



A **very** short talk about Apache Kylin Business Intelligence meets Big Data

Fabian Wilckens
EMEA Solutions Architect



The challenge today ...



Very quickly: OLAP

Online Analytical Processing

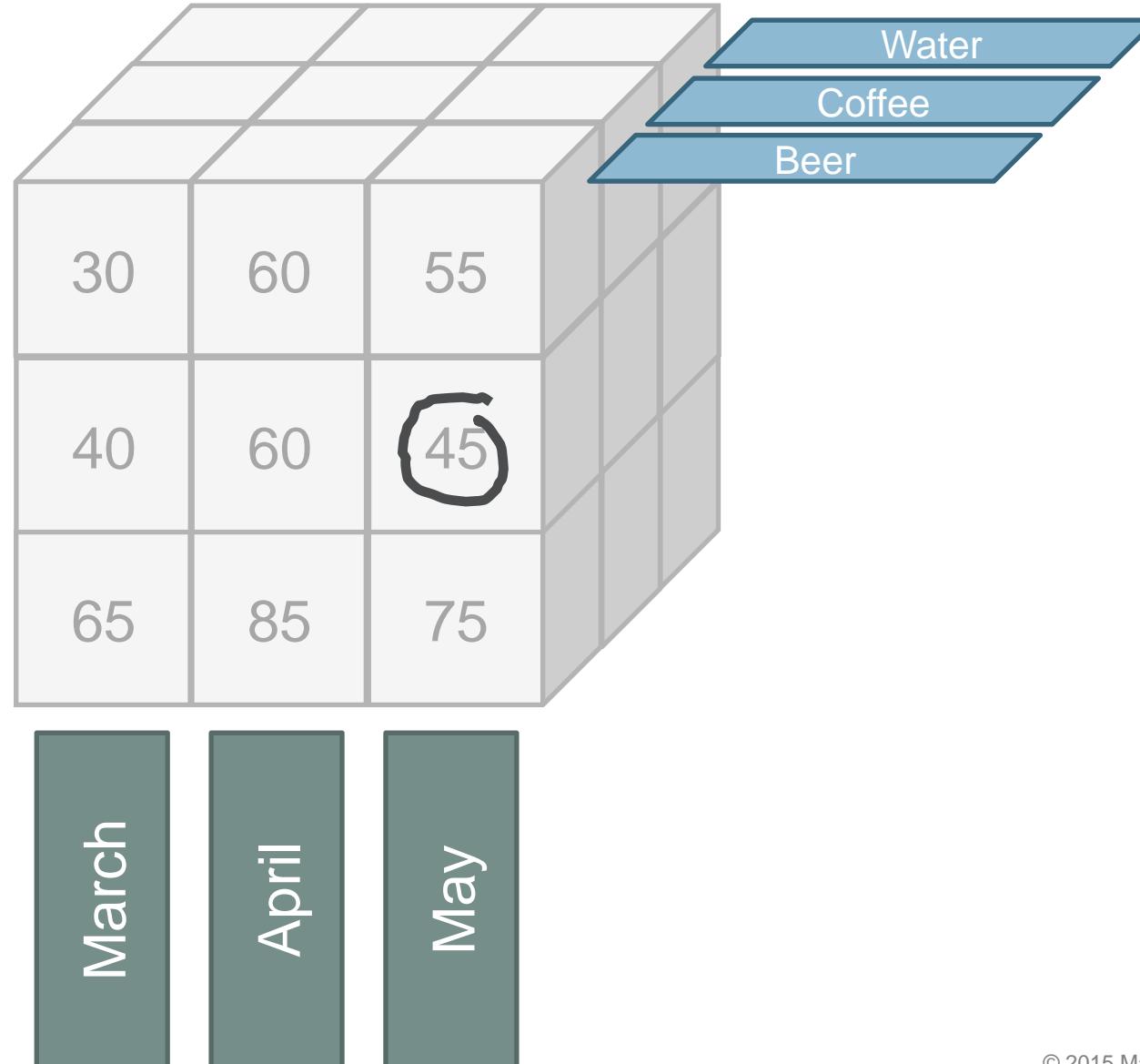


How many beers were ordered in Germany on a yearly basis?

