Apache Drill

作者：whoami

查询一切数据源NOSQL(NewSQL&&RDBMS)

- 敏捷
- 灵活
- 熟悉

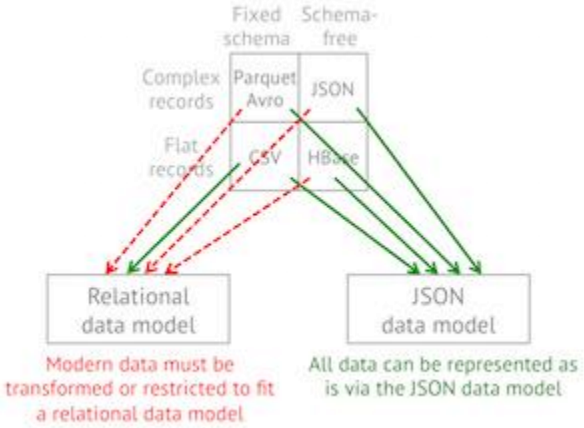
- Drill supports a variety of NoSQL databases and file systems, including HBase, MongoDB, MapR-DB, HDFS, MapR-FS, Amazon S3, Azure Blob Storage, Google Cloud Storage, Swift, NAS and local files. A single query can join data from multiple datastores. For example, you can join a user profile collection in MongoDB with a directory of event logs in Hadoop.
- Drill's datastore-aware optimizer automatically restructures a query plan to leverage the datastore's internal processing capabilities. In addition, Drill supports data locality, so it's a good idea to co-locate Drill and the datastore on the same nodes.

灵活SQL分析多种数据源

Kiss the overhead goodbye and enjoy data agility

Traditional query engines demand significant IT intervention before data can be queried. Drill gets rid of all that overhead so that users can just query the raw data in-situ. There's no need to load the data, create and maintain schemas, or transform the data before it can be processed. Instead, simply include the path to a Hadoop directory, MongoDB collection or S3 bucket in the SQL query.

Drill leverages advanced query compilation and re-compilation techniques to maximize performance without requiring up-front schema knowledge.

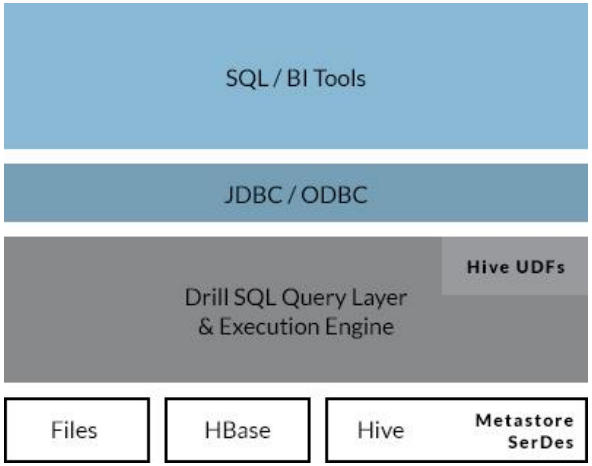



```
SELECT * FROM dfs.root.`/web/logs`;

SELECT country, count(*)
  FROM mongodb.web.users
  GROUP BY country;

SELECT timestamp
  FROM s3.root.`clicks.json`
  WHERE user_id = 'jdoe';
```

秉着“授人鱼不如授人以渔”理念

https://drill.apache.org/

集成多种BI工具



Drill supports standard SQL. Business users, analysts and data scientists can use standard BI/analytics tools such as Tableau, Qlik, MicroStrategy, Spotfire, SAS and Excel to interact with non-relational datastores by leveraging Drill's JDBC and ODBC drivers. Developers can leverage Drill's simple REST API in their custom applications to create beautiful visualizations.

Drill's virtual datasets allow even the most complex, non-relational data to be mapped into BI-friendly structures which users can explore and visualize using their tool of choice.

无等待...

```
$ curl -L "<url>" | tar xzf -
$ cd apache-drill-<version>
$ bin/drill-embedded
```

Scale from one laptop to 1000s of servers

We made it easy to download and run Drill on your laptop. It runs on Mac, Windows and Linux, and within a minute or two you'll be exploring your data. When you're ready for prime time, deploy Drill on a cluster of commodity servers and take advantage of the world's most scalable and high performance execution engine.

