

Efficient Big Data Exploration with SQL and Apache Drill

Jonatan Kazmierczak



Java User Group Switzerland, Zürich, 07.02.2017

About author



Jonatan.Kazmierczak (at) gmail (dot) com

- senior consultant
at **Atos Consulting Switzerland**
- creator of **Class Visualizer**
- top rated participant in contests
in programming and data science:
HackerRank, TopCoder, Google Code Jam
- working with Java and SQL for 20 years



About author – cont.

first rank in Java

Rank	Hacker ▼	Points	Country : Switzerland ✕
1	 jonatan_k	953.00	



www.hackerrank.com/leaderboard/java/practice/level/1/filter/country=Switzerland/page/1



www.hackerrank.com/jonatan_k

code jam

```
System.out.println("hello, world!");
```



topcoder™

Agenda

- Introduction
- Demo: starting with Drill
- Technical details
- Demo: deep dive into Drill
- Summary, Q & A

Introduction



Computers – before

www.amibay.com/showthread.php?71410-Atari-65XE-BOX-XC12-BOX-2-Quickshots



-- SQL and Apache Drill -- Jonatan Kazmierczak -- JUG CH 2017 --

Data – before



Data – now



-- SQL and Apache Drill -- Jonatan Kazmierczak -- JUG CH 2017 --

Computers – now



32GB RAM



3TB RAM

What is Apache Drill ?

- low latency distributed schema-free SQL query engine for large-scale datasets
- designed to scale to several thousands of nodes and query petabytes of data at the speeds required by BI/Analytics environments





Demo: starting with Drill



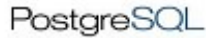
Technical details

Basic info

Website	drill.apache.org
Current version	1.9.0
Query language	SQL:2003
Interfaces	shell, web console, JDBC/ODBC, REST API, Java API, C++ API



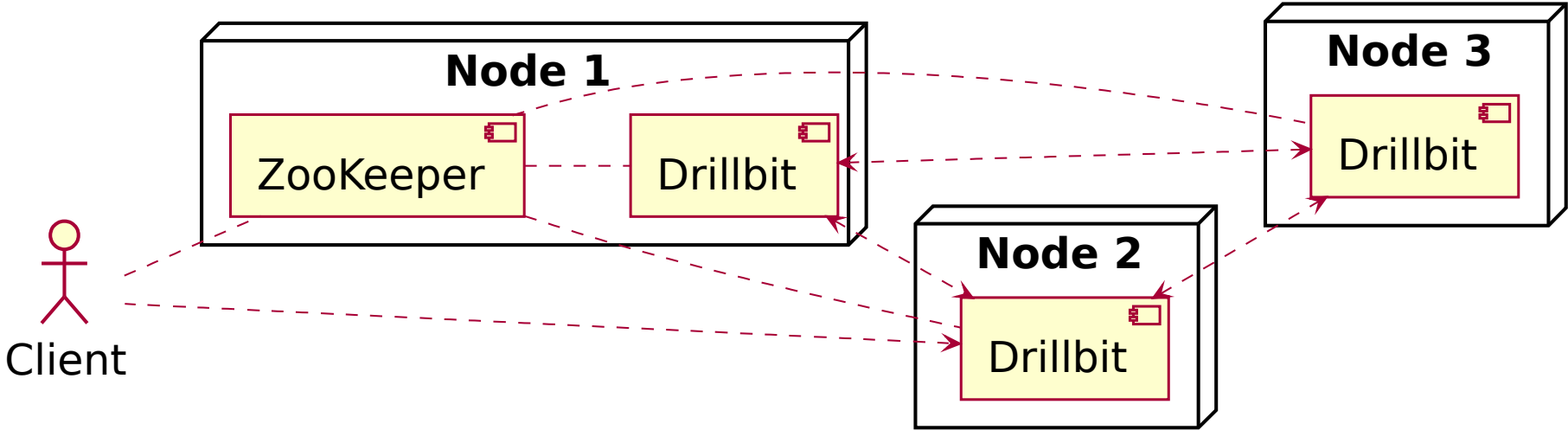
Supported data sources and formats



Features

- Dynamic schema discovery
- Flexible data model
- In-memory data processing (whenever possible)
- Extensible architecture
- Distributed and embedded mode

