Documentation

Welcome to the Apache Drill Documentation. If you do not see the navigation bar for the documentation, click the menu icon on the left at the top of any page.



If you're new to Drill, try out one of these resources:

* [Drill Introduction](http://drill.apache.org/docs/drill-introduction)
* [Tutorials](http://drill.apache.org/docs/tutorials-introduction)
* [Getting Started](javascript:%20void(0);)
* [Architecture](javascript:%20void(0);)
* [Tutorials](javascript:%20void(0);)
* [Install Drill](javascript:%20void(0);)
* [Configure Drill](javascript:%20void(0);)
* [Connect a Data Source](javascript:%20void(0);)
* [ODBC/JDBC Interfaces](javascript:%20void(0);)
* [Query Data](javascript:%20void(0);)
* [Performance Tuning](javascript:%20void(0);)
* [Log and Debug](javascript:%20void(0);)
* [SQL Reference](javascript:%20void(0);)
* [Data Sources and File Formats](javascript:%20void(0);)
* [Develop Custom Functions](javascript:%20void(0);)
* [Getting Started](javascript:%20void(0);)

# Drill Introduction

Jul 31, 2017

Drill is an Apache open-source SQL query engine for Big Data exploration(大数据检索的sql查询引擎). Drill is designed from the ground up（以下） to support high-performance analysis on the semi-structured（对半结构话的数据的高效分析） and rapidly evolving data coming from modern Big Data applications（以及从当今大数据应用中迅速的获取数据）, while still providing the familiarity and ecosystem of ANSI SQL（同时仍然提供我们熟悉的SQL生态系统这一行业标准的查询语言）, the industry-standard query language. Drill provides plug-and-play integration with existing Apache Hive and Apache HBase deployments（drill提供了与apache的hbase和hive等的集成插件）.

## What's New in Apache Drill 1.11

Drill 1.11 provides the following new features and improvements:

* Cryptography-related functions. (DRILL-5634)
* Spill to disk for the hash aggregate operator. (DRILL-5457)
* Format plugin support for PCAP files. (DRILL-5432)
* Ability to change the HDFS block Size for Parquet files. (DRILL-5379)
* Ability to store query profiles in memory. (DRILL-5481)
* Configurable CTAS directory and file permissions option. (DRILL-5391)
* Support for network encryption. (DRILL-4335)
* Relative paths stored in the metadata file. (DRILL-3867)
* Support for ANSI\_QUOTES. (DRILL-3510)

## What's New in Apache Drill 1.10

Drill 1.10 provides the following new features and improvements:

* Support for the [CREATE TEMPORARY TABLE AS (CTTAS)](http://drill.apache.org/docs/create-temporary-table-as-cttas/) command.
* A [JDBC connection option](http://drill.apache.org/docs/using-the-jdbc-driver/#using-the-jdbc-url-format-for-a-direct-drillbit-connection) that improves fault tolerance when connecting directly to a Drill node from a client.
* The [Web Console](http://drill.apache.org/docs/identifying-multiple-drill-versions-in-a-cluster) displays the Drill version and additional query profile statistics.
* Drill implicitly interprets the [INT96](http://drill.apache.org/docs/parquet-format/#about-int96-support/) timestamp data type in Parquet files.
* Support for Kerberos authentication between the client and drillbit.

## What's New in Apache Drill 1.9

Drill 1.9 provides the following new features:

* Asynchronous Parquet reader
* Parquet filter pushdown
* Dynamic UDF support
* HTTPD format plugin

## What's New in Apache Drill 1.8

Drill 1.8 provides the following new features and changes:

* Metadata cache pruning
* IF EXISTS parameter with the DROP TABLE and DROP VIEW commands
* DESCRIBE SCHEMA command
* Multi-byte delimiter support
* New parameters for filter selectivity estimates
* Changes to the configuration and launch scripts - See [Configuration and Launch Script Changes](http://drill.apache.org/docs/apache-drill-1-8-0-release-notes/#configuration-and-launch-script-changes)

## What's New in Apache Drill 1.7

Drill 1.7 provides the following new features:

* Monitoring via JMX
* Hive CHAR data type support
* HBase 1.x support

## What's New in Apache Drill 1.6

Drill 1.6 provides the following new features:

* Inbound impersonation
* Additional custom window frames

## What's New in Apache Drill 1.5

Drill 1.5 provides the following new features:

* Authentication and security for the Web interface and REST API
* Experimental query support for Apache Kudu (incubating)
* An improved memory allocator
* Configurable caching for Hive metadata