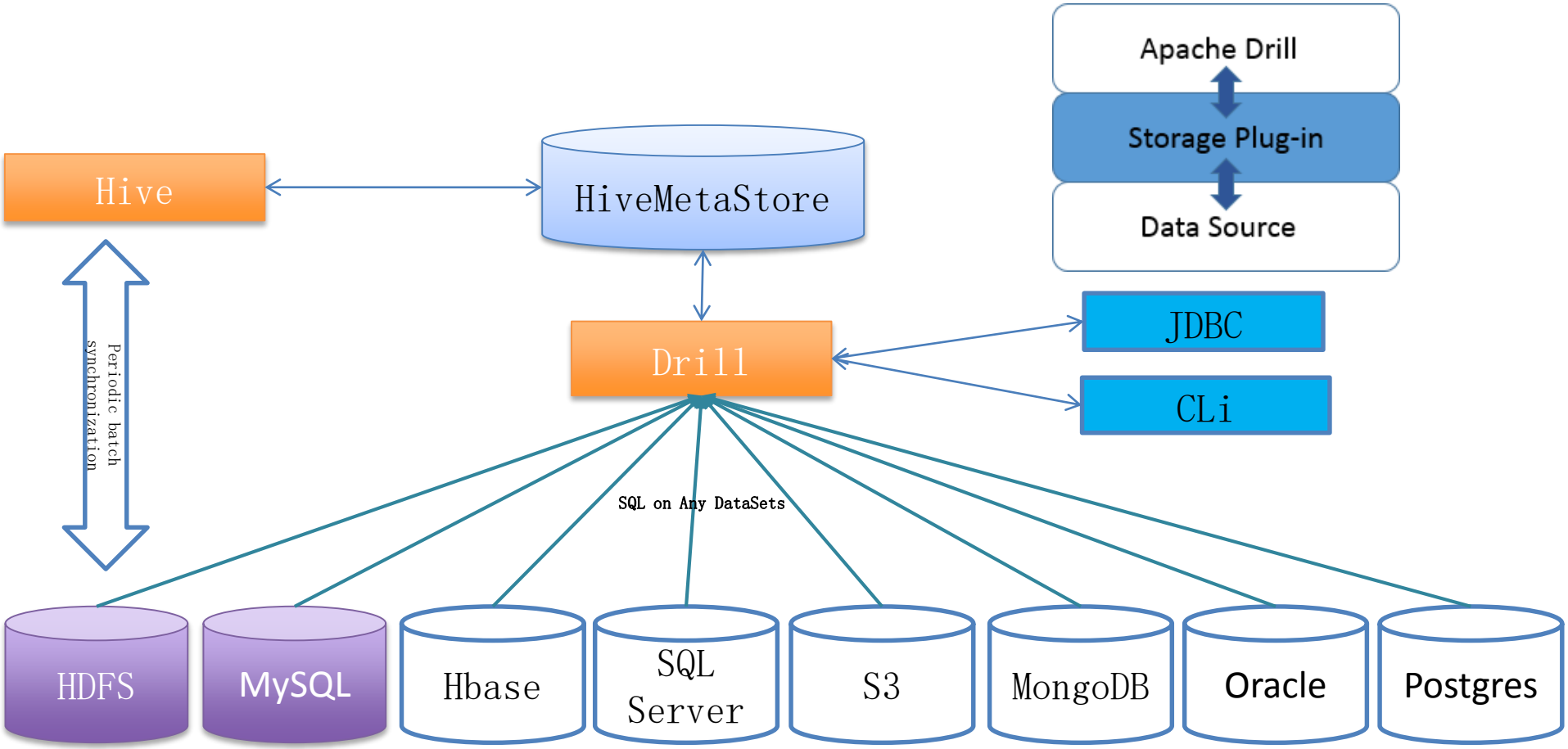
Drill's symmetrical architecture (all nodes are the same) and simple installation make it easy to deploy and operate very large clusters.

