Apache Drill
interactive, ad-hoc query at scale

Michael Hausenblas, Chief Data Engineer EMEA, MapR

Hadoop ecosystem - Open Source drives innovation and adoption in Big Data, 2013-01-05

Which workloads do **you** encounter in **your** environment?

http://www.flickr.com/photos/kevinomara/2866648330/ licensed under CC BY-NC-ND 2.0

APACHE DRILL

Batch processing

… for recurring tasks such as large-scale data mining, aggregation, ETL offloading, etc.

OLTP

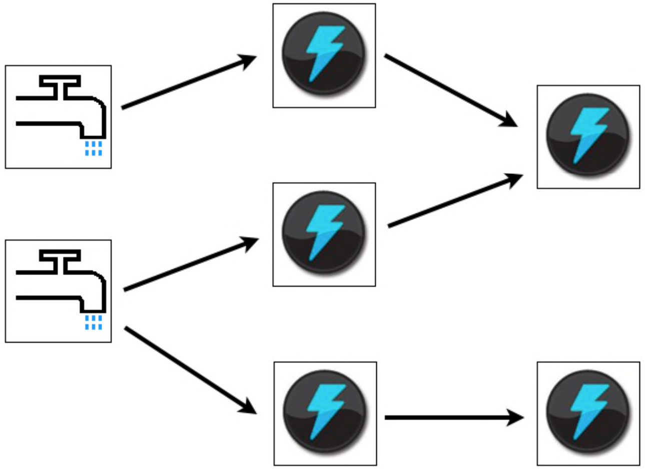
APACHE HBASE

Cassandra

… for example user-facing eCommerce transactions, real-time messaging at scale (FB) , etc.

Stream processing

Storm
Distributed and fault-tolerant realtime computation

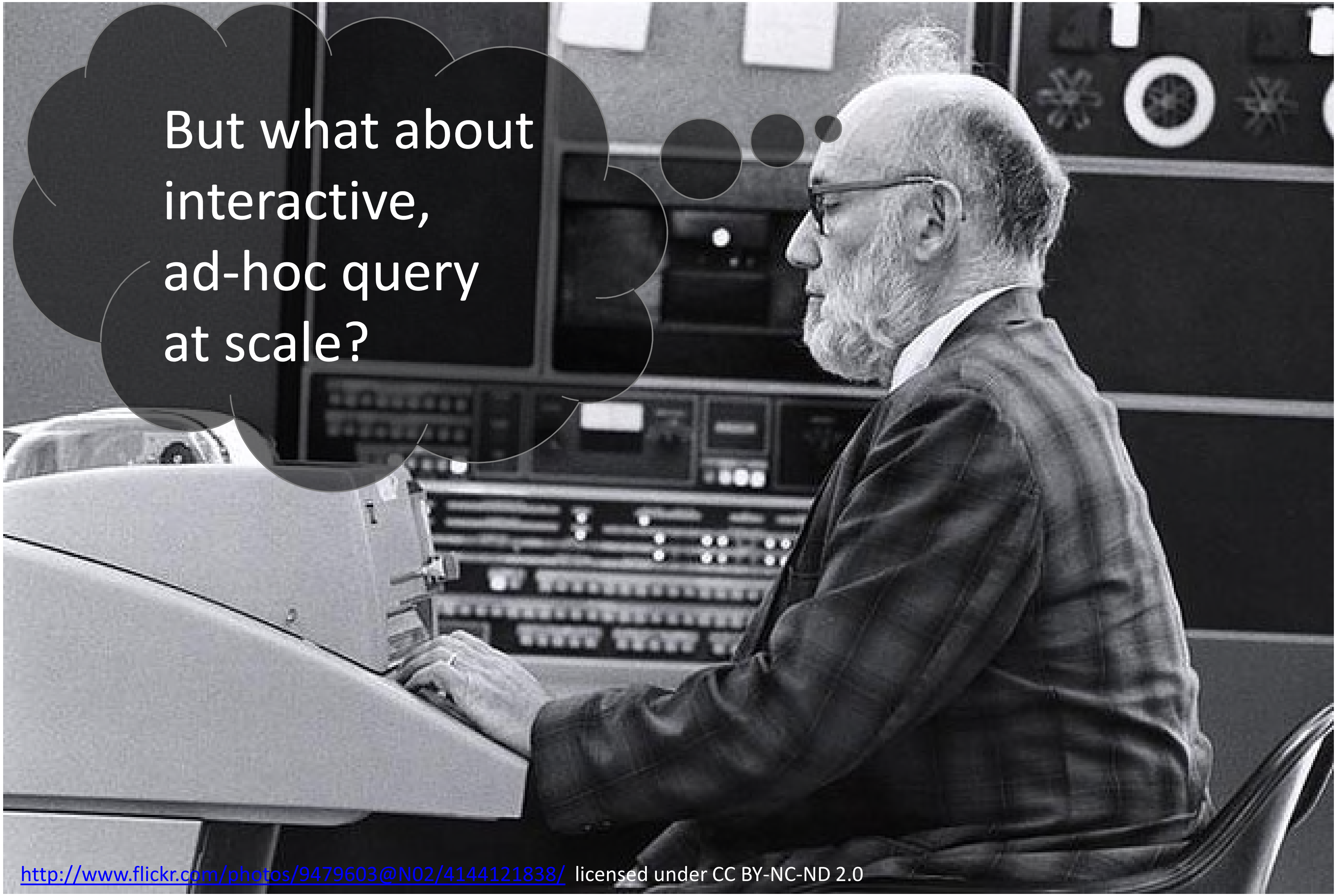


S4 distributed stream computing platform

… in order to handle stream sources such as social media feeds or sensor data (mobile phones, RFID, weather stations, etc.)

Search

… retrieval of items from semi-structured data formats (XML, JSON, etc.), documents (plain text, etc.) and datastores (MongoDB, CouchDB, etc.)



But what about interactive, ad-hoc query at scale?

http://www.flickr.com/photos/9479603@N02/4144121838/ licensed under CC BY-NC-ND 2.0

Interactive Query (?)

