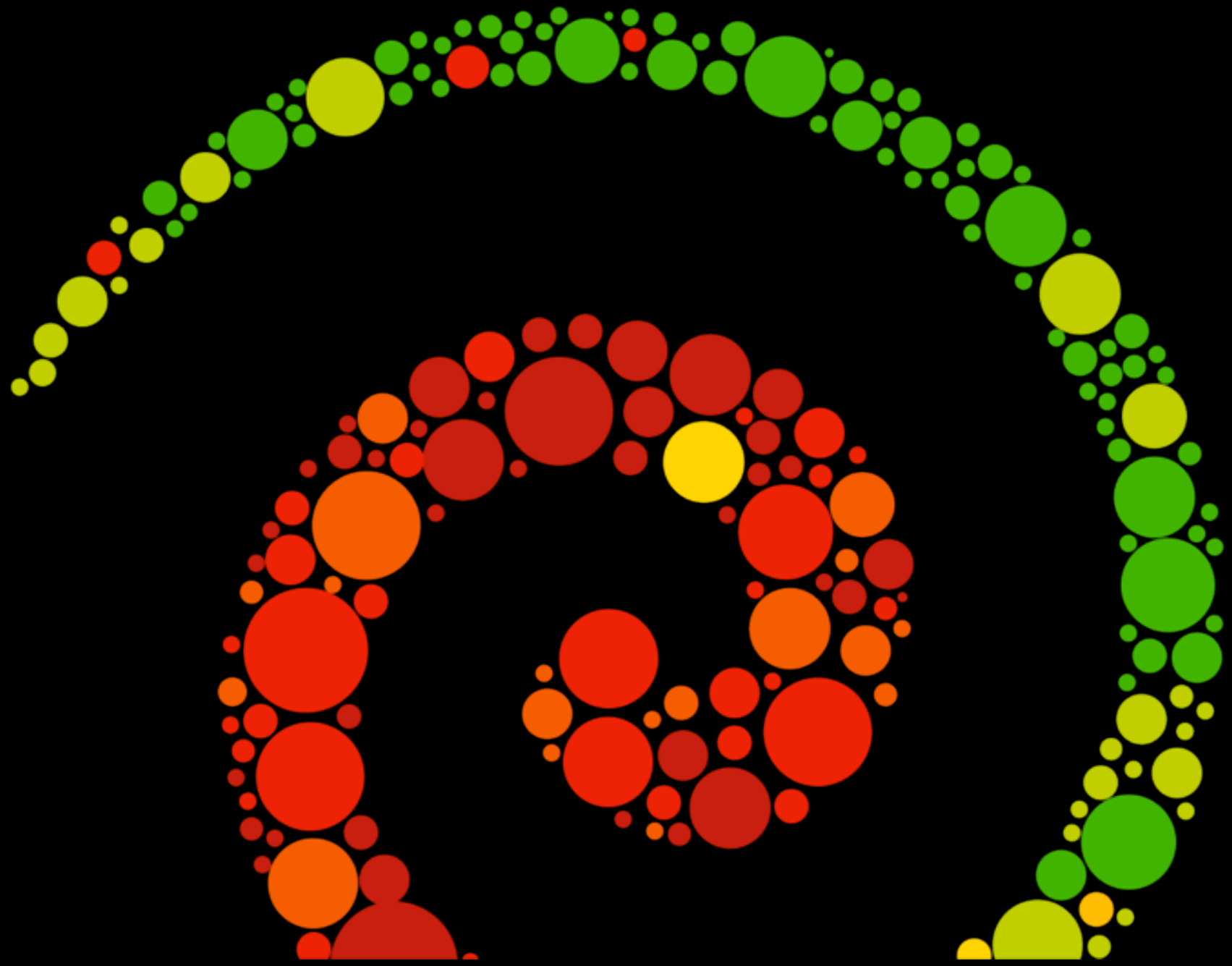
2013

Pentaho Community Meeting

Sintra . Portugal



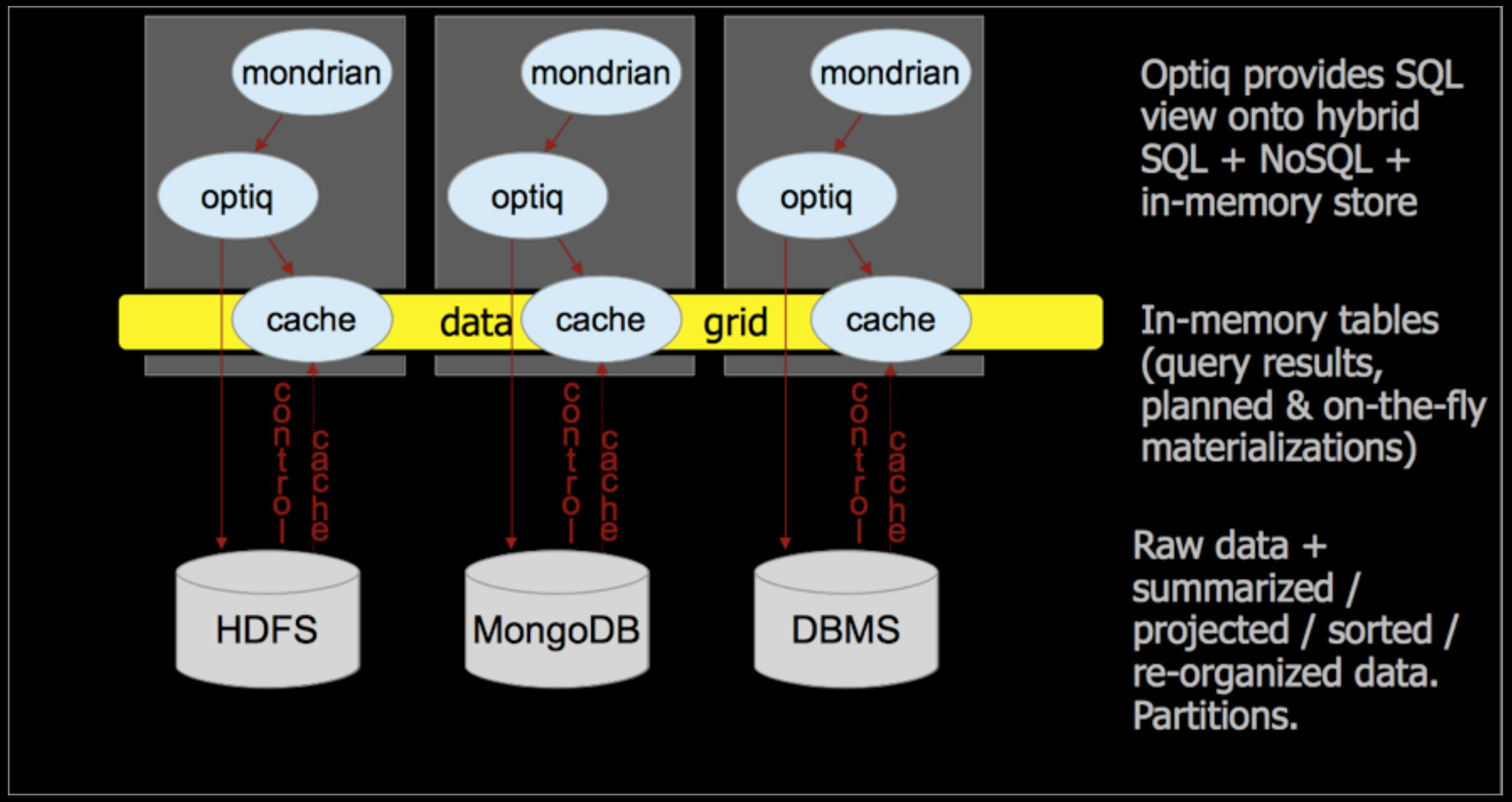
Optiq

A dynamic data management framework

@julianhyde

3 things

1. Databases are good.
2. It is hard to build analytics if your data is "all over the place."
3. Optiq makes heterogeneous data look and behave more like a database.



