



# How to **OPTIMIZE** your data lake

Dremio training series

1

2

3

4

# Dremio training series

## 1

### Dremio for users & developers

- System and concept overview
- Using the Dremio UI
- Dremio ANSI SQL
- Jobs, queues and reflections
- ODBC, JDBC, and REST clients

## 2

### Dremio for data engineers

- Semantic layer concepts
- Permissioning
- Overview of reflections
- Reflection best practices

## 3

### Query analysis & performance

- Using query profiles
- Troubleshooting performance
- Query profile examples
- Query acceleration techniques

## 4

### Dremio for administrators

- Deployment architecture
- Managing capacity and jobs
- System and query monitoring
- Metadata and reflection freshness
- Using the Dremio admin CLI

# System and concept overview

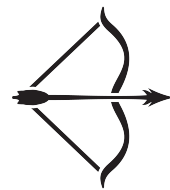
Including



Arrow

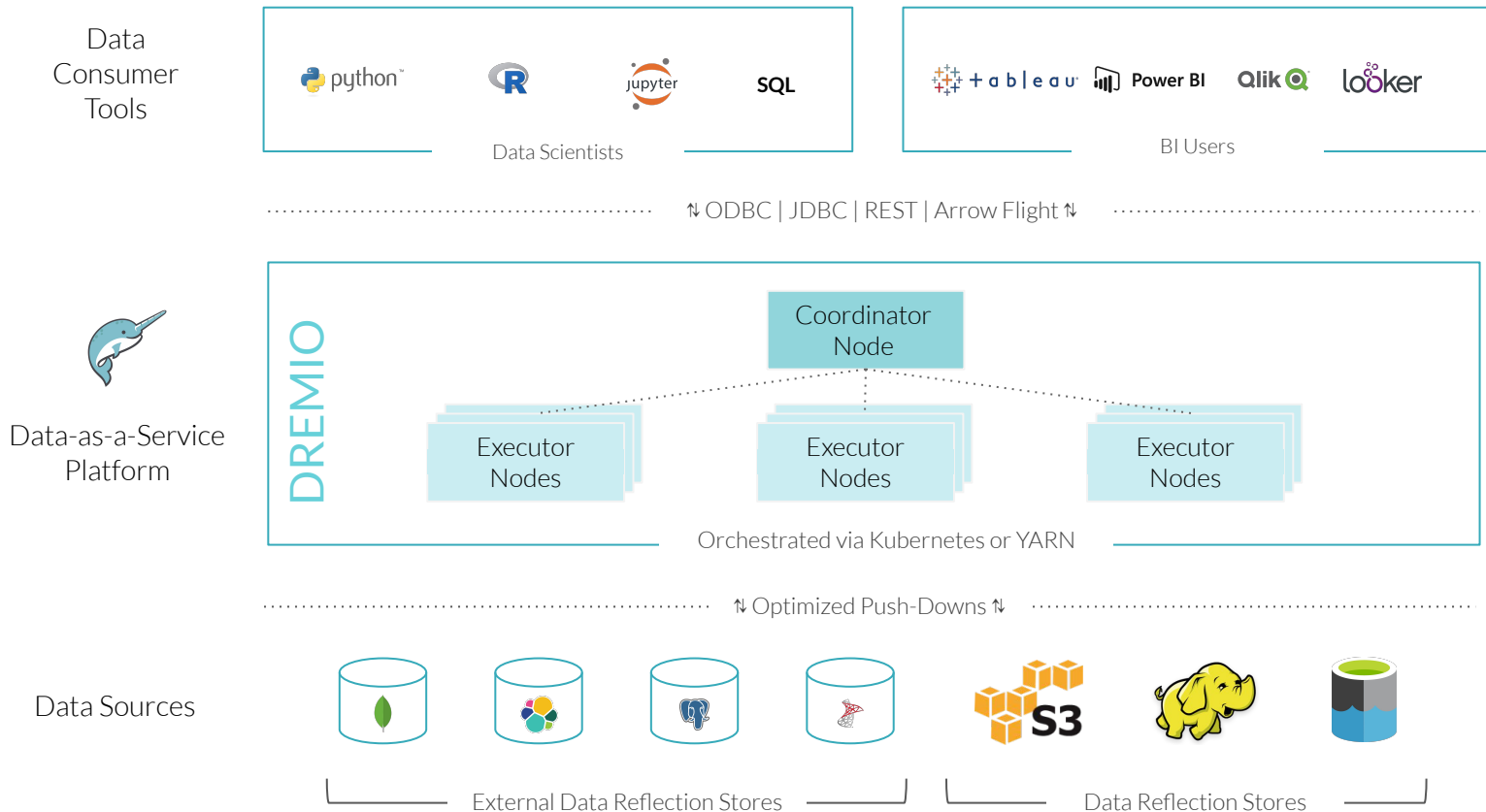


Parquet



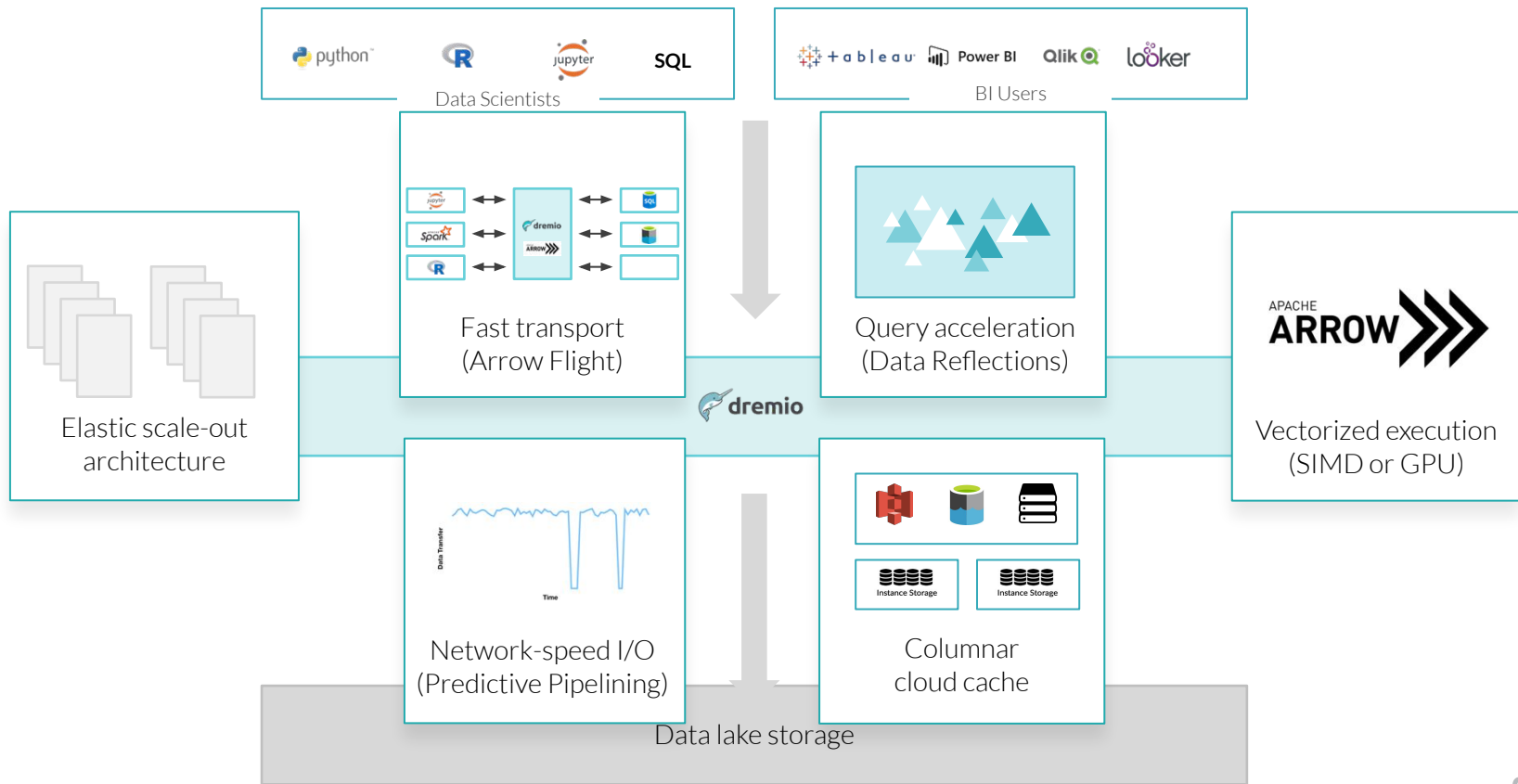
Gandiva

# Dremio deployment architecture





# How Dremio delivers data lake performance



# Key components of Dremio architecture

## Coordinator node

Responsible for:

- Hosting ODBC/JDBC & HTTP endpoints
- Query planning
- Managing metadata

Metadata includes information about data sources, virtual datasets, access controls, job profiles and more

- Persisted to file system
- No dependence on external database
- Uses RocksDB & Lucene

## Executor nodes

Responsible for:

- Query execution and data access
- Storing query results, user uploads, user downloads & data reflections (caches) in distributed file system

The number of executors in the cluster

- Scaled by workload size
- System can tolerate N-1 executor failures, but may be degraded before then

# Key objects

## Sources and datasets



**Source** - a data store connected to Dremio



**Physical DataSet (PDS)** - a source dataset  
- e.g. RDBMS table, HDFS directory



**Virtual DataSet (VDS)** - Similar to an RDBMS view, but with more features. VDSs have lineage



**Reflection** - Similar to a materialized view. Accelerates query performance. Transparent to users.

## Spaces



**Space** - a location that organizes shared VDSs



**Home space** - a private location for your own PDS and VDS



**Folder** - a subdirectory to organize items in a space

# Physical Dataset (PDS)

- Data from a configured source, or uploaded file with data format
- Immutable, data curation contained in upper-level VDS definitions
- Identified with Universal Resource Identifier which includes data lake prefix/directory structure
- Access controlled by user & group, normally restricted in Dremio to administrator

The screenshot displays the Dremio web interface. At the top, there's a navigation bar with 'Datasets', 'Jobs', and a search bar. The main area is split into a left sidebar and a central pane. The sidebar has 'Spaces' and 'Sources' sections. The 'Sources' section lists various data sources, with 'adlsv2' highlighted. The central pane shows the 'adlsv2.dremio' dataset structure, including a list of datasets like 'conditions-nyc', 'dist', 'employees', 'fs', 'nyctaxi', 'nytaxi\_10m', 'profiles\_query', 'trips', 'trips\_nyc', and 'yelp'. A right-hand pane shows a detailed view of the dataset, including a list of datasets like 'conditions-nyc', 'dist', 'employees', 'fs', 'nyctaxi', 'nytaxi\_10m', 'profiles\_query', 'trips', 'trips\_nyc', and 'yelp'. A green notification banner at the bottom right states 'Copied adlsv2.dremio.\"trips\_nyc\".' with a close button. A mouse cursor is pointing at the 'trips\_nyc' dataset entry.

# Virtual Datasets (VDS)

- View of the curated data from one or more PDS
- Data curation can include joins, filters, type casting, parsing, derived columns.
- Defined by SQL statements, which ideally include business logic for repeatable rapid execution
- Saved in Home space or shared space

The screenshot displays the Dremio interface with the following components:

- Header:** Dremio logo, tabs for Datasets, Jobs, and a search bar labeled "Search Catalog...". A "New Query" button is also present.
- Left Panel:** User profile "mhoerth@dremio.com".
  - Datasets:** A list of datasets with icons and counts. "Solution Architecture" is highlighted with 19 items.
  - Sources:** A list of data sources including adls, adlsv2, EMC-ECS (S3 Compatible), local-fs, postgres, s3, Samples, and sqlserver.
- Right Panel:** "Solution Architecture" view showing a hierarchy.
  - Name:** A dropdown menu.
  - Folders:** A list of folders including brock, Jonny, Louis, and Scott.
  - Datasets:** A list of datasets including Curated Employees, employee123, Employees Dataset (tagged "golden"), test\_emp, and testing.
  - Callouts:** Three teal callout boxes are present: "Folders" pointing to the brock folder, "Tags" pointing to the "golden" tag on the Employees Dataset, and "Copy the URI" pointing to the testing dataset icon.

