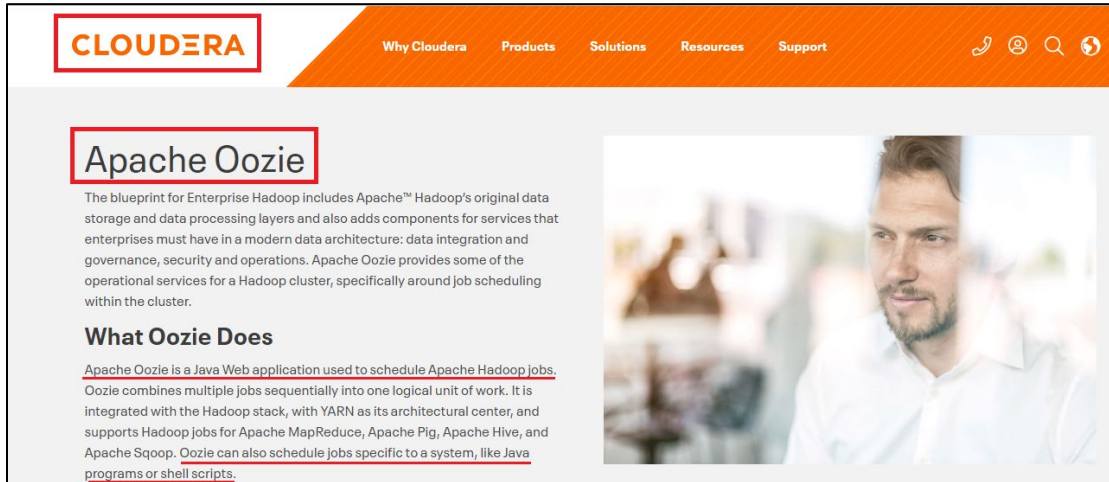
30. On information and belief, a significant portion of the operating revenue of Defendant is derived from the development, design, manufacture, distribution, licensing, sale, offering for sale, and use of Cloudera's data management and analytics products and services, including the Cloudera Platforms and their components. *See, e.g., Cloudera 2021 Annual Report* at 33 (“We generate revenue from subscriptions and services.”). For example, Defendant Cloudera utilizes its subsidiaries, distributors, resellers, vendors, contractors, dealers, installers, trainers, and other partners to provide data management and analytics products and services and related products and services to consumers. For the year 2020, Defendant reported \$794 million in revenue for the Subscription and Services combined. *See Cloudera 2021 Annual Report* at 37. For the year 2021, Defendant reported \$869 million in revenue for the Subscription and Services combined. *Id.* Cloudera reports that “[s]ales outside of the United States represented approximately 40%, 38% and 34% of our total revenue for the years ended January 31, 2021, 2020 and 2019, respectively.” *Id.* at 33. Thus, the majority of Cloudera's revenue derives from Cloudera's data management and analytics products and services distributed, licensed, sold and offered for sale by consumers in the United States.

31. The Asserted Patents cover Defendant's data management and analytics products and components, software, services, and processes related to same that cover various aspects of network systems extensible by users as subscribers to a network service, including such network systems that 1) allow a user to interact with the network by creating, copying, modifying, editing, and deleting agents to support consumption of network services and/or allow a user to provide for

automation of validation of equipment and/or processes via a user interface and validation processing engine; 2) server systems with network-connected distributed client systems to provide workload processing; 3) indexing of the location of data required to process workloads and processing of search results via a content aggregator; and 4) arranging data stored in such network systems in data files in a mixed format physical layout divided into fixed-sized fields and variable sized fields (collectively referred to herein as the “Accused Instrumentalities”). *See, e.g., Cloudera Data Platform (CDP)*, CLOUDERA, <https://www.cloudera.com/products/cloudera-data-platform.html> (“CDP delivers faster and easier data management and data analytics for data anywhere, with optimal performance, scalability, and security.”) (last visited Oct. 10, 2023). Defendant’s infringing Accused Instrumentalities include, but are not limited to, components of the Cloudera Platforms, including, but not limited to networks, methods, processes, software, firmware, distributions, infrastructure, environments, interfaces, hosts, tools, data connections, databases, resources, and related services provided to partners, users, customers, clients, and consumers via at least the Cloudera Enterprise, the Cloudera Data Platform, Data Hub, Runtime, Search, the Cloudera SDX Management Console, Cloudera Manager, CDH, Cloudera Flow Management, and Cloudera distributions of Apache Oozie, NiFi, YARN, Hue, Avro, Zookeeper and related data storage and compression techniques.

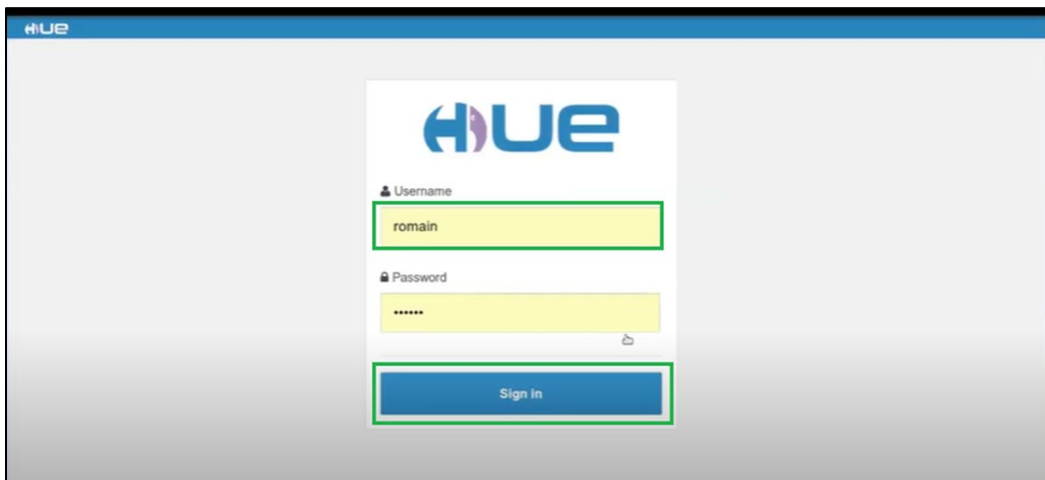
32. The Asserted Patents, including at least claim 37 of the ’733 patent, cover Accused Instrumentalities of Defendant that utilize Cloudera’s Apache Oozie, which, as described below, is a workflow scheduler system for managing and scheduling tasks in Cloudera’s Hadoop ecosystem (also known as CHP or CDP). Cloudera provides a web-based interface for interacting with Oozie editor to create, manage and schedule workflows. The Oozie editor allows a Cloudera user to create a scheduler agent that utilizes various Cloudera services to perform a task such as importing data

from HDFS for a period, deleting Internet history every week, etc. Further, Oozie uses YARN architecture to efficiently share resources, such as CPU and memory, to run the scheduling task.



<https://www.cloudera.com/products/open-source/apache-hadoop/apache-oozie.html>

33. As shown below, a user is admitted to the Cloudera network system (i.e., Hue) by passing login authentication.

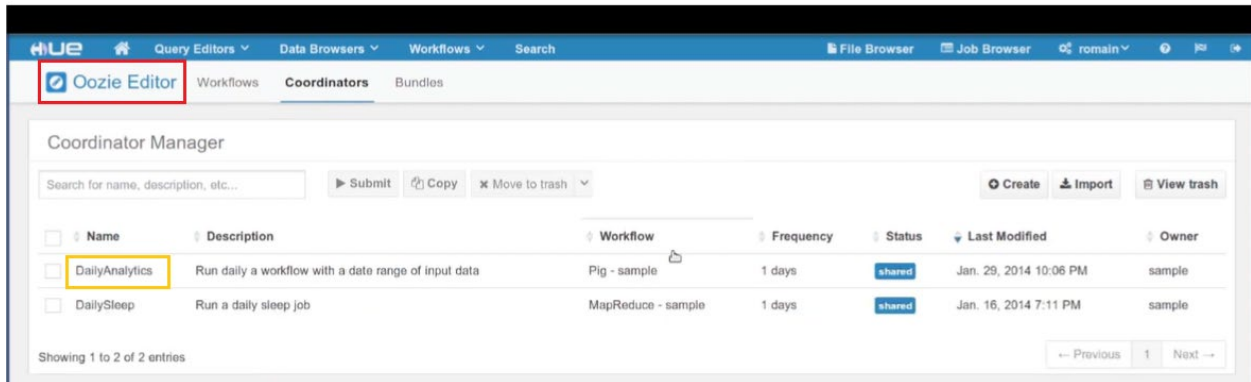


See *Hadoop Tutorial: Oozie crontab scheduling in Hue*, HUE VIDEOS, available via YouTube at [https://www.youtube.com/watch?v=Nnzd\\_q6vSHU](https://www.youtube.com/watch?v=Nnzd_q6vSHU).

34. Hue is a “web-based interactive query editor that enables you to interact with data warehouses.” See *Introduction to Hue*, CLOUDERA,

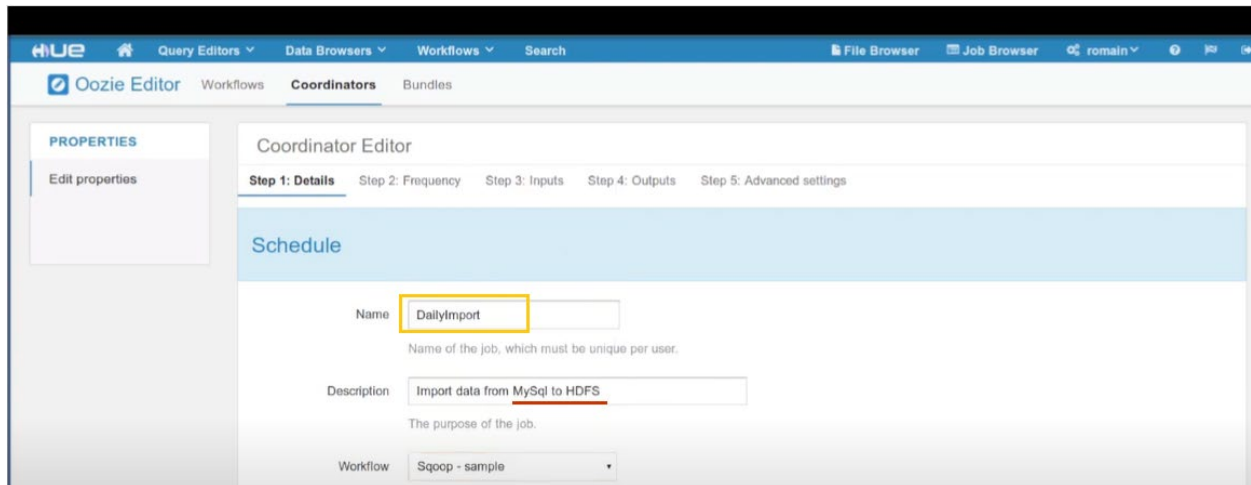
<https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/hue.html> (last visited Oct. 10, 2023).

35. Hue provides access to an “Oozie Editor allowing users to schedule workflows, e.g., “DailyAnalytics,” as shown below, among other types of Apache Hadoop jobs.



*Hadoop Tutorial: Oozie crontab scheduling in Hue*, HUE VIDEOS, available via YouTube at [https://www.youtube.com/watch?v=Nnzd\\_q6vSHU](https://www.youtube.com/watch?v=Nnzd_q6vSHU).

36. As shown below, the user creates an agent, via the Oozie Editor, which is operable to perform a task for the user, such as importing data from MySQL to HDFS.



*Hadoop Tutorial: Oozie crontab scheduling in Hue*, HUE VIDEOS, available via YouTube at [https://www.youtube.com/watch?v=Nnzd\\_q6vSHU](https://www.youtube.com/watch?v=Nnzd_q6vSHU).



37. Performance of the task consumes allocated resources using a YARN architecture. As explained below, YARN includes a “resource manager” that “[a]llocates cluster resources using a Scheduler.”

## Understanding YARN architecture

YARN, the Hadoop operating system, enables you to manage resources and schedule jobs in Hadoop.

YARN architecture and workflow

YARN has three main components:

- **ResourceManager** Allocates cluster resources using a Scheduler and ApplicationManager.
- ApplicationMaster: Manages the life-cycle of a job by directing the NodeManager to create or destroy a container for a job. There is only one ApplicationMaster for a job.
- NodeManager: Manages jobs or workflow in a specific node by creating and destroying containers in a cluster node.

<https://www.cloudera.com/products/open-source/apache-hadoop/apache-oozie.html>

38. The YARN resource manager allows for “allocating resources through scheduling limiting CPU usage,” among “multiple resource types.”

## Resource Scheduling and Management

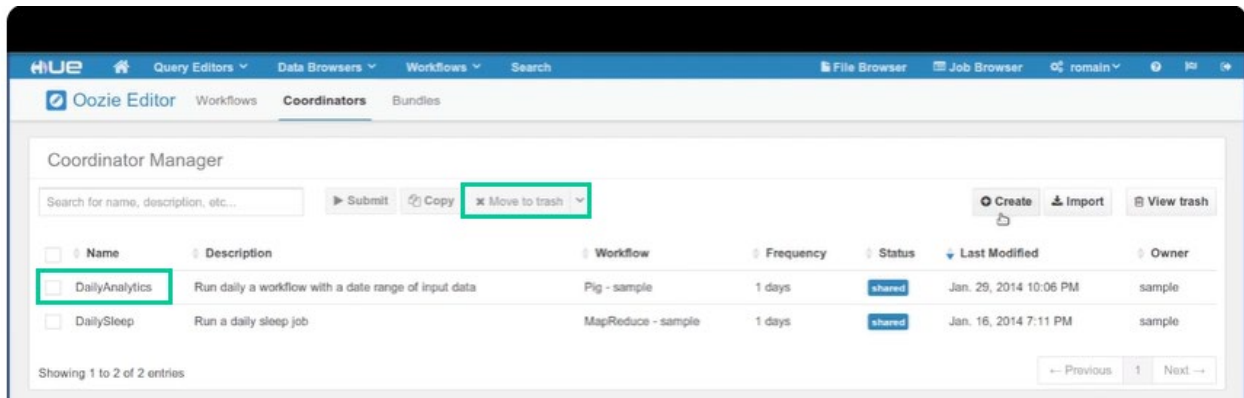
You can manage resources for the applications running on your cluster by allocating resources through scheduling, limiting CPU usage by configuring cgroups, and partitioning the cluster into subclusters using partitions, and launching applications on Docker containers.

The *CapacityScheduler* is responsible for scheduling. The *CapacityScheduler* is used to run Hadoop applications as a shared, multi-tenant cluster in an operator-friendly manner while maximizing the throughput and the utilization of the cluster.

The *ResourceCalculator* is part of the YARN *CapacityScheduler*. If you have only one type of resource, typically a CPU virtual core (vcore), use the `DefaultResourceCalculator`. If you have multiple resource types, use the `DominantResourceCalculator`.

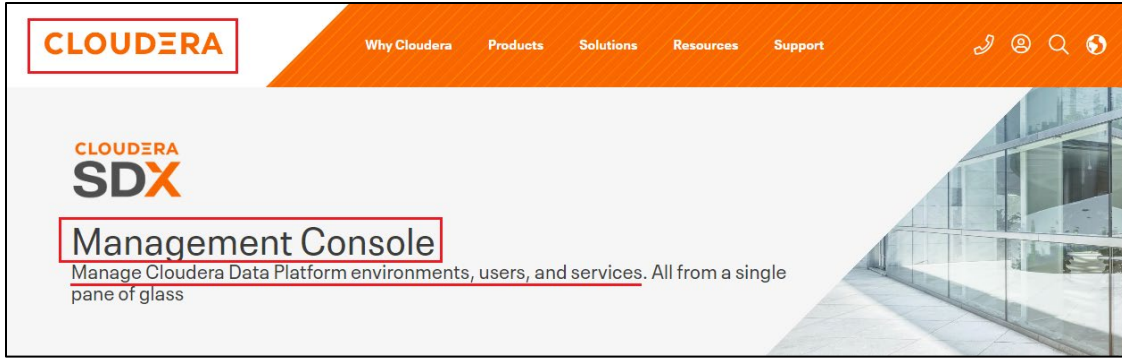
<https://docs.cloudera.com/cdp-private-cloud-base/7.1.6/yarn-allocate-resources/topics/yarn-cluster-management.html>

39. The user, via the Oozie Editor, can create, modify, or delete the agent (e.g., an Oozie workflow scheduler agent) within the network system. For example, the Oozie editor allows a Cloudera user to create a scheduler agent that utilizes various Cloudera services to perform a task such as importing data from HDFS for a period, deleting Internet history every week, etc. As shown below, these scheduler agents can be deleted (i.e., “move[d] to trash”) to stop its execution for the next run.

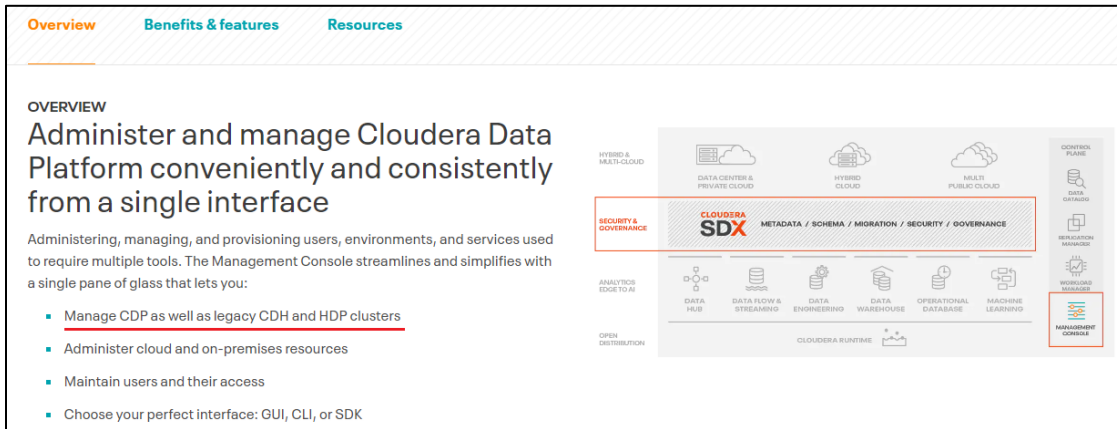


*Hadoop Tutorial: Oozie crontab scheduling in Hue*, HUE VIDEOS, available via YouTube at [https://www.youtube.com/watch?v=Nnzd\\_q6vSHU](https://www.youtube.com/watch?v=Nnzd_q6vSHU).

40. The Asserted Patents, including at least claim 24 of the '752 patent, cover Accused Instrumentalities of Defendant that practice a method comprising the steps of receiving, using a computing device (e.g., Cloudera server), data (e.g., cluster definition, cluster name, etc.) for creating a network-based agent (e.g., a cluster). As shown below, the Cloudera Management Console allows a user to create and manage clusters.

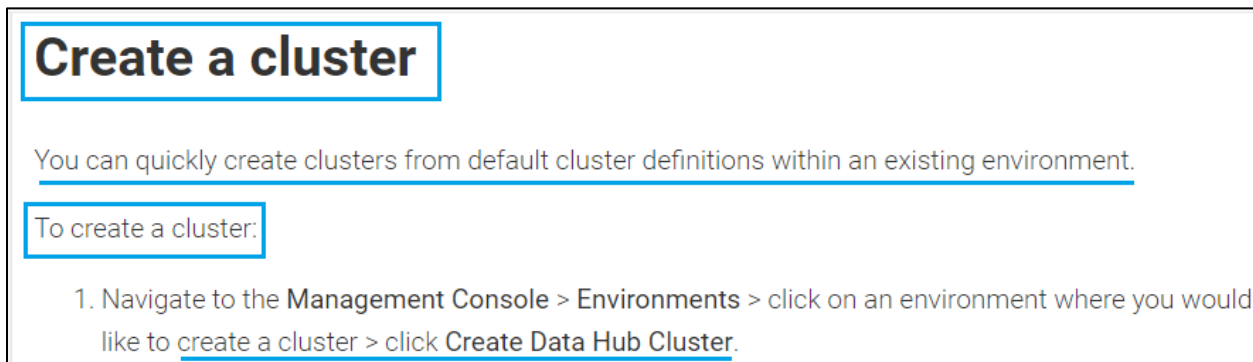


<https://www.cloudera.com/products/sdx/management-console.html>



<https://www.cloudera.com/products/sdx/management-console.html>

41. A cluster is a set of hosts running inter-dependent services. For creating a cluster, data such as cluster definition, number of nodes, types of service, cluster name, etc. are provided by the user.

