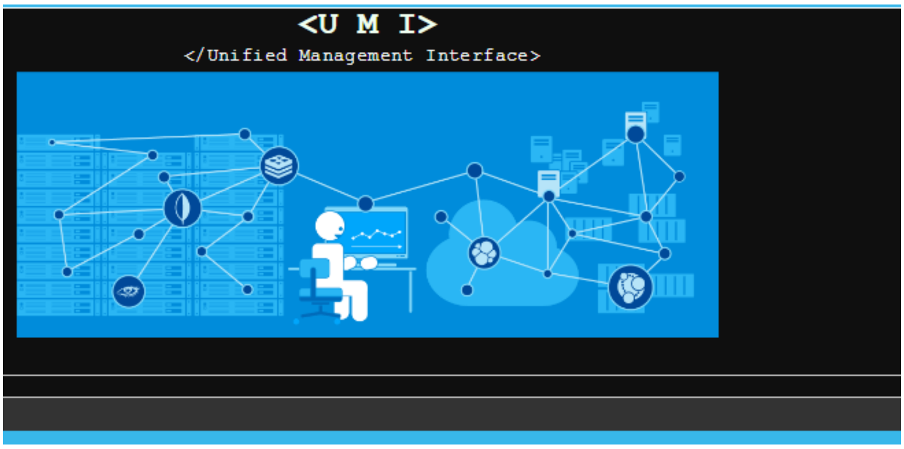
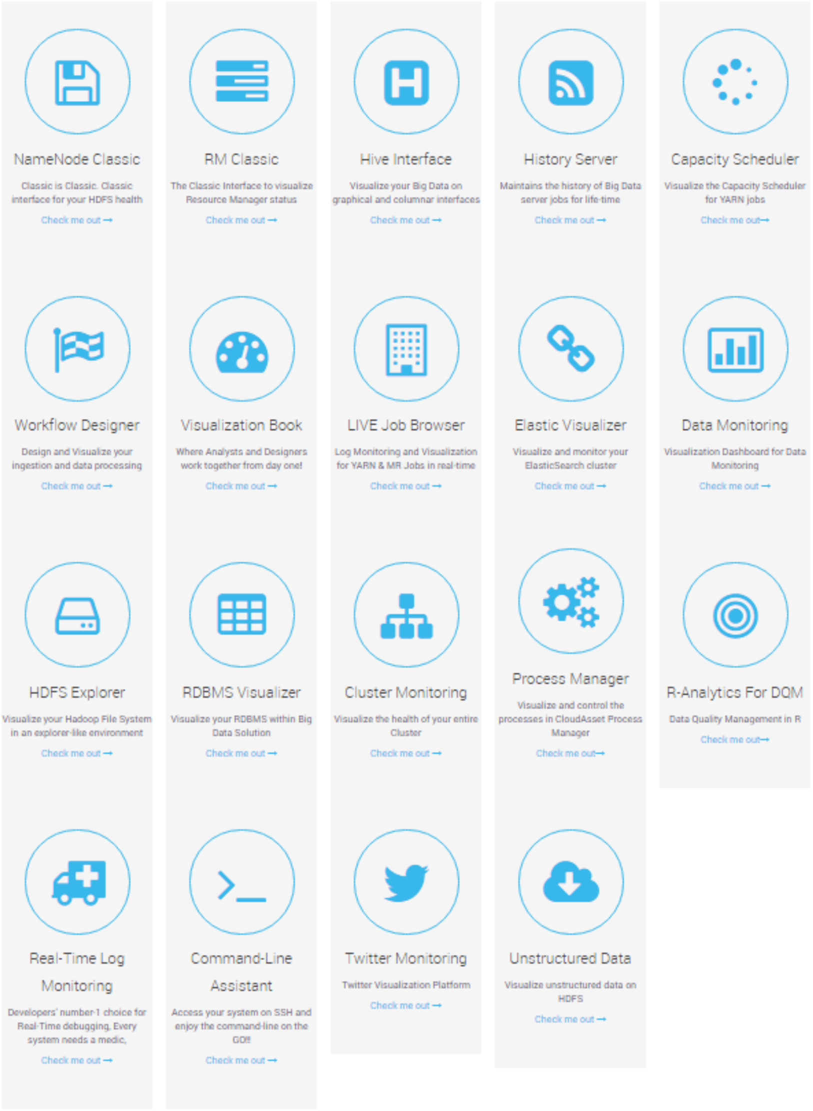
HYDRA BIG DATA PLATFORM