HIVE

TERADATA ASTER

Greenplum

ORACLE®

TERADATA MySQL®

low-latency

APACHE DRILL

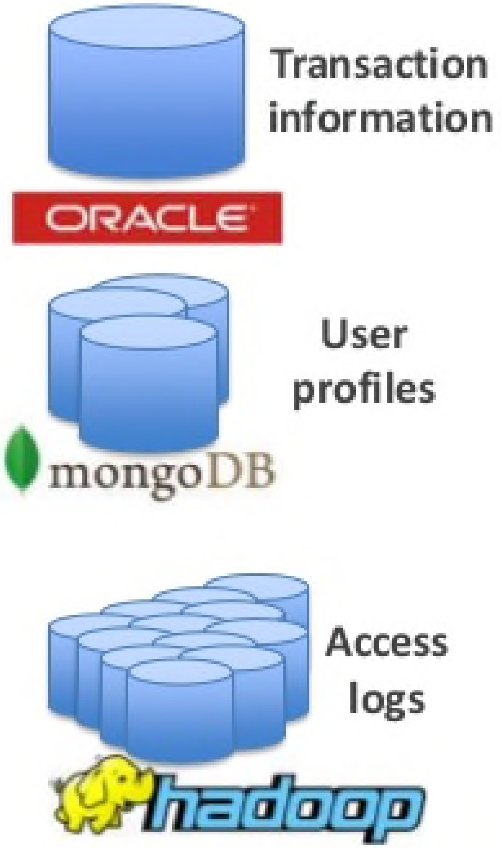
Impala

citusdata

HADAPT

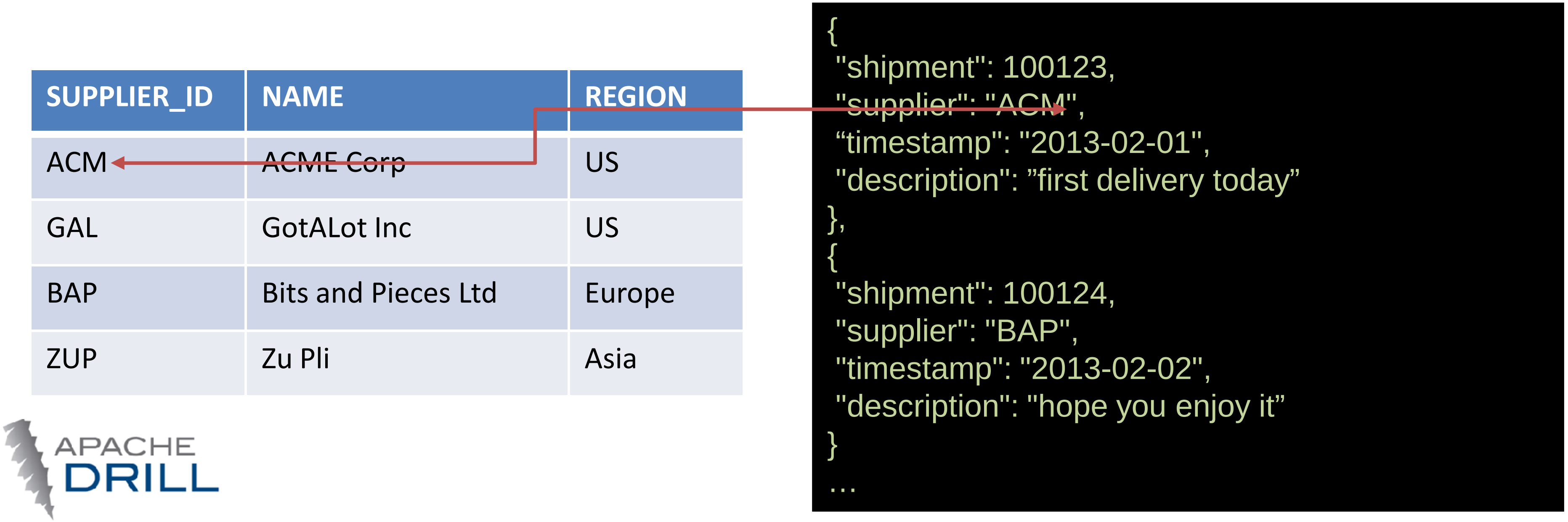
Use Case I

- Jane, a marketing analyst
- Determine target segments
- Data from different sources



Use Case II

- Logistics – supplier status
- Queries
  - How many shipments from supplier X?
  - How many shipments in region Y?



SUPPLIER_ID	NAME	REGION
ACM	ACME Corp	US
GAL	GotALot Inc	US
BAP	Bits and Pieces Ltd	Europe
ZUP	Zu Pli	Asia

```
{
 "shipment": 100123,
 "supplier": "ACM",
 “timestamp": "2013-02-01",
 "description": ”first delivery today”
},
{
 "shipment": 100124,
 "supplier": "BAP",
 "timestamp": "2013-02-02",
 "description": "hope you enjoy it”
}
…
```

Requirements

- Support for different data sources
- Support for different query interfaces
- Low-latency/real-time
- Ad-hoc queries
- Scalable, reliable

And now for something completely different …

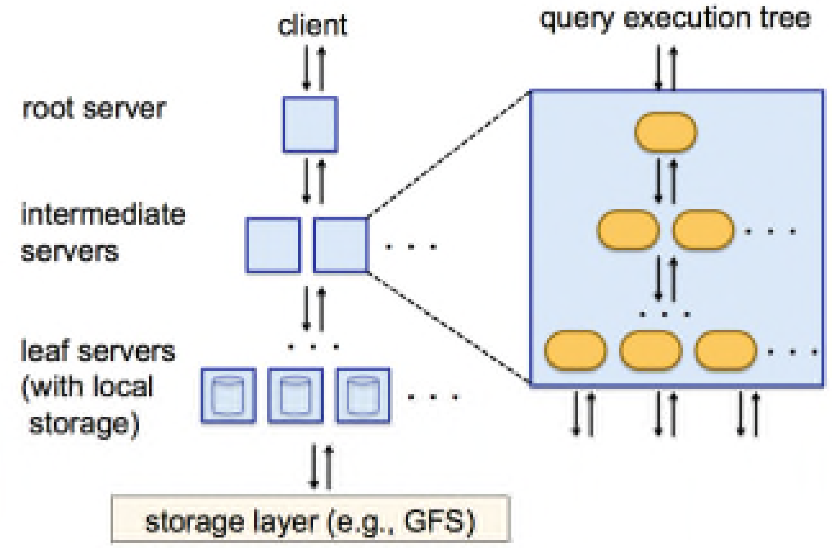
Google’s Dremel

> “
>
> Dremel is a scalable, interactive ad-hoc query system for analysis of read-only nested data. By combining multi-level execution trees and columnar data layout, it is capable of running aggregation queries over trillion-row tables in seconds. The system scales to thousands of CPUs and petabytes of data, and has thousands of users at Google.
>
> …
>
> ”