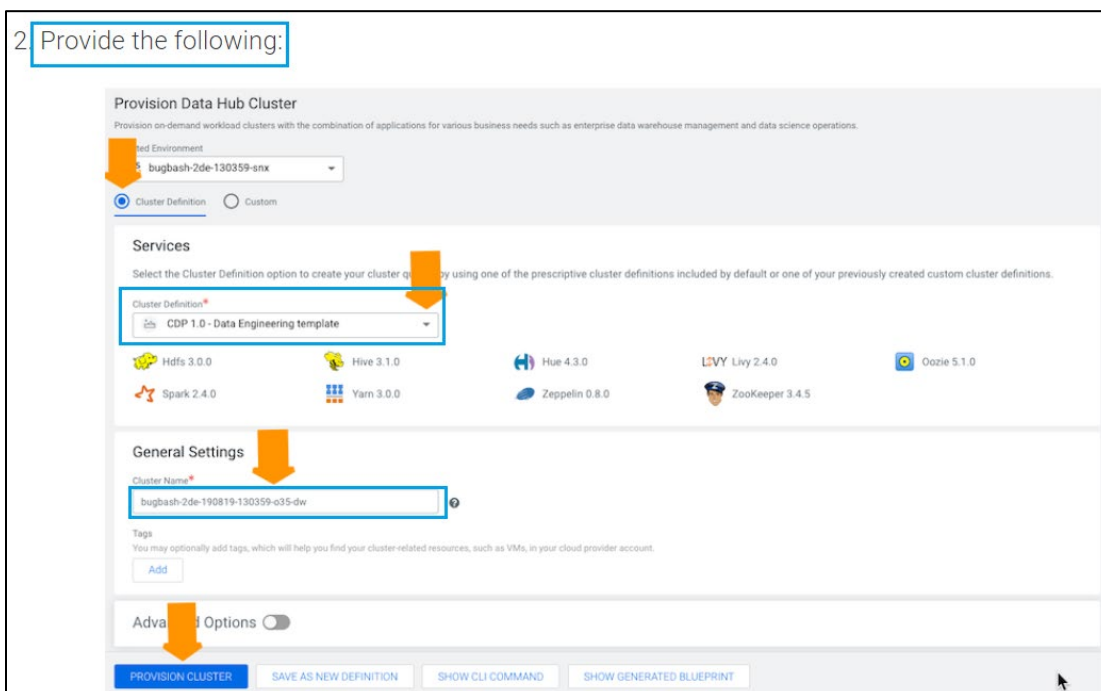


<https://docs.cloudera.com/data-hub/cloud/getting-started-tutorial/topics/dh-tutorial-create-cluster.html>

- a. Select Cluster Definition.
- b. In the **Services** section, select a specific cluster definition, for example Data Engineering for AWS.
- c. Under General Settings > Cluster Name, provide some name for your cluster.

<https://docs.cloudera.com/data-hub/cloud/getting-started-tutorial/topics/dh-tutorial-create-cluster.html>

42. When the user clicks on ‘Provision Cluster’, the creation of a cluster is triggered.



<https://docs.cloudera.com/data-hub/cloud/getting-started-tutorial/topics/dh-tutorial-create-cluster.html>

43. Triggering the provision of the cluster invokes execution of a network-based agent (i.e., starts a particular cluster). When the user clicks on ‘Provision Cluster’, the user is redirected to environment details page to access cluster details, as shown below.

## Monitor cluster creation

Once cluster creation has been triggered, you can monitor it from cluster details.

You are redirected to environment details > Data Hub Clusters and you can see a new entry corresponding to your cluster. Click on the entry corresponding to the name of your cluster to navigate to cluster details:

Environments / bugbash-2de-130359-snx / Clusters

bugbash-2de-130359-snx

US West (Oregon) - us-west-2

Actions ▾

<b>DATA LAKE NAME</b> bugbash-2de-190819-130359-o35-dd	<b>DATA LAKE STATUS</b> Running	<b>REASON</b> DataLake is running	<span>Atlas</span> <span>Ranger</span>
---	------------------------------------	--------------------------------------	--

Data Hub Clusters
Data Lake Cluster

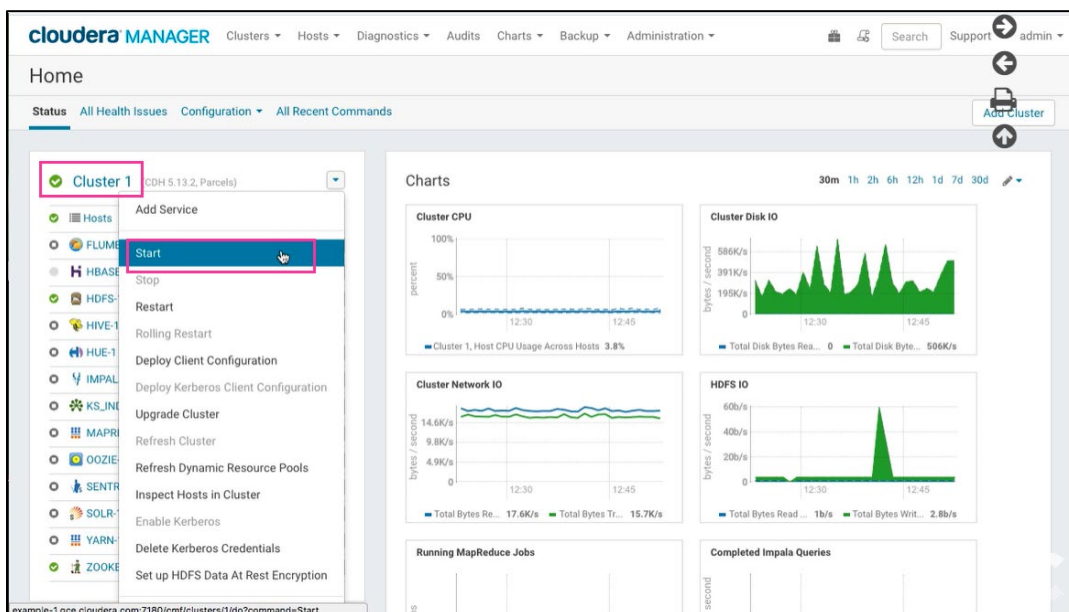
Data Hub Clusters
Create Data Hub Cluster

	Status	Name ↓	Data Hub Type	Version	Node Count	Created
<input type="checkbox"/>	Create in progress	bugbash-2de-190819-130359-o35-dw2	CDP 1.0 - Data Engineering: Apache Spark, Apache Hive, Apache Oozie	CDH 7.0.0	6	08/19/19, 7:06 AM PDT
<input checked="" type="checkbox"/>	Running	bugbash-2de-190819-130359-o35-dw1	CDP 1.0 - Data Engineering: Apache Spark, Apache Hive, Apache Oozie	CDH 7.0.0	6	08/19/19, 7:06 AM PDT

<https://docs.cloudera.com/data-hub/cloud/getting-started-tutorial/topics/dh-tutorial-monitor.html>

44. Cloudera Management Console allows a user to create and manage clusters by, for example, the user can click on “start” icon (e.g., hyperlinked with a URL that defines a type of event and identifies the network-based agent). In response to using the URL, the Cloudera server is

instructed to invoke and start execution of the network-based agent, i.e., the cluster. A new window opens which shows the status of starting the cluster.



[https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/cm\\_mc\\_start\\_stop\\_cluster.html](https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/cm_mc_start_stop_cluster.html)

Provision Data Hub Cluster

Provision on-demand workload clusters with the combination of applications for various business needs such as enterprise data warehouse management and data science operations.

Selected Environment  
bugbash-2de-130359-srx

Cluster Definition Custom

Services

Select the Cluster Definition option to create your cluster quickly using one of the prescriptive cluster definitions included by default or one of your previously created custom cluster definitions.

Cluster Definition\*  
CDP 1.0 - Data Engineering template

Hdfs 3.0.0 Hive 3.1.0 Hae 4.3.0 Ljvy 2.4.0 Oozie 5.1.0  
Spark 2.4.0 Yarn 3.0.0 Zeppelin 0.8.0 ZooKeeper 3.4.5

General Settings

Cluster Name\*  
bugbash-2de-190819-130359-035-dw

Tags  
You may optionally add tags, which will help you find your cluster-related resources, such as VMs, in your cloud provider account.  
Add

Advanced Options

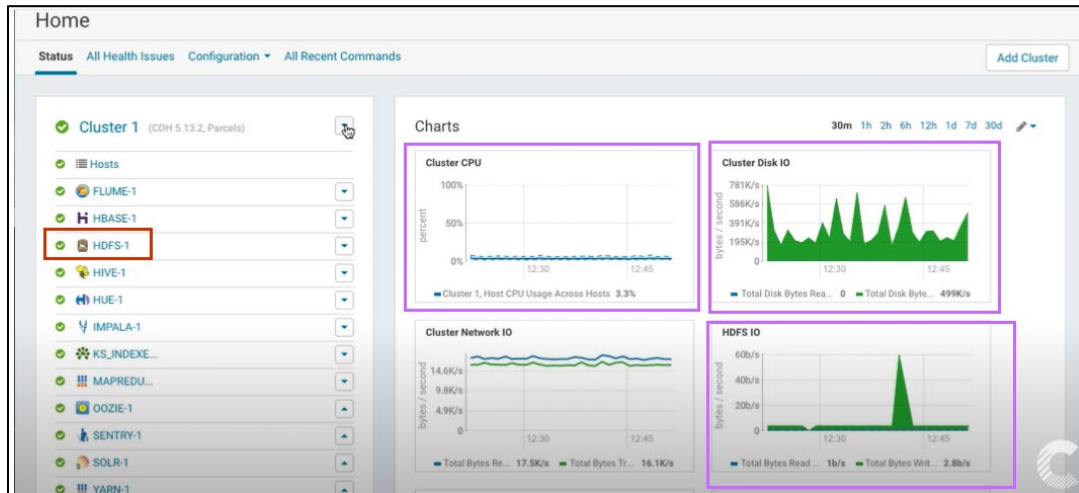
PROVISION CLUSTER SAVE AS NEW DEFINITION SHOW CLI COMMAND SHOW GENERATED BLUEPRINT

- a. Select Cluster Definition.
- b. In the Services section, select a specific cluster definition, for example Data Engineering for AWS.
- c. Under General Settings > Cluster Name, provide some name for your cluster.

3 Click on Provision Cluster to trigger cluster creation.

<https://docs.cloudera.com/data-hub/cloud/getting-started-tutorial/topics/dh-tutorial-create-cluster.html>

45. As shown below, the cluster executes multiple services such as HDFS, Hive, etc. as defined for the cluster by the user.



See *Stopping, Starting, and Restarting a Cluster*, CLUSTERA, INC., available via YouTube at [https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/cm\\_mc\\_start\\_stop\\_cluster.html](https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/cm_mc_start_stop_cluster.html).

46. The resources for these services are allocated as per the Cloudera YARN architecture.

### YARN Resource Allocation

You can manage your cluster capacity using the Capacity Scheduler in YARN. You can use the Capacity Scheduler's `DefaultResourceCalculator` or the `DominantResourceCalculator` to allocate available resources.

The fundamental unit of scheduling in YARN is the *queue*. The *capacity* of each queue specifies the percentage of cluster resources available for applications submitted to the queue. You can set up queues in a hierarchy that reflects the database structure, resource requirements, and access restrictions required by the organizations, groups, and individuals who use the cluster resources.

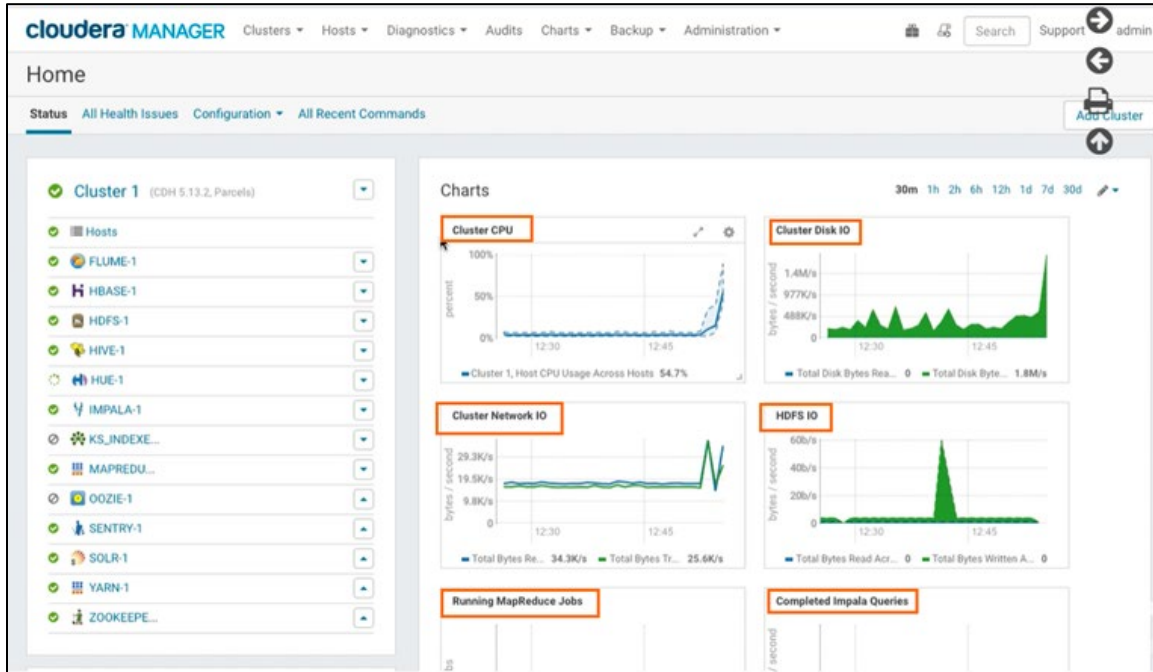
You can use the default resource calculator when you want the resource calculator to consider only the available memory for resource calculation. When you use the default resource calculator (`DefaultResourceCalculator`), resources are allocated based on the available memory.

[https://docs.cloudera.com/HDPDocuments/HDP3/HDP-3.1.0/data-operating-system/content/about\\_yarn\\_resource\\_allocation.html](https://docs.cloudera.com/HDPDocuments/HDP3/HDP-3.1.0/data-operating-system/content/about_yarn_resource_allocation.html)

47. A result of the cluster creation and start operations (e.g., resource utilization status) is communicated by the Cloudera server over a network communication link. As shown below,



when the services execute their specific tasks, the Cloudera management console communicates the resource utilization graphs to the user.



[https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/cm\\_mc\\_start\\_stop\\_cluster.html](https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/cm_mc_start_stop_cluster.html)

48. The Asserted Patents, including at least claim 11 of the '488 patent, cover Accused Instrumentalities of Defendant that practice in a computing environment (e.g., the Cloudera Data Platform) a method to automate the validation of equipment and/or processes for use in a pharmaceutical and/or bio-technology manufacturing facility. As shown below, the Cloudera Data Platform (CDP) is being used for pharma and biotech applications to automate validation equipment or processes, e.g., “deploy data lakes environments on-demand,” “manage [] healthcare data business at petabyte scale,” and “the implementation of the [CDP] enables complex machine

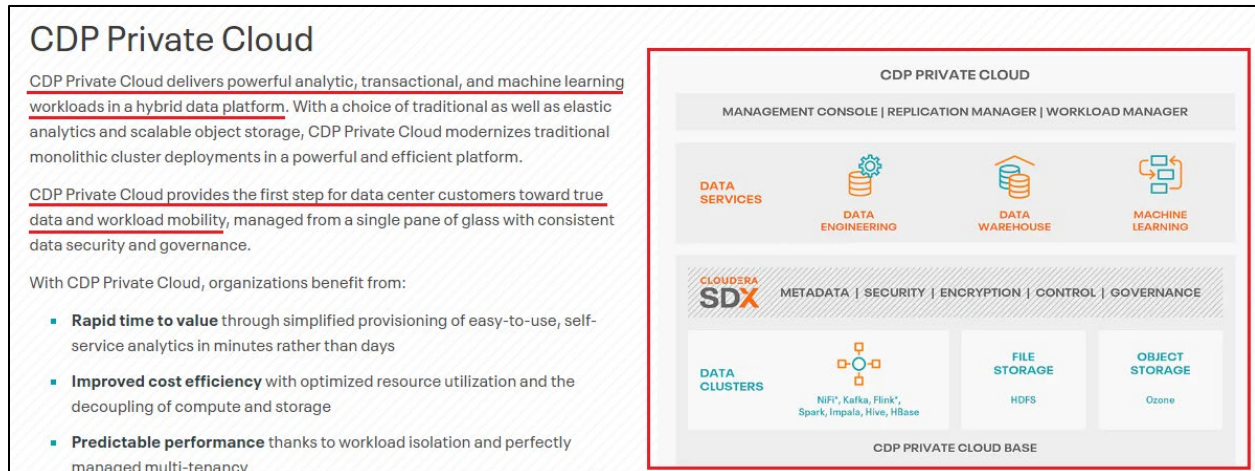
learning (ML) and artificial intelligence (AI) on petabytes of data to deliver actionable intelligence back to the point of care.”

**How Cloudera and IQVIA Help Pharma and Biotech Organizations**

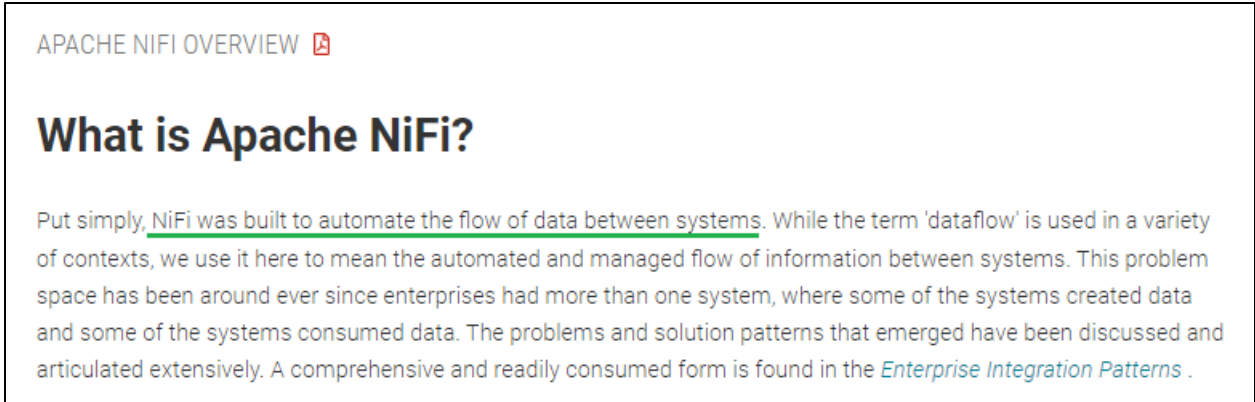
IQVIA's Platform-as-a-Service (PaaS) offering, built on Cloudera technology, offers clients the ability to deploy data lake environments on-demand. Clients can leverage the environment that IQVIA has built to power their business when they do not have the time or resources for an enterprise-level implementation. IQVIA builds and deploys the environments in days as opposed to weeks or months with all the same privacy, security and governance controls that were implemented to manage IQVIA's healthcare data business at petabyte scale. IQVIA's implementation of the Cloudera Data Platform (CDP) enables complex machine learning (ML) and artificial intelligence (AI) on petabytes of data to deliver actionable intelligence back to the point of care.

<https://www.cloudera.com/content/dam/www/marketing/resources/solution-briefs/cloudera-and-iqvia.pdf?daqp=true>

49. As shown below, the Cloudera Data Platform allows a user to automate the data flow validation process by using Apache NiFi.



<https://www.cloudera.com/products/cloudera-data-platform.html?tab=1>



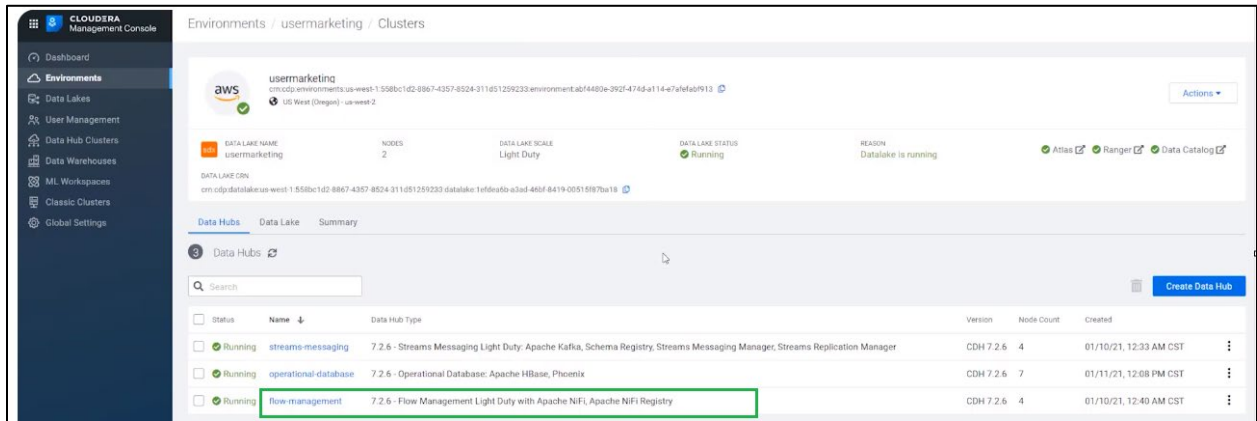
<https://docs.cloudera.com/cfm/2.0.1/nifi-overview/topics/nifi-what-is-apache-nifi.html>

50. As described below, Apache NiFi checks the validation of a processor properties for CDP.



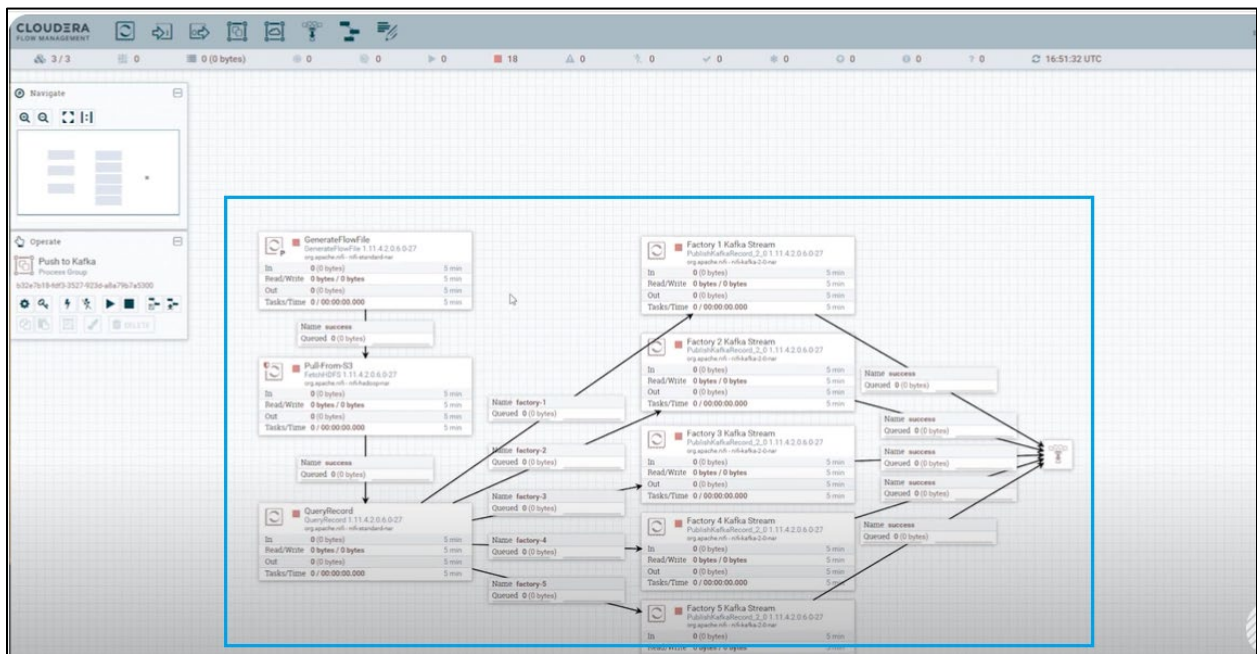
<https://docs.cloudera.com/cfm/2.1.3/nifi-dev-guide/topics/nifi-developer-guide-validator.html>

51. The Cludera Management Console displays to the user the Data Hub clusters that are running for a particular project. As shown below, a data hub relating to a data flow is established as a Flow Management instance utilizing Apache NiFi.