PRODUCT MANUAL





NameNode Classic
Classic is Classic. Classic interface for your HDFS health
Check me out →

RM Classic
The Classic Interface to visualize Resource Manager status
Check me out →

Hive Interface
Visualize your Big Data on graphical and columnar interfaces
Check me out →

History Server
Maintains the history of Big Data server jobs for life-time
Check me out →

Capacity Scheduler
Visualize the Capacity Scheduler for YARN jobs
Check me out→

Workflow Designer
Design and Visualize your ingestion and data processing
Check me out →

Visualization Book
Where Analysts and Designers work together from day one!
Check me out →

LIVE Job Browser
Log Monitoring and Visualization for YARN & MR Jobs in real-time
Check me out →

Elastic Visualizer
Visualize and monitor your ElasticSearch cluster
Check me out →

Data Monitoring
Visualization Dashboard for Data Monitoring
Check me out →

HDFS Explorer
Visualize your Hadoop File System in an explorer-like environment
Check me out →

RDBMS Visualizer
Visualize your RDBMS within Big Data Solution
Check me out →

Cluster Monitoring
Visualize the health of your entire Cluster
Check me out →

Process Manager
Visualize and control the processes in CloudAsset Process Manager
Check me out→

R-Analytics For DQM
Data Quality Management in R
Check me out→

Real-Time Log Monitoring
Developers' number-1 choice for Real-Time debugging. Every system needs a medic,
Check me out →

Command-Line Assistant
Access your system on SSH and enjoy the command-line on the GO!!
Check me out →

Twitter Monitoring
Twitter Visualization Platform
Check me out →

Unstructured Data
Visualize unstructured data on HDFS
Check me out →

ABOUT DOCUMENT

Hydra Big Data Platform Product Manual

Version	Date	Released By
Ver 0.5	14th January 2017	Ahsan Aftab
Ver 0.6	20th January 2017	Ahsan Aftab
Ver 1.0	24th January 2017	Ahsan Aftab

Authors

Name	Role
Ahsan Aftab	Lead Architect
Email	ahsan.aftab@alumni.ie.edu
Contact	+34-652890637