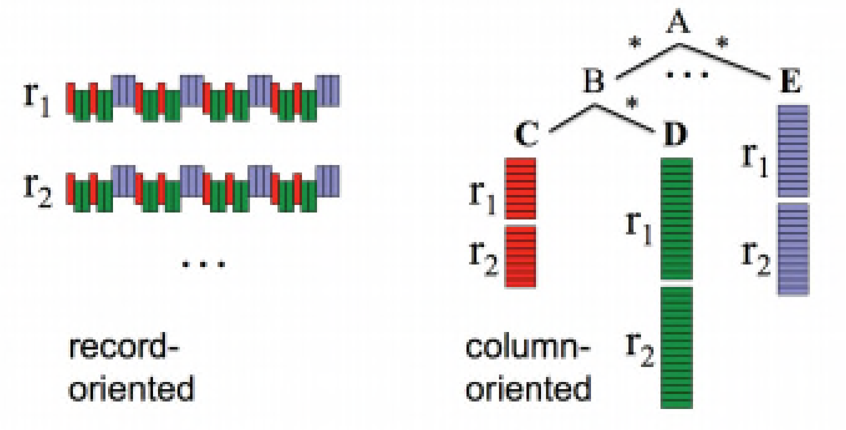
http://research.google.com/pubs/pub36632.html

Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, Theo Vassilakis, Proc. of the 36th Int'l Conf on Very Large Data Bases (2010), pp. 330-339

Google's Dremel

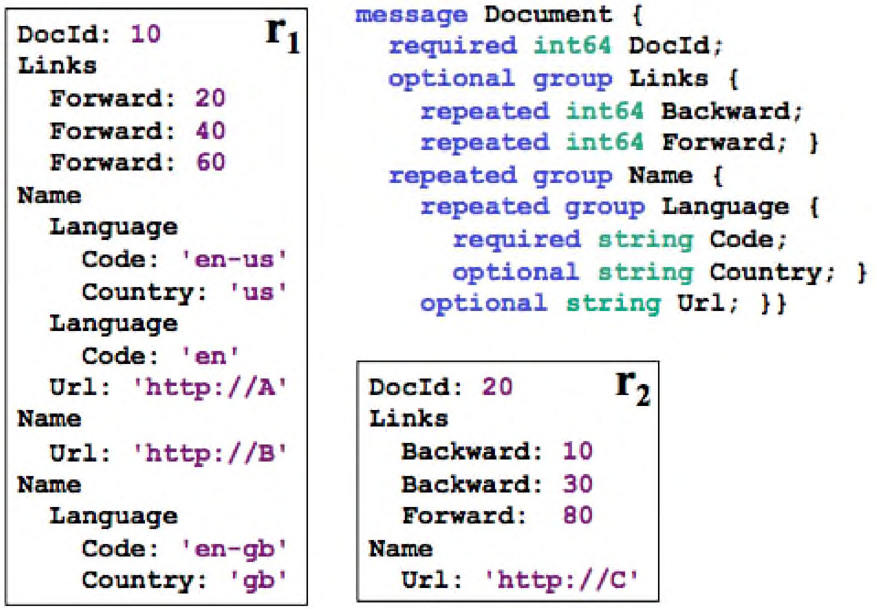


multi-level execution trees



columnar data layout

Google's Dremel

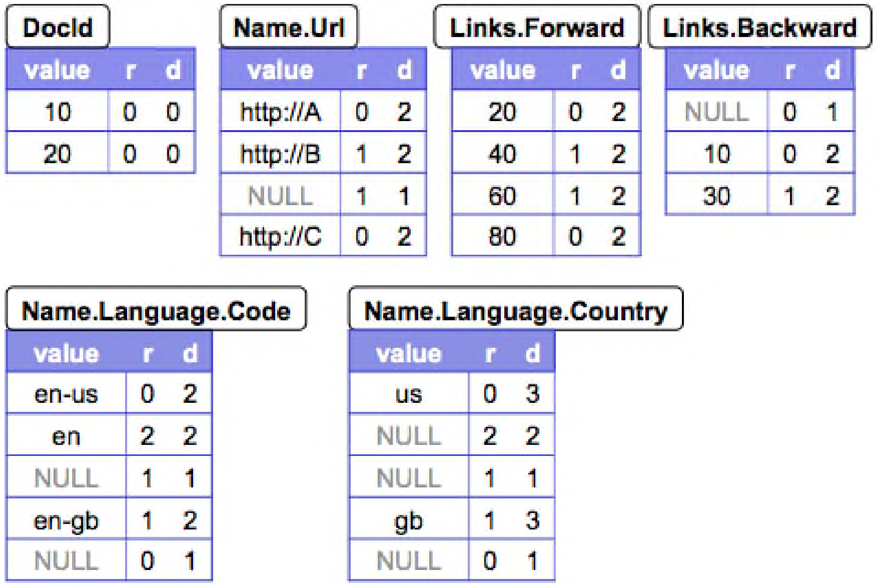


```
DocId: 10                 r1
Links
  Forward: 20
  Forward: 40
  Forward: 60
Name
  Language
    Code: 'en-us'
    Country: 'us'
  Language
    Code: 'en'
  Url: 'http://A'
Name
  Url: 'http://B'
Name
  Language
    Code: 'en-gb'
    Country: 'gb'
```

```
message Document {
  required int64 DocId;
  optional group Links {
    repeated int64 Backward;
    repeated int64 Forward; }
  repeated group Name {
    repeated group Language {
      required string Code;
      optional string Country; }
    optional string Url; }}
```

```
DocId: 20                 r2
Links
  Backward: 10
  Backward: 30
  Forward:  80
Name
  Url: 'http://C'
```