## What's New in Apache Drill 1.4

Drill 1.4 introduces the following improvements:

* [select with options](http://drill.apache.org/docs/plugin-configuration-basics/#using-the-formats-attributes-as-table-function-parameters) that you use in queries to change storage plugin settings
* Improved behavior when parsing CSV file header names
* A variable to set non-pretty, such as compact, printing of JSON
* Better drillbit.log files that include query text

Drill 1.4 fixes an error that occurred when you query a Hive table using the HBaseStorageHandler ([DRILL-3739](https://issues.apache.org/jira/browse/DRILL-3739)). To successfully query a Hive table using the HBaseStorageHandler, you need to configure the Hive storage plugin as described in the [Hive storage plugin documentation](http://drill.apache.org/docs/hive-storage-plugin/#connect-drill-to-the-hive-remote-metastore).

## What's New in Apache Drill 1.3

This releases fix issues and add a number of enhancements, including the following ones:

* [Enhanced Amazon S3 support](http://drill.apache.org/docs/s3-storage-plugin/)
* Hetrogeneous types  
  Support for columns that evolve from one data type to another over time.
* [Text file headers](http://drill.apache.org/docs/text-files-csv-tsv-psv/#using-a-header-in-a-file)
* [Sequence files support](http://drill.apache.org/docs/querying-sequence-files/)
* Enhancements related to querying Hive tables, MongoDB collections, and Avro files

## What's New in Apache Drill 1.2

This release of Drill fixes [many issues](http://drill.apache.org/docs/apache-drill-1-2-0-release-notes/) and introduces a number of enhancements, including the following ones:

* Support for JDBC data sources, such as MySQL, through a [new JDBC Storage plugin](https://issues.apache.org/jira/browse/DRILL-3180)
* Improvements in the Drill JDBC driver including inclusion of Javadocs and better application dependency compatibility
* Enhancements to Avro file formats  
  + [Support for complex data types](https://issues.apache.org/jira/browse/DRILL-3565), such as UNION and MAP
  + [Optimized Avro file processing](https://issues.apache.org/jira/browse/DRILL-3720) (block-wise)
* Partition pruning improvements
* A number of new [SQL window functions](http://drill.apache.org/docs/sql-window-functions)  
  + NTILE
  + LAG and LEAD
  + FIRST\_VALUE and LAST\_VALUE
* [HTTPS support](http://drill.apache.org/docs/configuring-web-console-and-rest-api-security/) for Web Console operations
* Performance improvements for [querying HBase](http://drill.apache.org/docs/querying-hbase/#querying-big-endian-encoded-data), which includes leveraging [ordered byte encoding](http://drill.apache.org/docs/querying-hbase/#leveraging-hbase-ordered-byte-encoding)
* [Optimized reads](http://drill.apache.org/docs/querying-hive/#optimizing-reads-of-parquet-backed-tables) of Parquet-backed, Hive tables
* Read support for the [Parquet INT96 type](http://drill.apache.org/docs/parquet-format/#about-int96-support) and a new TIMESTAMP\_IMPALA type used with the [CONVERT\_FROM](http://drill.apache.org/docs/supported-data-types/#data-types-for-convert_to-and-convert_from-functions) function decodes a timestamp from Hive or Impala.
* [Parquet metadata caching](http://drill.apache.org/docs/optimizing-parquet-metadata-reading/) to improve query performance on a large number of files
* DROP TABLE command
* Improved correlated subqueries
* Union Distinct
* Improved LIMIT processing

## What's New in Apache Drill 1.1

Many enhancements in Apache Drill 1.1 include the following key features:

* [SQL window functions](http://drill.apache.org/docs/sql-window-functions)
* [Partitioning data](http://drill.apache.org/docs/drill-introduction/) using the new [PARTITION BY](http://drill.apache.org/docs/partition-by-clause) clause in the CTAS command
* [Delegated Hive impersonation](http://drill.apache.org/docs/configuring-user-impersonation-with-hive-authorization/)
* Support for UNION and UNION ALL and better optimized plans that include UNION.

## What's New in Apache Drill 1.0

Apache Drill 1.0 offers the following new features:

* Many performance planning and execution [improvements](http://drill.apache.org/docs/performance-tuning-introduction/).
* Updated [Drill shell](http://drill.apache.org/docs/configuring-the-drill-shell) now formats query results.
* [Query audit logging](http://drill.apache.org/docs/getting-query-information/) for getting the query history on a Drillbit.
* Improved connection handling.
* New Errors tab in the Query Profiles UI that facilitates troubleshooting and distributed storing of profiles.
* Support for a new storage plugin input format: [Avro](http://avro.apache.org/docs/current/spec.html)

In this release, Drill disables the DECIMAL data type, including casting to DECIMAL and reading DECIMAL types from Parquet and Hive. You can [enable the DECIMAL type](http://drill.apache.org/docs/drill-introduction/docs/supported-data-types/#enabling-the-decimal-type), but this is not recommended.