Distributed setup



Sample query

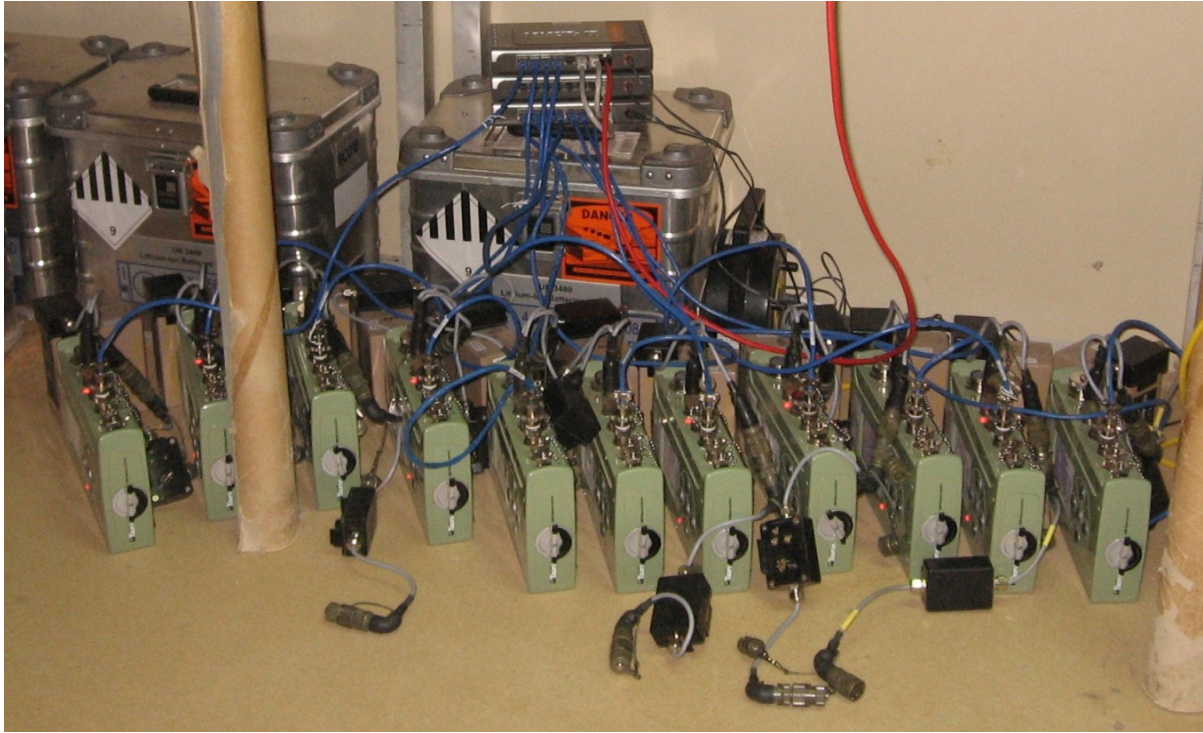
```
select * from dfs.demo.`countries.csv`
```