nested data + schema



DocId

value	r	d
10	0	0
20	0	0

Name.Url

value	r	d
http://A	0	2
http://B	1	2
NULL	1	1
http://C	0	2

Links.Forward

value	r	d
20	0	2
40	1	2
60	1	2
80	0	2

Links.Backward

value	r	d
NULL	0	1
10	0	2
30	1	2

Name.Language.Code

value	r	d
en-us	0	2
en	2	2
NULL	1	1
en-gb	1	2
NULL	0	1

Name.Language.Country

value	r	d
us	0	3
NULL	2	2
NULL	1	1
gb	1	3
NULL	0	1

column-striped representation

mapping nested data to tables

Google’s Dremel

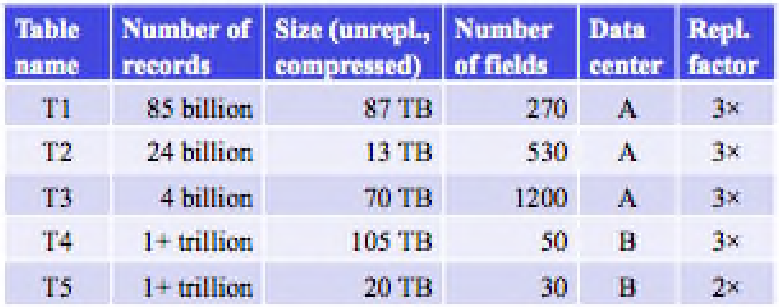
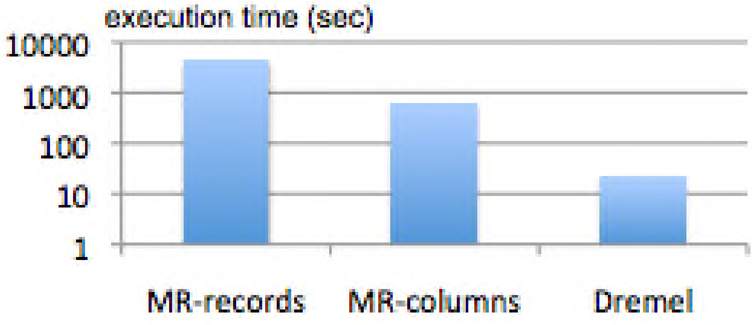
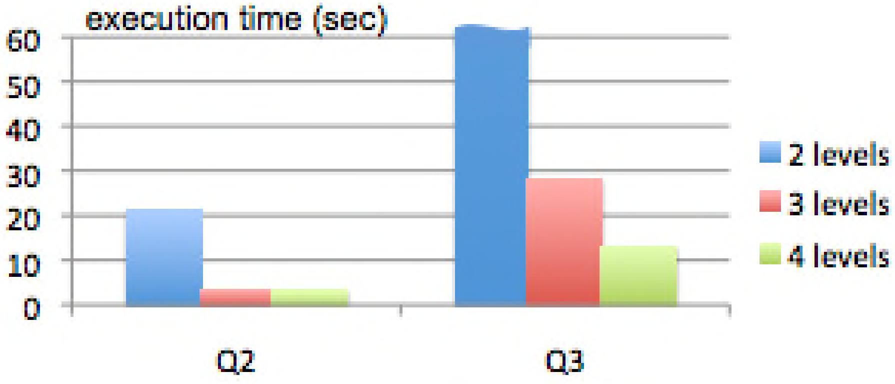


Table name	Number of records	Size (unrepl., compressed)	Number of fields	Data center	Repl. factor
T1	85 billion	87 TB	270	A	3×
T2	24 billion	13 TB	530	A	3×
T3	4 billion	70 TB	1200	A	3×
T4	1+ trillion	105 TB	50	B	3×
T5	1+ trillion	20 TB	30	B	2×