## Apache Drill Key Features

Key features of Apache Drill are:

* Low-latency SQL queries 低延迟的sql查询
* Dynamic queries on self-describing data in files (such as JSON, Parquet, text) and HBase tables, without requiring metadata definitions in the Hive metastore. 对于基于文件的自描述数据和hbase的表的动态查询，不需要在hive的metastore中的metadata定义
* ANSI SQL
* Nested data support 支持嵌套
* Integration with Apache Hive (queries on Hive tables and views, support for all Hive file formats and Hive UDFs) 与apache的hive集成（如查询hive的表和视图，支持所有的hive文件格式和hive的UDFs）
* BI/SQL tool integration using standard JDBC/ODBC drivers 集成了标准的JDBC和ODBC驱动的商业智能/sql 工具

## Quick Links

If you've never used Drill, visit these links to get a jump start:

* [Drill in 10 Minutes](http://drill.apache.org/docs/drill-in-10-minutes/)
* [Query Files](http://drill.apache.org/docs/querying-a-file-system)
* [Query HBase](http://drill.apache.org/docs/querying-hbase)
* [SQL Support](http://drill.apache.org/docs/sql-reference-introduction/)
* [Drill Tutorials](http://drill.apache.org/docs/tutorials-introduction)

# Why Drill

## Top 10 Reasons to Use Drill

## 1. Get started in minutes

It takes just a few minutes to get started with Drill. Untar（解压） the Drill software on your Linux, Mac, or Windows laptop and run a query on a local file. No need to set up any infrastructure or to define schemas. Just point to the data, such as data in a file, directory, HBase table, and drill（例如文件数据、文件目录、hbase的表和drill等）.

$ tar -xvf apache-drill-<version>.tar.gz

$ <install directory>/bin/drill-embedded

0: jdbc:drill:zk=local> SELECT \* FROM cp.`employee.json` LIMIT 5;

+--------------+----------------------------+---------------------+---------------+--------------+----------------------------+-----------+----------------+-------------+------------------------+----------+----------------+----------------------+-----------------+---------+-----------------------+

| employee\_id | full\_name | first\_name | last\_name | position\_id | position\_title | store\_id | department\_id | birth\_date | hire\_date | salary | supervisor\_id | education\_level | marital\_status | gender | management\_role |

+--------------+----------------------------+---------------------+---------------+--------------+----------------------------+-----------+----------------+-------------+------------------------+----------+----------------+----------------------+-----------------+---------+-----------------------+

| 1 | Sheri Nowmer | Sheri | Nowmer | 1 | President | 0 | 1 | 1961-08-26 | 1994-12-01 00:00:00.0 | 80000.0 | 0 | Graduate Degree | S | F | Senior Management |

| 2 | Derrick Whelply | Derrick | Whelply | 2 | VP Country Manager | 0 | 1 | 1915-07-03 | 1994-12-01 00:00:00.0 | 40000.0 | 1 | Graduate Degree | M | M | Senior Management |

| 4 | Michael Spence | Michael | Spence | 2 | VP Country Manager | 0 | 1 | 1969-06-20 | 1998-01-01 00:00:00.0 | 40000.0 | 1 | Graduate Degree | S | M | Senior Management |

| 5 | Maya Gutierrez | Maya | Gutierrez | 2 | VP Country Manager | 0 | 1 | 1951-05-10 | 1998-01-01 00:00:00.0 | 35000.0 | 1 | Bachelors Degree | M | F | Senior Management |

## 2. Schema-free JSON model

Drill is the world's first and only distributed SQL engine that doesn't require schemas（不需要scheme，即不需要数据结构）. It shares the same schema-free JSON model as MongoDB and Elasticsearch（它与mongodb和elasticsearch 一样使用无scheme的json）. No need to define and maintain schemas or transform data (ETL)（不需要保存表schem或者对数据进行转换）. Drill automatically understands the structure of the data（drill能够自动理解数据结构）.