Mondrian on Optiq on hybrid data + memory

Big Data

Friday, October 4, 13

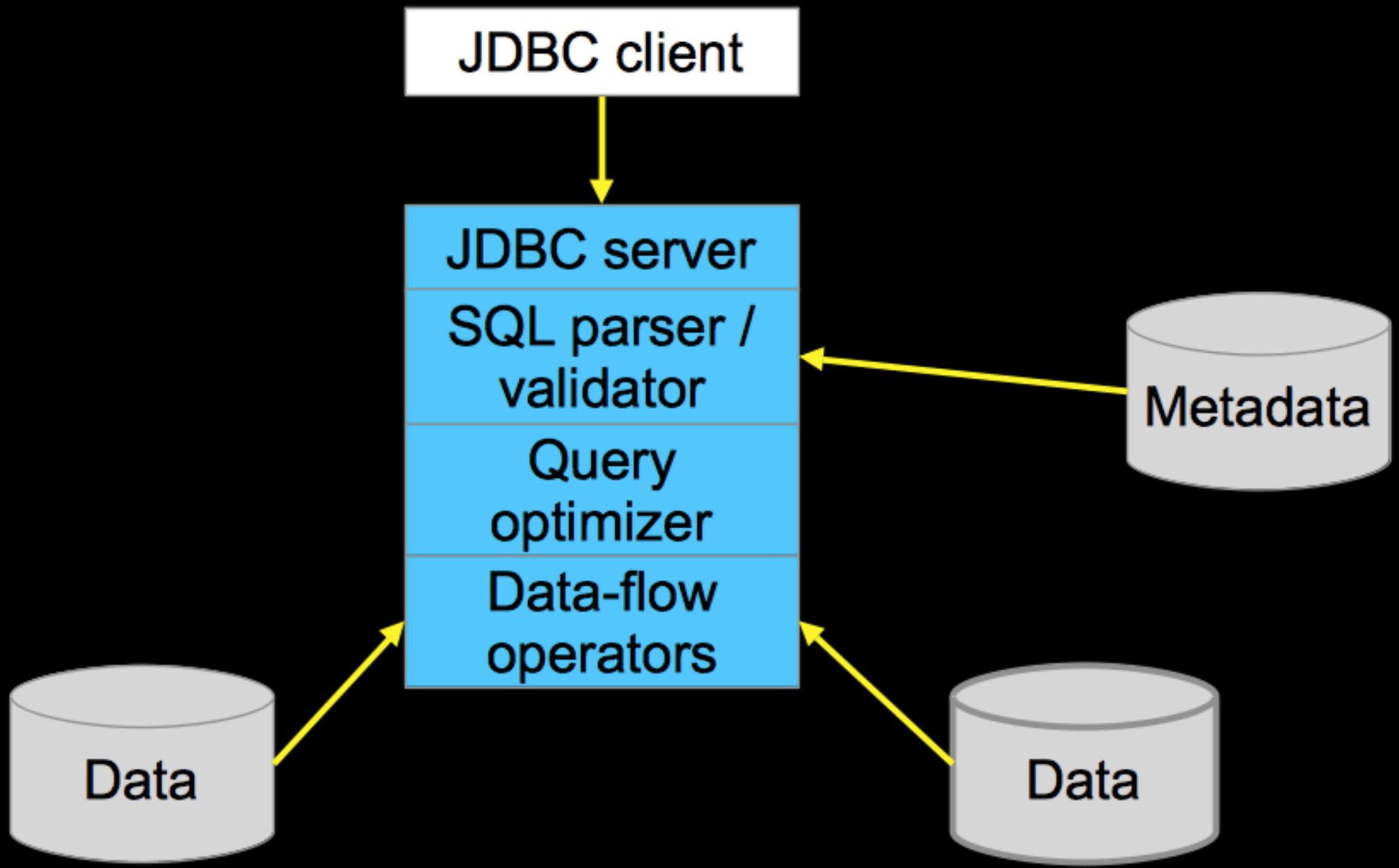
“Data all over the place”

- Different locations (HDFS, memory, DBMS)
- Different formats
- Different workloads
  - Transactional: Mainly write, targeted read
  - Analytic: Mainly read (bulk write)
- Query latency: Interactive vs. batch
- Data latency: Can we show out-of-date data?