experiments:
datasets & query performance



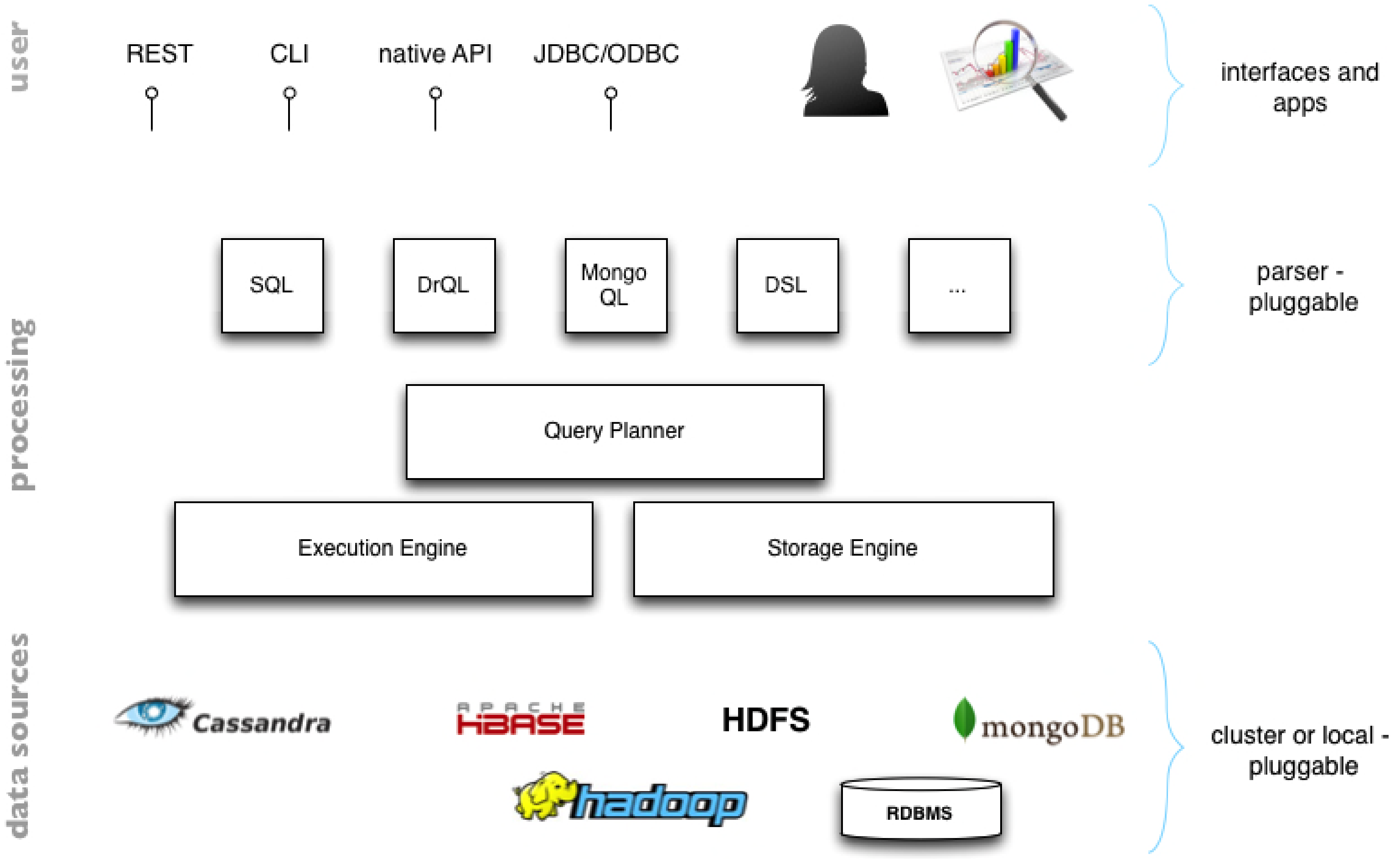


Back to Apache Drill ...

Apache Drill–key facts

- Inspired by Google’s **Dremel**
- Standard **SQL 2003** support
- Plug-able **data sources**
- **Nested data** is a first-class citizen
- **Schema** is **optional**
- **Community** driven, **open**, 100’s involved

High-level Architecture



user

REST

CLI

native API

JDBC/ODBC

interfaces and apps

processing

SQL

DrQL

Mongo QL

DSL

...