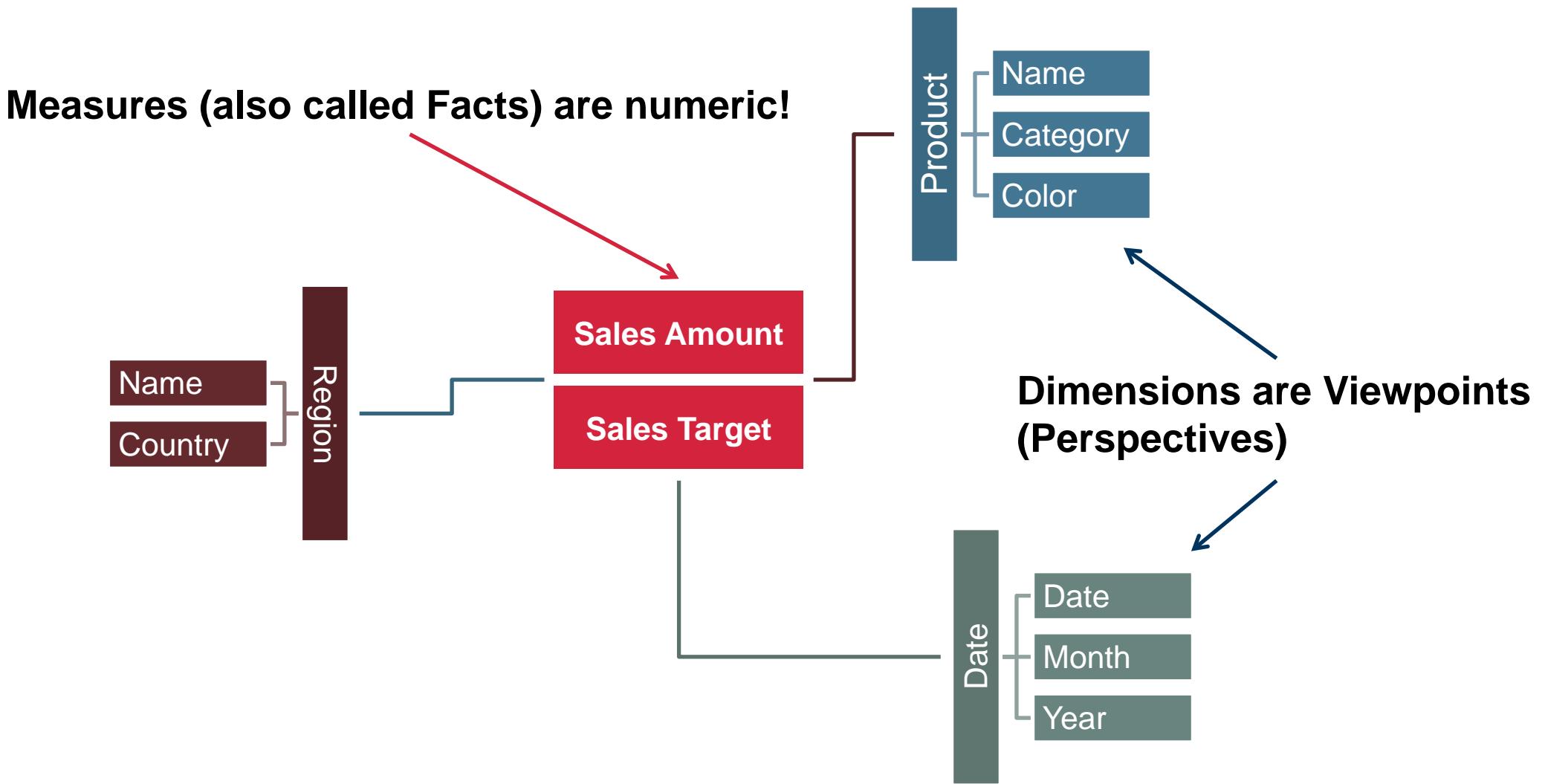
How many beers were ordered in Germany on a yearly basis,
broken down by months?