# PDS, VDS, and reflections

**PDS**



*Data from a source or uploaded  
from a data file*

**VDS**



*Like a view --  
Curated from one or more PDS*

**Reflection**



*Like a materialized view --  
accelerates queries on both PDS or VDS*

# Spaces enable data sharing and organization

- Shared locations for saving virtual datasets
- Sharing can be configured for each space
- Group datasets by a common theme such as a project or geographic region
- Enables building a semantic layer
- Home Space is private user work area
- Users will not see spaces for which they have no authorization

The screenshot shows the Dremio interface with the following components:

- Top Navigation Bar:** Includes the Dremio logo, 'Datasets', 'Jobs', a search bar labeled 'Search Catalog...', and a 'New Query' button.
- Spaces Section:** A sidebar on the left lists various spaces, including 'mhoerth' (Home space), 'AC Organization', 'Allied', 'aneesh', 'AppD', 'ara', 'ATC', 'Attrition Model', 'aws', 'baml', 'CapitalOne', and 'cars'.
- AC Organization Space Detail:** The right pane shows the contents of the 'AC Organization' space, which includes a directory named 'NEE Project' and several datasets like 'AC Employees2', 'Adobe Employees', 'CE Project', 'EOS Employees', 'Funcionarios iMasters', 'MarkEmployees', 'metest', 'My Employees', and 'Pfizer Reviews'.

Callout boxes provide additional context:

- Home space, private to the user:** Points to the 'mhoerth' space in the sidebar.
- Items in the space:** Points to the list of datasets within the 'AC Organization' space.
- Organize spaces with directories:** Points to the 'NEE Project' directory within the 'AC Organization' space.

# User uploaded data files

- Analyze auxiliary data on desktop / shared location by uploading it to Dremio
- Supported File Formats: Text (Delimited), JSON, Parquet, Excel, XLS
- Data is uploaded to user's personal space
- Can be shared with others by saving a Virtual Dataset in a shared Space

The screenshot displays the Dremio web interface. The top navigation bar includes the Dremio logo, tabs for 'Datasets', 'Jobs', and a search bar. The main content area is divided into a left sidebar and a central pane. The sidebar shows a tree view of 'Spaces' and 'Sources'. The 'Spaces' section is expanded, showing 'admin' as the 'Home space'. The central pane displays a list of datasets for the '@admin' user, including 'SF\_incidents2016', 'zip\_lookup', and 'zips'. A teal callout box labeled 'Home space' points to the 'admin' space in the sidebar. Another teal callout box labeled 'Upload button' points to the '+ [folder icon] [cloud icon]' button in the top right of the dataset list. An 'Add File: Browse for File (Step 1 of 2)' dialog is open in the foreground, featuring a large dashed box with a cloud and arrow icon, and the text 'Drop a local file here, or [browse](#).'

**Home space**

**Upload button**

Add File: Browse for File (Step 1 of 2)

Drop a local file here, or [browse](#).



# Promoting a file system source to a PDS

- Promote Individual File or Directory of similarly structured files into a Physical Dataset
- Upon promotion of Directory into PDS, sub-directory hierarchy is available as virtual columns (dir0, dir1 ..) in the PDS

The screenshot illustrates the Dremio interface for promoting a file system source to a PDS. The top panel shows the 'Samples (6).samples.dremio.com' dataset with files like 'SF weather 2018-2019.csv'. A teal callout 'Add format' points to the 'Add format' button. The bottom panel shows the 'Dataset Settings' dialog with 'Text (delimited)' format selected. A teal callout 'Verify' points to the 'Verify' button. The right panel shows the 'Resulting PDS' with the file 'SF weather 2018-2019.csv' promoted to a PDS.

**Dataset Settings**

Format: Text (delimited)

Field Delimiter: Comma, Quote: Double Quote, Comment: Number Sign, Line Delimiter: LF - Unix/Linux, Escape: Double Quote

Options: ☒ Extract Field Names, ☐ Skip First Line, ☒ Trim Field Names

Abc	STATION	Abc	NAME	Abc	LATITUDE	Abc	LONGITUDE	Abc	ELEVATION	Abc	DATE
USWB0023272	SAN FRANCISCO DOWNTOWN, CA	37.7785	-122.4269	45.7	2018-01-01						
USWB0023272	SAN FRANCISCO DOWNTOWN, CA	37.7785	-122.4269	45.7	2018-01-02						
USWB0023272	SAN FRANCISCO DOWNTOWN, CA	37.7785	-122.4269	45.7	2018-01-03						
USWB0023272	SAN FRANCISCO DOWNTOWN, CA	37.7785	-122.4269	45.7	2018-01-04						
USWB0023272	SAN FRANCISCO DOWNTOWN, CA	37.7785	-122.4269	45.7	2018-01-05						
USWB0023272	SAN FRANCISCO DOWNTOWN, CA	37.7785	-122.4269	45.7	2018-01-06						
USWB0023272	SAN FRANCISCO DOWNTOWN, CA	37.7785	-122.4269	45.7	2018-01-07						
USWB0023272	SAN FRANCISCO DOWNTOWN, CA	37.7785	-122.4269	45.7	2018-01-08						
USWB0023272	SAN FRANCISCO DOWNTOWN, CA	37.7785	-122.4269	45.7	2018-01-09						
USWB0023272	SAN FRANCISCO DOWNTOWN, CA	37.7785	-122.4269	45.7	2018-01-10						
USWB0023272	SAN FRANCISCO DOWNTOWN, CA	37.7785	-122.4269	45.7	2018-01-11						
USWB0023272	SAN FRANCISCO DOWNTOWN, CA	37.7785	-122.4269	45.7	2018-01-12						

# Moving to a new space

- Option to save a new VDS in your home space
- Iterate on the logic
- Move the VDS to a shared space when it's appropriate to share with others

The screenshot shows the Dremio web interface. At the top, there's a navigation bar with 'dremio', 'Datasets', 'Jobs', and a search bar. Below this, the 'Datasets' section is active, showing a list of datasets under the '@admin' user. The datasets listed are SF\_incidents2016, SFIncidents\_2016, zip\_lookup, zip\_lookup\_curated, zips, and zips\_curated. A teal callout bubble points to the 'SFIncidents\_2016' dataset with the text 'Choose to move the VDS ...'. Below the dataset list, there's a 'Move Dataset' dialog box. The 'Name' field is 'SFIncidents\_2016'. The 'Location' dropdown is expanded, showing options: '@admin', '311\_Cases', 'Chicaco-crime-data', and 'dremio'. A teal callout bubble points to the '311\_Cases' option with the text '... then choose the new location'. To the right of the main interface, there's a context menu with options: Query, Edit, Catalog, Explore Data Graph, Analyze With, Remove, Rename, Move, and Copy Path. Below the main interface, there's another 'Move Dataset' dialog box. The 'Name' field is 'SF\_incidents2016\_VDS1'. The 'Location' dropdown is expanded, showing options: '@admin', '311\_Cases', and 'Chicaco-crime-data'. A teal callout bubble points to the '311\_Cases' option with the text 'Beware of dependents'. A warning message at the top of this dialog box states: 'Changing the location of this dataset will disconnect 1 dependent datasets. Make a copy to preserve these connections.'

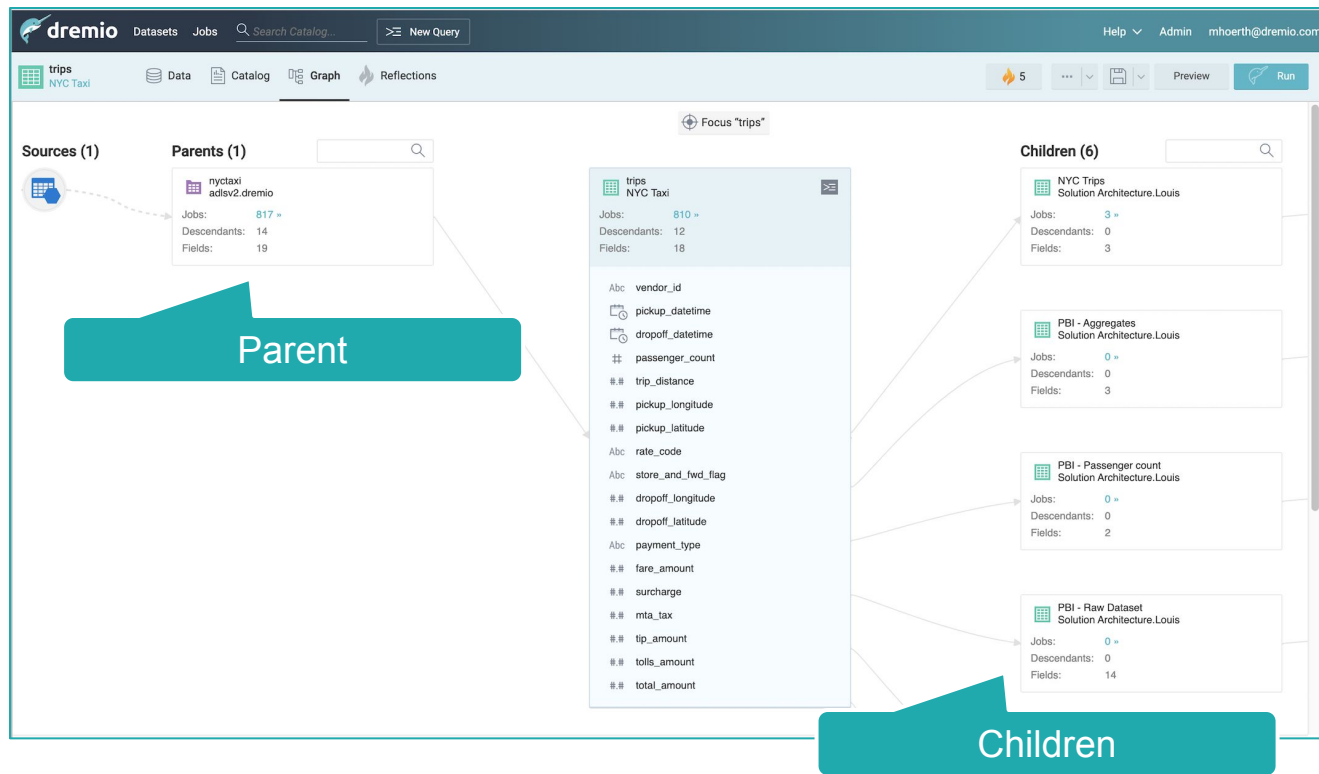
Choose to move the VDS ...

... then choose the new location

Beware of dependents

# Data lineage and searchable catalog

- See all downstream and upstream virtual and physical datasets
- Trace back to the original sources
- Fully navigate the lineage by clicking on a downstream or upstream dataset
- Add tags for additional richness in search



# Caching and querying source metadata

## Caching

- **Dataset Discovery:**  
interval for source DBs and tables. Lightweight.
- **Dataset Details:**  
Refreshes fields, types, files, etc. used for query planning.

## Querying

- Dremio catalogs, schemas, physical datasets, virtual datasets and columns c using **INFORMATION\_SCHEMA** queries.

The screenshot displays the Dremio web interface. On the left, the 'SQL Editor' shows a query: `select * from INFORMATION_SCHEMA."TABLES"`. Below the editor, a table of results is shown, listing various datasets and schemas. A teal callout box points to the query text with the text 'Select \* from INFORMATION\_SCHEMA."TABLES"'. On the right, the 'Edit Source' panel is visible, showing settings for 'Metadata', 'Dataset Handling', and 'Metadata Refresh'. A teal callout box points to the 'Dataset Discovery' section of the 'Metadata Refresh' settings, which includes a 'Fetch every' dropdown set to '1 Hour(s)'. Below the screenshot, two teal callout boxes are present: 'Querying' pointing to the SQL Editor and 'Caching' pointing to the 'Dataset Discovery' settings.