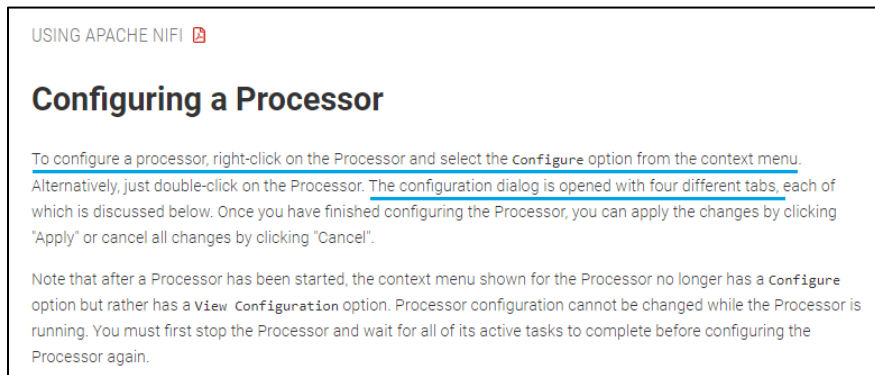
See *Collecting Data Using NiFi and Kafka on CDP Public Cloud*, CLUDERA, INC., available via YouTube at <https://www.youtube.com/watch?v=lrV-EwD4G8w>.

52. As shown below, Cludera Data Platform, via the Flow Management component, allows a user to automate the data flow validation process by using Apache NiFi.

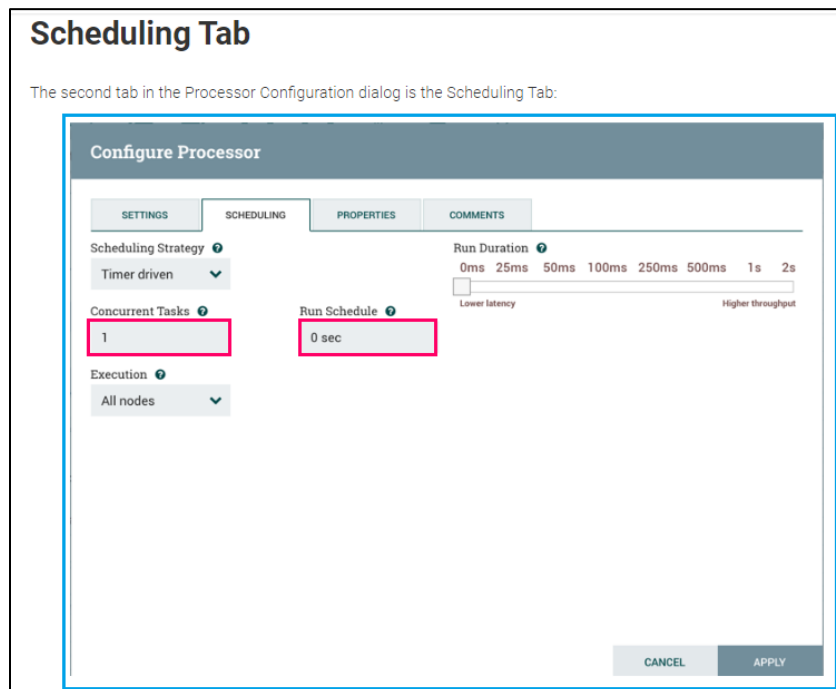


See *Collecting Data Using NiFi and Kafka on CDP Public Cloud*, CLUDERA, INC., available via YouTube at <https://www.youtube.com/watch?v=lrV-EwD4G8w>.

53. Cloudera’s Flow Management provides a user interface capable of accepting and/or displaying data representative of validation processing and/or validation workflow management information. The NiFi user interface, for example, provides the user, an option to enter values for configuration processor properties, settings and scheduling parameters that are required for creating a validation workflow. These configurations allow a user to automate the data flow validation process by using Apache NiFi.



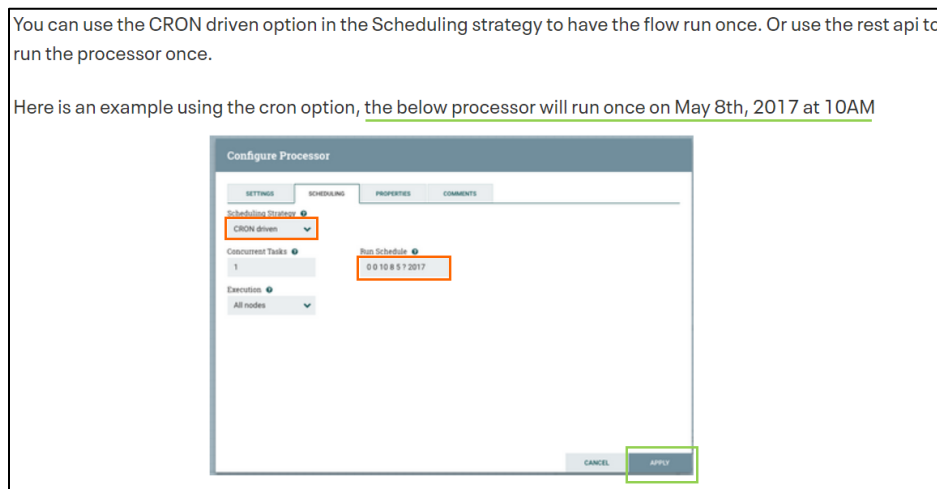
[https://docs.cloudera.com/cfm/2.1.3/nifi-user-guide/topics/nifi-user-guide-configuring\\_a\\_processor.html](https://docs.cloudera.com/cfm/2.1.3/nifi-user-guide/topics/nifi-user-guide-configuring_a_processor.html)



<https://docs.cloudera.com/cfm/2.1.3/nifi-user-guide/topics/nifi-user-guide-scheduling-tab.html>

54. The Cloudera Platforms further provide a validation processing engine (e.g., Cloudera server for Apache NiFi), said validation processing engine comprising at least one processing rule (e.g., rule requiring valid parameters, conditions, etc.) that operates on validation processing information selected through said user interface (e.g., Cloudera Apache NiFi UI) to produce validation protocol information (e.g., a validation property).

55. As shown below, Cloudera Data Platform allows a user to automate the data flow validation process by using Apache NiFi. For example, a “CRON driven option” allows a processor to run once at a scheduled time.



<https://community.cloudera.com/t5/Support-Questions/How-to-configure-a-processor-to-run-only-once/m-p/222956>

56. The NiFi user interface provides the user, an option to enter values for configuration properties, settings and scheduling parameters that are required for creating validation workflow. These values, such as the scheduling value shown below, should be valid values as specified by NiFi to create a validation property. The component is configured to run, only if the properties and other scheduling parameters are valid.

### Scheduling Strategy

CRON driven: When using the CRON driven scheduling mode, the Processor is scheduled to run periodically, similar to the Timer driven scheduling mode. However, the CRON driven mode provides significantly more flexibility at the expense of increasing the complexity of the configuration. The CRON driven scheduling value is a string of six required fields and one optional field, each separated by a space. These fields are:

<https://docs.cloudera.com/cfm/2.1.3/nifi-user-guide/topics/nifi-user-guide-scheduling-tab.html>


Field	Valid values
Seconds	0-59
Minutes	0-59
Hours	0-23
Day of Month	1-31
Month	1-12 or JAN-DEC
Day of Week	1-7 or SUN-SAT
Year (optional)	empty, 1970-2099

You typically specify values one of the following ways:

- **Number:** Specify one or more valid value. You can enter more than one value using a comma-separated list.
- **Range:** Specify a range using the <number>-<number> syntax.
- **Increment:** Specify an increment using <start value>/<increment> syntax. For example, in the Minutes field, 0/15 indicates the minutes 0, 15, 30, and 45.

<https://docs.cloudera.com/cfm/2.1.3/nifi-user-guide/topics/nifi-user-guide-scheduling-tab.html>

57. Moreover, each Processor, Reporting Task, or ControllerService uses “properties” defined by a “PropertyDescriptor.” A property includes “its name, description of the property, an optional default value, validation logic.”

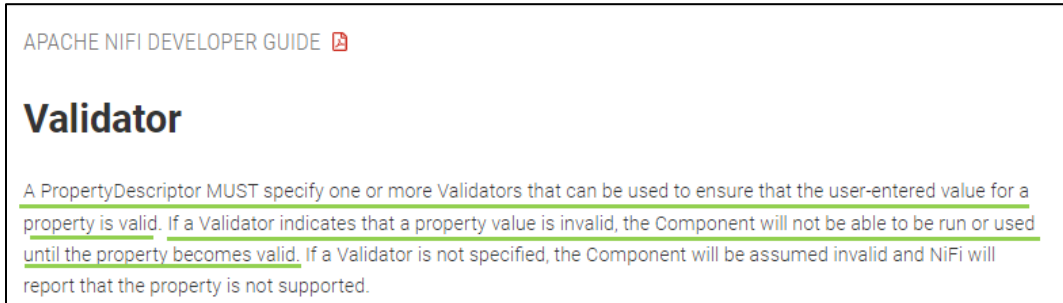
APACHE NIFI DEVELOPER GUIDE 

### PropertyDescriptor

PropertyDescriptor defines a property that is to be used by a Processor, ReportingTask, or ControllerService. The definition of a property includes its name, a description of the property, an optional default value, validation logic, and an indicator as to whether or not the property is required in order for the Processor to be valid. PropertyDescriptors are created by instantiating an instance of the `PropertyDescriptor.Builder` class, calling the appropriate methods to fill in the details about the property, and finally calling the `build` method.

[https://docs.cloudera.com/cfm/2.1.3/nifi-dev-guide/topics/nifi-developer-guide-property\\_descriptor.html](https://docs.cloudera.com/cfm/2.1.3/nifi-dev-guide/topics/nifi-developer-guide-property_descriptor.html)

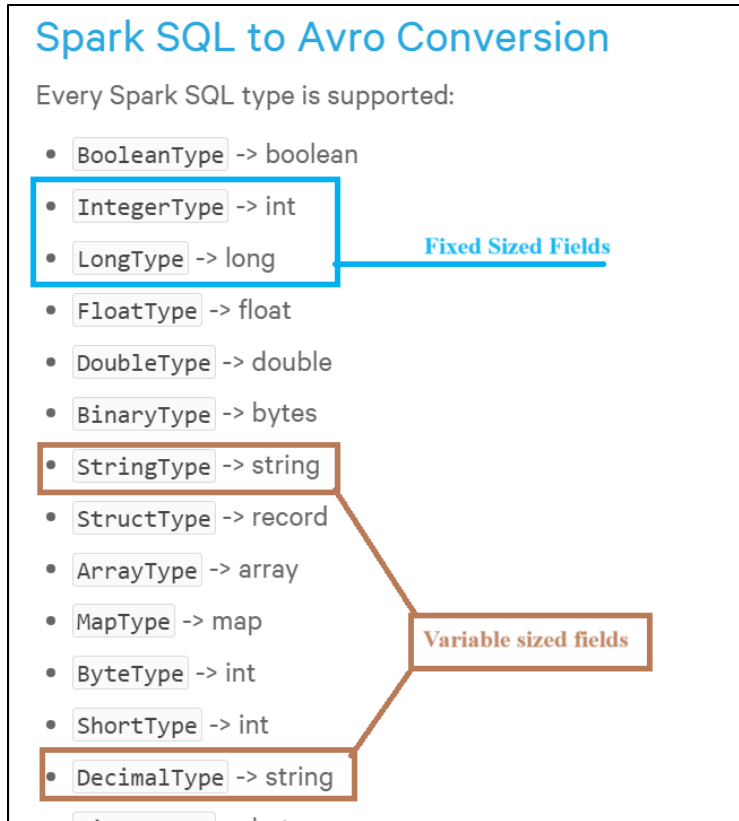
58. The PropertyDescriptor includes “one or more Validators [that] ensure that the user-entered value for a property is valid.”



<https://docs.cloudera.com/cfm/2.1.3/nifi-dev-guide/topics/nifi-developer-guide-validator.html>

59. The Asserted Patents, including at least claim 1 of the '897 patent, cover Accused Instrumentalities of Defendant that practice a method for improving compression of data comprising the steps of arranging data on a mix format physical layout. At least Defendant's Cloudera Enterprise platform arranges data according to at least an Avro data file schema. For example, the Cloudera Enterprise platform writes data, such as Spark SQL, to Avro files in a defined Avro data format, i.e., having a schema and a container file.





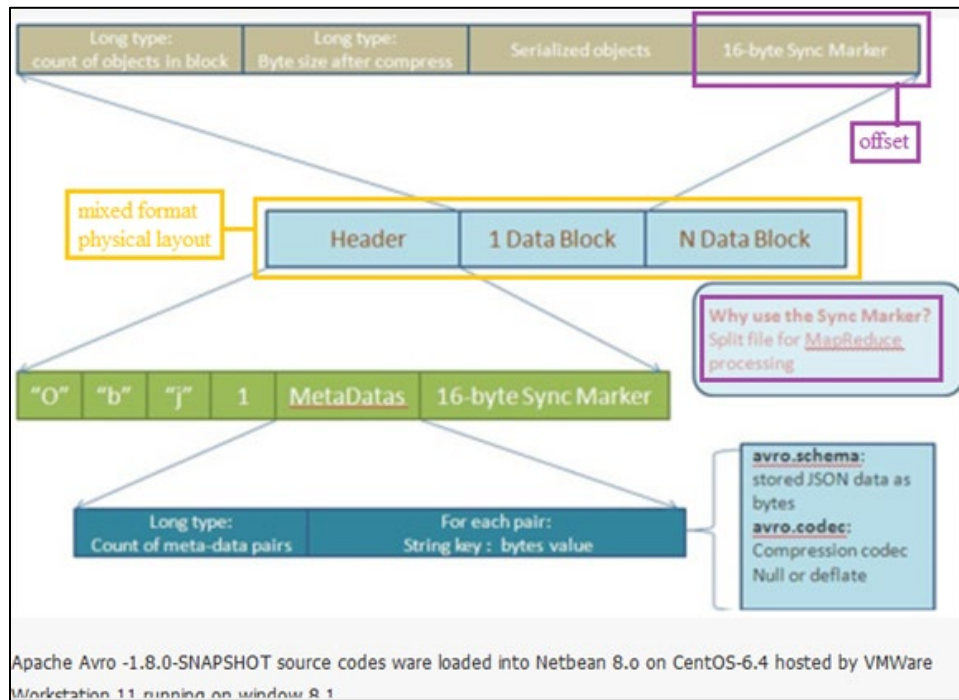
[https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/spark\\_avro.html#concept\\_hsz\\_nvn\\_45\\_fig\\_i4v\\_vp5\\_st](https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/spark_avro.html#concept_hsz_nvn_45_fig_i4v_vp5_st)

### Object Container Files

Avro includes a simple object container file format. A file has a schema, and all objects stored in the file must be written according to that schema, using binary encoding. Objects are stored in blocks that may be compressed. Synchronization markers are used between blocks to permit efficient splitting of files for MapReduce processing.

<https://avro.apache.org/docs/1.11.1/specification/>

60. As shown below, in the Avro data format, data is arranged into “data blocks,” which comprise a mixed format physical layout containing fixed-sized fields and variable-sized fields. Synchronization markers (i.e., offsets) are written between blocks, so that a file can be split.



<https://mingqin.wordpress.com/2014/12/29/apache-avro-object-container-file-format-examination/>

61. As shown below, the Cloudera Enterprise platform divides the data, for example SQL data, in the Avro data format into fixed-sized fields and variable-sized fields. According to the schema, the Avro container file, provided by the Cloudera Enterprise platform, stores data of different data types such as fixed-sized integer fields, i.e., int, long, etc., and variable sized strings, i.e., varchar string, etc.

A Schema is represented in JSON by one of:

- A JSON string, naming a defined type.
- A JSON object, of the form:

```
{"type": "typeName" ...attributes...}
```

where *typeName* is either a primitive or derived type name, as defined below. Attributes not defined in this document are permitted as metadata, but must not affect the format of serialized data.

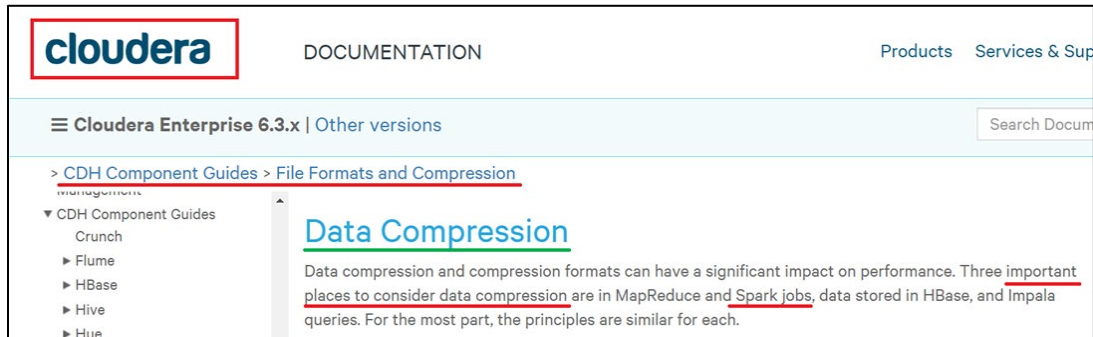
<https://avro.apache.org/docs/1.11.1/specification/>

The set of primitive type names is:

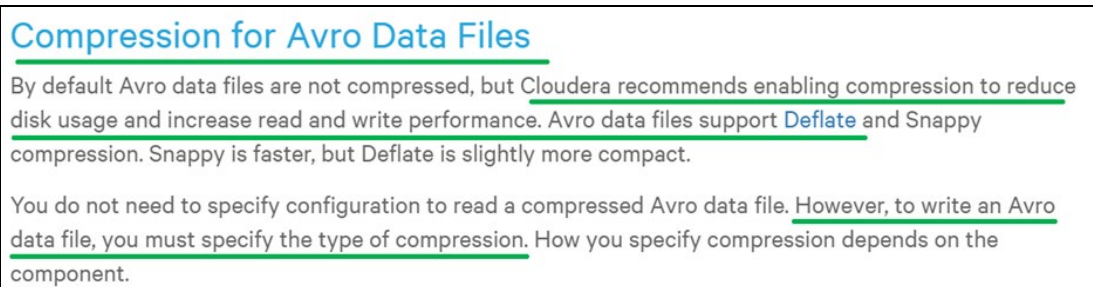
- *null*: no value
  - *boolean*: a binary value
  - *int*: 32-bit signed integer
  - *long*: 64-bit signed integer
  - *float*: single precision (32-bit) IEEE 754 floating-point number
  - *double*: double precision (64-bit) IEEE 754 floating-point number
  - *bytes*: sequence of 8-bit unsigned bytes
  - *string*: unicode character sequence
- Fixed-Sized Fields
- Variable-Sized Fields

<https://avro.apache.org/docs/1.11.1/specification/>

62. As shown below, the Cloudera Enterprise platform compresses the Avro data of the variable-sized fields and the fixed-sized fields. For example, Avro data files support “Deflate” compression.