Document Status

Final Draft - ready for official review

OVERVIEW

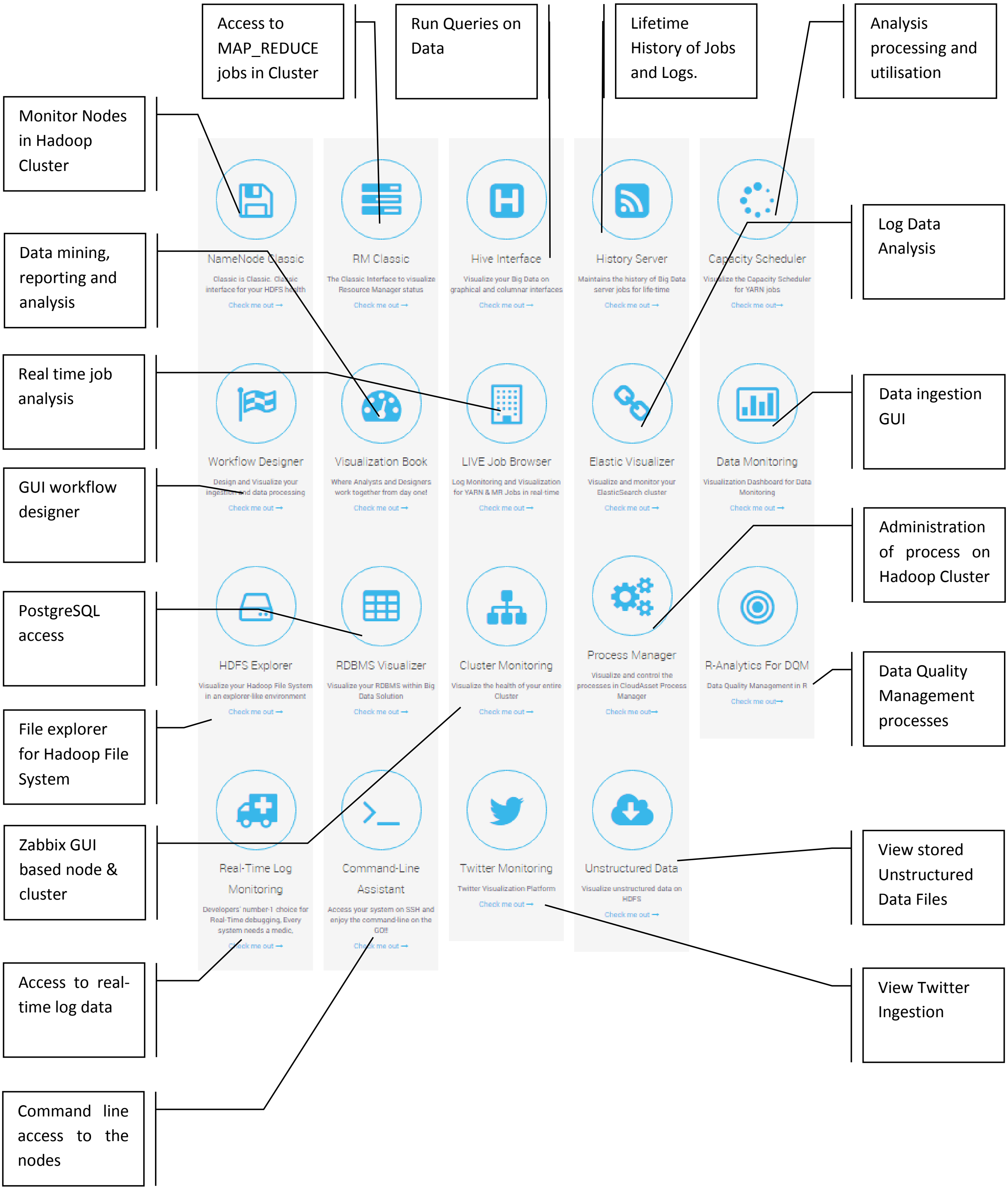
The Big Data Platform developed by me provides unprecedented capabilities to manage large and complex data sets for which traditional data processing methods and applications are inadequate. The technologies implemented will have a vast improvement on all aspects of data management – including data capture, curation, processing, analysis, search, sharing, transfer, querying and visualization.

This framework is further strengthened by addition of Big Data Framework Management components for HAII to monitor, manage and optimize data and hardware clusters, for both current requirements and future expansion.

FEATURES & BENEFITS

- Effectively manage large volumes of complex data
- Capture & utilize semi-structured and unstructured data
- Easily scalable as the needs for the system grow over time
- Significantly improve query response and quality
- Run data analytics
- Ensure improved data security and privacy
- Deliver enhanced data services for richer and more information applications
- Future ready and allow for experimentation with evolving technologies
- Provide complex data analysis
- Hardware - Configuration management
- Hardware Monitoring and alerting
- BI / DSS Ready

UMI SCHEMATIC



Access to MAP_REDUCE jobs in Cluster

Run Queries on Data

Lifetime History of Jobs and Logs.

