

Introduction to Hortonworks Sandbox

What is Hortonworks?

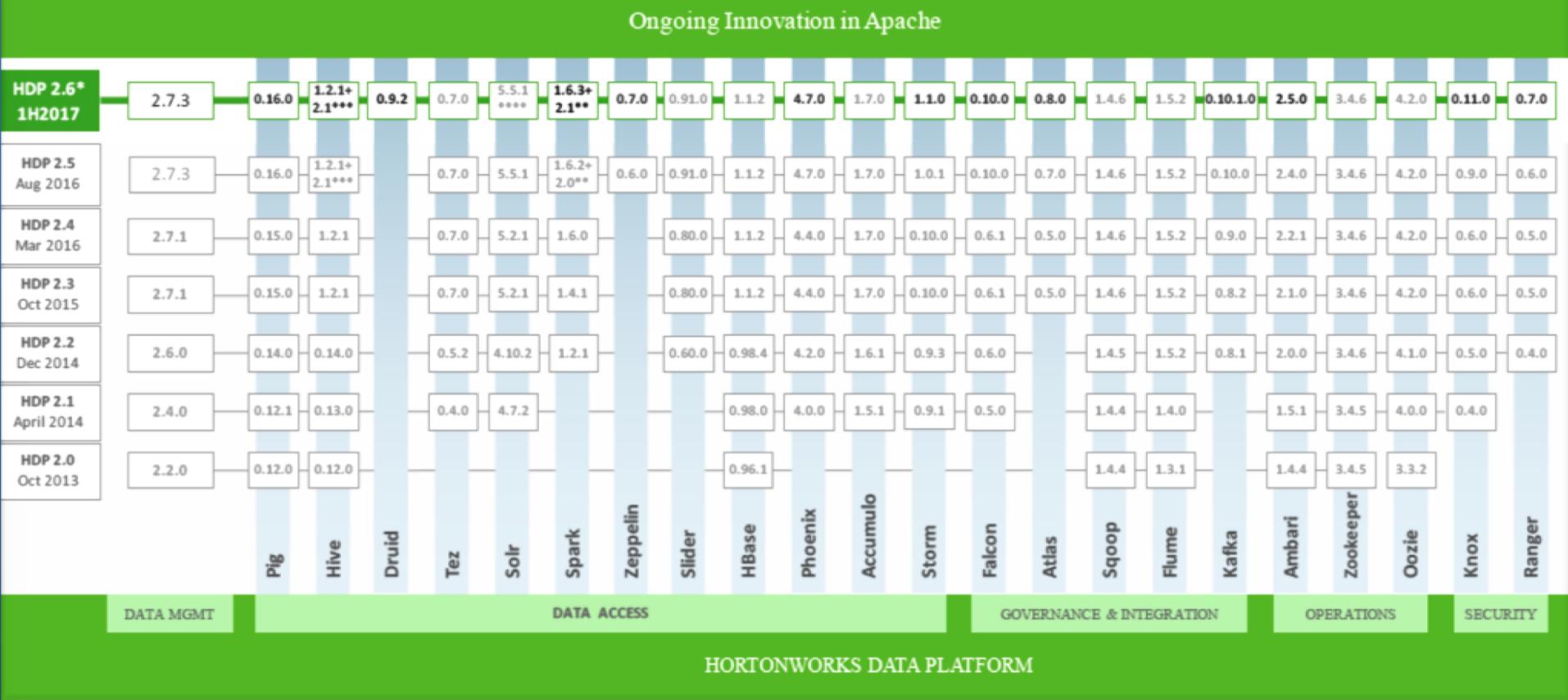
- Hortonworks is a business computer software company focusing on the development and support of Apache Hadoop.
- Hortonworks' product named Hortonworks Data Platform (HDP).
- HDP includes Apache Hadoop and is used for storing, processing, and analyzing large volumes of data from many sources and formats.
- Provides world-class support for Enterprise Hadoop from development to production.
- Is an open and single platform for any data and any workload.

Why Hortonworks Data Platform (HDP)?

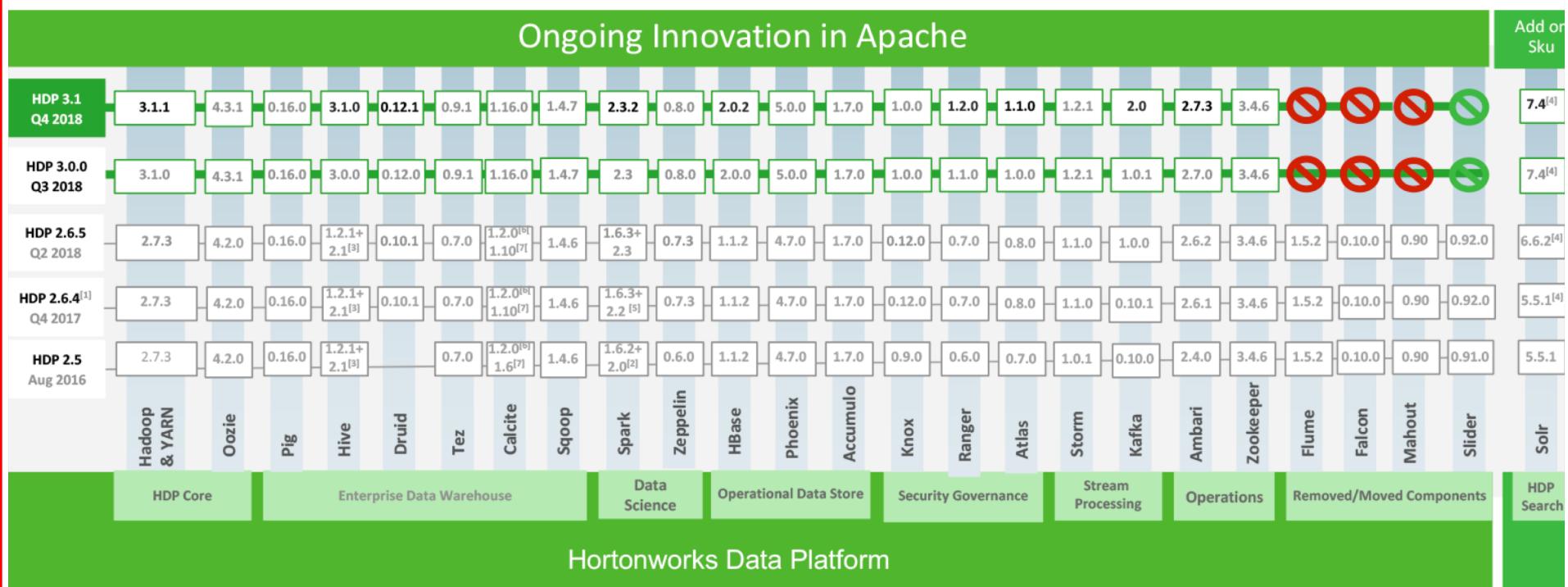
- Is complete, offering not just data processing and management, but the enterprise capabilities to match the demands of an enterprise spanning security, governance and operations. All of this is delivered in 100% open source.
- Is wholly integrated with existing data center investments and can be deployed in almost all computing environments.
- The platform includes various Apache Hadoop projects including the Hadoop Distributed File System (HDFS), MapReduce, Pig, Hive, HBase, Hue, Oozie, Ambari, Storm, Spark, Zookeeper and more.

