Optiq: a SQL front-end for everything

Julian Hyde @julianhyde

http://github.com/julianhyde/optiq
http://github.com/julianhyde/optiq-splunk

Pentaho Community Meetup
Amsterdam, 2012

http://www.flickr.com/photos/torkildr/3462606643


http://www.flickr.com/photos/sylvar/31436961/

“Big Data”

Right data, right time

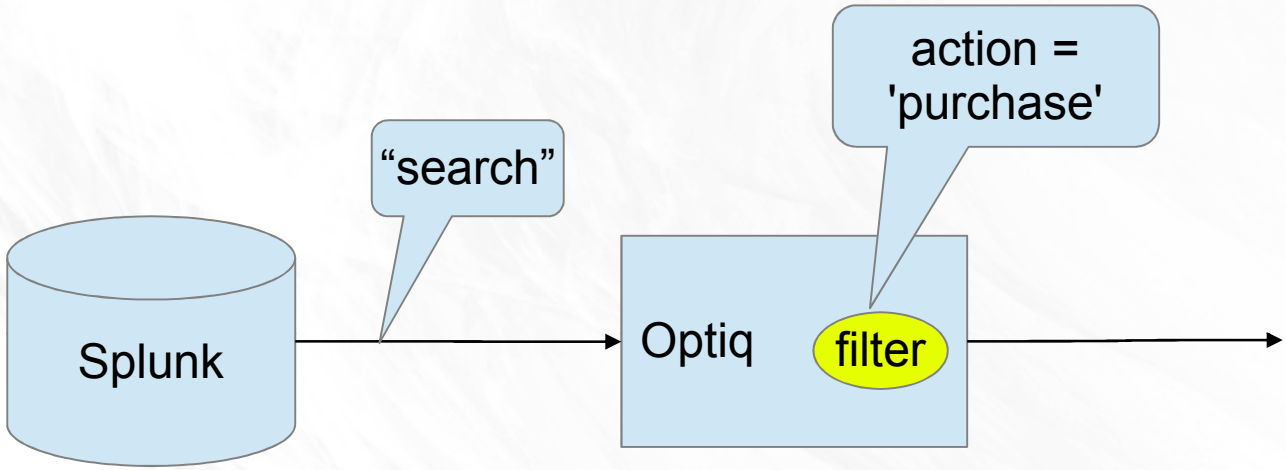
Diverse data sources / Performance / Suitable format

Use case: Splunk

- NoSQL database
- Every log file in the enterprise
- A single “table”
- A record for every line in every log file
- A column for every field that exists in any log file
- No schema

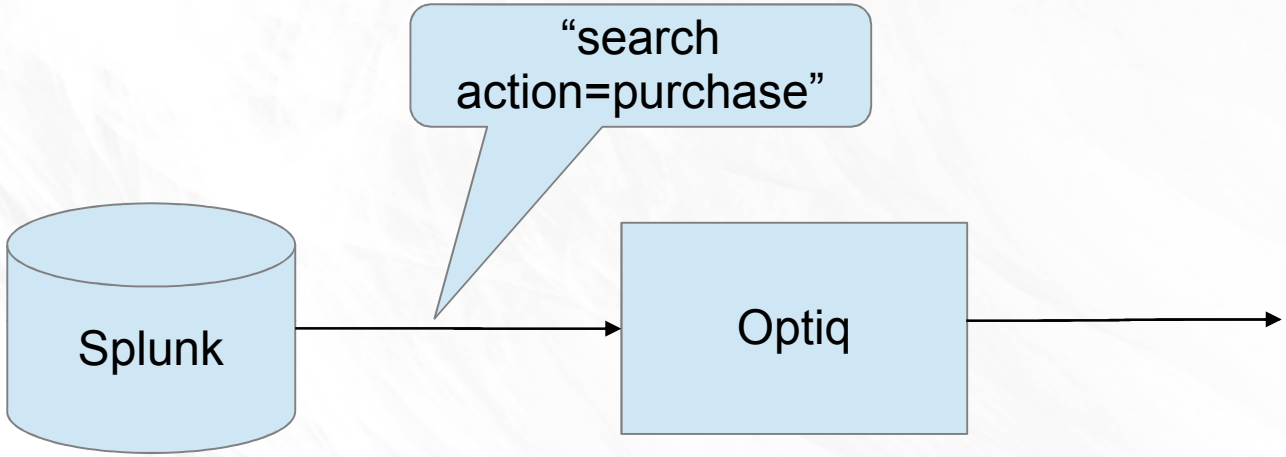
```
SELECT “source”, “product_id”, “http_code”
   FROM “splunk”.”splunk”
   WHERE “action” = 'purchase'
```

How do it (wrong)




```
SELECT “source”, “product_id”
FROM “splunk”.”splunk”
WHERE “action” = 'purchase'
```