OLAP Cubes

Berlin
Hamburg
Munich

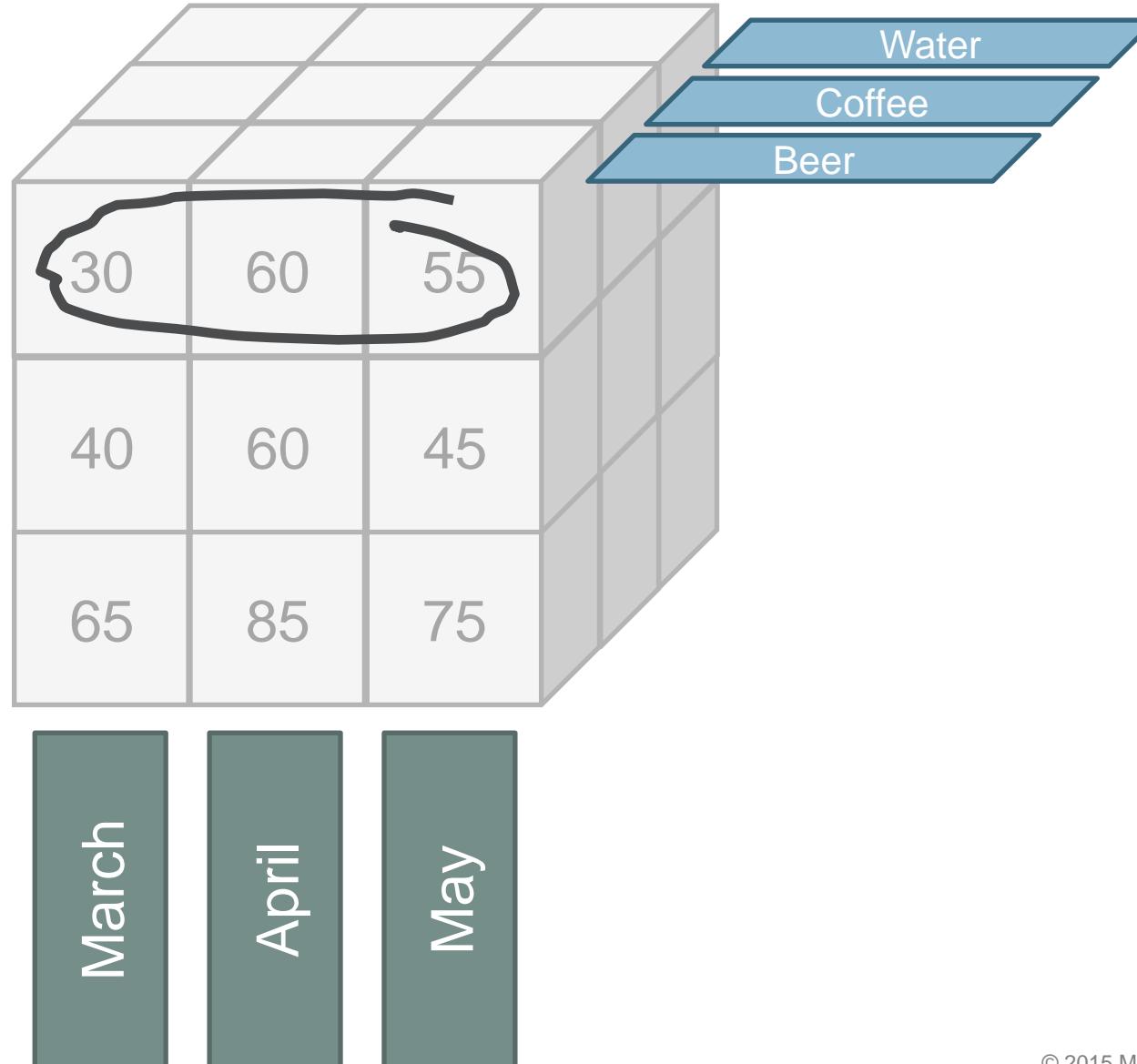


Cubes: Measures and Dimensions



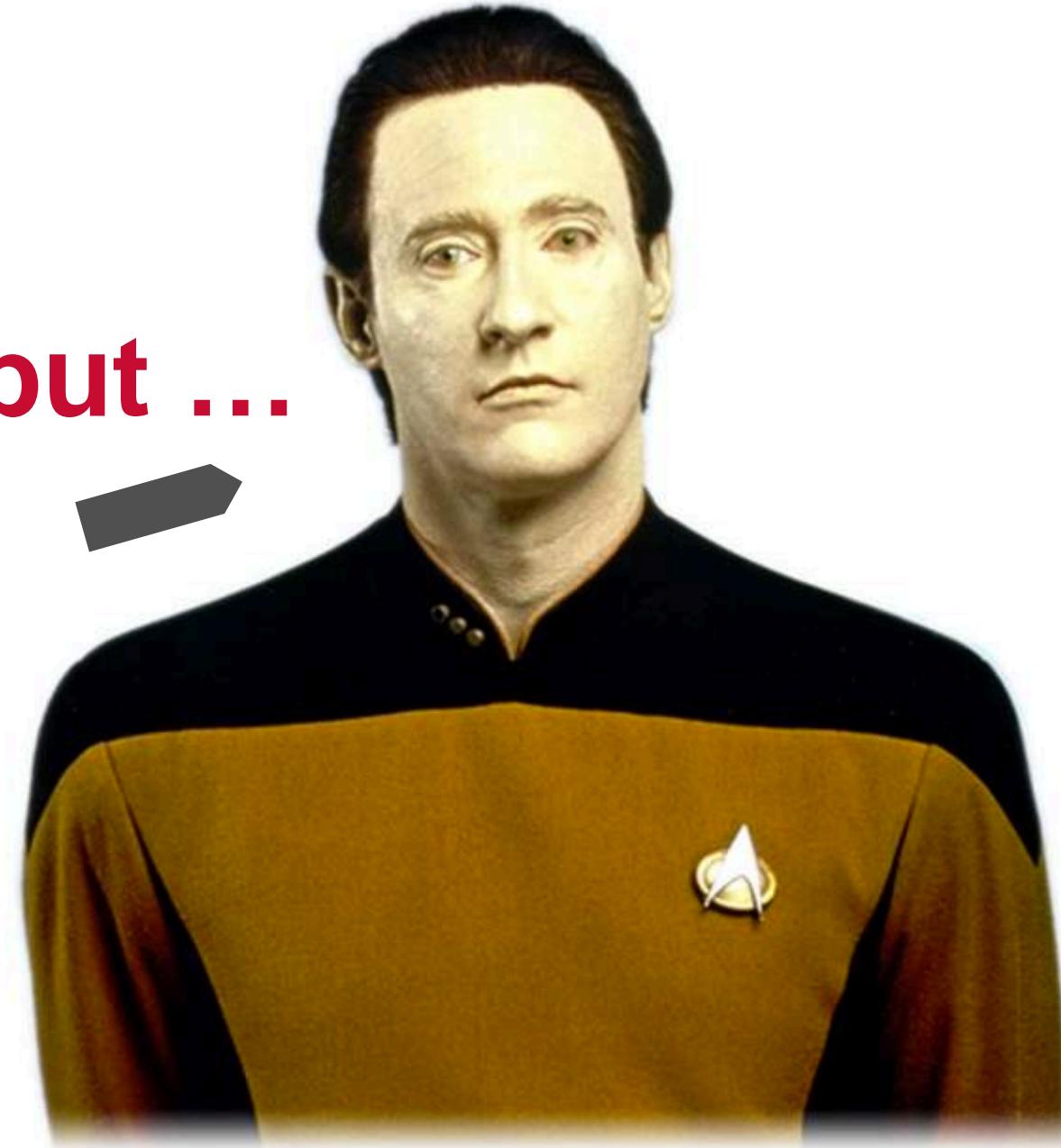
OLAP Cubes

Berlin
Hamburg
Munich



That's all great but ...

How about Big Data



What is Kylin?

kylin / 'ki:lɪn / 麒麟

--n. (in Chinese art) a mythical animal of composite form



Extreme OLAP Engine for Big Data

Kylin is an open source Distributed Analytics Engine from (originally from eBay) that provides SQL interface and multi-dimensional analysis (OLAP) on Hadoop for extremely large datasets

- Open Sourced on Oct 1st, 2014
- Accepted into incubation November, 2014
- Preparing for first Apache release