[https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/introduction\\_compression.html#concept\\_wlk\\_hgy\\_pv](https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/introduction_compression.html#concept_wlk_hgy_pv)



[https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/cdh\\_ig\\_avro\\_usage.html#topic\\_26](https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/cdh_ig_avro_usage.html#topic_26)

63. The Asserted Patents, including at least claim 1 of the '961 patent, cover Accused Instrumentalities of Defendant that practice a method of aggregating information content. Defendant provides a Cloudera Search tool that “is Apache Solr fully integrated in the Cloudera platform,” which “provides easy, natural language access to data stored in or ingested into Hadoop, HBase, or cloud storage.”

## Cloudera Search Overview

Cloudera Search provides easy, natural language access to data stored in or ingested into Hadoop, HBase, or cloud storage. End users and other web services can use full-text queries and faceted drill-down to explore text, semi-structured, and structured data as well as quickly filter and aggregate it to gain business insight without requiring SQL or programming skills.

Cloudera Search is Apache Solr fully integrated in the Cloudera platform, taking advantage of the flexible, scalable, and robust storage system and data processing frameworks included in Cloudera Data Platform (CDP). This eliminates the need to move large data sets across infrastructures to perform business tasks. It further enables a streamlined data pipeline, where search and text matching is part of a larger workflow.

Using Cloudera Search with the CDP infrastructure provides:

- Simplified infrastructure
- Better production visibility and control
- Quicker insights across various data types
- Quicker problem resolution
- Simplified interaction and platform access for more users and use cases beyond SQL
- Scalability, flexibility, and reliability of search services on the same platform used to run other types of workloads on the same data
- A unified security model across all processes with access to your data
- Flexibility and scale in ingest and pre-processing options

<https://docs.cloudera.com/runtime/7.2.0/search-overview/search-overview.pdf>

64. The Cloudera Search tool accesses the Solr API, as a content aggregator, “to provide scalable and reliable search services.” Search queries are transmitted by the Cloudera Search tool to the content aggregator, i.e., “submitted to Solr through the standard Solr API, or through the simple search GUI application.”

## Search Guide

Cloudera Search integrates with CDH and uses Apache Solr to provide scalable and reliable search services. Search makes these services available to end users through tools that use familiar access and querying models.

- Search integrates with the existing CDH ecosystem, so data can be stored, shared, and accessed using the various CDH components. This helps to prevent data silos and minimizes expensive data movement.
- Search provides access to data stored in CDH without requiring the Java skills required for MapReduce jobs or the SQL skills required for Impala queries.
- Search returns results typically within seconds, rather than the minutes or more that are often required for MapReduce jobs to complete.
- Search allows you to select the information you want to index. You can optimize indexes for completeness, size, data types, and so on.

<https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/search.html>

### How Cloudera Search Works

In near real-time indexing use cases, such as log or event stream analytics, Cloudera Search indexes events that are streamed through Apache Kafka, Spark Streaming, or HBase. Fields and events are mapped to standard Solr indexable schemas. Lucene indexes the incoming events and the index is written and stored in standard Lucene index files in HDFS.

The indexes are loaded from HDFS to Solr cores, exactly like Solr would have read from local disk. The difference in the design of Cloudera Search is the robust, distributed, and scalable storage layer of HDFS, which helps eliminate costly downtime and allows for flexibility across workloads without having to move data. Search queries can then be submitted to Solr through either the standard Solr API, or through a simple search GUI application, included in Cloudera Search, which can be deployed in Hue.

Cloudera Search batch-oriented indexing capabilities can address needs for searching across batch uploaded files or large data sets that are less frequently updated and less in need of near-real-time indexing. It can also be conveniently used for re-indexing (a common pain point in stand-alone Solr) or ad-hoc indexing for on-demand data exploration. Often, batch indexing is done on a regular basis (hourly, daily, weekly, and so on) as part of a larger workflow.

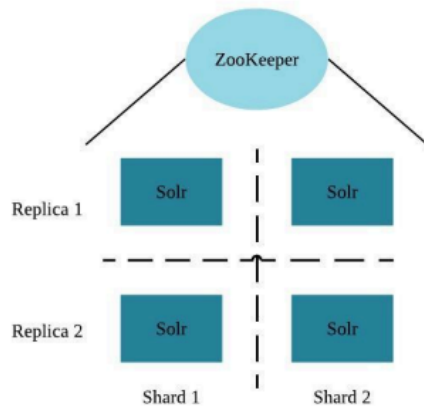
<https://docs.cloudera.com/runtime/7.2.0/search-overview/search-overview.pdf>

65. The Cloudera Search tool transmits the search queries from the content aggregator, i.e., the Solr API, to a “distributed service” operating on a set of servers (as remote agents). “[E]ach server is responsible for a portion of the searchable data.”

### Cloudera Search Architecture

Cloudera Search runs as a distributed service on a set of servers, and each server is responsible for a portion of the searchable data. The data is split into smaller pieces, copies are made of these pieces, and the pieces are distributed among the servers. This provides two main advantages:

- Dividing the content into smaller pieces distributes the task of indexing the content among the servers.
- Duplicating the pieces of the whole allows queries to be scaled more effectively and enables the system to provide higher levels of availability.



<https://docs.cloudera.com/runtime/7.2.0/search-overview/search-overview.pdf>

Each Cloudera Search server can handle requests independently. Clients can send requests to index documents or perform searches to any Search server, and that server routes the request to the correct server.

Each Search deployment requires:

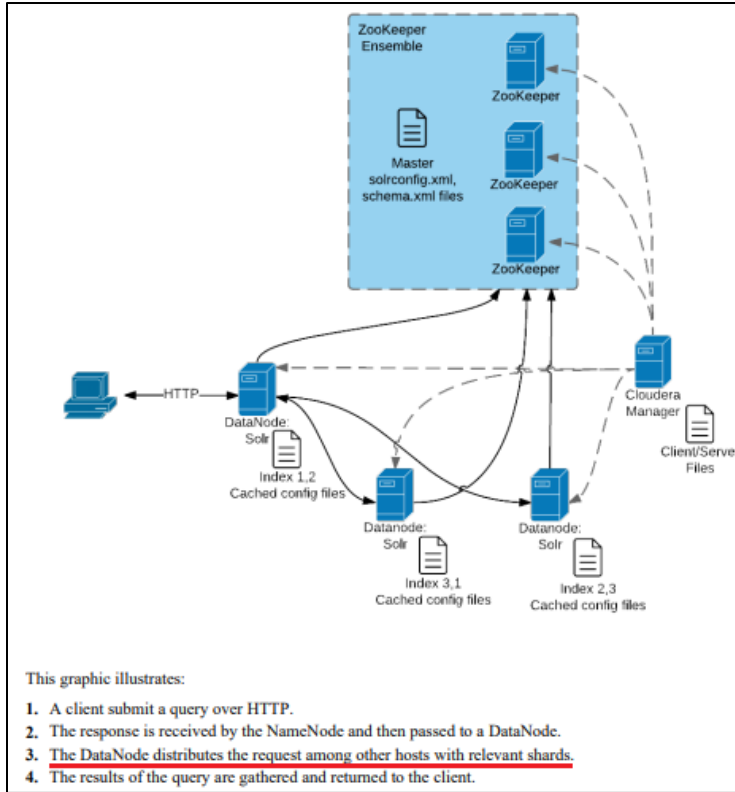
- ZooKeeper on at least one host. You can install ZooKeeper, Search, and HDFS on the same host.
- HDFS on at least one, but as many as all hosts. HDFS is commonly installed on all cluster hosts.
- Solr on at least one but as many as all hosts. Solr is commonly installed on all cluster hosts.

More hosts with Solr and HDFS provides the following benefits:

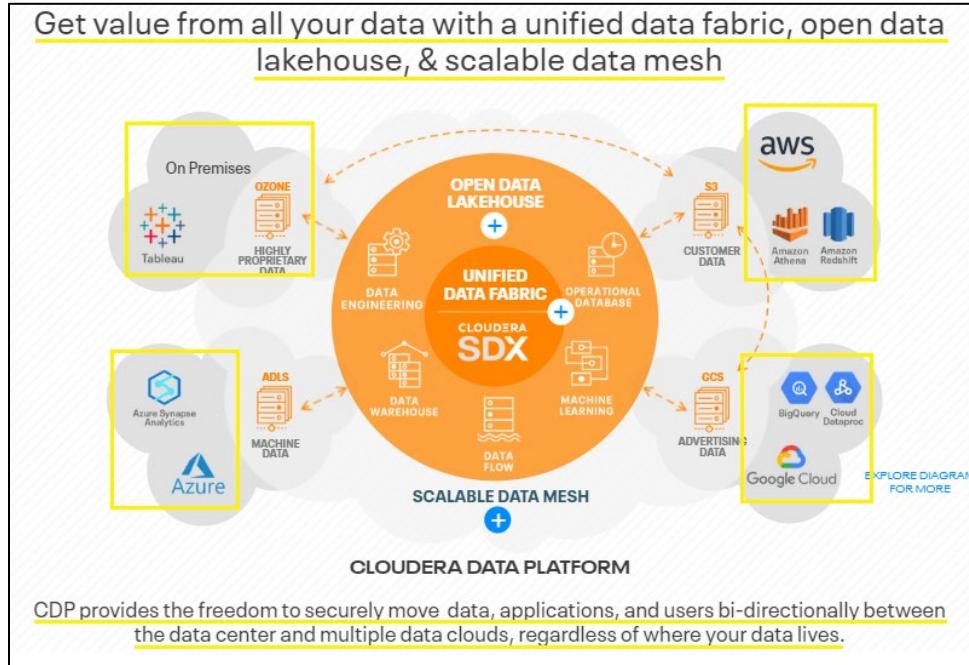
- More Search servers processing requests.
- More Search and HDFS collocation increasing the degree of data locality. More local data provides faster performance and reduces network traffic.

<https://docs.cloudera.com/runtime/7.2.0/search-overview/search-overview.pdf>

66. Cloudera Search’s distributed service operating on the servers, i.e., as remote agents, are located on one of a plurality of distinct networks. For example, the Cloudera Data Platform “provides the freedom to securely move data, applications, and users bi-directionally between the data center and multiple data clouds [i.e., distinct networks], regardless of where your data lives.” A plurality of cluster nodes (e.g., master node, worker nodes, and compute node) can be created on “on-premises” or on third-party cloud networks, each providing a distinct network for remote agents.



<https://docs.cloudera.com/runtime/7.2.0/search-overview/search-overview.pdf>



<https://www.cloudera.com/products/cloudera-data-platform.html>



### Data Hub overview

Data Hub is a service for launching and managing workload clusters powered by Cloudera Runtime (Cloudera's unified open source distribution including the best of CDH and HDP). Data Hub clusters can be created on AWS, Microsoft Azure, and Google Cloud Platform.

<https://docs.cloudera.com/data-hub/cloud/overview/dh-overview.pdf>

67. In the Cloudera Search architecture, “data is split into smaller pieces, copies are made of these pieces, and the pieces are distributed among the servers.” In responding to a search query, the “Data Node distributes the [search] request among other hosts with relevant shards.” “Each Cloudera Search server,” i.e., a remote agent, “handle[s] requests independently,” and “[c]lients can send requests to index documents or perform searches to any Search server, and that server routes the request to the correct server.”

Each Cloudera Search server can handle requests independently. Clients can send requests to index documents or perform searches to any Search server, and that server routes the request to the correct server.

Each Search deployment requires:

- ZooKeeper on at least one host. You can install ZooKeeper, Search, and HDFS on the same host.
- HDFS on at least one, but as many as all hosts. HDFS is commonly installed on all cluster hosts.
- Solr on at least one but as many as all hosts. Solr is commonly installed on all cluster hosts.

More hosts with Solr and HDFS provides the following benefits:

- More Search servers processing requests.
- More Search and HDFS collocation increasing the degree of data locality. More local data provides faster performance and reduces network traffic.

<https://docs.cloudera.com/runtime/7.2.0/search-overview/search-overview.pdf>

68. The Cloudera Search servers (i.e., agents) transmit the search results to the content aggregator (i.e., the Solr API). Via the Solr API, the Cloudera Search tool gathers the “results of the query” to process the results. For example, “[e]nd users and other web services [i.e., clients] can use full-text queries and faceted drill-down,” as rules and standards designated by the client “to explore text, semi-structured, and structured data as well as quickly filter and aggregate it to gain business insight.” This processed information is “returned to the client.”

## Cloudera Search Overview

Cloudera Search provides easy, natural language access to data stored in or ingested into Hadoop, HBase, or cloud storage. End users and other web services can use full-text queries and faceted drill-down to explore text, semi-structured, and structured data as well as quickly filter and aggregate it to gain business insight without requiring SQL or programming skills.

Cloudera Search is [Apache Solr](#) fully integrated in the Cloudera platform, taking advantage of the flexible, scalable, and robust storage system and data processing frameworks included in Cloudera Data Platform (CDP). This eliminates the need to move large data sets across infrastructures to perform business tasks. It further enables a streamlined data pipeline, where search and text matching is part of a larger workflow.

<https://docs.cloudera.com/runtime/7.2.0/search-overview/search-overview.pdf>

This graphic illustrates:

1. A client submit a query over HTTP.
2. The response is received by the NameNode and then passed to a DataNode.
3. The DataNode distributes the request among other hosts with relevant shards.
4. The results of the query are gathered and returned to the client.

Also notice that the:

<https://docs.cloudera.com/runtime/7.2.0/search-overview/search-overview.pdf>

69. The Asserted Patents, including claim 1 of the '474 patent, cover Accused Instrumentalities of Defendant that practice a method of operating a distributed processing system using a distributed search request processing system over a cloud network, i.e., the Cloudera Data Platform. For example, “Cloudera Search runs as a distributed service on a set of servers, and each server is responsible for a portion of the searchable data.” A multiplicity of distributed devices, i.e., Solr servers, process search requests, from a plurality of client systems.

## Cloudera Search Overview

Cloudera Search provides easy, natural language access to data stored in or ingested into Hadoop, HBase, or cloud storage. End users and other web services can use full-text queries and faceted drill-down to explore text, semi-structured, and structured data as well as quickly filter and aggregate it to gain business insight without requiring SQL or programming skills.

Cloudera Search is Apache Solr fully integrated in the Cloudera platform, taking advantage of the flexible, scalable, and robust storage system and data processing frameworks included in Cloudera Data Platform (CDP). This eliminates the need to move large data sets across infrastructures to perform business tasks. It further enables a streamlined data pipeline, where search and text matching is part of a larger workflow.

Using Cloudera Search with the CDP infrastructure provides:

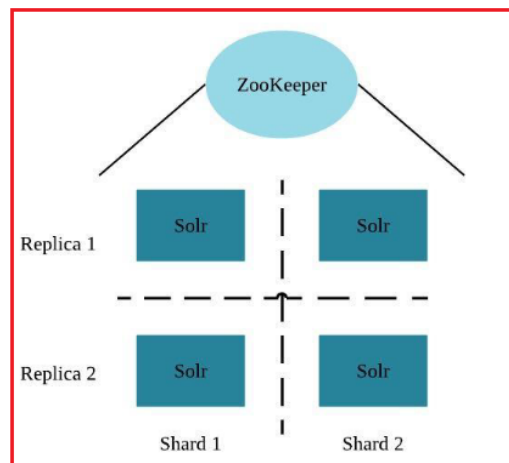
- Simplified infrastructure
- Better production visibility and control
- Quicker insights across various data types
- Quicker problem resolution
- Simplified interaction and platform access for more users and use cases beyond SQL
- Scalability, flexibility, and reliability of search services on the same platform used to run other types of workloads on the same data
- A unified security model across all processes with access to your data
- Flexibility and scale in ingest and pre-processing options

<https://docs.cloudera.com/runtime/7.2.0/search-overview/search-overview.pdf>

## Cloudera Search Architecture

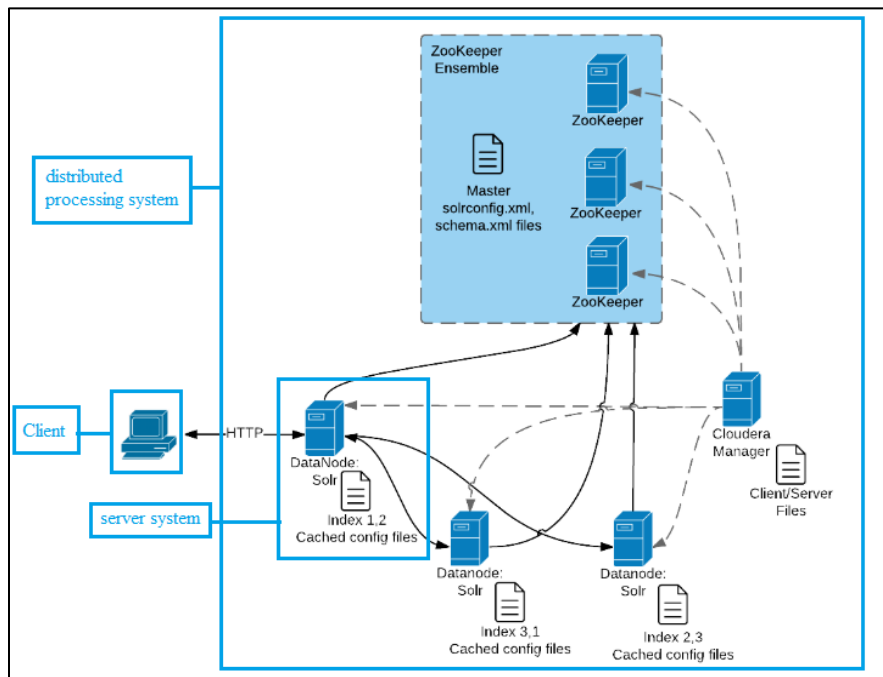
Cloudera Search runs as a distributed service on a set of servers, and each server is responsible for a portion of the searchable data. The data is split into smaller pieces, copies are made of these pieces, and the pieces are distributed among the servers. This provides two main advantages:

- Dividing the content into smaller pieces distributes the task of indexing the content among the servers.
- Duplicating the pieces of the whole allows queries to be scaled more effectively and enables the system to provide higher levels of availability.



<https://docs.cloudera.com/runtime/7.2.0/search-overview/search-overview.pdf>

70. As shown below, clients or users can send queries to a search server using hostname of the Solr Server and ports. The query is the “received by NameNode of an HDFS, and passed to data node.” The DataNode then distributes the query to other servers with relevant shards to process the search request, i.e., a first workload.



<https://docs.cloudera.com/runtime/7.2.0/search-overview/search-overview.pdf>

This graphic illustrates:

1. **A client submit a query over HTTP.**
2. The response is received by the NameNode and then passed to a DataNode.
3. The DataNode distributes the request among other hosts with relevant shards.
4. The results of the query are gathered and returned to the client.

<https://docs.cloudera.com/runtime/7.2.0/search-overview/search-overview.pdf>

71. As shown below, Cloudera Search indexes all ingested data to make it searchable. Lucene indexes data and stores it in Lucene index files in HDFS. Lucene files have parameters such as JavaInt or fields that indicate the location of data to be searched. These indexes are sent from HDFS to Solr cores or Solr servers (i.e., host distributed devices) that are used for searching. The HDFS is installed on at least one server in Cloudera search architecture but may be installed on all

servers. When a user sends a query via HTTP, the query is received by NameNode of an HDFS, and passed to the data node. The DataNode then distributes the query to servers with relevant shards.

### HDFS Key Features

HDFS is a fault-tolerant and self-healing distributed filesystem designed to turn a cluster of industry-standard servers into a massively scalable pool of storage. Developed specifically for large-scale data processing workloads where scalability, flexibility, and throughput are critical, HDFS accepts data in any format regardless of schema, optimizes for high-bandwidth streaming, and scales to proven deployments of 100PB and beyond.

<p><b>Hadoop Scalable:</b></p> <p>HDFS is designed for massive scalability, so you can store unlimited amounts of data in a single platform. As your data needs grow, you can simply add more servers to linearly scale with your business.</p>	<p><b>Flexibility:</b></p> <p>Store data of any type — structured, semi-structured, unstructured — without any upfront modeling. Flexible storage means you always have access to full-fidelity data for a wide range of analytics and use cases.</p>	<p><b>Reliability:</b></p> <p>Automatic, tunable replication means multiple copies of your data are always available for access and protection from data loss. Built-in fault tolerance means servers can fail but your system will remain available for all workloads.</p>
---	---	---

<https://www.cloudera.com/products/open-source/apache-hadoop/hdfs-mapreduce-yarn.html>

### How Cloudera Search Works

In near real-time indexing use cases, such as log or event stream analytics, Cloudera Search indexes events that are streamed through Apache Kafka, Spark Streaming, or HBase. Fields and events are mapped to standard Solr indexable schemas. Lucene indexes the incoming events and the index is written and stored in standard Lucene index files in HDFS.

The indexes are loaded from HDFS to Solr cores, exactly like Solr would have read from local disk. The difference in the design of Cloudera Search is the robust, distributed, and scalable storage layer of HDFS, which helps eliminate costly downtime and allows for flexibility across workloads without having to move data. Search queries can then be submitted to Solr through either the standard Solr API, or through a simple search GUI application, included in Cloudera Search, which can be deployed in Hue.

<https://docs.cloudera.com/runtime/7.2.0/search-overview/search-overview.pdf>

### How Search Uses Existing Infrastructure

Any data already in a CDP deployment can be indexed and made available for query by Cloudera Search. For data that is not stored in CDP, Cloudera Search provides tools for loading data into the existing infrastructure, and for indexing data as it is moved to HDFS or written to Apache HBase.