DREMIO	\$scratch	LI
DREMIO	\$scratch	te
DREMIO	Curated Datasets	Pr
DREMIO	Curated Datasets	Qu
DREMIO	exabean	ea
DREMIO	INFORMATION_SCHEMA	Ca
DREMIO	INFORMATION_SCHEMA	Co
DREMIO	INFORMATION_SCHEMA	So
DREMIO	INFORMATION_SCHEMA	Ta
DREMIO	INFORMATION_SCHEMA	Vi
DREMIO	NYC Taxi	te
DREMIO	NYC Taxi	te
DREMIO	paypal	pp
DREMIO	Product	SP
DREMIO	Security	do
DREMIO	Security	Me
DREMIO	Security	Te
DREMIO	sys	bo
DREMIO	sys	de
DREMIO	sys	te

# Using the Dremio UI

A few of the topics



Adding a column



Group By



Join

# The Dremio home screen

The screenshot shows the Dremio home interface. On the left is a dark sidebar with the Dremio logo and a 'Welcome to Dremio, please' message. The main area is light blue and contains a 'Datasets' section. A teal callout labeled 'Home space' points to the 'mhoerth@dremio.com' dataset entry. Another teal callout labeled 'Shared spaces' points to the 'Yelp' dataset entry. A third teal callout labeled 'Data sources' points to the 'adlsv2' dataset entry. A teal callout labeled 'You' points to the user profile 'mhoerth@dremio.com' in the top right header.

**Home space**

**Shared spaces**

**Data sources**

**You**

Name	Jobs	Action
Businesses	25	
Reviews	5	
taxiplus	9	
Tip	1	

Space	Jobs	Shared
Application	1	
Business	4	
Curated Datasets	2	
exabeam	1	
NYC Taxi	3	
paypal		
Preparation		
Product	1	
Security	8	
Solution Architecture	69	
Yelp	4	

Source	Jobs
adlsv2	17
ECS Data Lake	
EMC-ECS (S3 Compatible)	
postgres	8

# Viewing datasets

- View and edit the SQL that defines the dataset
- Run the query or “Preview” 10k rows
- Create new VDS built on this and possibly additional VDS
- Click the record count for a shortcut to the jobs page
- Column count of the dataset on the page
- Tabs to view the catalog and define reflections

The screenshot shows the Dremio SQL Editor interface. At the top, there are tabs for 'Businesses', 'Data', 'Catalog', 'Graph', and 'Reflections'. The 'Businesses' tab is active, showing a SQL query: `SELECT * FROM Businesses`. Below the query editor, there is a table preview with 14 columns: `business_id`, `name`, `address`, `city`, `state`, `postal_code`, `lat`, `lon`, `category`, `rating`, `review_count`, `price`, `hours`, and `phone`. The table contains 10 rows of data. A callout bubble points to the 'Preview or run query' button in the top right. Another callout bubble points to the 'Reference context' button in the bottom right. A third callout bubble points to the record count '100,000' in the bottom right, which is labeled 'Shortcut to jobs page'.

Preview or run query

Reference context

Shortcut to jobs page

business_id	name	address	city	state	postal_code	lat	lon	category	rating	review_count	price	hours	phone
15Wheh84yJXfytoV1LX0AQ	Arizona Biltmore Golf Club	2818 E Camino Acequia Drive	Phoenix	AZ	85016	33.45	-112.07	Golf Course	4.5	52214	15		
QXAEGFB4oINsVuTFxYKfQ	Emerald Chinese Restaurant	30 Eglinton Avenue W	Mississauga	ON	L5V 0B1	43.59	-79.64	Chinese Restaurant	4.0	397	10		
gnKjwl_1w79qi1V3IC_xQ0	Musashi Japanese Restaurant	10110 Johnston Rd, Ste 15	Charlotte	NC	28217	35.19	-80.91	Japanese Restaurant	4.0	925	10		
xvX2CttrVhyGz21dFg_0xw	Farmers Insurance - Paul Lore	15655 W Roosevelt St, Ste 237	Goodyear	AZ	85506	33.45	-112.07	Insurance Agency	4.0	1296	10		
Hhyx0kGAM07SYt1Q4wMFQ	Queen City Plumbing	4209 Stuart Andrew Blvd, Ste	Charlotte	NC	28217	35.19	-80.91	Plumbing	4.0	19001	10		
68dUKd8_8li37in4aw0SEA	The UPS Store	Credit Valley Town Plaza, F2	Mississauga	ON	L5V 0B1	43.59	-79.64	Shipping & Logistics	4.0	59947	10		
5JucpCfHzltJh5r1JabJ0g	Edgeworxx Studio	20 Douglas Woods Drive Southe	Calgary	AB	T2Z 1K4	50.94	-114.07	Art Studio	4.0	94364	10		
gbQn7vr_caG_A1ug5mGhWg	Supercuts	4545 E Tropicana Rd Ste 8, Tri	Las Vegas	NV	89121	36.09	-115.17	Hair Salon	4.0	8996	10		
Y6iyemLX_oyLRpnR38vgMA	Vita Bella Fine Day Spa	5940 W Union Hills Dr	Glendale	AZ	85308	33.65	-112.18	Spa	4.0	65481	10		
4GBVPiYRvzGh4K4TKRQ_rw	Options Salon & Spa	21689 Lorain Rd	Fairview Park	OH	44126	41.44	-80.82	Hair Salon	4.0	2526	10		
fcX0EzDxYeZq03LG10Xmg	Nucleus Information Service	1210 8th Street SW, Unit 220	Calgary	AB	T2R 1L3	51.04	-114.07	Information Service	4.0	1417	10		
1Dfx3zM-rW4n-31Kec8sJg	Taco Bell	2450 E Indian School Rd	Phoenix	AZ	85016	33.45	-112.07	Mexican Restaurant	4.0	49519	10		
5t3KVdMnFgAYm51lwLhmA	The Kilted Buffalo Langtree	119 Landings Dr, Ste 101	Moorestville	NC	28117	35.52	-77.40	Bar & Grill	4.0	988	10		
fweCY18FmbJXHCqLnwuk8w	Marco's Pizza	5981 Andrews Rd	Mentor-on-the-Lake	OH	44060	41.70	-83.50	Pizza Restaurant	4.0	706	10		

# Context vs. fully-qualified names

- PDS and VDS referenced by fully-qualified names, **source\space.folder.dataset\_name**
- Optionally set a context on each reference
- Include the full URI in **FROM <tablename>** references for maximum portability to another space
- Get the full URI from the dataset clipboard icon

The screenshot shows the Dremio web interface. At the top, there's a navigation bar with 'dremio', 'Datasets', 'Jobs', a search bar, and a 'New Query' button. Below this, there's a 'New Query' tab and a 'Data' tab. The 'SQL Editor' is visible on the left with the text '1 select \* from'. A 'Select Context' dialog is open in the center, showing a tree view of contexts. The 'Business' context is expanded, and 'Transportation' is selected. A teal callout bubble 'Push to change' points to the 'Transportation' folder. Below the dialog, the 'Business.Transportation' dataset details are shown. It includes a 'Name' dropdown and a list of datasets: 'cab weather data gj', 'NYC Trips' (with tags 'Gold', 'FHV', 'Cab', 'Taxi'), 'NYC Weather', and 'trips with weather gj'. A teal callout bubble 'Get URI' points to the clipboard icon next to 'trips with weather gj'. A grey box with 'Context: adlsv2' and a pencil icon is also visible.

SQL Editor

```
1 select * from
```

Select Context

Select New Context

- Home
- @mhoerth@dremio.com
- Application
- Business
  - Transportation
- Curated Datasets
  - NYC Taxi
  - Preparation
  - Product

Business.Transportation

Push to change

Context: adlsv2

Business.Transportation

Name

- cab weather data gj
- NYC Trips Gold FHV Cab Taxi
- NYC Weather
- trips with weather gj

Get URI



# Editing VDS

Curate data by applying data transformation functions

- Change type of a column
- Drop and rename columns
- Adding derived columns
- Split columns
- Extract strings from columns
- Handle complex fields (arrays/nested)
- Group by and join datasets

The image displays three screenshots of the Dremio interface, each illustrating a different data transformation function. The first screenshot shows the 'Filter or replace values' dialog, where users can select values to keep or exclude. The second screenshot shows the 'Or, calculate columns' dialog, where users can write a SQL case statement to calculate new columns. The third screenshot shows the 'Or, split values' dialog, where users can split a column into multiple columns.

**Filter or replace values**

Replace Extract Split Keep Only Exclude

Method: Values Pattern Custom Condition

Available Values Select: All | None 2 of 13 selected

Search values...

☐ NONE 48%

☒ ARREST 25%

☒ BOOKED 25%

☐ UNFOUNDED 1.3%

☐ .ILUVENII F BOOKED 0.48%

6686 matched values 6665 unmatched values

Apply Preview Cancel Result based on sample data

Abc	PdDistrict	Abc	Resolution
	SOUTHERN		ARREST
	SOUTHERN		BOOKED
	SOUTHERN		ARREST
	SOUTHERN		BOOKED
	BAYVIEW		ARREST
	BAYVIEW		BOOKED
	SOUTHERN		ARREST
	SOUTHERN		BOOKED
	SOUTHERN		ARREST
	SOUTHERN		BOOKED

**Or, calculate columns**

Add Calculated

```
1 case
2 when "DayOfWeek" = 'Monday' then 1
3 when "DayOfWeek" = 'Tuesday' then 2
4 when "DayOfWeek" = 'Wednesday' then 3
5 when "DayOfWeek" = 'Thursday' then 4
6 when "DayOfWeek" = 'Friday' then 5
```

New Field Name Options

DayOfWeek

Result based on sample data

Abc	DayOfWeek
	Friday
	Friday

**Or, split values**

Replace Extract Split Keep Only Exclude

Edit Selection Fixed String . Ignore Case

25459 matched values 74541 unmatched values

Position First New Field Name Descript Options Drop Source Field (Descript)

Apply Preview Cancel Result based on sample dataset

Abc	IncidentNum	Abc	Descript	Descript (new)	Abc	Day
	120058272		POSS OF PROHIBITED WEAPON	["POSS OF PROHIBITED WEAPON		Friday
	120058272		FIREARM, LOADED, IN VEHIC	["FIREARM"," LOADED, IN VEH		Friday
	141059263		WARRANT ARREST	["WARRANT ARREST"]		Monday
	160013662		LOST PROPERTY	["LOST PROPERTY"]		Tuesda

# Working with complex types

- Reference individual elements in complex type, or flatten
- Extract and unnest complex array types
- Extract from complex struct types
- Keep original column or replace

The screenshot shows the Dremio 'Extract' configuration panel. The 'Path' field is set to 'GoodForKids'. Below the path field, it indicates '34358 matched values' and '65642 unmatched values'. The 'New Field Name' is 'attributes\_1'. The 'Options' section has a checkbox for 'Drop Source Field (attributes)' which is unchecked. The 'Apply' button is highlighted. Below the configuration, a preview table shows the result of the extraction.

	attributes	attributes_1	categories
0	{ "GoodForKids": "True", "Rest": "False" }	False	Golf, Active Life
1	null	null	Specialty Food, Sushi Bars, Restaurants
1	{ "BusinessAcceptsCreditCard": "True", "BusinessParking": { "garage": "None", "street": "None" } }	null	Insurance, Financial Services
1	{ "BusinessAcceptsCreditCard": "True", "BusinessParking": { "garage": "None", "street": "None" } }	null	Plumbing, Shopping Centers
1	{ "BusinessAcceptsCreditCard": "True", "BusinessParking": { "garage": "None", "street": "None" } }	null	Shipping Centers
1	{ "BusinessAcceptsCreditCard": "True", "BusinessParking": { "garage": "None", "street": "None" } }	null	Beauty & Spas, Hair Salons, Hair Stylists
0	{ "BusinessAcceptsCreditCard": "True", "BusinessParking": { "garage": "None", "street": "None" } }	null	Nail Salons, Beauty & Spas, Nails
1	{ "BusinessAcceptsCreditCard": "True", "BusinessParking": { "garage": "None", "street": "None" } }	null	Beauty & Spas, Nails
1	{ "BusinessAcceptsCreditCard": "True", "BusinessParking": { "garage": "None", "street": "None" } }	null	Local Services, Other

A teal callout box with a struct icon and the text 'Struct type' points to the 'attributes' column in the preview table.