Analysis processing and utilisation

Monitor Nodes in Hadoop Cluster

Data mining, reporting and analysis

Real time job analysis

GUI workflow designer

PostgreSQL access

File explorer for Hadoop File System

Zabbix GUI based node & cluster

Access to real-time log data

Command line access to the nodes

Log Data Analysis

Data ingestion GUI

Administration of process on Hadoop Cluster

Data Quality Management processes

View stored Unstructured Data Files

View Twitter Ingestion

NameNode Classic

Classic is Classic. Classic interface for your HDFS health

Check me out →

RM Classic

The Classic Interface to visualize Resource Manager status

Check me out →

Hive Interface

Visualize your Big Data on graphical and columnar interfaces

Check me out →

History Server

Maintains the history of Big Data server jobs for life-time

Check me out →

Capacity Scheduler

Visualize the Capacity Scheduler for YARN jobs

Check me out→

Workflow Designer

Design and Visualize your ingestion and data processing

Check me out →

Visualization Book

Where Analysts and Designers work together from day one!

Check me out →

LIVE Job Browser

Log Monitoring and Visualization for YARN & MR Jobs in real-time

Check me out →

Elastic Visualizer

Visualize and monitor your ElasticSearch cluster

Check me out →

Data Monitoring

Visualization Dashboard for Data Monitoring

Check me out →

HDFS Explorer

Visualize your Hadoop File System in an explorer-like environment

Check me out →

RDBMS Visualizer

Visualize your RDBMS within Big Data Solution

Check me out →

Cluster Monitoring

Visualize the health of your entire Cluster

Check me out →

Process Manager

Visualize and control the processes in CloudAsset Process Manager

Check me out→

R-Analytics For DQM

Data Quality Management in R

Check me out→

Real-Time Log Monitoring

Developers' number-1 choice for Real-Time debugging. Every system needs a medic.

Check me out →

Command-Line Assistant

Access your system on SSH and enjoy the command-line on the GO!!

Check me out →

Twitter Monitoring

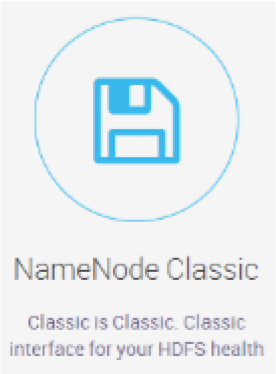
Twitter Visualization Platform

Check me out →

Unstructured Data

Visualize unstructured data on HDFS

Check me out →

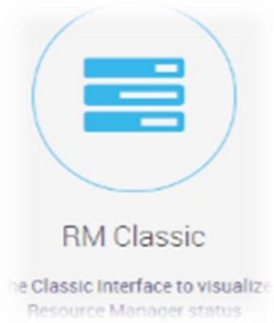


NameNode Classic

Classic is Classic. Classic interface for your HDFS health

The Namenode classic UI or the Namenode web interface is used to monitor the status of the Namenode or master node in a Hadoop Cluster. It’s a very useful tool to monitor and observe the basic health stats of the cluster.

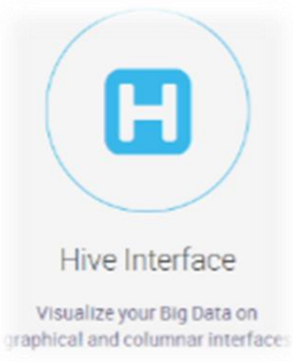
Overview section provides Block Pool Id, Cluster ID parameter, Configured Capacity, DFS used, NON-DFS used, DFS remaining, Live and Dead Nodes status, Safe-mode Indicator and Namenode Journal. Namenode Classic also features HDFS browser and Logs browser.



RM Classic

e Classic Interface to visualiz Resource Manager status

RM or Resource Manager Classic Interface provides real time access to running MAP-REDUCE jobs on Hadoop Cluster. The interface is a very helpful tool to visualize the on-going jobs, their status, cluster resource usage in terms of CPU and RAM, live and dead nodes status, and classification of running jobs as Successful, Suspended, Running or Killed. The Resource Manager can be a primary interface to explore processing related issues or performance tests. The jobs history is retained in Resource Manager for limited periods or unless the cluster is restarted. The retired jobs are archived and moved to History Server.