HDP 2.x

The HDP (2.x) distribution increasingly incorporates many innovations in the Hadoop ecosystem



HDP 3.x

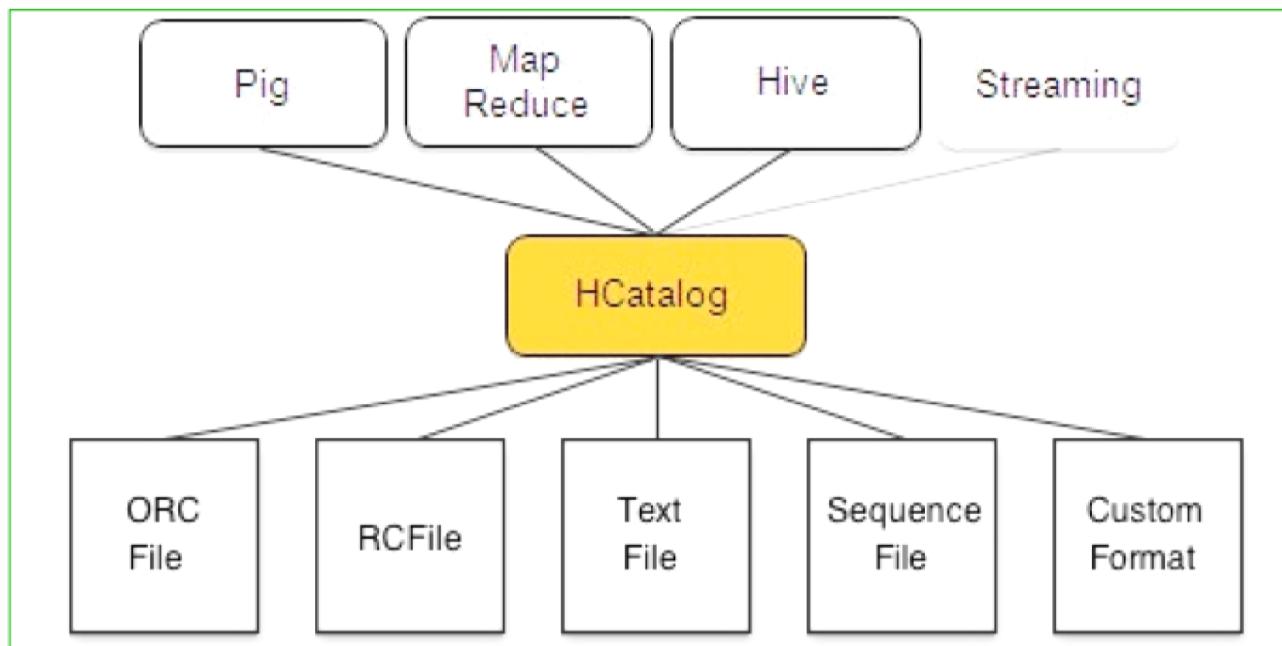


What is Hortonworks Sandbox?

- Hortonworks Sandbox is a personal, portable Hadoop environment that provides an easy and effective way to learn Enterprise Hadoop on a single-node cluster in a virtual machine.
- It comes with a pre-configured image.

HCatalog

- HCatalog is a table and storage management layer for Hadoop.
- It enables users with different data processing tools (Pig, MapReduce) to more easily read and write data.
- HCatalog's table abstraction presents users with a relational view of data in HDFS and ensures that users need not worry about where or in what format their data is stored.
- By default, HCatalog supports RCFile, CSV, JSON, and SequenceFile, and ORC file formats.



Ambari

A completely open framework for provisioning, managing and monitoring Apache Hadoop clusters

- Offers an intuitive collection of tools and APIs that mask the complexity of Hadoop, simplifying the operation of clusters no matter the size of the Hadoop cluster.
- Features:
 - Wizard-driven interface
 - API-driven installations
 - Granular service control
 - Configuration change history
 - Extensible framework
 - Customizable user interface
 - User views
 - File Browser for accessing HDFS.
 - Metastore Browser for accessing Hive metadata and HCatalog.
 - Hive Editor for developing and running Hive queries.
 - Pig Editor for submitting Pig scripts.
 - ...