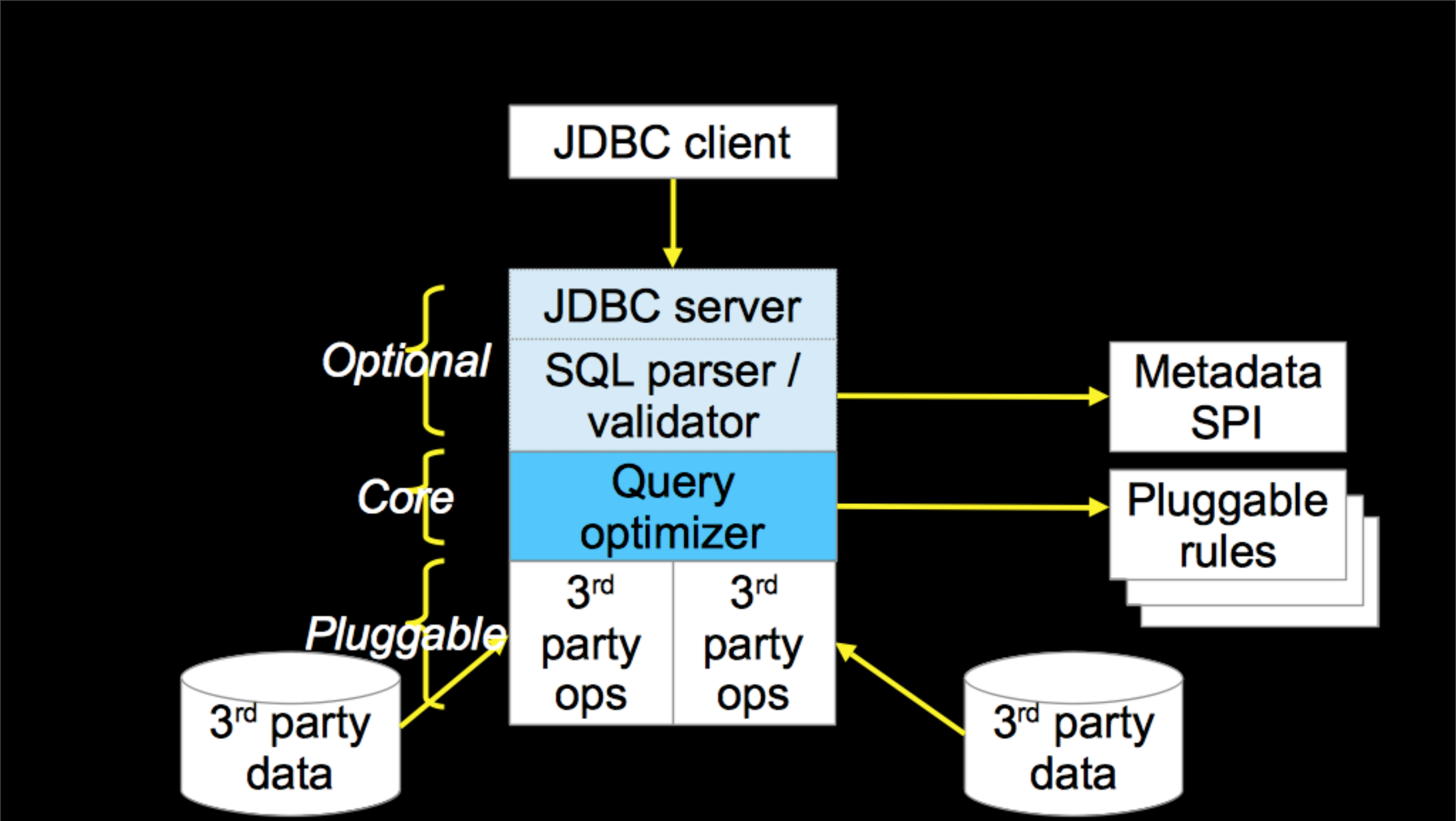
Databases are good

- Central point to manage data
- Simple, standard API for apps
- Powerful modeling techniques (e.g. star schemas)
- Data independence (i.e. tune your data after you write your application)
- Query optimization

Optiq



Conventional DB architecture



Optiq architecture

Examples

Example #1: CSV

- Uses CSV adapter (`optiq-csv`)
- Demo using `sqlline`
- Easy to run this for yourself:

```
$ git clone https://github.com/julianhyde/optiq-csv
$ cd optiq-csv
$ mvn install
$ ./sqlline
```