By leveraging existing infrastructure, Cloudera Search eliminates the need to create new, redundant structures. In addition, Cloudera Search uses services provided by CDP and Cloudera Manager in a way that does not interfere with other tasks running in the same environment. This way, you can reuse existing infrastructure without the cost and problems associated with running multiple services in the same set of systems.

<https://docs.cloudera.com/runtime/7.2.0/search-overview/search-overview.pdf>

72. As shown below, the Cloudera Search tool accesses the data to be searched from the address indicated by the HDFS (i.e., its NameNode), as part of processing a search query, submitted “over HTTP.” The DataNode “distributes the request among other hosts with relevant shards” of data. The results of the query are “gathered and returned to the client.”

This graphic illustrates:

1. A client submit a query over HTTP.
2. The response is received by the NameNode and then passed to a DataNode.
3. The DataNode distributes the request among other hosts with relevant shards.
4. The results of the query are gathered and returned to the client.

<https://docs.cloudera.com/runtime/7.2.0/search-overview/search-overview.pdf>

73. As shown below, the Cloudera Search tool updates an index (e.g., the HDFS index) to include a storage address of storage coupled to a Solr server (i.e., a host distributed device). For example, each Solr server may “store indexes in an HDFS filesystem.” Content in the Cloudera Data Platform can “be indexed on demand, or it can be updated and indexed continuously.” “In near real-time indexing use cases...Cloudera Search indexes events that are streamed through Apache Kafka, Spark Streaming, or HBase.”

## Solr and HDFS - the block cache

Cloudera Search enables Solr to store indexes in an HDFS filesystem. To maintain performance, an HDFS block cache has been implemented using Least Recently Used (LRU) semantics. This enables Solr to cache HDFS index files on read and write, storing the portions of the file in JVM direct memory (off heap) by default, or optionally in the JVM heap.

<https://docs.cloudera.com/cdp-private-cloud-base/latest/search-tuning/topics/search-tuning-hdfs-block-cache.html>

## Cloudera Search Tasks and Processes

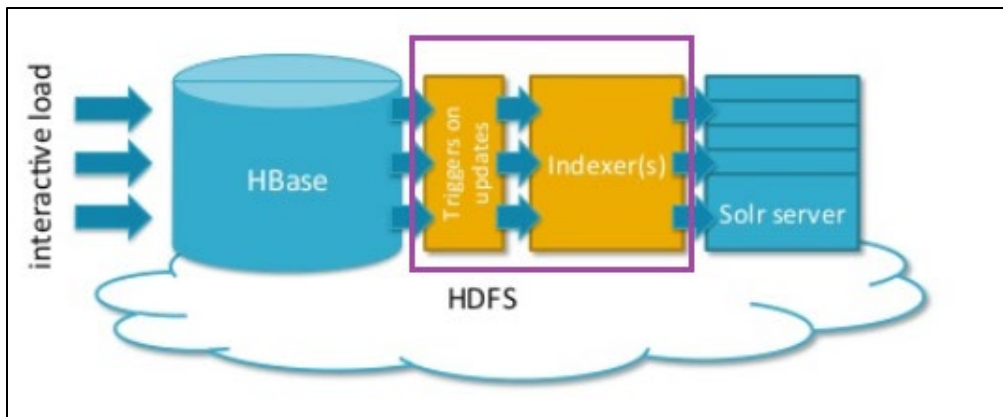
For content to be searchable, it must exist in Cloudera Data Platform (CDP) and be indexed. Content can either already exist in CDP and be indexed on demand, or it can be updated and indexed continuously. To make content searchable, first ensure that it is ingested or stored in CDP.

<https://docs.cloudera.com/runtime/7.2.0/search-overview/search-overview.pdf>

## How Cloudera Search Works

In near real-time indexing use cases, such as log or event stream analytics, Cloudera Search indexes events that are streamed through Apache Kafka, Spark Streaming, or HBase. Fields and events are mapped to standard Solr indexable schemas. Lucene indexes the incoming events and the index is written and stored in standard Lucene index files in HDFS.

<https://docs.cloudera.com/runtime/7.2.0/search-overview/search-overview.pdf>



<https://www.srcodes.com/nrt-near-real-time-indexing-cloudera-search-lily-hbase-indexer-morphline-apache-solr-lucene-tika-zookeeper/>

74. The Asserted Patents, including at least claims 2 and 14 of the '827 patent, cover Accused Instrumentalities of Defendant that practice a computer-implemented method that configures a distributed processing system, i.e., Cloudera's Data Hub service, with a plurality of distributed devices coupled to a network, i.e., nodes within clusters. As shown below, the Data Hub service provides a "cluster model in the cloud" that lets users "move existing workloads from on premises to the cloud or build directly in the cloud."

**OVERVIEW**

## Deploy a broad range of analytics in the public cloud quickly and easily.

CDP Data Hub is a powerful analytics service on **Cloudera Data Platform (CDP)** Public Cloud that makes it easier and faster to achieve high-value analytics from the Edge to AI in a familiar cluster model in the cloud. Featuring the widest range of analytical workloads—including streaming, ETL, data marts, databases, and machine learning—CDP Data Hub lets you easily move existing workloads from on premises to the cloud or build directly in the cloud.

The comprehensive, cloud-based solution is powered by Cloudera Runtime, a suite of integrated open source technologies, and built on **SDX**. It offers extensive choices in cluster shapes, workload types, pre-built templates, and configuration options, delivering an intuitive, customizable experience for users who are comfortable with traditional architectures.

<https://www.cloudera.com/products/data-hub.html?tab=1>

75. Cloudera’s Data Hub service includes a plurality of distributed devices that include client agents configured to process respective portions of a workload. For example, “Data Hub is a service for launching and managing workload clusters powered by Cloudera Runtime.” Cloudera’s NodeManager, as a client agent, “runs the components that are used for executing processing tasks.”

### Data Hub overview

Data Hub is a service for launching and managing workload clusters powered by Cloudera Runtime (Cloudera’s unified open source distribution including the best of CDH and HDP). Data Hub clusters can be created on AWS, Microsoft Azure, and Google Cloud Platform.

Data Hub includes a set of cloud optimized built-in templates for common workload types, as well as a set of options allowing for extensive customization based on your enterprise’s needs. Furthermore, it offers a set of convenient cluster management options such as cluster scaling, stop, restart, terminate, and more. All clusters are secured via wire encryption and strong authentication out of the box, and users can access cluster UIs and endpoints through a secure gateway powered by Apache Knox. Access to S3 cloud storage from Data Hub clusters is enabled by default (S3Guard is enabled and required in Runtime versions older than 7.2.2).

Data Hub provides complete workload isolation and full elasticity so that every workload, every application, or every department can have their own cluster with a different version of the software, different configuration, and running on different infrastructure. This enables a more agile development process.

Since Data Hub clusters are easy to launch and their lifecycle can be automated, you can create them on demand and when you don’t need them, you can return the resources to the cloud.

<https://docs.cloudera.com/data-hub/cloud/overview/dh-overview.pdf>

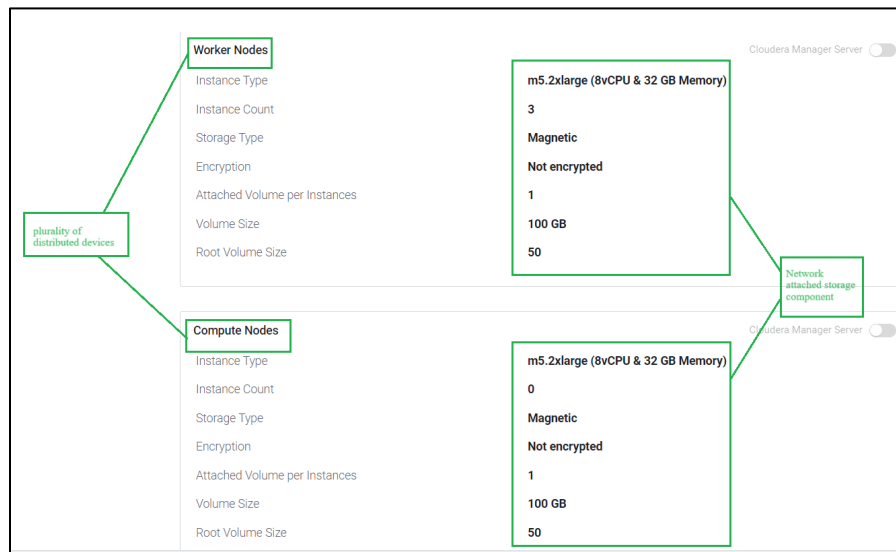


Host group	Description	Number of nodes
Master	The master host group runs the components for managing the cluster resources (including Cloudera Manager), storing intermediate data (e.g. HDFS), processing tasks, as well as other master components.	1
Worker	The worker host group runs the components that are used for executing processing tasks (such as NodeManager) and handling storing data in HDFS such as DataNode).	1+
Compute	The compute host group can optionally be used for running data processing tasks (such as NodeManager).	0+

plurality of distributed devices

<https://docs.cloudera.com/data-hub/cloud/overview/topics/dh-cluster-topology.html>

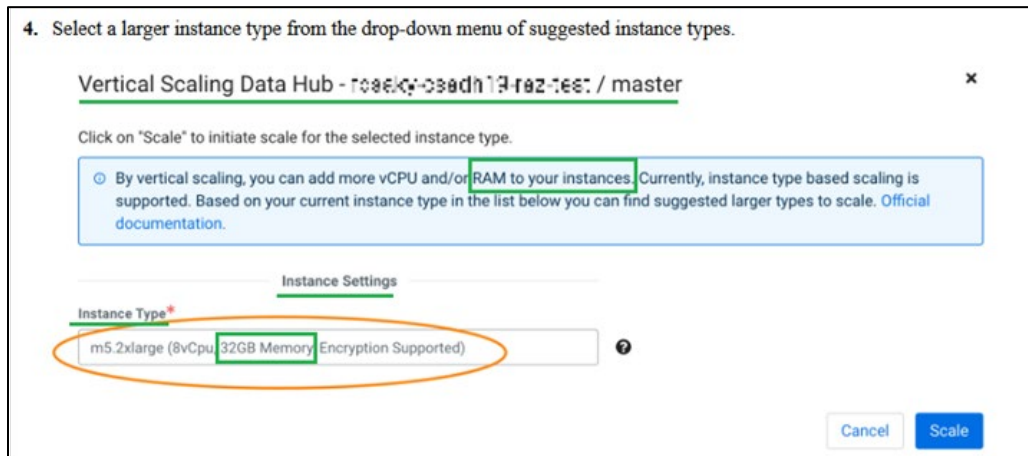
76. Cloudera’s Data Hub service includes client agents for particular distributed devices, e.g., YARN Node Managers hosted on worker or compute nodes within a multi-node cluster. As shown below, these nodes have corresponding software-based network attached storage (NAS) components, e.g., instances allocated to cluster nodes.



<https://www.cloudera.com/products/data-hub/cdp-tour-data-hub.html>

77. Cloudera’s software-based NAS components hosted on cluster nodes in Cloudera’s Data Hub service are configured to assess unused or under-utilized storage resources, e.g.,

resources, including storage, dedicated to each node that is an attached volume in in each instance. For example, the CDP provides a “Vertical Scaling Data Hub” that allows for the addition of “more vCPU and/or RAM to [a user’s] instances.” Moreover, “[s]electing a larger instance type adds more vCPU and/or RAM to your instances” and “[i]nstances can be scaled both up and down.”



<https://docs.cloudera.com/data-hub/cloud/manage-clusters/dh-manage-clusters.pdf>

## Vertically scaling instance types

If necessary, you can select a larger or smaller instance type for a Data Hub or Data Lake cluster after it has been deployed.

### Before you begin

You must stop the Data Lake or Data Hub cluster before you vertically scale any of the instances.

### About this task

Selecting a larger instance type adds more vCPU and/or RAM to your instances. Instances can be scaled both up and down, but scaling down to a smaller size requires 4 CPU and a minimum of 4 GB memory.

<https://docs.cloudera.com/data-hub/cloud/manage-clusters/topics/mc-vertically-scale-instances.html>

78. Cloudera’s Data Hub service represents the selected distributed devices comprised of software-based NAS component NAS devices with available storage resources related to unused and under-utilized storage resources. For example, Cloudera’s Data Hub service utilizes cluster nodes having an instance types that “configure the services on the cluster to use the additional or

reduced resources/memory.” Such instance types are represented as a NAS component having a storage resource, e.g., “100 GB Memory,” as shown below.

Parameter	Description
Cloudera Manager Server	You must select one node for Cloudera Manager Server by clicking the <input type="checkbox"/> button. The "Instance Count" for that host group must be set to "1". If you are using one of the default cluster templates, this is set by default.
Instance Type	Select an instance type. For information about instance types on AWS refer to <a href="#">Amazon EC2 Instance Types</a> in AWS documentation.
Instance Count	Enter the number of instances of a given type. Default is 1.
Storage Type	Select the volume type. The options vary by instance type and include: (1) Ephemeral (2) Magnetic (3) General Purpose SSD, (4) Throughput Optimized HDD. For more information about these options refer to <a href="#">Amazon EC2 Instance Store</a> in AWS documentation.  <div style="border: 1px solid #ccc; padding: 5px; background-color: #f9f9f9;"> <p><b>Note</b> Stopping and restarting Data Hub clusters using ephemeral storage is not supported.</p> </div>
Encryption	Under <b>Encryption Key</b> , you can select an existing encryption key. For more information, refer to <a href="#">EBS Encryption on AWS</a> .
Attached Volumes Per Instance	Enter the number of volumes attached per instance. Default is 1.
Volume Size	Enter the size in GB for each volume. Default is 100. <span style="float: right; border: 1px solid #ccc; padding: 2px;">storage resources</span>
Root Volume Size	This option allows you to increase or decrease the root volume size. Default is 100 GB. This option is

<https://docs.cloudera.com/data-hub/cloud/top-tasks/topics/dh-hardware-storage.html>

After you finish  
 After you have vertically scaled the cluster, configure the services on the cluster to use the additional or reduced resources/memory.

<https://docs.cloudera.com/data-hub/cloud/manage-clusters/topics/mc-vertically-scale-instances.html>

Master Nodes			
Instance Type	m5.2xlarge (8vCPU & 32 GB Memory)	Instance Count	1
Storage Type	Magnetic	Encryption	Not encrypted
Attached Volume per Instances	1	Volume Size	100 GB
Root Volume Size	50		

Worker Nodes			
Instance Type	m5.2xlarge (8vCPU & 32 GB Memory)	Instance Count	3
Storage Type	Magnetic	Encryption	Not encrypted
Attached Volume per Instances	1	Volume Size	100 GB
Root Volume Size	50		

Compute Nodes			
Instance Type	m5.2xlarge (8vCPU & 32 GB Memory)	Instance Count	0
Storage Type	Magnetic	Encryption	Not encrypted
Attached Volume per Instances	1	Volume Size	100 GB
Root Volume Size	50		

<https://www.cloudera.com/products/data-hub/cdp-tour-data-hub.html>

79. Cloudera’s Data Hub service processes data storage or access workloads by accessing data from or storing data for the client agent to a portion of the available amount of storage resources. As shown below, Cloudera’s Data Hub provides “hardware and storage” options that allow users to “customize the cloud provider specific cluster hardware and storage options.” Such settings include “instance type,” and “storage type,” and “volume size.” Moreover, Cloudera provides “Cloud Storage” options that allow users to “specify the base storage location used for YARN and Zeppelin.”

## Hardware and storage

The "Hardware and storage" options allow you to customize the cloud provider specific cluster hardware and storage options.

The following hardware and storage settings are available:

Parameter	Description
Cloudera Manager Server	You must select one node for Cloudera Manager Server by clicking the <input type="checkbox"/> button. The "Instance Count" for that host group must be set to "1". If you are using one of the default cluster templates, this is set by default.
Instance Type	Select an instance type. For information about instance types on AWS refer to <a href="#">Amazon EC2 Instance Types</a> in AWS documentation.
Instance Count	Enter the number of instances of a given type. Default is 1.
Storage Type	Select the volume type. The options vary by instance type and include: (1) Ephemeral (2) Magnetic (3) General Purpose SSD, (4) Throughput Optimized HDD. For more information about these options refer to <a href="#">Amazon EC2 Instance Store</a> in AWS documentation.  <div style="border: 1px solid #ccc; padding: 5px; background-color: #f9f9f9;"> <p><b>i Note</b> Stopping and restarting Data Hub clusters using ephemeral storage is not supported.</p> </div>
Encryption	Under <b>Encryption Key</b> , you can select an existing encryption key. For more information, refer to <a href="#">EBS Encryption on AWS</a> .
Attached Volumes Per Instance	Enter the number of volumes attached per instance. Default is 1.
Volume Size	Enter the size in GB for each volume. Default is 100.
Root Volume Size	This option allows you to increase or decrease the root volume size. Default is 100 GB. This option is useful if your custom image requires more space than the default 100 GB. If you use a custom Data Hub template specifying a root volume size smaller than 100GB, you may encounter an error.

<https://docs.cloudera.com/data-hub/cloud/top-tasks/topics/dh-hardware-storage.html>

## Cloud storage

The options on the "Cloud Storage" page allow you to optionally specify the base storage location used for YARN and Zeppelin.

<https://docs.cloudera.com/data-hub/cloud/top-tasks/topics/dh-cloud-storage.html>

80. Cloudera's Data Hub service enables a distributed device to function as a location distributed device to store location information associated with data stored by the distributed device through use of the client agents. For example, each worker engine "has a dedicated IP access with no possibility of port conflicts." Moreover, each host in a cluster has a "Name, IP address, [and]

rack ID.” These details provide location information associated with the data stored by the worker engine or host.

### Worker Network Communication

This section demonstrates some trivial examples of how two worker engines communicate with the master engine.

Workers are a low-level feature to help use higher level libraries that can operate across multiple hosts. As such, you will generally want to use workers only to launch the backends for these libraries.

To help you get your workers or distributed computing framework components talking to one another, every worker engine run includes an environmental variable `CML_MASTER_IP` with the fully addressable IP of the master engine. Every engine has a dedicated IP access with no possibility of port conflicts.

<https://docs.cloudera.com/machine-learning/saas/distributed-computing/topics/ml-worker-network-communication.html>

### Host Details

You can view details about each host from the status page for each host.

The host details include:

- Name, IP address, rack ID
- Health status of the host and last time the Cloudera Manager Agent sent a heartbeat to the Cloudera Manager Server
- Number of cores
- System load averages for the past 1, 5, and 15 minutes
- Memory usage
- File system disks, their mount points, and usage

**! Important**

If you have multiple mount points under the same device, then the available free space on that device is counted multiple times and adds to the total available disk space.

- Health test results for the host
- Charts showing a variety of metrics and health test results over time.
- Role instances running on the host and their health
- CPU, memory, and disk resources used for each role instance

[Viewing Host Details](#)

You can view detailed information about each host, such as name, IP address, and rack ID, and more from the All Hosts page.

81. In Cloudera’s Data Hub service, worker nodes that process workloads can function as stand-alone dedicated NAS devices. As shown below, hosts through the use of client agents (e.g., a Node Manager), can provide “complete workload isolation and full elasticity so that every workload, every application, or every department can have their own cluster with a different version of the software, different configuration, and running on different infrastructure.”

## Data Hub overview

Data Hub is a service for launching and managing workload clusters powered by Cloudera Runtime (Cloudera's unified open source distribution including the best of CDH and HDP). Data Hub clusters can be created on AWS, Microsoft Azure, and Google Cloud Platform.

Data Hub includes a set of cloud optimized built-in templates for common workload types, as well as a set of options allowing for extensive customization based on your enterprise's needs. Furthermore, it offers a set of convenient cluster management options such as cluster scaling, stop, restart, terminate, and more. All clusters are secured via wire encryption and strong authentication out of the box, and users can access cluster UIs and endpoints through a secure gateway powered by Apache Knox. Access to S3 cloud storage from Data Hub clusters is enabled by default (S3Guard is enabled and required in Runtime versions older than 7.2.2).

Data Hub provides complete workload isolation and full elasticity so that every workload, every application, or every department can have their own cluster with a different version of the software, different configuration, and running on different infrastructure. This enables a more agile development process.

Since Data Hub clusters are easy to launch and their lifecycle can be automated, you can create them on demand and when you don't need them, you can return the resources to the cloud.

The following diagram describes a simplified Data Hub architecture:

82. The Asserted Patents, including at least claim 1 of the '153 patent, cover Accused Instrumentalities of Defendant that practice a method of providing dynamic coordination of distributed client systems in a distributed computing platform. For example, Cloudera's Data Hub "is a service for launching and managing clusters powered by Cloudera Runtime," which "offers a set of convenient cluster management options such as cluster scaling, stop, restart, terminate, and more" and provides workload management "so that every workload, every application, or every department can have their own cluster with a different version of the software, different configuration, and running on different infrastructure." Each cluster provides a plurality of nodes (e.g., master node, worker nodes, and compute nodes), as a distributed computing platform of clusters and nodes that can be on the infrastructure of third-party cloud providers connected to the Cloudera Data Platform. The cluster nodes are configured with resources such as CPU cores, storage etc., which are then utilized by the node for processing workload. Cloudera provides its CDP management console for dynamic coordination of cluster resources on a server system (e.g., CDP servers) coupled to a network, e.g., the Cloudera Data Platform Data Cloud.

## Data Hub overview

Data Hub is a service for launching and managing workload clusters powered by Cloudera Runtime (Cloudera's unified open source distribution including the best of CDH and HDP). Data Hub clusters can be created on AWS, Microsoft Azure, and Google Cloud Platform.