Ambari

Ambari Sandbox 0 ops 0 alerts Dashboard Services Hosts Alerts Admin admin admin

HDFS

YARN

MapReduce2

Tez

Hive

HBase

Pig

Sqoop

Oozie

ZooKeeper

Falcon

Storm

Flume

Ambari Infra

Atlas

Kafka

Knox

Ranger

Metrics Heatmaps Config History

Metric Actions ▾ Last 1 hour ▾

HDFS Disk Usage  42%	DataNodes Live 1/1	HDFS Links NameNode Secondary NameNode 1 DataNodes More... ▾	Memory Usage No Data Available	Network Usage No Data Available
CPU Usage No Data Available	Cluster Load No Data Available	NameNode Heap  30%	NameNode RPC 1.33 ms	NameNode CPU WIO n/a
NameNode Uptime 566.5 s	HBase Master Heap n/a	HBase Links No Active Master 1 RegionServers n/a More... ▾	HBase Ave Load n/a	HBase Master Uptime n/a

Access the Linux Shell Web Client

- Start HDP Sandbox in VM and keep it running
- Go to <http://127.0.0.1:4200/> to access the shell web client (Or use ssh client via <ssh root@127.0.0.1 -p 2222>)
- Log in:
 - Username: **root**
 - Password: **hadoop** (very 1st time, change to a new password) or the updated password you set

Access Hortonworks Sandbox via Ambari

- Keep the Hortonworks Sandbox running in VM
- Go to <http://127.0.0.1:8080/>
- You may use the following users and passwords (that came with HDP) to log in

Username	Password	Roles
raj_ops	raj_ops	For infrastructure build and R&D activities
maria_dev	maria_dev	For preparing and getting insight from data
holger_gov	holger_gov	For the management of data elements
amy_ds	amy_ds	For exploratory data analysis, cleanup and transformation

Reset Ambari the **admin** Password

- Run this command at the prompt
(127.0.0.1:4200)
`ambari-admin-password-reset`
- Give **your new password** for Ambari.
It may take some time to complete. If your Ambari doesn't restart automatically, restart ambari service with command:
`ambari-agent restart`
- Next time you can access Ambari with
 - Username: admin
 - Password: **your new password**

Using Ambari: the User Views

The screenshot shows the Ambari User Views interface. At the top, there is a navigation bar with the Ambari logo, a Sandbox button, and a status bar showing 0 ops and 0 alerts. To the right of the status bar are links for Dashboard, Services, Hosts, Alerts, and Admin. On the far right, there is a user dropdown menu labeled "admin". Below the navigation bar, the main content area is titled "Your Views". It lists several views, each with a "View" button and a brief description:

- YARN Queue Manager (1.0.0)**
Manage YARN Capacity Scheduler Queues
- Files View (1.0.0)**
This view instance is auto created when the HDFS service is added to a cluster.
- Hive View (1.5.0)**
This view instance is auto created when the Hive service is added to a cluster.
- Hive View 2.0 (2.0.0)**
This view instance is auto created when the Hive service is added to a cluster.
- Pig View (1.0.0)**
User Interface to write and execute Pig scripts
- Storm View (0.1.0)**
Manage Storm
- Tez View (0.7.0.2.6.1.0-118)**
Monitor and debug all Tez jobs, submitted by Hive queries and Pig scripts (auto-created)
- Workflow Manager (1.0.0)**
Workflow manager for Apache Oozie

On the right side of the "Your Views" list, there is a vertical sidebar with a grid icon and a dropdown menu for "admin". The sidebar also lists additional views that are not currently selected:

- YARN Queue Manager
- Files View
- Hive View
- Hive View 2.0
- Pig View
- Storm View
- Tez View
- Workflow Manager

Ambari User Views: Files View

Ambari Sandbox 0 ops 0 alerts

Dashboard Services Hosts Alerts Admin

Home New Folder Upload Search (0)

Total: 15 files or folders

Name	Size	Last Modified	Owner	Group	Permission
apps	--	2017-07-28 10:06	hdfs	hdfs	drwxr-xr-x
ats	--	2017-07-28 10:00	yarn	hadoop	drwxr-xr-x
demo	--	2017-07-28 10:14	hdfs	hdfs	drwxr-xr-x
hdp	--	2017-07-28 10:00	hdfs	hdfs	drwxr-xr-x
livy-recovery	--	2017-07-28 10:02	livy	hdfs	drwx-----
livy2-recovery	--	2017-07-28 10:01	livy	hdfs	drwx-----
mapred	--	2017-07-28 10:00	mapred	hdfs	drwxr-xr-x
mr-history	--	2017-07-28 10:00	mapred	hadoop	drwxrwxrwx
ranger	--	2017-07-28 10:00	hdfs	hdfs	drwxr-xr-x
spark-history	--	2017-07-28 10:27	spark	hadoop	drwxrwxrwx
spark2-history	--	2019-02-20 00:44	spark	hadoop	drwxrwxrwx
stsci5065	--	2019-02-20 00:19	root	hdfs	drwxr-xr-x

Ambari User Views: Files View, Browse Files

The screenshot shows the Ambari user interface for browsing files. At the top, there is a navigation bar with links for Ambari, Sandbox, Dashboard, Services, Hosts, Alerts, and Admin. Below the navigation bar, the main content area shows a breadcrumb path: / > stsci5065 > data. A yellow box indicates "Total: 3 files or folders". On the left, there are icons for Home, Upload, and Refresh. On the right, there are buttons for "Select All" and "New Folder", and a search bar with the placeholder "Search in current directory...". The main table lists three files: 100-0.txt, 74-0.txt, and influenza.cds, with columns for Name, Size, Last Modified, Owner, Group, and Permissions.