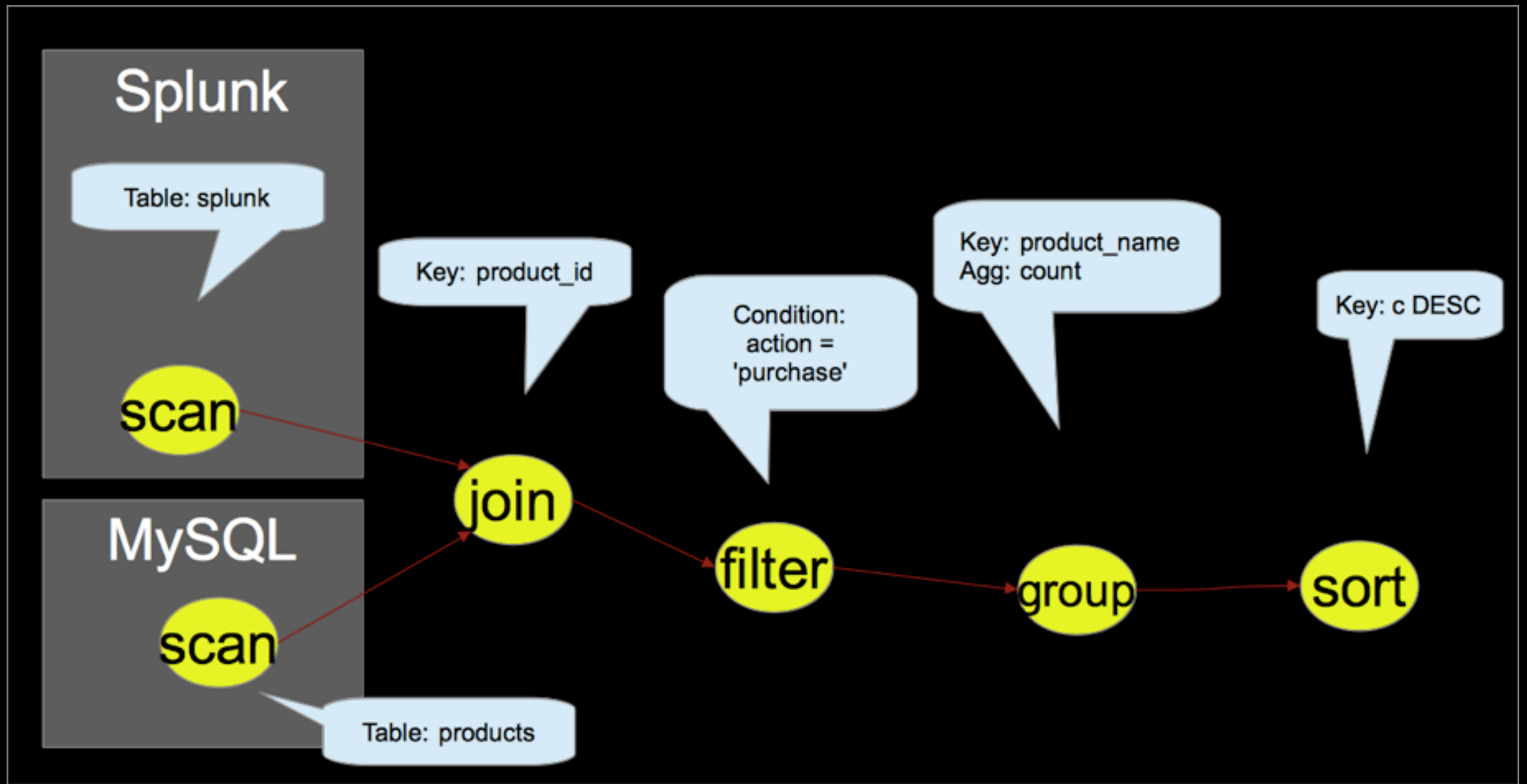
More adapters

Adapters	Embedded	Planned
CSV	Cascading (Lingual)	HBase (Phoenix)
JDBC	Apache Drill	Spark
MongoDB		Cassandra
Splunk		Mondrian
linq4j		