Goals

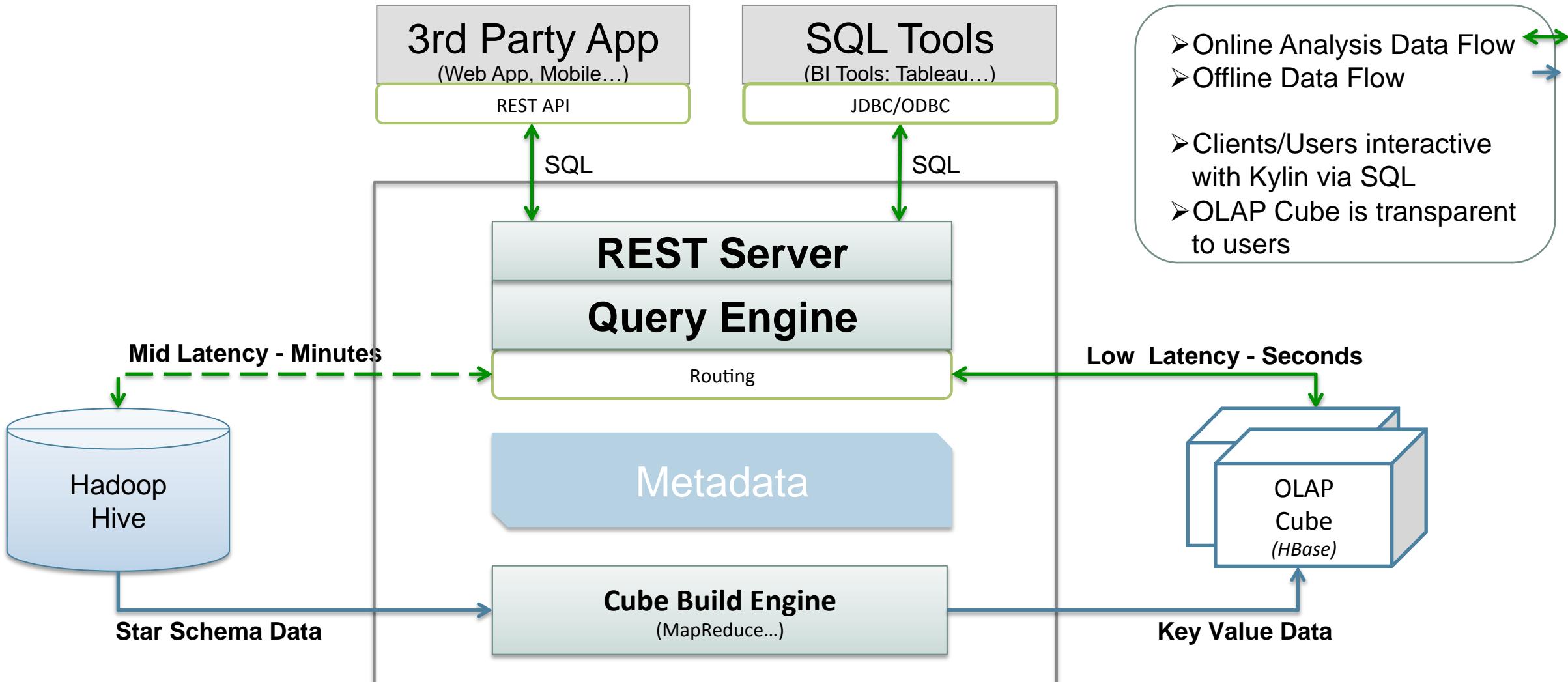
- Sub-second query latency on billions of rows
 - ANSI SQL for both analysts and engineers
 - Full OLAP capability to offer advanced functionality
 - Seamless Integration with BI Tools
-
- Support for high cardinality and dimensionality
 - High concurrency – thousands of end users
 - Distributed and scale out architecture for large data volume

Kylin Depends on Hadoop Eco-system

- Hive
 - Input source, pre-join star schema during cube building
- MapReduce
 - Aggregate metrics during cube building
- HDFS
 - Store intermediate files during cube building
- HBase
 - Store and query data cubes
- Calcite
 - SQL parsing, code generation, optimization



Kylin Architecture Overview



Kylin Highlights

- ***Extremely Fast OLAP Engine at Scale***

Kylin is designed to reduce query latency on Hadoop for 10+ billions of rows of data to seconds

- ***ANSI SQL Interface on Hadoop***

Kylin offers ANSI SQL on Hadoop and supports most ANSI SQL query functions

- ***Seamless Integration with BI Tools***

Kylin currently offers integration capability with BI Tools like Tableau.

- ***Interactive Query Capability***

Users can interact with Hadoop data via Kylin at sub-second latency

- ***MOLAP Cube***

User can define a data model and pre-build in Kylin with more than 10+ billions of raw data records



More Highlights

- Compression and Encoding Support
- Incremental Refresh of Cubes
- Approximate Query Capability for distinct Count (HyperLogLog)
- Leverage HBase Coprocessor for query latency
- Job Management and Monitoring
- Easy Web interface to manage, build, monitor and query cubes
- Security capability to set ACL at Cube/Project Level
- Support LDAP Integration



Cube Designer

Kylin Query Cubes Jobs Tables Admin Help Welcome, AD

Source Tables Cube Designer

test_kylin_cube_with_slr_empty READY 6.52 MB 10,000 2014-11-11 13:59:15 Action Action

Grid Visualization SQL JSON Access Notification HBase

The diagram illustrates the data flow for the cube. It starts with the source table **TEST_KYLIN_FACT**, which is connected via inner joins to four dimension tables: **TEST_CAL_DT**, **TEST_CATEGORY_GROUPINGS**, **TEST_SITES**, and **TEST_SELLER_TYPE_DIM**.