Name >	Size >	Last Modified >	Owner >	Group >	Permissions
100-0.txt	5.6 MB	2019-02-20 00:34	root	hdfs	-rw-r--r--
74-0.txt	418.1 kB	2019-02-20 00:34	root	hdfs	-rw-r--r--
influenza.cds	1.1 GB	2019-02-20 00:35	root	hdfs	-rw-r--r--

Ambari User Views: Files View, Click **Open** to Preview a File

The screenshot shows the Ambari user interface for viewing files. At the top, there's a navigation bar with links for Ambari, Sandbox, Dashboard, Services, Hosts, Alerts, Admin, and a user dropdown for 'admin'. Below the navigation bar, the main area shows a file tree under 'stsci5065 > data'. A file named '100-0.txt' is selected and highlighted in blue. A modal window titled 'File Preview' is open over the file list, displaying the contents of the selected file. The preview text is as follows:

Project Gutenberg's The Complete Works of William Shakespeare, by
William Shakespeare

This eBook is for the use of anyone anywhere in the United States and
most other parts of the world at no cost and with almost no
restrictions whatsoever. You may copy it, give it away or re-use it
under the terms of the Project Gutenberg License included with this
eBook or online at www.gutenberg.org. If you are not located in the
United States, you'll have to check the laws of the country where you
are located before using this ebook.

See at the end of this file: * CONTENT NOTE (added in 2017) *

Title: The Complete Works of William Shakespeare

Author: William Shakespeare

Ambari User Views: Files View, Click **Permissions** to Change Permissions of a File

The screenshot shows the Ambari interface with a red border around the main content area. At the top, there's a navigation bar with links for Ambari, Sandbox, Dashboard, Services, Hosts, Alerts, Admin, and a user icon for 'admin'. Below the navigation bar is a toolbar with icons for Open, Rename, Permissions, Delete, Copy, Select All, New Folder, Upload, and a search bar. The main area displays a file list for the directory 'stsci5065 > data'. The file '100-0.txt' is selected, and its details are shown: Name (100-0.txt), Size (5.6 kB), and Type (Text File). To the right of the file list is a table titled 'File Edit Permissions' showing permissions for User, Group, and Other. The 'User' row has 'Read' selected. The 'Group' and 'Other' rows also have 'Read' selected. Below the table are 'Cancel' and 'Save' buttons. In the background, there's a list of HDFS files: '100-0.txt', 'influenza.cds', and 'tom_sawyer.txt', each with their respective sizes and types.

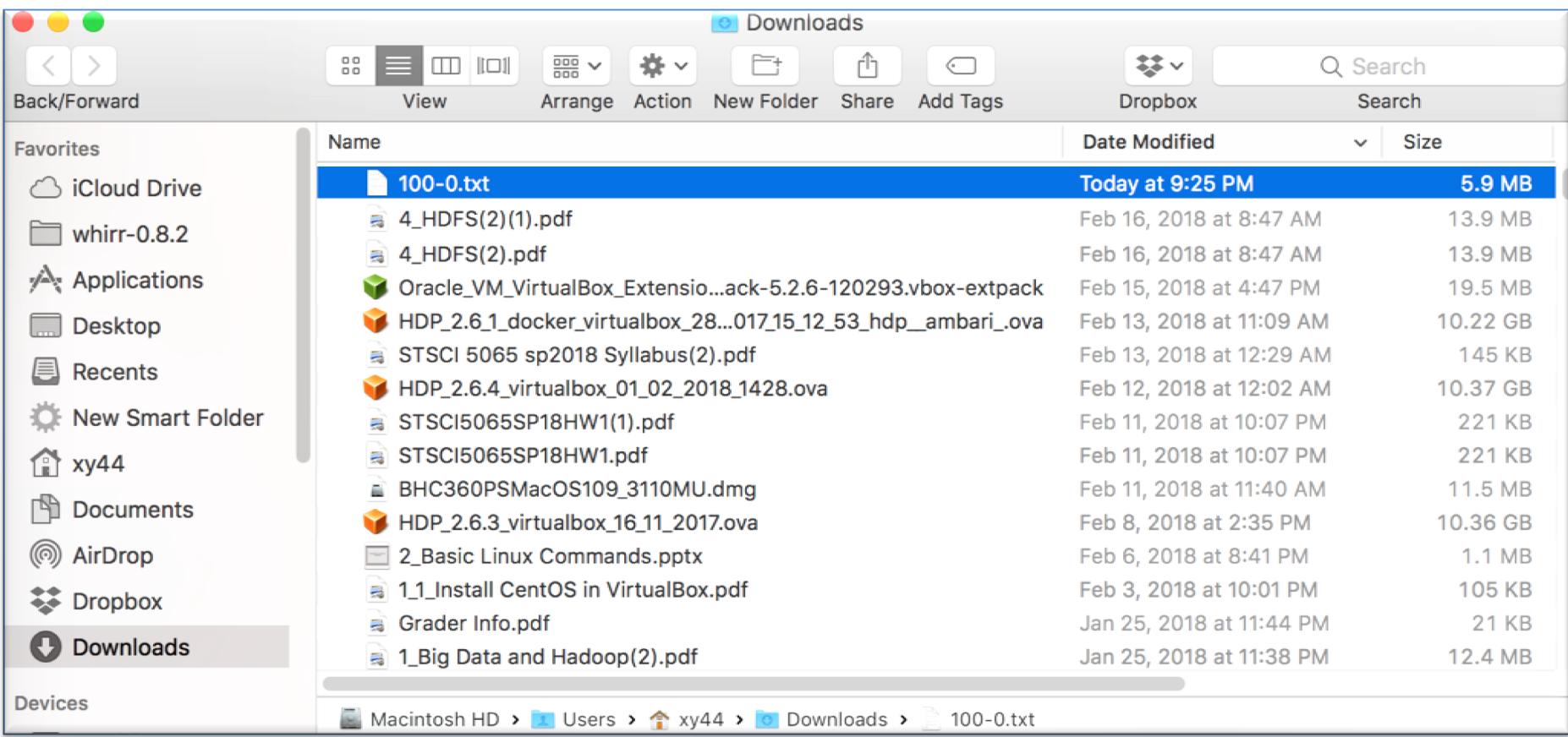
Name	Size	Type
100-0.txt	5.6 kB	Text File
influenza.cds	1.0 kB	CDH File
tom_sawyer.txt	422 B	Text File