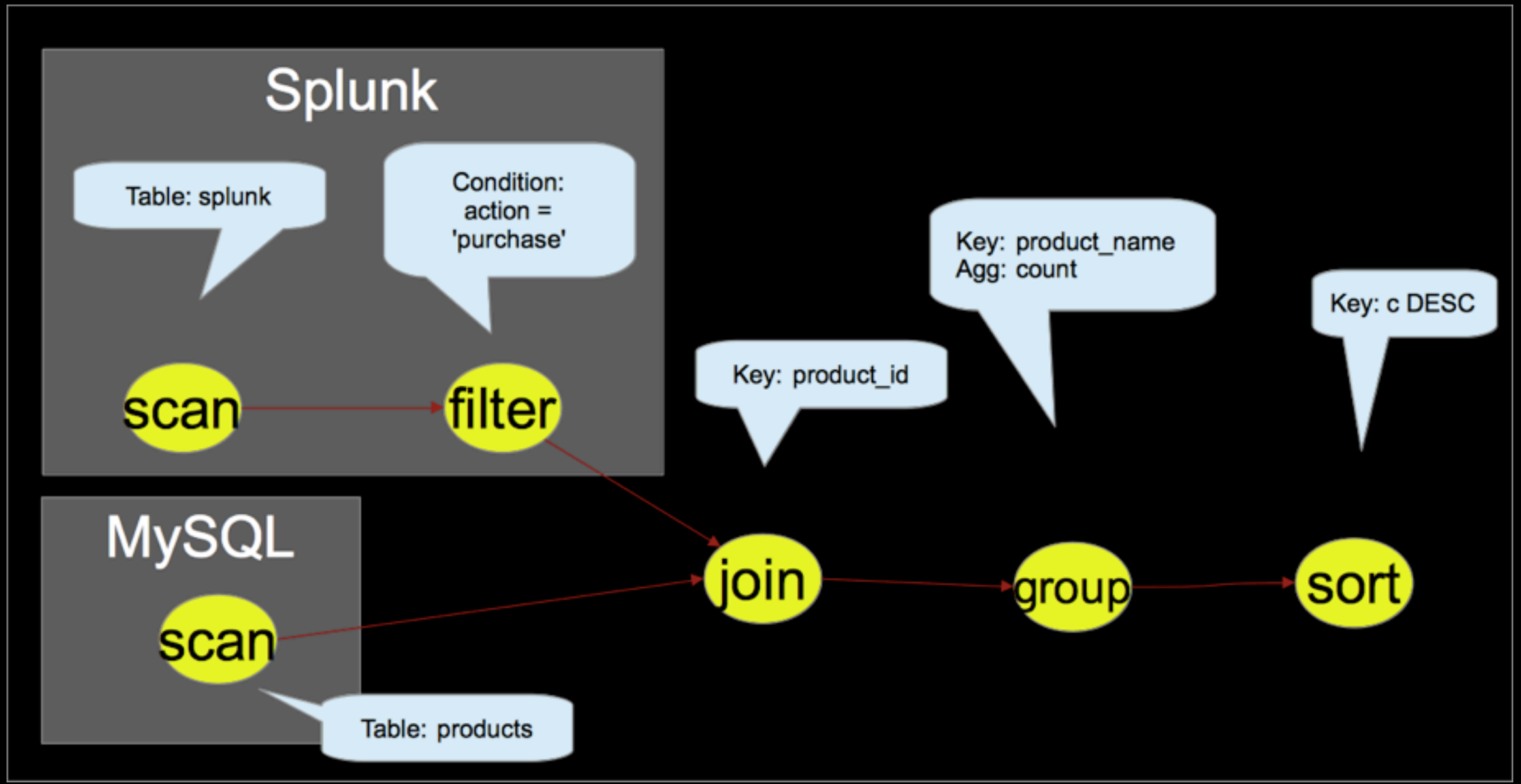
Example #2: Splunk + MySQL

```
SELECT p.product_name,
    COUNT(*) AS c
FROM splunk.splunk AS s
    JOIN mysql.products AS p
    ON s.product_id = p.product_id
WHERE s.action = 'purchase'
GROUP BY p.product_name
ORDER BY c DESC
```

Expression tree



Optimized tree



Analytics on heterogeneous data

Simple analytics problem

- 100M U.S. census records
- 1KB each record, 100GB total
- 4 SATA3 disks, total 1.2GB/s
- How to count all records in under 5s?

Simple analytics problem

- 100M U.S. census records
- 1KB each record, 100GB total
- 4 SATA3 disks, total 1.2GB/s
- How to count all records in under 5s?

Simple analytics problem

- 100M U.S. census records
- 1KB each record, 100GB total
- 4 SATA3 disks, total 1.2GB/s
- How to count all records in under 5s?

- Not possible?! It takes 80s just to read the data.

Solution: Cheat!