No more waiting for coffee

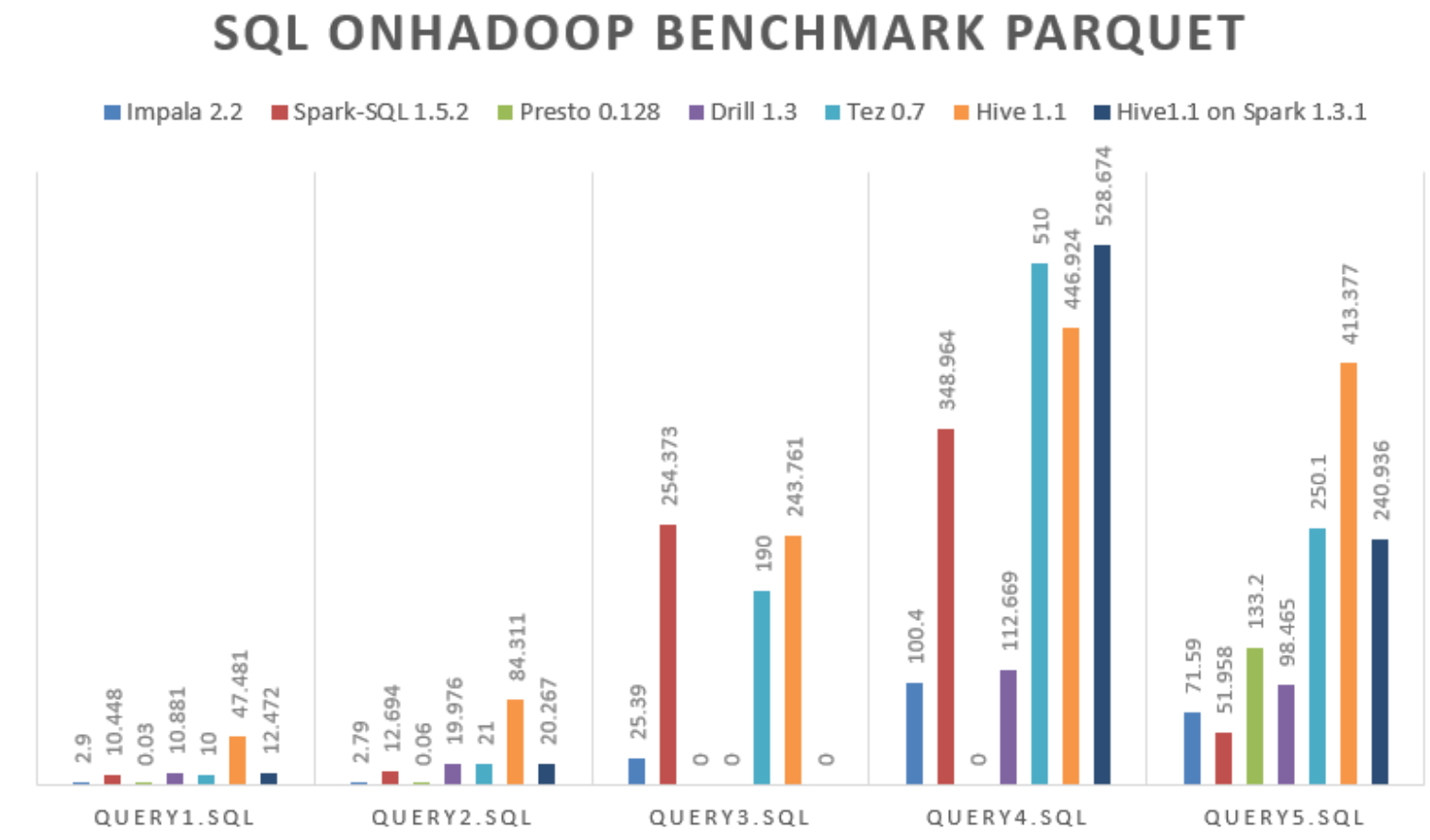
Drill isn't the world's first query engine, but it's the first that combines both flexibility and speed. To achieve this, Drill features a radically different architecture that enables record-breaking performance without sacrificing the flexibility offered by the JSON document model. Drill's design includes:

- Columnar execution engine (the first ever to support complex data!)
- Data-driven compilation and recompilation at execution time
- Specialized memory management that reduces memory footprint and eliminates garbage collections
- Locality-aware execution that reduces network traffic when Drill is co-located with the datastore
- Advanced cost-based optimizer that pushes processing into the datastore when possible

Drill 多数据源


Apache Drill
Storage Plug-in
Data Source
Hive
HiveMetaStore
Periodic batch synchronization
Drill
JDBC
CLi
SQL on Any DataSets
HDFS
MySQL
Hbase
SQL Server
S3
MongoDB
Oracle
Postgres

性能测试



执行时间为0s，说明执行失败，0是语法不支持,0.1执行失败，有些是框架本身不稳定(Hive on Spark && Presto)

性能测试

Node Num:4 MEM:16G CPU:8 Data: Parquet

SQL	Impala 2.2	Spark-SQL 1.5.2	Presto 0.128	Drill 1.3	Tez 0.7	Hive 1.1	Hive1.1 on Spark 1.3.1
query1.sql	2.9	10.448	0.03	10.881	10	47.481	12.472
query2.sql	2.79	12.694	0.06	19.976	21	84.311	20.267
query3.sql	25.39	254.373	0	0	190	243.761	0.1
query4.sql	100.4	348.964	0	112.669	510	446.924	528.674
query5.sql	71.59	51.958	133.2	98.465	250.1	413.377	240.936
备注			语法不支持				Hive on Spark不稳定,任务容易失败

执行时间为0s，说明执行失败，0是语法不支持,0.1执行失败，有些是框架本身不稳定(Hive on Spark && Presto)

SQL and Datasource： http://www.itweet.cn/2016/03/20/Impala-Hive-performance-tuning/

在线演示

Thank you

提问时间？

Blog： http://www.itweet.cn

PPT： https://github.com/itweet/course

Video： http://www.tudou.com/home/sparkjvm/