storage plugin

workspace

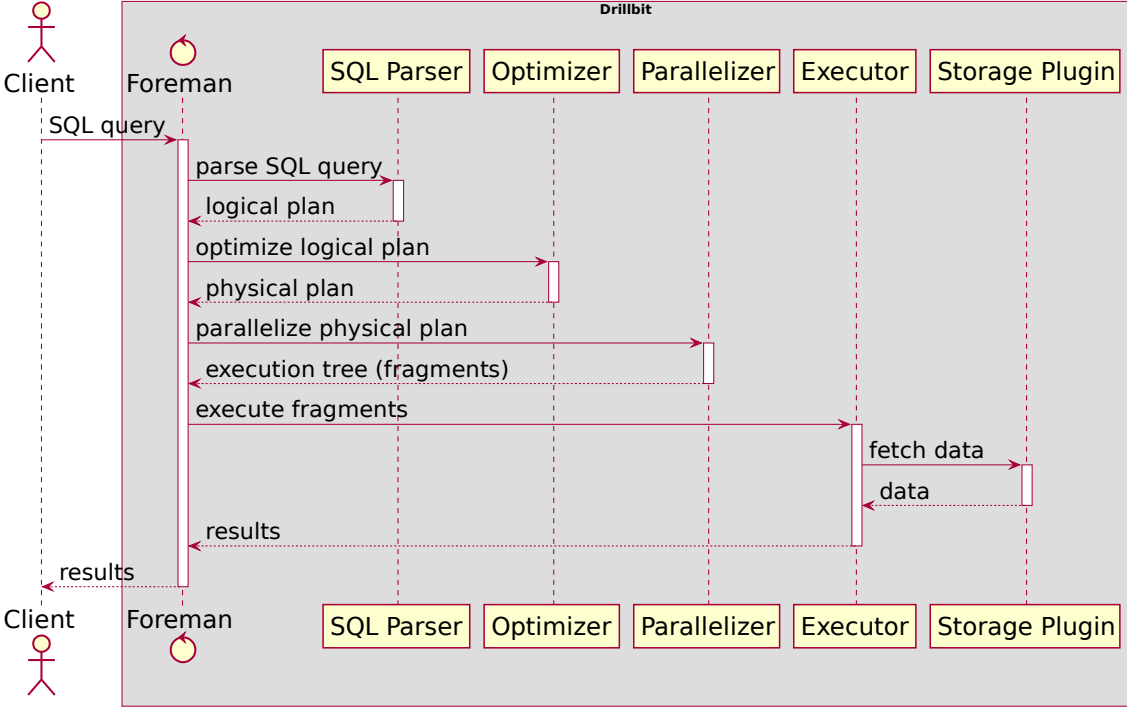
table / view / file / document

Config – storage plugins



-- SQL and Apache Drill -- Jonatan Kazmierczak -- JUG CH 2017 --

Query execution



Drill inside

over 0x2000 classes

The screenshot shows the Class Visualizer interface for a new project. The central pane displays a Relations Diagram for the `Drillbit` class. The diagram shows `Drillbit` as a central node with several dependencies and inheritance relationships. `Drillbit` inherits from `Object` and `AutoCloseable`. It has dependencies on `ShutdownThread`, `BootstrapContext`, `ClusterCoordinator`, `DrillbitContext`, `Logger`, `PersistentStoreProvider`, `RegistrationHandle`, `ServiceEngine`, `StoragePluginRegistry`, `String`, `WebServer`, `WorkManager`, `DrillConfig`, `RemoteServiceSet`, `ScanResult`, `StartupOptions`, and `DrillbitStartupException`. The `DrillConnectionImpl` class is also shown, with its own members and relations.

On the left, the Classes Filter shows a list of classes under the `foreman` package, including `ConnectionClosedListener`, `DrillbitStatusListener`, `ExceptionType`, `Foreman`, `ForemanException`, `ForemanResult`, `ForemanSetupException`, `FragmentData`, `FragmentStatusListener`, `FragmentSubmitFailures`, `FragmentSubmitListener`, `Inneriter`, `LoggedQuery`, `NodeTracker`, `Outeriter`, `PhysicalFromLogicalExplain`, `QueryManager`, `ResponseSendListener`, `Signal`, `SignalListener`, `SqlUnsupportedException`, `StateEvent`, `StateSwitch`, `SubmissionException`, `UnsupportedDataTypeException`, `UnsupportedFunctionException`, and `UnsupportedRelOperatorException`.

On the right, the Preview pane shows the details for the `Object` class, including its package (`org.apache.drill.exec.server`), its superclass (`Drillbit`), constants (`SYSTEM_OPTIONS_NAME`), fields (`context`, `coord`, `engine`, `isClosed`, `manager`, `registrationHandle`, `storageRegistry`, `storeProvider`, `webServer`), properties (`context`), constructors (`Drillbit`), and methods (`access$000`, `close`, `javaPropertiesToSystemOptions`, `main`, `run`, `start`, `start$DrillConfig`, `start$RemoteServiceSet`, `start$StartupOptions`, `stripQuotes`, `throwInvalidSystemOption`).



Demo: deep dive into Drill

Summary



Advantages

- Easy to start working with
- Concept of SQL-on-Anything
- Using standard SQL

Disadvantages

- Partially implemented or unfinished features
- Lacks in documentation

Use cases

- Data exploration
- Data transformation
- BI / Data analytics

Applicable

Not applicable

Questions



Thank you

Jonatan.Kazmierczak (at) gmail (dot) com

Son-of-God.info