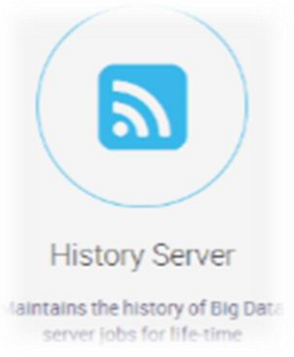


Hive Interface

Visualize your Big Data on raphical and columnar interface

The Hive interface, powered by Hue, is a primary tool to run user queries onto Hive data. The interface is very powerful in that the user queries can be designed, saved and tagged in this interface. Once the designed queries are finalized they can be added to notebooks for enhanced dashboards.

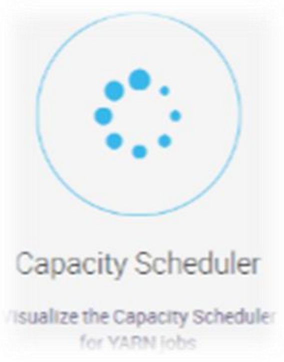
Hive interface also provides analytics for teams to analyse data on graphical interfaces. Thereby ensuring that the developers and analysts can start working on the available data from the first day. Hive interface also indicates queries status, runtime and retains the results for a few hours in temporary storage on HDFS.



aintains the history of Big Dat
server jobs for life-time

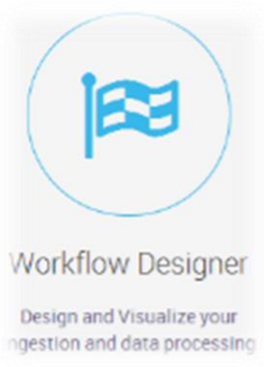
History server is the archive repository for Resource manager. Whenever a job is finished (successfully or otherwise), it is considered as retired and pushed to the History Server. The history server maintains the logs and jobs history for the lifetime of the cluster on HDFS. Thereby ensuring that historic checks and past job analysis may be conducted on any cluster whenever required. The history server also provides advanced counters and logs for analysing in depth the status of every job that ran on the cluster.

Can be a very useful tool for administrators, system analysts and Hadoop designers. The teams also use it during various performance tests on the cluster.



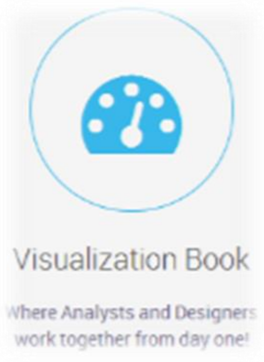
isualize the Capacity Schedule
for YARN jobs

Capacity Scheduler, the built in component within Resource Manager is a perfect tool to analyse the job queues with processing load and utilization % shown in real time for all running jobs on cluster. Capacity scheduler can be used to visualize the remaining processing capacity within the system at peak hours, investigate the priority queues, and analyse the running processes in depth. The display can be used to see the status of Available and Remaining Cores within the cluster, RAM/CPU utilization per node, per job and per queue. This tool is the perfect choice while performing upgrade tasks on the cluster.



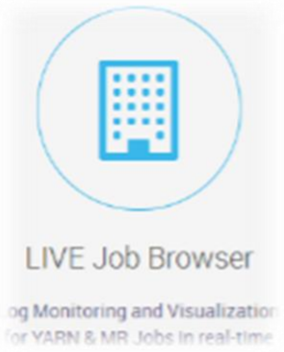
Design and Visualize your
ngestion and data processing

Workflow designer, powered by Hue and built over the top of Oozie is the most powerful and the most versatile tool in the stack. It allows users to create almost any workflow in a graphical user-interface, keep track of the workflow status, analyse logs in real time and schedule the workflows in an advanced job coordinator. Assisted with bundled feature of Oozie, it allows to define the groups or batches of workflows, their priority and their organization. We believe any solution can be integrated with Workflow designer thus allowing its users to visualize the ingestion, administration and mediation functions in a modern GUI.