parser - pluggable

Query Planner

Execution Engine

Storage Engine

data sources

Cassandra

HBASE

HDFS

mongoDB

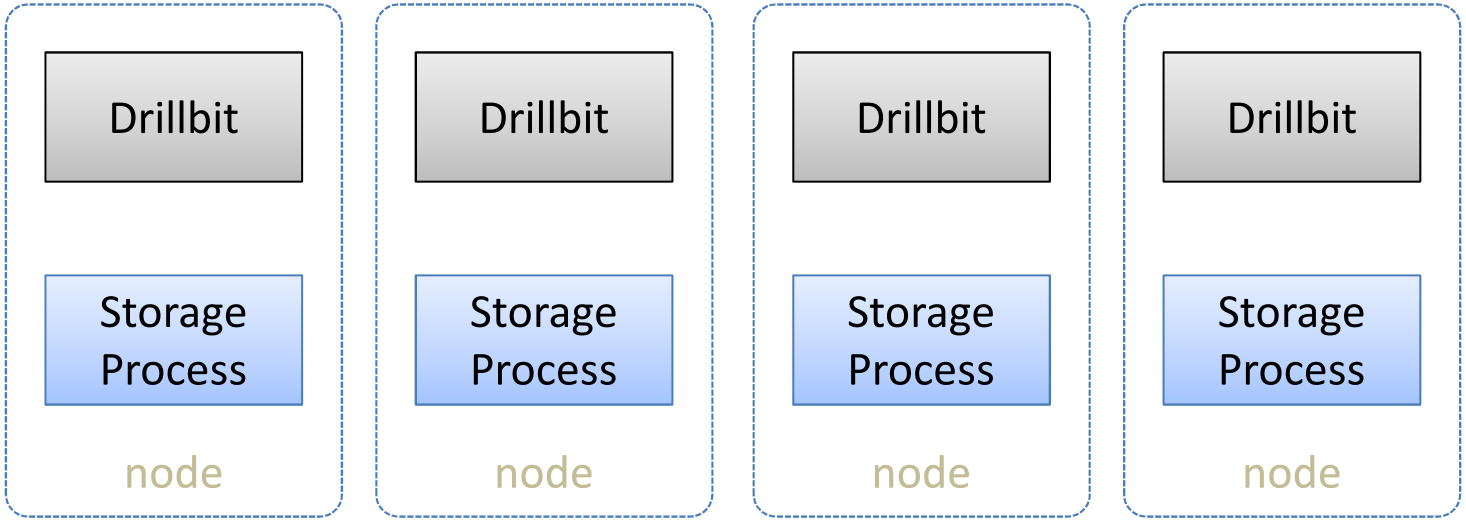
hadoop

RDBMS

cluster or local - pluggable

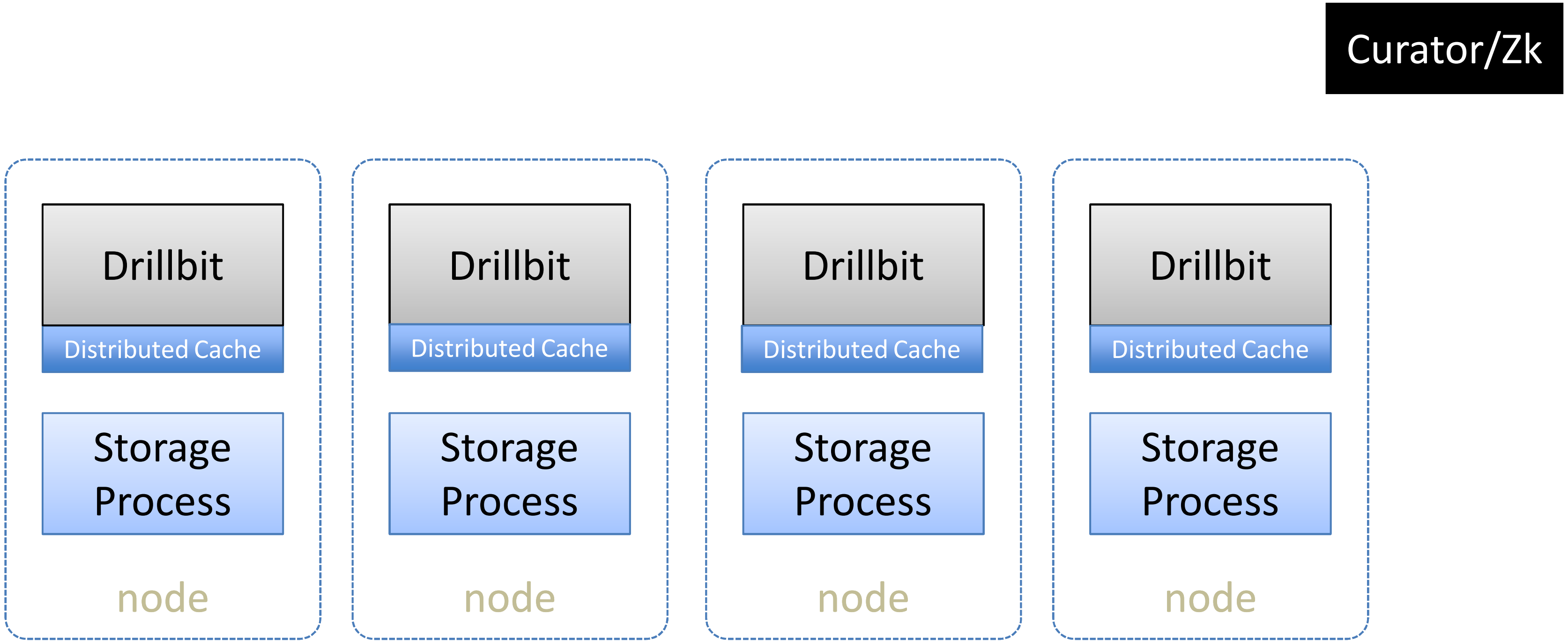
Wire-level Architecture

- Each node: **Drillbit** - maximize data locality
- Co-ordination, query planning, execution, etc, are **distributed**
- By default Drillbits hold all roles
- Any node can act as endpoint for a query



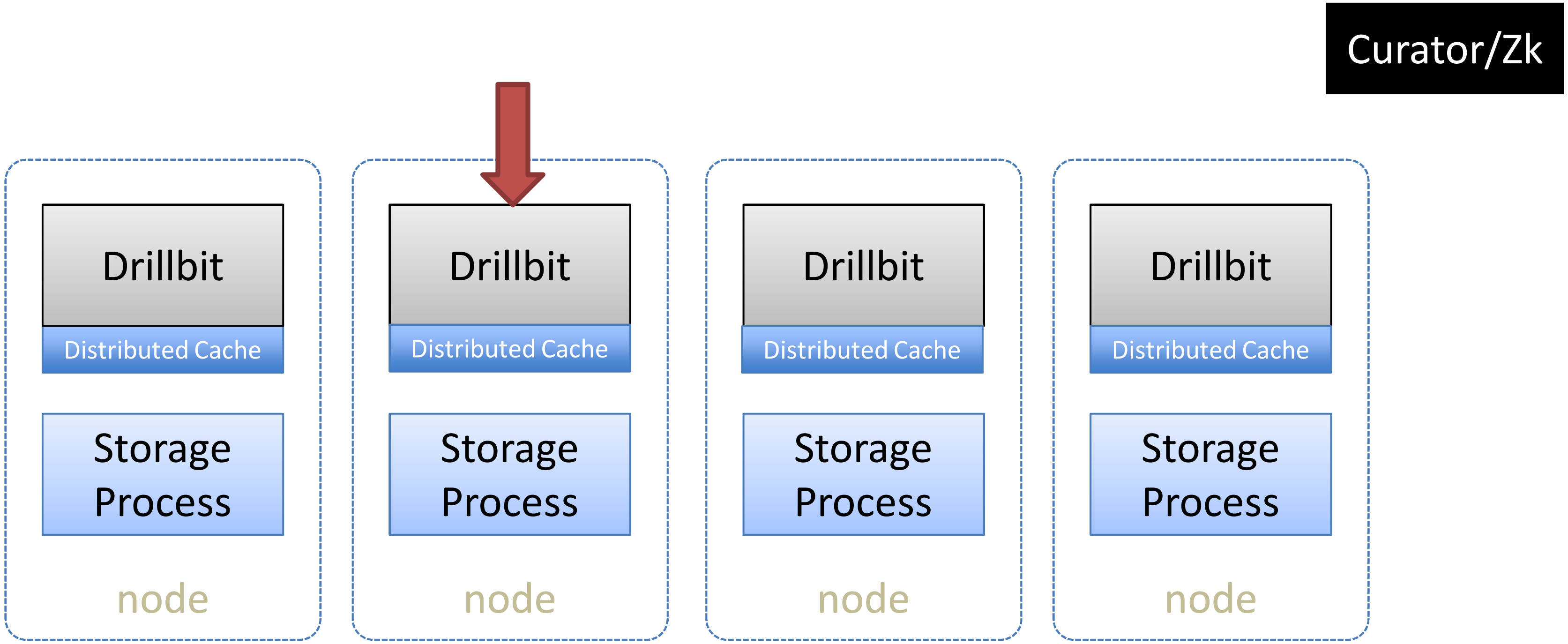
Wire-level Architecture

- **Zookeeper** for ephemeral cluster membership info
- **Distributed cache** (Hazelcast) for metadata, locality information, etc.