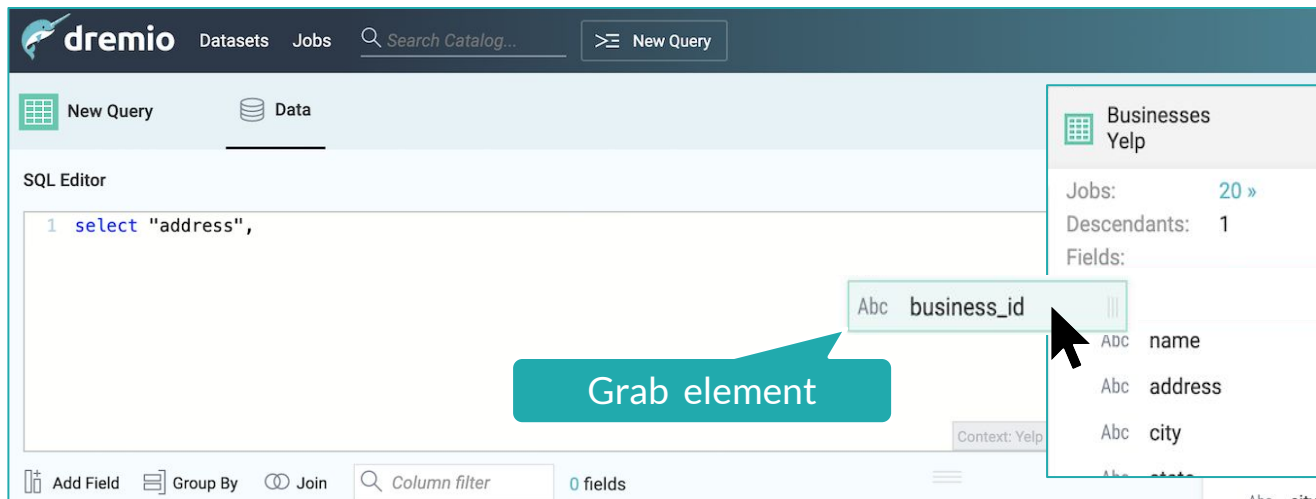
The screenshot shows the Dremio 'Extract' configuration panel. The 'Element' dropdown is set to 'd'. Below the configuration, it indicates '100000 matched values' and '0 unmatched values'. The 'New Field Name' is 'Descript\_1'. The 'Options' section has a checkbox for 'Drop Source Field (Descript)' which is unchecked. The 'Apply' button is highlighted. Below the configuration, a preview table shows the result of the extraction.

Descript	Descript_1	DayOfWeek
["POSS OF PROHIBITED WEAPON"]	POSS OF PROHIBITED WEAPON	Friday
["LOST PROPERTY"]	LOST PROPERTY	Friday
["BATTERY"]	BATTERY	Friday
["PAROLE VIOLATION"]	PAROLE VIOLATION	Saturday
["FIRE REPORT"]	FIRE REPORT	Saturday
["WARRANT ARREST"]	WARRANT ARREST	Saturday
["FOUND PERSON"]	FOUND PERSON	Friday
["AIDED CASE", "MENTAL DISTURBANCE"]	AIDED CASE	Saturday
["RESISTING ARREST"]	RESISTING ARREST	Monday
["AGGRAVATED ASSAULT WITH A DEADLY WEAPON"]	AGGRAVATED ASSAULT WITH A DEADLY WEAPON	Thursday

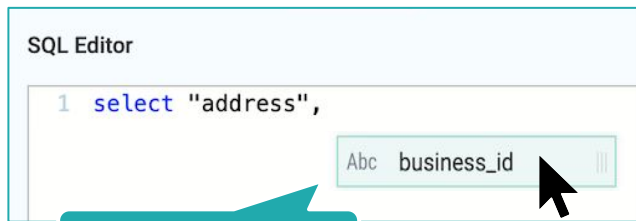
A teal callout box with an array icon and the text 'Array type' points to the 'Descript' column in the preview table.

# Drag and drop

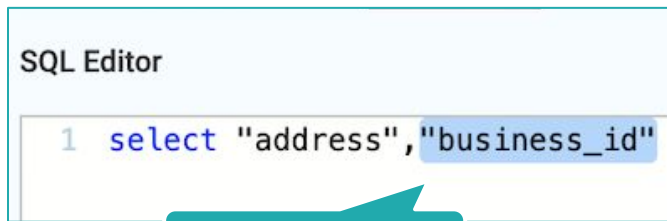
- Easily create selects, joins and “group by” queries with drag and drop
- Recommended joins based on query history and schema
- Group by dimensions and measures identified for easy query building



The screenshot shows the Dremio interface with a dark header bar containing the logo, navigation tabs (Datasets, Jobs), a search bar, and a 'New Query' button. Below the header, there are two tabs: 'New Query' (active) and 'Data'. The main area is the 'SQL Editor' with a text input containing '1 select "address",'. To the right, a sidebar shows a dataset named 'Businesses Yelp' with statistics (Jobs: 20, Descendants: 1) and a list of fields: 'business\_id', 'name', 'address', and 'city'. A teal callout bubble labeled 'Grab element' points to the 'business\_id' field in the sidebar.



This screenshot shows the 'business\_id' field from the sidebar being dragged into the SQL Editor. A teal callout bubble labeled 'Slide it' points to the field as it is being moved.



This screenshot shows the final state of the SQL Editor with the completed query: '1 select "address", "business\_id"'. A teal callout bubble labeled 'Done' points to the query text.

# VDS version tracking

- Every change to a virtual dataset is tracked
- You can see what the change was and who made it
- Go back to a particular version

The screenshot shows the Dremio VDS interface. At the top, there's a header with 'Help', 'Admin', and 'mhoerth@dremio.com'. Below that, a toolbar contains icons for a flame (0), a dropdown menu, a save icon, another dropdown, 'Preview', and a 'Run' button. The main area displays a SQL query with syntax highlighting. A teal callout bubble with the text 'Change tracker' points to the right sidebar. At the bottom, a status bar shows 'Job: Preview ⓘ | Records: 100 | Time: <1s'. Below the status bar is a table with columns 'area\_code', 'hire\_date', and 'job\_id'. The first row of data shows '515', '1996-02-17', and 'MK\_MAN'.

```
email, nested_0.phone_number AS phone_number, nested_0.area_code AS area_code,  
ed_0.commission_pct AS commission_pct, nested_0.manager_id AS man  
join_DEPARTMENTS.DEPARTMENT_NAME AS DEPARTMENT_NAME, join_DEPART  
  
, 'INDEX') AS area_code, hire_date, job_id, salary, commission_pct, manager_id,  
  
id = join_DEPARTMENTS.DEPARTMENT_ID
```

area_code	hire_date	job_id
515	1996-02-17	MK_MAN

# Export data using the clipboard

- Preview the dataset or run the query
- Easily export to other tools such as spreadsheets
- Start BI tools such as Tableau from the Dremio UI

The screenshot displays the Dremio web interface. At the top, there's a navigation bar with 'dremio', 'Datasets', 'Jobs', and a search bar. Below this, the 'Curated Employees' dataset is selected. The 'SQL Editor' tab is active, showing a complex SQL query that joins employee data with department information. A teal callout bubble with the text 'Save data to clipboard' points to the 'Run' button. Below the editor, a table preview shows the first few rows of the query results. At the bottom, an 'Untitled spreadsheet' is open, showing the same data in a tabular format with columns for employee details and department information.

**SQL Query:**

```
1 SELECT nested_0.first_name AS first_name, nested_0.last_name AS last_name, nested_0.email AS email, nested_0.phone_number AS phone_number, nested_0.area_code AS area_code,
2 nested_0.hire_date AS hire_date, nested_0.job_id AS job_id, nested_0.salary AS salary, nested_0.commission_pct AS commission_pct, nested_0.manager_id AS manager_id,
3 nested_0.department_id AS department_id, join_DEPARTMENTS.DEPARTMENT_ID AS DEPARTMENT_ID_0, join_DEPARTMENTS.DEPARTMENT_NAME AS DEPARTMENT_NAME, join_DEPARTMENTS.MANAGER_ID AS
4 MANAGER_ID_0, join_DEPARTMENTS.LOCATION_ID AS LOCATION_ID
5 FROM (
6 SELECT first_name, last_name, email, phone_number, extract_pattern(phone_number, '\d+', 0, 'INDEX') AS area_code, hire_date, job_id, salary, commission_pct, manager_id,
7 department_id
8 FROM postgres.public.employees
9 WHERE '1993-02-01' <= hire_date
10 ) nested_0
11 INNER JOIN sqlserver.employee.dbo.DEPARTMENTS AS join_DEPARTMENTS ON nested_0.department_id = join_DEPARTMENTS.DEPARTMENT_ID
```

**Table Preview:**

first_name	last_name	email	phone_number	area_code	hire_date	job_id
Michael	Hartstein	MHARTSTE	515.123.5555	515	1996-02-17	MK_MAN
Pat	Fay	PFAY	603.123.6666	603	1997-08-17	MK_REP
Shelley	Higgins	SHIGGINS	515.123.8080	515	1994-06-07	AC_MGR
William	Gietz	WGIEZT	515.123.8181	515	1994-06-07	AC_ACCOUNT

**Spreadsheet Data:**

	A	B	C	D	E	F	G	H	I	J
1	first_name	last_name	email	phone_number	area_code	hire_date	job_id	salary	commission_pct	manager_id
2	Michael	Hartstein	MHARTSTE	515.123.5555	515	1996-02-17	MK_MAN	13000		100
3	Pat	Fay	PFAY	603.123.6666	603	1997-08-17	MK_REP	6000		201
4	Shelley	Higgins	SHIGGINS	515.123.8080	515	1994-06-07	AC_MGR	12000		101
5	William	Gietz	WGIEZT	515.123.8181	515	1994-06-07	AC_ACCOUNT	8300		205
6	John	Russell	JRUSSEL	011.44.1344.429	11	1996-10-01	SA_MAN	14000	0.4	100
7	Charles	Johnson	CJOHNSON	011.44.1644.429	11	2000-01-04	SA_REP	6200	0.1	149

# Creating tables as select (CTAS)

- Tables created in file system source types or in \$scratch
- Parquet format
- \$scratch is readable and writable by all users.
- Must first enable exports for the source. If enabled, all users/groups who have access to the source can leverage CTAS and DROP TABLE on sources.

The Dremio SQL Editor interface shows a new query being executed. The SQL code is:

```
1 create table "mhoerth-sf"."binky-bucket".test.mytable
2 as select * from "@admin"."SF_incidents2016"
```

An annotation bubble points to the query with the text "CTAS to S3". Below the query editor, the results are displayed in a table format. The table has columns: Fragment, Records, Path, and Metadata. The first row shows a fragment of 1\_0, 150491 records, and a path of /binky-bucket/test/mytable/1\_0\_0.p.

Fragment	Records	Path	Metadata
1_0	150491	/binky-bucket/test/mytable/1_0_0.p	null

An annotation bubble points to the results table with the text "UI confirms result".

The AWS S3 console shows the bucket "binky-bucket" and the folder "test". The file "1\_0\_0.parquet" is visible in the list. An annotation bubble points to the file with the text "Visible on S3 console".