User	Read	Write	Execute
Group	Read	Write	Execute
Other	Read	Write	Execute

Cancel Save

Ambari User Views:

Files View, Click **Download** to Download a File to Your Local OS (Mac or Windows)



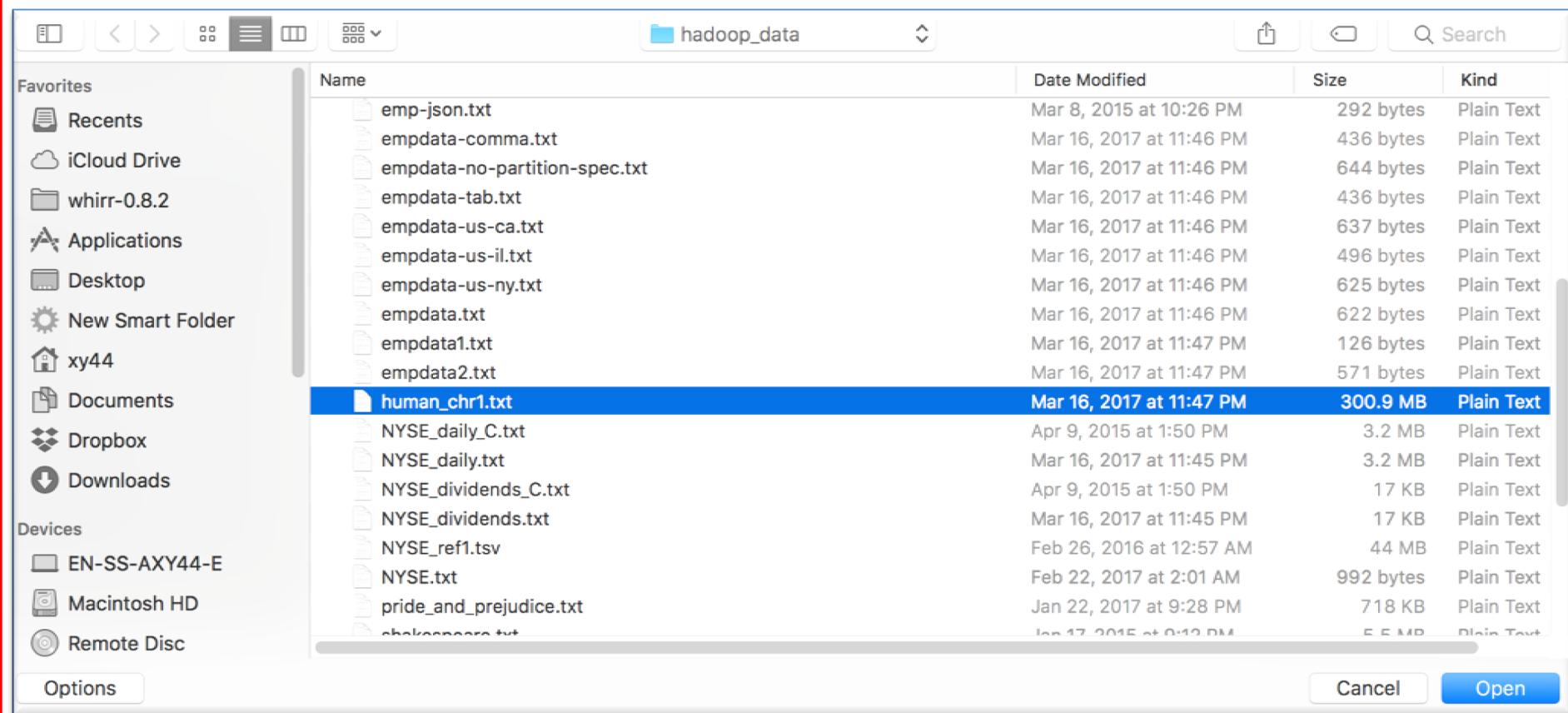
Ambari User Views:

Files View, Click **Upload** to Upload a File from Your Local OS (Mac or Windows) (1)

The screenshot shows the Ambari user interface for managing files. The top navigation bar includes links for Ambari, Sandbox, Dashboard, Services, Hosts, Alerts, Admin, and a user account for 'admin'. Below the navigation is a breadcrumb trail: / > stsci5065 > data. A message at the top center indicates '1 Files, 0 Folders selected'. On the left, a sidebar lists files: 100-0.txt (selected), influenza.cds, and tom_sawyer.txt. The main area displays a file listing with columns for Name, Group, and Permission. The 'Group' column for 100-0.txt is highlighted in blue. A central modal dialog is open, titled 'Upload file to /stsci5065/data'. It features a cloud icon with an upward arrow, a large dashed rectangular area for file selection, and the text 'Drag file to upload or click to browse'. A note below states 'Currently supports single file upload'. A 'Cancel' button is at the bottom right of the dialog. The entire screenshot is framed by a red border.