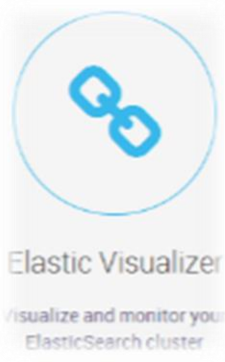


The Visualization book, powered by Hue, is a pre-defined visualization dashboard editor. It basically allows users to create and define their own dashboards in no time.

The Visualization Book or the Notebook can be integrated with some of the very advanced visualization tools in the market and at the same time can act as a stand-alone product. The ease of integration of queries into dashboards makes it an ideal tool for real time analysis, report generation and data mining in absolutely no time!

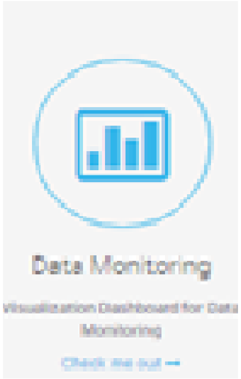

LIVE Job Browser
Log Monitoring and Visualization
for YARN & MR Jobs in real-time

Live Job Browser, the tool provides an advanced interface for Resource Manager Visualization. Based on the concept of classic RM UI, the Job Browser can be used to analyse the jobs status in real time, as well as analyse the logs and counters of each job in a very advanced interface. Search filters can be applied on various criteria, thereby allowing administrators and system designers to have a complete insight of their system running on Hadoop cluster. Live Job Browser is a detachable component within Hue, and can be integrated very easily with multiple clusters.



A web administration tool for ElasticSearch.

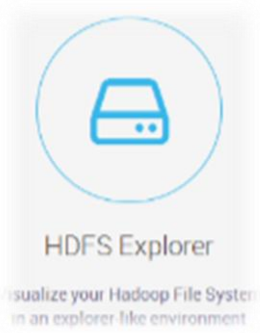
Offers a minimalistic GUI for easy performing of common management tasks of the Elasticsearch cluster. Enables features like cluster storage management, and querying for specific data retrieved by Logstash. This data includes data from Hadoop logs and Tweets from Twitter-API in JSON format.



A visualizer / dashboard tool for the Elastic stack

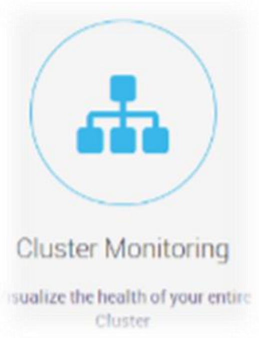
Kibana offers graph and dashboard creation for visualizing the data parsed from logs created by individual components of the platform. One example is to find ingestion problems quickly without the need to go through logs manually.

RDBMS visualizer, is a tool that has been integrated into Hue and can be used to analyse and access traditional databases like PostgreSQL, MySQL and oracle. The RDBMS visualizer enables us to visualize the ancestor databases from our Big Data Platform. It has also been used to deploy metastores for various subsystems of Hadoop. Metastores for subsystems are a key recommendation of Hadoop eco system. The metastores ensure that records are locked and shared amongst the resources in the most effective and efficient manner. This section can also be used to analyse source databases of HAII on PostgreSQL servers.


HDFS Explorer
isualize your Hadoop File Syste
in an explorer-like environment

HDFS explorer, is an advanced HDFS browser that allows its users to explore the HDFS and its directories like any of the most advanced file explorers. It can be used to upload, download or modify data on the HDFS. It can also be used to control the permissions on various directories in the file structure of HDFS.

HDFS file explorer is the number one choice of the developers who are analysing in depth the data that is residing onto their Hadoop cluster's HDFS. HDFS cannot be accessed by users without using Hadoop's URI, and therefore this tool eliminates the need of a file explorer for Hadoop.