Data Hub includes a set of cloud optimized built-in templates for common workload types, as well as a set of options allowing for extensive customization based on your enterprise's needs. Furthermore, it offers a set of convenient cluster management options such as cluster scaling, stop, restart, terminate, and more. All clusters are secured via wire encryption and strong authentication out of the box, and users can access cluster UIs and endpoints through a secure gateway powered by Apache Knox. Access to S3 cloud storage from Data Hub clusters is enabled by default (S3Guard is enabled and required in Runtime versions older than 7.2.2).

Data Hub provides complete workload isolation and full elasticity so that every workload, every application, or every department can have their own cluster with a different version of the software, different configuration, and running on different infrastructure. This enables a more agile development process.

Since Data Hub clusters are easy to launch and their lifecycle can be automated, you can create them on demand and when you don't need them, you can return the resources to the cloud.

<https://docs.cloudera.com/data-hub/cloud/overview/dh-overview.pdf>

Status	Name	Cloud Provider	Region	Data Lake	Time Created
Available	koeler	AWS	US East(N, Virginia)	Creating Stack	10/17/2019, 1:00:41 PM CDT
Available	abajwa-ire-01	AWS	EU (Ireland)	Running	10/17/2019, 12:29:26 PM CDT
Available	nismaily2	AWS	US East (Ohio)	Running	10/17/2019, 11:54:22 AM CDT
Available	brent	AWS	US West (N, California)	Running	10/17/2019, 11:15:07 AM CDT
Available	jazariah	AWS	US West (N, California)	Running	10/17/2019, 11:06:55 AM CDT
Available	cbove	AWS	US East(N, Virginia)	Running	10/17/2019, 11:00:34 AM CDT
Available	mchisam	AWS	US East (Ohio)	Running	10/17/2019, 11:05:30 AM CDT
Available	bhagan	AWS	US West (N, California)	Running	10/17/2019, 11:05:20 AM CDT
Available	lbrooks	AWS	EU (Ireland)	Running	10/17/2019, 11:05:15 AM CDT
Available	kat-bucket-njack	AWS	EU (Ireland)	Running	10/16/2019, 11:45:54 PM CDT
Available	ggoodson-demo-env	AWS	US East(N, Virginia)	Running	10/16/2019, 9:57:05 PM CDT
Available	cye-ohio-env	AWS	US East (Ohio)	Running	10/16/2019, 9:19:22 PM CDT
Creation Failed	psandbon5	AWS	EU (Ireland)	Provisioning Failed	10/16/2019, 4:13:19 PM CDT
Available	na-demo-env	AWS	US East(N, Virginia)	Not registered	10/16/2019, 1:57:14 PM CDT

<https://docs.cloudera.com/data-hub/cloud/top-tasks/topics/mc-creating-a-cluster.html>

## Workload clusters

All Data Hub clusters are workload clusters. These clusters are created for running specific workloads such as data engineering or data analytics.

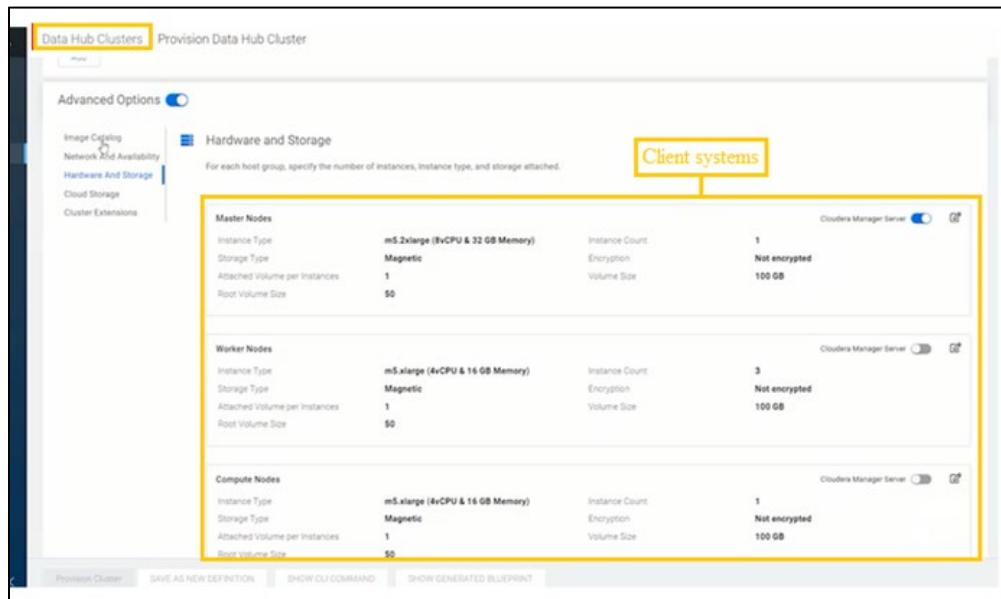
Data Hub clusters are powered by Cloudera Runtime. They can be ephemeral or long-running. Once a Data Hub cluster is created it can be managed by using the Management Console and Cloudera Manager.

<https://docs.cloudera.com/data-hub/cloud/overview/topics/dh-concept-workload-cluster.html>


83. Cloudera's Data Hub service provides the Cloudera Data Platform to manage a plurality of network-connected distributed client systems, i.e., the clusters and associated nodes. The nodes share resources having under-utilized capabilities and share intermediate data while



processing tasks associated with a workload. A new cluster can be created for processing a workload, in an environment (defining resources associated with an account), based on workload requirements. Nodes in a cluster (e.g., worker nodes) are configured with a Node Manager (e.g., YARN node manager) that is used for executing processing tasks. It also performs functions such as communicating with resource manager, checking resource utilization by the node, keeping track of node health etc.

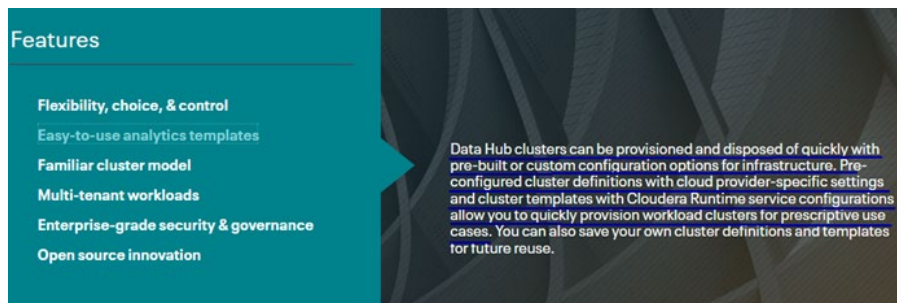


<https://docs.cloudera.com/data-hub/cloud/overview/topics/dh-overview.html>

Host group	Description	Number of nodes
Master	The master host group runs the components for managing the cluster resources (including Cloudera Manager), storing intermediate data (e.g. HDFS), processing tasks, as well as other master components.	1
Worker	The worker host group runs the components that are used for executing processing tasks (such as NodeManager) and handling storing data in HDFS such as DataNode).	1+ 
Compute	The compute host group can optionally be used for running	0+

<https://docs.cloudera.com/data-hub/cloud/overview/topics/dh-cluster-topology.html>

84. Cloudera’s Data Hub utilizes the CDP server system to distribute workloads for a project (i.e., data engineering and data analytics tasks) to clusters and their associated nodes, i.e., the client systems utilizing the NodeManager. The server system distributes initial project and poll parameters to the client systems. For example, initial project and poll parameters can be provided via at least “pre-built or custom configuration options for infrastructure.” Such “[p]re-configured cluster definitions with cloud provider-specific settings and cluster templates with Cloudera Runtime service configurations” allow users “to quickly provision workload clusters for prescriptive use cases.” Moreover, Cloudera’s Data Hub service provides autoscaling, via load-based or schedule based policies which “define policy parameters.” In a load-based policy, for example, auto-scaling “will scale the nodes within the selected range.” And “[a]fter an auto-scaling event occurs, the amount of time in minutes” as a parameter can be set “to wait before enforcing another scaling policy,” as a “cooldown” period.



The screenshot shows a 'Features' section with a teal background on the left and a dark background on the right. The teal section lists five features: Flexibility, choice, & control; Easy-to-use analytics templates; Familiar cluster model; Multi-tenant workloads; Enterprise-grade security & governance; and Open source innovation. The dark section contains a paragraph of text describing how Data Hub clusters can be provisioned and disposed of quickly with pre-built or custom configuration options for infrastructure.

**Features**

- Flexibility, choice, & control
- Easy-to-use analytics templates
- Familiar cluster model
- Multi-tenant workloads
- Enterprise-grade security & governance
- Open source innovation

Data Hub clusters can be provisioned and disposed of quickly with pre-built or custom configuration options for infrastructure. Pre-configured cluster definitions with cloud provider-specific settings and cluster templates with Cloudera Runtime service configurations allow you to quickly provision workload clusters for prescriptive use cases. You can also save your own cluster definitions and templates for future reuse.

<https://www.cloudera.com/products/data-hub.html?tab=3>

## Configuring autoscaling

To configure autoscaling, add a load-based or schedule-based policy to a cluster and define the policy parameters.

### Before you begin

If you are configuring a load-based autoscaling policy, you must set the YARN Node Decommission Timeout property to 30 seconds. Configure the following property in Cloudera Manager:

yarn\_resourcemanager\_nodemanager\_graceful\_decommission\_timeout\_secs

### Steps

1. From the CDP Management Console, click Data Hub Clusters and then select the cluster that you want to add an autoscaling policy to. You can add autoscaling policies to clusters after they have been created, not during the cluster creation process.
2. From the cluster details page, select the **Autoscale** tab and then click the slider button to enable autoscaling. Autoscaling is disabled by default.
3. Click **Add Autoscale Policy** and select **Load-Based** or **Schedule-Based**.

<https://docs.cloudera.com/data-hub/cloud/manage-clusters/topics/dh-configure-autoscaling.html>

**Add Autoscale Policy**

Load-Based  
Load-based auto-scaling will scale the nodes within the selected range.

Schedule-Based  
Schedule-based auto-scaling will scale the node on present schedules.

**Name\***  
Please enter the name

**Host Group\***  
compute

**Target\***  
min: 0 — max: 100

**Cooldown**  
After an auto scaling event occurs, the amount of time in minutes to wait before enforcing another scaling policy. This means that the scaling events scheduled during cooldown time are dropped.

2

Cancel Add

project and poll parameters

<https://docs.cloudera.com/data-hub/cloud/manage-clusters/topics/dh-autoscale.html>

85. Cloudera’s Data Hub service receives poll communications from the client systems (the nodes running the NodeManager) during processing of project workloads. For example, the NodeManager communicates pending (e.g., real-time) workload demand and available capacity of a given cluster, as part of applying an auto-scaling policy, to the Resource Manager—the NodeStatusUpdater “sends information about the resources available on the nodes” and “provide[s] updates on container statuses.” The NodeManager utilizes these communications to provide dynamic snapshot information of the current project status to the Resource Manager.

**1. NodeStatusUpdater**

On startup, this component registers with the RM and sends information about the resources available on the nodes. Subsequent NM-RM communication is to provide updates on container statuses - new containers running on the node, completed containers, etc.

In addition the RM may signal the NodeStatusUpdater to potentially kill already running containers.

<https://blog.cloudera.com/apache-hadoop-yarn-nodemanager/>

86. Cloudera’s Data Hub service analyzes poll communications to determine whether to make a modification to initial project and poll parameters. For example, the Resource Manager assesses “pending demand and available capacity,” as part of application of an auto-scaling policy. Based on the assessment, CDP auto scales a cluster, i.e., modifies the initial and poll parameters, by suspending or resuming nodes, “as workload demand requires.” Based on these parameters, autoscaling is performed as a response to the assessment, including decommissioning or addition of clusters is performed.

Load-based autoscaling suspends and resumes (stops and starts) instances on the cloud provider to increase or decrease capacity for nodes running NodeManagers (for example, the compute host group), based upon YARN’s assessment of pending demand and available capacity. Load-based autoscaling can help control costs while providing quick, on-demand cluster capacity when you need it (within a few minutes).

When you configure a load-based autoscaling policy, you choose a minimum and maximum number of nodes for the host group. The maximum number of nodes determines how many instances are provisioned, but these instances are suspended and resumed as the workload demand requires. The policy will not provision instances beyond the maximum range of nodes that you define, regardless of the demand on the cluster.

You also define a cooldown period, which is the amount of time in minutes to wait before another autoscaling operation is performed.

<https://docs.cloudera.com/data-hub/cloud/manage-clusters/topics/dh-autoscale.html>

Parameter	Description
Name	Enter a unique name for the policy.
Hostgroup	<u>Select the host group that you want to scale. The list of available host groups is determined by which host groups include services that can be scaled.</u>
Target	<u>Enter a minimum and maximum number of nodes for the policy. The maximum number of nodes determines how many instances are provisioned, but these instances are suspended and resumed as the workload demand requires. The policy will not provision instances beyond the maximum range of</u>
Parameter	Description
	nodes that you define, regardless of the demand on the cluster.

<https://docs.cloudera.com/data-hub/cloud/manage-clusters/dh-manage-clusters.pdf>

MANAGING CLUSTERS 

## Autoscaling clusters

Autoscaling is a feature that adjusts the capacity of cluster nodes running YARN by automatically increasing or decreasing, or suspending and resuming, the nodes in a host group. You can enable autoscaling based either on a schedule that you define, or the real-time demands of your workloads.

<https://docs.cloudera.com/data-hub/cloud/manage-clusters/topics/dh-autoscale.html>

87. Depending on the analysis of the poll communications, Cloudera's Data Hub service sends a poll response to the client systems to modify the initial and poll parameters. For example, as described above, the Cloudera Data Platform may perform scaling-up or down operation on cluster nodes based on the assessment, modifying the initial and poll parameters. Moreover, if the cluster has any node failures on instances running YARN ResourceManager, or the ClouderaManager node; or there is an ongoing cluster upgrade, auto scaling is not performed as per the user defined parameters (cool down time), also modifying the initial and poll parameters.

### General behaviors

- Clusters can perform one upscale or downscale operation at a time.
- A cluster will continue to accept jobs while it is running, regardless of any in-progress upscale or downscale operations.
- Only one autoscale policy type (either load-based or schedule-based) can be configured for a single host group, in a single cluster, at a time.
- Autoscaling is available for host groups with nodes running YARN NodeManager only (and optionally client/GATEWAY components, but not any other service components).
- If there are not enough nodes available to match the requested scale operation, the operation will proceed on however many nodes are available (for example, during a request for a 10 node scale-up, if the cluster loses 1 node, the operation will proceed with scaling-up 9 nodes instead of 10).
- Autoscaling will be disabled if the cluster has any node failures on instances running YARN ResourceManager, or the ClouderaManager node.
- After scaling down, nodes will show up as UNHEALTHY in Cloudera Manager. This is expected. While scaling down, stopped nodes are put into maintenance mode, to suppress alerts.

<https://docs.cloudera.com/data-hub/cloud/manage-clusters/topics/dh-autoscaling-behavior.html>

Cloudera Management Console has the following message: "Cloudera Manager reported health ... ..hostname1: [This host is in maintenance mode.], hostname2: [This host is in maintenance mode.]"

The cluster moves to 'Node Failure' state, and autoscaling is disabled. This typically indicates that some nodes are in the 'STARTED' state on the cloud-provider, but they are in 'Maintenance Mode' in Cloudera Manager. To remediate this, find the affected nodes in Cloudera Manager, and perform the steps in **Scaleup: Commission Services via ClouderaManager** above. Wait for a few minutes for the Cloudera Management Console to sync state. After this, the cluster should move to the running state, and autoscale is re-enabled. Alternately, you can stop the specific nodes on the cloud provider.

<https://docs.cloudera.com/data-hub/cloud/manage-clusters/topics/dh-autoscale-manual-recovery.html>

88. Cloudera's Data Hub repeats the receiving, analyzing and sending functions, described above, to dynamically coordinate project activities of the plurality of client systems during project operations. For example, each time autoscaling is performed, according to a scheduled timeframe, nodes can be scaled up or down based on workload requirements. Other project activities such as commissioning services on added nodes or stop instances on suspended nodes are performed on cluster nodes, each time an autoscaling function is performed. Also, if a node failure is detected, auto scaling is automatically disabled.

Load-based autoscaling suspends and resumes (stops and starts) instances on the cloud provider to increase or decrease capacity for nodes running NodeManagers (for example, the compute host group), based upon YARN's assessment of pending demand and available capacity. Load-based autoscaling can help control costs while providing quick, on-demand cluster capacity when you need it (within a few minutes).

When you configure a load-based autoscaling policy, you choose a minimum and maximum number of nodes for the host group. The maximum number of nodes determines how many instances are provisioned, but these instances are suspended and resumed as the workload demand requires. The policy will not provision instances beyond the maximum range of nodes that you define, regardless of the demand on the cluster. You also define a cooldown period, which is the amount of time in minutes to wait before another autoscaling operation is performed.

<https://docs.cloudera.com/data-hub/cloud/manage-clusters/topics/dh-autoscale.html>

**COUNT I**

(INFRINGEMENT OF U.S. PATENT NO. 6,839,733)

89. Plaintiff incorporates paragraphs 1 through 88 herein by reference.

90. Plaintiff BYTEWEAVR is the assignee of the '733 patent, entitled "Network system extensible by users," with ownership of all substantial rights in the '733 patent, including the right to exclude others and to enforce, sue, and recover damages for past and future infringements.

91. The '733 patent is valid, enforceable, and was duly issued in full compliance with Title 35 of the United States Code. The '733 patent issued from U.S. Patent Application No. 09/712,712. The '733 patent was granted on January 1, 2004 and expired on or about October 23, 2018.

92. Defendant has directly and/or indirectly infringed (by inducing infringement) one or more claims of the '733 patent in this District and elsewhere in Texas and the United States.

93. On information and belief, Defendant designs, develops, manufactures, imports, distributes, offers to sell, sells, and uses the Accused Instrumentalities, including via the activities of Cloudera and its alter egos, intermediaries, agents, distributors, importers, partners, customers, subsidiaries, affiliates, and/or consumers.

94. Defendant has directly infringed the '733 patent via 35 U.S.C. § 271(a) by making, offering for sale, selling, importing and/or using the Accused Instrumentalities, their components, and/or products containing the same that incorporate the fundamental technologies covered by the '733 patent to, for example, its alter egos, intermediaries, agents, distributors, importers, partners, customers, subsidiaries, affiliates, and/or consumers. Furthermore, on information and belief, Defendant develops and designs the Accused Instrumentalities for U.S. consumers, makes and sells the Accused Instrumentalities outside of the United States, delivers those products and services to



related entities, subsidiaries, distribution partners, resellers, vendors, installers, customers and other related service providers in the United States, or in the case that it delivers the Accused Instrumentalities outside of the United States it does so intending and/or knowing that those products are destined for the United States and/or designing those products for sale and use in the United States, thereby directly infringing the '733 patent. *See, e.g., Lake Cherokee Hard Drive Techs., L.L.C. v. Marvell Semiconductor, Inc.*, 964 F. Supp. 2d 653, 658 (E.D. Tex. 2013) (denying summary judgment and allowing presentation to jury as to “whether accused products manufactured and delivered abroad but imported into the United States market by downstream customers ... constitute an infringing sale under § 271(a)”).

95. Furthermore, Defendant Cloudera has directly infringed the '733 patent through its direct involvement in the activities of its subsidiaries, and related entities and other U.S.-based subsidiaries (e.g., Hortonworks, Inc., Cloudera (Government Solutions), Inc., and Eventador), members, segments, companies, and/or brands of Defendant Cloudera, including by designing the Accused Instrumentalities for U.S. consumers and selling and offering for sale the Accused Instrumentalities directly to its related entities and importing the Accused Instrumentalities into the United States for its related entities. On information and belief, U.S.-based members, segments, companies, and/or brands conduct activities that constitute direct infringement of the '733 patent under 35 U.S.C. § 271(a) by importing, offering for sale, selling, and/or using those Accused Instrumentalities in the U.S. on behalf of and for the benefit of Defendant. Defendant is vicariously liable for the infringing conduct of members, segments, companies, and/or brands of Cloudera (under both the alter ego and agency theories). On information and belief, Defendant Cloudera and other U.S. based subsidiaries, members, segments, companies, and/or brands of Cloudera are essentially the same company. Moreover, Cloudera, as the parent company, has the right and ability

to control the infringing activities of those entities such that Defendant receives a direct financial benefit from that infringement.

96. For example, Defendant infringes claim 37 of the '733 patent via the Accused Instrumentalities, namely data management and analytics products and components, software, services, and processes such as the Cloudera Platforms and their components, including the Cloudera Enterprise, the Cloudera Data Platform, Data Hub, Runtime, Search, the Cloudera SDX Management Console, Cloudera Manager, CDH, Cloudera Flow Management, and Cloudera distributions of Apache Oozie, NiFi, YARN, Hue, Avro, Zookeeper and related data storage and compression techniques.

97. Those Accused Instrumentalities include a “method” comprising the limitations of claim 37. The technology discussion above and the example Accused Instrumentalities provide context for Plaintiff’s allegations that each of those limitations are met. For example, the Accused Instrumentalities include the steps of admitting a user to a network system wherein at least one agent is operable to consume a service resource while utilizing a service to perform a task for the user; and allowing the user to create, modify, or delete the agent within the network system.