Ambari User Views:

Files View, Click Upload to Upload a File from Your Local OS (Mac or Windows) (2)



Ambari User Views: Files View, Click Upload to Upload a File from Your Local OS (Mac or Windows) (3)

The screenshot shows the Ambari interface for managing files. At the top, there's a navigation bar with various links like Apps, Apple, Hue, Pronounce, etc. Below it, the Ambari logo and 'Sandbox' status are displayed. The main area is titled 'data' under 'stsci5065'. A modal window is open, prompting to 'Upload file to /stsci5065/data'. Inside the modal, a progress bar shows '100%' completion for the file 'human_chr1.txt'. The file is listed in the background table along with others: '100-0.txt', 'influenza.cds', and 'tom_sawyer.txt'. The table also includes columns for Group, Permission, Size, Last Modified, and Owner. A search bar is visible on the right, and a notification icon shows '1'.

Name	Group	Permission	Size	Last Modified	Owner
100-0.txt	hdfs	-rw-r--r--			
influenza.cds	hdfs	-rw-r--r--			
human_chr1.txt	hdfs	-rw-r--r--	422.9 kB	2018-02-15 18:24	root
tom_sawyer.txt	hdfs	-rw-r--r--			

Ambari User Views:

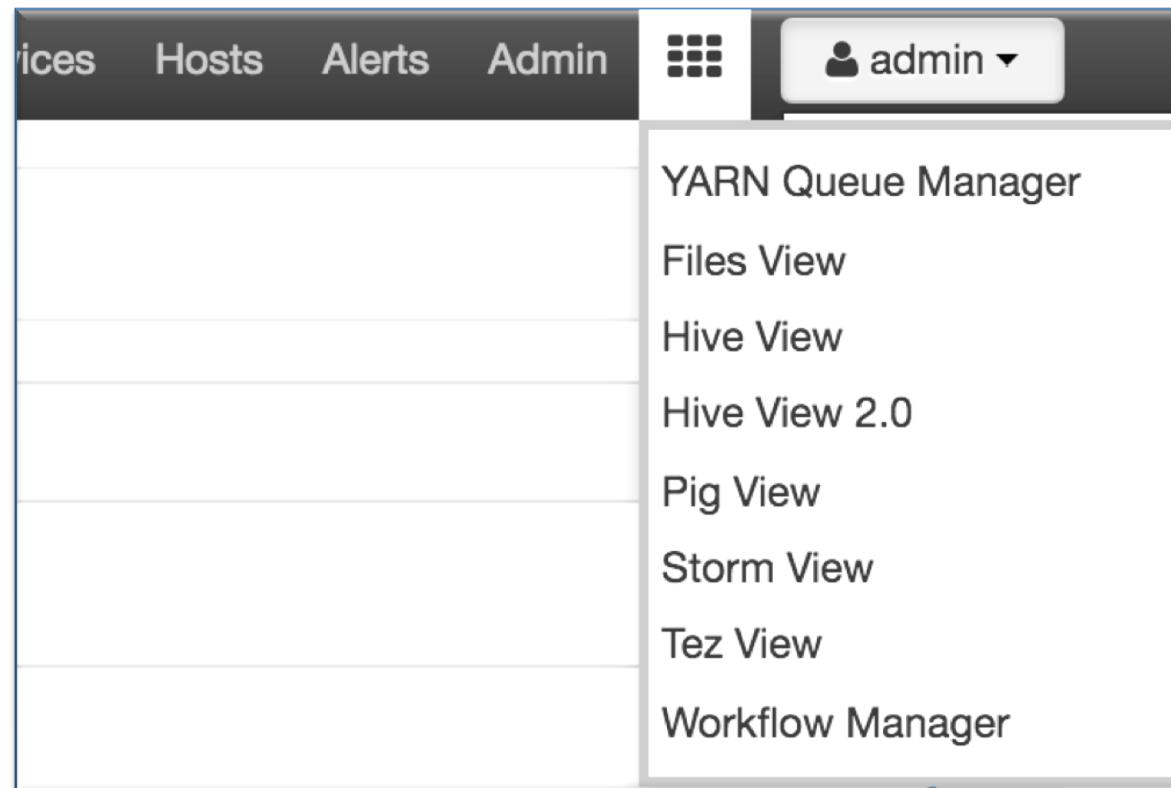
Files View, Click Upload to Upload a File from Your Local OS (Mac or Windows) (4)

The screenshot shows the Ambari user interface for viewing files. At the top, there is a navigation bar with links for Ambari, Sandbox, Dashboard, Services, Hosts, Alerts, Admin, and a user icon for 'admin'. Below the navigation bar, the main content area displays a file listing for the directory '/stsci5065/data'. A yellow box highlights the message 'Total: 4 files or folders'. On the right side of the file list are buttons for '+ Select All', 'New Folder', 'Upload', and a search bar with placeholder text 'Search in current directory...'. The file list table has columns for Name, Size, Last Modified, Owner, Group, and Permission. The rows show the following data:

Name	Size	Last Modified	Owner	Group	Permission
100-0.txt	5.6 MB	2019-02-20 00:34	root	hdfs	-rw-r--r--
74-0.txt	418.1 kB	2019-02-20 00:34	root	hdfs	-rw-r--r--
human_chr1.txt	287.0 MB	2019-02-20 00:58	admin	hdfs	-rw-r--r--
influenza.cds	1.1 GB	2019-02-20 00:35	root	hdfs	-rw-r--r--

Ambari User Views: Hive View

There are Two Hive Views



Ambari User Views: Hive View

The screenshot shows the Ambari Hive View interface. At the top, there is a navigation bar with links for Dashboard, Services, Hosts, Alerts, Admin, and a user icon for 'admin'. Below the navigation bar, there are tabs for Hive, Query, Saved Queries, History, UDFs, and Upload Table. The 'Hive' tab is selected.