Name	Last modified	Size	Storage class
1_0_0.parquet	Aug 29, 2019 10:01:34 PM GMT-0700	10.7 MB	Standard

# Dremio ANSI SQL

Including



Grammar

0101  
1101  
001

Types



Functions

# Example differences in grammar

## Hive

String constants quoted with non ANSI SQL double quotes

- `SELECT CONCAT (col1, "xyz")`
- `SELECT REGEXP_REPLACE (col1, "x", "y")`
- `CASE WHEN acquiring_portfolio = "chase" ...`

Otherwise invalid column names are escaped (``) in Hive but double-quoted in Dremio (")

- `SELECT ``DATE``, ``HOUR``, cast(``VALUE`` as INT)`
- `SELECT sum(sales) as ``2018_sales```

## Dremio

- `SELECT CONCAT (col1, 'xyz')`
- `SELECT REGEXP_REPLACE(col1, 'x', 'y')`
- `CASE WHEN acquiring_portfolio = 'chase' ...`

- `SELECT "DATE", "HOUR", cast("VALUE" as INT)`
- `SELECT sum(sales) as "2018_sales"`



# Implicit casting example of type difference

- Hive supports implicit casting in SELECT/ WHERE, JOIN, UNION etc clauses.
- Dremio requires explicit casting if the columns being compared, joined, unioned are of different type family (string<->numeric, string<->date...).
- Hive supports untyped NULLs while Dremio requires explicit casting

Hive	Dremio
<pre>-- Mixed types SELECT Coalesce (string_col, 1)</pre>	<pre>SELECT coalesce   (cast(string_col as INT) ,1) --or -- SELECT Coalesce(string_col,'1')</pre>
<pre>-- Mixed types SELECT sum(int_col * string_col)</pre>	<pre>SELECT sum(int_col * cast (string_col as int))</pre>
<pre>-- JOIN keys mixed types FROM table1 A INNER JOIN table2 B ON a.str_col=b.int_col</pre>	<pre>FROM table1 A INNER JOIN table2 B ON a.str_col=     cast(b.int_col as varchar)</pre>
<pre>-- Implicit NULL cast SELECT col1, NULL as col2, col3.....</pre>	<pre>SELECT col1, CAST(NULL as varchar) as col2, col3.....</pre>

# Example function differences

- Some function unique to other platforms may not exist on Dremio or differ in their input arguments

Hive	Dremio
<ul style="list-style-type: none"><li>TO_DATE(string_col)</li></ul>	<ul style="list-style-type: none"><li>TO_DATE(string_col, 'yyyy-mm-dd')</li><li>CAST(string_col as DATE)</li></ul>
<ul style="list-style-type: none"><li>SPLIT (col1, '_' ) [0]</li></ul>	<ul style="list-style-type: none"><li>SPLIT_PART (col1, '_' , 1)</li></ul>
<ul style="list-style-type: none"><li>Date_format (string, format)</li></ul>	<ul style="list-style-type: none"><li>TO_CHAR(CAST(string as DATE), format) -- no Date_format</li></ul>
<ul style="list-style-type: none"><li>CAST (col1 as String)</li></ul>	<ul style="list-style-type: none"><li>CAST (col1 as Varchar)</li></ul>
<ul style="list-style-type: none"><li>DAY (col1) function</li></ul>	<ul style="list-style-type: none"><li>EXTRACT(DAY from col1) -- no Day()</li></ul>
<ul style="list-style-type: none"><li>PMOD (col1)</li></ul>	<ul style="list-style-type: none"><li>ABS(MOD(col1)) -- no PMOD ()</li></ul>
<ul style="list-style-type: none"><li>SELECT current_time stamp()</li></ul>	<ul style="list-style-type: none"><li>SELECT current_timestamp -- w/no ()</li></ul>
<ul style="list-style-type: none"><li>MAP, LATERAL VIEW EXPLODE functions</li></ul>	<ul style="list-style-type: none"><li>Not supported</li></ul>
<ul style="list-style-type: none"><li>LEFT SEMI JOIN</li></ul>	<ul style="list-style-type: none"><li>Not supported</li></ul>

# Jobs in Dremio

Including



Job status



Accelerated queries



Profiles

# Jobs page

- All jobs are tracked
- Jobs can be filtered by
  - Status
  - Time of execution
  - Type
  - User
- Each user can see their own jobs, admin sees all jobs


The screenshot displays the Dremio Jobs page. At the top, there are tabs for 'Datasets', 'Jobs', and a search bar. Below the tabs, there are filters for 'Start Time', 'Status', 'UI, External Tools', 'Queue', and a 'Contains text...' search field. A table lists several jobs with columns for Dataset, User, Start Time, Duration, and End Time. A callout 'Overview tab' points to the 'Overview' tab in the job details panel. Another callout 'Job queue' points to the 'Queue' field in the job details. A third callout 'Original SQL' points to the 'SQL' section in the job details. The job details panel shows the following information:

- Completed** (with a flame icon)
- Overview** (selected), Details, Acceleration, Profile
- Summary** (with a mouse cursor icon):
  - Query Type: ODBC Client (execute prepared statement)
  - Duration: <1s
  - Start Time: 03/14/2020 14:59:06
  - End Time: 03/14/2020 14:59:07
  - User: gabe@dremio.com
  - Queue: Low Cost User Queries
  - Job ID: 2192ac55-4ce6-9f58-b554-bb749309df00
- Query**
- Parents**: taxiDell @gabe@dremio.com
- Accelerated By**: Aggregation Reflection @gabe@dremio.com.taxiDell
- Input**: Input Bytes: 2.42 MB, Input Records: 78,020
- Output**: Output Bytes: 1351, Output Records: 6
- SQL**:

```
1 SELECT AVG("taxiDell"."tip_amount") AS "avg_tip_amount_ok",
2   SUM(1) AS "sum_Number_of_Records_ok",
3   CAST({fn TRUNCATE(EXTRACT(YEAR FROM "taxiDell"."pickup_date"),0)} AS INTEGER) AS "yr_pickup_date_ok",
4 FROM "@gabe@dremio.com"."taxiDell" "taxiDell"
5 GROUP BY CAST({fn TRUNCATE(EXTRACT(YEAR FROM "taxiDell"."pickup_date"),0)} AS INTEGER)
```

# Introduction to job queues

- Runs each job in a queue with specific characteristics (such as memory limits, CPU priority, and queueing and runtime timeouts)
- Administrator can define rules that specify which query is assigned to which queue.
- Provides workload isolation and predictability for users and groups.

Overview	Details	Acceleration	Profile
<b>Summary</b>			
Query Type:	UI (preview)		
Duration:	<1s		
Start Time:	01/19/2020 17:43:11		
End Time:	01/19/2020 17:43:11		
User:	mhoerth@dremio.com		
Queue:	UI Previews		
Job ID:	21dafa4f-b655-8d33-cf8f-fdfdb4aa1200 		

Queue used to execute job

Queues			Rules		
Name ^	CPU Priority	Concurrency Limit	Order	Name	Rule
Critical	Critical	-	1	UI Previews	query_type() = 'UI Pre
High Cost Reflections	Background	1	2	High Cost Reflections	query_type() = 'Reflex
High Cost User Queries	Medium	10	3	Low Cost Reflections	query_type() = 'Reflex
Low Cost Reflections	Background	10	4	High Cost User Queries	query_cost() >= 3000
Low Cost User Queries	Medium	100	5	Low Cost User Queries	query_cost() < 30000
UI Previews	Critical	100	6	All Other Queries	All other queries

Queues and rules on admin page

# Job details

- Provides the query planning and execution time, and the amount of CPU used during data reading and writing phases.
- Lists the number of records returned and the data volume.
- Includes information on acceleration

The screenshot shows the Dremio interface with the 'Jobs' tab selected. A table lists several jobs with columns for Dataset, User, Start Time, Duration, and End Time. The first job is highlighted. To the right, the 'Details' tab is active, showing the query plan, read operations, and resource usage. Callouts highlight the 'Details tab', 'Resource use', and 'Records returned' sections.

**Jobs Table:**

Dataset	User	Start Time	Duration	End Time
taxiDell	gabe@dremio.com	03/14/2020 14:59:06	<1s	03/14/2020 14:59:0
taxiDell	gabe@dremio.com	03/14/2020 14:59:06	<1s	03/14/2020 14:59:0
taxiDell	gabe@dremio.com	03/14/2020 14:58:59	<1s	03/14/2020 14:58:5
taxiDell	gabe@dremio.com	03/14/2020 14:58:58	<1s	03/14/2020 14:58:5
taxitrips	EMC-ECS	03/14/2020 14:56:23	1s	03/14/2020 14:56:2
taxiDell	gabe@dremio.com	03/14/2020 14:44:10	<1s	03/14/2020 14:44:1
taxiDell	gabe@dremio.com	03/14/2020 14:44:09	<1s	03/14/2020 14:44:1
taxiDell	gabe@dremio.com	03/14/2020 14:43:57	<1s	03/14/2020 14:43:5
taxiDell	gabe@dremio.com	03/14/2020	<1s	03/14/2020 14:43:5

**Details tab:**

**Overview** | **Details** | Acceleration | Profile

**Plan**

Planning Time: <1s  
Enqueued Time: <1s

**Read**

(1) "\_accelerator"."eff44b9c-e89e-47ba-9164-8b65dec44137"

**Process**

Execution Time: <1s

Reading: 42% CPU  
Aggregate: 26% CPU  
Project: 26% CPU  
Client: <0.1% CPU  
Complex\_to\_JSON: <0.1% CPU

**Return**

Wait on Client: <1s  
Number of Records: 6  
Data Volume: 135

# Job acceleration

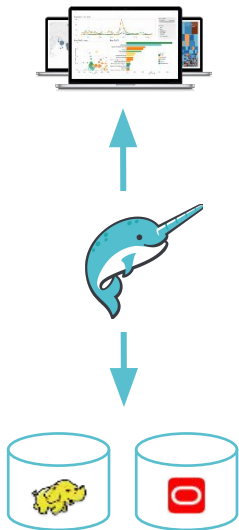
- The jobs page graphically shows queries which have been accelerated by a reflection.
- For each job, the page shows the reflections chosen, not chosen.
- Query profile provides additional information about the query. The profile can be viewed in the UI or downloaded and sent to Dremio.

The screenshot displays the Dremio interface. At the top, there's a navigation bar with 'dremio', 'Datasets', 'Jobs', and a search bar. Below this, the 'Jobs' page is shown with a table of job entries. A teal callout bubble labeled 'Accelerated query' points to a job entry that has a flame icon. The job entry shows details like 'Dataset', 'User', 'Start Time', and 'Duration'. To the right, a detailed view of a job is shown. A teal callout bubble labeled 'Acceleration tab' points to the 'Acceleration' tab in the job details view. The 'Acceleration' tab shows a 'Summary' section with the text 'Query was accelerated.' and an 'Accelerated By' section listing 'Raw Reflection' and 'NYC Taxi.trips'. Below this, a section titled 'Reflections Not Chosen' lists 'Pickup\_Dropoff', 'Aggregation Reflection', and 'Looker\_Map'. A teal callout bubble labeled 'Reflections chosen and not' points to this section.

Dataset	User	Start Time	Duration	Time
old trips km...	jason@dremio.com	03/13/2020 18:13:57	<1s	03/13/2020 18:13:57
old trips km...	jason@dremio.com	03/13/2020 18:13:51	<1s	03/13/2020 18:13:51
old trips km...	jason@dremio.com	03/13/2020 18:13:51	<1s	03/13/2020 18:13:51
old trips km...	jason@dremio.com	03/13/2020 18:13:51	<1s	03/13/2020 18:13:51
Catalog	jason@dremio.com	03/13/2020 18:13:50	<1s	03/13/2020 18:13:50
Unavailable	jason@dremio.com	03/13/2020 18:13:50	<1s	03/13/2020 18:13:50
old trips km...	jason@dremio.com	03/13/2020 18:13:47	<1s	03/13/2020 18:13:47
old trips km...	jason@dremio.com	03/13/2020 18:13:47	<1s	03/13/2020 18:13:47

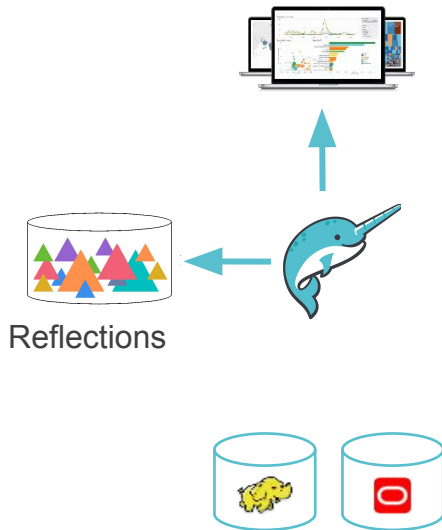
# How reflections work

## No reflections



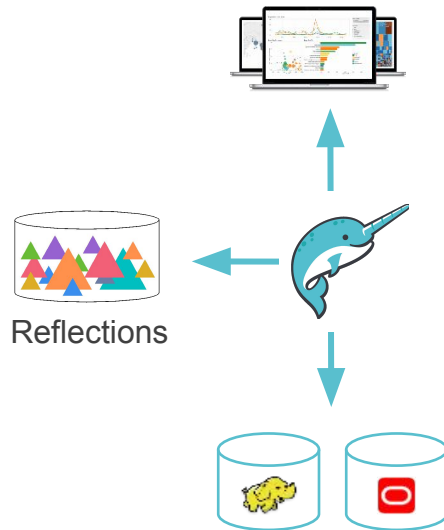
Query push-down

## Acceleration



Query satisfied by reflection

## Partial acceleration



Part of the query accelerated,  
remainder pushed down



# Job profile

- Shows when a thread is running, blocked and waiting for data, or sleeping. A thread is usually sleeping if the thread is ready to run but another thread is currently using the CPU.
- Shows details of query planning (logical and physical)
- Shows how reflections were evaluated and resulting matching
- Topic of separate module

The screenshot shows the Dremio interface with the 'Jobs' tab selected. A job is listed with a status of 'Completed'. A callout box labeled 'Profile tab' points to the 'Profile' tab in the job's detail view. The 'Job Profile' section is expanded, showing the 'Query and Planning' tab. The query text is visible, along with the 'Job Summary' and 'Threads' sections. The 'Threads' section includes a table with thread details.