Setting Advanced Setting Overview

Tips

1. Please indicate which type for refresh model
2. Leave as default if this cube always need full build
3. Please indicate partition column of Fact Table in Hive
4. Partition column accept expression like: concat(year, '-', month, '-', day)
5. Please indicate start date to just pull certain data from source

◀ Prev Next ▶

Job Management

The screenshot displays a composite view of Kylin's Job Management interface and its underlying Hadoop infrastructure.

Top Left: A large window titled "CUBE BUILD CONFIRM" shows the configuration for a build. It includes a "PARTITION DATE COLUMN" set to "test_kylin_fact.cal_dt", a "START DATE" of "Thursday, January 1, 1970", and a "END DATE". Below this, a table lists two jobs:

Job Name	Duration	Actions
test_kylin_cube_with_slr_empty - 19700101000000_19710101000000 - BUILD - PST 2014-11-05 03:01:20	0.00 mins	Action
test_kylin_cube_without_slr_empty - FULL_BUILD - BUILD - PST 2014-11-05 01:28:57	24.07 mins	Action

Bottom Left: A summary box provides details about the latest job:
SequenceID: 0
Status: FINISHED
Duration: 5.44 mins
Waiting: 0 seconds
Start At: 2014-11-05 17:29:08
End At: 2014-11-05 17:34:34
Data Size: 446.61 KB
MR Job:
job_1415168056392_0001

Middle Right: A modal window titled "Detail Information" shows the same job details as the summary box.

Bottom Right: A separate window titled "MapReduce Job job_1415168056392_0001" displays the Hadoop job's progress and logs. The logs show:

```
Logging initialized using configuration in file:/  
OK  
Time taken: 7.782 seconds  
OK  
Time taken: 3.618 seconds  
Query ID = root_20141105012929_23aldace-bo6b-4af  
Total jobs = 1  
14/11/05 01:31:13 WARN conf.Configuration: file:/  
14/11/05 01:31:13 WARN conf.Configuration: file:/  
Execution log at: /tmp/root/root_20141105012929_23aldace-bo6b-4af  
2014-11-05 01:31:18 Starting to launch local
```

The Hadoop interface also shows the Application Master's tasks and their status.

Query and Visualization

Results (1215)

Drag a column header here and drop it to group by that column.

LSTG_FORMAT...	WEEK_BEG_DT	META_CATEG...	PRICE
FP-GTC	2013-01-01	Sports McCard...	93.268450871...
Auction	2013-01-01	Business & Ind...	55.635632631...
Others	2013-01-01	Sporting Goods	82.533676494...
Auction	2013-01-01	Real Estate	20.347844733...
Auction	2013-01-01	Toys & Hobbies	4.0906740147

Source Table: EDW
TEST_CAL, TEST_CATE, TEST_KYLIN, TEST_SELL, TEST_SITE

EDW

Visualization Export

MIN ▾

Results (1215)

Status: All Start Time: 2014-11-12 17:33:48 Duration: 3.10s [Rerun](#) [Save](#)

Query String: Status: Success Project: onlyinner Cubes: test_kylin_cube_with_slr_empty

Results (1215)

Graph Type: Pie Chart

Dimensions: LSTG_FORMAT...

Metrics: PRICE

Pie Chart showing the distribution of metrics across five categories: ABIN, Auction, FP-GTC, FP-non GTC, and Others.