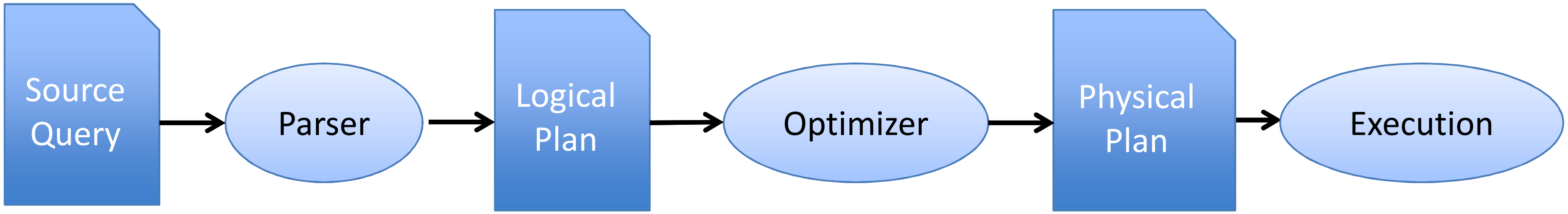


Wire-level Architecture

- **Originating Drillbit** acts as foreman, manages query execution, scheduling, locality information, etc.
- Streaming data **communication** avoiding SerDe



Principled Query Execution



SQL 2003
DrQL
MongoQL
DSL

parser API

```
query: [
 {
  @id: "log",
  op: "sequence",
  do: [
   {
    op: "scan",
    source: “logs”
   },
   {
    op:
      "filter",
    condition:
      "x > 3”
   },
```

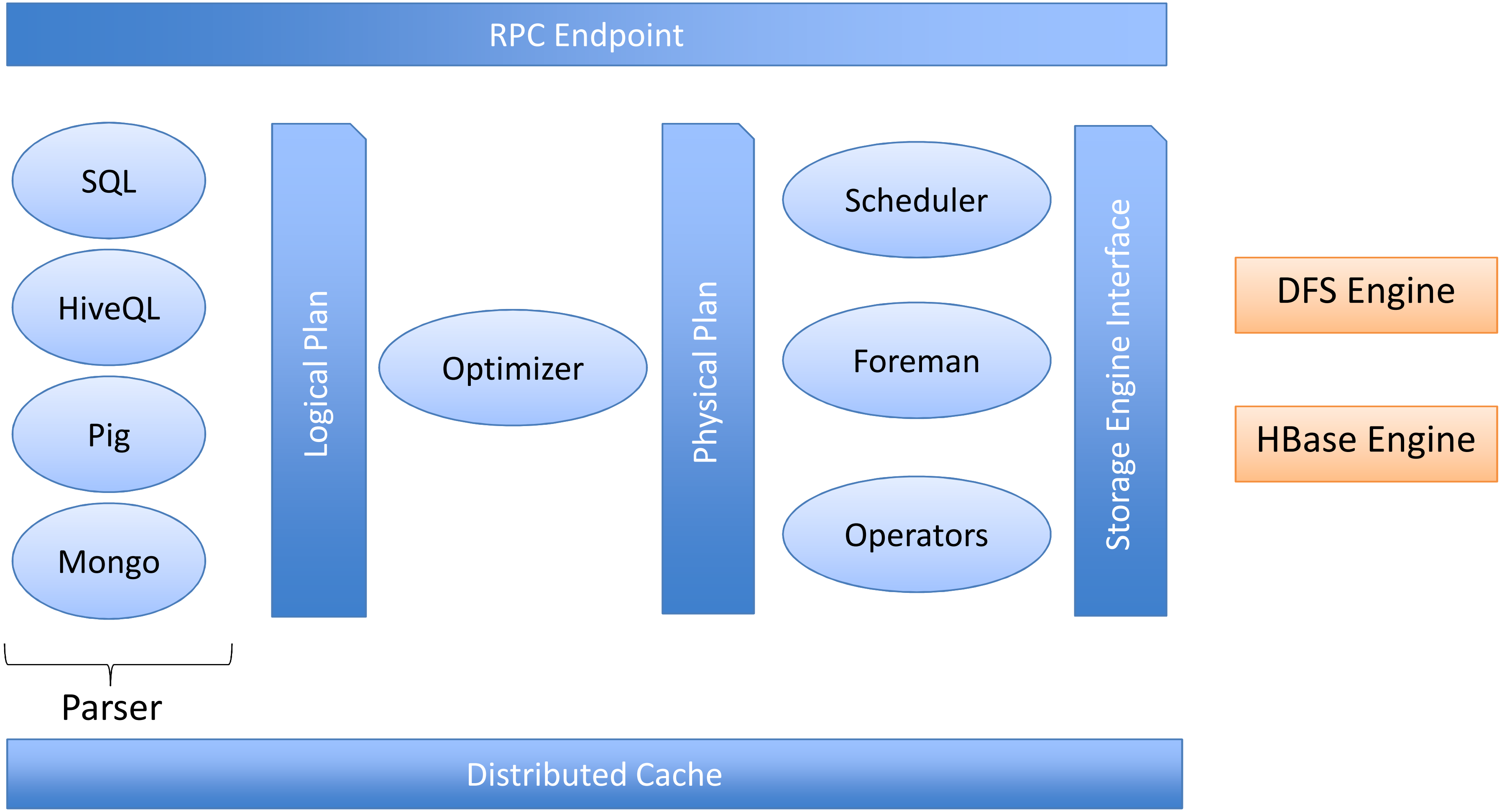
topology

scanner API

APACHE DRILL

MAPR TECHNOLOGIES

Drillbit Modules

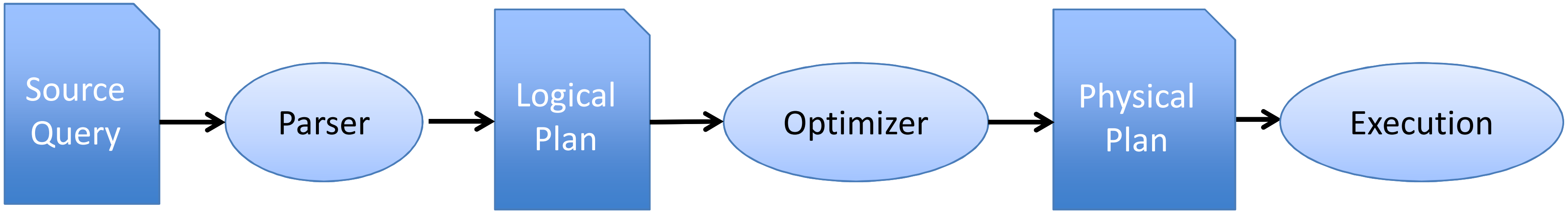


Key features

- Full SQL – ANSI SQL 2003
- Nested Data as first class citizen
- Optional Schema
- Extensibility Points …

Extensibility Points

- Source query → parser API
- Custom operators, UDF → logical plan
- Serving tree, CF, topology → physical plan/optimizer
- Data sources &formats → scanner API



… and Hadoop?