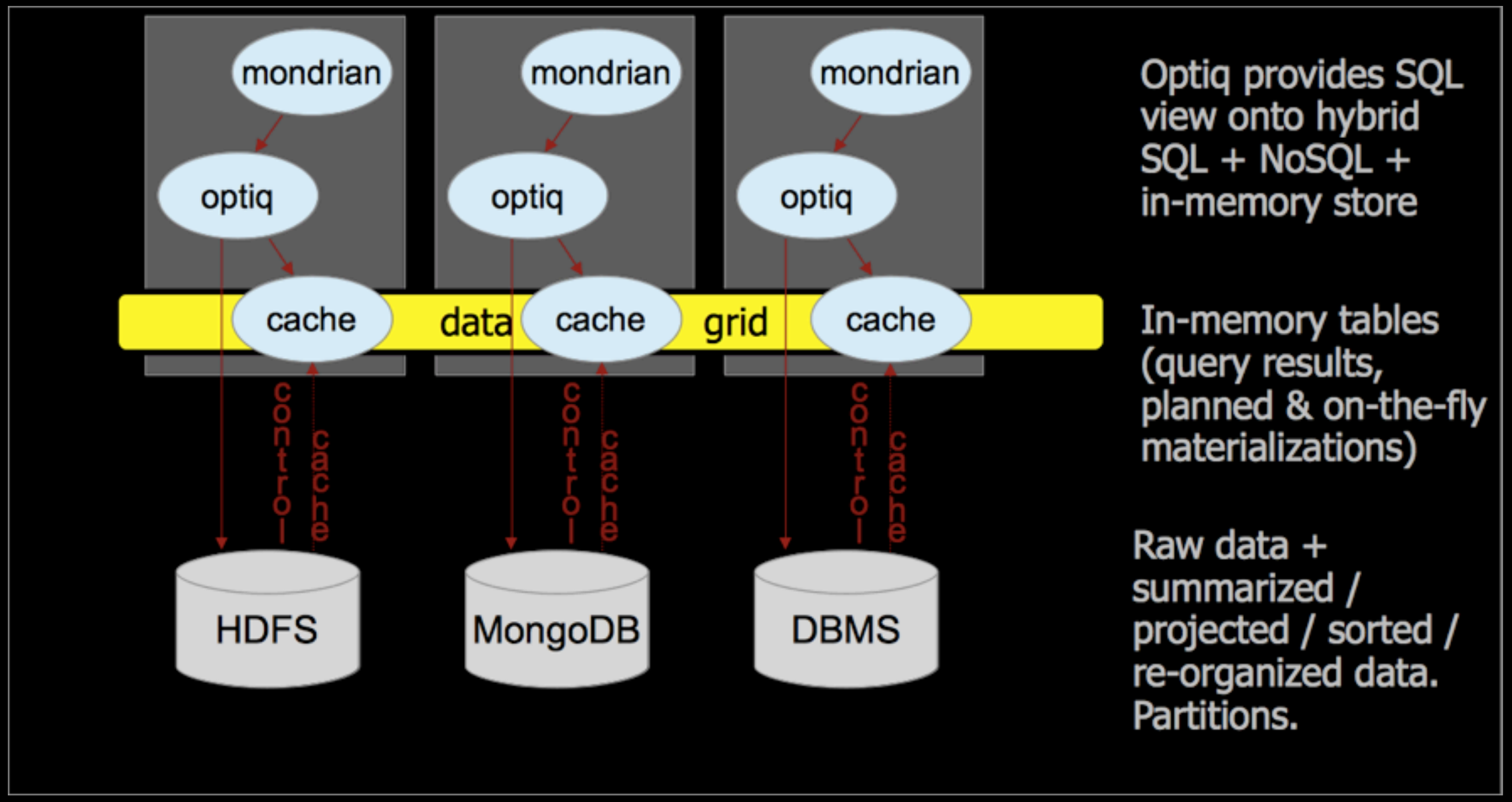
Solution: Cheat!