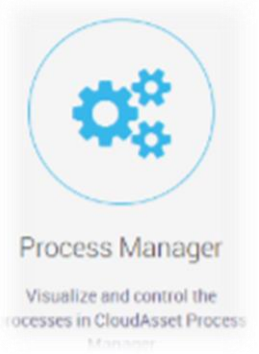


Cluster Monitoring

sualize the health of your entire
Cluster

Zabbix, a solution for monitoring the health of the cluster nodes

Zabbix provides graphs and alerts for monitoring the status and use of physical resources on each node. Different trigger values can be set, for example to give e-mail alerts to the administrators from disk space running low, CPU usage too high or any similar event.



Process Manager

Visualize and control the
ocesses in CloudAsset Process
Manager

CloudAsset Process Manager, allows its users to run and stop processes on entire Hadoop cluster from the same window. It’s a very advanced interface, built for cluster environments, thereby to give control to its users over every process within the cluster. The Process Flow Manager also supports advanced debugging messages, alerts, and Hadoop cluster control in almost no time.

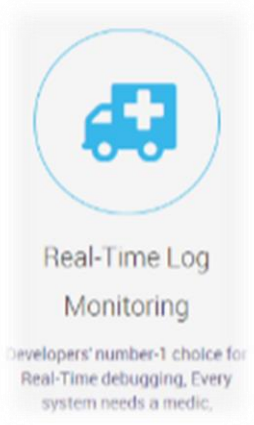
It offers a great interface to perform a general health check on the process of Hadoop cluster. Every service within our ecosystem will be controlled by this Process Flow Manager.

R-Analytics For DQM

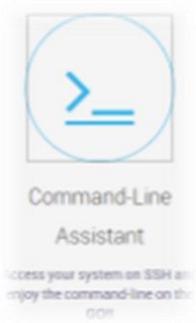
Data Quality Management in R

Data Quality Management is a tool for monitoring and analysing data from our various tables to detect sensor errors like Flat Values, Missing Gaps, Out of range values, Outlier, Inhomogeneity and Missing Patterns.

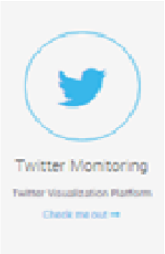
Using some of the cutting-edge technologies that are R, Spark and hdfs for analytics, and Shiny for visualization. Its user friendly graphical user interface helps intuitively observing the error values and help them to filter data and dig down into the issue and work out for its solution.



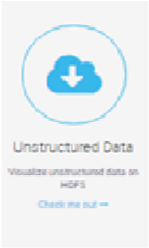
Another lightweight cloud application that offers real-time monitoring functionalities for Hadoop developers. Also allows filtering and searching from the logs in real-time. Can track logs from multiple nodes (Master and slaves alike) and updates to logs are shown as it happens. Also can help visualize logs from multiple services on the cloud simultaneously. This is a number one choice for analyzing multiple services or systems on cluster during debugging activities. Can also be used very effectively in the performance tuning of Hadoop cluster.



Command-Line Assistant, a web command-line SSH tool for accessing the nodes. Useful for doing any kind of maintenance that requires the command-line, without the need to install a separate ssh tool. The developers can now access their hadoop cluster from laptops, mobiles, workstations or any platform supported with a web browser. Thereby allowing administrators and system devops to work on the system from any location without hassle. The interface can be used to interact with the core of shell, hadoop and any service on the cluster.



Twitter Monitoring is constantly reading online tweets for any keywords, previously suggested by designers of process, and then it forwards those tweets to our big data platform. Those tweets are further segmented and filtered and made available as key statistical measure and key indicators on which analysis can be performed using our platform's capabilities. The solutions is highly flexible and can be easily integrated with the workflow



The unstructured data interface allows to access unstructured data directories on HDFS on a visual interface. Files can be copied, viewed, modified and analyzed in a general file explorer like application or structure. The data is synced with HAII's original unstructured data node, and it keeps populating data from that node in almost real time. The unstructured data can be further analyzed in the future solutions.