How do it (right)




```
SELECT “source”, “product_id”
FROM “splunk”.”splunk”
WHERE “action” = 'purchase'
```

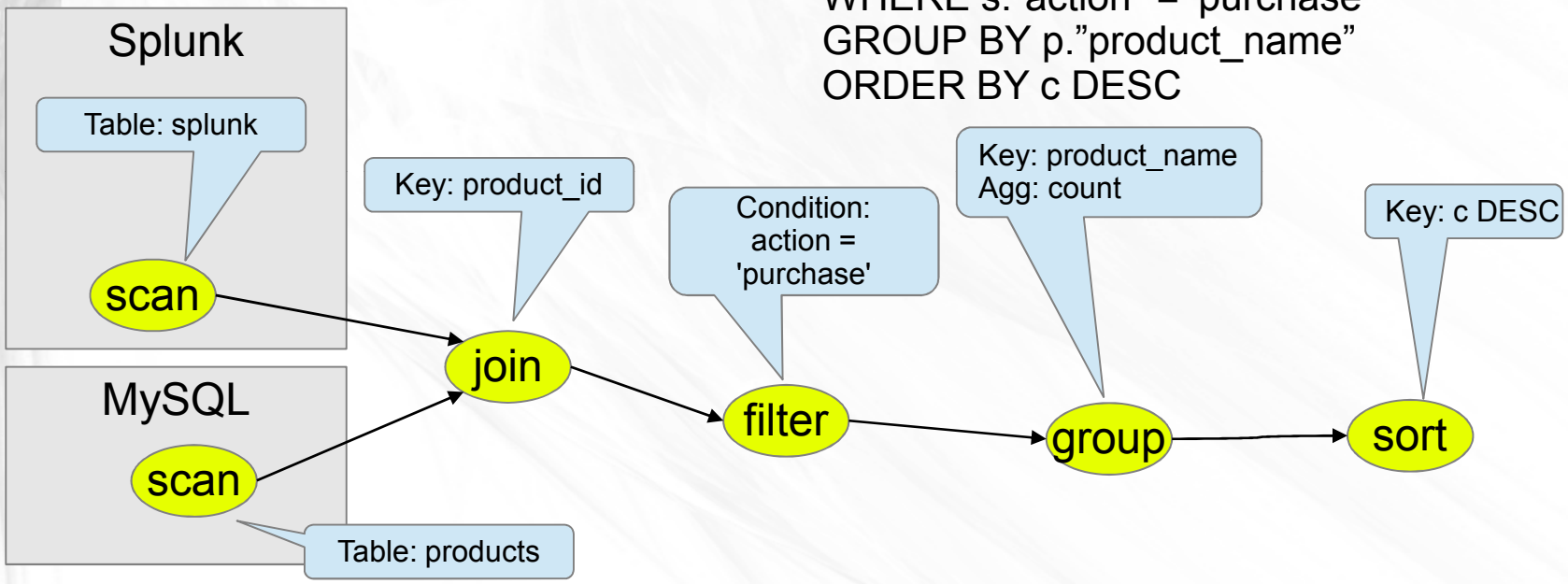
Example #2

Combining data from 2 sources (Splunk & MySQL)

Also possible: 3 or more sources; 3-way joins; unions

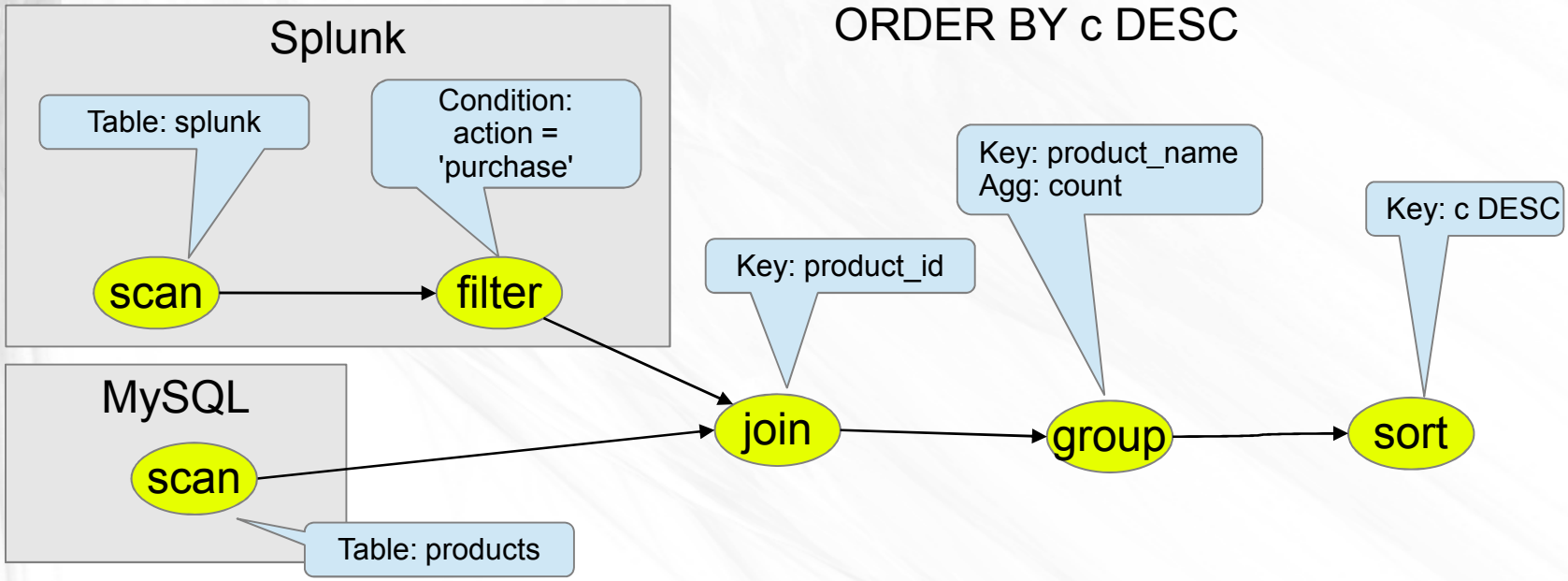
Expression tree

```
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)