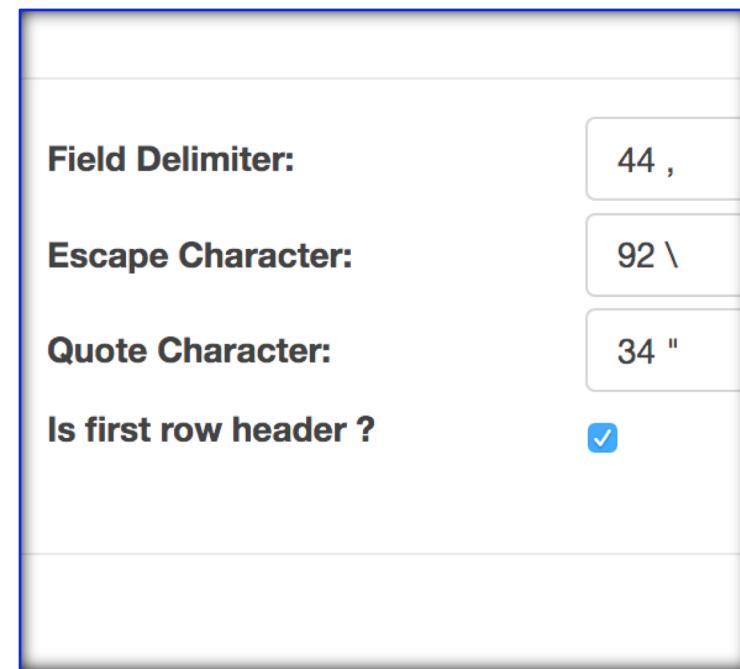
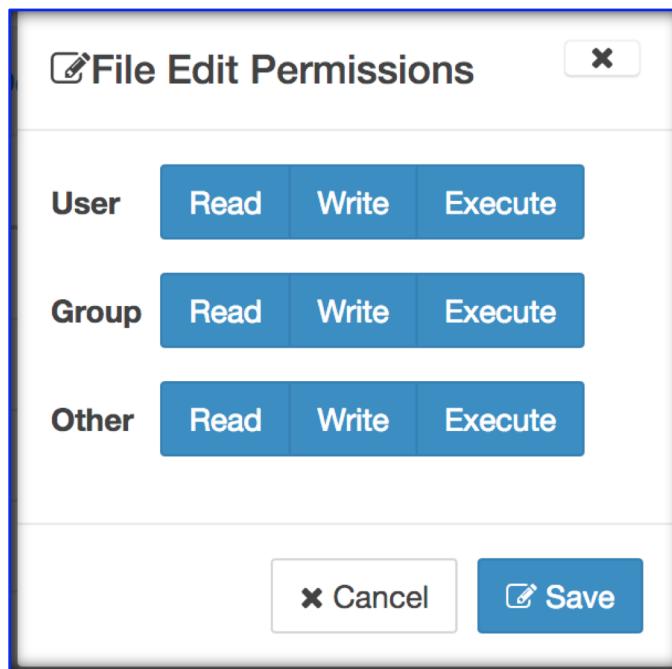
The main area is divided into two main sections: 'Database Explorer' on the left and 'Query Editor' on the right.

- Database Explorer:** This section contains a dropdown menu set to 'default', a search bar labeled 'Search tables...', and a list of databases: default, foodmart, and xademo.
- Query Editor:** This section has a 'Worksheet' tab with the number '1' indicating it is active. It includes a toolbar with icons for SQL, Settings, and Graph, and a sidebar with icons for Databases, Tables, and Functions. A red box highlights the 'SQL' icon in the sidebar.
- Bottom Navigation:** At the bottom of the interface are buttons for 'Execute', 'Explain', 'Save as...', and 'New Worksheet'.

Ambari User Views: Hive View

Prepare to Upload/Create a Table

- Change file permissions to 777 (often) in Files View
- Set right delimiter and import row header if any



Ambari User Views: Hive View

Upload/Create a Table

The screenshot shows the Ambari Hive View interface. At the top, there are tabs for Hive, Query, Saved Queries, History, UDFs, and Upload Table. The Upload Table tab is selected. The main area has two sections: 'Upload from Local' and 'Upload from HDFS'. In the 'Upload from Local' section, 'File type' is set to CSV, 'Database' to default, and 'Stored as' to ORC. In the 'Upload from HDFS' section, the 'HDFS Path' is /stsci5065/data/NYSE.csv, the 'Table name' is NYSE, and the 'Contains endlines?' checkbox is unchecked. Below these sections is a large table preview with columns: column1, column2, column3, and column4. The data rows are: NYSE, BCE, 2009-12-11, 0.386; NYSE, BCE, 2009-09-11, 0.373; NYSE, BCE, 2009-06-11, 0.35; NYSE, BCE, 2009-03-12, 0.309; NYSE, BCE, 2008-12-19, 0.296; NYSE, BCE, 2008-03-12, 0.363.

column1	column2	column3	column4
NYSE	BCE	2009-12-11	0.386
NYSE	BCE	2009-09-11	0.373
NYSE	BCE	2009-06-11	0.35
NYSE	BCE	2009-03-12	0.309
NYSE	BCE	2008-12-19	0.296
NYSE	BCE	2008-03-12	0.363

Ambari User Views: Hive View

Upload/Create a Table

The screenshot shows the Ambari Hive View interface. The top navigation bar includes links for Ambari, Sandbox, Dashboard, Services, Hosts, Alerts, Admin, and a user dropdown for 'admin'. Below the navigation is a menu bar with tabs: Hive, Query, Saved Queries, History, UDFs, and Upload Table. The main area is divided into two sections: 'Database Explorer' on the left and 'Query Editor' on the right.