98. At a minimum, Defendant has known of the '733 patent at least as early as the filing date of this Complaint.

99. Plaintiff BYTEWEAVR has been damaged as a result of Defendant’s infringing conduct described in this Count. Defendant is thus liable to BYTEWEAVR in an amount that adequately compensates BYTEWEAVR for its infringements, which, by law, cannot be less than a reasonable royalty, together with interest and costs as fixed by this Court under 35 U.S.C. § 284.

**COUNT II**

(INFRINGEMENT OF U.S. PATENT NO. 7,949,752)

100. Plaintiff incorporates paragraphs 1 through 99 herein by reference.

101. Plaintiff BYTEWEAVR is the assignee of the '752 patent, entitled "Network system extensible by users," with ownership of all substantial rights in the '752 patent, including the right to exclude others and to enforce, sue, and recover damages for past and future infringements.

102. The '752 patent is valid, enforceable, and was duly issued in full compliance with Title 35 of the United States Code. The '752 patent issued from U.S. Patent Application No. 10/995,159. The '752 patent was granted on May 24, 2011 and expired on or about Aug. 13, 2022.

103. Defendant has directly and/or indirectly infringed (by inducing infringement) one or more claims of the '752 patent in this District and elsewhere in Texas and the United States.

104. On information and belief, Defendant designs, develops, manufactures, imports, distributes, offers to sell, sells, and uses the Accused Instrumentalities, including via the activities of Cloudera and its alter egos, intermediaries, agents, distributors, importers, partners, customers, subsidiaries, affiliates, and/or consumers.

105. Defendant has directly infringed the '752 patent via 35 U.S.C. § 271(a) by making, offering for sale, selling, importing and/or using the Accused Instrumentalities, their components, and/or products containing the same that incorporate the fundamental technologies covered by the '752 patent to, for example, its alter egos, intermediaries, agents, distributors, importers, partners, customers, subsidiaries, affiliates, and/or consumers. Furthermore, on information and belief, Defendant develops and designs the Accused Instrumentalities for U.S. consumers, makes and sells the Accused Instrumentalities outside of the United States, delivers those products and services to related entities, subsidiaries, distribution partners, resellers, vendors, installers, customers and other

related service providers in the United States, or in the case that it delivers the Accused Instrumentalities outside of the United States it does so intending and/or knowing that those products are destined for the United States and/or designing those products for sale and use in the United States, thereby directly infringing the '752 patent. *See, e.g., Lake Cherokee Hard Drive Techs., L.L.C. v. Marvell Semiconductor, Inc.*, 964 F. Supp. 2d 653, 658 (E.D. Tex. 2013) (denying summary judgment and allowing presentation to jury as to “whether accused products manufactured and delivered abroad but imported into the United States market by downstream customers ... constitute an infringing sale under § 271(a)”).

106. Furthermore, Defendant Cloudera has directly infringed the '752 patent through its direct involvement in the activities of its subsidiaries, and related entities and other U.S.-based subsidiaries (e.g., Hortonworks, Inc., Cloudera (Government Solutions), Inc., and Eventador), members, segments, companies, and/or brands of Defendant Cloudera, including by designing the Accused Instrumentalities for U.S. consumers and selling and offering for sale the Accused Instrumentalities directly to its related entities and importing the Accused Instrumentalities into the United States for its related entities. On information and belief, U.S.-based members, segments, companies, and/or brands conduct activities that constitute direct infringement of the '752 patent under 35 U.S.C. § 271(a) by importing, offering for sale, selling, and/or using those Accused Instrumentalities in the U.S. on behalf of and for the benefit of Defendant. Defendant is vicariously liable for the infringing conduct of members, segments, companies, and/or brands of Cloudera (under both the alter ego and agency theories). On information and belief, Defendant Cloudera and other U.S. based subsidiaries, members, segments, companies, and/or brands of Cloudera are essentially the same company. Moreover, Cloudera, as the parent company, has the right and ability

to control the infringing activities of those entities such that Defendant receives a direct financial benefit from that infringement.

107. For example, Defendant infringes claim 24 of the '752 patent via the Accused Instrumentalities, namely data management and analytics products and components, software, services, and processes such as the Cloudera Platforms and their components, including the Cloudera Enterprise, the Cloudera Data Platform, Data Hub, Runtime, Search, the Cloudera SDX Management Console, Cloudera Manager, CDH, Cloudera Flow Management, and Cloudera distributions of Apache Oozie, NiFi, YARN, Hue, Avro, Zookeeper and related data storage and compression techniques.

108. Those Accused Instrumentalities include a “method” comprising the limitations of claim 24. The technology discussion above and the example Accused Instrumentalities provide context for Plaintiff’s allegations that each of those limitations are met. For example, the Accused Instrumentalities include the steps of receiving, using a computing device, data for creating a network-based agent; invoking, using the computing device, and in response to receiving a URL defining a type of event and identifying the network-based agent, execution of the network-based agent, wherein the invoking comprises using a service and a service resource configured to be consumed by the network-based agent for performing the operation, and wherein a discrete unit of the service resource is exhausted upon being consumed by the network-based agent; and communicating, using the computing device, a result of the operation over a network communication link.

109. At a minimum, Defendant has known of the '752 patent at least as early as the filing date of this Complaint.

110. Plaintiff BYTEWEAVR has been damaged as a result of Defendant's infringing conduct described in this Count. Defendant is thus liable to BYTEWEAVR in an amount that adequately compensates BYTEWEAVR for its infringements, which, by law, cannot be less than a reasonable royalty, together with interest and costs as fixed by this Court under 35 U.S.C. § 284.

### **COUNT III**

(INFRINGEMENT OF U.S. PATENT NO. 6,862,488)

111. Plaintiff incorporates paragraphs 1 through 110 herein by reference.

112. Plaintiff BYTEWEAVR is the assignee of the '488 patent, entitled "Automated validation processing and workflow management," with ownership of all substantial rights in the '488 patent, including the right to exclude others and to enforce, sue, and recover damages for past and future infringements.

113. The '488 patent is valid, enforceable, and was duly issued in full compliance with Title 35 of the United States Code. The '488 patent issued from U.S. Patent Application No. 10/190,368. The '488 patent was granted on March 1, 2005 and expired on or about April 9, 2023.

114. Defendant has directly and/or indirectly infringed (by inducing infringement) one or more claims of the '488 patent in this District and elsewhere in Texas and the United States.

115. On information and belief, Defendant designs, develops, manufactures, imports, distributes, offers to sell, sells, and uses the Accused Instrumentalities, including via the activities of Cloudera and its alter egos, intermediaries, agents, distributors, importers, partners, customers, subsidiaries, affiliates, and/or consumers.

116. Defendant has directly infringed the '488 patent via 35 U.S.C. § 271(a) by making, offering for sale, selling, importing and/or using the Accused Instrumentalities, their components, and/or products containing the same that incorporate the fundamental technologies covered by the

'488 patent to, for example, its alter egos, intermediaries, agents, distributors, importers, partners, customers, subsidiaries, affiliates, and/or consumers. Furthermore, on information and belief, Defendant develops and designs the Accused Instrumentalities for U.S. consumers, makes and sells the Accused Instrumentalities outside of the United States, delivers those products and services to related entities, subsidiaries, distribution partners, resellers, vendors, installers, customers and other related service providers in the United States, or in the case that it delivers the Accused Instrumentalities outside of the United States it does so intending and/or knowing that those products are destined for the United States and/or designing those products for sale and use in the United States, thereby directly infringing the '488 patent. *See, e.g., Lake Cherokee Hard Drive Techs., L.L.C. v. Marvell Semiconductor, Inc.*, 964 F. Supp. 2d 653, 658 (E.D. Tex. 2013) (denying summary judgment and allowing presentation to jury as to “whether accused products manufactured and delivered abroad but imported into the United States market by downstream customers ... constitute an infringing sale under § 271(a)”).

117. Furthermore, Defendant Cloudera has directly infringed the '488 patent through its direct involvement in the activities of its subsidiaries, and related entities and other U.S.-based subsidiaries (e.g., Hortonworks, Inc., Cloudera (Government Solutions), Inc., and Eventador), members, segments, companies, and/or brands of Defendant Cloudera, including by designing the Accused Instrumentalities for U.S. consumers and selling and offering for sale the Accused Instrumentalities directly to its related entities and importing the Accused Instrumentalities into the United States for its related entities. On information and belief, U.S.-based members, segments, companies, and/or brands conduct activities that constitute direct infringement of the '488 patent under 35 U.S.C. § 271(a) by importing, offering for sale, selling, and/or using those Accused Instrumentalities in the U.S. on behalf of and for the benefit of Defendant. Defendant is vicariously

liable for the infringing conduct of members, segments, companies, and/or brands of Cloudera (under both the alter ego and agency theories). On information and belief, Defendant Cloudera and other U.S. based subsidiaries, members, segments, companies, and/or brands of Cloudera are essentially the same company. Moreover, Cloudera, as the parent company, has the right and ability to control the infringing activities of those entities such that Defendant receives a direct financial benefit from that infringement.

118. For example, Defendant infringes claim 11 of the '488 patent via the Accused Instrumentalities, namely data management and analytics products and components, software, services, and processes such as the Cloudera Platforms and their components, including the Cloudera Enterprise, the Cloudera Data Platform, Data Hub, Runtime, Search, the Cloudera SDX Management Console, Cloudera Manager, CDH, Cloudera Flow Management, and Cloudera distributions of Apache Oozie, NiFi, YARN, Hue, Avro, Zookeeper and related data storage and compression techniques.

119. Those Accused Instrumentalities include, “[i]n a computing environment[,] a method to automate the validation of equipment and/or processes for use in a pharmaceutical and/or bio-technology manufacturing facility” comprising the limitations of claim 11. The technology discussion above and the example Accused Instrumentalities provide context for Plaintiff’s allegations that each of those limitations are met. For example, the Accused Instrumentalities include the steps of providing a user interface capable of accepting and/or displaying data representative of validation processing and/or validation workflow management information, wherein said user interface has at least one dialog box populated with validation processing and/or validation workflow management information; providing a validation processing engine, said validation processing engine comprising at least one processing rule that operates on validation



processing information selected through said user interface to produce validation protocol information.

120. At a minimum, Defendant has known of the '488 patent at least as early as the filing date of this Complaint.

121. Plaintiff BYTEWEAVR has been damaged as a result of Defendant's infringing conduct described in this Count. Defendant is thus liable to BYTEWEAVR in an amount that adequately compensates BYTEWEAVR for its infringements, which, by law, cannot be less than a reasonable royalty, together with interest and costs as fixed by this Court under 35 U.S.C. § 284.

#### **COUNT IV**

(INFRINGEMENT OF U.S. PATENT NO. 6,965,897)

122. Plaintiff incorporates paragraphs 1 through 121 herein by reference.

123. Plaintiff BYTEWEAVR is the assignee of the '897 patent, entitled "Data Compression Method and Apparatus," with ownership of all substantial rights in the '897 patent, including the right to exclude others and to enforce, sue, and recover damages for past and future infringements.

124. The '897 patent is valid, enforceable, and was duly issued in full compliance with Title 35 of the United States Code. The '897 patent issued from U.S. Patent Application No. 10/065,513. The '897 patent was granted on November 15, 2005 and expired on or about August 10, 2023.

125. Defendant has directly and/or indirectly infringed (by inducing infringement) one or more claims of the '897 patent in this District and elsewhere in Texas and the United States.

126. On information and belief, Defendant designs, develops, manufactures, imports, distributes, offers to sell, sells, and uses the Accused Instrumentalities, including via the activities

of Cloudera and its alter egos, intermediaries, agents, distributors, importers, partners, customers, subsidiaries, affiliates, and/or consumers.

127. Defendant has directly infringed the '897 patent via 35 U.S.C. § 271(a) by making, offering for sale, selling, importing and/or using the Accused Instrumentalities, their components, and/or products containing the same that incorporate the fundamental technologies covered by the '897 patent to, for example, its alter egos, intermediaries, agents, distributors, importers, partners, customers, subsidiaries, affiliates, and/or consumers. Furthermore, on information and belief, Defendant develops and designs the Accused Instrumentalities for U.S. consumers, makes and sells the Accused Instrumentalities outside of the United States, delivers those products and services to related entities, subsidiaries, distribution partners, resellers, vendors, installers, customers and other related service providers in the United States, or in the case that it delivers the Accused Instrumentalities outside of the United States it does so intending and/or knowing that those products are destined for the United States and/or designing those products for sale and use in the United States, thereby directly infringing the '897 patent. *See, e.g., Lake Cherokee Hard Drive Techs., L.L.C. v. Marvell Semiconductor, Inc.*, 964 F. Supp. 2d 653, 658 (E.D. Tex. 2013) (denying summary judgment and allowing presentation to jury as to “whether accused products manufactured and delivered abroad but imported into the United States market by downstream customers ... constitute an infringing sale under § 271(a)”).

128. Furthermore, Defendant Cloudera has directly infringed the '897 patent through its direct involvement in the activities of its subsidiaries, and related entities and other U.S.-based subsidiaries (e.g., Hortonworks, Inc., Cloudera (Government Solutions), Inc., and Eventador), members, segments, companies, and/or brands of Defendant Cloudera, including by designing the Accused Instrumentalities for U.S. consumers and selling and offering for sale the Accused

Instrumentalities directly to its related entities and importing the Accused Instrumentalities into the United States for its related entities. On information and belief, U.S.-based members, segments, companies, and/or brands conduct activities that constitute direct infringement of the '897 patent under 35 U.S.C. § 271(a) by importing, offering for sale, selling, and/or using those Accused Instrumentalities in the U.S. on behalf of and for the benefit of Defendant. Defendant is vicariously liable for the infringing conduct of members, segments, companies, and/or brands of Cloudera (under both the alter ego and agency theories). On information and belief, Defendant Cloudera and other U.S. based subsidiaries, members, segments, companies, and/or brands of Cloudera are essentially the same company. Moreover, Cloudera, as the parent company, has the right and ability to control the infringing activities of those entities such that Defendant receives a direct financial benefit from that infringement.

129. For example, Defendant infringes claim 1 of the '897 patent via the Accused Instrumentalities, namely data management and analytics products and components, software, services, and processes such as the Cloudera Platforms and their components, including the Cloudera Enterprise, the Cloudera Data Platform, Data Hub, Runtime, Search, the Cloudera SDX Management Console, Cloudera Manager, CDH, Cloudera Flow Management, and Cloudera distributions of Apache Oozie, NiFi, YARN, Hue, Avro, Zookeeper and related data storage and compression techniques.

130. Those Accused Instrumentalities include a “method for improving compression of data” comprising the limitations of claim 1. The technology discussion above and the example Accused Instrumentalities provide context for Plaintiff’s allegations that each of those limitations are met. For example, the Accused Instrumentalities include the steps of arranging the data on a mixed format physical layout having a plurality of fixed-sized fields, a plurality of variable-sized

fields and a plurality of offset slots, the fixed-sized fields being of a first size and the offset slots being of a second size; dividing the data on the mixed format physical layout into the fixed-sized fields and the variable sized fields; and compressing the data of the variable sized fields and the fixed-sized fields.

131. At a minimum, Defendant has known of the '897 patent at least as early as the filing date of this Complaint.

132. Plaintiff BYTEWEAVR has been damaged as a result of Defendant's infringing conduct described in this Count. Defendant is thus liable to BYTEWEAVR in an amount that adequately compensates BYTEWEAVR for its infringements, which, by law, cannot be less than a reasonable royalty, together with interest and costs as fixed by this Court under 35 U.S.C. § 284.

### **COUNT V**

(INFRINGEMENT OF U.S. PATENT NO. 6,999,961)

133. Plaintiff incorporates paragraphs 1 through 132 herein by reference.

134. Plaintiff BYTEWEAVR is the assignee of the '961 patent, entitled "Method of aggregating and distributing informal and formal knowledge using software agents," with ownership of all substantial rights in the '961 patent, including the right to exclude others and to enforce, sue, and recover damages for past and future infringements.

135. The '961 patent is valid, enforceable, and was duly issued in full compliance with Title 35 of the United States Code. The '961 patent issued from U.S. Patent Application No. 09/938,971. The '961 patent was granted on February 14, 2006 and expired on or about October 25, 2023.

136. Defendant has directly and/or indirectly infringed (by inducing infringement) one or more claims of the '961 patent in this District and elsewhere in Texas and the United States.

137. On information and belief, Defendant designs, develops, manufactures, imports, distributes, offers to sell, sells, and uses the Accused Instrumentalities, including via the activities of Cloudera and its alter egos, intermediaries, agents, distributors, importers, partners, customers, subsidiaries, affiliates, and/or consumers.

138. Defendant has directly infringed the '961 patent via 35 U.S.C. § 271(a) by making, offering for sale, selling, importing and/or using the Accused Instrumentalities, their components, and/or products containing the same that incorporate the fundamental technologies covered by the '961 patent to, for example, its alter egos, intermediaries, agents, distributors, importers, partners, customers, subsidiaries, affiliates, and/or consumers. Furthermore, on information and belief, Defendant develops and designs the Accused Instrumentalities for U.S. consumers, makes and sells the Accused Instrumentalities outside of the United States, delivers those products and services to related entities, subsidiaries, distribution partners, resellers, vendors, installers, customers and other related service providers in the United States, or in the case that it delivers the Accused Instrumentalities outside of the United States it does so intending and/or knowing that those products are destined for the United States and/or designing those products for sale and use in the United States, thereby directly infringing the '961 patent. *See, e.g., Lake Cherokee Hard Drive Techs., L.L.C. v. Marvell Semiconductor, Inc.*, 964 F. Supp. 2d 653, 658 (E.D. Tex. 2013) (denying summary judgment and allowing presentation to jury as to “whether accused products manufactured and delivered abroad but imported into the United States market by downstream customers ... constitute an infringing sale under § 271(a)”).

139. Furthermore, Defendant Cloudera has directly infringed the '961 patent through its direct involvement in the activities of its subsidiaries, and related entities and other U.S.-based subsidiaries (e.g., Hortonworks, Inc., Cloudera (Government Solutions), Inc., and Eventador),

members, segments, companies, and/or brands of Defendant Cloudera, including by designing the Accused Instrumentalities for U.S. consumers and selling and offering for sale the Accused Instrumentalities directly to its related entities and importing the Accused Instrumentalities into the United States for its related entities. On information and belief, U.S.-based members, segments, companies, and/or brands conduct activities that constitute direct infringement of the '961 patent under 35 U.S.C. § 271(a) by importing, offering for sale, selling, and/or using those Accused Instrumentalities in the U.S. on behalf of and for the benefit of Defendant. Defendant is vicariously liable for the infringing conduct of members, segments, companies, and/or brands of Cloudera (under both the alter ego and agency theories). On information and belief, Defendant Cloudera and other U.S. based subsidiaries, members, segments, companies, and/or brands of Cloudera are essentially the same company. Moreover, Cloudera, as the parent company, has the right and ability to control the infringing activities of those entities such that Defendant receives a direct financial benefit from that infringement.

140. For example, Defendant infringes claim 1 of the '961 patent via the Accused Instrumentalities, namely data management and analytics products and components, software, services, and processes such as the Cloudera Platforms and their components, including the Cloudera Enterprise, the Cloudera Data Platform, Data Hub, Runtime, Search, the Cloudera SDX Management Console, Cloudera Manager, CDH, Cloudera Flow Management, and Cloudera distributions of Apache Oozie, NiFi, YARN, Hue, Avro, Zookeeper and related data storage and compression techniques.

141. Those Accused Instrumentalities include “[a] method of aggregating information content” comprising the limitations of claim 1. The technology discussion above and the example Accused Instrumentalities provide context for Plaintiff’s allegations that each of those limitations

are met. For example, the Accused Instrumentalities include the steps of accessing a content aggregator; transmitting a search query to the content aggregator; transmitting the query from the content aggregator to a plurality of remote agents, wherein each of said agents is located on one of a plurality of distinct networks; searching each of said plurality of networks for content responsive to the query via its respective remote agent; transmitting a search result from each of said respective remote agents to the content aggregator; processing the plurality of search results into a processed information content via the aggregator, wherein said processing includes applying a rules and standard designated by a client, and transmitting said processed information content from said aggregator to said client.

142. At a minimum, Defendant has known of the '961 patent at least as early as the filing date of this Complaint.

143. Plaintiff BYTEWEAVR has been damaged as a result of Defendant's infringing conduct described in this Count. Defendant is thus liable to BYTEWEAVR in an amount that adequately compensates BYTEWEAVR for its infringements, which, by law, cannot be less than a reasonable royalty, together with interest and costs as fixed by this Court under 35 U.S.C. § 284.

## COUNT VI

(INFRINGEMENT OF U.S. PATENT NO. 7,082,474)

144. Plaintiff incorporates paragraphs 1 through 143 herein by reference.

145. Plaintiff BYTEWEAVR is the assignee of the '474 patent, entitled "Data sharing and file distribution method and associated distributed processing system," with ownership of all substantial rights in the '474 patent, including the right to exclude others and to enforce, sue, and recover damages for past and future infringements.

146. The '474 patent is valid, enforceable, and was duly issued in full compliance with Title 35 of the United States Code. The '474 patent issued from U.S. Patent Application No. 09/602,803. The '474 patent was granted on July 25, 2006 and expired on or about December 3, 2022.

147. Defendant has directly and/or indirectly infringed (by inducing infringement) one or more claims of the '474 patent in this District and elsewhere in Texas and the United States.

148. On information and belief, Defendant designs, develops, manufactures, imports, distributes, offers to sell, sells, and uses the Accused Instrumentalities, including via the activities of Cloudera and its alter egos, intermediaries, agents, distributors, importers, partners, customers, subsidiaries, affiliates, and/or consumers.