## 3. Query complex, semi-structured data in-situ

Using Drill's schema-free JSON model, you can query complex, semi-structured data in situ. No need to flatten or transform the data prior to or during query execution（不需要事先对数据进行转换）. Drill also provides intuitive extensions to SQL to work with nested data. Here's a simple query on a JSON file demonstrating how to access nested elements and arrays（嵌套的元素和数组）:

SELECT \* FROM (SELECT t.trans\_id,

t.trans\_info.prod\_id[0] AS prod\_id,

t.trans\_info.purch\_flag AS purchased

FROM `clicks/clicks.json` t) sq

WHERE sq.prod\_id BETWEEN 700 AND 750 AND

sq.purchased = 'true'

ORDER BY sq.prod\_id;

## 4. Real SQL -- not "SQL-like"

Drill supports the standard SQL:2003 syntax. No need to learn a new "SQL-like" language or struggle with a semi-functional BI tool. Drill supports many data types including DATE, INTERVAL, TIMESTAMP, and VARCHAR, as well as complex query constructs such as correlated sub-queries and joins in WHERE clauses. Here is an example of a TPC-H standard query that runs in Drill:

### TPC-H query 4

SELECT o.o\_orderpriority, COUNT(\*) AS order\_count

FROM orders o

WHERE o.o\_orderdate >= DATE '1996-10-01'

AND o.o\_orderdate < DATE '1996-10-01' + INTERVAL '3' month

AND EXISTS(

SELECT \* FROM lineitem l

WHERE l.l\_orderkey = o.o\_orderkey

AND l.l\_commitdate < l.l\_receiptdate

)

GROUP BY o.o\_orderpriority

ORDER BY o.o\_orderpriority;

## 5. Leverage standard BI tools

Drill works with standard BI tools. You can use your existing tools, such as Tableau, MicroStrategy, QlikView and Excel.

## 6. Interactive queries on Hive tables

Apache Drill lets you leverage your investments in Hive. You can run interactive queries with Drill on your Hive tables and access all Hive input/output formats (including custom SerDes). You can join tables associated with different Hive metastores（可以连接不同的hive元数据存储）, and you can join a Hive table with an HBase table or a directory of log files. Here's a simple query in Drill on a Hive table:

SELECT `month`, state, sum(order\_total) AS sales

FROM hive.orders

GROUP BY `month`, state

ORDER BY 3 DESC LIMIT 5;

## 7. Access multiple data sources

Drill is extensible（可扩展的）. You can connect Drill out-of-the-box to file systems (local or distributed, such as S3 and HDFS), HBase and Hive. You can implement a storage plugin to make Drill work with any other data source. Drill can combine data from multiple data sources on the fly in a single query, with no centralized metadata definitions（无需中心metadata的预先定义，drill即可连接多个数据源）. Here's a query that combines data from a Hive table, an HBase table (view) and a JSON file:

SELECT custview.membership, sum(orders.order\_total) AS sales

FROM hive.orders, custview, dfs.`clicks/clicks.json` c

WHERE orders.cust\_id = custview.cust\_id AND orders.cust\_id = c.user\_info.cust\_id

GROUP BY custview.membership

ORDER BY 2;

## 8. User-Defined Functions (UDFs) for Drill and Hive

Drill exposes a simple, high-performance Java API to build [custom user-defined functions](http://drill.apache.org/docs/develop-custom-functions/) (UDFs) for adding your own business logic to Drill（drill为用户定义自己的业务逻辑的构建提供了以非常简单的java api）. Drill also supports Hive UDFs(drill也支持hive的用户自定义函数，无需修改即可迁移到drill中). If you have already built UDFs in Hive, you can reuse them with Drill with no modifications.

## 9. High performance

Drill is designed from the ground up for high throughput and low latency（高产出低延迟）. It doesn't use a general purpose execution engine like MapReduce, Tez or Spark. As a result, Drill is flexible (schema-free JSON model) and performant（弹性的和表现好的）. Drill's optimizer leverages rule- and cost-based techniques, as well as data locality and operator push-down, which is the capability to push down query fragments into the back-end data sources. Drill also provides a columnar and vectorized execution engine（也提供了针对列和垂直列的执行引擎）, resulting in higher memory and CPU efficiency（导致了高效的内存和CPU效率）.

## 10. Scales from a single laptop to a 1000-node cluster

Drill is available as a simple download you can run on your laptop. When you're ready to analyze larger datasets, deploy Drill on your Hadoop cluster (up to 1000 commodity servers). Drill leverages the aggregate memory in the cluster to execute queries using an optimistic pipelined model, and automatically spills to disk when the working set doesn't fit in memory（使用了一个优化的管道模型，利用集群中的很大内存来执行查询，并且当内存不足时自动将数据写到磁盘）.