**Job Profile**

**Query and Planning**

Query Visualized Plan Planning Acceleration

```
SELECT AVG("taxiDell"."tip_amount") AS "avg_tip_amount_ok",  
SUM(1) AS "sum_Number_of_Records_ok",  
CAST({fn TRUNCATE(EXTRACT(YEAR FROM "taxiDell"."pickup_date"),0)} AS INTEGER) AS "yr_pickup_date_ok"  
FROM "gagabedremio.com"."taxiDell"."taxiDell"  
GROUP BY CAST({fn TRUNCATE(EXTRACT(YEAR FROM "taxiDell"."pickup_date"),0)} AS INTEGER)
```

**Job Summary**

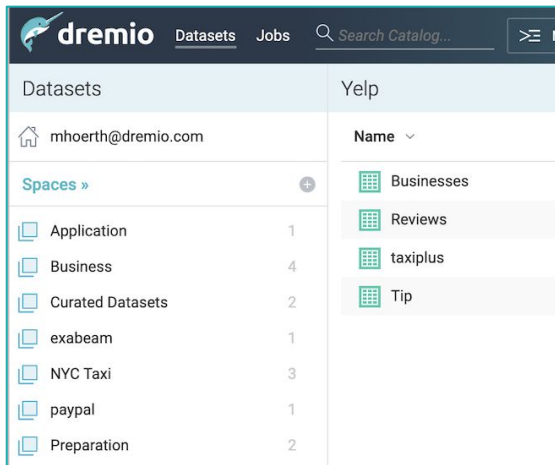
State: COMPLETED  
Coordinator: dremio-master-0.dremio-cluster-pod.default.svc.cluster.local  
Threads: 1  
Command Pool Wait: 1ms  
Planning Time: 3ms  
Resource Scheduling Time: 18ms

**Threads**

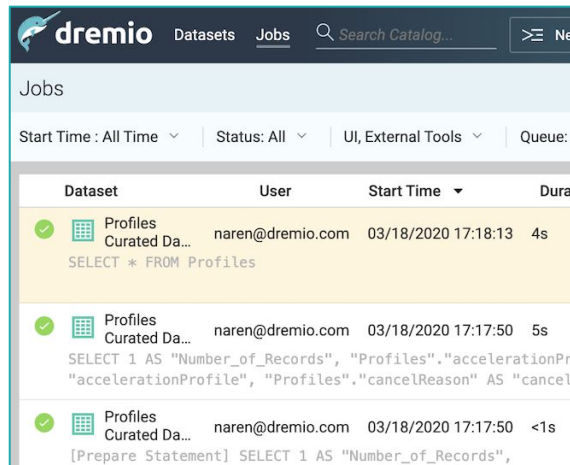
Overview

Phase	Thread Reporting	First Start	Last Start	First End	Last End	Min First-run	Avg First-run	Max First-run	Min Wall-clock	Avg Wall-clock	Max Wall-clock	Min Sleep	Avg Sleep	Max Sleep	Min Blocked
00-xx-xx	1 / 1	0.022s	0.022s	0.218s	0.218s	0.001s	0.001s	0.001s	0.196s	0.196s	0.196s	0.000s	0.000s	0.000s	0.170s

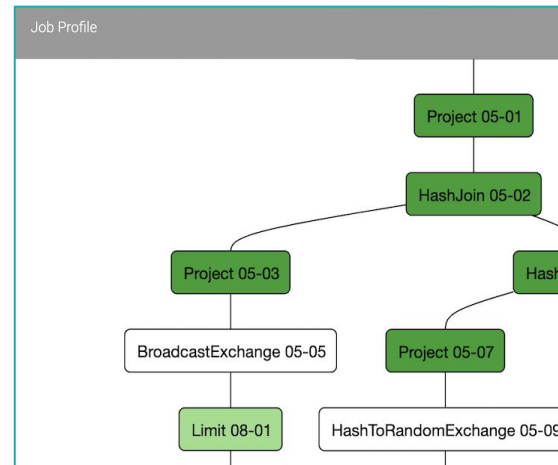
# Demo of Dremio



The screenshot shows the Dremio Datasets page. The top navigation bar includes the Dremio logo, 'Datasets', 'Jobs', and a search bar labeled 'Search Catalog...'. The main content area is divided into two sections: 'Datasets' on the left and 'Yelp' on the right. The 'Datasets' section lists various datasets with icons and counts: Application (1), Business (4), Curated Datasets (2), exabeam (1), NYC Taxi (3), paypal (1), and Preparation (2). The 'Yelp' section shows a table with columns 'Name' and 'Count', listing 'Businesses', 'Reviews', 'taxiplus', and 'Tip'.



The screenshot shows the Dremio Jobs page. The top navigation bar includes the Dremio logo, 'Datasets', 'Jobs', and a search bar labeled 'Search Catalog...'. The main content area is divided into two sections: 'Jobs' on the left and a table on the right. The 'Jobs' section shows filters for 'Start Time', 'Status', 'UI, External Tools', and 'Queue'. The table on the right lists jobs with columns 'Dataset', 'User', 'Start Time', and 'Duration'. The jobs are: 1. 'Profiles Curated Da...' by 'naren@dremio.com' at '03/18/2020 17:18:13' with a duration of '4s'. 2. 'Profiles Curated Da...' by 'naren@dremio.com' at '03/18/2020 17:17:50' with a duration of '5s'. 3. 'Profiles Curated Da...' by 'naren@dremio.com' at '03/18/2020 17:17:50' with a duration of '<1s'.



# ODBC, JDBC, and REST clients

Including



ODBC



JDBC



REST API

# Client integration



## BI tools

- Tableau
- Looker
- PowerBI
- Spotfire

SQL clients

- DBeaver
- DBVisualizer
- Jupyter
- R Studio
- Alteryx Designer

## DBeaver on Mac

## DBVisualizer on Windows



# Some tools integrate the driver

**Download**

Universal Database Tool

Free multi-platform database tool for developers, database administrators, analysts and all people who need to work with databases. Supports all popular databases: MySQL, PostgreSQL, SQLite, Oracle, DB2, SQL Server, Sybase, MS Access, Teradata, Firebird, Apache Hive, Phoenix, Presto, etc.

**Download**

**Choose Dremio**

Connection settings

Database connection settings.

Connection settings

- Initialization
- Shell Commands
- Client identification
- Transactions

General

- Metadata
- Errors and timeouts

Result Sets

- Editors
- Data Formatting
- Presentation

SQL Editor

- SQL Processing

**Provide Dremio host info**

JDBC URL: jdbc:dremio:direct=18.191.140.207:31010

Host: 18.191.140.207 Port: 31010

Database/Schema:

User name: admin

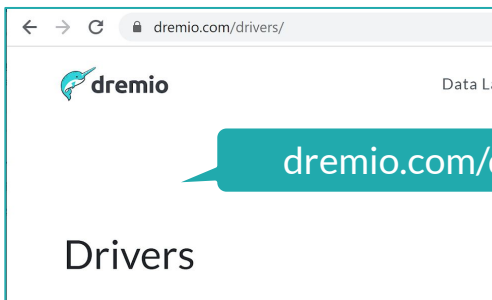
Password: .....

☒ Save password locally

**Use data**

Grid	IncidentNum	Category	Descrpt	DayOfWeek	Value
1	120058272	WEAPON LAWS	POSS OF PROHIBITED WEAPON	Friday	201
2	120058272	WEAPON LAWS	FIREARM, LOADED, IN VEHICLE, POSSESSION O	Friday	201
3	141059263	WARRANTS	WARRANT ARREST	Monday	201
4	160013662	NON-CRIMINAL	LOST PROPERTY	Tuesday	201
5	160002740	NON-CRIMINAL	LOST PROPERTY	Friday	201
6	160002869	ASSAULT	BATTERY	Friday	201
7	160003130	OTHER OFFENSES	PAROLE VIOLATION	Saturday	201
8	160003259	NON-CRIMINAL	FIRE REPORT	Saturday	201
9	160003970	WARRANTS	WARRANT ARREST	Saturday	201
10	160003641	MISSING PERSON	FOUND PERSON	Friday	201
11	160004053	NON-CRIMINAL	AIDED CASE, MENTAL DISTURBED	Saturday	201
12	160073014	OTHER OFFENSES	RESISTING ARREST	Monday	201
13	140776777	ASSAULT	AGGRAVATED ASSAULT WITH A GUN	Thursday	201
14	160004069	BURGLARY	BURGLARY, STORE UNDER CONSTRUCTION, FO	Saturday	201
15	160004150	STOLEN PROPERTY	STOLEN CHECKS, POSSESSION	Saturday	201
16	160004241	ROBBERY	ROBBERY, ARMED WITH A KNIFE	Saturday	201
17	160004558	ASSAULT	BATTERY WITH SERIOUS INJURIES	Saturday	201
18	160004655	ASSAULT			
19	160004637	LARCENY/THEFT			
20	160005421	FRAUD			
21	160005750	WARRANTS			
22	160005841	DRUG/N			
23	160005841	OTHER OFFENSES			
24	160006071	ASSAULT	THREATS AGAINST LIFE	Sunday	201
25	160006598	OTHER OFFENSES	VIOLATION OF STAY AWAY ORDER	Sunday	201
26	160006770	OTHER OFFENSES	LOST/STOLEN LICENSE PLATE	Sunday	201
27	160006786	BURGLARY	BURGLARY, VEHICLE (ARREST MADE)	Sunday	201
28	160006952	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Sunday	201
29	160007057	NON-CRIMINAL	AIDED CASE	Sunday	201
30	160007110	VEHICLE THEFT	STOLEN TRUCK	Sunday	201
31	160007132	VEHICLE THEFT	STOLEN AUTOMOBILE	Sunday	201