- HDFS can be a data source
- Complementary use cases*
- … use Apache Drill
  - Find record with specified condition
  - Aggregation under dynamic conditions
- … use MapReduce
  - Data mining with multiple iterations
  - ETL

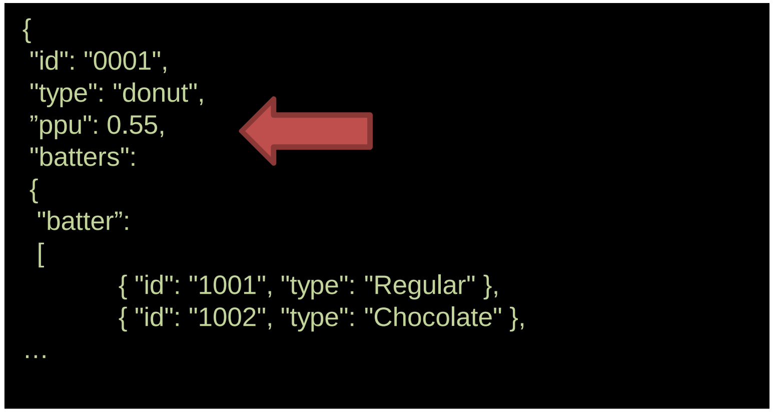
An Inside Look at Google BigQuery

Table of Contents

Google

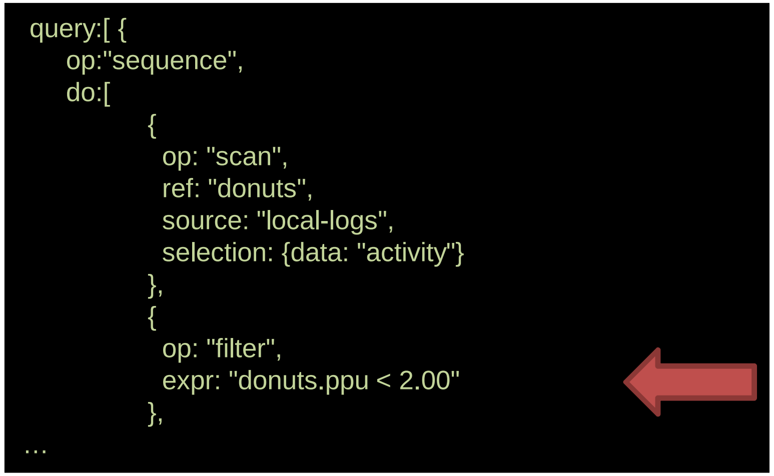
*) https://cloud.google.com/files/BigQueryTechnicalWP.pdf

Basic Demo



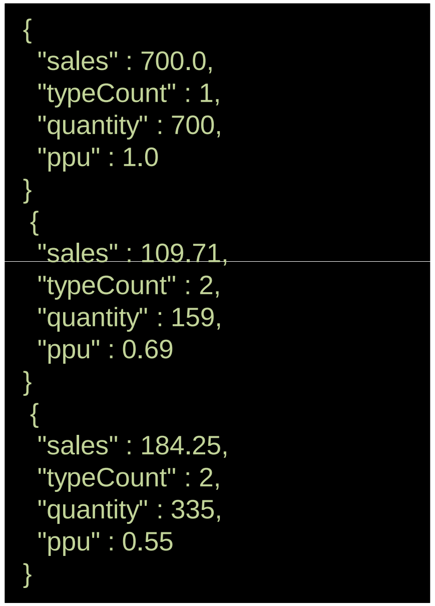
```
{
 "id": "0001",
 "type": "donut",
 ”ppu": 0.55,
 "batters":
 {
  "batter":
  [
           { "id": "1001", "type": "Regular" },
           { "id": "1002", "type": "Chocolate" },
…
```

data source: **donuts.json**



```
query:[ {
     op:"sequence",
     do:[
          {
            op: "scan",
            ref: "donuts",
            source: "local-logs",
            selection: {data: "activity"}
          },
          {
            op: "filter",
            expr: "donuts.ppu < 2.00"
          },
…
```

logical plan: **simple_plan.json**



```
{
  "sales" : 700.0,
  "typeCount" : 1,
  "quantity" : 700,
  "ppu" : 1.0
}
 {
  "sales" : 109.71,
  "typeCount" : 2,
  "quantity" : 159,
  "ppu" : 0.69
}
 {
  "sales" : 184.25,
  "typeCount" : 2,
  "quantity" : 335,
  "ppu" : 0.55
}
```

result: **out.json**

https://cwiki.apache.org/confluence/display/DRILL/Demo+HowTo

BE A PART OF IT!

Status

- Heavy development by multiple organizations

- Available
  - Logical plan (ADSP)
  - Reference interpreter
  - Basic SQL parser
  - Basic demo

Status

May 2013

- Full SQL support (+JDBC)
- Physical plan
- In-memory compressed data interfaces
- Distributed execution
- HBase and MySQL storage engine
- WebUI client

Contributing

Contributions appreciated (besides code drops)!

- Test data & test queries
- Use case scenarios (textual/SQL queries)
- Documentation
- Further schedule
  - Alpha Q2
  - Beta Q3

Kudos to …

- Julian Hyde, Pentaho
- Lisen Mu, XingCloud
- Tim Chen, Microsoft
- Chris Merrick, RJMetrics
- David Alves, UT Austin
- Sree Vaadi, SSS/NGData
- Jacques Nadeau, MapR
- Ted Dunning, MapR

Engage!

- Follow @ApacheDrill on Twitter
- Sign up at mailing lists (user | dev)
  http://incubator.apache.org/drill/mailing-lists.html
- Standing G+ hangouts every Tuesday at 5pm GMT
  http://j.mp/apache-drill-hangouts
- Keep an eye on http://drill-user.org/

Apache Drill User

KEEPING TRACK OF APACHE DRILL. FROM A GEEKS, FOR GEEKS.