- Compress data
- Column-oriented storage
- Store data in sorted order
- Put data in memory
- Cache previous query results
- Pre-compute (materialize) aggregates

How Optiq helps you to cheat

How Optiq helps you to cheat

- Materialized views
  - Pre-defined aggregate tables
  - Cached query results = In-memory tables
- Smart cache maintenance
  - Quickly bring materializations online & offline
  - Materializations over a subset of the data
- Spark distributed, in-memory processing & cache
- Application thinks it is talking to a single SQL database



Mondrian on Optiq on hybrid data + memory

Summary

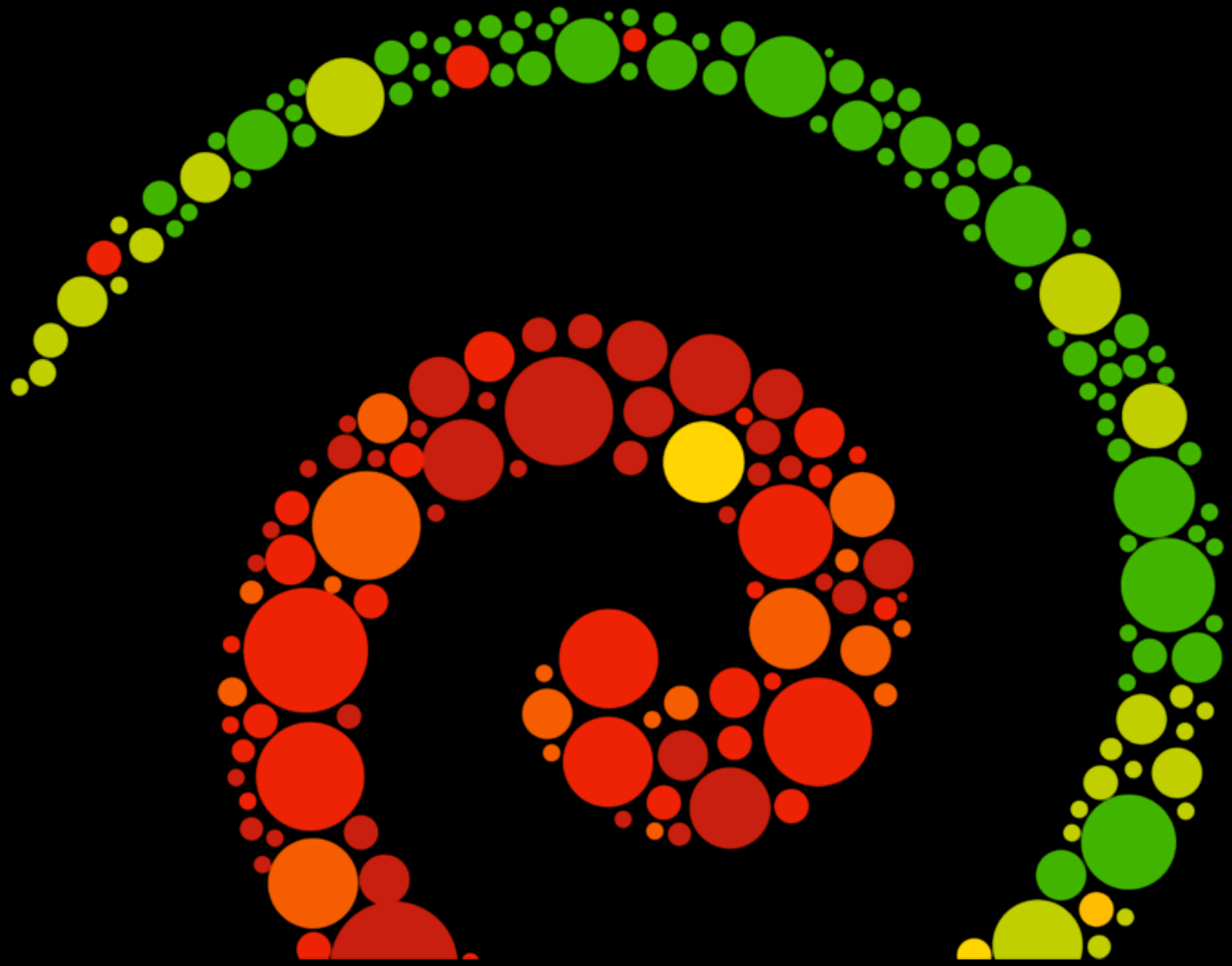
3 things (reprise)

1. **Databases are good.** Especially the flexibility that SQL gives us.
2. **It is hard to build analytics if your data is all over the place.** Different workloads (operational vs. analytic, small write vs. bulk read) require different data structures.
3. **Optiq is not a database**. But Optiq creates a federated data architecture that performs well, and looks like a database to your tools.

2013

Pentaho Community Meeting

Sintra . Portugal



@julianhyde

optiq https://github.com/julianhyde/optiq
mondrian http://mondrian.pentaho.com
blog http://julianhyde.blogspot.com