# For others, download driver from dremio.com



dremio.com/drivers

## Drivers

Download the JDBC or

ODBC Driver

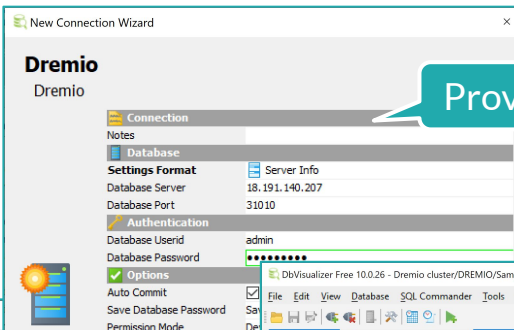
Windows and Mac ODBC

Download for Windows

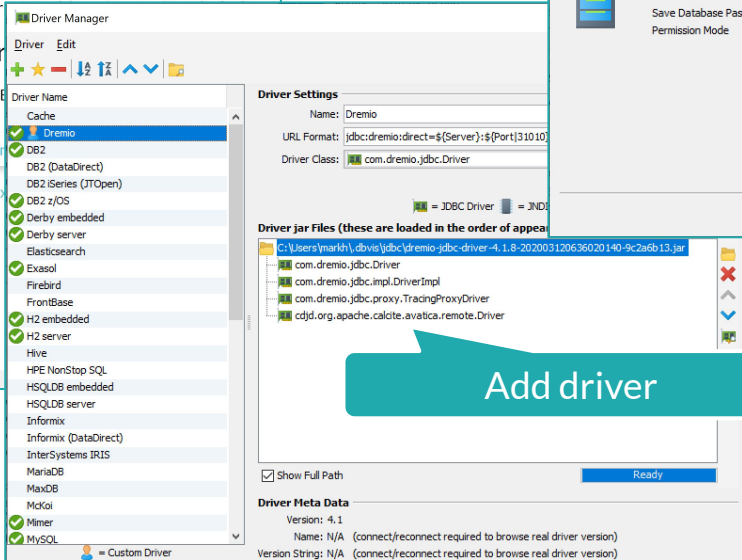
Windows 32 bit | Linux

JDBC Driver

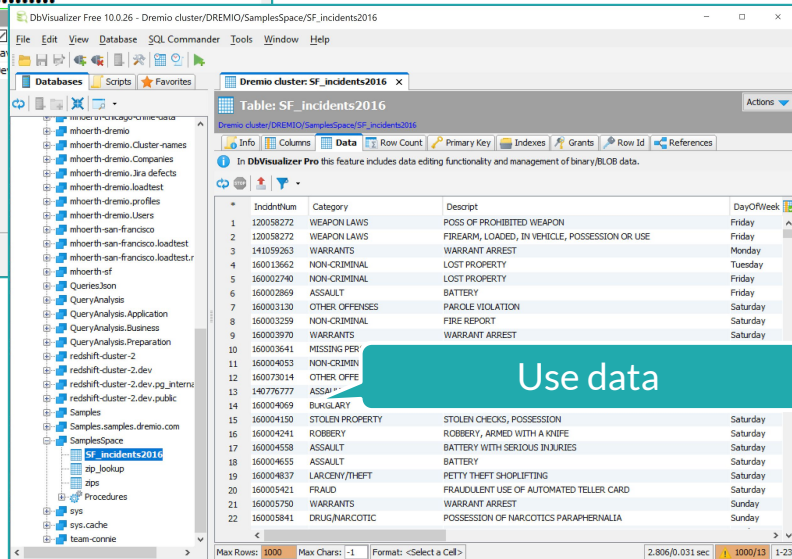
Download



Provide Dremio host info



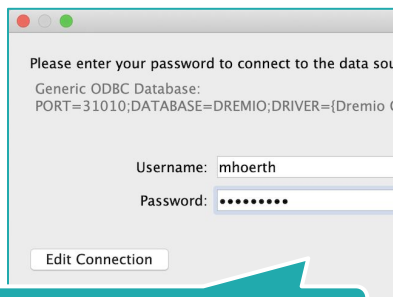
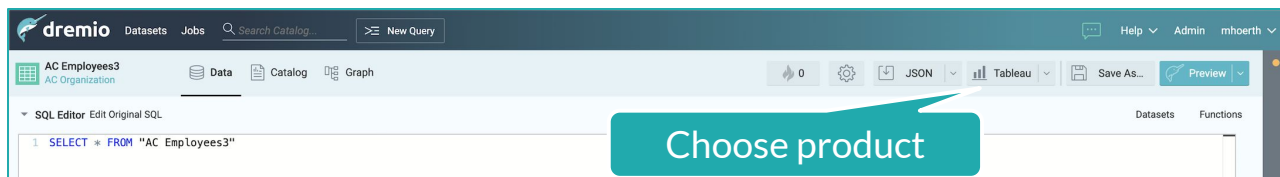
Add driver



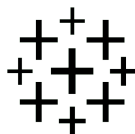
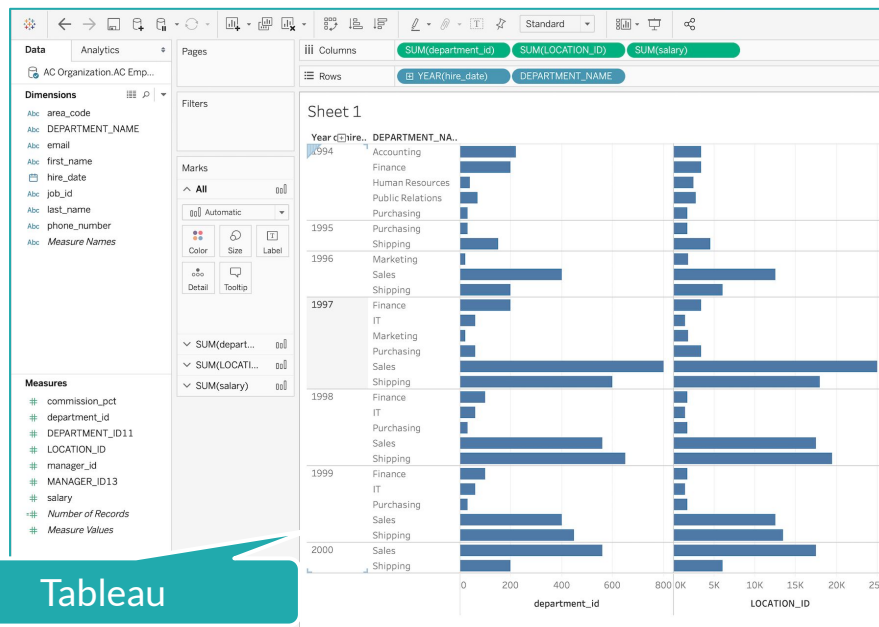
Use data

# Launch BI tools directly or from the UI

- Open Physical & Virtual Datasets in BI tools straight from the Dremio UI
- Select BI tool installed on client machine
- Clicking will download connection file



Dremio credentials



Tableau

# Dremio REST API

Many potential uses

- Integrate ETL processes
- Interact with sources, spaces and datasets.
- Trigger reflection refresh
- Submit queries
- Manage cluster resources and queues



**POST**

```
{ "userName": "mhoerth",  
  "password": "binky" }
```

```
{ "token":  
  "tokenstring",  
  ... }
```

**GET/POST/PUT/DELETE**

```
Authorization: _dremio tokenstring  
...
```



# REST API overview

All API URLs have the following base URL:

`{DREMIO_ORIGIN}/api/v3`

Versions prior to `v3` are considered internal and subject to change without version bumps.

Curly braces (`{ }`) are used to indicate sections of URLs where caller provides a value

*Legend of REST API calls:*

GET	Retrieve information about the resource
POST	Add a resource
PUT	Update a resource
DELETE	Delete a resource

## Account

GET	/user/{id}	Get a user by id
GET	/group/{id}	Get a group by id

## User

GET	/user/{id}/token	Get all user access tokens
POST	/usr/{id}/token	Create access token
DELETE	/user/{id}/token	Delete all access token
DELETE	/user/{id}/token/{tid}	Delete this access token

## Tokens

DELETE	/token	Delete all tokens
DELETE	/token/{tid}	Delete this token

# REST API overview

## Catalog

GET	/catalog	List all top-level containers
GET	/catalog/{id}	A specific source, space, folder, file or dataset
GET	/catalog/by-path/{path}	Get using the path
GET	/catalog/{id}/collaboration	Get tags and wiki content
POST	/catalog	Create a new catalog entry
POST	/catalog/{id}	Promote a file to a PDS
POST	/catalog/{id}/refresh	Refresh all reflections on PDS
POST	/catalog/{id}/collaboration	Create tags and wiki content
PUT	/catalog	Update existing dataset metadata
DELETE	/catalog	Delete catalog entry

## Jobs

GET	/job	Get job information
POST	/job	Post job information

## SQL

POST	/sql	Submit an SQL query
------	------	---------------------

## Reflections

GET	/reflection	Retrieve all reflections
GET	/reflection/{id}	Retrieve specific reflection
GET	/reflection/summary	Retrieve summary of all reflections
POST	/reflection	Create reflection
PUT	/reflection	Update specific reflection
DELETE	/reflection	Delete specific reflection

## Workloads

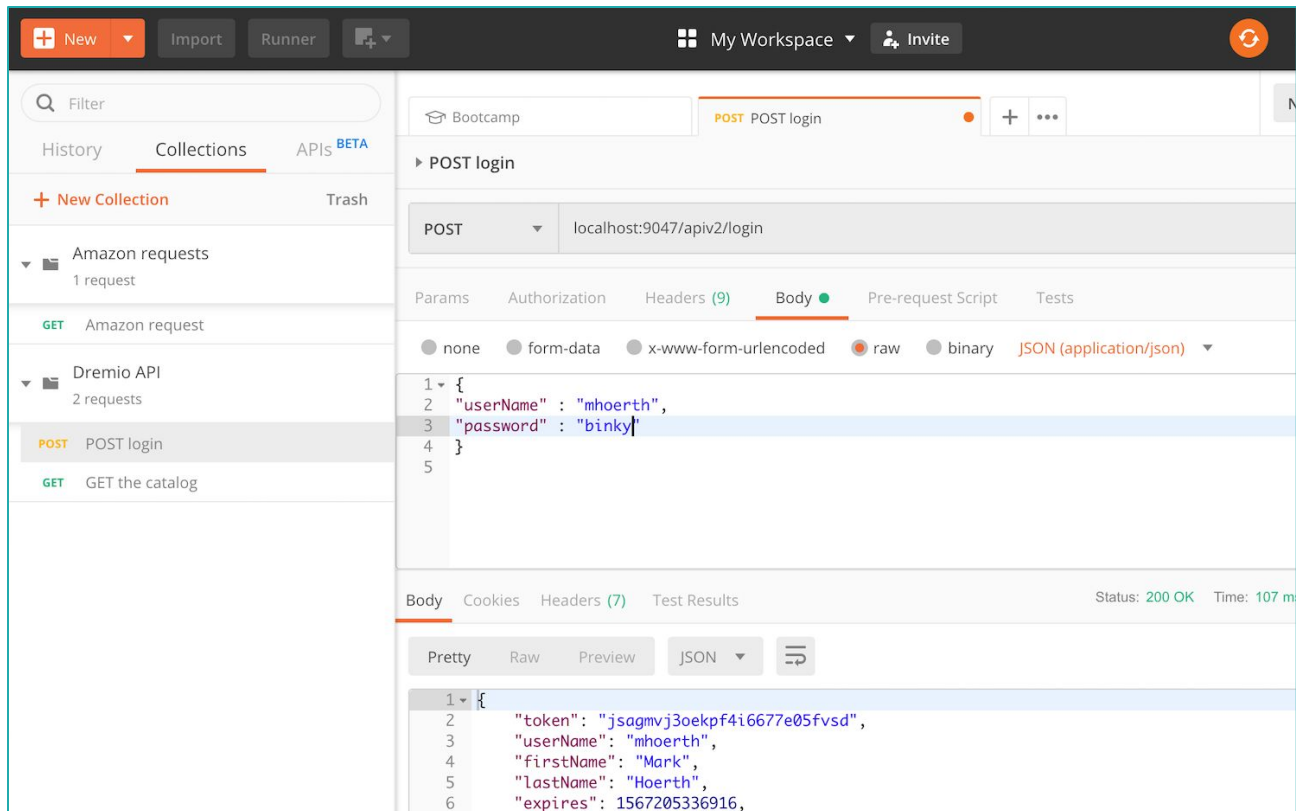
GET	/wlm/queue	Get a list of queues
POST	/wlm/queue	Create a new queue
PUT	/wlm/queue/{id}	Update queue attributes
DELETE	/wlm/queue/{id}	Delete a queue
GET	/wlm/rule	Get a list of rules
PUT	/wlm/rule	Create, update, delete rules