```
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
```

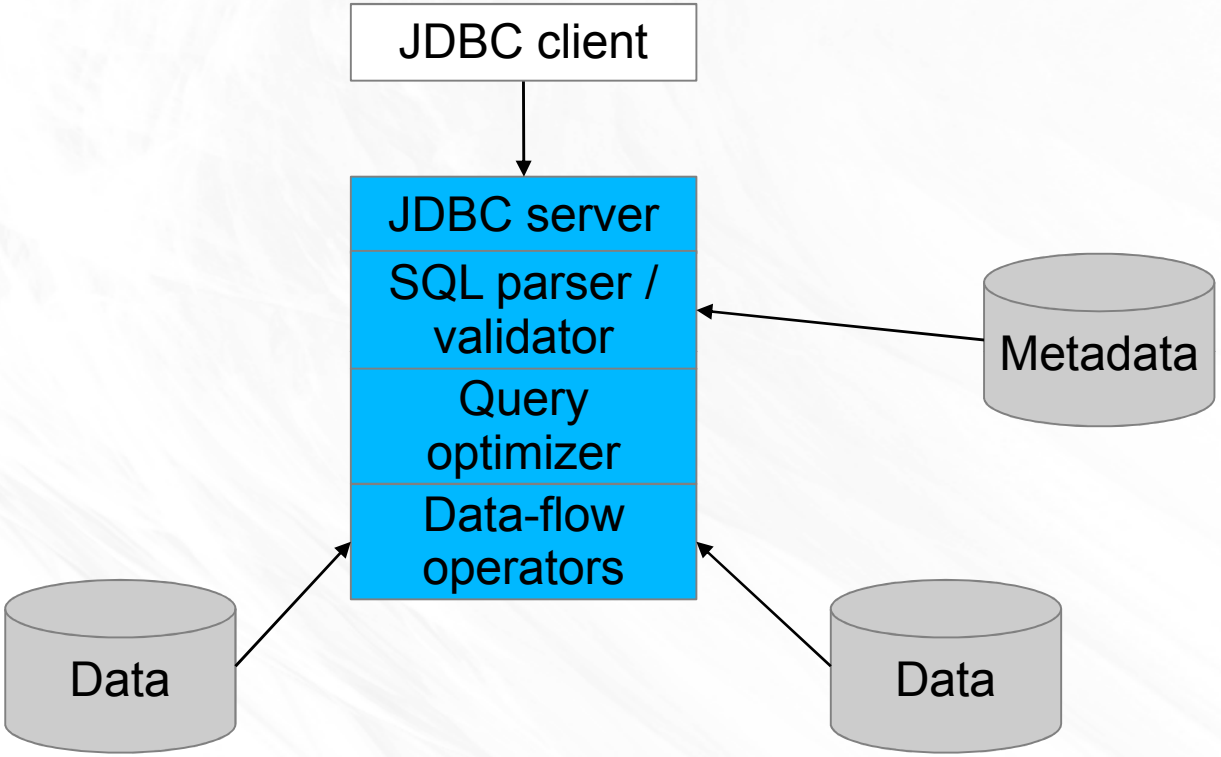


Optiq is not a database.

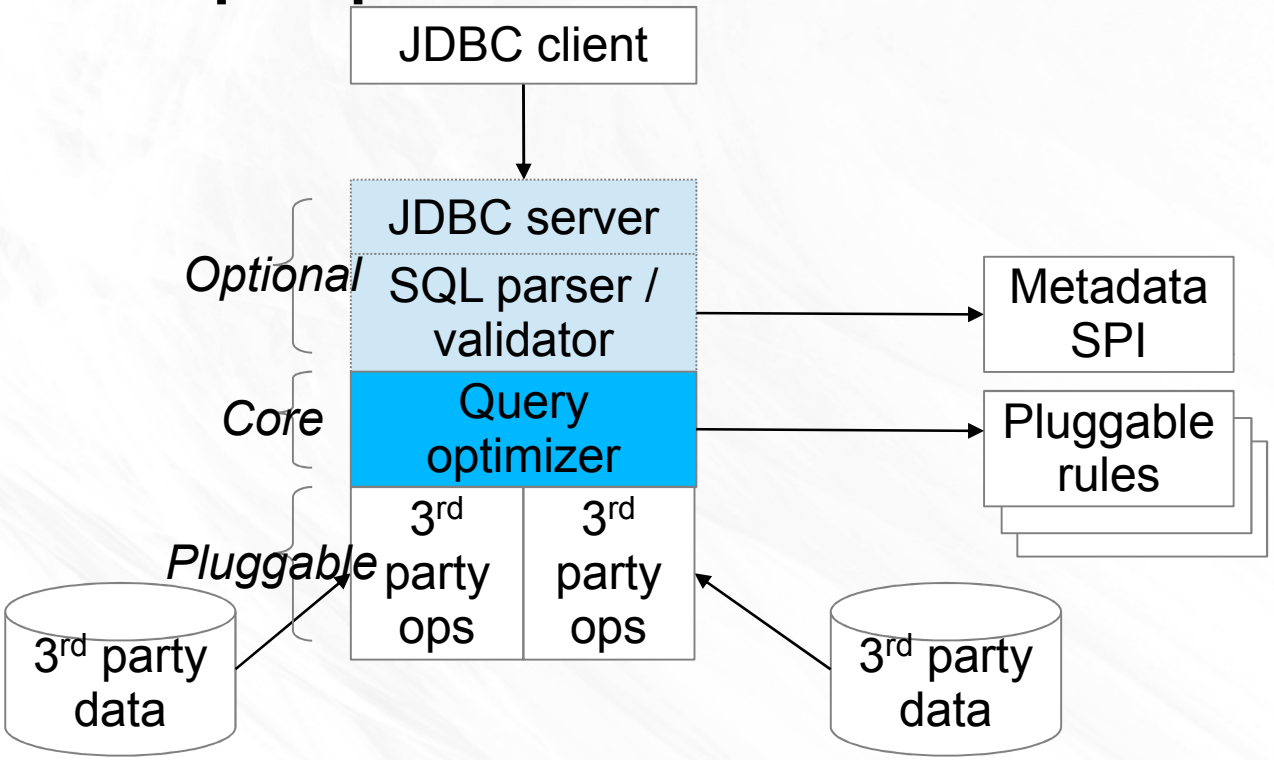
http://www.flickr.com/photos/torkildr/3462606643


http://www.flickr.com/photos/telstra-corp/5069403309/

Conventional database architecture



Optiq architecture



What is Optiq?

A really, really smart JDBC driver

Framework

Potential core of a data management system

Writing an adapter

Driver – if you want a vanity URL like “jdbc:splunk:”

Schema – describes what tables exist (Splunk has just one)

Table – what are the columns, and how to get the data. (Splunk's table has any column you like... just ask for it.)

Operators (optional) – non-relational operations

Rules (optional, but recommended) – improve efficiency by changing the question

Parser (optional) – to query via a language other than SQL


http://www.flickr.com/photos/walkercarpenter/4697637143/

Optiq roadmap ideas

Mondrian use Optiq to read from data sources such as Splunk & MongoDB, combine multiple data sources

Kettle integration: JDBC front-end; optimize jobs; push down filters & aggregations to data sources (e.g. SQL database)

Adapters: Cascading, MongoDB, Hbase, Apache Drill, …?

Front-ends: linq4j, Scala SLICK, Java8 streams

Contributions

Conclusions

Liberate your data!

Optiq is a framework

Build & share Optiq adapters

Questions?

@julianhyde

http://julianhyde.blogspot.com

http://github.com/julianhyde/optiq

http://github.com/julianhyde/optiq-splunk

Additional material: The following queries were used in the demo

```
select s."source", s."sourcetype"
    from "splunk"."splunk" as s;
```

```
select s."source",
    s."sourcetype", s."action"
    from "splunk"."splunk" as s
where s."action" = 'purchase';
```

```
select s."source",
    s."sourcetype", s."action"
```

```
select * from "mysql"."products";
```

```
select p."product_name",
    s."action"
from "splunk"."splunk" as s
 join "mysql"."products" as p
 on s."product_id" =
    p."product_id";
```