Database Explorer: This section displays a list of databases. The 'default' database is selected, indicated by a dropdown menu. A search bar below it contains the placeholder 'Search tables...'. The list of databases includes:

- default (selected)
- nyse (circled in red)
- sample_07
- sample_08
- foodmart
- xademo

A blue arrow points from the 'nyse' database entry in the list to a blue-bordered callout box containing the text 'Newly created table'.

Query Editor: This section features a 'Worksheet' tab labeled '1'. At the bottom of the editor are buttons for 'Execute', 'Explain', 'Upload', 'Save as...', and 'New Worksheet'.

Right Sidebar: A vertical sidebar on the right contains icons for information, SQL, settings, and monitoring, along with a 'TEZ' icon with a red '2' notification badge.

Ambari User Views: Hive View

Get a Description of a Table

The screenshot shows the Ambari Hive View interface. At the top, there's a navigation bar with links for Ambari, Sandbox, Dashboard, Services, Hosts, Alerts, and Admin. Below the navigation bar, there are tabs for Hive, Query, Saved Queries, History, UDFs, and Upload Table. The current tab is 'Query'.

In the main area, there are two panes: 'Database Explorer' on the left and 'Query Editor' on the right.

Database Explorer: Shows the selected database is 'foodmart'. There's a dropdown menu and a search bar labeled 'Search tables...'. A list of databases is provided, including default, foodmart, customer, inventory_fact_1998, product, sales_fact_dec_1998, store, and xademo.

Query Editor: The 'Worksheet' tab is active. It contains the SQL command: `1 describe customer;`. Below the worksheet are buttons for Execute, Explain, Upload, and Save as... The status of the query is shown as 'SUCCEEDED'.

Query Process Results (Status: SUCCEEDED): This section shows the results of the 'describe customer' query. It has tabs for Logs and Results, with 'Results' being the active tab. A 'Filter columns...' input field is present. The results table has columns: col_name, data_type, and comment. The data is as follows:

col_name	data_type	comment
customer_id	int	''
account_num	bigint	''
lname	varchar(30)	''
fname	varchar(30)	''

Ambari User Views: Hive View

An SQL Query

The screenshot shows the Ambari Hive View interface. At the top, there's a navigation bar with tabs for Ambari, Sandbox, Dashboard, Services, Hosts, Alerts, Admin, and a user icon labeled 'admin'. Below the navigation bar, there are tabs for Hive, Query, Saved Queries, History, UDFs, and Upload Table. The 'Query' tab is selected.

In the main area, there are two main sections: 'Database Explorer' on the left and 'Query Editor' on the right.

Database Explorer: This section shows a dropdown menu set to 'foodmart' and a search bar labeled 'Search tables...'. A list of databases is provided, including 'default', 'foodmart', 'customer', 'inventory_fact_1998', 'product', 'sales_fact_dec_1998', 'store', 'xademo', 'call_detail_records', 'customer_details', and 'recharge_details'. Each database name has a small blue icon next to it.

Query Editor: This section contains a 'Worksheet' area with the following content:

```
1 select * from customer where num_children_at_home > 4;
```

Below the worksheet are buttons for 'Execute', 'Explain', and 'Save as...', and a 'New Worksheet' button.

Query Process Results (Status: SUCCEEDED): This section shows the results of the executed query. It includes tabs for 'Logs' and 'Results', a 'Save results...' button, and a 'Filter columns...' input field. The results table has the following columns: customer.yearly_income, customer.gender, customer.total_children, customer.num_children_at_home, and customer.edu.

customer.yearly_income	customer.gender	customer.total_children	customer.num_children_at_home	customer.edu
30K - \$50K	M	5	5	High School C
10K - \$30K	F	5	5	Partial High Sc

Ambari User Views: Pig View

The screenshot shows the Ambari user interface for the Pig View. At the top, there is a navigation bar with the Ambari logo, the word "Sandbox" followed by "0 ops" and "0 alerts", and links for Dashboard, Services, Hosts, Alerts, Admin, and a user dropdown for "admin". On the left, a sidebar menu is open for the "group.pig" script, showing options to Save, Copy, or Delete the script. The main area displays the script code in a text editor. The code is as follows:

```
1 divs = load 'NYSE_dividens.txt' as (exchange:chararray,
2     symbol:chararray, date:chararray, dividends:float);
3 grpds = group divs by symbol;
4 avgdiv = foreach grpds generate group, AVG(divs.dividends);
5 dump avgdiv
6 |
```

Below the code, there are two buttons: "Execute on Tez" (disabled) and a large blue "Execute" button with a dropdown arrow. To the right of the execute buttons, the full path of the generated pig script is shown: "/tmp/.pigschemas/grouppig-2018-02-21_01-42.pig".

Ambari User Views: Pig View

The screenshot shows the Ambari user interface for a completed Apache Pig job named "group.pig". The top navigation bar includes links for Ambari, Sandbox, Dashboard, Services, Hosts, Alerts, Admin, and a user dropdown for "admin". The main content area displays the job details: "group.pig - Completed" with a green progress bar indicating completion. Below this, the Job ID is listed as "job_1519163059147_0003" and the start time as "2018-02-20 20:51". On the left, a sidebar provides options to "Save", "Copy", or "Delete" the script. Two sections are visible below: "Results" and "Logs". The "Results" section has a "Download" button. The "Logs" section shows the following log entries:

```
18/02/21 01:51:18 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/02/21 01:51:18 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/02/21 01:51:18 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
18/02/21 01:51:18 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
18/02/21 01:51:18 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
2018-02-21 01:51:18,148 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0.
2018-02-21 01:51:18,148 [main] INFO org.apache.pig.Main - Logging error messages to:
2018-02-21 01:51:18,676 [main] INFO org.apache.pig.impl.util.Utils - Default bootup
2018-02-21 01:51:18,782 [main] INFO org.apache.pig.backend.hadoop.executionengine.HT
```