## Vote

GET	/vote	List all votes summary
-----	-------	------------------------

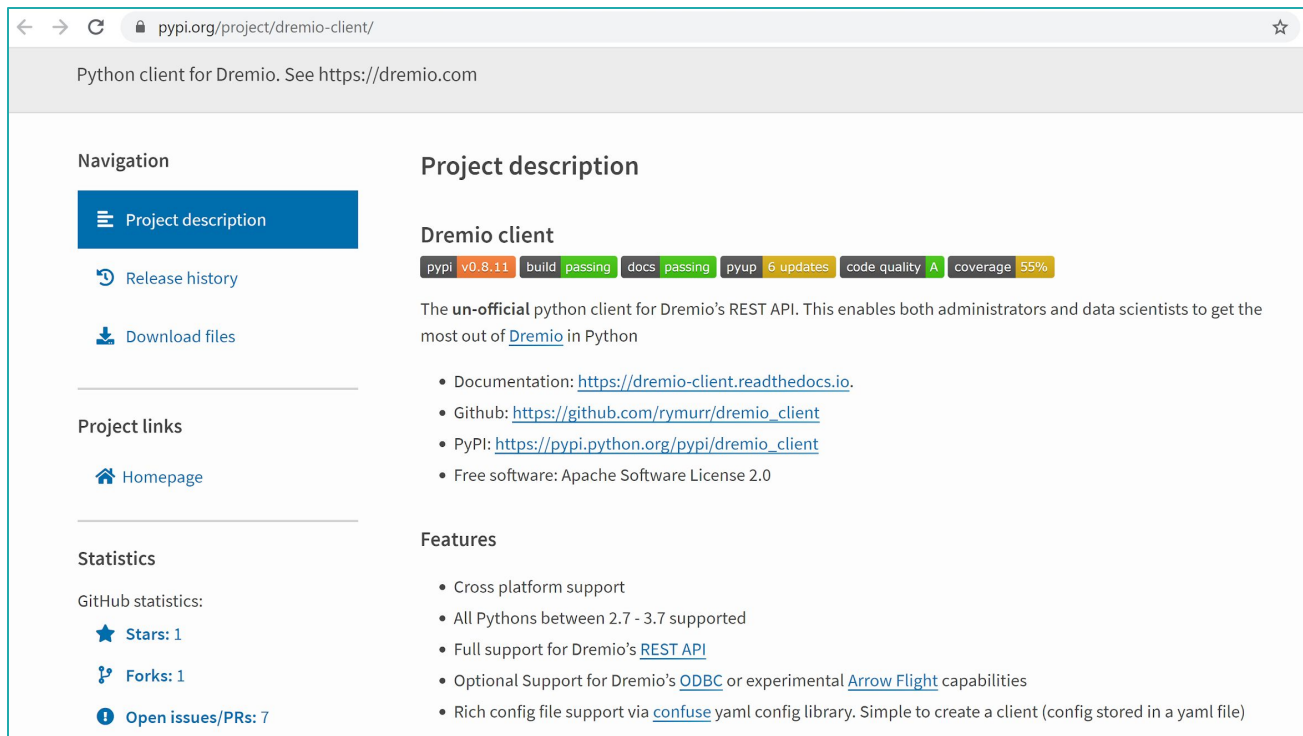
# Using the Dremio Rest API

- Install curl or an API development environment such as Postman, pictured here.
- Issue a login request to Dremio and receive an authorization token
- Using the token, get a list of the top level catalog entries from Dremio
- What other information can you get from the system?



# Dremio Python client

- All Pythons between 2.7 - 3.7 supported
- Full support for Dremio's REST API
- Optional Support for Dremio's ODBC or experimental Arrow Flight capabilities
- Rich config file support via confuse yaml config library. Simple to create a client (config stored in a yaml file)
- CLI interface for integration with scripts



The screenshot shows the PyPI project page for `dremio-client`. The browser address bar shows `https://pypi.org/project/dremio-client/`. The page title is "Python client for Dremio. See <https://dremio.com>".

**Navigation**

- Project description (selected)
- Release history
- Download files

**Project links**

- Homepage

**Statistics**

GitHub statistics:

- Stars: 1
- Forks: 1
- Open issues/PRs: 7

**Project description**

**Dremio client**

The **un-official** python client for Dremio's REST API. This enables both administrators and data scientists to get the most out of [Dremio](https://dremio.com) in Python

Badges: pypi v0.8.11, build passing, docs passing, pyup 6 updates, code quality A, coverage 55%

- Documentation: <https://dremio-client.readthedocs.io>.
- Github: [https://github.com/rymurr/dremio\\_client](https://github.com/rymurr/dremio_client)
- PyPI: [https://pypi.python.org/pypi/dremio\\_client](https://pypi.python.org/pypi/dremio_client)
- Free software: Apache Software License 2.0

**Features**

- Cross platform support
- All Pythons between 2.7 - 3.7 supported
- Full support for Dremio's [REST API](#)
- Optional Support for Dremio's [ODBC](#) or experimental [Arrow Flight](#) capabilities
- Rich config file support via [confuse](#) yaml config library. Simple to create a client (config stored in a yaml file)

<https://pypi.org/project/dremio-client/>

# Dremio Java client: Creating a VDS

In class VDSCreator { main() method:

```
for (int i = 0; i < sqlFiles.length; i++) {
    File sqlFile = sqlFiles[i].getCanonicalFile();
    StringBuffer sqlText = new StringBuffer();

    if (sqlFile.getName().endsWith(".sql")) {
        System.out.println("Processing SQL file: " +
            sqlFile.getName());

        // Obtain the SQL from the file
        List<String> lines = Files.readAllLines(sqlFile.
            getCanonicalFile().toPath(),
            Charset.defaultCharset());

        for (String line:lines) {
            sqlText.append(line);
        }

        System.out.println("VDS Definition: " + sqlText);

        payload = scrubText(sqlText.toString());
        response = dremio.postSQL(payload, null);
        System.out.println("Response: " + response);
    }
}
```

In class DremioAPI :

```
public String postSQL(String sql, String context) {
    /* Submits a SQL query.
     * https://docs.dremio.com/rest-api/
     * sources/post-source.html
     */

    String url = this.host + "/api/v3/sql";
    String body = null;
    if (context != null && context.length() > 0) {
        body = "{\"sql\": \"" + sql + "\", \"context\": \"" +
            context + "\"}";
    } else {
        body = "{\"sql\": \"" + sql + "\"}";
    }

    return doPost(url, body, true);
}
```

# Java doPost() part of Java client

```
private String doPost(String dremioURL, String bodyPayload, boolean withToken) {
    StringBuffer jsonString = new StringBuffer();
    try {

        // connection setup code

        if (withToken) {
            connection.setRequestProperty("Authorization", this.token);
        }
        OutputStreamWriter writer = new OutputStreamWriter(connection.getOutputStream());
        if (bodyPayload != null) {
            writer.write(bodyPayload);
        }
        writer.close();
        BufferedReader br = new BufferedReader(new InputStreamReader(connection.getInputStream()));

        String line;
        while ((line = br.readLine()) != null) {
            jsonString.append(line);
        }
        br.close();
        connection.disconnect()
    } catch (Exception e) {
        throw new RuntimeException(e.getMessage());
    }

    return jsonString.toString();
}
```

# Additional resources

Including



Dremio University



How Dremio and Tableau  
cloud data lake analytics  
InCrowd Sports  
Mar 3, 2020

In this webinar we explore how to accelerate performance for BI and Data Science workloads on the data lake. Connect with us on LinkedIn.

Blog

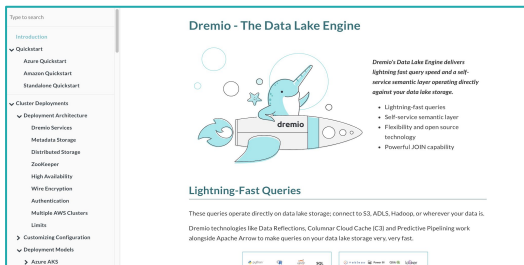


Documentation

# Dremio digital learning resources

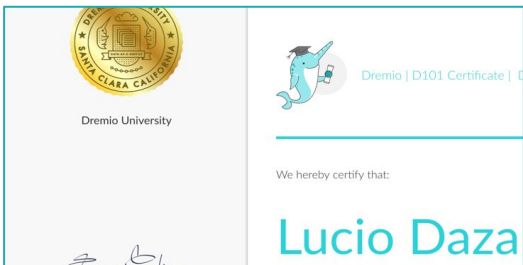
## Documentation

<https://docs.dremio.com>



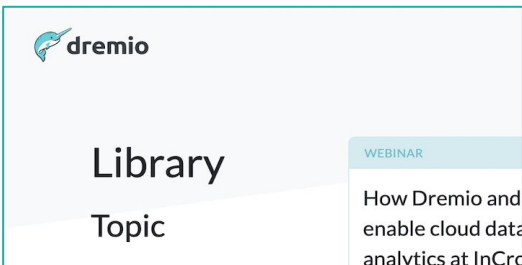
## Dremio University

<https://university.dremio.com>



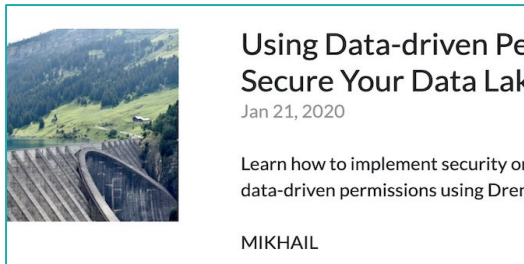
## Library

<https://www.dremio.com/library/>



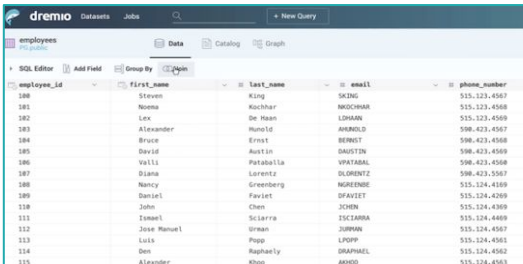
## Blog

<https://www.dremio.com/blog/>



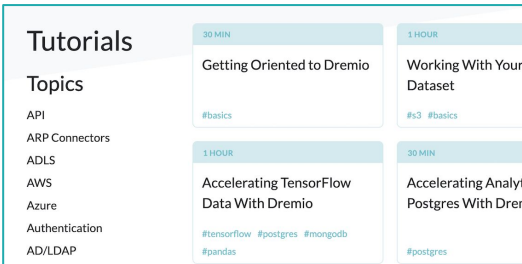
## Demos

<https://www.dremio.com/demos/>



## Tutorials

<https://www.dremio.com/tutorials/>





# Dremio University

- Guided, self-paced courses with hands-on labs in cloud-hosted Dremio instance
- Free to register - visit <https://university.dremio.com>
- Courses available: “Dremio Fundamentals”, “Data Reflections”, “Dremio for Data Consumers”
- Post questions in the [Dremio University topic on community.dremio.com](https://community.dremio.com)



Dremio University

Tomer Shiran  
FOUNDER & CEO  
Dremio Corp.

Gnarly The Narwhal  
DREMIO'S MASCOT  
Dremio Corp.



Dremio | D101 Certificate | Dremio University

We hereby certify that:

## Lucio Daza

successfully completed, and was awarded this platforms' Honor Code Certificate of Completion in:

Dremio | D101: Dremio Fundamentals

# Key digital training resources

		Dremio University	White Papers	Tutorials	Webinars	
Dremio Administrators	Data Engineers	Data Consumers	<ul style="list-style-type: none"><li><a href="#">Dremio Fundamentals</a></li><li><a href="#">Dremio for Data Consumers</a></li></ul>	<ul style="list-style-type: none"><li><a href="#">What Is a Data Lake Engine?</a></li><li><a href="#">What Is Data-as-a-Service?</a></li></ul>	<ul style="list-style-type: none"><li><a href="#">Getting Oriented to Dremio</a></li><li><a href="#">Working With Your First Dataset</a></li><li><a href="#">Combining Data From Multiple Datasets</a></li></ul>	
			<ul style="list-style-type: none"><li><a href="#">Data Reflections</a></li></ul>	<ul style="list-style-type: none"><li><a href="#">Dremio Security Architecture Guide</a></li><li><a href="#">Dremio Semantic Layer Best Practices</a></li><li><a href="#">BI on Big Data</a></li></ul>	<ul style="list-style-type: none"><li><a href="#">How To Share A Query Profile</a></li><li><a href="#">Data Curation With Dremio</a></li><li><a href="#">Using Dremio to Fix Data Inconsistency</a></li></ul>	<ul style="list-style-type: none"><li><a href="#">Data Science Across Data Sources With Apache Arrow</a></li><li><a href="#">Data Reflections: Accelerate your Queries Without Copies</a></li></ul>
				<ul style="list-style-type: none"><li><a href="#">Dremio Data Reflections Overview &amp; Best Practices</a></li><li><a href="#">Dremio Deployment Considerations Guide</a></li><li><a href="#">Dremio Architecture Guide</a></li></ul>	<ul style="list-style-type: none"><li><a href="#">Adding Users to Dremio</a></li><li><a href="#">Getting Started With Data Reflections</a></li></ul>	<ul style="list-style-type: none"><li><a href="#">Dremio 4.0 – Technical Deep Dive</a></li><li><a href="#">Running SQL-Based Workloads in the Cloud Using Apache Arrow</a></li></ul>

Click on the titles above to retrieve the resource.