149. Defendant has directly infringed the '474 patent via 35 U.S.C. § 271(a) by making, offering for sale, selling, importing and/or using the Accused Instrumentalities, their components, and/or products containing the same that incorporate the fundamental technologies covered by the '474 patent to, for example, its alter egos, intermediaries, agents, distributors, importers, partners, customers, subsidiaries, affiliates, and/or consumers. Furthermore, on information and belief, Defendant develops and designs the Accused Instrumentalities for U.S. consumers, makes and sells the Accused Instrumentalities outside of the United States, delivers those products and services to related entities, subsidiaries, distribution partners, resellers, vendors, installers, customers and other related service providers in the United States, or in the case that it delivers the Accused Instrumentalities outside of the United States it does so intending and/or knowing that those products are destined for the United States and/or designing those products for sale and use in the United States, thereby directly infringing the '474 patent. *See, e.g., Lake Cherokee Hard Drive Techs., L.L.C. v. Marvell Semiconductor, Inc.*, 964 F. Supp. 2d 653, 658 (E.D. Tex. 2013) (denying



summary judgment and allowing presentation to jury as to “whether accused products manufactured and delivered abroad but imported into the United States market by downstream customers ... constitute an infringing sale under § 271(a)”.

150. Furthermore, Defendant Cloudera has directly infringed the '474 patent through its direct involvement in the activities of its subsidiaries, and related entities and other U.S.-based subsidiaries (e.g., Hortonworks, Inc., Cloudera (Government Solutions), Inc., and Eventador), members, segments, companies, and/or brands of Defendant Cloudera, including by designing the Accused Instrumentalities for U.S. consumers and selling and offering for sale the Accused Instrumentalities directly to its related entities and importing the Accused Instrumentalities into the United States for its related entities. On information and belief, U.S.-based members, segments, companies, and/or brands conduct activities that constitute direct infringement of the '474 patent under 35 U.S.C. § 271(a) by importing, offering for sale, selling, and/or using those Accused Instrumentalities in the U.S. on behalf of and for the benefit of Defendant. Defendant is vicariously liable for the infringing conduct of members, segments, companies, and/or brands of Cloudera (under both the alter ego and agency theories). On information and belief, Defendant Cloudera and other U.S. based subsidiaries, members, segments, companies, and/or brands of Cloudera are essentially the same company. Moreover, Cloudera, as the parent company, has the right and ability to control the infringing activities of those entities such that Defendant receives a direct financial benefit from that infringement.

151. For example, Defendant infringes claim 1 of the '474 patent via the Accused Instrumentalities, namely data management and analytics products and components, software, services, and processes such as the Cloudera Platforms and their components, including the Cloudera Enterprise, the Cloudera Data Platform, Data Hub, Runtime, Search, the Cloudera SDX

Management Console, Cloudera Manager, CDH, Cloudera Flow Management, and Cloudera distributions of Apache Oozie, NiFi, YARN, Hue, Avro, Zookeeper and related data storage and compression techniques.

152. Those Accused Instrumentalities include a “method operating a distributed processing system having a network coupling a multiplicity of Host distributed devices for processing workloads for the distributed processing system, a plurality of Client systems requesting processing of the workloads, and a Server system for selectively distributing the workloads from the plurality of Client systems for processing by the distributed processing system” comprising the limitations of claim 1. The technology discussion above and the example Accused Instrumentalities provide context for Plaintiff’s allegations that each of those limitations are met. For example, the Accused Instrumentalities include the steps of receiving a request by the Server system from one of the plurality of Client systems to use the distributed processing system to process a first workload; sending the first workload to a first Host distributed device selected from the multiplicity of Host distributed devices; sending to the first Host distributed device an index of one or more data addresses defining a location of first data required to process the first workload; accessing the first data from a first data address selected from the one or more data addresses in the index; and updating the index to include a storage address of storage coupled to the first Host distributed device as a location of the first data.

153. At a minimum, Defendant has known of the ’474 patent at least as early as the filing date of this Complaint.

154. Plaintiff BYTEWEAVR has been damaged as a result of Defendant’s infringing conduct described in this Count. Defendant is thus liable to BYTEWEAVR in an amount that

adequately compensates BYTEWEAVR for its infringements, which, by law, cannot be less than a reasonable royalty, together with interest and costs as fixed by this Court under 35 U.S.C. § 284.

**COUNT VII**

(INFRINGEMENT OF U.S. PATENT NO. 8,275,827)

155. Plaintiff incorporates paragraphs 1 through 154 herein by reference.

156. Plaintiff BYTEWEAVR is the assignee of the '827 patent, entitled "Software-based network attached storage services hosted on massively distributed parallel computing networks," with ownership of all substantial rights in the '827 patent, including the right to exclude others and to enforce, sue, and recover damages for past and future infringements.

157. The '827 patent is valid, enforceable, and was duly issued in full compliance with Title 35 of the United States Code. The '827 patent issued from U.S. Patent Application No. 09/834,785.

158. Defendant has directly and/or indirectly infringed (by inducing infringement) one or more claims of the '827 patent in this District and elsewhere in Texas and the United States.

159. On information and belief, Defendant designs, develops, manufactures, imports, distributes, offers to sell, sells, and uses the Accused Instrumentalities, including via the activities of Cloudera and its alter egos, intermediaries, agents, distributors, importers, partners, customers, subsidiaries, affiliates, and/or consumers.

160. Defendant has directly infringed the '827 patent via 35 U.S.C. § 271(a) by making, offering for sale, selling, importing and/or using the Accused Instrumentalities, their components, and/or products containing the same that incorporate the fundamental technologies covered by the '827 patent to, for example, its alter egos, intermediaries, agents, distributors, importers, partners, customers, subsidiaries, affiliates, and/or consumers. Furthermore, on information and belief,

Defendant develops and designs the Accused Instrumentalities for U.S. consumers, makes and sells the Accused Instrumentalities outside of the United States, delivers those products and services to related entities, subsidiaries, distribution partners, resellers, vendors, installers, customers and other related service providers in the United States, or in the case that it delivers the Accused Instrumentalities outside of the United States it does so intending and/or knowing that those products are destined for the United States and/or designing those products for sale and use in the United States, thereby directly infringing the '827 patent. *See, e.g., Lake Cherokee Hard Drive Techs., L.L.C. v. Marvell Semiconductor, Inc.*, 964 F. Supp. 2d 653, 658 (E.D. Tex. 2013) (denying summary judgment and allowing presentation to jury as to “whether accused products manufactured and delivered abroad but imported into the United States market by downstream customers ... constitute an infringing sale under § 271(a)”).

161. Furthermore, Defendant Cloudera has directly infringed the '827 patent through its direct involvement in the activities of its subsidiaries, and related entities and other U.S.-based subsidiaries (e.g., Hortonworks, Inc., Cloudera (Government Solutions), Inc., and Eventador), members, segments, companies, and/or brands of Defendant Cloudera, including by designing the Accused Instrumentalities for U.S. consumers and selling and offering for sale the Accused Instrumentalities directly to its related entities and importing the Accused Instrumentalities into the United States for its related entities. On information and belief, U.S.-based members, segments, companies, and/or brands conduct activities that constitute direct infringement of the '827 patent under 35 U.S.C. § 271(a) by importing, offering for sale, selling, and/or using those Accused Instrumentalities in the U.S. on behalf of and for the benefit of Defendant. Defendant is vicariously liable for the infringing conduct of members, segments, companies, and/or brands of Cloudera (under both the alter ego and agency theories). On information and belief, Defendant Cloudera and

other U.S. based subsidiaries, members, segments, companies, and/or brands of Cloudera are essentially the same company. Moreover, Cloudera, as the parent company, has the right and ability to control the infringing activities of those entities such that Defendant receives a direct financial benefit from that infringement.

162. For example, Defendant infringes at least claims 2 and 14 of the '827 patent via the Accused Instrumentalities, namely data management and analytics products and components, software, services, and processes such as the Cloudera Platforms and their components, including the Cloudera Enterprise, the Cloudera Data Platform, Data Hub, Runtime, Search, the Cloudera SDX Management Console, Cloudera Manager, CDH, Cloudera Flow Management, and Cloudera distributions of Apache Oozie, NiFi, YARN, Hue, Avro, Zookeeper and related data storage and compression techniques.

163. Those Accused Instrumentalities include “[a] computer-implemented method” comprising the limitations of claim 1. The technology discussion above and the example Accused Instrumentalities provide context for Plaintiff’s allegations that each of those limitations are met. For example, the Accused Instrumentalities include the steps of configuring a distributed processing system of a plurality of distributed devices coupled to a network, wherein the plurality of distributed devices include respective client agents configured to process respective portions of a workload for the distributed processing system, wherein the respective client agents for particular distributed devices of the plurality of distributed devices have corresponding software-based network attached storage (NAS) components configured to assess unused or under-utilized storage resources in selected distributed devices of the plurality of distributed devices; representing with the corresponding software-based NAS component that the selected distributed devices respectively comprise NAS devices having an available amount of storage resources related to the unused and

under-utilized storage resources for the selected distributed devices; processing one or more of data storage or access workloads for the distributed processing system by accessing data from or storing data to at least a portion of the available amount of storage resources to provide NAS service to a client device coupled to the network; enabling at least one of the selected distributed devices to function as a location distributed device to store location information associated with data stored by the selected distributed devices through use of the respective client agents for the particular distributed device; and enabling at least one of the selected distributed devices to function as a stand-alone dedicated NAS device through use of the respective client agents for the particular distributed device.

164. At a minimum, Defendant has known of the '827 patent at least as early as the filing date of this Complaint.

165. On information and belief, since at least the above-mentioned date when Defendant was on notice of its infringement, Defendant has actively induced, under 35 U.S.C. § 271(b), importers, distribution partners, vendors, reseller partners, dealers, customers, installers, consumers, users and other related service providers that import, distribute, purchase, offer for sale, sell, or use the Accused Instrumentalities that include or are made using all of the limitations of one or more claims of the '827 patent to directly infringe one or more claims of the '827 patent by using, offering for sale, selling, and/or importing the Accused Instrumentalities. Since at least the date of notice provided above, Defendant conducts infringing activities with knowledge, or with willful blindness of the fact, that the induced acts constitute infringement of the '827 patent. On information and belief, Defendant intends to cause, and has taken affirmative steps to induce, infringement by importers, distribution partners, reseller partners, vendors, dealers, customers, installers, consumers, users, and other related service providers by at least, *inter alia*, the following: 1) sales

and marketing activities that promote the infringing use of the Accused Instrumentalities, 2) utilizing partners to create and/or maintain established distribution channels for the Accused Instrumentalities into and within the United States, 3) designing, developing, manufacturing the Accused Instrumentalities in conformity with U.S. laws, regulations, and market standards, 4) distributing or making available training, certifications, demos, webinars, events, resource libraries, documentation, instructions and/or manuals for the Accused Instrumentalities to purchasers and prospective buyers, 5) testing and certifying the features in the Accused Instrumentalities, and/or 6) providing technical support, upgrades and migrations, professional or tutorial services for the Accused Instrumentalities to purchasers in the United States. *See, e.g., Services & Support: Get the help you need*, CLOUDERA, <https://www.cloudera.com/about/services-and-support.html> (providing links where consumers may access “Support,” “Training,” “Professional services,” “Machine Learning Services,” a “Support Portal” and a “Community” for using Cloudera’s data management and analytics products and components, software, services, and processes) (last visited Oct. 11, 2023). Such support and services provide convenience, added functionality and value that induces partners and consumers to license, use, and incorporate the Defendant’s data management and analytics products and components, software, services, and processes into their own network systems and businesses. *See, e.g., Solutions Gallery*, CLOUDERA, <https://www.cloudera.com/solutions/gallery.html> (providing use cases for Cloudera’s products and services as examples of “Customer Analytics,” “IoT/ Connected Products,” “Security, Risk, & Compliance,” and “Modernize Architecture”) (last visited Oct. 11, 2023). Thus, these activities further infringe or induce infringement of the ’827 patent.

166. On information and belief, despite having knowledge of the ’827 patent and knowledge that it is directly and/or indirectly infringing one or more claims of the ’827 patent,

Defendant has nevertheless continued its infringing conduct and disregarded an objectively high likelihood of infringement. Each of Defendant's infringing activities relative to the '827 patent have been, and continue to be, willful, wanton, malicious, in bad-faith, deliberate, consciously wrongful, flagrant, characteristic of a pirate, and an egregious case of misconduct beyond typical infringement such that Plaintiff is entitled under 35 U.S.C. § 284 to enhanced damages up to three times the amount found or assessed.

167. Plaintiff BYTEWEAVR has been damaged as a result of Defendant's infringing conduct described in this Count. Defendant is thus liable to BYTEWEAVR in an amount that adequately compensates BYTEWEAVR for its infringements, which, by law, cannot be less than a reasonable royalty, together with interest and costs as fixed by this Court under 35 U.S.C. § 284.

### **COUNT VIII**

(INFRINGEMENT OF U.S. REISSUED PATENT NO. RE42153)

168. Plaintiff incorporates paragraphs 1 through 167 herein by reference.

169. Plaintiff BYTEWEAVR is the assignee of the '153 patent, entitled "Dynamic coordination and control of network connected devices for large-scale network site testing and associated architectures," with ownership of all substantial rights in the '153 patent, including the right to exclude others and to enforce, sue, and recover damages for past and future infringements.

170. The '153 patent is valid, enforceable, and was duly issued in full compliance with Title 35 of the United States Code. The '153 patent issued from U.S. Patent Application No. 10/190,368. The '153 patent was granted on March 1, 2005 and expired on or about March 26, 2022.

171. Defendant has directly and/or indirectly infringed (by inducing infringement) one or more claims of the '153 patent in this District and elsewhere in Texas and the United States.



172. On information and belief, Defendant designs, develops, manufactures, imports, distributes, offers to sell, sells, and uses the Accused Instrumentalities, including via the activities of Cloudera and its alter egos, intermediaries, agents, distributors, importers, partners, customers, subsidiaries, affiliates, and/or consumers.

173. Defendant has directly infringed the '153 patent via 35 U.S.C. § 271(a) by making, offering for sale, selling, importing and/or using the Accused Instrumentalities, their components, and/or products containing the same that incorporate the fundamental technologies covered by the '153 patent to, for example, its alter egos, intermediaries, agents, distributors, importers, partners, customers, subsidiaries, affiliates, and/or consumers. Furthermore, on information and belief, Defendant develops and designs the Accused Instrumentalities for U.S. consumers, makes and sells the Accused Instrumentalities outside of the United States, delivers those products and services to related entities, subsidiaries, distribution partners, resellers, vendors, installers, customers and other related service providers in the United States, or in the case that it delivers the Accused Instrumentalities outside of the United States it does so intending and/or knowing that those products are destined for the United States and/or designing those products for sale and use in the United States, thereby directly infringing the '153 patent. *See, e.g., Lake Cherokee Hard Drive Techs., L.L.C. v. Marvell Semiconductor, Inc.*, 964 F. Supp. 2d 653, 658 (E.D. Tex. 2013) (denying summary judgment and allowing presentation to jury as to “whether accused products manufactured and delivered abroad but imported into the United States market by downstream customers ... constitute an infringing sale under § 271(a)”).

174. Furthermore, Defendant Cloudera has directly infringed the '153 patent through its direct involvement in the activities of its subsidiaries, and related entities and other U.S.-based subsidiaries (e.g., Hortonworks, Inc., Cloudera (Government Solutions), Inc., and Eventador),

members, segments, companies, and/or brands of Defendant Cloudera, including by designing the Accused Instrumentalities for U.S. consumers and selling and offering for sale the Accused Instrumentalities directly to its related entities and importing the Accused Instrumentalities into the United States for its related entities. On information and belief, U.S.-based members, segments, companies, and/or brands conduct activities that constitute direct infringement of the '153 patent under 35 U.S.C. § 271(a) by importing, offering for sale, selling, and/or using those Accused Instrumentalities in the U.S. on behalf of and for the benefit of Defendant. Defendant is vicariously liable for the infringing conduct of members, segments, companies, and/or brands of Cloudera (under both the alter ego and agency theories). On information and belief, Defendant Cloudera and other U.S. based subsidiaries, members, segments, companies, and/or brands of Cloudera are essentially the same company. Moreover, Cloudera, as the parent company, has the right and ability to control the infringing activities of those entities such that Defendant receives a direct financial benefit from that infringement.

175. For example, Defendant infringes claim 1 of the '153 patent via the Accused Instrumentalities, namely data management and analytics products and components, software, services, and processes such as the Cloudera Platforms and their components, including the Cloudera Enterprise, the Cloudera Data Platform, Data Hub, Runtime, Search, the Cloudera SDX Management Console, Cloudera Manager, CDH, Cloudera Flow Management, and Cloudera distributions of Apache Oozie, NiFi, YARN, Hue, Avro, Zookeeper and related data storage and compression techniques.

176. Those Accused Instrumentalities include “[a] method of providing dynamic coordination of distributed client systems in a distributed computing platform” comprising the limitations of claim 1. The technology discussion above and the example Accused Instrumentalities

provide context for Plaintiff's allegations that each of those limitations are met. For example, the Accused Instrumentalities include the steps of providing at least one server system coupled to a network; providing a plurality of network-connected distributed client systems, the client systems having under-utilized capabilities and running a client agent program to provide workload processing for at least one project of a distributed computing platform; utilizing the server system to distribute workloads for the at least one project to the client systems and to distribute initial project and poll parameters to the client systems; receiving poll communications from the client systems during processing of project workloads by the client systems, wherein a dynamic snapshot information of current project status is provided based at least in part upon the poll communications; analyzing the poll communications to determine whether or not to make one or more modification to the initial project and poll parameters, wherein the modifications to the initial project and poll parameters utilize the dynamic snapshot information to determine whether to change how many client systems are active in the at least one project, and if a fewer number is desired, including within a polling response communications a reduction in the number of actively participating clients, and if a greater number is desired, adding client systems to active participation in the at least one project; sending the poll response communications to the client systems to modify the initial project and poll parameters depending upon one or more decisions reached in the analyzing step; and repeating the receiving, analyzing and sending steps to dynamically coordinate project activities of the plurality of client systems during project operations.

177. At a minimum, Defendant has known of the '153 patent at least as early as the filing date of this Complaint.

178. Plaintiff BYTEWEAVR has been damaged as a result of Defendant's infringing conduct described in this Count. Defendant is thus liable to BYTEWEAVR in an amount that

adequately compensates BYTEWEAVR for its infringements, which, by law, cannot be less than a reasonable royalty, together with interest and costs as fixed by this Court under 35 U.S.C. § 284.

### **CONCLUSION**

179. Plaintiff is entitled to recover from Defendant the damages sustained by Plaintiff as a result of Defendant's wrongful acts in an amount subject to proof at trial, which, by law, cannot be less than a reasonable royalty, together with interest and costs as fixed by this Court.

180. Plaintiff has incurred and will incur attorneys' fees, costs, and expenses in the prosecution of this action. The circumstances of this dispute may give rise to an exceptional case within the meaning of 35 U.S.C. § 285, and Plaintiff is entitled to recover its reasonable and necessary attorneys' fees, costs, and expenses.

### **JURY DEMAND**

181. Plaintiff hereby requests a trial by jury pursuant to Rule 38 of the Federal Rules of Civil Procedure.

### **PRAYER FOR RELIEF**

182. Plaintiff requests that the Court find in its favor and against Defendant, and that the Court grant Plaintiff the following relief:

- A. A judgment that Defendant have infringed the Asserted Patents as alleged herein, directly and/or indirectly by way of inducing infringement of such patents;
- B. A judgment for an accounting of damages sustained by Plaintiff as a result of the acts of infringement by Defendant;
- C. A judgment and order requiring Defendant to pay Plaintiff damages under 35 U.S.C. § 284, including up to treble damages as provided by 35 U.S.C. § 284, and any royalties determined to be appropriate;

- D. A judgment and order requiring Defendant to pay Plaintiff pre-judgment and post-judgment interest on the damages awarded;
- E. A judgment and order finding this to be an exceptional case and requiring Defendant to pay the costs of this action (including all disbursements) and attorneys' fees as provided by 35 U.S.C. § 285; and
- F. Such other and further relief as the Court deems just and equitable.

Dated: March 8, 2024

Respectfully submitted,

*/s/ Jeffrey R. Bragalone*

Jeffrey R. Bragalone (lead attorney)

Texas Bar No. 02855775

E-mail: jbragalone@bosfirm.com

Terry A. Saad

Texas Bar No. 24066015

E-mail: tsaad@bosfirm.com

Marcus Benavides

Texas Bar No. 24035574

E-mail: mbenavides@bosfirm.com

Brandon V. Zuniga

Texas Bar no. 24088720

E-mail: bzuniga@bosfirm.com

Mark Douglass

Texas Bar No. 24131184

Email: mdouglass@bosfirm.com

BRAGALONE OLEJKO SAAD PC

901 Main Street

Suite 3800

Dallas, Texas 75202

Telephone: (214) 785-6670

Facsimile: (214) 785-6680

**ATTORNEYS FOR PLAINTIFF  
BYTEWEAVR, LLC**

**CERTIFICATE OF SERVICE**

I hereby certify that on March 8, 2024, a copy of the foregoing document was filed electronically via the Court's CM/ECF system and therefore this document was served on all counsel who are deemed to have consented to electronic service.

*/s/ Terry A. Saad*  
\_\_\_\_\_  
TERRY A. SAAD