Results (1215)

Status: All Start Time: 2014-11-12 17:33:48 Duration: 3.10s [Rerun](#) [Save](#)

Query String: Status: Success Project: onlyinner Cubes: test_kylin_cube_with_slr_empty

Results (1215)

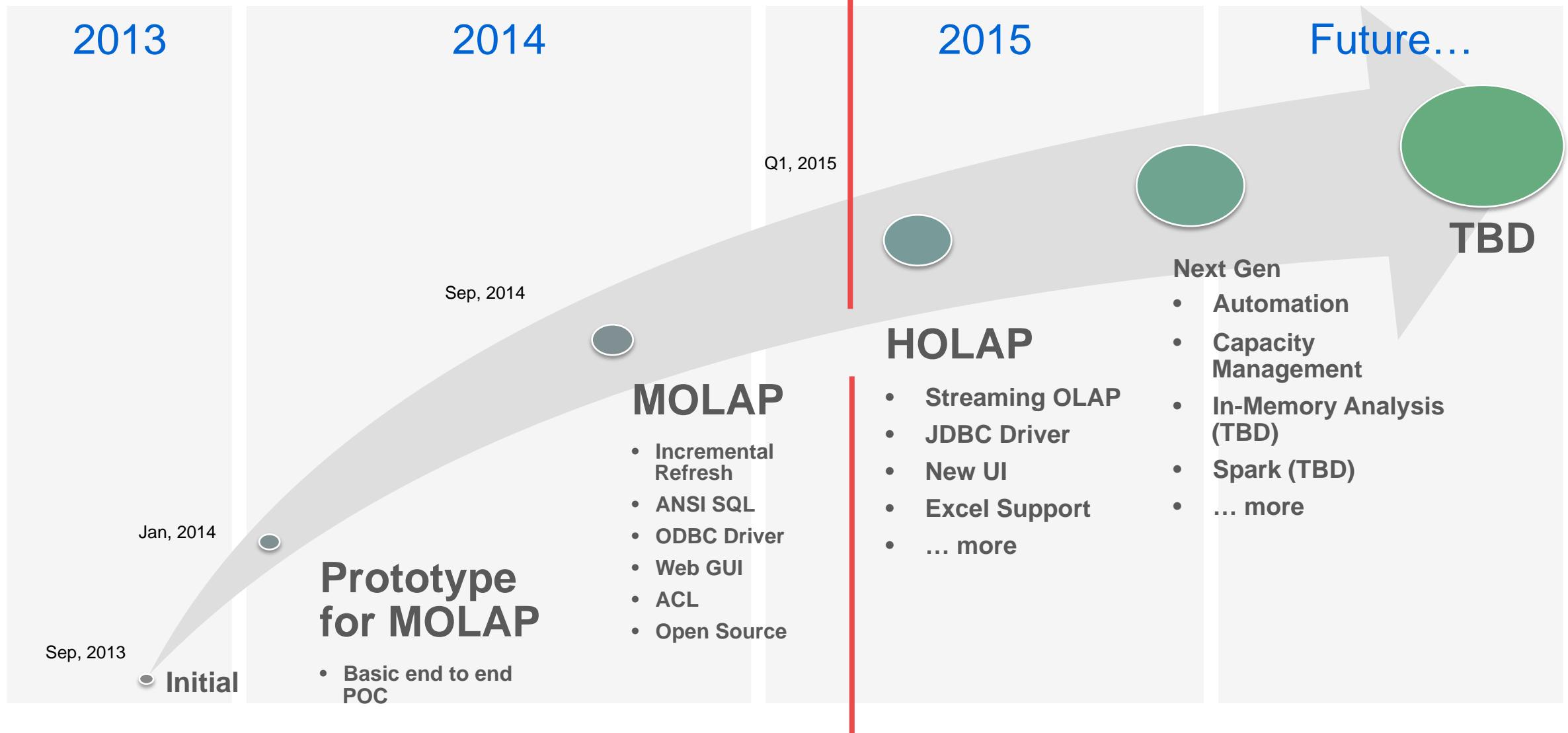
Graph Type: Line Chart

Dimensions: WEEK_BEG_DT

Metrics: PRICE

Line Chart showing the trend of Price over time from January 1, 2013, to May 26, 2013. The price starts at approximately 4000, peaks at about 7311885 on January 1, dips, and then fluctuates between 4000 and 6000.

Kylin History and Roadmap



Open Source

- Kylin Site:
 - <http://kylin.io>
- Twitter:
 - [@ApacheKylin](https://twitter.com/ApacheKylin)
- Source Code Repo:
 - <https://github.com/KylinOLAP>
- Google Group:
 - [Kylin OLAP](#)
- 微信
 - ApacheKylin





Fabian Wilckens
fwilckens@mapr.com



Free on-demand
Hadoop training leading to certification
Start becoming an expert now
mapr.com/training

